ABSTRACT

Title of Dissertation: MISSPECIFIED WEIGHTS IN WEIGHT-SMOOTHING METHODS
Xia Li, Doctor of Philosophy, 2018

Directed by: Professor Eric V. Slud
Department of Mathematics, Statistics Program

Misspecification happens for various reasons in weight adjustment procedures in survey data analysis. To study the consequences of weight misspecifications, we study the effects of using a multiplicative biasing factor to describe the weight adjustments and reflect the distributional change from design/initial weights to final weights. The necessary and sufficient condition of the Horvitz-Thompson (HT) estimator of a population total being consistent is then given in a superpopulation setting. When HT is consistent, we first investigate the bias in other estimators for population totals. We show the necessary condition for bias in Generalized Regression (GREG) estimator and the resulting bias formula in the superpopulation limiting sense. We also link the bias in a model-based estimator of Zheng and Little to the failure of extrapolated model-fitting outside the sample. Both findings are validated in simulation studies. Next we find that the biasing factor affects estimators so that one particular estimator may have the smallest variance under design weights but not under misspecified weights due to variance inflation. A preliminary

analysis on simulated samples drawn from a population of real American Community Survey (ACS) data illustrates the quality of fit of the biasing factor model we proposed to the ACS data with weights modified by a few calibration/raking steps.

MISSPECIFIED WEIGHTS IN WEIGHT-SMOOTHING METHODS

By

Xia Li

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
2018

Advisory Committee:
Professor Eric V. Slud, Chair/Advisor
Professor Maria Cameron
Professor Michael Fu
Dr. Barry Graubard
Professor Benjamin Kedem

Dedication

To my father Dr. Youxin Li, my mother Ms. Qiongying Wu, and my loving husband Yue Du.

Acknowledgments

Completing doctoral study and writing this thesis has been a tough task for me. I owe my gratitude to all the people who have made this thesis possible and because of whom my graduate experience has been one that I will cherish forever.

First, I would like to express my most sincere gratitude to my advisor, Professor Eric V. Slud, for his vision, advice and patience to help me proceed through my graduate studies. The completion of this thesis would not have been possible without his consistent guidance and support. He has always made himself available for help and advice. It has been a pleasure to work with and learn from such an extraordinary individual.

Special thanks to my other dissertation committee members, Professor Benjamin Kedem, Professor Maria Cameron, Professor Michael Fu and Dr. Barry Graubard, for taking the time to read the thesis, attend my final oral exam and provide valuable comments.

I would like to thank Professor Paul J. Smith and Professor Abram Kagan as well, who have been always willing to help me.

I would like to thank my peers within the department of Mathematics. Particularly I would like to thank Xuan Yao, who set a great role model for me and always encouraged me and struggled with me in the last few months. Thanks to my friends Ying Han and Yimei Fan for their best company and support.

I owe my deepest thanks to my family - my father and mother who have always trusted me, stood by me, and been proud of me.

Most importantly, doing the research and writing the thesis would be impossible without the support and understanding from my loving and supportive husband, Yue.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACS	American community survey
AV	Anticipated variance
CI	Confidence interval
GREG	Generalized regression estimator
HT	Horvitz-Thompson estimator
iid	Independent and identically distributed
PS	Pfeffermann-Sverchkov's estimator
PPS	Probability-proportional-to-size sampling
PPSWOR	Probability-proportional-to-size sampling without replacement
PSU	Primary subsampling unit
SRSWOR	Simple random sampling without replacement
SSU	Secondary subsampling unit
ZL	Zheng and Little's estimators

Chapter 1. Introduction

1.1 General introduction

Sample surveys are useful tools for understanding population characteristics such as quantitative information about economy and society in academic research, business decision making, and government planning. The population that we are interested in understanding is often called the *target population*. The population regarding which we have a whole list from which we could design the sampling scheme is often called *sampling frame*. In general, survey samples can be divided into two types: probability samples and non-probability samples. In this dissertation, we only focus on probability samples.

Sampling weights, or *survey weights* are important in constructing *unbiased* estimators with the collected the sample. One important example is the Horvitz-Thompson (HT) estimator, used to estimate population total of a measured attribute, which will be introduced in Section 1.1. Without using sampling weights, the estimates may reflect only nuances of a particular sample and may contain significant levels of bias. Ideally, the weight of a unit, which is a positive value associated with the unit in the sample, should be the size of the population subgroup that the

sampling unit represents in the target population.

Sampling weights are usually calculated or constructed in the following ways. 1). After the careful design of the sampling procedure, we can often calculate the *inclusion probabilities* for the sampling units. The *design weights*, or *initial weights* are defined as reciprocals of the inclusion probabilities. After collecting the sample, all the design weights usually are known within the sample. 2). After design weights are computed, *weight adjustments* are usually necessary due to *nonresponse*, the correction of frame deficiencies, and techniques (such as weight-trimming) used to reduce variances of estimators. 3). Those mentioned adjusting steps could be repeated if necessary. 4). If properly adjusted, the resulting sets of weights, called *final weights*, would enter into the analysis stage and be used to construct many different population quantities of interest.

Weight adjustments generally rely on extra model assumptions and external information. For example in adjusting for nonresponse, we have to assume a response-propensity model and in matching the population totals of covariates, we have to import extraneous information that we believe is reliable and accurate. The published literature explains, when those assumed models and imported information are truely reliable, the resulting set of weights will bring benefits such as reducing variance of estimators. We call such resulting set of weights *properly adjusted* sampling weights. However, it is important also to investigate the consequences when any of the model assumptions or exogenous information are not correct. For example, we may use a wrong response-propensity model that we believe to be correct, or use wrong totals that we consider to be accurate. In such a case we refer to the

weights are *misspecified*, or *inappropriately adjusted*. Whether such a set of weights could still produce design-unbiased estimators, or estimators with smaller variances, remain to be studied carefully.

In this dissertation, we will focus on six estimators, some of which are *design-based* and some of which are *model-based*. Therefore, it is also worth introducing the ideas of and the differences between design-based and model-based approaches. The design-based approach views all collected information including measurements of interest and auxiliary information as forming a big array of constants. When investigating biases and variances of the estimators, we view the randomnesses as coming only from whether the unit is sampled or not and take expectation with respect to the sample design. In this approach, a probability sample must be selected. In the model-based approach, we may view the collected information associated with each unit as forming a vector, which is a realization of an unknown underlying stochastic mechanism. We consider the population distributional structure in deciding on an estimator. This approach can be applied to either probability or non-probability samples. In this dissertation, the model-based methods we have considered are all applied to probability samples.

1.2 Overview

We consider probability sampling designs over a finite set of elements labeled by integers $\mathcal{U} = \{1, 2, \ldots, N\}$ with N being the finite population size. Let vector Y_i be the variable of interest associated with unit $i \in \mathcal{U}$. The set $\{Y_i\}_{i=1}^{N}$ is denoted by

$\mathcal{F}$, called a *finite population*. To simplify, we assume Y_i's are real numbers in this prospectus. Besides Y_i's, some auxiliary information, denoted by covariate column vector X_i, is associated with unit i. From now on let us define $\mathcal{F} = \{(Y_i, X_i)\}_{i \in \mathcal{U}}$.

We consider probability sampling designs where a random sample $\mathcal{S}$, a set of selected labels, is drawn from the finite population $\mathcal{F}$ according to the inclusion probabilities $\pi_i^0 = P(i \in \mathcal{S})$, for $i = 1, \ldots, N$. Let $I_{[i \in \mathcal{S}]}$ be the indicator variable for unit i, defined as below.

$$I_{[i \in \mathcal{S}]} = \begin{cases} 1 & \text{if unit } i \text{ is included in the sample } \mathcal{S}, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose the main interest of the sample survey is to estimate the population total of the outcome variable Y_i, defined as $t_Y = \sum_{i=1}^{N} Y_i$. The example of Y-total is not fully general, but many other parameters of interest can be expressed as functions of one or more attribute totals.

1.3 Superpopulation and pseudo-random samples

Design-based approaches view Y_i and X_i as elements of a big array and rows of a big matrix when constructing estimator of finite population parameter θ, say $\hat{\theta}$. The only randomness in the estimator $\hat{\theta}$ is viewed in this approach to come from $I_{[i \in \mathcal{S}]}$. Define π_i^0 be the inclusion probability with $P(i \in \mathcal{S}) = \pi_i^0$. It is sometimes sufficient to discuss *design-bias* and *design-variance* of estimator $\hat{\theta}$.

Another type of approach is called *model-based*. It is also common to for-

mulate the survey data analysis into a statistical parameter estimation problem. Commonly used statistical models include linear regression models and generalized linear models. Then it would be natural to ask questions about consistency of estimates of the model parameters and finite population parameters as well. This leads to the *superpopulation* set up. We may consider the finite population $\mathcal{F}$ to be generated by a unknown hypothetical underlying stochastic mechanism. For example, we consider $\mathcal{F} = \{(Y_i,\, X_i, \pi_i^0)\}_{i=1}^N$ *independently identically distributed* (*iid*) with a joint distribution function G, where π_i^0 could be viewed as depending on X_i (and possibly on other independent random variables) and therefore is random through X_i. The sample $(Y_i,\, X_i)_{i\in\mathcal{S}}$ then is drawn from the finite population $\mathcal{F}$. The unit i is included in the sample with probability π_i^0. Mathematical derivations on consistency, limiting distributions under this approach could be developed based on specific assumptions on G. We are considering *iid* superpopulation samples in this thesis, but it is worth mentioning that sometimes non-*iid* superpopulation samples are more practical for example in cluster sampling setting. Discussions could be found in Korn and Graubard (1998); Graubard and Korn (2002).

Although it often makes sense to view $(Y_i,\, X_i,\, \pi_i^0)$ as *iid* samples from a hypothetical distribution G, we may want to add more elements into $(Y_i,\, X_i)$ to incorporate other probabilistic mechanisms such as nonresponse or calibration. We then introduce the *pseudo-random* variable idea to represent all other probability procedures done after we collect the sampled units. For example, following the idea of Oh and Scheuren (1983), each unit $i \in \mathcal{F}$ is associated with a random variable R_i which is 1 if the unit would respond to the survey if sampled and 0 otherwise.

We would not see R_i for $i \notin \mathcal{S}$ just as we do not observe (Y_i, X_i) for $i \notin \mathcal{S}$. But hypothetically, we view R_i as a pre-generated random element before the probability sampling procedure and therefore R_i could be added into (Y_i, X_i) and $\mathcal{F}$ could be written as $\{(Y_i, X_i, R_i)\}_{i=1}^N$. All other probability mechanisms like calibration could be viewed as such pseudo-random variables and therefore are part of superpopulation *iid* samples. Later in Chapter 2 we will introduce the weight biasing factor incorporating all probability procedures related to weight modifications and/or adjustments. By borrowing the idea of pseudo-random survey variables, such a biasing factor could be viewed as a feature of superpopulation *iid* samples as well.

1.4 Brief introduction of six estimators considered

There are six estimators of t_Y considered in this thesis. All estimators introduced in this section have superscripts d, which implies that design weights $d_i = 1/\pi_i^0$ have been used. Later when we introduce the weight misspecification idea, we will use superscripts w to denote the estimators under misspecified weights.

- Horvitz-Thompson (HT) estimator is a design-unbiased estimator of the finite population total with unequal probabilities of inclusion, defined as

$$\hat{t}_Y^{HT,d} = \sum_{i \in \mathcal{S}} \frac{Y_i}{\pi_i^0}.$$

  Here *design-unbiased* is defined as

$$E_d\left(\hat{t}_Y^{HT,d} \,\middle|\, \mathcal{F}\right) = t_Y$$

where $E_d(\cdot|\,\mathcal{F})$, the design expectation, denotes the average over all possible samples under the design for the specific finite population $\mathcal{F}$ (Isaki and Fuller 1982; Fuller 2011). As long as we have $\pi_i^0 > 0$ for all $i \in \mathcal{U}$, this inverse probability weighting approach would provide unbiased estimation of the population total regardless of the sampling scheme adopted. Moreover, this universal unbiased property does not depend on any distributional model assumption of survey measurement Y_i and covariates X_i. When the outcome variable and the inclusion probability are weakly linearly correlated, the HT estimator could be very inefficient, i.e., could have large variance.

- Generalized regression estimator (GREG) (Särndal et al. 1992; Fuller 2002), given by (1.1), is *design-consistent* utilizing the association between covariate X_i and outcome Y_i when the total $t_X = \sum_{\mathcal{U}} X_i$ is known:

$$\hat{t}_Y^{GREG,\,d} = (N,\, t_X^{tr})\hat{\beta} \;=\; \hat{t}_Y^{HT,\,d} + \left(\begin{pmatrix} N \\ t_X \end{pmatrix} - \begin{pmatrix} \sum_{i\in\mathcal{S}} 1/\pi_i^0 \\ \hat{t}_X^{HT} \end{pmatrix} \right)^{tr} \hat{\beta} \qquad (1.1)$$

where $\hat{t}_X^{HT,\,d} = \sum_{i\in\mathcal{S}} X_i/\pi_i^0$ and

$$\hat{\beta} = \left(\sum_{\mathcal{S}} \frac{1}{\pi_i^0} \begin{pmatrix} 1 \\ X_i \end{pmatrix}^{\otimes 2} \right)^{-1} \left(\sum_{\mathcal{S}} \frac{1}{\pi_i^0} \begin{pmatrix} 1 \\ X_i \end{pmatrix} Y_i \right),$$

and the inverse is assumed to exist. Here the operator "$\otimes 2$" is defined as $\mathbf{x}^{\otimes 2} = \mathbf{x}\mathbf{x}'$. Following Fuller's definition (Fuller 2011), *design-consistency* of $\hat{t}_Y^{GREG}$ indicates that given a sequence of increasingly large finite populations $\{\mathcal{F}_N\}$ and an associated sequence of sample designs with increasing sample

size, for every $\epsilon > 0$,

$$\lim_{N\to\infty} P_d\left\{|\hat{t}_Y^{GREG} - t_Y| > \epsilon \middle| \mathcal{F}_N\right\} = 0,\ a.s.,$$

where the notation means that we condition on the realized finite population $\mathcal{F}_N$ and the probability $P_d\{\cdot \mid \mathcal{F}_N\}$ is with respect to the sample design, similar to E_d.

- Zheng and Little's methods (ZL). Zheng and Little (2003) considered smoothing the outcome variable by modeling Y_i against the inclusion probability using a p-spline function, defined in (1.2). One has to decide how delicate the p-spline model is by choosing the degree p and number of knots m. The exponent k also needs to be decided and commonly used values are $k = 0, 1/2$ or 1. Zheng and Little suggested using random-effect terms as coefficients $\gamma_{p+1}, \ldots, \gamma_{p+m}$. To simplify, only fixed effects were considered in this study.

$$\begin{aligned} Y_i &= \gamma_0 + \sum_{j=1}^{p} \gamma_j (\pi_i^0)^j + \sum_{l=1}^{m} \gamma_{p+l} (\pi_i^0 - \kappa_l)_+^p + \varepsilon_i, \\ &\text{where } \varepsilon_i \overset{iid}{\sim} \mathcal{N}\left(0, (\pi_i^0)^{2k}\sigma_\varepsilon^2\right). \end{aligned} \tag{1.2}$$

Let $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_{p+m})^{tr}$. After weighted least-squares estimates $\hat{\gamma}$ are obtained using (1.2), predicted Y_i values denoted by $\hat{Y}_i^{ZL,d}$, are given by

$$\hat{Y}_i^{ZL,d} = \hat{\gamma}_0 + \sum_{j=1}^{p} \hat{\gamma}_j (\pi_i^0)^j + \sum_{l=1}^{m} \hat{\gamma}_{p+l} (\pi_i^0 - \kappa_l)_+^p.$$

If π_i^0's are only known for the sample $\mathcal{S}$, the estimated population total is given by

$$\hat{t}_Y^{ZL1,\,d} = \sum_{i\in\mathcal{S}} \hat{Y}_i^{ZL,\,d}/\pi_i^0. \tag{1.3}$$

If π_i^0's are known for the whole population, the estimated population total could be given by

$$\hat{t}_Y^{ZL2,\,d} = \sum_{i\in\mathcal{S}} Y_i + \sum_{i\notin\mathcal{S}} \hat{Y}_i^{ZL,\,d}. \tag{1.4}$$

- Pfeffermann-Sverchkov's method (PS). Pfeffermann and Sverchkov (1999) considered smoothing the weights by a function of covariates and then applying the weighted least squares. We consider only least-squares estimation, which is semi-parametric in the sense that the residuals $Y_i - (1,\, X_i^{tr})\beta$ are assumed to have mean zero but are not otherwise distributionally restricted. The estimated coefficient vector has the form

$$\hat{\beta}^{PS,\,d} = \left(\sum_{i\in\mathcal{S}} \frac{d_i}{\tilde{d}_i}\begin{pmatrix}1\\ X_i\end{pmatrix}^{\otimes 2}\right)^{-1}\left(\sum_{i\in\mathcal{S}} \frac{d_i}{\tilde{d}_i}\begin{pmatrix}1\\ X_i\end{pmatrix}Y_i\right), \tag{1.5}$$

where the operator "$\otimes 2$" is defined as $\mathbf{x}^{\otimes 2} = \mathbf{x}\mathbf{x}^{tr}$ for column vector $\mathbf{x}$ and $\tilde{d}_i$ is obtained as the predictor of d_i from X_i under the regression model

$$\log(d_i - 1) = (1, X_i)^{tr}\beta \,+\, \epsilon_i, \qquad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

Here $d_i = 1/\pi_i^0$.

- Beaumont's method (B). Beaumont (2008) dealt with the inefficiency of the

HT estimator by smoothing weights against the outcome variable. Estimated smoothed weights $\hat{d}_i$ are obtained by a least-squares regression of $\log(d_i - 1)$ on vector h_i which is a known function of Y_i. Then the smoothed estimator of population total could be given by

$$\hat{t}_Y^B = \sum_{i \in \mathcal{S}} \hat{d}_i Y_i. \tag{1.6}$$

1.5 Brief introduction of sampling methods considered

Assume that all inclusion probabilities π_i^0's have been designed and computed. A practical issue in the simulation study is how to draw the sample S such that

$$P\left(i \in \mathcal{S}\right) \, = \, \pi_i^0.$$

Assume that we are interested in probability-proportional-to-size sampling without replacement (PPSWOR).

- One quick solution *Poisson sampling*, sampling procedure in which each element $i \in \mathcal{U}$ is chosen for inclusion in the sample S according to an independent Bernoulli trial. A detailed introduction was given in Section 3.2 of Särndal et al. (1992). In Poisson sampling,

$$I_{[i \in \mathcal{S}]} \overset{iid}{\sim} Bernoulli(\pi_i^0), \; i = 1, \ldots, N.$$

Since $I_{[i \in \mathcal{S}]}$'s are independent, the joint inclusion probability π_{ij}^0 for distinct units i, j is simply $\pi_i^0 \times \pi_j^0$. This simplifies some calculations in variance estimation. The disadvantage of Poisson sampling is that the actual sample size is random with expectation $\sum_{i=1}^N \pi_i^0$.

- Another easy method is *systematic sampling* (Madow 1949), available in R function `UPsystematic` in package `sampling` (Tillé and Matei 2016). When it is used in unequal probability sampling, rather than simply counting through elements of the population and selecting every k^{th} unit, each element is allocated a segment along a number line according to its inclusion probability. Then a random starting point from $Unif(0, 1)$ is generated. We then move along the number line in steps of 1 and select those elements into whose segments the successive steps fall.

- The third solution is called *rejective sampling*, proposed and summarized in Rao (1963), Carroll and Hartley (1964), and Hájek (1964). The idea is that we compute a probability vector a_i, $i = 1, \ldots, N$ and perform Poisson sampling. Samples with size not equal to n would be rejected so that we get a sample of size n exactly. Probability vector a_i is calculated beforehand, such that the overall inclusion probability is exactly π_i^0 for unit i.

- Similar to Hájek 's idea, Sampford (1967) selects units sequentially. The idea is to reject that sample if the same unit appears more than once and select a new sample. Assuming that V_i is the size variable associated with unit $i \in \mathcal{U}$, usually we set $\pi_i^0 = nV_i / \sum_{j=1}^N V_j$ to be the desired inclusion probability where

n represents the sample size. Using Sampford's idea, we start by selecting a unit from $\mathcal{U}$ with probability $\tilde{V}_i = V_i / \sum_{j=1}^{N} V_j$. Then the subsequent units are selected with probability proportional to

$$\frac{\tilde{V}_i}{1 - n\tilde{V}_i},$$

with replacement under the assumption that $n\tilde{V}_i < 1$ for all $i \in \mathcal{U}$. The whole sample is accepted only if it contains n distinct units. This sampling method is available in `UPsampford` function in R package `sampling` (Tillé and Matei 2016).

In this dissertation, all sampling procedures are always PPSWOR unless otherwise specified. When doing PPSWOR, Poisson sampling is used in simulations investigating bias in population totals; the Sampford method is used in simulations investigating variances; and systematic sampling is used in ACS-data simulations which will be explained clearly later in Chapter 6.

1.6 Outline of the dissertation

The rest of the dissertation is organized in the following way: in Chapter 2 we discuss weight misspecification, propose the idea of biasing factor and two classes of probabilistic model that the biasing factor may follow. At the end of Chapter 2, we discuss the condition under which that HT using misspecified weights is still consistent. In Chapter 3, we investigate the bias in GREG and in Chapter 4,

we examine the bias in ZL. In Chapter 5, the anticipated variance of considered estimators is considered under misspecified weights. A real data example based on American Community Survey (ACS) data is given in Chapter 6, showing a data analysis assessing one of the biasing factor models proposed in Chapter 2.

Chapter 2. Probabilistic Models for Weight Misspecification

2.1 Overview of weighting

Sampling weights play important roles in producing population estimates. In a probability sample survey when estimating a population total, the simple total $\tilde{t}_Y = \sum_{i \in \mathcal{S}} Y_i$, ignoring the sampling scheme would lead to severe levels of bias. Instead the HT estimator with the form $\hat{t}_Y^{HT,\,d} = \sum_{i \in \mathcal{S}} d_i Y_i$ is a design-unbiased estimator of population total, where d_i is the corresponding analysis weight. The general goal in weighting, or weight modification procedures, is to find a set of weights, w_i, that can be used in all analyses (including those with different attributes Y_i) to produce estimates for the target population under study. The HT estimate is one example; regression model analyses could also use the same set of weights if the same set of predictor variables remains suitable. If properly constructed, a set of weights can provide approximately unbiased and consistent estimates of many different population parameters of interest. As a result, one set of weights can serve many purposes, which is a major practical advantage.

