ABSTRACT

Title of dissertation: INVESTIGATION OF ALTERNATIVE CALIBRATION ESTIMATORS IN THE PRESENCE OF NONRESPONSE

Daifeng Han, Doctor of Philosophy, 2017

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

Calibration weighting is widely used to decrease variance, reduce nonresponse bias, and improve the face validity of survey estimates. In the purely sampling context, Deville & Särndal (1992) demonstrate that many alternative forms of calibration weighting are asymptotically equivalent, so for variance estimation purposes, the generalized regression (GREG) estimator can be used to approximate some general calibration estimators with no closed-form solutions such as raking. It is unclear whether this conclusion holds when nonresponse exists and single-step calibration weighting is used to reduce nonresponse bias (i.e., calibration is applied to the basic sampling weights directly without a separate nonresponse adjustment step).

In this dissertation, we first examine whether alternative calibration estimators may perform differently in the presence of nonresponse. More specifically, properties of three widely used calibration estimations, the GREG with only main effect covariates

(GREG_Main), poststratification, and raking, are evaluated. In practice, the choice between poststratification and raking are often based on sample sizes and availability of external data. Also, the raking variance is often approximated by a linear substitute containing residuals from a GREG_Main model. Our theoretical development and simulation work demonstrate that with nonresponse, poststratification, GREG_Main, and raking may perform differently and survey practitioners should examine both the outcome model and the response pattern when choosing between these estimators. Then we propose a distance measure that can be estimated for raking or GREG_Main from a given sample. Our analytical work shows that the distance measure follows a Chi-square probability distribution when raking or GREG_Main is unbiased. A large distance measure is a warning sign of potential bias and poor confidence interval coverage for some variables in a survey due to omitting a significant interaction term in the calibration process. Finally, we examine several alternative variance estimators for raking with nonresponse. Our simulation results show that when raking is model-biased, none of the linearization variance estimators under evaluation is unbiased. In contrast, the jackknife replication method performs well in variance estimation, although the confidence interval may still be centered in the wrong place if the point estimate is inaccurate.

INVESTIGATION OF ALTERNATIVE CALIBRATION ESTIMATORS IN THE PRESENCE OF NONRESPONSE

by

Daifeng Han

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
2017

Advisory Committee:
Dr. Richard Valliant, Chair/Advisor
Dr. Katharine G. Abraham
Dr. Ernesto Calvo
Dr. Jill DeMatteis
Dr. Keith Rust

DEDICATION

To my husband and better half, Bill Harden, for the many moments in the past few years that he kept his faith and sense of humor when I lost mine. You have always been there for me and showed me what unconditional love is.

To my son, Manny, who has expanded my horizon, stretched my limit, and forced me to strike a better balance between career and family. Seeing how you learn and grow every day inspires me to be a life-long learner and better person.

To my parents, Keming Han and Jinghua Li, who are among the most hardworking people I know. Thank you both for having brought me to this fascinating world and instilled within me good work ethics.

To my late husband, Huiping Xiang, whose sudden death about a decade ago still frequently reminds me how fragile life can be and how much I should cherish each day. Had I not had you, or had I not lost you, I would not be the person I am today. Thank you for having taught me so much and planted the seeds of continuous growth in my heart. Those seeds are now turning into blooming flowers.

ACKNOWLEDGMENTS

I am indebted to many people who have helped me in various ways during the process of my working on this dissertation.

I would like to express my deepest appreciation for my dissertation chair, Dr. Richard Valliant, for his tremendous dedication, patience, and support in every step of this journey, especially when I struggled to overcome the obstacles along the way.

I thank all the other members of my committee, Dr. Katharine G. Abraham, Dr. Ernesto Calvo, Dr. Jill DeMatteis, and Dr. Keith Rust, for their insightful advice. Special thanks to Dr. Rust and Dr. DeMatteis, who provided valuable specific suggestions on how to improve the research.

In addition, a deep-felt thank you to my colleagues and mentors at Westat, David Morganstein, Mike Brick, and Sharon Lohr, who showed me how interesting the world of survey statistics and methodology is, encouraged me to pursue this Ph.D. degree, and inspired me to learn continuously through real-world practice.

Finally, many colleagues and friends (unnamed here) have cheered me up as I faced critical life changes in more than a decade. It is their support and encouragement that have pulled me through in the moments I felt exhausted and had self-doubt. I cannot be more grateful for all these colleges and friends.

Table of Contents

List of Tables

List of Figures

Chapter 1. Literature Review

Calibration weighting was originally developed as a method for reducing sampling errors while retaining randomization consistency. Deville and Särndal (1992) introduce calibration estimators using the distance function approach. Later work by Särndal (2007) points out that there are two different approaches to take account of auxiliary information in estimation – a "calibration approach" and a "regression approach". The two approaches generate the same estimator, the generalized-regression (GREG) estimator, in the situation where the general least squares (GLS) distance function is used in the calibration approach and linear regression model is used in the regression approach. For the purpose of comparison, we use the term "general calibration estimators" to refer to the other estimators in the calibration estimator family covered by Deville and Särndal (1992), as opposed to the GREG estimator.

Although almost all surveys in practice are subject to frame deficiencies and nonresponse, the theories in Deville and Särndal (1992) are developed for the ideal situation where non-sampling errors do not exist. In this context (i.e., in the situation where non-sampling errors do not exist), Deville and Särndal (1992) show that many alternative forms of calibration weighting are asymptotically identical. This leads to a breakthrough in our understanding of some commonly used calibration estimators that do not have closed-form solutions, such as raking. As a result, the GREG estimator is often considered a good approximation of the general calibration estimators. However, non-sampling errors such as nonresponse almost always exist in real-world surveys. In the past decade,

Särndal and Lundström (1999, 2005), Kott (2006), Chang and Kott (2008), Kott and Chang (2010), and Kott and Liao (2012) have proposed different methods for using calibration to reduce nonresponse bias through one-step weighting, yet we still lack understanding of the empirical properties of the calibration estimators generated by these methods. For example, it is unclear whether the GREG estimator and the general calibration estimators are asymptotically equivalent when nonresponse is present in a survey and single-step calibration weighting is used to reduce potential nonresponse bias (i.e., calibration is applied to the basic sampling weights directly without a separate nonresponse adjustment step). In practice, the poststratification estimator (as a special case of the GREG estimator) and the raking estimator (as an example of the general calibration estimator) are both widely used in the government-sponsored surveys in the United States and European countries. Quite often survey practitioners choose between these two estimators based on the availability of the benchmark totals and the case counts in the survey requiring calibration. Such a decision rule is fully justifiable only if poststratification and raking can reduce nonresponse bias to a similar extent. However, no systematic research has been conducted on comparing the performance of the poststratification estimator and the raking estimator when calibration is used to correct nonresponse bias.

Our research expands the literature by relaxing the assumption of no non-sampling error. To keep the picture simple, we assume that the sampling frame has perfect coverage and there is no measurement error in surveys, so we can focus on the non-sampling error caused by nonresponse. Our goal is to evaluate the properties of some calibration

estimators when calibration is used to reduce nonresponse bias through a one-step weighting approach. The rest of this chapter is organized as follows: Sections 1.1 through 1.3 summarize the research on the properties of various calibration estimators, particularly those proposed by Deville and Särndal (1992), in the absence of nonresponse. Sections 1.4 describes the alternative single-step calibration methods in the literature. Section 1.5 explains the importance of choosing auxiliary variables to effectively reduce nonresponse bias. Section 1.6 points out the gaps in the existing literature and describes our research aims.

1.1 Two Approaches to Incorporate Auxiliary Information in Estimation

There are two systematic ways to take account of auxiliary information in estimation, labeled as the "regression approach" and the "calibration approach", although the distinction may not be completely clear-cut (Särndal 2007). In their original definition of the calibration estimator, Deville and Särndal (1992) require "minimum distance" between the calibration weights and the original sampling weights, subject to satisfying the calibration equation. In general, the term "calibration approach" often refers to creating estimators by benchmarking the auxiliary information to external controls.

Let y_k be the value of the variable of interest, *y*, for the *k*th population element, which is associated with a vector of auxiliary variables $\mathbf{x}_k = (x_{k1},\ldots x_{kp},\ldots,x_{kP})^{\mathrm{T}}$. For the elements $k \in s$, where s is the set of sample elements, we observe $(y_k, \mathbf{x}_k)$. For

simplicity, the population total of $\mathbf{x}$, $\mathbf{t}_x = \sum_U \mathbf{x}_k$, which is often referred to as the benchmark control vector, is assumed to be accurately known.

The objective is to estimate the population total $t_y = \sum_U y_k$. Let d_k be the basic sampling design weight calculated as the inverse of the inclusion probability π_k . The Horvitz-Thompson estimator is $\hat{t}_{y\pi} = \sum_s y_k / \pi_k = \sum_s d_k y_k$. The calibration estimator is defined as $\hat{t}_{yw} = \sum_s w_k y_k$, with weights w_k as close as possible, in an average sense based on a distance function, to the basic design weights d_k while respecting the calibration equation

$$\sum_s w_k \mathbf{x}_k = \mathbf{t}_x \tag{1.1}$$

Under a chosen distance function $G_k(w_k, d_k)$, this becomes an optimization problem. The goal is to find a set of weights $\{w_k\}_{k\in s}$ that minimizes $\sum_{k\in s} G_k(w_k, d_k)$ subject to (1.1). This leads to the Lagrange function

$$\Psi = \sum_{k\in s} G\left(w_k, d_k\right) + \boldsymbol{\lambda}^{\mathrm{T}}\left(\mathbf{t}_x - \sum_{k\in s} w_k \mathbf{x}_k\right) \tag{1.2}$$

which is minimized to find the optimal set of weights $\{w_k\}_{k\in s}$.

The calibration weights can be expressed as

$$w_k = d_k F_k(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}) \tag{1.3}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_P)^{\mathrm{T}}$ is the vector of Lagrange multipliers determined from (1.2). $\boldsymbol{\lambda}$ corresponds to a realized sample, but for simplicity we often use $\boldsymbol{\lambda}$ as the shorthand for $\boldsymbol{\lambda}_s$. $F_k(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda})$ is the inverse function of $g_k(w_k, d_k) = \partial G_k(w_k, d_k)/\partial w_k$, the first partial derivative of the distance function taken with respect to the calibrated weight. $F_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right)$ uniquely corresponds to $G_k(w_k, d_k)$. It is assumed that F_k is non-negative and convex, and that $F_k(0) = 1$, implying that when $w_k = d_k$ the distance between the basic design weights and calibrated weights is zero. Moreover, it is required that F_k' is continuous, one-to-one, and that $F_k'(1) = 0$ and $F_k''(1) > 0$, which makes $w_k = d_k$ a local minimum.

The Horvitz-Thompson estimator of $\mathbf{t}_x$ is $\hat{\mathbf{t}}_{x\pi} = \sum_s d_k \mathbf{x}_k$, so the calibration equation can be expressed as

$$\sum_s d_k F_k(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda})\mathbf{x}_k - \sum_s d_k \mathbf{x}_k = \mathbf{t}_x - \hat{\mathbf{t}}_{x\pi} \tag{1.4}$$

Define

$$\Phi_s(\boldsymbol{\lambda}) = \sum_s d_k \left\{F_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right) - 1\right\}\mathbf{x}_k \tag{1.5}$$

Then (1.4) can be written as

$$\Phi_s(\boldsymbol{\lambda}) = \mathbf{t}_x - \hat{\mathbf{t}}_{x\pi} \tag{1.6}$$

The task of obtaining w_k boils down to solving (1.6) for $\boldsymbol{\lambda}$. The calibration estimator of t_y is

$$\hat{t}_{yw} = \sum_s w_k y_k = \sum_s d_k F_k(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}) y_k \tag{1.7}$$

Depending on the distance function $G_k(w_k, d_k)$, iteration may be required to obtain a solution for $\boldsymbol{\lambda}$. With full response, the Horvitz-Thompson estimator $\hat{t}_{y\pi}$ using basic sampling weights d_k is unbiased. If the calibration weights w_k are as close as possible, according to $G_k(w_k, d_k)$, to the basic sampling weights d_k, then a realistic expectation is that the calibration weights will maintain near unbiasedness.

Although several distance functions are discussed in Deville and Särndal (1992), most theoretical research has focused on the GLS distance function $\sum_s (w_k - d_k)^2 / d_k q_k$, where $1/q_k$ is the positive weight associated with the *k*th term and is unrelated to d_k. With this distance function, the calibration equation has a closed-form solution. We obtain $\mathrm{F}_k(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}) = 1 + q_k \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}$, and the calibration estimator is the GREG estimator

$$\hat{t}_{yreg} = \sum_s d_k (1 + q_k \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}) y_k = \hat{t}_{y\pi} + (\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})^{\mathrm{T}} \hat{\mathbf{B}}_s \tag{1.8}$$

where

$$\boldsymbol{\lambda} = \mathbf{T}_s^{-1}(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}) \tag{1.9}$$

$$\hat{\mathbf{B}}_s = \mathbf{T}_s^{-1} \sum_s d_k q_k \mathbf{x}_k y_k \tag{1.10}$$

$$\mathbf{T}_s = \sum_s d_k q_k \mathbf{x}_k \mathbf{x}_k^{\mathrm{T}} \tag{1.11}$$

An alternative method for obtaining the calibration estimator is referred to as the "regression approach". With the regression approach, estimators are calculated by using an assisting model that closely represents the relationship between the outcome variable and the auxiliary variables. The assisting model is also referred to as the calibration model or the working prediction model by Kott (2006) to distinguish it from other models such as those used to address response propensity. The assisting model can have linear or nonlinear forms. When the assisting model is a linear regression model, the weight happens to be calibrated to the auxiliary controls and the estimator (which is the GREG estimator) is expressible as a linearly weighted sum with calibrated weights as a by-product. One advantage of the GREG estimator is that the calibrated weights are independent of any particular outcome variable y and can therefore be applied to all the variables of interest in a survey (Särndal, Swensson, and Wretman, 1992).

In summary, the central idea for the regression approach is to find an assisting model that fits the population data well. In practice, we are often interested in estimating totals for a number of survey variables, and it is unreasonable to assume that different outcome variables fit the same model. This is probably why survey statisticians often adopt a model-assisted approach rather than a model-based approach. In contrast, the calibration approach does not refer explicitly to any models, but emphasizes the linear weighting of the observed y values with weights made to confirm computable aggregates. The resulting calibrated weights are functions only of the auxiliary variables and not any of the outcome variables, so one set of final analysis weights is created instead of requiring weights specific to each variable within a set of key outcome variables. The two

approaches generate the same estimator, the GREG estimator, under the special situation where the GLS distance function is used in the calibration approach and linear regression is used in the regression approach (Särndal 2007).

Our research adopts the perspectives of both approaches. The weights are primarily justified by their consistency with the benchmark controls (which is the calibration approach). Although the calibration approach does not refer explicitly to any assisting models, we demonstrate that the performance of a calibration estimator in the presence of nonresponse depends on the choice of auxiliary vector and/or function form used in the calibration process, and this requires a modeling effort in some sense.

1.2 Distance Function Method versus Function Form Method

Under the umbrella of the calibration approach, two methods are discussed in the literature. Deville and Särndal (1992) initially require that the set of calibration weights $\{w_k\}_{k\in s}$ minimize some distance function $\sum_{k\in s} G_k(w_k, d_k)$ subject to satisfying the calibration equation — this is the "distance function method" described in Section 1.1. An alternative approach is the "function form method" (Estevao and Särndal 2006) or "instrument vector method" (Kott 2006). Just as the distance function approach can result in different sets of weights associated with different distance functions, the function form method can generate alternative sets of weights calibrated to the same auxiliary information using different function forms.

The function form method removes the limitation that the calibration weights minimize a distance function, and requires only that $\{w_k\}_{k\in s}$ satisfy the calibration equation and be of the function form $w_k = d_k F(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\lambda})$, where d_k is the design weight, and $\mathbf{z}_k$ is a vector with values defined for the units in the sample and sharing the dimension of the specified benchmark control vector $\mathbf{x}_k$. The vector $\mathbf{z}_k$ can be a specified function of $\mathbf{x}_k$ or of other background data about unit *k* (Särndal and Lundström 2005). The vector $\boldsymbol{\lambda}$ is determined from the calibration equation. The function $F(\cdot)$ plays a similar role as $G_k(w_k, d_k)$ does in the distance minimization method. For easy reference, we refer to $F(\cdot)$ as "weight adjustment function" or "adjustment function" in our research. One possible form of the weight adjustment function is $w_k = d_k(1+\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\lambda})$, and the corresponding calibration estimator is

$$\hat{t}_{ycal} = \sum_s d_k(1+\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\lambda})y_k \tag{1.12}$$

where

$$\boldsymbol{\lambda} = (\sum_s d_k \mathbf{x}_k \mathbf{z}_k^{\mathrm{T}})^{-1}(\mathbf{t}_x - \sum_s d_k \mathbf{x}_k) \tag{1.13}$$

The GREG estimator $\hat{t}_{yreg}$ defined in (1.8) is a special case of (1.12) obtained for $\mathbf{z}_k = q_k \mathbf{x}_k$.

We think that when nonresponse exists in a survey, it is more appropriate to understand the calibration process using the function form method rather than the distance function method. This is because in the presence of nonresponse, the Horvitz-Thompson estimator for the total of an outcome variable *y* using the basic design weights becomes

$\hat{t}_{y\pi} = \sum_r d_k y_k$, where r represents the responding set. This estimator is biased when $r \neq s$. If the calibration process aims to correct the nonresponse bias, it is neither necessary nor appropriate to require the calibrated weights to be "as close as possible" to the basic design weights based on a distance function.

More discussions about the weighting adjustment function $F(\cdot)$ are included in Section 1.3. When applying the function form method, survey practitioners face some practical questions. For example, is there any advantage to make the vector $\mathbf{z}_k$ in the weighting adjustment function $F(\cdot)$ differ from the calibration vector $\mathbf{x}_k$? How should the variables be chosen to include in $\mathbf{x}_k$ and $\mathbf{z}_k$? These questions have not been clearly answered by the existing literature. Särndal (2007, Section 4.3) gives an example showing that "even 'deliberately awkward choices' for $\mathbf{z}_k$ give surprisingly good results". However, the property of near-unbiasedness of the calibration estimator in this situation seems to depend on the assumption of no non-sampling error, which usually does not hold in practice.

1.3 Relationship between GREG Estimator and General Calibration Estimators in the Absence of Nonresponse Error

As described in Section 1.1, various calibration estimators can be derived with the aid of different distance measures under the same set of constraints on the auxiliary variables. Alternative distance functions are compared in Deville, Särndal, and Sautory (1993), Singh and Mohl (1996), and Stukel, Hidiroglou, and Särndal (1996). When there is no

non-sampling error, there are usually very small differences between the point estimates corresponding to the various distance functions, and changes in the distance function often have only a minor effect on the variance of the calibration estimator even if the sample size is rather small. The GREG estimator and the other members of the calibration estimator family (referred to as the "general calibration estimators") are compared in Deville and Särndal (1992). They conclude that the GREG estimator is a first approximation to the general calibration estimators, all the general calibration estimators are asymptotically equivalent to the GREG, and the variance estimator for the GREG can be used for the general calibration estimators.

Although the GREG estimator is a special case of the calibration estimator family when the function form is $F(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}) = 1 + q_k \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}$, we use $\hat{t}_{yreg}$ to denote the GREG estimator and $\hat{t}_{yw}$ to denote the other calibration estimators (i.e., the general calibration estimators) for the purpose of comparison.

Deville and Särndal (1992) consider a sequence of finite populations and sampling designs indexed by *n*, where *n* is the sample size (for a fixed-sized sampling design) or the expected sample size (for a random-sized sample design). The finite population size, *N*, tends to infinity with *n*. Several assumptions are made about the auxiliary vector **x**: (i) $\lim N^{-1}\mathbf{t}_x$ exists; (ii) $N^{-1}\left(\hat{\mathbf{t}}_{x\pi} - \mathbf{t}_x\right) = O_p(n^{-1/2})$, where the subscript p means probability induced by repeated sampling; and (iii) $n^{1/2}N^{-1}(\hat{\mathbf{t}}_{x\pi} - \mathbf{t}_x)$ converges in distribution to the multinormal $N(\mathbf{0}, \mathbf{A})$ where **A** is a covariance matrix**.** Two additional assumptions are

also made for proving their Results 3 through 5 (described in next paragraph): (iv) $\max\|\mathbf{x}_k\| = M < \infty$, where max is over n as well as over k; and (v) $\max F_k''(0) = M' < \infty$. Assumptions (i) through (iii) have two practical implications. First, the components of $N^{-1}\left(\hat{\mathbf{t}}_{x\pi} - \mathbf{t}_x\right)$ are considered small and qualities on the order of $N^{-2}\left\|\hat{\mathbf{t}}_{x\pi} - \mathbf{t}_x\right\|^2$ are considered negligible. Second, $\hat{\mathbf{t}}_{x\pi} - \mathbf{t}_x$ follows an approximately normal distribution with covariance matrix $n^{-1}N^2\mathbf{A}$ (where $\mathbf{A}$ can be viewed as a matrix that describes an asymptotic effect of the sampling design used for the survey), and this is to justify the use of the normal approximation in confidence intervals based on the point estimator. Assumption (iv) is usually satisfied in practice since covariates are bounded. Assumption (v) is verified for all the calibration estimators given in Deville and Särndal (1992).

Deville and Särndal (1992) show five results. Result 1 states that the calibration equation (1.6) has a unique solution belonging to an open neighborhood of $\mathbf{0}$, with probability tending to 1 as $n \to \infty$. Results 2 and 3 are about the magnitude of the Lagrange multiplier. They prove that $\boldsymbol{\lambda}_s = \mathbf{T}_s^{-1}\left(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}\right) + O_p\left(n^{-1}\right) = O_p(n^{-1/2})$, where $\mathbf{T}_s = \sum_s d_k q_k \mathbf{X}_k \mathbf{X}_k^{\mathrm{T}}$. So $\boldsymbol{\lambda}_s$ tends to $\mathbf{0}$ in design probability as $n \to \infty$. Result 4 indicates that the general calibration estimators are design-consistent, and the difference between the general calibration estimators and the Horvitz-Thompson estimator is asymptotically zero. That is, $N^{-1}\left(\hat{t}_{yw} - \hat{t}_{y\pi}\right) = O_p(n^{-1/2})$. Result 5 compares the general calibration estimators with the GREG estimator. For any weight adjustment function $F_k(\cdot)$ obeying their assumptions, $\hat{t}_{yw}$ given by equation (1.7) is asymptotically equivalent to the GREG

estimator given by equation (1.8), in the sense that $N^{-1}\left(\hat{t}_{yw}-\hat{t}_{yreg}\right)=O_p\left(n^{-1}\right)$. Results 4 and 5 together show that as $n \to \infty$, the difference between the general calibration estimators and the GREG estimator approaches zero faster than the difference between the general calibration estimators and the Horvitz-Thompson estimator. The asymptotic variance of $\hat{t}_{yw}$ is, thus, the same as that of the GREG estimator. The proofs for these five results are summarized in Appendix A.

These results have important practical implications because some general calibration estimators do not have a closed-form solution. For example, although the raking ratio estimator has a long history of use in survey practice, the variance of the raking estimator is difficult to derive even approximately. Deville and Särndal (1992) resolve the problem by using the property that the general calibration estimators and the GREG estimator are asymptotically equivalent. Thus, the large-sample variance of the raking ratio estimator can be calculated using the same formula as that for the GREG estimator, given in Särndal, Swensson, and Wretman (1992).

It is important to note that all the results in Deville and Särndal (1992) are derived under the assumptions i) $N^{-1}\left(\hat{\mathbf{t}}_{x\pi}-\mathbf{t}_x\right)=O_p(n^{-1/2})$; and ii) $n^{1/2}N^{-1}(\hat{\mathbf{t}}_{x\pi}-\mathbf{t}_x)$ converges in distribution to a multinormal distribution with mean of $\mathbf{0}$. That is, they require the Horvitz-Thompson estimator of the population total of the auxiliary vector $\mathbf{x}$ with the basic design weights to be approximately unbiased and consistent. This unbiasedness assumption is true in the purely sampling context, i.e., one uncontaminated by

nonresponse or undercoverage error. When non-sampling errors exist, the unbiasedness assumption above does not hold anymore, so it is unclear whether the GREG estimator is still asymptotically equivalent to other calibration estimators.

1.4 Calibration for Nonresponse Bias Reduction

There are several variations in the literature on how to adjust for nonresponse and calibrate the weights to benchmark controls. The conventional approach uses auxiliary information in two steps (Kalton and Flores-Cervantes 2003). In step (i), a response model is formed based on the patterns of correlation between the response probabilities and available auxiliary variables. The aim is to derive good proxies of the unknown response probabilities, so as to limit the nonresponse bias as much as possible. In step (ii), the goal is to select the auxiliary variables that best meet the dual purpose of reducing the sampling variance and of giving added protection against nonresponse bias. An alternative approach is to skip explicitly estimating the response propensity, but use calibration for nonresponse adjustment directly. The basic design weights are modified in a single step with two simultaneous goals: to reduce the nonresponse bias and to ensure the consistency between survey estimates and known population totals. The single-step weighting approach has the potential to simplify the derivation of the variance estimation formulas, so we adopt this approach in our work.

1.4.1 Alternative Single-Step Weighting Methods

A single-step weighting approach through calibration is first proposed by Särndal and Lundström (1999, 2005). The literature is expanded by Kott (2006), Chang and Kott (2008), Chang and Kott (2010), Kott and Liao (2012), and D'Arrigo and Skinner (2010) in the past decade.

Särndal & Lundström Method

In the Särndal & Lundström method, auxiliary controls can be available at the population level, the sample level, or both. At the level of the population *U*, let $\mathbf{x}_k^*$ denote a vector of dimension J^* such that the population vector total $\sum_U \mathbf{x}_k^*$ is known and for every $k \in r$ (where r is the set of respondents), the vector value $\mathbf{x}_k^*$ is known. At the level of the sample *s*, let $\mathbf{x}_k^o$ denote a vector of dimension J^o such that for every $k \in s$, the vector value $\mathbf{x}_k^o$ is known. During calibration, all the auxiliary controls from the population and/or the sample are included in the calibration equation, with the dual purpose of reducing both sampling error and nonresponse bias. The auxiliary vector $\mathbf{x}_k = \begin{pmatrix} \mathbf{x}_k^* \\ \mathbf{x}_k^o \end{pmatrix}$ has dimension $J^* + J^o$. The corresponding information input is $\mathbf{t}_x = \begin{pmatrix} \sum_U \mathbf{x}_k^* \\ \sum_s d_k \mathbf{x}_k^o \end{pmatrix}$. We seek a weighting system w_k for $k \in r$ that satisfies the calibration equation $\sum_r w_k \mathbf{x}_k = \mathbf{t}_x$. The calibrated weights are $w_k = d_k v_k$, where v_k corresponds to the weighting adjustment function $F(\cdot)$ described in Section 1.2 and can take different forms.

Although the distance function method is used in Lundström and Särndal (1999) for obtaining w_k, their later work adopts the function form method, which seems more appropriate when nonresponse is present and calibration is used to correct nonresponse bias. The calibration equation poses only weak constraint on the weights. Depending on the form v_k takes, there exist many sets of calibrated weights for a given auxiliary vector $\mathbf{x}_k$. Särndal and Lundström (2005) discuss two alternative schemes for defining the function form for v_k: (i) as a function of the auxiliary vector $\mathbf{x}_k$; and (ii) as a function of any vector $\mathbf{z}_k$ specified for $k \in r$ and with the same dimension as $\mathbf{x}_k$.

Under scheme (i), v_k should reflect the known individual characteristics of the element $k \in r$, summarized by the vector value $\mathbf{x}_k$. The calibration equation can be expressed as $\sum_r d_k F(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r) = \mathbf{t}_x$, where $\boldsymbol{\lambda}_r$ is a vector to be determined through the calibration equation. A simple function form is recommended that depends linearly on $\mathbf{x}_k$: $F(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r) = 1 + \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r$, where $\boldsymbol{\lambda}_r = (\sum_r d_k \mathbf{x}_k \mathbf{x}_k^{\mathrm{T}})^{-1}(\mathbf{t}_x - \sum_r d_k \mathbf{x}_k)$. An alternative scheme (i.e, scheme (ii)) is to define the weighting adjustment function using a vector $\mathbf{z}_k$ specified for $k \in r$ and with the same dimension as $\mathbf{x}_k$. The vector $\mathbf{z}_k$ can be a specified function of $\mathbf{x}_k$ or any background data about k. Only a linear function form based on $\mathbf{z}_k$ is considered by Särndal and Lundström (2005). The calibrated weights are $w_k = d_k(1 + \mathbf{z}_k^{\mathrm{T}}\boldsymbol{\lambda}_r)$, where $\boldsymbol{\lambda}_r = (\sum_r d_k \mathbf{x}_k \mathbf{z}_k^{\mathrm{T}})^{-1}(\mathbf{t}_x - \sum_r d_k \mathbf{x}_k)$. Särndal and Lundström (2005)

call the vector $\mathbf{z}_k$ an "instrument vector" for the calibration, but do not explain why an instrument vector $\mathbf{z}_k$ may be more desirable than the vector $\mathbf{x}_k$. Besides alerting the reader to this generality of the calibration approach, Särndal and Lundström (2005) give little information about how to choose $\mathbf{z}_k$ except to suggest that $\mathbf{z}_k = \mathbf{x}_k$ is the "standard choice". This gap is filled by some later work by Chang and Kott (2008), Kott and Chang (2010), and Kott and Liao (2012).

We can see that scheme (ii) is the generalization of scheme (i) in Särndal and Lundström (2005). When $\mathbf{z}_k = \mathbf{x}_k$, the two schemes give identical estimators. Furthermore, when $r = s$ (indicating full response) and $\mathbf{x}_k = \mathbf{x}_k^*$ (meaning that the auxiliary vector contains information only from external benchmarks and not from the sampling frame), the calibration estimator $w_k = d_k(1 + \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r)$ and the GREG estimator defined in equation (1.8) are identical.

Kott & Chang Method

Recent developments by Kott (2006), Chang and Kott (2008), Kott and Chang (2010), and Kott and Liao (2012) emphasize two possibilities: 1) the set of variables modeling the response mechanism (referred to as "model variables") being divergent from the benchmark variables in the calibration equation; and 2) using a nonlinear calibration weighting procedure to implicitly estimate a logistic response model. The vector for the benchmark controls in the calibration equation is still $\mathbf{x}_k$, with known population totals

$\mathbf{t}_x$. Unit nonresponse is viewed as an additional phase of Poisson sampling. Using the quasi-randomization perspective, each element k in the original sample is assumed to have a response probability $p_k(\cdot)$, which is a function of the response model covariate vector $\mathbf{z}_k$. Some components of the response-model vector $\mathbf{z}_k$ governing the unit response mechanism need not coincide with the components on the calibration vector $\mathbf{x}_k$. The components of $\mathbf{z}_k$ that are not components of $\mathbf{x}_k$ are called instrument variables. The reason to use a vector $\mathbf{z}_k$ that may be different from $\mathbf{x}_k$ is that sometimes the variables the response mechanism depends on are known only for respondents, not for the whole sample. For example, in an agriculture survey, the benchmark variables can be previous-census frame variables known for every farm in the population while the response model covariates are current-period variables known only for survey respondents. Kott (2006) still requires that the dimensions of $\mathbf{z}_k$ and $\mathbf{x}_k$ coincide. Chang and Kott (2008) expand the method such that it allows the number of benchmark variables (i.e., the dimension of $\mathbf{x}_k$) to exceed the number of response model covariates (i.e., the dimension of $\mathbf{z}_k$).

The statisticians can specify the function form for the response probability $p_k(\cdot)$ and the unknown parameters in the function can be estimated during the calibration process. Although in theory, the response propensity $p(\cdot)$ can take different forms, Kott and Chang's discussions are restricted to linear function of the response model covariates. For example, the response propensity for each responding unit k can be specified as $p(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\beta})$,

an unknown but estimable linear combination of the response model covariate vector $\mathbf{z}_k$. The input weight for the calibration equation is calculated as the product of basic design weight d_k and $1/p(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\beta})$, where the vector $\boldsymbol{\beta}$ can be estimated from the data using the calibration equation $\mathbf{t}_x = \sum_{k\in r} \frac{d_k}{p(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\beta})}\mathbf{x}_k$ through a nonlinear calibration process. This equation is sufficient to determine $\hat{\boldsymbol{\beta}}$ if the dimension of $\mathbf{x}_k$ equals the dimension of $\mathbf{z}_k$ (Kott 2006). On the other hand, when the dimension of $\mathbf{x}_k$ exceeds the dimension of $\mathbf{z}_k$, the calibration equation can be modified into a nonlinear regression-type equation $\mathbf{t}_x = \sum_{k\in r} \frac{d_k}{p(\mathbf{z}_k^{\mathrm{T}}\boldsymbol{\beta})}\mathbf{x}_k + \boldsymbol{\varepsilon}$, where $\mathbf{z}_k$ and $\mathbf{x}_k$ denote the vectors for response model covariates and benchmark variables respectively, $\mathbf{t}_x$ is the vector of calibration target values comprising the known population totals, and $\boldsymbol{\varepsilon}$ is the error term between the calibrated estimates and the population controls of the auxiliary variable (Chang and Kott 2008).

One main potential advantage of the Kott & Chang method is that it permits the use of variables that are observed only on the respondents, and thus may prove useful in the context of nonignorable nonresponse (Kott and Chang, 2010).

1.4.2 Properties of Calibration Estimators in the Presence of Nonresponse

Although several different calibration estimators are widely used in survey practice (e.g., poststratification and raking), there is not much literature about the properties of these

calibration estimators in the presence of nonresponse. It is unclear whether and how the different estimators may perform differently when the nonresponse mechanism is not MCAR.

In terms of bias, Särndal and Lundström (2005) claim that the single-step weighting through calibration approach meets the double objective of reducing sampling error and nonresponse error in the presence of powerful auxiliary information, but give little guidance about how to choose significant auxiliary variables. Although nonlinear adjustment function forms can be considered such as $F(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r) = \exp(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r)$, Särndal and Lundström (2005) suggest that the linear form will suffice due to its considerable computational advantage and the fact that it fits the routine production environment. However, little theoretical or empirical justification is provided to support this statement. During the discussion of confidence interval estimates, Särndal and Lundström (2005) point out that to trust the confidence interval, one must be reasonably assured that the bias of the point estimator is nearly zero; otherwise the confidence interval tends to be off-center and this will cause damage to the coverage rate. Kott and Liao (2012) claim that calibration weighting can provide "double protection" against the selection bias resulting from unit nonresponse. A statistician needs to assume an outcome model (which they refer to as "prediction model") and a response model (which they call "selection model") during calibration weighting. According to Kott and Liao (2012), if *either* an assumed linear prediction model *or* an implied unit selection model holds, the calibration estimator can be asymptotically unbiased "in some sense". It is unclear what Kott and Liao (2012) mean by "in some sense", so their conclusion is vague and may

require some unverifiable model assumptions. At the same time, the Kott & Chang method is computationally intensive, and thus may be difficult to implement in practice.

Lesage, Haziza, and D'Hautlfoeuille (2016) refer to the Kott & Chang method as instrument vector calibration. They lay out the conditions required for establishing the consistency of an instrumental calibration estimator. Let R_k denote the response indicator for unit k such that $R_k = 1$ if unit k is a respondent and $R_k = 0$ otherwise. Let $\left\{\left(\mathbf{x}_k^{\mathrm{T}},\, y_k,\, \mathbf{z}_k^{\mathrm{T}},\, r_k\right), k \in U\right\}$ be realizations of independent and identically distributed random vectors $\left\{\left(\mathbf{X}_k^{\mathrm{T}},\, Y_k,\, \mathbf{Z}_k^{\mathrm{T}},\, R_k\right), k \in U\right\}$. Assume that the response mechanism is described as

$$\mathrm{E}(R_k \mid \mathbf{Z}_k) = p(\mathbf{Z}_k^{\mathrm{T}}\boldsymbol{\beta}) \tag{1.14}$$

Lesage, Haziza, and D'Hautlfoeuille (2016) show that the instrumental vector calibration leads to negligible bias provided that the calibration function $F(\cdot)$ is correctly specified and the following two conditions (referred to as exclusion restriction conditions) are satisfied

$$R_k \perp \mathbf{X}_k \mid \mathbf{Z}_k \tag{1.15}$$

and

$$R_k \perp Y_k \mid \mathbf{Z}_k \tag{1.16}$$

That is, the response propensity is related to some instrument variables via (1.14) but, given the values of the instruments, $\mathbf{Z}_k$, response is unrelated to either the covariates, $\mathbf{X}_k$, or the analysis variables, $\mathbf{Y}_k$.

Lesage, Haziza, and D'Hautlfoeuille (2016) point out that although instrument vector calibration may be successful in reducing nonresponse bias, the estimator may be highly biased and/or unstable when the exclusion restriction conditions are not satisfied. For example, a violation of (1.15) may occur when there exists an unobserved variable U, independent of $\mathbf{Z}$ and Y, which is related to both R and $\mathbf{X}$. In practice, it is not possible to validate the choice of $F(\cdot)$ because the instrument variables are only available for the respondents. Also, it is not possible to check whether or not (1.15) and (1.16) hold. Ideally, the calibration variables should be those exhibiting a strong relationship with the instruments. Alternatively, one may use the one-step calibration procedure solely based on calibration variables for which the population total is known. Although one may not be successful in reducing the bias to the same extent as with instrument vector calibration in some situations, there is no risk of bias and variance amplification as the calibration variables coincide with the instruments, which in turn offer some protection against an unduly large bias and/or variance.

Regarding variance estimation, Särndal and Lundström (2005) show that the variance of a single-step calibration estimator is estimated as the sum of two components. That is, $\hat{V}(\hat{t}_{yw}) = \hat{V}_{SAM} + \hat{V}_{NR}$. The first component is the estimated sampling variance and the second component is the estimated nonresponse variance. Both components involve

estimating residuals — the differences between the observed values and the estimated values for the outcome variable. Using the notations defined for the Särndal & Lundström method in Section 1.4, the estimated sampling variance component is

$$\hat{V}_{SAM} = \sum\sum_r (d_k d_l - d_{kl})(v_k \hat{e}_k^*)(v_l \hat{e}_l^*) - \sum_r d_k (d_k - 1) v_k (v_k - 1)(\hat{e}_k^*)^2 \quad (1.17)$$

and the estimated nonresponse variance component is

$$\hat{V}_{NR} = \sum_r v_k (v_k - 1)(d_k \hat{e}_k)^2 \quad (1.18)$$

with

$$\hat{e}_k^* = y_k - (\mathbf{x}_k^*)^{\mathrm{T}} \mathbf{B}_{r;dv}^* \quad (1.19)$$

and

$$\hat{e}_k = y_k - \mathbf{x}_k^{\mathrm{T}} \mathbf{B}_{r;dv} = y_k - (\mathbf{x}_k^*)^{\mathrm{T}} \mathbf{B}_{r;dv}^* - (\mathbf{x}_k^o)^{\mathrm{T}} \mathbf{B}_{r;dv}^o \quad (1.20)$$

in which

$$\mathbf{B}_{r;dv} = \begin{pmatrix} \mathbf{B}_{r;dv}^* \\ \mathbf{B}_{r;dv}^o \end{pmatrix} = (\sum_r d_k v_k \mathbf{z}_k \mathbf{x}_k^{\mathrm{T}})^{-1} (d_k v_k \mathbf{z}_k y_k) \quad (1.21)$$

where $d_k = 1/\pi_k$ and $d_{kl} = 1/\pi_{kl}$ are the inverse of the first order inclusion probability and the inverse of the second order inclusion probability, respectively.

This variance estimator in Särndal and Lundström (2005) is based on a linear adjustment function $F(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r) = 1 + \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r$ so that the residual terms can be estimated using regression models. It is unclear how the variance should be estimated if a nonlinear function term is used for calibration (e.g., raking ratio adjustment).

D'Arrigo and Skinner (2010) evaluate the properties of the GREG estimator, raking ratio estimator, and maximum likelihood raking estimator as well as the performance of several linearization variance estimators in the presence of nonresponse. They define alternative forms of linearization variance estimators via the choices of (1) the weights applied to the residuals from the regression model; and (2) the weights used in the regression model to estimate regression coefficients and residuals. Their study displays few differences among the properties of the three calibration estimators for a given sampling scheme and nonresponse model. Among the linearization variance estimators, the approach that weights residuals by the design weight can be severely biased in the presence of nonresponse. The approach that weights residuals by the calibrated weight tends to display much less bias. Varying the choice of weights used to construct the regression coefficients has little impact. In the D'Arrigo and Skinner (2010) framework, the simulation is based on several variables from the British Labor Force Survey and German Survey of Income and Expenditure. Although the response model is discussed, there is no explicit information about the outcome variable model. It is unclear whether their conclusion will hold under different outcome variable models. More details about the forms of the linearization variance estimators in D'Arrigo and Skinner (2010) are included in Chapter 5.

1.5 Choosing Auxiliary Variables to Reduce Nonresponse Bias

In the single-step weighting approach, calibration is applied to the basic sampling weights directly without a separate nonresponse adjustment step, so Little and Vartivarian

(2005) offer a useful framework for thinking about how to choose auxiliary variables and/or calibration estimator. The situation in Little and Vartivarian (2005) is a very simple one – simple random sampling (SRS) with a negligible sampling fraction and an outcome variable (Y) with two values. This creates a 2×2 situation: response or nonresponse and the 2-value outcome variable. Two distributions are considered: the response distribution and the Y distribution. The properties of a nonresponse-adjusted mean estimator are evaluated across both the response and the superpopulation Y distributions. As shown in Table 1.1, four scenarios are assessed in Little and Vartivarian (2005) based on the association of the auxiliary variables with response and outcome. The following conclusions (quoting the original text from Little and Vartivarian (2005)) are reached:

L&V (i): "Substantial bias reduction requires adjustment cell variables that are related both to nonresponse and to the outcome of interest."

L&V (ii): "If the adjustment cell variables are unrelated to nonresponse, then weighting tends to have no impact on bias (an unweighted mean would also be unbiased), but reduces variance to the extent that the adjustment cell variables are good predictors of the outcome."

L&V (iii): "If adjustment cell variables are good predictors of nonresponse but unrelated to the outcome variable, then weighting increases variance without any reduction in bias."

L&V (iv): "If the adjustment cell variables are related to neither outcome nor nonresponse, then weighting affects neither bias nor variance."

Table 1.1 Summary of Little and Vartivarian (2005) Conclusions

Scenario	Association with Outcome	Association with Response	Bias	Variance
L&V (i)	High	High	↓	↓
L&V (ii)	High	Low	--	↓
L&V (iii)	Low	High	--	↑
L&V (iv)	Low	Low	--	--

Source: Little and Vartivarian (2005), Table 1.

However, the messages in Little and Vartivarian (2005) are not quite clear to the readers sometimes. For example, on the one hand, they assert that "[a] covariate for a weighting adjustment must have two characteristics to reduce nonresponse bias – it needs to be related to the probability of response, and it needs to be related to the survey outcome." On the other hand, they state that "the most important feature of variables for inclusion in weighting adjustment is that they are predictive of survey outcome; prediction of propensity to respond is a secondary, though useful, goal." The former statement seems to suggest that the outcome variable model and response model should play equally important roles in determining the appropriate covariates for nonresponse adjustment, while the latter seems to indicate that the outcome variable model should be the dominant factor. We suspect that this is because the variables that are predictive of response only have the potential to reduce nonresponse bias, but the variables that are predictive of

outcomes have the potential to reduce both nonresponse bias and sampling variance. So, if a variable is predictive of outcome it will reduce mean squared error (MSE) even if it is not predictive of response. But a variable that is only predictive of response will actually increase MSE.

Moreover, the descriptions in the main text and in Table 1 of Little and Vartivarian (2005) are not quite consistent. The text seems to address extreme conditions where the variables are either "related" or "not related" to the outcome and/or response, while Table 1 shows "high" and "low" correlations, which are the middle-ground conditions that we are more likely to see in reality.

Finally, Little and Vartivarian (2005) address only main effects and do not provide any explicit guidance about how to handle the interaction effects. Since the interaction terms of the main effect variables are not completely new variables, the conclusions in Little and Vartivarian (2005) do not shed light on the differences between the GREG estimator with only main effect terms, the poststratification estimator, and the raking estimator.

1.6 Gaps in the Literature and Research Aims

In the context of using calibration as a single-step weighting approach to reduce potential nonresponse bias, little evaluation has been conducted on the asymptotic properties of different calibration. For example, both the raking ratio estimator and the poststratification estimator are widely used in practice, the former as an example of the

general calibration estimators and the latter as a special case of the GREG estimator. Based on Deville and Särndal (1992), these two estimators are asymptotically equivalent in the absence of non-sampling errors. However, non-sampling errors such as nonresponse error almost always exist in surveys. It is important to re-examine conclusions in Deville and Särndal (1992) in the context of using calibration for nonresponse adjustment.

If the conclusions in Deville and Särndal (1992) do not hold when calibration is used for nonresponse adjustment, then the existing literature provides neither a good framework for comparing the performances of different calibration estimators, nor practical guidance for choosing the appropriate auxiliary vectors and/or function forms for calibration weighting. Although D'Arrigo and Skinner (2010) compare three calibration estimators in the presence of nonresponse, their conclusions are based on a limited number of outcome variables from two surveys, and thus may not hold up in terms of external validity. To understand how a calibration estimator may perform in the presence of nonresponse, we need to go beyond the purely design-based approach used in Deville and Särndal (1992) and examine the underlying models for population structure (i.e., what variables are correlated with the key outcome variable) and response mechanism (i.e., what variables are corrected with response propensity). Survey practitioners need guidelines for how to select the appropriate calibration estimator(s) for nonresponse adjustment, but there is not much research in this area. The work by Little and Vartivarian (2005) may help us define a framework for answering such questions, yet at discussed in Section 1.5, research is needed to address the issues about interaction terms

and refine the conclusions in Little & Vartivarian (2005) through some sensitivity analyses.

In this dissertation, our first research question is whether and how alternative calibration estimators may perform differently in the presence of nonresponse. More specifically, we want to evaluate the properties of three widely used calibration estimators over repeated sampling. Two chapters are dedicated to answering this question. The first chapter focuses on some design-based theoretical development. The second chapter contains a simulation study that compares the performance of three widely used calibration estimators – poststratification, raking, and GREG with only the main effect covariates.

The second research question is how the performance of a calibration estimator may vary by sample configuration. In the real-world survey practice, only one sample can be fielded and all the estimates are based on that particular sample, so it is important to study the properties of the calibration estimators conditioning on sample configuration. We propose a distance measure that can be calculated for a particular sample and may be related to the potential bias of a calibration estimator, which can be used as a diagnostic tool by survey practitioners.

The final chapter of this dissertation examines several alternative variance estimators for raking in the presence of nonresponse, including both the linearization method and replication method. We specify the outcome variable models and response models

explicitly so that the impact of these models on the performance of the variance estimators can be detectable.

Chapter 2. Analytical Work for Comparing the GREG Estimator and General Calibration Estimators with Nonresponse

Chapter 1 identifies some gaps in the existing literature on using calibration for reducing nonresponse bias. In this chapter, we attempt to fill in one gap by comparing the asymptotic properties of the general calibration estimators and GREG estimator when calibration is used for nonresponse adjustment through a single-step weighting approach. Given the risk of bias and variance amplification associated with the instrument vector calibration weighting (Lesage, Haziza, and D'Hautlfoeuille 2016), we use the Särndal & Lundström Method described in Section 1.4.1 and focus on the situation where the vector $\mathbf{z}_k$ used in the weighting adjustment function $F(\cdot)$ coincides with the calibration variable vector $\mathbf{x}_k$. In the presence of nonresponse, the Horvitz-Thompson estimator of the total for the auxiliary vector using the basic design weights is a function of the respondent set and can therefore be "far" from the benchmark control total. This violates one of the key assumptions in Deville and Särndal (1992), so it is unclear whether their conclusions about the relationship between the GREG estimator and the general calibration estimators still hold.

Section 2.1 below specifies the scope and assumptions underlying the theoretical derivation. Section 2.2 presents the analytical work using design-based approach and indicates that different calibration estimators are not necessarily asymptotically identical when calibration is applied on basic design weights directly to correct nonresponse bias. The setup and analytical work in this chapter largely follow the approach taken by Deville and Särndal (1992), which is purely design-based. The proofs in Deville and

Särndal (1992) are summarized in Appendix A, so we can refer to some of their equations during the presentation of our analytical work. We use the terms “new assumption” and “new result” to differentiate our assumptions and findings from those in Deville and Särndal (1992). Our theoretical results are applicable to a family of general calibration estimators discussed in Deville and Särndal (1992). At the end of the chapter, we point out the limitations of the purely design-based approach and emphasize the importance of examining the underlying models for the outcome variable and response propensity when comparing different calibration estimators.

2.1 Scope and Assumptions

First, we assume the analytic survey (i.e., the survey requiring calibration) and benchmark survey come from the same population U of size N. Although the benchmark control totals are often estimated and subject to sampling and non-sampling errors in practice, we assume that the total for the auxiliary vector $\mathbf{x}$ is accurately known and equal to the true population total.

Second, we assume that the analytic survey has no coverage or measurement error, but may suffer from nonresponse error that can bias the estimated parameters such as population totals. In the presence of nonresponse, the survey has a respondent set r of size n_r. We assume that no separate nonresponse adjustment is conducted prior to calibration, so the pre-calibration population estimates are calculated using only the basic design weights d_k. That is, the Horvitz-Thompson estimators of the population totals of

the auxiliary vector and outcome variable are $\hat{\mathbf{t}}_{r_{x\pi}} = \sum_r d_k \mathbf{x}_k$ and $\hat{t}_{r_{y\pi}} = \sum_r d_k y_k$ respectively. Under probability sampling, the Horvitz-Thompson estimator is unbiased if 100 percent participation rate is achieved.

Finally, although survey nonresponse is generally viewed as being caused by a random mechanism, for the simplicity of theoretical derivations in this chapter, we assume that each population member has fixed response propensity of either 1 or 0. (In later chapters, we do allow the response for each unit to be random so that the response propensities can be values between 0 and 1.) In the presence of nonresponse, the design-based expectation of the Horvitz-Thompson estimator reflects the characteristics of the "responding population" U_r of size N_r. We define $E_\pi(\hat{\mathbf{t}}_{r_{x\pi}}) = \mathbf{t}_{r_x}$ and $E_\pi(\hat{t}_{r_{y\pi}}) = t_{r_y}$, where E_π means design-based expectation, and $\mathbf{t}_{r_x}$ and t_{r_y} are the population totals of the auxiliary variable vector and the outcome variable for the respondent set U_r.

The theoretical derivation in this section requires the following assumptions. We refer to these as "new assumptions" in contrast of those in Deville and Särndal (1992).

New assumption (i): $\lim N_r^{-1}\mathbf{t}_{r_x}$ exist, but in general, $\lim N_r^{-1}\mathbf{t}_{r_x} \neq N^{-1}\mathbf{t}_x$.

New assumption (ii): $N_r^{-1}(\hat{\mathbf{t}}_{r_{x\pi}} - \mathbf{t}_{r_x}) \to \mathbf{0}$ in design probability. $N_r^{-1}\left(\hat{\mathbf{t}}_{r_{x\pi}} - \mathbf{t}_{r_x}\right) = O_p(n_r^{-1/2})$.

New assumption (iii). $n_r^{1/2} N_r^{-1} (\hat{\mathbf{t}}_{r_{x\pi}} - \mathbf{t}_{r_x})$ converges in distribution to the multinormal $N(\mathbf{0}, \mathbf{A})$, where $\mathbf{A}$ can be viewed as a matrix that describes an asymptotic effect of the sampling design used for the analytic survey.

Recall that one of the key assumptions in Deville and Särndal (1992) is that in the purely sampling context, the Horvitz-Thompson estimators of the population totals of the auxiliary vector approach the true values of the population as the sample size increases. That is, $N^{-1}\left(\hat{\mathbf{t}}_{x\pi} - \mathbf{t}_x\right) = O_p(n^{-1/2})$. Based on our new assumption (ii), the Horvitz-Thompson estimators from the respondent set approach only $\mathbf{t}_{r_x} = E_\pi(\hat{\mathbf{t}}_{r_{x\pi}})$. We know that $\mathbf{t}_{r_x} \neq \mathbf{t}_x$ in the presence of nonresponse. This has important implications in the theoretical derivation in Section 2.2.

2.2 Analytical Results Using Design-based Approach

In this section we re-examine the results in Deville and Särndal (1992) in the context of using calibration for nonresponse adjustment through single-step weighting. The input weights for the calibration equation are the basic design weights d_k. In this setup, the Horvitz-Thompson estimator $\hat{t}_{r_{y\pi}}$ using the basic sampling weights d_k is biased due to nonresponse, so calibration is used to reduce such bias to the extent possible. Conceptually, it is more appropriate to understand calibration from the perspective of the function form method than that of the distance function method. This is because our goal is not to obtain calibration weights that are as "close" to the basic design weights as

possible in order to maintain design unbiasedness, as required in Deville and Särndal (1992). We suspect that whether the calibration equation has a solution may depend on the how the response propensities differ by the benchmark control variables used in the calibration. We show that the vector for Lagrange multiplier determined from the calibration equation, $\boldsymbol{\lambda}_r$, consists of a term that is driven by the difference between the Horvitz-Thompson estimator of the auxiliary vector (using the basic design weights) for the respondent population total (denoted by $\hat{\mathbf{t}}_{r_{x\pi}}$) and the benchmark control total (denoted by $\mathbf{t}_x$). Unless nonresponse is negligible, this term does not decrease as the survey sample size increases, so $\boldsymbol{\lambda}_r$ may tend to a non-zero constant vector in design probability. Our analytical work results in the formulae for: (1) the difference between a general calibration estimator and Horvitz-Thompson estimator in the presence of nonresponse; and (2) the difference between a general calibration estimator and the GREG estimator in the presence of nonresponse. We prove that when nonresponse exists and calibration is used to reduce nonresponse bias through single-step weighting, the general calibration estimators and the GREG estimator are not asymptotically equivalent in general situations.

In the presence of nonresponse, the calibration equation is $\sum_r w_k \mathbf{x}_k = \mathbf{t}_x$ and the calibration estimator is $\hat{t}_{yw} = \sum_r w_k y_k$. Equations (1.5) and (1.6) in Chapter 1 should be modified into

$$\Phi_r\left(\boldsymbol{\lambda}_r\right) = \sum_r d_k \left\{F_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right) - 1\right\}\mathbf{x}_k \tag{2.1}$$

and

$$\Phi_r(\boldsymbol{\lambda}_r) = \mathbf{t}_x - \hat{\mathbf{t}}_{r_{x\pi}} = (\mathbf{t}_x - \mathbf{t}_{r_x}) + (\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}) \tag{2.2}$$

We know that $\mathbf{t}_{r_x} \neq \mathbf{t}_x$ in the presence of nonresponse, so the right-hand side of (2.2) contains a non-zero term that does not exist in equation (1.6) of Chapter 1. This non-zero term plays an important role in the discussions below. We have five new results in parallel to the ones in Deville and Särndal (1992).

New Result 1. As $n_r \to \infty$, whether equation (2.2) has a solution may depend on the difference between $\mathbf{t}_{r_x}$ and $\mathbf{t}_x$ as well as the function form $F_k(\cdot)$ used in the calibration.

For this result, we give intuitive explanations instead of strict proof. In the presence of nonresponse, equating (2.1) and (2.2) gives

$$N_r^{-1}\Phi_r(\boldsymbol{\lambda}_r) = N_r^{-1}\sum_r d_k F_k(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r)\mathbf{x}_k - N_r^{-1}\sum_r d_k \mathbf{x}_k = N_r^{-1}(\mathbf{t}_x - \mathbf{t}_{r_x}) + N_r^{-1}(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}) \tag{2.3}$$

The second term on the right-hand side of (2.3) is similar to that in Deville and Särndal (1992). $N_r^{-1}(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}) = O_p(n_r^{-1/2})$ and is asymptotically $\mathbf{0}$. However, when nonresponse exists, $\mathbf{t}_x \neq \mathbf{t}_{r_x}$ and the first term is $N_r^{-1}(\mathbf{t}_x - \mathbf{t}_{r_x}) = O(1)$. Due to this additional term, the right-hand side of (2.3) does not tend to $\mathbf{0}$, but becomes a non-zero constant vector as n_r increases.

A more intuitive way to understand this result is that in Deville and Särndal (1992), only "small" adjustments need to be made to the basic design weights to obtain the calibration weights, and that is essentially why the calibration equation almost always has a solution for large samples. When nonresponse exists, the Horvitz-Thompson estimator $\sum_r d_k \mathbf{x}_k$ may be "far" from the benchmark controls $\mathbf{t}_x$ and therefore "large" adjustments on the basic design weights may be required to satisfy the calibration constraints. In this situation, whether the calibration equation has a solution may depend on the difference between $\mathbf{t}_{r_x}$ and $\mathbf{t}_x$ as well as the function form $F_k(\cdot)$ used in the calibration. An empirical example is that for the same calibration constraints and respondent set $k \in r$, poststratification always has a solution but raking does not always converge.

New Result 2. Let $\boldsymbol{\lambda}_r$ be the solution to equation (2.3) if one exists. If $\mathbf{t}_x - \mathbf{t}_{r_x} \neq \mathbf{0}$, then $\boldsymbol{\lambda}_r = O_p(1)$ in general situations. This means that $\boldsymbol{\lambda}_r$ tends to a non-zero vector in design probability.

Proof: Define $\mathbf{z}_1 = N_r^{-1}(\mathbf{t}_x - \mathbf{t}_{r_x})$ and $\mathbf{z}_2 = N_r^{-1}(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{xr}})$, so $\boldsymbol{\lambda}_r = (N^{-1}\Phi_r)^{-1}(\mathbf{z}_1 + \mathbf{z}_2)$ if a solution to (2.3) exists. Since $N^{-1}\Phi_r(0) = 0$, we have $\boldsymbol{\lambda}_r - 0 = \left(N^{-1}\Phi_r\right)^{-1}(\mathbf{z}_1 + \mathbf{z}_2) - \left(N^{-1}\Phi_r\right)^{-1}(0)$. Following the notations in Deville and Särndal (1992), the inequality (A.3) in their Result 2 (refer to Appendix A) becomes

$$\|\boldsymbol{\lambda}_r\| \leq \|\mathbf{z}_1 + \mathbf{z}_2\| K(1-\beta)^{-1} \leq \|\mathbf{z}_1\| K(1-\beta)^{-1} + \|\mathbf{z}_2\| K(1-\beta)^{-1} \qquad (2.4)$$

where K is defined in (A.1) and $0 < \beta < \frac{1}{2}$.

