ABSTRACT

Title of dissertation: Collinearity Diagnostics for Complex Survey Data

Dan Liao
Doctor of Philosophy, 2010

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

Survey data are often used to fit models. The values of covariates used in modeling are not controlled as they might be in an experiment. Thus, collinearity among the covariates is an inevitable problem in the analysis of survey data. Although many books and articles have described the collinearity problem and proposed strategies to understand, assess and handle its presence, the survey literature has not provided appropriate diagnostic tools to evaluate its impact on the regression estimation when the survey complexities are considered. The goal of this research is to extend and adapt the conventional ordinary least squares collinearity diagnostics to complex survey data when a linear model or generalized linear model is used.

In this dissertation we have developed methods that generally have either a model-based or design-based interpretation. We assume that an analyst uses survey-weighted regression estimators to estimate both underlying model parameters (assuming a correctly specified model) and census-fit parameters in the finite population. Diagnostics statistics, variance inflation factors (VIFs), condition indexes and variance decomposition proportions are constructed to evaluate the impact of

collinearity and determine which variables are involved. Survey weights are components of the diagnostic statistics and the estimated variances of the coefficients are obtained from design-consistent estimators which account for complex design features, e.g. clustering and stratification.

Illustrations of these methods are given using data from a survey of mental health organizations and a household survey of health and nutrition. We demonstrate that specialized collinearity diagnostic statistics are needed to account for survey weights and complex finite population features that are reflected in the sample design and considered in the regression analysis.

Collinearity Diagnostics for Complex Survey Data

by

Dan Liao

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
2010

Advisory Committee:
Dr. Richard Valliant, Chair/Advisor
Dr. Michael Elliott
Dr. Wolfgang Jank
Dr. Stephen Miller
Dr. Nathaniel Schenker

Acknowledgments

I owe my gratitude to all the people who have made this thesis possible and because of whom my graduate experience has been one that I will cherish forever.

First and foremost I would like to express my deepest gratitude to my advisor, Dr. Richard Valliant, for his excellent guidance, caring, patience, and providing me with an excellent atmosphere for doing research. It has been a pleasure to work with and learn from such an extraordinary individual.

I would like to thank Dr. Michael Elliot, Dr. Wolfgang Jank, Dr. Stephen Miller and Dr. Nathaniel Schenker, for agreeing to serve on my thesis committee and giving me their precious time and expertise to better my work.

I must acknowledge as well the many friends, colleagues, students, and faculties who assisted, advised, and supported my study and research over the years in the Joint Program in Survey Methodology(JPSM). Especially, thank all the JPSM doctoral students for their friendship, wisdom and knowledge, which enlightened my life and encouraged me through the whole PhD study.

I would also like to acknowledge help and support from all the staff members in JPSM, especially, Rupa Jethwa Eapen, Sarah Gebremicael and Duane Gilbert.

Last but not least, I owe my deepest thanks to my family, who have always stood by me and guided me through my career and life.

Table of Contents

List of Tables v

List of Figures viii

1 Introduction 1
1.1 Regression Diagnostics . . . 1
1.2 Collinearity Diagnostics . . . 3
1.3 Regression Analysis with Complex Survey Data . . . 5
1.4 The Subject of This Dissertation . . . 8

2 Review of Traditional Techniques 10
2.1 Collinearity Diagnostics in Ordinary Least Squares . . . 10
2.1.1 Variance Inflation Factor . . . 11
2.1.2 Condition Indexes with Variance Decomposition . . . 13
2.1.2.1 Eigenvalues and Eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$ . . . 13
2.1.2.2 Singular-Value Decomposition, Condition Number and Condition Indexes . . . 14
2.1.2.3 Variance Decomposition Method . . . 16
2.2 Collinearity Diagnostics in Generalized Linear Model . . . 19
2.2.1 Definitions and Notations . . . 20
2.2.2 Condition Indexes with Variance Decomposition in GLM . . . 23

3 Variance Inflation Factor 26
3.1 VIF in Survey Weighted Least Squares Regression . . . 26
3.1.1 Survey-Weighted Least Squares Estimators . . . 26
3.1.2 Model Variance of Coefficient Estimates . . . 28
3.1.3 The Coefficient of Multiple Correlation . . . 29
3.1.4 Model-based VIF . . . 30
3.1.5 Intercept-Adjusted Model-based VIF . . . 37
3.1.6 Estimating the VIF with Known $\boldsymbol{V}$ in a Sample Selected from the Finite Population . . . 40
3.1.7 Estimating the VIF with Unknown $\boldsymbol{V}$ in a Sample Selected from the Finite Population . . . 46
3.1.7.1 VIF for A Model with Independent Errors . . . 47
3.1.7.2 VIF for A Model with Clustering . . . 50
3.1.7.3 VIF for A Model with Stratified Clustering . . . 54
3.1.7.4 Specialization for Stratified Models with No Clustering 62
3.1.8 VIF in Survey-Weighted Generalized Least Squares Regression 65
3.2 Experimental Study . . . 68
3.2.1 Introduction . . . 68
3.2.2 Survey of Mental Health Organizations . . . 68
3.2.2.1 Description of Study Population . . . 68
3.2.2.2 Collinearity in the Study Population . . . 71

3.2.2.3 Simulation . . . 76
3.2.3 National Health and Nutrition Examination Survey: 2001-2002 90
3.2.3.1 Description of the Data . . . 90
3.2.3.2 Collinearity Diagnostics for NHANES 2001-2002 . . . 90
3.2.3.3 Simulation . . . 96

4 Condition Index with Variance Decomposition Method 109
4.1 Adaptation in Survey-Weighted Least Squares . . . 110
4.1.1 Adaptation under the Model-Based Inference when $\boldsymbol{V}$ is known 110
4.1.2 Decomposition of the Estimated Variance in a Sample Selected from the Finite Population . . . 114
4.1.2.1 Variance Decomposition for A Model with Independent Errors . . . 114
4.1.2.2 Variance Decomposition for A Model with Clustering 116
4.1.2.3 Variance Decomposition for A Model with Stratified Clustering . . . 118
4.2 Experimental Study . . . 124
4.2.1 Survey of Mental Health Organizations . . . 127
4.2.2 National Health and Nutrition Examination Survey: 2001-2002 136

5 Collinearity Diagnostics in Generalized Linear Models 151
5.1 Survey-Weighted Generalized Linear Models . . . 151
5.2 Variance Inflation Factor in Generalized Linear Model . . . 153
5.2.1 Model-based VIF . . . 153
5.2.2 Design-based VIF for Three Typical Sampling Designs . . . 156
5.2.2.1 Linearization Variance Estimator for $\hat{\boldsymbol{\beta}}_{SW}$ . . . 156
5.2.2.2 VIF for $var_L(\hat{\boldsymbol{\beta}}_{SW})$ . . . 162
5.2.3 Logistic Model . . . 167
5.2.3.1 introduction . . . 167
5.2.3.2 Variance Inflation Factor in Logistic Model . . . 169
5.3 Condition Indexes with Variance Decomposition Method in Generalized Linear Model . . . 173
5.4 Experimental Study . . . 175

6 Conclusion and Discussion 185

A Inversion of Partitioned Matrices, $\dot{a}^{00}$, $\boldsymbol{a}^{(2\ldots p)1}$ 191

B1 Derivation of the Model Variance of $\hat{\beta}_{SWk}$ in M2 194

B2 Intercepted-adjusted VIF in OLS 196

C Linearization Variance Estimation of $\hat{\boldsymbol{\beta}}$ in Linear and Generalized Linear Models 197

D Expression of $Blkdiag(\boldsymbol{e}_i\boldsymbol{e}_i^T)$ and $Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)$ 199

List of Tables

3.1 Partitioning the Total Sum of Squares 30

3.2 R-Square for Several Selected Models in Different Samples in the Simulations 71

3.3 VIFs of the Three-Variable Regression and Five-Variable Regression in the Full Finite Population. $\hat{\boldsymbol{\beta}}_{OLS}$ used in all cases; three methods of variance estimation used. 75

3.4 Four Types of Linear Regressions and their VIF Formulas Used in the Simulation 80

3.5 Comparison between the VIFs from the Full Finite Population and the Average VIFs from 1,000 Simulations 81

3.6 Relative Biases in Estimated Model Parameters Using Different Regression Types and VIF Cut-off Values 85

3.7 Percentage of Samples where Models were Selected and Coverage of 95% Confidence Intervals Using Different Regression Types and VIF Cut-off Values 86

3.8 Coverage Rates of the Confidence Regions for the True Coefficient Values in the Artificial Population based on SMHO 88

3.9 Ratios of the Average Estimated $se(\hat{\boldsymbol{\beta}})$ to empirical $SE(\hat{\boldsymbol{\beta}})$ Using Different Regression Types and VIF Cut-off Values 89

3.10 Sample Size in each PSU in NHANES 2001-2002 Data File 91

3.11 VIFs of the Five Explanatory Variables in NHANES 2001-2002 Data File 94

3.12 The Final Model Obtained Using Different VIF Methods and VIF Cut-off Values 96

3.13 R-Square for Several Selected Models in Different Samples in the Simulations 99

3.14 Comparison between the VIFs from the Full Finite Population and the Average VIFs from 1000 Simulations 100

3.15 Percent Bias in Estimated Model Parameters Using Different Regression Types and VIF Cut-off Values 105

3.16 Percentage of Samples where Models were Selected and Coverage of 95% Confidence Intervals Using Different Regression Types and VIF Cut-off Values . . . 106

3.17 Coverage Rates of the Confidence Regions for the True Coefficient Values in the Artificial Population based on NHANES . . . 107

3.18 Ratios of the Average Estimated $se(\hat{\boldsymbol{\beta}})$ to Empirical $SE(\hat{\boldsymbol{\beta}})$ Using Different Regression Types and VIF Cut-off Values . . . 107

3.19 Ratios of the Average Estimated $se(\hat{\boldsymbol{\beta}})$ after Variable Elimination to the Average Estimated $se_6(\hat{\boldsymbol{\beta}})$ When All the Six Variables are in the Model . . . 108

4.1 Regression Analysis Output of SMHO Full Finite Population Data . . 131

4.2 Scaled Condition Indexes and Variance Decomposition Proportions: the Underlying Model in SMHO Full Finite Population Data . . . 131

4.3 Scaled Condition Indexes and Variance Decomposition Proportions: the Extended Model with Collinear Variables in SMHO Full Finite Population Data . . . 132

4.4 Regression Analysis Output of a Sample of SMHO Full Finite Population Data . . . 133

4.5 Scaled Condition Indexes and Variance Decomposition Proportions: the Underlying Model in a Sample of SMHO Full Finite Population Data . . . 134

4.6 Scaled Condition Indexes and Variance Decomposition Proportions: the Extended Model with Collinear Variables in a Sample of SMHO Full Finite Population Data . . . 135

4.7 Regression Analysis Output of NHANES Full Finite Population Data 140

4.8 Scaled Condition Indexes and Variance Decomposition Proportions: the underlying model in NHANES full finite population data set . . . 141

4.9 Scaled Condition Indexes and Variance Decomposition Proportions: the extended model with collinear variables in NHANES full finite population data set . . . 142

4.10 Regression Analysis Output of A Sample of NHANES Full Finite Population Data . . . 143

4.11 Scaled Condition Indexes and Variance Decomposition Proportions: the underlying model in a sample of NHANES full finite population . 144

4.12 Scaled Condition Indexes and Variance Decomposition Proportions: the extended model with collinear variables in a sample of NHANES full finite population 145

4.13 Regression Analysis Output: When Underweight is the Reference Category in the Model 148

4.14 Regression Analysis Output: When Normal/Overweight is the Reference Category in the Model 149

4.15 Largest Scaled Condition Indexes and Its Associated Variance Decomposition Proportions: When Underweight is the Reference Category in the Model 149

4.16 Largest Scaled Condition Indexes and Its Associated Variance Decomposition Proportions: When Underweight is the Reference Category in the Model 150

5.1 Characteristics of Some Common Univariate Distributions in the Exponential Family 167

5.2 Regression Models and their Collinearity Diagnostic Statistics used in this Experimental Study 180

5.3 Parameter Estimates with Their Associated Standard Errors in Two Models using Three Different Regression Methods 181

5.4 VIF values for Two Regression Models using Three Different Regression Methods 182

5.5 Scaled Condition Indexes and Variance Decomposition Proportions: the Identity Model using NHANES 2001-2002 Data File 183

5.6 Scaled Condition Indexes and Variance Decomposition Proportions: the Logistic Model using NHANES 2001-2002 Data File 184

List of Figures

3.1 Scatterplots and Correlation Coefficients of Five Explanatory Variables in the Artificial Population based on the SMHO Data Set . . . 74

3.2 Scatterplots and Correlation Coefficients of Five Variables in NHANES 2001-2002 data file 95

Chapter 1

Introduction

Over the last several decades, linear regression models, and generalized linear models, have played an important role in the analysis of experimental and observational data, due to their low computation costs, their intuitive plausibility in a wide variety of circumstances and their support by a broad and sophisticated body of statistical inference (Belsley et al., 1980). Given the data, a linear regression model seeks to describe an analysis variable as a function of some explanatory variables, which are assumed either as fixed numbers or random variables plus an error term. The errors are often assumed to be independent and normally distributed. Yet, in reality, these strong assumptions about the structure of data can be unreasonable and affect the validity and efficiency of the models and inferences about underlying parameters. Therefore, regression diagnostics are often needed to determine if the assumptions are reasonable, when a regression model is fitted for a given data set. These techniques have turned into an indispensable part of regression analysis.

1.1 Regression Diagnostics

Regression diagnostics are included in most of the statistical textbooks on linear models and have been extensively discussed in *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* by Belsley, Kuh, & Welsch

(1980) and *Regression Diagnostics* by Fox (1991). Many statistical packages, such as SAS®, SPSS®, Stata® and R®, also include diagnostic statistics as options in the linear regression modeling to aid analysts in locating influential points or measuring the presence and intensity of collinearity among the regression data.

Various sources of complexity can be encountered in a survey design. The sampling design, for instance, can involve multi-stage sampling and lead to positive intra-cluster correlation for a given study variable among units in the sample. Or, stratified sampling may be adopted, for which different mean and error variance-covariance structures may be appropriate in different strata. Unequal weights are often unavoidable in survey data due to varying inclusion probabilities and reweighting that accounts for nonresponse or noncontact adjustments, poststratification, and calibration adjustments. To account for the various complexities, analysts may use specialized methods in regression analysis for parameter and standard error estimation, as been discussed in some survey literature. Two of the most influential books are *Analysis of Complex Surveys* by Skinner et al. (1989) and *Analysis of Survey Data* by Chambers & Skinner (2003).

The need for diagnostic procedures may seem obvious, but survey literature gives limited attention to this problem. Skinner et al. (1989) and Chambers & Skinner (2003) mention diagnostics only in passing. Deville & Särndal (1992) and Potter (1990, 1993) discuss some possibilities for pinpointing or trimming extreme survey weights when the goal is estimating population totals and other simple descriptive statistics. Hulliger (1995) and Moreno-Rebollo & Muñoz Pichardo (1999) address the effect of outliers on the Horvitz-Thompson estimator of a population total. Outlier robust estimation techniques for totals are studied in Chambers (1996), Gwet & Rivest (1992), Welsh & Ronchetti (1998) and Duchesne (1999). A few references introduce some techniques for the evaluation of the quality of regression on complex survey data in the past decade. Elliot (2007), for instance, developed Bayesian methods for weight trimming of linear and generalized linear regression estimators in unequal probability-of-inclusion designs. Li (2007a,b); Li & Valliant (2006, 2009) adapted and extended a series of traditional diagnostic techniques to regression on complex survey data, mainly on identifying influential observations and influential groups of observations. Li's research covers residuals and leverages, DFBETA, DFBETAS, DFFIT, DFFITs, Cook's Distance and the forward search approach. However, none of this research touches upon diagnostics for collinearity when fitting models with survey data.

1.2 Collinearity Diagnostics

Collinearity of predictor variables in a linear regression refers to a situation where explanatory variables are correlated with each other. The terms, multicollinearity and ill conditioning are also used to denote the same situation. Collinearity is worrisome for both numerical and statistical reasons. The estimates of slope coefficients can be numerically unstable in some data sets in the sense that small changes in the $\boldsymbol{X}$'s or the $\boldsymbol{Y}$'s can produce large changes in the values of estimates. Correlation among the predictors can lead to slope estimates with large variances. In addition, when $\boldsymbol{X}$'s are strongly correlated, the R^2 in a regression can be large while the individual slope estimates are not statistically significant. Even if slope estimates are significant, they may have signs that are the opposite of what is expected. On the other hand, it is well known that collinearity need not harm forecasts, even if it has harmed structural estimation, as long as the pattern of collinearity continues into the forecast period. Belsley (1984a) conducted a case study to demonstrate that if, however, it is determined that collinearity exists that is unlikely to continue into the forecast period and has harmed structural estimates over the estimation period, then some means to improve structural estimates in line with prior information will likely result in more meaningful forecasts.

Sophisticated collinearity diagnostics for linear regression models have been developed over the last few decades to detect problems due to ill conditioned data. Conventional collinearity diagnostics are mainly aimed at ordinary or weighted least squares regressions through detecting the presence of correlated predictors and assessing the extent to which these relationships have degraded regression parameter estimates (Belsley et al., 1980). However, there has been only a limited amount of work on how to apply the method to the analysis of survey data, and on whether these tools need to be modified to account for complex sampling schemes and survey weights.

Conceptually, linear regression modeling brings statistical theory, discipline-specific theory and data together to increase our understanding of various phenomena and causal relationships. In experimental designs, it may be possible to create situations where the explanatory variables are orthogonal to each other. But, in observational data collected in surveys, predictors are almost always correlated with each other to some degree. This nonorthogonality leads to standard errors of slope estimates that are larger than would be the case with orthogonal predictors. Goals in the study of collinearity are to determine how serious this problem is, and, what, if anything, to do about it. An extensive literature in applied statistics provides valuable suggestions and guidelines for data analysts to diagnose the presence of collinearity (e.g. Farrar & Glauber 1967; Theil 1971; Belsley et al. 1980; Fox 1986). One of the most influential books is *Condition Diagnostics: Collinearity and Weak Data in Regression* by Belsley (1991). But relatively little research has been done to adapt those approaches or develop new ones in the analysis of survey data.

Weights for complex survey data account for unequal selection probabilities, nonresponse adjustments, poststratification, or other calibration adjustments. Chambers & Skinner (2003) reviewed the effects of ignoring informative sampling, in which the survey weights are correlated with the outcome variables and a model holding for the sample data is different from the model holding in the population. Our research will mainly focus on the collinearity diagnostics under informative sampling for analysts doing survey-weighted least squares regressions and survey-weighted generalized linear models. The goal of this research is to adapt and extend some of collinearity diagnostic techniques to account for sample designs and survey weights.

1.3 Regression Analysis with Complex Survey Data

Skinner et al. (1989) defined *complex survey data* as " the survey data arising from complex sampling schemes or reflecting associated underlying complex population structures." Some of the techniques that are used in selecting survey samples as stratification, clustering, selection in multiple stages, and sampling units selected with unequal probabilities. These features may also be related to structural features that need to be accounted for modeling. In the analysis of complex survey data, the OLS regression residuals and the Y's themselves may have different distributions in the sample than the ones in the finite population, which is due to substantial association of the Y's with some design variables. In general, sample designs possessing these properties are characterized as "informative" (see, e.g., Scott 1977). An analysis question is whether and how to take the informative design into account in the regression analysis. Little (2004) reviews some of the issues that should be considered when analyzing survey data.

Two uses of survey data are frequently addressed and contrasted in survey sampling literature: one is the descriptive uses for finite population parameters, like means or totals and another is the analytic uses for estimating model parameters. A basic distinction between them is the difference in the quantities to be estimated. Design-based and model-based approaches can be used for inference about either descriptive or analytic statistics. The design-based approach usually focuses on estimating summary measures for the finite population. The distribution used for inference is generated by the randomization mechanism used to select sample units. In the model-based approach, the distribution for inference is generated by a superpopulation model and only considers the random variation from the model. The probability distribution induced by the sampling design is treated as ignorable or non-informative (Rubin, 1976) when doing strictly model-based inference. However, some design features like stratification and clustering may be important to account for, even if an analyst has a purely model-based view. For example, the covariates in the model may include indicators for strata and model variances can account for clustering. When the design-based approach is applied to analytic inference, a model is also used since the parameters are always defined with respect to a certain model. To specify a parameter under design-based inference, a finite population parameter, θ_U, is defined corresponding to the model parameter θ.

Models may also be used in the design-based approach for improving the precision of estimates of descriptive parameters by including other auxiliary information, as is done with generalized regression estimators (Särndal, Swensson, & Wretman, 1992). Although models may be used to construct estimators, the distribution used for inference is the one generated by probability sampling. This hybrid approach is referred to as model-assisted. In model-based inference, the random variation needed for inference only comes from the distributions of the random effects and error terms in the model. However, when considering a statistical framework for a finite population, models with independent and identically distributed errors may be inappropriate when there is clustering or stratification in the population that is reflected in a complex sampling design. Thus, population features that are reflected in the sample design need to be considered in the model development, estimation and diagnostics. Elements of both the design-based and model-based approaches will be used in this thesis.

1.4 The Subject of This Dissertation

The remainder of this dissertation is organized as follows. Chapter 2 will review the conventional methods for diagnosing the presence of collinearity in regression analysis. Two approaches, variance inflation factors (VIFs) and condition indexes with variance decomposition method, are described for traditional linear models which is estimated by the ordinary least squares approach. Then, the maximum likelihood estimation for generalized linear models is reviewed and existing literature on collinearity diagnostics for generalized linear models is discussed. Chapter 3 and Chapter 4 will focus on modifying and adapting VIFs and condition indexes with the variance decomposition method to the survey setting. Estimates of VIFs developed here have both model-based and design-based justifications. The newly-adapted approaches are applied to real survey data and simulated data. Chapter 5 will extend these approaches to the class of generalized linear models. The logistic model will be taken as an example to illustrate how to apply these diagnostic statistics to a certain model. An experimental study is offered to investigate the difference and performance of traditional approaches and our new approaches. This study will conclude in Chapter 6 with a summary of limitations of the research, some possible remedies to the collinearity problems in complex survey data and suggestions for future research to advance this work. The new contributions in this dissertation are the adapted and modified collinearity diagnostic approaches which will be described in Chapters 3, 4 and 5. The development of these methods allows us to properly determine the presence of collinearity and evaluate its impact on linear regression and generalized linear regression using complex survey data.

Chapter 2

Review of Traditional Techniques

In many surveys, variables that are substantially correlated are collected for analysis. For example, total income and its components (e.g. wages and salaries, capital gains, interest and dividends) are collected in the Panel Survey of Income Dynamics (http://psidonline.isr.umich.edu/) to track economic well-being over time. When one explanatory variable is a linear combination of the others, this is known as perfect collinearity (or multicollinearity) and is easy to identify. Cases that are of interest in practice are ones where collinearity is less than perfect but still affects the precision of estimates (Kmenta 1986, sec.10.3). In this chapter, we will review some conventional collinearity diagnostic techniques that we will extend to the survey setting.

2.1 Collinearity Diagnostics in Ordinary Least Squares

Suppose the sample s has n units, on each of which p X's or predictors and one analysis variable Y are observed. The standard linear model in a nonsurvey setting is

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon} \tag{2.1}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of observations on a response or dependent variable; $\boldsymbol{X}$ is an $n \times p$ design matrix of fixed constants; $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters to be estimated; and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of statistically independent error terms with zero mean and constant variance σ^2 . We assume, for simplicity, that $\boldsymbol{X}$ has full column rank. The ordinary least squares (OLS) estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, for which the model variance is $Var_M(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Here, we use the subscript M to denote expectation under the model (and later, the subscript π to denote expectation under the design).

2.1.1 Variance Inflation Factor

Collinearities of explanatory variables inflate the model variance of the regression coefficients compared to having orthogonal $\boldsymbol{X}$'s. This effect can be seen in the formula for the variance of a specific estimated non-intercept coefficient $\hat{\beta}_k$ (Theil 1971),

$$Var_M(\hat{\beta}_k) = \frac{\sigma^2}{\sum_{i \in s} x_{ik}^2} \frac{1}{1 - R_k^2} \tag{2.2}$$

where R_k^2 is the square of the multiple correlation from the regression of the k^{th} column of $\boldsymbol{X}$ on the other columns. The term $\sigma^2 / \sum x_{ik}^2$ is the model variance of $\hat{\beta}_k$ if the k^{th} predictor were orthogonal to all other predictors, or equivalently, if there were only the k^{th} predictor in the regression. (Note that for the single-covariate interpretation to be correct, σ^2 must be the same in the full model and in the model containing only $\boldsymbol{x}_k$). The value of R_k^2 may be nonzero because the k^{th} predictor is correlated with one other explanatory variable or because of a more complex pattern of dependence between $\boldsymbol{x}_k$ and several other predictors. Consequently, the collinearity between $\boldsymbol{x}_k$ and some other explanatory variables can result in the inflation of the variance of $\hat{\beta}_k$ beyond what would be obtained with orthogonal $\boldsymbol{X}$'s. The second term in (2.2), $(1 - R_k^2)^{-1}$, is called the variance-inflation factor (VIF) (Theil, 1971).

In OLS, it is assumed that the residual variance is constant for all the values of the covariates. In some cases, it is desirable to weight cases differentially in a regression analysis to incorporate a nonconstant residual variance. This form of weighting is model-based and is called weighted least squares (WLS). Most of current statistical software packages, (e.g., SAS, STATA, S-Plus and R), use $(1-R^2_{k(WLS)})^{-1}$ as VIF for WLS, where $R^2_{k(WLS)}$ is the square of the multiple correlation from the WLS regression of the k^{th} column of $\boldsymbol{X}$ on the other columns. Fox & Monette (1992) also generalized this concept of variance inflation as a measure of collinearity to a subset of parameters in $\boldsymbol{\beta}$ and derived a *generalized variance-inflation factor* (GVIF). Furthermore, some interesting work has developed VIF-like measures, such as *collinearity indices* in Steward (1987) that are simply the square roots of the VIFs and *tolerance* defined as the inverse of VIF in Simon & Lesage (1988).

Although the derivation of VIF in OLS is relatively straightforward, opinions differ on how it should be used. Several rules of thumb have been proposed as signs of harmful collinearity. Marquardt (1970) treats a VIF of greater than 10 as a guideline for serious collinearity. The STATA manual (StataCorp 1997: 390) summarizing Chatterjee & Price (1991) says: "However, most analysts rely on informal rules of thumb applied to VIF. According to these rules, there is evidence of multi-collinearity if (1) the largest VIF is greater than 10 (some chose the more conservative threshold value of 30) or (2) the mean of all of the VIF's is considerably larger than 1." However, O'Brien (2007) examined several rules of thumb associated with VIF and found that threshold values of the VIF need to be evaluated in the context of several other factors that influence the stability of the estimates of the k^{th} regression coefficient. He especially pointed out the influence of collinearity on the estimation of σ^2 in (2.2) and relates it to the effect of the sample size and other effects that influence the variance of regression coefficients. Future research, beyond that in this dissertation, may be extended to this problem for survey data analysis.

2.1.2 Condition Indexes with Variance Decomposition

Belsley et al. (1980) applies numerical analysis techniques using condition indexes to signal the "near" dependencies in the data matrix $\boldsymbol{X}$. They used it in conjunction with the regression variance decomposition to not only uncover the variables causing collinearity, but also assess the degree to which the estimated coefficients are being degraded.

2.1.2.1 Eigenvalues and Eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$

When there is an exact (perfect) collinear relation in the $n \times p$ data matrix $\boldsymbol{X} \equiv (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$, we can find a set of values, $\boldsymbol{v} = (v_1, \ldots, v_p)$, not all zero, such that

$$v_1\boldsymbol{X}_1 + \cdots + v_p\boldsymbol{X}_p = \mathbf{0}, \quad \text{or } \boldsymbol{X}\boldsymbol{v} = \mathbf{0}. \tag{2.3}$$

However, in practice, when there exists no exact collinearity but some near dependencies in the data matrix, it may be possible to find one or more non-zero vector

$\boldsymbol{v}$('s) such that

$$\boldsymbol{Xv} = \boldsymbol{a}$$

with $\boldsymbol{a} \neq \boldsymbol{0}$ but small (close to $\boldsymbol{0}$). Alternatively, we might say that a near dependency exists if the length of vector $\boldsymbol{a}$, $\|\boldsymbol{a}\|$, is small. To normalize the problem of finding the set of $\boldsymbol{v}$'s that makes $\|\boldsymbol{a}\|$ small, we consider only $\boldsymbol{v}$ with unit length, that is, with $\|\boldsymbol{v}\| = 1$. Belsley (1991) discussed the connection of the eigenvalues and eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$ with the normalized vector $\boldsymbol{v}$ and $\|\boldsymbol{a}\|$. The minimum length $\|\boldsymbol{a}\|$ is simply the positive square root of the smallest eigenvalue of $\boldsymbol{X}^T\boldsymbol{X}$. The $\boldsymbol{v}$ that produces the $\boldsymbol{a}$ with minimum length must be the eigenvector of $\boldsymbol{X}^T\boldsymbol{X}$ that corresponds to the smallest eigenvalue.

2.1.2.2 Singular-Value Decomposition, Condition Number and Condition Indexes

The Singular-value decomposition (SVD) of matrix $\boldsymbol{X}$ is very closely allied to the eigensystem of $\boldsymbol{X}^T\boldsymbol{X}$, but with its own advantages. The $n \times p$ matrix $\boldsymbol{X}$ can be decomposed as $\boldsymbol{X} = \boldsymbol{U_1 D U_2}^T$, where $\boldsymbol{U_1}^T\boldsymbol{U_1} = \boldsymbol{U_2}^T\boldsymbol{U_2} = \boldsymbol{I}_p$ and $\boldsymbol{D} = diag(\mu_1, \ldots, \mu_p)$ is the diagonal matrix of singular values of $\boldsymbol{X}$, μ_k, $k = 1, \ldots, p$. Here, the three components in the decomposition are matrices with very special, highly exploitable properties: $\boldsymbol{U_1}$ is $n \times p$ (the same size as $\boldsymbol{X}$) and is column orthogonal; $\boldsymbol{U_2}$ is $p \times p$ and both row and column orthogonal; $\boldsymbol{D}$ is $p \times p$, nonnegative and diagonal. Belsley et al. (1980) felt that the SVD of $\boldsymbol{X}$ has several advantages over the eigensystem of $\boldsymbol{X}^T\boldsymbol{X}$, for the sake of both statistical usages and computational complexity. For statistical usages, $\boldsymbol{X}$ is the focus of our concern, not the cross-product matrix $\boldsymbol{X}^T\boldsymbol{X}$; besides, the lengths $\|\boldsymbol{a}\|$ of the linear combinations (2.3) of $\boldsymbol{X}$ that we seek to minimize above is properly defined in terms of the square roots of the eigenvalues of $\boldsymbol{X}^T\boldsymbol{X}$, that is to say, the singular values of $\boldsymbol{X}$. For computational complexity, in operating directly on the $n \times p$ data matrix $\boldsymbol{X}$, the singular value decomposition avoids the additional computational burden of forming $\boldsymbol{X}^T\boldsymbol{X}$, an operation involving np^2 unneeded sums and products and providing an unnecessary source of truncation error.

The condition number of $\boldsymbol{X}$ is defined as $\kappa(\boldsymbol{X}) = \mu_{max}/\mu_{min}$, where μ_{max} and μ_{min} are maximum and minimum singular values of $\boldsymbol{X}$. Condition indexes are defined as $\eta_k = \mu_{max}/\mu_k$. Empirically, if a value of κ or η exceeds a certain value, say, 10 to 30, it indicates that two or more columns of $\boldsymbol{X}$ have moderate or strong relations. But there is no absolute answer for "how small is small" or "how large is large " for condition indexes. Determination requires practical experience and depends on the purpose of each specific analysis. The simultaneous occurrence of several large η_k's is always remarkable for the existence of more than one near dependency.

One issue with SVD is whether the $\boldsymbol{X}$'s should be centered around their means. Marquardt (1980) states that the centering of observations removes the nonessential ill conditioning. In contrast, Belsley (1984b) argues that mean-centering typically masks the role of the constant term in any underlying near-dependencies. A typical case is a regression with dummy variables. For example, if gender is one of the independent variables in a regression and most of the cases are male (or female), then the dummy for gender can be strongly collinear with the intercept. The discussions following Belsley (1984b) illustrate the differences of opinion that occur among practitioners (Wood, 1984; Snee & Marquardt, 1984; Cook, 1984). Moreover, in linear regression analysis, the dummy variables can also play an important role as a possible source for multicollinearity. Wissmann et al. (2007) find that the multicollinearity with dummy variables may be reduced by choosing the correct reference category.

Another problem with the condition number is that it has its own scaling problems, see Steward (1987). By scaling down any column of $\boldsymbol{X}$, the condition number can be made arbitrarily large. This situation is known as *artificial ill-conditioning.* Belsley (1991) suggests to scale each column of the design matrix $\boldsymbol{X}$ using the Euclidean norm of each column before computing the condition number. This method is implemented in SAS and the package *perturb* of the statistical software R. Both use the root mean square of each column for scaling as its standard procedure. The condition number and condition indexes of the scaled matrix $\boldsymbol{X}$ are referred as *scaled condition number* and *scaled condition indexes* of the matrix $\boldsymbol{X}$. Similarly, the variance decomposition proportions relevant to the scaled $\boldsymbol{X}$ (which will be discussed in next section) can also be referred to as *scaled variance decomposition proportions.*

2.1.2.3 Variance Decomposition Method

To assess the extent to which near dependencies (i.e., having high condition indexes of $\boldsymbol{X}$ and $\boldsymbol{X}^T\boldsymbol{X}$) degrade the estimated variance of each regression coefficient, Belsley et al. (1980) reinterpreted and extended the work of Silvey (1969) by decomposing a coefficient variance into a sum of terms each of which is associated with a singular value. Recall that the model variance-covariance matrix of the OLS estimator $\hat{\boldsymbol{\beta}}$ is $Var_M(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Using the SVD, $\boldsymbol{X} = \boldsymbol{U_1}\boldsymbol{D}\boldsymbol{U_2}^T$, $Var_M(\hat{\boldsymbol{\beta}})$ can be written as:

$$Var_M(\hat{\boldsymbol{\beta}}) = \sigma^2[(\boldsymbol{U_1}\boldsymbol{D}\boldsymbol{U_2}^T)^T(\boldsymbol{U_1}\boldsymbol{D}\boldsymbol{U_2}^T)]^{-1} = \sigma^2\boldsymbol{U_2}\boldsymbol{D}^{-2}\boldsymbol{U_2}^T \tag{2.4}$$

and the k^{th} diagonal element in $Var_M(\hat{\boldsymbol{\beta}})$ is the estimated variance for the k^{th} coefficient, $\hat{\beta}_k$. According to (2.4), $Var_M(\hat{\beta}_k)$ can be expressed as:

$$Var(\hat{\beta}_k) = \sigma^2\Sigma_{j=1}^{p}\frac{u_{2kj}^2}{\mu_j^2} \tag{2.5}$$

where $\boldsymbol{U_2} = (u_{2kj})_{p\times p}$. Let $\phi_{kj} = \frac{u_{2kj}^2}{\mu_j^2}$, $\phi_k = \Sigma_{j=1}^{p}\phi_{kj}$ and $\boldsymbol{Q} = (\phi_{kj})_{p\times p} = (\boldsymbol{U}_2\boldsymbol{D}^{-1})\boldsymbol{\cdot}(\boldsymbol{U}_2\boldsymbol{D}^{-1})$, where $\boldsymbol{\cdot}$ is the Hadamard product. The variance-decomposition proportions are $\pi_{jk} = \phi_{jk}/\phi_k$, which is the proportion of the variance of the k^{th} regression coefficient associated with the j^{th} component of its decomposition in (2.5). Denote $\boldsymbol{\Pi} = (\pi_{jk})_{p\times p} = \boldsymbol{Q}^T\bar{\boldsymbol{Q}}^{-1}$, where $\bar{\boldsymbol{Q}}$ is the diagonal matrix with the row sums of $\boldsymbol{Q}$ on the main diagonal and 0 elsewhere.

In the variance decomposition (2.5), when other things are equal, a small singular value μ_j can lead to a large component of $Var(\hat{\beta}_k)$. However, if u_{2kj} is small too, then $Var(\hat{\beta}_k)$ may not be affected by a small μ_j. One extreme case is when $u_{2kj} = 0$. Suppose the k^{th} and j^{th} columns of $\boldsymbol{X}$ belong to separate orthogonal blocks. Let $\boldsymbol{X} \equiv [\boldsymbol{X}_1, \boldsymbol{X}_2]$ with $\boldsymbol{X}_1^T\boldsymbol{X}_2 = \mathbf{0}$ and let the singular-value decompositions of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ be given, respectively, as $\boldsymbol{X}_1 = \boldsymbol{U}_{1(1)}\boldsymbol{D}_{11}\boldsymbol{U}_{2(11)}^T$ and $\boldsymbol{X}_2 = \boldsymbol{U}_{1(2)}\boldsymbol{D}_{22}\boldsymbol{U}_{2(22)}^T$. Since $\boldsymbol{U}_{1(1)}$ and $\boldsymbol{U}_{1(2)}$ are the orthogonal bases for the space spanned by the columns of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ respectively, $\boldsymbol{X}_1^T\boldsymbol{X}_2 = \mathbf{0}$ implies $\boldsymbol{U}_{1(1)}^T\boldsymbol{U}_{1(2)} = \mathbf{0}$ and $\boldsymbol{U} \equiv [\boldsymbol{U}_{1(1)}, \boldsymbol{U}_{1(2)}]$ is column orthogonal. The singular value decomposition of $\boldsymbol{X}$ is simply $\boldsymbol{X} = \boldsymbol{U}_1\boldsymbol{D}\boldsymbol{U}_2^T$, with:

$$\boldsymbol{D} = \begin{bmatrix} \boldsymbol{D}_{11} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{D}_{22} \end{bmatrix} \tag{2.6}$$

and

$$\boldsymbol{U}_2 = \begin{bmatrix} \boldsymbol{U}_{2(11)} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{U}_{2(22)} \end{bmatrix}. \tag{2.7}$$

Thus $\boldsymbol{U}_{2(12)} = \mathbf{0}$. An analogous result clearly applies to any number of mutually orthogonal subgroups. Hence, if all the columns in $\boldsymbol{X}$ are orthogonal, all the $u_{2kj} = 0$ when $k \neq j$ and $\pi_{kj} = 0$ likewise.

The previous result implies that a high proportion of any variance can be associated with a large singular value even when there is no collinearity. The standard approach is to check a high condition index associated with a large proportions of the variance of two or more coefficients when diagnosing collinearity in linear regression models, since there must be two or more columns of $\boldsymbol{X}$ involved to make a near dependency. Belsley et al. (1980) suggested showing the matrix $\mathbf{\Pi}$ and condition indexes of $\boldsymbol{X}$ in a variance decomposition table as below. If two or more elements in the j^{th} row of matrix $\mathbf{\Pi}$ are relatively large and its associated condition index η_j

is large too, it signals that near dependencies are influencing regression estimates.

Condition	Proportions of variance			
Index	$Var_M(\hat{\beta}_1)$	$Var_M(\hat{\beta}_2)$	$\cdots$	$Var_M(\hat{\beta}_p)$
η_1	π_{11}	π_{12}	$\cdots$	π_{1p}
η_2	π_{21}	π_{22}	$\cdots$	π_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
η_p	π_{p1}	π_{p2}	$\cdots$	π_{pp}

2.2 Collinearity Diagnostics in Generalized Linear Model

The class of generalized linear models (GLMs) (McCullagh & Nelder, 1989; McCulloch & Searle, 2001) is a flexible generalization of ordinary least squares regression that allows the linear model to be related to the response variable via a link function and the magnitude of the variance of each measurement to be a function of its predicted value. Collinearity in GLMs can inflate variances of the estimated coefficients and cause poor prediction in certain regions of the regression space. Hauck & Donner (1977) also pointed out that collinearity may cause a nonsignificant Wald statistic even when the predictors are highly predictive in a logistic model and Væth (1985) noted that this effect of collinearity can happen for the other members of the family of the generalized linear models. In the past several decades, some literature discussed the collinearity problems in the logistic regression framework (see Schaefer et al., 1984; Schaefer, 1986; Marx & Smith, 1990b). Others explored this problem in the framework of GLMs (see Mackinnon & Puterman, 1990; Marx & Smith, 1990a; Weissfeld & Sereika, 1991; Lesaffre & Marx, 1993). All of these papers remark that collinearity in GLMs is not the collinear relations among the explanatory variables (or, equivalently, among the design matrix $\boldsymbol{X}$), but rather of weighted explanatory variables related to the observed information matrix. Lesaffre & Marx (1993), following a suggestion of Mackinnon & Puterman (1990), investigated the dependence of the ill conditioning problems on the particular value of $\boldsymbol{\beta}$. They concluded that in GLMs the response and the choice of the model also play a role in the degree of collinearity (ill-conditioning) of the information matrix. The term, ML-collinearity, is given in their paper to describe the scenario when the explanatory variables are not collinear, but at the maximum likelihood estimate of the model parameter there is collinearity among the weighted explanatory variables. These details will be given in the following sections.

2.2.1 Definitions and Notations

Suppose there are N observations in the population. The response variable $\boldsymbol{Y}_{N\times 1}$ is assumed to contain independent measurements from a distribution with mean μ_i and density from the exponential family or a family similar to exponential. That is, the density of y_i may be written as:

$$f_{Y_i}(y_i;\theta_i,\tau) = exp\left\{[y_i\theta_i - b(\theta_i)]/\tau^2 - c(y_i,\tau)\right\}, \quad i = 1,\ldots,N; \qquad (2.8)$$

where for convenience, we have written the distribution in what is called *canonical form.* Note that $E(y_i) = \mu_i = \partial b(\theta_i)/\partial\theta_i$ and $var(y_i) = \tau^2\partial^2 b(\theta_i)/\partial\theta_i^2 \equiv \tau^2 v(\mu_i)$, wherein we define $v(\mu_i)$ as $\partial^2 b(\theta_i)/\partial\theta_i^2$. Thus, the mean of y_i involves only the natural parameter θ_i, while τ denotes a nuisance scale parameter which is constant for all y_i (to conform to standard notation in the literature, we use μ_i to denote the mean of y_i. In the previous section μ denoted an eigenvalue of $\boldsymbol{X}$). Furthermore, to relate the parameters of the distribution to various predictors, define the linear predictor η_i, which is a transformation of the mean μ_i, by:

$$\eta_i = \dot{\boldsymbol{x}}_i \boldsymbol{\beta} = g(\mu_i) \tag{2.9}$$

where $\dot{\boldsymbol{x}}_i = (x_{1i}, ..., x_{pi})^T$ is the i^{th} row of the model matrix $\boldsymbol{X}_{N\times p}$, $\boldsymbol{\beta}$ is a p-column vector of parameters, $\eta_i = \theta_i$, and $g(\cdot)$ is a link function that is monotonic twice differentiable in the interior of an interval, $[\mu_{min}, \mu_{max}] = I_\mu \subset \Re$, where $\Re$ is the set of real numbers. $\boldsymbol{\beta}$ belongs to the parameter space P, which can be denoted as $P = \cap_i P_i$, $P_i = \{\boldsymbol{\beta} \in \Re^p | g^{-1}(\dot{\boldsymbol{x}}_i \boldsymbol{\beta}) \in I_\mu\}$. $\boldsymbol{\beta}$ belongs to the boundary of P, iff there is an i such that $g^{-1}(\dot{\boldsymbol{x}}_i \boldsymbol{\beta}) = \mu_{min}$ or μ_{max}. A model that can be written as in (2.9) is called a generalized linear model (GLM).

Maximum likelihood is used to estimate the regression parameters $\boldsymbol{\beta}$ and their functions, the linear predictors and the fitted values in succession (McCullagh & Nelder, 1989; McCulloch & Searle, 2001). Here, we will briefly review this method. Before we derive the maximum likelihood equations, two useful identities need to be addressed here:

$$\frac{\partial \theta_i}{\partial \mu_i} = \left(\frac{\partial \mu_i}{\partial \theta_i}\right)^{-1} = \left[\partial^2 b(\theta_i)/\partial \theta_i^2\right]^{-1} = \frac{1}{v(\mu_i)} \tag{2.10}$$

and, using the chain rule,

$$\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} = \frac{\partial \mu_i}{\partial g(\mu_i)} \frac{\partial g(\mu_i)}{\partial \boldsymbol{\beta}} = [\partial g(\mu_i)/\partial \mu_i]^{-1} \frac{\partial \dot{\boldsymbol{x}}_i \boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = [\partial g(\mu_i)/\partial \mu_i]^{-1} \dot{\boldsymbol{x}}_i. \tag{2.11}$$

The log-likelihood of the generalized linear model is the sum of the natural logarithm of each of the components defined by (2.8) over the N observations:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N} [y_i \theta_i - b(\theta_i)]/\tau^2 - \sum_{i=1}^{N} c(y_i, \tau). \tag{2.12}$$

By setting the first-order partials for $\boldsymbol{\beta}$ equal to zero, the maximum likelihood estimating (MLE) equations for $\boldsymbol{\beta}$ are given by:

$$\begin{aligned}
\frac{\partial \hat{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{1}{\tau^2} \sum_{i=1}^{N} \left[y_i \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} - \partial b(\theta_i)/\partial \theta_i \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \right] \\
&= \frac{1}{\tau^2} \sum_{i=1}^{N} (y_i - \mu_i) \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \\
&= \frac{1}{\tau^2} \sum_{i=1}^{N} (y_i - \mu_i) \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \\
&= \frac{1}{\tau^2} \sum_{i=1}^{N} \frac{(y_i - \mu_i)}{v(\mu_i) g'(\mu_i)} \dot{\boldsymbol{x}}_i \quad \text{using (2.10) and (2.11)} \\
&= \frac{1}{\tau^2} \sum_{i=1}^{N} (y_i - \mu_i) \gamma_i g'(\mu_i) \dot{\boldsymbol{x}}_i \\
&= \mathbf{0}^T
\end{aligned} \tag{2.13}$$

upon defining $\gamma_i = \{v(\mu_i)[g'(\mu_i)]^2\}^{-1}$ with $g'(\mu_i) = \partial g(\mu_i)/\partial \mu_i$.

We can also write this in matrix notation as

$$\frac{\partial \hat{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{1}{\tau^2}\boldsymbol{X}^T\boldsymbol{\Gamma}\boldsymbol{\Delta}(\boldsymbol{y}-\boldsymbol{\mu}) = \mathbf{0}^T, \tag{2.14}$$

with $\boldsymbol{\Gamma} = diag(\gamma_i)$, $\boldsymbol{\Delta} = diag(g'(\mu_i))$ and $\boldsymbol{\mu} = (\mu_i)_{n\times 1}$.

If $\boldsymbol{\beta}$ belongs to the boundary of P, there is at least one i, where $g^{-1}(\dot{\boldsymbol{x}}_i\boldsymbol{\beta}) = \mu_i = \mu_{min}$ or μ_{max}, as defined earlier. Then, $g'(\mu_i) = 0$, because g is a monotone and twice differentiable function in the interior of I_μ. Therefore, γ_i in (5.2) will be infinite and the MLE does not exist. This situation can occur when y_i is binary(0, 1) if all units are 0 or 1 for a particular configuration of $\boldsymbol{X}$'s.

2.2.2 Condition Indexes with Variance Decomposition in GLM

To derive the large-sample model variance of $\boldsymbol{\beta}$ and obtain its information matrix $\boldsymbol{I}(\boldsymbol{\beta})$, we can obtain the expected value of the second derivative of the log likelihood $l(\boldsymbol{\beta})$ at first:

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} = -\frac{1}{\tau^2}\boldsymbol{X}^T\boldsymbol{\Gamma}\boldsymbol{\Delta}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}^T} + \frac{1}{\tau^2}\boldsymbol{X}^T\frac{\partial \boldsymbol{\Gamma}\boldsymbol{\Delta}}{\partial \boldsymbol{\beta}^T}(\boldsymbol{y}-\boldsymbol{\mu}) \tag{2.15}$$

so that

$$\begin{aligned} -E\left[\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}\right] &= \frac{1}{\tau^2}\boldsymbol{X}^T\boldsymbol{\Gamma}\boldsymbol{\Delta}\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}^T} + \mathbf{0} \\ &= \frac{1}{\tau^2}\boldsymbol{X}^T\boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Delta}^{-1}\boldsymbol{X} \quad \text{using (2.11)} \\ &= \frac{1}{\tau^2}\boldsymbol{X}^T\boldsymbol{\Gamma}\boldsymbol{X}. \end{aligned} \tag{2.16}$$

Since

$$\begin{aligned} -E\left[\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\tau^2}\right] &= -E\left[\frac{\partial}{\partial\tau^2}\frac{1}{\tau^2}\boldsymbol{X}^T\boldsymbol{\Gamma}\boldsymbol{\Delta}(\boldsymbol{y}-\boldsymbol{\mu})\right] \\ &= \frac{1}{\tau^4}\boldsymbol{X}^T\boldsymbol{\Gamma}\boldsymbol{\Delta}E(\boldsymbol{y}-\boldsymbol{\mu}) \\ &= \mathbf{0}, \end{aligned} \tag{2.17}$$

the estimation of τ^2 does not affect the large-sample model variance of $\boldsymbol{\beta}$.

Under the model-based inference, the asymptotic variance-covariance matrix of $\boldsymbol{\beta}$ is (see McCulloch & Searle, 2001):

$$avar_M(\boldsymbol{\beta}) = [\boldsymbol{I}(\boldsymbol{\beta})]^{-1} = -E^{-1}\left[\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} + \frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\tau^2}\right] = \tau^2(\boldsymbol{X}^T\boldsymbol{\Gamma}\boldsymbol{X})^{-1} \tag{2.18}$$

where $avar_M$ stands for the limiting or asymptotic model variance.

In generalized linear models, we aim to estimate the parameters of interest, $\boldsymbol{\beta}$, so that we can also estimate the expectation of y_i, μ_i, conditional on $\dot{\boldsymbol{x}}_i$. Here we denote the estimated μ_i as $\hat{\mu}_i$, $\hat{\gamma}_i = \{v(\hat{\mu}_i)[g'(\hat{\mu}_i)]^2\}^{-1}$ with $g'(\hat{\mu}_i) = \partial g(\hat{\mu}_i)/\partial\hat{\mu}_i$, and $\hat{\boldsymbol{\Gamma}} = diag(\hat{\gamma}_i)$.

Lesaffre & Marx (1993) proved that when $\boldsymbol{X}^T\hat{\boldsymbol{\Gamma}}\boldsymbol{X}$ is singular, either $\boldsymbol{X}$ is not of full rank, or $\hat{\boldsymbol{\Gamma}}$ is not of full rank and $\hat{\boldsymbol{\beta}}$ belongs to the boundary of P, or both. If it is the first situation, there is an exact linear dependence among the explanatory variables. If it is the second situation, the MLE does not exist. In either of the two situations, there is an exact linear dependence in the constructed variables defined by the columns of $\hat{\boldsymbol{S}} = \hat{\boldsymbol{\Gamma}}^{1/2}\boldsymbol{X}$. The term, *ML-collinearity* is defined by Lesaffre &

Marx (1993), when the explanatory variables are not collinear but at the MLE there is collinearity among the constructed variables $\hat{\boldsymbol{S}}$.

Similar to the definition of condition indexes in linear regression, let $\hat{\lambda}_1, ..., \hat{\lambda}_p$ be the eigenvalues of $\boldsymbol{X}^T\hat{\boldsymbol{\Gamma}}\boldsymbol{X}$ in decreasing order. Define the information condition number $\kappa_{\Gamma X} = (\hat{\lambda}_1/\hat{\lambda}_p)^{1/2}$ and the information condition index, $\eta_{\Gamma X_j} = (\hat{\lambda}_1/\hat{\lambda}_j)^{1/2}$, $j = 1, ..., p$. Belsley & Oldford (1986) showed that this diagnostic can be used for log-likelihood conditioning. Belsley (1991) and Weissfeld & Sereika (1991) suggested using the condition number of the standardized $\boldsymbol{X}^T\hat{\boldsymbol{\Gamma}}\boldsymbol{X}$. While Marx & Smith (1990b) proposed another condition number based on $\hat{\Phi}^* = \hat{\boldsymbol{S}}^{*T}\hat{\boldsymbol{S}}^*$, with $\hat{\boldsymbol{S}}^*$ the centered and standardized version of the matrix $\hat{\boldsymbol{S}} = \hat{\boldsymbol{\Gamma}}^{1/2}\boldsymbol{X}$.

In view of the distinction between the two types of collinearity problems (collinearity among $\boldsymbol{X}$ and ML-collinearity), Lesaffre & Marx (1993) proposed to take $\eta_{\Gamma X_j}$ and $\kappa_{\Gamma X}$ as diagnostics for detecting ill-conditioned information with the columns of $\boldsymbol{X}$ standardized to unit length. They also recommended distinguishing the two types of collinearity by calculating the ratio $r_{\Gamma X} = \kappa_{\Gamma X}/\kappa_X$, where κ_X is the condition number of $\boldsymbol{X}$. It is suggested that if the ratio $r_{\Gamma X}$ is high, e.g. more than 5, and $\kappa_{\Gamma X} > 30$ there is ML-collinearity and if $\kappa_X > 30$ there is collinearity among the explanatory variables. A variance decomposition table, similar to the one for linear regression models in section 2.1.2.3, can also be used here to identify the collinearity relationships among the weighted explanatory variables, $\hat{\boldsymbol{S}}$.

Chapter 3

Variance Inflation Factor

In this chapter, we derive variance inflation factors for linear regression models fitted using survey weights. The approach used here is to begin with the model variance of a parameter estimator. The model variance is written in a way that displays the inflation of the variance of $\hat{\beta}_k$ due to nonorthogonality of the X predictors. We then construct estimates of the VIF that have either a model-based or design-based interpretation. Illustrations are given using data from a survey of mental health organizations and a survey of health and nutrition.

3.1 VIF in Survey Weighted Least Squares Regression

3.1.1 Survey-Weighted Least Squares Estimators

Suppose the underlying structural model in the superpopulation is $\boldsymbol{Y} = \boldsymbol{X}^T\boldsymbol{\beta} + \boldsymbol{e}$, where the error terms in the model have a general variance structure $\boldsymbol{e} \sim (0, \sigma^2\boldsymbol{V})$ with known $\boldsymbol{V}$ and σ^2. Define $\boldsymbol{W}$ to be the diagonal matrix of survey weights. We assume throughout that the survey weights are constructed in such a way that they can be used for estimating finite population totals. The survey weighted generalized least squares (SWGLS) estimator accounting for the general covariance matrix $\boldsymbol{V}$ is

$$\hat{\boldsymbol{\beta}}_{SWV} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{V}^{-1}\boldsymbol{Y}, \tag{3.1}$$

assuming $\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{V}^{-1}\boldsymbol{X}$ is invertible. The estimator in (3.1)was also recommended by Little (2004) as a way of accounting for both a design and the linear model. If $\boldsymbol{V}$ is not used in fitting, which is typical, the estimator is

$$\hat{\boldsymbol{\beta}}_{SW} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y},$$

which we denote as survey weighted least squares (SWLS). Fuller (2002) describes the properties of these estimators.

The estimate $\hat{\boldsymbol{\beta}}_{SWV}$ is approximately design-unbiased for the population parameter $\boldsymbol{B}_U = (\boldsymbol{X}_U^T\boldsymbol{V}_U^{-1}\boldsymbol{X}_U)^{-1}\boldsymbol{X}_U^T\boldsymbol{V}_U^{-1}\boldsymbol{Y}_U$, where N is the count of units in the finite population, the subscript U stands for the finite population, $\boldsymbol{Y}_U = (Y_1, ..., Y_N)^T$, $\boldsymbol{X}_U = (\boldsymbol{x}_1, ..., \boldsymbol{x}_p)$ with $\boldsymbol{x}_k$ the $N \times 1$ vector of values for covariate k, and $\sigma^2\boldsymbol{V}_U = var_M(\boldsymbol{Y}_U)$. Similarly, $\hat{\boldsymbol{\beta}}_{SW}$ is approximately design-unbiased for $\boldsymbol{B}_U = (\boldsymbol{X}_U^T\boldsymbol{X}_U)^{-1}$ $\boldsymbol{X}_U^T\boldsymbol{Y}_U$. Both $\hat{\boldsymbol{\beta}}_{SWV}$ and $\hat{\boldsymbol{\beta}}_{SW}$ are also model unbiased estimators of $\boldsymbol{\beta}$ under the model $\boldsymbol{Y} = \boldsymbol{X}^T\boldsymbol{\beta} + \boldsymbol{e}$ regardless of whether $Var_M(\boldsymbol{e}) = \sigma^2\boldsymbol{V}$ is specified correctly or not.

To diagnose collinearity, we can examine either the design-variance or the model-variance of the estimator of slope. As shown in the subsequent sections, these variances share some common terms and analyzing either provides generally similar information.

3.1.2 Model Variance of Coefficient Estimates

We first derive a VIF in SWLS regression. To simplify notation, we transform the design matrix as $\tilde{\boldsymbol{X}} = \boldsymbol{W}^{1/2}\boldsymbol{X}$ and the response vector as $\tilde{\boldsymbol{Y}} = \boldsymbol{W}^{1/2}\boldsymbol{Y}$. In accordance, the linear system $(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})\hat{\boldsymbol{\beta}}_{SW} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y}$ can be written as $(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})\hat{\boldsymbol{\beta}}_{SW} = \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}}$ and the parameter estimator is $\hat{\boldsymbol{\beta}}_{SW} = (\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}}$. We will also use $\tilde{\boldsymbol{e}} = \boldsymbol{W}^{1/2}\boldsymbol{e}$ with $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$ and $\hat{\tilde{\boldsymbol{e}}} = \boldsymbol{W}^{1/2}\hat{\boldsymbol{e}}$ with $\hat{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{SW}$.

The model variance of the parameter estimator $\hat{\boldsymbol{\beta}}_{SW}$, assuming $Var_M(\boldsymbol{e}) = \sigma^2\boldsymbol{V}$, can be expressed as

$$\begin{aligned} Var_M(\hat{\boldsymbol{\beta}}_{SW}) &= (\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T E(\tilde{\boldsymbol{e}}\tilde{\boldsymbol{e}}^T)\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1} \\ &= \sigma^2(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1} \\ &= \boldsymbol{A^{-1}BA^{-1}}\sigma^2 = \boldsymbol{G}\sigma^2 \end{aligned} \tag{3.2}$$

where $\boldsymbol{W^{1/2}VW^{1/2}} = \tilde{\boldsymbol{V}}$ and $E(\tilde{\boldsymbol{e}}\tilde{\boldsymbol{e}}^T) = Var_M(\tilde{\boldsymbol{e}}) = \sigma^2\boldsymbol{W^{1/2}VW^{1/2}} = \sigma^2\tilde{\boldsymbol{V}}$. We also define $\boldsymbol{A} = \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$, $\boldsymbol{B} = \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}\tilde{\boldsymbol{X}}$, and $\boldsymbol{G} = \boldsymbol{A^{-1}BA^{-1}}$.