Starting with the base or design weights, the common procedures of weight modifications include 1) adjustment for unknown eligibility, for example, distribut-

ing the total sample weights of the sampled units found not to be eligible among those determined to be eligible; 2) adjustment for nonresponse; 3) use of auxiliary data, for example calibration, to reduce variances and correct for frame deficiencies; 4) other changes including weight trimming and collapsing cells (Lohr 2009; Pfeffermann and Rao 2009; Valliant et al. 2013). From now on let d_i be the *design weight* or *base weight* for unit $i \in \mathcal{U}$, which is the weight before all modifications. Then $\pi_i^0 = P(i \in \mathcal{S}) = 1/d_i$ is the inclusion probability. Let w_i's be the *final weights* or *modified weights*, which are available in the released final data sets to the public and the microdata users. Let $\pi_i^F = 1/w_i$ denote the reciprocal of the modified weight for unit i. In general,

$$d_i \neq w_i,\ i \in \mathcal{S}\,. \tag{2.1}$$

2.2 Motivation for misspecified weights

A frequently used nonresponse adjustment method to handle nonresponse in sample surveys is propensity weighting, extending methods first introduced in Rosenbaum and Rubin (1983). Define ϕ_i as function of X_i to be the propensity score associated with unit $i \in \mathcal{U}$ which implies the probability of responding to the survey if the unit is sample. $\hat{\phi}_i$'s could be given from the predicted probability of response versus nonresponse through a logistic or probit model, which would be called the response-propensity model. Then the adjusted weight for unit i could be given by $d_i/\hat{\phi}_i$ if we go with propensity weighting or $d_i\left(\frac{1}{|c|}\sum_{j\in c}\hat{\phi}_j\right)^{-1}$ if we group

$\hat{\phi}_i$'s into C classes and $i \in$ class c, an idea of long standing in surveys, discussed for example by by Little (1986). Särndal and Lundström (2005) pointed out that nonresponse bias can be reduced without increasing variance if covariates that are highly associated with the response indicator and the survey outcome variable are used. However, such covariates often are difficult to find (Kreuter et al. 2010). If an incorrect or misspecified response-propensity model is used, then w_i and π_i^0 would generally not cancel out when taking the design expectation of the HT estimator given $\mathcal{F}$.

The calibration approach, as defined in Deville and Särndal (1992), Deville et al. (1993) and Särndal (2007), using population auxiliary information or auxiliary information from a larger sample, for example t_X in GREG, constructs estimators which may have good efficiency if linear combinations of the covariates X_i with known population totals are highly correlated with Y_i. A model misspecification issue may arise when we do not have a reliable source for the population totals of X_i. Let u_i be a variable with a known total $\sum_{\mathcal{U}} u_i$. When $\sum_{\mathcal{U}} u_i^2$ is unavailable, we have to exclude u_i^2 from X_i which may hurt the efficiency if Y_i truly depends on u_i^2. Another issue is that sometimes the auxiliary information may not be available and therefore needs to be imported from an outside source considered accurate enough. When the imported information deviates from the truth, the calibration may reduce the precision of survey estimates. Repeated weight trimmings may bring problems too. In the presence of extreme values, weight trimmings are desirable to reduce variance. If there is no additional raking or calibration performed after weight trimming, matched cell counts or population totals of covariates may be affected.

To summarize, misspecification happens when we calibrate on wrong totals and/or missing important totals, when w_i's are based on an incorrect response-propensity model in terms of covariates, or when the weights have been moved too much. Inappropriate modification steps may have been done in multiple steps performed in different possible orders. When such misspecifications happen, no theory guarantees that the estimator based on w_i's is still design-consistent. This problem is different from that of Ybarra and Lohr (2008) who discussed measurement error in auxiliary information in small area estimation models.

2.3 Probabilistic models of weight misspecification

We first, introduce a random variable $\eta > 0$ accounting for the random modification processes, and assume (w_i, d_i, η_i)'s satisfying

$$w_i = d_i \eta_i. \tag{2.2}$$

Let us denote $\mathcal{F} = \{(Y_i, X_i, d_i, \eta_i, w_i)\}_{i \in \mathcal{U}}$. Equation (2.2) says that the hypothetical random factor η_i's completely explain the changes of design weights due to modification procedures. Each modification step except for weight trimming more or less has covariates involved in it. Considering X_i to contain all auxiliary information that has been used in modification procedures, it is reasonable to assume that $(Y_i, X_i, d_i, \eta_i, w_i)$'s are *iid*. In practice, the user would like to maintain the average relative difference between the modified and design weights to be close to zero, which means $E(w_i/d_i) = 1$; and within each modifying stage, the survey practition-

ers would try to avoid extreme changes and extreme values of modified w_i. Therefore it is reasonable to further assume that η_i's are *iid* samples of some general distribution F_η which would give us reasonable values of η_i and w_i. Such *iid* assumption may not hold if weights have been adjusted due to the presence of stratum jumpers (Beaumont and Rivest 2007). That means, some units are wrongly classified into strata that they do not belong to and weight adjustments must be done to correct this. In this thesis, we exclude this possibility and still assume *iid* samples. These two assumptions, and an additional assumption are summarized as following,

A.1 $E(\eta) = 1$;

A.2 $\eta_i \overset{iid}{\sim} F_\eta$;

A.3 $E(\eta Y) = E(Y)$.

Next, Proposition 1 is given discussing when HT under w would still be unbiased.

Proposition 1. *Assume (Y_i, d_i, w_i, η_i)'s are superpopulation iid samples. Under unequal probability sampling, assume that unit $i \in \mathcal{U}$ is selected with probability $\pi_i^0 = 1/d_i$. Only the final weight $w_i = d_i\eta_i$ enters into analysis stage. Then HT is unbiased if and only if **A.3** holds.*

Proof of Proposition 1. HT estimator under w has expectation of

$$\begin{aligned} E\left(\hat{t}_Y^{HT,w}\right) &= E\left(\sum_{i=1}^{N} I_{[i\in\mathcal{S}]} d_i \eta_i Y_i\right) \\ &= E\left(\sum_{i=1}^{N} \eta_i Y_i\right). \end{aligned}$$

Therefore $\hat{t}_Y^{HT,w}$ is unbiased if and if only

$$E\left(\sum_{i=1}^{N}\eta_i Y_i\right) = E\left(\sum_{i=1}^{N} Y_i\right),$$

indicating that $E(\eta Y) = E(Y)$ is the necessary and sufficient condition for unbias of HT estimator under w. □

In the following chapters of this dissertation, we assume **A.1** through **A.3**, but will continue our discussion under the general assumption that HT remains unbiased.

Next, two general models of η_i will be introduced in Section 2.3.1 to 2.3.2, following the assumptions **A.1** to **A.3**.

2.3.1 Binary probabilistic model of weight misspecification

The first model considers an extreme case where $\eta = w/d$ only has two possible values, whose probability distribution depends on X. The biasing factor η and outcome variable Y are independent given the covariate X.

$$\eta = 1 - \zeta + 2\zeta I(X) \tag{2.3}$$

where $I(X)$ is an indicator function depending on X with expectation $1/2$, where ζ is a constant in $(0, 1)$ determining the level of misspecification.

2.3.2 Continuous probabilistic model of weight misspecification

A probabilistic model of weight misspecification is introduced as the following

$$\begin{cases} \eta_i = c(X_i) \cdot \left(m_\zeta(a(X_i))\right)^{-1} \cdot \exp\left(a(X_i)\zeta_i\right) \\ \zeta_i \overset{iid}{\sim} F_\zeta \\ \mathrm{E}[c(X_i)] = 1 \end{cases} \tag{2.4}$$

where $m_\zeta(\cdot)$ is the moment generating function of ζ, and ζ_i is independent of (Y_i, X_i) and of other random variables used in modeling. The probabilistic model (2.4) suggests that the modification procedures depend on the covariates X_i's and some other purely random factors which are accounted for by ζ_i. It is easy to see that (2.4) satisfy **A.1** and **A.3**. Here $a(\cdot)$ and $c(\cdot)$ are real-valued functions. Notice that directly from (2.4) we have

$$E(\eta_i|X_i) = c(X_i).$$

If $c(X_i) \equiv 1$, then we still have unbiased HT estimator by

$$E\left(\hat{t}_Y^{HT}\right) = E_p\left(E_d\left(\sum_{i=1}^{N} I_{[i\in\mathcal{S}]} w_i Y_i \,\middle|\, \mathcal{F}\right)\right) = \sum_{\mathcal{U}} E_p(Y_i). \tag{2.5}$$

The $E_p(\cdot)$ denotes the expectation with respect to the superpopulation model, following Isaki and Fuller (1982). If, however, $c(X_i) \neq 1$ for any X_i and $E\Big(Y_i c(X_i)\Big) \neq E_p(Y_i)$, then we lose unbiasedness of HT and the relative bias (RB) of HT estimator

is given by

$$RB(HT) = \frac{E\Big(\sum_{\mathcal{U}} I_{[i\in\mathcal{S}]} w_i Y_i\Big) - E\Big(\sum_{\mathcal{U}} Y_i\Big)}{E\Big(\sum_{\mathcal{U}} Y_i\Big)} = E\Big(Y_i(\eta_i - 1)\Big)\Big/E(Y_i).$$

2.4 Discussion

In this chapter, we introduced the idea of weight misspecification and proposed two classes of probabilistic model that the biasing factor may follow. Then we discussed the condition under which HT under misspecified weights is still consistent. In next three chapters, we will discuss how the biasing factor will affect our estimates of the population total. Specifically, Chapter 3 focuses on bias in GREG, Chapter 4 focuses on bias in ZL and Chapter 5 focuses on anticipated variances.

Chapter 3. Bias in Generalized Regression Estimator

As discussed in the previous chapter, HT under modified weights, $\hat{t}_Y^{HT,w}$, is still consistent as long as

$$E(\eta Y) = E(Y). \tag{3.1}$$

Similarly, the HT estimator of the population total of X under misspecified weights w_i, denoted by $\hat{t}_X^{HT,w}$, is consistent if we have the following

$$E(\eta X) = E(X). \tag{3.2}$$

As mentioned in Chapter 1, the generalized regression estimator (GREG) under modified weights, $\hat{t}_Y^{GREG,w}$, has the following form

$$\hat{t}_Y^{GREG,w} = \hat{t}_Y^{HT,w} + \left(\begin{pmatrix} N \\ t_X \end{pmatrix} - \begin{pmatrix} \hat{N}^{HT,w} \\ \hat{t}_X^{HT,w} \end{pmatrix} \right)^{tr} \hat{\beta}^{GREG,w}, \tag{3.3}$$

where $\hat{N}^{HT,w} = \sum_{i \in \mathcal{S}} w_i$ and X_i is not linearly degenerate in the sense that for fixed $\mathcal{F}$ there exists a constant vector c of the same dimension as X_i such that

$$\sum_{i \in \mathcal{U}} \left(X_i^{tr} c - 1 \right)^2 = 0.$$

If we have equations (3.1) and (3.2) hold, then the GREG estimator under w, $\hat{t}_Y^{GREG,w}$, is still consistent. In reality, it is possible that (3.2) does not hold but (3.1) holds, in which case the consistency of HT under w is guaranteed. The interpretation behind this situation is that, some covariates are not appropriately calibrated, as indicated by

$$E(\eta X) \neq E(X).$$

If this is true, then the second term in (3.3) might not have mean zero even with large sample size, implying that GREG under modified weights would not necessarily be consistent when $\hat{t}_Y^{HT,w}$ is consistent. Later in this chapter, Proposition 3 will be given explaining this in detail.

In this chapter, we discuss the potential consequences of using w in GREG when $\hat{t}_Y^{HT,w}$ is consistent. All of the following sections including simulation studies in this chapter assume (3.1). A brief review of GREG is given followed by the discussion of $\hat{t}_Y^{GREG,w}$. A bias formula of GREG in the limiting sense is given, showing that when the outcome model is wrongly specified and the inappropriately adjusted weights are being used, GREG may have serious bias. In the simulation study, an example of a misspecified $E(Y_i|X_i)$ is presented where only the correct main-effect terms are used in GREG. Under the misspecification model and parameters chosen guaranteeing the consistency of HT using w, the simulation results show that HT is always consistent as we expect while GREG sometimes becomes inconsistent when misspecified w_i's are used, indicating that misspecifying both the outcome model and sampling weights may lead to meaningful bias.

3.1 Brief review on generalized regression estimator

GREG (Särndal et al. 1992; Fuller 2002), given by (3.4), is *design-consistent* utilizing the association between covariate X_i and outcome Y_i when the total $t_X = \sum_{\mathcal{U}} X_i$ is assumed to be known. First, we address the case where the weights are properly specified in the sense of being equal to the inverse single-inclusion probabilities.

$$\begin{aligned} \hat{t}_Y^{GREG,\,d} &= (N,\, t_X^{tr})\, \hat{\beta}^{GREG,\,d} \\ &= \hat{t}_Y^{HT,\,d} + \left(\begin{pmatrix} N \\ t_X \end{pmatrix} - \begin{pmatrix} \hat{N}^{HT,\,d} \\ \hat{t}_X^{HT,\,d} \end{pmatrix} \right)^{tr} \hat{\beta}^{GREG,\,d} \end{aligned} \tag{3.4}$$

Let $\hat{N}^{HT,\,d} = \sum_{i \in \mathcal{S}} d_i$ and $\hat{t}_X^{HT,\,d} = \sum_{i \in \mathcal{S}} d_i X_i$ in (3.4). We follow Fuller's definition (Fuller 2011) of design-consistency. That is, given the finite population sequence $\mathcal{F}_N = \{(Y_i, X_i)\}_{i=1}^N$ indexed by N, a sequence of associated sample designs with sample size n tending to ∞, $\hat{t}_Y^{GREG,\,d}$ satisfies

$$\forall \epsilon > 0,\ \lim_{N \to \infty} P_d \left\{ \left| \hat{t}_Y^{GREG,\,d} - t_Y \right| > \epsilon \,\middle|\, \mathcal{F}_N \right\} = 0,\ a.s. \tag{3.5}$$

where the notation means that we condition on the realized finite population $\mathcal{F}_N$ and the probability $P_d\{\,\cdot \,|\, \mathcal{F}_N\}$ is with respect to the sample design. In (3.5), "a.s.", short for "almost surely", means that the property holds for all sequences except for a set of measure zero.

The estimated coefficient $\hat{\beta}^{GREG,\,d}$ has the following form,

$$\hat{\beta}^{GREG,\,d} = \left(\sum_{\mathcal{S}} d_i \begin{pmatrix} 1 \\ X_i \end{pmatrix}^{\otimes 2}\right)^{-1} \left(\sum_{\mathcal{S}} d_i \begin{pmatrix} 1 \\ X_i \end{pmatrix} Y_i\right) \tag{3.6}$$

. Assume that the element $(Y_i, X_i) \in \mathcal{F}$ are independent realizations of random vector (Y, X) following an unknown distribution and define $\mu = E(X)$, Fuller (2011) showed that $\hat{\beta}^{GREG,\,d}$ has a limit in probability, as $N,\, n \to \infty$

$$\begin{aligned} \beta^{GREG,\,d} &= E\left(\begin{pmatrix} 1 \\ X \end{pmatrix}^{\otimes 2}\right)^{-1} E\left(\begin{pmatrix} 1 \\ X \end{pmatrix} Y\right) \\ &= \begin{pmatrix} 1 & \mu^{tr} \\ \mu & E(X^{\otimes 2}) \end{pmatrix}^{-1} \begin{pmatrix} E(Y) \\ E(XY) \end{pmatrix}. \end{aligned} \tag{3.7}$$

As fully discussed in the literature (Estevao and Särndal 2000; Särndal 2007), there are various benefits of constructing GREG when possible:

- Usually we assume we know the exact total of X, t_X, or a very good estimate of t_X from other sources and it is often true that

$$\sum_{i \in \mathcal{S}} d_i X_i \neq t_X.$$

The calibration procedure reproduces exactly the known total for X. That is, the calibrated procedure adjusts $\{d_i\}_{i \in \mathcal{S}}$ to $\{\tilde{d}_i\}_{i \in \mathcal{S}}$ in such a way that

$$\sum_{i \in \mathcal{S}} \tilde{d}_i X_i = t_X. \tag{3.8}$$

It is a well known propoerty of GREG that $\hat{t}_Y^{GREG,\,d} = \sum_{i\in\mathcal{S}} \tilde{d}_i Y_i$ for a set of weights $\{\tilde{d}_i\}_{i\in\mathcal{S}}$ that are called "linearly calibrated weights" (Deville and Särndal 1992; Deville et al. 1993).

- Another important reason for constructing GREG at the analysis stage is that GREG has a smaller asymptotic variance than the HT estimator $\hat{t}_Y^{HT,\,d} = \sum_{i\in\mathcal{S}} d_i Y_i$ unless the covariates are all completely uncorrelated with Y.

3.2 GREG under misspecified weights

Under the misspecified weights w_i's, the estimated regression coefficient is given by

$$\hat{\beta}^{GREG,w} = \left(\sum_{\mathcal{S}} d_i\eta_i \begin{pmatrix}1\\X_i\end{pmatrix}^{\otimes 2}\right)^{-1}\left(\sum_{\mathcal{S}} d_i\eta_i \begin{pmatrix}1\\X_i\end{pmatrix} Y_i\right). \tag{3.9}$$

By including the biasing factor η_i into $\mathcal{F}$ and treating (Y_i, X_i, d_i, η_i) as realizations of independent samples of random vector (Y, X, d, η) following an unknown distribution and defining $\mu^* = E(\eta X)$ and $A = E(\eta X^{\otimes 2})$, we know that $\hat{\beta}^{GREG,\,w}$ has a limit in probability

$$\beta^{GREG,w} = \begin{pmatrix} 1 & \mu^{*tr} \\ \mu^* & A \end{pmatrix}^{-1} \begin{pmatrix} E(Y) \\ E(\eta XY) \end{pmatrix}. \tag{3.10}$$

So using w_i the GREG estimator divided by population size has a large-sample limit

$$\mu_Y^w = (1,\, \mu^{tr})\beta^{GREG,w}.$$

If $(1, \mu^{tr})\beta^{GREG,w} = E(Y)$, then GREG would still be design-consistent, even with misspecified weights w_i's used. The following two propositions state GREG's consistency in different situations. We will assume the following,

A.4 $E(\eta) = 1$ and $E(\eta Y) = E(Y)$;

A.5 $\hat{t}_Y^{HT,d}$, $\hat{t}_X^{HT,d}$ are design-consistent for finite population characteristics t_Y, t_X.

A.6 $N^{-1}\left(\sum_{i\in S} d_i\eta_i\binom{1}{X_i}^{\otimes 2}\right)$, and $N^{-1}\left(\sum_{i\in S} d_i\eta_i\binom{1}{X_i}Y_i\right)$ are design-consistent for finite population characteristics $\sum_{i=1}^N \eta_i\binom{1}{X_i}^{\otimes 2}\Big/N$, and $\sum_{i=1}^N \eta_i\binom{1}{X_i}Y_i\Big/N$.

Assumptions **A.5** and **A.6** are very weak assumptions, corresponding to laws of large numbers for the summed quantities and would fail only when there is extraordinarily strong dependence or imbalance in magnitude among the summands.

First let us consider the GREG under w_i with conditional mean of outcome variable Y given covariate X, $E(Y|X)$ being correctly specified. If η and Y are uncorrelated given X, then the GREG under w_i is still consistent. This is summarized in the following proposition. The proof can be found at the end of this chapter in Section 3.6.

Proposition 2. *Let $\mathcal{F}_N = \{(Y_i, X_i, d_i, \eta_i)\}_{i=1}^N$ be a sequence of identically distributed independent realizations of random vector (Y, X, d, η). Let $\pi_i^0 = 1/d_i$. Assume **A.4** to **A.6** and further assume the following:*

A.7 $E(Y|X) = (1, X^{tr})\beta$;

A.8 $E(\eta Y|X) = E(\eta|X)E(Y|X)$.

Then $(N,\, t_X^{tr})\hat{\beta}^{GREG,\,w}$ *converges to* $E(Y)$ *in probability, as both* n, N *go to* ∞.

However in practice, the specified outcome model might be wrong. That is, **A.7** may not hold. When the outcome model is wrong, it would be interesting to evaluate $\mu_Y^w = (1,\, \mu)^{tr}\beta^{GREG,\,w}$ to see if constructing GREG under w_i is still a good idea.

Based on the results on limiting distribution for the regression coefficients given by Fuller (2011), the following proposition states the general formula for the large-sample bias induced by regression coefficient $\hat{\beta}^{GREG,\,w}$. The proof can be found at the end of this chapter in Section 3.6.

Proposition 3. *Let* $\mathcal{F}_N = \{(Y_i, X_i, d_i, \eta_i)\}_{i=1}^N$ *be a sequence of identically distributed independent realizations of random vector* (Y, X, d, η). *Let* $\pi_i^0 = 1/d_i$ *and* $A = E(\eta X^{\otimes 2})$. *Define* $\mu = E(X)$, $\mu^* = E(\eta X)$ *and* $\Delta = 1 - \mu^{*\,tr} A^{-1}\mu^*$. *Let us assume* ***A.4*** *to* ***A.6*** *and the following:*

A.9 $E(X^{\otimes 2})$ *and* A *are invertible.*

Then $(N,\, t_X^{tr})\hat{\beta}^{GREG,\,w}/N$ *has a limit in probability as both* n, N *go to* ∞, *equal to*

$$\begin{aligned} &(1,\, \mu^{tr})\,\beta^{GREG,w} \\ &= E(Y) \;+\; (\mu - \mu^*)^{tr}\left\{\mathbf{I} + \frac{A^{-1}\mu^{*\otimes 2}}{\Delta}\right\} A^{-1}\, E\Big(\eta Y(X - \mu^*)\Big). \end{aligned} \tag{3.11}$$

Equation (3.11) gives the bias formula for GREG under misspecified weights. From (3.11), we see that $E(X) \neq E(\eta X)$ or $\mu \neq \mu^*$ when $\hat{t}_Y^{GREG,\,w}$ is inconsistent. With that being said, $E(X) \neq E(\eta X)$ is a necessary but not sufficient condition for

bias. From now on let us denote the bias term for GREG under w_i, equal to the right-hand side of (3.11) minus $E(Y)$, by $Bias(GREG, w)$.

3.3 Simulation models

This section presents the simulation models that will be used in this and later chapters. To be specific, we present the outcome models that describe the Y on X relationship and the propensity models that view inclusion probabilities as a function of covariate X. Chapters 3 and 4 discuss the bias in GREG and ZL, separately, and repeatedly use different versions of the outcome and propensity models that are described in this chapter. Again, let us assume outcome variable Y is a scalar and covariate X is p-dimensional. When discussing the dependence between outcome variables Y_i, covariate column vectors X_i and inclusion probabilities π_i^0, we treat the vectors (X_i, Y_i, π_i^0) as superpopulation *independent and identically distributed* (*iid*) samples. In simulation sections of later chapters, some variables for example X, may follow different distributions depending on the purpose of the simulation. But here, we focus on the relationship between the variables.