Since $\mathbf{z}_1 = O(1)$ and $\mathbf{z}_2 = O_p(n_r^{-1/2})$, inequality (2.4) implies that $\boldsymbol{\lambda}_r = O(1) + O_p(n_r^{-1/2})$. The second term tends to $\mathbf{0}$ as n_r increases. However, the first term is a non-zero constant vector in general situations, and does not decrease as n_r increases.

New Result 3. In general situations, $\boldsymbol{\lambda}_r = \mathbf{T}_r^{-1}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right) + O_p(1)$, where $\mathbf{T}_r = \sum_r d_k q_k \mathbf{x}_k \mathbf{x}_k^{\mathrm{T}}$.

Proof: We use $F_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right)$ to denote the adjustment function for a general calibration estimator. For the GREG estimator, the adjustment function takes the form $1 + q_k \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r$. The difference between the two adjustment functions is expressed as

$$\theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right) = F_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right) - (1 + q_k \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r) \tag{2.5}$$

From (2.1), (2.2), and (2.5), we obtain

$$(\mathbf{t}_x - \mathbf{t}_{r_x}) + (\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}) = \sum_r d_k \mathbf{x}_k \left\{ q_k \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r + \theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right)\right\} \tag{2.6}$$

Multiplying both sides of (2.6) by $\mathbf{T}_r^{-1}$ and rearranging the terms, we obtain

$$\boldsymbol{\lambda}_r - \mathbf{T}_r^{-1}(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}) = \mathbf{T}_r^{-1}(\mathbf{t}_x - \hat{\mathbf{t}}_{r_x}) - \mathbf{T}_r^{-1}\sum_r d_k \mathbf{x}_k \theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right) \tag{2.7}$$

An important assumption in Deville and Särndal (1992) is that $F_k''(0)$ is uniformly bounded, which is equivalent to $\theta\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right) = \max\ \theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right) = O\left(\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right)^2\right)$. Note that this assumption requires the condition that $\boldsymbol{\lambda}_r = O_p(n_r^{-1/2})$, which does not necessarily hold when $\mathbf{t}_x \neq \mathbf{t}_{r_x}$. However, given that $\max|\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r| < \infty$, when nonresponse in the analytic survey is not extremely severe, we can still assume that for any $\varepsilon > 0$, there exists K'' such that, for all *k*, $|\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r| < \varepsilon$ will imply that $\theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right) \le K''\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right)^2$.

Using (2.7) and the bound above on $\theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r\right)$, we have

$$\left\|\boldsymbol{\lambda}_r - \mathbf{T}_r^{-1}(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}})\right\| \le \left\|(N_r^{-1}\mathbf{T}_r)^{-1}\right\| K'' \left\{N_r^{-1}\sum_r d_k \left\|\mathbf{x}_k\right\|^3\right\}\left\|\boldsymbol{\lambda}_r\right\|^2 + \mathbf{T}_r^{-1}(\mathbf{t}_x - \mathbf{t}_{r_x}) \qquad (2.8)$$

We know that $\left\|(N_r^{-1}\mathbf{T}_r)^{-1}\right\| = O_p(1)$ and $N_r^{-1}\sum_r d_k \left\|\mathbf{x}_k\right\|^3 = O_p(1)$. Based on the New Result 2, $\left\|\boldsymbol{\lambda}_r\right\|^2 = O_p(1)$, so the first term of the right-hand side of (2.8) is $O_p(1)$. The second term of the right-hand side of (2.8) is also $O_p(1)$. Therefore we have $\boldsymbol{\lambda}_r = \mathbf{T}_r^{-1}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right) + O_p(1)$. Although $\mathbf{T}_r^{-1}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right)$ tends to $\mathbf{0}$ as $n_r \to \infty$, the magnitude of $\boldsymbol{\lambda}_r$ is $O_p(1)$ in general situations. Unless $\mathbf{t}_x = \mathbf{t}_{r_x}$, $\boldsymbol{\lambda}_r$ does not tend to $\mathbf{0}$ as $n_r \to \infty$.

New Result 4. The difference between the general calibration estimator and the Horvitz-Thompson estimator can be expressed in two ways.

In terms of totals:

$$\hat{t}_{r_{yw}} - \hat{t}_{r_{y\pi}} = \hat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_x - \mathbf{t}_{r_x}\right) + \hat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right) + (\mathbf{Y}_r - \hat{\mathbf{Y}}_r)^{\mathrm{T}}\mathbf{D}_r\boldsymbol{\theta}_r \tag{2.9}$$

where

$$\underset{p\times 1}{\hat{\mathbf{B}}_r} = \mathbf{T}_r^{-1}\mathbf{X}_r^{\mathrm{T}}\mathbf{D}_r\mathbf{Q}_r\mathbf{Y}_r = \left(\sum_r d_k q_k \mathbf{x}_k \mathbf{x}_k^{\mathrm{T}}\right)^{-1}\sum_r d_k q_k \mathbf{x}_k y_k$$

$$\underset{p\times p}{\mathbf{T}_r} = \mathbf{X}_r^{\mathrm{T}}\mathbf{D}_r\mathbf{Q}_r\mathbf{X}_r = \sum_r d_k q_k \mathbf{x}_k \mathbf{x}_k^{\mathrm{T}}$$

$$\underset{n_r\times p}{\mathbf{X}_r} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n_r 1} & \cdots & x_{n_r p} \end{pmatrix} = (\mathbf{x}_1^{\mathrm{T}}, \ldots \mathbf{x}_k^{\mathrm{T}}, \ldots, \mathbf{x}_{n_r}^{\mathrm{T}})^{\mathrm{T}}$$

$$\underset{n_r\times n_r}{\mathbf{Q}_r} = \begin{pmatrix} q_1 & & 0 \\ & \ddots & \\ 0 & & q_{n_r} \end{pmatrix}$$

$$\underset{n_r\times n_r}{\mathbf{D}_r} = \begin{pmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_{n_r} \end{pmatrix}$$

$$\mathbf{Y}_r = (y_1, \ldots y_k, \ldots, y_{n_r})^{\mathrm{T}}$$

$$\hat{\mathbf{Y}}_r = \mathbf{X}_r\hat{\mathbf{B}}_r$$

$$\boldsymbol{\theta}_r = (\theta_1, \ldots \theta_k, \ldots, \theta_{n_r})^{\mathrm{T}}$$

In terms of means:

$$\begin{aligned} &N^{-1}\hat{t}_{r_{yw}} - N_r^{-1}\hat{t}_{r_{y\pi}} \\ &= \hat{\mathbf{B}}_r^{\mathrm{T}}\boldsymbol{\mu}_x\left(1 - (\boldsymbol{\mu}_{r_x}/\boldsymbol{\mu}_x)\,p\right) - \hat{\boldsymbol{\mu}}_{r_{y\pi}}(1-p) + N^{-1}\hat{\mathbf{B}}_r^{\mathrm{T}}\left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}}\right) + N^{-1}(\mathbf{Y}_r - \hat{\mathbf{Y}}_r)^{\mathrm{T}}\mathbf{D}_r\boldsymbol{\theta}_r \end{aligned} \tag{2.10}$$

where $\boldsymbol{\mu}_{r_x}$ is the mean of the auxiliary vector for the respondent population, $\boldsymbol{\mu}_x$ is the true population mean, $\hat{\boldsymbol{\mu}}_{r_{y\pi}}$ is the Horvitz-Thompson estimator of the mean for the outcome variable estimated from the respondent set, and p is the response rate of the analytic survey.

Proof: If the calibration equation has a solution $\boldsymbol{\lambda}_r$, then from (2.5) the difference between the general calibration estimator and the Horvitz-Thompson estimator can be written as

$$\hat{t}_{r_{yw}} - \hat{t}_{r_{y\pi}} = \sum_r d_k y_k \{q_k \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r + \theta_k (\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r)\} \tag{2.11}$$

From (2.7),

$$\boldsymbol{\lambda}_r = \mathbf{T}_r^{-1} \left(\mathbf{t}_x - \mathbf{t}_{r_x} \right) + \mathbf{T}_r^{-1} \left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}} \right) - \mathbf{T}_r^{-1} \sum_r d_k \mathbf{x}_k \theta_k \left(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r \right) \tag{2.12}$$

Replacing the first occurrence of $\boldsymbol{\lambda}_r$ in (2.11) by the right-hand side of (2.12), we obtain

$$
\begin{aligned}
&\hat{t}_{r_{yw}} - \hat{t}_{r_{y\pi}} \\
&= \sum_r d_k y_k q_k \mathbf{x}_k^{\mathrm{T}} \left\{ \mathbf{T}_r^{-1} \left(\mathbf{t}_x - \mathbf{t}_{r_x} \right) + \mathbf{T}_r^{-1} \left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}} \right) - \mathbf{T}_r^{-1} \sum_r d_k \mathbf{x}_k \theta_k \left(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r \right) \right\} \\
&+ \sum_r d_k y_k \theta_k (\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r) \\
&= \sum_r d_k y_k q_k \mathbf{x}_k^{\mathrm{T}} \mathbf{T}_r^{-1} \left(\mathbf{t}_x - \mathbf{t}_{r_x} \right) + \sum_r d_k y_k q_k \mathbf{x}_k^{\mathrm{T}} \mathbf{T}_r^{-1} \left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}} \right) \\
&- \sum_r d_k y_k q_k \mathbf{x}_k^{\mathrm{T}} \mathbf{T}_r^{-1} \sum_r d_k \theta_k \left(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r \right) \mathbf{x}_k + \sum_r d_k y_k \theta_k (\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r) \\
&= \left(\mathbf{X}_r^{\mathrm{T}} \mathbf{D}_r \mathbf{Q}_r \mathbf{Y}_r \right)^{\mathrm{T}} \mathbf{T}_r^{-1} \left(\mathbf{t}_x - \mathbf{t}_{r_x} \right) + \left(\mathbf{X}_r^{\mathrm{T}} \mathbf{D}_r \mathbf{Q}_r \mathbf{Y}_r \right)^{\mathrm{T}} \mathbf{T}_r^{-1} \left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}} \right) \\
&- \left(\mathbf{X}_r^{\mathrm{T}} \mathbf{D}_r \mathbf{Q}_r \mathbf{Y}_r \right)^{\mathrm{T}} \mathbf{T}_r^{-1} \mathbf{X}_r^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r + \mathbf{Y}_r^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r \\
&= \left(\mathbf{X}_r^{\mathrm{T}} \mathbf{D}_r \mathbf{Q}_r \mathbf{Y}_r \right)^{\mathrm{T}} \mathbf{T}_r^{-1} \left(\mathbf{t}_x - \mathbf{t}_{r_x} \right) + \left(\mathbf{X}_r^{\mathrm{T}} \mathbf{D}_r \mathbf{Q}_r \mathbf{Y}_r \right)^{\mathrm{T}} \mathbf{T}_r^{-1} \left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}} \right) \\
&- \hat{\mathbf{B}}_r^{\mathrm{T}} \mathbf{X}_r^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r + \mathbf{Y}_r^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r \\
&= \hat{\mathbf{B}}_r^{\mathrm{T}} \left(\mathbf{t}_x - \mathbf{t}_{r_x} \right) + \hat{\mathbf{B}}_r^{\mathrm{T}} \left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}} \right) - \hat{\mathbf{Y}}_r^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r + \mathbf{Y}_r^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r \\
&= \hat{\mathbf{B}}_r^{\mathrm{T}} \left(\mathbf{t}_x - \mathbf{t}_{r_x} \right) + \hat{\mathbf{B}}_r^{\mathrm{T}} \left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}} \right) + (\mathbf{Y}_r - \hat{\mathbf{Y}}_r)^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r
\end{aligned}
\tag{2.13}
$$

Then the difference between two means is

$$
\begin{aligned}
&N^{-1} \hat{t}_{r_{yw}} - N_r^{-1} \hat{t}_{r_{y\pi}} \\
&= N^{-1} \sum_r d_k y_k \{ q_k \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r + \theta_k (\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r) \} + (N^{-1} - N_r^{-1}) \hat{t}_{r_{y\pi}} \\
&= N^{-1} \hat{\mathbf{B}}_r^{\mathrm{T}} \left(\mathbf{t}_x - \mathbf{t}_{r_x} \right) + N^{-1} \hat{\mathbf{B}}_r^{\mathrm{T}} \left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}} \right) \\
&+ N^{-1} (\mathbf{Y}_r - \hat{\mathbf{Y}}_r)^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r + (\hat{t}_{r_{y\pi}} / N_r)(N_r / N - 1) \\
&= \hat{\mathbf{B}}_r^{\mathrm{T}} \left(\mathbf{t}_x / N - (\mathbf{t}_{r_x} / N_r)(N_r / N) \right) + (\hat{t}_{r_{y\pi}} / N_r)(N_r / N - 1) \\
&+ N^{-1} \hat{\mathbf{B}}_r^{\mathrm{T}} \left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}} \right) + N^{-1} (\mathbf{Y}_r - \hat{\mathbf{Y}}_r)^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r \\
&= \hat{\mathbf{B}}_r^{\mathrm{T}} \left(\boldsymbol{\mu}_x - \boldsymbol{\mu}_{r_x} (N_r / N) \right) + \hat{\boldsymbol{\mu}}_{r_{y\pi}} (N_r / N - 1) + N^{-1} \hat{\mathbf{B}}_r^{\mathrm{T}} \left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}} \right) + N^{-1} (\mathbf{Y}_r - \hat{\mathbf{Y}}_r)^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r \\
&= \hat{\mathbf{B}}_r^{\mathrm{T}} \left(\boldsymbol{\mu}_x - \boldsymbol{\mu}_x (\boldsymbol{\mu}_{r_x} / \boldsymbol{\mu}_x) p \right) - \hat{\boldsymbol{\mu}}_{r_{y\pi}} (1 - p) + N^{-1} \hat{\mathbf{B}}_r^{\mathrm{T}} \left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}} \right) + N^{-1} (\mathbf{Y}_r - \hat{\mathbf{Y}}_r)^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r \\
&= \hat{\mathbf{B}}_r^{\mathrm{T}} \boldsymbol{\mu}_x \left(1 - (\boldsymbol{\mu}_{r_x} / \boldsymbol{\mu}_x) p \right) - \hat{\boldsymbol{\mu}}_{r_{y\pi}} (1 - p) + N^{-1} \hat{\mathbf{B}}_r^{\mathrm{T}} \left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_{x\pi}} \right) + N^{-1} (\mathbf{Y}_r - \hat{\mathbf{Y}}_r)^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r
\end{aligned}
\tag{2.14}
$$

where $p = N_r / N$ is the proportion of respondents in the population.

For the right-hand side of (2.14), the first two terms do not cancel out except in some special situations such as $\boldsymbol{\mu}_{r_x} = \boldsymbol{\mu}_x$ (indicating ignorable nonresponse) and $\widehat{\mathbf{B}}_r^{\mathrm{T}}\boldsymbol{\mu}_x = \hat{\boldsymbol{\mu}}_{r_{y\pi}}$ (meaning that the assisting linear regression model has perfect predicting power). The third term is $O_p\left(n_r^{-1/2}\right)$. The fourth term is a weighted sum of residuals $(\mathbf{Y}_r - \hat{\mathbf{Y}}_r)$, which has model-expectation 0 if *y* follows a linear model on the *x*'s based on the responding sample but not otherwise. Based on the New Result 3, we know $\boldsymbol{\theta}_r = O_p(1)$, so the fourth term does not necessarily diminish as n_r increases. Instead, its magnitude seems to depend on the variation of the outcome variable, the predicting power of the regression model underlying the GREG estimator, and the form of the weight adjustment function used in calibration. In general, the difference between general calibration estimator and Horvitz-Thompson estimator does not necessarily decrease as n_r increases.

New Result 5. The difference between the general calibration estimator and the GREG estimator is $N^{-1}\left(\hat{t}_{r_{yw}} - \hat{t}_{r_{yreg}}\right) = N^{-1}\mathbf{Y}_r^{\mathrm{T}}\mathbf{D}_r\boldsymbol{\theta}_r = O_p(1)$.

Proof: From (2.5) and (2.14), the general calibration estimator can be expressed as

$$
\begin{aligned}
\hat{t}_{r_{yw}} &= \sum_r d_k y_k + \sum_r d_k q_k \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r y_k + \sum_r d_k \theta_k \left(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r \right) y_k \\
&= \sum_r d_k y_k + \sum_r d_k q_k \mathbf{x}_k^{\mathrm{T}} \left\{ \mathbf{T}_r^{-1} \left(\mathbf{t}_x - \mathbf{t}_{r_x} \right) + \mathbf{T}_r^{-1} \left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_x} \right) - \mathbf{T}_r^{-1} \sum_r d_k \mathbf{x}_k \theta_k \left(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r \right) \right\} y_k \\
&\quad + \sum_r d_k \theta_k \left(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r \right) y_k \\
&= \sum_r d_k y_k + \sum_r d_k q_k \mathbf{x}_k^{\mathrm{T}} \mathbf{T}_r^{-1} \left(\mathbf{t}_x - \mathbf{t}_{r_x} \right) y_k + \sum_r d_k q_k \mathbf{x}_k^{\mathrm{T}} \mathbf{T}_r^{-1} \left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_x} \right) y_k \\
&\quad - \sum_r d_k q_k \mathbf{x}_k^{\mathrm{T}} \mathbf{T}_r^{-1} \sum_r d_k \mathbf{x}_k \theta_k \left(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r \right) y_k + \sum_r d_k \theta_k \left(\mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda}_r \right) y_k \\
&= \hat{t}_{r_{y\pi}} + \widehat{\mathbf{B}}_r^{\mathrm{T}} \left(\mathbf{t}_x - \mathbf{t}_{r_x} \right) + \widehat{\mathbf{B}}_r^{\mathrm{T}} \left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_x} \right) - \hat{\mathbf{Y}}_r^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r + \mathbf{Y}_r^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r
\end{aligned}
\tag{2.15}
$$

But the first four terms of the right-hand side of (2.15) is the GREG estimator

$$
\hat{t}_{r_{yreg}} = \hat{t}_{r_{y\pi}} + \widehat{\mathbf{B}}_r^{\mathrm{T}} \left(\mathbf{t}_x - \mathbf{t}_{r_x} \right) + \widehat{\mathbf{B}}_r^{\mathrm{T}} \left(\mathbf{t}_{r_x} - \hat{\mathbf{t}}_{r_x} \right) - \hat{\mathbf{Y}}_r^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r \tag{2.16}
$$

So $N^{-1} \left(\hat{t}_{r_{yw}} - \hat{t}_{r_{yreg}} \right) = N^{-1} \mathbf{Y}_r^{\mathrm{T}} \mathbf{D}_r \boldsymbol{\theta}_r = O_p(1)$.

The term $\boldsymbol{\theta}_r$ captures the difference between the weight adjustment function for any general calibration estimator and the weight adjustment function for the GREG estimator. When calibration is used for nonresponse adjustment, $\boldsymbol{\theta}_r = O_p(1)$ in general situations and does not tend to zero as the sample size n_r increases. As a result, the GREG estimator and the general calibration estimators are not asymptotically equivalent.

2.3 Summary

The results in this chapter are purely design-based and provide some initial insight on the difference between the general calibration estimators and the GREG estimator when

calibration is applied on the basic design weight directly to correct nonresponse bias. In contrast to the findings of Deville and Särndal (1992) in the absence of nonresponse, our theoretical analysis shows that in the presence of nonresponse, the general calibration estimators and the GREG estimator are not asymptotically equivalent in general situations. At the same time, there are some questions yet to answer. For example, what factors may affect the magnitude of the difference between two calibration estimators? Are there special situations where some forms of calibration estimators may yield asymptotically equivalent results? To further understand what drives the differences between the various calibration estimators, we need to go beyond the design-based approach and examine the underlying models for the outcome variable and the response mechanism. For example, a set of variables may be correlated with the outcome variable of interest. Another set of variables may be correlated with the response propensity. The question is how to incorporate these covariates in the calibration process to reduce potential nonresponse bias without increasing variance significantly. In the next chapter, we examine three widely used calibration estimators, poststratification, raking, and GREG estimator accounting for only the main effects of the auxiliary variables, in greater detail. We aim to provide a framework for evaluating calibration estimators using both design-based and model-based approaches.

Chapter 3. Comparison of Three Widely Used Calibration Estimators for Nonresponse Adjustment over Repeated Sampling

Chapter 2 contains some theoretical results about the difference between the GREG estimator and the general calibration estimators when nonresponse exists and calibration is used for nonresponse adjustment through a single-step weighting approach (i.e., calibration is applied to the basic sampling weights directly without a separate nonresponse adjustment step). The results in Chapter 2 show that the GREG estimator and general calibration estimators are not necessarily asymptotically equivalent when the nonresponse mechanism is not missing completely at random (MCAR). At the same time, to further understand what drives the differences between the various calibration estimators, it is necessary to go beyond the purely design-based approach and examine the underlying models for the outcome variable and the response mechanism.

In this chapter, we focus on three widely used calibration estimators in the situation where the auxiliary information is in the form of counts in a frequency table in two or more dimensions. We examine raking (as an example of the general calibration estimators), poststratification (as a special form of the GREG estimator that accounts for the interaction effects of the auxiliary variables), and the GREG estimator that accounts for only the main effects of the auxiliary variables. In practice, the choice between these estimators is often based on the availability of external data and the counts of respondents in cells formed by variables that may drive response propensities. This chapter uses a systematic approach to evaluate the performance of these three estimators through a simulation study. We compare the empirical biases, empirical variances, and coverage

rates of the 95 percent confidence intervals of these estimators over repeated sampling. We also provide a comprehensive framework to evaluate the impact of sampling, outcome variable structure, and nonresponse mechanism simultaneously. The findings demonstrate the importance of accounting for both the outcome variable model and the response model when choosing the appropriate calibration estimator. The results of this chapter also provide survey practitioners with some guidance for choosing between these widely used calibration estimators.

The content of this chapter is organized as follows: Section 3.1 defines the three calibration estimators in comparison. Sections 3.2 through 3.4 describe the scope, conceptual framework, scenarios, and steps for the simulation study. The evaluation criteria and anticipated results are presented in Sections 3.5 and 3.6. Section 3.7 shows the simulation results over repeated sampling, followed by some sensitivity analysis in Section 3.8. Section 3.9 summarizes the findings and discusses some potential work in the future.

3.1 Poststratification, Raking, and the GREG without Interaction Effects

Survey practitioners often face the issue of variable and function form selection when conducting calibration weighting to reduce nonresponse bias. For example, a set of covariates $\mathbf{X}_1$ may determine the outcome variable of interest while another set of covariates $\mathbf{X}_2$ may drive the response propensity. The relationship between $\mathbf{X}_1$ and $\mathbf{X}_2$ can fall into one of the three situations: 1) $\mathbf{X}_1$ and $\mathbf{X}_2$ are exactly the same; 2) $\mathbf{X}_1$ and

$\mathbf{X}_2$ are different but have overlapping components; or 3) $\mathbf{X}_1$ and $\mathbf{X}_2$ are completely different with no overlapping components. In practice, some the covariates associated with the response propensity are not be correlated with an outcome variable (the second situation above), so the question is what covariates should be included in the calibration process to reduce nonresponse without increasing variance significantly. In the single-step weighting approach, calibration is applied to the basic sampling weights directly without a separate nonresponse adjustment step, so Little and Vartivarian (2005) offer a useful framework for thinking about how to choose auxiliary variables and/or calibration estimator (e.g., poststratification versus raking). However, one of the limitations of Little and Vartivarian (2005) is that they address only main effects and do not provide any explicit guidance about how to handle the interaction effects.

The theoretical results in Chapter 2 indicate that the GREG estimator and general calibration estimators are not necessarily asymptotically equivalent when calibration is used for nonresponse adjustment. To further investigate the factors that may affect the difference between the GREG estimator and a general calibration estimator, we focus on three widely used calibration estimators: (1) poststratification estimator as a special case of the GREG estimator where both main and interaction effects of the categorical auxiliary variables are taken account of; (2) raking ratio estimator as an example of the general calibration estimators; and (3) the GREG estimator when only the main effects of the auxiliary variables are accounted for. For simplicity, we refer to the GREG estimator accounting for only main effects as "GREG_Main".

3.1.1 Poststratification Estimator

The poststratification estimator is model-unbiased under the group-mean assisting model. The auxiliary information consists of known cell counts in a frequency table in any number of dimensions. For simplicity, we consider a two-way table with r rows and c columns, and thus $r \times c$ mutually exclusive cells. The auxiliary vector $\mathbf{x}_k$ is composed of $rc-1$ entries of 0 and a single entry of 1 indicating the cell to which k belongs. The population U_{ij} in cell ij contains N_{ij} elements, i=1, …, r; j=1, …, c. So $N = \sum_{i=1}^{r}\sum_{j=1}^{c} N_{ij}$. Let r_{ij} denote the set of survey respondents in cell $U_{ij} \bigcap r$. For every k in r_{ij}, $F(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r) = N_{ij} \big/ \hat{N}_{rij}$, where $\hat{N}_{rij} = \sum_{k \in r_{ij}} d_k$. The calibration weights for all k in cell ij are calculated as

$$w_k = d_k \left(N_{ij} \big/ \hat{N}_{rij} \right), \qquad k \in r_{ij} \tag{3.1}$$

The poststratification estimator of a population total can be written as

$$\hat{t}_{yps} = \sum_{i=1}^{r}\sum_{j=1}^{c} N_{ij} \frac{\sum_{k \in r_{ij}} d_k y_k}{\sum_{k \in r_{ij}} d_k} \tag{3.2}$$

3.1.2 Raking Estimator

Sometimes survey practitioners do not have all the cell counts N_{ij}, but only marginal counts for the benchmark controls. In other cases, it may not be wise to use the full

cross-classification for adjusting the estimator when the sample sizes for some cells are small. One way to utilize the auxiliary information is to calibrate on known marginal counts, referred to as generalized raking (Deville and Särndal, 1993). For example, in the situation with two auxiliary variables with r and c categories respectively, it is only necessary to know the marginal population counts. The auxiliary vector takes the form of $\mathbf{x}_k = (\delta_{1\cdot k,\ ...,}\ \delta_{r\cdot k,}\ \delta_{\cdot 1k,\ ...,}\ \delta_{\cdot ck})^{\mathrm{T}}$, where $\delta_{i\cdot k} = 1$ if element k is in row i and 0 otherwise, and $\delta_{\cdot jk} = 1$ if k is in column j and 0 otherwise. Consequently, the benchmark control vector is $\sum_{\mathrm{U}} \mathbf{x}_k = (N_{1+},\ ...,\ N_{r+}, N_{+1},...,\ N_{+c})^{\mathrm{T}}$, where $N_{i+} = \sum_{j=1}^{c} N_{ij}$, $N_{+j} = \sum_{i=1}^{r} N_{ij}$. Define the vector of Lagrange multipliers $\boldsymbol{\lambda}_r = (u_{1,\ ...,}\ u_{r,}\ v_{1,\ ...,}\ v_c)^{\mathrm{T}}$, then $\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r = u_i + v_j$ and $F(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_r) = F(u_i + v_j)$ whenever k belongs to cell ij. With $\hat{N}_{rij} = \sum_{k \in r_{ij}} d_k$, the calibration equations are

$$\sum_{j=1}^{c} \hat{N}_{rij} F(u_i + v_j) = N_{i+}, \qquad i = 1,\ ...,\ r \tag{3.3}$$

and

$$\sum_{i=1}^{r} \hat{N}_{rij} F(u_i + v_j) = N_{+j}, \qquad j = 1,\ ...,\ c \tag{3.4}$$

These calibration equations do not have a closed-form solution for $\boldsymbol{\lambda}_r$. Deming and Stephan (1943) suggest an iterative proportional fitting procedure that adjusts one marginal at a time until convergence is achieved. Once $F(u_i + v_j)$ has been determined, the calibrated cell counts can be estimated as

$$\hat{N}_{rij}^{w} = \hat{N}_{rij} F(u_i + v_j), \quad i = 1,\ ...,\ r,\ j = 1,\ ...,\ c \tag{3.5}$$

The calibrated weights for all k in cell ij are

$$w_k = d_k \hat{N}^w_{rij} \big/ \hat{N}_{rij}, \qquad k \in r_{ij} \tag{3.6}$$

The corresponding calibration estimator of a population total is called raking ratio estimator

$$\hat{t}_{yrk} = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k\in r_{ij}} w_k y_k = \sum_{i=1}^{r}\sum_{j=1}^{c} \hat{N}^w_{rij} \frac{\sum_{k\in r_{ij}} d_k y_k}{\sum_{k\in r_{ij}} d_k} \tag{3.7}$$

3.1.3 GREG_Main Estimator

An alternative approach to take advantage of the marginal counts is to use the GREG estimator (Särndal, Swensson, and Wretman, 1992)

$$\hat{t}_{yg} = \hat{t}_{r_{y\pi}} + (\mathbf{t}_x - \hat{t}_{r_{x\pi}})^{\mathrm{T}} \hat{\mathbf{B}}_r \tag{3.8}$$

where $\hat{\mathbf{t}}_{r_{x\pi}}$ is the Horvitz-Thompson estimator of the population totals of the auxiliary vector from the respondent sample, $\hat{t}_{r_{y\pi}}$ is the Horvitz-Thompson estimator of the population total of the outcome variable from the respondent sample, and $\hat{\mathbf{B}}_r$ is estimated from the respondent sample as the solution of the weighted least squares equation.

$$\hat{\mathbf{t}}_{r_{x\pi}} = \sum_r d_k \mathbf{x}_k \tag{3.9}$$

$$\hat{t}_{r_{y\pi}} = \sum_r d_k y_k \tag{3.10}$$

$$\hat{\mathbf{B}}_r = \left(\sum_r \frac{\mathbf{x}_k \mathbf{x}_k^{\mathrm{T}}}{\sigma_k^2 \pi_k} \right)^{-1} \sum_r \frac{\mathbf{x}_k y_k^{\mathrm{T}}}{\sigma_k^2 \pi_k} \tag{3.11}$$

For the purpose of comparison, we examine the GREG estimator that accounts for only the main effects of the auxiliary variables, referred to as GREG_Main.

Deville and Särndal (1993) show that all calibration estimators built from the same set of covariates are asymptotically equivalent when there is full response. As a result, a GREG estimator that uses only main effects for a set of factor variables and a raking estimator that uses the margins for those factors should be approximately the same in large samples. Little and Wu (1991) and Little (1993) also show that the raking estimator has a Bayesian interpretation when cell means follow a main effects model and the probability of a unit's responding in cell *ij* is the product of row and column probabilities of response. Whether these results hold empirically when there is nonresponse is tested in subsequent sections.

3.1.4 Comparison between Poststratification, Raking, and GREG_Main

There are several motivations for comparing raking, GREG_Main, and poststratification. First, raking and GREG_Main share the same set of auxiliary variables, and their difference lies in the form of distance function $G(\cdot)$ and the corresponding adjustment function $F(\cdot)$. In the pure sampling context as discussed in Deville and Särndal (1992), these two estimators are asymptotically equivalent. That is, conditioning on the same set of auxiliary variables, the particular form of the distance function has negligible impact

on the asymptotic properties of the calibration estimator if non-sampling error does not exist. However, the conclusion in Deville and Särndal (1992) does not necessarily hold when nonresponse exists and calibration is used to reduce nonresponse bias. The theoretical results in Chapter 2 suggest that the difference between raking and GREG_Main could be as large as $O_p(1)$. The question is in what situation the two estimators tend to give very similar results and in what situations they tend to diverge significantly.

Second, GREG_Main and poststratification both belong to the GREG estimator family, although poststratification is usually not thought of in terms of calibration constraints and a distance function. GREG_Main accounts for only the main effects of the auxiliary variables while poststratification accounts for the interaction effects as well. The comparison of these two estimators shows the impact of the interaction terms in the outcome variable model and/or response model. The results can help us refine the guidelines in Little and Vartivarian (2005) for choosing auxiliary variables in nonresponse adjustment weighting.

Third, poststratification and raking are probably the two most commonly used calibration estimators in U.S. government surveys. From the practical perspective, a key difference between poststratification and raking is that the former fits a fully saturated model with both main and interaction effects of the auxiliary variables, while the latter fits a model including only the main effects. On the other hand, Deville, Särndal, and Sautory (1993) refer to poststratification as "complete poststratification" and raking ratio as "incomplete

poststratification". The literature shows that raking (through proportional fitting) preserves the multiplicative interaction effect in the sample data (Haberman 1979), although there is little theoretical or empirical research to suggest whether raking or GREG_Main may be superior in the situation that poststratification should be the most appropriate estimator. We attempt to investigate whether and to what extent raking can get closer to poststratification compared to how GREG_Main does.

3.2 Scope of Simulation Study

The simulation study aims to evaluate the empirical properties of the poststratification estimator, raking estimator, and GREG_Main estimator for finite population totals and means when calibration is used for nonresponse adjustment in a one-step weighting approach. We measure the magnitude of their differences in terms of empirical bias, variance, MSE, and coverage rate of 95 percent confidence intervals, under different model assumptions for the outcome variable and the nonresponse mechanism. The research is conducted in the following scope.

First, we evaluate estimates for population totals and means for a single outcome variable. In the presence of nonresponse, calibration is used to reduce the bias, variance, and MSE of the estimate for this single outcome variable.

Second, although Chapter 1 points out that it is possible to use a covariate vector $\mathbf{z}_k$ for the calibration adjustment function $F(\cdot)$ that is different from the auxiliary vector $\mathbf{x}_k$ in

the calibration equation, our evaluation focuses on the situation where $\mathbf{z}_k = \mathbf{x}_k$, which is the "standard choice" in Särndal and Lundström (2005).

Third, the outcome variable model and the response model contain the same main effect covariates. We also assume that there are only two main effect covariates and they are both categorical variables. In some scenarios, there is interaction between the two categorical variables because the effect of one variable depends on the value of the other variable. In these scenarios, an interaction term, assessed through either an additive model or a multiplicative model (to be discussed in greater detail in Section 3.3), is also included in the outcome variable model and/or response model.

Fourth, for the response mechanism, we assume missing at random (MAR). This means that the probability of response does not depend on the outcome variable once we control for the known covariates. The classes or cells defined by the covariates are response homogeneity groups.

Finally, the results focus on overall estimates in the context of SRS. Although practical surveys almost always involve complex sample designs, the SRS assumption allows us to focus on the impact of population structure and response mechanism on the performance of a calibration estimator. The findings about how to choose auxiliary variables and calibration estimators apply in general to complex designs, although the technical details become more complicated.

3.3 Outcome Variable Models and Response Models

One question to be answered through the simulation study is whether and how the three calibration estimators account for the interaction term in the outcome variable model and/or response propensity model. There are two types of interaction effects, referred to as "additive interaction" and "multiplicative" interaction respectively. Additive interaction model (also known as absolute difference model) evaluates whether the effects of one variable are the same across categories of another variable. The multiplicative interaction model (also known as relative difference model) examines whether the odds ratios or risk ratios by one variable are homogeneous across categories of another variable. In our simulation work, the interaction effect in the outcome variable model is assessed using an additive interaction model. For the response propensity model, the interaction effect is assessed through both multiplicative interaction model and additive interaction model because we are interested in evaluating the impacts of both types of interaction effects.

All the outcome variable models and response propensity models include two main effect covariates. In some scenarios, the models also include an interaction term in addition to the main effects. The alternative models for the outcome variable Y, Y_Main and Y_Additive_Interaction, are specified in (3.12) and (3.13).

Y_Main:

$$Y_{ijk} = \mu_Y + \alpha_{Yi} + \beta_{Yj} + \varepsilon_{Yijk},\ i = 1,\ 2;\ j = 1,\ 2;\ k = 1,\ \ldots,\ N_{ij} \tag{3.12}$$

Y_Additive_Interaction:

$$Y_{ijk} = \mu_Y + \alpha_{Yi} + \beta_{Yj} + \gamma_{Yij} + \varepsilon_{Yijk}, \quad i = 1,\ 2; j = 1,\ 2; k = 1,\ \ldots,\ N_{ij} \tag{3.13}$$

where N_{ij} is the population size in cell ij for the survey and $\varepsilon_{Yijk} \sim N(0, \sigma^2)$.

We refer to (3.12) as the "Y_Main" model because there are only main effect terms in the model. We refer to (3.13) as the "Y_Additive_Interaction" model because in addition to the main effects, a non-zero addictive interaction term is also included in the model.

For response, we use two models to describe the association and interaction patterns between two categorical random variables – a linear model that allows us to study the effect of additive interaction and a log-linear model that allows us to study the effect of multiplicative interaction. It matters what type of interaction (multiplicative versus additive) is included in the response model because the literature shows that raking forces the weights to conform to the marginal totals without perturbing the associations in the unadjusted table (Haberman 1979). That is, the raking process is expected to preserve the *multiplicative* interaction effect, but not necessarily the additive interaction effect, that already exists in the cell counts before calibration.

Linear response model:

$$R_{ij} = \mu_R + \alpha_{Ri} + \beta_{Rj} + \gamma_{Rij}, \qquad i = 1,\ 2; j = 1,\ 2 \tag{3.14}$$

Log-linear response model:

$$\log(R_{ij}) = \mu_R + \alpha_{Ri} + \beta_{Rj} + \gamma_{Rij}, \qquad i = 1,\ 2; j = 1,\ 2 \tag{3.15}$$

where R_{ij} is the response rate for cell ij.

In practice, both response models should satisfy the constraint $0 \le R_{ij} \le 1$. During the simulation, we choose not to manipulate the parameters in (3.14) and (3.15) directly. Instead, we randomly assign response flags (1 indicating response and 0 indicating nonresponse) for all the cases in each of the four cells formed by the two random categorical variables using the binomial distribution with parameters N_{ij} and R_{ij}, where N_{ij} and R_{ij} are the population count and response rate, respectively, for the cell ij. This method allows us to control the strength of both the main effects and the two types of interaction effects in the response model directly.

We assess the strength of additive interaction effect in the response model using $DIFF_{RR}$ defined in (3.16), which measures the extent to which the effect of the two variables together exceeds the effect of each considered individually.

$$\begin{aligned} DIFF_{RR} &= (R_{22} - R_{11}) - [(R_{21} - R_{11}) + (R_{12} - R_{11})] \\ &= R_{22} - R_{21} - R_{12} + R_{11} \end{aligned} \tag{3.16}$$

where R_{11}, R_{12}, R_{21}, and R_{22} are the response rates for the four cells formed by the categories of the two random independent variables.

Under model (3.14), equation (3.16) can be expressed as

$$\begin{aligned} DIFF_{RR} &= (\mu_R + \alpha_{R2} + \beta_{R2} + \gamma_{R22}) - (\mu_R + \alpha_{R2} + \beta_{R1} + \gamma_{R21}) \\ &\quad -(\mu_R + \alpha_{R1} + \beta_{R2} + \gamma_{R12}) + (\mu_R + \alpha_{R1} + \beta_{R1} + \gamma_{R11}) \\ &= \gamma_{R22} - \gamma_{R21} - \gamma_{R12} + \gamma_{R11} \end{aligned} \tag{3.17}$$

The sufficient condition for $DIFF_{RR} = 0$ is $\gamma_{R22} - \gamma_{R21} - \gamma_{R12} + \gamma_{R11} = 0$. One special situation is that when $\gamma_{Rij} = 0 \ \forall \ i, j$, we have $DIFF_{RR} = 0$. The nearer that $DIFF_{RR}$ is to zero, the lower is the effect of additive interaction.

To assess multiplicative interaction, we first define relative risks as in (3.18), (3.19), and (3.20).

$$RR_{22} = R_{22} / R_{11} \tag{3.18}$$

$$RR_{21} = R_{21} / R_{11} \tag{3.19}$$

$$RR_{12} = R_{12} / R_{11} \tag{3.20}$$

Then we calculate cross product ratio of response rates, CPR_{RR}, as in (3.21).

$$\begin{aligned} CPR_{RR} &= RR_{22} / (RR_{21} \times RR_{12}) \\ &= \frac{R_{22} \times R_{11}}{R_{21} \times R_{12}} \end{aligned} \tag{3.21}$$

Under model (3.15), equation (3.21) can be expressed as

$$\begin{aligned} CPR_{RR} &= \frac{R_{22} \times R_{11}}{R_{21} \times R_{12}} \\ &= \frac{e^{\mu_R + \alpha_{R2} + \beta_{R2} + \gamma_{R22}} e^{\mu_R + \alpha_{R1} + \beta_{R1} + \gamma_{R11}}}{e^{\mu_R + \alpha_{R2} + \beta_{R1} + \gamma_{R21}} e^{\mu_R + \alpha_{R1} + \beta_{R2} + \gamma_{R12}}} \\ &= e^{\gamma_{R22} - \gamma_{R21} - \gamma_{R12} + \gamma_{R11}} \end{aligned} \tag{3.22}$$

The sufficient condition for $CPR_{RR}=1$ is $\gamma_{R22}-\gamma_{R21}-\gamma_{R12}+\gamma_{R11}=0$. One special situation is that when $\gamma_{Rij}=0\ \forall\ i, j$, we have $CPR_{RR}=1$, indicating that the multiplicative interaction effect is zero; otherwise the multiplicative interaction effect is non-zero. The farther that CPR_{RR} is from one, the stronger is the effect of multiplicative interaction.

Here is a hypothetical example for the simulation setup. We are interested in a single outcome variable, income. Both the outcome variable and the response propensities can be fully explained by two dichotomous variables, education (high versus low) and age (young versus old), and possibly an interaction term between education and age. In the outcome variable and response rate models specified in (3.12) through (3.15), $\boldsymbol{\alpha}_Y=(\alpha_{Y1},\alpha_{Y2})$ and $\boldsymbol{\alpha}_R=(\alpha_{R1},\alpha_{R2})$ indicate the main effects of education on income and response rate respectively, $\boldsymbol{\beta}_Y=(\beta_{Y1},\ \beta_{Y2})$ and $\boldsymbol{\beta}_R=(\beta_{R1},\ \beta_{R2})$ indicate the main effects of age on income and response rate respectively, $\boldsymbol{\gamma}_Y=(\gamma_{Y11},\gamma_{Y12},\gamma_{Y21},\gamma_{Y22})$ denotes the additive interaction effect between education and age on income, and $\boldsymbol{\gamma}_R=(\gamma_{R11},\gamma_{R12},\gamma_{R21},\gamma_{R22})$ denotes the interaction effect (measured on either additive or multiplicative scale) between education and age on response rate. The interaction effect in the outcome variable model is assessed only on additive scale, so we have two scenarios: Y_Main and Y_Additive_Interaction. For the response model, 17 scenarios are created with different combinations of $DIFF_{RR}$ and CPR_{RR} values because we are interested in assessing the interaction term on both multiplicative scale and additive scale (see details in Section 3.4).

In poststratification, the weights are adjusted by four full-classification cells defined by the combinations of education and age. In raking, the weighting adjustment is conducted iteratively by using age and education as marginal controls until convergence is achieved. In GREG_Main, the calibration estimator is a function of the regression coefficient as the result of modeling the outcome variable income by only the main effects of education and age. Through the simulation study, we examine the performance of poststratification, raking, and GREG_Main under the scenarios created by the different outcome variable model and response model combinations. We evaluate the consistency between our results and those in Little & Vartivarian (2005) and refine their conclusions. At the same time, we attempt to expand Deville and Särndal (1992) and shed light on the empirical difference between the GREG estimators (i.e., GREG_Main and poststratification) and the raking estimator (as an example of the general calibration estimator) in the presence of nonresponse.

3.4 Simulation Scenarios and Steps

Several factors may affect the properties of and differences between the three calibration estimators under evaluation, including: (1) the number of simulation samples; (2) the overall sample size for the respondent sample and the distribution across the four subpopulations; (3) the substantive and statistical significance of the additive interaction effect in the outcome variable model; and (4) the strength of the multiplicative and additive interaction effect in the response model, measured by CPR_{RR} and $DIFF_{RR}$

respectively. In this section, we choose the simulation parameters so as to minimize the factors that could cloud our comparison between the three calibration estimators:

First, we use 1,000 simulation samples to evaluate the performance of the three calibration estimators over repeated sampling. Due to the large number of simulation iterations, it is unlikely that any observed differences between the estimators are due to chance.

Second, we know that the necessary condition for the additive interaction effect in the outcome variable model to be non-zero is $\gamma_{Y22} - \gamma_{Y21} - \gamma_{Y12} + \gamma_{Y11} \neq 0$. For Y_Main, we set $\gamma_{Yij} = 0 \ \forall \ i, j$ for simplicity. For Y_Additive_Interaction, the strength of the interaction effect can be controlled through the values for γ_{Yij}. Two criteria are followed when choosing the parameters for the outcome variable models. First, the random error terms in (3.12) and (3.13) should be very small such that the models have very strong predictive power. Second, the interaction effect in the Y_Additive_Interaction model should be substantively and statistically significant. Under these criteria, several sets of parameters for the outcome variable models have been used for test runs. The results associated with these different parameters lead to the same conclusions, so we choose to present the results based on only one set of parameters for Y_Main in (3.12) and Y_Additive_Interaction in (3.13), as shown below.

$$\mu_Y = 1000$$

$$\boldsymbol{\alpha}_Y = (\alpha_{Y1},\ \alpha_{Y2}) = (-200,\ 300)$$

$$\boldsymbol{\beta}_Y = (\beta_{Y1},\ \beta_{Y2}) = (-100,\ 150)$$

$$\boldsymbol{\gamma}_Y = (\gamma_{Y11}, \gamma_{Y12}, \gamma_{Y21}, \gamma_{Y22}) = (100,\ 300,\ 700,\ 1200)$$

$$\varepsilon_{Yijk} \sim N(0, 900)$$

Using these parameters, finite populations with approximately 40,000 units are generated for Y_Main and Y_Additive_Interaction, respectively. Then for each finite population, a regression model of the outcome variable is fitted (on both the main effect variables and the interaction term) to check the predictive power of the model as well as the strength of the interaction term. Table 3.1 shows the cell means, the R-squared value of the regression model, and the *p*-values for the interaction term for each finite population. The R-squared values are close to one for both Y_Main and Y_Additive_Interaction, indicating almost perfect prediction power of the outcome models. For Y_Main, the *p*-value for the interaction term is close to one. For Y_Additive_Interaction, the *p*-value for the interaction term is almost zero.

Table 3.1 Two Finite Populations Corresponding to Two Outcome Variable Models

Outcome Variable Model	$E_M(y_{11k})$	$E_M(y_{12k})$	$E_M(y_{21k})$	$E_M(y_{22k})$	R-squared for Overall Model	*p*-value for Interaction Term
Y_Main	700	950	1,200	1,450	0.9886	0.998
Y_Additive_Interaction	800	1,250	1,900	2,650	0.9979	<0.0001

Third, for the response model, we create 17 scenarios with different combinations of $DIFF_{RR}$ and CPR_{RR} values, as shown in Table 3.2. This is achieved by manipulating the four cell response rates ($R_{11}, R_{12}, R_{21}, R_{22}$). Models (3.14) and (3.15) are not used directly to generate responses. However, given a set of R_{ij}, CPR_{RR} and $DIFF_{RR}$ can be computed

to measure the strength of multiplicative interaction effect and the strength of additive interaction effect respectively. These scenarios can be divided into five categories: (1) full response (scenario S01), which serves as the evaluation baseline; (2) neither multiplicative nor additive interaction (scenario S02); (3) only additive interaction but no multiplicative interaction (scenario S03); (4) only multiplicative interaction but no additive interaction (scenarios S04 through S11); and (5) both types of interaction (scenarios S12 through S17). The direction and strength of the multiplicative interaction also vary among scenarios S4 through S17. These combinations allow us to understand whether and how these two different types of interaction effects in the response model affect the performance of poststratification, raking, and GREG_Main.

For Scenario S02, to ensure both additive independency and multiplicative independency for the response rates in a 2×2 table, the two conditions shown in (3.23) and (3.24) need to be satisfied.

$$DIFF_{RR} = R_{11} + R_{22} - R_{12} - R_{21} = 0 \tag{3.23}$$

$$CPR_{RR} = \frac{R_{22} \times R_{11}}{R_{21} \times R_{12}} = 1 \tag{3.24}$$

Putting (3.23) and (3.24) together, we have

$$
\begin{aligned}
& R_{11} + \frac{R_{21} \times R_{12}}{R_{11}} = R_{12} + R_{21} \\
& \Leftrightarrow R_{11} * R_{11} + R_{21} \times R_{12} = R_{11} * R_{12} + R_{11} * R_{21} \\
& \Leftrightarrow R_{11} * R_{11} - R_{11} * R_{12} = R_{11} * R_{21} - R_{21} \times R_{12} \\
& \Leftrightarrow R_{11}(R_{11} - R_{12}) = R_{21}(R_{11} - R_{12}) \\
& \Leftrightarrow (R_{11} - R_{21})(R_{11} - R_{12}) = 0 \\
& \Leftrightarrow R_{11} = R_{21} \text{ or } R_{11} = R_{12}
\end{aligned}
\tag{3.25}
$$

This means that the response rates should be independent of either the row variable or the column variable. That is, the response model essentially contains only one covariate, not two covariates.

Table 3.2 Scenarios for Response Models

Category	Scenario Number	R_{11}	R_{12}	R_{21}	R_{22}	CPR_{RR}	$DIFF_{RR}$
Full Response	S01	1.00	1.00	1.00	1.00	1.00	0.00
Neither Additive Interaction Nor Multiplicative Interaction	S02	0.45	0.45	0.30	0.30	1.00	0.00
Only Additive Interaction	S03	0.41	0.10	0.95	0.24	1.00	-0.41
Only Multiplicative Interaction	S04	0.12	0.48	0.02	0.38	4.75	0.00
	S05	0.26	0.94	0.06	0.74	3.41	0.00
	S06	0.28	0.92	0.08	0.72	2.74	0.00
	S07	0.32	0.88	0.12	0.68	2.06	0.00
	S08	0.40	0.80	0.20	0.60	1.50	0.00
	S09	0.46	0.74	0.26	0.54	1.29	0.00
	S10	0.54	0.66	0.34	0.46	1.11	0.00
	S11	0.56	0.64	0.36	0.44	1.07	0.00
Both Types of Interaction	S12	0.23	0.07	0.55	0.15	0.90	-0.24
	S13	0.20	0.10	0.52	0.18	0.69	-0.24
	S14	0.15	0.15	0.47	0.23	0.49	-0.24
	S15	0.09	0.21	0.41	0.29	0.30	-0.24
	S16	0.04	0.26	0.36	0.34	0.15	-0.24
	S17	0.02	0.58	0.66	0.74	0.04	-0.48

Finally, the respondent sample sizes by cell are determined by both the overall SRS sample size n and the response model. For each outcome variable and response model

combination, we vary the respondent sample sizes by using three alternative SRS sample sizes: n=8,000, n=2,000, and n=200. This allows us to not only evaluate the asymptotic properties of the calibration estimators, but also see the impact of small cell counts on poststratification in some scenarios.

In summary, the simulation study covers 102 scenarios formed by crossing two outcome variable scenarios, 17 response model scenarios, and three alternative SRS sample sizes. The following steps are used to evaluate the properties of the three calibration estimators over repeated sampling.

Step I: Generate two finite populations corresponding to the outcome variable models Y_Main and Y_Additive_Interaction in Table 3.1, respectively. Each finite population contains four subpopulations defined by the categories of the two auxiliary variables. The subpopulation sizes, N_{ij} , are determined through Poisson distribution with mean of 10,000. The overall size for each finite population is approximately 40,000, with approximately equal number of cases in each of the four cells. These two finite populations are used for repeated sampling in the steps that follow.

Step II: From each finite population, first select a simple random sample of size n , and then select a subsample of respondents from the simple random sample using one of the response models shown in Table 3.2. With the three alternative SRS sample sizes (n=8,000, n=2,000, and n=200) and 17 response model scenarios, this step results in 51 respondent samples from each finite population for each simulation iteration.

Step III: Conduct calibration on each respondent sample using poststratification, raking, and GREG_Main, respectively. Obtain the estimates for the outcome variable associated with the three calibration estimators.

Step IV: Repeat Steps II and III for 1,000 times. This results in 1,000 iterations for each of the 102 simulation scenarios.

Step V: For each of the 102 simulation scenarios, examine the empirical properties of poststratification, raking, and GREG_Main over repeated sampling using the 1,000 simulation samples and the evaluation criteria described in Section 3.5.

The simulation results over repeated sampling are reported in Section 3.7. We then conduct some sensitivity analysis in Section 3.8 by varying the predictive power of the outcome variable model.

The simulation is conducted in R (Lumley, 2005; R Development Core Team, 2015) because of its efficiency in handling matrix calculations and extensive capacity for analyzing survey data. The programs developed for the simulation studies are provided in Appendix B.

3.5 Evaluation Criteria

We examine the empirical properties of the calibration estimators using the repeated sampling approach (i.e., averaging across the 1,000 simulation samples). The empirical properties of the three calibration estimators under different outcome variable model, response model, and SRS sample size combinations are compared using several measures. The measures are described below in terms of totals. A similar set of measures can be used to evaluate the properties of the estimators in terms of overall means.

1. Relative bias $RelBias(\hat{t}_{yw}) = (1/S)\sum_{s=1}^{S}(\hat{t}_{yw_s} - t_y)\Big/t_y$

   where s indicates a particular sample, S is the total number of samples included. t_y is the true population total, and $\hat{t}_{yw_s}$ is the estimate from sample s using one of the three calibration estimators.

2. Empirical relative standard error

   $$EmpRelSE(\hat{t}_{yw}) = \sqrt{EmpVar(\hat{t}_{yw})}\Big/t_y = \sqrt{(1/S)\sum_{s=1}^{S}\left(\hat{t}_{yw_s} - E_p(\hat{t}_{yw_s})\right)^2}\Big/t_y$$

   where $E_p(\hat{t}_{yw}) = (1/S)\sum_{s=1}^{S}\hat{t}_{yw_s}$, the average value of $\hat{t}_{yw_s}$ over repeated sampling.

   $EmpVar(\hat{t}_{yw})$ is the empirical variance across the S simulated samples, not the average of the S estimated variances computed by the R software for all the simulated samples.

3. Relative square root of MSE

$$
\begin{aligned}
&RelRMSE(\hat{t}_{yw}) \\
&= \sqrt{\text{MSE}(\hat{t}_{yw})} \Big/ t_y \\
&= \sqrt{(1/S)\sum_{s=1}^{S}\left(\hat{t}_{yw_s} - E_p(\hat{t}_{yw_s})\right)^2 + \left((1/S)\sum_{s=1}^{S}(\hat{t}_{yw_s} - t_y)\right)^2} \Big/ t_y
\end{aligned}
$$

4. Coverage rate of the 95 percent confidence intervals

$$(1/S)\sum_{s=1}^{S} I\left(\left|\hat{z}_j\right| \le z_{1-\alpha/2}\right), \text{ where } \alpha = 0.05 \text{ and } \hat{z}_j = (\hat{t}_{yw_s} - t_y) \Big/ \sqrt{var(\hat{t}_{yw})}$$

where $var(\hat{t}_{yw})$ is the estimated variance for each simulated sample computed using the "calibrate" function in the R Survey package; $I(\square)$ is an indicator for whether $\left|\hat{z}_j\right| \le z_{1-\alpha/2}$. The method essentially estimates the variance of a linear substitute that is equivalent to the product of the calibrated weight and a residual calculated from a linear model of the outcome variable on a vector of auxiliary variables. For raking, the residual is based on a main effect model with the covariates being indicators for the raking categories of each dimension. We use $var(\hat{t}_{yw})$, the estimated variance from each simulated sample (instead of $EmpVar(\hat{t}_{yw})$, the empirical variance estimated from all the simulated samples), to obtain the 95 percent confidence interval because only one sample can be obtained for any survey in practice. The limitation of this approach is that it relies on the accuracy of the variance estimation method implemented in the "calibrate" function of the R Survey package. More details about the impact of the variance estimation method is discussed in Chapter 5.

5. Bias ratio, calculated as the ratio of $Bias(\hat{t}_{yw})$ and square root of $EmpVar(\hat{t}_{yw})$

$$BiasRatio(\hat{t}_{yw}) = (1/S)\sum_{s=1}^{S}(\hat{t}_{yw_s} - t_y) \Big/ \sqrt{(1/S)\sum_{s=1}^{S}\left(\hat{t}_{yw_s} - E_p(\hat{t}_{yw})\right)^2}$$

3.6 Expected Results from Simulation

We anticipate the results for the overall totals and those for the overall means to follow the same pattern because the denominator of the estimator for the overall mean is calibrated to the overall population count, which is a constant for any finite population. The properties of the three calibration estimators are expected to depend on the outcome variable model, response model, overall sample size, and existence of small cell counts.

3.6.1 Expected Impacts of Outcome Variable Model and Response Model

First of all, we anticipate the outcome variable model to be the primary driving factor for determining the performance of the calibration estimators. When the outcome variable model contains only the main effect terms (in the Y_Main scenarios), we expect the three calibration estimators to perform similarly well regardless of the form of the response model. The response model matters only when the outcome variable model includes an additive interaction term. In the Y_Additive_Interaction scenarios, the three calibration estimators are expected to perform differently. Moreover, the response model is expected to play an important role in the properties of the raking estimator and GREG_Main estimator. Out of the three estimators of interest, we find the exact theory about

GREG_Main difficult to develop. By definition, the GREG_Main estimator accounts for only the main effects. Under the Y_Additive_Interaction outcome model, GREG_Main is expected to be biased in all the response scenarios except S01 (with full response) and S02 (when the response model contains neither multiplicative interaction effect nor additive interaction effect, and thus depends on a single covariate essentially, as shown in (3.25)).

3.6.2 Theoretical Development about Poststratification

Poststratification accounts for the interaction term in the outcome variable model as shown below. As long as all the respondent cell counts are reasonably large, the poststratification estimator is expected to perform well regardless of the response model. We can prove that the poststratification estimator is model-unbiased.

