ABSTRACT

Title of dissertation: TOPICS IN MODEL-ASSISTED POINT AND VARIANCE ESTIMATION IN CLUSTERED SAMPLES

Timothy L. Kennel, Doctor of Philosophy, 2013

Dissertation directed by: Professor Richard Valliant
Joint Program in Survey Methodology

This dissertation describes three distinct research papers. Although each research topic is different and there is very little binding some of the chapters together, all three deal with innovations to model-assisted estimators. Moreover, all three papers explore different aspects of estimating totals, means, and rates from clustered samples. New estimators are presented. Their theoretical properties are explored; and, simulations are used to explore their design-based properties in realistic situations.

After an introductory chapter, we show how leverage adjustments can be made to sandwich variance estimators to improve variance estimates of Generalized Regression estimators in two-staged samples. In the third chapter, we explore multinomial logistic-assisted estimators of finite population totals in clustered samples. In the final chapter, we use generalized linear models to assist estimating finite population totals in cluster samples.

TOPICS IN MODEL-ASSISTED POINT AND VARIANCE ESTIMATION IN CLUSTERED SAMPLES

by

Timothy L. Kennel

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

Advisory Committee:
Professor Richard Valliant, Chair/Advisor
Professor Robert Croninger
Professor Michael R. Elliott
Professor Partha Lahiri
Dr. Tommy Wright

Preface

On the most practical sense, I wrote this dissertation to fulfill the requirements needed to graduate. But on another level, I have written this dissertation in thanks to the many great survey statisticians who have established and developed a wonderful and exciting field of study. From the great legends of Neyman, Kish, Cochran, Royall, Hansen, ... to the contemporary geniuses who have carefully written, edited, and creatively dreamed of new advances in the field of survey statistics, I am most grateful and humbled by the growth of this field.

This dissertation extensively draws upon the work of two great statisticians who passed on while I was writing this dissertation. I owe much to the landmark papers of great survey statisticians such as David Binder and Randy Sitter. Their work, along with their students and coauthors, has greatly influenced and shaped this dissertation.

Acknowledgments

I would not have been able to write this dissertation without the most general support of many kind-hearted, generous, and patient individuals.

First, and foremost, I wish to thank the U.S. Census Bureau for their most generous support. In addition to paying for my textbooks and tuition, they have given me time off to work to attend to my coursework and dissertation research while paying me my full salary. My bosses, Xijian Liu and Dawn Haines, have always respected my studies and my non-traditional schedule. It really has been a pleasure and joy to work for such flexible, empathetic, trusting, and generous managers. The Census Bureau truly is one of the greatest work-places and I feel greatly satisfied and honored to work there. In addition to my immediate supervisors, I wish to thank Cynthia C.Z. Clark for publicizing the JPSM program at the Census Bureau and enticing me into the Master's Degree program. Later Alan Tupek, Ruth Ann Killion, and the Methodology and Standards Council at the Census Bureau opened many doors for me to be the first Census Bureau employee to pursue a Ph.D. in Survey Methodology at the University of Maryland.

There are numerous professors at the Joint Program in Survey Methodology who have encouraged and guided me along the way. Early on, Katherine Abrahams, Roger Tourangeau, and Frauke Kreuter encouraged me to pursue Ph.D. in Survey Methodology. At the request of Richard Valliant, I took a series of statistics classes with Abram Kagan and Paul Smith at the University of Maryland. I owe much to Dr. Kagan and Dr. Smith for teaching me the foundations of classical statistics. Partha Lahiri's Small Area Estimation class did an excellent job firming my understanding of the design-based and model-based

statistical frameworks. I highly recommend all classes by those three professors. Without their expert lectures and challenging assignments, I would not have been able to write a statistical dissertation. Lastly, I wish to thank Robert Groves for building the JPSM program, passionately discussing the field of Survey Methodology with me, and for his candid advice and enthusiasm.

Merritt Gardner, my undergraduate, mathematics professor fostered my mathematical intuition and gave me many tools which more than adequately prepared me for my career and further studies. Professors Thomas Meyers and Duane Stoltzfus first introduced me to survey research methods and helped me to approach survey research from a scientific prospective.

Richard Valliant is a remarkable, patient, and superb advisor. He has helped me in numerous ways, from forming a dissertation topic to teaching me basic statistical techniques and the R computer language. I owe him much for carefully reading far too many drafts of my dissertation and clarifying the deep and rich field of statistics for me.

My dear wife has endured many of my academic escapades with a positive spirit and energy. She has given me the social and emotional support that I needed to persist in my studies. She was always extremely understanding and supportive of me on the hundreds of evenings and weekends that I dedicated to my doctoral work.

I feel very blessed to be surround by a community of supports who have worked tirelessly and often behind the scenes train and encourage me to be a Survey Methodologist. I look forward to helping future generations of students and colleagues.

It is impossible to remember all, and I apologize to those I've inadvertently left out.

Introduction

This dissertation describes three distinct research papers. Chapters 2, 3, and 4 separately introduce each of the three papers and research projects. When describing each topic, we have tried to clearly state the need for new research and our research plans.

Although each research topic is different and there is very little fusing some of the topics, all three deal with innovations to model-assisted estimators. Moreover, all three papers explore different aspects of estimating totals, means, and rates from complex surveys. Our methods also provide an underlying strategy used across all three papers. We construct new estimators, explore their theoretical properties, and use simulations to explore their behavior in common situations. All three papers also deal with clustered samples.

Every year, governments produce millions of estimates of totals, means, rates, and their standard errors based on carefully designed complex surveys. Many of these statistics are model-assisted estimators or derivations of model-assisted estimators. The primary motivation of all three papers is the search for improved estimation techniques that could improve the efficiency or accuracy of estimated descriptive statistics from complex surveys. Every estimation of a descriptive parameter like a total or mean has an implied underlying model for which it works best. If the implied model is incorrect, there is the potential for improvement by using an estimator whose model fits the data better. Much if this thesis looks at such improved estimators and how to estimate their variances.

The first paper applies the theory of leverage adjustments to create new and potentially better estimates of standard errors from cluster samples. A special focus of this

project is its heavy use of model-based theory to create design-consistent variance estimators.

In the past fifteen years, sampling frames in the United States have become incredibility rich with the availability of large national addresses lists derived from the Delivery Sequence File from the United States Postal Service. Such address lists have made it possible to use frame data to increase the precision of estimators. In the second and third papers, we investigate new estimation techniques for categorical data that use complete frame data to improve estimation.

In the second paper, we explore estimating finite population totals using a multinomial logistic assisting model. In a sense, the second paper is a specialization of the third paper, which focuses on generalized linear assisting models. Using nonlinear models to assist estimating totals has serious limitations; however, our research shows that the gains in precision from using nonlinear assisting models can greatly improve some estimates.

Although the focus of each paper is different, all three offer practical alternatives to common techniques used to estimate totals and standard errors from cluster samples. We hope that the research outlined in the following sections will spur further research aiming to bridge the divide between the model-based and design-based frameworks.

Contents

List of Tables xv

List of Figures xix

1 Introduction 1
- 1.1 Design-Based Framework . . . 2
  - 1.1.1 Brief History of the Design-Based Framework . . . 2
  - 1.1.2 Sample Design . . . 4
    - 1.1.2.1 Notation . . . 5
  - 1.1.3 Sampling Theory . . . 6
  - 1.1.4 Data . . . 12
  - 1.1.5 Estimation . . . 14
    - 1.1.5.1 With-Replacement Estimators . . . 14
    - 1.1.5.2 The π-Estimator . . . 16
  - 1.1.6 Empirical Properties of Design-Based Estimators . . . 21
  - 1.1.7 Theoretical Properties of Design-Based Estimators . . . 26
  - 1.1.8 Discussion . . . 30
- 1.2 Model-Assisted Frameworks . . . 33
  - 1.2.1 Generalized Regression Estimator . . . 33
    - 1.2.1.1 Point Estimator . . . 34
    - 1.2.1.2 Variance Estimator . . . 37
    - 1.2.1.3 Point Estimation in Two Staged Samples . . . 42
    - 1.2.1.4 Variance of GREG in Two Staged Samples . . . 43
  - 1.2.2 Generalized Difference Estimator . . . 47
  - 1.2.3 Calibrated Estimator . . . 48
  - 1.2.4 Model-Calibrated Estimator . . . 50
  - 1.2.5 Model-Calibrated Pseudoempirical Maximum Likelihood Estimator 52
  - 1.2.6 Pseudo Maximum Likelihood Estimation . . . 53
- 1.3 Conclusion . . . 56

2 Improved Variance Estimators for Generalized Regression (GREG) Estimators in Cluster Samples 57
- 2.1 Introduction . . . 57
- 2.2 Literature Review . . . 59
  - 2.2.1 Introduction . . . 60
  - 2.2.2 Linear Models . . . 62
    - 2.2.2.1 Parametrization . . . 62
    - 2.2.2.2 The Hat Matrix and Leverages . . . 64
    - 2.2.2.3 Residuals . . . 66
  - 2.2.3 Prediction Estimators . . . 68
  - 2.2.4 Discussion . . . 76
- 2.3 Theoretical Results . . . 79
- 2.4 Simulation . . . 86

2.4.1 Data . . . 87
2.4.1.1 Third Grade Population . . . 88
2.4.1.2 American Community Survey Population . . . 90
2.4.1.3 Simulated Population . . . 93
2.4.2 Results . . . 95
2.4.2.1 υ_g . . . 100
2.4.2.2 υ_{wr} and υ_{JL} . . . 103
2.4.2.3 υ_r and $\upsilon_r^\star$ . . . 104
2.4.2.4 υ_D and $\upsilon_D^\star$ . . . 105
2.4.2.5 υ_{Jack}, υ_J, υ_{J1}, $\upsilon_{Jack}^\star$, $\upsilon_J^\star$, and $\upsilon_{J1}^\star$ . . . 106
2.4.2.6 Summary . . . 108
2.5 Conclusion . . . 110

3 Multivariate Logistic-Assisted Estimators of Totals from Clustered Survey Samples in the Presence of Complete Auxiliary Information 112
3.1 Introduction . . . 112
3.1.1 Multinomial Logistic Regression . . . 114
3.1.2 Estimation of Totals for Multinomial Data in Poisson Samples . . 120
3.1.2.1 π-Estimator . . . 120
3.1.2.2 Generalized Difference Estimator . . . 121
3.1.2.3 Calibrated Estimator . . . 124
3.1.2.4 Model-Calibrated Estimator . . . 126
3.1.2.5 Model-Calibrated Maximum Pseudo Empirical Likelihood Estimator . . . 128
3.2 Main Results . . . 130
3.2.1 Generalized Difference Estimator . . . 130
3.2.1.1 Multivariate GREG . . . 131
3.2.1.2 Multinomial LGREG in Clustered Samples . . . 131
3.2.2 Model-Calibrated Estimator . . . 137
3.2.3 Model-Calibrated Maximum Pseudo Empirical Likelihood Estimator . . . 141
3.3 Simulation . . . 143
3.3.1 Populations . . . 145
3.3.1.1 Synthetic Population . . . 145
3.3.1.2 Postsecondary Majors Population . . . 146
3.3.1.3 Census Population . . . 149
3.3.2 Simulation Design . . . 151
3.3.2.1 Sample Design . . . 151
3.3.2.2 Number of Samples . . . 152
3.3.2.3 Estimation . . . 154
3.3.2.4 Measures . . . 155
3.3.3 Results . . . 157
3.3.3.1 Simulation Errors . . . 157
3.3.3.2 Point Estimators: Average Distance from True Value . . 159
3.3.3.3 Point Estimators: Mean Squared Error . . . 164

3.3.3.4 Point Estimators: Percent Relative Bias 168
3.3.3.5 Point Estimators: Summary Across All Populations . . 174
3.3.3.6 Variance Estimators of $\widehat{\mathbf{t}}_y^{lg}$ 176
3.3.3.7 Variance Estimators of $\widehat{\mathbf{t}}_{yc}^{mc}$ and $\widehat{\mathbf{t}}_{yc}^{peM}$ 185
3.4 Conclusion 190

4 Design-based Inference Assisted by Generalized Linear Models in Cluster Samples 193
4.1 Introduction 193
4.1.1 Generalized Linear Models 196
4.1.1.1 Likelihood 197
4.1.1.2 Link Functions 199
4.1.1.3 Parameter Estimation 200
4.1.1.4 Summary 204
4.1.2 Estimation of Totals for Categorical Data in Poisson Samples . . 204
4.1.2.1 The π Estimator 205
4.1.2.2 Projective Estimator 205
4.1.2.3 GLM-Assisted Difference Estimator 206
4.1.2.4 Calibrated Estimator 209
4.1.2.5 Model-Calibrated Estimator 209
4.1.2.6 Model-Calibrated Maximum Pseudoempirical Likelihood Estimator 210
4.2 Main Results 212
4.2.1 GLM-Assisted Difference Estimator 212
4.2.2 Model-Calibrated Estimator 218
4.2.3 Model-Calibrated Maximum Pseudoempirical Likelihood Estimator 222
4.3 Simulation 224
4.3.1 Population: 2000 Tract Level Planning Database 226
4.3.2 Models 229
4.3.3 Simulation Design 234
4.3.3.1 Sample Design 234
4.3.3.2 Number of Samples 235
4.3.3.3 Estimation 236
4.3.3.4 Measures 238
4.3.4 Results 239
4.3.4.1 Simulation Errors 239
4.3.4.2 Point Estimators 240
4.3.4.3 Variance Estimators for $\widehat{t}_y^{lg}$ 246
4.3.4.4 Variance Estimators for $\widehat{t}_y^{mc}$ and $\widehat{t}_y^{peM}$ 249
4.4 Conclusion 253

5 Conclusion 255

A Notes for Variance of Clustered GREG Paper 258
A.1 Some Asymptotic Results . . . 258
A.1.1 Proof that $\mathbf{A}_\pi = O(N)$ . . . 260
A.1.2 Proof that $\mathbf{g}_i = O(1)$ . . . 261
A.1.3 Proof that $\mathbf{H}_{ii} = O(n^{-1})$ . . . 262
A.1.4 Proof that $\mathscr{Q}_i = O(n^{-1})$ . . . 263
A.1.5 Proof that $G_i - \overline{G} \approx -\mathbf{1}^\top \mathbf{\Pi}_i^{-1} \widehat{\mathbf{y}}_i + \frac{1}{n} \sum_{i=1}^n \mathbf{1}^\top \mathbf{\Pi}_i^{-1} \widehat{\mathbf{y}}_i$ . . . 264
A.1.6 Proof that $K_i - \overline{K} \approx -n\mathbf{1}^\top \mathbf{\Pi}_i^{-1} \mathbf{X}_i \widehat{\mathbf{B}} + \sum_{i=1}^n \mathbf{1}^\top \mathbf{\Pi}_i^{-1} \mathbf{X}_i \widehat{\mathbf{B}}$ . . . 265
A.1.7 Proof that $F_i = o(1)$ . . . 266
A.1.8 Proof that $D_i = O\left(\frac{N}{n}\right)$ . . . 267
A.1.9 Proof that $\mathrm{var}_M(\mathbf{e}_i) \approx \mathbf{\Psi}_i$ . . . 268
A.2 Derivation of Sample Hat Matrix for Clustered GREG . . . 269
A.3 Model Variance of Clustered GREG . . . 272
A.4 Approximate Model Expectation of Sandwich Estimator for the Clustered GREG . . . 275
A.5 Delete-a-cluster Jackknife . . . 277
A.5.1 Proof that $\widehat{\mathbf{B}}_{(i)} = \widehat{\mathbf{B}} - \mathbf{Q}_i$ for cluster samples . . . 277
A.5.2 Jackknife variance estimator of clustered GREG in terms of leverages . . . 280
A.5.3 Jackknife variance estimator of clustered GREG in large samples 285
A.5.4 Further simplification for Jackknife variance estimator of clustered GREG in large samples . . . 286
A.6 Full Tables . . . 288
A.7 R code . . . 293

B Notes for LGREG Paper 297
B.1 Some Asymptotic Results . . . 297
B.2 Logistic Models . . . 299
B.2.1 Data . . . 300
B.2.2 Density Function . . . 303
B.2.3 Logistic Link Function . . . 306
B.2.4 Likelihood Equations . . . 308
B.2.5 Estimating Equations . . . 309
B.2.6 Residuals . . . 312
B.3 Derivation of Multivariate Calibration GREG . . . 314
B.4 LGREG . . . 317
B.4.1 Design Consistency of the Clustered LGREG Estimator . . . 317
B.4.2 Asymptotic variance of LGREG estimator . . . 319
B.4.3 Variance estimators of LGREG . . . 322
B.4.3.1 Linear substitute estimator . . . 322
B.4.3.2 With-replacement estimator . . . 324
B.4.3.3 Implicit differentiation variance estimator . . . 325
B.5 Model Calibration . . . 336
B.5.1 Construction of model calibration estimator . . . 336
B.5.2 Alternative forms of the model calibration estimator . . . 338

B.5.3 Design consistency of model calibration estimator 339
B.5.4 Asymptotic variance of model calibration estimator 342
B.5.5 Variance estimators of model calibration 346
B.5.5.1 Weighted variance estimator 346
B.5.5.2 With-replacement estimator 347
B.5.5.3 Implicit differentiation estimator 348
B.6 Model-Calibrated Maximum Pseudo Empirical Likelihood Estimator . . . 361
B.6.1 Estimation of Model-Calibrated Maximum Pseudo Empirical Likelihood Estimator 361
B.6.2 $\bar{t}_y^{peM}$ is asymptotically equal to $\bar{t}_y^{mc}$ 365
B.7 Simulation Results 374
B.7.1 Synthetic Population 374
B.7.1.1 Percent Simulation Coefficient of Variation Table . . . 374
B.7.1.2 Average Distance from True Value 375
B.7.1.3 Empirical Standard Deviation of Distance from True Value 376
B.7.1.4 Percent Relative Bias Table 377
B.7.1.5 Percent Relative Median Difference Table 378
B.7.1.6 Percent Relative Root Mean Squared Error Table 379
B.7.1.7 Percent Relative Root Median Squared Error Table . . . 380
B.7.1.8 Percent Relative Bias Table for LGREG Variance Estimators 381
B.7.1.9 Percent Relative Median Difference Table for Variance Estimators 382
B.7.1.10 Percent Relative Root Mean Squared Error Table for LGREG Variance Estimators 383
B.7.1.11 Percent Relative Root Median Squared Error Table for LGREG Variance Estimators 384
B.7.1.12 Average Distance from Empirical Value for Variance Estimators 385
B.7.1.13 Median Distance from Empirical Value for Variance Estimators 386
B.7.1.14 Standard Error of Average Distance from Empirical Value for Variance Estimators 387
B.7.1.15 95% Confidence Interval Coverage Table for Variance Estimators 388
B.7.1.16 Plots 389
B.7.2 Post-Secondary Population 394
B.7.2.1 Percent Simulation Coefficient of Variation Table . . . 394
B.7.2.2 Average Distance from True Value 395
B.7.2.3 Empirical Standard Deviation of Distance from True Value 396
B.7.2.4 Percent Relative Bias Table 397
B.7.2.5 Percent Relative Median Difference Table 398
B.7.2.6 Percent Relative Root Mean Squared Error Table 399

B.7.2.7 Percent Relative Root Median Squared Error Table . . . 400
B.7.2.8 Percent Relative Bias Table for LGREG Variance Estimators . . . 401
B.7.2.9 Percent Relative Median Difference Table for Variance Estimators . . . 402
B.7.2.10 Percent Relative Root Mean Squared Error Table for LGREG Variance Estimators . . . 403
B.7.2.11 Percent Relative Root Median Squared Error Table for LGREG Variance Estimators . . . 404
B.7.2.12 Average Distance from Empirical Value for Variance Estimators . . . 405
B.7.2.13 Median Distance from Empirical Value for Variance Estimators . . . 406
B.7.2.14 Standard Error of Average Distance from Empirical Value for Variance Estimators . . . 407
B.7.2.15 95% Confidence Interval Coverage Table for Variance Estimators . . . 408
B.7.2.16 Plots . . . 409
B.7.3 Census Population . . . 414
B.7.3.1 Percent Simulation Coefficient of Variation Table . . . 414
B.7.3.2 Average Distance from True Value . . . 415
B.7.3.3 Empirical Standard Deviation of Distance from True Value . . . 416
B.7.3.4 Percent Relative Bias Table . . . 417
B.7.3.5 Percent Relative Median Difference Table . . . 418
B.7.3.6 Percent Relative Root Mean Squared Error Table . . . 419
B.7.3.7 Percent Relative Root Median Squared Error Table . . . 420
B.7.3.8 Percent Relative Bias Table for Variance Estimators . . . 421
B.7.3.9 Percent Relative Median Difference Table for Variance Estimators . . . 422
B.7.3.10 Percent Relative Root Mean Squared Error Table for Variance Estimators . . . 423
B.7.3.11 Percent Relative Root Median Squared Error Table for Variance Estimators . . . 424
B.7.3.12 Average Distance from Empirical Value for Variance Estimators . . . 425
B.7.3.13 Median Distance from Empirical Value for Variance Estimators . . . 426
B.7.3.14 Standard Error of Average Distance from Empirical Value for Variance Estimators . . . 427
B.7.3.15 95% Confidence Interval Coverage Table for Variance Estimators . . . 428
B.7.3.16 Plots . . . 429
B.8 R Code . . . 430

C Notes for GLM-Assisted Estimation Paper 437
C.1 Derivation of Estimating Equations for Poisson Regression 437
C.1.1 Exponential Family 437
C.1.2 Link Function 437
C.1.3 Log Likelihood 438
C.1.4 Estimating Equations 438
C.2 Derivation of Estimating Equations for Binary Probit Regression 439
C.2.1 Exponential Family 439
C.2.2 Mean 441
C.2.3 Variance 441
C.2.4 Link Functions 443
C.2.5 Estimating Equations 443
C.3 Residuals for GLMs 444
C.4 GLM-Assisted Difference Estimator 446
C.4.1 Design Consistency of the Clustered GLM-Assisted Difference Estimator 446
C.4.2 Asymptotic Variance of the GLM-Assisted Difference Estimator . 447
C.4.3 Variance Estimators of the GLM-Assisted Difference Estimator . 448
C.4.3.1 Linear Substitute Estimator 448
C.4.3.2 With-replacement Estimator 449
C.4.3.3 Implicit Differentiation Variance Estimator 449
C.5 Model Calibration 452
C.5.1 Construction of the Model-Calibrated Estimator 452
C.5.2 Alternative Forms of the Model-Calibrated Estimator 453
C.5.3 Design Consistency of the Model-Calibrated Estimator 454
C.5.4 Asymptotic Variance of the Model-Calibrated Estimator 455
C.5.5 Variance Estimators of the Model-Calibrated Estimator 456
C.5.5.1 Linear Substitute Variance Estimators 456
C.5.5.2 With-Replacement Estimator 457
C.5.5.3 Implicit Differentiation Estimator 458
C.6 Model-Calibrated Maximum Pseudoempirical Likelihood Estimator . . . 460
C.6.1 Estimation of the Model-Calibrated Maximum Pseudoempirical Likelihood Estimator 460
C.6.2 $\bar{t}_y^{peM}$ is Asymptotically Equal to $\bar{t}_y^{mc}$ 461
C.7 Simulation Results 463
C.7.1 Simulation Coefficient of Variation 463
C.7.1.1 Simulation Coefficient of Variation of Count Response 464
C.7.1.2 Simulation Coefficient of Variation of Binary Response 465
C.7.1.3 Simulation Coefficient of Variation of Synthetic Response 466
C.7.2 Graphs for Point Estimators 467
C.7.2.1 Point Estimators of Count Response in Small Samples . 468
C.7.2.2 Point Estimators of Count Response in Large Samples . 469
C.7.2.3 Point Estimators of Binary Response in Small Samples 470
C.7.2.4 Point Estimators of Binary Response in Large Samples 471

C.7.2.5 Point Estimators of Synthetic Response in Small Samples 472
C.7.2.6 Point Estimators of Synthetic Response in Large Samples 473
C.7.3 Tables for Point Estimators 474
C.7.3.1 Point Estimators of Count Response in Small Samples . 474
C.7.3.2 Point Estimators of Count Response in Large Samples . 475
C.7.3.3 Point Estimators of Binary Response in Small Samples 476
C.7.3.4 Point Estimators of Binary Response in Large Samples 477
C.7.3.5 Point Estimators of Synthetic Response in Small Samples 478
C.7.3.6 Point Estimators of Synthetic Response in Large Samples 479
C.7.4 Graphs for Variance Estimators 480
C.7.4.1 Variance Estimators of Count Response in Small Samples 481
C.7.4.2 Variance Estimators of Count Response in Large Samples 482
C.7.4.3 Variance Estimators of Binary Response in Small Samples 483
C.7.4.4 Variance Estimators of Binary Response in Large Samples 484
C.7.4.5 Variance Estimators of Synthetic Response in Small Samples 485
C.7.4.6 Variance Estimators of Synthetic Response in Large Samples 486
C.8 R Code 487
C.8.1 Generation of Synthetic Variable 487
C.8.2 Simulation Program 488

List of Tables

1.1 Formulas for summary measures 26

2.1 Statistics of Interest for Clustered GREG Variance Simulation 86
2.2 Simulation Design 87
2.3 Simulation Results of Variance Estimators for Clustered GREG Estimate . 96
2.4 Variability of Sandwich Estimators for School Population 101
2.5 Coverage of Sandwich Estimators 102

3.1 Point Estimators 121
3.2 Variance Estimators Calculated in Simulations 145
3.3 Quartiles for Synthetic Population 146
3.4 Quartiles for Postsecondary Population 148
3.5 Quartiles for Census Population 150
3.6 Simulation Design 152
3.7 Logistic Regression Estimating Equations 154
3.8 Summary of empirical distributions 156
3.9 Number of Errors Found in Each Simulation 158
3.10 Average Distance from True Value for Synthetic Population (in thousands) 160
3.11 Average Distance from True Value for Postsecondary Population (in thousands) 161
3.12 Percent Relative Root Mean Squared Error of Total Estimators for Synthetic Population 166
3.13 Percent Relative Root Mean Squared Error for Postsecondary Population . 167
3.14 Percent Relative Bias for Synthetic Population 169
3.15 Percent Relative Bias for Census Population 170
3.16 Percent Relative Bias for Postsecondary Population 172
3.17 Quartiles for Percent Relative Difference of Math Estimators with Sample of Fixed SRS in Postsecondary Population 173
3.18 Quartiles for Percent Relative Difference of Health Estimators with Sample of Fixed SRS in Postsecondary Population 174
3.19 Percent Relative Difference of LGREG Variance Estimators for Synthetic Population 178
3.20 Relative bias of LGREG Variance Estimators for Postsecondary Population 179
3.21 Average Distance from Empirical Value for Standard Error Estimators in Postsecondary Population (in thousands) 180
3.22 Relative bias of LGREG Variance Estimators for Census Population . . . 181
3.23 Percent 95% Confidence Interval Coverage of LGREG Variance Estimators for Synthetic Population 182
3.24 Percent 95% Confidence Interval Coverage of LGREG Variance Estimators for Postsecondary Population 183
3.25 Percent 95% Confidence Interval Coverage of Finite Population Total Using Several variance Estimators of $\widehat{\mathbf{t}}_{yc}^{mc}$ for Postsecondary Population . . . 187

3.26 Percent 95% Confidence Interval Coverage of Finite Population Total Using Several Variance Estimators of $\widehat{\mathbf{t}}_{yc}^{mc}$ for Census Population 188
3.27 Percent Relative Root Mean Squared Error of Variance Estimators for Census Population 189

4.1 Distributions of the Exponential Family 198
4.2 Common Link Functions. Φ is the cumulative normal distribution function and $\mathscr{C}$ is the cumulative cauchy distribution function. 200
4.3 Point Estimators 205
4.4 Variance Estimators Calculated in Simulations 226
4.5 Edits for Tract Population 228
4.6 Quartiles for Tract Level Planning Dataset Population 229
4.7 Comparison of Finite Population Predictions when $\mathbf{B}$ is Known 233
4.8 Simulation Design 235
4.9 Simulation Design 237
4.10 Number of Errors Found in Each Simulation 239

A.1 Simulation Results of Variance Estimators for Clustered GREG Estimate . 288
A.2 Variability of Sandwich Estimators for School Population 289
A.3 Variability of Sandwich Estimators for ACS Population 290
A.4 Variability of Sandwich Estimators for Simulated Population 291
A.5 Confidence Interval Coverage of Variance Estimators 292

B.1 Distributions of the Exponential Family 306
B.2 Logistic Regression Estimating Equations 309
B.3 Percent Simulation Coefficient of Variation for Synthetic Population . . . 374
B.4 Average Distance from True Value for Synthetic Population (in thousands) 375
B.5 Empirical Standard Deviation of Distance from True Value for Synthetic Population (in thousands) 376
B.6 Percent Relative Bias for Synthetic Population 377
B.7 Percent Relative Median Difference for Synthetic Population 378
B.8 Percent Relative Root Mean Squared Error for Synthetic Population . . . 379
B.9 Percent Relative Root Mean Squared Error for Synthetic Population . . . 380
B.10 Percent Relative Bias of LGREG Variance Estimators for Synthetic Population 381
B.11 Percent Relative Median Difference of LGREG Variance Estimators for Synthetic Population 382
B.12 Percent Relative Root Mean Squared Error of LGREG Variance Estimators for Synthetic Population 383
B.13 Percent Relative Root Mean Squared Error of LGREG Variance Estimators for Synthetic Population 384
B.14 Average Distance from Empirical Value for Standard Error Estimators in Synthetic Population (in thousands) 385
B.15 Median Distance from Empirical Value for Standard Error Estimators in Synthetic Population (in thousands) 386

B.16 Standard Error of Average Distance from Empirical Value for Standard Error Estimators in Synthetic Population (in thousands) 387
B.17 Percent 95% Confidence Interval Coverage of LGREG Variance Estimators for Synthetic Population 388
B.18 Percent Simulation Coefficient of Variation for Post-Secondary Population 394
B.19 Average Distance from True Value for Post-Secondary Population (in thousands) 395
B.20 Empirical Standard Deviation of Distance from True Value for Post-Secondary Population (in thousands) 396
B.21 Percent Relative Bias for Post-Secondary Population 397
B.22 Percent Relative Median Difference for Post-Secondary Population 398
B.23 Percent Relative Root Mean Squared Error for Post-Secondary Population 399
B.24 Percent Relative Root Median Squared Error for Post-Secondary Population400
B.25 Percent Relative Bias of LGREG Variance Estimators for NSCG Population401
B.26 Percent Relative Median Difference of LGREG Variance Estimators for Post-Secondary Population 402
B.27 Percent Relative Root Mean Squared Error of LGREG Variance Estimators for Post-Secondary Population 403
B.28 Percent Relative Root Median Squared Error of LGREG Variance Estimators for Post-Secondary Population 404
B.29 Average Distance from Empirical Value for Standard Error Estimators in Post-Secondary Population (in thousands) 405
B.30 Median Distance from Empirical Value for Standard Error Estimators in Post-Secondary Population (in thousands) 406
B.31 Standard Error of Average Distance from Empirical Value for Standard Error Estimators in Post-Secondary Population (in thousands) 407
B.32 Percent 95% Confidence Interval Coverage of LGREG Variance Estimators for Post-Secondary Population 408
B.33 Percent Simulation Coefficient of Variation for Census Population 414
B.34 Average Distance from True Value for Census Population (in thousands) . 415
B.35 Empirical Standard Deviation of Distance from True Value for Census Population (in thousands) 416
B.36 Percent Relative Bias for Census Population 417
B.37 Percent Relative Median Difference for Census Population 418
B.38 Percent Relative Root Mean Squared Error for Census Population 419
B.39 Percent Relative Root Median Squared Error for Census Population . . . 420
B.40 Percent Relative Bias of Variance Estimators for Census Population . . . 421
B.41 Percent Relative Median Difference of Variance Estimators for Census Population 422
B.42 Percent Relative Root Mean Squared Error of Variance Estimators for Census Population 423
B.43 Percent Relative Root Median Squared Error of Variance Estimators for Census Population 424
B.44 Average Distance from Empirical Value for Standard Error Estimators in Census Population (in thousands) 425

B.45 Median Distance from Empirical Value for Standard Error Estimators in Census Population (in thousands) 426
B.46 Standard Error of Average Distance from Empirical Value for Standard Error Estimators in Census Population (in thousands) 427
B.47 Percent 95% Confidence Interval Coverage of LGREG Variance Estimators for Census Population 428

C.1 Simulation Coefficient of Variation for Point Estimators of Count Response. Estimates have been multiplied by $1,000,000$. 464
C.2 Simulation Coefficient of Variation for Point Estimators of Binary Response. Estimates have been multiplied by $1,000,000$. 465
C.3 Simulation Coefficient of Variation for Point Estimators of Synthetic Response. Estimates have been multiplied by $1,000,000$. 466
C.4 Relative Bias and Coefficient of Variation for Point Estimators of Count Response in Small Samples 474
C.5 Relative Bias and Coefficient of Variation for Point Estimators of Count Response in Large Samples 475
C.6 Relative Bias and Coefficient of Variation for Point Estimators of Binary Response in Small Samples 476
C.7 Relative Bias and Coefficient of Variation for Point Estimators of Binary Response in Large Samples 477
C.8 Relative Bias and Coefficient of Variation for Point Estimators of Synthetic Response in Small Samples 478
C.9 Relative Bias and Coefficient of Variation for Point Estimators of Synthetic Response in Large Samples 479

List of Figures

2.1 Scatter plot and residual plot for ACS population 91
2.2 Scatter plot and residual for simulated population 94
2.3 Boxplots of relative standard error estimates for SRS samples of size 25 from third grade population 98
2.4 Boxplots of relative standard error estimates for SRS samples of size 50 from third grade population 99

3.1 Density Plot of Distance Between Estimator and True Value for the Census Population 163
3.2 Box-and-Whisker Plot Showing Percent Relative Difference of Estimated Totals for y_1 of Synthetic Population under Small Fixed SRS 165
3.3 Box-and-Whisker Plot Showing Summary of All Point Estimators 175
3.4 Box-and-Whisker Plots Showing Percent Relative Difference of LGREG Variance Estimators for y_1 in Fixed SRS Samples from Synthetic Population. Small sample sizes on top. 177
3.5 Plot of $\widehat{t}^{lg}_{math}$ versus $v_{Binder}\left(\widehat{t}^{lg}_{math}\right)$ under Small Fixed SRS 178

4.1 Plot of predictions versus true values in for total non-mail returns. 230
4.2 Plot of predictions versus x_k for the binary response. Each point represents the true average rate for 260 units. Models were fitted using the entire population. 231
4.3 Plot of predictions versus x_k for the synthetic response. Each point represents the true average rate for 260 units. Models were fitted using the entire population. 232
4.4 Plot of Relative Bias and Coefficient of Variation for all estimators of total binary response in small samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. . . 242
4.5 Plot of Relative Bias and Coefficient of Variation for all estimators of total binary response in large samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. 244
4.6 Plot of Relative Bias and Confidence Interval Coverage of variance estimators for the GLM-assisted difference estimator of the binary response in small samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. 247
4.7 Plot of Relative Bias and Confidence Interval Coverage of variance estimators for the GLM-assisted difference estimator of the binary response in large samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. 248
4.8 Plot of the Relative Bias and Confidence Interval Coverage for all estimators of the total synthetic response in small samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. 251

4.9 Plot of Relative Bias and Confidence Interval Coverage of variance estimators for the GLM-assisted difference estimator of the synthetic variable in large samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. 252

B.1 Density Plot of Distance Between Estimator and True Value for Synthetic Population 389
B.2 Box and whisker plot showing percent relative difference of estimated totals for y_1 of synthetic population under small fixed SRS 390
B.3 Density Plot of Distance Between Variance Estimators of y_1 and Empirical Variance for Synthetic Population 391
B.4 Density Plot of Distance Between Variance Estimators of y_2 and Empirical Variance for Synthetic Population 392
B.5 Density Plot of Distance Between Variance Estimators of y_3 and Empirical Variance for Synthetic Population 393
B.6 Density Plot of Distance Between Estimator and True Value in the Post-Secondary Population 409
B.7 Plot of LGREG math estimates versus PML LGREG math variance estimates under small fixed SRS in the post-secondary population 410
B.8 Box and whisker plots showing percent relative difference of LGREG variance estimators for math in fixed SRS samples from post-secondary population including all outliers. Small sample sizes on top. 411
B.9 Box and whisker plots showing percent relative difference of LGREG variance estimators for **math** in fixed SRS samples from post-secondary population excluding all outliers. Outliers are 1.5 times the interquartile range beyond the first and third quartiles. Small sample sizes on top. Large samples on bottom. The empirical variance was calculated **with** the outliers. 412
B.10 Box and whisker plots showing percent relative difference of LGREG variance estimators for **math** in fixed SRS samples from post-secondary population excluding all outliers. Outliers are 1.5 times the interquartile range beyond the first and third quartiles. Small sample sizes on top. Large samples on bottom. The empirical variance was calculated **without** the outliers. 413
B.11 Density Plot of Distance Between Estimator and True Value in Census Population 429

C.1 Plot of Relative Bias and Coefficient of Variation for all estimators of Total Count in small samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. For numeric values of the points in this plot see Table C.4. 468
C.2 Plot of Relative Bias and Coefficient of Variation for all estimators of Total Count in large samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. For numeric values of the points in this plot see Table C.5. 469

C.3 Plot of Relative Bias and Coefficient of Variation for all estimators of total binary response in small samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. For numeric values of the points in this plot see Table C.6. 470
C.4 Plot of Relative Bias and Coefficient of Variation for all estimators of total binary response in large samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. For numeric values of the points in this plot see Table C.7. 471
C.5 Plot of Relative Bias and Coefficient of Variation for all estimators of total synthetic response in small samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. For numeric values of the points in this plot see Table C.8. 472
C.6 Plot of Relative Bias and Coefficient of Variation for all estimators of total synthetic response in large samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. For numeric values of the points in this plot see Table C.9. 473
C.7 Plot of Relative Bias and Confidence Interval Coverage of variance estimators for the GLM-assisted difference estimator of the count variable in small samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. 481
C.8 Plot of Relative Bias and Confidence Interval Coverage of variance estimators for the GLM-assisted difference estimator of the count response in large samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. 482
C.9 Plot of Relative Bias and Confidence Interval Coverage of variance estimators for the GLM-assisted difference estimator of the binary response in small samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. 483
C.10 Plot of Relative Bias and Confidence Interval Coverage of variance estimators for the GLM-assisted difference estimator of the binary response in large samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. 484
C.11 Plot of Relative Bias and Confidence Interval Coverage of variance estimators for the GLM-assisted difference estimator of the synthetic variable in small samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. 485
C.12 Plot of Relative Bias and Confidence Interval Coverage of variance estimators for the GLM-assisted difference estimator of the synthetic variable in large samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. 486

Chapter 1

Introduction

This dissertation comprises three distinct research papers. Although each research topic is different and there is very little unifying some of the topics together, all three deal with innovations to model-assisted estimators in complex sample designs. The three papers borrow ideas and techniques from the model-based framework to improve design-based estimation in complex surveys. This lengthy introduction is meant to provide much of the theoretical and technical background necessary to fully appreciate and understand the following chapters.

In the first section of this chapter, the design-based framework is introduced. All three following chapters deal with the analysis of clustered samples where sampling weights may be variable and units within clusters are correlated. In addition to briefly reviewing some key historical developments of the design-based framework, we discuss common sampling and estimation methods in the first section.

The second section of this chapter reviews several model-assisted methods used to improve design-based inference. Generalized regression, generalized difference, calibration, model-calibrated, and model-calibrated maximum pseudoempirical likelihood estimators are discussed. We conclude with an introduction to pseudomaximum likelihood estimation.

In the chapter, we describe the design-based and model-assisted paradigms, high-

light some of the landmark essays from each approach, and introduce some research combining both approaches. The goal of this chapter is to introduce and review many of the statistical methods and ideas which support the three following chapters.

1.1 Design-Based Framework

There are two broad inferential frameworks used to estimate finite population quantities: the design-based framework and the model-based framework. These two frameworks differ in their assumptions, data analysis techniques, and terminology. A careful understanding of these two frameworks is necessary to understand, discuss, and interpret the new research in the following chapters of this dissertation.

From the design-based framework, estimation and inference are taken with respect to repeated sampling. That is, individual characteristics of each unit in the population are considered fixed quantities and randomness is achieved through the sampling process.

1.1.1 Brief History of the Design-Based Framework

Concerned with hidden biases resulting from purposeful sampling techniques, survey statisticians in the middle of the twentieth century began to embrace the principle of using a random mechanism to select their samples instead of purposely picking which units to include in the sample. To analyze samples selected from a probability mechanism, survey statisticians developed the design-based approach. The design-based approach uses the sample design in estimation, rather than making assumptions about the distribution of population characteristics. The design-based theory was largely devel-

oped in the second quarter of the 20th century as a means to analyze randomized survey data that would be congruous with the survey design; however, the design-based theory has been criticized because estimates based on only one sample may be quite far from true population values, statistical analysis appears disjoint from classical statistics, analysis based entirely on the sampling distribution neglects valuable prior information, and poorly selected samples may produce ridiculous results.

In his landmark 1934 paper, Jerzy Neyman laid the foundation for sampling finite populations. He described methods and properties of simple random sampling, stratified random sampling, and multiple-staged stratified sampling. Additionally, he provided a method for sample allocation that minimized the sampling variance of a mean. Covering a variety of topics, Neyman (1934) passionately argued for using probability mechanisms to select samples.

Although Neyman (1934) covered a variety of topics, his paper is primarily a vociferous argument against purposeful sampling. For Neyman (1934, p. 585), randomization was essential to unbiased estimation. He carefully defined a *representative method of sampling* as a method that makes possible the estimation of unbiased results "*irrespective of the unknown properties of the population*." Regarding purposive samples, Neyman (1934, p. 586) boldly stated that they were

> not what I should call a representative method. Of course they may give sometimes perfect results, but these will be due rather to the uncontrollable intuition of the investigator and good luck than to the method itself.

Neyman also laid out the basic framework for cluster sampling. As Neyman (1934,

p. 570) astutely indicated,

> if it is impossible or difficult to organize a random sampling of the individuals forming the population to be studied, the difficulty may be overcome by sampling groups of individuals.

It is still common, especially in household surveys, to select the sample in several stages. For example, large national surveys often select a sample of counties. Then a segment or block of housing units within the sample counties is selected before housing units or even persons are selected. Motivated by practical sampling constraints, Hansen and Hurwitz (1943) set the fundamental theory for estimation from stratified, clustered, and multiple stage samples when the sample is selected with-replacement. Later, Horvitz and Thompson (1952) considered estimation from sampling without-replacement.

Certainly, Neyman compellingly argued for randomized sampling; however, it was not until the 1940's and 1950's that statisticians carefully laid out the theory and mathematics of the design-based paradigm. Together, Hansen and Hurwitz (1943), Yates (1949), Deming (1950), Narain (1951), Horvitz and Thompson (1952), Hansen et al. (1953a), Hansen et al. (1953b), Yates and Grundy (1953), Sen (1953), and Cochran (1953) established the design-based paradigm as the dominant sampling technique for the middle of the 20th century.

1.1.2 Sample Design

The sample design is at the heart of the design-based framework. Indeed, an understanding of the sample design is key to constructing and evaluating design-based and

model-assisted estimators. We begin this section by introducing the finite population and sample notation that we will employ throughout this dissertation. Then, we review the probability framework of sampling theory.

1.1.2.1 Notation

Consider a finite population of N primary sampling units denoted $\mathscr{U}_I = \{1, \ldots, i, \ldots, N\}$. When a two-staged sample is selected, a sample is selected from each sample primary sampling unit. The set of M_i elements in the ith primary sampling unit is denoted $\mathscr{U}_i = \{1, \ldots, k, \ldots, M_i\}$. Overall, there are $M = \sum_{i \in \mathscr{U}_I} M_i$ elements in the population. The full population of secondary sampling units is $\mathscr{U} = \{\mathscr{U}_1, \ldots, \mathscr{U}_k, \ldots, \mathscr{U}_N\}$.

Each unit in the population has a vector of covariates, $\mathbf{x}_k$, and a multivariate response vector $\mathbf{y}_k$. From the model-based framework, $\mathbf{y}_k$ is considered to be a realization of a multivariate random vector $\mathbf{Y}_k$. The sum of the unit level covariates in cluster i is denoted $\mathbf{t}_{xi} = \sum_{k \in \mathscr{U}_i} \mathbf{x}_k$. The full population total of the unit level covariates is $\mathbf{t}_{\mathbf{x}} = \sum_{k \in \mathscr{U}} \mathbf{x}_k$. Likewise, the sum of the unit level response vector in cluster i is $\mathbf{t}_{\mathbf{y}i} = \sum_{k \in \mathscr{U}_i} \mathbf{y}_k$ and the full population total is $\mathbf{t}_{\mathbf{y}} = \sum_{k \in \mathscr{U}} \mathbf{y}_k$. When our response is scalar, we write y_k in normal text; however, when our response is multivariate, we denote our response vector for unit k in boldface as $\mathbf{y}_k$.

In this dissertation, we consider the case where $\mathbf{x}_k$ is known for all units in the population, but $\mathbf{y}_k$ is only measured for sample units. This situation is becoming more and more common in the United States as information resellers create national sampling frames based on the United States Postal Service's Delivery Sequence File. Estevao and

Särndal (2006) describe several other cases, for example where auxiliary data may be available for all clusters in the population, but not all units in the population.

After a sample has been drawn, $\mathscr{U}$ can be partitioned into two mutually exclusive sets: the elements that have been selected for sample and the elements that have not been selected for sample. Let S be a random sample from $\mathscr{U}$ and $\mathfrak{s}$ be a realization of S. The set of all possible samples is denoted by $\mathscr{S}$. If the sample elements have been removed from the population, the remaining elements form a set called the non-sample, denoted with an r subscript. For example, t_{yr} is the total of variable y for the non-sample units.

1.1.3 Sampling Theory

The probability of selecting sample $\mathfrak{s}$ from $\mathscr{S}$ is $\mathscr{P}(\mathfrak{s})$ where $\mathscr{P}$ is a probability measure. That is,

$$P(S = \mathfrak{s}) = \mathscr{P}(\mathfrak{s}) \qquad \forall\, \mathfrak{s} \in \mathscr{S}$$

A sample design is simply the pair $(\mathscr{P}, \mathscr{S})$.

Samples can be selected either with-replacement or without-replacement. A sample design is *without-replacement* if no sample in $\mathscr{S}$ with a nonzero probability measure contains the same element of the population more than once. Alternatively, sample designs where at least one element in the population can appear in a single sample more than once are called *with-replacement* designs. Thus, in without-replacement designs an element can only appear in the sample one time; whereas, each element can appear in sample up to n times in with-replacement designs, where n is the number of draws used to select $\mathfrak{s}$.

In without-replacement sample designs, we can represent each sample as a vector of indicator variables where the kth indicator is set to 1 if the kth unit is included in the sample and set to 0 if the kth element is not included in a specific realized sample. The random sample inclusion indicator for element k is denoted δ_k. One specific realization of the N by 1 random vector $\boldsymbol{\delta}_S$ is denoted $\boldsymbol{\delta}_{\mathfrak{s}}$[1]. Thus, we can alternatively write the event $S = \mathfrak{s}$ as $\boldsymbol{\delta}_S = \boldsymbol{\delta}_{\mathfrak{s}}$. It is often beneficial to use this alternative notation because δ_k is a Bernoulli random variable.

For without-replacement sample designs, the probability that the kth element of the population will fall into sample is called the *first order inclusion probability* and denoted π_k. That is,

$$\begin{aligned}\pi_k &= P\left(k \in S\right) \\ &= P\left(\delta_k = 1\right) \\ &= \sum_{\mathscr{S} \ni k} \mathscr{P}\left(\mathfrak{s}\right)\end{aligned}$$

where the summation is over all samples that contain element k, that is all samples where $\delta_k = 1$. Based on this definition, we see that the first order inclusion probability is the expected value of the sample indicators. That is,

$$\pi_k = E\left(\delta_k\right)$$

For example, for simple random sampling without-replacement ($srswor$), there are $\binom{N}{n}$ different samples of size n and there are $\binom{N-1}{n-1}$ possible samples that contain element

[1] In this paper bold quantities designate vectors and matrices. Scalars are in normal type.

k. Thus, for $srswor$

$$\mathscr{P}(\mathfrak{s}) = \frac{1}{\binom{N}{n}} \qquad \forall\, \mathfrak{s} \in \mathscr{S}$$

and

$$\pi_k = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

Base weights are defined as the inverse of the probabilities of selection. Thus, the *base weight* is,

$$d_k = \frac{1}{\pi_k}$$

The probability that two elements, k and l, both fall into sample is called the *second order inclusion probability*, π_{kl} and defined by

$$\begin{aligned}
\pi_{kl} &= P\left(k \,\&\, l \in S\right) \\
&= P\left(\delta_k \delta_l = 1\right) \\
&= \sum_{\mathscr{S} \ni k \& l} \mathscr{P}(\mathfrak{s})
\end{aligned}$$

To measure the variability of estimators based on without-replacement sample designs, we need to know the variance and covariance of the sample membership indicators. Using the fact that δ_k is a Bernoulli random variable, it follows that the variance and covariance of the sample membership indicators are

$$\operatorname{var}(\delta_k) = \Delta_{kk} = \pi_k (1 - \pi_k)$$

$$\operatorname{cov}(\delta_k \delta_l) = \Delta_{kl} = \pi_{kl} - \pi_k \pi_l$$

For a simple random sample without-replacement $(srswr)$, these reduce to

$$\Delta_{kk} = \frac{n}{N}\left(1 - \frac{n}{N}\right)$$
$$\Delta_{kl} = \frac{n}{N}\frac{n-1}{N-1} - \left(\frac{n}{N}\right)^2$$

For two-staged samples, a probability sample of clusters, $\mathfrak{s}_I$, is selected from $\mathscr{U}_I$ according to some fixed sample design. Overall, n clusters are selected. The probability that cluster i is selected is $P\left(i \in \mathfrak{s}_1\right) = \pi_i$ and the joint inclusion probabilities of selection are $P\left(i, j \in \mathfrak{s}_I\right) = \pi_{ij}$. Furthermore, let $\Delta_{ij} = \pi_{ij} - \pi_i\pi_j$.

Within each sample cluster, a sample of units, $\mathfrak{s}_i$, is selected according to a fixed sample design. The cardinality of $\mathfrak{s}_i$ is m_i and $m = \sum_{i\in\mathfrak{s}} m_i$, the overall sample size of units. The probability that unit k is selected, given that cluster i was selected is $\pi_{k|i} = P\left(k \in \mathfrak{s}_i | i \in \mathfrak{s}\right)$ and the conditional joint inclusion probabilities of selection are $\pi_{kl|i} = P\left((k,l) \in \mathfrak{s}_i | i \in \mathfrak{s}\right)$. Also, let $\Delta_{ij} = \pi_{ij} - \pi_i\pi_j$ and $\Delta_{kl|i} = \pi_{kl|i} - \pi_{k|i}\pi_{l|i}$.

A similar theory has been developed for with-replacement sampling (see Särndal et al. 1992). One important difference between with and without-replacement designs is that the probability that the kth unit is selected on any given draw is denoted by p_k for with-replacement designs. In with-replacement designs, we let n be the total number of first-stage draws. In without-replacement samples, n is the total number of unique sample units, while in with-replacement samples a unique sample unit may be counted multiple times in n. According to Särndal et al. (1992, p. 51), the probability of including the kth element in a with-replacement sample is,

$$\pi_k = 1 - \left(1 - p_k\right)^n \approx np_k \tag{1.1}$$

Although with-replacement samples are rarely selected in practice, inference is often made assuming a with-replacement design because variance estimators simplify due to the independence of each successive draw. Depending on the sample design, using a with-replacement design when analyzing a without-replacement sample can result in conservative inference[2], although this is not always the case (see sec. 4.6 Särndal et al. 1992).

In $srswr$ and $srswor$ samples, every element in the population has the same probability of being selected. If auxiliary data are available about the size of all elements in the population, then a p-proportional to size (pps) or π-proportional to size (πps) sample can be selected. In this notation pps samples are selected with-replacement while $\pi - ps$ samples are selected without-replacement. If x_k designates the size of the kth element and $t_x = \sum_{k=1}^{N} x_k$, then pps designs select the kth element with probability $p_k = \frac{x_k}{t_x}$ on each draw; while the probability that the kth element will fall into a πps sample is $\pi_k = \frac{nx_k}{t_x}$.

In two-staged samples, clusters of units are first selected and then units within the sample units are sampled. First a frame is created so that population units are grouped into mutually exclusive clusters called *Primary Sampling Units* (PSUs). Then, a sample of PSUs is selected. In the second stage, a sample of units within sample PSUs is selected. When selecting a two-staged sample, the first and second stages can be selected using different techniques. However, point and variance estimation is simplified if the second-stage sample is invariant and independent of the first stage. According to Särndal et al. (1992), the second stage sample is *invariant* of the first stage sample if the same sample design is used to select $\mathfrak{s}_i$ every time the ith PSU is selected. That is, the probability

[2]Conservative inference means that the average estimated sampling error is larger than the true sampling error.

of selecting $\mathfrak{s}_i$ does not depend on $\mathfrak{s}_I$. Because of invariance, $E\left(\widehat{t_i}|\mathfrak{s}_I\right) = E\left(\widehat{t_i}\right)$ where $\widehat{t_i}$ is an estimated total for cluster i. A second stage sample is *independent* of the first stage sample if the sample design used to select $\mathfrak{s}_i$ does not depend on $\mathfrak{s}_j$. Because of independence $E\left(\widehat{t}|\mathfrak{s}_I\right) = \sum_{i\in\mathfrak{s}_I}\left(\widehat{t_i}|\mathfrak{s}_I\right)$.

Assuming invariance and independence between the first and second stage of sampling, the ultimate probabilities of selection are $\pi_k = \pi_i\pi_{k|i}$ for all k. Similarly, $d_k = d_1 d_{k|i}$ where d_k is the unconditional base weight, d_i is the base weight for cluster i, and $d_{k|i}$ is the base weight for unit k given that cluster i was selected.

The joint inclusion probabilities cannot be computed for some sample designs. And, even when the joint inclusion probabilities can be computed, they often are not computed because they are not needed to select the sample and require extensive computer resources. Consider that a single-staged sample of 1,000 elements would require storing a 1,000 by 1,000 triangular matrix of joint inclusion probabilities containing up to 500,500 unique elements.

Poisson sampling provides computational simplicity and data reduction. Poisson sampling is a broad class of sampling methods that can be used to select samples with a random sample size and unequal probabilities of inclusion. In Poisson samples the selection of each element is independent of all other selections. Thus, for a Poisson sample, the sample inclusion probability for primary sampling unit i is π_i and the probability that primary sampling unit i is not in sample is $1 - \pi_i$. Likewise, the sample inclusion probability for unit k given that primary sampling unit i was selected is $\pi_{k|i}$ whist the probably that unit k will not be included in sample given that primary sampling unit i was selected is $1 - \pi_{k|i}$. As a consequence, for Poisson samples $\pi_{ij} = \pi_i\pi_j$ for all $i \neq j$ and

$\pi_{kl|i} = \pi_{k|i}\pi_{l|i}$ for all $k \neq l$. These features of Poisson samples greatly simplify the calculations of Δ_{ij} and $\Delta_{kl|i}$. Systematic samples, with-replacement samples, and samples with a fixed sample design are not Poisson samples. If such common sample designs are used and $\pi_{ij} \approx \pi_i\pi_j$ and $\pi_{kl|i} \approx \pi_{k|i}\pi_{l|i}$, then we can simplify our estimation by assuming an approximate Poisson sample design.

Tillé (2006) identifies numerous algorithms that can be used to select a variety of samples. Many of these algorithms are programmed in the R sampling package (Tillé and Matei 2009).

1.1.4 Data

Each unit in the population has a set of characteristics associated with it. Some of these characteristics are known to the statistician prior to sampling, some are measured during the data collection, and some are never known. We let y_k be the key characteristic of interest from the kth unit, for $k = 1, 2, \ldots, m$. Combining the characteristic of interest into a sample vector gives,

$$\underset{m\times 1}{\mathbf{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

In many sample surveys, data are not collected for some sample units as a result of nonresponse or noncontacts. In such situations, there is a need to distinguish between the sample vector and the response vector. Because this dissertation deals entirely with sampling errors, nonsampling errors such as nonresponse are not considered. Thus, added

notation for response is not necessary.

If we select a clustered sample, we can combine our responses from the ith sample cluster into a vector denoted

$$\underset{m_i \times 1}{\mathbf{y}_i} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{m_i} \end{bmatrix}$$

In the case, when we are interested in a vector of C responses, we denote the vector of key characteristics for the kth unit as,

$$\underset{C \times 1}{\mathbf{y}_k} = \begin{bmatrix} y_{k1} \\ y_{k2} \\ \vdots \\ y_{kC} \end{bmatrix}$$

Also associated with the kth element of our sample is a nonrandom and fully known column vector in $\mathbb{R}^p$ of p explanatory variables, denoted $\mathbf{x}_k$. That is,

$$\underset{p \times 1}{\mathbf{x}_k} = \begin{bmatrix} x_{k1} \\ x_{k2} \\ \vdots \\ x_{kp} \end{bmatrix}$$

Moreover, we can combine all of the explanatory variables for our sample into a matrix,

giving

$$\underset{m\times p}{\mathbf{X}} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_m^\top \end{bmatrix}$$

When our sample or population is clustered, we can denote the matrix of auxiliary data for the ith sample cluster as,

$$\underset{m_i\times p}{\mathbf{X}_i} = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_{m_i}^\top \end{bmatrix}$$

1.1.5 Estimation

1.1.5.1 With-Replacement Estimators

Hansen and Hurwitz (1943) described sampling and estimation techniques for various with-replacement sample designs. With-replacement samples are often not drawn because they can result in the same unit appearing in the sample more than once. With-replacement samples can be thought of as a collection of independent draws from a population. Thus, many calculations simplify in with-replacement samples because each draw is independent of all other draws. This assumption of independence can greatly simplify estimation.

In the literature on with-replacement sample designs, the number of draws is often distinguished from the number of unique elements in sample. In this dissertation, we let

n be the total number of primary sampling unit draws prior to unduplication. In without-replacement samples, n simplifies to the total number of unique first-stage elements; but this is not the case in with-replacement samples.

Hansen and Hurwitz (1943) proposed estimating the total from a single-staged with-replacement sample with

$$\hat{t}_y^{pwr} = \frac{1}{n} \sum_{k \in \mathfrak{s}} \frac{y_k}{p_k} \tag{1.2}$$

We note that in with-replacement samples, $\mathfrak{s}$ may contain duplicates.

The variance of $\hat{t}_y^{pwr}$ in single-staged samples is,

$$\text{var}\left(\hat{t}_y^{pwr}\right) = \frac{1}{n} \sum_{k \in \mathscr{U}} \left(\frac{y_k}{p_k} - t_y\right)^2 p_k$$

Of course, y_k is not known for the entire population and t_y is also unknown. Thus, the variance is estimated by,

$$\upsilon_{wr}\left(\hat{t}_y^{pwr}\right) = \frac{1}{n\,(n-1)} \sum_{k \in \mathfrak{s}} \left(\frac{y_k}{p_k} - \hat{t}_y^{pwr}\right)^2 \tag{1.3}$$

In a clustered sample where the clusters are selected with-replacement, the pwr estimator is,

$$\hat{t}_y^{pwr} = \frac{1}{n} \sum_{i \in \mathfrak{s}} \frac{\hat{t}_{yi}}{p_i}$$

where $\hat{t}_{yi}$ is an estimate of the total for the ith cluster, p_i is the probability of drawing the ith cluster in a single draw, and n is the total number of draws from the population of clusters.

In multiple stages of sampling, the variance of $\hat{t}_y^{pwr}$ is,

$$\text{var}\left(\hat{t}_y^{pwr}\right) = \frac{1}{n} \sum_{i \in \mathscr{U}} \left(\frac{\hat{t}_{yi}}{p_i} - t_y\right)^2 p_i + \frac{1}{n} \sum_{i \in \mathscr{U}} \frac{V_i}{p_i}$$

where V_i is the variance of $\hat{t}_{yi}$ due to sampling within cluster i. Of course, $\hat{t}_{yi}$ is not known for the entire population and t_y is also unknown. An estimator of the variance is,

$$v_{wr}\left(\hat{t}_y^{pwr}\right) = \frac{1}{n\left(n-1\right)} \sum_{i \in \mathfrak{s}} \left(\frac{\hat{t}_{yi}}{p_i} - \hat{t}_y^{pwr} \right)^2 \tag{1.4}$$

Commonly, with-replacement variance estimators are used even when the first stage sample is selected without-replacement. As long as the sampling fraction is relatively small, the bias of using a with-replacement variance estimator is relatively small. Särndal et al. (1992, sec 4.6) discuss the classic with-replacement variance estimator of a total and provide some limitations for using the with-replacement variance estimator for samples selected without-replacement.

1.1.5.2 The π-Estimator

Point Estimation in Single Stage Samples

The first major development in the design-based framework was the introduction of the π-estimator, often attributed to Horvitz and Thompson (1952) and Narain (1951). This estimator expands each sample unit with a weight equal to the inverse of its probability of selection. By summing the weighted sample units, totals can be estimated. The π-*estimator* of a total for a sample of size n from a population of size N under a single

stage of sampling can be written in many ways including,

$$\begin{aligned}
\hat{t}_y^{\pi} &= \boldsymbol{\delta}_{\mathscr{U}}^{\top} \boldsymbol{\Pi}_{\mathscr{U}}^{-1} \mathbf{y}_{\mathscr{U}} = \mathbf{1}^{\top} \boldsymbol{\Pi}_s^{-1} \mathbf{y}_s \\
&= \sum_{k=1}^{N} \frac{\delta_k y_k}{\pi_k} = \sum_{k=1}^{n} \frac{y_k}{\pi_k} \\
&= \sum_{k \in \mathscr{U}} \frac{\delta_k y_k}{\pi_k} = \sum_{k \in \mathfrak{s}} \frac{\delta_k y_k}{\pi_k} \\
&= \sum_{\mathscr{U}} \delta_k d_k y_k = \sum_{\mathfrak{s}} d_k y_k
\end{aligned}$$

where $\boldsymbol{\Pi}$ is a diagonal matrix of the design-based selection probabilities. In invariant and independent clustered samples, $\boldsymbol{\Pi} = \operatorname{diag}(\pi_k) = \operatorname{diag}\left(\pi_{k|i}\pi_i\right)$ where $\pi_{k|i}$ is the probability of selecting unit k given that cluster i was already selected. Thus, the π-estimator requires weighting each element by the inverse of its unconditional probably of selection.

The π-estimator is relatively simple and does not require knowledge of any auxiliary information. It is important to note that the only random variable in the π-estimator is $\boldsymbol{\delta}$ or equivalently the $\mathfrak{s}$ subscript.

Variance Estimation in Single Stage Samples

The variance of the π-estimator depends on the sample design. Horvitz and Thompson (1952) constructed variance estimators of $\hat{t}_y^{\pi}$ under a variety of common sampling designs. Yates and Grundy (1953) and Sen (1953) generalized the variance estimators of Horvitz and Thompson to any measurable fixed single-stage sample design. All of these variance estimators require knowledge of the joint inclusion probabilities, which are often unknown at the time of analysis or cumbersome to calculate for sample designs with unequal probabilities of selection. The variance of the Horvitz-Thompson estimator in a

single-stage fixed sample size design is

$$\operatorname{var}\left(\hat{t}_y^{\pi}\right) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(\pi_i\pi_j - \pi_{ij}\right)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \tag{1.5}$$

$$= \sum_{i\in\mathscr{U}}\sum_{j\in\mathscr{U}}\Delta_{ij}\frac{y_i}{\pi_i}\frac{y_j}{\pi_j}. \tag{1.6}$$

For Poisson sampling $\pi_{ij} = \pi_i\pi_j$ for all $i \neq j$. Thus, the variance simplifies to

$$\operatorname{var}\left(\hat{t}_y^{\pi}\right) = \sum_{i\in\mathscr{U}}\pi_i\left(1-\pi_i\right)\left(\frac{y_i}{\pi_i}\right)^2. \tag{1.7}$$

In the absence of y_i for the full population, the variance must be estimated. An unbiased estimator using the linear substitution method is

$$\upsilon_e\left(\hat{t}_y^{\pi}\right) = \sum_{i=1}^{n}\sum_{j<i}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2.$$

For Poisson sampling our variance estimator simplifies to

$$\upsilon_e\left(\hat{t}_y^{\pi}\right) = \sum_{i\in\mathfrak{s}}\left[\frac{\pi_i\left(1-\pi_i\right)}{\pi_i}\left(\frac{y_i}{\pi_i}\right)^2\right].$$

Other variance estimators abound. Cumberland and Royall (1981) compared six variance estimators for the π-estimator. Four of the variance estimators were based on the design-based framework and the remaining two were bias-robust estimates motivated by the prediction theory. Cumberland and Royall (1981) derived theoretical differences between the estimators and then compared how well they performed on six populations using simulations. Using the design-based mean squared error as their standard, they found that the variance estimator proposed by Horvitz and Thompson (1952) was highly variable and in many samples failed to come close to measuring the true variance. Moreover, it often underestimated the true variance in samples of size 32. They concluded that,

"We believe these results show again that finite population inferences should be based on prediction models, not on the probability sampling distribution."

Point Estimation in Two Stage Samples

Point estimation from multiple stage samples can easily be made with the π-estimator. For example, the π-estimator in two stages of samples where both stages of sampling are selected without-replacement can be written as

$$\begin{aligned}
\hat{t}_y^{\pi} &= \sum_{i=1}^{N} \frac{\delta_i \hat{t}_i}{\pi_i} = \sum_{i=1}^{n} \frac{\hat{t}_i^{\pi}}{\pi_i} \\
&= \sum_{k=1}^{M} \frac{\delta_k y_k}{\pi_k} = \sum_{k=1}^{m} \frac{y_k}{\pi_k} \\
&= \sum_{i=1}^{N} \sum_{k=1}^{M_i} \frac{\delta_i \delta_{k|i} y_{ki}}{\pi_i \pi_{k|i}} = \sum_{i=1}^{n} \sum_{k=1}^{m_i} \frac{y_{ki}}{\pi_i \pi_{k|i}} \\
&= \sum_{k \in \mathscr{U}} \delta_k d_k y_k = \sum_{k \in \mathfrak{s}} d_k y_k \\
&= \sum_{i \in \mathfrak{s}_I} \frac{\hat{t}_{yi}^{\pi}}{\pi_i} = \sum_{i \in \mathfrak{s}_I} \sum_{k \in \mathfrak{s}_i} \frac{y_{ik}}{\pi_i \pi_{k|i}}
\end{aligned} \tag{1.8}$$

where

$$\widehat{t}_{yi}^{\pi} = \sum_{k \in \mathfrak{s}_i} \frac{y_{ik}}{\pi_{k|i}}. \tag{1.9}$$

and $\delta_{k|i}$ is the sample inclusion indicator for unit k within cluster i.

Variance Estimation in Two Stage Samples

Estimating the sampling variance of the π-estimator in cluster samples can be complicated because the unconditional probability that a unit is in sample depends on the first stage of sampling. Thus, multiple stages of sampling introduces covariance in the second stage of sampling. A prime concern for analyzing data from multiple-staged samples is dealing with the dependence introduced into the sample from the sample design. The

variance of the π-estimator is

$$\begin{aligned}\operatorname{var}\left(\widehat{t_y^\pi}\right) &= \operatorname{var}\left[E\left(\hat{t}_{yi}^\pi|\mathfrak{s}_I\right)|\mathfrak{s}_i\right] + E\left[\operatorname{var}\left(\hat{t}_{yi}^\pi|\mathfrak{s}_I\right)|\mathfrak{s}_i\right] \\ &= \sum_{i=1}^N \sum_{j=1}^N \left[\left(\pi_{ij} - \pi_i\pi_j\right)\frac{t_i}{\pi_i}\frac{t_j}{\pi_j}\right] + \sum_{i=1}^N \frac{\sum_{k=1}^{m_i}\sum_{l=1}^{m_i}\left(\pi_{kl|i} - \pi_{k|i}\pi_{l|i}\right)\frac{y_{k|i}}{\pi_{k|i}}\frac{y_{l|i}}{\pi_{l|i}}}{\pi_i}.\end{aligned}$$

Because t_i is not known for all clusters and y_k is not known for all elements in sample clusters, we cannot compute this variance. However, Särndal et al. (1992, sec 4.3) provide an unbiased estimator of the sampling variance,

$$\upsilon\left(t_y^\pi\right) = \sum_{i=1}^n \sum_{j=1}^n \left[\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}}\frac{\hat{t}_i^\pi}{\pi_i}\frac{\hat{t}_j^\pi}{\pi_j}\right] + \sum_{i=1}^n \frac{\sum_{k=1}^{m_i}\sum_{l=1}^{m_i}\frac{\pi_{kl|i}-\pi_{k|i}\pi_{l|i}}{\pi_{kl|i}}\frac{y_{k|i}}{\pi_{k|i}}\frac{y_{l|i}}{\pi_{l|i}}}{\pi_i}$$

If both stages of sample selection are independent Poisson samples, the variance of the π-estimator is

$$\operatorname{var}\left(\widehat{\mathbf{t}}_y^\pi\right) = \sum_{i\in\mathscr{U}_I}\left(\pi_i - \pi_i^2\right)\left(\frac{t_{yi}}{\pi_i}\right)^2 + \sum_{i\in\mathscr{U}_I}\frac{V_i}{\pi_i}$$

where

$$V_i = \operatorname{var}\left(\hat{t}_{yi}^\pi\right) = \sum_{k\in\mathscr{U}_i}\left(\pi_{k|i} - \pi_{k|i}^2\right)\left(\frac{y_k}{\pi_{k|i}}\right)^2.$$

This variance can be estimated by

$$\upsilon_e\left(\hat{t}_y^\pi\right) = \sum_{i\in\mathfrak{s}_I}\frac{\left(\pi_i - \pi_i^2\right)}{\pi_i}\left(\frac{\widehat{t_{yi}^\pi}}{\pi_i}\right)^2 + \sum_{i\in\mathfrak{s}_I}\frac{\widehat{V}_i}{\pi_i}$$

where

$$\widehat{V}_i = \sum_{k\in\mathfrak{s}_i}\frac{\left(\pi_{k|i} - \pi_{k|i}^2\right)}{\pi_{k|i}}\left(\frac{y_k}{\pi_{k|i}}\right)^2.$$

Commonly, with-replacement variance estimators are used even when the clusters were selected without-replacement. As long as the sampling fraction is relatively small,

the bias of using a with-replacement variance estimator is relatively small. Särndal et al. (1992, sec 4.6) discuss the classic with-replacement variance estimator of a total under multiple stages of sampling. For estimating the variance of the π-estimator, their with-replacement variance estimator is motivated by Equation (1.4). The with-replacement variance estimator for the π-estimator in two-staged samples is

$$v_{wr}\left(\hat{t}_y^{\pi}\right) = \frac{1}{n\left(n-1\right)} \sum_{i=1}^{n} \left(\frac{\hat{t}_{yi}^{\pi}}{p_i} - \hat{t}_y^{\pi} \right)^2 \tag{1.10}$$

where p_i is the probability of drawing the ith cluster in single draw and n is the total number of sample clusters. That is $p_i = \frac{\pi_i}{n}$.

Summary

The π-estimator is simple, versatile, design-unbiased, and requires no explicit parametric or model assumptions about the population. In many ways it is an attractive estimator. Unfortunately, the π-estimator has larger sampling variance than many other estimators that use auxiliary or frame data in estimation. In this sense, the π-estimator is not efficient. Moreover, sometimes the large variability of the π-estimator makes inference based on only one sample rather risky; despite, the fact that the π-estimator is design-unbiased. Indeed, estimates from a single sample may be far from the true value, especially if the probabilities of selection are negatively correlated with the characteristic of interest (see Basu (1971) and Little (2004)).

1.1.6 Empirical Properties of Design-Based Estimators

In the design-based theory, estimators are evaluated by how they perform over repeated sampling. Estimators that are close to the true value on average are more desirable

than those that are farther from the true value. Moreover, less variable estimators tend to be favored over highly variable estimators. To measure the performance of design-based estimators, statisticians tend to focus on bias, approximate bias, consistency, sampling variance, and the mean squared error.

The design-based expected value of an estimator, $\hat{\theta}$, denoted $\mathrm{E}\left(\hat{\theta}\right)$, is a weighted average of the estimator over all possible samples using a specified sample design

$$\mathrm{E}\left(\hat{\theta}\right) = \sum_{\mathfrak{s} \in \mathscr{S}} \mathscr{P}\left(\mathfrak{s}\right) \left[\hat{\theta}_{\mathfrak{s}}\right].$$

In a simulation where we select $\aleph$ samples, indexed by the letter ν, we can calculate the empirical expected value as

$$\widetilde{\mathrm{E}}\left(\hat{\theta}\right) = \frac{1}{\aleph} \sum_{\nu=1}^{\aleph} \hat{\theta}_{\nu}.$$

The *bias* of an estimator is a measure of how far the expected value of an estimator is from the true value, θ. The bias of an estimator is defined as

$$\mathrm{Bias} = \mathrm{E}\left(\hat{\theta} - \theta\right).$$

In a simulation, we can calculate the empirical bias as

$$\widetilde{\mathrm{Bias}} = \frac{1}{\aleph} \sum_{\nu=1}^{\aleph} \left(\hat{\theta}_{\nu} - \theta\right).$$

It is important to note that the bias of an estimator is a property of the estimator over all possible samples. An estimate from one sample may be very close to the true value, even if the estimator is biased. Alternatively, an estimate from one particular sample may be very far from the true value, even if the estimator is unbiased. Both the bias and variability of an estimator must be considered when evaluating an estimator. For this reason, Hansen et al. (1953a, p. 17) note that,

In many situations, an estimation procedure with a very small bias may be considerably more reliable than the best available unbiased estimating procedure.

Because the bias is sensitive to the scale of the estimator, the bias is often divided by θ. This quantity is called the *relative bias*,

$$\mathrm{RB}\left(\hat{\theta}\right) = \frac{\mathrm{E}\left(\hat{\theta}\right) - \theta}{\theta}.$$

In a simulation, we can calculate the empirical relative bias as

$$\widetilde{\mathrm{RB}}\left(\hat{\theta}\right) = \frac{\frac{1}{\aleph}\sum_{\nu=1}^{\aleph}\left(\hat{\theta}_{\nu}\right) - \theta}{\theta}.$$

The *sampling variance* is the average squared difference between estimates and the expected value of the estimator over all possible samples. That is,

$$\mathrm{var}\left(\hat{\theta}\right) = \sum_{\mathfrak{s}\in\mathscr{S}} \mathscr{P}\left(\mathfrak{s}\right)\left[\hat{\theta}_{\mathfrak{s}} - \mathrm{E}\left(\hat{\theta}\right)\right]^{2}.$$

In a simulation, we can calculate the empirical sampling variance as,

$$\widetilde{\mathrm{var}}\left(\hat{\theta}\right) = \frac{1}{\aleph}\sum_{\nu=1}^{\aleph}\left[\hat{\theta}_{\nu} - \frac{1}{\aleph}\sum_{\nu=1}^{\aleph}\left(\hat{\theta}_{\nu}\right)\right]^{2}.$$

The square root of this quantity is the *standard error*. The sampling variance and standard error are both measures of how variable an estimator is about its mean.

The *coefficient of variation* is the variance relative to the true value. That is,

$$\mathrm{CV}\left(\hat{\theta}\right) = \frac{\sqrt{\mathrm{var}\left(\hat{\theta}\right)}}{\theta}.$$

In a simulation, we can calculate the coefficient of variation with

$$\widetilde{\mathrm{CV}}\left(\hat{\theta}\right) = \frac{\sqrt{\widetilde{\mathrm{var}}\left(\hat{\theta}\right)}}{\widetilde{\mathrm{E}}\left(\hat{\theta}\right)}.$$

The *mean squared error* of an estimator is the average squared distance the estimator is from the true value over repeated samples. That is,

$$\mathrm{MSE}\left(\hat{\theta}\right) = \sum_{\mathfrak{s}\in\mathscr{S}} \mathscr{P}\left(\mathfrak{s}\right)\left[\hat{\theta}_{\mathfrak{s}} - \theta\right]^2.$$

The mean squared error can be decomposed into the variance plus the bias squared. The mean squared error is often used to evaluate an estimator's quality because it combines bias and variance. In a simulation, we can calculate the empirical mean squared error as,

$$\widetilde{\mathrm{MSE}}\left(\hat{\theta}\right) = \frac{1}{\aleph}\sum_{\nu=1}^{\aleph}\left[\hat{\theta}_{\nu} - \theta\right]^2.$$

The square root of the mean squared error is the *root mean squared error*. Dividing the root mean squared error by the estimator gives the *relative root mean squared error*.

$$\mathrm{RRMSE}\left(\hat{\theta}\right) = \frac{\sqrt{\sum_{\mathfrak{s}\in\mathscr{S}} \mathscr{P}\left(\mathfrak{s}\right)\left[\hat{\theta}_{\mathfrak{s}} - \theta\right]^2}}{\theta}.$$

Similarly, the empirical relative root mean squared error is

$$\widetilde{\mathrm{RRMSE}}\left(\hat{\theta}\right) = \frac{\sqrt{\frac{1}{\aleph}\sum_{\nu=1}^{\aleph}\left[\hat{\theta}_{\nu} - \theta\right]^2}}{\theta}.$$

Confidence intervals are often calculated to show the quality of an estimator. To summarize the quality of the confidence interval construction process, we can count the number of times that the true population value is below, within, and above the confidence

interval over repeated samples. A $100\,(1-\alpha)\,\%$ normal approximation confidence interval is based on the large sample approximation that

$$P\left(\left|\frac{\widehat{\theta}-\theta}{\sqrt{\operatorname{var}\left(\widehat{\theta}\right)}}\right| \leq z_{1-\frac{\alpha}{2}}\right) \approx 1-\alpha$$

where $z_{1-\frac{\alpha}{2}}$ is the $\left(1-\frac{\alpha}{2}\right)$ percentile of the normal distribution. The theory supporting this approximation requires that the number of first-stage sample units be large. When the sample size is only moderate, the t-distribution is often used. Thus, in moderate-sized samples, the approximate confidence interval is usually calculated by

$$\widehat{\theta} \pm t_{\frac{\alpha}{2},(n-1)}\sqrt{\operatorname{var}\left(\widehat{\theta}\right)}.$$

According to the central limit theorem, 95% of the 95% confidence intervals should include the true population total. Thus, the closer the confidence interval coverage rate is to 95%, the better. In cluster samples, the degrees of freedom for the t-distribution is approximated by $n-1$ where n is the number of sample clusters.

For each sample, we can create an empirical confidence interval by

$$\widehat{\theta}_{\nu} \pm t_{\frac{\alpha}{2},(n-1)}\sqrt{\widetilde{\operatorname{var}}\left(\widehat{\theta}\right)}.$$

For the full simulation, there will be $\aleph$ confidence intervals. The empirical confidence interval coverage rate is the percent of all samples where the true value is within the empirical confidence interval. Similarly the lower and upper empirical confidence interval noncoverage rates are the percent of samples where the true value is below or above the empirical confidence interval bounds.

Table 1.1 summarizes the empirical measures we commonly use to assess the properties of our estimators. Additionally, Table 1.1 also shows the *simulation coefficient of*

Table 1.1: Formulas for summary measures

Name	Summary Measure
Relative Bias	$\frac{\frac{1}{\aleph}\sum_{\nu=1}^{\aleph}\left(\widehat{\theta}_{\nu}-\theta\right)}{\theta}$
Coefficient of Variation	$\frac{\sqrt{\frac{1}{\aleph}\sum_{\nu=1}^{\aleph}\left(\widehat{\theta}_{\nu}-\frac{1}{\aleph}\sum_{\nu=1}^{\aleph}\theta\right)^{2}}}{\frac{1}{\aleph}\sum_{\nu=1}^{\aleph}\hat{\theta}_{\nu}}$
Relative Root Mean Squared Error	$\frac{\sqrt{\frac{1}{\aleph}\sum_{\nu=1}^{\aleph}\left(\widehat{\theta}_{\nu}-\theta\right)^{2}}}{\theta}$
Simulation Coefficient of Variation	$\frac{\frac{1}{\sqrt{\aleph}}\sqrt{\frac{1}{\aleph-1}\sum_{\nu=1}^{\aleph}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^{2}}}{\boldsymbol{\theta}}$

variation. The simulation coefficient of variation is the relative simulation error. It is an estimate of the standard error of $\widetilde{\mathrm{E}}\left(\hat{\theta}\right)$ divided by $\widetilde{\mathrm{E}}\left(\hat{\theta}\right)$. The simulation coefficient of variation can be used to compare and test the empirical performance of estimators within and among simulations.

1.1.7 Theoretical Properties of Design-Based Estimators

In addition to looking at the bias, variance, mean squared error, and confidence interval coverage, it is also instructive to investigate the behavior of an estimator as the sample size increases. Ideally, as the sample size increases, a sequence of estimates should get closer and closer to the population value. This property is called consistency. Hansen et al. (1953a, p. 20) explained that for consistent estimators

> if the sample size is sufficiently large one does not take a serious risk in using an estimate made from a sample drawn at random.

Consistency is especially useful when assessing nonlinear estimators. Whereas, bias and sampling variance can be computed for estimators that are linear in the sample indicators; for nonlinear functions, the expected value, bias, and variance can only be approximated.

In the case of nonlinear estimators, it is common to explore the large sample behavior of estimators. The conditions necessary to explore large sample theory in the design-based theory depend on the sample design and the structure of the finite population. One simple approach is to explore the estimator as n approaches N. For example, Cochran (1977, p. 21) claims that,

> A method of estimation is called *consistent* if the estimate becomes exactly equal to the population value when $n = N$, that is, when the sample consists of the whole population.

This definition of consistency has limited appeal because there are numerous common designs for which it is impossible for the sample and the population to be equivalent. For example, for πps the condition $x_k < t_x/n$ must be met for all k. When $n = N$, then $x_k = t_x/N = \bar{x}$. This will only occur when every element has the same value, thereby making πps indistinguishable from $srswor$. Särndal et al. (1992, p. 168) present more limitations of finite population consistency.

Alternative definitions of consistency rely on the asymptotic behavior of estimators as both the finite population and sample size increase. Since the design-based framework requires a finite population, and thereby a finite sample size, the design-based theory must be modified to explore the asymptotic properties of estimators. Numerous statisticians have relaxed various design-based assumptions to apply asymptotic theory to finite populations. Most of these descriptions revolve around the notion of a *superpopulation*. Definitions of the superpopulation usually involve relaxing two finite population assumptions. According to Hansen et al. (1953b, p. 74) consistency requires that

1. As the size of sample n increases, the size of population N will also increase, and for all n and N we will have $n < cN$, where $0 < c < 1$.

2. As the size of the population increases, the quantity θ that we want to estimate will remain constant.

Brewer (1979) and Isaki and Fuller (1982) provide elegant descriptions of the superpopulation framework for single-stage samples. Isaki and Fuller (1982) envision a series of nested populations. From each population, a sample is selected with respect to a fixed sample design. The sample size increases as the population size increases. Estimates are made from each sample. From the sequence of estimators, an infinite series is formed. From this framework, Särndal et al. (1992, p. 167) claim that,

An estimator $\hat{\theta}_\nu$ is *consistent* for θ, if for any fixed $\varepsilon > 0$,

$$\lim_{\nu \to \infty} P\left(\left|\hat{\theta}_\nu - \theta\right| > \varepsilon\right) = 0$$

where ν indexes the growing population. Särndal et al. (1992, p. 153) also use the superpopulation framework to investigate the bias of estimators as the sample grows,

An estimator $\hat{\theta}_\nu$ is *asymptotically unbiased* for θ if

$$\lim_{\nu \to \infty} \left[E\left(\hat{\theta}_\nu\right) - \theta\right] = 0$$

In multiple stages of sampling, the superpopulation structure must be further defined. Práŝková and Sen (2009) review many current asymptotic approaches to finite population sampling. For multiple stage sampling they describe two frameworks. In the first setup,

1. the number of population clusters increases to infinity,
2. the sampling rate for clusters does not approach 0, and
3. the sampling rate does not approach 1.

With these assumptions, the structure of the second and subsequent stages of sample selection is arbitrary. In the second framework,

1. the number of population clusters is fixed,
2. the secondary sampling units are selected with successively varying probabilities (with-replacement).

Prâšková and Sen (2009) note that the asymptotics of this framework are equivalent to stratified sampling under successive varying probabilities with-replacement.

In addition to consistency, an asymptotic framework allows us to linearize nonlinear estimators using the delta method. We use the delta method to explore the asymptotic bias and asymptotic variance. To compute the asymptotic bias and variance, the estimator is linearized and then the bias and variance of the linearized estimator are calculated.

The design bias, variance, mean squared error, consistency, and approximate bias are the primary measures used to evaluate design-based estimators. Calculating consistency and approximate bias requires using the superpopulation framework. For some estimators, it is impossible to analytically calculate the design bias, variance, and mean squared error without resorting to approximations such as linearization, numerical analysis, or simulations; however, consistency and approximate bias can be calculated under the superpopulation framework. In Appendix A.1 on page 258, we present details of the asymptotic framework used in this dissertation.

1.1.8 Discussion

Despite its prevalence, the design-based framework has not been impervious to criticism. Although an estimator could be unbiased under repeated sampling, an individual sample might produce estimates that are very different from the true parameters. Likewise, estimates of standard errors can be quite skewed for some samples. Neyman (1934, p. 586) acknowledged this when he wrote that probability sampling

> does not mean that we shall always get correct results when using this method. On the contrary, erroneous judgments must happen, but it is known how often they will happen in the long run.

Many of the critiques of the design-based framework emphasize that inference from unbalanced samples can be quite misleading. A balanced sample is one in which the sample moments were similar to population moments for covariates. For example, Cumberland and Royall (1988) showed that even though simple random sampling produced balanced samples on average, unbalanced samples were rather common. Given the possibility of an unbalanced sample, Royall (1970, p. 385) argued that "it is hard to give a useful general rigorous justification for letting a random device decide which units should be observed." Furthermore, in a simulation Cumberland and Royall showed that inferences from unbalanced samples were quite poor, even for large samples. Cumberland and Royall concluded that it was essential to select balanced samples through restricted sampling, systematic sampling, or stratified sampling, even when the sample size is large.

When the sample size is small, restricted, systematic, and stratified sample designs may not produce balanced samples. For small to moderate sized samples, the design-

based framework has limited appeal because design-based estimators tend to have large sampling errors and it is difficult to assure that the sample is balanced. Even under the ideal situation when the sample is balanced and the estimator is unbiased, the sample size or domain size may be too small and the sampling variance may be so large that the estimate is not usable or credible. In such cases mild assumptions about an underlying model may be able to greatly reduce the variability of point estimates under repeated sampling.

A slew of nonsampling errors can also threaten the balance of a sample. Survey methodologists are keenly aware that sampling error is only one source of many errors that occur in surveys. Nonsampling errors such as nonresponse and measurement errors often cannot be addressed without resorting to models. These errors have potential to add variability and bias to estimators. Unfortunately, the design-based assumption that everything is fixed, except for the sampling mechanism, is simplistic in the presence of non-sampling errors. In reality, random and systematic errors are introduced from many other sources including measurement errors, coverage errors, nonresponse errors, processing errors, and post-adjustment errors (Groves et al. 2004).

Little (2004) showed that some sample designs make assumptions which can lead to erroneous estimates. For example, πps samples implicitly assume a linear relationship between $\mathbf{y}$ and the size measures. In an often cited essay, Basu (1971) provided an example in which the π-estimator led to preposterous results. Design-based statisticians must be cautious when selecting probability proportional to size samples, because in cases where the size measure is inversely proportional to characteristics of interest, ridiculous results may arise. This is a particular concern when little information is available about some of

the characteristics of interest or when many measures are being estimated from a single survey.

In 1983, Hansen et al. defended the design-based theory for large samples. They argued,

> that modeling is an important tool for use in designing probability samples but that, with large samples, models can and should be used within the framework of probability-sampling inference. Thus, design decisions may be guided and evaluated by models, but inferences concerning population characteristics should be made on the basis of the induced randomization, at least when samples are reasonably large.

Naturally, if a sample is unbalanced, point estimates and estimates of their standard errors can be far from their expected values and the true population values as well. There are numerous methods used to select balanced samples, some of which involve probability sampling and others do not involve randomization at all. Certainly stratification and selecting samples proportional to size measures can help reduce the risk of selecting a sample that doesn't look like the population, but design-based estimation provides limited techniques to improve estimation given that a sample is not balanced. The model-based statisticians argue that it was much more important to have a balanced sample than a random sample.

1.2 Model-Assisted Frameworks

As its name implies, the model-assisted framework combines both the design-based and model-based frameworks. Models are used to reduce variance, but estimators are constructed so that they are approximately design-unbiased. By constructing asymptotically design-unbiased estimators, inference is protected against model misspecification. The main advantage of using a model to assist the design-based estimation is that estimators with smaller sampling errors can be constructed. This occurs when the model is correctly specified.

There are five model-assisted estimators covered in this dissertation: the generalized regression estimator, the generalized difference estimator, the calibration estimator, the model-calibrated estimator, and the model-calibrated maximum pseudoempirical likelihood estimator. All five estimators can produce approximately design-unbiased estimates and are quite efficient when the underlying models are correctly specified. We conclude this section with a description of the pseudomaximum likelihood estimation.

1.2.1 Generalized Regression Estimator

Generalized Regression (GREG) is widely used in the production of official statistics (Estevao et al. 1995) and is a popular method used to form descriptive statistics from survey data (Hidiroglou et al. 1995). Generalized Regression is attractive because it results in a common set of weights that can be used for all variables in a dataset, estimated totals from the survey can be made to match known population controls, and often the sampling variance of an estimator is reduced through borrowing strength from an assist-

ing model.

Cassel et al. (1976), Särndal (1980b), Särndal (1982), Isaki and Fuller (1982), and Wright (1983) laid the foundation for GREG estimation. Robinson and Särndal (1983) showed that the GREG is design-consistent and asymptotically design-unbiased in single-stage samples.

1.2.1.1 Point Estimator

Särndal (2007) formulates the GREG estimator, denoted $\hat{t}_y^{gr}$, in an easily interpretable way,

$$\hat{t}_y^{gr} = \sum_{\mathscr{U}} \hat{y}_k + \sum_{\mathfrak{s}} \frac{1}{\pi_k} (y_k - \hat{y}_k) \tag{1.11}$$

$$= \sum_{\mathfrak{s}} \frac{1}{\pi_k} y_k + \left(\sum_{\mathscr{U}} \hat{y}_k - \sum_{\mathfrak{s}} \frac{1}{\pi_k} \hat{y}_k \right). \tag{1.12}$$

When written in the form of (1.11) we see that the GREG estimator is the projective estimator, $\hat{t}_y^{pro} = \sum_{k \in \mathscr{U}} \hat{y}_k$, with a weighted residual adjustment. Predicted values are derived from a linear model

$$\hat{y}_k = \mathbf{x}_k^\top \widehat{\mathbf{B}}$$

where

$$\widehat{\mathbf{B}} = \left(\sum_{\mathfrak{s}} d_k q_k \mathbf{x}_k \mathbf{x}_k^\top \right)^{-1} \left(\sum_{\mathfrak{s}} d_k q_k \mathbf{x}_k y_k \right)$$

and q_k is chosen by the statistician. A familiar case of the GREG estimator occurs if $q_k = \frac{1}{\sigma_k^2}$. In this case, the $\hat{y}_k$'s will be the same predicted values from generalized least squares regression. Moreover, if $q_k = \frac{1}{\sigma^2 x_k}$ and there is just one auxiliary variable, then

$\hat{t}_y^{gr}$ reduces to the ratio estimator. According to Särndal (2007) and Valliant et al. (2000), q_k is commonly set to 1 for all k.

If we consider our population as being one realization of a superpopulation, then our focus shifts from estimating $\boldsymbol{\beta}$ to estimating $\mathbf{B}$, where $\boldsymbol{\beta}$ is the superpopulation model-based coefficient parameter and $\mathbf{B}$ is the realization of $\boldsymbol{\beta}$ for our finite population. That is, $\mathbf{B}$ is the maximum likelihood estimate of $\boldsymbol{\beta}$ that would be obtained if the sample contained the entire finite population. With a sample, $\mathbf{B}$ can be estimated. Lehtonen and Pahkinen (2004) review several techniques that incorporate the weights and sample design to estimate $\mathbf{B}$ from complex survey data. Binder (1983) and Firth and Bennett (1998) also focus on design-consistent methods to estimate the finite population quantity $\mathbf{B}$.

When written as Equation (1.12), we can easily see that the GREG estimator is equal to the π-estimator plus an adjustment factor equal to $\sum_{\mathscr{U}} \hat{y}_k - \sum_{\mathfrak{s}} \frac{1}{\pi_k}\hat{y}_k$. If the predicted population total, $\sum_{\mathscr{U}} \hat{y}_k$, is close to the weighted estimated population total, $\sum_{\mathfrak{s}} \frac{1}{\pi_k}\hat{y}_k$, then the adjustment will be small. However, if the predicted total is far from the weighted total, then the adjustment will move the GREG estimator away from the π-estimator and closer to the model estimate.

Although the GREG estimator is design-consistent regardless of the form of the assisting model, the sampling error of the GREG estimator is a function of the assisting model. Usually, the GREG estimator has smaller variance than the π-estimator because it makes use of auxiliary information. The gains in efficiency are a function of the relationship between $\mathbf{y}$ and $\mathbf{x}$. Särndal et al. (1992, p. 226) explain that

the adjustment term will often be negatively correlated with the error of the π-estimator. For samples in which the π-estimator alone gives a large error, the adjustment term will be about equally large as this error, but of the opposite sign, when the sample is fairly large and the linear relationship strong. Thus, the GREG will have a smaller error than the π-estimator.

Indeed, assisting models that fit the data well will generally result in estimators that have lower sampling variance than GREG estimators based on poorly fit assisting models. Särndal (2007) reviews many of the advantages to using the GREG estimator over design-based methods that are not assisted by a model.

Additional formulas for the GREG estimator abound. Särndal et al. (1992, p. 234) summarize at least five different forms of the GREG estimator in single-stage samples. For example, the GREG estimator can also be written in terms of matrices as

$$\hat{t}_y^{gr} = \hat{t}_y^{\pi} + \widehat{\mathbf{B}}^{\top} \left(\mathbf{t}_x - \hat{\mathbf{t}}_x^{\pi} \right)$$

where

$$\begin{aligned}
\mathbf{t}_x &= \left(\underset{N\times 1}{\mathbf{1}} \right)^{\top} \underset{N\times p}{\mathbf{X}} \\
\hat{\mathbf{t}}_x^{\pi} &= \left(\underset{n\times 1}{\mathbf{1}} \right)^{\top} \underset{n\times n}{\mathbf{\Pi}^{-1}} \underset{n\times 1}{\mathbf{X}} \\
\underset{p\times 1}{\widehat{\mathbf{B}}} &= \left(\mathbf{X}^{\top} \mathbf{\Pi}^{-1} \mathbf{Q} \mathbf{X} \right)^{-1} \mathbf{X}^{\top} \mathbf{\Pi}^{-1} \mathbf{Q} \mathbf{y}
\end{aligned}$$

and $\mathbf{Q}$ is a diagonal n by n matrix containing q_k for the kth element. Särndal (1980a) showed that the GREG estimator is a design-consistent estimator.

Särndal et al. (1992, p. 324) note that the GREG estimator of a total can be written

as:

$$\hat{t}_y^{gr} = \mathbf{g}^\top \mathbf{\Pi}^{-1} \mathbf{y} \tag{1.13}$$

where

$$\underset{n\times 1}{\mathbf{g}} = \mathbf{QX}\left(\mathbf{X}^\top \mathbf{\Pi}^{-1}\mathbf{QX}\right)^{-1}\left(\mathbf{t}_x - \hat{\mathbf{t}}_x^{\pi}\right) + \underset{n\times 1}{\mathbf{1}} \tag{1.14}$$

In this form, we see that we can interpret $\mathbf{g}$ as a vector of weight adjustments to the π-estimator. The new weights, $\mathbf{g}^\top \mathbf{\Pi}^{-1}$, are often called the calibration or GREG weights.

When the sample is selected in multiple stages, the general form is the same, provided that $\mathbf{\Pi}$ is a diagonal matrix containing the unconditional probabilities of selection and $\mathbf{Q}$ is a diagonal matrix.

1.2.1.2 Variance Estimator

Along with every point estimate, it is essential to also estimate the variability of the estimator. The sampling error is widely used to form confidence intervals, to test hypotheses, to assess the quality of the estimate, and to make inference to the finite population.

From the design-based framework, the GREG estimator is a nonlinear function in δ_k because the $\widehat{\mathbf{B}}_{\mathfrak{s}}$ term contains δ_k^{-1}. This nonlinearity makes it impossible to analytically calculate the design-based expectation of the GREG estimator in closed form. Särndal et al. (1992, p. 236) found the Taylor Series expansion of the GREG estimator and took the expectation of the linearized estimator. In this way, Särndal et al. (1992) showed that the GREG estimator was approximately unbiased. Using the Taylor series expansion,

Särndal et al. (1992, p. 235) found that the GREG estimator can be approximated by:

$$\hat{t}_y^{agr} = \sum_{\mathfrak{s}} \frac{1}{\pi_k} y_k + \sum_{\mathscr{U}} \mathbf{x}_k^\top \mathbf{B} - \sum_{\mathfrak{s}} \frac{1}{\pi_k} \mathbf{x}_k^\top \mathbf{B}$$

where

$$\underset{p\times 1}{\mathbf{B}} = \left(\mathbf{X}^\top \mathbf{Q} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Q} \mathbf{y}.$$

The linear approximation of the GREG estimator is an approximately design-unbiased estimator of the population total.

Särndal et al. (1992) then found the variance of the linearized GREG estimator to be

$$\mathrm{av}\left(\hat{t}_y^{gr}\right) = \sum_{k\in\mathscr{U}} \sum_{l\in\mathscr{U}} \left(\pi_{kl} - \pi_k \pi_l\right) \frac{E_k}{\pi_k} \frac{E_l}{\pi_l}$$

where

$$E_k = y_k - \mathbf{x}_k^\top \mathbf{B}. \tag{1.15}$$

This variance is called the asymptotic variance and denoted av.

Unless a complete sample is taken, y_k and $\mathbf{B}$ are not known for every element. Thus, this approximate variance must be estimated. Replacing the population quantities with the weighted sample estimates gives

$$\upsilon_e = \sum \sum_{\mathfrak{s}} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{e_k}{\pi_k} \frac{e_l}{\pi_l} \tag{1.16}$$

where

$$e_k = y_k - \mathbf{x}_k^\top \widehat{\mathbf{B}}. \tag{1.17}$$

Särndal et al. (1992, p. 176) remark that

> We caution that the Taylor linearization method has a tendency to lead to underestimated variances in not so large samples. The complexity of the statistic is a factor of importance. For a simple statistic, such as the weighted sample mean, the underestimation of the Taylor variance estimator may be without consequence even for modest sample sizes, but for complex statistics such as an estimator of a population variance, covariance, or correlation coefficient, fairly large samples may be required before the bias is negligible.

In an effort to not underestimate the variance, Särndal et al. (1989) and Estevao et al. (1995) recommend using the g-weights when estimating the variance. The preferred variance estimator is

$$\upsilon_g = \sum \sum_{\mathfrak{s}} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \frac{g_k e_k}{\pi_k} \frac{g_l e_l}{\pi_l} \tag{1.18}$$

where g_k is defined in Equation (1.14) and e_k is defined in Equation (1.17). Kott (1990) also proposed a Yates-Grundy type variance estimator for the GREG estimator in single-staged samples. Under regularity conditions, he showed that his variance estimator is design-consistent.

Under Poisson sampling, Särndal et al. (1989) propose estimating the asymptotic variance with

$$\upsilon_g \left(\hat{t}_y^{gr}\right) = \sum_{k \in \mathfrak{s}} \frac{\pi_k \left(1 - \pi_k\right)}{\pi_k} \left(g_k \frac{e_k}{\pi_k}\right)^2.$$

where g_k is defined in Equation (1.14) and e_k is defined in Equation (1.17).

When the joint inclusion probabilities are too difficult to compute or unknown to the analyst, both υ_e and υ_g cannot be calculated. However, if one assumes that the sample was

selected with-replacement, then v_e and v_g simplify. For example, for a $ppswr$ sample, the asymptotic variance can be estimated by

$$v_{wr} = \frac{1}{n\,(n-1)} \sum_{k \in \mathfrak{s}} \left(\frac{e_k}{\pi_k} - \frac{1}{n} \sum_{k \in \mathfrak{s}} \frac{e_k}{\pi_k} \right)^2$$

which is an application of Equation (1.3). Or, the asymptotic variance can be estimated by

$$v_{JL} = \frac{1}{n\,(n-1)} \sum_{k \in \mathfrak{s}} \left(\frac{g_k e_k}{\pi_k} - \frac{1}{n} \sum_{k \in \mathfrak{s}} \frac{g_k e_k}{\pi_k} \right)^2 .$$

The SUPERCARP software uses $v_{wr,g}$ for variance estimation (Hidiroglou et al. 1980). Yung and Rao (1996) also develop a simple variance estimator for the GREG estimator by linearizing the GREG estimator and then inserting the linearized GREG into the Jackknife formula. Their resulting estimator is equivalent to v_{JL}, the g-weighted with-replacement estimator.

Although v_{wr} and v_{JL} are popular variance estimators, they are based on the assumption that sample units were selected with-replacement. This may be problematic in practice, especially when the sampling fraction is rather high. Alternative variance estimation techniques could help correct for violating this assumption. Särndal et al. (1992, sec 4.6) note that with-replacement variance estimators have the potential to either over or under estimate the true sampling variance, depending on the sample design.

The delta method is often used to estimate the variance of GREG estimators, but such estimators tend to underestimate the true sampling error, especially in small to moderate sized samples. Alternative variance estimation techniques such as the jackknife and bootstrap are more attractive than linearization; but can be cumbersome to implement

and require extensive computational resources. Borrowing from the model-based theory of robust variance estimation, Valliant (2002) showed that leverage adjustments could be used to improve the linearized estimators in single-staged samples.

Valliant et al. (2000) applied sandwich estimation principles to develop variance estimators of the GREG estimator under one stage of sampling. For example, the basic sandwich variance estimator is

$$\upsilon_{r1} = \sum_{k \in \mathfrak{s}} a_k^2 e_k^2$$

where

$$a_k = \frac{g_k}{\pi_k} - 1.$$

A similar estimator is

$$\upsilon_{r2} = \sum_{k \in \mathfrak{s}} \frac{g_k^2}{\pi_k^2} e_k^2.$$

Using adjusted residuals, Valliant et al. (2000) showed that the following estimator approximates the jackknife,

$$\upsilon_J = \sum_{k \in \mathfrak{s}} \left[\frac{g_k e_k}{\pi_k (1 - h_{ii})} \right]^2 - \frac{1}{n} \left[\sum_{k \in \mathfrak{s}} \frac{g_k e_k}{\pi_k (1 - h_{ii})} \right]^2.$$

which could further be approximated by

$$\upsilon_J^\star = \sum_{k \in \mathfrak{s}} \left[\frac{g_k^2 e_k^2}{\pi_k^2 (1 - h_{ii})} \right].$$

Valliant (2002) reviewed the sampling literature for various variance estimators used with the GREG estimator and also constructed several variance estimators of his own. Using a simulation, he compared the root mean squared error, confidence interval coverage, and relative bias of a variety of variance estimators. Valliant (2002) concluded that

estimators can easily be constructed that are approximately unbiased for both the design-variance and, under certain models, the model-variance. Moreover, the dual-purpose estimators studied here are robust estimators of a model-variance even if the model that motivates the GREG has an incorrect variance parameter.

Further, Valliant (2002) noted that "a key feature of the best of these estimators is the adjustment of squared residuals by factors analogous to the leverages used in standard regression analysis." Thus, Valliant (2002) successfully used leverage adjustments to improve the variance estimation of GREG estimators in one stage of sampling. In Chapter 2, we develop similar variance estimators for two-staged samples.

1.2.1.3 Point Estimation in Two Staged Samples

Särndal et al. (1992, ch. 8) discuss three different GREG estimators that can be used in clustered samples. These three estimators depend on the available data. Case B occurs when unit level data are available for the complete sample and control totals are available for the population. In this case, they write the GREG estimator as

$$\hat{t}_y^{gr} = \sum_{k \in \mathscr{U}} \widehat{y}_k + \sum_{k \in \mathfrak{s}} \frac{e_{k\mathfrak{s}}}{\pi_k}.$$

To emphasize that this is estimated from a clustered sample design, we can also write this as

$$\hat{t}_y^{gr} = \sum_{i \in \mathscr{U}_I} \sum_{k \in \mathscr{U}_i} \widehat{y}_k + \sum_{\mathfrak{s}_I} \frac{1}{\pi_i} \sum_{\mathfrak{s}_i} \frac{y_k - \hat{y}_k}{\pi_{k|i}} \tag{1.19}$$

where

$$\widehat{y}_k = \mathbf{x}_k^\top \widehat{\mathbf{B}}.$$

In two-staged samples, the GREG estimator is also design-consistent.

1.2.1.4 Variance of GREG in Two Staged Samples

The GREG estimator is a nonlinear estimator with respect to the sample design because $\mathbf{A}_\pi^{-1} = \left(\mathbf{X}^\top \mathbf{\Pi}^{-1} \mathbf{Q} \mathbf{X}\right)^{-1}$ involves inverting sample quantities. As a nonlinear estimator, the exact design-based variance cannot be calculated. Instead, the asymptotic variance of $\hat{t}_y^{gr}$ is calculated using the delta method.

We denote the true sampling variance of the GREG estimator in clustered samples by $\mathrm{var}\left(\hat{t}_y^{gr}\right)$. Assuming that the second stage sample design is invariant and independent, Särndal et al. (1992, p. 325) show that the asymptotic variance of $\hat{t}_y^{gr}$ is

$$\begin{aligned}
\mathrm{av}\left(\hat{t}_y^{gr}\right) &= \sum_{i=1}^{N}\sum_{j=1}^{N} \Delta_{ij} \frac{t_{Ei}}{\pi_i}\frac{t_{Ej}}{\pi_j} + \sum_{i=1}^{N} \frac{\sum_{k\in U_i}\sum_{l\in U_i} \Delta_{kl|i} \frac{E_k}{\pi_{k|i}}\frac{E_l}{\pi_{l|i}}}{\pi_i} \\
&= \sum_{i\in\mathscr{U}_I}\sum_{j\in\mathscr{U}_I} \Delta_{ij} \frac{t_{Ei}}{\pi_i}\frac{t_{Ej}}{\pi_j} + \sum_{i\in\mathscr{U}_I} \frac{\sum_{k\in\mathscr{U}_i}\sum_{l\in\mathscr{U}_i} \Delta_{kl|i} \frac{E_k}{\pi_{k|i}}\frac{E_l}{\pi_{l|i}}}{\pi_i} \qquad (1.20) \\
&= \sum\sum_{\mathscr{U}_I} \Delta_{ij} \frac{t_{Ei}}{\pi_i}\frac{t_{Ej}}{\pi_{Ij}} + \sum_{\mathscr{U}_I} \frac{\sum\sum_{\mathscr{U}_i} \Delta_{kl|i} \frac{E_k}{\pi_{k|i}}\frac{E_l}{\pi_{l|i}}}{\pi_i}
\end{aligned}$$

where

$$t_{Ei} = \sum_{k=1}^{M_i} E_k$$

and E_k is defined in Equation (1.15).

Since the asymptotic variance depends on an unknown population vector, namely

$\mathbf{B}$, as well as nonsample units, the asymptotic variance cannot be calculated from a sample. Using the method of moments, a natural estimator for the asymptotic variance is,

$$\upsilon_e = \sum\sum_{\mathfrak{s}_I} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\hat{t}_{ei}}{\pi_{Ii}} \frac{\hat{t}_{ej}}{\pi_j} + \sum_{\mathfrak{s}_I} \frac{\sum\sum_{\mathfrak{s}_i} \frac{\Delta_{kl|i}}{\pi_{kl|i}} \frac{e_{k\mathfrak{s}}}{\pi_{k|i}} \frac{e_{l\mathfrak{s}}}{\pi_{l|i}}}{\pi_i} \tag{1.21}$$

where

$$\hat{t}_{ei} = \sum_{\mathfrak{s}_i} \frac{e_k}{\pi_{k|i}} \tag{1.22}$$

and e_k is defined Equation (1.17).

Särndal (1981) and Särndal (1982) proposed this type of variance estimator in single-staged samples. In 1989, Särndal et al. (1989) advocated that weighting the residuals by the g-weights improved inference. They also showed that the g-weighted variance estimator was design consistent. Extending this result to clustered samples, Särndal et al. (1992) propose estimating the asymptotic variance with,

$$\begin{aligned}
\upsilon_g &= \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\hat{t}^g_{e,i}}{\pi_i} \frac{\hat{t}^g_{e,j}}{\pi_j} + \sum_{i=1}^{n} \frac{\sum_{k\in s_i}\sum_{l\in s_i} \frac{\Delta_{kl|i}}{\pi_{kl|i}} \frac{g_k e_k}{\pi_{k|i}} \frac{g_l e_l}{\pi_{l|i}}}{\pi_i} \\
&= \sum_{i\in\mathfrak{s}_I}\sum_{j\in\mathfrak{s}_I} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\hat{t}^g_{ei}}{\pi_i} \frac{\hat{t}^g_{ej}}{\pi_j} + \sum_{i\in\mathfrak{s}_I} \frac{\sum_{k\in\mathfrak{s}_i}\sum_{l\in\mathfrak{s}_i} \frac{\Delta_{kl|i}}{\pi_{kl|i}} \frac{g_k e_k}{\pi_{k|i}} \frac{g_l e_l}{\pi_{l|i}}}{\pi_i} && (1.23) \\
&= \sum\sum_{\mathfrak{s}_I} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\hat{t}^g_{ei}}{\pi_i} \frac{\hat{t}^g_{ej}}{\pi_j} + \sum_{\mathfrak{s}_I} \frac{\sum\sum_{\mathfrak{s}_i} \frac{\Delta_{kl|i}}{\pi_{kl|i}} \frac{g_k e_k}{\pi_{k|i}} \frac{g_l e_l}{\pi_{l|i}}}{\pi_i} && (1.24)
\end{aligned}$$

where

$$\hat{t}^g_{e,i} = \sum_{\mathfrak{s}_i} \frac{g_{k\mathfrak{s}} e_{k\mathfrak{s}}}{\pi_{k|i}}. \tag{1.25}$$

Särndal et al. (1989) argue that υ_e tends to underestimate the true sampling error in practice for single-staged samples. For this reason, Särndal et al. (1992) recommend υ_g. All

of these variance estimators are cumbersome because they require knowledge of the first and second-stage joint inclusion probabilities.

If the first and second stage samples are selected using a Poisson sampling technique, then υ_g reduces to

$$\begin{aligned}\upsilon_g &= \sum_{i=1}^{n} \frac{(1-\pi_i)}{\pi_i^2} \left(\hat{t}_{e,i}^{g}\right)^2 + \sum_{i=1}^{n} \frac{1}{\pi_i} \sum_{k \in s_i} \frac{\left(1-\pi_{k|i}\right)}{\pi_{k|i}^2} g_k^2 e_k^2 \\ &= \sum_{i \in \mathfrak{s}_I} \frac{(1-\pi_i)}{\pi_i^2} \left(\hat{t}_{e,i}^{g}\right)^2 + \sum_{i \in \mathfrak{s}_I} \frac{1}{\pi_i} \sum_{k \in \mathfrak{s}_i} \frac{\left(1-\pi_{k|i}\right)}{\pi_{k|i}^2} g_k^2 e_k^2 \end{aligned} \tag{1.26}$$

where $\hat{t}_{e,i}^{g}$ is defined in Equation (1.25).

Commonly, with-replacement variance estimators are used even when the first stage sample was selected without-replacement. As long as the first-stage sampling fraction is relatively small, the bias of using a with-replacement variance estimator is relatively small. Särndal et al. (1992, sec 4.6) discuss the classic with-replacement variance estimator of a total under multiple stages of sampling. To construct a with-replacement variance estimator for the GREG estimator, we can modify (1.10) for the GREG estimator by replacing $\frac{\hat{t}_{yi}^{\pi}}{p_i}$ with $\hat{t}_{yi}^{gr\star} = \sum_{\mathscr{U}} \widehat{y}_k + \frac{\hat{t}_{ei}^{\star}}{p_i}$. Yung and Rao (1996) show that the variance of the GREG estimator in two-stages of with-replacement sampling can be estimated by

$$\begin{aligned}\upsilon_{wr} &= \frac{n}{(n-1)} \sum_{i=1}^{n} \left(\frac{\hat{t}_{e,i}}{\pi_i} - \frac{1}{n} \sum_{i=1}^{n} \frac{\hat{t}_{e,i}}{\pi_i} \right)^2 \\ &= \frac{1}{n(n-1)} \sum_{i \in \mathfrak{s}_I} \left(\sum_{\mathscr{U}} \widehat{y}_k + \frac{\hat{t}_{ei}}{p_i} - \hat{t}_y^{gr} \right)^2 \\ &= \frac{1}{n(n-1)} \sum_{i \in \mathfrak{s}_I} \left(\frac{\hat{t}_{ei}}{p_i} - \hat{t}_e \right)^2 . \end{aligned} \tag{1.27}$$

A more popular version of the with-replacement estimator uses the g-weights

$$\begin{aligned} v_{JL} &= \frac{1}{n\,(n-1)} \sum_{i \in \mathfrak{s}_I} \left(\frac{\hat{t}_{ei}^{g}}{p_i} - \hat{t}_{e}^{g} \right)^2 \\ &= \frac{n}{(n-1)} \sum_{i \in \mathfrak{s}_I} \left(\frac{\hat{t}_{ei}^{g}}{\pi_i} - \frac{1}{n} \sum_{k \in \mathfrak{s}_I} \frac{\hat{t}_{ei}^{g}}{\pi_i} \right)^2 . \end{aligned} \tag{1.28}$$

Although v_{wr} and v_{JL} are popular variance estimators, they are based on the assumption that sample clusters were selected with-replacement. Assuming that sample clusters are uncorrelated, even though they are not, may be problematic in practice. Alternative variance estimation techniques could help correct for violating this assumption. When the sampling fraction is large, the estimated variance is often multiplied by a finite correction factor of $1 - \frac{n}{N}$ to prevent the estimated variance from wildly overestimating the true variance. Särndal et al. (1992, sec 4.6) note that with-replacement variance estimators have the potential to either over or under estimate the true sampling variance, depending on the sample design.

Furthermore, v_{wr} and v_{JL} are estimators for the approximate variance of the GREG estimator. That is, they estimate $\mathrm{av}\left(\hat{t}_y^{gr}\right)$ rather than $\mathrm{var}\left(\hat{t}_y^{gr}\right)$. For this reason, neither v_{wr} nor v_{JL} include variability due to estimating $\mathbf{B}$. Thus, v_{wr} and v_{JL} may tend to underestimate the true variability of $\hat{t}_y^{gr}$ when there is considerable noise in estimating $\mathbf{B}$.

Because the GREG estimator is nonlinear, variance estimation is complicated. The delta method is often used to estimate the variance of the GREG estimator, but such estimators tend to underestimate the true sampling error, especially in small to moderate sized samples. Alternative variance estimation techniques such as the jackknife and bootstrap are more attractive than linearization; but can be cumbersome to implement and require extensive computational resources.

In conclusion, if covariates exist for all sample units and population controls for those covariates are available, Särndal et al. (1992) showed that the GREG estimator could be used to reduce the sampling variance of the π-estimator. They also showed that the GREG estimator was approximately design-unbiased for large samples. Using linearization, Särndal et al. provided a variance estimator for the GREG estimator. In official statistics, the GREG estimator is often used because it results in calibrated weights. Unlike the estimators from the prediction approach, the GREG estimator is approximately design-unbiased and incorporates unequal probabilities of selection into estimation. Indeed the GREG estimator has attractive design properties and usually has lower mean squared error than the π-estimator.

1.2.2 Generalized Difference Estimator

Chapters 3 and 4 focus on the generalized difference estimator. In many respects, the generalized difference estimator is equivalent to the GREG estimator, with the exception that an arbitrary model is used instead of the linear model. (Särndal et al. 1992, p. 222) define the difference estimator in a single-staged sample as

$$\hat{t}_y^{dif} = \sum_{i \in \mathscr{U}} \mathbf{a}^\top \mathbf{x}_i + \sum_{i \in \mathfrak{s}} \frac{1}{\pi_i} \left[y_i - \mathbf{a}^\top \mathbf{x}_i \right] \tag{1.29}$$

where $\mathbf{a}$ is an arbitrary vector known for all elements in the population.

The generalized difference estimator extends this estimator in two ways. First it treats $\mathbf{a}^\top \mathbf{x}_i$ as a prediction of y_i. That is, $\mathbf{a}^\top \mathbf{x}_i$ is replaced with $\mu\left(\mathbf{x}_i, \mathbf{B}\right)$ where μ is a function of known covariates $\mathbf{x}_i$ and a population parameter vector $\mathbf{B}$. In theory, gains in efficiency occur when $\mu\left(\mathbf{x}_i, \mathbf{B}\right)$ is a prediction of y_i based on a well fit model. The second

generalization is to replace $\mathbf{B}$ with an estimator. That is, actually, $\mu\left(\mathbf{x}_i, \widehat{\mathbf{B}}\right)$ is used in place of $\mathbf{a}^\top \mathbf{x}_i$. Thus, the generalized difference estimator is

$$\hat{t}_y^{gd} = \sum_{i \in \mathscr{U}} \mu\left(\mathbf{x}_i, \widehat{\mathbf{B}}\right) + \sum_{i \in \mathfrak{s}} \frac{1}{\pi_i}\left[y_i - \mu\left(\mathbf{x}_i, \widehat{\mathbf{B}}\right)\right] \tag{1.30}$$

where $\widehat{\mu}_i = \mu\left(\mathbf{x}_i, \widehat{\mathbf{B}}\right)$ is a prediction from an arbitrary model. When $\widehat{\mu}_i$ is estimated from a linear model, the generalized difference estimator reduces to the GREG estimator. However, when $\widehat{\mu}_i$ is found from some other model, it is not equivalent to the GREG estimator. Wu and Sitter (2001) summarize characteristics of the generalized difference estimator. Firth and Bennett (1998) also discuss properties of Equation (1.30), but they call it a difference estimator.

The variance of the generalized difference estimator is often found by using a GREG variance estimator with e_k defined as $y_k - \widehat{\mu}_k$ instead of the expression given in Equation (1.17).

In Chapters 3 and 4, we extend Equation (1.30) to cluster samples with multinomial logistic and general linear models for μ_k.

1.2.3 Calibrated Estimator

According to Deville and Särndal (1992), calibration estimators

> use calibrated weights, which are as close as possible, according to a given distance measure, to the original sampling design weights π_k^{-1} while also respecting a set of constraints, the calibration equations.

Typically the calibration equations are defined so that the weighed sum of auxiliary variables is equal to known population controls, that is, $\sum_{k \in \mathfrak{s}} w_k^{cal} \mathbf{x}_k = \mathbf{t}_x$ where w_k^{cal} is

the new calibration weight. The calibration property is especially attractive for official statistical agencies which seek to assure that key demographic estimates are consistent across surveys and equal to "known" population totals. Post-stratification, raking, and the generalized regression estimators are all examples of calibration estimators.

The primary analytic goal of calibration is to find a new vector of weights, $\mathbf{w}^{cal}$, that is minimal distance from the design weights, $\mathbf{d}$, and meets the constraints $\sum_{k \in \mathfrak{s}} w_k^{cal} \mathbf{x}_k = \mathbf{t}_x$. The calibrated weights depend on how one specifies the "distance" between the design weights and the calibrated weights. For example, Deville and Särndal (1992) show that the GREG is equivalent to calibration with a linear distance function equal to $\left(\mathbf{w}^{cal} - \mathbf{d}\right)^\top \mathbf{\Pi Q}^{-1} \left(\mathbf{w}^{cal} - \mathbf{d}\right)$.

Deville and Särndal (1992) also proved that calibration estimators are asymptotically equivalent to the GREG estimator, regardless of how one specifies the "distance." For this reason, Deville and Särndal (1992) suggest approximating the variance of calibrated estimators by simply using the GREG variance estimators.

Särndal (2007) reviewed several extensions of calibration to cluster samples. In cluster samples, the cluster weights, d_i, may be calibrated, the unit weights, d_k, may be calibrated, or both may be calibrated, depending on the available data and the analytic goals. Estevao and Särndal (2006) covered a number of different ways to calibrate data in cluster samples. When complete auxiliary data are available, the calibration estimator is

$$\hat{t}_y^{cal} = \left(\mathbf{w}^{cal}\right)^\top \mathbf{y} \tag{1.31}$$

where $\mathbf{w}^{cal}$ is found by minimizing

$$\left(\mathbf{w}^{cal} - \mathbf{d}\right)^{\top} \mathbf{\Pi Q}^{-1} \left(\mathbf{w}^{cal} - \mathbf{d}\right) = \sum_{k \in \mathfrak{s}} \frac{\left(w_k^{cal} - d_k\right)^2}{d_k q_k} \tag{1.32}$$

subject to the constraint

$$\sum_{k \in \mathscr{U}} \mathbf{x}_k = \sum_{k \in \mathfrak{s}} w_k^{cal} \mathbf{x}_k.$$

The variance of $\hat{t}_y^{cal}$ can be estimated with

$$\upsilon_e\left(\hat{t}_y^{cal}\right) = \sum_{i \in \mathfrak{s}_I} \sum_{j \in \mathfrak{s}_I} \frac{\Delta_{ij}}{\pi_{ij}} \hat{t}_{ei}^{cal} \hat{t}_{ej}^{cal} + \sum_{i \in \mathfrak{s}_I} \sum_{k \in \mathfrak{s}_i} \sum_{l \in \mathfrak{s}_i} \frac{\Delta_{kl|i}}{\pi_{kl|i}} w_k^{cal} e_k w_l^{cal} e_l$$

where $\hat{t}_{ei}^{cal} = \sum_{k \in \mathfrak{s}_i} w_k^{cal} e_k$.

As defined by Deville and Särndal (1992), the calibration constraints assure that the weighted auxiliary data equals known control totals. One advantage of this form of calibration is that one set of calibration weights can be created and used for all variables collected. Although calibration estimators are often more efficient than the π-estimator, further gains in efficiency can be made by building more specialized models for each response variable.

1.2.4 Model-Calibrated Estimator

Wu and Sitter (2001) extended calibration to cover nonlinear assisting models. They call their method, model-calibration. Instead of minimizing the distance between $\mathbf{d}$ and $\mathbf{w}^{cal}$ subject to $\sum_{k \in \mathfrak{s}} w_k^{cal} \mathbf{x}_k = \mathbf{t}_x$, they proposed minimizing the distance between $\mathbf{d}$ and $\mathbf{w}_k$ subject to $\frac{1}{N} \sum_{k \in \mathfrak{s}} w_k^{mc} = 1$ and $\sum_{k \in \mathfrak{s}} w_k^{mc} \widehat{\mu}_k = \sum_{k \in \mathscr{U}} \widehat{\mu}_k$ where $\widehat{\mu}_k$ is a prediction from a generalized linear model (GLM). After solving for w_k^{mc}, they estimated a finite

population total by $\hat{t}_y^{mc} = \sum_{k\in\mathfrak{s}} w_k^{mc} y_k$. With the linear distance measure, Equation (1.32) on page 50 t_y^{mc} can be explicitly written as

$$\hat{t}_y^{mc} = \widehat{\mathbf{t}}_y^{\pi} + \left(\sum_{k\in\mathscr{U}} \widehat{\mu}_k - \sum_{k\in\mathfrak{s}} d_k \widehat{\mu}_k \right) \widehat{\mathbf{B}}^{mc} \tag{1.33}$$

where

$$\widehat{\mathbf{B}}^{mc} = \frac{\sum_{k\in\mathfrak{s}} d_k q_k \left(\widehat{\mu}_k - \bar{\mu}\right)\left(y_k - \bar{y}\right)}{\sum_{k\in\mathfrak{s}} d_k q_k \left(\widehat{\mu}_k - \bar{\mu}\right)^2} \tag{1.34}$$

$$\bar{\mu}_c = \frac{\sum_{k\in\mathfrak{s}} d_k q_k \widehat{\mu}_k}{\sum_{k\in\mathfrak{s}} d_k q_k}. \tag{1.35}$$

Wu and Sitter (2001) also found the asymptotic variance of $\hat{t}_y^{mc}$ in single-staged samples to be

$$\operatorname{av}\left(\hat{t}_y^{mc}\right) = \sum_{k\in\mathscr{U}} \sum_{l\in\mathscr{U}} \Delta_{kl} \left(\frac{y_k - \widehat{\mu}_k \mathbf{B}^{mc}}{\pi_k} \right) \left(\frac{y_l - \widehat{\mu}_l \mathbf{B}^{mc}}{\pi_l} \right)$$

where

$$\mathbf{B}^{mc} = \frac{\sum_{k\in l\in\mathscr{U}} q_k \left(\mu_k - \bar{\mu}\right)\left(y_k - \bar{y}\right)}{\sum_{k\in l\in\mathscr{U}} q_k \left(\mu_k - \bar{\mu}\right)^2}$$

$$\bar{\mu} = \frac{1}{N} \sum_{k\in\mathscr{U}} \mu_k.$$

Under Poisson sampling the asymptotic variance simplifies to

$$\operatorname{av}\left(\hat{t}_y^{mc}\right) = \sum_{k\in\mathscr{U}} \pi_k \left(1 - \pi_k\right) \left(\frac{y_k - \mu_k \mathbf{B}^{mc}}{\pi_k} \right)^2$$

which can be estimated by

$$\upsilon_e\left(\hat{t}_y^{mc}\right) = \sum_{k\in\mathfrak{s}} \frac{\pi_k \left(1 - \pi_k\right)}{\pi_k} \left(\frac{y_k - \widehat{\mu}_k \widehat{\mathbf{B}}^{mc}}{\pi_k} \right)^2.$$

One advantage of the model-calibrated estimator is that it can improve design-based inference through nonlinear models. Since GLMs tend to fit data generated by nonlinear models better than linear regression, it seems advantageous to use model-calibration when analyzing nonlinear data. In this dissertation, general model-calibration is developed for two-staged samples. Kim et al. (2009) discuss nonparametric calibration in cluster samples, but they do not cover nonlinear models.

1.2.5 Model-Calibrated Pseudoempirical Maximum Likelihood Estimator

Chen and Qin (1993) describe the pseudoempirical likelihood approach under simple random sampling. Zhong and Rao (1996) and Chen and Sitter (1999) developed the pseudoempirical likelihood approach to complex survey designs. Wu and Sitter (2001) introduce model-calibrated constraints to the pseudoempirical likelihood approach.

The pseudoempirical likelihood approach is motivated by treating y_k in the population as a random variable with density p_k^{pe}. The empirical likelihood of $\mathbf{y}$ is

$$L\left(\mathbf{p}^{pe}\right) = \prod_{k \in \mathscr{U}} p_k^{pe}$$

and the log likelihood is

$$\ell\left(\mathbf{p}^{pe}\right) = \sum_{k \in \mathscr{U}} \log p_k^{pe}.$$

Unless a census is taken, the empirical likelihood must be estimated. Thus the pseudoempirical log likelihood is

$$\hat{\ell}\left(\mathbf{p}^{pe}\right) = \sum_{k \in \mathfrak{s}} d_k \log p_k^{pe}.$$

Following the theory of maximum likelihood, the pseudoempirical log likelihood is maximized. Furthermore, constraints are added to improve the efficiency of the estimators. The pseudoempirical log likelihood is maximized subject to

$$\sum_{k \in \mathfrak{s}} p_k^{pe} = 1$$

$$\sum_{k \in \mathfrak{s}} p_k^{pe} \mathbf{u}_k = 0.$$

Once we estimate p_k^{pe}, we can estimate the mean of our variable with

$$\widehat{\bar{y}}^{pe} = \sum_{k \in \mathfrak{s}} \hat{p}_k^{pe} y_k. \tag{1.36}$$

Chen and Qin (1993), Zhong and Rao (1996), and Chen and Sitter (1999) discuss estimators where $u_k = \mathbf{x}_k - \bar{\mathbf{x}}$, which reduces to the GREG weights. Wu and Sitter (2001) generalize this method to the case where $u_k = \widehat{\mu}_k - \frac{1}{N}\sum_{k=1}^{N} \widehat{\mu}_k$ where $\widehat{\mu}_k$ is a prediction from a generalized linear model.

1.2.6 Pseudo Maximum Likelihood Estimation

The method of maximum likelihood can be used to estimate superpopulation parameters. Binder (1983) extended this method to complex survey analysis by incorporating the survey weights into the log-likelihood equations. This general method that Binder (1983) described is called Pseudo Maximum Likelihood (PML) or implicit differentiation. PML uses linearization and estimating equations to produce design-consistent

estimators of finite population parameters. It is especially useful when the parameter of interest cannot be solved explicitly in closed form. Both Binder (1983) and Särndal et al. (1992, section 13.4) give several examples of how PML can be used to construct design-consistent estimators of coefficients from a nonlinear regression model. An advantage of implicit differentiation is that variance estimators can easily be computed from the estimating equations.

In this general method, the density of $\mathbf{y}_k$ is written as a function of explanatory variables and unknown superpopulation parameters, $\boldsymbol{\beta}$. We write this density as $f\left(y_k; \boldsymbol{\beta}\right)$. Then, following the theory of maximum likelihood, the likelihood of $\boldsymbol{\beta}$ is written as $L\left(\boldsymbol{\beta}; y_k\right)$. Maximizing this likelihood often involves first taking the log of the likelihood, denoted $\ell\left(\boldsymbol{\beta}\right)$. Then the derivative of the log-likelihood is taken and set to zero. Differentiating the log-likelihood gives us the minima and maxima of the likelihood. We call the derivative of the log likelihood the estimating equations. The maximum of the likelihood is the population parameter $\mathbf{B}$. In general, the estimating equations are written as

$$W\left(\beta\right) = \sum_{k \in \mathscr{U}} \mathbf{U}\left(y_k, \beta\right) - \mathbf{v}\left(\beta\right)$$

where $\mathbf{U}$ and $\mathbf{v}$ are determined by the likelihood function. Setting these estimating equations equal to zero and solving them gives the maximum and minimum points of the likelihood function. The solution to these estimating equations is the finite population parameter $\mathbf{B}$.

When a sample is selected, Binder (1983) showed that an unbiased estimate of these

estimation equations is,

$$\widehat{W}(\mathbf{B}) = \sum_{k \in \mathfrak{s}} d_k \mathbf{U}(y_k, \mathbf{B}) - \mathbf{V}(\mathbf{B}).$$

These equations are known as the pseudo log-likelihood estimating equations. Solving $\widehat{W}(\mathbf{B}) = \mathbf{0}$ for $\mathbf{B}$ gives the estimator $\widehat{\mathbf{B}}$ which is known as the pseudomaximum likelihood estimator.

More generally, we can replace the coefficient vector $\mathbf{B}$ in the previous equations with any population quantity $\boldsymbol{\theta}$. Using the delta method, Binder (1983) further showed that $\widehat{\boldsymbol{\theta}}$ is asymptotically normal under mild regularity conditions. Furthermore, the asymptotic variance of $\widehat{\boldsymbol{\theta}}$ is

$$\operatorname{av}\left(\widehat{\boldsymbol{\theta}}\right) = \left[\widehat{\mathbf{J}}^{-1}(\boldsymbol{\theta})\right] \left[\boldsymbol{\Sigma}(\boldsymbol{\theta})\right] \left[\widehat{\mathbf{J}}^{-1}(\boldsymbol{\theta})\right]^{\top}$$

where $\widehat{\mathbf{J}}(\boldsymbol{\theta})$ is the matrix of first order partial derivatives for the estimating equations taken with respect to $\boldsymbol{\theta}$. That is

$$\widehat{\mathbf{J}}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \widehat{\mathbf{W}}(\boldsymbol{\theta})$$

and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is symmetric matrix of design-based covariances among the sample estimating equations $\widehat{U}_k(\boldsymbol{\theta})$ and $\widehat{U}_l(\boldsymbol{\theta})$. That is

$$\boldsymbol{\Sigma}_{\mathbf{U}}(\boldsymbol{\theta}) = \operatorname{var}\left[\widehat{\mathbf{U}}(\boldsymbol{\theta})\right].$$

Specifically, we denote the element of in the ith row and the jth column by

$$\sigma_{kl} = \sum \sum_{\mathscr{U}} \Delta_{kl} \widehat{U}_k(y_k, \boldsymbol{\theta}) \widehat{U}_l(y_l, \boldsymbol{\theta})$$

where $\widehat{U}_k$ is the kth element of the vector $\mathbf{U}(\boldsymbol{\theta})$.

We can estimate our asymptotic variance by

$$\upsilon_{Binder}\left(\widehat{\boldsymbol{\theta}}\right) = \left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right]\left[\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right)\right]\left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{\top}$$

where

$$\widehat{\mathbf{J}}\left(\widehat{\boldsymbol{\theta}}\right) = \frac{\partial}{\partial\widehat{\boldsymbol{\theta}}}\widehat{\mathbf{W}}\left(\widehat{\boldsymbol{\theta}}\right)$$

and $\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right)$ is composed of

$$\widehat{\sigma}_{kl} = \sum\sum_{\mathfrak{s}}\frac{\Delta_{kl}}{\pi_{ij}}\widehat{U}_k\left(y_k,\widehat{\boldsymbol{\theta}}\right)\widehat{U}_l\left(y_l,\widehat{\boldsymbol{\theta}}\right).$$

Binder (1983), Roberts et al. (1987), Särndal et al. (1992, section 13.4), RTI (2004), and Lehtonen and Pahkinen (2004) all discuss pseudomaximum likelihood methods to estimate logistic regression models.

The pseudomaximum likelihood estimating equations result in design-based estimates of $\mathbf{B}$. However, more work must be done to estimate descriptive statistics, such as finite population means and totals.

1.3 Conclusion

This chapter introduced design-based and model-assisted estimation. As we saw, design-based inferences are made with respect to the sample design and all population quantities are treated as fixed constants. Model-assisted estimators borrow strength from models, but are design-consistent. Thus, they have attractive design-based and model-based properties. In the next three chapters of this dissertation, we use both model-based and design-based techniques to extend and improve model-assisted estimators.

Chapter 2

Improved Variance Estimators for Generalized Regression (GREG) Estimators in Cluster Samples

2.1 Introduction

Generalized regression (GREG) estimation is a common technique used to calibrate estimates, reduce sampling errors, and correct for nonsampling errors. Official surveys of household data often use generalized regression to calibrate sample-based estimates to population controls, assure consistent estimates of demographic characteristics across surveys, and reduce nonresponse and undercoverage errors. Generalized regression estimation is also frequently used because it tends to result in smaller sampling errors than other design-based estimators.

Because generalized regression estimation is frequently used in official statistics and policy analysis, it is critical to accurately and precisely measure the sampling variability of such estimates. Sampling error plays a central role in the analysis of survey data. Accurate estimates of sampling errors are necessary for confidence interval construction, hypothesis testing, quality assessment, design effect analysis, sample size determination, decision making, and inference. Inaccurate estimates of sampling errors can undermine decisions and threaten analysis.

Popular techniques used to estimate the sampling errors of calibrated totals in com-

plex samples either require extensive computational resources or tend to underestimate the true sampling errors, especially with small to moderate sample sizes. There are two popular techniques used to estimate the sampling variance of GREG estimators: linearization and replication. On the one hand, current linearization estimators (Särndal et al. (1989)) may not converge to the true sampling error fast enough to produce accurate results in small to moderate samples. As we noted in Chapter 1, Särndal et al. (1992, p. 176) remark that "For complex statistics such as an estimator of a population variance, covariance, or correlation coefficient, fairly large samples may be required before the bias is negligible." On the other hand, replication techniques such as the jackknife and the bootstrap can be computationally demanding.

Leverage-adjusted sandwich estimators provide an alternative approach to estimating design-based sampling errors that also have model-based justifications. From a model-based framework, Long and Ervin (2000) and MacKinnon and White (1985) demonstrated how the sandwich estimator could be used for variance estimation even when the variance component of the working model was misspecified. Valliant (2002) took this approach to estimate the design-based variance of GREG estimators under one stage of sampling. This paper extends Valliant's work to clustered sample designs.

In Section 1.2.1 on page 33, we introduced the GREG estimator and presented several common variance estimators for it. In the next section, we introduce the model-based framework. In the third section, we present our new research. We motivate and evaluate the sandwich variance estimator and several leverage adjustments to the sandwich variance estimator. In the fourth section, we show how the new variance estimators perform in several simulations. Lastly, we summarize our findings with a conclusion.

2.2 Literature Review

Concerned with some of the limitations of design-based estimation, some statisticians developed the prediction approach to estimate finite population parameters using models. In the prediction framework, estimation and inference are taken with respect to a working model, rather than the sample design. Individual characteristics of each unit in the population are considered random variables which can be modeled. Commonly, linear, generalized linear, and logistic models are built to predict the characteristics of non-sample units. The basic prediction estimator then combines the observed sample responses with the predicted responses for the non-sample units to form finite population estimates.

While the design-based framework is primarily nonparametric, classic statistical methods often rely on distribution or model assumptions. Although criticisms of the design-based theory emerged in the last quarter of the 20th century, the model-based framework for finite population estimation had been studied for decades prior to that. Model-based estimators borrow strength from prior or auxiliary information about the population to improve the efficiency of estimators. The model-based theory can add insights into the sample and population in ways that the design-based framework cannot. Indeed, the model-based theory offers much to the design-based analysis; however, if underlying parametric assumptions are violated, the model-based framework can produce misleading results.

2.2.1 Introduction

In the latter half of the 20th century, some statisticians began to explore the limits of the design-based framework. Their explorations led to critiques which spurred a new approach to estimating finite population parameters: the prediction paradigm. This new approach differs from the design-based approach by deemphasizing probability sampling, emphasizing balance, and relying on models.

With these criticisms in mind, statisticians drew upon a long history of classical statistics to form the prediction approach to estimating finite population parameters. Models are at the heart of the prediction framework.

The prediction approach combines sample survey data with predicted values to estimate finite population quantities. Driven by higher than expected nonresponse in small areas, Hansen et al. (1953a, p. 483 - 486) describe one of the first efforts to combine survey estimates with model predictions to estimate radio listening in 500 county areas. Hansen et al. (1953a) were concerned that significant nonresponse had threatened the integrity of their probability sample and opted to use models to help correct for nonresponse errors.

The classical model-based theory is much older than the design-based approach. Several hundred years old, pioneers such as Gauss, Bernoulli, Poisson, and Lexis all helped build classic statistical theory. It differs from the design-based framework by assuming that the population size is infinite and the characteristics of the units in the population are random rather than fixed. Thus, model-based statisticians treat the measured response, y_k, as an instance of a random variable Y_k. If the domain of Y_k is discrete, then

Hoel et al. (1971) denote the discrete density function of Y_k as

$$P\left(Y_k = y_k\right) = f_{Y_k}\left(y_k\right)$$

Casella and Berger (2002); Hoel et al. (1971); Hogg and Craig (1995); Shao (2003) all provide textbook definitions of discrete and continuous density functions as well as other fundamentals to probability theory and statistics. If the domain of Y_k is continuous, then the model-based expected value of a function, g of Y_k is defined as,

$$E_M g\left(Y_k\right) = \int_{-\infty}^{\infty} g\left(y_k\right) f_{Y_k}\left(y_k\right) dy_k$$

and model-based variance is defined as,

$$\mathrm{var}_M\left(Y_k\right) = E_M\left[Y_k - E_M\left(Y_k\right)\right]^2$$

If we conduct a survey and measure a key response variable for n units in our sample, we can put our response variables into a vector, denoted by $\mathbf{y}_{\mathfrak{s}}$. In the model-based framework, this response vector is one instance of the random vector $\mathbf{Y}_{\mathfrak{s}}$. Royall (1970) argued that the population characteristics can be thought of as random variables rather than fixed constants and that there is no loss of objectivity by thinking of the population characteristics as being random. Moreover, he argued that a characteristic being fixed at the time of the survey did not preclude the characteristic from being generated by a probability mechanism.

2.2.2 Linear Models

2.2.2.1 Parametrization

Linear models are commonly used to describe the relationship between multiple covariates and a continuous response variable. For example, the response from the kth unit can be modeled by

$$Y_k = \mathbf{x}_k^\top \boldsymbol{\beta} + \varepsilon_k \tag{2.1}$$

where Y_k is a random response variable and $\mathbf{x}_k$ is a non-random column vector of p auxiliary variables. Moreover, $\boldsymbol{\beta}$ is a p-valued column vector of model parameters called coefficients,

$$\underset{p\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix},$$

and $\boldsymbol{\varepsilon}_k$ is a random error term. If we additionally assume that all of the units in our sample, or population, can be described by the same model, then we can write the vector of responses in terms of the model,

$$\mathbf{Y} = \mathbf{X}^\top \boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.2}$$

where $\mathbf{Y}$ is a random vector of responses and $\mathbf{x}$ is a non-random full rank n by p matrix of auxiliary variables.

If we associate each sample element with a different axis, then $\mathbf{Y}$ is a random vector in n-dimensional space. Moreover, each column in $\mathbf{X}$ defines a fixed vector in n-dimensional space. If we let $\mathbf{b}$ be an arbitrary estimate of $\boldsymbol{\beta}$, then the set of linear

combinations $\mathbf{X}^\top \mathbf{b}$ determine the estimation space. Moreover, the random error vector $\boldsymbol{\varepsilon}$, is orthogonal to the $\mathbf{X}^\top \mathbf{b}$ space and intersects $\mathbf{Y}$.

It is beneficial to restrict $\mathbf{b}$ so that,

- $E_M(\boldsymbol{\varepsilon}) = \mathbf{0}$
- $\mathrm{var}_M(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$
- $\sigma^2 < \infty$

With these assumptions, the Gauss-Markov theorem claims that the best choice of $\mathbf{b}$ is $\hat{\boldsymbol{\beta}}$, which has the lowest variance among the class of unbiased linear estimators and is defined by,

$$\widehat{\boldsymbol{\beta}}(\mathbf{Y}) = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Y} \tag{2.3}$$

Because $\widehat{\boldsymbol{\beta}}$ depends on a random quantity, $\mathbf{Y}$, it is also random. When $\mathbf{Y}$ is replaced with one realization, $\mathbf{y}$, we write this nonrandom quantity,

$$\widehat{\boldsymbol{\beta}}(\mathbf{y}) = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}$$

Moreover, $\mathbf{X}^\top \widehat{\boldsymbol{\beta}}$ is a vector in the $\mathbf{X}^\top \mathbf{b}$ space that is of minimal distance from $\mathbf{y}$ and consequently also results in the minimal estimated error. That is, the length $\|\mathbf{Y} - \mathbf{X}^\top \widehat{\boldsymbol{\beta}}\|$ is minimized for any realization of $\mathbf{Y}$.

The covariance matrix of $\widehat{\boldsymbol{\beta}}$ is,

$$\mathrm{var}_M\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \left[\mathrm{var}_M(\mathbf{Y})\right] \mathbf{X}^\top \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$$

which can be written as

$$\mathrm{var}_M\left(\widehat{\boldsymbol{\beta}}\right) = \mathbf{a}^\top \boldsymbol{\Psi} \mathbf{a} \tag{2.4}$$

with

$$\underset{n\times n}{\mathbf{\Psi}} = \mathrm{var}_M\left(\mathbf{Y}\right)$$

$$\underset{n\times p}{\mathbf{a}} = \mathbf{X}^\top\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}.$$

Moreover, the variance of the components of $\widehat{\boldsymbol{\beta}}$ lies along the diagonal of $\mathrm{var}_M\left(\hat{\boldsymbol{\beta}}\right)$.

We notice that $\mathbf{a}^\top\mathbf{\Psi}\mathbf{a}$ looks like a sandwich with $\mathbf{a}$ being the bread and $\mathbf{\Psi}$ being the meat. Thus, $\mathbf{a}^\top\mathbf{\Psi}\mathbf{a}$ is called a *sandwich estimator*.

If the errors are homoscedastic; that is, if $\mathrm{var}_M\left(\mathbf{Y}\right) = \sigma^2\mathbf{I}$, then the variance reduces to,

$$\mathrm{var}_M\left(\widehat{\boldsymbol{\beta}}\right) = \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}.$$

If σ^2 is not known *a priori*, then it can be estimated by

$$\begin{aligned} s^2 &= \frac{1}{n-p}\left(\mathbf{y}-\mathbf{X}^\top\widehat{\boldsymbol{\beta}}\left(\mathbf{y}\right)\right)^\top\left(\mathbf{y}-\mathbf{X}^\top\widehat{\boldsymbol{\beta}}\left(\mathbf{y}\right)\right) \\ &= \frac{1}{n-p}\sum_{k\in\mathfrak{s}}\mathbf{r}_k\mathbf{r}_k^\top \end{aligned} \tag{2.5}$$

where $\mathbf{r}_k$ is a residual defined on in Equation (2.8) on page 67.

As seen in Rencher (2000, p. 135)

$$\upsilon_M\left(\widehat{\boldsymbol{\beta}}\right) = s^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$$

is an unbiased estimator of $\mathrm{var}_M\left(\widehat{\boldsymbol{\beta}}\right)$ when the errors are homoscedastic.

2.2.2.2 The Hat Matrix and Leverages

After estimating $\boldsymbol{\beta}$, one can predict $E_M\left(\mathbf{Y}\right)$ for all units in the population using the set of data about the full population, $\mathbf{X}_{\mathscr{U}}$, where the $\mathscr{U}$ subscript indicates that $\mathbf{X}$ contains

auxiliary data for all elements in the population. The values $\widehat{\mathbf{Y}}_{\mathscr{U}} = \mathbf{X}_{\mathscr{U}}^{\top}\hat{\boldsymbol{\beta}}$ are called the fitted or predicted values. The fitted values play an important role in residuals, population parameter estimation, and model error estimation. Using elementary linear algebra, one can relate the observed response variable to the fitted values by,

$$\begin{aligned}\widehat{\mathbf{Y}} &= \mathbf{X}^{\top}\left(\mathbf{X}^{\top}\mathbf{X}\right)^{-1}\mathbf{X}^{\top}\mathbf{Y} \\ &= \mathbf{H}\mathbf{Y}\end{aligned} \tag{2.6}$$

where $\mathbf{H}$ has dimension n by n. In Equation (2.6) we omit the population and sample subscripts because the equation hold for both the population and sample. Hoaglin and Welsch (1978) claim that John Tukey first called $\mathbf{H}$ the hat matrix because it puts the hat on $\mathbf{Y}$. Geometrically, we can interpret $\mathbf{H}$ as the matrix that projects $\mathbf{Y}$ onto the $\mathbf{X}^{\top}\mathbf{b}$ space. Since $\widehat{\mathbf{Y}} = \mathbf{X}^{\top}\widehat{\boldsymbol{\beta}}$, it follows that $\widehat{\mathbf{Y}}$ must lie in the $\mathbf{X}^{\top}\mathbf{b}$ space.

The hat matrix has several important uses. First, it plays an important part in the expected value of variance estimators. Secondly, the diagonal elements of $\mathbf{H}$ are called *leverages* and denoted h_{kk}. They illuminate the effect that $\mathbf{Y}$ has on $\widehat{\mathbf{Y}}$ by measuring how far an observation's covariates, $\mathbf{X}_k$, are from the expected value of all covariates, $\bar{\mathbf{X}}$. Rencher (2000, p. 218) writes the leverages as

$$h_{kk} = \frac{1}{n} + \left(\mathbf{x}_{1k} - \bar{\mathbf{x}}_1\right)^{\top}\left(\mathbf{X}_c^{\top}\mathbf{X}_c\right)^{-1}\left(\mathbf{x}_{1k} - \bar{\mathbf{x}}_1\right) \tag{2.7}$$

where

$$\mathbf{X}_c = \left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{X}_1$$

and $\mathbf{X}_1$ is $\mathbf{X}$ without the first intercept column, $\mathbf{J}$ is an n by n matrix of 1s, and $\mathbf{x}_{1k} - \bar{\mathbf{x}}_1$ is the kth row of $\mathbf{X}_c$. When the leverages are written in the form of Equation (2.7),

we see that the leverage is a standardized or Mahalanobis distance because $\mathbf{X}_c^\top \mathbf{X}_c$ is proportional to the sample covariance matrix. Thus, the leverage provides a measure of the relative distance between $\mathbf{x}_k$ and $\bar{\mathbf{x}}$. The leverages are bounded by 0 and 1 and sum to p. According to Kutner et al. (2005, p. 399), leverages do not indicate if an observation is an outlier or influential, but a common rule of thumb is to investigate units where the leverage is greater than $2\frac{p}{n}$ which is twice the average value of the leverages. If h_{kk} is relatively large, then the kth element plays a large role in the model estimation. In addition to their role in outlier detection, leverages can be used to form robust estimators. As we will see, using leverages to weight residuals is one way to standardize the residuals and prevent outliers from dominating variance estimates.