3.3.1 Outcome model

We consider a class of linear models as outcome models of the form

$$Y_i = \phi_0 + X_i^{tr}\phi_1 + W_i^{tr}\phi_2 + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2), \tag{3.12}$$

where W_i's include squares of and cross terms between components of X_i. When constructing GREG, it is interesting to evaluate GREG in the presence of both misspecified weights and misspecified outcome model. Given the form of (3.12), one easy form of outcome model misspecification is ignoring W_i, i.e. using

$$Y_i \;=\; (1, X_i^{tr})\beta \;+\; \varepsilon_i$$

as working model.

3.3.2 Propensity model

Assuming probability-proportional-to-size (PPS) sampling, let V_i represent the size associated with unit $i \in \mathcal{U}$, defined as

$$V_i = \frac{c_1}{c_2 + c_3 \left(E(Y_i|X_i)\right)^{\nu} + \delta_i}, \tag{3.13}$$

where δ_i's are iid $\mathcal{N}(0, \sigma_\delta^2)$, independent of (X_i, ε_i). In (3.13), c_1, c_2 and c_3 are constants, and ν could be $1/2$, 1 or 2. Parameters are chosen so that $V_i > 0$ for all $i \in \mathcal{U}$. The inclusion probability π_i^0 is taken proportional to size V_i, that is

$$\pi_i^0 \;=\; \frac{nV_i}{\sum_{j=1}^{N} V_j}$$

where N is the finite population size and n is the sample size. Such a size variable is roughly associated with inverse $E(Y_i|X_i)$ and hence depends on X_i through $E(Y_i|X_i)$.

So inclusion probabilities could be viewed approximately as a function of X_i.

3.4 Simulation studies

Simulation studies are conducted to study and illustrate how GREG estimates may be biased when both the wrong outcome model and misspecified adjusted weights are used, when HT is unbiased. First we simulate finite populations with different numbers of covariates, following the outcome model described in (3.12). Random probability-proportional-to-size samples are then drawn from the finite populations with inclusion probabilities $\pi_i^0 \propto V_i$ where V_i is the measure of size associated with unit $i \in \mathcal{U}$. Specifically, π_i^0 and V_i follow the propensity model (3.13). The binary biasing factor η_i, following (2.3), is taken to be

$$\eta_i = 1 - \zeta + 2\zeta I_{\{\gamma^{tr} X_i > \gamma^{tr} E(X_i)\}}, \tag{3.14}$$

where γ is a p-dimensional column vector and ζ is a chosen constant controlling the level of misspecification, with a larger value indicating a worse case of weight misspecification.

We simulate different population sizes, $10,000$ and $50,000$ with sample sizes 100 and 500 respectively. For each outcome model, propensity model and sample size, 50 realizations of finite frame population data are generated. For each randomly generated frame population, $1,000$ random samples are drawn with inclusion probabilities π_i^0 using Poisson sampling. So the inclusion indicator $I_{[i \in \mathcal{S}]}$'s independently follow Bernoulli-trial distribution with parameters π_i^0. For each sample drawn, HT

and GREG estimators are calculated using d_i and w_i, respectively.

While searching for examples showing that GREG estimates may be biased when both the outcome model and weights are misspecified, we search for parameters in outcome model (3.12) and biasing factor model (2.3) that maximize the relative bias of GREG under w_i,

$$\frac{\Big|Bias(GREG,\, w)\Big|}{E(Y)},$$

where $Bias(GREG,\, w)$ is given in Proposition 3, under some constraints. When maximizing the relative bias, one is able to see how bad the bias could be under w_i. Usually, a relative bias exceeding 5% should be considered as a warning sign. We have to point out that the current choices of parameters given in this section do not represent the worst relative biases since only local maximizers are searched. Our purpose is to illustrate that, within the class of outcome models (3.12), the relative bias of GREG might be as bad as 5% or worse with some choices of parameters, while HT is still unbiased under w_i. Therefore the first constraint we should impose is

$$E(\eta Y) \;=\; E(Y),$$

which guarantees that HT is unbiased under w_i. The second constraint is

$$\Big|\mu^{(j)} \;-\; \mu^{*\,(j)}\Big| \leq K\frac{\sigma_j}{\sqrt{n}}, \quad j = 1, \ldots, p, \tag{3.15}$$

where $x^{(j)}$ represents the j^{th} entry of vector x, $\mu^* = E(\eta X)$ and $\sigma_j^2 = Var\left(X_i^{(j)}\right)$ and K is constant to be chosen below. This constraint implies that the weights are

misspecified in a way that X might be miscalibrated to a moderate extent. From Proposition 3, we know that when bias in GREG under w_i is present, it must be true that

$$\mu \neq \mu^*. \tag{3.16}$$

It is not surprising that large errors would introduce bias. It would be more interesting to see if bias also exists when the miscalibration on X component totals is mild and is less likely to be identified by investigators or data users through preliminary statistical tests. Therefore, we put an upper bound of $K \times \sigma_j/\sqrt{n}$ on absolute bias in each X-component total. When $K = 1.96$, this upper bound is equivalent to a multiple p-fold Z-test at .05 significance level.

As mentioned, we searched parameters through optimization in such a way that GREG is biased under w_i. We should keep in mind that when d_i's are used, GREG maintains consistency even if the working model ignores important nonlinear and interaction terms in the present section, with moderate to large sample size n. When n is relatively small, we may still observe some bias in GREG.

When examining the simulation results, two things may be interesting besides checking relative biases:

1. Based on sampling theory, the investigator could construct a confidence interval (CI) for t_Y using $\hat{t}_Y^{GREG,w}$ ignoring the misspecified weights under the

Poisson sampling design, i.e.

$$\left(\hat{t}_Y^{GREG,\,w} \pm z_{1-\alpha/2} \sqrt{\sum_{i \in \mathcal{S}} (w_i^2 - w_i) \left(Y_i - X_i^{tr} \hat{\beta}^{GREG,\,w} \right)^2} \right), \tag{3.17}$$

where z_q represents the q^{th} quantile for standard normal distribution. So z_{1-q} satisfies $P\{Z \leq z_q\} = q$, where $0 < q < 1$ and $Z \sim \mathcal{N}(0,1)$. If the coverage probability of (3.17) is far lower than the nominal level $1-\alpha$ that the investigator would expect, then the investigator might report much higher confidence than is warranted.

2. With t_X assumed to be known, data users often would find that

$$\sum_{i \in \mathcal{S}} w_i X_i = t_X \tag{3.18}$$

do not hold for some prediction variables with known totals. But investigators may still believe that all covariates being used have been well calibrated if no significant test results indicate miscalibrations. Specifically, define $Z = \sum_{i=1}^{N} (w_i I_i - 1) X_i$ and consider the hypothesis test

$$H_0 : \text{no miscalibration, i.e. } E_d(Z) = 0$$

versus

$$H_1 : \text{there is miscalibration, i.e. } E_d(Z) \neq 0.$$

Test 1 Under H_0 in the Poisson sampling setting, the variance-covariance matrix

of Z, denoted by Σ_Z, has the form

$$\sum_{i \in \mathcal{U}} (d_i - 1) X_i^{\otimes 2},$$

which is estimated by Horvitz-Thompson style estimator

$$\widehat{\Sigma}_Z \;=\; \sum_{i \in \mathcal{S}} (d_i^2 - d_i) X_i^{\otimes 2}.$$

So the χ^2 test statistics

$$\mathcal{X}^2 \;=\; Z^{tr}\, \widehat{\Sigma}_Z^{-1}\, Z \;\sim\; \chi_p^2$$

under H_0. We reject H_0 if $p - value < .05$ where $p - value = 1 - F_{\chi_p^2}(\mathcal{X}^2)$, with $F_{\chi_p^2}$ representing the cumulative distribution function of χ_p^2. This test works well if the investigator is testing against many subtle miscalibrations, for example against alternatives with

$$|E(Z)| \;\geq\; (\lambda_0, \ldots, \lambda_0)^{tr},$$

where $\lambda_0 > 0$ is small.

Test 2 Reject H_0 if

$$\max_{1 \leq j \leq p} \left| \left(\widehat{\Sigma}_Z^{-1/2} Z \right)^{(j)} \right| \;>\; C,$$

where C is determined in such a way that

$$P_{d,\,H_0}\left\{\max_{1\le j\le p}\left|\widehat{\Sigma}_Z^{-1/2}Z\right|\le C\right\}\ge 1-\alpha.$$

The notation $P_{d,\,H_0}\{\cdot\}$ indicates the probability with respect to sampling design under H_0. In the present simulation, $C=\Phi^{-1}\left(\frac{(1-\alpha)^{1/p}+1}{2}\right)$. This test works well if the investigator is testing against alternatives in which one coordinate of $E_d(Z)$ is large, i.e.

$$\max_{1\le j\le p}(E_d(Z))^{(j)}\ge\lambda_1$$

3.4.1 Single covariate case

First we consider an outcome model with a single covariate, i.e. X_i's are real numbers. In (3.12), let $\phi_1=0$, $W_i=X_i^2$, $\sigma_\varepsilon^2=1$, the scalar variables $X_i\overset{iid}{\sim}\mathcal{N}(0,\sigma_x^2)$. In propensity model (3.13), $c_1=500$, $c_2=10$, $c_3=.1$, $\nu=1$, and $\sigma_\delta=1.5$. In the weight misspecification model, $I(X_i)=I_{\{X_i>0\}}$. Then HT under w_i is unbiased with

$$\begin{aligned}E(\eta Y_i)&=\phi_0+\phi_2E(\eta_iX_i^2)\\&=\phi_0+\phi_2\left((1-\zeta)\sigma_x^2+2\zeta\sigma_x^2/2\right)\\&=\phi_0+\phi_2\sigma_x^2=E(Y_i),\end{aligned}\tag{3.19}$$

and the bias for GREG under w is a function of $(\zeta, \sigma_x, \phi_0, \phi_2)$. Two sets of $(\sigma_x, \phi_0, \phi_2)$ given $\zeta = .3, .4$ are searched by optimizing the relative bias of GREG under w, as described in (3.20).

$$\max_{\sigma_x, \phi_0, \phi_2} \frac{\left| Bias(GREG, w) \right|}{E(Y)}. \tag{3.20}$$

Table 3.1 summarizes two choices of ζ, corresponding values of $(\sigma_x, \phi_0, \phi_2)$ and relative bias calculated from (3.11). Under both sets of parameters, the resulting relative bias values are greater than 5%. The last column of Table 3.1 shows $E(\eta X)$. It is worth pointing out that in the current single covariate case, (3.16) reduces to

$$2\zeta \frac{\sigma_x}{\sqrt{2\pi}} < K \frac{\sigma_x}{\sqrt{n}}$$

which leads to

$$n < \frac{2\pi}{\zeta^2} \tag{3.21}$$

if $K = 2$. This means that $n \leq 69$ if $\zeta = .3$, so that this kind of example could arise only in very small sample surveys. The highlight of Table 3.1 should be column 5, which refers to the theoretical relative bias in percentage in the limiting sense when both n and N go to infinity. We expect to see a good approximation of relative bias by the empirical average of relative bias when n is relatively large.

Table 3.1: Selected simulation parameters in single covariate case. Parameter searching was done under two different choices of ζ. Here $E(X_1) = 0$. Column 5 is the relative bias in the GREG estimator of t_Y under w_i in percent.

ζ	σ_x	ϕ_0	ϕ_2	Relative Bias (%)	$E(\eta X_1)$
0.3	2.45	0	1.82	-6.08	0.59
0.4	2.39	0	1.79	-11.34	0.76

Table 3.2: Simulation results for single covariate case.

			HT	GREG				
ζ	(n, N)	Weights	RB(%)	RB(%)	$Bias(\hat{\beta}_0)$	$Bias(\hat{\beta}_1)$	$CR_Y(\%)$	$RR_X(\%)$
0.3	(100,10000)	d	0.09	-3.15	-0.36	-0.03	87.56	4.98
		w	0.02	-8.72	-0.96	1.03	81.12	55.93
	(500,50000)	d	0.08	-0.60	-0.07	0.01	93.14	5.00
		w	0.11	-6.62	-0.73	1.12	79.88	99.82
0.4	(100,10000)	d	-0.11	-3.22	-0.29	0.03	87.10	4.91
		w	0.01	-13.82	-1.38	1.45	73.83	82.07
	(500,50000)	d	0.02	-0.63	-0.05	-0.00	92.88	4.75
		w	0.01	-11.80	-1.20	1.50	59.37	100.00

Table 3.2 summarizes the simulation results in the single covariate case. The first column represents the value of ζ from 3.14 defining η_i, controlling the level of misspecification; the second column shows the sample size (n) and frame population size (N); column 3 indicates which set of weights have been used in the analysis where d indicates that design weights are used and w means modified weights are used; columns 4-5 show the empirical average of percent relative bias in t_Y for HT and GREG where "RB" represents relative bias; and columns 6-7 show the empirical average of bias in the regression coefficients estimated from the GREG

method; column 8 is the empirical coverage rate for 95% confidence interval (3.17), where "CR" stands for "coverage rate"; column 9 is the empirical rejection rate of **Test 1**, where "RR" means "rejection rate". In the single covariate case with $p = 1$, **Test 1** and **Test 2** are equivalent.

As implied by (3.19), HT should be unbiased even when misspecified weights are used. This is also validated in Table 3.2 since HT has nearly zero relative bias across different choices of sample size. As for GREG, the nice properties of GREG guarantee that it is still consistent under d when important variables are dropped. From columns of β_0 and β_1 we can see that when the sample size is relatively small ($n = 100$), there is still some bias in regression coefficients under d, which leads to some degree of bias in GREG under d_i. But this bias is due to relatively small sample size, and could be easily corrected by large sample size. When n increases to 500, the biases in coefficients almost shrink to zero and the relative bias in GREG also disappears with larger n. But the bias we see in GREG under w_i is not corrected by large sample size. When misspecified weights w_i's are used, the estimated regression coefficients stay biased even when sample size is increasing. When n gets large, the empirical average of relative bias in GREG stays around 6.61% when $\zeta = .3$ and 11.79% when $\zeta = .4$, both of which match the relative biases shown in Table 3.1.

Columns 8-9 of Table 3.2 reflect what data users would experience when estimating Y-total using GREG. Ideally, the coverage rate in column 8 should be close to 95%, the nominal coverage of the CI for GREG. The observed coverage rate under d_i is lower than 95% when n is small but increases to around 93.2% when n is 500, implying that statistical inference based on d_i would give us results as expected

from theory. But things change when we use w_i to construct CI. When using w_i, the coverage rate is systematically lower than 95% and this empirical average drops further as the weight misspecification becomes worse. When $\zeta = .4$, the observed coverage rate drops as n grows. Column 9 shows the proportion of rejections in **Test 1** with nominal type I error being 5%. When using d_i and therefore there are no miscalibrations on X-totals, the rejection rates are all around 5%. When using w_i meaning there is some miscalibration of X-totals, the rejection rates are very high, even close to 100% when sample size is large. The high rejection rates reflect that the data users may be suspicious about the "bad" weights in practice, which is a good thing. This coincides with discussion in (3.21). When $n > 2\pi/\zeta^2$, we always have

$$\left|\mu - \mu^*\right| > \frac{2\sigma_x}{\sqrt{n}},$$

so that it is not surprising that miscalibrations on X-totals could be identified correctly under w_i when the number of covariates is small.

3.4.2 Multi-covariate case

Next we simulate a different model, now for $p = 10$. For the outcome model, let $\phi_0 = 1$, $\phi_1 = 3 \cdot \mathbf{1_{10}}$, $W_i = (X_i^{(1)\,2}, X_i^{(2)\,2}, X_i^{(3)\,2}, X_i^{(1)}X_i^{(2)}, X_i^{(1)}X_i^{(3)})^{tr}$, $\phi_2 = 2 \cdot \mathbf{1_5}$, where $\mathbf{1_q}$ is a $q-$dimension column vector with all entries being 1. All the covariates are independently normally distributed with mean zero. Again in propensity model (3.13), $c_1 = 500$, $c_2 = 10$, $c_3 = .1$, $\nu = 1$ and $\sigma_\delta = 1.5$. In the biasing factor, $I(X) = I_{\{\gamma^{tr}X>0\}}$. We take $\zeta = .4, .6$. When $\zeta = .4$, $Var(X_i^{(j)}) = 3$ for $j = 1, 2, 3$;

when $\zeta = .6$, $Var(X_i^{(j)}) = 5$ for $j = 1, 2, 3$. For the rest j, $Var(X_i^{(j)}) = 10$. When $n = 100$, $K = 2$ in (3.15) and $k = 5$ if $n = 500$. Given the above choices of parameters and ζ, the relative bias of GREG under w_i is a function of γ only. Four sets of γ are found by solving (3.22).

$$\begin{aligned} &\max_{\gamma} \frac{Bias(GREG,\, w)}{E(Y)} \\ &\text{where } E(\eta\, Y) \;=\; E(Y), \\ &\qquad \left|\mu^{(j)} \;-\; \mu^{*\,(j)}\right| \;\leq\; K\sqrt{\frac{Var(X^{(j)})}{n}}\,, \\ &\qquad j = 1,\, \ldots,\, 10. \end{aligned} \tag{3.22}$$

The parameter search results are summarized in Table 3.3. Again, the constraint $E(\eta\, Y) \;=\; E(Y)$ forces HT under w_i to be unbiased and the second constraint restricts the miscalibrations on X-total within a mild range. The resulting parameter γ_{opt} is found according to (3.22) based on choices of ζ, N, n and K. We chose $K = 5$ for larger sample size n because of the difficulty of optimization, which implies that some coordinate may have $\mu^{(j)} - \mu^{*\,(j)} > 1.96\sqrt{Var(X^{(j)})/n}$. We should keep in mind that, in the simulation results, it is very likely that we observe large rejection rates of both tests **Test 1** and **Test 2** when sample size is large. Table 3.4 records the resulting relative biases of GREG under w. We find that under our current examples of models, sample sizes and level of weight misspecifications, the bias of GREG under w_i could be worse than 5%.

Table 3.3: Selected simulation parameters in multi-covariate case. Parameter searching was done under four different choices of (ζ, K) and γ_{opt} was maximizer of (3.22). K in column 3 is as defined in (3.15).

ζ	(n, N)	K	γ_{opt}
0.4	(100, 10000)	2	123.93,98.35,98.35,-13.74,-13.74,-13.74,-13.74,-13.74,-13.74,-13.74
	(500, 50000)	5	-137.5,-93.31,-93.31,13.89,13.89,13.89,13.89,13.89,13.89,13.89
0.6	(100, 10000)	2	10.44,10.44,10.44,-4.78,-4.78,-4.78,-4.78,-4.78,4.15,4.15
	(500, 50000)	5	68.78,68.78,68.78,-29.57,-3.08,-29.7,-35.53,19.96,-12.74,-12.5

Table 3.4: The theoretical relative bias calculated from selected parameters in multi-covariate case. Parameter searching was done under four different choices of (ζ, K). K in column 3 is as defined in (3.15).

ζ	(n, N)	K	Relative Bias (%)
0.4	(100, 10000)	2	-5.41
	(500, 50000)	5	-5.44
0.6	(100, 10000)	2	-8.40
	(500, 50000)	5	-10.46

Table 3.5: Simulation results for multi-covariate case.

			HT	GREG				
ζ	(n, N)	Weights	RB(%)	RB(%)	$\max\lvert Bias(\hat{\beta}_j)\rvert$	$CR_Y(\%)$	$RR_X^{(1)}(\%)$	$RR_X^{(2)}(\%)$
0.4	(100, 10000)	d	-0.21	-3.99	0.73	82.19	5.50	5.52
		w	-0.03	-7.97	1.49	74.17	21.31	14.91
	(500, 50000)	d	-0.03	-0.85	0.17	92.37	5.29	5.30
		w	-0.06	-6.00	1.47	68.86	99.74	97.98
0.6	(100, 10000)	d	0.09	-5.07	1.55	79.75	4.93	5.33
		w	0.17	-11.08	3.41	65.16	61.95	21.68
	(500, 50000)	d	0.04	-1.08	0.32	91.40	5.07	4.91
		w	0.07	-10.85	3.34	41.00	100.00	99.99

Table 3.5 summarizes the simulation results in the multi-covariate case. Again,

"RB" stands for relative bias, "CR" means "coverage rate" and "RR" means "rejection rate". All relative biases mentioned in this table are empirical averages of percent relative bias in t_Y. Column 5 is the maximum of empirical average of biases in regression coefficients. Columns 8-9 are rejection rates of **Test 1** and **Test 2**, respectively. As expected, HT under w_i is unbiased across different choices of ζ and sample sizes since we designed the examples in this way. Similar to the results of the single covariate case in Table 3.2, GREG under d_i shows some biases when sample size is small but this bias diminishes with increasing n. As expected, GREG under w_i is biased with relative bias greater than 5%. The coverage rates of Y-total's CI under w_i are all far below 95%, indicating that it is very hard for data users to make a good statistical inference on Y-total using GREG under w_i. The rejection rates under d_i of two tests on miscalibration of X-total are about 5%, as expected. The same quantities under w_i tell different stories depending on sample size. When sample size is small, there is some chance that data users would not be able to tell if the miscalibration on X-totals exists when it does. When the sample size is relatively large, both rejection rates increase close to 1 which is a good sign, indicating that the data users may be able to tell something is wrong with the estimated X-total when there truly are some miscalibrations. Rejection rates together with the relative bias and coverage rate tell us that when sample size is small, it may be dangerous for data users estimating Y-total using GREG under w_i and any possible error originated from misspecified weights are hard to detect.

3.5 Discussion

Survey samplers often prefer GREG since it is more efficient than HT unless all the covariates used in GREG are irrelevant to the outcome variable. However sometimes we have to compromise with a working model ignoring some important terms if $\sum_{i\in\mathcal{U}} u_i^2$ or $\sum_{i\in\mathcal{U}} u_i v_i$ are not available. Our simulation shows that such ignorance may lead to serious bias in Y-total estimates when misspecified weights are used. When weight misspecification exists, we may have miscalibrations on X-total. When the number of covariates is small, it may be easy for data users to detect such a condition. When the number of covariates increases, i.e., the working model is more complex, data users might not be able to detect such errors with a relatively small sample size.