Assume the simulation population is a realization of the super population generated by (3.13). Let the population total of the outcome variable Y be $t_y = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k\in U_{ij}} y_k$. Then the expectation of t_y with respect to the outcome variable model is

$$
\begin{aligned}
\mathrm{E_M}(t_y) &= \mathrm{E_M}\Big(\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k\in U_{ij}} y_k\Big) \\
&= \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k\in U_{ij}} \mathrm{E_M}(y_k) \\
&= \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k\in U_{ij}} \mu_Y + \alpha_{Yi} + \beta_{Yj} + \gamma_{Yij} \\
&= \sum_{i=1}^{r}\sum_{j=1}^{c} N_{ij}(\mu_Y + \alpha_{Yi} + \beta_{Yj} + \gamma_{Yij})
\end{aligned}
\tag{3.26}
$$

Now define the following Horvitz-Thompson estimators

$$\hat{N}_{ij} = \sum_{k\in s_{ij}} d_k \tag{3.27}$$

$$\hat{N}_{rij} = \sum_{k\in r_{ij}} d_k = \sum_{k\in s_{ij}} \delta_k d_k \tag{3.28}$$

$$\hat{t}_{ysij} = \sum_{k\in s_{ij}} d_k y_k \tag{3.29}$$

$$\hat{t}_{yrij} = \sum_{k\in r_{ij}} d_k y_k = \sum_{k\in s_{ij}} \delta_k d_k y_k \tag{3.30}$$

where δ_k is the response indicator and d_k is the basic design weight for unit *k* in cell *ij*.

$$\delta_k = \begin{cases} 1 & \text{if response} \\ 0 & \text{if nonresponse} \end{cases}$$

$d_k = 1/\pi_k$, where π_k is the inclusion probability for unit *k* in cell *ij*.

The expectations of $\hat{N}_{rij}$ and $\hat{t}_{yrij}$ with respect to the response models (3.14) and (3.15) are

$$\mathrm{E_R}(\hat{N}_{rij}) = \sum_{k\in s_{ij}} \mathrm{E_R}(\delta_k d_k) = \sum_{k\in s_{ij}} d_k \mathrm{E_R}(\delta_k) = \sum_{k\in s_{ij}} d_k R_{ij} = R_{ij}\hat{N}_{ij} \tag{3.31}$$

and

$$\mathrm{E_R}(\hat{t}_{yrij}) = \sum_{k\in s_{ij}} \mathrm{E_R}(\delta_k d_k y_k) = \sum_{k\in s_{ij}} d_k y_k \mathrm{E_R}(\delta_k) = \sum_{k\in s_{ij}} d_k y_k R_{ij} = R_{ij}\hat{t}_{ysij} \tag{3.32}$$

The poststratification estimator in (3.2) can be expressed as

$$\hat{t}_{yps} = \sum_{i=1}^{r}\sum_{j=1}^{c} N_{ij} \frac{\sum_{k\in s_{ij}} \delta_k d_k y_k}{\sum_{k\in s_{ij}} \delta_k d_k} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{N_{ij}}{\hat{N}_{rij}} \hat{t}_{yrij} \tag{3.33}$$

Then using linear approximation, we can obtain

$$\mathrm{E_R}(\hat{t}_{yps}) = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{N_{ij}}{R_{ij}\hat{N}_{ij}} R_{ij}\hat{t}_{ysij} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{N_{ij}}{\hat{N}_{ij}} \hat{t}_{ysij} \tag{3.34}$$

and

$$\begin{aligned}
&\mathrm{E_M E_R}(\hat{t}_{yps}) \\
&\approx \mathrm{E_M}\Big(\sum_{i=1}^{r}\sum_{j=1}^{c} \frac{N_{ij}}{\hat{N}_{ij}} \hat{t}_{ysij}\Big) \\
&\approx \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{N_{ij}}{\hat{N}_{ij}} \mathrm{E_M}(\hat{t}_{ysij}) \\
&= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{N_{ij}}{\hat{N}_{ij}} \sum_{k\in s_{ij}} d_k \mathrm{E_M}(y_k) \\
&= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{N_{ij}}{\hat{N}_{ij}} \sum_{k\in s_{ij}} d_k (\mu_Y + \alpha_{Yi} + \beta_{Yj} + \gamma_{Yij}) \\
&= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{N_{ij}}{\hat{N}_{ij}} \hat{N}_{ij} (\mu_Y + \alpha_{Yi} + \beta_{Yj} + \gamma_{Yij}) \\
&= \sum_{i=1}^{r}\sum_{j=1}^{c} N_{ij} (\mu_Y + \alpha_{Yi} + \beta_{Yj} + \gamma_{Yij}) \\
&= \mathrm{E_M}(t_y)
\end{aligned} \tag{3.35}$$

So the poststratification estimator $\hat{t}_{yps}$ is model-unbiased under the outcome models and response models specified in (3.12) through (3.15).

3.6.3 Theoretical Development about Raking

Implications of Raking Maintaining Multiplicative Interaction Effect

Although raking does not explicitly account for the interaction term in the outcome variable model, the iterative proportional fitting algorithm forces the weights to conform to the marginal totals without perturbing the associations in the unadjusted table (Haberman 1979). To understand the implications of this in the setting of a 2×2 table, we define several cross product ratios of unweighted and weighted cell counts, including CPR_{pop} for the population cell counts, CPR_s for the weighted sample cell counts using the basic design weights, CPR_r for the weighted respondent cell counts using the basic design weights, and CPR_w for the weighted respondent cell counts using the calibrated weights from raking, poststratification, or GREG_Main.

$$CPR_{pop} = \frac{N_{11}N_{22}}{N_{12}N_{21}} \tag{3.36}$$

$$CPR_s = \frac{\hat{N}_{11}\hat{N}_{22}}{\hat{N}_{12}\hat{N}_{21}} = \frac{\sum_{k\in s_{11}} d_k \sum_{k\in s_{22}} d_k}{\sum_{k\in s_{12}} d_k \sum_{k\in s_{21}} d_k} \tag{3.37}$$

$$CPR_r = \frac{\hat{N}_{r11}\hat{N}_{r22}}{N_{r12}\hat{N}_{r21}} = \frac{\sum_{k\in r_{11}} d_k \sum_{k\in r_{22}} d_k}{\sum_{k\in r_{12}} d_k \sum_{k\in r_{21}} d_k} \tag{3.38}$$

$$CPR_w = \frac{\hat{N}^w_{r11}\hat{N}^w_{r22}}{\hat{N}^w_{r12}\hat{N}^w_{r21}} = \frac{\sum_{k\in r_{11}} w_k \sum_{k\in r_{22}} w_k}{\sum_{k\in r_{12}} w_k \sum_{k\in r_{21}} w_k} \tag{3.39}$$

where $\hat{N}^w_{rij}$ denotes the estimated population count in cell *ij* using the calibrated weights from raking, poststratification, or GREG_Main, and w_k is the calibrated weight for unit *k* in cell *ij*.

First, the proportional fitting process for raking makes the weights conform to the row control totals during each row iteration and conform to the column control totals during each column iteration. Let f_{im} denote the weighting adjustment factor for the *i*th row during the *m*th row iteration and f_{jn} denote the weighting adjustment factor for the *j*th column during the *n*th column iteration. It is important to note that f_{im} and f_{jn} are independent of each other. Assume that when raking converges, the total number of row iterations is *M* and the total number of column iterations is *N*. Usually $M = N \pm 1$ in practice. Then the overall weighting adjustment factor for unit *k* in cell *ij*, shown as $F(u_i + v_j)$ in (3.5), can be calculated as

$$F_{ij} = \prod_{m=1}^{M} f_{im} \prod_{n=1}^{N} f_{jn}, \qquad i = 1,2;\ j = 1,2 \tag{3.40}$$

We can re-write (3.39) as

$$
\begin{aligned}
CPR_w &= \frac{(F_{11}\sum_{k\in r_{11}} d_k)(F_{22}\sum_{k\in r_{22}} d_k)}{(F_{12}\sum_{k\in r_{12}} d_k)(F_{21}\sum_{k\in r_{21}} d_k)} \\
&= \frac{(\prod_{m=1}^{M} f_{1m}\prod_{n=1}^{N} f_{1n}\sum_{k\in r_{11}} d_k)(\prod_{m=1}^{M} f_{2m}\prod_{n=1}^{N} f_{2n}\sum_{k\in r_{22}} d_k)}{(\prod_{m=1}^{M} f_{1m}\prod_{n=1}^{N} f_{2n}\sum_{k\in r_{12}} d_k)(\prod_{m=1}^{M} f_{2m}\prod_{n=1}^{N} f_{1n}\sum_{k\in r_{21}} d_k)} \\
&= \frac{\sum_{k\in r_{11}} d_k \sum_{k\in r_{22}} d_k}{\sum_{k\in r_{12}} d_k \sum_{k\in r_{21}} d_k}
\end{aligned}
\tag{3.41}
$$

From (3.38) and (3.41), it is easy to see that, for raking,

$$CPR_w = CPR_r \tag{3.42}$$

That is, the cross product ratio of the weighted respondent cell counts before raking (using the basic design weights) is the same as that after raking (using the raked weights). Brick, Montaquila, and Roth (2003) also provide a numerical example showing that raking retains the cross product ratio of the observed case counts.

Second, in the SRS setting where $d_k = N/n$, if $n_{rij} = E_R\left(n_{rij}\right) = n_{ij}R_{ij}$, then (3.38) can be re-expressed as

$$CPR_r = \frac{\sum_{k\in s_{11}} \delta_k d_k \sum_{k\in s_{22}} \delta_k d_k}{\sum_{k\in s_{12}} \delta_k d_k \sum_{k\in s_{21}} \delta_k d_k} = \frac{n_{r11}n_{r22}}{n_{r12}n_{r21}} = \frac{(n_{11}R_{11})(n_{22}R_{22})}{(n_{12}R_{12})(n_{21}R_{21})} = CPR_s * CPR_{RR} \tag{3.43}$$

where CPR_{RR} is the cross product ratio of the response rates, as defined in (3.21).

Third, we know that CPR_s converges to CPR_{pop} as the cell sample sizes $n_{11}, n_{12}, n_{21}, n_{22} \to \infty$. From (3.42) and (3.43), it is easy to see that, for raking, CPR_w approaches the product of CPR_{pop} and CPR_{RR} as the cell sample sizes become large.

Finally, in the special situation where there is no multiplicative interaction term in the response model (as in the response scenario S03), $CPR_{RR} = 1$. So for raking, CPR_w approaches CPR_{pop} as $n_{11}, n_{12}, n_{21}, n_{22} \to \infty$. That is, raking maintains the internal interaction effect in the population cell counts. Therefore we expect the raking estimator to perform almost as well as poststratification when $CPR_{RR} = 1$. At the same time, the relative bias of raking is expected to increase as the multiplicative interaction term in the response model becomes stronger and CPR_{RR} becomes farther away from 1.

A Sufficient Condition in Weighting Adjustment for Raking to Be Unbiased

In the SRS setting, the *Y*-model expectation of the raking estimator is

$$\begin{aligned}
&\mathrm{E_M}(\hat{t}_{yrk}) \\
&= \sum_{i=1}^{2}\sum_{j=1}^{2} w_{ij} \sum_{k=1}^{n_{rij}} \mathrm{E_M}(y_{ijk}) \\
&= \sum_{i=1}^{2}\sum_{j=1}^{2} w_{ij} \sum_{k=1}^{n_{rij}} \mathrm{E_M}(\mu_Y + \alpha_{Yi} + \beta_{Yj} + \gamma_{Yij} + \varepsilon_{Yijk}) \\
&= \sum_{i=1}^{2}\sum_{j=1}^{2} w_{ij} \sum_{k=1}^{n_{rij}} (\mu_Y + \alpha_{Yi} + \beta_{Yj} + \gamma_{Yij}) \\
&= \hat{N}_{rij}^{w} \left(\mu_Y + \alpha_{Yi} + \beta_{Yj} + \gamma_{Yij}\right)
\end{aligned} \tag{3.44}$$

The *Y*-model expectation of the population total is

$$E_M\left(t_y\right)=\sum_{i}^{r}\sum_{j}^{c}\sum_{k=1}^{N_{ij}}y_{ijk}=\sum_{i}^{r}\sum_{j}^{c}N_{ij}\left(\mu_Y+\alpha_{Yi}+\beta_{Yj}+\gamma_{Yij}\right) \tag{3.45}$$

Therefore the *Y*-model bias of the raking estimator is

$$E_M\left(\hat{t}_{yrk}-t_y\right)=\sum_{i}^{r}\sum_{j}^{c}\left(\hat{N}_{rij}^{w}-N_{ij}\right)\left(\mu_Y+\alpha_{Yi}+\beta_{Yj}+\gamma_{Yij}\right) \tag{3.46}$$

In a general situation, $\gamma_{Yij}\neq 0$. A sufficient (but not necessary) condition for $E_M\left(\hat{t}_{yrk}-t_y\right)=0$ is $\hat{N}_{rij}^{w}=N_{ij}$. This is true regardless of whether the outcome variable model contains the interaction effect.

The response indicator for a unit *k* in cell *ij* is $\delta_{ijk}=\begin{cases}1 & \text{if response}\\ 0 & \text{if nonresponse}\end{cases}$. The base-weighted estimate of the number of units in cell *ij* based on the respondents in the cell is $\hat{N}_{rij}=\sum_{k=1}^{n_{rij}}d_{ijk}$. The expectation of $\hat{N}_{rij}$ over the response model is

$$E_R\left(\hat{N}_{rij}\left|s_{ij}\right.\right)=E_R\left(\sum_{k=1}^{n_{ij}}\delta_{ijk}d_{ijk}\right)=\sum_{k=1}^{n_{ij}}d_{ijk}R_{ij}=R_{ij}\hat{N}_{ij} \tag{3.47}$$

where R_{ij} is the response probability for all the units in cell *ij*.

Also, from (3.40), we know $\hat{N}_{rij}^{w}=F_{ij}\hat{N}_{rij}$, so the expectation over repeated sampling distribution (E_p) and response distribution (E_R) is

$$E_P E_R\left(\hat{N}_{rij}^{w}\,\middle|\,s_{ij}\right)=E_P E_R\left(F_{ij}\hat{N}_{rij}\,\middle|\,s_{ij}\right)=E_P\left(F_{ij}R_{ij}\hat{N}_{rij}\right)=R_{ij}E_P\left(F_{ij}\hat{N}_{rij}\right)=F_{ij}R_{ij}N_{ij} \tag{3.48}$$

The derivation above treats F_{ij} as fixed, which is loose since the weighting adjustment actually varies from sample to sample. This is probably acceptable if we think of F_{ij} as the converged value for a given initial sample, i.e., across the response distribution.

The compound bias (i.e., over repeated sampling, outcome variable, and response distributions) for raking is

$$\begin{aligned} &E_P E_R E_M\left(\hat{t}_{yrk}-t_y\right) \\ &=E_P E_R\sum_{i}^{r}\sum_{j}^{c}\left(\hat{N}_{rij}^{w}-N_{ij}\right)\left(\mu_Y+\alpha_{Yi}+\beta_{Yj}+\gamma_{Yij}\right) \\ &=\sum_{i}^{r}\sum_{j}^{c}\left(F_{ij}R_{ij}N_{ij}-N_{ij}\right)\left(\mu_Y+\alpha_{Yi}+\beta_{Yj}+\gamma_{Yij}\right) \end{aligned} \tag{3.49}$$

If $F_{ij}=R_{ij}^{-1}$ (i.e., the raking adjustment factor in cell *ij* is the inverse of the cell response probability), then $\hat{t}_{yrk}$ is unbiased across all three distributions. In this situation, raking achieves what poststratification does. Note that this is the sufficient condition, but not necessary condition, for $E_P E_R E_M\left(\hat{t}_{yrk}-t_{yU}\right)=0$.

3.6.4 Expected Impacts of Sample Sizes

We expect the sample sizes to have two effects on the performance of the estimators. First, for a biased and inconsistent estimator, we suspect that purely increasing the sample size (without improving the calibration model) does not necessarily improve the

effective coverage rate of the confidence interval. Although the variance tends to decrease as the sample size increases, the relative bias does not change with the sample size. Thus, the confidence interval tends to become narrower as the sample size becomes larger and the variance becomes smaller. As a result, the coverage rate of the 95 percent confidence intervals for a biased estimator is expected to become worse as the sample size increases.

Second, assuming that both estimators are unbiased, one advantage of raking over poststratification is that when the marginal counts are large but some cell counts (formed by crossing the categories of the auxiliary variables) are small, raking may be more stable than poststratification. We include some simulation scenarios to test this hypothesis. For example, in the response scenarios S04 and S17 under the Y_Main model, the marginal response rates are high enough but the smallest cell-level response rates are only 2 percent. In these scenarios, particularly with SRS n=200, we expect the empirical relative standard errors for raking to be smaller than those for poststratification.

3.7 Simulation Results

In this section we examine the empirical properties of poststratification, raking, and GREG_Main over repeated sampling. As expected, the results for the overall totals are almost exactly the same as those for the overall means; any differences are negligible and only due to rounding. Therefore we only show the properties of the estimators for the totals in the discussions below.

Tables 3.3 and Table 3.4 show the relative bias, empirical relative standard error, relative square root of MSE, bias ratio, and coverage rate of the 95 percent confidence intervals for the three estimators under the outcome variable models Y_Main and Y_Additive_Interaction respectively. In addition, the average respondent sample sizes by cell and average cross product ratios for various unweighted and weighted cell counts are also presented in the tables because the information helps explain the results. For each outcome variable scenario, three sets of results are presented for SRS sizes n=8,000, n=2,000, and n=200, respectively. Although the results corresponding to full response (in scenario S01) are not our focus, they serve as the evaluation baselines, and are thus included in the tables.

3.7.1 Impact of Outcome Variable Model and Response Model on Bias

The relative bias is not only an important evaluation measure itself, but also a factor affecting the bias ratio and coverage rate of the 95 percent confidence intervals, which are discussed in Section 3.7.3 in greater detail. Among all the simulation scenarios in Tables 3.3 and 3.4, none of the three estimators is associated with unacceptably high relative bias. The biggest relative bias is only approximately 4 percent – It occurs for GREG_Main (with SRS n=8,000 and n=2,000) when the outcome variable model contains a substantively and statistically significant additive interaction term (Y_Additive_Interaction) and the response model contains a strong interaction term (scenario S17 with $DIFF_{RR} = -0.48$ and $CPR_{RR} = 0.04$). Overall, the three estimators

can all achieve significant bias reduction through accounting for at least the main effects of the auxiliary variables.

At the same time, the data confirms our expectation that the performances of the estimators are affected by the outcome variable model and response model. It is clear that the outcome variable model is the primary driving factor for determining whether there are any substantial differences between the relative biases of the estimators. In Table 3.3 (for the Y_Main model), poststratification does not reduce the nonresponse bias further than raking or GREG_Main, and this is true regardless of the response model. The three calibration estimators yield very similar relative biases in all the response scenarios including those with strong interaction term, and any noticeable differences between them (such as in S17) can be attributed to random error, which is discussed in Section 3.7.2 in greater detail. The key to understanding this data pattern is that if an auxiliary variable is correlated only to nonresponse but uncorrelated to the outcome variable, then the differential response related to the auxiliary variable does not cause any nonresponse bias. For example, in a hypothetical survey targeting both males and females, if the outcome variable depends only on age, then differential response rates by gender do not make the overall estimate biased, as long as the distribution of age is independent of gender. In our simulation setting, although the interaction effect is present in some response models, it does not affect the outcome variable, and thus does not introduce any nonresponse bias *in addition to* the nonresponse bias that has already been caused by the main effects. Since no bias has been caused by the interaction effect in the first place, it does not help to include the interaction term in the calibration process.

This is why raking and GREG_Main perform almost as well as poststratification in terms of relative bias in all the Y_Main scenarios regardless of the response model.

In contrast, an interaction term in the response model plays an important role in the performance of a calibration estimator *conditioning on* the fact that the interaction effect is present in the outcome variable model. Figure 3.1 shows the absolute values of the relative biases for poststratification, raking, and GREG_Main in the various response scenarios when the outcome model is Y_Additive_Interaction and the SRS sample size is 8,000. Using the absolute values allow us to better understand the relationship between the magnitude of relative bias and the strength of the interaction term in the response model. We can see three patterns from Table 3.4 and Figure 3.1.

First, the response scenarios S01 and S02 are two special situations. S01 is for full response. S02 occurs only when the response rates are independent of either the row variable or the column variable, which means that the response rates are essentially driven by a single variable. Not surprisingly, the three estimators perform similarly well in S01 and S02.

Second, the response scenarios S03 (with $CPR_{RR} = 1$ and $DIFF_{RR} = -0.41$) and S04 (with $CPR_{RR} = 4.75$ and $DIFF_{RR} = 0$) form an interesting pair of contrasts because the former contains only additive interaction term and the latter contains only multiplicative interaction term. For example, in the scenario S03 with *n*=2,000, the magnitude of the relative bias for poststratification (0.002 percent) and that for raking (0.043 percent) are

comparable, while the magnitude for GREG_Main (1.17 percent) is substantially higher. In the scenario S04 with n=2,000, the response model contains a strong multiplicative interaction term, and neither GREG_Main nor raking can match the performance of poststratification. The absolute values of the relative biases for raking and GREG_Main are approximately 126 times and 215 times, respectively, as large as that for poststratification. These results are consistent with our anticipation that under the outcome variable model Y_Additive_Interaction, the performance of raking is affected by the multiplicative interaction effect rather than the additive interaction effect in the response model, yet GREG_Main is biased if there is either type of interaction effect in the response model.

Third, the results for the response scenarios S05 through S17 confirm our expectation that the biasedness of raking is associated with the strength of the multiplicative interaction term in the response model (measured by how far off CPR_{RR} is from 1). For example, in the Y_Additive_Interaction and n=8,000 scenario, the absolute value of the relative bias for raking increases from 0.06 percent to 1.71 percent as CPR_{RR} increases from 1.07 to 4.75, and increases from 0.04 percent to 3.09 percent as CPR_{RR} decreases from 0.90 to 0.04. In the response scenarios S11 and S12 with relatively weak multiplicative interaction term, raking may be considered an acceptable estimator in terms of relative bias and coverage rate of the 95 percent confidence intervals (92 percent in S11 and 100 percent in S12), but its performance becomes worse as CPR_{RR} moves farther away from 1. The performance of GREG_Main for response scenarios S05 through S17 follows a similar pattern except that it is generally more biased than raking.

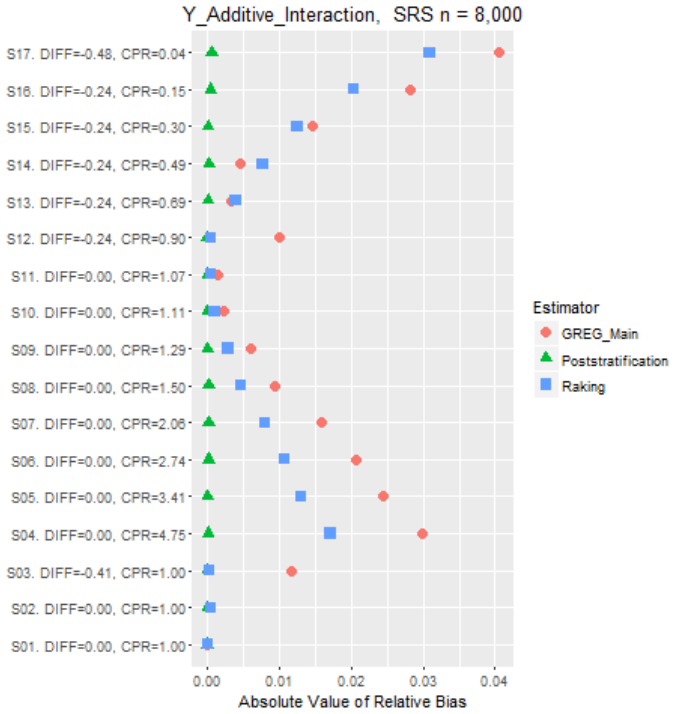


Figure 3.1 Absolute Values of Relative Biases for Poststratification, Raking, and GREG_Main for Outcome Model Y_Additive_Interaction, *n*=8,000, and Various Response Scenarios

Although the benchmark controls for raking are marginal totals instead of cell counts as in poststratification, raking performs almost as well as poststratification in the response scenario S03 (with $CPR_{RR} = 1$). The simulation results confirm our theoretical derivation about raking under Y_Additive_Interaction as discussed in Section 3.6.1. For example, in the Y_Additive_Interaction, *n*=8,000, and S03 scenario, the average cross product ratios of the cell counts for the population, respondent sample, poststratification, and raking are 0.99, 1.02, 0.99, and 1.02 respectively. For poststratification, $CPR_w = CPR_{pop} = 0.99$. The poststratification process forces the weighted cell counts to strictly align with the population cell counts and thus reduces nonresponse bias most effectively. For raking, $CPR_w = CPR_r = 1.02 \neq CPR_{pop}$. The difference between raking

and poststratification is caused by the random errors in the two-phase sampling process for selecting respondents, which is reflected in the difference between CPR_{pop} and CPR_r. This explains why raking performs almost as well (but not quite as well) as poststratification.

In theory, the cross product ratios of the unweighted and weighted cell counts for a given response scenario are independent of the outcome variable model, and the data in Tables 3.3 and 3.4 confirms this. Raking achieves bias reduction by forcing the weights to conform to the marginal totals while maintaining the association in the cell counts of the respondent sample. In contrast, GREG_Main fits a linear regression model for the outcome variable by accounting for only the main effects and excluding the interaction term. When the outcome model contains an interaction term, as in Y_Additive_Interaction, GREG_Main generally performs worse than raking. Raking and GREG_Main both include only main effects in their calibration equations, but are associated with different distance functions or function forms. The comparison between raking and GREG_Main shows that the form of distance function or function form matters much in some situations.

3.7.2 Impact of Outcome Variable Model and Small Cell Counts on Empirical Relative Standard Error

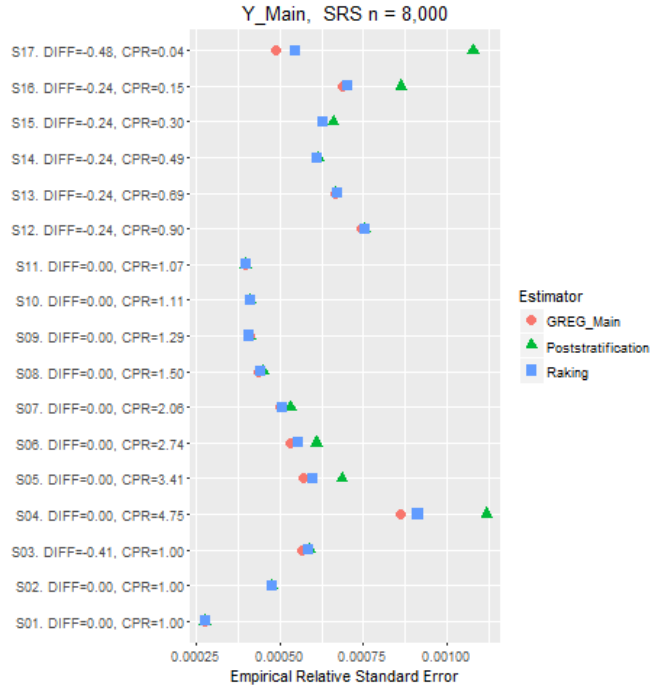
Figure 3.2 shows the empirical relative standard errors of the three estimators for various response scenarios with n=8,000. The top panel (a) is for the outcome variable model Y_Main and bottom panel (b) is for the outcome variable model Y_Additive_Interaction.

Before the simulation, we suspect that when the outcome variable model contains only main effects, using poststratification may increase the variance without reducing the bias further than raking and GREG_Main. This is confirmed by the simulation results shown in Table 3.3 and Figure 3.2(a). Under the Y_Main model, although the empirical relative standard errors for the three estimators in each response scenario are generally comparable, poststratification has a larger empirical relative standard error than raking and GREG_Main in the scenarios with small cell counts. For example, in the response scenario S17, the respondent sample size in one of the four cells is significantly smaller than those in the other three. The empirical relative standard errors for poststratification are significantly larger than those for raking and GREG_Main for all the SRS sample sizes (n=8,000, n=2,000, and n=200). This means the point estimate for poststratification is significantly less stable than those for raking and GREG_Main, which makes the coverage rate of the 95 percent confidence intervals significantly worse for poststratification (e.g., 88 percent for n=2,000 and 77 percent for n=200) . The results regarding Y_Main demonstrates that in the calibration process, including the interaction terms correlated only to the response propensity (but not to the outcome variable) can increase the variance without reducing the bias. One important implication is that when choosing the appropriate calibration estimator, survey practitioners should first focus on the outcome variable model, not the response model. We understand that in practice, survey practitioners not only need to create a single set of weights for a pool of outcome variables, but also often lack knowledge of the distribution of the outcome variables. This is probably why practitioners tend to model response and use that as the guidance

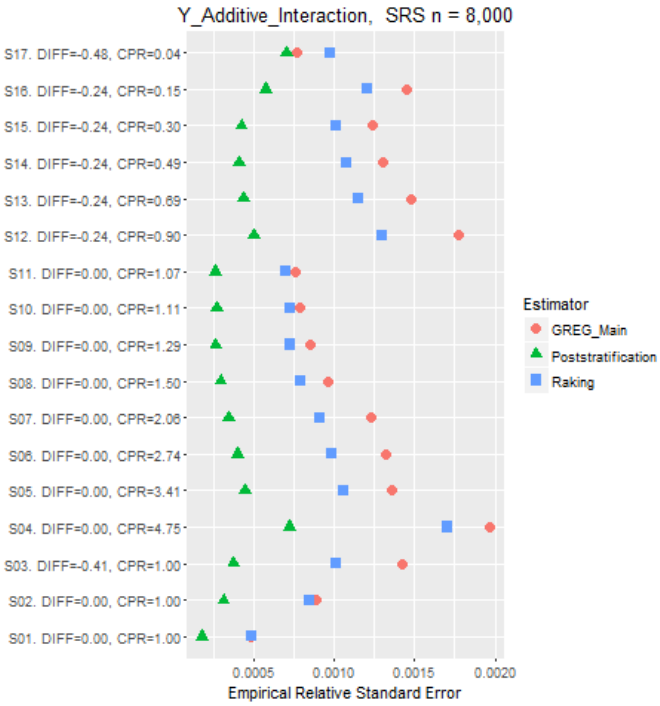
for choosing calibration covariates. However, clarifying the theoretical understanding is still critical for us to inform the selection of calibration variables in real-world surveys.

Table 3.4 shows that under the Y_Additive_Interaction outcome model, the empirical relative standard error for poststratification is noticeably smaller than those for raking and GREG_Main in all the response scenarios except S17 (where the small cell count makes poststratification unstable). This is true even for SRS n=200, under which we see some very small cell counts such as those in S04 and S16. Compared to raking and GREG_Main, the poststratification estimator is a better predictor of the outcome variable, and therefore is more stable as long as all the cell counts are large enough. At the same time, even for the response scenario S17, the empirical relative standard error for poststratification is no higher than those for raking and GREG_Main. Moreover, for the response scenarios S01, S02, and S03, although the three estimators perform similarly well in terms of bias reduction, poststratification outperforms the other two estimators in terms of variance reduction. All this shows that *when the outcome model has very high explanatory power and the interaction term in the model is substantively and statistically significant*, poststratification almost always outperforms raking and GREG_Main in terms of both bias reduction and variance reduction. Even in some situations with very small cell counts (e.g., the response scenario S17), using poststratification (when it is the most appropriate estimator based on the outcome model) does not necessarily lead to higher variance. In all the response scenarios including S17, the relative square root of MSE for poststratification is smaller than those for raking and GREG_Main.

Finally, although GREG_Main and raking are asymptotically equivalent when there is no nonresponse, Figure 3.2(b) illustrates that under Y_Additive_Interaction, raking has a consistently smaller empirical relative standard error than GREG_Main. In contrast, Figure 3.2(a) shows that the empirical relative standard errors for raking and GREG_Main under Y_Main are approximately equal. This means that when nonresponse exists, whether GREG_Main and raking become indistinguishable is affected by the underlying outcome variable model.



(a) Outcome Variable Model Y_Main



(b) Outcome Variable Model Y_Additive_Interaction

Figure 3.2 Empirical Relative Standard Errors for Poststratification, Raking, and GREG_Main for *n*=8,000 and Various Response Scenarios

3.7.3 Impact of Overall and Cell Sample Sizes on Bias Ratio and the Coverage Rate of 95 Percent Confidence Intervals

As discussed in Section 3.7.1, despite the differences between the three calibration estimators under the Y_Additive_Interaction model, these estimators can all achieve effective bias reduction through accounting for at least the main effects of the auxiliary variables. In Table 3.4, the relative biases are no more than approximately 4 percent even for the estimators that fail to appropriately account for the interaction term in the outcome variable model. Then the question is: Should survey practitioners be concerned about such small relative biases in raking and GREG_Main?

To answer this question, we should note that the coverage rate of the 95 percent confidence intervals can be poor even when the relative bias is small. In Table 3.4, for the SRS sample sizes n=8,000 and n=2,000, the coverage rates of the 95 percent confidence intervals for GREG_Main and raking are unacceptable in most of the response scenarios from S04 through S17. For example, the relative biases for raking and GREG_Main for n=8,000 and S10 are only approximately 1.0 percent and 2.3 percent respectively, but the coverage rates of the 95 percent confidence intervals are as low as 77 percent and 16 percent respectively.

Moreover, the coverage rate of the 95 percent confidence intervals becomes worse as the SRS sample size increases from 2,000 to 8,000. For example, in the response scenario S13, the coverage rates of the 95 percent confidence intervals for poststratification, GREG_Main, and raking are 95 percent, 86 percent, and 85 percent respectively for

n=2,000, but drop to 93 percent, 39 percent, and 27 percent respectively for *n*=8,000. When the SRS sample size drops to 200, we can see two opposite patterns depending on whether the calibration weighting involves some small subgroup counts. For poststratification, the coverage rates of the 95 percent confidence intervals for *n*=200 are generally worse than those for *n*=8,000 and *n*=2,000 because the estimator becomes unstable under *n*=200 due to some very small cell counts. For raking and GREG_Main, however, the coverage rates of the 95 percent confidence intervals for *n*=200 are noticeably better than those for *n*=8,000 and *n*=2,000 due to the larger variances under *n*=200 (which make the confidence intervals wider).

Figures 3.3 and 3.4 present the absolute values of bias ratios and the coverage rates of the 95 percent confidence intervals of the three estimators for the various response scenarios under the Y_Additive_Interaction model. The three panels (a), (b), and (c) correspond to *n*=8,000, *n*=2,000, and *n*=200, respectively. For raking and GREG_Main, the bias ratios increase as the SRS sample size increases (from 200 to 2,000, and then to 8,000), and this generally hurt the coverage rates of the 95 percent confidence intervals. On the other hand, increasing the overall sample size from 200 to 2,000 helps eliminate the small cell problem for poststratification to some extent. Given that the poststratification estimator is unbiased, eliminating the small cells during calibration weighting makes the estimate more stable, so the confidence interval is more likely to be centered at the population truth. This is why for the poststratification estimator, the coverage rate of the 95 percent confidence intervals for *n*=2,000 is better that for *n*=200. For example, for the response

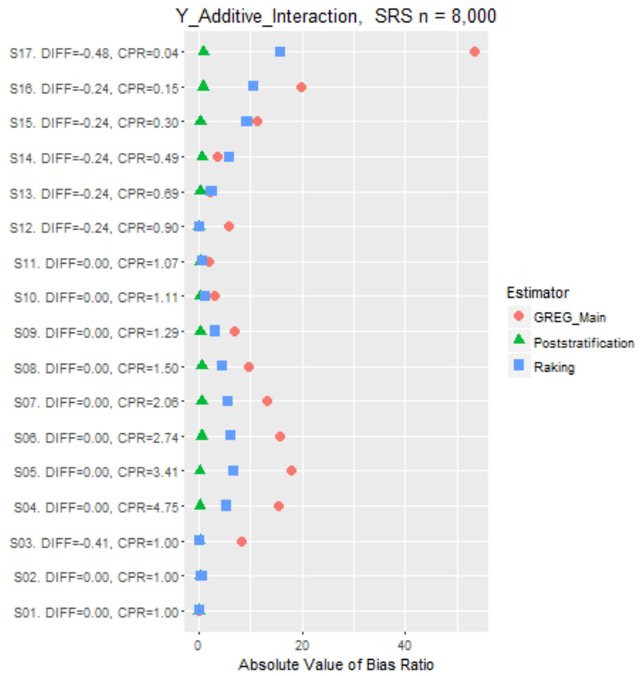
scenario S17, the coverage rates of the 95 percent confidence intervals for poststratification are 78 percent for n=200 and 89 percent for n=2,000, respectively.

How can the coverage rate of the 95 percent confidence intervals be unacceptably poor even when the relative bias is very low? Why may increasing sample size hurt the coverage rate of the 95 percent confidence intervals? The answer lies in the asymptotic property of the bias ratio. We can re-write the *t*-statistic into the summation of two terms

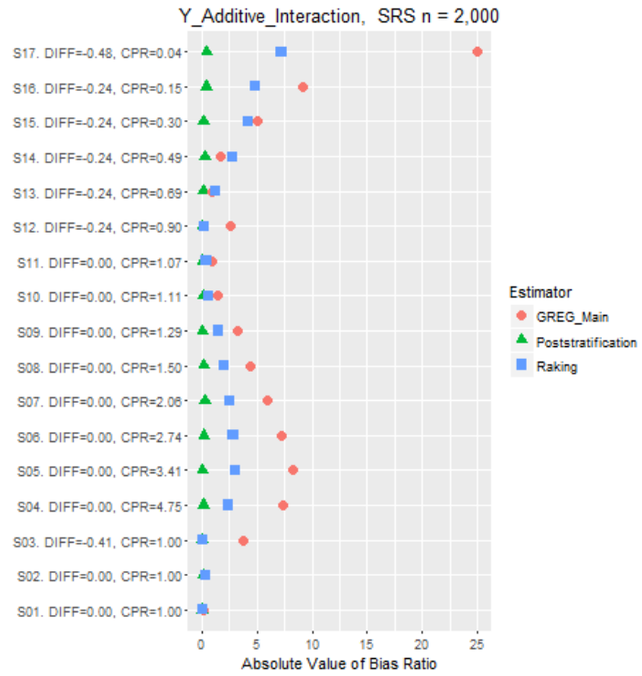
$$t\text{-}statistic = \frac{\hat{t}_{yw_s} - t_y}{\sqrt{\mathrm{Var}(\hat{t}_{yw})}} = \frac{\hat{t}_{yw_s} - \mathrm{E_p E_M}(\hat{t}_{yw})}{\sqrt{\mathrm{Var}(\hat{t}_{yw})}} + \frac{\mathrm{E_p E_M}(\hat{t}_{yw}) - t_y}{\sqrt{\mathrm{Var}(\hat{t}_{yw})}} \tag{3.50}$$

The first term on the right-hand side of (3.50) is asymptotically $N(0, 1)$ under standard conditions. The second term is the standardized bias or bias ratio. As the sample size increases, the denominator of the second term decreases. However, if the calibration estimator is model-biased as in the situation of GREG_Main and raking under Y_Additive_Interaction, the numerator in the second term of (3.50) stays constant instead of decreasing with the increase of sample size. As a result, a larger sample size makes the bias ratio higher, and thus leads to the *t-statistic* not being centered at zero. This hurts the coverage rate of the 95 percent confidence intervals. An important message to survey practitioners is that unless the calibration weighting process can be improved by incorporating more meaningful covariates, purely increasing the sample size does not help improve the performance of a calibration estimator that is model-biased. When the sample size is large, the coverage rate of the 95 confidence intervals for a biased

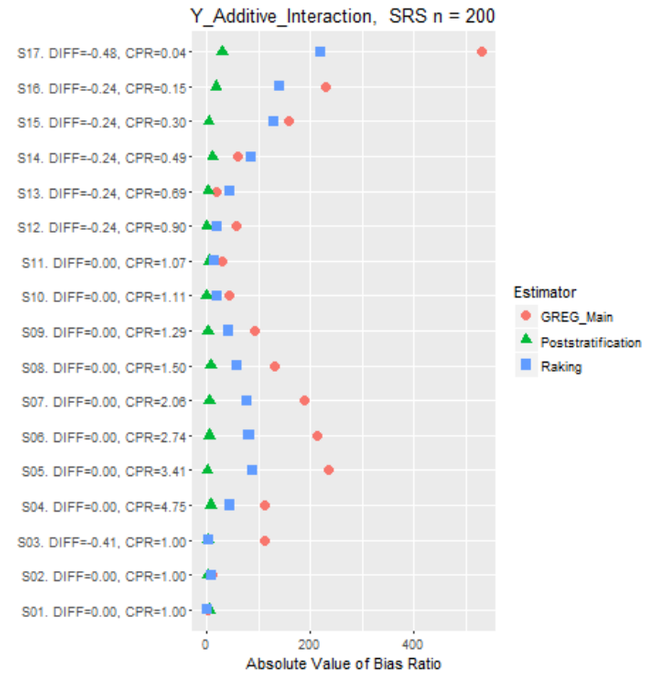
estimator can be unacceptably poor even when the relative bias is very small. However, this does not mean that we are advocating for small sample sizes in surveys. In practice, a bigger sample size allows for richer calibration models (e.g., with more variables, more categories for categorical variables, and more interaction terms), and thus more potential to reduce bias in practice.



(a) SRS Sample Size *n*=8,000

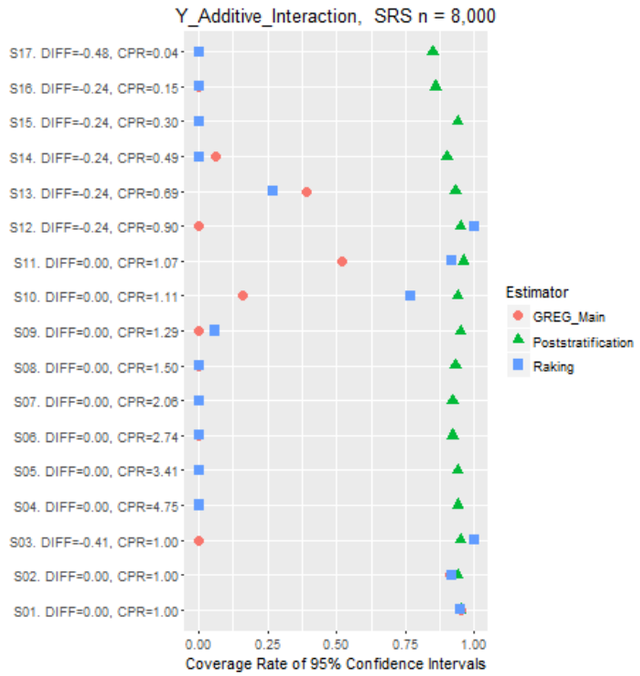


(b) SRS Sample Size *n*=2,000

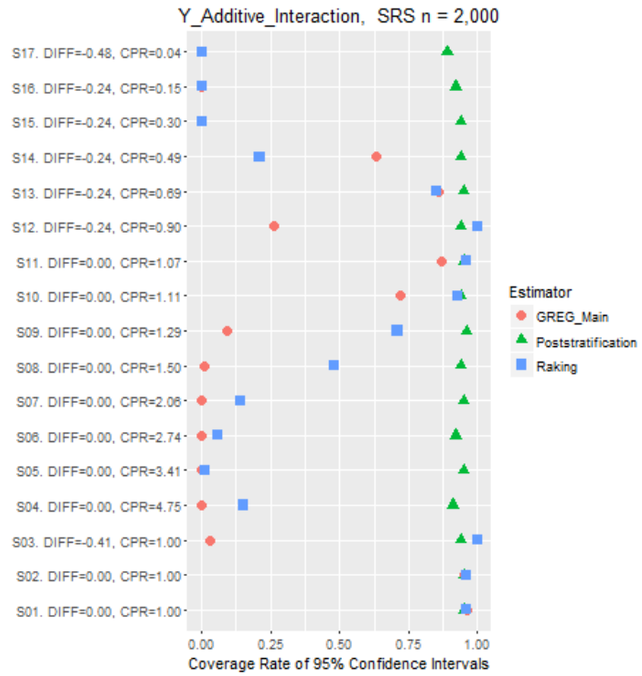


(c) SRS Sample Size *n*=200

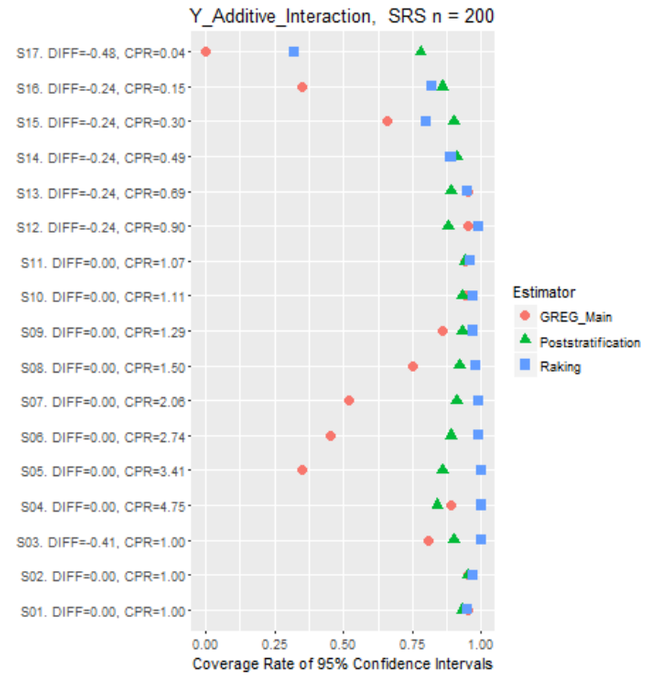
Figure 3.3 Bias Ratios for Poststratification, Raking, and GREG_Main under Y_Additive_Interaction and Various Response Scenarios



(a) SRS Sample Size *n*=8,000



(b) SRS Sample Size *n*=2,000



(c) SRS Sample Size *n*=200

Figure 3.4 Coverage Rates of 95 Percent Confidence Intervals for Poststratification, Raking, and GREG_Main under Y_Additive_Interaction and Various Response Scenarios

Table 3.3 Properties of Poststratification, Raking, and GREG_Main under Y_Main Outcome Variable Model

Outcome Variable Model: Y_Main	True Total $t_y \times 10^{-7}$	Relative Bias $RelBias(\hat{t}_{yw}) \times 10^5$				Empirical Relative Standard Error $EmpRelSE(\hat{t}_{yw}) \times 10^4$				Relative Square Root of MSE $RelRMSE(\hat{t}_{yw}) \times 10^4$			Bias Ratio $BiasRatio(\hat{t}_{yw}) \times 10^2$			Coverage Rate of 95% Confidence Intervals			Average Respondent Sample Sizes by Cell				Average Cross Product Ratios of Unweighted or Weighted Cell Counts				
		No calibration	Poststratification	GREG_Main	Raking	No calibration	Poststratification	GREG_Main	Raking	Poststratification	GREG_Main	Raking	Poststratification	GREG_Main	Raking	Poststratification	GREG_Main	Raking	n_{r11}	n_{r11}	n_{r11}	n_{r11}	Population	Respondent Sample	Poststratification	GREG_Main	Raking
SRS Sample Size n =8,000																											
S01. $DIFF_{RR}$=0.00, CPR_{RR}=1.00	4.28	5.6	1.1	1.1	1.1	26.2	2.7	2.7	2.7	2.2	2.2	2.2	3.9	3.9	3.9	0.95	0.95	0.95	2,018	2,006	2,008	1,968	0.99	0.99	0.99	0.99	0.99
S02. $DIFF_{RR}$=0.00, CPR_{RR}=1.00	4.28	-64746.0	5.1	5.3	5.4	48.5	4.8	4.8	4.8	3.8	3.8	3.8	10.9	11.4	11.5	0.95	0.95	0.95	903	875	590	589	0.99	1.03	0.99	1.03	1.03
S03. $DIFF_{RR}$=-0.41, CPR_{RR}=1.00	4.28	-56529.3	-8.2	-0.4	-7.9	53.7	5.9	5.7	5.9	4.7	4.5	4.7	-13.9	-0.7	-13.4	0.95	0.95	0.95	818	196	1,909	469	0.99	1.03	0.99	2.84	1.03
S04. $DIFF_{RR}$=0.00, CPR_{RR}=4.75	4.28	-74513.6	15.6	7.7	11.0	46.6	11.2	8.6	9.1	9.0	7.0	7.4	15.7	9.3	13.1	0.93	0.94	0.94	244	940	40	741	0.99	4.87	0.99	23.78	4.87
S05. $DIFF_{RR}$=0.00, CPR_{RR}=3.41	4.28	-48537.4	-11.8	-28.6	-20.9	55.1	6.9	5.7	6.0	5.5	5.1	5.0	-17.2	-50.6	-35.2	0.94	0.92	0.93	519	1,894	123	1,453	0.99	3.26	0.99	11.12	3.26
S06. $DIFF_{RR}$=0.00, CPR_{RR}=2.74	4.28	-48673.9	-26.0	-28.0	-27.0	56.7	6.1	5.3	5.5	5.3	4.8	4.9	-42.6	-52.0	-48.5	0.92	0.92	0.92	559	1,857	163	1,416	0.99	2.62	0.99	7.18	2.62
S07. $DIFF_{RR}$=0.00, CPR_{RR}=2.06	4.28	-49079.2	-29.6	-24.1	-26.8	53.6	5.3	5.0	5.1	4.9	4.5	4.6	-56.3	-48.7	-53.3	0.91	0.93	0.92	639	1,774	240	1,345	0.99	2.03	0.99	4.21	2.03
S08. $DIFF_{RR}$=0.00, CPR_{RR}=1.50	4.28	-49882.8	-21.1	-18.8	-19.9	53.5	4.5	4.3	4.4	4.0	3.8	3.9	-46.6	-42.4	-44.8	0.93	0.94	0.94	801	1,624	393	1,190	0.99	1.50	0.99	2.29	1.50
S09. $DIFF_{RR}$=0.00, CPR_{RR}=1.29	4.28	-50524.2	-5.9	-5.7	-5.8	51.2	4.1	4.1	4.1	3.3	3.3	3.3	-13.9	-13.6	-13.7	0.96	0.95	0.96	921	1,508	513	1,070	0.99	1.28	0.99	1.67	1.28
S10. $DIFF_{RR}$=0.00, CPR_{RR}=1.11	4.28	-51530.5	-3.6	-2.9	-3.3	50.2	4.1	4.1	4.1	3.3	3.3	3.3	-8.9	-7.2	-8.1	0.95	0.95	0.95	1,096	1,353	673	896	0.99	1.08	0.99	1.21	1.08
S11. $DIFF_{RR}$=0.00, CPR_{RR}=1.07	4.28	-51667.1	-6.5	-6.0	-6.3	48.4	4.0	4.0	4.0	3.2	3.2	3.2	-16.1	-14.8	-15.6	0.95	0.96	0.95	1,138	1,316	713	860	0.99	1.04	0.99	1.13	1.04
S12. $DIFF_{RR}$=-0.24, CPR_{RR}=0.90	4.28	-74290.7	0.4	-5.4	0.6	46.6	7.6	7.4	7.6	6.1	6.0	6.1	0.4	-7.5	0.7	0.95	0.95	0.95	463	128	1,114	292	0.99	0.95	0.99	2.46	0.95
S13. $DIFF_{RR}$=-0.24, CPR_{RR}=0.69	4.28	-73816.2	-15.3	-16.1	-14.6	47.0	6.7	6.7	6.7	5.3	5.4	5.4	-22.8	-24.2	-21.6	0.94	0.93	0.94	402	194	1,053	357	0.99	0.71	0.99	1.35	0.71
S14. $DIFF_{RR}$=-0.24, CPR_{RR}=0.49	4.28	-73532.9	-37.6	-39.5	-40.9	48.0	6.1	6.1	6.1	5.8	5.8	5.9	-59.7	-63.0	-65.1	0.92	0.90	0.90	301	288	936	458	0.99	0.51	0.99	0.68	0.51
S15. $DIFF_{RR}$=-0.24, CPR_{RR}=0.30	4.28	-72932.5	-16.4	-22.6	-21.8	47.8	6.6	6.3	6.3	5.4	5.3	5.3	-24.9	-36.3	-34.9	0.94	0.92	0.93	188	402	816	572	0.99	0.33	0.99	0.27	0.33
S16. $DIFF_{RR}$=-0.24, CPR_{RR}=0.15	4.28	-72729.5	-68.8	-54.5	-58.5	49.6	8.6	6.9	7.0	9.0	7.1	7.4	-81.5	-82.2	-85.8	0.87	0.87	0.86	80	496	714	660	0.99	0.15	0.99	0.06	0.15
S17. $DIFF_{RR}$=-0.48, CPR_{RR}=0.04	4.28	-43222.1	-92.3	0.8	-21.8	56.4	10.8	4.9	5.5	11.6	3.9	4.7	-89.2	1.7	-40.1	0.84	0.95	0.93	41	1,190	1,336	1,455	0.99	0.04	0.99	0.00	0.04
SRS Sample Size n =2,000																											
S01. $DIFF_{RR}$=0.00, CPR_{RR}=1.00	4.28	0.0	-0.9	-0.9	-0.9	58.6	6.0	6.0	6.0	4.8	4.8	4.8	-1.5	-1.5	-1.5	0.95	0.95	0.95	504	502	502	492	0.99	0.99	0.99	0.99	0.99
S02. $DIFF_{RR}$=0.00, CPR_{RR}=1.00	4.28	-64709.5	10.9	11.1	11.1	102.5	9.7	9.7	9.7	7.8	7.8	7.8	10.7	10.8	10.8	0.96	0.96	0.96	225	220	147	147	0.99	1.03	0.99	1.04	1.03
S03. $DIFF_{RR}$=-0.41, CPR_{RR}=1.00	4.28	-56528.1	-10.1	-1.5	-9.5	112.3	12.7	12.1	12.6	10.3	9.8	10.2	-7.8	-1.1	-7.3	0.95	0.96	0.95	204	49	478	117	0.99	1.04	0.99	2.93	1.04
S04. $DIFF_{RR}$=0.00, CPR_{RR}=4.75	4.28	-74531.9	9.4	12.6	11.9	102.3	26.4	18.6	20.2	20.9	14.9	16.3	9.3	6.5	7.5	0.91	0.95	0.94	62	234	10	185	0.99	5.44	0.99	33.29	5.44
S05. $DIFF_{RR}$=0.00, CPR_{RR}=3.41	4.28	-48461.3	-7.0	-20.1	-13.9	117.3	15.6	12.4	13.3	12.4	10.2	10.6	-4.2	-16.5	-10.9	0.93	0.95	0.94	130	473	31	364	0.99	3.35	0.99	11.90	3.35
S06. $DIFF_{RR}$=0.00, CPR_{RR}=2.74	4.28	-48700.2	-28.7	-26.5	-27.7	121.0	13.5	11.7	12.2	10.8	9.5	9.8	-21.3	-22.5	-22.6	0.93	0.93	0.93	140	462	40	355	0.99	2.72	0.99	7.71	2.72
S07. $DIFF_{RR}$=0.00, CPR_{RR}=2.06	4.28	-49014.6	-27.9	-22.3	-25.0	118.2	11.4	10.8	10.9	9.4	8.8	9.0	-24.3	-20.6	-22.8	0.94	0.94	0.95	160	444	60	336	0.99	2.06	0.99	4.37	2.06
S08. $DIFF_{RR}$=0.00, CPR_{RR}=1.50	4.28	-49816.3	-24.6	-21.9	-23.1	121.7	9.8	9.5	9.5	8.0	7.7	7.9	-24.9	-22.7	-23.8	0.94	0.94	0.94	201	406	99	298	0.99	1.51	0.99	2.33	1.51
S09. $DIFF_{RR}$=0.00, CPR_{RR}=1.29	4.28	-50484.2	-6.2	-6.1	-6.2	115.9	9.5	9.4	9.4	7.6	7.5	7.5	-6.7	-6.6	-6.7	0.94	0.94	0.94	230	376	128	269	0.99	1.29	0.99	1.70	1.29
S10. $DIFF_{RR}$=0.00, CPR_{RR}=1.11	4.28	-51493.1	-1.4	-0.7	-1.1	108.6	8.9	8.9	8.9	7.1	7.1	7.1	-1.7	-0.9	-1.3	0.95	0.94	0.94	274	337	169	225	0.99	1.09	0.99	1.22	1.09
S11. $DIFF_{RR}$=0.00, CPR_{RR}=1.07	4.28	-51694.3	-3.6	-2.8	-3.3	111.2	8.7	8.7	8.7	6.9	6.9	6.9	-4.0	-3.1	-3.7	0.95	0.95	0.95	284	329	178	214	0.99	1.05	0.99	1.13	1.05
S12. $DIFF_{RR}$=-0.24, CPR_{RR}=0.90	4.28	-74290.5	-1.3	-2.5	-0.7	101.3	16.8	16.2	16.8	13.3	12.8	13.2	-0.9	-1.7	-0.6	0.95	0.95	0.95	116	32	279	73	0.99	0.98	0.99	2.63	0.98
S13. $DIFF_{RR}$=-0.24, CPR_{RR}=0.69	4.28	-73782.9	-18.6	-18.6	-17.8	103.5	14.7	14.7	14.8	11.9	11.9	11.9	-12.6	-12.7	-12.0	0.94	0.94	0.94	100	48	264	90	0.99	0.72	0.99	1.41	0.72
S14. $DIFF_{RR}$=-0.24, CPR_{RR}=0.49	4.28	-73554.5	-40.6	-42.4	-43.7	104.4	13.8	13.8	13.8	11.5	11.5	11.5	-29.6	-31.0	-31.9	0.95	0.94	0.94	75	72	234	114	0.99	0.52	0.99	0.69	0.52
S15. $DIFF_{RR}$=-0.24, CPR_{RR}=0.30	4.28	-72960.3	-18.9	-23.7	-23.0	110.1	14.8	13.8	13.9	12.0	11.3	11.3	-13.8	-17.9	-17.4	0.94	0.95	0.95	47	100	204	142	0.99	0.33	0.99	0.28	0.33
S16. $DIFF_{RR}$=-0.24, CPR_{RR}=0.15	4.28	-72620.7	-62.4	-50.3	-53.8	108.9	19.2	14.7	15.2	15.8	12.3	12.8	-35.3	-34.8	-36.4	0.91	0.93	0.93	20	125	179	166	0.99	0.15	0.99	0.06	0.15
S17. $DIFF_{RR}$=-0.48, CPR_{RR}=0.04	4.28	-43249.3	-91.0	8.9	-15.1	126.5	23.9	10.6	11.7	20.1	8.6	9.4	-44.1	8.3	-12.2	0.88	0.95	0.95	10	299	333	363	0.99	0.04	0.99	0.00	0.04