2.2.2.3 Residuals

In the model-based theory, the ability of the model to predict a characteristic is often measured by the residual. A large residual indicates that the predicted value from the model is far from the observed value. Belsley et al. (1980) discuss how residuals can be used to detect outliers and influential observations; thereby, diagnosing problems in the working model specification and fit. Residuals also play an important role in estimating the model error, that is the variability between the model predictions and the observed population values. The residual is commonly formed as

$$\begin{aligned}\mathbf{R} &= \mathbf{Y} - \widehat{\mathbf{Y}} \\ &= \mathbf{Y} - \mathbf{X}^\top \widehat{\boldsymbol{\beta}}.\end{aligned}$$

Often the statistician differentiates the true error, denoted $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X}^\top \boldsymbol{\beta}$, from the residual, $\mathbf{R}$. Furthermore, the residual for a specific sample is

$$\begin{aligned} \mathbf{r} &= \mathbf{y} - \widehat{\mathbf{y}} \\ &= \mathbf{y} - \mathbf{X}^\top \widehat{\boldsymbol{\beta}} \end{aligned} \tag{2.8}$$

The true error variance plays an important role in model parameter estimation, diagnostics, inference, and finite population estimation. When the true error variance is unknown, one often relies on the residuals to estimate σ^2. It is assumed that

$$\mathrm{var}_M(\varepsilon) = \sigma^2 \mathbf{I}$$

Although one may be tempted to estimate σ^2 with $\frac{1}{n}\sum_{k \in \mathfrak{s}} r_k^2$, Theil (1971) showed that

$$\mathrm{var}_M(R_k) = \sigma^2 (1 - h_{kk}) \tag{2.9}$$

for all k. Since $0 \leq h_{kk} \leq 1$, we see that $\mathrm{var}_M(R_k)$ underestimates σ^2, the true error variance. As Kutner et al. (2005, p. 399) explain, units with large leverages have small residual variance, $\mathrm{var}_M(R_k)$, and units with small leverages will have large residual variance. "In the extreme case where $h_{kk} = 1$, the variance $\mathrm{var}_M(R_k)$ equals 0, so the fitted value $\widehat{Y}_k$ is forced to equal the observed value Y_k." In such an extreme case, it is clear that the kth unit dominated the model fitting process. If we solve Equation (2.9) for σ^2, we see that estimating σ^2 with R_k^2 will underestimate σ^2, especially when the leverages are large.

An additional problem with using residuals in model diagnostics, is that each R_k has a different variance because the variance of R_k depends on the leverages, h_{kk}. Thus, the variance of residuals of high-leverage observations will tend to be smaller than the

variance of residuals of units with smaller leverage. The lack of a common variance for all residuals makes it difficult to compare the residuals and find outliers. Fortunately, we can standardize the residuals by dividing them by $1 - h_{kk}$. Because the standardized residuals have common mean and variance, they can be used for testing for outliers and other diagnostic analysis. The *standardized residuals* with mean 0 and variance 1, denoted r_k^{std}, are

$$r_k^{std} = \frac{r_k}{\sigma\sqrt{1 - h_{kk}}}. \tag{2.10}$$

Since σ is rarely known, the estimated model error, s, is usually used instead of σ. These residuals are called the *studentized residuals*,

$$r_k^{stu} = \frac{r_k}{s\sqrt{1 - h_{kk}}}. \tag{2.11}$$

Rather than using s, which is based on the full sample, Belsley et al. (1980, p. 20) recommend using

$$r_k^{sd} = \frac{r_k}{s_{(k)}\sqrt{1 - h_{kk}}} \tag{2.12}$$

where $s_{(k)}$ is the estimated model error obtained from the full sample minus the kth unit. This residual is called the *studentized deleted residual*. Belsley et al. (1980) advocate the studentized deleted residuals because they have common variance and can easily be related to the t-distribution, thereby facilitating hypothesis testing of residuals. Moreover, if the kth unit is an outlier, then r_k^{sd} is likely to detect it because $s_{(k)}$ is not contaminated with the extremity of the kth unit.

2.2.3 Prediction Estimators

Valliant et al. (2000) provide an excellent introduction to prediction theory. This approach asserts that the characteristics observed in the population are generated by an underlying model, which is unknown to the statistician. Modelers seek to develop parsimonious models that closely resemble the underlying population distribution. These working models can be used to estimate a finite population total,

$$\widehat{T}_y^{pre} = \sum_{\mathfrak{s}} Y_k + \sum_{r} \widehat{Y}_k \tag{2.13}$$

where

$$\widehat{Y}_k = \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}\left(\mathbf{Y}\right).$$

Notice that $\widehat{T}_y^{pre}$ is a random variable. When we have a specific realization of $\mathbf{Y}_{\mathfrak{s}}$ our estimate is

$$\widehat{t}_y^{pre} = \sum_{\mathfrak{s}} y_k + \sum_{r} \widehat{y}_k$$

where

$$\widehat{y}_k = \mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}\left(\mathbf{y}\right).$$

There are several choices for estimating $\boldsymbol{\beta}$; however, Valliant et al. (2000, p. 29) provide a theorem for estimating $\boldsymbol{\beta}$ in such a way that $\hat{T}_y^{pre}$ is the the best linear unbiased predictor of T_y. If $\mathrm{var}_M\left(\mathbf{Y}\right) = \mathbf{Q}^{-1}$, then the unbiased linear estimator of the total with the minimum variance is obtained when

$$\widehat{\boldsymbol{\beta}}\left(\mathbf{Y}\right) = \left(\mathbf{X}^\top \mathbf{Q} \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{Q} \mathbf{Y}. \tag{2.14}$$

Under these assumptions, we see that we can also write our prediction estimator as

$$\begin{aligned}\widehat{T}_y^{pre} &= \mathbf{a}_{\mathfrak{s}}^\top \mathbf{Y}_{\mathfrak{s}} \\ &= \sum_{\mathfrak{s}} a_k Y_k \end{aligned} \tag{2.15}$$

where

$$\mathbf{a} = \mathbf{Q}_{\mathfrak{s}\mathfrak{s}} \left[\mathbf{Q}_{\mathfrak{s}r}^{-1} + \mathbf{X}_{\mathfrak{s}}\mathbf{A}_{\mathfrak{s}}^{-1}\left(\mathbf{X}_r^\top - \mathbf{X}_{\mathfrak{s}}^\top \mathbf{Q}_{\mathfrak{s}\mathfrak{s}}\mathbf{Q}_{\mathfrak{s}r}^{-1}\right)\right]\mathbf{1}_r + \mathbf{1}_{\mathfrak{s}} \tag{2.16}$$

$$\mathbf{Q}_{\mathscr{U}} = \begin{bmatrix} \mathbf{Q}_{\mathfrak{s}\mathfrak{s}} & \mathbf{Q}_{\mathfrak{s}r} \\ \mathbf{Q}_{r\mathfrak{s}} & \mathbf{Q}_{rr} \end{bmatrix} \tag{2.17}$$

and

$$\mathbf{A}_{\mathfrak{s}} = \mathbf{X}_{\mathfrak{s}}^\top \mathbf{Q}_{\mathfrak{s}\mathfrak{s}}\mathbf{X}_{\mathfrak{s}} \tag{2.18}$$

where r designates nonsample units. If we further assume that $\mathbf{Q}^{-1} = \sigma^2\mathbf{I}$, then

$$\mathbf{a} = \mathbf{X}_{\mathfrak{s}}\left(\mathbf{X}_{\mathfrak{s}}^\top\mathbf{X}_{\mathfrak{s}}\right)^{-1}\mathbf{X}_r^\top\mathbf{1}_r + \mathbf{1}_{\mathfrak{s}}.$$

If an estimator can be written in the form of Equation (2.15) and $\mathrm{var}_M(\mathbf{Y})$ is diagonal, then we can write the true variance of our estimator as

$$\mathrm{var}_M\left(\widehat{T}_y^{pre}\right) = \mathbf{a}^\top\mathbf{\Psi}\mathbf{a} = \sum_{\mathfrak{s}} a_k^2\psi_k$$

where $\mathbf{\Psi} = \mathrm{var}_M(\mathbf{Y})$ and $\psi_k = \mathrm{diag}(\mathbf{\Psi})$

Typically, homoscedastic errors are assumed so that $\psi_k = \sigma^2$ for all units. Moreover, σ^2 is estimated by s^2 as defined in Equation (2.5) on page 64.

Linear models can fail in the specification in the linear component, in the specification of the error component, or in the specification of both the linear and error components. The possibility of biased variance estimates is not trivial, given that the modeler

never knows the underlying population model and that it can be quite difficult to evaluate the fit of the variance component from some samples. As a measure of precaution against model failure, Royall and Herson (1973) argued that variance estimators should be robust. A *robust* variance estimator is a variance estimator that is model-unbiased under the working model and approximately unbiased when the variance component of the working model is misspecified. In prediction theory, estimation is made with respect to the working model. Of course, the modeler never knows the underlying population model. Thus, it is important to create estimators that are robust to model misspecifications. Valliant et al. (2000, chapters 5 and 9) and Royall and Cumberland (1981) describe a general strategy for constructing robust variance estimators in unclustered and clustered populations. Robust variance estimation protects against misspecification in the variance component of the working model. Thus, if one uses a robust variance estimator, there is less risk of a unreasonable variance estimate due to a misspecified variance component in the working model.

Concerned with estimating the variance of model parameters when errors are heteroscadastic, Eicker (1963, 1967), Huber (1967), and Hinkley (1977) developed the sandwich estimator and discussed its asymptotic properties. Situated within the model-based and asymptotic design-based frameworks, sandwich variance estimators seek to provide accurate estimates of standard errors even when the variance component of the working model is misspecified. Sandwich estimators are model-unbiased when the variance component of the working model is correctly specified and approximately model-unbiased otherwise. Traditional model-based variance estimators, such as those that assume homoscedastic errors, run the risk of being seriously biased when the model assumptions

are violated. If the errors are actually heteroscadastic, inferences can be quite poor. An alternative approach is to use a sandwich estimator. In this approach we replace r_k in Equation (2.5) with r_k^{std}, r_k^{stu}, or r_k^{sd} to approximate ψ_k.

If one does not want to assume homoscedastic errors, ψ_k can be estimated with the residual, r_k^2. This estimator is approximately unbiased for ψ_k, regardless of the specified variance of the working model. The statistical foundation for the sandwich estimator relies on the fact that $E_M\left(R_k^2\right) \approx \psi_k$ in large samples. White (1980) showed that R_k^2 was a consistent estimator of ψ_k. Although R_k^2 is a consistent estimator; for small and moderate sized samples, it underestimates ψ_k due to the leverages in Equation (2.9). In fact, units with larger leverages will underestimate ψ_k more than units with smaller leverages.

Seeking to correct for this fact Hinkley (1977), inflated r_k^2 with the factor $\frac{n}{n-p}$. Additionally, Horn et al. (1975) proposed using $\frac{r_k^2}{1-h_{kk}}$ to estimate ψ_k. Finally, Efron (1982) and MacKinnon and White (1985) suggest using $\frac{r_k^2}{(1-h_{kk})^2}$ to estimate ψ_k. Interestingly, this estimator is asymptotically equivalent to the jackknife variance estimator. Thus, leverage adjustments to the sandwich estimator can be used to approximate the jackknife, without taking up all of the computer resources needed for replication.

Using the adjusted residuals to improve variance estimates of prediction estimators, Valliant et al. (2000, p. 145) suggest estimating ψ_k in one of the following ways,

$$\begin{aligned}
\hat{\psi}_k^R &= r_k^2 \\
\hat{\psi}_k^D &= \frac{r_k^2}{1-h_{kk}} \\
\hat{\psi}_k^{J^*} &= \frac{r_k^2}{(1-h_{kk})^2}.
\end{aligned}$$

Naturally, h_{kk} will vary depending on the estimator and working model. Valliant et al.

(2000) show that replacing ψ_k with one of these estimators will provide an approximately unbiased estimate of the true error even if the variance component in the model is misspecified. Thus, the following estimators are considered robust estimators in single-stage samples

$$\upsilon_R = \sum_{\mathfrak{s}} a_k^2 r_k^2 \tag{2.19}$$

$$\upsilon_D = \sum_{\mathfrak{s}} a_k^2 \frac{r_k^2}{1 - h_{kk}} \tag{2.20}$$

$$\upsilon_{J^*} = \sum_{\mathfrak{s}} a_k^2 \frac{r_k^2}{(1 - h_{kk})^2}. \tag{2.21}$$

Similar robust estimators have been proposed, most of which can be thought of as adjustments to υ_{R_I} using leverages.

Valliant et al. (2000) dedicate one chapter to variance estimation in clustered samples. They focus on constructing robust model-based variance estimators for a variety of linear models in cluster samples. Section 9.5 deals with the regression estimator

$$\widehat{T}_y^{pre} = \mathbf{1}_{\mathfrak{s}}^\top \mathbf{Y}_{\mathfrak{s}} + \mathbf{1}_r^\top \mathbf{X}_r \widehat{\boldsymbol{\beta}}.$$

In their book, Valliant et al. (2000, chapter 5) use the method of adjusting variance estimators by the leverages to construct robust variance estimators of totals under a variety of working models. In chapter 9, Valliant et al. (2000) extend robust estimation to clustered samples. Specifically, they propose

$$\upsilon_R = \sum_{s_I} \left(\mathbf{g}_i^\top \mathbf{r}_i\right)^2$$

where $\mathbf{r}_i$ is a vector of residuals for the ith cluster and $\mathbf{g}_i$ is a vector of weights for the ith cluster. To avoid confusion, we note that these weights are not design weights based

on probabilities of selection, but rather model weights similar to those found in Equation (2.28). Using leverages, Valliant et al. (2000, p. 314) make an internal adjustment to υ_R to get

$$\upsilon_D = \sum_{\mathfrak{s}} \mathbf{g}_i^\top \mathbf{P}_i^{-1} \left(\mathbf{r}_i \mathbf{r}_i^\top\right) \mathbf{g}_i$$

where $\mathbf{P}_i = \mathbf{I}_{m_i} - \mathbf{H}_{ii}$ and

$$\underset{m\times m}{\mathbf{H}_{\mathfrak{s}}} = \mathbf{X}_{\mathfrak{s}} \mathbf{A}_{\mathfrak{s}}^{-1} \mathbf{X}_{\mathfrak{s}}^\top \mathbf{Q}_{\mathfrak{s}} \tag{2.22}$$

$$= \begin{bmatrix} \mathbf{X}_{\mathfrak{s}1} \mathbf{A}_{\mathfrak{s}}^{-1} \mathbf{X}_{\mathfrak{s}1}^\top \mathbf{Q}_{\mathfrak{s}1} & \dots & \mathbf{X}_{\mathfrak{s}1} \mathbf{A}_{\mathfrak{s}}^{-1} \mathbf{X}_{\mathfrak{s}n}^\top \mathbf{Q}_{\mathfrak{s}n} \\ \vdots & \ddots & \vdots \\ \mathbf{X}_{\mathfrak{s}n} \mathbf{A}_{\mathfrak{s}}^{-1} \mathbf{X}_{\mathfrak{s}1}^\top \mathbf{Q}_{\mathfrak{s}1} & \dots & \mathbf{X}_{\mathfrak{s}n} \mathbf{A}_{\mathfrak{s}}^{-1} \mathbf{X}_{\mathfrak{s}n}^\top \mathbf{Q}_{\mathfrak{s}n} \end{bmatrix} \tag{2.23}$$

where

$$\mathbf{A}_{\mathfrak{s}} = \mathbf{X}_{\mathfrak{s}}^\top \mathbf{Q}_{\mathfrak{s}} \mathbf{X}_{\mathfrak{s}}. \tag{2.24}$$

In Equation (2.22), we use the $\mathfrak{s}$ subscript to differentiate our sample matrices from population matrices. For example $\mathbf{X}_{\mathfrak{s}}$ is the matrix of auxiliary variables for all sample elements; while $\mathbf{X}_{\mathscr{U}}$ is the matrix of auxiliary variables for the full population. Usually $\mathbf{X}$ is the sample matrix of auxiliary variables, but sometimes we use $\mathbf{X}_{\mathfrak{s}}$ to emphasize that $\mathbf{X}$ is a sample quantity. Furthermore, in Equation (2.23) we write the matrix of covariates for sample cluster i as $\mathbf{X}_{\mathfrak{s}i}$ and the diagonal matrix of unit variances for cluster i as $\mathbf{Q}_{\mathfrak{s}i}^{-1}$. The block diagonal elements of $\mathbf{H}_{\mathfrak{s}}$ can be written as,

$$\underset{m_i\times m_i}{\mathbf{H}_{ii'}} = \mathbf{X}_{\mathfrak{s}i} \mathbf{A}_{\mathfrak{s}}^{-1} \mathbf{X}_{\mathfrak{s}i'}^\top \mathbf{Q}_{\mathfrak{s}i'}.$$

Valliant et al. (2000, p. 314) also propose two additional robust variance estimators of $\widehat{T}_y^{pre}$ in a clustered population

$$\upsilon_{J^*} = \sum_{i \in \mathfrak{s}_I} \mathbf{g}_i^\top \mathbf{P}_i^{-1} \left(\mathbf{r}_i \mathbf{r}_i^\top \right) \mathbf{P}_i^{-1} \mathbf{g}_i \tag{2.25}$$

$$\upsilon_J = \frac{n-1}{n} \left\{ \sum_{\mathfrak{s}_I} \left(\mathbf{a}_i^\top \mathbf{P}_i^{-1} \mathbf{r}_i \right)^2 - n^{-1} \left[\sum_{\mathfrak{s}_I} \mathbf{a}_i^\top \mathbf{P}_i^{-1} \mathbf{r}_i \right]^2 \right\} \tag{2.26}$$

where

$$\mathbf{a}_i^\top = \mathbf{1}_r^\top \mathbf{X}_r \mathbf{G} \mathbf{X}_{si}^\top \mathbf{W}_{si} \tag{2.27}$$

$$\mathbf{g} = \mathbf{W}_{\mathfrak{s}} \mathbf{X}_{\mathfrak{s}} \mathbf{G} \mathbf{X}_r^\top \mathbf{1}_r + \mathbf{1}_r \tag{2.28}$$

$$\mathbf{H}_{ii} = \mathbf{X}_{si} \mathbf{G} \mathbf{X}_{si}^\top \mathbf{W}_{si}. \tag{2.29}$$

Moreover, $\mathbf{G}$ is a solution to $\boldsymbol{\beta} = \mathbf{G}\mathbf{X}_{\mathfrak{s}}^\top \mathbf{W}_{\mathfrak{s}} \mathbf{Y}_{\mathfrak{s}}$ and $\mathbf{W}_{\mathfrak{s}}$ is a block diagonal working model covariance matrix. The vectors or matrices with the i subscript indicate the subset of the larger vector or matrix that is in the ith cluster.

Clearly, sandwich estimators will approximate ψ_k when the errors are homoscedastic or heteroscadastic because there is no explicit assumption about the distribution of the errors. This illustrates the concept of robust variance estimation. That is, even if the working model variances are misspecified, the sandwich estimator will still give an accurate estimate of the true parameter variance under expectation.

On the other hand, estimating ψ_k with r_k^2 may be less stable than $s^2 = \frac{1}{n-1} \sum r_k^2$ which gains stability by averaging over all of the unit variances. It is well known that sandwich estimators have larger mean squared error than other model-based estimators when the model is correctly specified because leverages add considerable variability to the

variance estimator (Carroll et al. 1998). Thus, when the working model is correctly specified, the sandwich estimator is not as efficient as standard parametric variance estimators. On the other hand, from a design-based framework, sandwich estimators are attractive because they can be more flexible and accurate than current estimators, especially in small or moderate sized samples. Moreover, the sandwich estimators are less computationally demanding than replication methods and can be constructed to give asymptotically similar estimates to the jackknife.

As we have seen, leverage adjusted sandwich estimators can be constructed to make estimation robust against misspecified error models. These adjustments can even be applied to variance estimators of finite population parameters in clustered populations. All of the estimators discussed are model-consistent for the true model-variance and many are approximately model-unbiased.

2.2.4 Discussion

In an early defense of modeling, Brewer argued that the prediction approach could produce accurate and precise estimates irrespective of the sample design. In fact, Brewer (1963, p. 98) proved that purposeful selection of samples could greatly reduce the mean squared error of estimators. For example, for populations generated by a linear model, the sample with the smallest mean squared error will be the "partial collection" with the maximum x_k values. Although Brewer found some significant advantages to purposeful sample selection, he also acknowledged the importance of building good models.

In social science problems, models are never known and must be posited. As

George Box once said, “all models are wrong, but some are useful” (Box and Draper 1987). Because of this element of subjectivity in modeling, Neyman (1934) advocated using the design-based framework over the model-based framework. Moreover, Hansen et al. (1983) remarked that

> if the assumed model does not accurately represent the state of nature, estimates of population parameters may be substantially biased, and statements about the sampling errors of those estimates may be very misleading. In attempts to avoid this problem, one may possibly relax the model, for example by including additional model parameters. However, even the relaxed model still may not represent the state of nature well enough to prevent misleading inferences.

Thus, Hansen et al. (1983) argued that model-based estimates were always sensitive to the model specification.

Hansen et al. (1983) were particularly concerned when the model fails to accurately describe the population. In such cases inference can be severely misleading. They also contend that model-dependent methods tend to underestimate the sampling variance of their estimators under repeated sampling. Thus, they argued, “model-dependent designs, including those that use ‘robust’ procedures, face the risk of substantially understating the mean squared error, even when the model appears to be satisfactory.” Hansen et al. (1983) further suggested that the robust modeling techniques of Royall and Herson (1973) made so few assumptions that they were nearly equivalent to design-based analysis. On this topic Hansen et al. (1983) stated, “the problems of model failure will remain unless

the designs are so robust as to be nearly model-independent, in which event they are essentially equivalent to probability-sampling designs."

In response to the threats of model misspecification, Royall and Herson (1973) introduced the notion of balanced samples. A balanced sample is a sample in which the distance from the center of the population auxiliary variables is close to the center of the sample auxiliary variables. That is, a sample is balanced if $\bar{\mathbf{x}}_{\mathfrak{s}} = \bar{\mathbf{x}}_{\mathscr{U}}$. The definition of balance is broad enough to extend to other moments of $\mathbf{x}$. As Cumberland and Royall (1988, p. 118) simply stated, "A sample is well balanced on an auxiliary variable x if the sample x-moments closely match the population x-moments."

Balanced samples are desirable because they support bias-robust estimation. A bias-robust estimate is one in which the estimate is unbiased even if the working model is misspecified. Specifically, balance, as discussed in Valliant et al. (2000), protects against leaving out higher order terms in the model. It does not necessarily protect against hidden regressors. In prediction theory, the estimator one uses and the type of balance one uses when sampling are closely tied together. Valliant et al. (2000) discusses bias robust estimation in great detail.

The prediction approach can improve estimates from off-balance samples by drawing strength from models. That is, if modeling is done correctly, point estimates from the prediction approach can be closer to the true population parameter even when the sample is off-balance. Moreover standard error estimates can be less biased and more stable than design-based estimates from off-balance samples. In general, the prediction approach offers protections against off-balanced samples by accounting for the configuration of the covariates in the sample.

The debate between the design-based and model-based frameworks continued throughout the 1980's and 1990's. In 2000, Valliant et al. (2000) carefully laid out the prediction framework. Valliant et al. (2000) contended that Hansen et al. (1983) misunderstood the model-based methodology when they empirically compared the design and model-based frameworks. Valliant et al. (2000) repeated the simulation in Hansen et al. (1983) and showed that the model-based framework provided better estimates under repeated sampling, even when the model failed.

Prediction theory offers much to survey statistics. First, there is a long and rich history of research from the model-based approach that can be used to improve design-based analysis. Second, models can be used to improve sample selection and reduce sampling errors. As Royall (1970) noted, purposeful samples can be designed that greatly reduce sampling error. Following up on this, Valliant et al. (2000) discussed how models could be used in the design-based paradigm to select balanced samples with smaller mean squared error than completely non-informed sampling. Lastly, the dependency on models can be relaxed through various robust estimation techniques.

In the next section, we use model-based sandwich variance estimators to improve variance estimators of the GREG estimator.

2.3 Theoretical Results

To motivate our new variance estimators, we take a model-based approach. As a working model, we assume that our response data are a linear combination of our auxiliary

data. That is,

$$\mathrm{E}_M\left(Y_k\right) = \mathbf{x}_k^\top \boldsymbol{\beta}$$

We further assume that elements are independent and uncorrelated between clusters, but dependent and correlated within clusters. Letting k and l denote elements within clusters and i and j, the covariance of two elements is

$$\mathrm{cov}_M\left(Y_k, Y_l\right) = \begin{cases} 0 & \forall\, i \neq j \\ \psi_{kl} & \forall\, i = j \end{cases}$$

Under these two assumptions, we derived the model-based variance of $\widehat{T}_y^{gr}$ in Appendix A.3 on page 272. The variance simplifies to,

$$\begin{aligned}
\mathrm{var}_M\left(\widehat{T}_y^{gr} - T_y\right) &= \mathrm{var}_M\left(\sum_{i=1}^{n} \mathbf{g}_i \boldsymbol{\Pi}_i^{-1} \mathbf{Y}_i - \sum_{k=1}^{N} Y_k\right) \\
&= \sum_{i=1}^{n} \mathbf{g}_i^\top \boldsymbol{\Pi}_i^{-1} \boldsymbol{\Psi}_i \boldsymbol{\Pi}_i^{-1} \mathbf{g}_i - 2\sum_{i=1}^{n} \mathbf{g}_i^\top \boldsymbol{\Pi}_i^{-1} \mathrm{var}_M\left(\mathbf{Y}_{\mathfrak{s}i}, \mathbf{Y}_{\mathscr{U}}\right) \mathbf{1} + \mathbf{1}^\top \boldsymbol{\Psi}_{\mathscr{U}} \mathbf{1} \\
&= L_1 - 2L_2 + L_3
\end{aligned}$$

where $\mathrm{var}_M\left(\mathbf{Y}_i\right) = \boldsymbol{\Psi}_i$, $\mathbf{g}_i$ denotes the set of all g_k weights in the ith cluster, $\mathbf{Y}_i$ denotes the set of all Y_k in the ith cluster, $\mathbf{Y}_{\mathscr{U}}$ denotes the unknown full population vector containing all values of Y_k, and $\mathrm{var}_M\left(\mathbf{Y}_{\mathscr{U}}\right) = \boldsymbol{\Psi}_{\mathscr{U}}$.

The model-based error variance of $\widehat{T}_y^{gr}$ requires knowledge of ψ_k for the full population. Without some strong assumptions that link the sample and nonsample covariance structures, ψ_k cannot be estimated from the sample. Fortunately, we show in Appendix A.3, that L_1 dominates the variance as the number of sample and population clusters increase. Specifically, as the number of population and sample clusters increase, we assume

Assumption 1. $\mathbf{g}_i \longrightarrow \mathbf{1}$.

Assumption 2. $\mathrm{var}_M\left(\widehat{\bar{t}}_i^{gr}\right)$ *is bounded where* $\widehat{\bar{t}}_i^{gr}$ *is the GREG estimate of the mean value of* y *for cluster* i.

Assumption 3. $N max \pi_i = O(n)$.

If these conditions are met, then $L_1 = O\left(\frac{N^2}{n}\right)$, $L_2 = O(N)$, and $L_3 = O(N)$. If $f = \frac{n}{N} \longrightarrow 0$ then L_1 will dominate because $L_1 = O\left(\frac{N}{f}\right)$. Thus,

$$\mathrm{av}_M\left(\widehat{T}_y^{gr} - T_y\right) = \sum_{i \in \mathfrak{s}_I} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{\Psi}_i \mathbf{\Pi}_i^{-1} \mathbf{g}_i \tag{2.30}$$

On the other hand, if the number of population clusters increases at the same rate as sample clusters, then L_1, L_2, and L_3 all contribute to the asymptotic variance.

Unless the true variance matrix of $\mathbf{Y}_{\mathfrak{s}}$ is known, $\mathbf{\Psi}_i$ must be estimated. One simple and common method to estimate $\mathbf{\Psi}_i$ is with residuals. In Appendix A.1.9 on page 268, we show that in large samples

$$\mathrm{var}(\mathbf{e}_i) \approx \mathbf{\Psi}_i$$

where $\mathbf{e}_i = \mathbf{Y}_i - \widehat{\mathbf{Y}}_i$. Substituting $\mathbf{e}_i\mathbf{e}_i^\top$ for $\mathbf{\Psi}_i$ in Equation (2.30) on page 81 yields the sandwich estimator

$$\upsilon_R = \sum_{i \in \mathfrak{s}_I} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{e}_i \mathbf{e}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{g}_i. \tag{2.31}$$

In Appendix A.4 on page 275, we show that υ_R is approximately unbiased for $\mathrm{av}_M\left(\widehat{T}_y^{gr} - T_y\right)$ in large samples. However, in small to moderate sized samples, υ_R will be biased because

$$\mathrm{E}_M\left(\mathbf{e}_i\mathbf{e}_i^\top\right) = \mathrm{var}_M(\mathbf{e}_i) = (\mathbf{I}_{n_i} - \mathbf{H}_{ii})\,\mathbf{\Psi}_i\,(\mathbf{I}_{n_i} - \mathbf{H}_{ii})^\top + \sum_{j \neq i} \mathbf{H}_{ij}\mathbf{\Psi}_j\mathbf{H}_{ij}^\top \tag{2.32}$$

where

$$\underset{m_i \times m_i}{\mathbf{H}_{ij}} = \mathbf{X}_i^\top \mathbf{A}_\pi^{-1} \mathbf{X}_j \mathbf{Q}_i \mathbf{\Pi}_i^{-1}. \tag{2.33}$$

In Appendix A.1.3, we show that $\mathbf{H} = O\left(n^{-1}\right)$, which further suggests that $\mathrm{var}_M\left(\mathbf{e}_i\right) \approx \mathbf{\Psi}_i$. $\mathbf{H}$ is known as the survey weighted hat matrix. Li and Valliant (2009); Valliant (2002) show that the survey weighted hat matrix is

$$\begin{aligned} \underset{m \times m}{\mathbf{H}} &= \mathbf{X} \mathbf{A}_\pi^{-1} \mathbf{X}^\top \mathbf{Q} \mathbf{\Pi}^{-1} \\ &= \begin{bmatrix} \mathbf{X}_1 \mathbf{A}_\pi^{-1} \mathbf{X}_1^\top \mathbf{Q}_1 \mathbf{\Pi}_1^{-1} & \dots & \mathbf{X}_1 \mathbf{A}_\pi^{-1} \mathbf{X}_n^\top \mathbf{Q}_n \mathbf{\Pi}_n^{-1} \\ \vdots & \ddots & \vdots \\ \mathbf{X}_n \mathbf{A}_\pi^{-1} \mathbf{X}_1^\top \mathbf{Q}_1 \mathbf{\Pi}_1^{-1} & \dots & \mathbf{X}_n \mathbf{A}_\pi^{-1} \mathbf{X}_n^\top \mathbf{Q}_n \mathbf{\Pi}_n^{-1} \end{bmatrix} \end{aligned} \tag{2.34}$$

where

$$\mathbf{A}_\pi = \mathbf{X}^\top \mathbf{\Pi}^{-1} \mathbf{Q} \mathbf{X}. \tag{2.35}$$

The diagonal elements of $\mathbf{H}$ are called the leverages and denoted, $h_{kk} = \frac{q_k \mathbf{x}_k^\top \mathbf{A}_\pi^{-1} \mathbf{x}_k}{\pi_k}$. The off diagonal components of the survey weighted hat matrix are $h_{kl} = \frac{q_l \mathbf{x}_k^\top \mathbf{A}_\pi^{-1} \mathbf{x}_l}{\pi_l}$. In single-stage samples Li and Valliant (2009) argue that the leverages can be large if $\frac{q_k}{\pi_k}$ is relatively large or if $\mathbf{x}_k$ is relatively far from $\bar{\mathbf{x}}$. We define the portion of the hat matrix associated with cluster i as

$$\underset{m_i \times m_i}{\mathbf{H}_{ii}} = \mathbf{X}_i \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{Q}_i \mathbf{\Pi}_i^{-1}.$$

And the off diagonal elements of $\mathbf{H}$ as,

$$\underset{m_i \times m_j}{\mathbf{H}_{ij}} = \mathbf{X}_i \mathbf{A}_\pi^{-1} \mathbf{X}_j^\top \mathbf{Q}_j \mathbf{\Pi}_j^{-1}.$$

To adjust for the fact that $\mathbf{e}_i\mathbf{e}_i^\top$ in a biased estimator of $\mathbf{\Psi}_i$ in small to moderate samples, we make leverage adjustments to $\mathbf{e}_i\mathbf{e}_i^\top$. The basic sandwich estimator can be improved with leverage adjustments.

If $\mathbf{\Pi}_i^{-1}\mathbf{Q}_i = c\mathbf{\Psi}_i^{-1}$ for some constant c, then

$$\mathrm{var}_M\left(\mathbf{e}_i\right) = \left(\mathbf{I} - \mathbf{H}_{ii}\right)\mathbf{\Psi}_i.$$

Solving for $\mathbf{\Psi}_i$ and substituting into (2.31) gives,

$$\upsilon_D = \sum_{i\in\mathfrak{s}_I} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i\mathbf{e}_i^\top\mathbf{\Pi}_i^{-1}\mathbf{g}_i. \tag{2.36}$$

Since, all elements of $\mathbf{H} = O\left(n^{-1}\right)$, we see that $\upsilon_D \approx \mathrm{av}_M\left(\widehat{T}_y^{gr} - T_y\right)$. One undesirable feature of υ_D is that it can be negative. This is a result of some clusters having negative estimates of υ_{Di} where $\upsilon_{Di} = \mathbf{g}_i^\top\mathbf{\Pi}_i^{-1}\left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i\mathbf{e}_i^\top\mathbf{\Pi}_i^{-1}\mathbf{g}_i$. For such clusters, replacing υ_{Di} with υ_{Ri} will assure a positive variance estimator, where $\upsilon_{Ri} = \mathbf{g}_i^\top\mathbf{\Pi}_i^{-1}\mathbf{e}_i\mathbf{e}_i^\top\mathbf{\Pi}_i^{-1}\mathbf{g}_i$. That is, we replace $\mathbf{I}_{n_i} - \mathbf{H}_{ii}$ with $\mathbf{I}_{n_i}$ when υ_{Di} is negative.

The jackknife is a popular variance estimation technique. Krewski and Rao (1981) present several asymptotically equivalent ways of writing the jackknife. The following form is a convenient starting point for the calculations that follow. Commonly, the jackknife is written as

$$\upsilon_{Jack} = \frac{n-1}{n}\sum_{i\in\mathfrak{s}_I}\left(\widehat{T}_{y(i)}^{gr} - \widehat{T}_{y(\cdot)}^{gr}\right)^2. \tag{2.37}$$

where $\widehat{T}_{y(i)}^{gr}$ is the value of the GREG estimator after removing cluster i and $\widehat{T}_{y(\cdot)}^{g}$ is the average of all $\widehat{T}_{y(i)}^{gr}$ estimates. When written as Equation (2.37), it is apparent that $\widehat{T}_{y(i)}^{gr}$ must be calculated for each cluster. Using Equation (2.37) can be computationally demanding

because n different estimates of $\widehat{T}^{gr}_{y(i)}$ must be computed. Alternatively, in Appendix A.5.2 on page 280, we show that the jackknife can be written as,

$$\upsilon_{Jack} = \frac{n}{n-1}\left[\sum_{i\in\mathfrak{s}_I}\left(D_i - \overline{D}\right)^2 - 2\sum_{i\in\mathfrak{s}_I}\left(D_i - \overline{D}\right)F_i + \sum_{i\in\mathfrak{s}_I}F_i^2\right]$$

where

$$\begin{aligned}
F_i &= \left(G_i - \overline{G}\right) - \frac{1}{n}\left(K_i - \overline{K}\right)\\
D_i &= \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\left(\mathbf{I} - \mathbf{H}_{ii}\right)^{-1}\mathbf{r}_i\\
K_i &= \left(\mathbf{1}^\top\mathbf{X} - n\mathbf{1}^\top\mathbf{\Pi}_i^{-1}\mathbf{X}_i\right)\left(\widehat{\mathbf{B}} - \mathbf{Q}_i\right)\\
\overline{K_i} &= \frac{1}{n}\sum_{i=1}^{n}K_i\\
G_i &= \mathbf{1}^\top\mathbf{\Pi}_i^{-1}\left(\mathbf{I} - \mathbf{H}_{ii}\right)^{-1}\left[\mathbf{H}_{ii}\mathbf{Y}_i - \widehat{\mathbf{Y}}_i\right]\\
\overline{G_i} &= \frac{1}{n}\sum_{i=1}^{n}G_i\\
\mathscr{Q}_i &= \mathbf{A}^{-1}\mathbf{X}_i^\top\mathbf{Q}_i\mathbf{\Pi}_i^{-1}\left(\mathbf{I} - \mathbf{H}_{ii}\right)^{-1}\mathbf{r}_i.
\end{aligned}$$

This form of the jackknife results in a significant reduction in computations since only one GREG estimate is needed, rather than n estimates.

In Appendix A.5.3 on page 285, we show that in large samples υ_{Jack} can be approximated by

$$\upsilon_{J1} = \frac{n}{n-1}\sum_{i\in\mathfrak{s}_I}\left(D_i - \overline{D}\right)^2.$$

Even further approximating the jackknife reveals that in large samples the jackknife

simplifies to

$$\upsilon_J = \sum_{i \in \mathfrak{s}_I} D_i^2 \tag{2.38}$$

$$= \sum_{i \in \mathfrak{s}_I} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1} \mathbf{e}_i \mathbf{e}_i^\top \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1} \mathbf{\Pi}_i^{-1} \mathbf{g}_i. \tag{2.39}$$

In Appendix A.5.3 on page 285, we show that υ_J is an approximation to the jackknife in large samples. An alternative motivation of υ_J relies on the asymptotic variance of $\mathbf{e}$. If $\sum_{j \neq i} \mathbf{H}_{ij} \mathbf{\Psi}_j \mathbf{H}_{ij}^\top$ is negligible, then we can set it equal to 0 in Equation (2.32). Then setting the result equal to $\mathbf{e}_i \mathbf{e}_i^\top$ and solving for $\mathbf{\Psi}_i$ gives an expression for $\mathbf{\Psi}_i$. Substituting this expression into Equation (2.31) on page 81 also gives υ_J.

None of these sandwich estimators includes finite population correction factors. Thus, they may tend to overestimate the sampling variance when a large proportion of the sample clusters is selected. To account for the finite population, we can further adjust all of the variance estimators in an *ad hoc* fashion by multiplying the variance estimators by a finite population correction factor, denoted f_{pc}, as developed by Kott (1988). This results in the following new estimators

$$\upsilon_R^\star = f_{pc} \sum_{i \in \mathfrak{s}_I} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{e}_i \mathbf{e}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{g}_i$$

$$\upsilon_D^\star = f_{pc} \sum_{i \in \mathfrak{s}_I} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1} \mathbf{e}_i \mathbf{e}_i^\top \mathbf{\Pi}_i^{-1} \mathbf{g}_i$$

$$\upsilon_{Jack}^\star = f_{pc} \frac{n}{n-1} \left[\sum_{i \in \mathfrak{s}_I} \left(D_i - \overline{D}\right)^2 - 2 \sum_{i \in \mathfrak{s}_I} \left(D_i - \overline{D}\right) F_i + \sum_{i \in \mathfrak{s}_I} F_i^2 \right]$$

$$\upsilon_{J1}^\star = f_{pc} \frac{n}{n-1} \sum_{i \in \mathfrak{s}_I} \left(D_i - \overline{D}\right)^2$$

$$\upsilon_J^\star = f_{pc} \sum_{i \in \mathfrak{s}_I} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1} \mathbf{e}_i \mathbf{e}_i^\top \left(\mathbf{I}_{n_i} - \mathbf{H}_{ii}\right)^{-1} \mathbf{\Pi}_i^{-1} \mathbf{g}_i$$

When a simple random sample is selected in the first stage, $f_{pc} = 1 - \frac{n}{N}$. According to Kott (1988), an appropriate correction when the first stage is selected with probabilities proportional to x is $f_{pc} = 1 - n\sum_{i=1}^{N} p_i^2$ where p_i is a single draw probability.

2.4 Simulation

We performed three simulation studies to test the performance of the new variance estimators in different populations. In each simulation, we estimated the quantities indicated in Table 2.1. To evaluate the variance estimators, we calculated their empirical bias, empirical variance, empirical mean squared error, and confidence interval coverage probabilities.

Table 2.1: Statistics of Interest for Clustered GREG Variance Simulation

Statistic	Description
$\widehat{t_y^{\pi}}$	The estimated total from the Horvitz-Thompson Estimator
$\widehat{t_y^{gr}}$	The estimated total from the GREG
υ_E	The empirical variance
υ_g	The design-based asymptotic variance estimator from Särndal et al. (1992)
υ_{wr}	The with-replacement variance estimator
υ_{JL}	The g-weighted with-replacement variance estimator from Yung and Rao (1996)
υ_R	The sandwich estimator
υ_D	The first leverage adjusted sandwich estimator
υ_{Jack}	The jackknife variance estimator
υ_{J1}	The first approximation to the jackknife variance estimator
υ_J	The second approximation to the jackknife variance estimator
$\upsilon_R^{\star}$	The sandwich estimator with a finite population adjustment
$\upsilon_D^{\star}$	The first leverage adjusted sandwich estimator with a finite population adjustment
$\upsilon_{Jack}^{\star}$	The jackknife variance estimator with a finite population adjustment
$\upsilon_{J1}^{\star}$	The first approximation to the jackknife variance estimator with a finite population adjustment
$\upsilon_J^{\star}$	The approximation to the jackknife variance estimator with a finite population adjustment

2.4.1 Data

We conducted simulations on three different populations to assess the performance of the variance estimators under a variety of situations. In the first population, we investigated the performance of the variance estimators when the first-stage sampling fraction was large and the sample size was moderate. The focus of the second simulation study was on the performance of the variance estimators under a relatively messy dataset and a small first-stage sample size. The final simulation study shows the performance of the variance estimators in large samples.

Table 2.2 summarizes the sample designs for the 18 simulation studies.

Table 2.2: Simulation Design

	Population	First Stage Sample	n	Second Stage Sample	Iterations
1	Third Grade	srswor	25	$m_i = 5$	$1,000$
2	Third Grade	srswor	50	$m_i = 5$	$1,000$
3	Third Grade	srswor	25	$f_i = \frac{675}{2,427}$	$1,000$
4	Third Grade	srswor	50	$f_i = \frac{675}{2,427}$	$1,000$
5	Third Grade	ppswor	25	$m_i = 5$	$1,000$
6	Third Grade	ppswor	50	$m_i = 5$	$1,000$
7	ACS	srswor	3	$m_i = 9$	$5,000$
8	ACS	srswor	15	$m_i = 9$	$5,000$
9	ACS	srswor	3	$f_i = \frac{30,430}{194,329}$	$5,000$
10	ACS	srswor	15	$f_i = \frac{30,430}{194,329}$	$5,000$
11	ACS	ppswor	3	$m_i = 9$	$5,000$
12	ACS	ppswor	15	$m_i = 9$	$5,000$
13	Simulated	srswor	300	$m_i = 2$	$1,000$
14	Simulated	srswor	1,500	$m_i = 2$	100
15	Simulated	srswor	300	$f_i = \frac{60,000}{195,164}$	$1,000$
16	Simulated	srswor	1,500	$f_i = \frac{60,000}{195,164}$	100
17	Simulated	ppswor	300	$m_i = 2$	$1,000$
18	Simulated	ppswor	1,500	$m_i = 2$	100

2.4.1.1 Third Grade Population

The first simulation study used the Third Grade population from Appendix B.6 of Valliant et al. (2000). This dataset contained the mathematics achievement scores for 2,427 third graders in 135 schools. The relatively small number of schools in this population and the fairly constant number of students in each school made it ideal for studying samples with large sampling fractions.

We used GREG to estimate the average mathematics achievement score for third graders in the population of schools. Altogether, we selected 6,000 samples using six sample designs. In the first sample design, we selected 1,000 simple random samples of 25 schools without replacement. Within each sampled school, we selected exactly five students. Because the number of students in each school varied from school to school, this sample design resulted in different unconditional probabilities of selection, but a fixed sample size of 125 students. The second sample design was similar to the first, except we selected 50 schools. Selecting 50 of the 135 schools resulted in a large first-stage sampling fraction of 0.37, necessitating a finite population correction factor.

In the third sample design, we selected 1,000 simple random samples of 25 schools without replacement. Within each sampled school, we selected students at a constant rate of $\frac{675}{2{,}427}$, yielding 1,000 samples with random sizes centered around 125 students. The result of this design was that each student had the same unconditional probability of selection. The fourth sample design was similar to the third, except we selected 50 schools. The sample sizes were also random under this design, with an average of 250 students. Since the third and fourth sample designs resulted in every unit getting the

same chance of selection, these sample designs are labeled srs epsem (equal probability selection mechanism).

In the fifth design, we selected 1,000 samples of 25 schools with probabilities proportional to the number of students in each school. Within each sampled school, we selected exactly five students, yielding 1,000 samples with exactly 125 students each. The sixth sample design was similar to the fifth, except we selected 50 schools. We selected 1,000 samples of size 250 students using this design. In the fifth and sixth designs, each student had the same unconditional probability of selection. Like the second and fourth sample designs, this sample design also had a large sampling fraction and warranted the need for a finite population correction factor to adjust the variance estimators. Altogether, we selected 6,000 samples; 1,000 from each design.

From each sample, we estimated the average achievement scores for the finite population using a GREG estimator. The assisting model was meant to replicate the clustered linear regression model in Section 9.6 of Valliant et al. (2000). The eleven explanatory variables used to model each student's math achievement score were: an intercept, sex (male or female), ethnicity (White/Asian, Black, Native American/Other, or Hispanic), language spoken at home (Always, Sometimes/Never), and type of community (Outskirts of a town or city, Village/City), and school enrollment. The total mathematics achievement estimated with the GREG estimator was divided by the number of students in the population, $2,427$, to get the average achievement score. The average achievement score for the population was 477.7019. For the full population, the R-squared was 0.9735, indicating a very strong linear relationship.

2.4.1.2 American Community Survey Population

The second simulation study used Census 2000 Summary File 3 data and American Community Survey (ACS) 2005 - 2009 Summary File data. The goal was to estimate the total number of housing units in Alabama as reported in the ACS Summary File. Block group level counts from Census 2000 were used as covariates in the assisting model.

To create the population, first all block group data was extracted from the ACS Summary File and the Census 2000 Summary File 3. Then, the two files were merged at the block group level. Block groups with 1,000 or more housing units in Census 2000 were removed because such large block groups had different characteristics than the majority of blocks. In many sampling designs such large units would be placed in a separate stratum. Also, block groups with extreme growth in the total number of housing units were also removed. Specifically block groups that had gained more than 10 units over twice the 2000 census count were removed.

Clusters were defined as counties and block groups were treated as units. At first glance it may seem odd to treat the block group as a unit. However, these simulations are motivated by the common task of selecting a sample of blocks, listing them, and then using the listings to estimate the total number of housing units in the finite population.

Clusters with fewer than 10 block groups or more than 120 block groups in them were removed from the frame of clusters. Overall, there were 61 clusters containing a total of 2,051 block groups and 1,109,499 housing units in the edited dataset. Altogether, six counties and 1,278 block groups containing 1,030,471 housing units were removed from the Alabama file.

Figure 2.1 shows two scatterplots. The first plot shows the total number of housing units in the block group as reported on the ACS summary file as a function of the 2000 census housing unit count. Each point represents one of the 2,051 block groups in the finite population. The red line is a nonparametric smoother, indicating a strong relationship between the two variables. The plot also shows some evidence of heteroscedasticity because the points appear to fan out as the 2000 census count increases. The second plot shows the residuals obtained by regressing the 2000 census housing unit count on the ACS housing unit count as a function of the ACS count. As the number of housing units reported on the ACS file increases, the model predictions appear to seriously underestimate the true number of housing units. This suggests considerable heteroscedasticity in variance.

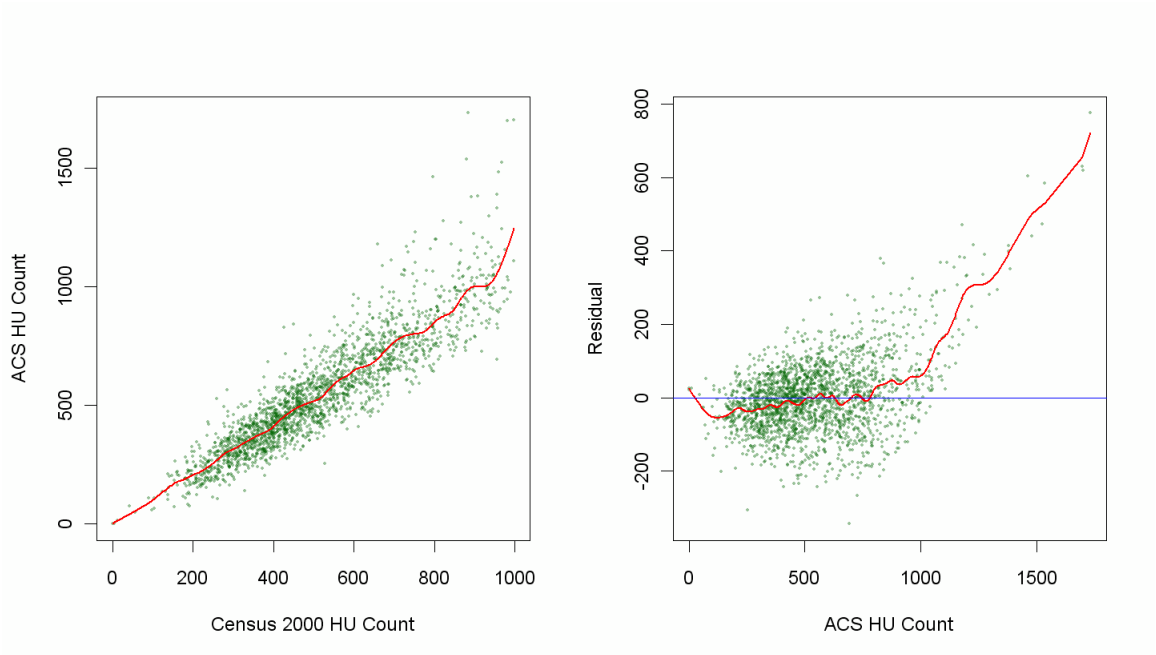

Figure 2.1: Scatter plot and residual plot for ACS population

As in the first simulation study, we tested six different sample designs. We selected

30,000 samples using six different selection mechanisms. In the first sample design, we selected 5,000 simple random samples of 3 clusters without replacement. In large national surveys, it is not uncommon to select a small number of primary sampling units in each strata. In this case, we treat our population of counties in Alabama as a design stratum and select three counties within that stratum. Many surveys select as few as one or two counties in each stratum which is emulated in the sample designs where only three clusters are selected. Within each cluster, we selected nine block groups. This design resulted in a constant sample size of 27 block groups. The second sample design was similar to the first, with the exception that 15 clusters were selected. The first two sample designs resulted in highly variable weights.

The third and fourth sample designs were created so that the unconditional probabilities of selection would be constant, even though the sample size was somewhat variable. In the third sample design, 5,000 simple random samples of 3 clusters were selected without replacement. Within each cluster, we selected block groups at a constant rate of $\frac{30{,}430}{194{,}329}$, yielding 5,000 samples with random sizes centered around 27 block groups. The fourth sample design was similar to the third, except we selected 15 clusters. We selected 5,000 samples using this design. The sample sizes were also random under this design, with an average of 135 sample units.

In the fifth design, we selected 5,000 samples of 3 clusters with probabilities proportional to the number of block groups in each cluster. Within each cluster, we selected exactly nine block groups, yielding 5,000 samples with exactly 27 block groups. The fourth sample design was similar to the third, except we selected 15 block groups. We selected 5,000 samples of 15 clusters using this design. Like the third and fourth de-

signs, these designs resulted in unconditional equal selection probabilities. Altogether, we selected 30,000 samples; 5,000 from each design.

From each sample, we estimated the total number of housing units in the finite population using a GREG estimator. The assisting model included an intercept and the Census 2000 count of housing units (H0340001). For the full population, the R-squared was 0.819, indicating a strong linear relationship.

2.4.1.3 Simulated Population

A population was created with a large number of clusters to assess the asymptotic characteristics of the variance estimators. Generated using a classic linear model, a total of 30,000 clusters were created, each with a random number of units. The number of units in each cluster was determined by adding three to a uniform random integer between 0 and 7. This created clusters ranging in size from 3 to 10 units. Altogether, the population contained 195,164 units within 30,000 clusters. For each unit, a positive covariate was created by exponentiating a standard normal variate and multiplying it by 1,000. Thus, $x_k \sim 1000 \exp N(0,1)$ where $N(0,1)$ is a normal random variate with mean of 0 and standard deviation of 1. A random response was created such that $y_k \sim N(1,000 + 2x_k, \frac{x_k}{2})$. Figure 2.2 shows scatter plots depicting the relationship between x_k and y_k for the finite population.

We selected 3,300 samples using six different probably selection mechanisms. In the first sample design, we selected 1,000 simple random samples of 300 clusters without replacement. Within each cluster, we selected 3 sample units, yielding 1,000 samples

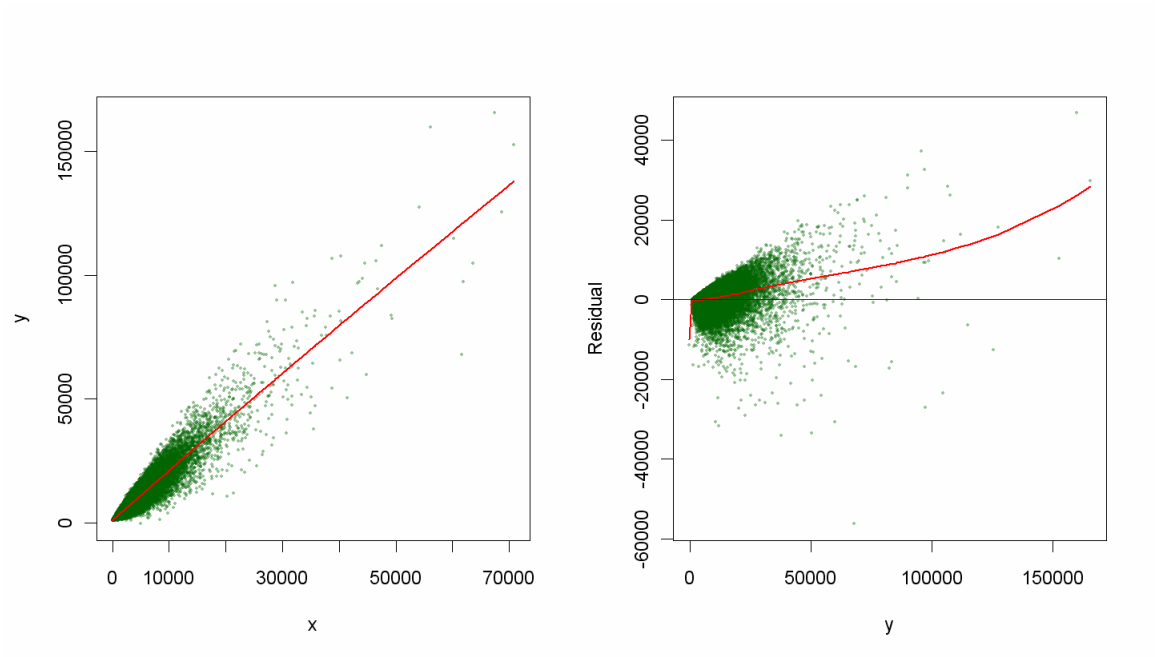

Figure 2.2: Scatter plot and residual for simulated population

with exactly 900 units each. In the second sample design, we selected 100 simple random samples of 1,500 clusters without replacement. Within each cluster, we selected 3 sample units, yielding 100 samples with exactly 4,500 units. We only selected 100 samples due to the excessive amount of computer time it took to select each sample.

In the third sample design, we selected 1,000 simple random samples of 300 clusters without replacement. Within each sample cluster, we selected units at a constant rate of $\frac{60{,}000}{195{,}164}$, yielding 1,000 samples with random sizes centered around 900 units. The fourth sample design was similar to the first, except we selected 1,500 clusters. We selected 100 samples using this design. The sample sizes were also random under this design, with an average of 4,500 units in each sample.

In the fifth design, we selected 1,000 samples of 300 clusters with probabilities proportional to the number of units in each cluster. Within each cluster, we selected

exactly three units, yielding 1,000 samples with exactly 900 units each. The last sample design was similar to the fifth, except we selected 1,500 clusters. We selected 100 samples of size 1,500 using this design.

From each sample, we estimated the total of the response using a GREG estimator. The true finite population total was 839,149,969. The assisting model included an intercept and $\mathbf{x}$. For the full population, the R-squared was 0.953, indicating a very strong linear relationship. Figure 2.2 shows a scatter plot of the population as well as a residual plot based on an ordinary least squares regression of x_k on y_k for the full population. There is clear evidence of heteroscedasticity of errors.

2.4.2 Results

We explored the bias, variability, and confidence interval coverage of the new and existing variance estimators. Appendix A.6 on page 288 shows tables documenting the full results of all simulations. In this section, we discuss the full results, but only show tables for some of the simulations.

Table 2.3 shows the central tendency of the π-estimator and the GREG estimator as well as the average value of the square root of the new variance estimators for the Third Grade Population across all simulations. We see that the π-estimator and the GREG estimator give similar estimates on average. Moreover, the estimates tend to be close to the true population values. The true mean for the third grade population was 477.7019. The true totals for the ACS and simulated populations are 1,109,499 and 839,149,969 respectively. On average, both the π-estimator and the GREG estimator are close to the

Table 2.3: Simulation Results of Variance Estimators for Clustered GREG Estimate

Estimator	srs fixed		srs epsem		pps epsem	
	Third Grade Population					
	$n = 25$	$n = 50$	$n = 25$	$n = 50$	$n = 25$	$n = 50$
Average $\frac{\hat{t}_y^{\pi}}{N}$	477.2	477.1	476.3	476.9	477.3	477.8
rmse $\frac{\hat{t}_y^{\pi}}{N}$	25.8	16.3	44.9	31.3	12.0	7.3
Average $\frac{\hat{t}_y^{gr}}{N}$	474.3	476.4	476.9	477.2	477.5	477.9
rmse $\frac{\hat{t}_y^{gr}}{N}$	14.8	8.2	10.7	7.1	11.0	6.4
$\sqrt{v_g}$	12.4	7.6	9.0	6.3	8.8	6.1
$\sqrt{v_{wr}}$	12.5	8.5	9.3	7.2	9.3	7.0
$\sqrt{v_{JL}}$	13.3	8.7	9.7	7.3	9.6	7.1
$\sqrt{v_r}$	13.2	8.7	9.5	7.3	9.4	7.0
$\sqrt{v_D}$	15.5	9.3	10.9	7.8	10.6	7.5
$\sqrt{v_J}$	18.9	10.0	12.7	8.4	12.0	7.9
$\sqrt{v_{Jack}}$	18.2	9.8	12.4	8.3	11.8	7.8
$\sqrt{v_{J1}}$	19.0	10.0	12.9	8.5	12.2	8.0
$\sqrt{v_r^{\star}}$	11.9	6.9	8.6	5.8	8.4	5.5
$\sqrt{v_D^{\star}}$	14.0	7.3	9.8	6.2	9.5	5.8
$\sqrt{v_J^{\star}}$	17.0	7.9	11.4	6.7	10.8	6.2
$\sqrt{v_{Jack}^{\star}}$	16.4	7.8	11.2	6.6	10.6	6.1
$\sqrt{v_{J1}^{\star}}$	17.1	7.9	11.7	6.7	11.0	6.2

true finite population quantities they estimate. However, the GREG estimator is much more efficient.

In every simulation study, the root mean squared error of the GREG estimator was less than the root mean squared error of the Horvitz-Thompson estimator. Such gains in efficiency will not always occur, but can be expected when the covariates are highly correlated to the response variable. This emphasizes the importance of building good assisting models and obtaining auxiliary data that is highly correlated to the response variable.

The sample size and design also effect the efficiency of the GREG estimator. For the Third Grade population, the srs epsem sample design is most efficient of the three when 25 clusters are selected; while the srs design is the least efficient. However, when 50 clusters

are selected, the pps design is most efficient. As evidenced in the simulation studies, there is no general sample design that will uniformly work best for all populations. Moreover, for some populations, one sample design may perform better for some estimators than others.

The estimates from the ACS population (see Table A.1 on page 288) with the simple random sample of 3 clusters and 9 units in each cluster stand out. The inverses of the probabilities of selection vary quite a bit for this sample design. The variability of these weights, coupled with some extreme observations in the population, has caused instability for some of the variance estimators. Namely, υ_J, υ_{Jack}, υ_{J1}, $\upsilon_J^\star$, $\upsilon_{Jack}^\star$, $\upsilon_{J1}^\star$ are rather absurd estimates. All six of these estimators contain g_k^2 terms which can be quite large and seriously inflate the variance estimators when multiplied by large sampling weights. As indicated by the median value of these estimators, they tend to give reasonable estimates in general; but sometimes are unreasonably far from the true value.

All variance estimates fluctuated from sample to sample, since they depended on sample quantities. To show the variability of the estimators, we created boxplots depicting the estimated standard errors as a fraction of the empirical standard error. For instance, we calculated 1,000 estimates of υ_R from the Third Grade Population with simple random samples of 25 clusters. Taking the square root of each variance estimate, gave 1,000 standard error estimates. Further, dividing each of the 1,000 standard error estimates by the empirical standard error, $\sqrt{\frac{1}{1{,}000}\sum_{\nu=1}^{1{,}000}\left[t_{y,\nu}^{gr} - \frac{1}{1{,}000}\sum_{\nu=1}^{1{,}000} t_{y,\nu}^{gr}\right]^2}$, gave 1,000 relative standard error estimates. An estimate of 1 represents that the estimated variance was equal to the empirical variance, while an estimate of 1.5 indicates that the estimated variance was 1.5 times larger than then empirical variance. The boxplots in Figure 2.3 depict

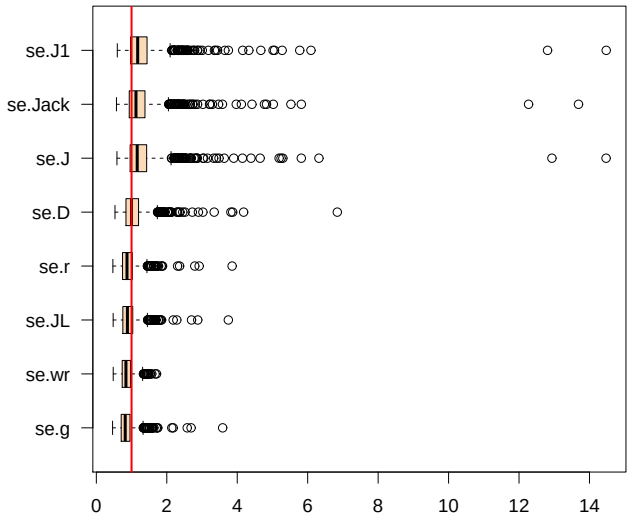

Figure 2.3: Boxplots of relative standard error estimates for SRS samples of size 25 from third grade population

these relative standard errors for some of the estimators. As we see, some samples yield large standard error estimates, even though the majority of samples are much closer to the empirical variance.

Additionally, Figure, 2.4 shows boxplots for the simple random samples of size 50 clusters from the Third Grade Population. Clearly, there are fewer outliers as the sample size increases and the spread of the estimators decreases.

To quantify the variability of the estimated standard errors, we calculated the standard error of the estimated standard errors as well as the root mean squared error of the standard errors. For example, the empirical standard error of the sandwich estimator is $\sqrt{\frac{1}{1{,}000}\sum_{\nu=1}^{1{,}000}\left[se_{R,\nu}-\frac{1}{1{,}000}\sum_{\nu=1}^{1{,}000}se_{R,\nu}\right]^2}$ and the root mean squared error of the standard errors is $\sqrt{\frac{1}{1{,}000}\sum_{\nu=1}^{1{,}000}\left[se_{R,\nu}-\frac{1}{1{,}000}se_E\right]^2}$ where se_E is the empirical standard error.

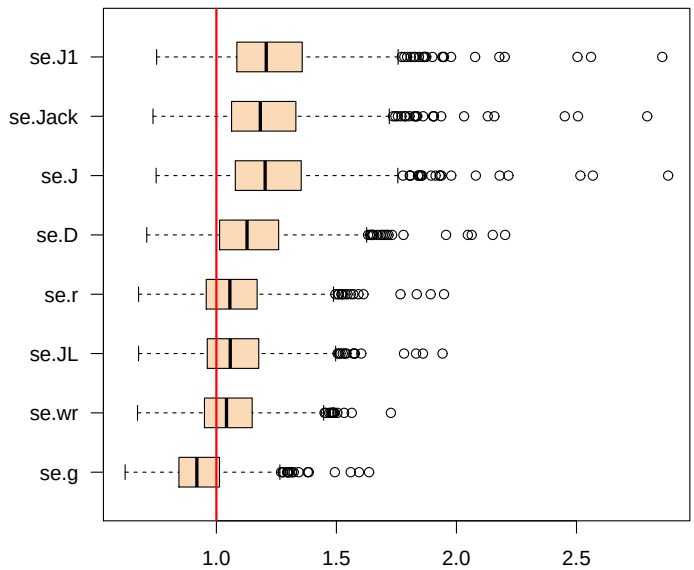

Figure 2.4: Boxplots of relative standard error estimates for SRS samples of size 50 from third grade population

Table 2.4 shows the standard deviation and the root mean squared error of the estimated standard errors of the GREG estimator for the Third Grade simulations selected using the srs and pps designs. Additionally, the minimum, maximum, and quartiles of the variance estimators are shown in Table 2.4. The first column names the estimator. The second column shows the mean value of the estimator across all 1,000 samples. For example, the element in the second column of the first row was calculated by $\frac{1}{1{,}000}\sum_{\nu=1}^{1{,}000}\sqrt{\upsilon_{g\nu}}$. The third column shows the empirical standard error of the standard error estimator. For $\sqrt{\upsilon_g}$, this simplifies to $\frac{1}{1{,}000}\sum_{\nu=1}^{1{,}000}\sqrt{\upsilon_{g\nu}} - \overline{\sqrt{\upsilon_g}}$. The fourth column shows the empirical root mean squared error, which was calculated by $\frac{1}{1{,}000}\sum_{\nu=1}^{1{,}000}\sqrt{\upsilon_{g\nu}} - \sqrt{\upsilon_E}$ where $\sqrt{\upsilon_E}$ is the square root of the empirical variance of t_y^{gr}. The final columns show summary statistics for the relative square root estimators. For example, for $\sqrt{\upsilon_g}$, we first divided $\sqrt{\upsilon_{g\nu}}$

by $\sqrt{\upsilon_E}$. We then ordered these 1,000 estimates and determined the minimum, maximum, quartiles, and mean of the 1,000 estimates. Table 2.4 show the variability and range of values for the variance estimators. Generally, variance estimators with smaller root mean squared error are preferred to estimators with larger root mean squared error. Appendix A.6 on page 288 shows the full results for all simulations.

Lastly, Table 2.5 shows the 95% confidence interval coverage for all of the estimators when the distribution of the GREG estimator is assumed to be normal. That is, we computed, $\left[t_y^{gr} - 1.96\sqrt{\upsilon}, t_y^{gr} + 1.96\sqrt{\upsilon}\right]$ and noted how often the true value fell below, above, and inside this range. In addition to the new and old estimators, Table 2.5 also shows the confidence interval coverage attained when the empirical variance, υ_E, was used to form the confidence intervals. Ideally, the population mean should be within the estimated 95% confidence interval for 95% of the samples. Furthermore, the true mean should be below the 95% confidence interval for 2.5% of the samples and above the confidence interval for an equal number of samples.

2.4.2.1 υ_g

Särndal et al. (1992) discuss the properties of υ_g in clustered samples. Although Särndal et al. (1992) show that υ_g is asymptotically unbiased, they note that complex estimators, such as υ_g, can be rather slow in converging to the true variance. Our simulations confirm this finding. υ_g consistently underestimates the empirical variance in simulations. Table 2.4 shows the average and median values of $\sqrt{\upsilon_g}$ as a percent of the empirical standard error, that is $\frac{\sqrt{\upsilon_g}}{\sqrt{\upsilon_E}}$, for the Third Grade Population. Invariably, these values are less

Table 2.4: Variability of Sandwich Estimators for School Population

$\hat{\theta}$	$\bar{\hat{\theta}}$	se $\hat{\theta}$	rmse $\hat{\theta}$	Distribution of $\hat{\theta}/\sqrt{v_E}$ Min	1st Qu.	Median	Mean	3rd Qu.	Max
srs n = 25									
$\sqrt{v_g}$	12.42	3.54	4.06	0.46	0.71	0.82	0.86	0.96	3.59
$\sqrt{v_{wr}}$	12.49	2.72	3.32	0.48	0.73	0.84	0.87	0.97	1.71
$\sqrt{v_{JL}}$	13.32	3.79	3.94	0.48	0.75	0.88	0.92	1.03	3.75
$\sqrt{v_r}$	13.22	3.88	4.06	0.47	0.74	0.87	0.92	1.02	3.85
$\sqrt{v_D}$	15.52	5.86	5.96	0.53	0.84	1.00	1.08	1.20	6.84
$\sqrt{v_J}$	18.88	11.38	12.22	0.59	0.96	1.16	1.31	1.43	14.47
$\sqrt{v_{Jack}}$	18.19	10.69	11.34	0.57	0.93	1.13	1.26	1.38	13.69
$\sqrt{v_{J1}}$	18.98	11.23	12.12	0.59	0.97	1.17	1.32	1.44	14.48
$\sqrt{v_r^\star}$	11.93	3.51	4.29	0.42	0.67	0.79	0.83	0.92	3.48
$\sqrt{v_D^\star}$	14.01	5.29	5.30	0.48	0.76	0.90	0.97	1.08	6.17
$\sqrt{v_J^\star}$	17.04	10.27	10.60	0.53	0.87	1.05	1.18	1.29	13.06
$\sqrt{v_{Jack}^\star}$	16.42	9.65	9.85	0.52	0.84	1.02	1.14	1.25	12.35
$\sqrt{v_{J1}^\star}$	17.14	10.14	10.49	0.54	0.88	1.06	1.19	1.30	13.07
srs n = 50									
$\sqrt{v_g}$	7.56	1.10	1.21	0.62	0.84	0.92	0.94	1.01	1.64
$\sqrt{v_{wr}}$	8.51	1.25	1.33	0.67	0.95	1.04	1.06	1.15	1.73
$\sqrt{v_{JL}}$	8.69	1.36	1.50	0.68	0.96	1.06	1.08	1.18	1.94
$\sqrt{v_r}$	8.66	1.38	1.50	0.68	0.96	1.06	1.07	1.17	1.95
$\sqrt{v_D}$	9.27	1.57	1.98	0.71	1.01	1.13	1.15	1.26	2.20
$\sqrt{v_J}$	9.97	1.86	2.66	0.75	1.08	1.20	1.24	1.35	2.88
$\sqrt{v_{Jack}}$	9.80	1.81	2.51	0.74	1.06	1.18	1.22	1.33	2.79
$\sqrt{v_{J1}}$	10.01	1.84	2.68	0.75	1.09	1.21	1.24	1.36	2.86
$\sqrt{v_r^\star}$	6.87	1.09	1.61	0.54	0.76	0.84	0.85	0.93	1.55
$\sqrt{v_D^\star}$	7.35	1.25	1.43	0.56	0.80	0.89	0.91	1.00	1.75
$\sqrt{v_J^\star}$	7.91	1.47	1.48	0.59	0.86	0.95	0.98	1.07	2.29
$\sqrt{v_{Jack}^\star}$	7.78	1.43	1.46	0.58	0.84	0.94	0.97	1.06	2.22
$\sqrt{v_{J1}^\star}$	7.94	1.46	1.47	0.60	0.86	0.96	0.99	1.08	2.27
pps n = 25									
$\sqrt{v_g}$	8.84	1.44	2.62	0.48	0.71	0.79	0.80	0.88	1.33
$\sqrt{v_{wr}}$	9.30	1.40	2.22	0.51	0.76	0.84	0.84	0.92	1.30
$\sqrt{v_{JL}}$	9.57	1.65	2.21	0.50	0.76	0.86	0.87	0.96	1.46
$\sqrt{v_r}$	9.38	1.62	2.32	0.49	0.75	0.84	0.85	0.94	1.43
$\sqrt{v_D}$	10.55	1.95	2.01	0.53	0.83	0.94	0.96	1.06	1.66
$\sqrt{v_J}$	12.00	2.47	2.65	0.59	0.94	1.06	1.09	1.21	2.15
$\sqrt{v_{Jack}}$	11.76	2.42	2.53	0.57	0.92	1.04	1.07	1.18	2.10
$\sqrt{v_{J1}}$	12.24	2.52	2.79	0.60	0.96	1.08	1.11	1.23	2.19
$\sqrt{v_r^\star}$	8.41	1.45	3.00	0.43	0.67	0.76	0.76	0.84	1.30
$\sqrt{v_D^\star}$	9.46	1.74	2.35	0.47	0.75	0.84	0.86	0.95	1.51
$\sqrt{v_J^\star}$	10.76	2.21	2.22	0.52	0.84	0.95	0.98	1.08	1.90
$\sqrt{v_{Jack}^\star}$	10.55	2.16	2.22	0.51	0.82	0.93	0.96	1.06	1.86
$\sqrt{v_{J1}^\star}$	10.98	2.25	2.25	0.53	0.86	0.97	1.00	1.10	1.93
pps n = 50									
$\sqrt{v_g}$	6.10	0.61	0.69	0.72	0.88	0.95	0.95	1.01	1.28
$\sqrt{v_{wr}}$	6.98	0.71	0.90	0.78	1.00	1.09	1.09	1.16	1.47
$\sqrt{v_{JL}}$	7.11	0.83	1.07	0.81	1.01	1.11	1.11	1.19	1.52
$\sqrt{v_r}$	7.04	0.82	1.02	0.80	1.00	1.09	1.09	1.18	1.50
$\sqrt{v_D}$	7.45	0.91	1.37	0.84	1.06	1.15	1.16	1.25	1.64
$\sqrt{v_J}$	7.90	1.02	1.79	0.88	1.11	1.22	1.23	1.33	1.83
$\sqrt{v_{Jack}}$	7.82	1.01	1.72	0.88	1.10	1.21	1.22	1.31	1.81
$\sqrt{v_{J1}}$	7.98	1.03	1.87	0.89	1.13	1.23	1.24	1.34	1.85
$\sqrt{v_r^\star}$	5.49	0.64	1.14	0.62	0.78	0.85	0.85	0.92	1.16
$\sqrt{v_D^\star}$	5.81	0.71	0.95	0.65	0.82	0.90	0.90	0.97	1.28
$\sqrt{v_J^\star}$	6.16	0.80	0.85	0.68	0.87	0.95	0.96	1.03	1.43
$\sqrt{v_{Jack}^\star}$	6.10	0.79	0.86	0.67	0.86	0.94	0.95	1.02	1.42
$\sqrt{v_{J1}^\star}$	6.22	0.81	0.84	0.69	0.88	0.96	0.97	1.04	1.44

Table 2.5: Coverage of Sandwich Estimators

Estimator	Third Grade			ACS			Simulation		
	Lower	Middle	Upper	Lower	Middle	Upper	Lower	Middle	Upper
	srs n = 25			srs n = 3			srs n = 300		
$\sqrt{v_E}$	3.9	94.4	1.7	3.9	95.3	0.8	2.7	95.0	2.3
$\sqrt{v_g}$	9.0	89.0	2.0	17.9	78.1	4.1	4.4	93.4	2.2
$\sqrt{v_{wr}}$	7.8	89.5	2.7	23.5	69.5	6.9	3.9	92.8	3.3
$\sqrt{v_{JL}}$	7.1	91.1	1.8	22.0	72.1	5.8	4.4	93.4	2.2
$\sqrt{v_r}$	7.3	90.9	1.8	18.3	77.2	4.5	4.4	93.4	2.2
$\sqrt{v_D}$	4.5	94.5	1.0	10.8	87.0	2.2	3.7	94.2	2.1
$\sqrt{v_J}$	2.5	97.2	0.3	4.9	94.1	1.0	3.6	94.4	2.0
$\sqrt{v_{Jack}}$	2.6	97.0	0.4	11.8	85.3	3.0	3.6	94.4	2.0
$\sqrt{v_{J1}}$	2.3	97.4	0.3	6.3	92.1	1.6	3.6	94.4	2.0
$\sqrt{v_r^\star}$	9.8	87.9	2.3	18.9	76.4	4.8	4.4	93.4	2.2
$\sqrt{v_D^\star}$	6.7	91.8	1.5	11.4	86.3	2.3	3.8	94.1	2.1
$\sqrt{v_J^\star}$	4.0	95.3	0.7	5.2	93.7	1.0	3.6	94.4	2.0
$\sqrt{v_{Jack}^\star}$	4.7	94.6	0.7	12.1	84.9	3.0	3.7	94.2	2.1
$\sqrt{v_{J1}^\star}$	3.9	95.4	0.7	6.5	91.8	1.6	3.6	94.3	2.1
	srs n = 50			srs n = 15			srs n = 1,500		
$\sqrt{v_E}$	3.7	94.7	1.6	4.3	94.3	1.4	1.0	96.0	3.0
$\sqrt{v_g}$	6.2	92.4	1.4	8.7	89.8	1.6	1.0	95.0	4.0
$\sqrt{v_{wr}}$	4.5	94.5	1.0	9.3	88.5	2.2	1.0	96.0	3.0
$\sqrt{v_{JL}}$	3.8	95.6	0.6	9.0	89.0	2.0	1.0	95.0	4.0
$\sqrt{v_r}$	4.0	95.4	0.6	8.2	90.3	1.6	1.0	95.0	4.0
$\sqrt{v_D}$	3.1	96.4	0.5	6.4	92.6	1.0	1.0	95.0	4.0
$\sqrt{v_J}$	2.2	97.5	0.3	5.2	94.3	0.5	1.0	95.0	4.0
$\sqrt{v_{Jack}}$	2.3	97.4	0.3	6.8	92.0	1.2	1.0	95.0	4.0
$\sqrt{v_{J1}}$	2.1	97.6	0.3	5.8	93.4	0.8	1.0	95.0	4.0
$\sqrt{v_r^\star}$	8.2	89.0	2.8	11.4	85.9	2.8	1.0	95.0	4.0
$\sqrt{v_D^\star}$	7.4	90.9	1.7	9.4	88.6	2.0	1.0	95.0	4.0
$\sqrt{v_J^\star}$	5.9	93.1	1.0	7.3	91.3	1.4	1.0	95.0	4.0
$\sqrt{v_{Jack}^\star}$	5.9	93.0	1.1	9.4	88.5	2.1	1.0	95.0	4.0
$\sqrt{v_{J1}^\star}$	5.8	93.1	1.1	8.0	90.4	1.6	1.0	95.0	4.0

than 1 in all simulations for the Third Grade Population, indicating a tendency to underestimate the empirical standard error. Only in the largest samples of 1,500 clusters in the simulated population does v_g overestimate the empirical variance (see Table A.4).

The fact that v_g tends to underestimate the empirical variance also impacts inferences. Table 2.5 indicates that inferences with v_g might lead to overstating the significance of statistics. Indeed, confidence interval coverage tends to be less than the nominal 95%. Compared to competing estimators, v_g tends to perform among the worst in terms of confidence interval coverage. Moreover, confidence interval coverage is skewed, with the true value being below the confidence interval more often than it is above the confidence interval.

Despite its tendency to underestimate the empirical variance, v_g has some attractive features. First, it does not overestimate the empirical variance when the sampling fraction is large. Indeed, Table 2.4 shows that on average v_g is lower than the empirical variance, even when the sampling fraction is over 0.3. Second, it is often less variable than other estimators and many times has the smallest root mean squared error. In fact, for the samples of 50 clusters from the Third Grade Population, v_g has the smallest root mean squared error for all three sample designs. It is consistently among the best in terms of root mean squared error for the other two populations as well.

2.4.2.2 v_{wr} and v_{JL}

The only difference between v_{wr} and v_{JL} is that the residuals in v_{JL} have been adjusted with g-weights, while v_{wr} does not use the adjusted residuals. Because they share so much in common, estimates from both estimators are similar. In the Third Grade population, v_{JL} tends to be larger than v_{wr}; whilst, v_{wr} tends to be larger in the ACS population. Unless the sampling fraction is large or the sample size is large, both estimators tend to underestimate the empirical variance. In the ACS and simulated populations, v_{JL} appears to be better than v_{wr} in terms of mean squared error. The opposite is the case for the Third Grade population.

As long as the sampling fraction is small, v_{wr} and v_{JL} tend to outperform the leverage adjusted variance estimators in terms of mean squared error. This is primarily related to the fact that v_{wr} and v_{JL} are less variable than the leverage adjusted variance estimators. On the other hand, v_{wr} and v_{JL} tend to underestimate the empirical variance in small

samples. As shown in Table 2.5, the bias of v_{wr} and v_{JL} in small samples plays a roll in inference. For the smaller sample sizes, confidence intervals based on v_{wr} and v_{JL} tend to exclude the true value more often than the nominal 5% rate. Furthermore, the confidence intervals tend to be below the true value when $\hat{t}_y^{gr}$ is less than the true total. Further evidence of this can be seen in the positive correlation between $\hat{t}_y^{gr}$ and the two variance estimators, indicating that the variance estimators tend to be larger when $\hat{t}_y^{gr}$ exceeds the true value.

When the sampling fraction is large, v_{wr} and v_{JL} can slightly overestimate the empirical variance. However, this positive bias tends to be less than the overestimation that can be expected from the leverage adjusted variance estimators. Even though v_{wr} and v_{JL} do not have finite population correction factors, they tend to be competitive with the leverage adjusted variance estimators that have a finite population correction adjustment. The tendency of v_{wr} and v_{JL} to underestimate the empirical variance works to their advantage when the first-stage sampling fraction is rather large. In such situations, the mean squared error of v_{wr} and v_{JL} tend to be among the best of the variance estimators included in the simulation. According to the simulations, confidence interval coverage also improves as the first-stage sample size increase, regardless of the sampling fraction.

2.4.2.3 v_r and $v_r^\star$

As expected from the theory, v_r and $v_r^\star$ are biased in small samples. However, as the sample size increases and the sampling fraction remains small, the bias of v_r and $v_r^\star$ decreases. When the sampling fraction is large, v_r tends to slightly overestimate the

empirical variance, while $v_r^\star$ tends to underestimate it.

v_r is similar to v_{wr} and v_{JL} and often between the two values. In terms of the root mean squared error and confidence interval coverage, v_r is comparable to v_{wr} and v_{JL}.

Unless the sample size is very large, $v_r^\star$ tends to severely underestimate the empirical variance and is not attractive for that reason. When the first-stage sampling fraction is large and the first-stage sample size is small or moderate, v_r seems to outperform $v_r^\star$ in terms of bias, variability, root mean squared error, and confidence interval coverage.

2.4.2.4 v_D and $v_D^\star$

Of the leverage adjusted variance estimators without a finite population correction term, v_D seems to fare the best. In the small to moderate samples, v_D tends to slightly overestimate the empirical variance, but this overestimation tends to be smaller than that of v_{Jack}, v_J, and v_{J1}. Thus, v_D is a somewhat conservative variance estimator. In terms of the root mean squared error, v_D tends to outperform v_{Jack}, v_J, and v_{J1}; sometimes having as much as half the root mean squared error as the Jackknife estimators. Although v_g, v_{wr}, v_{JL}, and v_r often have smaller root mean squared error than v_D, they tend to underestimate the empirical variance and underestimate the confidence interval coverage more frequently than v_D. For these reasons, v_D should be considered when the first-stage sample size is small.

The advantage of $v_D^\star$ when the first-stage sampling fraction is large is not clear. The finite population correction factor adjustment seems to deflate v_D too much, resulting in understating the empirical variance. On the other hand, the adjustment adds stability to

the variance estimator, which can lead to reductions in the root mean squared error over v_D. Furthermore, the confidence interval coverage of v_D tends to be closer to the nominal rate than $v_D^\star$, even when the first-stage sampling fraction is large.

One feature of v_D and $v_D^\star$ is that both cluster specific contributions, $v_{D,i}$ and $v_{D,i}^\star$, as well as the overall variance estimates can be negative, if adjustments are not made. Negative estimates were more common when the second stage sample sizes were small and the weights were quite variable. For example, for the ACS population, almost 28% of the simple random samples of 3 clusters and $m_i = 9$ resulted in at least one negative variance contribution for a cluster. More commonly, about 10% of the samples contained at least one negative variance estimate for a cluster. In the Third Grade population, 16% to 27% of the samples had at least one negative value of v_{Di}. In the simulated population with large sample sizes, v_{Di} was negative in less than 5% of the samples. With the *ad hoc* correction of setting $I_i - H_{ii}$ to I_i, v_D is one of the most attractive variance estimators because it tends to slightly overestimate the empirical variance, has some of the best confidence interval coverage, and has reasonable root mean squared error.

2.4.2.5 v_{Jack}, v_J, v_{J1}, $v_{Jack}^\star$, $v_J^\star$, and $v_{J1}^\star$

The jackknife variance estimators tend to overestimate the empirical variance. In terms of ordering: v_{Jack} tends to be less than v_J, which is often less than v_{J1}. v_{J1} is often the largest of the variance estimators, sometimes well over 20% larger than the empirical variance. In small samples, all three not only have undesirably large positive bias, but are also highly variable. All of the jackknife estimators involve the $\left(I_i - H_{(ii)}\right)^{-1}$ factors that

tend to inflate the variance estimates. Some samples yield jackknife variance estimates that are well over ten times the empirical variance. This is especially true in the small samples, although the root mean squared error of the jackknife variance estimators is often larger than the other variance estimators, regardless of the sample size.

The finite population correction adjustments certainly help the overestimation problem, but often overcompensate, resulting in underestimates of the variance when the first-stage sample size is large. Interestingly, the estimators with the finite population correction adjustment seem to fare best when the first-stage sampling fraction is small. In such cases, the adjustments reduce the overestimation significantly; yet, often still slightly overestimate the empirical variance, enabling conservative inference. Although slightly improved from the jackknife variance estimators without the adjustments, $\upsilon^{\star}_{Jack}$, $\upsilon^{\star}_{J}$, and $\upsilon^{\star}_{J1}$ still have some of the largest root mean squared errors.

In terms of confidence interval coverage, the jackknife estimators frequently are closer to the nominal coverage rate when compared to all other estimators. In small samples υ_J and υ_{J1} come closest to the nominal coverage rate followed by υ_{Jack}, which tends to be lower than the nominal rate in the ACS population. Although they tend to slightly overstate the confidence interval coverage, $\upsilon^{\star}_{Jack}$, $\upsilon^{\star}_{J}$, and $\upsilon^{\star}_{J1}$ are nonetheless attractive estimators in terms of confidence interval coverage and tend to outperform other estimates.

2.4.2.6 Summary

All variance estimators perform similarly across all three sample designs. When the sample size is held constant, the ordering of the variance estimators is very similar from sample design to sample design. This is not to say that the sample design has minimal effect on the variance estimators. Indeed, the sample design impacts the central tendency and variability of the estimators. For the three populations used in this study, the design with the unequal probabilities of selection produced the most variable estimates compared to the other two designs. This finding will not be true for all populations. Generally, the more optimal sample designs should have less variable variance estimators.

In the ACS population, we explored the performance of the variance estimators when the first-stage sample size was small; either 3 or 15. Because there were only 61 clusters in the ACS population, the samples with 15 clusters had a large first-stage sampling fraction of 0.25. In the smaller samples $\upsilon_r^\star$, υ_g, υ_{wr}, υ_r, and υ_{JL} all underestimated the empirical root mean squared error of the GREG estimator on average. Even when the sampling fraction was 0.25, these estimators still underestimated the empirical sampling variance. On the other extreme, υ_J and υ_{J1} frequently overestimated the empirical root mean squared error. When the first-stage sampling fraction was small, $\upsilon_J^\star$ and $\upsilon_{J1}^\star$ tended to overestimate the empirical variance; while slightly underestimating it when the first-stage sampling fraction was large. In small samples, no variance estimator performed perfectly; however, υ_D and υ_{Jack} tend to be close to the empirical root mean squared error on average. The confidence interval coverage of the leverage-adjusted sandwich estimators, as well as the jackknife, were often closer to the nominal coverage rate than the

linearized and with-replacement estimators. The cost in the improvement in confidence interval coverage seems to have come at the expense of the root mean squared error of the estimators. Indeed, the root mean squared error of the leverage-adjusted estimators tends to be larger than the linearized and with-replacement estimators. Although some of the new estimators are less biased than the established estimators, they are more variable, especially in small samples.

In the Third Grade population, we explored the performance of the variance estimators when the first-stage sample size was moderate; either 25 or 50. Again, the larger sample size had a large sampling fraction well over 0.3. With this large sampling fraction, υ_{wr}, υ_{JL}, υ_r, υ_D, υ_{Jack}, υ_J, and υ_{J1} all regularly overestimated the empirical root mean squared error of $\frac{1}{N}\hat{t}_y^{gr}$. This result was somewhat expected because none of these estimators had a finite population correction factor adjustment. On the other extreme, every estimator with Kott's finite population correction adjustment underestimated the empirical variance on average when the sampling fraction was large. With the large sampling fraction, υ_g, υ_{wr}, υ_{JL}, and υ_r tend to have smaller root mean squared error and slightly overestimate the empirical root mean squared error. Furthermore, their confidence interval coverage is close to the nominal value, making them clearly the better estimators when the sample size is moderate and the sampling fraction is large.

On the other hand, with a moderate sample size and a small first-stage sampling fraction, υ_D and $\upsilon_{Jack}^{\star}$ tend to be the closest to the empirical root mean squared error. The exception is with the pps samples where $\upsilon_J^{\star}$, and $\upsilon_{J1}^{\star}$ are closer to the empirical root mean squared error. Of these four estimators, υ_D has the smallest root mean squared error. All four are among the best in terms of confidence interval coverage as well. Although none

of these leverage adjusted sandwich estimators have the lowest root mean squared error, they are all close to the root mean squared error of $\hat{t}_y^{gr}$ on average, have confidence interval coverage close to the nominal rate, and are slightly more variable than the linearized estimators.

From the Simulated dataset, we see how the variance estimators perform in large samples. All of the variance estimators are asymptotically unbiased and should provide confidence interval coverage close to the nominal value. Of course, υ_J, υ_{Jack}, and υ_{J1} also continue to overestimate the empirical root mean squared error, but are more reasonable estimators in large samples. In terms of the bias and confidence interval coverage, all of the estimators are practically the same in large samples. However, in terms of variability, the jackknife variance estimators are consistently more variable than the other estimators. The estimators with the smallest root mean squared errors are υ_g, υ_r, and $\upsilon_r^\star$.

2.5 Conclusion

Accurately estimating sampling errors for GREG estimators in complex samples can be a challenge. Yet, estimates of sampling errors are essential to solving many problems and making inferences to a population. In this chapter, we constructed new sandwich variance estimators for the GREG in two-staged samples and evaluated their model-based and design-based properties.

Sandwich estimators provide an alternative technique to estimating the variance of GREG estimators in complex samples. At the expense of inflating the root mean squared error of the variance estimator, leverage-adjusted sandwich estimators can be constructed

with confidence interval coverage that is closer to the nominal value in small to moderate samples. Depending on the sample design and population characteristics, leverage-adjusted sandwich estimators can produce less biased variance estimates and better inferences when compared to the standard methods. This study investigated and assessed the sandwich variance estimation technique for calculating standard errors of GREG estimators in complex samples.

Chapter 3

Multivariate Logistic-Assisted Estimators of Totals from Clustered Survey Samples in the Presence of Complete Auxiliary Information

3.1 Introduction

The collection of categorical data in complex surveys is ubiquitous. Demographic, crime, employment, health, discrete-choice, brand preference, satisfaction, and political opinion questions often ask the respondent to select one or more options from a finite set of categories. Analyzing categorical data from a complex survey usually requires specialized techniques. In this chapter, we extend some calibrated estimators of multinomial data developed for single-staged samples to complex two-staged sample designs.

Data collected in multiple stages is also common. In an effort to reduce travel and other field costs, multiple-staged samples are generally selected in large face to face surveys. However, the analysis of clustered data is frequently more complicated than data collected in a single stage.

The sample design impacts data analysis, estimation, and inference. If the sample design is not taken into account, point estimators, variance estimators, and test statistics may be misleading. For this reason, estimators based on single-staged samples are rarely appropriate for multi-staged sample designs. Clustered samples also differ from single-staged samples in the level of data that may be available. Auxiliary data may be available

at the unit level, at the cluster level, at both the cluster and unit level, or not at all. The level of covariates and whether they are available for the sample only or for the full population also impacts how one constructs estimators.

In this paper, we focus on the case where auxiliary data are available for all units in the population. We call this the case of complete unit auxiliaries. Although not always the case, auxiliary data are often available for all units in the population. Address based sampling frames, national population registers, marketing databases, and professional organizations often contain a wealth of data about all units on the sampling frame. When such data are available, it is often advantageous to calibrate sample totals to known frame totals. Calibrated estimators often have lower nonsampling and sampling errors when compared to more naive estimators.

Current estimators of totals from clustered samples are not well suited for multinomial data. One of the key characteristics of multinomial data is that the response options are conditional on a known number. The linear assisting model does not preserve this important characteristic of multinomial data. For example, when estimating the proportion of persons who are employed, unemployed, and not in the labor force, linear models may give proportions that do not add up to 1 and they may give individual predictions that are negative or greater than 1. Assisting models specifically built to analyze categorical data can improve point estimation and reduce sampling errors.

Our new research in this chapter extends calibrated logistic-assisted point estimates of totals to two-staged samples and evaluates serval variance estimators of the logistic-assisted calibration estimators. We develop and compare three different kinds of logistic-assisted point estimators: the logistic general regression (LGREG) estimator, the model-

calibration estimator, and the model-calibrated maximum pseudoempirical likelihood estimator. We also propose several variance estimators for these logistic-assisted estimators.

3.1.1 Multinomial Logistic Regression

Logistic regression is a popular tool used to analyze binary, binomial, percent, and multinomial response data. It is widely used in medical and epidemiological studies, economics, survey methodology, and a host of other fields. Unlike linear regression, logistic regression is well suited to the analysis of binary and binomial data because predicted values are bounded, the interpretation of coefficients is closely linked to the odds ratio, and the variance of the observations does not need to be independent of the mean.

Numerous textbooks and papers devote attention to the model fitting, parameter estimation, and interpretation of logistic regression (see Agresti (2002), Bishop et al. (2007), McCullagh and Nelder (1999), Hosmer and Lemeshow (2000), Hilbe (2009), and Shao (2003)). All of these introductory texts focus on estimating superpopulation parameters, such as $\boldsymbol{\beta}$, from logistic regression models.

Hilbe (2009, p. 270) argues that the term *Logistic Regression* is used to describe several different kinds of models that can be characterized by the distribution of the response variable. Here we provide results for multinomial logistic regression. In Appendix B.2 on page 299 we provide specific results for binary and binomial logistic regression.

The multinomial distribution is a powerful distribution commonly used to analyze univariate and multivariate count, percent, and binary data. The multinomial distribution is often used to model discrete vector-valued response data, such as responses to questions

with multiple choice options.

For example, Agresti (2002) gives an example where the multinomial distribution is used to estimate the probability that a fatal transportation accident in Italy will be in an automobile, airplane, or railway. He also describes how multinomial logistic regression can be used to predict what percent of an alligator's diet will be from fish, invertebrates, reptiles, birds, and other animals. The multinomial distribution is also used to model discrete-choice data. A discrete-choice model predicts one of several outcomes based on covariates. Agresti (2002) notes that the multinomial distribution has been used to model choice of transportation to work, choice of brands, whether a person will buy a house, condominium, or rent, and where one will shop. The binary and binomial distributions are perhaps the most common types of multinomial random variables, but they do not embody the full range of analytical possibilities of the multinomial distribution.

Multinomial random variables can be written as vectors of counts or percents. For example, consider transportation deaths in Italy. We let c index categories (automobile, airplane, and railway) and k index the 20 regions in Italy. Let C be the total number of categories and M be the total number of regions. Here, $C = 3$ and $M = 20$. Further, suppose there are z_k transportation deaths in region k. Let $y_{car,k}$ be the total number of automobile deaths, $y_{plane,k}$ be the total number of airplane deaths, and $y_{train,k}$ be the total number of train deaths in region k. The response vector for Lombardy can be written as $\mathbf{y}_{Lombardy} = [y_{car,Lombardy}, y_{plane,Lombardy}, y_{train,Lombardy}]^\top$. Similarly, the response can be written as a percent by replacing y_{ck} with $z_k \mathfrak{p}_{ck}$ where $\mathfrak{p}_{ck} = \frac{y_{ck}}{z_k}$. The measured percent of deaths in the three categories is $\mathfrak{p}_{car,k} = \frac{y_{car,k}}{z_k}$, $\mathfrak{p}_{plane,k} = \frac{y_{plane,k}}{z_k}$, and $\mathfrak{p}_{train,k} = \frac{y_{train,k}}{z_k}$, respectively.

Often z_k is set to be 1 so that $\mathbf{y}_k$ is a random vector with $C-1$ elements equal to 0 and exactly one element equal to 1. For example, if there are five age classifications then $\mathbf{y}_k = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}^\top$ for a person in the youngest age group, $\mathbf{y}_k = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \end{bmatrix}^\top$ for a sample unit in the middle age group, and $\mathbf{y}_k = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \end{bmatrix}^\top$ for a sample unit in the oldest age group. However, if each element of $\mathbf{y}_k$ is a count, the z_k is the sum of all elements in $\mathbf{y}_k$. For example, if $\mathbf{y}_k = \begin{bmatrix} 8 & 3 & 6 & 7 \end{bmatrix}^\top$, then $z_k = 24$.

Examples of categorical data that can be modeled with the multinomial distribution abound. For example, if we categorize the labor force status as: not in the labor force, employed, or unemployed, then we can use multinomial logistic regression to model the respondent's labor force status. Another example would be to model the mode of transportation one takes to work where the options are: car, bike, public transportation, walk, or another mode. Although z_k is frequently 1, we will more generally consider the case where z_k is a positive integer.

Assuming that $\mathbf{y}_k$ is distributed as a multinomial random vector, the probability mass function for $\mathbf{y}_k$ is

$$f\left(\mathbf{y}_k | z_k; \mathfrak{p}_k\right) = \frac{z_k!}{\prod_{c=1}^{C} y_{ck}!} \prod_{c=1}^{C} \mathfrak{p}_{ck}^{y_{ck}} \tag{3.1}$$

Since, $\sum_{c=1}^{C} \mathfrak{p}_{ck} = 1$, one of the categories is redundant. This redundancy leads to estimation problems. As a solution, we employ the baseline categorization by replacing y_{Ck} with $z_k - \sum_{c=1}^{C-1} y_{ck}$. The full rank probability mass function is

$$\begin{aligned} f\left(\mathbf{y}_k | z_k; \mathfrak{p}_k\right) &= \frac{z_k!}{\prod_{c=1}^{C} y_{ck}!} \left(1 - \sum_{c=1}^{C-1} \mathfrak{p}_{ck}\right)^{z_k - \sum_{c=1}^{C-1} y_{ck}} \prod_{c=1}^{C-1} \mathfrak{p}_{ck}^{y_{ck}} \\ &= \frac{z_k!}{\prod_{c=1}^{C} y_{ck}!} \left(1 - \sum_{c=1}^{C-1} \mathfrak{p}_{ck}\right)^{z_k} \prod_{c=1}^{C-1} \left(\frac{\mathfrak{p}_{ck}}{1 - \sum_{c=1}^{C-1} \mathfrak{p}_{ck}}\right)^{y_{ck}}. \end{aligned} \tag{3.2}$$

where $\mathfrak{p}_k$ is the underlying parameter vector for the kth unit. It is clear from Equation (3.2) that we only need to estimate $C-1$ parameters for each unit rather than C parameters. According to Shao (2003, p. 98) this is a member of the full rank exponential dispersion family. We note that $f : \mathbb{R}^C \to \mathbb{R}^1$. That is, f maps a C dimensional response vector to a scalar quantity, the real numbers.

If a sample of size m is selected and the units are independent of each other, then we can write the joint density as,

$$f\left(\mathbf{y}|\mathbf{z};\mathfrak{p}\right) = \prod_{k=1}^{m} f\left(\mathbf{y}_k|z_k;\mathfrak{p}_k\right)$$

For example, if transportation deaths are independent and identically distributed across regions in the Italian example, then the probability mass function for all 20 regions is

$$f\left(\mathbf{y}|\mathbf{z};\mathfrak{p}\right) = \prod_{k=1}^{20}\left[\frac{z_k!}{\prod_{c=1}^{3} y_{ck}}\prod_{c=1}^{3}\mathfrak{p}_{ck}^{y_{ck}}\right].$$

The binomial and Bernoulli distributions are both examples of the multinomial distribution. When $C=2$, then Equation (3.1) reduces to the binomial distribution. Also, when $C=2$ and $z_k=1$, then Equation (3.1) reduces to the Bernoulli distribution. Thus, the multinomial distribution encompasses a variety of common distributions used to analyze categorical data. Appendix B.2 on page 299 reviews Bernoulli and binomial logistic regression.

If covariates are available, we might be able to improve our estimates with a generalized linear model. Agresti (2002) describes the logistic generalized linear model for categorical data. We consider the case where we model all of the categories with the same set of covariates, but allow the coefficients to differ among the sampling units. Fahrmeir and Tutz (2001, p. 79) call this type of design matrix *global* because the covariates do not

depend on each category. Category-specific design matrices can also be constructed, but are not considered in this dissertation. As an example of a global design matrix, suppose we have two covariates in our model, an intercept and a variable with the person's age. Thus $p = 2$. The kth unit's explanatory variable is captured in a vector

$$\mathbf{x}_k = \begin{bmatrix} x_{\text{intercept, k}} \\ \\ x_{\text{age, k}} \end{bmatrix}$$

For a sample of size m, there are m of the $\mathbf{x}_k$ vectors.

The model-based expected value of a multinomial random variable, $\mathbf{Y}_k$ is $\boldsymbol{\mu}_k = E\left(\mathbf{Y}_k\right)$. In the model-based framework, covariates are used to model the expected response, μ_{ck} in the following way

$$E_M\left(Y_{ck}\right) = \mu\left(\mathbf{x}_k, \boldsymbol{\beta}_c\right) = \frac{z_k e^{\mathbf{x}_k^\top \boldsymbol{\beta}_c}}{1 + \sum_{c=1}^{C-1} e^{\mathbf{x}_k^\top \boldsymbol{\beta}_c}}$$

where $\boldsymbol{\beta}_c$ is a superpopulation vector containing parameters for the cth category. In the design-based framework

$$\mu_{ck} = \mu\left(\mathbf{x}_k, \mathbf{B}_c\right) = \frac{z_k e^{\mathbf{x}_k^\top \mathbf{B}_c}}{1 + \sum_{c=1}^{C-1} e^{\mathbf{x}_k^\top \mathbf{B}_c}} \tag{3.3}$$

where $\mathbf{B}_c$ is the p-dimensional finite population parameter vector for the cth category. For our multivariate response vector, we have

$$\underset{C\times 1}{\boldsymbol{\mu}_k} = \boldsymbol{\mu}\left(\mathbf{X}_k, \mathbf{B}\right) = \frac{z_k e^{\mathbf{X}_k \text{vec}(\mathbf{B})}}{1 + \sum_{c=1}^{C-1} e^{\mathbf{x}_k^\top \mathbf{B}_c}} \tag{3.4}$$

where $\mathbf{B}$ is a p by $C-1$ dimensional matrix containing the finite population parameters for all $C-1$ independent categories, the vec function stacks columns of a matrix into one column vector (see Harville (1997), Searle (1982), or Seber (2008)), and

$$\underset{C\times(C-1)\cdot p}{\mathbf{X}_k} = \begin{bmatrix} \mathbf{x}_k^\top & & & \\ & \mathbf{x}_k^\top & & \\ & & \ddots & \\ & & & \mathbf{x}_k^\top \\ 0 & & \cdots & 0 \end{bmatrix} \tag{3.5}$$

$$= \widetilde{\mathbf{I}}_C \bigotimes \mathbf{x}_k^\top$$

and

$$\widetilde{\mathbf{I}} = \begin{bmatrix} \underset{(C-1)\times(C-1)}{\mathbf{I}} \\ \underset{(C-1)\times 1}{\mathbf{0}}^\top \end{bmatrix}.$$

If we do not know $\mathbf{B}$, we can estimate it from a sample. In this case

$$\underset{C\times 1}{\widehat{\boldsymbol{\mu}}_k} = \boldsymbol{\mu}\left(\mathbf{X}_k, \widehat{\mathbf{B}}\right) = \frac{z_k e^{\mathbf{X}_k \mathrm{vec}\left(\widehat{\mathbf{B}}\right)}}{1 + \sum_{c=1}^{C-1} e^{\mathbf{x}_k^\top \widehat{\mathbf{B}}_c}} \tag{3.6}$$

where $\widehat{\mathbf{B}}_c$ is an estimate of $\mathbf{B}_c$ and $\widehat{\mathbf{B}}$ is an estimate of $\mathbf{B}$.

The formulation of $\mathbf{X}_k$ for multinomial logistic regression in Equation (3.5) is unique to this paper. Both Agresti (2002) and Fahrmeir and Tutz (2001) write $\mathbf{X}_k$ without the final row of $\mathbf{0}^\top$. The benefit of including $\mathbf{0}^\top$ in $\mathbf{X}_k$ is that predictions can be made for all C categories, including the baseline category. It is more standard notation to obtain responses for the baseline by defining $\mathbf{B}_C = \mathbf{0}$, but the notation used throughout this text allows us the advantage of writing $\mathbf{B}$ as a full rank matrix that produces estimates for all C

categories. When the $\mathbf{0}^\top$ row is not included in $\mathbf{X}_k$, then either a separate expression must be used to estimate μ_{kC} or $\mathbf{B}$ must be written so that the final column is $\mathbf{0}$. Redefining $\mathbf{B}$ in this way is not desirable because it makes $\mathbf{B}$ less than full rank.

In Appendix B.2.5 on page 309, we show that a sample weighted estimate of the finite population parameter $\mathbf{B}_c$ can be found by iteratively solving the estimating equations

$$\sum_{k \in \mathfrak{s}} d_k \left\{ \left[y_{ck} - \frac{z_k e^{\mathbf{x}_k^\top \mathbf{B}_c}}{1 + \sum_{c=1}^{C-1} e^{\mathbf{x}_k^\top \mathbf{B}_c}} \right] \mathbf{x}_k^\top \right\} = \mathbf{0}.$$

3.1.2 Estimation of Totals for Multinomial Data in Poisson Samples

In this section, we review several point estimators that can be used to estimate totals of multinomimal data. We open with the most basic design-based estimator, the π-estimator. Then, we introduce two generalized difference estimators, the GREG and logistic general regression (LGREG) estimators. We then discuss three different types of calibration estimators. We begin with the traditional calibration estimator and then discuss the model-calibration estimator. We conclude with two model-calibrated maximum pseudoempirical likelihood estimators. Table 3.1 shows the estimators that follow. The following introductions are rather brief and primarily focus on estimating multinomial totals from logistic models. More details about these estimators in general be found in Section 1.2 on page 33.

3.1.2.1 π-Estimator

The π-estimator is design-unbiased and simple to compute. However, the variability of this estimator from sample to sample tends to be larger than competing estimators that

Table 3.1: Point Estimators

Statistic	Description
$\hat{\mathbf{t}}_y^{\pi}$	π-Estimator
$\hat{\mathbf{t}}_y^{gr}$	GREG / Calibration Estimator
$\hat{\mathbf{t}}_y^{lg}$	LGREG Estimator
$\hat{\mathbf{t}}_y^{mc}$	Model-Calibration Estimator
$\hat{\mathbf{t}}_y^{peM}$	Pseudo-Empirical Maximum Likelihood Estimator using M
$\hat{\mathbf{t}}_y^{pe\widehat{M}}$	Pseudo-Empirical Maximum Likelihood Estimator using $\widehat{M}$

make use of covariates, especially for small and moderate-sized samples. Also, compared to other estimators, the π-estimator does not have the calibration property, a very important property for official statistics. Thus, the π-estimator is not preferred to alternative estimators, such as the Generalized Difference Estimator.

3.1.2.2 Generalized Difference Estimator

Wu and Sitter (2001) defined the generalized difference estimator for multivariate responses in single-stage samples as

$$\hat{t}_{yc}^{gd} = \sum_{k \in \mathscr{U}} \widehat{\mu}_{ck} + \sum_{k \in \mathfrak{s}} d_k \left(y_{ck} - \widehat{\mu}_{ck} \right) \tag{3.7}$$

where $\widehat{\mu}_{ck}$ is an estimate of $E_M \left(y_{ck} | \mathbf{x}_k, \widehat{\mathbf{B}}_c \right)$ under some working model. For example, $\widehat{\mu}_{ck}$ could be an estimate from a linear, logistic, or nonparametric model. For single-staged samples, Wu and Sitter (2001) proved that $\hat{t}_{yc}^{gd}$ is a design-consistent estimator for t_{yc} with asymptotic variance

$$\text{av} \left(\hat{t}_{yc}^{gd} \right) = \sum_{k \in \mathscr{U}} \sum_{l \in \mathscr{U}} \Delta_{kl} \left(\frac{y_{ck} - \mu_{ck}}{\pi_k} \right) \left(\frac{y_{lc} - \mu_{lc}}{\pi_l} \right)$$

where $\mu_{ck} = E_M\left(y_{ck}|\mathbf{x}_k, \mathbf{B}_c\right)$. With a sample, the asymptotic variance can be estimated by

$$\upsilon_e\left(\hat{t}_{yc}^{gd}\right) = \sum_{k\in\mathfrak{s}}\sum_{l\in\mathfrak{s}} \frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{y_{ck}-\widehat{\mu}_{ck}}{\pi_k}\right)\left(\frac{y_{lc}-\widehat{\mu}_{lc}}{\pi_l}\right). \tag{3.8}$$

When using a linear working model, $\widehat{\mu}_{ck} = \mathbf{x}_k^\top \widehat{\mathbf{B}}_c$ and Equation (3.7) reduces to the General REGression (GREG) estimator. See Section 1.2.1 on page 33 for an introduction to the GREG estimator.

Although the GREG estimator is approximately design-unbiased, it is motivated by a linear relationship between the response variable and the covariates. Even when the linear model assumptions are violated, as is the case with multinomial data, the GREG estimator is still design-consistent. This property of the GREG makes it a popular estimator for both continuous and categorical data. Moreover, it results in one general set of calibrated weights that can be used for a variety of dependent variables. Lastly, it does not require auxiliary information for the complete frame. It only requires covariates for sample units and control totals for the population. Despite these benefits, there may be more efficient estimators that use more appropriate models when dealing with categorical data.

In 1998, Lehtonen and Veijanen described one way to use an assisting logistic regression model to estimate totals when the response data are characterized by the binary, binomial, or multinomial distributions. Lehtonen and Veijanen (1998) claim that their estimator, called the logistic GREG (LGREG) estimator, is design-consistent and has smaller mean squared error than the GREG estimator in some situations. Wu and Sitter (2001) further proved that the generalized difference estimator is design-consistent under

certain assumptions. In one stage of sampling, the LGREG estimator is equivalent to Equation (3.7) on page 121 with $\widehat{\mu}_{ck}$ defined in Section 3.1.1 on page 114. That is

$$\widehat{\mathbf{t}}_y^{lg} = \sum_{k \in \mathscr{U}} \widehat{\boldsymbol{\mu}}_k + \sum_{k \in \mathfrak{s}} d_k \left(\mathbf{y}_k - \widehat{\boldsymbol{\mu}}_k \right) \tag{3.9}$$

where

$$\widehat{\boldsymbol{\mu}}_k = \frac{z_k e^{\mathbf{X}_k \mathrm{vec}\left(\widehat{\mathbf{B}}\right)}}{1 + \sum_{c=1}^{C-1} e^{\mathbf{x}_k^\top \widehat{\mathbf{B}}_c}}$$

If the estimated logistic regression model coefficients, $\widehat{\mathbf{B}}_c$, are calculated using weighted pseudomaximum likelihood estimating equations, then the LGREG estimator will be a design-consistent estimator of the population total under a variety of sample designs, including multiple stage samples. Since the first summation in Equation (3.9) is over the entire universe, $\mathbf{x}_k$ must be known for all units in the population. For this reason, using the LGREG estimator requires a sampling frame complete with all explanatory variables used in the assisting model for all units in the population. Many address-based sampling frames, business registers, and trade association lists contain a wealth of covariates for all "known" units.

Lehtonen and Veijanen (1998) recommend estimating the variance of $\hat{t}_{yc}^{lg}$ in single-stage samples with

$$\begin{aligned} \upsilon_e \left(\hat{t}_{yc}^{lg} \right) &= \sum_{k \in \mathfrak{s}} \sum_{l \in \mathfrak{s}} \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{e_{ck}}{\pi_k} \right) \left(\frac{e_{cl}}{\pi_l} \right) \\ &= \sum_{k \in \mathfrak{s}} \sum_{l \in \mathfrak{s}} \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{y_{ck} - \mu_{ck}}{\pi_k} \right) \left(\frac{y_{cl} - \mu_{cl}}{\pi_l} \right). \end{aligned}$$

Lehtonen and Veijanen (1998) exclusively focused on single-stage samples. Their variance estimator will generally underestimate the sampling error in clustered samples because it does not account for the correlation between clusters. Moreover, in small samples,

it may poorly estimate the variability of $\hat{t}_{yc}^{lg}$ because it estimates the asymptotic variance of $\hat{t}_{yc}^{lg}$ rather than the exact variance of $\hat{t}_{yc}^{lg}$. The variance estimator proposed by Lehtonen and Veijanen (1998) also requires knowledge of joint inclusion probabilities, which often are impossible to compute or unavailable to data analysts. Of course, if a Poisson sample is selected, Δ_{kl} conveniently reduces to 0 when $k \neq l$ and $\pi_k (1 - \pi_k)$ when $k = l$.

The generalized difference estimator is a broad class of design-consistent estimators that includes both the GREG and LGREG estimators. Generalized difference estimators have many advantages over the π-estimator. Hitherto, the properties of the generalized difference estimator have not been explored for a logistic-assisting model when the sample was selected from a clustered design.

3.1.2.3 Calibrated Estimator

According to Deville and Särndal (1992), calibration estimators

> use calibrated weights, which are as close as possible, according to a given distance measure, to the original sampling design weights d_k while also respecting a set of constraints, the calibration equations.

Typically the calibration equations are formulated so that the weighed sum of auxiliary variables is equal to known population controls, that is, $\sum_{k \in \mathfrak{s}} w_k^{cal} \mathbf{x}_k = \mathbf{t}_x$ where w_k^{cal} is the new calibration weight. The calibration property is especially attractive for official statistical agencies which seek to assure that key demographic estimates are consistent across surveys and equal to "known" population totals. Post-stratification, raking, and the general regression estimators are all examples of calibration estimators.

The primary analytic goal of calibration is to find a new vector of weights, $\mathbf{w}^{cal}$, that is minimal distance from the design weights, $\mathbf{d}$, and meets the constraints $\sum_{k\in\mathfrak{s}} w_k^{cal}\mathbf{x}_k = \mathbf{t}_x$. The calibrated weights depend on how one specifies the "distance" between the design weights and the calibrated weights. For example, Deville and Särndal (1992) show that the GREG estimator is equivalent to calibration with a linear distance function equal to $\sum_{k\in\mathfrak{s}} \frac{\left(w_k^{cal}-d_k\right)^2}{d_k q_k}$ where q_k is chosen by the statistician, often selected to be 1, $\frac{1}{\sigma^2}$, or $\frac{1}{\sigma^2 x_k}$.

Deville and Särndal (1992) also proved that calibration estimators are asymptotically equivalent to the GREG estimator, regardless of how one specifies the "distance." For this reason, Deville and Särndal (1992) suggest approximating the variance of calibrated estimators by simply using the GREG variance estimators.

Särndal (2007) reviewed several extensions of calibration to cluster samples. In cluster samples, the cluster weights, d_i, may be calibrated, the unit weights, d_k, may be calibrated, or both may be calibrated, depending on the available data and the analytic goals. Estevao and Särndal (2006) covered a number of different ways to calibrate data in cluster samples. When complete auxiliary data are available, the calibration estimator is

$$\hat{t}_{yc}^{cal} = \sum_{k\in\mathfrak{s}} w_k^{cal} y_{ck} \tag{3.10}$$

where w_k^{cal} is found by minimizing

$$\sum_{k\in\mathfrak{s}} \frac{\left(w_k^{cal} - d_k\right)^2}{d_k q_k} \tag{3.11}$$

subject to the constraint

$$\sum_{k\in\mathscr{U}} \mathbf{x}_k = \sum_{k\in\mathfrak{s}} w_k^{cal}\mathbf{x}_k.$$

The variance of $\hat{t}_{yc}^{cal}$ can be estimated with

$$\upsilon_e\left(\hat{t}_{yc}^{cal}\right) = \sum_{i\in\mathfrak{s}_I}\sum_{j\in\mathfrak{s}_I}\frac{\Delta_{ij}}{\pi_{ij}}\hat{t}_{eci}^{cal}\hat{t}_{ecj}^{cal} + \sum_{i\in\mathfrak{s}_I}\sum_{k\in\mathfrak{s}_i}\sum_{l\in\mathfrak{s}_i}\frac{\Delta_{kl|i}}{\pi_{kl|i}}w_k^{cal}e_{ck}w_l^{cal}e_{cl}$$

where $\hat{t}_{eci}^{cal} = \sum_{k\in\mathfrak{s}_i} w_k^{cal}e_{ck}$.

As defined by Deville and Särndal (1992), the calibration constraints assure that the weighted auxiliary data equals known control totals. One advantage of this form of calibration is that one set of calibration weights can be created and used for all variables collected. Although calibration estimators are often more efficient than the π-estimator, further gains in efficiency can be made by building more specialized models.

3.1.2.4 Model-Calibrated Estimator

Wu and Sitter (2001) extended calibration to cover nonlinear assisting models. They call their method, model-calibration. Instead of minimizing the distance between d_k and w_k^{cal} subject to $\sum_{k\in\mathfrak{s}} w_k^{cal}\mathbf{x}_k = \mathbf{t}_x$, they proposed minimizing the distance between d_k and w_{ck}^{mc} subject to $\frac{1}{N}\sum_{k\in\mathfrak{s}} w_{ck}^{mc} = 1$ and $\sum_{k\in\mathfrak{s}} w_{ck}^{mc}\widehat{\mu}_{ck} = \sum_{k\in\mathscr{U}}\widehat{\mu}_{ck}$ where $\widehat{\mu}_{ck}$ is a prediction from a generalized linear model. After solving for w_{ck}^{mc}, they estimated a finite population total by $\widehat{t}_{yc}^{mc} = \sum_{k\in\mathfrak{s}} w_{ck}^{mc}y_{ck}$. When the linear distance measure (see Equation (3.11) on page 125) is used, $\widehat{t}_{yc}^{mc}$ can be explicitly written as

$$\widehat{t}_{yc}^{mc} = \widehat{\mathbf{t}}_{yc}^{\pi} + \left(\sum_{k\in\mathscr{U}}\widehat{\mu}_{ck} - \sum_{k\in\mathfrak{s}} d_k\widehat{\mu}_{ck}\right)\widehat{\mathbf{B}}_c^{mc} \tag{3.12}$$

where

$$\widehat{\mathbf{B}}_c^{mc} = \frac{\sum_{k\in\mathfrak{s}} d_k q_k\left(\widehat{\mu}_{ck} - \bar{\mu}_c\right)\left(y_{ck} - \bar{y}_c\right)}{\sum_{k\in\mathfrak{s}} d_k q_k\left(\widehat{\mu}_{ck} - \bar{\mu}_c\right)^2} \tag{3.13}$$

$$\bar{\mu}_c = \frac{\sum_{k\in\mathfrak{s}} d_k q_k\widehat{\mu}_{ck}}{\sum_{k\in\mathfrak{s}} d_k q_k}. \tag{3.14}$$

Wu and Sitter (2001) also found the asymptotic variance of $\widehat{t}_{yc}^{mc}$ in single-staged samples to be

$$\operatorname{av}\left(\widehat{t}_{yc}^{mc}\right) = \sum_{k\in\mathscr{U}}\sum_{l\in\mathscr{U}} \Delta_{kl}\left[\left(\frac{y_{ck}-\widehat{\mu}_{ck}\mathbf{B}_c^{mc}}{\pi_k}\right)\left(\frac{y_{cl}-\widehat{\mu}_{cl}\mathbf{B}_c^{mc}}{\pi_l}\right)\right]^2$$

where

$$\mathbf{B}_c^{mc} = \frac{\sum_{k\in l\in\mathscr{U}} q_k\left(\mu_{ck}-\bar{\mu}_c\right)\left(y_{ck}-\bar{y}_c\right)}{\sum_{k\in l\in\mathscr{U}} q_k\left(\mu_{ck}-\bar{\mu}_c\right)^2}$$

$$\bar{\mu}_c = \frac{1}{N}\sum_{k\in\mathscr{U}}\mu_{ck}.$$

Under Poisson sampling the asymptotic variance simplifies to

$$\operatorname{av}\left(\hat{t}_{yc}^{mc}\right) = \sum_{k\in\mathscr{U}} \pi_k\left(1-\pi_k\right)\left(\frac{y_{ck}-\mu_{ck}\mathbf{B}_c^{mc}}{\pi_k}\right)^2.$$

which can be estimated by

$$\upsilon_e\left(\hat{t}_{yc}^{mc}\right) = \sum_{k\in\mathfrak{s}} \frac{\pi_k\left(1-\pi_k\right)}{\pi_k}\left(\frac{y_{ck}-\widehat{\mu}_{ck}\widehat{\mathbf{B}}_c^{mc}}{\pi_k}\right)^2.$$

One advantage of the model calibrated estimator is that it can improve design-based inference by using nonlinear models. Since logistic regression fits data generated by the multinomial distribution better than linear regression, it seems advantageous to use model-calibration when analyzing multinomial data. This paper extends previous literature by developing model-calibration for multivariate response data in two-staged samples. Previous papers have only dealt with scalar response data in single-staged samples. Kim et al. (2009) discuss nonparametric calibration in cluster samples, but they do not cover multivariate response data nor nonlinear models, such as the logistic model.

Of course there are some disadvantages to model-calibration. First, complete data are needed for all sample and nonsample units. Frames rich in auxiliary data are becoming more popular with address based sampling frames, but such frames are not always available. Second, model-calibration results in a new set of weights for each response variable. In large multipurpose surveys, one set of calibrated weights that can be used for all response variables is preferred. Model calibrated weights are not general and each response variable requires a different set of weights. Finally, even though predictions of μ_{ck} are bounded by 0 and z_k, there is no guarantee $\widehat{t}_{yc}^{mc}$ will be bounded by 0 and $\sum_{k\in\mathscr{U}} z_k$. Thus, some estimates of $\widehat{t}_{yc}^{mc}$ could be negative or larger than possible.

3.1.2.5 Model-Calibrated Maximum Pseudo Empirical Likelihood Estimator

Rao and Wu (2009) review the history and motivation of empirical likelihood methods. The pseudoempirical likelihood approach is motivated by treating y_k as a random variable with density p_k^{pe}. The empirical likelihood of $\mathbf{y}$ is

$$L\left(\mathbf{p}^{pe}\right) = \prod_{k\in\mathscr{U}} p_k^{pe}$$

and the log likelihood is

$$\ell\left(\mathbf{p}^{pe}\right) = \sum_{k\in\mathscr{U}} \log p_k^{pe}.$$

Unless a census is taken, the empirical likelihood must be estimated. Thus the pseudoempirical log likelihood is

$$\hat{\ell}\left(\mathbf{p}^{pe}\right) = \sum_{k \in \mathfrak{s}} d_k \log p_k^{pe}. \tag{3.15}$$

Following the theory of maximum likelihood, the pseudoempirical log likelihood is maximized. Furthermore, constraints are added to improve the efficiency of the estimators. The pseudoempirical log likelihood is maximized subject to

$$\sum_{k \in \mathfrak{s}} p_k^{pe} = 1 \tag{3.16}$$

$$\sum_{k \in \mathfrak{s}} p_k^{pe} \mathbf{u}_k = 0 \tag{3.17}$$

where $\mathbf{u}_k$ is a function of the calibration variables (examples follow). Our restricted optimization problem is to maximize Equation (3.15) subject to Equations (3.16) and (3.17).

Once we estimate p_k^{pe}, we can estimate the mean of our variable with

$$\widehat{\bar{y}}_c^{pe} = \sum_{k \in \mathfrak{s}} \hat{p}_k^{pe} y_{ck}.$$

To date, estimators of totals using the model-calibrated pseudoempirical likelihood method have not been discussed in the literature, although Sitter and Wu (2002) discusses totals of quadratic functions.

Chen and Qin (1993), Zhong and Rao (1996), and Chen and Sitter (1999) discuss maximum pseudoempirical likelihood estimators where $\mathbf{u}_k = \mathbf{x}_k - \bar{\mathbf{x}}$, which reduces to the GREG weights. Wu and Sitter (2001) extend this method for calibration with nonlinear models. Their model-calibration maximum pseudoempirical likelihood method

uses the model-calibrated function, $\mathbf{u}_k = \mu_k - \frac{1}{N}\sum_{k\in\mathscr{U}} \mu_k$ where μ_k is a prediction from a generalized linear model.

Wu and Sitter (2001) showed that $\widehat{\bar{y}}_c^{pe}$ is asymptotically equivalent to $\widehat{\bar{y}}_c^{mc}$. Therefore, the variance of $\widehat{\bar{y}}_c^{pe}$ could be estimated with $\upsilon_e\left(\widehat{\bar{y}}_c^{mc}\right)$, although Wu and Sitter (2001) recommend using the jackknife variance estimator.

One advantage of the model-calibration maximum pseudoempirical likelihood method is that the weights p_k^{pe} are forced to be positive. Like LGREG and model-calibrated estimation, model-calibrated pseudoempirical maximum likelihood estimation requires complete data and every response variable needs a different set of p_k^{pe} adjustments.

3.2 Main Results

In this section, we extend the logistic-assisted estimators to accommodate multivariate response vectors. We also derive variances for our estimators in cluster samples. Using a common asymptotic design-based framework, we show that the logistic general regression estimator is asymptotically unbiased in cluster samples. We also show that the model-calibration maximum pseudoempirical likelihood estimator is asymptotically equivalent to the model-calibration estimator in cluster samples.

3.2.1 Generalized Difference Estimator

In this section, we present results for the multivariate GREG and LGREG estimators in clustered samples. Derivations, proofs, and technical details supporting this section are in Appendices B.3 and B.4.

3.2.1.1 Multivariate GREG

In Appendix B.3 on page 314, we use the calibration technique with a chi-squared distance measure to form a multivariate calibration estimator. To date, calibration estimators have only been studied for scalar responses. As expected, our multivariate estimator has the general form of a GREG estimator. The multivariate GREG estimator is

$$\underset{C\times 1}{\widehat{\boldsymbol{t}}_y^{gr}} = \widehat{\mathbf{t}}_y^{\pi} + \widehat{\mathbf{B}}_{\mathbf{yx}}\left(\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x}}^{\pi}\right) \tag{3.18}$$

where

$$\underset{C\times p}{\widehat{\mathbf{B}}_{\mathbf{yx}}} = \sum_{k\in\mathfrak{s}} \frac{q_k \mathbf{y}_k \mathbf{x}_k^{\top}}{\pi_k} \left(\sum_{k\in\mathfrak{s}} \frac{q_k \mathbf{x}_k \mathbf{x}_k^{\top}}{\pi_k}\right)^{-1}. \tag{3.19}$$

As in the univariate case, we can also write the GREG as

$$\hat{\boldsymbol{t}}_y^{gr} = \mathbf{y}^{\top}\mathbf{\Pi}^{-1}\mathbf{g}$$

where

$$\underset{n\times 1}{\mathbf{g}} = \mathbf{1} + \mathbf{Q}\mathbf{X}_{\mathfrak{s}}\left(\mathbf{X}_{\mathfrak{s}}^{\top}\mathbf{\Pi}^{-1}\mathbf{Q}\mathbf{X}_{\mathfrak{s}}\right)^{-1}\left(\mathbf{t}_{\mathbf{x}\mathscr{U}} - \hat{\mathbf{t}}_{\mathbf{x}\mathfrak{s}}\right).$$

Although Equation (3.18) is more compact than writing the GREG estimator for each category, estimates using Equation (3.18) will be equivalent to estimating each category separately.

3.2.1.2 Multinomial LGREG in Clustered Samples

Assuming a logistic multinomial model, we extend the generalized difference estimator to two-staged samples and explore characteristics of the estimator. The variability

of the generalized difference estimator depends on the fit of the assisting model. If the data are more aptly described by a logistic model than a linear model, the LGREG estimator will be a more efficient estimator than the GREG estimator.

Equation (1.19) on page 42 showed one way to express the generalized difference estimator for a multivariate response obtained from a clustered sample using a linear assisting model. If we now replace the linear assisting model with a multinomial logistic assisting model in Equation (1.19), we have the clustered LGREG estimator

$$\widehat{\mathbf{t}}_y^{lg} = \sum_{\mathscr{U}} \widehat{\boldsymbol{\mu}}_k + \sum_{\mathfrak{s}} d_k \left[\mathbf{y}_k - \widehat{\boldsymbol{\mu}}_k\right] \tag{3.20}$$

where $\widehat{\boldsymbol{\mu}}_k$ is defined in Equation (3.6).

Since $\widehat{\mathbf{t}}_y^{lg}$ is a function of $\widehat{\boldsymbol{\mu}}_k$ and $\widehat{\boldsymbol{\mu}}_k$ is a function of $\widehat{\mathbf{B}}$, one needs to compute $\widehat{\mathbf{B}}$ in order to use $\widehat{\mathbf{t}}_y^{lg}$. In single-stage samples, Lehtonen and Veijanen (1998) suggest using implicit differentiation to estimate $\mathbf{B}_c$. The same general technique can be applied to clustered samples. Specifically, $\mathbf{B}_c$ can be estimated by numerically solving the estimating equations

$$\sum_{i\in\mathfrak{s}_I}\sum_{k\in\mathfrak{s}_i} d_k \left\{ \left[y_{ck} - \frac{z_k e^{\mathbf{x}_k^\top \mathbf{B}_c}}{1+\sum_{c=1}^{C-1} e^{\mathbf{x}_k^\top \mathbf{B}_c}} \right] \mathbf{x}_k^\top \right\} = \mathbf{0}$$

for $\mathbf{B}_c$.

Alternatively, one can simultaneously compute $\widehat{\mathbf{t}}_y^{lg}$ and $\mathrm{vec}\left(\widehat{\mathbf{B}}\right)$ using implicit differentiation. This is accomplished by adding C estimating equations to the estimating equations for $\mathbf{B}$. Let our parameter vector be

$$\underset{[C+(C-1)\cdot p]\times 1}{\boldsymbol{\theta}} = \begin{bmatrix} \mathbf{t}_y^{lg} \\ \mathrm{vec}\left(\mathbf{B}\right) \end{bmatrix}$$

In Appendix B.4.3.3 on page 325, we show that $\boldsymbol{\theta}$ can be estimated using the pseudomaximum likelihood estimating equations

$$\underset{[C+(C-1)\cdot p]\times 1}{\mathbf{W}(\boldsymbol{\theta})} = \begin{bmatrix} \sum_{\mathfrak{s}} d_i d_{k|i} \left(\mathbf{y}_k - \boldsymbol{\mu}_k\right) - \left(\mathbf{t}_y^{lg} - \sum_{\mathscr{U}} \boldsymbol{\mu}_k\right) \\ \sum_{\mathfrak{s}} d_k \mathbf{x}_k \left[y_{1k} - \mu_{1k}\right] \\ \vdots \\ \sum_{\mathfrak{s}} d_k \mathbf{x}_k \left[y_{(C-1)k} - \mu_{(C-1)k}\right] \end{bmatrix}$$

to simultaneously solve for $\mathbf{t}_y^{lg}$ and $\mathrm{vec}\left(\mathbf{B}\right)$. This is done by setting $\mathbf{W}\left(\boldsymbol{\theta}\right) = 0$ and numerically solving for $\boldsymbol{\theta}$.

To determine the asymptotic properties of $\mathbf{t}_y^{lg}$ in cluster samples, we must make some general assumptions which describe our asymptotic framework. Using three assumptions, Wu and Sitter (2001) showed that the LGREG estimator is asymptotically design-unbiased in single-staged samples. Furthermore, under a fourth assumption, Wu and Sitter (2001) calculated the asymptotical variance of the LGREG estimator in single-staged samples. We extend the four assumptions presented in Wu and Sitter (2001) to cluster samples and calculate the asymptotic bias and variance of the LGREG estimator.

First, we assume that our estimated coefficients are consistent estimators of the finite population coefficients. Moreover, we also assume that as the number of clusters increase, the finite population coefficients approach the superpopulation parameters. Technically,

Assumption 4. *As our population and sample sizes increase, our finite population parameter vector,* $\mathbf{B}_N$ *and our weighted estimator,* $\widehat{\mathbf{B}}$ *get closer and closer to a constant, namely our superpopulation parameter,* $\boldsymbol{\beta}$*. That is,* $\widehat{\mathbf{B}} = \mathbf{B}_N + O_p\left(n^{-\frac{1}{2}}\right)$ *and* $\mathbf{B}_N \longrightarrow \boldsymbol{\beta}$

Second, we assume that our LGREG function is smooth, differentiable, and that the LGREG mean function is bounded. That is,

Assumption 5. *For each* $\mathbf{x}_k$, $\frac{\partial}{\partial \mathbf{t}} \boldsymbol{\mu}(\mathbf{x}_k, \mathbf{t})$ *is continuous in* $\mathbf{t}$ *and* $|\frac{\partial}{\partial \mathbf{t}} \boldsymbol{\mu}(\mathbf{x}_k, \mathbf{t})| \leq h(\mathbf{x}_k, \boldsymbol{\theta})$ *for* $\mathbf{t}$ *in a neighborhood of* $\boldsymbol{\theta}$*, and* $N^{-1} \sum_{i=1}^{N} h(\mathbf{x}_k, \boldsymbol{\theta}) = O(1)$*, where* $h(\mathbf{x}_k, \boldsymbol{\theta})$ *is a finite scalar.*

Third, we let our basic design weights be bounded in such a way that means generated using the basic design weights are asymptotically normally distributed.

Assumption 6. *The* π*-estimators for certain population means are asymptotically normally distributed.*

Lastly, to compute the asymptotic variance of the LGREG estimator, we will need to assume that the second derivative of the LGREG function is smooth, continuous, and bounded.

Assumption 7. *For each* $\mathbf{x}_k$, $\frac{\partial^2}{\partial \mathbf{t} \partial \mathbf{t}^\top} \mu(\mathbf{x_k}, \mathbf{t})$ *is continuous in* $\mathbf{t}$ *and* $\max_{k,l} |\frac{\partial^2}{\partial \mathbf{t} \partial \mathbf{t}^\top} \mu(\mathbf{x_k}, \mathbf{t})| \leq h(\mathbf{x}_k, \boldsymbol{\theta})$ *for* $\mathbf{t}$ *in the neighborhood of* $\boldsymbol{\theta}$ *and* $N^{-1} \sum_{k=1}^{N} h(\mathbf{x}_k, \boldsymbol{\theta}) = O(1)$.

Theorem 3.1. *Under Assumptions 4, 5, and 6,* $\hat{t}_{yc}^{lg}$ *is asymptotically design-unbiased for* t_{yc} *in two-staged samples. Furthermore, under Assumption 7, the asymptotic variance of* $\widehat{\mathbf{t}}_{yc}^{lg}$ *is*

$$\mathrm{av}_{II}\left(\hat{\mathbf{t}}_y^{lg}\right) = \sum_{i \in \mathscr{U}_I} \sum_{j \in \mathscr{U}_I} \left(\Delta_{ij} d_i d_j \mathbf{t}_{ei} \mathbf{t}_{ej}^\top\right) + \sum_{i \in \mathscr{U}_I} \left[d_i \left(\sum_{k \in \mathscr{U}_i} \sum_{l \in \mathscr{U}_i} \Delta_{kl|i} d_{k|i} d_{l|i} \mathbf{e}_k \mathbf{e}_l^\top \right) \right] \tag{3.21}$$

where

$$\mathbf{t}_{ei} = \sum_{k \in \mathscr{U}_i} \mathbf{e}_k. \tag{3.22}$$

Furthermore, this asymptotic variance can be estimated by

$$\upsilon_{wr}\left(\mathbf{t}_y^{lg}\right) = \frac{n}{(n-1)} \sum_{i \in \mathfrak{s}_I}^{n} \left(\widehat{\mathbf{t}}_{\widehat{e}i}^{\pi} - \frac{1}{n}\mathbf{t}_{\widehat{\mathbf{e}}}^{\pi}\right)\left(\widehat{\mathbf{t}}_{\widehat{e}i}^{\pi} - \frac{1}{n}\mathbf{t}_{\widehat{\mathbf{e}}}^{\pi}\right)^{\top} \tag{3.23}$$

where

$$\widehat{\mathbf{t}}_{\widehat{e}i}^{\pi} = \sum_{k \in \mathfrak{s}_i} d_k \widehat{\mathbf{e}}_k \tag{3.24}$$

$$\widehat{\mathbf{t}}_{\widehat{e}}^{\pi} = \sum_{k \in \mathfrak{s}} d_k \widehat{\mathbf{e}}_k = \sum_{i \in \mathfrak{s}_I} \widehat{\mathbf{t}}_{\widehat{e}i}^{\pi} \tag{3.25}$$

$$\widehat{\mathbf{e}}_k = \mathbf{y}_k - \widehat{\boldsymbol{\mu}}_k. \tag{3.26}$$

or by,

$$\upsilon_e\left(\hat{\mathbf{t}}_y^{lg}\right) = \sum_{i \in \mathfrak{s}_I}\sum_{j \in \mathfrak{s}_I}\left(\frac{\Delta_{ij}}{\pi_{ij}} d_i d_j \widehat{\mathbf{t}}_{ei}\widehat{\mathbf{t}}_{ej}^{\top}\right) + \sum_{i \in \mathfrak{s}_I}\left[d_i\left(\sum_{k \in \mathfrak{s}_i}\sum_{l \in \mathfrak{s}_i}\frac{\Delta_{kl|i}}{\pi_{kl|i}} d_{k|i} d_{l|i}\widehat{\mathbf{e}}_k\widehat{\mathbf{e}}_l^{\top}\right)\right] \tag{3.27}$$

where

$$\widehat{\mathbf{t}}_{ei} = \sum_{k \in \mathfrak{s}_i} d_{k|i}\widehat{\mathbf{e}}_k \tag{3.28}$$

or by,

$$\upsilon_{Binder}\left(\hat{\mathbf{t}}_y^{lg}\right) = \left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right]\left[\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right)\right]\left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{\top} \tag{3.29}$$

where $\widehat{\mathbf{J}}\left(\widehat{\boldsymbol{\theta}}\right)$ *and* $\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right)$ *are defined in Appendix B.4.3.3. These variance estimators are described in more detail below.*

In Appendix B.4.1 on page 317, we prove that $\widehat{\mathbf{t}}_y^{lg}$ is asymptotically design-unbiased for $\mathbf{t}_y$ in two-staged samples. In Appendix B.4.2 on page 319, we prove that the asymptotic variance of $\hat{t}_{yc}^{lg}$ is

$$\mathrm{av}_{II}\left(\hat{\mathbf{t}}_y^{lg}\right) = \sum_{i\in\mathscr{U}_I}\sum_{j\in\mathscr{U}_I}\left(\Delta_{ij}d_i d_j \mathbf{t}_{ei}\mathbf{t}_{ej}^\top\right) + \sum_{i\in\mathscr{U}_I}\left[d_i\left(\sum_{k\in\mathscr{U}_i}\sum_{l\in\mathscr{U}_i}\Delta_{kl|i}d_{k|i}d_{l|i}\mathbf{e}_k\mathbf{e}_l^\top\right)\right] \quad (3.30)$$

where

$$\mathbf{t}_{ei} = \sum_{k\in\mathscr{U}_i}\mathbf{e}_k. \quad (3.31)$$

Lastly, in Appendix B.4.3 on page 322, we construct υ_{wr}, υ_e, and υ_{Binder}. The with-replacement variance estimator, υ_{wr}, is based on the assumption that the clusters were selected with-replacement. This estimator will usually approximate the variance in without-replacement samples when the fraction of sample clusters to total clusters is small. The classic survey weighted residual variance estimator, υ_e, requires knowledge of joint inclusion probabilities of selection.

When the point estimator can be written in terms of a g-weight, Särndal et al. (1989) use these weights in the variance estimator. Alas, $\widehat{\mathbf{t}}_y^{lg}$ cannot be written as a linear combination involving a g-weight. Thus, we do not propose a g-weighted adjustment to υ_e.

The final variance estimator is the implicit differentiation variance estimator proposed by Binder (1983). The $\mathbf{J}$ matrix on the outsides of this estimator is the jacobian of the estimating equations, $\mathbf{W}(\boldsymbol{\theta})$, with respect to the parameters, $\boldsymbol{\theta}$. That is, $\widehat{\mathbf{J}}(\boldsymbol{\theta}) = \frac{\partial}{\partial(\mathrm{vec}\boldsymbol{\theta})^\top}\widehat{\mathbf{W}}(\boldsymbol{\theta})$. The middle term in this estimator, $\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right)$, is an estimate of the variance of the sample weighted estimating equations. That is $\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right) = av\left(\sum_{\mathfrak{s}}\widehat{\mathbf{U}}_k(\boldsymbol{\theta})\right)$ where $\widehat{\mathbf{U}}_k(\boldsymbol{\theta})$ is the weighted portion of the estimating equations as shown in Appendix

B.4.3 on page 322. In the simulation, we use a with-replacement variance estimator to estimate this variance.

We constructed a logistic regression point estimator for a multinomial response variable selected from clustered samples. In Appendix B.4.1 on page 317 we prove that our estimator is design-consistent for the true finite population total. In Appendix B.4.2 on page 319 we calculate the asymptotic variance of the LGREG estimator. Finally, in Appendix B.4.3 on page 322 we construct three variance estimators of the asymptotic variance. Results from these proofs are summarized in Theorem 3.1 on page 134.

3.2.2 Model-Calibrated Estimator

In this section, we extend the model-calibration estimator with a logistic multinomial model to two-staged samples and explore asymptotic characteristics of the estimator.

Equation (1.31) on page 49 presented the calibration estimator in two-staged samples. If we replace the constraints in the calibration estimator with the multinomial logistic model-calibrated constraints, we obtain a model-calibrated estimator for clustered samples. Doing so gives

$$\underset{C\times 1}{\widehat{\mathbf{t}}_y^{mc}} = \mathbf{y}^\top \mathbf{w}^{mc} \tag{3.32}$$

where $\underset{n\times 1}{\mathbf{w}^{mc}}$ is found by minimizing the chi-squared distance between the design weights and the model-calibration weights,

$$\frac{1}{2}\left(\mathbf{d}-\mathbf{w}^{mc}\right)^\top \mathbf{\Pi}\mathbf{Q}^{-1}\left(\mathbf{d}-\mathbf{w}^{mc}\right) \tag{3.33}$$

subject to the constraint

$$\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\mathbf{w}^{mc} = \underline{\boldsymbol{\mu}}_{\mathscr{U}}^{\top}\mathbf{1} \tag{3.34}$$

where

$$\underline{\boldsymbol{\mu}} = \begin{bmatrix} \mathbf{1} & \boldsymbol{\mu} \end{bmatrix} \tag{3.35}$$

Notice that if the first column of $\underset{n\times(C+1)}{\underline{\boldsymbol{\mu}}_{\mathfrak{s}}}$ is $\mathbf{1}$, then the following constraint is also obtained.

$$\mathbf{1}^{\top}\mathbf{w}^{mc} = N. \tag{3.36}$$

Our restricted objective function is

$$\phi = \frac{1}{2}\left(\mathbf{d} - \mathbf{w}^{mc}\right)^{\top}\boldsymbol{\Pi}\mathbf{Q}^{-1}\left(\mathbf{d} - \mathbf{w}^{mc}\right) - \boldsymbol{\lambda}^{\top}\left(\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\mathbf{w}^{mc} - \underline{\boldsymbol{\mu}}_{\mathscr{U}}^{\top}\mathbf{1}\right)$$

where $\boldsymbol{\lambda}$ is a $C+1$ by 1 vector of Lagrange multipliers. We show in Appendix B.5 on page 336 that minimizing this equation gives us the model-calibrated estimator for two-staged samples

$$\widehat{\mathbf{t}}_{y}^{mc} = \mathbf{y}^{\top}\mathbf{w}^{mc} \tag{3.37}$$

$$= \hat{\boldsymbol{t}}_{y} + \widehat{\mathbf{B}}_{\mathbf{y}\underline{\boldsymbol{\mu}}}\left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^{\top}\mathbf{1} - \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\mathbf{d}\right) \tag{3.38}$$

where

$$\underset{C\times(C+1)}{\widehat{\mathbf{B}}_{\mathbf{y}\underline{\boldsymbol{\mu}}}} = \mathbf{y}^{\top}\boldsymbol{\Pi}^{-1}\mathbf{Q}\underline{\boldsymbol{\mu}}_{\mathfrak{s}}\left(\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\boldsymbol{\Pi}^{-1}\mathbf{Q}\underline{\boldsymbol{\mu}}_{\mathfrak{s}}\right)^{-1}.$$

We can also write our estimator as

$$\widehat{\mathbf{t}}_{y}^{mc} = \mathbf{y}^{\top}\boldsymbol{\Pi}^{-1}\mathbf{g} \tag{3.39}$$

where

$$\underset{n\times 1}{\mathbf{g}} = \mathbf{1} + \mathbf{Q}^\top \underline{\boldsymbol{\mu}}_{\mathfrak{s}} \left(\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \boldsymbol{\Pi}^{-1} \mathbf{Q} \underline{\boldsymbol{\mu}}_{\mathfrak{s}} \right)^{-1} \left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top \mathbf{1} - \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \mathbf{d} \right).$$

Details of this minimization are in Appendix B.5.1 on page 336. Alternative forms of $\widehat{\mathbf{t}}_y^{mc}$ are in Appendix B.5.2 on page 338. Although all elements of $\mathbf{y}^\top$ and $\boldsymbol{\Pi}^{-1}$ are nonnegative, $\mathbf{g}$ could be negative, especially when $\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \mathbf{d}$ is larger than $\underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top \mathbf{1}$. When $\mathbf{g}$ is negative, negative estimates of $\widehat{\mathbf{t}}_y^{mc}$ are possible, but undesirable.

Since $\widehat{\mathbf{t}}_y^{mc}$ is a nonlinear function of sample inclusion indicators, the exact variance of $\hat{\mathbf{t}}_y^{mc}$ cannot be determined. However, under our asymptotic framework, we can compute the asymptotic variance of $\hat{\mathbf{t}}_y^{mc}$. Furthermore, we can construct variance estimators of this asymptotic variance. The following theorem reports the asymptotic variance of $\hat{\mathbf{t}}_y^{mc}$ and presents three estimators of this asymptotic variance.

Theorem 3.2. *Under Assumption 4, the asymptotic variance of $\hat{t}_y^{mc}$ is*

$$\operatorname{av}\left(\hat{\mathbf{t}}_y^{mc}\right) = \sum_{i\in\mathscr{U}_I} \sum_{j\in\mathscr{U}_I} \left(\Delta_{ij} d_i d_j \mathbf{t}_{ei} \mathbf{t}_{ej}^\top\right) + \sum_{i\in\mathscr{U}_I} \left[d_i \left(\sum_{k\in\mathscr{U}_i} \sum_{l\in\mathscr{U}_i} \Delta_{kl|i} d_{k|i} d_{l|i} \mathbf{e}_k \mathbf{e}_l^\top \right) \right] \tag{3.40}$$

where

$$\mathbf{t}_{ei} = \sum_{k\in\mathscr{U}_i} \mathbf{e}_k. \tag{3.41}$$

The asymptotic variance of $\widehat{\mathbf{t}}_y^{mc}$ can be estimated by

$$\upsilon_g\left(\hat{\mathbf{t}}_y^{mc}\right) = \sum_{i\in\mathfrak{s}_I} \sum_{j\in\mathfrak{s}_I} \left(d_{ij} \Delta_{ij} d_i d_j \hat{\mathbf{t}}_{g\hat{\mathbf{e}}i} \hat{\mathbf{t}}_{g\hat{\mathbf{e}}j}^\top\right) + \sum_{i\in\mathfrak{s}_I} \left[d_i \left(\sum_{k\in\mathfrak{s}_i} \sum_{l\in\mathfrak{s}_i} d_{kl|i} \Delta_{kl|i} d_{k|i} d_{l|i} g_k \hat{\mathbf{e}}_k g_l \hat{\mathbf{e}}_l^\top \right) \right] \tag{3.42}$$

where $\hat{t}_{g\hat{\mathbf{e}}i} = \sum_{\mathfrak{s}_i} \frac{g_k \hat{\mathbf{e}}_k}{\pi_{k|i}}$ *or by*

$$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right) = \frac{n}{(n-1)} \sum_{i \in \mathfrak{s}_I} \left[d_i \sum_{k \in \mathfrak{s}_i} \left(d_{k|i}\hat{\mathbf{e}}_{k|i}\right) - \frac{1}{n} \sum_{k \in \mathfrak{s}} \left(d_k \hat{\mathbf{e}}_k\right) \right] \left[d_i \sum_{k \in \mathfrak{s}_i} \left(d_{k|i}\hat{\mathbf{e}}_{k|i}\right) - \frac{1}{n} \sum_{k \in \mathfrak{s}} \left(d_k \hat{\mathbf{e}}_k\right) \right]^\top \tag{3.43}$$

where $\hat{t}_{e,i} = \sum_{\mathfrak{s}_I} \frac{e_k}{\pi_{k|i}}$*, or by*

$$\upsilon_{Binder}\left(\widehat{\boldsymbol{\theta}}\right) = \left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right] \left[\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right)\right] \left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right]^\top \tag{3.44}$$

where $\widehat{\mathbf{J}}\left(\widehat{\boldsymbol{\theta}}\right)$ *and* $\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right)$ *are defined in Appendix B.5.5.3 on page 348.*

See Appendix B.5 on page 336 for the proof of Theorem 3.2.