3.6 Proofs

3.6.1 Proof of Proposition 2

Proof: The estimated regression coefficient under w_i has the form

$$\hat{\beta}^{GREG,\,w} = \left(\frac{1}{N}\sum_{\mathcal{S}} d_i\eta_i \begin{pmatrix}1\\ X_i\end{pmatrix}^{\otimes 2}\right)^{-1}\left(\frac{1}{N}\sum_{\mathcal{S}} d_i\eta_i \begin{pmatrix}1\\ X_i\end{pmatrix}Y_i\right)$$

which is design consistent for

$$\beta_N^{GREG,\,w} = \left(\sum_{i=1}^{N} \eta_i \begin{pmatrix}1\\ X_i\end{pmatrix}^{\otimes 2}\right)^{-1}\left(\sum_{i=1}^{N} \eta_i \begin{pmatrix}1\\ X_i\end{pmatrix}Y_i\right).$$

by the design consistency of $\frac{1}{N}\sum_{\mathcal{S}} d_i\eta_i \begin{pmatrix}1\\ X_i\end{pmatrix}^{\otimes 2}$ and $\frac{1}{N}\sum_{\mathcal{S}} d_i\eta_i \begin{pmatrix}1\\ X_i\end{pmatrix}Y_i$ assumed in **A.5** and **A.6**. Also we have $\beta_N^{GREG,w} - \beta^{GREG,w} = O_p\left(N^{-1/2}\right)$ under superpopulation iid sample assumption, where

$$\beta^{GREG,w} = \begin{pmatrix} 1 & \mu^{*tr} \\ \mu^* & A \end{pmatrix}^{-1} E\left(\eta\begin{pmatrix}1\\ X\end{pmatrix}Y\right).$$

Then we know that

$$\hat{\beta}^{GREG,w} \to \beta^{GREG,w}, \quad n, N \to \infty,$$

indicating that

$$\hat{t}_Y^{GREG,w}/N \xrightarrow{p} (1,\, \mu^{tr})\beta^{GREG,w}. \tag{3.23}$$

Under **A.7** and **A.8**,

$$\begin{aligned}
\beta^{GREG,w} &= E\left(\eta\begin{pmatrix}1\\ X\end{pmatrix}^{\otimes 2}\right)^{-1} E\left(\eta\begin{pmatrix}1\\ X\end{pmatrix}Y\right) \\
&= E\left(\eta\begin{pmatrix}1\\ X\end{pmatrix}^{\otimes 2}\right)^{-1} E\left(E(\eta|X)\begin{pmatrix}1\\ X\end{pmatrix}E(Y|X)\Big|X\right) \\
&= E\left(\eta\begin{pmatrix}1\\ X\end{pmatrix}^{\otimes 2}\right)^{-1} E\left(\eta\begin{pmatrix}1\\ X\end{pmatrix}^{\otimes 2}\right)\beta \\
&= \beta.
\end{aligned} \tag{3.24}$$

Equations (3.23) and (3.24) guarantee that $\hat{t}_Y^{GREG,w}/N$ is consistent for $E(Y)$.

□

3.6.2 Proof of Proposition 3

Proof: By similar argument, we have

$$\hat{\beta}^{GREG,\,w} \to \beta^{GREG,\,w}, \quad n, N \to \infty,$$

indicating that $\hat{t}_Y^{GREG,\,w}/N$ has a limit $(1,\, \mu^{tr})\beta^{GREG,\,w}$.

By the block matrix inverse (Bernstein 2005),

$$\begin{pmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{pmatrix}^{-1}$$

$$= \begin{pmatrix} (\mathbf{B}_{11} - \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{21})^{-1} & -(\mathbf{B}_{11} - \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{21})^{-1}\mathbf{B}_{12}\mathbf{B}_{22}^{-1} \\ \mathbf{B}_{22}^{-1}\mathbf{B}_{21}(\mathbf{B}_{11} - \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{21})^{-1} & \mathbf{B}_{22}^{-1} + \mathbf{B}_{22}^{-1}\mathbf{B}_{21}(\mathbf{B}_{11} - \mathbf{B}_{12}\mathbf{B}_{22}^{-1}\mathbf{B}_{21})^{-1}\mathbf{B}_{12}\mathbf{B}_{22}^{-1} \end{pmatrix}.$$

Let $\mathbf{B}_{11} = 1$, $\mathbf{B}_{12} = \mathbf{B}_{21}^{tr} = \mu^{*\,tr}$ and $\mathbf{B}_{22} = A$, then

$$\begin{pmatrix} 1 & \mu^{*\,tr} \\ \mu^{*\,tr} & A \end{pmatrix}^{-1} = \begin{pmatrix} \Delta^{-1} & -\Delta^{-1}\mu^{*\,tr}A^{-1} \\ -A^{-1}\mu^{*}\Delta^{-1} & \left(\mathbf{I} + \frac{A^{-1}\mu^{*\otimes 2}}{\Delta}\right)A^{-1} \end{pmatrix}, \tag{3.25}$$

where the lower right block follows from

$$A^{-1} + A^{-1}\mu^{*\otimes 2}A^{-1}/\Delta = \left(\mathbf{I} + \frac{A^{-1}\mu^{*\otimes 2}}{\Delta}\right)A^{-1}.$$

Therefore **A.4** together with (3.25) imply that

$$\begin{aligned}\begin{pmatrix}1\\ \mu\end{pmatrix}^{tr}\beta^{GREG,\,w} &= (1,\mu^{tr})\begin{pmatrix}\Delta^{-1} & -\Delta^{-1}\mu^{*\,tr}A^{-1}\\ -A^{-1}\mu^{*}\Delta^{-1} & \left(\mathbf{I}+\frac{A^{-1}\mu^{*\otimes 2}}{\Delta}\right)A^{-1}\end{pmatrix}\begin{pmatrix}E(\eta Y)\\ E(\eta XY)\end{pmatrix}\\ &= \frac{EY}{\Delta}-\frac{\mu^{*\,tr}A^{-1}E(\eta XY)}{\Delta}-\frac{\mu^{tr}A^{-1}\mu^{*}EY}{\Delta}\\ &\quad+\mu^{tr}\left(\mathbf{I}+\frac{A^{-1}\mu^{*\otimes 2}}{\Delta}\right)A^{-1}E(\eta XY). \qquad (3.26)\end{aligned}$$

Repeatedly using the fact that $1=\Delta+\mu^{*\,tr}A^{-1}\mu^{*}$, we further simply (3.26) as

$$\begin{aligned}&\frac{EY}{\Delta}-\frac{\mu^{*\,tr}A^{-1}E(\eta XY)}{\Delta}-\frac{\mu^{tr}A^{-1}\mu^{*}EY}{\Delta}+\mu^{tr}\left(\mathbf{I}+\frac{A^{-1}\mu^{*\otimes 2}}{\Delta}\right)A^{-1}E(\eta XY)\\ =\;& E(Y)\frac{\Delta+\mu^{*\,tr}A^{-1}\mu^{*}}{\Delta}-E(Y)\mu^{tr}A^{-1}\mu^{*}\frac{\Delta+\mu^{*\,tr}A^{-1}\mu^{*}}{\Delta}\\ &-\mu^{*\,tr}A^{-1}E(\eta XY)\frac{\Delta+\mu^{*\,tr}A^{-1}\mu^{*}}{\Delta}+\mu^{tr}\left(\mathbf{I}+\frac{A^{-1}\mu^{*\otimes 2}}{\Delta}\right)A^{-1}E(\eta XY)\\ =\;& E(Y)+\mu^{*\,tr}A^{-1}\mu^{*}E(Y)/\Delta-\mu^{tr}A^{-1}\mu^{*}E(Y)-\frac{\mu^{tr}A^{-1}\mu^{*\otimes 2}A^{-1}}{\Delta}\mu^{*}E(Y)\\ &-\mu^{*\,tr}A^{-1}E(\eta XY)-\frac{\mu^{*\,tr}A^{-1}\mu^{*\otimes 2}A^{-1}}{\Delta}E(\eta XY)\\ &+\mu^{tr}\left(\mathbf{I}+\frac{A^{-1}\mu^{*\otimes 2}}{\Delta}\right)A^{-1}E(\eta XY)\\ =\;& E(Y)-\mu^{*\,tr}\left(\mathbf{I}+\frac{A^{-1}\mu^{*\otimes 2}}{\Delta}\right)A^{-1}E(\eta XY)+\mu^{tr}\left(\mathbf{I}+\frac{A^{-1}\mu^{*\otimes 2}}{\Delta}\right)A^{-1}E(\eta XY)\\ &+\mu^{*\,tr}A^{-1}E(\eta Y\mu^{*})\frac{\Delta+\mu^{*\,tr}A^{-1}\mu^{*}}{\Delta}-\mu^{tr}A^{-1}E(\eta Y\mu^{*})\\ &-\frac{\mu^{tr}A^{-1}\mu^{*\otimes 2}A^{-1}}{\Delta}E(\eta Y\mu^{*})\\ =\;& E(Y)+(\mu-\mu^{*})^{tr}\left(\mathbf{I}+\frac{A^{-1}\mu^{*\otimes 2}}{\Delta}\right)A^{-1}E(\eta XY)\\ &+(\mu^{*}-\mu)^{tr}A^{-1}E(\eta Y\mu^{*})+(\mu^{*}-\mu)^{tr}\frac{A^{-1}\mu^{*\otimes 2}}{\Delta}A^{-1}E(\eta Y\mu^{*})\\ =\;& E(Y)\;+\;(\mu-\mu^{*})^{tr}\left\{\mathbf{I}+\frac{A^{-1}\mu^{*\otimes 2}}{\Delta}\right\}A^{-1}\,E\Big(\eta Y(X-\mu^{*})\Big).\end{aligned}$$

which completes the proof. □

Chapter 4. Bias in Zheng and Little's Methods

4.1 Spline models based on truncated power functions

In practice, we often encounter the problem that we only have some general knowledge about a function, say $g(x)$, and do not know g fully. In this case, we want to find a nice approximation of $g(x)$, say $\tilde{g}(x)$. One way to construct $\tilde{g}$ is *spline* approximation and smoothing. Assuming we are estimating $g(\cdot)$ over the interval $[a, b]$, we may subdivide the interval as

$$a \leq \tau_1 \leq \tau_2 \leq \cdots \leq \tau_m \leq b,$$

and on each subinterval use polynomials with low degree. Often we impose some piecewise or global continuity restrictions on $\tilde{g}(\cdot)$ and its derivatives to achieve smoothness. Then at any point $x \in [a, b]$, the approximation of $g(x)$ is a sum of one or more piecewise polynomials evaluated at x. Assume that the approximation, $\tilde{g}(x)$, is formed as the linear combination of linearly independent functions $\{b_j(x)\}_{j=1}^{J}$,

$$\tilde{g}(x) = \sum_{j=1}^{J} c_j b_j(x).$$

Such $\tilde{g}(x)$ is called a "spline". Functions $b_j(x)$ are called spline basis functions.

One type of commonly used spline basis functions is called "truncated power functions" (or power functions, TPF). For m chosen knots and degree p, there are $k+p+1$ basis elements $b_j(x)$:

$$1, x, \ldots, x^p, (x-\tau_1)_+^p, \ldots, (x-\tau_m)_+^p,$$

where $x_+ = \max\{x, 0\}$ denotes the positive part of x. A convenient feature of TPF is that after p is chosen, adding or deleting knots is equivalent to adding or removing the basis function $(x-\tau_i)_+^p$ for some i's. Formal methods for choosing knots include stepwise idea (Gentle 2009) and regularization method (Ruppert 2002; Ruppert et al. 2003).

Often we estimate coefficients c_j's by least-squares with a roughness penalty. As summarized in Schoenberg (1964, 1988), the solutions of such optimization problems, within broad classes of potential solutions, are in fact splines. Such a regression spline fit is called a penalized spline, or p-spline model fit. In this chapter, the smoothing method proposed by Zheng and Little (2003) is discussed and examined.

4.2 Zheng and Little's methods

As always let π_i^0 be the actual inclusion probability for unit $i \in \mathcal{U}$ which is sometimes known to the users. Let $d_i = 1/\pi_i^0$ be the design weight for unit i. Assume w_i's are the modified or final weights after all the weight adjustment procedures. Define $\pi_i^F = 1/w_i$. Final weights w_i are always available in the final

data analysis stage. According to Zheng & Little's idea, one may replace Y_i in the HT estimator using a spline-model-based estimator of $E(Y_i|\pi_i^0)$, which is more robust to misspecification than simpler parametric models and still provides more efficient estimation of t_Y than the HT estimator. Consider the model

$$\begin{aligned} Y_i &= \gamma_0 + \sum_{j=1}^{p} \gamma_j (\pi_i^0)^j + \sum_{l=1}^{m} \gamma_{p+l} ((\pi_i^0)^j - \kappa_l)_+^p + \varepsilon_i \\ &\text{where } \varepsilon_i \overset{iid}{\sim} \mathcal{N}\left(0, (\pi_i^0)^{2k} \sigma_\varepsilon^2\right) \end{aligned} \tag{4.1}$$

Let us further denote $b_j(\pi_i^0)$'s to be the spline basis functions

$$\begin{cases} b_0(\pi_i^0) = 1, \\ b_j(\pi_i^0) = (\pi_i^0)^j,\ j = 1, \ldots, p, \\ b_{p+l}(\pi_i^0) = (\pi_i^0 - \kappa_l)_+^p,\ l = 1, \ldots, m. \end{cases} \tag{4.2}$$

and denote

$$\mathbf{b}(\pi_i^0) \ = \ (b_0(\pi_i^0), \ldots, b_{p+m}(\pi_i^0))^{tr}.$$

Then $E(Y_i|\pi_i^0)$ could be estimated by

$$\hat{Y}_i^{ZL,\,d} \ = \ \mathbf{b}^{tr}(\pi_i^0)\hat{\gamma}^{ZL,\,d}, \tag{4.3}$$

where $\hat{\gamma}^{ZL,\,d}$ is the solution to the estimating equation

$$\sum_{i \in \mathcal{S}} d_i^{2k} b_j(\pi_i^0)(Y_i - \mathbf{b}^{tr}(\pi_i^0)\gamma) \ = \ 0, \quad j = 0, 1, 2, \ldots, p+m. \tag{4.4}$$

We can see that the solution $\hat{\gamma}^{ZL,\,d}$ is the coefficient set of a spline-fit to the sample survey data as a function of inclusion probabilities $\pi_i^0 = d_i$.

If we replace d_i's and π_i^0's with w_i's and π_i^F's in (4.1) – (4.4), we get $\hat{Y}_i^{ZL,\,w}$,

$$\hat{Y}_i^{ZL,\,w} = \mathbf{b}^{tr}(\pi_i^F)\hat{\gamma}^{ZL,\,w},$$

where $\hat{\gamma}^{ZL,\,w}$ is the solution to the estimating equation

$$\sum_{i\in\mathcal{S}} w_i^{2k} b_j(\pi_i^F)(Y_i - \mathbf{b}^{tr}(\pi_i^F)\gamma) = 0, \quad j = 0, 1, 2, \ldots, p+m. \tag{4.5}$$

Zheng & Little suggested that sample quantiles of inclusion probabilities within sample could be chosen as knots κ_l's and that one could take $k = 0, 1/2$ or 1. After estimating the coefficients of the p-spline by least squares, Zheng and Little suggested two ways to construct estimates of t_Y:

(1). If inclusion probabilities are only known for the sampled units, ZL1, the first estimator of t_Y constructed by Zheng and Little, is given by

$$\begin{aligned} \hat{t}_Y^{ZL1,\,d} &= \sum_{i\in\mathcal{S}} d_i \hat{Y}_i^{ZL,\,d} \\ \hat{t}_Y^{ZL1,\,w} &= \sum_{i\in\mathcal{S}} w_i \hat{Y}_i^{ZL,\,w} \end{aligned} \tag{4.6}$$

(2). If inclusion probabilities are known for all the units in finite frame population,

ZL2, the second type of estimator of t_Y constructed by Zheng and Little is

$$\begin{aligned} \hat{t}_Y^{ZL2,\,d} &= \sum_{i\in\mathcal{S}} Y_i + \sum_{i\notin\mathcal{S}} \hat{Y}_i^{ZL,\,d} \\ \hat{t}_Y^{ZL2,\,w} &= \sum_{i\in\mathcal{S}} Y_i + \sum_{i\notin\mathcal{S}} \hat{Y}_i^{ZL,\,w} \end{aligned} \tag{4.7}$$

According to Zheng and Little (2003), k could be chosen from $0, 1/2$ and 1. In general no matter which set of weights is used, ZL1 would be either HT precisely or very close to HT under some conditions. This statement is made precise in Proposition 4. The proof can be found at the end of this chapter, Section 4.5.1.

Proposition 4. *Let $\{(Y_i, d_i, \eta_i, w_i)\}_{i=1}^N$ be a finite universe of real-valued elements, and let π_i^0 and η_i be the corresponding inclusion probability and biasing factor, respectively, associated with $i \in \mathcal{U}$. Let $\mathcal{S}$ be the set of selected indices. Define $d_i = 1/\pi_i^0$, $\pi_i^F = 1/w_i$, where d_i and w_i satisfy*

$$w_i = d_i\eta_i.$$

Let us then further assume the following:

A.10 *The degree of the spline model $p \geq 1$.*

A.11 *$(Y_i, d_i, \eta_i, w_i) \overset{iid}{\sim}$ unknown distribution, $i = 1, \ldots, N$.*

A.12 *Weights are uniformly bounded from above. That is, $\exists\, a > 0$ s.t.*

$$1 \leq d_i,\, w_i \leq 1/a, \quad \text{for all } i \in \mathcal{U}\,.$$

When constructing ZL1 following (4.6) to estimate t_Y, ZL1 is either exactly equal to HT when $k = 1/2$ or 1, or approximately equal to HT when $k = 0$ with sufficiently large p, m, under d_i or w_i.

Unlike ZL1, ZL2 is formed as the summation of Y_i's both within and outside the sample, based on extrapolation when $i \notin \mathcal{S}$. Therefore, the success of $ZL2$ heavily relies on the quality of spline fitting outside of the sample. This is summarized in Proposition 5. The proof can be found at the end of this chapter, Section 4.5.2.

Proposition 5. *Let $\{Y_i\}_{i\in\mathcal{U}}$ be a finite universe of real-valued elements, and let π_i^0 and η_i be the corresponding inclusion probability and biasing factor respectively associated with $i \in \mathcal{U}$. Let $\mathcal{S}$ be the set of sampled indices. Define $d_i = 1/\pi_i^0$, $\pi_i^F = 1/w_i$, where d_i and w_i satisfy*

$$w_i = d_i\eta_i.$$

Let $\hat{\gamma}^{ZL,d}$ and $\hat{\gamma}^{ZL,w}$ be the solution to (4.4) and (4.5), respectively. Then the following condition,

$$E\left(\sum_{i\notin\mathcal{S}} d_i^{2k} b_j(\pi_i^0)\big(Y_i - \mathbf{b}^{tr}(\pi_i^0)\hat{\gamma}^{ZL,d}\big)\right) = 0, \qquad j = 0,\ldots,p+m, \tag{4.8}$$

is sufficient for consistency of ZL2 under d_i in t_Y. Replacing d_i with w_i in (4.8), then we get a sufficient condition for the consistency of ZL2 under w_i,

$$E\left(\sum_{i\notin\mathcal{S}} w_i^{2k} b_j(\pi_i^F)\big(Y_i - \mathbf{b}^{tr}(\pi_i^F)\hat{\gamma}^{ZL,w}\big)\right) = 0, \qquad j = 0,\ldots,p+m. \tag{4.9}$$

According to Proposition 5, if $\hat{t}_Y^{ZL,w}$ is not consistent, we should observe that

$$\sum_{i \notin \mathcal{S}} w_i^{2k} b_j(\pi_i^F)\big(Y_i \; - \; \mathbf{b}^{tr}(\pi_i^F)\hat{\gamma}^{ZL,w}\big)$$

is far from zero for at least one $j = 0, \ldots, p+m$. In simulation, we should observe that the corresponding empirical average is far from zero too.

4.3 Simulation Studies

The idea of the following simulation study is to show that even if the model is a good fit within the sample, extrapolating the estimated model to the units outside of sample could be dangerous and could lead to inconsistent estimators. In other words, ZL2 under w_i must have

$$E\left(\sum_{i \notin \mathcal{S}} w_i^{2k} b_j(\pi_i^F)(Y_i - \hat{Y}_i^{ZL,w})\right) = 0$$

to be consistent.

4.3.1 Simulation models

In the simulation studies in this chapter, we follow the class of outcome and propensity models that have been already given in early section (3.12). Assuming the covariate vectors X_i are of dimension 3, specifically we consider

$$Y_i \; = \; \beta_0 \; + \sum_{j=1}^{3} \beta_j X_i^{(j)} + \sum_{j<l} \beta_{jl} X_i^{(j)} X_i^{(l)} \; + \; K(X_i) e_i \tag{4.10}$$

where $K(X_i) = .5 + .5\sum_{j=1}^{3} X_i^{(j)} + .5\sum_{j<l} X_i^{(j)} X_i^{(l)}$, $e_i \overset{iid}{\sim} \mathcal{N}(0,1)$. The independent components of X_i, $X_i^{(1)}$, $X_i^{(2)}$ and $X_i^{(3)}$ follow $Unif(1,5)$, $Unif(1,6)$, and $Unif(1,3)$ distributions, respectively.

Following the propensity model that has been given in (3.13), the size variable associated with unit $i \in \mathcal{U}$ is defined as $V_i = 1/\left(10 + (E(Y_i|X_i))^2 + \delta_i\right)$ with $\delta_i \overset{iid}{\sim} \mathcal{N}(0, 1.5^2)$. Following the probability-proportional-to-size (PPS) sampling idea, the inclusion probability π_i^0 is then defined as

$$\pi_i^0 = nV_i \Big/ \sum_{j=1}^{N} V_j \tag{4.11}$$

To take care of the weight misspecification, we adopted the biasing factor model (2.4) introduced in Chapter 2, as given in

$$\eta_i = \frac{\exp\{K(X_i)\zeta/15\}}{m_\zeta\left(K(X_i)/15\right)} \tag{4.12}$$

where $m_\zeta(\cdot)$ is the moment generating function of ζ and ζ_i is independent of $(X_i, Y_i, V_i, e_i, \delta_i)$. The function $K(\cdot)$ in (4.12) is the same with $K(\cdot)$ in (4.10). Therefore under (4.12), $E(\eta_i) = 1$ and $E(\eta_i|X_i) = E\Big(\exp\{a(X_i)\zeta_i\}\Big)/m_\zeta(a(X_i)) = 1$. We also have $E(\eta_i Y_i) = E(E(Y_i|X_i) \cdot E(\eta_i|X_i)) = E(Y_i)$ since Y_i and η_i are independent given X_i, so that HT is still consistent under w_i.