Table 3.3 Properties of Poststratification, Raking, and GREG_Main under Y_Main Outcome Variable Model (Continued)

Outcome Variable Model: Y_Main	True Total	Relative Bias				Empirical Relative Standard Error				Relative Square Root of MSE			Bias Ratio			Coverage Rate of 95% Confidence Intervals			Average Respondent Sample Sizes by Cell				Average Cross Product Ratios of Unweighted or Weighted Cell Counts				
	$t_y \times 10^{-7}$	$RelBias(\hat{t}_{yw}) \times 10^5$				$EmpRelSE(\hat{t}_{yw}) \times 10^4$				$RelRMSE(\hat{t}_{yw}) \times 10^4$			$BiasRatio(\hat{t}_{yw}) \times 10^2$						n_{r11}	n_{r11}	n_{r11}	n_{r11}					
		No calibration	Poststratification	GREG_Main	Raking	No calibration	Poststratification	GREG_Main	Raking	Poststratification	GREG_Main	Raking	Poststratification	GREG_Main	Raking	Poststratification	GREG_Main	Raking					Population	Respondent Sample	Poststratification	GREG_Main	Raking
SRS Sample Size n =200																											
S01. $DIFF_{RR}$=0.00, CPR_{RR}=1.00	4.28	-102.8	-1.1	-1.4	-1.4	191.9	19.6	19.6	19.6	15.6	15.6	15.6	-0.7	-0.9	-0.8	0.95	0.95	0.95	51	50	50	49	0.99	1.04	0.99	1.04	1.04
S02. $DIFF_{RR}$=0.00, CPR_{RR}=1.00	4.28	-64820.9	5.5	6.3	5.8	349.0	33.3	33.1	33.2	26.7	26.6	26.6	1.8	2.0	1.8	0.94	0.95	0.95	23	22	15	15	0.99	1.18	0.99	1.21	1.18
S03. $DIFF_{RR}$=-0.41, CPR_{RR}=1.00	4.28	-56834.2	8.9	12.2	6.0	361.8	43.3	40.5	42.2	34.8	32.9	33.9	2.2	3.5	2.1	0.90	0.93	0.93	20	5	47	12	0.99	1.21	0.99	4.23	1.21
S04. $DIFF_{RR}$=0.00, CPR_{RR}=4.75	4.28	-74070.9	-8.8	-9.7	-7.2	339.2	57.3	54.1	54.7	45.8	42.7	43.5	2.8	-1.1	1.1	0.85	0.92	0.91	6	23	2	18	0.99	2.15	0.99	5.07	2.15
S05. $DIFF_{RR}$=0.00, CPR_{RR}=3.41	4.28	-48436.5	-17.6	-34.4	-27.4	396.0	48.6	41.5	43.0	38.5	32.9	34.1	-4.0	-8.8	-7.0	0.86	0.92	0.92	13	47	4	36	0.99	3.25	0.99	12.15	3.25
S06. $DIFF_{RR}$=0.00, CPR_{RR}=2.74	4.28	-48672.5	-41.5	-37.6	-38.6	393.4	44.5	38.6	40.0	35.4	30.8	32.0	-10.5	-9.6	-9.2	0.88	0.93	0.93	14	46	4	35	0.99	3.01	0.99	10.40	3.01
S07. $DIFF_{RR}$=0.00, CPR_{RR}=2.06	4.28	-49001.5	-37.5	-33.7	-36.0	377.3	39.9	35.9	36.8	32.1	28.7	29.7	-12.7	-11.0	-12.0	0.90	0.94	0.93	16	45	6	33	0.99	2.34	0.99	6.32	2.34
S08. $DIFF_{RR}$=0.00, CPR_{RR}=1.50	4.28	-49877.0	-28.6	-26.8	-28.2	366.1	32.3	31.0	31.3	25.8	24.6	25.0	-9.0	-8.6	-9.0	0.94	0.94	0.94	20	41	10	30	0.99	1.75	0.99	3.11	1.75
S09. $DIFF_{RR}$=0.00, CPR_{RR}=1.29	4.28	-50689.5	-4.3	-7.3	-5.5	371.1	31.2	30.5	30.5	25.0	24.4	24.5	-1.7	-2.6	-2.0	0.93	0.94	0.94	23	38	13	27	0.99	1.44	0.99	2.15	1.44
S10. $DIFF_{RR}$=0.00, CPR_{RR}=1.11	4.28	-51615.6	-3.8	-2.5	-2.8	361.7	29.4	29.0	29.0	23.6	23.3	23.3	-1.6	-1.2	-1.3	0.94	0.95	0.95	28	34	17	22	0.99	1.17	0.99	1.34	1.17
S11. $DIFF_{RR}$=0.00, CPR_{RR}=1.07	4.28	-51573.7	-7.0	-9.1	-9.1	363.6	29.0	28.8	28.9	23.2	23.2	23.2	-2.2	-2.9	-2.9	0.95	0.95	0.95	29	33	18	22	0.99	1.17	0.99	1.30	1.17
S12. $DIFF_{RR}$=-0.24, CPR_{RR}=0.90	4.28	-73980.4	-1.8	-10.7	-0.2	326.4	57.0	54.7	56.6	44.9	43.1	44.7	0.1	-2.6	-1.7	0.87	0.90	0.90	12	4	28	7	0.99	1.00	0.99	2.75	1.00
S13. $DIFF_{RR}$=-0.24, CPR_{RR}=0.69	4.28	-73757.2	-30.2	-30.7	-30.3	330.4	49.8	48.4	49.3	39.7	38.8	39.3	-7.6	-7.2	-6.9	0.91	0.92	0.92	10	5	26	9	0.99	0.86	0.99	1.95	0.86
S14. $DIFF_{RR}$=-0.24, CPR_{RR}=0.49	4.28	-73372.5	-24.2	-20.7	-22.7	347.6	47.6	46.5	46.6	37.8	36.9	37.2	-6.4	-4.9	-5.1	0.92	0.93	0.93	7	7	23	12	0.99	0.62	0.99	0.96	0.62
S15. $DIFF_{RR}$=-0.24, CPR_{RR}=0.30	4.28	-72749.4	-31.7	-36.1	-35.0	326.4	47.6	43.8	44.0	38.1	35.1	35.2	-7.7	-7.9	-7.7	0.91	0.94	0.94	5	10	20	14	0.99	0.41	0.99	0.39	0.41
S16. $DIFF_{RR}$=-0.24, CPR_{RR}=0.15	4.28	-72578.8	-100.6	-67.3	-77.4	336.7	56.4	47.1	48.0	45.6	37.6	38.6	-24.1	-15.4	-17.7	0.83	0.92	0.92	3	12	18	16	0.99	0.25	0.99	0.14	0.25
S17. $DIFF_{RR}$=-0.48, CPR_{RR}=0.04	4.28	-43295.2	-119.1	-17.2	-46.6	394.9	53.0	34.7	36.8	43.5	27.5	29.7	-35.9	-5.0	-12.5	0.77	0.93	0.92	2	29	33	36	0.99	0.09	0.99	0.02	0.09

Table 3.4 Properties of Poststratification, Raking, and GREG_Main under Y_Additive_Interaction Outcome Variable Model

	True Total	Relative Bias				Empirical Relative Standard Error				Relative Square Root of MSE			Bias Ratio			Coverage Rate of 95% Confidence Intervals			Average Respondent Sample Sizes by Cell				Average Cross Product Ratios of Unweighted or Weighted Cell Counts				
	$t_y \times 10^{-7}$	$RelBias(\hat{t}_{yw}) \times 10^5$				$EmpRelSE(\hat{t}_{yw}) \times 10^4$				$RelRMSE(\hat{t}_{yw}) \times 10^4$			$BiasRatio(\hat{t}_{yw}) \times 10^2$						n_{r11}	n_{r11}	n_{r11}	n_{r11}					
		No calibration	Poststratification	GREG_Main	Raking	No calibration	Poststratification	GREG_Main	Raking	Poststratification	GREG_Main	Raking	Poststratification	GREG_Main	Raking	Poststratification	GREG_Main	Raking					Population	Respondent Sample	Poststratification	GREG_Main	Raking
SRS Sample Size n =8,000																											
S01. $DIFF_{RR}$=0.00, CPR_{RR}=1.00	6.56	6.4	0.5	2.9	2.9	39.3	1.8	4.9	4.9	1.4	4.0	4.0	2.6	5.8	5.8	0.95	0.95	0.95	2,018	2,003	2,009	1,969	0.99	0.99	0.99	0.99	0.99
S02. $DIFF_{RR}$=0.00, CPR_{RR}=1.00	6.56	-64934.7	3.9	44.5	45.8	55.1	3.1	8.9	8.5	2.5	7.9	7.6	12.8	51.3	52.8	0.94	0.91	0.92	902	878	590	587	0.99	1.02	0.99	1.02	1.02
S03. $DIFF_{RR}$=-0.41, CPR_{RR}=1.00	6.56	-61333.8	-5.1	1162.2	31.8	48.9	3.8	14.2	10.1	3.0	116.2	8.4	-13.3	812.1	16.9	0.95	0.00	1.00	816	197	1,909	469	0.99	1.02	0.99	2.82	1.02
S04. $DIFF_{RR}$=0.00, CPR_{RR}=4.75	6.56	-71372.7	8.5	2982.1	1712.3	55.2	7.3	19.7	17.0	5.8	298.2	171.2	13.2	1537.0	522.0	0.94	0.00	0.00	244	941	41	738	0.99	4.81	0.99	23.37	4.81
S05. $DIFF_{RR}$=0.00, CPR_{RR}=3.41	6.56	-42358.2	-6.8	2444.8	1314.2	64.7	4.5	13.6	10.6	3.6	244.5	131.4	-15.3	1786.9	668.7	0.94	0.00	0.00	519	1,892	123	1,456	0.99	3.28	0.99	11.17	3.28
S06. $DIFF_{RR}$=0.00, CPR_{RR}=2.74	6.56	-42899.6	-18.4	2074.0	1078.7	63.1	4.0	13.2	9.9	3.5	207.4	107.9	-46.3	1573.1	625.5	0.92	0.00	0.00	560	1,855	163	1,418	0.99	2.64	0.99	7.22	2.64
S07. $DIFF_{RR}$=0.00, CPR_{RR}=2.06	6.56	-44017.1	-18.7	1576.0	799.7	63.5	3.4	12.3	9.1	3.1	157.6	80.0	-54.5	1319.9	567.8	0.92	0.00	0.00	640	1,775	239	1,345	0.99	2.04	0.99	4.25	2.04
S08. $DIFF_{RR}$=0.00, CPR_{RR}=1.50	6.56	-46330.6	-13.5	936.0	462.2	58.7	3.0	9.6	7.9	2.6	93.6	46.2	-45.7	958.1	441.7	0.93	0.00	0.00	802	1,624	392	1,188	0.99	1.50	0.99	2.30	1.50
S09. $DIFF_{RR}$=0.00, CPR_{RR}=1.29	6.56	-48051.7	-4.7	593.7	288.3	57.3	2.7	8.5	7.3	2.2	59.4	28.8	-17.2	690.5	324.8	0.95	0.00	0.06	921	1,508	513	1,071	0.99	1.28	0.99	1.67	1.28
S10. $DIFF_{RR}$=0.00, CPR_{RR}=1.11	6.56	-50614.7	-3.4	228.9	101.6	54.8	2.7	7.8	7.3	2.2	22.9	10.7	-12.7	296.3	130.6	0.94	0.16	0.77	1,096	1,352	674	897	0.99	1.08	0.99	1.21	1.08
S11. $DIFF_{RR}$=0.00, CPR_{RR}=1.07	6.56	-51108.0	-5.0	143.7	56.6	57.4	2.6	7.6	7.0	2.2	14.6	7.4	-19.2	188.5	73.9	0.96	0.52	0.92	1,138	1,317	714	859	0.99	1.04	0.99	1.12	1.04
S12. $DIFF_{RR}$=-0.24, CPR_{RR}=0.90	6.56	-76999.7	-1.0	1013.4	-41.6	43.8	5.0	17.7	13.0	4.0	101.3	11.0	-1.8	567.7	-21.2	0.95	0.00	1.00	464	128	1,113	292	0.99	0.96	0.99	2.48	0.96
S13. $DIFF_{RR}$=-0.24, CPR_{RR}=0.69	6.56	-75908.1	-10.8	334.8	-394.3	45.0	4.4	14.8	11.5	3.6	33.6	39.4	-24.9	221.3	-243.3	0.93	0.39	0.27	401	194	1,053	357	0.99	0.71	0.99	1.35	0.71
S14. $DIFF_{RR}$=-0.24, CPR_{RR}=0.49	6.56	-74650.8	-26.6	-457.6	-770.9	48.0	4.1	13.0	10.8	3.9	45.8	77.1	-64.9	-355.2	-594.7	0.90	0.06	0.00	301	288	937	458	0.99	0.51	0.99	0.68	0.51
S15. $DIFF_{RR}$=-0.24, CPR_{RR}=0.30	6.56	-73004.1	-10.6	-1445.9	-1247.2	49.8	4.2	12.4	10.1	3.5	144.6	124.7	-24.6	-1115.4	-931.1	0.94	0.00	0.00	188	403	817	570	0.99	0.33	0.99	0.27	0.33
S16. $DIFF_{RR}$=-0.24, CPR_{RR}=0.15	6.56	-71838.2	-47.0	-2817.3	-2036.1	55.5	5.8	14.5	12.1	6.1	281.7	203.6	-85.1	-1990.6	-1056.1	0.86	0.00	0.00	81	497	714	660	0.99	0.15	0.99	0.06	0.15
S17. $DIFF_{RR}$=-0.48, CPR_{RR}=0.04	6.56	-40220.0	-59.3	-4058.7	-3089.0	66.2	7.0	7.7	9.7	7.4	405.9	308.9	-88.0	-5346.8	-1567.3	0.85	0.00	0.00	40	1,191	1,335	1,457	0.99	0.04	0.99	0.00	0.04
SRS Sample Size n =2,000																											
S01. $DIFF_{RR}$=0.00, CPR_{RR}=1.00	6.56	-36.5	-1.1	-7.6	-7.6	85.1	3.9	10.5	10.4	3.1	8.3	8.3	-2.6	-7.1	-7.1	0.95	0.96	0.96	504	503	502	491	0.99	0.98	0.99	0.98	0.98
S02. $DIFF_{RR}$=0.00, CPR_{RR}=1.00	6.56	-64879.4	6.1	51.9	54.1	114.9	6.7	19.0	18.0	5.4	15.7	15.0	9.2	27.5	28.6	0.95	0.95	0.96	226	219	148	147	0.99	1.04	0.99	1.04	1.04
S03. $DIFF_{RR}$=-0.41, CPR_{RR}=1.00	6.56	-61350.0	-2.1	1173.1	42.8	110.3	8.6	31.9	23.3	6.9	117.3	19.0	-2.5	374.6	8.3	0.94	0.03	1.00	205	49	477	117	0.99	1.04	0.99	2.95	1.04
S04. $DIFF_{RR}$=0.00, CPR_{RR}=4.75	6.56	-71308.3	14.0	3014.4	1759.1	121.7	16.1	43.4	38.0	12.7	301.4	175.9	13.7	733.3	242.6	0.91	0.00	0.15	61	235	10	185	0.99	5.40	0.99	32.81	5.40
S05. $DIFF_{RR}$=0.00, CPR_{RR}=3.41	6.56	-42333.9	-5.2	2451.4	1324.8	138.0	9.6	30.8	24.4	7.7	245.1	132.5	-4.9	826.7	307.0	0.95	0.00	0.01	130	473	31	364	0.99	3.39	0.99	12.16	3.39
S06. $DIFF_{RR}$=0.00, CPR_{RR}=2.74	6.56	-42936.4	-15.2	2059.4	1071.8	139.6	9.1	29.9	22.7	7.2	205.9	107.2	-17.3	717.9	283.9	0.92	0.00	0.06	139	464	41	354	0.99	2.66	0.99	7.49	2.66
S07. $DIFF_{RR}$=0.00, CPR_{RR}=2.06	6.56	-43992.0	-18.7	1560.0	789.6	134.7	7.2	26.8	19.9	5.9	156.0	79.0	-24.6	599.0	256.1	0.95	0.00	0.14	160	444	60	336	0.99	2.04	0.99	4.32	2.04
S08. $DIFF_{RR}$=0.00, CPR_{RR}=1.50	6.56	-46301.0	-11.7	938.1	463.6	134.7	6.6	21.6	17.5	5.3	93.8	46.4	-18.3	439.2	201.9	0.94	0.01	0.48	201	406	98	297	0.99	1.51	0.99	2.33	1.51
S09. $DIFF_{RR}$=0.00, CPR_{RR}=1.29	6.56	-48077.9	-3.0	599.7	294.6	131.3	5.8	18.4	15.9	4.6	60.0	29.8	-5.0	318.4	151.2	0.96	0.09	0.71	230	376	128	268	0.99	1.29	0.99	1.70	1.29
S10. $DIFF_{RR}$=0.00, CPR_{RR}=1.11	6.56	-50583.2	-4.4	235.2	105.6	128.3	6.0	17.1	15.9	4.7	25.0	15.6	-7.6	138.5	61.5	0.94	0.72	0.93	274	338	168	225	0.99	1.09	0.99	1.23	1.09
S11. $DIFF_{RR}$=0.00, CPR_{RR}=1.07	6.56	-51081.1	-3.2	151.0	65.5	117.1	5.7	16.0	14.9	4.5	18.2	13.2	-5.5	90.5	39.0	0.95	0.87	0.96	285	328	179	215	0.99	1.05	0.99	1.14	1.05
S12. $DIFF_{RR}$=-0.24, CPR_{RR}=0.90	6.56	-76925.1	-7.6	1018.6	-46.0	92.2	11.1	40.0	28.8	8.8	101.9	23.2	-6.4	259.0	-13.7	0.94	0.26	1.00	116	32	280	73	0.99	0.98	0.99	2.67	0.98
S13. $DIFF_{RR}$=-0.24, CPR_{RR}=0.69	6.56	-75919.3	-9.3	320.4	-405.8	98.9	9.4	31.8	24.7	7.6	36.8	41.9	-9.7	94.4	-117.2	0.95	0.86	0.85	101	49	264	89	0.99	0.71	0.99	1.37	0.71
S14. $DIFF_{RR}$=-0.24, CPR_{RR}=0.49	6.56	-74698.1	-22.9	-453.0	-765.6	104.8	9.2	28.4	23.7	7.5	46.6	76.6	-25.5	-162.8	-271.4	0.94	0.63	0.21	75	72	234	114	0.99	0.52	0.99	0.69	0.52
S15. $DIFF_{RR}$=-0.24, CPR_{RR}=0.30	6.56	-72982.5	-11.1	-1436.1	-1236.8	109.6	9.4	28.1	23.2	7.5	143.6	123.7	-12.2	-507.8	-422.4	0.94	0.00	0.00	47	101	204	143	0.99	0.33	0.99	0.28	0.33
S16. $DIFF_{RR}$=-0.24, CPR_{RR}=0.15	6.56	-71851.7	-46.7	-2821.4	-2041.0	111.4	12.4	31.9	26.1	10.6	282.1	204.1	-40.5	-916.7	-482.0	0.92	0.00	0.00	20	124	179	165	0.99	0.15	0.99	0.06	0.15
S17. $DIFF_{RR}$=-0.48, CPR_{RR}=0.04	6.56	-40277.1	-60.2	-4058.9	-3103.3	143.4	15.3	16.3	21.3	13.2	405.9	310.3	-44.9	-2504.9	-721.5	0.89	0.00	0.00	10	298	334	363	0.99	0.04	0.99	0.00	0.04

Table 3.4 Properties of Poststratification, Raking, and GREG_Main under Y_Additive_Interaction Outcome Variable Model (Continued)

	True Total	Relative Bias				Empirical Relative Standard Error				Relative Square Root of MSE			Bias Ratio			Coverage Rate of 95% Confidence Intervals			Average Respondent Sample Sizes by Cell				Average Cross Product Ratios of Unweighted or Weighted Cell Counts				
	$t_y \times 10^{-7}$	$RelBias(\hat{t}_{yw}) \times 10^5$				$EmpRelSE(\hat{t}_{yw}) \times 10^4$				$RelRMSE(\hat{t}_{yw}) \times 10^4$			$BiasRatio(\hat{t}_{yw}) \times 10^2$														
		No calibration	Poststratification	GREG_Main	Raking	No calibration	Poststratification	GREG_Main	Raking	Poststratification	GREG_Main	Raking	Poststratification	GREG_Main	Raking	Poststratification	GREG_Main	Raking	n_{r11}	n_{r11}	n_{r11}	n_{r11}	Population	Respondent Sample	Poststratification	GREG_Main	Raking
SRS Sample Size n=200																											
S01. $DIFF_{RR}$=0.00, CPR_{RR}=1.00	6.56	-13.4	5.9	6.1	6.3	281.4	13.3	35.0	34.8	10.6	27.9	27.8	4.7	1.7	1.8	0.93	0.95	0.95	50	50	50	49	0.99	1.03	0.99	1.03	1.03
S02. $DIFF_{RR}$=0.00, CPR_{RR}=1.00	6.56	-64875.6	4.7	59.2	56.3	361.4	21.2	63.9	59.4	17.0	50.7	47.0	2.3	9.4	9.0	0.95	0.95	0.97	22	22	15	15	0.99	1.17	0.99	1.20	1.17
S03. $DIFF_{RR}$=-0.41, CPR_{RR}=1.00	6.56	-61223.1	5.3	1148.4	42.3	344.1	28.5	94.5	71.5	22.4	125.9	57.4	2.1	111.4	-3.2	0.90	0.81	1.00	20	5	48	12	0.99	1.21	0.99	4.23	1.21
S04. $DIFF_{RR}$=0.00, CPR_{RR}=4.75	6.56	-71024.8	10.1	1476.0	713.2	373.4	38.4	89.6	74.1	30.6	155.8	88.1	7.0	112.4	45.0	0.84	0.89	1.00	6	23	2	18	0.99	2.16	0.99	5.04	2.16
S05. $DIFF_{RR}$=0.00, CPR_{RR}=3.41	6.56	-41934.6	-9.0	2191.6	1179.2	464.2	30.6	78.3	62.3	24.2	219.7	119.8	-1.4	234.3	88.0	0.86	0.35	1.00	13	47	4	37	0.99	3.36	0.99	12.27	3.36
S06. $DIFF_{RR}$=0.00, CPR_{RR}=2.74	6.56	-42853.4	-16.1	1935.2	1014.8	440.7	28.6	83.1	63.2	22.8	194.4	104.3	-5.8	212.4	81.6	0.89	0.45	0.99	14	46	4	36	0.99	2.91	0.99	9.68	2.91
S07. $DIFF_{RR}$=0.00, CPR_{RR}=2.06	6.56	-44064.6	-13.4	1584.5	833.6	440.5	25.4	83.3	62.8	20.3	161.0	89.2	-4.7	189.7	78.5	0.91	0.52	0.99	16	44	6	34	0.99	2.46	0.99	6.72	2.46
S08. $DIFF_{RR}$=0.00, CPR_{RR}=1.50	6.56	-46327.7	-15.5	935.2	462.0	424.5	22.1	71.3	57.5	17.6	98.5	59.4	-7.0	131.6	57.4	0.92	0.75	0.98	20	41	10	30	0.99	1.70	0.99	3.03	1.70
S09. $DIFF_{RR}$=0.00, CPR_{RR}=1.29	6.56	-48007.3	5.2	590.2	283.3	422.2	20.4	63.7	54.6	16.2	70.0	49.5	2.7	92.1	40.9	0.93	0.86	0.97	23	38	13	27	0.99	1.40	0.99	1.98	1.40
S10. $DIFF_{RR}$=0.00, CPR_{RR}=1.11	6.56	-50734.8	-1.2	252.9	123.3	409.6	19.2	56.6	51.3	15.3	48.9	42.0	0.0	43.0	20.2	0.93	0.94	0.97	27	33	17	23	0.99	1.20	0.99	1.39	1.20
S11. $DIFF_{RR}$=0.00, CPR_{RR}=1.07	6.56	-50924.1	9.7	171.3	81.8	404.5	19.1	55.0	51.1	15.3	45.5	41.2	5.1	30.1	14.1	0.94	0.94	0.96	28	33	18	22	0.99	1.15	0.99	1.27	1.15
S12. $DIFF_{RR}$=-0.24, CPR_{RR}=0.90	6.56	-76696.3	2.7	729.0	-197.3	303.5	35.8	104.2	81.9	28.2	106.5	67.2	0.6	56.4	-19.2	0.88	0.95	0.99	11	4	28	7	0.99	1.03	0.99	2.84	1.03
S13. $DIFF_{RR}$=-0.24, CPR_{RR}=0.69	6.56	-75889.2	-7.1	280.6	-429.4	321.9	33.6	103.7	82.2	26.8	86.0	74.2	-3.1	18.6	-43.8	0.89	0.95	0.95	10	5	26	9	0.99	0.83	0.99	1.90	0.83
S14. $DIFF_{RR}$=-0.24, CPR_{RR}=0.49	6.56	-74643.7	-29.9	-491.7	-779.8	326.6	31.5	99.8	81.3	25.2	90.9	93.8	-11.4	-58.5	-86.6	0.91	0.89	0.89	7	7	23	12	0.99	0.64	0.99	0.98	0.64
S15. $DIFF_{RR}$=-0.24, CPR_{RR}=0.30	6.56	-72937.3	-7.8	-1428.3	-1249.4	353.7	32.6	89.9	74.6	25.9	147.9	128.0	-3.7	-158.6	-130.6	0.90	0.66	0.80	5	10	21	14	0.99	0.38	0.99	0.36	0.38
S16. $DIFF_{RR}$=-0.24, CPR_{RR}=0.15	6.56	-71873.2	-46.3	-2235.1	-1642.0	356.8	36.7	72.9	65.6	29.3	223.9	165.0	-17.8	-229.6	-140.0	0.86	0.35	0.82	3	12	18	16	0.99	0.26	0.99	0.15	0.26
S17. $DIFF_{RR}$=-0.48, CPR_{RR}=0.04	6.56	-40296.6	-64.3	-3437.8	-2476.1	436.3	34.0	34.6	38.5	27.5	343.8	247.6	-29.4	-531.6	-220.0	0.78	0.00	0.32	2	30	33	36	0.99	0.09	0.99	0.02	0.09

3.8 Sensitivity Analysis

In Section 3.7, the parameters for the models Y_Main and Y_Additive_Interaction are chosen in a way to maximize the possibility of detecting the impact of the outcome model. The R-squared values for the outcome variable models shown in Table 3.1 are exceptionally high. In the real world, however, it is often unrealistic to expect the outcome variable models to have such strong predictive power.

In this section, we use the Y_Additive_Interaction scenario to conduct some sensitivity analysis by lowering the overall predictive power of the outcome variable model while keeping the response propensity models unchanged. Operationally, we achieve this by increasing the variance of the random error terms in outcome variable model. That is, we set $\varepsilon_{Yijk} \sim N(0, 250000)$ for the Y_Additive_Interaction model in (3.13) while keeping the values for μ_Y, $\boldsymbol{\alpha}_Y$, $\boldsymbol{\beta}_Y$, and $\boldsymbol{\gamma}_Y$ the same as specified in Section 3.4. As a result, the expected cell means for the outcome variable used in the sensitivity analysis remain the same as those shown in the Y_Additive_Interaction row of Table 3.1. The R-squared value for the overall model drops to 0.6348 and the *p*-value for the interaction term remains less than 0.0001 (meaning that the interaction term is still highly statistically significantly in an overall model with less explanatory power). The response scenarios for the sensitivity analysis are the same as those shown in Table 3.2. Using these model specifications, the simulation steps described in Section 3.5 are repeated. Then the properties of the three calibration estimators over repeated sampling are evaluated using

the criteria described in Section 3.6. The detailed results from the sensitivity analysis are presented in Table 3.5.

Figures 3.5 through 3.8 compare the results in Section 3.7 (shown in Table 3.4 for the Y_Additive_Interaction model) and those from the sensitivity analysis for the SRS sample size n=8,000. The four figures show the impact of the overall predictive power of the outcome variable model (measured by R-squared value) on the absolute value of relative bias, empirical relative standard error, absolute value of the bias ratio, and coverage rate of the 95 percent confidence intervals, respectively. For simplicity, we sometimes refer to the Y_Additive_Interaction model in Section 3.7 (with R^2= 0.9979) as the "high R-squared setup" and the Y_Additive_Interaction model for the sensitivity analysis (with R^2=0.6348) as the "medium R-squared setup". In Figures 3.5 through 3.8, the results for the medium R-squared setup are shown on the top panel and the results for the high R-squared setup are shown on the bottom panel. In general, these figures show that as the predictive power of the outcome model Y_Additive_Interaction decreases, the differences between poststratification and the other two estimators become smaller. We see several patterns from Table 3.4, Table 3.5, and Figures 3.5 through 3.8.

First, when the R-squared value for the Y_Additive_Interaction model decreases from 0.9979 to 0.6348, the impact on the empirical relative biases for raking and GREG_Main are generally negligible; any noticeable changes can be attributed to simulation variation (especially for the SRS sample size n=200). In contrast, the empirical relative biases for poststratification increase approximately 17 times for all the response scenarios. This is

because the variance of the random error term in the outcome variable model increases from 900 for the high R-squared setup to 250,000. The square root of the ratio between 250,000 and 900 is approximately 17. In general, the *empirical* relative bias for poststratification moves farther away from zero as the R-squared value for the Y_Additive_Interaction model decreases. We know that the poststratification estimator is model-unbiased, so any change in the empirical relative bias is actually a reflection of increased empirical variance.

Second, the empirical relative standard errors increase for all the three calibration estimators as the predictive power of the outcome variable model decreases, yet the biggest increases occur in poststratification. The differences in the empirical relative standard errors between the three estimators diminish almost completely from the high R-squared setup to the medium R-squared setup. Moreover, in some situations with small cell counts (e.g., the response scenarios S16 and S17), the empirical relative standard errors for poststratification are larger than those for raking and GREG_Main in the medium R-squared setup, which are not seen in the high R-squared setup. Recall that in the R-squared setup, poststratification almost always outperforms raking and GREG_Main in terms of both bias reduction and variance reduction. In the medium R-squared setup, the effect of further bias reduction may not outweigh the drawback of increased variance for poststratification when the calibration process involves some small cells. For example, in the response scenario S17, the relative square root of MSE for poststratification is larger than those for raking and GREG_Main.

Third, the bias ratio for poststratification remains unchanged for each response scenario as the R-squared value for the outcome variable model changes. This, again, is because in theory, poststratification is model-unbiased, so the observed empirical relative bias for poststratification actually reflects the magnitude of the empirical relative standard error. For raking and GREG_Main, the bias ratio decreases as the R-squared value for the outcome variable model decreases because the empirical relative standard error increases to a greater extent than the increase in empirical relative bias.

Finally, the coverage rate of the 95 percent confidence intervals is independent of the R-squared value of the outcome variable model for poststratification, but improves for raking and GREG_Main as the predictive power of the Y_Additive_Interaction model becomes weaker. The latter is largely due to increased standard errors of the estimators, which make the bias ratios smaller and confidence intervals wider.

The sensitivity analysis shows that the results in Section 3.7 may be highly sensitive to the model specifications for the outcome variable, or more specifically, the predictive power of the outcome variable model. Although the differences between poststratification, raking, and GREG_Main under the Y_Additive_Interaction model in Section 3.7 are very revealing, those conclusions are based on the assumptions that the outcome variable models have almost perfect predictive power (R-squared value being approximately 0.99). When the R-squared value for the Y_Additive_Interaction model drops to a reasonably high level (approximately 0.65), the poststratification estimator still outperforms raking and GREG_Main in terms of bias and MSE (except in the situations

with small cell counts), although the differences between poststratification and the other two estimators decrease significantly. In the real world, it is not rare to have an outcome model with the R-squared value being under 0.50. This is probably why survey practitioners often use raking or GREG_Main in place of poststratification, and the differences between poststratification and the other two estimators are not expected to be detrimental. Moreover, raking and GREG_Main may have smaller MSEs than poststratification when small cells are involved in the poststratification weighting.

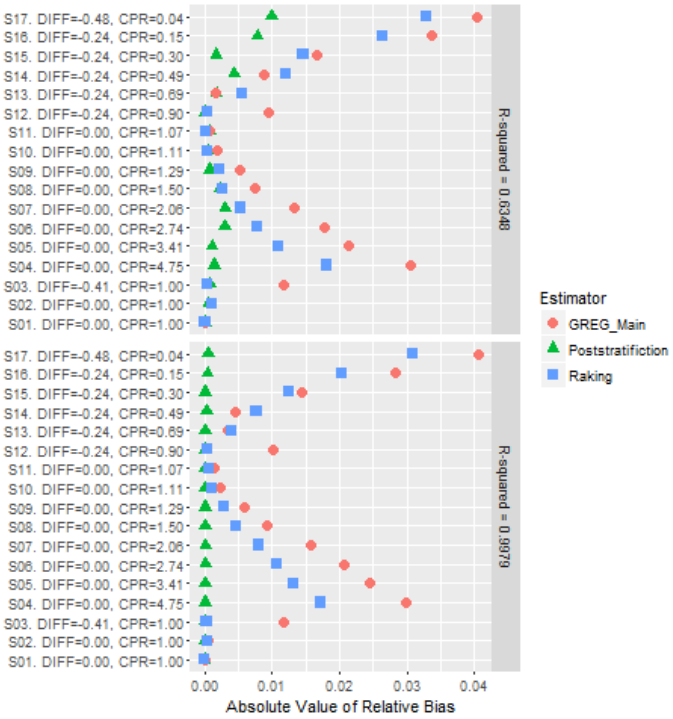


Figure 3.5 Impact of Predictive Power of Outcome Variable Model on Absolute Value of Relative Bias for Y_Additive_Interaction Model and *n*=8,000

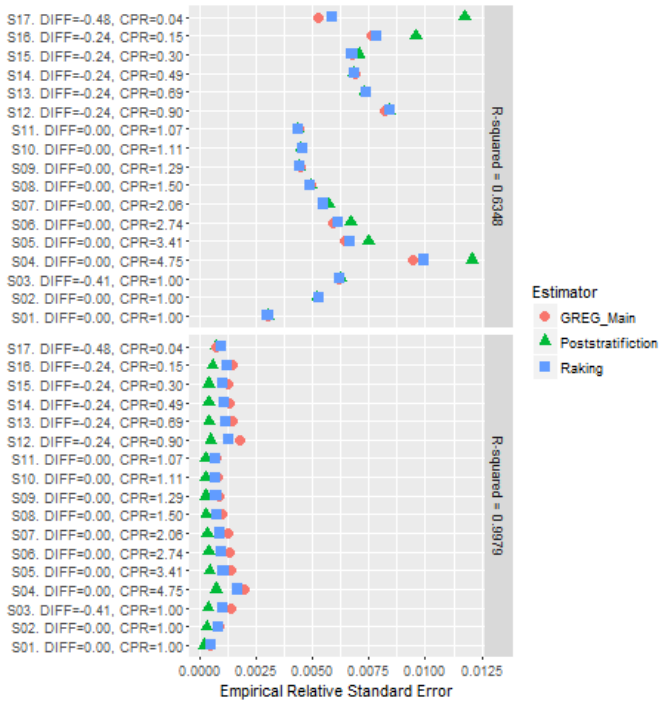


Figure 3.6 Impact of Predictive Power of Outcome Variable Model on Empirical Relative Standard Error for Y_Additive_Interaction Model and *n*=8,000

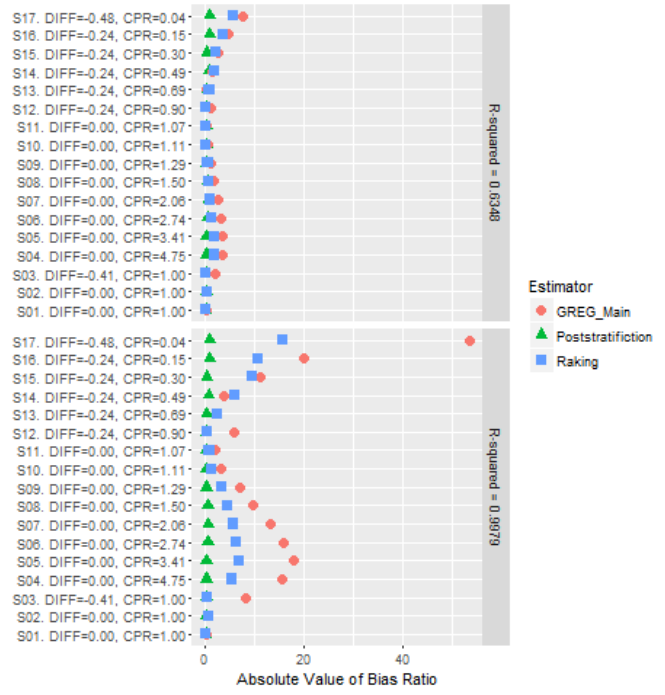


Figure 3.7 Impact of Predictive Power of Outcome Variable Model on Absolute Value of Bias Ratio for Y_Additive_Interaction Model and *n*=8,000

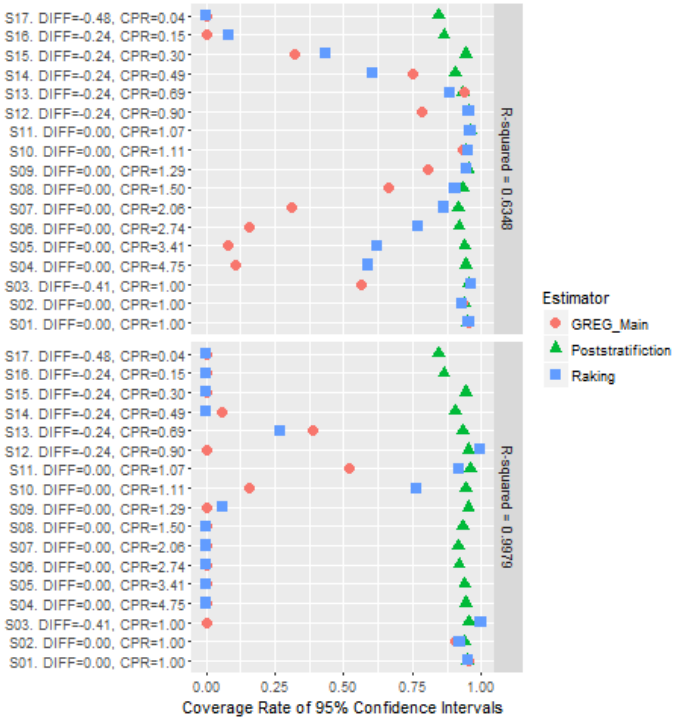


Figure 3.8 Impact of Predictive Power of Outcome Variable Model on Coverage Rate of 95 Percent Confidence Intervals for Y_Additive_Interaction Model and *n*=8,000

Table 3.5 Properties of Poststratification, Raking, and GREG_Main under the Y_Additive_Interaction Model for Sensitivity Analysis

	True Total	Relative Bias				Empirical Relative Standard Error				Relative Square Root of MSE			Bias Ratio			Coverage Rate of 95% Confidence Intervals			Average Respondent Sample Sizes by Cell				Average Cross Product Ratios of Unweighted or Weighted Cell Counts				
	$t_y \times 10^{-7}$	$RelBias(\hat{t}_{yw}) \times 10^5$				$EmpRelSE(\hat{t}_{yw}) \times 10^4$				$RelRMSE(\hat{t}_{yw}) \times 10^4$			$BiasRatio(\hat{t}_{yw}) \times 10^2$						n_{r11}	n_{r11}	n_{r11}	n_{r11}					
		No calibration	Poststratification	GREG_Main	Raking	No calibration	Poststratification	GREG_Main	Raking	Poststratification	GREG_Main	Raking	Poststratification	GREG_Main	Raking	Poststratification	GREG_Main	Raking					Population	Respondent Sample	Poststratification	GREG_Main	Raking
SRS Sample Size n =8,000																											
S01. $DIFF_{RR}$=0.00, CPR_{RR}=1.00	6.57	14.0	8.0	10.4	10.4	48.7	30.2	30.3	30.3	24.0	24.2	24.2	2.6	3.4	3.4	0.95	0.96	0.96	2,018	2,003	2,009	1,969	0.99	0.99	0.99	0.99	0.99
S02. $DIFF_{RR}$=0.00, CPR_{RR}=1.00	6.57	-64913.2	65.1	107.5	108.9	58.4	52.2	52.8	52.8	41.4	42.5	42.4	12.8	20.9	21.1	0.94	0.93	0.93	902	878	590	587	0.99	1.02	0.99	1.02	1.02
S03. $DIFF_{RR}$=-0.41, CPR_{RR}=1.00	6.57	-61312.0	-84.8	1162.5	-44.8	52.3	62.5	61.6	62.1	50.1	117.9	49.4	-13.3	182.3	-6.7	0.95	0.56	0.96	816	197	1,909	469	0.99	1.02	0.99	2.82	1.02
S04. $DIFF_{RR}$=0.00, CPR_{RR}=4.75	6.57	-71344.2	141.6	3046.7	1806.0	57.2	120.7	94.7	99.2	96.5	304.7	183.2	13.2	325.2	178.0	0.94	0.11	0.59	244	941	41	738	0.99	4.81	0.99	23.37	4.81
S05. $DIFF_{RR}$=0.00, CPR_{RR}=3.41	6.57	-42445.9	-113.5	2142.6	1102.1	68.6	74.6	64.2	66.5	60.6	214.3	112.5	-15.3	342.0	166.0	0.94	0.08	0.62	519	1,892	123	1,456	0.99	3.28	0.99	11.17	3.28
S06. $DIFF_{RR}$=0.00, CPR_{RR}=2.74	6.57	-42981.3	-306.7	1772.8	783.7	65.9	66.8	59.3	61.1	58.4	177.3	84.4	-46.3	296.7	125.5	0.92	0.16	0.77	560	1,855	163	1,418	0.99	2.64	0.99	7.22	2.64
S07. $DIFF_{RR}$=0.00, CPR_{RR}=2.06	6.57	-44070.4	-311.8	1339.1	535.4	66.9	57.3	54.5	54.8	51.7	134.1	63.4	-54.5	243.0	94.9	0.92	0.31	0.86	640	1,775	239	1,345	0.99	2.04	0.99	4.25	2.04
S08. $DIFF_{RR}$=0.00, CPR_{RR}=1.50	6.57	-46394.2	-223.9	752.5	265.2	62.2	49.3	49.1	49.2	43.7	77.8	44.6	-45.7	153.5	53.8	0.93	0.66	0.91	802	1,624	392	1,188	0.99	1.50	0.99	2.30	1.50
S09. $DIFF_{RR}$=0.00, CPR_{RR}=1.29	6.57	-48076.7	-79.0	517.9	213.3	60.3	44.2	44.7	44.6	35.8	57.7	39.9	-17.2	111.6	45.8	0.95	0.81	0.95	921	1,508	513	1,071	0.99	1.28	0.99	1.67	1.28
S10. $DIFF_{RR}$=0.00, CPR_{RR}=1.11	6.57	-50622.4	-55.9	183.9	52.5	58.2	44.9	45.4	45.4	36.6	39.2	36.8	-12.7	41.2	11.8	0.94	0.93	0.95	1,096	1,352	674	897	0.99	1.08	0.99	1.21	1.08
S11. $DIFF_{RR}$=0.00, CPR_{RR}=1.07	6.57	-51133.6	-83.9	71.5	-19.6	60.9	43.6	44.0	43.9	35.8	36.1	35.6	-19.2	16.1	-4.4	0.96	0.96	0.96	1,138	1,317	714	859	0.99	1.04	0.99	1.12	1.04
S12. $DIFF_{RR}$=-0.24, CPR_{RR}=0.90	6.57	-76975.9	-16.5	938.8	-53.7	46.4	83.9	82.3	84.3	66.9	103.4	67.5	-1.8	115.4	-6.0	0.95	0.78	0.96	464	128	1,113	292	0.99	0.96	0.99	2.48	0.96
S13. $DIFF_{RR}$=-0.24, CPR_{RR}=0.69	6.57	-75900.0	-180.2	156.7	-555.7	47.2	73.0	73.4	73.9	59.8	59.3	75.3	-24.9	21.2	-74.4	0.93	0.94	0.88	401	194	1,053	357	0.99	0.71	0.99	1.35	0.71
S14. $DIFF_{RR}$=-0.24, CPR_{RR}=0.49	6.57	-74742.4	-442.6	-890.4	-1215.1	49.7	68.3	69.1	68.8	64.9	94.9	123.2	-64.9	-128.9	-175.4	0.90	0.75	0.61	301	288	937	458	0.99	0.51	0.99	0.68	0.51
S15. $DIFF_{RR}$=-0.24, CPR_{RR}=0.30	6.57	-73035.1	-176.3	-1669.5	-1463.0	51.6	70.7	67.9	67.6	58.2	167.0	146.6	-24.6	-242.1	-211.5	0.94	0.32	0.44	188	403	817	570	0.99	0.33	0.99	0.27	0.33
S16. $DIFF_{RR}$=-0.24, CPR_{RR}=0.15	6.57	-71912.1	-783.1	-3358.0	-2632.3	57.8	95.9	75.9	78.2	101.5	335.8	263.2	-85.1	-460.0	-346.4	0.86	0.00	0.08	81	497	714	660	0.99	0.15	0.99	0.06	0.15
S17. $DIFF_{RR}$=-0.48, CPR_{RR}=0.04	6.57	-40225.6	-987.8	-4029.6	-3291.9	69.8	117.2	52.6	59.0	123.9	403.0	329.2	-88.0	-761.7	-549.5	0.85	0.00	0.00	40	1,191	1,335	1,457	0.99	0.04	0.99	0.00	0.04
SRS Sample Size n =2,000																											
S01. $DIFF_{RR}$=0.00, CPR_{RR}=1.00	6.57	-52.9	-17.5	-23.9	-23.9	108.5	65.4	65.8	65.8	52.1	52.4	52.4	-2.6	-3.6	-3.5	0.95	0.95	0.95	504	503	502	491	0.99	0.98	0.99	0.98	0.98
S02. $DIFF_{RR}$=0.00, CPR_{RR}=1.00	6.57	-64845.5	101.5	150.2	152.4	121.8	112.0	112.8	112.7	89.7	91.1	91.0	9.2	13.4	13.6	0.95	0.95	0.95	226	219	148	147	0.99	1.04	0.99	1.04	1.04
S03. $DIFF_{RR}$=-0.41, CPR_{RR}=1.00	6.57	-61321.9	-35.5	1226.8	13.1	118.7	143.3	139.6	143.9	114.8	152.8	114.7	-2.5	88.4	1.3	0.94	0.87	0.95	205	49	477	117	0.99	1.04	0.99	2.95	1.04
S04. $DIFF_{RR}$=0.00, CPR_{RR}=4.75	6.57	-71255.5	233.3	3172.3	1944.1	124.6	267.6	205.1	215.7	210.8	328.5	238.4	13.7	157.1	92.3	0.91	0.64	0.86	61	235	10	185	0.99	5.40	0.99	32.81	5.40
S05. $DIFF_{RR}$=0.00, CPR_{RR}=3.41	6.57	-42416.9	-86.6	2160.9	1128.3	148.4	160.4	135.4	140.4	128.6	222.2	145.1	-4.9	158.5	79.0	0.95	0.65	0.89	130	473	31	364	0.99	3.39	0.99	12.16	3.39
S06. $DIFF_{RR}$=0.00, CPR_{RR}=2.74	6.57	-43002.7	-253.5	1760.6	801.7	146.8	151.1	135.6	139.8	120.0	189.3	129.1	-17.3	135.8	60.1	0.92	0.72	0.90	139	464	41	354	0.99	2.66	0.99	7.49	2.66
S07. $DIFF_{RR}$=0.00, CPR_{RR}=2.06	6.57	-44043.7	-311.4	1335.3	531.8	142.8	119.6	115.9	115.2	98.8	149.0	102.7	-24.6	110.9	43.5	0.95	0.81	0.95	160	444	60	336	0.99	2.04	0.99	4.32	2.04
S08. $DIFF_{RR}$=0.00, CPR_{RR}=1.50	6.57	-46353.8	-195.5	781.4	293.1	143.7	109.7	110.4	110.0	88.1	109.4	90.7	-18.3	73.2	27.4	0.94	0.87	0.93	201	406	98	297	0.99	1.51	0.99	2.33	1.51
S09. $DIFF_{RR}$=0.00, CPR_{RR}=1.29	6.57	-48092.7	-50.3	553.1	247.6	138.0	97.3	96.6	96.4	77.2	90.5	79.9	-5.0	54.7	24.4	0.96	0.93	0.96	230	376	128	268	0.99	1.29	0.99	1.70	1.29
S10. $DIFF_{RR}$=0.00, CPR_{RR}=1.11	6.57	-50601.6	-73.6	174.6	41.3	139.4	99.1	101.0	100.6	78.1	80.8	79.3	-7.6	17.9	4.2	0.94	0.92	0.94	274	338	168	225	0.99	1.09	0.99	1.23	1.09
S11. $DIFF_{RR}$=0.00, CPR_{RR}=1.07	6.57	-51087.4	-52.5	107.2	18.2	127.3	94.3	95.1	95.0	74.9	75.8	75.3	-5.5	11.1	1.9	0.95	0.95	0.95	285	328	179	215	0.99	1.05	0.99	1.14	1.05
S12. $DIFF_{RR}$=-0.24, CPR_{RR}=0.90	6.57	-76913.1	-127.2	860.4	-164.5	98.0	183.9	177.6	183.8	145.7	158.7	145.6	-6.4	48.8	-7.9	0.94	0.93	0.94	116	32	280	73	0.99	0.98	0.99	2.67	0.98
S13. $DIFF_{RR}$=-0.24, CPR_{RR}=0.69	6.57	-75889.9	-154.5	174.5	-548.1	104.9	157.3	157.7	160.1	126.0	126.9	135.2	-9.7	10.8	-33.5	0.95	0.95	0.94	101	49	264	89	0.99	0.71	0.99	1.37	0.71
S14. $DIFF_{RR}$=-0.24, CPR_{RR}=0.49	6.57	-74779.3	-381.1	-826.9	-1152.5	111.7	153.6	153.0	152.9	125.3	139.1	155.6	-25.5	-54.9	-76.2	0.94	0.91	0.87	75	72	234	114	0.99	0.52	0.99	0.69	0.52
S15. $DIFF_{RR}$=-0.24, CPR_{RR}=0.30	6.57	-73016.3	-185.4	-1662.5	-1455.8	113.4	156.1	152.4	151.4	124.9	187.2	172.6	-12.2	-111.1	-97.1	0.94	0.80	0.83	47	101	204	143	0.99	0.33	0.99	0.28	0.33
S16. $DIFF_{RR}$=-0.24, CPR_{RR}=0.15	6.57	-71935.2	-777.0	-3434.6	-2687.6	116.9	207.0	164.3	167.9	176.9	345.3	276.5	-40.5	-215.9	-163.0	0.92	0.43	0.64	20	124	179	165	0.99	0.15	0.99	0.06	0.15
S17. $DIFF_{RR}$=-0.48, CPR_{RR}=0.04	6.57	-40290.2	-1001.8	-4031.8	-3309.8	151.3	255.2	111.9	124.5	220.4	403.2	331.0	-44.9	-350.1	-257.7	0.89	0.06	0.30	10	298	334	363	0.99	0.04	0.99	0.00	0.04

Table 3.5 Properties of Poststratification, Raking, and GREG_Main under the Y_Additive_Interaction Model for Sensitivity Analysis (Continued)

	True Total	Relative Bias				Empirical Relative Standard Error				Relative Square Root of MSE			Bias Ratio			Coverage Rate of 95% Confidence Intervals			Average Respondent Sample Sizes by Cell				Average Cross Product Ratios of Unweighted or Weighted Cell Counts				
	$t_y \times 10^{-7}$	$RelBias(\hat{t}_{yw}) \times 10^5$				$EmpRelSE(\hat{t}_{yw}) \times 10^4$				$RelRMSE(\hat{t}_{yw}) \times 10^4$			$BiasRatio(\hat{t}_{yw}) \times 10^2$														
		No calibration	Poststratification	GREG_Main	Raking	No calibration	Poststratification	GREG_Main	Raking	Poststratification	GREG_Main	Raking	Poststratification	GREG_Main	Raking	Poststratification	GREG_Main	Raking	n_{r11}	n_{r11}	n_{r11}	n_{r11}	Population	Respondent Sample	Poststratification	GREG_Main	Raking
SRS Sample Size n=200																											
S01. $DIFF_{RR}$=0.00, CPR_{RR}=1.00	6.57	78.9	97.7	94.8	94.8	347.5	222.0	223.7	223.7	175.9	177.3	177.3	4.7	4.4	4.4	0.93	0.93	0.93	50	50	50	49	0.99	1.03	0.99	1.03	1.03
S02. $DIFF_{RR}$=0.00, CPR_{RR}=1.00	6.57	-64867.7	78.9	112.6	109.5	384.2	353.6	354.2	353.5	283.6	285.0	284.5	2.3	3.2	3.2	0.95	0.95	0.95	22	22	15	15	0.99	1.17	0.99	1.20	1.17
S03. $DIFF_{RR}$=-0.41, CPR_{RR}=1.00	6.57	-61156.2	88.7	1321.7	135.7	379.9	474.5	448.3	464.3	372.9	371.9	366.1	2.1	29.4	4.8	0.90	0.94	0.93	20	5	48	12	0.99	1.21	0.99	4.23	1.21
S04. $DIFF_{RR}$=0.00, CPR_{RR}=4.75	6.57	-70950.0	168.8	1682.9	885.0	392.5	639.1	609.9	616.3	508.6	504.7	494.9	7.0	33.5	21.5	0.84	0.91	0.92	6	23	2	18	0.99	2.16	0.99	5.04	2.16
S05. $DIFF_{RR}$=0.00, CPR_{RR}=3.41	6.57	-42044.0	-150.6	1955.5	990.2	485.3	509.7	442.9	453.3	402.9	387.0	370.5	-1.4	48.4	27.6	0.86	0.90	0.92	13	47	4	37	0.99	3.36	0.99	12.27	3.36
S06. $DIFF_{RR}$=0.00, CPR_{RR}=2.74	6.57	-42929.7	-268.2	1753.3	809.7	465.6	476.7	416.0	428.5	378.9	361.3	348.6	-5.8	44.0	22.6	0.89	0.92	0.93	14	46	4	36	0.99	2.91	0.99	9.68	2.91
S07. $DIFF_{RR}$=0.00, CPR_{RR}=2.06	6.57	-44090.7	-222.8	1408.9	647.6	461.7	423.5	392.6	395.5	338.6	332.5	319.5	-4.7	37.7	18.8	0.91	0.92	0.94	16	44	6	34	0.99	2.46	0.99	6.72	2.46
S08. $DIFF_{RR}$=0.00, CPR_{RR}=1.50	6.57	-46395.2	-257.2	723.4	234.7	457.4	367.8	360.8	361.2	293.0	288.6	286.5	-7.0	21.5	7.8	0.92	0.92	0.93	20	41	10	30	0.99	1.70	0.99	3.03	1.70
S09. $DIFF_{RR}$=0.00, CPR_{RR}=1.29	6.57	-47975.4	86.9	677.7	370.3	442.5	340.0	340.8	339.5	270.3	273.4	270.1	2.7	20.7	11.5	0.93	0.93	0.93	23	38	13	27	0.99	1.40	0.99	1.98	1.40
S10. $DIFF_{RR}$=0.00, CPR_{RR}=1.11	6.57	-50730.3	-19.6	223.8	98.5	444.6	319.2	322.0	321.6	254.8	258.0	257.7	0.0	7.6	3.7	0.93	0.94	0.94	27	33	17	23	0.99	1.20	0.99	1.39	1.20
S11. $DIFF_{RR}$=0.00, CPR_{RR}=1.07	6.57	-50851.5	161.1	324.3	235.2	433.0	318.6	321.0	320.7	254.6	257.0	256.1	5.1	10.0	7.2	0.94	0.95	0.95	28	33	18	22	0.99	1.15	0.99	1.27	1.15
S12. $DIFF_{RR}$=-0.24, CPR_{RR}=0.90	6.57	-76648.1	45.0	765.6	-166.2	324.5	596.7	574.4	590.2	468.8	453.3	463.6	0.6	14.1	-1.0	0.88	0.92	0.91	11	4	28	7	0.99	1.03	0.99	2.84	1.03
S13. $DIFF_{RR}$=-0.24, CPR_{RR}=0.69	6.57	-75865.7	-117.4	205.2	-525.4	344.3	558.4	551.0	558.3	445.9	438.3	446.5	-3.1	3.9	-9.2	0.89	0.92	0.91	10	5	26	9	0.99	0.83	0.99	1.90	0.83
S14. $DIFF_{RR}$=-0.24, CPR_{RR}=0.49	6.57	-74721.9	-496.9	-1019.3	-1303.7	343.9	523.7	520.1	518.5	420.0	421.5	425.3	-11.4	-22.6	-28.0	0.91	0.92	0.92	7	7	23	12	0.99	0.64	0.99	0.98	0.64
S15. $DIFF_{RR}$=-0.24, CPR_{RR}=0.30	6.57	-72944.8	-129.5	-1486.2	-1302.0	366.5	543.0	507.9	509.3	431.9	424.5	420.6	-3.7	-32.1	-28.6	0.90	0.92	0.92	5	10	21	14	0.99	0.38	0.99	0.36	0.38
S16. $DIFF_{RR}$=-0.24, CPR_{RR}=0.15	6.57	-71996.3	-770.4	-2876.9	-2303.7	369.9	610.8	510.8	522.1	488.2	465.7	454.9	-17.8	-60.6	-50.1	0.86	0.89	0.90	3	12	18	16	0.99	0.26	0.99	0.15	0.26
S17. $DIFF_{RR}$=-0.48, CPR_{RR}=0.04	6.57	-40384.3	-1070.7	-3603.0	-2883.7	465.6	566.0	356.4	381.0	458.5	416.0	381.4	-29.4	-101.3	-80.8	0.78	0.85	0.87	2	30	33	36	0.99	0.09	0.99	0.02	0.09

3.9 Summary of Findings

This chapter compares the empirical properties of three widely used calibration estimators – poststratification, raking, and GREG_Main. The simulation results show that in the presence of nonresponse, the conclusions in Deville and Särndal (1992) that all the calibration estimators should perform approximately the same in large samples do not necessarily hold. The speed at which the calibration estimators that use the same set of covariates but different adjustment functions become equivalent also depends on the underlying outcome variable model. The differences between poststratification, raking, and GREG_Main can be either substantive or negligible depending on the outcome variable model and response model. We demonstrate the importance of accounting for the outcome variable model and response model when choosing the appropriate calibration estimator. The outcome variable model should be the driving factor. If a significant and strong interaction effect is present in the outcome variable model and the overall predictive power of the model is very strong (with R-squared value being close to 1), then poststratification outperforms the other two calibration estimators except in the special situation that the response model does not include a multiplicative interaction term, in which case raking performs almost equally well as poststratification. Raking preserves the multiplicative interaction effect that is internal in the data before calibration while GREG_Main does not, and this is why raking can be less biased than GREG_Main when the response model contains a strong multiplicative interaction term.

