ABSTRACT

Title of Dissertation: WEIGHT ADJUSTMENT METHODS AND THEIR IMPACT ON SAMPLE-BASED INFERENCE

Kimberly Henry, Doctor of Philosophy, 2011

Dissertation directed by: Dr. Richard Valliant, Research Faculty Member
Joint Program in Survey Methodology,
University of Michigan

Weighting samples is important to reflect not only sample design decisions made at the planning stage, but also practical issues that arise during data collection and cleaning that necessitate weighting adjustments. Adjustments to base weights are used to account for these planned and unplanned eventualities. Often these adjustments lead to variations in the survey weights from the original selection weights (i.e., the weights based solely on the sample units' probabilities of selection). Large variation in survey weights can cause inferential problems for data users. A few extremely large weights in a sample dataset can produce unreasonably large estimates of national- and domain-level estimates and their variances in particular samples, even when the estimators are unbiased over many samples. Design-based and model-based methods have been developed to adjust such extreme weights; both approaches aim to trim weights such that the overall mean square error (MSE) is lowered by decreasing the variance more than increasing the square of the bias. Design-based methods tend to be ad hoc, while

Bayesian model-based methods account for population structure but can be computationally demanding. I present three research papers that expand the current weight trimming approaches under the goal of developing a broader framework that connects gaps and improves the existing alternatives. The first paper proposes more in-depth investigations of and extensions to a newly developed method called generalized design-based inference, where we condition on the realized sample and model the survey weight as a function of the response variables. This method has potential for reducing the MSE of a finite population total estimator in certain circumstances. However, there may be instances where the approach is inappropriate, so this paper includes an in-depth examination of the related theory. The second paper incorporates Bayesian prior assumptions into model-assisted penalized estimators to produce a more efficient yet robust calibration-type estimator. I also evaluate existing variance estimators for the proposed estimator. Comparisons to other estimators that are in the literature are also included. In the third paper, I develop summary- and unit-level diagnostic tools that measure the impact of variation of weights and of extreme individual weights on survey-based inference. I propose design effects to summarize the impact of variable weights produced under calibration weighting adjustments under single-stage and cluster sampling. A new diagnostic for identifying influential, individual points is also introduced in the third paper.

WEIGHT ADJUSTMENTS METHODS AND THEIR IMPACT ON SAMPLE-BASED INFERENCE

by

Kimberly A. Henry

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 in Philosophy
2011

Advisory Committee:

Dr. Richard Valliant, Chair
Dr. Michael Elliott
Dr. Partha Lahiri
Dr. Jill Montaquila
Dr. Paul Smith

Acknowledgements

I would first like to express sincere appreciation and gratitude to my thesis advisor, Dr. Richard Valliant. This research would not have been possible without your guidance, expertise, and mentoring. I am eternally grateful for all that you have taught me and look forward to collaboration on future projects.

I would also like to acknowledge the efforts and contributions of my committee members, Dr. Michael Elliott, Dr. Jill Montaquila, Dr. Partha Lahiri, and Dr. Paul Smith. This includes their research collaboration efforts and thorough review of my thesis. I recognize their separate but equally important areas of expertise, each of which this dissertation research benefitted from their involvement.

Last, I appreciate the support of my family, friends, and coworkers. In particular, I would like to thank my husband, Eric Henry. I truly appreciate his patience, understanding, and uncanny ability to be completely ignored during the years I dedicated to pursuing this research.

Table of Contents

List of Tables

List of Figures

Introduction

Weighting samples is important to reflect not only sample design decisions made at the planning stage and use of auxiliary data to improve the efficiency of estimators, but also practical issues that arise during data collection and cleaning that necessitate weighting adjustments. Planned and unplanned adjustments to survey weights are used to account for these practical issues. Often these adjustments lead to variations in the survey weights that may be inefficient when making finite population estimates.

The standard, theoretically-based methods for weighting are summarized by Kalton and Flores-Cervantes (2003). However, when large variation in survey weights exists, it can potentially produce unreasonable point estimates of population means and totals and decrease the precision of these point estimates. Survey practitioners commonly face this problem when producing weights for analysis datasets (Kish 1990; Liu *et al*. 2004; Chowdhury *et al*. 2007). Even when the estimators are unbiased, extreme weights can produce inefficient estimates. Trimming or truncating large weights can reduce unreasonably large estimated totals and substantially reduce the variability due to the weights. This reduces variance at the expense of introducing bias; if the variance reduction is larger than the squared bias increase, then the net result is an overall decrease in mean square error (MSE) of the estimate. The various existing trimming methods use either design-based or model-based approaches to meet this MSE-reduction goal.

Variation in survey weights can arise at the sample design, data collection, and post-data collection stages of sampling. First, intentional differential base weights, the inverse of the probability of selection, are created under different sampling designs. For example, multiple survey analysis objectives may lead to disproportionate sampling of population subgroups. Issues that occur during data collection can also impact probabilities of selection, e.g., in area probability samples, new construction developments with a large number of housing units that were not originally listed may be discovered. These are usually subsampled to reduce interviewer workloads, but this subsampling may create a subset of units with extremely large base weights. Another example is subsampling cases for nonresponse follow up; subsequent weighted analysis incorporates subsampling adjustments. Last, post-data collection adjustments to base survey weights are also commonly used to account for multistage sampling. Examples include subsampling persons within households (Liu *et al.* 2004), adjusting for nonresponse to the survey (Oh and Scheuren 1983), calibrating to external population totals to control for nonresponse and coverage error (Holt and Smith 1979; Särndal *et al*. 1992; Bethlehem 2002; Särndal and Lundström 2005), and combining information across multiple frames (such as telephone surveys collected from landline telephone and cell phone frames, e.g., Cochran 1967; Hartley 1962).

Often multiple adjustments are performed at each stage of sample design, selection, and data editing. For example, the Bureau of Transportation Statistics (2002) used the following steps to produce weights in their National Transportation Availability and Use Survey, a complex sample using in-person household interviews to assess people's access to public and private transportation in the U.S.:

- Household-level weights: base weights for stratified, multistage cluster sampling; unknown residential status adjustment; screener nonresponse adjustment; subsampling households for persons with and without disabilities; multiple telephone adjustment; poststratification.

- Person-level weights: the initial weight is the product of the household-level weight from above and a subsampling adjustment for persons within the household; an extended interview nonresponse adjustment; trimming for disability status; raking; and a non-telephone adjustment.

This example also illustrates that there can be a need for multiple weights based on the level of analysis (here household-level vs. person-level) that is desired from the sample data.

Next I provide some examples of various weighting adjustment methods that are performed in sample surveys. This is not an all-encompassing list; refer to the post-data collection adjustment references and Chapters 4, 8, 16, and 25 in Pfeffermann and Rao (2009) for more detail.

Example 0.1. *Cell-based Nonresponse Adjustments*. These adjustments involve categorizing the sample dataset into cells using covariates available for both respondents and nonrespondents that are believed to be highly correlated with response propensity and key survey variables. Assuming that nonresponse is constant within each cell ("missing at random;" age/race/sex are often used in household surveys), the reciprocal of the cell-based response rate is used to increase the weights of all units within the cell. Propensity models across all cells using the cell-based covariates can also be used to predict the response rate. Other nonresponse weighting adjustments are discussed in Brick and Montaquila (2009).

▪

Example 0.2. *Dual Frame Adjustments*. For surveys that use multiple frames, e.g., telephone and area probability samples or landline and cell phone surveys, additional weighting adjustments may be used to account for units that are contained in more than one frame. Often composite estimators, which are a weighted average of the separate frame estimates (Hartley 1962) are used, incorporating the known or estimated overlap of units on both frames. These can be explicitly expressed as adjustments to each sample unit's weight.

▪

Example 0.3. *Poststratification*. Here survey weights are adjusted such that they add up to external population counts by available domains. This widely-used approach allows us to correct the imbalance than can occur between the sample design and sample completion, i.e., if the sample respondent distribution within the external categories differs from the population (which can occur if, e.g., more women respond than men; historically young black males contribute to undercoverage), as well as reduce potential bias in the sample-based estimates. Denoting the poststrata by $d = 1, \ldots, D$, the poststratification estimator for a total involves adjusting the base-weighted domain totals ($\hat{T}_d$) by the ratio of known (N_d) to estimated ($\hat{N}_d$) domain sizes: $\hat{T}_{PS} = \sum_{d=1}^{D} \frac{N_d}{\hat{N}_d} \hat{T}_d$.

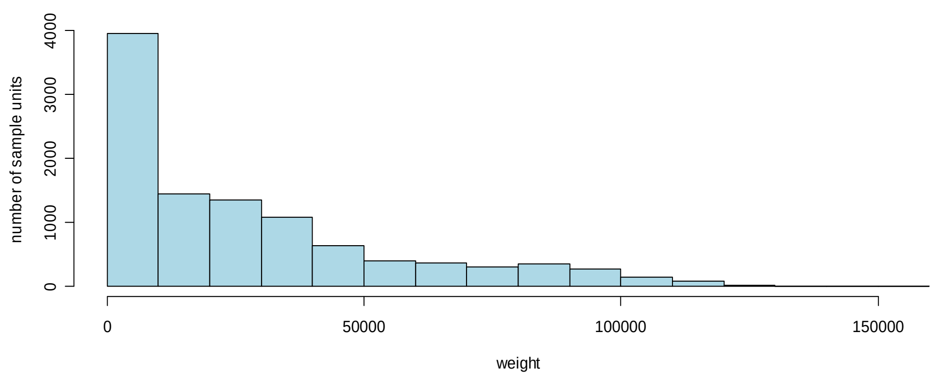
Example 0.4. *Calibration Adjustments*. Case weights resulting from calibration on benchmark auxiliary variables can be defined with a global regression model for the survey variables (Huang and Fuller 1978; Bardsley and Chambers 1984; Bethlehem and Keller 1987; Särndal *et al*. 1992; Sverchkov and Pfeffermann 2004; Beaumont 2008; Kott 2009). Deville and Särndal (1992) proposed a model-assisted calibration approach that involves minimizing a distance function between the base weights and final weights to obtain an optimal set of survey weights. Here "optimal" means that the final weights produce totals that match external population totals for the auxiliary variables $\mathbf{X}$ within a margin of error. Specifying alternative distance functions produces alternative estimators; a linear distance function produces the *general regression estimator (GREG)* $\hat{T}_{GREG} = \hat{T}_{HT} + \hat{\mathbf{B}}^T \left(\mathbf{T}_X - \hat{\mathbf{T}}_{XHT} \right) = \sum_{i \in s} \frac{g_i y_i}{\pi_i}$, where $\hat{\mathbf{T}}_{XHT} = \sum_{i \in s} w_i \mathbf{X}_i = \sum_{i \in s} \frac{\mathbf{X}_i}{\pi_i}$ is the vector of Horvitz-Thompson totals for the auxiliary variables, $\mathbf{T}_X = \sum_{i=1}^{N} \mathbf{X}_i$ is the corresponding vector of known totals, $\hat{\mathbf{B}}^T = \mathbf{A}_s^{-1} \mathbf{X}_s^T \mathbf{V}_{ss}^{-1} \mathbf{\Pi}_s^{-1} \mathbf{y}_s$, with $\mathbf{A}_s = \mathbf{X}_s^T \mathbf{V}_{ss}^{-1} \mathbf{\Pi}_s^{-1} \mathbf{X}_s$, $\mathbf{X}_s^T$ is the matrix of $\mathbf{X}_i$ values in the sample, $\mathbf{V}_{ss} = diag\left(v_i\right)$ is the diagonal of the variance matrix specified under the model, and $\mathbf{\Pi}_s = diag\left(\pi_i\right)$ is the diagonal matrix of the probabilities of selection for the sample units. In the second expression for the GREG estimator, $g_i = 1 + \left(\mathbf{T}_X - \hat{\mathbf{T}}_{XHT} \right)^T \mathbf{A}_s^{-1} \mathbf{X}_i v_i^{-1}$ is called the "g-weight."

The GREG estimator for a total is model-unbiased under the associated working model and is approximately design-unbiased when the sample size is large (Deville and Särndal 1992). When the model is correct, the GREG estimator achieves efficiency gains; if the model is incorrect, then the efficiency gains will be dampened (or nonexistent) but the approximate design-unbiased property still holds. One disadvantage to the GREG approach is that the resulting weights can be negative or less than one. Calibration can also introduce considerable variation in the survey weights. To overcome the first problem, extensions to limit the range of calibration weights have been developed that involve either using a bounded distance function (Rao and Singh 1999; Singh and Mohl 1996; Theberge 1999) or bounding the range of the weights using an optimization method (such as quadratic programming, Isaki *et al.* 1992). Chambers (1996) proposed penalized calibration optimization function to produce non-negative weights and methods that impose additional constraints on the calibration equations.

The effect of such weighting adjustments, when applied across the sample's design strata and PSUs, is that the variability of the weights in increased (Kish 1965, 1992; Kalton and Kasprzyk 1986). This problem, which the Bayesian weight trimming methods currently do not address, frequently increases the variances of the sample-based estimators, thereby decreasing precision. A few very large weights can also be created such that the product $w_i y_i$ creates an unusually large estimate of the population total.

To see why weight trimming may be needed, consider the following empirical example from the 2005-2006 National Health and Nutrition Examination (NHANES) public-use file dataset (NHANESa). The interview-level weights for 10,348 people have post-stratification adjustments to control totals estimated from the Current Population Survey (NHANESb). These weights range from 1,225 to 152,162 and are quite skewed:

Figure 0.1. Histogram of NHANES 2005-2006 Adult Interview Survey Weights



The weights in the tails of the distribution, such as the one shown in Figure 0.1, can lead to overly large totals with associated large variances, particularly when the weight is combined with a large survey response value. This problem can increase for domain-level estimates, particularly in establishment data, when variables of interest can also be highly skewed toward zero. On the other hand, large weights for some subgroups of units may be necessary to produce estimators that are, in some sense, unbiased.

Beaumont and Rivest (2008) estimate establishment sample units that were assigned large design weights based on incorrect small measures of size accounted for 20-30% of an estimated domain total. These outlying weights also drive variance estimates.

There is limited literature and theory on design-based weight trimming methods, most of which are not peer-reviewed publications and focus on issues specific to a single survey or estimator. The most cited work is by Potter (1988; 1990), who presents an overview of alternative procedures and applies them in simulations. Other studies involve a particular survey (e.g., Battaglia *et al*. 2004; Chowdhury *et al*. 2007; Griffin 1995; Liu *et al*. 2004; Pedlow *et al*. 2003; Reynolds and Curtin 2009). All design-based methods involve establishing an upper cutoff point for large weights, reducing weights larger than the cutoff to its value, then "redistributing" the weight above the cutoff to the non-trimmed cases. This ensures that the weights before and after trimming sum to the same totals (Kalton and Flores-Cervantes 2003). The methods vary by how the cutoff is chosen. There are three general approaches: (1) ad hoc methods that do not use the survey response variables or an explicit model for the weight to determine the cutoff (e.g., trimming weights that exceed five times the median weight to this value); (2) Cox and McGrath's (1981) method that uses the empirical MSE of a particular estimator and variable of interest; and (3) methods assuming that the right-tail of the weights follow some skewed parametric distribution, then use the cutoff associated with an arbitrarily small probability from the empirical distribution (Chowdhury *et al*. 2007; Potter 1988).

Alternatively, Bayesian methods that pool or group data together have been recently proposed for weight trimming. There are two complementary approaches: "weight pooling" and "weight smoothing." While both use models that appear similar, weight pooling is the Bayesian extension of design-based trimming and weight smoothing is the Bayesian extension of classical random effect smoothing. In weight pooling models, cases are grouped into strata, some of which are collapsed into groups, and the group weight replaces the original weights. In weight smoothing, a model that treats the group means as random effects smoothes the survey response values. In this approach, the influence of large weights on the estimated mean and its variance is reduced under the smoothing model. In both methods, Bayesian models are used to average the means across all possible trimming points, which are obtained by varying the smoothing cut point. Both methods can produce variable-dependent weights. In addition, these methods have been developed from a very theoretical framework, for specific inferences, and the model may be difficult to apply and validate in practice.

Other forms of inference produce indirect weighting adjustments using models. In particular, penalized (*p*-) spline estimators have been recently developed to produce more robust estimators of a total. Zheng and Little (2003, 2005) used *p*-spline estimators to improve estimates of totals from probability-proportional-to-size (pps) samples. Breidt *et al.* (2005) proposed and developed a model-assisted *p*-spline approach that produced a GREG) estimator that was more robust to misspecification of the linear model, resulting in minimum loss in efficiency compared to alternative GREGs. In the model-prediction approach (Valliant *et. al* 2000), models are incorporated to improve estimators of totals. The *p*-spline estimators are a specific case of a robust prediction estimator.

All weight trimming methods have the potential to “distort” (reduce) the amount of information contained in weights related to the units’ analytic importance to data users (reflected in the inverse of the probabilities of selection), and nonresponse/undercoverage assumptions. Some weighting adjustment methods, such as nonresponse or calibration adjustments, are designed to reduce the bias and/or variances. In some cases, variable weights can be more efficient and their beneficial bias/variance reductions could be needlessly removed through arbitrary trimming of large weights. Thus, there is a need for diagnostic measures of the impact of weight trimming on survey inference that extend past the existing "design effect" type of summary measures, most of which do not incorporate the survey variable of interest. The current methods do not quantify such "loss of information;" i.e., there is no indication of how various methods' distortion of the original weight distribution potentially impacts sample-based inference.

This proposal includes three separate but related papers that attempt to address some gaps in the area of weight trimming. In particular, all three papers aim to provide a more in-depth understanding of how different weighting adjustment and trimming methods impact survey-based inference. First, I extend a newly developed approach of weight trimming in the areas of variance estimation, model sensitivity to different kinds of survey variables, and explore some robust methods to estimate the model parameters. Second, I propose extending the model-assisted *p*-spline estimation approach to use the Bayesian approach (following Zheng and Little 2003; 2005) and incorporate prior distributions and data-based estimators for the model components. Third, I explore diagnostic measures to gauge the impact of alternative weighting adjustments on sample-based inference. Specifically, I propose design effect measures to gauge the impact of calibration weights under unit- and cluster-level sampling for different variables of interest. I also apply a regression-based diagnostic to flag units within a given sample that are more or less influential when their weights are trimmed or non-trimmed.

Paper 1: Extending the Generalized Design-based Weight Trimming Approach for Estimating a Population Total

Abstract: Here I propose three extensions to the generalized design-based inference approach to weight trimming: (1) develop an appropriate design-consistent and model-robust variance estimator; (2) illustrate some of the limitations of the current method; and (3) use nonparametric methods to estimate the model parameters. These methods are proposed to produce new generalized design-based estimators of totals with trimmed weights and their variances that also overcome weaknesses in the existing methods.

1.1. Introduction, Research Plan, and Research Hypotheses

1.1.1. Introduction

In this paper, I extend a new method called generalized design-based inference to develop estimators with trimmed weights. Under this approach, the survey response variables and weights are both treated as random variables. Based on the data from the one sample at hand, the weights are replaced with their expected values under a model. The model for predicting the weights is based on the survey response variables. Preliminary theory and empirical evaluations have demonstrated this method can produce trimmed weights that reduce the mean square error (MSE) of an estimated total (Beaumont 2008; Beaumont and Alavi 2004; Beaumont and Rivest 2009; Sverchkov and Pfeffermann 2004; 2009). However, the proposed variance estimators have also not been fully evaluated and the model has only been proposed for element sampling designs. The method's dependence on an underlying model makes the smoothed weights, the estimated totals, and their variances susceptible to influential values.

First, Beaumont (2008) provides a general method for estimating the variance of the smoothed weight estimators of totals and proves that one variance estimator is design-consistent for a simple model, but does not prove that the estimator always yields a positive variance estimate. While he suggests using the replication-based Rao-Wu (1988) bootstrap variance estimator and a design-based MSE estimator, empirically I show that these methods did not perform well against model-based variance estimators. I propose two variance estimators: a robust sandwich variance estimator and more appropriate variance estimator under the sample design and weights model. I evaluate the robust variance estimator for Beaumont's estimator under a general model against Beaumont's and the Rao-Wu bootstrap variance estimators.

Second, since the generalized design-based method involves using the survey response variables to smooth the weights, I focus on illustrating how this method performs using different types of survey response variables. I also consider a special form of survey response variable called zero-inflated variables. These are variables with many observations having a value of zero, as well as positive values that appear to follow some particular distribution. These variables typically are modeled using a two-part model that first predicts a zero/nonzero value (i.e., a binary response), then predicts the specific values for the observations predicted to have nonzero values.

Last, I address influential values since outliers in the survey response variables (the predictors), the weights (the dependent variable), or a combination of both can unduly influence the weights predicted from a generalized design-based model. I propose to use methods developed in the nonparametric (NP) regression literature to protect the predicted weights from model misspecification caused by such influential values. The methods also relax the linearity assumption underlying the Beaumont (2008) approach. I consider three specific NP methods: MM estimation, least median of squares, and least trimmed squares. Nonparametric literature demonstrates that these particular methods, most of which use estimation criteria based on different versions of squared residuals, produce model coefficient estimates that are more robust to outliers than parametric alternatives like least squares and more efficient than other NP methods. However, one method has not yet been identified as "best" (uniformly superior) among them in the NP literature, so I consider these alternatives. I also demonstrate empirically that the NP totals can outperform the Beaumont estimators when nonresponse adjustments are applied to base weights such that a few outlying large weights are produced.

The generalized design-based approach has a strong limitation in that the model specified on the weights must be appropriate. Since all weights are trimmed with the model, this method has potential for over-trimming the weights. I demonstrate empirically in simulations how sensitive this approach is to model misspecification, both in producing biased estimates of totals and poor estimates of their variances. In addition, I demonstrate the difficulty in identifying one model for different kinds of survey response variables.

1.1.2. Research Plan and Hypotheses

Theoretical properties of the generalized design-based estimator will be established with respect to both the sample design and the weights model; the current focus has been restricted to properties that hold only under the model. I demonstrate theoretically that my proposed variance estimator is more robust to misspecification of the variance component in the weights model. I also use a result from Valliant *et al*. (2000) that this estimator is positively biased when the weights model does not hold. However, in this circumstance, the variance estimator underestimates the MSE. To evaluate the proposed robust variance estimator under a linear model for the weights, I will prove that both are design-consistent under a general working model.

When fitting the generalized design-based models to zero-inflated variables, I expect models that ignore the "zero-inflation" aspect of these variables to produce biased results, since the zero values will attenuate the model coefficients and thus further over-smooth the weights. However, the proposed two-part weights model reduces this bias at the expense of increasing the variance of estimated totals.

To evaluate the proposed robust prediction methods, I first establish the theoretical properties of the NP smoothed estimators of finite population totals under generalized design-based models (by analogy of their properties established in the regression literature that prove the model parameter estimators are consistent and asymptotically unbiased). I also evaluate the estimators in a simulation study investigating the impact of model misspecification and varying weights under single-stage sample designs, and samples with simple nonresponse adjustments applied to the base weights.

Modeling the survey weights is a practical and simple method to implement under a wide variety of sample designs, variables of interest, and weighting adjustments. These nonparametric methods should produce generalized design-based estimators with lower MSE than the design-based alternatives when there are outliers in the weights and/or the survey response variables, when the underlying model is nonlinear, and when the correlation between the survey response variables and weight is high. The existing Beaumont model-based methods do not account for these conditions. There are design- and other model-based methods that attempt to account for these conditions by controlling on auxiliary variables, but they have not been fully compared and it may be possible to improve them.

Generally, I hypothesize that the proposed estimators will have lower efficiency (higher variance) than the parametric generalized design-based approaches when the model holds, but more robustness (i.e., have lower bias) when it does not hold or in data containing influential observations. This allows me to evaluate the performance of the proposed estimators against alternative design- and model-based methods that have been proposed in the related literature.