The matrix of predictors can be written as $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, where $\boldsymbol{x}_k$ is the $n \times 1$ vector of values of explanatory variable k for the n sample units. If the columns of X are orthogonal, then $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} = diag(\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k)$ and $\boldsymbol{A}^{-1} = diag(1/\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k)$. The ij^{th} element of $\boldsymbol{G}$ then becomes $\tilde{\boldsymbol{x}}_i^T\tilde{\boldsymbol{V}}\tilde{\boldsymbol{x}}_j/(\tilde{\boldsymbol{x}}_i^T\tilde{\boldsymbol{x}}_i)^2$. Thus, when the $\boldsymbol{X}$'s are orthogonal, the model variance of $\hat{\beta}_{SW_k}$ is

$$Var_M(\hat{\beta}_{SW_k}) = \sigma^2\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{V}}\tilde{\boldsymbol{x}}_k/(\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k)^2, \tag{3.3}$$

a fact we will use later.

More generally, the model variance of $\hat{\beta}_{SW_k}$, the coefficient of the k^{th} explanatory variable, is

$$Var_M(\hat{\beta}_{SW_k}) = \boldsymbol{i}_k' Var_M(\hat{\boldsymbol{\beta}}_{SW})\boldsymbol{i}_k = \sigma^2 \boldsymbol{i}_k' \boldsymbol{G} \boldsymbol{i}_k \tag{3.4}$$

where $\boldsymbol{i}_k$ is a $p \times 1$ vector with 1 in position k and 0's elsewhere. Therefore, we are interested in the kth diagonal element of matrix $\boldsymbol{G}$ to explore the impact of other explanatory variables on $Var_M(\hat{\beta}_{SW_k})$.

3.1.3 The Coefficient of Multiple Correlation

As noted in Section 2.1.1, the multiple correlation coefficient, R_k, determines the VIF in OLS regression. To derive the VIF for survey-weighted (SW) regression, we will look into the multiple correlation coefficient in SWLS regression. Similar to the linear model variance analysis in OLS, we can show that:

$$\tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{Y}} = \hat{\boldsymbol{\beta}}_{SW}^T \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} \hat{\boldsymbol{\beta}}_{SW} + \hat{\tilde{\boldsymbol{e}}}^T \hat{\tilde{\boldsymbol{e}}}$$

The total sum of squares, $SST_{SW} = \tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{Y}}$, in SWLS can be partitioned as shown in Table 3.1. The sum of squares due to fitting a general mean SSM_{SW} is $SSM_{SW} = \hat{N}\bar{\tilde{Y}}^2$, where $\hat{N} = \sum_{i \in s} w_i$ and $\bar{\tilde{Y}}$ denotes $\sum_{i \in s} \tilde{Y}_i / \hat{N} = \sum_{i \in s} w_i Y_i / \hat{N}$, the weighted mean of $\boldsymbol{Y}$. The total sum of squares, corrected for the mean, is $SST_{SW_m} = SST_{SW} - SSM_{SW}$; the sum of squares for fitting the model, corrected for the mean, is $SSR_{SW_m} = SSR_{SW} - SSM_{SW} = \hat{\boldsymbol{\beta}}_{SW}^T \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} \hat{\boldsymbol{\beta}}_{SW} - \hat{N}\bar{\tilde{Y}}^2$, and the residual error sum of squares is defined as $SSE_{SW} = \hat{\tilde{\boldsymbol{e}}}^T \hat{\tilde{\boldsymbol{e}}} = SST_{SW_m} - SSR_{SW_m}$, which is also equal to $SST_{SW} - SSR_{SW}$.

Table 3.1: Partitioning the Total Sum of Squares

Type	Sum of Squares	Sum of Squares corrected for mean
Mean	$SSM_{SW} = \hat{N}\bar{\tilde{Y}}^2$	
Regression	$SSR_{SW} = \hat{\boldsymbol{\beta}}_{SW}^T \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} \hat{\boldsymbol{\beta}}_{SW}$	$SSR_{SW_m} = \hat{\boldsymbol{\beta}}_{SW}^T \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} \hat{\boldsymbol{\beta}}_{SW} - \hat{N}\bar{\tilde{Y}}^2$
Error	$SSE_{SW} = SST_{SW} - SSR_{SW}$	$SSE_{SW} = SST_{SW_m} - SSR_{SW_m}$
Total	$SST_{SW} = \tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{Y}}$	$SST_{SW_m} = \tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{Y}} - \hat{N}\bar{\tilde{Y}}^2$

The *multiple correlation coefficient* R_{SW} is the nonnegative square root of $R^2_{SW} = \frac{SSR_{SW}}{SST_{SW}} = \frac{\hat{\boldsymbol{\beta}}_{SW}^T \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} \hat{\boldsymbol{\beta}}_{SW}}{\tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{Y}}}$, which represents the proportion of $\tilde{\boldsymbol{Y}}^T \tilde{\boldsymbol{Y}}$ that is accounted for by the explanatory variables. To separate the effects of the mean (i.e. the *intercept*, or *constant term*) in the model, we also define the *coefficient of determination, corrected for the mean*, as $R^2_{SW_m} = SSR_{SW_m}/SST_{SW_m} = 1 - SSE_{SW}/SST_{SW_m}$.

3.1.4 Model-based VIF

In this section, our aim is to examine the underlying impact of collinearity on the diagonal elements of $Var_M(\hat{\boldsymbol{\beta}}_{SW})$ and obtain a new form of VIF from the model-based perspective. First, we will study the two featured matrices, $\boldsymbol{A}$ and $\boldsymbol{G}$ defined in (3.2). Our goal is to rewrite $Var_M(\hat{\beta}_{SW_k})$ in a way that reflects any correlation between $\tilde{\boldsymbol{x}}_k$ and the other independent variables.

Similar to the derivation of conventional OLS VIF in Theil (1971), the sum of squares and cross products matrix of the observation matrix $[\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{X}}]$ is:

$$\dot{\boldsymbol{A}}_{(p+1)\times(p+1)} = \begin{pmatrix} \tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{Y}} & \tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{X}} \\ \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}} & \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} \end{pmatrix}$$

When there is a unique solution for $\hat{\boldsymbol{\beta}}_{SW}$ in the standard linear model, $\dot{\boldsymbol{A}}$ is full-rank and thus has an inverse matrix that we will write as $\dot{\boldsymbol{A}}^{-1} = [\dot{a}^{ij}]$ where $i, j = 0, 1, ..., p$. Using the formula for the inverse of a partitioned matrix, the first element $\dot{a}^{00}$ of $\dot{\boldsymbol{A}}^{-1}$, can be shown to be (see *Appendix A*)

$$\begin{aligned} \dot{a}^{00} &= \frac{1}{\tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}}} = \frac{1}{\tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{Y}} - \boldsymbol{\beta}^T\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\boldsymbol{\beta}} \\ &= \frac{1}{SST_{SW} - SSR_{SW}} = \frac{1}{(1-R^2_{SW})SST_{SW}} = \frac{1}{(1-R^2_{SW})\tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{Y}}} \end{aligned} \tag{3.5}$$

Notice that the denominator is a sum of squared errors (SSE_{SW}) in the regression.

Analogously, if we run a regression of one of the explanatory variables $\tilde{\boldsymbol{x}}_k = \boldsymbol{W}^{1/2}\boldsymbol{x}_k$ on the $p-1$ other explanatory variables, the sum of squares and cross products matrix is $\boldsymbol{A} = \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$, which can be partitioned as

$$\boldsymbol{A}_{p\times p} = \begin{pmatrix} \tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k & \tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{X}}_{(k)} \\ \tilde{\boldsymbol{X}}_{(k)}^T\tilde{\boldsymbol{x}}_k & \tilde{\boldsymbol{X}}_{(k)}^T\tilde{\boldsymbol{X}}_{(k)} \end{pmatrix} \tag{3.6}$$

where the columns of $\tilde{\boldsymbol{X}}$ are reordered so that $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{x}}_k \ \tilde{\boldsymbol{X}}_{(k)})$ with $\tilde{\boldsymbol{X}}_{(k)}$ being the $n \times (p-1)$ matrix containing all columns except the k^{th} column of $\tilde{\boldsymbol{X}}$.

Parallel to (3.5), the upper-left element of $\boldsymbol{A}^{-1}$ is:

$$a^{kk} = \boldsymbol{i}_k^T \boldsymbol{A}^{-1} \boldsymbol{i}_k = \boldsymbol{i}_k^T (\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}})^{-1} \boldsymbol{i}_k = \frac{1}{(1 - R^2_{SW_{(k)}}) SST_{SW_{(k)}}} = \frac{1}{(1 - R^2_{SW_{(k)}}) \tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{x}}_k} \tag{3.7}$$

where $R^2_{SW_{(k)}} = \frac{\hat{\beta}^T_{SW_{(k)}} \tilde{\boldsymbol{X}}^T_{(k)} \tilde{\boldsymbol{X}}_{(k)} \hat{\beta}_{SW_{(k)}}}{SST_{SW_{(k)}}}$ with $\hat{\beta}_{SW_{(k)}} = (\tilde{\boldsymbol{X}}^T_{(k)} \tilde{\boldsymbol{X}}_{(k)})^{-1} \tilde{\boldsymbol{X}}^T_{(k)} \tilde{\boldsymbol{x}}_k$ is the coefficient of determination corresponding to the regression of $\boldsymbol{x}_k$ on the $p-1$ other explanatory variables. The term $SST_{SW_{(k)}} = \tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{x}}_k$, is the total sum of squares in this regression.

The term $(1 - R^2_{SW_{(k)}})^{-1}$ in (3.7) is the VIF that will be produced by standard statistical packages when a weighted least squares regression is run. Under the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim (0, \sigma^2 \boldsymbol{W}^{-1})$, expression (3.7) is equal to $Var_M(\hat{\beta}_{SW_k})/\sigma^2$. However, this is not appropriate for survey-weighted least squares regressions because the variance of $\hat{\boldsymbol{\beta}}_{SW}$ has the more complex form in (3.2).

The matrix $\boldsymbol{G} = \boldsymbol{A}^{-1} \boldsymbol{B} \boldsymbol{A}^{-1}$ can be expressed as:

$$\boldsymbol{G} = \begin{pmatrix} a^{kk} & \boldsymbol{a}^{k(k)} \\ \boldsymbol{a}^{(k)k} & \boldsymbol{A}^{(k)(k)} \end{pmatrix} \begin{pmatrix} b_{kk} & \boldsymbol{b}_{k(k)} \\ \boldsymbol{b}_{(k)k} & \boldsymbol{B}_{(k)(k)} \end{pmatrix} \begin{pmatrix} a^{kk} & \boldsymbol{a}^{k(k)} \\ \boldsymbol{a}^{(k)k} & \boldsymbol{A}^{(k)(k)} \end{pmatrix} \tag{3.8}$$

where the inverse matrix $\boldsymbol{A}^{-1} = [a^{hk}]$, $h, k = 1, ..., p$, $\boldsymbol{a}^{k(k)}$ is defined as the k^{th} row of $\boldsymbol{A}^{-1}$ excluding a^{kk}, $(a^{k1}, \ldots, a^{k(k-1)}, a^{k(k+1)}, \ldots, a^{kp})$, $\boldsymbol{a}^{(k)k} = [\boldsymbol{a}^{k(k)}]^T$ and $\boldsymbol{A}^{(k)(k)}$ is defined as the $(k-1) \times (k-1)$ part of matrix $\boldsymbol{A}^{-1}$ excluding the k^{th} row and column. The partitioned version of $\boldsymbol{B}$ is

$$\boldsymbol{B} = \begin{pmatrix} b_{kk} & \boldsymbol{b}_{k(k)} \\ \\ \boldsymbol{b}_{(k)k} & \boldsymbol{B}_{(k)(k)} \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{V}} \tilde{\boldsymbol{x}}_k & \tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{V}} \tilde{\boldsymbol{X}}_{(k)} \\ \\ \tilde{\boldsymbol{X}}_{(k)}^T \tilde{\boldsymbol{V}} \tilde{\boldsymbol{x}}_k & \tilde{\boldsymbol{X}}_{(k)}^T \tilde{\boldsymbol{V}} \tilde{\boldsymbol{X}}_{(k)} \end{pmatrix}. \tag{3.9}$$

By virtue of the symmetry of $\boldsymbol{A}$ and $\boldsymbol{B}$, the k^{th} diagonal element of $\boldsymbol{G}$ is

$$g^{kk} = a^{kk}(a^{kk} b_{kk} + 2\boldsymbol{b}_{k(k)} \boldsymbol{a}^{(k)k}) + \boldsymbol{a}^{(k)k^T} \boldsymbol{B}_{(k)(k)} \boldsymbol{a}^{(k)k}. \tag{3.10}$$

Using the partitioned inverse of matrix $\boldsymbol{A}$, which represents $(\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}})^{-1}$, it can be shown that (see *Appendix A*)

$$\boldsymbol{a}^{(k)k} = -a^{kk} (\tilde{\boldsymbol{X}}_{(k)}^T \tilde{\boldsymbol{X}}_{(k)})^{-1} \tilde{\boldsymbol{X}}_{(k)} \tilde{\boldsymbol{x}}_k = -a^{kk} \hat{\boldsymbol{\beta}}_{SW(k)}. \tag{3.11}$$

Substituting $a^{(k)k}$ in (3.10), g^{kk} can be compactly expressed in terms of a^{kk}, $\hat{\boldsymbol{\beta}}_{SW(k)}$ and the lower right component of matrix $\boldsymbol{B}$:

$$\begin{aligned} g^{kk} &= (a^{kk})^2 (b_{kk} - 2\boldsymbol{b}_{k(k)} \hat{\boldsymbol{\beta}}_{SW(k)} + \hat{\boldsymbol{\beta}}_{SW(k)}^T \boldsymbol{B}_{(k)(k)} \hat{\boldsymbol{\beta}}_{SW(k)}) \\ &= \left(\frac{1}{1 - R_{SW(k)}^2} \frac{1}{\tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{x}}_k} \right)^2 \times \\ &\quad (\tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{V}} \tilde{\boldsymbol{x}}_k - 2\tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{V}} \tilde{\boldsymbol{X}}_{(k)} \hat{\boldsymbol{\beta}}_{SW(k)} + \hat{\boldsymbol{\beta}}_{SW(k)}^T \tilde{\boldsymbol{X}}_{(k)} \tilde{\boldsymbol{V}} \tilde{\boldsymbol{X}}_{(k)} \hat{\boldsymbol{\beta}}_{SW(k)}) \\ &= \frac{(\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{X}}_{(k)} \hat{\boldsymbol{\beta}}_{SW(k)})^T \tilde{\boldsymbol{V}} (\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{X}}_{(k)} \hat{\boldsymbol{\beta}}_{SW(k)})}{\left[(1 - R_{SW(k)}^2) \tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{x}}_k\right]^2}. \end{aligned} \tag{3.12}$$

As shown in (3.5), the term in brackets in the denominator above is the sum of squared errors $SSE_{SW(k)}$ in SWLS, which can be rewritten as $(\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{X}}_{(k)} \hat{\boldsymbol{\beta}}_{SW(k)})^T (\tilde{\boldsymbol{x}}_k -$

$\tilde{\boldsymbol{X}}_{(k)}\hat{\boldsymbol{\beta}}_{SW(k)})$. The term g^{kk} can be then expressed in terms of a^{kk} with an adjustment (denoted as ζ_k) involving variance matrix $\tilde{\boldsymbol{V}}$:

$$
\begin{aligned}
g^{kk} &= \frac{(\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{X}}_{(k)}\hat{\boldsymbol{\beta}}_{SW(k)})^T \tilde{\boldsymbol{V}} (\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{X}}_{(k)}\hat{\boldsymbol{\beta}}_{SW(k)})}{(\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{X}}_{(k)}\hat{\boldsymbol{\beta}}_{SW(k)})^T (\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{X}}_{(k)}\hat{\boldsymbol{\beta}}_{SW(k)})} \frac{1}{1-R^2_{SW(k)}} \frac{1}{\tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{x}}_k} \\
&= \frac{\tilde{\boldsymbol{e}}_{xk}^T \tilde{\boldsymbol{V}} \tilde{\boldsymbol{e}}_{xk}}{\tilde{\boldsymbol{e}}_{xk}^T \tilde{\boldsymbol{e}}_{xk}} a^{kk} \\
&= \frac{\zeta_k}{1-R^2_{SW(k)}} \frac{1}{\tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{x}}_k} = \zeta_k a^{kk}
\end{aligned}
\tag{3.13}
$$

where $\tilde{\boldsymbol{e}}_{xk} = \tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{X}}_{(k)}\hat{\boldsymbol{\beta}}_{SW(k)}$ is the residual from regressing $\tilde{\boldsymbol{x}}_k$ on $\tilde{\boldsymbol{X}}_{(k)}$ and $\zeta_k = \frac{(\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{X}}_{(k)}\hat{\boldsymbol{\beta}}_{SW(k)})^T \tilde{\boldsymbol{V}} (\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{X}}_{(k)}\hat{\boldsymbol{\beta}}_{SW(k)})}{(\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{X}}_{(k)}\hat{\boldsymbol{\beta}}_{SW(k)})^T (\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{X}}_{(k)}\hat{\boldsymbol{\beta}}_{SW(k)})} = \frac{\tilde{\boldsymbol{e}}_{xk}^T \tilde{\boldsymbol{V}} \tilde{\boldsymbol{e}}_{xk}}{\tilde{\boldsymbol{e}}_{xk}^T \tilde{\boldsymbol{e}}_{xk}}$.

Consequently, in SWLS estimation, the model variance of $\hat{\beta}_{SWk}$ is the k^{th} diagonal element of $Var_M(\hat{\boldsymbol{\beta}}_{SW})$ and can be written as:

$$
\begin{aligned}
Var_M(\hat{\beta}_{SW_k}) &= g^{kk}\sigma^2 = \zeta_k a^{kk} \sigma^2 \\
&= \frac{\zeta_k}{1-R^2_{SW(k)}} \frac{\sigma^2}{\tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{x}}_k} \\
&= \frac{\zeta_k}{1-R^2_{SW(k)}} \frac{\tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{x}}_k}{\tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{V}} \tilde{\boldsymbol{x}}_k} \frac{\sigma^2 \tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{V}} \tilde{\boldsymbol{x}}_k}{(\tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{x}}_k)^2} \\
&= \frac{\zeta_k \varrho_k}{1-R^2_{SW(k)}} \frac{\sigma^2 \tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{V}} \tilde{\boldsymbol{x}}_k}{(\tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{x}}_k)^2}
\end{aligned}
\tag{3.14}
$$

where $\varrho_k = \frac{\tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{x}}_k}{\tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{V}} \tilde{\boldsymbol{x}}_k}$ and ζ_k, ϱ_k are two adjustment coefficients involving $\tilde{\boldsymbol{V}}$.

Notice that ζ_k can also be rewritten as $\zeta_k = \frac{\boldsymbol{e}_{xk}^T \boldsymbol{WVW} \boldsymbol{e}_{xk}}{\boldsymbol{e}_{xk}^T \boldsymbol{W} \boldsymbol{e}_{xk}}$, where $\boldsymbol{e}_{xk} = \boldsymbol{x}_k - \boldsymbol{X}_{(k)}\hat{\boldsymbol{\beta}}_{SW(k)}$ is the residual from SWLS regressing $\boldsymbol{x}_k$ on $\boldsymbol{X}_{(k)}$ and $\varrho_k = \frac{\tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{x}}_k}{\boldsymbol{x}_k^T \boldsymbol{WVW} \boldsymbol{x}_k}$. Hence, ζ_k and ϱ_k depends on $\boldsymbol{W}$ and $\boldsymbol{V}$.

Recall that when the column of $\boldsymbol{X}$ are orthogonal, in OLS (2.2), the model variance of $\hat{\beta}_k$ is $\sigma^2/\boldsymbol{x}_k^T\boldsymbol{x}_k$ and in WLS, when $\boldsymbol{V} = \boldsymbol{W}^{-1}$, the model variance of $\hat{\beta}_k$ can be expressed as $\sigma^2/\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k$. However, the model variance of $\hat{\beta}_{SW_k}$ in SWLS under orthogonality is $\sigma^2\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{V}}\tilde{\boldsymbol{x}}_k/(\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k)^2$ as shown in (3.3). Thus, the variance under orthogonality is inflated by $\frac{\zeta_k\varrho_k}{1-R^2_{SW(k)}}$ times when incorporating the other $p-1$ explanatory variables in SWLS. The model-based VIF in SWLS includes not only the multiple correlation coefficient $R^2_{SW(k)}$ but also two adjustment coefficients, ζ_k, ϱ_k that are not present in the OLS and WLS cases.

We can prove that the range of ζ_k and ϱ_k are related to the minimum and maximum singular values of $\tilde{\boldsymbol{V}}$ as below.

Define two vectors both of which have unit length, $\boldsymbol{e}^* = \frac{\tilde{\boldsymbol{e}}_{xk}}{\left(\tilde{\boldsymbol{e}}_{xk}^T\tilde{\boldsymbol{e}}_{xk}\right)^{1/2}}$ and $\boldsymbol{x}^* = \frac{\tilde{\boldsymbol{x}}_k}{\left[\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k\right]^{1/2}}$. Note that the symmetry of $\tilde{\boldsymbol{V}}$ implies that its singular-value decomposition takes the form $\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$, where $\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}_p$ and $\boldsymbol{D}$ is a diagonal matrix with nonnegative diagonal elements, $\mu_1, \mu_2, ..., \mu_p$, which are the singular values of $\tilde{\boldsymbol{V}}$. Now, ζ_k can then be written as

$$\zeta_k = \boldsymbol{e}^{*T}\tilde{\boldsymbol{V}}\boldsymbol{e}^* = \boldsymbol{e}^{*T}\boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{e}^* \equiv \boldsymbol{\iota}^T\boldsymbol{D}\boldsymbol{\iota},$$

where $\|\boldsymbol{\iota}\| \equiv \|\boldsymbol{U}^T\boldsymbol{e}^*\| = 1$, and hence,

$$\mu_{min}(\tilde{\boldsymbol{V}}) = \mu_{min}(\tilde{\boldsymbol{V}})\boldsymbol{\iota}^T\boldsymbol{\iota} \leq \zeta_k \leq \mu_{max}(\tilde{\boldsymbol{V}})\boldsymbol{\iota}^T\boldsymbol{\iota} = \mu_{max}(\tilde{\boldsymbol{V}}). \tag{3.15}$$

ϱ_k can also be written in the similar form:

$$\varrho_k = \frac{1}{\boldsymbol{x}^{*T}\tilde{\boldsymbol{V}}\boldsymbol{x}^*}. \tag{3.16}$$

Similar to the preceding procedures, the denominator of (3.16) is smaller than $\mu_{max}(\tilde{\boldsymbol{V}})$ and larger than $\mu_{min}(\tilde{\boldsymbol{V}})$, and hence,

$$\frac{1}{\mu_{max}(\tilde{\boldsymbol{V}})} \leq \varrho_k \leq \frac{1}{\mu_{min}(\tilde{\boldsymbol{V}})}. \tag{3.17}$$

Combining (3.15) and (3.17) together, the joint coefficient $\zeta_k\varrho_k$ is bounded in the range of:

$$\frac{\mu_{min}(\tilde{\boldsymbol{V}})}{\mu_{max}(\tilde{\boldsymbol{V}})} \leq \zeta_k\varrho_k \leq \frac{\mu_{max}(\tilde{\boldsymbol{V}})}{\mu_{min}(\tilde{\boldsymbol{V}})}.$$

Notice that when $\tilde{\boldsymbol{V}} = \boldsymbol{I}$, $\zeta_k = \varrho_k = 1$ and (3.14) reduces to $\frac{1}{1-R^2_{SW(k)}}\frac{\sigma^2}{\boldsymbol{x}_k^T\boldsymbol{W}\boldsymbol{x}_k}$, which is the model variance of the WLS estimates when $\boldsymbol{V}$ is diagonal and $\boldsymbol{W}$ is correctly specified as $\boldsymbol{W} = \boldsymbol{V}^{-1}$. In that unusual case, the VIF computed by software packages will be appropriate for SWLS. However, rarely will it be reasonable to think that $\boldsymbol{W} = \boldsymbol{V}^{-1}$ in survey estimation. If $\tilde{\boldsymbol{V}} \neq \boldsymbol{I}$, then ζ_k and ϱ_k are not equal to 1 and a specialized calculation of the VIF is still needed. When $\boldsymbol{V} = \boldsymbol{I}$, which is the usual application considered by analysts, $\tilde{\boldsymbol{V}} = \boldsymbol{W}$, $\zeta_k = \frac{\tilde{\boldsymbol{e}}_{xk}^T\boldsymbol{W}\tilde{\boldsymbol{e}}_{xk}}{\tilde{\boldsymbol{e}}_{xk}^T\tilde{\boldsymbol{e}}_{xk}}$, $\varrho_k = \frac{\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k}{\tilde{\boldsymbol{x}}_k^T\boldsymbol{W}\tilde{\boldsymbol{x}}_k}$ and $\frac{\mu_{min}(\tilde{\boldsymbol{V}})}{\mu_{max}(\tilde{\boldsymbol{V}})} = \frac{w_{min}}{w_{max}}$, where w_{min} is the minimum value of survey weights and w_{max} is the maximum value of survey weights. In this case, the range of $\zeta_k\varrho_k$ is bounded by $[\frac{w_{min}}{w_{max}}, \frac{w_{max}}{w_{min}}]$. When all the survey weights are equal to 1, $\zeta_k\varrho_k = 1$ and the VIF produced by standard software, $(1-R^2_{SW})^{-1}$, does not need to be adjusted in SWLS;

however, when the range of the survey weights is large, $\zeta_k \varrho_k$ can be very small or very large. In this case the VIF produced by standard software is not appropriate and a special calculation is needed. These facts will be shown in our experimental studies.

3.1.5 Intercept-Adjusted Model-based VIF

In the discussion above, the value of VIF measures the degree of variance inflation of the parameter estimator for the k^{th} explanatory variable caused by correlation with the other $p-1$ variables (which include the intercept if the model contains one). In other words, referring to its formulation (2.2) in OLS and (3.14) in SWLS, we always compare the model variances of β_k in the regression with all the other $p-1$ variables (full model) to the one with only k^{th} variable and no intercept (single variable with no intercept model). Cook (1984) observes that this may not be the most useful choice of reference model. In reality, analysts often include the intercept in the model and are more interested in evaluating the collinearity effects of variance inflation by comparing the full model to the regression with both the k^{th} variable and intercept (single variable with intercept model). For reference, we denote the three models as:

M1: Full Model:

$\boldsymbol{Y} = \boldsymbol{\beta_0} + \boldsymbol{x_1}\boldsymbol{\beta_1} + ... + \boldsymbol{x_p}\boldsymbol{\beta_p} + \boldsymbol{e}$;

M2: Single variable with intercept model:

$\boldsymbol{Y} = \boldsymbol{\beta_0} + \boldsymbol{x_k}\boldsymbol{\beta_k} + \boldsymbol{e}$;

M3: Single variable with no intercept model:

$\boldsymbol{Y} = \boldsymbol{x_k}\boldsymbol{\beta_k} + \boldsymbol{e}$;

where $\boldsymbol{x}_k$, $k = 1, ..., p$, is the $n \times 1$ vector for the k^{th} explanatory variable and $\boldsymbol{\beta}_0$ stands for the intercept. Note that a special case of M1 would be a model with no intercept.

In the previous section, the VIF is derived to assess the degree of variance inflation of coefficient $\boldsymbol{x_k}$ from M3 to M1, which includes the other explanatory variables and the intercept (if model has it). In this section, we are interested in measuring the variance inflation from M2 to M1.

First of all, using (3.14), we can obtain the variance of $\hat{\beta}_{SW_k}$ in M2 as: (see Appendix B1)

$$Var_{M2}(\hat{\beta}_{SW_k}) = \frac{(\tilde{\boldsymbol{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)^T \tilde{\boldsymbol{V}} (\tilde{\boldsymbol{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)}{(\tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{x}}_k - \hat{N}\bar{\tilde{x}}_k^2)^2} = \frac{(\tilde{\boldsymbol{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)^T \tilde{\boldsymbol{V}} (\tilde{\boldsymbol{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)}{SST^2_{SW_m(k)}} \tag{3.18}$$

where $\bar{\tilde{x}}_k = \sum_{i \in s} w_i x_{ki}/\hat{N}$.

Analogous to the Table 3.1 partitioning of the total sum of squares, the denominator of formula for g^{kk} (3.13) is $(\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{X}}_{(k)}\hat{\boldsymbol{\beta}}_{SW(k)})^T(\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{X}}_{(k)}\hat{\boldsymbol{\beta}}_{SW(k)}) = SSE_{SW(k)}$ of the SWLS regression of $\tilde{\boldsymbol{x}}_k$ on $\tilde{\boldsymbol{X}}_{(k)}$, which can also be written as $SST_{SW_{m(k)}} - SSR_{SW_{m(k)}} = (1 - R^2_{SW_{m(k)}})SST_{SW_{m(k)}}$. The term g^{kk} in $Var_{M1}(\hat{\beta}_{SW_k})$ can then be expressed in terms of a sum of squares corrected for the mean: $g^{kk} = \frac{\zeta_k}{1-R^2_{SWm(k)}} \frac{1}{SST_{SW_{m(k)}}}$ with ζ_k defined the same as in (3.13). To decompose the model variance of $\hat{\beta}_{SWk}$ under M1 into a new form depending on $Var_{M2}(\hat{\beta}_{SW_k})$ and adjustment terms as we did in the previous section, we insert the new form of g^{kk} in

(3.14) to obtain:

$$
\begin{aligned}
Var_{M1}(\hat{\beta}_{SW_k}) &= \boldsymbol{g}^{kk}\sigma^2 \\
&= \frac{\zeta_k}{1-R^2_{SWm(k)}}\frac{\sigma^2}{SST_{SW_{m(k)}}} \\
&= \frac{\zeta_k}{1-R^2_{SWm(k)}}\frac{SST_{SW_{m(k)}}}{(\tilde{\boldsymbol{x}}_k-\tilde{\mathbf{1}}\bar{\tilde{x}}_k)^T\tilde{\boldsymbol{V}}(\tilde{\boldsymbol{x}}_k-\tilde{\mathbf{1}}\bar{\tilde{x}}_k)}\frac{\sigma^2(\tilde{\boldsymbol{x}}_k-\tilde{\mathbf{1}}\bar{\tilde{x}}_k)^T\tilde{\boldsymbol{V}}(\tilde{\boldsymbol{x}}_k-\tilde{\mathbf{1}}\bar{\tilde{x}}_k)}{SST^2_{SW_{m(k)}}} \\
&= \frac{\zeta_k}{1-R^2_{SWm(k)}}\frac{(\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k-\hat{N}\bar{\tilde{x}}_k^2)}{(\tilde{\boldsymbol{x}}_k-\tilde{\mathbf{1}}\bar{\tilde{x}}_k)^T\tilde{\boldsymbol{V}}(\tilde{\boldsymbol{x}}_k-\tilde{\mathbf{1}}\bar{\tilde{x}}_k)}\frac{\sigma^2(\tilde{\boldsymbol{x}}_k-\tilde{\mathbf{1}}\bar{\tilde{x}}_k)^T\tilde{\boldsymbol{V}}(\tilde{\boldsymbol{x}}_k-\tilde{\mathbf{1}}\bar{\tilde{x}}_k)}{(\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k-\hat{N}\bar{\tilde{x}}_k^2)^2} \\
&= \frac{\zeta_k\varrho_{mk}}{1-R^2_{SWm(k)}}\frac{\sigma^2(\tilde{\boldsymbol{x}}_k-\tilde{\mathbf{1}}\bar{\tilde{x}}_k)^T\tilde{\boldsymbol{V}}(\tilde{\boldsymbol{x}}_k-\tilde{\mathbf{1}}\bar{\tilde{x}}_k)}{(\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k-\hat{N}\bar{\tilde{x}}_k^2)^2} \\
&= \frac{\zeta_k\varrho_{mk}}{1-R^2_{SWm(k)}}Var_{M2}(\hat{\beta}_{SW_k})
\end{aligned}
\tag{3.19}
$$

where ζ_k was defined in (3.13), while ϱ_{mk} is changed into a new form, $\varrho_{mk} = \frac{(\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k-\hat{N}\bar{\tilde{x}}_k^2)}{(\tilde{\boldsymbol{x}}_k-\tilde{\mathbf{1}}\bar{\tilde{x}}_k)^T\tilde{\boldsymbol{V}}(\tilde{\boldsymbol{x}}_k-\tilde{\mathbf{1}}\bar{\tilde{x}}_k)}$, that is corrected for the mean $\bar{\tilde{x}}_k$.

As in the previous section, we can bound the VIF in (3.19). Define a vector with unit length, $\boldsymbol{x}_m^* = \frac{(\tilde{\boldsymbol{x}}_k-\tilde{\mathbf{1}}\bar{\tilde{x}}_k)}{\left[(\tilde{\boldsymbol{x}}_k-\tilde{\mathbf{1}}\bar{\tilde{x}}_k)^T(\tilde{\boldsymbol{x}}_k-\tilde{\mathbf{1}}\bar{\tilde{x}}_k)\right]^{1/2}}$. ϱ_{mk} can be written as

$$\varrho_{mk} = \frac{1}{\boldsymbol{x}_m^{*T}\tilde{\boldsymbol{V}}\boldsymbol{x}_m^*}. \tag{3.20}$$

Similar to the preceding section,

$$\frac{1}{\mu_{max}(\tilde{\boldsymbol{V}})} \leq \varrho_{mk} \leq \frac{1}{\mu_{min}(\tilde{\boldsymbol{V}})}, \tag{3.21}$$

and

$$\frac{\mu_{min}(\tilde{\boldsymbol{V}})}{\mu_{max}(\tilde{\boldsymbol{V}})} \leq \zeta_k \varrho_{mk} \leq \frac{\mu_{max}(\tilde{\boldsymbol{V}})}{\mu_{min}(\tilde{\boldsymbol{V}})}.$$

The model variance of $\hat{\beta}_{SW_k}$ is inflated by $\frac{\zeta_k \varrho_{mk}}{1-R^2_{SWm(k)}}$ compared to its variance in the model with only the explanatory variable $\tilde{\boldsymbol{x}}_k$ and intercept (M2). The new intercept-adjusted VIF retains some properties of the original VIF from previous section. The conventional intercept-adjusted VIF in OLS is the variance inflation factor for comparing M2 to M1, which is equal to $(1 - R^2_{m(k)})^{-1}$ (see Appendix B2), where $R^2_{m(k)}$ is the coefficient of determination, corrected for the mean and similar to $R^2_{SW_m}$ in Section 3.1.3. When $\tilde{\boldsymbol{V}} = \boldsymbol{I}$, $\zeta_k = 1$, $\varrho_{mk} = 1$ and the intercept-adjusted VIF in (3.19) for SWLS is equal to the conventional intercept-adjustment VIF. When $\boldsymbol{V} = \boldsymbol{I}$, which is the application used by most analysts, $\tilde{\boldsymbol{V}} = \boldsymbol{W}$, $\zeta_k = \frac{\tilde{\boldsymbol{e}}_{xk} \boldsymbol{W} \tilde{\boldsymbol{e}}_{xk}}{\tilde{\boldsymbol{e}}_{xk} \tilde{\boldsymbol{e}}_{xk}}$, $\varrho_{mk} = \frac{(\tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{x}}_k - \hat{N} \bar{\tilde{x}}_k^2)}{(\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{1}} \bar{\tilde{x}}_k)^T \boldsymbol{W} (\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{1}} \bar{\tilde{x}}_k)}$ and $\frac{\mu_{min}(\tilde{\boldsymbol{V}})}{\mu_{max}(\tilde{\boldsymbol{V}})} = \frac{w_{min}}{w_{max}}$. The range of $\zeta_k \varrho_{mk}$ also depends on the range of survey weights as did $\zeta_k \varrho_k$.

3.1.6 Estimating the VIF with Known $\boldsymbol{V}$ in a Sample Selected from the Finite Population

To derive a design-based VIF, we will use an approach similar to the one for pseudo-maximum likelihood least squares (PMLEs). The model-variance of $\hat{\beta}_{SWk}$ when the full finite population is in the sample will be the starting point. We will then substitute design-based estimates of each the components of the model variance.

In the previous section, expressions were given for $Var_M(\hat{\beta}_k)$ in (3.14) and

(3.19). When the entire finite population is in the sample, the diagonal matrix of unit survey weights, $\boldsymbol{W}$, is equal to the $N \times N$ identity matrix. The model-based variance of population coefficient $\hat{\boldsymbol{\beta}}_{U(k)}$ can be expressed as:

$$\begin{aligned} Var_M(\hat{\boldsymbol{\beta}}_{U(k)}) &= \frac{\zeta_{Uk}\varrho_{Uk}}{1-R^2_{U(k)}}\frac{\sigma^2\boldsymbol{x}^T_{Uk}\boldsymbol{V}\boldsymbol{x}_{Uk}}{(\boldsymbol{x}^T_{Uk}\boldsymbol{x}_{Uk})^2} \\ &= \frac{\zeta_{Uk}\varrho_{Umk}}{1-R^2_{Um(k)}}\frac{\sigma^2(\boldsymbol{x}_{Uk}-\mathbf{1}_N\bar{x}_{Uk})^T\boldsymbol{V}(\boldsymbol{x}_{Uk}-\mathbf{1}_N\bar{x}_{Uk})}{(\boldsymbol{x}^T_{Uk}\boldsymbol{x}_{Uk}-N\bar{x}^2_{Uk})^2} \end{aligned} \tag{3.22}$$

where $\boldsymbol{x}_{Uk} = (x_{k1}, \ldots, x_{kN})^T$ is the population vector for k^{th} variable, $\mathbf{1}_N$ is a column vector consisting of N unit elements, $\boldsymbol{e}_{Uxk} = \boldsymbol{x}_{Uk} - \boldsymbol{X}_{U(k)}\hat{\boldsymbol{\beta}}_{U(k)}$,

$$\zeta_{Uk} = \frac{(\boldsymbol{x}_{Uk}-\boldsymbol{X}_{U(k)}\hat{\boldsymbol{\beta}}_{U(k)})^T\boldsymbol{V}(\boldsymbol{x}_{Uk}-\boldsymbol{X}_{U(k)}\hat{\boldsymbol{\beta}}_{U(k)})}{(\boldsymbol{x}_{Uk}-\boldsymbol{X}_{U(k)}\hat{\boldsymbol{\beta}}_{SW(k)})^T(\boldsymbol{x}_{Uk}-\tilde{\boldsymbol{X}}_{U(k)}\hat{\boldsymbol{\beta}}_{U(k)})} = \frac{\boldsymbol{e}^T_{Uxk}\boldsymbol{V}\boldsymbol{e}_{Uxk}}{\boldsymbol{e}^T_{Uxk}\boldsymbol{e}_{Uxk}},\ \varrho_k = \frac{\boldsymbol{x}^T_{Uk}\boldsymbol{x}_{Uk}}{\boldsymbol{x}^T_{Uk}\boldsymbol{V}\boldsymbol{x}_{Uk}},$$

$$R^2_{U(k)} = \frac{\hat{\boldsymbol{\beta}}^T_{U(k)}\boldsymbol{X}^T_{U(k)}\boldsymbol{X}_{U(k)}\hat{\boldsymbol{\beta}}_{U(k)}}{\boldsymbol{x}^T_{Uk}\boldsymbol{x}_{Uk}},$$

and their intercept-adjusted forms,

$$\varrho_{mk} = \frac{(\boldsymbol{x}^T_{Uk}\boldsymbol{x}_{Uk}-N\bar{x}^2_{Uk})}{(\boldsymbol{x}_{Uk}-\mathbf{1}_N\bar{x}_{Uk})^T\boldsymbol{V}(\boldsymbol{x}_{Uk}-\mathbf{1}_N\bar{x}_{Uk})} = \frac{(\boldsymbol{x}_{Uk}-\mathbf{1}_N\bar{x}_{Uk})^T(\boldsymbol{x}_{Uk}-\mathbf{1}_N\bar{x}_{Uk})}{(\boldsymbol{x}_{Uk}-\mathbf{1}_N\bar{x}_{Uk})^T\boldsymbol{V}(\boldsymbol{x}_{Uk}-\mathbf{1}_N\bar{x}_{Uk})},$$

$$R^2_{Um(k)} = \frac{\hat{\boldsymbol{\beta}}^T_{U(k)}\boldsymbol{X}^T_{U(k)}\boldsymbol{X}_{U(k)}\hat{\boldsymbol{\beta}}_{U(k)}}{(\boldsymbol{x}^T_{Uk}\boldsymbol{x}_{Uk}-N\bar{x}^2_{Uk})} = \frac{\hat{\boldsymbol{\beta}}^T_{U(k)}\boldsymbol{X}^T_{U(k)}\boldsymbol{X}_{U(k)}\hat{\boldsymbol{\beta}}_{U(k)}}{(\boldsymbol{x}_{Uk}-\mathbf{1}_N\bar{x}_{Uk})^T(\boldsymbol{x}_{Uk}-\mathbf{1}_N\bar{x}_{Uk})}.$$

Thus, (3.22) is a finite population parameter consisting of totals that can be estimated using design-based approach methods. To derive the design-based estimates of the VIF in (3.22), $\frac{\zeta_{Uk}\varrho_{Uk}}{1-R^2_{U(k)}}$ (or its intercept-adjustment version, $\frac{\zeta_{Uk}\varrho_{Umk}}{1-R^2_{Um(k)}}$), we will derive the design-based estimates of its three components: ζ_{Uk}, ϱ_{Uk} (or ϱ_{Umk})

and $R^2_{U(k)}$ (or $R^2_{Um(k)}$) at first.

The numerator of ζ_{Uk} in (3.22) is given by:

$$\boldsymbol{e}_{Uxk}^T \boldsymbol{V} \boldsymbol{e}_{Uxk} = \begin{pmatrix} e_{xk1} & \dots & e_{xkN} \end{pmatrix} \begin{pmatrix} v_{11} & \cdots & v_{1N} \\ \cdots & \cdots & \cdots \\ v_{N1} & \cdots & v_{NN} \end{pmatrix} \begin{pmatrix} e_{xk1} \\ \vdots \\ e_{xkN} \end{pmatrix} \tag{3.23}$$
$$= \sum_{i=1}^{N} \sum_{j=1}^{N} e_{xki} v_{ij} e_{xkj}$$

where e_{xki} is the i^{th} element of $\boldsymbol{e}_{Uxk}$.

Suppose we know the sampling probability for the i^{th} unit as π_i and the joint selection probability of units i and j as π_{ij}, or equivalently, $w_i = w_{ii} = \pi_i^{-1} = \pi_{ii}^{-1}$ and $w_{ij} = \pi_{ij}^{-1}$. Define the $n \times n$ sample matrix $\boldsymbol{W}^* = (w_{ij})$. A design-based estimate of (3.23) is:

$$\sum_{i \in s} \sum_{j \in s} \frac{1}{\pi_{ij}} e_{xki} v_{ij} e_{xkj} = \sum_{i \in s} \sum_{j \in s} e_{xki} v_{ij}^* e_{xkj} = \boldsymbol{e}_{xk}^T \boldsymbol{V}^* \boldsymbol{e}_{xk}, \tag{3.24}$$

where $\boldsymbol{V}^* = (v_{ij}^*)_{n \times n} = (v_{ij}/\pi_{ij})_{n \times n} = (v_{ij} w_{ij})_{n \times n} = \boldsymbol{W^*} \cdot \boldsymbol{V}$ with the dot again denoting Hadamard product. If $\boldsymbol{V}$ is diagonal, then $\boldsymbol{e}_{Uxk}^T \boldsymbol{V} \boldsymbol{e}_{Uxk} = \sum_{i \in U} e_{xki}^2 v_i$, which can be estimated by $\boldsymbol{e}_{xk}^T \tilde{\boldsymbol{V}} \boldsymbol{e}_{xk}$, with $\tilde{\boldsymbol{V}} = \boldsymbol{W}\boldsymbol{V} = diag(w_i v_i),\ i \in s$.

Analogously, the denominator of ζ_{Uk} in (3.22) has its design-based estimate as: $\sum_{i \in s} w_i e_{xki}^2 = \boldsymbol{e}_{xk}^T \boldsymbol{W} \boldsymbol{e}_{xk}$, since $\boldsymbol{e}_{Uxk}^T \boldsymbol{e}_{Uxk} = \sum_{i \in U} e_{Uxki}^2$. Hence, the design-based estimate of ζ_{UK} is:

$$\hat{\zeta}_k = \frac{\boldsymbol{e}_{xk}^T \boldsymbol{V}^* \boldsymbol{e}_{xk}}{\boldsymbol{e}_{xk}^T \boldsymbol{W} \boldsymbol{e}_{xk}} = \frac{\boldsymbol{e}_{xk}^T \boldsymbol{W^*} \cdot \boldsymbol{V} \boldsymbol{e}_{xk}}{\boldsymbol{e}_{xk}^T \boldsymbol{W} \boldsymbol{e}_{xk}}. \tag{3.25}$$

Similarly, we can obtain the design-based estimate of ϱ_{Uk}:

$$\hat{\varrho}_k = \frac{\boldsymbol{x}_k^T \boldsymbol{W} \boldsymbol{x}_k}{\boldsymbol{x}_k^T \boldsymbol{V}^* \boldsymbol{x}_k} = \frac{\boldsymbol{x}_k^T \boldsymbol{W} \boldsymbol{x}_k}{\boldsymbol{x}_k^T \boldsymbol{W}^* \boldsymbol{\cdot} \boldsymbol{V} \boldsymbol{x}_k}, \tag{3.26}$$

and since the design-based estimate of $\bar{x}_{Uk}$ is $\bar{\bar{x}}_k = \frac{\sum_{i \in s} w_i x_{ki}}{\sum_{i \in s} w_i} = \frac{\sum_{i \in s} w_i x_{ki}}{\hat{N}}$ that is defined after (3.18), the design-based estimate of ϱ_{Umk} is:

$$\hat{\varrho}_{mk} = \frac{(\boldsymbol{x}_k - \boldsymbol{1}_n \bar{\bar{x}}_k)^T \boldsymbol{W} (\boldsymbol{x}_k - \boldsymbol{1}_n \bar{\bar{x}}_k)}{(\boldsymbol{x}_k - \boldsymbol{1}_n \bar{\bar{x}}_k)^T \boldsymbol{V}^* (\boldsymbol{x}_k - \boldsymbol{1}_n \bar{\bar{x}}_k)} = \frac{\boldsymbol{x}_k^T \boldsymbol{W} \boldsymbol{x}_k - \hat{N} \bar{\bar{x}}_k^2}{(\boldsymbol{x}_k - \boldsymbol{1}_n \bar{\bar{x}}_k)^T \boldsymbol{W}^* \boldsymbol{\cdot} \boldsymbol{V} (\boldsymbol{x}_k - \boldsymbol{1}_n \bar{\bar{x}}_k)}, \tag{3.27}$$

where $\boldsymbol{1}_n$ is a column vector consisting of n unit elements.

In the numerator of $R^2_{U(k)}$:

$$\boldsymbol{X}_{U(k)}^T \boldsymbol{X}_{U(k)} = \begin{pmatrix} \boldsymbol{x}_{U1}^T \\ \vdots \\ \boldsymbol{x}_{Up}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{x}_{U1} & \dots & \boldsymbol{x}_{Up} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_{U1}^T \boldsymbol{x}_{U1} & \dots & \boldsymbol{x}_{U1}^T \boldsymbol{x}_{Up} \\ \dots & \dots & \dots \\ \boldsymbol{x}_{Up}^T \boldsymbol{x}_{U1} & \dots & \boldsymbol{x}_{Up}^T \boldsymbol{x}_{Up} \end{pmatrix} \tag{3.28}$$

where column k of $\boldsymbol{X}_U$ is implicitly omitted. The jl^{th} element of the matrix is $\boldsymbol{x}_{Uj}^T \boldsymbol{x}_{Ul} = \sum_{i \in U} x_{ji} x_{li}$, whose design-based estimate is $\sum_{i \in s} w_i x_{ji} x_{li} = \boldsymbol{x}_j^T \boldsymbol{W} \boldsymbol{x}_k$. Accordingly, $\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}$ is a design-based estimator of $\boldsymbol{X}_{U(k)}^T \boldsymbol{X}_{U(k)}$. So, a design-based estimate of the numerator of $R^2_{U(k)}$ would be: $\hat{\boldsymbol{\beta}}_{SW(k)}^T \boldsymbol{X}_{(k)}^T \boldsymbol{W} \boldsymbol{X}_{(k)} \hat{\boldsymbol{\beta}}_{SW(k)}$, where $\hat{\boldsymbol{\beta}}_{SW(k)} = (\boldsymbol{X}_{(k)}^T \boldsymbol{W} \boldsymbol{X}_{(k)})^{-1} \boldsymbol{X}_{(k)}^T \boldsymbol{W} \boldsymbol{x}_k$. The design-based estimate of $R^2_{U(k)}$ can then be written as:

$$\hat{R}^2_{(k)} = \frac{\hat{\boldsymbol{\beta}}_{SW(k)}^T \boldsymbol{X}_{(k)}^T \boldsymbol{W} \boldsymbol{X}_{(k)} \hat{\boldsymbol{\beta}}_{SW(k)}}{\boldsymbol{x}_k^T \boldsymbol{W} \boldsymbol{x}_k}, \tag{3.29}$$

and

$$\hat{R}^2_{m(k)} = \frac{\hat{\boldsymbol{\beta}}^T_{SW(k)} \boldsymbol{X}^T_{(k)} \boldsymbol{W} \boldsymbol{X}_{(k)} \hat{\boldsymbol{\beta}}_{SW(k)}}{(\boldsymbol{x}_k - \mathbf{1}_n \bar{\bar{x}}_k)^T \boldsymbol{W} (\boldsymbol{x}_k - \mathbf{1}_n \bar{\bar{x}}_k)} = \frac{\hat{\boldsymbol{\beta}}^T_{SW(k)} \boldsymbol{X}^T_{(k)} \boldsymbol{W} \boldsymbol{X}_{(k)} \hat{\boldsymbol{\beta}}^T_{SW(k)}}{\boldsymbol{x}_k \boldsymbol{W} \boldsymbol{x}_k - \hat{N} \bar{\bar{x}}^2_k}. \tag{3.30}$$

Let the design-based estimate of VIF to be

$$\widehat{\mathrm{VIF}}_k = \frac{\hat{\zeta}_k \hat{\varrho}_k}{1 - \hat{R}^2_{(k)}}, \tag{3.31}$$

and its intercept-adjustment version to be

$$\widehat{\mathrm{VIF}}_{mk} = \frac{\hat{\zeta}_k \hat{\varrho}_{mk}}{1 - \hat{R}^2_{m(k)}}. \tag{3.32}$$

As long as $\hat{\zeta}_k$, $\hat{\varrho}_k$, $\hat{R}^2_{(k)}$, $\hat{\varrho}_{mk}$ and $\hat{R}^2_{m(k)}$ are consistent estimators of the universe quantities in the sense that

$$\begin{aligned} &\hat{\zeta}_k - \zeta_{Uk} \xrightarrow{p} 0, \quad \hat{\varrho}_k - \varrho_{Uk} \xrightarrow{p} 0, \quad \hat{R}^2_{(k)} - R^2_{U(k)} \xrightarrow{p} 0, \\ &\hat{\varrho}_{mk} - \varrho_{Umk} \xrightarrow{p} 0, \quad \hat{R}^2_{m(k)} - R^2_{Um(k)} \xrightarrow{p} 0, \end{aligned} \tag{3.33}$$

then $\widehat{\mathrm{VIF}}_k$ is a consistent estimator of VIF_{Uk} and $\widehat{\mathrm{VIF}}_{mk}$ is a consistent estimator of VIF_{Umk}:

$$\begin{aligned} &\widehat{\mathrm{VIF}}_k - \mathrm{VIF}_{Uk} \xrightarrow{p} 0, \\ &\widehat{\mathrm{VIF}}_{mk} - \mathrm{VIF}_{Umk} \xrightarrow{p} 0. \end{aligned} \tag{3.34}$$

Technical conditions are required to make this argument more formal.

Next, we compare $\widehat{\text{VIF}}_k$ with the model-based VIF_k in (3.14), which is $\text{VIF}_k = \frac{\zeta_k \varrho_k}{1-R^2_{SW(k)}}$. The model-based estimate, $\zeta_k = \frac{\boldsymbol{e}_{xk}^T \boldsymbol{W}\boldsymbol{V}\boldsymbol{W}\boldsymbol{e}_{xk}}{\boldsymbol{e}_{xk}^T \boldsymbol{W}\boldsymbol{e}_{xk}}$, differs from the design-based estimate $\hat{\zeta}_k$, which has $\boldsymbol{W}^* \cdot \boldsymbol{V}$ in the numerator. Similarly, the model-based estimate, $\varrho_k = \frac{\boldsymbol{x}_k^T \boldsymbol{W}\boldsymbol{x}_k}{\boldsymbol{x}_k^T \boldsymbol{W}\boldsymbol{V}\boldsymbol{W}\boldsymbol{x}_k}$, is different from the design-based estimate $\hat{\varrho}_k$ derived above and the model-based estimate of ϱ_{mk} is different from the design-based estimate $\hat{\varrho}_{mk}$.

On the other hand, we can demonstrate the fact that the model-based estimate $R^2_{SW(k)}$ is the same as the design-based estimate:

$$\begin{aligned} R^2_{SW(k)} &= \frac{\hat{\boldsymbol{\beta}}^T_{SW_{(k)}} \tilde{\boldsymbol{X}}^T_{(k)} \tilde{\boldsymbol{X}}_{(k)} \hat{\boldsymbol{\beta}}_{SW_{(k)}}}{\tilde{\boldsymbol{X}}^T_k \tilde{\boldsymbol{X}}_k} \\ &= \frac{\hat{\boldsymbol{\beta}}^T_{SW(k)} \boldsymbol{X}^T_{(k)} \boldsymbol{W} \boldsymbol{X}_{(k)} \hat{\boldsymbol{\beta}}_{SW(k)}}{\boldsymbol{x}_k^T \boldsymbol{W} \boldsymbol{x}_k} = \hat{R}^2_{(k)}. \end{aligned} \tag{3.35}$$

This indicates that the design-based of the conventional VIF approach can be expressed as $(1 - R^2_{SW(k)})^{-1}$ and its intercept-adjustment version can be expressed as $(1 - R^2_{SWm(k)})^{-1}$. However, the model-based VIF_k in (3.14) and VIF_{mk} in (3.19) may be numerically different from $\widehat{\text{VIF}}_k$ and $\widehat{\text{VIF}}_{mk}$ in (3.31), due to the differences between ζ_k, ϱ_k, ϱ_{mk} and $\hat{\zeta}_k$, $\hat{\varrho}_k$, $\hat{\varrho}_{mk}$. If $\boldsymbol{W}^* = \boldsymbol{W}$ and $\boldsymbol{W} = \boldsymbol{V}^{-1}$, then $\zeta_k = \varrho_k = \varrho_{mk} = 1$ in (3.14) and (3.19) so that the model-based VIF and VIF_{mk} in (3.14) and (3.19) reduce to $(1 - R^2_{SW(k)})^{-1}$ and $(1 - R^2_{SWm(k)})^{-1}$, respectively. In contrast, $\hat{\zeta}_k$ in (3.25), $\hat{\varrho}_k$ in (3.26) and $\hat{\zeta}_{mk}$ in (3.27) do not reduce to 1, i.e. the design-based $\widehat{\text{VIF}}_k$ and $\widehat{\text{VIF}}_{mk}$ in (3.31) and (3.32) still involve the ζ and ϱ (or ϱ_m) adjustments.

3.1.7 Estimating the VIF with Unknown $\boldsymbol{V}$ in a Sample Selected from the Finite Population

The discussion in section 3.1.6 assumed that the covariance matrix $\boldsymbol{V}$ was known. In practice, $\boldsymbol{V}$ must be estimated. In this section we present estimators that are appropriate for models with independent errors and for models that have a clustered covariance structure. In both instances, estimators can be constructed using squared residuals that are approximately model-unbiased under very general variance-covariance structures.

There is a natural correspondence between particular sampling plans and certain types of models. A model with independent but heteroscedastic errors is often appropriate for a design where single-stage sampling is used. A model where units within clusters are correlated, but units in different clusters are uncorrelated may naturally describe a population where two-stage cluster sampling is used. The model variance in (3.2) has the form $Var_M(\hat{\boldsymbol{\beta}}_{SW}) = \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}\sigma^2 = \boldsymbol{G}\sigma^2$ with $\boldsymbol{B}$ containing the $\boldsymbol{V}$ matrix. As shown in this section estimators of this variance, which contain estimators of $\boldsymbol{V}$, can be constructed that have both model-based and design-based justification.