One interesting finding is that for a large sample, a small relative bias associated with an inappropriate calibration estimator can still lead to very poor coverage rate of the 95 percent confidence intervals. This is because the bias remains constant while the standard error decreases as the sample size increases, so a larger sample size tends to make the bias ratio higher.

The sensitivity analysis suggests that the differences between poststratification, raking, and GREG_Main are highly sensitive to the model specifications for the outcome variable. As the predictive power of the outcome variable model decreases, the advantage of poststratification over raking and GREG_Main becomes less substantial.

We understand that in practice, response propensity model often tends to drive the selection of auxiliary variables to be used in calibration. Quite often, survey practitioners either lack the knowledge of the outcome variable(s) or need to create a single set of weights for analyzing a range of outcome variables. Despite the practical limitations, a better understanding of the impacts of the outcome variable model and response model can provide a good framework for us to examine the variable and function form selection issues in calibration weighting. For example, using paradata for nonresponse adjustment has been a popular topic in the recent survey literature (Kreuter et al. 2010, Kreuter 2013). It is important to evaluate to what extent the paradata (for example, the number of call attempts to reach a target respondent) may be correlated with the outcome variable(s) (for example, employment status, tobacco use, mental health status). Including in the calibration model any paradata that correlates only to the response propensities but not to the outcome variable(s) does not help reduce potential nonresponse bias.

Chapter 4. A Proposed Distance Measure Related to the Potential Bias of Raking and GREG_Main

Chapter 3 compares the empirical properties of the poststratification, raking, and GREG_Main estimators over repeated sampling. In the real-world survey practice, only one sample can be fielded and all the estimates are based on that particular sample, so it is important to understand the properties of the calibration estimators conditioning on sample configuration. In this chapter, we propose a distance measure that is related to the magnitude of bias for raking and GREG_Main when the outcome variable model contains an interaction term (referred to as Y_Additive_Interaction in Chapter 3). For a particular sample, survey practitioners can use this distance measure as a diagnostic tool to gauge the *potential* impact of failing to incorporate a significant interaction term in the calibration process. Section 4.1 presents the general theory of the proposed distance measure. Section 4.2 discusses the application of the proposed distance measure in the SRS 2×2 table setting. Sections 4.3 and 4.4 show the simulation results over repeated sampling and conditioning on samples grouped by the proposed distance measure, respectively, followed by a summary of conclusions and limitations in Section 4.5.

4.1 General Theory

The distance measure we propose applies to raking and GREG_Main. It helps gauge the potential impact of omitting a significant interaction term between two auxiliary covariates in the calibration process. Assume that the two main effect variables have I and J categories, respectively. Based on (3.49), the potential bias of the raking estimator

is related to how much the estimated cell counts using the raked weights differ from the population counts (i.e., $\hat{N}_{ij}^{w} - N_{ij}$). A statistic that summarizes the differences is the distance measure *DIST*, defined as

$$DIST = m\left(\hat{\mathbf{N}} - \mathbf{N}\right)^{\mathrm{T}} \mathbf{V}(\hat{\mathbf{N}})^{-1}\left(\hat{\mathbf{N}} - \mathbf{N}\right) = m\left(\hat{\mathbf{p}} - \mathbf{p}\right)^{\mathrm{T}} \mathbf{V}(\hat{\mathbf{p}})^{-1}\left(\hat{\mathbf{p}} - \mathbf{p}\right) \tag{4.1}$$

where

$\mathbf{N} = \left(N_{11},\ldots,N_{1(J-1),}N_{21,},\ldots,N_{2(J-1),}\ldots,N_{ij},\ldots,N_{(I-1)(J-1)}\right)^{\mathrm{T}}$ is the vector of population benchmark totals for the cells defined by the two auxiliary variables, assuming that the cross-classification between the two variables are available;

$\hat{\mathbf{N}} = \left(\hat{N}_{11}^{w},\ldots,\hat{N}_{1(J-1),}^{w}\hat{N}_{21,}^{w},\ldots,\hat{N}_{2(J-1),}^{w}\ldots,\hat{N}_{ij}^{w},\ldots,\hat{N}_{(I-1)(J-1)}^{w}\right)^{\mathrm{T}}$ is the vector of estimated population totals from raking or GREG_Main for the cells defined by the two auxiliary variables;

$\mathbf{p} = \left(p_{11},\ldots,p_{1(J-1),}p_{21,},\ldots,p_{2(J-1),}\ldots,p_{ij},\ldots,p_{(I-1)(J-1)}\right)^{\mathrm{T}}$,

in which $p_{ij} = N_{ij} \Big/ \sum_{i=1}^{I}\sum_{j=1}^{J} N_{ij}$;

$\hat{\mathbf{p}} = \left(\hat{p}_{11},\ldots,\hat{p}_{1(J-1),}\hat{p}_{21,},\ldots,\hat{p}_{2(J-1),}\ldots,\hat{p}_{ij},\ldots,\hat{p}_{(I-1)(J-1)}\right)^{\mathrm{T}}$,

in which $\hat{p}_{ij} = \hat{N}_{ij}^{w} \Big/ \sum_{i=1}^{I}\sum_{j=1}^{J} \hat{N}_{ij}^{w} \approx \hat{N}_{ij}^{w} \Big/ \sum_{i=1}^{I}\sum_{j=1}^{J} N_{ij}$ because $\sum_{i=1}^{I}\sum_{j=1}^{J} \hat{N}_{ij}^{w} \approx \sum_{i=1}^{I}\sum_{j=1}^{J} N_{ij}$ for raking (when the process converges) and GREG_Main; and

$\mathbf{V}(\hat{\mathbf{N}})/m$ and $\mathbf{V}(\hat{\mathbf{p}})/m$ are the true variance-covariance matrices for $\hat{\mathbf{N}}$ and $\hat{\mathbf{p}}$, in which *m* is the number of sampled primary sampling units in a complex sample design and the sample size in an SRS design. That is, $\mathrm{Var}(\hat{\mathbf{N}}) = \mathbf{V}(\hat{\mathbf{N}})/m$ and $\mathrm{Var}(\hat{\mathbf{p}}) = \mathbf{V}(\hat{\mathbf{p}})/m$.

Note that there are only $(I-1)\times(J-1)$ elements, but not $I\times J$ elements, in $\hat{\mathbf{N}}-\mathbf{N}$ and $\hat{\mathbf{p}}-\mathbf{p}$ because $\sum_{j=1}^{J}(\hat{N}_{ij}^{w}-N_{ij})\approx 0$ and $\sum_{i=1}^{I}(\hat{N}_{ij}^{w}-N_{ij})\approx 0$ after the raking process (when the process converges) or the GREG_Main calibration process. *DIST* has a similar form as a generalized Wald statistic (Rao and Scott 1981). Whether the distance measure has the same value regardless of which set of $(I-1)\times(J-1)$ categories are used to construct the statistic needs to be further examined through some analytical work.

Our first goal is to obtain the probability distribution of the proposed distance measure in (4.1) under the null hypothesis H_o: $E(\hat{\mathbf{N}})=\mathbf{N}$ or $E(\hat{\mathbf{p}})=\mathbf{p}$, so we can use the statistical properties of the known probability distribution to make inference.

Based on Krewski and Rao (1981), we have $\sqrt{m}\hat{\mathbf{p}}\sim N(\mathbf{p},\mathbf{V})$ asymptotically (i.e., as $\mathbf{N}$ approaches infinity) under the null hypothesis. Their result does apply to multistage sample design with potentially varying probabilities at each stage but with the assumption that the primary sampling units are selected with replacement. Now define a vector $\mathbf{z}=\sqrt{m}(\hat{\mathbf{p}}-\mathbf{p})$. Under the null hypothesis $E(\hat{\mathbf{p}})=\mathbf{p}$, we have $\mathbf{z}\sim AN(\mathbf{0},\mathbf{V}(\hat{\mathbf{p}}))$. The distance measure *DIST* can be expressed as a quadratic form in $\mathbf{z}$ and $\mathbf{V}(\hat{\mathbf{p}})^{-1}$. Also, since $\mathbf{V}(\hat{\mathbf{p}})$ is positive definite and symmetric, it can be factored as $\mathbf{V}(\hat{\mathbf{p}})=\mathbf{L}\mathbf{L}^{\mathrm{T}}$, where $\mathbf{L}$ is a nonsingular, lower triangular matrix. Assuming that $\mathbf{V}(\hat{\mathbf{p}})$ is invertible, the distance measure in (4.1) can be re-written as

$$
\begin{aligned}
DIST &= \mathbf{z}^{\mathrm{T}}\mathbf{V}(\hat{\mathbf{p}})^{-1}\mathbf{z} \\
&= \mathbf{z}^{\mathrm{T}}\left(\mathbf{L}\mathbf{L}^{\mathrm{T}}\right)^{-1}\mathbf{z} \\
&= \mathbf{z}^{\mathrm{T}}\left(\mathbf{L}^{-1}\right)^{\mathrm{T}}\mathbf{L}^{-1}\mathbf{z}
\end{aligned}
\tag{4.2}
$$

We further define $\mathbf{w} = \mathbf{L}^{-1}\mathbf{z} = \sqrt{m}\mathbf{L}^{-1}\left(\hat{\mathbf{p}} - \mathbf{p}\right)$. Then, (4.2) can be re-written as

$$
DIST = \mathbf{w}^{\mathrm{T}}\mathbf{w} = \sum_{i=1}^{I-1}\sum_{j=1}^{J-1} W_{ij}^2 \tag{4.3}
$$

Under the null hypothesis $E(\hat{\mathbf{p}}) = \mathbf{p}$, we have

$$
E\left(\mathbf{w}\right) = E\left(\sqrt{m}\mathbf{L}^{-1}\left(\hat{\mathbf{p}} - \mathbf{p}\right)\right) = \mathbf{0} \tag{4.4}
$$

and

$$
\begin{aligned}
\mathrm{Var}\left(\mathbf{w}\right) &= \mathrm{Var}\left(\sqrt{m}\mathbf{L}^{-1}\left(\hat{\mathbf{p}} - \mathbf{p}\right)\right) \\
&= m\mathbf{L}^{-1}\,\mathrm{Var}\left(\hat{\mathbf{p}} - \mathbf{p}\right)\left(\mathbf{L}^{-1}\right)^{\mathrm{T}} \\
&= m\mathbf{L}^{-1}\,\mathrm{Var}\left(\hat{\mathbf{p}}\right)\left(\mathbf{L}^{-1}\right)^{\mathrm{T}} \\
&= m\mathbf{L}^{-1}\,\frac{\mathbf{V}\left(\hat{\mathbf{p}}\right)}{m}\left(\mathbf{L}^{-1}\right)^{\mathrm{T}} \\
&= \mathbf{L}^{-1}\mathbf{L}\mathbf{L}^{\mathrm{T}}\left(\mathbf{L}^{-1}\right)^{\mathrm{T}} \\
&= \mathbf{I}
\end{aligned}
\tag{4.5}
$$

Under the null hypothesis $E(\hat{\mathbf{p}}) - \mathbf{p} = 0$, we know $\mathbf{z} = \sqrt{m}\left(\hat{\mathbf{p}} - \mathbf{p}\right) \sim AN\left(\mathbf{0}, \mathbf{V}\right)$. Therefore, $\mathbf{w} \sim AN\left(\mathbf{0}, \mathbf{I}\right)$ and W_{ij}^2 in (4.3) are independent Chi-square(1) random variables. The

probability distribution of the distance measure *DIST* is the same as that of $\sum_{i=1}^{I-1}\sum_{j=1}^{J-1} W_{ij}^2$, which is central χ^2 with $(I-1)\times(J-1)$ degrees of freedom.

Our second question is: if $E(\hat{\mathbf{p}})-\mathbf{p}=\boldsymbol{\Delta}$ for some non-zero $\boldsymbol{\Delta}$ (i.e., the null hypothesis should be rejected), then what is the distribution of *DIST* defined in (4.1)? At a given relative bias level, the distance measure tends to increase with the sample size. So the question is how large the distance measure should be to make it practically important. This involves the power theory about the distance measure.

Suppose that $\sqrt{m}\left(\hat{\mathbf{p}}-\mathbf{p}\right)\sim AN\left(\boldsymbol{\Delta},\mathbf{V}\right)$. Define $\mathbf{z}=\sqrt{m}\left(\hat{\mathbf{p}}-\mathbf{p}\right)$, $\mathbf{V}(\hat{\mathbf{p}})=\mathbf{L}\mathbf{L}^{\mathrm{T}}$, and $\mathbf{w}=\mathbf{L}^{-1}\mathbf{z}$ as in the earlier proof, where $\mathbf{L}$ is a nonsingular, lower triangular matrix. When $E(\hat{\mathbf{p}})-\mathbf{p}=\boldsymbol{\Delta}$ for some non-zero $\boldsymbol{\Delta}$, the distance measure still has the forms shown in (4.2) and (4.3), where the variance-covariance matrix for $\mathbf{w}$ is shown in (4.5) and the mean of $\mathbf{w}$ is

$$E\left(\mathbf{w}\right)=E\left(\mathbf{L}^{-1}\mathbf{z}\right)=\mathbf{L}^{-1}E\left(\mathbf{z}\right)=\mathbf{L}^{-1}\sqrt{m}E\left(\hat{\mathbf{p}}-\mathbf{p}\right)=\mathbf{L}^{-1}\sqrt{m}\boldsymbol{\Delta} \qquad (4.6)$$

That is, $\mathbf{w}\sim AN\left(\mathbf{L}^{-1}\sqrt{m}\boldsymbol{\Delta},\ \mathbf{I}\right)$. According to Searle (1971, Section 2.4h), *DIST* □ noncentral χ^2 with $(I-1)\times(J-1)$ degrees of freedom when $E(\hat{\mathbf{p}})-\mathbf{p}=\boldsymbol{\Delta}$ for some non-zero $\boldsymbol{\Delta}$. A noncentral χ^2 distribution involves the noncentrality parameter δ (which is a scaler) as shown in (4.7).

$$
\begin{aligned}
\delta &= \frac{1}{2}\left(E\left(\mathbf{w}\right)\right)^{\mathrm{T}} E\left(\mathbf{w}\right) \\
&= \frac{1}{2}\left(\mathbf{L}^{-1}\sqrt{m}\boldsymbol{\Delta}\right)^{\mathrm{T}}\left(\mathbf{L}^{-1}\sqrt{m}\boldsymbol{\Delta}\right) \\
&= \frac{1}{2} m\boldsymbol{\Delta}^{\mathrm{T}}\left(\mathbf{L}^{-1}\right)^{\mathrm{T}}\mathbf{L}^{-1}\boldsymbol{\Delta} \\
&= \frac{1}{2} m\boldsymbol{\Delta}^{\mathrm{T}}\mathbf{V}(\hat{\mathbf{p}})^{-1}\boldsymbol{\Delta} \\
&= \frac{1}{2}\boldsymbol{\Delta}^{\mathrm{T}}\mathrm{Var}(\hat{\mathbf{p}})^{-1}\boldsymbol{\Delta}
\end{aligned}
\tag{4.7}
$$

In practice, we can specify a level of relative bias (for an estimator of cell population count) that is important to detect, say b=0.10 (i.e., 10 percent relative bias). For simplicity, we assume that the same b value is specified for all the cells ij. That is, $b = E(\hat{p}_{ij} - p_{ij})/p_{ij} = \Delta_{ij}/p_{ij}$ and $\boldsymbol{\Delta} = E(\hat{\mathbf{p}}) - \mathbf{p} = b\mathbf{p}$. Then, we can evaluate how much power the *DIST* test has at the specified relative bias level b.

We can calculate the noncentrality parameter for a given b using

$$
\begin{aligned}
\delta &= \frac{1}{2} m(b\mathbf{p})^{\mathrm{T}}\mathbf{V}(\hat{\mathbf{p}})^{-1}(\mathbf{p}b) \\
&= \frac{mb^2}{2}\mathbf{p}^{\mathrm{T}}\mathbf{V}(\hat{\mathbf{p}})^{-1}\mathbf{p} \\
&= \frac{mb^2}{2}\mathbf{p}^{\mathrm{T}}\mathrm{Var}(\hat{\mathbf{p}})^{-1}\mathbf{p}
\end{aligned}
\tag{4.8}
$$

Then, Power = $\Pr(DIST > c) = \Pr(\chi^2_{(I-1)(J-1),\delta} > c)$, where c is a critical point for a central χ^2 with $(I-1)\times(J-1)$ degrees of freedom and $\chi^2_{(I-1)(J-1),\delta}$ is a noncentral χ^2 with $(I-1)\times(J-1)$ degrees of freedom and the noncentral parameter in (4.7). In practice, we

can evaluate the power for a range of specified relative bias levels, such as $-0.2 \le b \le 0.2$, using the corresponding δ's calculated through (4.8). Note that $\mathbf{p}$ has to be obtained from external sources. We do not know the true variance-covariance matrices Var($\hat{\mathbf{N}}$) and Var($\hat{\mathbf{p}}$), and thus need to estimate the values from an achieved sample. For a consistent variance estimator, the estimated variance approaches the true variance as the sample size approaches infinity. At the same time, survey practitioners may face small sample size problems in the real world sometimes, which make the estimated variance unstable. In our simulation study, we include some small sample size scenarios (with SRS n=200) to help us understand whether the proposed distance measure can really be useful in practice.

4.2 Application in the 2×2 Table Setting

In the 2×2 table setting, conditions (4.9) through (4.12) are satisfied as the result of raking or GREG_Main calibration.

$$\hat{N}_{11}^{w} + \hat{N}_{12}^{w} = N_{11} + N_{12} \tag{4.9}$$

$$\hat{N}_{11}^{w} + \hat{N}_{21}^{w} = N_{11} + N_{21} \tag{4.10}$$

$$\hat{N}_{12}^{w} + \hat{N}_{22}^{w} = N_{12} + N_{22} \tag{4.11}$$

$$\hat{N}_{21}^{w} + \hat{N}_{22}^{w} = N_{21} + N_{22} \tag{4.12}$$

That is,

$$\hat{N}_{11}^{w} - N_{11} = -(\hat{N}_{12}^{w} - N_{12}) = -(\hat{N}_{21}^{w} - N_{21}) = \hat{N}_{22}^{w} - N_{22} \tag{4.13}$$

Given that $Var(\hat{N}_{ij}^{w} - N_{ij}) = Var(\hat{N}_{ij}^{w})$, we have

$$Var(\hat{N}_{11}^{w}) = Var(\hat{N}_{12}^{w}) = Var(\hat{N}_{21}^{w}) = Var(\hat{N}_{22}^{w}) \tag{4.14}$$

As the result of (4.13) and (4.14), $DIST = \dfrac{(\hat{N}_{ij}^{w} - N_{ij})^2}{\mathrm{Var}\left(\hat{N}_{ij}^{w}\right)}$ is the same regarless of which category is deleted in a 2×2 table.

As discussed in Chapter 3, the outcome variable model may be Y_Main or Y_Additive_Interaction depending on whether the model contains an interaction term. To facilitate the discussions in this chapter, we use $Y_{ijk} = \mu_Y + \alpha_{Yi} + \beta_{Yj} + \gamma_{Yij} + \varepsilon_{Yijk}$ as the general form for both Y_Main and Y_Additive_Interaction. That is, $\gamma_{Yij} = 0$ for Y_Main and $\gamma_{Yij} \neq 0$ for Y_Additive_Interaction.

A poststratification, raking, or GREG_Main calibration estimator for a total associated with a 2×2 table can be expressed as

$$\hat{t}_{yw} = \sum_{i=1}^{2}\sum_{j=1}^{2} w_{ij} \sum_{k=1}^{n_{ij}} y_{ijk} \tag{4.15}$$

where w_{ij} is the calibrated weight for a unit k in cell ij.

Under the general form for the outcome variable model, the model expectation of the calibration estimator $\hat{t}_{yw}$ can be expressed as

$$
\begin{aligned}
&\mathrm{E_M}(\hat{t}_{yw}) \\
&= \sum_{i=1}^{2}\sum_{j=1}^{2} w_{ij} \sum_{k=1}^{n_{ij}} \mathrm{E_M}(y_{ijk}) \\
&= \sum_{i=1}^{2}\sum_{j=1}^{2} w_{ij} \sum_{k=1}^{n_{ij}} \mathrm{E_M}(\mu_Y + \alpha_{Yi} + \beta_{Yj} + \gamma_{Yij} + \varepsilon_{Yijk}) \\
&= \sum_{i=1}^{2}\sum_{j=1}^{2} w_{ij} \sum_{k=1}^{n_{ij}} (\mu_Y + \alpha_{Yi} + \beta_{Yj} + \gamma_{Yij}) \\
&= \mu_Y \sum_{i=1}^{2}\sum_{j=1}^{2} w_{ij} n_{ij} + \sum_{i=1}^{2} \alpha_{Yi} \sum_{j=1}^{2} w_{ij} n_{ij} + \sum_{j=1}^{2} \beta_{Yj} \sum_{i=1}^{2} w_{ij} n_{ij} + \sum_{i=1}^{2}\sum_{j=1}^{2} \gamma_{Yij} w_{ij} n_{ij} \\
&= \mu_Y \hat{N}^w + \sum_{i=1}^{2} \alpha_{Yi} \hat{N}^w_{i\cdot} + \sum_{j=1}^{2} \beta_{Yj} \hat{N}^w_{\cdot j} + \sum_{i=1}^{2}\sum_{j=1}^{2} \gamma_{Yij} \hat{N}^w_{ij}
\end{aligned}
\tag{4.16}
$$

Then the model bias of the estimator $\hat{t}_{yw}$ is

$$
\begin{aligned}
&\mathrm{E_M}(\hat{t}_{yw} - t_y) \\
&= \mu_Y (\hat{N}^w - N) + \sum_{i=1}^{2} \alpha_{Yi} (\hat{N}^w_{i\cdot} - N_{i\cdot}) + \sum_{j=1}^{2} \beta_{Yj} (\hat{N}^w_{\cdot j} - N_{\cdot j}) + \sum_{i=1}^{2}\sum_{j=1}^{2} \gamma_{Yij} (\hat{N}^w_{ij} - N_{ij})
\end{aligned}
\tag{4.17}
$$

During the calibration process, raking (when converged), poststratification, and GREG_Main can all force the estimated row totals and column totals to be equal or approximately equal to the marginal control totals. That is, the terms $(\hat{N}^w - N)$, $(\hat{N}^w_{i\cdot} - N_{i\cdot})$, and $(\hat{N}^w_{\cdot j} - N_{\cdot j})$ are expected to be zero regardless of which of the three calibration estimators is used, making the first three terms in (4.17) zero. However, whether the fourth term in (4.17) is zero may depend on the outcome variable model, response model, and calibration process. If the outcome model is Y_Main, then $\gamma_{Yij} = 0$ and the fourth term is zero regardless of the value for $(\hat{N}^w_{ij} - N_{ij})$. If the outcome variable

model is Y_Additive_Interaction (with $\gamma_{Yij} \neq 0$), then poststratification forces $(\hat{N}_{ij}^{w} - N_{ij})$ to be zero and therefore is model-unbiased, but GREG_Main and raking are model-biased except in some special situations. One special situation is $\gamma_{Y11} - \gamma_{Y12} - \gamma_{Y21} + \gamma_{Y21} = 0$, which makes $\sum_{i=1}^{2}\sum_{j=1}^{2} \gamma_{Yij}(\hat{N}_{ij}^{w} - N_{ij}) = 0$ due to the condition in (4.13).

Although we normally do not know the values for μ_Y, α_{Yi}, β_{Yj}, and γ_{Yij} in the outcome variable model, we can compute $(\hat{N}_{ij}^{w} - N_{ij})$ as long as the classification and corresponding cell totals for the population are available. The larger the magnitude of $(\hat{N}_{ij}^{w} - N_{ij})$ is, the more severe the potential bias is for raking and GREG_Main. In a national survey of general population, for example, the marginal control totals $N_{i\cdot}$ and $N_{\cdot j}$ can probably be obtained from either the Census or the Census population projections or estimates. The cross-classification control totals N_{ij} may be estimated from some large samples such as American Community Survey and Current Population Survey. These estimated totals are often treated as known population truth during the calibration process (Dever 2008). However, quite often, raking or GREG_Main is used in practice mainly because only the marginal control totals, but not the cross-classification cell totals, are available. In this situation, the proposed distance measure is still useful for conducitng sensitivity analysis. For example, survey practitioners can create a set of hypothetical cross-classification cell totals based on various assumptions about the interaction effect between the auxiliary variables, and then use the hypothetical cross-clasification cell totals to compute the distance measures. The range of the estimated

distance measure can help us gauge the potential impact of the interaction effect tin the control totals.

Given the conditions in (4.13) and (4.14), we can use the information from any of the four cells to compute the distance measure. The estimated distance measure in the SRS 2×2 table setting is

$$\hat{DIST}_s = \frac{(\hat{N}_{11}^w - N_{11})^2}{\text{var}\left(\hat{N}_{11}^w\right)} = \frac{(\hat{N}_{12}^w - N_{12})^2}{\text{var}\left(\hat{N}_{12}^w\right)} = \frac{(\hat{N}_{21}^w - N_{21})^2}{\text{var}\left(\hat{N}_{21}^w\right)} = \frac{(\hat{N}_{22}^w - N_{22})^2}{\text{var}\left(\hat{N}_{22}^w\right)} \tag{4.18}$$

where $\text{var}\left(\hat{N}_{ij}^w\right)$ is the estimated value of $\text{Var}\left(\hat{N}_{ij}^w\right)$ from the sample, i=1, 2; j=1, 2.

The distance measure in (4.18) follows a Chi-square distribution with one degree of freedom. On the one hand, the term in (4.17) that is related to the potential bias of the calibration estimator, $(\hat{N}_{ij}^w - N_{ij})$, is not a function of sample size for any particular sample (since $\hat{N}_{ij}^w$ is fixed for a given sample). On the other hand, the estimated distance measure in (4.18) is a function of the sample size because its denominator is the estimated variance of the estimated population size for cell *ij*. As contradictory as this may seem, we choose to define the distance measure in the general form shown in (4.1) for two main reasons. First, although we can obtain the distribution of $(\hat{N}_{ij}^w - N_{ij})$ across all the iterations in a simulation study, only one sample can be obtained in practice. We have no knowledge of the distribution of $(\hat{N}_{ij}^w - N_{ij})$, and thus no decision rule for determining whether a value is "too large" or not. The distance measure we propose, however, follows a known probability distribution (i.e., Chi-square distribution) under the

null hypothesis, so it allows us to make statistical inference from a single sample. Second, as discussed in Chapter 3, bias ratio may be a more important indicator of the performance of a calibration estimator than absolute or relative bias. The former can be more revealing than the latter because a large bias ratio often means an unacceptable coverage rate of the 95 percent confidence intervals even when the bias is small. The bias ratio is a function of the sample size, with the order of the square root of the order for the proposed distance measure. That is, we suspect that the proposed distance measure has the advantage of being highly correlated to the bias ratio under Y_Additive_Interaction.

One way to use the proposed distance measure is to compare the estimated distance measure from a given sample to the critical values of the Chi-square distribution. For example, $\text{Prob}(0.004 < \chi^2(1) < 3.84) = 0.95$ and $\text{Prob}(0.000 < \chi^2(1) < 6.63) = 0.99$, so the upper tail critical values for Chi-square distribution with one degree of freedom is 3.84 at 5 percent significance level and 6.63 at 1 percent significance level. If the estimated distance measure from a SRS 2×2 table, $\hat{DIST}_s$, is 5.0, then we consider it "too large" at 5 percent significant level, but not "too large" at 1 percent significance level. Knowing whether the estimated distance measure is "too large" can help us determine whether the raking estimator or GREG_Main estimator is *potentially* biased. On the one hand, it is important to note that "$\hat{DIST}_s$ not being too large" is a sufficient yet not a unnecessary condition for the model-unbiasedness of raking and GREG_Main estimators. As (4.17) shows, in the Y_Main scenario, $\gamma_{Yij} = 0$, so raking and GREG_Main are

unbiased regardless of the value of $(\hat{N}_{ij}^{w} - N_{ij})$ or the estimated distance measure. The bias (or more accurately, bias ratio) of a raking estimator or GREG_Main estimator is associated with the distance measure only under Y_Additive_Interaction outcome model. On the other hand, a real-world survey contains a number of key outcome measures and it is rare that none of the outcome measures is governed by a Y_Additive_Interaction model. If raking or GREG_Main is used for the calibration weighting of a given sample, then a large value of the estimated distance measure is probably a warning sign of potential bias for some variables due to omitting a significant interaction term in the calibration process.

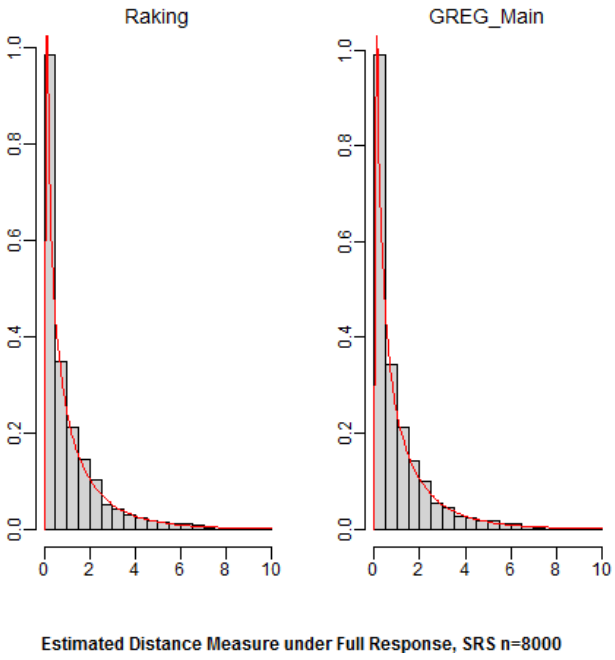
4.3 Simulation Results over Repeated Sampling

This section demonstrates the properties of the proposed distance measure and its relationships with bias and bias ratio over repeated sampling. All the simulation work is based on the Y_Additive_Interaction model with R^2=0.9979 (shown in Table 3.1) because it is under this outcome model that raking and GREG_Main may be severely biased. We do not cover the Y_Additive_Main scenario because when the outcome variable model does not include an interaction term, all the three calibration estimators are expected to be unbiased despite the magnitude of the proposed distance measure. Section 4.3.1 examines the empirical distributions of the proposed distance measures for raking and GREG_Main under full response, in which the null hypothesis $E(\hat{\mathbf{N}}) = \mathbf{N}$ or $E(\hat{\mathbf{p}}) = \mathbf{p}$ is true. Section 4.3.2 evaluates the relationships between the strength of the multiplicative interaction term in the response model, the proposed distance measure, and

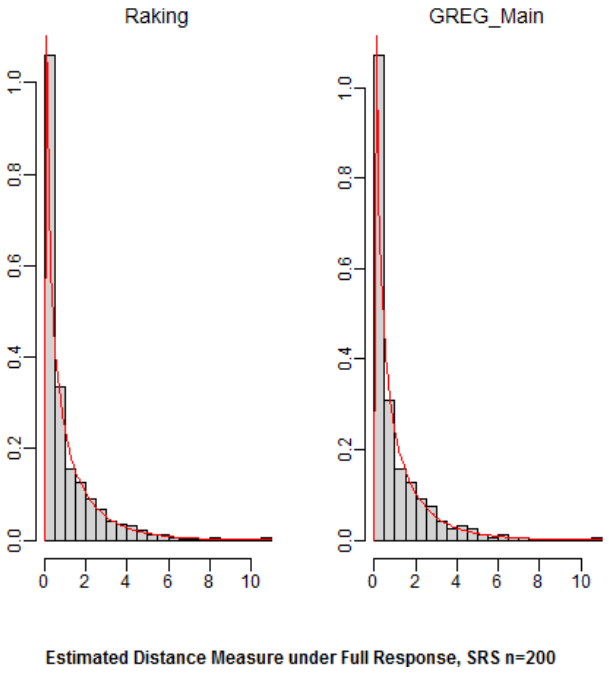
the empirical bias and bias ratio of the calibration estimator over repeated sampling (i.e., averaging across the 1,000 simulated samples for each response scenario). The simulation scenarios and procedure are very similar to what is described in Chapter 3, with all the 17 response scenarios included. Two alternative SRS sample sizes, n=8,000 and n=200, are used to evaluate whether a small sample size may affect the usefulness of the proposed distance measure. In addition to the evaluation parameters described in Chapter 3, the distance measures for raking and GREG_Main are also estimated from each simulated sample using (4.18).

4.3.1 Distribution of Estimated Distance Measure under Full Response

When there is full response, raking and GREG_Main are both unbiased regardless of the outcome variable model. If the theory presented in Section 4.1 holds, then we expect that in the response scenario S01 (which is full response, as described in Chapter 3), the estimated distance measure should follow Chi-squared distribution with one degree of freedom. Figure 4.1 shows the histograms of the estimated distance measures for raking and GREG_Main over the 1,000 simulated samples. Panels (a) and (b) are for two SRS sample sizes, n=8,000 and n=200, respectively. The distributions of the estimated distance measures seem to align well with the $\chi^2(1)$ distribution curve shown in red.



(a) SRS Sample Size *n*=8,000



(b) SRS Sample Size *n*=200

Figure 4.1 Histograms of Estimated Distance Measures for Raking and GREG_Main under Full Response against Chi-square Distribution with One Degree of Freedom

Table 4.1 shows some key statistics of the estimated distance measures for raking and GREG_Main under full response, for SRS sample sizes n=8,000 and n=200, respectively. First, we examine the empirical mean $E_p(\hat{DIST}) = (1/S)\sum_{s=1}^{S} \hat{DIST}_s$ and the empirical variance $EmpVar(\hat{DIST}) = (1/S)\sum_{s=1}^{S}\left(\hat{DIST}_s - E_p(\hat{DIST})\right)^2$. The empirical means of the estimated distance measures range from approximately 0.98 to approximately 1.02. The empirical variances of the estimated distance measures range from approximately 1.84 to approximately 1.93. These values are reasonably close to the mean and the variance of the $\chi^2(1)$ probability distribution (i.e., the mean should be one and the variance should be two). Second, for n=8,000, the 95th percentiles of the estimated distance measures are very close to 3.84 (approximately 3.88 for raking and approximately 3.84 for GREG_Main) and the 99th percentiles are reasonably close to 6.63 (approximately 6.22 for both raking and GREG_Main). For n=200, the numbers are slightly more off (approximately 3.80 and 5.72 for raking and approximately 3.96 and 6.12 for GREG_Main) probably due to a smaller sample size. Across the simulation iterations, the proportions of samples with the estimated distance measure larger than 3.84 are close to 5 percent (ranging from approximately 5.0 percent to approximately 5.5 percent). The estimated proportions of samples with the estimated distance measure larger than 6.63 are not far from 1 percent (all approximately 0.7 percent). Finally, we use a one-sample Kolmogorov-Smirnov test to compare the distribution of the estimated distance measure to the $\chi^2(1)$ probability distribution. The p-values are all larger than 0.05, so the distributions of these estimated distance measures are not significantly different (at the 5

percent significance level) from the $\chi^2(1)$ distribution. The *p*-values for *n*=8,000 (approximately 0.16 for raking and approximately 0.12 for GREG_Main) are lower than those for *n*=200 (approximately 0.33 for both raking and GREG_Main) mainly because they are associated with a larger sample.

Table 4.1 Statistics of Estimated Raking and GREG_Main Distance Measures under Full Response

	SRS Sample Size *n*=8,000		SRS Sample Size *n*=200	
	Raking	GREG_Main	Raking	GREG_Main
Empirical Mean $E_p(\hat{DIST})$	1.02	1.02	0.98	1.00
Empirical Variance $EmpVar(\hat{DIST})$	1.84	1.84	1.86	1.93
95th percentile of $\hat{DIST}$	3.88	3.84	3.80	3.96
99th percentile of $\hat{DIST}$	6.22	6.22	5.72	6.12
Percent of samples with $\hat{DIST}_s > 3.84$	5.1%	5.0%	5.0%	5.5%
Percent of samples with $\hat{DIST}_s > 6.63$	0.7%	0.8%	0.7%	0.7%
p-value for One-Sample Kolmogorov-Smirnov Test of $\hat{DIST}$ Distribution against $\chi^2(1)$	0.16	0.12	0.33	0.33

4.3.2 Interaction Effect in Response Model, Distance Measure, and Bias

For each SRS sample size and response scenario combination, we calculate the average relative bias, average bias ratio, coverage rate of the 95 percent confidence intervals, and some statistics about the estimated distance measures over the 1,000 simulated samples. The results are shown in Table 4.2 and Figures 4.2 through 4.4, from which we can draw four conclusions.

The first conclusion is that the magnitude of distance measure is positively correlated with the strength of the interaction term in the response model. This relationship is

clearly manifested in the scenarios with SRS sample size n=8,000. The expectations of the estimated distance measures are close to one for raking (which is the mean of the $\chi^2(1)$ distribution) when $CPR_{RR} = 1$ (in the response scenarios S01, S02, and S03). As CPR_{RR} moves away from one, the estimated distance measures for raking generally become larger. For example, the expectation of the estimated distance measures is 12.7 when $CPR_{RR} = 1.29$ (in the response scenario S09), but increases to 28.0 when $CPR_{RR} = 4.75$ (in the response scenario S04) and to 243.9 when $CPR_{RR} = 0.04$ (in the response scenario S17). The GREG_Main distance measure follows a similar pattern except that it is driven by not only the multiplicative interaction effect, but also the additive interaction effect, in the response model. The correlation between the distance measure and the strength of the interaction term in the response model can also be observed for the response scenarios under the SRS sample size n=200, although the range of the estimated distance measures are much smaller for n=200 than that for n=8,000. The smaller range is due to two reasons. First, the numerator of the distance measure does not depend on the sample size, but the denominator (which is the variance of an estimator) increases as the sample size becomes smaller. Second, the variance in the denominator of the distance measure is estimated using a linearization method implemented in the R Survey package. This method tends to overestimate the variance for raking as the multiplicative interaction effect in the response model becomes stronger (more details are provided in Chapter 5), and the impact of such overestimation seems more noticeable for n=200 than for n=8,000.

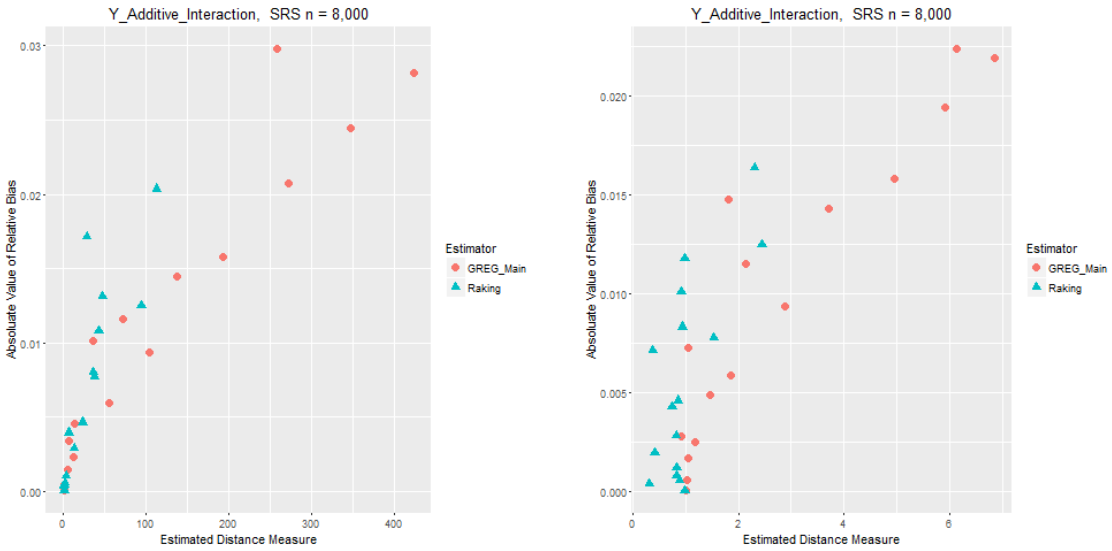
The second conclusion is that the proportion of samples with extreme distance measure also depends on the strength of the interaction effect in the response model. For S01 and S02, only approximately four to eight percent of the simulated samples have estimated distance measures larger than 3.84 and approximately one to two percent of the simulated samples have estimated distance measures larger than 6.63. These percentages largely reflect the magnitude of Type I error. As the interaction effect in the response model becomes stronger, the proportion of samples with extreme distance measure increases. For raking with SRS sample size n=8,000, when $CPR_{RR} = 1.07$ (in the response scenario S11), only approximately 11 percent of the samples have $\hat{DIST}_{raking_s} > 3.84$ and approximately 3 percent of the samples have $\hat{DIST}_{raking_s} > 6.63$. This means that for the majority of the simulated samples, the estimated distance measures fall within the range of the 95 percent or 99 percent confidence interval of the $\chi^2(1)$ distribution. For these majority of samples, the raking estimator does a reasonably good job in terms of reducing bias and producing accurate confidence interval estimate. When CPR_{RR} increases to as large as 1.50 or decreases to as small as 0.49, the expectation of the estimated distance measure for SRS sample size n=8,000 becomes much larger than 3.84 or 6.63 (being 22.9 for $CPR_{RR} = 1.50$ and 37.5 for $CPR_{RR} = 0.49$). This indicates noticeable bias, large bias ratio, and unacceptable coverage rate of the 95 percent confidence intervals for raking and GREG_Main.

The third conclusion is that there is positive correlation between the magnitude of bias as well as bias ratio and the distance measure. To demonstrate this more clearly, we take

the absolute values of the relative biases and bias ratios for raking and GREG_Main, and then plot them against the corresponding distance measures for all the response scenarios except S17. The magnitude of bias and distance measure for S17 is much larger than that for the other response scenarios, so we exclude such extreme data points to improve the representational value of the graphs. Figure 4.2 shows the relationship between the absolute value of relative bias and the distance measure. Figure 4.3 shows the relationship between the absolute value of bias ratio and the distance measure. For both figures, panel (a) is for the SRS sample size n=8,000 and panel (b) is for the SRS sample size n=200. The data patterns for the two sample sizes are similar. Although the absolute value of the relative bias and the absolute value of the bias ratio both increase as the distance measure becomes larger, the distance measure is a more precise predictor of the latter (shown in Figure 4.3) than the former (shown in Figure 4.2). As discussed in Section 4.2, the bias ratio and the distance measure are both functions involving both absolute bias and sample size. This is why the data points in Figure 4.3 reveal a clearer pattern than those in Figure 4.2. We suspect that if we plot the square of the bias ratio against the distance measure, we are likely to see a positive linear relationship between the two.

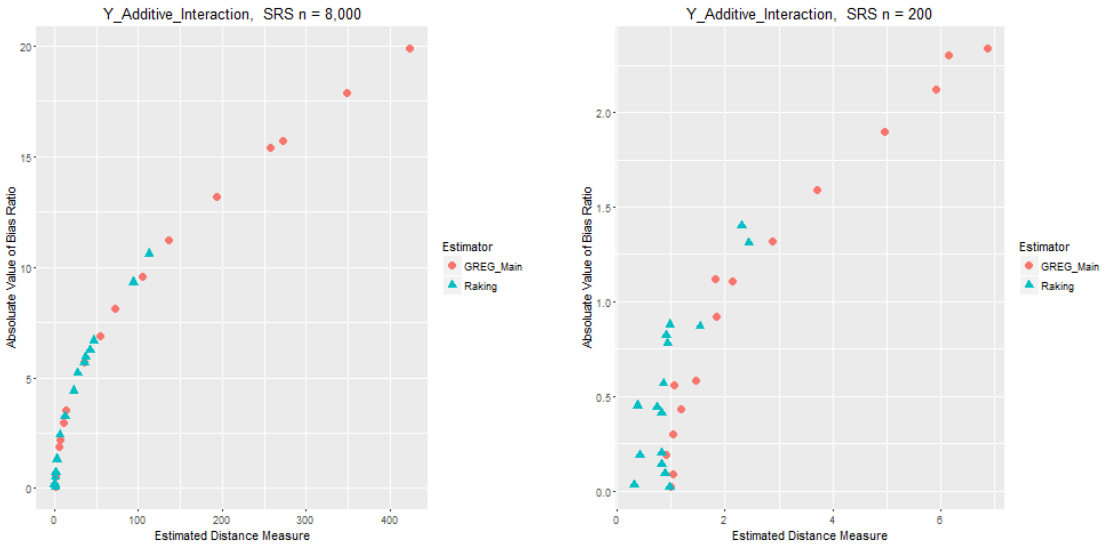
The fourth conclusion is that, as shown in Figure 4.3, in each response scenario with CPR_{RR} being away from one, the distance measure for the SRS sample size n=200 is substantially smaller than that for n=8,000. This is due to the larger variance in the denominator of the distance measure for a smaller sample size. The same pattern holds for the bias ratio. That is, although the magnitude of the bias does not depend on the

sample size, the distance measure and bias ratio both decrease as the sample size decreases, resulting in a better coverage rate of the confidence intervals under the given bias level. As shown in Figure 4.4 (a), the coverage rates of the 95 percent confidence intervals are unacceptable for most response scenarios under n=8,000. In contrast, the coverage rates of the 95 percent confidence intervals for n=200 are close to 95 percent except for a few response scenarios with CPR_{RR} being far away from one. Instead of predicting the bias level, the proposed distance measure is actually a good indicator of the bias ratio and the quality of the coverage rate of the 95 percent confidence intervals. If the estimated distance measure is "too large", then the survey practitioner should be warned of the possibly poor coverage rate of the confidence intervals.



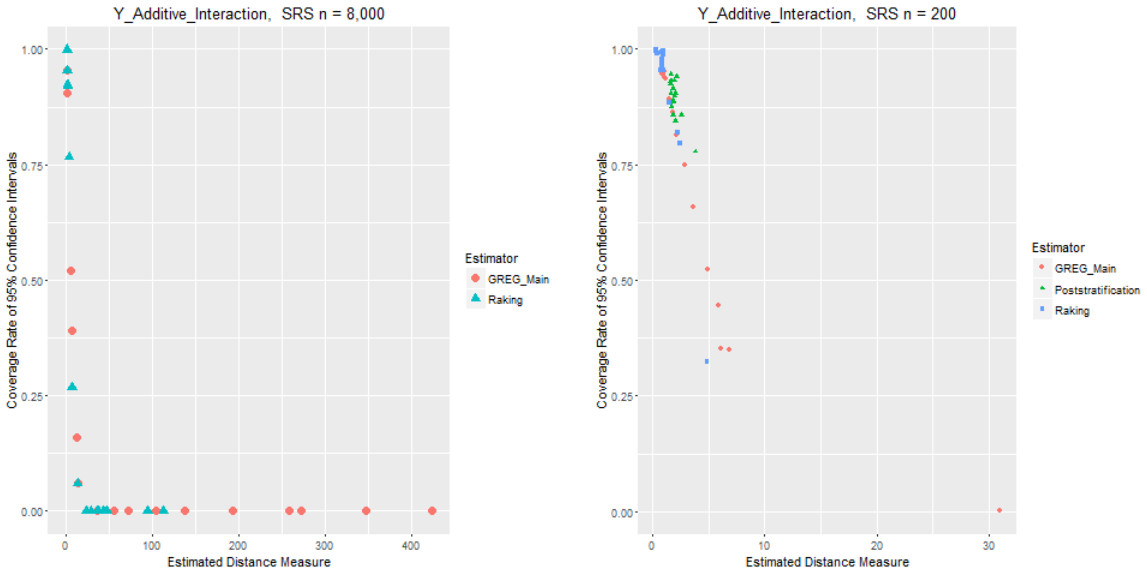
(a) SRS Sample Size *n*=8,000 (b) SRS Sample Size *n*=200

Figure 4.2 Absolute Values of Relative Biases versus Estimated Distance Measures under Y_Additive_Interaction and Various Response Scenarios



(a) SRS Sample Size *n*=8,000 (b) SRS Sample Size *n*=200

Figure 4.3 Absolute Values of Bias Ratios versus Estimated Distance Measures under Y_Additive_Interaction and Various Response Scenarios



(a) SRS Sample Size *n*=8,000 (b) SRS Sample Size *n*=200

Figure 4.4 Coverage Rates of 95% Confidence Intervals versus Estimated Distance Measures under Y_Additive_Interaction and Various Response Scenarios

Table 4.2 Relative Bias, Bias Ratio, Coverage Rate of 95 Percent Confidence Intervals, and Statistics about Estimated Distance Measure over Repeated Sampling

	Raking						GREG_Main					
	Relative Bias	Bias Ratio	Coverage Rate of 95% Confidence Intervals	Distance Measure	% samples with extreme distance measure		Relative Bias	Bias Ratio	Coverage Rate of 95% Confidence Intervals	Distance Measure	% samples with extreme distance measure	
Sample Size and Response Scenario	$RelBias(\hat{t}_{yw})\times 10^5$	$BiasRatio(\hat{t}_{yw})\times 10^2$		$E_p(D\hat{I}ST_{raking})$	$D\hat{I}ST_{raking_s} > 3.84$	$D\hat{I}ST_{raking_s} > 6.63$	$RelBias(\hat{t}_{yw})\times 10^5$	$BiasRatio(\hat{t}_{yw})\times 10^2$		$E_p(D\hat{I}ST_{GREG\_Main})$	$D\hat{I}ST_{GREG\_Main_s} > 3.84$	$D\hat{I}ST_{GREG\_Main_s} > 6.63$
SRS Sample Size n =8,000												
S01. $DIFF_{RR}$ =0.00, CPR_{RR} =1.00	2.9	5.8	95%	1.0	5%	1%	2.9	5.8	95%	1.0	5%	1%
S02. $DIFF_{RR}$ =0.00, CPR_{RR} =1.00	45.8	52.8	92%	1.2	7%	1%	44.5	51.3	91%	1.3	8%	2%
S03. $DIFF_{RR}$ =-0.41, CPR_{RR} =1.00	31.8	16.9	100%	0.4	0%	0%	1162.2	812.1	0%	72.3	100%	100%
S04. $DIFF_{RR}$ =0.00, CPR_{RR} =4.75	1712.3	522.0	0%	28.0	100%	100%	2982.1	1537.0	0%	257.9	100%	100%
S05. $DIFF_{RR}$ =0.00, CPR_{RR} =3.41	1314.2	668.7	0%	47.1	100%	100%	2444.8	1786.9	0%	347.7	100%	100%
S06. $DIFF_{RR}$ =0.00, CPR_{RR} =2.74	1078.7	625.5	0%	42.5	100%	100%	2074.0	1573.1	0%	272.2	100%	100%
S07. $DIFF_{RR}$ =0.00, CPR_{RR} =2.06	799.7	567.8	0%	36.0	100%	100%	1576.0	1319.9	0%	193.5	100%	100%
S08. $DIFF_{RR}$ =0.00, CPR_{RR} =1.50	462.2	441.7	0%	22.9	100%	100%	936.0	958.1	0%	104.3	100%	100%
S09. $DIFF_{RR}$ =0.00, CPR_{RR} =1.29	288.3	324.8	6%	12.7	97%	88%	593.7	690.5	0%	54.8	100%	100%
S10. $DIFF_{RR}$ =0.00, CPR_{RR} =1.11	101.6	130.6	77%	2.9	28%	10%	228.9	296.3	16%	11.2	88%	72%
S11. $DIFF_{RR}$ =0.00, CPR_{RR} =1.07	56.6	73.9	92%	1.6	11%	3%	143.7	188.5	52%	5.3	57%	29%
S12. $DIFF_{RR}$ =-0.24, CPR_{RR} =0.90	-41.6	-21.2	100%	0.4	0%	0%	1013.4	567.7	0%	35.6	100%	100%
S13. $DIFF_{RR}$ =-0.24, CPR_{RR} =0.69	-394.3	-243.3	27%	6.6	74%	42%	334.8	221.3	39%	6.6	68%	44%
S14. $DIFF_{RR}$ =-0.24, CPR_{RR} =0.49	-770.9	-594.7	0%	37.5	100%	100%	-457.6	-355.2	6%	13.6	94%	82%
S15. $DIFF_{RR}$ =-0.24, CPR_{RR} =0.30	-1247.2	-931.1	0%	94.5	100%	100%	-1445.9	-1115.4	0%	136.9	100%	100%
S16. $DIFF_{RR}$ =-0.24, CPR_{RR} =0.15	-2036.1	-1056.1	0%	112.7	100%	100%	-2817.3	-1990.6	0%	423.8	100%	100%
S17. $DIFF_{RR}$ =-0.48, CPR_{RR} =0.04	-3089.0	-1567.3	0%	243.9	100%	100%	-4058.7	-5346.8	0%	3379.6	100%	100%
SRS Sample Size n =200												
S01. $DIFF_{RR}$ =0.00, CPR_{RR} =1.00	6.3	1.8	95%	1.0	5%	1%	6.1	1.7	95%	1.0	6%	1%
S02. $DIFF_{RR}$ =0.00, CPR_{RR} =1.00	56.3	9.0	97%	0.9	4%	1%	59.2	9.4	95%	1.0	5%	1%
S03. $DIFF_{RR}$ =-0.41, CPR_{RR} =1.00	42.3	-3.2	100%	0.3	0%	0%	1148.4	111.4	81%	2.1	19%	5%
S04. $DIFF_{RR}$ =0.00, CPR_{RR} =4.75	713.2	45.0	100%	0.4	0%	0%	1476.0	112.4	89%	1.8	12%	2%
S05. $DIFF_{RR}$ =0.00, CPR_{RR} =3.41	1179.2	88.0	100%	1.0	0%	0%	2191.6	234.3	35%	6.9	69%	43%
S06. $DIFF_{RR}$ =0.00, CPR_{RR} =2.74	1014.8	81.6	99%	0.9	1%	0%	1935.2	212.4	45%	5.9	59%	35%
S07. $DIFF_{RR}$ =0.00, CPR_{RR} =2.06	833.6	78.5	99%	0.9	1%	0%	1584.5	189.7	52%	5.0	49%	26%
S08. $DIFF_{RR}$ =0.00, CPR_{RR} =1.50	462.0	57.4	98%	0.9	2%	0%	935.2	131.6	75%	2.9	29%	10%
S09. $DIFF_{RR}$ =0.00, CPR_{RR} =1.29	283.3	40.9	97%	0.8	3%	0%	590.2	92.1	86%	1.9	16%	5%
S10. $DIFF_{RR}$ =0.00, CPR_{RR} =1.11	123.3	20.2	97%	0.8	3%	1%	252.9	43.0	94%	1.2	6%	2%
S11. $DIFF_{RR}$ =0.00, CPR_{RR} =1.07	81.8	14.1	96%	0.8	3%	1%	171.3	30.1	94%	1.0	6%	1%
S12. $DIFF_{RR}$ =-0.24, CPR_{RR} =0.90	-197.3	-19.2	99%	0.4	1%	0%	729.0	56.4	95%	1.1	4%	0%
S13. $DIFF_{RR}$ =-0.24, CPR_{RR} =0.69	-429.4	-43.8	95%	0.7	4%	1%	280.6	18.6	95%	0.9	4%	1%
S14. $DIFF_{RR}$ =-0.24, CPR_{RR} =0.49	-779.8	-86.6	89%	1.5	10%	4%	-491.7	-58.5	89%	1.5	10%	3%
S15. $DIFF_{RR}$ =-0.24, CPR_{RR} =0.30	-1249.4	-130.6	80%	2.4	20%	7%	-1428.3	-158.6	66%	3.7	38%	18%
S16. $DIFF_{RR}$ =-0.24, CPR_{RR} =0.15	-1642.0	-140.0	82%	2.3	15%	4%	-2235.1	-229.6	35%	6.1	69%	37%
S17. $DIFF_{RR}$ =-0.48, CPR_{RR} =0.04	-2476.1	-220.0	32%	4.9	68%	16%	-3437.8	-531.6	0%	30.9	100%	100%

4.4 Simulation Results Conditioning on Samples Grouped by Distance Measure

Both Chapter 3 and Section 4.3 show that under the outcome variable model Y_Additive_Interaction, the *average* relative bias and coverage rate of the 95 percent confidence intervals for raking over repeated sampling may be acceptable in some response scenarios with weak multiplicative interaction effort (e.g., S11 and S12). However, we have only one sample for a survey in the real world, so it is important to understand how a calibration estimator may perform for a given sample. That is, although a calibration estimator may perform reasonably well on average (over repeated sampling), we may still end up with an "unlucky" sample with poor performance in practice. In this section, we demonstrate the value of the proposed distance measure in helping identify such samples. We use the combination of the outcome variable model Y_Additive_Interaction and response scenario S11 to evaluate the properties of poststratification, raking, and GREG_Main conditioning on samples defined by the proposed distance measure. Given the fact that the coverage rates of the 95 percent confidence intervals are acceptable for most of the response scenarios with SRS sample size n=200 (discussed in Section 4.3), the simulation work in this section is based on only two SRS sample sizes: n=8,000 and n=2,000. The simulation setup is similar to that in Section 4.3 except that the total number of simulated samples is increased to 10,000 to warrant a large number of samples in each group defined by the range of distance measures. The 10,000 simulated samples are sorted by the estimated distance measure for the calibration estimator of interest and then partitioned into 20 groups. For example, to compare raking with poststratification, we first estimate the raking distance

measure for each sample, $\hat{DIST}_{raking_s}$. Then we sort all the 10,000 samples ascendingly by $\hat{DIST}_{raking_s}$ and partition the sorted samples into 20 groups with approximately 500 samples in each group. These groups are referred to as "0^{th}-5^{th} percentile distance measure group", "5^{th}-10^{th} percentile distance measure group", …, and "95^{th}-100^{th} percentile distance measure group". Finally, we examine the properties of the poststratification estimator and raking estimator for the samples in each of the 20 distance groups using the evaluation parameters described in Chapter 3. A similar procedure is used to compare GREG_Main and poststratification conditioning on samples defined by the GREG_Main distance measure $\hat{DIST}_{GREG\_Main_s}$. The results in this section warn us of the potential consequence of using an "almost appropriate but not quite appropriate" calibration estimator for a possibly "unlucky" sample in the real world.