The first variance estimator, $\upsilon_g\left(\hat{\mathbf{t}}_y^{mc}\right)$ is the standard weighted residual variance estimator with a g-weight adjustment. In Appendix B.5 on page 336, we also develop the weighted residual variance estimator without the g-weighted adjustment, but do not report results here because Särndal et al. (1989) showed that in general the g-weighted variance estimator had better properties than the estimator without the g-weights. The second estimator is the classic with-replacement variance estimator adjusted for the model-calibration estimator. When the fraction of sample clusters to total clusters is small, the with-replacement variance estimator usually comes close to the variance in without-replacement samples. The clear advantage of the with-replacement variance estimator is its simplicity. The final variance estimator is the implicit differentiation variance estimator proposed by Binder (1983). The middle term in this estimator $\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right)$ is an estimate of the variance of the sample weighted estimating equations. In the simulation, we use a with-replacement variance estimator to estimate this variance.

In summary, we constructed a model-calibrated point estimator for a multinomial response variable selected from clustered samples. Our estimator is asymptotically unbiased and design-consistent. We also calculated the asymptotic variance of the model-calibration estimator and constructed three variance estimators of the asymptotic variance.

3.2.3 Model-Calibrated Maximum Pseudo Empirical Likelihood Estimator

Assuming a logistic multinomial model, we extend the pseudoempirical calibration estimator to two-staged samples and explore asymptotic characteristics of the estimator.

Equation (1.36) on page 53 shows the pseudoempirical calibration estimator of a mean in single-stage samples. Total estimates can be constructed with

$$\hat{t}_{yc}^{pe,N} = N \sum_{i \in \mathfrak{s}} p_k^{pe} y_{ci}, \tag{3.45}$$

or

$$\hat{t}_{yc}^{pe,\widehat{N}} = \widehat{N} \sum_{i \in \mathfrak{s}} p_i^{pe} y_{ci}, \tag{3.46}$$

where

$$\widehat{N} = \sum_{i \in \mathfrak{s}} d_i.$$

When complete auxiliary data are available, N is known and $\hat{t}_{yc}^{pe,N}$ will be the preferable estimator. Since there may be considerable sampling error in estimating $\widehat{N}$, $\hat{t}_{yc}^{pe,\widehat{N}}$ may be significantly more variable than $\hat{t}_{yc}^{pe,N}$. One exception is for sample designs where $\widehat{N} = N$, in which case the two estimators are equivalent.

Modifying these estimators for clustered samples yields the new pseudoempirical likelihood calibration estimators

$$\hat{t}_{yc}^{peM} = M \sum_{i \in \mathfrak{s}_I} \sum_{k \in \mathfrak{s}_i} p_{ik}^{pe} y_{cik}, \tag{3.47}$$

or

$$\hat{t}_{yc}^{pe\widehat{M}} = \widehat{M} \sum_{i \in \mathfrak{s}_I} \sum_{k \in \mathfrak{s}_i} p_{ik}^{pe} y_{cik}, \tag{3.48}$$

where

$$\widehat{M} = \sum_{i \in \mathfrak{s}_I} \sum_{k \in \mathfrak{s}_i} d_k$$

and p_{ik}^{pe} is found by maximizing

$$\sum_{\mathfrak{s}_I} \sum_{\mathfrak{s}_i} d_{ik} \log\left(p_{ik}^{pe}\right) \tag{3.49}$$

subject to

$$\sum_{\mathfrak{s}_I} \sum_{\mathfrak{s}_i} p_{ik}^{pe} = 1 \tag{3.50}$$

$$\sum_{\mathfrak{s}_I} \sum_{\mathfrak{s}_i} p_{ik}^{pe} \mathbf{u}_{ik} = 0 \tag{3.51}$$

with

$$\mathbf{u}_{ik} = \boldsymbol{\mu}_{ik} - \frac{1}{M} \sum_{k \in \mathscr{U}} \boldsymbol{\mu}_{ik} \tag{3.52}$$

and $\boldsymbol{\mu}_{ik}$ is the two-staged version of (3.4).

Our restricted optimization problem is to maximize

$$\sum_{\mathfrak{s}_I} \sum_{\mathfrak{s}_{Ii}} d_{ik} \log\left(p_{ik}^{pe}\right) - \lambda_1 \left(\sum_{\mathfrak{s}_I} \sum_{\mathfrak{s}_{Ii}} p_{ik}^{pe} - 1 \right) - \lambda_2 \left(\sum_{\mathfrak{s}_I} \sum_{\mathfrak{s}_{Ii}} p_{ik}^{pe} \mathbf{u}_{ik} \right)$$

Where λ_1 and λ_2 are the Lagrange multipliers. Unlike the model-calibration estimator, we cannot write the solution to this estimator explicitly. Numerical solutions are needed.

Theorem 3.3. *Under Assumptions 4, 5, 17, 18 and 19, $\widehat{\mathbf{t}}_y^{peM}$ is asymptotically design-unbiased for $\mathbf{t}_y$ in two-staged samples. Furthermore, the asymptotic variance of $\widehat{\mathbf{t}}_y^{peM}$ is equivalent to the asymptotic variance of $\widehat{\mathbf{t}}_y^{mc}$ and can be estimated with the variance estimators for $\widehat{\mathbf{t}}_y^{mc}$.*

See Appendix B.6.2 on page 365 for the proof of Theorem 3.3.

In our simulation, we compare both $\hat{t}_{yc}^{peM}$ and $\hat{t}_{yc}^{pe\hat{M}}$; although do not see any advantages of $\hat{t}_{yc}^{pe\hat{M}}$ over $\hat{t}_{yc}^{peM}$.

In Appendix B.6 on page 361 we maximize the pseudoempirical likelihood subject to our model calibration constraints to create the model-calibrated maximum pseudoempirical likelihood estimator. We also prove that $\widehat{t}_{yc}^{peM}$ is asymptotically equivalent to $\widehat{t}_{yc}^{mc}$.

3.3 Simulation

We performed several simulation studies to compare the design-based properties of the three new types of logistic-assisted estimators in two-staged samples. We selected both small and large samples from three sampling frames, one generated from a multinomial logistic model and two representing fairly realistic situations where the data are difficult to model. From each sampling frame, we repeatedly selected six two-staged samples.

Fixed SRS In the first set of samples, we selected a fixed number of clusters. Then, we selected a fixed number of units within each sample cluster. For selecting the

clusters and units, we used a simple random sample without-replacement algorithm. We call this method Fixed SRS because in both stages of sampling, we selected a fixed number of units. Because our cluster sizes varied from cluster to cluster, this design resulted in unequal weights. The second sample design was the same as the first, with the exception that the number of sample clusters selected was larger.

Rate SRS In the third and fourth set of samples, we selected a fixed number of clusters, but selected units in sample clusters at a constant rate. This design resulted in random sample sizes, but all sample units had the same base weight. We call this sample design Rate SRS because units within sample clusters were selected at a constant rate. The third and fourth sample designs differed in the number of clusters selected.

Fixed PPS Finally, in the fifth and sixth set of samples, a sample of clusters was selected with probabilities proportional to the number of units in the cluster. Then a fixed number of units in each sample cluster was selected using a simple random sample without-replacement algorithm. This method resulted in a fixed sample size and equal weights. The fifth and sixth sample designs differed in the number of clusters selected. The Fixed PPS sample design is common in area frame sampling.

For each sample, we estimated the total of our multivariate response vector using the estimators in Table 3.1 on page 121. We repeated this process for thousands of samples. Then, we calculated the empirical bias, empirical variance, and the empirical relative root mean squared error of the estimators in Table 3.1. For each sample, we also estimated the asymptotic variance and confidence interval coverage of $\widehat{\mathbf{t}}_y^{lg}$ and $\widehat{\mathbf{t}}_y^{mc}$ using the variance

estimators in Table 3.2.

Table 3.2: Variance Estimators Calculated in Simulations

Statistic	Description
$\upsilon_{wr}\left(\hat{\mathbf{t}}_y^{lg}\right)$	With Replacement Variance Estimator of $\hat{\mathbf{t}}_y^{lg}$
$\upsilon_{e}\left(\hat{\mathbf{t}}_y^{lg}\right)$	Without Replacement Variance Estimator of $\hat{\mathbf{t}}_y^{lg}$
$\upsilon_{Binder}\left(\hat{\mathbf{t}}_y^{lg}\right)$	Binder's Variance Estimator of $\hat{\mathbf{t}}_y^{lg}$
$\upsilon_{wr}\left(\hat{\mathbf{t}}_y^{mc}\right)$	With Replacement Variance Estimator of $\hat{\mathbf{t}}_y^{mc}$
$\upsilon_{e}\left(\hat{\mathbf{t}}_y^{mc}\right)$	Without Replacement Variance Estimator of $\hat{\mathbf{t}}_y^{mc}$
$\upsilon_{g}\left(\hat{\mathbf{t}}_y^{mc}\right)$	g-weighted Without Replacement Variance Estimator of $\hat{\mathbf{t}}_y^{mc}$
$\upsilon_{Binder}\left(\hat{\mathbf{t}}_y^{mc}\right)$	Binder's Variance Estimator of $\hat{\mathbf{t}}_y^{mc}$

We included $\upsilon_{e}\left(\hat{\mathbf{t}}_y^{mc}\right)$ in our simulations even though Särndal et al. (1989) clearly advocated using the g-weighted variance estimator. We include it for comparison purposes, even though we expect $\upsilon_{g}\left(\hat{\mathbf{t}}_y^{mc}\right)$ to perform better than $\upsilon_{e}\left(\hat{\mathbf{t}}_y^{mc}\right)$ in most cases.

3.3.1 Populations

3.3.1.1 Synthetic Population

For the first set of simulations, we generated a clustered population of multinomial random variables.

First we generated $N = 30{,}000$ clusters of size $M_i = 11 + \lambda_i$ where λ_i is a random draw from an exponential distribution with parameter 0.25. To assure that M_i was an integer, we rounded λ_i to the nearest whole number. Overall, the pseudo population contained $M = 450{,}265$ units. On average, each cluster contained about 15 units.

Next, we generated our auxiliary variable using a hierarchial process to simulate a clustering effect. For each unit, we created an auxiliary variable using the model $x_k = \delta_i + \varepsilon_k$ where δ_i was a draw for the ith cluster from the standard normal distribution

and ε_k was a draw from a normal distribution with mean of 0 and a standard deviation of 0.1. Using this model, we assured that all units within the same cluster had the same superpopulation mean, but different superpopulation means with units from other clusters.

Table 3.3: Quartiles for Synthetic Population

Variable	Minimum	First Quartile	Median	Mean	Third Quartile	Maximum	Total
y_1	0	9	28	53	71	1,082	23,807,066
y_2	0	3	8	13	18	229	6,012,970
y_3	0	4	18	44	56	1,099	19,745,614
z	10	39	79	110	149	1,275	49,565,650
x	-4.15	-0.68	0	0	0.68	4.31	-1,670
Units Per Cluster	11	12	14	15	17	49	450,265

Using the auxiliary variable, x_k we generated random response variables. First, we created a random number, z_k. We set $z_k = 10 + \lambda_k$ where λ_k was a draw from an exponential distribution with parameter of 0.01. To assure that z_k was an integer, we rounded λ_k to the nearest whole number. For each of the 450,265 units, we created a vector of length 3 containing the superpopulation parameters, π_{1k}, π_{2k}, and π_{3k}. The probabilities generating the multinomial random vector were set to be: $\pi_{1k} = \frac{e^{.5+3x_k}}{1+e^{.5+3x_k}+e^{-.5+2x_k}}$, $\pi_{2k} = \frac{e^{-.5+2x_k}}{1+e^{.5+3x_k}+e^{-.5+2x_k}}$, and $\pi_{3k} = 1 - (\pi_{1k} + \pi_{2k})$. Using these parameters, we generated a random vector of length 3 using the `rmultinomial()` function from the mc2d package in R (Pouillot and Delignette-Muller 2010). The sum of the three random elements was set to be z_k. Table 3.3 shows summary statistics for the ideal population.

3.3.1.2 Postsecondary Majors Population

The second sampling frame was derived from the Integrated Postsecondary Education Data System (IPEDS) [1]. This system contains survey and census data for over 7,000 postsecondary educational institutions in the United States.

[1] See nces.ed.gov/ipeds/datacenter/Default.aspx.

We started by downloading the 2009 Completions Dataset (C2009 A). This dataset contained the total number of degrees conferred upon graduating students in 2009 by major field of study. Since there were scores of major categories, we collapsed the majors into four broad categories:

- mathematics (Major series starting with 27),
- health (Major series starting with 51),
- business (Major series starting with 52), and
- all remaining series.

Overall, there were 6,912 institutions conferring degrees in 2009. There were 16,560 mathematics degrees awarded, 783,008 health degrees given, 727,290 business degrees earned, and 2,698,440 other degrees conferred.

The 2009 Completions Dataset was then merged with two institutional characteristics datasets (hd2009 and ic2009) and an enrollments dataset (efest2009) to get auxiliary data about all postsecondary institutions. We used these auxiliary variables to edit the sampling frame. Specifically, we removed all institutions where the number of students enrolled (TOTENRL) was greater than 25,000. We also removed institutions with missing values in the institutional size (INSTSIZE), less than one year certificate indicator (LEVEL1), or type of board controlling the institution (CONTROL) variables. We further removed all institutions in congressional districts that had fewer than 6 postsecondary institutions or more than 50 institutions. The resulting dataset contained 6,354 institutions giving 16,560 math degrees, 649,978 health degrees, 500,444 business degrees, and 1,883,457 other degrees. We defined a cluster as a congressional district. After all editing the final population contained 6,354 units in 406 clusters.

All types of degrees were considered, including associate, undergraduate, and graduate degrees. Furthermore, students who graduated with multiple majors or minors were counted several times, once for each major and once for each minor. Thus a student who majored in math and history and had a minor in education would be counted three times, once as a math graduate and twice for the two other fields of study.

Table 3.4: Quartiles for Postsecondary Population

Variable	Minimum	First Quartile	Median	Mean	Third Quartile	Maximum	Total
Math	0	0	0	3	0	354	16,560
Health	0	0	27	102	131	2,861	649,978
Business	0	0	5	79	66	7,938	500,444
Other	0	22	86	296	316	7,841	1,883,457
z	1	61	188	480	547	9,697	3,050,986
Total Enrollment	1	138	587	2,322	2,401	24,919	14,753,916
Level 1	0	0	0	0.49	1	1	3,119
Control	0	0	0	0.29	1	1	1,820
Units Per Cluster	7	11	15	16	19	38	6,354

Our assisting model contained an intercept and three variables: TOTENRL, LEVEL1, and CONTROL. The IPEDS Enrollment Dataset contained the TOTENRL variable. This variable was collected from each postsecondary institution and contains an early estimate of the institution's fall enrollment for full and part time students. The LEVEL1 variable indicates if the postsecondary institution grants postsecondary awards, certificates, or diplomas for less than one academic year of study. Overall, 48 percent of postsecondary institutions responded as offering such certificates. The CONTROL variable indicates how the postsecondary institution is governed: Public (CONTROL = 1), Private not-for profit (CONTROL = 2), or Private for-profit (CONTROL = 3). Of the 7,316 postsecondary institutions, 2,148 were public, 1,952 were private not-for-profit, 3,176 were private for-profit, and 40 were not applicable. More information on all three of these variables can be found on the IPEDS website.

The motivating sample design is to select a sample of congressional districts in the first stage of sampling. Then, within selected clusters, a sample of postsecondary institutions is sampled. A survey is then conducted in the sample institutions to collect the total number of degrees that are awarded within the four fields: math, health, business, and other. We assume that the frame has rudimentary auxiliary information about all institutions, such as the total number of students enrolled, whether a one year certificate is awarded, and the type of board controlling the institution. Our goal is to estimate the total number students graduating with majors in the four fields.

The number of degrees awarded in various fields has a major impact on the national economy. The demand for skilled scientists, health care professionals, and other technical jobs has grown considerably and the number of students graduating with degrees in these fields impacts the future of these professions and the nation's ability to provide necessary services. Table 3.4 shows summary statistics for the postsecondary population.

3.3.1.3 Census Population

The final sampling frame was derived from Census 2000 data. We downloaded Census 2000 housing unit and population data from Summary File 3 for California, Florida, and New York from the US Census Bureau's website. We then subset the data to block groups with at least one occupied housing unit and one person. Furthermore, all "orphan" counties, counties containing only one valid block group, were removed. We obtained the following variables: Total Number of Occupied Housing Units in the block group (H007001), Total Number of Housing Units owned in the block (H007002), Total Num-

ber of Housing Units being rented in the block (H007003), and the Percent of persons living at or below the poverty line [(P088002 + P088003 + P088004) / P088001].

To assure that we had no certainty clusters for the PPS samples, we divided clusters with over 600 block groups. The splitting was done so that the first 600 block groups were considered the first cluster while each subsequent set of 600 block groups was considered a new cluster. The final cluster for the county had the remaining block groups. Furthermore, to assure that each within cluster sample would not have any certainties, counties with fewer than 9 block groups were removed.

We used the dataset to estimate the total number of rental housing units in California, Florida, and New York.

The motivating sample design is to select counties in the first stage of sampling. Then, within sample counties, a set of block groups is selected. The block group is treated as the ultimate sampling unit. A survey is then conducted within the sample block groups to determine the total number of rental units in each sample block group.

Overall, this population had 214 primary sampling units (counties) and 44,018 ultimate sampling units (block groups). There were $9,356,962$ renter occupied housing units and $13,437,067$ owner occupied housing units in the three states. Table 3.5 shows descriptive statistics about this population.

Table 3.5: Quartiles for Census Population

Variable	Minimum	First Quartile	Median	Mean	Third Quartile	Maximum	Total
Renter (y_1)	0	61	136	213	281	6,343	9,356,962
Owner (y_2)	0	145	250	305	391	4,960	13,437,067
z	1	310	434	518	623	11,130	22,794,019
Poverty Rate (x)	0	0.05	0.10	0.14	0.20	1	6,260
Units Per Cluster	10	49	102	206	346	600	44,018

We used a binomial logistic regression assisting model to estimate the total number

of rental units in California, Florida, and New York. The assisting model contained an intercept and the percent of persons living at or below the poverty line in each block group.

3.3.2 Simulation Design

3.3.2.1 Sample Design

From each of the three sampling frames described, we selected simple random sample without-replacement (SRSWOR) and πps samples of clusters. Within each cluster, we selected a sample of units. We then estimated the total of the response variables using the estimators in Table 3.1 on page 121.

We used the `UPrandomsystematic()` and `UPpoisson()` functions in the `sampling` package of R to select all the samples (Tillé and Matei 2009). To select a systematic random sample, the `UPrandomsystematic()` function sorts the population into a random order and then selects a sample with probabilities proportional to a size measure. This function selects without-replacement samples to achieve a fixed sample size. We used the `UPrandomsystematic()` function to select both stages of the Fixed SRS and Fixed PPS samples. We also used it to select the first stage of the Rate SRS samples. The `UPpoisson()` function selects a Poisson sample and was used to select the second stage of the Rate SRS samples.

We tested how the estimators performed under the three realistic sample designs described at the beginning of Section 3.3 on page 143.

In the Synthetic population, we selected samples of 20 and 1,500 clusters. From

each cluster, a sample of nine units were selected. In the postsecondary population, samples of 10 and 50 clusters were selected. From each sample cluster, four units were randomly selected. From the Census population, we selected either 5 or 50 clusters and about nine units in each cluster. Table 3.6 summarizes the different designs used to select the samples.

Table 3.6: Simulation Design

	Simulation	First Stage Sample	n	Second Stage Sample	Number of Samples
1	Synthetic	srswor	20	$m_i = 2$	$2,000$
2	Synthetic	srswor	1,500	$m_i = 2$	$2,000$
3	Synthetic	srswor	20	$f_i = \frac{60,000}{195,164}$	$2,000$
4	Synthetic	srswor	1,500	$f_i = \frac{60,000}{195,164}$	$2,000$
5	Synthetic	ppswor	20	$m_i = 2$	$2,000$
6	Synthetic	ppswor	1,500	$m_i = 2$	$2,000$
7	Postsecondary Majors	srswor	10	$m_i = 4$	$10,000$
8	Postsecondary Majors	srswor	50	$m_i = 4$	$10,000$
9	Postsecondary Majors	srswor	10	$f_i = \frac{675}{2,427}$	$10,000$
10	Postsecondary Majors	srswor	50	$f_i = \frac{675}{2,427}$	$10,000$
11	Postsecondary Majors	ppswor	10	$m_i = 4$	$10,000$
12	Postsecondary Majors	ppswor	50	$m_i = 4$	$10,000$
13	Census Population	srswor	5	$m_i = 9$	$5,000$
14	Census Population	srswor	50	$m_i = 9$	$5,000$
15	Census Population	srswor	5	$f_i = \frac{30,430}{194,329}$	$5,000$
16	Census Population	srswor	50	$f_i = \frac{30,430}{194,329}$	$5,000$
17	Census Population	ppswor	5	$m_i = 9$	$5,000$
18	Census Population	ppswor	50	$m_i = 9$	$5,000$

3.3.2.2 Number of Samples

Our point and variance estimators varied from sample to sample. To summarize our simulations, we created means and variances of our estimators. For example, consider estimator $\widehat{\theta}_\nu$ from sample ν. The average of our $\widehat{\theta}_\nu$ estimators across all $\aleph$ samples is $\overline{\widehat{\theta}} = \frac{1}{\aleph}\sum_{\nu=1}^{\aleph}\widehat{\theta}_\nu$ and an estimate of the standard error of this mean is $se\left(\widehat{\theta}\right) = \frac{1}{\sqrt{\aleph}}\sqrt{\frac{1}{\aleph-1}\sum_{\nu=1}^{\aleph}\left(\widehat{\theta}_\nu - \overline{\widehat{\theta}}\right)^2}$. This standard error, $se\left(\widehat{\theta}\right)$, is called the simulation

error. Notice that the simulation error is different from the empirical standard deviation of the estimated total, $se\left(\widehat{\theta}\right) = \sqrt{\frac{1}{\aleph-1}\sum_{\nu=1}^{\aleph}\left(\widehat{\theta}_{\nu} - \overline{\widehat{\theta}}\right)^2}$, which does not depend on $\aleph$ in the denominator. Clearly, the more samples we select, the more confidence we will have in the mean of the totals and the standard error of the totals.

We calculated the number of samples needed for the average of the GREG estimator across the repeated samples to have a coefficient of variation (CV) of 0.005 in the Synthetic population, 0.0061 in the Post-secondary population, and 0.007 in the Census population. To simplify our calculations, we assumed samples were selected using simple random sampling with-replacement. That is, we calculated the number of samples to be

$$\text{Number of Samples} = \frac{\frac{\left[se\left(\widehat{\theta}\right)\right]^2}{\bar{\hat{t}}^2}}{CV_0^2}$$

where CV_0 is the target coefficient of variation, $\hat{\sigma}^2$ is an empirical estimate of the standard deviation of the GREG estimator obtained from 100 samples and $\bar{\hat{t}}$ is the average of the GREG estimator for the 100 samples. The number of samples needed was rounded up to the nearest thousand.

In the synthetic population, the maximum ratio of $\hat{\sigma}$ to $\bar{\hat{t}}$ was 0.22, which means $1,922$ samples were needed to achieve a CV of 0.005. As we see from Table 3.6 on page 152, we conservatively selected $2,000$ samples from this population. In the post-secondary population, the maximum ratio of $\hat{\sigma}$ to $\bar{\hat{t}}$ was 0.61, which means $9,887$ samples were needed to achieve a CV of 0.0061. As we see from Table 3.6, we selected $10,000$ samples from this population. In the census population, the maximum ratio of $\hat{\sigma}$ to $\bar{\hat{t}}$ was 0.49, which means $4,903$ samples were needed to achieve a CV of 0.007. As we see from Table 3.6, we selected $5,000$ samples from this population to be on the safe side.

3.3.2.3 Estimation

We estimated the total of each response variable using the estimators in Table 3.1 on page 121. We repeated this process for all samples.

Predictions from a linear model play a key part of the GREG estimator. We used the `lm()` function in R with a weights option to predict the fitted values which we used in the GREG estimation. Each category was independently estimated, so there was no assurance that the sum of the response variables would equal a fixed constant. Our linear model contained the same covariates as the logistic models.

The remaining estimators required predicting $\boldsymbol{\mu}_k$. To calculate $\boldsymbol{\mu}_k$, we first estimated $\mathbf{B}$, the parameters obtained from running a multinomial logistic regression model on the full population. We used the pseudomaximum likelihood method to estimate $\mathbf{B}$ (see Binder (1983)). We first estimated $\boldsymbol{\beta}$, the superpopulation parameter associated with the assisting model, using the `vglm()` function in the VGAM package of R (Yee 2012). Then, we used the value of $\widehat{\boldsymbol{\beta}}$ as a starting point to minimize the logistic pseudo-log likelihood. Table 3.7 shows the pseudomaximum log-likelihood estimating equations that were used to estimate $\mathbf{B}$. These estimating equations were solved numerically using the `optim()` function in R (R Development Core Team 2012). Appendix B.2.5 on page 309 describes the pseudoempirical maximum likelihood method for estimating $\mathbf{B}$. Table 3.7 shows both the sample pseudo log-likelihood estimating equations as well as the derivative of them. With $\widehat{\mathbf{B}}$, $\mathbf{X}_k$, and z_k we calculated $\widehat{\boldsymbol{\mu}}_k$ for all elements on the frame.

Table 3.7: Logistic Regression Estimating Equations

Distribution of Response	Sample Pseudo Log Likelihood $\widehat{L}(\mathbf{B})$	Gradient $\widehat{\ell}(\mathbf{B})$
MN $(\mathbf{p}_k; z_k)$	$\sum_{\mathfrak{s}} d_k \left[\mathbf{y}_k^\top \left(\mathbf{X}_k \mathrm{vec}\left(\widehat{\mathbf{B}}\right)\right) - z_k \ln\left(1 + \sum_{c=1}^{C-1} e^{\mathbf{x}_k \mathrm{vec}(\widehat{\mathbf{B}})}\right)\right]$	$\sum_{\mathfrak{s}} d_k \left(\mathbf{y}_k - z_k \frac{e^{\mathbf{x}_k \mathrm{vec}(\widehat{\mathbf{B}})}}{1+\sum_{c=1}^{C-1} e^{\mathbf{x}_k \mathrm{vec}(\widehat{\mathbf{B}})}}\right) \mathbf{X}_k$

We used an explicit form of the model-calibration estimator to make estimates. For inverting matrices in the model-calibration estimator, we used the `solve()` function in R. For the model-calibrated maximum pseudoempirical likelihood, we used the `Lag2` function provided by Changbao Wu[2].

For the implicit differentiation variance estimators, we formed estimating equations and used the `jacobian()` function in R to numerically calculate the Jacobian of the survey weighted estimating equations (Gilbert 2012).

Appendix B.8 on page 430 contains the code used to select the samples and estimate all parameters.

3.3.2.4 Measures

To evaluate the point estimators, we calculated the average distance between the estimated total vector and the population value, the percent relative empirical bias, the percent relative empirical median difference, the percent relative root empirical mean squared error, and the percent relative root median squared error for the point estimators. For the variance estimators, we also calculated the confidence interval coverage. Appendix B.7 on page 374 contains these measures for all simulations.

Let ν index $\aleph$ samples. Also, let θ_c be a true population parameter for category c and $\theta_{c\nu}$ be a point estimator of $\widehat{\theta}_c$ based on sample ν. Table 3.8 shows formulas for the summary measures of the point estimators we calculated.

To evaluate the variance estimators of $\widehat{\theta}_c$, we replaced $\widehat{\theta}_{c\nu}$ with the variance estimator. Furthermore, for the percent relative empirical bias and the percent relative root

[2]See http://www.math.uwaterloo.ca/ cbwu/Rcodes/LagrangeM2.txt.

Table 3.8: Summary of empirical distributions

Estimator	Equation
Percent Simulation Coefficient of Variation	$100 \cdot \frac{\frac{1}{\sqrt{\aleph}}\sqrt{\frac{1}{\aleph-1}\sum_{\nu=1}^{\aleph}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^2}}{\boldsymbol{\theta}}$
Average Distance	$\lVert\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\rVert$
Percent Relative Bias	$100 \cdot \frac{\frac{1}{\aleph}\sum_{\nu=1}^{\aleph}\left(\widehat{\theta}_{c\nu}-\theta_c\right)}{\theta_c}$
Percent Median Relative Bias	$100 \cdot \frac{\text{median}\left(\widehat{\theta}_{c\nu}-\theta_c\right)}{\theta_c}$
Percent Relative Root Mean Squared Error	$100 \cdot \frac{\sqrt{\frac{1}{\aleph}\sum_{\nu=1}^{\aleph}\left(\widehat{\theta}_{c\nu}-\theta_c\right)^2}}{\theta_c}$
Percent Relative Root Median Squared Error	$100 \cdot \frac{\sqrt{\text{median}\left(\widehat{\theta}_{c\nu}-\theta_c\right)^2}}{\theta_c}$

empirical mean squared error, θ_c was replaced with the empirical variance of θ_c

$$\upsilon_{emp}\left(\theta_c\right) = \frac{1}{\aleph}\sum_{\nu=1}^{\aleph}\left(\widehat{\theta}_{c\nu} - \theta_c\right)^2.$$

For the percent relative empirical median difference and the percent relative root empirical median squared error, $\widehat{\theta}_c$ was replaced with empirical median squared error of $\widehat{\theta}_c$

$$m\upsilon_{emp}\left(\theta_c\right) = \text{median}\left(\widehat{\theta}_{c\nu} - \theta_c\right)^2.$$

To evaluate the performance of the variance estimators, we constructed confidence intervals using the variance estimators. We calculated the percent of samples in which the confidence intervals contained the true population value. Confidence intervals were created using the *t*-distribution with $n-1$ degrees of freedom where n is the number of sample clusters. Specifically, we counted the number of samples where

$$\frac{\left|\widehat{t}_c - t_c\right|}{\sqrt{\upsilon\left(\widehat{t}_c\right)}} \leq t_{n-1,0.975}.$$

Dividing this count by the number of samples and multiplying by 100 gave us the percent confidence interval coverage.

We expect that about 95% of the confidence intervals should contain the true value.

We note that using $n-1$ degrees of freedom is a commonly used approximation, but not exact. In large samples, errors associated with using this approximation are negligible.

3.3.3 Results

3.3.3.1 Simulation Errors

In general, we designed our populations and sample designs to limit the risk of encountering a problem estimating certain quantities. Nevertheless, when modeling the response data, several problems arose, especially in small samples.

There are several practical problems that may hinder using one of the logistic-assisted estimators. Point estimation is not possible if

1. Responses in one of the categories is the same for all sample units. This is common with rare characteristics where the characteristic is not observed in sample.
2. $\mathbf{X}$ is not full rank. Of course this can easily be fixed in practice by removing the dependent variable or using a generalized inverse when inverting functions of $\mathbf{X}$.
3. μ_k cannot be predicted for a non-sample unit. For example, if none of the sample units has one level of a covariate used to model μ_k.

Furthermore, implicit differentiation variance estimators are not possible if

4. $(\boldsymbol{\mu}\mathbf{w})^\top \boldsymbol{\mu}$ is not full rank,
5. the jacobian of LGREG estimating equations is not full rank, or
6. the jacobian of MCAL estimating equations is not full rank.

Table 3.9 shows the number of errors encountered in each simulation. The simulation numbers correspond to the simulations in Table 3.6 on page 152. As soon as an error was encountered, the sample was thrown out and a new sample was selected to replace the skipped sample. For this reason, the counts in Table 3.9 are not mutually exclusive.

Table 3.9: Number of Errors Found in Each Simulation

Simulation	Error 1	Error 2	Error 3	Error 4	Error 5	Error 6
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	0	0	0	0
6	0	0	0	0	0	0
7	5	0	9	0	251	20
8	0	0	0	0	0	0
9	9	1	14	0	367	40
10	0	0	0	0	0	0
11	3	0	5	0	255	21
12	0	0	0	0	0	0
13	0	0	0	0	0	0
14	0	0	0	0	0	0
15	0	0	0	0	0	0
16	0	0	0	0	0	0
17	0	0	0	0	0	0
18	0	0	0	0	0	0

For example, if $\mathbf{y}_{math} = 0$ for all sample units and $\mathbf{X}$ was not full rank, only Error 1 would be recorded in Table 3.9.

As we see, problems were only encountered in the small samples from the postsecondary sample. For the model-calibration estimator, there could be negative or near zero calibrated weights. This was the result of unstable estimates of $\mathbf{B}$ for one or more variables. For the Postsecondary population, this often had the negative effect of inflating the estimates of t_{math} and attenuating the other estimates. Because the sum of the categories for each unit was fixed, instability in estimating one of the categories adversely impacted the other categories. In general, errors were more frequent with variance estimation than with point estimators.

There are numerous techniques that can be used to correct for the common model-

ing errors encountered when estimating rare characteristics or data from small samples. Since the focus of this paper is not on small area estimation, we did not employ such techniques, but consider it a worthwhile endeavor for future research.

To summarize, Table 3.9 shows that very few critical errors prevented us from making inference from the logistic-assisted estimators.

3.3.3.2 Point Estimators: Average Distance from True Value

There are numerous ways to measure the performance of point estimators. In this dissertation, we consider two measures: the average distance and the relative root mean squared error. The average distance summarizes the performance of all categories into one measure, while the relative root mean squared error measures the performance of each estimator for each category separately. In this section we focus on the average distance between the estimator and the true value for all estimators across the three populations. In Section 3.3.3.3 on page 164 we present results for the mean squared error of our point estimators. We begin with some introductory comments about the average distance. Then we report our results, focusing on how the estimators perform in large samples and how the estimators perform in small samples. For the small samples, we investigate the estimators when the assisting model is correctly specified and when the assisting models do not fit the data.

Consider a C-dimensional space where each dimension is defined by a category. Let $\widehat{\mathbf{t}}_{y\nu}$ be a C-dimensional estimate of the finite population vector $\mathbf{t}_y$ from sample ν. The Euclidian distance between an estimate and the true value is the norm, $\|\widehat{\mathbf{t}}_{y\nu} - \mathbf{t}_y\| =$

$\sqrt{\sum_{c=1}^{C}\left(\widehat{t}_{yc}-t_{yc}\right)^2}$. This measure summarizes of how far the estimator is from the true value. We calculated the norm for each sample and then summarized the norms by calculating the mean and variance of the norms across all $\aleph$ samples.

The norm does not equally weight each category. Instead, larger categories tend to dominate the norm in practice. For example, consider a 2-dimensional estimate, $(10, 500)$, of the vector $(50, 500)$. In this case the estimate is 40 units from the true value. On the other hand, an estimate of $(50, 100)$ will be 400 units from the true value. In both cases, one of the categories was 20% of the true value, but the distance between the estimate and the true value was driven by the larger category.

Table 3.10: Average Distance from True Value for Synthetic Population (in thousands)

	Small Samples			Large Samples		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
$\hat{\mathbf{t}}_y^{\pi}$	7,317.3	8,715.5	6,848.9	844.1	1023.7	777.0
$\hat{\mathbf{t}}_{yc}^{gd}$	5,419.9	5,597.9	5,385.2	638.0	633.0	609.7
$\hat{\mathbf{t}}_y^{lg}$	431.9	430.7	424.7	47.6	47.1	46.8
$\hat{\mathbf{t}}_y^{mc}$	468.0	482.9	469.3	47.6	47.1	46.8
$\hat{\mathbf{t}}_y^{peM}$	484.8	549.2	494.2	47.6	47.1	46.8
$\hat{\mathbf{t}}_y^{pe\widehat{M}}$	1,641.6	4,176.2	494.2	182.1	475.9	46.8

We calculated the average distance between the estimators and the true finite population total for all estimators across the three populations. See Appendix B.7 on page 374 for tables of all results as well as tables showing the standard error of the average distance. Here we only present results for the Synthetic and Postsecondary populations. Tables 3.10 and 3.11 show the average distance between each estimator and the finite population total for all estimators in the Synthetic and Postsecondary populations. Table

B.34 on page 415 in the appendix shows results for the Census population.

In large samples, the logistic-assisted estimators often perform better than the GREG and π-estimators. Regardless of the population, $\hat{\mathbf{t}}_y^{lg}$, $\hat{\mathbf{t}}_y^{mc}$, and $\hat{\mathbf{t}}_y^{peM}$ were closer to the true population value on average in the large samples. In fact, these three logistic-assisted estimators were often at least 30% closer to the population total than the GREG estimator in the large samples. Across all three populations and all three sample designs, we see that $\hat{\mathbf{t}}_y^{lg}$, $\hat{\mathbf{t}}_y^{mc}$, and $\hat{\mathbf{t}}_y^{peM}$ are all about the same in large samples. Thus, our simulations suggest that these three estimators are asymptotically equivalent.

In the Synthetic population, the logistic model fits the data very well. Table 3.10 shows that the logistic-assisted estimators perform better than the GREG and π-estimators in both small and large samples. In fact, the logistic-assisted estimators can be much better than the other estimators. All three logistic-assisted estimators are over 90% closer to the true population total vector on average in the small samples. On average, $\hat{\mathbf{t}}_y^{lg}$ is closest to the true population value in small samples.

Table 3.11: Average Distance from True Value for Postsecondary Population (in thousands)

	Small Samples			Large Samples		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
$\hat{\mathbf{t}}_y^{\pi}$	576	602	539	262	268	242
$\hat{\mathbf{t}}_{yc}^{gd}$	367	371	365	164	160	157
$\hat{\mathbf{t}}_y^{lg}$	274	284	284	107	103	102
$\hat{\mathbf{t}}_y^{mc}$	898	1,084	1,067	114	111	111
$\hat{\mathbf{t}}_y^{peM}$	364	372	374	116	115	114
$\hat{\mathbf{t}}_y^{\widehat{peM}}$	429	497	374	154	188	114

When the assisting model does not fit the data as well, we also see that the logistic-assisted estimators are often closer to the finite population value than the GREG and π-estimators in small samples, although there are exceptions. For example, in Table 3.11, we see that $\hat{\mathbf{t}}_y^{mc}$ tends to be farther from the finite population value than all other estimators. As we will see in Section 3.3.3.4, this is primarily driven by instability of $\hat{t}_{math}^{mc}$ and $\hat{t}_{health}^{mc}$. Although $\hat{\mathbf{t}}_y^{mc}$ performs poorly for the small samples in the Postsecondary population, it performs well in the Synthetic and Census populations. In general $\hat{\mathbf{t}}_y^{lg}$ tends to be closer to the finite population total than other estimators.

Figure 3.1 shows density plots of the norms in the Census simulations. In both the small and large samples, we see that the distribution of the distance between the finite population total vector and $\widehat{\mathbf{t}}_y^{\pi}$, $\widehat{\mathbf{t}}_y^{gd}$, and $\widehat{\mathbf{t}}_y^{pe\widehat{M}}$ is much wider than the distribution for the logistic-assisted estimators. Since the bulk of the mass under the densities for $\hat{\mathbf{t}}_y^{lg}$, $\hat{\mathbf{t}}_y^{mc}$, and $\hat{\mathbf{t}}_y^{peM}$ are closer to 0, we conclude that these estimators are consistently better than the other estimators.

Our simulations also support the theoretical result that $\hat{\mathbf{t}}_y^{peM} = \hat{\mathbf{t}}_y^{pe\widehat{M}}$ in Fixed PPS samples. Tables 3.10 and 3.11 and Figure 3.1 all show that $\hat{\mathbf{t}}_y^{peM} = \hat{\mathbf{t}}_y^{pe\widehat{M}}$ in Fixed PPS samples. In fact, this result is consistent in both the small and large samples.

In summary, the three logistic-assisted estimators: $\hat{\mathbf{t}}_y^{lg}$, $\hat{\mathbf{t}}_y^{mc}$, and $\hat{\mathbf{t}}_y^{peM}$ tend to outperform the GREG and π-estimators. In terms of how close these estimators are to the true population totals, $\hat{\mathbf{t}}_y^{lg}$ is clearly the best estimator, especially in smaller samples. In large samples, our empirical results support our theoretical findings that $\hat{\mathbf{t}}_y^{lg}$, $\hat{\mathbf{t}}_y^{mc}$, and $\hat{\mathbf{t}}_y^{peM}$ are asymptotically equivalent. We also see that estimating $\widehat{M}$ in $\hat{\mathbf{t}}_y^{pe\widehat{M}}$ can reduce the performance of the model-calibrated maximum pseudoempirical likelihood estimator.

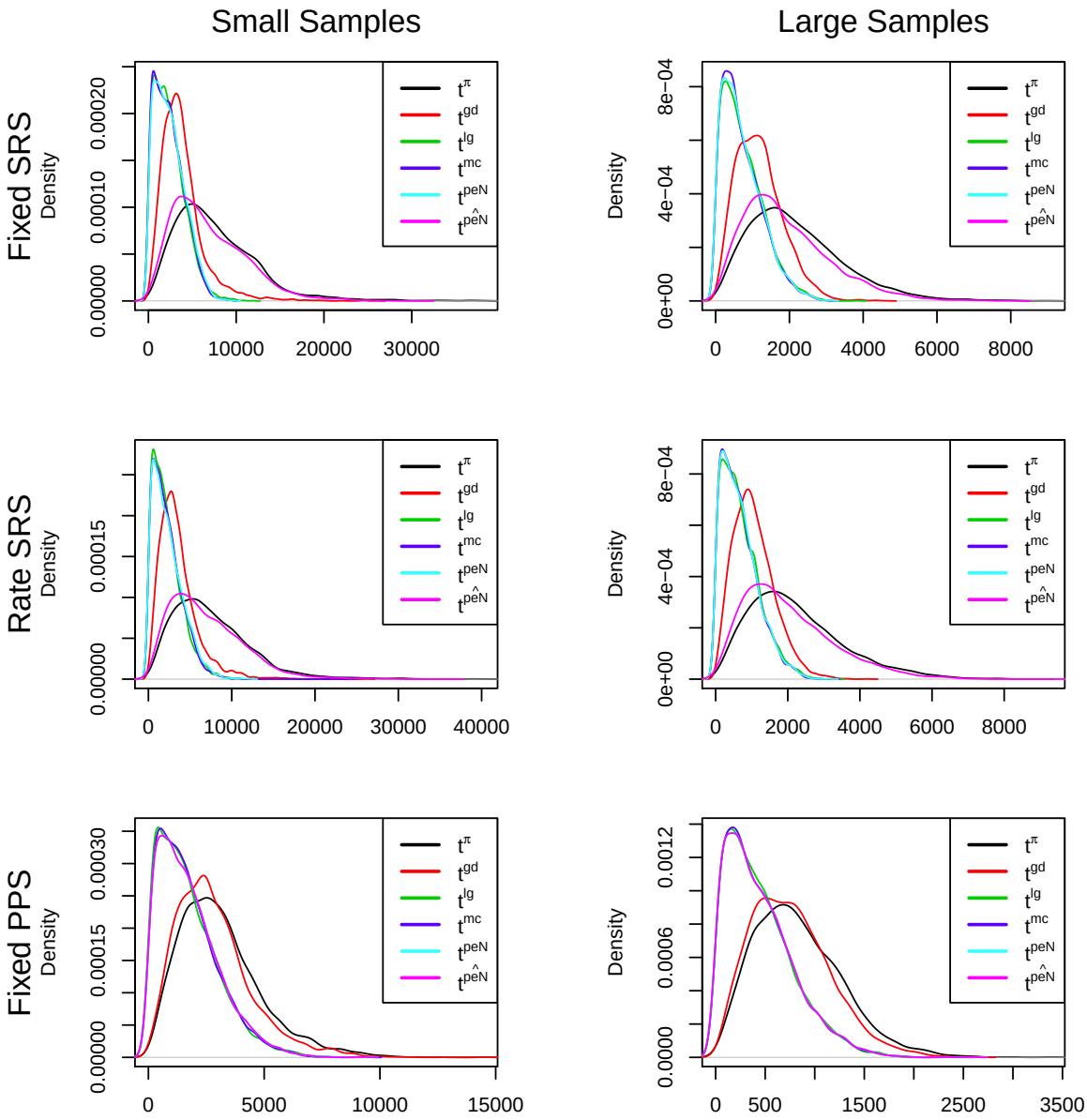

Figure 3.1: Density Plot of Distance Between Estimator and True Value for the Census Population

3.3.3.3 Point Estimators: Mean Squared Error

The relative root mean squared error is often used to summarize and compare the performance of estimators because it incorporates both bias and variance. In this section, we report the observed percent relative root mean squared error for all estimators in all populations, focusing on how the estimators perform in large samples and then how the estimators perform in small samples. For the small samples, we investigate the estimators when the assisting model is correctly specified and when the assisting models do not fit the data. In most situations the logistic-assisted estimators outperform the GREG and π-estimators. However, their dominance is not uniform. For estimators with extreme and highly influential estimates, we also refer to the empirical interquartile range.

Appendix B.7 on page 374 shows the empirical percent relative root mean squared error for all estimators in all three populations. In this section, we report results for the Synthetic and Postsecondary populations. Tables 3.12 and 3.13 on pages 166 and 167 show the percent relative root mean squared error for each estimator in the Synthetic and Postsecondary populations. Table B.38 on page 419 in the appendix shows results for the Census population. In addition to the percent relative root mean squared error, Appendix B.7 also shows the percent relative root median squared error for all estimators in all populations.

When the sample size is small and the model fits very well, $\hat{\mathbf{t}}_y^{lg}$, $\hat{\mathbf{t}}_y^{mc}$, and $\hat{\mathbf{t}}_y^{peM}$ outperform the other estimators. Furthermore, $\hat{\mathbf{t}}_y^{pe\widehat{M}}$ is less variable than the π and GREG estimators. Figure 3.2 shows box-and-whisker plots for y_1 in the small Fixed SRS samples of the Synthetic population. These plots summarize the empirical distribution of our

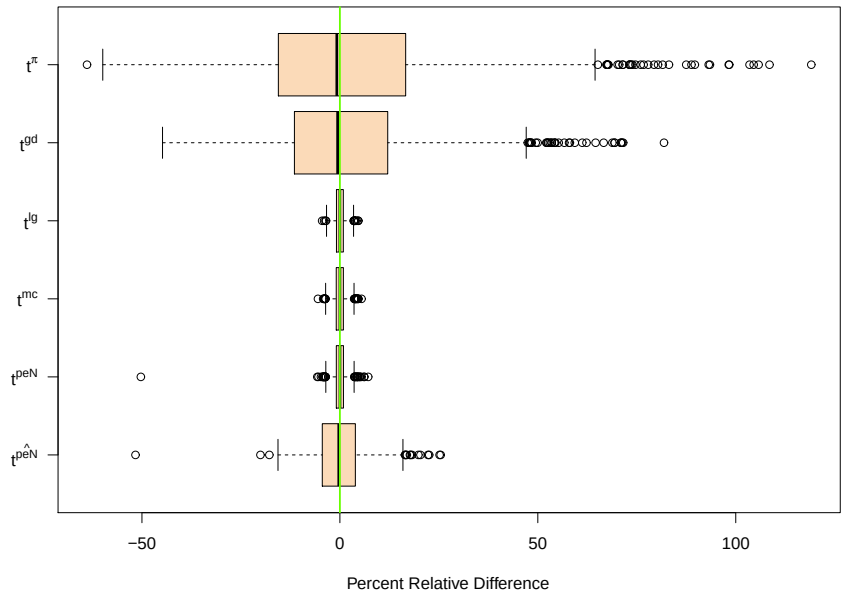

Figure 3.2: Box-and-Whisker Plot Showing Percent Relative Difference of Estimated Totals for y_1 of Synthetic Population under Small Fixed SRS

point estimators. It is quite clear from Figure 3.2 that $\hat{\mathbf{t}}_y^{lg}$, $\hat{\mathbf{t}}_y^{mc}$, and $\hat{\mathbf{t}}_y^{peM}$ produce less variable estimators when the model fits the population very well, even in small samples. As expected, $\hat{\mathbf{t}}_y^{pe\widehat{M}}$ is more variable than $\hat{\mathbf{t}}_y^{peM}$ because $\hat{\mathbf{t}}_y^{pe\widehat{M}}$ requires estimating M in addition to estimating $\mathbf{t_y}$.

Table 3.12 shows the percent relative root mean squared error for all estimators in the Synthetic population. Table 3.12 shows that the logistic-assisted estimators perform very well when the model is correctly specified. In fact, the relative root mean squared error is often one tenth of the relative root mean squared error of the GREG estimator.

In small samples when the assisting model is less than ideal, the performance of the logistic-assisted estimators is mixed. As we saw in the Synthetic population, $\hat{\mathbf{t}}_y^{lg}$, $\hat{\mathbf{t}}_y^{mc}$, and $\hat{\mathbf{t}}_y^{peM}$ outperform the π and GREG estimators in terms of mean squared error for small samples in the Census population.

Table 3.12: Percent Relative Root Mean Squared Error of Total Estimators for Synthetic Population

	Small Samples								
	Fixed SRS			Rate SRS			Fixed PPS		
Estimator	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
$\hat{\mathbf{t}}_y^{\pi}$	25.3	21.2	29.0	29.9	25.9	33.5	23.4	19.6	26.9
$\hat{\mathbf{t}}_{yc}^{gd}$	18.7	21.5	20.6	18.6	22.5	22.1	18.9	21.1	20.4
$\hat{\mathbf{t}}_y^{lg}$	1.3	4.3	1.4	1.3	4.3	1.4	1.3	4.2	1.4
$\hat{\mathbf{t}}_y^{mc}$	1.4	4.7	1.5	1.5	5.0	1.6	1.4	4.8	1.6
$\hat{\mathbf{t}}_y^{peM}$	1.8	4.9	1.8	3.0	5.3	3.5	1.9	5.0	2.0
$\hat{\mathbf{t}}_y^{pe\widehat{M}}$	6.2	7.8	6.2	16.4	16.9	16.5	1.9	5.0	2.0
	Large Samples								
$\hat{\mathbf{t}}_y^{\pi}$	2.8	2.4	3.3	3.5	3.0	4.0	2.6	2.2	3.1
$\hat{\mathbf{t}}_{yc}^{gd}$	2.2	2.3	2.4	2.2	2.3	2.4	2.1	2.2	2.3
$\hat{\mathbf{t}}_y^{lg}$	0.1	0.5	0.2	0.1	0.5	0.1	0.1	0.5	0.2
$\hat{\mathbf{t}}_y^{mc}$	0.1	0.5	0.2	0.1	0.5	0.1	0.1	0.5	0.2
$\hat{\mathbf{t}}_y^{peM}$	0.1	0.5	0.2	0.1	0.5	0.1	0.1	0.5	0.2
$\hat{\mathbf{t}}_y^{pe\widehat{M}}$	0.7	0.8	0.7	1.9	1.9	1.9	0.1	0.5	0.2

On the other hand, Table 3.13 shows extremely large estimates of the percent relative root mean squared error for $\widehat{t}_{math}^{lg}$ and $\widehat{t}_{math}^{mc}$ in small samples from the Postsecondary population. As we will see in Section 3.3.3.4, math majors are rare which causes some extreme and influential estimates in the logistic-assisted estimators for the small samples in the Postsecondary population. Yet, Table 3.17 shows that the interquartile range for $\hat{\mathbf{t}}_{math}^{lg}$ and $\hat{\mathbf{t}}_{math}^{mc}$ is only slightly larger than the interquartile range for the other estimators in the small Fixed SRS samples from the Postsecondary population. Table 3.13 confirms that some large estimates of $\hat{\mathbf{t}}_y^{lg}$ and $\hat{\mathbf{t}}_y^{mc}$ have adversely inflated measures of variance and mean squared error.

The percent relative root median squared error is more robust against the influence of outliers than the percent relative root mean squared error. Appendix B.7.2 on page 394

Table 3.13: Percent Relative Root Mean Squared Error for Postsecondary Population

	Small Samples											
	Fixed SRS				Rate SRS				Fixed PPS			
Estimator	Math	Health	Business	Other	Math	Health	Business	Other	Math	Health	Business	Other
$\hat{\mathbf{t}}_y^{\pi}$	70.7	29.9	44.7	33.3	68.8	32.0	47.9	34.2	66.3	28.0	45.0	30.1
$\hat{\mathbf{t}}_{yc}^{gd}$	56.4	27.7	36.3	18.3	57.7	27.6	37.7	18.9	56.5	27.5	35.4	18.1
$\hat{\mathbf{t}}_y^{lg}$	691.6	24.9	34.2	11.4	811.4	25.2	35.4	12.0	761.9	25.0	36.3	11.9
$\hat{\mathbf{t}}_y^{mc}$	748.2	325.4	330.3	138.7	867.3	350.0	331.1	130.3	841.0	370.1	359.1	136.3
$\hat{\mathbf{t}}_y^{peM}$	56.4	29.7	37.0	19.2	55.5	29.8	37.6	19.5	55.3	29.4	37.8	19.6
$\hat{\mathbf{t}}_y^{pe\widehat{M}}$	57.5	31.9	38.6	22.0	57.6	34.4	40.9	25.7	55.3	29.4	37.8	19.6
	Large Samples											
$\hat{\mathbf{t}}_y^{\pi}$	30.6	13.5	20.4	14.7	30.1	14.2	20.9	14.9	29.5	12.4	19.9	13.3
$\hat{\mathbf{t}}_{yc}^{gd}$	25.5	11.7	15.8	7.8	25.8	11.3	15.9	7.7	25.4	11.1	15.3	7.5
$\hat{\mathbf{t}}_y^{lg}$	25.9	10.7	12.6	4.2	26.9	10.2	12.1	4.1	25.6	10.2	11.9	4.0
$\hat{\mathbf{t}}_y^{mc}$	26.6	11.5	13.8	4.5	29.8	11.1	13.6	4.5	28.6	11.0	13.7	4.5
$\hat{\mathbf{t}}_y^{peM}$	25.9	11.7	14.1	4.7	27.9	11.5	14.1	4.9	27.9	11.4	14.0	4.8
$\hat{\mathbf{t}}_y^{pe\widehat{M}}$	26.5	12.8	14.8	7.0	28.9	13.8	15.9	9.4	27.9	11.4	14.0	4.8

contains tables showing the relative root median squared error for the Postsecondary population. In terms of the relative root median squared error, all of the estimators perform similarly in the small samples pulled from the Postsecondary population. When compared to $\hat{\mathbf{t}}_y^{gd}$, sometimes the logistic-assisted estimators performed better than the GREG; other times, they did not.

Even though $\hat{\mathbf{t}}_y^{pe\widehat{M}}$ and $\hat{\mathbf{t}}_y^{peM}$ are very similar, all of our results show that estimating M increases the root mean squared error of the model-calibrated maximum pseudoempirical likelihood estimator. In general, there is little reason for using $\hat{\mathbf{t}}_y^{pe\widehat{M}}$. Since one needs complete auxiliary data for both $\hat{\mathbf{t}}_y^{pe\widehat{M}}$ and $\hat{\mathbf{t}}_y^{peM}$, M should be available and used. The one exception is when $\widehat{M} = M$, which we see in the Fixed PPS samples.

In large samples, we see clear advantages to the logistic-assisted estimators. The relative root mean squared error of the three logistic-assisted estimators is smaller than the relative root mean squared error of the π-estimator, and often less than the GREG. In

general, the difference between the logistic-assisted estimators is quite small, supporting the fact that they are all asymptotically equivalent. As we would expect, the mean squared error decreases as the sample size increases. This characteristic further suggests that all of the estimators are design-consistent.

Despite the fact that our assisting models are relatively simple and do not fit very well in the Postsecondary and Census populations, we see clear advantages to using $\hat{\mathbf{t}}_y^{lg}$, $\hat{\mathbf{t}}_y^{mc}$, and $\hat{\mathbf{t}}_y^{peM}$ in moderate to large samples. In many cases, the root mean squared error of the logistic estimators is at least 20% less than the variances of the GREG in the Postsecondary population. Examples like this clearly show that the logistic-assisted estimators are worth further study and use.

In conclusion, the logistic-assisted estimators tend to outperform the π and GREG estimators in large samples. In small samples, we noted that $\hat{\mathbf{t}}_y^{lg}$ and $\hat{\mathbf{t}}_y^{mc}$ can give nonsensical results, especially for rare characteristics. In large samples, we saw that the π-estimator can be more than 35 times larger than the percent relative root mean squared error of the logistic-assisted estimators when the model is correctly specified. Even the more modest reductions in relative root mean squared error in the Postsecondary and Census populations can result in major reductions in sample size and cost.

3.3.3.4 Point Estimators: Percent Relative Bias

We have already proved that $\hat{\mathbf{t}}_y^{lg}$, $\hat{\mathbf{t}}_y^{mc}$, $\hat{\mathbf{t}}_y^{peM}$, and $\hat{\mathbf{t}}_y^{pe\widehat{M}}$ are asymptotically unbiased. All of our simulations supported this fact. Tables 3.14, 3.16, and 3.15 show the relative bias of all estimators for all samples in each of the three populations. Regardless of the

population and the sample design, the relative bias of all of the estimators is near zero in the large sample sizes.

Our estimates of relative bias in the smaller samples show that the relative bias of the logistic-assisted estimators can be more biased than the π-estimator, but the direction of the bias is not the same for all categories. Furthermore, the magnitude of the bias tends to be rather small, suggesting that any bias will not be meaningful in many cases.

Table 3.14: Percent Relative Bias for Synthetic Population

	Small Samples								
	Fixed SRS			Rate SRS			Fixed PPS		
Estimator	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
$\hat{\mathbf{t}}_y^{\pi}$	2.0	1.2	-0.3	0.1	0.0	-0.3	-0.4	-0.1	0.7
$\hat{\mathbf{t}}_{yc}^{gd}$	1.2	4.0	-0.4	0.6	4.7	-0.6	-0.2	2.8	-0.4
$\hat{\mathbf{t}}_y^{lg}$	0.0	-0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0
$\hat{\mathbf{t}}_y^{mc}$	0.0	-0.1	0.0	0.0	0.1	0.0	0.0	0.2	0.0
$\hat{\mathbf{t}}_y^{peM}$	0.0	-0.1	0.0	0.0	0.0	-0.1	0.0	0.2	-0.1
$\hat{\mathbf{t}}_y^{pe\widehat{M}}$	0.0	0.0	0.0	-0.5	-0.5	-0.6	0.0	0.2	-0.1
	Large Samples								
$\hat{\mathbf{t}}_y^{\pi}$	0.0	0.0	-0.1	0.1	0.1	0.1	0.0	0.0	0.1
$\hat{\mathbf{t}}_{yc}^{gd}$	0.0	0.0	-0.1	0.0	0.0	0.0	0.0	0.0	0.0
$\hat{\mathbf{t}}_y^{lg}$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\hat{\mathbf{t}}_y^{mc}$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\hat{\mathbf{t}}_y^{peM}$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\hat{\mathbf{t}}_y^{pe\widehat{M}}$	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0

When the model fits the data very well, as in the Synthetic population, the bias of estimators tends to be small in both the small and large samples. The empirical estimates of percent relative bias for the small samples in Table 3.14 show that $\hat{\mathbf{t}}_y^{lg}$, $\hat{\mathbf{t}}_y^{mc}$, $\hat{\mathbf{t}}_y^{peM}$, and $\hat{\mathbf{t}}_y^{pe\widehat{M}}$ can be unbiased in small samples under a variety of sample designs when the assisting model is correctly specified.

Whereas Table 3.14 shows the percent relative bias for the estimators in an ideal model setting, Table 3.15, shows results in a more realistic situation. When the model isn't a great fit, one can expect some minor bias in small sample sizes from the logistic-assisted estimators. This bias is neither systematically negative nor positive. Furthermore, if there is any bias, it tends to be small. The empirical bias of the LGREG estimator is always less than five percent of the true value in our Census simulations. Although we did find a small relative bias of -4.9 percent for the LGREG estimator in small Fixed SRS samples, we see that this bias tends to disappear as the number of sample clusters increases. Under a very simple model, containing only one covariate and an intercept, we see clear benefits to the LGREG estimator over the GREG and π-estimators.

Table 3.15: Percent Relative Bias for Census Population

	Small Samples					
	Fixed SRS		Rate SRS		Fixed PPS	
Estimator	Renter	Owner	Renter	Owner	Renter	Owner
$\hat{\mathbf{t}}_y^{\pi}$	0.3	0.1	0.6	0.6	0.6	-0.1
$\hat{\mathbf{t}}_{yc}^{gd}$	-2.0	2.6	-2.6	2.1	1.9	-0.3
$\hat{\mathbf{t}}_y^{lg}$	-4.9	3.4	-4.2	2.9	0.2	-0.2
$\hat{\mathbf{t}}_y^{mc}$	-5.5	3.8	-3.8	2.6	0.7	-0.5
$\hat{\mathbf{t}}_y^{peM}$	-7.5	5.2	-4.8	3.1	0.2	-0.2
$\hat{\mathbf{t}}_y^{pe\widehat{M}}$	-3.0	2.3	0.5	1.3	0.2	-0.2
	Large Samples					
$\hat{\mathbf{t}}_y^{\pi}$	0.5	0.6	0.2	0.1	0.1	0.0
$\hat{\mathbf{t}}_{yc}^{gd}$	-0.2	0.1	0.0	0.2	0.2	0.0
$\hat{\mathbf{t}}_y^{lg}$	-0.4	0.2	-0.3	0.2	0.0	0.0
$\hat{\mathbf{t}}_y^{mc}$	-0.4	0.3	-0.3	0.2	0.1	0.0
$\hat{\mathbf{t}}_y^{peM}$	-0.6	0.4	-0.4	0.3	0.0	0.0
$\hat{\mathbf{t}}_y^{pe\widehat{M}}$	0.3	1.0	0.0	0.2	0.0	0.0

On the other hand, Table 3.16 tells a more complex story. We first notice very large

estimates of bias for $\hat{\mathbf{t}}^{lg}_{math}$. The number of math degrees is small compared the other three categories. As we saw in Table 3.4 on page 148, math degrees only accounted for $16,560$ of the $3,050,986$ total degrees conferred. Furthermore, these $16,560$ math degrees were not uniformly distributed across the $6,912$ postsecondary institutions. In fact, over 75% of the postsecondary institutions did not grant any math degrees. Math degrees are rare and tend to be concentrated in a few academic institutions. The infrequency of math degrees caused numeric instability in estimating the parameters in our logistic models. It is well know that logistic models can have problems when estimating probabilities at the extremes, near 0 and 1.

We would expect quite a bit of variability when making national estimates from such a small number of institutions. Since only a few institutions granted math degrees, we would expect many of our estimates of total math majors to be quite low. On the other hand, we would also expect some very large predictions of math majors when universities with large math programs such as Columbia and the University of Los Angeles fall into sample.

A closer inspection of the small Fixed SRS samples confirms some estimates of $\hat{\mathbf{t}}^{lg}_{math}$ and $\hat{\mathbf{t}}^{mc}_{math}$ are not reasonable. Table 3.17 shows the quartiles for the percent relative difference of the math estimators in the Fixed SRS samples. Table 3.17 shows that the median value of the percentage relative error of $\hat{\mathbf{t}}^{lg}_{math}$ and $\hat{\mathbf{t}}^{mc}_{math}$ across the $10,000$ samples is near zero, indicating that usually $\hat{\mathbf{t}}^{lg}_{math}$ and $\hat{\mathbf{t}}^{mc}_{math}$ are close to the true value. However, $\hat{\mathbf{t}}^{lg}_{math}$ and $\hat{\mathbf{t}}^{mc}_{math}$ produce some very large estimates which inflate the bias estimates considerably. The large difference between the median and mean for these estimators further suggests some extreme estimates are skewing the estimate of bias. One should certainly

Table 3.16: Percent Relative Bias for Postsecondary Population

	Small Samples											
	Fixed SRS				Rate SRS				Fixed PPS			
Estimator	Math	Health	Business	Other	Math	Health	Business	Other	Math	Health	Business	Other
$\hat{\mathbf{t}}_y^{\pi}$	0.6	-0.1	0.3	0.5	1.7	0.7	1.1	1.4	2.2	-0.1	0.9	0.7
$\hat{\mathbf{t}}_{yc}^{gd}$	-6.4	0.5	-3.3	-1.8	-5.2	0.3	-3.5	-1.5	-4.1	0.4	-2.8	-1.4
$\hat{\mathbf{t}}_y^{lg}$	171.7	1.5	3.5	-3.0	216.3	1.1	3.4	-3.1	210.3	1.1	3.8	-3.2
$\hat{\mathbf{t}}_y^{mc}$	160.1	-21.3	0.0	5.9	210.0	-25.8	-3.3	7.9	206.8	-28.2	4.1	6.8
$\hat{\mathbf{t}}_y^{peM}$	-11.3	0.4	-4.1	-6.1	-11.4	0.1	-4.5	-6.2	-11.0	-0.9	-4.9	-6.6
$\hat{\mathbf{t}}_y^{pe\widehat{M}}$	-11.6	0.2	-4.5	-6.4	-10.7	0.7	-3.9	-5.5	-11.0	-0.9	-4.9	-6.6
	Large Samples											
$\hat{\mathbf{t}}_y^{\pi}$	0.0	0.1	0.1	0.0	-0.2	0.0	0.0	0.0	-0.2	0.0	-0.1	-0.1
$\hat{\mathbf{t}}_{yc}^{gd}$	-1.5	0.1	-0.9	-0.5	-1.2	0.1	-0.7	-0.4	-1.1	0.0	-0.8	-0.4
$\hat{\mathbf{t}}_y^{lg}$	3.2	0.6	0.4	-0.4	1.8	0.3	0.0	-0.1	1.8	0.3	-0.1	-0.1
$\hat{\mathbf{t}}_y^{mc}$	0.0	0.6	0.4	-0.3	1.1	0.5	0.2	-0.2	1.0	0.6	-0.1	-0.2
$\hat{\mathbf{t}}_y^{peM}$	-1.0	0.7	0.4	-0.4	-0.2	0.6	0.2	-0.4	0.0	0.6	0.0	-0.3
$\hat{\mathbf{t}}_y^{pe\widehat{M}}$	-0.9	0.7	0.4	-0.4	-0.2	0.5	0.2	-0.4	0.0	0.6	0.0	-0.3

be cautious about using model-assisted techniques in small samples of rare characteristics.

Despite some instability and extreme estimates, the median relative difference for $\hat{\mathbf{t}}_{math}^{lg}$ and $\hat{\mathbf{t}}_{math}^{mc}$ is closer to zero than the competing estimators. Although these estimators are extremely biased, most of the time they are pretty close to the true value, even in small samples of rare characteristics.

From Table 3.16, we also see evidence of bias when estimating $\hat{\mathbf{t}}_{health}^{mc}$ in small samples. As we noted in the theoretical results, $\hat{\mathbf{t}}_{health}^{mc}$ can be negative. When the percent relative difference is -100%, the estimate will be 0. Estimates less than -100% will be negative.

Table 3.18 confirms our theoretical note that $\hat{\mathbf{t}}_{health}^{mc}$ can produce negative estimates. Table 3.18 also shows that $\hat{\mathbf{t}}_{health}^{mc}$ is sometimes 70 times larger than the true population value. When $\sum_{k\in\mathfrak{s}} d_k \underline{\boldsymbol{\mu}}_{yk}$ is much larger than $\sum_{k\in\mathscr{U}} \underline{\boldsymbol{\mu}}_{yk}$, then $\hat{\mathbf{t}}_y^{mc}$ may be negative.

Table 3.17: Quartiles for Percent Relative Difference of Math Estimators with Sample of Fixed SRS in Postsecondary Population

Estimator	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
			Small Samples			
$\hat{\mathbf{t}}_y^{\pi}$	-98.0	-47.2	-15.3	0.7	28.4	596.9
$\hat{\mathbf{t}}_{yc}^{gd}$	-350.1	-42.8	-15.5	-6.4	18.9	526.1
$\hat{\mathbf{t}}_y^{lg}$	-99.3	-28.9	3.7	171.7	61.6	10,326.1
$\hat{\mathbf{t}}_y^{mc}$	-11,268.1	-36.4	-3.0	160.1	52.1	11,244.0
$\hat{\mathbf{t}}_y^{peM}$	-99.8	-47.2	-20.9	-11.3	11.3	547.5
$\hat{\mathbf{t}}_y^{p\widehat{eM}}$	-99.8	-47.9	-21.8	-11.6	10.7	623.2
			Large Samples			
$\hat{\mathbf{t}}_y^{\pi}$	-73.4	-21.9	-4.4	0.0	17.3	183.0
$\hat{\mathbf{t}}_{yc}^{gd}$	-65.3	-19.2	-4.8	-1.5	12.1	221.9
$\hat{\mathbf{t}}_y^{lg}$	-57.0	-13.6	0.1	3.2	15.7	317.8
$\hat{\mathbf{t}}_y^{mc}$	-149.2	-16.5	-3.8	0.0	11.8	320.2
$\hat{\mathbf{t}}_y^{peM}$	-77.5	-17.3	-4.5	-0.9	10.9	275.7
$\hat{\mathbf{t}}_y^{p\widehat{eM}}$	-76.4	-17.8	-4.5	-0.9	11.0	266.6

Likewise, when $\sum_{k\in\mathfrak{s}} d_k \underline{\boldsymbol{\mu}}_{yk}$ is much smaller than $\sum_{k\in\mathscr{U}} \underline{\boldsymbol{\mu}}_{yk}$, then $\hat{\mathbf{t}}_y^{mc}$ may be much larger than the true value. When estimating $\hat{\mathbf{t}}_y^{mc}$ in small samples, one should monitor the difference between $\sum_{k\in\mathfrak{s}} d_k \underline{\boldsymbol{\mu}}_{yk}$ and $\sum_{k\in\mathscr{U}} \underline{\boldsymbol{\mu}}_{yk}$ and compare $\hat{\mathbf{t}}_y^{mc}$ to other estimators. Despite the potential for extreme estimates of $\hat{\mathbf{t}}_y^{mc}$, Table 3.18 shows that about half the time $\hat{\mathbf{t}}_{health}^{mc}$ is less than the finite population value and about half the time it is greater than the true finite population total. Thus, in terms of the median relative bias, $\hat{\mathbf{t}}_{health}^{mc}$ is unbiased, even in small samples.

Table 3.16 also shows that the bias of the logistic-assisted estimators tends to be negligible in realistic situations with large samples. In fact, we found that the empirical bias of the logistic-assisted estimators is always less than 3.5% of the true value in large samples across all three populations. Although we did find a relative bias of 3.2 percent for $\hat{\mathbf{t}}_{y,math}^{lg}$ in Fixed SRS samples, we expect this bias to disappear as the number of sample

Table 3.18: Quartiles for Percent Relative Difference of Health Estimators with Sample of Fixed SRS in Postsecondary Population

Estimator	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
			Small Samples			
$\hat{\mathbf{t}}_y^{\pi}$	-69.6	-21.2	-4.3	-0.1	17.1	192.9
$\hat{\mathbf{t}}_{yc}^{GD}$	-108.4	-18.0	-2.7	0.6	15.4	297.5
$\hat{\mathbf{t}}_y^{LG}$	-64.0	-15.7	-0.4	1.6	15.8	212.8
$\hat{\mathbf{t}}_y^{mc}$	-26,354.7	-21.3	-0.4	-27.3	21.7	7,127.9
$\hat{\mathbf{t}}_y^{peM}$	-87.4	-19.6	-2.4	0.4	16.4	292.9
$\hat{\mathbf{t}}_y^{pe\hat{M}}$	-89.0	-21.8	-4.0	0.1	17.0	369.4
			Large Samples			
$\hat{\mathbf{t}}_y^{\pi}$	-48.8	-9.3	-0.5	0.1	8.5	59.6
$\hat{\mathbf{t}}_{yc}^{GD}$	-44.1	-8.0	-0.5	0.1	7.6	50.7
$\hat{\mathbf{t}}_y^{LG}$	-37.7	-6.7	0.1	0.6	7.5	47.3
$\hat{\mathbf{t}}_y^{mc}$	-77.5	-7.1	0.1	0.6	7.8	91.9
$\hat{\mathbf{t}}_y^{peM}$	-42.4	-7.2	0.0	0.7	7.6	77.3
$\hat{\mathbf{t}}_y^{pe\hat{M}}$	-42.2	-8.1	-0.2	0.7	8.2	86.3

clusters increases.

In summary, the logistic-assisted estimators are unbiased. We found clear empirical evidence to support this in large samples. In smaller samples, one should be cautious when using logistic-assisted estimators. Estimates of rare characteristics should be thoroughly reviewed. Furthermore, one should check model diagnostics to see if a few extreme observations are artificially inflating or deflating estimates. Yet, assuming the estimates have been vetted, the logistic-assisted estimators seem to estimate about what they should, even in small samples.

3.3.3.5 Point Estimators: Summary Across All Populations

In general, we found that all of our point estimators were centered around the finite population total. Furthermore, $\hat{\mathbf{t}}_y^{lg}$, $\hat{\mathbf{t}}_y^{mc}$, and $\hat{\mathbf{t}}_y^{peM}$ tend to have smaller mean squared error

than the other estimators.

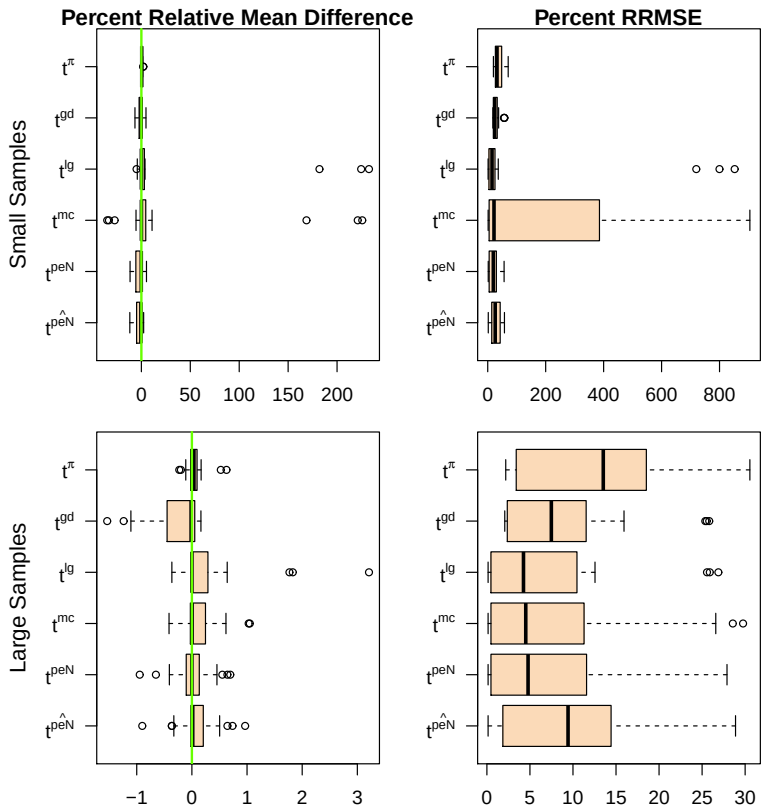

Figure 3.3: Box-and-Whisker Plot Showing Summary of All Point Estimators

Figure 3.3 shows box-and-whisker plots summarizing each simulation. Each box-and-whisker plot is based on 27 estimates, one for each category in each sample design. For example, the first box-and-whisker plot in the upper left plot shows the the average percent relative difference for the π-estimator for the three categories in the synthetic simulation, the four categories in the postsecondary design, and the two categories in the Census design for the Fixed SRS, Rate SRS, and PPS sample designs.

The outliers in the upper right quadrant of Figure 3.3 confirm that one should be cautious when estimating $\hat{\mathbf{t}}_y^{lg}$ and $\hat{\mathbf{t}}_y^{mc}$ in some small samples. However, in many situations, the logistic-assisted estimators will be stable and centered around the true value. Certainly, as the sample size increases, the difference between the estimators and the finite

population total decreases.

If we look at the median value of the percent relative root mean squared error in the large samples, we see strong evidence that the logistic-assisted estimators outperform the other estimators.

Although one needs to be careful with using the logistic-assisted estimators, they can be several times more efficient than common estimators, such as the π-estimator and GREG estimator.

3.3.3.6 Variance Estimators of $\widehat{\mathbf{t}}_y^{lg}$

As we have seen, $\widehat{\mathbf{t}}_y^{lg}$ is often more efficient that competing estimators. In addition to estimating finite population totals, it is often essential to estimate the variability of the estimator in repeated samples. In this section, we compare three different variance estimators for $\widehat{\mathbf{t}}_y^{lg}$. Specifically, we compare $\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$, $\upsilon_e\left(\widehat{\mathbf{t}}_y^{lg}\right)$, and $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$. Certainly jackknife, bootstrap, and other resampling variance estimators could be constructed, but they are not explored in this dissertation. In this section, we show that none of the variance estimators performs especially well in small samples. However, of the three estimators, $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$ generally performs better than $\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$ and $\upsilon_e\left(\widehat{\mathbf{t}}_y^{lg}\right)$.

We begin our analysis with small samples in the Synthetic population. Figure 3.4 shows box-and-whisker plots of the relative difference for the three variance estimators of $\mathbf{t}_{y_1}^{lg}$. In the small samples, the median value of $\upsilon_{Binder}\left(\mathbf{t}_{y_1}^{lg}\right)$ is closest to the empirical variance; however, the distribution of all three variance estimators are quite similar. We also see one sample produced estimates of $\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$ and $\upsilon_e\left(\widehat{\mathbf{t}}_y^{lg}\right)$ well over seven times

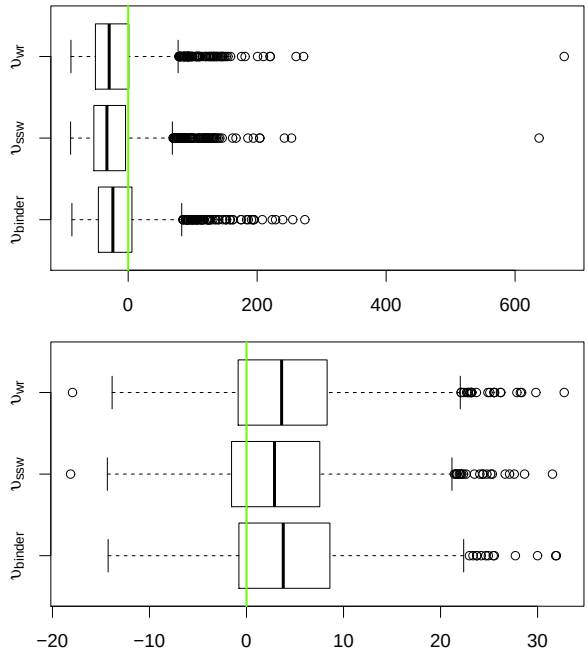

Figure 3.4: Box-and-Whisker Plots Showing Percent Relative Difference of LGREG Variance Estimators for y_1 in Fixed SRS Samples from Synthetic Population. Small sample sizes on top.

the empirical variance. In most of the small samples, all three estimators underestimated the empirical variance.

Table 3.19 shows the relative bias of the three variance estimators for all categories in the three samples from the Synthetic population. In the small samples, all of the variance estimators tend to underestimate the empirical variance. Consistently, $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$ seems to be less biased than the other two estimators in small samples. In large samples, $\upsilon_e\left(\widehat{\mathbf{t}}_y^{lg}\right)$ and $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$ are the least biased. Appendix B.7.1 shows estimates of the percent relative root mean squared error for all three variance estimators. The mean squared error for all three variance estimators are about the same as suggested in Figure 3.4. Table 3.19 shows that even in ideal conditions, there are opportunities for improvement.

Table 3.19: Percent Relative Difference of LGREG Variance Estimators for Synthetic Population

	Small Samples								
	Fixed SRS			Rate SRS			Fixed PPS		
Estimator	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-18.6	-15.0	-23.9	-26.0	-27.1	-30.5	-20.2	-19.9	-22.9
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-22.6	-19.3	-27.7	-29.7	-30.8	-34.0	-24.2	-23.9	-26.7
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-13.8	-9.8	-16.1	-12.3	-14.8	-16.5	-10.3	-11.0	-13.2
Large Samples									
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	3.9	-3.2	4.6	-4.1	-7.1	2.5	-1.6	-1.8	-2.4
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	3.2	-3.8	3.9	-4.8	-7.7	1.8	-2.3	-2.5	-3.1
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	4.0	-3.1	4.7	-4.0	-7.0	2.7	-1.4	-1.6	-2.3

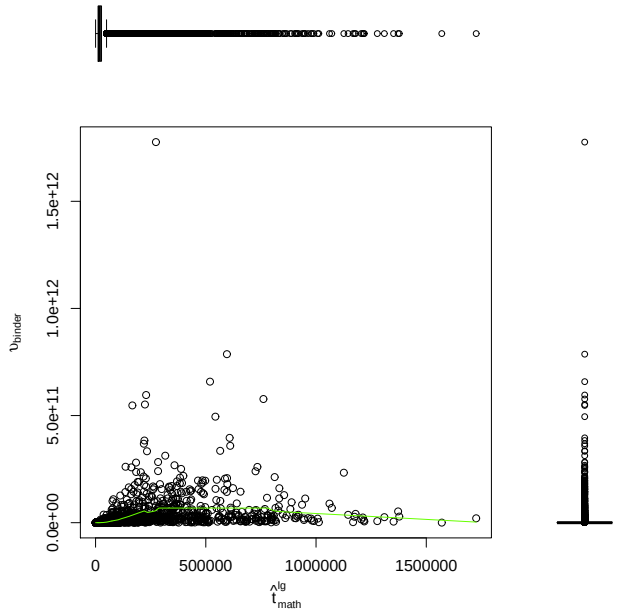

Figure 3.5: Plot of $\widehat{t}_{math}^{lg}$ versus $\upsilon_{Binder}\left(\widehat{t}_{math}^{lg}\right)$ under Small Fixed SRS

When the assisting model accurately describes the population as seen in the Synthetic population, the three variance estimators perform similarly in small and large samples. However, when the assisting models do not fit the data well, we see differences between the estimators. One might also expect poor variance estimates when point estimates are extreme. In Section 3.3.3.4 on page 168 we noted that some estimates of $\widehat{t}^{lg}_{math}$ were extremely large. In Figure 3.5, we plot estimates of $\widehat{t}^{lg}_{math}$ and $v_{Binder}\left(\mathbf{t}^{lg}_{math}\right)$ in small Fixed SRS samples. As we see, extremely large values of $\widehat{t}^{lg}_{math}$ do not necessarily correspond to inaccurate estimates of $v_{Binder}\left(\mathbf{t}^{lg}_{math}\right)$.

Table 3.20: Relative bias of LGREG Variance Estimators for Postsecondary Population

	Small Samples											
	Fixed SRS				Rate SRS				Fixed PPS			
Estimator	Math	Health	Business	Other	Math	Health	Business	Other	Math	Health	Business	Other
$v_{wr}\left(\widehat{\mathbf{t}}^{lg}_{y}\right)$	-99.9	-49.0	-76.9	-66.6	-99.9	-51.7	-79.3	-70.4	-99.9	-50.8	-80.0	-69.9
$v_{e}\left(\widehat{\mathbf{t}}^{lg}_{y}\right)$	-99.9	-54.3	-79.3	-70.0	-99.9	-56.7	-81.4	-73.5	-99.9	-55.9	-82.0	-73.0
$v_{Binder}\left(\widehat{\mathbf{t}}^{lg}_{y}\right)$	-65.1	-34.1	-53.1	-47.1	-68.1	-37.2	-53.5	-51.1	-64.6	-33.9	-55.0	-49.5
Large Samples												
$v_{wr}\left(\widehat{\mathbf{t}}^{lg}_{y}\right)$	-53.2	-14.2	-30.0	-18.9	-59.4	-10.7	-30.8	-20.1	-55.1	-10.7	-26.7	-17.6
$v_{e}\left(\widehat{\mathbf{t}}^{lg}_{y}\right)$	-55.5	-18.3	-33.3	-22.8	-61.6	-15.2	-34.1	-24.3	-57.5	-15.3	-30.4	-21.9
$v_{Binder}\left(\widehat{\mathbf{t}}^{lg}_{y}\right)$	-33.8	-12.2	-28.8	-18.9	-33.2	-7.1	-27.3	-17.2	-28.7	-7.4	-23.3	-14.4

Table 3.20 shows the average relative difference between the variance estimator and the empirical difference for the Postsecondary population. All of the variance estimators systematically underestimate the empirical variance of $\widehat{t}^{lg}_{yc}$. The magnitude of the bias seems to decrease as the sample size increases. Of course, in some instances the empirical variance includes some very large estimates of $\widehat{t}^{lg}_{yc}$ which have a large influence in inflating the empirical variance. In the Postsecondary population, v_{Binder} is the least biased.

In addition to calculating the average relative difference between the variance es-

Table 3.21: Average Distance from Empirical Value for Standard Error Estimators in Postsecondary Population (in thousands)

	Small Samples			Large Samples		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	201.1	223.8	215.5	32.5	29.6	27.7
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	205.9	228.5	220.6	33.2	30.2	28.3
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	181.6	202.9	196.3	28.4	26.5	24.9

timators and the empirical variance, we also calculated the percent relative root mean squared error of the variance estimators. These results are in Appendix B.7.2 which starts on page 394. The pseudoempirical maximum likelihood estimator, $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$, is more variable than the other estimators. Even though $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$ clearly is less biased than the other two estimators, the fact that it is more variable than the other estimators makes it less attractive as a variance estimator. None of the estimators is centered around the empirical variance and none of them are highly reliable in the Postsecondary population.

We now turn our attention to the Census population. In the Postsecondary population, $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$ was the least biased of the estimators. In the Census population, we see less conclusive results. In the small samples, $\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$ seems to be the least biased of the three estimators. As we see in Table 3.22, all three variance estimators underestimate the empirical variance in small samples.

Appendix B.7.3 which starts on page 414 shows estimates of the mean squared error of the three variance estimators in the Census population. Unlike in the Postsecondary population, $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$ has the smallest percent relative root mean squared error of the

Table 3.22: Relative bias of LGREG Variance Estimators for Census Population

	Small Samples					
	Fixed SRS		Rate SRS		Fixed PPS	
Estimator	Renter	Owner	Renter	Owner	Renter	Owner
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-42.2	-42.2	-44.1	-44.1	-16.3	-16.3
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-54.1	-54.1	-55.9	-55.9	-34.8	-34.8
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-60.3	-60.3	-59.0	-59.0	-20.9	-20.9
	Large Samples					
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	11.7	11.7	11.4	11.4	51.0	51.0
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-5.7	-5.7	-11.1	-11.1	2.1	2.1
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	7.5	7.5	8.8	8.8	50.0	50.0

three estimators in the small samples from the Census population.

The bias and mean squared error of the variance estimators can be used to discriminate between the variance estimators. In many applications, the variance estimators are primarily used to create confidence intervals and test hypotheses. In this respect, it is useful to measure the confidence interval coverage obtained when using the point estimator along with the variance estimator.

We calculated the empirical confidence interval coverage for all three estimators in all three populations. Tables 3.23 and 3.24 show the confidence interval coverage for the Synthetic and Postsecondary populations. See Table B.47 in Appendix B.7.3 on page 428 for the confidence interval coverage of the estimators in the Census population. In all cases, confidence intervals created using $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$ were closer to the nominal 95% level in small samples than the other two variance estimators. In the large samples, $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$ also performed quite well. In the Synthetic population, all three estimators

Table 3.23: Percent 95% Confidence Interval Coverage of LGREG Variance Estimators for Synthetic Population

	Small Samples								
	Fixed SRS			Rate SRS			Fixed PPS		
Estimator	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	89.6	91.2	87.5	87.2	88.1	86.2	90.2	91.0	89.0
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	89.2	90.5	86.7	86.3	87.4	85.7	89.8	90.4	88.3
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	92.6	92.8	91.6	91.2	91.2	90.2	92.7	93.5	92.0
Large Samples									
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	95.5	95.1	95.5	94.3	94.3	95.0	95.0	95.3	94.7
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	95.5	95.1	95.4	94.2	94.2	95.0	94.8	95.1	94.7
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	95.5	95.2	95.3	94.3	94.3	95.2	95.2	95.3	94.9

performed very close to their nominal level in the large samples. In the Postsecondary population, $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$ was also closest to the nominal level in the large samples. With the exception of the Fixed PPS samples in the Census population, $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$ was also closer to the nominal confidence interval coverage than the competing estimators. Since the sampling fraction was quite large for the Fixed PPS samples in the Census population, we would expect confidence intervals to exceed the nominal coverage.

As the sample size increases, the mean squared error of the point and variance estimators decreases. This fact indicates that our variance estimators are consistent. In terms of confidence interval coverage, this feature of our estimators means that the confidence interval coverage will get closer to the nominal level as the sample size increases (assuming the sampling fraction is small). In all three populations, the empirical confidence intervals get closer to the nominal value, providing further evidence that our variance estimators behave as we would hope. In the simulated population, we see that the larger

Table 3.24: Percent 95% Confidence Interval Coverage of LGREG Variance Estimators for Postsecondary Population

	Small Samples											
	Fixed SRS				Rate SRS				Fixed PPS			
Estimator	Math	Health	Business	Other	Math	Health	Business	Other	Math	Health	Business	Other
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	52.2	81.6	66.0	74.7	48.7	80.0	65.1	72.9	50.0	82.0	65.8	74.2
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	50.8	80.0	64.2	73.1	47.5	78.3	63.4	71.1	48.5	80.2	64.0	72.3
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	76.3	86.9	78.1	83.8	75.5	85.2	76.2	82.2	76.7	86.6	77.3	83.5
Large Samples												
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	81.4	91.4	84.8	90.4	81.1	92.0	84.5	89.7	80.9	92.6	85.8	90.8
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	80.5	90.8	83.8	89.7	80.2	91.3	83.7	89.0	80.1	92.1	85.1	90.1
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	86.3	92.6	88.0	91.9	86.5	93.2	87.1	91.5	86.8	93.4	88.7	92.4

samples of $1,500$ clusters and $3,000$ units were large enough to get close to the nominal confidence interval coverage. In the Postsecondary population, the larger samples contained only 200 units in 50 clusters. For those samples, confidence interval coverage was three to fifteen points lower than the nominal level. In the Census population, the larger samples contained 450 sampling units in 50 clusters and confidence interval coverage was up to 5 points lower than the nominal level of 95%.

In the Synthetic and Postsecondary populations, the bias of the variance estimators decreases as the sample size increases. The one exception is with the large Fixed PPS samples in the Census population. In large samples, the finite population correction factor was not included in the with-replacement and Binder estimators. In the large Fixed PPS samples, the probability of selecting some of the primary sampling units was close to one. The large sampling rate resulted in significant reductions in variance for the Fixed PPS samples. Since $\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$ and $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$ could not react to the reductions in variance due to the high sampling rate, they grossly overestimated the empirical variance. When

the finite population correction factor is large, $\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$ and $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$ are expected to overestimate the empirical variance unless a finite population correction factor is used to adjust the estimators. For the Binder estimator, this could easily be done by using an estimator of $\mathbf{\Sigma}_{\widehat{\mathbf{U}}}$ adapted to the specific design used instead of the with-replacement estimator. In the large samples, $\upsilon_e\left(\widehat{\mathbf{t}}_y^{lg}\right)$ is clearly the better estimator in terms of bias and relative root mean squared error in the Census population. In the Synthetic and Postsecondary populations, the three estimators are similar in large samples.

As expected, the variance estimators are much less variable in the large samples. All three variance estimators are similar; although, the median value of $\upsilon_e\left(\widehat{\mathbf{t}}_y^{lg}\right)$ is closer to the empirical variance than the other two variance estimators. In the larger samples, the three variance estimators sometimes overstate the variance.

Of the three variance estimators for $\widehat{\mathbf{t}}_y^{lg}$ that we compared, $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$ had the best empirical confidence interval coverage and was generally less biased than the other estimators. However, it was considerably more variable than the other estimators. In large samples where a nontrivial proportion of the sample has been selected, estimators that make use of a finite population correction factor should be used. This could be accomplished by using $\upsilon_e\left(\widehat{\mathbf{t}}_y^{lg}\right)$ or by making adjustments to $\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$ or $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$.

Even though $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$ often has better properties than the other two estimators, there is much room for better estimators. Indeed, none of the variance estimators perform especially well in small samples. In large samples all three variance estimators are about the same.

3.3.3.7 Variance Estimators of $\widehat{\mathbf{t}}_{yc}^{mc}$ and $\widehat{\mathbf{t}}_{yc}^{peM}$

In this section, we compare the four variance estimators for $\widehat{\mathbf{t}}_{yc}^{mc}$. Earlier, we proved that $\widehat{\mathbf{t}}_{y}^{peM}$ was asymptotically equivalent to $\widehat{\mathbf{t}}_{y}^{mc}$. Thus, our four variance estimators can also be used to estimate the variance of $\widehat{\mathbf{t}}_{y}^{peM}$.

We begin with several introductory remarks about the bias and consistency of the four variance estimators. Then, we investigate the variance estimators when the assisting model accurately predicts the response variable. Next, we discuss the performance of the four variance estimators when the model is less accurate. For the Postsecondary and Census populations, we focus on the confidence interval coverage and percent relative root mean squared error in the small samples followed by a similar analysis for large samples.

Tables showing the empirical relative bias of the variance estimators are in Appendix B.7 which starts on page 374. In general, the relative bias decreases as the sample size increases, suggesting that the variance estimators are asymptotically unbiased. Furthermore, the relative root mean squared error tends to decrease as the sample size increases, suggesting that the variance estimators are consistent. In all cases, the confidence interval coverage gets closer to the nominal value when the sample size increases, suggesting that asymptotically inference from $\widehat{\mathbf{t}}_{yc}^{mc}$ and any of the four variance estimators is of a high quality. Unfortunately, the rate of convergence appears to be slower than desired. Unless the assisting model fits the data very well or the sample size is very large, one should not strongly rely on the accuracy or precision of the four variance estimators.

In Appendix B.7.1 on page 374, we show results from our simulations for all four

variance estimators in the Synthetic population. In general, all four variance estimators perform similarly. In the small samples $\upsilon_g\left(\widehat{\mathbf{t}}_y^{mc}\right)$ is the least biased and the most variable of the variance estimators. It also comes the closest to the nominal confidence interval coverage rate, although 95% confidence intervals for all estimators only cover the true value between 80 and 89 percent of the time. In terms of the mean squared error $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$ performs the best. For the larger samples, all four variance estimators perform about the same and no one estimator outperforms the others.

When the assisting model does not fit the data well, results are less encouraging. In the small samples from the Census population, confidence interval coverage for all variance estimators is less than 75% for the Fixed and Rate SRS samples. In the small samples from the Postsecondary population, only $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$ has confidence interval coverage regularly above 75%, making it the best variance estimator for small samples in this population. Table 3.25 shows confidence interval coverage of the four variance estimators in the Postsecondary population.

As we see in Table 3.25, $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$ has the best confidence interval coverage in small samples. In the Postsecondary population, the other three variance estimators have extremely poor confidence interval coverage and probably should not be considered in small samples with poor assisting models. On the other hand, the mean squared error of $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$ is many times larger than the mean squared error of competing variance estimators. Even though confidence interval coverage of $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$ is quite attractive, one should expect some variance estimates to be far from the true variance in small samples.

Table 3.26 shows confidence interval coverage for all four variance estimators in the

Table 3.25: Percent 95% Confidence Interval Coverage of Finite Population Total Using Several variance Estimators of $\widehat{\mathbf{t}}_{yc}^{mc}$ for Postsecondary Population

	Small Samples											
	Fixed SRS				Rate SRS				Fixed PPS			
Estimator	Math	Health	Business	Other	Math	Health	Business	Other	Math	Health	Business	Other
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	46.7	63.0	51.5	59.4	43.5	59.2	49.1	55.3	43.8	60.6	49.0	56.6
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	45.2	61.3	49.5	57.7	42.0	57.6	47.4	53.9	42.4	58.8	47.4	54.8
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	56.4	80.7	67.9	76.6	53.8	78.3	67.2	74.4	55.9	78.8	68.0	76.0
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	78.3	89.9	82.0	87.8	78.0	88.7	82.1	87.3	79.3	90.2	82.7	88.4
	Large Samples											
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	78.8	88.6	79.8	86.5	77.5	88.6	79.5	86.0	78.1	89.7	80.8	86.8
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	77.9	87.9	78.7	85.7	76.7	87.8	78.7	85.2	77.1	89.0	80.1	86.0
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	80.9	90.3	82.5	88.5	80.0	91.0	83.3	88.7	80.2	91.5	84.2	89.3
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	83.0	90.6	83.8	89.0	82.5	91.3	84.1	88.9	82.9	92.0	85.4	90.0

Census population. Results in the small samples from this population are not consistent with findings from the Postsecondary population. In the Census population, $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$ does not have the best confidence interval coverage. Instead, $\upsilon_g\left(\widehat{\mathbf{t}}_y^{mc}\right)$ has the best confidence interval coverage. In the SRS samples it has the best confidence interval coverage. In the Fixed PPS samples, it was only 0.1 point less than the best estimator, $\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$.

Särndal et al. (1989) argue that the g-weighted variance estimator has better properties in small and moderate samples than υ_e when estimating the variance of $\widehat{\mathbf{t}}_y^{gr}$. For the model calibrated estimator, results are mixed when comparing $\upsilon_g\left(\widehat{\mathbf{t}}_y^{mc}\right)$ to $\upsilon_e\left(\widehat{\mathbf{t}}_y^{mc}\right)$. In the small samples, $\upsilon_e\left(\widehat{\mathbf{t}}_y^{mc}\right)$ has smaller mean squared error in the Synthetic and Postsecondary populations, but not in the Census population. Although we found that $\upsilon_g\left(\widehat{\mathbf{t}}_y^{mc}\right)$ was less biased in the small samples for all populations, it is considerably more variable than $\upsilon_e\left(\widehat{\mathbf{t}}_y^{mc}\right)$.

In small samples, we found mixed results in the Postsecondary and Census popu-

Table 3.26: Percent 95% Confidence Interval Coverage of Finite Population Total Using Several Variance Estimators of $\widehat{\mathbf{t}}_{yc}^{mc}$ for Census Population

	Small Samples					
	Fixed SRS		Rate SRS		Fixed PPS	
Estimator	Renter	Owner	Renter	Owner	Renter	Owner
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_{y}^{mc}\right)$	62.3	62.3	65.0	65.0	88.0	88.0
$\upsilon_{e}\left(\widehat{\mathbf{t}}_{y}^{mc}\right)$	59.7	59.7	62.2	62.2	84.8	84.8
$\upsilon_{g}\left(\widehat{\mathbf{t}}_{y}^{mc}\right)$	68.1	68.1	71.1	71.1	87.9	87.9
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_{y}^{mc}\right)$	61.1	61.1	64.0	64.0	87.7	87.7
	Large Samples					
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_{y}^{mc}\right)$	90.7	90.7	91.7	91.7	97.1	97.1
$\upsilon_{e}\left(\widehat{\mathbf{t}}_{y}^{mc}\right)$	89.0	89.0	89.3	89.3	94.0	94.0
$\upsilon_{g}\left(\widehat{\mathbf{t}}_{y}^{mc}\right)$	90.9	90.9	90.5	90.5	94.3	94.3
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_{y}^{mc}\right)$	90.6	90.6	91.7	91.7	97.1	97.1

lations. In the Postsecondary population, $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_{y}^{mc}\right)$ had the best confidence interval coverage, but had poor mean squared error properties. In the Census population, $\upsilon_{g}\left(\widehat{\mathbf{t}}_{y}^{mc}\right)$ had the best confidence interval coverage and also the lowest mean squared error.

In large samples, 95% confidence intervals based on $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_{y}^{mc}\right)$ tend to cover the true value at a rate closer to 95% compared to the other estimators. Yet, the difference in coverage rates for the four variance estimators is much smaller in the large samples than in the small samples. Given the similarity of the variance estimators in terms of confidence interval coverage, one might consider selecting a variance estimator based on the mean squared error in large samples. Table 3.27 shows the percent relative root mean squared error of the variance estimators selected from the Census population. In large samples from the Census population, $\upsilon_{g}\left(\widehat{\mathbf{t}}_{y}^{mc}\right)$ has the smallest root mean squared error.

On the other hand, $\upsilon_g\left(\widehat{\mathbf{t}}_y^{mc}\right)$ has several times larger mean squared error than $\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$ and $\upsilon_e\left(\widehat{\mathbf{t}}_y^{mc}\right)$ in the Postsecondary population. In both populations, $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$ has the highest or close to the highest mean squared error.

Table 3.27: Percent Relative Root Mean Squared Error of Variance Estimators for Census Population

	Small Samples					
	Fixed SRS		Rate SRS		Fixed PPS	
Estimator	Renter	Owner	Renter	Owner	Renter	Owner
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	119.3	119.3	112.4	112.4	76.3	76.3
$\upsilon_e\left(\widehat{\mathbf{t}}_y^{mc}\right)$	104.9	104.9	100.5	100.5	71.2	71.2
$\upsilon_g\left(\widehat{\mathbf{t}}_y^{mc}\right)$	79.6	79.6	88.6	88.6	63.4	63.4
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	120.2	120.2	112.2	112.2	76.3	76.3
	Large Samples					
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	63.5	63.5	66.2	66.2	58.3	58.3
$\upsilon_e\left(\widehat{\mathbf{t}}_y^{mc}\right)$	52.0	52.0	53.2	53.2	24.7	24.7
$\upsilon_g\left(\widehat{\mathbf{t}}_y^{mc}\right)$	41.2	41.2	42.4	42.4	22.0	22.0
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	63.4	63.4	66.1	66.1	57.8	57.8

Given the diverging results between the Postsecondary and Census populations, we conclude that none of the four variance estimator is uniformly the best. Indeed, in the large samples results are inconclusive. The variance estimators seem to behave differently in different populations and under different assisting models.

We developed four variance estimators for $\widehat{\mathbf{t}}_{yc}^{mc}$. As we saw in the previous section, none of the variance estimators performs exceptionally well. When the assisting model accurately describes the population, all four variance estimators perform similarly. When the assisting model does not fit the population very well, $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$ has the best confi-

dence interval coverage in small samples even though its mean squared error is seemingly large. In the larger samples with poor to moderate fitting assisting models, all four variance estimators have good confidence interval coverage; although, $\upsilon_e\left(\mathbf{t}_y^{mc}\right)$ or $\upsilon_g\left(\mathbf{t}_y^{mc}\right)$ tend to be less variable than the other estimators.

3.4 Conclusion

In this paper, we constructed four new point estimators of multinomial response data in clustered samples. In the process of developing these estimators, we also extended the GREG estimator for multivariate response data. Under a common asymptotic framework with regularity assumptions, we proved all four estimators are asymptotically unbiased. Additionally, we calculated the asymptotic variance of two point estimators and proved that the third point estimator was asymptotically equivalent to one of the other estimators. We also constructed with-replacement, survey weighted residual, and implicit differentiation variance estimators of the asymptotic variance for three of the logistic-assisted estimators.

Using a simulation, we compared the three new logistic-assisted point estimators to the π and GREG estimators. In terms of relative bias, we found that all of the estimators appear to be unbiased in large samples. In general, the logistic-assisted point estimators have smaller mean squared errors than the π and GREG estimators. We found strong evidence of benefits to the logistic-assisted estimators in small and large samples in a variety of populations including two public use datasets. Indeed, the logistic-assisted estimators have the potential be much more efficient than the GREG and π-estimators

and warrant further research. One disadvantage of the logistic-assisted estimators is that they can be unstable in small samples with rare characteristics.

In our simulation, we also compared the three new variance estimators of the LGREG estimator to the empirical variance of the LGREG estimator. On average, we found that confidence interval coverage of the variance estimators were close to the nominal level in large samples; although confidence intervals constructed from one sample may be much larger or smaller than what they should be. Overall, the Binder variance estimator had the best confidence interval coverage in both small and large samples; although, the mean squared error of the Binder estimator was larger than the competing estimators in some samples. Unfortunately none of the variance estimators consistently have both attractive confidence interval coverage and small mean squared error. Estimating the variance of the LGREG estimator is difficult and careful attention should be given to this topic in the future.

We also compared the four new variance estimators for the model-calibration and model-calibration maximum pseudoempirical likelihood estimators. When the assisting model accurately describes the population, all four variance estimators perform similarly. When the assisting model does not fit the population very well, $\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$ has the best confidence interval coverage in small samples even though its mean squared error is seemingly large. In the larger samples with poor to moderate fitting assisting models, all four variance estimators have good confidence interval coverage; although, $\upsilon_e\left(\mathbf{t}_y^{mc}\right)$ or $\upsilon_g\left(\mathbf{t}_y^{mc}\right)$ tend to be less variable than the other estimators. In general, we found mixed results and recommend further research on adjustments to the variance estimators as well as replication variance estimators.

In conclusion, the logistic-assisted point estimators have many advantages over the GREG estimator for categorical data in clustered samples, even when the assisting model fit is less than ideal. We presented several variance estimators for the logistic-assisted estimators, although more research is needed to improve variance estimation of the logistic-assisted point estimators.

Chapter 4

Design-based Inference Assisted by Generalized Linear Models in Cluster Samples

4.1 Introduction

As already noted in the third chapter, GREG is a powerful and widely used estimation technique, but in some situations can be improved with assisting models that fit the data better than the classic linear model. Whereas in Chapter 3, we focused on logistic regression; in this chapter, we broaden our scope to a powerful family of assisting models called Generalized Linear Models (GLMs). GLMs can be used to model any variable whose distribution is a member of the exponential dispersion family. Since this family includes the normal, Bernoulli, binomial, multinomial, Poisson, and negative binomial distributions, it is a versatile family used to model continuous, binary, and count data. Linear, logistic, probit, complementary-log-log, Poisson, and negative binomial regression are all examples of GLMs.

This chapter provides the theory needed to use GLMs to assist design-based estimation in cluster samples. Such research has the potential to produce more precise estimates than GREG estimators and thereby increase the quality of estimates and improve hypothesis tests.

This chapter generalizes many parts of Chapter 3. Furthermore, like the previous

chapters, this chapter was written to be a self-contained article. For this reason, some parts of the previous chapter are repeated. Readers of this chapter who have already read Chapter 3 are encouraged to be patient and understanding of the overlap between these two chapters.

Unlike the previous chapter, the focus of this chapter is limited to the case where an arbitrary function of a scalar response variable is linear in auxiliary variables. That is, we consider assisting models where $g(y) = \mathbf{X}\boldsymbol{\beta}$. Chapter 3 was restricted to the logit link function. The estimators in this chapter should be useful for people who may prefer to use assisting log, probit, or complementary log-log models. The log link is often used for count and rate data because it is the canonical link for the Poisson distribution. Probit regression is preferred in some disciplines because if of its relationship to the normal distribution. In practice, probit and logistic regression are often very similar, although the logistic distribution has slightly less mass in the tails. When the data are skewed, the complementary log-log model is often preferred because it is not symmetric around the mean. Also, cauchit models have heavier tails than the logit link function, so they can be used to model binomial data when the probability of success is quite variable among units or at the extremes.

Current estimators of totals from clustered samples are not well suited for categorical data. For example, the GREG assisting model is based on a linear model which may not fit binary data as well as a probit or log-log model. One of the key characteristics of binary data is that the response options are bounded between 0 and 1. Linear models do not preserve this important characteristic of binary data. When estimating the proportion of persons who are employed, the linear assisting model may produce negative rates or

rates over 100%. The implied predictions for individual elements may also be outside the range $[0, 1]$ when using a linear assisting model. Models specifically built to analyze binary data can improve point estimation and reduce sampling errors. In fact, assisting models that fit the data well generally result in estimators that have lower sampling variance than estimators based on poorly fit assisting models.

Data collected in multiple stages is also common. In an effort to reduce travel and other field costs, multiple-staged samples are generally selected in large face to face surveys. However, the analysis of clustered data is frequently more complicated than data collected in a single stage.

The sample design impacts data analysis, estimation, and inference. If the sample design is not taken into account, point estimators, variance estimators, and test statistics may be misleading. For this reason, estimators based on single-staged samples are rarely appropriate for multi-staged sample designs. Clustered samples also differ from single-staged samples in the level of data that may be available. Auxiliary data may be available at the unit level, at the cluster level, at both the cluster and unit level, or not at all. The level of covariates and whether they are available for the sample only or for the full population also impacts how one constructs estimators.

In this paper, we present the case where auxiliary data are available for all units in the population. We call this the case of complete unit auxiliaries. Auxiliary data are often available for all units in the population. Address based sampling frames, national population registers, marketing databases, and professional organizations often contain a wealth of data about all units on the sampling frame. When such data are available, it is often advantageous to calibrate sample totals to known population totals. Calibrated

estimators often have lower nonsampling and sampling errors when compared to simpler estimators.

Previous research has focused on calibrated GLM-assisted point estimators in single-staged samples. In this chapter, we extend these results to two-staged samples and construct variance estimators appropriate for two-staged samples. We develop and compare three different kinds of GLM-assisted point estimators: the generalized difference estimator, the model-calibration estimator, and the model-calibrated maximum pseudoempirical likelihood estimator. We also propose several variance estimators for these estimators.

4.1.1 Generalized Linear Models

Nelder and Wedderburn (1972) first introduced generalized linear models (GLMs) and provided many of the necessary details needed to estimate their parameters. Since then, numerous textbooks and papers have devoted much attention to the model fitting, parameter estimation, and application of GLMs (see Agresti (2002), Bishop et al. (2007), McCullagh and Nelder (1999), McCulloch and Searle (2004), and Shao (2003)).

One of the reasons GLMs are so popular is that they can be tailored to fit a variety of categorical response variables. For example, when modeling a percent, one might want the fitted value to be bounded between 0% and 100%. Unfortunately, the fitted values from linear regression are unbounded. In this case, a nonlinear function that bounds the fitted values between 0% and 100% may be more appropriate. GLMs provide an alternative set of models that may better suit situations where the linear regression assumptions may not hold.

4.1.1.1 Likelihood

In this review of GLMs, we primarily draw upon the notation and logic of Shao (2003, sec 4.4). We begin with a scalar response for the kth sample unit.

We take the model-based framework where the kth observation is a random element drawn from some density function or probability mass function denoted by $f\left(y_k; \eta_k, \phi_k\right)$. In this case, η_k is an unknown superpopulation parameter unique for the kth unit. It is called the *natural parameter*. And ϕ_k is an unknown scalar-valued superpopulation parameter called the *dispersion parameter*.

Many common densities can be written as a member of the *exponential dispersion family*, which is defined as

$$f\left(y_k; \eta_k, \phi_k\right) = e^{\frac{[y_k \eta_k - \zeta(\eta_k)]}{\phi_k} + h(y_k, \phi_k)} \tag{4.1}$$

Table 4.1 shows how several common distributions can be written in terms of the exponential family. As we see, the exponential dispersion family covers a wide variety of popular distributions including the normal, Bernoulli, binomial, and Poisson distributions. The normal distribution is often used to model continuous data, the Bernoulli distribution is often used to model binary data, the binomial distribution is often used to model percent data, and the Poisson distribution is often used to model count and rate data. Indeed, the natural dispersion family covers a wide range of modeling possibilities.

Although GLMs were developed to model response data generated by a member of the exponential dispersion family, in many situations GLMs are used to model data regardless of the underlying superpopulation model generating the data.

If a sample of size n is selected and the units are independent of each other, then

Table 4.1: Distributions of the Exponential Family

Name	Density	η_k	$\zeta(\eta_k)$	ϕ_k	$h(y_k, \phi_k)$
Normal	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_k-\mu)^2}{2\sigma^2}}$	μ	$\frac{\mu^2}{2}$	σ^2	$-\frac{1}{2}\left[\frac{y_k^2}{\sigma^2} + \ln(2\pi\sigma^2)\right]$
Bernoulli	$\pi_k^{y_k}(1-\pi_k)^{1-y_k}$	$\ln\left(\frac{\pi_k}{1-\pi_k}\right)$	$\ln(1+e^{\eta_k})$	1	0
Binomial	$\begin{pmatrix} z_k \\ z_k p_k \end{pmatrix} \pi_k^{z_k p_k}(1-\pi_k)^{z_k - z_k p_k}$	$\ln\left(\frac{\pi_k}{1-\pi_k}\right)$	$\ln(1+e^{\eta_k})$	$\frac{1}{z_k}$	$\ln\begin{pmatrix} z_k \\ z_k p_k \end{pmatrix}$
Poisson	$\frac{e^{-\mu_k}\mu_k^{y_k}}{y_k!}$	$\ln\mu_k$	e^{η_k}	1	-ln $(y_k!)$
Gamma	$\frac{1}{\nu^\mu}\frac{1}{\Gamma(\mu)} y_k^{\mu-1} e^{-\frac{y_k}{\nu}}$	$\frac{1}{\mu_k}$	$-\ln(-\eta_k)$	ν^{-1}	$\nu\ln(\nu y_k) - \ln(y_k) - \ln[\Gamma(\nu)]$
Inverse Gaussian	$\left[\frac{\sigma^2}{2\pi y_k^3}\right]^{\frac{1}{2}} e^{\frac{-\sigma^2(y_k-\mu)^2}{2\mu^2 y_k}}$	$\frac{1}{\mu_k^2}$	$-(-2\eta_k)^{\frac{1}{2}}$	σ^2	$-\frac{1}{2}\left[\ln(2\pi\phi y_k^3) + \frac{1}{\phi y_k}\right]$

we can write the joint density as,

$$f(\mathbf{y}; \boldsymbol{\eta}, \boldsymbol{\phi}) = \prod_{k=1}^{n} f(y_k; \eta_k, \phi_k)$$

Moreover, we define the likelihood as

$$\begin{aligned} L &= f(\boldsymbol{\eta}, \boldsymbol{\phi}; \mathbf{y}) \\ &= \prod_{k=1}^{n} f(\eta_k, \phi_k; y_k) \\ &= \prod_{k=1}^{n} L_k \end{aligned}$$

Ideally, we would like to estimate η_k and ϕ_k for each sample unit; however, we only have n realizations of Y_k to estimate both sets of parameters. One solution to this problem is to assume that all elements in the sample have the exact same density. That is, if we assume that $\eta_k = \eta$ and $\phi_k = \phi$ for all units, then we can easily use maximum likelihood or quasi-maximum likelihood to solve for η and ϕ. Once we estimate these parameters, we can easily compute various characteristics of our distribution. Many research questions and problems can be solved in this manner.

4.1.1.2 Link Functions

Often, assuming that η_k and ϕ_k are the same for all units is too restrictive. An alternative and more flexible approach is to use auxiliary data to model each η_k. This approach leads to the generalized linear model. Suppose we have a p-dimensional vector of covariates for the kth unit, denoted $\mathbf{x}_k$. Let $\boldsymbol{\beta}$ be the p-dimensional vector of coefficients.

We relate a linear combination of our covariates, denoted $\boldsymbol{\gamma}_k = \boldsymbol{\beta}^\top \mathbf{x}_k$, to $\boldsymbol{\eta_k}$ by way of a link function. That is,

$$\begin{aligned}\eta_k &= (g \circ \mu)^{-1} \left(\boldsymbol{\beta}^\top \mathbf{x}_k\right) \\ &= \mu^{-1} \left(g^{-1} \left(\boldsymbol{\beta}^\top \mathbf{x}_k\right)\right).\end{aligned}$$

Also

$$\gamma_k = g\left(\mu_k\right) = \boldsymbol{\beta}^\top \mathbf{x}_k$$

where g is called the *link* function and μ_k is the mean function. Table 4.2 shows some other common link functions. In general, the mean function is

$$\mu_k\left(\eta_k\right) = E_M\left(Y_k\right) = \frac{\partial \zeta\left(\eta_k\right)}{\partial \eta_k}.$$

And the model variance of Y_k, denoted $\boldsymbol{\Sigma}\left(\eta_k, \phi_k\right)$, is

$$\boldsymbol{\Sigma}\left(\eta_k, \phi_k\right) = \operatorname{var}_M\left(Y_k\right) = \phi_k \frac{\partial^2 \zeta\left(\eta_k\right)}{\partial \eta_k \partial \eta_k}.$$

If $g\left(\mu_k\right) = \eta_k$, then g is called the *canonical* link. In this case g and μ are inverse functions and $\eta_k = \boldsymbol{\beta}^\top \mathbf{x}_k$.

In addition to modeling η_k with covariates, we also reduce the number of dispersion parameters we need to estimate. In his mathematical statistics book, Shao (2003)

Table 4.2: Common Link Functions. Φ is the cumulative normal distribution function and $\mathscr{C}$ is the cumulative cauchy distribution function.

Link Function	η_k	μ_k
Cauchit	$\eta_k = \mathscr{C}^{-1}\left(\mathbf{x}_k^\top \mathbf{B}\right)$	$\mu_k = \mathscr{C}\left(\mathbf{x}_k^\top \mathbf{B}\right)$
Complementary Log-Log	$\eta_k = -\ln\left[-\ln\left(1-\mu_k\right)\right]$	$\mu_k = 1 - e^{-e^{\mathbf{x}_k^\top \mathbf{B}}}$
Identity	$\eta_k = \mu_k$	$\mu_k = \mathbf{x}_k^\top \mathbf{B}$
Inverse Square	$\eta_k = \frac{1}{\mu_k^2}$	$\mu_k = \frac{1}{\sqrt{\mathbf{x}_k^\top \mathbf{B}}}$
Log	$\eta_k = \ln\left(\mu_k\right)$	$\mu_k = e^{\left(\mathbf{x}_k^\top \mathbf{B}\right)}$
Logit	$\eta_k = \frac{\ln(\mu_k)}{1-\mu_k}$	$\mu_k = \frac{e^{\mathbf{x}_k^\top \mathbf{B}}}{1+e^{\mathbf{x}_k^\top \mathbf{B}}}$
Log-Log	$\eta_k = -\ln\left[-\ln\left(\mu_k\right)\right]$	$\mu_k = e^{-e^{-\mathbf{x}_k^\top \mathbf{B}}}$
Probit	$\eta_k = \Phi^{-1}\left(\mathbf{x}_k^\top \mathbf{B}\right)$	$\mu_k = \Phi\left(\mathbf{x}_k^\top \mathbf{B}\right)$
Reciprocal	$\eta_k = \frac{1}{\mu_k}$	$\mu_k = \frac{1}{\mathbf{x}_k^\top \mathbf{B}}$

proposes assuming that $\phi_k = \frac{\phi}{\omega_k}$ for some known scaling factor ω_k. Thus, we only need to estimate one value of ϕ; but the dispersion can vary from unit to unit through ω_k. In this dissertation, we do not consider models that involve estimating ϕ_k. Extending the results in this dissertation to models with variable dispersion parameters may be a fruitful area for future research.

4.1.1.3 Parameter Estimation

By modeling η_k with covariates and estimating a common dispersion parameter, we can use maximum likelihood estimation or quasi-maximum likelihood estimation to form estimating equations which are numerically solved for $\boldsymbol{\beta}$. Shao (2003) shows that if a density or probability mass function is a member of the exponential dispersion family, then the maximum likelihood of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi)$ can be found by maximizing the log-likelihood

functions,

$$\ell(\boldsymbol{\beta}, \phi) = \ln L = \sum_{k \in \mathscr{U}} \left[\ln \left[h\left(y_k, \frac{\phi}{\omega_k}\right) \right] + \frac{\psi\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right) y_k - \zeta\left(\psi\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right)\right)}{\frac{\phi}{\omega_k}} \right] \tag{4.2}$$

where

$$\psi\left(\mathbf{x}^\top \boldsymbol{\beta}\right) = \mu^{-1}\left(g^{-1}\left(\mathbf{x}^\top \boldsymbol{\beta}\right)\right)$$

$$\mu(\eta) = \zeta'(\eta).$$

Many numerical optimizers can be used to find the values of $\boldsymbol{\beta}$ and ϕ that maximize ℓ.

Another method to maximize ℓ would be to simultaneously differentiate ℓ with respect to $\boldsymbol{\beta}$ and ϕ. We call the derivatives of the log-likelihood our estimating equations, denoted $w(\boldsymbol{\beta})$. Setting them equal to zero and solving for $\boldsymbol{\beta}$ and ϕ gives us our maximum likelihood estimators $\widehat{\boldsymbol{\beta}}$ and $\widehat{\phi}$. According to Shao (2003, p. 281), the estimating equations for $\boldsymbol{\beta}$ are

$$\begin{aligned}
w(\boldsymbol{\beta}) &= \frac{\partial \ell}{\partial \boldsymbol{\beta}} \\
&= \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{k=1}^{n} \left[\ln \left[h\left(y_k, \frac{\phi}{\omega_k}\right) \right] + \frac{\psi\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right) y_k - \zeta\left(\psi\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right)\right)}{\frac{\phi}{\omega_k}} \right] \\
&= \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{k=1}^{n} \left[\frac{\psi\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right) y_k - \zeta\left(\psi\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right)\right)}{\frac{\phi}{\omega_k}} \right] \\
&= \frac{1}{\phi} \sum_{k=1}^{n} \omega_k \frac{\partial}{\partial \boldsymbol{\beta}} \left[\psi\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right) y_k - \zeta\left(\psi\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right)\right) \right]
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial \psi\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right) y_k}{\partial \boldsymbol{\beta}} &= \left(\frac{\partial \gamma_k}{\partial \boldsymbol{\beta}^\top}\right) \frac{\partial \psi(\gamma_k)}{\partial \gamma_k} y_k \\
&= \mathbf{x}_k^\top \frac{\partial \psi(\gamma_k)}{\partial \gamma_k} y_k
\end{aligned}$$

and

$$\frac{\partial}{\partial \boldsymbol{\beta}}\left[\zeta\left(\psi\left(\boldsymbol{\beta}^{\top}\mathbf{x}_k\right)\right)\right] = \left[\frac{\partial \gamma_k}{\partial \boldsymbol{\beta}}\right]\left[\frac{\partial \psi\left(\gamma_k\right)}{\partial \gamma_k^{\top}}\right]\frac{\partial \zeta\left(\psi\left(\boldsymbol{\beta}^{\top}\mathbf{x}_k\right)\right)}{\partial \psi}$$
$$= \mathbf{x}_k\left[\frac{\partial \psi\left(\gamma_k\right)}{\partial \gamma_k^{\top}}\right]\mu_k\left(\psi\left(\boldsymbol{\beta}^{\top}\mathbf{x}_k\right)\right).$$

Thus, our estimating equestions for $\boldsymbol{\beta}$ are

$$w\left(\boldsymbol{\beta}\right) = \frac{1}{\phi}\sum_{k\in\mathscr{U}}\left\{\left[y_k - \mu_k\left(\psi\left(\boldsymbol{\beta}^{\top}\mathbf{x}_k\right)\right)\right]\left[\frac{\partial \psi\left(\gamma_k\right)}{\partial \gamma_k}\right]\omega_k\mathbf{x}_k\right\}. \tag{4.3}$$

Recall that $g\left(\mu\left(\eta_k\right)\right) = \boldsymbol{\beta}^{\top}\mathbf{x}_k$. Thus, $\eta_k = \mu^{-1}\left(g^{-1}\left(\boldsymbol{\beta}^{\top}\mathbf{x}_k\right)\right)$, which is equal to $\psi\left(\boldsymbol{\beta}^{\top}\mathbf{x}_k\right)$. So, $\mu\left(\psi\left(\boldsymbol{\beta}^{\top}\mathbf{x}_k\right)\right) = \mu\left(\eta_k\right) = \mu_k$. Therefore, we can simplify out estimating equations to

$$w\left(\boldsymbol{\beta}\right) = \frac{1}{\phi}\sum_{k\in\mathscr{U}}\left\{\left[y_k - \mu_k\right]\left[\frac{\partial \psi\left(\gamma_k\right)}{\partial \gamma_k}\right]\omega_k\mathbf{x}_k\right\}.$$

The unsimplified derivative, $\left[\frac{\partial\psi(\gamma_k)}{\partial\gamma_k}\right]$, depends on the link function. We can write $\frac{\partial\psi(\gamma_k)}{\partial\gamma_k}$ as $\frac{\partial\eta_k}{\partial\gamma_k} = \frac{\partial\eta_k}{\partial\mu_k}\frac{\partial\mu_k}{\partial\gamma_k}$. By definition $\frac{\partial\mu_k}{\partial\eta_k} = \frac{\mathrm{var}(y_k)}{\phi_k}$. Thus, we further simplify to

$$w\left(\boldsymbol{\beta}\right) = \frac{1}{\phi}\sum_{k\in\mathscr{U}}\left\{\left[y_k - \mu_k\right]\left[\frac{\phi_k}{\mathrm{var}\left(y_k\right)}\frac{\partial \mu_k}{\partial \gamma_k}\right]\omega_k\mathbf{x}_k\right\}. \tag{4.4}$$

The sample weighted version of Equation (4.4) gives the pseudomaximum likelihood estimating equations

$$\widehat{w}\left(\boldsymbol{\beta}\right) = \frac{1}{\phi}\sum_{k\in\mathfrak{s}} d_k\left\{\left[y_k - \mu_k\right]\left[\frac{\phi_k}{\mathrm{var}\left(y_k\right)}\frac{\partial \mu_k}{\partial \gamma_k}\right]\omega_k\mathbf{x}_k\right\}. \tag{4.5}$$

When g is the canonical link function, we can further simplify our estimating equations to

$$\widehat{w}\left(\boldsymbol{\beta}\right) = \frac{1}{\phi}\sum_{k\in\mathfrak{s}} d_k\left\{\left[y_k - \mu_k\right]\omega_k\mathbf{x}_k\right\}. \tag{4.6}$$

If we have a dispersion parameter to estimate, the estimating equation for ϕ is

$$\frac{\partial \ell}{\partial \phi} = \sum_{k \in \mathscr{U}} \left\{ \frac{\partial \ln h\left(y_k, \frac{\phi}{\omega_k}\right)}{\partial \phi} - \frac{\omega_k \left[\psi\left(\mathbf{x}_k^\top \boldsymbol{\beta}\right) y_k - \zeta\left(\psi\left(\mathbf{x}_k^\top \boldsymbol{\beta}\right)\right)\right]}{\phi^2} \right\}.$$

Setting the estimating equations equal to zero and solving for our parameters gives us the maximum likelihood estimators. To determine if we have the maximum or minimum, we must consider the second derivative,

$$\frac{\partial^2 \ell_k}{\partial \eta_k \partial \eta_k} = \frac{-\zeta''\left(\eta_k\right)}{\phi}.$$

Specifically, the second derivative must be negative. Using our sample, the solution to the estimating equations is $\widehat{\beta}$ [1].

In Appendix C on page 437, we give two examples of GLMS. In Appendix C.1 on page 437, we derive estimating equations for a Poisson random variable with a log link function. Then, in Appendix C.2 on page 439, we derive estimating equations for a Bernoulli random variable with a probit link function.

Residuals play an important part of evaluating the fit of models. Although GLMs can be fit to many different kinds of data, GLM-assisted estimators will perform best if the model fits the data well. Like linear models, residuals can be used to assess the fit of GLMs. However, the form of the residuals is slightly different from GLMs than for linear regression. Appendix C.3 introduces several different residuals for GLMs.

Since Nelder and Wedderburn (1972), many statisticians have extended and adapted the theory of GLMs. Of chief importance to the field of sampling statistics, Binder (1983) described how finite population parameters, such as $\mathbf{B}$, could be estimated from sam-

[1] Unfortunately, sometimes the solution to the estimating equations gives a value outside the range of possible values. These "boundary" cases have been well studied and documented.

ple data. His results apply to model parameters, including dispersion parameters, of all GLMs. Also, Firth and Bennett (1998) explored the calibration properties of projective and predictive estimators constructed using the pseudomaximum likelihood estimates of $\mathbf{B}$. Lehtonen and Pahkinen (2004), and others have explored many of the properties of design-based estimators for the coefficients of generalized linear models from complex survey data.

4.1.1.4 Summary

Rather than assuming that all units in the population were generated by one set of parameters, such as a common mean and variance, GLMs use auxiliary data to model the population parameters. This allows each unit to have a different mean and variance structure which can lead to increased flexibility and model fit over linear models.

4.1.2 Estimation of Totals for Categorical Data in Poisson Samples

In this section, we review several model-assisted point estimators that can be used to estimate finite population totals. Our review introduces the estimators presented in Chapter 3, with the exception that we have a scalar response instead of a multivariate response. We briefly review the π and GREG estimators. Then, we introduce the projective estimator and the generalized difference estimator. We then discuss three different types of calibration estimators. We begin with the traditional calibration estimator and then discuss the model-calibration estimator. We conclude with two model-calibrated maximum pseudoempirical likelihood estimators. Table 4.3 shows the estimators that follow.

Table 4.3: Point Estimators

Statistic	Description
$\hat{t}_y^{\pi}$	π-Estimator
$\hat{t}_y^{pr}$	Projective Estimator / Regression Estimator
$\hat{t}_y^{gd}$	Generalized Difference Estimator
$\hat{t}_y^{mc}$	Generalized Model-Calibration Estimator
$\hat{t}_y^{peM}$	Genearlized Pseudo-Empirical Maximum Likelihood Estimator using M
$\hat{t}_y^{pe\hat{M}}$	Genearlized Pseudo-Empirical Maximum Likelihood Estimator using $\widehat{M}$

4.1.2.1 The π Estimator

In section 1.1.5.2 on page 16, we introduced the π-estimator

$$\hat{t}_y^{\pi} = \sum_{k \in \mathfrak{s}} \frac{y_k}{\pi_k}$$

with variance in single-staged samples

$$\text{var}\left(\hat{t}_y^{\pi}\right) = \sum_{k \in \mathscr{U}} \sum_{l \in \mathscr{U}} \Delta_{kl} \frac{y_k}{\pi_k} \frac{y_l}{\pi_l}.$$

The π-estimator is design-unbiased and simple to compute. However, the variability of this estimator from sample to sample tends to be larger than competing estimators that make use of covariates, especially for small and moderate-sized samples. Also, compared to other estimators, the π-estimator does not have the calibration property, a very important property for official statistics. Thus, the π-estimator is not preferred over alternative estimators, such as the Generalized Difference Estimator.

4.1.2.2 Projective Estimator

Perhaps the simplest design-consistent point estimator, the projective estimator is simply the sum of predictions for the complete population. Firth and Bennett (1998)

define the projective estimator as

$$t_y^{pr} = \sum_{k \in \mathscr{U}} a_k \widehat{y}_k \tag{4.7}$$

where a_k is known for the full population prior to sampling and $\widehat{y}_k$ are predictions. For a GLM, we define the projective estimator as

$$t_y^{pr} = \sum_{k \in \mathscr{U}} \widehat{\mu}_k \tag{4.8}$$

where $\widehat{\mu}_k = \mu\left(\mathbf{x}_k, \widehat{\mathbf{B}}\right)$ is a prediction based on a GLM.

The projective estimator is not universally design-consistent. Firth and Bennett (1998) provide some conditions that make the projective estimator developed from a GLM design-consistent. Specifically, they argue that an estimator of a finite population total will be design-consistent if the model and model fitting procedure are correctly aligned. For example, a GLM with a canonical link fitted with maximum likelihood estimation will be design-consistent. In this chapter, we primarily focus on non-canonical links. In that case, Firth and Bennett (1998) show that a sufficient condition for design-consistency is

$$\sum_{k \in \mathfrak{s}} \frac{1}{\pi_k} a_k \left(y_k - \widehat{\mu}_k\right) = 0 \tag{4.9}$$

for all possible samples. In the GLMs presented in this paper, the estimating equations for $\mathbf{B}$ do not always simplify to Equation (4.9).

4.1.2.3 GLM-Assisted Difference Estimator

For finite population prediction, the estimated coefficients can be used to predict values of μ_k for all units in the population, as long as the covariates are available for all

population units. From the model-based framework, these fitted values can be used to construct projective and predictive estimates of finite population totals as long as the explanatory variables are known for all units in the population. Valliant (1985) and Valliant et al. (2000) discuss the nonlinear predictive estimator in single-staged samples. In many samples, projective estimators are equivalent to the generalized difference estimator.

For the generalized difference estimator, the projective estimator is adjusted based on weighted residuals. Equivalently, the generalized difference estimator can be thought of as the π-estimator with an adjustment based on the difference between the projective total and the weighted total. Wu and Sitter (2001) defined the generalized difference estimator in single stage samples as

$$\hat{t}_y^{gd} = \sum_{k \in \mathscr{U}} \widehat{\mu}_k + \sum_{k \in \mathfrak{s}} d_k \left(y_k - \widehat{\mu}_k \right) \tag{4.10}$$

where $\widehat{\mu}_k$ is an estimate of $E_M \left(y_k | \mathbf{x}_k, \mathbf{B} \right)$ under some working model. For example, $\widehat{\mu}_k$ could be an estimate from a linear, logistic, or nonparametric model. For single-staged samples, Wu and Sitter (2001) proved that $\hat{t}_y^{gd}$ is a design-consistent estimator for t_y with asymptotic variance

$$\text{av} \left(\hat{t}_y^{gd} \right) = \sum_{k \in \mathscr{U}} \sum_{l \in \mathscr{U}} \Delta_{kl} \left(\frac{y_k - \mu_k}{\pi_k} \right) \left(\frac{y_l - \mu_l}{\pi_l} \right)$$

where $\mu_k = E_M \left(y_k | \mathbf{x}_k, \mathbf{B} \right)$. With a sample, the asymptotic variance can be estimated by

$$\upsilon_e \left(\hat{t}_y^{gd} \right) = \sum_{k \in \mathfrak{s}} \sum_{l \in \mathfrak{s}} \frac{\Delta_{kl}}{\pi_{kl}} \left(\frac{y_k - \widehat{\mu}_k}{\pi_k} \right) \left(\frac{y_l - \widehat{\mu}_l}{\pi_l} \right). \tag{4.11}$$

When using a linear working model, $\widehat{\mu}_k = \mathbf{x}_k^\top \widehat{\mathbf{B}}$, Equation (4.10) reduces to the Generalized REGression (GREG) estimator

$$\hat{t}_y^{gr} = \widehat{t}_y^{\pi} + \left(\mathbf{t}_x - \widehat{\mathbf{t}}_x^{\pi} \right) \widehat{\mathbf{B}}$$

where

$$\widehat{\mathbf{B}} = \left(\mathbf{X}^\top \mathbf{\Pi}^{-1} \mathbf{Q} \mathbf{X}\right)^{-1} \left(\mathbf{X}^\top \mathbf{\Pi}^{-1} \mathbf{Q} \mathbf{y}\right).$$

and $\mathbf{Q}$ is a matrix determined by the analyst, often set to the identity matrix, $\mathbf{I}$, or a diagonal matrix with elements equal to $\frac{1}{\sigma^2}$ or $\frac{1}{\sigma_k^2}$. In section 1.2.1 on page 33, we introduce the GREG and discuss properties of this estimator.

Särndal (1980a) showed that the GREG estimator was a design-consistent estimator and that Equation (4.11) could be used to estimate the asymptotic variance of the GREG estimator. Although the GREG estimator is approximately design-unbiased, it is motivated by a linear relationship between the response variable and the covariates. Even when the linear model assumptions are violated, the GREG estimator is still design-consistent. Moreover, it results in one general set of calibrated weights that can be used for a variety of dependent variables. Lastly, it does not require auxiliary information for the complete frame. It only requires covariates for sample units and control totals for the population. Properties of GREG estimators have been discussed for complex survey designs. Despite these benefits, there may be more efficient estimators that use more appropriate models when dealing with data that doesn't easily fit a linear model.

The generalized difference estimator is a broad class of design-consistent estimators that includes the GREG estimator. It has many advantages over the π-estimator. In Appendix C.4 on page 446, we prove that the generalized difference estimator with a GLM-assisting model is design-unbiased when the sample was selected from a clustered design. Furthermore, in Appendix C.4.2 we derive the asymptotic variance of the generalized difference in clustered designs.

4.1.2.4 Calibrated Estimator

In section 1.2.3 on page 48, we discussed the construction and properties of calibration estimators. In calibration, a new set of weights is found my minimizing the distance between the base weights and the new set of weights, subject to calibration constraints, usually $\sum_{k\in\mathfrak{s}} w_k^{cal}\mathbf{x}_k = \mathbf{t}_x$. The calibrated weights depend on how one specifies the "distance" between the design weights and the calibrated weights.

If a linear distance is used, then Deville and Särndal (1992) showed that the calibration estimator is equivalent to the GREG estimator. For this reason, we only consider the GREG estimator in this chapter.

4.1.2.5 Model-Calibrated Estimator

Wu and Sitter (2001) extended calibration to cover nonlinear assisting models. They call their method, model-calibration. Instead of minimizing the distance between $\mathbf{d}$ and $\mathbf{w}^{cal}$ subject to $\sum_{k\in\mathfrak{s}} w_k^{cal}\mathbf{x}_k = \mathbf{t}_x$, they proposed minimizing the distance between $\mathbf{d}$ and $\mathbf{w}_k^{mc}$ subject to $\frac{1}{N}\sum_{k\in\mathfrak{s}} w_k^{mc} = 1$ and $\sum_{k\in\mathfrak{s}} w_k^{mc}\widehat{\mu}_k = \sum_{k\in\mathscr{U}} \widehat{\mu}_k$. After solving for w_k^{mc}, the model-calibrated estimator is explicitly written as

$$\hat{t}_y^{mc} = \widehat{\mathbf{t}}_y^{\pi} + \left(\sum_{k\in\mathscr{U}} \widehat{\mu}_k - \sum_{k\in\mathfrak{s}} d_k\widehat{\mu}_k\right)\widehat{\mathbf{B}}^{mc} \tag{4.12}$$

where

$$\widehat{\mathbf{B}}^{mc} = \frac{\sum_{k\in\mathfrak{s}} d_k q_k \left(\widehat{\mu}_k - \bar{\mu}\right)\left(y_k - \bar{y}\right)}{\sum_{k\in\mathfrak{s}} d_k q_k \left(\widehat{\mu}_k - \bar{\mu}\right)^2} \tag{4.13}$$

$$\bar{\mu}_c = \frac{\sum_{k\in\mathfrak{s}} d_k q_k \widehat{\mu}_k}{\sum_{k\in\mathfrak{s}} d_k q_k}. \tag{4.14}$$

In Section 1.2.4 on page 50, we review more details about the model-calibrated estimator.

One advantage of the model-calibrated estimator is that it can improve design-based inference by using nonlinear models. Since GLMs tend to fit data generated by nonlinear models better than linear regression, it seems advantageous to use model-calibration when analyzing nonlinear data. In this dissertation, general model-calibration is developed for two-stage samples. Kim et al. (2009) discuss nonparametric calibration in cluster samples, but they do not cover nonlinear models.

Of course there are some disadvantages to general model-calibration. First, complete data are needed for all sample and nonsample units. Frames rich in auxiliary data are becoming more popular with address based sampling frames, but such frames are not always available. Furthermore, even when they exist, complete data frames are not always up-to-date, accurate, or contain variables useful for modeling. Second, model-calibration results in a new set of weights for each response variable. In large multipurpose surveys, one set of calibrated weights that can be used for all response variables is preferred. Model-calibrated weights are not general and each response variable requires a different set of weights. Finally, even though predictions of μ_k are often bounded, there is no guarantee $\widehat{t}_y^{mc}$ will be bounded. Thus, some estimates of $\widehat{t}_{yc}^{mc}$ could be negative or larger than possible.

4.1.2.6 Model-Calibrated Maximum Pseudoempirical Likelihood Estimator

Rao and Wu (2009) review the history and motivation of empirical likelihood methods. The pseudoempirical likelihood approach is motivated by treating y_k in the popula-

tion as a random variable with density of p_k^{pe}. In section, 1.2.5 on page 52, we introduce the model-calibrated maximum pseudoempirical likelihood estimator. According to Wu and Sitter (2001), the model-calibrated maximum pseudoempirical likelihood estimator for a finite population mean in single-staged samples with a GLM assisting model is

$$\widehat{\bar{y}}^{pe} = \sum_{k \in \mathfrak{s}} \hat{p}_k^{pe} y_k \tag{4.15}$$

where $\hat{p}_k^{pe}$ is found my maximizing

$$\hat{\ell}\left(\mathbf{p}^{pe}\right) = \sum_{k \in \mathfrak{s}} d_k \log p_k^{pe} \tag{4.16}$$

subject to

$$\sum_{k \in \mathfrak{s}} p_k^{pe} = 1 \tag{4.17}$$

$$\sum_{k \in \mathfrak{s}} p_k^{pe} \mathbf{u}_k = 0 \tag{4.18}$$

where

$$\mathbf{u}_k = \mu_k - \frac{1}{N} \sum_{k \in \mathscr{U}} \mu_k$$

To date, estimators of totals using the model-calibrated pseudoempirical likelihood method have not been discussed in the literature, although Sitter and Wu (2002) discuss totals of quadratic functions.

Wu and Sitter (2001) showed that $\widehat{\bar{y}}^{pe}$ is asymptotically equivalent to $\widehat{\bar{y}}^{mc}$. Therefore, the variance of $\widehat{\bar{y}}^{pe}$ could be estimated with $\upsilon_e\left(\widehat{\bar{y}}^{mc}\right)$, although Wu and Sitter (2001) recommend using the jackknife variance estimator.

One advantage of the model-calibration maximum pseudoempirical likelihood method is that the weights p_k^{pe} are forced to be positive. On the other hand, this method requires complete data and every response variable will need a different p_k^{pe} weight.

4.2 Main Results

In this section, we extend the GLM-assisted estimators to two-stage samples. We also derive variances for our estimators. Using a common asymptotic design-based framework, we show that the generalized difference estimator is asymptotically unbiased in cluster samples. We also show that the GLM model-calibrated maximum pseudoempirical likelihood estimator is asymptotically equivalent to the GLM model-calibrated estimator in cluster samples.

4.2.1 GLM-Assisted Difference Estimator

In this section, we present results for the GLM-assisted difference estimator in clustered samples. Using a generalized linear assisting model, we extend the GLM-assisted difference estimator to two-stage samples and explore characteristics of the estimator. The variability of the GLM-assisted difference estimator depends on the fit of the assisting model. If the data are more aptly described by a GLM than a linear model, the GLM-assisted difference estimator will be more efficient than the GREG estimator.

Equation (1.19) on page 42 showed one way to express the GREG estimator obtained from a clustered sample. If we now replace the linear model in Equation (1.19)

with a GLM, we have the clustered GLM-assisted difference estimator

$$\widehat{\mathbf{t}}_y^{gd} = \sum_{\mathscr{U}} \widehat{\mu}_k + \sum_{\mathfrak{s}} d_k \left[y_k - \widehat{\mu}_k \right] \tag{4.19}$$

where $\widehat{\mu}_k$ is defined in Table 4.2 on page 200.

Since $\widehat{t}_y^{gd}$ is a function of $\widehat{\mu}_k$ and $\widehat{\mu}_k$ is a function of $\widehat{\mathbf{B}}$, one needs to compute $\widehat{\mathbf{B}}$ in the process of estimating $\widehat{\mathbf{t}}_y^{gd}$. Specifically, $\mathbf{B}$ can be estimated by numerically solving the GLM estimating equations reported in Equation (4.5). In two staged-samples, these estimating equations can be written as

$$\widehat{w}\left(\boldsymbol{\beta}\right) = \frac{1}{\phi} \sum_{i \in \mathfrak{s}_I} \sum_{k \in \mathfrak{s}_i} d_k \left\{ \left[y_k - \mu_k \right] \left[\frac{\phi_k}{\operatorname{var}\left(y_k\right)} \frac{\partial \mu_k}{\partial \gamma_k} \right] \omega_k \mathbf{x}_k \right\} = \mathbf{0}. \tag{4.20}$$

Alternatively, one can simultaneously compute $\widehat{\mathbf{t}}_y^{gd}$ and $\widehat{\mathbf{B}}$ using implicit differentiation. This is accomplished by adding one estimating equation to the estimating equations for $\mathbf{B}$. Let our parameter vector be

$$\underset{1+p}{\boldsymbol{\theta}} = \begin{bmatrix} t_y^{gd} \\ \\ \mathbf{B} \end{bmatrix}$$

In Appendix C.4.3.3 which starts on page 449, we show that $\boldsymbol{\theta}$ can be estimated by simultaneously solving the pseudomaximum likelihood estimating equations,

$$\underset{1+p}{\mathbf{W}\left(\boldsymbol{\theta}\right)} = \begin{bmatrix} \sum_{\mathfrak{s}} d_i d_{k|i} \left(\mathbf{y}_k - \boldsymbol{\mu}_k \right) - \left(\mathbf{t}_y^{gd} - \sum_{\mathscr{U}} \boldsymbol{\mu}_k \right) \\ \frac{1}{\phi} \sum_{\mathfrak{s}} d_k \left\{ \left[y_k - \mu_k \right] \left[\frac{\phi_k}{\operatorname{var}(y_k)} \frac{\partial \mu_k}{\partial \gamma_k} \right] \omega_k \mathbf{x}_k \right\} \end{bmatrix}$$

for $\mathbf{t}_y^{gd}$ and $\mathbf{B}$. This is done by setting $\mathbf{W}\left(\boldsymbol{\theta}\right) = \mathbf{0}$ and numerically solving for $\boldsymbol{\theta}$.

To determine the asymptotic properties of $\mathbf{t}_y^{gd}$ in cluster samples, we must make some general assumptions which describe our asymptotic framework. Using three assumptions, Wu and Sitter (2001) showed that the GLM-assisted difference estimator was

asymptotically design-unbiased in single-staged samples. Furthermore, under a fourth assumption, Wu and Sitter (2001) calculated the asymptotic variance of the GLM-assisted difference estimator in single-staged samples. We extend the four assumptions presented in Wu and Sitter (2001) to cluster samples and present the asymptotic bias and variance of the GLM-assisted difference estimator. Details in our proofs and derivations are in Appendix C.4 on page 446.

First, we assume that our estimated coefficients are consistent estimators of the finite population coefficients. Moreover, we also assume that as the number of clusters increase, the finite population coefficients approach the superpopulation parameters. Technically,

Assumption 8. $\|\widehat{\mathbf{B}} - \mathbf{B}\| = O_p\left(n^{-\frac{1}{2}}\right)$ *and* $\mathbf{B} \to \boldsymbol{\beta}$.

Second, we assume that our estimating function is smooth, differentiable, and that the estimator mean function is bounded. That is,

Assumption 9. *For each* $\mathbf{x}_k$, $\frac{\partial}{\partial \mathbf{t}} \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{t}\right)$ *is continuous in* $\mathbf{t}$ *and* $\left|\frac{\partial}{\partial \mathbf{t}} \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{t}\right)\right| \leq h\left(\mathbf{x}_k, \boldsymbol{\theta}\right)$ *for* $\mathbf{t}$ *in a neighborhood of* $\boldsymbol{\theta}$*, and* $N^{-1}\sum_{i=1}^{N} h\left(\mathbf{x}_k, \boldsymbol{\theta}\right) = O\left(1\right)$*, where* $h\left(\mathbf{x}_k, \boldsymbol{\theta}\right)$ *is a finite scalar.*

Third, we let our basic design weights be bounded in such a way that means generated using the basic design weights are asymptotically normally distributed.

Assumption 10. *The basic design weights,* $d_k = \frac{1}{\pi_k}$*, satisfy that the* π*-estimators for certain population means are asymptotically normally distributed.*

Lastly, to compute the asymptotic variance of the GLM-assisted difference estimator, we will need to assume that the second derivative of the GLM-assisted difference

estimating function is smooth, continuous, and bounded.

Assumption 11. *For each* $\mathbf{x}_k$, $\frac{\partial^2}{\partial \mathbf{t} \partial \mathbf{t}^\top} \mu\left(\mathbf{x_k}, \mathbf{t}\right)$ *is continuous in* $\mathbf{t}$ *and* $\max_{k,l} \left| \frac{\partial^2}{\partial \mathbf{t} \partial \mathbf{t}^\top} \mu\left(\mathbf{x_k}, \mathbf{t}\right) \right| \leq h\left(\mathbf{x}_k, \boldsymbol{\theta}\right)$ *for* $\mathbf{t}$ *in the neighborhood of* $\boldsymbol{\theta}$ *and* $N^{-1} \sum_{k=1}^{N} h\left(\mathbf{x}_k, \boldsymbol{\theta}\right) = O\left(1\right)$.