1.2. Literature Review

This section mixes summaries of existing methods of estimation and examples of alternative weight trimming approaches proposed in the related design-based, generalized design-based, and nonparametric regression literature, respectively.

1.2.1. Existing Design-based Approaches

The most common sample-based inference for a finite population total involves the Horvitz-Thompson (HT, 1952) estimator. This section briefly introduces the HT estimator and some examples of methods that trim the HT weights.

The Horvitz-Thompson Estimator

For s denoting a probability sample of size n drawn from a population of N units, the *Horvitz-Thompson estimator* (HT) for a finite population total of the variable of interest

y is

$$\hat{T}_{HT} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} w_i y_i \ . \tag{1.1}$$

Here the inverse of the probability of selection, $\pi_i = P(i \in s)$, is used as the weight, $w_i = \pi_i^{-1}$. It is established (e.g., Horvitz and Thompson 1952) that this estimator is unbiased for the finite population total in repeated πps sampling, but can be quite inefficient due to variation in the selection probabilities if π_i and y_i are not closely related. The design-based variance of (1.1) is

$$Var\left(\hat{T}_{HT}\right) = \sum_{i \in U} \sum_{j \in U} \left(\pi_{ij} - \pi_i \pi_j\right) \frac{y_i y_j}{\pi_i \pi_j} . \tag{1.2}$$

where π_{ij} is the joint selection probability of units i and j in the population set U. Influential observations in estimating a population total using (1.1) and the variance estimator associated with the variance in (1.2) arise simply due to the combination of probabilities of selection and survey variable values. Alternative sample designs, such as probability proportional to some available measure of size, introduce variable probabilities of selection in (1.1). The variability in selection probabilities can increase under complex multistage sampling and multiple weighting adjustments. Thus, the HT-based estimates from one particular sample may be far from the true total value, particularly if the probabilities of selection are negatively correlated with the characteristic of interest (see discussion in Little 2004).

Examples of existing design-based trimming methods are presented in the subsequent part of this section. In all methods, outlier weights are flagged in the survey dataset, usually through data inspection, editing, and/or computation of domain-level estimates, then trimmed to some arbitrary value. The remaining portion of the weight, called the "excess weight," is then "redistributed" to other survey units. This increases the weights on non-trimmed cases; if the increase is slight, then the associated bias is small. Redistributing the weight is done to ensure that the weights after trimming still add up to target population sizes. The underlying assumption is that decreasing the variability caused by the outlying weights offsets the increase in bias incurred by units that absorb the excess weight. The most extreme windsorized value for outlying weights is one; other possibilities include using weights within another stratum, adjacent weighting class, or some percentile value from an assumed weight distribution.

Examples of Design-Based Weight Trimming Methods

The alternative design-based methods differ in how the cutoff boundary to identify outlying weights is chosen, but they can be grouped into ad hoc methods (Ex. 1.1 and 1.2), methods that use the empirical MSE of the estimator of interest (Ex. 1.3), and methods that assume a specified parametric (skewed) distribution for the weight (Ex. 1.4 and 1.5).

Example 1.1. *The NAEP method*. To reduce extremely large HT-estimator weights in (1.1), Potter (1988) proposed trimming all weights $w_i > \sqrt{c\sum_{i \in s} w_i^2 / n}$ to this cutoff value. This method was used to trim weights in the 1986 National Association of Educational Progress sample (Johnson *et al*. 1987). The other sample units' weights are adjusted to reproduce the original weighted sum in (1). The value of c is "arbitrary and is chosen empirically by looking at values of $nw_i^2 / \sum_{i \in s} w_i^2$" (p. 457 in Potter 1988). The sum of squared adjusted weights is computed iteratively until no weights exceed the cutoff value, then the windsorized weights replace w_i in (1.1) to estimate the total. Potter (1990) claims this method outperformed other MSE-minimizing alternatives, despite the fact it does not incorporate the survey response variables of interest.

▪

Example 1.2. *The NIS method*. Chowdhury *et al.* (2007) describe the weight trimming method used to estimate proportions in the U.S. National Immunization Survey (NIS).

The "current" (at the time of the article) cutoff value was $\text{median}(w_i) + 6IQR(w_i)$, where $IQR(w_i)$ denotes the inter-quartile range of the weights. Versions of this cutoff (e.g., a constant times the median weight or other percentiles of the weights) have been used by other survey organizations (Battaglia *et al.* 2004; Pedlow *et al.* 2003; NCES 2003; Appendix A in Reynolds and Curtin 2009).

■

Example 1.3. *Cox and McGrath's MSE Trimming Method.* Cox and McGrath (1981) proposed using the empirical MSE for a sample mean estimated at various trimming levels:

$$\widehat{MSE}(\bar{y}_t) = (\bar{y}_t - \bar{y}_w)^2 + \widehat{Var}(\bar{y}_t) + 2\sqrt{\widehat{Var}(\bar{y}_t)\widehat{Var}(\bar{y}_w)}, t = 1,\ldots,T \text{ , (1.3)}$$

where t denotes the trimming level ranging from $t = 1$ for the unweighted sample mean estimator to $t = T$ denoting the fully-weighted sample estimator $\bar{y}_w$ (the sample-based estimate of the mean with no weights trimmed). Assuming that $\bar{y}_w$ is the true population mean, expression (1.3) is calculated for possible values of t, which correspond to different weight trimming cutoffs, and the cutoff associated with the minimum MSE value in (1.3) is chosen as "optimal." Potter (1988) also used this approach, estimating the MSE for a few survey variables at twenty trimming levels. He determined the "optimal" trimming by ranking the MSE (from 1 to 20) for each variable/trimming level combination, calculating the average rank across variables, and identifying the trimming level with the lowest average rank.

■

Example 1.4. *Inverse Beta Distribution Method.* Potter (1988) also considered a method that assumes the survey weights (w) follow an inverted Beta, $IB(\alpha,\beta)$, distribution:

$$f(w) = \frac{n\left(\frac{1}{nw}\right)^{\alpha+1}\left(1-\frac{1}{nw}\right)^{\beta-1}\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}, \frac{1}{n} \le w \le \infty, \qquad (1.4)$$

where $\Gamma(\bullet)$ denotes the gamma function. The IB distribution was proposed since it is a right-tailed skewed distribution. The mean and variance of the empirical IB distribution generated from the sample weights are used to estimate the IB model parameters. The trimming level is then set according to some pre-specified level in the cumulative IB distribution and weights in the tail of the distribution are trimmed.

■

Example 1.5. *Exponential Distribution Method.* Chowdhury *et al*. (2007) propose an alternative weight trimming method to the ad hoc method in Ex. 1.2. They assume that the weights in the right-tail of the weight distribution follow an exponential distribution, $Exp(\lambda)$,

$$f(w) = \lambda e^{-\lambda w} = \mu^{-1} e^{-w\mu^{-1}}, 0 \le w \le \infty, \tag{1.5}$$

where $\lambda = \mu^{-1}$ and μ is the mean of the weights within the right-tail of the weight distribution. Using the $Exp(\lambda)$ cumulative distribution function, for p denoting an arbitrarily small probability, they obtain the trimming level $-\mu \log(p)$. In application to NIS data, they assume $p = 0.01$, using the cutoff 4.6μ. They also try to account for "influential weights" (above the median) in estimating μ by adjusting the trimming level to $median(w_i) + \frac{\log(p)}{n} \sum_{i \in s} Z_i$, where $Z_i = w_i - median(w_i)$ for weights exceeding the median weight and zero otherwise. They also use Fuller's (1991) minimum MSE estimator to estimate μ to avoid extreme values influencing $\frac{1}{n} \sum_{i \in s} Z_i$ and derive the bias/MSE for children's vaccination rates (proportions). While they found the proposed method produced estimates with lower variance than the Ex. 1.2 method, the offset in the empirical MSE estimates was negligible.

■

Additional methods intended to bound the survey weights exist. Two general approaches to bounding weights have been proposed in the calibration literature: bounding the range of the weights or bounding their change before and after calibration. Deville and Sarndal (1992) describe a form of calibration that involves simply bounding the weighting adjustment factors. Isaki *et al*. (1992) show that quadratic programming can easily accomplish this bounding of the calibration weights themselves, rather than bounding the adjustment factors. These are illustrated next.

Example 1.6. *Bounding the Range of Weights*. One method proposed to bound the range of weights uses quadratic programming (Isaki *et al.* 1992). Quadratic programming seeks the vector $\mathbf{k}$ to minimize the function

$$\Phi = \frac{1}{2} \mathbf{k}' \mathbf{\Sigma} \mathbf{k} - \mathbf{z}' \mathbf{k} \tag{1.6}$$

subject to the constraint

$$\mathbf{c}' \mathbf{k} \ge c_0, \tag{1.7}$$

where $\boldsymbol{\Sigma}$ is a symmetric matrix of constants and $\mathbf{z}$ a vector of constants. For $\mathbf{d} = (d_1, \ldots, d_n)^T$ and $\mathbf{w} = (w_1, \ldots, w_n)^T$ denoting the set of input and final weights, respectively, calibration weights are produced when minimizing a distance function specified for (1.6) subject to the (1.7) constraints that the final weights reproduce population control totals but fall within specified bounds: $\sum_{i \in s} w_i \mathbf{X}_i = \mathbf{T}_x$ and $L \le w_i \le U$.

The extent to which the constraints affect the input weights depends on which units are randomly sampled. In addition, there is no developed theory that a consistent and asymptotically unbiased variance estimator is produced when the weights are constrained using quadratic programming.

▪

Example 1.7. *Bounding the Relative Change in Weights.* Another weight bounding method is to constrain the adjustment factors by which weights are changed (see Singh and Mohl 1996 for a summary). Folsom and Singh (2000) propose minimizing a constrained distance function using the generalized exponential model for poststratification. For the unit-specified upper and lower bounds L_i, U_i and centering constant C_i such that $L_i < C_i < U_i$, the bounded adjustment factor for the weights is

$$a_i(\boldsymbol{\lambda}) = \frac{L_i(U_i - C_i) + U_i(C_i - L_i)\exp\left(A_i \mathbf{X}_i^T \boldsymbol{\lambda}\right)}{(U_i - C_i) + (C_i - L_i)\exp\left(A_i \mathbf{X}_i^T \boldsymbol{\lambda}\right)}, \tag{1.8}$$

where $A_i = \dfrac{U_i - L_i}{(U_i - C_i)(C_i - L_i)}$ can control the behavior of (1.8). For example, as $L_i \to 1, C_i \to 2, U_i \to \infty$, $a_i(\boldsymbol{\lambda}) \to 1 + \exp\left(\mathbf{X}_i^T \boldsymbol{\lambda}\right)$. It can be shown that the resulting estimator with $C_i = 1$ is asymptotically equivalent to the GREG estimator. This method is incorporated in SUDAAN's proc wtadjust (RTI 2010).

▪

As Examples 1.1-1.7 illustrate, design-based weight trimming methods vary widely. Most are simple to understand (relative to the model-based approaches, see Paper 2) and implement in practice. All methods aim to change the most extreme weight values to make the largest reduction in the variance such that the overall MSE of the estimator is reduced. However, these methods are ad hoc, data-driven, and estimator-driven, so one method that works well in a particular sample may not work in other samples. No one method appears in every application paper. Redistribution also requires careful judgment by the weight trimmer (Kish 1990). The empirical MSE method is the most theoretical method – from a design-based perspective – but it is also variable-dependent. To produce one set of weights for multiple variables in practice, some ad hoc compromise – like Potter's average MSE across variables (Ex. 1.3) – must be used. In addition, the sample at hand may not produce very accurate estimates of the MSE or the weights' distribution function.

1.2.2. The Generalized Design-based Approach

A recently developed weight trimming approach uses a model to trim large weights on highly influential or outlier observations. This method was first formulated in a Bayesian framework by Sverchkov and Pfeffermann (2004, 2009); independently Beaumont and Alavi (2004) propose a similar method by extending bounded calibration (e.g., Singh and Mohl 1996) to improve the efficiency and MSE of the general regression estimator. These articles separately examine specific examples of models; the general framework and theory for estimating finite population totals was developed later by Beaumont (2008). For applications, Beaumont and Rivest (2008) use an analysis-of-variance model for "stratum jumpers," units that received incorrect base weights due to incorrect information at the time of sample selection. Beaumont and Rivest (2009) describe this method as a general approach for handling outliers in survey data.

Before introducing this method, I provide a general discussion on the approach and introduce the notation. Generally, within a given observed sample, we fit a model between the weights and the survey response variables. The weights predicted from the model then replace the weights and estimate the total. The hope is that using regression predictions of the weights will eliminate extreme weights. The underlying theory uses the properties of the model with respect to the weights; this is very different from conventional "model-based" approaches (i.e., the Bayesian modeling and superpopulation modeling approaches), where the properties are with respect to a model fit to the survey variable. Also, since this Paper aims at further understanding the method, much of the Beaumont theory is extended to incorporate the sample design as well as the weights model. Thus the related Beaumont theory to do this extension, is detailed when necessary. However, the extended theory is identified as such throughout this section and proposed methods are described in Sec. 1.3.

For the notation, denote M as the model proposed for the weights, and π the design used to select the sample. The model M trims weights by removing variability in them. This is different from the superpopulation model approach (see Valliant *et. al* 2000), where a model describes the relationships between a survey response variable and a set of auxiliary variables. In the generalized design-based approach, only one model is fit and one set of smoothed weights is produced for all variables.

Also, denote $\mathbf{I} = \left(I_1, \ldots, I_N\right)^T$ as the vector of sample inclusion indicators, i.e., I_i is the sample inclusion indicator (1 for units in the sample, 0 otherwise), and $\mathbf{Y} = \left(Y_1, \ldots, Y_N\right)^T$ the values of the survey response variable y. Generalized design-based inference is defined as "any inference that is conditional on $\mathbf{Y}$ but not $\mathbf{I}$" (p. 540 in Beaumont 2008). Noninformative probability sampling is assumed, such that $p\left(\mathbf{I} \middle| \mathbf{Z}, \mathbf{Y}\right) = p\left(\mathbf{I} \middle| \mathbf{Z}\right)$. For inferential purposes, we also consider $\mathbf{Z} = \left(Z_1, \ldots, Z_N\right)^T$, the vector of design-variables, and $\mathbf{H}_i = \mathbf{H}_i\left(\mathbf{y}_i\right)$, a vector of specified functions of different y-values for unit i. Beaumont (2008) makes specific inferences (i.e., taking expectations) with respect to the joint distribution of $\mathbf{Z}$ and $\mathbf{I}$, conditional on $\mathbf{Y}$, denoted by $F_{\mathbf{Z},\mathbf{I}|\mathbf{Y}}$. Despite confusing notation, $\mathbf{I}$ is thus the only random quantity (not $\mathbf{Z}$). In order to evaluate the estimators with respect to both the sample design and the model for the weights, I denote such expectations by $E_F = E_M\left[E_\pi(\bullet)\right]$ or $E_F = E_\pi\left[E_M(\bullet)\right]$. In the simple case of one design variable z_i and one response variable y_i, we denote the smoothed weight by $\tilde{w}_i = E_M\left(w_i \middle| I_i, z_i, y_i\right)$.

By definition, $\hat{T}_{HT} = \sum_{i \in s} w_i y_i$, where $w_i = \pi_i^{-1}$, is the HT estimator in (1.1). For particular single-stage sample designs, such as probability proportional to size sampling, this weight can vary considerably due to varying selection probabilities and result in a few extreme outliers. The Beaumont (2008) estimator, proposed to reduce the variability in the w_i's, replaces them with their expected value under the weights model:

$$
\begin{aligned}
\tilde{T}_B &= E_M\left(\hat{T}_{HT} \middle| \mathbf{I}, \mathbf{Y}\right) \\
&= \sum_{i \in s} E_M\left(w_i \middle| \mathbf{I}, \mathbf{Y}\right) y_i \,. \\
&= \sum_{i \in s} \tilde{w}_i y_i
\end{aligned}
\tag{1.9}
$$

Beaumont (2008) gave two examples for the model M, the linear and exponential model. Examples for our simple one-survey variable model (Beaumont provides equivalent expressions for multiple y-variables) are given next.

Example 1.8. *Linear model.* $E_M\left(w_i \middle| \mathbf{I}, \mathbf{Y}\right) = \mathbf{H}_i^T \boldsymbol{\beta} + v_i^{1/2} \varepsilon_i$, where $\mathbf{H}_i$ and $v_i > 0$ are known functions of $\mathbf{y}_i$, the errors are $\varepsilon_i \overset{\text{i.i.d.}}{\sim} \left(0, \sigma^2\right)$, and $\boldsymbol{\beta}, \sigma^2$ are unknown model parameters. This model produces the smoothed weight $\hat{w}_i = \mathbf{H}_i^T \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the generalized LS estimate of $\boldsymbol{\beta}$.

∎

Example 1.9. *Exponential model.* $E_M\left(w_i \middle| \mathbf{I},\mathbf{Y}\right)=1+\exp\left(\mathbf{H}_i^T\boldsymbol{\beta}+v_i^{1/2}\varepsilon_i\right)$, where $\mathbf{H}_i$, v_i, ε_i, $\boldsymbol{\beta}$, and σ^2 are given in Ex. 1.8. The exponential model produces the smoothed weight $\hat{w}_i=1+\exp\left(\mathbf{H}_i^T\hat{\boldsymbol{\beta}}\right)\sum_{i\in s}\frac{\exp\left(v_i^{1/2}\hat{\varepsilon}_i\right)}{n}$, where $\hat{\varepsilon}_i=\frac{\log\left(w_i-1\right)-\mathbf{H}_i^T\hat{\boldsymbol{\beta}}}{v_i^{1/2}}$.

■

Since $\tilde{w}_i=E_M\left(w_i \middle| \mathbf{I},\mathbf{Y}\right)$ is unknown, we estimate it with $\hat{w}_i$, found by fitting a model fit to the sample data. The estimator for the finite population total is then

$$\hat{T}_B=\sum_{i\in s}\hat{w}_i y_i . \tag{1.10}$$

If our model for the weights is correct (see expressions (A.1) and (A.2) in Appendix 1 for details), then

$$\begin{aligned} E_M\left(\hat{T}_B \middle| \mathbf{Y}\right) &= E_M\left(\sum_{i\in s}\hat{w}_i y_i \middle| \mathbf{Y}\right) \\ &= \sum_{i\in s}\tilde{w}_i y_i \\ &= \tilde{T}_B \end{aligned} . \tag{1.11}$$

Beaumont (2008) demonstrates that several properties hold under the generalized design-based approach. These properties are listed and detailed in Appendix 1. First, (see (A.3) and (A.4)) the HT estimator is always unbiased across the weights model and sample designs, i.e., $E_F\left(\hat{T}_{HT} \middle| \mathbf{Y}\right)=E_\pi\left[E_M\left(\hat{T}_{HT} \middle| \mathbf{Z},\mathbf{Y}\right) \middle| \mathbf{Y}\right]=T$. Also, if the model for the weights is correct, then the Beaumont estimator is also unbiased (A.6). However, I show that the Beaumont estimator is biased when the weights model does not hold ((A.7), with examples under specific weights models in (A.9) and (A.10)). Third, under relaxed assumptions, Beaumont also showed that these estimators are also consistent ((A.11)-(A.16)).

Beaumont also derived some special properties of his estimator under a linear model for the weights. These properties are detailed in Appendix 2, since this model is used in the Sec. 1.4.4 variance estimation evaluation. Here, it is simpler to use matrix notation, which I also use for the variance estimator in Sec. 1.3.2. Under the linear model, $E_M\left(w_i \middle| \mathbf{I}, \mathbf{Y}\right) = \mathbf{H}_i^T \boldsymbol{\beta} + v_i^{1/2} \boldsymbol{\varepsilon}$, again where $\mathbf{H}_i = \mathbf{H}_i\left(\mathbf{y}_i\right)$ is a vector of specified functions of multiple y-values for unit i. The predicted weights are $\hat{w}_i = \mathbf{H}_i^T \hat{\boldsymbol{\beta}}$, such that $\hat{T}_B = \hat{\mathbf{w}}^T \mathbf{y}_s$ is the vector of estimated totals, where $\hat{\mathbf{w}} = \mathbf{H}^T \hat{\boldsymbol{\beta}} = \left(\hat{w}_1 \ldots \hat{w}_n\right)^T$ is the vector of predicted weights, $\mathbf{y}_s = \left(\mathbf{y}_1 \ldots \mathbf{y}_n\right)^T$ is the matrix of y-values for the sample units, and $\mathbf{H} = \left[\mathbf{H}_1\boldsymbol{\beta} \quad \mathbf{H}_2\boldsymbol{\beta} \quad \cdots \quad \mathbf{H}_n\boldsymbol{\beta}\right]^T$ is the $n \times p$ matrix with rows of the $p \times 1$ vector $\mathbf{H}_i$. We then denote

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\mathbf{H}^T \mathbf{V}^{-1} \mathbf{H}\right)^{-1} \mathbf{H}^T \mathbf{V}^{-1} \mathbf{w} \\ &= \mathbf{A}^{-1} \mathbf{H}^T \mathbf{V}^{-1} \mathbf{w} \end{aligned} . \tag{1.12}$$

where $\mathbf{V} = diag\left(v_i\right)$ is the variance matrix specified under the model for the weights. Under the model, $\hat{\boldsymbol{\beta}}$ is unbiased for the parameter $\boldsymbol{\beta}$ and has variance $Var_M\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2 \mathbf{A}^{-1}$, where $\mathbf{A} = \mathbf{H}^T \mathbf{V}^{-1} \mathbf{H}$. Under the linear model, Beaumont drops the term $Var_\pi\left[\tilde{T}_B \middle| \mathbf{Y}\right]$ and derives that the variances of the HT (A.22) and Beaumont estimator (A.24) are

$$\begin{aligned} Var_B\left(\hat{T}_{HT} \middle| \mathbf{Y}\right) &= E_\pi\left[Var_M\left(\hat{T}_{HT} \middle| \mathbf{Z}, \mathbf{Y}\right) \middle| \mathbf{Y}\right] + Var_\pi\left(\tilde{T}_B \middle| \mathbf{Z}, \mathbf{Y}\right) \\ &= Var_M\left(\hat{T}_{HT} \middle| \mathbf{Z}, \mathbf{Y}\right) \\ &= \sigma^2 \sum\nolimits_{i \in s} v_i y_i^2 \\ &= \sigma^2 \mathbf{y}_s^T \mathbf{V} \mathbf{y}_s \end{aligned} , \tag{1.13}$$

where $\mathbf{V} = diag\left(v_i\right)$, and

$$Var_B\left(\hat{T}_B \middle| \mathbf{Y}\right) = \sigma^2 \sum_{i \in s} y_i \mathbf{H}_i \left[\sum_{i \in s} \frac{\mathbf{H}_i \mathbf{H}_i^T}{v_i}\right]^{-1} \sum_{i \in s} \mathbf{H}_i^T y_i \\ = \sigma^2 \mathbf{y}_s^T \mathbf{D} \mathbf{y}_s \quad (1.14)$$

where $\mathbf{D} = \mathbf{H}\mathbf{A}^{-1}\mathbf{H}^T$ has elements $D_{ij}, i = 1, \ldots, n, j = 1, \ldots, n$, $\mathbf{A} = \mathbf{H}^T \mathbf{V}^{-1} \mathbf{H}$, and the subscript B denotes Beaumont's expressions. It can be shown that under the model ((A.27)–(A.29)), Beaumont's estimator is always as efficient or more efficient than the HT estimator under the model, i.e., $Var_B\left(\hat{T}_{HT} \middle| \mathbf{Z}, \mathbf{Y}\right) \geq Var_B\left(\hat{T}_B \middle| \mathbf{Z}, \mathbf{Y}\right)$. For estimation purposes, Beaumont ignores the E_π expectation and considers the difference in the theoretical variances with respect only to the model under the one realized sample. This is a severely limiting estimation approach. I extend this to incorporate the E_π expectation and develop a variance estimator that accounts for the sample design and weights model. That is, I derive the variances under both the weights models and sample design in (A.24) and (A.26). Theoretically, if the weights model is correct, then $Var_F\left(\hat{T}_{HT} \middle| \mathbf{Z}, \mathbf{Y}\right) \geq Var_F\left(\hat{T}_B \middle| \mathbf{Z}, \mathbf{Y}\right)$ also holds (A.31).