3.1.7.1 VIF for A Model with Independent Errors

In unequal-weighted single-stage with-replacement sampling, consider a linear model in which the Y_i's are independent but whose variances differ among the units:

$$Y_i = \dot{\boldsymbol{x}}_i \boldsymbol{\beta} + \varepsilon_i, \quad \varepsilon_i \sim ind(0, \psi_i), \tag{3.36}$$

where in this case $\dot{\boldsymbol{x}}_i = (x_{1i}, ..., x_{pi})^T$, the vector of covariates for the i^{th} unit and ψ_i is an unknown variance parameter and here the variance matrix $\boldsymbol{V} = diag(\psi_i)$, is the diagonal matrix with ψ_i on the main diagonal. The model variance of survey weighted estimator, $\hat{\boldsymbol{\beta}}_{SW}$, is

$$Var_M(\hat{\boldsymbol{\beta}}_{SW}) = \boldsymbol{A}^{-1} \left(\sum_{i=1}^{n} \dot{\boldsymbol{x}}_i w_i \psi_i w_i \dot{\boldsymbol{x}}_i^T \right) \boldsymbol{A}^{-1}. \tag{3.37}$$

The associated residual for i^{th} unit is $e_i = Y_i - \dot{\boldsymbol{x}}_i \hat{\boldsymbol{\beta}}_{SW}$. Under certain regularity conditions, $E_M(e_i^2) \approx \psi_i$ and e_i^2 is an approximately model-unbiased estimator of ψ_i (Valliant et al., 2000). Therefore, we can use e_i^2 to estimate the unknown variance elements ψ_i in (3.37) and use $diag(e_i^2)$ to estimate the unknown $\boldsymbol{V}$. The estimated model variance of $\hat{\boldsymbol{\beta}}_{SW}$ is:

$$var_M(\hat{\boldsymbol{\beta}}_{SW}) = \boldsymbol{A}^{-1} \left(\sum_{i=1}^{n} \dot{\boldsymbol{x}}_i w_i e_i^2 w_i \dot{\boldsymbol{x}}_i^T \right) \boldsymbol{A}^{-1} = \boldsymbol{A}^{-1} \boldsymbol{X}^T \boldsymbol{W} diag(e_i^2) \boldsymbol{W} \boldsymbol{X} \boldsymbol{A}^{-1}. \tag{3.38}$$

Here, note that $\boldsymbol{W}^{1/2} diag(e_i^2) \boldsymbol{W}^{1/2}$ estimates the matrix $\tilde{\boldsymbol{V}}$ and $\boldsymbol{X}^T \boldsymbol{W} diag(e_i^2) \boldsymbol{W} \boldsymbol{X}$ estimates the matrix $\boldsymbol{B}$ in Section 3.1.4.

Hence, the corresponding VIF is

$$\frac{\hat{\zeta}_k \hat{\varrho}_k}{1 - R^2_{SW(k)}} \tag{3.39}$$

with

$$\begin{aligned} \hat{\zeta}_k &= \frac{\tilde{\boldsymbol{e}}^T_{xk} \boldsymbol{W}^{1/2} diag(e_i^2) \boldsymbol{W}^{1/2} \tilde{\boldsymbol{e}}_{xk}}{\tilde{\boldsymbol{e}}^T_{xk} \tilde{\boldsymbol{e}}_{xk}} \\ &= \frac{\boldsymbol{e}^T_{xk} \boldsymbol{W} diag(e_i^2) \boldsymbol{W} \boldsymbol{e}_{xk}}{\boldsymbol{e}^T_{xk} \boldsymbol{W} \boldsymbol{e}_{xk}}, \end{aligned} \tag{3.40}$$

where $\boldsymbol{e}_{xk} = \boldsymbol{x}_k - \boldsymbol{X}_{(k)} \hat{\boldsymbol{\beta}}_{SW(k)}$, and

$$\hat{\varrho}_k = \frac{\tilde{\boldsymbol{x}}^T_k \tilde{\boldsymbol{x}}_k}{\tilde{\boldsymbol{x}}^T_k \boldsymbol{W}^{1/2} diag(e_i^2) \boldsymbol{W}^{1/2} \tilde{\boldsymbol{x}}_k} = \frac{\boldsymbol{x}^T_k \boldsymbol{W} \boldsymbol{x}_k}{\boldsymbol{x}^T_k \boldsymbol{W} diag(e_i^2) \boldsymbol{W} \boldsymbol{x}_k}.$$

Based on the results in Section 3.1.5, the intercept-adjusted $\hat{\varrho}_k$ and $R^2_{SW(k)}$ are:

$$\hat{\varrho}_{mk} = \frac{(\tilde{\boldsymbol{x}}^T_k \tilde{\boldsymbol{x}}_k - \hat{N} \bar{\tilde{x}}^2_k)}{(\tilde{\boldsymbol{x}}_k - \tilde{\mathbf{1}} \bar{\tilde{x}}_k)^T \boldsymbol{W} diag(e_i^2) \boldsymbol{W} (\tilde{\boldsymbol{x}}_k - \tilde{\mathbf{1}} \bar{\tilde{x}}_k)},$$

and

$$R^2_{SWm(k)} = \frac{\hat{\boldsymbol{\beta}}^T_{SW(k)} \boldsymbol{X}^T_{(k)} \boldsymbol{X}_{(k)} \hat{\boldsymbol{\beta}}_{SW(k)}}{(\tilde{\boldsymbol{x}}^T_k \tilde{\boldsymbol{x}}_k - \hat{N} \bar{\tilde{x}}^2_k)}.$$

As shown below, (3.38) is approximately equal to the design-based linearization estimator computed by many survey software packages. Using the design-based linearization variance estimator in this single-stage design, the design consistent variance estimator is: (See *Appendix.* C)

$$var_L(\hat{\boldsymbol{\beta}}_{SW}) = \boldsymbol{A}^{-1}\left[\frac{n}{n-1}\sum_{i=1}^{n}(\boldsymbol{z}_i^* - \bar{\boldsymbol{z}}^*)(\boldsymbol{z}_i^* - \bar{\boldsymbol{z}}^*)^T\right]\boldsymbol{A}^{-1} \tag{3.41}$$

where $\boldsymbol{z}_i^* = w_i(y_i - \dot{\boldsymbol{x}}_i^T\hat{\boldsymbol{\beta}}_{SW})\dot{\boldsymbol{x}}_i = w_ie_i\dot{\boldsymbol{x}}_i$ which has a $p \times 1$ dimension and $e_i = y_i - \dot{\boldsymbol{x}}_i^T\hat{\boldsymbol{\beta}}_{SW}$ is the residual for the i^{th} unit.

Since the estimating equation in the finite population is,

$$\sum_{i=1}^{N}(y_i - \dot{\boldsymbol{x}}_i^T\boldsymbol{\beta}_{SW})\dot{\boldsymbol{x}}_i = 0,$$

the mean of $\boldsymbol{z}_i^*$ is:

$$\bar{\boldsymbol{z}}^* = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i^* = \frac{1}{n}\sum_{i=1}^{n}w_i(y_i - \dot{\boldsymbol{x}}_i^T\boldsymbol{\beta}_{SW})\dot{\boldsymbol{x}}_i = 0.$$

Applying this to (3.41), we have:

$$\begin{aligned} var_L(\hat{\boldsymbol{\beta}}_{SW}) &= \boldsymbol{A}^{-1}\left[\frac{n}{n-1}\sum_{i=1}^{n}\boldsymbol{z}_i^*\boldsymbol{z}_i^{*T}\right]\boldsymbol{A}^{-1} \\ &= \frac{n}{n-1}\sum_{i=1}^{n}\boldsymbol{A}^{-1}\dot{\boldsymbol{x}}_iw_ie_i^2w_i\dot{\boldsymbol{x}}_i^T\boldsymbol{A}^{-1} \\ &= \frac{n}{n-1}\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{W}diag(e_i^2)\boldsymbol{W}\boldsymbol{X}\boldsymbol{A}^{-1}. \end{aligned} \tag{3.42}$$

As the sample size increases, $\frac{n}{n-1} \to 1$, so that $var_L(\hat{\boldsymbol{\beta}}_{SW})$ and $var_M(\hat{\boldsymbol{\beta}}_{SW})$ are approximately the same. Thus, $var_L(\hat{\boldsymbol{\beta}}_{SW})$ and $var_M(\hat{\boldsymbol{\beta}}_{SW})$ are approximiately design-unbiased under with-replacement sampling and model-unbiased under model (3.36). Moreover, since the factor $(n/n-1)$ in var_L is only decided by the sample size, the VIF as the factor of variance inflation will then be the same whether we used var_L or var_M as the estimator of variance.

3.1.7.2 VIF for A Model with Clustering

For a multi-stage sampling design, suppose that there are $i = 1, ..., N$ clusters in the population and $t = 1, ..., M_i$ units in cluster i. Under the model-based inference, it is assumed that units in different clusters are independent in a model. Under the design-based inference, analogously, we assume that the first-stage sample units are selected with replacement, which means that the sample indicators for units in different clusters are independent. As in the case of estimation under the independence model, we can construct a simple sandwich estimator that is consistent under a reasonably general variance specification and find an estimator for the variance matrix $\boldsymbol{V}$. Consider the model:

$$\begin{aligned} Y_{it} &= \dot{\boldsymbol{x}}_{it}^T\boldsymbol{\beta} + \varepsilon_{it} \quad i = 1, ..., N \quad t = 1, ..., M_i, \\ Cov_M(\varepsilon_{it}, \varepsilon_{i't'}) &= \begin{cases} \sigma_{itt}^2 & i = i', t = t' \\ \sigma_{itt'}^2 & i = i', t \neq t' \\ 0 & i \neq i', t \neq t'. \end{cases} \end{aligned} \tag{3.43}$$

Within each cluster, Y_{it}'s are correlated within the clusters while independent across different clusters. In the clustered sample, suppose n clusters are selected out of N clusters and m_i units are selected out of M_i units in the selected cluster i. The total number of sample units is $m = \sum_{i \in s} m_i$. $\hat{\boldsymbol{\beta}}_{SW}$ can be written as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{SW} &= \sum_{i \in s} \sum_{t \in s_i} \boldsymbol{A}^{-1} \dot{\boldsymbol{x}}_{it} w_{it} Y_{it} \\ &= \sum_{i \in s} \boldsymbol{A}^{-1} \boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{Y}_i,\end{aligned} \tag{3.44}$$

where s is the set of sample clusters and s_i is the set of sample units within sample cluster i. In (3.44), $\boldsymbol{X}_i$ is the $m_i \times p$ matrix of covariates for sample units in cluster i, $\boldsymbol{W}_i = diag(w_{ij}),\ j \in s_i$ is the $m_i \times m_i$ matrix of weights for cluster i, and $\boldsymbol{Y}_i$ is the $m_i \times 1$ vector of response variables in cluster i.

The model variance of $\hat{\boldsymbol{\beta}}_{SW}$ is:

$$Var_M(\hat{\boldsymbol{\beta}}_{SW}) = \boldsymbol{A}^{-1} \left[\sum_{i \in s} \boldsymbol{X}_i^T \boldsymbol{W}_i \boldsymbol{V}_i \boldsymbol{W}_i \boldsymbol{X}_i \right] \boldsymbol{A}^{-1}, \tag{3.45}$$

where $\boldsymbol{V}_i = Var_M(\boldsymbol{Y}_i)$.

Denote the cluster-level residuals as a vector, $\boldsymbol{e}_i = \boldsymbol{Y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}_{SW}$. As the number of sampled clusters gets large, $n \to \infty$, $E_M(\boldsymbol{e}_i \boldsymbol{e}_i^T) \doteq Var_M(\boldsymbol{Y}_i)$ (Valliant et al. (2000), sec.9.5.1), and the model-based variance estimator for $\hat{\boldsymbol{\beta}}_{SW}$ has a similar sandwich form as the one in the single-stage model,

$$\begin{aligned}var_M(\hat{\boldsymbol{\beta}}_{SW}) &= \boldsymbol{A}^{-1} \left[\sum_{i \in s} \boldsymbol{X}_i^T \boldsymbol{W}_i (\boldsymbol{e}_i \boldsymbol{e}_i^T) \boldsymbol{W}_i \boldsymbol{X}_i \right] \boldsymbol{A}^{-1} \\ &= \boldsymbol{A}^{-1} \boldsymbol{X}^T \boldsymbol{W} Blkdiag(\boldsymbol{e}_i \boldsymbol{e}_i^T) \boldsymbol{W} \boldsymbol{X} \boldsymbol{A}^{-1},\end{aligned} \tag{3.46}$$

where $Blkdiag(\boldsymbol{e}_i \boldsymbol{e}_i^T)$ is an $m \times m$ block diagonal matrix with $\boldsymbol{e}_i \boldsymbol{e}_i^T$ on the main diagonal position and 0 elsewhere (see *Appendix.* D for more details).

Similar to the single-stage sampling model, $Blkdiag(\boldsymbol{e}_i\boldsymbol{e}_i^T)$ estimates the variance matrix $\boldsymbol{V}$, $\boldsymbol{W}^{1/2}Blkdiag(\boldsymbol{e}_i\boldsymbol{e}_i^T)\boldsymbol{W}^{1/2}$ estimates the matrix $\tilde{\boldsymbol{V}}$ and $\boldsymbol{X}^T\boldsymbol{W}Blkdiag(\boldsymbol{e}_i\boldsymbol{e}_i^T)\boldsymbol{W}\boldsymbol{X}$ estimates the matrix $\boldsymbol{B}$ in Section 3.1.4. The corresponding VIF is

$$\frac{\hat{\zeta}_k\hat{\varrho}_k}{1-R^2_{SW(k)}} \tag{3.47}$$

with

$$\begin{aligned}\hat{\zeta}_k &= \frac{\tilde{\boldsymbol{e}}_{xk}^T\boldsymbol{W}^{1/2}Blkdiag(\boldsymbol{e}_i\boldsymbol{e}_i^T)\boldsymbol{W}^{1/2}\tilde{\boldsymbol{e}}_{xk}}{\tilde{\boldsymbol{e}}_{xk}^T\tilde{\boldsymbol{e}}_{xk}} \\ &= \frac{\boldsymbol{e}_{xk}^T\boldsymbol{W}Blkdiag(\boldsymbol{e}_i\boldsymbol{e}_i^T)\boldsymbol{W}\boldsymbol{e}_{xk}}{\boldsymbol{e}_{xk}^T\boldsymbol{W}\boldsymbol{e}_{xk}}\end{aligned} \tag{3.48}$$

and

$$\hat{\varrho}_k = \frac{\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k}{\tilde{\boldsymbol{x}}_k^T\boldsymbol{W}^{1/2}Blkdiag(\boldsymbol{e}_i\boldsymbol{e}_i^T)\boldsymbol{W}^{1/2}\tilde{\boldsymbol{x}}_k} = \frac{\boldsymbol{x}_k^T\boldsymbol{W}\boldsymbol{x}_k}{\boldsymbol{x}_k^T\boldsymbol{W}Blkdiag(\boldsymbol{e}_i\boldsymbol{e}_i^T)\boldsymbol{W}\boldsymbol{x}_k}.$$

Based on the results in Section 3.1.5, the intercept-adjusted $\hat{\varrho}_k$ and $R^2_{SW(k)}$ are:

$$\hat{\varrho}_{mk} = \frac{(\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k - \hat{N}\bar{\tilde{x}}_k^2)}{(\tilde{\boldsymbol{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)^T\boldsymbol{W}Blkdiag(\boldsymbol{e}_i\boldsymbol{e}_i^T)\boldsymbol{W}(\tilde{\boldsymbol{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{x}}_k)},$$

and

$$R^2_{SWm(k)} = \frac{\hat{\boldsymbol{\beta}}_{SW(k)}^T\boldsymbol{X}_{(k)}^T\boldsymbol{X}_{(k)}\hat{\boldsymbol{\beta}}_{SW(k)}}{(\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k - \hat{N}\bar{\tilde{x}}_k^2)}.$$

When the number of sampled clusters n is large, the estimate in (3.46) is also approximately equal to the design-based linearization estimator. We show this below.

In the design-based linear regression, the estimating equations in the finite population are:

$$\begin{aligned}&\sum_{i=1}^{N}\sum_{t=1}^{M_i}(y_{it}-\dot{\boldsymbol{x}}_{it}^T\hat{\boldsymbol{\beta}}_{SW})x_{kit}=0, \quad k=1,\ldots,p;\\&\sum_{i=1}^{N}\boldsymbol{x}_{Uki}^T(\boldsymbol{Y}_{Ui}-\boldsymbol{X}_{Ui}^T\hat{\boldsymbol{\beta}}_{SW})=0,\end{aligned} \tag{3.49}$$

where $\boldsymbol{Y}_{Ui}=(Y_{i1},...,Y_{iM_i})^T$, $\boldsymbol{X}_{Ui}=(X_{i1},...,X_{iM_i})$ and $\boldsymbol{x}_{Uki}=(x_{ki1},...,x_{kiM_i})^T$.

Let $\boldsymbol{z}_i=\boldsymbol{X}_i^T(\boldsymbol{Y}_i-\boldsymbol{X}_i\hat{\boldsymbol{\beta}}_{SW})$. Define a weighted vector of residuals for cluster i as, $\boldsymbol{z}_i^*=\boldsymbol{X}_i^T\boldsymbol{W}_i(\boldsymbol{Y}_i-\boldsymbol{X}_i\hat{\boldsymbol{\beta}}_{SW})$. The design-based linearization variance estimator can be given as:

$$var_L(\hat{\boldsymbol{\beta}}_{SW})=\boldsymbol{A}^{-1}\left[\frac{n}{n-1}\sum_{i=1}^{n}(\boldsymbol{z}_i^*-\bar{\boldsymbol{z}}^*)(\boldsymbol{z}_i^*-\bar{\boldsymbol{z}}^*)^T\right]\boldsymbol{A}^{-1} \tag{3.50}$$

where

$$\bar{\boldsymbol{z}}^*=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{z}_i^*=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i^T\boldsymbol{W}_i(\boldsymbol{Y}_i-\boldsymbol{X}_i\hat{\boldsymbol{\beta}}_{SW})=0$$

by the property of estimating equations (3.49).

Thus, the linearization variance estimator can be written as,

$$\begin{aligned}var_L(\hat{\boldsymbol{\beta}}_{SW})&=\boldsymbol{A}^{-1}\left[\frac{n}{n-1}\sum_{i=1}^{n}\boldsymbol{z}_i^*\boldsymbol{z}_i^{*T}\right]\boldsymbol{A}^{-1}\\&=\frac{n}{n-1}\sum_{i=1}^{n}\boldsymbol{A}^{-1}\boldsymbol{X}_i\boldsymbol{W}_i(\boldsymbol{e}_i\boldsymbol{e}_i^T)\boldsymbol{W}_i\boldsymbol{X}_i^T\boldsymbol{A}^{-1}\\&=\frac{n}{n-1}\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{W}Blkdiag(\boldsymbol{e}_i\boldsymbol{e}_i^T)\boldsymbol{W}\boldsymbol{X}\boldsymbol{A}^{-1}.\end{aligned} \tag{3.51}$$

In large cluster samples, when $\frac{n}{n-1}\to 1$, $var_L(\hat{\boldsymbol{\beta}}_{SW})$ and $var_M(\hat{\boldsymbol{\beta}}_{SW})$ are then approximately the same. $Blkdiag(\boldsymbol{e}_i\boldsymbol{e}_i^T)$ can also approximately estimate the variance matrix $\boldsymbol{V}$ and $\boldsymbol{X}^T\boldsymbol{W}Blkdiag(\boldsymbol{e}_i\boldsymbol{e}_i^T)\boldsymbol{W}\boldsymbol{X}$ can approximately estimate the matrix $\boldsymbol{B}$. Similar to the single stage sampling, $var_M(\hat{\boldsymbol{\beta}}_{SW})$ and $var_L(\hat{\boldsymbol{\beta}}_{SW})$ are approxmiately model-unbiased under (3.43) and design-unbiased when clusters are sampled with replacement. Since the factor $(n/n-1)$ in var_L is only decided by the sample size, the VIF as the factor of variance inflation for var_L is the same as the VIF for var_M, which is listed in (3.47).

3.1.7.3 VIF for A Model with Stratified Clustering

Suppose that in a stratified multistage sampling design, there are $h = 1, ..., H$ strata in the population, $i = 1, ..., N_h$ clusters in the corresponding stratum h and $t = 1, ..., M_{hi}$ units in cluster hi. Denote the set of sample clusters in stratum h by s_h. The total number of sample units in stratum h is $m_h = \sum_{i \in s_h} m_{hi}$ and the total in the sample is $m = \sum_{h=1}^{H} m_h$. We select $i = 1, ..., n_h$ clusters in stratum h and $t = 1, ..., m_{hi}$ units in cluster hi. Clusters are assumed to be selected with replacement within strata and independently between strata. We will consider two linear models: one assumes that there are common intercept and slopes across strata; another assumes that there are different linear models, or different parameters in each stratum.

The first model can be expressed as:

$$\begin{aligned}
&E_M(Y_{hit}) = \boldsymbol{x}_{hit}^T\boldsymbol{\beta} \quad h = 1,\ldots,H, \quad i = 1,\ldots,N_h, \quad t = 1,\ldots,M_{hi} \\
&Cov_M(Y_{hit}, Y_{hi't'}) = 0 \quad i \neq i' \\
&Cov_M(Y_{hit}, Y_{h'i't'}) = 0 \quad h \neq h'.
\end{aligned} \tag{3.52}$$

The estimator of the regression parameter is a modification of (3.44):

$$\hat{\boldsymbol{\beta}}_{SW} = \sum_{h=1}^{H}\sum_{i\in s_h} \boldsymbol{A}^{-1}\boldsymbol{X}_{hi}^T\boldsymbol{W}_{hi}\boldsymbol{Y}_{hi} \tag{3.53}$$

where $\boldsymbol{X}_{hi}(m_{hi} \times p)$, $\boldsymbol{W}_{hi}(m_{hi} \times m_{hi})$ and $\boldsymbol{Y}_{hi}(m_{hi} \times 1)$ are defined by analogy to $\boldsymbol{X}_i$, $\boldsymbol{W}_i$ and $\boldsymbol{Y}_i$ in the previous section. The model variance of $\hat{\boldsymbol{\beta}}_{SW}$ is:

$$Var_M(\hat{\boldsymbol{\beta}}_{SW}) = \boldsymbol{A}^{-1}\left[\sum_{h=1}^{H}\sum_{i\in s_h} \boldsymbol{X}_{hi}^T\boldsymbol{W}_{hi}\boldsymbol{V}_{hi}\boldsymbol{W}_{hi}\boldsymbol{X}_{hi}\right]\boldsymbol{A}^{-1}, \tag{3.54}$$

where $\boldsymbol{V}_{hi} = Var_M(\boldsymbol{Y}_{hi})$.

The model-based sandwich variance estimator is:

$$\begin{aligned}
var_M(\hat{\boldsymbol{\beta}}_{SW}) &= \boldsymbol{A}^{-1}\left[\sum_{h=1}^{H}\sum_{i\in s_h} \boldsymbol{X}_{hi}^T\boldsymbol{W}_{hi}(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)\boldsymbol{W}_{hi}\boldsymbol{X}_{hi}\right]\boldsymbol{A}^{-1} \\
&= \boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{W}Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)\boldsymbol{W}\boldsymbol{X}\boldsymbol{A}^{-1}
\end{aligned} \tag{3.55}$$

where $\boldsymbol{e}_{hi} = \boldsymbol{Y}_{hi} - \boldsymbol{X}_{hi}\hat{\boldsymbol{\beta}}_{SW}$ and the matrix $m \times m$, $Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)$, has $\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T$ on the main diagonal position and 0 elsewhere (See *Appendix* D for more details).

Similar to the previous section, $Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)$ is an estimator for the variance matrix $\boldsymbol{V}$, $\boldsymbol{W}^{1/2}Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)\boldsymbol{W}^{1/2}$ estimates the matrix $\tilde{\boldsymbol{V}}$ and $\boldsymbol{X}^T\boldsymbol{W}Blkdiag$ $(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)\boldsymbol{W}\boldsymbol{X}$ estimates the matrix $\boldsymbol{B}$ in Section 3.1.4. Therefore, if we replace $Blkdiag(\boldsymbol{e}_i\boldsymbol{e}_i^T)$ in (3.47) with $Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)$, we can obtain the estimated VIF for model (3.52). Here, the VIF can be estimated by:

$$\frac{\hat{\zeta}_k\hat{\varrho}_k}{1-R^2_{SW(k)}} \tag{3.56}$$

with

$$\begin{aligned}\hat{\zeta}_k &= \frac{\tilde{\boldsymbol{e}}_{xk}^T\boldsymbol{W}^{1/2}Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)\boldsymbol{W}^{1/2}\tilde{\boldsymbol{e}}_{xk}}{\tilde{\boldsymbol{e}}_{xk}^T\tilde{\boldsymbol{e}}_{xk}} \\ &= \frac{\boldsymbol{e}_{xk}^T\boldsymbol{W}Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)\boldsymbol{W}\boldsymbol{e}_{xk}}{\boldsymbol{e}_{xk}^T\boldsymbol{W}\boldsymbol{e}_{xk}}\end{aligned} \tag{3.57}$$

and

$$\hat{\varrho}_k = \frac{\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k}{\tilde{\boldsymbol{x}}_k^T\boldsymbol{W}^{1/2}Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)\boldsymbol{W}^{1/2}\tilde{\boldsymbol{x}}_k} = \frac{\boldsymbol{x}_k^T\boldsymbol{W}\boldsymbol{x}_k}{\boldsymbol{x}_k^T\boldsymbol{W}Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)\boldsymbol{W}\boldsymbol{x}_k}.$$

The intercept-adjusted $\hat{\varrho}_k$ and $R^2_{SW(k)}$ are:

$$\hat{\varrho}_{mk} = \frac{(\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k - \hat{N}\bar{\tilde{x}}_k^2)}{(\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{1}}\bar{\tilde{x}}_k)^T\boldsymbol{W}Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)\boldsymbol{W}(\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{1}}\bar{\tilde{x}}_k)},$$

and

$$R^2_{SWm(k)} = \frac{\hat{\boldsymbol{\beta}}_{SW(k)}^T\boldsymbol{X}_{(k)}^T\boldsymbol{X}_{(k)}\hat{\boldsymbol{\beta}}_{SW(k)}}{(\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k - \hat{N}\bar{\tilde{x}}_k^2)}.$$

Note that $\hat{\zeta}_k$ incorporates both the effect of clustering on the model errors (through $Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)$) and how closely related $\boldsymbol{x}_k$ is to the other $\boldsymbol{x}$'s (through

$\boldsymbol{e}_{xk}$). The term $\hat{\varrho}_k$ reflects clustering (through $Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)$ in its denominator) but tends to move in the opposite direction from $\hat{\zeta}_k$.

Next, as proven in the previous section, we can show that the estimate in (3.55) is also approximately equal to the design-based linearization estimator for the corresponding stratified multistage design. The components in it can correspondingly estimate the $\tilde{\boldsymbol{V}}$ in $var_L(\hat{\boldsymbol{\beta}}_{SW})$ for this sampling design as they did in $var_M(\hat{\boldsymbol{\beta}}_{SW})$.

After accounting for stratification, the linearization variance estimator in this case becomes:

$$\begin{aligned} var_L(\hat{\boldsymbol{\beta}}_{SW}) &= \boldsymbol{A}^{-1}\left[\sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{i\in s_h}(\boldsymbol{z}_{hi}^* - \bar{\boldsymbol{z}}_h^*)(\boldsymbol{z}_{hi}^* - \bar{\boldsymbol{z}}_h^*)^T\right]\boldsymbol{A}^{-1} \\ &= \boldsymbol{A}^{-1}\left[\sum_{h=1}^{H}\frac{n_h}{n_h-1}(\sum_{i\in s_h}\boldsymbol{z}_{hi}^*\boldsymbol{z}_{hi}^{*T} - n_h\bar{\boldsymbol{z}}_h^*\bar{\boldsymbol{z}}_h^{*T})\right]\boldsymbol{A}^{-1} \\ &= \sum_{h=1}^{H}\boldsymbol{A}^{-1}\left(\frac{n_h}{n_h-1}\sum_{i\in s_h}\boldsymbol{z}_{hi}^*\boldsymbol{z}_{hi}^{*T}\right)\boldsymbol{A}^{-1} - \sum_{h=1}^{H}\boldsymbol{A}^{-1}\left(\frac{n_h^2}{n_h-1}\bar{\boldsymbol{z}}_h^*\bar{\boldsymbol{z}}_h^{*T}\right)\boldsymbol{A}^{-1} \end{aligned} \tag{3.58}$$

where $\boldsymbol{z}_{hi}^* = \boldsymbol{X}_{hi}^T\boldsymbol{W}_{hi}\boldsymbol{e}_{hi}$, and $\bar{\boldsymbol{z}}_h^* = \frac{1}{n_h}\sum_{i\in s}\boldsymbol{z}_{hi}^*$. This expression can be reduced to the formula for a single-stage stratified design when the primary sampling unit (PSU) sizes are all equal to 1, $m_{hi} = 1$. The estimator $var_L(\hat{\boldsymbol{\beta}}_{SW})$ is consistent and approximately design-unbiased under a design where PSUs are selected with replacement (Fuller, 2002).

As shown before, we have

$$\boldsymbol{A}^{-1}\left(\frac{n_h}{n_h-1}\sum_{i\in s}\boldsymbol{z}^*_{hi}\boldsymbol{z}^{*T}_{hi}\right)\boldsymbol{A}^{-1}=\boldsymbol{A}^{-1}\left[\frac{n_h}{n_h-1}\boldsymbol{X}_h^T\boldsymbol{W}_h Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)\boldsymbol{W}_h\boldsymbol{X}_h\right]\boldsymbol{A}^{-1} \tag{3.59}$$

where $\boldsymbol{X}_h$ is the $\sum_{i\in s_h} m_{hi}\times p$ matrix of covariates for sample units in stratum h and $\boldsymbol{W}_h$ is the $\sum_{i\in s_h} m_{hi}\times\sum_{i\in s_h} m_{hi}$ matrix of weights for sample units in stratum h.

Let $\boldsymbol{z}^*_h=(\boldsymbol{z}^*_{h1},...,\boldsymbol{z}^*_{hn_h})$ and $\boldsymbol{E}_h=diag(\boldsymbol{e}_{h1},...,\boldsymbol{e}_{hn_h})$. Then $\boldsymbol{z}^*_h=\boldsymbol{X}_h^T\boldsymbol{W}_h\boldsymbol{E}_h$. Let $\boldsymbol{l}$ be a column vector consisting of m_h unit elements. The stratum mean of $\boldsymbol{z}^*_{hi}$ is:

$$\bar{\boldsymbol{z}}^*_h=\frac{1}{n_h}\boldsymbol{z}^*_h\boldsymbol{l},$$

with this notation we have:

$$\begin{aligned}\bar{\boldsymbol{z}}^*_h\bar{\boldsymbol{z}}^{*T}_h&=\frac{1}{n_h^2}\boldsymbol{z}^*_h\boldsymbol{l}\boldsymbol{l}^T\boldsymbol{z}^{*T}_h\\&=\frac{1}{n_h^2}\boldsymbol{X}_h^T\boldsymbol{W}_h\boldsymbol{E}_h\boldsymbol{l}\boldsymbol{l}^T\boldsymbol{E}_h^T\boldsymbol{W}_h\boldsymbol{X}_h\\&=\frac{1}{n_h^2}\boldsymbol{X}_h^T\boldsymbol{W}_h\boldsymbol{e}_h\boldsymbol{e}_h^T\boldsymbol{W}_h\boldsymbol{X}_h\end{aligned} \tag{3.60}$$

where $\boldsymbol{e}_h=(\boldsymbol{e}_{h1},\boldsymbol{e}_{h2},...,\boldsymbol{e}_{hn_h})^T$ is a vector of unit residuals in stratum h and $\boldsymbol{e}_h\boldsymbol{e}_h^T$

is a symmetric $m_h \times m_h$ matrix:

$$\begin{pmatrix} \boldsymbol{e}_{h1}\boldsymbol{e}_{h1}^T & \boldsymbol{e}_{h1}\boldsymbol{e}_{h2}^T & \dots & \boldsymbol{e}_{h1}\boldsymbol{e}_{hn_h}^T \\ \boldsymbol{e}_{h2}\boldsymbol{e}_{h1}^T & \boldsymbol{e}_{h2}\boldsymbol{e}_{h2}^T & \dots & \boldsymbol{e}_{h2}\boldsymbol{e}_{hn_h}^T \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{e}_{hn_h}\boldsymbol{e}_{h1}^T & \boldsymbol{e}_{hn_h}\boldsymbol{e}_{h2}^T & \dots & \boldsymbol{e}_{hn_h}\boldsymbol{e}_{hn_h}^T \end{pmatrix}.$$

Substituting (3.59) and (3.60) in (3.58) gives:

$$\begin{aligned} var_L(\hat{\boldsymbol{\beta}}) &= \sum_{h=1}^{H} \boldsymbol{A}^{-1} \quad \left[\frac{n_h}{n_h-1}\boldsymbol{X}_h^T\boldsymbol{W}_h Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)\boldsymbol{W}_h\boldsymbol{X}_h\right]\boldsymbol{A}^{-1} \\ &- \sum_{h=1}^{H} \boldsymbol{A}^{-1}\left[\frac{1}{n_h-1}\boldsymbol{X}_h^T\boldsymbol{W}_h\boldsymbol{e}_h\boldsymbol{e}_h^T\boldsymbol{W}_h\boldsymbol{X}_h\right]\boldsymbol{A}^{-1} \\ &= \sum_{h=1}^{H} \boldsymbol{A}^{-1}\frac{n_h}{n_h-1}\boldsymbol{X}_h^T\boldsymbol{W}_h\left[Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T) - \frac{1}{n_h}\boldsymbol{e}_h\boldsymbol{e}_h^T\right]\boldsymbol{W}_h\boldsymbol{X}_h\boldsymbol{A}^{-1}. \end{aligned} \tag{3.61}$$

$\sum_{h=1}^{H} \frac{n_h}{n_h-1}\boldsymbol{X}_h^T\boldsymbol{W}_h\left[Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T) - \frac{1}{n_h}\boldsymbol{e}_h\boldsymbol{e}_h^T\right]\boldsymbol{W}_h\boldsymbol{X}_h$ estimates matrix $\boldsymbol{B}$ and $Blkdiag\left\{\frac{n_h}{n_h-1}\left[Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T) - \frac{1}{n_h}\boldsymbol{e}_h\boldsymbol{e}_h^T\right]\right\}$ estimates matrix $\boldsymbol{V}$ in $var_L(\hat{\boldsymbol{\beta}}_{SW})$. The VIF for $var_L(\hat{\boldsymbol{\beta}}_{SW})$ in this stratified multistage sampling design, therefore, VIF can be estimated by:

$$\frac{\hat{\zeta}_k\hat{\varrho}_k}{1-R^2_{SW(k)}} \tag{3.62}$$

where

$$\hat{\zeta}_k = \frac{\boldsymbol{e}_{xk}^T\boldsymbol{W}Blkdiag\left\{\frac{n_h}{n_h-1}\left[Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T) - \frac{1}{n_h}\boldsymbol{e}_h\boldsymbol{e}_h^T\right]\right\}\boldsymbol{W}\boldsymbol{e}_{xk}}{\boldsymbol{e}_{xk}^T\boldsymbol{W}\boldsymbol{e}_{xk}}, \tag{3.63}$$

and

$$\hat{\varrho}_k = \frac{\boldsymbol{x}_k^T \boldsymbol{W} \boldsymbol{x}_k}{\boldsymbol{x}_k^T \boldsymbol{W} Blkdiag\left\{\frac{n_h}{n_h-1}\left[Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T) - \frac{1}{n_h}\boldsymbol{e}_h\boldsymbol{e}_h^T\right]\right\}\boldsymbol{W}\boldsymbol{x}_k},$$

with $\boldsymbol{e}_{xk} = \boldsymbol{x}_k - \boldsymbol{X}_{(k)}\hat{\boldsymbol{\beta}}_{SW(k)}$.

The intercept-adjusted $\hat{\varrho}_k$ and $R^2_{SW(k)}$ are:

$$\hat{\varrho}_{mk} = \frac{(\tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{x}}_k - \hat{N}\bar{\bar{x}}_k^2)}{(\tilde{\boldsymbol{x}}_k - \tilde{\mathbf{1}}\bar{\bar{x}}_k)^T \boldsymbol{W} Blkdiag\left\{\frac{n_h}{n_h-1}\left[Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T) - \frac{1}{n_h}\boldsymbol{e}_h\boldsymbol{e}_h^T\right]\right\}\boldsymbol{W}(\tilde{\boldsymbol{x}}_k - \tilde{\mathbf{1}}\bar{\bar{x}}_k)},$$

and

$$R^2_{SWm(k)} = \frac{\hat{\boldsymbol{\beta}}_{SW(k)}^T \boldsymbol{X}_{(k)}^T \boldsymbol{X}_{(k)} \hat{\boldsymbol{\beta}}_{SW(k)}}{(\tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{x}}_k - \hat{N}\bar{\bar{x}}_k^2)}.$$

According to (3.61), the model-based expectation of $var_L(\hat{\boldsymbol{\beta}})$ is:

$$\begin{aligned} E_M\left[var_L(\hat{\boldsymbol{\beta}}_{SW})\right] = \sum_{h=1}^{H} \frac{n_h}{n_h-1} & \left\{E_M\left[\boldsymbol{A}^{-1}\boldsymbol{X}_h^T\boldsymbol{W}_h Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)\boldsymbol{W}_h\boldsymbol{X}_h\boldsymbol{A}^{-1}\right]\right. \\ & \left. -E_M\left[\frac{1}{n_h}\boldsymbol{A}^{-1}\boldsymbol{X}_h^T\boldsymbol{W}_h\boldsymbol{e}_h\boldsymbol{e}_h^T\boldsymbol{W}_h\boldsymbol{X}_h\boldsymbol{A}^{-1}\right]\right\}. \end{aligned} \tag{3.64}$$

Note that $E_M(\boldsymbol{e}_h\boldsymbol{e}_h^T) = E_M\left[Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)\right]$ because $E_M(\boldsymbol{e}_{hi}\boldsymbol{e}_{h'i'}^T) = 0$ when $(hi) \neq (h'i')$. As in Valliant et al. (2000), $E_M(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T) \doteq \boldsymbol{V}_{hi}$. Next, define $\boldsymbol{V}_h = Blkdiag(\boldsymbol{V}_{hi})$, $i \in s_h$. Thus, $E_M(\boldsymbol{e}_h\boldsymbol{e}_h^T) = E_M\left[Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)\right] \doteq \boldsymbol{V}_h$. Substituting in (3.64), we have

$$\begin{aligned} E_M\left[var_L(\hat{\boldsymbol{\beta}}_{SW})\right] &\doteq \sum_{h=1}^{H} \boldsymbol{A}^{-1}\boldsymbol{X}_h^T\boldsymbol{W}_h\boldsymbol{V}_h\boldsymbol{W}_h\boldsymbol{X}_h\boldsymbol{A}^{-1} \\ &= Var_M(\hat{\boldsymbol{\beta}}_{SW}). \end{aligned} \tag{3.65}$$

Consequently, $var_L(\hat{\boldsymbol{\beta}}_{SW})$ and $var_M(\hat{\boldsymbol{\beta}}_{SW})$ are approximately model and design unbiased.

Under both of the situations, the model-based expectation of $var_L(\hat{\boldsymbol{\beta}}_{SW})$ is equal to $var_M(\hat{\boldsymbol{\beta}}_{SW})$ and so the model-based expectation of the design-based VIF in (3.56) is the same as the model-based VIF. When n_h is large, $var_L(\hat{\boldsymbol{\beta}}_{SW})$ and $var_M(\hat{\boldsymbol{\beta}}_{SW})$ are approximately the same and so are their VIFs.

As suggested in Little (2004), if we incorporate the stratification in the model and assume different linear models, or different slope parameters, $\boldsymbol{\beta}_{SWh}$, in each stratum, the model in each stratum is:

$$\begin{aligned} &E_M(Y_{hit}) = \boldsymbol{x}_{hit}^T \boldsymbol{\beta}_{SWh} \quad h = 1, \ldots, H, \quad i = 1, \ldots, N_h, \quad t = 1, \ldots, M_{hi} \\ &Cov_M(Y_{hit}, Y_{hi't'}) = 0 \quad i \neq i'. \end{aligned} \tag{3.66}$$

Within each stratum, the estimation of regression parameters and their variances is the same as that for model (3.43) in the previous section. In each stratum, the model is:

$$\hat{\boldsymbol{\beta}}_{SWh} = \boldsymbol{A}_h^{-1} \sum_{i \in s} \boldsymbol{X}_{hi}^T \boldsymbol{W}_{hi} \boldsymbol{Y}_{hi},$$

and

$$var_M(\hat{\boldsymbol{\beta}}_{SWh}) = \boldsymbol{A}_h^{-1} \boldsymbol{X}_h^T \boldsymbol{W}_h Blkdiag(\boldsymbol{e}_{hi} \boldsymbol{e}_{hi}^T) \boldsymbol{W}_h \boldsymbol{X}_h \boldsymbol{A}_h^{-1}$$

where $\boldsymbol{A}_h = \boldsymbol{X}_h^T \boldsymbol{W}_h \boldsymbol{X}_h$ and $\boldsymbol{e}_{hi} = \boldsymbol{Y}_{hi} - \boldsymbol{X}_{hi} \hat{\boldsymbol{\beta}}_{SWh}$. The VIF for $\hat{\boldsymbol{\beta}}_{SWh}$ is similar to (3.47) and the design-based linearization variance estimator for $\hat{\boldsymbol{\beta}}_{SWh}$ is similar to (3.50), but with a stratum subscript h. When n_h is large, $var_M(\hat{\boldsymbol{\beta}}_{SWh})$ and

$var_L(\hat{\boldsymbol{\beta}}_{SWh})$ are approximately the same. Collinearity diagnostics will be conducted independently within each stratum for this setting.

3.1.7.4 Specialization for Stratified Models with No Clustering

When the sampling design is a stratified single stage sampling design, suppose that there are $h = 1, ..., H$ strata in the population and $i = 1, ..., N_h$ units in the corresponding stratum h. We select $i = 1, ..., n_h$ units with replacement in stratum h and independently between strata. Denote the set of sample units in stratum h by s_h. This is a special case of the sampling design discussed in Section 3.1.7.3, in which each cluster just has one single unit in it. Here, we can consider a stratified model with no clustering for this design, given as:

$$\begin{aligned} &E_M(Y_{hi}) = \boldsymbol{x}_{hi}^T\boldsymbol{\beta} \quad h = 1, \ldots, H, \quad i = 1, \ldots, N_h \\ &Cov_M(Y_{hi}, Y_{h'i'}) = 0 \quad i \neq i'. \end{aligned} \tag{3.67}$$

Correspondingly, the estimator of the regression parameter is:

$$\hat{\boldsymbol{\beta}}_{SW} = \sum_{h=1}^{H} \sum_{i \in s_h} \boldsymbol{A}^{-1} \dot{\boldsymbol{x}}_{hi}^T w_{hi} Y_{hi}. \tag{3.68}$$

Under the model-based inference, for this special case, the block diagonal matrix $Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)$ in the model-based VIF estimator in (3.56) can then be simplified as a diagonal matrix $diag(e_{hi}^2)$, with $e_{hi} = Y_{hi} - \boldsymbol{X}_{hi}\hat{\boldsymbol{\beta}}_{SW}$, and the model-based VIF can then be estimated by:

$$\frac{\hat{\zeta}_k \hat{\varrho}_k}{1 - R^2_{SW(k)}} \tag{3.69}$$

with

$$\begin{aligned}\hat{\zeta}_k &= \frac{\tilde{\boldsymbol{e}}_{xk}^T \boldsymbol{W}^{1/2} diag(e_{hi}^2) \boldsymbol{W}^{1/2} \tilde{\boldsymbol{e}}_{xk}}{\tilde{\boldsymbol{e}}_{xk}^T \tilde{\boldsymbol{e}}_{xk}} \\ &= \frac{\boldsymbol{e}_{xk}^T \boldsymbol{W} diag(e_{hi}^2) \boldsymbol{W} \boldsymbol{e}_{xk}}{\boldsymbol{e}_{xk}^T \boldsymbol{W} \boldsymbol{e}_{xk}}\end{aligned}$$

and

$$\hat{\varrho}_k = \frac{\tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{x}}_k}{\tilde{\boldsymbol{x}}_k^T \boldsymbol{W}^{1/2} diag(e_{hi}^2) \boldsymbol{W}^{1/2} \tilde{\boldsymbol{x}}_k} = \frac{\boldsymbol{x}_k^T \boldsymbol{W} \boldsymbol{x}_k}{\boldsymbol{x}_k^T \boldsymbol{W} diag(e_{hi}^2) \boldsymbol{W} \boldsymbol{x}_k}.$$

The intercept-adjusted $\hat{\varrho}_k$ and $R^2_{SW(k)}$ are:

$$\hat{\varrho}_{mk} = \frac{(\tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{x}}_k - \hat{N} \bar{\tilde{x}}_k^2)}{(\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{1}} \bar{\tilde{x}}_k)^T \boldsymbol{W} diag(e_{hi}^2) \boldsymbol{W} (\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{1}} \bar{\tilde{x}}_k)},$$

and

$$R^2_{SWm(k)} = \frac{\hat{\boldsymbol{\beta}}_{SW(k)}^T \boldsymbol{X}_{(k)}^T \boldsymbol{X}_{(k)} \hat{\boldsymbol{\beta}}_{SW(k)}}{(\tilde{\boldsymbol{x}}_k^T \tilde{\boldsymbol{x}}_k - \hat{N} \bar{\tilde{x}}_k^2)}.$$

Under the design-based inference,the estimator for $\boldsymbol{V}$ in the linearization variance estimator,

$$Blkdiag \left\{ \frac{n_h}{n_h - 1} \left[Blkdiag(\boldsymbol{e}_{hi} \boldsymbol{e}_{hi}^T) - \frac{1}{n_h} \boldsymbol{e}_h \boldsymbol{e}_h^T \right] \right\},$$

can be simplified as

$$Blkdiag\left\{\frac{n_h}{n_h-1}\left[diag(e_{hi}^2)-\frac{1}{n_h}\boldsymbol{e}_h\boldsymbol{e}_h^T\right]\right\},$$

where $\boldsymbol{e}_h = (e_{h1}, e_{h2}, ..., e_{hn_h})^T$ is a vector of n_h unit residuals in stratum h. Substituting it into the design-based VIF estimator for stratified with clustering design, expressed in (3.62), we can estimate the design-based VIF for this special case by:

$$\frac{\hat{\zeta}_k\hat{\varrho}_k}{1-R^2_{SW(k)}} \tag{3.70}$$

where

$$\hat{\zeta}_k = \frac{\boldsymbol{e}_{xk}^T\boldsymbol{W}Blkdiag\left\{\frac{n_h}{n_h-1}\left[diag(e_{hi}^2)-\frac{1}{n_h}\boldsymbol{e}_h\boldsymbol{e}_h^T\right]\right\}\boldsymbol{W}\boldsymbol{e}_{xk}}{\boldsymbol{e}_{xk}^T\boldsymbol{W}\boldsymbol{e}_{xk}},$$

and

$$\hat{\varrho}_k = \frac{\boldsymbol{x}_k^T\boldsymbol{W}\boldsymbol{x}_k}{\boldsymbol{x}_k^T\boldsymbol{W}Blkdiag\left\{\frac{n_h}{n_h-1}\left[diag(e_{hi}^2)-\frac{1}{n_h}\boldsymbol{e}_h\boldsymbol{e}_h^T\right]\right\}\boldsymbol{W}\boldsymbol{x}_k}.$$

The intercept-adjusted $\hat{\varrho}_k$ and $R^2_{SW(k)}$ are:

$$\hat{\varrho}_{mk} = \frac{(\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k - \hat{N}\bar{\tilde{x}}_k^2)}{(\tilde{\boldsymbol{x}}_k-\tilde{\mathbf{1}}\bar{\tilde{x}}_k)^T\boldsymbol{W}Blkdiag\left\{\frac{n_h}{n_h-1}\left[diag(e_{hi}^2)-\frac{1}{n_h}\boldsymbol{e}_h\boldsymbol{e}_h^T\right]\right\}\boldsymbol{W}(\tilde{\boldsymbol{x}}_k-\tilde{\mathbf{1}}\bar{\tilde{x}}_k)},$$

and

$$R^2_{SWm(k)} = \frac{\hat{\boldsymbol{\beta}}_{SW(k)}^T\boldsymbol{X}_{(k)}^T\boldsymbol{X}_{(k)}\hat{\boldsymbol{\beta}}_{SW(k)}}{(\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k - \hat{N}\bar{\tilde{x}}_k^2)}.$$

Note that the model-based VIF estimator here in (3.69) is the same as the model-based VIF estimator in Section 4.1.2.1, (3.39), because the model in (3.67) is a special case of model in (3.36) with independent errors since (3.67) assumes the units in each strata are independent with each other. However, under the design-based inference, the VIF estimators in (3.39) and (3.70) are different due to their different sampling designs.

3.1.8 VIF in Survey-Weighted Generalized Least Squares Regression

In the survey-weighted generalized least squares (SWGLS) regression, we adjust the weight matrix $\boldsymbol{W}$ in the SWVLS regression by defining $\ddot{\boldsymbol{W}} = \boldsymbol{W}\boldsymbol{V}^{\mathbf{-1}}$ as a new weight matrix. Similar to Section 3.1.2, we transform the design matrix as $\tilde{\tilde{\boldsymbol{X}}} = \ddot{\boldsymbol{W}}^{1/2}\boldsymbol{X}$ and the response vector as $\tilde{\tilde{\boldsymbol{Y}}} = \ddot{\boldsymbol{W}}^{1/2}\boldsymbol{Y}$. In accordance, the parameter estimator for SWVLS in (3.1) can then be written as

$$\hat{\boldsymbol{\beta}}_{SWV} = (\tilde{\tilde{\boldsymbol{X}}}^T\tilde{\tilde{\boldsymbol{X}}})^{-1}\tilde{\tilde{\boldsymbol{X}}}^T\tilde{\tilde{\boldsymbol{Y}}}.$$

We will also use $\tilde{\tilde{\boldsymbol{e}}} = \ddot{\boldsymbol{W}}^{1/2}\boldsymbol{e}$ with $\hat{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{SWV}$ and $\hat{\tilde{\tilde{\boldsymbol{e}}}} = \ddot{\boldsymbol{W}}^{1/2}\hat{\boldsymbol{e}}$ with $\hat{\boldsymbol{e}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{SWV}$.

The model variance of the parameter estimator $\hat{\boldsymbol{\beta}}_{SWV}$ can be expressed as

$$\begin{aligned} Var_M(\hat{\boldsymbol{\beta}}_{SWV}) &= (\tilde{\tilde{\boldsymbol{X}}}^T\tilde{\tilde{\boldsymbol{X}}})^{-1}\tilde{\tilde{\boldsymbol{X}}}^T E(\tilde{\tilde{\boldsymbol{e}}}^T\tilde{\tilde{\boldsymbol{e}}})\tilde{\tilde{\boldsymbol{X}}}(\tilde{\tilde{\boldsymbol{X}}}^T\tilde{\tilde{\boldsymbol{X}}})^{-1} \\ &= \sigma^2(\tilde{\tilde{\boldsymbol{X}}}^T\tilde{\tilde{\boldsymbol{X}}})^{-1}\tilde{\tilde{\boldsymbol{X}}}^T\boldsymbol{W}\tilde{\tilde{\boldsymbol{X}}}(\tilde{\tilde{\boldsymbol{X}}}^T\tilde{\tilde{\boldsymbol{X}}})^{-1} \\ &= \boldsymbol{A}^{\mathbf{-1}}\boldsymbol{B}\boldsymbol{A}^{\mathbf{-1}}\sigma^2 = \boldsymbol{G}\sigma^2 \end{aligned} \tag{3.71}$$

where

$$\begin{aligned}
E(\tilde{\tilde{\boldsymbol{e}}}^T\tilde{\tilde{\boldsymbol{e}}}) &= Var_M(\tilde{\tilde{\boldsymbol{e}}}) \\
&= \sigma^2\ddot{\boldsymbol{W}}^{1/2}\boldsymbol{V}\ddot{\boldsymbol{W}}^{1/2} \\
&= \sigma^2\boldsymbol{W}^{1/2}\boldsymbol{V}^{-1/2}\boldsymbol{V}\boldsymbol{V}^{-1/2}\boldsymbol{W}^{1/2} \\
&= \sigma^2\boldsymbol{W}^{1/2}\boldsymbol{V}^{-1/2}\boldsymbol{V}^{1/2}\boldsymbol{V}^{1/2}\boldsymbol{V}^{-1/2}\boldsymbol{W}^{1/2} \\
&= \sigma^2\boldsymbol{W},
\end{aligned} \tag{3.72}$$

since $Var_M(\boldsymbol{e}) = \sigma^2\boldsymbol{V}$. We also define $\boldsymbol{A} = \tilde{\tilde{\boldsymbol{X}}}^T\tilde{\tilde{\boldsymbol{X}}}$, $\boldsymbol{B} = \tilde{\tilde{\boldsymbol{X}}}^T\boldsymbol{W}\tilde{\tilde{\boldsymbol{X}}}$, and $\boldsymbol{G} = \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}$.

Similar to the SWLS case, the model variance of $\hat{\boldsymbol{\beta}}_{SWV_k}$, the coefficient of the k^{th} explanatory variable, is

$$Var_M(\hat{\beta}_{SWV_k}) = \boldsymbol{i}_k' Var_M(\hat{\boldsymbol{\beta}}_{SWV})\boldsymbol{i}_k = \sigma^2\boldsymbol{i}_k'\boldsymbol{G}\boldsymbol{i}_k = \sigma^2 g^{kk}. \tag{3.73}$$

Analogous to (3.7), the k^{th} diagonal element of $\boldsymbol{A}^{-1}$ is:

$$a^{kk} = \boldsymbol{i}_k^T\boldsymbol{A}^{-1}\boldsymbol{i}_k = \boldsymbol{i}_k^T(\tilde{\tilde{\boldsymbol{X}}}^T\tilde{\tilde{\boldsymbol{X}}})^{-1}\boldsymbol{i}_k = \frac{1}{(1-R^2_{SWV_{(k)}})SST_{SWV_{(k)}}} = \frac{1}{(1-R^2_{SWV_{(k)}})\tilde{\tilde{\boldsymbol{x}}}_k^T\tilde{\tilde{\boldsymbol{x}}}_k} \tag{3.74}$$

where $R^2_{SWV_{(k)}} = \hat{\beta}^T_{SWV_{(k)}}\tilde{\tilde{\boldsymbol{X}}}^T_{(k)}\tilde{\tilde{\boldsymbol{X}}}_{(k)}\hat{\beta}_{SWV_{(k)}}/\tilde{\tilde{\boldsymbol{x}}}_k^T\tilde{\tilde{\boldsymbol{x}}}_k$ with $\hat{\beta}_{SWV_{(k)}} = (\tilde{\tilde{\boldsymbol{X}}}^T_{(k)}\tilde{\tilde{\boldsymbol{X}}}_{(k)})^{-1}\tilde{\tilde{\boldsymbol{X}}}^T_{(k)}\tilde{\tilde{\boldsymbol{x}}}_k$ is the coefficient of determination corresponding to the SWVLS regression of $\tilde{\tilde{\boldsymbol{x}}}_k$ on the $p-1$ other explanatory variables. The term $SST_{SWV_{(k)}} = \tilde{\tilde{\boldsymbol{x}}}_k^T\tilde{\tilde{\boldsymbol{x}}}_k$, is the total sum of squares in this regression.

The partitioned version of $\boldsymbol{B}$ in SWVLS is:

$$\boldsymbol{B} = \begin{pmatrix} b_{kk} & \boldsymbol{b}_{k(k)} \\ \boldsymbol{b}_{(k)k} & \boldsymbol{B}_{(k)(k)} \end{pmatrix} = \begin{pmatrix} \tilde{\tilde{\boldsymbol{x}}}_k^T \boldsymbol{W} \tilde{\tilde{\boldsymbol{x}}}_k & \tilde{\tilde{\boldsymbol{x}}}_k^T \boldsymbol{W} \tilde{\tilde{\boldsymbol{X}}}_{(k)} \\ \tilde{\tilde{\boldsymbol{X}}}_{(k)}^T \boldsymbol{W} \tilde{\tilde{\boldsymbol{x}}}_k & \tilde{\tilde{\boldsymbol{X}}}_(k)^T \boldsymbol{W} \tilde{\tilde{\boldsymbol{X}}}_{(k)}. \end{pmatrix} \tag{3.75}$$

Analogous to (3.12), g^{kk} can be compactly expressed in terms of a^{kk}, $\hat{\boldsymbol{\beta}}_{SWV(k)}$ and the lower right component of matrix $\boldsymbol{B}$ and the model variance of $\hat{\boldsymbol{\beta}}_{SWV_k}$ is:

$$\begin{aligned} Var_M(\hat{\beta}_{SWV_k}) &= \sigma^2 g^{kk} \\ &= \sigma^2 (a^{kk})^2 (b_{kk} - 2\boldsymbol{b}_{k(k)} \hat{\boldsymbol{\beta}}_{SWV(k)} + \hat{\boldsymbol{\beta}}_{SWV(k)}^T \boldsymbol{B}_{(k)(k)} \hat{\boldsymbol{\beta}}_{SWV(k)}) \\ &= \sigma^2 \left(\frac{1}{1 - R^2_{SWV(k)}} \frac{1}{\tilde{\tilde{\boldsymbol{x}}}_k^T \tilde{\tilde{\boldsymbol{x}}}_k} \right)^2 \times \\ &\quad (\tilde{\tilde{\boldsymbol{x}}}_k^T \boldsymbol{W} \tilde{\tilde{\boldsymbol{x}}}_k - 2\tilde{\tilde{\boldsymbol{x}}}_k^T \boldsymbol{W} \tilde{\boldsymbol{X}}_{(k)} \hat{\boldsymbol{\beta}}_{SWV(k)} + \hat{\boldsymbol{\beta}}_{SWV(k)}^T \tilde{\boldsymbol{X}}_{(k)} \boldsymbol{W} \tilde{\boldsymbol{X}}_{(k)} \hat{\boldsymbol{\beta}}_{SWV(k)}) \\ &= \sigma^2 \frac{(\tilde{\tilde{\boldsymbol{x}}}_k - \tilde{\tilde{\boldsymbol{X}}}_{(k)} \hat{\boldsymbol{\beta}}_{SWV(k)})^T \boldsymbol{W} (\tilde{\tilde{\boldsymbol{x}}}_k - \tilde{\tilde{\boldsymbol{X}}}_{(k)} \hat{\boldsymbol{\beta}}_{SWV(k)})}{\left[(1 - R^2_{SWV(k)}) \tilde{\tilde{\boldsymbol{x}}}_k^T \tilde{\tilde{\boldsymbol{x}}}_k \right]^2} \\ &= \frac{\zeta_k \varrho_k}{1 - R^2_{SWV(k)}} \frac{\sigma^2 \tilde{\tilde{\boldsymbol{x}}}_k^T \boldsymbol{W} \tilde{\tilde{\boldsymbol{x}}}_k}{(\tilde{\tilde{\boldsymbol{x}}}_k^T \tilde{\tilde{\boldsymbol{x}}}_k)^2}, \end{aligned} \tag{3.76}$$

where $\frac{\sigma^2 \tilde{\tilde{\boldsymbol{x}}}_k^T \boldsymbol{W} \tilde{\tilde{\boldsymbol{x}}}_k}{(\tilde{\tilde{\boldsymbol{x}}}_k^T \tilde{\tilde{\boldsymbol{x}}}_k)^2}$ is the model variance of $\hat{\boldsymbol{\beta}}_{SWV(k)}$ when the columns of $\boldsymbol{X}$ are orthogonal, $\zeta_k = \frac{(\tilde{\tilde{\boldsymbol{x}}}_k - \tilde{\tilde{\boldsymbol{X}}}_{(k)} \hat{\boldsymbol{\beta}}_{SWV(k)})^T \boldsymbol{W} (\tilde{\tilde{\boldsymbol{x}}}_k - \tilde{\tilde{\boldsymbol{X}}}_{(k)} \hat{\boldsymbol{\beta}}_{SWV(k)})}{(\tilde{\tilde{\boldsymbol{x}}}_k - \tilde{\tilde{\boldsymbol{X}}}_{(k)} \hat{\boldsymbol{\beta}}_{SWV(k)})^T (\tilde{\tilde{\boldsymbol{x}}}_k - \tilde{\tilde{\boldsymbol{X}}}_{(k)} \hat{\boldsymbol{\beta}}_{SWV(k)})}$, $\varrho_k = \frac{\tilde{\tilde{\boldsymbol{x}}}_k^T \tilde{\tilde{\boldsymbol{x}}}_k}{\tilde{\tilde{\boldsymbol{x}}}_k^T \boldsymbol{W} \tilde{\tilde{\boldsymbol{x}}}_k}$. ζ_k, ϱ_k are two adjustment coefficients involving $\boldsymbol{W}$. These terms are bounded using the maximum eigenvalue of the matrix of survey weights:

$$\zeta_k \leq \mu_{max}(\boldsymbol{W})$$

$$\varrho_k \geq \frac{1}{\mu_{max}(\boldsymbol{W})}.$$

Since $\boldsymbol{W}$ is diagonal, $\mu_{max}(\boldsymbol{W}) = max_{k\in s}(w_k)$, i.e. the largest survey weight.