4.3.2 Simulation cases

In Section 4.3.1, there are parameters in the outcome and biasing factor models which will be defined here. Three different sets of choices of those parameters form three cases, which will later be related to the seriousness of bias in Zheng and Little's estimators in simulation results.

Case 4.1 In (4.10) the coefficient vector $\beta = (3, 0, 0, 0, 0, 0, 0)$, that is, this is a "mean-only" model. In (4.12), $\zeta_i \overset{iid}{\sim} \mathcal{N}(0, .5^2)$ on interval $(-1.8, .8)$.

Case 4.2 In (4.10) the coefficient vector $\beta = (3, 3, 3, 3, 3, 3, 3)$, so the conditional mean include main effects and also two-way interaction terms. In (4.12), $\zeta_i \overset{iid}{\sim} \mathcal{N}(0, .5^2)$ on interval $(-1.5, .3)$.

Case 4.3 In (4.10) the coefficient vector $\beta = (3, 3, 3, 3, 0, 0, 0)$, so the conditional mean only includes the main effects. In (4.12), $\zeta_i \overset{iid}{\sim} \mathcal{N}(0, .8^2)$ on interval $(-2, 1.2)$.

The choice of size variable model already gives some advantages to ZL1 and ZL2. Above parameter choices in three cases are based on how well the spline model fit would be outside the sample. From **Case 4.1** to **Case 4.3**, the spline model fit w_i gets worse outside of sample. Figure 4.1 shows the comparison of model fitting between using d_i and w_i within sample. Each row represents one case and. The left column shows the model fitting under d_i and the right shows w_i. From Figure 4.1, we barely could see significant difference between the two sets of models, using d_i and w_i. Figure 4.2 has exactly the same display but all the data points are from not sampled units. We can clearly observe that, extrapolating the fitted model to

not sampled part still works very good when using d_i. But the left column except the top subplot shows more noisy pattern and the fitted model may not predict well outside of the sample.

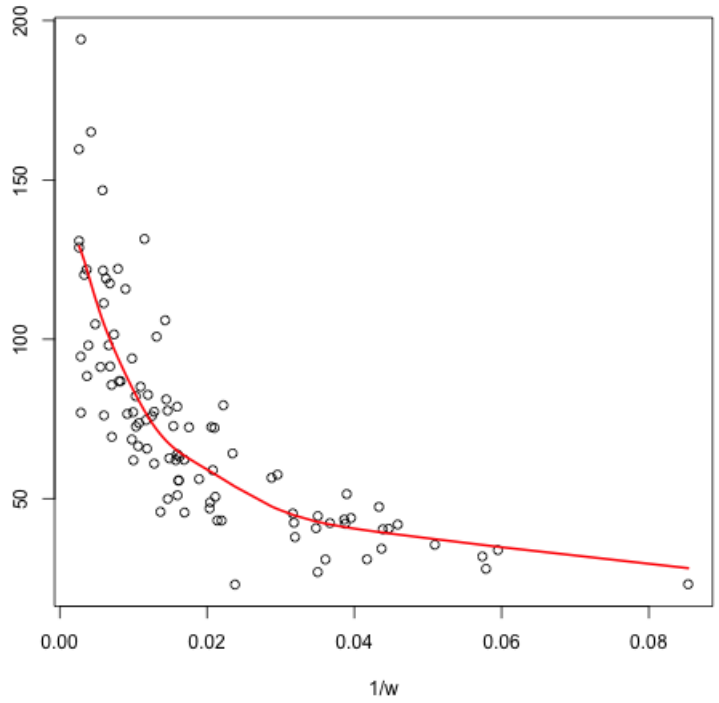

Figure 4.1: Scatter plots showing different levels of spline model fitting within sample. X-axis is inverse sample weight where "d" refers to design weight and "w" refers to misspecified weight. Y-axis is outcome variable, Y. Each row represents one case. The solid line shows the fitted outcome variable using the corresponding set of weights.

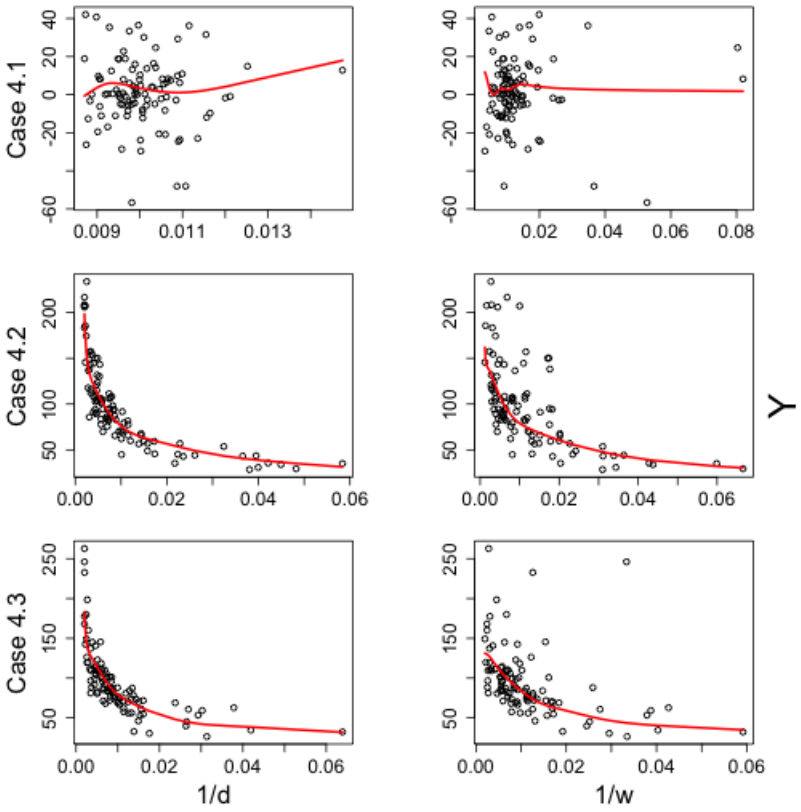

Figure 4.2: Scatter plots showing different levels of spline model fitting outside the sample. X-axis is inverse sample weight where "d" refers to design weight and "w" refers to misspecified weight. Y-axis is outcome variable, Y. Each row represents one case. The solid line shows the fitted outcome variable using the corresponding set of weights.

4.3.3 Simulation results

In the p-spline model, we take degree $p = 3$, number of knots $m = 5$, and $k = 0$. There are 50 frame population data sets generated. For each generated frame population data set, 1000 samples are drawn using PPSWOR sampling. We use Poisson sampling again here, for simplicity. For each sample $\mathcal{S}$, two sets of estimators (HT, ZL1 and ZL2) are computed to estimate t_Y, one under d_i and the other set under w_i.

Table 4.1: Simulation results relating weight misspecification to bias in ZL2. RB refers to average relative bias in ZL2, defined as bias in ZL2 divided by true population total in Y. EE in $\mathcal{S}$ is the average of left hand side of (4.5) and EE outside of $\mathcal{S}$ is the average to estimate the expectation in (4.8)

		RB(%)			EE in sample		EE outside of sample	
Case	Weights	HT	ZL1	ZL2	`min`	`max`	`min`	`max`
Case 4.1	d	0.05	0.05	0.04	0	0	0.00	0.42
	w	0.07	0.06	-0.83	0	0	15.98	297.92
Case 4.2	d	0.02	0.00	-0.03	0	0	0.01	3.66
	w	-0.12	-0.16	-5.77	0	0	0.03	2670.75
Case 4.3	d	-0.01	-0.01	-0.00	0	0	0.00	0.73
	w	-0.24	-0.26	-12.16	0	0	9569.37	39082.10

Table 4.1 summarizes the simulation results of **Case 4.1** to **Case 4.3**, as defined above. All the numbers in the table are empirical averages. The first column shows which case the results refer to. Column 2 indicates which set of weights have been used while d stands for design weights and w represents misspecified weights. In columns 3-5, RB stands for "relative bias" with percentage in parentheses indicating that the empirical relative biases are divided by t_Y and multiplied by 100. We simulate only in settings where HT under w_i is consistent. By Proposition 4, ZL1 should also be exactly or approximately consistent under w_i. All percent relative biases in HT and ZL1 in Table 4.1 are close to zero as expected. Zheng and Little's second estimator, ZL2, shows a different story. When using d_i, ZL2 is consistent as other estimators. But ZL2 under w_i shows some biases, in **Case 4.2** and **Case 4.3**.

EE in sample is calculated as the following

$$\begin{aligned}&\frac{1}{R}\sum_{r=1}^{R} U_{\mathcal{S}}(a,j,r)\\ =&\frac{1}{R}\sum_{r=1}^{R}\sum_{i\in\mathcal{S}} a_i^{2k} b_j(1/a_i)\left(Y_i - \mathbf{b}^{tr}(1/a_i)\hat{\gamma}^{ZL,a,r}\right), \quad j=0,\ldots,p+m,\end{aligned} \tag{4.13}$$

where $a = d$ or w, r represents the r^{th} replication, $\hat{\gamma}^{ZL,a,r}$ is the estimated coefficient in the r^{th} replication, j denotes the spline basis index and R is the total number of replications. Column 6 is the minimum value of (4.13) over j and column 7 is the maximum value of (4.13) over j. We know that for all r, j, $U_{\mathcal{S}}(d,j,r)$ and $U_{\mathcal{S}}(w,j,r)$ are expected to be zero since we estimate coefficients by solving for (4.4) and (4.5). On the other hand, EE outside the sample is calculated as

$$\begin{aligned}&\frac{1}{R}\sum_{r=1}^{R} U_{\mathcal{U}/\mathcal{S}}(a,j,r)\\ =&\frac{1}{R}\sum_{r=1}^{R}\sum_{i\notin\mathcal{S}} a_i^{2k} b_j(1/a_i)\left(Y_i - \mathbf{b}^{tr}(1/a_i)\hat{\gamma}^{ZL,a,r}\right), \quad j=0,\ldots,p+m,\end{aligned} \tag{4.14}$$

Column 8 is the minimum value of (4.14) over j and column 9 is the maximum value of (4.14) over j. First when using d_i, all values of (4.14) are zeros or nearly zeros, indicating that extrapolating to the whole frame population might be a good idea since the fitted model also fits the non-sampled data. That is why we observe consistency in ZL2 under d_i. When using w_i, (4.14) could be very large at least for one j. By Proposition 5, non-zero values mean possible bias in ZL2 and large values indicate definite bias in ZL2. The simulation results support our couclusion in Proposition 5.

4.4 Discussion

In this chapter, we examined two estimators proposed by Zheng & Little in their 2003 paper (Zheng and Little 2003). We illustrated and showed why the first estimator should work well even under w_i under the conditions given in Proposition 4. Next, we investigated the second estimator, ZL2, which utilizes all weights in the finite population using the extrapolation idea. In Proposition 5, we linked the bias in ZL2 to the estimating equation system outside the sample. In the simulation studies, we created three cases, all of which guarantee that HT is consistent so that ZL1 would be consistent too by Proposition 4. In simulation results, we observed that from **Case 4.1** to **Case 4.3**, the percent relative bias in ZL2 was non-zero while the estimating equation values are also non-zero, which validated the conclusion in Proposition 5 too.

4.5 Proof of propositions

4.5.1 Proof of Proposition 4

Proof: The following proof applies to ZL1 under d_i. The arguments of ZL1 under w_i would be exactly the same.

- When $k = 1/2$ or 1, the estimating equation $j = 0$ when $k = 1/2$ and equation $j = 1$ when $k = 1$ in the system (4.4) and **A.10** guarantees

$$\sum_{i \in S} \frac{Y_i - \hat{Y}_i^{ZL,\,d}}{\pi_i^0} = 0, \tag{4.15}$$

implying that $\hat{t}_Y^{HT,d} = \hat{t}_Y^{ZL1,d}$.

- If we take $k = 0$, we have

$$\sum_{i \in \mathcal{S}} b_j(\pi_i^0) \left(Y_i - \hat{Y}_i^{ZL,d} \right) = 0, \quad \text{for } j = 0, 1, \ldots, p+m. \tag{4.16}$$

by (4.4). By **A.12**, $g(u) = 1/u$ is a continuous function on the interval $[a, 1]$. Then by Weierstrass approximation theorem (De Branges 1959; De Boor et al. 1978), $g(u)$ can be uniformly approximated as closely as desired by the spline basis,

$$1, u, \ldots, u^p, (u - \tau_1)_+^p, \ldots, (u - \tau_m)_+^p.$$

That means, $\forall\, \varepsilon > 0$, there exist a degree p, number of knots m and a set of coefficients $c = (c_0, \ldots, c_{p+m})$ such that

$$\sup_{a \le u \le 1} \left| \frac{1}{u} - \sum_{j=0}^{p+m} b_j(u) c_j \right| \le \sqrt{\varepsilon}. \tag{4.17}$$

Replace π_i^0 with u in (4.17) and let $\Omega_i = d_i - \sum_{j=0}^{p+m} b_j(\pi_i^0) c_j$. Then we get

$$|\Omega_i| \le \sqrt{\varepsilon}, \qquad i \in \mathcal{S}\,. \tag{4.18}$$

Spline functions can also be used to approximate the function $E(Y_i|\pi_i^0)$, so that for the same ε, there exists another set of coefficients $\gamma^* = (\gamma_0^*, \ldots, \gamma_{p+m}^*)$ such

that

$$E\left(Y_i - \sum_{j=0}^{p+m} b_j(\pi_i^0)\gamma_j\right)^2$$
$$= E\left(Y_i - E(Y_i|\pi_i^0)\right)^2 + E\left(E(Y_i|\pi_i^0) - \sum_{j=0}^{p+m} b_j(\pi_i^0)\gamma\right)^2$$

is minimized and

$$E\left(E(Y_i|\pi_i^0) - \sum_{j=0}^{p+m} b_j(\pi_i^0)\gamma_j^*\right)^2 \leq \varepsilon, \tag{4.19}$$

using the fact that π_i^0 are *iid* samples. By Jensen's inequality and (4.19), we also have

$$E\left|E(Y_i|\pi_i^0) - \sum_{j=0}^{p+m} b_j(\pi_i^0)\gamma_j^*\right| \leq \sqrt{\varepsilon}. \tag{4.20}$$

On the other hand, solving (4.16) is equivalently to minimize

$$\sum_{i\in\mathcal{S}}\left(\sum_{j=0}^{p+m} b_j(\pi_i^0)\gamma_j - Y_i\right)^2.$$

Let $\gamma^{\mathcal{U}}$ be the minimizer of

$$\sum_{i\in\mathcal{U}}\left(\sum_{j=0}^{p+m} b_j(\pi_i^0)\gamma_j - Y_i\right)^2.$$

Then the least-square estimate $\hat{\gamma}^{ZL,d}$ converges to the population parameter $\gamma^{\mathcal{U}}$ in probability, as n, N go to ∞. So for the same ε and chosen p, m, we

have

$$\max_j \left|\hat{\gamma}_j^{ZL,\,d} - \gamma_j^{\mathcal{U}}\right| \leq \frac{\sqrt{\varepsilon}}{p+m} \sum_{j,\, a\leq u\leq 1} b_j(u). \tag{4.21}$$

for sufficiently large n and N, with probability greater than $1-\varepsilon$. The population parameter $\gamma^{\mathcal{U}}$ and coefficient γ^* satisfy

$$\max_j \left|\gamma_j^{\mathcal{U}} - \gamma_j^*\right| \leq \frac{\sqrt{\varepsilon}}{p+m} \sum_{j,\, a\leq u\leq 1} b_j(u). \tag{4.22}$$

for sufficiently large N under superpopulation *iid* sample assumption. We could take large N so that both (4.21) and (4.22) hold.

Then by (4.16) and (4.18), we have

$$\begin{aligned}
&\frac{1}{N}\Big|\sum_{i\in\mathcal{S}} d_i\left(\hat{Y}_i^{ZL,\,d} - Y_i\right)\Big| \\
=\ &\frac{1}{N}\Big|\sum_{i\in\mathcal{S}} d_i\left(\hat{Y}_i^{ZL,\,d} - Y_i\right) - \sum_{j=0}^{p+m} c_j \sum_{i\in\mathcal{S}} b_j(\pi_i^0)\left(\hat{Y}_i^{ZL,\,d} - Y_i\right)\Big| \\
=\ &\frac{1}{N}\Big|\sum_{i\in\mathcal{S}} d_i\left(\hat{Y}_i^{ZL,\,d} - Y_i\right) - \sum_{i\in\mathcal{S}}\left\{\sum_{j=0}^{p+m} \hat{c}_j b_j(\pi_i^0)\right\}\left\{\hat{Y}_i^{ZL,\,d} - Y_i\right\}\Big| \\
=\ &\frac{1}{N}\Big|\sum_{i\in\mathcal{S}} \Omega_i\Big\{\hat{Y}_i^{ZL,\,d} - \sum_{j=0}^{p+m} b_j(\pi_i^0)\gamma_j^{\mathcal{U}} + \sum_{j=0}^{p+m} b_j(\pi_i^0)\gamma_j^{\mathcal{U}} - \sum_{j=0}^{p+m} b_j(\pi_i^0)\gamma_j^* \\
&\qquad\qquad \sum_{j=0}^{p+m} b_j(\pi_i^0)\gamma_j^* - E(Y_i|\pi_i^0) + (Y_i|\pi_i^0) - Y_i\Big\}\Big| \\
\leq\ &A_1 + A_2 + A_3 + A_4,
\end{aligned} \tag{4.23}$$

where in (4.23), A_1 to A_4 are defined as

$$
\begin{aligned}
A_1 &= \frac{1}{N}\Big|\sum_{i\in\mathcal{S}}\Omega_i\left(\sum_{j=0}^{p+m} b_j(\pi_i^0)\hat{\gamma}_j^{ZL,\,d} - \sum_{j=0}^{p+m} b_j(\pi_i^0)\gamma_j^{\mathcal{U}}\right)\Big|,\\
A_2 &= \frac{1}{N}\Big|\sum_{i\in\mathcal{S}}\Omega_i\left(\sum_{j=0}^{p+m} b_j(\pi_i^0)\gamma_j^{\mathcal{U}} - \sum_{j=0}^{p+m} b_j(\pi_i^0)\gamma_j^{*}\right)\Big|,\\
A_3 &= \frac{1}{N}\Big|\sum_{i\in\mathcal{S}}\Omega_i\left(\sum_{j=0}^{p+m} b_j(\pi_i^0)\gamma_j^{*} - E(Y_i|\pi_i^0)\right)\Big|,\\
A_4 &= \frac{1}{N}\Big|\sum_{i\in\mathcal{S}}\Omega_i\left(E(Y_i|\pi_i^0) - Y_i\right)\Big|.
\end{aligned}
$$

The proof will be done as long as (4.23) is sufficiently small with large probability and sufficiently large n, N. By (4.21),

$$
A_1 \le \frac{n}{N}\sqrt{\varepsilon}, (p+m)\max_j|\hat{\gamma}^{ZL,\,d} - \gamma_j^{\mathcal{U}}|\sup_{j,\,a\le u\le 1} b_j(u) \le \varepsilon, \tag{4.24}
$$

with probability greater than $1-\varepsilon$ and sufficiently large n and N. Similarly we have

$$
A_2 \le \varepsilon, \tag{4.25}
$$

with probability greater than $1-\varepsilon$ and sufficiently large N. In A_3 by (4.20), the fact that $E(Y_i|\pi_i^0) - \sum_{j=1}^{p+m} b_j(\pi_i^0)\gamma^*$ are *iid* samples and the law of large

numbers, we have

$$
\begin{aligned}
A_3 \leq & \sqrt{\varepsilon}\frac{1}{N}\sum_{i=1}^{N}\Big|\{\sum_{j=0}^{p+m} b_j(\pi_i^0)\gamma_j^* - E(Y_i|\pi_i^0)\}\Big| \\
\leq & \sqrt{\varepsilon}\Big\{\frac{1}{N}\sum_{i=1}^{N}\Big|\Big\{\sum_{j=0}^{p+m} b_j(\pi_i^0)\gamma_j^* - E(Y_i|\pi_i^0)\Big\} - \\
& E\Big(\sum_{j=0}^{p+m} b_j(\pi_i^0)\gamma_j^* - E(Y_i|\pi_i^0)\Big)\Big| + \\
& E\Big|\Big(\sum_{j=0}^{p+m} b_j(\pi_i^0)\gamma_j^* - E(Y_i|\pi_i^0)\Big)\Big|\Big\} \\
\leq & \sqrt{\varepsilon}(\sqrt{\varepsilon} + \sqrt{\varepsilon}) = 2\varepsilon
\end{aligned}
\tag{4.26}
$$

with probability greater than $1-\varepsilon$ and sufficiently large n, N. In A_4, by **A.11** and the fact that Ω_i is a function of π_i^0, $\Omega_i\Big(E(Y_i|\pi_i^0) - Y_i\Big)$ are *iid* samples with mean zero too, we have

$$
A_4 \leq \varepsilon, \tag{4.27}
$$

with probability greater than $1-\varepsilon$, for sufficiently large n, N, by the law of large numbers.

Therefore by (4.24) to (4.27), we have

$$
\begin{aligned}
& \frac{1}{N}\Big|\sum_{i\in\mathcal{S}} d_i\left(\hat{Y}_i^{ZL,\,d} - Y_i\right)\Big| \\
\leq & \; A_1 + A_2 + A_3 + A_4 \\
\leq & \; 5\,\varepsilon
\end{aligned}
$$

with probability greater than $1 - 4\varepsilon$ for sufficiently large n and N. Therefore we proved that the discrepancy between $\hat{t}_Y^{ZL1,\,d}$ and $\hat{t}_Y^{HT,\,d}$ is negligible in probability when p, m can be chosen arbitrarily large, as population size and sample size increase.

Using the same arguments on w_i, we have $\hat{t}_Y^{ZL1,\,w} = \hat{t}_Y^{HT,\,w}$ when $k = 1/2$ or 1 and $\hat{t}_Y^{ZL1,\,w} \approx \hat{t}_Y^{HT,\,w}$ when $k = 0$.

□

4.5.2 Proof of Proposition 5

Proof: Again the following proof is based on ZL2 under d_i. The arguments of ZL2 under w_i would be exactly the same.