Theorem 4.1. *Under Assumptions 8, 9, and 10,* $\hat{t}_y^{gd}$ *is asymptotically design-unbiased for* t_y *in two-stage samples. Furthermore, under Assumption 11, the asymptotic variance of* $\widehat{t}_y^{gd}$ *is*

$$\operatorname{av}\left(\hat{\mathbf{t}}_y^{gd}\right) = \sum_{i \in \mathscr{U}_I} \sum_{j \in \mathscr{U}_I} \left(\Delta_{ij} d_i d_j t_{ei} t_{ej}^\top\right) + \sum_{i \in \mathscr{U}_I} \left[d_i \left(\sum_{k \in \mathscr{U}_i} \sum_{l \in \mathscr{U}_i} \Delta_{kl|i} d_{k|i} d_{l|i} e_k e_l^\top \right) \right] \tag{4.21}$$

where

$$t_{ei} = \sum_{k \in \mathscr{U}_i} e_k \tag{4.22}$$

$$e_k = y_k - \mu\left(\mathbf{x}_k, \mathbf{B}\right). \tag{4.23}$$

Furthermore, this asymptotic variance can be estimated by

$$v_{wr}\left(t_y^{gd}\right) = \frac{n}{(n-1)} \sum_{i \in \mathfrak{s}_I} \left(\widehat{t}_{\widehat{e}i}^{\pi} - \frac{1}{n} t_{\widehat{e}}^{\pi} \right) \left(\widehat{t}_{\widehat{e}i}^{\pi} - \frac{1}{n} t_{\widehat{e}}^{\pi} \right)^\top \tag{4.24}$$

where

$$\widehat{t}_{\widehat{e}i}^{\pi} = \sum_{k \in \mathfrak{s}_i} d_k \widehat{e}_k \tag{4.25}$$

$$\widehat{t}_{\widehat{e}}^{\pi} = \sum_{k \in \mathfrak{s}} d_k \widehat{e}_k = \sum_{i \in \mathfrak{s}_I} \widehat{t}_{\widehat{e}i}^{\pi} \tag{4.26}$$

$$\widehat{e}_k = y_k - \widehat{\mu}_k. \tag{4.27}$$

or by,

$$\upsilon_e\left(\hat{t}_y^{gd}\right) = \sum_{i\in\mathfrak{s}_I}\sum_{j\in\mathfrak{s}_I}\left(\frac{\Delta_{ij}}{\pi_{ij}}d_i d_j \widehat{t}_{ei}\widehat{t}_{ej}^{\top}\right) + \sum_{i\in\mathfrak{s}_I}\left[d_i\left(\sum_{k\in\mathfrak{s}_i}\sum_{l\in\mathfrak{s}_i}\frac{\Delta_{kl|i}}{\pi_{kl|i}}d_{k|i}d_{l|i}\widehat{e}_k\widehat{e}_l^{\top}\right)\right] \tag{4.28}$$

where

$$\widehat{t}_{ei} = \sum_{k\in\mathfrak{s}_i} d_{k|i}\widehat{e}_k \tag{4.29}$$

or by,

$$\upsilon_{Binder}\left(\hat{\mathbf{t}}_y^{gd}\right) = \left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right]\left[\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right)\right]\left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{\top} \tag{4.30}$$

where $\widehat{\mathbf{J}}\left(\widehat{\boldsymbol{\theta}}\right)$ *and* $\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right)$ *are defined in Appendix C.4.3.3 on page 449.*

In Appendix C.4.1 on page 446, we prove that $\widehat{t}_y^{gd}$ is design-consistent for t_y in two-stage samples. In Appendix C.4.2 on page 447, we prove that the asymptotic variance of $\hat{t}_y^{gd}$ is

$$\text{av}\left(\hat{t}_y^{gd}\right) = \sum_{i\in\mathscr{U}_I}\sum_{j\in\mathscr{U}_I}\left(\Delta_{ij}d_i d_j t_{ei}t_{ej}^{\top}\right) + \sum_{i\in\mathscr{U}_I}\left[d_i\left(\sum_{k\in\mathscr{U}_i}\sum_{l\in\mathscr{U}_i}\Delta_{kl|i}d_{k|i}d_{l|i}e_k e_l\right)\right] \tag{4.31}$$

where

$$t_{ei} = \sum_{k\in\mathscr{U}_i} e_k \tag{4.32}$$

and e_k is defined in Equation (4.23). Lastly, in Appendix C.4.3 on page 448, we construct υ_{wr}, υ_e, and υ_{Binder}. The with-replacement variance estimator, υ_{wr}, is based on the

assumption that the clusters were selected with-replacement. This estimator will usually approximate the variance in without-replacement samples when the fraction of sample clusters to total clusters is small. The classic survey weighted residual variance estimator, υ_e, requires knowledge of joint inclusion probabilities of selection.

When the point estimator can be written in terms of a g-weight, Särndal et al. (1989) use these weights in the variance estimator. Alas, $\widehat{t}_y^{gd}$ cannot be written as a linear combination involving a g-weight. Thus, we do not propose a g-weighted adjustment to υ_e.

The final variance estimator is the implicit differentiation variance estimator proposed by Binder (1983). The $\mathbf{J}$ matrix on the outsides of this estimator is the jacobian of the estimating equations, $\mathbf{W}(\boldsymbol{\theta})$, with respect to the parameters, $\boldsymbol{\theta}$. That is, $\widehat{\mathbf{J}}(\boldsymbol{\theta}) = \frac{\partial}{\partial(\boldsymbol{\theta})^\top}\widehat{\mathbf{W}}(\boldsymbol{\theta})$. The middle term in this estimator, $\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right)$, is an estimate of the variance of the sample weighted estimating equations. That is $\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right) = \upsilon\left(\sum_{\mathfrak{s}} \widehat{\mathbf{U}}_k(\boldsymbol{\theta})\right)$ where $\widehat{\mathbf{U}}_k(\boldsymbol{\theta})$ is the weighted portion of the estimating equations as shown in Appendix C.4.3. In the simulation, we use a with-replacement variance estimator to estimate this variance.

We constructed a GLM-assisted difference estimator for a scalar response variable selected from clustered samples. In Appendix C.4.1 on page 446 we prove that our estimator is design consistent for the true finite population total. In Appendix C.4.2 on page 447 we calculate the asymptotic variance of the estimator. Finally, in Appendix C.4.3 on page 448 we construct three variance estimators of the asymptotic variance. Results from these proofs are summarized in Theorem 4.1 on page 215.

4.2.2 Model-Calibrated Estimator

In this section, we extend the model-calibration estimator to two-stage samples and explore asymptotic characteristics of the estimator.

Equation (1.31) on page 49 presented the calibration estimator in two-stage samples. If we replace the constraints in the calibration estimator with the model-calibrated constraints, we obtain a model-calibrated estimator for clustered samples. Doing so gives

$$t_y^{mc} = \mathbf{y}^\top \mathbf{w}^{mc} \tag{4.33}$$

where $\underset{n\times 1}{\mathbf{w}^{mc}}$ is found by minimizing the chi-squared distance between the design weights and the model-calibration weights,

$$\frac{1}{2}\left(\mathbf{d} - \mathbf{w}^{mc}\right)^\top \mathbf{\Pi Q}^{-1}\left(\mathbf{d} - \mathbf{w}^{mc}\right) \tag{4.34}$$

which can be written in scalar form as

$$\frac{1}{2}\sum_{k\in\mathfrak{s}} \frac{\left(d_k - w_k^{mc}\right)^2}{d_k q_k} \tag{4.35}$$

subject to the constraint

$$\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \mathbf{w}^{mc} = \underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top \mathbf{1} \tag{4.36}$$

where

$$\underline{\boldsymbol{\mu}} = \begin{bmatrix} \mathbf{1} & \boldsymbol{\mu} \end{bmatrix} \tag{4.37}$$

Notice that if the first column of $\underset{n\times 2}{\underline{\boldsymbol{\mu}}_{\mathfrak{s}}}$ is $\mathbf{1}$, then the following constraint is also obtained.

$$\mathbf{1}^\top \mathbf{w}^{mc} = N. \tag{4.38}$$

Our restricted objective function is

$$\phi = \frac{1}{2}\left(\mathbf{d} - \mathbf{w}^{mc}\right)^{\top} \mathbf{\Pi}\mathbf{Q}^{-1}\left(\mathbf{d} - \mathbf{w}^{mc}\right) - \boldsymbol{\lambda}^{\top}\left(\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\mathbf{w}^{mc} - \underline{\boldsymbol{\mu}}_{\mathscr{U}}^{\top}\mathbf{1}\right)$$

where $\boldsymbol{\lambda}$ is a 2 by 1 vector of Lagrange multipliers. Minimizing this equation gives us the model-calibrated estimator for two-stage samples

$$\widehat{t}_y^{mc} = \mathbf{y}^{\top}\mathbf{w}^{mc} \tag{4.39}$$

$$= \hat{t}_y + \widehat{\mathbf{B}}_{\mathbf{y}\underline{\boldsymbol{\mu}}}\left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^{\top}\mathbf{1} - \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\mathbf{d}\right) \tag{4.40}$$

where

$$\underset{1\times 2}{\widehat{\mathbf{B}}_{\mathbf{y}\underline{\boldsymbol{\mu}}}} = \mathbf{y}^{\top}\mathbf{\Pi}^{-1}\mathbf{Q}\underline{\boldsymbol{\mu}}_{\mathfrak{s}}\left(\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\mathbf{\Pi}^{-1}\mathbf{Q}\underline{\boldsymbol{\mu}}_{\mathfrak{s}}\right)^{-1}.$$

We can also write our estimator as

$$\widehat{t}_y^{mc} = \mathbf{y}^{\top}\mathbf{\Pi}^{-1}\mathbf{g} \tag{4.41}$$

where

$$\underset{n\times 1}{\mathbf{g}} = \mathbf{1} + \mathbf{Q}\underline{\boldsymbol{\mu}}_{\mathfrak{s}}\left(\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\mathbf{\Pi}^{-1}\mathbf{Q}\underline{\boldsymbol{\mu}}_{\mathfrak{s}}\right)^{-1}\left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^{\top}\mathbf{1} - \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\mathbf{d}\right).$$

Details of this minimization are in Appendix C.5.1 on page 452. Alternative forms of $\widehat{t}_y^{mc}$ are in Appendix C.5.2 on page 453. Although all elements of $\mathbf{y}$ and $\mathbf{\Pi}^{-1}$ are nonnegative, $\mathbf{g}$ could be negative, especially when $\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\mathbf{d}$ is larger than $\underline{\boldsymbol{\mu}}_{\mathscr{U}}^{\top}\mathbf{1}$. When $\mathbf{g}$ is negative, negative estimates of $\widehat{t}_y^{mc}$ are possible.

Since $\widehat{t}_y^{mc}$ is a nonlinear function of sample inclusion indicators, the exact variance of $\hat{t}_y^{mc}$ cannot be determined. However, under our asymptotic framework, we can compute

the asymptotic variance of $\hat{t}_y^{mc}$. Furthermore, we can construct variance estimators of this asymptotic variance. The following theorem reports the asymptotic variance of $\hat{t}_y^{mc}$ and presents three estimators of this asymptotic variance.

Theorem 4.2. *The model-calibrated estimator, $\hat{t}_y^{mc}$, is design-consistent for the true population total. Furthermore, under Assumption 8 on page 214, the asymptotic variance of $\hat{t}_y^{mc}$ is*

$$\operatorname{av}\left(\hat{t}_y^{mc}\right) = \sum_{i\in\mathscr{U}_I}\sum_{j\in\mathscr{U}_I}\left(\Delta_{ij}d_id_jt_{ei}t_{ej}\right) + \sum_{i\in\mathscr{U}_I}\left[d_i\left(\sum_{k\in\mathscr{U}_i}\sum_{l\in\mathscr{U}_i}\Delta_{kl|i}d_{k|i}d_{l|i}e_ke_l\right)\right] \tag{4.42}$$

where

$$t_{ei} = \sum_{k\in\mathscr{U}_i} e_k \tag{4.43}$$

and e_k is defined in Equation (4.23)*. The asymptotic variance of $\widehat{t}_y^{mc}$ can be estimated by*

$$\upsilon_g\left(\hat{t}_y^{mc}\right) = \sum_{i\in\mathfrak{s}_I}\sum_{j\in\mathfrak{s}_I}\left(d_{ij}\Delta_{ij}d_id_j\hat{t}_{g\hat{e}i}\hat{t}_{g\hat{e}j}\right) + \sum_{i\in\mathfrak{s}_I}\left[d_i\left(\sum_{k\in\mathfrak{s}_i}\sum_{l\in\mathfrak{s}_i}d_{kl|i}\Delta_{kl|i}d_{k|i}d_{l|i}g_kg_l\hat{e}_k\hat{e}_l\right)\right] \tag{4.44}$$

where $\hat{t}_{g\hat{e}i} = \sum_{\mathfrak{s}_i}\frac{g_k\hat{e}_k}{\pi_{k|i}}$ or by

$$\upsilon_{wr}\left(\widehat{t}_y^{mc}\right) = \frac{n}{(n-1)}\sum_{i\in\mathfrak{s}_I}\left[d_i\sum_{k\in\mathfrak{s}_i}\left(d_{k|i}\hat{e}_{k|i}\right) - \frac{1}{n}\sum_{k\in\mathfrak{s}}\left(d_k\hat{e}_k\right)\right]\left[d_i\sum_{k\in\mathfrak{s}_i}\left(d_{k|i}\hat{e}_{k|i}\right) - \frac{1}{n}\sum_{k\in\mathfrak{s}}\left(d_k\hat{e}_k\right)\right]^\top \tag{4.45}$$

where $\hat{t}_{e,i} = \sum_{\mathfrak{s}_I}\frac{\hat{e}_k}{\pi_{k|i}}$, or by

$$\upsilon_{Binder}\left(\widehat{\boldsymbol{\theta}}\right) = \left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right]\left[\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right)\right]\left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right]^\top \tag{4.46}$$

where $\widehat{\mathbf{J}}\left(\widehat{\boldsymbol{\theta}}\right)$ and $\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right)$ are defined in Appendix C.5.5.3.

In Appendix C.5.3, we prove that the model-calibrated estimator is design-consistent in two-stage samples. See Appendix C.5.4 for a derivation of the asymptotic variance of the model-calibrated estimator. In Appendix C.5.5 on page 456 we derive the three variance estimators noted in Theorem 4.2.

The first variance estimator, $\upsilon_g\left(\hat{t}_y^{mc}\right)$ is the standard weighted residual variance estimator with a g-weight adjustment. In Appendix C.5.5.1 on page 456, we also develop the weighted residual variance estimator without the g-weighted adjustment, but do not report results here because Särndal et al. (1989) showed that in general the g-weighted variance estimator had better properties than the estimator without the g-weights. The second estimator is the classic with-replacement variance estimator. When the fraction of sample clusters to total clusters is small, the with-replacement variance estimator usually comes close to the variance in without-replacement samples. The clear advantage of the with-replacement variance estimator is its simplicity. The final variance estimator is the implicit differentiation variance estimator proposed by Binder (1983). The middle term in this estimator $\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right)$ is an estimate of the variance of the sample weighted estimating equations. In the simulation, we use a with-replacement variance estimator to estimate this variance.

We constructed a model-calibration point estimator for scalar responses selected from clustered samples and proved that it was design-consistent. We also calculated the asymptotic variance of the model-calibration estimator and constructed three variance estimators of the asymptotic variance.

4.2.3 Model-Calibrated Maximum Pseudoempirical Likelihood Estimator

We extend the generalized linear model-calibrated maximum pseudoempirical likelihood estimator to two-stage samples and explore asymptotic characteristics of the estimator.

Equation (1.36) on page 53 shows the model-calibrated maximum pseudoempirical estimator of a mean in one-stage samples. Total estimates can be constructed with,

$$\hat{t}_y^{peM} = M \sum_{i \in \mathfrak{s}} p_k^{pe} y_i, \tag{4.47}$$

or

$$\hat{t}_y^{pe\widehat{M}} = \widehat{M} \sum_{i \in \mathfrak{s}} p_i^{pe} y_i, \tag{4.48}$$

where

$$\widehat{M} = \sum_{k \in \mathfrak{s}} d_k$$

When complete auxiliary data are available, M is known and $\hat{t}_y^{peM}$ will be the preferable estimator. Since there may be considerable sampling error in estimating $\widehat{M}$, $\hat{t}_y^{pe\widehat{M}}$ may be significantly more variable than $\hat{t}_y^{peM}$. One exception is for sample designs where $\widehat{M} = M$, in which case the two estimators are equivalent. For example, $\widehat{M}$ reduces to M in probability proportional to size samples where clusters are selected with probabilities $n\frac{M_i}{M}$ and units within clusters are selected with probabilities $\frac{m}{M_i}$.

Modifying the single-stage model-calibrated maximum pseudoempirical estimator for clustered samples yields the new model-calibrated maximum pseudoempirical esti-

mator

$$\hat{t}_y^{peM} = M \sum_{i \in \mathfrak{s}_I} \sum_{k \in \mathfrak{s}_i} p_{ik}^{pe} y_{ik}, \tag{4.49}$$

or

$$\hat{t}_y^{pe\widehat{M}} = \widehat{M} \sum_{i \in \mathfrak{s}_I} \sum_{k \in \mathfrak{s}_i} p_{ik}^{pe} y_{ik}, \tag{4.50}$$

where

$$\widehat{M} = \sum_{i \in \mathfrak{s}_I} \sum_{k \in \mathfrak{s}_i} d_k$$

and p_{ik}^{pe} is found by maximizing

$$\sum_{\mathfrak{s}_I} \sum_{\mathfrak{s}_i} d_{ik} \log\left(p_{ik}^{pe}\right) \tag{4.51}$$

subject to

$$\sum_{\mathfrak{s}_I} \sum_{\mathfrak{s}_i} p_{ik}^{pe} = 1 \tag{4.52}$$

$$\sum_{\mathfrak{s}_I} \sum_{\mathfrak{s}_i} p_{ik}^{pe} u_{ik} = 0 \tag{4.53}$$

where

$$u_{ik} = \mu_{ik} - \frac{1}{M} \sum_{k \in \mathscr{U}} \mu_{ik}. \tag{4.54}$$

In Appendix C.6.1 on page 460 which references B.6.1, we maximize the pseudoempirical likelihood subject to our model calibration constraints to create the model-calibrated maximum pseudoempirical likelihood estimator. In Appendix C.6.2 we prove that $\widehat{t}_y^{peM}$ is asymptotically equivalent to $\widehat{t}_y^{mc}$.

Theorem 4.3. *Under Assumptions 4 through 7 on page 134, $\widehat{t}_y^{pe}$ is asymptotically design-unbiased for t_y in two-stage samples. Furthermore, the asymptotic variance of $\widehat{t}_y^{pe}$ is equivalent to the asymptotic variance of $\widehat{t}_y^{mc}$ and can be estimated with the variance estimators for $\widehat{t}_y^{mc}$.*

See Appendix C.6.2 on page 461 for the proof of Theorem 4.3. Since the variance of $\widehat{t}_y^{pe}$ is asymptotically equivalent to the variance of $\widehat{t}_y^{mc}$, we do not construct new variance estimators for $\widehat{t}_y^{pe}$. Instead, we recommend using one of the four variance estimators we already constructed for $\widehat{t}_y^{mc}$.

In our simulation, we compare both $\hat{t}_y^{peM}$ and $\hat{t}_y^{pe\hat{M}}$; although we do not see any advantages of $\hat{t}_y^{pe\hat{M}}$ over $\hat{t}_y^{peM}$.

In this section, we derived the two-stage version of the generalized linear model-calibrated maximum pseudoempirical likelihood estimator and proved that it is asymptotically equivalent to the GLM model-calibrated estimator. Our estimator could not be written in closed form, so numerical methods will be necessary to estimate p_{ik}^{pe}.

4.3 Simulation

We performed a simulation study to compare the design-based properties of the three new types of GLM-assisted estimators in two-stage samples. We selected both small and large samples from a population derived form Census data.

From our sampling frame, we repeatedly selected six types of two-stage samples.

Fixed SRS. In the first and second sets of samples, we selected a fixed set of clusters.

Then, we selected a fixed number of units within each sample cluster. For select-

ing the clusters and units, we used a simple random sample without-replacement algorithm. We call this method Fixed SRS because in both stages of sampling, we selected a fixed number of units. Because our cluster sizes varied, this design resulted in unequal weights. The second sample design was the same as the first, with the exception that the number of sample clusters selected was larger.

Rate SRS. In the third and fourth sets of samples, we selected a fixed set of clusters, but selected units in sample clusters at a constant rate. This design resulted in random sample sizes, but all sample units had the same base weight. We call this sample design Rate SRS because units within sample clusters were selected at a constant rate. The third and fourth sample designs differed in the number of clusters selected.

Fixed PPS. Finally, in the fifth and sixth sets of samples, a sample of clusters was selected with probabilities proportional to the number of units in the cluster. Then a fixed number of units in each sample cluster was selected using a simple random sample without-replacement algorithm. This method resulted in a fixed sample size and equal weights. The fifth and sixth samples differed in the number of clusters selected.

The goal of these simulations was to assess how the design-based empirical bias and variance of the new estimators compared to the bias and variance of the π-estimator and the GREG estimator using a similar model. For each sample, we estimated the total of our multivariate response vector using the estimators in Table 4.3. We repeated this process for ten thousand samples. For the point estimators in Table 4.4, we calculated the relative empirical bias and empirical coefficient of variation. For the variance estimators

in Table 4.4, we calculated the relative bias and confidence interval coverage of $\widehat{t}_y^{gd}$ and $\widehat{t}_y^{mc}$.

Table 4.4: Variance Estimators Calculated in Simulations

Statistic	Description
$\upsilon_{wr}\left(\hat{t}_y^{gd}\right)$	With Replacement Variance Estimator of $\hat{t}_y^{gd}$
$\upsilon_{e}\left(\hat{t}_y^{gd}\right)$	Without Replacement Variance Estimator of $\hat{t}_y^{gd}$
$\upsilon_{Binder}\left(\hat{t}_y^{gd}\right)$	Binder's Variance Estimator of $\hat{t}_y^{gd}$
$\upsilon_{wr}\left(\hat{t}_y^{mc}\right)$	With Replacement Variance Estimator of $\hat{t}_y^{mc}$
$\upsilon_{e}\left(\hat{t}_y^{mc}\right)$	Without Replacement Variance Estimator of $\hat{t}_y^{mc}$
$\upsilon_{g}\left(\hat{t}_y^{mc}\right)$	g-weighted Without Replacement Variance Estimator of $\hat{t}_y^{mc}$
$\upsilon_{Binder}\left(\hat{t}_y^{mc}\right)$	Binder's Variance Estimator of $\hat{t}_y^{mc}$

We included $\upsilon_{e}\left(\hat{t}_y^{mc}\right)$ in our simulations even though Särndal et al. (1989) clearly advocated using the g-weighted variance estimator. We include it for comparison purposes, even though we expect $\upsilon_{g}\left(\hat{t}_y^{mc}\right)$ to perform better than $\upsilon_{e}\left(\hat{t}_y^{mc}\right)$ in most cases.

In the large samples, we used an *ad hoc* finite population correction factor of $1-\frac{n}{N}$ for υ_{wr} and υ_{Binder}. For example, we multiplied $\upsilon_{wr}\left(\hat{t}_y^{gd}\right)$ by $1-\frac{35}{136}$ in the large samples where we selected 35 clusters.

4.3.1 Population: 2000 Tract Level Planning Database

This section describes the 2000 Tract Level Planning Dataset (TLPD) from the US Census Bureau, the pseudo-population used to evaluate the GLM-assisted estimators in clustered samples.

This pseudo-population came from the second version of the US Census Bureau's Tract Level Planning Database with Census 2000 Data (Bruce and Robinson 2006). This dataset contained the mail return rates from the 2000 Census for every tract in the 50 states and the District of Columbia. According to Bruce and Robinson (2006), "census tracts

are delineated for all metropolitan areas and counties. Tracts usually have between $2,500$ and $8,000$ people, though some have very small populations. When first delineated, tracts are designed to be homogeneous with respect to population characteristics, economic statistics, and living conditions. The spatial size of tracts varies widely depending on the diversity of settlement." Along with this data, the database also contained tract level summary data from the 2000 Census and the American Community Survey. This dataset played a central role in estimating the budget for the 2010 Census and in developing the marketing campaign for the 2010 Census.

We edited the Tract Level Planning Database to make it suitable for our simulation. The first edit was to remove all nonrepresentative tracts which were flagged on the database.

The second edit was used to remove outliers which threatened the fit of the GLMs in small samples. We used a linear model with an intercept and the hard to count score to fit the mail return rate. If the leverage points in this model were greater than 0.00005, we removed the tract.

The third edit was to remove all counties with either less than 80 tracts or more than 500 tracts. This requirement was needed to assure that each county had enough units within it for the second-stage sample. It also assured that no counties were selected with certainty when selecting a PPS sample.

Clusters were defined as counties and tracts were used as units. The final dataset contained $21,642$ tracts in 136 counties. Table 4.5 summarizes the edits to the database.

We estimated three totals:

Table 4.5: Edits for Tract Population

Description	SAS code
Remove nonrepresentative tracts	FLAG $\neq$ " "
Remove tracts missing mail return rate	Mail Return Rate $\neq$.
Remove tracts missing housing units in single structures	Pct Single U Strc $\neq$.
Remove tracts missing poverty rate	Pct Prs Blw Pov Lev $\neq$.
Remove tracts where the poverty rate is over 40 percent	Pct Prs Blw Pov Lev $\leq$ 40
Remove tracts missing percent white population	Pct White $\neq$.
Remove tracts where model leverage is over 0.00005	lev $<$ 0.00005
Remove counties with less than 80 tracts	$M_i \geq 80$
Remove counties with more than 500 tracts	$M_i \leq 500$

Count: Non-mail returns in the US. The tract level non-mail return rate was defined as the total number of occupied housing units that did not respond to the 2000 Census by mail divided by the total number of occupied housing units. We used the total number of housing units, the housing unit vacancy rate, and the mail return rates to calculate this rate.

Binary: Tracts with a mail return rate less than or equal to 75 percent.

Synthetic: Simulated binary variable. We used the method described by Oman and Zucker (2001) to generate a clustered binary response variable with a total similar to the number of tracts with a mail participation rate greater than 25 percent. We generated the variable so that the correlation of units within clusters would be about 0.09 and fit our GLM. The code we used to generate our random variable is in Appendix C.8.1 on page 487.

All of our assisting models used the same set of covariates:

- an intercept and
- the standardized hard to count score.

The hard to count score was standardized so that it would have a mean of 0 and a standard deviation of 1. Table 4.6 shows summary statistics for the key variables in the population.

Table 4.6: Quartiles for Tract Level Planning Dataset Population

Variable	Minimum	First Quartile	Median	Mean	Third Quartile	Maximum	Total
Total non-mail returns	0	210	323	372	474	2615	8,049,846
Occupied housing units	1	1,141	1,579	1,693	2,114	11,170	36,647,789
Tracts with a nonparticipation rate over 25%	0	0	0	0.3372	1	1	7,298
Synthetic response	0	0	0	0.3276	1	1	7,090
Standardized Hard to Count Score	-1.1865	-0.8964	-0.2195	0	0.8442	2.1014	0
Tracts per county	81	101.5	133.5	159.1	178.5	493	21,642

4.3.2 Models

We employed five different link functions to predict the count of non-mail returns in each tract:

- the identity link,
- the complementary log-log link,
- the probit link,
- the cauchit link, and
- the log link.

For the identity link, our model did not include the total number of occupied housing units in the tract. For the other four link functions, our model was,

$$\frac{\mathscr{E} y_k}{z_k} = \mu_k = g\left(\mathbf{x}_k^\top \boldsymbol{\beta}\right) \tag{4.55}$$

where y_k is the total number of non-mail returns and z_k is the total number of occupied housing units in the tract. Table 4.2 on page 200 provides the form of g for all five link functions we used to model this variable.

Using the complete population, we plotted predictions for all five models against the standardized hard to count score. Rather than showing all of the predictions, we

plotted the lowess smoother through the points. Figure 4.1 shows the plot. The black line in Figure 4.1 shows the perfect model where $y_k = \widehat{\mu}_k$. Since the yellow, green, blue, and purple lines representing the four nonlinear models are nearly on top of each other, we conclude that all four models perform similarly. Given that they are fairly close to the black line, we also see that individual predictions from the four nonlinear models are fairly close to the true value. In general, the models slightly overestimate the response when x_k is small and large, while underestimating y_k when x_k is around zero. The red linear regression line completely overestimates the individual predictions.

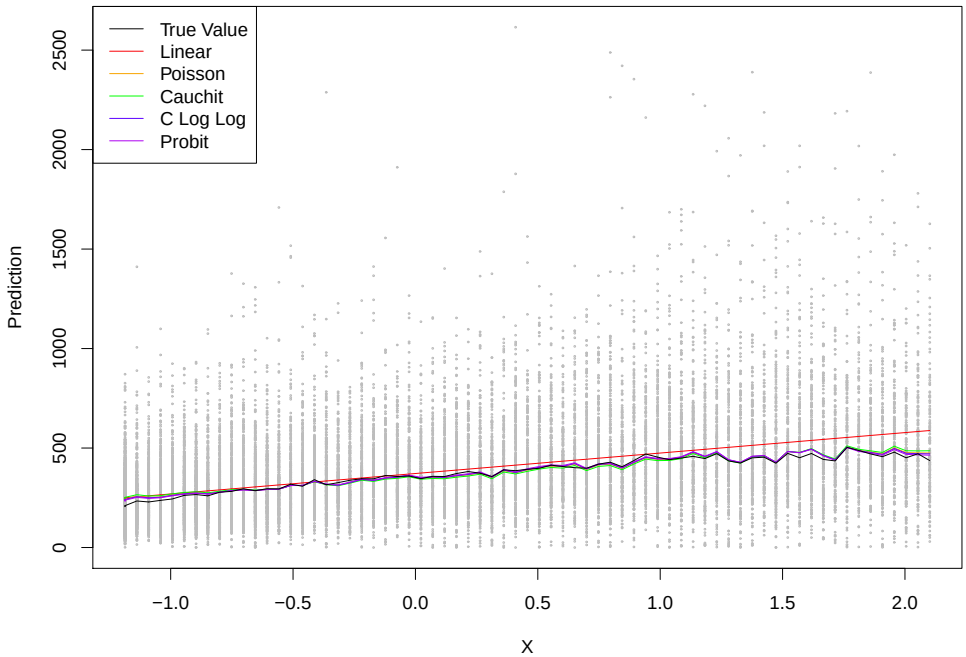

Figure 4.1: Plot of predictions versus true values in for total non-mail returns.

Even though the nonlinear models seem to do a better job at predicting individual responses, this does not necessarily mean that estimates of totals based on the nonlinear models will be better in terms of bias and variance than estimates based on the linear

model. Given that some of the values in the population are quite large, some overestimation might be advantageous in repeated samples. Table 4.7 shows total estimates based on fitting our models to the finite population. As we see, estimates of totals using the linear model are quite accurate, despite less than desirable predictions at the unit level.

We used four assisting GLMs to estimate the total number of tracts that had a non-mail return rate greater than 25%: a probit model, a cauchit model, a complementary log-log model, and a linear model. Our response was a binary outcome, taking on the value of 0 if the non-mail return rate was 25% or less and a value of 1 if the non-mail return rate was greater than 25%. Since our outcome variable could only take on one of two values, the log link function was not appropriate because $\log(0)$ is undefined.

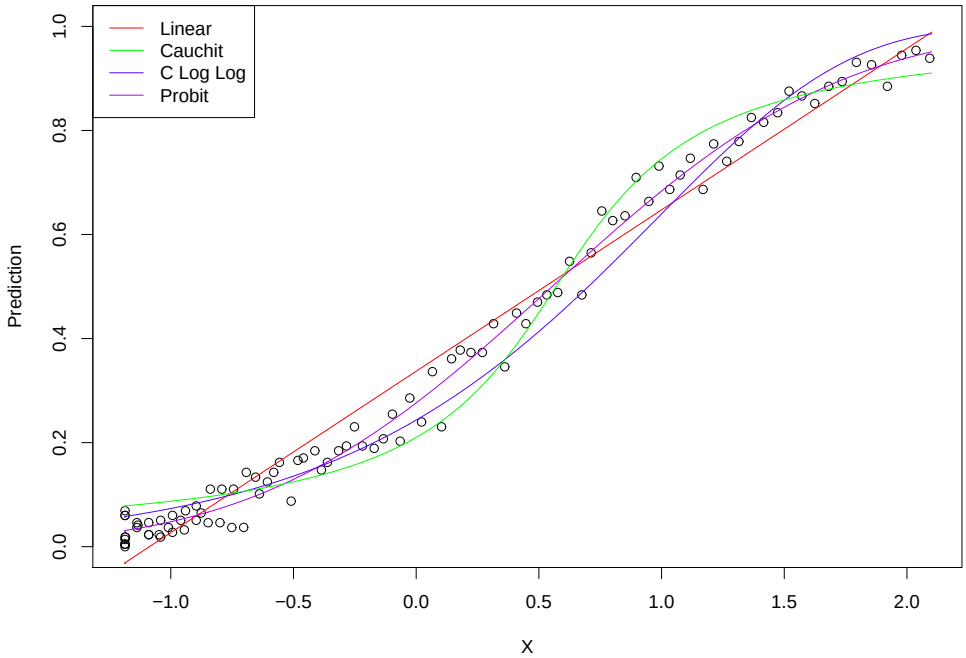

Figure 4.2: Plot of predictions versus x_k for the binary response. Each point represents the true average rate for 260 units. Models were fitted using the entire population.

To assess the fit of the four models, we divided the population into 100 equal sized groups. To form the groups we first sorted the population based on the standardized hard-to-count score, $\mathbf{x}_k$. Then, going down our ordered list, each group was determined by sequentially taking the next 260 units. Within each group, we calculated the percent of units with a success. Each dot in Figure 4.3 represents the true mean for a group of 260 units. Then we drew the fitted GLM lines over those points. Figure 4.3 shows the resulting graphs for all four models.

From Figure 4.3, we see that the linear model is the worst fit of the four GLMs. The probit and cauchit models seem to fit the data quite well, while the complementary log-log model seems to overestimate the true values at the extremes and underestimate the true value in the middle.

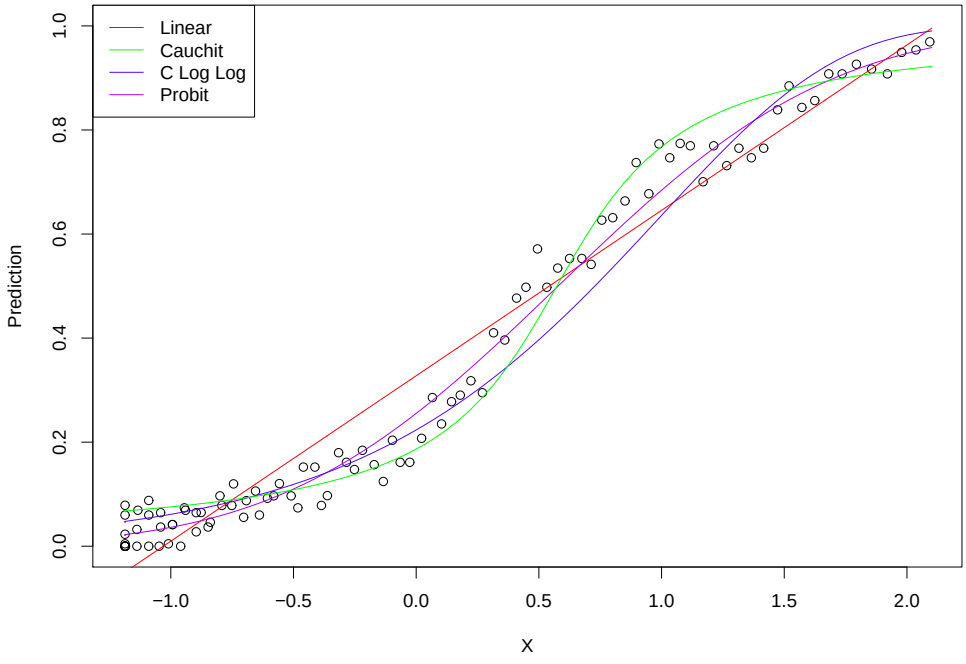

Figure 4.3: Plot of predictions versus x_k for the synthetic response. Each point represents the true average rate for 260 units. Models were fitted using the entire population.

To assess the fit of the synthetic response variable, we repeated the analysis for the binary response with the exception that the synthetic response variable was used in the models. As we see in Figure 4.3, the linear model does not seem to adequately predict the response variable. Indeed, the population seems to have more of an s-shape than a linear-shape. This is expected since a clustered logistic model was used to generate the response variable. The probit model appears to fit the data the best. When estimating totals, it is especially important to estimate the larger values in the population accurately. The cauchit model seems to underestimate these large values while the complementary log-log model appears to overestimate the larger values.

We calculated $\mathbf{B}$ by fitting our models to the complete finite population. In many estimation techniques, summing all the fitted values from the finite population should equal the sum of the true values. That is, we expect $\sum_{k\in\mathscr{U}} \mu\left(\mathbf{x}_k, \mathbf{B}\right) = \sum_{k\in\mathscr{U}} y_k$. Table 4.7 shows values of $\sum_{k\in\mathscr{U}} \mu\left(\mathbf{x}_k, \mathbf{B}\right)$ using each of the link functions. For the binary and synthetic totals, this property is violated for the complementary log-log and cauchit models. For this reason, we would expect some bias for all models using these link functions. Additionally, we would also expect some small bias in the cauchit models for the count data.

Table 4.7: Comparison of Finite Population Predictions when $\mathbf{B}$ is Known

Link Function	Count Total	Binary Total	Synthetic Total
True Value	8,049,846	7,298	7,090
Identity	8,049,846	7,298	7,090
Log	8,047,178		
Probit	8,049,221	7,294	7,088
Complementary Log Log	8,048,782	7,231	7,012
Cauchit	8,057,717	7,465	7,344

Firth and Bennett (1998) discuss this property in greater detail and suggest estimation techniques to assure that $\sum_{k\in\mathscr{U}} \mu\left(\mathbf{x}_k, \mathbf{B}\right) = \sum_{k\in\mathscr{U}} y_k$. Since the complementary-log-log and cauchit link functions are not canonical, we do not expect $\sum_{k\in\mathscr{U}} \mu\left(\mathbf{x}_k, \mathbf{B}\right)$ to be equal to $\sum_{k\in\mathscr{U}} y_k$ unless we include such a constraint in our estimator. When a canonical link is used, the population estimating equations for $\mathbf{B}$ shown in Equation (4.4) reduce to

$$w\left(\boldsymbol{\beta}\right) = \sum_{k=1}^{N} \left\{\left[y_k - \hat{\mu}_k\right] \mathbf{x}_k\right\} \tag{4.56}$$

If our model has an intercept and the canonical link is used, then the first element of $\mathbf{x}_k$ is 1 and our property is held. However, in the case of probit, cauchit, log, and complementary log-log models, our estimating equations will not reduce to Equation (4.56) and we can not guarantee that $\sum_{k\in\mathscr{U}} \mu\left(\mathbf{x}_k, \mathbf{B}\right)$ will equal $\sum_{k\in\mathscr{U}} y_k$.

4.3.3 Simulation Design

4.3.3.1 Sample Design

From our sampling frame, we selected simple random sample without-replacement (SRSWOR) and πps samples of clusters. Within each cluster, we selected a sample of units. We then estimated the total of the response variables using the estimators in Table 4.3.

We used the `UPrandomsystematic()` and `UPpoisson()` functions in the `sampling` package of R to select all the samples (Tillé and Matei 2009). The R function called `UPrandomsystematic()` selects a randomized systematic sample by sorting

the population into a random order and then selecting a sample with probabilities proportional to a size measure. This function selects without-replacement samples to achieve a fixed sample size. We used the `UPrandomsystematic()` function to select both stages of the Fixed SRS and Fixed PPS samples. We also used it to select the first stage of the Rate SRS samples. The `UPPoisson()` function selects a Poisson sample and was used to select the second stage of the Rate SRS samples.

Table 4.8: Simulation Design

	Simulation	First Stage Sample	n	Second Stage Sample	Number of Samples
1	SRSWOR Fixed	srswor	5	$m_i = 60$	$10,000$
2	SRSWOR Fixed	srswor	35	$m_i = 60$	$10,000$
3	SRSWOR Rate	srswor	5	$f_i = \frac{9,360}{26,023}$	$10,000$
4	SRSWOR Rate	srswor	35	$f_i = \frac{9,360}{26,023}$	$10,000$
5	PPSWOR Fixed	ppswor	5	$m_i = 60$	$10,000$
6	PPSWOR Fixed	ppswor	35	$m_i = 60$	$10,000$

We tested how the estimators performed under the three realistic sample designs described at the beginning of Section 4.3.

We selected samples of 5 and 35 clusters. From each cluster, a sample of 60 units was selected. Table 4.8 summarizes the different designs used to select the samples.

4.3.3.2 Number of Samples

In each of the six simulations, we selected $10,000$ samples. In Chapter 3, we determined the number of samples needed to attain target coefficients of variation. In this chapter, we set the number of samples at $10,000$ and report the simulation coefficient of variation

$$CV_{sim} = \frac{\frac{1}{\sqrt{\aleph}}\sqrt{\frac{1}{\aleph-1}\sum_{\nu=1}^{\aleph}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^2}}{\boldsymbol{\theta}}$$

Our point and variance estimators varied from sample to sample. Consider estimator $\widehat{\theta}_\nu$ from sample ν. The average of our $\widehat{\theta}_\nu$ estimators across all $\aleph$ samples is $\overline{\widehat{\theta}} = \frac{1}{\aleph}\sum_{\nu=1}^{\aleph}\widehat{\theta}_\nu$ and an estimate of the standard error of this mean is $se\left(\overline{\widehat{\theta}}\right) = \frac{\sqrt{\frac{1}{\aleph-1}\sum_{\nu=1}^{\aleph}\left(\widehat{\theta}_\nu-\overline{\widehat{\theta}}\right)^2}}{\sqrt{\aleph}}$. This standard error, $se\left(\overline{\widehat{\theta}}\right)$, is called the simulation error. Notice that the simulation error is different from the empirical standard deviation of the estimated total, $se\left(\widehat{\theta}\right) = \sqrt{\frac{1}{\aleph}\sum_{\nu=1}^{\aleph}\left(\widehat{\theta}_\nu - \overline{\widehat{\theta}}\right)^2}$, which does not depend on $\aleph$ in the denominator. Clearly, the more samples we select, the more confidence we will have in the mean of the totals and the standard error of the totals.

Appendix C.7.1 on page 463 shows tables reporting the simulation coefficient of variation for all point estimators in the six simulations. In all cases the simulation error was less than 0.0036% of the true population total.

4.3.3.3 Estimation

We estimated the total of each response variable using the estimators in Table 4.3. We repeated this process for all samples.

Predictions from a linear model play a key part of the GREG estimator. We used the `lm()` function in R with a weights option to predict the fitted values which were used in the GREG estimation. Our linear model contained the same covariate as the other GLM models. We did not use an offset for the GREG estimation.

The remaining estimators required predicting μ_k. To calculate μ_k, we first estimated $\mathbf{B}$, the parameters obtained from running a GLM on the full population. We used iterated weighted least squares to estimate $\mathbf{B}$ using the `glm()` function in R. To determine con-

vergence, we used a tolerance of $\sqrt{.\text{Machine\$double.eps}}$. To assist estimation, we used the true population value of $\mathbf{B}$ as a starting point for computing $\widehat{\mathbf{B}}$ from a sample. With $\widehat{\mathbf{B}}$ based on our sample and $\mathbf{x}_k$ for the complete population, we calculated $\widehat{\mu}_k$ for all elements on the frame.

We used an explicit form of the model-calibration estimator to make estimates. For inverting matrices in the model-calibration estimator, we used the `solve()` function in R. For the model-calibrated maximum pseudoempirical likelihood, we used the `Lag2` function provided by Changbao Wu[2].

For the implicit differentiation variance estimators, we formed estimating equations and used the `jacobian()` function in R to numerically calculate the Jacobian of the survey weighted estimating equations (Gilbert 2012).

Appendix C.8.2 on page 488 contains the code used to select the samples and estimate all parameters.

Table 4.9: Simulation Design

Estimator	Total Non-Mail Returns		Non-Mail Return Rate over 25%		Simulated Response		Total
	Point	Variance	Point	Variance	Point	Variance	
$\hat{t}_y^{\pi}$	1	0	1	0	1	0	3
$\hat{t}_y^{gr}$	1	0	1	0	1	0	3
$\hat{t}_y^{pr}$	5	0	4	0	4	0	13
$\hat{t}_y^{gd}$	5	15	4	12	4	12	52
$\hat{t}_y^{mc}$	5	20	4	16	4	16	65
$\hat{t}_y^{peM}$	5	0	4	0	4	0	13
$\hat{t}_y^{pe\hat{M}}$	5	0	4	0	4	0	13
Total	27	35	22	28	22	28	162

For each sample, we made 162 estimates. Table 4.9 summarizes the estimates made for each sample. For the Count response variable, we estimated the projective, generalized difference, model calibrated, and two model-calibrated maximum pseudoempirical

[2]See http://www.math.uwaterloo.ca/ cbwu/Rcodes/LagrangeM2.txt.

likelihood estimators with five GLM assisting models. For each of the five GLM assisting models, we calculated three variance estimators for the difference estimators and four variance estimators for the model-calibrated estimator. Since there were three variance estimators for each of the five $\hat{t}_y^{gd}$ estimators, there are 15 total variance estimates for $\hat{t}_{count}^{gd}$ in each sample. We only used four assisting models for the other two response variables, so there are only 12 variance estimators.

4.3.3.4 Measures

To evaluate the point estimators, we calculated the relative bias and coefficient of variation of the estimators. For the variance estimators, we calculated the relative bias and confidence interval coverage. Section 1.1.6 on page 21 describes the empirical relative bias and coefficient of variation in more detail. Appendix C.7 on page 463 contains plots or tables showing these measures for all simulations.

To evaluate the performance of the variance estimators, we constructed confidence intervals using the variance estimators. We calculated the percent of samples in which the confidence intervals contained the true population value. Confidence intervals were created using the *t*-distribution with $n - 1$ degrees of freedom where n was the number of sample clusters. Section 1.1.6 on page 21 describes methods we used to calculate the relative bias and confidence interval coverage. We used a nominal coverage level of 95%. Thus, we expect that about 95% of the confidence intervals should cover the true value. We note that using $n - 1$ degrees of freedom is a commonly used approximation, but not exact. In large samples, errors associated with using this approximation are negligible.

4.3.4 Results

4.3.4.1 Simulation Errors

In general, we designed our populations and samples to limit the risk of encountering a problem estimating certain quantities. Nevertheless, when modeling the response data, several problems arose, especially in small samples.

There are several practical problems that may hinder using one of the GLM-assisted estimators. Point estimation is not possible if

1. All of the responses in the sample are the same. This is common with rare characteristics where the characteristic is not observed in sample.
2. $\mathbf{X}$ is not full rank. Of course this can easily be fixed in practice by removing the dependent variable or using a generalized inverse when inverting functions of $\mathbf{X}$.
3. μ_k cannot be predicted for a non-sample unit. For example, if none of the sample units has one level of a covariate used to model μ_k or if one of the models fails to converge.

Furthermore, implicit differentiation variance estimators are not possible if

4. $(\boldsymbol{\mu}\mathbf{w})^\top \boldsymbol{\mu}$ is not full rank,
5. the jacobian of difference estimator estimating equations is not full rank, or
6. the jacobian of MCAL estimating equations is not full rank.

Table 4.10: Number of Errors Found in Each Simulation

Simulation	Error 1	Error 2	Error 3	Error 4	Error 5	Error 6
1	0	0	0	164	491	0
2	0	0	0	0	0	0
3	0	0	0	203	497	0
4	0	0	0	0	0	0
5	0	0	0	158	488	0
6	0	0	0	0	0	0

Table 4.10 shows the number of samples that were thrown out and replaced by a new sample. The simulation numbers correspond to the simulations in Table 4.8. As soon

as an error was encountered, the sample was thrown out and a new sample was selected to replace the skipped sample. For this reason, the counts in Table 4.10 are not mutually exclusive. For example, if $\mathbf{y}_{math} = 0$ for all sample units and $\mathbf{X}$ was not full rank, only Error 1 would be recorded in Table 4.10.

As we see, problems were only encountered in the small samples. In the small samples, about 7% of the samples were rejected. Since the within-cluster samples were reasonably large and none of the variables were rare, we would not expect the first error to occur. The fact that our covariate was continuous helped mitigate the frequency of the second and third errors. All of the critical errors were a result of trying to invert matrices for t^{mc} and the Binder variance estimators. Certainly, there are plenty of alternative point and variance estimators that can be used if one encounters a problem with t^{mc} and the Binder variance estimators in practice.

In addition to removing some samples because of the errors previously mentioned, we also looked for extreme estimates that could threaten our estimates of bias and variance. All samples in all six simulations conformed to our expectations and did not warrant removal. There were samples which generated estimates that were far from the true values, but these estimates were not excessively large enough to meaningfully alter our summary measures.

4.3.4.2 Point Estimators

In this section we report and discuss results about the point estimators from the six simulations. We focus on the relative bias and coefficient of variation of all point

estimators in Table 4.3 on page 205. In general, the bias of all estimators was relatively low. Furthermore, any detectable bias was related to both the general form of the estimator and the assisting model. As we will show in plots of the coefficient of variation, the GLM-assisted estimators tended to be more efficient than the π-estimator. Sometimes they were also more efficient that the GREG estimator, but the performance of the GLM-assisted estimators depends on the model fit.

Table 4.7 on page 233 shows estimates of totals when models are fit to the full population. Ideally, all of the estimates should be equal to the population total. As we see, this was not the case for all estimators. Given that the log, complementary-log-log, and cauchit GLMs in Table 4.7 were sometimes biased, we would expect to see some bias in the estimators using these link functions. At the end of Section 4.3.2, we describe why we do not expect all estimators with noncanonical link functions to be unbiased.

Appendix C.7.2 on page 467 contains plots depicting the relative bias and coefficient of variation for all estimators in all samples. Appendix C.7.3 on page 474 contains tables with the relative bias and coefficients of variation for all estimators. The plots and tables tend to be similar, so we only present plots for the binary variable in this section.

In Figure 4.4, we see that the relative bias of the binary response was small and about the same for all estimators. As the figures in Appendix C.7.2 show, the estimators tended to be unbiased in the small and large samples. Of course, there were exceptions with the generalized difference estimator and projective estimators based on the cauchit, complementary-log-log, and log links.

We see most clearly in Figure 4.4 that the bias of the projective estimator is stubbornly large with the cauchit link and negative with the complementary log-log link, re-

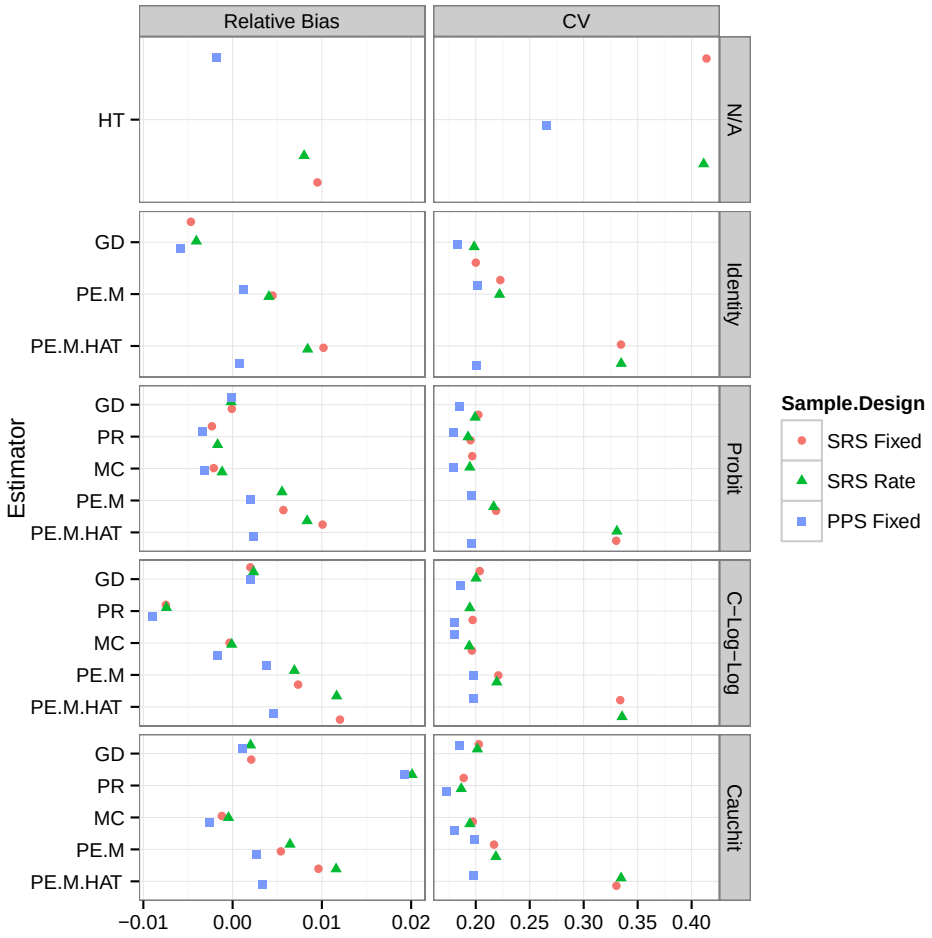

Figure 4.4: Plot of Relative Bias and Coefficient of Variation for all estimators of total binary response in small samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other.

gardless of the sample size. This bias was expected, given that the estimating equations for these two link functions are not calibrated and the projective estimator is not calibrated. Table 4.7 highlights the fact that a calibrated estimator must be used with these two link functions in order to produce approximately unbiased estimates.

Indeed, the bias of the projective estimator for the noncanonical link functions illustrates the fundamental thesis of Firth and Bennett (1998). That is, the bias of estimators depends on the form of the estimator as well as the process to estimate the parameters. Even though the estimating equations for the cauchit and complementary log-log link functions do not simplify to Equation (4.56), the generalized difference, model-calibrated, and model-calibrated maximum pseudoempirical likelihood estimators remain unbiased with these link functions because the constraint in Equation (4.56) is part of the model calibration process. Thus, the form of the estimator can overcome any bias associated with using a noncanonical link. Figure 4.4 clearly shows that if the estimating equations for $\mathbf{B}$ do not simplify to Equation (4.56), then these conditions should be built into the estimator through calibration. For this reason, the generalized difference, model-calibrated, and model-calibrated maximum pseudoempirical likelihood estimators add extra protection against biased predictions.

Figure 4.5 shows that in large samples, the generalized difference estimator, model-calibrated, and model-calibrated maximum pseudoempirical likelihood estimators are unbiased whilst the projective estimator remains above the true binary response total in large samples with the cauchit link. Since the estimator is inherently biased, we do not expect the bias to decrease as the sample size increases. Figures 4.4 and 4.5 confirm this.

In terms of the variability of our estimators, we see that they were always less

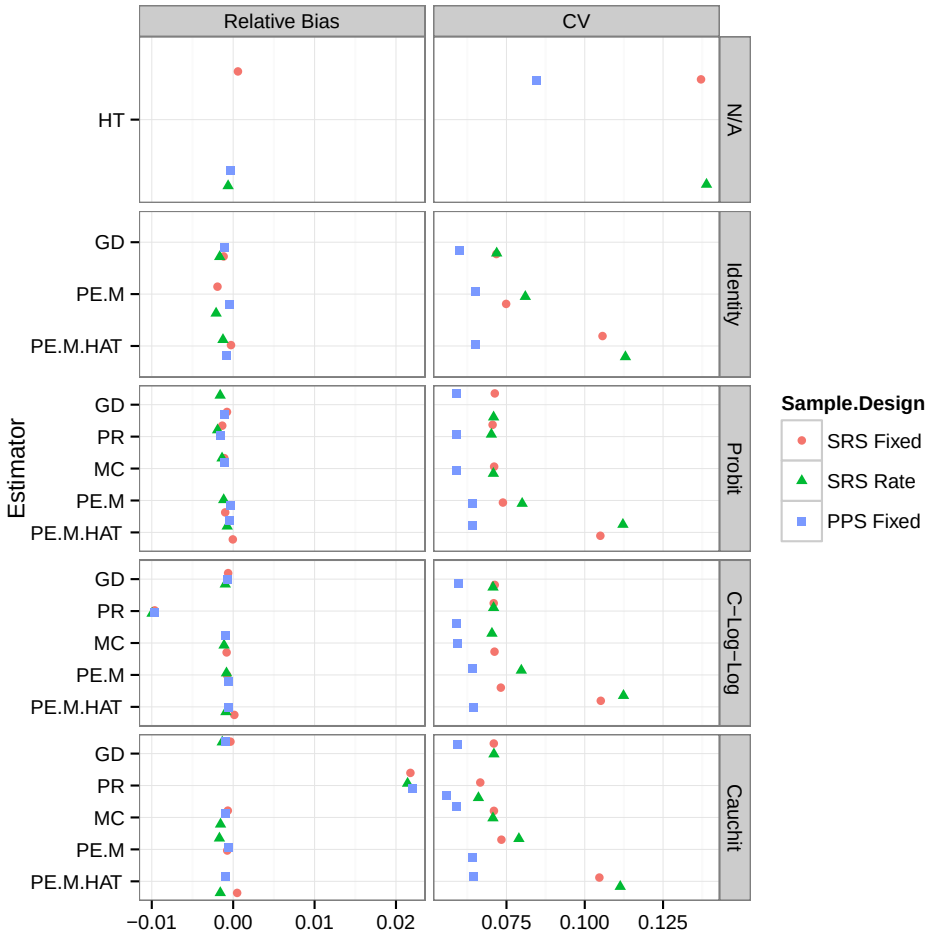

Figure 4.5: Plot of Relative Bias and Coefficient of Variation for all estimators of total binary response in large samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other.

variable than the π-estimator, conditional on the sample design. As we noted earlier, the model-calibrated maximum pseudomaximum likelihood estimator which uses $\widehat{M}$ is much more variable than the corresponding estimator based on M. The added variability due to estimating M, is quite large especially in small samples.

Estimators in the SRS Fixed and SRS Rate sample designs are regularly more variable than the same estimators in the probability proportional to size samples. Indeed, probability proportional to size samples tend to be more efficient than simple random samples, especially in clustered samples where the measure of size is correlated with the response.

In Figure C.2 on page 469, we see that the GLM-assisted estimators tend to be more efficient than the GREG estimator for the count variable. For the binary and synthetic responses, all of the model-assisted estimators are about as variable as the generalized difference estimator.

Our simulations showed that $\hat{\mathbf{t}}^{gd}$, $\hat{\mathbf{t}}^{mc}$, and $\hat{\mathbf{t}}^{peM}$ are relatively unbiased and have similar or smaller variances than the traditional GREG estimator. When complete auxiliary data are available for the population and GLMs fit the data better than the classic linear model, there can be significant gains in efficiently to using one of the GLM-assisted estimators. Since $\hat{\mathbf{t}}^{pr}$ is the simplest of the estimators and it performed as well as the competing GLM-assisted estimators, it may be preferable to the other estimators. In their simulation, Firth and Bennett (1998) also found that the projective estimator behaved similarly to the generalized difference estimator. On the other hand, if the estimating equations for μ_k significantly deviate from Equation (4.56) as is common is many non-canonical link functions, our simulations show that the generalized difference, model-

calibration, and model-calibrated maximum pseudoempirical likelihood estimators can be used to produce approximately unbiased estimates.

4.3.4.3 Variance Estimators for $\widehat{t}_y^{lg}$

In this section we report and discuss results of the variance estimators for the GLM-assisted difference estimator in the six simulations. We focus on the relative bias and the confidence interval coverage of $\upsilon_{wr}\left(\hat{t}_y^{gd}\right)$, $\upsilon_e\left(\hat{t}_y^{gd}\right)$, and $\upsilon_{Binder}\left(\hat{t}_y^{gd}\right)$.

In general, all three variance estimators are similar. As expected, confidence interval coverage tends to improve as the sample size increases. All estimators tend to perform better in the probability proportional to size samples where the totals tend to be more efficiently estimated. Despite some exceptions and small differences, the Binder and with-replacement variance estimators tend to have less bias than υ_e and have better confidence interval coverage. Although frequently used, in the six simulations υ_e tended to perform the worst.

In small samples, all three estimators underestimate the empirical variance. As a result, confidence interval coverage is less than the nominal value in small samples. Figure 4.6 shows the relative bias and confidence interval coverage of the three variance estimators of $\widehat{t}_{binary}^{lg}$ in the small samples. The top half of each box shows the variance estimators of $\widehat{t}_{binary}^{lg}$ whilst the bottom half shows the variance estimators of $\widehat{t}_{binary}^{mc}$.

The squares represent the probability proportional to size samples. In general, bias in the Fixed PPS samples is closest to 0 and confidence interval coverage is closest to 95%.

The purple shapes represent the Binder variance estimators and the green shapes are for the weighted residual variance estimators. Although there are exceptions, the Binder and with-replacement variance estimators tended to have smaller bias and better confidence interval coverage than the other estimators, conditional on the sample design. Results in the small samples are similar, regardless of the link function.

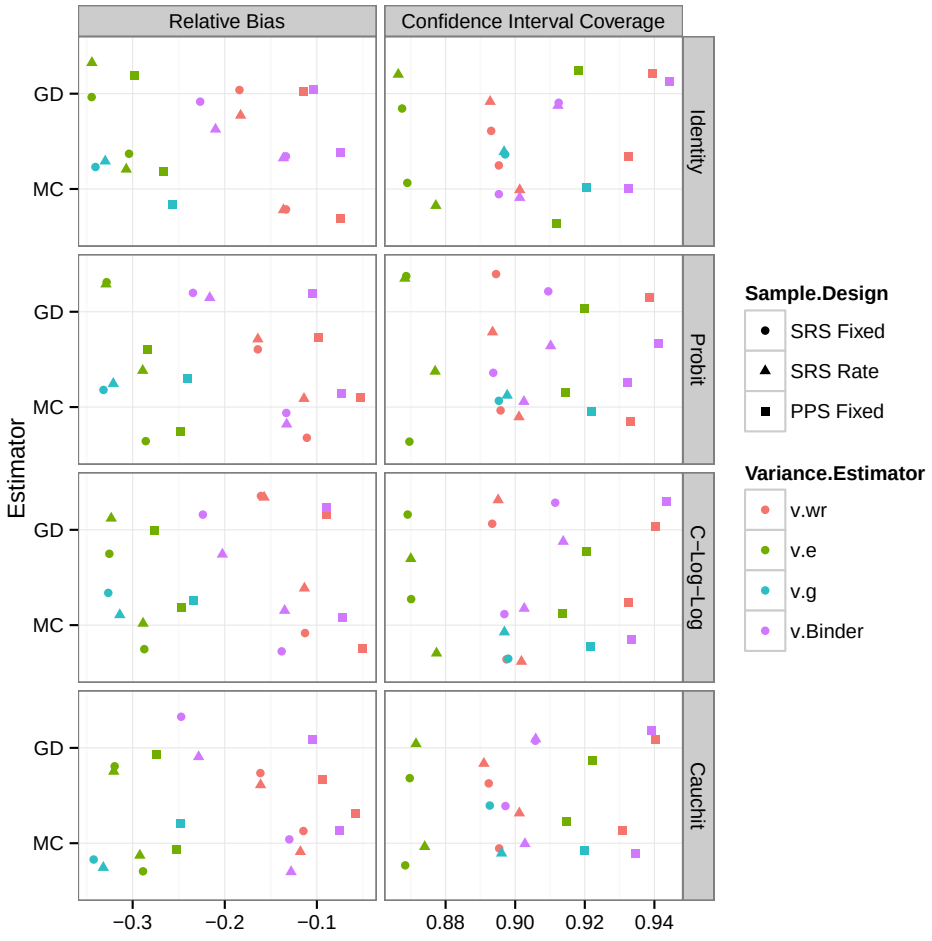

Figure 4.6: Plot of Relative Bias and Confidence Interval Coverage of variance estimators for the GLM-assisted difference estimator of the binary response in small samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other.

In large samples, we see evidence of slight negative bias for the variance estimators in SRS Fixed and SRS Rate samples. In the PPS Fixed samples, the with-replacement

and Binder estimators tend to overestimate the empirical variance by 5 to 10 percent in large samples.

As we seen in Figure 4.7, the variance estimators of the generalized difference estimator in large samples are similar, regardless of the link function.

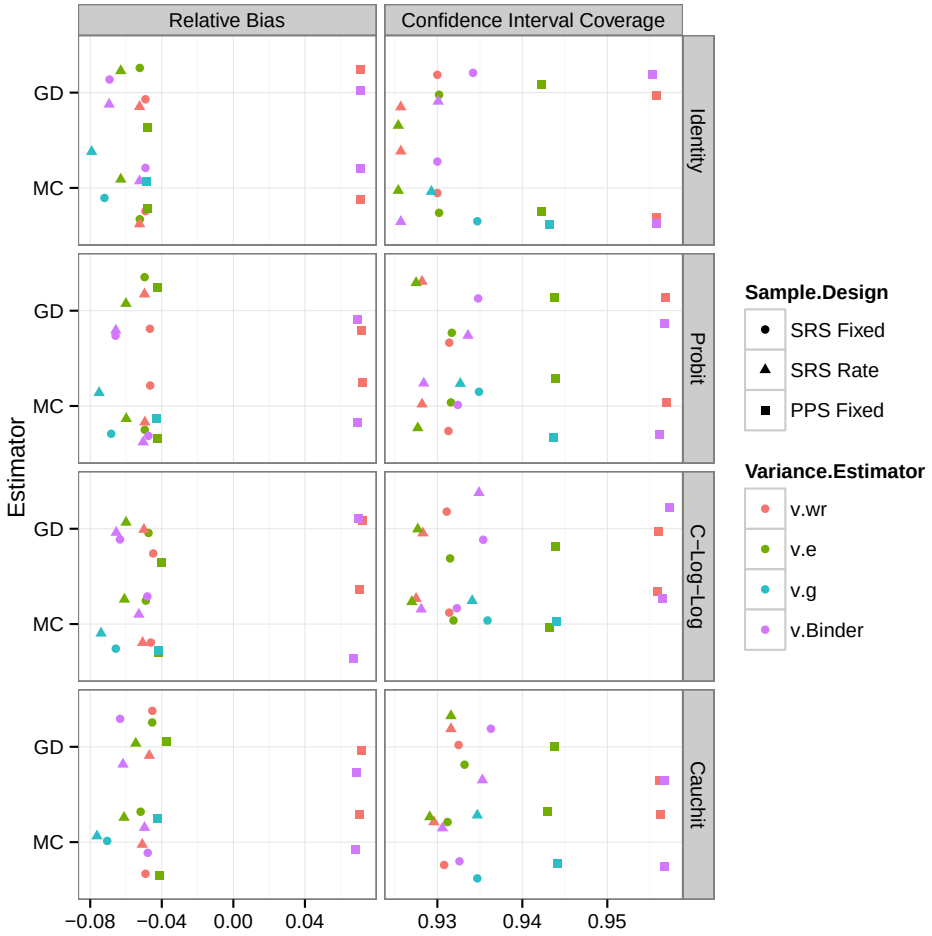

Figure 4.7: Plot of Relative Bias and Confidence Interval Coverage of variance estimators for the GLM-assisted difference estimator of the binary response in large samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other.

In terms of confidence interval coverage, we see improved coverage as the sample size increases. Regardless of the link function, the Binder variance estimator performed the best in terms of confidence interval coverage in the large samples. The with-

replacement and weighted residual estimators are similar, but the with-replacement estimator is often closer to 95% than the weighted residual estimator.

In conclusion, the Binder estimator seems to have the best confidence interval coverage in small and large samples. The performance of all three variance estimators is sensitive to the sample design.

4.3.4.4 Variance Estimators for $\widehat{t}_y^{mc}$ and $\widehat{t}_y^{peM}$

In this section we report and discuss results of the variance estimators for the model-calibrated and model-calibrated maximum pseudoempirical likelihood estimators in the six simulations. Since these two estimators are asymptotically equivalent, the same variance estimator can be used to estimate the variance of both estimators. We focus on the relative bias and the confidence interval coverage of $\upsilon_{wr}\left(\hat{t}_y^{mc}\right)$, $\upsilon_{ssw,e}\left(\hat{t}_y^{mc}\right)$, $\upsilon_g\left(\hat{t}_y^{mc}\right)$, and $\upsilon_{Binder}\left(\hat{t}_y^{mc}\right)$.

In general, results are very similar to those for the variance estimators of the GLM-assisted difference estimator. All four variance estimators are similar. As expected, confidence interval coverage tended to improve as the sample size increases. All estimators performed differently in the probability proportional to size samples. Although there are exceptions and the difference was small, the Binder and with-replacement variance estimators were very similar to each other in terms of their bias and confidence interval coverage. In fact, with the linear link, they seem to be equivalent. In terms of confidence interval coverage $\upsilon_g\left(\hat{t}_y^{mc}\right)$ tended to be better than $\upsilon_e\left(\hat{t}_y^{mc}\right)$, even though the bias of $\upsilon_e\left(\hat{t}_y^{mc}\right)$ tended to be smaller than $\upsilon_g\left(\hat{t}_y^{mc}\right)$. Although υ_g is commonly used, we found

that the Binder and with-replacement estimators performed better in terms of confidence interval coverage for the six simulations.

In small samples, all four estimators underestimate the empirical variance. As a result, confidence interval coverage is less than the nominal coverage. Figure 4.8 shows the relative bias and confidence interval coverage of the four variance estimators of $\widehat{t}^{mc}_{synthetic}$ in the small samples. The top half of each box shows the variance estimators of $\widehat{t}^{lg}_{synthetic}$ whilst the bottom half shows the variance estimators of $\widehat{t}^{mc}_{synhetic}$. With only a few exceptions, the weighted residual variance estimators tend to be between 65% to 80% of the empirical variance. On the other hand, the with-replacement and Binder variance estimators tend to be between 85% and 95% of the empirical variance.

The squares represent the probability proportional to size samples. In general all four variance estimators were relatively unbiased and their confidence interval coverage was close to 95% in the small samples.

The purple shapes represent the Binder variance estimators and the green shapes are for the weighted residual variance estimators. Although there are exceptions, the Binder variance and with-replacement variance estimators have smaller bias and better confidence interval coverage than the weighted residual variance estimators, conditional on the sample design. Results in the small samples are similar, regardless of the link function.

In the small samples, the confidence interval coverage was often between 88% and 94%. At the lowest end, we see $v_e\left(\hat{t}^{mc}_y\right)$ with confidence interval coverage consistently around 88% in the SRS cluster samples. At the high end, we see the Binder and with-replacement variance estimators around 95% in the probability proportional to size sam-

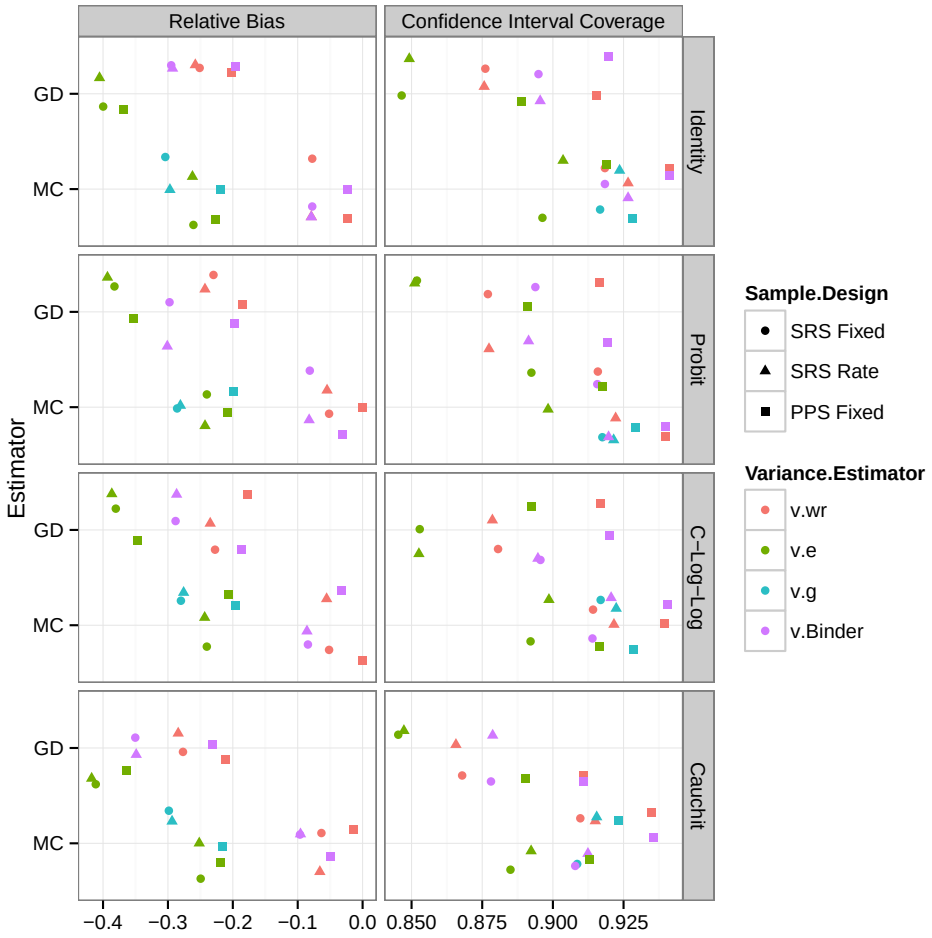

Figure 4.8: Plot of the Relative Bias and Confidence Interval Coverage for all estimators of the total synthetic response in small samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other.

ples. In the SRS samples, the Binder and with-replacement confidence intervals tend to contain the true value between 90 and 92 percent of the time in small samples.

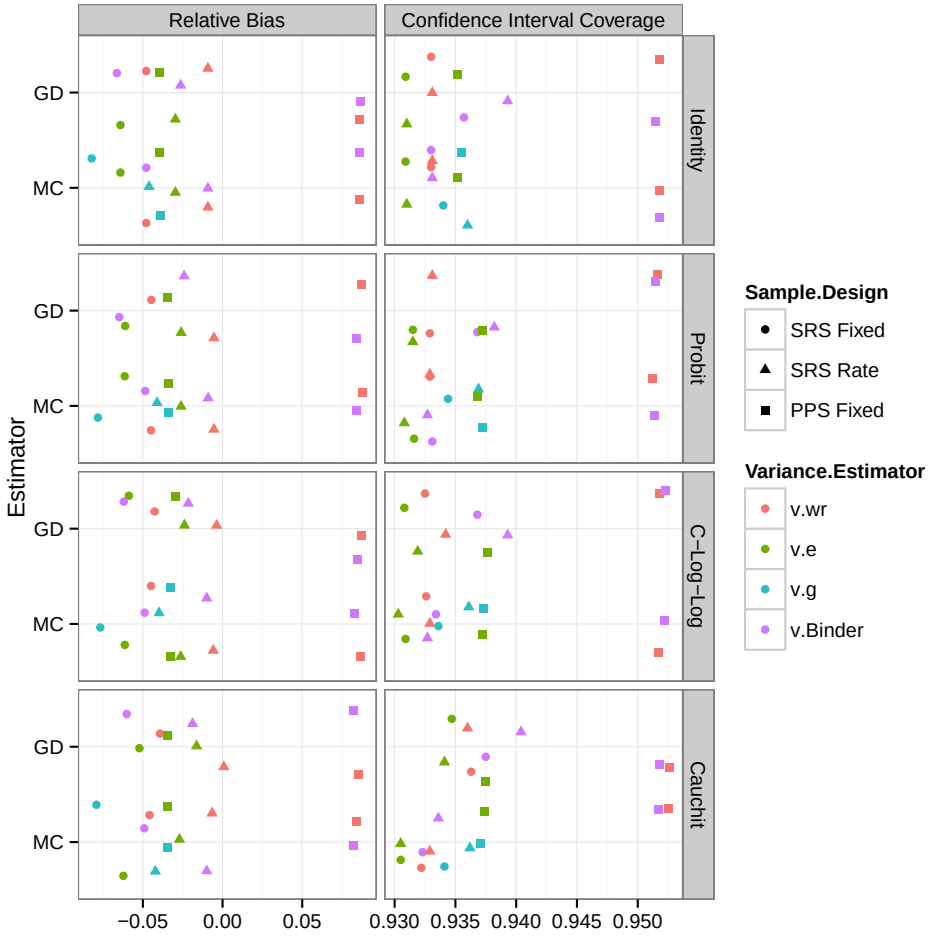

Figure 4.9: Plot of Relative Bias and Confidence Interval Coverage of variance estimators for the GLM-assisted difference estimator of the synthetic variable in large samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other.

Results for the four variance estimators of the model-calibrated estimator of the synthetic response in large samples are in Figure 4.9. In the large samples, we see that the variance estimators tended to be unbiased and the confidence interval coverage was between 93 and 96 percent. This represents improvement from the smaller samples.

Comparing the confidence interval coverage in the large samples, we see that the

g-weighted residual variance estimator was closer to the nominal 95% coverage rate than the other estimators, conditional on the sample design. This estimator did not perform as well in the smaller samples. Thus, as the sample size increases, this estimator improves.

In conclusion, we see that our variance estimators have different properties depending on the sample design and the size of the sample. In the small samples, the Binder estimator had the best confidence interval coverage, whilst in the large samples, the g-weighted residual variance estimator was closer to the nominal value. In the large probability proportional to size samples, the with-replacement and Binder estimators were positively biased and overestimate the nominal confidence interval coverage.

4.4 Conclusion

In this chapter, we extended the work of Wu and Sitter (2001) to cluster samples. Specifically, we proved that the GLM-assisted difference and model-calibrated estimators were design-consistent in cluster samples. We also derived the asymptotic variance of the two estimators and constructed with-replacement, weighted residual, and Binder variance estimators for the clustered GLM-assisted difference and model-calibrated estimators.

In a simulation, we compared the point and variance estimators under three sample designs in both small and large sample sizes. We found that all of the new point estimators could be more efficient than the π-estimator and sometimes less variable than the GREG estimator. The performance of the variance estimators depends on the sample size and sample design. In general, we found that the new Binder variance estimator was competitive with the more traditional variance estimators.

In the future, we would like to explore similar point estimators based on nonparametric models such as random trees and neural networks instead of generalized linear models. In some cases, we expect these alternative models to more closely fit our sample data. These new models are commonly used in industrial settings, but are much less common in survey statistics. Other extensions would be to compare replication variance estimators to the ones presented in this chapter. Calibrating at different levels of geography and incorporating dispersion parameters into the modeling process may also improve estimation in cluster samples.

Chapter 5

Conclusion

Calibration and generalized regression are frequently used to estimate totals from clustered samples drawn from finite populations. In this dissertation, we borrowed from the classical model-based theory to develop new calibrated point estimators and improved variance estimators. This dissertation showed that the model-based theory could be used to construct estimators with attractive design-based properties.

In Chapter 2, we focused on estimating the variance of the generalized regression (GREG) estimator in cluster samples. After deriving the model-based variance of the GREG estimator, we created five asymptotically unbiased estimators of that variance using leverage adjustments to the sandwich estimator. Furthermore, we proved that some of our new variance estimators were asymptotically equivalent to the delete-a-cluster Jackknife. We then evaluated the design-based properties of the new variance estimators in a large simulation involving three different populations and three different sample designs for both small and large samples. In general, our new variance estimators have better confidence interval coverage than more established estimators. On the other hand, the new variance estimators tend to be more complex and variable than the established variance estimators. We also found that the new variance estimators performed differently across simulations and even within each simulation. Although they are not uniformly better than established estimators, we showed that the new variance estimators should be worth

consideration and used when the circumstances warrant it.

Generalized regression estimators are frequently used to calibrate totals in cluster samples, regardless of the distribution of the underlying data. In Chapters 3 and 4, we derived new calibration estimators that are tailored to data generated from any member of the exponential distribution family. We proved that all of our new variance estimators are asymptotically unbiased. We also derived the asymptotic variance of all new estimators and constructed with-replacement, weighted residual, and implicit differentiation variance estimators of the asymptotic variance of our new estimators. In simulation studies we explored the performance of our new point and variance estimators for different populations and samples. In Chapter 3, we found examples where our new multinomial logistic-assisted estimators had much lower mean squared error than the GREG estimator. In Chapter 4, we were able to slightly improve upon the GREG estimator using generalized linear models (GLMs).

As already noted, our estimators come at a price. The sandwich variance estimators introduced in Chapter 2 are more variable and complicated than the GREG estimator. The calibrated multinomial and GLM estimators require complete auxiliary information for the entire population and do not result in one vector of weights that can support estimates for all response variables in a dataset. In spite of these limitations, our new estimators have clear theoretical and practical advantages over established estimators in some situations.

This dissertation has demonstrated that elements of the model-based framework can be used to improve design-based point and variance estimators in cluster samples. Since the model-based framework is so vast, there are many opportunities and avenues for

further research. Clear extensions of the papers in this dissertation include cluster-level models. In cases where complete information is unavailable for the population, but is available for all clusters, new calibrated point and variance estimators could be developed to improve estimation. The theoretical and applied properties of such new estimators would need to be derived and explored as we have done in this dissertation with our new estimators.

Appendix A

Notes for Variance of Clustered GREG Paper

A.1 Some Asymptotic Results

In our asymptotic framework, we assume that the number of population and sample clusters approach infinity; however, the number of population clusters increases at a faster rate than the number of sample clusters. We write this assumption as

Assumption 12. $\frac{n}{N} \longrightarrow 0$ *as* $n \longrightarrow \infty$ *and* $M \longrightarrow \infty$.

Additionally, the number of elements in each cluster is bounded so that no single cluster dominates. We write this assumption as

Assumption 13. *All* M_i *are bounded.*

We select our sample so that no element dominates our sample. That is, the unconditional probabilities of selection are bounded and approach 0 as the population size increases. We write this as,

Assumption 14. $\pi_k = O\left(\frac{m}{M}\right) \quad \forall k\,.$

This assumption can also be written as $N\max\pi_k = O\left(n\right)$. That is, our weights are roughly $O\left(\frac{N}{n}\right)$.

Fuller (2009, p. 40) describes an elegant asymptotic framework based on a sequence of nested populations. As the sample and population sizes increase, Fuller (2009) assumes

that the elements of $\mathbf{X}$ and $\mathbf{Q}^{-1}$ are all bounded. That is, we assume that

$$\lim_{\nu\to\infty} M_\nu^{-1}\sum_{k=1}^{M_\nu}\left(\mathbf{x}_{k\nu}, \mathbf{x}_{k\nu}^2\right) = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$$
$$\lim_{\nu\to\infty} M_\nu^{-1}\sum_{k=1}^{M_\nu}\left(q_{k\nu}^{-1}, q_{k\nu}^{-2}\right) = (\theta_3, \theta_4)$$
$$\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1^2 > 0$$

and

$$\theta_4 - \theta_3^2 > 0.$$

for a sequence of nested populations indexed by ν. For short, we write these assumptions by

Assumption 15. *All elements of* $\mathbf{X}$ *and* $\mathbf{Q}^{-1}$ *are bounded.*

Together, these assumptions imply that $\boldsymbol{t}_x$ and $\widehat{\boldsymbol{t}}_x$ are $O\left(M\right)$.

Furthermore, we make a similar assumption that the first and second moments of our response variable are bounded. That is,

$$\lim_{\nu\to\infty} M_\nu^{-1}\sum_{k=1}^{M_\nu}\left(y_{k\nu}, y_{k\nu}^2\right) = (\theta_5, \theta_6)$$
$$\theta_6 - \theta_5^2 > 0.$$

This can also be written as,

Assumption 16. *All elements of* $\mathbf{y}$ *and* $\boldsymbol{\Psi}$ *are bounded.*

A.1.1 Proof that $\mathbf{A}_\pi = O(N)$

By definition

$$\mathbf{A}_\pi = \mathbf{X}_s^\top \mathbf{Q}_s \mathbf{\Pi}^{-1} \mathbf{X}_s$$

By Assumption 15, $\mathbf{X}$ and $\mathbf{Q}$ are $O(1)$, elementwise. Since $\mathbf{\Pi} = O\left(\frac{m}{M}\right)$ elementwise by Assumption 14, we have $\mathbf{\Pi}^{-1} = O\left(\frac{M}{m}\right)$. Moreover, $\mathbf{A}_\pi$ can be written as the sum of m terms. Therefore

$$\begin{aligned}
\mathbf{A}_\pi &= \mathbf{X}_s^\top \mathbf{Q}_s \mathbf{\Pi}^{-1} \mathbf{X}_s \\
&= O\left(m \cdot 1 \cdot 1 \cdot \frac{M}{m} \cdot 1\right) \\
&= O(M).
\end{aligned}$$

Thus every element of $\mathbf{A}_\pi$ is $O(M)$ Since $M = N\bar{M}$, each element of $\mathbf{A}_\pi$ is also $O(N)$.

A.1.2 Proof that $\mathbf{g}_i = O(1)$

By definition

$$\mathbf{g}_i = \mathbf{Q}_i \mathbf{X}_i \mathbf{A}_s^{-1} \left(\mathbf{t}_x - \widehat{\mathbf{t}}_x \right) + \mathbf{1}$$

In Section A.1.1, we showed that $\mathbf{A}_s^{-1} = O\left(\frac{1}{M}\right)$. By Assumption 15, we see that $\boldsymbol{t}_x = O(M)$, and $\widehat{\boldsymbol{t}}_x = O(M)$. Lastly, $\mathbf{Q}_i$ and $\mathbf{X}_i$ are bounded by Assumption 15. Thus

$$\begin{aligned}
\mathbf{g}_i &= \mathbf{Q}_i \mathbf{X}_i \mathbf{A}_s^{-1} \left(\mathbf{t}_x - \widehat{\mathbf{t}}_x \right) + \mathbf{1} \\
&= O(1)\, O(1)\, O\left(\frac{1}{M}\right) [O(M) - O(M)] + O(1) \\
&= O\left(\frac{1}{M}\right) O(M) \\
&= O(1)
\end{aligned}$$

A.1.3 Proof that $\mathbf{H}_{ii} = O\left(n^{-1}\right)$

$$\underset{m_i \times m_j}{\mathbf{H}_{ij}} = \mathbf{X}_i \mathbf{A}_\pi^{-1} \mathbf{X}_j^\top \mathbf{Q}_j \mathbf{\Pi}_j^{-1}$$

In Section A.1.1, we showed that each element of $\mathbf{A}_\pi$ is $O\left(N\right)$. Thus, each element of $\mathbf{A}_\pi^{-1}$ is $O\left(\frac{1}{N}\right)$. By Assumption 15, all elements of $\mathbf{X}_i$ and $\mathbf{Q}_j$ are $O\left(1\right)$. By Assumption 14, each element of $\mathbf{\Pi}_j^{-1}$ is $O\left(\frac{M}{m}\right)$. Thus,

$$\begin{aligned}
\mathbf{H}_{ij} &= \mathbf{X}_i \mathbf{A}_\pi^{-1} \mathbf{X}_j^\top \mathbf{Q}_j \mathbf{\Pi}_j^{-1} \\
&= O\left(1\right) O\left(\frac{1}{M}\right) O\left(1\right) O\left(1\right) O\left(\frac{M}{m}\right) \\
&= O\left(\frac{1}{m}\right) \\
&= O\left(\frac{1}{n\bar{m}}\right) \\
&= O\left(\frac{1}{n}\right)
\end{aligned}$$

Thus every element of $\mathbf{H}_{ij}$ is $O\left(n^{-1}\right)$.

A.1.4 Proof that $\mathscr{Q}_i = O\left(n^{-1}\right)$

We now consider the asymptotic behavior of $\mathscr{Q}_i$. Starting with the definition of $\mathscr{Q}_i$ gives,

$$\mathscr{Q}_i = \mathbf{A}_\pi^{-1}\mathbf{X}_i^\top\mathbf{Q}_i\mathbf{\Pi}_i^{-1}\left(\mathbf{I} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i$$

Since, $\mathbf{H}_{ii} = O\left(n^{-1}\right)$, it follows that $\left(\mathbf{I} - \mathbf{H}_{ii}\right) \approx \mathbf{I}$. Thus,

$$\mathscr{Q}_i \approx \mathbf{A}_\pi^{-1}\mathbf{X}_i^\top\mathbf{Q}_i\mathbf{\Pi}_i^{-1}\mathbf{e}_i$$

In Section A.1.1, we showed that $\mathbf{A}_\pi^{-1} = O\left(\frac{1}{M}\right)$. By Assumption 15, $\mathbf{X}_i^\top$ and $\mathbf{Q}_i$ are $O\left(1\right)$. By Assumption 14, each element of $\mathbf{\Pi}^{-1}$ is $O\left(\frac{M}{m}\right)$. Lastly, by assumption 16, we see that $\mathbf{e}_i = (1)$. With these assumptions,

$$\begin{aligned}
\mathscr{Q}_i &\approx \mathbf{A}_\pi^{-1}\mathbf{X}_i^\top\mathbf{Q}_i\mathbf{\Pi}_i^{-1}\mathbf{e}_i \\
&= O\left(\frac{1}{M}\right) O\left(1\right) O\left(1\right) O\left(\frac{M}{m}\right) O\left(1\right) \\
&= O\left(\frac{1}{m}\right) \\
&= O\left(\frac{1}{n}\right)
\end{aligned}$$

Thus every element of $\mathscr{Q}_i$ is $O\left(n^{-1}\right)$.

A.1.5 Proof that $G_i - \overline{G} \approx -\mathbf{1}^\top \mathbf{\Pi}_i^{-1} \widehat{\mathbf{y}}_i + \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}^\top \mathbf{\Pi}_i^{-1} \widehat{\mathbf{y}}_i$

We now consider what happens to the elements of $\mathbf{G}_i$ as our sample and population sizes increase. Starting with the definition of $\mathbf{G}_i$ gives,

$$G_i - \overline{G} = \mathbf{1}^\top \mathbf{\Pi}_i^{-1} \left(\mathbf{I} - \mathbf{H}_{ii}\right)^{-1} \left[\mathbf{H}_{ii}\mathbf{y}_i - \widehat{\mathbf{y}}_i\right] - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}^\top \mathbf{\Pi}_i^{-1} \left(\mathbf{I} - \mathbf{H}_{ii}\right)^{-1} \left[\mathbf{H}_{ii}\mathbf{y}_i - \widehat{\mathbf{y}}_i\right]$$

In section, A.1.3, we showed that $\mathbf{H}_{ii} = O\left(n^{-1}\right)$. Thus, $\mathbf{H}_{ii}$ will approach $\mathbf{0}$ in large samples. Also, by Assumption 16, $\mathbf{y}_i$ is bounded. Using these two properties, we can subsitute to obtain

$$\begin{aligned} G_i - \overline{G} &\approx \mathbf{1}^\top \mathbf{\Pi}_i^{-1} \left(\mathbf{I}\right)^{-1} \left[-\widehat{\mathbf{y}}_i\right] - \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}^\top \mathbf{\Pi}_i^{-1} \left(\mathbf{I}\right)^{-1} \left[-\widehat{\mathbf{y}}_i\right] \\ &= -\mathbf{1}^\top \mathbf{\Pi}_i^{-1} \widehat{\mathbf{y}}_i + \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}^\top \mathbf{\Pi}_i^{-1} \widehat{\mathbf{y}}_i \end{aligned}$$

A.1.6 Proof that $K_i - \overline{K} \approx -n\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\mathbf{X}_i\widehat{\mathbf{B}} + \sum_{i=1}^n \mathbf{1}^\top \mathbf{\Pi}_i^{-1}\mathbf{X}_i\widehat{\mathbf{B}}$

We now consider what happens to the elements of $\mathbf{K}_i$ as our sample and population sizes increase. Starting with the definition of $\mathbf{K}_i$ gives,

$$K_i - \overline{K} = \left(\mathbf{1}^\top\mathbf{X} - n\mathbf{1}^\top\mathbf{\Pi}_i^{-1}\mathbf{X}_i\right)\left(\widehat{\mathbf{B}} - \mathscr{Q}_i\right) - \frac{1}{n}\sum_{i=1}^n \left(\mathbf{1}^\top\mathbf{X} - n\mathbf{1}^\top\mathbf{\Pi}_i^{-1}\mathbf{X}_i\right)\left(\widehat{\mathbf{B}} - \mathscr{Q}_i\right)$$

In section, A.1.4, we showed that $\mathbf{Q}_i = O\left(n^{-1}\right)$. Since $\widehat{\mathbf{B}} = O\left(1\right)$,

$$\begin{aligned}
K_i - \overline{K} &\approx \left(\mathbf{1}^\top\mathbf{X} - n\mathbf{1}^\top\mathbf{\Pi}_i^{-1}\mathbf{X}_i\right)\widehat{\mathbf{B}} - \frac{1}{n}\sum_{i=1}^n \left(\mathbf{1}^\top\mathbf{X} - n\mathbf{1}^\top\mathbf{\Pi}_i^{-1}\mathbf{X}_i\right)\widehat{\mathbf{B}} \\
&= \mathbf{1}^\top\mathbf{X}\widehat{\mathbf{B}} - n\mathbf{1}^\top\mathbf{\Pi}_i^{-1}\mathbf{X}_i\widehat{\mathbf{B}} - \mathbf{1}^\top\mathbf{X}\widehat{\mathbf{B}} + \sum_{i=1}^n \mathbf{1}^\top\mathbf{\Pi}_i^{-1}\mathbf{X}_i\widehat{\mathbf{B}} \\
&= -n\mathbf{1}^\top\mathbf{\Pi}_i^{-1}\mathbf{X}_i\widehat{\mathbf{B}} + \sum_{i=1}^n \mathbf{1}^\top\mathbf{\Pi}_i^{-1}\mathbf{X}_i\widehat{\mathbf{B}}.
\end{aligned}$$

A.1.7 Proof that $F_i = o(1)$

We now consider what happens to the elements of $\mathbf{F}_i$ under these conditions. Starting with the definition of $\mathbf{F}_i$ and making substitutions for G_i and K_i in sections A.1.5 and A.1.6 gives

$$\begin{aligned}
F_i &= \left(G_i - \overline{G}\right) - \frac{1}{n}\left(K_i - \overline{K}\right) \\
&\approx \left[-\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\widehat{\mathbf{y}}_i + \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\widehat{\mathbf{y}}_i\right] - \frac{1}{n}\left[-n\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\mathbf{X}_i\widehat{\mathbf{B}} + \sum_{i=1}^{n}\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\mathbf{X}_i\widehat{\mathbf{B}}\right] \\
&= -\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\widehat{\mathbf{y}}_i + \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\widehat{\mathbf{y}}_i + \mathbf{1}^\top \mathbf{\Pi}_i^{-1}\mathbf{X}_i\widehat{\mathbf{B}} - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\mathbf{X}_i\widehat{\mathbf{B}}
\end{aligned}$$

Since $\widehat{\mathbf{y}}_i = \mathbf{X}_i\widehat{\mathbf{B}}$, we have

$$\begin{aligned}
F_i &\approx -\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\widehat{\mathbf{y}}_i + \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\widehat{\mathbf{y}}_i + \mathbf{1}^\top \mathbf{\Pi}_i^{-1}\widehat{\mathbf{y}}_i - \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\widehat{\mathbf{y}}_i \\
&= 0
\end{aligned}$$

Thus, $F_i \approx 0$ and $F_i = o(1)$.

A.1.8 Proof that $D_i = O\left(\frac{N}{n}\right)$

Starting with the definition of D_i gives,

$$D_i = \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \left(\mathbf{I} - \mathbf{H}_{ii}\right)^{-1} \mathbf{e}_i.$$

Since $\mathbf{g}_i^\top$, $\left(\mathbf{I} - \mathbf{H}_{ii}\right)^{-1}$, and $\mathbf{e}_i$ are all bounded and $\mathbf{\Pi}_k = O\left(\frac{N}{n}\right)$, elementwise, we have

$$\begin{aligned} D_i &= \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \left(\mathbf{I} - \mathbf{H}_{ii}\right)^{-1} \mathbf{e}_i \\ &= O\left(1\right) O\left(\frac{N}{n}\right) O\left(1\right) O\left(1\right) \\ &= O\left(\frac{N}{n}\right). \end{aligned}$$

A.1.9 Proof that $\mathrm{var}_M(\mathbf{e}_i) \approx \mathbf{\Psi}_i$

We first rewrite the residual in terms of the hat matrix. Of course, the residual for cluster i is defined as,

$$\mathbf{e}_i = \mathbf{y}_i - \hat{\mathbf{y}}_i.$$

Writing the predicted value in terms of the hat matrix gives

$$\begin{aligned}
&= \mathbf{y}_i - \sum_{j=1}^{n} \mathbf{H}_{ij}\mathbf{y}_j \\
&= \mathbf{y}_i - \mathbf{H}_{ii}\mathbf{y}_i + \sum_{j \neq i} \mathbf{H}_{ij}\mathbf{y}_j \\
&= \left(\underset{m_i \times m_i}{\mathbf{I}} - \mathbf{H}_{ii} \right) \mathbf{y}_i - \sum_{j \neq i} \mathbf{H}_{ij}\mathbf{y}_j
\end{aligned}$$

Now consider the variance of $\mathbf{e}_i$

$$\begin{aligned}
\mathrm{var}_M(\mathbf{e}_i) &= \mathrm{var}_M\left[(\mathbf{I} - \mathbf{H}_{ii})\,\mathbf{y}_i - \sum_{j \neq i} \mathbf{H}_{ij}\mathbf{y}_j \right] \\
&= (\mathbf{I} - \mathbf{H}_{ii})\,\mathrm{var}_M(\mathbf{y}_i)\,(\mathbf{I} - \mathbf{H}_{ii})^\top + \sum_{j \neq i} \mathbf{H}_{ij}\mathrm{var}_M(\mathbf{y}_j)\,\mathbf{H}_{ij}^\top \\
&= (\mathbf{I} - \mathbf{H}_{ii})\,\mathbf{\Psi}_i\,(\mathbf{I} - \mathbf{H}_{ii})^\top + \sum_{j \neq i} \mathbf{H}_{ij}\mathbf{\Psi}_j\mathbf{H}_{ij}^\top.
\end{aligned}$$

In Section A.1.3 we showed that $\mathbf{H}_{ii} = O\left(\frac{1}{n}\right)$. Furthermore, by Assumption 16, $\mathbf{\Psi}_i = O(1)$, elementwise. Thus

$$\begin{aligned}
\mathrm{var}_M(\mathbf{e}_i) &= \mathbf{\Psi}_i + O\left(\frac{1}{n^2}\right) \\
&\cong \mathbf{\Psi}_i.
\end{aligned}$$

A.2 Derivation of Sample Hat Matrix for Clustered GREG

The sample hat matrix for a clustered population, denoted $\mathbf{H}_s$, satisfies the following equation:

$$\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}.$$

As we can see, the hat matrix puts the "hat" on $\mathbf{y}$. The hat matrix has several important uses. First, it plays an important part in taking the expected value of variance estimators. Secondly, its diagonal elements, called leverages, are a measure of an observation's influence on the regression model. The leverages can be used to determine outliers and form robust estimators. Finally, the hat matrix simplifies calculating the expectation of sandwich estimators.

To write the hat matrix, we first consider the prediction from our survey weighted linear model. The predicted value is

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\mathbf{B}}$$

Since all quantities are conditional on the sample, we remove the s subscript. We also assume that this model describes each element, cluster, and the full sample. That is $y_k = \mathbf{x}_k^\top \mathbf{B}$ for all k, $\mathbf{y}_i = \mathbf{X}_i^\top \mathbf{B}$ for all i, and $\mathbf{y}_s = \mathbf{X}\widehat{\mathbf{B}}$

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\mathbf{B}}$$

Substituting for $\widehat{\mathbf{B}}$ gives

$$\begin{aligned}\widehat{\mathbf{y}} &= \mathbf{X}\left(\mathbf{X}^\top \mathbf{Q}\boldsymbol{\Pi}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^\top \mathbf{Q}\boldsymbol{\Pi}^{-1}\mathbf{y} \\ &= \mathbf{X}\mathbf{A}_\pi^{-1}\mathbf{X}^\top \mathbf{Q}\boldsymbol{\Pi}^{-1}\mathbf{y}\end{aligned}$$

where

$$\mathbf{A}_\pi = \mathbf{X}^\top \mathbf{Q} \mathbf{\Pi}^{-1} \mathbf{X} = \sum_{k \in s} \frac{q_k \mathbf{x}_k \mathbf{x}_k^\top}{\pi_k}$$

and the elments of $\mathbf{A}_\pi$ are

$$a_{rc} = \sum_{k=1}^{n} \frac{q_{kk} x_{kr} x_{kc}}{\pi_{kk}}$$

$$\mathbf{Q} = \operatorname{diag}(\mathbf{q})$$

and

$$\mathbf{\Pi} = \operatorname{diag}(\pi_k)$$

Now, if we let

$$\mathbf{H} = \mathbf{X} \mathbf{A}_\pi^{-1} \mathbf{X}^\top \mathbf{Q} \mathbf{\Pi}^{-1}$$

Then, we see that $\mathbf{H}$ is the hat matrix. That is,

$$\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

In this case $\mathbf{H}$ is a full sample m by m matrix where $m = \sum_{i \in s} m_i$. For each cluster, we have,

$$\hat{\mathbf{y}}_i = \sum_{j=1}^{n} \mathbf{H}_{ij} \mathbf{y}_j$$

Where $\mathbf{H}$ is defined in terms of cluster components,

$$\begin{aligned} \underset{m \times m}{\mathbf{H}} &= \mathbf{X} \mathbf{A}_\pi^{-1} \mathbf{X}^\top \mathbf{Q} \mathbf{\Pi}^{-1} \\ &= \begin{bmatrix} \mathbf{X}_1 \mathbf{A}_\pi^{-1} \mathbf{X}_1^\top \mathbf{Q}_1 \mathbf{\Pi}_1^{-1} & \dots & \mathbf{X}_1 \mathbf{A}_\pi^{-1} \mathbf{X}_n^\top \mathbf{Q}_n \mathbf{\Pi}_n^{-1} \\ \vdots & \ddots & \vdots \\ \mathbf{X}_n \mathbf{A}_\pi^{-1} \mathbf{X}_1^\top \mathbf{Q}_1 \mathbf{\Pi}_1^{-1} & \dots & \mathbf{X}_n \mathbf{A}_\pi^{-1} \mathbf{X}_n^\top \mathbf{Q}_n \mathbf{\Pi}_n^{-1} \end{bmatrix} \end{aligned}$$

The block diagonal elements of $\mathbf{H}$ can be written as,

$$\underset{m_i \times m_i}{\mathbf{H}_{ii}} = \mathbf{X}_i \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{Q}_i \mathbf{\Pi}_i^{-1}$$

And the off diagonal elements of $\mathbf{H}$ can be written as,

$$\underset{m_i \times m_j}{\mathbf{H}_{ij}} = \mathbf{X}_i \mathbf{A}_\pi^{-1} \mathbf{X}_j^\top \mathbf{Q}_j \mathbf{\Pi}_j^{-1}$$

Li and Valliant (2009) discuss properties of the survey weighted hat matrix in single stage samples.

A.3 Model Variance of Clustered GREG

Before considering the sampling error of the GREG, we first consider the structure of the population element level covariance matrix (i.e. the variance of y_k for all elements in the population). From the design-based framework, it is common to assume that clusters were selected with-replacement. Thus, the indicators for whether different clusters are in the same sample are uncorrelated, but the indicators for elements within clusters may be correlated. The model-based parallel is to assume that units in different clusters are independent under the model. Both the design-based and model-based perspectives commonly require that $\mathrm{var}\left(\mathbf{y}\right)$ is a block diagonal matrix with each block corresponding to a cluster. For the population, the model covariance matrix is,

$$\begin{aligned}\mathrm{var}\left(\mathbf{y}_{\mathscr{U}}\right) &= \underset{M\times M}{\boldsymbol{\Psi}} \\ &= \begin{bmatrix} \boldsymbol{\Psi}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Psi}_N \end{bmatrix}\end{aligned}$$

where

$$\underset{M_i\times M_i}{\boldsymbol{\Psi}_i} = \mathrm{var}\left(y_{ik}, y_{il}\right) = \mathrm{E}\left[\left(y_{ik} - \mathrm{E}\left(y_{ik}\right)\right)\left(y_{il} - \mathrm{E}\left(y_{il}\right)\right)\right]$$

Thus, we are assuming that elements within clusters are correlated, but not among different clusters.

As a preliminary to the proof, we also consider some notation. Let $\mathbf{g}_i$ be the vector of g-weights for the elements in the ith cluster. That is, $\mathbf{g}_i = \mathbf{Q}_i\mathbf{X}_i\mathbf{A}_s^{-1}\left(\mathbf{t}_x - \widehat{\mathbf{t}}_x\right) + \mathbf{1}$. Furthermore, let $\mathbf{y}_{si}$ be the vector of all sample elements in cluster i and $\mathbf{y}_i$ be the vector

of all elements in cluster i.

The variance of the GREG, with respect to the working model is:

$$\operatorname{var}\left(\widehat{t}_y^g - t_y\right) = \operatorname{var}\left(\sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\mathbf{y}_{si} - \sum_{i\in\mathscr{U}} \mathbf{1}_i^\top \mathbf{y}_i\right)$$

Calculating the variance, with respect to our model, gives

$$\operatorname{var}\left(\widehat{t}_y^g - t_y\right) = \sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\mathbf{\Psi}_{si}\mathbf{\Pi}_i^{-1}\mathbf{g}_i - 2\operatorname{cov}\left(\sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\mathbf{y}_{si}, \sum_{i\in\mathscr{U}} \mathbf{1}_i^\top \mathbf{y}_i\right) + \mathbf{1}^\top\mathbf{\Psi}\mathbf{1}$$

Since $\sum_{i\in\mathscr{U}} \mathbf{1}_i^\top \mathbf{y}_i = \sum_{i\in s} \mathbf{1}_i^\top \mathbf{y}_i + \sum_{i\in r} \mathbf{1}_i^\top \mathbf{y}_i$, we have,

$$\begin{aligned}\operatorname{var}\left(\widehat{t}_y^g - t_y\right) &= \sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\mathbf{\Psi}_{si}\mathbf{\Pi}_i^{-1}\mathbf{g}_i - 2\operatorname{cov}\left(\sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\mathbf{y}_{si}, \sum_{i\in s} \mathbf{1}_i^\top \mathbf{y}_i\right) \\ &\quad - 2\operatorname{cov}\left(\sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\mathbf{y}_{si}, \sum_{i\in r} \mathbf{1}_i^\top \mathbf{y}_i\right) + \mathbf{1}^\top\mathbf{\Psi}\mathbf{1}\end{aligned}$$

Under our working model, the covariance between the sample and nonsample clusters is zero. Thus, we simplify to

$$\begin{aligned}\operatorname{var}\left(\widehat{t}_y^g - t_y\right) &= \sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\mathbf{\Psi}_{si}\mathbf{\Pi}_i^{-1}\mathbf{g}_i - 2\operatorname{cov}\left(\sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\mathbf{y}_{si}, \sum_{i\in s} \mathbf{1}_i^\top \mathbf{y}_i\right) + \mathbf{1}^\top\mathbf{\Psi}\mathbf{1} \\ &= \sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\mathbf{\Psi}_{si}\mathbf{\Pi}_i^{-1}\mathbf{g}_i - 2\sum_{i\in s}\left[\mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\operatorname{cov}\left(\mathbf{y}_{si}, \mathbf{y}_i\right)\mathbf{1}_i^\top\right] + \mathbf{1}^\top\mathbf{\Psi}\mathbf{1} \\ &= L_1 - 2L_2 + L_3\end{aligned}$$

In Section A.1.1, we showed that $\mathbf{A}_s^{-1} = O\left(\frac{1}{M}\right)$. In Section A.1.2, we showed that $\mathbf{g}_i = O\left(1\right)$.

Now, we consider the order of L_1. Both $\mathbf{g}_i$ and $\mathbf{\Psi}_{si}$ are bounded. Thus, $\mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\mathbf{\Psi}_{si}\mathbf{\Pi}_i^{-1}\mathbf{g}_i = O\left(\frac{M}{m}\right)$ and $L_1 = \sum_{i\in s} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\mathbf{\Psi}_{si}\mathbf{\Pi}_i^{-1}\mathbf{g}_i = O\left(m\frac{M^2}{m^2}\right) = O\left(\frac{M^2}{m}\right)$.

Also, consider the order of L_2. Since $\mathbf{\Psi}_{si}$ is bounded, it follows that $\operatorname{cov}\left(\mathbf{y}_{si}, \mathbf{y}_i\right) = O\left(1\right)$ as well. Since $\mathbf{g}_i$ is also bounded, we have, $\mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\operatorname{cov}\left(\mathbf{y}_{si}, \mathbf{y}_i\right)\mathbf{1}_i^\top = O\left(m_i M_i \frac{M}{m}\right)$.

The m_i and M_i terms come from the fact that $\mathbf{y}_{si}$ is an m_i dimensional vector and $\mathbf{y}_i$ is an M_i dimensional vector. The full summation $2\sum_{i\in s}\left[\mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\mathrm{cov}\left(\mathbf{y}_{si},\mathbf{y}_i\right)\mathbf{1}_i^\top\right] = O\left(n\bar{m}\bar{M}\frac{M}{m}\right)$ where $\bar{m}$ is the average number of elements selected in each cluster and $\bar{M}$ is the average number of elements per cluster. By Assumption 13, both $\bar{m}$ and $\bar{M}$ are bounded. The total number of elements in sample is $m = n\bar{m}$. Therefore, $2\sum_{i\in s}\left[\mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\mathrm{cov}\left(\mathbf{y}_{si},\mathbf{y}_i\right)\mathbf{1}_i^\top\right] = O\left(n\bar{m}\bar{M}\frac{M}{m}\right) = O\left(m\bar{M}\frac{M}{m}\right) = O\left(M\right)$.

Lastly, consider L_3. Under our working model and Assumption 15, $\mathbf{\Psi}$ has $n\bar{M}^2$ bounded terms. That is L_3 is the sum of N different M_i by M_i cluster matrices. Since M_i is also bounded, $L_3 = O\left(N\right)$.

When we compare the three terms, we see that L_1 dominates. For example, consider the ratio $\frac{L_3}{L_1} = O\left(\frac{M}{\frac{M^2}{m}}\right) = O\left(\frac{m}{M}\right)$. By Assumption 12, this approaches zero, suggesting that L_1 dominates.

A.4 Approximate Model Expectation of Sandwich Estimator for the Clustered GREG

We would like to take the expected value of our sandwich estimator.

$$\mathrm{E}\left(\upsilon_R\right) = \mathrm{E}\left[\sum_{i=1}^{n} \mathbf{g}_{si}^{\top}\mathbf{\Pi}_{si}^{-1}\mathbf{e}_{si}\mathbf{e}_{si}^{\top}\left(\mathbf{\Pi}_{si}^{-1}\right)^{\top}\mathbf{g}_{si}\right]$$

We now drop the sample subscripts, since we are dealing exclusively with sample quantities. Since, $\mathbf{g}$ and $\mathbf{\Pi}$ are constants with respect to our model, our expectation simplifies to

$$\mathrm{E}\left(\upsilon_R\right) = \sum_{i=1}^{n} \mathbf{g}_{i}^{\top}\mathbf{\Pi}_{i}^{-1}\mathrm{E}\left[\mathbf{e}_{i}\mathbf{e}_{i}^{\top}\right]\left(\mathbf{\Pi}_{i}^{-1}\right)^{\top}\mathbf{g}_{i}$$

By the definition of variance, we have

$$\mathrm{E}\left(\upsilon_R\right) = \sum_{i=1}^{n} \mathbf{g}_{i}^{\top}\mathbf{\Pi}_{i}^{-1}\mathrm{var}\left(\mathbf{e}_{i}\right)\left(\mathbf{\Pi}_{i}^{-1}\right)^{\top}\mathbf{g}_{i}$$

The model variance of $\mathbf{e}_i$ is complicated by the fact that $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ are not independent. To solve this problem, we recall from Section A.1.9 that $\mathbf{e}_i = \left(\underset{m_i \times m_i}{\mathbf{I}} - \mathbf{H}_{ii}\right)\mathbf{y}_i - \sum_{j \neq i}\mathbf{H}_{ij}\mathbf{y}_j$.

$$\mathrm{E}\left(\upsilon_R\right) = \sum_{i=1}^{n} \mathbf{g}_{i}^{\top}\mathbf{\Pi}_{i}^{-1}\mathrm{var}\left[\left(\mathbf{I} - \mathbf{H}_{ii}\right)\mathbf{y}_i - \sum_{j \neq i}\mathbf{H}_{ij}\mathbf{y}_j\right]\left(\mathbf{\Pi}_{i}^{-1}\right)^{\top}\mathbf{g}_{i}$$

As long as our clusters are independent, we can take the variance of each term separately

$$\begin{aligned}\mathrm{E}\left(\upsilon_R\right) &= \sum_{i=1}^{n} \mathbf{g}_{i}^{\top}\mathbf{\Pi}_{i}^{-1}\left\{\mathrm{var}\left[\left(\mathbf{I} - \mathbf{H}_{ii}\right)\mathbf{y}_i\right] + \mathrm{var}\left[\sum_{j \neq i}\mathbf{H}_{ij}\mathbf{y}_j\right]\right\}\left(\mathbf{\Pi}_{i}^{-1}\right)^{\top}\mathbf{g}_{i} \\ &= \sum_{i=1}^{n} \mathbf{g}_{i}^{\top}\mathbf{\Pi}_{i}^{-1}\left[\left(\mathbf{I} - \mathbf{H}_{ii}\right)\mathrm{var}\left(\mathbf{y}_i\right)\left(\mathbf{I} - \mathbf{H}_{ii}\right)^{\top} + \sum_{j \neq i}\mathbf{H}_{ij}\mathrm{var}\left(\mathbf{y}_j\right)\mathbf{H}_{ij}^{\top}\right]\left(\mathbf{\Pi}_{i}^{-1}\right)^{\top}\mathbf{g}_{i}\end{aligned}$$

Using the nested population asymptotic framework, we now take the limit of our expectation as the number of sample and population clusters increase. Assuming that $\mathbf{A}_\pi = O(N)$ elementwise, $\mathbf{A}_\pi^{-1} = O(N^{-1})$, and that m_i is bounded for all clusters, then $\mathbf{H}_{ij} = O(n^{-1})$. The first term in the brackets is $O(n^{-1})$ while the second term is $O(n^{-2})$. Thus, our asymptotic variance as n increases is

$$\begin{aligned}
\mathrm{E}(v_R) &= \sum_{i=1}^{n} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} \mathrm{var}(\mathbf{y}_i) \left(\mathbf{\Pi}_i^{-1}\right)^\top \mathbf{g}_i \\
&= L_1
\end{aligned}$$

Therefore, as long as our clusters are independent and $\mathbf{H}_{ij} = O(n^{-1})$, the sandwich estimator is approximately unbiased for the true model variance in large samples.

Although the sandwich estimator is unbiased for the true model variance in large samples, it is not unbiased for moderate and small samples. When the sample size is small to moderate, we can find leverage adjustments to make the sandwich estimator unbiased.

A.5 Delete-a-cluster Jackknife

A.5.1 Proof that $\widehat{\mathbf{B}}_{(i)} = \widehat{\mathbf{B}} - \mathbf{Q}_i$ for cluster samples

First, let the subscript of (i) denote removal of the ith cluster from the full sample matrix or vector. For example, $\widehat{\mathbf{B}}_{(i)}$ is an estimate of $\mathbf{B}$ based on all sample clusters, except cluster i. According to Yung and Rao (1996),

$$\widehat{\mathbf{B}}_{(i)} = \mathbf{A}_{\pi(i)}^{-1}\widehat{\mathbf{b}}_{(i)}$$

where

$$\mathbf{A}_{\pi(i)} = \sum_{j\neq i}\sum_{k=1}^{m_j}\frac{q_{ik}\mathbf{x}_{ik}\mathbf{x}_{ik}^\top}{\pi_{ik}} = \mathbf{X}_{(i)}^\top\mathbf{Q}_{(i)}\mathbf{\Pi}_{(i)}^{-1}\mathbf{X}_{(i)}$$
$$\widehat{\mathbf{b}}_{(i)} = \sum_{j\neq i}\sum_{k=1}^{m_j}\frac{q_{ik}\mathbf{x}_{ik}y_{ik}}{\pi_{ik}} = \mathbf{X}_{(i)}^\top\mathbf{Q}_{(i)}\mathbf{\Pi}_{(i)}^{-1}\mathbf{y}_{(i)}.$$

Thus,

$$\widehat{\mathbf{B}}_{(i)} = \left(\mathbf{X}_{(i)}^\top\mathbf{Q}_{(i)}\mathbf{\Pi}_{(i)}^{-1}\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}_{(i)}^\top\mathbf{Q}_{(i)}\mathbf{\Pi}_{(i)}^{-1}\mathbf{y}_{(i)}.$$

Substituting $\mathbf{W}_{(i)} = \mathbf{Q}_{(i)}\mathbf{\Pi}_{(i)}^{-1}$, we have,

$$\widehat{\mathbf{B}}_{(i)} = \left(\mathbf{X}_{(i)}^\top\mathbf{W}_{(i)}\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}_{(i)}^\top\mathbf{W}_{(i)}\mathbf{y}_{(i)}.$$

Although not necessary, we assume that $\mathbf{X}$ and $\mathbf{X}_{(i)}$ are full column rank to simplify our calculations. According to Lemma 9.5.1 in Valliant et al. (2000), we have

$$\begin{aligned}\widehat{\mathbf{B}}_{(i)} &= \left(\mathbf{A}_\pi^{-1} + \mathbf{A}_\pi^{-1}\mathbf{X}_i^\top\mathbf{W}_i\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\mathbf{X}_i\mathbf{A}_\pi^{-1}\right)\mathbf{X}_{(i)}^\top\mathbf{W}_{(i)}\mathbf{y}_{(i)}\\ &= \mathbf{A}_\pi^{-1}\mathbf{X}_{(i)}^\top\mathbf{W}_{(i)}\mathbf{y}_{(i)} + \mathbf{A}_\pi^{-1}\mathbf{X}_i^\top\mathbf{W}_i\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\mathbf{X}_i\mathbf{A}_\pi^{-1}\mathbf{X}_{(i)}^\top\mathbf{W}_{(i)}\mathbf{y}_{(i)}.\end{aligned}$$

Since $\mathbf{X}_{(i)}^\top \mathbf{W}_{(i)} \mathbf{y}_{(i)} = \mathbf{X}^\top \mathbf{W}\mathbf{y} - \mathbf{X}_i^\top \mathbf{W}_i \mathbf{y}_i$, and $\widehat{\mathbf{B}} = \mathbf{A}_\pi^{-1} \mathbf{X}^\top \mathbf{W}\mathbf{y}$

$$
\begin{aligned}
\widehat{\mathbf{B}}_{(i)} &= \mathbf{A}_\pi^{-1} \left(\mathbf{X}^\top \mathbf{W}\mathbf{y} - \mathbf{X}_i^\top \mathbf{W}_i \mathbf{y}_i \right) \\
&\quad + \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \left(\mathbf{I} - \mathbf{H}_{ii} \right)^{-1} \mathbf{X}_i \mathbf{A}_\pi^{-1} \left(\mathbf{X}^\top \mathbf{W}\mathbf{y} - \mathbf{X}_i \mathbf{W}_i \mathbf{y}_i \right) \\
&= \mathbf{A}_\pi^{-1} \mathbf{X}^\top \mathbf{W}\mathbf{y} - \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \mathbf{y}_i \\
&\quad + \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \left(\mathbf{I} - \mathbf{H}_{ii} \right)^{-1} \mathbf{X}_i \mathbf{A}_\pi^{-1} \mathbf{X}^\top \mathbf{W}\mathbf{y} - \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \left(\mathbf{I} - \mathbf{H}_{ii} \right)^{-1} \mathbf{X}_i \mathbf{A}_\pi^{-1} \mathbf{X}_i \mathbf{W}_i \mathbf{y}_i \\
&= \widehat{\mathbf{B}} - \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \mathbf{y}_i + \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \left(\mathbf{I} - \mathbf{H}_{ii} \right)^{-1} \mathbf{X}_i \widehat{\mathbf{B}} \\
&\quad - \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \left(\mathbf{I} - \mathbf{H}_{ii} \right)^{-1} \mathbf{X}_i \mathbf{A}_\pi^{-1} \mathbf{X}_i \mathbf{W}_i \mathbf{y}_i.
\end{aligned}
$$

Since $\widehat{\mathbf{y}}_i = \mathbf{X}_i^\top \widehat{\mathbf{B}}$ and $\mathbf{H}_{ii} = \mathbf{X}_i^\top \mathbf{A}_\pi^{-1} \mathbf{X}_i \mathbf{W}_i$,

$$
\begin{aligned}
\widehat{\mathbf{B}}_{(i)} &= \widehat{\mathbf{B}} - \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \mathbf{y}_i + \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \left(\mathbf{I} - \mathbf{H}_{ii} \right)^{-1} \widehat{\mathbf{y}}_i \\
&\quad - \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \left(\mathbf{I} - \mathbf{H}_{ii} \right)^{-1} \mathbf{X}_i \mathbf{A}_\pi^{-1} \mathbf{X}_i \mathbf{W}_i \mathbf{y}_i \\
&= \widehat{\mathbf{B}} - \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \mathbf{y}_i + \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \left(\mathbf{I} - \mathbf{H}_{ii} \right)^{-1} \widehat{\mathbf{y}}_i - \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \left(\mathbf{I} - \mathbf{H}_{ii} \right)^{-1} \mathbf{H}_{ii} \mathbf{y}_i.
\end{aligned}
$$

Since $\mathbf{I} = \left(\mathbf{I} - \mathbf{H}_{ii} \right)^{-1} \left(\mathbf{I} - \mathbf{H}_{ii} \right)$,

$$
\begin{aligned}
\widehat{\mathbf{B}}_{(i)} &= \widehat{\mathbf{B}} - \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \left(\mathbf{I} - \mathbf{H}_{ii} \right)^{-1} \left(\mathbf{I} - \mathbf{H}_{ii} \right) \mathbf{y}_i \\
&\quad + \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \left(\mathbf{I} - \mathbf{H}_{ii} \right)^{-1} \widehat{\mathbf{y}}_i - \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \left(\mathbf{I} - \mathbf{H}_{ii} \right)^{-1} \mathbf{H}_{ii} \mathbf{y}_i \\
&= \widehat{\mathbf{B}} - \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \left(\mathbf{I} - \mathbf{H}_{ii} \right)^{-1} \mathbf{y}_i + \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \left(\mathbf{I} - \mathbf{H}_{ii} \right)^{-1} \mathbf{H}_{ii} \mathbf{y}_i \\
&\quad + \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \left(\mathbf{I} - \mathbf{H}_{ii} \right)^{-1} \widehat{\mathbf{y}}_i - \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \left(\mathbf{I} - \mathbf{H}_{ii} \right)^{-1} \mathbf{H}_{ii} \mathbf{y}_i \\
&= \widehat{\mathbf{B}} - \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \left(\mathbf{I} - \mathbf{H}_{ii} \right)^{-1} \mathbf{y}_i + \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \left(\mathbf{I} - \mathbf{H}_{ii} \right)^{-1} \widehat{\mathbf{y}}_i \\
&= \widehat{\mathbf{B}} - \mathbf{A}_\pi^{-1} \mathbf{X}_i^\top \mathbf{W}_i \left(\mathbf{I} - \mathbf{H}_{ii} \right)^{-1} \left(\mathbf{y}_i - \widehat{\mathbf{y}}_i \right).
\end{aligned}
$$

Since $\mathbf{e}_i = \mathbf{y}_i - \widehat{\mathbf{y}}_i$,

$$\widehat{\mathbf{B}}_{(i)} = \widehat{\mathbf{B}} - \mathbf{A}_{\pi}^{-1}\mathbf{X}_i^{\top}\mathbf{W}_i\left(\mathbf{I} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i.$$

Recall that $\mathbf{W}_i = \mathbf{Q}^{-1}\mathbf{\Pi}_i^{-1}$. Letting $\mathscr{Q}_i = \mathbf{A}_{\pi}^{-1}\mathbf{X}_i^{\top}\mathbf{Q}_i\mathbf{\Pi}_i^{-1}\left(\mathbf{I} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i$, we obtain

$$\widehat{\mathbf{B}}_{(i)} = \widehat{\mathbf{B}} - \mathscr{Q}_i.$$

A.5.2 Jackknife variance estimator of clustered GREG in terms of leverages

We now simplify the delete-a-cluster Jackknife variance estimator of the clustered GREG. The estimated total after removing the ith cluster is defined as

$$
\begin{aligned}
\widehat{t}^{gr}_{y(i)} &= \frac{n}{n-1}\widehat{t}^{\pi}_{y,(i)} + \left[\mathbf{t}_x - \frac{n}{n-1}\widehat{\mathbf{t}}^{\pi}_{x,(i)}\right]\widehat{\mathbf{B}}_{(i)} \\
&= \frac{n}{n-1}\mathbf{1}^\top\mathbf{\Pi}^{-1}_{(i)}\mathbf{y}_{(i)} + \left[\mathbf{1}^\top\mathbf{X} - \frac{n}{n-1}\mathbf{1}^\top\mathbf{\Pi}^{-1}_{i}\mathbf{X}_{(i)}\right]\widehat{\mathbf{B}}_{(i)} \\
&= \frac{n}{n-1}\left(\mathbf{1}^\top\mathbf{\Pi}^{-1}\mathbf{y} - \mathbf{1}^\top\mathbf{\Pi}^{-1}_{i}\mathbf{y}_i\right) + \left[\mathbf{1}^\top\mathbf{X} - \frac{n}{n-1}\left(\mathbf{1}^\top\mathbf{\Pi}^{-1}\mathbf{X} - \mathbf{1}^\top\mathbf{\Pi}^{-1}_{i}\mathbf{X}_i\right)\right]\left(\widehat{\mathbf{B}} - \mathscr{Q}_i\right) \\
&= \frac{n\mathbf{1}^\top\mathbf{\Pi}^{-1}\mathbf{y}}{n-1} - \frac{n\mathbf{1}^\top\mathbf{\Pi}^{-1}_{i}\mathbf{y}_i}{n-1} + \left[\mathbf{1}^\top\mathbf{X} - \frac{n\mathbf{1}^\top\mathbf{\Pi}^{-1}\mathbf{X}}{n-1} + \frac{n\mathbf{1}^\top\mathbf{\Pi}^{-1}_{i}\mathbf{X}_i}{n-1}\right]\left(\widehat{\mathbf{B}} - \mathscr{Q}_i\right).
\end{aligned}
$$

Since $\frac{n}{n-1} - \frac{1}{n-1} = 1$, we have,

$$
\begin{aligned}
\widehat{t}^{gr}_{y(i)} &= \frac{n\mathbf{1}^\top\mathbf{\Pi}^{-1}\mathbf{y}}{n-1} - \frac{n\mathbf{1}^\top\mathbf{\Pi}^{-1}_{i}\mathbf{y}_i}{n-1} \\
&\quad + \left[\left(\frac{n}{n-1} - \frac{1}{n-1}\right)\mathbf{1}^\top\mathbf{X} - \frac{n\mathbf{1}^\top\mathbf{\Pi}^{-1}\mathbf{X}}{n-1} + \frac{n\mathbf{1}^\top\mathbf{\Pi}^{-1}_{i}\mathbf{X}_i}{n-1}\right]\left(\widehat{\mathbf{B}} - \mathscr{Q}_i\right) \\
&= \frac{n\mathbf{1}^\top\mathbf{\Pi}^{-1}\mathbf{y}}{n-1} - \frac{n\mathbf{1}^\top\mathbf{\Pi}^{-1}_{i}\mathbf{y}_i}{n-1} \\
&\quad + \left[\frac{n\mathbf{1}^\top\mathbf{X}}{n-1} - \frac{n\mathbf{1}^\top\mathbf{\Pi}^{-1}\mathbf{X}}{n-1} - \frac{\mathbf{1}^\top\mathbf{X}}{n-1} + \frac{n\mathbf{1}^\top\mathbf{\Pi}^{-1}_{i}\mathbf{X}_i}{n-1}\right]\left(\widehat{\mathbf{B}} - \mathscr{Q}_i\right) \\
&= \frac{n\mathbf{1}^\top\mathbf{\Pi}^{-1}\mathbf{y}}{n-1} - \frac{n\mathbf{1}^\top\mathbf{\Pi}^{-1}_{i}\mathbf{y}_i}{n-1} \\
&\quad + \left[\frac{n}{n-1}\left(\mathbf{1}^\top\mathbf{X} - \mathbf{1}^\top\mathbf{\Pi}^{-1}\mathbf{X}\right) - \frac{1}{n-1}\left(\mathbf{1}^\top\mathbf{X} - n\mathbf{1}^\top\mathbf{\Pi}^{-1}_{i}\mathbf{X}_i\right)\right]\left(\widehat{\mathbf{B}} - \mathbf{Q}_i\right) \\
&= \frac{n\mathbf{1}^\top\mathbf{\Pi}^{-1}\mathbf{y}}{n-1} - \frac{n\mathbf{1}^\top\mathbf{\Pi}^{-1}_{i}\mathbf{y}_i}{n-1} \\
&\quad + \frac{n}{n-1}\left(\mathbf{1}^\top\mathbf{X} - \mathbf{1}^\top\mathbf{\Pi}^{-1}\mathbf{X}\right)\left(\widehat{\mathbf{B}} - \mathscr{Q}_i\right) - \frac{1}{n-1}\left(\mathbf{1}^\top\mathbf{X} - n\mathbf{1}^\top\mathbf{\Pi}^{-1}_{i}\mathbf{X}_i\right)\left(\widehat{\mathbf{B}} - \mathscr{Q}_i\right).
\end{aligned}
$$

Letting $K_i = \left(\mathbf{1}^\top \mathbf{X} - n\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\mathbf{X}_i\right)\left(\widehat{\mathbf{B}} - \mathscr{Q}_i\right)$, we have

$$\begin{aligned}
\widehat{t}^{gr}_{y(i)} &= \frac{n\mathbf{1}^\top \mathbf{\Pi}^{-1}\mathbf{y}}{n-1} - \frac{n\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\mathbf{y}_i}{n-1} + \frac{n}{n-1}\left(\mathbf{1}^\top \mathbf{X} - \mathbf{1}^\top \mathbf{\Pi}^{-1}\mathbf{X}\right)\left(\widehat{\mathbf{B}} - \mathscr{Q}_i\right) - \frac{1}{n-1}K_i \\
&= \frac{n\mathbf{1}^\top \mathbf{\Pi}^{-1}\mathbf{y}}{n-1} - \frac{n\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\mathbf{y}_i}{n-1} \\
&\quad + \frac{n}{n-1}\left(\mathbf{1}^\top \mathbf{X} - \mathbf{1}^\top \mathbf{\Pi}^{-1}\mathbf{X}\right)\widehat{\mathbf{B}} - \frac{n}{n-1}\left(\mathbf{1}^\top \mathbf{X} - \mathbf{1}^\top \mathbf{\Pi}^{-1}\mathbf{X}\right)\mathscr{Q}_i - \frac{1}{n-1}K_i.
\end{aligned}$$

By definition, $\hat{t}^{gr}_y = \mathbf{1}^\top \mathbf{\Pi}^{-1}\mathbf{y} + \left(\mathbf{1}^\top \mathbf{X} - \mathbf{1}^\top \mathbf{\Pi}^{-1}\mathbf{X}\right)\widehat{\mathbf{B}}$. We simplify to

$$\widehat{t}^{gr}_{y(i)} = \frac{n}{n-1}\hat{t}^{gr}_y - \frac{n\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\mathbf{y}_i}{n-1} - \frac{n}{n-1}\left(\mathbf{1}^\top \mathbf{X} - \mathbf{1}^\top \mathbf{\Pi}^{-1}\mathbf{X}\right)\mathscr{Q}_i - \frac{1}{n-1}K_i.$$

Adding and subtracting $\frac{n}{n-1}\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\left(\mathbf{I} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i$ gives

$$\begin{aligned}
\widehat{t}^{gr}_{y(i)} &= \frac{n}{n-1}\hat{t}^{gr}_y - \frac{1}{n-1}K_i \\
&\quad + \frac{n}{n-1}\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\left(\mathbf{I} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i - \frac{n\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\mathbf{y}_i}{n-1} \\
&\quad - \frac{n}{n-1}\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\left(\mathbf{I} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i - \frac{n}{n-1}\left(\mathbf{1}^\top \mathbf{X} - \mathbf{1}^\top \mathbf{\Pi}^{-1}\mathbf{X}\right)\mathscr{Q}_i.
\end{aligned}$$

Applying the definitions of $\mathscr{Q}_i$ from section A.5.1 and $\mathbf{K}_i$ from section 2.3, and multiplying by $\left(\mathbf{I} - \mathbf{H}_{ii}\right)^{-1}\left(\mathbf{I} - \mathbf{H}_{ii}\right)$ gives

$$\begin{aligned}
\widehat{t}^{gr}_{y(i)} &= \frac{n}{n-1}\hat{t}^{gr}_y - \frac{1}{n-1}K_i \\
&\quad + \frac{n}{n-1}\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\left(\mathbf{I} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i - \frac{n\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\left(\mathbf{I} - \mathbf{H}_{ii}\right)^{-1}\left(\mathbf{I} - \mathbf{H}_{ii}\right)\mathbf{y}_i}{n-1} \\
&\quad - \frac{n}{n-1}\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\left(\mathbf{I} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i - \frac{n}{n-1}\left(\mathbf{1}^\top \mathbf{X} - \mathbf{1}^\top \mathbf{\Pi}^{-1}\mathbf{X}\right)\mathbf{A}^{-1}\mathbf{X}_i^\top \mathbf{Q}_i \mathbf{\Pi}_i^{-1}\left(\mathbf{I} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i. \\
&= \frac{n}{n-1}\hat{t}^{gr}_y - \frac{1}{n-1}K_i \\
&\quad + \frac{n}{n-1}\mathbf{1}^\top \mathbf{\Pi}_i^{-1}\left(\mathbf{I} - \mathbf{H}_{ii}\right)^{-1}\left[\mathbf{e}_i - \left(\mathbf{I} - \mathbf{H}_{ii}\right)\mathbf{y}_i\right] \\
&\quad - \frac{n}{n-1}\left[\mathbf{1}^\top + \left(\mathbf{1}^\top \mathbf{X} - \mathbf{1}^\top \mathbf{\Pi}^{-1}\mathbf{X}\right)\mathbf{A}^{-1}\mathbf{X}_i^\top \mathbf{Q}_i\right]\mathbf{\Pi}_i^{-1}\left(\mathbf{I} - \mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i.
\end{aligned}$$

Since $\mathbf{g}_i^\top = \mathbf{1}^\top + \left(\mathbf{1}^\top\mathbf{X} - \mathbf{1}^\top\mathbf{\Pi}^{-1}\mathbf{X}\right)\mathbf{A}^{-1}\mathbf{X}_i^\top\mathbf{Q}_i$

$$
\begin{aligned}
\widehat{t}_{y(i)}^{gr} &= \frac{n}{n-1}\hat{t}_y^{gr} - \frac{1}{n-1}K_i \\
&\quad + \frac{n}{n-1}\mathbf{1}^\top\mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\left[\mathbf{e}_i - \left(\mathbf{I}-\mathbf{H}_{ii}\right)\mathbf{y}_i\right] \\
&\quad - \frac{n}{n-1}\mathbf{g}_i^\top\mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i \\
&= \frac{n}{n-1}\hat{t}_y^{gr} - \frac{n}{n-1}\mathbf{g}_i^\top\mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i - \frac{1}{n-1}K_i \\
&\quad + \frac{n}{n-1}\mathbf{1}^\top\mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\left[\mathbf{e}_i - \left(\mathbf{I}-\mathbf{H}_{ii}\right)\mathbf{y}_i\right].
\end{aligned}
$$

Since $\mathbf{e}_i = \mathbf{y}_i - \widehat{\mathbf{y}}_i$

$$
\begin{aligned}
\widehat{t}_{y(i)}^{gr} &= \frac{n}{n-1}\hat{t}_y^{gr} - \frac{n}{n-1}\mathbf{g}_i^\top\mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i - \frac{1}{n-1}K_i \\
&\quad + \frac{n}{n-1}\mathbf{1}^\top\mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\left[\mathbf{y}_i - \widehat{\mathbf{y}}_i - \left(\mathbf{I}-\mathbf{H}_{ii}\right)\mathbf{y}_i\right] \\
&= \frac{n}{n-1}\hat{t}_y^{gr} - \frac{n}{n-1}\mathbf{g}_i^\top\mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i - \frac{1}{n-1}K_i \\
&\quad + \frac{n}{n-1}\mathbf{1}^\top\mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\left[\mathbf{y}_i - \widehat{\mathbf{y}}_i - \mathbf{I}\mathbf{y}_i + \mathbf{H}_{ii}\mathbf{y}_i\right] \\
&= \frac{n}{n-1}\hat{t}_y^{gr} - \frac{n}{n-1}\mathbf{g}_i^\top\mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i - \frac{1}{n-1}K_i \\
&\quad + \frac{n}{n-1}\mathbf{1}^\top\mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\left[-\widehat{\mathbf{y}}_i + \mathbf{H}_{ii}\mathbf{y}_i\right] \\
&= \frac{n}{n-1}\hat{t}_y^{gr} - \frac{n}{n-1}\mathbf{g}_i^\top\mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i - \frac{1}{n-1}K_i \\
&\quad + \frac{n}{n-1}\mathbf{1}^\top\mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\left[\mathbf{H}_{ii}\mathbf{y}_i - \widehat{\mathbf{y}}_i\right].
\end{aligned}
$$

Letting $G_i = \mathbf{1}^\top\mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\left[\mathbf{H}_{ii}\mathbf{y}_i - \widehat{\mathbf{y}}_i\right]$, we have

$$
\widehat{t}_{y(i)}^{gr} = \frac{n}{n-1}\hat{t}_y^{gr} - \frac{n}{n-1}\mathbf{g}_i^\top\mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i + \frac{n}{n-1}G_i - \frac{1}{n-1}K_i.
$$

To construct the Jackknife, we need to find the difference between the estimate having deleted a cluster and the average of the estimates. That is, we need to compute $\Delta_{J(i)} = \hat{t}^g_{y(i)} - \hat{t}^g_{y(\cdot)}$ where $\hat{t}^g_{y(\cdot)} = \frac{1}{n}\sum_{i=1}^n \hat{t}^g_{y(i)}$.

Substituting for $\widehat{t}^{gr}_{y(i)}$ gives

$$\begin{aligned}
\Delta_{J(i)} &= \frac{n}{n-1}\hat{t}^{gr}_y - \frac{n}{n-1}\mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i + \frac{n}{n-1}G_i - \frac{1}{n-1}K_i \\
&\quad - \frac{1}{n}\sum_{i=1}^n \left[\frac{n}{n-1}\hat{t}^{gr}_y - \frac{n}{n-1}\mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i + \frac{n}{n-1}G_i - \frac{1}{n-1}K_i\right] \\
&= \frac{n}{n-1}\hat{t}^{gr}_y - \frac{n}{n-1}\mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i + \frac{n}{n-1}G_i - \frac{1}{n-1}K_i \\
&\quad - \frac{n}{n-1}\hat{t}^{gr}_y + \frac{1}{n-1}\sum_{i=1}^n \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i - \frac{n}{n-1}\overline{G} + \frac{1}{n-1}\overline{K} \\
&= -\frac{n}{n-1}\left(\mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i - \frac{1}{n}\sum_{i=1}^n \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i\right) \\
&\quad + \frac{n}{n-1}\left(G_i - \overline{G}\right) - \frac{1}{n-1}\left(K_i - \overline{K}\right).
\end{aligned}$$

Letting $D_i = \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i$, we have

$$\begin{aligned}
\Delta_{J(i)} &= -\frac{n}{n-1}\left(D_i - \overline{D}\right) + \frac{n}{n-1}\left(G_i - \overline{G}\right) - \frac{1}{n-1}\left(K_i - \overline{K}\right) \\
&= -\frac{n}{n-1}\left(D_i - \overline{D}\right) + \frac{n}{n-1}\left[\left(G_i - \overline{G}\right) - \frac{1}{n}\left(K_i - \overline{K}\right)\right].
\end{aligned}$$

Letting $F_i = \left(G_i - \overline{G}\right) - \frac{1}{n}\left(K_i - \overline{K}\right)$ gives

$$\Delta_{J(i)} = -\frac{n}{n-1}\left(D_i - \overline{D}\right) + \frac{n}{n-1}F_i.$$

We can now simplify the Jackknife variance estimator as

$$
\begin{aligned}
v_{Jack} &= \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{t}^{g}_{y(i)} - \hat{t}^{g}_{y(\cdot)} \right)^2 \\
&= \frac{n-1}{n} \sum_{i=1}^{n} \left(\Delta_{J(i)} \right)^2 \\
&= \frac{n-1}{n} \sum_{i=1}^{n} \left(-\frac{n}{n-1} \left(D_i - \overline{D} \right) + \frac{n}{n-1} F_i \right)^2 \\
&= \frac{n}{n-1} \sum_{i=1}^{n} \left(-\left(D_i - \overline{D} \right) + F_i \right)^2 \\
&= \frac{n}{n-1} \sum_{i=1}^{n} \left[\left(D_i - \overline{D} \right)^2 - 2 \left(D_i - \overline{D} \right) F_i + F_i^2 \right] \\
&= \frac{n}{n-1} \left[\sum_{i=1}^{n} \left(D_i - \overline{D} \right)^2 - 2 \sum_{i=1}^{n} \left(D_i - \overline{D} \right) F_i + \sum_{i=1}^{n} F_i^2 \right]
\end{aligned}
$$

A.5.3 Jackknife variance estimator of clustered GREG in large samples

In section A.5.2 we showed that

$$\upsilon_{Jack} = \frac{n}{n-1}\left[\sum_{i=1}^{n}\left(D_i - \overline{D}\right)^2 - 2\sum_{i=1}^{n}\left(D_i - \overline{D}\right)F_i + \sum_{i=1}^{n}F_i^2\right]$$

We now consider υ_{Jack} when the number of sample and population clusters is large. In section A.1.7 we show that $F_i \approx 0$. Thus, we conclude that $\sum_{i=1}^{n} F_i^2 \approx 0$.

In section A.1.8 we show that $D_i = O\left(\frac{N}{n}\right)$; however, since $F_i \approx 0$, we conclude that $-2\sum_{i=1}^{n}\left(D_i - \overline{D}\right)F_i \approx 0$. Lastly,

$$\begin{aligned}
\sum_{i=1}^{n}\left(D_i - \overline{D}\right)^2 &= \sum_{i=1}^{n}\left[O\left(\frac{N}{n}\right)\right]^2 \\
&= \sum_{i=1}^{n} O\left(\frac{N^2}{n^2}\right) \\
&= nO\left(\frac{N^2}{n^2}\right) \\
&= O\left(\frac{N^2}{n}\right)
\end{aligned}$$

Thus, the first term of υ_J dominates and

$$\begin{aligned}
\upsilon_{Jack} &\approx \frac{n}{n-1}\left[\sum_{i=1}^{n}\left(D_i - \overline{D}\right)^2\right] \\
&= \upsilon_{J1}
\end{aligned}$$

A.5.4 Further simplification for Jackknife variance estimator of clustered GREG in large samples

We begin with v_{J1} and show that it is asymptotically equivalent to v_J. In section A.5.3, we showed that the Jackknife is asymptotically equivalent to v_{J1} where,

$$v_{J1} = \frac{n}{n-1}\left[\sum_{i=1}^{n}\left(D_i - \overline{D}\right)^2\right].$$

We divide by N^2 to keep v_{J1} from approaching infinity as n and N get large. Further, substituting for D_i gives

$$\begin{aligned}\frac{1}{N^2}v_{J1} &= \frac{1}{N^2}\frac{n}{n-1}\left[\sum_{i=1}^{n}\left(\mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i - \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i\right)^2\right] \\ &= \frac{n}{n-1}\left\{\sum_{i=1}^{n}\left[\mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i\right]^2 - \frac{1}{n}\left[\sum_{i=1}^{n}\mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i\right]^2\right\}.\end{aligned}$$

Now, we show that this second term converges in probability to zero. First, we write the second term in terms of full sample vectors

$$\frac{1}{n}\left[\sum_{i=1}^{n}\mathbf{g}_i^\top \mathbf{\Pi}_i^{-1}\left(\mathbf{I}-\mathbf{H}_{ii}\right)^{-1}\mathbf{e}_i\right]^2 = \frac{1}{n}\mathbf{g}^\top \mathbf{\Pi}^{-1}\left(\mathbf{I}-\text{blkdiag}\left(\mathbf{H}\right)\right)^{-1}\mathbf{e}\mathbf{e}^\top\left[\mathbf{I}-\text{blkdiag}\left(\mathbf{H}\right)\right]^{-1}\mathbf{\Pi}^{-1}\mathbf{g}$$

which is the square of $B = \frac{1}{\sqrt{n}}\mathbf{g}^\top \mathbf{\Pi}^{-1}\left(\mathbf{I}-\text{blkdiag}\left(\mathbf{H}\right)\right)^{-1}\mathbf{e} = \frac{1}{\sqrt{n}}\mathbf{g}^\top \mathbf{\Pi}^{-1}\left(\mathbf{U}\right)^{-1}\mathbf{e}$ with $\mathbf{U} = \mathbf{I} - \text{blkdiag}\left(\mathbf{H}\right)$. Since the expected value of $\mathbf{e}$ is 0 with respect to our model, the expected value of B is also 0. Now, the model variance of B is

$$\begin{aligned}\text{var}_M\left(B\right) &= \text{var}_M\left[\frac{1}{\sqrt{n}}\mathbf{g}^\top \mathbf{\Pi}^{-1}\mathbf{U}^{-1}\mathbf{e}\right] \\ &= \frac{1}{n}\mathbf{g}^\top \mathbf{\Pi}^{-1}\mathbf{U}^{-1}\text{var}_M\left(\mathbf{e}\right)\mathbf{U}^{-1}\mathbf{\Pi}^{-1}\mathbf{g}.\end{aligned}$$

Since $\mathbf{e} = \left(\mathbf{I}-\mathbf{H}\right)\mathbf{y}$, our variance can be written as

$$\text{var}_M\left(B\right) = \frac{1}{n}\mathbf{g}^\top \mathbf{\Pi}^{-1}\mathbf{U}^{-1}\left(\mathbf{I}-\mathbf{H}\right)\mathbf{\Psi}\left(\mathbf{I}-\mathbf{H}\right)\mathbf{U}^{-1}\mathbf{\Pi}^{-1}\mathbf{g}$$

Under our assumptions $\mathbf{g} = O(1)$, $(\mathbf{I} - \mathbf{H}) \approx \mathbf{I}$, $\mathbf{U}^{-1} \approx \mathbf{I}$, $\mathbf{\Psi} = O(1)$, $\mathbf{\Pi}^{-1} = O\left(\frac{N}{n}\right)$, and $\mathrm{var}_M(B)$ is the sum of $m = n\bar{m}$ terms. Thus,

$$\begin{aligned}\mathrm{var}_M(B) &= O\left(\frac{n}{n} \cdot \frac{N}{n} \cdot \frac{N}{n}\right) \\ &= O\left(\frac{N^2}{n^2}\right).\end{aligned}$$

Consequently

$$\frac{1}{N^2}\mathrm{var}_M(B) = O\left(\frac{1}{n^2}\right).$$

Since $\frac{1}{N^2}\frac{1}{n}\left[\sum_{i=1}^{n} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} (\mathbf{I} - \mathbf{H}_{ii})^{-1} \mathbf{e}_i\right]^2$ is the square of a term with mean 0 and variance approaching 0, as the number of clusters increases, we conclude that

$$\frac{1}{n}\left[\sum_{i=1}^{n} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} (\mathbf{I} - \mathbf{H}_{ii})^{-1} \mathbf{e}_i\right]^2 \xrightarrow{\text{prob}} 0$$

by Chebyshev's inequality. That is, $\frac{1}{n}\left[\sum_{i=1}^{n} \mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} (\mathbf{I} - \mathbf{H}_{ii})^{-1} \mathbf{e}_i\right]^2$ converges to 0 in probability as the number of clusters gets large. In contrast, the first term of $\frac{1}{N^2}\mathrm{var}_M(B) = O\left(\frac{1}{n}\right)$. Given this, an alternative large sample form of υ_J is simply

$$\upsilon_J = \frac{n}{n-1}\sum_{i=1}^{n}\left[\mathbf{g}_i^\top \mathbf{\Pi}_i^{-1} (\mathbf{I} - \mathbf{H}_{ii})^{-1} \mathbf{e}_i\right]^2$$

We also see that this is equivalent to υ_J in (2.39).