Analogous to Section 3.1.5, the intercept-adjusted ϱ_k and $R^2_{SW(k)}$ are:

$$\varrho_{mk} = \frac{(\tilde{\tilde{\boldsymbol{x}}}_k^T \tilde{\tilde{\boldsymbol{x}}}_k - \hat{N}\bar{\tilde{\tilde{x}}}_k^2)}{(\tilde{\tilde{\boldsymbol{x}}}_k - \tilde{\tilde{\mathbf{1}}}\bar{\tilde{\tilde{x}}}_k)^T \boldsymbol{W} (\tilde{\tilde{\boldsymbol{x}}}_k - \tilde{\tilde{\mathbf{1}}}\bar{\tilde{\tilde{x}}}_k)}, \quad \text{where } \hat{N} = \tilde{\tilde{\mathbf{1}}}^T \tilde{\tilde{\mathbf{1}}} \text{ and } \bar{\tilde{\tilde{x}}}_k = \tilde{\tilde{\mathbf{1}}}^T \tilde{\tilde{\boldsymbol{x}}}_k / \hat{N},$$

and

$$R^2_{SWm(k)} = \frac{\hat{\boldsymbol{\beta}}^T_{SWV(k)} \boldsymbol{X}^T_{(k)} \boldsymbol{X}_{(k)} \hat{\boldsymbol{\beta}}_{SWV(k)}}{(\tilde{\tilde{\boldsymbol{x}}}_k^T \tilde{\tilde{\boldsymbol{x}}}_k - \hat{N}\bar{\tilde{\tilde{x}}}_k^2)}.$$

3.2 Experimental Study

3.2.1 Introduction

To investigate the effectiveness of the proposed and modified collinearity diagnostics techniques, we will apply them to real survey data and then conduct appropriate evaluations. Two survey data sets are employed: the 1998 Survey of Mental Health Organizations (SMHO) and the 2001-2002 National Health and Nutrition Examination Survey (NHANES).

3.2.2 Survey of Mental Health Organizations

3.2.2.1 Description of Study Population

The 1998 Survey of Mental Health Organizations (SMHO) data file contains N=875 mental health organizations. The survey is described in more detail in Li &

Valliant (2009) and Manderscheid & Henderson (2002). Three variables were chosen as explanatory variables for our true model, including the number of clients/patients enrolled on the first day of the reporting year (FIRST), the number of additions of patients or clients during the reporting year (ADDS) and the number of clients/patients on the open and active rolls at the end of the reporting year (EOYCNT). They are relatively low-correlated with each other. An organization type (DSTRAT) variable is selected as the stratum variable in the later sampling selection, which has four categories: (1) psychiatric hospitals, (2) residential hospitals, (3) general hospitals and (4) Department of Veteran Affairs medical centers, and other organizations. To construct a population for this study that is less affected by influential points, cases with extreme values of auxiliary variables were excluded. In the retained population, FIRST ranges from 3 to 400, ADDS ranges from 10 to 6000 and EOYCNT ranges from 10 to 6000. A total of 410 cases remained. Due to the large value ranges, the square-root transformation was used for ADDS and FIRST, and the fourth root transformation was used for EOYCNT. In the models that follow, these transformations were used, but we will still refer to them as ADDS, FIRST and EOYCNT to simply the notation.

To create the study population, a bootstrap method was applied by taking N=2,000 random observations with replacement from this data set. From the sampled set, values of the Y variable were generated by the three explanatory variables using Gamma distributions: $Y_i \sim Gamma(\alpha_i, \beta_i)$, with shape parameters $\alpha_i = (\boldsymbol{x}_i^T\boldsymbol{\beta})^2/\sigma^2 v_i$ and scale parameters $\beta_i = \sigma^2 v_i/\boldsymbol{x}_i^T\boldsymbol{\beta}$, $\boldsymbol{x}_i = (1, \text{FIRST}_i, \text{ADDS}_i, \text{EOYCNT}_i)^T$, $\boldsymbol{\beta} = (4, 1, 1.5, 2.5)^T$, $\sigma^2 = 10$, and $v_i = \text{DSTRAT}_i \times \text{ADDS}_i$, where i

stands for the i^{th} case in the data set and DSTRAT$_i$ is coded 1, 2, 3 or 4 depending on the stratum. That is, the underlying model is:

$$Y_i = \beta_0 + \beta_{\text{FIRST}} \times \text{FIRST}_i + \beta_{\text{ADDS}} \times \text{ADDS}_i \ + \beta_{\text{EOYCNT}} \times \text{EOYCNT}_i + \varepsilon_i$$

with independent errors and $Var_M(Y_i) = \sigma^2 v_i$. The population variance matrix is then $\boldsymbol{V} = diag(v_i)$.

To create variables that are collinear with FIRST, ADDS and EOYCNT, two independent variables were constructed:

$$X_{1i} = C_{1i} + \ddot{e}_{1i},$$
$$X_{2i} = C_{2i} + \ddot{e}_{2i}$$

where $C_{1i} = \text{FIRST}_i + 0.25\text{ADDS}_i + 0.05\text{EOYCNT}_i$ and $C_{2i} = 0.2\text{FIRST}_i + 0.05\text{ADDS}_i + \text{EOYCNT}_i$.

The error terms have means of 0 and

$$Var_M(\ddot{e}_{1i}) = (1/\gamma_1^2 - 1)\tau_i var(C_{1i}),$$
$$Var_M(\ddot{e}_{2i}) = (1/\gamma_2^2 - 1)\tau_i var(C_{2i})$$

with $\tau_i = \frac{v_i - min(v_i)+1}{max(v_i)-min(v_i)+1}$, $\gamma_1 = 0.90$, $\gamma_2 = 0.80$ and $var(C_{ki}),\ k = 1, 2.$, denotes the variance of C_{ki} across all 2,000 cases in the finite population.

Let $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{C}_1$, and $\boldsymbol{C}_2$ be the 2000×1 population vectors of X_{1i}, X_{2i}, C_{1i} and C_{2i}. In our generated population, $\boldsymbol{X}_1$ and $\boldsymbol{C}_1$ have correlation equal to 0.97. $\boldsymbol{X}_2$ and $\boldsymbol{C}_2$ have correlation equal to 0.93.

As noted earlier, when there are collinear variables, several alternative models may give similar fits as measured by R^2. Table 3.2 lists the values of R^2 for several combinations of independent variables. The value of R^2 was computed as $1 - SSE/SST$, where $SSE = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$, $\boldsymbol{Y}$ and $\boldsymbol{X}$ are the finite population values of y-vector and $\boldsymbol{X}$-matrix, and $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. The total sum of squares was computed as $SST = \boldsymbol{Y}^T\boldsymbol{Y}$. In the simulation described in section 3.2.2.3, the models that included the intercept, FIRST, ADDS and EOYCNT (F+A+E) or the intercept, FIRST, ADDS, EOYCNT, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ (F+A+E+$\boldsymbol{X}_1$+$\boldsymbol{X}_2$) were selected over 94% of the time regardless of method of VIF calculation. Thus, including or excluding the collinear variables has little effect on this measure of overall fit.

Table 3.2: R-Square for Several Selected Models in Different Samples in the Simulations

Variable	Model					
	1	2	3	4	5	6
Intercept	√	√	√	√	√	√
FIRST	√	√	√	√		√
ADDS	√	√	√	√	√	
EOYCNT	√	√	√	√	√	√
$\boldsymbol{X}_1$		√		√	√	√
$\boldsymbol{X}_2$			√	√		
$R2$	0.4282	0.4285	0.4283	0.4285	0.4283	0.3938

3.2.2.2 Collinearity in the Study Population

We will identify the presence of collinearity and their influence on the estimation of the true model for our study population at first, before we start to evaluate the performance of our diagnostic methods in the sample subsets. Figure 3.1 displays the pairwise scatterplots, single variable histograms and correlation coefficients r. The three variables in the true model, FIRST, ADDS, EOYCNT, have low correlations in the scatterplots that are all smaller than 0.25. However, $\boldsymbol{X}_1$, as an additional variable not in the true model, has relatively high correlations with two variables, FIRST and ADDS, in the true model ($r_{\text{FIRST},\boldsymbol{X}_1} = 0.72$ and $r_{\text{ADDS},\boldsymbol{X}_1} = 0.74$) and another additional variable, $\boldsymbol{X}_2$, has a moderate or relatively high correlation with all the three variables in the true model ($r_{\text{FIRST},\boldsymbol{X}_2} = 0.53$, $r_{\text{ADDS},\boldsymbol{X}_2} = 0.54$) and $r_{\text{EOYCNT},\boldsymbol{X}_2} = 0.77$).

Table 3.3 shows results of regressions, including estimated slopes and standard errors, fitted using the full finite population. Thus, we treated the population as a random realization from the model so that fitted slopes are estimates that have variances. The table includes the estimated coefficients and VIF values from the regression of $\boldsymbol{Y}$ on the three variables in the true model (FIRST, ADDS and EOYCNT) and on the five variables including collinear variables, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Recall that, when we generated $\boldsymbol{Y}$ in the previous section, the variance-covariance matrix is set as $\boldsymbol{V} = \sigma^2 \times diag(\text{DSTRAT}_i * \text{ADDS}_i)$. To estimate the model variances and standard errors of OLS coefficients, three alternatives were used:

OLS1: OLS formulas used to estimate $\boldsymbol{\beta}$, the variance of $\boldsymbol{\beta}_{OLS}$ and the VIF are estimated from the standard packages;

OLS2: OLS used to estimate $\boldsymbol{\beta}$; the variance of $\boldsymbol{\beta}_{OLS}$ and the VIF are estimated under the model in section 3.2.2.1 assuming that heteroscedastic variances are known;

OLS3: OLS used to estimate $\boldsymbol{\beta}$; the variance of $\boldsymbol{\beta}_{OLS}$ and the VIF are estimated using an estimator of $\boldsymbol{V}$ matrix based on squared model residuals.

The standard errors obtained from OLS1 are smaller than their corresponding standard errors from OLS2 and OLS3, due to the fact that OLS1 assumes the homoscedasticity of the errors and only uses $\hat{\sigma}^2$ in the estimation of standard errors, while OLS2 and OLS3 allow the heterogeneity of the errors and use $\sigma^2\boldsymbol{V}$ or $\hat{\boldsymbol{V}}$ in the estimation of standard errors. The coefficients of three variables in the true model are significant in the three-variable model using all of the methods. However, in the five-variable model, the coefficient of FIRST became insignificant. Although not shown in Table 3.3, the coefficients of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are also significant when there is only one of them in the model without all the other explanatory variables. But both $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are insignificant in the five-variable model. Their significance is affected by the high collinearity in the five-variable model.

Parallel to each OLS regression, the intercept-adjusted VIF values are computed according to their ways of model variance estimation. For OLS1, we used $\text{VIF} = \frac{1}{1-R^2_{m(k)}}$, which is the VIF formula used in the standard packages; For OLS2, we used $\text{VIF} = \frac{\zeta_k \varrho_k}{1-R^2_{SWm(k)}}$ in (3.19); For OLS3, we used $\text{VIF} = \frac{\hat{\zeta}_k \hat{\varrho}_k}{1-R^2_{SWm(k)}}$ in (3.69). The VIF values from regressions with three variables in the upper tier of Table 3.3 are all close to 1, as a result of the low correlations of the $\boldsymbol{X}$'s in the underlying model. OLS3 has VIF values that are even slightly smaller than 1. In the five variable regression, the last column in Table 3.3, which is the VIF for the OLS $\hat{\boldsymbol{\beta}}$ computed using an estimate of $\boldsymbol{V}$, is the best reflection of the increase in variance due to collinearity in the realized finite population.

In the five variable regression, $\boldsymbol{X}_1$ has the highest VIF value compared to the other explanatory variables and its VIF value in OLS2 is the highest (22.06) out of its VIF values in all the three OLSs of regressions. In OLS1, ADDS has a higher VIF value than FIRST, comparing 9.22 to 8.73. However, in OLS2 and OLS3, the VIF values of FIRST are slightly higher than the ones of ADDS. This implies that neglecting the correct covariance structure does not create a consistent pattern in VIFs. We may either underestimate or overestimate VIF values when we ignore $\boldsymbol{V}$ or $\hat{\boldsymbol{V}}$ in the computation of VIF and instead use the VIFs produced in the standard packages.

Figure 3.1: Scatterplots and Correlation Coefficients of Five Explanatory Variables in the Artificial Population based on the SMHO Data Set

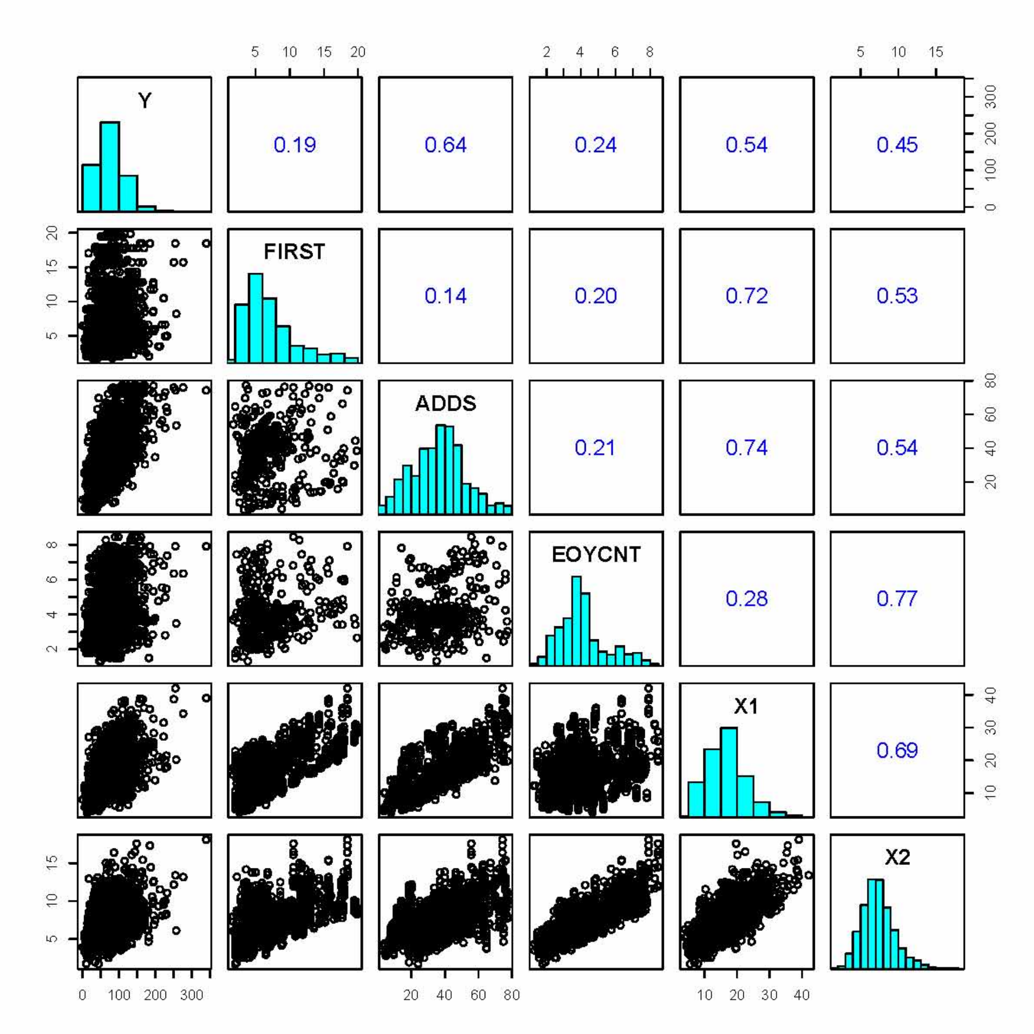

Table 3.3: VIFs of the Three-Variable Regression and Five-Variable Regression in the Full Finite Population. $\hat{\boldsymbol{\beta}}_{OLS}$ used in all cases; three methods of variance estimation used.

Variable	**Underlying true model parameters**	**Coeff.**	**OLS1** estimated σ^2		**OLS2** known $\sigma^2\boldsymbol{V}$		**OLS3** estimated $\boldsymbol{V}$	
			SE	**VIF**[a]	SE	**VIF**[b]	SE	**VIF**[c]
Intercept	4	4.83	2.45*[d]		2.78		2.92	
FIRST	1	0.83	0.18***	1.05	0.17***	1.03	0.18***	0.95
ADDS	1.5	1.50	0.04***	1.07	0.05***	0.98	0.05***	0.93
EOYCNT	2.5	2.58	0.49***	1.08	0.60***	1.01	0.57***	0.96
Intercept	4	4.79	2.46		2.78		2.93	
FIRST	1	0.40	0.50	8.73	0.61	12.66	0.58	10.30
ADDS	1.5	1.39	0.13***	9.22	0.15***	11.45	0.15***	10.09
EOYCNT	2.5	2.33	0.95*	4.01	1.19*	3.99	1.16*	4.06
$\boldsymbol{X}_1$		0.39	0.44	16.66	0.55	22.06	0.53	16.15
$\boldsymbol{X}_2$		0.21	0.81	7.41	1.02	7.82	1.04	7.09

[a] $\text{VIF} = \frac{1}{1-R^2_{(k)}}$

[b] $\text{VIF} = \frac{\zeta_k \varrho_k}{1-R^2_{(k)}}$, with $\zeta_k = \frac{\boldsymbol{e}_{xk}^T \boldsymbol{V} \boldsymbol{e}_{xk}}{\boldsymbol{e}_{xk}^T \boldsymbol{e}_{xk}}$ and $\varrho_k = \frac{\boldsymbol{x}_k^T \boldsymbol{x}_k}{\boldsymbol{x}_k^T \boldsymbol{V} \boldsymbol{x}_k}$

[c] $\text{VIF} = \frac{\hat{\zeta}_k \hat{\varrho}_k}{1-R^2_{(k)}}$, with $\hat{\zeta}_k = \frac{\boldsymbol{e}_{xk}^T \hat{\boldsymbol{V}} \boldsymbol{e}_{xk}}{\boldsymbol{e}_{xk}^T \boldsymbol{e}_{xk}}$ and $\hat{\varrho}_k = \frac{\boldsymbol{x}_k^T \boldsymbol{x}_k}{\boldsymbol{x}_k^T \hat{\boldsymbol{V}} \boldsymbol{x}_k}$

[d] p values of significance: *$p = 0.05$; **$p = 0.01$; ***$p = 0.005$.

3.2.2.3 Simulation

To study the properties of the different estimations of VIFs, we conducted a simulation study. The sample design used in this simulation is a stratified single-stage sample from the study population. Strata were formed based on four organization types, which are defined by stratum variable DSTRAT. Fifty sample units in each stratum were then selected with probability proportional to the rounded value of the size of EOYCNT multiplied by 100 so that in each sample, 200 cases were drawn without replacement. The method of selection was the one due to Hartley & Rao (1962), in which the population order is randomized and a systematic sample of units is selected. Sample weights were calculated as the inverses of the selection probabilities. The variables used in our analysis are: Y (generated), FIRST, ADDS, EOYCNT, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$.

To evaluate the performance of our new VIF methods, we performed 1,000 simulations, in which each sample was drawn using the sampling design described above and then four types of regressions with their corresponding VIF methods were applied on this sample. In each simulation, We used cut-off values for the VIF equal to 2, 7, 10 separately to delete collinear variables and keep the final reduced model by the backward elimination.

The four types of regressions and their VIF methods are listed in Table 3.4. TYPE1 is a ordinary least squares (OLS) regression with estimated σ^2 (unknown); TYPE2 is a weighted least squares (WLS) regression with estimated σ^2 (unknown) and assuming $\boldsymbol{V} = \boldsymbol{W}^{-1}$, where $\boldsymbol{W}$ is the survey weight matrix. The VIFs for the first two types of regressions can be obtained from the standard statistical packages. Note that assuming $\boldsymbol{V} = \boldsymbol{W}^{-1}$ is incorrect since $\boldsymbol{V}$ in the underlying model depends on DSTRAT and ADDS not EOYCNT, which was used to determine selection probabilities.

TYPE3 is a SWLS with known $\sigma^2\boldsymbol{V}$. We computed both its model-based VIFs (as derived in Section 3.1.5) and design-based VIFs (as derived in Section 3.1.6). and TYPE4 is a SWLS with estimated $\hat{\boldsymbol{V}}$, when $\sigma^2\boldsymbol{V}$ is unknown. We computed its model-based VIFs and design-based VIFs by (3.69) and (3.70) separately, in correspondence with the stratified single-stage sampling design in this study. TYPE4 is the most practical situation because $\boldsymbol{V}$ is always unknown.

Using different regression types, their corresponding VIF formulas and VIF cut-off values, summary statistics across the 1,000 simulations include:

1) The average VIFs of the models with all five explanatory variables (these should be compared to the VIFs from the full finite population in Table 3.5).

2) The percentages of samples where different final models were selected (shown in Table 3.7).

3) The percentage of 95% confidence intervals of $\hat{\beta}_{ki}$ that include the full finite population parameters β_{Uk} of the underlying model, where k stands for a given variable (intercept, FIRST, ADDS or EOYCNT). $\hat{\beta}_{ki}$ is the estimate of parameter for variable k in the final models selected in sample i. The finite population parameters of the underlying model, $\boldsymbol{\beta}_U = (\text{Intercept}, \beta_{\text{FIRST}U}, \beta_{\text{ADDS}U}, \beta_{\text{EOYCNT}U})^T = (4.83, 0.83, 1.50, 2.58)^T$, were shown in Table 3.3. The confidence intervals for β_k in sample i were computed as $\hat{\beta}_{ki} \pm 1.96\sqrt{v(\hat{\beta}_{ki})}$. The confidence interval coverage was

computed among samples that included variable k in the model (shown in Table 3.7).

4) The percentage of 95% confidence regions based on $\hat{\boldsymbol{\beta}}$ that include the full finite population parameter $\boldsymbol{\beta}_U$ of the underlying model. $\hat{\boldsymbol{\beta}}_i$ is the estimate of the parameter vector of the final model selected in sample i, $\hat{\boldsymbol{\beta}}_i = (\text{Intercept}, \hat{\beta}_{\text{FIRST}U}, \hat{\beta}_{\text{ADDS}U}, \hat{\beta}_{\text{EOYCNT}U}, \hat{\beta}_{\boldsymbol{X}_1 U}, \hat{\beta}_{\boldsymbol{X}_2 U})^T$. When one variable is omitted in the final model, we set its estimated parameter to zero. The finite population parameters of the underlying model,

$$\begin{aligned}\boldsymbol{\beta}_U &= (\text{Intercept}, \beta_{\text{FIRST}U}, \beta_{\text{ADDS}U}, \beta_{\text{EOYCNT}U}, \beta_{\boldsymbol{X}_1 U}, \beta_{\boldsymbol{X}_2 U}) \\ &= (4.83, 0.83, 1.50, 2.58, 0, 0)^T\end{aligned}$$

.

The confidence regions of $\hat{\boldsymbol{\beta}}$ were computed as

$$n(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_U)^T var_M^{-1}(\hat{\boldsymbol{\beta}}_i)(\hat{\boldsymbol{\beta}}_i - \boldsymbol{\beta}_U) < \chi^2_{6,0.05}$$

among 1,000 simulations (shown in Table 3.8). The variance estimator $var_M(\hat{\beta}_i)$ was the unclustered version of (3.55) which is approximately the same as the design-based linearization estimator in (3.61).

5) The average parameter estimates and their relative biases to the underlying model. For a given variable, the relative bias was estimated by $relbias(\hat{\beta}_k) = bias(\hat{\beta}_k)/\boldsymbol{\beta}_{Uk} = (\bar{\hat{\beta}}_k - \beta_{Uk})/\beta_{Uk}$, where $\bar{\hat{\beta}}_k = \sum_{i=1}^{S_k} \hat{\beta}_i / S_k$ and S_k is the number

of samples that included variable k in the model.

6) The ratios of average estimated standard errors of model parameter estimates to the empirical standard errors. For a given variable k, the average estimated standard error of $\hat{\beta}_{ki}$ was calculated as $se(\hat{\beta}_k) = \sum_i se(\hat{\beta}_{ki})/S_k$, where $se(\hat{\beta}_{ki})$ is the estimated standard error of $\hat{\beta}_{ki}$ which was calculated at the i^{th} simulation. The empirical standard error of $\hat{\beta}_k$ was defined as $SE(\hat{\beta}_{ki}) = \sqrt{\sum_i(\hat{\beta}_{ki} - \bar{\hat{\beta}}_k)/S_k}$ (shown in Table 3.9).

Table 3.4: Four Types of Linear Regressions and their VIF Formulas Used in the Simulation

TYPE	Parameter Estimation	Model-Based Variance Estimation of $\hat{\boldsymbol{\beta}}$	VIF fomula
TYPE1	OLS with $\hat{\sigma}^2$ $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, $Var_M(\boldsymbol{Y}) = \sigma^2$	$\hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$	$\mathrm{VIF} = \frac{1}{1-R^2_{(k)}}$
TYPE2	WLS with $\hat{\sigma}^2$ and $\boldsymbol{W}$ $\hat{\boldsymbol{\beta}}_{SW} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y}$, $Var_M(\boldsymbol{Y}) = \sigma^2\boldsymbol{V} = \sigma^2\boldsymbol{W}^{-1}$	$\hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$	$\mathrm{VIF} = \frac{1}{1-R^2_{SW(k)}}$
TYPE3	SWLS with $\sigma^2\boldsymbol{V}$ and $\boldsymbol{W}$ $\hat{\boldsymbol{\beta}}_{SW} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y}$, $Var_M(\boldsymbol{Y}) = \sigma^2\boldsymbol{V}$	$\sigma^2(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{V}\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$	MB : $\mathrm{VIF} = \frac{\zeta_k\varrho_k}{1-R^2_{SW(k)}}$, with $\zeta_k = \frac{\boldsymbol{e}^T_{xk}\boldsymbol{W}\boldsymbol{V}\boldsymbol{W}\boldsymbol{e}_{xk}}{\boldsymbol{e}^T_{xk}\boldsymbol{W}\boldsymbol{e}_{xk}}$ and $\varrho_k = \frac{\boldsymbol{x}^T_k\boldsymbol{W}\boldsymbol{x}_k}{\boldsymbol{x}^T_k\boldsymbol{W}\boldsymbol{V}\boldsymbol{W}\boldsymbol{x}_k}$ DB : $\widehat{\mathrm{VIF}} = \frac{\hat{\zeta}_k\hat{\varrho}_k}{1-\hat{R}^2_{SW(k)}}$, with $\hat{\zeta}_k = \frac{\boldsymbol{e}^T_{xk}\boldsymbol{W}^{*}\cdot\boldsymbol{V}\boldsymbol{e}_{xk}}{\boldsymbol{e}^T_{xk}\boldsymbol{W}\boldsymbol{e}_{xk}}$ and $\hat{\varrho}_k = \frac{\boldsymbol{x}^T_k\boldsymbol{W}\boldsymbol{x}_k}{\boldsymbol{x}^T_k\boldsymbol{W}^{*}\cdot\boldsymbol{V}\boldsymbol{x}_k}$
TYPE4	SWLS with $\hat{\boldsymbol{V}}$ and $\boldsymbol{W}$ $\hat{\boldsymbol{\beta}}_{SW} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y}$, $Var_M(\boldsymbol{Y}) = \sigma^2\boldsymbol{V}$	$\hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\hat{\boldsymbol{V}}\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$	MB : $\widehat{\mathrm{VIF}} = \frac{\hat{\zeta}_k\hat{\varrho}_k}{1-R^2_{SW(k)}}$, with $\hat{\zeta}_k = \frac{\boldsymbol{e}^T_{xk}\boldsymbol{W}\hat{\boldsymbol{V}}\boldsymbol{W}\boldsymbol{e}_{xk}}{\boldsymbol{e}^T_{xk}\boldsymbol{W}\boldsymbol{e}_{xk}}$ and $\hat{\varrho}_k = \frac{\boldsymbol{x}^T_k\boldsymbol{W}\boldsymbol{x}_k}{\boldsymbol{x}^T_k\boldsymbol{W}\hat{\boldsymbol{V}}\boldsymbol{W}\boldsymbol{x}_k}$ where $\hat{\boldsymbol{V}} = diag(e^2_{hi})$ DB : $\widehat{\mathrm{VIF}} = \frac{\hat{\zeta}_k\hat{\varrho}_k}{1-R^2_{SW(k)}}$, with $\hat{\zeta}_k = \frac{\boldsymbol{e}^T_{xk}\boldsymbol{W}\hat{\boldsymbol{V}}\boldsymbol{W}\boldsymbol{e}_{xk}}{\boldsymbol{e}^T_{xk}\boldsymbol{W}\boldsymbol{e}_{xk}}$ and $\hat{\varrho}_k = \frac{\boldsymbol{x}^T_k\boldsymbol{W}\boldsymbol{x}_k}{\boldsymbol{x}^T_k\boldsymbol{W}\hat{\boldsymbol{V}}\boldsymbol{W}\boldsymbol{x}_k}$ where $\hat{\boldsymbol{V}} = \sum_{h=1}^{H}\frac{n_h}{n_h-1}\left[diag(e^2_{hi}) - \frac{1}{n_h}\boldsymbol{e}_h\boldsymbol{e}^T_h\right]$

Table 3.5: Comparison between the VIFs from the Full Finite Population and the Average VIFs from 1,000 Simulations

	OLS	Using the true $\boldsymbol{V}$ or estimated $\hat{\boldsymbol{V}}$ in the regression							
	OLS1	TYPE1 OLS	OLS2 known $\boldsymbol{V}$	OLS3 est. $\hat{\boldsymbol{V}}$	TYPE2 WLS	TYPE3 SWLS with $\boldsymbol{V}$		TYPE4 SWLS with $\hat{\boldsymbol{V}}$	
						Model-based	Design-based	Model-based	Design-based
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	Pop.[a]	Mean[b]	**Pop.**	**Pop.**	Mean	Mean	Mean	Mean	Mean
FIRST	**8.73**	10.94	**12.66**	**10.3**	9.10	13.72	13.15	12.43	12.43
ADDS	**9.22**	11.10	**11.45**	**10.09**	9.61	11.71	11.84	11.20	11.20
EOYCNT	**4.01**	4.53	**3.99**	**4.06**	4.14	4.31	4.08	4.46	4.46
$\boldsymbol{X}_1$	**16.66**	18.63	**22.06**	**16.15**	17.44	24.63	23.02	21.04	21.04
$\boldsymbol{X}_2$	**7.41**	8.44	**7.82**	**7.09**	7.67	8.67	8.06	8.61	8.61

[a]VIFs from the full finite population
[b]The average VIFs from 1000 simulations

Table 3.5 reports the average VIFs of the five-variable model from 1,000 simulations and compares them with the VIFs of the five-variable model from the full finite population. There are some correspondences between the columns for population values and simulation means in Table 3.5, listed as bellow:

Simulation mean	Corresponding population values
2	none
5	**1**
6,7	**3**
8,9	**4**

The population VIFs in column **1** would be correct for the OLS estimates of $\boldsymbol{\beta}$ if $\boldsymbol{V} = \sigma^2\boldsymbol{I}$, i.e. the underlying model has homogeneous variances which, as noted earlier, is not the case. The column **5** VIFs, which use the survey weights, are design-based estimates of the (incorrect) population VIFs in column **1**. The simulation means in column **2** are OLS VIFs and do not estimate, in a repeated-sampling sense, any of the population VIFs in Table 3.3. Recall that in Section 3.1.6, the derivation of the design-based VIFs with known $\boldsymbol{V}$ in a sample selected from the full finite population aimed to estimate the model-based VIFs when the full finite population is in the samples, by substituting design-based estimates of each of the components of the model variance. Thus, the simulation means of the TYPE3 VIFs in column **7** estimate the population VIFs in column **3**, which are computed using the underlying $\boldsymbol{V}$. Numerically, the model-based VIFs in column **6** are very close to those in column **7**. The means of the TYPE4 VIFs in columns **8** and **9** use the survey weights and $\hat{\boldsymbol{V}}$; they approximately estimate the population VIFs in column **4** which are computed using the population $\hat{\boldsymbol{V}}$.

If survey weights are used to estimate $\boldsymbol{\beta}$, i.e, $\hat{\boldsymbol{\beta}}_{SW}$ is used, the VIF estimates that will robustly reflect an unknown underlying variance structure are those in columns **8** and **9**. Note that the means in columns **8** and **9** are the same to two decimal places since $\hat{\boldsymbol{V}} = diag(e_{hi}^2)$ in (3.69) and $\hat{\boldsymbol{V}} = \sum_{h=1}^{H} \frac{n_h}{n_h-1} \left[diag(e_{hi}^2) - \frac{1}{n_h} \boldsymbol{e}_h \boldsymbol{e}_h^T \right]$ in (3.70) are numerically very close.

In Table 3.5, the explanatory variable, $\boldsymbol{X}_1$, always had the highest VIF value across the ones in the full finite population and the average VIFs from 1,000 simulation no matter which regression types were used. The high VIF is caused by its very high correlation with the linear combination, $\boldsymbol{C1}$, of three explanatory variables in the true model (FIRST, ADDS and EOYCNT). This is consistent with the results shown in Table 3.7, in which $\boldsymbol{X}_1$ was deleted in most of simulations when VIF cut-off value is 10 or smaller. However, the variable that had the second highest VIF value varied across regression types. In OLS1, TYPE1(OLS) and TYPE2(WLS), the VIF values were obtained from the standard packages and the explanatory variable, ADDS, had the second highest VIF value. When, using the other regression methods, OLS2, OLS3, TYPE3 (SWLS with $\boldsymbol{V}$) and TYPE4 (SWLS with $\hat{\boldsymbol{V}}$), the VIF values were obtained from our new approaches. The explanatory variable, FIRST, had the second highest VIF values. This difference indicates that the collinearity among these five variables had different impact on estimating parameters of interest in the model, when different regression methods were used.

Starting from the five-variable model, the backward elimination method was adopted to delete collinear variables and select a reduced final model. The percentages of samples where different final models were selected are listed in Table 3.7. For comparison, the last tier in the table includes results of VIF $= \infty$, i.e. not dropping of collinear variables. When the VIF cut-off value was equal to 2, the two collinear variables, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, were deleted and the underlying model with FIRST, ADDS and EOYCNT was selected in almost all the simulations, no matter which regression type and inference approach were used. Only when using the VIFs for TYPE4 (SWLS with $\hat{\boldsymbol{V}}$), quite a few simulations selected models other than the underlying models. As noted earlier, TYPE4 VIFs are the ones that would be used in practice because $\boldsymbol{V}$ is never known. When VIF cut-off value was equal to 7, the reduced final model varied across different regression types and inference approaches in some simulations. Most of simulations still selected the underlying model, but due to the looser collinearity criteria, in some other simulations, $\boldsymbol{X}_2$ was kept in the final model. When cut-off value was equal to 10, $\boldsymbol{X}_2$ was kept in most of simulations. The cut-off value equal to 2 is always regarded as "too tight" as a criteria for VIF diagnostics method and a cut-off value equal to 7, 10 or even higher is recommended in the literature and is probably more acceptable to the analysts. But when using large VIF cut-off values, $\boldsymbol{X}_2$ will not be detected as a collinear variable, although its correlation of 0.93 with the linear combination $\boldsymbol{C}_2$ is quite high. Thus, the VIF diagnostics method might not be effective to detect some moderate or relatively high collinearity in the regression estimation, especially when the cut-off value is large. Other diagnostics method, such like condition indexes with variance decomposition method should be performed to detect these near-dependencies.

In some simulations, when using the same VIF cut-off values, a different final model was selected depending on the regression type and inference approach. For example, if an analyst ignores the sample design and uses OLS estimates with a cutoff of 10, the correct model (F+A+E) is selected only 7.8% of the time. If the survey-weighted least squares is used along with the TYPE4 design based VIF, F+A+E is selected 27.5% of the time. This difference confirms that the impact of data collinearity varies among the different regression methods. The variance of a given explanatory variable can be inflated more by the same collinear data in some regression methods than in others. The ranking of VIF values among these explanatory variables can also be different when different regression methods are used. One explanatory variable may have the largest VIF caused by the collinear data using, say, OLS while another explanatory variable may have the largest VIF when SWLS and a corresponding VIF are used. The collinearity we detected here is not data collinearity but system collinearity which is not only associated with the data collinearity but also conditional on the regression method, which involves survey weights, sampling design and variance matrix ($\boldsymbol{V}$ or $\hat{\boldsymbol{V}}$).

Table 3.6 lists the relative biases in the estimated model parameters. Variable, EOYCNT, had noticeably large negative biases, especially when VIF cut-off

Table 3.6: Relative Biases in Estimated Model Parameters Using Different Regression Types and VIF Cut-off Values

Coefficient	TYPE1: OLS	TYPE2: WLS	TYPE3: SWLS with $\boldsymbol{V}$		TYPE4: SWLS with $\hat{\boldsymbol{V}}$	
			Model-based	Design-based	Model-based	Design-based
Cut-off values for VIF=2						
Intercept	29.58%	-0.29%	-0.29%	-0.29%	0.41%	0.41%
FIRST	2.60%	2.62%	2.62%	2.62%	-3.02%	-3.05%
ADDS	0.33%	0.24%	0.24%	0.24%	0.13%	0.11%
EOYCNT	-12.55%	-2.14%	-2.14%	-2.14%	-2.21%	-2.20%
Cut-off values for VIF=7						
Intercept	34.48%	3.63%	3.87%	3.78%	5.07%	5.19%
FIRST	1.35%	-0.54%	-0.25%	-0.70%	-12.53%	-13.07%
ADDS	0.18%	-0.20%	-0.15%	-0.24%	-1.03%	-1.00%
EOYCNT	-13.93%	-4.14%	-3.89%	-4.59%	-14.77%	-14.77%
Cut-off values for VIF=10						
Intercept	39.34%	4.49%	4.12%	4.46%	4.81%	4.63%
FIRST	1.25%	-7.94%	-7.56%	-8.40%	-14.28%	-15.25%
ADDS	0.76%	-0.71%	-0.66%	-0.77%	-1.30%	-1.44%
EOYCNT	-6.44%	-10.73%	-10.36%	-11.67%	-13.65%	-14.75%

Note: relative bias was computed among samples that included a given variable in the model.

value is large. This is related to the fact that different variables were retained in the model using the VIF criteria. The larger the VIF cut-off value is, the more samples will keep the collinear variable, $\boldsymbol{X}_2$ in the final reduced model. As shown in Figure 3.1, EOYCNT and $\boldsymbol{X}_2$ have a positive correlation equal to 0.77 in the study population. The involvement of $\boldsymbol{X}_2$ in the model will degrade the effect of EOYCNT on the dependent variable and cause the negative bias of the estimated model parameter of EOYCNT. Furthermore, since $\boldsymbol{X}_2$ is also positively correlated with other explanatory variables in the underlying model, when the VIF cut-off value is high, FIRST and ADDS also have negative biases. It is interesting to see that when VIF cut-off value equals to 10, although more samples in TYPE4 chose the underlying model (25.90% under model-based inference and 27.50% under design-

Table 3.7: Percentage of Samples where Models were Selected and Coverage of 95% Confidence Intervals Using Different Regression Types and VIF Cut-off Values

Final model	Coefficient	TYPE1: OLS	TYPE2: WLS	TYPE3: SWLS with $\boldsymbol{V}$ Model-based	TYPE3: SWLS with $\boldsymbol{V}$ Design-based	TYPE4: SWLS with $\hat{\boldsymbol{V}}$ Model-based	TYPE4: SWLS with $\hat{\boldsymbol{V}}$ Design-based
Cut-off values for VIF=2							
Percentage of samples where model was selected							
F+A+E[a]		100.00%	100.00%	100.00%	100.00%	98.30%	98.20%
Other		0.00%	0.00%	0.00%	0.00%	1.70%	1.80%
Coverage of 95% confidence intervals[b]							
	Intercept	93.60%	92.80%	95.60%	95.60%	95.30%	95.20%
	FIRST	95.80%	95.50%	95.50%	95.50%	93.45%	93.34%
	ADDS	93.70%	91.70%	95.90%	95.90%	93.44%	93.54%
	EOYCNT	95.20%	92.80%	96.90%	96.90%	95.70%	95.70%
Cut-off values for VIF=7							
Percentage of samples where model was selected							
F+A+E		89.30%	69.60%	92.70%	81.10%	71.30%	71.80%
F+A+E+$\boldsymbol{X_2}$[c]		10.70%	30.40%	7.30%	18.90%	26.60%	26.10%
Other		0.00%	0.00%	0.00%	0.00%	2.10%	2.10%
Coverage of 95% confidence intervals							
	Intercept	90.50%	90.80%	94.90%	94.90%	94.50%	94.50%
	FIRST	95.30%	95.30%	95.80%	95.80%	92.60%	92.60%
	ADDS	94.20%	92.00%	95.90%	96.00%	93.03%	92.92%
	EOYCNT	93.90%	91.30%	96.80%	96.40%	94.10%	93.90%
Cut-off values for VIF=10							
Percentage of samples where model was selected							
F+A+E		7.80%	1.60%	13.00%	5.30%	25.90%	27.50%
F+A+E+$\boldsymbol{X_2}$		92.20%	98.30%	87.00%	94.70%	69.00%	67.20%
Other		0.00%	0.10%	0.00%	0.00%	5.10%	5.30%
Coverage of 95% confidence intervals							
	Intercept	91.90%	92.50%	95.60%	95.70%	94.80%	94.90%
	FIRST	92.70%	93.30%	96.70%	96.90%	93.30%	93.40%
	ADDS	94.40%	94.30%	97.20%	97.50%	94.98%	94.88%
	EOYCNT	90.90%	90.60%	94.60%	95.00%	93.30%	93.70%
Cut-off values for VIF=∞							
Percentage of samples where model was selected							
F+A+E+$\boldsymbol{X}_1$+$\boldsymbol{X}_2$[d]		100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Coverage of 95% confidence intervals							
	Intercept	92.50%	93.60%	96.80%	96.80%	95.80%	95.80%
	FIRST	89.50%	88.30%	95.50%	95.50%	91.70%	91.70%
	ADDS	90.00%	89.30%	95.90%	95.90%	92.70%	92.70%
	EOYCNT	91.00%	89.60%	96.10%	96.10%	94.50%	94.50%

[a]The underlying model: Y=Intercept+FIRST+ADDS+EOYCNT

[b]Note: confidence interval coverage was computed among samples that included a given variable in the model.

[c]Model: Y=Intercept+FIRST+ADDS+EOYCNT+$\boldsymbol{X_2}$

[d]Model: Y=Intercept+FIRST+ADDS+EOYCNT+$\boldsymbol{X}_1$+$\boldsymbol{X}_2$

based inference) than the ones in TYPE3 (13.00% under model-based inference and 5.30% under design-based inference) as shown in Table 3.7, the relative biases of FIRST and EOYCNT are larger in TYPE4 than the ones in TYPE3. Note that in TYPE4, 5.10% of the samples under the model-based inference and 5.30% of the samples under the design-based inference selected some models other than F+A+E or F+A+E+$\boldsymbol{X}_2$. In those samples, the final models in TYPE4 left one or more explanatory variables in the underlying model out of the final model but included $\boldsymbol{X}_1$ and/or $\boldsymbol{X}_2$ in the final model. A more detailed examination of the simulation results (not shown in Table 3.6), revealed that omitting variables in the underlying model led to even more severe biases in the estimated model parameters of the three explanatory variables in the underlying model, comparing to the samples where the final models included FIRST, ADDS and EOYCNT.

The coverage rates of the confidence intervals and the confidence regions on the true coefficient values are presented in Table 3.7 and Table 3.8. Empirical coverage rates are computed among samples that include a given variable in the model. For example, in Table 3.7, the coverage of 93.4% for FIRST with (TYPE4, cutoff=10, design-based) is among samples where FIRST was retained in the model. If samples that excluded a variable were counted as non-coverage, the percentage in Table 3.7 would be somewhat lower. TYPE3 (using SWLS with $\boldsymbol{V}$) always had the largest coverage rates of both confidence interval and confidence region than the other types. Some of its coverage rates were even above the nominal level (95%). This is also consistent with the results in Table 3.9 where the standard errors were overestimated for the regressions compared to the empirical standard errors in TYPE3. The overestimation for TYPE3 is related to using the variance matrix $\boldsymbol{V}$ in the variance estimation, which describes the random variation of the superpopulation from which our study population was drawn. When an estimated $\boldsymbol{V}$ is used in TYPE4, the standard errors are typically underestimated. TYPE2 (using WLS) always had the smallest coverage rates, by assuming $\boldsymbol{V} = \boldsymbol{W}^{-1}$. As we stated before, this assumption is improper in this study. This illustrates that if analysts intend to improve the variance estimation and related diagnostic methods by using the survey weights in WLS, as implemented in software for non-survey data, they will not only fail to achieve their goals but even make the estimation worse than the OLS estimation. In Table 3.9, some of the standard errors were underestimated in the TYPE1 (OLS) and TYPE4 regressions, in part due to the fact that the standard variance estimates did not account the possibility that the selected set of independent variables can differ from one sample to another. This is similar to the situation in stepwise regression (Hurvich & Tsai, 1990; Zhang, 1992). The same reason can also make the coverage rates of their confidence intervals and confidence regions smaller than the nominal level.

Table 3.8: Coverage Rates of the Confidence Regions for the True Coefficient Values in the Artificial Population based on SMHO

Cut-off value	TYPE1: OLS	TYPE2: WLS	TYPE3: SWLS with $\boldsymbol{V}$		TYPE4: SWLS with $\hat{\boldsymbol{V}}$	
			Model-based	Design-based	Model-based	Design-based
2	93.70%	92.50%	97.40%	97.40%	92.80%	92.50%
7	91.70%	89.80%	97.20%	97.30%	91.10%	91.00%
10	91.70%	89.60%	96.90%	97.00%	91.10%	91.20%
∞	87.70%	85.30%	96.60%	96.60%	89.30%	89.50%

Note: If a variable in the underlying model was left out of the final model for a sample, the region was counted as not covering the true parameters.

Table 3.9: Ratios of the Average Estimated $se(\hat{\boldsymbol{\beta}})$ to empirical $SE(\hat{\boldsymbol{\beta}})$ Using Different Regression Types and VIF Cut-off Values

Coefficient	TYPE1: OLS	TYPE2: WLS	TYPE3: SWLS with $\boldsymbol{V}$		TYPE4: SWLS with $\hat{\boldsymbol{V}}$	
			Model-based	Design-based	Model-based	Design-based
Cut-off values for VIF=2						
Intercept	94.48%	91.06%	104.98%	104.98%	103.36%	103.63%
FIRST	105.09%	105.60%	105.23%	105.23%	77.36%	77.60%
ADDS	95.73%	90.30%	105.02%	105.02%	95.73%	96.07%
EOYCNT	101.79%	91.25%	111.97%	111.97%	102.99%	103.30%
Cut-off values for VIF=7						
Intercept	88.45%	87.35%	101.05%	101.10%	99.15%	99.49%
FIRST	97.97%	97.90%	106.19%	104.28%	78.19%	77.16%
ADDS	92.75%	87.45%	104.37%	104.42%	92.57%	93.00%
EOYCNT	94.10%	82.21%	108.67%	104.70%	93.12%	92.96%
Cut-off values for VIF=10						
Intercept	94.15%	90.08%	103.52%	103.67%	101.41%	101.72%
FIRST	91.85%	93.82%	104.08%	105.07%	86.51%	87.16%
ADDS	92.58%	94.95%	114.00%	115.27%	97.28%	97.26%
EOYCNT	84.49%	83.94%	104.36%	105.54%	94.10%	94.25%

Note: ratio was computed among samples that included a given variable in the model.

3.2.3 National Health and Nutrition Examination Survey: 2001-2002

3.2.3.1 Description of the Data

The NHANES survey is conducted by the National Center for Health Statistics (NCHS) to assess the health and nutritional status of adults and children in the United States by both interviews and physical examinations. The interview includes demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by medical personnel. Thus, NHANES provides various quantitative and qualitative variables which are meaningful for regression analysis. NHANES uses a complex, multistage, probability sampling design, oversampling of certain population subgroups is done to increase the reliability and precision of health status indicator estimates for these groups. The data set used in our study is a subset of 2001-2002 data composed of respondents aged between 18 and 65. Observations with missing values are excluded from the sample which finally contains 3,011 complete respondents. The final weights in our sample range from 1,528.281 to 211,850.664, with a ratio of 139:1. The design of the sample can be approximated by the stratified selection of 30 PSUs from 15 strata, with 2 PSUs within each stratum. The sample size in each PSU is listed in Table 3.10.

3.2.3.2 Collinearity Diagnostics for NHANES 2001-2002

Five of the body measurement variables collected in the NHANES 2001-2002 physical examinations are: body weight(BMXWT)(kg), body mass index (BMI)(kg/m**2),

Table 3.10: Sample Size in each PSU in NHANES 2001-2002 Data File

	PSU	
Stratum	**1**	**2**
1	70	75
2	116	110
3	78	84
4	139	113
5	90	96
6	131	98
7	96	98
8	90	128
9	109	97
10	131	132
11	117	117
12	67	115
13	105	104
14	82	105
15	52	66

waist circumference(BMXWAIST)(cm), thigh circumference(BMXTHICR)(cm) and subscapular skinfold(BMXSUB)(mm). In this data file, all of the five variables are moderately or highly correlated with each other, as shown in Figure 3.2. Since BMI is defined as weight in kilograms over the squre of height in meters, it is highly correlated with the other physical measurements. If we use these variables or a subset of them as the explanatory variables, this could well create a collinearity problem in a regression. Here, to make an example, we regressed systolic blood pressure(mm Hg) on all the five variables and one demographic variable, patient's age (AGE). As shown in Figure 3.2, the correlation coefficients between patient's age and other five explanatory variables are relatively small, and all of these six explanatory variables are positively correlated with the dependent variable, systolic blood pressure. We compared the VIF values of the explanatory variables in the model by using the conventional methods and our new methods, which accounted for the sampling design (stratified clustering sampling) and survey weights (provided in the NHANES

2001-2002 data file) in Table 3.11.

Four types of regressions are compared in this section:

Type 1: using OLS, without considering the variance-covariance matrix and survey weights;

Type 2: using WLS, by including the survey weights in the estimation;

Type 3: using SWLS with estimated variance-covariance matrix $\hat{\boldsymbol{V}}$ without considering the stratified cluster sample design; here, $\hat{\boldsymbol{V}}$ is a diagonal matrix, equal to $diag(e_{hit}^2)$, where hit stands for the t^{th} unit in h^{th} stratum and i^{th} PSU, $h = 1, ..., H$, $i = 1, ..., n_h$, $t = 1, ..., m_{hi}$; both of the model-based and design-based VIF values were estimated using (3.39);

Type 4: using SWLS with estimated variance-covariance matrix $\hat{\boldsymbol{V}}$ accounting for the stratified cluster sample design; according to (3.55), $\hat{\boldsymbol{V}}$ is equal to $Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)$; the model-based VIF values were estimated by replacing $Blkdiag(\boldsymbol{e}_i\boldsymbol{e}_i^T)$ in (3.56) with $Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)$ and the design-based VIF values were estimated by (3.62). Since $n_h/(n_h - 1) = 2$, a constant, and $\boldsymbol{e}_h\boldsymbol{e}_h^T/n_h$ has a lower order of magnitude than $Blkdiag(e_{hi}e_{hi}^T)$, the model-based and design-based VIFs are approximately equal.

As in earlier sections, these alternatives account for design and population structure to varying degrees. Type 1 OLS corresponds to what analyst would do who ignores all design features. Type 2 WLS accounts for survey weights but nothing else. Type 3 SWLS accounts for survey weights and a heterogeneous error structure but ignores strata and clusters. Type 4 SWLS accounts completely for design and population structure. The VIF values of the first two types of regressions were obtained from the standard packages. Note that in WLS, the survey weights are treated as if they are inversely proportional to model variances when computing standard errors and VIFs. As observed earlier, this is generally incorrect for survey data.

As seen in Table 3.11, if we use the VIF methods from standard packages, either accounting for the survey weights (WLS) or not (OLS), BMI has the highest VIF value and will be deleted from the list of explanatory variables if the VIF cut-off value is smaller than or equal to 9. However, using our new methods which partially or completely account for unequal variances and clustering in Type 3 and Type 4, waist circumference had the highest VIF value instead of BMI. In Type 3 when we used SWLS but not considering clustering, the VIF values were relatively higher than the ones in the other types of regressions. But when we used SWLS and considering clustering in the variance-covariance matrix in Type 4, the VIF values were much smaller than the ones in the other types of regressions. When approximately accounted for, the extra noise due to clustering implies that collinearity is less of a concern than suggested by the statistics that ignore clustering. The model-based and design-based VIF values in Type 3 were estimated by the same equation, (3.39), and thus they are equal to each other. The Type 4 model-based VIFs were estimated as in (3.56) using $Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)$ as shown in (3.57). The design-based VIFs in Type 4 are computed using (3.62), . Since $n_h/(n_h-1) \doteq 1$ and $e_h e_h^T/n_h$ has a lower order of magnitude than $Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)$, the model-based and design-based VIFs are approximately equal.

In Table 3.12, the final models are listed when we used different VIF methods and VIF cut-off values to select variables by backward elimination. When we used the VIFs given from the standard packages, Type 1 (OLS) and Type 2 (WLS) selected the same final models, no matter which VIF cut-off value is used (2, 7, or 10). When VIF cut-off value was equal to 2, Type 4 selected the same final model with OLS and WLS by including patient's age, thigh circumference and subscapular skinfold in the model, while Type 3 selected a different final model with patient's age, body weight and subscapular skinfold. When VIF cut-off value was equal to 7, Type 1 and 2 added waist circumference in the final model; while, Type 4 added body weight in the final model. Because all the VIF values in Type 4 were smaller than 5 (as listed in Table 3.11), when the VIF cut-off value is 7, all the explanatory variables were kept in the final model and none of them were deleted. When the cut-off value is equal to 10, BMI and the waist circumference are deleted in Type 3, since all the other VIF values were all smaller than 10 as shown in Table 3.11.

Table 3.11: VIFs of the Five Explanatory Variables in NHANES 2001-2002 Data File

			Regression Type			
Variable	**TYPE 1**	**TYPE 2**	**TYPE 3**		**TYPE 4**	
	OLS	WLS	SWLS with $\hat{\boldsymbol{V}}$, not considering clustering		SWLS with $\hat{\boldsymbol{V}}$, considering clustering	
			Model-based	Design-based	Model-based	Design-based
age	1.30	1.29	1.18	1.18	1.28	1.28
body Weight	7.42	8.58	9.10	9.10	4.20	4.18
body mass index	**9.30**	**9.70**	10.02	10.02	4.78	4.78
waist circumference	8.02	9.14	**10.15**	**10.15**	**5.81**	**5.80**
thigh circumference	5.94	6.18	6.55	6.55	2.65	2.64
subscapular skinfold	2.43	2.37	2.28	2.28	1.53	1.53

Figure 3.2: Scatterplots and Correlation Coefficients of Five Variables in NHANES 2001-2002 data file

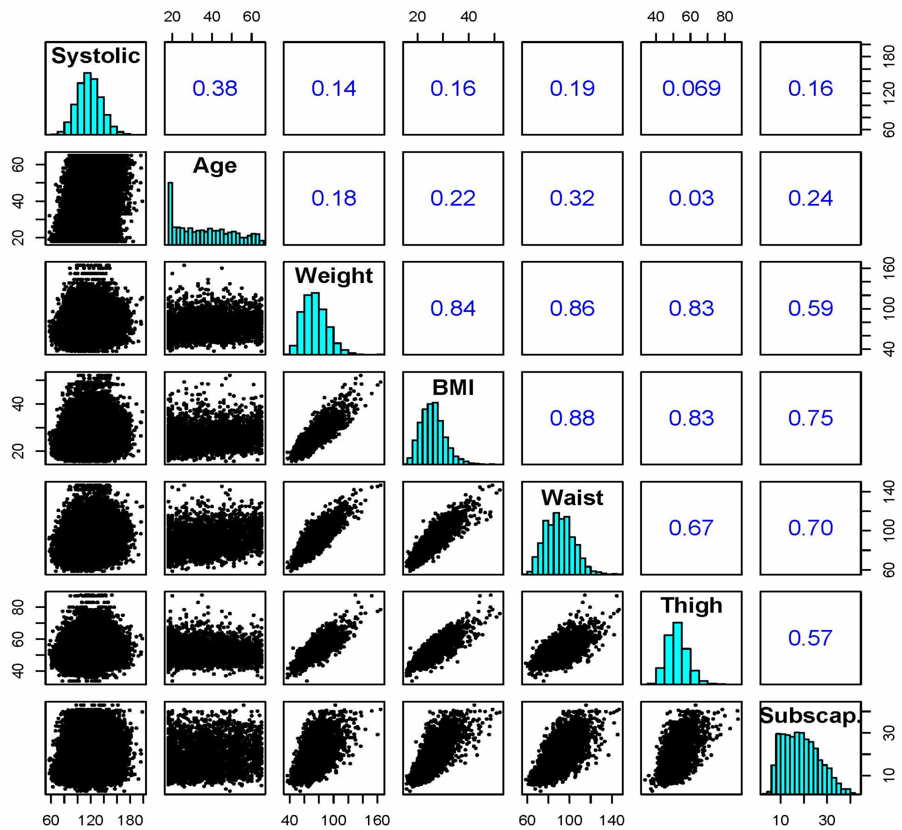

Table 3.12: The Final Model Obtained Using Different VIF Methods and VIF Cut-off Values

	Regression Type					
Variable	**TYPE 1**	**TYPE 2**	**TYPE 3**		**TYPE 4**	
	OLS	WLS	SWLS with $\hat{\boldsymbol{V}}$, not considering clustering		SWLS with $\hat{\boldsymbol{V}}$, considering clustering	
			Model-based	Design-based	Model-based	Design-based
			VIF cut-off value=2			
age	√	√	√	√	√	√
body Weight			√	√		
body mass index						
waist circumference						
thigh circumference	√	√			√	√
subscapular skinfold	√	√	√	√	√	√
			VIF cut-off value=7			
age	√	√	√	√	√	√
body Weight			√	√	√	√
body mass index					√	√
waist circumference	√	√			√	√
thigh circumference	√	√	√	√	√	√
subscapular skinfold	√	√	√	√	√	√
			VIF cut-off value=10			
age	√	√	√	√	√	√
body Weight	√	√	√	√	√	√
body mass index	√	√	√	√	√	√
waist circumference	√	√			√	√
thigh circumference	√	√	√	√	√	√
subscapular skinfold	√	√	√	√	√	√

3.2.3.3 Simulation

Population

To create the study population for our simulation, we performed several steps to create a large population to use in the simulation, starting with the NHANES sample described in Section 3.2.3.1. The general idea was to create a population with a larger number of PSUs than 30 while preserving an intracluster correlation structure similar to that found in NHANES.