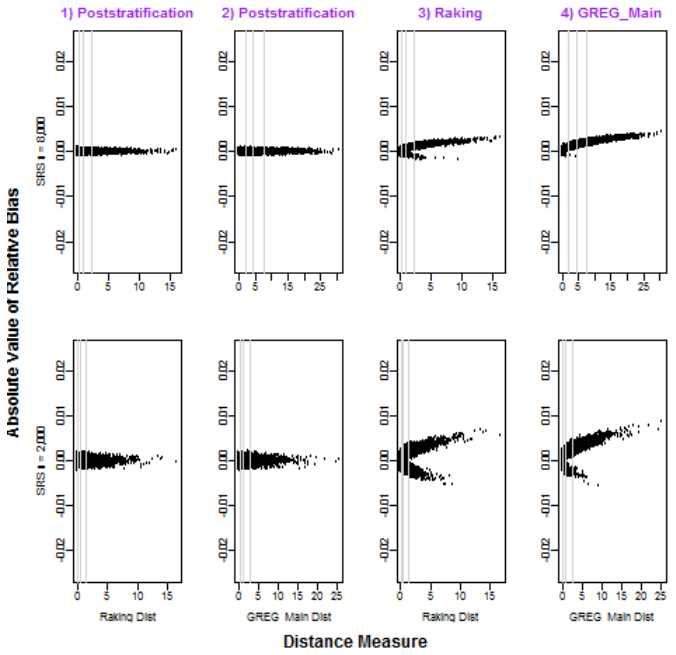
Figure 4.5 illustrates the properties of the three calibration estimators conditioning on samples grouped by the distance measure for the Y_Additive_Interaction and S11 combination. The response model S11 has $DIFF_{RR} = 0$ and $CPR_{RR} = 1.07$, meaning there is no additive interaction effect and almost no multiplicative interaction effect in the model. The top panel (a) of Figure 4.5 shows the relationship between the relative bias and the distance measure, and the bottom panel (b) shows the relationship between the bias ratio and the distance measure. Within Figure 4.5(a), the two rows from top to bottom correspond to SRS sample sizes n=8,000 and n=2,000 respectively. The four columns from left to right show the relationships: 1) between poststratification relative bias and raking distance measure; 2) between poststratification relative bias and

GREG_Main distance measure; 3) between relative bias and distance measure for raking; and 4) between relative and distance measure for GREG_Main. The three grey lines in each of the eight embedded subfigures indicate the upper limits of the distance measures for the 25^{th} percentile, 50^{th} percentile, and 75^{th} percentile, respectively, of the 10,000 sorted samples. The two left columns (both labeled "Poststratification") in Figure 4.5(a) demonstrate that the relative bias for poststratification remains very small (actually zero in theory) and is independent of the magnitude of the raking and GREG_Main distance measure. The two right columns (labeled "Raking" and "GREG_Main") show that for both raking and GREG_Main, there is a positive relationship between the absolute value of relative bias and the distance measure. For all the three estimators, the "bands" of the relative biases become wider as the SRS sample size decreases because the variances of the estimators become larger.

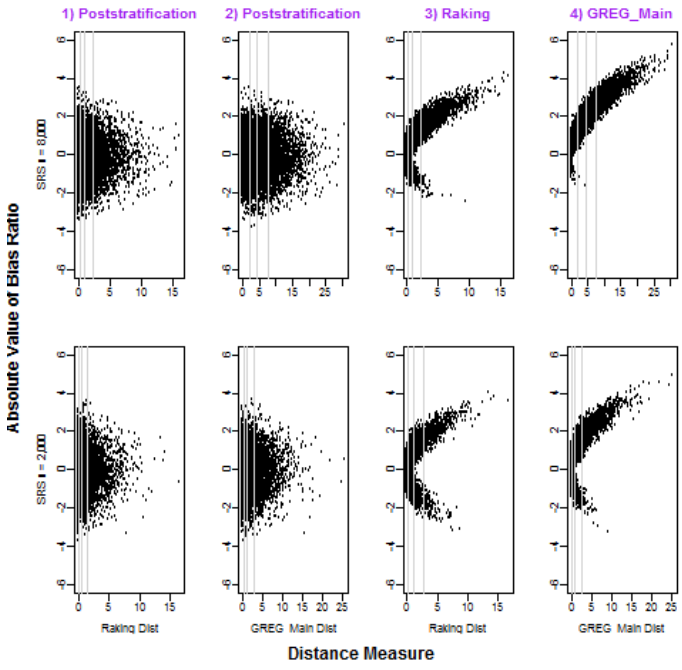
Figure 4.5(b) has the same structure as Figure 4.5(a) except that the y-axis for each subfigure in Figure 4.5(b) is bias ratio instead of relative bias. The bias ratios for poststratification generally fall within the range of [-2, 2] and are independent of the distance measures. For raking and GREG_Main, the absolute value of the bias ratio increases as the distance measure becomes larger. For example, Table 4.2 above shows that for the combination of Y_Additive_Interaciton, S11, and n=8,000, the average bias ratios over all the simulated samples are approximately 0.74 for raking and approximately 1.89 for GREG_Main. When the samples are sorted ascendingly by the estimated distance measure and divided into 20 distance groups, the average absolute values of the average bias ratios by distance group range from 0.04 (for the 0^{th}-5^{th}

percentile distance group) to 2.49 (for the 95th-100th distance group) for raking and from 0.11 to 3.80 for GREG_Main. As discussed earlier, the absolute value of the bias ratio increases as the SRS sample size increases because the larger sample size decreases sample variances. This is reflected in Figure 4.5(b) by the ranges of bias ratios for raking and GREG_Main being wider with n=8,000 than with n=2,000.

The coverage rates of the 95 percent confidence intervals by distance group under the Y_Additive_Interaction, n=8,000, and S11 combination are presented in Table 4.3 for raking and Table 4.4 for GREG_Main, respectively. Although the average coverage rate of the 95 percent confidence intervals over all the simulated samples is as good as approximately 92 percent for raking (see Table 3.4 in Chapter 3), Table 4.3 demonstrates that the coverage rates for raking vary substantially by the distance measure. For samples in the 0th to 60th percentile distance groups, the average coverage rates of the 95 percent confidence intervals are 100 percent (i.e., over coverage). However, the coverage rates drop to under 84 percent for the samples in the 80th to 100th percentile distance groups (corresponding to $\hat{DIST}_{raking_s} > 2.76$). If a survey practitioner happens to obtain a sample from the 95th-100th percentile distance group (corresponding to $\hat{DIST}_{raking_s} > 5.56$), then the coverage rate of the 95 percent confidence intervals is only 9 percent. In contrast, the coverage rates of the 95 percent confidence intervals for poststratification are essentially independent of the groups of samples defined by the raking distance measure.



(a) Relationship between Relative Bias and Distance Measure



(b) Relationship between Bias Ratio and Distance Measure

Figure 4.5 Properties of Poststratification, Raking, and GREG_Main Conditioning on Samples Grouped by Distance Measure under Outcome Model Y_Additive_Interaction and Response Model S11

Table 4.3 Properties of Raking and Poststratification Conditioning on Estimated Raking Distance Measure under Outcome Model Y_Additive_Interaction, SRS n=8,000, and Response Model S11

Distance Group Based on $\hat{DIST}_{raking_s}$	Range of $\hat{DIST}_{raking_s}$	Mean of $\hat{DIST}_{raking_s}$	Coverage Rate of 95% Confidence Intervals	
			Poststratification	Raking
0th – 5th percentile	$(6.23\times10^{-9},0.00784]$	0.00	96%	100%
5th – 10th percentile	(0.00784,0.034]	0.02	96%	100%
10th – 5th percentile	(0.034,0.0749]	0.05	95%	100%
15th – 20th percentile	(0.0749,0.138]	0.10	95%	100%
20th – 25th percentile	(0.138,0.215]	0.17	97%	100%
25th – 30th percentile	(0.215,0.312]	0.26	95%	100%
30th – 35th percentile	(0.312,0.432]	0.37	95%	100%
35th – 40th percentile	(0.432,0.556]	0.49	95%	100%
40th – 45th percentile	(0.556,0.699]	0.63	95%	100%
45th – 50th percentile	(0.699,0.879]	0.79	95%	100%
50th – 55th percentile	(0.879,1.09]	0.98	94%	100%
55th – 60th percentile	(1.09,1.32]	1.20	95%	100%
60th – 65th percentile	(1.32,1.57]	1.45	95%	99%
65th – 70th percentile	(1.57,1.9]	1.74	96%	99%
70th – 75th percentile	(1.9,2.27]	2.08	95%	96%
75th – 80th percentile	(2.27,2.76]	2.51	95%	92%
80th – 85th percentile	(2.76,3.33]	3.03	94%	84%
85th – 90th percentile	(3.33,4.18]	3.71	95%	71%
90th – 95th percentile	(4.18,5.56]	4.81	95%	46%
95th – 100th percentile	(5.56,16.3]	7.60	96%	9%

Table 4.4 shows the coverage rates of the 95 percent confidence intervals for GREG_Main and poststratification for the various groups defined by the GREG_Main distance measure. The average coverage rate of the 95 percent confidence intervals over the 10,000 simulated samples is only approximately 52 percent for GREG_Main (see Table 3.4 in Chapter 3). In Table 4.4, the coverage rates become unacceptable for the samples in the 45th to 100th percentile distance groups (corresponding to $\hat{DIST}_{GREG\_Main_s} > 3.86$) due to the combined effect of a biased estimator (so the confidence interval is centered at a wrong point) and very small variance associated with large sample size (so the confidence interval is very narrow).

Table 4.4 Properties of GREG_Main and Poststratification Conditioning on Estimated GREG_Main Distance Measure under Outcome Model Y_Additive_Interaction, SRS n=8,000, and Response Model S11

Distance Group Based on $\hat{DIST}_{GREG\_Main_s}$	Range of $\hat{DIST}_{GREG\_Main_s}$	Mean of $\hat{DIST}_{GREG\_Main_s}$	Coverage Rate of 95% Confidence Intervals	
			Poststratification	GREG_Main
0th – 5th percentile	$(2.43\times10^{-7}, 0.231]$	0.08	97%	100%
5th – 10th percentile	(0.231,0.667]	0.44	96%	100%
10th – 5th percentile	(0.667,1.13]	0.90	94%	100%
15th – 20th percentile	(1.13,1.58]	1.35	95%	100%
20th – 25th percentile	(1.58,2]	1.79	96%	99%
25th – 30th percentile	(2,2.5]	2.24	94%	96%
30th – 35th percentile	(2.5,2.92]	2.71	96%	90%
35th – 40th percentile	(2.92,3.39]	3.15	96%	86%
40th – 45th percentile	(3.39,3.86]	3.63	96%	72%
45th – 50th percentile	(3.86,4.39]	4.13	94%	61%
50th – 55th percentile	(4.39,4.96]	4.67	95%	46%
55th – 60th percentile	(4.96,5.52]	5.25	94%	35%
60th – 65th percentile	(5.52,6.12]	5.82	98%	21%
65th – 70th percentile	(6.12,6.8]	6.46	93%	14%
70th – 75th percentile	(6.8,7.58]	7.20	96%	6%
75th – 80th percentile	(7.58,8.51]	8.03	95%	3%
80th – 85th percentile	(8.51,9.68]	9.06	95%	1%
85th – 90th percentile	(9.68,11.2]	10.41	94%	0%
90th – 95th percentile	(11.2,13.6]	12.26	97%	0%
95th – 100th percentile	(13.6,30.8]	16.94	94%	0%

4.5 Conclusions and Limitations

Chapter 3 may give the readers the impression that raking should be a good calibration estimator even for the outcome variable model Y_Additive_Interaction as long as the multiplicative interaction term in the response model is weak. Such a conclusion is only based on the average properties of the estimator over repeated sampling, and thus can be misleading. This is because in the real world, a survey practitioner can usually obtain only one sample and all the outcome measures must be estimated from this sample.

This chapter shows that choosing an "almost appropriate but not quite appropriate" estimator can be detrimental to the bias ratio and the coverage rate of the 95 percent confidence intervals for some "unlucky" samples, especially when the sample size is large (assuming that the same calibration weighting strategy is used regardless of the sample size). The distance measure we propose can help identify such "unlucky" samples to some extent. Through both theoretical development and simulation work, we prove that the proposed distance measure follows Chi-square probability distribution under the null hypothesis that the expected values of the estimated cell counts equal the cell benchmark controls. The proposed distance measure can be estimated from an achieved sample, and then compared to the critical values in a Chi-square distribution table to determine whether it is "too large". On the one hand, we need to emphasize that the outcome variable model is the most critical factor and a large distance measure does not necessarily indicate significant bias for raking or GREG_Main. On the other hand, a real-world survey usually contains multiple key outcome measures, and it is often unlikely that the interaction term exists in none of the key outcome variable models. If the estimated distance measure is "too large", then it is a warning sign of potential bias for raking or GREG_Main due to excluding a significant interaction term during calibration.

Finally, we need to point out that the variance term $\text{var}\left(\hat{N}_{11}^{w}\right)$ in (4.18) is estimated from each simulated sample, so the conclusions in Sections 4.3 and 4.4 may depend on the accuracy of the variance estimation method implemented in the "calibrate" function of the R Survey package. To check the validity of these conclusions, the empirical

variance of $\hat{N}_{11}$ is calculated over all the simulated samples and compared to the estimated variance for some response scenarios. It is found that for raking, the estimated variance tends to be noticeably larger than the empirical variance under the response models with strong multiplicative interaction effect, making the estimated distance measure smaller than the true value. Despite this limitation, the conclusions about the relationships between the strength of the interaction effect in the response model, the distance measure, and the bias and bias ratio of the calibration estimator under the Y_Additive_Interaction model still hold. We plan to further investigate the variance estimation issue for raking in Chapter 5.

Chapter 5. Comparison of Alternative Variance Estimators for Raking

Several evaluation measures presented in Chapters 3 and 4 involve estimated variances using the "calibrate" function in the R Survey package. For example, the estimated variance from the sample, $var(\hat{t}_{yw})$, is used to obtain the 95 percent confidence interval for each simulated sample in Chapter 3. This approach (instead of using the empirical variance from all the simulation samples) is chosen because in practice, only one sample can be obtained for a survey and the variance has to be estimated from the sample. The "calibrate" function in the R Survey package estimates the variance of a linear substitute that is equivalent to the product of the calibrated weight and a residual calculated from a linear model of the outcome variable on a vector of auxiliary variables. For raking, the residual is based on a main effects model with the covariates being indictors for the raking categories of each dimension. The limitation of using the estimated variance from the sample is that the results rely on the accuracy of the variance estimation method implemented in the "calibrate" function.

In Chapter 4, the denominator of the distance measure shown in (4.18), $var(\hat{N}_{ij}^{w})$, is also computed from each simulation sample, so the conclusions about the distance measure may depend on the accuracy of the variance estimation method as well. During the validity check (described in Section 4.5 of Chapter 4), it is found that for raking, the estimated variance tends to be significantly larger than the empirical variance for the response scenarios with strong multiplicative interaction effect (see more details in Section 5.6 of this chapter). Although the bias in the estimated variance does not change

the general conclusions in Chapter 4, it motivates us to further investigate the variance estimation issue for raking.

5.1 Background and General Research Method

Since iteration is needed to solve the calibration equations for raking, survey practitioners often approximate the variance of the estimated total $\hat{t}_{yrk}$ by the variance of the "converged" estimator, i.e., the hypothetical estimator arising from an infinite number of iterations, represented by $\text{var}(\sum_r w_i y_i)$, where w_i is the "converged" weights (Deville, Särndal, and Sautory 1993). In practice, a linear model, $y_k = \mathbf{B}\mathbf{x}_k + \varepsilon_k,\ \varepsilon_k \sim iid\ N(0,\ \sigma^2)$, is fitted for the outcome variable y on a vector of auxiliary variables $\mathbf{x}$. For raking, the linear model is a main effects model with the covariates being indictors for the raking categories of each dimension. A linearization variance estimator is obtained by approximating $\text{var}(\sum_r w_i y_i)$ by $\text{var}(\sum_r d_i \hat{z}_i)$ for a "linearized variable" z_i, where $\hat{z}_i = (y_i - \hat{\mathbf{B}}\mathbf{x}_i) f_i$, with f_i being the weighting adjustment factor applied to the basic design weight d_i when weighting (Deville 1999, D'Arrigo and Skinner 2010). Several choices of the factor f_i are available and discussed later in this chapter. D'Arrigo and Skinner (2010) define alternative forms of linearization variance estimators for an estimated total via different choices of weights applied to not only the residuals but also the estimated regression coefficients used in calculating the residuals. Their empirical work results in two conclusions. First, the

approach that weights residuals by the basic design weight can be severely biased in the presence of nonresponse. In contrast, the approach that weights residuals by the calibrated weight tends to display much less bias. Second, varying the choice of weights used to construct the regression coefficients has little impact. In the D'Arrigo and Skinner (2010) framework, however, the simulation is based on a few selected variables from the British Labor Force Survey and German Survey of Income and Expenditure. It is unclear what models may govern the outcome variables (i.e., whether there are strong interaction effects in the outcome models and/or whether the outcome models have very strong explanatory power). Although response models are discussed in their work, there is no explicit manipulation of the strength of the multiplicative interaction term in the response model. We know from Chapter 3 that both the outcome model and the response model may affect the performance of a raking estimator. Now the question is whether and how these models may impact the performance of a variance estimator for raking. The existing literature does not provide a clear answer to this question.

Given the perceived bias of the linearization variance estimator for raking in Chapter 4, it is worthwhile to evaluate how alternative variance estimators for raking may perform in the presence of nonresponse under different outcome models and response models. We will specify the outcome models and response models explicitly to show the impacts of these models on the performance of the variance estimators. We will also vary the R-squared values of the outcome variable models to help us understand how the results may hold in practice.

One challenge in the attempt to obtain a variance estimator for raking is that the raking ratio estimator does not have a closed form solution. Because of this, it is unclear how to obtain an analytical solution for the linearization variance estimator. Therefore our approach is to use a simulation to obtain the empirical approximation to the distribution of the variance of a raking estimator. During the simulation study, we repeatedly draw a sample, rake, and compute an estimate and estimated variances using the variance estimators under evaluation. Then for each variance estimator, we compute the mean (across simulation iterations) of the variance estimates, and then estimate the empirical bias of the variance estimator by comparing the mean of the variance estimates to the empirical variance of the estimates.

5.2 Variance Estimators under Evaluation

Shao (1996) and Wolter (2007) provide detailed discussions of both replication and linearization approximation methods used for variance estimation from sample surveys. In this research, we first re-evaluate the properties of the four linearization variance estimators proposed by D'Arrigo and Skinner (2010), and then examine the performance of a replication variance estimator for raking.

5.2.1 Four Linearization Variance Estimators

D'Arrigo and Skinner (2010) show that the linearization variance estimator for a raking estimator for a total $\hat{t}_{yrk}$ can be expressed as

$$\text{var}(\hat{t}_{yrk}) \approx \text{var}(\sum_r d_i \hat{z}_i) \tag{5.1}$$

where $\hat{z}_i = (y_i - \hat{\mathbf{B}}\mathbf{x}_i) f_i$ is treated as a fixed variable. That is, the variance of the raking estimator $\hat{t}_{yrk}$ is approximately equal to the variance of $\sum_r d_i \left(y_i - \hat{\mathbf{B}}\mathbf{x}_i \right) f_i$. d_i is the basic design weights. $\hat{\mathbf{B}}$ is the vector for the regression coefficients in the weighted regression model for obtaining the GREG estimator. f_i is the weighting adjustment factor applied to the basic design weight d_i when weighting the residuals $e_i = y_i - \hat{\mathbf{B}}\mathbf{x}_i$ from the regression model. Therefore, the variance of $\hat{t}_{yrk}$ depends on not only the variance of the residuals e_i, but also the weighting adjustment factor f_i.

A number of choices of $\hat{\mathbf{B}}$ and f_i are discussed in D'Arrigo and Skinner (2010). Two options are considered for $\hat{\mathbf{B}}$ depending on what weights are used in the regression model:

1) $\hat{\mathbf{B}}^{bwt} = \left(\sum_r d_i y_i \mathbf{x}_i^{\mathrm{T}}\right)\left(\sum_r d_i \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right)^{-1}$ when the regression model uses base weights. The corresponding residual from the regression model is $e_i^{bwt} = y_i - \hat{\mathbf{B}}^{bwt}\mathbf{x}_i$.
2) $\hat{\mathbf{B}}^{rkwt} = \left(\sum_r w_i y_i \mathbf{x}_i^{\mathrm{T}}\right)\left(\sum_r w_i \mathbf{x}_i \mathbf{x}_i^{\mathrm{T}}\right)^{-1}$ when the regression model uses raked weights $w_i = d_i F(\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\lambda}})$. The corresponding residual from the regression model is $e_i^{rkwt} = y_i - \hat{\mathbf{B}}^{rkwt}\mathbf{x}_i$.

The weighting adjustment factor f_i determines how the residuals are weighted. Two natural choices are:

1) base-weighted residuals, where $f_i = 1$, and

2) calibration-weighted residuals, where $f_i = F(\mathbf{x}_i^{\mathrm{T}}\hat{\boldsymbol{\lambda}}) = w_i / d_i$.

In summary, the variance of a raking ratio estimator can be estimated as

$$\mathrm{var}(\hat{t}_{yrk}) = \mathrm{var}(\sum_r d_i f_i e_i) = \frac{n_r}{n_r - 1}\sum_{i\in r}\left(d_i f_i e_i - \frac{1}{n_r}\sum\nolimits_{i'\in r} d_{i'} f_{i'} e_{i'} \right)^2 \qquad (5.2)$$

where

$$f_i = \begin{cases} 1 \\ w_i / d_i \end{cases}$$

$$e_i = \begin{cases} y_i - \hat{\mathbf{B}}^{bwt}\mathbf{x}_i \\ y_i - \hat{\mathbf{B}}^{rkwt}\mathbf{x}_i \end{cases}$$

As part of this research, we will obtain the mean of the variance estimates (over all the simulation iterations) for raking using each of the following four linearization variance estimators discussed in D'Arrigo and Skinner (2000). These estimators are based on different choices for $\hat{\mathbf{B}}$ and f_i, as summarized in Table 5.1. Among these four estimators, "BWT.Residual_RKWT.Regression" is probably the least intuitive one, so we include it mainly for completeness.

Table 5.1 Four Linearization Variance Estimators and Their Labels

Description of how weights are used in variance estimator	Choice for f_i	Choice for $\hat{\mathbf{B}}$	Label for Easy Reference
Base weights to weight up residuals & base weights to obtain regression coefficients	1	$\hat{\mathbf{B}}^{bwt}$	BWT.Residual_BWT.Regression
Base weights to weight up residuals & raked weights to obtain regression coefficients	1	$\hat{\mathbf{B}}^{rkwt}$	BWT.Residual_RKWT.Regression
Raked weights to weight up residuals & base weights to obtain regression coefficients	w_i/d_i	$\hat{\mathbf{B}}^{bwt}$	RKWT.Residual_BWT.Regression
Raked weights to weight up residuals & raked weights to obtain regression coefficients	w_i/d_i	$\hat{\mathbf{B}}^{rkwt}$	RKWT.Residual_RKWT.Regression

5.2.2 Replication Variance Estimator

Replication variance estimation consists of repeatedly calculating estimates for subgroups of the full sample and then computing the variance among these "replicate" estimates. One main advantage of the replication method is that it provides a simple way to account for adjustments that are made in weighting. By separately computing the weighting adjustments for each replicate, it is possible to reflect the effect of variability of weight adjustments in the estimates of variance. Replication also has some disadvantages. For example, the method is computationally intensive and, in the case of the jackknife, inappropriate for quantile estimation.

The key motivation for considering the replication method is that the raking ratio estimator does not have a closed form solution, so that the linearization method of variance estimation may not correctly account for all sources of variation in an estimator. A good alternative may be to use the replication method to approximate the variance.

The Jackknife 1 (JK1) method is appropriate for our simulation study because the sample design (described in Chapter 3) involves no explicit stratification. To implement the JK1 method for raking, we first form replicates that are random subsets of equal or nearly equal size, with each subset resembling the full sample. Then raking is performed separately on the full sample as well as on each replicate, and the estimate of interest is calculated from the full sample and each replicate. Finally, the variation between the replicate estimates and the full-sample estimate is used to estimate the variance for the full sample. Assuming that the finite population correction factor can be ignored, the JK1 variance estimator for an estimated total (using a raking estimator) takes the form

$$\text{var}(\hat{t}_{yrk}) = \frac{G-1}{G}\sum_{g=1}^{G}(\hat{t}_{yrk(g)} - \hat{t}_{yrk})^2 \qquad (5.3)$$

where $\hat{t}_{yrk}$ is the full-sample estimate, $\hat{t}_{yrk(g)}$ is the estimate of t_y based on the observations included in the g-th replicate, and G is the total number of replicates formed.

5.3 Simulation Setup

5.3.1 Simulation Scenarios

The simulation study aims to compare the properties of several alternative variance estimators for raking for the estimate of a finite population total under different outcome

variable models and response scenarios. The scope of the study is described in Section 3.3 of Chapter 3. The SRS sample size is 8,000 for the simulation conducted in this chapter. The simulation scenarios are determined by the combination of three factors.

First, there are two outcome variable models: Y_Main as specified in (3.12) versus Y_Additive_Interaction, as specified in (3.13). The Y_Main model contains only main effect terms, while the Y_Additive_Interaction model contains a non-zero additive interaction term in addition to the main effects.

Second, the predictive power of each outcome variable model is varied as in the sensitivity analysis in Chapter 3. The R-squared value of the model is either close to one (i.e., the high R-squared setup) or approximately 0.65 (i.e., the medium R-squared setup).

Third, the strength of the multiplicative interaction effect in the response model is varied because the simulation results in Chapter 3 shows that it is the multiplicative interaction effect (not the additive interaction effect) in the response model that affects the performance of the raking estimator under the Y_Additive_Interaction outcome variable model. For the evaluation of the variance estimators, we choose only seven of the 17 response scenarios from Table 3.3 because the replication method is computationally intensive. These seven response scenarios still represent a gradual change of the strength of the multiple interaction effect, with the CPR_{RR} ranging from 0.04 to 4.75.

The total number of simulation scenarios is 2×2×7=28. Table 5.2 summarizes these 28 scenarios defined by four outcome variable models and seven response models.

Table 5.2 Simulation Scenarios for Comparing Variance Estimators

Scenario		Model Parameters
Outcome Variable Model	Y_Int with R^2=0.9979	$\mu_Y = 1000,\ \boldsymbol{\alpha}_Y = (\alpha_{Y1},\ \alpha_{Y2}) = (-200,\ 300),\ \boldsymbol{\beta}_Y = (\beta_{Y1},\ \beta_{Y2}) = (-100,\ 150),$ $\boldsymbol{\gamma}_Y = (\gamma_{Y11}, \gamma_{Y12}, \gamma_{Y21}, \gamma_{Y22}) = (100,\ 300,\ 700,\ 1200),\ \varepsilon_{Yijk} \sim N(0, 900)$
	Y_Main with R^2=0.9886	$\mu_Y = 1000,\ \boldsymbol{\alpha}_Y = (\alpha_{Y1},\ \alpha_{Y2}) = (-200,\ 300),\ \boldsymbol{\beta}_Y = (\beta_{Y1},\ \beta_{Y2}) = (-100,\ 150),$ $\varepsilon_{Yijk} \sim N(0, 900)$
	Y_Int with R^2=0.6348	$\mu_Y = 1000,\ \boldsymbol{\alpha}_Y = (\alpha_{Y1},\ \alpha_{Y2}) = (-200,\ 300),\ \boldsymbol{\beta}_Y = (\beta_{Y1},\ \beta_{Y2}) = (-100,\ 150),$ $\boldsymbol{\gamma}_Y = (\gamma_{Y11}, \gamma_{Y12}, \gamma_{Y21}, \gamma_{Y22}) = (100,\ 300,\ 700,\ 1200),\ \varepsilon_{Yijk} \sim N(0, 250000)$
	Y_Main with R^2=0.6813	$\mu_Y = 1000,\ \boldsymbol{\alpha}_Y = (\alpha_{Y1},\ \alpha_{Y2}) = (-200,\ 300),\ \boldsymbol{\beta}_Y = (\beta_{Y1},\ \beta_{Y2}) = (-100,\ 150),$ $\varepsilon_{Yijk} \sim N(0, 40000)$
Response Model	S04.CPR=4.75	R_{11}=0.12, R_{12}=0.48, R_{21}=0.02, R_{22}=0.38
	S06.CPR=2.74	R_{11}=0.28, R_{12}=0.92, R_{21}=0.08, R_{22}=0.72
	S08.CPR=1.50	R_{11}=0.40, R_{12}=0.80, R_{21}=0.20, R_{22}=0.60
	S11.CPR=1.07	R_{11}=0.56, R_{12}=0.64, R_{21}=0.36, R_{22}=0.44
	S15.CPR=0.30	R_{11}=0.09, R_{12}=0.21, R_{21}=0.41, R_{22}=0.29
	S16.CPR=0.15	R_{11}=0.04, R_{12}=0.26, R_{21}=0.36, R_{22}=0.34
	S17.CPR=0.04	R_{11}=0.02, R_{12}=0.58, R_{21}=0.66, R_{22}=0.74

5.3.2 Simulation Steps and Evaluation Criteria

We compare the properties of six variance estimators for raking, including the four linearization variance estimators in Table 5.1, the JK1 variance estimator (with 80 replicate groups), and the variance estimator implemented in the "calibrate" function of the R Survey package. The last one is referred to as "Lumley estimator"; it is included in the evaluation because the existing documentation does not provide much technical detail about the method. For each of the 28 simulation scenarios, the following steps are used to evaluate each of the variance estimators.

First, we repeatedly draw a sample, rake, and compute an estimate and estimated variance $\text{var}(\hat{t}_{yrk})$ using the variance estimators under evaluation. We also compute the estimated relative standard error, $RelSE(\hat{t}_{yrk}) = \sqrt{\text{var}(\hat{t}_{yrk})} \Big/ t_y$.

Second, we compute the mean (across the *S* simulation iterations) of the estimated relative standard errors. This mean is denoted $E_p(RelSE(\hat{t}_{yrk}))$.

Third, we compute the empirical relative standard error across the *S* simulated samples, $EmpRelSE(\hat{t}_{yw}) = \sqrt{EmpVar(\hat{t}_{yw})} \Big/ t_y = \sqrt{(1/S)\sum_{s=1}^{S}\left(\hat{t}_{yw_s} - E_p(\hat{t}_{yw})\right)^2} \Big/ t_y$, where $E_p(\hat{t}_{yw}) = (1/S)\sum_{s=1}^{S}\hat{t}_{yw_s}$, the average value of $\hat{t}_{yw_s}$ over repeated sampling.

Finally, we compare $E_p(RelSE(\hat{t}_{yrk}))$ against $EmpRelSE(\hat{t}_{yw})$ by calculating the ratio between $E_p(RelSE(\hat{t}_{yrk}))$ and $EmpRelSE(\hat{t}_{yw})$ (referred to as "ratio of estimated standard error versus empirical standard error).

5.4 Theoretical Development and Expected Results from Simulation

5.4.1 General Formula for Raking Variance

In the presence of nonresponse, the variance comes from three sources, including the outcome variable, response, and sampling distributions. We are interested in finding

$E_P E_R V_M(\hat{t}_{yrk})$, the expectation of the model variance for raking over both response distribution (E_R) and sampling distribution (E_P).

To obtain the variance formula for raking, a linear model, $y_k = \mathbf{B}\mathbf{x}_k + \varepsilon_k,\ \varepsilon_k \sim iid\ N(0,\ \sigma^2)$, is fitted for the outcome variable y on a vector of auxiliary variables $\mathbf{x}$. Then as shown in (5.1), the model variance for raking (conditioning on sampling and response) can be approximated by the variance of a linear substitute

$$\begin{aligned} & V_M(\hat{t}_{yrk} \,|\, s, r) \\ & \approx V_M\Big(\sum_{k\in r} d_k z_k \,|\, s, r\Big) \\ & = V_M\Big(\sum_{k\in r} d_k f_k \varepsilon_k\Big) \\ & = \sum_{k\in r} d_k^2 f_k^2 V_M(\varepsilon_k) \end{aligned} \tag{5.4}$$

where d_i is the basic design weight, f_k is the weighting adjustment factor applied to d_i, and $\varepsilon_k = y_k - \mathbf{B}\mathbf{x}_k$. Both d_i and f_k are treated as fixed variables.

Now define the response indicator $\delta_{Rk} = \begin{cases} 1 & \text{if response} \\ 0 & \text{if nonresponse} \end{cases}$. The expectation of the model variance for raking over response distribution (still conditioning on the initial random sample) is

$$\begin{aligned} & E_R\left(V_M(\hat{t}_{yrk} \,|\, s)\right) \\ & \approx E_R\left(\sum_{k\in s} \delta_{Rk} d_k^2 f_k^2 V_M(\varepsilon_k)\right) \\ & = \sum_{k\in s} R_k d_k^2 f_k^2 V_M(\varepsilon_k) \end{aligned} \tag{5.5}$$

where R_k is the response propensity for unit *k*.

Finally, define the sampling indicator $\delta_{pk} = \begin{cases} 1 & \text{if sampled} \\ 0 & \text{if not sampled} \end{cases}$. The expectation of the model variance for raking over response and random sampling distributions is

$$\begin{aligned} & E_P E_R V_M (\hat{t}_{yrk}) \\ & \approx E_P \left(\sum_{k \in U} \delta_{pk} R_k d_k^2 f_k^2 V_M (\varepsilon_k) \right) \\ & = \sum_{k \in U} \frac{1}{d_k} R_k d_k^2 f_k^2 V_M (\varepsilon_k) \\ & = \sum_{k \in U} R_k d_k f_k^2 V_M (\varepsilon_k) \end{aligned} \tag{5.6}$$

This is the general formula for the variance for raking.

5.4.2 Variance Estimator for a Special Situation When Raking Is Unbiased

In Chapter 3, we prove that if $F_{ij} = R_{ij}^{-1}$ (i.e., the raking adjustment factor in cell *ij* is the inverse of the cell response probability), then $\hat{t}_{yrk}$ is unbiased across the outcome variable, response, and repeated sampling distributions. Note that this is a sufficient condition, but not necessary condition, for $E_P E_R E_M \left(\hat{t}_{yrk} - t_{yU} \right) = 0$.

In our simulation setup, all the units *k* in cell *ij* have the same response propensity. Also, when raking is converged, all the units *k* in cell *ij* also have the same weighting adjustment factor. Therefore a sufficient condition for the raking estimator to be

unbiased would be $f_k = R_k^{-1}$, which is essentially Condition C in D'Arrigo and Skinner (2010).

When $f_k = R_k^{-1}$, we can simplify (5.6) into two forms

$$E_P E_R V_M(\hat{t}_{yrk}) \approx \sum_{k \in U} d_k f_k V_M(\varepsilon_k) \tag{5.7}$$

or

$$E_P E_R V_M(\hat{t}_{yrk}) \approx \sum_{k \in U} \frac{1}{\pi_k R_k} V_M(\varepsilon_k) \tag{5.8}$$

where $\pi_k R_k$ is the product of selection probability and response propensity. That is, $\pi_k R_k$ represents the probability of the unit being observed.

In practice, f_i and $V_M(\varepsilon_k)$ can be estimated from the responding sample, so (5.7) can be used to estimate the variance for raking *under a sufficient condition in which the raking estimator is unbiased* (as discussed in Section 5.4.2). Under this sufficient condition, the approach in (5.7) is consistent with the RKWT.Residual_BWT.Regression approach and the RKWT.Residual_RKWT.Regression approach (see Table 5.1) in D'Arrigo and Skinner (2010).

There are two remaining questions. First, when the sufficient condition for raking to be unbiased is not satisfied or when the raking estimator is biased, how do the four linearization variance estimators in D'Arrigo and Skinner (2010) perform? Second, although we know that $V_M(\varepsilon_k)$ can be estimated using e_i^2, it is unclear whether basic

design weights or raked weights should be used in the regression model for obtaining the residual e_i. The simulation results can help shed light on these two questions.

5.5 Simulation Results

Table 5.3 compares the estimated relative standard errors using different variance estimation methods. The first column shows all the simulation scenarios defined by different outcome variable model specifications and response models with varying strength of multiplicative interaction effect. The second column is for the empirical relative standard error. The remaining columns show the ratio of the estimated relative standard error versus the empirical relative standard error for each of the four linearization variance estimators in D'Arrigo and Skinner (2010), the Lumley result using the "calibrate" function in the R Survey package, and the result using JK1 replication method with 80 replicate groups, respectively. Some cells are shown in color font to help us identify and explain the data pattern. Recall that when the outcome variable model contains a significant interaction term, the raking estimator $\hat{t}_{yrk}$ is unbiased only in the response scenarios with CPR_{RR} being close to 1. Several conclusions can be drawn from Table 5.3.

First, the BWT.Residual_BWT.Regression estimates and BWT.Residual_RKWT.Regression estimates are much smaller than the estimates in the "Empirical" column. The underestimation is due to the basic design weights not weighting the sum in (5.2) to a high enough level to account for nonresponse. As a

result, using the basic design weights to weight up the squared residuals from the regression model (regardless of what weights are used in the regression model to obtain the coefficients) results in underestimating the variance for an estimated total. For example, the ratio of BWT.Residual_BWT.Regression to the empirical relative standard error for the outcome model "Y_Int with R-squared = 0.9979" and response model S04 combination is only 0.09.

Second, under the outcome models "Y_Main with R-squared = 0.9886", "Y_Int with R-squared = 0.6348", and "Y_Main with R-squared = 0.6813", the RKWT.Residual_BWT.Regression estimates (using basic design weights to obtain regression coefficients and raked weights to weight up the residuals from the regression model) align well with the "Empirical" estimates (shown in purple font). That is, the RKWT.Residual_BWT.Regression approach performs well under two types of outcome variable models: (1) when the outcome variable model contains only the main effect covariates; or (2) when the outcome variable model contains both main effect and interaction terms, but the overall explanatory power of the model is not close to being perfect. In contrast, using raked weights in the regression model to obtain residuals (the RKWT.Residual_RKWT.Regression estimates) leads to over-estimated relative standard errors under these outcome variable models unless the response model contains almost no multiplicative interaction effect (shown in green font).

Third, under the outcome variable model "Y_Int with R-squared = 0.9979", both RKWT.Residual_BWT.Regression and RKWT.Residual_RKWT.Regression are biased

variance estimators except for the response model S11 with CPR_{RR}=1.07 (shown in red font). That is, unless the response model contains almost no multiplicative interaction effect, none of the linearization variance estimators in D'Arrigo and Skinner (2010) performs well when the outcome variable model contains strong interaction effect and the model has almost perfect prediction power.

Fourth, for the scenarios in which the RKWT.Residual_BWT.Regression variance estimator and RKWT.Residual_RKWT.Regression variance estimator are biased (shown in green font and red font), the magnitude of the bias seems to be positively correlated with the strength of the multiplicative interaction effect in the response model.

Fifth, the "Lumley" column shows the estimates using the "calibrate" function in the R Survey package. Our simulation results show that Lumley estimates are consistent with the RKWGT.Residual_BWGT.Regression approach.

Finally, the replication method clearly outperforms all the linearization variance estimation methods in D'Arrigo and Skinner (2010) in the scenarios that the raking estimator is biased. The JK1 relative standard error aligns well with the empirical relative standard error regardless of the outcome variable model and response model. However, despite the unbiased variance estimator using JK1, the confidence intervals do not cover at the correct rate when the raking estimator $\hat{t}_{yrk}$ is biased. This is because the confidence intervals tend to center at the wrong place due to the bias of the point estimator.

Table 5.3 Comparison of Estimated Relative Standard Errors Using Different Variance Estimation Methods for SRS sample size *n*=8,000

Outcome Variable Model and Response Model	Ratio of Estimated Relative Standard Error versus Empirical Relative Standard Error						
	Empirical Relative Standard Error x 10^4	Linearization Method in D'Arrigo and Skinner (2010)				Lumley	JK1
		BWT.Residual_ BWT.Regression	BWT.Residual_ RKWT.Regression	RKWT.Residual_ BWT.Regression	RKWT.Residual_ RKWT.Regression		
Y_Int with R-squared = 0.9979							
S04: R_{11}=0.12, R_{12}=0.48, R_{21}=0.02, R_{22}=0.38, CPR=4.75	17.26	0.09	0.15	2.17	2.83	2.18	1.10
S06: R_{11}=0.28, R_{12}=0.92, R_{21}=0.08, R_{22}=0.72, CPR=2.74	10.80	0.25	0.35	1.80	1.97	1.80	1.02
S08: R_{11}=0.40, R_{12}=0.80, R_{21}=0.20, R_{22}=0.60, CPR=1.50	8.57	0.40	0.45	1.35	1.30	1.35	1.02
S11: R_{11}=0.56, R_{12}=0.64, R_{21}=0.36, R_{22}=0.44, CPR=1.07	7.62	0.50	0.51	1.11	1.09	1.11	1.04
S15: R_{11}=0.09, R_{12}=0.21, R_{21}=0.41, R_{22}=0.29, CPR=0.30	11.11	0.22	0.24	1.35	1.31	1.35	1.06
S16: R_{11}=0.04, R_{12}=0.26, R_{21}=0.36, R_{22}=0.34, CPR=0.15	12.80	0.16	0.19	1.69	1.68	1.70	1.01
S17: R_{11}=0.02, R_{12}=0.58, R_{21}=0.66, R_{22}=0.74, CPR=0.04	10.59	0.19	0.28	2.10	2.81	2.10	1.04
Y_Main with R-squared = 0.9886							
S04: R_{11}=0.12, R_{12}=0.48, R_{21}=0.02, R_{22}=0.38, CPR=4.75	10.03	0.15	0.15	1.01	2.27	1.01	1.01
S06: R_{11}=0.28, R_{12}=0.92, R_{21}=0.08, R_{22}=0.72, CPR=2.74	6.09	0.36	0.36	1.01	1.79	1.01	1.00
S08: R_{11}=0.40, R_{12}=0.80, R_{21}=0.20, R_{22}=0.60, CPR=1.50	5.08	0.43	0.43	0.98	1.23	0.98	0.98
S11: R_{11}=0.56, R_{12}=0.64, R_{21}=0.36, R_{22}=0.44, CPR=1.07	4.29	0.52	0.52	1.05	1.10	1.05	1.06
S15: R_{11}=0.09, R_{12}=0.21, R_{21}=0.41, R_{22}=0.29, CPR=0.30	6.26	0.25	0.25	1.10	1.32	1.10	1.10
S16: R_{11}=0.04, R_{12}=0.26, R_{21}=0.36, R_{22}=0.34, CPR=0.15	7.19	0.22	0.22	1.03	1.46	1.03	1.04
S17: R_{11}=0.02, R_{12}=0.58, R_{21}=0.66, R_{22}=0.74, CPR=0.04	5.91	0.37	0.37	1.01	1.99	1.01	1.01
Y_Int with R-squared = 0.6348							
S04: R_{11}=0.12, R_{12}=0.48, R_{21}=0.02, R_{22}=0.38, CPR=4.75	109.16	0.15	0.16	1.07	2.31	1.07	1.02
S06: R_{11}=0.28, R_{12}=0.92, R_{21}=0.08, R_{22}=0.72, CPR=2.74	66.92	0.36	0.36	1.04	1.80	1.04	1.00
S08: R_{11}=0.40, R_{12}=0.80, R_{21}=0.20, R_{22}=0.60, CPR=1.50	56.26	0.43	0.43	0.98	1.22	0.98	0.97
S11: R_{11}=0.56, R_{12}=0.64, R_{21}=0.36, R_{22}=0.44, CPR=1.07	47.08	0.52	0.52	1.06	1.11	1.06	1.06
S15: R_{11}=0.09, R_{12}=0.21, R_{21}=0.41, R_{22}=0.29, CPR=0.30	68.79	0.25	0.25	1.11	1.32	1.11	1.10
S16: R_{11}=0.04, R_{12}=0.26, R_{21}=0.36, R_{22}=0.34, CPR=0.15	79.85	0.21	0.21	1.05	1.46	1.05	1.03
S17: R_{11}=0.02, R_{12}=0.58, R_{21}=0.66, R_{22}=0.74, CPR=0.04	65.75	0.36	0.36	1.03	1.99	1.03	1.00
Y_Main with R-squared = 0.6813							
S04: R_{11}=0.12, R_{12}=0.48, R_{21}=0.02, R_{22}=0.38, CPR=4.75	66.92	0.15	0.15	1.01	2.27	1.01	1.01
S06: R_{11}=0.28, R_{12}=0.92, R_{21}=0.08, R_{22}=0.72, CPR=2.74	40.63	0.36	0.36	1.01	1.79	1.01	1.00
S08: R_{11}=0.40, R_{12}=0.80, R_{21}=0.20, R_{22}=0.60, CPR=1.50	33.88	0.43	0.43	0.98	1.23	0.98	0.98
S11: R_{11}=0.56, R_{12}=0.64, R_{21}=0.36, R_{22}=0.44, CPR=1.07	28.60	0.52	0.52	1.05	1.10	1.05	1.06
S15: R_{11}=0.09, R_{12}=0.21, R_{21}=0.41, R_{22}=0.29, CPR=0.30	41.75	0.25	0.25	1.10	1.32	1.10	1.10
S16: R_{11}=0.04, R_{12}=0.26, R_{21}=0.36, R_{22}=0.34, CPR=0.15	47.98	0.22	0.22	1.03	1.46	1.03	1.04
S17: R_{11}=0.02, R_{12}=0.58, R_{21}=0.66, R_{22}=0.74, CPR=0.04	39.39	0.37	0.37	1.01	1.99	1.01	1.01

To illustrate the data distribution, we plot the ratio of the estimated (relative) standard error versus the empirical (relative) standard error, $E_p(RelSE(\hat{t}_{yrk}))/EmpRelSE(\hat{t}_{yw})$, for RKWGT.Residual_BWGT.Regression and RKWGT.Residual_RKWGT.Regression under each of the 28 simulation scenarios. Figure 5.1 shows the distribution of these ratios grouped by the outcome variable model and reveal three patterns.

First, for the response model S11 with CPR_{RR}=1.07 (when the raking estimator is almost unbiased regardless of the outcome variable model), the ratio of the estimated standard error versus the empirical standard error is close to 1 regardless of the outcome variable model.

Second, for the outcome variable models "Y_Main with R-squared = 0.9886", "Y_Int with R-squared = 0.6348", and "Y_Main with R-squared = 0.6813", all the ratios for RKWT.Residual_BWT.Regression are close to 1 while the ratio for RKWT.Residual_RKWT.Regression increases as the CPR value moves away from 1 to either 0.04 or 4.75.

Finally, for the outcome variable model "Y_Int with R-squared = 0.9979", the ratio for RKWT.Residual_BWT.Regression and the ratio for RKWT.Residual_RKWT.Regression both increase as the CPR value moves away from 1 to either 0.04 or 4.75.

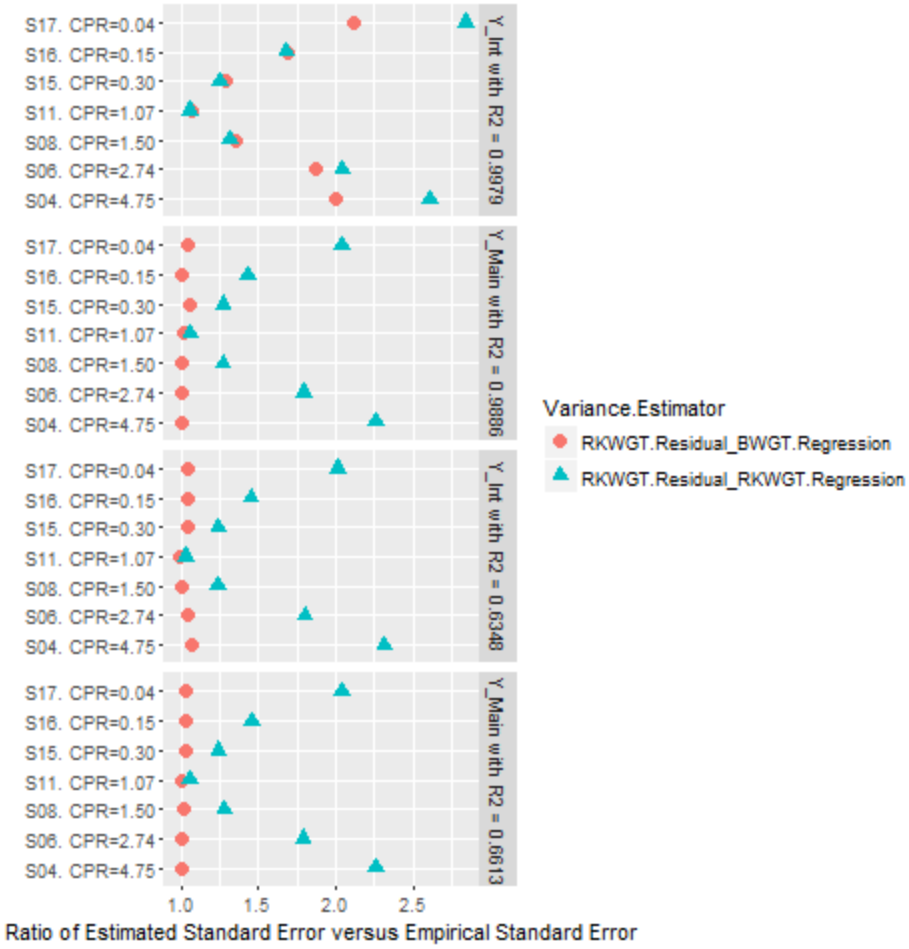


Figure 5.1 Ratio of Estimated (Relative) Standard Error versus Empirical (Relative) Standard Error for RKWGT.Residual_BWGT.Regression and RKWGT.Residual_RKWGT.Regression under Different Outcome Variable Models and Response Models

The results in Valliant, Dorfman, and Royall (2000, Section 5.6) on model-based variance estimation are relevant to the results for "Y_Int with R-squared = 0.9979". The outcome model has almost perfect explanatory power and contains a substantively and statistically significant interaction term. However, the raking estimator implicitly fits a *Y*-model with main effects only, which is a misspecified model. In this case, a variance estimator based on squared residuals from a misspecified model is expected to have two properties. First, the variance estimator is expected to overestimate the variance of the

raking estimator, but still be the same order of magnitude as the variance. Second, the MSE of the raking estimator is expected to be of a higher order of magnitude than the variance estimator due to the bias of $\hat{t}_{yrk}$.

The results in Table 5.3 and the plots in Figure 5.1 are consistent with the first point above. Both RKWGT.Residual_BWGT.Regression and RKWGT.Residual_RKWGT.Regression produce standard error estimates that are larger than the empirical standard error, although, as noted above, the degree of overestimation is substantially larger for the latter.

Table 5.4 shows the ratio of the estimated (relative) standard error versus the square root of empirical (relative) MSE, $E_p(RelSE(\hat{t}_{yrk}))/RelRMSE(\hat{t}_{yw})$, for RKWGT.Residual_BWGT.Regression and RKWGT.Residual_RKWGT.Regression under the outcome model "Y_Int with R-squared = 0.9979" and various response models. The only response scenario with these ratios larger than one is S11. This is because when the CPR_{RR} is close to one (i.e., CPR_{RR}=1.07 for S11), the bias of the raking estimator is negligible, so the empirical MSE is approximately the same as the empirical variance. For all the other rows, the data pattern is consistent with the second point above. The estimated variance based on RKWGT.Residual_BWGT.Regression and RKWGT.Residual_RKWGT.Regression are less than the empirical MSE. That is, overestimating the standard error is not sufficient to produce good estimates of the actual MSE because the MSE has a higher order than the variance due to the bias of the raking estimator.

Table 5.4 Ratio of Estimated (Relative) Standard Error versus Square Root of Empirical (Relative) MSE for RKWGT.Residual_BWGT.Regression and RKWGT.Residual_RKWGT.Regression under Outcome Model "Y_Int with R-squared = 0.9979" and Various Response Models

Y_Int with R-squared = 0.9979	$E_p(RelSE(\hat{t}_{yrk}))/RelRMSE(\hat{t}_{yw})$	
	RKWGT.Residual_ BWGT.Regression	RKWGT.Residual_ RKWGT.Regression
S04: R_{11}=0.12, R_{12}=0.48, R_{21}=0.02, R_{22}=0.38, CPR=4.75	0.22	0.28
S06: R_{11}=0.28, R_{12}=0.92, R_{21}=0.08, R_{22}=0.72, CPR=2.74	0.18	0.20
S08: R_{11}=0.40, R_{12}=0.80, R_{21}=0.20, R_{22}=0.60, CPR=1.50	0.25	0.24
S11: R_{11}=0.56, R_{12}=0.64, R_{21}=0.36, R_{22}=0.44, CPR=1.07	1.14	1.13
S15: R_{11}=0.09, R_{12}=0.21, R_{21}=0.41, R_{22}=0.29, CPR=0.30	0.12	0.12
S16: R_{11}=0.04, R_{12}=0.26, R_{21}=0.36, R_{22}=0.34, CPR=0.15	0.11	0.11
S17: R_{11}=0.02, R_{12}=0.58, R_{21}=0.66, R_{22}=0.74, CPR=0.04	0.07	0.10

D'Arrigo and Skinner (2010) conclude that both RKWT.Residual_BWT.Regression and RKWT.Residual_RKWT.Regression are nearly unbiased estimators. However, their research does not explicitly investigate the impact of the outcome variable model or the strength of the multiplicative interaction term in the response model. The conclusions from our simulation study can help refine those in D'Arrigo and Skinner (2010). It is interesting that using the raked weights in the regression model to obtain regression coefficient is actually proposed by Deville and Särndal (1992, equation 3.4). A similar idea is also reflected in formula (1.21). D'Arrigo and Skinner (2000) explain that this approach may be more practical than using the basic design weights to compute $\hat{\mathbf{B}}$ because the users of survey data files usually have access to the raked weights, but not the basic design weights. However, the linearization variance estimator RKWT.Residual_RKWT.Regression is biased in all our simulation scenarios except when CPR_{RR} is approximately 1 (i.e., there is almost no multiplicative interaction effect in the response model). This has two indications for the survey organizations in practice. First, serious consideration should be given to producing replicate weights. Second, the

basic design weights should be included in the public use file to facilitate the correct implementation of the linearization variance estimation method.

5.6 Re-Examination of Conclusions about Raking in Chapters 3 and 4

In this chapter, our simulation results show that under some outcome variable model and response model combinations, all of the linearization variance estimators in D'Arrigo and Skinner (2010) perform poorly, one of which (i.e., RKWGT.Residual_BWGT.Regression) appears to be the variance estimation method implemented in the "calibrate" function of the R Survey package (referred to as "Lumley method"). On the other hand, the variance estimator using the JK1 replication method is unbiased regardless of the outcome variable model and response model.

Some conclusions about raking in Chapters 3 and 4 are based on the Lumley variance estimation method. In this section, we re-examine those results by using the JK1 replication method to estimate the variance for raking. The measures of interest include relative standard error, bias ratio, coverage rate of the 95 percent confidence intervals, and distance measure. Due to the intensive computational work involved in the replication method, we select only a limited number of scenarios for re-evaluation: "Y_Main with R^2=0.9886" and "Y_Int with R^2=0.9979" combined with the response models shown in Table 5.2 for the SRS sample size n=8,000. Two new finite populations are generated using the outcome model parameters described in Section 3.4 of Chapter 3. Then simulated samples are drawn from the new finite populations for this re-evaluation

study. Due to the variation across simulation iterations, it is normal that the results using the Lumley method in this Section may not be exactly the same as those in Tables 3.4 and Table 3.5 in Chapter 3.

Table 5.5 compares the evaluation measures involving variance estimation using the Lumley method to those using the JK1 replication method. The first several columns in Table 5.5 are about the relative standard error, bias ratio, and coverage rate of the 95 percent confidence intervals for the estimated total for the outcome variable $\hat{t}_y$. Whether there is any nonegligible difference between the Lumley method and the JK1 replication method depends on the outcome variable model and response model. Under "Y_Main with R^2=0.9886", the raking estimator is nearly unbiased regardless of the response model (as shown in Chapter 3). The Lumley method and JK1 replication method yield approximately equal estimates for the relative standard error, bias ratio, and effective confidence interval coverage rate. Under "Y_Int with R^2=0.9979", the magnitude of the bias of the raking estimator is positively correlated with the strength of the multiplicative interaction effect in the response model (as shown in Chapter 3). Table 5.5 shows that when the CPR_{RR} is away from 1 for the response model, the Lumley method tends to over-estimate the variance for raking. This makes the estimated relative standard error too big, the estimated bias ratio too small, and the estimated confidence interval too wide. Despite these inaccurate estimates due to the bias in the variance estimator for raking, all the conclusions about raking in Chapter 3 still hold in general.