A.6 Full Tables

Table A.1: Simulation Results of Variance Estimators for Clustered GREG Estimate

Estimator	srs fixed		srs epsem		pps epsem	
	Third Grade Population					
	n = 25	n = 50	n = 25	n = 50	n = 25	n = 50
Average $\frac{\hat{t}_y^\pi}{N}$	477.2	477.1	476.3	476.9	477.3	477.8
rmse $\frac{\hat{t}_y^\pi}{N}$	25.8	16.3	44.9	31.3	12.0	7.3
Average $\frac{\hat{t}_y^g}{N}$	474.3	476.4	476.9	477.2	477.5	477.9
rmse $\frac{\hat{t}_y^g}{N}$	14.8	8.2	10.7	7.1	11.0	6.4
$\sqrt{v_g}$	12.4	7.6	9.0	6.3	8.8	6.1
$\sqrt{v_{wr}}$	12.5	8.5	9.3	7.2	9.3	7.0
$\sqrt{v_{JL}}$	13.3	8.7	9.7	7.3	9.6	7.1
$\sqrt{v_r}$	13.2	8.7	9.5	7.3	9.4	7.0
$\sqrt{v_D}$	15.5	9.3	10.9	7.8	10.6	7.5
$\sqrt{v_J}$	18.9	10.0	12.7	8.4	12.0	7.9
$\sqrt{v_{Jack}}$	18.2	9.8	12.4	8.3	11.8	7.8
$\sqrt{v_{J1}}$	19.0	10.0	12.9	8.5	12.2	8.0
$\sqrt{v_r^\star}$	11.9	6.9	8.6	5.8	8.4	5.5
$\sqrt{v_D^\star}$	14.0	7.3	9.8	6.2	9.5	5.8
$\sqrt{v_J^\star}$	17.0	7.9	11.4	6.7	10.8	6.2
$\sqrt{v_{Jack}^\star}$	16.4	7.8	11.2	6.6	10.6	6.1
$\sqrt{v_{J1}^\star}$	17.1	7.9	11.7	6.7	11.0	6.2
	ACS Population (numbers in thousands)					
	n = 3	n = 15	n = 3	n = 15	n = 3	n = 15
Average $\frac{\hat{t}_y^\pi}{N}$	1119.1	1108.2	1112.9	1113.9	1111.5	1109.0
rmse $\frac{\hat{t}_y^\pi}{N}$	425.8	166.3	449.0	181.5	126.5	51.2
Average $\frac{\hat{t}_y^g}{N}$	1081.7	1103.3	1104.5	1108.5	1106.4	1108.5
rmse $\frac{\hat{t}_y^g}{N}$	105.9	30.4	46.0	20.2	43.3	18.8
$\sqrt{v_g}$	71.1	26.6	25.7	17.6	27.7	17.5
$\sqrt{v_{wr}}$	67.6	27.1	29.5	19.1	33.5	19.1
$\sqrt{v_{JL}}$	64.3	26.6	30.6	19.0	32.9	18.9
$\sqrt{v_r}$	71.5	27.8	24.9	18.3	26.9	18.3
$\sqrt{v_D}$	115.0	31.1	34.6	19.7	34.7	19.3
$\sqrt{v_J}$	929.7	35.2	54.0	21.3	46.5	20.3
$\sqrt{v_{Jack}}$	517.8	31.6	41.7	20.6	37.2	19.6
$\sqrt{v_{J1}}$	929.3	34.0	62.5	22.0	55.8	21.0
$\sqrt{v_r^\star}$	69.7	24.1	24.3	15.9	25.9	14.8
$\sqrt{v_D^\star}$	112.1	27.0	33.7	17.1	33.5	15.7
$\sqrt{v_J^\star}$	906.6	30.5	52.6	18.5	44.8	16.5
$\sqrt{v_{Jack}^\star}$	504.9	27.4	40.7	17.9	35.9	15.9
$\sqrt{v_{J1}^\star}$	906.2	29.6	60.9	19.1	53.8	17.1
	Simulated Population (numbers in millions)					
	n = 300	n = 1,500	n = 300	n = 1,500	n = 300	n = 1,500
Average $\frac{\hat{t}_y^\pi}{N}$	838.9	838.7	838.1	843.1	838.7	839.1
rmse $\frac{\hat{t}_y^\pi}{N}$	39.9	15.8	48.0	23.7	34.9	15.9
Average $\frac{\hat{t}_y^g}{N}$	838.6	839.1	838.8	840.0	839.4	839.1
rmse $\frac{\hat{t}_y^g}{N}$	12.5	4.8	10.8	4.4	10.3	5.0
$\sqrt{v_g}$	11.6	5.0	10.2	4.7	10.2	4.7
$\sqrt{v_{wr}}$	11.7	5.1	10.2	4.8	10.2	4.8
$\sqrt{v_{JL}}$	11.6	5.1	10.3	4.7	10.3	4.8
$\sqrt{v_r}$	11.7	5.1	10.3	4.7	10.3	4.8
$\sqrt{v_D}$	12.1	5.1	10.5	4.8	10.5	4.8
$\sqrt{v_J}$	12.9	5.2	10.8	4.8	10.8	4.8
$\sqrt{v_{Jack}}$	12.9	5.2	10.8	4.8	10.7	4.8
$\sqrt{v_{J1}}$	12.9	5.2	10.8	4.8	10.8	4.8
$\sqrt{v_r^\star}$	11.6	5.0	10.2	4.6	10.2	4.6
$\sqrt{v_D^\star}$	12.0	5.0	10.4	4.6	10.4	4.7
$\sqrt{v_J^\star}$	12.8	5.0	10.7	4.7	10.7	4.7
$\sqrt{v_{Jack}^\star}$	12.8	5.0	10.7	4.7	10.7	4.7
$\sqrt{v_{J1}^\star}$	12.8	5.0	10.7	4.7	10.7	4.7

Table A.2: Variability of Sandwich Estimators for School Population

$\hat{\theta}$	$\bar{\hat{\theta}}$	se $\hat{\theta}$	rmse $\hat{\theta}$	Distribution of $\hat{\theta}/\sqrt{v_E}$ Min	1st Qu.	Median	Mean	3rd Qu.	Max
srs n = 25									
$\sqrt{v_g}$	12.42	3.54	4.06	0.46	0.71	0.82	0.86	0.96	3.59
$\sqrt{v_{wr}}$	12.49	2.72	3.32	0.48	0.73	0.84	0.87	0.97	1.71
$\sqrt{v_{JL}}$	13.32	3.79	3.94	0.48	0.75	0.88	0.92	1.03	3.75
$\sqrt{v_r}$	13.22	3.88	4.06	0.47	0.74	0.87	0.92	1.02	3.85
$\sqrt{v_D}$	15.52	5.86	5.96	0.53	0.84	1.00	1.08	1.20	6.84
$\sqrt{v_J}$	18.88	11.38	12.22	0.59	0.96	1.16	1.31	1.43	14.47
$\sqrt{v_{Jack}}$	18.19	10.69	11.34	0.57	0.93	1.13	1.26	1.38	13.69
$\sqrt{v_{J1}}$	18.98	11.23	12.12	0.59	0.97	1.17	1.32	1.44	14.48
$\sqrt{v_r^\star}$	11.93	3.51	4.29	0.42	0.67	0.79	0.83	0.92	3.48
$\sqrt{v_D^\star}$	14.01	5.29	5.30	0.48	0.76	0.90	0.97	1.08	6.17
$\sqrt{v_J^\star}$	17.04	10.27	10.60	0.53	0.87	1.05	1.18	1.29	13.06
$\sqrt{v_{Jack}^\star}$	16.42	9.65	9.85	0.52	0.84	1.02	1.14	1.25	12.35
$\sqrt{v_{J1}^\star}$	17.14	10.14	10.49	0.54	0.88	1.06	1.19	1.30	13.07
srs n = 50									
$\sqrt{v_g}$	7.56	1.10	1.21	0.62	0.84	0.92	0.94	1.01	1.64
$\sqrt{v_{wr}}$	8.51	1.25	1.33	0.67	0.95	1.04	1.06	1.15	1.73
$\sqrt{v_{JL}}$	8.69	1.36	1.50	0.68	0.96	1.06	1.08	1.18	1.94
$\sqrt{v_r}$	8.66	1.38	1.50	0.68	0.96	1.06	1.07	1.17	1.95
$\sqrt{v_D}$	9.27	1.57	1.98	0.71	1.01	1.13	1.15	1.26	2.20
$\sqrt{v_J}$	9.97	1.86	2.66	0.75	1.08	1.20	1.24	1.35	2.88
$\sqrt{v_{Jack}}$	9.80	1.81	2.51	0.74	1.06	1.18	1.22	1.33	2.79
$\sqrt{v_{J1}}$	10.01	1.84	2.68	0.75	1.09	1.21	1.24	1.36	2.86
$\sqrt{v_r^\star}$	6.87	1.09	1.61	0.54	0.76	0.84	0.85	0.93	1.55
$\sqrt{v_D^\star}$	7.35	1.25	1.43	0.56	0.80	0.89	0.91	1.00	1.75
$\sqrt{v_J^\star}$	7.91	1.47	1.48	0.59	0.86	0.95	0.98	1.07	2.29
$\sqrt{v_{Jack}^\star}$	7.78	1.43	1.46	0.58	0.84	0.94	0.97	1.06	2.22
$\sqrt{v_{J1}^\star}$	7.94	1.46	1.47	0.60	0.86	0.96	0.99	1.08	2.27
srs epsem n = 25									
$\sqrt{v_g}$	8.96	1.73	2.42	0.41	0.72	0.83	0.84	0.94	1.63
$\sqrt{v_{wr}}$	9.32	1.90	2.32	0.41	0.76	0.86	0.87	0.97	1.60
$\sqrt{v_{JL}}$	9.67	1.96	2.20	0.40	0.78	0.89	0.91	1.02	1.78
$\sqrt{v_r}$	9.48	1.92	2.26	0.39	0.76	0.87	0.89	1.00	1.74
$\sqrt{v_D}$	10.86	2.44	2.44	0.45	0.86	0.99	1.02	1.15	2.00
$\sqrt{v_J}$	12.65	3.32	3.87	0.52	0.97	1.14	1.19	1.33	2.78
$\sqrt{v_{Jack}}$	12.39	3.24	3.67	0.51	0.96	1.11	1.16	1.30	2.71
$\sqrt{v_{J1}}$	12.90	3.38	4.05	0.53	1.00	1.16	1.21	1.36	2.82
$\sqrt{v_r^\star}$	8.55	1.74	2.73	0.36	0.69	0.79	0.80	0.91	1.57
$\sqrt{v_D^\star}$	9.80	2.20	2.36	0.41	0.78	0.90	0.92	1.04	1.80
$\sqrt{v_J^\star}$	11.42	3.00	3.09	0.47	0.88	1.02	1.07	1.20	2.51
$\sqrt{v_{Jack}^\star}$	11.18	2.93	2.97	0.46	0.86	1.00	1.05	1.18	2.45
$\sqrt{v_{J1}^\star}$	11.65	3.05	3.20	0.48	0.90	1.05	1.09	1.23	2.55
srs epsem n = 50									
$\sqrt{v_g}$	6.34	0.83	1.11	0.61	0.82	0.88	0.90	0.96	1.49
$\sqrt{v_{wr}}$	7.22	1.06	1.07	0.65	0.92	1.00	1.02	1.11	1.51
$\sqrt{v_{JL}}$	7.35	1.08	1.11	0.66	0.94	1.03	1.04	1.13	1.79
$\sqrt{v_r}$	7.28	1.06	1.08	0.66	0.93	1.02	1.03	1.12	1.77
$\sqrt{v_D}$	7.80	1.27	1.46	0.69	0.98	1.08	1.10	1.20	2.06
$\sqrt{v_J}$	8.41	1.55	2.05	0.72	1.04	1.16	1.19	1.29	2.43
$\sqrt{v_{Jack}}$	8.32	1.53	1.98	0.72	1.03	1.14	1.18	1.28	2.40
$\sqrt{v_{J1}}$	8.49	1.57	2.11	0.73	1.06	1.17	1.20	1.30	2.45
$\sqrt{v_r^\star}$	5.77	0.84	1.54	0.52	0.74	0.81	0.82	0.89	1.40
$\sqrt{v_D^\star}$	6.19	1.01	1.33	0.55	0.78	0.86	0.88	0.95	1.63
$\sqrt{v_J^\star}$	6.67	1.23	1.29	0.57	0.83	0.92	0.94	1.02	1.93
$\sqrt{v_{Jack}^\star}$	6.60	1.22	1.30	0.57	0.82	0.91	0.93	1.01	1.91
$\sqrt{v_{J1}^\star}$	6.74	1.24	1.28	0.58	0.84	0.93	0.95	1.03	1.95
pps n = 25									
$\sqrt{v_g}$	8.84	1.44	2.62	0.48	0.71	0.79	0.80	0.88	1.33
$\sqrt{v_{wr}}$	9.30	1.40	2.22	0.51	0.76	0.84	0.84	0.92	1.30
$\sqrt{v_{JL}}$	9.57	1.65	2.21	0.50	0.76	0.86	0.87	0.96	1.46
$\sqrt{v_r}$	9.38	1.62	2.32	0.49	0.75	0.84	0.85	0.94	1.43
$\sqrt{v_D}$	10.55	1.95	2.01	0.53	0.83	0.94	0.96	1.06	1.66
$\sqrt{v_J}$	12.00	2.47	2.65	0.59	0.94	1.06	1.09	1.21	2.15
$\sqrt{v_{Jack}}$	11.76	2.42	2.53	0.57	0.92	1.04	1.07	1.18	2.10
$\sqrt{v_{J1}}$	12.24	2.52	2.79	0.60	0.96	1.08	1.11	1.23	2.19
$\sqrt{v_r^\star}$	8.41	1.45	3.00	0.43	0.67	0.76	0.76	0.84	1.30
$\sqrt{v_D^\star}$	9.46	1.74	2.35	0.47	0.75	0.84	0.86	0.95	1.51
$\sqrt{v_J^\star}$	10.76	2.21	2.22	0.52	0.84	0.95	0.98	1.08	1.90
$\sqrt{v_{Jack}^\star}$	10.55	2.16	2.22	0.51	0.82	0.93	0.96	1.06	1.86
$\sqrt{v_{J1}^\star}$	10.98	2.25	2.25	0.53	0.86	0.97	1.00	1.10	1.93
pps n = 50									
$\sqrt{v_g}$	6.10	0.61	0.69	0.72	0.88	0.95	0.95	1.01	1.28
$\sqrt{v_{wr}}$	6.98	0.71	0.90	0.78	1.00	1.09	1.09	1.16	1.47
$\sqrt{v_{JL}}$	7.11	0.83	1.07	0.81	1.01	1.11	1.11	1.19	1.52
$\sqrt{v_r}$	7.04	0.82	1.02	0.80	1.00	1.09	1.09	1.18	1.50
$\sqrt{v_D}$	7.45	0.91	1.37	0.84	1.06	1.15	1.16	1.25	1.64
$\sqrt{v_J}$	7.90	1.02	1.79	0.88	1.11	1.22	1.23	1.33	1.83
$\sqrt{v_{Jack}}$	7.82	1.01	1.72	0.88	1.10	1.21	1.22	1.31	1.81
$\sqrt{v_{J1}}$	7.98	1.03	1.87	0.89	1.13	1.23	1.24	1.34	1.85
$\sqrt{v_r^\star}$	5.49	0.64	1.14	0.62	0.78	0.85	0.85	0.92	1.16
$\sqrt{v_D^\star}$	5.81	0.71	0.95	0.65	0.82	0.90	0.90	0.97	1.28
$\sqrt{v_J^\star}$	6.16	0.80	0.85	0.68	0.87	0.95	0.96	1.03	1.43
$\sqrt{v_{Jack}^\star}$	6.10	0.79	0.86	0.67	0.86	0.94	0.95	1.02	1.42
$\sqrt{v_{J1}^\star}$	6.22	0.81	0.84	0.69	0.88	0.96	0.97	1.04	1.44

Table A.3: Variability of Sandwich Estimators for ACS Population

$\hat{\theta}$	$\bar{\hat{\theta}}$	se $\hat{\theta}$	rmse $\hat{\theta}$	Distribution of $\hat{\theta}/\sqrt{v_E}$					
				Min	1st Qu.	Median	Mean	3rd Qu.	Max
srs n = 3									
$\sqrt{v_g}$	71.12	158.95	161.95	0.05	0.26	0.43	0.70	0.73	90.25
$\sqrt{v_{wr}}$	67.62	92.53	98.78	0.00	0.18	0.38	0.66	0.79	12.78
$\sqrt{v_{JL}}$	64.34	142.76	147.69	0.01	0.23	0.40	0.63	0.69	81.58
$\sqrt{v_r}$	71.46	162.27	165.15	0.01	0.26	0.43	0.70	0.74	92.10
$\sqrt{v_D}$	114.97	234.84	235.16	0.01	0.35	0.62	1.13	1.18	93.12
$\sqrt{v_J}$	929.73	13859.21	13882.51	0.03	0.51	1.03	9.10	2.93	8223.46
$\sqrt{v_{Jack}}$	517.79	7227.47	7238.69	0.01	0.33	0.69	5.07	1.81	4338.69
$\sqrt{v_{J1}}$	929.32	13858.97	13882.25	0.01	0.51	1.05	9.09	2.92	8223.46
$\sqrt{v_r^*}$	69.68	158.23	161.53	0.01	0.25	0.42	0.68	0.72	89.81
$\sqrt{v_D^*}$	112.10	228.99	229.18	0.01	0.34	0.61	1.10	1.15	90.80
$\sqrt{v_J^*}$	906.58	13514.11	13536.68	0.03	0.50	1.00	8.87	2.86	8018.02
$\sqrt{v_{Jack}^*}$	504.90	7047.51	7058.30	0.01	0.33	0.67	4.94	1.76	4231.08
$\sqrt{v_{J1}^*}$	906.18	13513.88	13536.43	0.01	0.50	1.02	8.87	2.85	8018.02
srs n = 15									
$\sqrt{v_g}$	26.58	10.90	11.34	0.27	0.65	0.82	0.89	1.05	5.15
$\sqrt{v_{wr}}$	27.08	13.16	13.42	0.14	0.61	0.83	0.91	1.10	5.49
$\sqrt{v_{JL}}$	26.64	11.52	11.93	0.18	0.63	0.82	0.90	1.07	4.85
$\sqrt{v_r}$	27.76	12.37	12.53	0.18	0.65	0.85	0.93	1.11	5.67
$\sqrt{v_D}$	31.05	15.85	15.90	0.19	0.71	0.93	1.04	1.24	9.61
$\sqrt{v_J}$	35.17	21.64	22.31	0.21	0.77	1.02	1.18	1.39	17.41
$\sqrt{v_{Jack}}$	31.61	18.79	18.88	0.19	0.70	0.92	1.06	1.25	14.64
$\sqrt{v_{J1}}$	34.04	20.69	21.14	0.21	0.75	0.99	1.14	1.35	16.59
$\sqrt{v_r^*}$	24.11	10.75	12.13	0.16	0.57	0.74	0.81	0.97	4.92
$\sqrt{v_D^*}$	26.96	13.76	14.04	0.17	0.62	0.81	0.91	1.08	8.35
$\sqrt{v_J^*}$	30.54	18.79	18.81	0.18	0.67	0.89	1.03	1.21	15.11
$\sqrt{v_{Jack}^*}$	27.45	16.32	16.48	0.17	0.60	0.80	0.92	1.09	12.71
$\sqrt{v_{J1}^*}$	29.56	17.97	17.97	0.18	0.65	0.86	0.99	1.17	14.41
srs epsem n = 3									
$\sqrt{v_g}$	25.69	16.17	25.71	0.07	0.31	0.48	0.56	0.72	4.02
$\sqrt{v_{wr}}$	29.46	23.66	28.68	0.00	0.30	0.50	0.64	0.82	3.99
$\sqrt{v_{JL}}$	30.56	20.86	25.76	0.00	0.34	0.57	0.67	0.88	4.95
$\sqrt{v_r}$	24.95	17.03	26.83	0.00	0.28	0.47	0.55	0.72	4.05
$\sqrt{v_D}$	34.57	24.23	26.65	0.00	0.38	0.64	0.76	0.99	5.86
$\sqrt{v_J}$	53.95	45.57	46.31	0.00	0.56	0.95	1.18	1.52	20.74
$\sqrt{v_{Jack}}$	41.70	34.03	34.26	0.00	0.43	0.74	0.91	1.18	14.10
$\sqrt{v_{J1}}$	62.48	50.84	53.54	0.00	0.65	1.11	1.37	1.76	21.16
$\sqrt{v_r^*}$	24.33	16.61	27.05	0.00	0.27	0.46	0.53	0.70	3.94
$\sqrt{v_D^*}$	33.71	23.62	26.48	0.00	0.37	0.62	0.74	0.96	5.72
$\sqrt{v_J^*}$	52.61	44.44	44.97	0.00	0.54	0.92	1.15	1.48	20.22
$\sqrt{v_{Jack}^*}$	40.67	33.18	33.55	0.00	0.42	0.72	0.89	1.15	13.75
$\sqrt{v_{J1}^*}$	60.93	49.58	51.86	0.00	0.63	1.08	1.33	1.72	20.63
srs epsem n = 15									
$\sqrt{v_g}$	17.63	5.39	5.96	0.39	0.67	0.83	0.87	1.03	2.20
$\sqrt{v_{wr}}$	19.14	7.30	7.37	0.27	0.67	0.88	0.95	1.18	2.60
$\sqrt{v_{JL}}$	18.97	6.66	6.77	0.31	0.69	0.88	0.94	1.13	2.53
$\sqrt{v_r}$	18.32	6.43	6.69	0.30	0.67	0.85	0.91	1.09	2.45
$\sqrt{v_D}$	19.73	7.22	7.23	0.31	0.71	0.91	0.98	1.18	2.76
$\sqrt{v_J}$	21.30	8.16	8.23	0.33	0.76	0.98	1.06	1.28	3.12
$\sqrt{v_{Jack}}$	20.57	7.87	7.88	0.32	0.73	0.95	1.02	1.23	3.01
$\sqrt{v_{J1}}$	22.04	8.44	8.64	0.34	0.78	1.01	1.09	1.32	3.23
$\sqrt{v_r^*}$	15.91	5.59	7.03	0.26	0.58	0.74	0.79	0.95	2.12
$\sqrt{v_D^*}$	17.13	6.27	6.97	0.27	0.62	0.79	0.85	1.03	2.39
$\sqrt{v_J^*}$	18.50	7.08	7.28	0.29	0.66	0.85	0.92	1.11	2.71
$\sqrt{v_{Jack}^*}$	17.86	6.84	7.22	0.28	0.63	0.82	0.89	1.07	2.62
$\sqrt{v_{J1}^*}$	19.14	7.32	7.40	0.30	0.68	0.88	0.95	1.15	2.80
pps epsem n = 3									
$\sqrt{v_g}$	27.68	15.25	21.75	0.08	0.38	0.57	0.64	0.82	2.61
$\sqrt{v_{wr}}$	33.55	22.56	24.53	0.01	0.41	0.66	0.78	1.01	4.27
$\sqrt{v_{JL}}$	32.90	20.09	22.57	0.01	0.42	0.68	0.76	1.01	3.31
$\sqrt{v_r}$	26.86	16.41	23.14	0.01	0.34	0.55	0.62	0.82	2.70
$\sqrt{v_D}$	34.74	21.64	23.22	0.01	0.44	0.71	0.80	1.06	3.49
$\sqrt{v_J}$	46.49	30.06	30.24	0.02	0.58	0.92	1.08	1.41	5.45
$\sqrt{v_{Jack}}$	37.18	24.06	24.80	0.01	0.46	0.74	0.86	1.13	3.98
$\sqrt{v_{J1}}$	55.76	36.09	38.22	0.02	0.69	1.11	1.29	1.70	5.97
$\sqrt{v_r^*}$	25.91	15.79	23.40	0.01	0.33	0.53	0.60	0.79	2.62
$\sqrt{v_D^*}$	33.51	20.83	22.97	0.01	0.42	0.68	0.78	1.02	3.38
$\sqrt{v_J^*}$	44.84	28.97	29.02	0.02	0.56	0.89	1.04	1.35	5.33
$\sqrt{v_{Jack}^*}$	35.86	23.18	24.31	0.01	0.44	0.72	0.83	1.09	3.81
$\sqrt{v_{J1}^*}$	53.78	34.77	36.35	0.02	0.66	1.08	1.25	1.64	5.71
pps n = 15									
$\sqrt{v_g}$	17.47	4.41	4.59	0.37	0.75	0.89	0.93	1.08	1.93
$\sqrt{v_{wr}}$	19.07	6.06	6.07	0.30	0.77	0.96	1.02	1.23	2.33
$\sqrt{v_{JL}}$	18.93	5.70	5.70	0.32	0.78	0.96	1.01	1.21	2.28
$\sqrt{v_r}$	18.29	5.51	5.53	0.30	0.75	0.93	0.98	1.17	2.20
$\sqrt{v_D}$	19.27	5.97	5.99	0.31	0.79	0.98	1.03	1.23	2.37
$\sqrt{v_J}$	20.33	6.49	6.68	0.33	0.82	1.02	1.08	1.30	2.58
$\sqrt{v_{Jack}}$	19.64	6.27	6.33	0.31	0.80	0.99	1.05	1.25	2.50
$\sqrt{v_{J1}}$	21.04	6.72	7.10	0.34	0.85	1.06	1.12	1.34	2.67
$\sqrt{v_r^*}$	14.85	4.50	5.95	0.25	0.61	0.76	0.79	0.94	1.90
$\sqrt{v_D^*}$	15.65	4.88	5.78	0.25	0.64	0.79	0.83	1.00	2.04
$\sqrt{v_J^*}$	16.51	5.31	5.76	0.26	0.67	0.83	0.88	1.05	2.20
$\sqrt{v_{Jack}^*}$	15.95	5.12	5.84	0.25	0.65	0.80	0.85	1.02	2.12
$\sqrt{v_{J1}^*}$	17.08	5.49	5.74	0.27	0.69	0.86	0.91	1.09	2.28

Table A.4: Variability of Sandwich Estimators for Simulated Population

$\hat{\theta}$	$\overline{\hat{\theta}}$	se $\hat{\theta}$	rmse $\hat{\theta}$	Distribution of $\hat{\theta}/\sqrt{v_E}$					
				Min	1st Qu.	Median	Mean	3rd Qu.	Max
srs n = 300									
$\sqrt{v_g}$	11.63	2.59	2.73	0.62	0.81	0.88	0.93	1.00	3.24
$\sqrt{v_{wr}}$	11.72	3.11	3.21	0.56	0.79	0.88	0.94	1.01	4.51
$\sqrt{v_{JL}}$	11.64	2.58	2.72	0.62	0.81	0.88	0.93	1.00	3.23
$\sqrt{v_r}$	11.65	2.59	2.73	0.62	0.81	0.88	0.93	1.01	3.25
$\sqrt{v_D}$	12.10	3.78	3.79	0.63	0.82	0.90	0.97	1.04	4.88
$\sqrt{v_J}$	12.92	8.20	8.21	0.63	0.83	0.92	1.03	1.07	13.07
$\sqrt{v_{Jack}}$	12.86	8.17	8.17	0.63	0.83	0.92	1.03	1.06	13.03
$\sqrt{v_{J1}}$	12.91	8.20	8.21	0.63	0.83	0.92	1.03	1.07	13.07
$\sqrt{v_r^\star}$	11.59	2.58	2.73	0.62	0.80	0.88	0.93	1.00	3.23
$\sqrt{v_D^\star}$	12.04	3.76	3.78	0.63	0.81	0.90	0.96	1.03	4.85
$\sqrt{v_J^\star}$	12.85	8.16	8.17	0.63	0.83	0.92	1.03	1.06	13.01
$\sqrt{v_{Jack}^\star}$	12.80	8.12	8.13	0.63	0.82	0.91	1.02	1.06	12.97
$\sqrt{v_{J1}^\star}$	12.84	8.16	8.16	0.63	0.83	0.92	1.03	1.06	13.01
srs n = 1,500									
$\sqrt{v_g}$	5.05	0.46	0.51	0.87	0.98	1.02	1.05	1.10	1.40
$\sqrt{v_{wr}}$	5.08	0.48	0.54	0.88	0.99	1.03	1.05	1.10	1.39
$\sqrt{v_{JL}}$	5.08	0.46	0.53	0.87	0.98	1.03	1.05	1.10	1.41
$\sqrt{v_r}$	5.08	0.46	0.53	0.87	0.98	1.03	1.05	1.10	1.41
$\sqrt{v_D}$	5.13	0.49	0.57	0.88	0.99	1.04	1.06	1.11	1.43
$\sqrt{v_J}$	5.17	0.52	0.62	0.89	0.99	1.05	1.07	1.13	1.46
$\sqrt{v_{Jack}}$	5.17	0.52	0.62	0.89	0.99	1.05	1.07	1.13	1.46
$\sqrt{v_{J1}}$	5.17	0.52	0.62	0.89	0.99	1.05	1.07	1.13	1.46
$\sqrt{v_r^\star}$	4.95	0.45	0.47	0.85	0.96	1.01	1.03	1.08	1.37
$\sqrt{v_D^\star}$	5.00	0.48	0.50	0.86	0.96	1.01	1.03	1.08	1.40
$\sqrt{v_J^\star}$	5.04	0.51	0.55	0.87	0.97	1.02	1.04	1.10	1.42
$\sqrt{v_{Jack}^\star}$	5.04	0.51	0.55	0.87	0.97	1.02	1.04	1.10	1.42
$\sqrt{v_{J1}^\star}$	5.04	0.51	0.55	0.87	0.97	1.02	1.04	1.10	1.42
srs epsem n = 300									
$\sqrt{v_g}$	10.24	1.51	1.62	0.68	0.85	0.92	0.95	1.01	1.80
$\sqrt{v_{wr}}$	10.20	1.55	1.67	0.67	0.84	0.92	0.94	1.02	1.81
$\sqrt{v_{JL}}$	10.27	1.52	1.61	0.69	0.86	0.92	0.95	1.01	1.80
$\sqrt{v_r}$	10.26	1.52	1.62	0.68	0.85	0.92	0.95	1.01	1.80
$\sqrt{v_D}$	10.49	1.75	1.78	0.69	0.87	0.93	0.97	1.03	2.10
$\sqrt{v_J}$	10.77	2.12	2.12	0.69	0.88	0.95	0.99	1.06	2.57
$\sqrt{v_{Jack}}$	10.75	2.11	2.11	0.69	0.88	0.95	0.99	1.05	2.56
$\sqrt{v_{J1}}$	10.79	2.12	2.12	0.69	0.88	0.95	1.00	1.06	2.57
$\sqrt{v_r^\star}$	10.21	1.51	1.63	0.68	0.85	0.92	0.94	1.00	1.79
$\sqrt{v_D^\star}$	10.44	1.74	1.78	0.68	0.86	0.93	0.96	1.03	2.09
$\sqrt{v_J^\star}$	10.72	2.11	2.11	0.69	0.87	0.95	0.99	1.05	2.55
$\sqrt{v_{Jack}^\star}$	10.70	2.10	2.10	0.69	0.87	0.95	0.99	1.05	2.55
$\sqrt{v_{J1}^\star}$	10.74	2.11	2.11	0.69	0.88	0.95	0.99	1.05	2.56
srs epsem n = 1,500									
$\sqrt{v_g}$	4.69	0.38	0.50	0.92	1.02	1.06	1.07	1.10	1.42
$\sqrt{v_{wr}}$	4.77	0.44	0.59	0.93	1.03	1.07	1.09	1.16	1.45
$\sqrt{v_{JL}}$	4.73	0.38	0.52	0.93	1.03	1.07	1.08	1.11	1.43
$\sqrt{v_r}$	4.73	0.38	0.52	0.93	1.03	1.07	1.08	1.11	1.43
$\sqrt{v_D}$	4.76	0.41	0.57	0.93	1.03	1.07	1.09	1.12	1.49
$\sqrt{v_J}$	4.80	0.45	0.62	0.93	1.04	1.08	1.10	1.13	1.56
$\sqrt{v_{Jack}}$	4.80	0.45	0.62	0.93	1.04	1.08	1.10	1.13	1.56
$\sqrt{v_{J1}}$	4.80	0.45	0.62	0.93	1.04	1.08	1.10	1.13	1.56
$\sqrt{v_r^\star}$	4.61	0.38	0.44	0.91	1.00	1.04	1.05	1.09	1.40
$\sqrt{v_D^\star}$	4.64	0.40	0.48	0.91	1.01	1.05	1.06	1.09	1.45
$\sqrt{v_J^\star}$	4.68	0.44	0.53	0.91	1.01	1.05	1.07	1.10	1.52
$\sqrt{v_{Jack}^\star}$	4.68	0.44	0.53	0.91	1.01	1.05	1.07	1.10	1.52
$\sqrt{v_{J1}^\star}$	4.68	0.44	0.53	0.91	1.01	1.05	1.07	1.10	1.52
pps n = 300									
$\sqrt{v_g}$	10.25	1.44	1.44	0.70	0.90	0.98	1.00	1.06	1.82
$\sqrt{v_{wr}}$	10.23	1.55	1.55	0.65	0.89	0.98	1.00	1.07	1.71
$\sqrt{v_{JL}}$	10.28	1.45	1.45	0.70	0.90	0.98	1.00	1.07	1.82
$\sqrt{v_r}$	10.27	1.44	1.44	0.70	0.90	0.98	1.00	1.07	1.82
$\sqrt{v_D}$	10.49	1.61	1.63	0.71	0.91	1.00	1.02	1.09	1.89
$\sqrt{v_J}$	10.76	1.87	1.93	0.71	0.93	1.01	1.05	1.11	2.04
$\sqrt{v_{Jack}}$	10.74	1.87	1.92	0.71	0.92	1.01	1.04	1.11	2.04
$\sqrt{v_{J1}}$	10.77	1.87	1.94	0.71	0.93	1.01	1.05	1.12	2.04
$\sqrt{v_r^\star}$	10.21	1.44	1.44	0.70	0.89	0.97	0.99	1.06	1.81
$\sqrt{v_D^\star}$	10.43	1.60	1.61	0.70	0.91	0.99	1.01	1.08	1.88
$\sqrt{v_J^\star}$	10.70	1.86	1.90	0.71	0.92	1.01	1.04	1.11	2.03
$\sqrt{v_{Jack}^\star}$	10.68	1.86	1.90	0.71	0.92	1.01	1.04	1.11	2.03
$\sqrt{v_{J1}^\star}$	10.71	1.86	1.91	0.71	0.92	1.01	1.04	1.11	2.03
pps n = 1,500									
$\sqrt{v_g}$	4.72	0.37	0.50	0.79	0.89	0.92	0.93	0.98	1.24
$\sqrt{v_{wr}}$	4.75	0.34	0.45	0.75	0.89	0.93	0.94	0.98	1.12
$\sqrt{v_{JL}}$	4.75	0.38	0.48	0.80	0.89	0.93	0.94	0.99	1.25
$\sqrt{v_r}$	4.75	0.38	0.48	0.80	0.89	0.93	0.94	0.98	1.25
$\sqrt{v_D}$	4.79	0.41	0.48	0.80	0.90	0.93	0.95	0.99	1.31
$\sqrt{v_J}$	4.82	0.44	0.50	0.80	0.90	0.94	0.96	1.00	1.37
$\sqrt{v_{Jack}}$	4.82	0.44	0.50	0.80	0.90	0.94	0.95	1.00	1.37
$\sqrt{v_{J1}}$	4.82	0.44	0.50	0.80	0.90	0.94	0.96	1.00	1.38
$\sqrt{v_r^\star}$	4.62	0.36	0.56	0.77	0.87	0.90	0.92	0.96	1.22
$\sqrt{v_D^\star}$	4.65	0.39	0.56	0.78	0.87	0.90	0.92	0.96	1.27
$\sqrt{v_J^\star}$	4.69	0.43	0.56	0.78	0.88	0.91	0.93	0.97	1.34
$\sqrt{v_{Jack}^\star}$	4.69	0.43	0.56	0.78	0.88	0.91	0.93	0.97	1.34
$\sqrt{v_{J1}^\star}$	4.69	0.43	0.56	0.78	0.88	0.91	0.93	0.97	1.34

Table A.5: Confidence Interval Coverage of Variance Estimators

Finite CI	Finite			ACS			Simulation		
	Lower	Middle	Upper	Lower	Middle	Upper	Lower	Middle	Upper
	srs n = 25			srs n = 3			srs n = 300		
$\sqrt{v_E}$	3.9	94.4	1.7	3.9	95.3	0.8	2.7	95.0	2.3
$\sqrt{v_g}$	9.0	89.0	2.0	17.9	78.1	4.1	4.4	93.4	2.2
$\sqrt{v_{wr}}$	7.8	89.5	2.7	23.5	69.5	6.9	3.9	92.8	3.3
$\sqrt{v_{JL}}$	7.1	91.1	1.8	22.0	72.1	5.8	4.4	93.4	2.2
$\sqrt{v_r}$	7.3	90.9	1.8	18.3	77.2	4.5	4.4	93.4	2.2
$\sqrt{v_D}$	4.5	94.5	1.0	10.8	87.0	2.2	3.7	94.2	2.1
$\sqrt{v_J}$	2.5	97.2	0.3	4.9	94.1	1.0	3.6	94.4	2.0
$\sqrt{v_{Jack}}$	2.6	97.0	0.4	11.8	85.3	3.0	3.6	94.4	2.0
$\sqrt{v_{J1}}$	2.3	97.4	0.3	6.3	92.1	1.6	3.6	94.4	2.0
$\sqrt{v_r^*}$	9.8	87.9	2.3	18.9	76.4	4.8	4.4	93.4	2.2
$\sqrt{v_D^*}$	6.7	91.8	1.5	11.4	86.3	2.3	3.8	94.1	2.1
$\sqrt{v_J^*}$	4.0	95.3	0.7	5.2	93.7	1.0	3.6	94.4	2.0
$\sqrt{v_{Jack}^*}$	4.7	94.6	0.7	12.1	84.9	3.0	3.7	94.2	2.1
$\sqrt{v_{J1}^*}$	3.9	95.4	0.7	6.5	91.8	1.6	3.6	94.3	2.1
	srs n = 50			srs n = 15			srs n = 1,500		
$\sqrt{v_E}$	3.7	94.7	1.6	4.3	94.3	1.4	1.0	96.0	3.0
$\sqrt{v_g}$	6.2	92.4	1.4	8.7	89.8	1.6	1.0	95.0	4.0
$\sqrt{v_{wr}}$	4.5	94.5	1.0	9.3	88.5	2.2	1.0	96.0	3.0
$\sqrt{v_{JL}}$	3.8	95.6	0.6	9.0	89.0	2.0	1.0	95.0	4.0
$\sqrt{v_r}$	4.0	95.4	0.6	8.2	90.3	1.6	1.0	95.0	4.0
$\sqrt{v_D}$	3.1	96.4	0.5	6.4	92.6	1.0	1.0	95.0	4.0
$\sqrt{v_J}$	2.2	97.5	0.3	5.2	94.3	0.5	1.0	95.0	4.0
$\sqrt{v_{Jack}}$	2.3	97.4	0.3	6.8	92.0	1.2	1.0	95.0	4.0
$\sqrt{v_{J1}}$	2.1	97.6	0.3	5.8	93.4	0.8	1.0	95.0	4.0
$\sqrt{v_r^*}$	8.2	89.0	2.8	11.4	85.9	2.8	1.0	95.0	4.0
$\sqrt{v_D^*}$	7.4	90.9	1.7	9.4	88.6	2.0	1.0	95.0	4.0
$\sqrt{v_J^*}$	5.9	93.1	1.0	7.3	91.3	1.4	1.0	95.0	4.0
$\sqrt{v_{Jack}^*}$	5.9	93.0	1.1	9.4	88.5	2.1	1.0	95.0	4.0
$\sqrt{v_{J1}^*}$	5.8	93.1	1.1	8.0	90.4	1.6	1.0	95.0	4.0
	srs epsem n = 25			srs epsem n = 3			srs epsem n = 300		
$\sqrt{v_E}$	2.6	95.1	2.3	1.8	95.1	3.1	2.4	94.7	2.9
$\sqrt{v_g}$	6.6	89.4	4.0	22.3	67.6	10.1	2.7	93.9	3.4
$\sqrt{v_{wr}}$	6.7	89.8	3.5	24.1	66.8	9.1	3.1	93.3	3.6
$\sqrt{v_{JL}}$	5.3	91.8	2.9	19.1	72.5	8.3	2.6	94.1	3.3
$\sqrt{v_r}$	5.7	91.2	3.1	24.1	64.8	11.1	2.7	93.9	3.4
$\sqrt{v_D}$	4.9	92.9	2.2	17.4	75.9	6.7	2.5	94.3	3.2
$\sqrt{v_J}$	2.6	96.0	1.4	10.1	86.8	3.0	2.3	94.9	2.8
$\sqrt{v_{Jack}}$	2.9	95.6	1.5	14.5	80.6	4.9	2.3	94.9	2.8
$\sqrt{v_{J1}}$	2.2	96.7	1.1	8.4	89.2	2.4	2.3	94.9	2.8
$\sqrt{v_r^*}$	8.0	87.1	4.9	24.7	63.9	11.4	2.7	93.8	3.5
$\sqrt{v_D^*}$	5.6	91.7	2.7	17.9	75.1	7.0	2.5	94.3	3.2
$\sqrt{v_J^*}$	4.7	93.4	1.9	10.6	86.2	3.3	2.3	94.8	2.9
$\sqrt{v_{Jack}^*}$	4.9	93.2	1.9	15.1	79.7	5.2	2.3	94.7	3.0
$\sqrt{v_{J1}^*}$	3.7	94.5	1.8	8.6	88.8	2.6	2.3	94.9	2.8
	srs epsem n = 50			srs epsem n = 15			srs epsem n = 1,500		
$\sqrt{v_E}$	2.5	95.1	2.4	1.8	95.4	2.7	3.0	94.0	3.0
$\sqrt{v_g}$	5.5	91.6	2.9	10.1	87.8	2.1	3.0	96.0	1.0
$\sqrt{v_{wr}}$	3.8	94.7	1.5	10.7	87.9	1.4	3.0	95.0	2.0
$\sqrt{v_{JL}}$	3.0	95.6	1.4	9.8	88.8	1.3	3.0	96.0	1.0
$\sqrt{v_r}$	3.1	95.5	1.4	10.6	87.8	1.6	3.0	96.0	1.0
$\sqrt{v_D}$	2.5	96.7	0.8	9.3	89.6	1.1	3.0	96.0	1.0
$\sqrt{v_J}$	2.1	97.7	0.2	8.2	91.0	0.8	3.0	96.0	1.0
$\sqrt{v_{Jack}}$	2.2	97.5	0.3	8.8	90.3	0.9	3.0	96.0	1.0
$\sqrt{v_{J1}}$	2.1	97.7	0.2	7.6	91.8	0.6	3.0	96.0	1.0
$\sqrt{v_r^*}$	8.3	87.3	4.4	13.6	83.2	3.2	3.0	96.0	1.0
$\sqrt{v_D^*}$	6.7	90.7	2.6	12.3	85.4	2.3	3.0	96.0	1.0
$\sqrt{v_J^*}$	5.0	92.9	2.1	10.8	87.7	1.5	3.0	96.0	1.0
$\sqrt{v_{Jack}^*}$	5.3	92.5	2.2	11.7	86.4	1.8	3.0	96.0	1.0
$\sqrt{v_{J1}^*}$	4.9	93.1	2.0	10.3	88.5	1.2	3.0	96.0	1.0
	pps n = 25			pps n = 3			pps n = 300		
$\sqrt{v_E}$	2.1	95.0	2.9	1.4	94.8	3.8	2.9	94.2	2.9
$\sqrt{v_g}$	7.1	88.3	4.6	19.6	73.0	7.4	2.9	93.9	3.2
$\sqrt{v_{wr}}$	6.3	89.3	4.4	17.6	76.6	5.9	3.1	93.6	3.3
$\sqrt{v_{JL}}$	5.7	90.6	3.7	17.0	76.8	6.2	2.9	93.9	3.2
$\sqrt{v_r}$	6.2	89.9	3.9	21.4	70.0	8.6	2.9	93.9	3.2
$\sqrt{v_D}$	4.7	92.7	2.6	16.4	78.3	5.4	2.7	94.7	2.6
$\sqrt{v_J}$	3.0	95.5	1.5	11.1	86.3	2.7	2.6	95.0	2.4
$\sqrt{v_{Jack}}$	3.2	95.1	1.7	15.7	79.6	4.7	2.6	95.0	2.4
$\sqrt{v_{J1}}$	2.9	95.6	1.5	8.4	89.5	2.1	2.6	95.0	2.4
$\sqrt{v_r^*}$	8.6	85.8	5.6	22.2	68.6	9.3	2.9	93.9	3.2
$\sqrt{v_D^*}$	6.1	90.3	3.6	17.0	77.4	5.7	2.7	94.4	2.9
$\sqrt{v_J^*}$	4.7	93.0	2.3	11.4	85.6	3.0	2.6	95.0	2.4
$\sqrt{v_{Jack}^*}$	4.9	92.7	2.4	16.3	78.6	5.0	2.6	95.0	2.4
$\sqrt{v_{J1}^*}$	4.3	93.4	2.3	8.8	88.9	2.3	2.6	95.0	2.4
	pps n = 50			pps n = 9			pps n = 1,500		
$\sqrt{v_E}$	2.6	94.7	2.7	2.2	95.2	2.6	2.0	95.0	3.0
$\sqrt{v_g}$	3.5	93.3	3.2	7.9	90.9	1.2	2.0	92.0	6.0
$\sqrt{v_{wr}}$	2.3	96.4	1.3	7.9	91.3	0.8	3.0	92.0	5.0
$\sqrt{v_{JL}}$	2.1	96.6	1.3	7.6	91.6	0.8	2.0	92.0	6.0
$\sqrt{v_r}$	2.1	96.6	1.3	8.3	90.7	1.0	2.0	92.0	6.0
$\sqrt{v_D}$	1.9	97.0	1.1	7.4	91.9	0.7	2.0	92.0	6.0
$\sqrt{v_J}$	1.6	97.7	0.7	6.8	92.7	0.5	2.0	92.0	6.0
$\sqrt{v_{Jack}}$	1.6	97.7	0.7	7.3	92.0	0.7	2.0	92.0	6.0
$\sqrt{v_{J1}}$	1.6	97.8	0.6	6.3	93.3	0.4	2.0	92.0	6.0
$\sqrt{v_r^*}$	5.5	89.8	4.7	12.7	84.5	2.7	2.0	92.0	6.0
$\sqrt{v_D^*}$	4.5	91.6	3.9	11.9	85.9	2.3	2.0	92.0	6.0
$\sqrt{v_J^*}$	3.4	93.6	3.0	10.8	87.6	1.6	2.0	92.0	6.0
$\sqrt{v_{Jack}^*}$	3.8	93.0	3.2	11.6	86.3	2.1	2.0	92.0	6.0
$\sqrt{v_{J1}^*}$	3.2	93.9	2.9	9.8	88.9	1.4	2.0	92.0	6.0

A.7 R code

```
# I had to alter the UPsystematic function so that it would work.
# I changed trunc(n) to round(n)
UPsystematic.round <- function (pik, eps = 1e-06)
{
    if (any(is.na(pik)))
        stop("there are missing values in the pik vector")
    n = sum(pik)
    if (abs(n - round(n)) < 1e-03)
        n = round(n)
    else stop("the sum of pik is not integer")
    list = pik > eps & pik < 1 - eps
    pik1 = pik[list]
    N = length(pik1)
    a = (c(0, cumsum(pik1)) - runif(1, 0, 1))%%1
    s1 = as.integer(a[1:N] > a[2:(N + 1)])
    s = pik
    s[list] = s1
    s
}

UPrandomsystematic.alt <- function (pik, eps = 1e-06)
{
    if (any(is.na(pik)))
        stop("there are missing values in the pik vector")
    N = length(pik)
    v = sample(N, N)
    s = numeric(N)
    s[v] = UPsystematic.round(pik[v], eps)
    s
}

UPrandomsystematic.alt2 <- function (x, eps = 1e-06)
{
  X.I.ii <- UPrandomsystematic.alt(x$pi.II.all)
  subset(x, X.I.ii == 1)
}

UPoi <- function (x)
{
  X.I.ii <- UPpoisson(x$pi.II.all)
  sa.mp <- subset(x, X.I.ii == 1)
  if(nrow(sa.mp) > 0) return(subset(x, X.I.ii == 1))
}

make.cv <- function(data) {
  empirical.variance <- var(data[,  "total.greg"])
  cv.sandwich <- 100 * (data[, "sandwich"] - empirical.variance ) / empirical.variance
  cv.wr <- 100 * (data[, "v.wr"] - empirical.variance ) / empirical.variance
  print(cbind(cv.sandwich, cv.wr))
}


greg.sim <- function(X.Pop, Y.Pop, clus.id,  Q, a, b, iterations, seed, smp, smp2)
{
  cat("Begin Intro", format(Sys.time(), "%X"), "\n")

  load(file = "C:\\Documents and Settings\\Tim\\My Documents\\Data\\seed.Rdata")
  set.seed(seed)

  Pop.1 <- cbind(X.Pop, Y.Pop, clus.id)

  # Create the measures of size
  mos.1 <- as.vector(by(Pop.1, Pop.1[,"clus.id"], nrow))

  # M.clus is the total number of clusters in the population
  M.clus <- length(unique(Pop.1[, "clus.id"]))

  # Create the first stage sampling probabilities
  pi.I.pps <- a * mos.1  / nrow(Pop.1)
  pi.I.srs <- rep(a / M.clus, M.clus)
  if(smp == "srs") pi.I <- pi.I.srs else pi.I <- pi.I.pps

  pi.II.fixed <- b / mos.1
  pi.II.rate <- (b * sum(M.clus)) / sum(mos.1)
  if(smp2 == "fixed") pi.II.all <- pi.II.fixed else pi.II.all <- pi.II.rate

  pi.k.all <- pi.I * pi.II.all

  # Get the number of columns in X and Y
  X.dim <- ncol(X.Pop)

  # Recode the clusterid
  c.id <- c(1: M.clus)
  clus.conversion <- cbind(unique(Pop.1[, "clus.id"]), c.id, pi.I, pi.II.all, pi.k.all)
  X.clusid <- merge(x = Pop.1, y = clus.conversion, by.x = "clus.id", by.y = 1)

  w.n <- 1 / X.clusid[, "pi.k.all"]
  w.n.II <- 1 / X.clusid[, "pi.II.all"]
  ind <- X.clusid[, "clus.id"]
```

```
  # Create a list of cluster auxilliaries
  X.clus <- split(X.clusid, clus.id)

  t.y.pi <- matrix(0, nrow = iterations, ncol = 1)
  t.y.greg <- matrix(0, nrow = iterations, ncol = 1)

  v.ssw <- matrix(0, nrow = iterations, ncol = 1)
  v.wr <- matrix(0, nrow = iterations, ncol = 1)
  v.JL  <- matrix(0, nrow = iterations, ncol = 1)
  v.r <- matrix(0, nrow = iterations, ncol = 1)
  v.D  <- matrix(0, nrow = iterations, ncol = 1)
  v.J  <- matrix(0, nrow = iterations, ncol = 1)
  v.Jack <- matrix(0, nrow = iterations, ncol = 1)
  v.J1 <- matrix(0, nrow = iterations, ncol = 1)
  v.r.star <- matrix(0, nrow = iterations, ncol = 1)
  v.D.star  <- matrix(0, nrow = iterations, ncol = 1)
  v.J.star  <- matrix(0, nrow = iterations, ncol = 1)
  v.Jack.star <- matrix(0, nrow = iterations, ncol = 1)
  v.J1.star <- matrix(0, nrow = iterations, ncol = 1)

  v.D.error <- matrix(0, nrow = iterations, ncol = 1)

  for(j in 1: iterations)
  {
    ## Sampling begins here
    # Select the first stage sample without replacement
    samp.clus <- UPrandomsystematic.alt(clus.conversion[,"pi.I"])
    X.clus.sample <- X.clus[c.id[samp.clus >= 1]]

    # Select the second stage sample
    if(smp2 == "rate") X.sample.f <- lapply(X.clus.sample, UPoi) else X.sample.f <- lapply(X.clus.sample, UPrandomsystematic.alt2)

    a.f <- sapply(X.sample.f, length)
    b.f <- names(a.f[a.f>1])
    X.sample <- X.sample.f[c(b.f)]

    ## Estimation begins here
    # Population Totals
    T.x <- colSums(X.Pop)

    # Create Unclustered data
    # Note that the sample elements can be repeated
    # Note: There may be some duplicates
    if(smp2 == "fixed") sample.id <- c(sapply(X = X.sample, FUN = rownames, simplify = T, USE.NAMES = T))
        else sample.id <- unique(as.vector(do.call(cbind, (sapply(X = X.sample, FUN = rownames, simplify = F, USE.NAMES = T)))))

    # Sample X and Y values
    # Note: There may be some duplicates when the first stage is selected with replacement
    X.samp <- X.Pop[as.numeric(sample.id),]
    Y.samp <- Y.Pop[as.numeric(sample.id),]
    w.k <- w.n[as.numeric(sample.id)]
    w.k.2 <- w.n.II[as.numeric(sample.id)]

#    ind.1 <- as.matrix(X.clusid[as.numeric(sample.id), "c.id"])
     ind.1 <- factor(ind[as.numeric(sample.id)])

    samp.pi.I <- subset(pi.I, samp.clus == 1)
    samp.pi.I.list <- split(samp.pi.I, f = seq(1:length(samp.pi.I)))

    w.k.clus <- split(w.k, ind.1)

    # Estiamted Covariate Totals
    T.hat.pi.x <- colSums(X.samp * w.k)

    # The Estimated A Inverse matrix
    A.pi.s.inv <- try(solve((t(X.samp) * (w.k)) %*% X.samp))

    # The Estimated B matrix
    B.hat <- A.pi.s.inv %*% t(X.samp) %*% ((w.k) * Y.samp)
    beta.hat <- t(t(coefficients(lm(Y.samp ~ X.samp -1, weights = w.k))))

    # Calculate the g weights
    g.k <- t(1 + (t(T.x - T.hat.pi.x) %*% A.pi.s.inv %*% (t(X.samp))))
    g.i <- split(g.k, ind.1)

    sandwich.clus <- NULL
    t.Ei <- NULL

    # Hat Matrix
    HAT <- (X.samp %*% A.pi.s.inv %*% t(X.samp) * w.k)
    HAT.ii <- lapply(1:length(g.i), function(i, x)
              as.matrix(x[[i]][,2:(X.dim + 1)])
              %*% A.pi.s.inv %*%
              t(as.matrix(x[[i]][,2:(X.dim + 1)])) * (1/ as.vector(x[[i]][,"pi.k.all"])), x = X.sample)

    ## Sandwich Estimator
    ## For ACS 100, goal is 3.340642e+12
    # Notice beta.hat rather than B.hat
    e.k <- Y.samp - X.samp %*% B.hat
    e.i <- split(e.k, ind.1)
    t.E.i <- by(g.k * w.k * e.k, INDICES = ind.1, sum, simplify = T)
```

```
    t.E.i.L <- by(w.k * e.k, INDICES = ind.1, sum, simplify = T)
    t.y.sand <- sum(t.E.i^2)

    v.ssw.i <- lapply(1:length(t.E.i),
                  function(i, X.sample, g.i, e.i)
                  sum(((1 - X.sample[[i]][,"pi.II.all"] ) / X.sample[[i]][,"pi.II.all"]^2) * g.i[[i]]^2 * e.i[[i]]^2),
                  X.sample = X.sample, g.i = g.i, e.i = e.i)

    t.hat.e.i <- lapply(1:length(t.E.i),
                  function(i, g.i, e.i, X.sample)
                  sum((g.i[[i]] * e.i[[i]]) / (X.sample[[i]][,"pi.II.all"])),
                  X.sample = X.sample, g.i = g.i, e.i = e.i)


    v.ssw.clus <- lapply(1:length(t.E.i),
                  function(i, v.ssw.i, t.hat.e.i, samp.pi.I.list)
                  ((1 - samp.pi.I.list[[i]]) / samp.pi.I.list[[i]]^2) * t.hat.e.i[[i]]^2 + (1 / samp.pi.I.list[[i]]) * v.ssw.i[[i]],
                  v.ssw.i = v.ssw.i, t.hat.e.i = t.hat.e.i, samp.pi.I.list = samp.pi.I.list)

    v.D.i.1 <- lapply(1:length(t.E.i),
                  function(i, g.i, w.k.clus, HAT.ii, e.i)
                  t(as.matrix(g.i[[i]]) * w.k.clus[[i]]) %*%
                  ginv(diag(nrow(HAT.ii[[i]])) - as.matrix(HAT.ii[[i]])) %*%
                  as.matrix(e.i[[i]]) %*% t(as.matrix(e.i[[i]])) %*%
                  (as.matrix(g.i[[i]] * w.k.clus[[i]])),
                  g.i = g.i, w.k.clus = w.k.clus, HAT.ii = HAT.ii, e.i = e.i)


    v.D.I.i <- lapply(1:length(t.E.i),
                  function(i, g.i, w.k.clus, HAT.ii, e.i)
                  t(as.matrix(g.i[[i]]) * w.k.clus[[i]]) %*%
                  as.matrix(e.i[[i]]) %*% t(as.matrix(e.i[[i]])) %*%
                  (as.matrix(g.i[[i]] * w.k.clus[[i]])),
                  g.i = g.i, w.k.clus = w.k.clus, HAT.ii = HAT.ii, e.i = e.i)


    v.D.i <- lapply(1:length(t.E.i),
                  function(i, v.D.i, v.D.I.i)
                  ifelse(v.D.i[[i]] <=0, v.D.I.i[[i]], v.D.i[[i]]),
                  v.D.i = v.D.i.1, v.D.I.i = v.D.I.i)

    v.D.i.err <- lapply(1:length(t.E.i),
                  function(i, v.D.i)
                  ifelse(v.D.i[[i]] <=0, 1, 0),
                  v.D.i = v.D.i.1)


    v.J.i <- lapply(1:length(t.E.i),
                  function(i, g.i, w.k.clus, HAT.ii, e.i)
                  t(as.matrix(g.i[[i]]) * w.k.clus[[i]]) %*%
                  ginv(diag(nrow(HAT.ii[[i]])) - as.matrix(HAT.ii[[i]])) %*%
                  as.matrix(e.i[[i]]) %*% t(as.matrix(e.i[[i]])) %*%
                  t(ginv(diag(nrow(HAT.ii[[i]])) - as.matrix(HAT.ii[[i]]))) %*%
                  (as.matrix(g.i[[i]] * w.k.clus[[i]])),
                  g.i = g.i, w.k.clus = w.k.clus, HAT.ii = HAT.ii, e.i = e.i)

    D.i <- lapply(1:length(t.E.i),
                  function(i, g.i, w.k.clus, HAT.ii, e.i)
                  t(as.matrix(g.i[[i]]) * w.k.clus[[i]]) %*%
                  ginv(diag(nrow(HAT.ii[[i]])) - as.matrix(HAT.ii[[i]])) %*%
                  as.matrix(e.i[[i]]),
                  g.i = g.i, w.k.clus = w.k.clus, HAT.ii = HAT.ii, e.i = e.i)

    v.J1.i <- lapply(1:length(t.E.i),
                  function(i, D.i)
                  (D.i[[i]] - mean(sapply(D.i, mean)))^2,
                  D.i = D.i)


### Jackknife
    Q.i <- lapply(1:length(g.i),
              function(i, A.pi.s.inv, X.sample, w.k.clus, HAT.ii, e.i)
                A.pi.s.inv %*% t(as.matrix(X.sample[[i]][,2:(X.dim + 1)]) * w.k.clus[[i]]) %*%
                ginv(diag(nrow(HAT.ii[[i]])) - as.matrix(HAT.ii[[i]])) %*%
                as.matrix(e.i[[i]]),
               A.pi.s.inv = A.pi.s.inv, X.sample = X.sample, w.k.clus = w.k.clus, HAT.ii = HAT.ii, e.i = e.i)

    G.i <- lapply(1:length(g.i),
              function(i, w.k.clus, HAT.ii, X.sample, B.hat)
                t(w.k.clus[[i]]) %*%
                ginv(diag(nrow(HAT.ii[[i]])) - as.matrix(HAT.ii[[i]])) %*%
                (as.matrix(HAT.ii[[i]]) %*% as.matrix(X.sample[[i]][,"Y.Pop"]) -
                as.matrix(X.sample[[i]][,2:(X.dim + 1)] ) %*% B.hat
                ),
                w.k.clus = w.k.clus, HAT.ii = HAT.ii, X.sample = X.sample, B.hat = B.hat)

    K.i <- lapply(1:length(g.i),
              function(i, T.x, w.k.clus, a, X.sample, B.hat, Q.i)
                (T.x - a * w.k.clus[[i]] %*% as.matrix(X.sample[[i]][,2:(X.dim + 1)])) %*%
                (B.hat - Q.i[[i]]),
                T.x = T.x, w.k.clus = w.k.clus, a = a, X.sample = X.sample, B.hat = B.hat, Q.i = Q.i)
```

```
    F.i <- lapply(1:length(g.i),
              function(i, G.i, a, K.i)
                (G.i[[i]] - mean(sapply(G.i, mean))) - (1/a) * (K.i[[i]] - mean(sapply(K.i, mean))),
                G.i = G.i, a = a, K.i = K.i)

    v.Jack.i <- lapply(1:length(t.E.i),
                  function(i, D.i, F.i)
                  (D.i[[i]] - mean(sapply(D.i, mean)))^2 -
                  2 * (D.i[[i]] - mean(sapply(D.i, mean))) * F.i[[i]] +
                  F.i[[i]]^2,
                  D.i = D.i, F.i = F.i)

 # Calculate the finite population correction factor
    fpc.srs <- 1 - a / M.clus
    fpc.pps <- 1 - 2 * sum(samp.pi.I / a) + a * sum(( pi.I.pps / a)^2 )
    if(smp == "srs") fpc <- fpc.srs else fpc <- fpc.pps

    v.ssw[j, 1] <- sum(sapply(v.ssw.clus, sum))
    v.wr[j, 1] <- (a / (a-1)) * sum((t.E.i.L - mean(t.E.i.L))^2)
    v.JL[j, 1] <- (a / (a-1)) * sum((t.E.i - mean(t.E.i))^2)
    v.r[j, 1] <- t.y.sand
    v.D[j, 1] <- sum(sapply(v.D.i, sum))
    v.J[j, 1] <- sum(sapply(v.J.i, sum))
    v.Jack[j, 1] <- (a / (a-1)) * sum(sapply(v.Jack.i, sum))
    v.J1[j, 1] <- (a / (a-1)) * sum(sapply(v.J1.i, sum))

    v.r.star[j, 1] <- fpc * t.y.sand
    v.D.star[j, 1] <- fpc * sum(sapply(v.D.i, sum))
    v.J.star[j, 1] <- fpc * sum(sapply(v.J.i, sum))
    v.Jack.star[j, 1] <- fpc * (a / (a-1)) * sum(sapply(v.Jack.i, sum))
    v.J1.star[j, 1] <- fpc * (a / (a-1)) * sum(sapply(v.J1.i, sum))

    v.D.error[j, 1] <- sum(sapply(v.D.i.err, sum))


    # Calculate the GREG
    t.y.greg[j, 1] <- sum((w.k) * (g.k) * Y.samp)

    # Pi Estimator
    t.y.pi[j, 1] <- sum((w.k) * Y.samp)

    if((j %% 10) == 0)
    {
        cat(j, format(Sys.time(), "%X"),
               "  True:  ", sum(Y.Pop[,1]),
               "  Mean t.y.pi", mean(t.y.pi[1:j,]),
               "  Mean t.y.greg: ", mean(t.y.greg[1:j,]), "\n",
               "  se t.y.greg: ", sqrt(var(t.y.greg[1:j,])),
               "  v.ssw: ", sqrt(mean(v.ssw[1:j,])),
               "  v.wr: ", sqrt(mean(v.wr[1:j,])),
               "  v.JL: ", sqrt(mean(v.JL[1:j,])),
               "  v.r: ", sqrt(mean(v.r[1:j,])),
               "  v.D: ", sqrt(mean(v.D[1:j,])),
               "  v.Jack: ", sqrt(mean(v.Jack[1:j,])),
               "  v.J1: ", sqrt(mean(v.J1[1:j,])),
               "  v.J: ", sqrt(mean(v.J[1:j,])), "\n")
    }
  }
  list(total.greg = t.y.greg, total.pi = t.y.pi,
       v.ssw = v.ssw, v.wr = v.wr, v.JL = v.JL,
       v.r = v.r, v.D = v.D, v.J = v.J, v.Jack = v.Jack, v.J1 = v.J1,
       v.r.star = v.r.star, v.D.star = v.D.star, v.J.star = v.J.star, v.Jack.star = v.Jack.star, v.J1.star = v.J1.star,
       v.D.error = v.D.error)
}
```

Appendix B

Notes for LGREG Paper

B.1 Some Asymptotic Results

In this section, we discuss four characteristics of our asymptotic framework

- the mechanism generating the clusters in our finite population
- the rate with which clusters and units are added to the sequence of finite populations
- the rate with which the finite population increases, with respect to the sample
- the sample design

Cluster Generation

Fuller (2009) describes two methods for generating the series of finite populations, a superpopulation framework and a fixed sequence framework. We take the superpopulation framework. In this framework, a sequence of finite populations is generated from a random mechanism. That is, new clusters are added to our growing finite population according to some superpopulation model.

Our superpopulation model, requires that the cluster totals, $\mathbf{t}_{iy}$ and $\mathbf{t}_{ix}$, are generated from an arbitrary model with finite first and second moments. This prevents any one cluster from dominating the population. We assume that each $\mathbf{t}_{iy}$ is independent of every other cluster total and identically distributed from the distribution function $F\left\{\mathbf{t}_{iy}\right\}$ such

that

$$E\left\{\mathbf{t}_{iy}\right\} = \boldsymbol{\mu}$$

$$E\left\{\left(\widehat{\mathbf{t}}_{iy} - \boldsymbol{\mu}\right)\left(\widehat{\mathbf{t}}_{iy} - \boldsymbol{\mu}\right)^{\top}\right\} = \boldsymbol{\Sigma}_{yy}$$

These two restrictions imply the following

$$\lim_{N\longrightarrow\infty}\mathbf{S}_{yyN} = \boldsymbol{\Sigma}_{yy}$$

$$\text{var}\left\{\bar{\mathbf{t}}_n - \bar{\mathbf{t}}_N | F_N\right\} = O_p\left(n_N^{-1}\right)$$

$$\bar{\mathbf{t}}_n - \bar{\mathbf{t}}_N | F_N = O_p\left(n_N^{-\frac{1}{2}}\right)$$

where

$$\mathbf{S}_{yyN} = \frac{1}{N-1}\sum_{i\in\mathscr{U}_I}\left(\mathbf{t}_{iN} - \bar{\mathbf{t}}_N\right)\left(\mathbf{t}_{iN} - \bar{\mathbf{t}}_N\right)^{\top}$$

$$\bar{\mathbf{t}}_N = \frac{1}{N}\sum_{i\in\mathscr{U}_I}\widehat{\mathbf{t}}_{iN}$$

$$\bar{\mathbf{t}}_n = \frac{1}{n_N}\sum_{i\in\mathfrak{s}_I}\widehat{\mathbf{t}}_{iN}$$

$$\widehat{\mathbf{t}}_{iN} = \sum_{k\in\mathfrak{s}_i} d_{k|i}\mathbf{y}_{kN}.$$

Here F_N is the finite population with N clusters, $d_{k|i}$ inverse of the probability of selecting unit k given that cluster i was selected, and n_N is the average number of clusters in sample from the population of size N. Note that we often omit the N subscript for simplicity of notation. However, when we need to emphasize that the sample size changes as the population grows, we will use the n_N notation.

Similarly, we also place restrictions on our covariates. That is $\mathbf{t}_{1x}, ..., \mathbf{t}_{Nx}\ iid\left(\boldsymbol{\mu}_x, \Sigma_{xx}\right)$.

Relationship between Cluster and Unit Growth

Each finite population has N clusters where $N \longrightarrow \infty$. The number of units in each cluster may vary from cluster to cluster, however the cluster size must be bounded such that $M_i = O(1) \forall i$. The necessary condition to prevent any cluster from growing faster than the population is Assumption 13 of Appendix A.1 on page 258.

Relationship between Population and Sample Growth

Assumption 12 in Appendix A.1 on page 258 dictates the relationship between our population and sample sizes as the population increases. Both the sample size and the population sizes increase to infinity, but the population grows at a faster rate.

Sample Design

Lastly, we place some conditions on our sampling mechanism. As the sample and population sizes grow, we do not want the probability of selecting any unit to increase. This requirement is written as Assumption 14 in Appendix A.1 on page 258.

The superpopulation framework presented here is similar in practice to the framework presented in Appendix A.1 with the exception of the description of how the clusters are generated. Practically, the two frameworks are similar, even though there are philosophical differences between the superpopulation and fixed sequence frameworks.

B.2 Logistic Models

In this section, we introduce binary, binomial, and multinomial logistic regression. After presenting these three density functions, we use the theory of maximum likelihood and generalized linear models to derive estimating equations for the population parameter vector $\mathbf{B}$. We conclude with some notes about model fitting and residuals.

The most important difference between linear regression and logistic regression is that in logistic regression a nonlinear transformation of the expected value of the response variable is related to explanatory variables, while in linear regression the expected value of the observed response variable is linearly related to explanatory variables. The simple **linear** regression model for the kth unit can be written as

$$E_M\left(Y_k\right) = \mu_k = \mathbf{x}_k^\top \boldsymbol{\beta} \tag{B.1}$$

where $\mathbf{x}_k$ and $\boldsymbol{\beta}$ are p-dimensional vectors. For **logistic** regression, the corresponding regression model is

$$E_M\left(Y_k\right) = \mu_k = g\left(\mathbf{x}_k^\top \boldsymbol{\beta}\right) \tag{B.2}$$

where g is called the link function.

B.2.1 Data

Shao (2003, sec 4.4) describes the Generalized Linear Model for a multivariate response. Although the notation is a bit complicated for multivariate response data, such complexity is necessary to understand the mathematics of the multinomial distribution. We begin with a C-dimensional response vector for the kth sample unit. For example, the number of grams of salt, sugar, and fat that unit k eats a day can be put into a vector,

$$\mathbf{Y}_k = \begin{bmatrix} Y_{k,\text{salt}} \\ Y_{k,\text{sugar}} \\ Y_{k,\text{fat}} \end{bmatrix}$$

In this case, we see that $C = 3$.

As another example, consider a time usage study. Respondents are asked to record how many minutes are spent eating, sleeping, and doing other activities for a day. In this case, the total number of minutes in the day is known and fixed at $1,440$. Assuming that what is performed at one minute is independent of what happens in the next minute, the sum of all minutes for the day in each category is a multinomial random vector. One example of $\mathbf{Y}_k$ is

$$\mathbf{y}_k = \begin{bmatrix} y_{eating,k} \\ y_{sleeping,k} \\ y_{other,k} \end{bmatrix} = \begin{bmatrix} 90 \\ 480 \\ 870 \end{bmatrix}$$

In this case, there are $C = 3$ categories. Furthermore, the sum of all categories in a multinomial random variable is a known constant, denoted z_k. In this time usage example, $z_k = 1,440$.

Categorical responses are ubiquitous. If we categorize the labor force status as: not in the labor force, employed, or unemployed, then we can use multinomial logistic regression to model which category a person is in. Another example would be to model the mode of transportation one takes to work where the options are: car, bike, public transportation, walk, or another mode. Questions where respondents must select one in a series of options can be modeled by a multinomial distribution. For example, the American Community Survey asks "Which FUEL is used MOST for heating this house, apartment, or mobile home?" followed by nine response options. In this case, $C = 9$ and $z_k = 1$.

Often z_k is fixed to be 1 so that $\mathbf{Y}_k$ is a random vector with $C - 1$ elements equal to 0 and exactly one element equal to 1. For example, if there are 5 age classi-

fications then $\mathbf{y}_k = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}^\top$ for a person in the youngest age group, $\mathbf{y}_k = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \end{bmatrix}^\top$ for a sample unit in the middle age group, and $\mathbf{y}_k = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 \end{bmatrix}^\top$ for a sample unit in the oldest age group. Notice, that $\mathbf{y}_k$ is C dimensional. One of the C categories is redundant since $z_k = \sum_{c=1}^{C} y_{kc}$. To make all elements of $\mathbf{y}_k$ we remove one category, called the baseline category. Typically, the last category, C, is set as the baseline category. The full rank $C-1$ dimensional response vector is denoted $\mathbf{y}_{\star k}$.

Both the Bernoulli and binomial distributions are examples of the multinomial distribution. If $C = 2$ and $z_k = 1$, then $\mathbf{Y}_k$ is a Bernoulli random variable. If $C = 1$ and $z_k \geq 1$ then, $\mathbf{Y}_k$ is a binomial random variable.

In binomial logistic regression, the response variable, Y_k, is a binomial random variable that can be any natural number from 0 to z_k. The binomial distribution is characterized by the number of successful events that occurred in a fixed number of independent trials. The total number of trials, z_k, can be different from one sample unit to another, but must be a known nonrandom quantity. For example, if school enrollment is fixed and known, the total number of students receiving a free or reduced lunch can be modeled with the binomial distribution. If the total number of mailable households in every Census tract is known, then the total number of households that would not participate in a mail census can be modeled by a binomial distribution.

An alternative formulation of the binomial distribution is to divide the total number of events by the total number of trials. This results in a response variable that is a proportion or percent. For example, the percent of one's income that is spent on groceries can be modeled by the binomial distribution. Thus, binomial logistic regression is often used to

model counts or proportions. Table B.1 shows the two different parameterizations of the binomial density function. Binary logistic regression is a special case of binomial logistic regression when the total number of trials for all units in the population is fixed at 1.

In binary logistic regression, the response variable is a Bernoulli random variable. Logistic regression is commonly used when the response can take on one of two values. For example, it can be used to model the presence or absence of a disease, whether a person has smoked at least 100 cigarettes, whether an elementary school student is proficient at math or not, whether a person is in the labor force or not, whether a person has been a victim of a violent crime, whether a housing unit is vacant or not, and whether someone was satisfied with a product or not. In binary logistic regression, the response variable, y_k, can take on one of two values, usually written as 0 for failure or 1 for success.

B.2.2 Density Function

Many common densities can be written as a members of the *exponential dispersion family*, which is defined as

$$f\left(\mathbf{y}_k; \boldsymbol{\eta}_k, \phi_k\right) = e^{\frac{\left[\mathbf{y}_k^\top \boldsymbol{\eta}_k - \zeta(\boldsymbol{\eta}_k)\right]}{\phi_k} + h(\mathbf{y}_k, \phi_k)} \tag{B.3}$$

We note that $f : \mathbb{R}^C \to \mathbb{R}^1$. That is, f maps a C-dimensional vector to the real numbers. In this section, we prove that the multinomial, binomial, and Bernoulli distributions are a member of the exponential dispersion family.

One important characteristic of all members of the exponential family is that

$$E_M\left(\mathbf{Y}_k\right) = \boldsymbol{\mu}_k = \mu\left(\boldsymbol{\eta}_k\right) = \frac{\partial}{\boldsymbol{\eta}_k} \zeta\left(\eta_k\right) \tag{B.4}$$

and

$$\mathrm{var}_M\left(\mathbf{Y}_k\right) = \psi_k \frac{\partial\partial}{\boldsymbol{\eta}_k \boldsymbol{\eta}_k^{\top}} \zeta\left(\eta_k\right) \tag{B.5}$$

The multinomial distribution can be motivated by a conditional Poisson distribution. The multinomial distribution is the distribution of a set of independent Poisson variables conditional on the sum of the Poisson variables (McCullagh and Nelder (1999)). For example, if $Y_{k,1}$, $Y_{k,2}$ and $Y_{k,3}$ are independent Poisson random variables and $z_k = \sum_{c=1}^{3} Y_{k,c}$, then the multinomial density is

$$f\left(y_{k,1}, y_{k,2}, y_{k,3} | z_k; \pi_{k,1}, \pi_{k,2}, \pi_{k,3}\right) = \frac{z_k!}{\prod_{c=1}^{C} y_{k,c}} \prod_{c=1}^{C} {\pi_{k,c}}^{y_{k,c}}$$

Where $\pi_{k,c}$ is the probability that an individual trial for the kth unit will be in the cth category. Unfortunately, this is not a full rank density because one of the C categories is redundant. That is, $1 = \sum_{c=1}^{C} \pi_{k,c}$. To remove this redundancy, one of the categories is chosen to be the baseline. The remaining probabilities are divided by the baseline. Doing so gives us the full rank density. That is, the full rank probability mass function for a multinomial random vector with the final category as the baseline is,

$$f\left(\mathbf{y}_k | z_k; \boldsymbol{\pi}_{\star k}\right) = \frac{z_k!}{\prod_{c \in C} y_{k,c}} \prod_{c \in C} {\pi_{k,c}}^{y_{k,c}}$$

We now prove that the multinomial density is a member of the exponential family.

$$
\begin{aligned}
f\left(\mathbf{y}_k|z_k;\boldsymbol{\pi}_{\star k}\right) &= \frac{z_k!}{\prod_{c\in C} y_{k,c}} \prod_{c\in C_\star} \pi_{k,C}^{y_{k,C}} \pi_{k,c}^{y_{k,c}} \\
&= \frac{z_k!}{\prod_{c\in C} y_{k,c}} \pi_{k,C}^{y_{k,C}} \prod_{c\in C_\star} \pi_{k,c}^{y_{k,c}} \\
&= \frac{z_k!}{\prod_{c\in C} y_{kc}} \pi_{k,C}^{z_k-\sum_{c\in C_\star} y_{kc}} \prod_{c\in C_\star} \pi_{kc}^{y_{kc}} \\
&= \frac{z_k!}{\prod_{c\in C} y_{kc}} \frac{\pi_{k,C}^{z_k}}{\pi_{k,C}^{\sum_{c\in C_\star} y_{kc}}} \prod_{c\in C_\star} \pi_{kc}^{y_{kc}} \\
&= \frac{z_k!}{\prod_{c\in C} y_{kc}} \pi_{k,C}^{z_k} \prod_{c\in C_\star} \frac{\pi_{kc}^{y_{kc}}}{\pi_{k,C}^{\sum_{c\in C_\star} y_{kc}}} \\
&= \frac{z_k!}{\prod_{c\in C} y_{kc}} e^{z_k \ln\left(\pi_{k,C}\right)+\sum_{c\in C_\star} y_{kc}\ln(\pi_{kc})-\sum_{c\in C_\star} y_{kc}\ln\left(\pi_{k,C}\right)} \\
&= \frac{z_k!}{\prod_{c\in C} y_{kc}} e^{z_k \ln\left(\pi_{k,C}\right)+\sum_{c\in C_\star} y_{kc}\ln\left(\frac{\pi_{kc}}{\pi_{k,C}}\right)} \\
&= \frac{z_k!}{\prod_{c\in C} y_{kc}} e^{\sum_{c\in C_\star} y_{kc}\ln\left(\frac{\pi_{kc}}{\pi_{k,C}}\right)--z_k \ln\left(\pi_{k,C}\right)}. \qquad \text{(B.6)}
\end{aligned}
$$

Comparing Equation (B.6) to Equation (B.3), we see the multinomial distribution is a member of the exponential family with parameters given in Table B.1. Shao (2003, p. 98) also confirms that the multinomial distribution is a member of the full rank exponential dispersion family with the parameters in Table B.1. We note that $f : \mathbb{R}^{C_\star} \to \mathbb{R}^1$. That is, f maps a $C_\star$ dimensional vector to the real numbers.

When $C = 2$, the multinomial density simplifies to the binomial density. Furthermore, when $C = 2$ and $z_k = 1$, the multinomial distribution simplifies to the Bernoulli density. Table B.1 shows that the binomial and Bernoulli distributions are members of the exponential family.

Table B.1: Distributions of the Exponential Family

Name	Density	η_k	$\zeta(\eta_k)$	ϕ_k	$h(y_k, \phi_k)$
Bernoulli	$\pi_k^{y_k}(1-\pi_k)^{1-y_k}$	$\ln\left(\frac{\pi_k}{1-\pi_k}\right)$	$\ln(1+e^{\eta_k})$	1	0
Binomial-Percent	$\binom{z_k}{z_k p_k}\pi_k^{z_k p_k}(1-\pi_k)^{z_k - z_k p_k}$	$\ln\left(\frac{\pi_k}{1-\pi_k}\right)$	$\ln(1+e^{\eta_k})$	$\frac{1}{z_k}$	$\ln\binom{z_k}{z_k p_k}$
Binomial-Count	$\binom{z_k}{y_k}\pi_k^{y_k}(1-\pi_k)^{z_k - y_k}$	$\ln\left(\frac{\pi_k}{1-\pi_k}\right)$	$\ln(1+e^{\eta_k})$	$\frac{1}{z_k}$	$\ln\binom{z_k}{y_k}$
Multinomial-Percent	$\binom{z_k}{z_k p_{ck}}\pi_k^{z_k p_{ck}}(1-\pi_{ck})^{z_k - z_k p_{ck}}$	$\ln\left(\frac{\pi_{ck}}{1-\pi_{ck}}\right)$	$\ln(1+e^{\eta_{ck}})$	$\frac{1}{z_k}$	$\ln\binom{z_k}{z_k p_{ck}}$
Multinomial-Count	$\binom{z_k}{y_{ck}}\pi_{ck}^{y_{ck}}(1-\pi_{ck})^{z_k - y_{ck}}$	$\ln\left(\frac{\pi_{ck}}{1-\pi_{ck}}\right)$	$\ln(1+e^{\eta_{ck}})$	$\frac{1}{z_k}$	$\ln\binom{z_k}{y_{ck}}$

B.2.3 Logistic Link Function

If we assume that μ_k is the same for all elements in the population, we could use maximum likelihood to estimate μ_k. This simplifying assumption is often too restrictive. An alternative and more flexible assumption is to build a model for each $\boldsymbol{\mu}_k$. This approach leads to the generalized linear model. If explanatory variables are available, we can use them to model $\boldsymbol{\mu}_k$. Suppose we have a p-dimensional vector of covariates about the kth unit, denoted $\mathbf{x}_k$. Let $\boldsymbol{\beta}_c$ be the p-dimensional vector of coefficients associated with the cth category. Moreover, let $\boldsymbol{\beta}$ be the $p \times c$ matrix of coefficients associated with all categories. In the food example, if age is the covariate, then

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{\text{salt,intercept}} & \beta_{\text{sugar,intercept}} & \beta_{\text{fat,intercept}} \\ \beta_{\text{salt,age}} & \beta_{\text{sugar,age}} & \beta_{\text{fat,age}} \end{bmatrix}$$

The link function relates μ_{kc} to the linear predictor $\mathbf{x}_k^\top \boldsymbol{\beta}_c$. Specifically, the link function is defined as

$$g(\mu(\eta_{kc})) = \mathbf{x}_k^\top \boldsymbol{\beta}_c.$$

The multivariate logit link function is defined as

$$g\left[\mu\left(\eta_{kc}\right)\right] = \ln\left(\frac{\mu\left(\eta_{kc}\right)}{z_k - \sum_{c\in C_\star}\mu\left(\eta_{kc}\right)}\right) = \mathbf{x}_k^\top\boldsymbol{\beta}_c.$$

One important attribute of the logit link is that g and μ are inverse functions. When $g^{-1} = \mu$, then g is said to be the **canonical** link and $\boldsymbol{\eta}_{kc} = \mathbf{x}_k^\top\boldsymbol{\beta}_c$. Relating the explanatory variables to the expected response by using the logit link gives,

$$\mu_{kc} = z_k\frac{e^{\mathbf{x}_k^\top\boldsymbol{\beta}_c}}{1+\sum_{c\in C_\star}e^{\mathbf{x}_k^\top\boldsymbol{\beta}_c}} \tag{B.7}$$

Since, we can write our density function in Equation (B.6) as

$$\begin{aligned} f\left(\mathbf{y}_k|z_k;\boldsymbol{\pi}_{\star k}\right) &= \frac{z_k!}{\prod_{c\in C}y_{kc}}e^{\sum_{c\in C_\star}y_{kc}\ln\left(\frac{\pi_{kc}}{\pi_{k,C}}\right)--z_k\ln\left(\pi_{k,C}\right)} \\ &= \frac{z_k!}{\prod_{c\in C}y_{kc}}e^{\sum_{c\in C_\star}y_{kc}\ln\left(\frac{\mu_{kc}}{\mu_{k,C}}\right)--z_k\ln\left(\mu_{k,C}\right)} \end{aligned} \tag{B.8}$$

where μ_{kc} is defined in Equation (B.7).

In summary, for multinomial regression

$$E_M\left(Y_{kc}\right) = \mu_{kc} = \frac{z_k e^{\mathbf{x}_{kc}^\top\boldsymbol{\beta}_c}}{1+\sum_{\mathrm{c}=1}^{C-1}e^{\mathbf{x}_{kc}^\top\boldsymbol{\beta}_c}} \tag{B.9}$$

which can also be written as

$$g\left(\mu_{kc}\right) = \mathbf{x}_k^\top\boldsymbol{\beta}_c = \ln\frac{\mu_{kc}}{1-\mu_{kc}} \tag{B.10}$$

where g is the link function. In binary logistic regression, $z_k = 1$ and $C = 2$. Thus the link function is

$$g\left(\mu_k\right) = \mathbf{x}_k^\top\boldsymbol{\beta} = \ln\frac{\mu_k}{1-\mu_k} \tag{B.11}$$

and

$$E_M\left(Y_k\right) = \frac{e^{\mathbf{x}_k^\top \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_k^\top \boldsymbol{\beta}}} \tag{B.12}$$

For **binomial** logistic regression, the corresponding link function is

$$g\left(\mu_k\right) = \mathbf{x}_k^\top \boldsymbol{\beta} = \ln \frac{\mu_k}{z_k - \mu_k} \tag{B.13}$$

and

$$E_M\left(Y_k\right) = \mu_k = \frac{z_k e^{\mathbf{x}_k^\top \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_k^\top \boldsymbol{\beta}}} \tag{B.14}$$

One advantage of logistic regression is that $g\left(\mu_k\right) = \eta_k$ is the *log odds*. It is the natural log of the odds, the probability of a success to that of a failure for a binary response.

B.2.4 Likelihood Equations

We now use our density functions to write our log-likelihood equations. Taking the log of our likelihood in Equation (B.8) gives us our population log-likelihood. Agresti (2002)[p. 273] and Kutner et al. (2005)[p. 614] show that the log-likelihood for multinomial logestic regression is

$$\ell = \sum_{k \in \mathscr{U}} \left[\ln \left(\frac{C z_k!}{y_{k1}! y_{k2}! \cdots y_{kC}!} \right) + \mathbf{x}_{kC}^\top \boldsymbol{\beta}_C y_{kC} - z_k \ln \left(1 + \sum_{c=1}^{C-1} e^{\mathbf{x}_{kc}^\top \boldsymbol{\beta}_c} \right) \right] \tag{B.15}$$

The values of $\boldsymbol{\beta}$ that maximize Equation (B.15) can be found numerically.

For binomial logistic regression the log-likelihood simplifies to

$$\ell = \sum_{k \in \mathscr{U}} \left[\begin{pmatrix} z_k \\ \\ y_k \end{pmatrix} \mathbf{x}^\top \boldsymbol{\beta} y_k - z_k \ln \left(1 + e^{\mathbf{x}_k^\top \boldsymbol{\beta}} \right) \right].$$

Lastly, the log-likelihood for Bernoulli logistic regression is

$$\ell = \sum_{k \in \mathscr{U}} \mathbf{x}_k^\top \boldsymbol{\beta} y_k - \ln\left(1 + e^{\mathbf{x}_k^\top \boldsymbol{\beta}}\right)$$

Table B.2 summarizes the log-likelihood equations that are used to estimate $\boldsymbol{\beta}$. These estimating equations can be solved numerically using any nonlinear optimization program. Alternatively, they can also be solved by differentiating them, setting them equal to 0, and solving for $\boldsymbol{\beta}$. Either way, the solution to the maximum likelihood equations is denoted, $\widehat{\boldsymbol{\beta}}$. Table B.2 shows both the log-likelihood estimating equations as well as the derivative of them for the three different cases of logistic regression. Derivations for these estimating equations and other techniques to estimate $\boldsymbol{\beta}$ can be found in numerous sources.

Table B.2: Logistic Regression Estimating Equations

Distribution of Response	Log Likelihood $\widehat{L}(\boldsymbol{\beta})$	Estimating Equations
Bernoulli	$\sum_{k \in \mathscr{U}} \left[y_k \left(\mathbf{x}_k^\top \boldsymbol{\beta}\right) - \ln\left(1 + e^{\mathbf{x}_k^\top \boldsymbol{\beta}}\right)\right]$	$\sum_{k \in \mathscr{U}} \left(y_k - \frac{e^{\mathbf{x}_k^\top \boldsymbol{\beta}}}{1+e^{\mathbf{x}_k^\top \boldsymbol{\beta}}}\right) \mathbf{x}_k$
Binomial	$\sum_{k \in \mathscr{U}} \left[y_k \left(\mathbf{x}_k^\top \boldsymbol{\beta}\right) - z_k \ln\left(1 + e^{\mathbf{x}_k^\top \boldsymbol{\beta}}\right)\right]$	$\sum_{k \in \mathscr{U}} \left(y_k - z_k \frac{e^{\mathbf{x}_k^\top \boldsymbol{\beta}}}{1+e^{\mathbf{x}_k^\top \boldsymbol{\beta}}}\right) \mathbf{x}_k$
Multinomial	$\sum_{k \in \mathscr{U}} \left[\mathbf{y}_k^\top \left(\mathbf{x}_k^\top \boldsymbol{\beta}\right) - z_k \ln\left(1 + e^{\mathbf{x}_k^\top \boldsymbol{\beta}}\right)\right]$	$\sum_{k \in \mathscr{U}} \left(\mathbf{y}_k - z_k \frac{e^{\mathbf{x}_k^\top \boldsymbol{\beta}}}{1+e^{\mathbf{x}_k^\top \boldsymbol{\beta}}}\right) \mathbf{x}_k^\top$

B.2.5 Estimating Equations

The values of $\boldsymbol{\beta}_c$ that maximize and minimize the log-likelihood can be computed by finding the roots of the derivatives with respect to $\boldsymbol{\beta}_c$. We call the derivatives of the log-likelihood our estimating equations. Differentiating the log-likelihood gives the p

estimating equations for $\boldsymbol{\beta}_c$

$$\sum_{k \in \mathscr{U}} \left\{ \left[y_{kc} - \frac{z_k e^{\mathbf{x}_k^\top \boldsymbol{\beta}_c}}{1 + \sum_{c=1}^{C-1} e^{\mathbf{x}_k^\top \boldsymbol{\beta}_c}} \right] \mathbf{x}_k^\top \right\} = \mathbf{0}.$$

If we treat our finite population as one possible population generated from the super-population model, then $\mathbf{B}_c = \widehat{\boldsymbol{\beta}}_c$ is the solution to our estimating equation. Unless we take a census, we cannot compute $\mathbf{B}_c$. Instead, we estimate it using the pseudomaximum likelihood approach. That is we estimate $\mathbf{B}_c$ by solving

$$\sum_{k \in \mathfrak{s}} d_k \left\{ \left[y_{kc} - \frac{z_k e^{\mathbf{x}_k^\top \mathbf{B}_c}}{1 + \sum_{c=1}^{C-1} e^{\mathbf{x}_k^\top \mathbf{B}_c}} \right] \mathbf{x}_k^\top \right\} = \mathbf{0}.$$

There is no analytic solution to these estimating equations. Numerical methods such as Newton-Raphson, Fisher-scoring, or Iterative Reweighted Least Squares are often used to solve these estimating equations. (Shao 2003, p. 283) shows that the Newton-Raphson and Fisher-scoring methods are identical. Both methods require iterating:

$$\widehat{\boldsymbol{\beta}}^{(t+1)} = \widehat{\boldsymbol{\beta}}^{(t)} \left[M_n \left(\widehat{\boldsymbol{\beta}}^{(t)} \right) \right]^{-1} s_n \left(\widehat{\boldsymbol{\beta}}^{(t)} \right)$$

where

$$M_n \left(\widehat{\boldsymbol{\beta}}^{(t)} \right) = \sum_{k \in \mathfrak{s}} \left[\psi' \left(\boldsymbol{\beta}^\top \mathbf{x}_k \right) \right]^2 \zeta'' \left[\psi \left(\boldsymbol{\beta}^\top \mathbf{x}_k \right) \right] \omega_k \mathbf{x}_k \mathbf{x}_k^\top$$

$$s_n \left(\boldsymbol{\beta} \right) = \phi \frac{\partial \ln \ell \left(\boldsymbol{\beta} \right)}{\partial \boldsymbol{\beta}}$$

For multinomial logistic regression without a dispersion parameter, these equations simplify to

$$
\begin{aligned}
M_n\left(\widehat{\boldsymbol{\beta}}^{(t)}\right) &= \sum_{k\in\mathfrak{s}} \operatorname{var}\left(\mathbf{y}_{\star k}\right)\mathbf{x}_k\mathbf{x}_k^\top \\
&= \frac{\partial^2 \ln \ell\left(\boldsymbol{\beta}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} \\
&= \mathbf{J}\left(\boldsymbol{\beta}\right) \\
s_n\left(\boldsymbol{\beta}\right) &= \sum_{k\in\mathfrak{s}} \left[\mathbf{y}_k - \boldsymbol{\mu}_k\right]\mathbf{x}_k
\end{aligned}
$$

We note that $\mathbf{J}\left(\boldsymbol{\beta}\right)$ is the Jacobian of the estimating equations. Also, since $\boldsymbol{\beta}_c$ is a vector of length p and there are $C_\star$ unique categories, we will need to solve $p \cdot C_\star$ equations for $p \cdot C_\star$ unknowns. That is, each of the $C_\star$ independent categories will have a set of p coefficients that we will need to estimate. The solution to the pseudo maximum likelihood estimating equation is denoted $\widehat{\mathbf{B}}_c$ To determine if we have the maximum or minimum, we must consider the second derivative,

$$
\frac{\partial^2 \ell_k}{\partial\boldsymbol{\beta}_c\partial\boldsymbol{\beta}_c^\top}
$$

Specifically, the second derivative must be negative. Using our sample, the solution to the estimating equations is $\widehat{\beta}$ [1].

For binomial logistic regression, the estimating equations are,

$$
w = \sum_{k\in\mathfrak{s}} d_k \left\{\left[y_k - \frac{z_k e^{\mathbf{x}_k^\top\boldsymbol{\beta}}}{1+e^{\mathbf{x}_k^\top\boldsymbol{\beta}}}\right]\mathbf{x}_k\right\} = 0.
$$

[1] Sometimes the solution to the estimating equations gives a value outside the range of possible values. These "boundary" cases have been well studied and documented.

Furthermore, the estimating equations for $\boldsymbol{\beta}$ in binary logistic regression are

$$w = \sum_{k \in \mathfrak{s}} d_k \left\{ \left[y_k - \frac{e^{\mathbf{x}_k^\top \boldsymbol{\beta}}}{1 + e^{\mathbf{x}_k^\top \boldsymbol{\beta}}} \right] \mathbf{x}_k \right\} = 0.$$

B.2.6 Residuals

As in any model building process, it is important to diagnose the fit of the model and deal with influential observations and outliers. Agresti (2002), Hilbe (2009), Hosmer and Lemeshow (2000), and Kutner et al. (2005) discuss various goodness of fit statistics and residuals for logistic regression models.

For multinomial logistic regression, the Pearson residual is defined as

$$\begin{aligned} \mathbf{r}_k^p &= \left(\mathbf{Y}_{k\star} - \hat{\boldsymbol{\mu}}_k\right)^\top \left\{\text{diag}\left[\widehat{\text{var}}\left(\mathbf{Y}_{k\star}\right)\right]\right\}^{-\frac{1}{2}} \\ &= \left(\mathbf{Y}_{k\star} - z_k \hat{\boldsymbol{\pi}}_k\right)^\top \left\{ z_k \left[\text{diag}\left(\widehat{\boldsymbol{\pi}}_{k\star}\right) - \text{diag}\left(\widehat{\boldsymbol{\pi}}_{k\star} \widehat{\boldsymbol{\pi}}_{k\star}^\top\right)\right]\right\}^{-\frac{1}{2}} \end{aligned}$$

As in linear regression, residuals are useful in assessing the fit of binary logistic regression models. The Pearson residuals for a binomial random variable are

$$r_k^p = \frac{y_k - z_k \widehat{\pi}_k}{\sqrt{z_k \widehat{\pi}_k \left(1 - \widehat{\pi}_k\right)}}$$

where

$$\widehat{\pi}_k = \frac{e^{\mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}}}{1 + e^{\mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}}}$$

The Pearson residual with $\boldsymbol{\beta}$ instead of $\widehat{\boldsymbol{\beta}}$ is asymptotically standard normal; however, Agresti (2002, sec 6.2.1) argues that the Pearson residual using $\widehat{\boldsymbol{\beta}}$ will be slightly less variable than the standard normal. To correct for the fact that $\sum_{k \in \mathfrak{s}} \left(r_k^p\right)^2$ will underesti-

mate σ_y^2, Agresti (2002) suggests using the standardized residual,

$$r_k^{sp} = \frac{y_k - z_k\widehat{\pi}_k}{\sqrt{z_k\widehat{\pi}_k\left(1-\widehat{\pi}_k\right)\left(1-\widehat{h}_k\right)}}$$

where $\widehat{h}_k$ is the leverage, taken from the diagonal of

$$\mathbf{H} = \boldsymbol{\mu}^{\frac{1}{2}}\mathbf{X}\left(\mathbf{X}^\top\boldsymbol{\mu}\mathbf{X}\right)^{-1}\mathbf{X}^\top\boldsymbol{\mu}^{\frac{1}{2}}$$

The deviance residual provides an alternative residual which has similar properties to the Pearson residual. The deviance residual is

$$r_k^d = \sqrt{2\left(y_k \ln\frac{y_k}{z_k\widehat{\pi}_k} + \left(z_k - y_k\right)\ln\frac{z_k - y_k}{z_k - z_k\widehat{\pi}_k}\right)} \times \operatorname{sign}\left(y_k - z_k\widehat{\pi}_k\right)$$

Like the Pearson residual, this residual can also be standardized.

The Pearson residuals for a Bernoulli random variable are

$$r_k^p = \frac{y_k - \widehat{\pi}_k}{\sqrt{\widehat{\pi}_k\left(1-\widehat{\pi}_k\right)}}$$

The Pearson residual with $\boldsymbol{\beta}$ instead of $\widehat{\boldsymbol{\beta}}$ is asymptotically standard normal; however, Agresti (2002, sec 6.2.1) argues that the Pearson residual using $\widehat{\boldsymbol{\beta}}$ will be slightly less variable than the standard normal. To correct for the fact that $\sum_{k\in\mathfrak{s}}\left(r_k^p\right)^2$ will underestimate σ_y^2, Agresti (2002) suggests using the standardized residual,

$$r_k^{sp} = \frac{y_k - \widehat{\pi}_k}{\sqrt{\widehat{\pi}_k\left(1-\widehat{\pi}_k\right)\left(1-\widehat{h}_k\right)}}$$

The deviance residual provides an alternative residual which has similar properties to the Pearson residual. The deviance residual is

$$r_k^d = \sqrt{2\left(y_k \ln\frac{y_k}{\widehat{\pi}_k} + \left(1 - y_k\right)\ln\frac{1 - y_k}{1 - \widehat{\pi}_k}\right)} \times \operatorname{sign}\left(y_k - \widehat{\pi}_k\right)$$

Like the Pearson residual, this residual can also be standardized.

B.3 Derivation of Multivariate Calibration GREG

Let $\underset{n\times C}{\mathbf{y}}$ be a matrix containing the responses for C variables. For example, $\mathbf{y}$ could contain the response to a categorical question with C possible outcomes or $\mathbf{y}$ could contain the responses to C unrelated questions. Also, let $\underset{C\times 1}{\widehat{\boldsymbol{t}}_y^{gr}}$ be a vector estimating the total of the C responses. The calibrated estimated total is

$$\widehat{\boldsymbol{t}}_y^{gr} = \mathbf{y}^\top \mathbf{w}^{gr}$$

where $\underset{n\times 1}{\mathbf{w}_k^{gr}}$ is found by minimizing the objective function

$$\frac{1}{2}\left(\mathbf{d} - \mathbf{w}^{gr}\right)^\top \boldsymbol{\Pi}\mathbf{Q}^{-1}\left(\mathbf{d} - \mathbf{w}^{gr}\right)$$

subject to the constraint

$$\mathbf{X}_{\mathfrak{s}}^\top \mathbf{w}^{gr} = \mathbf{X}_{\mathscr{U}}^\top \mathbf{1}.$$

Notice that if the first column of $\mathbf{X}$ is $\mathbf{1}$, then the following constraint is also obtained.

$$\mathbf{1}^\top \mathbf{w}^{gr} = N.$$

Using the method of Lagrange multipliers, our estimating equation is

$$\phi = \frac{1}{2}\left(\mathbf{d} - \mathbf{w}^{gr}\right)^\top \boldsymbol{\Pi}\mathbf{Q}^{-1}\left(\mathbf{d} - \mathbf{w}^{gr}\right) - \boldsymbol{\lambda}^\top\left(\mathbf{X}_{\mathfrak{s}}^\top \mathbf{w}^{gr} - \mathbf{X}_{\mathscr{U}}^\top \mathbf{1}\right)$$

where $\boldsymbol{\lambda}$ is a p by 1 vector of Lagrange multipliers and p is the rank of $\mathbf{X}$.

Differentiating ϕ with respect to $\mathbf{w}^{gr}$ gives

$$\frac{\partial \phi}{\partial \mathbf{w}^{gr}} = \boldsymbol{\Pi}\mathbf{Q}^{-1}\left(\mathbf{w}^{gr} - \mathbf{d}\right) - \mathbf{X}_{\mathfrak{s}}\boldsymbol{\lambda}.$$

Setting the partial derivative equal to 0 and solving for $\mathbf{w}^{gr}$ gives

$$
\begin{aligned}
0 &= \left(\mathbf{Q}\mathbf{\Pi}^{-1}\right)\left(\mathbf{\Pi}\mathbf{Q}^{-1}\right)\left(\mathbf{w}^{gr} - \mathbf{d}\right) - \left(\mathbf{Q}\mathbf{\Pi}^{-1}\right)\mathbf{X}_{\mathfrak{s}}\boldsymbol{\lambda} \\
&= \mathbf{w}^{gr} - \mathbf{d} - \left(\mathbf{Q}\mathbf{\Pi}^{-1}\right)\mathbf{X}_{\mathfrak{s}}\boldsymbol{\lambda} \\
\mathbf{w}^{gr} &= \mathbf{d} + \left(\mathbf{Q}\mathbf{\Pi}^{-1}\right)\mathbf{X}_{\mathfrak{s}}\boldsymbol{\lambda}.
\end{aligned}
$$

Plugging $\mathbf{w}^{gr}$ into our constraint and solving for $\boldsymbol{\lambda}$ gives

$$
\begin{aligned}
\mathbf{X}_{\mathfrak{s}}^{\top}\mathbf{w}^{gr} &= \mathbf{X}_{\mathscr{U}}^{\top}\mathbf{1} \\
\mathbf{X}_{\mathfrak{s}}^{\top}\left(\mathbf{d} + \left(\mathbf{Q}\mathbf{\Pi}^{-1}\right)\mathbf{X}_{\mathfrak{s}}\boldsymbol{\lambda}\right) &= \mathbf{X}_{\mathscr{U}}^{\top}\mathbf{1} \\
\mathbf{X}_{\mathfrak{s}}^{\top}\mathbf{d} + \mathbf{X}_{\mathfrak{s}}^{\top}\left(\mathbf{Q}\mathbf{\Pi}^{-1}\right)\mathbf{X}_{\mathfrak{s}}\boldsymbol{\lambda} &= \mathbf{X}_{\mathscr{U}}^{\top}\mathbf{1} \\
\mathbf{X}_{\mathfrak{s}}^{\top}\left(\mathbf{Q}\mathbf{\Pi}^{-1}\right)\mathbf{X}_{\mathfrak{s}}\boldsymbol{\lambda} &= \mathbf{X}_{\mathscr{U}}^{\top}\mathbf{1} - \mathbf{X}_{\mathfrak{s}}^{\top}\mathbf{d} \\
\boldsymbol{\lambda} &= \left(\mathbf{X}_{\mathfrak{s}}^{\top}\left(\mathbf{Q}\mathbf{\Pi}^{-1}\right)\mathbf{X}_{\mathfrak{s}}\right)^{-1}\left(\mathbf{X}_{\mathscr{U}}^{\top}\mathbf{1} - \mathbf{X}_{\mathfrak{s}}^{\top}\mathbf{d}\right) \\
&= \left(\mathbf{X}_{\mathfrak{s}}^{\top}\left(\mathbf{Q}\mathbf{\Pi}^{-1}\right)\mathbf{X}_{\mathfrak{s}}\right)^{-1}\left(\mathbf{t}_{\mathbf{x}\mathscr{U}} - \hat{\mathbf{t}}_{\mathbf{x}\mathfrak{s}}\right).
\end{aligned}
$$

Now, substituting $\boldsymbol{\lambda}$ into the function for $\mathbf{w}^{gr}$ gives

$$
\begin{aligned}
\mathbf{w}^{gr} &= \mathbf{d} + \left(\mathbf{Q}\mathbf{\Pi}^{-1}\right)\mathbf{X}_{\mathfrak{s}}\boldsymbol{\lambda} \\
&= \mathbf{d} + \left(\mathbf{Q}\mathbf{\Pi}^{-1}\right)\mathbf{X}_{\mathfrak{s}}\left(\mathbf{X}_{\mathfrak{s}}^{\top}\left(\mathbf{Q}\mathbf{\Pi}^{-1}\right)\mathbf{X}_{\mathfrak{s}}\right)^{-1}\left(\mathbf{t}_{\mathbf{x}\mathscr{U}} - \hat{\mathbf{t}}_{\mathbf{x}\mathfrak{s}}\right).
\end{aligned}
$$

Assuming that $\mathbf{\Pi}$ and $\mathbf{Q}^{-1}$ are invertible and commutable, we can write our estimator in

the following ways

$$
\begin{aligned}
\hat{\boldsymbol{t}}_y^{gr} &= \mathbf{y}^\top \mathbf{w}^{gr} \\
&= \mathbf{y}^\top \left[\mathbf{d} + \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \mathbf{X}_{\mathfrak{s}} \left(\mathbf{X}_{\mathfrak{s}}^\top \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \mathbf{X}_{\mathfrak{s}} \right)^{-1} \left(\mathbf{t}_{\mathbf{x}\mathscr{U}} - \hat{\mathbf{t}}_{\mathbf{x}\mathfrak{s}} \right) \right] \\
&= \mathbf{y}^\top \left[\boldsymbol{\Pi}^{-1}\mathbf{1} + (\boldsymbol{\Pi})^{-1} \mathbf{Q}\mathbf{X}_{\mathfrak{s}} \left(\mathbf{X}_{\mathfrak{s}}^\top \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \mathbf{X}_{\mathfrak{s}} \right)^{-1} \left(\mathbf{t}_{\mathbf{x}\mathscr{U}} - \hat{\mathbf{t}}_{\mathbf{x}\mathfrak{s}} \right) \right] \\
&= \mathbf{y}^\top \boldsymbol{\Pi}^{-1} \left[\mathbf{1} + \mathbf{Q}\mathbf{X}_{\mathfrak{s}} \left(\mathbf{X}_{\mathfrak{s}}^\top \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \mathbf{X}_{\mathfrak{s}} \right)^{-1} \left(\mathbf{t}_{\mathbf{x}\mathscr{U}} - \hat{\mathbf{t}}_{\mathbf{x}\mathfrak{s}} \right) \right] \\
&= \mathbf{y}^\top \boldsymbol{\Pi}^{-1} \mathbf{g}
\end{aligned}
$$

where

$$
\underset{n\times 1}{\mathbf{g}} = \mathbf{1} + \mathbf{Q}\mathbf{X}_{\mathfrak{s}} \left(\mathbf{X}_{\mathfrak{s}}^\top \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \mathbf{X}_{\mathfrak{s}} \right)^{-1} \left(\mathbf{t}_{\mathbf{x}\mathscr{U}} - \hat{\mathbf{t}}_{\mathbf{x}\mathfrak{s}} \right).
$$

Moreover, we can also write our estimator as

$$
\begin{aligned}
\widehat{\boldsymbol{t}}_y^{gr} &= \mathbf{y}^\top \left[\mathbf{d} + \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \mathbf{X}_{\mathfrak{s}} \left(\mathbf{X}_{\mathfrak{s}}^\top \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \mathbf{X}_{\mathfrak{s}} \right)^{-1} \left(\mathbf{t}_{\mathbf{x}\mathscr{U}} - \hat{\mathbf{t}}_{\mathbf{x}\mathfrak{s}} \right) \right] \\
&= \mathbf{y}^\top \mathbf{d} + \mathbf{y}^\top \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \mathbf{X}_{\mathfrak{s}} \left(\mathbf{X}_{\mathfrak{s}}^\top \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \mathbf{X}_{\mathfrak{s}} \right)^{-1} \left(\mathbf{t}_{\mathbf{x}\mathscr{U}} - \hat{\mathbf{t}}_{\mathbf{x}\mathfrak{s}} \right) \\
&= \widehat{\mathbf{t}}_y + \widehat{\mathbf{B}}_{\mathbf{yx}} \left(\mathbf{t}_{\mathbf{x}\mathscr{U}} - \hat{\mathbf{t}}_{\mathbf{x}\mathfrak{s}} \right)
\end{aligned}
$$

where

$$
\underset{C\times p}{\widehat{\mathbf{B}}_{\mathbf{yx}}} = \mathbf{y}^\top \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \mathbf{X}_{\mathfrak{s}} \left(\mathbf{X}_{\mathfrak{s}}^\top \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \mathbf{X}_{\mathfrak{s}} \right)^{-1}.
$$

Here, $\mathbf{B}_{\mathbf{yx}}$ is a matrix of finite population parameter slope coefficients obtained from regressing x on y using ordinary least squares with all units in the population. Furthermore, $\widehat{\mathbf{B}}_{\mathbf{yx}}$ is a weighted estimate of $\mathbf{B}_{\mathbf{yx}}$ using sample values.

B.4 LGREG

B.4.1 Design Consistency of the Clustered LGREG Estimator

By the mean value theorem, there is a point, $\mathbf{B}^{\star}$, such that

$$\boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}_N\right) + \left[\left.\frac{\partial \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{t}\right)}{\partial \mathbf{t}}\right|_{\mathbf{t}=\mathbf{B}^{\star}}\right]^{\top}\left(\hat{\mathbf{B}} - \mathbf{B}_N\right)$$

Summing and dividing by N gives,

$$N^{-1}\sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = N^{-1}\sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}_N\right) + N^{-1}\sum_{\mathscr{U}}\left[\frac{\partial \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{t}\right)}{\partial \mathbf{t}}|\mathbf{t} = \mathbf{B}^{\star}\right]^{\top}\left(\hat{\mathbf{B}} - \mathbf{B}_N\right).$$

To show consistency, we must make some assumptions. Here, we borrow two assumptions from Wu and Sitter (2001). First, we assume that our derivative is locally continuous around $\mathbf{B}_N$ and that our derivative is bounded as the sample and population grow. This is a fairly mild regularity condition which is Assumption 5 in Section 3.2.1.1 on page 134.

Second, we assume that our estimator of $\mathbf{B}_N$ is consistent. Most standard estimation techniques, including those recommended in this thesis, share this property. We write this assumption as Assumption 4 in Section 3.2.1.1 on page 133.

Under Assumption 4 in Section 3.2.1.1, $N^{-1}\sum_{\mathscr{U}}\left[\frac{\partial \boldsymbol{\mu}(\mathbf{x}_k, \mathbf{t})}{\partial \mathbf{t}}\Big|_{\mathbf{t}=\mathbf{B}^{\star}}\right]^{\top} = O\left(1\right)$. Also, under Assumption 5 in Section 3.2.1.1, we see that $\left(\hat{\mathbf{B}} - \mathbf{B}_N\right) = O_p\left(n^{-\frac{1}{2}}\right)$. Thus

$$N^{-1}\sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = N^{-1}\sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}_N\right) + O_p\left(n^{-\frac{1}{2}}\right).$$

So, as the sample size gets large, $N^{-1}\sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right)$ converges to the true population quantity $N^{-1}\sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}_N\right)$.

Now, we consider weighted totals. The weighted average for our initial expression is,

$$N^{-1}\sum_{\mathfrak{s}} d_k \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = N^{-1}\sum_{\mathfrak{s}} d_k \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}_N\right) + N^{-1}\sum_{\mathfrak{s}} d_k \left[\frac{\partial \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{t}\right)}{\partial \mathbf{t}}\bigg|_{\mathbf{t}=\mathbf{B}^\star}\right]^\top \left(\hat{\mathbf{B}} - \mathbf{B}_N\right).$$

Under Assumption 14 in Appendix A.1 on page 258, our weights are roughly $O\left(\frac{N}{n}\right)$ in large samples. Summing these weights for the sample gives $N^{-1}\sum_s \frac{N}{n} = O\left(N^{-1} n \frac{N}{n}\right) = O\left(1\right)$. Under Assumptions 4 and 5 in Section 3.2.1.1 as well as the assumption that none of our weights dominate, we see that

$$N^{-1}\sum_{\mathfrak{s}} d_k \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = N^{-1}\sum_{\mathfrak{s}} d_k \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}_N\right) + O_p\left(n^{-\frac{1}{2}}\right).$$

So, as the sample size gets large, $N^{-1}\sum_{\mathfrak{s}} d_k \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right)$ converges to the estimated value when the true coefficients are used.

Using the asymptotic normality of Assumption 6 in Section 3.2.1.1 on page 134,

$$N^{-1}\sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}_N\right) - N^{-1}\sum_{\mathfrak{s}} d_k \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}_N\right) = O_p\left(n^{-\frac{1}{2}}\right).$$

To summarize, we showed that

$$N^{-1}\sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = N^{-1}\sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}_N\right) + O_p\left(n^{-\frac{1}{2}}\right)$$

$$N^{-1}\sum_{\mathfrak{s}} d_k \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = N^{-1}\sum_{\mathfrak{s}} d_k \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}_N\right) + O_p\left(n^{-\frac{1}{2}}\right)$$

$$N^{-1}\sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}_N\right) - N^{-1}\sum_{\mathfrak{s}} d_k \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}_N\right) = O_p\left(n^{-\frac{1}{2}}\right).$$

Together these three equations imply that

$$N^{-1}\sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) - N^{-1}\sum_{\mathfrak{s}} d_k \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = O_p\left(n^{-\frac{1}{2}}\right)$$

In terms of matrices, we have

$$\left(\hat{\boldsymbol{\mu}}_{\mathscr{U}}^{\top}\mathbf{1} - \hat{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\mathbf{d}\right) = O_p\left(n^{-\frac{1}{2}}\right).$$

where $\hat{\boldsymbol{\mu}}_{\mathscr{U}}$ is the matrix of predictions based on $\boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right)$ for all elements of the population and $\hat{\boldsymbol{\mu}}_{\mathfrak{s}}$ is the corresponding matrix for all sample units.

Now, consider our estimator,

$$\begin{aligned}
\widehat{\mathbf{t}}_y^{lg} &= \sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) + \sum_{\mathfrak{s}} d_k\left[\mathbf{y}_k - \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right)\right] \\
&= \sum_{\mathfrak{s}} d_k\mathbf{y}_k + \sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) - \sum_{\mathfrak{s}} \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) \\
&= \sum_{\mathfrak{s}} d_k\mathbf{y}_k + O_p\left(n^{-\frac{1}{2}}\right) \\
&= \widehat{t}_y + O_p\left(n^{-\frac{1}{2}}\right).
\end{aligned}$$

Recall that $\widehat{\mathbf{t}}_y = \mathbf{t}_y + O_p\left(n^{-\frac{1}{2}}\right)$. Thus

$$\widehat{\mathbf{t}}_y^{lg} = \mathbf{t}_y + O_p\left(n^{-\frac{1}{2}}\right).$$

Since $\widehat{\mathbf{t}}_y^{lg} = \mathbf{t}_y + O_p\left(n^{-\frac{1}{2}}\right)$, we conclude that $\widehat{\mathbf{t}}_y^{lg}$ is a consistent estimator of $\mathbf{t}_y$. Furthermore, it is asymptotically centered around the Horvitz-Thompson estimator, an unbiased estimator.

B.4.2 Asymptotic variance of LGREG estimator

Introduction

Wu and Sitter (2001) argue that the asymptotic variance of the LGREG can be obtained by repeating their proof of the asymptotic variance of the model calibration

estimator with $\widehat{\mathbf{B}} = \mathbf{1}$ and $\widehat{\mathbf{B}}_N = \mathbf{1}$. Here we follow that general outline by extending our derivation of the asymptotic variance of the model calibration estimator for multinomial logistic regression in Appendix B.5.4 on page 342.

We show that the asymptotic variance of the LGREG for a multinomial response in two staged samples is

$$\operatorname{av}\left(\hat{\mathbf{t}}_y^{lg}\right) = \sum \sum_{\mathscr{U}_I} \Delta_{ij} \frac{t_{e_i}}{\pi_{Ii}} \frac{t_{e_j}}{\pi_{Ij}} + \sum_{\mathscr{U}_I} \frac{\sum_{\mathscr{U}_i} \sum_{\mathscr{U}_i} \Delta_{kl|i} \frac{e_{k|i}}{\pi_{k|i}} \frac{e_{l|i}}{\pi_{l|i}}}{\pi_{Ii}}.$$

Our proof is identical to that shown in Wu and Sitter (2001) with a few minor exceptions. Rather than assuming a generalized linear model, we explicitly use a multinomial logistic model. This means that our dependent variable and coefficients are matrices, and our model has an explicit form. Furthermore, whereas Wu and Sitter (2001) provide results for single stage sampling, we derive the variance when samples were selected in two stages.

Proof

A second order Taylor series expansion of $\mu\left(\mathbf{x}, \widehat{\mathbf{B}}\right)$ at $\widehat{\mathbf{B}} = \mathbf{B}_N$ is

$$\begin{aligned}\boldsymbol{\mu}\left(\mathbf{x}_k, \widehat{\mathbf{B}}\right) &= \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}_N\right) + \left[\left.\frac{\partial \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{t}\right)}{\partial \mathbf{t}}\right|_{\mathbf{t}=\mathbf{B}_N}\right]^\top \left(\widehat{\mathbf{B}} - \mathbf{B}_N\right) \\ &\quad + \left(\hat{\mathbf{B}} - \mathbf{B}_N\right)^\top \left[\left.\frac{\partial^2 \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{t}\right)}{\partial \mathbf{t} \partial \mathbf{t}^\top}\right|_{\mathbf{t}=\mathbf{B}^\star}\right]^\top \left(\widehat{\mathbf{B}} - \mathbf{B}_N\right)\end{aligned}$$

Summing and dividing by N gives,

$$\begin{aligned}\frac{1}{N} \sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) &= \frac{1}{N} \sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}_N\right) + \frac{1}{N} \sum_{\mathscr{U}} \left[\left.\frac{\partial \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{t}\right)}{\partial \mathbf{t}}\right|_{\mathbf{t}=\mathbf{B}_N}\right]^\top \left(\widehat{\mathbf{B}} - \mathbf{B}_N\right) \\ &\quad + \frac{1}{N} \sum_{\mathscr{U}} \left(\widehat{\mathbf{B}} - \mathbf{B}_N\right)^\top \left[\left.\frac{\partial^2 \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{t}\right)}{\partial \mathbf{t} \partial \mathbf{t}^\top}\right|_{\mathbf{t}=\mathbf{B}^\star}\right]^\top \left(\widehat{\mathbf{B}} - \mathbf{B}_N\right).\end{aligned}$$

where $\mathbf{B}^\star$ is a point between $\hat{\mathbf{B}}$ and $\mathbf{B}_N$.

Now, we borrow another assumption from Wu and Sitter (2001). We assume that our second derivative is locally continuous around $\mathbf{B}_N$ and that our second derivative is bounded as the sample and population grow. This is a fairly mild regularity condition. Now, by Assumptions 5 and 7 in Section 3.2.1.1, we see that

$$\frac{1}{N}\sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = \frac{1}{N}\sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}_N\right) + \frac{1}{N}\sum_{\mathscr{U}} \left[\frac{\partial \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{t}\right)}{\partial \mathbf{t}}\bigg|_{\mathbf{t}=\mathbf{B}_N}\right]^\top \left(\hat{\mathbf{B}} - \mathbf{B}_N\right) + O_p\left(n^{-1}\right)$$

Renaming the first derivative gives

$$\frac{1}{N}\sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = \frac{1}{N}\sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}_N\right) + \frac{1}{N}\sum_{\mathscr{U}} \left[\mathbf{K}\left(\mathbf{x}_k, \mathbf{B}_N\right)\right]^\top \left(\hat{\mathbf{B}} - \mathbf{B}_N\right) + O_p\left(n^{-1}\right).$$

Furthermore, by our assumptions about the sample design

$$\frac{1}{N}\sum_{\mathfrak{s}} d_k\boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = \frac{1}{N}\sum_{\mathfrak{s}} d_k\boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}_N\right) + \frac{1}{N}\sum_{\mathfrak{s}} \left[d_k\mathbf{K}\left(\mathbf{x}_i, \mathbf{B}_N\right)\right]^\top \left(\hat{\mathbf{B}} - \mathbf{B}_N\right) + O_p\left(n^{-1}\right).$$

By Assumptions 4, 5, and 6 in Section 3.2.1.1, we have $\hat{\mathbf{B}} - \mathbf{B}_N = O\left(n^{-\frac{1}{2}}\right)$ and $\frac{1}{N}\sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}_N\right) - \frac{1}{N}\sum_{\mathscr{U}} d_k\boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}_N\right) = O_p\left(n^{-\frac{1}{2}}\right)$. Therefore,

$$\frac{1}{N}\sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) - \frac{1}{N}\sum_{s} d_k\boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = \frac{1}{N}\sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}\right) - \frac{1}{N}\sum_{s} d_k\boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}\right) + O_p\left(n^{-1}\right).$$

Now, consider our estimator,

$$\widehat{\mathbf{t}}_y^{lg} = \widehat{\boldsymbol{t}}_y + \widehat{\boldsymbol{\mu}}_{\mathscr{U}}^\top \mathbf{1} - \widehat{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \mathbf{d}$$

Replacing $\widehat{\boldsymbol{\mu}}$ with $\boldsymbol{\mu}$ gives

$$\begin{aligned}
\widehat{\mathbf{t}}_y^{lg} &= \hat{\boldsymbol{t}}_y - \boldsymbol{\mu}_{\mathfrak{s}}^\top \mathbf{d} + \boldsymbol{\mu}_{\mathscr{U}}^\top \mathbf{1} + o_p\left(n^{-\frac{1}{2}}\right) \\
&= \left[\mathbf{y}^\top - \boldsymbol{\mu}_{\mathfrak{s}}^\top\right] \mathbf{d} + \boldsymbol{\mu}_{\mathscr{U}}^\top \mathbf{1} + o_p\left(n^{-\frac{1}{2}}\right) \\
&= \mathbf{e}^\top \mathbf{d} + \boldsymbol{\mu}_{\mathscr{U}}^\top \mathbf{1} + o_p\left(n^{-\frac{1}{2}}\right)
\end{aligned}$$

where $\mathbf{e} = \mathbf{y} - \boldsymbol{\mu}_{\mathfrak{s}}$.

Thus, the asymptotic variance of $\widehat{\mathbf{t}}_y^{lg}$ will be equivalent to the asymptotic variance of a weighted sum of $\mathbf{e}_k$. In this way, we can use the formulas for the variance of the Horvitz-Thompson estimator in clustered designs by substituting $\mathbf{y}_{k*}$ with $\mathbf{e}_k$. For invariant and independent sample designs, the asymptotic variance is

$$\begin{aligned}
\text{av}\left(\hat{\mathbf{t}}_y^{lg}\right) &= \text{var}\left(\sum_{i\in\mathfrak{s}_I}\left(d_i\widehat{\mathbf{t}}_{ei}\right)\right) \\
&= \text{var}\left[\text{E}\left[\sum_{i\in\mathfrak{s}_I}\left(d_i\widehat{\mathbf{t}}_{ei}\right)|\mathfrak{s}_I\right]\mathfrak{s}_i\right] + \text{E}\left[\text{var}\left[\sum_{i\in\mathfrak{s}_i}\left(d_i\widehat{\mathbf{t}}_{ei}\right)|\mathfrak{s}_I\right]|\mathfrak{s}_i\right] \\
&= \sum_{i\in\mathfrak{s}_I}\text{var}\left(d_i\mathbf{t}_{ei}|\mathfrak{s}_i\right) + \text{E}\left[\sum_{i\in\mathfrak{s}_I}\left(d_i^2\text{var}\left(\sum_{k\in\mathfrak{s}_i}d_{k|i}\mathbf{e}_{k|i}|\mathfrak{s}_I\right)\right)|\mathfrak{s}_i\right] \\
&= \sum_{i\in\mathscr{U}_I}\sum_{j\in\mathscr{U}_I}\left(\Delta_{ij}d_id_j\mathbf{t}_{ei}\mathbf{t}_{ej}^\top\right) + \text{E}\left[\sum_{i\in\mathfrak{s}_I}\left(d_i^2\left(\sum_{k\in\mathscr{U}_i}\sum_{l\in\mathscr{U}_i}\Delta_{kl|i}d_{k|i}d_{l|i}\mathbf{e}_{k|i}\mathbf{e}_{l|i}^\top\right)\right)|\mathfrak{s}_i\right] \\
&= \sum_{i\in\mathscr{U}_I}\sum_{j\in\mathscr{U}_I}\left(\Delta_{ij}d_id_j\mathbf{t}_{ei}\mathbf{t}_{ej}^\top\right) + \sum_{i\in\mathscr{U}_I}\left[d_i\left(\sum_{k\in\mathscr{U}_i}\sum_{l\in\mathscr{U}_i}\Delta_{kl|i}d_{k|i}d_{l|i}\mathbf{e}_k\mathbf{e}_l^\top\right)\right] \qquad \text{(B.16)}
\end{aligned}$$

where

$$\mathbf{t}_{ei} = \sum_{k\in\mathscr{U}_i}\mathbf{e}_k. \qquad \text{(B.17)}$$

For category c, we write the variance as

$$\text{av}_{II}\left(\hat{\mathbf{t}}_{yc}^{lg}\right) = \sum\sum_{\mathscr{U}_I}\Delta_{ij}\frac{t_{e_{ci}}}{\pi_i}\frac{t_{e_{cj}}}{\pi_j} + \sum_{\mathscr{U}_I}\frac{\sum_{\mathscr{U}_i}\sum_{\mathscr{U}_i}\Delta_{kl|i}\frac{e_{ck}}{\pi_{k|i}}\frac{e_{cl}}{\pi_{l|i}}}{\pi_i}.$$

B.4.3 Variance estimators of LGREG

B.4.3.1 Linear substitute estimator

Equation (B.16) on page 322 shows the asymptotic variance of the LGREG. A naive variance estimator of the asymptotic variance of $\hat{\mathbf{t}}_y^{lg}$ can be formed by replacing

$\mathbf{e}_k = \mathbf{y}_k - \boldsymbol{\mu}_k$ in the asymptotic variance with the sample estimator $\widehat{\mathbf{e}}_k - \mathbf{y}_k - \widehat{\boldsymbol{\mu}}_k$ and replacing unknown population sums with weighed sample sums. Doing so, gives the linear substitute variance estimator

$$\upsilon_e\left(\hat{\mathbf{t}}_y^{lg}\right) = \sum_{i\in\mathfrak{s}_I}\sum_{j\in\mathfrak{s}_I}\left[\frac{\Delta_{ij}}{\pi_{ij}}d_i d_j \widehat{\mathbf{t}}_{\widehat{e}i}^{\pi}\left(\widehat{\mathbf{t}}_{\widehat{e}j}^{\pi}\right)^{\top}\right] + \sum_{i\in\mathfrak{s}_I}\left[d_i\left(\sum_{k\in\mathfrak{s}_i}\sum_{l\in\mathfrak{s}_i}\frac{\Delta_{kl|i}}{\pi_{kl|i}}d_{k|i}d_{l|i}\widehat{\mathbf{e}}_k\widehat{\mathbf{e}}_l^{\top}\right)\right]$$

where

$$\widehat{\mathbf{t}}_{\widehat{e}i}^{\pi} = \sum_{k\in\mathfrak{s}_i} d_k\widehat{\mathbf{e}}_k \tag{B.18}$$

$$\widehat{\mathbf{e}}_k = \mathbf{y}_k - \widehat{\boldsymbol{\mu}}_k. \tag{B.19}$$

For category c, the estimator is

$$\upsilon_e\left(\hat{t}_{yc}^{lg}\right) = \sum\sum_{\mathfrak{s}_I}\frac{\Delta_{ij}}{\pi_{ij}}\frac{\widehat{t}_{\widehat{e}ci}^{\pi}}{\pi_i}\frac{\widehat{t}_{\widehat{e}cj}^{\pi}}{\pi_j} + \sum_{\mathfrak{s}_I}\frac{\sum_{\mathfrak{s}_i}\sum_{\mathfrak{s}_i}\frac{\Delta_{kl|i}}{\pi_{kl|i}}\frac{\widehat{e}_{ck}}{\pi_{k|i}}\frac{\widehat{e}_{cl}}{\pi_{l|i}}}{\pi_i}$$

where

$$\widehat{t}_{\widehat{e}ci}^{\pi} = \sum_{k\in\mathfrak{s}_i} d_k\widehat{e}_{ck}$$

$$\widehat{e}_{ck} = y_{ck} - \widehat{\mu}_{ck}.$$

If the first and second stage samples are selected using a Poisson sampling technique, then υ_e reduces to

$$\sum_{i\in\mathfrak{s}_I}\left[\frac{(1-\pi_i)}{\pi_i^2}\widehat{\mathbf{t}}_{ei}\widehat{\mathbf{t}}_{ei}^{\top}\right] + \sum_{i\in\mathfrak{s}_I}\left[d_i\left(\sum_{k\in\mathfrak{s}_i}\sum_{l\in\mathfrak{s}_i}\frac{\left(1-\pi_{k|i}\right)}{\pi_{k|i}^2}\widehat{\mathbf{e}}_k\widehat{\mathbf{e}}_k^{\top}\right)\right]$$

For category c, the estimator is

$$\sum_{i\in\mathfrak{s}_I}\frac{(1-\pi_i)}{\pi_i^2}\hat{t}_{eci}^2 + \sum_{i\in\mathfrak{s}_I}\frac{1}{\pi_i}\sum_{k\in\mathfrak{s}_i}\frac{\left(1-\pi_{k|i}\right)}{\pi_{k|i}^2}e_{ck}^2.$$

Whereas υ_e generally requires knowledge of the first and second-stage joint inclusion probabilities; in Poisson samples, it does not. Furthermore, some simplicity is gained in Poisson samples because the double summation reduces to a single summation.

B.4.3.2 With-replacement estimator

Commonly, with-replacement variance estimators are used even when the first stage sample is selected without-replacement. As long as the sampling fraction is relatively small, the bias of using a with-replacement variance estimator is relatively small. Furthermore, any bias in the with-replacement variance estimator tends to be positive, thus making the with-replacement variance estimator conservative. Särndal et al. (1992, sec 4.6) discuss the classic with-replacement variance estimator of a total and provide some limitations for using the with-replacement variance estimator for samples selected without-replacement.

We begin with the with-replacement variance estimator for the Horvitz-Thompson estimator in a clustered design

$$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{\pi}\right) = \frac{1}{n\left(n-1\right)}\sum_{i\in\mathfrak{s}_I}^{n}\left(\frac{\widehat{\mathbf{t}}_{yi}^{\pi}}{p_i} - \widehat{\mathbf{t}}_y^{\pi}\right)^2 \tag{B.20}$$

where $p_i = \frac{\pi_i}{n}$ is the probability of drawing the ith primary sampling unit in single draw, n is the total number of primary sampling sample units, and $\widehat{\mathbf{t}}_{yi}^{\pi} = \sum_{k\in\mathfrak{s}_i}\frac{y_{k|i}}{\pi_{k|i}}$.

We can modify Equation (B.20) for the LGREG by replacing $\frac{\widehat{\mathbf{t}}_{yi}^{\pi}}{p_i}$ with $\widehat{\mathbf{t}}_{yi}^{lg} = \sum_{k\in\mathscr{U}}\widehat{\boldsymbol{\mu}}_k + \sum_{k\in\mathfrak{s}_i}\left(\frac{d_{k|i}\widehat{\mathbf{e}}_k}{p_i}\right)$ and $\widehat{\mathbf{t}}_y^{\pi}$ with $\widehat{\mathbf{t}}_y^{lg} = \sum_{k\in\mathscr{U}}\widehat{\boldsymbol{\mu}}_k + \sum_{k\in\mathfrak{s}} d_k\widehat{\mathbf{e}}_k$. These two substitutions are

motivated by Equation (3.20) on page 132. Now consider

$$\begin{aligned}
\widehat{\mathbf{t}}_{yi}^{lg} - \widehat{\mathbf{t}}_{y}^{lg} &= \sum_{k\in\mathscr{U}} \widehat{\boldsymbol{\mu}}_k + \sum_{k\in\mathfrak{s}_i} \left(\frac{d_{k|i}\widehat{\mathbf{e}}_k}{p_i}\right) - \left[\sum_{k\in\mathscr{U}} \widehat{\boldsymbol{\mu}}_k + \sum_{k\in\mathfrak{s}} d_k\widehat{\mathbf{e}}_k\right] \\
&= \sum_{k\in\mathfrak{s}_i} nd_i d_{k|i}\widehat{\mathbf{e}}_k - \sum_{k\in\mathfrak{s}} d_k\widehat{\mathbf{e}}_k \\
&= nd_i \sum_{k\in\mathfrak{s}_i} d_{k|i}\widehat{\mathbf{e}}_k - \mathbf{t}_{\widehat{\mathbf{e}}}^{\pi}
\end{aligned}$$

Thus, following the logic in Särndal et al. (1992, sec 4.6), the with-replacement variance estimator is

$$\begin{aligned}
v_{wr}\left(\mathbf{t}_y^{lg}\right) &= \frac{1}{n\left(n-1\right)} \sum_{i\in\mathfrak{s}_I}^{n} \left(nd_i \sum_{k\in\mathfrak{s}_i} \left(d_{k|i}\widehat{\mathbf{e}}_k\right) - \mathbf{t}_{\widehat{\mathbf{e}}}^{\pi}\right) \left(nd_i \sum_{k\in\mathfrak{s}_i} \left(d_{k|i}\widehat{\mathbf{e}}_k\right) - \mathbf{t}_{\widehat{\mathbf{e}}}^{\pi}\right)^{\top} \\
&= \frac{n}{n^2\left(n-1\right)} \sum_{i\in\mathfrak{s}_I}^{n} \left(nd_i \sum_{k\in\mathfrak{s}_i} \left(d_{k|i}\widehat{\mathbf{e}}_k\right) - \mathbf{t}_{\widehat{\mathbf{e}}}^{\pi}\right) \left(nd_i \sum_{k\in\mathfrak{s}_i} \left(d_{k|i}\widehat{\mathbf{e}}_k\right) - \mathbf{t}_{\widehat{\mathbf{e}}}^{\pi}\right)^{\top} \\
&= \frac{n}{\left(n-1\right)} \sum_{i\in\mathfrak{s}_I}^{n} \left(d_i \sum_{k\in\mathfrak{s}_i} \left(d_{k|i}\widehat{\mathbf{e}}_k\right) - \frac{1}{n}\mathbf{t}_{\widehat{\mathbf{e}}}^{\pi}\right) \left(d_i \sum_{k\in\mathfrak{s}_i} \left(d_{k|i}\widehat{\mathbf{e}}_k\right) - \frac{1}{n}\mathbf{t}_{\widehat{\mathbf{e}}}^{\pi}\right)^{\top} \\
&= \frac{n}{\left(n-1\right)} \sum_{i\in\mathfrak{s}_I}^{n} \left(d_i\widehat{\mathbf{t}}_{\widehat{e}i}^{\pi} - \frac{1}{n}\mathbf{t}_{\widehat{\mathbf{e}}}^{\pi}\right) \left(d_i\widehat{\mathbf{t}}_{\widehat{e}i}^{\pi} - \frac{1}{n}\mathbf{t}_{\widehat{\mathbf{e}}}^{\pi}\right)^{\top} \qquad \text{(B.21)}
\end{aligned}$$

where

$$\widehat{\mathbf{t}}_{\widehat{\mathbf{e}}}^{\pi} = \sum_{k\in\mathfrak{s}} \left(d_k\widehat{\mathbf{e}}_k\right) \qquad \text{(B.22)}$$

and $\widehat{\mathbf{t}}_{\widehat{\mathbf{e}}i}^{\pi}$ is defined in Equation (B.18) and $\widehat{\mathbf{e}}_k$ is defined in Equation (B.19) on page 323.

B.4.3.3 Implicit differentiation variance estimator

Introduction

An additional alternative variance estimator uses implicit differentiation. First described by Binder (1983), implicit differentiation uses linearization and estimating equa-

tions to produce design-consistent estimators of finite population parameters. Implicit differentiation is especially useful when the parameter of interest cannot be solved explicitly in closed form. Both Binder (1983) and Särndal et al. (1992)[section 13.4] give several examples of how implicit differentiation can be used to construct design-consistent variance estimators of $\mathbf{B}$ from a logistic regression model. However, neither discuss estimating the variance of logistic assisted estimators of totals. An advantage of implicit differentiation is that variance estimators can easily be computed from the estimating equations.

An understanding of estimating equations is essential to developing implicit differentiation variance estimators. In the method that follows, the population parameter vector is slightly different from other parameter vectors that form the basis of this variance estimation method. For this reason, the first section presents the population parameter vector. Then, population and survey weighted estimating equations are constructed. The survey weighted estimating equations are used to estimate $\mathbf{t}_y^{lg}$. After defining the estimating equations, the implicit differentiation method is used to derive the asymptotic variance of $\mathbf{t}_y^{lg}$. An estimator of this asymptotic variance requires a rather complex differentiation. We end with the components of this differentiation.

Parameters

For the multivariate LGREG with the logit link, we begin by defining a vector with our parameters of interest, called $\boldsymbol{\theta}$. This vector contains both $\mathbf{t}_y^{lg}$ and $\mathrm{vec}\,(\mathbf{B})$, which are the parameters we would obtain if we had a complete census with neither sampling nor

nonsampling errors. That is,

$$\underset{(C+(C-1)\cdot p)\times 1}{\boldsymbol{\theta}} = \begin{bmatrix} \underset{C\times 1}{\mathbf{t}_y^{lg}} \\ \\ \underset{(C-1)\cdot p\times 1}{\mathrm{vec}\left(\mathbf{B}\right)} \end{bmatrix}.$$

Unless the complete population is measured without error, our parameter vector is unknown. We denote an estimate of our parameter vector as

$$\widehat{\boldsymbol{\theta}} = \begin{bmatrix} \widehat{\mathbf{t}}_y^{lg} \\ \\ \mathrm{vec}\left(\widehat{\mathbf{B}}\right) \end{bmatrix}.$$

$\widehat{\mathbf{t}}_y^{lg}$ is defined as

$$\begin{aligned} \widehat{\mathbf{t}}_y^{lg} &= \sum_{\mathscr{U}} \frac{z_k e^{\mathbf{X}_k^\top \mathrm{vec}\left(\widehat{\mathbf{B}}\right)}}{1+\sum_{c=1}^{C-1} e^{\widehat{\mathbf{B}}_c^\top \mathbf{x}_k}} + \sum_{\mathfrak{s}} d_k \left[\mathbf{y}_k - \frac{z_k e^{\mathbf{X}_k^\top \mathrm{vec}\left(\widehat{\mathbf{B}}\right)}}{1+\sum_{c=1}^{C-1} e^{\widehat{\mathbf{B}}_c^\top \mathbf{x}_k}}\right] \\ &= \sum_{\mathscr{U}} \widehat{\boldsymbol{\mu}}_k + \sum_{\mathfrak{s}} d_k \left[\mathbf{y}_k - \widehat{\boldsymbol{\mu}}_k\right] \end{aligned}$$

where

$$\begin{aligned} \underset{C\times 1}{\widehat{\boldsymbol{\mu}}_k} &= \frac{z_k e^{\mathbf{X}_k^\top \mathrm{vec}\left(\widehat{\mathbf{B}}\right)}}{1+\sum_{c=1}^{C-1} e^{\widehat{\mathbf{B}}_c^\top \mathbf{x}_k}} \\ \widehat{\mu}_{ck} &= \frac{z_k e^{\widehat{\mathbf{B}}_c^\top \mathbf{x}_k}}{1+\sum_{c=1}^{C-1} e^{\widehat{\mathbf{B}}_c^\top \mathbf{x}_k}} \end{aligned}$$

and $\mathbf{X}_k$ is defined in Equation 3.5. That is

$$\underset{C\times(C-1)\cdot p}{\mathbf{X}_k} = \begin{bmatrix} \mathbf{x}_k^\top & & & \\ & \mathbf{x}_k^\top & & \\ & & \ddots & \\ & & & \mathbf{x}_k^\top \\ 0 & & \cdots & 0 \end{bmatrix}$$

and $\widehat{\mathbf{B}}$ is a $p \times (C-1)$ matrix where each column is a vector of coefficients associated with the cth category.

Population Estimating Equations

To motivate the implicit differentiation estimator, we write our estimator when every unit in the finite population is included in the sample. This is the ideal situation when all population quantities are known and a full census of the population is taken. In this case, $\mathscr{U} = \mathfrak{s}$ and $d_k = 1$ for all k and we can write our estimator as

$$\mathbf{t}_y^{lg} = \sum_{\mathscr{U}} \boldsymbol{\mu}_k + \sum_{\mathscr{U}} [\mathbf{y}_k - \boldsymbol{\mu}_k].$$

Of course in this ideal situation, $\mathbf{t}_y^{lg}$ reduces to $\mathbf{t}_y$ and no estimation is necessary. Nevertheless, we use this equation to motivate our sample estimating equations. An estimating equation for $\mathbf{t}_y^{lg}$ is

$$\mathbf{0} = \sum_{\mathscr{U}} \boldsymbol{\mu}_k + \sum_{\mathscr{U}} [\mathbf{y}_k - \boldsymbol{\mu}_k] - \mathbf{t}_y^{lg}.$$

The coefficient vector for category c, denoted $\mathbf{B}_c$, is the solution to

$$\mathbf{0} = \sum_{\mathscr{U}} \mathbf{x}_k \left[y_{ck} - \mu_{ck}\right].$$

The coefficient matrix, $\mathbf{B}$, is the solution to

$$\mathbf{0} = \sum_{\mathscr{U}} \mathbf{x}_k \left[\mathbf{y}_k^\top - \boldsymbol{\mu}_k^\top\right].$$

Thus, our parameter vector is the solution to $\mathbf{W}(\boldsymbol{\theta}) = \mathbf{0}$ where

$$\underset{(C+C\cdot p-p)\times 1}{\mathbf{W}(\boldsymbol{\theta})} = \begin{bmatrix} \sum_{\mathscr{U}} (\mathbf{y}_k - \boldsymbol{\mu}_k) - \left(\mathbf{t}_y^{lg} - \sum_{\mathscr{U}} \boldsymbol{\mu}_k\right) \\ \sum_{\mathscr{U}} \mathbf{x}_k \left[y_{k1} - \mu_{k1}\right] \\ \vdots \\ \sum_{\mathscr{U}} \mathbf{x}_k \left[y_{kC-1} - \mu_{kC-1}\right] \end{bmatrix}$$

which we write as

$$\mathbf{W}(\boldsymbol{\theta}) = \left(\sum_{\mathscr{U}} \mathbf{U}_k\right) - \mathbf{v}$$

where

$$\mathbf{U}_k = \begin{bmatrix} (\mathbf{y}_k - \boldsymbol{\mu}_k) \\ \mathbf{x}_k (y_{k1} - \mu_{k1}) \\ \vdots \\ \mathbf{x}_k (y_{kC-1} - \mu_{kC-1}) \end{bmatrix}$$

$$\mathbf{v} = \begin{bmatrix} \mathbf{t}_y^{lg} - \sum_{\mathscr{U}} \boldsymbol{\mu}_k \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \end{bmatrix}.$$

Survey Weighted Estimating Equations

We only measure $\mathbf{y}_k$ for the sample units. Thus, we cannot compute $\mathbf{W}(\boldsymbol{\theta})$. Yet, we can estimate $\mathbf{W}(\boldsymbol{\theta})$ from our sample. The weighted estimate of our estimating equations

is

$$\underset{(C+C-1\cdot p)\times 1}{\widehat{\mathbf{W}}(\boldsymbol{\theta})} = \left(\sum_{\mathfrak{s}} \widehat{\mathbf{U}}_k(\boldsymbol{\theta})\right) - \mathbf{v}$$
$$= \widehat{\mathbf{U}}(\boldsymbol{\theta}) - \mathbf{v}$$

where

$$\widehat{\mathbf{U}}_k(\boldsymbol{\theta}) = \begin{bmatrix} d_k\left[\mathbf{y}_k - \boldsymbol{\mu}_k\right] \\ d_k\mathbf{x}_k\left(y_{k1} - \mu_{k1}\right) \\ \vdots \\ d_k\mathbf{x}_k\left(y_{kC-1} - \mu_{kC-1}\right) \end{bmatrix}$$

and

$$\widehat{\mathbf{U}}(\boldsymbol{\theta}) = \sum_{\mathfrak{s}} \widehat{\mathbf{U}}_k(\boldsymbol{\theta}).$$

The value of $\boldsymbol{\theta}$ that solves the estimating equations, $\widehat{\mathbf{W}}(\boldsymbol{\theta})$, is denoted $\widehat{\boldsymbol{\theta}}$. Also, let $\boldsymbol{\Sigma}_{\widehat{\mathbf{U}}}(\boldsymbol{\theta})$ be the asymptotic design variance of $\sum_{\mathfrak{s}} \widehat{\mathbf{U}}_k$. That is

$$\underset{(C+C\cdot p-p)\times(C+C\cdot p-p)}{\boldsymbol{\Sigma}_{\widehat{\mathbf{U}}}(\boldsymbol{\theta})} \approx \mathrm{var}\left(\sum_{\mathfrak{s}} \widehat{\mathbf{U}}_k(\boldsymbol{\theta})\right)$$

Derivation of Variance Estimator

Simultaneously solving for $\mathrm{vec}(\mathbf{B})$ and $\mathbf{t}_y^{lg}$ has the advantage that it simplifies variance estimation. Moreover, it results in the complete covariance matrix containing the estimated covariances between $\hat{\mathbf{t}}_y^{lg}$ and $\mathrm{vec}\left(\widehat{\mathbf{B}}\right)$.

We now turn to estimating the variance of $\widehat{\boldsymbol{\theta}}$. Under mild regularity conditions, a linear approximation of our estimating equations is

$$\widehat{\mathbf{W}}\left(\widehat{\boldsymbol{\theta}}\right) \approx \widehat{\mathbf{W}}(\boldsymbol{\theta}) + \widehat{\mathbf{J}}(\boldsymbol{\theta})\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = \mathbf{0}$$

where

$$\underset{(C+C\cdot p-p)\times(C+C\cdot p-p)}{\widehat{\mathbf{J}}\left(\boldsymbol{\theta}\right)} = \frac{\partial}{\partial\left(\mathrm{vec}\boldsymbol{\theta}\right)^{\top}}\widehat{\mathbf{W}}\left(\boldsymbol{\theta}\right).$$

According to Lemma 1 in Binder (1983), we can also write our linearization as

$$\widehat{\mathbf{W}}\left(\widehat{\boldsymbol{\theta}}\right) \approx \widehat{\mathbf{W}}\left(\boldsymbol{\theta}\right) + \mathbf{J}\left(\boldsymbol{\theta}\right)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = \mathbf{0}$$

where

$$\mathbf{J}\left(\boldsymbol{\theta}\right) = \frac{\partial}{\partial\left(\mathrm{vec}\boldsymbol{\theta}\right)^{\top}}\mathbf{W}\left(\boldsymbol{\theta}\right).$$

Using our linearization, we can derive an asymptotic variance estimator

$$\begin{aligned}
\widehat{\mathbf{W}}\left(\widehat{\boldsymbol{\theta}}\right) &\approx \widehat{\mathbf{W}}\left(\boldsymbol{\theta}\right) + \mathbf{J}\left(\boldsymbol{\theta}\right)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = \mathbf{0} \\
-\mathbf{J}\left(\boldsymbol{\theta}\right)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) &\approx \widehat{\mathbf{W}}\left(\boldsymbol{\theta}\right)
\end{aligned}$$

Assuming $\mathbf{J}\left(\boldsymbol{\theta}\right)$ is invertible

$$\begin{aligned}
\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) &\approx -\mathbf{J}^{-1}\left(\boldsymbol{\theta}\right)\widehat{\mathbf{W}}\left(\boldsymbol{\theta}\right) \\
\mathrm{var}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) &\approx \mathbf{J}^{-1}\left(\boldsymbol{\theta}\right)\mathrm{var}\left[\widehat{\mathbf{U}}\left(\boldsymbol{\theta}\right)\right]\left[\mathbf{J}^{-1}\left(\boldsymbol{\theta}\right)\right]^{\top} \\
\mathrm{var}\left(\widehat{\boldsymbol{\theta}}\right) &\approx \mathbf{J}^{-1}\left(\boldsymbol{\theta}\right)\left[\boldsymbol{\Sigma}_{\widehat{\mathbf{U}}}\left(\boldsymbol{\theta}\right)\right]\left[\mathbf{J}^{-1}\left(\boldsymbol{\theta}\right)\right]^{\top}
\end{aligned}$$

Thus, the asymptotic variance of $\widehat{\boldsymbol{\theta}}$ is $\mathbf{J}^{-1}\left(\boldsymbol{\theta}\right)\left[\boldsymbol{\Sigma}_{\widehat{\mathbf{U}}}\left(\boldsymbol{\theta}\right)\right]\left[\mathbf{J}^{-1}\left(\boldsymbol{\theta}\right)\right]^{\top}$, which we denote as $\mathrm{V}\left(\boldsymbol{\theta}\right)$. Because the asymptotic variance of $\widehat{\boldsymbol{\theta}}$ is a function of $\boldsymbol{\theta}$, we write our asymptotic variance as $\mathrm{V}\left(\boldsymbol{\theta}\right)$. We note that $\mathbf{J}\left(\boldsymbol{\theta}\right)$ must be invertible. Furthermore, $\boldsymbol{\Sigma}_{\widehat{\mathbf{U}}}\left(\boldsymbol{\theta}\right) = \mathrm{var}\left[\widehat{\mathbf{U}}\left(\boldsymbol{\theta}\right)\right]$.

We usually do not know $\boldsymbol{\theta}$ nor $\mathbf{J}$; thus, we substitute them for estimated quantities. Moreover, $\widehat{\boldsymbol{\Sigma}}_{\widehat{\mathbf{U}}}\left(\boldsymbol{\theta}\right)$ is an estimate of the design-based variance of $\widehat{\mathbf{U}}\left(\boldsymbol{\theta}\right)$. That is,

$\widehat{\boldsymbol{\Sigma}}_{\widehat{\mathbf{U}}}(\boldsymbol{\theta}) = \upsilon\left[\widehat{\mathbf{U}}(\boldsymbol{\theta})\right]$. So, our final variance estimator is

$$\upsilon_{Binder}\left(\widehat{\boldsymbol{\theta}}\right) = \left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right]\left[\widehat{\boldsymbol{\Sigma}}_{\widehat{\mathbf{U}}}\left(\widehat{\boldsymbol{\theta}}\right)\right]\left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{\top}.$$

The exact form of $\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)$ with depend on the assisting model. In the next sections, we explore the major components of $\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)$ and $\widehat{\boldsymbol{\Sigma}}_{\widehat{\mathbf{U}}}\left(\widehat{\boldsymbol{\theta}}\right)$ for the LGREG.