First, since there are 15 strata each with 2 PSUs in the original data set, we merged three adjacent numbered strata into one and obtained 5 large strata each with 6 PSUs;

Second, in each of the five large strata, we randomly drew 100 PSUs out of the 6 PSUs with replacement;

Third, within each selected PSU hi, we randomly drew m_{hi} sample units with replacement, where m_{hi} is the number of units in PSU hi; since the number of units varies across different PSUs in the original data set, the sample size varies across different selected PSUs;

Fourth, from this sampled data set, the values of the Y variable were generated by the three explanatory variables using Gamma distributions: $Y_{hit} \sim Gamma(\alpha_{hit}, \beta_{hit})$, with shape parameters $\alpha_{hit} = \mu^2(Y_{hit})/\sigma^2$ and scale parameters $\beta_{hit} = \sigma^2/\mu^2(Y_{hit})$, where hit stands for the t^{th} unit in the i^{th} PSU within stratum h and $\sigma^2 = 250$. So that, the model is:

$$\begin{aligned} Y_{hit} &= \mu(Y_{hit}) + \varepsilon_{hit} \\ &= \beta_0 + \beta_{\text{AGE}} * \text{AGE}_{hit} + \beta_{\text{BMXTHICR}} * \text{BMXTHICR}_{hit} + \\ &\quad \beta_{\text{BMXSUB}} * \text{BMXSUB}_{hit} + \nu_{hi} + \varepsilon_{hit} \end{aligned} \tag{3.77}$$

where $\boldsymbol{\beta} = c(90, 0.47, 0.17, 0.08)$, ν_{hi} is caused by the random effect of PSUs, $E(\nu_{hi}) = 0$, $Var(\nu_{hi}) = \frac{\rho}{1-\rho} \times \sigma^2$, $\rho = 0.10$, $\sigma^2 = 250$, ε_{hit} is an error term with mean 0, $Var(\varepsilon_{hit}) = \sigma^2$, $Var(Y_{hit}) = \sigma^2/(1-\rho)$, $Cov(Y_{hit}, Y_{hit'}) = Var(\nu_{hi}) = \rho \times Var(Y_{hit})$, when $t \neq t'$, and $Cov(Y_{hit}, Y_{h'i't'}) = 0$, when $i \neq i'$. Finally, there are 49,894 units in this study population, which has 5 strata with 100 PSUs in each. The vector of the regression coefficients in this finite population is: $\boldsymbol{\beta}_U =$

$(90.08, 0.47, 0.16, 0.11)^T$.

Sample Design

A stratified clustered sampling design was conducted in the simulations. First, within each stratum in the study population, 6 PSUs were drawn out of 100 PSUs by simple random sampling with replacement; then within each sampled PSU, 5 men and 20 women were randomly selected without replacement. The selection probability for a person of gender g in PSU hi was then

$$\pi_{hit} = \frac{n_h}{N_h} \cdot \frac{m_{hi(g)}}{M_{hi(g)}}$$

where $m_{hi(g)} = 5$ for men and 20 for women and $M_{hi(g)}$ was the population count of persons of gender g in the PSU hi.

The finite population values of R^2 are shown in Table 3.13 for several models that selected in different samples in the simulations. The value of R^2 was computed as $1 - SSE/SST$, where $SSE = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$, $\boldsymbol{Y}$ and $\boldsymbol{X}$ are the finite population values of Y-vector and $\boldsymbol{X}$-matrix, and $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. The total sum of squares was computed as $SST = \boldsymbol{Y}^T\boldsymbol{Y}$. As shown in the table, these models give similar fits as measured by R^2. Thus, including or excluding the collinear variables has little effect on this measure of overall fit.

Simulation Results

Similar to the simulations we did for the SMHO dataset, 1,000 simulations were performed, in which a sample was drawn using the sampling design described above and four types of regressions (shown in Table 3.4) with their corresponding VIF

Table 3.13: R-Square for Several Selected Models in Different Samples in the Simulations

Variable	Model				
	1	2	3	4	5
Intercept	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
Age	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
BMXWT		$\surd$		$\surd$	$\surd$
BMXBMI					$\surd$
BMXWAIST			$\surd$	$\surd$	$\surd$
BMXTHICR	$\surd$		$\surd$	$\surd$	$\surd$
BMXSUB	$\surd$	$\surd$	$\surd$	$\surd$	$\surd$
R^2	0.1498	0.1491	0.1498	0.1498	0.1498

methods were applied on this sample. We also used cut-off value for the VIF equal to 2, 7, 10 separately to select variables by the backward elimination and compared the summary statistics described in Section 3.2.2.3 when different regression types and VIF cut-off values were used.

The VIFs for TYPE1(OLS) and TYPE2(WLS) are obtained from the standard statistical packages. For TYPE3, (SWLS with known $\sigma^2\boldsymbol{V}$), We computed both its model-based VIFs (as derived in Section 3.1.5) and design-based VIFs (as derived in Section 3.1.6). For TYPE4 (a SWLS with estimated $\hat{\boldsymbol{V}}$), we estimated the model-based VIFs by (3.56) and the design-based VIFs by (3.62). In TYPE4, the model-based VIFs and design-based VIFs are quite close due to their similar formulas, making their summary statistics quite close to each other as shown in the following tables.

Table 3.14 compares the average VIFs of the five-variable model from 1,000 simulations and the VIFs of the five-variable model from the full finite population. BMXBMI always haS the highest VIF value across the ones in the full finite population and the average VIFs. BMXWT and BMXWAIST also have large VIFs

Table 3.14: Comparison between the VIFs from the Full Finite Population and the Average VIFs from 1000 Simulations

	OLS		Using the true $\boldsymbol{V}$ or estimated $\hat{\boldsymbol{V}}$ or weight matrix $\boldsymbol{W}$ in the regression						
	OLS1	TYPE1 OLS	OLS2 known $\boldsymbol{V}$	OLS3 est. $\hat{\boldsymbol{V}}$	TYPE2 WLS	TYPE3 SWLS with $\boldsymbol{V}$		TYPE4 SWLS with $\hat{\boldsymbol{V}}$	
						Model-based	Design-based	Model-based	Design-based
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
	Pop.[a]	Mean[b]	**Pop.**	**Pop.**	Mean	Mean	Mean	Mean	Mean
AGE	**1.31**	1.19	**1.36**	**1.34**	1.32	1.16	1.38	1.17	1.17
BMXWT	**7.48**	8.09	**7.96**	**8.33**	7.65	7.31	7.78	8.14	8.14
BMXBMI	**9.43**	9.71	**10.94**	**12.97**	9.60	9.67	11.23	10.94	10.93
BMXWAIST	**8.13**	7.18	**9.17**	**9.38**	8.34	7.15	9.73	7.89	7.89
BMXTHICR	**6.04**	6.05	**7.00**	**7.81**	6.18	6.09	7.02	6.73	6.73
BMXSUB	**2.49**	2.52	**2.91**	**2.65**	2.50	2.56	2.95	2.56	2.56

Note: column (**1, 2, 5**): ignoring clustering; column (**3, 4, 6, 7, 8, 9**) accounting for clustering.
[a]VIFs from the full finite population
[b]The average VIFs from 1000 simulations

in all the columns. Among the three population VIFs, BMXWAIST has a higher VIF than BMXWT, while among the average VIFs, the difference between the VIFs of these two variables varied. In TYPE2 (WLS) and design-based TYPE3 (SWLS with $\boldsymbol{V}$), BMXWAIST has a higher VIF than BMXWT and it is consistent with the differences among population VIFs. But in other regression types and inference approaches, BMXWT has a higher VIF than BMXWAIST. Another noticeable difference in this table is that the population VIFs in OLS2 and OLS3 are relatively larger than the other ones in OLS1. These phenomena demonstrate that taking account of the variance structure ($\boldsymbol{V}$ or $\hat{\boldsymbol{V}}$), sampling design and survey weights in the regression can influence the impact of collinearity in $\boldsymbol{X}$ on the parameter estimation (the collinearity among the explanatory variables in $\boldsymbol{X}$ inflates the estimated standard errors of parameters to different degrees across different regression types and inference approaches). Hence, it is necessary to diagnose the effects of collinearity on the regression estimation according to given regression method and inference approach.

As discussed in Section 3.2.2.3, there are some correspondences between the columns for population values and simulation means. Here, in Table 3.14, the average VIFs in design-based TYPE3 (in column **7**) correspond to the population VIFs in OLS2 (in column **3**) and the average VIFs in TYPE2 (in column **5**) correspond to the population VIFs in OLS1 (in column **1**). These two sets of values are quite close in the table. Unlike the other correspondences we summarized for the SMHO data set in Section 3.2.2.3, the other average VIFs do not match with any population VIFs according to their VIF formulation when $\boldsymbol{V}$ is not diagonal.

Starting from the six-variable model, the backward elimination method was adopted to delete collinear variables and select a reduced final model. The percentages of samples where different final models were selected are listed in Table 3.16. When the VIF cut-off value was equal to 2, in the first three columns, including TYPE1 (OLS), TYPE2 (WLS) and model-based TYPE3 (SWLS with $\boldsymbol{V}$), the underlying model (Y=Age+BMXTHICR+BMXSUB) were selected almost in all the 1,000 samples. But in TYPE4 (SWLS with $\hat{\boldsymbol{V}}$) and design-based TYPE3, less than half of samples selected the underlying model. Some samples selected the model that includes BMXWT but eliminate BMXTHICR, and some samples selected the models with other sets of variables. When VIF cut-off value was equal to 7, the looser criterion let BMXWAIST, and sometimes BMXWT, remain in the models selected by TYPE1, TYPE2 and model-based TYPE3. In TYPE4 and design-based TYPE3, some of the samples will select other models that may have BMXBMI, or not having BMXWAIST or BMXWT or BMXTHICR in the model. When the VIF cut-off value is 10, most of samples in TYPE1, TYPE2 and model-based TYPE3 did not delete any of the variables due to the high cut-off value, while around 30 to 35.3 percent of the samples deleted BMXBMI in the model. In TYPE4 and design-based TYPE3, among 1,000 samples, some of them kept all the variables in the final model and some of them deleted BMXBMI. But there were still between 25 to 30.20 percent of samples when other models were selected.

Table 3.15 lists the relative biases in the estimated parameters for the true, underlying model. The biases are computed among samples that included a given predictor in the selected model. Here, the larger the VIF cut-off value gets, the larger the negative biases of all the coefficients were, since more collinear variables were kept in the model as shown in Table 3.16. These collinear variables are all positively correlated with the explanatory variables in the underlying model as shown in Figure 3.2, especially with BMXTHICR and BMXSUB. This is also the reason why the negative biases of the three explanatory variables in the underlying model are larger in TYPE1, TYPE2 and model-based TYPE3 than the other regression types when VIF cut-off value is equal to 10. Referring to Table 3.16, these three types kept all the six variables in the model in most of sample and all the three collinear variables can cause negative biases in the estimated parameters for the three predictors in the underlying model.

The coverage rates of the confidence intervals and the confidence regions on the true coefficient values are presented in Table 3.16 and Table 3.17. For comparison results when all variables are retained (VIF$=\infty$) are also shown. In Table 3.16, coverage rates are among samples that included a given predictor. In Table 3.17

if a predictor in the underlying model were omitted, the region was counted as not covering. Similar to the study of SMHO data set, TYPE3 (using SWLS with $\boldsymbol{V}$, model-based) always had the largest coverage rates of confidence interval and confidence region than the other types. Some of its coverage rates were even above the nominal level (95%). This is also consistent with the results in Table 3.18 (discussed more below) where the empirical standard errors were overestimated for the regressions compared to the empirical standard errors in model-based TYPE3. The only exception here is that in the design-based TYPE3, the coverage rates of the confidence interval of the intercept (83.20%) and the confidence region (37.5%) were relatively small when cutoff value is 2, because 56.90 percent of the samples in design-based TYPE3 selected the model with AGE, BMXWT and BMXSUB without BMXTHICR. The estimation of intercept can be different when different explanatory variables are in the model and the confidence region will not cover the true coefficient values when the variable is not in the model. TYPE2 (using WLS) always has the smallest coverage rates (except for VIF=2), by assuming $\boldsymbol{V} = \boldsymbol{W}^{-1}$. As we stated before, this assumption is incorrect in this study.

In Table 3.18, some of the standard errors were underestimated in the TYPE1 (OLS) and TYPE4 regressions, at least in part due to the fact that the standard variance estimator does not account for the possibility that the selected set of independent variables can differ from one sample to another (Hurvich & Tsai, 1990; Zhang, 1992). The same reason can also make the coverage rates of their confidence intervals and confidence regions smaller than the nominal level. Note that when VIF cut-off value is equal to 10, the standard errors of BMXTHICR and BMXSUB were overestimated in the TYPE1 and TYPE4, due to the fact that the presence of the collinear variables in the model inflated their standard errors to certain degrees that even counteract and exceed the reasons for underestimation.

The purpose of using VIFs is to diagnose and assess the variance inflation caused by the collinearity in $\boldsymbol{X}$. Thus, it is also interesting to compare the standard error of a given parameter estimate when it is in the reduced model obtained by using certain VIF criteria, with its standard error when it is in the six-variable full model without any variable elimination procedure. The ratios between these two standard errors are given in Table 3.19. As the VIF cut-off value get smaller and more collinear variables are excluded from the final model, the ratios get smaller for all the explanatory variables including the intercept. In other words, including extraneous, collinear variables increases the standard errors of parameter estimates for the variables that should be in the model. BMXTHICR, had the smallest ratio among all the explanatory variables no matter which regression type and VIF cut-off value is used.

Thus, the variance of BMXTHICR is inflated the most by the collinear variables. When these collinear variables are eliminated in the model, the variance of BMXTHICR gets much smaller (note that the ratio is around 50 percent when VIF cut-off value is equal to 2). So, in the six-variable full model, the other collinear variables can dramatically influence the estimation of BMXTHICR. In some samples, its effect on dependent variable, $\boldsymbol{Y}$, was to change its coefficient estimate from significant into insignificant and change the sign of the coefficient from positive to negative. These are well-known consequences of collinearity in OLS regression. Age is the only variable whose ratios are close to 100 percent. Its standard error did not get inflated too much by the existence of other extraneous variables in the model, due to its relatively low correlation with all the other variables as shown in Figure 3.2.

Table 3.15: Percent Bias in Estimated Model Parameters Using Different Regression Types and VIF Cut-off Values

Coefficient	TYPE1: OLS	TYPE2: WLS	TYPE3: SWLS with $\boldsymbol{V}$		TYPE4: SWLS with $\hat{\boldsymbol{V}}$	
			Model-based	Design-based	Model-based	Design-based
Cut-off values for VIF=2						
Intercept	0.18%	0.01%	-0.12%	3.21%	1.21%	1.20%
AGE	0.04%	-0.48%	-0.41%	-2.27%	-1.40%	-1.37%
BMXTHICR	-1.06%	2.25%	2.11%	6.09%	4.24%	4.32%
BMXSUB	-10.26%	-2.57%	-2.98%	6.63%	3.68%	3.56%
Cut-off values for VIF=7						
Intercept	0.42%	0.03%	0.15%	0.33%	0.67%	0.65%
AGE	0.28%	-0.20%	-0.21%	-0.22%	-0.28%	-0.30%
BMXTHICR	-10.27%	-5.53%	-7.35%	-10.23%	-23.86%	-24.00%
BMXSUB	-10.81%	0.63%	0.54%	-1.21%	-10.93%	-10.92%
Cut-off values for VIF=10						
Intercept	0.41%	-1.15%	-1.11%	-0.44%	-0.27%	-0.27%
AGE	-0.05%	-0.07%	-0.09%	-0.17%	-0.29%	-0.29%
BMXTHICR	-65.67%	-65.95%	-66.55%	-24.58%	-33.77%	-33.99%
BMXSUB	-68.66%	-68.27%	-70.95%	-30.01%	-37.93%	-38.06%

Note: relative bias was computed among samples that included a given variable in the model.

Table 3.16: Percentage of Samples where Models were Selected and Coverage of 95% Confidence Intervals Using Different Regression Types and VIF Cut-off Values

Final model	Coefficient	TYPE1: OLS	TYPE2: WLS	TYPE3: SWLS with $\boldsymbol{V}$		TYPE4: SWLS with $\hat{\boldsymbol{V}}$	
				Model-based	Design-based	Model-based	Design-based
Cut-off values for VIF=2							
Percentage of samples where model was selected							
Age+THICR+SUB[a]		100.00%	97.60%	99.90%	39.30%	49.50%	49.30%
Age+WT+SUB[b]		0.00%	2.40%	0.00%	56.90%	11.60%	11.80%
Other		0.00%	0.00%	0.10%	3.80%	38.90%	38.90%
Coverage of 95% confidence intervals[c]							
	Intercept	93.20%	89.10%	95.40%	83.20%	88.30%	88.40%
	Age	94.30%	87.70%	95.10%	94.70%	92.70%	92.70%
	BMXTHICR	92.70%	88.93%	95.40%	96.21%	90.32%	90.28%
	BMXSUB	91.40%	86.60%	93.59%	93.39%	90.68%	90.79%
Cut-off values for VIF=7							
Percentage of samples where model was selected							
Age+WAIST+THICR+SUB[d]		95.90%	69.50%	66.70%	23.60%	41.00%	40.80%
Age+WT+WAIST+THICR+SUB[e]		4.10%	25.00%	33.30%	23.40%	24.40%	24.40%
Other		0.00%	5.50%	0.00%	53.00%	34.60%	34.80%
Coverage of 95% confidence intervals							
	Intercept	93.90%	88.70%	96.60%	95.40%	90.90%	91.00%
	Age	94.90%	88.50%	94.90%	95.20%	92.40%	92.60%
	BMXTHICR	92.60%	89.30%	95.00%	95.33%	91.96%	92.17%
	BMXSUB	93.30%	86.10%	95.00%	94.20%	91.00%	91.00%
Cut-off values for VIF=10							
Percentage of samples where model was selected							
Age+WT+WAIST+THICR+SUB		35.30%	31.80%	32.10%	44.80%	30.30%	30.10%
All Six Variables[f]		64.20%	66.60%	67.80%	29.40%	39.70%	39.70%
Other		0.50%	1.60%	0.10%	25.80%	30.00%	30.20%
Coverage of 95% confidence intervals							
	Intercept	93.30%	89.40%	94.70%	95.60%	91.40%	91.50%
	Age	94.50%	89.50%	96.00%	96.50%	93.90%	94.00%
	BMXTHICR	92.50%	89.20%	95.20%	95.39%	92.42%	92.41%
	BMXSUB	94.50%	87.20%	95.30%	95.90%	92.00%	92.00%
Cut-off values for VIF=∞							
Percentage of samples where model was selected							
All Six Variables		100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Coverage of 95% confidence intervals							
	Intercept	92.50%	86.70%	90.60%	90.60%	92.90%	92.80%
	Age	91.20%	86.20%	90.50%	90.50%	92.00%	92.00%
	BMXTHICR	93.20%	87.60%	91.30%	91.30%	92.90%	92.90%
	BMXSUB	85.60%	79.60%	85.80%	85.80%	89.30%	89.30%

[a]The underlying model: Y=Age+BMXTHICR+BMXSUB

[b]The model: Y=Age+BMXWT+BMXTHICR

[c]Note: confidence interval coverage was computed among samples that included a given variable in the model.

[d]The model: Y=Age+BMXWAIST+BMXTHICR+BMXSUB

[e]The model: Y=Age+BMXWT+BMXWAIST+BMXTHICR+BMXSUB

[f]The model: Y=Age+BMXWT+BMXBMI+BMXWAIST+BMXTHICR+BMXSUB

Table 3.17: Coverage Rates of the Confidence Regions for the True Coefficient Values in the Artificial Population based on NHANES

Cut-off value	TYPE1: OLS	TYPE2: WLS	TYPE3: SWLS with $\boldsymbol{V}$		TYPE4: SWLS with $\hat{\boldsymbol{V}}$	
			Model-based	Design-based	Model-based	Design-based
2	77.70%	58.90%	94.70%	37.50%	57.40%	56.90%
7	77.70%	62.90%	96.00%	94.40%	75.30%	75.30%
10	79.40%	61.30%	94.20%	94.40%	77.00%	76.90%
∞	54.10%	40.7%	72.80%	72.80%	73.50%	73.70%

Note: If a variable in the underlying model was left out of the final model for a sample, the region was counted as not covering the true parameters.

Table 3.18: Ratios of the Average Estimated $se(\hat{\boldsymbol{\beta}})$ to Empirical $SE(\hat{\boldsymbol{\beta}})$ Using Different Regression Types and VIF Cut-off Values

Coefficient	TYPE1: OLS	TYPE2: WLS	TYPE3: SWLS with $\boldsymbol{V}$		TYPE4: SWLS with $\hat{\boldsymbol{V}}$	
			Model-based	Design-based	Model-based	Design-based
Cut-off values for VIF=2						
Intercept	93.37%	81.25%	101.19%	85.81%	86.93%	86.68%
AGE	101.71%	79.34%	103.38%	102.25%	97.58%	97.34%
BMXTHICR	91.56%	79.25%	98.91%	98.02%	87.51%	87.45%
BMXSUB	92.92%	75.97%	97.06%	96.95%	89.41%	89.25%
Cut-off values for VIF=7						
Intercept	95.56%	81.97%	101.68%	99.12%	89.36%	89.19%
AGE	97.06%	80.48%	99.91%	99.66%	94.61%	94.54%
BMXTHICR	92.21%	83.35%	101.13%	101.08%	98.05%	97.68%
BMXSUB	93.01%	75.92%	96.71%	96.10%	91.14%	91.12%
Cut-off values for VIF=10						
Intercept	94.88%	82.04%	102.24%	100.92%	93.25%	93.21%
AGE	97.92%	83.88%	103.94%	103.47%	99.51%	99.49%
BMXTHICR	147.63%	142.36%	185.40%	118.04%	123.80%	124.18%
BMXSUB	144.11%	126.84%	168.86%	114.88%	113.14%	113.45%

Note: ratio was computed among samples that included a given variable in the model.

Table 3.19: Ratios of the Average Estimated $se(\hat{\boldsymbol{\beta}})$ after Variable Elimination to the Average Estimated $se_6(\hat{\boldsymbol{\beta}})$ When All the Six Variables are in the Model

Coefficient	TYPE1: OLS	TYPE2: WLS	TYPE3: SWLS with $\boldsymbol{V}$		TYPE4: SWLS with $\hat{\boldsymbol{V}}$	
			Model-based	Design-based	Model-based	Design-based
Cut-off values for VIF=2						
Intercept	62.44%	61.84%	62.77%	45.40%	54.48%	54.39%
AGE	95.03%	91.15%	95.74%	95.04%	94.25%	94.22%
BMXTHICR	51.06%	49.89%	50.95%	51.02%	49.64%	49.64%
BMXSUB	80.71%	80.27%	79.69%	78.85%	79.05%	79.06%
Cut-off values for VIF=7						
Intercept	65.99%	70.06%	72.28%	73.32%	71.49%	71.55%
AGE	98.72%	98.80%	98.79%	98.66%	97.83%	97.81%
BMXTHICR	59.35%	65.83%	65.45%	74.61%	69.85%	69.88%
BMXSUB	92.33%	91.31%	90.70%	87.45%	90.33%	90.31%
Cut-off values for VIF=10						
Intercept	96.26%	95.94%	96.23%	87.01%	82.48%	82.44%
AGE	99.66%	99.70%	99.82%	99.46%	98.66%	98.65%
BMXTHICR	93.39%	93.18%	92.90%	85.43%	81.71%	81.66%
BMXSUB	97.37%	97.67%	97.38%	94.22%	94.35%	94.31%

Note: For a given variable k, $se(\hat{\beta}_k) = \sum_i se(\hat{\beta}_{ki})/S_k$, where S_k is the number of samples that included variable k in the model and $se(\hat{\beta}_{ki})$ is the estimated standard error of $\hat{\beta}_{ki}$ which was calculated at the i^{th} simulation in the model after variable elimination. While $se_6(\hat{\beta}_k) = \sum_i se_6(\hat{\beta}_{ki})/1000$, where $se_6(\hat{\beta}_{ki})$ is the estimated standard error of $\hat{\beta}_{ki}$ which was calculated at the i^{th} simulation when all the six variables are in the model.

Chapter 4

Condition Index with Variance Decomposition Method

A high VIF value indicates a sign of harmful collinearity, particularly on the inflated variances-one of the collinearity's best known side effects in the least-squares context. But VIFs are not able to diagnose the number of near dependencies that are present in the data, which limit its effectiveness on developing remedies for collinearity. (As we discussed in the experimental studies in the previous chapter, using VIFs, we can delete the variables with the highest VIFs using backward selection method, but can not identify the near dependencies among certain predictors.) However, the method described in this chapter that is based on the eigensystem of the survey weighted data matrix, is able to be used to form a set of condition indexes that allow us to determine the strength and number of near dependencies, and the newly-developed variance decomposition proportions can allow us to determine variable involvement in the regressions using survey-weighted leat squares. It provides more information on the influence of collinearity and will assist us to develop some other efficient remedies for collinearity issues.

4.1 Adaptation in Survey-Weighted Least Squares

4.1.1 Adaptation under the Model-Based Inference when $\boldsymbol{V}$ is known

In survey-weighted least squares (SWLS), we are more interested in the collinear relations among the columns in the matrix $\tilde{\boldsymbol{X}} = \boldsymbol{W}^{1/2}\boldsymbol{X}$ instead of $\boldsymbol{X}$ in OLS, since the design matrix $\boldsymbol{X}$ is weighted by $\boldsymbol{W}^{1/2}$. The singular value decomposition of $\tilde{\boldsymbol{X}}$ is $\tilde{\boldsymbol{X}} = \boldsymbol{U_1}\boldsymbol{D}\boldsymbol{U_2}^T$, where $\boldsymbol{U_1}$, $\boldsymbol{U_2}$ and $\boldsymbol{D}$ are usually different from the ones of $\boldsymbol{X}$, due to the unequal survey weights.

The condition number of $\tilde{\boldsymbol{X}}$ is defined as $\kappa(\tilde{\boldsymbol{X}}) = \mu_{max}/\mu_{min}$, where μ_{max} and μ_{min} are maximum and minimum singular values of $\tilde{\boldsymbol{X}}$. The condition number of $\tilde{\boldsymbol{X}}$ is usually different from the condition number of the data matrix $\boldsymbol{X}$ due to the unequal survey weights. Condition indexes are defined as

$$\eta_k = \mu_{max}/\mu_k, \quad k = 1, ..., p \tag{4.1}$$

where μ_k is one of the singular values of $\tilde{\boldsymbol{X}}$. The scaled condition indexes and condition numbers are the condition indexes and condition numbers of the scaled $\tilde{\boldsymbol{X}}$.

Based on the extrema of the ratio of quadratic forms (Lin, 1984), the condition number $\kappa(\tilde{\boldsymbol{X}})$ is bounded in the range of:

$$\frac{w_{min}^{1/2}}{w_{max}^{1/2}}\kappa(\boldsymbol{X}) \leq \kappa(\tilde{\boldsymbol{X}}) \leq \frac{w_{max}^{1/2}}{w_{min}^{1/2}}\kappa(\boldsymbol{X}), \tag{4.2}$$

where w_{min} and w_{max} are the minimum and maximum survey weights.

This expression indicates that if the survey weights do not vary too much, the condition number in SWLS resembles the one in OLS. While if it is an unequal-weighted sampling design with a wide range of survey weights, the condition number can be very different between SWLS and OLS. When SWLS has a large condition number, OLS might not. Recall that in section 3.1.2, we showed that in the case of exact linear dependence among the columns of $\boldsymbol{X}$, the columns of $\tilde{\boldsymbol{X}}$ will also be linearly dependent. In this extreme case at least one eigenvalue of $\boldsymbol{X}$ will be zero, and both $\kappa(\boldsymbol{X})$ and $\kappa(\tilde{\boldsymbol{X}})$ will be infinite.

Large values of κ or of the η_k's of, say, 10 to 30, are usually interpreted in OLS as signals that two or more columns of $\boldsymbol{X}$ have moderate to strong dependencies. How well the OLS rules-of-thumb apply to survey data needs to be studied.

We showed in Chapter 3 that the model variance of the parameter estimator under a model with $Var_M(\boldsymbol{Y}) = \boldsymbol{V}$ is:

$$Var_M(\hat{\boldsymbol{\beta}}_{SW}) = \sigma^2(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{V}}\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}. \tag{4.3}$$

Using the SVD of $\tilde{\boldsymbol{X}}$, we can rewrite $Var_M(\hat{\boldsymbol{\beta}}_{SW})$ as

$$Var_M(\hat{\boldsymbol{\beta}}_{SW}) = \sigma^2\boldsymbol{U_2}\boldsymbol{D}^{-1}\boldsymbol{U_1}^T\tilde{\boldsymbol{V}}\boldsymbol{U_1}\boldsymbol{D}^{-1}\boldsymbol{U_2}^T. \tag{4.4}$$

Define the $p \times p$ matrix $\boldsymbol{Z} = \boldsymbol{U_1}^T \tilde{\boldsymbol{V}} \boldsymbol{U_1} = (z_{ij})_{p\times p}$, then

$$z_{ij} = \sum_{a=1}^{n} u_{1aj} \sum_{b=1}^{n} u_{1bi} \tilde{v}_{ba} \tag{4.5}$$

with $\boldsymbol{U_1} = (u_{1ij}), i, j = 1, \ldots, p$ and $\tilde{\boldsymbol{V}} = (\tilde{v}_{ab}), a, b = 1, \ldots, n$.

The k^{th} diagonal element of $Var_M(\hat{\boldsymbol{\beta}}_{SW})$, which is the variance of $\hat{\beta}_{SWk}$, is

$$Var_M(\hat{\beta}_{SWk}) = \sigma^2 \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{u_{2ki}}{\mu_i} \frac{u_{2kj}}{\mu_j} z_{ij} = \sigma^2 \sum_{i=1}^{p} \frac{u_{2ki} q_{ki}}{\mu_i}, \tag{4.6}$$

where u_{2ki} is an element of $\boldsymbol{U_2}_{p\times p}$, $\boldsymbol{Q} = (q_{ki})_{p\times p} = \left(\boldsymbol{U_1}^T \tilde{\boldsymbol{V}} \boldsymbol{U_1} \boldsymbol{D}^{-1} \boldsymbol{U_2}^T\right)^T = \left(\boldsymbol{Z}\boldsymbol{D}^{-1}\boldsymbol{U_2}^T\right)^T$, and $q_{ki} = \sum_{j=1}^{p} u_{2kj} z_{ij}/\mu_j$. The proportional contribution of i^{th} singular value to the variance of the k^{th} regression parameter is then

$$\pi_{ik} = \frac{u_{2ki} q_{ki}}{\mu_i} / Var_M(\hat{\beta}_{SWk}). \tag{4.7}$$

Note that the terms $\frac{u_{2ki}q_{ki}}{\mu_i}$ are the elements of the matrix $\boldsymbol{Q}^* = (\boldsymbol{U}_2 \boldsymbol{D}^{-1}) \cdot \boldsymbol{Q}$. The proportion π_{ik} is found by dividing the ki^{th} elements of $\boldsymbol{Q}^*$ by its row sums. Denote $\boldsymbol{\Pi} = (\pi_{jk})_{p\times p} = \boldsymbol{Q}^{*T} \bar{\boldsymbol{Q}}^{*-1}$, where $\bar{\boldsymbol{Q}}^*$ is the diagonal matrix with the row sums of $\boldsymbol{Q}^*$ on the main diagonal and 0 elsewhere.

Analogous to the suggestion in Belsley et al. (1980) for OLS regression, a variance decomposition table can be formed:

Condition	Proportions of variance			
Index	$Var_M(\hat{\beta}_{SW1})$	$Var_M(\hat{\beta}_{SW2})$	$\cdots$	$Var_M(\hat{\beta}_{SWp})$
μ_1	π_{11}	π_{12}	$\cdots$	π_{1p}
μ_2	π_{21}	π_{22}	$\cdots$	π_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
μ_p	π_{p1}	π_{p2}	$\cdots$	π_{pp}

When two or more independent variables are collinear (or "nearly dependent"), one singular value should make a large contribution to the variance of the parameter estimates associated with those variables.

If $\boldsymbol{V} = \boldsymbol{W}^{-1}$ as assumed in WLS, $\tilde{\boldsymbol{V}} = \boldsymbol{I}$, (4.5) can then be rewritten as: $z_{ij} = \sum_{t=1}^{n} u_{1tj}u_{1ti}$ and since $\boldsymbol{U_1}$ is column orthogonal, $z_{ij} = 1$ when $i = j$, while $z_{ij} = 0$ when $i \neq j$. The variance decomposition in (4.6) has the same form as (2.4) in OLS, (note that $\boldsymbol{U_2}$ is still different from the one in OLS, which is one component of the SVD of $\tilde{\boldsymbol{X}}$ instead of $\boldsymbol{X}$). Another special case here is when $\boldsymbol{V} = \boldsymbol{I}$ and the survey weights are equal. However, when the survey weights are unequal, even when $\boldsymbol{V} = \boldsymbol{I}$, the variance decomposition in (4.6) is different from (2.4) in OLS since $\tilde{\boldsymbol{V}} \neq \boldsymbol{I}$. In the next section, we will consider some special models that take the population features such as clusters and strata into account in this variance decomposition.

4.1.2 Decomposition of the Estimated Variance in a Sample Selected from the Finite Population

Estimating the variance of $\hat{\boldsymbol{\beta}}_{SW}$ in (4.3) can be accomplished by substituting an estimator of $\tilde{\boldsymbol{V}} = \boldsymbol{W}^{1/2}\boldsymbol{V}\boldsymbol{W}^{1/2}$. In this section, we present estimators that are appropriate for different models and sample designs. The general form of the model variance estimators is:

$$var_M(\hat{\boldsymbol{\beta}}_{SW}) = \sigma^2 \boldsymbol{U_2}\boldsymbol{D}^{-1}\hat{\boldsymbol{Z}}\boldsymbol{D}^{-1}\boldsymbol{U_2}^T. \tag{4.8}$$

Since $\tilde{\boldsymbol{X}} = \boldsymbol{W}^{1/2}\boldsymbol{X} = \boldsymbol{U_1}\boldsymbol{D}^{-1}\boldsymbol{U_2}^T$, the matrices $\boldsymbol{U}_2$ and $\boldsymbol{D}$ do not have to be estimated but are determined from the form of the model and matrix of survey weights.

4.1.2.1 Variance Decomposition for A Model with Independent Errors

As discussed in the preceding chapter, in a model with independent errors, the estimated model variance of $\hat{\boldsymbol{\beta}}_{SW}$ is:

$$\begin{aligned} var_M(\hat{\boldsymbol{\beta}}_{SW}) &= \boldsymbol{A}^{-1}\left(\sum_{i=1}^{n} \dot{\boldsymbol{x}}_i w_i e_i^2 w_i \dot{\boldsymbol{x}}_i^T\right)\boldsymbol{A}^{-1} \\ &= \boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{W} diag(e_i^2)\boldsymbol{W}\boldsymbol{X}\boldsymbol{A}^{-1} \\ &= \boldsymbol{A}^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{W}^{1/2} diag(e_i^2)\boldsymbol{W}^{1/2}\tilde{\boldsymbol{X}}\boldsymbol{A}^{-1}, \end{aligned} \tag{4.9}$$

where $e_i = Y_i - \dot{\boldsymbol{x}}_i\hat{\boldsymbol{\beta}}_{SW}$ as defined in Chapter 3. In (4.9) the estimated variance-covariance matrix $\hat{\hat{\boldsymbol{V}}}$ is $\boldsymbol{W}^{1/2}diag(e_i^2)\boldsymbol{W}^{1/2}$. Hence, $\hat{\hat{v}}_{ij} = w_i e_i^2$ when $i = j$, while $\hat{\hat{v}}_{ij} = 0$ when $i \neq j$. Applying these results to (4.5), we can estimate matrix $\boldsymbol{Z}$ by $\hat{\boldsymbol{Z}} = \boldsymbol{U_1}^T\boldsymbol{W}^{1/2}diag(e_i^2)\boldsymbol{W}^{1/2}\boldsymbol{U_1}$ and the estimated $\hat{z}_{ij}$ is:

$$\hat{z}_{ij} = \sum_{a=1}^{n} u_{1aj}u_{1ai}w_a e_a^2. \tag{4.10}$$

Substituting (4.10) in (4.6), we can decompose the estimated model variance of $\hat{\beta}_{SWk}$:

$$var_M(\hat{\beta}_{SWk}) = \sum_{i=1}^{p}\sum_{j=1}^{p}\frac{u_{2ki}}{\mu_i}\frac{u_{2kj}}{\mu_j}\hat{z}_{ij} = \sum_{i=1}^{p}\frac{u_{2ki}\hat{q}_{ki}}{\mu_i} \tag{4.11}$$

with $\hat{\boldsymbol{Q}} = (\hat{q}_{ki})_{p\times p} = \left(\boldsymbol{U_1}^T\hat{\hat{\boldsymbol{V}}}\boldsymbol{U_1}\boldsymbol{D}^{-1}\boldsymbol{U_2}^T\right)^T = \left(\hat{\boldsymbol{Z}}\boldsymbol{D}^{-1}\boldsymbol{U_2}^T\right)^T$ and $\hat{q}_{ki} = \sum_{j=1}^{p} u_{2kj}\hat{z}_{ij}/\mu_j$. Let $\hat{\boldsymbol{Q}}^* = (\boldsymbol{U}_2\boldsymbol{D}^{-1}) \cdot \hat{\boldsymbol{Q}}$. The proportion of the estimated variance of the k^{th} regression coefficient associated with the i^{th} component of its decomposition is:

$$\hat{\pi}_{ik} = \frac{u_{2ki}\hat{q}_{ki}}{\mu_i}/var_M(\hat{\beta}_{SWk}), \tag{4.12}$$

and $\hat{\boldsymbol{\Pi}} = (\hat{\pi}_{jk})_{p\times p} = \hat{\boldsymbol{Q}}^{*T}\hat{\hat{\boldsymbol{Q}}}^{*-1}$, where $\hat{\hat{\boldsymbol{Q}}}^*$ is the diagonal matrix with the row sums of $\hat{\boldsymbol{Q}}^*$ on the main diagonal and 0 elsewhere.

Using the design-based linearization variance estimator in a single-stage with-replacement design, a design consistent variance estimator is:

$$
\begin{aligned}
var_L(\hat{\boldsymbol{\beta}}_{SW}) &= \frac{n}{n-1}\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{W}diag(e_i^2)\boldsymbol{W}\boldsymbol{X}\boldsymbol{A}^{-1} \\
&= \frac{n}{n-1}\boldsymbol{A}^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{W}^{1/2}diag(e_i^2)\boldsymbol{W}^{1/2}\tilde{\boldsymbol{X}}\boldsymbol{A}^{-1} \\
&= \frac{n}{n-1}var_M(\hat{\beta}_{SWk}).
\end{aligned} \tag{4.13}
$$

Its decomposition is the same as (4.11) for the estimated model variance of $\hat{\beta}_{SWk}$ multiplied by the factor $(n/n-1)$:

$$
var_L(\hat{\beta}_{SWk}) = \frac{n}{n-1}\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{u_{2ki}}{\mu_i}\frac{u_{2kj}}{\mu_j}\hat{z}_{ij} = \frac{n}{n-1}\sum_{i=1}^{p}\frac{u_{2ki}\hat{q}_{ki}}{\mu_i}. \tag{4.14}
$$

The variance-decomposition proportions of $var_L(\hat{\beta}_{SWk})$ are also the same as the ones for $var_M(\hat{\beta}_{SWk})$:

$$
\hat{\pi}_{ik} = \frac{n}{n-1}\frac{u_{2ki}\hat{q}_{ki}}{\mu_i}/var_L(\hat{\beta}_{SWk}) = \frac{u_{2ki}\hat{q}_{ki}}{\mu_i}/var_M(\hat{\beta}_{SWk}). \tag{4.15}
$$

4.1.2.2 Variance Decomposition for A Model with Clustering

For a multi-stage sampling design, suppose we have a two-stage sampling design. For model-based inference, assume that model 3.43 holds, i.e. units in different clusters are uncorrelated. For design-based inference, assume the first-stage sample units are selected with replacement. In the clustered sample, suppose n clusters are selected out of N clusters and m_l units are selected out of M_l units in the selected cluster l. The total number of sample units is $m = \sum_{l\in s} m_l$, where s is the set of sample clusters. The model-based variance estimator for $\hat{\boldsymbol{\beta}}_{SW}$ has a similar sandwich form to the one in the single-stage model,

$$
\begin{aligned}
var_M(\hat{\boldsymbol{\beta}}_{SW}) &= \boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{W}Blkdiag(\boldsymbol{e}_l\boldsymbol{e}_l^T)\boldsymbol{W}\boldsymbol{X}\boldsymbol{A}^{-1} \\
&= \boldsymbol{A}^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{W}^{1/2}Blkdiag(\boldsymbol{e}_l\boldsymbol{e}_l^T)\boldsymbol{W}^{1/2}\tilde{\boldsymbol{X}}\boldsymbol{A}^{-1},
\end{aligned}
\tag{4.16}
$$

where $\boldsymbol{e}_l = \boldsymbol{Y}_l - \boldsymbol{X}_l\hat{\boldsymbol{\beta}}_{SW}$ as defined in Chapter 3 and $Blkdiag(\boldsymbol{e}_l\boldsymbol{e}_l^T)$ is an $m \times m$ block diagonal matrix with $\boldsymbol{e}_l\boldsymbol{e}_l^T$ on the main diagonal position and 0 elsewhere. Here, the estimated covariance-variance matrix is $\hat{\tilde{\boldsymbol{V}}} = \boldsymbol{W}^{1/2}Blkdiag(\boldsymbol{e}_l\boldsymbol{e}_l^T)\boldsymbol{W}^{1/2}$ and thus the estimated matrix $\boldsymbol{Z}$ is $\boldsymbol{U_1}^T\boldsymbol{W}^{1/2}Blkdiag(\boldsymbol{e}_l\boldsymbol{e}_l^T)\boldsymbol{W}^{1/2}\boldsymbol{U_1}$. When unit i and j are in different clusters, $\hat{\tilde{v}}_{ij}$ is equal to 0; while when unit i and j are in the same cluster, $\hat{\tilde{v}}_{ij} = w_i^{1/2}w_j^{1/2}e_ie_j$. Applying these results to (4.5), the estimated $\hat{z}_{ij}$ is:

$$
\begin{aligned}
\hat{z}_{ij} &= \sum_{l\in s}\left(\sum_{a\in s_l}u_{1aj}\sum_{b\in s_l}u_{1bi}w_b^{1/2}w_a^{1/2}e_be_a\right) \\
&= \sum_{l\in s}\left(\sum_{a\in s_l}u_{1aj}u_{1ai}w_ae_a^2 + \sum_{a\in s_l}u_{1aj}\sum_{b\in s_l, b\neq a}u_{1bi}w_b^{1/2}w_a^{1/2}e_be_a\right),
\end{aligned}
\tag{4.17}
$$

where s_l is the set of all the sample units in the selected cluster l. The matrix of the $\hat{z}_{ij}$ can also be written as $\hat{\boldsymbol{Z}} = \boldsymbol{U_1}^T\hat{\tilde{\boldsymbol{V}}}\boldsymbol{U_1}$ where, as defined in Section 2.1.2.2, $\boldsymbol{U_1}$ is the part of the singular value decomposition of $\tilde{\boldsymbol{X}}$.

The decomposition of the estimated model variance of $\hat{\beta}_{SWk}$ is given by (4.9) and (4.11) with $\hat{z}_{ij}$ defined by (4.17).

Under the design-based inference, the linearization variance estimator for $\hat{\boldsymbol{\beta}}_{SW}$

is,

$$
\begin{aligned}
var_L(\hat{\boldsymbol{\beta}}_{SW}) &= \frac{n}{n-1}\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{W}Blkdiag(\boldsymbol{e}_l\boldsymbol{e}_l^T)\boldsymbol{W}\boldsymbol{X}\boldsymbol{A}^{-1} \\
&= \frac{n}{n-1}\boldsymbol{A}^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{W}^{1/2}Blkdiag(\boldsymbol{e}_l\boldsymbol{e}_l^T)\boldsymbol{W}^{1/2}\tilde{\boldsymbol{X}}\boldsymbol{A}^{-1} \qquad (4.18) \\
&= \frac{n}{n-1}var_M(\hat{\beta}_{SWk}).
\end{aligned}
$$

Similar to the decompositions for single-stage sampling design, its decomposition can be expressed as (4.14) with the estimated $\hat{z}_{ij}$ in (4.17). The variance-decomposition proportions are also the same as the ones in $var_M(\hat{\beta}_{SWk})$ for a multi-stage sampling design.

4.1.2.3 Variance Decomposition for A Model with Stratified Clustering

In a stratified multistage sampling design, suppose that there are $h = 1, ..., H$ strata in the population, we select $l = 1, ..., n_h$ clusters in stratum h and $t = 1, ..., m_{hl}$ units in cluster hl. The total number of units in the sample is $m = \sum_h \sum_{l \in s_h} m_{hl}$ where s_h is the set of sample clusters in stratum h and the total number of sample units in stratum h is $m_h = \sum_{l \in s_h} m_{hl}$. Clusters are assumed to be selected with replacement within strata and independently between strata. We will consider two linear models: one assumes that there are common intercept and slopes across strata; while another assumes that there are different linear models, or different parameters in each stratum.

Model 1: Common intercept and slopes across strata:

The general formulas for the model variance in (4.3) and (4.4) still apply. In this case, $\boldsymbol{U}_1$ in $\boldsymbol{Z} = \boldsymbol{U_1}^T\tilde{\boldsymbol{V}}\boldsymbol{U_1}$ is $m \times p$ since m is the total number of sample units across all strata and clusters. The model-based sandwich variance estimator is:

$$\begin{aligned} var_M(\hat{\boldsymbol{\beta}}_{SW}) &= \boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{W}Blkdiag(\boldsymbol{e}_{hl}\boldsymbol{e}_{hl}^T)\boldsymbol{W}\boldsymbol{X}\boldsymbol{A}^{-1} \\ &= \boldsymbol{A}^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{W}^{1/2}Blkdiag(\boldsymbol{e}_{hl}\boldsymbol{e}_{hl}^T)\boldsymbol{W}^{1/2}\tilde{\boldsymbol{X}}\boldsymbol{A}^{-1} \\ &= \boldsymbol{A}^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{W}^{1/2}\boldsymbol{P}\boldsymbol{W}^{1/2}\tilde{\boldsymbol{X}}\boldsymbol{A}^{-1}, \end{aligned} \tag{4.19}$$

where $\boldsymbol{e}_{hl} = \boldsymbol{Y}_{hl} - \boldsymbol{X}_{hl}\hat{\boldsymbol{\beta}}_{SW}$ is the m_{hl} vector of residuals for cluster hl and the $m \times m$ matrix $\boldsymbol{P} = Blkdiag(\boldsymbol{e}_{hl}\boldsymbol{e}_{hl}^T)$ has $\boldsymbol{e}_{hl}\boldsymbol{e}_{hl}^T$ on the main diagonal position and 0 elsewhere. The estimated covariance-variance matrix is $\hat{\tilde{\boldsymbol{V}}} = \boldsymbol{W}^{1/2}\boldsymbol{P}\boldsymbol{W}^{1/2}$. When unit i and j are in different clusters, $\hat{\tilde{v}}_{ij}$ is equal to 0; when unit i and j are in the same cluster, $\hat{\tilde{v}}_{ij} = w_i^{1/2}w_j^{1/2}e_ie_j$. Substituting these results into (4.5), the matrix $\hat{\boldsymbol{Z}}$ is estimated by:

$$\hat{\boldsymbol{Z}} = \boldsymbol{U_1}^T\boldsymbol{W}^{1/2}\boldsymbol{P}\boldsymbol{W}^{1/2}\boldsymbol{U_1} \tag{4.20}$$

and its element $\hat{z}_{ij}$ is:

$$\hat{z}_{ij} = \sum_{h=1}^{H}\left[\sum_{l\in s_h}\left(\sum_{a\in s_{hl}} u_{1aj}u_{1ai}w_a e_a^2 + \sum_{a\in s_{hl}} u_{1aj}\sum_{b\in s_{hl}, b\neq a} u_{1bi}w_b^{1/2}w_a^{1/2}e_be_a\right)\right], \tag{4.21}$$

where s_h in the set of all the selected clusters in stratum H and s_{hl} is the set of all the sample units in the selected cluster hl.

Substituting in (4.6), we can also decompose the estimated model variance of $\hat{\beta}_{SWk}$ in (4.19) in a similar way as (4.11) for the single-stage sampling design.

Under the design-based inference, the linearization variance estimator for $\hat{\boldsymbol{\beta}}_{SW}$ was given in (3.61) of Chapter 3 and is repeated here for convenience,

$$
\begin{aligned}
var_L(\hat{\boldsymbol{\beta}}_{SW}) &= \sum_{h=1}^{H} \boldsymbol{A}^{-1} \left\{ \frac{n_h}{n_h - 1} \boldsymbol{X}_h^T \boldsymbol{W}_h \left[Blkdiag(\boldsymbol{e}_{hl} \boldsymbol{e}_{hl}^T) - \frac{1}{n_h} \boldsymbol{e}_h \boldsymbol{e}_h^T \right] \boldsymbol{W}_h \boldsymbol{X}_h \right\} \boldsymbol{A}^{-1} \\
&= \sum_{h=1}^{H} \boldsymbol{A}^{-1} \tilde{\boldsymbol{X}}_h^T \boldsymbol{W}_h^{1/2} \left\{ \frac{n_h}{n_h - 1} \left[\boldsymbol{P} - \frac{1}{n_h} \boldsymbol{e}_h \boldsymbol{e}_h^T \right] \right\} \boldsymbol{W}_h^{1/2} \tilde{\boldsymbol{X}}_h \boldsymbol{A}^{-1}.
\end{aligned}
\tag{4.22}
$$

As observed in Section 3.1.7.3, $E_M \left[var_L(\hat{\boldsymbol{\beta}}_{SW}) \right] \doteq Var_M(\hat{\boldsymbol{\beta}}_{SW})$, and both (4.19) and (4.22) are approximately unbiased for the model variance.

Unlike in the single-stage sampling or multi-stage sampling design, the estimator of $\tilde{\boldsymbol{V}}$ is different from the one above used in (4.19):

$$
\tilde{\boldsymbol{V}} = Blkdiag \left[\frac{n_h}{n_h - 1} \boldsymbol{W}_h^{1/2} \left(\boldsymbol{P} - \frac{1}{n_h} \boldsymbol{e}_h \boldsymbol{e}_h^T \right) \boldsymbol{W}_h^{1/2} \right]. \tag{4.23}
$$

When unit i and j are in different strata, $\hat{\tilde{v}}_{ij} = 0$; when unit i and j are in the same strata but different clusters, $\hat{\tilde{v}}_{ij} = -\frac{1}{n_h - 1} w_i^{1/2} w_j^{1/2} e_i e_j$; when unit i and j are in the same strata and same cluster, $\hat{\tilde{v}}_{ij} = w_i^{1/2} w_j^{1/2} e_i e_j$. The matrix $\hat{\boldsymbol{Z}}$ is estimated by:

$$
\hat{\boldsymbol{Z}} = \boldsymbol{U_1}^T \tilde{\boldsymbol{V}} \boldsymbol{U_1} = \boldsymbol{U_1}^T Blkdiag \left[\frac{n_h}{n_h - 1} \boldsymbol{W}_h^{1/2} \left(\boldsymbol{P} - \frac{1}{n_h} \boldsymbol{e}_h \boldsymbol{e}_h^T \right) \boldsymbol{W}_h^{1/2} \right] \boldsymbol{U_1}, \tag{4.24}
$$

and its element $\hat{z}_{ij}$ is:

$$\begin{aligned}\hat{z}_{ij} =& \sum_{h=1}^{H}\left[\sum_{l\in s_h}\left(\sum_{a\in s_{hl}} u_{1aj}u_{1ai}w_a e_a^2 + \sum_{a\in s_{hl}} u_{1aj} \sum_{b\in s_{hl},b\neq a} u_{1bi}w_b^{1/2}w_a^{1/2}e_b e_a\right)\right] \\ & - \sum_{h=1}^{H}\frac{1}{n_h}\left[\sum_{l,l'\in s_h,l\neq l'}\left(\sum_{a\in s_{hl}} u_{1aj}\sum_{b\in s_{hl'}} u_{1bi}w_b^{1/2}w_a^{1/2}e_b e_a\right)\right].\end{aligned} \tag{4.25}$$

Analogous to the variance decomposition for the model variance, we substitute (4.25) into (4.6) and decompose the estimated design-based variance of $\hat{\beta}_{SWk}$ in (4.22) as:

$$var_L(\hat{\beta}_{SWk}) = \sum_{i=1}^{p}\sum_{j=1}^{p}\frac{u_{2ki}}{\mu_i}\frac{u_{2kj}}{\mu_j}\hat{z}_{ij} = \sum_{i=1}^{p}\frac{u_{2ki}\hat{q}_{ki}}{\mu_i} \tag{4.26}$$

with $\hat{z}_{ij}$ estimated in (4.25). The variance-decomposition proportions are:

$$\hat{\pi}_{ik} = \frac{u_{2ki}\hat{q}_{ki}}{\mu_i}/var_L(\hat{\beta}_{SWk}), \tag{4.27}$$

where

$$\begin{aligned}\hat{\boldsymbol{Q}} = (\hat{q}_{ki})_{p\times p} &= \boldsymbol{U_1}^T Blkdiag\left[\frac{n_h}{n_h-1}\boldsymbol{W}_h^{1/2}\left(\boldsymbol{P}-\frac{1}{n_h}\boldsymbol{e}_h\boldsymbol{e}_h^T\right)\boldsymbol{W}_h^{1/2}\right]\boldsymbol{U_1}\boldsymbol{D}^{-1}\boldsymbol{U_2}^T \\ &= \hat{\boldsymbol{Z}}\boldsymbol{D}^{-1}\boldsymbol{U_2}^T,\end{aligned} \tag{4.28}$$

The elements of $\hat{\boldsymbol{Q}}$ use $\hat{z}_{ij}$ in (4.25) instead of the ones in (4.21) for $var_M(\hat{\boldsymbol{\beta}})$.

When the sampling design is a stratified single stage sampling design, suppose that there are $h = 1, ..., H$ strata in the population and $l = 1, ..., N_h$ units in the corresponding stratum h. We select $l = 1, ..., n_h$ units with replacement in stratum h

and independently between strata. This is a special case of the stratified multistage sampling design, in which each cluster just has one single unit in it, $m_{hl} = 1$. The matrix $\boldsymbol{P}$ in (4.20) and (4.24) can be simplified as the diagonal matrix $diag(e_{hl}^2)$ with the residual terms $e_{hl} = Y_{hl} - \dot{\boldsymbol{x}}_{hl}\hat{\boldsymbol{\beta}}_{SW}$ and the vector of covariates for hl^{th} unit is $\dot{\boldsymbol{x}}_{hl} = (x_{1hl}, ..., x_{phl})^T$. Substituting in (4.6) correspondingly, we can also decompose both of the estimated model and design-based variances of $\hat{\beta}_{SWk}$ in a similar way as (4.11) for the single-stage sampling design.

Model 2: Different linear models, or different parameters in each stratum

The model for the mean of unit i in cluster l and stratum h is $E_M(Y_{hli}) = \dot{\boldsymbol{x}}_{hli}\boldsymbol{\beta}_h$, where $\dot{\boldsymbol{x}}_{hli}$ is the $p \times 1$ vector of predictor variables for unit (hlk) and $\boldsymbol{\beta}_h$ is a $p \times 1$ parameter vector specific to stratum h. Within each stratum, the estimation of regression parameters and their variances is for a unstratified multi-stage sampling design and thus its variance decomposition is the same as we discussed in the previous section for the model in a multi-stage sampling design. In this case, there are Hp slope parameters in the model, where H is the total number of strata. We will do a separate decomposition within each stratum.

In each stratum, the model variance of $\hat{\boldsymbol{\beta}}_{SWh}$ is estimated by:

$$var_M(\hat{\boldsymbol{\beta}}_{SWh}) = \boldsymbol{A}_h^{-1}\boldsymbol{X}_h^T\boldsymbol{W}_h\boldsymbol{P}_h\boldsymbol{W}_h\boldsymbol{X}_h\boldsymbol{A}_h^{-1}, \tag{4.29}$$

where $\boldsymbol{A}_h = \boldsymbol{X}_h^T\boldsymbol{W}_h\boldsymbol{X}_h$, $\boldsymbol{e}_{hl} = \boldsymbol{Y}_{hi} - \boldsymbol{X}_{hi}\hat{\boldsymbol{\beta}}_{SWh}$ and the $m_h \times m_h$ matrix $\boldsymbol{P}_h$ is the block diagonal matrix with $\boldsymbol{e}_{hl}\boldsymbol{e}_{hl}^T$ on the main diagonal position and 0 elsewhere.