In order to estimate the variance of the Beaumont estimator, we start with the Beaumont estimators to motivate the proposed variance estimator and theory in Sec. 1.4.1 and 1.4.2, respectively. For variance estimation, Beaumont makes the following assumption: $E_\pi\left(\hat{T}_B \middle| \mathbf{I}, \mathbf{Y}\right) = \tilde{T}_B + O_p\left(N/\sqrt{n}\right)$. From this, $E_M\left(\hat{T}_B \middle| \mathbf{I}, \mathbf{Y}\right) \approx \tilde{T}_B$, since $\hat{T}_B - T = O_p\left(N/\sqrt{n}\right)$ (see A.13). This assumption only holds if $\mathbf{y}_s$ is bounded and

$E_M\left(\hat{w}_i \middle| \mathbf{I},\mathbf{Y}\right) = \tilde{w}_i + O_p\left(N/n^{3/2}\right)$, (not $E_M\left(\hat{w}_i \middle| \mathbf{I},\mathbf{Y}\right) = \tilde{w}_i + \frac{N}{n} o_p\left(n^{-1/2}\right)$ as in Beaumont 2008) with equality holding under the linear weights model. Also, after some algebra ((A.31)–(A.34)), we can approximate $Var_F\left(\hat{T}_B \middle| \mathbf{Y}\right)$ in (1.14) with

$$Var_F\left(\hat{T}_B \middle| \mathbf{Y}\right) \approx E_\pi\left[Var_M\left(\hat{T}_{HT} \middle| \mathbf{Z},\mathbf{Y}\right) \middle| \mathbf{Y}\right] + E_\pi\left[\left\{Var_M\left(\hat{T}_B \middle| \mathbf{I},\mathbf{Y}\right) - Var_M\left(\hat{T}_{HT} \middle| \mathbf{I},\mathbf{Y}\right)\right\} \middle| \mathbf{Y}\right] \tag{1.15}$$

To estimate the variance in (1.15), Beaumont proposes

$$var_B\left(\hat{T}_B \middle| \mathbf{Y}\right) = var_\pi\left(\hat{T}_{HT} \middle| \mathbf{Z},\mathbf{Y}\right) + \left\{var_M\left(\hat{T}_B \middle| \mathbf{I},\mathbf{Y}\right) - var_M\left(\hat{T}_{HT} \middle| \mathbf{I},\mathbf{Y}\right)\right\}, \tag{1.16}$$

where $var_\pi\left(\hat{T}_{HT} \middle| \mathbf{Z},\mathbf{Y}\right)$ is a design-consistent variance estimator for $Var_\pi\left(\hat{T}_{HT} \middle| \mathbf{Z},\mathbf{Y}\right)$, and $var_M\left(\hat{T}_B \middle| \mathbf{I},\mathbf{Y}\right)$ and $var_M\left(\hat{T}_{HT} \middle| \mathbf{I},\mathbf{Y}\right)$ are consistent variance estimators with respect to the model M for the weights. In the last component of Beaumont's estimator (1.16), again the expectation with respect to the design is ignored; the estimators are conditional only on the weights model being correct. Also, the component $var_\pi\left(\hat{T}_{HT} \middle| \mathbf{Z},\mathbf{Y}\right)$ in (1.16) is not an accurate estimator of $E_\pi\left[Var_M\left(\hat{T}_B \middle| \mathbf{Z},\mathbf{Y}\right) \middle| \mathbf{Y}\right]$ in (1.15). My proposed variance estimator in Sec. 1.3.1 is a more appropriate estimator, with respect to the sample design, of (1.15) than (1.16).

From the theoretical variance under the linear model (see (A.32)–(A.36)), the Beaumont variance estimator for the difference term in braces in (1.16) is

$$var_M\left(\hat{T}_B \middle| \mathbf{Z},\mathbf{Y}\right) - var_M\left(\hat{T}_{HT} \middle| \mathbf{Z},\mathbf{Y}\right) = -\hat{\sigma}^2\left[\sum\nolimits_{i \in s} v_i y_i \left(y_i - \frac{\mathbf{H}_i^T \hat{\Omega}}{v_i}\right)\right], \tag{1.17}$$

where $\hat{\Omega} = \sum_{i \in s} y_i \mathbf{H}_i^T \left[\sum_{i \in s} \frac{\mathbf{H}_i \mathbf{H}_i^T}{v_i} \right]^{-1} \sum_{i \in s} \mathbf{H}_i^T y_i$ and $\hat{\sigma}^2$ is a model-consistent estimator of σ^2. I provide an alternative to this estimator in Sec. 1.3..

To implement the variance estimators in practice, Beaumont proposes using design-based variance estimators for $var_\pi\left(\hat{T}_{HT} \middle| \mathbf{Z}, \mathbf{Y}\right)$ and the bootstrap approach for the other terms. Note that if we were to use the variance estimator $\hat{\sigma}^2 = \hat{\sigma}_E^2 = \frac{1}{n-1} \sum_{i=1}^n \left(w_i - \hat{w}_i\right)^2$, then we can also consider (1.17) as a form of the sandwich variance estimator.

Beaumont did not prove that his variance estimator is always positive, but the first component is $O\left(N^2/n\right)$ while the second component is $O(n)$, so the second component can be expected to be much smaller in magnitude. Although not directly expressing the bias in his estimator, Beaumont acknowledges a presence of potential bias and also proposed considering the design-based mean square error (A.37) rather than the variance:

$$MSE\left(\hat{T}_B \middle| \mathbf{Y}\right) = E_\pi\left[Var_M\left(\hat{T}_B \middle| \mathbf{Z}, \mathbf{Y}\right) \middle| \mathbf{Y} \right] + B_M^2, \tag{1.18}$$

where $B_M = E_M\left[\left(\hat{T}_B - T\right) \middle| \mathbf{Z}, \mathbf{Y} \right]$ is the model-based bias of the estimator $\hat{T}_B$. Again, Beaumont proposes using a standard design-based method to estimate the variance $Var_M\left(\hat{T}_B \middle| \mathbf{Z}, \mathbf{Y}\right)$. I used the bootstrap variance estimation in the Sec. 1.4.1 empirical evaluation. While the design-based bias $\hat{T}_B - \hat{T}_{HT}$ is an unbiased estimator of the bias

B_M, $\left(\hat{T}_B - \hat{T}_{HT}\right)^2$ is not an unbiased estimator of the squared bias. Thus, to estimate (1.18), Beaumont proposes:

$$\begin{aligned} mse\left(\hat{T}_B \middle| \mathbf{Z},\mathbf{Y}\right) &= var_M\left(\hat{T}_B \middle| \mathbf{Z},\mathbf{Y}\right) + \hat{B}_M^2 \\ &= var_M\left(\hat{T}_B \middle| \mathbf{Z},\mathbf{Y}\right) + \max\left[0, \left(\hat{T}_B - \hat{T}_{HT}\right)^2 - var_\pi\left[\left(\hat{T}_B - \hat{T}_{HT}\right)\middle| \mathbf{Z},\mathbf{Y}\right]\right] \end{aligned} \tag{1.19}$$

where $var_\pi\left[\left(\hat{T}_B - \hat{T}_{HT}\right)\middle| \mathbf{Z},\mathbf{Y}\right]$ is a design-consistent estimator of $Var_\pi\left[\left(\hat{T}_B - \hat{T}_{HT}\right)\middle| \mathbf{Z},\mathbf{Y}\right]$. To ensure that $mse\left(\hat{T}_B \middle| \mathbf{Z},\mathbf{Y}\right) \le var_M\left(\hat{T}_{HT} \middle| \mathbf{Z},\mathbf{Y}\right)$ in (1.19), since $Var_M\left(\hat{T}_{HT} \middle| \mathbf{Z},\mathbf{Y}\right) \ge Var_M\left(\hat{T}_B \middle| \mathbf{Z},\mathbf{Y}\right)$, Beaumont proposes the design-based MSE estimator

$$mse_D\left(\hat{T}_B\right) = \min\left[mse\left(\hat{T}_B \middle| \mathbf{Z},\mathbf{Y}\right), var_M\left(\hat{T}_{HT} \middle| \mathbf{Z},\mathbf{Y}\right)\right]. \tag{1.20}$$

Since the Beaumont estimator can be model- and design-biased (see Appendices (A.7)-(A.10)), it is reasonable to consider an MSE estimator instead of the variance. However, in practice it is difficult to estimate the MSE. I demonstrate in Sec. 4.4 that the bias component in estimator (1.19) may perform poorly, and can drive the estimates of (1.19). When this occurs, estimator (1.20) is equivalent to the model-based variance of the HT estimator. In addition, both estimators (1.19) and (1.20) are ad hoc.

The generalized design-based inference approach appears to have performed well as a weight trimming method in some preliminary simulations and case studies. Beaumont and Alavi (2004), Beaumont (2008), and Beaumont and Rivest (2009) use this approach to produce estimators with lower MSE's than the untrimmed HT estimator. However, in a preliminary study of his proposed variance and MSE estimators, Beaumont found that incorrect model specification produced a biased variance estimator and slightly biased MSE estimator. This motivates my proposed "robust" model-based variance estimator in Sec. 1.3.1, which is robust to the specification of the weights model variance component.

In addition, Beaumont (2008) developed the models and theoretical properties for non-self-weighting element sampling designs; these models may not extend to complex sample designs like cluster sampling. These models obviously cannot be fit to self-weighting designs such as simple random sampling and stratified sampling with proportional allocation, since it is not possible to model a constant weight as a function of the survey response variables. Weight trimming can be a concern in these designs if the product $w_i y_i$ produced an influential value and trimming the weight was desired instead of editing the y_i value. Also, there is still dependence on an underlying model, which motivates the use of nonparametric methods to produce trimmed weights that are more robust to outliers.

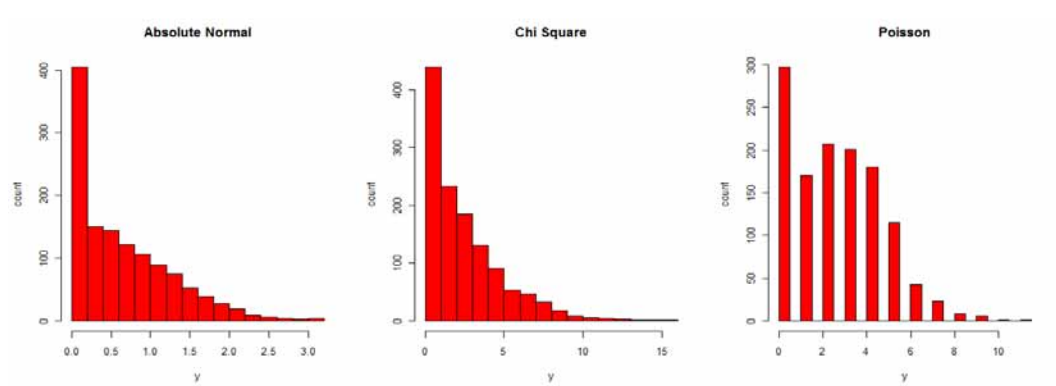
There are also several circumstances where a particular weights model may be severely inappropriate. For a simple illustration, suppose we use a linear model (Ex. 1.8) when a probability proportional to size sample with respect to some auxiliary variable x_i and $x_i \propto y_i$. The Horvitz-Thompson selection probability here is $\pi_i = nx_i / N\bar{X}$, which means $w_i = \pi_i^{-1} \propto x_i^{-1}$. Fitting a linear weights model between w_i and y_i corresponds to modeling x_i^{-1} as a function of x_i, which is clearly wrong. However, in this case, $H_i = y_i^{-1}$ is a more appropriate model. However, empirically I demonstrate in Sec. 1.4 that this model can produce inefficient totals.

Since the underlying theory and my evaluation studies indicate the Beaumont estimator is very sensitive to specification of the weights model, there is thus a need to lay out guidelines of how to choose an appropriate model. Last, Beaumont's approach for weight trimming offers no guarantee of the "redistribution of weights" property that the sum of the trimmed weights equals the sum of the untrimmed weights; in all studies this empirically did not hold. This design-based appealing property can be achieved using simple post-stratification adjustments. This suggests that this method combined with design-based weight adjustment methods can be used for improved inference. However, it also indicates that this method should not be the sole weight-adjustment method used, as it can easily over-trim HT weights (particularly when the weights model is incorrectly specified).

1.2.3. Zero-Inflated Variables

A special kind of survey variable of interest y_i is a zero-inflated variable. This variable, when plotted in a histogram, has a spike of values at zero, but some distribution for positive (or negative) values. This kind of variable is considered here to illustrate how sensitive the Beaumont method is to model failure for different types of survey variables. Some hypothetical examples are given in Figure 1.1.

Figure 1.1. Examples of Zero-Inflated Distributions



This type of variable is often modeled using a two-part model:

$$\begin{aligned}\Pr\left(y_i = 0\right) &= \phi + \left(1-\phi\right) f\left(0\right) \\ \Pr\left(y_i > 0\right) &= \left(1-\phi\right) f\left(y_i\right)\end{aligned}, \tag{1.21}$$

where $0 < \phi < 1$ and $f\left(y_i\right)$ is some specified statistical distribution. Most commonly this model is applied to discrete data, but it can also be applied to continuous variables. The means and variances are calculated using the nonzero values. Two examples follow.

Ex 1.10. *Zero-inflated Poisson distribution*. For $y_1,\ldots,y_n \overset{\text{ind}}{\sim} ZIPoi\left(\phi,\lambda\right)$, the distribution is $\Pr\left(y_i = 0\right) = \phi + \left(1-\phi\right)e^{-\lambda}$; $\Pr\left(y_i > 0\right) = \left(1-\phi\right)e^{-\lambda}\lambda^y / y!$, and the sample mean is $\bar{y} = \left(1-\phi\right)\lambda$.

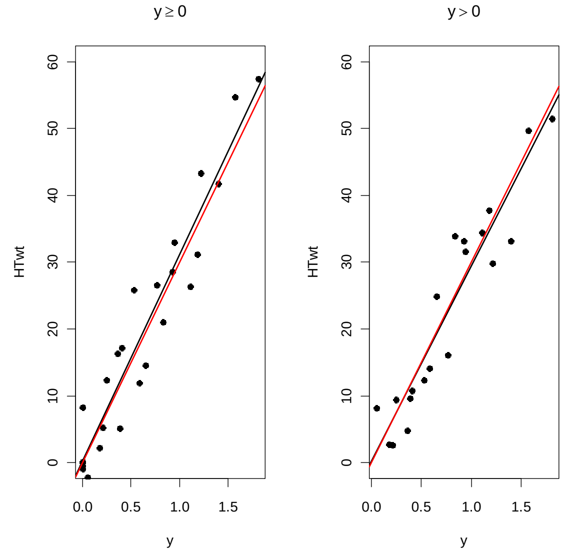
■

Ex 1.11. *Zero-inflated Binomial distribution*. For $y_1,\ldots,y_n \overset{\text{ind}}{\sim} ZIB\left(\phi,n,p\right)$, we have $\Pr\left(y_i = 0\right) = \phi + \left(1-\phi\right)\left(1-p\right)^n$; $\Pr\left(y_i > 0\right) = \left(1-\phi\right)\binom{n}{y} p^y \left(1-p\right)^{n-y}$, and the sample mean is $\bar{y} = \left(1-\phi\right)\mu$ and variance $\phi\left(V + \mu^2\right) - \phi^2\mu^2$, where μ, V are the mean and variance of the nonzero y-values.

■

Applying this model to data involves first predicting the proportion of zero/nonzero values, then fitting the model conditional on nonzero values (Thas and Rayner 2005). Here, if we were to fit a generalized design-based model and a particular y_i follows a model like (1.21), then we would introduce model misspecification error in both the estimated total and its estimated variance. To illustrate, consider the following simple (hypothetical) example in Figure 1.2 on the following page. Here the impact when including the cases with $y_i = 0$ is the slope coefficient is attenuated, while the intercept is increased. In this particular example, excluding the zero values and smoothing the HT weights produced a total six percent lower than the total produced when including them (the 45-degree line is in red, the fitted line in black). Section 1.4.2 is a simulation study to gauge the impact of this on estimating totals.

Figure 1.2. Linear Regression of Zero and Nonzero y -values vs. HT Weights, 45-Degree Line in Red, Fitted Line in Black



1.2.4. Nonparametric Regression Methods

Several "robust" methods have been developed for estimating linear models. The first, and thus most-developed, method is M-estimation (Hampel *et al.* 1986; Huber 1981), which uses maximum likelihood estimation in models with relaxed parametric assumptions. For regression models, the linearity assumption and Normal distribution assumption for the residuals can be relaxed (Hollander and Wolfe 1999; Rousseeuw and Leroy 1987). The specific methods I propose to use are described in Sec. 3.

Nonparametric methods have been proposed in other areas of survey estimation, particularly outlier detection and correction (e.g., Zaslavsky *et al.* 2001), but not weight trimming. Chambers (1996) uses M-estimation to predict the total of non-sample units under a superpopulation approach; while gains in MSE can be obtained, he noted that it is difficult in practice to produce these estimates since they require choosing a loss function. Beaumont and Alavi (2004) and Beaumont and Rivest (2009) use various M-estimation methods to produce a more robust estimator of the generalized regression model parameters and found substantial reductions in MSE of estimated totals. Beaumont (2008) uses a semi-parametric method, a penalized spline, and produced estimated totals with comparable MSE gains to other generalized design-based methods in simulations. Elliott and Little (2000) use a penalized spline to estimate Bayesian weight smoothing parameters that are more robust to general model misspecification (of the functional form of the model, not influential observations). Zheng and Little also show that a penalized spline model with Bayesian priors on the unknown model parameters can produce more accurate estimates from pps samples. This preliminary work indicates nonparametric methods offer robustness to generalized design-based models that could produce an improved weight trimming method. My proposal extends these preliminary results and incorporates methods that the nonparametric regression literature has demonstrated perform better than penalized spline or M-estimators. The methods I focus on here also can outperform parametric methods, like LS or Best Linear Unbiased Prediction (BLUP) when observations have values that influence these estimates (e.g., McKean 2004; Rousseeuw 1984; 1997; Rousseeuw and Van Driessen 1999).

There are several examples of nonparametric regression methods being used for sample-based inference. Kuo (1988) applied NP regression to sample data to estimate the finite population distribution function; Dorfman and Hall (1993) and Kuk (1993) further developed this theory and methods. Dorfman (2000) applied NP regression to sample data to estimate a finite population total; Chambers *et al.* (1993) used NP calibration and

Chambers (1996) used multivariate NP regression calibration with ridge regression to accomplish the same goal. Breidt and Opsomer (2000) estimated totals using local linear regression with design-based weights for the original model fit and residual adjustments in a method that paralleled the GREG estimator. While these methods lead to various weight adjustments, none of these methods have addressed trimming weights.

1.3. Proposed Methods

1.3.1. Variance Estimation

Here I propose variance estimation improvements to the theoretical variance given in (1.15). A variance estimator using the expectation with respect to both the sample design and weights model is an improvement over Beaumont's estimator in (1.16) and (1.29). In particular, the first component in (1.16), $var_{\pi}\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)$, is not an unbiased estimator of the theoretical $E_{\pi}\left[Var_{M}\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right]$. Here matrix-based notation is simpler. Under a general model $M\left(\mathbf{y}_{s},\mathbf{V}\right)$, if the true model is $M\left(\mathbf{y}_{s},\mathbf{\Psi}\right)$, then the proposed variance estimator is robust to using the working model $M\left(\mathbf{y}_{s},\mathbf{V}\right)$. For $\hat{T}_{B}=\sum_{i\in s}w_{i}y_{i}=\mathbf{w}^{T}\mathbf{y}_{s}$, $\mathbf{V}=diag\left(v_{i}\right)$, denoting the matrix of the variance specified in our model $E_{M}\left(\mathbf{w}\right)=\mathbf{H}^{T}\boldsymbol{\beta}$, we can write $\hat{\mathbf{w}}=\mathbf{H}^{T}\hat{\boldsymbol{\beta}}=\mathbf{H}\left(\mathbf{H}^{T}\mathbf{V}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^{T}\mathbf{V}^{-1}\mathbf{w}$. For example (from A.24 and A.26), the theoretical variances of the HT and Beaumont estimators under the linear weights model and sample design are given on the following page.

$$\begin{aligned} Var_F\left(\hat{T}_{HT}\middle|\mathbf{Y}\right) &= E_\pi\left[Var_M\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right] + Var_\pi\left[\tilde{T}_B\middle|\mathbf{Y}\right] \\ &= \sigma^2\mathbf{y}_U^T\left(\mathbf{V}\bullet diag\left(\mathbf{\Pi}\right)\right)\mathbf{y}_U + Var_\pi\left[\tilde{T}_B\middle|\mathbf{Y}\right] \end{aligned}, \quad (1.22)$$

$$\begin{aligned} Var_F\left(\hat{T}_B\middle|\mathbf{Y}\right) &= E_\pi\left[Var_M\left(\hat{T}_B\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right] + Var_\pi\left[\tilde{T}_B\middle|\mathbf{Y}\right] \\ &= \sigma^2\mathbf{y}_U^T\left(\mathbf{\Pi}\bullet\mathbf{D}\right)\mathbf{y}_U + Var_\pi\left[\tilde{T}_B\middle|\mathbf{Y}\right] \end{aligned}, \quad (1.23)$$

where $\mathbf{\Pi}$ is the $N \times N$ matrix of the selection probabilities (with diagonal elements being the first-order probabilities), $\bullet$ denotes a Hadamard product; and $\mathbf{D} = \mathbf{H}\mathbf{A}^{-1}\mathbf{H}^T$ .To estimate the variance of (1.15), we first rewrite expression (1.15) as

$$\begin{aligned} Var\left(\hat{T}_B\middle|\mathbf{Y}\right) &= \sum_{i\in s}\sigma^2 v_i y_i^2 + Var_\pi\left(\tilde{T}_B\middle|\mathbf{Y}\right) \\ &= \sum_{i\in s}\psi_i y_i^2 + Var_\pi\left(\tilde{T}_B\middle|\mathbf{Y}\right) \end{aligned}. \quad (1.24)$$

To estimate (1.24), we can use

$$var\left(\hat{T}_B\middle|\mathbf{Y}\right) = \sum_{i\in s}\hat{\psi}_i y_i^2 + Var_\pi\left(\tilde{T}_B\middle|\mathbf{Y}\right), \quad (1.25)$$

where $\hat{\psi}_i$ is an estimator of the component ψ_i. Estimation of the two (1.15) components are next examined separately, though similar estimators are proposed.

First (1.15) Component Variance Estimation

The first variance component for the linear weights model, from (1.22), is

$$\begin{aligned} E_\pi\left[Var_M\left(\hat{T}_{HT}\middle|\mathbf{Y}\right)\right] &= \sigma^2\mathbf{y}_U^T\mathbf{V}diag\left(\mathbf{\Pi}\right)\mathbf{y}_U \\ &= \sum_{i\in U} y_i^2\pi_i\psi_i \end{aligned}, \quad (1.26)$$

Expression (1.26) is still a finite population total. Thus, we can estimate it with

$$\begin{aligned} \hat{E}_\pi\left[Var_M\left(\hat{T}_{HT}\middle|\mathbf{Y}\right)\right] &= \hat{\sigma}^2\mathbf{y}_s^T\mathbf{V}diag\left(\mathbf{\Pi}\right)\left[diag\left(\mathbf{\Pi}\right)\right]^{-1}\mathbf{y}_s \\ &= \sum_{i\in s} y_i^2\hat{\psi}_i \end{aligned}, \quad (1.27)$$

where $\hat{\psi}_i = \hat{e}_i^2 / (1-h_{ii})^2$ and $\hat{e}_i^2 = (w_i - \hat{w}_i)^2 = \left(w_i - \mathbf{H}_i^T \hat{\boldsymbol{\beta}}\right)^2$. Since this term also appears in the second model component, more details are given in the following section.