Simplification of Jacobian

To simplify our variance estimator, we must simplify $\widehat{\mathbf{J}} = \frac{\partial}{\partial\boldsymbol{\theta}}\widehat{\mathbf{W}}(\boldsymbol{\theta})$. We first partition $\widehat{\mathbf{J}}$ into four blocks,

$$\underset{(C+C\cdot p-p)\times(C+C\cdot p-p)}{\widehat{\mathbf{J}}} = \begin{bmatrix} \mathscr{A} & \mathscr{B} \\ \mathscr{C} & \mathscr{D} \end{bmatrix}$$

where

$$\begin{aligned}
\underset{C\times C}{\mathscr{A}} &= \frac{\partial}{\partial\left[\mathrm{vec}\left(\widehat{\mathbf{t}}_y^{lg}\right)\right]^{\top}}\left[\sum_{\mathfrak{s}} d_k\left(\mathbf{y}_k - \boldsymbol{\mu}_k\right) - \mathbf{t}_y^{lg} + \sum_{\mathscr{U}}\boldsymbol{\mu}_k\right] \\
\underset{C\times(C\cdot p-p)}{\mathscr{B}} &= \frac{\partial}{\partial\left[\mathrm{vec}\left(\mathbf{B}\right)\right]^{\top}}\left[\sum_{\mathfrak{s}} d_k\left(\mathbf{y}_k - \boldsymbol{\mu}_k\right) - \mathbf{t}_y^{lg} + \sum_{\mathscr{U}}\boldsymbol{\mu}_k\right] \\
\underset{(C\cdot p-p)\times C}{\mathscr{C}} &= \frac{\partial}{\partial\left[\mathrm{vec}\left(\widehat{\mathbf{t}}_y^{lg}\right)\right]^{\top}}\left[\sum_{\mathfrak{s}} d_k\mathbf{x}_k\left(\mathbf{y}_k^{\top} - \boldsymbol{\mu}_k^{\top}\right)\right] \\
\underset{(C\cdot p-p)\times(C\cdot p-p)}{\mathscr{D}} &= \frac{\partial}{\left[\mathrm{vec}\left(\mathbf{B}\right)\right]^{\top}}\left[\sum_{\mathfrak{s}} d_k\mathbf{x}_k\left(\mathbf{y}_k^{\top} - \boldsymbol{\mu}_k^{\top}\right)\right]
\end{aligned}$$

Simplification of $\mathscr{A}$

Simplifying $\mathscr{A}$ is rather easy,

$$\begin{aligned}
\mathscr{A} &= \frac{\partial}{\partial \left[\mathrm{vec}\left(\widehat{\mathbf{t}}_y^{lg}\right)\right]^\top} \left[\sum_{\mathfrak{s}} d_k \left(\mathbf{y}_k - \boldsymbol{\mu}_k\right) - \mathbf{t}_y^{lg} + \sum_{\mathscr{U}} \boldsymbol{\mu}_k\right] \\
&= \frac{\partial}{\partial \left[\mathrm{vec}\left(\widehat{\mathbf{t}}_y^{lg}\right)\right]^\top} \left(-\mathbf{t}_y^{lg}\right) \\
&= -\mathbf{I}_C.
\end{aligned}$$

Simplification of $\mathscr{B}$

To simplify $\mathscr{B}$, we extend a proof in Agresti (2002)[section 14.4]. Let $\boldsymbol{\mu}_k = z_k \mathfrak{p}$. Further, let B_i be the ith element of $\mathrm{vec}\left(\mathbf{B}\right)$ and let $\mathbf{B}_h$ be one column of $\mathbf{B}$. Agresti (2002) simplifies $\frac{\partial \mathfrak{p}_{kc}}{\partial B_i}$. Extending those results for $\frac{\partial \mu_{kc}}{\partial B_i}$ gives,

$$\begin{aligned}
\frac{\partial \mu_{kc}}{\partial B_i} &= \frac{\partial z_k \mathfrak{p}_{kc}}{\partial B_i} \\
&= \frac{\partial}{\partial B_i} \frac{z_k e^{\mathbf{x}_k^\top \mathbf{B}_c}}{1 + \sum_{h=1}^{C-1} e^{\mathbf{x}_k^\top \mathbf{B}_h}} \\
&= z_k \frac{\left[1 + \sum_{h=1}^{C-1} e^{\mathbf{x}_k \mathbf{B}_h}\right] \left[e^{\mathbf{x}_k \mathbf{B}_c}\right] x_{kci} - \left[e^{\mathbf{x}_k \mathbf{B}_c}\right] \left[1 + \sum_{h=1}^{C-1} x_{khi} e^{\mathbf{x}_k \mathbf{B}_h}\right]}{\left(1 + \sum_{h=1}^{C-1} e^{\mathbf{x}_k \mathbf{B}_h}\right)^2} \\
&= z_k \mathfrak{p}_{kc} x_{kci} - z_k \mathfrak{p}_{kc} \sum_{h=1}^{C-1} x_{khi} \mathfrak{p}_{kh}.
\end{aligned}$$

In matrix notation, we have

$$\frac{\partial \boldsymbol{\mu}_k}{\partial \left(\mathrm{vec}\mathbf{B}\right)^\top} = z_k \left[\mathrm{diag}\left(\mathfrak{p}_k\right) - \mathfrak{p}_k \mathfrak{p}_k^\top\right] \mathbf{X}_k \tag{B.23}$$

Thus,

$$\begin{aligned}
\mathscr{B} &= \frac{\partial}{\partial \left[\mathrm{vec}\left(\mathbf{B}\right)\right]^{\top}} \left[\sum_{\mathfrak{s}} d_k \left(\mathbf{y}_k - \boldsymbol{\mu}_k\right) - \sum_{\mathscr{U}} \left[\frac{1}{N}\mathbf{t}_y^{lg} - \boldsymbol{\mu}_k\right]\right] \\
&= \frac{\partial}{\partial \left[\mathrm{vec}\left(\mathbf{B}\right)\right]^{\top}} \left[\sum_{\mathscr{U}} \boldsymbol{\mu}_k - \sum_{\mathfrak{s}} d_k \boldsymbol{\mu}_k\right] \\
&= \sum_{\mathscr{U}} z_k \left[\mathrm{diag}\left(\mathfrak{p}_k\right) - \mathfrak{p}_k \mathfrak{p}_k^{\top}\right] \mathbf{X}_k - \sum_{\mathfrak{s}} d_k z_k \left[\mathrm{diag}\left(\mathfrak{p}_k\right) - \mathfrak{p}_k \mathfrak{p}_k^{\top}\right] \mathbf{X}_k.
\end{aligned}$$

Simplification of $\mathscr{C}$

Since the third derivative, $\mathscr{C}$, does not contain $\widehat{\mathbf{t}}_y^{lg}$ it is quite simple to calculate

$$\begin{aligned}
\mathscr{C} &= \frac{\partial}{\partial \left[\mathrm{vec}\left(\widehat{\mathbf{t}}_y^{lg}\right)\right]^{\top}} \sum_{\mathfrak{s}} d_k \mathbf{x}_k \left(\mathbf{y}_k^{\top} - \boldsymbol{\mu}_k^{\top}\right) \\
&= \mathbf{0}.
\end{aligned}$$

Simplification of $\mathscr{D}$

The final derivative, $\mathscr{D}$, is a common component in estimating the variance of $\mathbf{B}$. According to Agresti

$$\begin{aligned}
\mathscr{D} &= \frac{\partial}{\left[\mathrm{vec}\left(\mathbf{B}\right)\right]^{\top}} \sum_{\mathfrak{s}} d_k \mathbf{x}_k \left(\mathbf{y}_k^{\top} - \boldsymbol{\mu}_k^{\top}\right) \\
&= -\sum_{\mathfrak{s}} d_k \frac{\partial \mathbf{x}_k \boldsymbol{\mu}_k^{\top}}{\left[\mathrm{vec}\left(\mathbf{B}\right)\right]^{\top}} \\
&= -\sum_{\mathfrak{s}} d_k z_k \mathbf{X}_k^{\top} \left[\mathrm{diag}\left(\mathfrak{p}_k\right) - \mathfrak{p}_k \mathfrak{p}_k^{\top}\right] \mathbf{X}_k
\end{aligned}$$

Therefore,

$$\widehat{\mathbf{J}} = \begin{bmatrix} -\mathbf{I} & \sum_{\mathfrak{s}} d_k z_k \left[\mathrm{diag}\left(\mathfrak{p}_k\right) - \mathfrak{p}_k \mathfrak{p}_k^{\top}\right] \mathbf{X}_k - \sum_{\mathscr{U}} z_k \left[\mathrm{diag}\left(\mathfrak{p}_k\right) - \mathfrak{p}_k \mathfrak{p}_k^{\top}\right] \mathbf{X}_k \\ \mathbf{0} & -\sum_{\mathfrak{s}} d_k z_k \mathbf{X}_k^{\top} \left[\mathrm{diag}\left(\mathfrak{p}_k\right) - \mathfrak{p}_k \mathfrak{p}_k^{\top}\right] \mathbf{X}_k \end{bmatrix}$$

Simplification of $\Sigma_{\widehat{U}}\left(\widehat{\boldsymbol{\theta}}\right)$

$\boldsymbol{\Sigma}$ is the design-based variance of the estimating equations. That is $\boldsymbol{\Sigma} = \mathrm{var}\left[\widehat{\mathbf{U}}\left(\boldsymbol{\theta}\right)\right]$. An estimator of this variance is denoted $\widehat{\boldsymbol{\Sigma}} = \widehat{\mathrm{var}}\left[\widehat{\mathbf{U}}\left(\boldsymbol{\theta}\right)\right]$. We note that $\widehat{\boldsymbol{\Sigma}}$ is a function of $\widehat{\boldsymbol{\theta}}$.

Perhaps the simplest way to estimate $\boldsymbol{\Sigma}$ is to assume that clusters were selected with-replacement. Consider the expansion estimator for the total of the estimating equations for cluster i,

$$\widehat{\mathbf{U}}_{i+}\left(\boldsymbol{\theta}\right) = \sum_{k \in \mathfrak{s}_i} \widehat{\mathbf{U}}_k\left(\boldsymbol{\theta}\right)$$

Also, recall that under with-replacement sampling, the single draw probability for cluster i is $p_i = \frac{\pi_{Ii}}{n_I}$. In this case, Särndal et al. (1992)[p. 154] show that an unbiased estimator of $\boldsymbol{\Sigma}$ is

$$\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right) = \frac{n}{n-1}\left\{\sum_{\mathfrak{s}_I}\left[\widehat{\mathbf{t}}_{\hat{U}i} - \frac{1}{n}\sum_{i \in \mathfrak{s}_I}\widehat{\mathbf{t}}_{\hat{U}i}\right]\right\}\left\{\sum_{\mathfrak{s}_I}\left[\widehat{\mathbf{t}}_{\hat{U}i} - \frac{1}{n}\sum_{i \in \mathfrak{s}_I}\widehat{\mathbf{t}}_{\hat{U}i}\right]\right\}^{\top}$$

On the other hand, if the first stage sample is selected without-replacement, then we can extend the classic design variance formulas from Särndal et al. (1992)[p. 137] to the multivariate case. The variance of the estimating equations will be

$$\begin{aligned}\boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}\right) = &\sum\sum_{\mathscr{U}_I}\frac{\pi_{Iij} - \pi_{Ii}\pi_{Ij}}{\pi_{Ii}\pi_{Ij}}\widehat{\mathbf{U}}_{i+}\left(\widehat{\boldsymbol{\theta}}\right)\widehat{\mathbf{U}}_{j+}\left(\widehat{\boldsymbol{\theta}}\right)^{\top} \\ &+ \sum_{\mathscr{U}_I}\frac{1}{\pi_i}\sum\sum_{\mathscr{U}_i}\frac{\pi_{kl|i} - \pi_{k|i}\pi_{l|i}}{\pi_{k|i}\pi_{l|i}}\widehat{\mathbf{U}}_{k|i}\left(\widehat{\boldsymbol{\theta}}\right)\widehat{\mathbf{U}}_{l|i}\left(\widehat{\boldsymbol{\theta}}\right)^{\top}\end{aligned}$$

Resampling techniques may also be used to estimate $\boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}\right)$ in cluster samples.

B.5 Model Calibration

B.5.1 Construction of model calibration estimator

The model-calibrated estimated total is

$$\underset{C\times 1}{\widehat{\mathbf{t}}_y^{mc}} = \mathbf{y}^\top \mathbf{w}^{mc}$$

where $\underset{n\times 1}{\mathbf{w}^{mc}}$ is found by minimizing the weighted distance between the design weights and the model calibration weights,

$$\frac{1}{2}\left(\mathbf{d}-\mathbf{w}^{mc}\right)^\top \mathbf{\Pi Q}^{-1}\left(\mathbf{d}-\mathbf{w}^{mc}\right)$$

subject to the constraint

$$\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \mathbf{w}^{mc} = \underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top \mathbf{1}$$

where

$$\underline{\boldsymbol{\mu}} = \begin{bmatrix} \mathbf{1} & \boldsymbol{\mu} \end{bmatrix}$$

Notice that if the first column of $\underset{n\times(C+1)}{\underline{\boldsymbol{\mu}}_{\mathfrak{s}}}$ is $\mathbf{1}$, then the following constraint is also obtained.

$$\mathbf{1}^\top \mathbf{w}^{mc} = N.$$

Our restricted objective function is

$$\phi = \frac{1}{2}\left(\mathbf{d}-\mathbf{w}^{mc}\right)^\top \mathbf{\Pi Q}^{-1}\left(\mathbf{d}-\mathbf{w}^{mc}\right) - \boldsymbol{\lambda}^\top\left(\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \mathbf{w}^{mc} - \underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top \mathbf{1}\right)$$

where $\boldsymbol{\lambda}$ is a $C+1$ by 1 vector of Lagrange multipliers.

Differentiating ϕ with respect to $\mathbf{w}^{mc}$ gives

$$\frac{\partial \phi}{\partial \mathbf{w}^{mc}} = \mathbf{\Pi Q}^{-1}\left(\mathbf{w}^{mc} - \mathbf{d}\right) - \underline{\boldsymbol{\mu}}_{\mathfrak{s}}\boldsymbol{\lambda}.$$

Setting the partial derivative equal to 0 and solving for $\mathbf{w}^{mc}$ gives

$$\begin{aligned}
\underset{n\times 1}{\mathbf{0}} &= \left(\mathbf{Q\Pi}^{-1}\right)\mathbf{\Pi Q}^{-1}\left(\mathbf{w}^{mc} - \mathbf{d}\right) - \left(\mathbf{Q\Pi}^{-1}\right)\underline{\boldsymbol{\mu}}_{\mathfrak{s}}\boldsymbol{\lambda} \\
&= \mathbf{w}^{mc} - \mathbf{d} - \left(\mathbf{Q\Pi}^{-1}\right)\underline{\boldsymbol{\mu}}_{\mathfrak{s}}\boldsymbol{\lambda} \\
\mathbf{w}^{mc} &= \mathbf{d} + \left(\mathbf{Q\Pi}^{-1}\right)\underline{\boldsymbol{\mu}}_{\mathfrak{s}}\boldsymbol{\lambda}.
\end{aligned}$$

Plugging $\mathbf{w}^{mc}$ into our constraint and solving for $\boldsymbol{\lambda}$ gives

$$\begin{aligned}
\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\mathbf{w}^{mc} &= \underline{\boldsymbol{\mu}}_{\mathscr{U}}^{\top}\mathbf{1} \\
\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\left(\mathbf{d} + \left(\mathbf{Q\Pi}^{-1}\right)\underline{\boldsymbol{\mu}}_{\mathfrak{s}}\boldsymbol{\lambda}\right) &= \underline{\boldsymbol{\mu}}_{\mathscr{U}}^{\top}\mathbf{1} \\
\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\mathbf{d} + \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\left(\mathbf{Q\Pi}^{-1}\right)\underline{\boldsymbol{\mu}}_{\mathfrak{s}}\boldsymbol{\lambda} &= \underline{\boldsymbol{\mu}}_{\mathscr{U}}^{\top}\mathbf{1} \\
\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\left(\mathbf{Q\Pi}^{-1}\right)\underline{\boldsymbol{\mu}}_{\mathfrak{s}}\boldsymbol{\lambda} &= \underline{\boldsymbol{\mu}}_{\mathscr{U}}^{\top}\mathbf{1} - \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\mathbf{d} \\
\underset{C\times 1}{\boldsymbol{\lambda}} &= \left(\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\left(\mathbf{Q\Pi}^{-1}\right)\underline{\boldsymbol{\mu}}_{\mathfrak{s}}\right)^{-1}\left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^{\top}\mathbf{1} - \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\mathbf{d}\right).
\end{aligned}$$

Thus, our calibrated weights are

$$\begin{aligned}
\mathbf{w}^{mc} &= \mathbf{d} + \left(\mathbf{Q\Pi}^{-1}\right)\underline{\boldsymbol{\mu}}_{\mathfrak{s}}\boldsymbol{\lambda} \\
&= \mathbf{d} + \left(\mathbf{Q\Pi}^{-1}\right)\underline{\boldsymbol{\mu}}_{\mathfrak{s}}\left(\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\left(\mathbf{Q\Pi}^{-1}\right)\underline{\boldsymbol{\mu}}_{\mathfrak{s}}\right)^{-1}\left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^{\top}\mathbf{1} - \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\mathbf{d}\right).
\end{aligned}$$

Our final estimator is

$$
\begin{aligned}
\hat{t}_y^{mc} &= \mathbf{y}^\top \mathbf{w}^{mc} \\
&= \mathbf{y}^\top \left[\mathbf{d} + \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \underline{\boldsymbol{\mu}}_{\mathfrak{s}} \left(\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \underline{\boldsymbol{\mu}}_{\mathfrak{s}} \right)^{-1} \left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top \mathbf{1} - \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \mathbf{d} \right) \right] \\
&= \hat{\boldsymbol{t}}_y + \mathbf{y}^\top \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \underline{\boldsymbol{\mu}}_{\mathfrak{s}} \left(\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \underline{\boldsymbol{\mu}}_{\mathfrak{s}} \right)^{-1} \left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top \mathbf{1} - \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \mathbf{d} \right) \\
&= \hat{\boldsymbol{t}}_y + \widehat{\mathbf{B}}_{\underline{\boldsymbol{\mu}}\mathbf{y}} \left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top \mathbf{1} - \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \mathbf{d} \right)
\end{aligned}
$$

where

$$
\underset{C\times C+1}{\widehat{\mathbf{B}}_{\underline{\boldsymbol{\mu}}\mathbf{Y}}} = \mathbf{y}^\top \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \underline{\boldsymbol{\mu}}_{\mathfrak{s}} \left(\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \underline{\boldsymbol{\mu}}_{\mathfrak{s}} \right)^{-1}
$$

B.5.2 Alternative forms of the model calibration estimator

If $\boldsymbol{\Pi}^{-1}$ and $\mathbf{Q}$ are invertible and commutable, then our estimator can be written as:

$$
\begin{aligned}
\hat{t}_y^{mc} &= \mathbf{y}^\top \left[\mathbf{d} + \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \underline{\boldsymbol{\mu}}_{\mathfrak{s}} \left(\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \underline{\boldsymbol{\mu}}_{\mathfrak{s}} \right)^{-1} \left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top \mathbf{1} - \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \mathbf{d} \right) \right] \\
&= \mathbf{y}^\top \left[\boldsymbol{\Pi}^{-1}\mathbf{1} + \boldsymbol{\Pi}^{-1}\mathbf{Q}\underline{\boldsymbol{\mu}}_{\mathfrak{s}} \left(\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \underline{\boldsymbol{\mu}}_{\mathfrak{s}} \right)^{-1} \left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top \mathbf{1} - \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \mathbf{d} \right) \right] \\
&= \mathbf{y}^\top \boldsymbol{\Pi}^{-1} \left[\mathbf{1} + \mathbf{Q}\underline{\boldsymbol{\mu}}_{\mathfrak{s}} \left(\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \underline{\boldsymbol{\mu}}_{\mathfrak{s}} \right)^{-1} \left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top \mathbf{1} - \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \mathbf{d} \right) \right] \\
&= \mathbf{y}^\top \boldsymbol{\Pi}^{-1} \mathbf{g}
\end{aligned}
$$

where

$$
\underset{n\times 1}{\mathbf{g}} = \mathbf{1} + \mathbf{Q}^{-1}\underline{\boldsymbol{\mu}}_{\mathfrak{s}} \left(\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right) \underline{\boldsymbol{\mu}}_{\mathfrak{s}} \right)^{-1} \left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top \mathbf{1} - \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \mathbf{d} \right).
$$

We can compactly write our estimator if we let

$$\begin{aligned}
\mathbf{\Pi}_\star &= \mathbf{\Pi}\mathbf{Q}^{-1} \\
\underset{C\times 1}{\mathbf{t}_{\underline{\widehat{\boldsymbol{\mu}}}}} &= \underline{\widehat{\boldsymbol{\mu}}}_{\mathscr{U}}^\top \mathbf{1} = \sum_{\mathscr{U}} \underline{\widehat{\boldsymbol{\mu}}}_k \\
\underset{C\times 1}{\widehat{\mathbf{t}}_{\underline{\widehat{\boldsymbol{\mu}}}}} &= \underline{\widehat{\boldsymbol{\mu}}}_{\mathfrak{s}}^\top \mathbf{d} = \sum_{\mathfrak{s}} d_k \underline{\widehat{\boldsymbol{\mu}}}_k \\
\underset{C\times C}{\widehat{\mathbf{A}}} &= \underline{\widehat{\boldsymbol{\mu}}}_{\mathfrak{s}}^\top \mathbf{\Pi}_\star^{-1} \underline{\widehat{\boldsymbol{\mu}}}_{\mathfrak{s}} = \sum_{\mathfrak{s}} \frac{d_k}{q_k} \underline{\widehat{\boldsymbol{\mu}}}_k \underline{\widehat{\boldsymbol{\mu}}}_k^\top .
\end{aligned}$$

For convenience, we let

$$\begin{aligned}
\underset{C\times 1}{\mathbf{t}_{\underline{\boldsymbol{\mu}}}} &= \underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top \mathbf{1} = \sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}_k \\
\widehat{\mathbf{t}}_{\underline{\boldsymbol{\mu}}}(\mathbf{B}) &= \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \mathbf{d} = \sum_{\mathfrak{s}} d_k \underline{\boldsymbol{\mu}}_k \\
\underset{C\times C}{\mathbf{A}} &= \underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top \underline{\boldsymbol{\mu}}_{\mathscr{U}} = \sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}_k \underline{\boldsymbol{\mu}}_k^\top \\
\widehat{\mathbf{A}}(\mathbf{B}) &= \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \mathbf{\Pi}_\star^{-1} \underline{\boldsymbol{\mu}}_{\mathfrak{s}} = \sum_{\mathfrak{s}} \frac{d_k}{q_k} \underline{\boldsymbol{\mu}}_k \underline{\boldsymbol{\mu}}_k^\top .
\end{aligned}$$

The model calibration estimator of a finite population total is

$$\begin{aligned}
\widehat{\mathbf{t}}_y^{mc} &= \mathbf{y}^\top \left[\mathbf{d} + \left(\mathbf{Q}\mathbf{\Pi}^{-1}\right) \underline{\widehat{\boldsymbol{\mu}}}_{\mathfrak{s}} \left(\underline{\widehat{\boldsymbol{\mu}}}_{\mathfrak{s}}^\top \left(\mathbf{Q}\mathbf{\Pi}^{-1}\right) \underline{\widehat{\boldsymbol{\mu}}}_{\mathfrak{s}} \right)^{-1} \left(\underline{\widehat{\boldsymbol{\mu}}}_{\mathscr{U}}^\top \mathbf{1} - \underline{\widehat{\boldsymbol{\mu}}}_{\mathfrak{s}}^\top \mathbf{d} \right) \right] . \\
&= \widehat{\mathbf{t}}_y + \mathbf{y}_{\mathfrak{s}}^\top \underline{\widehat{\boldsymbol{\mu}}}_{\mathfrak{s}} \left[\widehat{\mathbf{A}} \right]^{-1} \left(\mathbf{t}_{\underline{\widehat{\boldsymbol{\mu}}}} - \widehat{\mathbf{t}}_{\underline{\widehat{\boldsymbol{\mu}}}} \right) \\
&= \sum_{\mathfrak{s}} \frac{d_k}{q_k} \mathbf{y}_k \left[1 + \underline{\widehat{\boldsymbol{\mu}}}_k^\top \left[\widehat{\mathbf{A}} \right]^{-1} \left(\mathbf{t}_{\underline{\widehat{\boldsymbol{\mu}}}} - \widehat{\mathbf{t}}_{\underline{\widehat{\boldsymbol{\mu}}}} \right) \right] .
\end{aligned}$$

B.5.3 Design consistency of model calibration estimator

In this section, we prove that the model calibration estimator is design consistent for the true value. A similar proof is in Appendix B.4.1 on page 317. Because these two proofs are so similar, much of this proof is repeated in both sections.

Wu and Sitter (2001) show that $\hat{t}_y^{mc}$ is design consistent and asymptotically unbiased for a generalized linear model with a univariate response. Here, we extend their proof to a multivariate response with a logistic regression model.

By the mean value theorem, there is a point, $\mathbf{B}^\star$, such that

$$\underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \mathbf{B}_N\right) + \left[\frac{\partial \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \mathbf{t}\right)}{\partial \mathbf{t}}\bigg|_{\mathbf{t}=\mathbf{B}^\star}\right]^\top \left(\hat{\mathbf{B}} - \mathbf{B}_N\right)$$

Summing and dividing by N gives,

$$N^{-1}\sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = N^{-1}\sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \mathbf{B}_N\right) + N^{-1}\sum_{\mathscr{U}} \left[\frac{\partial \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \mathbf{t}\right)}{\partial \mathbf{t}}|\mathbf{t} = \mathbf{B}^\star\right]^\top \left(\hat{\mathbf{B}} - \mathbf{B}_N\right).$$

To show consistency, we must make some assumptions. Here, we borrow two assumptions from Wu and Sitter (2001). First, we assume that our derivative is locally continuous around $\mathbf{B}_N$ and that our derivative is bounded as the sample and population grow. This is a fairly mild regularity condition. Specifically, we make Assumption 5 in Section 3.2.1.1 on page 134.

Second, we assume that our estimator of $\mathbf{B}_N$ is consistent. Most standard estimation techniques, including those recommended in this thesis, share this property. Our specific assumption is, Assumption 4 in Section 3.2.1.1.

Under Assumption 5 in Section 3.2.1.1, $N^{-1}\sum_{\mathscr{U}} \left[\frac{\partial \underline{\boldsymbol{\mu}}(\mathbf{x}_k, \mathbf{t})}{\partial \mathbf{t}}\Big|_{\mathbf{t}=\mathbf{B}^\star}\right]^\top = O\left(1\right)$. Also, under Assumption 4 in Section 3.2.1.1, we see that $\left(\hat{\mathbf{B}} - \mathbf{B}_N\right) = O_p\left(n^{-\frac{1}{2}}\right)$. Thus

$$N^{-1}\sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = N^{-1}\sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \mathbf{B}_N\right) + O_p\left(n^{-\frac{1}{2}}\right).$$

So, as the sample size gets large, $N^{-1}\sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right)$ converges to the true population quantity $N^{-1}\sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \mathbf{B}_N\right)$.

Now, we consider weighted totals. The weighted average for our initial expression is,

$$N^{-1}\sum_{\mathfrak{s}} d_k \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = N^{-1}\sum_{\mathfrak{s}} d_k \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \mathbf{B}_N\right) + N^{-1}\sum_{\mathfrak{s}} d_k \left[\frac{\partial \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \mathbf{t}\right)}{\partial \mathbf{t}}\bigg|_{\mathbf{t}=\mathbf{B}^\star}\right]^\top \left(\hat{\mathbf{B}} - \mathbf{B}_N\right).$$

By Assumption 14 in Appendix A.1 on page 258, our weights are roughly $O\left(\frac{N}{n}\right)$. Summing these weights for the sample gives $N^{-1}\sum_s \frac{N}{n} = N^{-1}n\frac{N}{n} = O\left(1\right)$. Under Assumptions 5 and 4 in Section 3.2.1.1 as well as the assumption that none of our weights dominate, we see that

$$N^{-1}\sum_{\mathfrak{s}} d_k \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = N^{-1}\sum_{\mathfrak{s}} d_k \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \mathbf{B}_N\right) + O_p\left(n^{-\frac{1}{2}}\right).$$

So, as the sample size gets large, $N^{-1}\sum_{\mathfrak{s}} d_k \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right)$ converges to the estimated value when the true coefficients are used.

By Assumption 6 in Section 3.2.1.1,

$$N^{-1}\sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \mathbf{B}_N\right) - N^{-1}\sum_{\mathfrak{s}} d_k \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \mathbf{B}_N\right) = O_p\left(n^{-\frac{1}{2}}\right).$$

To summarize, we showed that

$$N^{-1}\sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = N^{-1}\sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \mathbf{B}_N\right) + O_p\left(n^{-\frac{1}{2}}\right)$$

$$N^{-1}\sum_{\mathfrak{s}} d_k \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = N^{-1}\sum_{\mathfrak{s}} d_k \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \mathbf{B}_N\right) + O_p\left(n^{-\frac{1}{2}}\right)$$

$$N^{-1}\sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \mathbf{B}_N\right) - N^{-1}\sum_{\mathfrak{s}} d_k \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \mathbf{B}_N\right) = O_p\left(n^{-\frac{1}{2}}\right).$$

Together these three equations imply that

$$N^{-1}\sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) - N^{-1}\sum_{\mathfrak{s}} d_k \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = O_p\left(n^{-\frac{1}{2}}\right)$$

In terms of matrices, we have

$$\left(\underline{\hat{\boldsymbol{\mu}}}_{\mathscr{U}}^{\top}\mathbf{1} - \underline{\hat{\boldsymbol{\mu}}}_{\mathfrak{s}}^{\top}\mathbf{d}\right) = O_p\left(n^{-\frac{1}{2}}\right)$$

Now, consider our estimator,

$$\begin{aligned}\hat{\mathbf{t}}_y^{mc} &= \hat{\mathbf{t}}_y + \widehat{\mathbf{B}}_{\underline{\boldsymbol{\mu}}\mathbf{Y}}\left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^{\top}\mathbf{1} - \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\mathbf{d}\right) \\ &= \hat{\mathbf{t}}_y + \widehat{\mathbf{B}}_{\underline{\boldsymbol{\mu}}\mathbf{Y}} O_p\left(n^{-\frac{1}{2}}\right)\end{aligned}$$

Furthermore, $\hat{\mathbf{B}}_{\underline{\boldsymbol{\mu}}\mathbf{Y}} = O_p(1)$. Thus,

$$\hat{\mathbf{t}}_y^{mc} = \hat{\mathbf{t}}_y + O_p\left(n^{-\frac{1}{2}}\right)$$

Since $\hat{\mathbf{t}}_y = \mathbf{t}_y + O_p\left(n^{-\frac{1}{2}}\right)$, we see that $\hat{\mathbf{t}}_y^{mc}$ is a consistent estimator. Furthermore, it is asymptotically centered around the Horvitz-Thompson estimator, an unbiased estimator.

B.5.4 Asymptotic variance of model calibration estimator

Wu and Sitter (2001) calculate the asymptotic variance of $\hat{t}_y^{mc}$ for a scalar under a generalized linear model. Here, we extend their argument for multivariate responses under a logistic regression model.

Following the outline of the proof in Wu and Sitter (2001), we show that

$$\begin{aligned}\hat{\hat{\mathbf{t}}}_y^{mc} &= \hat{\hat{\boldsymbol{t}}}_y + \mathbf{B}_{\underline{\boldsymbol{\mu}}\mathbf{Y}}\left(N^{-1}\underline{\boldsymbol{\mu}}_{\mathscr{U}}^{\top}\mathbf{1} - N^{-1}\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\mathbf{d}\right) + o_p(1) \\ &= N^{-1}\mathbf{e}_{\mathfrak{s}}^{\top}\mathbf{d} + \mathbf{B}_{\underline{\boldsymbol{\mu}}\mathbf{Y}}\left(N^{-1}\underline{\boldsymbol{\mu}}_{\mathscr{U}}^{\top}\mathbf{1}\right) + o_p(1)\end{aligned}$$

where

$$\mathbf{e} = \mathbf{y} - \mathbf{B}_{\underline{\boldsymbol{\mu}}\mathbf{Y}}\underline{\boldsymbol{\mu}}_k$$

Further, we show that, the asymptotic variance of the model calibration estimator for a multinomial response in two staged samples is

$$\mathrm{av}_{II}\left(\hat{\mathbf{t}}_y^{mc}\right) = \sum\sum_{\mathscr{U}_I} \Delta_{ij} \frac{t_{e_i}}{\pi_{Ii}} \frac{t_{e_j}}{\pi_{Ij}} + \sum_{\mathscr{U}_I} \frac{\sum_{\mathscr{U}_i} \sum_{\mathscr{U}_i} \Delta_{kl|i} \frac{e_{k|i}}{\pi_{k|i}} \frac{e_{l|i}}{\pi_{l|i}}}{\pi_{Ii}}.$$

Our proof is identical to that shown in Wu and Sitter (2001) with a few minor exceptions. Rather than assuming a generalized linear model, we explicitly use a multinomial logistic model. This means that our dependent variable and coefficients are matrices, and our model has an explicit form. Furthermore, whereas Wu and Sitter (2001) provide results for single stage sampling, we focus the variance when samples were selected in two stages.

Proof

A second order Taylor series expansion of $\mu\left(\mathbf{x}, \widehat{\mathbf{B}}\right)$ at $\widehat{\mathbf{B}} = \mathbf{B}_N$ is

$$\begin{aligned}\underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \mathbf{B}_N\right) &+ \left[\frac{\partial \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \mathbf{t}\right)}{\partial \mathbf{t}}\bigg|_{\mathbf{t}=\mathbf{B}_N}\right]^\top \left(\hat{\mathbf{B}} - \mathbf{B}_N\right) \\ &+ \left(\hat{\mathbf{B}} - \mathbf{B}_N\right)^\top \left[\frac{\partial^2 \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \mathbf{t}\right)}{\partial \mathbf{t} \partial \mathbf{t}^\top}\bigg|_{\mathbf{t}=\mathbf{B}^\star}\right]^\top \left(\hat{\mathbf{B}} - \mathbf{B}_N\right)\end{aligned}$$

Summing and dividing by N gives,

$$\begin{aligned}\frac{1}{N}\sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) = \frac{1}{N}\sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \mathbf{B}_N\right) &+ \frac{1}{N}\sum_{\mathscr{U}} \left[\frac{\partial \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \mathbf{t}\right)}{\partial \mathbf{t}}\bigg|_{\mathbf{t}=\mathbf{B}_N}\right]^\top \left(\hat{\mathbf{B}} - \mathbf{B}_N\right) \\ &+ \frac{1}{N}\sum_{\mathscr{U}} \left(\hat{\mathbf{B}} - \mathbf{B}_N\right)^\top \left[\frac{\partial^2 \underline{\boldsymbol{\mu}}\left(\mathbf{x}_k, \mathbf{t}\right)}{\partial \mathbf{t} \partial \mathbf{t}^\top}\bigg|_{\mathbf{t}=\mathbf{B}^\star}\right]^\top \left(\hat{\mathbf{B}} - \mathbf{B}_N\right).\end{aligned}$$

where $\mathbf{B}^\star$ is a point between $\hat{\mathbf{B}}$ and $\mathbf{B}_N$.

Now, we borrow another assumption from Wu and Sitter (2001). We assume that our second derivative is locally continuous around $\mathbf{B}_N$ and that our second derivative is

bounded as the sample and population grow. This is a fairly mild regularity condition. This assumption is written as Assumption 7 in Section 3.2.1.1 on page 134.

Now, by Assumptions 4 and 7 in in Section 3.2.1.1, we see that

$$\frac{1}{N}\sum_{\mathscr{U}}\underline{\boldsymbol{\mu}}\left(\mathbf{x}_k,\hat{\mathbf{B}}\right)=\frac{1}{N}\sum_{\mathscr{U}}\underline{\boldsymbol{\mu}}\left(\mathbf{x}_k,\mathbf{B}_N\right)+\frac{1}{N}\sum_{\mathscr{U}}\left[\frac{\partial\underline{\boldsymbol{\mu}}\left(\mathbf{x}_k,\mathbf{t}\right)}{\partial\mathbf{t}}\bigg|_{\mathbf{t}=\mathbf{B}_N}\right]^{\top}\left(\hat{\mathbf{B}}-\mathbf{B}_N\right)+O_p\left(n^{-1}\right)$$

Renaming the first derivative gives

$$=\frac{1}{N}\sum_{\mathscr{U}}\underline{\boldsymbol{\mu}}\left(\mathbf{x}_k,\mathbf{B}_N\right)+\frac{1}{N}\sum_{\mathscr{U}}\left[\mathbf{K}\left(\mathbf{x}_k,\mathbf{B}_N\right)\right]^{\top}\left(\hat{\mathbf{B}}-\mathbf{B}_N\right)+O_p\left(n^{-1}\right).$$

Furthermore, by our assumptions about the sample design

$$\frac{1}{N}\sum_{\mathscr{U}}d_k\underline{\boldsymbol{\mu}}\left(\mathbf{x}_k,\hat{\mathbf{B}}\right)=\frac{1}{N}\sum_{\mathscr{U}}d_k\underline{\boldsymbol{\mu}}\left(\mathbf{x}_k,\mathbf{B}_N\right)+\frac{1}{N}\sum_{\mathscr{U}}\left[d_k\mathbf{K}\left(\mathbf{x}_i,\mathbf{B}_N\right)\right]^{\top}\left(\hat{\mathbf{B}}-\mathbf{B}_N\right)+O_p\left(n^{-1}\right).$$

By Assumptions 5, 4, and 6 in in Section 3.2.1.1, we have $\hat{\mathbf{B}}-\mathbf{B}_N=O\left(n^{-\frac{1}{2}}\right)$ and $\frac{1}{N}\sum_{\mathscr{U}}\underline{\boldsymbol{\mu}}\left(\mathbf{x}_k,\mathbf{B}_N\right)-\frac{1}{N}\sum_{\mathfrak{s}}d_k\underline{\boldsymbol{\mu}}\left(\mathbf{x}_k,\mathbf{B}_N\right)=O_p\left(n^{-\frac{1}{2}}\right)$. Therefore,

$$\frac{1}{N}\sum_{\mathscr{U}}\underline{\boldsymbol{\mu}}\left(\mathbf{x}_k,\hat{\mathbf{B}}\right)-\frac{1}{N}\sum_{\mathfrak{s}}d_k\underline{\boldsymbol{\mu}}\left(\mathbf{x}_k,\hat{\mathbf{B}}\right)=\frac{1}{N}\sum_{\mathscr{U}}\underline{\boldsymbol{\mu}}\left(\mathbf{x}_k,\mathbf{B}\right)-\frac{1}{N}\sum_{\mathfrak{s}}d_k\underline{\boldsymbol{\mu}}\left(\mathbf{x}_k,\mathbf{B}\right)+O_p\left(n^{-1}\right).$$

Now, consider our estimator,

$$\hat{\mathbf{t}}_y^{mc}=\hat{\boldsymbol{t}}_y+\widehat{\mathbf{B}}_{\underline{\boldsymbol{\mu}}\mathbf{Y}}\left(\underline{\hat{\boldsymbol{\mu}}}_{\mathscr{U}}^{\top}\mathbf{1}-\underline{\hat{\boldsymbol{\mu}}}_{\mathfrak{s}}^{\top}\mathbf{d}\right)$$

Replacing $\underline{\widehat{\boldsymbol{\mu}}}$ with $\underline{\boldsymbol{\mu}}$ gives

$$\hat{\mathbf{t}}_y^{mc}=\hat{\boldsymbol{t}}_y+\widehat{\mathbf{B}}_{\underline{\boldsymbol{\mu}}\mathbf{Y}}\left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^{\top}\mathbf{1}-\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\mathbf{d}\right)+o_p\left(n^{-\frac{1}{2}}\right)$$

Furthermore, $\hat{\mathbf{B}}_{\underline{\boldsymbol{\mu}}\mathbf{Y}} = O_p(1)$. Thus,

$$\begin{aligned}
\hat{\mathbf{t}}_y^{mc} &= \hat{\boldsymbol{t}}_y + \mathbf{B}_{\underline{\boldsymbol{\mu}}\mathbf{Y}}\left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top \mathbf{1} - \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \mathbf{d}\right) + o_p\left(n^{-\frac{1}{2}}\right) \\
&= \hat{\boldsymbol{t}}_y - \mathbf{B}_{\underline{\boldsymbol{\mu}}\mathbf{Y}}\left(\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \mathbf{d}\right) + \mathbf{B}_{\underline{\boldsymbol{\mu}}\mathbf{Y}}\left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top \mathbf{1}\right) + o_p\left(n^{-\frac{1}{2}}\right) \\
&= \left[\mathbf{y}^\top - \mathbf{B}_{\underline{\boldsymbol{\mu}}\mathbf{Y}}\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top\right]\mathbf{d} + \mathbf{B}_{\underline{\boldsymbol{\mu}}\mathbf{Y}}\left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top \mathbf{1}\right) + o_p\left(n^{-\frac{1}{2}}\right) \\
&= \mathbf{e}^\top \mathbf{d} + \mathbf{B}_{\underline{\boldsymbol{\mu}}\mathbf{Y}}\left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top \mathbf{1}\right) + o_p\left(n^{-\frac{1}{2}}\right)
\end{aligned}$$

where $\mathbf{e} = \mathbf{y} - \mathbf{B}_{\underline{\boldsymbol{\mu}}\mathbf{Y}}\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top$.

Since $\mathbf{B}_{\underline{\boldsymbol{\mu}}\mathbf{Y}}\left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top \mathbf{1}\right)$ is a constant with respect to the sample design, the asymptotic variance of $\hat{\mathbf{t}}_y^{mc}$ will be equivalent to the asymptotic variance of a weighted sum of $\mathbf{e}_k$. In this way, we can use the formulas for the variance of the Horvitz-Thompson estimator in clustered designs by substituting $\mathbf{y}_k$ with $\mathbf{e}_k$. So, as long as our sample design is independent and invariant, the asymptotic variance is

$$\begin{aligned}
\mathrm{av}_{II}\left(\hat{\mathbf{t}}_y^{mc}\right) &= \mathrm{var}\left(\mathbf{e}^\top \mathbf{d}\right) \\
&= \mathrm{var}\left(\sum_{i\in\mathfrak{s}_I}\left(d_i\widehat{\mathbf{t}}_{ei}\right)\right) \\
&= \mathrm{var}\left[\mathrm{E}\left[\sum_{i\in\mathfrak{s}_I}\left(d_i\widehat{\mathbf{t}}_{ei}\right)|\mathfrak{s}_i\right]|\mathfrak{s}_i\right] + \mathrm{E}\left[\mathrm{var}\left[\sum_{i\in\mathfrak{s}_i}\left(d_i\widehat{\mathbf{t}}_{ei}\right)|\mathfrak{s}_I\right]|\mathfrak{s}_i\right] \\
&= \sum_{i\in\mathfrak{s}_I}\mathrm{var}\left(d_i\mathbf{t}_{ei}\right)|\mathfrak{s}_i + \mathrm{E}\left[\sum_{i\in\mathfrak{s}_I}\left(d_i^2\mathrm{var}\left(\sum_{k\in\mathfrak{s}_i}d_{k|i}\mathbf{e}_{k|i}|\mathfrak{s}_I\right)\right)|\mathfrak{s}_i\right] \\
&= \sum_{i\in\mathscr{U}_I}\sum_{j\in\mathscr{U}_I}\left(\Delta_{ij}d_id_j\mathbf{t}_{ei}\mathbf{t}_{ej}^\top\right) + \mathrm{E}\left[\sum_{i\in\mathfrak{s}_I}\left(d_i^2\left(\sum_{k\in\mathscr{U}_i}\sum_{l\in\mathscr{U}_i}\Delta_{kl|i}d_{k|i}d_{l|i}\mathbf{e}_{k|i}\mathbf{e}_{l|i}^\top\right)\right)|\mathfrak{s}_i\right] \\
&= \sum_{i\in\mathscr{U}_I}\sum_{j\in\mathscr{U}_I}\left(\Delta_{ij}d_id_j\mathbf{t}_{ei}\mathbf{t}_{ej}^\top\right) + \sum_{i\in\mathscr{U}_I}\left[d_i\left(\sum_{k\in\mathscr{U}_i}\sum_{l\in\mathscr{U}_i}\Delta_{kl|i}d_{k|i}d_{l|i}\mathbf{e}_k\mathbf{e}_l^\top\right)\right]
\end{aligned} \tag{B.24}$$

where

$$\mathbf{t}_{ei} = \sum_{k \in \mathscr{U}_i} \mathbf{e}_k. \tag{B.25}$$

For category c, we write the variance as

$$\operatorname{av}\left(\hat{\mathbf{t}}_{yc}^{mc}\right) = \sum \sum_{\mathscr{U}_I} \Delta_{ij} \frac{t_{e_{ci}}}{\pi_i} \frac{t_{e_{cj}}}{\pi_j} + \sum_{\mathscr{U}_I} \frac{\sum_{\mathscr{U}_i} \sum_{\mathscr{U}_i} \Delta_{kl|i} \frac{e_{ck}}{\pi_{k|i}} \frac{e_{cl}}{\pi_{l|i}}}{\pi_i}.$$

B.5.5 Variance estimators of model calibration

B.5.5.1 Weighted variance estimator

If we simply estimate the totals in Equation B.24 on page 345, we get an estimator for the asymptotic variance of $\hat{\mathbf{t}}_y^{mc}$,

$$\upsilon_e\left(\hat{\mathbf{t}}_y^{mc}\right) = \sum_{i \in \mathfrak{s}_I} \sum_{j \in \mathfrak{s}_I} \left(d_{ij} \Delta_{ij} d_i d_j \hat{\mathbf{t}}_{\hat{\mathbf{e}}i} \hat{\mathbf{t}}_{\hat{\mathbf{e}}j}^\top\right) + \sum_{i \in \mathfrak{s}_I} \left[d_i \left(\sum_{k \in \mathfrak{s}_i} \sum_{l \in \mathfrak{s}_i} d_{kl|i} \Delta_{kl|i} d_{k|i} d_{l|i} \hat{\mathbf{e}}_k \hat{\mathbf{e}}_l^\top \right) \right]$$

where

$$\hat{\mathbf{t}}_{\hat{\mathbf{e}}i} = \sum_{k \in \mathfrak{s}_i} d_{k|i} \hat{\mathbf{e}}_{k|i}$$

$$\hat{\mathbf{e}}_k = \mathbf{y}_k - \hat{\boldsymbol{\mu}}_k.$$

Särndal et al. (1989) argue that this estimator tends to underestimate the true sampling error in practice for single-staged samples. For this reason, Särndal et al. (1992) prefer a variant of υ_e based on an adjustment to the residuals.

Using the weighted residual technique advocated in Särndal et al. (1989), we replace $\hat{\mathbf{e}}_k$ with $g_k \hat{\mathbf{e}}_k$, where g_k is the kth element in the vector

$$\mathbf{g} = \left[\mathbf{1} + \mathbf{Q} \underline{\hat{\boldsymbol{\mu}}}_{\mathfrak{s}} \left(\underline{\hat{\boldsymbol{\mu}}}_{\mathfrak{s}}^\top \left(\mathbf{Q} \boldsymbol{\Pi}^{-1} \right) \underline{\hat{\boldsymbol{\mu}}}_{\mathfrak{s}} \right)^{-1} \left(\underline{\hat{\boldsymbol{\mu}}}_{\mathscr{U}}^\top \mathbf{1} - \underline{\hat{\boldsymbol{\mu}}}_{\mathfrak{s}}^\top \mathbf{d} \right) \right].$$

That is

$$g_k = \left[1 + q_k \underline{\hat{\boldsymbol{\mu}}}_k^\top \left[\sum_{k\in\mathfrak{s}} d_k q_k^{-1} \underline{\hat{\boldsymbol{\mu}}}_k \underline{\hat{\boldsymbol{\mu}}}_k^\top\right]^{-1} \left[\sum_{k\in\mathscr{U}} \underline{\hat{\boldsymbol{\mu}}}_k - \sum_{k\in\mathfrak{s}} d_k \underline{\hat{\boldsymbol{\mu}}}_k\right]\right].$$

Thus, the weighted residual estimator is

$$\upsilon_g\left(\hat{\mathbf{t}}_y^{mc}\right) = \sum_{i\in\mathfrak{s}_I}\sum_{j\in\mathfrak{s}_I}\left(d_{ij}\Delta_{ij}d_i d_j \hat{\mathbf{t}}_{g\hat{\mathbf{e}}i}\hat{\mathbf{t}}_{g\hat{\mathbf{e}}j}^\top\right) + \sum_{i\in\mathfrak{s}_I}\left[d_i\left(\sum_{k\in\mathfrak{s}_i}\sum_{l\in\mathfrak{s}_i} d_{kl|i}\Delta_{kl|i}d_{k|i}d_{l|i}g_k\hat{\mathbf{e}}_k g_l\hat{\mathbf{e}}_l^\top\right)\right]$$

where

$$\hat{t}_{g\hat{\mathbf{e}}i} = \sum_{\mathfrak{s}_i}\frac{g_k\hat{\mathbf{e}}_k}{\pi_{k|i}}.$$

If the first and second stage samples are selected using a Poisson sampling technique, then $\upsilon_g\left(\hat{\mathbf{t}}_y^{mc}\right)$ reduces to

$$\upsilon_g\left(\hat{\mathbf{t}}_y^{mc}\right) = \sum_{i\in\mathfrak{s}_I}\frac{(1-\pi_i)}{\pi_i^2}\hat{\mathbf{t}}_{g\hat{\mathbf{e}}i}\hat{\mathbf{t}}_{g\hat{\mathbf{e}}i}^\top + \sum_{i\in\mathfrak{s}_I}\frac{1}{\pi_i}\sum_{k\in\mathfrak{s}_i}\frac{\left(1-\pi_{k|i}\right)}{\pi_{k|i}^2}g_k^2\hat{\mathbf{e}}_k\hat{\mathbf{e}}_k^\top$$

B.5.5.2 With-replacement estimator

Commonly, with-replacement variance estimators are used even when the first stage sample is selected without-replacement. As long as the first-stage sampling fraction is relatively small, the bias of using a with-replacement variance estimator is relatively small. Särndal et al. (1992, sec 4.6) discuss the classic with-replacement variance estimator of a total under multiple stages of sampling.

For estimating the variance of the Horvitz-Thompson estimator in a clustered design, the with-replacement variance estimator is Equation (B.20),

$$\upsilon_{wr}\left(\widehat{t}_y^\pi\right) = \frac{1}{n(n-1)}\sum_{i\in\mathfrak{s}_I}^{n}\left(\frac{\widehat{t}_{yi}^\pi}{p_i} - \widehat{t}_y^\pi\right)^\top\left(\frac{\widehat{t}_{yi}^\pi}{p_i} - \widehat{t}_y^\pi\right)$$

where $p_k = \frac{\pi_k}{n}$ is the probability of drawing the ith primary sampling unit in single draw and n is the total number of primary sampling sample units. We can modify Equation (B.20) on page 324 for the Model Calibration Estimator by replacing $\frac{\widehat{t}^{\pi}_{yi}}{p_i}$ with $\widehat{\mathbf{t}}^{\pi}_{\widehat{\mathbf{e}}i} = \sum_{k \in \mathfrak{s}_i} \left(\frac{\widehat{\mathbf{e}}_{k|i}}{p_i} \right)$. This substitution is motivated by Equation (B.24) on page 345. Now consider

$$
\begin{aligned}
\upsilon_{wr}\left(\widehat{\mathbf{t}}^{mc}_{y}\right) &= \frac{1}{n\left(n-1\right)} \sum_{i \in \mathfrak{s}_I} \left(\frac{\widehat{\mathbf{t}}^{\pi}_{\widehat{\mathbf{e}}i}}{p_i} - \widehat{\mathbf{t}}^{\pi}_{\widehat{\mathbf{e}}} \right) \left(\frac{\widehat{\mathbf{t}}^{\pi}_{\widehat{\mathbf{e}}i}}{p_i} - \widehat{\mathbf{t}}^{\pi}_{\widehat{\mathbf{e}}} \right)^{\top} \\
&= \frac{1}{n\left(n-1\right)} \sum_{i \in \mathfrak{s}_I} \left(\frac{n\widehat{\mathbf{t}}^{\pi}_{\widehat{\mathbf{e}}i}}{\pi_i} - \widehat{\mathbf{t}}^{\pi}_{\widehat{\mathbf{e}}} \right) \left(\frac{n\widehat{\mathbf{t}}^{\pi}_{\widehat{\mathbf{e}}i}}{\pi_i} - \widehat{\mathbf{t}}^{\pi}_{\widehat{\mathbf{e}}} \right)^{\top} \\
&= \frac{n^2}{n\left(n-1\right)} \sum_{i \in \mathfrak{s}_I} \left(\frac{\widehat{\mathbf{t}}^{\pi}_{\widehat{\mathbf{e}}i}}{\pi_i} - \frac{1}{n}\widehat{\mathbf{t}}^{\pi}_{\widehat{\mathbf{e}}} \right) \left(\frac{\widehat{\mathbf{t}}^{\pi}_{\widehat{\mathbf{e}}i}}{\pi_i} - \frac{1}{n}\widehat{\mathbf{t}}^{\pi}_{\widehat{\mathbf{e}}} \right)^{\top} \\
&= \frac{n}{\left(n-1\right)} \sum_{i \in \mathfrak{s}_I} \left(\frac{\widehat{\mathbf{t}}^{\pi}_{\widehat{\mathbf{e}}i}}{\pi_i} - \frac{1}{n}\widehat{\mathbf{t}}^{\pi}_{\widehat{\mathbf{e}}} \right) \left(\frac{\widehat{t}^{\pi}_{\mathbf{e}i}}{\pi_i} - \frac{1}{n}\widehat{\mathbf{t}}^{\pi}_{\widehat{\mathbf{e}}} \right)^{\top} \\
&= \frac{n}{\left(n-1\right)} \sum_{i \in \mathfrak{s}_I} \left[d_i \sum_{k \in \mathfrak{s}_i} \left(d_{k|i} \hat{\mathbf{e}}_{k|i} \right) - \frac{1}{n} \sum_{k \in \mathfrak{s}} \left(d_k \hat{\mathbf{e}}_k \right) \right] \left[d_i \sum_{k \in \mathfrak{s}_i} \left(d_{k|i} \hat{\mathbf{e}}_{k|i} \right) - \frac{1}{n} \sum_{k \in \mathfrak{s}} \left(d_k \hat{\mathbf{e}}_k \right) \right]^{\top}
\end{aligned}
\tag{B.26}
$$

B.5.5.3 Implicit differentiation estimator

Introduction

In this section, we derive the implicit differentiation estimator for the model calibrated estimator. We follow a similar process to the construction of the implicit differentiation estimator for the LGREG in Appendix B.4.3.3 on page 325. We begin by repeating much of the text in Appendix B.4.3.3, but then move into the unique derivation for the model calibrated estimator.

First described by Binder (1983), implicit differentiation uses linearization and es-

timating equations to produce design-consistent estimators of finite population parameters. Implicit differentiation is especially useful when the parameter of interest cannot be solved explicitly in closed form. Both Binder (1983) and Särndal et al. (1992)[section 13.4] give several examples of how implicit differentiation can be used to construct design-consistent estimators of $\mathbf{B}$ from a logistic regression model. However, no authors have use the implicit differentiation method to construct variance estimators of model calibrated totals. The derivative of the model calibration estimator with respect to $\mathbf{B}$ is significantly more involved and complex than the LGREG derivatives.

Parameters

For the multivariate model calibration estimator with the logit link, we begin by defining a vector with our parameters of interest

$$\underset{C+(C-1)\cdot p\times 1}{\boldsymbol{\theta}} = \begin{bmatrix} \underset{C\times 1}{\mathbf{t}_y^{mc}} \\ \\ \underset{(C-1)\cdot p\times 1}{\mathrm{vec}\left(\mathbf{B}\right)} \end{bmatrix}.$$

Unless the complete population is measured without error, our parameter vector is unknown. Denote an estimate of the parameter vector as

$$\widehat{\boldsymbol{\theta}} = \begin{bmatrix} \widehat{\mathbf{t}}_y^{mc} \\ \\ \mathrm{vec}\left(\widehat{\mathbf{B}}\right) \end{bmatrix}.$$

The model calibration estimator of a finite population total is

$$\widehat{\mathbf{t}}_y^{mc} = \mathbf{y}^\top \left[\mathbf{d} + \left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right)\underline{\boldsymbol{\mu}}_{\mathscr{S}}\left(\underline{\boldsymbol{\mu}}_{\mathscr{S}}^\top\left(\mathbf{Q}\boldsymbol{\Pi}^{-1}\right)\underline{\boldsymbol{\mu}}_{\mathscr{S}}\right)^{-1}\left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top\mathbf{1} - \underline{\boldsymbol{\mu}}_{\mathscr{S}}^\top\mathbf{d}\right)\right].$$

If we set $\mathbf{Q}^{-1} = \mathbf{I}$ and make some substitutions, we have

$$\begin{aligned}\widehat{\mathbf{t}}_y^{mc} &= \widehat{\mathbf{t}}_y + \mathbf{y}_{\mathfrak{s}}^\top \widehat{\underline{\boldsymbol{\mu}}}_{\mathfrak{s}} \left[\widehat{\mathbf{A}}\left(\mathbf{B}\right)\right]^{-1} \left(\mathbf{t}_{\underline{\boldsymbol{\mu}}} - \widehat{\mathbf{t}}_{\underline{\boldsymbol{\mu}}}\left(\mathbf{B}\right)\right) \\ &= \sum_{\mathfrak{s}} d_k \mathbf{y}_k \left[1 + \widehat{\underline{\boldsymbol{\mu}}}_k^\top \left[\widehat{\mathbf{A}}\left(\mathbf{B}\right)\right]^{-1} \left(\mathbf{t}_{\underline{\boldsymbol{\mu}}} - \widehat{\mathbf{t}}_{\underline{\boldsymbol{\mu}}}\left(\mathbf{B}\right)\right)\right].\end{aligned}$$

Population Estimating Equations

To motivate the implicit differentiation estimator, we write our estimator when every unit in the finite population is included in the sample. This is the ideal situation when all population quantities are known and a full census of the population is taken. In this case, $\mathscr{U} = \mathfrak{s}$ and $d_k = 1$ for all k and we can write our estimator as

$$\mathbf{t}_y^{mc} = \sum_{\mathscr{U}} \mathbf{y}_k \left[1 + \underline{\boldsymbol{\mu}}_k^\top \mathbf{A}^{-1} \left(\mathbf{t}_{\underline{\boldsymbol{\mu}}} - \mathbf{t}_{\underline{\boldsymbol{\mu}}}\right)\right].$$

Of course in this ideal situation, $\mathbf{t}_y^{mc}$ reduces to $\mathbf{t}_y$ and no estimation is necessary. Nevertheless, we use this equation to motivate our sample estimating equations. An estimating equation for $\mathbf{t}_y^{mc}$ is

$$\mathbf{0} = \sum_{\mathscr{U}} \mathbf{y}_k \left[1 + \underline{\boldsymbol{\mu}}_k^\top \mathbf{A}^{-1} \left(\mathbf{t}_{\underline{\boldsymbol{\mu}}} - \mathbf{t}_{\underline{\boldsymbol{\mu}}}\right)\right] - \mathbf{t}_y^{mc}.$$

The coefficient vector for category c, called $\mathbf{B}_c$, is the solution to

$$\mathbf{0} = \sum_{\mathscr{U}} \left[y_{kc} - \frac{z_k e^{\mathbf{B}_c^\top \mathbf{x}_k}}{1 + \sum_{c=1}^{C-1} e^{\mathbf{B}_c^\top \mathbf{x}_k}}\right] \mathbf{x}_k.$$

Thus, our parameter vector is the solution to $\mathbf{W}(\boldsymbol{\theta}) = \mathbf{0}$ where

$$
\begin{aligned}
\underset{C+(C-1)\cdot p\times 1}{\mathbf{W}(\boldsymbol{\theta})} &= \left(\sum_{\mathscr{U}} \mathbf{U}_k(\boldsymbol{\theta})\right) - \mathbf{v} \\
&= \sum_{\mathscr{U}} \begin{bmatrix} \mathbf{y}_k \left[1 + \underline{\boldsymbol{\mu}}_k^\top \mathbf{A}^{-1}\left(\mathbf{t}_{\underline{\boldsymbol{\mu}}} - \mathbf{t}_{\underline{\boldsymbol{\mu}}}\right)\right] \\ \left(y_{k1} - \frac{z_k e^{\mathbf{B}_1^\top \mathbf{x}_k}}{1+\sum_{c=1}^{C-1} e^{\mathbf{B}_c^\top \mathbf{x}_k}}\right)\mathbf{x}_k \\ \vdots \\ \left(y_{kC-1} - \frac{z_k e^{\mathbf{B}_{C-1}^\top \mathbf{x}_k}}{1+\sum_{c=1}^{C-1} e^{\mathbf{B}_c^\top \mathbf{x}_k}}\right)\mathbf{x}_k \end{bmatrix} - \begin{bmatrix} \mathbf{t}_y^{mc} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.
\end{aligned}
$$

Survey Weighted Estimating Equations

We only measure $\mathbf{y}_k$ for the sample units. Thus, we cannot compute $\mathbf{W}(\boldsymbol{\theta})$. Yet, we can estimate $\mathbf{W}(\boldsymbol{\theta})$ from our sample. The weighted estimate of our estimating equations is

$$
\begin{aligned}
\underset{C+(C-1)\cdot p\times 1}{\widehat{\mathbf{W}}(\boldsymbol{\theta})} &= \left(\sum_{\mathfrak{s}} \widehat{\mathbf{U}}_k(\boldsymbol{\theta})\right) + \mathbf{v} \\
&= \sum_{\mathfrak{s}} \begin{bmatrix} d_k\mathbf{y}_k \left[1 + \underline{\boldsymbol{\mu}}_k^\top \left[\widehat{\mathbf{A}}(\mathbf{B})\right]^{-1}\left(\mathbf{t}_{\underline{\boldsymbol{\mu}}} - \widehat{\mathbf{t}}_{\underline{\boldsymbol{\mu}}}(\mathbf{B})\right)\right] \\ d_k\left(y_{k1} - \frac{z_k e^{\mathbf{B}_1^\top \mathbf{x}_k}}{1+\sum_{c=1}^{C-1} e^{\mathbf{B}_c^\top \mathbf{x}_k}}\right)\mathbf{x}_k \\ \vdots \\ d_k\left(y_{kC-1} - \frac{z_k e^{\mathbf{B}_{C-1}^\top \mathbf{x}_k}}{1+\sum_{c=1}^{C-1} e^{\mathbf{B}_c^\top \mathbf{x}_k}}\right)\mathbf{x}_k \end{bmatrix} - \begin{bmatrix} \mathbf{t}_y^{mc} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.
\end{aligned}
$$

The value of $\boldsymbol{\theta}$ that solves the estimating equations, $\widehat{\mathbf{W}}(\boldsymbol{\theta})$, is denoted $\widehat{\boldsymbol{\theta}}$. That is,

$$\underset{C+(C-1)\cdot p\times 1}{\widehat{\mathbf{W}}\left(\widehat{\boldsymbol{\theta}}\right)} = \left(\sum_{\mathfrak{s}} \widehat{\mathbf{U}}_k\left(\widehat{\boldsymbol{\theta}}\right)\right) + \mathbf{v}$$

$$= \sum_{\mathfrak{s}} \begin{bmatrix} d_k \mathbf{y}_k \left[1 + \widehat{\underline{\boldsymbol{\mu}}}_k^\top \left[\widehat{\mathbf{A}}\left(\widehat{\mathbf{B}}\right)\right]^{-1} \left(\mathbf{t}_{\underline{\boldsymbol{\mu}}} - \widehat{\mathbf{t}}_{\underline{\boldsymbol{\mu}}}\left(\widehat{\mathbf{B}}\right)\right)\right] \\ d_k\left(y_{k1} - \frac{z_k e^{\widehat{\mathbf{B}}_1^\top \mathbf{x}_k}}{1+\sum_{c=1}^{C-1} e^{\widehat{\mathbf{B}}_c^\top \mathbf{x}_k}}\right)\mathbf{x}_k \\ \vdots \\ d_k\left(y_{kC-1} - \frac{z_k e^{\widehat{\mathbf{B}}_{C-1}^\top \mathbf{x}_k}}{1+\sum_{c=1}^{C-1} e^{\widehat{\mathbf{B}}_c^\top \mathbf{x}_k}}\right)\mathbf{x}_k \end{bmatrix} - \begin{bmatrix} \widehat{\mathbf{t}}_y^{mc} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Simultaneously solving for $\mathrm{vec}(\mathbf{B})$ and $\mathbf{t}_y^{mc}$ has the advantage that it simplifies variance estimation. Moreover, it results in the complete covariance matrix containing the estimated covariances between $\hat{\mathbf{t}}_y^{mc}$ and $\mathrm{vec}\left(\widehat{\mathbf{B}}\right)$.

Derivation of Variance Estimator

We now turn to estimating the variance of $\widehat{\boldsymbol{\theta}}$. Under mild regularity conditions, a linear approximation of our estimating equations is

$$\widehat{\mathbf{W}}\left(\widehat{\boldsymbol{\theta}}\right) \approx \widehat{\mathbf{W}}(\boldsymbol{\theta}) + \widehat{\mathbf{J}}(\boldsymbol{\theta})\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = \mathbf{0}$$

where

$$\underset{(C+C\cdot p-p)\times(C+C\cdot p-p)}{\widehat{\mathbf{J}}(\boldsymbol{\theta})} = \frac{\partial}{\partial\left(\mathrm{vec}\boldsymbol{\theta}\right)^\top}\widehat{\mathbf{W}}(\boldsymbol{\theta}).$$

According to Lemma 1 in Binder (1983), we can also write our linearizaiton as

$$\widehat{\mathbf{W}}\left(\widehat{\boldsymbol{\theta}}\right) \approx \widehat{\mathbf{W}}(\boldsymbol{\theta}) + \mathbf{J}(\boldsymbol{\theta})\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = \mathbf{0}$$

where

$$\mathbf{J}(\boldsymbol{\theta}) = \frac{\partial}{\partial\left(\mathrm{vec}\boldsymbol{\theta}\right)^\top}\mathbf{W}(\boldsymbol{\theta}).$$

Using our linearization, we can compute an asymptotic variance estimator

$$\begin{aligned}
\widehat{\mathbf{W}}\left(\widehat{\boldsymbol{\theta}}\right) &\approx \widehat{\mathbf{W}}\left(\boldsymbol{\theta}\right) + \mathbf{J}\left(\boldsymbol{\theta}\right)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) = \mathbf{0} \\
-\mathbf{J}\left(\boldsymbol{\theta}\right)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) &\approx \widehat{\mathbf{W}}\left(\boldsymbol{\theta}\right) \\
\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) &\approx -\mathbf{J}^{-1}\left(\boldsymbol{\theta}\right)\widehat{\mathbf{W}}\left(\boldsymbol{\theta}\right) \\
\operatorname{var}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) &\approx \mathbf{J}^{-1}\left(\boldsymbol{\theta}\right)\operatorname{var}\left[\widehat{\mathbf{W}}\left(\boldsymbol{\theta}\right)\right]\left[\mathbf{J}^{-1}\left(\boldsymbol{\theta}\right)\right]^{\top} \\
\operatorname{var}\left(\widehat{\boldsymbol{\theta}}\right) &\approx \mathbf{J}^{-1}\left(\boldsymbol{\theta}\right)\left[\boldsymbol{\Sigma}\left(\boldsymbol{\theta}\right)\right]\left[\mathbf{J}^{-1}\left(\boldsymbol{\theta}\right)\right]^{\top} \\
&\approx \mathrm{V}\left(\boldsymbol{\theta}\right).
\end{aligned}$$

Because the asymptotic variance of $\widehat{\boldsymbol{\theta}}$ is a function of $\boldsymbol{\theta}$, we write our asymptotic variance as $\mathrm{V}\left(\boldsymbol{\theta}\right)$. We note that $\mathbf{J}\left(\boldsymbol{\theta}\right)$ must be invertible. Furthermore, $\boldsymbol{\Sigma}\left(\boldsymbol{\theta}\right) = \operatorname{var}\left[\widehat{\mathbf{W}}\left(\boldsymbol{\theta}\right)\right]$ and $\widehat{\boldsymbol{\Sigma}}\left(\boldsymbol{\theta}\right)$ is an estimate of the design-based variance of $\widehat{\mathbf{W}}\left(\boldsymbol{\theta}\right)$. That is, $\widehat{\boldsymbol{\Sigma}}\left(\boldsymbol{\theta}\right) = \upsilon\left[\widehat{\mathbf{W}}\left(\boldsymbol{\theta}\right)\right]$.

We usually do not know $\boldsymbol{\theta}$ nor $\mathbf{J}$; thus, we substitute them for estimated quantities.

$$\upsilon_{Binder}\left(\widehat{\boldsymbol{\theta}}\right) = \left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right]\left[\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right)\right]\left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{\top}.$$

Partitioning the Jacobian into Blocks

To simplify our variance estimator, we must simplify $\widehat{\mathbf{J}} = \frac{\partial}{\partial \mathrm{vec}\boldsymbol{\theta}}\widehat{\mathbf{W}}\left(\boldsymbol{\theta}\right)$. We first partition $\widehat{\mathbf{J}}$ into four blocks,

$$\widehat{\mathbf{J}} = \begin{bmatrix} \mathscr{A} & \mathscr{B} \\ \mathscr{C} & \mathscr{D} \end{bmatrix}$$

where

$$
\begin{aligned}
\underset{C\times C}{\mathscr{A}} &= \frac{\partial}{\partial\left(\text{vec}\widehat{\mathbf{t}}_y^{mc}\right)^\top}\sum_{\mathfrak{s}}\left[d_k\mathbf{y}_k\left[1+\underline{\boldsymbol{\mu}}_k^\top\widehat{\mathbf{A}}^{-1}\left(\mathbf{t}_{\underline{\boldsymbol{\mu}}}-\widehat{\mathbf{t}}_{\underline{\boldsymbol{\mu}}}\right)\right]-\frac{1}{n}\widehat{\mathbf{t}}_y^{mc}\right] \\
\underset{C\times(C\cdot p-p)}{\mathscr{B}} &= \frac{\partial}{\partial\left(\text{vec}\mathbf{B}\right)^\top}\sum_{\mathfrak{s}}\left[d_k\mathbf{y}_k\left[1+\underline{\boldsymbol{\mu}}_k^\top\widehat{\mathbf{A}}^{-1}\left(\mathbf{t}_{\underline{\boldsymbol{\mu}}}-\widehat{\mathbf{t}}_{\underline{\boldsymbol{\mu}}}\right)\right]-\frac{1}{n}\widehat{\mathbf{t}}_y^{mc}\right] \\
\underset{(C\cdot p-p)\times C}{\mathscr{C}} &= \frac{\partial}{\partial\left(\text{vec}\widehat{\mathbf{t}}_y^{mc}\right)^\top}\sum_{\mathfrak{s}}\left[d_k\left(y_{k1}-\frac{z_k e^{\mathbf{B}_1^\top\mathbf{x}_k}}{1+\sum_{c=1}^{C-1}e^{\mathbf{B}_c^\top\mathbf{x}_k}}\right)\mathbf{x}_k\right] \\
\underset{(C\cdot p-p)\times(C\cdot p-p)}{\mathscr{D}} &= \frac{\partial}{\left(\text{vec}\partial\mathbf{B}\right)^\top}\sum_{\mathfrak{s}}\left[d_k\left(y_{k1}-\frac{z_k e^{\mathbf{B}_1^\top\mathbf{x}_k}}{1+\sum_{c=1}^{C-1}e^{\mathbf{B}_c^\top\mathbf{x}_k}}\right)\mathbf{x}_k\right]
\end{aligned}
$$

By the same logic in Appendix B.4.3.3 on page 325, we see that

$$
\mathscr{A} = -\mathbf{I}_C
$$

$$
\mathscr{C} = \mathbf{0}
$$

and

$$
\mathscr{D} = -\sum_{\mathfrak{s}} d_k z_k \mathbf{X}_k^\top\left[\text{diag}\left(\boldsymbol{\mathfrak{p}}_k\right)-\boldsymbol{\mathfrak{p}}_k\boldsymbol{\mathfrak{p}}_k^\top\right]\mathbf{X}_k
$$

We now turn our attention to $\mathscr{C}$.

Derivative of estimating equation for $\mathbf{t}_y^{mc}$ with respect to vec $(\mathbf{B})$

To calculate our derivative, we first apply the chain rule. Then we apply the product rule and simplify the three components of the product rule.

First, let

$$
\mathbf{f}\left(\underline{\boldsymbol{\mu}}\right) = \sum_{k\in\mathfrak{s}}\mathbf{f}_k\left(\underline{\boldsymbol{\mu}}\right)
$$

where

$$\mathbf{f}_k\left(\underline{\boldsymbol{\mu}}\right) = \mathbf{y}_k\left[d_k + \frac{d_k}{q_k}\underline{\boldsymbol{\mu}}_k^\top\left(\sum_{\mathfrak{s}}\frac{d_k}{q_k}\underline{\boldsymbol{\mu}}_k\underline{\boldsymbol{\mu}}_k^\top\right)^{-1}\left(\sum_{\mathscr{U}}\underline{\boldsymbol{\mu}}_k - \sum_{\mathfrak{s}} d_k\underline{\boldsymbol{\mu}}_k\right)\right].$$

Note that $\underline{\boldsymbol{\mu}}_k$, $\mathbf{f}_k\left(\underline{\boldsymbol{\mu}}\right)$, and $\mathbf{f}\left(\underline{\boldsymbol{\mu}}\right)$ are all C column vectors. Using the chain rule

$$\begin{aligned}\mathscr{B} &= \frac{\partial \mathbf{f}\left(\underline{\boldsymbol{\mu}}\right)}{\partial\left(\mathrm{vec}\mathbf{B}\right)^\top}\\ &= \sum_{k\in\mathfrak{s}}\frac{\partial \mathbf{f}_k\left(\underline{\boldsymbol{\mu}}\right)}{\partial\left(\mathrm{vec}\underline{\boldsymbol{\mu}}\right)^\top}\frac{\partial \mathrm{vec}\left(\underline{\boldsymbol{\mu}}\right)}{\partial\left(\mathrm{vec}\mathbf{B}\right)^\top}.\end{aligned}$$

We now simplify the derivative on the right. Following that, we simplify $\frac{\partial \mathbf{f}_k\left(\underline{\boldsymbol{\mu}}\right)}{\partial\left(\mathrm{vec}\underline{\boldsymbol{\mu}}\right)^\top}$.

Derivative of $\mathrm{vec}\left(\underline{\boldsymbol{\mu}}\right)$ **with respect to** $\left(\mathrm{vec}\mathbf{B}\right)^\top$

Since $\underline{\boldsymbol{\mu}}$ is a $n \times (C+1)$ matrix and $\mathbf{B}$ is a $(C-1)\times p$ matrix, our derivative will have $p\,(C-1)$ rows and $n\,(C+1)$ columns.

Recall that the first column of $\underline{\boldsymbol{\mu}}$ is $\mathbf{1}$. That is

$$\underline{\boldsymbol{\mu}} = \begin{bmatrix}\mathbf{1} & \boldsymbol{\mu}\end{bmatrix}$$

Since, $\mathbf{1}$ is a constant with respect to $\mathrm{vec}\left(\mathbf{B}\right)$, the derivative on the right simplifies to

$$\begin{aligned}\frac{\partial \mathrm{vec}\left(\underline{\boldsymbol{\mu}}\right)}{\partial\left(\mathrm{vec}\mathbf{B}\right)^\top} &= \frac{\partial}{\partial\left(\mathrm{vec}\mathbf{B}\right)^\top}\begin{bmatrix}\mathbf{1}\\ \boldsymbol{\mu}\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{0}\\ \frac{\partial}{\partial\left(\mathrm{vec}\mathbf{B}\right)^\top}\boldsymbol{\mu}\end{bmatrix}\end{aligned}$$

The derivative of $\boldsymbol{\mu}$ with respect to $\mathbf{B}$ was shown in Equation B.23 on page 333. That is

$$\frac{\partial \boldsymbol{\mu}_k}{\partial\left(\mathrm{vec}\mathbf{B}\right)^\top} = z_k\left[\mathrm{diag}\left(\mathfrak{p}_k\right) - \mathfrak{p}_k\mathfrak{p}_k^\top\right]\mathbf{X}_k$$

Thus the derivative of $\text{vec}\left(\underline{\boldsymbol{\mu}}\right)$ with respect to $\left(\text{vec}\mathbf{B}\right)^\top$ is

$$\frac{\partial \text{vec}\left(\underline{\boldsymbol{\mu}}\right)}{\partial\left(\text{vec}\mathbf{B}\right)^\top} = \begin{bmatrix} \mathbf{0} \\ \\ z_k\left[\text{diag}\left(\mathfrak{p}_k\right) - \mathfrak{p}_k\mathfrak{p}_k^\top\right]\mathbf{X}_k \end{bmatrix}$$

Derivative of $\mathbf{f}_k\left(\underline{\boldsymbol{\mu}}\right)$ with respect to $\left(\text{vec}\underline{\boldsymbol{\mu}}\right)^\top$

Since $\mathbf{f}_k\left(\underline{\boldsymbol{\mu}}\right)$ is a $C \times 1$ matrix and $\mathbf{B}$ is a $(C-1) \times p$ matrix, our derivative will have $p\left(C-1\right)$ rows and C columns. Simplifying gives

$$\begin{aligned}
\frac{\partial \mathbf{f}_k\left(\underline{\boldsymbol{\mu}}\right)}{\partial\left(\text{vec}\underline{\boldsymbol{\mu}}\right)^\top} &= \frac{\partial}{\partial\left(\text{vec}\underline{\boldsymbol{\mu}}\right)^\top}\left\{\mathbf{y}_k\left[d_k + \frac{d_k}{q_k}\underline{\boldsymbol{\mu}}_k^\top\left(\sum_{\mathfrak{s}}\frac{d_k}{q_k}\underline{\boldsymbol{\mu}}_k\underline{\boldsymbol{\mu}}_k^\top\right)^{-1}\left(\sum_{\mathscr{U}}\underline{\boldsymbol{\mu}}_k - \sum_{\mathfrak{s}} d_k\underline{\boldsymbol{\mu}}_k\right)\right]\right\} \\
&= \left[\frac{\partial}{\partial\left(\text{vec}\underline{\boldsymbol{\mu}}\right)^\top}\frac{d_k}{q_k}\mathbf{y}_k\underline{\boldsymbol{\mu}}_k^\top\left(\sum_{\mathfrak{s}}\frac{d_k}{q_k}\underline{\boldsymbol{\mu}}_k\underline{\boldsymbol{\mu}}_k^\top\right)^{-1}\left(\sum_{\mathscr{U}}\underline{\boldsymbol{\mu}}_k - \sum_{\mathfrak{s}} d_k\underline{\boldsymbol{\mu}}_k\right)\right] \\
&= \frac{d_k}{q_k}\mathbf{y}_k\frac{\partial}{\partial\left(\text{vec}\underline{\boldsymbol{\mu}}\right)^\top}\underline{\boldsymbol{\mu}}_k^\top\left[\left(\sum_{\mathfrak{s}}\frac{d_k}{q_k}\underline{\boldsymbol{\mu}}_k\underline{\boldsymbol{\mu}}_k^\top\right)^{-1}\left(\sum_{\mathscr{U}}\underline{\boldsymbol{\mu}}_k - \sum_{\mathfrak{s}} d_k\underline{\boldsymbol{\mu}}_k\right)\right].
\end{aligned}$$

Now applying the product rule from Seber (2008)[p. 360]

$$\begin{aligned}
\frac{\partial \mathbf{f}_k\left(\underline{\boldsymbol{\mu}}\right)}{\partial\left(\text{vec}\underline{\boldsymbol{\mu}}\right)^\top} = \frac{d_k}{q_k}\mathbf{y}_k & \\
&\left\{\left(\sum_{\mathscr{U}}\underline{\boldsymbol{\mu}}_k - \sum_{\mathfrak{s}} d_k\underline{\boldsymbol{\mu}}_k\right)^\top\left[\left(\sum_{\mathfrak{s}}\frac{d_k}{q_k}\underline{\boldsymbol{\mu}}_k\underline{\boldsymbol{\mu}}_k^\top\right)^{-1}\right]^\top\frac{\partial\underline{\boldsymbol{\mu}}_k}{\partial\left(\text{vec}\underline{\boldsymbol{\mu}}\right)^\top}\right. \\
&+\left[\text{vec}\left(\underline{\boldsymbol{\mu}}_k\left(\sum_{\mathscr{U}}\underline{\boldsymbol{\mu}}_k - \sum_{\mathfrak{s}} d_k\underline{\boldsymbol{\mu}}_k\right)^\top\right)\right]^\top\frac{\partial\text{vec}\left[\left(\sum_{\mathfrak{s}}\frac{d_k}{q_k}\underline{\boldsymbol{\mu}}_k\underline{\boldsymbol{\mu}}_k^\top\right)^{-1}\right]}{\partial\left(\text{vec}\underline{\boldsymbol{\mu}}\right)^\top} \\
&\left.+\underline{\boldsymbol{\mu}}_k^\top\left(\sum_{\mathfrak{s}}\frac{d_k}{q_k}\underline{\boldsymbol{\mu}}_k\underline{\boldsymbol{\mu}}_k^\top\right)^{-1}\frac{\partial\left(\sum_{\mathscr{U}}\underline{\boldsymbol{\mu}}_k - \sum_{\mathfrak{s}} d_k\underline{\boldsymbol{\mu}}_k\right)}{\partial\left(\text{vec}\underline{\boldsymbol{\mu}}\right)^\top}\right\}
\end{aligned} \tag{B.27}$$

First Derivative

The derivative on the right hand side of the first term in Equation B.27 simplifies to

a $(C+1) \times n \cdot (C+1)$ matrix

$$\frac{\partial \underline{\boldsymbol{\mu}}_k}{\partial \left(\mathrm{vec}\underline{\boldsymbol{\mu}}\right)^\top} = \underset{(C+1)\times(C+1)}{\breve{\mathbf{I}}} \bigotimes \underset{1\times n}{\mathbf{I}_k}.$$

where $\breve{\mathbf{I}}$ is the identity matrix with the upper left element replaced with 0 and $\mathbf{I}_k$ is the kth row of the n by n identity matrix.

Second Derivative

We now simplify the derivative on the right side of the second term in Equation B.27. According to Seber (2008)[p. 363], our second derivative can be written as

$$\frac{\partial \mathrm{vec}\left[\left(\sum_{\mathfrak{s}} \frac{d_k}{q_k} \underline{\boldsymbol{\mu}}_k \underline{\boldsymbol{\mu}}_k^\top\right)^{-1}\right]}{\partial \underline{\boldsymbol{\mu}}^\top} = -\left[\left(\left(\sum_{\mathfrak{s}} \frac{d_k}{q_k} \underline{\boldsymbol{\mu}}_k \underline{\boldsymbol{\mu}}_k^\top\right)^{-1}\right)^\top \bigotimes \left(\sum_{\mathfrak{s}} \frac{d_k}{q_k} \underline{\boldsymbol{\mu}}_k \underline{\boldsymbol{\mu}}_k^\top\right)^{-1}\right] \mathbf{M}$$

where

$$\underset{(C+1)^2\times n(C+1)}{\mathbf{M}} = \frac{\partial \mathrm{vec}\left(\sum_{\mathfrak{s}} \frac{d_k}{q_k} \underline{\boldsymbol{\mu}}_k \underline{\boldsymbol{\mu}}_k^\top\right)}{\partial \underline{\boldsymbol{\mu}}^\top}. \tag{B.28}$$

Let $\mathbf{M}_{ck}$ be a $(C+1) \times (C+1)$ matrix and let i index the rows of $\mathbf{M}_{ck}$ and j index the columns of $\mathbf{M}_{ck}$. We define $\mathbf{M}_{ck}$ as

$$\begin{aligned} \mathbf{M}_{ck} &= \frac{\partial \left(\sum_{\mathfrak{s}} \frac{d_k}{q_k} \underline{\boldsymbol{\mu}}_k \underline{\boldsymbol{\mu}}_k^\top\right)}{\partial \mu_{ck}} \\ &= \frac{d_k}{q_k} \left(\underline{\boldsymbol{\mu}}_k \boldsymbol{\iota}_{ck}^\top + \boldsymbol{\iota}_{ck} \underline{\boldsymbol{\mu}}_k^\top\right) \end{aligned} \tag{B.29}$$

where $\boldsymbol{\iota}_{ck}$ is an $C-1$ dimensional vector with a 1 in the position of category c and zeros elsewhere. That is $\boldsymbol{\iota}_{ck}$ is the cth column of $\mathbf{I}_{C^\star}$. At the element level, $\mathbf{M}_{ck}$ can also be written as

$$M_{kc,ij} = \begin{cases} 2\frac{d_k}{q_k}\mu_{ck} & \text{where } i = c \text{ and } j = c \\ \frac{d_k}{q_k}\mu_{kj} & \text{where } i = c \text{ and } j \neq c \\ \frac{d_k}{q_k}\mu_{ki} & \text{where } i \neq c \text{ and } j = c \\ 0 & \text{where } i \neq c \text{ and } j \neq c \end{cases} \tag{B.30}$$

So, the ck^{th} column of $\frac{\partial \text{vec}\left[\left(\sum_{\mathfrak{s}} \frac{d_k}{q_k}\underline{\boldsymbol{\mu}}_k\underline{\boldsymbol{\mu}}^\top\right)^{-1}\right]}{\partial \underline{\boldsymbol{\mu}}_k^\top}$ will be $\text{vec}\left(\boldsymbol{M}_{ck}\right)$.

Third Derivative

Lastly, our final derivative in Equation B.27 on page 356 is an $C-1 \times n \cdot C-1$ matrix

$$\begin{aligned} \frac{\partial \sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}_k - \sum_{\mathfrak{s}} d_k \underline{\boldsymbol{\mu}}_k}{\partial \left(\text{vec}\underline{\boldsymbol{\mu}}\right)^\top} &= \underset{C-1\times C-1}{\mathbf{I}} \bigotimes \underset{n\times 1}{\mathbf{1}}^\top - \underset{C-1\times C-1}{\mathbf{I}} \bigotimes \underset{n\times 1}{\mathbf{d}}^\top \\ &= \mathbf{I} \bigotimes \left(\mathbf{1} - \mathbf{d}\right)^\top. \end{aligned}$$

Summary of Derivatives We decomposed our Jacobian into four blocks

$$\left[\widehat{\mathbf{J}}\left(\boldsymbol{\theta}\right)\right]^{-1} = \begin{bmatrix} \mathscr{A} & \mathscr{B} \\ \mathscr{C} & \mathscr{D} \end{bmatrix}.$$

Blocks $\mathscr{A}$, $\mathscr{C}$, and $\mathscr{D}$ were previously found in Appendix B.4.3.3. We showed that

$$
\begin{aligned}
\mathscr{B} &= \sum_{k\in\mathfrak{s}} \frac{\partial \mathbf{f}_k\left(\underline{\boldsymbol{\mu}}\right)}{\partial\left(\mathrm{vec}\underline{\boldsymbol{\mu}}\right)^\top} \frac{\partial \mathrm{vec}\left(\underline{\boldsymbol{\mu}}\right)}{\partial\left(\mathrm{vec}\mathbf{B}\right)^\top} \\
&= \sum_{k\in\mathfrak{s}} \frac{\partial \mathbf{f}_k\left(\underline{\boldsymbol{\mu}}\right)}{\partial\left(\mathrm{vec}\underline{\boldsymbol{\mu}}\right)^\top} \begin{bmatrix} \mathbf{0} \\ z_k\left[\mathrm{diag}\left(\mathfrak{p}_k\right) - \mathfrak{p}_k\mathfrak{p}_k^\top\right]\mathbf{X}_k \end{bmatrix} \\
&= \sum_{k\in\mathfrak{s}} \frac{d_k}{q_k}\mathbf{y}_k \Bigg\{ \left(\sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}_k - \sum_{\mathfrak{s}} d_k\underline{\boldsymbol{\mu}}_k\right)^\top \left[\left(\sum_{\mathfrak{s}} \frac{d_k}{q_k}\underline{\boldsymbol{\mu}}_k\underline{\boldsymbol{\mu}}_k^\top\right)^{-1}\right]^\top \frac{\partial \underline{\boldsymbol{\mu}}_k}{\partial\left(\mathrm{vec}\underline{\boldsymbol{\mu}}\right)^\top} \\
&\quad + \left[\mathrm{vec}\left(\underline{\boldsymbol{\mu}}_k\left(\sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}_k - \sum_{\mathfrak{s}} d_k\underline{\boldsymbol{\mu}}_k\right)^\top\right)\right]^\top \frac{\partial \mathrm{vec}\left[\left(\sum_{\mathfrak{s}} \frac{d_k}{q_k}\underline{\boldsymbol{\mu}}_k\underline{\boldsymbol{\mu}}_k^\top\right)^{-1}\right]}{\partial\left(\mathrm{vec}\underline{\boldsymbol{\mu}}\right)^\top} \\
&\quad + \underline{\boldsymbol{\mu}}_k^\top\left(\sum_{\mathfrak{s}} \frac{d_k}{q_k}\underline{\boldsymbol{\mu}}_k\underline{\boldsymbol{\mu}}_k^\top\right)^{-1} \frac{\partial\left(\sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}_k - \sum_{\mathfrak{s}} d_k\underline{\boldsymbol{\mu}}_k\right)}{\partial\left(\mathrm{vec}\underline{\boldsymbol{\mu}}\right)^\top} \Bigg\} \begin{bmatrix} \mathbf{0} \\ z_k\left[\mathrm{diag}\left(\mathfrak{p}_k\right) - \mathfrak{p}_k\mathfrak{p}_k^\top\right]\mathbf{X}_k \end{bmatrix}
\end{aligned}
$$

where

$$
\begin{aligned}
\frac{\partial \underline{\boldsymbol{\mu}}_k}{\partial\left(\mathrm{vec}\underline{\boldsymbol{\mu}}\right)^\top} &= \underset{(C+1)\times(C+1)}{\breve{\mathbf{I}}} \bigotimes \underset{1\times n}{\mathbf{I}_k} \\
\frac{\partial \mathrm{vec}\left[\left(\sum_{\mathfrak{s}} \frac{d_k}{q_k}\underline{\boldsymbol{\mu}}_k\underline{\boldsymbol{\mu}}_k^\top\right)^{-1}\right]}{\partial \underline{\boldsymbol{\mu}}^\top} &= -\left[\left(\left(\sum_{\mathfrak{s}} \frac{d_k}{q_k}\underline{\boldsymbol{\mu}}_k\underline{\boldsymbol{\mu}}_k^\top\right)^{-1}\right)^\top \bigotimes \left(\sum_{\mathfrak{s}} \frac{d_k}{q_k}\underline{\boldsymbol{\mu}}_k\underline{\boldsymbol{\mu}}_k^\top\right)^{-1}\right]\mathbf{M} \\
\frac{\partial \sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}_k - \sum_{\mathfrak{s}} d_k\underline{\boldsymbol{\mu}}_k}{\partial\left(\mathrm{vec}\underline{\boldsymbol{\mu}}\right)^\top} &= \mathbf{I} \bigotimes \left(\mathbf{1} - \mathbf{d}\right)^\top
\end{aligned}
$$

and $\mathbf{M}$ is defined in Equations (B.28), (B.29), and (B.30).

Inverse of Jacobian

Inverting the Jacobian can be done using a standard statistical package. Recall, that

the inverse of a block matrix is

$$\left[\widehat{\mathbf{J}}\left(\boldsymbol{\theta}\right)\right]^{-1} = \begin{bmatrix} \mathscr{A} & \mathscr{B} \\ \mathscr{C} & \mathscr{D} \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} \left(\mathscr{A} - \mathscr{B}\mathscr{D}^{-1}\mathscr{C}\right)^{-1} & -\mathscr{A}\mathscr{B}\left(\mathscr{D} - \mathscr{C}\mathscr{A}^{-1}\mathscr{B}\right)^{-1} \\ -\mathscr{D}^{-1}\mathscr{C}\left(\mathscr{A} - \mathscr{B}\mathscr{D}^{-1}\mathscr{C}\right)^{-1} & \left(\mathscr{D} - \mathscr{C}\mathscr{A}^{-1}\mathscr{B}\right)^{-1} \end{bmatrix}$$

Simplification of $\boldsymbol{\Sigma}_{\widehat{U}}\left(\widehat{\boldsymbol{\theta}}\right)$

$\boldsymbol{\Sigma}_{\widehat{U}}$ is the design-based variance of the estimating equations. That is $\boldsymbol{\Sigma}_{\widehat{U}} = \mathrm{var}\left[\widehat{\mathbf{U}}\left(\boldsymbol{\theta}\right)\right]$. An estimator of this variance is denoted $\widehat{\boldsymbol{\Sigma}}_{\widehat{U}} = \upsilon\left[\widehat{\mathbf{U}}\left(\boldsymbol{\theta}\right)\right]$. We note that $\widehat{\boldsymbol{\Sigma}}_{\widehat{U}}$ is a function of $\widehat{\boldsymbol{\theta}}$.

Perhaps the simplest way to estimate $\boldsymbol{\Sigma}_{\widehat{U}}$ is to assume that clusters were selected with-replacement. Consider the π-estimator for the total of the estimating equations for cluster i,

$$\widehat{\mathbf{U}}_i\left(\boldsymbol{\theta}\right) = \sum_{k \in \mathfrak{s}_i} \widehat{\mathbf{U}}_k\left(\boldsymbol{\theta}\right)$$

Also, recall that under with-replacement sampling, the single draw probability for cluster i is $p_i = \frac{\pi_i}{n}$. In this case, Särndal et al. (1992)[p. 154] show that an unbiased estimator of $\boldsymbol{\Sigma}_{\widehat{U}}$ is

$$\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right) = \frac{1}{n\left(n-1\right)} \sum_{i \in \mathfrak{s}_I} \left[\frac{1}{p_i}\widehat{\mathbf{U}}_i\left(\widehat{\boldsymbol{\theta}}\right) - \widehat{\mathbf{U}}\left(\widehat{\boldsymbol{\theta}}\right)\right]\left[\frac{1}{p_i}\widehat{\mathbf{U}}_i\left(\widehat{\boldsymbol{\theta}}\right) - \widehat{\mathbf{U}}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{\top}$$

On the other hand, if the first stage sample is selected without replacement, then we can extend the classic design variance formulas from Särndal et al. (1992)[p. 137] to the

multivariate case. The variance of the estimating equations will be

$$\boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}\right) = \sum\sum_{\mathscr{U}_I} \frac{\pi_{Iij} - \pi_{Ii}\pi_{Ij}}{\pi_{Ii}\pi_{Ij}} \widehat{\mathbf{U}}_i\left(\widehat{\boldsymbol{\theta}}\right) \widehat{\mathbf{U}}_j\left(\widehat{\boldsymbol{\theta}}\right)^{\top}$$
$$+ \sum_{\mathscr{U}_I} \frac{1}{\pi_{Ii}} \sum\sum_{\mathscr{U}_i} \frac{\pi_{kl|i} - \pi_{k|i}\pi_{l|i}}{\pi_{k|i}\pi_{l|i}} \widehat{\mathbf{U}}_{k|i}\left(\widehat{\boldsymbol{\theta}}\right) \widehat{\mathbf{U}}_{l|i}\left(\widehat{\boldsymbol{\theta}}\right)^{\top}$$

There are numerous other techniques one may employ to estimate $\boldsymbol{\Sigma}\left(\widehat{\boldsymbol{\theta}}\right)$ in cluster samples.

B.6 Model-Calibrated Maximum Pseudo Empirical Likelihood Estimator

Chen and Qin (1993) describe the pseudoempirical likelihood approach under simple random sampling. Zhong and Rao (1996) and Chen and Sitter (1999) extend the pseudoempirical likelihood approach to complex survey designs. In 2001, Wu and Sitter (2001) extended the pseudoempirical maximum likelihood estimator by integrating a generalized linear model into the calibration weighting.

B.6.1 Estimation of Model-Calibrated Maximum Pseudo Empirical Likelihood Estimator

Let $\mathbf{y}_k$ be a multivariate response for the kth sample unit. Also, let $\hat{\mathbf{t}}_y^{peM}$ be an estimate of the total of $\mathbf{t}_y = \sum_{\mathscr{U}} \mathbf{y}_k$. The pseudoempirical estimated total in a two-staged

sample is

$$\begin{aligned}\underset{(C-1)\times 1}{\hat{\mathbf{t}}_y^{peM}} &= M\sum_{\mathfrak{s}} p_k \mathbf{y}_k \\ &= M \underset{m\times(C-1)}{\mathbf{y}}^{\top} \underset{n\times 1}{\mathbf{p}}\end{aligned}$$

where $\mathbf{p}$ is found by maximizing the objective function

$$\underset{m\times 1}{\mathbf{d}}^{\top} \underset{m\times 1}{\log\mathbf{p}} = \sum_{\mathfrak{s}} d_k \log\left(p_k\right)$$

subject to the constraints

$$\begin{aligned}\underset{m\times(C-1)}{\mathbf{u}}^{\top}\mathbf{p} &= \underset{(C-1)\times 1}{\mathbf{0}} \\ \underset{m\times 1}{\mathbf{1}}^{\top}\mathbf{p} &= 1.\end{aligned}$$

Our estimating equation is

$$\boldsymbol{\psi} = \mathbf{d}^{\top}\log\mathbf{p} - \lambda_1\left(\mathbf{1}^{\top}\mathbf{p} - 1\right) - \underset{(C-1)\times 1}{\boldsymbol{\lambda}_2}^{\top}\left(\mathbf{u}^{\top}\mathbf{p}\right).$$

Differentiating $\boldsymbol{\phi}$ with respect to p_k gives

$$\frac{\partial\boldsymbol{\phi}}{\partial p_k} = \underset{m\times m}{\operatorname{diag}\left(\frac{1}{p_k}\right)}\underset{m\times 1}{\mathbf{d}} - \lambda_1\underset{m\times 1}{\mathbf{1}} - \underset{m\times(C-1)}{\mathbf{u}}\underset{(C-1)\times 1}{\boldsymbol{\lambda}_2}.$$

Setting the partial derivatives equal to 0 and solving for p_k gives

$$\begin{aligned}\underset{m\times 1}{\mathbf{0}} &= \operatorname{diag}\left(\frac{1}{p_k}\right)\mathbf{d} - \lambda_1\mathbf{1} - \mathbf{u}\boldsymbol{\lambda}_2 \\ \lambda_1\mathbf{1} + \mathbf{u}\lambda_2 &= \operatorname{diag}\left(\frac{1}{p_k}\right)\mathbf{d} \\ \operatorname{diag}\left(p_k\right)\left[\lambda_1\mathbf{1} + \mathbf{u}\boldsymbol{\lambda}_2\right] &= \mathbf{d}\end{aligned}$$

Thus,

$$p_k = \frac{d_k}{\lambda_1 + \boldsymbol{\lambda}_2^\top \mathbf{u}_k}.$$

We now solve for λ_1. Starting with our derivative, we have

$$\begin{aligned}
\underset{n\times 1}{\mathbf{0}} &= \operatorname{diag}\left(\frac{1}{p_k}\right)\mathbf{d} - \lambda_1 \mathbf{1} - \mathbf{u}\boldsymbol{\lambda}_2 \\
&= \mathbf{p}^\top \operatorname{diag}\left(\frac{1}{p_k}\right)\mathbf{d} - \mathbf{p}^\top \lambda_1 \mathbf{1} - \mathbf{p}^\top \mathbf{u}\boldsymbol{\lambda}_2 \\
&= \mathbf{p}^\top \operatorname{diag}\left(\frac{1}{p_k}\right)\mathbf{d} - \lambda_1 \mathbf{p}^\top \mathbf{1} - \mathbf{p}^\top \mathbf{u}\boldsymbol{\lambda}_2 \\
&= \mathbf{1}^\top \mathbf{d} - \lambda_1 \mathbf{p}^\top \mathbf{1} \\
&= \mathbf{1}^\top \mathbf{d} - \lambda_1 \\
&= \mathbf{1}^\top \mathbf{d} - \lambda_1 \\
\lambda_1 &= \mathbf{1}^\top \mathbf{d} = \sum_{\mathfrak{s}} d_k \\
&= \hat{M}.
\end{aligned}$$

Simplifying our function for p_k gives,

$$\begin{aligned}
p_k &= \frac{d_k}{\lambda_1 + \boldsymbol{\lambda}_2^\top \mathbf{u}_k} \\
&= \frac{d_k}{\hat{M} + \boldsymbol{\lambda}_2^\top \mathbf{u}_k} \\
&= \frac{\frac{d_k}{\hat{M}}}{1 + \frac{1}{\hat{M}} \boldsymbol{\lambda}_2^\top \mathbf{u}_k} \\
&= \frac{d_k^\star}{1 + \boldsymbol{\lambda}_{2\star}^\top \mathbf{u}_k}
\end{aligned}$$

where

$$d_k^\star = \frac{d_k}{\hat{M}}$$
$$\boldsymbol{\lambda}_{2\star} = \frac{1}{\hat{M}}\boldsymbol{\lambda}_2$$

We now write an estimating equation for λ_2. Starting with our formula for p_k gives

$$p_k = \frac{d_k}{\lambda_1 + \boldsymbol{\lambda}_2^\top \mathbf{u}_k}$$
$$\mathbf{u}_k p_k = \frac{d_k \mathbf{u}_k}{\lambda_1 + \boldsymbol{\lambda}_2^\top \mathbf{u}_k}$$
$$\sum_{\mathfrak{s}} \mathbf{u}_k p_k = \sum_{\mathfrak{s}} \frac{d_k \mathbf{u}_k}{\lambda_1 + \boldsymbol{\lambda}_2^\top \mathbf{u}_k}.$$

Considering our constraint, we have

$$\begin{aligned}
\underset{(C-1)\times 1}{\mathbf{0}} &= \sum_{\mathfrak{s}} \frac{d_k \mathbf{u}_k}{\lambda_1 + \boldsymbol{\lambda}_2^\top \mathbf{u}_k} \\
&= \sum_{\mathfrak{s}} \frac{d_k \mathbf{u}_k}{\hat{N} + \boldsymbol{\lambda}_2^\top \mathbf{u}_k} \\
&= \sum_{\mathfrak{s}} \frac{\frac{d_k}{\hat{N}} \mathbf{u}_k}{1 + \frac{1}{\hat{N}} \boldsymbol{\lambda}_2^\top \mathbf{u}_k} \\
&= \sum_{\mathfrak{s}} \frac{d_k^\star \mathbf{u}_k}{1 + \boldsymbol{\lambda}_{2\star}^\top \mathbf{u}_k}
\end{aligned}$$

Thus, we can solve for $\mathbf{p}$ by simultaneously solving

$$p_k = \frac{d_k^\star}{1 + \boldsymbol{\lambda}_{2\star}^\top \mathbf{u}_k}$$
$$\mathbf{0} = \sum_{\mathfrak{s}} \frac{d_k^\star \mathbf{u}_k}{1 + \boldsymbol{\lambda}_{2\star}^\top \mathbf{u}_k}$$

where

$$d_k^\star = \frac{d_k}{\hat{M}}$$
$$\boldsymbol{\lambda}_{2\star} = \frac{1}{\hat{M}}\boldsymbol{\lambda}_2$$

Once p_k has been estimated for all units in the population, we can construct the model-calibrated maximum pseudoempirical likelihood estimator as

$$\widehat{\mathbf{t}}_y^{peM} = M \sum_{\mathfrak{s}} p_k \mathbf{y}_k$$

or

$$\widehat{\mathbf{t}}_y^{pe\widehat{M}} = \widehat{M} \sum_{\mathfrak{s}} p_k \mathbf{y}_k$$

B.6.2 $\bar{t}_y^{peM}$ is asymptotically equal to $\bar{t}_y^{mc}$

In addition to Assumptions 4 and 5 in Section 3.2.1.1 on page 133, we make three assumptions from Wu (1999). The first two assumptions are discussed in Appendices 1 and 2 of Chen and Sitter (1999). In Appendix A2.3, Chen and Sitter verify the scalar version of these two conditions for ultimate cluster sampling. We closely follow their proof here, extending it to the case of multivariate response variables in cluster samples. In general, we follow the logic of Chen and Sitter (1999), but replace their scalar functions with elementwise matrix operations. We begin with three assumptions.

Assumption 17. $\mathbf{e}^\star = \max_{k \in \mathfrak{s}} |\mathbf{e}_k| = o_p\left(n^{\frac{1}{2}}\right)$, *elementwise.*

Means of random variables are usually $O_p(1)$. In this case, we would expect $\mathbf{e}^\star = o_p(n^\varepsilon)$ for any positive ε. In this proof, we assume an even more relaxed assumption. Assumption 17. That is, $\mathbf{e}^\star$ could be bounded, but it can also get smaller as the sample size increases.

Assumption 18. $\left[\sum_{\mathfrak{s}} \left(d_k^\star \mathbf{e}_k \mathbf{e}_k^\top\right)\right]^{-1} \sum_{\mathfrak{s}} d_k^\star \mathbf{e}_k = O_p\left(n^{-\frac{1}{2}}\right)$, *elementwise.*

Assumption 18 follows from Assumption 17, and the fact that $d_k = O\left(n^{-1}\right)$.

Assumption 19. $h^\star = \max_{k \in \mathfrak{s}} |h_i| = o_p\left(n\right)$.

Let $\underset{(C-1)\times 1}{\mathbf{e}_k} = \underset{(C-1)\times 1}{\boldsymbol{\mu}}\left(\underset{p\times 1}{\mathbf{x}_k}, \underset{p\times(C-1)}{\mathbf{B}_M}\right) - \frac{1}{M}\sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \mathbf{B}_M\right)$ and $\mathbf{h}_k = \mathbf{h}\left(\mathbf{x}_k, \mathbf{B}_M\right)$ where $\mathbf{h}\left(\mathbf{x}_k, \mathbf{B}_M\right)$ is defined in Assumption 5. Also let $\underset{(C-1)\times 1}{\mathbf{u}_k} = \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) - \frac{1}{M}\sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right)$.

If we treat the last category, C, as the baseline, Agresti (2002)[p. 271] shows that

$$\mu_{ck} = \frac{z_k e^{\mathbf{x}_k^\top \mathbf{B}_c}}{1 + \sum_{c=1}^{C-1} e^{\mathbf{x}_k^\top \mathbf{B}_c}}$$

and the baseline can be determined with $\mathbf{B}_C$ as a zero vector.

We employ the notation of Seber (2008)[p. 371] for matrix differentials. Specifically, let

$$\frac{\mathrm{d}\boldsymbol{\mu}_k}{\mathbf{B}} = \begin{bmatrix} \frac{\partial \mu_{k1}}{\partial \mathbf{B}} \\ \frac{\partial \mu_{k2}}{\partial \mathbf{B}} \\ \vdots \\ \frac{\partial \mu_{kC-1}}{\partial \mathbf{B}} \end{bmatrix} = \boldsymbol{\mu}_k \otimes \frac{\partial}{\partial \mathbf{B}}$$

For example, if there are five categories and two covariates, then

$$\frac{\partial \mu_{k1}}{\partial \mathbf{B}} = \left[\begin{pmatrix} \frac{\partial \mu_{k1}}{\partial B_{11}} & \frac{\partial \mu_{k1}}{\partial B_{12}} & \frac{\partial \mu_{k1}}{\partial B_{13}} & \frac{\partial \mu_{k1}}{\partial B_{14}} \\ \frac{\partial \mu_{k1}}{\partial B_{21}} & \frac{\partial \mu_{k1}}{\partial B_{22}} & \frac{\partial \mu_{k1}}{\partial B_{23}} & \frac{\partial \mu_{k1}}{\partial B_{24}} \end{pmatrix}\right]$$

where B_{cp} is the pth coefficient for category c. If $\boldsymbol{\mu}_k$ is $(C-1)\times 1$ and $\mathbf{B}$ is $p \times (C-1)$, then $\frac{\mathrm{d}\boldsymbol{\mu}_k}{\mathbf{B}}$ will be a $(C-1)\,p \times (C-1)$ matrix.

Let $\mathbf{B}^\star \in \left(\hat{\mathbf{B}}, \mathbf{B}_M\right)$ mean that each component of $\mathbf{B}^\star$ is between the corresponding elements of $\hat{\mathbf{B}}$ and $\mathbf{B}_M$. For example, for row r and column c, $\mathbf{B}^\star \in \left(\hat{\mathbf{B}}, \mathbf{B}_M\right)$ means that $\hat{B}_{cr} \leq B^\star_{cr} \leq B_{crM}$ for all r and c.

By the mean value theorem, there exists a matrix, $\mathbf{B}^\star$ where each element is either in the interval $\left(\hat{\mathbf{B}}, \mathbf{B}_M\right)$ or $\left(\mathbf{B}_M, \hat{\mathbf{B}}\right)$ such that

$$\mathbf{u}_k = \mathbf{e}_k + \underbrace{\left\{ \frac{\mathrm{d}\boldsymbol{\mu}\left(\mathbf{x}_k, \boldsymbol{\theta}\right)}{\mathrm{d}\boldsymbol{\theta}} \bigg|_{\boldsymbol{\theta}=\mathbf{B}^\star} \right\}^\top}_{p(C-1)\times(C-1)} \underbrace{\mathrm{vec}\left(\hat{\mathbf{B}} - \mathbf{B}_M\right)}_{p(C-1)\times 1} - \left\{ \frac{1}{N} \sum_{\mathscr{U}} \frac{\mathrm{d}\boldsymbol{\mu}\left(\mathbf{x}_k, \boldsymbol{\theta}\right)}{\mathrm{d}\boldsymbol{\theta}} \bigg|_{\boldsymbol{\theta}=\mathbf{B}^\star} \right\}^\top \mathrm{vec}\left(\hat{\mathbf{B}} - \mathbf{B}_M\right).$$

This implies that each category vector taken from $\mathbf{u}_k$ can be linearized as

$$\begin{aligned}\mathbf{u}_{ck} = \mathbf{e}_{ck} &+ \sum_{c=1}^{C-1} \underbrace{\left\{ \frac{\mathrm{d}\mu\left(\mathbf{x}_k, \boldsymbol{\theta}_c\right)}{\mathrm{d}\boldsymbol{\theta}_c} \bigg|_{\boldsymbol{\theta}_c=\mathbf{B}_c^\star} \right\}^\top}_{p\times 1} \underbrace{\left(\hat{\mathbf{B}}_c - \mathbf{B}_{cM}\right)}_{p\times 1} \\ &- \sum_{c=1}^{C-1} \left\{ \frac{1}{M} \sum_{\mathscr{U}} \frac{\mathrm{d}\boldsymbol{\mu}\left(\mathbf{x}_k, \boldsymbol{\theta}_c\right)}{\mathrm{d}\boldsymbol{\theta}_c} \bigg|_{\boldsymbol{\theta}_c=\mathbf{B}_c^\star} \right\}^\top \left(\hat{\mathbf{B}}_c - \mathbf{B}_{cM}\right)\end{aligned}$$

By Assumption 4 on page 133, $\hat{\mathbf{B}} - \mathbf{B} = O_p\left(n^{-\frac{1}{2}}\right)$. Also, by Assumption 5, we see that $\left\{ \frac{1}{M} \sum_{\mathscr{U}} \frac{\mathrm{d}\boldsymbol{\mu}(\mathbf{x}_k, \mathbf{B})}{\mathrm{d}\mathbf{B}} \bigg|_{\mathbf{B}=\mathbf{B}^\star} \right\}^\top = O\left(1\right)$. Thus, we have,

$$\mathbf{u}_k = \mathbf{e}_k + O_p\left(n^{-\frac{1}{2}}\right).$$

By Assumption 17 on page 365, $\mathbf{u}^\star = \max_{k\in\mathfrak{s}} |\mathbf{u}_k| = o_p\left(n^{\frac{1}{2}}\right)$. Similarly, Assumption 18 can be restated in terms of $\mathbf{u}_k$.

Now, from our estimating equation for λ_2, we have

$$
\begin{aligned}
\mathbf{0} &= \sum_{\mathfrak{s}} \frac{d_k^\star \mathbf{u}_k}{1 + \boldsymbol{\lambda}_{2\star}^\top \mathbf{u}_k} \\
&= \sum_{\mathfrak{s}} \frac{d_k^\star \mathbf{u}_k + d_k^\star \mathbf{u}_k \mathbf{u}_k^\top \boldsymbol{\lambda}_{2\star} - d_k^\star \mathbf{u}_k \mathbf{u}_k^\top \boldsymbol{\lambda}_{2\star}}{1 + \boldsymbol{\lambda}_{2\star}^\top \mathbf{u}_k} \\
&= \sum_{\mathfrak{s}} \left[\frac{d_k^\star \mathbf{u}_k + d_k^\star \mathbf{u}_k \mathbf{u}_k^\top \boldsymbol{\lambda}_{2\star}}{1 + \boldsymbol{\lambda}_{2\star}^\top \mathbf{u}_k} - \frac{d_k^\star \mathbf{u}_k \mathbf{u}_k^\top \boldsymbol{\lambda}_{2\star}}{1 + \boldsymbol{\lambda}_{2\star}^\top \mathbf{u}_k} \right] \\
&= \sum_{\mathfrak{s}} \left[\frac{d_k^\star \mathbf{u}_k \left(1 + \mathbf{u}_k^\top \boldsymbol{\lambda}_{2\star}\right)}{1 + \boldsymbol{\lambda}_{2\star}^\top \mathbf{u}_k} - \frac{d_k^\star \mathbf{u}_k \mathbf{u}_k^\top \boldsymbol{\lambda}_{2\star}}{1 + \boldsymbol{\lambda}_{2\star}^\top \mathbf{u}_k} \right] \\
&= \sum_{\mathfrak{s}} \left[d_k^\star \mathbf{u}_k - \frac{d_k^\star \mathbf{u}_k \mathbf{u}_k^\top \boldsymbol{\lambda}_{2\star}}{1 + \boldsymbol{\lambda}_{2\star}^\top \mathbf{u}_k} \right] \\
&= \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k - \sum_{\mathfrak{s}} \frac{d_k^\star \mathbf{u}_k \mathbf{u}_k^\top \boldsymbol{\lambda}_{2\star}}{1 + \boldsymbol{\lambda}_{2\star}^\top \mathbf{u}_k} \\
&= \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k - \sum_{\mathfrak{s}} \left[\frac{d_k^\star \mathbf{u}_k \mathbf{u}_k^\top}{1 + \boldsymbol{\lambda}_{2\star}^\top \mathbf{u}_k} \right] \boldsymbol{\lambda}_{2\star},
\end{aligned}
$$

implyiing that

$$
\sum_{\mathfrak{s}} \left[\frac{d_k^\star \mathbf{u}_k \mathbf{u}_k^\top}{1 + \boldsymbol{\lambda}_{2\star}^\top \mathbf{u}_k} \right] \boldsymbol{\lambda}_{2\star} = \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k.
$$

Let $\mathbf{u}^\star$ be a matrix containing the maximum values of the absolute value of each component of $\mathbf{u}$ across the full sample. For example, the first element of $\mathbf{u}^\star$ is $\max_{k \in \mathfrak{s}} \mid \mathbf{u}_{11k} \mid$. Replacing $\mathbf{u}_k$ with $\mathbf{u}^\star$ gives us an upper bound,

$$
\begin{aligned}
\sum_{\mathfrak{s}} \left[d_k^\star \mathbf{u}_k \mathbf{u}_k^\top \right] \frac{|\boldsymbol{\lambda}_{2\star}|}{1 + |\boldsymbol{\lambda}_{2\star}|^\top \mathbf{u}^\star} &\leq \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k \\
\frac{|\boldsymbol{\lambda}_{2\star}|}{1 + |\boldsymbol{\lambda}_{2\star}|^\top \mathbf{u}^\star} &\leq \left[\sum_{\mathfrak{s}} \left(d_k^\star \mathbf{u}_k \mathbf{u}_k^\top \right) \right]^{-1} \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k.
\end{aligned}
$$

Considering Assumption 18 and the relationship between e_k and u_k we have

$$
\left[\sum_{\mathfrak{s}} \left(d_k^\star \mathbf{u}_k \mathbf{u}_k^\top \right) \right]^{-1} \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k = O_p \left(n^{-\frac{1}{2}} \right).
$$

Hence,

$$\frac{\mid \boldsymbol{\lambda}_{2\star} \mid}{1+ \mid \boldsymbol{\lambda}_{2\star}^{\top} \mid \mathbf{u}^{\star}} \leq O_p\left(n^{-\frac{1}{2}}\right)$$

$$\mid \boldsymbol{\lambda}_{2\star} \mid \leq \left[1+ \mid \boldsymbol{\lambda}_{2\star}^{\top} \mid \mathbf{u}^{\star}\right] O_p\left(n^{-\frac{1}{2}}\right)$$

$$\mid \boldsymbol{\lambda}_{2\star} \mid - \mathbf{u}^{\star\top} \mid \boldsymbol{\lambda}_{2\star} \mid O_p\left(n^{-\frac{1}{2}}\right) \leq O_p\left(n^{-\frac{1}{2}}\right)$$

$$\left[1 - \mathbf{u}_k^{\star\top} O_p\left(n^{-\frac{1}{2}}\right)\right] \mid \boldsymbol{\lambda}_{2\star} \mid \leq O_p\left(n^{-\frac{1}{2}}\right)$$

By Assumption 17 on page 365, we have

$$\left[1 - o_p\left(n^{\frac{1}{2}}\right) O_p\left(n^{-\frac{1}{2}}\right)\right] \mid \boldsymbol{\lambda}_{2\star} \mid \leq O_p\left(n^{-\frac{1}{2}}\right)$$

$$\left[1 - o_p\left(1\right)\right] \mid \boldsymbol{\lambda}_{2\star} \mid \leq O_p\left(n^{-\frac{1}{2}}\right)$$

Thus, we must have

$$\boldsymbol{\lambda}_{2\star} = O_p\left(n^{-\frac{1}{2}}\right).$$

Furthermore,

$$\begin{aligned}
\frac{1}{1 + \boldsymbol{\lambda}_{2\star}^{\top}\mathbf{u}^{\star}} &= \frac{1 + \boldsymbol{\lambda}_{2\star}^{\top}\mathbf{u}^{\star} - \boldsymbol{\lambda}_{2\star}^{\top}\mathbf{u}^{\star}}{1 + \boldsymbol{\lambda}_{2\star}^{\top}\mathbf{u}^{\star}} \\
&= 1 - \frac{\boldsymbol{\lambda}_{2\star}^{\top}\mathbf{u}^{\star}}{1 + \boldsymbol{\lambda}_{2\star}^{\top}\mathbf{u}^{\star}} \\
&= 1 - \frac{O_p\left(n^{-\frac{1}{2}}\right) o_p\left(n^{\frac{1}{2}}\right)}{1 + O_p\left(n^{-\frac{1}{2}}\right) o_p\left(n^{\frac{1}{2}}\right)} \\
&= 1 - \frac{o_p\left(1\right)}{1 + o_p\left(1\right)} \\
&= 1 + o_p\left(1\right).
\end{aligned}$$

The term $o_p\left(1\right)$ is uniform over $k \in \mathfrak{s}$. This gives $\sum_{\mathfrak{s}} \left[\frac{d_k^{\star}\mathbf{u}_k\mathbf{u}_k^{\top}}{1+\boldsymbol{\lambda}_{2\star}^{\top}\mathbf{u}_k}\right] = \left(\sum_{\mathfrak{s}} d_k^{\star}\mathbf{u}_k\mathbf{u}_k^{\top}\right)\left[1 + o_p\left(1\right)\right]$.

Consider that

$$
\begin{aligned}
\sum_{\mathfrak{s}} \left[\frac{d_k^\star \mathbf{u}_k \mathbf{u}_k^\top}{1 + \boldsymbol{\lambda}_{2\star}^\top \mathbf{u}_k} \right] \boldsymbol{\lambda}_{2\star} &= \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k \\
\boldsymbol{\lambda}_{2\star} &= \left[\sum_{\mathfrak{s}} \left[\frac{d_k^\star \mathbf{u}_k \mathbf{u}_k^\top}{1 + \boldsymbol{\lambda}_{2\star}^\top \mathbf{u}_k} \right] \right]^{-1} \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k \\
\boldsymbol{\lambda}_{2\star} &= \left\{ \left(\sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k \mathbf{u}_k^\top \right) \left[1 + o_p(1) \right] \right\}^{-1} \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k \\
\boldsymbol{\lambda}_{2\star} &= \left(\sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k \mathbf{u}_k^\top \right)^{-1} \frac{1}{1 + o_p(1)} \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k \\
\boldsymbol{\lambda}_{2\star} \left[1 + o_p(1) \right] &= \left(\sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k \mathbf{u}_k^\top \right)^{-1} \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k \\
\boldsymbol{\lambda}_{2\star} + \boldsymbol{\lambda}_{2\star} o_p(1) &= \left(\sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k \mathbf{u}_k^\top \right)^{-1} \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k \\
\boldsymbol{\lambda}_{2\star} &= \left(\sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k \mathbf{u}_k^\top \right)^{-1} \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k - \boldsymbol{\lambda}_{2\star} o_p(1) \\
\boldsymbol{\lambda}_{2\star} &= \left(\sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k \mathbf{u}_k^\top \right)^{-1} \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k - O_p\left(n^{-\frac{1}{2}} \right) o_p(1)
\end{aligned}
$$

Thus, we obtain

$$
\boldsymbol{\lambda}_{2\star} = \left(\sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k \mathbf{u}_k^\top \right)^{-1} \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k - o_p\left(n^{-\frac{1}{2}} \right).
$$

It now follows that

$$
\begin{aligned}
\hat{\tilde{\mathbf{t}}}_y^{pe} &= \sum_{\mathfrak{s}} p_k \mathbf{y}_k \\
&= \sum_{\mathfrak{s}} \frac{d_k^\star}{1 + \boldsymbol{\lambda}_{2\star}^\top \mathbf{u}_k} \mathbf{y}_k \\
&= \sum_{\mathfrak{s}} d_k^\star \mathbf{y}_k - \sum_{\mathfrak{s}} \left[\frac{d_k^\star \mathbf{u}_k \mathbf{y}_k^\top}{1 + \boldsymbol{\lambda}_{2\star}^\top \mathbf{u}_k} \right] \boldsymbol{\lambda}_{2\star} \\
&= \sum_{\mathfrak{s}} d_k^\star \mathbf{y}_k - \sum_{\mathfrak{s}} \left[d_k^\star \mathbf{u}_k \mathbf{y}_k^\top \right] (1 + o_p(1)) \boldsymbol{\lambda}_{2\star} \\
&= \sum_{\mathfrak{s}} d_k^\star \mathbf{y}_k - \sum_{\mathfrak{s}} \left[d_k^\star \mathbf{u}_k \mathbf{y}_k^\top \right] \boldsymbol{\lambda}_{2\star} - \sum_{\mathfrak{s}} \left[d_k^\star \mathbf{u}_k \mathbf{y}_k^\top \right] o_p(1) \boldsymbol{\lambda}_{2\star} \\
&= \sum_{\mathfrak{s}} d_k^\star \mathbf{y}_k - \sum_{\mathfrak{s}} \left[d_k^\star \mathbf{u}_k \mathbf{y}_k^\top \right] \boldsymbol{\lambda}_{2\star} - \\
&\qquad \sum_{\mathfrak{s}} \left[d_k^\star \mathbf{u}_k \mathbf{y}_k^\top \right] o_p(1) \left[\left(\sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k \mathbf{u}_k^\top \right)^{-1} \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k - o_p\left(n^{-\frac{1}{2}}\right) \right] \\
&= \sum_{\mathfrak{s}} d_k^\star \mathbf{y}_k - \sum_{\mathfrak{s}} \left[d_k^\star \mathbf{u}_k \mathbf{y}_k^\top \right] \boldsymbol{\lambda}_{2\star} - \sum_{\mathfrak{s}} \left[d_k^\star \mathbf{u}_k \mathbf{y}_k^\top \right] o_p(1) \left[O_p\left(n^{-\frac{1}{2}}\right) - o_p\left(n^{-\frac{1}{2}}\right) \right]
\end{aligned}
$$

$\sum_{\mathfrak{s}} \left[d_k^\star \mathbf{u}_k \mathbf{y}_k^\top\right]$ is the mean of bounded terms. Thus,

$$
\begin{aligned}
\hat{\bar{\mathbf{t}}}_y^{pe} &= \sum_{\mathfrak{s}} d_k^\star \mathbf{y}_k - \sum_{\mathfrak{s}} \left[d_k^\star \mathbf{u}_k \mathbf{y}_k^\top\right] \boldsymbol{\lambda}_{2\star} - O_p(1)\, o_p(1) \left[O_p\left(n^{-\frac{1}{2}}\right) - o_p\left(n^{-\frac{1}{2}}\right)\right] \\
&= \sum_{\mathfrak{s}} d_k^\star \mathbf{y}_k - \sum_{\mathfrak{s}} \left[d_k^\star \mathbf{u}_k \mathbf{y}_k^\top\right] \boldsymbol{\lambda}_{2\star} - O_p(1)\, o_p(1)\, O_p\left(n^{-\frac{1}{2}}\right) \\
&= \sum_{\mathfrak{s}} d_k^\star \mathbf{y}_k - \sum_{\mathfrak{s}} \left[d_k^\star \mathbf{u}_k \mathbf{y}_k^\top\right] \boldsymbol{\lambda}_{2\star} - o_p\left(n^{-\frac{1}{2}}\right) \\
&= \sum_{\mathfrak{s}} d_k^\star \mathbf{y}_k - \sum_{\mathfrak{s}} \left[d_k^\star \mathbf{u}_k \mathbf{y}_k^\top\right] \left[\left(\sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k \mathbf{u}_k^\top\right)^{-1} \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k - o_p\left(n^{-\frac{1}{2}}\right)\right] - o_p\left(n^{-\frac{1}{2}}\right) \\
&= \sum_{\mathfrak{s}} d_k^\star \mathbf{y}_k - \sum_{\mathfrak{s}} \left[d_k^\star \mathbf{u}_k \mathbf{y}_k^\top\right] \left(\sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k \mathbf{u}_k^\top\right)^{-1} \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k - \sum_{\mathfrak{s}} \left[d_k^\star \mathbf{u}_k \mathbf{y}_k^\top\right] o_p\left(n^{-\frac{1}{2}}\right) - o_p\left(n^{-\frac{1}{2}}\right) \\
&= \sum_{\mathfrak{s}} d_k^\star \mathbf{y}_k - \sum_{\mathfrak{s}} \left[d_k^\star \mathbf{u}_k \mathbf{y}_k^\top\right] \left(\sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k \mathbf{u}_k^\top\right)^{-1} \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k - O_p(1)\, o_p\left(n^{-\frac{1}{2}}\right) - o_p\left(n^{-\frac{1}{2}}\right) \\
&= \sum_{\mathfrak{s}} d_k^\star \mathbf{y}_k - \sum_{\mathfrak{s}} \left[d_k^\star \mathbf{u}_k \mathbf{y}_k^\top\right] \left(\sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k \mathbf{u}_k^\top\right)^{-1} \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k - o_p\left(n^{-\frac{1}{2}}\right) \\
&= \sum_{\mathfrak{s}} d_k^\star \mathbf{y}_k - \frac{1}{\hat{M}} \sum_{\mathfrak{s}} \left[d_k \mathbf{u}_k \mathbf{y}_k^\top\right] \hat{M} \left(\sum_{\mathfrak{s}} d_k \mathbf{u}_k \mathbf{u}_k^\top\right)^{-1} \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k - o_p\left(n^{-\frac{1}{2}}\right) \\
&= \sum_{\mathfrak{s}} d_k^\star \mathbf{y}_k - \sum_{\mathfrak{s}} \left[d_k \mathbf{u}_k \mathbf{y}_k^\top\right] \left(\sum_{\mathfrak{s}} d_k \mathbf{u}_k \mathbf{u}_k^\top\right)^{-1} \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k - o_p\left(n^{-\frac{1}{2}}\right) \\
&= \sum_{\mathfrak{s}} d_k^\star \mathbf{y}_k - \hat{\mathbf{B}}_{\mathbf{u},\mathbf{y}} \sum_{\mathfrak{s}} d_k^\star \mathbf{u}_k - o_p\left(n^{-\frac{1}{2}}\right) \\
&= \sum_{\mathfrak{s}} d_k^\star \mathbf{y}_k - \hat{\mathbf{B}}_{\mathbf{u},\mathbf{y}} \sum_{\mathfrak{s}} d_k^\star \left[\boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) - \frac{1}{M} \sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right)\right] - o_p\left(n^{-\frac{1}{2}}\right) \\
&= \sum_{\mathfrak{s}} d_k^\star \mathbf{y}_k + \hat{\mathbf{B}}_{\mathbf{u},\mathbf{y}} \left[\frac{1}{M} \sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) - \sum_{\mathfrak{s}} d_k^\star \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right)\right] - o_p\left(n^{-\frac{1}{2}}\right) \\
&= \frac{1}{\hat{M}} \sum_{\mathfrak{s}} d_k \mathbf{y}_k + \hat{\mathbf{B}}_{\mathbf{u},\mathbf{y}} \left[\frac{1}{M} \sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) - \frac{1}{\hat{M}} \sum_{\mathfrak{s}} d_k \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right)\right] - o_p\left(m^{-\frac{1}{2}}\right)
\end{aligned}
$$

If we replace $\hat{M}$ with M, we have

$$\begin{aligned}
\hat{\bar{\mathbf{t}}}_y^{pe} &= \frac{1}{M}\left\{\sum_{\mathfrak{s}} d_k \mathbf{y}_k + \hat{\mathbf{B}}_{\mathbf{u},\mathbf{y}}\left[\sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) - \sum_{\mathfrak{s}} d_k \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right)\right]\right\} - o_p\left(n^{-\frac{1}{2}}\right) \\
&= \frac{1}{N}\widehat{\mathbf{t}}_y^{mc} - o_p\left(n^{-\frac{1}{2}}\right) \\
&= \widehat{\bar{\mathbf{t}}}_y^{mc} - o_p\left(n^{-\frac{1}{2}}\right)
\end{aligned}$$

where

$$\hat{\mathbf{B}}_{\mathbf{u},\mathbf{y}} = \sum_{\mathfrak{s}} \left[d_k \mathbf{u}_k \mathbf{y}_k^\top\right]\left(\sum_{\mathfrak{s}} d_k \mathbf{u}_k \mathbf{u}_k^\top\right)^{-1}$$

We conclude that $\hat{\mathbf{t}}_y^{peM}$ is asymptotically equivalent to the model calibration estimator and propose using the model calibration variance estimators to estimate the variance of $\hat{\mathbf{t}}_y^{peM}$.

In so far as $\hat{M}$ can be replaced by M, we also conclude that $\hat{\mathbf{t}}_y^{pe\widehat{M}}$ is asymptotically equivalent to the model calibration estimator.

B.7 Simulation Results

This section contains tables and graphs summarizing our analysis of the simulations for the Multinomial Logistic Assisted Estimators. Formulas for the summary measures that follow can be found in Table 3.8 of Section 3.3.2.4 on page 156.