The design-based linearization variance estimator is:

$$\begin{aligned} var_L(\hat{\boldsymbol{\beta}}_{SWh}) &= \frac{n_h}{n_h-1}\boldsymbol{A}_h^{-1}\boldsymbol{X}_h^T\boldsymbol{W}_h\boldsymbol{P}_h\boldsymbol{W}_h\boldsymbol{X}_h\boldsymbol{A}_h^{-1} \\ &= \frac{n_h}{n_h-1}\boldsymbol{A}_h^{-1}\tilde{\boldsymbol{X}}_h^T\boldsymbol{W}_h^{1/2}\boldsymbol{P}_h\boldsymbol{W}_h^{1/2}\tilde{\boldsymbol{X}}_h\boldsymbol{A}_h^{-1} \\ &= \frac{n_h}{n_h-1}var_M(\hat{\beta}_{SWh}). \end{aligned} \tag{4.30}$$

Let the SVD of $\boldsymbol{X}_h$ to be:

$$\boldsymbol{X}_h = \boldsymbol{U}_{1h}\boldsymbol{D}_h\boldsymbol{U}_{h2}^T$$

where $\boldsymbol{D}_h = diag(\mu_{h1}, \ldots, \mu_{hp})$ is the diagonal matrix of singular values of $\boldsymbol{X}_h$, $\boldsymbol{U}_{\boldsymbol{h1}} = (u_{1hkj})$ and $\boldsymbol{U}_{\boldsymbol{h2}} = (u_{2hkj})$.

Within each stratum, the variance decomposition of $var_M(\hat{\boldsymbol{\beta}}_{SWh})$ in (4.29) is:

$$var_M(\hat{\beta}_{SWhk}) = \sum_{i=1}^{p}\sum_{j=1}^{p}\frac{u_{2hki}}{\mu_{hi}}\frac{u_{2hkj}}{\mu_{hj}}\hat{z}_{hij} = \sum_{i=1}^{p}\frac{u_{2hki}\hat{q}_{hki}}{\mu_{hi}} \tag{4.31}$$

with

$$\hat{z}_{hij} = \sum_{l\in s_h}\left(\sum_{a\in s_{hl}} u_{1haj}u_{1hai}w_a e_a^2 + \sum_{a\in s_{hl}} u_{1haj}\sum_{b\in s_{hl}, b\neq a} u_{1hbi}w_b^{1/2}w_a^{1/2}e_b e_a\right)$$

in

$$\hat{\boldsymbol{Z}}_{h(p\times p)} = (\hat{z}_{hij}) = \boldsymbol{U_{1h}}^T\boldsymbol{W}_h^{1/2}\boldsymbol{P}_h\boldsymbol{W}_h^{1/2}\boldsymbol{U_{1h}},$$

and

$$\hat{q}_{hki} = \sum_{j=1}^{p} u_{2hkj} \hat{z}_{hij} / \mu_{hj}$$

in

$$\hat{\boldsymbol{Q}}_{h(p \times p)} = (\hat{q}_{hki}) = \boldsymbol{U_{1h}}^T \boldsymbol{W}_h^{1/2} \boldsymbol{P}_h \boldsymbol{W}_h^{1/2} \boldsymbol{U}_{1h} \boldsymbol{D}_h^{-1} \boldsymbol{U}_{2h}^T = \hat{\boldsymbol{Z}}_h \boldsymbol{D}_h^{-1} \boldsymbol{U}_{2h}^T.$$

Therefore, the estimated proportion of the variance of the k^{th} regression coefficient associated with the i^{th} component of its decomposition in stratum h is: $\hat{\pi}_{hik} = \frac{u_{2hki} \hat{q}_{hki}}{\mu_{hi}} / var_M(\hat{\beta}_{SWhk})$.

Analogously, the linearization variance estimator $var_L(\hat{\boldsymbol{\beta}}_{SWh})$ in (4.30) can be decomposed:

$$var_L(\hat{\beta}_{SWhk}) = \frac{n_h}{n_h - 1} \sum_{i=1}^{p} \sum_{j=1}^{p} \frac{u_{2hki}}{\mu_{hi}} \frac{u_{2hkj}}{\mu_{hj}} \hat{z}_{hij} = \frac{n_h}{n_h - 1} \sum_{i=1}^{p} \frac{u_{2hki} \hat{q}_{hki}}{\mu_{hi}}; \tag{4.32}$$

its variance-decomposition proportions in each stratum are also the same as the ones for $var_M(\hat{\beta}_{SWhk})$:

$$\hat{\pi}_{hik} = \frac{n_h}{n_h - 1} \frac{u_{2hki} \hat{q}_{hki}}{\mu_{hi}} / var_L(\hat{\beta}_{SWhk}) = \frac{u_{2hki} \hat{q}_{hki}}{\mu_{hi}} / var_M(\hat{\beta}_{SWhk}). \tag{4.33}$$

4.2 Experimental Study

Belsley (1991) suggests that the presence of degrading collinearity requires the joint occurrence of high variance-decomposition proportions for two or more coefficients associated with a single condition index deemed to be large. Knowledge of what constitutes these high values must be determined empirically. In this section, two data sets will be used as our experimental studies to provide such experience. One is the SMHO data set with generated Y's and another is the data set generated from NHANES 2001-2002 data set, both of which have been described in the previous chapter and used for the simulation studies.

In each study, we will start from the generated study population and obtain the diagnostic statistics (scaled condition indexes and scaled variance decomposition proportions) for the two models compared in the previous chapter, of which one is the underlying model with relatively independent explanatory variables and another is the extended model with collinear variables. Parallel to our analysis in the previous chapter, three types of regressions are fitted using the full finite population and correspondingly their diagnostic statistics are computed:

OLS1: OLS formulas are used to estimate $\boldsymbol{\beta}$, the variance of $\hat{\boldsymbol{\beta}}_{OLS}$ and the diagnostic statistics are obtained using standard methods;

OLS2: OLS are used to estimate $\boldsymbol{\beta}$; the variance of $\hat{\boldsymbol{\beta}}_{OLS}$ and the diagnostic statistics are estimated assuming that heteroscedastic variances are known; the scaled condition indexes are estimated using (4.1) and the scaled variance decomposition proportions are estimated using (4.7) with $\boldsymbol{W} = \boldsymbol{I}$;

OLS3: OLS used to estimate $\boldsymbol{\beta}$; the variance of $\hat{\boldsymbol{\beta}}_{OLS}$ and the diagnostic statistics are estimated using an estimator of $\boldsymbol{V}$ matrix, denoted as $\hat{\boldsymbol{V}}$; the scaled condition indexes are estimated using (4.1) and the scaled variance decomposition proportions are estimated using (4.12) with z_{ij} in (4.21).

Then, we will draw one sample from the full finite population using the sampling scheme described in the previous chapter. Four types of regressions will be fitted for this particular sample to demonstrate different diagnostic methods for different regression methods and compare their results:

TYPE1: OLS regression with estimated σ^2 (unknown); the diagnostic statistics are obtained using standard methods;

TYPE2: WLS regression with estimated σ^2 (unknown) and assuming $\boldsymbol{V} = \boldsymbol{W}^{-1}$; the scaled condition indexes are estimated using (4.1) and the scaled variance decomposition proportions are estimated using (4.7);

TYPE3: SWLS with known $\sigma^2\boldsymbol{V}$; the scaled condition indexes are estimated using (4.1) and the scaled variance decomposition proportions are estimated using (4.7);

TYPE4: SWLS with estimated $\hat{\boldsymbol{V}}$, when $\sigma^2\boldsymbol{V}$ is unknown; the scaled condition indexes are estimated using (4.1); the scaled variance decomposition proportions are estimated using (4.12) with z_{ij} in (4.21) for the model-based variance estimators and with z_{ij} in (4.25) for the design-based variance estimators.

Selected tables displaying the scaled condition indexes and the variance decomposition proportions are reported for each study. A proportion larger than 0.3 will be highlighted in the table. Belsley et al. (1980) suggest that a condition index of 10 signals that collinearity has a moderate effect on standard errors; an index of 100 would indicate a serious effect. In this study, we consider a scaled condition index greater than 10 to be relatively large, and ones greater than 30 as large and remarkable. Furthermore, the large scaled variance-decomposition proportions (greater than 0.3) associated with each high scaled condition index identifies those variates that are involved in that near dependency. The magnitude of these proportions in conjunction with a large scaled condition index provides a measure of the degree to which the corresponding regression estimate has been degraded by the presence of collinearity. We also report the regression analysis output of the underlying models and extended models when different regression methods are applied to provide a link between the magnitudes of scaled condition indexes and these more familiar notations.

4.2.1 Survey of Mental Health Organizations

Table 4.1 presents the regression analysis output of the SMHO full finite population data using the three regression types, OLS1, OLS2 (with $\boldsymbol{V}$) and OLS3 (with $\hat{\boldsymbol{V}}$). Using OLS1, in the underlying model, all the parameter estimates of the three explanatory variables in the underlying model (FIRST, ADDS and EOYCNT) and the intercept are significant by a standard t-test. However, in the extended model, after including $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ in the model, the standard errors of the parameter estimates are inflated in different degrees along with the changes of their p-values. The most remarkable change is the estimated coefficient of FIRST, whose value changes from 0.83 in the underlying model to 0.40 in the extended model, the standard error is inflated from 0.18 to 0.50 and the p-value turns from a very significant value (smaller than 0.005) to an insignificant value (larger than 0.05). The significance of the intercept is also changed from significant to insignificant when the two collinear variables enter in the model, although the value and standard error of the intercept are just changed slightly. In contrast, the coefficients of the other two variables are still significant, while their values get smaller and their standard errors get larger in the extended model. It should be noted here that collinearity will degrade the regression estimates by inflating the standard errors and affecting the value of coefficients but this degradation need not be great enough actually to cause trouble for some purposes, such as testing the significance of the coefficients based on the p-values. Belsley (1991, chap. 3) gives more discussion about the difference between harmful and degrading collinearity. In OLS2 and OLS3, the standard errors of all the three explanatory variables in the underlying model are inflated in the extended model as in OLS1, although the degrees of inflation are different due to their different ways of variance estimation. Note that the underlying model and extended model give very similar fits as measured by R^2 here. Thus, including or excluding the collinear variables has little effect on this measure of overall fit.

Although Table 4.1 shows the changes of coefficients and their standard errors in the model fitting when the collinearity in the design matrix gets stronger, it can not tell us how many near dependencies exist among the ill-conditioned data and which variables are involved in them. To better understand the collinearity problem in the data, the analysis of the scaled condition indexes and variance decomposition proportions for the SMHO full finite population is reported in Table 4.2 and Table 4.3. Table 4.2 gives the diagnostic statistics for the underlying model. In all three regression types, no severe collinearity appears in the underlying model with all the scaled condition indexes smaller than 10, although there is a relatively small near-dependency between the intercept term and EOYCNT with a scaled condition index of 9. Note that the values of $u_{2ki}\hat{q}_{ki}$ in (4.14) can be negative, leading to the negative π_{ik} in Table 4.2.

When the two collinear variables, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$, enter in the model, two near dependencies appear in the model as shown in Table 4.3, one dominant with a large scaled condition index of 36 and one moderate with a scaled condition index of 26. Scanning across the bottom two rows, we can see that the dominant relation involves FIRST, ADDS and $\boldsymbol{X1}$ and in all regression types, approximately 90% of the variance for the coefficients of these three variables are associated with the largest scaled condition index. The weaker relation involves another two variables, EOYCNT and $\boldsymbol{X}_2$. Similarly, around 75% of the variance for the coefficient of EOYCNT and 95% of the variance for the coefficient of $\boldsymbol{X}_2$ are associated with the second largest scaled condition index. Although the third largest scaled condition index is 10, it only involves the intercept term and unlikely add too much to our knowledge about collinearity. It is likely that some of the other variables could be moderately correlated with the intercept (for example, EOYCNT as shown in Table 4.2), but their effects possibly are masked by the two stronger near dependencies.

We also applied the regression analysis and collinearity diagnostics to a sample selected from the SMHO full finite population. Selecting a sample will illustrate changes in the diagnostics due to unequal survey weights. The sample was selected using stratified single-stage sampling, in which strata were formed based on four organization types and fifty sample units in each stratum were selected with probability proportional to the rounded value of the size of EOYCNT multiplied by 100. Table 4.4 reports the regression analysis output of this sample. The results are quite similar across all the four regression types. FIRST and ADDS are significant in the underlying model while the intercept and EOYCNT are insignificant. In the extended model, FIRST becomes insignificant due to the variance inflation caused by its collinearity with the two additional variables, $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. Comparing the extended model with the underlying model, the standard errors of all the three explanatory variables are inflated in different degrees due to the positive relationship between the three explanatory variables and two collinear variables.

Just as in the full finite population, the scaled condition indexes are relatively small (less than 10) in the underlying model as seen in Table 4.5 but have two large values in the extended model as seen in Table 4.6. For each of the four regression types and their corresponding methods to obtain the scaled condition indexes and variance decomposition proportions, one dominant near dependency has a large scaled condition index of 36, which involves FIRST, ADDS and $\boldsymbol{X}_1$, and another has a scaled condition index of 25 (in TYPE1) or 24 (in all the other regression types), which involves EOYCNT and $\boldsymbol{X}_2$. The differences in the condition indexes between TYPE1 and all the other types are caused by their different ways of computation. TYPE1 uses the singular values of matrix $\boldsymbol{X}$, while the other types use the singular values of matrix $\tilde{\boldsymbol{X}}$. The scaled variance decomposition proportions are slightly different across different regression types but are close to each other. Around 90% of the variance for the coefficients of FIRST and ADDS and 80% of the variance for the coefficient of $\boldsymbol{X}_1$ are associated with the largest scaled condition index. Around 60% of the variance for the coefficient of EOYCNT and 70% of the variance for the coefficient of $\boldsymbol{X}_2$ are associated with the second largest scaled condition index.

Table 4.1: Regression Analysis Output of SMHO Full Finite Population Data

Regression Type	Model Type	Intercept	FIRST	ADDS	EOYCNT	$\boldsymbol{X1}$	$\boldsymbol{X2}$	R^2
OLS1	Underlying[a]	4.83*[b]	0.83***	1.50***	2.58***			0.4282
		(2.45)[c]	(0.18)	(0.04)	(0.49)			
	Extended[d]	4.79	0.40	1.39***	2.33*	0.39	0.22	0.4285
		(2.46)	(0.50)	(0.13)	(0.95)	(0.44)	(0.81)	
OLS2: with $\boldsymbol{V}$	Underlying	4.83	0.83***	1.50***	2.58***			0.4282
		(2.78)	(0.18)	(0.05)	(0.60)			
	Extended	4.79	0.40	1.39***	2.33*	0.39	0.22	0.4285
		(2.78)	(0.61)	(0.15)	(1.19)	(0.55)	(1.02)	
OLS3: with	Underlying	4.83	0.83***	1.50***	2.58***			0.4282
		(2.94)	(0.18)	(0.05)	(0.57)			
estimated $\hat{\boldsymbol{V}}$	Extended	4.79	0.40	1.39***	2.33*	0.39	0.22	0.4285
		(2.94)	(0.58)	(0.15)	(1.16)	(0.53)	(1.04)	

[a]The underlying model: $Y_i = \text{Intercept} + \text{FIRST}_i + \text{ADDS}_i + \text{EOYCNT}_i$
[b]p-value: *, 0.05; **, 0.01; ***, 0.005
[c]Standard errors are in parentheses under parameter estimates.
[d]The extended model: $Y_i = \text{Intercept} + \text{FIRST}_i + \text{ADDS}_i + \text{EOYCNT}_i + \boldsymbol{X}_{1i} + \boldsymbol{X}_{2i}$

Table 4.2: Scaled Condition Indexes and Variance Decomposition Proportions: the Underlying Model in SMHO Full Finite Population Data

Scaled Condition	Scaled Proportion of the Variance of			
Index	Intercept	FIRST	ADDS	EOYCNT
OLS1				
1	0.005	0.014	0.010	0.007
5	0.011	**0.880**	0.188	0.020
6	0.047	0.083	**0.705**	0.357
9	**0.937**	0.023	0.097	**0.616**
OLS2: with $\boldsymbol{V}$				
1	-0.021	-0.013	0.046	0.026
5	0.005	**0.897**	0.198	0.023
6	0.016	0.089	**0.636**	0.287
9	**1.000**	0.026	0.120	**0.664**
OLS3: with estimated $\hat{\boldsymbol{V}}$				
1	-0.021	0.002	0.048	0.026
5	0.014	**0.840**	0.150	0.016
6	0.039	0.106	**0.654**	0.259
9	**0.969**	0.052	0.148	**0.699**

Table 4.3: Scaled Condition Indexes and Variance Decomposition Proportions: the Extended Model with Collinear Variables in SMHO Full Finite Population Data

Scaled Condition	Scaled Proportion of the Variance of					
Index	Intercept	FIRST	ADDS	EOYCNT	$\boldsymbol{X}_1$	$\boldsymbol{X}_2$
OLS1						
1	0.002	0.001	0.001	0.001	0.000	0.000
6	0.017	0.100	0.015	0.012	0.001	0.001
7	0.034	0.000	0.059	0.067	0.007	0.003
10	**0.929**	0.000	0.002	0.075	0.000	0.020
26	0.015	0.001	0.000	**0.731**	0.098	**0.852**
36	0.004	**0.898**	**0.924**	0.115	**0.893**	0.125
OLS2: with $\boldsymbol{V}$						
1	-0.015	-0.001	0.003	0.004	0.000	0.000
6	0.007	0.066	0.013	0.012	0.001	0.000
7	0.009	0.000	0.038	0.046	0.007	0.003
10	**0.976**	0.000	0.002	0.079	0.000	0.018
26	0.020	0.000	0.000	**0.754**	0.110	**0.866**
36	0.004	**0.934**	**0.945**	0.107	**0.882**	0.112
OLS3: with estimated $\hat{\boldsymbol{V}}$						
1	-0.015	-0.002	0.002	0.003	0.001	0.001
6	0.020	0.055	0.016	0.010	0.003	0.000
7	0.026	0.000	0.034	0.040	0.010	0.002
10	**0.961**	0.001	-0.003	0.054	0.002	0.029
26	0.012	0.002	0.001	**0.779**	0.099	**0.845**
36	-0.003	**0.944**	**0.950**	0.115	**0.886**	0.123

Table 4.4: Regression Analysis Output of a Sample of SMHO Full Finite Population Data

Regression Type	Model Type	Intercept	FIRST	ADDS	EOYCNT	$\boldsymbol{X1}$	$\boldsymbol{X2}$	R^2
TYPE1	Underlying[a]	-3.35	1.36*[b]	1.56***	2.59			0.4668
OLS		(8.27)[c]	(0.53)	(0.13)	(1.46)			
	Extended[d]	-3.39	1.25	1.53***	3.15	0.24	-0.61	0.4670
		(8.34)	(1.62)	(0.40)	(2.85)	(1.38)	(2.45)	
TYPE2	Underlying	-8.83	1.46*	1.64***	2.95			0.4658
WLS		(8.55)	(0.58)	(0.13)	(1.54)			
	Extended	-8.88	1.60	1.67***	3.69	0.02	-0.74	0.4660
		(8.59)	(1.61)	(0.40)	(2.98)	(1.34)	(2.55)	
TYPE3	Underlying	-8.83	1.46*	1.64***	2.95			0.4658
SWLS		(9.25)	(0.58)	(0.17)	(1.83)			
with $\boldsymbol{V}$	Extended	-8.88	1.60	1.67***	3.69	0.02	-0.74	0.4660
		(9.25)	(2.01)	(0.53)	(3.55)	(1.81)	(3.25)	
TYPE4	Underlying	-8.83	1.46*	1.64***	2.95			0.4658
SWLS		(9.32)	(0.56)	(0.14)	(1.66)			
with est. $\hat{\boldsymbol{V}}$	Extended	-8.88	1.60	1.67***	3.69	0.02	-0.74	0.4660
model-based		(9.35)	(1.61)	(0.42)	(3.48)	(1.37)	(3.02)	
TYPE4	Underlying	-8.83	1.46*	1.64***	2.95			0.4658
SWLS		(9.33)	(0.56)	(0.14)	(1.68)			
with est. $\hat{\boldsymbol{V}}$	Extended	-8.88	1.60	1.67***	3.69	0.02	-0.74	0.4660
design-based		(9.36)	(1.61)	(0.42)	(3.52)	(1.36)	(3.05)	

[a]The underlying model: $Y_i = \text{Intercept} + \text{FIRST}_i + \text{ADDS}_i + \text{EOYCNT}_i$
[b]p-value: *, 0.05; **, 0.01; ***, 0.005
[c]Standard errors are in parentheses under parameter estimates.
[d]The extended model: $Y_i = \text{Intercept} + \text{FIRST}_i + \text{ADDS}_i + \text{EOYCNT}_i + \boldsymbol{X}_{1i} + \boldsymbol{X}_{2i}$

Table 4.5: Scaled Condition Indexes and Variance Decomposition Proportions: the Underlying Model in a Sample of SMHO Full Finite Population Data

Scaled Condition	Scaled Proportion of the Variance of			
Index	Intercept	FIRST	ADDS	EOYCNT
TYPE1: OLS				
1	0.005	0.014	0.011	0.007
4	0.001	**0.521**	**0.313**	0.001
6	0.045	**0.399**	**0.550**	**0.473**
9	**0.949**	0.066	0.125	**0.520**
TYPE2: WLS				
1	0.005	0.014	0.009	0.007
4	0.008	**0.785**	0.206	0.008
6	0.012	0.151	**0.520**	**0.530**
9	**0.975**	0.050	0.265	**0.455**
TYPE3: SWLS with $\boldsymbol{V}$				
1	-0.005	-0.020	0.045	0.001
4	0.014	**0.840**	0.244	-0.007
6	-0.002	0.114	**0.542**	**0.569**
9	**0.993**	0.065	0.169	**0.438**
TYPE4: SWLS with $\hat{\boldsymbol{V}}$, model-based				
1	-0.010	-0.005	0.049	0.011
4	0.017	**0.767**	0.190	-0.002
6	0.003	0.146	**0.507**	**0.487**
9	**0.990**	0.092	0.255	**0.504**
TYPE4: SWLS with $\hat{\boldsymbol{V}}$, design-based				
1	-0.011	-0.002	0.048	0.012
4	0.016	**0.758**	0.184	-0.001
6	0.004	0.153	**0.509**	**0.488**
9	**0.991**	0.091	0.259	**0.500**

Table 4.6: Scaled Condition Indexes and Variance Decomposition Proportions: the Extended Model with Collinear Variables in a Sample of SMHO Full Finite Population Data

Scaled Condition	Scaled Proportion of the Variance of					
Index	Intercept	FIRST	ADDS	EOYCNT	$\boldsymbol{X}_1$	$\boldsymbol{X}_2$
TYPE1: OLS						
1	0.002	0.001	0.001	0.001	0.000	0.000
5	0.001	0.057	0.033	0.000	0.000	0.000
7	0.044	0.019	0.031	0.094	0.010	0.004
10	**0.914**	0.000	0.002	0.060	0.001	0.019
25	0.014	0.002	0.002	**0.622**	0.142	**0.756**
36	0.026	**0.921**	**0.932**	0.224	**0.846**	0.220
TYPE2: WLS						
1	0.002	0.001	0.001	0.001	0.000	0.000
5	0.009	0.102	0.022	0.003	0.000	0.000
7	0.014	0.005	0.037	0.101	0.009	0.004
10	**0.953**	0.000	0.009	0.053	0.002	0.018
24	0.010	0.002	0.007	**0.575**	0.162	**0.687**
36	0.012	**0.891**	**0.925**	0.267	**0.826**	0.290
TYPE3: SWLS with $\boldsymbol{V}$						
1	-0.004	0.000	0.004	0.000	-0.001	0.000
5	0.014	0.045	0.033	0.005	0.001	-0.001
7	-0.002	0.006	0.049	0.067	0.004	0.013
10	**0.972**	0.000	0.002	0.043	0.001	0.021
24	0.009	0.009	0.014	**0.642**	0.203	**0.720**
36	0.011	**0.941**	**0.899**	0.244	**0.792**	0.247
TYPE4: SWLS with $\hat{\boldsymbol{V}}$, model-based						
1	-0.017	-0.001	0.003	0.007	0.001	-0.001
5	0.010	0.043	0.039	-0.001	0.003	0.001
7	0.003	0.002	0.015	0.064	0.015	0.004
10	**0.966**	0.000	0.004	0.079	0.008	0.003
24	0.032	-0.008	-0.003	**0.602**	0.166	**0.715**
36	0.006	**0.963**	**0.942**	0.251	**0.807**	0.278
TYPE4: SWLS with $\hat{\boldsymbol{V}}$, design-based						
1	-0.008	0.000	0.004	0.003	0.001	-0.001
5	0.017	0.051	0.039	-0.003	0.003	0.001
7	0.006	0.005	0.035	0.065	0.013	0.007
10	**0.948**	0.000	0.005	0.063	0.005	0.007
24	0.027	-0.009	-0.005	**0.607**	0.156	**0.705**
36	0.010	**0.954**	**0.922**	0.266	**0.824**	0.282

4.2.2 National Health and Nutrition Examination Survey: 2001-2002

Table 4.7 presents the regression analysis output of the NHANES full finite population data using the three regression types, OLS1, OLS2 (with $\boldsymbol{V}$) and OLS3 (with $\hat{\boldsymbol{V}}$). In all three regression types, although all the parameter estimates of the three explanatory variables in the underlying model (age, thigh circumference and subscapular skinfold) and the intercept are significant by a standard t-test in both underlying and extended models, their standard errors are inflated in different degrees due to the collinearity caused by including other collinear variables in the model. Note that the underlying model and extended model give very similar fits as measured by R^2 here. Thus, including or excluding the collinear variables has little effect on this measure of overall fit.

To examine the collinearity among all the explanatory variables in this finite population, first, we obtained the scaled condition indexes and variance decomposition proportions of the underlying model given in Table 4.8. One large condition index of 29 indicates a near-dependency among the predictors. In OLS1, this condition index is associated with two large (> 0.3) variance decomposition proportions for intercept and thigh circumference (0.963 and 0.9774 respectively). In OLS2 and OLS3, this condition index is not only associated with the two large scaled variance decomposition proportions for intercept and thigh circumference as in OLS1, but also associated with the one for subscapular skinfold. This implies that the near-dependency among intercept, thigh circumference and maybe with subscapular skinfold significantly affected the variances of the estimated associated parameters in the underlying model. Furthermore, recalling our analysis in the VIF chapter, this near-dependency in the underlying model has not been detected using the VIF method due to two shortcomings of VIF. One is its inadequacy to diagnose the influence of collinearity between intercept and other variables. Researchers usually want to keep the intercept in the model by default and will neglect to investigate the variance inflation of certain predictor caused by the intercept. Another is the deficiency of VIF when detecting moderate collinearity, such like the collinearity between thigh circumference and subscapular skinfold here.

Table 4.9 lists the scaled condition indexes and variance decomposition proportions of the extended model, which also includes the other three collinear variables, body weight, BMI and waist circumference. The largest condition index is 84 and is associated with the intercept, thigh circumference and the three collinear variables. The second largest condition index is 40, but only involves waist circumference. It is likely that waist circumference is correlated with some of the other variables whose effects may be masked by the strongest near dependency. The third largest condition index is 38 and is associated with body weight and BMI.

We also applied the regression analysis and collinearity diagnostics to a sample selected from the NHANES full finite population. The sample was selected by a stratified clustered sampling. First, within each stratum in the study population, 6 PSUs were drawn out of 100 PSUs by simple random sampling with replacement; then within each sampled PSU, 5 men and 20 women were randomly selected without replacement. The selection probability for a person of gender g in PSU hi was then

$$\pi_{hit} = \frac{n_h}{N_h} \cdot \frac{m_{hi(g)}}{M_{hi(g)}}$$

where $m_{hi(g)} = 5$ for men and 20 for women and $M_{hi(g)}$ was the population count of persons of gender g in the PSU.

Table 4.10 reports the regression analysis output of this sample. The results are quite similar across all the four regression types. In TYPE1, the intercept, age and subscapular skinfold are significant from fitting only the variables in the underlying model but thigh circumference is insignificant. This might be due to the correlation between the intercept and thigh circumference. In the extended model, the standard errors of the coefficients are inflated in different degrees while the three collinear variables are all insignificant. In TYPE2, TYPE3 and TYPE4, only the intercept and age are significant in the underlying model. In TYPE2 and TYPE4, the intercept, age, subscapular and body weight are significant while the other variables are all insignificant in the extended model. In TYPE3, only the intercept, age and body are significant. As we can see in this table, the collinearity among the explanatory variables inflates the standard errors of the coefficients in the underlying model no matter which regression method are applied, while this collinearity plays slightly different role in the regression estimation when different regression methods of variance estimation are used.

In this sample, we report the scaled condition indexes and variance decomposition proportions of the underlying model (in Table 4.11) and the extended model (in Table 4.12). In the underlying model, TYPE1 uses the condition indexes that are defined by the eigenvalues of the design matrix $\boldsymbol{X}$. One strong near dependency with a scaled condition index of 28 involves the intercept, thigh circumference and subscapular skinfold. TYPE2, TYPE3 and TYPE4 use the condition indexes that are defined by the eigenvalues of the weighted design matrix $\tilde{\boldsymbol{X}}$. One strong near dependency is also shown with a scaled condition index of 30. In TYPE2 and TYPE3, this near dependency also involves the intercept, thigh circumference and subscapular skinfold. However, in TYPE4, this near dependency only involves the intercept and thigh circumference. In the extended model, we can see at least one dominant near dependency with a scaled condition index of 80 in TYPE1 or 87 in other three regression types. Clearly, at least the intercept, body weight, BMI, waist circumference and thigh circumference are involved in this dominant near dependency. In TYPE1, the second largest scaled condition index is 39 and is associated with a near dependency which involves body weight and BMI. Further, the intercept, waist circumference and thigh circumference could conceivably be involved in this weaker near dependency, their effects possibly being masked by the dominant dependency. In TYPE2, TYPE3 and TYPE4, it seems that waist circumference is the only variable likely involved in the second strongest near dependency with a scaled condition indexes of 41. Although the other explanatory variables involved in the dominant dependency might also be involved in the weaker near dependency, their effects possibly are masked by the dominant dependency.

Table 4.7: Regression Analysis Output of NHANES Full Finite Population Data

Regression Type	Model Type	Intercept	age	thigh circumference	subscapular skinfold	body weight	BMI	waist circumference	R^2
OLS1	Underlying[a]	90.01***[b]	0.47***	0.16***	0.11***				0.1498
		(0.75)[c]	(0.01)	(0.02)	(0.01)				
	Extended[d]	89.55***	0.47***	0.18***	0.12***	0.00	-0.07	0.01	0.1498
		(1.19)	(0.01)	(0.03)	(0.02)	(0.01)	(0.05)	(0.02)	
OLS2: with $\boldsymbol{V}$	Underlying	90.01***	0.47***	0.16***	0.11***				0.1498
		(0.90)	(0.01)	(0.02)	(0.01)				
	Extended	89.55***	0.47***	0.18***	0.12***	0.00	-0.07	0.01	0.1498
		(1.37)	(0.01)	(0.04)	(0.02)	(0.01)	(0.05)	(0.02)	
OLS3: with estimated $\hat{\boldsymbol{V}}$	Underlying	90.01***	0.47***	0.16***	0.11***				0.1498
		(0.90)	(0.01)	(0.02)	(0.01)				
	Extended	89.55***	0.47***	0.18***	0.12***	0.00	-0.07	0.01	0.1498
		(1.40)	(0.01)	(0.04)	(0.02)	(0.01)	(0.06)	(0.02)	

[a] The underlying model: $Y_i = \text{Intercept} + \text{age}_i + \text{thighcircumference}_i + \text{subscapularskinfold}_i$

[b] p-value: *, 0.05; **, 0.01; ***, 0.005

[c] Standard errors are in parentheses under parameter estimates.

[d] The extended model: $Y_i = \text{Intercept} + \text{age}_i + \text{bodyweight}_i + \text{BMI}_i + \text{waistcircumference}_i + \text{thighcircumference}_i + \text{subscapularskinfold}_i$

Table 4.8: Scaled Condition Indexes and Variance Decomposition Proportions: the underlying model in NHANES full finite population data set

Scaled Condition Condition Index	Scaled Proportion of the Variance of Intercept	age	thigh circumference	subscapular skinfold
OLS1				
1	0.001	0.008	0.001	0.006
6	0.000	**0.657**	0.002	**0.316**
7	0.037	0.240	0.020	**0.395**
29	**0.963**	0.096	**0.977**	0.284
OLS2: with $\boldsymbol{V}$				
1	0.018	0.014	0.000	0.007
6	0.000	**0.630**	0.001	0.264
7	0.039	0.258	0.032	**0.412**
29	**0.943**	0.098	**0.968**	**0.318**
OLS3: with estimated $\hat{\boldsymbol{V}}$				
1	0.019	0.027	-0.001	0.016
6	0.000	**0.627**	-0.003	0.272
7	0.033	0.236	0.033	**0.375**
29	**0.949**	0.111	**0.972**	**0.336**

Table 4.9: Scaled Condition Indexes and Variance Decomposition Proportions: the extended model with collinear variables in NHANES full finite population data set

Scaled Condition Index	Scaled Proportion of the Variance of intercept	age	body weight	BMI	waist circumference	thigh circumference	subscapular skinfold
OLS1							
1	0.000	0.002	0.000	0.000	0.000	0.000	0.001
8	0.000	**0.680**	0.002	0.001	0.000	0.000	0.072
9	0.008	0.086	0.001	0.000	0.001	0.002	**0.420**
19	0.079	0.018	0.152	0.005	0.001	0.002	0.207
38	0.061	0.004	**0.300**	**0.514**	0.002	0.006	0.249
40	0.005	0.178	0.040	0.004	**0.371**	0.189	0.047
84	**0.847**	0.032	**0.506**	**0.477**	**0.625**	**0.801**	0.004
OLS2: with $\boldsymbol{V}$							
1	0.005	0.004	0.001	0.000	0.000	0.000	0.002
8	0.000	**0.616**	0.002	0.002	0.000	0.000	0.045
9	0.003	0.099	-0.002	0.000	0.003	0.005	**0.432**
19	0.098	0.017	0.132	0.002	0.002	-0.001	0.187
38	0.056	-0.004	**0.324**	**0.515**	-0.002	0.001	0.255
40	0.009	0.231	0.061	0.015	**0.432**	0.249	0.072
84	**0.830**	0.037	**0.483**	**0.467**	**0.566**	**0.747**	0.008
OLS3: with estimated $\hat{\boldsymbol{V}}$							
1	0.005	0.008	0.001	0.001	0.000	-0.001	0.003
8	0.000	**0.625**	-0.002	0.003	0.000	0.000	0.052
9	0.001	0.081	-0.002	0.000	0.002	0.007	**0.455**
19	0.105	0.033	0.147	-0.012	0.003	-0.005	0.182
38	0.033	-0.009	**0.339**	**0.514**	0.000	-0.002	0.243
40	0.009	0.224	0.064	0.013	**0.408**	0.236	0.055
84	**0.848**	0.038	**0.454**	**0.481**	**0.587**	**0.764**	0.011

Table 4.10: Regression Analysis Output of A Sample of NHANES Full Finite Population Data

Regression Type	Model Type	Intercept	age	thigh circumference	subscapular skinfold	body weight	BMI	waist circumference	R^2
TYPE1	Underlying[a]	98.93***[b]	0.41***	-0.01	0.29**				0.1533
OLS		(5.66)[c]	(0.04)	(0.12)	(0.10)				
	Extended[d]	103.62***	0.41***	-0.10	0.34**	0.17	-0.36	-0.04	0.1580
		(9.06)	(0.05)	(0.23)	(0.12)	(0.10)	(0.35)	(0.12)	
TYPE2	Underlying	94.22***	0.46***	0.10	0.16				0.1598
WLS		(6.17)	(0.05)	(0.13)	(0.10)				
	Extended	107.61***	0.44***	-0.21	0.26*	0.29**	-0.32	-0.13	0.1727
		(9.98)	(0.05)	(0.25)	(0.12)	(0.10)	(0.37)	(0.13)	
TYPE3	Underlying	94.22***	0.46***	0.10	0.16				0.1598
SWLS		(7.67)	(0.06)	(0.16)	(0.13)				
with $\boldsymbol{V}$	Extended	107.61***	0.44***	-0.21	0.26	0.29*	-0.32	-0.13	0.1727
		(12.56)	(0.06)	(0.32)	(0.15)	(0.13)	(0.46)	(0.15)	
TYPE4	Underlying	94.22***	0.46***	0.10	0.16				0.1598
SWLS		(5.68)	(0.05)	(0.11)	(0.10)				
with est. $\hat{\boldsymbol{V}}$	Extended	107.61***	0.44***	-0.21	0.26*	0.29*	-0.32	-0.13	0.1727
model-based		(9.91)	(0.05)	(0.28)	(0.12)	(0.12)	(0.42)	(0.13)	
TYPE4	Underlying	94.22***	0.46***	0.10	0.16				0.1598
SWLS		(5.68)	(0.05)	(0.11)	(0.10)				
with est. $\hat{\boldsymbol{V}}$	Extended	107.61***	0.44***	-0.21	0.26*	0.29*	-0.32	-0.13	0.1727
design-based		(9.91)	(0.05)	(0.28)	(0.12)	(0.12)	(0.42)	(0.13)	

[a]The underlying model: $Y_i = \text{Intercept} + \text{age}_i + \text{thighcircumference}_i + \text{subscapularskinfold}_i$

[b]p-value: *, 0.05; **, 0.01; ***, 0.005

[c]Standard errors are in parentheses under parameter estimates.

[d]The extended model: $Y_i = \text{Intercept} + \text{age}_i + \text{bodyweight}_i + \text{BMI}_i + \text{waistcircumference}_i + \text{thighcircumference}_i + \text{subscapularskinfold}_i$

Table 4.11: Scaled Condition Indexes and Variance Decomposition Proportions: the underlying model in a sample of NHANES full finite population

Scaled Condition Index	Scaled Proportion of the Variance of Intercept	age	thigh circumference	subscapular skinfold
TYPE1: OLS				
1	0.001	0.007	0.001	0.005
6	0.000	**0.666**	0.002	0.293
7	0.045	0.240	0.019	**0.352**
28	**0.955**	0.086	**0.978**	**0.350**
TYPE2: WLS				
1	0.001	0.006	0.001	0.005
7	0.006	**0.326**	0.000	**0.573**
7	0.028	**0.558**	0.021	0.083
30	**0.965**	0.110	**0.979**	**0.339**
TYPE3: SWLS with $\boldsymbol{V}$				
1	0.003	0.017	0.002	-0.002
7	0.003	0.291	0.000	**0.539**
7	0.027	**0.590**	0.032	0.110
30	**0.967**	0.102	**0.966**	**0.353**
TYPE4: SWLS with $\hat{\boldsymbol{V}}$, model-based				
1	0.001	0.048	-0.004	0.043
7	0.042	0.218	-0.002	**0.649**
7	0.052	**0.708**	0.028	0.187
30	**0.904**	0.026	**0.978**	0.121
TYPE4: SWLS with $\hat{\boldsymbol{V}}$, design-based				
1	0.002	0.048	-0.004	0.044
7	0.043	0.218	-0.002	**0.650**
7	0.052	**0.709**	0.028	0.188
30	**0.904**	0.025	**0.978**	0.119

Table 4.12: Scaled Condition Indexes and Variance Decomposition Proportions: the extended model with collinear variables in a sample of NHANES full finite population

Scaled Condition Index	Scaled Proportion of the Variance of intercept	age	body weight	BMI	waist circumference	thigh circumference	subscapular skinfold
TYPE1: OLS							
1	0.000	0.002	0.000	0.000	0.000	0.000	0.001
8	0.000	**0.707**	0.002	0.001	0.000	0.000	0.070
9	0.011	0.096	0.001	0.000	0.001	0.002	**0.393**
19	0.087	0.023	0.146	0.014	0.001	0.001	**0.325**
36	0.018	0.065	0.065	0.185	0.235	0.118	0.030
39	0.015	0.053	**0.350**	**0.352**	0.153	0.049	0.176
80	**0.869**	0.055	**0.437**	**0.449**	**0.610**	**0.829**	0.004
TYPE2: WLS							
1	0.000	0.002	0.000	0.000	0.000	0.000	0.001
8	0.000	**0.723**	0.002	0.001	0.000	0.001	0.019
9	0.008	0.002	0.000	0.000	0.001	0.001	**0.453**
19	0.068	0.010	0.160	0.006	0.001	0.002	0.217
36	0.050	0.055	0.160	**0.366**	0.062	0.038	0.183
41	0.002	0.160	0.191	0.112	**0.325**	0.118	0.116
87	**0.873**	0.049	**0.486**	**0.515**	**0.611**	**0.839**	0.012
TYPE3: SWLS with $\boldsymbol{V}$							
1	0.002	0.003	0.004	-0.001	0.000	-0.001	-0.001
8	0.001	**0.750**	0.004	0.002	-0.001	0.001	0.014
9	0.009	0.006	0.001	0.000	-0.001	0.002	**0.479**
19	0.063	0.016	0.187	0.002	0.001	0.003	0.216
36	0.051	0.036	0.173	**0.361**	0.055	0.043	0.181
41	0.001	0.127	0.181	0.121	**0.324**	0.120	0.101
87	**0.873**	0.063	**0.450**	**0.515**	**0.623**	**0.831**	0.011
TYPE4: SWLS with $\hat{\boldsymbol{V}}$, model-based							
1	0.007	-0.001	0.011	0.003	-0.003	-0.007	0.009
8	0.001	**0.800**	0.011	0.001	-0.002	0.001	0.011
9	0.012	0.013	-0.002	0.000	0.004	0.004	**0.572**
19	0.023	-0.020	0.241	-0.002	0.003	0.014	0.067
36	0.038	0.049	0.227	**0.385**	0.078	0.036	0.141
41	0.000	0.091	0.125	0.135	0.292	0.135	0.190
87	**0.920**	0.068	**0.390**	**0.479**	**0.627**	**0.819**	0.012
TYPE4: SWLS with $\hat{\boldsymbol{V}}$, design-based							
1	0.007	-0.001	0.011	0.003	-0.003	-0.007	0.010
8	0.001	**0.799**	0.010	0.001	-0.002	0.001	0.011
9	0.012	0.013	-0.002	0.000	0.004	0.004	**0.574**
19	0.021	-0.020	0.240	-0.002	0.003	0.014	0.065
36	0.038	0.050	0.227	**0.384**	0.078	0.036	0.141
41	-0.001	0.092	0.124	0.135	0.292	0.135	0.189
87	**0.921**	0.067	**0.390**	**0.479**	**0.628**	**0.818**	0.011

Reference Level for Categorical Variables

As been discussed in the previous section, in linear regression analysis, the dummy variables can also play an important role as a possible source for collinearity. The choice of reference level for a categorical variable may affect the degree of collinearity in the data. To be more specific, choosing a category that has a low frequency as the reference may give rise to collinearity with the intercept term. This phenomenon carries over to survey data analysis as we now illustrate.

To explore this influence, we now add a categorical variable (body type) in the underlying model. Body type has three categories defined by respondent's BMI (underweight: BMI$\leq$18.5; normal or overweight: 18.5$<$BMI$<$30; obese: BMI$\geq$30) and hence we need two dummy variables to represent them as BT1 and BT2. Because thigh circumference was diagnosed as having a problematic correlation with the intercept in Table 4.8, we will leave it out of this study model to avoid its influence on diagnosing the near dependency between the dummy variables and the intercept. The model considered here is:

$$\begin{aligned} Y_{hit} = \beta_0 + \beta_{\text{age}} * \text{age}_{hit} + \beta_{\text{BT1}} * \text{BT1}_{hit} + \\ \beta_{\text{BT2}} * \text{BT2}_{hit} + \beta_{\text{BMXSUB}} * \text{BMXSUB}_{hit} + \varepsilon_{hit} \end{aligned} \tag{4.34}$$

where subscript hit stands for the t^{th} unit in the selected PSU hi and BMXSUB is the variable name of subscapular skinfold. Only 3% of the respondents are underweight. If we choose underweight as a reference category, then we expect a near dependency between the corresponding dummy variable and intercept term. Around 79% of the respondents are normal or overweight. If we choose normal/overweight as a reference category, then we expect no severe collinearity between dummy variables and the intercept and the variance of intercept is smaller.

From the output in Table 4.13, we can see that both normal/overweight and obesity are positively related to Y when choosing underweight as the reference category in the full finite population while these relationships are not significant in the selected sample. This result is consistent with the output in Table 4.14. Underweight has a significant negative relationship with Y and obesity has a significant positive relationship with Y relative to the baseline of normal/overweight in the full finite population. However, these relationships are not significant in the selected sample. Note that the standard error of the intercept when choosing underweight as the reference category (in Table 4.13) is much smaller than the value of the standard error of the intercept when choosing normal/overweight as the reference category (in Table 4.13).

The scaled condition indexes and variance decomposition proportions show a medium degree of collinearity (condition index larger than 10) between the dummy variables for body type and the intercept, when choosing underweight as the reference category, in both the full finite population and the selected sample. Only the last row for the largest condition index is printed in Table 4.15. For the full finite population data, the intercept and the two dummy variables for body types are involved with the near dependency with the largest scaled condition index of 16, although the scaled variance decomposition proportions vary among different regression types. Similarly, for the selected sample data, the intercept and the two dummy variables for body types are involved with the near dependency with the largest scaled condition index. The largest scaled condition index is 14 in TYPE1 (OLS) without considering the survey weights, and is a little larger, 15, in other regression types when taking the survey weights into account. Besides, the scaled variance decomposition proportions vary among different regression types and inference approaches.

In contrast, choosing the normal/overweight as the reference category lowers the scaled condition indexes. The scaled condition indexes and variance decomposition proportions are stated in Table 4.16. Only the last row is printed in Table 4.16. There is no remarkable near-dependency in both full finite population and the selected sample, no matter which regression type or inference approach is used.

Table 4.13: Regression Analysis Output: When Underweight is the Reference Category in the Model

Regression Type	Intercept	age	BT1 (overweight)	BT2 (obesity)	subscapular skinfold
		In the Full Finite Population			
OLS1	96.65***	0.46***	1.42***	2.60**	0.15***
	(0.47)	(0.01)	(0.46)	(0.52)	(0.01)
OLS2	96.65***	0.46***	1.42***	2.60***	0.15***
	(0.56)	(0.01)	(0.48)	(0.56)	(0.01)
OLS3	96.65***	0.46***	1.42***	2.60***	0.15***
	(0.55)	(0.01)	(0.48)	(0.53)	(0.01)
		In One Selected Sample			
TYPE1	102.45***	0.41***	-5.44	-6.73	0.36***
OLS	(3.20)	(0.04)	(3.10)	(3.64)	(0.09)
TYPE2	99.27***	0.45***	-0.76	-2.10	0.25**
WLS	(3.53)	(0.05)	(3.44)	(3.91)	(0.09)
TYPE3	99.27***	0.45***	-0.76	-2.10	0.25*
SWLS with $\boldsymbol{V}$	(4.44)	(0.06)	(4.08)	(4.80)	(0.12)
TYPE4	99.27***	0.45***	-0.76	-2.10	0.25*
SWLS with $\hat{\boldsymbol{V}}$, model-based	(4.07)	(0.05)	(3.51)	(4.03)	(0.10)
TYPE4	99.27***	0.45***	-0.76	-2.10	0.25*
SWLS with $\hat{\boldsymbol{V}}$, design-based	(3.81)	(0.05)	(3.10)	(3.61)	(0.10)

Table 4.14: Regression Analysis Output: When Normal/Overweight is the Reference Category in the Model

Regression Type	Intercept	age	BT1 (overweight)	BT2 (obesity)	subscapular skinfold
			In the Full Finite Population		
OLS1	98.07***	0.46***	-1.42***	1.18**	0.15***
	(0.27)	(0.01)	(0.46)	(0.23)	(0.01)
OLS2	98.07***	0.46***	-1.42***	1.18**	0.15***
	(0.39)	(0.01)	(0.48)	(0.25)	(0.01)
OLS3	98.07***	0.46***	-1.42***	1.18**	0.15***
	(0.38)	(0.01)	(0.48)	(0.24)	(0.01)
			In One Selected Sample		
TYPE1	97.01***	0.41***	5.44	-1.29	0.36***
OLS	(2.20)	(0.04)	(3.10)	(1.70)	(0.09)
TYPE2	98.51***	0.45***	0.76	-1.34	0.25**
WLS	(2.12)	(0.05)	(3.44)	(1.71)	(0.09)
TYPE3	98.51***	0.45***	0.76	-1.34	0.25*
SWLS with $\boldsymbol{V}$	(3.03)	(0.06)	(4.08)	(2.32)	(0.12)
TYPE4	98.51***	0.45***	0.76	-1.34	0.25*
SWLS with $\hat{\boldsymbol{V}}$, model-based	(2.31)	(0.05)	(3.51)	(1.73)	(0.10)
TYPE4	98.51***	0.45***	0.76	-1.34	0.25*
SWLS with $\hat{\boldsymbol{V}}$, design-based	(2.28)	(0.05)	(3.11)	(1.67)	(0.10)

Table 4.15: Largest Scaled Condition Indexes and Its Associated Variance Decomposition Proportions: When Underweight is the Reference Category in the Model

Scaled Condition Index	Scaled Proportion of the Variance of				
	Intercept	age	BT1(overweight)	BT2(obesity)	subscapular skinfold
			In the Full Finite Population		
OLS1					
16	**0.912**	0.023	**0.909**	**0.705**	0.055
OLS2: with $\boldsymbol{V}$					
16	**0.824**	0.027	**0.907**	**0.694**	0.002
OLS3: with estimated $\hat{\boldsymbol{V}}$					
16	**0.831**	0.029	**0.897**	**0.716**	0.001
			In One Selected Sample		
TYPE1: OLS					
14	**0.875**	0.045	**0.649**	**0.870**	0.011
TYPE2: WLS					
15	**0.907**	0.047	**0.715**	**0.898**	0.015
TYPE3: SWLS with $\boldsymbol{V}$					
15	**0.854**	0.055	**0.657**	**0.891**	0.013
TYPE4: SWLS with $\hat{\boldsymbol{V}}$, model-based					
15	**0.870**	0.086	**0.759**	**0.967**	0.006
TYPE4: SWLS with $\hat{\boldsymbol{V}}$, design-based					
15	**0.870**	0.086	**0.759**	**0.967**	0.006

Table 4.16: Largest Scaled Condition Indexes and Its Associated Variance Decomposition Proportions: When Underweight is the Reference Category in the Model

Scaled Condition Index	Scaled Proportion of the Variance of				
	Intercept	age	BT1(overweight)	BT2(obesity)	subscapular skinfold
			In the Full Finite Population		
OLS1					
8	0.955	0.214	0.071	0.178	0.591
OLS2: with $\boldsymbol{V}$					
8	0.829	0.246	0.073	0.200	0.631
OLS3: with estimated $\hat{\boldsymbol{V}}$					
8	0.819	0.209	0.070	0.187	0.590
			In One Selected Sample		
TYPE1: OLS					
9	0.958	0.227	0.105	0.198	0.592
TYPE2: WLS					
7	0.970	0.255	0.075	0.135	0.440
TYPE3: SWLS with $\boldsymbol{V}$					
7	0.944	0.288	0.084	0.156	0.492
TYPE4: SWLS with $\hat{\boldsymbol{V}}$, model-based					
7	0.991	0.268	0.028	0.132	0.485
TYPE4: SWLS with $\hat{\boldsymbol{V}}$, design-based					
7	0.991	0.269	0.028	0.131	0.485

Chapter 5

Collinearity Diagnostics in Generalized Linear Models

The class of generalized linear models (GLMs) is an extension of traditional linear models that allows the linear model to be related to the response variable via a linear or nonlinear link function and allows the response probability distribution to be any member of an exponential family of distributions. Although we developed the collinearity diagnostics for linear models using complex survey data in the previous two chapters, these collinearity diagnostics are not adequate for the detection of collinearity in GLMs, as will be discussed below. The purposes of this chapter are to investigate the sources and consequences of collinearity in GLMs.

5.1 Survey-Weighted Generalized Linear Models

In the complex survey design, denote the sampled data set as s and suppose there are n observations in s. The log-likelihood of the generalized linear model in (2.12) is estimated by summing the natural logarithm of each of the components defined by (2.8) over the n observations weighted by the survey weights w_i:

$$\hat{l}(\boldsymbol{\beta}) = \sum_{i \in s} w_i [y_i \theta_i - b(\theta_i)]/\tau^2 - \sum_{i \in s} w_i c(y_i, \tau). \tag{5.1}$$

By setting the first-order partials for $\boldsymbol{\beta}$ equal to zero, the pseudo-maximum likelihood estimating equations for $\boldsymbol{\beta}$ are given by:

$$\begin{aligned}
\frac{\partial \hat{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{1}{\tau^2} \sum_{i \in s} w_i \left[y_i \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} - \partial b(\theta_i)/\partial \theta_i \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \right] \\
&= \frac{1}{\tau^2} \sum_{i \in s} w_i (y_i - \mu_i) \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \\
&= \frac{1}{\tau^2} \sum_{i \in s} w_i (y_i - \mu_i) \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \\
&= \frac{1}{\tau^2} \sum_{i \in s} w_i \frac{(y_i - \mu_i)}{v(\mu_i) g'(\mu_i)} \dot{\boldsymbol{x}}_i \quad \text{using (2.10) and (2.11)} \\
&= \frac{1}{\tau^2} \sum_{i \in s} w_i (y_i - \mu_i) \gamma_i g'(\mu_i) \dot{\boldsymbol{x}}_i \\
&= \mathbf{0}^T
\end{aligned} \tag{5.2}$$

upon defining $\gamma_i = \{v(\mu_i)[g'(\mu_i)]^2\}^{-1}$ with $g'(\mu_i) = \partial g(\mu_i)/\partial \mu_i$.

We can also write this in matrix notation as

$$\frac{\partial \hat{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{1}{\tau^2} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\Gamma} \boldsymbol{\Delta} (\boldsymbol{y} - \boldsymbol{\mu}) = \mathbf{0}^T, \tag{5.3}$$

with $\boldsymbol{W} = diag(w_i)$, $\boldsymbol{\Gamma} = diag(\gamma_i)$, $\boldsymbol{\Delta} = diag\,[g'(\mu_i)]$ and $\boldsymbol{\mu} = (\mu_i)_{n \times 1}$. The value of $\boldsymbol{\beta}$ that solves the survey-weighted estimating equations will be denoted by $\hat{\boldsymbol{\beta}}_{SW}$. To distinguish the generalized linear models in the complex survey design from the ordinary generalized linear models, we will refer to *survey-weighted generalized linear models* (SWGLM) since the survey weights are incorporated in the pseudo-maximum likelihood estimation.

5.2 Variance Inflation Factor in Generalized Linear Model

5.2.1 Model-based VIF

To derive the large-sample model variance of $\hat{\boldsymbol{\beta}}_{SW}$ and obtain its information matrix $\boldsymbol{I}(\hat{\boldsymbol{\beta}}_{SW})$, we can obtain the expected value of the second derivative of the log likelihood $\hat{l}(\boldsymbol{\beta})$ at first:

$$\frac{\partial^2 \hat{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = -\frac{1}{\tau^2} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\Gamma} \boldsymbol{\Delta} \frac{\boldsymbol{\mu}}{\partial \boldsymbol{\beta}^T} + \frac{1}{\tau^2} \boldsymbol{X}^T \boldsymbol{W} \frac{\partial \boldsymbol{\Gamma} \boldsymbol{\Delta}}{\partial \boldsymbol{\beta}^T} (\boldsymbol{y} - \boldsymbol{\mu}) \tag{5.4}$$

so that

$$\begin{aligned} -E\left(\frac{\partial^2 \hat{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right) &= \frac{1}{\tau^2} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\Gamma} \boldsymbol{\Delta} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}^T} + \mathbf{0} \\ &= \frac{1}{\tau^2} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\Gamma} \boldsymbol{\Delta} \boldsymbol{\Delta}^{-1} \boldsymbol{X} \quad \text{using (2.11)} \\ &= \frac{1}{\tau^2} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\Gamma} \boldsymbol{X}. \end{aligned} \tag{5.5}$$

Since

$$\begin{aligned} -E\left(\frac{\partial^2 \hat{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \tau^2}\right) &= -E\left[\frac{\partial}{\partial \tau^2} \frac{1}{\tau^2} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\Gamma} \boldsymbol{\Delta} (\boldsymbol{y} - \boldsymbol{\mu})\right] \\ &= \frac{1}{\tau^4} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\Gamma} \boldsymbol{\Delta} E(\boldsymbol{y} - \boldsymbol{\mu}) \\ &= \mathbf{0}, \end{aligned} \tag{5.6}$$

the estimation of τ^2 does not affect the large-sample variance of $\hat{\boldsymbol{\beta}}_{SW}$.

Under the model-based inference, the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{SW}$ is (see McCulloch & Searle, 2001):

$$avar_M(\hat{\boldsymbol{\beta}}_{SW}) = [\hat{\boldsymbol{I}}(\boldsymbol{\beta})]^{-1} = \left\{ -E\left[\frac{\partial^2 \hat{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} + \frac{\partial^2 \hat{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \tau^2} \right] \right\}^{-1} = \tau^2 (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\Gamma} \boldsymbol{X})^{-1} \tag{5.7}$$

where $avar_M$ stands for the limiting or asymptotic model variance. Notice that the model variance does not have the same sandwich form as for linear models (see expression (3.2)), because if $E_M(y_i)$ is specified correctly, then so is $Var_M(y)_i$. This was not true for a linear model.

In generalized linear models, we aim to estimate the parameters of interest, $\boldsymbol{\beta}$, so that we can also estimate the expectation of y_i, μ_i, conditional on $\dot{\boldsymbol{x}}_i$. Here we denote the estimated μ_i as $\hat{\mu}_i$, $\hat{\gamma}_i = \{v(\hat{\mu}_i)[g'(\hat{\mu}_i)]^2\}^{-1}$ with $g'(\hat{\mu}_i) = \partial g(\hat{\mu}_i)/\partial \hat{\mu}_i$, and $\hat{\boldsymbol{\Gamma}} = diag(\hat{\gamma}_i)$.