The last three columns in Table 5.5 are about the estimated distance measure. The data patterns for "Y_Main with R^2=0.9886" and "Y_Int with R^2=0.9979" are essentially the same because the distance measure depends on only the estimated totals for the auxiliary variables, but not the outcome variable. When the CPR_{RR} is away from 1 for the response model, the estimated variances for the estimated cell counts using the Lumley method tend to be noticeably larger than the estimated variances using the JK1 replication method (which are close to the empirical variances). This makes the estimated distance measure using the Lumley method noticeably smaller than that using the JK1 replication method. For example, the ratio of the latter to the former is approximately 3.2 for the response scenario S16 with CPR_{RR} being 0.15. For the SRS sample size n=8,000, the estimated distance measures are larger than the critical value 3.84 for rejecting the null hypothesis (i.e., $\text{Prob}(0.004 < \chi^2(1) < 3.84) = 0.95$), regardless of the variance estimation method, for all the response models with CPR_{RR} being away from one. Thus, the Lumley results and JK1 results (despite their difference in the specific values for the estimated distance measure) are likely to lead to the same conclusions. For the SRS sample size n=200, however, the conclusions about the distance measure and potential bias for raking may be sensitive to the variance estimation method. For example, Table 4.2 in Chapter 4 shows that for n=200, the estimated distance measure using the Lumley method is approximately 2.3 for the response scenario S16, which is smaller than the critical value 3.84. However, if we do a crude adjustment by using the ratio between the JK1 method and Lumley method for the SRS sample size n=8,000 (which is 3.2 as described above), then the "corrected" distance measure should be approximately 7.3. This corrected distance measure is larger than the critical value 3.84, and thus can probably explain why

the effective coverage rate of the 95 percent confidence intervals is only 82 percent. Despite the limitation in Chapter 4 that is caused by the variance estimation method, the conclusions about the properties of the distance measure in Chapter 4 still hold. That is, the distance measure can help identify particular samples where the raking estimator is likely to be biased, and consequently, the confidence interval coverage is likely to be poor.

Table 5.5 Comparison of Some Evaluation Measures in Chapters 3 and 4 Using Lumley Method and JK1 Replication Method for SRS Sample Size n=8,000

Outcome Variable model	Response Scenario	Properties of Raking Estimator Depending on How $Var(\hat{t}_{yw})$ Is Estimated							Distance Measure Depending on How $Var(\hat{N}_{11}^{w})$ Is Estimated		
		Relative Standard Error $RelSE(\hat{t}_{yw})\times 10^4$			Bias Ratio $BiasRatio(\hat{t}_{yw})\times 10^2$		Coverage Rate of 95% Confidence Intervals		$E_p(D\hat{I}ST)$		
		Empirical	Lumley	JK1	Lumley	JK1	Lumley	JK1	Empirical	Lumley	JK1
Y_Main with R^2=0.9886	S04. CPR_{RR}=4.75	10.1	10.1	10.1	21.8	21.8	96%	94%	101.9	21.4	94.9
	S06. CPR_{RR}=2.74	6.1	6.1	6.1	23.6	23.5	95%	95%	133.4	36.0	124.4
	S08. CPR_{RR}=1.50	4.9	5.0	5.0	-17.5	-17.7	96%	95%	33.8	17.4	33.2
	S11. CPR_{RR}=1.07	4.4	4.5	4.5	5.9	6.0	96%	95%	2.0	1.6	1.9
	S15. CPR_{RR}=0.30	6.5	6.9	6.9	-29.7	-29.7	95%	95%	154.4	87.7	155.7
	S16. CPR_{RR}=0.15	7.3	7.4	7.4	-41.8	-41.8	94%	93%	310.1	98.0	307.4
	S17. CPR_{RR}=0.04	5.7	5.9	6.0	4.0	3.6	96%	96%	933.9	193.6	893.6
Y_Int with R^2=0.9979	S04. CPR_{RR}=4.75	18.6	37.5	19.0	457.7	907.5	0%	0%	95.1	21.3	93.1
	S06. CPR_{RR}=2.74	10.4	19.5	11.1	592.3	1048.0	0%	0%	138.7	36.0	125.3
	S08. CPR_{RR}=1.50	8.5	11.6	8.8	389.9	518.0	0%	0%	34.9	17.5	33.3
	S11. CPR_{RR}=1.07	7.9	8.4	8.0	84.2	90.0	89%	85%	1.9	1.7	2.0
	S15. CPR_{RR}=0.30	11.7	15.0	11.8	-901.7	-1153.3	0%	0%	146.7	88.3	155.6
	S16. CPR_{RR}=0.15	12.8	21.7	12.9	-968.5	-1634.2	0%	0%	306.9	97.4	307.7
	S17. CPR_{RR}=0.04	10.6	22.4	11.1	-1376.1	-2796.6	0%	0%	971.6	193.2	890.6

Chapter 6. Conclusions and Future Work

6.1 Conclusions

This dissertation investigates the properties of several widely used calibration estimators in the presence of nonresponse. In the purely sampling context, Deville & Särndal (1992) demonstrate that many alternative forms of calibration weighting are asymptotically equivalent, so the GREG estimator can be used to approximate some general calibration estimators with no closed-form solutions such as raking. Our research in this dissertation shows that this conclusion does not necessarily hold when nonresponse exists and single-step calibration weighting is used to reduce nonresponse bias. With nonresponse, the differences between poststratification, raking, and GREG_Main can be either substantive or negligible depending on the outcome variable model and response model, so it is important to examine these models to the extent possible when choosing the appropriate calibration estimator. First, the outcome variable model is the dominant factor. If a significant interaction effect is present in the outcome model and the overall predictive power of the model is very strong (with R-squared value being close to 1), then poststratification (which is comparable to a GREG model with interaction terms) outperforms raking and GREG_Main except in the special situation that the response model does not include a multiplicative interaction term, in which case raking performs almost equally well as poststratification. Second, raking preserves the multiplicative interaction effect that is internal in the data before calibration while GREG_Main does not, so raking tends to be less biased than GREG_Main when the response model

contains a strong multiplicative interaction term. Third, for a large sample, a small relative bias associated with an inappropriate calibration estimator can still lead to very poor coverage rate of the 95 percent confidence intervals. Finally, as the predictive power of the outcome variable model decreases, the advantage of poststratification over raking and GREG_Main in bias reduction becomes less substantial. Moreover, if the predictive power of the outcome model with the interaction term is not extremely high and poststratification involves some very small cell counts, then the MSE may be higher for poststratification than for raking and/or GREG_Main.

Our research also yields a proposed distance measure that can help gauge the potential bias of raking and GREG_Main for a given sample. The distance measure follows the Chi-square probability distribution when raking or GREG_Main is unbiased. In practice, the distance measure is computable as long as the classification and corresponding cell totals for the population are available. A large estimated distance measure is a warning sign of potential bias and poor confidence interval coverage for some variables in a survey due to omitting a significant interaction term in the calibration process.

The last part of our research is an empirical evaluation of several variance estimators for raking with nonresponse, including linearization and replication methods. Our simulation results refine the conclusions in D'Arrigo and Skinner (2010) by demonstrating the impact of outcome model and response model on the performance of several linearization variance estimators. We show that when raking is model-biased, none of the linearization variance estimators in D'Arrigo and Skinner (2010) is unbiased. In contrast,

the jackknife replication method performs well in variance estimation, although the confidence interval may still be centered in the wrong place if the point estimate is biased. Our research has two indications for the survey organizations in practice. First, serious consideration should be given on producing replicate weights. Second, the basic design weights should be included in the public use file to facilitate the correct implementation of the linearization variance estimation method.

6.2 Future Work

Our dissertation presents a comprehensive framework for comparing the various calibration estimators in the presence of nonresponse. We choose a limited scope for our empirical work (as described in Section 3.2 of Chapter 3) such that the results can clearly demonstrate the impact of outcome models and response models. The real-world surveys often involve complex sample design and calibration estimators based on an array of variables with multiple categories. Future improvement and extension to our work may include:

1. Empirical research on the settings that are more complicated than a 2×2 table. In such settings, in addition to $\chi^2_{(I-1)(J-1)}$ statistic, the Cramér's V may be used to measure association between variables with more than two categories.

2. Empirical investigation on how the power theory for the distance measure may work.

3. Theoretical and empirical work to show whether and why the choice of categories does not affect the value of the proposed distance measure under complex sample designs (because under a complex sample design, each cell estimate may have a different design effect). Chapter 4 shows that the choice of the cell does not matter for a 2×2 table, but further work is needed for tables with more categories.

4. Theoretical development for the raking variance estimators when the main effects outcome model does not hold.

5. Examination of domain estimators for total and mean.

Appendix A. Summary of Proofs in Deville & Särndal (1992)

Deville & Särndal (1992) consider a sequence of finite populations and sampling designs indexed by *n*, where *n* is the sample size (for a fixed-sized sampling design) or the expected sample size (for a random-sized sample design). The finite population size, *N*, tends to infinity with *n*. The calibration weight is calculated as $w_k = d_k F_k(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda})$, where $F_k(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda})$ is non-negative and convex with $F_k(0)=1$ and $F_k^{'}(0)>0$. Several assumptions are made about the auxiliary vector **x**: (i) lim $N^{-1}\mathbf{t}_x$ exists; (ii) $N^{-1}\left(\hat{\mathbf{t}}_{x\pi}-\mathbf{t}_x\right)=O_p(n^{-1/2})$; and (iii) $n^{1/2}N^{-1}(\hat{\mathbf{t}}_{x\pi}-\mathbf{t}_x)$ converges in distribution to the multinormal $N(\mathbf{0},\mathbf{A})$, where **A** can be reviewed as a matrix that describes an asymptotic effect of the sampling design used in the survey. Two more assumptions are added for proving Results 3-5: (iv) $\max\|\mathbf{x}_k\|=M<\infty$, where max is over *n* as well as over *k*; and (v) $\max F_k^{"}(0)=M^{'}<\infty$. All the distance functions given in Deville & Särndal (1992) satisfy these conditions.

Result 1. The calibration equation has a unique solution belonging to the open neighborhood of **0**, with probability tending to 1 as $n \to \infty$.

Proof: $G_k(w,d)$ is defined on an interval $D_k(d)$ containing d. $g_k(w,d)=\partial G_k(w,d)/\partial w$ maps $D_k(d)$ onto an interval $Im_k(d)$ in a one-to-one fashion. $w_k = d_k F_k(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda})$, where $F_k(\cdot)$ is the reciprocal mapping of $g_k(\cdot, d)$ that maps $Im_k(d)$ onto $D_k(d)$.

Equation (1.4) defines a function of $\boldsymbol{\lambda}$ on $C_n = \bigcap_{k \in U_n} \{\boldsymbol{\lambda} : \mathbf{x}_k^{\mathrm{T}} \boldsymbol{\lambda} \in Im_k(d_k)\}$, where $\cap$ is over $k \in U_n$, the finite population associated with the (expected) sample size *n*. The interior C_n^0 of C_n is an open convex set containing **0** for every *n*. Then $C^* = \bigcap_{n=1}^{\infty} C_n^0$ is convex, and we assume it is also open. Let $E_n(\cdot)$ denote expectation with respect to the sampling design indexed by *n*. For $\boldsymbol{\lambda} \in C^*$, $N^{-1} E_n \left\{ \Phi_s(\boldsymbol{\lambda}) \right\}$ is a well defined continuously differentiable function. Assumption (iii) is that $n^{1/2} N^{-1} (\hat{\mathbf{t}}_{x\pi} - \mathbf{t}_x)$ converges in distribution to the multinormal $N(\mathbf{0}, \mathbf{A})$. For equation (1.4) to hold, it is necessary that for $\boldsymbol{\lambda} \in C^*$, $N^{-1} E_n \left\{ \Phi_s(\boldsymbol{\lambda}) \right\}$ converges to a fixed function denoted Φ, and the convergence is uniform on every compact set in C^*. Let $\Phi_s'(\boldsymbol{\lambda}) = \partial \Phi_s(\boldsymbol{\lambda}) / \partial \boldsymbol{\lambda}$. We can obtain $N^{-1} \Phi_s(0) = 0$, $N^{-1} \Phi_s'(0) = N^{-1} \mathbf{T}_s, \Phi(0) = 0, \Phi'(0) = \mathbf{T} = \lim N^{-1} \sum_U \mathbf{x}_k \mathbf{x}_k^{\mathrm{T}}$.

Φ maps C^* onto an open neighborhood of **0** in R^J. Let *B* be a closed sphere with radius *r* contained in this neighborhood, and let *A* be the compact set $\Phi^{-1}(B)$. The inverse function Φ^{-1} is defined on *B*, continuous and continuously differentiable. $\left\| \Phi^{-1}(\mathbf{x}) \right\|$ is bounded on *B*. All functions $N^{-1} \Phi_s'(\boldsymbol{\lambda})$ are defined on C^* and therefore on *A*. Let P_n denote probability with respect to the sampling design indexed by *n*. For every $\varepsilon > 0$, $P_n(\left\| N^{-1} \Phi_s - \Phi \right\|_A < \varepsilon) \to 1$ when *n* increases.

Let

$$K = \max_{\mathbf{x}\in B} \left\| (\Phi^{-1})'(\mathbf{x}) \right\| \tag{A.1}$$

Let $\Phi_1 = N^{-1}\Phi_s$ for some functions verifying $\left\|\Phi_1 - \Phi\right\|_A \le \beta r$, $\left\|\Phi_1' - \Phi'\right\|_A \le \beta K$, with $0 < \beta < \frac{1}{2}$. The probability of this event tends to 1 as *n* increases.

Let B_1 be the sphere $\|\mathbf{x}\| < (1-\beta)r$ in R^J . Now Φ_1 maps the frontier of *A* onto the crown $(1-\beta)r \le \|\mathbf{x}\| \le (1+\beta)r$. Consequently, $\Phi_1(A)$ covers the sphere B_1. In other words, for every $\mathbf{x} \in B_1$, the equation $\Phi_1(\boldsymbol{\lambda}) = \mathbf{x}$ has a (unique) solution. Because $\left\|\Phi_1' - \Phi'\right\|_A \le \beta K$ for every $\boldsymbol{\lambda}$ in *C*, $\left(\Phi_1^{-1}\right)'(\mathbf{x})$ exists for every $\mathbf{x} \in B_1$. Moreover,

$$\left\|\Phi_1^{-1}(\mathbf{x})\right\| \le \|\mathbf{x}\| K (1-\beta)^{-1} \tag{A.2}$$

Conclusion: $N^{-1}\left(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}\right)$ belongs to B_1 with a probability tending to 1. $N^{-1}\Phi_s(\boldsymbol{\lambda})$ has an inverse function on B_1 with a probability tending to 1. Equation (1.5) can be written as $N^{-1}\Phi_s(\boldsymbol{\lambda}) = N^{-1}(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})$, which has a unique solution with probability tending to 1.

Result 2. Let $\boldsymbol{\lambda}_s$ be the solution to equation (1.5) if one exist; otherwise let $\boldsymbol{\lambda}_s$ be an arbitrary fixed value. $\boldsymbol{\lambda}_s = O_p(n^{-1/2})$, so $\boldsymbol{\lambda}_s$ tends to **0** in design probability.

Proof: Define $\mathbf{z} = N^{-1}\left(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}\right)$. Let $\boldsymbol{\lambda}_s = \left(N^{-1}\Phi_s\right)^{-1}(\mathbf{z})$ if $\mathbf{z}$ belongs to B_1; otherwise $\boldsymbol{\lambda}_s$ is arbitrarily defined. Since $N^{-1}\Phi_s(0) = 0$, we have $\boldsymbol{\lambda}_s - 0 = \left(N^{-1}\Phi_s\right)^{-1}(\mathbf{z}) - \left(N^{-1}\Phi_s\right)^{-1}(0)$, so

$$\|\boldsymbol{\lambda}_s\| \le \|\mathbf{z}\| K(1-\beta)^{-1} \tag{A.3}$$

Inequality (A.3) holds with probability tending to 1 when n increases. Since $\mathbf{z} = O_p(n^{-1/2})$, there exists a constant $K^{'}$ such that $P_n(\|\mathbf{z}\| \le K^{'} n^{-1/2}) \to 1$. Applying this to (A.3), we obtain $P_n(\|\boldsymbol{\lambda}_s\| \le K^{'} K(1-\beta)^{-1} n^{-1/2}) \to 1$, which implies $\boldsymbol{\lambda}_s = O_p(n^{-1/2})$.

Result 3. $\boldsymbol{\lambda}_s = \mathbf{T}_s^{-1}\left(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}\right) + O_p\left(n^{-1}\right)$.

Proof: Let the difference between the adjustment functions for a general calibration estimator and the GREG estimator be

$$\theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right) = F_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right) - (1 + q_k \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}) \tag{A.4}$$

where $1 + q_k \mathbf{x}_k^T \boldsymbol{\lambda}$ is the adjustment function for the GREG estimator.

The assumption is that $\theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right) = O\left(\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right)^2\right)$ holds uniformly, which is equivalent to the assumption that $F_k^{"}(0)$ is uniformly bounded. Thus$\theta\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right) = \max \theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right) = O\left(\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right)^2\right)$.

Otherwise, for any $\varepsilon > 0$, there exists K^* such that, for all k, $|\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}| < \varepsilon$ will imply that $\theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right) \le K^*\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right)^2$.

From (1.3) and (A.4), the calibration equation can be rewritten as

$$\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi} = \sum_s d_k \mathbf{x}_k \left\{ q_k \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda} + \theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right)\right\} \tag{A.5}$$

Multiplying both sides of (A.5) by $\mathbf{T}_s^{-1}$ and rearranging the terms, we obtain

$$\boldsymbol{\lambda}_s - \mathbf{T}_s^{-1}\left(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}\right) = -\mathbf{T}_s^{-1}\sum_s d_k \mathbf{x}_k \theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_s\right) \tag{A.6}$$

For $\boldsymbol{\lambda}_s$ sufficiently small,

$$\left\|\boldsymbol{\lambda}_s - \mathbf{T}_s^{-1}\left(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi}\right)\right\| \le \left\|(N^{-1}\mathbf{T}_s)^{-1}\right\| K^* \left\{ N^{-1}\sum_s d_k \left\|\mathbf{x}_k\right\|^3 \right\} \left\|\boldsymbol{\lambda}_s\right\|^2 \tag{A.7}$$

We know that $\left\|(N^{-1}\mathbf{T}_s)^{-1}\right\| = O_p(1)$, and assumption (ii) indicates $N^{-1}\sum_s d_k \left\|\mathbf{x}_k\right\|^3 = O_p(1)$. From Result 2, $\left\|\boldsymbol{\lambda}_s\right\|^2 = O_p\left(n^{-1}\right)$. So Result 3 follows.

Result 4. The calibration estimator given in (1.6) is design-consistent, and $N^{-1}\left(\hat{t}_{yw} - \hat{t}_{y\pi}\right) = O_p(n^{-1/2})$.

Proof: If equation (1.5) has a solution $\boldsymbol{\lambda}_s$, then from (A.4)

$$\hat{t}_{yw} - \hat{t}_{y\pi} = \sum_s d_k y_k \mathrm{F}_k(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_s) - \sum_s d_k y_k = \sum_s d_k y_k \{q_k \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_s + \theta_k(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_s)\} \tag{A.8}$$

Given the assumption that $\max\ \theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right) = O\left(\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right)^2\right)$, we obtain

$$\begin{aligned} & N^{-1}\left|\hat{t}_{yw} - \hat{t}_{y\pi}\right| \\ & \le N^{-1}\left|\sum_s d_k y_k q_k \mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}\right| + N^{-1}\left|\sum_s d_k y_k \theta_k(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda})\right| \\ & \le N^{-1}\left\{(\sum_s d_k q_k \left|y_k\right|\left\|\mathbf{x}_k\right\|)\left\|\boldsymbol{\lambda}\right\|\right\} + O_p\left(n^{-1}\right) \end{aligned} \tag{A.9}$$

where $N^{-1}\left\{\sum_s d_k q_k \left|y_k\right|\mathbf{x}_k\right\} = O_p(1)$ and $\boldsymbol{\lambda}_s = O_p(n^{-1/2})$.

Then, Result 4 follows.

Result 5. For any $F_k(\cdot)$ obeying the assumptions, $\hat{t}_{yw}$ given by (1.6) is asymptotically equivalent to the regression estimator given by (1.7), in the sense that $N^{-1}\left(\hat{t}_{yw} - \hat{t}_{yreg}\right) = O_p\left(n^{-1}\right)$.

From (1.6) and (A.4),

$$N^{-1}\hat{t}_{yw} = N^{-1}\hat{t}_{y\pi} + N^{-1}(\mathbf{t}_x - \hat{\mathbf{t}}_{x\pi})^{\mathrm{T}}\hat{\mathbf{B}}_s + O_p\left(n^{-1}\right) + N^{-1}\sum_s d_k y_k \theta_k\left(\mathbf{x}_k^{\mathrm{T}}\boldsymbol{\lambda}_s\right) \tag{A.10}$$

The first two terms of the right side equal $N^{-1}\hat{t}_{yreg}$ as given in equation (1.7). The last term is $O_p\left(n^{-1}\right)$. Therefore, $n^{1/2}N^{-1}\left(\hat{t}_{yw} - \hat{t}_{yreg}\right) = O_p\ (\ n^{-1/2}\ )$, with zero asymptotic variance.

Appendix B. R Programs and Functions for Simulation Work

B.1 A Program for Creating the Population, Conducting Simple Random Sampling, Respondent Sampling, and Calibration, and Obtaining Evaluation Measures

This program is used in Chapters 3 and 4.

```
library(sampling)
library(survey)
library(plyr)
require(MASS)

#############################################################################
#Function for generating population, control totals, response indicator
#############################################################################

pop.and.control <- function (seed, lambda, lambda_i, lambda_j,
              lambda_ij, yseed, ymu, yalpha, ybeta, ygamma, ysigma,
              rmeans) {

  # popcnt -- population count in each of the four cells
  # pop -- a "dataset" for the population
  # totals.xvar1xvar2 -- a 2*2 matrix showing the crosstab of
              xvar1*xvar2, xvar1 and xvar2 are both categorical
              variables
  # totals.xvar1 -- a vector showing the tab of xvar1
  # totals.xvar2 -- a vector showing the tab of xvar2

# Generate popcnt
  set.seed(seed)
  popcnt <- matrix(nrow = 2,ncol = 2)  # population counts in cells
  popcnt[1,1] <- rpois(n=1,
              lambda=lambda+lambda_i[1]+lambda_j[1]+lambda_ij[1,1])
  popcnt[1,2] <- rpois(n=1,
              lambda=lambda+lambda_i[1]+lambda_j[2]+lambda_ij[1,2])
  popcnt[2,1] <- rpois(n=1,
              lambda=lambda+lambda_i[2]+lambda_j[1]+lambda_ij[2,1])
  popcnt[2,2] <- rpois(n=1,
              lambda=lambda+lambda_i[2]+lambda_j[2]+lambda_ij[2,2])

# Generate pop
  pop <- matrix(nrow = sum(popcnt), ncol = 5) # dataset for population
  colnames(pop) <- c("xvar1", "xvar2", "xvar12", "y", "respflag")
  pop <- as.data.frame(pop)

  # xvar1 and xvar2 are both categorical variables
  # Values for xvar1
  pop[1:sum(popcnt[1,]), "xvar1"] <- 1
  pop[sum(popcnt[1,],1):sum(popcnt), "xvar1"] <- 2
  # Values for xvar2
  pop[1:popcnt[1,1], "xvar2"] <- 1
```

```
  pop[sum(popcnt[1,1],1):sum(popcnt[1,]), "xvar2"] <- 2
  pop[sum(popcnt[1,],1):sum(popcnt[1,],popcnt[2,1]), "xvar2"] <- 1
  pop[sum(popcnt[1,],popcnt[2,1],1):sum(popcnt), "xvar2"] <- 2

  # Create xvar12 for QC
  pop$xvar12 <- pop$xvar1*10 + pop$xvar2

  # Create x11 for obtainging variance of N11 later

  # For standardized distance measure later
  pop$x11 <-ifelse(pop$xvar12==11, 1, 0)

  # Turn xvar1 and xvar2 into factor vars to be used for calibration
  pop$xvar1 <- as.factor(pop$xvar1)
  pop$xvar2 <- as.factor(pop$xvar2)

  totals.xvar1xvar2 <- xtabs(~xvar1 + xvar2, data = pop)

  rk.control.xvar1<-data.frame(xvar1=c(1,2), Freq = c(sum(popcnt[1,]),
              sum(popcnt[2,])))
  rk.control.xvar2<-data.frame(xvar2=c(1,2), Freq = c(sum(popcnt[,1]),
              sum(popcnt[,2])))

  # greg.control.xvar1 <- table(pop$xvar1)
  # greg.control.xvar2 <- table(pop$xvar2)

  # Generate values for y (ygamma vector index: 1=[1,1], 2=[1,2],
              3=[2,1], 4=[2,2])
  set.seed(yseed)
  pop[1:popcnt[1,1],"y"] <- rnorm(n = popcnt[1,1], mean = ymu +
              yalpha[1] + ybeta[1] + ygamma[1], sd = ysigma)
  pop[sum(popcnt[1,1],1):sum(popcnt[1,]),"y"] <- rnorm(n = popcnt[1,2],
              mean = ymu + yalpha[1] + ybeta[2] + ygamma[2], sd =
              ysigma)
  pop[sum(popcnt[1,],1):sum(popcnt[1,],popcnt[2,1]),"y"] <- rnorm(n =
              popcnt[2,1], mean = ymu + yalpha[2] + ybeta[1] +
              ygamma[3], sd = ysigma)
  pop[sum(popcnt[1,],popcnt[2,1],1):sum(popcnt),"y"] <- rnorm(n =
              popcnt[2,2], mean = ymu + yalpha[2] + ybeta[2] +
              ygamma[4], sd = ysigma)

  # Generate values for response flag (rmean index: 1=[1,1], 2=[1,2],
              3=[2,1], 4=[2,2])
  pop[1:popcnt[1,1], "respflag"] <- rbinom(n = popcnt[1,1], size=1,
              prob = rmeans[1])
  pop[sum(popcnt[1,1],1):sum(popcnt[1,]), "respflag"] <- rbinom(n =
              popcnt[1,2], size=1, prob = rmeans[2])
  pop[sum(popcnt[1,],1):sum(popcnt[1,],popcnt[2,1]),"respflag"] <-
              rbinom(n = popcnt[2,1], size=1, prob = rmeans[3])
  pop[sum(popcnt[1,],popcnt[2,1],1):sum(popcnt),"respflag"] <- rbinom(n
              = popcnt[2,2], size=1, prob = rmeans[4])
  return(list(pop=pop, totals.xvar1xvar2=totals.xvar1xvar2,
              rk.control.xvar1=rk.control.xvar1,
              rk.control.xvar2=rk.control.xvar2))
}
```

```
##############################################################################
# Functions for simple random sampling and for sampling of respondents
##############################################################################

srs.smp <- function(srsseed, popdata, n){
  srs.bad <- FALSE

  N <- nrow(popdata)
  s <- srswor(n, N)
  bwgt <- rep (N/n, n)
  f1 <- rep (n/N, n)

  srs.smp <- data.frame(popdata[s==1,], bwgt, f1)
  srs.totals <- xtabs(~xvar1 + xvar2, data = srs.smp)

  if (srs.totals[1, 1]<2 | srs.totals[1, 2]<2 | srs.totals[2, 1]<2 |
              srs.totals[2, 2]<2){
    srs.bad <- TRUE
  }
  return(list(srs.bad=srs.bad, srs.smp=srs.smp, srs.totals=srs.totals))
}

resp.smp <- function (srsdata){
  resp.bad <- FALSE

  resp.indic <- srsdata["respflag"] > 0
  resp.smp <- srsdata[resp.indic==1, ]

  resp.totals <- xtabs(~xvar1 + xvar2, data = resp.smp)

  if (resp.totals[1, 1]<2 | resp.totals[1, 2]<2 | resp.totals[2, 1]<2 |
              resp.totals[2, 2]<2){
    resp.bad <- TRUE
  }
  return(list(resp.bad=resp.bad, resp.smp=resp.smp,
              resp.totals=resp.totals))
}

##############################################################################
# Function for calibration and obtaining summary stats from each sample
##############################################################################

calib <- function(respinfo, popinfo, srsinfo){

  # Form design object
  dsgn <- svydesign(
    ids = ~0, # No cluster
    strata = NULL, # No strata
    fpc = ~f1,
    weights = ~bwgt,
    data = respinfo$resp.smp)

  # Calibration
  ps.dsgn <- postStratify(design = dsgn, strata = ~xvar1 + xvar2,
              population = popinfo$totals.xvar1xvar2, partial=TRUE)
  ps.wgt <- weights(ps.dsgn)
```

```
rk.dsgn.rake <- rake(design = dsgn, sample.margins =
            list(~xvar1,~xvar2), population.margins =
            list(popinfo$rk.control.xvar1, popinfo$rk.control.xvar2),
            control = list(maxit = 50))
rk.wgt.rake <- weights(rk.dsgn.rake)

rk.dsgn <- calibrate(design = dsgn, formula = ~xvar1 + xvar2,
            population = c('(Intercept)'=nrow(popinfo$pop),
            xvar12=sum(popinfo$totals.xvar1xvar2[2,]),
            xvar22=sum(popinfo$totals.xvar1xvar2[,2])),
            calfun="raking")
rk.wgt <- weights(rk.dsgn)

greg.dsgn <- calibrate(design = dsgn, formula = ~xvar1 + xvar2,
            population = c('(Intercept)'=nrow(popinfo$pop),
            xvar12=sum(popinfo$totals.xvar1xvar2[2,]),
            xvar22=sum(popinfo$totals.xvar1xvar2[,2])),
            calfun="linear")
greg.wgt <- weights(greg.dsgn)

# fit regression models and get R-squared
# Models including only main effects
ps.xx.main <- lm(y ~ xvar1+xvar2, data=respinfo$resp.smp, weights =
            ps.wgt)
ps.yy.main <- summary(ps.xx.main)
ps.R2.main <- ps.yy.main$r.squared

rk.xx.main <- lm(y ~ xvar1+xvar2, data=respinfo$resp.smp, weights =
            rk.wgt)
rk.yy.main <- summary(rk.xx.main)
rk.R2.main <- rk.yy.main$r.squared

greg.xx.main <- lm(y ~ xvar1+xvar2, data=respinfo$resp.smp, weights =
            greg.wgt)
greg.yy.main <- summary(greg.xx.main)
greg.R2.main <- greg.yy.main$r.squared

# Model including main effect and interaction
ps.xx.int <- lm(y ~ xvar1*xvar2, data=respinfo$resp.smp, weights =
            ps.wgt)
ps.yy.int <- summary(ps.xx.int)
ps.R2.int <- ps.yy.int$r.squared

rk.xx.int <- lm(y ~ xvar1*xvar2, data=respinfo$resp.smp, weights =
            rk.wgt)
rk.yy.int <- summary(rk.xx.int)
rk.R2.int <- rk.yy.int$r.squared

greg.xx.int <- lm(y ~ xvar1*xvar2, data=respinfo$resp.smp, weights =
            greg.wgt)
greg.yy.int <- summary(greg.xx.int)
greg.R2.int <- greg.yy.int$r.squared

 # SSW page 266 variance estimate formula for PS
Nc <- as.vector(popinfo$totals.xvar1xvar2)
nc <- as.vector(respinfo$resp.totals)
sc2data <- aggregate(y~xvar1*xvar2, data=respinfo$resp.smp, FUN=var)
```

```
sc2 <- as.vector(sc2data$y)
ps.total.se.SSW <- sqrt(sum((Nc*Nc/nc)*sc2))    # estimate using SSW

# ybarc <- as.vector(aggregate(y~xvar1*xvar2, data=respinfo$resp.smp,
            FUN=mean))

#####################
# Summary statistics
#####################

# Estimate for outcome variable, associated SE, CI, and CI coverage
# Mean
pop.mean <- mean(popinfo$pop[,"y"])
resp.mean <- svymean(~y, dsgn)
ps.mean <- svymean(~y, ps.dsgn)
rk.mean <- svymean(~y, rk.dsgn)
rk.mean.rake <- svymean(~y, rk.dsgn.rake)
greg.mean <- svymean(~y, greg.dsgn)

resp.mean.se <- SE(svymean(~y, dsgn))
ps.mean.se <- SE(svymean(~y, ps.dsgn))
rk.mean.se <- SE(svymean(~y, rk.dsgn))
rk.mean.se.rake <- SE(svymean(~y, rk.dsgn.rake))
greg.mean.se <- SE(svymean(~y, greg.dsgn))

ps.mean.CI <- confint(svymean(~y, ps.dsgn))
rk.mean.CI <- confint(svymean(~y, rk.dsgn))
greg.mean.CI <- confint(svymean(~y, greg.dsgn))

ps.mean.CI.coverage
        <- ifelse(ps.mean.CI[1]<=pop.mean & pop.mean<=ps.mean.CI[2],
            1, 0)
rk.mean.CI.coverage
        <- ifelse(rk.mean.CI[1]<=pop.mean & pop.mean<=rk.mean.CI[2],
            1, 0)
greg.mean.CI.coverage
        <- ifelse(greg.mean.CI[1]<=pop.mean &
            pop.mean<=greg.mean.CI[2], 1, 0)

# Total
pop.total <- sum(popinfo$pop[,"y"])
resp.total <- svytotal(~y, dsgn)
ps.total <- svytotal(~y, ps.dsgn)
rk.total <- svytotal(~y, rk.dsgn)
rk.total.rake <- svytotal(~y, rk.dsgn.rake)
greg.total <- svytotal(~y, greg.dsgn)

resp.total.se <- SE(svytotal(~y, dsgn))   # Lumley estimates of SEs.
ps.total.se <- SE(svytotal(~y, ps.dsgn))
rk.total.se <- SE(svytotal(~y, rk.dsgn))
rk.total.se.rake <- SE(svytotal(~y, rk.dsgn.rake))
greg.total.se <- SE(svytotal(~y, greg.dsgn))

ps.total.CI <- confint(svytotal(~y, ps.dsgn))
ps.total.CI.SSW <- c(ps.total-1.96*ps.total.se.SSW,
            ps.total+1.96*ps.total.se.SSW)
         # estimate using SSW
```

```
rk.total.CI <- confint(svytotal(~y, rk.dsgn))
greg.total.CI <- confint(svytotal(~y, greg.dsgn))

# CI coverage
ps.total.CI.coverage<- ifelse(ps.total.CI[1]<=pop.total &
            pop.total<=ps.total.CI[2], 1, 0)
ps.total.CI.coverage.SSW <- ifelse(ps.total.CI.SSW[1]<=pop.total &
            pop.total<=ps.total.CI.SSW[2], 1, 0)  # estimate using
            SSW
rk.total.CI.coverage <- ifelse(rk.total.CI[1]<=pop.total &
            pop.total<=rk.total.CI[2], 1, 0)
greg.total.CI.coverage <- ifelse(greg.total.CI[1]<=pop.total &
            pop.total<=greg.total.CI[2], 1, 0)

# Estimate N, Nr, Nc, Nrc as well as the difference from population
            truth
ps.Nrc <- svytable(~xvar1 + xvar2, ps.dsgn)
rk.Nrc <- svytable(~xvar1 + xvar2, rk.dsgn)
greg.Nrc <- svytable(~xvar1 + xvar2, greg.dsgn)

ps.Diff.Nrc = ps.Nrc - popinfo$totals.xvar1xvar2
rk.Diff.Nrc = rk.Nrc - popinfo$totals.xvar1xvar2
greg.Diff.Nrc = greg.Nrc - popinfo$totals.xvar1xvar2

# Odds ratios

pop.OR  <-
            (popinfo$totals.xvar1xvar2[1,1]*popinfo$totals.xvar1xvar2
            [2,2])/(popinfo$totals.xvar1xvar2[1,2]*popinfo$totals.xva
            r1xvar2[2,1])
resp.OR  <-
            (respinfo$resp.totals[1,1]*respinfo$resp.totals[2,2])/(re
            spinfo$resp.totals[1,2]*respinfo$resp.totals[2,1])

ps.OR  <- (ps.Nrc[1,1]*ps.Nrc[2,2])/(ps.Nrc[1,2]*ps.Nrc[2,1])
rk.OR  <- (rk.Nrc[1,1]*rk.Nrc[2,2])/(rk.Nrc[1,2]*rk.Nrc[2,1])
greg.OR  <-
            (greg.Nrc[1,1]*greg.Nrc[2,2])/(greg.Nrc[1,2]*greg.Nrc[2,1
            ])

# Distance measure for Chapter 3 (We experimented with this)
ps.Distance.Chap3 = sqrt(sum(ps.Diff.Nrc^2))
rk.Distance.Chap3 = sqrt(sum(rk.Diff.Nrc^2))
greg.Distance.Chap3 = sqrt(sum(greg.Diff.Nrc^2))

# Distance measure for Chapter 4 (This is what we propose)
ps.Distance.Chap4.est <- (ps.Diff.Nrc[1,1]/SE(svytotal(~x11,
            ps.dsgn)))^2
rk.Distance.Chap4.est <- (rk.Diff.Nrc[1,1]/SE(svytotal(~x11,
            rk.dsgn)))^2
greg.Distance.Chap4.est <- (greg.Diff.Nrc[1,1]/SE(svytotal(~x11,
            greg.dsgn)))^2

# vector to return, for estimates of means
results.mean <- vector(length=14)

results.mean[1] <- pop.mean
```

```
results.mean[2] <- resp.mean
results.mean[3] <- ps.mean
results.mean[4] <- greg.mean
results.mean[5] <- rk.mean
results.mean[6] <- rk.mean.rake

results.mean[7] <- resp.mean.se
results.mean[8] <- ps.mean.se
results.mean[9] <- greg.mean.se
results.mean[10] <- rk.mean.se
results.mean[11] <- rk.mean.se.rake

results.mean[12] <- ps.mean.CI.coverage
results.mean[13] <- greg.mean.CI.coverage
results.mean[14] <- rk.mean.CI.coverage

# vector to return, for estimates of totals
results.total <- vector(length=16)

results.total[1] <- pop.total
results.total[2] <- resp.total
results.total[3] <- ps.total
results.total[4] <- greg.total
results.total[5] <- rk.total
results.total[6] <- rk.total.rake

results.total[7] <- resp.total.se
results.total[8] <- ps.total.se
results.total[9] <- ps.total.se.SSW
results.total[10] <- greg.total.se
results.total[11] <- rk.total.se
results.total[12] <- rk.total.se.rake

results.total[13] <- ps.total.CI.coverage
results.total[14] <- ps.total.CI.coverage.SSW
results.total[15] <- greg.total.CI.coverage
results.total[16] <- rk.total.CI.coverage


# vector to return, for sample sizes, auxiliary info, and diff term
results.common <- vector(length=33)

results.common[1] <- respinfo$resp.totals[1,1]
results.common[2] <- respinfo$resp.totals[1,2]
results.common[3] <- respinfo$resp.totals[2,1]
results.common[4] <- respinfo$resp.totals[2,2]

results.common[5] <- ps.Distance.Chap3
results.common[6] <- rk.Distance.Chap3
results.common[7] <- greg.Distance.Chap3

results.common[8] <- ps.Distance.Chap4.est
results.common[9] <- rk.Distance.Chap4.est
results.common[10] <- greg.Distance.Chap4.est


results.common[11] <- pop.OR
```

```
  results.common[12] <- resp.OR
  results.common[13] <- ps.OR
  results.common[14] <- greg.OR
  results.common[15] <- rk.OR

  results.common[16] <- ps.Diff.Nrc[1,1]
  results.common[17] <- ps.Diff.Nrc[1,2]
  results.common[18] <- ps.Diff.Nrc[2,1]
  results.common[19] <- ps.Diff.Nrc[2,2]

  results.common[20] <- rk.Diff.Nrc[1,1]
  results.common[21] <- rk.Diff.Nrc[1,2]
  results.common[22] <- rk.Diff.Nrc[2,1]
  results.common[23] <- rk.Diff.Nrc[2,2]

  results.common[24] <- greg.Diff.Nrc[1,1]
  results.common[25] <- greg.Diff.Nrc[1,2]
  results.common[26] <- greg.Diff.Nrc[2,1]
  results.common[27] <- greg.Diff.Nrc[2,2]

  results.common[28] <- ps.R2.main
  results.common[29] <- rk.R2.main
  results.common[30] <- greg.R2.main

  results.common[31] <- ps.R2.int
  results.common[32] <- rk.R2.int
  results.common[33] <- greg.R2.int

  return (t(c(results.mean, results.total, results.common)))
}

#############################################################################
# Function for calling srs.smp, resp.smp, calib during each simulation
#############################################################################

srs.resp.calib <- function (benchmark, k, srs.size){

  S <- k      # number of good samples to keep
  s <- 1
  bad.smp <- 0

  # An empty matrix to store results

  rslt <- matrix(nrow=S, ncol=  sum(14, 16, 33))
  colnames(rslt) <- c("pop.mean",
                      "resp.mean",
                      "ps.mean",
                      "greg.mean",
                      "rk.mean",
                      "rk.mean.rake",

                      "resp.mean.se",
                      "ps.mean.se",
                      "greg.mean.se",
                      "rk.mean.se",
                      "rk.mean.se.rake",
```

```
"ps.mean.CI.coverage",
"greg.mean.CI.coverage",
"rk.mean.CI.coverage",

"pop.total",
"resp.total",
"ps.total",
"greg.total",
"rk.total",
"rk.total.rake",

"resp.total.se",
"ps.total.se",
"ps.total.se.SSW",
"greg.total.se",
"rk.total.se",
"rk.total.se.rake",

"ps.total.CI.coverage",
"ps.total.CI.coverage.SSW",
"greg.total.CI.coverage",
"rk.total.CI.coverage",

"respcnt11",
"respcnt12",
"respcnt21",
"respcnt22",

"ps.Distance.Chap3",
"rk.Distance.Chap3",
"greg.Distance.Chap3",

"ps.Distance.Chap4.est",
"rk.Distance.Chap4.est",
"greg.Distance.Chap4.est",

"pop.OR",
"resp.OR",
"ps.OR",
"greg.OR",
"rk.OR",

"ps.Diff.Nrc11",
"ps.Diff.Nrc21",
"ps.Diff.Nrc12",
"ps.Diff.Nrc22",

"rk.Diff.Nrc11",
"rk.Diff.Nrc21",
"rk.Diff.Nrc12",
"rk.Diff.Nrc22",

"greg.Diff.Nrc11",
"greg.Diff.Nrc21",
"greg.Diff.Nrc12",
"greg.Diff.Nrc22",
```

```
                      "ps.R2.main",
                      "rk.R2.main",
                      "greg.R2.main",
                      "ps.R2.int",
                      "rk.R2.int",
                      "greg.R2.int")
  while (s <= S){

  keep.sw <- TRUE

  # draw srs sample and respondent sample
  srssmp <- srs.smp(popdata=benchmark$pop, n=srs.size)
  if (srssmp$srs.bad==TRUE){
    bad.smp <- bad.smp + 1
    keep.sw <- FALSE
  }
  else {
    # assign respondent
    respsmp <- resp.smp (srsdata=srssmp$srs.smp)
  }

  if (respsmp$resp.bad==TRUE){
    bad.smp <- bad.smp + 1
    keep.sw <- FALSE
  }

  else {

  # calibration and save summary statisticis
  rslt[s, ] <- calib(respinfo=respsmp, popinfo=benchmark,
              srsinfo=srssmp)

  # increase sample counter
  s <- s + 1
  }
 }
 return (list(bad.smp=bad.smp, rslt=rslt))
}
```

B.2 A Program Calling the Program in B.1

This program is used in Chapters 3 through 5.

```
library(ResourceSelection)
# kk: repetition of simulation
# nn: SRS sample size
kk <- 1000
nn <- 8000

# Call pop.and.control to generate population (x, y and response model)
              and control totals
```

```
# Scenarios under "Y main":

# 1 100% response rate
ymain.r01 <- pop.and.control(seed=41151515, lambda=10000, lambda_i=c(0,
              0), lambda_j=c(0, 0), lambda_ij=matrix(c(0,0,0,0), nrow =
              2, ncol = 2), yseed=15157552, ymu=1000, yalpha=c(-
              200,300), ybeta=c(-100,150), ygamma=c(0,0,0,0),
              ysigma=30, rmeans = c(1, 1, 1, 1))
summary (ymain.r01$pop)
ddply(ymain.r01$pop,~xvar12,summarise,mean=mean(y))
ddply(ymain.r01$pop,~xvar12,summarise,mean=mean(respflag))

# check additive independence in Y model
ymain.fit <- lm(y ~ xvar1*xvar2, data=ymain.r01$pop)
summary(ymain.fit)

################################################
# 2  0.45000.4500 0.3000      0.3000      0.0000      1.0000
ymain.r02 <- pop.and.control(seed=41151515, lambda=10000, lambda_i=c(0,
              0), lambda_j=c(0, 0), lambda_ij=matrix(c(0,0,0,0), nrow =
              2, ncol = 2), yseed=15157552, ymu=1000, yalpha=c(-
              200,300), ybeta=c(-100,150), ygamma=c(0,0,0,0),
              ysigma=30, rmeans = c(0.4500, 0.4500, 0.3000,  0.3000))
summary (ymain.r02$pop)
ddply(ymain.r02$pop,~xvar12,summarise,mean=mean(y))
ddply(ymain.r02$pop,~xvar12,summarise,mean=mean(respflag))

# check additive independence in R model
rmain.fit <- lm(respflag ~ xvar1*xvar2, data=ymain.r02$pop)
summary(rmain.fit)

# check multiplicative independence in R model
logistic.rmain.fit <- glm(respflag ~ xvar1 + xvar2 + xvar1*xvar2,
              family=binomial(link='logit'), data=ymain.r02$pop)
summary(logistic.rmain.fit)

anova(logistic.rmain.fit, test="Chisq")

# Hosmer-Lemeshow Goodness of Fit: computed on data after the
              observations have been segmented into groups
# based on having similar predicted probabilities. It examines whether
              the observed proportions of events
# are similar to the predicted probabilities of occurence in subgroups
              of the data set using a pearson chi square test.
# Small values with large p-values indicate a good fit to the data
              while large values with p-values below 0.05 indicate a
              poor fit.

hoslem.test(ymain.r02$pop$respflag, fitted(logistic.rmain.fit))

################################################
# 3  0.40800.1020 0.9520      0.2380      -0.4080     1.0000
ymain.r03 <- pop.and.control(seed=41151515, lambda=10000, lambda_i=c(0,
              0), lambda_j=c(0, 0), lambda_ij=matrix(c(0,0,0,0), nrow =
              2, ncol = 2), yseed=15157552, ymu=1000, yalpha=c(-
              200,300), ybeta=c(-100,150), ygamma=c(0,0,0,0),
```

```
              ysigma=30, rmeans = c(0.4080,  0.1020,  0.9520,
                  0.2380))

# 4  0.12000.4800 0.0200      0.3800      0.0000      4.7500
ymain.r04 <- pop.and.control(seed=41151515, lambda=10000, lambda_i=c(0,
              0), lambda_j=c(0, 0), lambda_ij=matrix(c(0,0,0,0), nrow =
              2, ncol = 2), yseed=15157552, ymu=1000, yalpha=c(-
              200,300), ybeta=c(-100,150), ygamma=c(0,0,0,0),
              ysigma=30, rmeans = c(0.1200,  0.4800,  0.0200,
                  0.3800))

# 5  0.26000.9400 0.0600      0.7400      0.0000      3.4113
ymain.r05 <- pop.and.control(seed=41151515, lambda=10000, lambda_i=c(0,
              0), lambda_j=c(0, 0), lambda_ij=matrix(c(0,0,0,0), nrow =
              2, ncol = 2), yseed=15157552, ymu=1000, yalpha=c(-
              200,300), ybeta=c(-100,150), ygamma=c(0,0,0,0),
              ysigma=30, rmeans = c(0.2600,  0.9400,  0.0600,
                  0.7400))

# 6  0.28000.9200 0.0800      0.7200      0.0000      2.7391
ymain.r06 <- pop.and.control(seed=41151515, lambda=10000, lambda_i=c(0,
              0), lambda_j=c(0, 0), lambda_ij=matrix(c(0,0,0,0), nrow =
              2, ncol = 2), yseed=15157552, ymu=1000, yalpha=c(-
              200,300), ybeta=c(-100,150), ygamma=c(0,0,0,0),
              ysigma=30, rmeans = c(0.2800,  0.9200,  0.0800,
                  0.7200))

# 7  0.32000.8800 0.1200      0.6800      0.0000      2.0606
ymain.r07 <- pop.and.control(seed=41151515, lambda=10000, lambda_i=c(0,
              0), lambda_j=c(0, 0), lambda_ij=matrix(c(0,0,0,0), nrow =
              2, ncol = 2), yseed=15157552, ymu=1000, yalpha=c(-
              200,300), ybeta=c(-100,150), ygamma=c(0,0,0,0),
              ysigma=30, rmeans = c(0.3200,  0.8800,  0.1200,
                  0.6800))

# 8  0.40000.8000 0.2000      0.6000      0.0000      1.5000
ymain.r08 <- pop.and.control(seed=41151515, lambda=10000, lambda_i=c(0,
              0), lambda_j=c(0, 0), lambda_ij=matrix(c(0,0,0,0), nrow =
              2, ncol = 2), yseed=15157552, ymu=1000, yalpha=c(-
              200,300), ybeta=c(-100,150), ygamma=c(0,0,0,0),
              ysigma=30, rmeans = c(0.4000,  0.8000,  0.2000,
                  0.6000))

# 9  0.46000.7400 0.2600      0.5400      0.0000      1.2911
ymain.r09 <- pop.and.control(seed=41151515, lambda=10000, lambda_i=c(0,
              0), lambda_j=c(0, 0), lambda_ij=matrix(c(0,0,0,0), nrow =
              2, ncol = 2), yseed=15157552, ymu=1000, yalpha=c(-
              200,300), ybeta=c(-100,150), ygamma=c(0,0,0,0),
              ysigma=30, rmeans = c(0.4600,  0.7400,  0.2600,
                  0.5400))

# 10  0.54000.6600      0.3400      0.4600      0.0000      1.1070
ymain.r10 <- pop.and.control(seed=41151515, lambda=10000, lambda_i=c(0,
              0), lambda_j=c(0, 0), lambda_ij=matrix(c(0,0,0,0), nrow =
              2, ncol = 2), yseed=15157552, ymu=1000, yalpha=c(-
              200,300), ybeta=c(-100,150), ygamma=c(0,0,0,0),
```

```
              ysigma=30, rmeans = c(0.5400,  0.6600,  0.3400,
                  0.4600))

# 11  0.56000.6400      0.3600      0.4400      0.0000      1.0694
ymain.r11 <- pop.and.control(seed=41151515, lambda=10000, lambda_i=c(0,
              0), lambda_j=c(0, 0), lambda_ij=matrix(c(0,0,0,0), nrow =
              2, ncol = 2), yseed=15157552, ymu=1000, yalpha=c(-
              200,300), ybeta=c(-100,150), ygamma=c(0,0,0,0),
              ysigma=30, rmeans = c(0.5600,  0.6400,  0.3600,
                  0.4400))

# 12  0.23000.0700      0.5500      0.1500      -0.2400     0.8961
ymain.r12 <- pop.and.control(seed=41151515, lambda=10000, lambda_i=c(0,
              0), lambda_j=c(0, 0), lambda_ij=matrix(c(0,0,0,0), nrow =
              2, ncol = 2), yseed=15157552, ymu=1000, yalpha=c(-
              200,300), ybeta=c(-100,150), ygamma=c(0,0,0,0),
              ysigma=30, rmeans = c(0.2300,  0.0700,  0.5500,
                  0.1500))

# 13  0.20000.1000      0.5200      0.1800      -0.2400     0.6923
ymain.r13 <- pop.and.control(seed=41151515, lambda=10000, lambda_i=c(0,
              0), lambda_j=c(0, 0), lambda_ij=matrix(c(0,0,0,0), nrow =
              2, ncol = 2), yseed=15157552, ymu=1000, yalpha=c(-
              200,300), ybeta=c(-100,150), ygamma=c(0,0,0,0),
              ysigma=30, rmeans = c(0.2000,  0.1000,  0.5200,
                  0.1800))

# 14  0.15000.1500      0.4700      0.2300      -0.2400     0.4894
ymain.r14 <- pop.and.control(seed=41151515, lambda=10000, lambda_i=c(0,
              0), lambda_j=c(0, 0), lambda_ij=matrix(c(0,0,0,0), nrow =
              2, ncol = 2), yseed=15157552, ymu=1000, yalpha=c(-
              200,300), ybeta=c(-100,150), ygamma=c(0,0,0,0),
              ysigma=30, rmeans = c(0.1500,  0.1500,  0.4700,
                  0.2300))

# 15  0.09000.2100      0.4100      0.2900      -0.2400     0.3031
ymain.r15 <- pop.and.control(seed=41151515, lambda=10000, lambda_i=c(0,
              0), lambda_j=c(0, 0), lambda_ij=matrix(c(0,0,0,0), nrow =
              2, ncol = 2), yseed=15157552, ymu=1000, yalpha=c(-
              200,300), ybeta=c(-100,150), ygamma=c(0,0,0,0),
              ysigma=30, rmeans = c(0.0900,  0.2100,  0.4100,
                  0.2900))

# 16  0.04000.2600      0.3600      0.3400      -0.2400     0.1453
ymain.r16 <- pop.and.control(seed=41151515, lambda=10000, lambda_i=c(0,
              0), lambda_j=c(0, 0), lambda_ij=matrix(c(0,0,0,0), nrow =
              2, ncol = 2), yseed=15157552, ymu=1000, yalpha=c(-
              200,300), ybeta=c(-100,150), ygamma=c(0,0,0,0),
              ysigma=30, rmeans = c(0.0400,  0.2600,  0.3600,
                  0.3400))

# 17  0.02000.5800      0.6600      0.7400      -0.4800     0.0387
ymain.r17 <- pop.and.control(seed=41151515, lambda=10000, lambda_i=c(0,
              0), lambda_j=c(0, 0), lambda_ij=matrix(c(0,0,0,0), nrow =
              2, ncol = 2), yseed=15157552, ymu=1000, yalpha=c(-
              200,300), ybeta=c(-100,150), ygamma=c(0,0,0,0),
```

```
                ysigma=30, rmeans = c(0.0200,  0.5800,  0.6600,
                    0.7400))


# The procedure is repeated for 17 response scenarios under "Y int".
# The R code is not shown here.

#########################################################################
# Call function srs.resp.calib (select sample, calibrate, and save
                summary statistics)
#########################################################################

# popinfo, srsinfo, respinfo: dataset plus some control totals
# popdata, srsdata: dataset
# k: repetition of simulation
# srs.size: SRS sample size

ymain.r01.out <- srs.resp.calib (benchmark=ymain.r01, k=kk,
                srs.size=nn)
# Repeated for ymain.r02.out through ymain.r17.out.  R code is not
                shown.

yint.r01.out <- srs.resp.calib (benchmark=yint.r01, k=kk, srs.size=nn)
# Repeated for yint.r02.out through yint.r17.out.  R code is not shown.

save(ymain.r01, ymain.r02, ymain.r03, ymain.r04, ymain.r05, ymain.r06,
                ymain.r07, ymain.r08, ymain.r09, ymain.r10, ymain.r11,
                ymain.r12, ymain.r13, ymain.r14, ymain.r15, ymain.r16,
                ymain.r17,
     yint.r01, yint.r02, yint.r03, yint.r04, yint.r05, yint.r06,
                yint.r07, yint.r08, yint.r09, yint.r10, yint.r11,
                yint.r12, yint.r13, yint.r14, yint.r15, yint.r16,
                yint.r17,
     ymain.r01.out, ymain.r02.out, ymain.r03.out, ymain.r04.out,
                ymain.r05.out, ymain.r06.out, ymain.r07.out,
                ymain.r08.out, ymain.r09.out, ymain.r10.out,
                ymain.r11.out, ymain.r12.out, ymain.r13.out,
                ymain.r14.out, ymain.r15.out, ymain.r16.out,
                ymain.r17.out,
     yint.r01.out, yint.r02.out, yint.r03.out, yint.r04.out,
                yint.r05.out, yint.r06.out, yint.r07.out, yint.r08.out,
                yint.r09.out, yint.r10.out, yint.r11.out, yint.r12.out,
                yint.r13.out, yint.r14.out, yint.r15.out, yint.r16.out,
                yint.r17.out,
                file="D:\\Dissertation\\CompareThreeCalibrationEstimators
                \\Simulation\\SRS1.RData")
```

B.3 A Program for Saving Results from Each Simulated Sample and Evaluation Measures over All the Simulated Samples