Second (1.15) Component Variance Estimation

Here I propose variance estimation improvements to the second component in estimator (1.15). Here, a matrix-based notation is simpler. Under a general model $M(\mathbf{y}_s, \mathbf{V})$, if the true model is $M(\mathbf{y}_s, \boldsymbol{\Psi})$, then the proposed variance estimator is robust to using the working model $M(\mathbf{y}_s, \mathbf{V})$. For $\hat{T}_B = \sum_{i \in s} w_i y_i = \mathbf{w}^T \mathbf{y}_s$, $\mathbf{V} = diag(v_i)$ denoting the matrix of the variance specified in our model, $E_M(\mathbf{w}) = \mathbf{H}^T \boldsymbol{\beta}$, we can write $\hat{\mathbf{w}} = \mathbf{H}^T \hat{\boldsymbol{\beta}} = \mathbf{H}\left(\mathbf{H}^T \mathbf{V}^{-1} \mathbf{H}\right)^{-1} \mathbf{H}^T \mathbf{V}^{-1} \mathbf{w}$. From this, it can be shown (A.21) that the variance of $\hat{T}_B$ under the model M is

$$\begin{aligned} Var_M\left(\hat{T}_B \middle| \mathbf{I}, \mathbf{Y}\right) &= Var_M\left(\hat{\mathbf{w}}^T \mathbf{y}_s\right) \\ &= \mathbf{H}^T \mathbf{A}^{-1} \mathbf{H}^T \mathbf{V}^{-1} \mathbf{H}^T \mathbf{A}^{-1} \mathbf{H} \end{aligned}, \tag{1.28}$$

where $\mathbf{A} = \mathbf{H}^T \mathbf{V}^{-1} \mathbf{H} = \sum_{i \in s} \frac{\mathbf{H}_i^T \mathbf{H}_i}{v_i}$ in Beaumont's (2008) notation in the $\hat{\Omega}$ expression in (1.17). If we assume that the variance parameter in $\mathbf{V}$ is incorrectly specified, then an appropriate variance estimator is

$$var_M\left(\hat{T}_B \middle| \mathbf{I}, \mathbf{Y}\right) = \mathbf{H}^T \mathbf{A}^{-1} \mathbf{H}^T \hat{\boldsymbol{\Psi}}_E^{-1} \mathbf{H}^T \mathbf{A}^{-1} \mathbf{H}, \tag{1.29}$$

where $diag\left(\hat{\boldsymbol{\Psi}}_E\right) = \hat{e}_i^2$ has elements that are the residuals under the model. This "sandwich" variance estimator $\hat{\boldsymbol{\Psi}}_E$ is approximately model-unbiased for $\mathbf{V}$, and it is a natural "first-choice" estimator since $E_M\left(\hat{e}_i^2\right) \approx \psi_i^2$ in large samples (White 1980). Note that under the linear model for the weights, this sandwich estimator form of Beaumont's estimator can be written similar to (1.17) as

$$var_M\left(\hat{T}_B\middle|\mathbf{Y}\right)=var_\pi\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)-\hat{\sigma}_E^2\sum_{i\in s}v_i\left(y_i-\frac{H_i}{v_i}\hat{\Omega}\right)^2 . \qquad (1.30)$$

While $\hat{e}_i$ is a consistent estimator of ψ_i, it can underestimate ψ_i in small or moderate sample sizes since $Var_M\left(\hat{e}_i\right)=v_i\sigma^2\left(1-h_{ii}\right)$, where the leverage of unit i, denoted h_{ii}, is the diagonal element of $\mathbf{H}\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T$ and $0\le h_{ii}\le 1$. It is well-established (White 1980, Horn *et al.* 1975, Efron 1982; MacKinnon and White 1985) that the leverage contribution to the variance is not constant: a few units with larger leverages (and thus more "influence") will contribute more to the variance underestimation in ψ_i than units with small leverages. Thus, the variance of $\hat{e}_i$ for large h_{ii}'s tend to be smaller than the variance of $\hat{e}_i$ in observations with small h_{ii} values.

Several correction factors have been proposed to overcome this non-constant variance in e_i (e.g., see above references). Of these, Efron (1982) and MacKinnon and White (1985) propose using $\hat{e}_i^2/\left(1-h_{ii}\right)^2$. The resulting variance estimator has also been shown to be asymptotically equivalent to the jackknife replication variance estimator (Valliant *et al.* 2000 p. 141), which is a conventional design-based variance estimation method. Adopting this here gives

$$var_{B*}\left(\hat{T}_B\middle|\mathbf{I},\mathbf{Y}\right)=\mathbf{y}_s^T\mathbf{A}^{-1}\mathbf{H}^T\hat{\mathbf{\Psi}}_{E*}^{-1}\mathbf{H}^T\mathbf{A}^{-1}\mathbf{y}_s , \qquad (1.31)$$

where $diag\left(\hat{\mathbf{\Psi}}_{E*}\right)=\hat{\psi}_i^2=\left(w_i-\hat{w}_i\right)^2/\left(1-h_{ii}\right)^2$. If the model errors are misspecified, then (1.31) will still give an accurate estimate of the true variance component (1.28) in expectation (see Sec. 1.3.2 for more detail). Combining (1.31) with (1.27) gives the following total variance under the linear weights model as follows:

$$var_{B*}\left(\hat{T}_B\middle|\mathbf{Y}\right)=\sum_{i\in s}\hat{\psi}_i y_i^2-\left[\sum_{i\in s}\hat{\psi}_i\left(y_i-\frac{H_i\hat{\Omega}}{v_i}\right)^2\right], \tag{1.32}$$

where $\hat{\psi}_i=\left(w_i-\hat{w}_i\right)^2\big/\left(1-h_{ii}\right)^2$ and v_i is from the working weights model. The variance estimator in (1.32) estimates the components of (1.15) with respect to both the sample design and the weights model. As a result, it is an improvement over Beaumont's estimator in (1.16). In particular, the first component in (1.16) $var_{\pi}\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)$ is not an unbiased estimator of the theoretical $E_{\pi}\left[Var_M\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right]$.

1.3.2. Additional Variance and MSE Theoretical Properties When the Weights Model Does Not Hold

As noted earlier, when the weights model is incorrectly specified, the Beaumont estimator can be biased for the finite population total. Under this circumstance, it is reasonable to consider MSE estimation, but this is difficult to achieve in practice. The weights model also impacts variance estimation. In particular, I show here that when the linear weights model is incorrect, the variance estimator in Sec. 1.3.1. is positively biased. However, the variance estimator will still underestimate the MSE. The theory presented here is largely borrowed from parallel results in Valliant *et al*. (2000, p.150-151) and is detailed in Appendix 6.

Suppose that the working weights model M is used, when the true weights model is actually $\tilde{M}$. If the variance component specification in model M is wrong, then we have

$$E_{\tilde{M}}\left[var_B\left(\hat{T}_B\middle|\mathbf{Y}\right)\right]\cong\sum\nolimits_{i\in s}v_i y_i^2 E_{\tilde{M}}\left[\left(w_i-\hat{w}_i\right)^2\right]$$
$$\cong\sum\nolimits_{i\in s}v_i y_i^2\psi_i+\sum\nolimits_{i\in s}v_i y_i^2\left[E_{\tilde{M}}\left(w_i-\hat{w}_i\right)\right]^2 \quad . \qquad (1.33)$$
$$=Var_{\tilde{M}}\left(\hat{T}_B\middle|\mathbf{Y}\right)+\sum\nolimits_{i\in s}v_i y_i^2\left[E_{\tilde{M}}\left(w_i-\hat{w}_i\right)\right]^2$$

Both of the components in (1.33), the model-variance and the positive bias term, have the same order of magnitude, $O_p\left(N^2/n\right)$. This is the same order of magnitude as the variance component in the MSE (1.18). However, the bias component (the second component in (1.18)) has order $O_p\left(N^2\right)$, which is higher than $O_p\left(N^2/n\right)$. This means that when the weights model does not hold, the variance estimator will be positively biased but will still underestimate the true MSE.

1.3.3. Nonparametric Generalized Design-Based Weight Smoothing

Here I propose new generalized design-based estimators of totals that are similar to the HT estimator in (1.1), but with different weights $\widehat{w}_i$:

$$\hat{T}_{NP}=\sum\nolimits_{i\in s}\widehat{w}_i y_i \ . \qquad (1.34)$$

To protect the estimator (1.34) from influential values in the weights and survey response variables, the smoothed weights $\widehat{w}_i$ are developed from NP models. I focus on influential values in the survey response variable (the predictors), the weight (dependent variable), or both since they can influence the generalized design-based smoothed weights. Different NP models can be used to fit the weights and therefore produce different sets of smoothed weights. Specifically, $\tilde{w}_i$ is the weight predicted from the nonlinear model η in

$$\log\left[E_\eta\left(w_i\middle|\mathbf{I},\mathbf{Y}\right)\right]-1=\mathbf{H}^T\boldsymbol{\gamma}\ , \qquad (1.35)$$

where the log transformation in (1.35) is used to reduce the skewness in the survey weights (e.g., Fuller (1991) and Section 1.4 evalation studies). From (1.35), the predicted weights are $\hat{w}_i = 1 + \exp\left(\mathbf{H}_i^T \hat{\boldsymbol{\gamma}}\right)$, which also avoids the problem of producing negative weights (Valliant 2004). I propose to use three specific nonparametric (NP) alternatives to produce $\hat{\boldsymbol{\gamma}}$, as summarized in Table 1.1. The notation used in the table is discussed following the table.

Table 1.1. Proposed Nonparametric Estimators

Method	$\boldsymbol{\gamma}$-Estimator	Additional Terms
MM	$\hat{\boldsymbol{\gamma}}_{MM} = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \rho_1\left(\frac{\log(w_i - 1) - \mathbf{H}_i^T \boldsymbol{\gamma}}{\hat{\sigma}_S^2}\right)$	$\hat{\sigma}_S^2 = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^p}\left[\hat{\sigma}_S^2(\boldsymbol{\gamma})\right]$, where we solve for $\hat{\sigma}_S^2(\boldsymbol{\gamma})$ in $\frac{1}{n}\sum_{i=1}^{n} \rho_0\left(\frac{\log(w_i - 1) - \mathbf{H}_i^T \boldsymbol{\gamma}}{\hat{\sigma}_S^2(\boldsymbol{\gamma})}\right) = b$
LMS	$\hat{\boldsymbol{\gamma}}_{LMS} = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^p} median\left[\log(w_i - 1) - \mathbf{H}_i^T \boldsymbol{\gamma}\right]^2$	(none)
LTS	$\hat{\boldsymbol{\gamma}}_{LTS} = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^p} \sum_{i=1}^{q}\left[\log(w_i - 1) - \mathbf{H}_i^T \boldsymbol{\gamma}\right]_{(i)}^2$	$q = \left\lfloor \frac{n + p + 1}{2} \right\rfloor$

The predicted weights are then less influenced by outliers in the weight and the survey response variables. Each method has associated strengths and weaknesses, but all were developed to be more robust than the M-estimation method. While the tradeoff for such robustness can be lower efficiency (higher variance), the NP literature has demonstrated that these four robust estimators can be more efficient than alternatives. This bias/efficiency tradeoff is quantified empirically in simulations (Sec. 1.4). Note that the proposed methods do not have closed-form estimates for the $\boldsymbol{\gamma}$-parameters, so iterative methods must be used to find solutions in practice.

Note that the Table 1.1 estimators are similar to the least squares (LS) estimator. For LS, we find the argument (arg) $\boldsymbol{\gamma}$ that minimizes the total squared residuals under model (1.35) over the p -dimensional space of real numbers ($\mathbb{R}^p$), i.e.,

$$\hat{\boldsymbol{\gamma}}_{LS} = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^p} \sum_{i\in s}\left[\log(w_i - 1) - \mathbf{H}_i^T\boldsymbol{\gamma}\right]^2 . \tag{1.36}$$

However, $\hat{\boldsymbol{\gamma}}_{LS}$ is not robust to outliers. The "degree of robustness" can be measured by the break-down point (Hampel 1971; Donoho 1982; Donoho and Huber 1983). The finite sample break-down point measures the maximum fraction of outliers within a given sample that is allowed without the estimator going to infinity. To solve this lack of robustness, Rousseeuw (1984) proposed the LTS and LMS estimators, which have high asymptotic break-down points (0.5). M-estimators have a low break-down point (0), but Rousseeuw and Yohai (1984) proposed robust M-estimates of the residual scale, the S-estimates. Yohai (1985) introduced the MM estimator to have a high break-down point (0.5 asymptotically for conventional choices of ρ_0, ρ_1) and still be efficient under Normal errors.

The first NP method in Table 1.1 is the *MM estimator*, which is a combination of its predecessors, the M- and S-estimators. Huber (1981) introduced "maximum likelihood type" (M) estimation to conduct robust regression analysis for nonlinear equations. In this, we assume to have n independent observations from a location family with probability density function $f\left(y-\mu\right)$ for some function that is symmetric around μ (the "location parameter" at the center of the distribution, not necessarily the mean). The function ρ is designed to dampen the effect of extreme values in the residuals $\log\left(w_i - 1\right) - \mathbf{H}_i^T\hat{\boldsymbol{\gamma}}$. M-estimators use $\log\left(f\right)$ as ρ, the MLE solves for μ in

$\min_{\mu} \sum_{i=1}^{n} -\log\left[f\left(y_i - \mu\right)\right] = \min_{\mu} \sum_{i=1}^{n} \rho\left[f\left(y_i - \mu\right)\right]$. If ρ' exists, then the M-estimator is obtained by solving $\sum_{i=1}^{n} \rho'\left(y_i - \hat{\mu}\right) = 0$ for $\hat{\mu}$. M-estimators are robust to outliers in the response variable, but are as sensitive to covariate outliers as the LS estimators. This method also requires choosing the loss function (or "scaling factor") ρ', which can be difficult in practice (Chambers 1986; Beaumont and Rivest 2009).

Alternatively, "scale" (S) estimation fits a line that minimizes a robust estimate of the scale (i.e., the unknown variance parameter) of the residuals. This method was developed to improve the lack of covariate-robustness with M-estimation. While S-estimators are outlier-resistant, they can be inefficient (Rousseeuw and Leroy 1986, 2003; Stromberg 1993).

MM estimation is a compromise of the M- and S-estimation methods, to overcome their problems and retain their benefits (Yohai 1987). If we denote $G_0\left(\mathbf{H}\right)$ and $F_0\left(\mathbf{e}\right)$ as the distributions of the $\mathbf{y}_s$-variables and errors, respectively, then the joint distribution of $\mathbf{w}, \mathbf{H}$ is given by $H_0\left(\mathbf{w}, \mathbf{H}\right) = G_0\left(\mathbf{H}\right) F_0\left(\mathbf{w} - \mathbf{H}^T \boldsymbol{\gamma}\right)$. MM estimators use two loss functions, denoted by ρ_0 and ρ_1, which determine the estimator's theoretical properties (the "breakdown point" and efficiency, respectively, Huber 1981). First we obtain the S-estimate for the model variance of the errors, σ^2 as $\hat{\sigma}_S^2 = \arg\min_{\boldsymbol{\gamma} \in \mathbb{R}^p} \left[\hat{\sigma}_S^2\left(\boldsymbol{\gamma}\right)\right]$, where $\hat{\sigma}_S^2\left(\boldsymbol{\gamma}\right)$ is obtained by solving $\frac{1}{n} \sum_{i=1}^{n} \rho_0\left(\frac{\log(w_i - 1) - \mathbf{H}_i^T \boldsymbol{\gamma}}{\hat{\sigma}_S^2\left(\boldsymbol{\gamma}\right)} \right) = b$ and b is a value that must be set in advance. The loss function ρ_0 is assumed to be an even, continuous, and non-decreasing function on

$[0, \infty)$, with the properties that $\rho_0(0) = 0$ and $\sup_{\mathbf{e} \in \mathbb{R}} \rho_0(\mathbf{e}) = 1$. The recommended choice of $b = E_{F_0}[\rho_0(\varepsilon_1)]$ ensures that $\hat{\sigma}_S^2$ is a consistent estimator (Salibian-Barrera 2006).

The MM estimator for the parameters in $\boldsymbol{\gamma}$ is any local minimizer of a specified function $f(\boldsymbol{\gamma})$ that maps from the *p*-dimension real space to the one-dimensional space of positive numbers, denoted by $f(\boldsymbol{\gamma}): \mathbb{R}^p \mapsto \mathbb{R}^+$. This leads to the Table 1.1 estimator. MM estimators are more robust to outliers in the dependent variable and covariates and are more efficient than S- and M-estimators by themselves (Smyth and Hawkins 2000, Stromberg 1993, Tatsuoka and Tyler 2000; Yohai 1987). However, they require careful choice of the loss functions ρ_0 in $\hat{\sigma}_S^2$ and ρ_1 in the function $f(\boldsymbol{\gamma})$. This may be difficult to validate in practice, but conventional choices perform relatively well compared to the Beaumont estimators in my Sec. 1.4.1 evaluation study.

The second NP method I consider is the *least median of squares (LMS) estimator*, which minimizes the median of the squared residuals, as shown in Table 1.1. Since the median is very robust against outliers, LMS estimators are the "highest breakdown" estimators. This means they are the most robust estimators to outliers in both the dependent and covariate variables (Rousseeuw 1984, Rousseeuw and Van Driesen 1999, Rousseeuw and Leroy 2003; Rousseeuw and Ryan 1997, 2008). The LMS estimator also transforms properly under certain transformations and has no rescaling factors like the MM estimator, so it is widely used in many applications (Rousseeuw 1984, 1997). However, it may not be the most efficient estimator (Ripley 2004) and is sensitive to data values that are close to the median (Davies 1993; Edelsbrunner and Souvaine 1990).

The last NP method is *least-trimmed squares estimation*, which was also developed as a more robust alternative to M-estimation. It also produces faster convergence in the iterative methods used to obtain the parameter solutions for the MM and LMS estimators, which do not have closed-forms. The *least-trimmed squares (LTS) estimator* is the LS estimator taken over the smallest $q = \lfloor (n+p+1)/2 \rfloor$ squared residuals. This method's proponents argue that the LMS estimator is very robust to outliers in both the dependent and covariate variables, needs no rescaling, is more efficient than LMS estimators, and is just as resistant to outliers (Rousseeuw and Van Driesen 1999). However, the NP literature and practitioners have not embraced this method as being the "best" overall alternative. I use simulations and a case study (Sec. 1.4) to identify which methods seem most promising to estimate generalized design-based model parameters.

Based on the related literature, I summarize the expected performance of the proposed estimators against LS estimator, in Table 1.2. Each method can be compared using outlier-robustness, efficiency, and how easily the method can be implemented in practice in terms of the estimators "breakdown point," the fraction of the sample that is allowed to be outlying without the estimator being undefined. The "Yes/No" rating means "Yes" under particular conditions and "No" in others. It is apparent that there is no "one best" estimator clearly identified in the NP literature theory and applications.

Table 1.2. Summary of Proposed Nonparametric Estimator Properties

	Robust to Outliers		**Efficient**	**Break-down Point**
Estimator	*Survey Response*	*Weights*		*(asymptotic)*
LS	No	No	Yes	n^{-1}
MM	Yes	Yes	Yes/No	0.5
LMS	Yes	Yes	No	0.5
LTS	Yes	Yes	Yes/No	0.5

Theoretically, for the NP methods, if $E_{\eta}\left(\hat{\boldsymbol{\gamma}}_{NP}\right)=\boldsymbol{\gamma}$, then under the nonlinear model η for the weights, the corresponding NP total is unbiased. That is, if

$$\begin{aligned} E_{\eta}\left(\hat{T}_{NP}\middle|\mathbf{Y}\right) &= \sum\nolimits_{i\in s} E_{\eta}\left(\mathbf{H}_i^T\hat{\boldsymbol{\gamma}}_{NP}\middle|\mathbf{Y}\right) \\ &= \sum\nolimits_{i\in s}\mathbf{H}_i^T\boldsymbol{\gamma} y_i \\ &= \sum\nolimits_{i\in s}\tilde{w}_i y_i \\ &= \tilde{T}_B \end{aligned}, \qquad (1.37)$$

then the same results from the Beaumont estimator applies here. The MM (Yohai 1985, 1987; Huber 1981), LTS (Čižek 2004; Andrew 1987, 1992; Arcones and Yu 1994; Yu 1994), and LMS (Zinde-Walsh 2002; Gelfand and Vilenkin 1964) estimators have all been proven in the related literature to be asymptotically model-unbiased and consistent for $\boldsymbol{\gamma}$.

1.4. Evaluation Studies

I conducted four evaluation studies, each related to an approach proposed in Section 1.3, so this section is divided into four such sections. I focus on estimation of totals in the first three evaluations, then estimates of their variances in the fourth. First, I use the simulations designed by Beaumont to initially gauge the performance of the NP estimators proposed in 1.3.2. Second, I demonstrate how alternative weight models can estimate totals of zero-inflated survey variables. Third, I demonstrate how the proposed NP estimators are improvements over the Beaumont estimators for weights that use nonresponse adjustments. Last, I use simulations to empirically evaluate my proposed variance estimator against Beaumont’s variance and MSE estimators.

1.4.1. Extending Beaumont's Study for Nonparametric Estimators

Here I mimic Beaumont's simulation study of alternative totals, which involved a pseudo population of 10,000 observations with four variables of interest: $z_i \sim \exp(30)+0.5$, $y_{1i} = 30+\varepsilon_{1i}$, $y_{2i} = 30+0.1498 z_i + \varepsilon_{2i}$, $y_{3i} = 30+2.9814 z_i + \varepsilon_{3i}$, where $\varepsilon_{1i},\varepsilon_{2i},\varepsilon_{3i} \overset{\text{ind}}{\sim} N(0,2000)$. The slope coefficients creating y_{1i}, y_{2i}, y_{3i} were chosen to vary their correlation with z_i: $\rho(y_1,z)=0, \rho(y_2,z)=0.1$, and $\rho(y_3,z)=\sqrt{0.8}$. Figure 1.3 shows the scatterplots and histograms of the pseudopopulation values.