With $\boldsymbol{A} = \boldsymbol{X}^T \boldsymbol{W} \hat{\boldsymbol{\Gamma}} \boldsymbol{X}$, the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{SW}$ can be estimated by:

$$avar_M(\hat{\boldsymbol{\beta}}_{SW}) = \hat{\tau}^2 \boldsymbol{A}^{-1}, \tag{5.8}$$

and its k^{th} element on the main diagonal is the asymptotic model variance of $\hat{\beta}_{SWk}$, which, equivalently, can be estimated by:

$$avar_M(\hat{\beta}_{SWk}) = \hat{\tau}^2 a^{kk}, \tag{5.9}$$

where a^{kk} is the k^{th} element on the main diagonal of matrix $\boldsymbol{A}^{-1}$. In some models, τ must be estimated. For the analysis here, no estimate is needed, because the VIF given below does not include τ.

As illustrated in (3.7) in Chapter 3, a^{kk} can be expressed as the inverse of the sum of squared errors (SSE) from the weighted least squares regression of $\boldsymbol{x}_k$ on the $p-1$ other explanatory variables with the weight matrix $\boldsymbol{W}\hat{\boldsymbol{\Gamma}}$:

$$a^{kk} = \boldsymbol{i}_k^T \boldsymbol{A}^{-1} \boldsymbol{i}_k = \boldsymbol{i}_k^T (\boldsymbol{X}^T \boldsymbol{W}\hat{\boldsymbol{\Gamma}}\boldsymbol{X})^{-1} \boldsymbol{i}_k = \frac{1}{(1 - R^2_{W\Gamma_{(k)}})\boldsymbol{x}_k^T \boldsymbol{W}\hat{\boldsymbol{\Gamma}}\boldsymbol{x}_k} \tag{5.10}$$

where $R^2_{W\Gamma_{(k)}} = \frac{\hat{\boldsymbol{\beta}}^T_{W\Gamma(k)} \boldsymbol{X}^T \boldsymbol{W}\hat{\boldsymbol{\Gamma}}\boldsymbol{X}\hat{\boldsymbol{\beta}}_{W\Gamma(k)}}{\boldsymbol{x}_k^T \boldsymbol{W}\hat{\boldsymbol{\Gamma}}\boldsymbol{x}_k}$ with $\hat{\beta}_{W\Gamma_{(k)}} = (\boldsymbol{X}^T_{(k)} \boldsymbol{W}\hat{\boldsymbol{\Gamma}}\boldsymbol{X}_{(k)})^{-1} \boldsymbol{X}^T_{(k)} \boldsymbol{W}\hat{\boldsymbol{\Gamma}}\boldsymbol{x}_k$, which is the coefficient of determination in this regression and ranges from 0 to 1.

When there is only $\boldsymbol{x}_k$ in the regression, according to (5.8), the asymptotic variance of $\hat{\beta}_{SWk}$ is:

$$avar_M(\hat{\beta}_{SWk}) = \hat{\tau}^2 (\boldsymbol{x}_k^T \boldsymbol{W}\hat{\boldsymbol{\Gamma}}\boldsymbol{x}_k)^{-1}. \tag{5.11}$$

Comparing (5.10) and (5.11), we can see that the asymptotic model-based variance of $\hat{\beta}_{SWk}$ is inflated

$$\frac{1}{1 - R^2_{W\Gamma_{(k)}}} \tag{5.12}$$

times in (5.10) by taking all the other $p-1$ explanatory variables into the regression with $\boldsymbol{x}_k$. This variance inflation factor is always larger than or equal to 1, because of the range of $R^2_{W\Gamma_{(k)}}$.

Under the model-based inference, the variance of y_i in GLMs is proportional to $v(\mu_i)$. For instance, the variance of y_i in a logistic model is $p_i(1-p_i)$ under the binomial distribution assumption. This assumption is violated when the data are clustered in the population, which results in an *over-dispersion*, i.e., the variance of y_i exceeds variance $v(\mu_i)$. A typical way to handle clustering under model-based inference is using generalized linear mixed models (GLMMs) by treating clusters as random effects. Agresti (2002, chap. 12) and Shao (2003, chap. 4) discussed more details about this approach. In this dissertation, we only treat the general GLMs and will develop appropriate diagnostics for GLMMs in our future research.

5.2.2 Design-based VIF for Three Typical Sampling Designs

5.2.2.1 Linearization Variance Estimator for $\hat{\boldsymbol{\beta}}_{SW}$

Under the design-based inference, we can use the linearization variance estimator for $\hat{\boldsymbol{\beta}}_{SW}$ via the "Binder" method (Binder, 1983). The estimating equations for $\hat{\boldsymbol{\beta}}_{SW}$ in generalized linear models are obtained by setting $\frac{\partial \hat{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ equal to zero. Given by (5.2), they can be expressed as:

$$\begin{aligned} \frac{\partial \hat{l}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{1}{\tau^2} \sum_{i \in s} w_i (y_i - \mu_i) \hat{\gamma}_i g'(\mu_i) \dot{\boldsymbol{x}}_i \\ &= \mathbf{0}^T. \end{aligned} \tag{5.13}$$

Let $\boldsymbol{z}_i = (y_i - \mu_i)\gamma_i g'(\mu_i)\dot{\boldsymbol{x}}_i = e_i \gamma_i g'(\mu_i)\dot{\boldsymbol{x}}_i$, where $e_i = y_i - \mu_i$ is the residual for unit i, and $\hat{\boldsymbol{z}}_i^* = w_i \boldsymbol{z}_i$, $\hat{\boldsymbol{z}}^* = \sum_{i \in s} \boldsymbol{z}_i^* = \frac{1}{\hat{\tau}^2} \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{\Gamma} \boldsymbol{\Delta} (\boldsymbol{y} - \boldsymbol{\mu})$ with $\boldsymbol{\mu} = (\mu_i)_{n \times 1}$ and $\hat{\boldsymbol{\Delta}} = diag[g'(\mu_i)]$. The estimating equations in (5.13) can then be briefly written as $\boldsymbol{z}^* = \mathbf{0}$. Note that to evaluate $\hat{\boldsymbol{z}}_i^*$, sample estimates of the μ_i must be used.

Using the "Binder" method, the linearized estimated variance-covariance matrix for $\hat{\boldsymbol{\beta}}_{SW}$ under design-based inference has the form :

$$var_L(\hat{\boldsymbol{\beta}}_{SW}) = \hat{\boldsymbol{J}}(\hat{\boldsymbol{\beta}}_{SW})^{-1} var_\pi(\hat{\boldsymbol{z}}^*) \hat{\boldsymbol{J}}(\hat{\boldsymbol{\beta}}_{SW})^{-1} \tag{5.14}$$

where $\hat{\boldsymbol{J}}(\hat{\boldsymbol{\beta}}_{SW})$ is the partial derivative (Jacobian matrix) of the estimating equations for $\hat{\boldsymbol{\beta}}_{SW}$, evaluated at the sample estimate of $\boldsymbol{\beta}$, that is $\frac{\partial^2 \hat{l}(\boldsymbol{\beta}_{SW})}{\partial \boldsymbol{\beta}_{SW} \partial \boldsymbol{\beta}_{SW}^T}$ in (5.4).

As shown in (5.5), the expectation of the Jacobian matrix, $\hat{\boldsymbol{J}}(\boldsymbol{\beta})$, is equal to $-\hat{\boldsymbol{I}}(\boldsymbol{\beta})$ in large samples, and thus, under the design-based inference, the estimated asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{SW}$ can be given as:

$$\begin{aligned} avar_L(\hat{\boldsymbol{\beta}}_{SW}) &= \hat{\tau}^4[\boldsymbol{X}^T\boldsymbol{W}\hat{\boldsymbol{\Gamma}}\boldsymbol{X}]^{-1} var_\pi(\hat{\boldsymbol{z}}^*)[\boldsymbol{X}^T\boldsymbol{W}\hat{\boldsymbol{\Gamma}}\boldsymbol{X}]^{-1} \\ &= \hat{\tau}^4 \boldsymbol{A}^{-1} var_\pi(\hat{\boldsymbol{z}}^*)\boldsymbol{A}^{-1}. \end{aligned} \tag{5.15}$$

When a design-unbiased or consistent estimate of $var_\pi(\hat{\boldsymbol{z}}^*)$ is substituted in (5.15), the linearization variance estimate for $\hat{\boldsymbol{\beta}}_{SW}$ will also approximately design-unbiased and consistent. To derive the design-based VIF, We need to integrate our diagnostics method with a particular sampling design, because the estimation of the design-based variance estimator $var_\pi(\hat{\boldsymbol{z}}^*)$ is related to the sampling design. Here, we will discuss three typical sampling designs in practice, which are unequal-weighted single stage sampling design, multistage sampling design and stratified sampling design. First, we will present the estimated $var_\pi(\hat{\boldsymbol{z}}^*)$ in these three designs respectively.

A. Unequal-weighted Single Stage Sampling Design

In an unequal-weighted single-stage design, in which units are selected with replacement, the design consistent variance estimator for $\boldsymbol{z}^*$ is:

$$\begin{aligned}
var_\pi(\hat{\boldsymbol{z}}^*) &= \frac{n}{n-1}\sum_{i\in s}(\boldsymbol{z}_i^* - \bar{\boldsymbol{z}}^*)(\boldsymbol{z}_i^* - \bar{\boldsymbol{z}}^*)^T \\
&= \frac{n}{n-1}\sum_{i\in s}\boldsymbol{z}_i^*\boldsymbol{z}_i^{*T} \\
&= \frac{n}{n-1}\sum_{i\in s}\dot{\boldsymbol{x}}_i w_i\hat{\gamma}_i g'(\mu_i)e_i^2 g'(\mu_i)\hat{\gamma}_i w_i\dot{\boldsymbol{x}}_i^T \\
&= \frac{n}{n-1}\boldsymbol{X}^T\boldsymbol{W}\hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\Delta}}diag(e_i^2)\hat{\boldsymbol{\Delta}}\hat{\boldsymbol{\Gamma}}\boldsymbol{W}\boldsymbol{X},
\end{aligned} \tag{5.16}$$

employing the fact that $\bar{\boldsymbol{z}}^* = \frac{1}{n}\sum_{i=1}^n \boldsymbol{z}_i^* = \boldsymbol{0}^T$, based on the estimating equations in (5.13).

Substituting this result for $var_\pi(\hat{\boldsymbol{z}}^*)$ in (5.15), the linearization variance estimator $avar_L(\hat{\boldsymbol{\beta}}_{SW})$ is:

$$avar_L(\hat{\boldsymbol{\beta}}_{SW}) = \hat{\tau}^4\frac{n}{n-1}\boldsymbol{A}^{-1}\boldsymbol{X}^T\boldsymbol{W}\hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\Delta}}diag(e_i^2)\hat{\boldsymbol{\Delta}}\hat{\boldsymbol{\Gamma}}\boldsymbol{W}\boldsymbol{X}\boldsymbol{A}^{-1}. \tag{5.17}$$

B. Multistage Sampling Design

For a multi-stage sampling design, suppose that it is a two-stage sampling design, units in different clusters are independent in a model under the model-based inference and the first-stage sample units are selected with replacement. In the clustered sample, suppose n clusters are selected out of N clusters and m_i units are selected out of M_i units in the selected cluster i. Under the design-based inference, the estimating equations in a sample are:

$$\begin{aligned}
&\sum_{i\in s}\sum_{t\in s_i}(y_{it}-\mu_{it})\hat{\gamma}_{it}g'(\mu_{it})\dot{\boldsymbol{x}}_{it}=0, \quad k=1,\dots,p;\\
&\sum_{i\in s}\boldsymbol{X}_i^T\boldsymbol{W}_i\hat{\boldsymbol{\Gamma}}_i\hat{\boldsymbol{\Delta}}_i(\boldsymbol{y}_i-\boldsymbol{\mu}_i)=\boldsymbol{0};
\end{aligned} \tag{5.18}$$

where s is the set of sampled clusters, s_i is the set of sampled units in selected i^{th} cluster, $\boldsymbol{y}_i=(y_{i1},...,y_{im_i})^T$, $\boldsymbol{\mu}_i=(\mu_{i1},...,\mu_{im_i})^T$, $\boldsymbol{W}_i=diag(w_{i1},...,w_{im_i})$, $\hat{\boldsymbol{\Gamma}}_i=diag(\gamma_{i1},...,\gamma_{im_i})$, $\hat{\boldsymbol{\Delta}}_i=diag(g'(\mu_{i1}),...,g'(\mu_{im_i}))$ and $\boldsymbol{X}_i(m_i\times p)=(\boldsymbol{x}_{1i},...,\boldsymbol{x}_{pi})$ with $\boldsymbol{x}_{ki}=(x_{ki1},...,x_{kim_i})^T$.

Let $\boldsymbol{z}_i=\boldsymbol{X}_i^T\hat{\boldsymbol{\Gamma}}_i\hat{\boldsymbol{\Delta}}_i(\boldsymbol{y}_i-\boldsymbol{\mu}_i)$. Define a weighted vector of residuals for cluster i as, $\boldsymbol{z}_i^*=\boldsymbol{X}_i^T\boldsymbol{W}_i\hat{\boldsymbol{\Gamma}}_i\hat{\boldsymbol{\Delta}}_i(\boldsymbol{y}_i-\boldsymbol{\mu}_i)$. The design consistent variance estimator for $\boldsymbol{z}^*$ can be given as:

$$\begin{aligned}
var_\pi(\hat{\boldsymbol{z}}^*)&=\frac{n}{n-1}\sum_{i=1}^n(\boldsymbol{z}_i^*-\bar{\boldsymbol{z}}^*)(\boldsymbol{z}_i^*-\bar{\boldsymbol{z}}^*)^T\\
&=\frac{n}{n-1}\sum_{i=1}^n\boldsymbol{z}_i^*\boldsymbol{z}_i^{*T}\\
&=\frac{n}{n-1}\boldsymbol{X}^T\boldsymbol{W}_i\hat{\boldsymbol{\Gamma}}_i\hat{\boldsymbol{\Delta}}_i(\boldsymbol{e}_i\boldsymbol{e}_i^T)\hat{\boldsymbol{\Delta}}_i\hat{\boldsymbol{\Gamma}}_i\boldsymbol{W}_i\boldsymbol{X}_i\\
&=\frac{n}{n-1}\boldsymbol{X}^T\boldsymbol{W}\hat{\boldsymbol{\Gamma}}\hat{\boldsymbol{\Delta}}Blkdiag(\boldsymbol{e}_i\boldsymbol{e}_i^T)\hat{\boldsymbol{\Delta}}\hat{\boldsymbol{\Gamma}}\boldsymbol{W}\boldsymbol{X}
\end{aligned} \tag{5.19}$$

where $\boldsymbol{e}_i=\boldsymbol{y}_i-\boldsymbol{\mu}_i$ and $\bar{\boldsymbol{z}}^*=\frac{1}{n}\sum_{i=1}^n\boldsymbol{z}_i^*=\boldsymbol{0}$ because of the property of estimating equations (5.18).

Substituting (5.19) into (5.15), the linearization variance estimator $avar_L(\hat{\boldsymbol{\beta}}_{SW})$

is:

$$avar_L(\hat{\boldsymbol{\beta}}_{SW}) = \hat{\tau}^4 \frac{n}{n-1} \boldsymbol{A}^{-1} \boldsymbol{X}^T \boldsymbol{W} \hat{\boldsymbol{\Gamma}} \hat{\boldsymbol{\Delta}} Blkdiag(\boldsymbol{e}_i \boldsymbol{e}_i^T) \hat{\boldsymbol{\Delta}} \hat{\boldsymbol{\Gamma}} \boldsymbol{W} \boldsymbol{X} \boldsymbol{A}^{-1}. \tag{5.20}$$

C. Stratified Multistage Sampling Design

In a stratified multistage sampling design, suppose that there are $h = 1, ..., H$ strata in the population, we select $i = 1, ..., n_h$ clusters in stratum h and $t = 1, ..., m_{hi}$ units in cluster hi. Clusters are assumed to be selected with replacement within strata and independently between strata. We will consider two linear models: one assumes that there are common intercept and slopes across strata; while another assumes that there are different linear models, or different parameters in each stratum.

In the first model, under the design-based inference, the estimating equations in a sample are:

$$\begin{aligned} &\sum_{h=1}^{H} \sum_{i \in s_h} \sum_{t \in s_{hi}} (y_{hit} - \mu_{hit}) \hat{\gamma}_{hit} g'(\mu_{hit}) \dot{\boldsymbol{x}}_{hit} = 0, \quad k = 1, \ldots, p; \\ &\sum_{h=1}^{H} \sum_{i \in s_h} \boldsymbol{X}_{hi}^T \boldsymbol{W}_{hi} \hat{\boldsymbol{\Gamma}}_{hi} \hat{\boldsymbol{\Delta}}_{hi} (\boldsymbol{y}_{hi} - \boldsymbol{\mu}_{hi}) = \boldsymbol{0}; \end{aligned} \tag{5.21}$$

where s_h is the set of sampled clusters in stratum h, s_{hi} is the set of sampled units in selected cluster hi, $\boldsymbol{y}_{hi} = (y_{hi1}, ..., y_{him_{hi}})^T$, $\boldsymbol{\mu}_{hi} = (\mu_{hi1}, ..., \mu_{him_{hi}})^T$, $\boldsymbol{W}_{hi} = diag(w_{hi1}, ..., w_{him_{hi}})$, $\hat{\boldsymbol{\Gamma}}_{hi} = diag(\gamma_{hi1}, ..., \gamma_{him_i})$, $\hat{\boldsymbol{\Delta}}_{hi} = diag(g'(\mu_{hi1}), ..., g'(\mu_{him_i}))$ and $\boldsymbol{X}_{hi}(m_{hi} \times p) = (\boldsymbol{x}_{1hi}, ..., \boldsymbol{x}_{phi})$ with $\boldsymbol{x}_{khi} = (x_{khi1}, ..., x_{khim_{hi}})^T$.

Let $\boldsymbol{z}_{hi} = \boldsymbol{X}_{hi}^T \hat{\boldsymbol{\Gamma}}_{hi} \hat{\boldsymbol{\Delta}}_{hi} (\boldsymbol{y}_{hi} - \boldsymbol{\mu}_{hi})$. Define a weighted vector of residuals for cluster i as, $\boldsymbol{z}_{hi}^* = \boldsymbol{X}_{hi}^T \boldsymbol{W}_{hi} \hat{\boldsymbol{\Gamma}}_{hi} \hat{\boldsymbol{\Delta}}_{hi} (\boldsymbol{y}_{hi} - \boldsymbol{\mu}_{hi})$. The design consistent variance estimator for $\boldsymbol{z}^*$ in this stratified multistage sampling design is:

$$
\begin{aligned}
var_\pi(\hat{\boldsymbol{z}}^*) &= \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i \in s} (\boldsymbol{z}_{hi}^* - \bar{\boldsymbol{z}}_h^*)(\boldsymbol{z}_{hi}^* - \bar{\boldsymbol{z}}_h^*)^T \\
&= \sum_{h=1}^{H} \frac{n_h}{n_h - 1} (\sum_{i \in s} \boldsymbol{z}_{hi}^* \boldsymbol{z}_{hi}^{*T} - n_h \bar{\boldsymbol{z}}_h^* \bar{\boldsymbol{z}}_h^{*T}) \\
&= \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \sum_{i \in s} \boldsymbol{z}_{hi}^* \boldsymbol{z}_{hi}^{*T} - \sum_{h=1}^{H} \frac{n_h^2}{n_h - 1} \bar{\boldsymbol{z}}_h^* \bar{\boldsymbol{z}}_h^{*T} \\
&= \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \boldsymbol{X}_h^T \boldsymbol{W}_h \hat{\boldsymbol{\Gamma}}_h \hat{\boldsymbol{\Delta}}_h \left[Blkdiag(\boldsymbol{e}_{hi} \boldsymbol{e}_{hi}^T) - \frac{1}{n_h} \boldsymbol{e}_h \boldsymbol{e}_h^T \right] \hat{\boldsymbol{\Delta}}_h \hat{\boldsymbol{\Gamma}}_h \boldsymbol{W}_h \boldsymbol{X}_h,
\end{aligned}
\tag{5.22}
$$

where $\boldsymbol{X}_h^T (\sum_{i \in s} m_{hi} \times p) = (\boldsymbol{X}_{h1}, ..., \boldsymbol{X}_{hn_h})^T$, $\boldsymbol{W}_h = diag(\boldsymbol{W}_{hi})_i$, $\hat{\boldsymbol{\Gamma}}_h = diag(\hat{\boldsymbol{\Gamma}}_{hi})_i$, $\hat{\boldsymbol{\Delta}}_h = diag(\hat{\boldsymbol{\Delta}}_{hi})_i$, $\boldsymbol{e}_{hi} = \boldsymbol{y}_{hi} - \boldsymbol{\mu}_{hi}$ and $\boldsymbol{e}_h = (\boldsymbol{e}_{h1}, \boldsymbol{e}_{h2}, ..., \boldsymbol{e}_{hn_h})^T$ is a vector of unit residuals in stratum h.

Correspondingly, the linearization variance estimator $avar_L(\hat{\boldsymbol{\beta}}_{SW})$ is:

$$
\begin{aligned}
avar_L(\hat{\boldsymbol{\beta}}_{SW}) = \hat{\tau}^4 \sum_{h=1}^{H} &\boldsymbol{A}^{-1} \frac{n_h}{n_h - 1} \boldsymbol{X}_h^T \boldsymbol{W}_h \hat{\boldsymbol{\Gamma}}_h \hat{\boldsymbol{\Delta}}_h \left[Blkdiag(\boldsymbol{e}_{hi} \boldsymbol{e}_{hi}^T) - \frac{1}{n_h} \boldsymbol{e}_h \boldsymbol{e}_h^T \right] \times \\
&\hat{\boldsymbol{\Delta}}_h \hat{\boldsymbol{\Gamma}}_h \boldsymbol{W}_h \boldsymbol{X}_h \boldsymbol{A}^{-1} \\
= \hat{\tau}^4 &\boldsymbol{A}^{-1} \boldsymbol{X}^T \boldsymbol{W} \hat{\boldsymbol{\Gamma}} \hat{\boldsymbol{\Delta}} Blkdiag \left\{ \frac{n_h}{n_h - 1} \left[Blkdiag(\boldsymbol{e}_{hi} \boldsymbol{e}_{hi}^T) - \frac{1}{n_h} \boldsymbol{e}_h \boldsymbol{e}_h^T \right] \right\} \times \\
&\hat{\boldsymbol{\Delta}} \hat{\boldsymbol{\Gamma}} \boldsymbol{W} \boldsymbol{X} \boldsymbol{A}^{-1}.
\end{aligned}
\tag{5.23}
$$

When n_h is large,

$$var_\pi(\hat{\boldsymbol{z}}^*) = \sum_{h=1}^{H} \frac{n_h}{n_h-1} \boldsymbol{X}_h^T \boldsymbol{W}_h \hat{\boldsymbol{\Gamma}}_h \hat{\boldsymbol{\Delta}}_h Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T) \hat{\boldsymbol{\Delta}}_h \hat{\boldsymbol{\Gamma}}_h \boldsymbol{W}_h \boldsymbol{X}_h, \tag{5.24}$$

and

$$avar_L(\hat{\boldsymbol{\beta}}_{SW}) = \hat{\tau}^4 \sum_{h=1}^{H} \boldsymbol{A}^{-1} \frac{n_h}{n_h-1} \boldsymbol{X}_h^T \boldsymbol{W}_h \hat{\boldsymbol{\Gamma}}_h \hat{\boldsymbol{\Delta}}_h Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T) \hat{\boldsymbol{\Delta}}_h \hat{\boldsymbol{\Gamma}}_h \boldsymbol{W}_h \boldsymbol{X}_h \boldsymbol{A}^{-1}. \tag{5.25}$$

If we incorporate the stratification in the model and assume different linear models, or different slope parameters, $\boldsymbol{\beta}_h$, in each stratum. Within each stratum, the estimation of regression parameters, their variances and corresponding VIF values is the same as that for the multistage sampling design described in 5.2.2.1.B above. Collinearity diagnostics will be conducted independently within each stratum for this setting.

5.2.2.2 VIF for $var_L(\hat{\boldsymbol{\beta}}_{SW})$

Note that if we let $\hat{\boldsymbol{V}} = \frac{n}{n-1} diag(e_i^2)$ in the unequal weighted single stage sampling design or $\hat{\boldsymbol{V}} = \frac{n}{n-1} Blkdiag(\boldsymbol{e}_i\boldsymbol{e}_i^T)$ in the multistage sampling design or $\hat{\boldsymbol{V}} = Blkdiag\left\{\frac{n_h}{n_h-1}\left[Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T) - \frac{1}{n_h}\boldsymbol{e}_h\boldsymbol{e}_h^T\right]\right\}$ in the stratified multistage sampling design, the linearization variance estimator $avar_L(\hat{\boldsymbol{\beta}}_{SW})$ can be rewritten as:

$$avar_L(\hat{\boldsymbol{\beta}}_{SW}) = \hat{\tau}^4 \boldsymbol{A}^{-1} \boldsymbol{X}^T \boldsymbol{W} \hat{\boldsymbol{\Gamma}} \hat{\boldsymbol{\Delta}} \hat{\boldsymbol{V}} \hat{\boldsymbol{\Delta}} \hat{\boldsymbol{\Gamma}} \boldsymbol{W} \boldsymbol{X} \boldsymbol{A}^{-1}. \tag{5.26}$$

Denote $\boldsymbol{V}_\Delta = \hat{\boldsymbol{\Delta}}\hat{\boldsymbol{V}}\hat{\boldsymbol{\Delta}}$, $\boldsymbol{W}_\Gamma = \boldsymbol{W}\hat{\boldsymbol{\Gamma}} = diag(w_i\hat{\gamma}_i)$, $\boldsymbol{B} = \boldsymbol{X}^T\boldsymbol{W}_\Gamma\boldsymbol{V}_\Delta\boldsymbol{W}_\Gamma\boldsymbol{X}$, $\boldsymbol{G} = \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}$ and we have $avar_L(\hat{\boldsymbol{\beta}}_{SW}) = \hat{\tau}^4\boldsymbol{G}$.

The linearization variance estimator for $\hat{\beta}_k$ is the k^{th} element on the main diagonal of $avar_L(\hat{\boldsymbol{\beta}}_{SW})$, so it is directly related to the k^{th} element on the main diagonal of matrix $\boldsymbol{G}$ as:

$$avar_L(\hat{\beta}_{SWk}) = \hat{\tau}^4 g_{kk}, \tag{5.27}$$

upon defining $\boldsymbol{G} = (g_{ij})$.

Partitioning the inverse of matrix $\boldsymbol{A}$, the lower right component of $\boldsymbol{A}^{-1}$ is:

$$\boldsymbol{a}^{(k)k} = -a^{kk}(\boldsymbol{X}_{(k)}^T\boldsymbol{W}_\Gamma\boldsymbol{X}_{(k)})^{-1}\boldsymbol{X}_{(k)}\boldsymbol{W}_\Gamma\boldsymbol{x}_k = -a^{kk}\hat{\boldsymbol{\beta}}_{W\Gamma(k)}. \tag{5.28}$$

A partitioned version of $\boldsymbol{B}$ can be expressed as

$$\boldsymbol{B} = \begin{pmatrix} b_{kk} & \boldsymbol{b}_{k(k)} \\ \boldsymbol{b}_{(k)k} & \boldsymbol{B}_{(k)(k)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_k^T\boldsymbol{W}_\Gamma\boldsymbol{V}_\Delta\boldsymbol{W}_\Gamma\boldsymbol{x}_k & \boldsymbol{x}_k^T\boldsymbol{W}_\Gamma\boldsymbol{V}_\Delta\boldsymbol{W}_\Gamma\boldsymbol{X}_{(k)} \\ \boldsymbol{X}_{(k)}^T\boldsymbol{W}_\Gamma\boldsymbol{V}_\Delta\boldsymbol{W}_\Gamma\boldsymbol{x}_k & \boldsymbol{X}_{(k)}^T\boldsymbol{W}_\Gamma\boldsymbol{V}_\Delta\boldsymbol{W}_\Gamma\boldsymbol{X}_{(k)} \end{pmatrix}. \tag{5.29}$$

Using (5.10), (5.28) and steps analogous to the derivation of VIF in linear regression in Chapter 3, g_{kk} can be decomposed into a term with $R^2_{W\Gamma(k)}$ and two adjustment coefficients:

$$
\begin{aligned}
g^{kk} &= a^{kk}(a^{kk}b_{kk} + 2\boldsymbol{b}_{k(k)}\boldsymbol{a}^{(k)k}) + \boldsymbol{a}^{(k)k^T}\boldsymbol{B}_{(k)(k)}\boldsymbol{a}^{(k)k} \\
&= (a^{kk})^2(b_{kk} - 2\boldsymbol{b}_{k(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)} + \hat{\boldsymbol{\beta}}^T_{W\Gamma(k)}\boldsymbol{B}_{(k)(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)}) \\
&= \left(\frac{1}{1-R^2_{W\Gamma(k)}}\frac{1}{\boldsymbol{x}_k^T\boldsymbol{W}_\Gamma\boldsymbol{x}_k}\right)^2 \times \\
&\quad (\boldsymbol{x}_k^T\boldsymbol{W}_\Gamma\boldsymbol{V}_\Delta\boldsymbol{W}_\Gamma\boldsymbol{x}_k - 2\boldsymbol{x}_k^T\boldsymbol{W}_\Gamma\boldsymbol{V}_\Delta\boldsymbol{W}_\Gamma\boldsymbol{x}_{(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)} + \hat{\boldsymbol{\beta}}^T_{W\Gamma(k)}\boldsymbol{X}_{(k)}\boldsymbol{W}_\Gamma\boldsymbol{V}_\Delta\boldsymbol{W}_\Gamma\boldsymbol{x}_{(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)}) \\
&= \frac{1}{1-R^2_{W\Gamma(k)}}\frac{(\boldsymbol{x}_k-\boldsymbol{X}_{(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)})^T\boldsymbol{W}_\Gamma\boldsymbol{V}_\Delta\boldsymbol{W}_\Gamma(\boldsymbol{x}_k-\boldsymbol{X}_{(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)})}{(\boldsymbol{x}_k-\boldsymbol{X}_{(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)})^T\boldsymbol{W}_\Gamma(\boldsymbol{x}_k-\boldsymbol{X}_{(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)})}\frac{1}{\boldsymbol{x}_k^T\boldsymbol{W}_\Gamma\boldsymbol{x}_k} \\
&= \frac{1}{1-R^2_{W\Gamma(k)}}\frac{\boldsymbol{e}_{xk}^T\boldsymbol{W}_\Gamma\boldsymbol{V}_\Delta\boldsymbol{W}_\Gamma\boldsymbol{e}_{xk}}{\boldsymbol{e}_{xk}^T\boldsymbol{W}_\Gamma\boldsymbol{e}_{xk}}\frac{\boldsymbol{x}_k^T\boldsymbol{W}_\Gamma\boldsymbol{x}_k}{\boldsymbol{x}_k^T\boldsymbol{W}_\Gamma\boldsymbol{V}_\Delta\boldsymbol{W}_\Gamma\boldsymbol{x}_k}\frac{\boldsymbol{x}_k^T\boldsymbol{W}_\Gamma\boldsymbol{V}_\Delta\boldsymbol{W}_\Gamma\boldsymbol{x}_k}{(\boldsymbol{x}_k^T\boldsymbol{W}_\Gamma\boldsymbol{x}_k)^2} \\
&= \frac{\hat{\zeta}_k\hat{\varrho}_k}{1-R^2_{W\Gamma(k)}}\frac{\boldsymbol{x}_k^T\boldsymbol{W}_\Gamma\boldsymbol{V}_\Delta\boldsymbol{W}_\Gamma\boldsymbol{x}_k}{(\boldsymbol{x}_k^T\boldsymbol{W}_\Gamma\boldsymbol{x}_k)^2}
\end{aligned}
\tag{5.30}
$$

where $\boldsymbol{e}_{xk} = \boldsymbol{x}_k - \boldsymbol{X}_{(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)}$ is the residual from WLS regressing $\boldsymbol{x}_k$ on $\boldsymbol{X}_{(k)}$ using weight matrix $\boldsymbol{W}_\Gamma$, $\hat{\zeta}_k = \frac{(\boldsymbol{x}_k-\boldsymbol{X}_{(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)})^T\boldsymbol{W}_\Gamma\boldsymbol{V}_\Delta\boldsymbol{W}_\Gamma(\boldsymbol{x}_k-\boldsymbol{X}_{(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)})}{(\boldsymbol{x}_k-\boldsymbol{X}_{(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)})^T\boldsymbol{W}_\Gamma(\boldsymbol{x}_k-\boldsymbol{X}_{(k)}\hat{\boldsymbol{\beta}}_{W\Gamma(k)})} = \frac{\boldsymbol{e}_{xk}^T\boldsymbol{W}_\Gamma\boldsymbol{V}_\Delta\boldsymbol{W}_\Gamma\boldsymbol{e}_{xk}}{\boldsymbol{e}_{xk}^T\boldsymbol{W}_\Gamma\boldsymbol{e}_{xk}}$, and $\hat{\varrho}_k = \frac{\boldsymbol{x}_k^T\boldsymbol{W}_\Gamma\boldsymbol{x}_k}{\boldsymbol{x}_k^T\boldsymbol{W}_\Gamma\boldsymbol{V}_\Delta\boldsymbol{W}_\Gamma\boldsymbol{x}_k}$. Refer to (3.13) in linear VIF chapter (Chapter 3, Section 3.1.4), which is the direct analog. Note that if we assume $\boldsymbol{V}_\Delta = \boldsymbol{W}_\Gamma^{-1}$, the VIF for design-based approach will be similar to the one used for model-based approach as mentioned in Section 5.2.1, $\frac{1}{1-R^2_{W\Gamma(k)}}$. But this assumption is always untrue in reality.

Replacing g^{kk} in (5.27) with (5.30), the linearization variance estimator for the estimated coefficient of the k^{th} explanatory variables in the full model with all the explanatory variables is:

$$avar_L(\hat{\beta}_{SWk}) = \hat{\tau}^4 g_{kk} = \hat{\tau}^4 \frac{\hat{\zeta}_k \hat{\varrho}_k}{1 - R^2_{W\Gamma(k)}} \frac{\boldsymbol{x}_k^T \boldsymbol{W}_\Gamma \boldsymbol{V}_\Delta \boldsymbol{W}_\Gamma \boldsymbol{x}_k}{(\boldsymbol{x}_k^T \boldsymbol{W}_\Gamma \boldsymbol{x}_k)^2}. \tag{5.31}$$

Consider a model with only $\boldsymbol{x}_k$, using (5.15), the linearization variance estimator for $\hat{\beta}_{SWk}$ is:

$$var_{0L}(\hat{\beta}_{SWk}) = \hat{\tau}^4 \frac{\boldsymbol{x}_k^T \boldsymbol{W}_\Gamma \boldsymbol{V}_\Delta \boldsymbol{W}_\Gamma \boldsymbol{x}_k}{(\boldsymbol{x}_k^T \boldsymbol{W}_\Gamma \boldsymbol{x}_k)^2}. \tag{5.32}$$

Hence, $var_{0L}(\hat{\beta}_{SWk})$ is inflated by

$$\frac{\hat{\zeta}_k \hat{\varrho}_k}{1 - R^2_{W\Gamma(k)}} \tag{5.33}$$

times when taking other explanatory variables in the model, comparing $var_{0L}(\hat{\beta}_{SWk})$ in (5.32) to $avar_L(\hat{\beta}_{SWk})$ in (5.31). This assumes that τ is the same in the model that includes only $\boldsymbol{x}_k$ and the one that includes all $\boldsymbol{x}$'s. The term $\frac{\hat{\zeta}_k \hat{\varrho}_k}{1 - R^2_{W\Gamma(k)}}$ is called the VIF for the k^{th} explanatory variable in the model.

Furthermore, when there is a model with only $\boldsymbol{x}_k$ and an intercept, using (5.30), the linearization variance estimator for $\hat{\beta}_{SWk}$ can be written as:

$$var_{1L}(\hat{\beta}_{SWk}) = \hat{\tau}^4 \frac{(\boldsymbol{x}_k - \mathbf{1}\bar{\bar{x}}_k)^T \boldsymbol{W}_\Gamma \boldsymbol{V}_\Delta \boldsymbol{W}_\Gamma (\boldsymbol{x}_k - \mathbf{1}\bar{\bar{x}}_k)}{(\boldsymbol{x}_k^T \boldsymbol{W}_\Gamma \boldsymbol{x}_k - \hat{N}\bar{\bar{x}}_k^2)^2} \tag{5.34}$$

where $\hat{N} = \sum_{i \in s} w_i \hat{\gamma}_i$ and $\bar{\bar{x}}_k = \sum_{i \in s} w_i \hat{\gamma}_i x_{ki} / \hat{N}$.

The variance $avar_L(\hat{\beta}_{SWk})$ in (5.31) can be decomposed into several factors correspondingly:

$$
\begin{aligned}
avar_L(\hat{\beta}_{SWk}) &= g^{kk} \\
&= \frac{\hat{\zeta}_k}{1 - R^2_{W\Gamma m(k)}} \frac{\boldsymbol{x}_k^T \boldsymbol{W}_\Gamma \boldsymbol{x}_k - \hat{N}\bar{\bar{x}}_k^2}{(\boldsymbol{x}_k - \mathbf{1}\bar{\bar{x}}_k)^T \boldsymbol{W}_\Gamma \boldsymbol{V}_\Delta \boldsymbol{W}_\Gamma (\boldsymbol{x}_k - \mathbf{1}\bar{\bar{x}}_k)} \\
&\quad \frac{(\boldsymbol{x}_k - \mathbf{1}\bar{\bar{x}}_k)^T \boldsymbol{W}_\Gamma \boldsymbol{V}_\Delta \boldsymbol{W}_\Gamma (\boldsymbol{x}_k - \mathbf{1}\bar{\bar{x}}_k)}{(\boldsymbol{x}_k^T \boldsymbol{W}_\Gamma \boldsymbol{x}_k - \hat{N}\bar{\bar{x}}_k^2)^2} \\
&= \frac{\hat{\zeta}_k \hat{\varrho}_{km}}{1 - R^2_{W\Gamma m(k)}} var_{1L}(\hat{\beta}_{SWk})
\end{aligned}
\tag{5.35}
$$

where $R^2_{W\Gamma m(k)} = \frac{\hat{\boldsymbol{\beta}}^T_{W\Gamma(k)} \boldsymbol{X}^T \boldsymbol{W}_\Gamma \boldsymbol{X} \hat{\boldsymbol{\beta}}_{W\Gamma(k)} - \hat{N}\bar{\bar{x}}_k^2}{\boldsymbol{x}_k^T \boldsymbol{W}_\Gamma \boldsymbol{x}_k - \hat{N}\bar{\bar{x}}_k^2}$ is the coefficient of determination corrected for the mean in the WLS regression when weight matrix is $\boldsymbol{W}_\Gamma$, and $\hat{\varrho}_{km} = \frac{\boldsymbol{x}_k^T \boldsymbol{W}_\Gamma \boldsymbol{x}_k - \hat{N}\bar{\bar{x}}_k^2}{(\boldsymbol{x}_k - \mathbf{1}\bar{\bar{x}}_k)^T \boldsymbol{W}_\Gamma \boldsymbol{V}_\Delta \boldsymbol{W}_\Gamma (\boldsymbol{x}_k - \mathbf{1}\bar{\bar{x}}_k)}$.

The term

$$
\frac{\hat{\zeta}_k \hat{\varrho}_{km}}{1 - R^2_{W\Gamma m(k)}} \tag{5.36}
$$

is called the intercept-adjusted VIF for the k^{th} explanatory variable in the model.

The design-based approach for fitting GLMs as we discussed above is one of the typical methods by using the survey weights directly in the estimation equations. Some alternative design-based approaches have also been developed for the fitting of GLM under information sampling. Other than the pseudo-likelihood approach, Pfeffermann & Sverchkov (2003) considered another two approaches, which require the modeling and estimation of the expectation of the sampling weights, either as a function of the outcome and the explanatory variables, or as a function of only the explanatory variables. The expectation of the sampling weights will be used as the weights in the estimation equations and thus when these two approaches are used, we can simply treat these weights as the $\boldsymbol{W}$ matrix in (5.33) or (5.36) to obtain the VIFs in these estimations.

Some common univariate distributions in the exponential family and their special link functions are listed in Table 5.1. Substituting their corresponding elements of matrix $\boldsymbol{\Delta}$ and $\boldsymbol{\Gamma}$ in (5.33) and (5.36), we can obtain their VIF formulas respectively. To demonstrate the techniques in a special GLM model, we will take logistic model as an example in the following section.

Table 5.1: Characteristics of Some Common Univariate Distributions in the Exponential Family

	Normal	Poisson	Binomial		Gamma	Inverse Gaussian
Notation	$N(\mu_i, \sigma^2)$	$P(\mu_i)$	$B(m, \mu_i)/m$		$G(\mu_i, v)$	$IG(\mu_i, \sigma^2)$
Range of y	$(-\infty, \infty)$	$(0(1), \infty)$	$\frac{(0(1),m)}{m}$		$(0, \infty)$	$(0, \infty)$
τ^2	σ^2	1	$1/m$		v^{-1}	σ^2
Link Function	identity	log	logit	probit	reciprocal	μ_i^{-2}
Canonical link $g(\mu_i)$	μ_i	$log(\mu_i)$	$log(\frac{\mu_i}{1-\mu_i})$	${}^a\Phi^{-1}(\mu_i)$	μ_i^{-1}	μ_i^{-2}
Variance function $v(\mu_i)$	1	μ_i	$\mu_i(1-\mu_i)$	$\mu_i(1-\mu_i)$	μ_i^2	μ_i^3
${}^b g'(\mu_i)$	1	μ_i^{-1}	$[\mu_i(1-\mu_i)]^{-1}$	${}^c\varphi^{-1}(\dot{\boldsymbol{x}}_i^T\boldsymbol{\beta})$	$-\mu_i^{-2}$	$-2\mu_i^{-3}$
${}^d\gamma_i = \{v(\mu_i)[g'(\mu_i)]^2\}^{-1}$	1	μ_i	$\mu_i(1-\mu_i)$	$\frac{\varphi^2(\dot{\boldsymbol{x}}_i^T\boldsymbol{\beta})}{\mu_i(1-\mu_i)}$	μ_i^2	$\frac{1}{4}\mu_i^3$
${}^e\gamma_i g'(\mu_i)$	1	1	1	$\frac{\varphi(\dot{\boldsymbol{x}}_i^T\boldsymbol{\beta})}{\mu_i(1-\mu_i)}$	-1	$-\frac{1}{2}$

${}^a\Phi$ is the cdf of the standard normal distribution
${}^b\boldsymbol{\Delta} = diag[g'(\mu_i)]$
${}^c\varphi$=Φ' is the pdf of the standard normal distribution and $\dot{\boldsymbol{x}}_i^T\boldsymbol{\beta} = g(\mu_i)$
${}^d\boldsymbol{\Gamma} = diag(\gamma_i)$
${}^e\boldsymbol{\Gamma\Delta} = diag[\gamma_i g'(\mu_i)]$

5.2.3 Logistic Model

5.2.3.1 introduction

Logistic regression (sometimes called the logistic model or logit model) is a generalized linear model used to model binary (or, more generally, multi-category class or ordinal) variables as a function of predictor variables that are categorical, continuous, or both. It is used extensively in the medical and social sciences as well as marketing applications or other business analysis fields. Here, I will use it as an example to illustrate the collinearity diagnostics for generalized linear models in the analysis of complex survey data.

Let p_i denote the probability that unit i has a certain characteristic (binary). Let $y_i = 1$, if sample unit i has the characteristic of interest, and $y_i = 0$, if not. Then, under the model, $P(y_i = 1) = p_i$. In logistic regression, we try to model the log odds (also called logit) of p_i's as linear function of predictor variables, just as in the linear regression:

$$log\left(\frac{p_i}{1-p_i}\right) = \dot{\boldsymbol{x}}_i^T\boldsymbol{\beta}$$

where $\dot{\boldsymbol{x}}_i = (x_{1i}, ..., x_{pi})^T$, is the vector of covariates for the i^{th} unit, $\boldsymbol{\beta}_{p\times 1}$ is the column vector of the parameters of interest and $log\left[p_i/(p_i-1)\right]$ is the log odds of p_i.

The distribution of y_i in the logistic regression is a Bernoulli or binary distribution and can be expressed as:

$$f(y_i|p_i) = p_i^{y_i}(1-p_i)^{1-y_i}.$$

This is a simple probability distribution from exponential family that can be easily put into exponential form as (2.8). We have

$$f(y_i|p_i) = e^{ylog(\frac{p_i}{1-p_i})+log(1-p_i)} = e^{y\theta - log(1+e^\theta)}, \tag{5.37}$$

where the canonical parameters are $\theta = log\left(\frac{p_i}{1-p_i}\right)$ and $b(\theta) = log(1+e^{\theta})$. In this case, $\mu_i = p_i$, $\tau = 1$ and $c(y_i, \tau) = 0$. The variance function is $var(y_i) = \partial^2 b(\theta_i)/\partial\theta_i^2 \equiv v(\mu_i) = p_i(1-p_i)$. The link function is $g(p_i) = log\left(\frac{p_i}{1-p_i}\right) = \dot{\boldsymbol{x}}_i^T\boldsymbol{\beta}$ and its first-order partial for p_i is $g'(p_i) = \frac{1}{p_i(1-p_i)}$. Hence, γ_i in (5.2) here is $p_i(1-p_i)$, the matrix $\boldsymbol{\Gamma}$ in (5.3) is $diag[p_i(1-p_i)]$ and $\boldsymbol{\Delta}$ in (5.3) is $diag\left[\frac{1}{p_i(1-p_i)}\right]$.

To estimate $\hat{\boldsymbol{\beta}}$ and make inference based on the data obtained from the complex survey design, we use a weighted sample likelihood function:

$$\hat{L}(\hat{\boldsymbol{\beta}}_{SW}) = \prod_{i\in s} p_i^{w_i y_i}(1-p_i)^{w_i(1-y_i)}, \tag{5.38}$$

where w_i is the survey weight for unit i.

Taking the logarithm of (5.38) and substituting

$$log(p_i/(1-p_i)) = log(p_i) - log(1-p_i) = \dot{\boldsymbol{x}}_i^T\hat{\boldsymbol{\beta}}_{SW}$$

gives the log-likelihood function evaluated at the SW estimator as:

$$\hat{l}(\hat{\boldsymbol{\beta}}_{SW}) = log\hat{L}(\hat{\boldsymbol{\beta}}_{SW}) = \sum_{i\in s} w_i\left[y_i\dot{\boldsymbol{x}}_i^T\hat{\boldsymbol{\beta}}_{SW} + log(1-p_i)\right]. \tag{5.39}$$

5.2.3.2 Variance Inflation Factor in Logistic Model

The estimating equations in logistic regression can be given by:

$$\frac{\partial\hat{l}(\hat{\boldsymbol{\beta}}_{SW})}{\partial\hat{\boldsymbol{\beta}}_{SW}} = \sum_{i\in s} w_i(y_i-\hat{p}_i)\dot{\boldsymbol{x}}_i^T = \boldsymbol{X}\boldsymbol{W}(\boldsymbol{y}-\hat{\boldsymbol{p}}) = \boldsymbol{0}^T \tag{5.40}$$

with $\hat{\boldsymbol{p}} = (\hat{p}_i)_{n\times 1}$. Note that this is consistent with the estimating equations for GLM in (5.3) since $\boldsymbol{\Gamma}\boldsymbol{\Delta} = \boldsymbol{I}$.

Model-based VIF

Based on the estimating equations in (5.40), we can directly show that the second derivative of the log likelihood $\hat{l}(\hat{\boldsymbol{\beta}}_{SW})$ is:

$$\frac{\partial^2 \hat{l}(\hat{\boldsymbol{\beta}}_{SW})}{\partial \hat{\boldsymbol{\beta}}_{SW} \partial \hat{\boldsymbol{\beta}}_{SW}^T} = -\sum_{i\in s} w_i \hat{p}_i (1-\hat{p}_i) \dot{\boldsymbol{x}}_i \dot{\boldsymbol{x}}_i^T = -\boldsymbol{X}^T \boldsymbol{W} diag[\hat{p}_i(1-\hat{p}_i)]\boldsymbol{X}, \qquad (5.41)$$

and under the model-based inference, the asymptotic variance-covariance matrix of $\hat{\boldsymbol{\beta}}_{SW}$ is:

$$avar_M(\hat{\boldsymbol{\beta}}_{SW}) = [\boldsymbol{I}(\hat{\boldsymbol{\beta}}_{SW})]^{-1} = -E\left[\frac{\partial^2 \hat{l}(\hat{\boldsymbol{\beta}}_{SW})}{\partial \hat{\boldsymbol{\beta}}_{SW} \partial \hat{\boldsymbol{\beta}}_{SW}^T}\right] = \boldsymbol{X}^T \boldsymbol{W} diag[\hat{p}_i(1-\hat{p}_i)]\boldsymbol{X} \qquad (5.42)$$

where *avar* stands for the limiting or asymptotic variance. This is also consistent with the model-based estimated variance of $\hat{\boldsymbol{\beta}}_{SW}$ in (5.8) since $\hat{\boldsymbol{\Gamma}} = diag[\hat{p}_i(1-\hat{p}_i)]$.

According to the derivation of model-based VIF in GLM (shown in section 5.2.1), the model-based VIF for logistic regression is:

$$\frac{1}{1 - R^2_{W\Gamma_{(k)}}} \qquad (5.43)$$

where $R^2_{W\Gamma_{(k)}} = \frac{\hat{\boldsymbol{\beta}}^T_{W\Gamma(k)} \boldsymbol{X}^T \boldsymbol{W} \hat{\boldsymbol{\Gamma}} \boldsymbol{X} \hat{\boldsymbol{\beta}}_{W\Gamma(k)}}{\boldsymbol{x}_k^T \boldsymbol{W} \hat{\boldsymbol{\Gamma}} \boldsymbol{x}_k}$ with $\hat{\beta}_{W\Gamma_{(k)}} = (\boldsymbol{X}^T_{(k)} \boldsymbol{W} \hat{\boldsymbol{\Gamma}} \boldsymbol{X}_{(k)})^{-1} \boldsymbol{X}^T_{(k)} \boldsymbol{W} \hat{\boldsymbol{\Gamma}} \boldsymbol{x}_k$ and $\hat{\boldsymbol{\Gamma}} = diag[\hat{p}_i(1-\hat{p}_i)]$.

Design-based VIF for Three Typical Sampling Designs

Let $\boldsymbol{z}_i = (y_i - \hat{p}_i)\dot{\boldsymbol{x}}_i^T = e_i\dot{\boldsymbol{x}}_i^T$ and $\boldsymbol{z}_i^* = w_i\boldsymbol{z}_i$, where $e_i = y_i - \hat{p}_i$ is the residual for unit i. Via the "Binder" method (Binder, 1983), the linearized estimated variance-covariance matrix for $\hat{\boldsymbol{\beta}}_{SW}$ under design-based inference has the form:

$$var_L(\hat{\boldsymbol{\beta}}_{SW}) = \hat{\boldsymbol{J}}(\hat{\boldsymbol{\beta}})^{-1} var_L(\hat{\boldsymbol{z}}^*)\hat{\boldsymbol{J}}(\hat{\boldsymbol{\beta}})^{-1} \tag{5.44}$$

where $\hat{\boldsymbol{z}}^* = \sum_{i\in s}\boldsymbol{z}_i^*$ is the weighted total of the $\boldsymbol{z}_i$'s and $\hat{\boldsymbol{J}}(\hat{\boldsymbol{\beta}})$ is the partial derivative (Jacobian matrix) of estimating equations for $\hat{\boldsymbol{\beta}}_{SW}$, that is $\frac{\partial^2 \hat{l}(\boldsymbol{\beta}_{SW})}{\partial\boldsymbol{\beta}_{SW}\partial\boldsymbol{\beta}_{SW}^T}$ in (5.41). Hence,

$$\hat{\boldsymbol{J}}(\hat{\boldsymbol{\beta}}) = -\sum_{i\in s} w_i\hat{p}_i(1-\hat{p}_i)\dot{\boldsymbol{x}}_i\dot{\boldsymbol{x}}_i^T = -\boldsymbol{X}^T\boldsymbol{W} diag[\hat{p}_i(1-\hat{p}_i)]\boldsymbol{X}, \tag{5.45}$$

which is also called the information matrix when discussing model fitting and assessment of fit.

Substituting (5.45) in (5.44), the design-based estimated variance for $\hat{\boldsymbol{\beta}}_{SW}$ is consistent with (5.15) in GLM and can be written as:

$$\begin{aligned} var_L(\hat{\boldsymbol{\beta}}_{SW}) &= [\boldsymbol{X}^T\boldsymbol{W}\hat{\boldsymbol{\Gamma}}\boldsymbol{X}]^{-1} var_\pi(\hat{\boldsymbol{z}}^*)[\boldsymbol{X}^T\boldsymbol{W}\hat{\boldsymbol{\Gamma}}\boldsymbol{X}]^{-1} \\ &= \boldsymbol{A}^{-1} var_\pi(\hat{\boldsymbol{z}}^*)\boldsymbol{A}^{-1}, \end{aligned} \tag{5.46}$$

where $\boldsymbol{A} = \boldsymbol{X}^T\boldsymbol{W}\hat{\boldsymbol{\Gamma}}\boldsymbol{X}$ with $\hat{\boldsymbol{\Gamma}} = diag[\hat{p}_i(1-\hat{p}_i)]$.

Following the derivation in Section 5.2.2 and referring to the design-based variance estimator for $\hat{\boldsymbol{\beta}}_{SW}$ in (5.47), if we let $\hat{\boldsymbol{V}} = \frac{n}{n-1} diag(e_i^2)$ in the unequal weighted single stage sampling design, or $\hat{\boldsymbol{V}} = \frac{n}{n-1} Blkdiag(\boldsymbol{e}_i \boldsymbol{e}_i^T)$ in the multistage sampling design, or $\hat{\boldsymbol{V}} = Blkdiag \left\{ \frac{n_h}{n_h - 1} \left[Blkdiag(\boldsymbol{e}_{hi} \boldsymbol{e}_{hi}^T) - \frac{1}{n_h} \boldsymbol{e}_h \boldsymbol{e}_h^T \right] \right\}$ in the stratified multistage sampling design, the linearization variance estimator $var_L(\hat{\boldsymbol{\beta}}_{SW})$ in the logistic regression in (5.46) can be rewritten as:

$$var_L(\hat{\boldsymbol{\beta}}_{SW}) = \boldsymbol{A}^{-1} \boldsymbol{X}^T \boldsymbol{W}_\Gamma \boldsymbol{V}_\Delta \boldsymbol{W}_\Gamma \boldsymbol{X} \boldsymbol{A}^{-1}, \tag{5.47}$$

with $\boldsymbol{W}_\Gamma = diag\left[w_i \hat{p}_i (1 - \hat{p}_i)\right]$ and $\boldsymbol{V}_\Delta = diag\left[w_i \hat{p}_i (1 - \hat{p}_i)\right]^{-1} \hat{\boldsymbol{V}} diag\left[w_i \hat{p}_i (1 - \hat{p}_i)\right]^{-1}$.

According to the VIF functions for GLM as shown in (5.33) and (5.36), the VIF for the k^{th} explanatory variable in the logistic regression is:

$$\frac{\hat{\zeta}_k \hat{\varrho}_k}{1 - R^2_{W\Gamma(k)}} \tag{5.48}$$

where $\boldsymbol{e}_{xk} = \boldsymbol{x}_k - \boldsymbol{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)}$ is the residual from WLS regressing $\boldsymbol{x}_k$ on $\boldsymbol{X}_{(k)}$ using weight matrix $\boldsymbol{W}_\Gamma = \boldsymbol{W}\hat{\boldsymbol{\Gamma}}$, $\hat{\zeta}_k = \frac{(\boldsymbol{x}_k - \boldsymbol{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)})^T \boldsymbol{W}_\Gamma \boldsymbol{V}_\Delta \boldsymbol{W}_\Gamma (\boldsymbol{x}_k - \boldsymbol{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)})}{(\boldsymbol{x}_k - \boldsymbol{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)})^T \boldsymbol{W}_\Gamma (\boldsymbol{x}_k - \boldsymbol{X}_{(k)} \hat{\boldsymbol{\beta}}_{W\Gamma(k)})} = \frac{\boldsymbol{e}_{xk}^T \boldsymbol{W}_\Gamma \boldsymbol{V}_\Delta \boldsymbol{W}_\Gamma \boldsymbol{e}_{xk}}{\boldsymbol{e}_{xk}^T \boldsymbol{W}_\Gamma \boldsymbol{e}_{xk}}$, and $\hat{\varrho}_k = \frac{\boldsymbol{x}_k^T \boldsymbol{W}_\Gamma \boldsymbol{x}_k}{\boldsymbol{x}_k^T \boldsymbol{W}_\Gamma \boldsymbol{V}_\Delta \boldsymbol{W}_\Gamma \boldsymbol{x}_k}$; and the intercept-adjusted VIF for the k^{th} explanatory variable in the logistic regression is:

$$\frac{\hat{\zeta}_k \hat{\varrho}_{km}}{1 - R^2_{W\Gamma m(k)}} \tag{5.49}$$

where $R^2_{W\Gamma m(k)} = \frac{\hat{\boldsymbol{\beta}}^T_{W\Gamma(k)} \boldsymbol{X}^T \boldsymbol{W}_\Gamma \boldsymbol{X} \hat{\boldsymbol{\beta}}_{W\Gamma(k)} - \hat{N} \bar{\bar{x}}_k^2}{\boldsymbol{x}_k^T \boldsymbol{W}_\Gamma \boldsymbol{x}_k - \hat{N} \bar{\bar{x}}_k^2}$ is the coefficient of determination corrected for the mean in the WLS regression when weight matrix is $\boldsymbol{W}_\Gamma$, and $\hat{\varrho}_{km} =$

$\frac{\boldsymbol{x}_k^T \boldsymbol{W}_\Gamma \boldsymbol{x}_k - \hat{N}\bar{\tilde{x}}_k^2}{(\boldsymbol{x}_k - \boldsymbol{1}\bar{\tilde{x}}_k)^T \boldsymbol{W}_\Gamma \boldsymbol{V}_\Delta \boldsymbol{W}_\Gamma (\boldsymbol{x}_k - \boldsymbol{1}\bar{\tilde{x}}_k)}$ with $\hat{N} = \sum_{i \in s} w_i \hat{p}_i (1 - \hat{p}_i)$, and $\bar{\tilde{x}}_k = \sum_{i \in s} w_i \hat{p}_i (1 - \hat{p}_i) x_{ki} / \hat{N}$.