This program is used in Chapters 3 and 4.

```
library(sampling)
library(survey)

######################################################################
# Function for creating measures from the S good samples
######################################################################

all.info <- function(datain){
  datain <- data.frame(datain)
  attach(datain)

  # For mean: relative bias
  resp.rel.bias.mean <- (resp.mean - pop.mean)/pop.mean
  ps.rel.bias.mean <- (ps.mean - pop.mean)/pop.mean
  rk.rel.bias.mean <- (rk.mean - pop.mean)/pop.mean
  rk.rel.bias.mean.rake <- (rk.mean.rake - pop.mean)/pop.mean
  greg.rel.bias.mean <- (greg.mean - pop.mean)/pop.mean

  # For mean: relative square root of mse
  ps.rel.sqrt.mse.mean <- sqrt((ps.mean - pop.mean)^2)/pop.mean
  rk.rel.sqrt.mse.mean <- sqrt((rk.mean - pop.mean)^2)/pop.mean
  greg.rel.sqrt.mse.mean <- sqrt((greg.mean - pop.mean)^2)/pop.mean

  # For mean: bias ratio or t-statitics
  ps.bias.ratio.mean = (ps.mean - pop.mean) / ps.mean.se
  rk.bias.ratio.mean = (rk.mean - pop.mean) / rk.mean.se
  greg.bias.ratio.mean = (greg.mean - pop.mean) / greg.mean.se

  # Total: relative bias
  resp.rel.bias.total <- (resp.total - pop.total)/pop.total
  ps.rel.bias.total  <- (ps.total - pop.total)/pop.total
  rk.rel.bias.total <- (rk.total - pop.total)/pop.total
  rk.rel.bias.total.rake <- (rk.total.rake - pop.total)/pop.total
  greg.rel.bias.total <- (greg.total - pop.total)/pop.total

  # Total: relative square root of mse
  ps.rel.sqrt.mse.total <- sqrt((ps.total - pop.total)^2)/pop.total
  rk.rel.sqrt.mse.total <- sqrt((rk.total - pop.total)^2)/pop.total
  greg.rel.sqrt.mse.total <- sqrt((greg.total - pop.total)^2)/pop.total

  # Total: bias ratio or t-statitics
  ps.bias.ratio.total = (ps.total - pop.total) / ps.total.se
  rk.bias.ratio.total = (rk.total - pop.total) / rk.total.se
  greg.bias.ratio.total = (greg.total - pop.total) / greg.total.se

  # Distance Measure using empirical variance, for Chapter 4
  ps.Distance.Chap4.EmpVar <- ps.Diff.Nrc11^2/var(ps.Diff.Nrc11)
  rk.Distance.Chap4.EmpVar <- rk.Diff.Nrc11^2/var(rk.Diff.Nrc11)
  greg.Distance.Chap4.EmpVar <- greg.Diff.Nrc11^2/var(greg.Diff.Nrc11)
```

```
rk.Distance.est.LG384 <- ifelse(rk.Distance.Chap4.est>3.84, 1, 0)
rk.Distance.est.LG663 <- ifelse(rk.Distance.Chap4.est>6.63, 1, 0)

greg.Distance.est.LG384 <- ifelse(greg.Distance.Chap4.est>3.84, 1, 0)
greg.Distance.est.LG663 <- ifelse(greg.Distance.Chap4.est>6.63, 1, 0)

#####################################################################
# variable for examing samples by groups

rk.ranges <- quantile(rk.Distance.Chap4.est, c(0, .05, .10, .15, .20,
             .25, .30, .35, .40, .45, .50, .55, .60, .65, .70, .75,
             .80, .85, .90, .95, 1), na.rm=TRUE)
rk.grp <- cut(rk.Distance.Chap4.est, rk.ranges, include.LOWEST=TRUE)
rk.grp.num <- as.numeric(rk.grp)

greg.ranges <- quantile(greg.Distance.Chap4.est, c(0, .05, .10, .15,
             .20, .25, .30, .35, .40, .45, .50, .55, .60, .65, .70,
             .75, .80, .85, .90, .95, 1), na.rm=TRUE)
greg.grp <- cut(greg.Distance.Chap4.est, greg.ranges,
             include.LOWEST=TRUE)
greg.grp.num <- as.numeric(greg.grp)

datafinal <- data.frame(  datain,
                          resp.rel.bias.mean,
                          ps.rel.bias.mean,
                          rk.rel.bias.mean,
                          rk.rel.bias.mean.rake,
                          greg.rel.bias.mean,

                          ps.rel.sqrt.mse.mean,
                          rk.rel.sqrt.mse.mean,
                          greg.rel.sqrt.mse.mean,

                          ps.bias.ratio.mean,
                          rk.bias.ratio.mean,
                          greg.bias.ratio.mean,

                          resp.rel.bias.total,
                          ps.rel.bias.total,
                          rk.rel.bias.total,
                          rk.rel.bias.total.rake,
                          greg.rel.bias.total,

                          ps.rel.sqrt.mse.total,
                          rk.rel.sqrt.mse.total,
                          greg.rel.sqrt.mse.total,

                          ps.bias.ratio.total,
                          rk.bias.ratio.total,
                          greg.bias.ratio.total,

                          ps.Distance.Chap4.EmpVar,
                          rk.Distance.Chap4.EmpVar,
                          greg.Distance.Chap4.EmpVar,

                          rk.Distance.est.LG384,
```

```
                            rk.Distance.est.LG663,
                            greg.Distance.est.LG384,
                            greg.Distance.est.LG663,

                            rk.grp,
                            rk.grp.num,
                            greg.grp,
                            greg.grp.num)
  detach(datain)

  return(datafinal)
}

#########################################################
# Function for generating overall summary statitics
#########################################################

overall <- function (datain, stat){

  attach(datain)

  # For mean: relative bias, relative standard error, relative square
              root of mse, bias ratio
  pop.mean <- mean(pop.mean)
  resp.rel.bias.mean <- mean(resp.rel.bias.mean)
  ps.rel.bias.mean <- mean(ps.rel.bias.mean)
  rk.rel.bias.mean <- mean(rk.rel.bias.mean)
  rk.rel.bias.mean.rake <- mean(rk.rel.bias.mean.rake)
  greg.rel.bias.mean <- mean(greg.rel.bias.mean)

  resp.rel.se.mean <- sqrt(var(resp.mean))/pop.mean  # empirical
  ps.rel.se.mean <- sqrt(var(ps.mean))/pop.mean
  rk.rel.se.mean <- sqrt(var(rk.mean))/pop.mean
  rk.rel.se.mean.rake <- sqrt(var(rk.mean.rake))/pop.mean
  greg.rel.se.mean <- sqrt(var(greg.mean))/pop.mean

  ps.rel.sqrt.mse.mean <- mean(ps.rel.sqrt.mse.mean)
  rk.rel.sqrt.mse.mean <- mean(rk.rel.sqrt.mse.mean)
  greg.rel.sqrt.mse.mean <- mean(greg.rel.sqrt.mse.mean)

  ps.bias.ratio.mean <- mean(ps.bias.ratio.mean)
  rk.bias.ratio.mean <- mean(rk.bias.ratio.mean)
  greg.bias.ratio.mean <- mean(greg.bias.ratio.mean)

  # For total: relative bias, relative standard error, relative square
              root of mse, bias ratio
  pop.total <- mean(pop.total)
  resp.rel.bias.total <- mean(resp.rel.bias.total)
  ps.rel.bias.total <- mean(ps.rel.bias.total)
  rk.rel.bias.total <- mean(rk.rel.bias.total)
  rk.rel.bias.total.rake <- mean(rk.rel.bias.total.rake)
  greg.rel.bias.total <- mean(greg.rel.bias.total)

  resp.rel.se.total <- sqrt(var(resp.total))/pop.total  # empirical
  ps.rel.se.total <- sqrt(var(ps.total))/pop.total
  rk.rel.se.total <- sqrt(var(rk.total))/pop.total
  rk.rel.se.total.rake <- sqrt(var(rk.total.rake))/pop.total
```

```
greg.rel.se.total <- sqrt(var(greg.total))/pop.total

ps.rel.sqrt.mse.total <- mean(ps.rel.sqrt.mse.total)
rk.rel.sqrt.mse.total <- mean(rk.rel.sqrt.mse.total)
greg.rel.sqrt.mse.total <- mean(greg.rel.sqrt.mse.total)

ps.bias.ratio.total <- mean(ps.bias.ratio.total)
rk.bias.ratio.total <- mean(rk.bias.ratio.total)
greg.bias.ratio.total <- mean(greg.bias.ratio.total)

CheckLumley.ps.rel.se.total <- mean(ps.total.se/pop.total)   # Lumley
            estimate for QC.
CheckLumley.ps.rel.se.total.SSW <- mean(ps.total.se.SSW/pop.total)  #
            SSW for QC.

# respondent Sample sizes
respcnt11 <- mean(respcnt11)
respcnt12 <- mean(respcnt12)
respcnt21 <- mean(respcnt21)
respcnt22 <- mean(respcnt22)

# Odds ratios
pop.OR = mean (pop.OR)
resp.OR = mean (resp.OR)
ps.OR = mean (ps.OR)
rk.OR = mean (rk.OR)
greg.OR = mean (greg.OR)

# CI coverage
ps.CI.coverage.mean <- mean(ps.mean.CI.coverage)
rk.CI.coverage.mean <- mean(rk.mean.CI.coverage)
greg.CI.coverage.mean <- mean(greg.mean.CI.coverage)

ps.CI.coverage.total <- mean(ps.total.CI.coverage)
ps.CI.coverage.total.SSW <- mean(ps.total.CI.coverage.SSW)
rk.CI.coverage.total <- mean(rk.total.CI.coverage)
greg.CI.coverage.total <- mean(greg.total.CI.coverage)

# Distance Measure (using empirical variance)
ps.Distance.Chap4.EmpVar <- mean(ps.Distance.Chap4.EmpVar)
rk.Distance.Chap4.EmpVar <- mean(rk.Distance.Chap4.EmpVar)
greg.Distance.Chap4.EmpVar <-mean(greg.Distance.Chap4.EmpVar)

# Distance Measure -- estimated from sample, and then taking average
ps.Distance.Chap4.est <- mean(ps.Distance.Chap4.est)
rk.Distance.Chap4.est <- mean(rk.Distance.Chap4.est)
greg.Distance.Chap4.est <- mean(greg.Distance.Chap4.est)

rk.Distance.est.LG384 <- mean(rk.Distance.est.LG384)
rk.Distance.est.LG663 <- mean(rk.Distance.est.LG663)

greg.Distance.est.LG384 <- mean(greg.Distance.est.LG384)
greg.Distance.est.LG663 <- mean(greg.Distance.est.LG663)

# R-squared for main-effect model and full model (including
            interaction term)
ps.R2.main <- mean(ps.R2.main)
```

```
  rk.R2.main <- mean(rk.R2.main)
  greg.R2.main <- mean(greg.R2.main)
  ps.R2.int <- mean(ps.R2.int)
  rk.R2.int <- mean(rk.R2.int)
  greg.R2.int <- mean(greg.R2.int)

  # compare analytical results and simulation results
  # ratio.rk_greg = mean(rk.total - greg.total ) / mean(diff.rk_greg)
  # ratio.rk_ps = mean(rk.total - ps.total ) / mean(diff.rk_ps)

  evalmean <- cbind (pop.mean, resp.rel.bias.mean, ps.rel.bias.mean,
              greg.rel.bias.mean, rk.rel.bias.mean,
              rk.rel.bias.mean.rake, resp.rel.se.mean, ps.rel.se.mean,
              greg.rel.se.mean, rk.rel.se.mean, rk.rel.se.mean.rake,
                    ps.rel.sqrt.mse.mean, greg.rel.sqrt.mse.mean,
              rk.rel.sqrt.mse.mean, ps.bias.ratio.mean,
              greg.bias.ratio.mean, rk.bias.ratio.mean,
                    ps.CI.coverage.mean, greg.CI.coverage.mean,
              rk.CI.coverage.mean, respcnt11, respcnt12, respcnt21,
              respcnt22, pop.OR, resp.OR, ps.OR, greg.OR, rk.OR,
                    ps.Distance.Chap4.EmpVar,
              rk.Distance.Chap4.EmpVar, greg.Distance.Chap4.EmpVar,
                    ps.Distance.Chap4.est, rk.Distance.Chap4.est,
              greg.Distance.Chap4.est,
                    rk.Distance.est.LG384, rk.Distance.est.LG663,
              greg.Distance.est.LG384, greg.Distance.est.LG663,
                    ps.R2.main, rk.R2.main, greg.R2.main, ps.R2.int,
              rk.R2.int, greg.R2.int)

  evaltotal <- cbind (pop.total, resp.rel.bias.total,
              ps.rel.bias.total, greg.rel.bias.total,
              rk.rel.bias.total, rk.rel.bias.total.rake,
                    resp.rel.se.total, ps.rel.se.total,
              greg.rel.se.total, rk.rel.se.total, rk.rel.se.total.rake,
                    ps.rel.sqrt.mse.total, greg.rel.sqrt.mse.total,
              rk.rel.sqrt.mse.total,
                    ps.bias.ratio.total, greg.bias.ratio.total,
              rk.bias.ratio.total,
                    ps.CI.coverage.total, greg.CI.coverage.total,
              rk.CI.coverage.total,
                    respcnt11, respcnt12, respcnt21, respcnt22,
                    pop.OR, resp.OR, ps.OR, greg.OR, rk.OR,
                    ps.Distance.Chap4.EmpVar,
              rk.Distance.Chap4.EmpVar, greg.Distance.Chap4.EmpVar,
                    ps.Distance.Chap4.est, rk.Distance.Chap4.est,
              greg.Distance.Chap4.est,
                    rk.Distance.est.LG384, rk.Distance.est.LG663,
              greg.Distance.est.LG384, greg.Distance.est.LG663,
                    ps.R2.main, rk.R2.main, greg.R2.main, ps.R2.int,
              rk.R2.int, greg.R2.int,
                    CheckLumley.ps.rel.se.total,
              CheckLumley.ps.rel.se.total.SSW,
              ps.CI.coverage.total.SSW)

  detach(datain)
  ifelse(stat==1, return(evalmean), return(evaltotal))
}
```

```
#############################################################
# Function for calculating group summary statistics
#############################################################

groups.rk <- function (datain, stat){

  attach(datain)

  ps.grp.Distance.rkgrp <-
              by(datain$ps.Distance.Chap4.est,INDICES=rk.grp, mean)
  rk.grp.Distance <- by(datain$rk.Distance.Chap4.est,INDICES=rk.grp,
              mean)

  # Odds ratio
  pop.grp.OR.rkgrp <- by(datain$pop.OR,INDICES=rk.grp, mean)
  resp.grp.OR.rkgrp <- by(datain$resp.OR,INDICES=rk.grp, mean)
  ps.grp.OR.rkgrp <- by(datain$ps.OR,INDICES=rk.grp, mean)
  rk.grp.OR <- by(datain$rk.OR,INDICES=rk.grp, mean)

  # R-squared
  ps.grp.R2.main.rkgrp <- by(datain$ps.R2.main,INDICES=rk.grp, mean)
  ps.grp.R2.int.rkgrp <- by(datain$ps.R2.int,INDICES=rk.grp, mean)
  rk.grp.R2.main <- by(datain$rk.R2.main,INDICES=rk.grp, mean)
  rk.grp.R2.int <- by(datain$rk.R2.int,INDICES=rk.grp, mean)

  # For mean: bias
  ps.grp.rel.bias.mean.rkgrp <-
              by(datain$ps.rel.bias.mean,INDICES=rk.grp, mean)
  rk.grp.rel.bias.mean <- by(datain$rk.rel.bias.mean,INDICES=rk.grp,
              mean)

  # For mean: bias ratio
  ps.grp.bias.ratio.mean.rkgrp <-
              by(datain$ps.bias.ratio.mean,INDICES=rk.grp, mean)
  rk.grp.bias.ratio.mean <-
              by(datain$rk.bias.ratio.mean,INDICES=rk.grp, mean)

  # For mean: CI coverage
  ps.grp.CI.coverage.mean.rkgrp <-
              by(datain$ps.mean.CI.coverage,INDICES=rk.grp, mean)
  rk.grp.CI.coverage.mean <-
              by(datain$rk.mean.CI.coverage,INDICES=rk.grp, mean)

  bygroup.mean <- data.frame(cbind(ps.grp.rel.bias.mean.rkgrp,
              rk.grp.rel.bias.mean, ps.grp.bias.ratio.mean.rkgrp,
              rk.grp.bias.ratio.mean, ps.grp.CI.coverage.mean.rkgrp,
              rk.grp.CI.coverage.mean, ps.grp.Distance.rkgrp,
              rk.grp.Distance, pop.grp.OR.rkgrp, resp.grp.OR.rkgrp,
              ps.grp.OR.rkgrp, rk.grp.OR, ps.grp.R2.main.rkgrp,
              ps.grp.R2.int.rkgrp, rk.grp.R2.main, rk.grp.R2.int))


  # For total: bias
  ps.grp.rel.bias.total.rkgrp <-
              by(datain$ps.rel.bias.total,INDICES=rk.grp, mean)
```

```
  rk.grp.rel.bias.total <- by(datain$rk.rel.bias.total,INDICES=rk.grp,
              mean)

  # For total: bias ratio
  ps.grp.bias.ratio.total.rkgrp <-
              by(datain$ps.bias.ratio.total,INDICES=rk.grp, mean)
  rk.grp.bias.ratio.total <-
              by(datain$rk.bias.ratio.total,INDICES=rk.grp, mean)

  # For total: CI coverage
  ps.grp.CI.coverage.total.rkgrp <-
              by(datain$ps.total.CI.coverage,INDICES=rk.grp, mean)
  rk.grp.CI.coverage.total <-
              by(datain$rk.total.CI.coverage,INDICES=rk.grp, mean)

  ps.grp.CI.coverage.total.rkgrp.SSW <-
              by(datain$ps.total.CI.coverage.SSW,INDICES=rk.grp, mean)

  bygroup.total <- data.frame(cbind(ps.grp.rel.bias.total.rkgrp,
              rk.grp.rel.bias.total, ps.grp.bias.ratio.total.rkgrp,
              rk.grp.bias.ratio.total, ps.grp.CI.coverage.total.rkgrp,
              rk.grp.CI.coverage.total, ps.grp.Distance.rkgrp,
              rk.grp.Distance, pop.grp.OR.rkgrp, resp.grp.OR.rkgrp,
              ps.grp.OR.rkgrp, rk.grp.OR, ps.grp.R2.main.rkgrp,
              ps.grp.R2.int.rkgrp, rk.grp.R2.main, rk.grp.R2.int,
              ps.grp.CI.coverage.total.rkgrp.SSW))
  detach(datain)

  ifelse(stat==1, return(bygroup.mean), return(bygroup.total))
}

groups.greg <- function (datain, stat){

  attach(datain)

  ps.grp.Distance.greggrp <-
              by(datain$ps.Distance.Chap4.est,INDICES=greg.grp, mean)
  greg.grp.Distance <-
              by(datain$greg.Distance.Chap4.est,INDICES=greg.grp, mean)

  # Odds ratio
  pop.grp.OR.greggrp <- by(datain$pop.OR,INDICES=greg.grp, mean)
  resp.grp.OR.greggrp <- by(datain$resp.OR,INDICES=greg.grp, mean)
  ps.grp.OR.greggrp <- by(datain$ps.OR,INDICES=greg.grp, mean)
  greg.grp.OR <- by(datain$greg.OR,INDICES=greg.grp, mean)

  # R-squared
  ps.grp.R2.main.greggrp <- by(datain$ps.R2.main,INDICES=greg.grp,
              mean)
  ps.grp.R2.int.greggrp <- by(datain$ps.R2.int,INDICES=greg.grp, mean)
  greg.grp.R2.main <- by(datain$greg.R2.main,INDICES=greg.grp, mean)
  greg.grp.R2.int <- by(datain$greg.R2.int,INDICES=greg.grp, mean)

  # For mean: bias
  ps.grp.rel.bias.mean.greggrp <-
              by(datain$ps.rel.bias.mean,INDICES=greg.grp, mean)
```

```
greg.grp.rel.bias.mean <-
            by(datain$greg.rel.bias.mean,INDICES=greg.grp, mean)

# For mean: bias ratio
ps.grp.bias.ratio.mean.greggrp <-
            by(datain$ps.bias.ratio.mean,INDICES=greg.grp, mean)
greg.grp.bias.ratio.mean <-
            by(datain$greg.bias.ratio.mean,INDICES=greg.grp, mean)

# For mean: CI coverage
ps.grp.CI.coverage.mean.greggrp <-
            by(datain$ps.mean.CI.coverage,INDICES=greg.grp, mean)
greg.grp.CI.coverage.mean <-
            by(datain$greg.mean.CI.coverage,INDICES=greg.grp, mean)

bygroup.mean <- data.frame(cbind(ps.grp.rel.bias.mean.greggrp,
            greg.grp.rel.bias.mean, ps.grp.bias.ratio.mean.greggrp,
            greg.grp.bias.ratio.mean,ps.grp.CI.coverage.mean.greggrp,
            greg.grp.CI.coverage.mean, ps.grp.Distance.greggrp,
            greg.grp.Distance, pop.grp.OR.greggrp,
            resp.grp.OR.greggrp, ps.grp.OR.greggrp, greg.grp.OR,
            ps.grp.R2.main.greggrp, ps.grp.R2.int.greggrp,
            greg.grp.R2.main, greg.grp.R2.int))


# For total: bias
ps.grp.rel.bias.total.greggrp <-
            by(datain$ps.rel.bias.total,INDICES=greg.grp, mean)
greg.grp.rel.bias.total <-
            by(datain$greg.rel.bias.total,INDICES=greg.grp, mean)

# For total: bias ratio
ps.grp.bias.ratio.total.greggrp <-
            by(datain$ps.bias.ratio.total,INDICES=greg.grp, mean)
greg.grp.bias.ratio.total <-
            by(datain$greg.bias.ratio.total,INDICES=greg.grp, mean)

# For total: CI coverage
ps.grp.CI.coverage.total.greggrp <-
            by(datain$ps.total.CI.coverage,INDICES=greg.grp, mean)
greg.grp.CI.coverage.total <-
            by(datain$greg.total.CI.coverage,INDICES=greg.grp, mean)

ps.grp.CI.coverage.total.greggrp.SSW <-
            by(datain$ps.total.CI.coverage.SSW,INDICES=greg.grp,
            mean)

bygroup.total <- data.frame(cbind(ps.grp.rel.bias.total.greggrp,
            greg.grp.rel.bias.total, ps.grp.bias.ratio.total.greggrp,
            greg.grp.bias.ratio.total,
            ps.grp.CI.coverage.total.greggrp,
            greg.grp.CI.coverage.total, ps.grp.Distance.greggrp,
            greg.grp.Distance, pop.grp.OR.greggrp,
            resp.grp.OR.greggrp, ps.grp.OR.greggrp, greg.grp.OR,
            ps.grp.R2.main.greggrp, ps.grp.R2.int.greggrp,
            greg.grp.R2.main, greg.grp.R2.int,
```

```
              ps.grp.CI.coverage.total.greggrp.SSW))
  detach(datain)

  ifelse(stat==1, return(bygroup.mean), return(bygroup.total))
}
```

B.4 A Program for Calling the Program in B.3 to Produce Results over Repeated Sampling

This program is used in Chapters 3 and 4.

```
# Set up an empty matrix to store overall summary statistics
eval <- matrix (nrow=68, ncol=49)
  colnames(eval) <- c("ID", "pop.truth",
                      "resp.bias", "ps.bias",  "greg.bias", "rk.bias",
                "rk.bias.rake", "resp.se", "ps.se", "greg.se", "rk.se",
                "rk.se.rake", "ps.sqrt.mse",  "greg.sqrt.mse",
                "rk.sqrt.mse", "ps.bias.ratio", "greg.bias.ratio",
                "rk.bias.ratio", "ps.CI.coverage", "greg.CI.coverage",
                "rk.CI.coverage", "respcnt11", "respcnt12", "respcnt21",
                "respcnt22", "pop.OR", "resp.OR", "ps.OR", "greg.OR",
                "rk.OR", "ps.Distance.Chap4.EmpVar",
                "rk.Distance.Chap4.EmpVar", "greg.Distance.Chap4.EmpVar",
                 "ps.Distance.Chap4.est", "rk.Distance.Chap4.est",
                "greg.Distance.Chap4.est","rk.Distance.est.LG384",
                "rk.Distance.est.LG663", "greg.Distance.est.LG384",
                "greg.Distance.est.LG663","ps.R2.main", "rk.R2.main",
                "greg.R2.main", "ps.R2.int", "rk.R2.int", "greg.R2.int",
                 "CheckLumley.ps.se", "CheckLumley.ps.se.SSW",
                "ps.CI.coverage.total.SSW")

  eval[1, "ID"] <-  "MEAN -- Y_Main, R_scenario  1: R11=1.0000,
                R12=1.0000, R21=1.0000, R22=1.0000, DIFF=0.0000,
                OR=1.0000"
  # R code not shown for filling in MEAN – Y_Main, R_scenarios 2
                through 16.
  eval[17, "ID"] <- "MEAN -- Y_Main, R_scenario 17: R11=0.0200,
                R12=0.5800, R21=0.6600, R22=0.7400, DIFF=-0.4800,
                OR=0.0387"

  eval[18, "ID"] <-  "MEAN -- Y_Interaction, R_scenario  1: R11=1.0000,
                R12=1.0000, R21=1.0000, R22=1.0000, DIFF=0.0000,
                OR=1.0000"
   # R code not shown for filling in MEAN – Y_Interaction, R_scenarios
                2 through 16.
   eval[34, "ID"] <- "MEAN -- Y_Interaction, R_scenario 17: R11=0.0200,
                R12=0.5800, R21=0.6600, R22=0.7400, DIFF=-0.4800,
                OR=0.0387"
```

```
   eval[35, "ID"] <-  "TOTAL -- Y_Main, R_scenario  1: R11=1.0000,
               R12=1.0000, R21=1.0000, R22=1.0000, DIFF=0.0000,
               OR=1.0000"
   # R code not shown for filling in TOTAL – Y_MAIN, R_scenarios 2
               through 16.
   eval[51, "ID"] <- "TOTAL -- Y_Main, R_scenario 17: R11=0.0200,
               R12=0.5800, R21=0.6600, R22=0.7400, DIFF=-0.4800,
               OR=0.0387"

   eval[52, "ID"] <-  "TOTAL -- Y_Interaction, R_scenario  1:
               R11=1.0000, R12=1.0000, R21=1.0000, R22=1.0000,
               DIFF=0.0000, OR=1.0000"
   # R code not shown for filling in TOTAL – Y_Interaction, R_scenarios
               2 through 16.
   eval[68, "ID"] <- "TOTAL -- Y_Interaction, R_scenario 17:
               R11=0.0200, R12=0.5800, R21=0.6600, R22=0.7400, DIFF=-
               0.4800, OR=0.0387"


# Obtain results
# Y main, R 100%
  ymain.r01.rslt <- all.info(datain=ymain.r01.out$rslt)
  # R code not shown for ymain.r02.rslt through ymain.r16.rslt
  ymain.r17.rslt <- all.info(datain=ymain.r17.out$rslt)


  yint.r01.rslt <- all.info(datain=yint.r01.out$rslt)
  # R code not shown for yint.r02.rslt through yint.r16.rslt
  yint.r17.rslt <- all.info(datain=yint.r17.out$rslt)

  eval[1,2:46] <- overall(datain=ymain.r01.rslt, stat=1)
  …
  eval[17,2:46] <- overall(datain=ymain.r17.rslt, stat=1)

  eval[18,2:46] <- overall(datain=yint.r01.rslt, stat=1)
  …
  eval[34,2:46] <- overall(datain=yint.r17.rslt, stat=1)

  eval[35,2:49] <- overall(datain=ymain.r01.rslt, stat=2)
  …
  eval[51,2:49] <- overall(datain=ymain.r17.rslt, stat=2)

  eval[52,2:49] <- overall(datain=yint.r01.rslt, stat=2)
  …
  eval[68,2:49] <- overall(datain=yint.r17.rslt, stat=2)

overall.result <- data.frame(eval)
write.csv(overall.result, file =
               "D:\\Dissertation\\CompareThreeCalibrationEstimators\\Sim
               ulation\\OverallResult_DistanceMeasureIncluded.csv")
```

B.5 A Program for Calling the Programs in B.1 and B.3 to Produce Results Conditioning on Samples Grouped by Estimated Distance Measure

This program is used in Chpater 4.

```
library(ResourceSelection)

kk <- 10000   # kk: repetition of simulation
nn1 <- 200   # nn1, nn2, nn3: SRS sample size
nn2 <- 2000
nn3 <- 8000

# Call pop.and.control to generate population (x, y and response model)
              and control totals
# ymu=1000, yalpha=c(-200,300), ybeta=c(-100,150),
              ygamma=c(100,700,300,1200)
# We use only Y_Additive_Interaction model here.

# response scenarios:
# Scenario  p11   p12   p21   p22   diff  OR
# 11 0.56000.6400  0.3600      0.4400      0.0000      1.0694

# Scenarios under "Y int":
# 11  0.56000.6400      0.3600      0.4400      0.0000      1.0694
yint.r11 <-  pop.and.control(seed=41151515, lambda=10000, lambda_i=c(0,
              0), lambda_j=c(0, 0), lambda_ij=matrix(c(0,0,0,0), nrow =
              2, ncol = 2), yseed=15157552, ymu=1000, yalpha=c(-
              200,300), ybeta=c(-100,150), ygamma=c(100,700,300,1200),
              ysigma=30, rmeans = c(0.5600,  0.6400,  0.3600,
                  0.4400))

# Call function srs.resp.calib (select sample, calibrate, and save
              summary statistics)
# popinfo, srsinfo, respinfo: dataset plus some control totals
# popdata, srsdata: dataset
# k: repetition of simulation
# srs.size: SRS sample size

yint.r11.out.200 <- srs.resp.calib (benchmark=yint.r11, k=kk,
              srs.size=nn1)
yint.r11.out.2000 <- srs.resp.calib (benchmark=yint.r11, k=kk,
              srs.size=nn2)
yint.r11.out.8000 <- srs.resp.calib (benchmark=yint.r11, k=kk,
              srs.size=nn3)

# SRS n = 8000, Y_Interaction, response scenario S11
  yint.r11.rslt.8000 <- all.info(datain=yint.r11.out.8000$rslt)
  grp.mean.rk <- groups.rk(datain=yint.r11.rslt.8000, stat=1)
  grp.mean.greg <- groups.greg(datain=yint.r11.rslt.8000, stat=1)
  grp.total.rk <- groups.rk(datain=yint.r11.rslt.8000, stat=2)
  grp.total.greg <- groups.greg(datain=yint.r11.rslt.8000, stat=2)

# SRS n = 2000, Y_Interaction, response scenario S11
```

```
  yint.r11.rslt.2000 <- all.info(datain=yint.r11.out.2000$rslt)
  grp.mean.rk <- groups.rk(datain=yint.r11.rslt.2000, stat=1)
  grp.mean.greg <- groups.greg(datain=yint.r11.rslt.2000, stat=1)
  grp.total.rk <- groups.rk(datain=yint.r11.rslt.2000, stat=2)
  grp.total.greg <- groups.greg(datain=yint.r11.rslt.2000, stat=2)

  save(yint.r11.rslt.8000, yint.r11.rslt.2000,
              file="D:\\Dissertation\\CompareThreeCalibrationEstimators
              \\Simulation\\ConditioningOnSample\\CondOnSmp.RData")

# Graphs relative bias and bias ratio by type of samples based on
              distance measure.  I do this only for totals because I
              don't plan to include graphs for the means in the
              writing.

# Rel bias vs distance

png(file =
              "D:\\Dissertation\\CompareThreeCalibrationEstimators\\Sim
              ulation\\ConditioningOnSample\\Yint_R11_RelBias.png")
#win.metafile(filename="D:\\Dissertation\\CompareThreeCalibrationEstima
              tors\\Simulation\\ConditioningOnSample\\Yint_R11_RelBias.
              emf")
par(mfrow=c(2,4), oma=c(2, 2, 2, 0), mar=c(3, 3, 2, 1), mgp=c(2, 0.5,
              0))

attach(yint.r11.rslt.8000)

table(greg.grp)
greg.grp.max <- by (greg.Distance.Chap4.est, INDICES=greg.grp.num, max)
greg.v.lines.20 <- as.vector(greg.grp.max)
greg.v.lines <- greg.v.lines.20[c(5, 10, 15)]

table(rk.grp)
rk.grp.max <- by (rk.Distance.Chap4.est, INDICES=rk.grp.num, max)
rk.v.lines.20 <- as.vector(rk.grp.max)
rk.v.lines <- rk.v.lines.20[c(5, 10, 15)]

h.lines <- c(-1.96, -1.64, -1.28, 0, 1.28,1.64, 1.96)

plot(rk.Distance.Chap4.est, ps.rel.bias.total, ylab="SRS n = 8,000",
              xlab="", ylim=c(-0.025, 0.025), cex=0.3)
abline(v=rk.v.lines, col = "lightgray")
title(main="1) Poststratification", col.main="purple", font.main=2,
              line=1)

plot(greg.Distance.Chap4.est, ps.rel.bias.total, ylab="", xlab="",
              ylim=c(-0.025, 0.025), cex=0.3)
abline(v=greg.v.lines, col = "lightgray")
title(main="2) Poststratification", col.main="purple", font.main=2,
              line=1)

plot(rk.Distance.Chap4.est, rk.rel.bias.total, ylab="", xlab="",
              ylim=c(-0.025, 0.025), cex=0.3)
abline(v=rk.v.lines, col = "lightgray")
title(main="3) Raking", col.main="purple", font.main=2, line=1)
```

```
plot(greg.Distance.Chap4.est, greg.rel.bias.total, ylab="", xlab="",
              ylim=c(-0.025, 0.025), cex=0.3)
abline(v=greg.v.lines, col = "lightgray")
title(main="4) GREG_Main", col.main="purple", font.main=2, line=1)

detach(yint.r11.rslt.8000)

# R code for SRS sample size 2000 is not shown here, but similar to the
              R code for SRS sample size 8000.

mtext("Absolute Value of Relative Bias", side=2, font=2, line=1,
              outer=TRUE)
mtext("Distance Measure", side=1, font=2, line=1, outer=TRUE)

dev.off()

# Bias ratio vs Distance.Chap4.est

png(file =
              "D:\\Dissertation\\CompareThreeCalibrationEstimators\\Sim
              ulation\\ConditioningOnSample\\Yint_R11_BiasRatio.png")
#win.metafile(filename="D:\\Dissertation\\CompareThreeCalibrationEstima
              tors\\Simulation\\ConditioningOnSample\\Yint_R11_BiasRati
              o.emf")
par(mfrow=c(2,4), oma=c(2, 2, 2, 0), mar=c(3, 3, 2, 1), mgp=c(2, 0.5,
              0))

########################################
attach(yint.r11.rslt.8000)

table(greg.grp)
greg.grp.max <- by (greg.Distance.Chap4.est, INDICES=greg.grp.num, max)
greg.v.lines.20 <- as.vector(greg.grp.max)
greg.v.lines <- greg.v.lines.20[c(5, 10, 15)]

table(rk.grp)
rk.grp.max <- by (rk.Distance.Chap4.est, INDICES=rk.grp.num, max)
rk.v.lines.20 <- as.vector(rk.grp.max)
rk.v.lines <- rk.v.lines.20[c(5, 10, 15)]

h.lines <- c(-1.96, -1.64, -1.28, 0, 1.28,1.64, 1.96)

plot(rk.Distance.Chap4.est, ps.bias.ratio.total, ylab="SRS n = 8,000",
              xlab="", ylim=c(-6, 6), cex=0.3)
abline(v=rk.v.lines, col = "lightgray")
title(main="1) Poststratification", col.main="purple", font.main=2,
              line=1)

plot(greg.Distance.Chap4.est, ps.bias.ratio.total, ylab="", xlab="",
              ylim=c(-6, 6), cex=0.3)
abline(v=greg.v.lines, col = "lightgray")
title(main="2) Poststratification", col.main="purple", font.main=2,
              line=1)

plot(rk.Distance.Chap4.est, rk.bias.ratio.total, ylab="", xlab="",
              ylim=c(-6, 6), cex=0.3)
abline(v=rk.v.lines, col = "lightgray")
```

```
title(main="3) Raking", col.main="purple", font.main=2, line=1)

plot(greg.Distance.Chap4.est, greg.bias.ratio.total, ylab="", xlab="",
              ylim=c(-6, 6), cex=0.3)
abline(v=greg.v.lines, col = "lightgray")
title(main="4) GREG_Main", col.main="purple", font.main=2, line=1)

detach(yint.r11.rslt.8000)

# R code for SRS sample size 2000 is not shown here, but similar to the
              R code for SRS sample size 8000.

mtext("Absolute Value of Bias Ratio", side=2, font=2, line=1,
              outer=TRUE)
mtext("Distance Measure", side=1, font=2, line=1, outer=TRUE)

dev.off()
```

B.6 A Function to Adapt the Program in B.1 for Comparing Measures from Different Raking Variance Estimation Methods

This function is used in Chapter 5.

```
#####################################################################
# Function for SRS sampling from population and respondent sampling
#####################################################################

srs.smp <- function(srsseed, popdata, n, repnum){

  srs.bad <- FALSE

  N <- nrow(popdata)
  s <- srswor(n, N)
  bwgt <- rep (N/n, n)
  f1 <- rep (n/N, n)

  srs.smp <- data.frame(popdata[s==1,], bwgt, f1)
  srs.totals <- xtabs(~xvar1 + xvar2, data = srs.smp)

  # Form design object for JKn
  # randomize the order of the sample (although this may not be
              necessary for SRS sample, but I did this to make sure)
  # "sample" can randomize the order of the vector
  # "1:n" add a sequential number indicating the order of the record
              after random sorting
  number <- 1:n
  srs.smp.JK1 <- data.frame(srs.smp[sample(1:n), ], number,
              psu=ceiling(number/(n/repnum)))

  # TS design
  TS.dsgn <- svydesign(
```

```
    ids = ~0, # No cluster
    strata = NULL, # No strata
    # fpc = ~f1,
    weights = ~bwgt,
    data = srs.smp)

  #JK1
  dsgn <- svydesign(
    ids = ~psu,
    strata = NULL, # No strata
    # fpc = ~f1,
    weights = ~bwgt,
    data = srs.smp.JK1)

  JK1.dsgn <- as.svrepdesign(design=dsgn, type="JK1")

  if (srs.totals[1, 1]<2 | srs.totals[1, 2]<2 | srs.totals[2, 1]<2 |
              srs.totals[2, 2]<2){
    srs.bad <- TRUE
  }
  return(list(srs.bad=srs.bad, srs.smp=srs.smp, srs.totals=srs.totals,
              TS.dsgn=TS.dsgn, JK1.dsgn=JK1.dsgn))
}

resp.smp <- function (srsdata, TS.dsgn, JK1.dsgn){

  resp.bad <- FALSE

  resp.indic <- srsdata["respflag"] > 0
  resp.smp <- srsdata[resp.indic==1, ]

  resp.totals <- xtabs(~xvar1 + xvar2, data = resp.smp)

  # design objects for response sample
  TS.dsgn.resp <- subset(TS.dsgn, respflag>0)
  JK1.dsgn.resp <- subset(JK1.dsgn, respflag>0)

  if (resp.totals[1, 1]<2 | resp.totals[1, 2]<2 | resp.totals[2, 1]<2 |
              resp.totals[2, 2]<2){
    resp.bad <- TRUE
  }
  return(list(resp.bad=resp.bad, resp.smp=resp.smp,
              resp.totals=resp.totals, TS.dsgn.resp=TS.dsgn.resp,
              JK1.dsgn.resp=JK1.dsgn.resp))
}

#####################################################################
# Function for calibration and obtaining summary statistics
#####################################################################

calib <- function(respinfo, popinfo, srsinfo, TS.dsgn.resp,
              JK1.dsgn.resp){

  # Calibration TS approach
  TS.ps.dsgn <- postStratify(design = TS.dsgn.resp, strata = ~xvar1 +
              xvar2, population = popinfo$totals.xvar1xvar2,
              partial=TRUE)
```

```
TS.ps.wgt <- weights(TS.ps.dsgn)

TS.rk.dsgn <- calibrate(design = TS.dsgn.resp, formula = ~xvar1 +
            xvar2, population = c('(Intercept)'=nrow(popinfo$pop),
            xvar12=sum(popinfo$totals.xvar1xvar2[2,]),
            xvar22=sum(popinfo$totals.xvar1xvar2[,2])),
            calfun="raking")
TS.rk.wgt <- weights(TS.rk.dsgn)

TS.greg.dsgn <- calibrate(design = TS.dsgn.resp, formula = ~xvar1 +
            xvar2, population = c('(Intercept)'=nrow(popinfo$pop),
            xvar12=sum(popinfo$totals.xvar1xvar2[2,]),
            xvar22=sum(popinfo$totals.xvar1xvar2[,2])),
            calfun="linear")
TS.greg.wgt <- weights(TS.greg.dsgn)

# Calibration JK1 approach
JK1.ps.dsgn <- postStratify(design = JK1.dsgn.resp, strata = ~xvar1 +
            xvar2, population = popinfo$totals.xvar1xvar2,
            partial=TRUE)
JK1.ps.wgt <- weights(JK1.ps.dsgn)

JK1.rk.dsgn <- calibrate(design = JK1.dsgn.resp, formula = ~xvar1 +
            xvar2, population = c('(Intercept)'=nrow(popinfo$pop),
            xvar12=sum(popinfo$totals.xvar1xvar2[2,]),
            xvar22=sum(popinfo$totals.xvar1xvar2[,2])),
            calfun="raking")
JK1.rk.wgt <- weights(JK1.rk.dsgn)

JK1.greg.dsgn <- calibrate(design = JK1.dsgn.resp, formula = ~xvar1 +
            xvar2, population = c('(Intercept)'=nrow(popinfo$pop),
            xvar12=sum(popinfo$totals.xvar1xvar2[2,]),
            xvar22=sum(popinfo$totals.xvar1xvar2[,2])),
            calfun="linear")
JK1.greg.wgt <- weights(JK1.greg.dsgn)

######################
# Summary statistics
######################
# Total
pop.total <- sum(popinfo$pop[,"y"])
rk.total <- rk.total <- svytotal(~y, TS.rk.dsgn)

rk.Nrc <- svytable(~xvar1 + xvar2, TS.rk.dsgn)
rk.Diff.Nrc = rk.Nrc - popinfo$totals.xvar1xvar2

#### TS approach ####
TS.rk.total.se <- SE(svytotal(~y, TS.rk.dsgn))           # SE using
            Lumley TS approach
TS.rk.total.CI <- confint(svytotal(~y, TS.rk.dsgn))          # CI
            using Lumley TS approach
TS.rk.total.CI.coverage <- ifelse(TS.rk.total.CI[1]<=pop.total &
            pop.total<=TS.rk.total.CI[2], 1, 0)     # CI coverage
            using Lumley TS approach
TS.rk.Distance <- (rk.Diff.Nrc[1,1]/SE(svytotal(~x11, TS.rk.dsgn)))^2
            # Distance measure Lumley TS approach
```

```
#### JK1 approach ####
JK1.rk.total.se <- SE(svytotal(~y, JK1.rk.dsgn))         # SE using
            JK1 approach
JK1.rk.total.CI <- confint(svytotal(~y, JK1.rk.dsgn))          # CI
            using JK1 approach
JK1.rk.total.CI.coverage <- ifelse(JK1.rk.total.CI[1]<=pop.total &
            pop.total<=JK1.rk.total.CI[2], 1, 0)     # CI coverage
            using JK1 approach
JK1.rk.Distance <- (rk.Diff.Nrc[1,1]/SE(svytotal(~x11,
            JK1.rk.dsgn)))^2   # Distance measrue JK1 approach

#### SEs using four different approaches in DArrigo and Skinner ####
# obtain residuals #
resp.resid <- data.frame(respinfo$resp.smp, TS.rk.wgt)

# residuals from regression model using base weights
bwgt.reg.resid <- residuals(svyglm(y~xvar1+xvar2,
            design=TS.dsgn.resp))

# residuals from regression model using raked weights
rkwgt.reg.resid <- residuals(svyglm(y~xvar1+xvar2,
            design=(svydesign(ids = ~0, # No cluster

            strata = NULL, # No strata

            # fpc = ~f1,

            weights = ~TS.rk.wgt,

            data = resp.resid))))

# Create design objects #
resp.resid.wgt <- data.frame(resp.resid, bwgt.reg.resid,
            rkwgt.reg.resid)

# Design object with base weights (to be used for weighting
            residuals)
bwgt.resid.dsgn <- svydesign(ids = ~0, # No cluster
                             strata = NULL, # No strata
                             # fpc = ~f1,
                             weights = ~bwgt,
                             data = resp.resid.wgt)

# Design object with raked weights (to be used for weighting
            residuals)
rkwgt.resid.dsgn <- svydesign(ids = ~0, # No cluster
                              strata = NULL, # No strata
                              # fpc = ~f1,
                              weights = ~TS.rk.wgt,
                              data = resp.resid.wgt)

# base weights for weighting residuals & base weights for regression
            model
Bresid.Breg.rk.total.se <- SE(svytotal(~bwgt.reg.resid,
            bwgt.resid.dsgn))
```

```
  # base weights for weighting residuals & raked weights for regression
              model
  Bresid.RKreg.rk.total.se <- SE(svytotal(~rkwgt.reg.resid,
              bwgt.resid.dsgn))

  # raked weights for weighting residuals & base weights for regression
              model
  RKresid.Breg.rk.total.se <- SE(svytotal(~bwgt.reg.resid,
              rkwgt.resid.dsgn))

  # raked weights for weighting residuals & raked weights for
              regression model
  RKresid.RKreg.rk.total.se <- SE(svytotal(~rkwgt.reg.resid,
              rkwgt.resid.dsgn))

  # vector to return, for estimates of totals
  results.total <- vector(length=19)

  results.total[1] <- pop.total
  results.total[2] <- rk.total
  results.total[3] <- TS.rk.total.se
  results.total[4] <- TS.rk.total.CI.coverage
  results.total[5] <- TS.rk.Distance

  results.total[6] <- JK1.rk.total.se
  results.total[7] <- JK1.rk.total.CI.coverage
  results.total[8] <- JK1.rk.Distance

  results.total[9] <- Bresid.Breg.rk.total.se
  results.total[10] <- Bresid.RKreg.rk.total.se
  results.total[11] <- RKresid.Breg.rk.total.se
  results.total[12] <- RKresid.RKreg.rk.total.se

  results.total[13] <- mean(bwgt.reg.resid)
  results.total[14] <- var(bwgt.reg.resid)
  results.total[15] <- mean(bwgt.reg.resid^2)

  results.total[16] <- mean(rkwgt.reg.resid)
  results.total[17] <- var(rkwgt.reg.resid)
  results.total[18] <- mean(rkwgt.reg.resid^2)

  results.total[19] <- rk.Diff.Nrc[1,1]
  return (t(results.total))
}

#########################################################################
# Function for calling srs.smp, resp.smp, calib during each simulation
#########################################################################

srs.resp.calib <- function (benchmark, k, srs.size, repnum){

  S <- k      # number of good samples to keep
  s <- 1
  bad.smp <- 0

  # An empty matrix to store results.
  rslt <- matrix(nrow=S, ncol=19)
```

```
 colnames(rslt) <- c("pop.total",
                     "rk.total",

                     "TS.rk.total.se",
                     "TS.rk.total.CI.coverage",
                     "TS.rk.Distance",

                     "JK1.rk.total.se",
                     "JK1.rk.total.CI.coverage",
                     "JK1.rk.Distance",

                     "Bresid.Breg.rk.total.se",
                     "Bresid.RKreg.rk.total.se",
                     "RKresid.Breg.rk.total.se",
                     "RKresid.RKreg.rk.total.se",

                     "mean.bwgt.reg.resid",
                     "var.bwgt.reg.resid",
                     "mean.bwgt.reg.resid.squared",

                     "mean.rkwgt.reg.resid",
                     "var.rkwgt.reg.resid",
                     "mean.rkwgt.reg.resid.squared",
                     "rk.Diff.Nrc11")

 while (s <= S){

 keep.sw <- TRUE

 # draw srs sample and respondent sample
 srssmp <- srs.smp(popdata=benchmark$pop, n=srs.size, repnum=repnum)
 if (srssmp$srs.bad==TRUE){
   bad.smp <- bad.smp + 1
   keep.sw <- FALSE
 }
 else {
   # assign respondent
   respsmp <- resp.smp (srsdata=srssmp$srs.smp,
              TS.dsgn=srssmp$TS.dsgn, JK1.dsgn=srssmp$JK1.dsgn)
 }

 if (respsmp$resp.bad==TRUE){
   bad.smp <- bad.smp + 1
   keep.sw <- FALSE
 }

 else {

 # calibration and save summary statisticis
 rslt[s, ] <- calib(respinfo=respsmp, popinfo=benchmark,
              srsinfo=srssmp, TS.dsgn.resp=respsmp$TS.dsgn.resp,
              JK1.dsgn.resp=respsmp$JK1.dsgn.resp)

 # increase sample counter
 s <- s + 1
 }
}
```

```
 return (list(bad.smp=bad.smp, rslt=rslt))
}
```

B.7 A Program in Chapter 5 for Calling the Program That is Adopted from the Program in B.1 with the Function in B.6.

```
library(sampling)
library(survey)

#####################################################################
# Function for creating measures from the S good samples
#####################################################################

all.info <- function(datain){

  datain <- data.frame(datain)
  attach(datain)

  # Total: relative bias
  rk.rel.bias.total <- (rk.total - pop.total)/pop.total

  # Total: relative square root of mse
  rk.rel.sqrt.mse.total <- sqrt((rk.total - pop.total)^2)/pop.total

  # Total: bias ratio or t-statitics using TS
  TS.rk.bias.ratio.total = (rk.total - pop.total) / TS.rk.total.se

  # Total: bias ratio or t-statitics using JK1
  JK1.rk.bias.ratio.total = (rk.total - pop.total) / JK1.rk.total.se

  # Distance Measure using empirical variance from simulation, for
              Chapter 4
  Emp.rk.Distance <- rk.Diff.Nrc11^2/var(rk.Diff.Nrc11)

  TS.rk.Distance.LG384 <- ifelse(TS.rk.Distance>3.84, 1, 0)
  TS.rk.Distance.LG663 <- ifelse(TS.rk.Distance>6.63, 1, 0)

  JK1.rk.Distance.LG384 <- ifelse(JK1.rk.Distance>3.84, 1, 0)
  JK1.rk.Distance.LG663 <- ifelse(JK1.rk.Distance>6.63, 1, 0)

  datafinal <- data.frame(  datain,
                            rk.rel.bias.total,
                            rk.rel.sqrt.mse.total,
                            TS.rk.bias.ratio.total,
                            JK1.rk.bias.ratio.total,
                            Emp.rk.Distance,
                            TS.rk.Distance.LG384,
                            TS.rk.Distance.LG663,
                            JK1.rk.Distance.LG384,
                            JK1.rk.Distance.LG663)

  detach(datain)
```

```
  return(datafinal)
}

#########################################################
# Function for generating overall summary statitics
#########################################################

overall <- function (datain){

  attach(datain)

  # For total: relative bias, relative standard error, relative square
              root of mse, bias ratio
  pop.total <- mean(pop.total)

  rk.rel.bias.total <- mean(rk.rel.bias.total)   # relative bias

  rk.rel.sqrt.mse.total <- mean(rk.rel.sqrt.mse.total)  # MSE

  Emp.rk.rel.se.total <- sqrt(var(rk.total))/pop.total      # SE's
  TS.rk.rel.se.total <- mean(TS.rk.total.se)/pop.total
  JK1.rk.rel.se.total <- mean(JK1.rk.total.se)/pop.total

  TS.rk.bias.ratio.total <- mean(TS.rk.bias.ratio.total)     # Bias
              ratios
  JK1.rk.bias.ratio.total <- mean(JK1.rk.bias.ratio.total)

  TS.rk.CI.coverage.total <- mean(TS.rk.total.CI.coverage)   # CI
              coverage
  JK1.rk.CI.coverage.total <- mean(JK1.rk.total.CI.coverage)

  Emp.rk.Distance <- mean(Emp.rk.Distance)   # Distance measure and
              extreme values
  TS.rk.Distance <- mean(TS.rk.Distance)
  JK1.rk.Distance <- mean(JK1.rk.Distance)

  TS.rk.Distance.LG384 <- mean(TS.rk.Distance.LG384)
  JK1.rk.Distance.LG384 <- mean(JK1.rk.Distance.LG384)

  TS.rk.Distance.LG663 <- mean(TS.rk.Distance.LG663)
  JK1.rk.Distance.LG663 <- mean(JK1.rk.Distance.LG663)

  ########## Four SEs based on DArrigo and Skinner and diagnostics for
              residuals from regression model
  # Using base weights on residuals
  Bresid.Breg.rk.rel.se.total <-
              mean(Bresid.Breg.rk.total.se)/pop.total
  Bresid.RKreg.rk.rel.se.total <-
              mean(Bresid.RKreg.rk.total.se)/pop.total

  # Using raked weights on residuals
  RKresid.Breg.rk.rel.se.total <-
              mean(RKresid.Breg.rk.total.se)/pop.total
  RKresid.RKreg.rk.rel.se.total <-
              mean(RKresid.RKreg.rk.total.se)/pop.total

  # Mean of residuals from regression model
```

```
  mean.bwgt.reg.resid <- mean(mean.bwgt.reg.resid)
  mean.rkwgt.reg.resid <- mean(mean.rkwgt.reg.resid)

  # Variance of residuals from regression model
  var.bwgt.reg.resid <- mean(var.bwgt.reg.resid)
  var.rkwgt.reg.resid <- mean(var.rkwgt.reg.resid)

  # Mean of squared residuals from regression model (we calculate this
              because we are not sure if the residuals would sum up to
              zero)
  mean.bwgt.reg.resid.squared <- mean(mean.bwgt.reg.resid.squared)
  mean.rkwgt.reg.resid.squared <- mean(mean.rkwgt.reg.resid.squared)

  detach(datain)

  ###################################################
  evaltotal <- cbind (pop.total, rk.rel.bias.total,
              rk.rel.sqrt.mse.total,
                     Emp.rk.rel.se.total, TS.rk.rel.se.total,
              JK1.rk.rel.se.total,
                     TS.rk.bias.ratio.total, JK1.rk.bias.ratio.total,
                     TS.rk.CI.coverage.total,
              JK1.rk.CI.coverage.total,
                     Emp.rk.Distance, TS.rk.Distance, JK1.rk.Distance,
                     TS.rk.Distance.LG384, JK1.rk.Distance.LG384,
                     TS.rk.Distance.LG663, JK1.rk.Distance.LG663,
                     Bresid.Breg.rk.rel.se.total,
              Bresid.RKreg.rk.rel.se.total,
                     RKresid.Breg.rk.rel.se.total,
              RKresid.RKreg.rk.rel.se.total,
                     mean.bwgt.reg.resid, mean.rkwgt.reg.resid,
                     var.bwgt.reg.resid, var.rkwgt.reg.resid,
                     mean.bwgt.reg.resid.squared,
                     mean.rkwgt.reg.resid.squared)

  return(evaltotal)
}
```

References

Agresti, A. (2013). *Categorical Data Analysis, 3rd Edition*. John Wiley & Sons: New York.

Brick, J.M, Montaquila, J., and Roth, S (2003). Identifying problems with raking estimators. *Proceedings of the Section on Survey Research Methods, American Statistical Association,* 710-717.

Chang, T., and Kott, P. (2008). Using Calibration Weighting to Adjust for Nonresponse under a Plausible Model. *Biometrika,* 95, 555-571.

D'Arrigo, J., and Skinner, C. (2010). Linearization Variance Estimation for Generalized Raking Estimators in the Presence of Nonresponse. *Survey Methodology*, 36, 181-192.

Deming, W.E. (1943). *Statistical Adjustment of Data*. John Wiley & Sons: New York.

Dever, J. (2008). Sampling Weight Calibration with Estimated Control Totals. (Dissertation)

Dever, J., and Valliant, R. (2010). A Comparison of Variance Estimators for Poststratification to Estimated Control Totals. *Survey Methodology*, 36, 45-56.

Deville, J-C., and Särndal, C-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87(418), 376-382.

Deville, J-C., Särndal, C-E., and Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, 88(423), 1013-1020.

Deville, J-C. (1999). Variance Estimation for Complex Statistics and Estimators: Linearization and Residual Techniques. *Survey Methodology,* 25, 193-203.

Estevao, V. M., and Särndal, C-E. (2000). A Functional Form Approach to Calibration. *Journal of Official Statistics*, 16, 379-399.

Estevao, V.M., and Särndal, C.E. (2006). Survey Estimates by Calibration on Complex Auxiliary Information. *International Statistical Review*, 74, 127-147.

Fuller, W. A. (2000). Two-phase Sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 23-30.

Haberman, S. (1979). *Analysis of Qualitative Data.* Academic Press: New York.

Kalton G., and Flores-Cervantes, I. (2003). Weighting Methods. *Journal of Official Statistics*, 19, 81-97.

Kott, P.S. (1994). A note on Handling Nonresponse in Surveys. *Journal of the American Statistical Association*, 89, 693-696.

Kott, P. S. (2006). Using Calibration Weighting to Adjust for Nonresponse and Coverage Errors. *Survey Methodology*, 32, 133-142.

Kott, P. S., and Chang, T. (2010). Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse. *Journal of the American Statistical Association*, 105(491), 1265-1275.

Kott, P. S., and Liao, D. (2012). Providing Double Protection for Unit Nonresponse with a Nonlinear Calibration-Weighting Routine. *Survey Research Methods*, 6, 105-111.

Kreuter, F. (2013). *Improving Surveys with Paradata: Analytic Uses of Process Information.* Hoboken, NJ: Wiley.

Kreuter, F., Olsen, K., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casa-Cordero, C., Raghunathan, T.E. (2010). Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust for Non-response: Examples from Multiple Surveys. *Journal of the Royal Statistical Society*, Series A, 173, 389-407.

Krewski and Rao (1981). Inference from Stratified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods. *The Annals of Statistics*, 9(5).

Lesage, E., Haziza, D., and D'Hautlfoeuille, X. (2016). A cautionary tale on instrument vector calibration for the treatment of unit nonresponse in surveys. In revision for *Journal of the American Statistical Association. (The following electronic copy of the paper was found and last accessed on April 27, 2017: http://www.crest.fr/ckfinder/userfiles/files/Pageperso/xdhaultfoeuille/DH_H_L.pdf.)*

Little, R. J. A. (1993). Post-Stratification: A Modeler's Perspective. *Journal of the American Statistical Association*, 88, 1001-1012.

Little, R. J. A., and Vartivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology*, 31, 161-168.

Little, R. J. A., and Wu, M. M. (1991). Models for Contingency Tables with Known Margins When Target and Sample Populations Differ. *Journal of the American Statistical Association*, 86, 87-95.

Lumley, T. (2010). *Complex Surveys: A Guide to Analysis Using R.* John Wiley & Sons: New Jersey.

Lundström, S., and Särndal, C-E. (1999). Calibration as a Standard Method for Treatment of Nonresponse. *Journal of Official Statistics*, 15, 305-327

R Development Core Team (2015). *R: A Language and Environment for Statistical Computing.* R Development Core Team, Vienna, Austria. URL http://www.R-project.org

Rao, J.N.K. and Scott, A.J. (1981). The Analysis of Categorical Data from Complex Sample Surveys: Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables. *Journal of the American Statistical Association*, 76, 221-230.

Särndal, C-E. (2007). The Calibration Approach in Survey Theory and Practice. *Survey Methodology*, 33, 99-119.

Särndal, C-E. and Lundström, S. (2005). *Estimation in Surveys with Nonresponse*. John Wiley & Sons: Chichester.

Särndal, C-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer.

Searle, S. R. (1971). *Linear Models*. John Wiley & Sons: New York.

Shao, J. (1996). Resampling Methods in Sample Surveys (with Discussion. *Statistics*, 27, 203-254.

Singh, A.C., and Mohl, C.A. (1996). Understanding Calibration Estimators in Survey Sampling. *Survey Methodology*, 22, 107-115.

Slud, E. and Thibaudeau, Y. (2009). Simultaneous Calibration and Nonresponse Adjustment. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 2263-2272.

Stukel, D.M., Hidiroglou, M.A. and Särndal, C-E. (1996). Variance Estimation for Calibration Estimators: A comparison of Jackknifing versus Taylor Linearization. *Survey Methodology*, 22, 117-125.

Valliant, R., Dever, J., and Kreuter, F. (2013). *Practical Tools for Designing and Weighting Survey Samples*. Springer: New York.

Valliant, R., Dorfman, A.H., and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. John Wiley & Sons: New York.

WesVar 4.0 User's Guide (Appendix A). Rockville, MD: Westat. Anonymous. (2006).

Wolter, K.M. (2007). *Introduction to Variance Estimation, 2nd Edition*. New York: Springer-Verlag.