Figure 1.3. Beaumont Simulated Population and Loess Lines

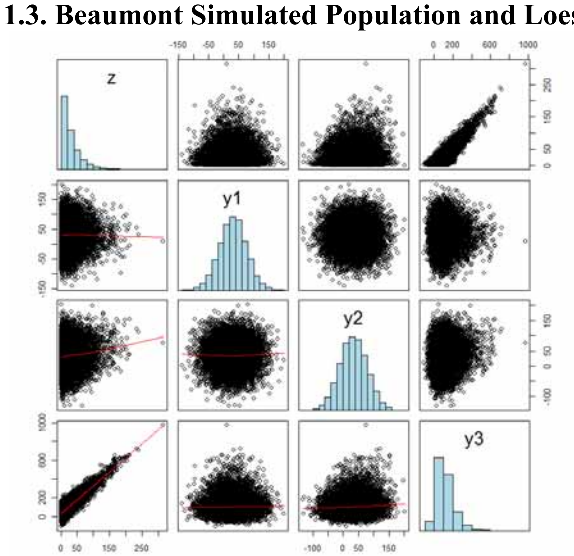
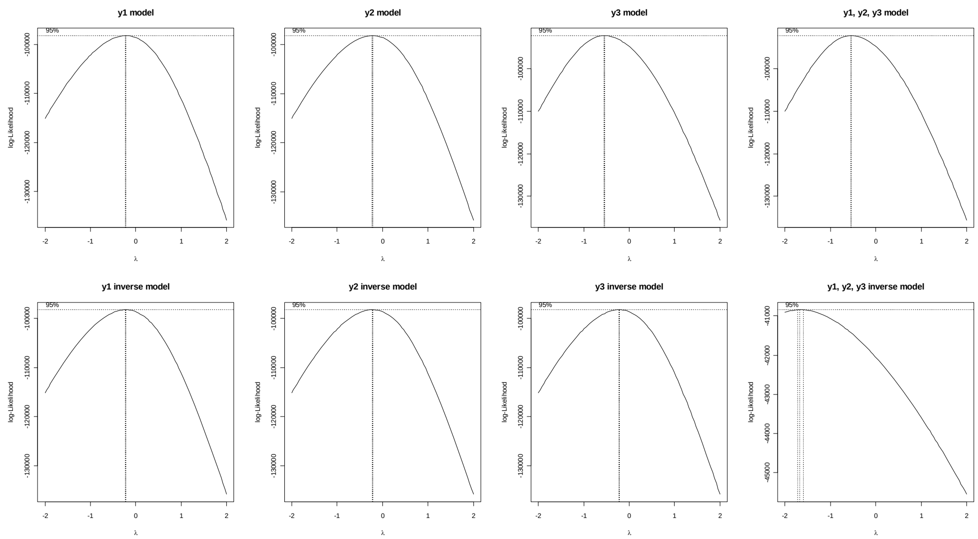


Figure 1.3 demonstrates Beaumont's unusual choice of simulation data; in particular, for the variable y_3, $pp(z)$ sampling is not the most efficient choice (as the variance of y_3 decreases with z, however we select units with larger y_3 values with higher probabilities of selection). There also seems to be no relationship between z and y_1 or y_2, so this may be a difficult pseudopopulation to find one weights model that is appropriate for all three variables of interest. I illustrate this by calculating the probabilities of selection and HT weights for all units in the population, then conducting

a Box-Cox transformation for w_i using y_1, y_2, y_3 and $y_1^{-1}, y_2^{-1}, y_3^{-1}$ regressed separately and together on w_i. The likelihood plots for a sample of size 100 are in Figure 1.4.

Figure 1.4. Population Box-Cox Transformation Plots for Modeling HT Weights as a Function of different y-variables, Beaumont Simulated Population



These plots suggest that different functions of w_i should be used for different y-variables. For the linear model, the plots variables y_1 and y_2 having the likelihood function maximized at $\lambda = -0.2$, corresponding to $w_i^{-0.2}$, suggests that a log function may be appropriate enough. However, for y_3 and y_1, y_2, y_3, $\lambda = -0.7$ suggests that $w_i^{-0.7}$ may be more appropriate. For the function $H_i = y_i^{-1}$, an intuitive choice for probability proportional to size sampling, separately y_1, y_2, y_3 each suggest that a log transformation is appropriate, but together they imply $w_i^{-1.5}$ may be more appropriate. Based on these mixed results and a realistic scenario in which these underlying relationships are unknown in practice; a linear model (with no transformation on w_i)

model was fit with $H_i = y_i^{-1}$ and an exponential weights model (i.e., $\log(w_i - 1)$) was fit to no transformation on y_1, y_2, y_3.

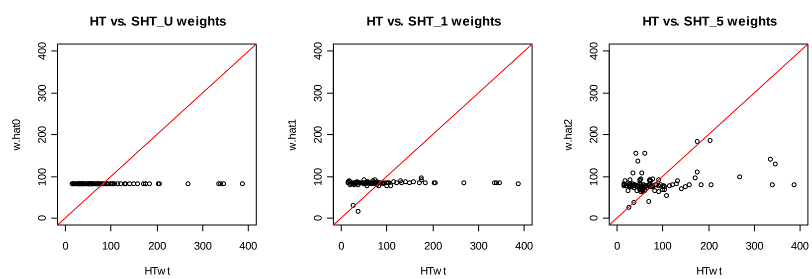
Two thousand samples of size 100 were drawn using Sampford's $pps(z)$ method (I used a smaller sample size than Beaumont's 500 and fewer simulation iterations instead of his 10,000 to reduce simulation computation time). The following estimators (using Beaumont's naming convention in the "" below) were included:

- HT: $\hat{T}_k^{HT} = \sum_{i \in s} \frac{y_{ki}}{\pi_i}, k = 1,2,3$, where $\pi_i = 100 z_i \big/ \sum_{i \in U} z_i$;
- Hajék: $\hat{T}_k^{HJ} = \frac{N \hat{T}_k^{HT}}{\hat{N}}$, where $\hat{N} = \sum_{i \in s} w_i$;
- "SHT_U": $\hat{T}_k^{SHT\_0} = \sum_{i \in s} \hat{w}_i y_{ki}$, where w_i is predicted from an intercept-only model;
- "SHT_1": $\hat{T}_k^{SHT\_1} = \sum_{i \in s} \hat{w}_i y_{ki}$, where w_i is predicted from a linear or exponential weights model;
- "SHT_5": $\hat{T}_k^{SHT\_5} = \sum_{i \in s} \hat{w}_i y_{ki}$, where w_i is predicted from a fifth-order polynomial model using all y_1, y_2, y_3 and stepwise variable selection to retain "only the most important predictors" (p. 547). Beaumont did not indicate how the stepwise selection was done; I used backwards selection and the AIC measure to select the predictors.
- The three proposed nonparametric estimators, denoted $\hat{T}_k^{MM}$, $\hat{T}_k^{LMS}$, and $\hat{T}_k^{LTS}$ in (1.30), that use weights predicted from NP models fit to all y_1, y_2, y_3.

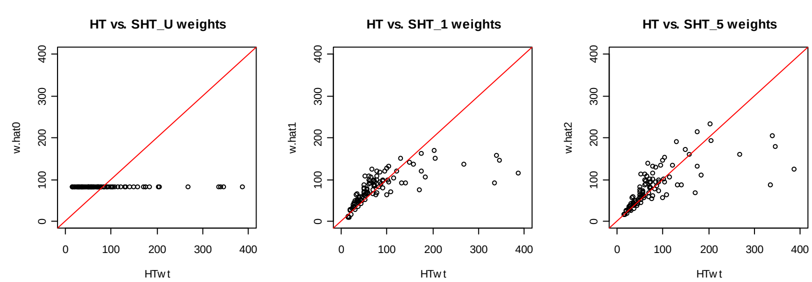
Three weight models were used: the multivariate equivalents to Ex. 2.7 (the "Linear Model"), Ex. 2.8 (the "Exponential Model"), and the heuristic choice of modeling the inverse of the weights described at the end of Sec. 1.2.2 (the "Inverse Model"). For the linear and inverse models, which Beaumont omitted, I chose $H_i = y_i^{-1}$ and $H_i = y_i$, respectively. Like Beaumont, I use $H_i = y_i$ for the exponential model. Figure 1.5 shows an example of how the various methods produce smoothed weights under these two models, compared to the HT weights prior to smoothing.

Figure 1.5. One-Sample Examples of Weights Before/After Smoothing: Beaumont Methods

Linear Model: $w_i = \beta_0 + \beta_1 H_{1i} + \beta_2 H_{2i} + \beta_3 H_{3i}\,, H_i = y_i^{-1}\,,\ v_i = 1$



Exponential Model:, $\log\left(w_i - 1\right) = \beta_0 + \beta_1 H_{1i} + \beta_2 H_{2i} + \beta_3 H_{3i}\ \ H_i = y_i\,, v_i = 1$

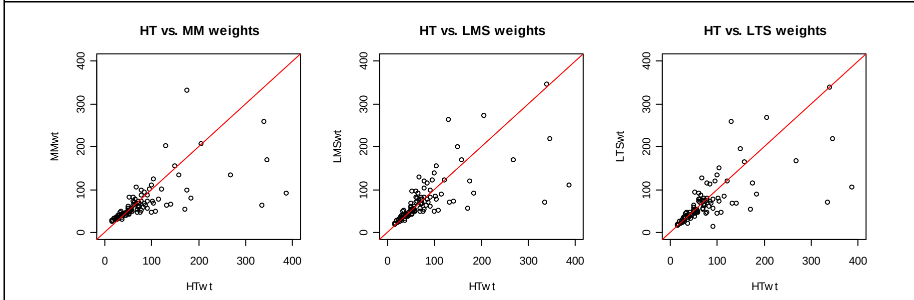


Inverse Model: $w_i^{-1} = \beta_0 + \beta_1 H_{1i} + \beta_2 H_{2i} + \beta_3 H_{3i}\,, H_i = y_i\,, v_i = 1$

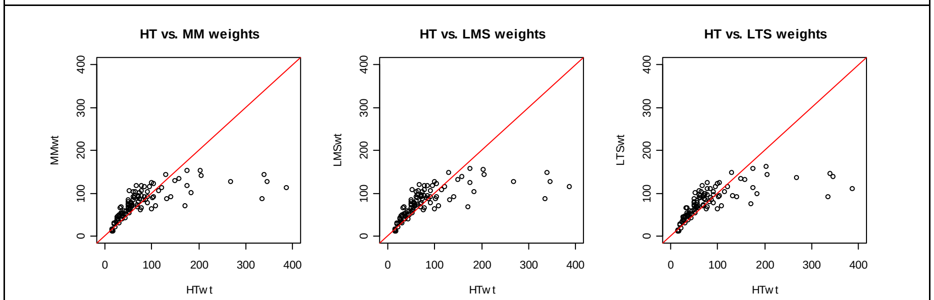
The pps sampling method produces a skewed HT weight distribution, as units with higher values of z_i are selected with higher probabilities of selection and thus have smaller HT weights. The SHT_U weights are the most smoothed, as all weights are assigned the intercept value, the average of the HT weights. The linear SHT_1 and polynomial regression SHT_5 models both reduced the variation in the weights, though for this example the linear weights model resembles the SHT_U weights. Similar plots for the proposed NP methods are shown in Figure 1.6. They produce similar weight distributions to SHT_1, with a smaller range of trimmed weights.

Figure 1.6. One-Sample Examples of Weights Before/After Smoothing: NP Methods

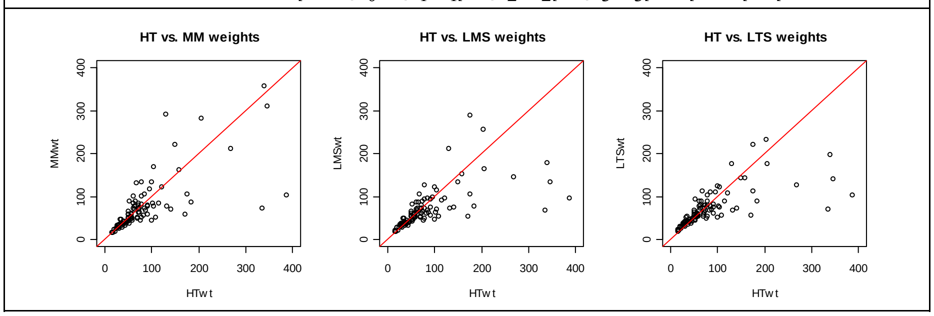
Linear Model: $w_i = \beta_0 + \beta_1 H_{1i} + \beta_2 H_{2i} + \beta_3 H_{3i}\,, H_i = y_i^{-1}\,,\ v_i = 1$



Exponential Model:, $\log\left(w_i - 1\right) = \beta_0 + \beta_1 H_{1i} + \beta_2 H_{2i} + \beta_3 H_{3i}\ \ H_i = y_i\,, v_i = 1$



Inverse Model: $w_i^{-1} = \beta_0 + \beta_1 H_{1i} + \beta_2 H_{2i} + \beta_3 H_{3i}\,, H_i = y_i\,, v_i = 1$



To evaluate the estimated totals across the alternatives, for comparability, I use the same evaluation measures as Beaumont (2008):

- *Percentage Relative Bias:* $Relbias\left(\hat{T}_k\right) = 100\left(2000 T_k\right)^{-1} \sum_{b=1}^{2000}\left(\hat{T}_{bk} - T_k\right)$,
- *Variance Ratio (to HT Variance):* $VarRatio\left(\hat{T}_k\right) = \dfrac{2000^{-1}\sum_{b=1}^{2000}\left(\hat{T}_{bk} - \bar{\hat{T}}_{bk}\right)^2}{2000^{-1}\sum_{b=1}^{2000}\left(\hat{T}_{bk}^{HT} - \bar{\hat{T}}_{bk}^{HT}\right)^2}$,
- *Relative Root Mean Square Error:* $RelRMSE\left(\hat{T}_k\right) = \dfrac{\sqrt{\sum_{b=1}^{2000}\left(\hat{T}_{bk} - T_k\right)^2}}{\sqrt{\sum_{b=1}^{2000}\left(\hat{T}_{bk}^{HT} - T_k\right)^2}}$,

where $\hat{T}_{bk}$ is an estimate of the true total $T_k = \sum_{i \in U} y_{ki}$ for variable y_k on simulation sample $b\,(b = 1,\ldots,2000)$, $\bar{\hat{T}}_{bk} = 2000^{-1}\sum_{b=1}^{2000} \hat{T}_{bk}$, and $\bar{\hat{T}}_{bk}^{HT} = 2000^{-1}\sum_{b=1}^{2000} \hat{T}_{bk}^{HT}$ (Note: Beaumont incorrectly labeled the percentage relative root mean square error as the relative efficiency). These measures are shown in Table 1.3 on the following page, along with the associated estimates from Beaumont's simulation results (who used the exponential model and the same simulation conditions except sample size and number of iterations, from Table 1 on p. 547 in Beaumont 2008).

Some general comments can be made from Table 1.3. First, as expected, the HT estimator is nearly unbiased across the 2,000 samples, but has a large variance relative to some of the model-based alternatives. Both the Beaumont and proposed NP estimators have nonzero biases, but the magnitude of the bias in the NP estimators is generally equal or less than the corresponding Beaumont estimators. This implies that if the weights model is incorrectly specified, then both the Beaumont and NP methods can over- or under-trim the weights, producing a bias in the estimated totals in some cases (e.g., linear, exponential, and inverse models for y_1). In other cases, the Beaumont and NP methods give seriously biased estimates of the total (e.g., linear and inverse models for y_3 and exponential model for y_3 with any NP method). For the NP methods, the presence of influential observations with unusually large weights would lead the NP weights models to "pull back" the regression line, resulting in smaller regression coefficients and thus more smoothed weights. This over-smoothing also increases in both models as the relationship between the weight and variable of interest is stronger (shown by the larger biases in y_3).

Table 1.3. Beaumont Simulation Results

Percentage Relative Bias Measures

	My Results			**Beaumont Results***		
Design-based Estimators	y_1	y_2	y_3	y_1	y_2	y_3
HT	0.1	-0.4	0.2	0.0	0.2	-0.1
Hajék	0.8	0.5	-0.7			

	My Results									**Beaumont Results***		
Model-based Estimators	*Linear Model*			*Exponential Model*			*Inverse Model*			*Exponential Model*		
	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
SHT_U	1.5	14.2	70.6	1.5	14.2	70.6	-47.6	-41.1	-12.1	-0.8	12.1	73.3
SHT_1	1.6	13.9	69.2	-8.0	-5.1	-7.4	-1.2	-4.5	-11.5	-9.1	-5.7	8.3
SHT_5	-0.3	11.2	66.3	-5.3	-4.5	-0.1	-18.8	-18.8	-13.3	-6.1	-4.4	0.2
MM	-7.0	-2.6	-11.1	-8.9	-6.0	-9.4	-0.5	-3.5	-11.0			
LMS	-8.6	-4.2	-15.2	-8.9	-5.3	-12.8	-1.8	-2.0	-8.0			
LTS	-8.2	-3.2	-16.0	-8.7	-5.2	-13.0	-3.2	-3.4	-8.1			

Variance Ratio Measures

	My Results			**Beaumont Results***		
Design-based Estimators	y_1	y_2	y_3	y_1	y_2	y_3
HT	1.00	1.00	1.00	1.00	1.00	1.00
Hajék	0.67	0.69	2.77			

	My Results									**Beaumont Results***		
Model-based Estimators	*Linear Model*			*Exponential Model*			*Inverse Model*			*Exponential Model*		
	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
SHT_U	0.46	0.59	9.52	0.46	0.59	9.52	0.07	0.06	0.62	0.45	1.43	4.33
SHT_1	0.48	0.60	9.22	0.39	0.41	1.18	2.10	1.84	0.46	0.77	0.59	1.69
SHT_5	0.52	0.63	9.17	0.51	0.52	0.79	0.93	0.55	0.36	0.64	0.58	0.84
MM	863.58	58.00	4.24	0.36	0.35	1.64	2.30	2.11	0.58			
LMS	872.22	59.99	5.07	0.45	0.51	2.81	2.81	2.76	3.08			
LTS	927.79	62.87	2.14	0.44	0.48	2.43	2.62	2.52	3.22			

* Beaumont used 10,000 Samford samples of size n=500 drawn from a population of 50,000.

As shown in Table 1.3, the model-based estimators, despite the nonzero biases, have smaller variances than the HT estimator, with the exception of the NP methods being very inefficient under the linear weights model. They are also very inefficient, by a lesser extent, under the inverse weights model. The NP estimators are also as efficient or more so (as measured by the variance) than the Beaumont estimators under the exponential and linear weights models for the variable y_3 (though not as efficient at the HT estimator). The NP estimators are very inefficient under the linear weights model for estimating the totals for y_1 and y_2; the population plot in Figure 1.16 shows that there is a very weak relationship between these variables to z_i, which is inversely related to the HT weights. The Beaumont estimators' efficiency is not impacted by this.

Table 1.4 contains the relative RMSE's, a more comprehensive summary measure, from my simulation (Beaumont did not report this result).

Table 1.4. Relative Root Mean Square Errors, My Simulation Results

Design-based Estimators	y_1			y_2			y_3		
HT	1.00			1.00			1.00		
Hajék	0.67			0.70			2.84		
Model-based Estimators	*Linear Model*			*Exponential Model*			*Inverse Model*		
	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
SHT_U	0.47	0.84	0.47	0.84	80.83	80.83	2.36	2.18	2.14
SHT_1	0.48	0.84	0.46	0.44	1.97	77.57	2.10	1.87	2.33
SHT_5	0.52	0.79	0.54	0.55	0.79	72.06	0.92	0.99	2.89
MM	863.41	58.00	0.44	0.40	3.07	6.23	2.29	2.12	2.54
LMS	872.07	60.00	0.53	0.55	5.44	8.78	2.82	2.76	4.11
LTS	927.62	62.87	0.52	0.51	5.16	6.09	2.63	2.53	4.29

The RMSE's provide more insight into the magnitude of the total errors of these estimators, both bias and variance. While the Beaumont estimators are more biased under the linear model than the NP estimators, the variances of the NP estimators drive their large relative RMSE's in Table 1.4 for y_1 and y_2. Since the relationship of the weights and y_1 and y_2 is weaker, the Exponential model produced totals with lower RMSE's. However, the inverse weights model is inefficient for all estimators, with the exception of SHT_5 for y_1 and y_2. These inefficiencies drove the RMSE's. Here also the NP method results are comparable to the Beaumont estimators. For the variable y_3, which has a stronger relationship to the weights, the only estimator with lower RMSE than the HT estimator is the fifth-order polynomial regression-based SHT_5 estimator under the Exponential weights model, but the RMSE under the Beaumont estimators and linear model are very high for y_3. Generally, the SHT_5 had the lowest RMSE's across the variables and weights model, with the exception of y_3 under the exponential model.

Figures 1.7 through 1.9 at the end of this section show the boxplots of the 2,000 estimated totals for each variable. In each plot, the true population total is represented by a horizontal line. It is clear how the bias shifts the distribution of the totals, while inefficiency "stretches" them out. From Figure 1.7, it is obvious that the inefficiency in the MM, LMS, and LTS estimators is caused by a few outlying totals (fewer than the HT estimator). Further investigation showed that in these particular samples, the iterative algorithms did not converge within 50 iterations; further research could establish the conditions under which this happened and produce some guidelines when applying the NP methods in practice. If this occurs within one sample in practice, an unreasonably large total can be identified by comparing the NP estimator to the HT estimator with unadjusted weights. The same problem caused a few outlying totals, and thus larger sampling variance, for these estimators for y_2 in Figure 1.8. This did not occur for the MM estimator for y_3, but did occur for the LMS and LTS estimators. For y_2, the NP methods with the linear weights model over-smoothes the weights. While the bias is reduced in the exponential model, a few outlying totals contributed to the large Table 1.3 variance ratios for the LMS and LTS estimators. The inverse weights model again is most inefficient. For y_3, the estimator SHT_U was positively biased. While the NP methods are negatively biased under both models, their range of totals is much smaller. Again, a few outlying totals lead to the increased variance ratios for the LMS and LTS estimators. Thus, while it seems that the NP estimators generally had lower bias than the Beaumont estimators, a few outlying totals made these estimators inefficient across the simulation samples. In addition, biases in the totals were produced in particular cases, (e.g., using a linear model). In these cases, a "safer" design-based estimatoris the Hajék. This estimator is not only easier to produce, but also here performed as well or better in terms of the RMSE of the estimated totals across the variables.