ZL2 under d_i has the following expectation,

$$\begin{aligned} E\left(\hat{t}_Y^{ZL2,\,d}\right) &= E\left(\sum_{i\in\mathcal{S}} Y_i + \sum_{i\notin\mathcal{S}} \hat{Y}_i^{ZL,\,d}\right) \\ &= E\left(\sum_{i=1}^{N} Y_i + \sum_{i\notin\mathcal{S}} \left(\mathbf{b}^{tr}(\pi_i^0)\hat{\gamma}^{ZL,\,d} - Y_i\right)\right) \\ &= E\left(\sum_{i\in\mathcal{U}} Y_i\right) + E\left(\sum_{i\notin\mathcal{S}} \left(\mathbf{b}^{tr}(\pi_i^0)\hat{\gamma}^{ZL,\,d} - Y_i\right)\right), \end{aligned} \tag{4.28}$$

where $E\left(\sum_{i\notin\mathcal{S}} \left(\mathbf{b}^{tr}(\pi_i^0)\hat{\gamma}^{ZL,\,d} - Y_i\right)\right)$ forms the bias part.

If $k = 0$, the equation of $j = 0$ in (4.8) implies

$$E\left(\sum_{i\notin\mathcal{S}} \left(Y_i - \mathbf{b}^{tr}(\pi_i^0)\hat{\gamma}^{ZL,\,d}\right)\right) = 0.$$

The equation of $j = 1$ if $k = 1/2$ and equation of $j = 2$ if $k = 1$ in (4.8) lead to zero-bias too. So ZL2 under d_i is consistent when (4.8) holds. □

Chapter 5. Inflated Variance

In this chapter, we continue to assume that the outcome variable Y_i, covariate vector X_i, design weight d_i, biasing factor η_i and inclusion indicator $I_{[i\in\mathcal{S}]}$ form the superpopulation iid vector $(Y_i, X_i, d_i, \eta_i, I_{[i\in\mathcal{S}]})$ which is the element of finite population $\mathcal{F}$. In Chapter 2, a necessary and sufficient condition under which HT using w is still consistent was given as,

$$E(\eta\, Y) \;=\; E(Y),$$

where the expectation is taken with respect to superpopulation distribution. In Chapter 3, we also showed that GREG under w_i remains consistent if the conditional mean of Y given X is correctly specified. In Chapter 4, we linked the bias in ZL to an estimating equation system outside the sample. In this chapter, we focus on variances of population total estimators. It is easy to see that the biasing factor introduces extra noise, so that estimators may have inflated variance. We are interested in knowing whether the biasing factor affects some estimators so that one particular estimator may have the smallest variance under d_i but not under w_i due to variance inflation.

5.1 Design variance and anticipated variance

In survey sampling, there are two types of variances associated with estimators. From a purely design-based point of view, we look at design variance, taken with respect to sampling design, treating $I_{[i\in\mathcal{S}]}$ as the only random variables. For example, the HT estimator has design variance

$$\begin{aligned} V_d\left(\hat{t}_Y^{HT,d}\right) &= V\left(\hat{t}_Y^{HT,d}\middle|\mathcal{F}\right) \\ &= E_d\left(\left(\hat{t}_Y^{HT,d}-t_Y\right)^2\middle|\mathcal{F}\right) \\ &= \sum_{i\in\mathcal{U}}\sum_{j\in\mathcal{U}}\frac{\pi_{ij}^0-\pi_i^0\pi_j^0}{\pi_i^0\,\pi_j^0}Y_iY_j \end{aligned}$$

where π_{ij}^0 denotes the joint inclusion probability for unit $i,j\in\mathcal{U}$, $P(i\,\&\,j\,\in\,\mathcal{S})$.

The definition of anticipated variance was first introduced by Isaki and Fuller (1982) and was also summarized in Fuller (2011). The anticipated variance (AV) for an estimator $\hat{\theta}$ estimating the population parameter θ_N is given by

$$AV\{\hat{\theta}\,-\,\theta_N\}\,=\,E_p\{E_d[(\hat{\theta}\,-\,\theta_N)^2\mid\mathcal{F}]\}\,-\,E_p\{E_d(\hat{\theta}-\theta_N\mid\mathcal{F})\} \tag{5.1}$$

Again as we have mentioned earlier, the notation $E_p(\cdot)$ denotes expectation with respect to the superpopulation distribution, and $E_d(\cdot)$ indicates expectation with respect to the sampling design. In this chapter, we focus on anticipated variance and compare the AV's across different methods under weight misspecification.

5.2 Important results on optimal weighting

Model-based methods like estimators introduced by Zheng and Little (2003) and Pfeffermann and Sverchkov (1999) rely on model assumptions. When weights are properly specified, model-based estimators tend to be more efficient than HT when the required distributional assumptions hold. In this section, we cite and summarize important results on optimal weighting. Both GREG and PS involve estimating the regression coefficient vector β by solving the estimating equation of the form

$$\sum_{i \in \mathcal{S}} u_i \begin{pmatrix} 1 \\ X_i \end{pmatrix} \left(Y_i - \beta^{tr} \begin{pmatrix} 1 \\ X_i \end{pmatrix} \right) = 0, \tag{5.2}$$

where u_i are the working weights. GREG uses sampling weights and PS uses sampling weights divided by estimated values from a model. Magee (1998) considers the following outcome model

$$\begin{aligned} Y &= (1, X^{tr})\beta + \varepsilon, \quad \text{where} \\ E(\varepsilon \,|\, X) &= 0, \\ E(\varepsilon^2 \,|\, X) &= \sigma^2(X). \end{aligned} \tag{5.3}$$

and investigates the survey-weighted least squares regression (5.2). Let d_i be the usual inverse sample inclusion probabilities. Magee considered weighting by d_i multiplied by a_i, i.e.,

$$u_i = d_i a_i,$$

where a_i is a function of covariates. This a_i should be chosen to minimize the asymptotic variance, given by Magee in the form

$$\left(\sum_{i\in\mathcal{S}} d_i a_i X_i^{\otimes 2}\right)^{-1} \sum_{i\in\mathcal{S}} d_i^2 a_i^2 \hat{\varepsilon}_i^2 X_i^{\otimes 2} \left(\sum_{i\in\mathcal{S}} d_i a_i X_i^{\otimes 2}\right)^{-1}. \tag{5.4}$$

It can be shown that the optimal a_i should be taken proportional to $1/\left(d_i \times \sigma^2(X_i)\right)$ within (5.4). Equivalently speaking, the optimal noninformative weights $u_i^* = d_i a_i$ should be proportional to

$$u_i^* \propto \frac{1}{\sigma^2(X_i)}. \tag{5.5}$$

"Noninformative" means the outcome variable Y_i and inclusion probability π_i^0 are conditionally independent given covariate X_i,

$$Y_i \perp\!\!\!\perp \pi_i^0 \,\Big|\, X_i.$$

The result in (5.5) implies that, if we know the variance structure clearly, then we should weight by the inverse of the conditional variance function.

Considering the optimal weighting from this angle, if the outcome model and propensity model imply that $d_i \times \sigma^2(X_i) \approx$ constant, then the optimal weighting should be $u_i^* \approx d_i$, implying that using d_i's as in GREG would give the optimal asymptotic variance. At the same time, PS estimates β by taking $u_i = d_i/\hat{d}_i$, where $\hat{d}_i$ is estimated from regression model by taking X_i as regressors. As long as $d_i/\hat{d}_i$ is somewhat different from d_i, PS then gives suboptimal asymptotic variance.

On the other hand, if the working weights for PS, $u_i = d_i/\hat{d}_i$, is roughly proportional to the optimal weights given by Magee, $1/\sigma^2(X)$, then PS should outperform GREG.

5.3 Simulated cases in comparing anticipated variances

In this section, we explore four cases (**Case 1** to **Case 4**) to study how weight misspecifications affect the performance of estimators. Under each case, there is always at least one estimator performing well under d_i, i.e., the corresponding variance is relatively small compared to other estimators. However that good estimator turns out to perform not so well under w_i, i.e., the corresponding variance is inflated by misspecified weights and therefore that estimator is outperformed by some of the rest estimators. Let N be the finite population size and n be the sample size. Define $f = n/N$ to be the sampling fraction. Again, we consider probability-proportional-to-size (PPS) sampling and take the inclusion probability to be $\pi_i^0 = nV_i/\sum_{j=1}^N V_j = fV_i/\bar{V}$ where V_j is the size measurement associated with unit $j \in \mathcal{U}$.

We will give four cases **Case 1** to **Case 4**. Each case is defined by an outcome model, a propensity model and a biasing factor model. All the biasing factor models in the displayed examples follow (2.4). Explanations will be given in each case why a particular estimator should have smaller variance under the described outcome and propensity model than the others. Simulation studies follow in Section 5.4 and 5.5, validating that particular estimator performs well under d_i but loses its advantage

under w_i.

Case 1 GREG is favored and better than PS under d_i. Consider the outcome model that Y_i

$$\begin{aligned} Y_i &= 5 + X_i + \varepsilon_i, \quad \text{where} \\ X_i &\overset{iid}{\sim} Unif(5,\, 25) \\ \varepsilon_i &\overset{iid}{\sim} \mathcal{N}\left(0,\, \frac{15^2}{5 + (X_i - 15)^2/2}\right) \end{aligned} \tag{5.6}$$

The outcome model Y and the scalar covariate X_i are nearly perfectly linearly correlated. The propensity model is taken to be

$$\begin{aligned} V_i &= \frac{\delta_i}{5 + (X_i - 15)^2/2}, \quad \text{where} \\ \delta_i &\overset{iid}{\sim} \mathcal{LN}(-.5^2/2,\, .5^2), \end{aligned} \tag{5.7}$$

and δ_i is independent of $(Y_i,\, X_i)$. The outcome model (5.6) and propensity model (5.7) show that $d_i \times \sigma^2(X_i) \approx$ constant, which means that d_i in GREG should be the optimal weighting in (5.2). Figure 5.1 displays scatter plots of population sample in this case. Plot (a) shows that given the large error, $Y - X$ still has mild and clear linear correlation. Plot (b) is the scatter plot between inclusion probabilities and Y_i indicating that ZL might not be very appealing since the dependence of Y_i on π_i^0 is weak. Plot (c) shows the dependence of design weights d_i on Y_i, which Beaumont's estimator relies on. Plot (b)-(c) shows that the model-based estimators ZL and B should not be the optimal in terms of variance among those considered estimators.

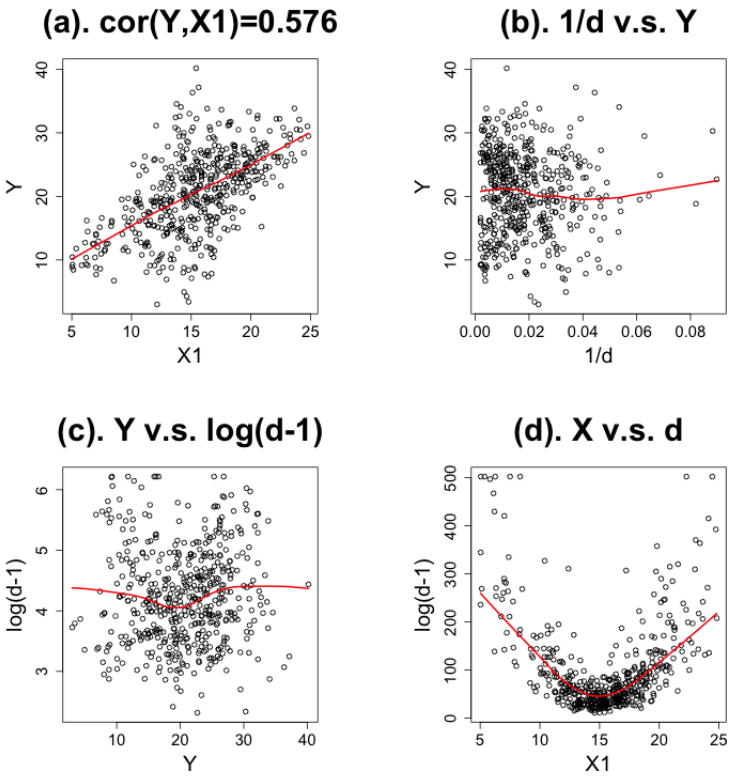

Figure 5.1: Scatter plots showing example defined in **Case 1** favoring GREG over other considered estimators. The red solid line is local polynomial regression curve fitted by `loess` defaults in `R`.

Plot (d) shows that d_i is approximately a quadratic function of X_i. Implied by this, if we consider the working model

$$d_i = \phi_0 + \phi_1 X_i + \phi_2 X_i^2 + e_i, \tag{5.8}$$

where e_i is a purely random normally distributed noise term, then $d_i/\hat{d}_i$ should be close to one and differ from d_i in values. This means that PS is not optimal either in this case.

Figure 5.2 displays the scatter plots showing various relationships between outcome variable, coavariate and weights in **Case 2**. Similarly, we see from plot (b)-(c) that the methods of Zheng and Little and Beaumont are less appealing since distributional assumptions are not satisfied.

The biasing factor model is taken to be

$$\begin{aligned} \eta_i &= \left(m_\zeta\big(a(X_i)\big)\right)^{-1} \exp\Big\{a(X_i)\zeta_i\Big\}, \quad \text{where} \\ \zeta_i &\overset{iid}{\sim} Unif(-.54,\, .35), \\ a(X_i) &= 1\Big/\left\{-.03\left(.5(X_i-15)^2+10\right)\right\}. \end{aligned} \tag{5.9}$$

In (5.9), ζ_i is independent of $(Y_i, X_i, \varepsilon, \delta_i)$ and $E(\eta_i|X_i) = 1$. According to (3.11) given in Proposition 3, we expect that GREG is consistent under w_i. Also η_i and Y_i are conditionally independent given X_i,

$$E(\eta_i Y_i) = E(E(\eta_i|X_i)E(Y_i|X_i)) = E(Y_i).$$

Then HT is consistent by (1).

Case 2 PS is favored over GREG. Again we take $p = 1$ so that X_i are real numbers. Assume that the outcome model and propensity model are exactly the same as in (5.6) and (5.7) respectively except that $\varepsilon\, iid \sim \mathcal{N}(0,\, 8^2)$ and $\delta_i\, iid \sim$

$\mathcal{LN}(-.2^2/2,\ .2^2)$. Specifically, the outcome model follows

$$\begin{aligned} Y_i &= 5 + X_i + \varepsilon_i, \quad \text{where} \\ X_i &\overset{iid}{\sim} Unif(5, 25) \\ \varepsilon_i &\overset{iid}{\sim} \mathcal{N}\left(0,\ 8^2\right) \end{aligned} \tag{5.10}$$

indicating that $\sigma^2(X)$ is a constant. So according to Magee's results, the optimal weights should be a constant. The propensity model follows

$$\begin{aligned} V_i &= \frac{\delta_i}{5 + (X_i - 15)^2/2}, \quad \text{where} \\ \delta_i &\overset{iid}{\sim} \mathcal{LN}(-.2^2/2,\ .2^2), \end{aligned} \tag{5.11}$$

meaning that again, we could obtain good estimates of d_i denoted by $\hat{d}_i$ by working with (5.8). So $d_i/\hat{d}_i$ should be close to one, and then PS should be roughly optimal. On the other hand, GREG uses d_i which should be very different from $d_i/\hat{d}_i$ so GREG will not be as good as PS. From Figure 5.2 we can see that the dependence between inclusion probabilities and Y_i, and the dependence between Y_i and design weights, are both weak. So it is expected that the two model-based methods, those of Zheng and Little and of Beaumont, do not work very well.

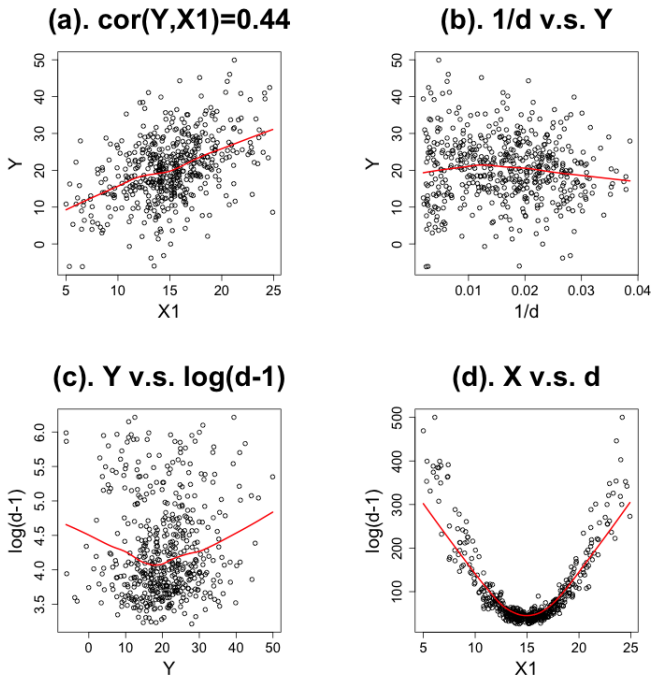

Figure 5.2: Scatter plots showing situations defined in **Case 2** favoring PS over other considered estimators. The red solid line is local polynomial regression curve fitted by `loess` defaults in `R`.

The biasing factor model in this case is again taken to be

$$\begin{aligned} \eta_i &= \big(m_\zeta\big(a(X_i)\big)\big)^{-1} \exp\Big\{a(X_i)\zeta_i\Big\}, \quad \text{where} \\ \zeta_i &\overset{iid}{\sim} Unif(-1.73,\, .6), \\ a(X_i) &= 1\Big/\left\{.38\left((X_i-15)^2+10\right)\right\}. \end{aligned} \tag{5.12}$$

Except for a few different parameter choices, this biasing factor model is exactly the same as that in **Case 1**. So again HT under w_i is consistent.

Case 3 ZL2 is expected to work very well when the dependence between Y_i and inclusion probabilities is strong and easily described by a spline model. In this

case, assume again that the covariate X_i is a scalar. Consider the outcome model,

$$\begin{aligned} Y_i &= 200 - 22X_i + 2X_i^2 - .05X_i^3 + \varepsilon_i, \quad \text{where} \\ X_i &\overset{iid}{\sim} Unif(5,\, 25), \\ \varepsilon_i &\overset{iid}{\sim} \mathcal{N}\left(0,\, \frac{X_i^2}{4}\right). \end{aligned} \tag{5.13}$$

From the outcome model the relationship between Y_i and X_i is seen to be clear and strong, but the linear correlation is weak. Also $\sigma^2(X_i) = X_i^2/4$ so the optimal weights $u_i^* \propto 1/X_i^2$. The propensity model follows

$$\begin{aligned} V_i &= \frac{\delta_i}{200 - 22X_i^{(1)} + 2X_i^{(1)\,2} - .05X_i^{(1)\,3}} \quad \text{where} \\ \delta_i &\overset{iid}{\sim} \mathcal{LN}(-.2^2/2,\, .2^2). \end{aligned} \tag{5.14}$$

From (5.14) and Figure 5.3, Y_i and π_i^0 have a nice functional relationship, which means that ZL2 may outperform the rest of the estimators. From plot (d) of Figure 5.3, the dependence between X_i and weights are strong, so if we choose a good model to obtain $\hat{d}_i$ when constructing PS, we should have $d_i/\hat{d}_i$ close to one, which is very different from u_i^*, then PS is not optimal; if we happen to use a bad model to obtain $\hat{d}_i$, the estimation itself that PS gives might be very bad. On the other hand, d_i are proportional to X_i^2 or X_i^3 so they are proportional to u_i^*, but considering that the linear correlation between Y_i and X_i is not very strong, we expect that GREG should work better than PS but no better than HT in this particular case.

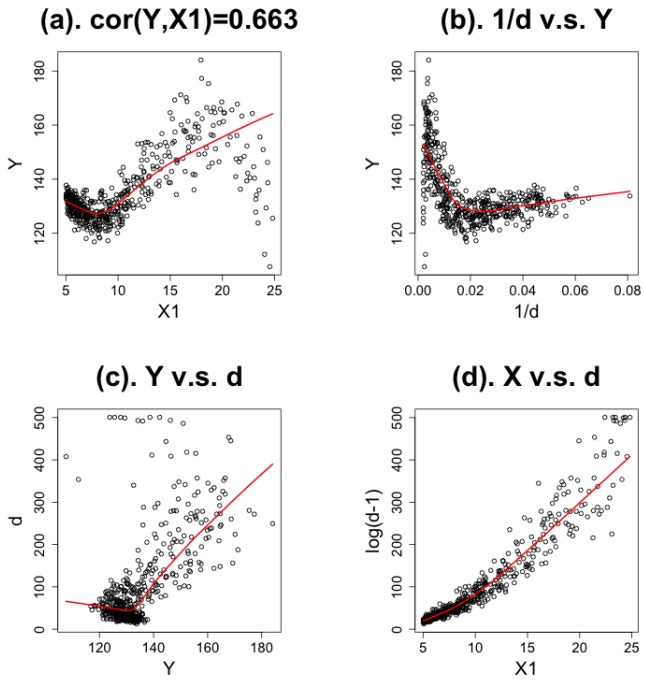

Figure 5.3: Scatter plots showing situations defined in **Case 3** favoring ZL over other considered estimators. The red solid line is local polynomial regression curve fitted by `loess` defaults in `R`.

The biasing factor model follows

$$\eta_i = (m_\zeta(1))^{-1} \exp\left\{\zeta_i\right\}, \quad \text{where}$$
$$\zeta_i \overset{iid}{\sim} \mathcal{N}(0, .8^2) \text{ on interval } (-3.5, 1). \tag{5.15}$$

In (5.15), η_i does not depend on covariate X_i so it is a purely random noise factor. So GREG and HT should be consistent under w_i.

Case 4 Beaumont's estimator is not the best in this case, but it is the second best. Assume that X_i is a 2-dimensional vector where only the first component $X_i^{(1)}$ shows up in the conditional mean $E(Y_i|X_i)$. We use the same conditional mean

function of X_i as in **Case 3** but different $\sigma^2(X_i)$. Specifically, the outcome model follows

$$\begin{aligned} Y_i \;&=\; 200 \;-\; 22X_i^{(1)} \;+\; 2X_i^{(1)\,2} \;-\; .05X_i^{(1)\,3} \;+\; \varepsilon_i, \quad \text{where} \\ X_i^{(1)} &\overset{iid}{\sim} Unif(5, 25),\; X_2^{(2)}\, iid \sim Unif(.5, 1.5), \\ \varepsilon_i &\overset{iid}{\sim} \mathcal{N}\left(0,\; \frac{\left[X_i^{(1)} X_i^{(2)}\right]^2}{16}\right). \end{aligned} \tag{5.16}$$

Then $\sigma^2(X_i) \;=\; \left[X_i^{(1)} X_i^{(2)}\right]^2 \Big/ 16$ and the optimal weights $u_i^* \;\propto\; 1 \Big/ \left[X_i^{(1)} X_i^{(2)}\right]$. The propensity model gives Beaumont's method some advantages since it imposes strong dependence between $E(Y_i|X_i)$ and size variable V_i,

$$\begin{aligned} V_i \;&=\; \frac{\delta_i}{200 \;-\; 22X_i^{(1)} \;+\; 2X_i^{(1)\,2} \;-\; .05X_i^{(1)\,3}} \quad \text{where} \\ \delta_i &\overset{iid}{\sim} \mathcal{LN}(-.1^2/2,\; .1^2). \end{aligned} \tag{5.17}$$

From Figure **Case 4**, we can see that both $\pi_i^0 - Y_i$ and $Y_i - d_i$ have strong dependence that can be easily modeled. For similar reason, PS estimator may not work very well.