5.3 Condition Indexes with Variance Decomposition Method in Generalized Linear Model

Condition indexes with variance decomposition method is another popular collinearity diagnostic tool as been suggested in Belsley (1984b) for linear model. To detect problematic collinearities in other members of the class of GLMs, Mackinnon & Puterman (1990), Weissfeld & Sereika (1991) and Lesaffre & Marx (1993) modified this approach on the scaled information matrix in GLM instead of the design matrix in linear models, so that we can detect near dependencies (singularities) in the information matrix which affect the stability of the estimates of $\boldsymbol{\beta}$.

In the survey-weighted generalized linear model, we will apply similar approach to the information matrix $\boldsymbol{I}(\hat{\boldsymbol{\beta}}_{SW})$ evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{SW}$. As shown in (5.8), the information matrix $\boldsymbol{I}(\hat{\boldsymbol{\beta}}_{SW}) = \boldsymbol{X}^T \boldsymbol{W} \hat{\boldsymbol{\Gamma}} \boldsymbol{X}$. To be more specific, this approach is developed based on the the singular value decomposition of $\boldsymbol{I}(\hat{\boldsymbol{\beta}})$. Denote the weighted matrix $\boldsymbol{W}^{1/2} \hat{\boldsymbol{\Gamma}}^{1/2} \boldsymbol{X} = \boldsymbol{W}_\Gamma^{1/2} \boldsymbol{X}$ as $\tilde{\boldsymbol{X}}$. The singular value decomposition of $\tilde{\boldsymbol{X}}$ is $\tilde{\boldsymbol{X}} = \boldsymbol{U_1} \boldsymbol{D} \boldsymbol{U_2}^T$, where $\boldsymbol{U_1}$, $\boldsymbol{U_2}$ and $\boldsymbol{D}$ are usually different from the ones of $\boldsymbol{X}$, due to the unequal survey weights and $\hat{\gamma}_i$'s across different observed units.

The condition number of $\tilde{\boldsymbol{X}}$ is defined as $\kappa(\tilde{\boldsymbol{X}}) = \mu_{max}/\mu_{min}$, where μ_{max} and μ_{min} are maximum and minimum singular values of $\tilde{\boldsymbol{X}}$. The condition number of $\tilde{\boldsymbol{X}}$ is usually different from the condition number of the data matrix $\boldsymbol{X}$ due to the unequal survey weights and $\hat{\gamma}_i$'s. Condition indexes are defined as

$$\eta_k = \mu_{max}/\mu_k, \quad k = 1, ..., p \tag{5.50}$$

where μ_k is one of the singular values of $\tilde{\boldsymbol{X}}$. The scaled condition indexes and condition numbers are the condition indexes and condition numbers of the scaled $\tilde{\boldsymbol{X}}$.

Based on the extrema of the ratio of quadratic forms (Lin, 1984), the condition number $\kappa(\tilde{\boldsymbol{X}})$ is bounded in the range of:

$$\frac{(w \cdot \hat{\gamma})_{min}^{1/2}}{(w \cdot \hat{\gamma})_{max}^{1/2}} \kappa(\boldsymbol{X}) \leq \kappa(\tilde{\boldsymbol{X}}) \leq \frac{(w \cdot \hat{\gamma})_{max}^{1/2}}{(w \cdot \hat{\gamma})_{min}^{1/2}} \kappa(\boldsymbol{X}), \tag{5.51}$$

where $(w \cdot \hat{\gamma})_{min}$ and $(w \cdot \hat{\gamma})_{max}$ are the minimum and maximum values of $w_i \cdot \hat{\gamma}_i$ across different observed units.

This expression indicates that if the working weights used in the SWGLM (i.e. $w_i \cdot \hat{\gamma}_i$) do not vary too much, the condition number in SWGLM resembles the one in OLS. While if it is an unequal-weighted sampling design with a wide range of survey weights or the $\hat{\gamma}_i$ has a wide range, the condition number can be very different. When SWGLM has large condition number, OLS might not.

In Section 4.1.2, if we replace $\boldsymbol{W}$ and $\hat{\boldsymbol{V}}$ with $\boldsymbol{W}_\Gamma$ and $\boldsymbol{V}_\Delta$ defined in Section 5.2.2.2 respectively, we can obtain the variance decomposition proportions for survey-weighted generalized linear models.

5.4 Experimental Study

We will now illustrate the foregoing techniques and investigate their behavior using dietary intake data from the National Health and Nutrition Examination Survey(NHANES) 2001-2002. The dietary intake data are used to estimate the types and amounts of foods and beverages consumed during the 24-hour period prior to the interview (midnight to midnight), and to estimate intakes of energy, nutrients, and other food components from those foods and beverages. NHANES uses a complex, multistage, probability sampling design, oversampling of certain population subgroups is done to increase the reliability and precision of health status indicator estimates for these groups. Among the respondents who received the in-person interview in the mobile examination center (MEC), around 94% provided complete dietary intakes. The survey weights were constructed by taking MEC sample weights and further adjusting for the additional nonresponse and the differential allocation by day of the week for the dietary intake data collection. These weights are more variable than the MEC weights. The data set used in our study is a subset of 2001-2002 data composed of respondents aged between 18 and 65. Observations with missing values in the selected variables are excluded from the sample which finally contains 3,217 complete respondents. The final weights in our sample range from 617.8693 to 341,097.2373, with a ratio of 552:1. The design of the sample can be approximated by the stratified selection of 30 PSUs from 15 strata, with 2 PSUs within each stratum.

For this empirical study, the normal identity (linear) and binary logistic models are considered. The explanatory variables considered include two dummy variables for race (white and black, while treating the other ethnic groups as the reference group), the interaction term between gender (male=1, female=0) and age, and nine daily total nutrition intake variables (calorie(kcal), protein(gm), carbohydrate(gm), sugar(gm), total fat(gm), total saturated fatty acids(gm), total monounsaturated fatty acids(gm), total polyunsaturated fatty acids(gm) and alcohol(gm)). Logistic regression models are fit to these data with obese or non-obese respondent as a binary response variable [1], while normal identity models are fit to these data with respondent's Body Index Mass (BMI) as the dependent variable, where the values of BMI ranges from 15.41 to 52.09.

Three regression methods are applied and compared in this study. The first one uses *generalized least squares* (GLS) method but ignores sampling complexities including the weighting. The second one uses *generalized weighted least squares* (GWLS) method, which incorporates the survey weights but assumes units are independent, i.e. it ignores strata and clustering, and $\boldsymbol{V}_\Delta = \boldsymbol{W}_\Gamma^{-1}$. The third one is a design-based method that uses the actual, possibly complex, sampling design as described in section 5.2.2. We referto it as *survey weighted generalized least squares* (SWGLS). The weighted matrices, coefficient variance estimators and collinearity diagnostics of these three methods in the two regression models (identity and logistic) are listed in Table 5.2.

The results from fitting each of the two models using three different regression

[1] Obese is defined by the respondent's Body Index Mass (BMI) value. An adult who has a BMI of 30 or higher is considered obese.

methods are presented in Table 5.3. The models with all the explanatory variables are fitted at first. Then, a reduced model with less near-dependency problem is fitted with two dummy variables for race (other race and black, treating white as the reference group), gender, age and carbohydrate. Here, the reference group of race variables in the reduced model is changed from other ethnic groups to white so that the collinearity between the dummy variables and intercept can be reduced as we discussed in Chapter 4. The standard errors of coefficients in the identity models are relatively larger than those in the logistic models. The absence of some other correlated variables, including the interaction term between age and gender and other total nutrition intake variables, makes the standard errors of all the coefficients get smaller. Both age and carbohydrate are significant in all the full regression models (with all the explanatory variables). In the reduced regression models with fewer predictors, age stays significant in all the models. The absolute values of the coefficients of carbohydrate in the reduced models are also smaller, which leads the corresponding p-values get larger. Hence, in some regression models, carbohydrate is not significant anymore although the associated standard errors are smaller than those in the full models. It demonstrates that collinearity can not only inflate the estimated variance of coefficients but can influence the values of coefficients and their corresponding p-values.

Table 5.4 reports the VIF values for the two types of models using the three difference regression methods. The VIF formulas for these regression models are listed in Table 5.2. Calorie has the largest VIF values in all the regression models due to its high near-dependency with all the other total nutrition variables. Since total calories is based on the sum of the foods a person consumed, this dependence is expected. In the three identity models, most VIF values in SWGLS are larger than those in GLS and GWLS. Therefore, if we use standard packages to obtain VIF values for SWGLS, which may give you a set of VIF values in GLS or GWLS, we will underestimate the collinearity problem in this regression analysis. Nevertheless, in the three logistic models, most VIF values in SWGLS are smaller than those in GLS and GWLS. Hence, the specialized VIFs developed will give a better picture of the degree of collinearity in the survey-weighted regression analysis. In summary, although the data in both types of models are the same, the impact of collinearity on regression estimation can still be underestimated or overestimated in the design-based SWGLS if survey complexities are overlooked. The extent of the error in estimating VIFs depends on the response and the choice of the model in the class of GLMs.

Table 5.5 and Table 5.6 presents the scaled condition indexes and variance decomposition proportions for the two types of models using GLS and SWGLS. Consistently in all the four regression models, calorie, protein, carbohydrate, total fat and alcohol are involved in the dominant near-dependency; while total fat, total saturated fatty acids, total monounsaturated fatty acids and total polyunsaturated fatty acids are involved in the secondary near-dependency; and intercept, the interaction term between gender and age are involved in the third near-dependency. In Table 5.5, when using GLS in the identity model, the condition indexes corresponding to these three remarkable near-dependencies are smaller than the ones when SWGLS is used. But in logistic models as shown in Table 5.6, the condition indexes when GLS is used are larger than the ones when SWGLS is used for the two most significant near-dependencies. This phenomenon emphasizes that the difference of the impact of collinearity between GLS and SWGLS should not only depend on the survey weights and designs but also depend on the response and the choice of the model in the class GLMs.

Table 5.2: Regression Models and their Collinearity Diagnostic Statistics used in this Experimental Study

Model	$\boldsymbol{Y}$	Regression Type	$\boldsymbol{W}_\Gamma$ [a]	Variance Estimation of $\hat{\boldsymbol{\beta}}$	VIF fomula	Matrix for Condion Indexes [b]
Identity	BMI	GLS	$\boldsymbol{I}$	$\hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ [c]	$\text{VIF} = \frac{1}{1-R^2_{(k)}}$	$\boldsymbol{X}^T\boldsymbol{X}$
(Normal)		GWLS	$\boldsymbol{W}$ [d]	$\hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$	$\text{VIF} = \frac{1}{1-R^2_{SW(k)}}$	$\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$
		(model-based)				
		SWGLS	$\boldsymbol{W}$	$\hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\hat{\boldsymbol{V}}\boldsymbol{W}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}$	$\text{VIF} = \frac{\hat{\zeta}_k\hat{\varrho}_k}{1-R^2_{SW(k)}}$	$\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X}$
		(design-based)		$\hat{\boldsymbol{V}} = \sum_{h=1}^H \frac{n_h}{n_h-1}\left[Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T) - \frac{1}{n_h}\boldsymbol{e}_h\boldsymbol{e}_h^T\right]$	with $\hat{\zeta}_k = \frac{\boldsymbol{e}_{xk}^T\boldsymbol{W}\hat{\boldsymbol{V}}\boldsymbol{W}\boldsymbol{e}_{xk}}{\boldsymbol{e}_{xk}^T\boldsymbol{W}\boldsymbol{e}_{xk}}$,	
				$\hat{\varrho}_k = \frac{\boldsymbol{x}_k^T\boldsymbol{W}\boldsymbol{x}_k}{\boldsymbol{x}_k^T\boldsymbol{W}\hat{\boldsymbol{V}}\boldsymbol{W}\boldsymbol{x}_k}$		
Logistic	Obese	GLS	$\hat{\boldsymbol{\Gamma}}$ [e]	$(\boldsymbol{X}^T\hat{\boldsymbol{\Gamma}}\boldsymbol{X})^{-1}$ [f]	$\text{VIF} = \frac{1}{1-R^2_{\Gamma(k)}}$	$\boldsymbol{X}^T\boldsymbol{\Gamma}\boldsymbol{X}$
(Binary)	/Non-obese	GWLS	$\boldsymbol{W}\hat{\boldsymbol{\Gamma}}$ [g]	$(\boldsymbol{X}^T\boldsymbol{W}_\Gamma\boldsymbol{X})^{-1}$	$\text{VIF} = \frac{1}{1-R^2_{W\Gamma(k)}}$	$\boldsymbol{X}^T\boldsymbol{W}_\Gamma\boldsymbol{X}$
		(model-based)				
		SWGLS	$\boldsymbol{W}\hat{\boldsymbol{\Gamma}}$	$(\boldsymbol{X}^T\boldsymbol{W}_\Gamma\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}_\Gamma\boldsymbol{V}\Delta\boldsymbol{W}_\Gamma\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}_\Gamma\boldsymbol{X})^{-1}$	$\text{VIF} = \frac{\hat{\zeta}_k\hat{\varrho}_k}{1-R^2_{W\Gamma(k)}}$	$\boldsymbol{X}^T\boldsymbol{W}_\Gamma\boldsymbol{X}$
		(design-based)		$\boldsymbol{V}_\Delta = \hat{\boldsymbol{\Delta}}\hat{\boldsymbol{V}}\hat{\boldsymbol{\Delta}}$ [h]	with $\hat{\zeta}_k = \frac{\boldsymbol{e}_{xk}^T\boldsymbol{W}_\Gamma\boldsymbol{V}_\Delta\boldsymbol{W}_\Gamma\boldsymbol{e}_{xk}}{\boldsymbol{e}_{xk}^T\boldsymbol{W}_\Gamma\boldsymbol{e}_{xk}}$,	
				$= \hat{\boldsymbol{\Delta}}\left\{\sum_{h=1}^H \frac{n_h}{n_h-1}\left[Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T) - \frac{1}{n_h}\boldsymbol{e}_h\boldsymbol{e}_h^T\right]\right\}\hat{\boldsymbol{\Delta}}$	$\hat{\varrho}_k = \frac{\boldsymbol{x}_k^T\boldsymbol{W}_\Gamma\boldsymbol{x}_k}{\boldsymbol{x}_k^T\boldsymbol{W}_\Gamma\boldsymbol{V}_\Delta\boldsymbol{W}_\Gamma\boldsymbol{x}_k}$	

[a] In all the regression models, the parameters are estimated by: $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{W}_\Gamma\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}_\Gamma\boldsymbol{Y}$.

[b] The eigenvalues of this matrix will be used to compute the Condition Indexes for the corresponding regression model.

[c] $\tau^2 = \sigma^2$ in identity model

[d] It is the diagonal matrix with survey weights w_i on the main diagonal.

[e] $\hat{\boldsymbol{\Gamma}} = diag\,[\hat{p}_i(1-\hat{p}_i)]$

[f] $\tau^2 = 1$ in logistic model

[g] It is the diagonal matrix with survey weights $w_i * \gamma_i$ on the main diagonal.

[h] $\hat{\boldsymbol{\Delta}} = diag\left\{[\hat{p}_i(1-\hat{p}_i)]^{-1}\right\}$

Table 5.3: Parameter Estimates with Their Associated Standard Errors in Two Models using Three Different Regression Methods

Variable	Normal (Identity Model)			Binary (Logistic Model)		
	GLS	GWLS model based	SWGLS design based	GLS	GWLS model based	SWGLS design based
Intercept	24.28***	24.44***	24.44***	-2.03***	-2.02***	-2.02***
	(0.41)	(0.48)	(0.85)	(0.23)	(0.26)	(0.41)
White	-0.18	0.15	0.15	0.21	0.34*	0.34*
	(0.20)	(0.24)	(0.28)	(0.11)	(0.13)	(0.15)
Black	-0.46	0.19	0.19	0.19	0.46*	0.46*
	(0.25)	(0.34)	(0.38)	(0.13)	(0.18)	(0.23)
Gender	-0.52	-0.53	-0.53	-0.55*	-0.30	-0.30
	(0.48)	(0.56)	(0.74)	(0.28)	(0.30)	(0.36)
Age	7.4e-02***	5.4e-02***	5.4e-02***	2.2e-02***	1.6e-02**	1.6e-02*
	(8.9e-03)	(9.9e-03)	(1.7e-02)	(4.5e-03)	(7.3e-03)	(7.0e-03)
Gender*Age	1.1e-02	2.5e-02	2.5e-02	7.1e-03	5.0e-03	5.0e-03
	(3.6e-03)	(1.4e-02)	(1.6e-02)	(6.4e-03)	(5.3e-03)	(7.8e-03)
Calorie	-4.4e-02**	1.6e-02***	1.6e-02***	5.8e-03**	5.9e-03**	5.9e-03**
	(1.6e-02)	(3.5e-03)	(4.4e-03)	(2.0e-03)	(1.9e-03)	(1.9e-03)
Protein	-4.8e-02**	-6.6e-02***	-6.6e-02***	-2.3e-02**	-2.4e-02**	-2.4e-02**
	(1.4e-02)	(1.5e-02)	(1.9e-02)	(8.9e-03)	(8.2e-03)	(8.2e-03)
Carbohydrate	3.5e-03***	-6.4e-02***	-6.4e-02***	-2.4e-02**	-2.4e-02**	-2.4e-02**
	(1.9e-03)	(1.4e-02)	(1.8e-02)	(8.0e-03)	(7.4e-03)	(7.6e-03)
Sugar	-0.12	9.2e-04	9.2e-04	7.6e-04	-6.7e-04	-6.7e-04
	(2.9e-02)	(2.0e-03)	(3.6e-03)	(1.1e-03)	(1.1e-03)	(2.2e-03)
Total Fat	4.4e-03***	-0.12***	-0.12***	-5.0e-02**	-5.0e-02***	-5.0e-02*
	(2.9e-02)	(2.7e-02)	(4.5e-02)	(1.7e-02)	(1.5e-02)	(2.0e-02)
Total Saturated Fatty Acids	3.2e-02	-3.0e-02	-3.0e-02	-1.0e-02	-5.3e-03	-5.3e-03
	(2.8e-02)	(2.8e-02)	(4.4e-02)	(1.7e-02)	(1.5e-02)	(2.0e-02)
Total Monounsaturated Fatty Acids	2.0e-02	9.4e-03	9.4e-03	1.9e-03	6.5e-03	6.5e-03
	(2.7e-02)	(2.7e-02)	(4.3e-02)	(1.6e-02)	(1.4e-02)	(2.2e-02)
Total Polyunsaturated Fatty Acids	-8.2e-02	-8.8e-03	-8.8e-03	4.2e-03	1.8e-03	1.8e-03
	(2.5e-02)	(2.6e-02)	(4.0e-02)	(1.6e-02)	(1.4e-02)	(1.7e-02)
Alcohol	1.2e-02**	-0.11***	-0.11***	-4.5e-02**	-4.3e-02**	-4.3e-02**
	(1.2e-02)	(2.4e-02)	(3.1e-02)	(1.4e-02)	(1.3e-02)	(1.5e-02)
R^2	0.0574	0.0449	0.0449			
Intercept	24.27***	24.42***	24.42***	-1.97***	-1.73	-1.73***
	(0.34)	(0.38)	(0.67)	(0.19)	(0.20)	(0.33)
Other Race	-1.24**	-0.81	-0.81	-0.82**	-0.55	-0.55
	(0.44)	(0.43)	(0.64)	(0.31)	(0.22)	(0.47)
Black	-0.21	0.31	0.31	0.12	0.25	0.25
	(0.21)	(0.28)	(0.28)	(0.11)	(0.15)	(0.17)
Gender	-7.6e-02	0.53**	0.53**	-0.30**	-7.4e-02	-7.4e-02
	(0.18)	(0.18)	(0.20)	(0.10)	(0.10)	(7.8e-02)
Age	7.0e-02***	5.8e-02***	5.8e-02***	2.3e-02***	1.6e-02	1.6e-02**
	(6.2e-03)	(7.0e-03)	(1.0e-02)	(3.3e-03)	(3.63e-03)	(4.6e-03)
Carbohydrate	-1.8e-03**	-2.3e-03***	-2.3e-03*	-7.9e-04*	-9.0e-04	-9.0e-04
	(6.4e-04)	(6.7e-04)	(9.8e-04)	(3.7e-04)	(3.48e-04)	(5.3e-04)
R^2	0.0520	0.0326	0.0326			

Table 5.4: VIF values for Two Regression Models using Three Different Regression Methods

Variable	Normal (Identity Model) GLS	Normal (Identity Model) GWLS model based	Normal (Identity Model) SWGLS design based	Binary (Logistic Model) GLS	Binary (Logistic Model) GWLS model based	Binary (Logistic Model) SWGLS design based
White	1.38	1.52	2.21	1.44	1.68	2.85
Black	1.36	1.47	2.87	1.43	1.65	3.26
Gender	8.00	10.53	10.57	9.13	11.40	11.00
Age	10.31	10.47	7.31	11.23	10.85	10.11
Gender*Age	16.19	19.25	18.42	17.26	19.99	22.98
Calorie	2171.23	1885.14	2342.16	1980.31	1893.56	933.14
Protein	76.21	64.87	74.35	75.21	69.76	60.26
Carbohydrate	567.66	490.05	594.10	524.24	458.74	374.36
Sugar	4.70	4.53	4.78	4.49	4.22	4.90
Total Fat	297.85	237.04	413.25	304.10	272.04	298.48
Total Saturated Fatty Acids	37.57	32.43	38.08	36.66	36.19	36.31
Total Monounsaturated Fatty Acids	45.38	37.20	69.48	45.64	42.11	62.29
Total Polyunsaturated Fatty Acids	15.11	12.57	16.53	16.44	14.35	15.78
Alcohol	108.36	94.88	124.11	63.55	72.71	8.83

Table 5.5: Scaled Condition Indexes and Variance Decomposition Proportions: the Identity Model using NHANES 2001-2002 Data File

Scaled Condition Index	Scaled Proportion of the Variance of Intercept	White	Black	Gender	Age	Gender*Age	Calorie	Protein	Carbohydrate	Sugar	Total Fat	Sat.fat[a]	Mono.fat[b]	Poly.fat[c]	Alcohol
	Normal Identity Model without considering survey weights and survey design, GLS														
1	.[d]	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	.	.	0.325	.	.	.	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
6	.	0.661	0.391	.	.	.	.	.	.	.	.	.	.	.	.
9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
15	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
21	.	.	.	.	.	.	.	.	.	0.439	.	.	.	.	.
22	.	.	.	.	.	.	.	.	.	0.344	.	.	.	.	.
58	0.877	.	.	0.939	0.902	0.903	.	.	.	.	.	.	.	.	.
95	.	.	.	.	.	.	.	.	.	.	0.498	0.675	0.658	0.679	.
418	.	.	.	.	.	.	0.999	0.962	0.992	.	0.501	.	.	.	0.988
	Normal Identity Model with considering survey weights and survey design, SWGLS														
1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
9	.	0.322	.	.	.	.	.	.	.	.	.	.	.	.	.
10	.	0.377	.	.	.	.	.	.	.	.	.	.	.	.	.
11	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
12	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
17	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
22	.	.	.	.	.	.	.	.	.	0.575	.	.	.	.	.
23	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
72	0.834	.	.	0.856	0.869	0.874	.	.	.	.	.	.	.	.	.
88	.	.	.	.	.	.	.	.	.	.	0.677	0.882	0.780	0.837	.
404	.	.	.	.	.	.	1.000	0.997	0.957	.	0.316	.	.	.	1.000

[a] Total Saturated Fatty Acids
[b] Total Monounsaturated Fatty Acids
[c] Total Polyunsaturated Fatty Acid
[d] The scaled variance decomposition proportions smaller than 0.3 is omitted in this table.

Table 5.6: Scaled Condition Indexes and Variance Decomposition Proportions: the Logistic Model using NHANES 2001-2002 Data File

Scaled	Scaled Proportion of the Variance of														
Condition Index	Intercept	White	Black	Gender	Age	Gender*Age	Calorie	Protein	Carbohydrate	Sugar	Total Fat	Sat.fat[a]	Mono.fat[b]	Poly.fat[c]	Alcohol
Binary Logistic Model without considering survey weights and survey design, GLS															
1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	.	.	0.345	.	.	.	.	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
7	.	0.792	0.499	.	.	.	.	.	.	.	.	.	.	.	.
9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
10	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
12	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
16	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
20	.	.	.	.	.	.	.	.	.	0.368	.	.	.	.	.
22	.	.	.	.	.	.	.	.	.	0.392	.	.	.	.	.
74	0.909	.	.	0.955	0.929	0.919	.	.	.	.	.	.	.	.	.
100	.	.	.	.	.	.	.	.	.	.	0.505	0.705	0.689	0.721	.
409	.	.	.	.	.	.	0.998	0.962	0.991	.	0.493	.	.	.	0.975
Binary Logistic Model with considering survey weights and survey design, SWGLS															
1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
3	.	.	0.442	.	.	.	.	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
4	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
6	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
9	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
11	.	0.449	.	.	.	.	.	.	.	.	.	.	.	.	.
12	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
13	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
18	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
22	.	.	.	.	.	.	.	.	.	0.506	.	.	.	.	.
23	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
79	0.931	.	.	0.969	0.933	1.000	.	.	.	.	.	.	.	.	.
101	.	.	.	.	.	.	.	.	.	.	0.602	0.775	0.651	0.698	.
416	.	.	.	.	.	.	0.996	1.000	0.940	.	0.367	.	.	.	0.925

[a] Total Saturated Fatty Acids
[b] Total Monounsaturated Fatty Acids
[c] Total Polyunsaturated Fatty Acid

Chapter 6

Conclusion and Discussion

Belsley (1991) stated that: "... in nonexperimental sciences, ..., collinearity is a natural flaw in the data set resulting from the uncontrollable operations of the data-generating mechanism and is simply a painful and unavoidable fact of life." When using survey data, few analysts have escaped the problem of collinearity in regression estimation and the presence of this problem encumbers precise statistical explanation about the relationships between predictors and responses. Its side effects are: high standard errors, low or high t-statistics, and illogical or overly sensitive parameter estimates. Many books and articles describe the collinearity problem and propose strategies to understand, assess and handle its presence. But few provide appropriate diagnostics to take the survey complexities into account in the analysis of complex survey data. The collinearity diagnostics developed in the preceding chapters, consisting of variance inflation factors (VIFs), condition indexes, and variance decomposition proportions for survey-weighted linear models and generalized linear models, allow us to detect the presence of collinearity in the complex survey data and to determine the variables involved in each near-dependency. Furthermore, the newly-developed diagnostic statistics allow us to evaluate how much each of the regression estimates is degraded by the presence of collinearity.

In ordinary least squares (OLS) or generalized least squares (GLS), the model parameter estimators and their variance estimators, denoted as $\hat{\boldsymbol{\theta}}$, are functions of data matrix $\boldsymbol{X}$ and response variable $\boldsymbol{Y}$, which can be expressed as $\hat{\boldsymbol{\theta}} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{Y})$ where $\boldsymbol{f}$ stands for a certain function. The collinearity diagnostics in these regressions deal with the existence of the nearly linear relationship among columns of $\boldsymbol{X}$ and evaluate their impact on the estimation of $\hat{\boldsymbol{\theta}}$. Thus, these diagnostic statistics are only determined by the three factors here, $\boldsymbol{f}$, $\boldsymbol{X}$ and $\boldsymbol{Y}$. To be more specific, in OLS, the collinearity diagnostics, VIFs, condition indexes and variance decomposition proportions depend on the structure of the data matrix $\boldsymbol{X}$; while in GLS, these diagnostic statistics depend on the data matrix $\boldsymbol{X}$, the response variable $\boldsymbol{Y}$ (through its covariance matrix) and the choice of the model (defined in $\boldsymbol{f}$). The details have been discussed in Chapter 2. However, in a linear model using survey-weighted least squares (SWLS) or survey-weighted generalized linear models (SWGLMs), we aim to estimate the parameters in the finite population, $\boldsymbol{\theta}_U$, using a selected sample. Not only the data matrix $\boldsymbol{X}$ and the response variable $\boldsymbol{Y}$, but also the survey weights $\boldsymbol{W}$ and the variance-covariance matrix $\boldsymbol{V}$ which reflects the heterogeneous errors, sample designs and features of the finite population, are involved in the regression estimation, that is $\hat{\boldsymbol{\theta}} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{V})$. Thus, the collinearity diagnostics consistently need to incorporate these factors when assessing the influence of collinearity among the given data. Unavoidably, the survey weights, sample designs, features of the finite population (reflected in $\boldsymbol{V}$) and the methods used for inference (defined in $\boldsymbol{f}$) will influence our decision on the pattern and harmfulness of collinearity among certain predictors.

In this dissertation we have developed methods that generally have either a model-based or design-based interpretation. For example, in deriving VIFs for linear models, we have considered the case of an analyst who estimates model parameters using survey-weighted least squares in order to account for an informative sample design. The SWLS parameter estimates will estimate both underlying model parameters (assuming a correctly specified model) and census-fit parameters in the finite population. In the case of linear regression, a type of sandwich variance estimator will estimate both the model variance and design variance of the SWLS slope estimator. The model or design variance of $\hat{\beta}_k$, an estimator of slope associated with the predictor $\boldsymbol{x}_k$, is inflated somewhat when different predictors are correlated with each other composed to what the variance would be if $\boldsymbol{x}_k$ were orthogonal to the other predictors. The measure of inflation, the VIF, is composed of terms that must be estimated from the sample. Our approach has been to substitute estimators that have both a model and design interpretation.

In Chapter 3, we relax the assumption in standard weighted least squares (WLS) that $\boldsymbol{V}^{-1} = \boldsymbol{W}$, which is always untrue in the analysis of complex survey data. The VIFs for SWLS are obtained by multiplying the inverse of $1 - R^2_{SWk}$ by two adjustment coefficients, ζ_k and ϱ_k. These coefficients are decided by the data matrix $\boldsymbol{X}$, the survey weight matrix $\boldsymbol{W}$ and the variance-covariance matrix $\boldsymbol{V}$ (or $\hat{\boldsymbol{V}}$ when $\boldsymbol{V}$ is unknown). The adjusted VIFs can precisely reflect the degree of the variance inflation caused by the linear relationship between the k^{th} predictor and the other predictors in the linear model when SWLS is used. In the simulation studies, we demonstrate the application of the new approaches and compare the new VIFs with the VIFs obtained from the conventional methods.

New VIFs are needed because the standard ones for OLS and WLS usually do not reflect features of the population that also occur in the sample design. For example, clustering affects the variance of a regression parameter estimator, but is not incorporated in the OLS and WLS VIFs. The new VIFs can be either larger or smaller than the VIFs from WLS or OLS. Using different simulation samples that are drawn from the same finite population by the same sampling design, the VIFs can be different from the standard ones. One remedy for collinearity is to remove $\boldsymbol{x}$'s found to be collinear with others. We demonstrated that using the backward selection method, different reduced models can be selected in different simulation samples. This introduces problems for variance estimation that we have not attempted to solve here.

In Chapter 4, the condition indexes are extended for SWLS and proper variance decomposition proportions are developed for the variance estimators in SWLS. As for VIFs, special methods are needed for SWLS estimates that account for stratification. clustering and unequal weighting. Two experimental studies are conducted to contrast the collinearity diagnostic statistics among different regression methods and an example is provided to show that the choice of reference category for a categorical variable may affect the degree of collinearity in the data.

In Chapter 5, we extend the VIFs, condition indexes and variance decomposition proportions developed in the previous two chapters to the GLMs. The new methods are applied to a logistic model as an example. In the experimental study, we demonstrate the application of our new approaches and show that in GLMs, the response and the choice of the model together with the data matrix and survey complexities play a role in the degree of collinearity in the data.

The major effect of collinearity is degrading the regression estimates but degrading collinearity need not to be harmful. Some cutoffs have been suggested as signs of severe collinearity, such as a VIF value larger than 7 or 10, a condition index larger than 10 or 30, or a variance decomposition proportion larger than 0.30 or 0.50, but these suggestions are ad hoc. In fact, the research here does not suggest that the rule-of-thumb of 7, 10, or 30 or 0.30 or 0.50 need to be changed when analyzing survey data. Our conclusion is the same as the one that for non-survey data: examining VIFs or condition indexes or variance decomposition proportions alone is inadequate. A practical example is that when the standard error of the k^{th} regression coefficient is very small and the coefficient is significant, adding some collinear variables may not change that significance even if the VIF is 10 or even larger. One the other hand, collinearity can be harmful if it causes a predictor to have an illogical sign. To decide if the collinearity is harmful, we not only need to obtain accurate collinearity diagnostics for the regression models, but also need to consider the variances of parameters that are of interest. Especially when we are handling survey data, if the sample size is small, the variances of parameters are relatively large and the coefficients can be more sensitive to the presence of collinearity; while when the sample size is very large, the variances of parameters are relatively small enough and the coefficients are more tolerant to the presence of collinearity. Therefore, to determine the appropriate cutoffs for collinearity diagnostics, these factors should be considered.

The corrective action taken for fixing collinearity in this dissertation is simply drop the most collinear variables and obtain a reduced model by the backward selection method. But, in reality, it is not appropriate to drop variables which have a theoretical substantive role in someone's statistical analysis. The most direct and obvious means for solving collinearity is through the collection and use of additional data. If the sample size is large enough, significant predictors will be identified even if the variances of their estimated coefficients are inflated by collinearity. But when it is impossible to obtain new data, some other techniques are needed. A very common approach is to transform the data, by standardizing, centering, taking logarithms, or computing first differences (on time series data). Or, some specialized techniques such as mixed Bayesian and ridge regression procedures can be used, but require extra (prior) information beyond what is available to many analysts. The applications of these remedies on survey data are limited and need to be fully evaluated in the future research.

Appendix A

Inversion of Partitioned Matrices, $\dot{a}^{00}$, $\boldsymbol{a}^{(2...p)1}$

We repeatedly use the formula for the inverse of a partitioned matrix (see Theil, 1971, Section 1.2). Consider the following nonsingular and symmetric matrix:

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{P}_1 & \boldsymbol{R}_1 \\ \boldsymbol{R}_1^T & \boldsymbol{Q}_1 \end{pmatrix}$$

where $\boldsymbol{P}_1$ and $\boldsymbol{Q}_1$ are nonsingular symmetric submatrices. The inverse of $\boldsymbol{D}$ is

$$\boldsymbol{E} = \begin{pmatrix} \boldsymbol{P}^1 & \boldsymbol{R}^1 \\ \boldsymbol{R}^{1^T} & \boldsymbol{Q}^1 \end{pmatrix}$$

where

$$\boldsymbol{P}^1 = (\boldsymbol{P}_1 - \boldsymbol{R}_1 \boldsymbol{Q}_1^{-1} \boldsymbol{R}_1^T)^{-1} \tag{A.1}$$

$$\boldsymbol{R}^{1^T} = -\boldsymbol{Q}^1 \boldsymbol{R}_1^T \boldsymbol{P}^{-1}$$

The derivation of an element of the inverse matrix can be shown using the results obtained from the inversion of a partitioned matrix. We begin with

$$\dot{\boldsymbol{A}}_{(p+1)\times(p+1)} = \begin{pmatrix} \tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{Y}} & \tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{X}} \\ \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}} & \tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}} \end{pmatrix}$$

Using (A.1), the upper left element $\dot{a}^{00}$ of $\dot{\boldsymbol{A}}^{-1}$ can then be shown to be:

$$\dot{a}^{00} = (\tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}})^{-1} = \frac{1}{\tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{Y}}^T\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{Y}}}.$$

Some rearrangement yields (3.5).

To derive $\boldsymbol{a}^{(k)(k)}$ and $\boldsymbol{a}^{(k)k}$ in (3.8), use the form of $\boldsymbol{A}$ in (3.6) and write the inverse of $\boldsymbol{A}$ as:

$$\boldsymbol{A}^{-1} = \begin{pmatrix} a^{kk} & \boldsymbol{a}^{k(k)} \\ \boldsymbol{a}^{(k)k} & \boldsymbol{a}^{(k)(k)} \end{pmatrix} = \begin{pmatrix} \tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k & \tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{X}}_{(k)} \\ \tilde{\boldsymbol{X}}_{(k)}^T\tilde{\boldsymbol{x}}_k & \tilde{\boldsymbol{X}}_{(k)}^T\tilde{\boldsymbol{X}}_{(k)} \end{pmatrix}^{-1}$$

Using the formula derived above, $\boldsymbol{P}^1 = (\boldsymbol{P}_1 - \boldsymbol{R}_1\boldsymbol{Q}_1^{-1}\boldsymbol{R}_1^T)^{-1}$, for the inverse of a partitioned matrix:

$$a^{kk} = [\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{X}}_{(k)}\tilde{\boldsymbol{A}}_{(k)}^{-1}\tilde{\boldsymbol{X}}_{(k)}^T\tilde{\boldsymbol{x}}_k]^{-1} = \frac{1}{1 - R^2_{SW(k)}}\frac{1}{\tilde{\boldsymbol{x}}_k^T\tilde{\boldsymbol{x}}_k}$$

with $\tilde{\boldsymbol{A}}_{(k)} = \tilde{\boldsymbol{X}}_{(k)}^T\tilde{\boldsymbol{X}}_{(k)}$.

The vector of $\boldsymbol{a}^{(k)k}$ in $\boldsymbol{A}^{-1}$ corresponds to $\boldsymbol{R}^{1^T}$ in $\boldsymbol{E}$, and thus, using $\boldsymbol{R}^{1^T} = -\boldsymbol{Q}^1\boldsymbol{R}_1^T\boldsymbol{P}^{-1}$, we have:

$$\boldsymbol{a}^{(k)k} = -a^{kk}(\tilde{\boldsymbol{X}}_{(k)}^T \tilde{\boldsymbol{X}}_{(k)})^{-1} \tilde{\boldsymbol{X}}_{(k)} \tilde{\boldsymbol{x}}_k = -a^{kk} \hat{\boldsymbol{\beta}}_{SW(k)}.$$

Appendix B1

Derivation of the Model Variance of $\hat{\beta}_{SWk}$ in M2

Consider M2: $\boldsymbol{Y} = \boldsymbol{\beta_0} + \tilde{\boldsymbol{x}}_k\beta_k + \boldsymbol{e}$ can also be written as: $\boldsymbol{Y} = \tilde{\mathbf{1}}\beta_0 + \tilde{\boldsymbol{x}}_k\beta_k + \boldsymbol{e}$ where $\tilde{\mathbf{1}} = \boldsymbol{W}^{1/2} * \mathbf{1} = (w_1^{1/2}, \ldots, w_n^{1/2})^T$ and $\mathbf{1}$ is a column vector consisting of n unit elements. When $\tilde{\mathbf{1}}$ is regressed on $\tilde{\boldsymbol{x}}_k$, the estimated coefficient of $\tilde{\boldsymbol{x}}_k$, $\hat{\beta}_{SW(k)}$ is:

$$
\begin{aligned}
\hat{\beta}_{SWk} &= (\tilde{\mathbf{1}}^T\tilde{\mathbf{1}})^{-1}\tilde{\mathbf{1}}^T\tilde{\boldsymbol{x}}_k \\
&= \left((w_1^{1/2}, \ldots, w_n^{1/2}) \begin{pmatrix} w_1^{1/2} \\ \vdots \\ w_n^{1/2} \end{pmatrix} \right)^{-1} (w_1^{1/2}, \ldots, w_n^{1/2})\tilde{\boldsymbol{x}}_k \\
&= \frac{\sum_{i\in s} w_i x_{ki}}{\sum_{i\in s} w_i}
\end{aligned}
\tag{B1.1}
$$

Let $\hat{N} = \sum_{i\in s} w_i$ and $\bar{\bar{x}}_k = \sum_{i\in s} w_i x_{ki}/\hat{N}$, then $\hat{\beta}_{SW(k)} = \bar{\bar{x}}_k$.

Using (3.13) and (3.14), the variance of $\hat{\beta}_{SW(k)}$ can be written as:

$$
\begin{aligned}
Var_{M2}(\hat{\beta}_{SW_k}) &= \frac{(\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{X}}_{(k)}\hat{\boldsymbol{\beta}}_{SW(k)})^T\boldsymbol{W}\boldsymbol{V}(\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{X}}_{(k)}\hat{\boldsymbol{\beta}}_{SW(k)})}{[(\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{X}}_{(k)}\hat{\boldsymbol{\beta}}_{SW(k)})^T(\tilde{\boldsymbol{x}}_k - \tilde{\boldsymbol{X}}_{(k)}\hat{\boldsymbol{\beta}}_{SW(k)})]^2} \\
&= \frac{(\tilde{\boldsymbol{x}}_k - \tilde{\mathbf{1}}\hat{\boldsymbol{\beta}}_{SW(k)})^T\boldsymbol{W}\boldsymbol{V}(\tilde{\boldsymbol{x}}_k - \tilde{\mathbf{1}}\hat{\boldsymbol{\beta}}_{SW(k)})}{[(\tilde{\boldsymbol{x}}_k - \tilde{\mathbf{1}}\hat{\boldsymbol{\beta}}_{SW(k)})^T(\tilde{\boldsymbol{x}}_k - \tilde{\mathbf{1}}\hat{\boldsymbol{\beta}}_{SW(k)})]^2}
\end{aligned}
\tag{B1.2}
$$

In the denominator, because $\hat{\beta}_{SW(k)} = \bar{\bar{x}}_k$, we have $(\tilde{\boldsymbol{x}}_k - \tilde{\mathbf{1}}\hat{\boldsymbol{\beta}}_{SW(k)})^T(\tilde{\boldsymbol{x}}_k - \tilde{\mathbf{1}}\hat{\boldsymbol{\beta}}_{SW(k)}) = \tilde{\boldsymbol{x}}_k{}^T\tilde{\boldsymbol{x}}_k - \hat{N}\bar{\bar{x}}_k^2$. It is also SST_{SWm} from partitioning the total sum of squares in Table 3.1 as the total sum of squares, corrected for the mean. The form of $Var_{M2}(\hat{\beta}_{SW_k})$ can then be simplified as:

$$Var_{M2}(\hat{\beta}_{SW_k}) = \frac{(\tilde{\boldsymbol{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{\boldsymbol{x}}}_k)^T \boldsymbol{W}\boldsymbol{V}(\tilde{\boldsymbol{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{\boldsymbol{x}}}_k)}{(\tilde{\boldsymbol{x}}_k{}^T \tilde{\boldsymbol{x}}_k - \hat{N}\bar{\tilde{\boldsymbol{x}}}_k^2)^2} = \frac{(\tilde{\boldsymbol{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{\boldsymbol{x}}}_k)^T \boldsymbol{W}\boldsymbol{V}(\tilde{\boldsymbol{x}}_k - \tilde{\mathbf{1}}\bar{\tilde{\boldsymbol{x}}}_k)}{SST^2_{SW_m}}. \tag{B1.3}$$

Appendix B2

Intercepted-adjusted VIF in OLS

When we take the intercept into account, the equation expression of $\dot{a}^{00}$ in (3.5) can also be written as:

$$\begin{aligned}\dot{a}^{00} &= \frac{1}{SST - (SSM + SSR_m)} \\ &= \frac{1}{SST_m - SSR_m} = \frac{1}{(1 - R_m^2)SST_m}\end{aligned} \tag{B2.1}$$

and hence

$$Var_M(\hat{\beta}_k) = \frac{\sigma^2}{SST_{m(k)}} \frac{1}{(1 - R^2_{m(k)})} \tag{B2.2}$$

where $R^2_{m(k)}$ is the multiple correlation coefficient, corrected for mean, corresponding to the regression of x_k on the $p - 1$ other explanatory variables and $SST_{m(k)}$ is the total sum of squares in this regression, corrected for mean. Therefore, the second term in (B2.2) is the intercept-adjusted VIF in OLS.

Appendix C

Linearization Variance Estimation of $\hat{\boldsymbol{\beta}}$ in Linear and Generalized Linear Models

As shown in Binder (1983), the finite population estimating equations for $\boldsymbol{\beta}$ in either linear models or GLMs have the form $\mathbf{w}_U(\boldsymbol{\beta}) = 0$ where $\mathbf{w}_U$ is a sum over the full finite population. Denote the survey-weighted estimate of $\mathbf{w}_U(\boldsymbol{\beta})$ by $\hat{\mathbf{w}}(\boldsymbol{\beta}) = 0$. For example, in linear regression, $\hat{\mathbf{w}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} w_i \dot{\boldsymbol{x}}_i (y_i - \dot{\boldsymbol{x}}_i^T \boldsymbol{\beta}) = \mathbf{0}$ with $\hat{\boldsymbol{\beta}}_{SW}$ being the solution to the equations. The estimating equations evaluated at $\hat{\boldsymbol{\beta}}_{SW}$ are denoted $\hat{\mathbf{w}}(\boldsymbol{\beta}_{SW})$.

To obtain the linearization variance of $\hat{\boldsymbol{\beta}}_{SW}$, first of all, we take a Taylor expansion of $\hat{\mathbf{w}}(\hat{\boldsymbol{\beta}}_{SW})$, at $\hat{\boldsymbol{\beta}}_{SW} = \boldsymbol{\beta}$:

$$\hat{\mathbf{w}}(\hat{\boldsymbol{\beta}}_{SW}) \doteq \hat{\mathbf{w}}(\boldsymbol{\beta}) + \frac{\partial \hat{\mathbf{w}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} (\hat{\boldsymbol{\beta}}_{SW} - \boldsymbol{\beta}) = \mathbf{0}. \tag{C.1}$$

Solving for $\hat{\mathbf{w}}(\hat{\boldsymbol{\beta}}_{SW})$ and taking the variance leads to

$$\Sigma(\boldsymbol{\beta}) = \left[\frac{\partial \hat{\mathbf{w}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right] Var(\hat{\boldsymbol{\beta}}_{SW}) \left[\frac{\partial \hat{\mathbf{w}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]^T \tag{C.2}$$

where $\Sigma(\boldsymbol{\beta}) = Var[\hat{\mathbf{w}}(\boldsymbol{\beta})]$. Solving for $Var(\hat{\boldsymbol{\beta}}_{SW})$ gives

$$Var(\hat{\boldsymbol{\beta}}_{SW}) = \left[\frac{\partial \hat{\mathbf{w}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]^{-1} \Sigma(\hat{\boldsymbol{\beta}}) \left[\frac{\partial \hat{\mathbf{w}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right]^{-1}. \tag{C.3}$$

In the case of linear regression

$$\frac{\partial \hat{\mathbf{w}}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{n} w_i \dot{\boldsymbol{x}}_i (y_i - \dot{\boldsymbol{x}}_i^T \boldsymbol{\beta}) = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}. \tag{C.4}$$

To estimate $Var(\hat{\boldsymbol{\beta}}_{SW})$, it is necessary to substitute $\hat{\boldsymbol{\beta}}_{SW}$ wherever $\boldsymbol{\beta}$ appears in (C.3). The form of the estimator of $\Sigma(\boldsymbol{\beta})$ depends on the sample design as discussed in Chapter 3 and 5.

Appendix D

Expression of $Blkdiag(\boldsymbol{e}_i\boldsymbol{e}_i^T)$ and $Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)$

Explicitly, $Blkdiag(\boldsymbol{e}_i\boldsymbol{e}_i^T)$ is a matrix with $\boldsymbol{e}_i\boldsymbol{e}_i^T$ on the main diagonal position and 0 elsewhere. In the h^{th} stratum, it can be expressed as:

$$Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T) = \begin{pmatrix} \boldsymbol{e}_{h1}\boldsymbol{e}_{h1}^T & 0 & 0 & \dots & 0 \\ 0 & \boldsymbol{e}_{h2}\boldsymbol{e}_{h2}^T & 0 & \dots & 0 \\ 0 & 0 & \boldsymbol{e}_{h3}\boldsymbol{e}_{h3}^T & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \dots & \boldsymbol{e}_{hn_h}\boldsymbol{e}_{hn_h}^T \end{pmatrix}, \quad i \in s_h, i = 1, \dots, n_h$$

$$= \begin{pmatrix} e_{h11}e_{h11} & \dots & e_{h11}e_{h1m_{h1}} & 0 & \dots & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \dots & \dots & \vdots \\ e_{h1m_{h1}}e_{h11} & \dots & e_{h1m_{h1}}e_{h1m_{h1}} & \vdots & \dots & \dots & 0 \\ 0 & \dots & 0 & \ddots & \dots & \dots & 0 \\ \vdots & \dots & \dots & \vdots & e_{hn_h1}e_{hn_h1} & \dots & e_{hn_h1}e_{hn_hm_{hn_h}} \\ \vdots & \dots & \dots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & e_{hn_hm_{hn_h}}e_{hn_h1} & \dots & e_{hn_hm_{hn_h}}e_{hn_hm_{hn_h}} \end{pmatrix},$$

where m_{hi} stands for the number of selected units in the cluster i which belongs to stratum h.

For example, if there is 2 selected units in each selected cluster and 2 clusters are selected in the h^{th} stratum.

$$Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T) = \begin{pmatrix} \boldsymbol{e}_{h1}\boldsymbol{e}_{h1}^T & 0 \\ 0 & \boldsymbol{e}_{h2}\boldsymbol{e}_{h2}^T \end{pmatrix}, \quad i \in s_h, i = 1, \ldots, 2$$

$$= \begin{pmatrix} e_{h11}e_{h11} & e_{h11}e_{h12} & 0 & 0 \\ e_{h12}e_{h11} & e_{h12}e_{h12} & 0 & 0 \\ 0 & 0 & e_{h21}e_{h21} & e_{h21}e_{h22} \\ 0 & 0 & e_{h22}e_{h21} & e_{h22}e_{h22} \end{pmatrix}.$$

The same definition for $Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T)$, which has $\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T$ on the main diagonal position and 0 elsewhere. We have:

$$Blkdiag(\boldsymbol{e}_{hi}\boldsymbol{e}_{hi}^T) =$$

$$\begin{pmatrix} Blkdiag(\boldsymbol{e}_{1i}\boldsymbol{e}_{1i}^T)_{i\in s_1} & 0 & 0 & \ldots & 0 \\ 0 & Blkdiag(\boldsymbol{e}_{2i}\boldsymbol{e}_{2i}^T)_{i\in s_2} & 0 & \ldots & 0 \\ 0 & 0 & Blkdiag(\boldsymbol{e}_{3i}\boldsymbol{e}_{3i}^T)_{i\in s_3} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \ldots & Blkdiag(\boldsymbol{e}_{Hi}\boldsymbol{e}_{Hi}^T)_{i\in s_H} \end{pmatrix},$$

$i \in s_h, \quad h = 1, \ldots, H.$

Bibliography

Agresti, A. (2002). *Categorical Data Analysis*. Wiley Series in Probability and Statistics. New York: Wiley Interscience, 2nd ed.

Belsley, D. A. (1984a). Collinearity and forecasting. *Journal of Forecasting*, *38*, 73–93.

Belsley, D. A. (1984b). Demeaning conditioning diagnostics through centering. *The American Statistician*, *38*(2), 73–77.

Belsley, D. A. (1991). *Conditioning Diagnostics, Collinearity and Weak Data in Regression*. New York: John Wiley and Sons.

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley Series in Probability and Statistics. New York: Wiley Interscience.

Belsley, D. A., & Oldford, R. W. (1986). The general problem of ill conditioning and its role in statistical analysis. *Computational Statistics and Data*, *4*, 104–120.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, *51*, 279–292.

Chambers, R. L. (1996). Outlier robust finite population estimation. *Journal of the American Statistical Association*, *81*, 1063–1069.

Chambers, R. L., & Skinner, C. J. (2003). *Analysis of Survey Data*. Wiley Series in Survey Methodology. New York: Wiley Interscience.

Chatterjee, S., & Price, B. (1991). *Regression Analysis by Example*. New York: John Wiley, 2nd ed.

Cook, R. D. (1984). Comment on demeaning conditioning diagnostics trhough centering. *The American Statistician*, *2*, 88–90.

Deville, J., & Särndal, C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*, 376–382.

Duchesne, P. (1999). Robust calibration estimators. *Survey Methodology*, *25*, 43–56.

Elliot, M. (2007). Bayesian weight trimming for generalized linear regression models. *Survey Methodology*, *33*, 23–34.

Farrar, D. E., & Glauber, R. R. (1967). Multicollinearity in regression analysis. *Review of Economics and Statistics*, *49*, 92–107.

Fox, J. (1986). *Linear Statistical Models and Related Methods, With Applications to Social Research*. New York: John Wiley.

Fox, J. (1991). *Regression Diagnostics: An introduction*. Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-079. Newbury Park: Sage.

Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Journal Of the American Staitstical Association*, *87*(417), 178–183.

Fuller, W. A. (2002). Regression estimation for survey samples. *Survey Methodology*, *28*(1), 5–23.

Gwet, J., & Rivest, L. (1992). Outlier resistant alternatives to the ratio estimator. *Journal of the American Statistical Association*, *87*, 1174–1182.

Hartley, H. O., & Rao, J. N. K. (1962). Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*, *33*, 350–374.

Hauck, W. W., & Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Jounal of the American Statistical Association*, *72*(360), 851–853.

Hulliger, B. (1995). Outlier robust Horvitz-Thompson estimators. *Survey Methodology*, *21*, 79–87.

Hurvich, C. M., & Tsai, C. (1990). The impact of model selection on inference in linear regression. *The American Statistician*, *44*, 214–217.

Kmenta, J. (1986). *Elements of Econometrics*. New York: Macmillan, 2nd ed.

Lesaffre, E., & Marx, B. D. (1993). Collineraity in generalized linear regression. *Communications in Statistics-Theory and Methods*, *22*(7), 1933–1952.

Li, J. (2007a). Linear regression diagnostics in cluster samples. *ASA Proceedings of the Section on Survey Research Methods*, (pp. 3341–3348).

Li, J. (2007b). Regression diagnostics for complex survey data. Unpublished doctoral dissertation, University of Maryland.

Li, J., & Valliant, R. (2006). Influence analysis in linear regression with sampling weights. *ASA Proceedings of the Section on Survey Research Methods*, (pp. 3330–3337).

Li, J., & Valliant, R. (2009). Survey weighted hat matrix and leverages. *Survey Methodology*, *35*(1), 15–24.

Lin, C. (1984). Extrema of quadratic forms and statistical applications. *Communications in Statistics-Theory and Methods*, *13*, 1517 – 1520.

Little, R. (2004). To model or not to model? competing modes of inference for finite population sampling. *Journal of American Statistical Association*, *99*(466), 546–556.

Mackinnon, M. J., & Puterman, M. L. (1990). Collinearity in generalized linear models. *Communications in Statistics-Theory and Methods*, *18*(9), 3463–3472.

Manderscheid, R. W., & Henderson, M. J. (2002). Mental health, united states, 2002. dhhs publication no. sma04-3938. *Survey Methodology*. Available at: http : //mentalhealth.samhsa.gov/publications/allpubs/SMA04 − 3938/ AppendixA.asp.

Marquardt, D. W. (1970). Generalized inverses ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics*, *12*, 591–612.

Marquardt, D. W. (1980). A critique on some ridge regression methodsl comment 'you should standardize the predictor variables in your regression models'. *Journal of the American Statistical Association*, *75*(369), 87–91.

Marx, B. D., & Smith, E. P. (1990a). Principal component estimation for generalized linear regression. *Biometrika*, *77*(1), 23–32.

Marx, B. D., & Smith, E. P. (1990b). Weighted multicollinearity in logistic regression: diagnostics and biased estimation techniques with an example from lake acidification. *Canadian Journal of Fisheries and Aquatic Sciences*, *47*(6), 1128–1135.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman and Hall, 2nd ed.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. Wiley Series in Probability and Statistics. New York: John Wiley & Sons, Inc.

Moreno-Rebollo, M. n.-R. A., J. L., & Muñoz Pichardo, J. (1999). Influence diagnostic in survey sampling: Conditional bias. *Biometrika*, *86*, 923–928.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality and Quantity*, *41*, 673–690.

Pfeffermann, D., & Sverchkov, M. (2003). Fitting generalized linear models under informative sampling. In R. Chambers, & C. Skinner (Eds.) *Analysis of Survey Data*, chap. 12, (pp. 175–195). Wiley Series in Survey Methodology.

Potter, F. J. (1990). A study of procedures to identify and trim extreme sampling weights. *ASA Proceedings of the Section on Survey Research Methods*, (pp. 225–230).

Potter, F. J. (1993). The effect of weight trimming on nonlinear survey estimates. *ASA Proceedings of the Section on Survey Research Methods*, (pp. 758–763).

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Schaefer, R. L. (1986). Alternative estimators in logistic regression when the data are collinear. *Journal of Statistical Computations and Simulations*, *25*, 75–91.

Schaefer, R. L., Roi, L. D., & Wolfe, R. A. (1984). A ridge logistic estimator. *Communications in Statistics-Theory and Methods*, *13*(1), 99–113.

Shao, J. (2003). *Mathematics Statistics*. Springer, 2nd ed.

Silvey, S. D. (1969). Multicollinearity and imprecise estimation. *Journal of the Royal Statistical Society*, *31*.

Simon, S. D., & Lesage, J. P. (1988). The impact of collinearity involving the intercept term on the numerical accuracy of regression. *Computer Science in Economics and Management*, *1*, 137–152.

Skinner, C. J., Holt, D., & Smith, T. M. F. (1989). *Analysis of Complex Surveys*. Wiley Series in Survey Methodology. Wiley Interscience.

Snee, R. D., & Marquardt, D. W. (1984). Collinearity diagnostics depend on the domain of prediction, and model, and the data. *The American Statistician*, *2*, 88–90.

Steward, G. W. (1987). Collinearity and least squares regression. *Statistical Science*, *2*(1), 68–84.

Theil, H. (1971). *Principles of Econometrics*. New York: John Wiley.

Væth, M. (1985). On the use of Wald's test in exponential families. *International Statistical Review*, *53*(2), 199–214.

Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: John Wiley.

Weissfeld, L. A., & Sereika, S. M. (1991). A multicollinearity diagnostic for generalized linear models. *Communications in Statistics-Theory and Methods*, *20*(4), 1183–1198.

Welsh, A. H., & Ronchetti, E. (1998). Bias-calibrated estimation from sample surveys containing outliers. *Journal of the Royal Statistical Society: Series B, Methodological*, *60*.

Wissmann, M., Toutenburg, H., & Shalabh (2007). Role of categorical variables in multicollinearity in the linear regression model. Technical Report Number 008, Department of Statistics, University of Munich. Available at `http://epub.ub.uni-muenchen.de/2081/1/report008_statistics.pdf`.

Wood, F. S. (1984). Effect of centerng on collinearity and interpretation of the constant. *The American Statistician*, *2*, 88–90.

Zhang, P. (1992). Influence after variable selection in linear regression models. *Biometrica*, *79*, 741–746.