B.7.1 Synthetic Population

B.7.1.1 Percent Simulation Coefficient of Variation Table

Table B.3: Percent Simulation Coefficient of Variation for Synthetic Population

	Small Samples								
	Fixed SRS			Rate SRS			Fixed PPS		
Estimator	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
$\hat{\mathbf{t}}_y^{\pi}$	0.567	0.475	0.648	0.669	0.580	0.749	0.524	0.439	0.601
$\hat{\mathbf{t}}_{yc}^{gd}$	0.418	0.480	0.461	0.417	0.504	0.494	0.422	0.472	0.457
$\hat{\mathbf{t}}_y^{lg}$	0.029	0.096	0.032	0.029	0.097	0.032	0.028	0.095	0.032
$\hat{\mathbf{t}}_y^{mc}$	0.031	0.105	0.035	0.033	0.112	0.036	0.032	0.108	0.035
$\hat{\mathbf{t}}_y^{peN}$	0.041	0.109	0.040	0.066	0.119	0.078	0.042	0.111	0.046
$\hat{\mathbf{t}}_y^{pe\hat{N}}$	0.140	0.175	0.138	0.366	0.378	0.369	0.042	0.111	0.046
	Large Samples								
$\hat{\mathbf{t}}_y^{\pi}$	0.063	0.054	0.075	0.078	0.067	0.088	0.058	0.049	0.070
$\hat{\mathbf{t}}_{yc}^{gd}$	0.048	0.051	0.055	0.048	0.052	0.053	0.046	0.049	0.052
$\hat{\mathbf{t}}_y^{lg}$	0.003	0.011	0.003	0.003	0.011	0.003	0.003	0.010	0.003
$\hat{\mathbf{t}}_y^{mc}$	0.003	0.011	0.003	0.003	0.011	0.003	0.003	0.010	0.003
$\hat{\mathbf{t}}_y^{peN}$	0.003	0.011	0.003	0.003	0.011	0.003	0.003	0.010	0.003
$\hat{\mathbf{t}}_y^{pe\hat{N}}$	0.015	0.019	0.015	0.042	0.043	0.041	0.003	0.010	0.003

B.7.1.2 Average Distance from True Value

Table B.4: Average Distance from True Value for Synthetic Population (in thousands)

	Small Samples			Large Samples		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
$\hat{\mathbf{t}}_y^{\pi}$	7,317.3	8,715.5	6,848.9	844.1	1023.7	777.0
$\hat{\mathbf{t}}_{yc}^{gd}$	5,419.9	5,597.9	5,385.2	638.0	633.0	609.7
$\hat{\mathbf{t}}_y^{lg}$	431.9	430.7	424.7	47.6	47.1	46.8
$\hat{\mathbf{t}}_y^{mc}$	468.0	482.9	469.3	47.6	47.1	46.8
$\hat{\mathbf{t}}_y^{peM}$	484.8	549.2	494.2	47.6	47.1	46.8
$\hat{\mathbf{t}}_y^{pe\hat{M}}$	1,641.6	4,176.2	494.2	182.1	475.9	46.8

B.7.1.3 Empirical Standard Deviation of Distance from True Value

Table B.5: Empirical Standard Deviation of Distance from True Value for Synthetic Population (in thousands)

	Small Samples			Large Samples		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
$\hat{\mathbf{t}}_y^{\pi}$	4,153.1	4,570.1	3,703.8	435.6	528.3	412.9
$\hat{\mathbf{t}}_{yc}^{gd}$	2,942.3	3,037.9	3,016.9	330.8	321.8	323.0
$\hat{\mathbf{t}}_y^{lg}$	233.0	237.0	228.7	24.7	24.7	24.5
$\hat{\mathbf{t}}_y^{mc}$	257.0	289.5	266.5	24.8	24.6	24.5
$\hat{\mathbf{t}}_y^{peM}$	405.2	876.6	459.4	24.8	24.6	24.5
$\hat{\mathbf{t}}_y^{pe\hat{M}}$	1,101.7	3,058.7	459.4	121.7	341.0	24.5

B.7.1.4 Percent Relative Bias Table

Table B.6: Percent Relative Bias for Synthetic Population

	Small Samples								
	Fixed SRS			Rate SRS			Fixed PPS		
Estimator	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
$\hat{\mathbf{t}}_y^{\pi}$	2.0	1.2	-0.3	0.1	0.0	-0.3	-0.4	-0.1	0.7
$\hat{\mathbf{t}}_y^{gd}$	1.2	4.0	-0.4	0.6	4.7	-0.6	-0.2	2.8	-0.4
$\hat{\mathbf{t}}_y^{lg}$	0.0	-0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0
$\hat{\mathbf{t}}_y^{mc}$	0.0	-0.1	0.0	0.0	0.1	0.0	0.0	0.2	0.0
$\hat{\mathbf{t}}_y^{peM}$	0.0	-0.1	0.0	0.0	0.0	-0.1	0.0	0.2	-0.1
$\hat{\mathbf{t}}_y^{peM\hat{}}$	0.0	0.0	0.0	-0.5	-0.5	-0.6	0.0	0.2	-0.1
	Large Samples								
$\hat{\mathbf{t}}_y^{\pi}$	0.0	0.0	-0.1	0.1	0.1	0.1	0.0	0.0	0.1
$\hat{\mathbf{t}}_y^{gd}$	0.0	0.0	-0.1	0.0	0.0	0.0	0.0	0.0	0.0
$\hat{\mathbf{t}}_y^{lg}$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\hat{\mathbf{t}}_y^{mc}$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\hat{\mathbf{t}}_y^{peM}$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\hat{\mathbf{t}}_y^{peM\hat{}}$	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0

B.7.1.5 Percent Relative Median Difference Table

Table B.7: Percent Relative Median Difference for Synthetic Population

	Small Samples								
	Fixed SRS			Rate SRS			Fixed PPS		
Estimator	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
$\hat{\mathbf{t}}_y^{\pi}$	-0.8	-0.5	-3.4	-2.0	-1.9	-3.6	-2.0	-1.5	-1.9
$\hat{\mathbf{t}}_y^{gd}$	-0.6	2.7	-2.3	-0.5	2.7	-3.0	-1.7	0.8	-1.8
$\hat{\mathbf{t}}_y^{lg}$	0.0	0.1	0.0	0.0	0.0	0.0	-0.1	0.0	0.0
$\hat{\mathbf{t}}_y^{mc}$	0.0	-0.1	0.0	0.0	0.1	0.0	0.0	0.1	0.0
$\hat{\mathbf{t}}_y^{peM}$	0.0	-0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.0
$\hat{\mathbf{t}}_y^{pe\hat{M}}$	-0.3	-0.5	-0.3	-0.9	-0.7	-1.0	0.0	0.1	0.0
	Large Samples								
$\hat{\mathbf{t}}_y^{\pi}$	-0.1	0.0	-0.1	0.1	-0.1	0.0	0.0	0.0	0.1
$\hat{\mathbf{t}}_y^{gd}$	0.0	0.0	-0.1	0.0	0.1	0.0	0.0	0.0	0.1
$\hat{\mathbf{t}}_y^{lg}$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\hat{\mathbf{t}}_y^{mc}$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\hat{\mathbf{t}}_y^{peM}$	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
$\hat{\mathbf{t}}_y^{pe\hat{M}}$	0.0	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0

B.7.1.6 Percent Relative Root Mean Squared Error Table

Table B.8: Percent Relative Root Mean Squared Error for Synthetic Population

	Small Samples								
	Fixed SRS			Rate SRS			Fixed PPS		
Estimator	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
$\hat{\mathbf{t}}_y^{\pi}$	25.3	21.2	29.0	29.9	25.9	33.5	23.4	19.6	26.9
$\hat{\mathbf{t}}_y^{gd}$	18.7	21.5	20.6	18.6	22.5	22.1	18.9	21.1	20.4
$\hat{\mathbf{t}}_y^{lg}$	1.3	4.3	1.4	1.3	4.3	1.4	1.3	4.2	1.4
$\hat{\mathbf{t}}_y^{mc}$	1.4	4.7	1.6	1.5	5.0	1.6	1.4	4.8	1.6
$\hat{\mathbf{t}}_y^{peM}$	1.8	4.9	1.8	3.0	5.3	3.5	1.9	5.0	2.0
$\hat{\mathbf{t}}_y^{peM\hat{}}$	6.2	7.8	6.2	16.4	16.9	16.5	1.9	5.0	2.0
	Large Samples								
$\hat{\mathbf{t}}_y^{\pi}$	2.8	2.4	3.3	3.5	3.0	4.0	2.6	2.2	3.1
$\hat{\mathbf{t}}_y^{gd}$	2.2	2.3	2.4	2.1	2.3	2.4	2.1	2.2	2.4
$\hat{\mathbf{t}}_y^{lg}$	0.1	0.5	0.2	0.1	0.5	0.2	0.1	0.5	0.2
$\hat{\mathbf{t}}_y^{mc}$	0.1	0.5	0.2	0.1	0.5	0.2	0.1	0.5	0.2
$\hat{\mathbf{t}}_y^{peM}$	0.1	0.5	0.2	0.1	0.5	0.2	0.1	0.5	0.2
$\hat{\mathbf{t}}_y^{peM\hat{}}$	0.7	0.8	0.7	1.9	1.9	1.8	0.1	0.5	0.2

B.7.1.7 Percent Relative Root Median Squared Error Table

Table B.9: Percent Relative Root Mean Squared Error for Synthetic Population

	Small Samples								
	Fixed SRS			Rate SRS			Fixed PPS		
Estimator	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
$\hat{\mathbf{t}}_y^{\pi}$	15.9	13.4	19.5	19.7	16.9	22.7	15.9	12.9	18.1
$\hat{\mathbf{t}}_y^{gd}$	11.7	13.9	14.1	12.3	14.0	14.3	12.1	13.4	12.8
$\hat{\mathbf{t}}_y^{lg}$	0.9	2.7	1.0	0.9	2.8	0.9	0.8	2.8	0.9
$\hat{\mathbf{t}}_y^{mc}$	0.9	3.1	1.0	0.9	3.1	1.1	0.9	3.1	1.0
$\hat{\mathbf{t}}_y^{peM}$	0.9	3.2	1.0	0.9	3.2	1.1	1.0	3.1	1.1
$\hat{\mathbf{t}}_y^{pe\hat{M}}$	4.2	5.3	4.1	10.8	11.5	10.9	1.0	3.1	1.1
	Large Samples								
$\hat{\mathbf{t}}_y^{\pi}$	1.9	1.6	2.2	2.4	2.0	2.6	1.8	1.4	2.1
$\hat{\mathbf{t}}_y^{gd}$	1.4	1.5	1.6	1.4	1.5	1.6	1.4	1.5	1.6
$\hat{\mathbf{t}}_y^{lg}$	0.1	0.3	0.1	0.1	0.3	0.1	0.1	0.3	0.1
$\hat{\mathbf{t}}_y^{mc}$	0.1	0.3	0.1	0.1	0.3	0.1	0.1	0.3	0.1
$\hat{\mathbf{t}}_y^{peM}$	0.1	0.3	0.1	0.1	0.3	0.1	0.1	0.3	0.1
$\hat{\mathbf{t}}_y^{pe\hat{M}}$	0.5	0.6	0.5	1.2	1.3	1.3	0.1	0.3	0.1

B.7.1.8 Percent Relative Bias Table for LGREG Variance Estimators

Table B.10: Percent Relative Bias of LGREG Variance Estimators for Synthetic Population

	Small Samples								
	Fixed SRS			Rate SRS			Fixed PPS		
Estimator	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-18.6	-15.0	-23.9	-26.0	-27.1	-30.5	-20.2	-19.9	-22.9
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-22.6	-19.3	-27.7	-29.7	-30.8	-34.0	-24.2	-23.9	-26.7
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-13.8	-9.8	-16.1	-12.3	-14.8	-16.5	-10.3	-11.0	-13.2
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-37.5	-37.2	-40.2	-49.6	-50.5	-49.4	-41.9	-44.3	-41.6
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-40.7	-40.3	-43.2	-52.1	-53.0	-51.9	-44.8	-47.1	-44.6
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-31.2	-30.1	-31.9	-33.4	-34.1	-34.2	-30.0	-32.7	-30.8
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-35.2	-36.5	-36.3	-46.3	-49.5	-43.7	-39.6	-43.8	-37.8
	Large Samples								
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	3.9	-3.2	4.6	-4.1	-7.1	2.5	-1.6	-1.8	-2.4
$\upsilon_{e}\left(\mathbf{t}_y^{lg}\right)$	3.2	-3.8	3.9	-4.8	-7.7	1.8	-2.3	-2.5	-3.1
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	4.0	-3.1	4.7	-4.0	-7.0	2.7	-1.4	-1.6	-2.3
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	4.0	-3.6	4.2	-4.2	-7.2	2.5	-2.0	-1.6	-2.7
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	3.2	-4.2	3.5	-5.0	-7.8	1.7	-2.7	-2.3	-3.4
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	3.4	-4.0	3.8	-4.6	-7.5	2.1	-2.4	-2.0	-3.2
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	4.0	-3.6	4.3	-4.2	-7.2	2.5	-1.9	-1.6	-2.6

B.7.1.9 Percent Relative Median Difference Table for Variance Estimators

Table B.11: Percent Relative Median Difference of LGREG Variance Estimators for Synthetic Population

	Small Samples								
	Fixed SRS			Rate SRS			Fixed PPS		
Estimator	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	54.0	83.2	53.8	40.7	50.7	38.6	65.1	61.1	58.5
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	46.3	74.0	46.1	33.7	43.2	31.6	56.9	53.1	50.6
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	66.5	97.7	69.0	57.9	70.3	58.9	77.5	73.7	74.2
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	32.7	24.7	18.5	12.4	10.1	-0.9	21.3	22.9	15.0
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	26.1	18.5	12.6	6.8	4.5	-5.9	15.3	16.8	9.2
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	42.8	35.8	28.3	28.1	26.4	16.9	31.0	32.8	22.9
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	37.3	26.0	26.3	19.4	12.7	10.3	25.6	24.6	21.8
	Large Samples								
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	131.3	112.6	120.6	110.8	91.3	125.1	105.9	118.3	104.9
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	129.7	111.2	119.1	109.3	89.9	123.5	104.4	116.7	103.4
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	131.7	112.9	120.6	111.0	91.4	125.6	106.2	118.6	105.1
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	129.9	112.6	123.1	114.7	88.5	121.0	102.6	121.5	105.6
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	128.3	111.1	121.6	113.3	87.1	119.5	101.2	119.9	104.0
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	128.9	111.5	121.8	113.5	87.6	119.5	101.2	120.3	104.1
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	129.9	112.5	123.3	114.9	88.5	120.9	102.6	121.5	105.5

B.7.1.10 Percent Relative Root Mean Squared Error Table for LGREG Variance Estimators

Table B.12: Percent Relative Root Mean Squared Error of LGREG Variance Estimators for Synthetic Population

	Small Samples								
	Fixed SRS			Rate SRS			Fixed PPS		
Estimator	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	51.4	49.6	50.2	50.5	48.8	53.5	43.0	41.7	46.8
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	50.9	48.9	50.2	50.8	49.3	53.8	43.4	42.2	47.1
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	48.5	47.9	53.0	60.9	54.5	62.8	49.9	46.0	52.0
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	51.6	50.3	53.0	57.7	57.8	58.9	50.1	51.3	51.9
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	52.8	51.6	54.2	59.2	59.4	60.2	51.8	53.1	53.4
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	55.7	57.2	64.0	75.4	75.5	72.5	63.7	60.4	64.0
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	50.5	50.0	52.1	55.9	57.2	56.9	48.7	50.9	50.5
	Large Samples								
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	8.1	7.0	9.0	7.6	9.2	7.8	5.5	5.2	6.3
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	7.6	7.2	8.5	7.8	9.6	7.4	5.7	5.4	6.6
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	7.9	6.7	8.7	7.4	9.1	7.8	5.5	5.1	6.3
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	8.1	7.2	8.8	7.6	9.2	7.8	5.6	5.2	6.5
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	7.6	7.4	8.3	7.9	9.7	7.4	5.8	5.3	6.7
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	7.6	7.1	8.1	7.7	9.4	7.5	5.7	5.2	6.6
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	8.1	7.2	8.8	7.6	9.2	7.7	5.6	5.2	6.4

B.7.1.11 Percent Relative Root Median Squared Error Table for LGREG Variance Estimators

Table B.13: Percent Relative Root Mean Squared Error of LGREG Variance Estimators for Synthetic Population

	Small Samples								
	Fixed SRS			Rate SRS			Fixed PPS		
Estimator	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	56.6	81.2	56.9	50.6	53.9	50.4	66.1	61.2	60.4
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	50.7	72.1	52.2	46.4	50.1	47.9	58.3	53.5	53.8
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	67.7	95.7	67.9	60.4	70.3	63.7	77.9	72.8	73.5
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	42.0	38.1	41.0	40.5	38.0	40.7	36.3	35.5	37.9
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	40.1	35.9	39.0	39.4	37.2	40.2	34.0	32.5	36.1
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	48.4	43.4	42.7	44.8	43.7	43.9	42.0	42.8	40.2
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	44.9	38.7	42.4	41.7	38.5	41.5	39.0	36.2	40.7
	Large Samples								
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	132.7	112.1	120.4	109.6	90.6	125.3	106.3	118.6	104.8
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	131.1	110.6	118.9	108.1	89.2	123.7	104.7	117.1	103.3
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	133.1	112.3	120.4	109.8	90.7	125.8	106.5	119.0	105.1
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	129.9	112.6	123.1	114.7	88.5	121.0	102.6	121.5	105.6
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	128.3	111.1	121.6	113.3	87.1	119.5	101.2	119.9	104.0
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	128.9	111.5	121.8	113.5	87.6	119.5	101.2	120.3	104.1
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	129.9	112.5	123.3	114.9	88.5	120.9	102.6	121.5	105.5

B.7.1.12 Average Distance from Empirical Value for Variance Estimators

Table B.14: Average Distance from Empirical Value for Standard Error Estimators in Synthetic Population (in thousands)

	Small Samples			Large Samples		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	122.2	136.4	110.8	1.9	2.0	1.4
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	125.5	140.6	114.8	1.9	2.1	1.4
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	115.8	132.0	111.4	1.9	2.0	1.4
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	158.4	199.8	161.2	1.9	2.0	1.4
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	164.8	207.0	168.7	1.9	2.1	1.4
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	155.3	193.6	162.4	1.8	2.0	1.4
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	153.8	191.6	156.2	1.9	2.0	1.4

B.7.1.13 Median Distance from Empirical Value for Variance Estimators

Table B.15: Median Distance from Empirical Value for Standard Error Estimators in Synthetic Population (in thousands)

	Small Samples			Large Samples		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	122.8	106.8	113.8	18.3	16.7	16.7
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	114.5	100.7	105.1	18.1	16.5	16.5
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	133.8	126.0	127.7	18.3	16.7	16.8
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	92.8	95.0	87.3	18.3	16.6	16.8
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	89.5	93.3	84.3	18.1	16.4	16.6
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	99.9	105.7	97.2	18.1	16.4	16.6
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	95.6	98.4	91.2	18.3	16.6	16.8

B.7.1.14 Standard Error of Average Distance from Empirical Value for Variance Estimators

Table B.16: Standard Error of Average Distance from Empirical Value for Standard Error Estimators in Synthetic Population (in thousands)

	Small Samples			Large Samples		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	53.4	56.5	46.1	0.9	0.9	0.7
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	53.7	57.4	47.2	0.9	0.9	0.7
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	52.0	62.5	49.2	0.9	0.8	0.7
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	59.5	71.8	58.2	0.9	0.9	0.7
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	60.2	71.9	58.6	0.8	0.9	0.7
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	62.3	82.7	66.6	0.8	0.8	0.7
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	58.3	69.4	57.8	0.9	0.9	0.7

B.7.1.15 95% Confidence Interval Coverage Table for Variance Estimators

Table B.17: Percent 95% Confidence Interval Coverage of LGREG Variance Estimators for Synthetic Population

	Small Samples								
	Fixed SRS			Rate SRS			Fixed PPS		
Estimator	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	89.6	91.2	87.5	87.2	88.1	86.2	90.2	91.0	89.0
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	89.2	90.5	86.7	86.3	87.4	85.7	89.8	90.4	88.3
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	92.6	92.8	91.6	91.2	91.2	90.2	92.7	93.5	92.0
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	86.2	87.0	84.7	82.5	83.3	80.8	86.1	86.4	85.8
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	85.3	86.0	83.8	81.7	82.5	79.9	85.3	85.5	84.9
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	88.2	88.7	87.4	86.8	87.0	85.5	88.7	89.0	88.4
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	87.2	87.2	85.4	83.9	83.5	83.2	87.0	86.7	86.9
	Large Samples								
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	95.5	95.1	95.5	94.3	94.3	95.0	95.0	95.3	94.7
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	95.5	95.1	95.4	94.2	94.2	95.0	94.8	95.1	94.7
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	95.4	95.2	95.3	94.3	94.3	95.2	95.2	95.3	94.9
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	95.2	94.8	95.5	94.2	94.2	95.1	94.8	95.2	94.7
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	95.2	94.8	95.3	94.2	94.0	95.0	94.8	95.1	94.7
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	95.3	95.0	95.4	94.0	94.2	95.0	95.2	95.2	94.8
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	95.2	94.8	95.4	94.2	94.2	95.0	94.8	95.2	94.7

B.7.1.16 Plots

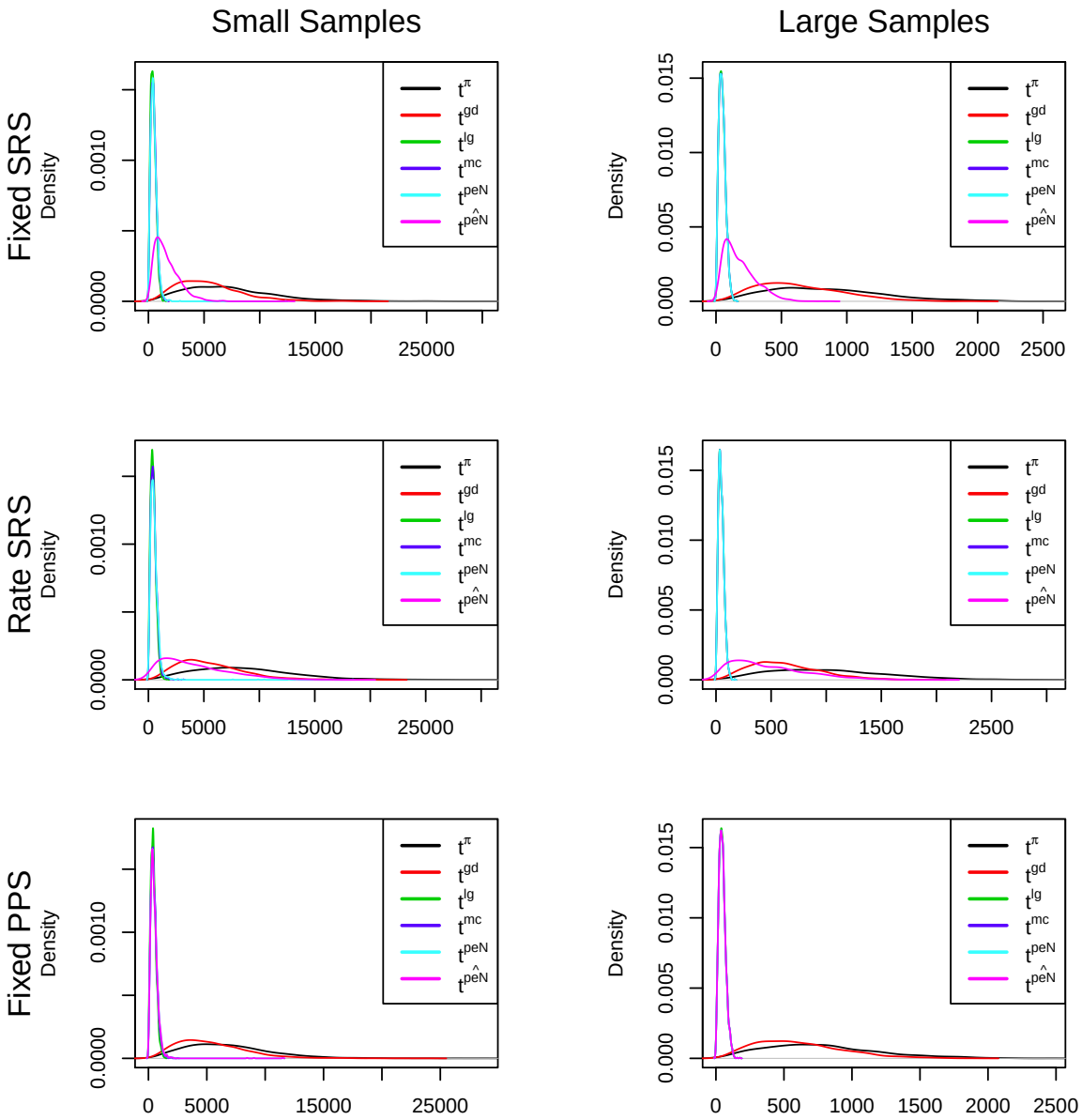

Figure B.1: Density Plot of Distance Between Estimator and True Value for Synthetic Population

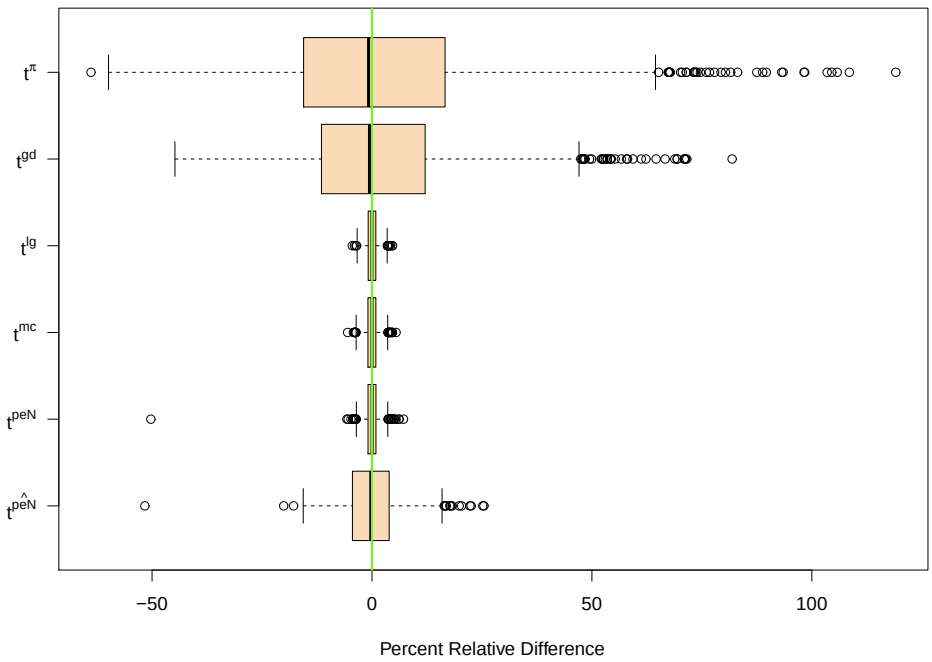

Figure B.2: Box and whisker plot showing percent relative difference of estimated totals for y_1 of synthetic population under small fixed SRS

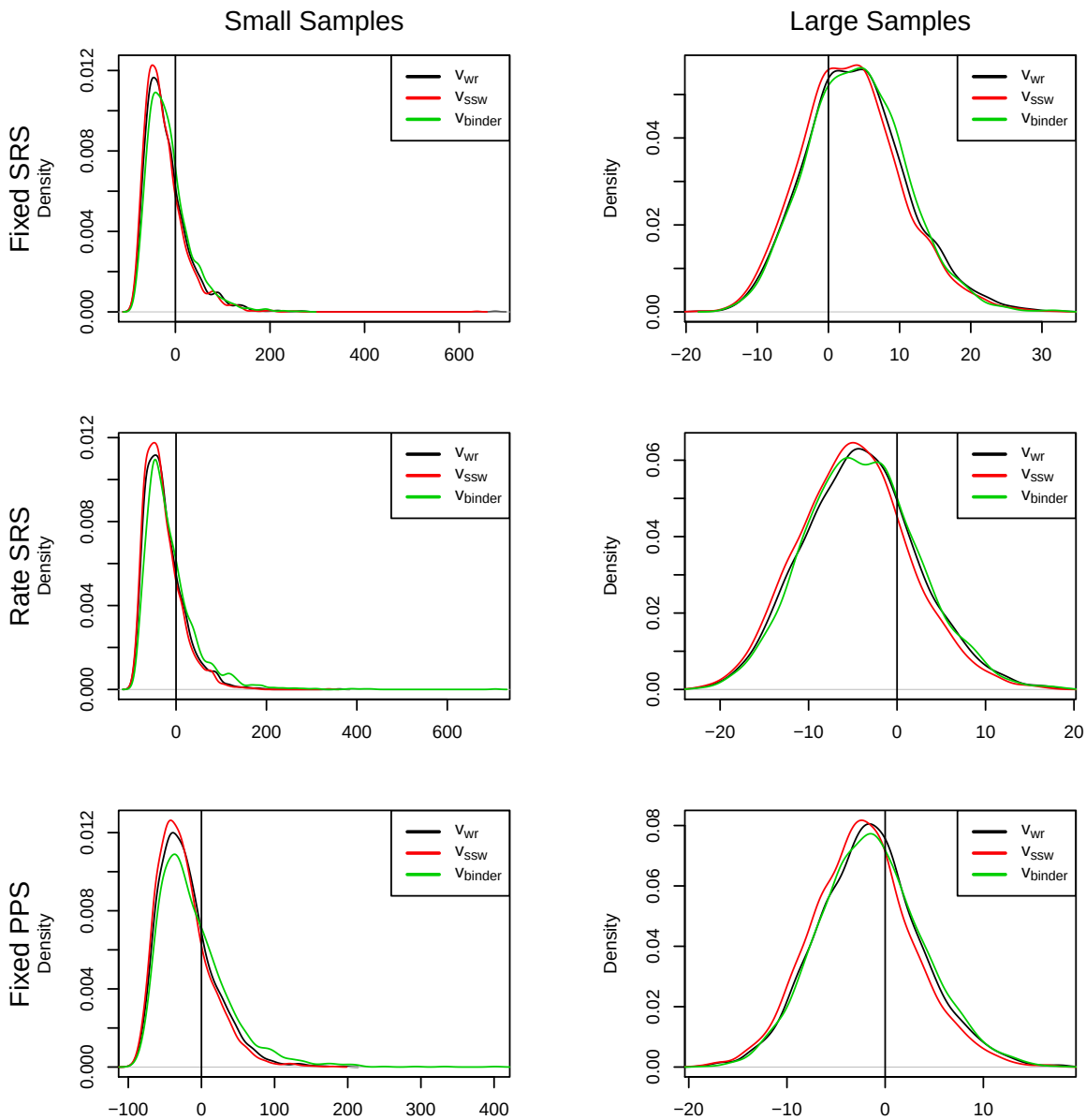

Figure B.3: Density Plot of Distance Between Variance Estimators of y_1 and Empirical Variance for Synthetic Population

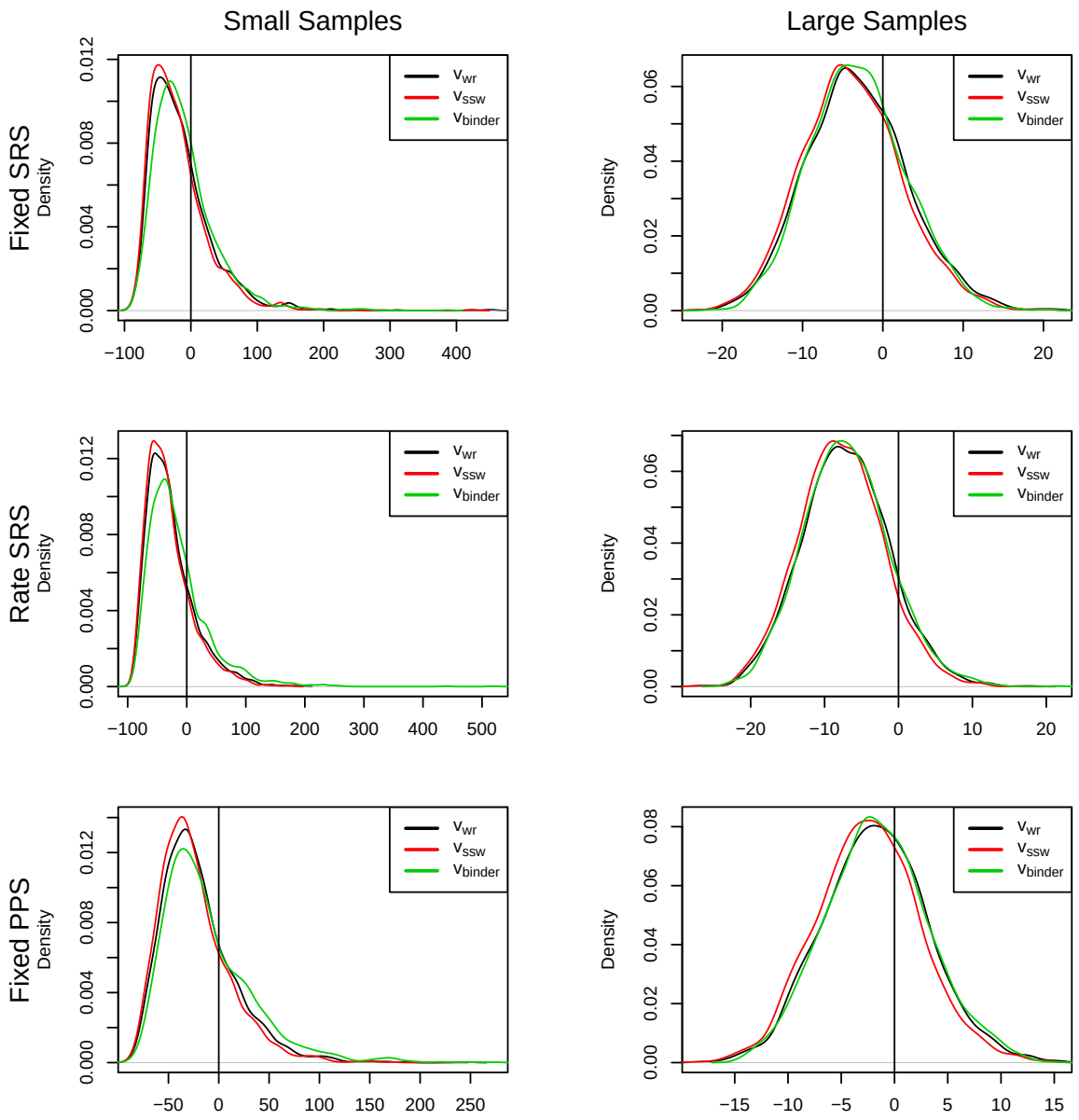

Figure B.4: Density Plot of Distance Between Variance Estimators of y_2 and Empirical Variance for Synthetic Population

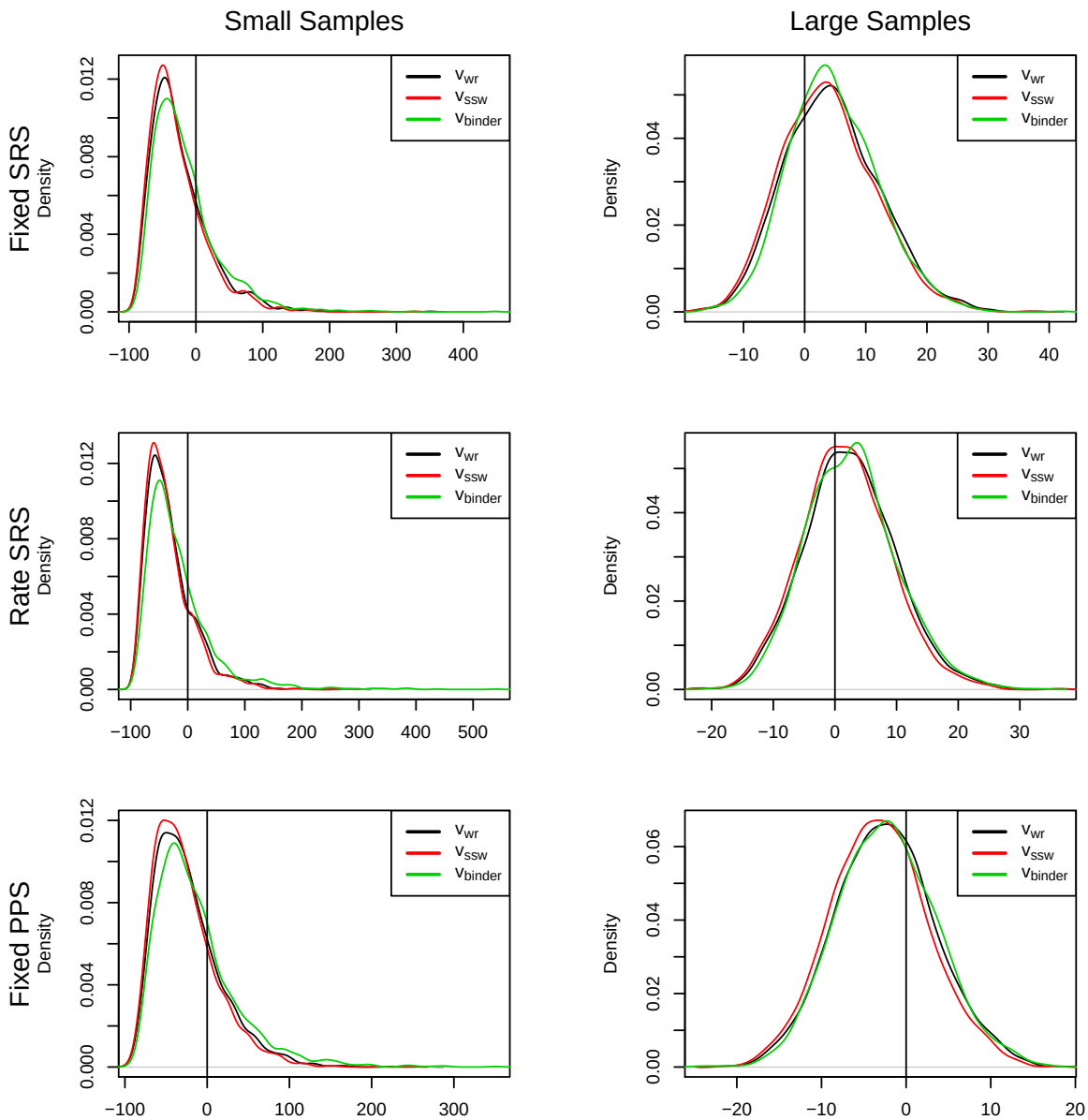

Figure B.5: Density Plot of Distance Between Variance Estimators of y_3 and Empirical Variance for Synthetic Population

B.7.2 Post-Secondary Population

B.7.2.1 Percent Simulation Coefficient of Variation Table

Table B.18: Percent Simulation Coefficient of Variation for Post-Secondary Population

	Small Samples											
	Fixed SRS				Rate SRS				Fixed PPS			
Estimator	Math	Health	Business	Other	Math	Health	Business	Other	Math	Health	Business	Other
$\hat{\mathbf{t}}_y^{\pi}$	0.706	0.299	0.448	0.334	0.688	0.320	0.477	0.343	0.664	0.280	0.451	0.302
$\hat{\mathbf{t}}_{yc}^{GD}$	0.564	0.277	0.363	0.183	0.577	0.277	0.377	0.189	0.566	0.274	0.354	0.181
$\hat{\mathbf{t}}_y^{LG}$	7.198	0.249	0.342	0.115	8.522	0.251	0.355	0.121	7.999	0.250	0.363	0.120
$\hat{\mathbf{t}}_y^{mc}$	7.768	4.497	3.649	1.658	9.048	5.458	3.569	1.878	8.888	4.296	4.065	1.531
$\hat{\mathbf{t}}_y^{peN}$	0.565	0.297	0.371	0.193	0.555	0.299	0.376	0.196	0.554	0.294	0.378	0.197
$\hat{\mathbf{t}}_y^{pe\hat{N}}$	0.575	0.319	0.387	0.221	0.576	0.344	0.410	0.258	0.554	0.294	0.378	0.197
	Large Samples											
$\hat{\mathbf{t}}_y^{\pi}$	0.306	0.135	0.204	0.147	0.302	0.142	0.209	0.149	0.295	0.124	0.199	0.133
$\hat{\mathbf{t}}_{yc}^{GD}$	0.255	0.117	0.158	0.078	0.258	0.113	0.159	0.076	0.254	0.111	0.153	0.075
$\hat{\mathbf{t}}_y^{LG}$	0.259	0.107	0.126	0.042	0.269	0.102	0.121	0.041	0.256	0.102	0.119	0.040
$\hat{\mathbf{t}}_y^{mc}$	0.266	0.115	0.139	0.046	0.298	0.111	0.136	0.045	0.286	0.110	0.137	0.045
$\hat{\mathbf{t}}_y^{peN}$	0.259	0.117	0.141	0.047	0.279	0.115	0.141	0.049	0.279	0.114	0.140	0.048
$\hat{\mathbf{t}}_y^{pe\hat{N}}$	0.265	0.128	0.148	0.070	0.289	0.138	0.159	0.094	0.279	0.114	0.140	0.048

B.7.2.2 Average Distance from True Value

Table B.19: Average Distance from True Value for Post-Secondary Population (in thousands)

	Small Samples			Large Samples		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
$\hat{\mathbf{t}}_y^{\pi}$	576	602	539	262	268	242
$\hat{\mathbf{t}}_{yc}^{gd}$	367	371	365	164	160	157
$\hat{\mathbf{t}}_y^{lg}$	274	284	284	107	103	102
$\hat{\mathbf{t}}_y^{mc}$	898	1,084	1,067	114	111	111
$\hat{\mathbf{t}}_y^{peM}$	364	372	374	116	115	114
$\hat{\mathbf{t}}_y^{pe\hat{M}}$	429	497	374	154	188	114

B.7.2.3 Empirical Standard Deviation of Distance from True Value

Table B.20: Empirical Standard Deviation of Distance from True Value for Post-Secondary Population (in thousands)

	Small Samples			Large Samples		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
$\hat{\mathbf{t}}_y^{\pi}$	388.5	394.0	341.4	160.4	163.2	142.5
$\hat{\mathbf{t}}_{yc}^{gd}$	221.8	238.7	214.0	82.3	83.5	79.4
$\hat{\mathbf{t}}_y^{lg}$	200.1	216.3	212.3	60.7	57.9	56.7
$\hat{\mathbf{t}}_y^{mc}$	3,638.8	3,577.1	3,807.6	68.4	68.5	69.7
$\hat{\mathbf{t}}_y^{peM}$	264.9	263.3	262.3	71.9	76.3	75.9
$\hat{\mathbf{t}}_y^{pe\hat{M}}$	259.9	281.1	262.3	78.5	102.7	75.9

B.7.2.4 Percent Relative Bias Table

Table B.21: Percent Relative Bias for Post-Secondary Population

	Small Samples											
	Fixed SRS				Rate SRS				Fixed PPS			
Estimator	Math	Health	Business	Other	Math	Health	Business	Other	Math	Health	Business	Other
$\hat{\mathbf{t}}_y^{\pi}$	0.6	-0.1	0.3	0.5	1.7	0.7	1.1	1.4	2.2	-0.1	0.9	0.7
$\hat{\mathbf{t}}_{yc}^{gd}$	-6.4	0.5	-3.3	-1.8	-5.2	0.3	-3.5	-1.5	-4.1	0.4	-2.8	-1.4
$\hat{\mathbf{t}}_y^{lg}$	171.7	1.5	3.5	-3.0	216.3	1.1	3.4	-3.1	210.3	1.1	3.8	-3.2
$\hat{\mathbf{t}}_y^{mc}$	160.1	-21.3	0.0	5.9	210.0	-25.8	-3.3	7.9	206.8	-28.2	4.1	6.8
$\hat{\mathbf{t}}_y^{peM}$	-11.3	0.4	-4.1	-6.1	-11.4	0.1	-4.5	-6.2	-11.0	-0.9	-4.9	-6.6
$\hat{\mathbf{t}}_y^{pe\hat{M}}$	-11.6	0.2	-4.5	-6.4	-10.7	0.7	-3.9	-5.5	-11.0	-0.9	-4.9	-6.6
	Large Samples											
$\hat{\mathbf{t}}_y^{\pi}$	0.0	0.1	0.1	0.0	-0.2	0.0	0.0	0.0	-0.2	0.0	-0.1	-0.1
$\hat{\mathbf{t}}_{yc}^{gd}$	-1.5	0.1	-0.9	-0.5	-1.2	0.1	-0.7	-0.4	-1.1	0.0	-0.8	-0.4
$\hat{\mathbf{t}}_y^{lg}$	3.2	0.6	0.4	-0.4	1.8	0.3	0.0	-0.1	1.8	0.3	-0.1	-0.1
$\hat{\mathbf{t}}_y^{mc}$	0.0	0.6	0.4	-0.3	1.1	0.5	0.2	-0.2	1.0	0.6	-0.1	-0.2
$\hat{\mathbf{t}}_y^{peM}$	-1.0	0.7	0.4	-0.4	-0.2	0.6	0.2	-0.4	0.0	0.6	0.0	-0.3
$\hat{\mathbf{t}}_y^{pe\hat{M}}$	-0.9	0.7	0.4	-0.4	-0.2	0.5	0.2	-0.4	0.0	0.6	0.0	-0.3

B.7.2.5 Percent Relative Median Difference Table

Table B.22: Percent Relative Median Difference for Post-Secondary Population

	Small Samples											
	Fixed SRS				Rate SRS				Fixed PPS			
Estimator	Math	Health	Business	Other	Math	Health	Business	Other	Math	Health	Business	Other
$\hat{\mathbf{t}}_y^{\pi}$	-15.3	-4.3	-8.0	-4.2	-12.2	-2.8	-7.1	-2.3	-10.8	-2.8	-6.7	-2.5
$\hat{\mathbf{t}}_{yc}^{gd}$	-15.5	-2.8	-8.3	-2.8	-14.4	-2.5	-8.8	-2.5	-13.1	-2.6	-7.9	-2.1
$\hat{\mathbf{t}}_y^{lg}$	3.7	-0.4	-2.2	-2.0	4.4	-1.1	-2.6	-2.0	5.4	-1.2	-2.8	-2.2
$\hat{\mathbf{t}}_y^{mc}$	-3.0	-0.3	-2.1	-1.0	-0.6	-0.9	-2.7	-0.9	-0.2	-1.1	-3.0	-1.0
$\hat{\mathbf{t}}_y^{peM}$	-20.9	-2.3	-8.8	-3.6	-21.1	-2.7	-9.3	-4.0	-20.4	-3.6	-9.8	-4.1
$\hat{\mathbf{t}}_y^{pe\hat{M}}$	-21.8	-4.0	-10.0	-6.2	-21.3	-3.1	-10.2	-6.0	-20.4	-3.6	-9.8	-4.1
	Large Samples											
$\hat{\mathbf{t}}_y^{\pi}$	-4.4	-0.5	-2.2	-1.0	-3.7	-0.7	-2.2	-0.8	-4.3	-0.5	-2.5	-0.8
$\hat{\mathbf{t}}_{yc}^{gd}$	-4.8	-0.5	-3.0	-0.9	-4.2	-0.2	-2.9	-0.8	-4.2	-0.6	-2.8	-0.6
$\hat{\mathbf{t}}_y^{lg}$	0.1	0.1	-0.6	-0.2	-1.8	0.0	-1.0	0.0	-1.5	0.0	-1.1	0.0
$\hat{\mathbf{t}}_y^{mc}$	-3.8	0.1	-0.8	-0.2	-3.2	-0.1	-1.1	0.0	-2.7	0.0	-1.2	0.0
$\hat{\mathbf{t}}_y^{peM}$	-4.5	0.0	-0.9	-0.2	-3.9	-0.2	-1.2	-0.1	-3.6	-0.1	-1.4	-0.1
$\hat{\mathbf{t}}_y^{pe\hat{M}}$	-4.5	-0.2	-1.0	-0.5	-3.8	-0.1	-1.0	-0.3	-3.6	-0.1	-1.4	-0.1

B.7.2.6 Percent Relative Root Mean Squared Error Table

Table B.23: Percent Relative Root Mean Squared Error for Post-Secondary Population

					Small Samples							
		Fixed SRS				Rate SRS				Fixed PPS		
Estimator	Math	Health	Business	Other	Math	Health	Business	Other	Math	Health	Business	Other
$\hat{\mathbf{t}}_y^{\pi}$	70.7	29.9	44.7	33.3	68.8	32.0	47.9	34.2	66.3	28.0	45.0	30.1
$\hat{\mathbf{t}}_{yc}^{gd}$	56.4	27.7	36.3	18.3	57.7	27.6	37.7	18.9	56.5	27.5	35.4	18.1
$\hat{\mathbf{t}}_y^{lg}$	691.6	24.9	34.2	11.4	811.4	25.2	35.4	12.0	761.9	25.0	36.3	11.9
$\hat{\mathbf{t}}_y^{mc}$	748.2	325.4	330.3	138.7	867.3	350.0	331.1	130.3	841.0	370.1	359.1	136.3
$\hat{\mathbf{t}}_y^{peM}$	56.4	29.7	37.0	19.2	55.5	29.8	37.6	19.5	55.3	29.4	37.8	19.6
$\hat{\mathbf{t}}_y^{pe\hat{M}}$	57.5	31.9	38.6	22.0	57.6	34.4	40.9	25.7	55.3	29.4	37.8	19.6
					Large Samples							
$\hat{\mathbf{t}}_y^{\pi}$	30.6	13.5	20.4	14.7	30.1	14.2	20.9	14.9	29.5	12.4	19.9	13.3
$\hat{\mathbf{t}}_{yc}^{gd}$	25.5	11.7	15.8	7.8	25.8	11.3	15.9	7.7	25.4	11.1	15.3	7.5
$\hat{\mathbf{t}}_y^{lg}$	25.9	10.7	12.6	4.2	26.9	10.2	12.1	4.1	25.6	10.2	11.9	4.0
$\hat{\mathbf{t}}_y^{mc}$	26.6	11.5	13.8	4.5	29.8	11.1	13.6	4.5	28.6	11.0	13.7	4.5
$\hat{\mathbf{t}}_y^{peM}$	25.9	11.7	14.1	4.7	27.9	11.5	14.1	4.9	27.9	11.4	14.0	4.8
$\hat{\mathbf{t}}_y^{pe\hat{M}}$	26.5	12.8	14.8	7.0	28.9	13.8	15.9	9.4	27.9	11.4	14.0	4.8

B.7.2.7 Percent Relative Root Median Squared Error Table

Table B.24: Percent Relative Root Median Squared Error for Post-Secondary Population

	Small Samples											
	Fixed SRS				Rate SRS				Fixed PPS			
Estimator	Math	Health	Business	Other	Math	Health	Business	Other	Math	Health	Business	Other
$\hat{\mathbf{t}}_y^{\pi}$	41.0	19.6	28.1	21.5	41.0	21.2	28.1	22.9	39.6	18.9	25.6	20.2
$\hat{\mathbf{t}}_{yc}^{gd}$	34.6	17.0	20.7	11.5	34.9	17.4	20.4	11.7	34.6	17.2	20.2	11.6
$\hat{\mathbf{t}}_y^{lg}$	37.8	15.7	18.6	6.5	41.2	15.9	18.5	6.8	40.4	15.8	18.8	6.7
$\hat{\mathbf{t}}_y^{mc}$	40.2	21.4	23.8	8.6	45.1	23.2	25.2	9.3	45.0	23.2	25.5	9.4
$\hat{\mathbf{t}}_y^{peM}$	35.4	18.2	22.3	9.3	35.9	18.5	22.6	9.8	35.7	18.7	22.4	10.1
$\hat{\mathbf{t}}_y^{pe\hat{M}}$	36.6	19.7	23.5	13.6	37.7	21.8	25.7	17.4	35.7	18.7	22.4	10.1
	Large Samples											
$\hat{\mathbf{t}}_y^{\pi}$	20.2	8.9	13.6	9.9	19.4	9.7	13.6	10.0	19.2	8.4	12.6	9.0
$\hat{\mathbf{t}}_{yc}^{gd}$	16.4	7.8	10.3	5.3	16.0	7.6	9.6	5.1	15.8	7.4	9.7	5.0
$\hat{\mathbf{t}}_y^{lg}$	14.5	7.0	8.1	2.8	14.2	6.8	7.8	2.7	14.2	6.9	7.8	2.7
$\hat{\mathbf{t}}_y^{mc}$	14.5	7.4	8.8	2.9	14.9	7.0	8.2	2.8	14.7	7.1	8.3	2.8
$\hat{\mathbf{t}}_y^{peM}$	14.6	7.4	8.8	3.0	14.9	7.2	8.3	2.9	15.0	7.2	8.4	2.9
$\hat{\mathbf{t}}_y^{pe\hat{M}}$	15.1	8.1	9.4	4.6	15.9	9.0	9.9	6.3	15.0	7.2	8.4	2.9

B.7.2.8 Percent Relative Bias Table for LGREG Variance Estimators

Table B.25: Percent Relative Bias of LGREG Variance Estimators for NSCG Population

	Small Samples											
	Fixed SRS				Rate SRS				Fixed PPS			
Estimator	Math	Health	Business	Other	Math	Health	Business	Other	Math	Health	Business	Other
$v_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-99.9	-49.0	-76.9	-66.6	-99.9	-51.7	-79.3	-70.4	-99.9	-50.8	-80.0	-69.9
$v_e\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-99.9	-54.3	-79.3	-70.0	-99.9	-56.7	-81.4	-73.5	-99.9	-55.9	-82.0	-73.0
$v_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-65.1	-34.1	-53.1	-47.1	-68.1	-37.2	-53.5	-51.1	-64.6	-33.9	-55.0	-49.5
$v_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-99.9	-99.9	-99.8	-99.9	-99.9	-99.9	-99.8	-99.9	-99.9	-99.9	-99.9	-99.9
$v_e\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-99.9	-99.9	-99.9	-99.9	-99.9	-99.9	-99.8	-99.9	-99.9	-99.9	-99.9	-99.9
$v_g\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-85.8	-48.1	-38.8	-58.6	-88.3	-68.2	-52.7	-67.8	-87.7	-33.3	-55.6	-47.4
$v_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-11.6	338.8	135.6	303.4	-17.7	459.0	377.4	462.8	-16.0	202.4	139.3	221.2
	Large Samples											
$v_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-53.2	-14.2	-30.0	-18.9	-59.4	-10.7	-30.8	-20.1	-55.1	-10.7	-26.7	-17.6
$v_e\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-55.5	-18.3	-33.3	-22.8	-61.6	-15.2	-34.1	-24.3	-57.5	-15.3	-30.4	-21.9
$v_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-33.8	-12.2	-28.8	-18.9	-33.2	-7.1	-27.3	-17.2	-28.7	-7.4	-23.3	-14.4
$v_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-60.5	-30.7	-49.4	-37.0	-69.9	-31.2	-51.8	-39.8	-67.5	-29.6	-51.9	-41.0
$v_e\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-62.5	-34.0	-51.8	-40.1	-71.6	-34.7	-54.2	-43.0	-69.2	-33.2	-54.3	-44.2
$v_g\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-54.8	-25.1	-45.6	-33.4	-62.5	-22.4	-43.1	-32.4	-60.5	-20.7	-43.3	-33.0
$v_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-30.0	-1.6	-14.9	-8.3	-42.9	2.7	-18.7	-9.9	-42.2	7.0	-21.5	-9.2

B.7.2.9 Percent Relative Median Difference Table for Variance Estimators

Table B.26: Percent Relative Median Difference of LGREG Variance Estimators for Post-Secondary Population

	Small Samples											
	Fixed SRS				Rate SRS				Fixed PPS			
Estimator	Math	Health	Business	Other	Math	Health	Business	Other	Math	Health	Business	Other
$v_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-80.3	-8.3	-56.8	-34.1	-84.7	-11.3	-56.1	-38.8	-83.4	-7.5	-55.2	-36.8
$v_e\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-82.2	-17.6	-61.1	-40.7	-86.2	-20.3	-60.5	-45.0	-85.1	-17.1	-59.7	-43.2
$v_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-42.5	16.1	-24.2	1.8	-46.4	8.8	-27.6	-5.0	-44.8	12.7	-25.3	-3.8
$v_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-84.2	-59.1	-79.1	-67.4	-88.3	-66.1	-81.4	-73.1	-87.7	-64.5	-81.7	-72.4
$v_e\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-85.8	-63.2	-81.2	-70.7	-89.5	-69.6	-83.2	-75.8	-88.9	-68.1	-83.6	-75.2
$v_g\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-70.6	-27.5	-58.6	-42.1	-76.1	-35.6	-61.3	-48.1	-74.8	-33.1	-60.7	-47.8
$v_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-8.3	30.5	-11.8	12.9	-13.0	29.1	-4.7	10.4	-6.6	34.0	-3.2	12.3
	Large Samples											
$v_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	9.8	73.3	24.6	56.2	14.6	81.3	28.8	62.2	12.7	77.7	31.7	64.8
$v_e\left(\widehat{\mathbf{t}}_y^{lg}\right)$	4.3	65.4	18.6	48.8	8.5	72.4	22.7	54.3	6.8	68.5	25.3	56.6
$v_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	25.8	80.0	40.4	66.9	35.7	89.4	45.5	73.3	32.9	84.2	49.1	74.3
$v_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	7.2	46.8	-4.3	29.7	0.1	54.7	3.6	32.2	1.4	56.0	6.4	33.1
$v_e\left(\widehat{\mathbf{t}}_y^{mc}\right)$	1.8	39.9	-8.6	23.5	-5.1	47.1	-1.0	25.8	-4.0	48.0	1.5	26.5
$v_g\left(\widehat{\mathbf{t}}_y^{mc}\right)$	11.9	51.9	5.9	36.3	4.5	61.1	14.7	39.4	6.1	60.5	17.8	39.2
$v_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	27.1	62.6	16.9	45.3	20.8	73.6	28.2	50.9	21.8	73.9	29.6	51.7

B.7.2.10 Percent Relative Root Mean Squared Error Table for LGREG Variance Estimators

Table B.27: Percent Relative Root Mean Squared Error of LGREG Variance Estimators for Post-Secondary Population

	Small Samples											
	Fixed SRS				Rate SRS				Fixed PPS			
Estimator	Math	Health	Business	Other	Math	Health	Business	Other	Math	Health	Business	Other
$v_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	99.9	72.1	88.2	76.8	99.9	71.1	89.7	78.2	99.9	67.5	89.2	76.1
$v_e\left(\widehat{\mathbf{t}}_y^{lg}\right)$	99.9	71.9	88.2	77.9	99.9	71.4	89.6	79.5	99.9	68.5	89.3	77.8
$v_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	247.6	126.6	116.0	97.5	212.0	104.1	146.3	90.5	208.0	144.2	131.9	86.5
$v_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	99.9	99.9	99.8	99.9	99.9	99.9	99.8	99.9	99.9	99.9	99.9	99.9
$v_e\left(\widehat{\mathbf{t}}_y^{mc}\right)$	99.9	99.9	99.9	99.9	99.9	99.9	99.8	99.9	99.9	99.9	99.9	99.9
$v_g\left(\widehat{\mathbf{t}}_y^{mc}\right)$	215.8	1,319.6	1,390.1	1,168.7	173.2	637.2	608.4	528.9	161.5	2,001.3	839.0	986.9
$v_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	1,497.8	16,225.5	5,061.6	14,337.5	1,715.2	25,766.1	16,031.5	23,574.3	1,144.8	5,023.7	4,166.2	5,225.9
	Large Samples											
$v_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	70.2	47.2	75.8	53.9	71.1	41.7	68.9	47.1	70.0	39.3	71.2	44.3
$v_e\left(\widehat{\mathbf{t}}_y^{lg}\right)$	70.5	46.1	74.1	52.6	71.8	40.5	67.3	46.4	70.5	38.4	69.2	43.7
$v_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	192.9	43.6	64.9	44.0	256.5	42.0	61.1	41.9	221.7	40.0	61.4	40.1
$v_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	72.0	46.8	70.4	51.9	75.7	44.0	66.5	50.2	74.4	42.1	66.6	49.5
$v_e\left(\widehat{\mathbf{t}}_y^{mc}\right)$	72.7	47.5	70.4	52.6	76.6	45.1	66.9	51.5	75.3	43.4	67.0	51.1
$v_g\left(\widehat{\mathbf{t}}_y^{mc}\right)$	108.2	69.6	75.3	69.5	125.1	90.8	92.2	72.6	105.4	88.6	95.7	84.7
$v_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	298.9	291.4	511.2	338.1	270.3	353.0	416.4	256.2	193.3	662.0	235.1	403.8

B.7.2.11 Percent Relative Root Median Squared Error Table for LGREG Variance Estimators

Table B.28: Percent Relative Root Median Squared Error of LGREG Variance Estimators for Post-Secondary Population

	Small Samples											
	Fixed SRS				Rate SRS				Fixed PPS			
Estimator	Math	Health	Business	Other	Math	Health	Business	Other	Math	Health	Business	Other
$v_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	83.2	49.8	66.5	51.3	86.2	50.8	65.8	53.4	84.8	48.4	65.2	49.9
$v_e\left(\widehat{\mathbf{t}}_y^{lg}\right)$	84.0	49.4	68.5	53.1	87.2	50.4	68.0	55.4	85.9	47.4	67.2	52.5
$v_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	85.9	49.2	60.0	46.6	89.5	49.8	62.7	49.2	88.5	48.7	61.2	47.0
$v_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	85.7	62.8	80.2	68.6	88.8	68.0	82.2	74.1	88.1	65.7	82.2	72.8
$v_e\left(\widehat{\mathbf{t}}_y^{mc}\right)$	86.9	65.7	82.0	71.5	89.8	70.7	83.9	76.5	89.2	68.8	83.9	75.5
$v_g\left(\widehat{\mathbf{t}}_y^{mc}\right)$	85.8	65.8	79.0	67.6	89.4	71.3	82.5	74.0	88.9	70.4	82.5	73.3
$v_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	89.4	75.7	83.0	75.8	90.9	81.0	85.0	80.9	90.8	80.3	85.8	79.4
	Large Samples											
$v_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	42.8	73.3	44.0	56.3	43.3	81.3	43.8	62.2	42.3	77.7	44.5	64.8
$v_e\left(\widehat{\mathbf{t}}_y^{lg}\right)$	41.3	65.4	42.6	49.5	41.5	72.4	42.0	54.4	40.7	68.5	42.0	56.7
$v_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	46.7	80.0	46.3	66.9	48.9	89.4	48.9	73.3	49.1	84.2	51.1	74.3
$v_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	42.1	47.6	38.7	35.7	40.4	54.9	37.8	37.2	39.9	56.1	37.1	36.3
$v_e\left(\widehat{\mathbf{t}}_y^{mc}\right)$	40.9	41.9	38.8	32.5	39.2	47.6	37.4	33.1	39.0	48.4	36.2	32.3
$v_g\left(\widehat{\mathbf{t}}_y^{mc}\right)$	42.5	52.0	37.0	38.1	41.1	61.1	38.5	40.8	41.9	60.5	38.8	40.0
$v_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	48.7	62.7	43.3	46.6	47.4	73.6	44.8	51.6	46.7	73.9	44.5	52.0

B.7.2.12 Average Distance from Empirical Value for Variance Estimators

Table B.29: Average Distance from Empirical Value for Standard Error Estimators in Post-Secondary Population (in thousands)

	Small Samples			Large Samples		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	201.1	223.8	215.5	32.5	29.6	27.7
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	205.9	228.5	220.6	33.2	30.2	28.3
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	181.6	202.9	196.3	28.4	26.5	24.9
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	4,494.8	5163.4	4342.6	41.4	39.9	39.5
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	4,502.8	5171.1	4,350.5	42.7	41.3	41.0
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	4,436.3	5,014.4	4,245.9	39.7	39.2	39.0
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	4,770.1	5,539.1	4,633.7	43.4	42.1	41.9

B.7.2.13 Median Distance from Empirical Value for Variance Estimators

Table B.30: Median Distance from Empirical Value for Standard Error Estimators in Post-Secondary Population (in thousands)

	Small Samples			Large Samples		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	129.0	129.8	130.4	58.0	56.4	56.2
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	127.2	128.3	128.1	56.8	55.1	54.8
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	134.0	134.8	136.6	58.2	57.2	56.7
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	169.4	183.4	183.3	57.8	56.0	55.6
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	169.9	184.3	184.0	57.2	55.4	55.1
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	183.8	201.7	204.4	57.5	56.3	56.4
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	206.5	228.4	232.7	59.9	58.6	58.7

B.7.2.14 Standard Error of Average Distance from Empirical Value for Variance Estimators

Table B.31: Standard Error of Average Distance from Empirical Value for Standard Error Estimators in Post-Secondary Population (in thousands)

	Small Samples			Large Samples		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	36.4	36.9	35.1	15.1	12.8	12.6
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	35.8	36.1	34.5	14.6	12.5	12.2
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	54.7	59.6	59.0	12.8	11.5	11.3
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	58.1	60.0	53.9	14.5	13.7	13.3
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	54.9	56.5	50.7	14.3	13.7	13.3
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	2,442.8	2032.9	2,493.1	17.0	18.9	18.5
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	8,322.0	11,532.2	6,546.6	46.2	43.9	45.0

B.7.2.15 95% Confidence Interval Coverage Table for Variance Estimators

Table B.32: Percent 95% Confidence Interval Coverage of LGREG Variance Estimators for Post-Secondary Population

	Small Samples											
	Fixed SRS				Rate SRS				Fixed PPS			
Estimator	Math	Health	Business	Other	Math	Health	Business	Other	Math	Health	Business	Other
$v_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	52.0	81.5	65.9	74.5	48.4	79.9	64.8	72.6	49.7	81.9	65.6	74.0
$v_e\left(\widehat{\mathbf{t}}_y^{lg}\right)$	50.6	79.9	64.1	72.9	47.2	78.2	63.1	70.8	48.2	80.2	63.8	72.1
$v_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	76.2	86.8	78.1	83.7	75.3	85.1	76.0	82.0	76.5	86.6	77.2	83.3
$v_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	46.7	63.0	51.5	59.4	43.5	59.2	49.1	55.3	43.8	60.6	49.0	56.6
$v_e\left(\widehat{\mathbf{t}}_y^{mc}\right)$	45.2	61.3	49.5	57.7	42.0	57.6	47.4	53.9	42.4	58.8	47.4	54.8
$v_g\left(\widehat{\mathbf{t}}_y^{mc}\right)$	56.4	80.7	67.9	76.6	53.8	78.3	67.2	74.4	55.9	78.8	68.0	76.0
$v_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	78.3	89.9	82.0	87.8	78.0	88.7	82.1	87.3	79.3	90.2	82.7	88.4
Large Samples												
$v_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	81.4	91.4	84.8	90.4	81.1	92.0	84.5	89.7	80.9	92.6	85.8	90.8
$v_e\left(\widehat{\mathbf{t}}_y^{lg}\right)$	80.5	90.8	83.8	89.7	80.2	91.3	83.7	89.0	80.1	92.1	85.1	90.1
$v_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	86.3	92.6	88.0	91.9	86.5	93.2	87.1	91.5	86.8	93.4	88.7	92.4
$v_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	78.8	88.6	79.8	86.5	77.5	88.6	79.5	86.0	78.1	89.7	80.8	86.8
$v_e\left(\widehat{\mathbf{t}}_y^{mc}\right)$	77.9	87.9	78.7	85.7	76.7	87.8	78.7	85.2	77.1	89.0	80.1	86.0
$v_g\left(\widehat{\mathbf{t}}_y^{mc}\right)$	80.9	90.3	82.5	88.5	80.0	91.0	83.3	88.7	80.2	91.5	84.2	89.3
$v_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	83.0	90.6	83.8	89.0	82.5	91.3	84.1	88.9	82.9	92.0	85.4	90.0

B.7.2.16 Plots

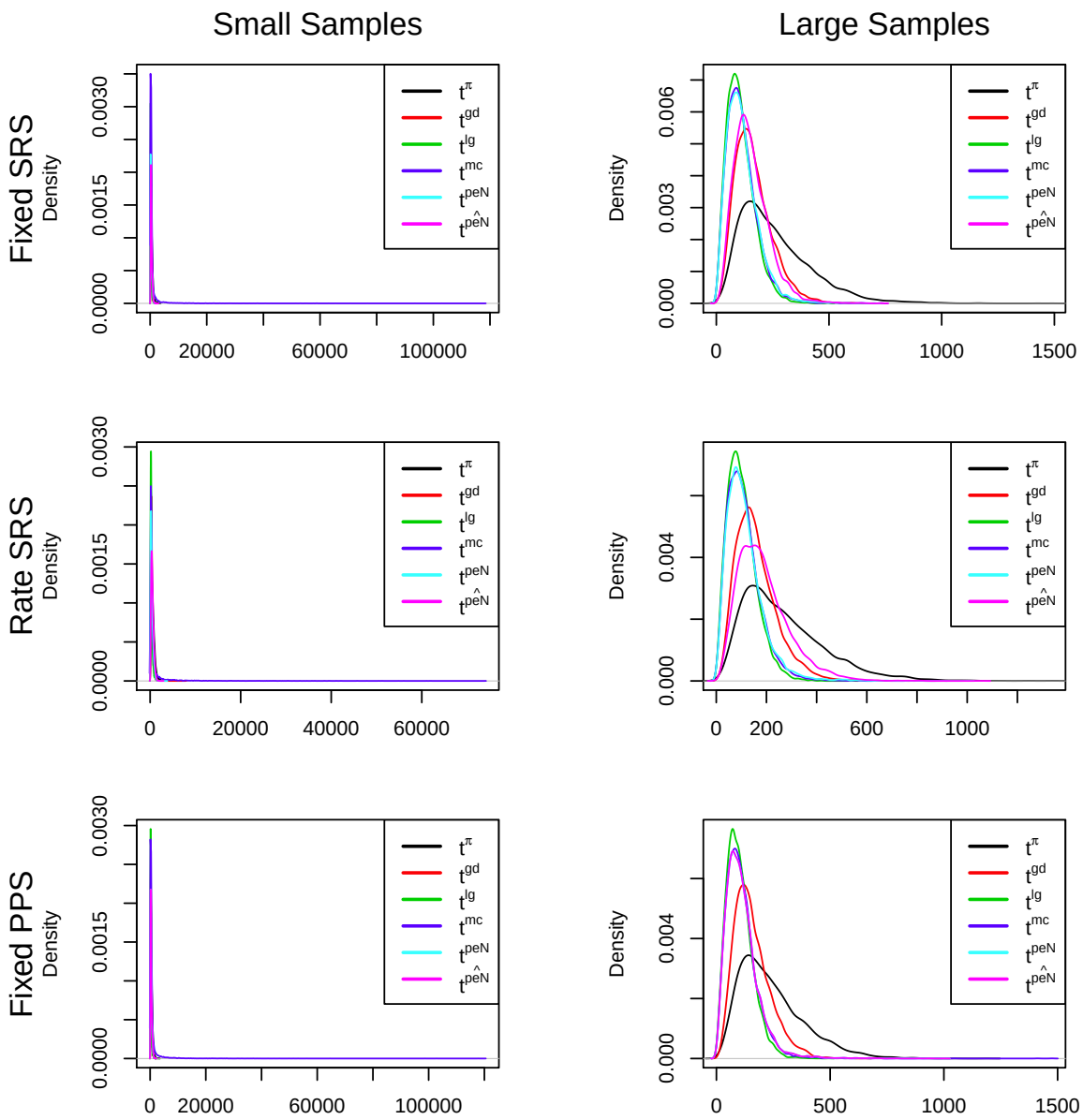

Figure B.6: Density Plot of Distance Between Estimator and True Value in the Post-Secondary Population

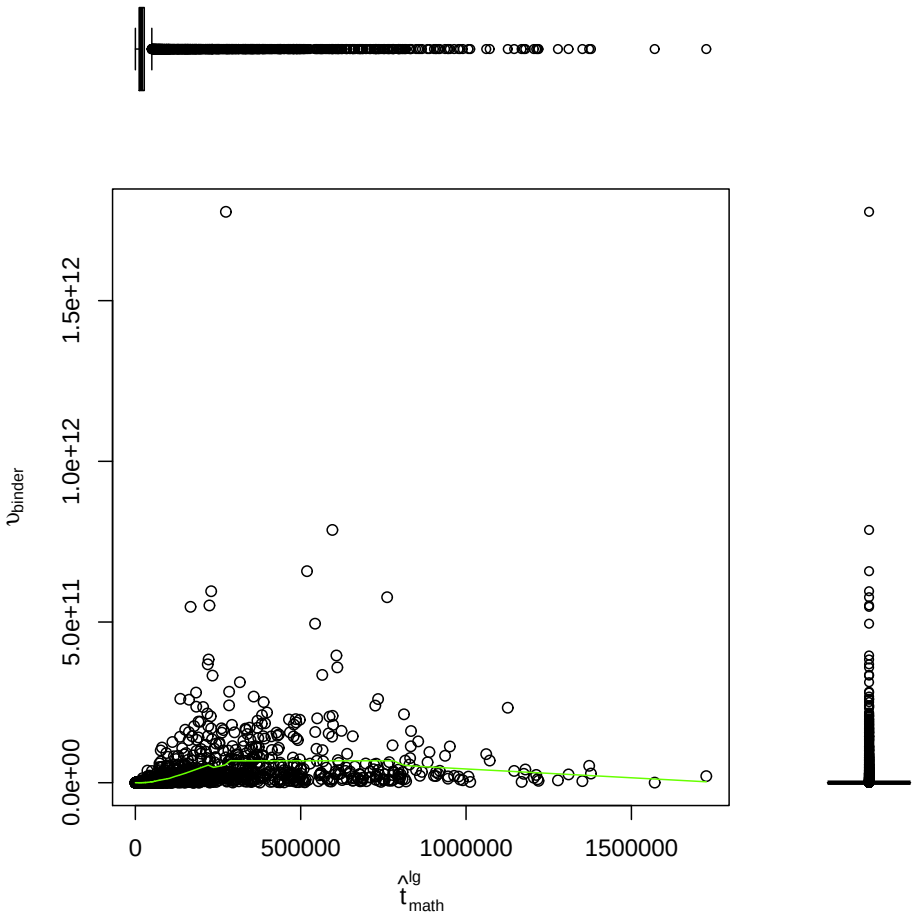

Figure B.7: Plot of LGREG math estimates versus PML LGREG math variance estimates under small fixed SRS in the post-secondary population

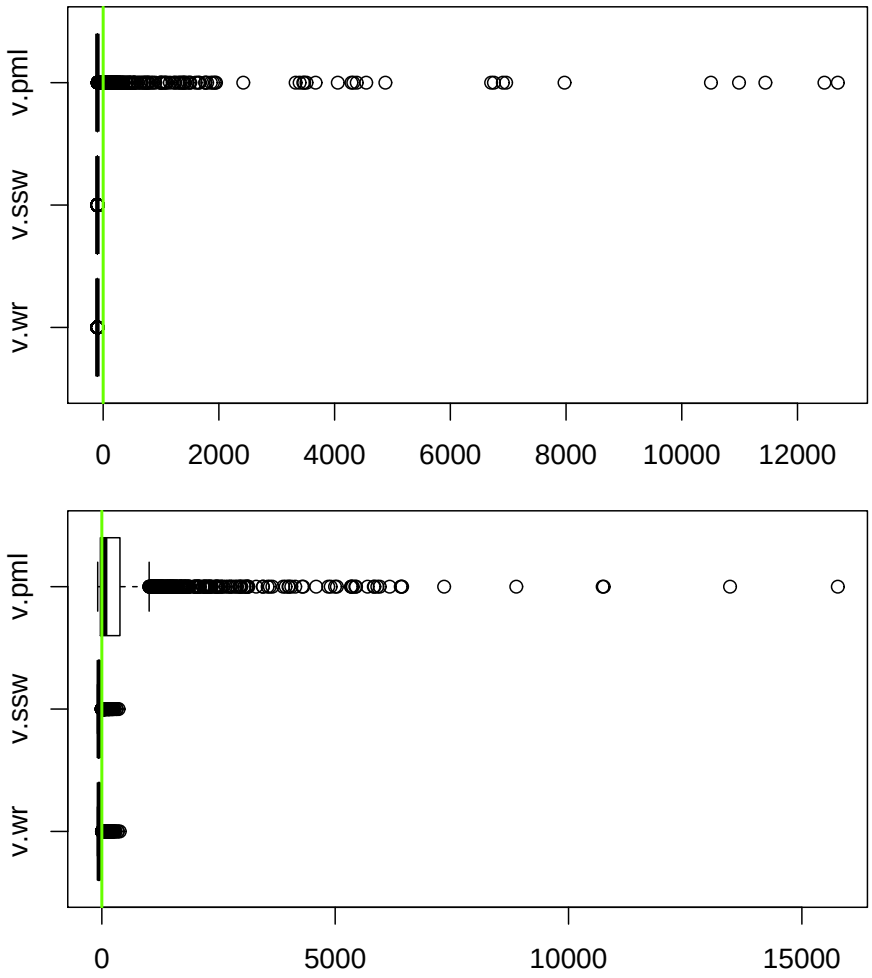

Figure B.8: Box and whisker plots showing percent relative difference of LGREG variance estimators for math in fixed SRS samples from post-secondary population including all outliers. Small sample sizes on top.

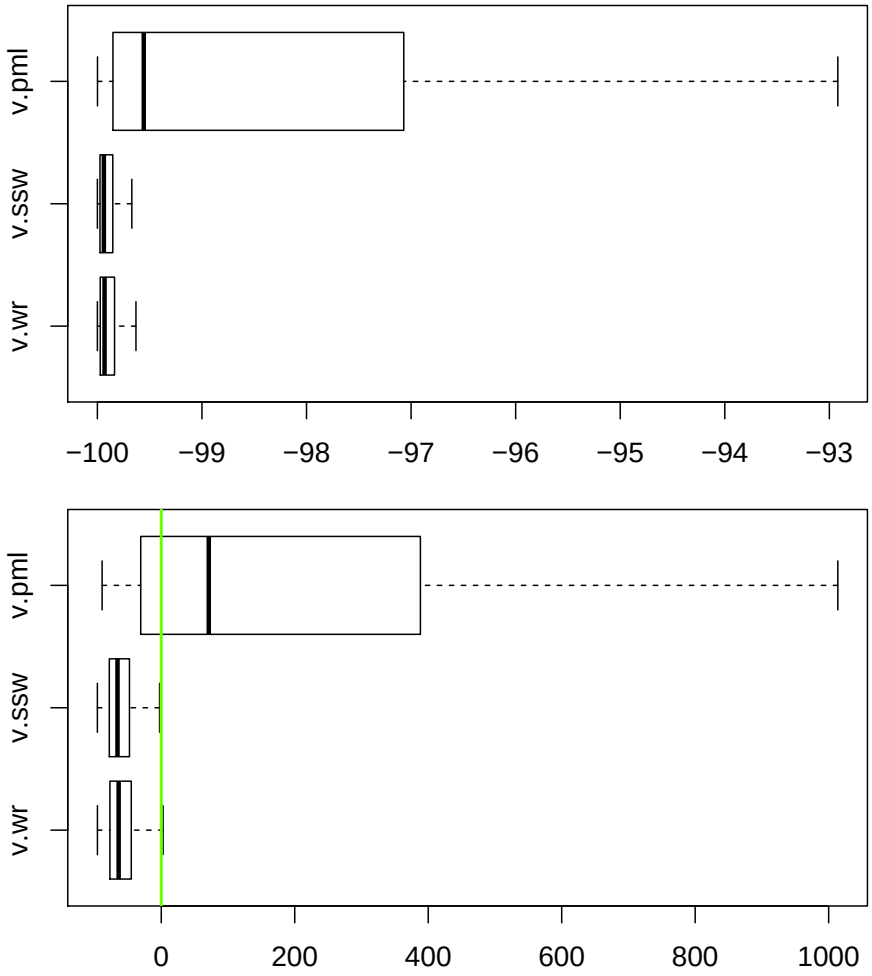

Figure B.9: Box and whisker plots showing percent relative difference of LGREG variance estimators for **math** in fixed SRS samples from post-secondary population excluding all outliers. Outliers are 1.5 times the interquartile range beyond the first and third quartiles. Small sample sizes on top. Large samples on bottom. The empirical variance was calculated **with** the outliers.

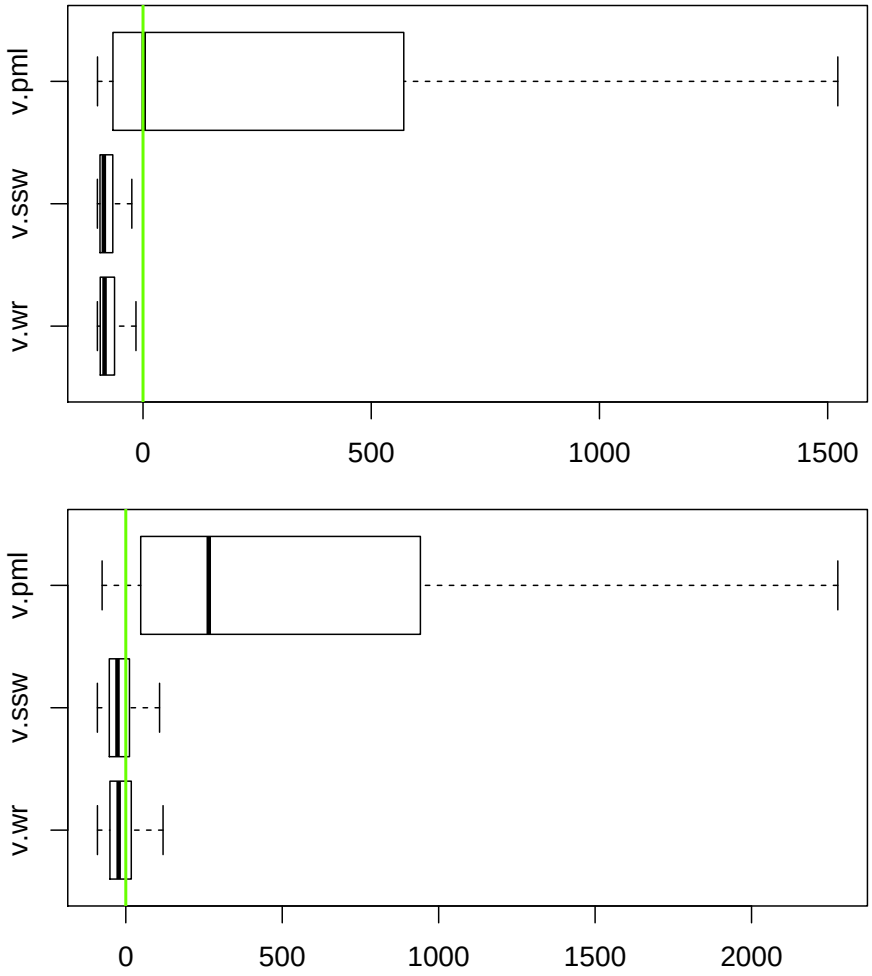

Figure B.10: Box and whisker plots showing percent relative difference of LGREG variance estimators for **math** in fixed SRS samples from post-secondary population excluding all outliers. Outliers are 1.5 times the interquartile range beyond the first and third quartiles. Small sample sizes on top. Large samples on bottom. The empirical variance was calculated **without** the outliers.

B.7.3 Census Population

B.7.3.1 Percent Simulation Coefficient of Variation Table

Table B.33: Percent Simulation Coefficient of Variation for Census Population

	Small Samples					
	Fixed SRS		Rate SRS		Fixed PPS	
Estimator	Renter	Owner	Renter	Owner	Renter	Owner
$\hat{\mathbf{t}}_y^{\pi}$	0.926	0.688	0.896	0.698	0.356	0.284
$\hat{\mathbf{t}}_{yc}^{GD}$	0.496	0.339	0.432	0.341	0.339	0.250
$\hat{\mathbf{t}}_y^{LG}$	0.314	0.219	0.301	0.210	0.229	0.160
$\hat{\mathbf{t}}_y^{mc}$	0.310	0.216	0.306	0.213	0.229	0.160
$\hat{\mathbf{t}}_y^{peN}$	0.317	0.221	0.316	0.220	0.234	0.163
$\hat{\mathbf{t}}_y^{pe\hat{N}}$	0.796	0.638	0.816	0.649	0.234	0.163
	Large Samples					
$\hat{\mathbf{t}}_y^{\pi}$	0.262	0.193	0.262	0.203	0.091	0.076
$\hat{\mathbf{t}}_{yc}^{GD}$	0.146	0.098	0.127	0.091	0.085	0.069
$\hat{\mathbf{t}}_y^{LG}$	0.099	0.069	0.092	0.064	0.060	0.042
$\hat{\mathbf{t}}_y^{mc}$	0.096	0.067	0.089	0.062	0.060	0.042
$\hat{\mathbf{t}}_y^{peN}$	0.097	0.068	0.090	0.063	0.060	0.042
$\hat{\mathbf{t}}_y^{pe\hat{N}}$	0.228	0.177	0.236	0.189	0.060	0.042

B.7.3.2 Average Distance from True Value

Table B.34: Average Distance from True Value for Census Population (in thousands)

	Small Samples			Large Samples		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
$\hat{\mathbf{t}}_y^{\pi}$	7,602	7,551	3,104	2,179	2,233	832
$\hat{\mathbf{t}}_{yc}^{gd}$	3,838	3,576	2,821	1,185	1,060	760
$\hat{\mathbf{t}}_y^{lg}$	2,387	2,223	1,704	747	691	450
$\hat{\mathbf{t}}_y^{mc}$	2,376	2,265	1,720	722	670	449
$\hat{\mathbf{t}}_y^{peM}$	2,442	2,336	1,753	729	676	451
$\hat{\mathbf{t}}_y^{pe\hat{M}}$	6,853	6,931	1,753	1,929	2,020	451

B.7.3.3 Empirical Standard Deviation of Distance from True Value

Table B.35: Empirical Standard Deviation of Distance from True Value for Census Population (in thousands)

	Small Samples			Large Samples		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
$\hat{\mathbf{t}}_y^{\pi}$	4,739.8	4,697.4	1,789.5	1,277.4	1,311.2	442.6
$\hat{\mathbf{t}}_{yc}^{gd}$	2,528.1	2,428.8	1,651.3	629.7	567.0	402.6
$\hat{\mathbf{t}}_y^{lg}$	1,719.3	1,728.0	1,299.7	554.4	514.8	337.9
$\hat{\mathbf{t}}_y^{mc}$	1,655.7	1,746.5	1,282.5	535.8	501.0	340.0
$\hat{\mathbf{t}}_y^{peM}$	1,682.3	1,808.0	1,308.3	542.5	505.0	342.0
$\hat{\mathbf{t}}_y^{pe\hat{M}}$	4,184.0	4,373.6	1,308.3	1,182.7	1,254.5	342.0

B.7.3.4 Percent Relative Bias Table

Table B.36: Percent Relative Bias for Census Population

	Small Samples					
	Fixed SRS		Rate SRS		Fixed PPS	
Estimator	Renter	Owner	Renter	Owner	Renter	Owner
$\hat{\mathbf{t}}_y^{\pi}$	0.3	0.1	0.6	0.6	0.6	-0.1
$\hat{\mathbf{t}}_{yc}^{gd}$	-2.0	2.6	-2.6	2.1	1.9	-0.3
$\hat{\mathbf{t}}_y^{lg}$	-4.9	3.4	-4.2	2.9	0.2	-0.2
$\hat{\mathbf{t}}_y^{mc}$	-5.5	3.8	-3.8	2.6	0.7	-0.5
$\hat{\mathbf{t}}_y^{peM}$	-7.5	5.2	-4.8	3.1	0.2	-0.2
$\hat{\mathbf{t}}_y^{peM\hat{}}$	-3.0	2.3	0.5	1.3	0.2	-0.2
	Large Samples					
$\hat{\mathbf{t}}_y^{\pi}$	0.5	0.6	0.2	0.1	0.1	0.0
$\hat{\mathbf{t}}_{yc}^{gd}$	-0.2	0.1	0.0	0.2	0.2	0.0
$\hat{\mathbf{t}}_y^{lg}$	-0.4	0.2	-0.3	0.2	0.0	0.0
$\hat{\mathbf{t}}_y^{mc}$	-0.4	0.3	-0.3	0.2	0.1	0.0
$\hat{\mathbf{t}}_y^{peM}$	-0.6	0.4	-0.4	0.3	0.0	0.0
$\hat{\mathbf{t}}_y^{peM\hat{}}$	0.3	1.0	0.0	0.2	0.0	0.0

B.7.3.5 Percent Relative Median Difference Table

Table B.37: Percent Relative Median Difference for Census Population

	Small Samples					
	Fixed SRS		Rate SRS		Fixed PPS	
Estimator	Renter	Owner	Renter	Owner	Renter	Owner
$\hat{\mathbf{t}}_y^{\pi}$	-11.8	-6.5	-9.3	-4.9	-2.8	-1.8
$\hat{\mathbf{t}}_{yc}^{gd}$	-8.2	-0.7	-5.2	-1.5	-1.3	-2.2
$\hat{\mathbf{t}}_y^{lg}$	-7.0	4.9	-5.2	3.6	-1.0	0.7
$\hat{\mathbf{t}}_y^{mc}$	-6.9	4.8	-4.3	3.0	-0.1	0.1
$\hat{\mathbf{t}}_y^{peM}$	-9.4	6.6	-5.2	3.4	-0.6	0.4
$\hat{\mathbf{t}}_y^{peM\hat{}}$	-10.7	-1.7	-6.7	-2.6	-0.6	0.4
	Large Samples					
$\hat{\mathbf{t}}_y^{\pi}$	-0.2	-0.1	-0.8	-0.4	-0.1	-0.2
$\hat{\mathbf{t}}_{yc}^{gd}$	-0.7	-0.2	-0.8	-0.1	0.0	-0.1
$\hat{\mathbf{t}}_y^{lg}$	-0.6	0.4	-0.5	0.3	0.0	0.0
$\hat{\mathbf{t}}_y^{mc}$	-0.6	0.4	-0.4	0.2	0.1	-0.1
$\hat{\mathbf{t}}_y^{peM}$	-0.8	0.6	-0.5	0.4	0.0	0.0
$\hat{\mathbf{t}}_y^{peM\hat{}}$	-0.3	0.6	-0.7	-0.2	0.0	0.0

B.7.3.6 Percent Relative Root Mean Squared Error Table

Table B.38: Percent Relative Root Mean Squared Error for Census Population

	Small Samples					
	Fixed SRS		Rate SRS		Fixed PPS	
Estimator	Renter	Owner	Renter	Owner	Renter	Owner
$\hat{\mathbf{t}}_y^{\pi}$	65.5	48.6	63.4	49.3	25.2	20.1
$\hat{\mathbf{t}}_{yc}^{gd}$	35.1	23.9	30.5	24.1	24.0	17.7
$\hat{\mathbf{t}}_y^{lg}$	22.2	15.5	21.3	14.8	16.2	11.3
$\hat{\mathbf{t}}_y^{mc}$	21.9	15.2	21.6	15.1	16.2	11.3
$\hat{\mathbf{t}}_y^{peM}$	22.4	15.6	22.3	15.5	16.5	11.5
$\hat{\mathbf{t}}_y^{peM\hat{}}$	56.3	45.1	57.7	45.9	16.5	11.5
	Large Samples					
$\hat{\mathbf{t}}_y^{\pi}$	18.5	13.7	18.5	14.3	6.5	5.4
$\hat{\mathbf{t}}_{yc}^{gd}$	10.3	7.0	9.0	6.4	6.0	4.8
$\hat{\mathbf{t}}_y^{lg}$	7.0	4.9	6.5	4.5	4.2	3.0
$\hat{\mathbf{t}}_y^{mc}$	6.8	4.7	6.3	4.4	4.2	3.0
$\hat{\mathbf{t}}_y^{peM}$	6.9	4.8	6.4	4.4	4.3	3.0
$\hat{\mathbf{t}}_y^{peM\hat{}}$	16.1	12.5	16.7	13.4	4.3	3.0

B.7.3.7 Percent Relative Root Median Squared Error Table

Table B.39: Percent Relative Root Median Squared Error for Census Population

	Small Samples					
	Fixed SRS		Rate SRS		Fixed PPS	
Estimator	Renter	Owner	Renter	Owner	Renter	Owner
$\hat{\mathbf{t}}_y^{\pi}$	45.1	32.5	43.3	33.9	15.9	13.2
$\hat{\mathbf{t}}_{yc}^{gd}$	21.1	14.5	17.7	14.0	14.7	11.6
$\hat{\mathbf{t}}_y^{lg}$	15.8	11.0	14.0	9.8	10.9	7.6
$\hat{\mathbf{t}}_y^{mc}$	16.2	11.3	14.5	10.1	11.1	7.7
$\hat{\mathbf{t}}_y^{peM}$	16.6	11.6	14.9	10.3	11.4	7.9
$\hat{\mathbf{t}}_y^{peM\hat{}}$	40.3	30.9	40.9	31.4	11.4	7.9
	Large Samples					
$\hat{\mathbf{t}}_y^{\pi}$	12.6	9.1	12.5	9.7	4.3	3.6
$\hat{\mathbf{t}}_{yc}^{gd}$	6.9	4.6	5.9	4.3	4.0	3.2
$\hat{\mathbf{t}}_y^{lg}$	4.8	3.3	4.5	3.1	2.9	2.0
$\hat{\mathbf{t}}_y^{mc}$	4.5	3.2	4.4	3.0	2.9	2.0
$\hat{\mathbf{t}}_y^{peM}$	4.6	3.2	4.4	3.1	2.9	2.0
$\hat{\mathbf{t}}_y^{peM\hat{}}$	10.9	8.4	11.6	8.9	2.9	2.0

B.7.3.8 Percent Relative Bias Table for Variance Estimators

Table B.40: Percent Relative Bias of Variance Estimators for Census Population

	Small Samples					
	Fixed SRS		Rate SRS		Fixed PPS	
Estimator	Renter	Owner	Renter	Owner	Renter	Owner
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-42.2	-42.2	-44.1	-44.1	-16.3	-16.3
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-54.1	-54.1	-55.9	-55.9	-34.8	-34.8
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-60.3	-60.3	-59.0	-59.0	-20.9	-20.9
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-53.5	-53.5	-56.5	-56.5	-33.0	-33.0
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-63.0	-63.0	-65.7	-65.7	-47.7	-47.7
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-67.2	-67.2	-68.0	-68.0	-44.5	-44.5
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-55.0	-55.0	-57.8	-57.8	-35.1	-35.1
	Large Samples					
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	11.7	11.7	11.4	11.4	51.0	51.0
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-5.7	-5.7	-11.1	-11.1	2.1	2.1
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	7.5	7.5	8.8	8.8	50.0	50.0
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	4.5	4.5	5.2	5.2	42.2	42.2
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-11.9	-11.9	-15.8	-15.8	-2.6	-2.6
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-14.1	-14.1	-17.8	-17.8	-2.0	-2.0
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	3.9	3.9	4.7	4.7	41.7	41.7

B.7.3.9 Percent Relative Median Difference Table for Variance Estimators

Table B.41: Percent Relative Median Difference of Variance Estimators for Census Population

	Small Samples					
	Fixed SRS		Rate SRS		Fixed PPS	
Estimator	Renter	Owner	Renter	Owner	Renter	Owner
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-64.6	-64.6	-57.9	-57.9	22.4	22.4
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-70.9	-70.9	-66.5	-66.5	-3.7	-3.7
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	-50.6	-50.6	-40.0	-40.0	31.0	31.0
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-72.0	-72.0	-66.9	-66.9	-0.2	-0.2
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-77.1	-77.1	-73.5	-73.5	-21.7	-21.7
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-62.3	-62.3	-57.5	-57.5	-7.8	-7.8
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	-74.2	-74.2	-69.3	-69.3	-4.3	-4.3
	Large Samples					
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	90.2	90.2	77.0	77.0	217.0	217.0
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	63.6	63.6	43.3	43.3	116.2	116.2
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	90.9	90.9	80.7	80.7	217.3	217.3
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	96.0	96.0	78.2	78.2	202.8	202.8
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	68.1	68.1	44.3	44.3	107.1	107.1
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	74.3	74.3	53.0	53.0	110.7	110.7
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	93.7	93.7	76.7	76.7	201.7	201.7

B.7.3.10 Percent Relative Root Mean Squared Error Table for Variance Estimators

Table B.42: Percent Relative Root Mean Squared Error of Variance Estimators for Census Population

	Small Samples					
	Fixed SRS		Rate SRS		Fixed PPS	
Estimator	Renter	Owner	Renter	Owner	Renter	Owner
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	148.2	148.2	140.9	140.9	96.6	96.6
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	124.1	124.1	118.8	118.8	80.8	80.8
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	78.4	78.4	77.8	77.8	74.5	74.5
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	119.3	119.3	112.4	112.4	76.3	76.3
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	104.9	104.9	100.5	100.5	71.2	71.2
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	79.6	79.6	88.6	88.6	63.4	63.4
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	120.2	120.2	112.2	112.2	76.3	76.3
	Large Samples					
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	80.1	80.1	83.8	83.8	68.8	68.8
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	63.6	63.6	64.6	64.6	26.6	26.6
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	67.4	67.4	74.2	74.2	65.4	65.4
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	63.5	63.5	66.2	66.2	58.3	58.3
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	52.0	52.0	53.2	53.2	24.7	24.7
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	41.2	41.2	42.4	42.4	22.0	22.0
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	63.4	63.4	66.1	66.1	57.8	57.8

B.7.3.11 Percent Relative Root Median Squared Error Table for Variance Estimators

Table B.43: Percent Relative Root Median Squared Error of Variance Estimators for Census Population

	Small Samples					
	Fixed SRS		Rate SRS		Fixed PPS	
Estimator	Renter	Owner	Renter	Owner	Renter	Owner
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	87.2	87.2	85.0	85.0	61.2	61.2
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	86.8	86.8	83.9	83.9	54.5	54.5
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	70.2	70.2	67.5	67.5	62.2	62.2
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	86.8	86.8	84.9	84.9	56.9	56.9
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	87.0	87.0	85.0	85.0	53.5	53.5
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	72.0	72.0	69.3	69.3	51.6	51.6
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	87.9	87.9	85.6	85.6	56.0	56.0
	Large Samples					
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	90.2	90.2	77.0	77.0	217.0	217.0
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	64.7	64.7	47.6	47.6	116.2	116.2
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	90.9	90.9	80.7	80.7	217.3	217.3
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	96.0	96.0	78.2	78.2	202.8	202.8
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	68.4	68.4	47.7	47.7	107.1	107.1
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	74.3	74.3	53.1	53.1	110.7	110.7
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	93.7	93.7	76.7	76.7	201.7	201.7

B.7.3.12 Average Distance from Empirical Value for Variance Estimators

Table B.44: Average Distance from Empirical Value for Standard Error Estimators in Census Population (in thousands)

	Small Samples			Large Samples		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	1,742.0	1,633.5	792.8	238.7	228.1	130.1
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	1,769.2	1,675.8	836.9	228.9	230.0	57.6
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	1,463.0	1,376.9	712.9	205.5	200.1	126.8
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	1,711.1	1,684.3	792.5	204.2	194.4	112.2
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	1,753.3	1,743.3	864.3	198.8	198.9	56.1
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	1,514.5	1,528.2	775.0	163.0	162.9	49.4
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	1,739.2	1,704.5	802.6	204.5	194.8	111.2

B.7.3.13 Median Distance from Empirical Value for Variance Estimators

Table B.45: Median Distance from Empirical Value for Standard Error Estimators in Census Population (in thousands)

	Small Samples			Large Samples		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	1,124.7	967.4	454.0	240.1	196.0	296.8
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	1,132.2	968.6	417.9	183.0	131.6	178.9
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	801.4	687.2	459.8	239.4	201.2	297.1
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	1,147.1	1,044.7	438.1	240.1	194.0	280.7
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	1,157.8	1,051.9	428.6	181.6	130.2	166.6
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	868.9	774.7	395.1	192.4	137.5	171.2
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	1,173.9	1,055.3	437.2	235.5	191.1	279.5

B.7.3.14 Standard Error of Average Distance from Empirical Value for Variance Estimators

Table B.46: Standard Error of Average Distance from Empirical Value for Standard Error Estimators in Census Population (in thousands)

	Small Samples			Large Samples		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	952.6	898.8	542.6	193.8	185.9	92.1
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	872.9	818.5	478.4	157.0	139.1	43.5
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	697.2	678.4	468.7	168.0	168.1	85.9
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	842.4	827.5	483.4	153.6	147.3	80.7
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	788.4	774.0	460.1	131.8	118.8	41.0
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	657.1	682.1	439.8	107.8	99.3	37.3
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	845.1	827.2	485.1	153.4	146.8	80.1

B.7.3.15 95% Confidence Interval Coverage Table for Variance Estimators

Table B.47: Percent 95% Confidence Interval Coverage of LGREG Variance Estimators for Census Population

	Small Samples					
	Fixed SRS		Rate SRS		Fixed PPS	
Estimator	Renter	Owner	Renter	Owner	Renter	Owner
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	65.1	65.1	68.0	68.0	91.3	91.3
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	63.0	63.0	65.5	65.5	88.9	88.9
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	71.9	71.9	75.1	75.1	92.1	92.1
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	62.3	62.3	65.0	65.0	88.0	88.0
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	59.7	59.7	62.2	62.2	84.8	84.8
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	68.1	68.1	71.1	71.1	87.9	87.9
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	61.1	61.1	64.0	64.0	87.7	87.7
	Large Samples					
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	91.3	91.3	92.0	92.0	97.9	97.9
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	90.0	90.0	89.7	89.7	94.7	94.7
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{lg}\right)$	93.1	93.1	93.5	93.5	97.8	97.8
$\upsilon_{wr}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	90.7	90.7	91.7	91.7	97.1	97.1
$\upsilon_{e}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	89.0	89.0	89.3	89.3	94.0	94.0
$\upsilon_{g}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	90.9	90.9	90.5	90.5	94.3	94.3
$\upsilon_{Binder}\left(\widehat{\mathbf{t}}_y^{mc}\right)$	90.6	90.6	91.7	91.7	97.1	97.1

B.7.3.16 Plots

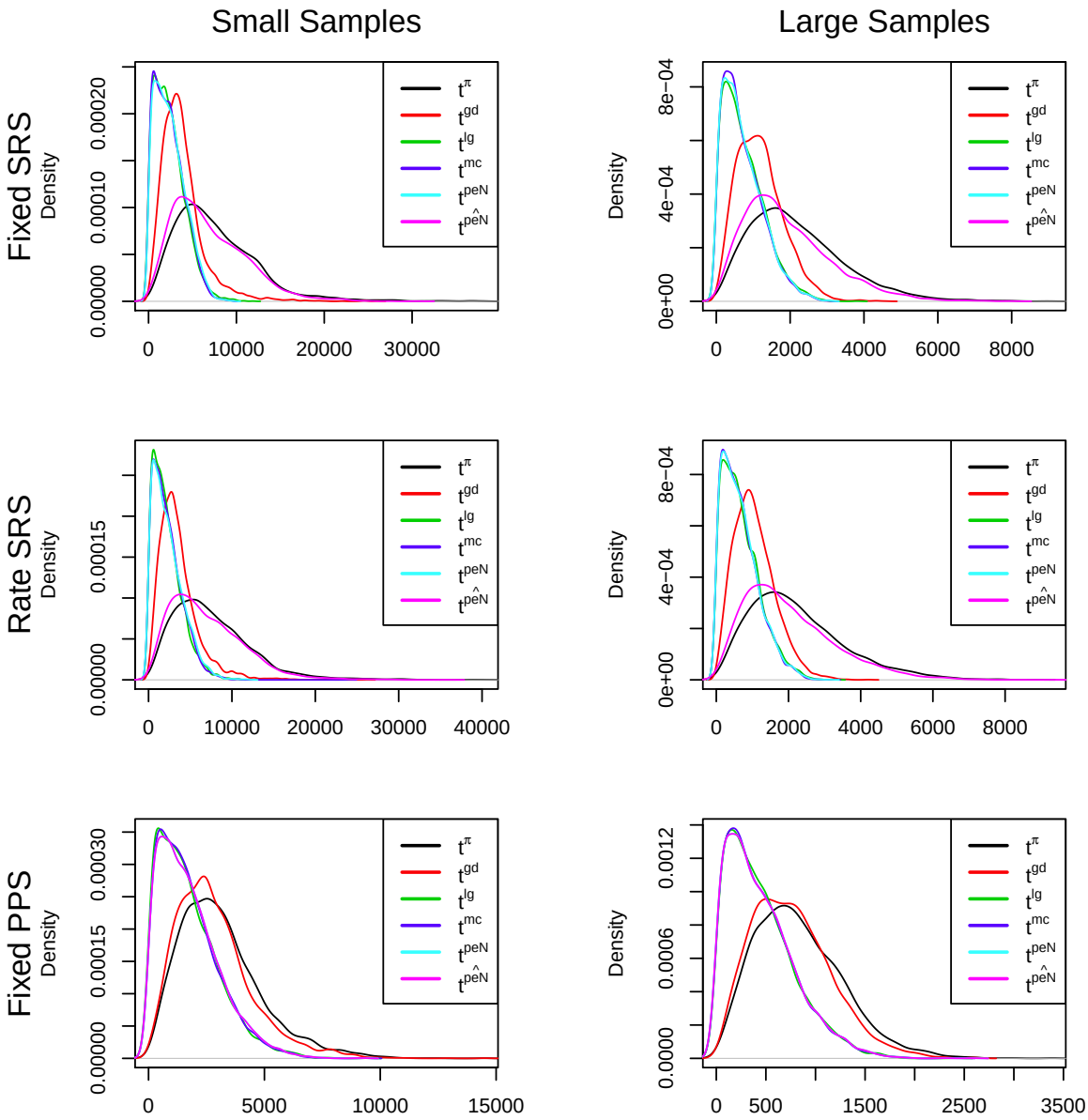

Figure B.11: Density Plot of Distance Between Estimator and True Value in Census Population

B.8 R Code

```
# I had to alter the UPsystematic function so that it would work.
# I changed trunc(n) to round(n)
UPsystematic.round <- function (pik, eps = 1e-06)
{
    if (any(is.na(pik)))
        stop("there are missing values in the pik vector")
    n = sum(pik)
    if (abs(n - round(n)) < 1e-03)
        n = round(n)
    else stop("the sum of pik is not integer")
    list = pik > eps & pik < 1 - eps
    pik1 = pik[list]
    N = length(pik1)
    a = (c(0, cumsum(pik1)) - runif(1, 0, 1))%%1
    s1 = as.integer(a[1:N] > a[2:(N + 1)])
    s = pik
    s[list] = s1
    s
}

UPrandomsystematic.alt <- function (pik, eps = 1e-06)
{
    if (any(is.na(pik)))
        stop("there are missing values in the pik vector")
    N = length(pik)
    v = sample(N, N)
    s = numeric(N)
    s[v] = UPsystematic.round(pik[v], eps)
    s
}

UPrandomsystematic.alt2 <- function (x, eps = 1e-06)
{
  X.I.ii <- UPrandomsystematic.alt(x$pi.II.all)
  subset(x, X.I.ii == 1)
}

UPoi <- function (x)
{
  X.I.ii <- UPpoisson(x$pi.II.all)
  sa.mp <- subset(x, X.I.ii == 1)
  if(nrow(sa.mp) > 0) return(subset(x, X.I.ii == 1))
}

Lag1=function(u,ds,mu){
  dif=1
  tol=1e-8
  if(min(u-mu)>=0 | max(u-mu)<=0){
    dif=0
    M=0
  }
  L=-1/max(u-mu)
  R=-1/min(u-mu)
  while(dif>tol){
    M=(L+R)/2
    glam=sum((ds*(u-mu))/(1+M*(u-mu)))
    if(glam>0) L=M
    if(glam<0) R=M
    dif=abs(glam)
  }
  return(M)
}



Lag2=function(u,ds,mu){
  n=length(ds)
  u=u-rep(1,n)%*%t(mu)
  M=0*mu
  dif=1
  tol=1e-8
  k=0
  while(dif>tol & k<=50){
  D1=t(u)%*%((ds/(1+u%*%M))*rep(1,n))
  DD=-t(u)%*%(c((ds/(1+u%*%M)^2))*u)
  D2=solve(DD,D1,tol=1e-40)
  dif=max(abs(D2))
  rule=1
  while(rule>0){
    rule=0
    if(min(1+t(M-D2)%*%t(u))<=0) rule=rule+1
    if(rule>0) D2=D2/2
               }
    M=M-D2
    k=k+1
                    }
  if(k>=50) M=0*mu
  return(as.vector(M))
}
```

```
LGREG.sim <- function(X.Pop, Y.Pop, clus.id, a, b, iterations, seed, smp, smp2)
{
  cat("Begin Intro", format(Sys.time(), "%X"), "\n")

  load(file = "C:\\Documents and Settings\\Tim\\My Documents\\Data\\seed.Rdata")
  set.seed(seed)

  Pop.1 <- cbind(X.Pop, Y.Pop, clus.id)
  z.Pop <- rowSums(Y.Pop)

  # Get the population size
  M.1 <- nrow(Pop.1)

  # Get the number of columns in X and Y
  X.dim <- ncol(X.Pop)
  Y.dim <- ncol(Y.Pop)

  # Create the measures of size
  mos.1 <- as.vector(by(Pop.1, Pop.1[,"clus.id"], nrow))

  # M.clus is the total number of clusters in the population
  M.clus <- length(unique(Pop.1[, "clus.id"]))

  # Create the first stage sampling probabilities
  pi.I.pps <- a * mos.1  / nrow(Pop.1)
  pi.I.srs <- rep(a / M.clus, M.clus)
  if(smp == "srs") pi.I <- pi.I.srs else pi.I <- pi.I.pps

  pi.II.fixed <- b / mos.1
  pi.II.rate <- (b * sum(M.clus)) / sum(mos.1)
  if(smp2 == "fixed") pi.II.all <- pi.II.fixed else pi.II.all <- pi.II.rate

  pi.k.all <- pi.I * pi.II.all


  # Recode the clusterid
  c.id <- c(1: M.clus)
  clus.conversion <- cbind(unique(Pop.1[, "clus.id"]), c.id, pi.I, pi.II.all, pi.k.all)
  X.clusid <- merge(x = Pop.1, y = clus.conversion, by.x = "clus.id", by.y = 1)

  w.n <- 1 / X.clusid[, "pi.k.all"]
  w.n.II <- 1 / X.clusid[, "pi.II.all"]
  ind <- X.clusid[, "clus.id"]

  # Create a list of cluster auxilliaries
  X.clus <- split(X.clusid, clus.id)

  t.HT <- matrix(0, nrow = iterations, ncol = (Y.dim))
  t.PROJ.glm <- matrix(0, nrow = iterations, ncol = (Y.dim))
  t.PROJ.wglm <- matrix(0, nrow = iterations, ncol = (Y.dim))
  t.PROJ.pml <- matrix(0, nrow = iterations, ncol = (Y.dim))
  t.GREG <- matrix(0, nrow = iterations, ncol = (Y.dim))
  v.LGREG.wr <- matrix(0, nrow = iterations, ncol = (Y.dim))
  v.LGREG.ssw <- matrix(0, nrow = iterations, ncol = (Y.dim))
  t.LGREG.pml <- matrix(0, nrow = iterations, ncol = (Y.dim))
  t.MCAL.pml <- matrix(0, nrow = iterations, ncol = (Y.dim))
  t.MCAL.solve.pml <- matrix(0, nrow = iterations, ncol = (Y.dim))
  t.MCAL.solnp.log <- matrix(0, nrow = iterations, ncol = (Y.dim))
  v.MCAL.wr <- matrix(0, nrow = iterations, ncol = (Y.dim))
  v.MCAL.ssw <- matrix(0, nrow = iterations, ncol = (Y.dim))
  v.MCAL.ssw.g <- matrix(0, nrow = iterations, ncol = (Y.dim))
  t.PEMLE.pml <- matrix(0, nrow = iterations, ncol = (Y.dim))
  t.PEMLE.pml.N <- matrix(0, nrow = iterations, ncol = (Y.dim))
  t.PEMLE.pml.w <- matrix(0, nrow = iterations, ncol = (Y.dim))
  t.PEMLE.glm <- matrix(0, nrow = iterations, ncol = (Y.dim))

  t.PEML.v <- vector(length = Y.dim)
  v.LGREG.pml <- matrix(0, nrow = iterations, ncol = (Y.dim))
  v.LGREG.pml.10 <- matrix(0, nrow = iterations, ncol = (Y.dim))
  v.MCAL.pml.old <- matrix(0, nrow = iterations, ncol = (Y.dim))
  v.MCAL.pml <- matrix(0, nrow = iterations, ncol = (Y.dim))

  cat("End Intro", format(Sys.time(), "%X"), "\n")

  j <- 1
  j.master <- 0
  error.1 <- NULL
  error.2 <- NULL
  error.3 <- NULL
  error.4 <- matrix(0, nrow = iterations, ncol = (Y.dim))
  error.5 <- matrix(0, nrow = iterations, ncol = (Y.dim))
  error.6 <- matrix(0, nrow = iterations, ncol = (Y.dim))
  error.7 <- matrix(0, nrow = iterations, ncol = (Y.dim))
  error.8 <- NULL
  error.9 <- NULL
  error.10 <- NULL

#  for(j in 1: iterations)
  while(j < iterations +1)
  {
```

```
    j.master <- j.master +1

    ## Sampling begins here
    # Select the first stage sample without replacement
    samp.clus <- UPrandomsystematic.alt(clus.conversion[,"pi.I"])
    X.clus.sample <- X.clus[c.id[samp.clus >= 1]]

    # Select the second stage sample
    if(smp2 == "rate") X.sample.f <- lapply(X.clus.sample, UPoi) else X.sample.f <- lapply(X.clus.sample, UPrandomsystematic.alt2)

    # Vector of sample clusters including zero clusters
    if(smp2 == "rate") a.f <- sapply(X.sample.f, length) else a.f <- sapply(X.sample.f, nrow)

    # Number of sample clusters including zeros
    n.clus.samp.z <- length(a.f)

    # Vector of sample clusters excluding zero clusters
    a.g <- subset(a.f, a.f > 0)

    # Number of sample clusters excluding zeros
    n.clus.samp <- length(a.g)

    # Vector of nonzero clusters
    a.1 <- ifelse(a.f > 0, 1, 0)
    a.n <- c(1:n.clus.samp.z)
    a.n1 <- a.1 * a.n
    a.i <- subset(a.n1, a.n1 > 0)

    # Create Unclustered data
    # Note that the sample elements can be repeated
    # Note: There may be some duplicates
if(smp2 == "fixed") {
  Fixed.id <- lapply(a.i,
           function(i, X.sample.f)
                   c(as.numeric(rownames(X.sample.f[[i]]))),
                   X.sample.f= X.sample.f)
  sample.id <- c(sapply(X = Fixed.id, FUN = rbind, simplify = T, USE.NAMES = T))
}
else {
  sample.id <- as.numeric(unique(as.vector(do.call(c, (sapply(X = X.sample.f, FUN = rownames, simplify = F, USE.NAMES = T))))))
}
    # Matrix of sample units in sample clusters
    X.sample <- X.clusid[sample.id, ]

    # List of nonzero sample cluster names
    b.f <- as.numeric(names(a.g))


    # Cluster probabilities of selection for nonzero sample clusters
    samp.clus.pi <- pi.I[b.f]

    # Cluster probabilities of selection for nonzero sample clusters repeated for each category
    samp.clus.pi.cat <- samp.clus.pi %x% matrix(rep(1, ncol(Y.Pop)), ncol = ncol(Y.Pop))

    ## Estimation begins here
    # Population Totals
    T.x <- colSums(X.Pop)

    # Sample X and Y values
    # Note: There may be some duplicates when the first stage is selected with replacement
    X.samp <- X.Pop[as.numeric(sample.id),]
    Y.samp <- Y.Pop[as.numeric(sample.id),]
    z.samp <- rowSums(Y.samp)
    w.k <- w.n[as.numeric(sample.id)]
    w.k.2 <- w.n.II[as.numeric(sample.id)]

     ind.1 <- factor(ind[as.numeric(sample.id)])

    samp.pi.I <- subset(pi.I, samp.clus == 1)
    samp.pi.I.list <- split(samp.pi.I, f = seq(1:length(samp.pi.I)))

    # Cluster level weight for nonzero clusters
    w.k.clus <- split(w.k, ind.1)

    # Number of units in sample
    n.samp <- length(w.k)


    # Skip if there are data problems
    error.1[j.master] <- ifelse(any(colSums(Y.samp) ==0), 1, 0)
    error.2[j.master] <- ifelse(qr(X.samp)$rank < ncol(X.samp), 1, 0)
    if(any(colSums(Y.samp) ==0)) next
    if(qr(X.samp)$rank < ncol(X.samp)) next

    # Pi Estimator
    t.y.pi <- colSums((w.k) * Y.samp)
    t.HT[j, ] <- t.y.pi

    ## Estimate beta
    # LM
    lm.1 <- lm(Y.samp ~ X.samp - 1, weights = w.k)
    beta.lm <- matrix(coefficients(lm.1), nrow = X.dim)
```

```
# GLM
logit.glm <- try(vglm(Y.samp ~ X.samp -1, multinomial))
beta.glm <- matrix(coefficients(logit.glm), nrow = X.dim, ncol = (Y.dim - 1), byrow = TRUE)
b.dim <- length(beta.glm)

# Pseudo Maximum Likelihood
L.beta <- function(par){
  var.id <- 1
  mu.k <-  X.samp %*% matrix(c(par[1:b.dim]), nrow = ncol(X.samp), ncol = (Y.dim -1), byrow = FALSE)
  -sum(w.k * rowSums(Y.samp[, 1: (Y.dim - 1)] * (mu.k)) - z.samp * w.k* log(1 + rowSums(exp( mu.k)) ))
}
min.L <- optim(par = c(beta.glm), fn = L.beta)
beta.pml <- matrix(min.L$par, nrow = ncol(X.samp), ncol = (Y.dim-1), byrow = FALSE)

##### Sample prediction
# LM
Samp.fit.lm <- fitted.values(lm.1)

# GLM
Samp.fit.glm <- z.samp * fitted.values(logit.glm)

# Pseudo Maximum Likelihood
Samp.fit.pml.1 <- z.samp * exp(X.samp %*% beta.pml) / (1 + rowSums(exp(X.samp %*% beta.pml)))
Samp.fit.pml.2 <- z.samp / (1 + rowSums(exp(X.samp %*% beta.pml)))
Samp.fit.pml <- cbind(Samp.fit.pml.1, Samp.fit.pml.2)

### Sample Residuals
clus.resid <- t(sapply(by(w.k * (Y.samp - Samp.fit.pml), INDICES = ind.1, colSums, simplify = T), FUN = identity))
mean.resid <- matrix(rep(colMeans(clus.resid), n.clus.samp), nrow = n.clus.samp, ncol = Y.dim, byrow = TRUE)

### LGREG Variance Estimators
v.LGREG.wr[j, ] <- (a / (a - 1)) * colSums((clus.resid - mean.resid)^2)

ssw.clus <- colSums((1 - samp.clus.pi.cat) * (clus.resid)^2)
ssw.within
    <- t(1 / samp.clus.pi) %*%
        t(sapply(by(w.k.2^2 * (1 - 1/w.k.2) * (Y.samp - Samp.fit.pml)^2, INDICES = ind.1, colSums, simplify = T), FUN = identity))
v.LGREG.ssw[j, ] <- ssw.clus + ssw.within

##### Population prediction
# LM
Pop.fit.lm <- X.Pop %*% beta.lm

# GLM
Pop.fit.glm.1 <- z.Pop * exp(X.Pop %*% beta.glm) / (1 + rowSums(exp(X.Pop %*% beta.glm)))
Pop.fit.glm.2 <- z.Pop / (1 + rowSums(exp(X.Pop %*% beta.glm)))
Pop.fit.glm <- cbind(Pop.fit.glm.1, Pop.fit.glm.2)

## Stop if any estimates are infinity or bad
error.3[j.master] <- ifelse(any(is.na(colSums(Pop.fit.glm))) ==TRUE, 1, 0)
if(any(is.na(colSums(Pop.fit.glm))) ==TRUE) next

# Pseudo Maximum Likelihood
Pop.fit.pml.1 <- z.Pop * exp(X.Pop %*% beta.pml) / (1 + rowSums(exp(X.Pop %*% beta.pml)))
Pop.fit.pml.2 <- z.Pop / (1 + rowSums(exp(X.Pop %*% beta.pml)))
Pop.fit.pml <- cbind(Pop.fit.pml.1, Pop.fit.pml.2)

## PML Weighted Projective Estimator
t.PROJ.pml[j, ] <- colSums(Pop.fit.pml)
t.PROJ.glm[j, ] <- colSums(Pop.fit.glm)

## GREG
t.GREG[j, ] <- colSums(Pop.fit.lm, na.rm = TRUE) -  colSums(w.k * (Y.samp - Samp.fit.lm), na.rm = TRUE)

## LGREG
# Using PML
t.LGREG.pml[j, ] <- colSums(Pop.fit.pml, na.rm = TRUE) -  colSums(w.k * (Y.samp - Samp.fit.pml), na.rm = TRUE)

## Calibration
# Same as GREG

## Model Calibration
# Using PML
# Just use mu with intercept
samp.mu <- cbind(1, Samp.fit.pml)
pop.mu <- cbind(1, Pop.fit.pml)

A.mu <- t(samp.mu * w.k) %*% samp.mu
error.10[j.master]
    <- ifelse( (any(abs(eigen(A.mu, only.values = TRUE)$values) <= .Machine$double.eps) || qr(A.mu)$rank < ncol(A.mu)), 1, 0)
if(error.10[j.master] == 1) next

t.MC.pml <- t.y.pi + (colSums(pop.mu) - colSums(samp.mu * w.k)) %*% ( solve(A.mu) %*% t(samp.mu * w.k) %*% Y.samp)
t.MCAL.pml[j, ] <- t.MC.pml

### Model Calibration Residuals
clus.resid.mc
    <- t(sapply(by(w.k * (Y.samp - samp.mu %*% (
        solve(A.mu) %*% t(samp.mu * w.k) %*% Y.samp)), INDICES = ind.1, colSums, simplify = T), FUN = identity))
mean.resid.mc <- matrix(rep(colMeans(clus.resid.mc), n.clus.samp), nrow = n.clus.samp, ncol = Y.dim, byrow = TRUE)
```

```
    ### Model Calibration Variance Estimator
    v.MCAL.wr[j, ] <- (a / (a - 1)) * colSums((clus.resid.mc - mean.resid.mc)^2)

    ssw.clus.MCAL <- colSums((1 - samp.clus.pi.cat) * (clus.resid.mc)^2)
    ssw.within.MCAL
        <- t(1 / samp.clus.pi) %*%
            t(sapply(by(w.k.2^2 * (1 - 1/w.k.2) * (Y.samp - samp.mu %*%
            ( solve(A.mu) %*% t(samp.mu * w.k)%*% Y.samp))^2, INDICES = ind.1, colSums, simplify = T), FUN = identity))
    v.MCAL.ssw[j, ] <- ssw.clus.MCAL + ssw.within.MCAL

    g.k <- c(1 + (colSums(pop.mu) - colSums(samp.mu * w.k)) %*% ( solve(A.mu) %*% t(samp.mu)))
    clus.gresid.mc
        <- t(sapply(by((w.k * g.k) * (Y.samp - samp.mu %*%
        ( solve(A.mu) %*% t(samp.mu * w.k) %*% Y.samp)), INDICES = ind.1, colSums, simplify = T), FUN = identity))
    ssw.gclus.MCAL <- colSums((1 - samp.clus.pi.cat) * (clus.gresid.mc)^2)
    ssw.gwithin.MCAL
        <- t(1 / samp.clus.pi) %*% t(sapply(by(w.k.2^2 * (1 - 1/w.k.2) * g.k^2 * (Y.samp - samp.mu %*%
        ( solve(A.mu) %*% t(samp.mu * w.k)%*% Y.samp))^2, INDICES = ind.1, colSums, simplify = T), FUN = identity))
    v.MCAL.ssw.g[j, ] <- ssw.gclus.MCAL + ssw.gwithin.MCAL

####### Variance of Estimation Equations (Start)
#### Common Estimates
### Create parameter vector
  theta.pml <- c(t(t.LGREG.pml[j,]), t(t.MC.pml), beta.pml)

### Length of parameters
  LGREG.start <- 1
  LGREG.end <- Y.dim

  MCAL.start <- Y.dim + 1
  MCAL.end <- 2* Y.dim

  beta.start <- 2 * Y.dim + 1
  beta.end <- 2 * Y.dim + b.dim


## Create estimating equation function for estimating theta.pml
# The output of this function is the sum of the estimating equations for all units
W.est <- function(par){
  Samp.fit.pml.1
    <- z.samp * exp(X.samp %*% matrix(c(par[(beta.start): beta.end]), nrow = ncol(X.samp), ncol = (Y.dim -1), byrow = FALSE)) /
    (1 + rowSums(exp(X.samp %*% matrix(c(par[beta.start: beta.end]), nrow = ncol(X.samp), ncol = (Y.dim -1), byrow = FALSE))))
  Samp.fit.pml.2
    <- z.samp /
    (1 + rowSums(exp(X.samp %*% matrix(c(par[(beta.start): (beta.end)]), nrow = ncol(X.samp), ncol = (Y.dim -1), byrow = FALSE))))
  mu.k <- cbind(Samp.fit.pml.1, Samp.fit.pml.2)
  samp.mu <- cbind(1, Samp.fit.pml.1, Samp.fit.pml.2)
  A.mu <- t(samp.mu * w.k) %*% samp.mu

  Pop.fit.pml.1
    <- z.Pop * exp(X.Pop %*% matrix(c(par[(beta.start): (beta.end)]), nrow = ncol(X.samp), ncol = (Y.dim -1), byrow = FALSE)) /
    (1 + rowSums(exp(X.Pop %*% matrix(c(par[(beta.start): (beta.end)]), nrow = ncol(X.samp), ncol = (Y.dim -1), byrow = FALSE))))
  Pop.fit.pml.2
    <- z.Pop /
    (1 + rowSums(exp(X.Pop %*% matrix(c(par[(beta.start): (beta.end)]), nrow = ncol(X.samp), ncol = (Y.dim -1), byrow = FALSE))))
  mu.k.pop <- cbind(Pop.fit.pml.1, Pop.fit.pml.2)
  pop.mu <- cbind(1, Pop.fit.pml.1, Pop.fit.pml.2)

  z.LGREG <- colSums(w.k * (Y.samp - mu.k)) - (par[LGREG.start: (LGREG.end)] - colSums(mu.k.pop))
  z.MCAL
    <- t.y.pi + (colSums(pop.mu) - colSums(samp.mu * w.k)) %*%
    ( solve(A.mu) %*% t(samp.mu * w.k) %*% Y.samp) - par[MCAL.start: (MCAL.end)]
  z.beta <- t(Y.samp[, -Y.dim]) %*% (w.k * X.samp) - t(mu.k[, -Y.dim]) %*% (w.k * X.samp)
  c(z.LGREG, z.MCAL, t(z.beta))
}

# Jacobian
Jacob <- jacobian(W.est, theta.pml)

## Cluster Level Estimating Equations for B
# The output of this function is the sum of the estimating equations for B within each cluster
# Unit Level Estimating Equations for B
X.samp.unit <- split(X.samp, f = c(1:n.samp))
resid.unit <- split(w.k * (Y.samp[, 1: (Y.dim -1)] - Samp.fit.pml[, 1: (Y.dim -1)]), f = c(1:n.samp))
Est.Eq <- lapply(1:n.samp,
          function(i, X.samp.unit, resid.unit)
                  t(t(c(t(t(X.samp.unit[[i]])) %*% resid.unit[[i]]))),
                  X.samp.unit = X.samp.unit, resid.unit = resid.unit)
Est.Eq.Matrix <- t(sapply(X = Est.Eq, FUN = identity, simplify = T, USE.NAMES = T))

# Cluster Level Estimating Equations for B
t.W.clus <- t(sapply(by(Est.Eq.Matrix, ind.1, colSums), FUN = identity, simplify = T, USE.NAMES = T))

#### LGREG
Jacob.LGREG <- Jacob[c(LGREG.start:LGREG.end, beta.start:beta.end), c(LGREG.start:LGREG.end, beta.start:beta.end)]
error.8[j.master]
    <- ifelse( (any(abs(eigen(Jacob.LGREG, only.values = TRUE)$values) <= .Machine$double.eps) ||
        qr(Jacob.LGREG)$rank < ncol(Jacob.LGREG)), 1, 0)
if( error.8[j.master] == 1) next

  # Invert the LGREG jacobian of the estimating equations with theta.pml as the input
  J.inv.LGREG <- solve(Jacob.LGREG, tol = 1e-23)
```

```
  ## Cluster Level Estimating Equations for LGREG
  # The output of this function is the sum of the LGREG estimating equations within each cluster
  z.LGREG.b <- (w.k * (Y.samp - Samp.fit.pml))
  z.LGREG <- t(sapply(by(z.LGREG.b, INDICES = ind.1, colSums, simplify = T), FUN = identity))

  ## Combine Cluster Level Estimating Equations
  W.LGREG.all <- cbind(z.LGREG, t.W.clus)

  # Mean of Cluster Level Estimating Equations
  W.LGREG.all.mean <- t(colMeans(W.LGREG.all)) %x% t(t(rep(1,nrow(W.LGREG.all))))

  # Covariance Matrix
  Sigma.LGREG.j <- (a / (a - 1)) * t(W.LGREG.all - W.LGREG.all.mean) %*% (W.LGREG.all - W.LGREG.all.mean)

  # Variance of LGREG
  var.LGREG.matrix <- diag(J.inv.LGREG %*% Sigma.LGREG.j %*% t(J.inv.LGREG))[(1: Y.dim)]
  v.LGREG.pml[j, ] <- var.LGREG.matrix

#### MCAL
# MCAL Jacobian
Jacob.MCAL <- Jacob[c(MCAL.start:beta.end), c(MCAL.start:beta.end)]
error.9[j.master]
    <- ifelse( (any(abs(eigen(Jacob.MCAL, only.values = TRUE)$values) <= .Machine$double.eps) ||
        qr(Jacob.MCAL)$rank < ncol(Jacob.MCAL)), 1, 0)
if(error.9[j.master] == 1) next

  # Invert the MCAL jacobian of the estimating equations with theta.pml as the input
  J.inv.MCAL <- solve(Jacob.MCAL, tol =1e-23)

  ## Cluster Level Estimating Equations for MCAL
  # The output of this function is the sum of the LGREG estimating equations within each cluster
  z.MCAL <- clus.resid.mc

  ## Cluster Level Estimating Equations for B: Same as LGREG
  ## Combine Cluster Level Estimating Equations
  W.MCAL.all <- cbind(z.MCAL, t.W.clus)

  # Mean of Cluster Level Estimating Equations
  W.MCAL.all.mean <- matrix(rep(colMeans(W.MCAL.all), n.clus.samp), nrow = n.clus.samp, byrow = TRUE)

  # Covariance Matrix
  Sigma.MCAL.j <- (a / (a - 1)) * t(W.MCAL.all - W.MCAL.all.mean) %*% (W.MCAL.all - W.MCAL.all.mean)

  # Variance of MCAL
  var.MC.matrix <- diag(J.inv.MCAL %*% Sigma.MCAL.j %*% t(J.inv.MCAL))[(1: Y.dim)]
  v.MCAL.pml[j, ] <- var.MC.matrix

####### Test: Variance of Estimation Equations (End)

    ## Pseudo Empirical Maximimum Likelihood
    m.1 <- nrow(X.Pop)
    ds <- w.k / sum(w.k)

    # Using GLM
    u <- Samp.fit.glm
    mu <- colMeans(Pop.fit.glm)
    lambda.1 <- Lag2(u = u, ds = ds, mu = mu)
    mu.matrix <- matrix(rep(colMeans(Pop.fit.glm), length(w.k)), nrow = length(w.k), ncol = Y.dim, byrow = TRUE)
    p.i <- (ds) / (1 + (u - mu.matrix) %*% t(t(lambda.1)) )
    t.PEMLE.glm[j, ] <- M.1 * t(p.i) %*% (Y.samp)

    # Using PML: Mean
    u <- Samp.fit.pml
    mu <- colMeans(Pop.fit.pml)
    lambda.1 <- Lag2(u = u, ds = ds, mu = mu)
    mu.matrix <- matrix(rep(colMeans(Pop.fit.pml), length(w.k)), nrow = length(w.k), ncol = Y.dim, byrow = TRUE)
    p.i <- (ds) / (1 + (u - mu.matrix) %*% t(t(lambda.1)) )
    t.PEMLE.pml.N[j, ] <- M.1 * t(p.i) %*% (Y.samp)
    t.PEMLE.pml[j, ] <- t(p.i * sum(w.k)) %*% (Y.samp)
    t.PEMLE.pml.w[j, ] <- (1 / (length(w.k))) * t(w.k/p.i) %*% (Y.samp)


    if(((j) %% 10) == 0)
    {
        cat(j, format(Sys.time(), "%X"), "\n",
                "  True:                  ", sum(Y.Pop[,1]), "\n",
                "  Mean t.HT              ", round(mean(t.HT[1:j,1])), "\n",
                "  Mean t.GREG:           ", round(mean(t.GREG[1:j,1])), "\n",
                "  Mean t.LGREG.pml:      ", round(mean(t.LGREG.pml[1:j,1])), "\n",
                "  Mean t.PROJ.pml:       ", round(mean(t.PROJ.pml[1:j,1])), "\n",
                "  Mean t.MCAL.pml:       ", round(mean(t.MCAL.pml[1:j,1])), "\n",
                "  Mean t.PEMLE.pml.N:    ", round(mean(t.PEMLE.pml.N[1:j,1])), "\n",
                "  Mean t.PEMLE.pml:      ", round(mean(t.PEMLE.pml[1:j,1])), "\n",
                "  Mean t.PEMLE.pml.w:    ", round(mean(t.PEMLE.pml.w[1:j,1])), "\n", "\n",
                "  se t.HT                ", round(sqrt(var(t.HT[1:j,1]))), "\n",
                "  se t.GREG:             ", round(sqrt(var(t.GREG[1:j,1]))), "\n", "\n",
                "  se t.LGREG.pml:        ", round(sqrt(var(t.LGREG.pml[1:j,1]))), "\n",
                "  se.wr t.LGREG.pml:     ", round(sqrt(mean(v.LGREG.wr[1:j,1], na.rm=TRUE))), "\n",
                "  se.ssw t.LGREG.pml:    ", round(sqrt(mean(v.LGREG.ssw[1:j,1], na.rm=TRUE))), "\n",
                "  se.pml t.LGREG.pml:    ", round(sqrt(mean(v.LGREG.pml[1:j,1], na.rm=TRUE))), "\n", "\n",
                "  se t.PROJ.pml:         ", round(sqrt(var(t.PROJ.pml[1:j,1]))), "\n",
```

```
              "  se t.MCAL.pml:          ", round(sqrt(var(t.MCAL.pml[1:j,1], na.rm=TRUE))), "\n",
              "  se.wr t.MCAL.pml:       ", round(sqrt(mean(v.MCAL.wr[1:j,1], na.rm=TRUE))), "\n",
              "  se.ssw.e t.MCAL.pml:    ", round(sqrt(mean(v.MCAL.ssw[1:j,1], na.rm=TRUE))), "\n",
              "  se.ssw.g t.MCAL.pml:    ", round(sqrt(mean(v.MCAL.ssw.g[1:j,1], na.rm=TRUE))), "\n",
              "  se.MCAL.pml:            ", round(sqrt(mean(v.MCAL.pml[1:j,1], na.rm=TRUE))), "\n", "\n",
              "  se t.PEMLE.pml.N:       ", round(sqrt(var(t.PEMLE.pml.N[1:j,1]))), "\n",
              "  se t.PEMLE.pml.w:       ", round(sqrt(var(t.PEMLE.pml.w[1:j,1]))), "\n",
              "  se t.PEMLE.pml:         ", round(sqrt(var(t.PEMLE.pml[1:j,1]))), "\n",
              "  se t.PEMLE.glm:         ", round(sqrt(var(t.PEMLE.glm[1:j,1]))),
              "\n")
    }

    j <- j + 1
    print(j.master)

  }
  list(t.HT, t.GREG,
       t.LGREG.pml,
       t.MCAL.pml,
       t.PEMLE.pml.N, t.PEMLE.pml, t.PEMLE.pml.w,
       v.LGREG.wr, v.LGREG.ssw, v.LGREG.pml,
       v.MCAL.wr, v.MCAL.ssw, v.MCAL.ssw.g, v.MCAL.pml,
       error.1, error.2, error.3, error.4, error.5, error.6, error.7, error.8, error.9, error.10)
}
```

Appendix C

Notes for GLM-Assisted Estimation Paper

C.1 Derivation of Estimating Equations for Poisson Regression

C.1.1 Exponential Family

The probability mass function for a Poisson random variable is

$$\begin{aligned} f\left(y_k;\mu_k\right) &= \frac{e^{-\mu_k}\mu_k^{y_k}}{y_k!} \\ &= e^{y_k \ln\mu_k - \mu - \ln y_k!} \\ &= e^{y_k\eta_k - e^{\eta_k} - \ln y_k!} \end{aligned}$$

This is a member of the natural exponential family with

$$\eta_k = \ln\mu_k$$

$$\zeta\left(\eta_k\right) = e^{\eta_k}$$

$$\phi_k = 1$$

$$h\left(y_k,\phi_k\right) = -\ln y_k!$$

C.1.2 Link Function

Now, we can use any link function that we desire; however, the log link is commonly used because it is the canonical link and simplifies calculations. With the log link

$g(\mu_k) = \ln\mu_k$. Likewise the inverse link function is $g^{-1}\left(\mathbf{x}_k^\top \boldsymbol{\beta}\right) = e^{\mathbf{x}_k^\top \boldsymbol{\beta}}$. We can easily see that μ is the inverse of g, thus $\psi\left(\mathbf{x}_k^\top \boldsymbol{\beta}\right) = \mathbf{x}_k^\top \boldsymbol{\beta}$.

C.1.3 Log Likelihood

Equation (4.2) on page 201 shows the log-likelihood for a GLM. We just showed that $\phi_k = 1$. Thus, $\frac{\phi}{\omega_k} = 1$. Substituting the quantities we found so far gives

$$\begin{aligned}
\ell = \ln L &= \sum_{k\in\mathscr{U}} \left[\ln\left[h\left(y_k, \frac{\phi}{\omega_k}\right)\right] + \frac{\psi\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right) y_k - \zeta\left(\psi\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right)\right)}{\frac{\phi}{\omega_k}}\right] \\
&= \sum_{k\in\mathscr{U}} \ln\left(-\ln y_k!\right) + \left[\left(\mathbf{x}_k^\top \mathbf{B}\right) y_k - \zeta\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right)\right] \\
&= \sum_{k\in\mathscr{U}} \ln\left(-\ln y_k!\right) + \left[\left(\mathbf{x}_k^\top \mathbf{B}\right) y_k - e^{\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right)}\right] \\
&= \sum_{k\in\mathscr{U}} \left[\left(\mathbf{x}_k^\top \mathbf{B}\right) y_k - e^{\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right)} - \ln y_k!\right]
\end{aligned}$$

C.1.4 Estimating Equations

Equation (4.3) on page 202 shows the estimating equations for a GLM. Applying this to the Poisson model with log link gives,

$$w(\boldsymbol{\beta}) = \frac{1}{\phi} \sum_{k\in\mathscr{U}} \left\{\left[y_k - \mu_k\left(\psi\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right)\right)\right] \left[\frac{\partial\psi(\gamma_k)}{\partial\gamma_k}\right] \omega_k \mathbf{x}_k\right\}.$$

We do not have a dispersion parameter, so

$$w(\boldsymbol{\beta}) = \sum_{k\in\mathscr{U}} \left\{\left[y_k - \mu_k\left(\psi\left(\boldsymbol{\beta}^\top \mathbf{x}_k\right)\right)\right] \left[\frac{\partial\psi(\gamma_k)}{\partial\gamma_k}\right] \mathbf{x}_k\right\}.$$

Since $\psi\left(\mathbf{x}_k^\top \boldsymbol{\beta}\right) = \mathbf{x}_k^\top \boldsymbol{\beta}$

$$w(\boldsymbol{\beta}) = \sum_{k\in\mathscr{U}} \left\{\left[y_k - \mu_k\left(\mathbf{x}_k^\top \boldsymbol{\beta}\right)\right] \left[\frac{\partial\psi(\gamma_k)}{\partial\gamma_k}\right] \mathbf{x}_k\right\}.$$

Since $\psi\left(\mathbf{x}_k^\top \boldsymbol{\beta}\right) = \mathbf{x}_k^\top \boldsymbol{\beta}$ and γ_k is defined as $\mathbf{x}_k^\top \boldsymbol{\beta}$, we simplify to

$$w\left(\boldsymbol{\beta}\right) = \sum_{k \in \mathscr{U}} \left\{ \left[y_k - \mu_k \left(\mathbf{x}_k^\top \boldsymbol{\beta} \right) \right] \mathbf{x}_k \right\}.$$

If we simply write $\mu_k \left(\mathbf{x}_k^\top \boldsymbol{\beta} \right)$ as μ_k we have

$$w\left(\boldsymbol{\beta}\right) = \sum_{k \in \mathscr{U}} \left\{ \left[y_k - \mu_k \right] \mathbf{x}_k \right\}.$$

A sample estimator of this is

$$\hat{w}\left(\boldsymbol{\beta}\right) = \sum_{k \in \mathfrak{s}} d_k \left\{ \left[y_k - \mu_k \right] \mathbf{x}_k \right\}.$$

Thus, our pseudomaximum likelihood estimation equations of $\mathbf{B}$ for Poisson regression with a log link are

$$0 = \sum_{k \in \mathfrak{s}} d_k \left\{ \left[y_k - \mu_k \right] \mathbf{x}_k \right\}.$$

Shao (2003) describes numeric methods that can be used to solve this equation for $\mathbf{B}$.

C.2 Derivation of Estimating Equations for Binary Probit Regression

C.2.1 Exponential Family

The probability mass function for a Bernoulli random variable is

$$f\left(y_k; \pi_k\right) = \pi_k^{y_k} \left(1 - \pi_k\right)^{1 - y_k}$$

which can be written as

$$
\begin{aligned}
&= \pi_k^{y_k} \left(1-\pi_k\right)^1 \left(1-\pi_k\right)^{-y_k} \\
&= \pi_k^{y_k} \left(1-\pi_k\right)^{-y_k} \left(1-\pi_k\right)^1 \\
&= \left(\frac{\pi_k}{1-\pi_k}\right)^{y_k} \left(1-\pi_k\right) \\
&= e^{\ln\left[\left(\frac{\pi_k}{1-\pi_k}\right)^{y_k}(1-\pi_k)\right]} \\
&= e^{\ln\left[\left(\frac{\pi_k}{1-\pi_k}\right)^{y_k}\right]+\ln(1-\pi_k)} \\
&= e^{y_k\ln\left[\left(\frac{\pi_k}{1-\pi_k}\right)\right]+\ln(1-\pi_k)} \\
&= e^{y_k\ln\left[\left(\frac{\pi_k}{1-\pi_k}\right)\right]-\ln\left(\frac{1}{1-\pi_k}\right)} \\
&= e^{y_k\ln\left[\left(\frac{\pi_k}{1-\pi_k}\right)\right]-\ln\left(\frac{1-\pi_k+\pi_k}{1-\pi_k}\right)} \\
&= e^{y_k\ln\left[\left(\frac{\pi_k}{1-\pi_k}\right)\right]-\ln\left(1+\frac{\pi_k}{1-\pi_k}\right)} \\
&= e^{y_k\ln\left[\left(\frac{\pi_k}{1-\pi_k}\right)\right]-\ln\left(1+e^{\ln\left(\frac{\pi_k}{1-\pi_k}\right)}\right)}.
\end{aligned}
$$

This is a member of the exponential dispersion family with

$$
\begin{aligned}
\eta_k &= \ln\left(\frac{\pi_k}{1-\pi_k}\right) \\
\zeta\left(\eta_k\right) &= \ln\left(1+e^{\eta_k}\right) \\
\phi_k &= 1 \\
h\left(y_k, \phi_k\right) &= 0.
\end{aligned}
$$

C.2.2 Mean

For exponential families

$$\begin{aligned}
\mu_k(\eta_k) &= \frac{\partial \zeta}{\partial \eta_k} \\
&= \frac{\partial}{\partial \eta_k} \ln\left(1 + e^{\eta_k}\right) \\
&= \frac{1}{1 + e^{\eta_k}} e^{\eta_k}.
\end{aligned}$$

Writing this in terms of π_k gives

$$\begin{aligned}
\mu_k(\eta_k) &= \frac{1}{1 + e^{\ln\left(\frac{\pi_k}{1-\pi_k}\right)}} e^{\ln\left(\frac{\pi_k}{1-\pi_k}\right)} \\
&= \frac{1}{1 + \frac{\pi_k}{1-\pi_k}} \left(\frac{\pi_k}{1-\pi_k}\right) \\
&= \frac{1}{\frac{1+\pi_k-\pi_k}{1-\pi_k}} \left(\frac{\pi_k}{1-\pi_k}\right) \\
&= \frac{1}{\frac{1}{1-\pi_k}} \left(\frac{\pi_k}{1-\pi_k}\right) \\
&= (1-\pi_k) \left(\frac{\pi_k}{1-\pi_k}\right) \\
&= \pi_k.
\end{aligned}$$

Since $\mu_k(\eta_k) = \pi_k$ and $\eta_k = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$, we see that $\eta_k(\mu_k) = \ln\left(\frac{\mu_k}{1-\mu_k}\right) = \mu_k^{-1}(\eta_k)$.

C.2.3 Variance

$$\operatorname{var}(y_k) = \zeta''(\eta_k)\, a(\phi).$$

Since $\phi_k = 1$

$$
\begin{aligned}
\text{var}\left(y_k\right) &= \zeta''\left(\eta_k\right) \\
&= \frac{\partial}{\partial \eta_k} \frac{e^{\eta_k}}{1+e^{\eta_k}} \\
&= \frac{e^{\eta_k}\left(1+e^{\eta_k}\right) - e^{\eta_k}\left(e^{\eta_k}\right)}{\left(1+e^{\eta_k}\right)^2} \\
&= \frac{e^{\eta_k} + e^{2\eta_k} - e^{2\eta_k}}{\left(1+e^{\eta_k}\right)^2} \\
&= \frac{e^{\eta_k}}{\left(1+e^{\eta_k}\right)^2}. \qquad \text{(C.1)}
\end{aligned}
$$

Writing this in terms of μ_k gives

$$
\begin{aligned}
\text{var}\left(y_k\right) &= \frac{e^{\ln\left(\frac{\mu_k}{1-\mu_k}\right)}}{\left(1+e^{\ln\left(\frac{\mu_k}{1-\mu_k}\right)}\right)^2} \\
&= \frac{\frac{\mu_k}{1-\mu_k}}{\left(1+\frac{\mu_k}{1-\mu_k}\right)^2} \\
&= \frac{\frac{\mu_k}{1-\mu_k}}{\left(\frac{1-\mu_k+\mu_k}{1-\mu_k}\right)^2} \\
&= \frac{\frac{\mu_k}{1-\mu_k}}{\left(\frac{1}{1-\mu_k}\right)^2} \\
&= \frac{\mu_k}{1-\mu_k}\left(1-\mu_k\right)^2 \\
&= \mu_k\left(1-\mu_k\right). \qquad \text{(C.2)}
\end{aligned}
$$

C.2.4 Link Functions

The probit link which is defined as

$$g\left(\mu\left(\eta_k\right)\right) = \Phi^{-1}\left(\mu_k\left(\eta_k\right)\right)$$

$$= \gamma_k$$

where Φ^{-1} is the inverse of the standard normal distribution.

Solving our link function for μ_k gives

$$\mu_k = \Phi\left(\gamma_k\right) = \frac{1}{\sqrt{2\pi}} e^{\frac{-\gamma_k^2}{2}}.$$

C.2.5 Estimating Equations

To simplify our estimating equations for the probit link, we must simplify $\frac{\partial \mu_k}{\partial \gamma_k}$. For the probit link $\mu_k = \frac{1}{\sqrt{2\pi}} e^{\frac{\gamma_k^2}{2}}$. Differentiating this with respect to η_k gives

$$\frac{\partial \mu_k}{\partial \gamma_k} = \frac{\partial}{\partial \gamma_k} \frac{1}{\sqrt{2\pi}} e^{\frac{-\gamma_k^2}{2}} \tag{C.3}$$

$$= \frac{1}{\sqrt{2\pi_k}} e^{\frac{-\gamma_k^2}{2}} \tag{C.4}$$

$$= \Phi\left(\gamma_k\right). \tag{C.5}$$

Using Equation (C.3), we write γ_k in terms of μ_k

$$\frac{\partial \mu_k}{\partial \gamma_k} = \Phi\left(\Phi^{-1}\left(\mu_k\right)\right) = \mu_k. \tag{C.6}$$

Equation (4.5) on page 202 shows the pseudomaximum likelihood estimating equations for a GLM. Applying this to the Bernoulli model with a probit link gives,

$$\hat{w}\left(\boldsymbol{\beta}\right) = \frac{1}{\phi} \sum_{k \in \mathfrak{s}} d_k \left\{ \left[y_k - \mu_k\right] \left[\frac{\phi_k}{\operatorname{var}\left(y_k\right)} \frac{\partial \mu_k}{\partial \gamma_k}\right] \omega_k \mathbf{x}_k \right\}.$$

We do not have a dispersion parameter, so

$$\hat{w}\left(\boldsymbol{\beta}\right) = \sum_{k \in \mathfrak{s}} d_k \left\{ \left[y_k - \mu_k\right] \left[\frac{1}{\operatorname{var}\left(y_k\right)} \frac{\partial \mu_k}{\partial \gamma_k}\right] \mathbf{x}_k \right\}.$$

Subsituting the variance from Equation (C.2) gives

$$\hat{w}\left(\boldsymbol{\beta}\right) = \sum_{k \in \mathfrak{s}} d_k \left\{ \frac{\left[y_k - \mu_k\right]}{\mu_k\left(1 - \mu_k\right)} \frac{\partial \mu_k}{\partial \gamma_k} \mathbf{x}_k \right\}.$$

Since we found in Equation (C.6) that $\frac{\partial \mu_k}{\partial \gamma_k} = \mu_k$

$$\begin{aligned}
\hat{w}\left(\boldsymbol{\beta}\right) &= \sum_{k \in \mathfrak{s}} d_k \left\{ \frac{\left[y_k - \mu_k\right]}{\pi_k\left(1 - \pi_k\right)} \mu_k \mathbf{x}_k \right\} \\
&= \sum_{k \in \mathfrak{s}} d_k \left\{ \frac{\left[y_k - \mu_k\right]}{\left(1 - \mu_k\right)} \mathbf{x}_k \right\}.
\end{aligned}$$

Thus, our pseudomaximum likelihood estimation equations of $\mathbf{B}$ for Bernoulli regression with a probit link are

$$0 = \sum_{k \in \mathfrak{s}} d_k \left\{ \frac{\left[y_k - \mu_k\right]}{\left(1 - \mu_k\right)} \mathbf{x}_k \right\}.$$

Shao (2003) describes numeric methods that can be used to solve these equations for $\mathbf{B}$.

C.3 Residuals for GLMs

Models rarely fit the data perfectly. As in linear regression, we can use residuals to assess the fit of our model. The deviance residual is defined as

$$r_k^d = \sqrt{d_k} \times \operatorname{sign}\left(y_k - \widehat{\mu}_k\right)$$

where

$$d_k = 2\omega_k \left[y_k\left(\widetilde{\eta}_k - \widehat{\eta}_k\right) - \zeta\left(\widetilde{\eta}_k\right) + \zeta\left(\widehat{\eta}_k\right)\right]$$

and $\widehat{\eta}_k$ is the maximum likelihood estimate of η_k using the GLM and $\widetilde{\eta}_k$ is the estimate of η_k using $\mathbf{y}_k$ instead of $\mathbf{x}_k^\top \widehat{\boldsymbol{\beta}}$.

The Pearson residual is defined as

$$r_k^p = \frac{y_k - \hat{\mu}_k}{\sqrt{\upsilon_M\left(Y_k\right)}}$$

The standardized Pearson residual is

$$r_k^{sp} = \frac{y_k - \hat{\mu}_k}{\sqrt{\upsilon_M\left(Y_k\right)\left(1 - \hat{h}_k\right)}}$$

where $\hat{h}_k$ are the diagonal elements of

$$\mathbf{H} = \mathbf{W}^{\frac{1}{2}}\mathbf{X}\left(\mathbf{X}^\top\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{W}^{\frac{1}{2}}$$

and $\mathbf{W}$ is a diagonal matrix with each element

$$w_k = \frac{\left(\frac{\partial \mu_k}{\partial \eta_k}\right)^2}{\upsilon_M\left(Y_k\right)}$$

It can also be shown that

$$r_k^{sp} = \frac{r_k^p}{\sqrt{\left(1 - \hat{h}_k\right)}}$$

C.4 GLM-Assisted Difference Estimator

C.4.1 Design Consistency of the Clustered GLM-Assisted Difference Estimator

In Section B.4.1 on page 317, we proved that the generalized difference estimator was design-consistent for the true population total in clustered samples under the mild regularity conditions presented in Section 3.2.1.1. The proof in Section B.4.1 was general and did not use any specific link function, thus the proof holds for any arbitrary link function as long as the assumptions hold. Furthermore, the proof in Section B.4.1 was for a multivariate response variable. The case of a scalar response variable is also covered under the proof by treating $\mathbf{y}_k$ and $\boldsymbol{\mu}_k$ as univariate.

Therefore, based on the proof in Section B.4.1, we conclude that

$$\begin{aligned}
\widehat{t}_y^{gd} &= \sum_{\mathscr{U}} \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) + \sum_{\mathfrak{s}} d_k \left[\mathbf{y}_k - \boldsymbol{\mu}\left(\mathbf{x}_k, \hat{\mathbf{B}}\right)\right] \\
&= \widehat{t}_y + O_p\left(n^{-\frac{1}{2}}\right). \\
&= t_y + O_p\left(n^{-\frac{1}{2}}\right).
\end{aligned}$$

and that $\widehat{t}_y^{gd}$ is a consistent estimator of t_y. Furthermore, $\widehat{t}_y^{gd}$ is asymptotically centered around the Horvitz-Thompson estimator, an unbiased estimator.

C.4.2 Asymptotic Variance of the GLM-Assisted Difference Estimator

In Section B.4.2 on page 319, we proved that the asymptotic variance of the generalized difference estimator with a multivariate response in clustered samples was

$$\operatorname{av}\left(\widehat{\mathbf{t}}_y^{lg}\right) = \sum\sum_{\mathscr{U}_I} \Delta_{ij} \frac{t_{e_i}}{\pi_{Ii}} \frac{t_{e_j}}{\pi_{Ij}} + \sum_{\mathscr{U}_I} \frac{\sum_{\mathscr{U}_i} \sum_{\mathscr{U}_i} \Delta_{kl|i} \frac{e_{k|i}}{\pi_{k|i}} \frac{e_{l|i}}{\pi_{l|i}}}{\pi_{Ii}}.$$

The proof in Section B.4.2 was general and did not use any specific link function, thus the proof holds for any arbitrary link function as long as the asymptotic assumptions hold. Furthermore, the proof in Section B.4.2 was for a multivariate response variable. The case of a scalar response variable is also covered under the proof by treating $\mathbf{y}_k$ and $\boldsymbol{\mu}_k$ as univariate.

Therefore, based on the proof in Section B.4.2, we conclude that

$$\begin{aligned}
\operatorname{av}\left(\hat{t}_y^{gd}\right) &= \operatorname{var}\left(\sum_{i\in\mathfrak{s}_I} d_i \widehat{t}_{ei}\right) \\
&= \sum_{i\in\mathscr{U}_I}\sum_{j\in\mathscr{U}_I} \left(\Delta_{ij} d_i d_j t_{ei} t_{ej}\right) + \sum_{i\in\mathscr{U}_I}\left[d_i\left(\sum_{k\in\mathscr{U}_i}\sum_{l\in\mathscr{U}_i} \Delta_{kl|i} d_{k|i} d_{l|i} e_k e_l\right)\right] \qquad \text{(C.7)} \\
&= \sum\sum_{\mathscr{U}_I} \Delta_{ij} \frac{t_{e_i}}{\pi_i} \frac{t_{e_j}}{\pi_j} + \sum_{\mathscr{U}_I} \frac{\sum_{\mathscr{U}_i} \sum_{\mathscr{U}_i} \Delta_{kl|i} \frac{e_k}{\pi_{k|i}} \frac{e_l}{\pi_{l|i}}}{\pi_i}.
\end{aligned}$$

where

$$\begin{aligned}
e_k &= y_k - \widehat{\mu}_k \\
t_{ei} &= \sum_{k\in\mathscr{U}_i} e_k. \qquad \text{(C.8)}
\end{aligned}$$

C.4.3 Variance Estimators of the GLM-Assisted Difference Estimator

C.4.3.1 Linear Substitute Estimator

In Section B.4.3.1 on page 322, we constructed a linear substitute variance estimator for the asymptotic variance of the generalized difference estimator with a multivariate response in clustered samples. The proof in Section B.4.3.1 was general and did not use any specific link function, thus the proof holds for any arbitrary link function. Furthermore, the proof in Section B.4.3.1 was for a multivariate response variable. The case of a scalar response variable is also covered under the proof by treating $\mathbf{y}_k$ and $\boldsymbol{\mu}_k$ as scalars.

Therefore, based on the derivation in Section B.4.3.1, we conclude that the linear substitute variance estimator for the scalar-valued generalized difference estimator with an arbitrary link function in clustered samples is,

$$\upsilon_e\left(\hat{t}_y^{gd}\right) = \sum\sum_{\mathfrak{s}_I} \frac{\Delta_{ij}}{\pi_{ij}} \frac{\widehat{t}_{\widehat{e}i}^{\pi}}{\pi_i} \frac{\widehat{t}_{\widehat{e}j}^{\pi}}{\pi_j} + \sum_{\mathfrak{s}_I} \frac{\sum_{\mathfrak{s}_i}\sum_{\mathfrak{s}_i} \frac{\Delta_{kl|i}}{\pi_{kl|i}} \frac{\widehat{e}_k}{\pi_{k|i}} \frac{\widehat{e}_l}{\pi_{l|i}}}{\pi_i}$$

where

$$\widehat{t}_{\widehat{e}i}^{\pi} = \sum_{k\in\mathfrak{s}_i} d_k \widehat{e}_k$$

$$\widehat{e}_k = y_k - \widehat{\mu}_k.$$

If the first and second stage samples are selected using a Poisson sampling technique, then υ_e reduces to

$$\sum_{i\in\mathfrak{s}_I} \frac{(1-\pi_i)}{\pi_i^2} \hat{t}_{ei}^2 + \sum_{i\in\mathfrak{s}_I} \frac{1}{\pi_i} \sum_{k\in\mathfrak{s}_i} \frac{\left(1-\pi_{k|i}\right)}{\pi_{k|i}^2} e_k^2.$$

C.4.3.2 With-replacement Estimator

In Section B.4.3.2 on page 324, we constructed a with-replacement variance estimator for the asymptotic variance of the generalized difference estimator with a multivariate response in clustered samples. The proof in Section B.4.3.2 was general and did not use any specific link function, thus the variance estimator holds for any arbitrary link function. Furthermore, the derivation in Section B.4.3.2 was for a multivariate response variable. The case of a scalar response variable is also covered under the proof by treating $\mathbf{y}_k$ and $\boldsymbol{\mu}_k$ as univariate.

Thus, by Section B.4.3.2, the with-replacement variance estimator is

$$\upsilon_{wr}\left(t_y^{gd}\right) = \frac{n}{(n-1)} \sum_{i \in \mathfrak{s}_I}^{n} \left(d_i \widehat{t}_{\widehat{e}i}^{\pi} - \frac{1}{n} t_{\widehat{e}}^{\pi} \right)^2 \tag{C.9}$$

where

$$\widehat{t}_{\widehat{e}}^{\pi} = \sum_{k \in \mathfrak{s}} (d_k \widehat{e}_k) \widehat{t}_{\widehat{e}i}^{\pi} \tag{C.10}$$

$$= \sum_{k \in \mathfrak{s}_i} d_k \widehat{e}_k \tag{C.11}$$

and

$$\widehat{e}_k = y_k - \widehat{\mu}_k. \tag{C.12}$$

C.4.3.3 Implicit Differentiation Variance Estimator

In Section B.4.3.3 on page 325, we showed that for multinomial logistic regression

$$\upsilon_{Binder}\left(\widehat{\boldsymbol{\theta}}\right) = \left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right] \left[\widehat{\boldsymbol{\Sigma}}_{\widehat{\mathbf{U}}}\left(\widehat{\boldsymbol{\theta}}\right)\right] \left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{\top}$$

where

$$\widehat{\boldsymbol{\theta}} = \begin{bmatrix} \widehat{\mathbf{t}}_y^{gd} \\ \mathrm{vec}\left(\widehat{\mathbf{B}}\right) \end{bmatrix}$$

$$\widehat{\mathbf{t}}_y^{gd} = \sum_{\mathscr{U}} \widehat{\boldsymbol{\mu}}_k + \sum_{\mathfrak{s}} d_k \left[\mathbf{y}_k - \widehat{\boldsymbol{\mu}}_k\right]$$

$$\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right) = \frac{n}{n-1}\left\{\sum_{\mathfrak{s}_I}\left[\widehat{\mathbf{t}}_{\hat{U}i} - \frac{1}{n}\sum_{i\in\mathfrak{s}_I}\widehat{\mathbf{t}}_{\hat{U}i}\right]\right\}\left\{\sum_{\mathfrak{s}_I}\left[\widehat{\mathbf{t}}_{\hat{U}i} - \frac{1}{n}\sum_{i\in\mathfrak{s}_I}\widehat{\mathbf{t}}_{\hat{U}i}\right]\right\}^\top$$

and

$$\mathbf{J}\left(\boldsymbol{\theta}\right) = \frac{\partial}{\partial\left(\mathrm{vec}\boldsymbol{\theta}\right)^\top}\mathbf{W}\left(\boldsymbol{\theta}\right). \tag{C.13}$$

Although the simplification of $\mathbf{J}\left(\boldsymbol{\theta}\right)$ depends on the link function, the general form of Equation C.13 holds under basic regularity conditions, regardless of the link function. When y_k is a univariate response and μ_k is based on a GLM, the Binder estimator still holds. Specifically the estimator will be

$$\upsilon_{Binder}\left(\widehat{\boldsymbol{\theta}}\right) = \left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right]\left[\widehat{\boldsymbol{\Sigma}}_{\hat{\mathbf{U}}}\left(\widehat{\boldsymbol{\theta}}\right)\right]\left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right]^\top$$

where

$$\widehat{\boldsymbol{\theta}} = \begin{bmatrix} \widehat{t}_y^{gd} \\ \\ \widehat{\mathbf{B}} \end{bmatrix}$$

$$\widehat{t}_y^{gd} = \sum_{\mathscr{U}} \widehat{\mu}_k + \sum_{\mathfrak{s}} d_k \left[y_k - \widehat{\mu}_k \right]$$

$$\widehat{\boldsymbol{\Sigma}} \left(\widehat{\boldsymbol{\theta}} \right) = \frac{n}{n-1} \left\{ \sum_{\mathfrak{s}_I} \left[\widehat{t}_{\hat{U}i} - \frac{1}{n} \sum_{i \in \mathfrak{s}_I} \widehat{t}_{\hat{U}i} \right] \right\}^2$$

$$\widehat{\mathbf{U}}_k \left(\boldsymbol{\theta} \right) = \begin{bmatrix} d_k \left[y_k - \mu_k \right] \\ \\ \frac{d_k}{\phi} \left\{ \left[y_k - \mu_k \right] \left[\frac{\phi_k}{\mathrm{var}(y_k)} \frac{\partial \mu_k}{\partial \gamma_k} \right] \omega_k \mathbf{x}_k \right\} \end{bmatrix}$$

$$\widehat{\mathbf{U}} \left(\boldsymbol{\theta} \right) = \sum_{\mathfrak{s}} \widehat{\mathbf{U}}_k \left(\boldsymbol{\theta} \right)$$

and

$$\mathbf{J} \left(\boldsymbol{\theta} \right) = \frac{\partial}{\partial \left(\boldsymbol{\theta} \right)} \mathbf{W} \left(\boldsymbol{\theta} \right).$$

To simplify $\mathbf{J}\left(\boldsymbol{\theta}\right)$, further calculus is needed and will depend on the link function. In practice, numeric derivatives can be used.

C.5 Model Calibration

C.5.1 Construction of the Model-Calibrated Estimator

In Section B.5.1 on page 336, we constructed the model-calibrated estimator for a multivariate response in clustered samples.

The logic in Section B.5.1 was general and did not make any references to any specific link function, thus the derivation holds for any arbitrary link function. Furthermore, the calculations in Section B.5.1 was for a multivariate response variable. The case of a univariate response variable is also covered under the derivation by treating $\mathbf{y}_k$ and $\boldsymbol{\mu}_k$ as scalars.

Therefore, based on the calculations in Section B.5.1, we conclude that the model-calibrated estimated total is

$$\begin{aligned}\widehat{t}_y^{mc} &= \mathbf{y}^\top \mathbf{w}^{mc} \\ &= \widehat{t}_y + \widehat{\mathbf{B}}_{\underline{\boldsymbol{\mu}}\mathbf{y}} \left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top \mathbf{1} - \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \mathbf{d} \right)\end{aligned}$$

where

$$\underline{\boldsymbol{\mu}} = \begin{bmatrix} \mathbf{1} & \boldsymbol{\mu} \end{bmatrix}$$

and

$$\underset{1\times 2}{\widehat{\mathbf{B}}_{\underline{\boldsymbol{\mu}}\mathbf{Y}}} = \mathbf{y}^\top \left(\boldsymbol{\Pi}^{-1}\mathbf{Q} \right) \underline{\boldsymbol{\mu}}_{\mathfrak{s}} \left(\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \left(\boldsymbol{\Pi}^{-1}\mathbf{Q} \right) \underline{\boldsymbol{\mu}}_{\mathfrak{s}} \right)^{-1}$$

C.5.2 Alternative Forms of the Model-Calibrated Estimator

In Section B.5.2, we derived several alternative forms of the the model-calibrated estimator. Like previous sections in this appendix, we refer to those results since the derivations were general and did not make specific reference to any formula for μ_k. Again, we present results from Appendix B.5.2 for univariate response variables.

The model-calibrated estimator can be written in the following alternative form

$$\hat{t}_y^{mc} = \mathbf{y}^\top \mathbf{\Pi}^{-1} \mathbf{g}$$

where

$$\underset{n\times 1}{\mathbf{g}} = \mathbf{1} + \mathbf{Q}\underline{\boldsymbol{\mu}}_{\mathfrak{s}} \left(\underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \left(\mathbf{\Pi}^{-1}\mathbf{Q} \right) \underline{\boldsymbol{\mu}}_{\mathfrak{s}} \right)^{-1} \left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^\top \mathbf{1} - \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^\top \mathbf{d} \right).$$

We can compactly write our estimator if we let

$$\mathbf{\Pi}_\star = \mathbf{\Pi}\mathbf{Q}^{-1}$$

$$\underset{2\times 1}{\mathbf{t}_{\underline{\widehat{\boldsymbol{\mu}}}}} = \underline{\widehat{\boldsymbol{\mu}}}_{\mathscr{U}}^\top \mathbf{1} = \sum_{\mathscr{U}} \underline{\widehat{\boldsymbol{\mu}}}_k$$

$$\underset{2\times 1}{\widehat{\mathbf{t}}_{\underline{\widehat{\boldsymbol{\mu}}}}} = \underline{\widehat{\boldsymbol{\mu}}}_{\mathfrak{s}}^\top \mathbf{d} = \sum_{\mathfrak{s}} d_k \underline{\widehat{\boldsymbol{\mu}}}_k$$

$$\underset{2\times 2}{\widehat{\mathbf{A}}} = \underline{\widehat{\boldsymbol{\mu}}}_{\mathfrak{s}}^\top \mathbf{\Pi}_\star^{-1} \underline{\widehat{\boldsymbol{\mu}}}_{\mathfrak{s}} = \sum_{\mathfrak{s}} \frac{d_k}{q_k} \underline{\widehat{\boldsymbol{\mu}}}_k \underline{\widehat{\boldsymbol{\mu}}}_k^\top.$$

For convenience, we let

$$\underset{2\times 1}{\mathbf{t}_{\underline{\boldsymbol{\mu}}}} = \underline{\boldsymbol{\mu}}_{\mathscr{U}}^{\top}\mathbf{1} = \sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}_k$$

$$\widehat{\mathbf{t}}_{\underline{\boldsymbol{\mu}}}\left(\mathbf{B}\right) = \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\mathbf{d} = \sum_{\mathfrak{s}} d_k \underline{\boldsymbol{\mu}}_k$$

$$\underset{2\times 2}{\mathbf{A}} = \underline{\boldsymbol{\mu}}_{\mathscr{U}}^{\top}\underline{\boldsymbol{\mu}}_{\mathscr{U}} = \sum_{\mathscr{U}} \underline{\boldsymbol{\mu}}_k \underline{\boldsymbol{\mu}}_k^{\top}$$

$$\widehat{\mathbf{A}}\left(\mathbf{B}\right) = \underline{\boldsymbol{\mu}}_{\mathfrak{s}}^{\top}\boldsymbol{\Pi}_{\star}^{-1}\underline{\boldsymbol{\mu}}_{\mathfrak{s}} = \sum_{\mathfrak{s}} \frac{d_k}{q_k}\underline{\boldsymbol{\mu}}_k \underline{\boldsymbol{\mu}}_k^{\top}.$$

The model-calibrated estimator of a finite population total is

$$\begin{aligned}
\widehat{t}_y^{mc} &= \mathbf{y}^{\top}\left[\mathbf{d} + \left(\boldsymbol{\Pi}^{-1}\mathbf{Q}\right)\underline{\widehat{\boldsymbol{\mu}}}_{\mathfrak{s}}\left(\underline{\widehat{\boldsymbol{\mu}}}_{\mathfrak{s}}^{\top}\left(\boldsymbol{\Pi}^{-1}\mathbf{Q}\right)\underline{\widehat{\boldsymbol{\mu}}}_{\mathfrak{s}}\right)^{-1}\left(\underline{\widehat{\boldsymbol{\mu}}}_{\mathscr{U}}^{\top}\mathbf{1} - \underline{\widehat{\boldsymbol{\mu}}}_{\mathfrak{s}}^{\top}\mathbf{d}\right)\right]. \\
&= \widehat{t}_y + \mathbf{y}_{\mathfrak{s}}^{\top}\underline{\widehat{\boldsymbol{\mu}}}_{\mathfrak{s}}\left[\widehat{\mathbf{A}}\right]^{-1}\left(\mathbf{t}_{\underline{\widehat{\boldsymbol{\mu}}}} - \widehat{\mathbf{t}}_{\underline{\widehat{\boldsymbol{\mu}}}}\right) \\
&= \sum_{\mathfrak{s}} \frac{d_k}{q_k}\mathbf{y}_k\left[1 + \underline{\widehat{\boldsymbol{\mu}}}_k^{\top}\left[\widehat{\mathbf{A}}\right]^{-1}\left(\mathbf{t}_{\underline{\widehat{\boldsymbol{\mu}}}} - \widehat{\mathbf{t}}_{\underline{\widehat{\boldsymbol{\mu}}}}\right)\right].
\end{aligned}$$

C.5.3 Design Consistency of the Model-Calibrated Estimator

In Section B.5.3 on page 339, we proved that the model-calibrated estimator was design-consistent for the true population total in clustered samples under the mild regularity conditions presented in Section 3.2.1.1.