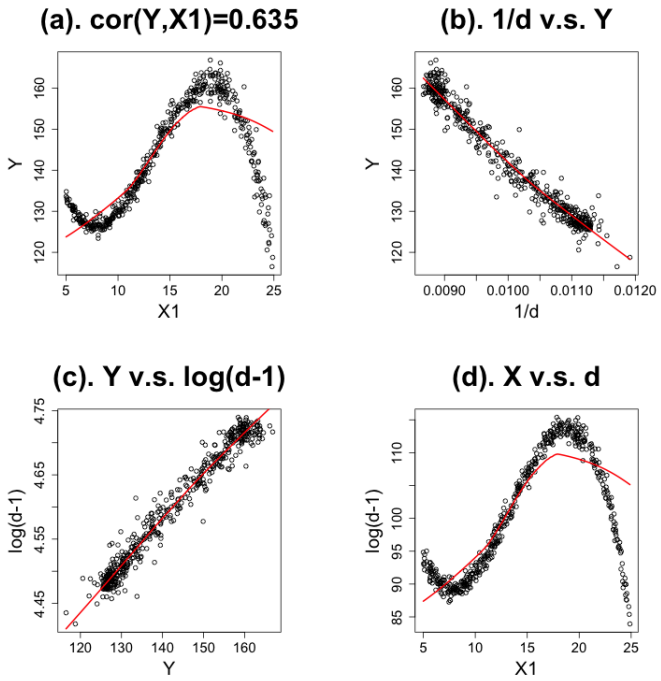

Figure 5.4: Scatter plots showing situations defined in **Case 4** favoring Beaumont's method over other considered estimators. The red solid line is local polynomial regression curve fitted by `loess` defaults in `R`.

The biasing factor model follows

$$\begin{aligned} \eta_i &= (m_\zeta(1))^{-1} \exp(\zeta_i), \quad \text{where} \\ \zeta_i \, iid &\sim Unif(-.54, .35). \end{aligned} \tag{5.18}$$

Again η_i is a purely random variable, independent of $(Y_i, X_i, \varepsilon, \delta_i)$. So that both HT and GREG are consistent under w_i.

In summary, **Case 1** to **Case 4** define four cases by defining the outcome, propensity and biasing factor models. Under each case, there is always at least one estimator performing well under d_i, i.e., the corresponding variance is relatively

small compared to other estimators. We then will use the simulation results in next section to illustrate that how these advantage of a particular estimator may be damaged by using w_i, i.e. the anticipated variance may be inflated so that that particular estimator may not be the best among all six considered estimators. The associated simulation results are summarized in next section.

5.4 Simulation results without weight misspecification

Before discussing the simulation results, it is very clear that most estimators across different cases should be consistent, resulting in nearly zero percent relative bias. When comparing the anticipated variances, we expect that GREG, PS, ZL2 and B perform the best (or second best) in **Case 1** to **Case 4** respectively. In this and the next section, the Sampford PPS sampling method previously discussed in Section 1.5 is used. Table 5.1 summarizes the comparative results under d_i. The first column specifies which estimator is summarized in that row. Columns 2-3 refers to **Case 1** defined in previous section, the rest columns are defined similarly. Again, RB stands for "relative bias" and RMSE means "relative root mean square error". All numbers are empirical average divided by population total of Y.

Table 5.1: Simulation results comparing anticipated variances without misspecification. Case 1-4 are as defined in **Case 1** to **Case 4**.

	Case 1		Case 2		Case 3		Case 4	
Method	RB(%)	RMSE(%)	RB(%)	RMSE(%)	RB(%)	RMSE(%)	RB(%)	RMSE(%)
HT	0.00	2.51	0.04	5.12	-0.08	0.74	-0.01	0.41
ZL1	0.00	2.51	0.04	5.12	-0.08	0.74	-0.01	0.41
ZL2	-0.00	2.56	0.03	3.30	-0.11	0.63	-0.45	2.16
B	-0.62	2.36	-0.12	5.06	-0.08	0.74	-0.00	0.41
PS	-0.01	1.33	-0.02	1.80	0.90	1.09	-0.26	0.44
GREG	0.00	1.06	0.02	2.38	0.02	0.76	0.00	0.34

All relative biases are close to zero, indicating that comparing anticipated variances is equivalent to comparing relative RMSE taken with respect to both the design and model. Within **Case 1**, we see that HT, ZL1, ZL2 and B have about the same level of RMSE while both PS and GREG have much smaller RMSE. Remember that in this case, the linear correlation between Y_i and X_i was relatively strong and according to Magee's results, GREG uses the optimal weights so that GREG should perform better than PS. Column 3 shows that the RMSE of PS is about 20% larger than the RMSE of GREG, indicating that GREG, indeed, has the smallest RMSE among six estimators. We expect PS to be the best in **Case 2**. Column 5 shows that HT, ZL1 and B are not efficient compared with PS and GREG. Also, the RMSE of PS is about 27.4% larger than the RMSE of GREG. Simulation validates that PS has the smallest variance in this case. Same stories apply to **Case 3** and **Case 4**. We see that in column 7, ZL2 has much smaller RMSE than the rest and in column 9, B is not the best but roughly the second best.

5.5 Simulation results with weight misspecification

Table 5.2 summarizes the comparing results under w_i. The display is exactly the same as the previous table except that w_i are used. All estimators except B in **Case 2** are consistent. The only observed inconsistency in B in this case may be due to model assumption failure, compounded by weight misspecification.

In column 3, we see that with all RMSEs increase, GREG is still the best among all with the smallest RMSE. In **Case 1**, we did not find a misspecification situation that makes GREG perform worse than others. In column 5, we see that PS and GREG have about the same RMSEs, outperforming the rest. Previously when there was no misspecification, PS was significantly better than GREG. In column 7, ZL2 is not the best anymore with increased RMSE. With weight misspecification, GREG performs better than ZL2. In column 9, B is not the second best with weight misspecification.

Table 5.2: Simulation results comparing anticipated variances with misspecification. Case 1-4 are as defined in **Case 1** to **Case 4**.

	Case 1		Case 2		Case 3		Case 4	
Method	RB(%)	RMSE(%)	RB(%)	RMSE(%)	RB(%)	RMSE(%)	RB(%)	RMSE(%)
HT	-0.04	7.55	0.11	6.05	-0.27	7.80	-0.05	4.66
ZL1	-0.04	7.55	0.11	6.05	-0.27	7.80	-0.05	4.66
ZL2	0.00	2.26	0.03	2.95	-1.58	1.68	-0.67	0.76
B	1.19	8.56	3.14	9.03	-0.27	7.80	-0.05	4.66
PS	-0.01	2.01	0.04	2.61	0.94	1.18	-0.26	0.44
GREG	0.00	1.33	0.05	2.67	0.07	0.88	0.01	0.35

5.6 Discussion

In this chapter, we examined anticipated variance and compared all six estimators by simulation. We first gave four cases defined by outcome, propensity, and biasing factor models. In each case, exactly one estimator had the smallest or the second smallest variances under d_i. The biasing factor model was chosen to invalidate the major model assumption that the winner needs in order to perform the best. If under the given biasing factor model, the weight misspecifications do alter the rankings of estimators, then we should see the variance of that the winning estimator is not necessarily the smallest under w_i. We gave one weight misspecification scenario for each case. From the simulation we found that PS and ZL2 were strongly affected by weight misspecification while we did not observe such big changes in GREG and B. This does not mean GREG and B are robust against weight misspecification though.

Chapter 6. ACS Simulation

In Chapter 2 we expressed the weight-misspecification biasing factor as the ratio of final modified weights to initial or design weights. Two probabilistic models, (2.3) and (2.4), were also presented to model the biasing factor on covariates X. In this chapter, we will discuss whether these two proposed models are useful in an illustrative example for describing the weight modification procedures of nonresponse and miscalibration, i.e. we would like to assess the quality of those proposed models in describing biasing factor η using covariates X. With this purpose, we take a realistic dataset as a data frame and draw samples from it. After a few steps of weight modifications, we then compute the η_i by taking the modified weights divided by initial weights.

The rest of this chapter is organized as follows: Section 6.1 describes the extraction of a real data set which will be treated as the finite frame population from which a sample is drawn; Section 6.2 describes the multiple-stage sampling procedure, the weight modification procedure that we consider, and some thoughts on the model fitting.

6.1 Description of underlying ACS data

The American Community Survey (ACS) is the largest ongoing household survey that the Census Bureau administers as the key source of information about American population and housing characteristics. The ACS is weighted to account for selection and housing unit nonresponse and hence all missing item responses have been removed (Ramirez and Ennis 2010). The ACS microdata from the year of 2000 to the present is available at `IPUMS USA` (Ruggles et al. 2007), which preserves and harmonizes census microdata and provides easy access to this data. In this section, 2016 ACS data was extracted from `IPUMS USA`*, which is a 1-in-100 national random sample of the population, of size 3,156,487. We use variable `STATEICP` to further refine the data set into a sample of 2016 ACS from Maryland only. The number of observations decreases to 59,408. The following variables are considered, most of which are commonly seen and used demographic and geographic variables. Recoding is also done for the main purpose of simplifying categories. The detailed information for variable importing and recoding is as follows:

- Household level
  - `County`, county code where the household was enumerated;
  - `Ownership of dwelling`, categorized into "N/A", "Owned or being bought (loan)" and "Rented".
  - `House acreage`, categorized into "N/A", "less than 10 acres", "10 acres

*https://usa.ipums.org/usa/sampdesc.shtml, section of ACS 2016 sample, accessed on Jan 30, 2018

or more";

- Person level
  - `Person weight`, person weight denoted by w_i^0 for the unit i in our extracted 2016 ACS Maryland sample;
  - `Race`, coded as 1 if white, 2 if black/African American/Negro, 3 if American Indian or Alaska Native, 4 if Chinese, 5 if Japanese, 6 if other Asian or Pacific Islander, 7 if other race, 8 if two major races and 9 if three or more major races. *Hispanic origin is assessed in a separate question, see below. To simplify, variable Race and Hispanic are recoded into a single variable.*
  - `Hispanic origin`, coded as 0 if not Hispanic, 1 if Mexican, 2 if Puerto Rican, 3 if Cuban, 4 if other. We combined and recoded `Race` and `Hispanic origin`. As long as the individual self-identify as being of Hispanic origin, the new race is recoded into "Hispanic". Among the individuals who did not self-identified as Hispanic, `Race` 4 to 6 are combined into "Asian or Pacific Islander", `Race` $3, 7$ and 9 are combined into "Others" due to very low frequencies.
  - `Sex`, coded as 1 if male, 2 if female;
  - `Age`, integers representing the age in years. Usually the age interval would be categorized into ≤ 17, $18-24$, $25-44$, $45-54$, $55-64$ and ≥ 65; in this chapter, the age intervals are categorized into ≤ 24, $25-54$ and ≥ 55 for simplicity.

– `Marital status`, categorized as 1 if "Married, spouse present", 2 if "Married, spouse absent", 3 if "Separated", 4 if "Divorced", 5 if "Widowed" and 6 if "Never married/single". We recode this variable and combine 1 and 2 into "Married", 3 to 5 into "Separated/Divorced/Widowed" and leave "Never married/single" as is;

– `Employment status`, categorized into "N/A", "Employed", "Unemployed" and "Not in labor force". Categories except for "Employed" are combined into "Others".

– `Number of own children in the household`, assume that this is the outcome variable of interest, denoted by Y_i for unit i.

Let us denote $\mathcal{U}$ as the universe of all Maryland residents with size N. Assume that we are interested in estimating the average number of own children in the household in Maryland state, i.e.,

$$\sum_{i \in \mathcal{U}} Y_i / N.$$

6.2 Data example based on ACS

6.2.1 Overview of the data example

In the present data example, we will imitate sampling and weight adjustment procedures restricted to data from the state of Maryland. Then we will assess the biasing factor model quality as defined in Chapter 2.

As mentioned above, all residents in Maryland form the finite frame population $\mathcal{U}$. The data set with number of rows $59,408$ that were extracted from `IPUMS-USA` form an initial sample of $\mathcal{U}$, denoted by $\mathcal{U}^*$. We consider ACS to be a preliminary sampling stage and then design a PPSWOR/PPSWOR/SRSWOR sampling procedure and further draw random sample $\mathcal{S}$ from $\mathcal{U}^*$ with sub-selection probability $\pi(i)$ for $i \in \mathcal{U}^*$. Counties are considered as primary subsampling units (PSU); groups clustered from person weights within county are considered as secondary subsampling units (SSU). The overall weights of sampled units in $\mathcal{S}$ are calculated as

$$w_i^1 = w_i^0/\pi_i, \tag{6.1}$$

where the form of π_i will be introduced below in Section 6.2.2 and w_i^0 are person weights available from ACS. Therefore all w_i^0, π_i and w_i^1 are known for all $i \in \mathcal{U}^*$. The person weights, w_i^0, have been used in three ways in this data example. First w_i^0 reflects the sampling procedure of ACS; secondly, w_i^0 have been used as stratification variable since person weights also reflect geographical information; thirdly, later w_i^0 are used in defining a measure of size (MOS) for further sampling stages.

Taking w_i^1 as input weights, we then perform raking and linear calibration for several rounds and the final weights after the last round are denoted by w_i^F. So the biasing factor would be

$$\eta_i = w_i^F/w_i^1, \quad i \in \mathcal{S}\,.$$

We would like to assess the quality of (2.4), as a model for η_i in the previous line. Specifically, we would like to assess the following model

$$\eta_i = \Big(m_\zeta(a(X_i))\Big)^{-1} \cdot \exp\Big(a(X_i)\zeta_i\Big), \tag{6.2}$$

where $\zeta_i \overset{iid}{\sim} F_\zeta$. Here F_ζ is an unknown distribution and $m_\zeta(\cdot)$ is the moment generating function of the bounded random variable ζ.

Since all covariates considered here are categorical variables, let us use C_l to denote the cell l, where all units in the same cell have the same value a_l for the function $a(X)$ of the covariate vector. Then (6.2) reduces to

$$\begin{aligned} \eta_i &= \frac{\exp(a_l\zeta_i)}{m_l}, \quad i \in C_l \cap \mathcal{S}, \\ m_l &= m_\zeta(a_l), \quad l = 1, \dots L. \end{aligned} \tag{6.3}$$

where L represents the total number of cells considering all levels of covariates and m_l and a_l are just unknown parameters, with some intrinsic restrictions.

6.2.2 Subsampling scheme

Counties are treated as primary subsampling units (PSU). Let us use j to denote the index of PSUs, $j = 1, 2, \dots, J$. All units are grouped into K strata based on the person weight w_i^0, using quantiles of w_i^0 as cutoff points. Let us use k denote the index of secondary subsampling units (SSU). Let U_j be the universe of SSUs within PSU j and U_{jk} be the universe of units within SSU k, PSU j. Next,

define the following quantities,

$$\begin{aligned}
\hat{N}_{jk} &= \sum_{i \in U_{jk}} w_i^0, \\
\hat{N}_j &= \sum_{k \in U_j} \hat{N}_{jk}, \\
\hat{N} &= \sum_{j=1}^{J} \hat{N}_j = \sum_{i \in \mathcal{U}^*} w_i^0,
\end{aligned}$$

where $\hat{N}_{ij}$, $\hat{N}_j$ and $\hat{N}$ actually estimate the sizes of U_{ij}, U_j and $\mathcal{U}$ respectively. As mentioned above, the adopted subsampling scheme is PPSWOR/PPSWOR/SRSWOR.

- First stage: m PSUs are selected. The selection probability of PSU j is defined as

  $$\pi(j) = m \cdot \frac{N_j}{N}.$$

  Selecting PSUs can be achieved using `cluster` function with method "systematic", in R pacakge `sampling`.

- Second stage: within each selected PSU, ν SSUs are selected. SSU k within PSU j has the (conditional) inclusion probability

  $$\pi(k|j) = \nu \cdot \frac{N_{jk}}{N_j}$$

  Selecting SSUs can be achieved using `strata` function with method "systematic".

- Third stage: within each selected SSU, q individuals are selected via simple

random sampling with replacement (SRSWOR). For each individual $i \in U_{jk}$, the inclusion probability is

$$\pi(i|j,k) \;=\; q/N_{jk}$$

Again this could be done via `strata` function with method "SRSWOR".

In the present simulation study, we take $K = 6$, $m = 10$, $\nu = 3$, $q = 200$. So in total the sample is is of the size $n = m \cdot \nu \cdot q = 6000$.

6.2.3 Weight adjustment procedures

Assume that the person weights, w_i^0, were already adjusted for nonresponse. Therefore, we only consider raking and linear calibration. Assume that we have three rounds of raking and calibration in the subsample. The outcome variable Y_i of interest is `number of own children in the household`, therefore one may want to consider adjusting for `ownership of dwelling`, `marital status` and `employment status` besides the commonly considered `sex`, `race`, `Hispanic origin` and `age`. The three rounds of weight adjustments are: 1) raking to marginal totals of `sex`, `race`; 2) linear calibration on `age`; and 3) raking to marginal totals of `ownership of dwelling`, `marital status` and `employment status`. The final weights are denoted by w_i^F for $i \in \mathcal{S}$. These could be done via the `calibrate` function with `calfun` = "linear" or "raking". All marginal totals we use in raking or calibration, are defined as internal HT estimated totals using design weights w_i^1 defined in (6.1).

Table 6.1: Summary of weight changes after three rounds of adjustments. Weight ratios are define as modified weights divided by design weights.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Design weights	42.86	583.96	862.40	1002.92	1308.34	4696.42
Weight ratios after round 1	0.90	0.93	0.99	0.99	1.04	1.11
Weight ratios after round 2	0.87	0.95	0.98	0.99	1.03	1.14
Weight ratios after round 3	0.76	0.94	0.98	0.99	1.03	1.20

Table 6.1 summarizes the distribution of design weights w_i^1 and weight changes after each further round of adjustment. Weight ratios after each round are defined as the modified weights after that round divided by the design weights w_i^1. The averages of weight ratios after each round are about 1 which satisfy our assumption **A.1** in Chapter 2 Section 2.3.

6.2.4 Preliminary model fitting for the biasing factor

Assume that we consider `race`, `age` and `sex` in model (6.3), then the total number of "cells" is $L = 5 \cdot 2 \cdot 3 = 30$. The numbers of observations are very different across cells. The minimum number is 8, and the maximum number is 712. Table 6.2 summarizes the distribution of cell counts. Imbalanced cell counts indicate imbalanced contributions from different cells when fitting (6.3).

Table 6.2: Distribution of cell counts. The cells are defined according to the values of `age`, `sex` and `race`.

	Cell Counts			
	≤ 20	21-100	101-200	≥ 200
Frequency	4	13	3	10

6.2.4.1 Assuming location-scale family

In (6.3), we have not assumed the distribution of ζ. If given a known and easy-to-work-with distribution such as normal distribution, we could easily establish the relationship between m_l and a_l. But ζ_i are not necessarily normally distributed. Instead of assuming that ζ_i follows a specific distribution, let us consider F_ζ to belong to a *location-scale family*. Then the model (6.3) is equivalent to

$$\frac{\log(\eta_i) + \log(m_l)}{a_l} = \zeta_i, \tag{6.4}$$

where $m_l = m_\zeta(a_l)$, $l = 1, \ldots, L$. We could view $-\log(m_l)$ as cell mean and a_l as cell standard deviation. Although m_l and a_l have a definite relationship under (6.3), we may still estimate $-\log(m_l)$ by the sample cell mean

$$-\log(\tilde{m}_l), = \frac{1}{|C_l \cap \mathcal{S}|} \sum_{i \in C_l \cap \mathcal{S}} \log(\eta_i) \tag{6.5}$$

and a_l by the sample cell standard deviation

$$s_l = \sqrt{\frac{1}{|C_l \cap \mathcal{S}| - 1} \sum_{i \in C_l \cap \mathcal{S}} (\log(\eta_i) + \log(\tilde{m}_l))^2}. \tag{6.6}$$

The cell residuals $r_i \overset{iid}{\sim} (0, 1)$ are given by

$$r_i = \frac{\log(\eta_i) + \log(\tilde{m}_l)}{s_l}, \; i \in C_l \cap \mathcal{S}. \tag{6.7}$$

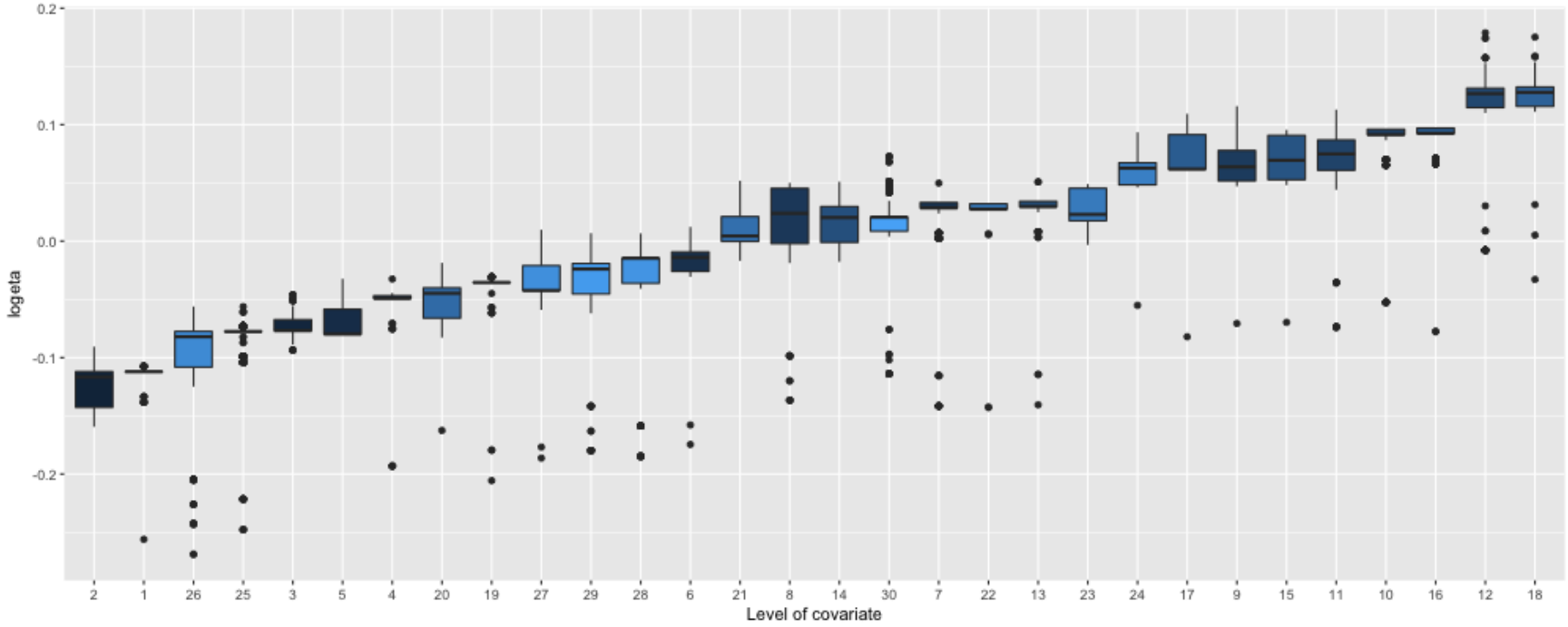

Figure 6.1: Box plot of cell residuals by level of covariates, sorted by increasing order of cell means $\tilde{m}_l$.