Figure 1.7. Side-by-Side Boxplots of Bias in Estimated Totals: y_1

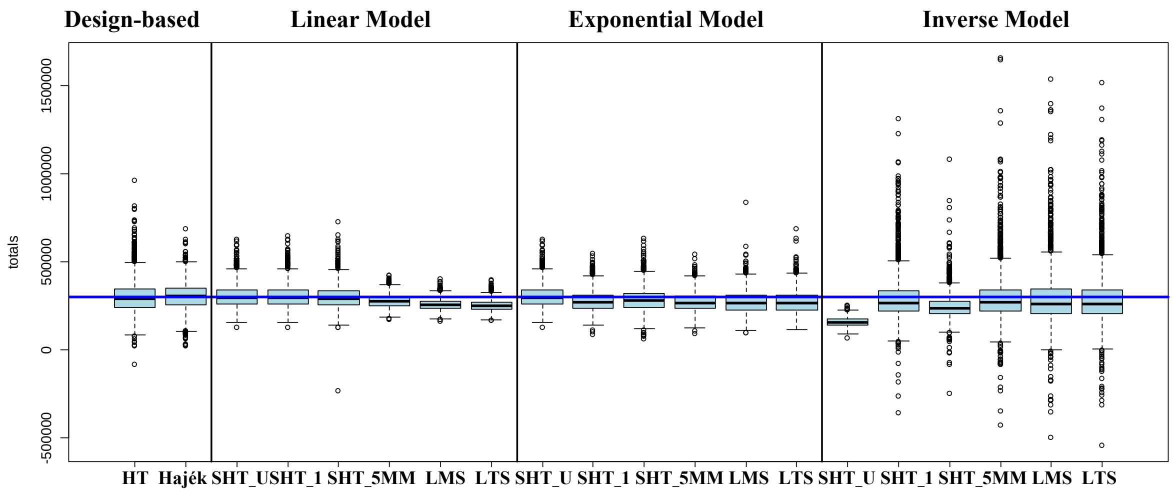


Figure 1.8. Side-by-Side Boxplots of Bias in Estimated Totals: y_2

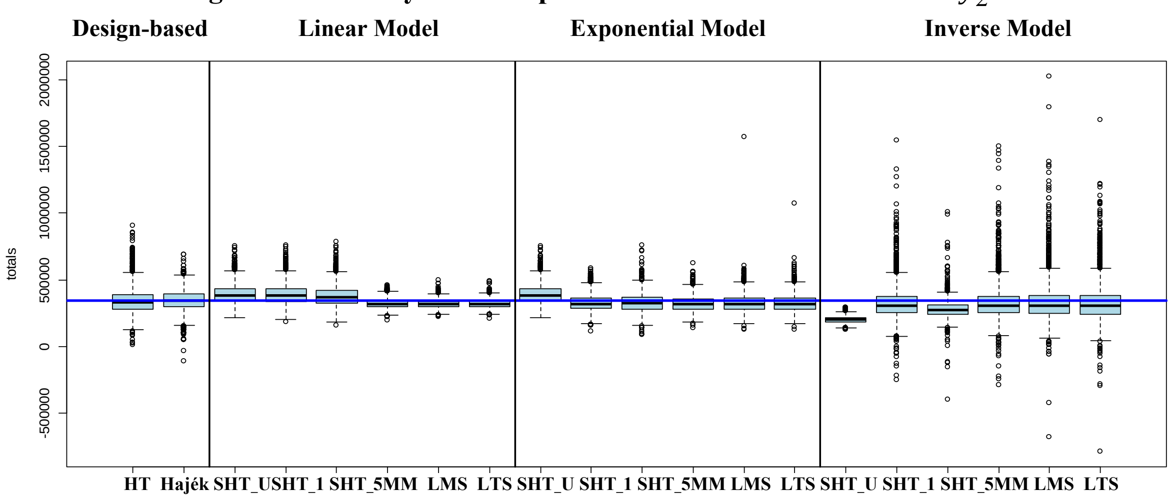
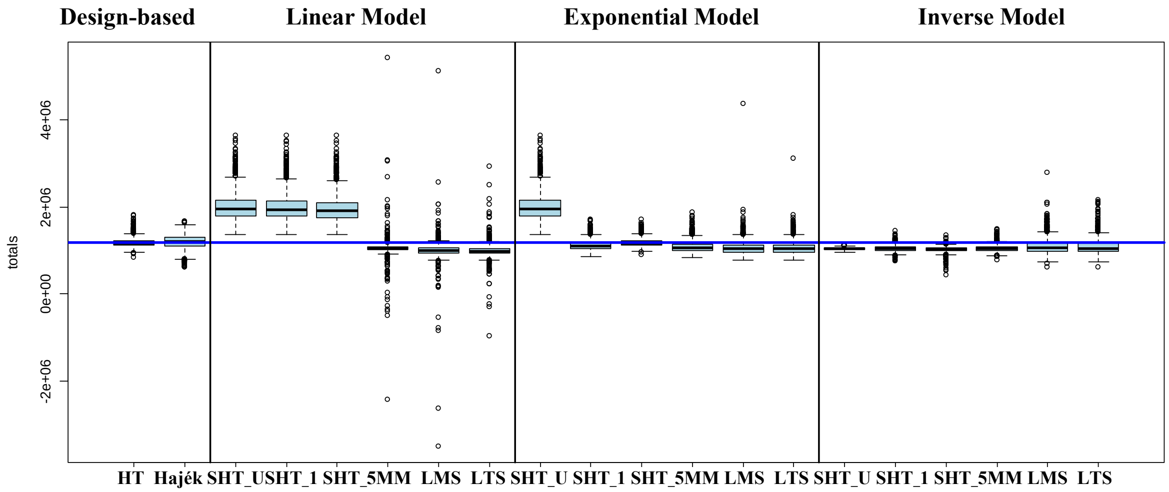


Figure 1.9. Side-by-Side Boxplots of Bias in Estimated Totals: y_3



1.4.2. The Impact of Model Specification: Accounting for Zero-Inflated Variables

This simulation study aims to gain an initial understanding of how sensitive the generalized design-based models are to zero-inflated survey variables. This kind of variable is considered since it differs from the continuous variables produced in Sec. 1.4.1 and can illustrate how particular weights models fit to different kinds of survey variables, i.e., continuous or categorical. Since the focus here is on the functional form of the weights models, and in Sec. 1.4.1 it was found that the NP methods did not produce superior estimators of totals over the Beaumont estimators when the weights model was misspecified, the NP methods are not examined here.

I select 1,000 $\pi ps(z_i)$ samples of size $n = 500$ from the pseudopopulation $\{z_{1i}, y_{1i}, y_{2i}, y_{3i}\}$ of size $N = 10{,}000$, where $y_{1i} = \beta_1 z_i + \varepsilon_{1i}$, $\beta_1 = 10$, $z_{1i} \sim Gamma(3,4)$, y_{2i} is a zero-inflated Exponential variable:

$$\Pr(y_{2i} = 0) = \phi_1 + (1-\phi_1)\lambda$$
$$\Pr(y_{2i} > 0) = (1-\phi_1)\lambda e^{-\lambda y_{2i}},$$

where $\phi_1 = 0.1$ and $\lambda = \frac{1}{N}\sum_{i \in U} z_i$, and y_{3i} is a zero-inflated Poisson variable:

$$\Pr\left(y_{3i} = 0\right) = \phi_2 + \left(1 - \phi_2\right)e^{-\lambda}$$
$$\Pr\left(y_{3i} > 0\right) = \left(1 - \phi_2\right)\frac{e^{-\lambda}\lambda^{y_{3i}}}{y_{3i}!},$$

where $\phi_2 = 0.4$ and $\lambda = \frac{1}{N}\sum_{i \in U} z_i$. That is, ten percent of the y_2-values and forty percent of the y_3-values in the population are zero. Ninety and sixty percent follow the specified Exponential and Poisson models, respectively. The errors follow the same, but separate, Normal distributions: $\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i} \overset{\text{ind}}{\sim} N\left(0, \sqrt{50}\right)$. Figure 1.10 shows the pseudopopulation scatterplots and histograms. In Figure 1.10, we can see how the variables y_2 and y_3 have a concentrated mass of values at zero, then the nonzero values follow the separate distribution. To see the potential relationship between y_1, y_2, y_3 and the HT weights produced from a $pp(z)$ sample of size 500, I calculated the associated probabilities of selection for all 10,000 population units.

Figure 1.10. Simulated Zero-Inflated Population Values and Loess Lines

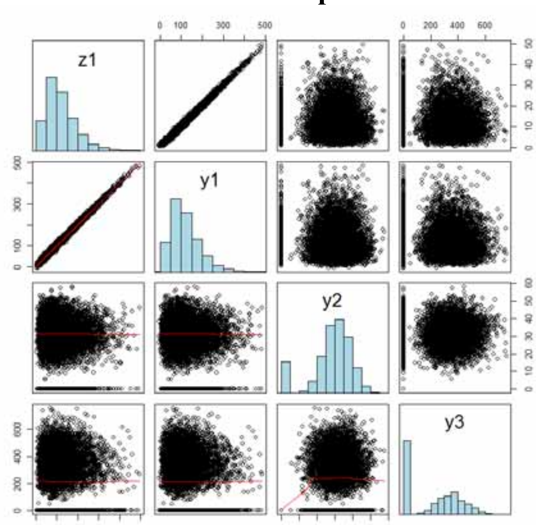
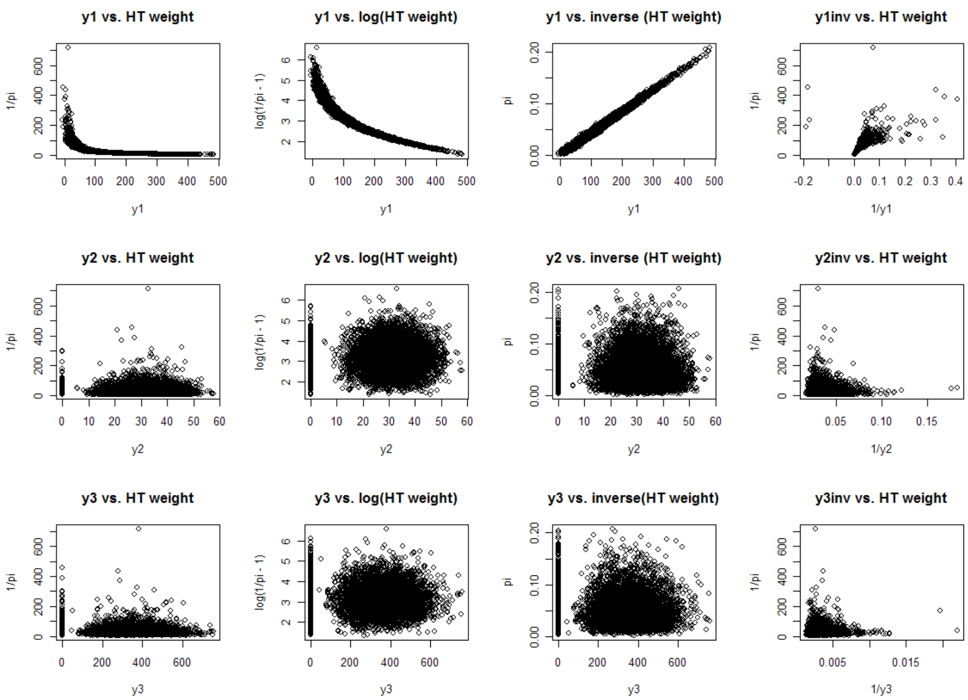


Figure 1.11 shows these HT weights and various transformations of them vs. y_1, y_2, y_3 and $y_1^{-1}, y_2^{-1}, y_3^{-1}$:

Figure 1.11. Transformations of Simulated Population Values vs. Transformations of Population HT Weights



The variables y_2 and y_3 have a weak relationship with the variable z in Figure 1.10 and there is a very weak relationship with these variables to the HT weights in Figure 1.11. Also, the strongest relationship is between the inverse of the HT weights vs. y_1. From this, I include both "incorrectly" specified models and a "correct" model, as well as one that accounts for the zeroes in the y_2, y_3 values.

The predicted weights are produced when ignoring the zero-inflation and using the linear (M1), exponential (M2), and inverse (M3) weights model, and fitting the inverse weights model to just the nonzero values of y_2, y_3. This leads to four weight models:

- M1-Linear: $w_i = \beta_0 + \beta_1 H_{1i} + \beta_2 H_{2i} + \beta_3 H_{3i}$, $H_{ki} = y_{ki}^{-1}, k = 1,2,3$, $v_i = y_{1i}$

- M2-Exponential: $\log(w_i - 1) = \beta_0 + \beta_1 H_{1i} + \beta_2 H_{2i} + \beta_3 H_{3i}, H_{ki} = y_{ki}, v_i = y_{1i}$
- M3-Inverse: $w_i^{-1} = \beta_0 + \beta_1 H_{1i} + \beta_2 H_{2i} + \beta_3 H_{3i}$, $H_{ki} = y_{ki}$, $v_i = y_{1i}$
- M4-Z-inverse: model M3 fit to the data with $y_2 > 0, y_3 > 0$,

where all the β-coefficients were estimated using WLS (and the weights $1/|y_1|$), the M2 and M3 coefficients are estimated from the non-zero values. The zero-values of y_2, y_3 had to be treated differently for different weights models: for the linear model, zero values were replaced with a value of one, while zeroes retain their value for the exponential and inverse weights models. Note that I include an intercept $\hat{\beta}_0$ while the true population model does not have an intercept between the y-variables and z. The intercept is included to capture any undue relationship caused by sampling or simulation error. The model-independent HT estimator is also included for comparison.

Figure 1.12 shows boxplots of the weights produced under the models, while plots of the weights before and after smoothing for one particular sample are in Figure 1.13.

Figure 1.12. One-Sample Boxplot Distributions of Weights Under Different Models, One-Sample Example

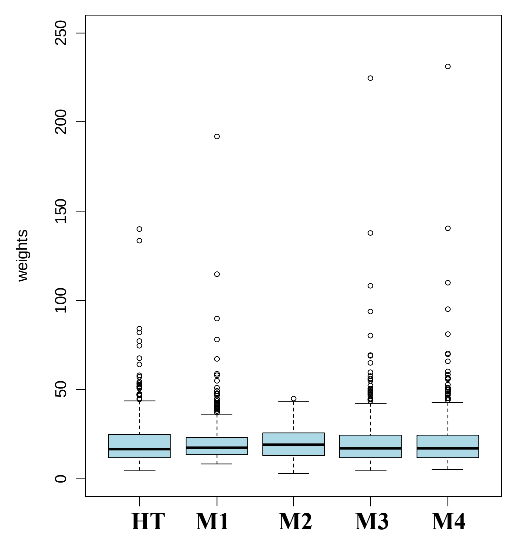
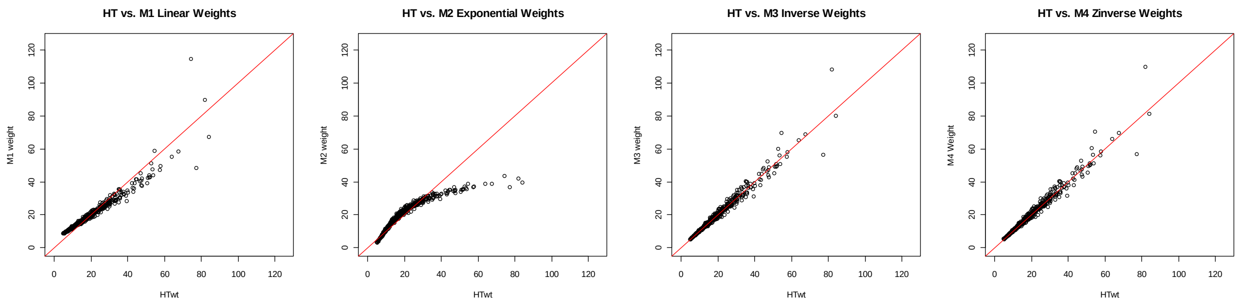


Figure 1.13. One-Sample Weights Before and After Smoothing, One-Sample Example



Here we see that the Exponential weights model leads to the most severe smoothing of the HT weights in this example; smaller weights are actually increased more and larger weights are severely decreased. For this particular sample, there is also negligible difference between the weights produced using all (in M3) vs. the nonzero (M4) y_2 and y_3 values.

To evaluate the totals estimated under the different methods, I use the percentage relative bias, the ratio of the empirical variance to that of the HT estimator, and the empirical mean square error (RMSE) relative to the HT estimator's RMSE across the 2,000 simulation samples:

- $RelBias\left(\hat{T}_k\right) = 100\left(2000T_k\right)^{-1}\sum_{b=1}^{2000}\left(\hat{T}_{bk} - T_k\right)$,
- $VarRatio\left(\hat{T}_k\right) = \dfrac{2000^{-1}\sum_{b=1}^{2000}\left(\hat{T}_{bk} - \bar{\hat{T}}_k\right)^2}{2000^{-1}\sum_{b=1}^{2000}\left(\hat{T}_{bk}^{HT} - \bar{\hat{T}}_k^{HT}\right)^2}$,
- $RelRMSE\left(\hat{T}_k\right) = \dfrac{\sqrt{\sum_{b=1}^{2000}\left(\hat{T}_{bk} - T_k\right)^2}}{\sqrt{\sum_{b=1}^{2000}\left(\hat{T}_{bk}^{HT} - T_k\right)^2}}$,

where $\hat{T}_{bk}$ is an estimate of the true total $T_k = \sum_{i \in U} y_{ki}$ for variable y_k in simulation sample b ($b = 1, \ldots, 2000$), $k = 1, 2, 3$ is the variable index, $\bar{\hat{T}}_k = 2000^{-1}\sum_{b=1}^{2000}\hat{T}_{bk}$ and

$\bar{\hat{T}}_k^{HT} = 2000^{-1}\sum_{b=1}^{2000}\hat{T}_{bk}^{HT}$.

In addition to the four model-based totals, evaluation measures were computed for the HT and Hajék estimators, where $\hat{T}_{HJ} = N\hat{T}_{HT}/\hat{N}$, where $\hat{N} = \sum_{i\in s} w_i$. The Hajék estimator is included since the relationship between z and y_2, y_3 is weaker (Valliant *et al*., 2000). The Hajék estimator is also recommended in the design-based literature as providing some protection against the effects of extreme weights. Results are shown in Table 1.5 on the following page.

The Table 1.5 results for y_1 suggest that using a zero-inflated adjustment model produces less efficient estimates for non-zero inflated variables (with model M3 being an exception). The bias in estimating y_1 is largest in the Beaumont estimator with an incorrectly specified weights model (the M2 Exponential). For this variable, despite lower biases, it also appears that the proposed Z-inverse estimator can be less efficient than the HT estimator and Beaumont estimators. Among the Beaumont estimators, the one using the inverse weights model was generally the least biased. The Beaumont estimator with the exponential model was the most efficient for y_2 and y_3 among the Beaumont estimators, but worse than the Hajék. These large variances produced larger RMSE's than the Beaumont models for all three variables, but in general though, the zero-inflated models seem to produce comparable results. If the M4-Z-Inverse estimator's efficiency could be improved, then the RMSE results would be lower than the alternatives. As expected from the relationship shown in Figure 1.9, the Hajék estimator performed well for y_2 and y_3, but not y_1.

Table 1.5. Zero-Inflated Variables Simulation Results

	Bias			Variance Ratio			RelRMSE		
Estimator	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
<u>Design-based</u>:									
HT	0.01	-0.06	-0.09	1.00	1.00	1.00	1.00	1.00	1.00
Hajék	0.18	0.03	0.03	81.70	0.21	0.43	82.00	0.21	0.43
<u>Beaumont</u>*:									
M1-Linear	4.97	-0.08	-0.15	335.23	0.99	0.99	559.87	0.99	0.99
M2-Exponential	3.08	-3.08	-3.18	3.03	0.67	0.63	3.03	1.45	1.12
M3-Inverse	-0.24	0.61	-0.34	0.84	1.17	1.16	1.34	1.20	1.17
M4-Z-inverse	0.16	0.62	0.21	32.02	2.71	1.90	32.23	2.92	1.97

* Models: M1 Linear: $w_i = \beta_0 + \beta_1 H_{1i} + \beta_2 H_{2i} + \beta_3 H_{3i}$, $H_i = y_i^{-1}$, $v_i = |y_{1i}|$;

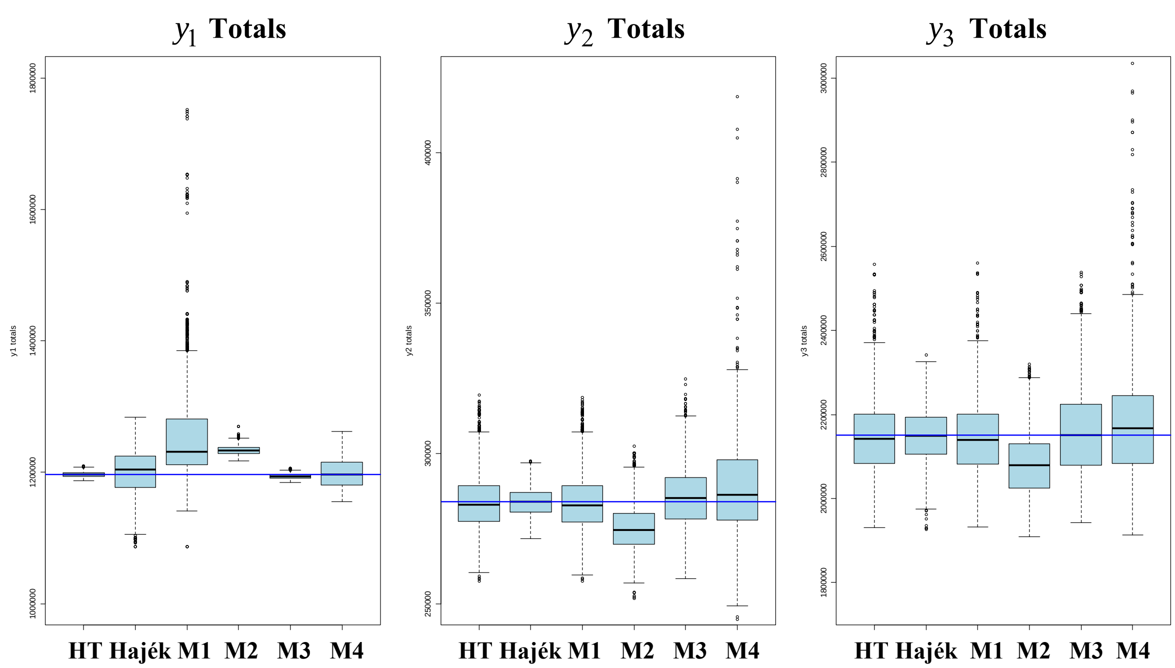
M2 Exponential: $\log(w_i - 1) = \beta_0 + \beta_1 H_{1i} + \beta_2 H_{2i} + \beta_3 H_{3i}$, $H_i = y_i$, $v_i = |y_{1i}|$;

M3 Inverse:, $w_i^{-1} = \beta_0 + \beta_1 H_{1i} + \beta_2 H_{2i} + \beta_3 H_{3i}$ $H_i = y_i$, $v_i = |y_{1i}|$;

M4 Z-inverse: model M3 fit to the data with $y_2 > 0, y_3 > 0$.

This evaluation study also indicates that the generalized design-based totals can be sensitive to the weights model fit to different types of survey response variables. We can easily see this in looking at the sampling distribution of the estimated totals across the simulation samples; Figures 1.14 shows the boxplots for y_1, y_2, y_3 (note a difference in scale). The true population total is shown as a horizontal line over each histogram.

Figure 1.14. Boxplot Sampling Distributions of y_1, y_2, y_3 Totals



The plots for all three variables in Figure 1.14 show the same general patterns: the linear model M1 produces totals with the largest sampling variance and the exponential model is not centered around the true population total (illustrating the bias). The HT estimator and M3 Inverse model have the sampling distributions centered close to the true population total, with the smallest sampling variance. The Z-inverse M4 estimator is the least biased, but its inefficiency is shown in the plots. For y_2 and y_3, the sampling variance is caused by a few outlying totals; identifying these cases is a first step towards improving the efficiency in this estimator for these variables.

To see how similar or dissimilar the totals estimated from each simulation sample are, Figures 1.15 through 1.17 on the following page show the pairwise scatterplots of the six alternative totals for the 2,000 samples for all three variables (note a scale difference). These show how the alternative Beaumont models adjust the HT weights for y_1 such that there is almost no discernible pattern between the totals before and after the adjustments, but the totals are much closer for y_2 and y_3.