The proof in Section B.5.3 was general and did not use any specific link function, thus the proof holds for any arbitrary link function as long as the assumptions hold. Furthermore, the proof in Section B.5.3 was for a multivariate response variable. The case of a scalar response variable is also covered under the proof by treating $\mathbf{y}_k$ and $\boldsymbol{\mu}_k$ as scalars.

Therefore, based on the proof in Sections B.5.3 and B.4.1, we conclude that

$$\begin{aligned}\widehat{t}_y^{mc} &= \widehat{t}_y^{\pi} + \widehat{\mathbf{B}}_{\underline{\boldsymbol{\mu}}\mathbf{Y}} \left(\underline{\boldsymbol{\mu}}_{\mathscr{U}}^{\top} \mathbf{1} - \underline{\boldsymbol{\mu}}_{\mathscr{S}}^{\top} \mathbf{d} \right) \\ &= \widehat{t}_y^{\pi} + \widehat{\mathbf{B}}_{\underline{\boldsymbol{\mu}}\mathbf{Y}} O_p \left(n^{-\frac{1}{2}} \right)\end{aligned}$$

Furthermore, $\hat{\mathbf{B}}_{\underline{\boldsymbol{\mu}}\mathbf{Y}} = O_p(1)$. Thus,

$$\widehat{t}_y^{mc} = \widehat{t}_y^{\pi} + O_p \left(n^{-\frac{1}{2}} \right)$$

Since $\widehat{t}_y^{\pi} = t_y + O_p \left(n^{-\frac{1}{2}} \right)$, we see that $\hat{t}_y^{mc}$ is a consistent estimator. Furthermore, it is asymptotically centered around the π-estimator, an unbiased estimator.

C.5.4 Asymptotic Variance of the Model-Calibrated Estimator

In Section B.5.4 on page 342, we proved that the asymptotic variance of the model calibration estimator with a multivariate response in clustered samples was

$$\text{av}\left(\hat{t}_y^{mc}\right) = \sum \sum_{\mathscr{U}_I} \Delta_{ij} \frac{t_{e_i}}{\pi_i} \frac{t_{e_j}}{\pi_j} + \sum_{\mathscr{U}_I} \frac{\sum_{\mathscr{U}_i} \sum_{\mathscr{U}_i} \Delta_{kl|i} \frac{e_k}{\pi_{k|i}} \frac{e_l}{\pi_{l|i}}}{\pi_i} \tag{C.14}$$

where

$$\begin{aligned} e_k &= y_k - \widehat{\mu}_k \\ t_{ei} &= \sum_{k \in \mathscr{U}_i} e_k \end{aligned}$$

The proof in Section B.5.4 was general and did not use any specific link function, thus the proof holds for any arbitrary link function as long as the asymptotic assumptions hold. Furthermore, the proof in Section B.5.4 was for a multivariate response variable. The case of a scalar response variable is also covered under the proof by treating $\mathbf{y}_k$ and $\boldsymbol{\mu}_k$ as scalars.

Therefore, based on the proof in Section B.5.4, we conclude that the asymptotic variance of the model-calibration estimator in clustered samples where μ_k is estimated with a GLM is Equation (C.14).

C.5.5 Variance Estimators of the Model-Calibrated Estimator

C.5.5.1 Linear Substitute Variance Estimators

If we simply estimate the totals in Equation (C.14), we get an estimator for the asymptotic variance of $\hat{t}_y^{mc}$

$$\upsilon_e\left(\hat{t}_y^{mc}\right) = \sum_{i\in\mathfrak{s}_I}\sum_{j\in\mathfrak{s}_I}\left(d_{ij}\Delta_{ij}d_i d_j \hat{t}_{\hat{e}i}\hat{t}_{\hat{e}j}\right) + \sum_{i\in\mathfrak{s}_I}\left[d_i\left(\sum_{k\in\mathfrak{s}_i}\sum_{l\in\mathfrak{s}_i} d_{kl|i}\Delta_{kl|i}d_{k|i}d_{l|i}\hat{e}_k\hat{e}_l\right)\right]$$

where

$$\hat{t}_{\hat{\mathbf{e}}i} = \sum_{k\in\mathfrak{s}_i} d_{k|i}\hat{e}_{k|i}$$

$$\hat{e}_k = y_k - \hat{\mu}_k.$$

Särndal et al. (1989) argue that this estimator tends to underestimate the true sampling error in practice for single-staged samples. For this reason, Särndal et al. (1992) prefer a variant of υ_e based on an adjustment to the residuals.

Using the weighted residual technique advocated in Särndal et al. (1989), we replace $\hat{e}_k$ with $g_k\hat{e}_k$, where g_k is the kth element in the vector

$$\mathbf{g} = \left[\mathbf{1} + \mathbf{Q}\underline{\hat{\boldsymbol{\mu}}}_{\mathfrak{s}}\left(\underline{\hat{\boldsymbol{\mu}}}_{\mathfrak{s}}^{\top}\left(\boldsymbol{\Pi}^{-1}\mathbf{Q}\right)\underline{\hat{\boldsymbol{\mu}}}_{\mathfrak{s}}\right)^{-1}\left(\underline{\hat{\boldsymbol{\mu}}}_{\mathscr{U}}^{\top}\mathbf{1} - \underline{\hat{\boldsymbol{\mu}}}_{\mathfrak{s}}^{\top}\mathbf{d}\right)\right].$$

That is

$$g_k = \left[1 + q_k \underline{\hat{\boldsymbol{\mu}}}_k^\top \left[\sum_{k\in\mathfrak{s}} d_k q_k \underline{\hat{\boldsymbol{\mu}}}_k \underline{\hat{\boldsymbol{\mu}}}_k^\top\right]^{-1} \left[\sum_{k\in\mathscr{U}} \underline{\hat{\boldsymbol{\mu}}}_k - \sum_{k\in\mathfrak{s}} d_k \underline{\hat{\boldsymbol{\mu}}}_k\right]\right].$$

Thus, the weighted residual estimator is

$$\upsilon_g\left(\hat{t}_y^{mc}\right) = \sum_{i\in\mathfrak{s}_I}\sum_{j\in\mathfrak{s}_I}\left(d_{ij}\Delta_{ij}d_i d_j \hat{t}_{g\hat{e}i}\hat{t}_{g\hat{e}j}\right) + \sum_{i\in\mathfrak{s}_I}\left[d_i\left(\sum_{k\in\mathfrak{s}_i}\sum_{l\in\mathfrak{s}_i} d_{kl|i}\Delta_{kl|i}d_{k|i}d_{l|i}g_k\hat{e}_k g_l\hat{e}_l\right)\right]$$

where

$$\hat{t}_{g\hat{e}i} = \sum_{\mathfrak{s}_i}\frac{g_k\hat{e}_k}{\pi_{k|i}}.$$

If the first and second stage samples are selected using a Poisson sampling technique, then $\upsilon_g\left(\hat{t}_y^{mc}\right)$ reduces to

$$\upsilon_g\left(\hat{t}_y^{mc}\right) = \sum_{i\in\mathfrak{s}_I}\frac{(1-\pi_i)}{\pi_i^2}\hat{t}_{g\hat{e}i}\hat{t}_{g\hat{e}i} + \sum_{i\in\mathfrak{s}_I}\frac{1}{\pi_i}\sum_{k\in\mathfrak{s}_i}\frac{\left(1-\pi_{k|i}\right)}{\pi_{k|i}^2}g_k^2\hat{e}_k\hat{e}_k$$

C.5.5.2 With-Replacement Estimator

In Section B.5.5.2 on page 347, we constructed a with-replacement variance estimator for the asymptotic variance of the model-calibrated estimator with a multivariate response in clustered samples. The proof in Section B.5.5.2 was general and did not use any specific link function, thus the variance estimator holds for any arbitrary link function. Furthermore, the derivation in Section B.5.5.2 was for a multivariate response variable. The case of a scalar response variable is also covered under the proof by treating $\mathbf{y}_k$ and $\boldsymbol{\mu}_k$ as univariate.

Thus, by Section B.5.5.2, the with-replacement variance estimator is

$$\upsilon_{wr}\left(\mathbf{t}_y^{mc}\right) = \frac{n}{(n-1)}\sum_{i\in\mathfrak{s}_I}\left(\frac{\widehat{t}_{\widehat{e}i}^{\pi}}{\pi_i} - \frac{1}{n}\widehat{t}_{\widehat{e}}^{\pi}\right)^2$$

where

$$\widehat{t}_{\widehat{e}}^{\pi} = \sum_{k \in \mathfrak{s}} \left(d_k \widehat{e}_k \right)$$

$$\widehat{t}_{\widehat{e}i}^{\pi}$$

$$= \sum_{k \in \mathfrak{s}_i} d_{k|i} \widehat{e}_{k|i}$$

and

$$\widehat{e}_k = y_k - \widehat{\mu}_k.$$

C.5.5.3 Implicit Differentiation Estimator

In Section B.5.5.3 on page 348, we showed that for multinomial logistic regression

$$\upsilon_{Binder}\left(\widehat{\boldsymbol{\theta}}\right) = \left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right]\left[\widehat{\boldsymbol{\Sigma}}_{\widehat{\mathbf{U}}}\left(\widehat{\boldsymbol{\theta}}\right)\right]\left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{\top}$$

where

$$\widehat{\boldsymbol{\theta}} = \begin{bmatrix} \widehat{\mathbf{t}}_y^{mc} \\ \mathrm{vec}\left(\widehat{\mathbf{B}}\right) \end{bmatrix}$$

$$\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right) = \frac{n}{n-1}\left\{\sum_{\mathfrak{s}_I}\left[\widehat{\mathbf{t}}_{\hat{U}i} - \frac{1}{n}\sum_{i \in \mathfrak{s}_I}\widehat{\mathbf{t}}_{\hat{U}i}\right]\right\}\left\{\sum_{\mathfrak{s}_I}\left[\widehat{\mathbf{t}}_{\hat{U}i} - \frac{1}{n}\sum_{i \in \mathfrak{s}_I}\widehat{\mathbf{t}}_{\hat{U}i}\right]\right\}^{\top}$$

$$\mathbf{J}\left(\boldsymbol{\theta}\right) = \frac{\partial}{\partial\left(\mathrm{vec}\boldsymbol{\theta}\right)^{\top}}\mathbf{W}\left(\boldsymbol{\theta}\right).$$

Although the simplification of $\mathbf{J}\left(\boldsymbol{\theta}\right)$ depends on the link function, the general form above holds under basic regularity conditions, regardless of the link function. When y_k is a univariate response and μ_k is based on a GLM, the Binder estimator still holds. Specifically

the estimator will be

$$\upsilon_{Binder}\left(\widehat{\boldsymbol{\theta}}\right) = \left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right]\left[\widehat{\boldsymbol{\Sigma}}_{\widehat{\mathbf{U}}}\left(\widehat{\boldsymbol{\theta}}\right)\right]\left[\widehat{\mathbf{J}}^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right]^{\top}$$

where

$$\widehat{\boldsymbol{\theta}} = \begin{bmatrix} \widehat{t}_y^{mc} \\ \\ \widehat{\mathbf{B}} \end{bmatrix}$$

$$\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right) = \frac{n}{n-1}\left\{\sum_{\mathfrak{s}_I}\left[\widehat{t}_{\hat{U}i} - \frac{1}{n}\sum_{i\in\mathfrak{s}_I}\widehat{t}_{\hat{U}i}\right]\right\}^2$$

$$\widehat{\mathbf{U}}_k\left(\boldsymbol{\theta}\right) = \begin{bmatrix} d_k\mathbf{y}_k\left[1 + \underline{\boldsymbol{\mu}}_k^{\top}\left[\widehat{\mathbf{A}}\left(\mathbf{B}\right)\right]^{-1}\left(\mathbf{t}_{\underline{\boldsymbol{\mu}}} - \widehat{\mathbf{t}}_{\underline{\boldsymbol{\mu}}}\left(\mathbf{B}\right)\right)\right] \\ \frac{d_k}{\phi}\left\{\left[y_k - \mu_k\right]\left[\frac{\phi_k}{\mathrm{var}(y_k)}\frac{\partial\mu_k}{\partial\gamma_k}\right]\omega_k\mathbf{x}_k\right\} \end{bmatrix}$$

$$\widehat{\mathbf{U}}\left(\boldsymbol{\theta}\right) = \sum_{\mathfrak{s}}\widehat{\mathbf{U}}_k\left(\boldsymbol{\theta}\right)$$

and

$$\mathbf{J}\left(\boldsymbol{\theta}\right) = \frac{\partial}{\partial\left(\boldsymbol{\theta}\right)}\mathbf{W}\left(\boldsymbol{\theta}\right).$$

To simplify $\mathbf{J}\left(\boldsymbol{\theta}\right)$, further calculus is needed and will depend on the link function. In practice, numeric derivatives can be used.

C.6 Model-Calibrated Maximum Pseudoempirical Likelihood Estimator

C.6.1 Estimation of the Model-Calibrated Maximum Pseudoempirical Likelihood Estimator

In Section B.6.1, which starts on page 361, we showed that the model-calibrated maximum pseudoempirical likelihood estimator of a multivariate total from a clustered sample is

$$\widehat{\mathbf{t}}_y^{peM} = M \sum_{\mathfrak{s}} p_k \mathbf{y}_k$$

or

$$\widehat{\mathbf{t}}_y^{pe\widehat{M}} = \widehat{M} \sum_{\mathfrak{s}} p_k \mathbf{y}_k$$

where p_k is computed by iteratively solving

$$\begin{aligned} p_k &= \frac{d_k^\star}{1 + \boldsymbol{\lambda}_{2\star}^\top \mathbf{u}_k} \\ \mathbf{0} &= \sum_{\mathfrak{s}} \frac{d_k^\star \mathbf{u}_k}{1 + \boldsymbol{\lambda}_{2\star}^\top \mathbf{u}_k} \end{aligned}$$

and

$$\begin{aligned} d_k^\star &= \frac{d_k}{\hat{M}} \\ \boldsymbol{\lambda}_{2\star} &= \frac{1}{\hat{M}} \boldsymbol{\lambda}_2. \end{aligned}$$

Section B.6.1 derived the model-calibrated maximum pseudoempirical likelihood estimator for a multivariate response under a GLM. Since a scalar response is a subset of a multivariate response, we can apply the derivation in Section B.6.1 to the scalar

response by replacing $\mathbf{y}_k$ with y_k and $\boldsymbol{\mu}_k$ with μ_k. Thus, the model-calibrated maximum pseudoempirical likelihood estimator is

$$\widehat{t}_y^{peM} = M \sum_{\mathfrak{s}} p_k y_k$$

or

$$\widehat{t}_y^{pe\widehat{M}} = \widehat{M} \sum_{\mathfrak{s}} p_k y_k$$

where p_k is computed by iteratively solving

$$\begin{aligned} p_k &= \frac{d_k^\star}{1 + \lambda_{2\star}^\top \mathbf{u}_k} \\ \mathbf{0} &= \sum_{\mathfrak{s}} \frac{d_k^\star \mathbf{u}_k}{1 + \lambda_{2\star}^\top u_k} \end{aligned}$$

and

$$\begin{aligned} d_k^\star &= \frac{d_k}{\hat{M}} \\ \lambda_{2\star} &= \frac{1}{\hat{M}} \lambda_2 \\ u_k &= \mu_k - \frac{1}{M} \sum_{\mathscr{U}} \mu_k. \end{aligned}$$

C.6.2 $\bar{t}_y^{peM}$ is Asymptotically Equal to $\bar{t}_y^{mc}$

In Section B.6.2 on page 365, we showed that $\bar{t}_y^{peM}$ was asymptotically equal to $\bar{t}_y^{mc}$. Although we made reference to our link function in Section B.6.2, our proof did not rely on a specific link function. Thus, we proved that the model-calibrated maximum pseudoempirical likelihood estimator was equivalent to the model-calibrated estimator for any GLM that met our assumptions. Furthermore, since a scalar response is a special case of a multivariate response, our proof easily applies to univariate responses.

Thus, by the proof in Section B.6.2, we conclude that

$$\begin{aligned}
\hat{\hat{t}}_y^{pe} &= \frac{1}{M}\left\{\sum_{\mathfrak{s}} d_k y_k + \hat{B}_{\mathbf{u},\mathbf{y}}\left[\sum_{\mathscr{U}} \mu\left(\mathbf{x}_k, \hat{\mathbf{B}}\right) - \sum_{\mathfrak{s}} d_k \mu\left(\mathbf{x}_k, \hat{\mathbf{B}}\right)\right]\right\} - o_p\left(n^{-\frac{1}{2}}\right) \\
&= \frac{1}{N}\widehat{t}_y^{mc} - o_p\left(n^{-\frac{1}{2}}\right) \\
&= \widehat{\bar{t}}_y^{mc} - o_p\left(n^{-\frac{1}{2}}\right)
\end{aligned}$$

where

$$\hat{B}_{\mathbf{u},\mathbf{y}} = \frac{\sum_{\mathfrak{s}}\left[d_k u_k y_k^\top\right]}{\sum_{\mathfrak{s}} d_k u_k^2}$$

We conclude that $\hat{t}_y^{peM}$ is asymptotically equivalent to the model-calibrated estimator and propose using the model-calibrated variance estimators to estimate the variance of $\hat{t}_y^{peM}$.

In so far as $\hat{M}$ can be replaced by M, we also conclude that $\hat{t}_y^{pe\widehat{M}}$ is asymptotically equivalent to the model-calibrated estimator. Depending on the measure of size, $\hat{M}$ is often equal to M in probability proportional to size samples. But in general, using $\hat{M}$ will add variance to the model-calibrated maximum pseudoempirical likelihood estimator.

C.7 Simulation Results

This section contains tables and graphs summarizing our analysis of the simulations for the GLM-assisted estimators. Formulas for the summary measures that follow can be found in Table 1.1 of Section 1.1.6.

C.7.1 Simulation Coefficient of Variation

Here we present estimates of the simulation coefficient of variation, which we define as

$$CV_{sim} = \frac{\frac{1}{\sqrt{\aleph}}\sqrt{\frac{1}{\aleph-1}\sum_{\nu=1}^{\aleph}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^2}}{\boldsymbol{\theta}}$$

Because the simulation coefficients of variation were so small, we multiplied them by $1,000,000$.

C.7.1.1 Simulation Coefficient of Variation of Count Response

Table C.1: Simulation Coefficient of Variation for Point Estimators of Count Response. Estimates have been multiplied by $1,000,000$.

	Small Samples			Large Samples		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
			No Link Function			
$\widehat{\mathrm{t}}^{\pi}$	26.7	26.7	12.2	8.7	9.0	3.9
			Identity Link Function			
$\widehat{\mathrm{t}}^{gd}$	11.0	11.2	10.1	3.9	3.9	3.3
$\widehat{\mathrm{t}}^{peM}$	10.9	11.1	10.0	3.9	3.9	3.3
$\widehat{\mathrm{t}}^{pe\widehat{M}}$	24.3	24.4	10.0	8.0	8.2	3.3
			Probit Link Function			
$\widehat{\mathrm{t}}^{pr}$	6.3	6.3	5.9	2.3	2.3	1.9
$\widehat{\mathrm{t}}^{gd}$	6.5	6.5	6.0	2.3	2.3	1.9
$\widehat{\mathrm{t}}^{mc}$	6.3	6.3	5.9	2.3	2.3	1.9
$\widehat{\mathrm{t}}^{peM}$	6.3	6.3	5.9	2.3	2.3	1.9
$\widehat{\mathrm{t}}^{pe\widehat{M}}$	24.0	24.1	5.9	7.8	8.1	1.9
			Complementary Log-Log Link Function			
$\widehat{\mathrm{t}}^{pr}$	6.3	6.3	5.9	2.3	2.3	1.9
$\widehat{\mathrm{t}}^{gd}$	6.5	6.5	6.0	2.3	2.3	1.9
$\widehat{\mathrm{t}}^{mc}$	6.3	6.3	5.9	2.3	2.3	1.9
$\widehat{\mathrm{t}}^{peM}$	6.3	6.3	5.9	2.3	2.3	1.9
$\widehat{\mathrm{t}}^{pe\widehat{M}}$	24.0	24.2	5.9	7.8	8.1	1.9
			Log Link Function			
$\widehat{\mathrm{t}}^{pr}$	6.4	6.3	5.9	2.3	2.3	1.9
$\widehat{\mathrm{t}}^{gd}$	6.6	6.5	6.0	2.3	2.3	1.9
$\widehat{\mathrm{t}}^{mc}$	6.3	6.3	5.9	2.3	2.3	1.9
$\widehat{\mathrm{t}}^{peM}$	6.3	6.3	5.9	2.3	2.3	1.9
$\widehat{\mathrm{t}}^{pe\widehat{M}}$	24.0	24.2	5.9	7.8	8.1	1.9
			Cauchit Link Function			
$\widehat{\mathrm{t}}^{pr}$	6.4	6.4	5.9	2.3	2.3	1.9
$\widehat{\mathrm{t}}^{gd}$	6.6	6.6	6.0	2.3	2.3	1.9
$\widehat{\mathrm{t}}^{mc}$	6.4	6.4	5.9	2.3	2.3	1.9
$\widehat{\mathrm{t}}^{peM}$	6.4	6.4	5.9	2.3	2.3	1.9
$\widehat{\mathrm{t}}^{pe\widehat{M}}$	24.1	24.2	5.9	7.9	8.1	1.9

C.7.1.2 Simulation Coefficient of Variation of Binary Response

Table C.2: Simulation Coefficient of Variation for Point Estimators of Binary Response. Estimates have been multiplied by $1,000,000$.

	Small Samples			Large Samples		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
			No Link Function			
$\widehat{\mathbf{t}}^{\pi}$	41.4	41.1	26.5	13.7	13.9	8.4
			Identity Link Function			
$\widehat{\mathbf{t}}^{gd}$	20.0	19.8	18.4	7.2	7.2	6.0
$\widehat{\mathbf{t}}^{peM}$	22.3	22.2	20.1	7.5	8.1	6.5
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	33.5	33.6	20.1	10.6	11.3	6.5
			Probit Link Function			
$\widehat{\mathbf{t}}^{pr}$	19.5	19.3	17.9	7.1	7.0	5.9
$\widehat{\mathbf{t}}^{gd}$	20.3	20.0	18.4	7.1	7.1	5.9
$\widehat{\mathbf{t}}^{mc}$	19.6	19.4	18.0	7.1	7.1	5.9
$\widehat{\mathbf{t}}^{peM}$	21.9	21.6	19.6	7.4	8.0	6.4
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	33.1	33.0	19.6	10.5	11.2	6.4
			Complementary Log-Log Link Function			
$\widehat{\mathbf{t}}^{pr}$	19.7	19.5	18.1	7.2	7.1	6.0
$\widehat{\mathbf{t}}^{gd}$	20.4	20.1	18.5	7.1	7.1	5.9
$\widehat{\mathbf{t}}^{mc}$	19.7	19.4	18.0	7.1	7.1	5.9
$\widehat{\mathbf{t}}^{peM}$	22.2	22.0	19.8	7.3	8.0	6.4
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	33.5	33.6	19.8	10.5	11.2	6.4
			Cauchit Link Function			
$\widehat{\mathbf{t}}^{pr}$	19.0	18.8	17.4	7.0	6.9	6.0
$\widehat{\mathbf{t}}^{gd}$	20.3	20.1	18.4	7.1	7.1	5.9
$\widehat{\mathbf{t}}^{mc}$	19.7	19.5	18.0	7.1	7.1	5.9
$\widehat{\mathbf{t}}^{peM}$	21.7	21.9	19.9	7.3	7.9	6.4
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	33.1	33.5	19.9	10.5	11.1	6.4

C.7.1.3 Simulation Coefficient of Variation of Synthetic Response

Table C.3: Simulation Coefficient of Variation for Point Estimators of Synthetic Response. Estimates have been multiplied by $1,000,000$.

	Small Samples			Large Samples		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
			No Link Function			
$\widehat{\mathbf{t}}^{\pi}$	41.3	40.9	31.3	14.0	14.1	10.5
			Identity Link Function			
$\widehat{\mathbf{t}}^{gd}$	27.0	26.6	25.4	10.5	10.3	8.8
$\widehat{\mathbf{t}}^{peM}$	28.3	28.1	26.4	10.6	10.8	9.1
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	34.3	34.6	26.4	11.6	12.1	9.1
			Probit Link Function			
$\widehat{\mathbf{t}}^{pr}$	26.5	26.2	25.1	10.6	10.4	8.9
$\widehat{\mathbf{t}}^{gd}$	29.3	29.2	27.5	10.4	10.2	8.8
$\widehat{\mathbf{t}}^{mc}$	26.4	26.1	25.0	10.4	10.2	8.8
$\widehat{\mathbf{t}}^{peM}$	27.9	27.6	26.1	10.6	10.6	9.1
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	34.0	34.0	26.1	11.5	12.0	9.1
			Complementary Log-Log Link Function			
$\widehat{\mathbf{t}}^{pr}$	26.7	26.4	25.4	10.6	10.4	9.1
$\widehat{\mathbf{t}}^{gd}$	29.5	29.2	27.6	10.4	10.2	8.8
$\widehat{\mathbf{t}}^{mc}$	26.5	26.2	25.0	10.4	10.2	8.8
$\widehat{\mathbf{t}}^{peM}$	28.5	27.9	26.4	10.5	10.6	9.1
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	35.0	34.5	26.4	11.5	11.9	9.1
			Cauchit Link Function			
$\widehat{\mathbf{t}}^{pr}$	25.1	24.8	23.2	9.8	9.6	8.4
$\widehat{\mathbf{t}}^{gd}$	30.0	29.8	27.8	10.5	10.3	8.8
$\widehat{\mathbf{t}}^{mc}$	26.4	26.1	25.1	10.4	10.2	8.8
$\widehat{\mathbf{t}}^{peM}$	27.7	27.6	26.1	10.5	10.7	9.1
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	34.0	34.0	26.1	11.5	11.9	9.1

C.7.2 Graphs for Point Estimators

The following six plots show the relative bias and coefficient of variation for the point estimators of the three response variables in the small and large samples. We first show estimates of the count variable in the samples where we only selected 5 clusters. Then we show similar results for the samples with 35 clusters. Following that, we show results for estimates of the the binary response variable in small and large samples. The last two graphs are for estimates of the synthetic response variable in small and large samples.

The Horvitz-Thompson estimator is labeled HT while the generalized difference estimator is labeled GD. With the identity link and the sample designs we employed, GD is equivalent to the projective estimator (PR), the GREG estimator, and the model-calibrated (MC) estimator. For the other links. the projective estimator, the generalized difference estimator, and the model-calibrated estimators are different. The graphs also show the performance of the two model-calibrated maximum pseudoempirical likelihood estimators, PE.M and PE.M.HAT.

Following the plots are six tables showing numeric values for all estimates in the plots that follow.

C.7.2.1 Point Estimators of Count Response in Small Samples

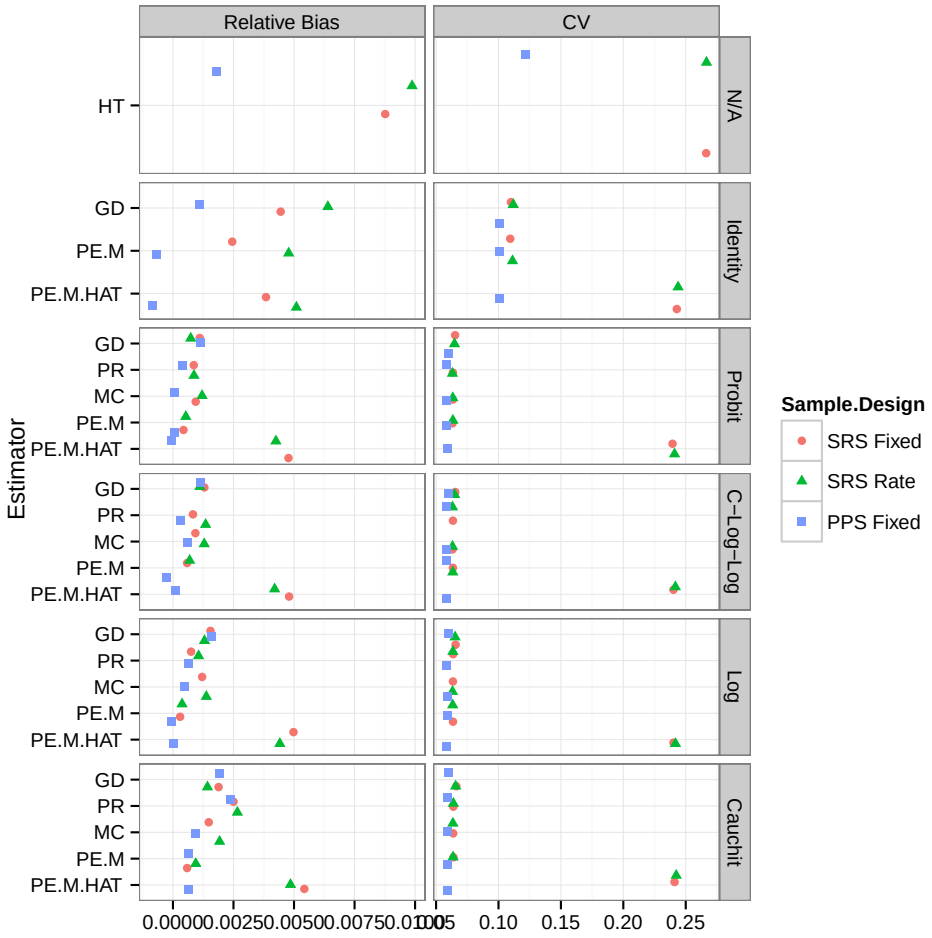

Figure C.1: Plot of Relative Bias and Coefficient of Variation for all estimators of Total Count in small samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. For numeric values of the points in this plot see Table C.4.

C.7.2.2 Point Estimators of Count Response in Large Samples

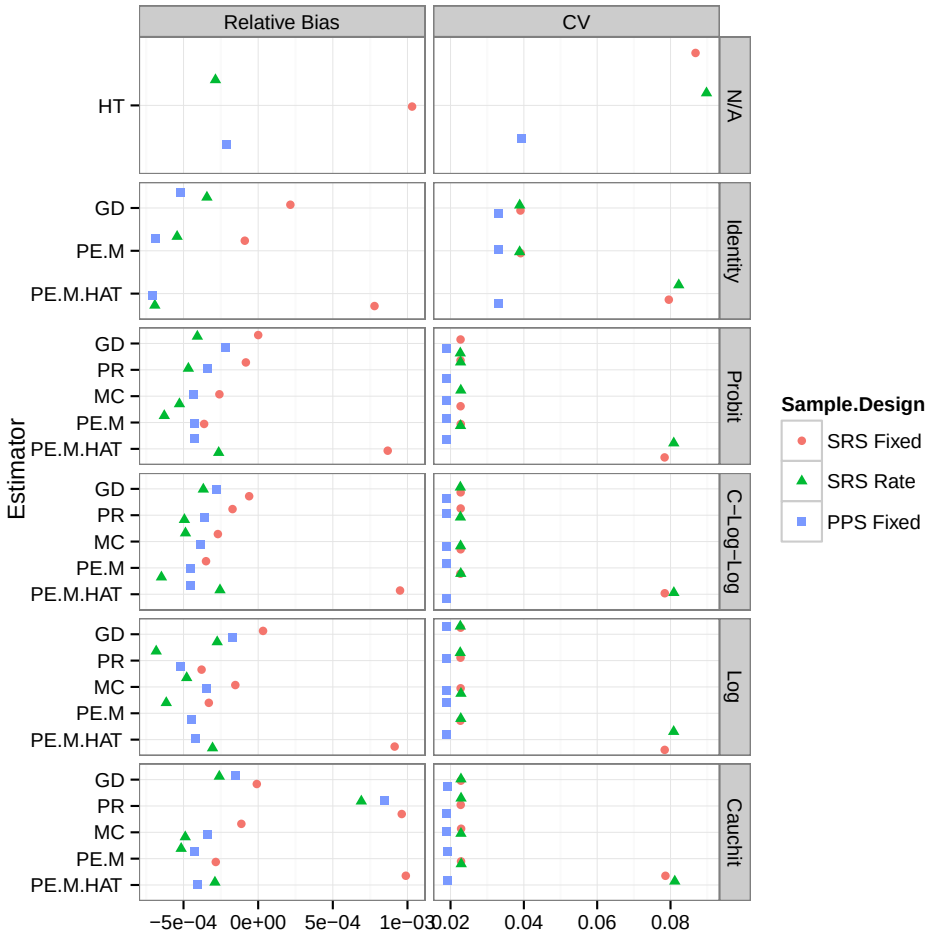

Figure C.2: Plot of Relative Bias and Coefficient of Variation for all estimators of Total Count in large samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. For numeric values of the points in this plot see Table C.5.

C.7.2.3 Point Estimators of Binary Response in Small Samples

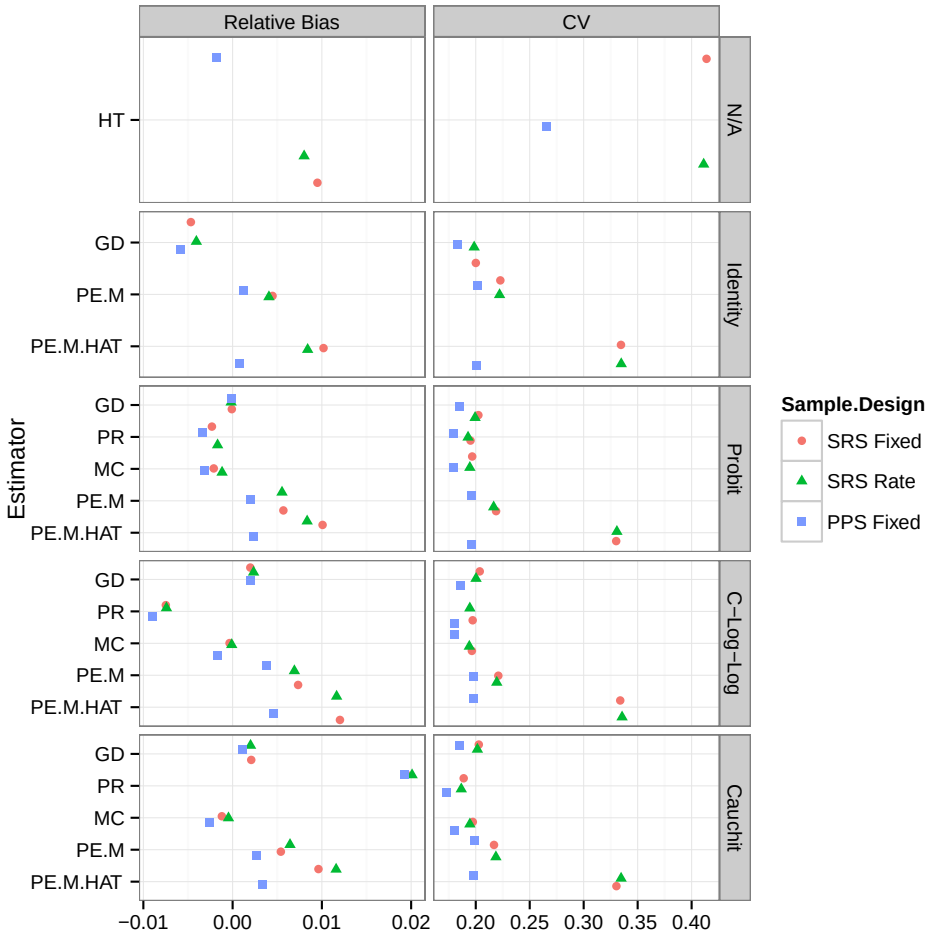

Figure C.3: Plot of Relative Bias and Coefficient of Variation for all estimators of total binary response in small samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. For numeric values of the points in this plot see Table C.6.

C.7.2.4 Point Estimators of Binary Response in Large Samples

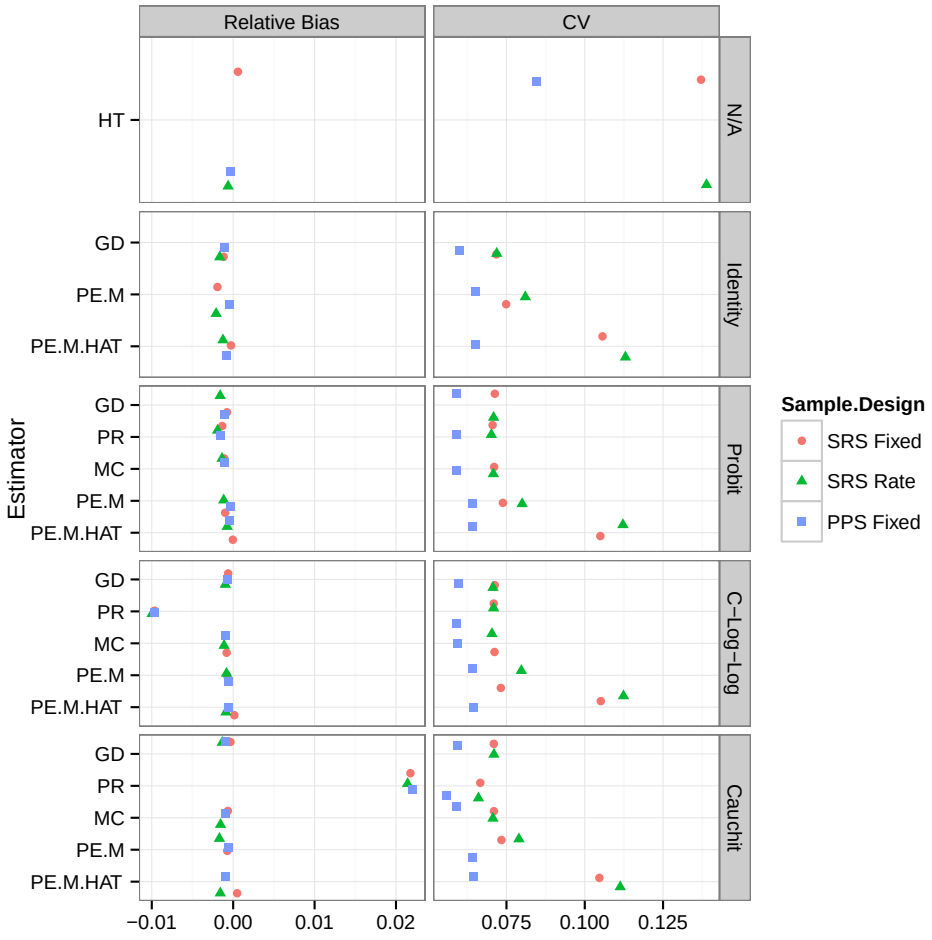

Figure C.4: Plot of Relative Bias and Coefficient of Variation for all estimators of total binary response in large samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. For numeric values of the points in this plot see Table C.7.

C.7.2.5 Point Estimators of Synthetic Response in Small Samples

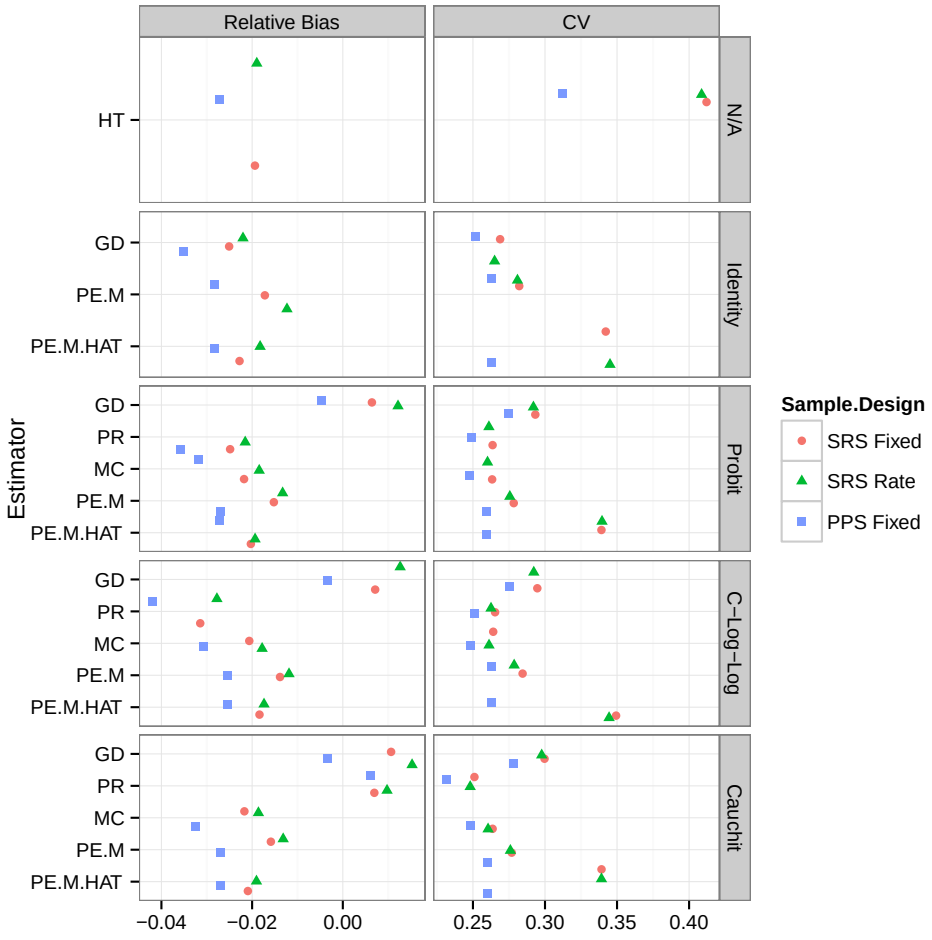

Figure C.5: Plot of Relative Bias and Coefficient of Variation for all estimators of total synthetic response in small samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. For numeric values of the points in this plot see Table C.8.

C.7.2.6 Point Estimators of Synthetic Response in Large Samples

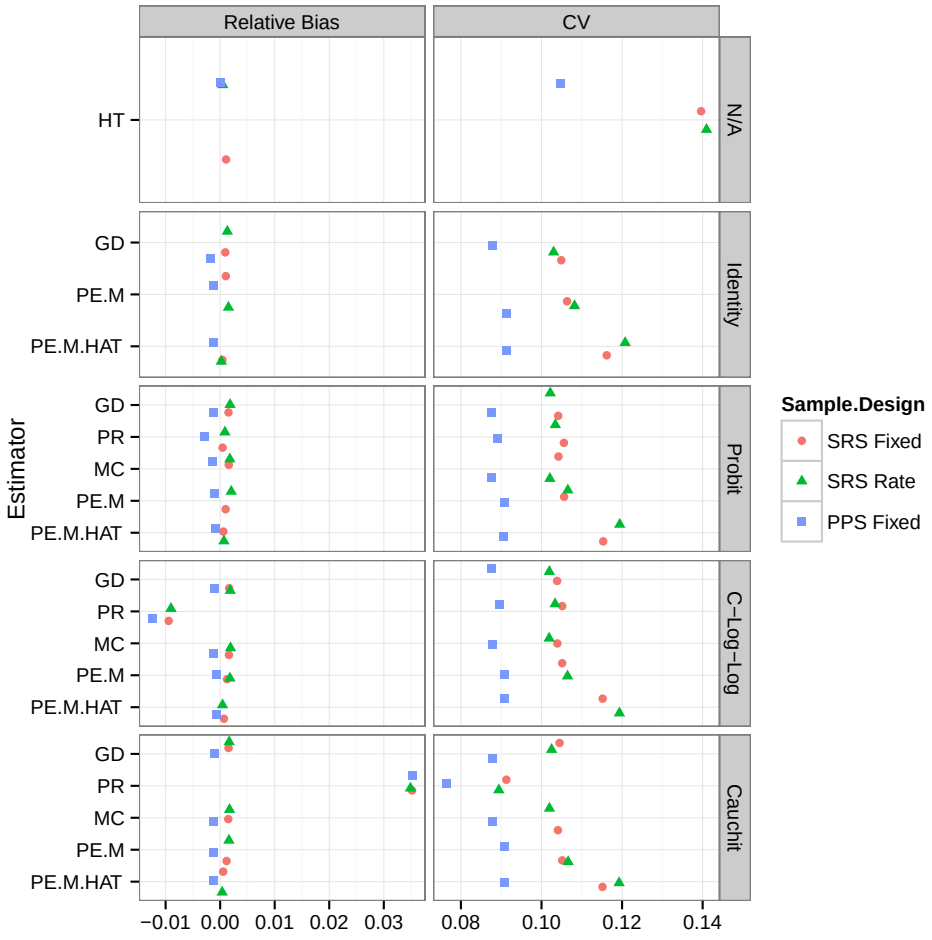

Figure C.6: Plot of Relative Bias and Coefficient of Variation for all estimators of total synthetic response in large samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other. For numeric values of the points in this plot see Table C.9.

C.7.3 Tables for Point Estimators

C.7.3.1 Point Estimators of Count Response in Small Samples

Table C.4: Relative Bias and Coefficient of Variation for Point Estimators of Count Response in Small Samples

	Relative Bias			Coefficient of Variation		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
			No Link Function			
$\widehat{\mathrm{t}}^{\pi}$	0.009	0.010	0.002	0.266	0.267	0.122
			Identity Link Function			
$\widehat{\mathrm{t}}^{gd}$	0.005	0.006	0.001	0.110	0.112	0.101
$\widehat{\mathrm{t}}^{peM}$	0.002	0.005	-0.001	0.109	0.111	0.100
$\widehat{\mathrm{t}}^{pe\widehat{M}}$	0.004	0.005	-0.001	0.243	0.244	0.100
			Probit Link Function			
$\widehat{\mathrm{t}}^{pr}$	0.001	0.001	0.000	0.063	0.063	0.059
$\widehat{\mathrm{t}}^{gd}$	0.001	0.001	0.001	0.065	0.065	0.060
$\widehat{\mathrm{t}}^{mc}$	0.001	0.001	0.000	0.063	0.063	0.059
$\widehat{\mathrm{t}}^{peM}$	0.000	0.000	0.000	0.063	0.063	0.059
$\widehat{\mathrm{t}}^{pe\widehat{M}}$	0.005	0.004	0.000	0.240	0.241	0.059
			Complementary Log-Log Link Function			
$\widehat{\mathrm{t}}^{pr}$	0.001	0.001	0.001	0.063	0.063	0.059
$\widehat{\mathrm{t}}^{gd}$	0.001	0.001	0.001	0.065	0.065	0.060
$\widehat{\mathrm{t}}^{mc}$	0.001	0.001	0.000	0.063	0.063	0.059
$\widehat{\mathrm{t}}^{peM}$	0.000	0.001	0.000	0.063	0.063	0.059
$\widehat{\mathrm{t}}^{pe\widehat{M}}$	0.005	0.004	0.000	0.240	0.241	0.059
			Log Link Function			
$\widehat{\mathrm{t}}^{pr}$	0.001	0.001	0.000	0.064	0.063	0.059
$\widehat{\mathrm{t}}^{gd}$	0.001	0.001	0.001	0.066	0.065	0.060
$\widehat{\mathrm{t}}^{mc}$	0.001	0.001	0.001	0.063	0.063	0.059
$\widehat{\mathrm{t}}^{peM}$	0.001	0.001	0.000	0.063	0.063	0.059
$\widehat{\mathrm{t}}^{pe\widehat{M}}$	0.005	0.004	0.000	0.240	0.241	0.059
			Cauchit Link Function			
$\widehat{\mathrm{t}}^{pr}$	0.003	0.003	0.002	0.064	0.064	0.059
$\widehat{\mathrm{t}}^{gd}$	0.002	0.002	0.002	0.066	0.066	0.060
$\widehat{\mathrm{t}}^{mc}$	0.001	0.002	0.001	0.064	0.064	0.059
$\widehat{\mathrm{t}}^{peM}$	0.001	0.001	0.000	0.064	0.064	0.059
$\widehat{\mathrm{t}}^{pe\widehat{M}}$	0.005	0.005	0.000	0.241	0.242	0.059

C.7.3.2 Point Estimators of Count Response in Large Samples

Table C.5: Relative Bias and Coefficient of Variation for Point Estimators of Count Response in Large Samples

	Relative Bias			Coefficient of Variation		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
			No Link Function			
$\widehat{\mathrm{t}}^{\pi}$	0.001	0.000	0.000	0.087	0.090	0.039
			Identity Link Function			
$\widehat{\mathrm{t}}^{gd}$	0.000	0.000	-0.001	0.039	0.039	0.033
$\widehat{\mathrm{t}}^{peM}$	0.000	-0.001	-0.001	0.039	0.039	0.033
$\widehat{\mathrm{t}}^{pe\widehat{M}}$	0.001	-0.001	-0.001	0.080	0.082	0.033
			Probit Link Function			
$\widehat{\mathrm{t}}^{pr}$	0.000	0.000	0.000	0.023	0.023	0.019
$\widehat{\mathrm{t}}^{gd}$	0.000	0.000	0.000	0.023	0.023	0.019
$\widehat{\mathrm{t}}^{mc}$	0.000	-0.001	0.000	0.023	0.023	0.019
$\widehat{\mathrm{t}}^{peM}$	0.000	-0.001	0.000	0.023	0.023	0.019
$\widehat{\mathrm{t}}^{pe\widehat{M}}$	0.001	0.000	0.000	0.078	0.081	0.019
			Complementary Log-Log Link Function			
$\widehat{\mathrm{t}}^{pr}$	0.000	0.000	0.000	0.023	0.023	0.019
$\widehat{\mathrm{t}}^{gd}$	0.000	0.000	0.000	0.023	0.023	0.019
$\widehat{\mathrm{t}}^{mc}$	0.000	0.000	0.000	0.023	0.023	0.019
$\widehat{\mathrm{t}}^{peM}$	0.000	-0.001	0.000	0.023	0.023	0.019
$\widehat{\mathrm{t}}^{pe\widehat{M}}$	0.001	0.000	0.000	0.078	0.081	0.019
			Log Link Function			
$\widehat{\mathrm{t}}^{pr}$	0.000	-0.001	-0.001	0.023	0.023	0.019
$\widehat{\mathrm{t}}^{gd}$	0.000	0.000	0.000	0.023	0.023	0.019
$\widehat{\mathrm{t}}^{mc}$	0.000	0.000	0.000	0.023	0.023	0.019
$\widehat{\mathrm{t}}^{peM}$	0.000	-0.001	0.000	0.023	0.023	0.019
$\widehat{\mathrm{t}}^{pe\widehat{M}}$	0.001	0.000	0.000	0.078	0.081	0.019
			Cauchit Link Function			
$\widehat{\mathrm{t}}^{pr}$	0.001	0.001	0.001	0.023	0.023	0.019
$\widehat{\mathrm{t}}^{gd}$	0.000	0.000	0.000	0.023	0.023	0.019
$\widehat{\mathrm{t}}^{mc}$	0.000	0.000	0.000	0.023	0.023	0.019
$\widehat{\mathrm{t}}^{peM}$	0.000	-0.001	0.000	0.023	0.023	0.019
$\widehat{\mathrm{t}}^{pe\widehat{M}}$	0.001	0.000	0.000	0.079	0.081	0.019

C.7.3.3 Point Estimators of Binary Response in Small Samples

Table C.6: Relative Bias and Coefficient of Variation for Point Estimators of Binary Response in Small Samples

	Relative Bias			Coefficient of Variation		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
			No Link Function			
$\widehat{\mathbf{t}}^{\pi}$	0.010	0.009	-0.001	0.414	0.411	0.265
			Identity Link Function			
$\widehat{\mathbf{t}}^{gd}$	-0.005	-0.005	-0.006	0.200	0.198	0.183
$\widehat{\mathbf{t}}^{peM}$	0.005	0.004	0.001	0.223	0.222	0.201
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	0.010	0.008	0.001	0.335	0.335	0.201
			Probit Link Function			
$\widehat{\mathbf{t}}^{pr}$	-0.002	-0.002	-0.004	0.195	0.193	0.179
$\widehat{\mathbf{t}}^{gd}$	-0.001	0.000	0.000	0.203	0.200	0.184
$\widehat{\mathbf{t}}^{mc}$	-0.002	-0.001	-0.003	0.196	0.194	0.180
$\widehat{\mathbf{t}}^{peM}$	0.005	0.005	0.002	0.219	0.216	0.196
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	0.010	0.009	0.002	0.331	0.330	0.196
			Complementary Log-Log Link Function			
$\widehat{\mathbf{t}}^{pr}$	-0.007	-0.007	-0.009	0.197	0.195	0.181
$\widehat{\mathbf{t}}^{gd}$	0.002	0.002	0.002	0.204	0.201	0.185
$\widehat{\mathbf{t}}^{mc}$	0.000	0.000	-0.001	0.197	0.194	0.180
$\widehat{\mathbf{t}}^{peM}$	0.007	0.007	0.004	0.221	0.220	0.198
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	0.012	0.012	0.004	0.334	0.336	0.198
			Cauchit Link Function			
$\widehat{\mathbf{t}}^{pr}$	0.019	0.021	0.019	0.189	0.187	0.173
$\widehat{\mathbf{t}}^{gd}$	0.002	0.002	0.001	0.203	0.201	0.184
$\widehat{\mathbf{t}}^{mc}$	-0.001	-0.001	-0.002	0.197	0.195	0.180
$\widehat{\mathbf{t}}^{peM}$	0.005	0.007	0.003	0.216	0.219	0.199
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	0.010	0.011	0.003	0.330	0.335	0.199

C.7.3.4 Point Estimators of Binary Response in Large Samples

Table C.7: Relative Bias and Coefficient of Variation for Point Estimators of Binary Response in Large Samples

	Relative Bias			Coefficient of Variation		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
			No Link Function			
$\widehat{\mathbf{t}}^{\pi}$	0.001	-0.001	0.000	0.137	0.139	0.084
			Identity Link Function			
$\widehat{\mathbf{t}}^{gd}$	-0.001	-0.002	-0.001	0.072	0.072	0.060
$\widehat{\mathbf{t}}^{peM}$	-0.002	-0.002	-0.001	0.075	0.081	0.065
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	0.000	-0.001	-0.001	0.106	0.113	0.065
			Probit Link Function			
$\widehat{\mathbf{t}}^{pr}$	-0.001	-0.002	-0.001	0.071	0.070	0.059
$\widehat{\mathbf{t}}^{gd}$	-0.001	-0.001	-0.001	0.071	0.071	0.059
$\widehat{\mathbf{t}}^{mc}$	-0.001	-0.001	-0.001	0.071	0.071	0.059
$\widehat{\mathbf{t}}^{peM}$	-0.001	-0.001	-0.001	0.074	0.080	0.064
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	0.000	-0.001	-0.001	0.105	0.112	0.064
			Complementary Log-Log Link Function			
$\widehat{\mathbf{t}}^{pr}$	-0.010	-0.010	-0.010	0.071	0.071	0.059
$\widehat{\mathbf{t}}^{gd}$	-0.001	-0.001	-0.001	0.071	0.071	0.059
$\widehat{\mathbf{t}}^{mc}$	-0.001	-0.001	-0.001	0.071	0.071	0.059
$\widehat{\mathbf{t}}^{peM}$	-0.001	-0.001	-0.001	0.073	0.080	0.064
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	0.000	-0.001	-0.001	0.105	0.112	0.064
			Cauchit Link Function			
$\widehat{\mathbf{t}}^{pr}$	0.022	0.021	0.022	0.067	0.066	0.056
$\widehat{\mathbf{t}}^{gd}$	0.000	-0.001	-0.001	0.071	0.071	0.059
$\widehat{\mathbf{t}}^{mc}$	-0.001	-0.001	-0.001	0.071	0.071	0.059
$\widehat{\mathbf{t}}^{peM}$	-0.001	-0.002	-0.001	0.073	0.079	0.064
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	0.000	-0.001	-0.001	0.105	0.111	0.064

C.7.3.5 Point Estimators of Synthetic Response in Small Samples

Table C.8: Relative Bias and Coefficient of Variation for Point Estimators of Synthetic Response in Small Samples

	Relative Bias			Coefficient of Variation		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
			No Link Function			
$\widehat{\mathbf{t}}^{\pi}$	-0.019	-0.019	-0.027	0.412	0.408	0.312
			Identity Link Function			
$\widehat{\mathbf{t}}^{gd}$	-0.025	-0.022	-0.035	0.269	0.265	0.252
$\widehat{\mathbf{t}}^{peM}$	-0.017	-0.012	-0.028	0.282	0.281	0.263
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	-0.023	-0.018	-0.028	0.342	0.345	0.263
			Probit Link Function			
$\widehat{\mathbf{t}}^{pr}$	-0.025	-0.022	-0.036	0.264	0.261	0.249
$\widehat{\mathbf{t}}^{gd}$	0.007	0.012	-0.005	0.293	0.292	0.275
$\widehat{\mathbf{t}}^{mc}$	-0.022	-0.019	-0.032	0.263	0.260	0.248
$\widehat{\mathbf{t}}^{peM}$	-0.015	-0.013	-0.027	0.278	0.276	0.260
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	-0.020	-0.019	-0.027	0.339	0.340	0.260
			Complementary Log-Log Link Function			
$\widehat{\mathbf{t}}^{pr}$	-0.031	-0.028	-0.042	0.265	0.263	0.251
$\widehat{\mathbf{t}}^{gd}$	0.007	0.013	-0.004	0.295	0.292	0.276
$\widehat{\mathbf{t}}^{mc}$	-0.021	-0.018	-0.031	0.264	0.261	0.249
$\widehat{\mathbf{t}}^{peM}$	-0.014	-0.012	-0.025	0.284	0.279	0.263
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	-0.018	-0.017	-0.025	0.349	0.344	0.263
			Cauchit Link Function			
$\widehat{\mathbf{t}}^{pr}$	0.007	0.010	0.006	0.251	0.248	0.232
$\widehat{\mathbf{t}}^{gd}$	0.011	0.015	-0.003	0.300	0.298	0.278
$\widehat{\mathbf{t}}^{mc}$	-0.022	-0.018	-0.032	0.263	0.261	0.249
$\widehat{\mathbf{t}}^{peM}$	-0.016	-0.013	-0.027	0.277	0.276	0.260
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	-0.021	-0.019	-0.027	0.339	0.339	0.260

C.7.3.6 Point Estimators of Synthetic Response in Large Samples

Table C.9: Relative Bias and Coefficient of Variation for Point Estimators of Synthetic Response in Large Samples

	Relative Bias			Coefficient of Variation		
Estimator	Fixed SRS	Rate SRS	Fixed PPS	Fixed SRS	Rate SRS	Fixed PPS
			No Link Function			
$\widehat{\mathbf{t}}^{\pi}$	0.001	0.001	0.000	0.140	0.141	0.105
			Identity Link Function			
$\widehat{\mathbf{t}}^{gd}$	0.001	0.001	-0.002	0.105	0.103	0.088
$\widehat{\mathbf{t}}^{peM}$	0.001	0.002	-0.001	0.106	0.108	0.091
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	0.000	0.000	-0.001	0.116	0.121	0.091
			Probit Link Function			
$\widehat{\mathbf{t}}^{pr}$	0.000	0.001	-0.003	0.106	0.104	0.089
$\widehat{\mathbf{t}}^{gd}$	0.001	0.002	-0.001	0.104	0.102	0.088
$\widehat{\mathbf{t}}^{mc}$	0.001	0.002	-0.001	0.104	0.102	0.088
$\widehat{\mathbf{t}}^{peM}$	0.001	0.002	-0.001	0.106	0.106	0.091
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	0.001	0.001	-0.001	0.115	0.120	0.091
			Complementary Log-Log Link Function			
$\widehat{\mathbf{t}}^{pr}$	-0.009	-0.009	-0.013	0.105	0.103	0.090
$\widehat{\mathbf{t}}^{gd}$	0.002	0.002	-0.001	0.104	0.102	0.088
$\widehat{\mathbf{t}}^{mc}$	0.002	0.002	-0.001	0.104	0.102	0.088
$\widehat{\mathbf{t}}^{peM}$	0.001	0.002	-0.001	0.105	0.106	0.091
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	0.001	0.000	-0.001	0.115	0.119	0.091
			Cauchit Link Function			
$\widehat{\mathbf{t}}^{pr}$	0.035	0.035	0.035	0.091	0.089	0.076
$\widehat{\mathbf{t}}^{gd}$	0.002	0.002	-0.001	0.104	0.102	0.088
$\widehat{\mathbf{t}}^{mc}$	0.002	0.002	-0.001	0.104	0.102	0.088
$\widehat{\mathbf{t}}^{peM}$	0.001	0.002	-0.001	0.105	0.107	0.091
$\widehat{\mathbf{t}}^{pe\widehat{M}}$	0.001	0.000	-0.001	0.115	0.119	0.091

C.7.4 Graphs for Variance Estimators

The following six figures show the relative bias of the new variance estimators as well as the empirical confidence interval coverage obtained by using the point estimator and the estimated standard error. Variance estimators are shown for the GLM-assisted difference estimator (GD) and the model-calibrated (MC) estimator. Colors are used to distinguish the four variance estimators: v_{wr} (v.wr), v_e (v.ssw.e), v_g (v.ssw.g), and v_{Binder} (v.Binder). Shapes are used to distinguish between the three different sample designs.

Altogether, there are six plots. We first show variance estimators for the count variable in the samples where we only selected 5 counties. Then we show similar results for the samples with 35 counties. Following that, we show results for estimates of the binary response variable in small and large samples. The last two graphs are for estimates of the synthetic response variable in small and large samples.

Since these plots contain very little overplotting, tables do not follow the plots.

C.7.4.1 Variance Estimators of Count Response in Small Samples

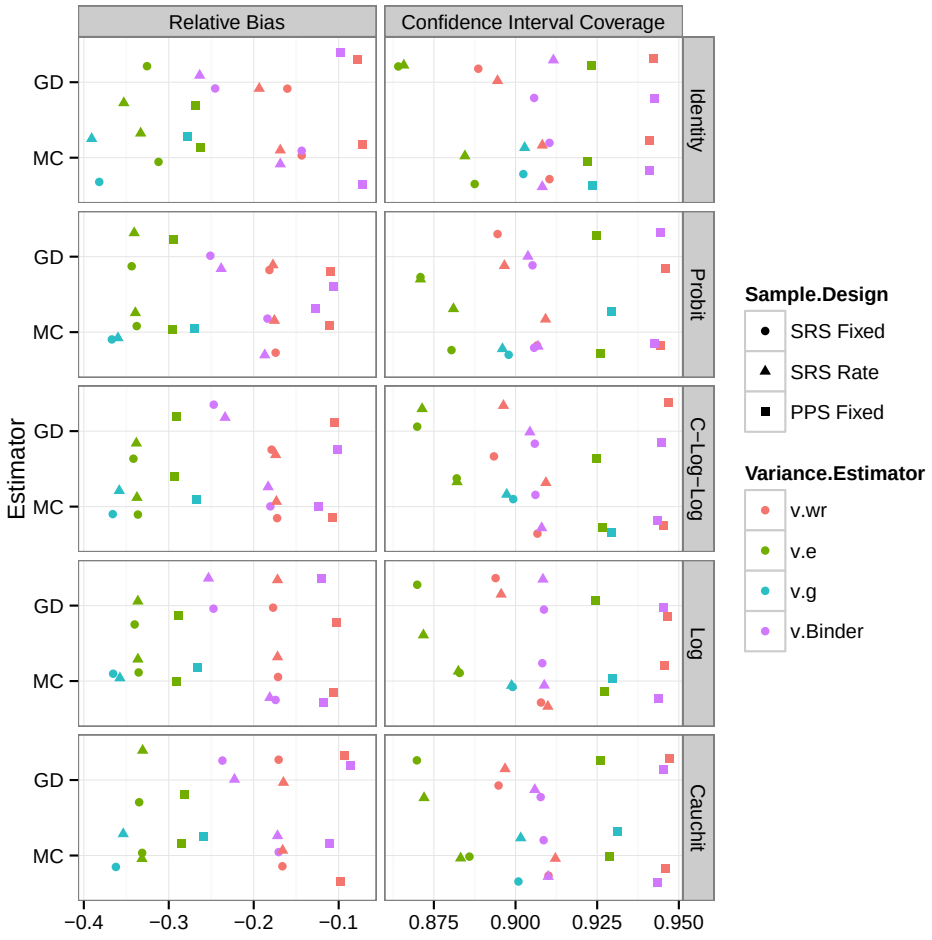

Figure C.7: Plot of Relative Bias and Confidence Interval Coverage of variance estimators for the GLM-assisted difference estimator of the count variable in small samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other.

C.7.4.2 Variance Estimators of Count Response in Large Samples

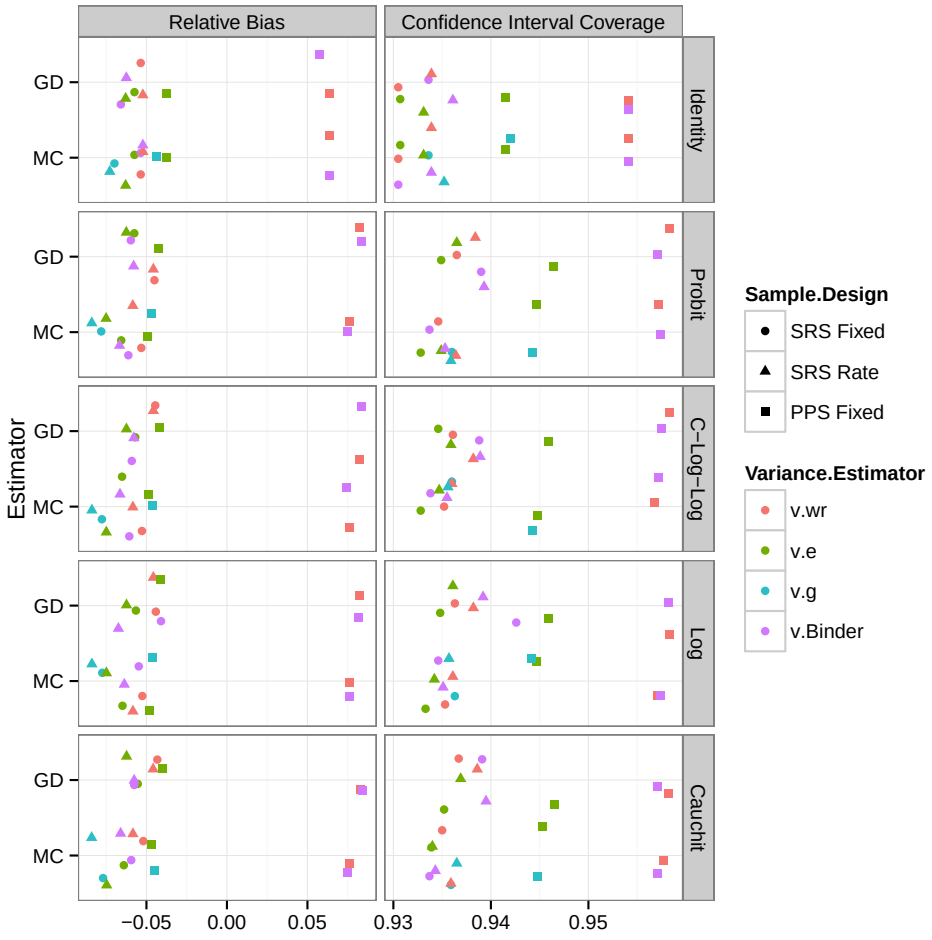

Figure C.8: Plot of Relative Bias and Confidence Interval Coverage of variance estimators for the GLM-assisted difference estimator of the count response in large samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other.

C.7.4.3 Variance Estimators of Binary Response in Small Samples

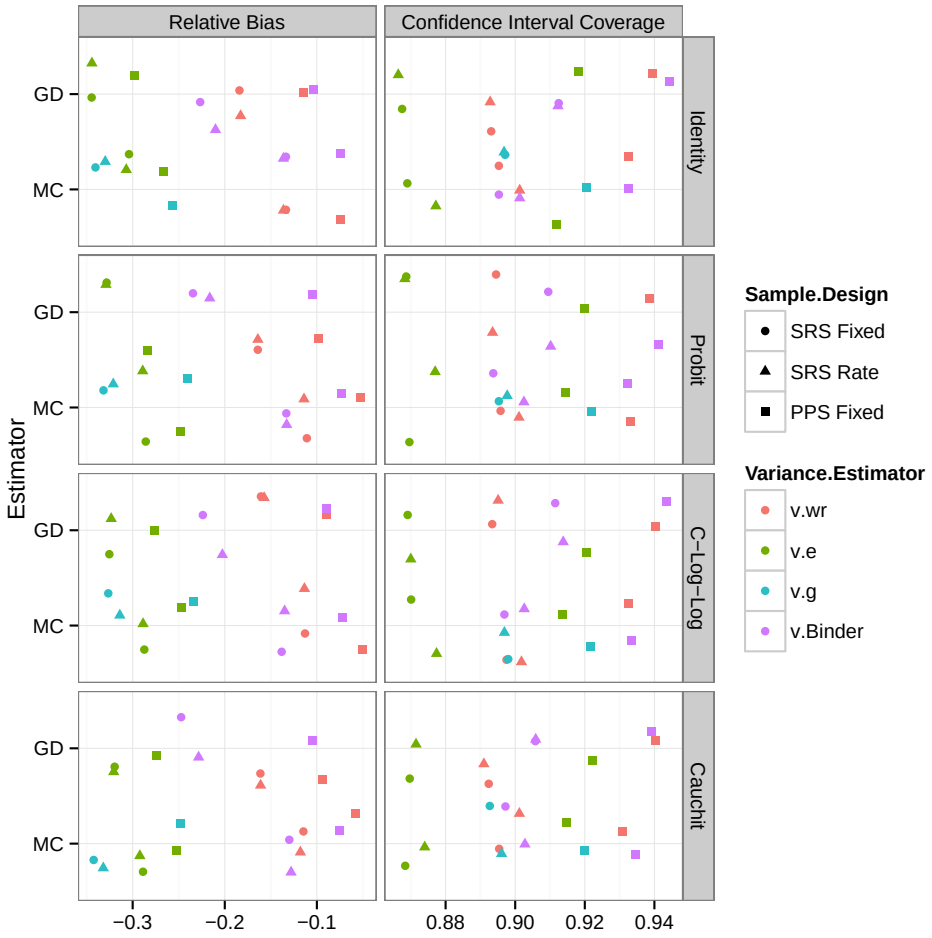

Figure C.9: Plot of Relative Bias and Confidence Interval Coverage of variance estimators for the GLM-assisted difference estimator of the binary response in small samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other.

C.7.4.4 Variance Estimators of Binary Response in Large Samples

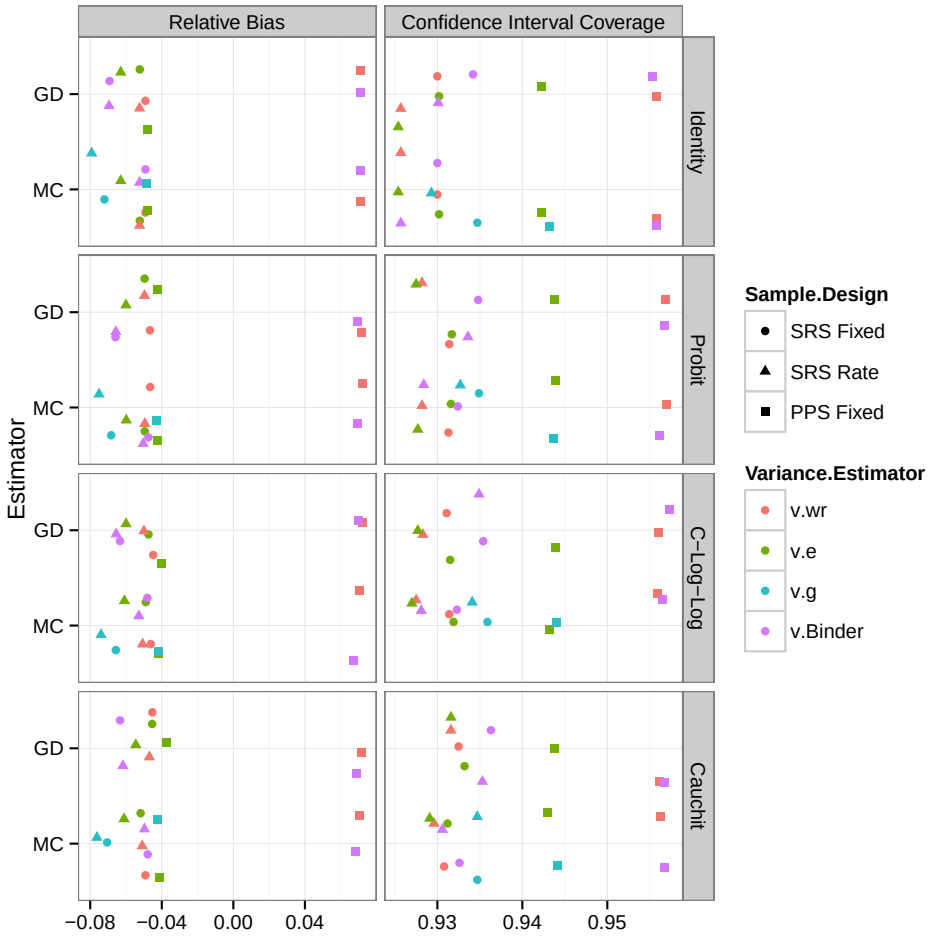

Figure C.10: Plot of Relative Bias and Confidence Interval Coverage of variance estimators for the GLM-assisted difference estimator of the binary response in large samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other.

C.7.4.5 Variance Estimators of Synthetic Response in Small Samples

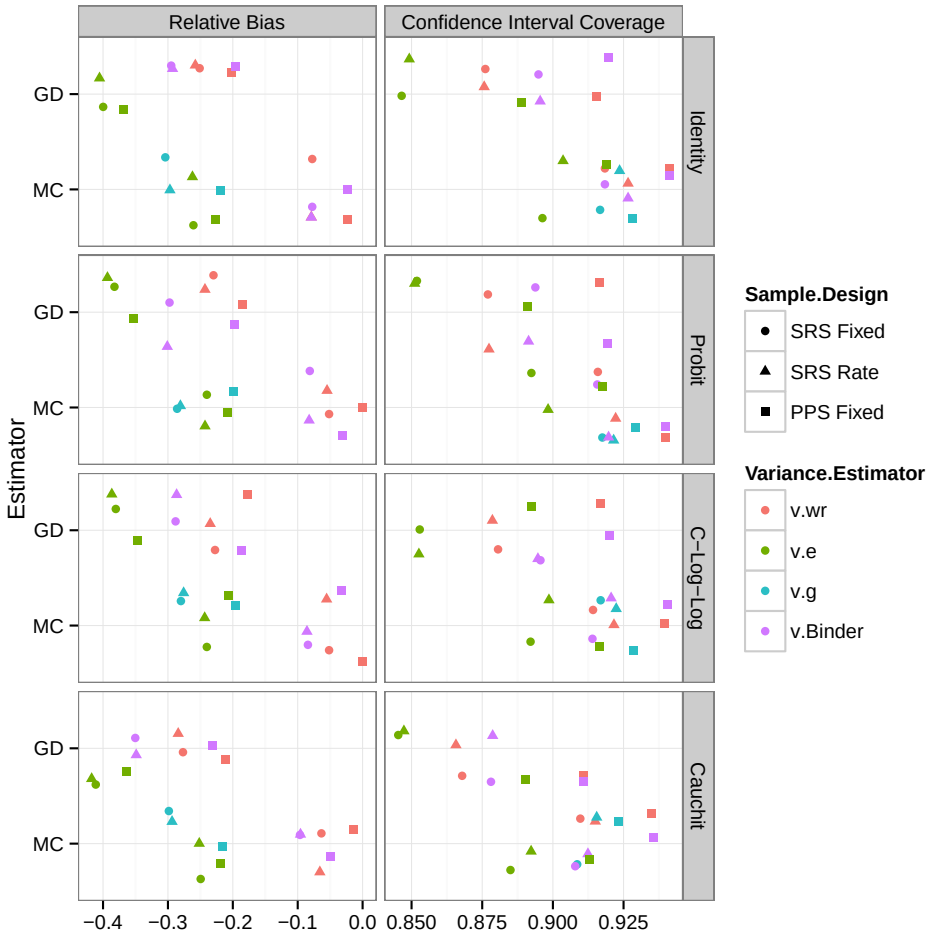

Figure C.11: Plot of Relative Bias and Confidence Interval Coverage of variance estimators for the GLM-assisted difference estimator of the synthetic variable in small samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other.

C.7.4.6 Variance Estimators of Synthetic Response in Large Samples

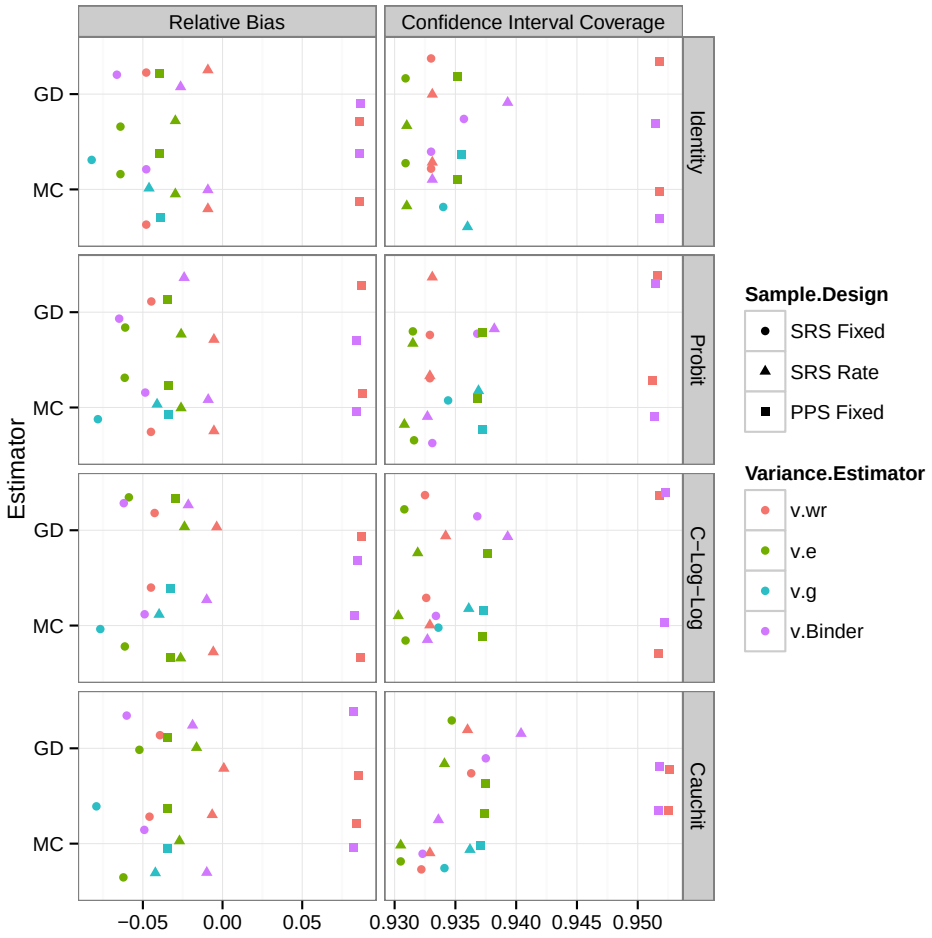

Figure C.12: Plot of Relative Bias and Confidence Interval Coverage of variance estimators for the GLM-assisted difference estimator of the synthetic variable in large samples. Points have been jittered along the vertical axis to prevent plotting several points on top of each other.

C.8 R Code

C.8.1 Generation of Synthetic Variable

```
# Save the seed to repeat the experiment
runif(1)
load(file = "C:\\Documents and Settings\\Tim\\My Documents\\Data\\seed.Rdata")
#set.seed(.Random.seed[24])
set.seed(seed)

# Pick Gamma
  s.gamma <- .9

# Generate e_0
  s.e.0.j <- NULL
  for(j in 1: length(unique(clus.id))) {
    s.e.0.j[j] <- rlogis(n = 1, location = 0, scale = 1)
  }
  s.e.0.2 <- cbind(unique(clus.id), s.e.0.j)

  s.e.0 <- merge(x = clus.id, y = s.e.0.2, by.x = "clus.id", by.y = "clus.id")

# Generate e_i and s.U.i
  s.e.i <- NULL
  s.U.i <- NULL
  for(j in 1: length(Y.Bin)) {
    s.e.i[j] <- rlogis(n = 1, location = 0, scale = 1)
    s.U.i[j] <- rbern(n = 1, prob = s.gamma)
  }


# Calculate Z
  s.Z.i <- s.U.i * s.e.0[,2] + (1 - s.U.i) * s.e.i


# Calculate theta_i: Linear Predictor
  data.POP <- data.frame(cbind(Y.Bin, X.Pop, clus.id))

  split.POP <- split(data.POP, clus.id)

  logit.glm <- NULL
  logit.out <- NULL
  E.1 <- function(data){
             logit.glm <- try(glm(data[,1] ~ as.matrix(data[,2:3]) -1, family = binomial(link = "logit")))
             logit.glm$linear.predictors
  }

  E.1(split.POP[[1]])


  Est.Eq <- lapply(1:length(unique(clus.id)),
           function(i, split.POP)
           E.1(split.POP[[i]]), split.POP = split.POP)

Est.Eq.Matrix <- t(sapply(X = Est.Eq, FUN = identity, simplify = T, USE.NAMES = T))

s.theta <- c(unsplit(Est.Eq.Matrix, clus.id))

# Calculate Y
  Y.Best <- as.matrix(as.vector(ifelse(s.Z.i < s.theta,1,0)))
```

C.8.2 Simulation Program

```
PGREG.sim <- function(X.Pop, Y.Pop, offset, clus.id, a, b, iterations, seed, smp, smp2)
{
  cat("Begin Intro", format(Sys.time(), "%X"), "\n")

  load(file = "C:\\Documents and Settings\\Tim\\My Documents\\Data\\seed.Rdata")
  set.seed(seed)

  Y.NR <- Y.Pop[, "Y.NR"]
  Y.Rate <- Y.Pop[, "Y.Rate"]
  Y.Bin <- Y.Pop[, "Y.Bin"]
  Y.Best <- Y.Pop[, "Y.Best"]

  X.1 <- X.Pop[, 1]
  X.2 <- X.Pop[, 2]

  Pop.1 <- cbind(X.Pop, Y.Pop, clus.id)

  # Get the population size
  M.1 <- nrow(Pop.1)

  # Get the number of columns in X and Y
  X.dim <- ncol(X.Pop)
  Y.dim <- ncol(Y.NR)

  # Create the measures of size
  mos.1 <- as.vector(by(Pop.1, Pop.1[,"clus.id"], nrow))

  # M.clus is the total number of clusters in the population
  M.clus <- length(unique(Pop.1[, "clus.id"]))

  # Create the first stage sampling probabilities
  pi.I.pps <- a * mos.1  / nrow(Pop.1)
  pi.I.srs <- rep(a / M.clus, M.clus)
  if(smp == "srs") pi.I <- pi.I.srs else pi.I <- pi.I.pps

  pi.II.fixed <- b / mos.1
  pi.II.rate <- (b * sum(M.clus)) / sum(mos.1)
  if(smp2 == "fixed") pi.II.all <- pi.II.fixed else pi.II.all <- pi.II.rate

  pi.k.all <- pi.I * pi.II.all

  # Recode the clusterid
  c.id <- c(1: M.clus)
  clus.conversion <- cbind(unique(Pop.1[, "clus.id"]), c.id, pi.I, pi.II.all, pi.k.all)
  X.clusid <- merge(x = Pop.1, y = clus.conversion, by.x = "clus.id", by.y = 1)

  w.n <- 1 / X.clusid[, "pi.k.all"]
  w.n.II <- 1 / X.clusid[, "pi.II.all"]
  ind <- X.clusid[, "clus.id"]

  ### GLM
  # Count
  Pop.glm.probit.count <- try(glm(cbind(Y.NR, Y.Resp) ~ X.Pop -1, family = binomial(link = "probit")))
  Pop.glm.probit.count.beta <- coefficients(Pop.glm.probit.count)
  b.dim <- length(Pop.glm.probit.count.beta)

  Pop.glm.cloglog.count <- try(glm(cbind(Y.NR, Y.Resp) ~ X.Pop -1, family = binomial(link = "cloglog")))
  Pop.glm.cloglog.count.beta <- coefficients(Pop.glm.cloglog.count)

  Pop.glm.poisson.count <- try(glm((Y.NR / offset)~ X.Pop -1, family = poisson(link = "log")))
  Pop.glm.poisson.count.beta <- coefficients(Pop.glm.poisson.count)

  Pop.glm.cauchit.count <- try(glm(cbind(Y.NR, Y.Resp) ~ X.Pop -1, family = binomial(link = "cauchit")))
  Pop.glm.cauchit.count.beta <- coefficients(Pop.glm.cauchit.count)

  Pop.glm.identity.count <- try(glm((Y.NR/offset) ~ X.Pop -1, family = gaussian(link = "identity")))
  Pop.glm.identity.count.beta <- coefficients(Pop.glm.identity.count)

  # Binary
  Pop.glm.probit.binary <- try(glm(Y.Bin ~ X.Pop -1, family = binomial(link = "probit")))
  Pop.glm.probit.binary.beta <- coefficients(Pop.glm.probit.binary)

  Pop.glm.cloglog.binary <- try(glm(Y.Bin ~ X.Pop -1, family = binomial(link = "cloglog")))
  Pop.glm.cloglog.binary.beta <- coefficients(Pop.glm.cloglog.binary)

  Pop.glm.cauchit.binary <- try(glm(Y.Bin ~ X.Pop -1, family = binomial(link = "cauchit")))
  Pop.glm.cauchit.binary.beta <- coefficients(Pop.glm.cauchit.binary)

  Pop.glm.identity.binary <- try(glm(Y.Bin ~ X.Pop -1, family = gaussian(link = "identity")))
  Pop.glm.identity.binary.beta <- coefficients(Pop.glm.identity.binary)

  # Best
  Pop.glm.probit.best <- try(glm(Y.Best ~ X.Pop -1, family = binomial(link = "probit")))
  Pop.glm.probit.best.beta <- coefficients(Pop.glm.probit.best)

  Pop.glm.cloglog.best <- try(glm(Y.Best ~ X.Pop -1, family = binomial(link = "cloglog")))
  Pop.glm.cloglog.best.beta <- coefficients(Pop.glm.cloglog.best)

  Pop.glm.cauchit.best <- try(glm(Y.Best ~ X.Pop -1, family = binomial(link = "cauchit")))
```

```
  Pop.glm.cauchit.best.beta <- coefficients(Pop.glm.cauchit.best)

  Pop.glm.identity.best <- try(glm(Y.Best ~ X.Pop -1, family = gaussian(link = "identity")))
  Pop.glm.identity.best.beta <- coefficients(Pop.glm.identity.best)

  Y.col <- 15

  # Create a list of cluster auxiliaries
  X.clus <- split(X.clusid, clus.id)

  t.HT <- matrix(0, nrow = iterations, ncol = 3)
  t.PROJ <- matrix(0, nrow = iterations, ncol = Y.col)
  t.GREG <- matrix(0, nrow = iterations, ncol = 5)
  t.GGREG <- matrix(0, nrow = iterations, ncol = Y.col)
  t.MCAL <- matrix(0, nrow = iterations, ncol = Y.col)
  t.PEMLE.N <- matrix(0, nrow = iterations, ncol = Y.col)
  t.PEMLE.N.hat <- matrix(0, nrow = iterations, ncol = Y.col)

  v.GGREG.wr <- matrix(0, nrow = iterations, ncol = Y.col)
  v.GGREG.ssw <- matrix(0, nrow = iterations, ncol = Y.col)
  v.GGREG.pml <- matrix(0, nrow = iterations, ncol = Y.col)

  v.MCAL.wr <- matrix(0, nrow = iterations, ncol = Y.col)
  v.MCAL.ssw <- matrix(0, nrow = iterations, ncol = Y.col)
  v.MCAL.ssw.g <- matrix(0, nrow = iterations, ncol = Y.col)
  v.MCAL <- matrix(0, nrow = iterations, ncol = Y.col)

  cat("End Intro", format(Sys.time(), "%X"), "\n")

  j <- 1
  j.master <- 0
  error.1  <- NULL
  error.2  <- NULL
  error.3  <- NULL
  error.4  <- matrix(0, nrow = iterations * 1.4 + 10, ncol = Y.col)
  error.5  <- matrix(0, nrow = iterations * 1.4 + 10, ncol = Y.col)
  error.6 <- matrix(0, nrow = iterations * 1.4 + 10, ncol = Y.col)

#  for(j in 1: iterations)
  while(j < iterations +1)
  {
    j.master <- j.master +1

    ## Sampling begins here
    # Select the first stage sample without replacement
    samp.clus <- UPrandomsystematic.alt(clus.conversion[,"pi.I"])
    X.clus.sample <- X.clus[c.id[samp.clus >= 1]]

    # Select the second stage sample
    if(smp2 == "rate") X.sample.f <- lapply(X.clus.sample, UPoi) else X.sample.f <- lapply(X.clus.sample, UPrandomsystematic.alt2)

    # Vector of sample clusters including zero clusters
    if(smp2 == "rate") a.f <- sapply(X.sample.f, length) else a.f <- sapply(X.sample.f, nrow)

    # Number of sample clusters including zeros
    n.clus.samp.z <- length(a.f)

    # Vector of sample clusters excluding zero clusters
    a.g <- subset(a.f, a.f > 0)

    # Number of sample clusters excluding zeros
    n.clus.samp <- length(a.g)

    # Vector of nonzero clusters
    a.1 <- ifelse(a.f > 0, 1, 0)
    a.n <- c(1:n.clus.samp.z)
    a.n1 <- a.1 * a.n
    a.i <- subset(a.n1, a.n1 > 0)

    # Create Unclustered data
    # Note that the sample elements can be repeated
    # Note: There may be some duplicates
if(smp2 == "fixed") {
  Fixed.id <- lapply(a.i,
            function(i, X.sample.f)
                    c(as.numeric(rownames(X.sample.f[[i]]))),
                    X.sample.f= X.sample.f)
  sample.id <- c(sapply(X = Fixed.id, FUN = rbind, simplify = T, USE.NAMES = T))
}
else {
  sample.id <- as.numeric(unique(as.vector(do.call(c, (sapply(X = X.sample.f, FUN = rownames, simplify = F, USE.NAMES = T))))))
}
    # Matrix of sample units in sample clusters
    X.sample <- X.clusid[sample.id, ]

    # List of nonzero sample cluster names
    b.f <- as.numeric(names(a.g))

    # Cluster probabilities of selection for nonzero sample clusters
    samp.clus.pi <- pi.I[b.f]

    # Cluster probabilities of selection for nonzero sample clusters repeated for each category
```

```
  samp.clus.pi.cat <- samp.clus.pi %x% matrix(rep(1, Y.col), ncol = Y.col)

  ## Estimation begins here
  # Population Totals
  T.x <- colSums(X.Pop)

  # Sample X and Y values
  # Note: There may be some duplicates when the first stage is selected with replacement
  X.samp <- X.Pop[as.numeric(sample.id),]
  X.1.samp <- X.1[as.numeric(sample.id)]
  X.2.samp <- X.2[as.numeric(sample.id)]
  X.3.samp <- X.3[as.numeric(sample.id)]
  X.4.samp <- X.4[as.numeric(sample.id)]
  offset.samp <- offset[as.numeric(sample.id)]
  Y.samp <- Y.Pop[as.numeric(sample.id),]
  Y.NR.samp <-    Y.samp[, "Y.NR"]
  Y.Resp.samp <-  Y.samp[, "Y.Resp"]
  Y.Rate.samp <-  Y.samp[, "Y.Rate"]
  Y.Bin.samp <-   Y.samp[, "Y.Bin"]
  Y.Best.samp <-  Y.samp[, "Y.Best"]
  w.k <- w.n[as.numeric(sample.id)]
  w.k.2 <- w.n.II[as.numeric(sample.id)]

   ind.1 <- factor(ind[as.numeric(sample.id)])

  samp.pi.I <- subset(pi.I, samp.clus == 1)
  samp.pi.I.list <- split(samp.pi.I, f = seq(1:length(samp.pi.I)))

  # Cluster level weight for nonzero clusters
  w.k.clus <- split(w.k, ind.1)

  # Number of units in sample
  n.samp <- length(w.k)

  # Skip if there are data problems
  error.1[j.master] <- ifelse(min(colSums(Y.samp)) ==0, 1, 0)
  error.2[j.master] <- ifelse(qr(X.samp)$rank < ncol(X.samp), 1, 0)
  if(min(colSums(Y.samp)) ==0) next
  if(qr(X.samp)$rank < ncol(X.samp)) next

  ## Pi Estimator
  t.NR.pi <- sum((w.k) * Y.NR.samp)
  t.Bin.pi <- sum((w.k) * Y.Bin.samp)
  t.Best.pi <- sum((w.k) * Y.Best.samp)
  t.HT[j, ] <- cbind(t.NR.pi, t.Bin.pi, t.Best.pi)

  ## GREG
  lm.NR   <- lm((Y.NR.samp) ~ X.1.samp + X.2.samp - 1, offset = offset.samp, weights = w.k)
  lm.NR.alt <- lm(Y.NR.samp ~ X.1.samp + X.2.samp - 1, weights = w.k)
  lm.NR.alt.2 <- lm((Y.NR.samp / offset.samp) ~ X.1.samp + X.2.samp - 1, weights = w.k)
  lm.Bin  <- lm(Y.Bin.samp ~ X.1.samp + X.2.samp - 1, weights = w.k)
  lm.Best <- lm(Y.Best.samp ~ X.1.samp + X.2.samp - 1, weights = w.k)

  Samp.fit.lm.NR     <- lm.NR$fitted.values
  Samp.fit.lm.NR.alt <- lm.NR.alt$fitted.values
  Samp.fit.lm.NR.alt.2 <- offset.samp * lm.NR.alt.2$fitted.values
  Samp.fit.lm.Bin    <- lm.Bin$fitted.values
  Samp.fit.lm.Best   <- lm.Best$fitted.values

  Samp.lm.NR.beta    <- lm.NR$coefficients
  Samp.lm.NR.beta.alt    <- lm.NR.alt$coefficients
  Samp.lm.NR.beta.alt.2  <- lm.NR.alt.2$coefficients
  Samp.lm.Bin.beta   <- lm.Bin$coefficients
  Samp.lm.Best.beta  <- lm.Best$coefficients

  Pop.fit.lm.NR      <- predict(lm.NR, newdata =data.frame(X.1.samp = X.1, X.2.samp = X.2, offset.samp = offset), type = "response")
  Pop.fit.lm.NR.alt  <- predict(lm.NR.alt, newdata =data.frame(X.1.samp = X.1, X.2.samp = X.2), type = "response")
  Pop.fit.lm.NR.alt.2  <- offset * predict(lm.NR.alt.2, newdata =data.frame(X.1.samp = X.1, X.2.samp = X.2), type = "response")
  Pop.fit.lm.Bin     <- predict(lm.Bin, newdata =data.frame(X.1.samp = X.1, X.2.samp = X.2), type = "response")
  Pop.fit.lm.Best    <- predict(lm.Best, newdata =data.frame(X.1.samp = X.1, X.2.samp = X.2), type = "response")

  t.NR.GREG      <- sum(Pop.fit.lm.NR, na.rm = TRUE) +  sum(w.k * (Y.NR.samp - Samp.fit.lm.NR), na.rm = TRUE)
  t.NR.GREG.alt  <- sum(Pop.fit.lm.NR.alt, na.rm = TRUE) +  sum(w.k * (Y.NR.samp - Samp.fit.lm.NR.alt), na.rm = TRUE)
  t.NR.GREG.alt.2  <- sum(Pop.fit.lm.NR.alt.2, na.rm = TRUE) +  sum(w.k * (Y.NR.samp - Samp.fit.lm.NR.alt.2), na.rm = TRUE)
  t.Bin.GREG     <- sum(Pop.fit.lm.Bin, na.rm = TRUE) +  sum(w.k * (Y.Bin.samp - Samp.fit.lm.Bin), na.rm = TRUE)
  t.Best.GREG    <- sum(Pop.fit.lm.Best, na.rm = TRUE) +  sum(w.k * (Y.Best.samp - Samp.fit.lm.Best), na.rm = TRUE)

  t.GREG[j, ] <- cbind(t.NR.GREG, t.NR.GREG.alt, t.NR.GREG.alt.2, t.Bin.GREG, t.Best.GREG)

##### Sample prediction
# Count
samp.glm.probit.count <-
  try(glm(cbind(Y.NR.samp, Y.Resp.samp) ~ X.1.samp + X.2.samp -1, family = quasibinomial(link = "probit"),
      weights = w.k, start = Pop.glm.probit.count.beta))
samp.glm.probit.count.beta <- coefficients(samp.glm.probit.count)
samp.fit.probit.count <- offset.samp * fitted.values(samp.glm.probit.count)
Pop.fit.probit.count <-
  offset * predict(samp.glm.probit.count, newdata =data.frame(X.1.samp = X.1, X.2.samp = X.2), type = "response")

samp.glm.cloglog.count <-
  try(glm(cbind(Y.NR.samp, Y.Resp.samp) ~ X.1.samp + X.2.samp -1, family = quasibinomial(link = "cloglog"),
  weights = w.k, start = Pop.glm.cloglog.count.beta))
```

```
samp.glm.cloglog.count.beta <- coefficients(samp.glm.cloglog.count)
samp.fit.cloglog.count <- offset.samp * fitted.values(samp.glm.cloglog.count)
Pop.fit.cloglog.count <-
  offset * predict(samp.glm.cloglog.count, newdata =data.frame(X.1.samp = X.1, X.2.samp = X.2), type = "response")

samp.glm.poisson.count <-
  try(glm((Y.NR.samp / offset.samp)~ X.1.samp + X.2.samp -1, family = quasipoisson(link = "log"),
  weights = w.k, start = Pop.glm.poisson.count.beta))
samp.glm.poisson.count.beta <- coefficients(samp.glm.poisson.count)
samp.fit.poisson.count <- offset.samp * fitted.values(samp.glm.poisson.count)
Pop.fit.poisson.count <-
  offset * predict(samp.glm.poisson.count, newdata =data.frame(X.1.samp = X.1, X.2.samp = X.2), type = "response")

samp.glm.cauchit.count <-
  try(glm(cbind(Y.NR.samp, Y.Resp.samp) ~ X.1.samp + X.2.samp -1, family = quasibinomial(link = "cauchit"),
  weights = w.k, start = Pop.glm.cauchit.count.beta))
samp.glm.cauchit.count.beta <- coefficients(samp.glm.cauchit.count)
samp.fit.cauchit.count <- offset.samp * fitted.values(samp.glm.cauchit.count)
Pop.fit.cauchit.count <-
  offset * predict(samp.glm.cauchit.count, newdata =data.frame(X.1.samp = X.1, X.2.samp = X.2), type = "response")

# Binary
samp.glm.probit.binary <-
  try(glm(Y.Bin.samp ~ X.1.samp + X.2.samp -1, family = quasibinomial(link = "probit"),
  weights = (w.k), start = Pop.glm.probit.binary.beta))
samp.glm.probit.binary.beta <- coefficients(samp.glm.probit.binary)
samp.fit.probit.binary <- fitted.values(samp.glm.probit.binary)
Pop.fit.probit.binary <-
  predict(samp.glm.probit.binary, newdata =data.frame(X.1.samp = X.1, X.2.samp = X.2), type = "response")

samp.glm.cloglog.binary <-
  try(glm(Y.Bin.samp ~ X.1.samp + X.2.samp -1, family = quasibinomial(link = "cloglog"),
  weights = (w.k), start = Pop.glm.cloglog.binary.beta))
samp.glm.cloglog.binary.beta <- coefficients(samp.glm.cloglog.binary)
samp.fit.cloglog.binary <- fitted.values(samp.glm.cloglog.binary)
Pop.fit.cloglog.binary <-
  predict(samp.glm.cloglog.binary, newdata =data.frame(X.1.samp = X.1, X.2.samp = X.2), type = "response")

samp.glm.cauchit.binary <-
  try(glm(Y.Bin.samp ~ X.1.samp + X.2.samp -1, family = quasibinomial(link = "cauchit"),
  weights = (w.k), start = Pop.glm.cauchit.binary.beta))
samp.glm.cauchit.binary.beta <- coefficients(samp.glm.cauchit.binary)
samp.fit.cauchit.binary <- fitted.values(samp.glm.cauchit.binary)
Pop.fit.cauchit.binary <-
  predict(samp.glm.cauchit.binary, newdata =data.frame(X.1.samp = X.1, X.2.samp = X.2), type = "response")

# Best
samp.glm.probit.best <-
  try(glm(Y.Best.samp ~ X.1.samp + X.2.samp -1, family = quasibinomial(link = "probit"),
  weights = (w.k), start = Pop.glm.probit.best.beta))
samp.glm.probit.best.beta <- coefficients(samp.glm.probit.best)
samp.fit.probit.best <- fitted.values(samp.glm.probit.best)
Pop.fit.probit.best <-
  predict(samp.glm.probit.best, newdata =data.frame(X.1.samp = X.1, X.2.samp = X.2), type = "response")

samp.glm.cloglog.best <-
  try(glm(Y.Best.samp ~ X.1.samp + X.2.samp -1, family = quasibinomial(link = "cloglog"),
  weights = (w.k), start = Pop.glm.cloglog.best.beta))
samp.glm.cloglog.best.beta <- coefficients(samp.glm.cloglog.best)
samp.fit.cloglog.best <- fitted.values(samp.glm.cloglog.best)
Pop.fit.cloglog.best <-
  predict(samp.glm.cloglog.best, newdata =data.frame(X.1.samp = X.1, X.2.samp = X.2), type = "response")

samp.glm.cauchit.best <-
  try(glm(Y.Best.samp ~ X.1.samp + X.2.samp -1, family = quasibinomial(link = "cauchit"),
  weights = (w.k), start = Pop.glm.cauchit.best.beta))
samp.glm.cauchit.best.beta <- coefficients(samp.glm.cauchit.best)
samp.fit.cauchit.best <- fitted.values(samp.glm.cauchit.best)
Pop.fit.cauchit.best <-
  predict(samp.glm.cauchit.best, newdata =data.frame(X.1.samp = X.1, X.2.samp = X.2), type = "response")

Samp.beta <- cbind(
        samp.glm.probit.count.beta,  samp.glm.cloglog.count.beta,  samp.glm.poisson.count.beta,  samp.glm.cauchit.count.beta,
        Samp.lm.NR.beta, Samp.lm.NR.beta.alt, Samp.lm.NR.beta.alt.2,
        samp.glm.probit.binary.beta, samp.glm.cloglog.binary.beta, samp.glm.cauchit.binary.beta, Samp.lm.Bin.beta,
        samp.glm.probit.best.beta,   samp.glm.cloglog.best.beta,   samp.glm.cauchit.best.beta,   Samp.lm.Best.beta)
Samp.fit <- cbind(
      samp.fit.probit.count, samp.fit.cloglog.count, samp.fit.poisson.count, samp.fit.cauchit.count,
      Samp.fit.lm.NR, Samp.fit.lm.NR.alt, Samp.fit.lm.NR.alt.2,
      samp.fit.probit.binary, samp.fit.cloglog.binary, samp.fit.cauchit.binary, Samp.fit.lm.Bin,
      samp.fit.probit.best, samp.fit.cloglog.best, samp.fit.cauchit.best, Samp.fit.lm.Best)
Pop.fit <- cbind(
      Pop.fit.probit.count, Pop.fit.cloglog.count, Pop.fit.poisson.count, Pop.fit.cauchit.count,
      Pop.fit.lm.NR, Pop.fit.lm.NR.alt, Pop.fit.lm.NR.alt.2,
      Pop.fit.probit.binary, Pop.fit.cloglog.binary, Pop.fit.cauchit.binary, Pop.fit.lm.Bin,
      Pop.fit.probit.best, Pop.fit.cloglog.best, Pop.fit.cauchit.best, Pop.fit.lm.Best)
Y.samp.all <- cbind(
      Y.NR.samp, Y.NR.samp, Y.NR.samp, Y.NR.samp, Y.NR.samp, Y.NR.samp, Y.NR.samp,
      Y.Bin.samp, Y.Bin.samp, Y.Bin.samp, Y.Bin.samp,
      Y.Best.samp, Y.Best.samp, Y.Best.samp, Y.Best.samp)
Y.Pop.all <- cbind(Y.NR, Y.NR, Y.NR, Y.NR, Y.NR, Y.NR, Y.NR, Y.Bin, Y.Bin, Y.Bin, Y.Bin,
  Y.Best, Y.Best, Y.Best, Y.Best)
```

```
  Y.col <- ncol(Samp.beta)

  ## Projective Estimator
  t.PROJ[j, ] <- colSums(Pop.fit)

  ## Stop if any estimates are infinity or bad
    error.3[j.master] <- ifelse(any(is.na(sum(Pop.fit))) ==TRUE, 1, 0)
    if(any(is.na(colSums(Pop.fit))) ==TRUE) next

  ### Sample Residuals
    clus.resid <- t(t((sapply(by(w.k * (Y.samp.all - Samp.fit), INDICES = ind.1, colSums, simplify = T), FUN = identity))))
    mean.resid <- matrix(rep(t(t((1 / a) * colSums(w.k * (Y.samp.all - Samp.fit)))), n.clus.samp),
        ncol = n.clus.samp, nrow = Y.col, byrow = FALSE)

  ### GGREG Variance Estimators
    v.GGREG.wr[j, ] <- (a / (a - 1)) * rowSums((clus.resid - mean.resid)^2)

    ssw.clus <- t(rowSums(t(1 - samp.clus.pi.cat) * (clus.resid)^2))
    ssw.within <-  t(t(t(
        sapply(by(w.k.2^2 * (1 - 1/w.k.2) * (Y.samp.all - Samp.fit)^2,
        INDICES = ind.1, colSums, simplify = T), FUN = identity))) %*% t(t(1 / samp.clus.pi)))
    v.GGREG.ssw[j, ] <- ssw.clus + ssw.within

    ##### Population prediction
    ## GGREG
    # Using PML
    t.GGREG[j, ] <- colSums(Pop.fit, na.rm = TRUE) -  colSums(w.k * (Y.samp.all - Samp.fit), na.rm = TRUE)

    ## Calibration
    # Same as GREG

    ## Model Calibration
    # Using PML
    # Just use mu with intercept
    k <- 1
    while(k < Y.col + 1)
    {
      samp.mu.k <- cbind(1, Samp.fit[,k])
      pop.mu.k <- cbind(1, Pop.fit[,k])

      A.mu.k <- t(samp.mu.k * w.k) %*% samp.mu.k

      error.4[j.master, k] <- ifelse( (sum(abs(eigen(A.mu.k, only.values = TRUE)$values) <= .Machine$double.eps) ||
                                  qr(A.mu.k)$rank < ncol(A.mu.k)), 1, 0)
      if( error.4[j.master, k] == 1) break

      t.MC <- (t(w.k) %*% Y.samp.all[,k]) + (colSums(pop.mu.k) -
               colSums(samp.mu.k * w.k)) %*% ( solve(A.mu.k) %*% t(samp.mu.k * w.k) %*% Y.samp.all[,k])
      t.MCAL[j, k] <- t.MC

    ### Model Calibration Residuals
    clus.resid.mc <- t(t(sapply(by(w.k * (Y.samp.all[,k] - samp.mu.k %*% ( solve(A.mu.k) %*% t(samp.mu.k * w.k) %*%
        Y.samp.all[,k])), INDICES = ind.1, colSums, simplify = T), FUN = identity)))
    mean.resid.mc <- matrix(rep(colMeans(clus.resid.mc), n.clus.samp), nrow = n.clus.samp, ncol = 1, byrow = FALSE)

    ### Model Calibration Variance Estimator
    v.MCAL.wr[j, k] <- (a / (a - 1)) * colSums((clus.resid.mc - mean.resid.mc)^2)

    ssw.clus.MCAL <- colSums((1 - samp.clus.pi.cat[,k]) * (clus.resid.mc)^2)
    ssw.within.MCAL <- t(1 / samp.clus.pi) %*% t(t(sapply(by(w.k.2^2 * (1 - 1/w.k.2) * (Y.samp.all[,k] - samp.mu.k %*%
        ( solve(A.mu.k) %*% t(samp.mu.k * w.k)%*% Y.samp.all[,k]))^2, INDICES = ind.1, sum, simplify = T), FUN = identity)))
    v.MCAL.ssw[j, k] <- ssw.clus.MCAL + ssw.within.MCAL

    g.k <- c(1 + (colSums(pop.mu.k) - colSums(samp.mu.k * w.k)) %*% ( solve(A.mu.k) %*% t(samp.mu.k)))
    clus.gresid.mc <- t(sapply(by((w.k * g.k) * (Y.samp.all[,k] -
        samp.mu.k %*% ( solve(A.mu.k) %*% t(samp.mu.k * w.k) %*% Y.samp.all[,k])),
        INDICES = ind.1, colSums, simplify = T), FUN = identity))
    ssw.gclus.MCAL <- (clus.gresid.mc)^2 %*% (1 - samp.clus.pi.cat[,k])
    ssw.gwithin.MCAL <- t(1 / samp.clus.pi) %*% t(t(sapply(by(w.k.2^2 * (1 - 1/w.k.2) * g.k^2 * (Y.samp.all[,k] -
        samp.mu.k %*% ( solve(A.mu.k) %*% t(samp.mu.k * w.k)%*% Y.samp.all[,k]))^2,
        INDICES = ind.1, sum, simplify = T), FUN = identity)))
    v.MCAL.ssw.g[j, k] <- ssw.gclus.MCAL + ssw.gwithin.MCAL
    k <- k + 1
    }
     if(sum(error.4[j.master,]) > 0) next

####### Variance of Estimation Equations (Start)
#### Common Estimates
### Create parameter vector
    k <- 1
    theta.k <- NULL
    while(k < Y.col + 1)
    {
      theta.k[[k]] <- c(t(t.GGREG[j,k]), t(t.MCAL[j,k]), Samp.beta[, k])
      k <- k+1
    }

### Length of parameters
  GGREG.start <- 1
  GGREG.end <- 1

  MCAL.start <- 2
```

```
  MCAL.end <- 2

  beta.start <- 3
  beta.end <- 2 + b.dim

## Create estimating equation function for estimating theta.pml
# The output of this function is the sum of the estimating equations for all units
W.probit <- function(par){
    beta <- t(t(par[(beta.start): beta.end]))
    LGREG.par <- par[1]
    MC.par <- par[2]

    mu.k <- pnorm(X.samp %*% beta)
    samp.mu <- cbind(1, mu.k)
    A.mu <- t(samp.mu * w.k) %*% samp.mu

    mu.k.pop <- pnorm(X.Pop %*% beta)
    pop.mu <- cbind(1, mu.k.pop)

  z.LGREG <- colSums(mu.k.pop) + sum(w.k * (Y.samp.i - mu.k)) - LGREG.par
  z.MCAL <- t.y.pi + (colSums(pop.mu) - colSums(samp.mu * w.k)) %*% ( solve(A.mu) %*% t(samp.mu * w.k) %*% Y.samp.i) - MC.par
  z.beta <- t(Y.samp.i) %*% (w.k * X.samp) - t(mu.k) %*% (w.k * X.samp)
  c(z.LGREG, z.MCAL, t(z.beta))
}
W.probit.o <- function(par){
    beta <- t(t(par[(beta.start): beta.end]))
    LGREG.par <- par[1]
    MC.par <- par[2]

    mu.k <- offset.samp * pnorm(X.samp %*% beta)
    samp.mu <- cbind(1, mu.k)
    A.mu <- t(samp.mu * w.k) %*% samp.mu

    mu.k.pop <- offset * pnorm(X.Pop %*% beta)
    pop.mu <- cbind(1, mu.k.pop)

  z.LGREG <- colSums(mu.k.pop) + sum(w.k * (Y.samp.i - mu.k)) - LGREG.par
  z.MCAL <- t.y.pi + (colSums(pop.mu) - colSums(samp.mu * w.k)) %*% ( solve(A.mu) %*% t(samp.mu * w.k) %*% Y.samp.i) - MC.par
  z.beta <- t(Y.samp.i) %*% (w.k * X.samp) - t(mu.k) %*% (w.k * X.samp)
  c(z.LGREG, z.MCAL, t(z.beta))
}

W.cloglog <- function(par){
    beta <- t(t(par[(beta.start): beta.end]))
    GGREG.par <- par[1]
    MC.par <- par[2]

    mu.k <- 1 - exp(- exp(X.samp %*% beta))
    samp.mu <- cbind(1, mu.k)
    A.mu <- t(samp.mu * w.k) %*% samp.mu

    mu.k.pop <- 1 - exp(- exp(X.Pop %*% beta))
    pop.mu <- cbind(1, mu.k.pop)

  z.GGREG <- colSums(mu.k.pop) + sum(w.k * (Y.samp.i - mu.k)) - GGREG.par
  z.MCAL <- t.y.pi + (colSums(pop.mu) - colSums(samp.mu * w.k)) %*% ( solve(A.mu) %*% t(samp.mu * w.k) %*% Y.samp.i) - MC.par
  z.beta <- t(Y.samp.i) %*% (w.k * X.samp) - t(mu.k) %*% (w.k * X.samp)
  c(z.GGREG, z.MCAL, t(z.beta))
}
W.cloglog.o <- function(par){
    beta <- t(t(par[(beta.start): beta.end]))
    GGREG.par <- par[1]
    MC.par <- par[2]

    mu.k <- offset.samp * (1 - exp(- exp(X.samp %*% beta)))
    samp.mu <- cbind(1, mu.k)
    A.mu <- t(samp.mu * w.k) %*% samp.mu

    mu.k.pop <- offset * (1 - exp(- exp(X.Pop %*% beta)))
    pop.mu <- cbind(1, mu.k.pop)

  z.GGREG <- colSums(mu.k.pop) + sum(w.k * (Y.samp.i - mu.k)) - GGREG.par
  z.MCAL <- t.y.pi + (colSums(pop.mu) - colSums(samp.mu * w.k)) %*% ( solve(A.mu) %*% t(samp.mu * w.k) %*% Y.samp.i) - MC.par
  z.beta <- t(Y.samp.i) %*% (w.k * X.samp) - t(mu.k) %*% (w.k * X.samp)
  c(z.GGREG, z.MCAL, t(z.beta))
}

W.poisson <- function(par){
    beta <- t(t(par[(beta.start): beta.end]))
    GGREG.par <- par[1]
    MC.par <- par[2]

    mu.k <- offset.samp * exp(X.samp %*% beta)
    samp.mu <- cbind(1, mu.k)
    A.mu <- t(samp.mu * w.k) %*% samp.mu

    mu.k.pop <- offset * exp(X.Pop %*% beta)
    pop.mu <- cbind(1, mu.k.pop)

  z.GGREG <- colSums(mu.k.pop) + sum(w.k * (Y.samp.i - mu.k)) - GGREG.par
  z.MCAL <- t.y.pi + (colSums(pop.mu) - colSums(samp.mu * w.k)) %*% ( solve(A.mu) %*% t(samp.mu * w.k) %*% Y.samp.i) - MC.par
  z.beta <- t(Y.samp.i) %*% (w.k * X.samp) - t(mu.k) %*% (w.k * X.samp)
```

```
  c(z.GGREG, z.MCAL, t(z.beta))
}

W.cauchit <- function(par){
    beta <- t(t(par[(beta.start): beta.end]))
    LGREG.par <- par[1]
    MC.par <- par[2]

    mu.k <- pcauchy(X.samp %*% beta)
    samp.mu <- cbind(1, mu.k)
    A.mu <- t(samp.mu * w.k) %*% samp.mu

    mu.k.pop <- pcauchy(X.Pop %*% beta)
    pop.mu <- cbind(1, mu.k.pop)

  z.LGREG <- colSums(mu.k.pop) + sum(w.k * (Y.samp.i - mu.k)) - LGREG.par
  z.MCAL <- t.y.pi + (colSums(pop.mu) - colSums(samp.mu * w.k)) %*% ( solve(A.mu) %*% t(samp.mu * w.k) %*% Y.samp.i) - MC.par
  z.beta <- t(Y.samp.i) %*% (w.k * X.samp) - t(mu.k) %*% (w.k * X.samp)
  c(z.LGREG, z.MCAL, t(z.beta))
}
W.cauchit.o <- function(par){
    beta <- t(t(par[(beta.start): beta.end]))
    LGREG.par <- par[1]
    MC.par <- par[2]

    mu.k <- offset.samp * pcauchy(X.samp %*% beta)
    samp.mu <- cbind(1, mu.k)
    A.mu <- t(samp.mu * w.k) %*% samp.mu

    mu.k.pop <- offset * pcauchy(X.Pop %*% beta)
    pop.mu <- cbind(1, mu.k.pop)

  z.LGREG <- colSums(mu.k.pop) + sum(w.k * (Y.samp.i - mu.k)) - LGREG.par
  z.MCAL <- t.y.pi + (colSums(pop.mu) - colSums(samp.mu * w.k)) %*% ( solve(A.mu) %*% t(samp.mu * w.k) %*% Y.samp.i) - MC.par
  z.beta <- t(Y.samp.i) %*% (w.k * X.samp) - t(mu.k) %*% (w.k * X.samp)
  c(z.LGREG, z.MCAL, t(z.beta))
}

W.identity <- function(par){
    beta <- t(t(par[(beta.start): beta.end]))
    GGREG.par <- par[1]
    MC.par <- par[2]

    mu.k <- (X.samp %*% beta)
    samp.mu <- cbind(1, mu.k)
    A.mu <- t(samp.mu * w.k) %*% samp.mu

    mu.k.pop <- (X.Pop %*% beta)
    pop.mu <- cbind(1, mu.k.pop)

  z.GGREG <- colSums(mu.k.pop) + sum(w.k * (Y.samp.i - mu.k)) - GGREG.par
  z.MCAL <- t.y.pi + (colSums(pop.mu) - colSums(samp.mu * w.k)) %*% ( solve(A.mu) %*% t(samp.mu * w.k) %*% Y.samp.i) - MC.par
  z.beta <- t(Y.samp.i) %*% (w.k * X.samp) - t(mu.k) %*% (w.k * X.samp)
  c(z.GGREG, z.MCAL, t(z.beta))
}
W.identity.o <- function(par){
    beta <- t(t(par[(beta.start): beta.end]))
    GGREG.par <- par[1]
    MC.par <- par[2]

    mu.k <- (X.samp %*% beta) + offset.samp
    samp.mu <- cbind(1, mu.k)
    A.mu <- t(samp.mu * w.k) %*% samp.mu

    mu.k.pop <- (X.Pop %*% beta) + offset
    pop.mu <- cbind(1, mu.k.pop)

  z.GGREG <- colSums(mu.k.pop) + sum(w.k * (Y.samp.i - mu.k)) - GGREG.par
  z.MCAL <- t.y.pi + (colSums(pop.mu) - colSums(samp.mu * w.k)) %*% ( solve(A.mu) %*% t(samp.mu * w.k) %*% Y.samp.i) - MC.par
  z.beta <- t(Y.samp.i) %*% (w.k * X.samp) - t(mu.k) %*% (w.k * X.samp)
  c(z.GGREG, z.MCAL, t(z.beta))
}
W.identity.alt <- function(par){
    beta <- t(t(par[(beta.start): beta.end]))
    GGREG.par <- par[1]
    MC.par <- par[2]

    mu.k <- offset.samp * (X.samp %*% beta)
    samp.mu <- cbind(1, mu.k)
    A.mu <- t(samp.mu * w.k) %*% samp.mu

    mu.k.pop <- offset * (X.Pop %*% beta)
    pop.mu <- cbind(1, mu.k.pop)

  z.GGREG <- colSums(mu.k.pop) + sum(w.k * (Y.samp.i - mu.k)) - GGREG.par
  z.MCAL <- t.y.pi + (colSums(pop.mu) - colSums(samp.mu * w.k)) %*% ( solve(A.mu) %*% t(samp.mu * w.k) %*% Y.samp.i) - MC.par
  z.beta <- t(Y.samp.i) %*% (w.k * X.samp) - t(mu.k) %*% (w.k * X.samp)
  c(z.GGREG, z.MCAL, t(z.beta))
}

# Jacobian
    t.y.pi <- t.HT[j,1]
```

```
    Y.samp.i <- Y.samp[,1]
    Jacob <- NULL
    Jacob[[1]] <- jacobian(W.probit.o, theta.k[[1]])
    Jacob[[2]] <- jacobian(W.cloglog.o, theta.k[[2]])
    Jacob[[3]] <- jacobian(W.poisson, theta.k[[3]])
    Jacob[[4]] <- jacobian(W.cauchit.o, theta.k[[4]])
    Jacob[[5]] <- jacobian(W.identity.o, theta.k[[5]])
    Jacob[[6]] <- jacobian(W.identity, theta.k[[6]])
    Jacob[[7]] <- jacobian(W.identity.alt, theta.k[[7]])

    t.y.pi <- t.HT[j,2]
    Y.samp.i <- Y.samp[,"Y.Bin"]
    Jacob[[8]] <- jacobian(W.probit, theta.k[[8]])
    Jacob[[9]] <- jacobian(W.cloglog, theta.k[[9]])
    Jacob[[10]] <- jacobian(W.cauchit, theta.k[[10]])
    Jacob[[11]] <- jacobian(W.identity, theta.k[[11]])

    t.y.pi <- t.HT[j,3]
    Y.samp.i <- Y.samp[,"Y.Best"]
    Jacob[[12]] <- jacobian(W.probit, theta.k[[12]])
    Jacob[[13]] <- jacobian(W.cloglog, theta.k[[13]])
    Jacob[[14]] <- jacobian(W.cauchit, theta.k[[14]])
    Jacob[[15]] <- jacobian(W.identity, theta.k[[15]])

## Cluster Level Estimating Equations for B
# The output of this function is the sum of the estimating equations for B within each cluster
# Unit Level Estimating Equations for B
    X.samp.unit <- split(X.samp, f = c(1:n.samp))
    resid.unit <- w.k * (Y.samp.all - Samp.fit)

    Est.Eq <- NULL
    Est.Eq.Matrix <- NULL
    t.W.clus <- NULL
    Jacob.GGREG <- NULL
    J.inv.GGREG <- NULL
    z.GGREG.b <- NULL
    z.GGREG <- NULL
    W.GGREG.all <- NULL
    W.GGREG.all.mean <- NULL
    Sigma.GGREG.j <- NULL
    var.GGREG.matrix <- NULL

    k <- 1
    while(k < Y.col + 1)
    {
      Est.Eq <- lapply(1:n.samp,
              function(i, X.samp.unit, resid.unit)
                      t(t(c(t(t(X.samp.unit[[i]])) %*% resid.unit[i, k]))),
                      X.samp.unit = X.samp.unit, resid.unit = resid.unit)
      Est.Eq.Matrix <- t(sapply(X = Est.Eq, FUN = identity, simplify = T, USE.NAMES = T))

    # Cluster Level Estimating Equations for B
      t.W.clus[[k]] <- t(sapply(by(Est.Eq.Matrix, ind.1, colSums), FUN = identity, simplify = T, USE.NAMES = T))

    #### GGREG
      Jacob.GGREG[[k]] <- Jacob[[k]][c(GGREG.start:GGREG.end, beta.start:beta.end), c(GGREG.start:GGREG.end, beta.start:beta.end)]
      error.5[j.master, k] <- ifelse( (any(abs(eigen(Jacob.GGREG[[k]], only.values = TRUE)$values) <= .Machine$double.eps) ||
                                      qr(Jacob.GGREG[[k]])$rank < ncol(Jacob.GGREG[[k]])), 1, 0)
      if( error.5[j.master, k] == 1) break

    # Invert the GGREG jacobian of the estimating equations with theta.pml as the input
      J.inv.GGREG[[k]] <- solve(Jacob.GGREG[[k]], tol = 1e-23)

    ## Cluster Level Estimating Equations for GGREG
    # The output of this function is the sum of the GGREG estimating equations within each cluster
      z.GGREG.b[[k]] <- (w.k * (Y.samp.all[,k] - Samp.fit[,k]))
      z.GGREG[[k]] <- t(t(sapply(by(z.GGREG.b[[k]], INDICES = ind.1, sum, simplify = T), FUN = identity)))

    ## Combine Cluster Level Estimating Equations
      W.GGREG.all[[k]] <- cbind(z.GGREG[[k]], t.W.clus[[k]])

    # Mean of Cluster Level Estimating Equations
      W.GGREG.all.mean[[k]] <- t(colMeans(W.GGREG.all[[k]])) %x% t(t(rep(1,nrow(W.GGREG.all[[k]]))))

    # Covariance Matrix
      Sigma.GGREG.j[[k]] <- (a / (a - 1)) * t(W.GGREG.all[[k]] - W.GGREG.all.mean[[k]]) %*% (W.GGREG.all[[k]] - W.GGREG.all.mean[[k]])

    # Variance of GGREG
      var.GGREG.matrix[[k]] <- diag(J.inv.GGREG[[k]] %*% Sigma.GGREG.j[[k]] %*% t(J.inv.GGREG[[k]]))[1]
      v.GGREG.pml[j, k] <- var.GGREG.matrix[[k]]

      k <- k+1
    }
      if( sum(error.5[j.master, ]) > 0) next

#### MCAL
# MCAL Jacobian
    Jacob.MCAL <- NULL
    J.inv.MCAL <- NULL
    z.MCAL <- NULL
    W.MCAL.all <- NULL
    W.MCAL.all.mean <- NULL
```

```
    Sigma.MCAL.j <- NULL
    var.MC.matrix <- NULL

    k <- 1
    while(k < Y.col + 1)
    {
      samp.mu.k <- cbind(1, Samp.fit[,k])
      A.mu.k <- t(samp.mu.k * w.k) %*% samp.mu.k

      Jacob.MCAL <- Jacob[[k]][c(MCAL.start:beta.end), c(MCAL.start:beta.end)]
      error.6[j.master] <- ifelse( (any(abs(eigen(Jacob.MCAL, only.values = TRUE)$values) <= .Machine$double.eps) ||
                                    qr(Jacob.MCAL)$rank < ncol(Jacob.MCAL)), 1, 0)
      if(error.6[j.master, k] == 1) next

        # Invert the MCAL jacobian of the estimating equations with theta.pml as the input
        J.inv.MCAL <- solve(Jacob.MCAL, tol =1e-23)

        ## Cluster Level Estimating Equations for MCAL
        # The output of this function is the sum of the GGREG estimating equations within each cluster
        z.MCAL <- t(t(sapply(by(w.k * (Y.samp.all[,k] - samp.mu.k %*% ( solve(A.mu.k) %*%
            t(samp.mu.k * w.k) %*% Y.samp.all[,k])), INDICES = ind.1, colSums, simplify = T), FUN = identity)))

        ## Cluster Level Estimating Equations for B: Same as GGREG
        ## Combine Cluster Level Estimating Equations
        W.MCAL.all <- cbind(z.MCAL, t.W.clus[[k]])

        # Mean of Cluster Level Estimating Equations
        W.MCAL.all.mean <- matrix(rep(colMeans(W.MCAL.all), n.clus.samp), nrow = n.clus.samp, byrow = TRUE)

        # Covariance Matrix
        Sigma.MCAL.j <- (a / (a - 1)) * t(W.MCAL.all - W.MCAL.all.mean) %*% (W.MCAL.all - W.MCAL.all.mean)

        # Variance of MCAL
        var.MC.matrix <- diag(J.inv.MCAL %*% Sigma.MCAL.j %*% t(J.inv.MCAL))[1]
        v.MCAL[j, k] <- var.MC.matrix
      k <- k+1
    }

####### Variance of Estimation Equations (End)

    ## Pseudoempirical Maximimum Likelihood
    m.1 <- nrow(X.Pop)
    ds <- w.k / sum(w.k)

    # Using PML: Mean
    k <- 1
    while(k < Y.col + 1)
    {
      u <- Samp.fit[,k]
      mu <- mean(Pop.fit[,k])
      lambda.1 <- Lag2(u = u, ds = ds, mu = mu)
      mu.matrix <- rep(mean(Pop.fit[,k]), length(w.k))
      p.i <- (ds) / (1 + (u - mu.matrix) %*% t(t(lambda.1)) )
      t.PEMLE.N[j, k] <- M.1 * t(p.i) %*% (Y.samp.all[,k])
      t.PEMLE.N.hat[j, k] <- t(p.i * sum(w.k)) %*% (Y.samp.all[,k])
#     t.PEMLE.pml.w[j, k] <- (1 / (length(w.k))) * t(w.k/p.i) %*% (Y.samp.all[[k]])

      k <- k+1
    }

    if(((j) %% 10) == 0)
    {
        cat(j, format(Sys.time(), "%X"), "\n",
               "  True:               ", sum(Y.Pop[,1]), "\n",
               "  Mean t.HT           ", round(mean(t.HT[1:j,1])), "\n",
               "  Mean t.GREG:        ", round(mean(t.GREG[1:j,1])), "\n",
               "  Mean t.PROJ:        ", round(mean(t.PROJ[1:j,1])), "\n",
               "  Mean t.GGREG:       ", round(mean(t.GGREG[1:j,1])), "\n",
               "  Mean t.MCAL:        ", round(mean(t.MCAL[1:j,1])), "\n",
               "  Mean t.PEMLE.N:     ", round(mean(t.PEMLE.N[1:j,1])), "\n",
               "  Mean t.PEMLE.N.hat: ", round(mean(t.PEMLE.N.hat[1:j,1])), "\n",
               "  se t.HT             ", round(sqrt(var(t.HT[1:j,1]))), "\n",
               "  se t.GREG:          ", round(sqrt(var(t.GREG[1:j,1]))), "\n", "\n",
               "  se t.GGREG:         ", round(sqrt(var(t.GGREG[1:j,1]))), "\n",
               "  se.wr t.GGREG:      ", round(sqrt(mean(v.GGREG.wr[1:j,1], na.rm=TRUE))), "\n",
               "  se.ssw t.GGREG:     ", round(sqrt(mean(v.GGREG.ssw[1:j,1], na.rm=TRUE))), "\n",
               "  se.pml t.GGREG:     ", round(sqrt(mean(v.GGREG.pml[1:j,1], na.rm=TRUE))), "\n", "\n",
               "  se t.PROJ:          ", round(sqrt(var(t.PROJ[1:j,1]))), "\n",
               "  se t.MCAL:          ", round(sqrt(var(t.MCAL[1:j,1], na.rm=TRUE))), "\n",
               "  se.wr t.MCAL:       ", round(sqrt(mean(v.MCAL.wr[1:j,1], na.rm=TRUE))), "\n",
               "  se.ssw.e t.MCAL:    ", round(sqrt(mean(v.MCAL.ssw[1:j,1], na.rm=TRUE))), "\n",
               "  se.ssw.g t.MCAL:    ", round(sqrt(mean(v.MCAL.ssw.g[1:j,1], na.rm=TRUE))), "\n",
               "  se.MCAL.pml:        ", round(sqrt(mean(v.MCAL[1:j,1], na.rm=TRUE))), "\n", "\n",
               "  se t.PEMLE.N:       ", round(sqrt(var(t.PEMLE.N[1:j,1]))), "\n",
               "  se t.PEMLE.N.hat:   ", round(sqrt(var(t.PEMLE.N.hat[1:j,1]))), "\n",
               "\n")
    }

    j <- j + 1
    print(j.master)

  }
```

```
  list(t.HT,
       t.GREG,
       t.PROJ,
       t.GGREG,
       t.MCAL,
       t.PEMLE.N,
       t.PEMLE.N.hat,
       v.GGREG.wr,
       v.GGREG.ssw,
       v.GGREG.pml,
       v.MCAL.wr,
       v.MCAL.ssw,
       v.MCAL.ssw.g,
       v.MCAL,
       error.1, error.2, error.3, error.4, error.5, error.6)
}
```

Bibliography

Agresti, A. (2002), *Categorical Data Analysis*, John Wiley & Sons. 114, 115, 117, 119, 196, 308, 312, 313, 333, 366

Basu, D. (1971), "An Essay on the Logical Foundations of Survey Sampling, Part I," *Foundations of Statistical Inference*, 203–242. 21, 31

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley. 66, 68

Binder, D. A. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys," *International Statistical Review / Revue Internationale de Statistique*, 51, 279–292. 35, 53, 54, 55, 56, 136, 140, 154, 203, 217, 221, 325, 326, 331, 348, 349, 352

Bishop, Y. M., Fienberg, S. E., and Holland, P. W. (2007), *Discrete Multivariate Analysis*, New York: Springer Verlag. 114, 196

Box, G. E. P. and Draper, N. R. (1987), *Empirical Model-Building and Response Surfaces*, Wiley, New York. 77

Brewer, K. (1963), "Ratio Estimation and Finite Populations: Some Results Deductible from the Assumption of an Underlying Stochastic Process," *Australian & New Zealand Journal of Statistics*, 5, 93–105. 76

Brewer, K. R. W. (1979), "A Class of Robust Sampling Designs for Large-Scale Surveys," *Journal of the American Statistical Association*, 74, 911–915. 28

Bruce, A. and Robinson, J. G. (2006), "Tract-Level Planning Database with Census 2000 Data," *US Department of Commerce, US Census Bureau: Washington, DC http://www. census. gov/procur/www/2010communications/library. html*. 226

Carroll, R. J., Wang, S., Simpson, D. G., Stromberg, A. J., and Ruppert, D. (1998), "The Sandwich (Robust Covariance Matrix) Estimator," *Unpublished manuscript*, available at https://www.stat.tamu.edu/ftp/pub/rjcarroll/sandwich.pdf. 76

Casella, G. and Berger, R. L. (2002), *Statistical Inference*, Duxbury Press Belmont, Calif. 61

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976), "Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations," *Biometrika*, 63, 615–620, with comments by T. M. F. Smith and a reply by the authors. 34

Chen, J. and Qin, J. (1993), "Empirical Likelihood Estimation for Finite Populations and the Effective Usage of Auxiliary Information," *Biometrika*, 80, 107. 52, 53, 129, 361

Chen, J. and Sitter, R. R. (1999), "A Pseudo Empirical Likelihood Approach to the Effective use of Auxiliary Information in Complex Surveys," *Statistica Sinica*, 9, 385–406. 52, 53, 129, 361, 365

Cochran, W. M. (1953), *Sampling Techniques*, New York. 4

— (1977), *Sampling Techniques*, Wiley and Sons, Inc, third edition ed. 27

Cumberland, W. G. and Royall, R. M. (1981), "Prediction Models and Unequal Probability Sampling," *Journal of the Royal Statistical Society. Series B (Methodological)*, 43, 353–367. 18

— (1988), "Does Simple Random Sampling Provide Adequate Balance?" *Journal of the Royal Statistical Society. Series B (Methodological)*, 50, 118–124. 30, 78

Deming, W. E. (1950), *Some Theory of Sampling*, Wiley, New York. 4

Deville, J.-C. and Särndal, C.-E. (1992), "Calibration Estimators in Survey Sampling," *Journal of the American Statistical Association*, 87, 376–382. 48, 49, 50, 124, 125, 126, 209

Efron, B. (1982), *The Jackknife, the Bootstrap and other Resampling Plans*, CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia: Society for Industrial and Applied Mathematics (SIAM), 1982. 72

Eicker, F. (1963), "Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions," *Annals of Mathematical Statistics*, 34, 447–56. 71

— (1967), "Limit Theorems for Regressions with Unequal and Dependent Errors," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 59–82. 71

Estevao, V. M., Hidiroglou, M. A., and Sarndal, C.-E. (1995), "Methodological Principles for a Generalized Estimation System at Statistics Canada," *Journal of Official Statistics*, 11, 181–204. 33, 39

Estevao, V. M. and Särndal, C.-E. (2006), "Survey Estimates by Calibration on Complex Auxiliary Information," *International Statistical Review*, 74, 127–147. 5, 49, 125

Fahrmeir, L. and Tutz, G. (2001), *Multivariate Statistical Modelling Based on Generalized Linear Models*, Springer Verlag. 117, 119

Firth, D. and Bennett, K. E. (1998), "Robust Models in Probability Sampling," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 60, 3–21. 35, 48, 204, 205, 206, 233, 243, 245

Fuller, W. A. (2009), *Sampling Statistics*, John Wiley & Sons, Inc. 258, 297

Gilbert, P. (2012), *numDeriv: Accurate Numerical Derivatives*, r package version 2012.3-1. 155, 237

Groves, R. M., Fowler, Jr., F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2004), *Survey Methodology*, Wiley Series in Survey Methodology, Hoboken, NJ: Wiley-Interscience [John Wiley & Sons]. 31

Hansen, M. H. and Hurwitz, W. N. (1943), "On the Theory of Sampling from Finite Populations," *The Annals of Mathematical Statistics*, 14, 333–362. 4, 14, 15

Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953a), *Sample Survey Methods and Theory*, vol. Volume I: Methods and Applications, New York, NY (EUA). John Wiley. 4, 22, 26, 60

— (1953b), *Sample Survey Methods and Theory*, vol. Volume II: Theory, New York, NY (EUA). John Wiley. 4, 27

Hansen, M. H., Madow, W. G., and Tepping, B. J. (1983), "An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys," *Journal of the American Statistical Association*, 78, 776–793. 32, 77, 79

Harville, D. A. (1997), *Matrix Algebra from a Statistician's Perspective*, Springer, New York. 119

Hidiroglou, M. A., Fuller, W. A., and Hickman, R. D. (1980), *Super Carp*, Survey Section, Statistical Laboratory, Iowa State University. 40

Hidiroglou, M. A., Särndal, C. E., and Binder, D. A. (1995), "Weighting and Estimation in Business Surveys," *Business Survey Methods*, 477–502. 33

Hilbe, J. M. (2009), *Logistic regression models*, New York: Chapman & Hall/CRC Press. 114, 312

Hinkley, D. V. (1977), "Jackknifing in Unbalanced Situations," *Technometrics*, 19, 285–292. 71, 72

Hoaglin, D. C. and Welsch, R. E. (1978), "The Hat Matrix in Regression and ANOVA," *The American Statistician*, 32, 17–22. 65

Hoel, P. G., Port, S. C., and Stone, C. J. (1971), *Introduction to probability theory*, Houghton Mifflin. 61

Hogg, R. V. and Craig, A. T. (1995), *Introduction to Mathematical Statistics, 1995*, Prentice-Hall, Inc. 61

Horn, S. D., Horn, R. A., and Duncan, D. B. (1975), "Estimating Heteroscedastic Variances in Linear Models," *Journal of the American Statistical Association*, 70, 380–385. 72

Horvitz, D. G. and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47, 663–685. 4, 16, 17, 18

Hosmer, D. W. and Lemeshow, S. (2000), *Applied logistic regression*, New York: Wiley-Interscience. 114, 312

Huber, P. J. (1967), "The behavior of maximum likelihood estimates under nonstandard conditions," *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics*, 1, 221–233. 71

Isaki, C. T. and Fuller, W. A. (1982), "Survey Design Under the Regression Superpopulation Model," *Journal of the American Statistical Association*, 77, 89–96. 28, 34

Kim, J.-Y., Breidt, F. J., and Opsomer, J. D. (2009), "Nonparametric Regression Estimation of Finite Population Totals under Two-Stage Sampling," Tech. Rep. 2009/4, Department of Statistics, Colorado State University. 52, 127, 210

Kott, P. S. (1988), "Model-Based Finite Population Correction for the Horvitz-Thompson Estimator," *Biometrika*, 75, pp. 797–799. 85, 86

— (1990), "Estimating the conditional variance of a design consistent regression estimator," *Journal of statistical planning and inference*, 24, 287–296. 39

Krewski, D. and Rao, J. (1981), "Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods," *The Annals of Statistics*, 9, 1010–1019. 83

Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2005), *Applied linear statistical models*, McGraw-Hill. 66, 67, 308, 312

Lehtonen, R. and Pahkinen, E. (2004), *Practical methods for design and analysis of complex surveys*, Wiley. 35, 56, 204

Lehtonen, R. and Veijanen, A. (1998), "Logistic generalized regression estimators," *Survey Methodology*, 24, 51–55. 122, 123, 124, 132

Li, J. and Valliant, R. (2009), "Survey weighted hat matrix and leverages," *Survey Methodology*, 35, 15–25. 82, 271

Little, R. J. (2004), "To model or not to model? Competing modes of inference for finite population sampling," *Journal of the American Statistical Association*, 99, 546–556. 21, 31

Long, J. S. and Ervin, L. H. (2000), "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model," *The American Statistician*, 54, 217–224. 58

MacKinnon, J. G. and White, H. (1985), "Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties," *Journal of Econometrics*, 29, 305–325. 58, 72

McCullagh, P. and Nelder, J. A. (1999), *Generalized linear models. (Monographs on Statistics and Applied Probability)*, Boca Raton: Chapman and Hall/CRC. 114, 196, 304

McCulloch, C. E. and Searle, S. R. (2004), *Generalized, linear, and mixed models*, Wiley-Interscience. 196

Narain, R. D. (1951), "On sampling without replacement with varying probabilities," *Journal of the Indian Society of Aricultural Statistics*, 3, 169–174. 4, 16

Nelder, J. A. and Wedderburn, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society. Series A (General)*, 135, 370–384. 196, 203

Neyman, J. (1934), "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," *Journal of the Royal Statistical Society*, 97, 558–625. 3, 4, 30, 77

Oman, S. and Zucker, D. (2001), "Modelling and generating correlated binary variables," *Biometrika*, 88, 287–290. 228

Pouillot, R. and Delignette-Muller, M.-L. (2010), "Evaluating variability and uncertainty in microbial qunatitative risk assessment using two R packages," *International Journal of Food Microbiology*, 142, 330–40. 146

Prášková, Z. and Sen, P. K. (2009), "Asymptotics in Finite Population Sampling," *Handbook of Statistics*, 29, 489–522. 28, 29

R Development Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. 154

Rao, J. and Wu, C. (2009), "Empirical likelihood methods," *Handbook of Statistics*, 29, 189–207. 128, 210

Rencher, A. C. (2000), *Linear models in statistics*, Wiley Series in Probability and Statistics, New York: John Wiley & Sons Inc., a Wiley-Interscience Publication. 64, 65

Roberts, G., Rao, J. N. K., and Kumar, S. (1987), "Logistic Regression Analysis of Sample Survey Data," *Biometrika*, 74, 1–12. 56

Robinson, P. M. and Särndal, C.-E. (1983), "Asymptotic properties of the generalized regression estimator in probability sampling," *Sankhyā Ser. B*, 45, 240–248. 34

Royall, R. M. (1970), "On Finite Population Sampling Theory Under Certain Linear Regression Models," *Biometrika*, 57, 377–387. 30, 61, 79

Royall, R. M. and Cumberland, W. G. (1981), "The Finite-Population Linear Regression Estimator and Estimators of its Variance–An Empirical Study," *Journal of the American Statistical Association*, 76, 924–930. 71

Royall, R. M. and Herson, J. (1973), "Robust Estimation in Finite Populations I," *Journal of the American Statistical Association*, 68, 880–889. 71, 77, 78

RTI (2004), "SUDAAN Manual Release 9.0," *Research Triangle Park, NC, Research Triangle Institute*. 56

Särndal, C.-E. (1980a), "On π-inverse weighting versus best linear unbiased weighting in probability sampling," *Biometrika*, 67, 639–650. 36, 208

— (1980b), "Two model-based inference arguments in survey sampling," *Austral. J. Statist.*, 22, 341–348. 34

— (1981), "Frameworks for inference in survey sampling with applications to small area estimation and adjustment for nonresponse," *Bulletin of the International Statistical Institute*, 49, 494–513. 44

— (1982), "Implications of survey design for generalized regression estimation of linear functions," *J. Statist. Plann. Inference*, 7, 155–170. 34, 44

— (2007), "The calibration approach in survey theory and practice," *Survey Methodology*, 33, 99 – 119. 34, 35, 36, 49, 125

Särndal, C.-E., Swensson, B., and Wretman, J. H. (1989), "The weighted residual technique for estimating the variance of the general regression estimator of the finite population total," *Biometrika*, 76, 527–537. 39, 44, 58, 136, 140, 145, 187, 217, 221, 226, 346, 456

— (1992), *Model assisted survey sampling*, Springer Series in Statistics, New York: Springer-Verlag. 9, 10, 16, 20, 21, 27, 28, 35, 36, 37, 38, 40, 42, 43, 44, 45, 46, 47, 54, 56, 58, 86, 100, 324, 325, 326, 335, 346, 347, 349, 360, 456

Searle, S. R. (1982), *Matrix algebra useful for statistics*, John Wiley New York. 119

Seber, G. A. F. (2008), *A matrix handbook for statisticians*, vol. 746, Wiley-Interscience. 119, 356, 357, 366

Sen, A. R. (1953), "On the estimate of the variance in sampling with varying probabilities," *Journal of the Indian Society of Agricultural Statistics*, 5, 119–127. 4, 17

Shao, J. (2003), *Mathematical Statistics*, New York: Springer Verlag. 61, 114, 117, 196, 197, 199, 200, 201, 300, 305, 310, 439, 444

Sitter, R. R. and Wu, C. (2002), "Efficient estimation of quadratic finite population functions in the presence of auxiliary information," *Journal of the American Statistical Association*, 97, 535–543. 129, 211

Theil, H. (1971), *Principles of Econometrics*, John Wiley, New York. 67

Tillé, Y. (2006), *Sampling algorithms*, Springer Verlag. 12

Tillé, Y. and Matei, A. (2009), *sampling: Survey Sampling*, r package version 2.3. 12, 151, 234

Valliant, R. (1985), "Nonlinear Prediction Theory and the Estimation of Proportions in a Finite Population," *Journal of the American Statistical Association*, 80, 631–641. 207

— (2002), "Variance estimation for the general regression estimator," *Survey Methodology*, 28, 103–114. 41, 42, 58, 82

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000), *Finite population sampling and inference: A prediction approach*, Wiley Series in Probability and Statistics: Survey Methodology Section, Wiley-Interscience, New York. 35, 41, 68, 69, 71, 72, 73, 74, 75, 78, 79, 88, 89, 207, 277

White, H. (1980), "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity," *Econometrica*, 48, 817–838. 72

Wright, R. L. (1983), "Finite Population Sampling With Multivariate Auxiliary Information," *Journal of the American Statistical Association*, 78, 879–884. 34

Wu, C. (1999), "The Effective Use of Complete Auxiliary Information from Survey Data," Ph.D. thesis, Simon Fraser University. 365

Wu, C. and Sitter, R. R. (2001), "A Model-Calibration Approach to Using Complete Auxiliary Information from Survey Data," *Journal of the American Statistical Association*, 96, 185–193. 48, 50, 51, 52, 53, 121, 122, 126, 127, 129, 130, 133, 207, 209, 211, 213, 214, 253, 317, 319, 320, 321, 339, 340, 342, 343, 361

Yates, F. (1949), *Sampling for Censuses and Surveys*, Griffin. 4

Yates, F. and Grundy, P. M. (1953), "Selection Without Replacement from Within Strata with Probability Proportional to Size," *Journal of the Royal Statistical Society. Series B (Methodological)*, 15, 253–261. 4, 17

Yee, T. W. (2012), *VGAM: Vector Generalized Linear and Additive Models*, R package version 0.8-6. 154

Yung, W. and Rao, J. N. K. (1996), "Jackknife linearization variance estimators under stratified multi-stage sampling," *Survey Methodology*, 22, 23–31. 40, 45, 86, 277

Zhong, C. X. B. and Rao, J. N. K. (1996), "Empirical Likelihood Inference under Stratified Random Sampling Using Auxiliary Information," in *ASA Proceedings of the Section on Survey Research Methods*, American Statistical Association, pp. 798–803. 52, 53, 129, 361