Figure 6.1 displays the boxplots of r_i within each cell C_l defined by covariates, sorted by increasing order of cell means $\log(\tilde{m}_l)$ in (6.5). From the figure we see that the ranges within cells differ from each other.

Figure 6.2 shows the histogram, with estimated density function and normal quantile-quantile (QQ) plot, of cell residuals r_i defined above. We can see that the residuals r_i thin tails and are skewed.

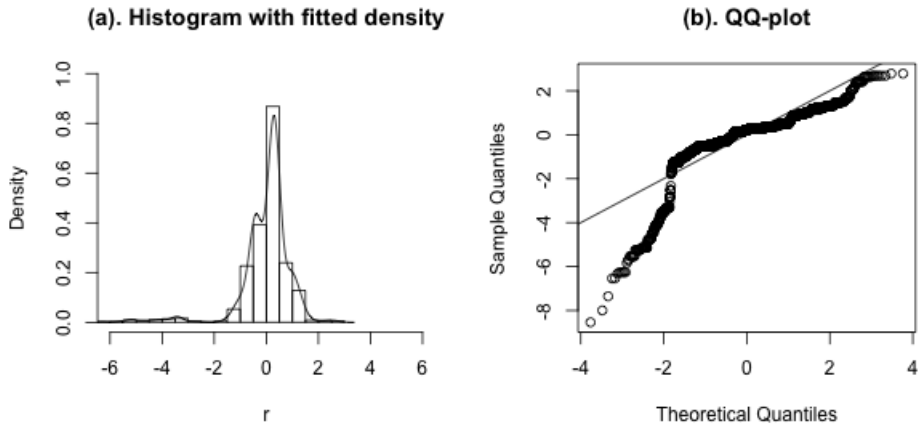

Figure 6.2: Histogram with fitted density and normal quantile-quantile plot of cell residuals. Gaussian kernel and bandwith .2 were chosen by visual inspection. Density is estimated at 2^{10} equally spaced points.

If ζ_i have expectation μ_ζ and variance σ_ζ^2, we may assume $\mu_\zeta = 0$ without the loss of generality since

$$\frac{\exp(a(\zeta' + \mu_\zeta))}{m_{\zeta'+\mu_\zeta}(a)} = \frac{\exp(a\mu_\zeta)\exp(a\zeta')}{\exp(a\mu_\zeta)m_{\zeta'}(a)} = \frac{\exp(a\zeta')}{m_{\zeta'}(a)},$$

where $\zeta' \sim (0, \sigma_\zeta^2)$. From (6.4) and the fact that $r_i \overset{iid}{\sim} (0, 1)$, we know that

$$\zeta_i \sim \sigma_\zeta r_i, \qquad \forall\, i \in \mathcal{S}\,. \tag{6.8}$$

Let us further assume that $a = (a_1, \ldots, a_L)$ satisfies,

$$\|a\|^2/L \;=\; \left(a_1^2 + \cdots + a_L^2\right)^2/L \;=\; 1,$$

then by (6.8) we know that

$$s_l^2 \;=\; a_l^2\sigma_\zeta^2, \quad l = 1, \ldots, L.$$

Therefore we have the following

$$a_l \;=\; s_l/\sigma_\zeta, \tag{6.9}$$

$$\sigma_\zeta^2 \;=\; \|s\|/L, \tag{6.10}$$

$$\zeta_i \;\sim\; \sigma_\zeta r_i \;=\; r_i\sqrt{\|s\|^2/L}, \tag{6.11}$$

$$m_l \;=\; E(\exp(a_l\zeta)) \;=\; E(\exp(s_l r_i)). \tag{6.12}$$

Let $\hat{f}(t)$ be the estimated density function of r_i. Then m_l can be estimated by a Trapezoidal rule approximation to $\int \exp(s_l t)\hat{f}(t)dt$,

$$\hat{m}_l \;=\; \sum_{j=1}^{2^{10}-1} \left(\exp(s_l t_j)\hat{f}(t_j) \;+\; \exp(s_l t_{j+1})\hat{f}(t_{j+1})\right)\frac{t_{j+1}-t_j}{2}, \quad l \;=\; 1, \ldots, L. \tag{6.13}$$

Fig 6.3 shows the fit of the density curve to the cell residuals r_i of the ACS data. The X-axis is the cell standard error, s_l, defined in (6.6) and the Y-axis is the cell mean, $\log(\tilde{m}_l)$,defined in (6.5). The dots show the pairs $(s_l, \log(\tilde{m}_l))$. The curve shows $(s_l, -log(\hat{m}_l))$ where $\hat{m}_l$ is estimated as (6.13). If the fitting is good, the

points should be close to the curve. We see that all the points are scattered around the curve, indicating a less appealing fit of the model (6.3) to the ACS data.

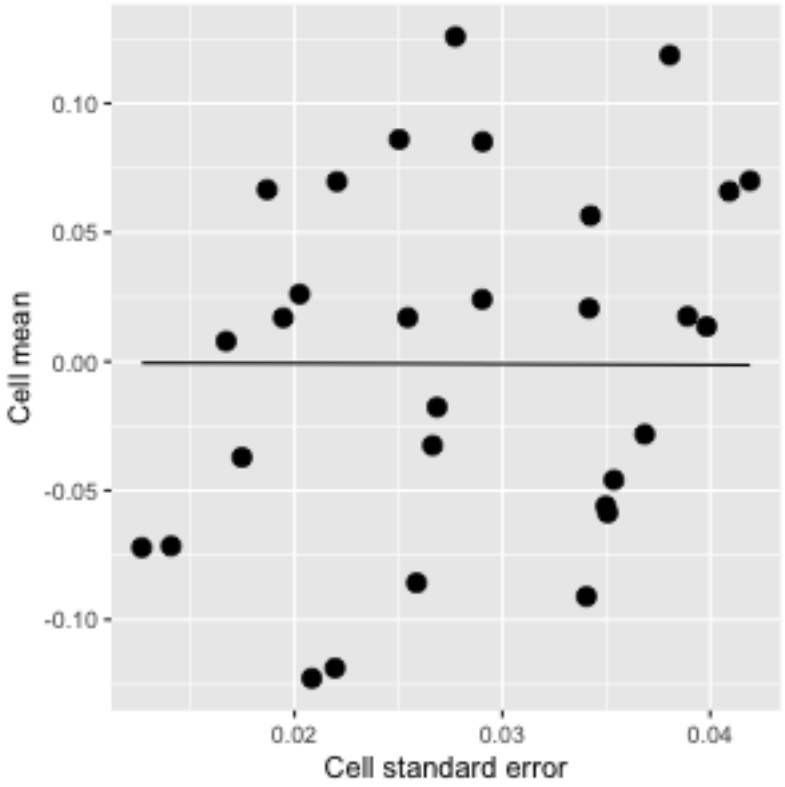

Figure 6.3: Comparing observed $(s_l, \log(\tilde{m}_l))$ with fitted curve

6.2.4.2 Model fit without covariate

A possible reason for the bad fit seen above might be the weak relationship between η_i and the selected covariates `age`, `sex` and `race`. If the covariates used in the weight adjustment procedures are ignored, and we consider instead

$$\eta_i = \frac{\exp(\lambda Z_i)}{m_Z(\lambda)} = \frac{\exp(\lambda Z_i)}{\exp(\lambda^2/2)} \quad \text{where } Z_i \sim \mathcal{N}(0,1),\ \lambda > 0, \tag{6.14}$$

followed by

$$\log(\eta_i) \overset{iid}{\sim} \mathcal{N}\big(-\lambda^2/2, \lambda^2\big).$$

Then we can estimate λ by maximizing the log-likelihood function with respect to λ,

$$-n\log(\lambda) - \sum_{i=1}^{n} \frac{(\log(\eta_i) + \lambda^2/2)^2}{2\lambda^2}. \tag{6.15}$$

Define

$$\hat{\zeta}_i = \frac{\log(\eta_i) + \hat{\lambda}^2/2}{\hat{\lambda}}. \tag{6.16}$$

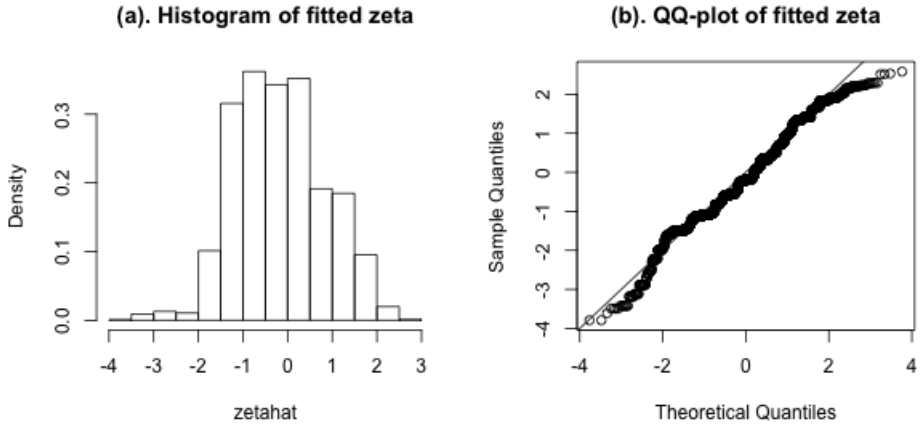

Figure 6.4: Histogram and normal quantile-quantile plot of $\hat{\zeta}_i$ in (6.16) checking normal assumption when fitting ignoring covariates. The line in (b) is the 45 degree line.

Figure 6.4 displays the histogram and QQ-plot of $\hat{\zeta}_i$ defined in (6.16). From the histogram we see that the distribution of $\hat{\zeta}_i$ may be skewed. But from the

normal QQ plot, we see that normal assumption may be acceptable when ignoring all covariates.

6.3 Discussion

The main purpose of this chapter was to check if the biasing factor models that we have proposed in Chapter 2 are good to explain the possible weight adjustment procedures in ACS data. A sample of ACS data was obtained and refined and treated as sampling frame. A PPSWOR/PPSWOR/SRSWOR sampling scheme was adopted to draw sample from the sampling frame. We considered raking and linear calibrations in three rounds. The biasing factors η_i were obtained at the end and model (2.4) was examined.

A preliminary analysis showed that a location-scale family assumption may not hold with covariates chosen in the current ACS sample. The unsatisfied fit showed in Figure 6.2 may be due to the model (6.3) reduced from (2.4), imbalanced cells, or inappropriately chosen set of covariates. We may explore other biasing factor models following the form

$$\eta = \frac{g(X, \zeta)}{E(g(X, \zeta))}.$$

To have balanced cells, we may design sampling procedure to achieve desired sampling rate within each cell (for example, in each age group by race). When ignoring covariates and fitting (6.3) under normal assumption, the normal QQ plot in Figure 6.4 showed a better fit to 45 degree line, indicating that the set of covariates we

have chosen (`race`, `age`, `sex`) may not be very good and another set of covariates may fit the data and explain the weight adjustment steps better.

Chapter 7. Contribution and Future Work

7.1 Contributions

- In Chapter 2, we proposed the idea of biasing factors which can be interpreted as the effect of multiplicative weight adjustments in survey data analysis. Two classes of probabilistic models for these biasing factors were proposed. The necessary and sufficient superpopulation condition for the Horvitz-Thompson (HT) estimator to be consistent under misspecified weights was given. To the best of our knowledge, the present study is the first research work modeling the distributional change from design to final weights and investigating the consequences on bias and variance introduced by such biasing factors.

- In Chapter 3, we examined the bias in the Generalized Regression Estimator (GREG) under the condition that HT is still consistent. We showed that when the conditional mean of the outcome variable is correctly specified, GREG is still consistent under misspecified weights. We then explored the bias in GREG when both the conditional mean and weights are misspecified, in the limiting sense. The formula implies that when there is bias in GREG with misspecified weights, it must be true that some of the covariates are miscalibrated. Sim-

ulation studies were done showing how misspecified weights could invalidate the consistency of GREG estimators of survey totals. Especially, data users might be misled by their estimates, since the 95% CI coverage rates for t_Y under misspecified weights were all much smaller than the nominal probability, 95%. But we did find out in simulations that when both the number of covariates and the sample size were moderate to large, it might be easy for the investigators to detect the miscalibrations among X totals. However, this finding was limited to the distributional assumptions that we have made in the simulations.

- In Chapter 4, we examined the bias in two estimators of Zheng and Little under the condition that HT is still consistent. Zheng and Little proposed two estimators, ZL1 and ZL2. We showed that ZL1 is either exactly the same as HT, or very close to HT, under certain conditions. Therefore we mainly focused on investigating the bias in ZL2 under misspecified weights. We then linked theoretically the non-zero estimating equation values outside the sample to the bias in model-based estimator of Zheng and Little. Simulation results showed that large biases in ZL2 under misspecified weights were always associated with non-zero estimating equations outside sample, indicating that misspecified weights might have changed the pattern of dependence between outcome variable and inclusion probability, so that the model fit well within the sample but not outside the sample.

- In Chapter 5, we studied anticipated variance under misspecified weights. It

is not surprising that a biasing factor would introduce extra noise so that the variances of estimators would be increased, in general. Our focus was on how the misspecification of weights would affect the relative performance of different estimators. The examples we found showed that the best estimator under design weights may not perform well under misspecified weights.

- In Chapter 6, a real data example was given, trying to assess the biasing factor model that we have proposed in Chapter 2. The biasing factors based on real ACS data sample (as true population) were obtained after we extracted the ACS data, did further PPSWOR/PPSWOR/SRSWOR sampling and performed three rounds of weight adjustments. Assuming a location-scale family of distributions for the log biasing factors, we examined the cell residuals and found out that the distribution might be skewed and thin-tailed, indicating that the normal assumption was not suitable in this data example. One should try other distributions, or evaluate the moment generating function values numerically, which is left for future work.

7.2 Summary

- Proposition 3 provides a necessary condition for bias in GREG under w_i. It implies that if inaccurate population totals have been used in calibrations, using the resulting set of weights may lead to bias. This suggests that at the weight adjustment stage, one should be very careful about the imported population total of X. It is possible that the source of information is out-of-date

and hence does not reflect the truth any more. Or people may import large-scale accurate survey information when calibrating on a smaller-scale survey. Therefore when projecting the accurate X-totals onto smaller-scale survey, we may produce error if the effective target population in the smaller survey may be different from that of the known total. At the analysis stage when all weight adjustments have been done, data users may perform statistical tests to check if the X-totals have been well calibrated. Simulation studies show that it may be possible to detect miscalibrations on X-totals in a large survey, i.e., when sample size n is large. It might be difficult to do so in a small survey. If data users identify some miscalibrations on X-totals, one may continue to use GREG with those miscalibrated covariates dropped from the working model, or proceed with other estimators.

- The usage of model-based methods like Zheng and Little's or Beaumont's depend on the model assumptions heavily. Our investigations show that when the fitted model do not fit the data outside of sample well, the second estimator of Zheng and Little may have bias. The extrapolation idea requires strong model assumptions which we are not able to observe and examine. When weights are inappropriately adjusted, it is possible that the dependence between weights and other variables have been affected and therefore the model assumptions may not hold. At the analysis stage, it is highly recommended that data users check model assumptions carefully within the sample. Methods utilizing the extrapolation idea might be dangerous since we are not able

to examine the data that we do not observe.

7.3 Future work

- In any of the simulation work that we have done here, more classes of outcome and propensity models should be tried. For example as indicated on the previous page, the ability of simple hypothesis tests to detect the miscalibrations in X totals in simulations was limited to the distributional assumptions that we have made. It may still be true that under certain conditions, it is still hard for investigators to detect such miscalibrations, which may lead to bias in GREG.

- In examples like the ACS real data example in Chapter 6, a more systematic study of models of the biasing factors could be attempted. First we have found that with the covariates (`race`, `age`, `sex`) chosen, the normal distributional assumption does not work well for ACS data. The location-scale family assumption did not fit the ACS data very well either. As discussed in Section 6.3, the bad fit might be due to the model (6.3) reduced from (2.4), imbalanced cells, or inappropriately chosen set of covariates. We may explore other biasing factor models following the form

$$\eta = \frac{g(X, \zeta)}{E(g(X, \zeta))}.$$

We may explore other biasing factor models following the form

$$\eta = \frac{g(X, \zeta)}{E(g(X, \zeta))}.$$

To have balanced cells, we may design sampling procedure to achieve desired sampling rate within each cell (for example, in each age group by race). When ignoring covariates and fitting (6.3) under normal assumption, the normal QQ plot in Figure 6.4 showed a better fit to 45 degree line, indicating that the set of covariates we have chosen (`race`, `age`, `sex`) may not be very good and another set of covariates may fit the data and explain the weight adjustment steps better.

References

Beaumont, J. F. (2008), "A new approach to weighting and inference in sample surveys," *Biometrika*, 95, 539–553.

Beaumont, J.-F. and Rivest, L.-P. (2007), "A Weight Smoothing Method for Dealing with Stratum Jumpers in Business Surveys," *A A*, 9, 1.

Bernstein, D. S. (2005), *Matrix Mathematics: Theory, Facts, and Formulas with Application to Linear Systems Theory*, Princeton University Press Princeton.

Carroll, J. and Hartley, H. (1964), "The symmetric method of unequal probability sampling without replacement," *Biometrics*, 20, 908–909.

De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C., and De Boor, C. (1978), *A Practical Guide to Splines*, vol. 27, Springer-Verlag New York.

De Branges, L. (1959), "The Stone-Weierstrass theorem," *Proceedings of the American Mathematical Society*, 10, 822–824.

Deville, J.-C. and Särndal, C.-E. (1992), "Calibration Estimators in Survey Sampling," *Journal of the American Statistical Association*, 87, 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, 1013–1020.

Estevao, V. M. and Särndal, C.-E. (2000), "A functional form approach to calibration," *Journal of Official Statistics*, 16, 379.

Fuller, W. A. (2002), "Regression estimation for survey samples," *Survey Methodology*, 28, 5–23.

— (2011), *Sampling Statistics*, John Wiley & Sons.

Gentle, J. E. (2009), *Computational Statistics*, Springer Publishing Company, Incorporated, 1st ed.

Graubard, B. I. and Korn, E. L. (2002), "Inference for superpopulation parameters using sample surveys," *Statistical Science*, 73–96.

Hájek, J. (1964), "Asymptotic theory of rejective sampling with varying probabilities from a finite population," *The Annals of Mathematical Statistics*, 1491–1523.

Isaki, C. T. and Fuller, W. A. (1982), "Survey design under the regression superpopulation model," *Journal of the American Statistical Association*, 77, 89–96.

Korn, E. L. and Graubard, B. I. (1998), "Variance estimation for superpopulation parameters," *Statistica Sinica*, 1131–1151.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R. M., and 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 (Statistics in Society)*, 173, 389–407.

Little, R. J. (1986), "Survey nonresponse adjustments for estimates of means," *International Statistical Review/Revue Internationale de Statistique*, 139–157.

Lohr, S. (2009), *Sampling: Design and Analysis*, Nelson Education.

Madow, W. G. (1949), "On the theory of systematic sampling, II," *The Annals of Mathematical Statistics*, 333–354.

Magee, L. (1998), "Improving survey-weighted least squares regression," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60, 115–126.

Oh, H. L. and Scheuren, F. J. (1983), "Weighting adjustment for unit nonresponse," *Incomplete Data in Sample Surveys*, 2, 143–184.

Pfeffermann, D. and Rao, C. (2009), *Handbook of Statistics Vol. 29A Sample Surveys: Design, Methods and Applications*, North-Holland.

Pfeffermann, D. and Sverchkov, M. (1999), "Parametric and semi-parametric estimation of regression models fitted to survey data," *Sankhyā: The Indian Journal of Statistics, Series B*, 61, 166–186.

Ramirez, R. R. and Ennis, S. R. (2010), *Item nonresponse, allocation, and data editing of the question on Hispanic origin in the American Community Survey (ACS): 2000 to 2007*, US Census Bureau.

Rao, J. (1963), "On three procedures of unequal probability sampling without replacement," *Journal of the American Statistical Association*, 58, 202–215.

Rosenbaum, P. R. and Rubin, D. B. (1983), "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 41–55.

Ruggles, S., Genadek, K., Goeken, R., Grover, J., and Sobek, M. (2007), "Integrated Public Use Microdata Series: Version 7.0," University of Minnesota. https://doi.org/10.18128/D010.V7.0.

Ruppert, D. (2002), "Selecting the Number of Knots for Penalized Splines," *Journal of Computational & Graphical Statistics*, 11, 735–757.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge University Press.

Sampford, M. (1967), "On sampling without replacement with unequal probabilities of selection," *Biometrika*, 54, 499–513.

Särndal, C., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer-Verlag.

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.

Schoenberg, I. J. (1964), "Spline functions and the problem of graduation," *Proceedings of the National Academy of Sciences*, 52, 947–950.

— (1988), "Contributions to the problem of approximation of equidistant data by analytic functions," in *IJ Schoenberg Selected Papers*, Springer, pp. 3–57.

Tillé, Y. and Matei, A. (2016), *Sampling: Survey Sampling*, r package version 2.8.

Valliant, R., Dever, J. A., and Kreuter, F. (2013), *Practical Tools for Designing and Weighting Survey Samples*, Springer.

Ybarra, L. M. and Lohr, S. L. (2008), "Small area estimation when auxiliary information is measured with error," *Biometrika*, 95, 919–931.

Zheng, H. and Little, R. J. (2003), "Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples," *Journal of Official Statistics*, 19, 99–117.