Figure 1.15. Pairwise Scatterplots of y_1 Totals

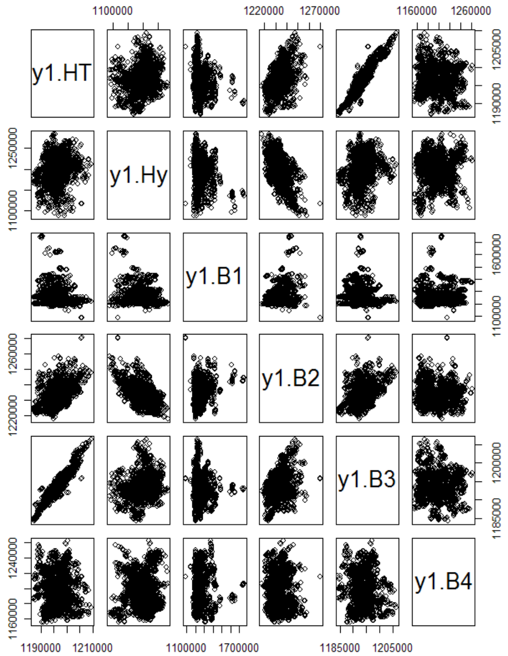


Figure 1.16. Pairwise Scatterplots of y_2 Totals

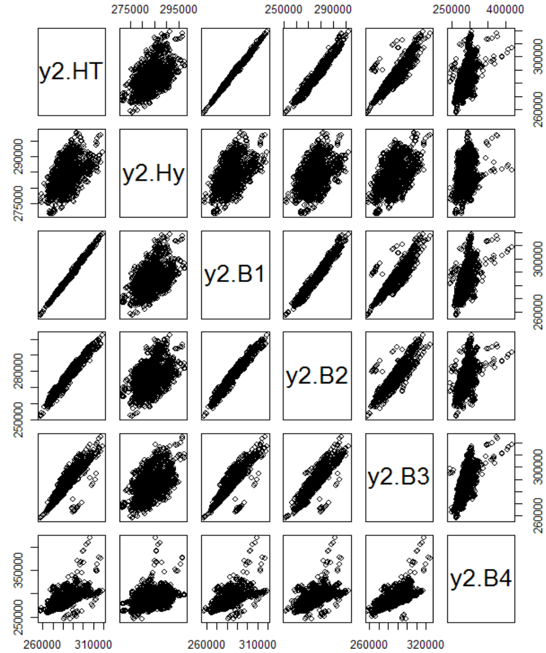
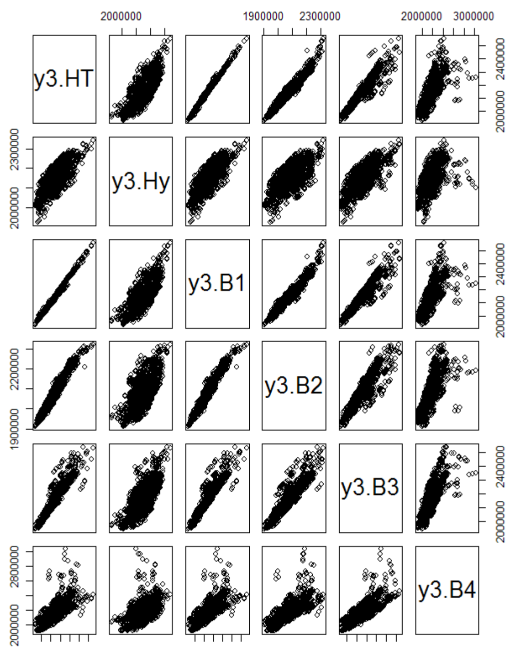


Figure 1.17. Pairwise Scatterplots of y_3 Totals

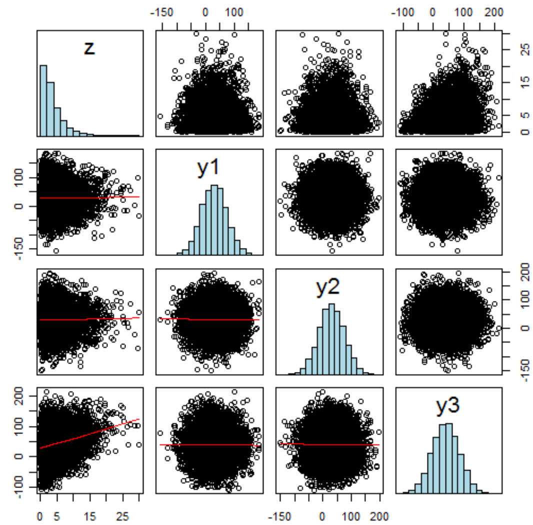


1.4.3. Performance of Beaumont and NP Estimators with Outlying Weights

In this simulation, a population similar to Beaumont's in Sec. 1.4.1 was created. However, here the variable z was created using two chi-square distributions, $\chi^2(23)$ for

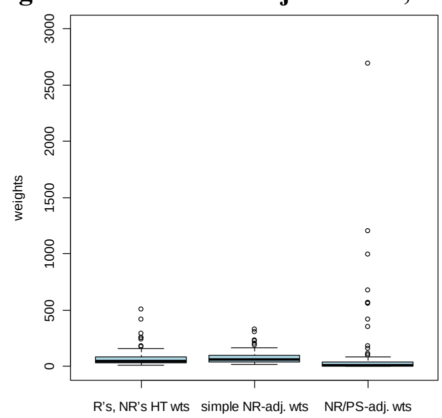
9,500 units and $\chi^2(40)$ for 500. This was done to intentionally vary the probabilities of selection, and thus the HT weights, further than those obtained in the Sec. 1.4.1 study. The variables y_1, y_2, y_3 were then created similar to how they were in Sec. 1.4.1, i.e., $y_{1i} = 30 + \varepsilon_{1i}$, $y_{2i} = 30 + 0.1498 z_i + \varepsilon_{2i}$, but here $\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i} \overset{\text{ind}}{\sim} N(0, 200)$. Figure 1.18 shows the pseudopopulation plot of these revised variables of interest:

Figure 1.18. Revised Beaumont Simulation Population and Loess Lines



Again, two thousand Sampford's $pps(z)$ samples of size 100 were drawn. The Sec. 1.4.1 simulation was also extended by perturbing some of the largest HT weights to mimic nonresponse adjustments that produce more varied weights. To do this, first all sample units with the largest HT weights (defined by weights exceeding the 95$^{\text{th}}$ percentile in the empirical weight distribution) were regarded as nonrespondents, and thus dropped from each sample. This is done to mimic methods used in establishment studies, where nonrespondents with the highest probabilities of selection are contacted more extensively to collect their responses, while smaller unit nonrespondents are adjusted with weights or imputation. Then a propensity model using the variable y_3 (the variable most correlated with z, thus mimicking a variable that is correlated with the response propensity) was used to adjust the HT weights. These nonresponse weights were then rescaled using an overall adjustment to force the weights to sum to N (see Ex. 0.3), resulting in a more skewed weight distribution. Figure 1.19 shows boxplots of these perturbed weights (labeled "NR/PS-adj. wts") compared to the original HT weights for all sample units ("R's, NR's HT wts") and the HT weights after a simple overall response rate adjustment ("simple NR-adj. wts") is applied.

Figure 1.19. HT Weights Before/After Adjustments, One-Sample Example



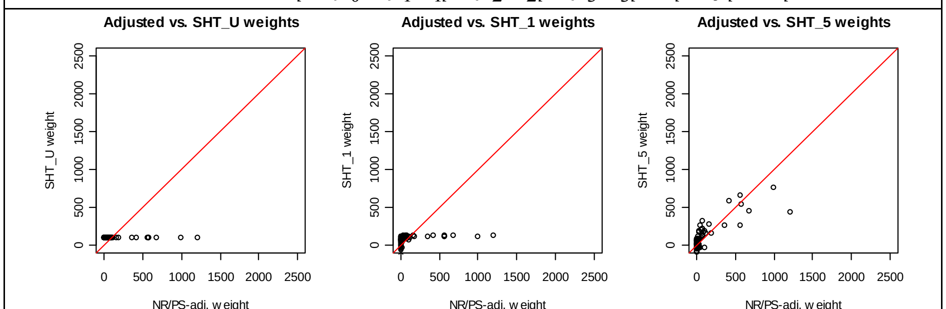
The following estimators were included:

- HT with simple NR-adjustment: $\hat{T}_k^{HT} = \sum_{i \in s} \frac{y_{ki}}{r\pi_i}, k = 1,2,3$, where $\pi_i = 100 z_i \big/ \sum_{i \in U} z_i$ is the selection probability for respondent i and $r \approx 0.77$ is the overall response rate. These weights are labeled "simple NR-adj. wts" in Figure 1.19.
- Hajék: $\hat{T}_k^{HJ} = N\hat{T}_k^{HT} \big/ \hat{N}$, where $\hat{N} = \sum_{i \in s} \left(r\pi_i\right)^{-1}$ uses the simple NR-adjusted weights, labeled "NR/PS-adj. wts" in Figure 1.19;
- "SHT_U": $\hat{T}_k^{SHT\_0} = \sum_{i \in s} \hat{w}_i y_{ki}$, where the nonresponse/PS weight w_i is predicted from an intercept-only model;
- "SHT_1": $\hat{T}_k^{SHT\_1} = \sum_{i \in s} \hat{w}_i y_{ki}$;
- "SHT_5": $\hat{T}_k^{SHT\_5} = \sum_{i \in s} \hat{w}_i y_{ki}$, where w_i is predicted from a fifth-order polynomial model with y_1, y_2, y_3 and backwards selection with the AIC measure;
- The three NP estimators $\hat{T}_k^{MM}$, $\hat{T}_k^{LMS}$, and $\hat{T}_k^{LTS}$.

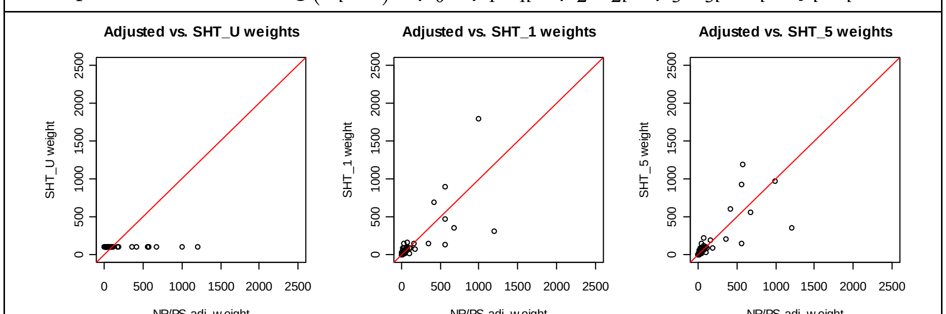
The same three weight models, linear, exponential, and inverse, were used. Here the estimators SHT_U, SHT_1, SHT_5, and the NP estimators used the weights labeled "NR/PS-adj. wts" in Figure 1.19. For the linear and inverse models, I chose $H_i = y_i^{-1}$ and $H_i = y_i$ for the exponential model. Figure 1.20 shows the one-sample examples of the various weights before and after smoothing; the NP examples are in Figure 1.21.

Figure 1.20. Beaumont Weights Before/After Smoothing, One-Sample Example

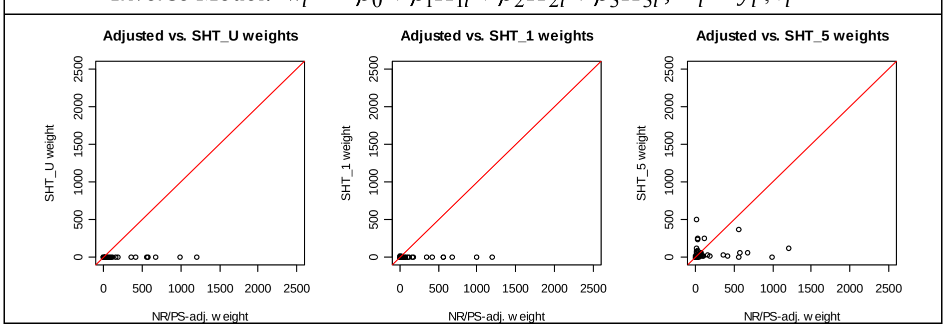
Linear Model: $w_i = \beta_0 + \beta_1 H_{1i} + \beta_2 H_{2i} + \beta_3 H_{3i}\,, H_i = y_i^{-1}\,,\ v_i = 1$



Exponential Model: $\log(w_i - 1) = \beta_0 + \beta_1 H_{1i} + \beta_2 H_{2i} + \beta_3 H_{3i}\,, H_i = y_i\,, v_i = 1$

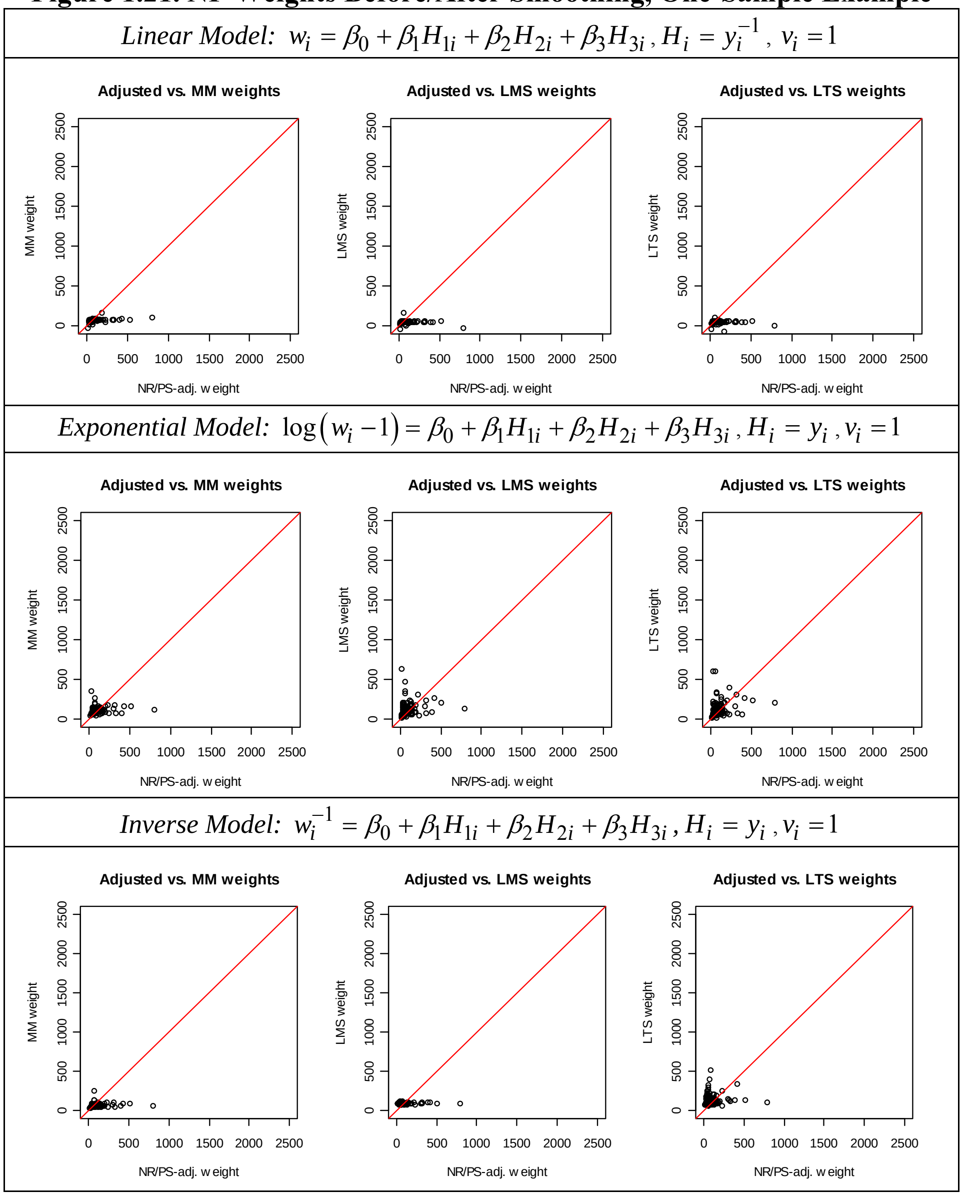


Inverse Model: $w_i^{-1} = \beta_0 + \beta_1 H_{1i} + \beta_2 H_{2i} + \beta_3 H_{3i}\,, H_i = y_i\,, v_i = 1$



Again, the SHT_U weights are smoothed to the same common weight value, as are the linear and inverse model in SHT_1. The polynomial regression SHT_5 reduces the variation in the weights, while the NP methods are similar to the Beaumont methods' weights, with a smaller range of trimmed weights. The linear model induces more trimming than the exponential model, particularly in larger HT weights.

Figure 1.21. NP Weights Before/After Smoothing, One-Sample Example



I use the same evaluation measures as in Sec. 1.4.1 to evaluate the estimated totals:

- $Relbias\left(\hat{T}_k\right) = 100\left(2000T_k\right)^{-1} \sum_{b=1}^{2000}\left(\hat{T}_{bk} - T_k\right),$

- $VarRatio\left(\hat{T}_k\right) = \dfrac{2000^{-1}\sum_{b=1}^{2000}\left(\hat{T}_{bk} - \hat{\bar{T}}_{bk}\right)^2}{2000^{-1}\sum_{b=1}^{2000}\left(\hat{T}_{bk}^{HT} - \hat{\bar{T}}_{bk}^{HT}\right)^2}$,
- $RelRMSE\left(\hat{T}_k\right) = \dfrac{\sqrt{\sum_{b=1}^{2000}\left(\hat{T}_{bk} - T_k\right)^2}}{\sqrt{\sum_{b=1}^{2000}\left(\hat{T}_{bk}^{HT} - T_k\right)^2}}$,

where $\hat{T}_{bk}$ is an estimate of the true total $T_k = \sum_{i \in U} y_{ki}$ for variable y_k on simulation sample b ($b = 1, \ldots, 2000$), $\hat{\bar{T}}_{bk} = 2000^{-1}\sum_{b=1}^{2000}\hat{T}_{bk}$, and $\hat{\bar{T}}_{bk}^{HT} = 2000^{-1}\sum_{b=1}^{2000}\hat{T}_{bk}^{HT}$.

These measures are shown in Table 1.6. Some general comments can be made from these results. The HT estimator with the simple nonresponse adjustment and Hajék estimators are nearly unbiased across the 2,000 samples. Both the Beaumont and proposed NP estimators have nonzero biases. In other cases, the model-based methods give seriously biased estimates of the total (e.g., linear and inverse models). The unusually large weights are less influential in the NP weights models, which results in generally lower biases in the estimated totals, particularly with the exponential model. However, the NP methods are still susceptible to "pulling back" the regression line in the present of influential observations, resulting in smaller regression coefficients and thus smaller smoothed weights. This increases as the relationship between the weight and variable of interest is stronger (shown by large negative biases, particularly in the linear model). None of the model-based alternatives produce unbiased estimates of the total for y_3, which was the variable used to produce the nonresponse weights.

Again, the model-based estimators can have smaller variances than the HT estimator, with the exception of the NP methods being very inefficient under the linear weights model. All of the model-based methods are most inefficient for the variable y_3.

The MM estimator is as efficient or more so (as measured by the lower variance ratios) than the Beaumont and HT estimators under the exponential and inverse weights models for the variables y_1 and y_2. The NP estimators are very inefficient under the linear weights model for estimating the totals for y_2 and y_3.

The RMSE's describe the magnitude of the total errors of these estimators. Again, we see that the bias does not contribute as much to the MSE as the variances. Specifically, while the Beaumont estimators are generally more biased than the NP estimators, their variances drive their large relative RMSE's under the linear model. The same occurred for the Beaumont estimators under the inverse weights model. Since the relationship of the weights and y_1 and y_2 is weaker, the Exponential model produced totals with lower RMSE's. In particular, the RMSE's of the NP estimators here are lowest. However, the inverse weights model is inefficient for all estimators, with the exception of SHT_5 for y_1 and y_2. Here also the NP method results are comparable to the Beaumont estimators. For the variable y_3, which has a stronger relationship to the weights, the only estimator with lower RMSE than the HT estimator is the fifth-order polynomial regression-based SHT_5 estimator under the Exponential weights model, but the RMSE under the Beaumont estimators and linear model are very high for y_3. Generally, the inverse model had the lowest RMSE's for this variable, but all of the RMSE's across the models were at least double the RMSE of the HT estimator.

Table 1.6. Outlying Weights Simulation Results

Percentage Relative Bias Measures									
Design-based Estimators	y_1			y_2			y_3		
HT	-0.6			-0.4			-0.1		
Hajék	-0.5			0.4			0.1		
Model-based Estimators	*Linear Model*			*Exponential Model*			*Inverse Model*		
	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
SHT_U	-0.5	2.2	25.1	0.4	3.1	26.2	-79.7	-79.2	-74.7
SHT_1	-5.5	4.9	54.5	-7.7	7.8	98.4	-78.5	-77.8	-73.1
SHT_5	-8.7	6.1	92.5	-4.4	13.3	112.2	-45.6	-43.0	-9.7
MM	-34.8	-32.6	-13.2	1.1	4.0	35.2	-35.1	-33.1	-14.1
LMS	-52.5	-52.5	-38.1	11.8	15.6	50.3	1.5	4.5	35.0
LTS	-53.5	-54.5	-41.2	13.4	17.1	52.2	4.5	8.0	39.8
Variance Ratio Measures									
Design-based Estimators	y_1			y_2			y_3		
HT	1.00			1.00			1.00		
Hajék	0.15			0.14			0.15		
Model-based Estimators	*Linear Model*			*Exponential Model*			*Inverse Model*		
	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
SHT_U	0.10	0.10	0.13	0.10	0.10	0.13	1.20	1.26	2.22
SHT_1	1.89	0.38	6.52	3.66	4.13	38.97	1.36	1.35	2.28
SHT_5	2.21	0.91	11.70	3.03	2.62	32.05	4.43	4.54	15.15
MM	3.14	11.66	29.38	0.22	0.23	0.59	0.39	0.43	1.39
LMS	3.33	17.47	47.46	1.73	1.93	5.85	4.48	4.51	9.74
LTS	3.24	15.10	42.04	2.00	2.12	7.02	4.75	4.92	10.77
Relative RMSE Measures									
Design-based Estimators	y_1			y_2			y_3		
HT	1.00			1.00			1.00		
Hajék	0.15			0.14			0.34		
Model-based Estimators	*Linear Model*			*Exponential Model*			*Inverse Model*		
	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
SHT_U	0.10	0.12	2.76	0.10	0.14	3.02	28.69	28.39	25.43
SHT_1	2.02	0.48	17.82	3.92	4.39	72.23	28.04	27.50	24.45
SHT_5	2.54	1.07	45.62	3.11	3.38	79.02	13.39	12.52	12.55
MM	8.86	16.73	23.98	0.23	0.30	6.28	6.20	5.68	2.03
LMS	16.35	30.67	44.27	2.38	3.09	16.48	4.49	4.60	13.43
LTS	16.87	29.30	41.12	2.84	3.52	18.32	4.84	5.22	15.94

Figures 1.22 through 1.24 at the end of this section show the boxplots of the 2,000 estimated totals for each variable. The true population total is represented by a horizontal reference line. The bias shifts the boxplot distributions away from the reference line, while the inefficiency "stretches" them out. From Figure 1.22, like the corresponding

Figure 1.7 in Sec. 1.4.1, the inefficiency in the MM, LMS, and LTS estimators is caused by a few outlying totals. This caused larger variances for these estimators in Table 1.6. This did not occur for the MM estimator for y_3, but did occur for the LMS and LTS estimators. These are also reflected in Figures 1.23 and 1.24. For y_2, we see in Figure 1.23 that the NP methods with the linear weights model produces very biased estimates. The bias is reduced in the exponential model, although a few outlying totals contributed to the large Table 1.6 variance ratios for the LMS and LTS estimators. And the inverse weights model again is most inefficient. For y_3, the overall smoothed weight estimator SHT_U was positively biased; neither of these estimators depend on the weights model. While the NP methods are negatively biased under both models, their range of totals is much smaller. Their inter-quartile ranges are also comparable to that of the HT estimator. Again, we see that a few outlying totals lead to the increased variance ratios for the LMS and LTS estimators for the exponential weights model and the inefficiency in the inverse model.

Figure 1.22. Side-by-Side Boxplots of Bias in Estimated Totals: y_1

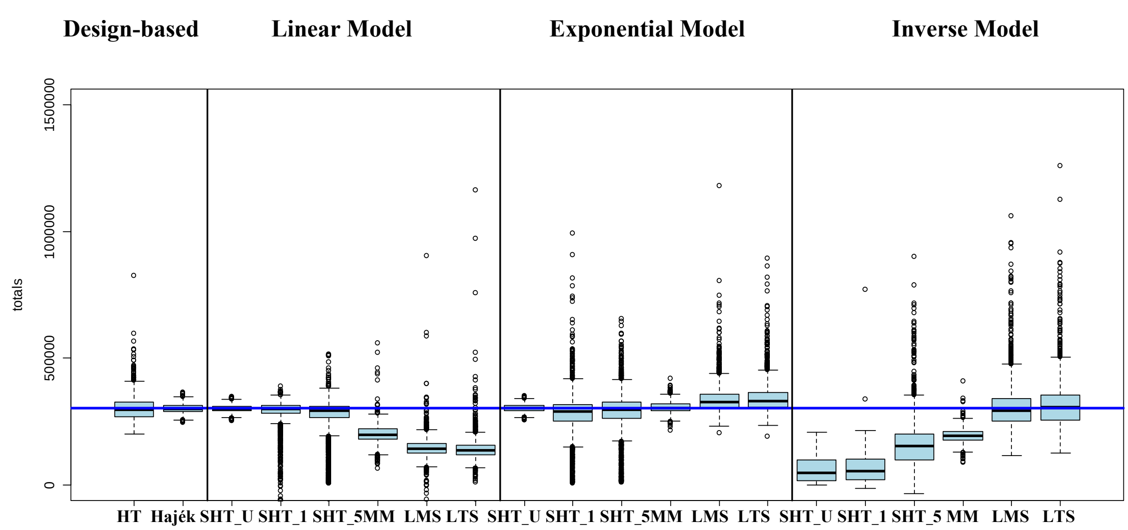


Figure 1.23. Side-by-Side Boxplots of Bias in Estimated Totals: y_2

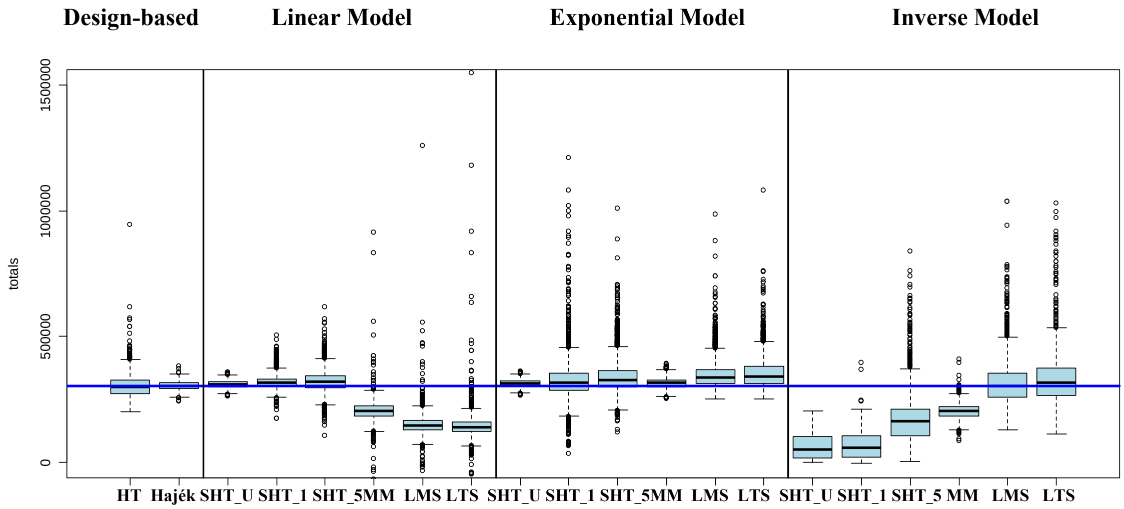
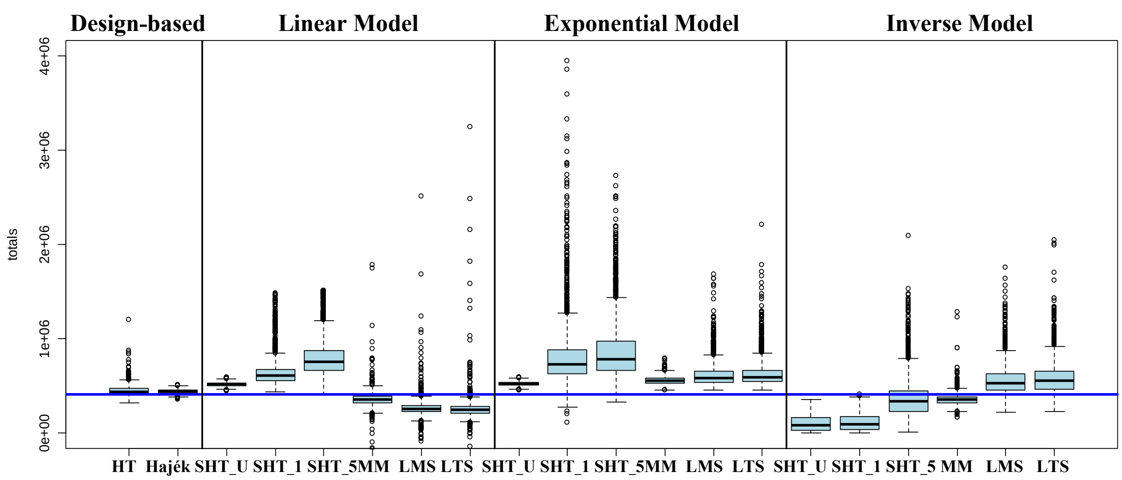


Figure 1.24. Side-by-Side Boxplots of Bias in Estimated Totals: y_3



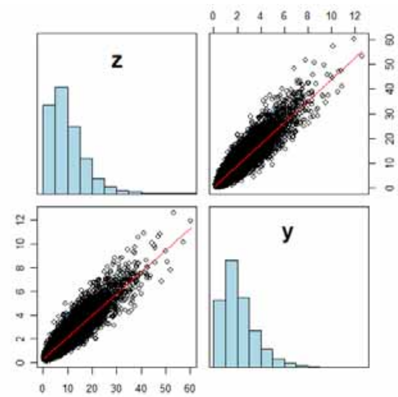
1.4.4. Evaluating Alternative Variance Estimators

The goal of this simulation is to compare the proposed robust variance estimator against Beaumont's proposed variance and MSE estimators for a simple model for the weights. To test their performance, I select 10,000 $\pi ps(z)$ samples of size $n = 100, 500$, and 1,000 from a pseudopopulation $\{z_i, y_i\}$ of size 10,000. The z_i's are assumed to be known for each unit i in the finite population. This is an unstratified version of the population described in Hansen *et al.* (1983), which follows a superpopulation with the following structure:

$$E_M(y_i \mid z_i) = \beta_0 + \beta_1 z_i, \; Var_M(y_i \mid z_i) = \sigma^2 z_i . \tag{1.38}$$

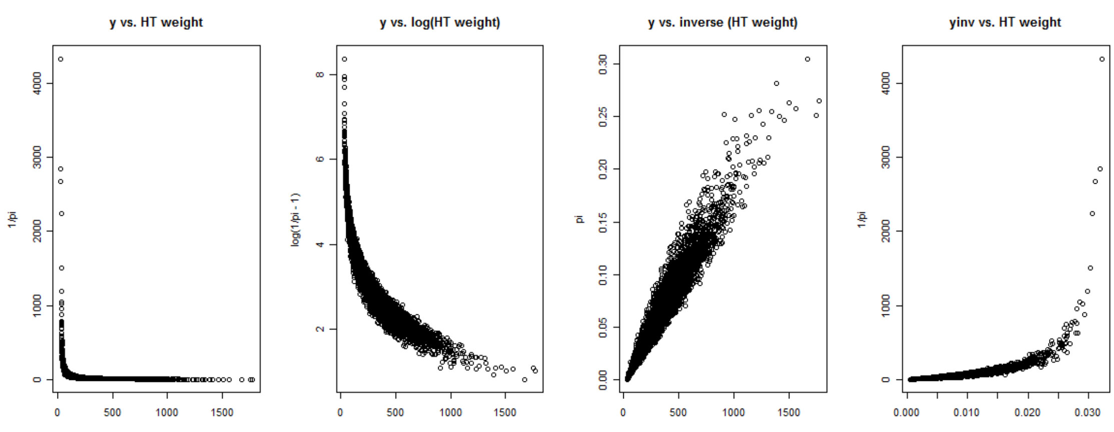
The population plot is given in Figure 1.25 below.

Figure 1.25. Simulated Population and Loess Lines for Variance Estimation Evaluation



From the plot, we see that y_i is linearly related to z_i and its variance increases with the size of z_i. I calculated the probabilities of selection for a $pp(z)$ sample of size 500. The population-level plot of the associated HT weights and various transformation of them vs. y_1, y_2, y_3 and $y_1^{-1}, y_2^{-1}, y_3^{-1}$ are shown in Figure 1.26 (note a scale difference).

Figure 1.26. Simulated Population Values vs. Population HT Weights for a *ppswor* Sample of Size 500

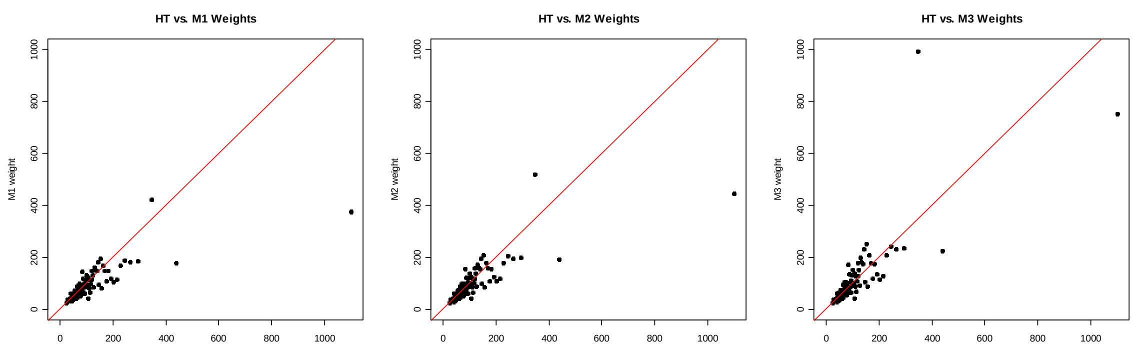


Since the variance estimator theory depends on the functional form of the weights model being correct, the "correct model" based on the population plot in Figure 1.26 involves the inverse transformation of the weight modeled as a function of the untransformed y_i-values. However, since Beaumont's variance estimator depends on the model residuals, I focus on the simple inverse weights model (linear under the inverse transformation) with $H_i = y_i$ (the "correct" model under this pps sampling) and vary the error specification:

$$\begin{aligned}
&\text{M1: } w_i^{-1} = \beta_0 + \beta_1 H_i + \varepsilon_i \,,\\
&\text{M2: } w_i^{-1} = \beta_0 + \beta_1 H_i + y_i \varepsilon_i \,,\\
&\text{M3: } w_i^{-1} = \beta_0 + \beta_1 H_i + y_i^2 \varepsilon_i \,.
\end{aligned}$$

Thus, of these three models, given the pseudopopulation model(1.38), models M1 and M3 are misspecfied in the error component and M2 is correctly specified. These three models lead to Beaumont estimators of the total $T = \sum_{i \in U} y_i$ denoted by $\hat{T}_{B1}, \hat{T}_{B2}, \hat{T}_{B3}$, respectively. To see how these models smooth the weights, Figure 1.27 shows examples of the weights before and after smoothing under each model for one particular sample.

Figure 1.27. Alternative Weights Before and After Smoothing, One-Sample Example



Five variance estimators are compared:

- *Rao-Wu (1988) bootstrap* (labeled "boot" in Figure 1.29): $var_{BOOT}\left(\hat{T}_B\right)=\frac{n}{(n-1)B}\sum_{b=1}^{B}\left(\hat{T}_b-\hat{T}_B\right)^2$, where using $\hat{T}_b$ is a realized estimate from bootstrap sample b and $\hat{T}_B$ is a Beaumont estimator. I used 100 bootstrap samples (with-replacement simple random samples of the same size as the original sample, 100, 500, or 1000);

- *sandwich Beaumont variance* ("sand" in Figure 1.29): $var_B = var_\pi\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)-\hat{\sigma}_E^2\sum_{i\in s}v_i\left(H_i-\frac{H_i}{v_i}\hat{\Omega}\right)^2$, where $\hat{\sigma}_E^2=\frac{1}{n-1}\sum_{i\in s}e_i^2$, $\hat{\Omega}=\sum_{i\in s}\left(\frac{H_i^2}{v_i}\right)^{-1}\sum_{i\in s}H_i^2$, $H_i=y_i$, and $v_i=1,y_i,y_i^2$,

- *proposed robust variance estimator* ("rob" in Figure 1.29): $var_{B*}=var_{B*}\left(\hat{T}_B\middle|\mathbf{Y}\right)=\sum_{i\in s}\hat{\psi}_i y_i^2-\hat{\sigma}_{E*}^2\left[\sum_{i\in s}v_i\left(y_i-\frac{H_i\hat{\Omega}}{v_i}\right)^2\right]$, where $\hat{\sigma}_{E*}^2=\frac{1}{n-1}\sum_{i=1}^{n}\hat{\psi}_i$ and $\hat{\psi}_i=\left(w_i-\hat{w}_i\right)^2\big/\left(1-h_{ii}\right)^2$;

- *Beaumont MSE* ("mse"): $mse\left(\hat{T}_B\middle|\mathbf{Z},\mathbf{Y}\right)=var_M\left(\hat{T}_B\middle|\mathbf{Z},\mathbf{Y}\right)+\max\left[0,\left(\hat{T}_B-\hat{T}_{HT}\right)^2-var_\pi\left[\left(\hat{T}_B-\hat{T}_{HT}\right)\middle|\mathbf{Z},\mathbf{Y}\right]\right]$, where $var_\pi\left[\left(\hat{T}_B-\hat{T}_{HT}\right)\middle|\mathbf{Z},\mathbf{Y}\right]$ is calculated using the Rao-Wu bootstrap;

- *Design MSE* ("Dmse"): $mse_D\left(\hat{T}_B\right)=\min\left[mse\left(\hat{T}_B\middle|\mathbf{Z},\mathbf{Y}\right),var_\pi\left(\hat{T}_{HT}\right)\right]$.

The design MSE I use is a more conservative estimate than Beaumont's design-MSE measure (in expression (1.20)); I use the measure here since Beaumont's MSE has more complicated terms (in particular the one involving $\tilde{T}_B$, which cannot be calculated). However, the Beaumont mse is far larger than Dmse, as will be illustrated below.

I use the following five evaluation measures to compare the above alternatives produced for $B = 10,000$ simulation samples:

- *Relative bias*: the average distance between the variance estimator $var_b\left(\hat{T}_B\right)$ and empirical variance of $\hat{T}_B$, denoted $v\left(\hat{T}_B\right) = B^{-1}\sum_{b=1}^{B}\left(\hat{T}_B - B^{-1}\sum_{b=1}^{B}\hat{T}_B\right)^2$:

$$RB\left(var\left(\hat{T}_B\right)\right) = \left[B^{-1}\sum_{b=1}^{B}\left(var_b\left(\hat{T}_B\right) - v\left(\hat{T}_B\right)\right)\right] \Big/ v\left(\hat{T}_B\right),$$

- *Empirical CV*: the standard error of the variance estimator, expressed as a percentage of the empirical variance:

$$CV\left(var\left(\hat{T}_B\right)\right) = \sqrt{B^{-1}\sum_{b=1}^{B}\left(var_b\left(\hat{T}_B\right) - B^{-1}\sum_{b=1}^{B}var_b\left(\hat{T}_B\right)\right)^2} \Big/ v\left(\hat{T}_B\right);$$

- *Empirical RelRMSE*: the mean square error of the variance estimator, expressed as a percentage of the empirical variance:

$$RelRMSE\left(var\left(\hat{T}_B\right)\right) = \sqrt{B^{-1}\sum_{b=1}^{B}\left(var_b\left(\hat{T}_B\right) - v\left(\hat{T}_B\right)\right)^2} \Big/ v\left(\hat{T}_B\right);$$

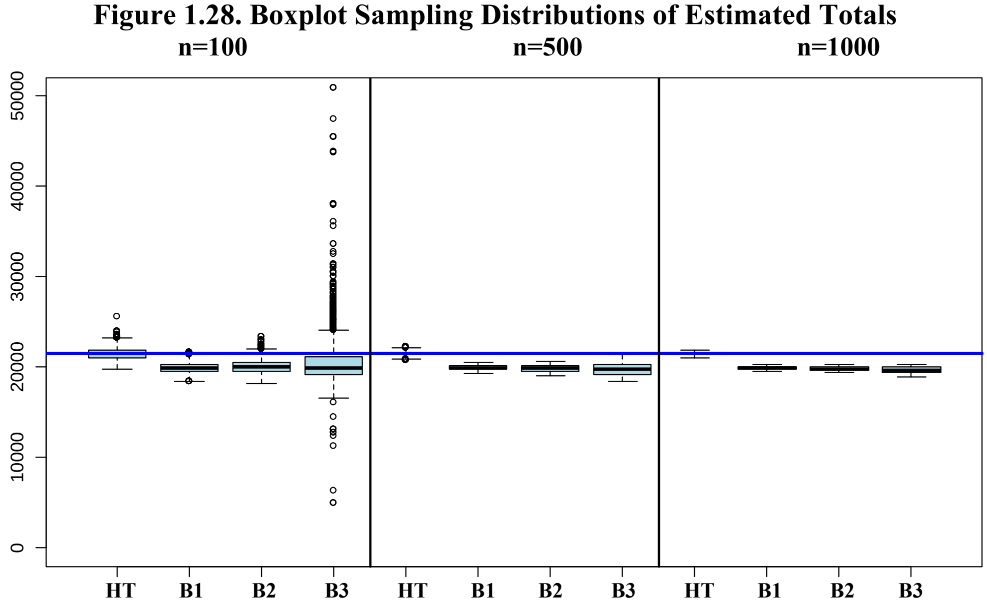
- *95% CI Coverage rate*: the percentage of the B simulated confidence intervals that contain the true population total: $\left|\hat{T}_B - T\right| \Big/ \sqrt{var\left(\hat{T}_B\right)} \le z_{\alpha/2} = 1.96$;

- *Average CI width*: the average width of the 95% confidence intervals: $\frac{2}{B}\sum_{b=1}^{B}1.96\sqrt{var\left(\hat{T}_B\right)}$.

Table 1.7 at the end of this section shows results from the 10,000 simulated samples for each sample size. Here, the most noticeable –and unexpected– result is how poorly the bootstrap variance estimator performs, particularly for $n = 100$, in terms of large biases and large empirical CVs, both of which produce larger RMSE's. For larger sample sizes, the bootstrap is more comparable to the other results. The bootstrap is also very sensitive to extreme outliers, which occurred in the $n = 500$ case. For example, the relative bias in the median of the bootstrap variance estimates for the $n = 100$ sample size and models M1, M2, M3 were 0.40%, 0.89%, and -0.75%, respectively. The corresponding bias using the means across the 10,000 samples, which are in Table 1.7, were 210.2%, 1,551.4%, and 3,095.6%.

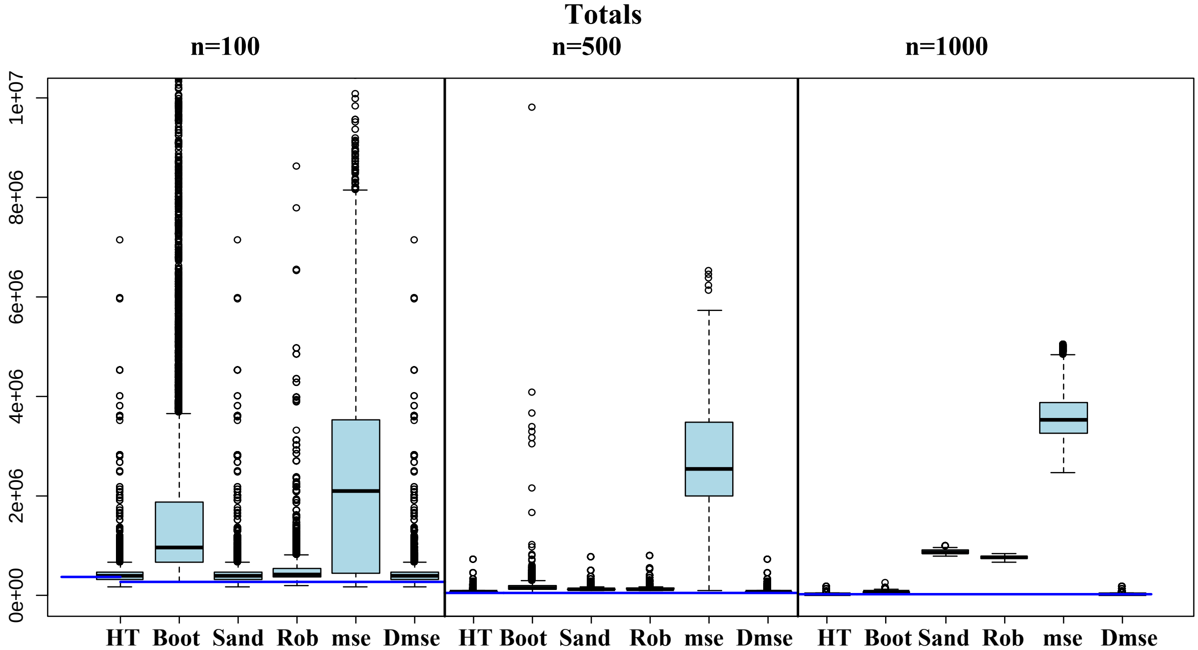
For estimation of the totals, in terms of repeated sample-based inference, the 95% confidence interval coverage rates and average CI widths indicate that the model-based alternatives perform poorly compared to the alternatives. Despite the "correct" weights model used (correct in the functional component) in producing var_B and var_{B*} most of the Table 1.7 confidence interval coverage rates were close to zero. To see why this happened, Figure 1.28 shows boxplots of the alternative totals. Here we see that the Beaumont estimators produced biased totals, which produced the poorer CI results, not the variance estimates. As the sample size increases, the variance of the Beaumont estimators decreases while the bias becomes nearly constant. In such a case, confidence interval coverage is asymptotically zero. It is also notable that six of the 10,000 simulation samples contained a negative estimate for the total produced using model M3 due to negative weights; these were omitted to avoid skewing the M3 results.

Figure 1.28. Boxplot Sampling Distributions of Estimated Totals



For the alternative variance estimators, the model-based variance estimators appear to be the most efficient, as they generally had the lowest empirical CV's, relatively low biases, and the lowest RMSEs under model M2, the correct model. As a result, the model-based alternatives appear to be better alternatives at estimating the variance. To see why the Beaumont MSE estimator performed poorly, Figure 1.29 contains boxplots of the sampling distributions of the variance estimators for the samples using weights model M2 (the M1 and M3 plots were similar). The empirical variances of each associated estimator are shown as horizontal reference line segments. All of the variance estimators are very skewed. In addition, Beaumont's MSE estimator ("mse") is very biased due to the bias in the totals. The design-MSE estimator ("Dmse") did not have this problem, since this is bounded above by the HT estimator's design-variance. Here this minimum $var_{\pi}\left(\hat{T}_{HT}\right)$ was always used.

Figure 1.29. Boxplot Sampling Distributions of Variance Estimates, HT and Beaumont M2 Totals



For the other Table 1.7 results, the RMSE's of the alternative estimators decreased as the sample sizes increased. The Beaumont variance estimators had the lowest RMSE's when using the correct model M2, but not the other models (whose RMSE's were larger than that of the HT estimator). As expected, the average width of the 95% confidence intervals decreases from $n = 100$ to $n = 500$ to $n = 1000$ for the HT variance and the model-based variance estimators var_B, var_{B*}, and $mse(\hat{T}_B)$. Generally, across the alternative weights models, model M2 with the "correct" error specification among the three alternatives had the best performance. While the results are closer than models that use an incorrect specification for the weights (Sec. 1.4.1 and 1.4.2), the variance estimators are sensitive to the choice of the model.

In evaluating the proposed robust variance estimator that uses leverage-based adjustments, the results are similar to the Beaumont sandwich variance estimator. Both variance estimators are positively biased, which is apparent as the sample size increases, with the robust variance estimator's bias being lower (particularly when the incorrect models M1 and M3 are used). This occurred since these variance estimators do not have finite population correction factors (fpc's) like the HT and bootstrap variances. Thus, while the variance of the variance decreases with the sample size – shown by lower CVs and smaller CI widths – the bias of the model-based variance estimators increase. However, the overall impact is that the RMSE's of the model-based variance estimators are lower for the smaller sample size $n = 100$ using the correct model M2. However, the bootstrap and Beaumont's design MSE, both of which have fpc's, have the smallest RMSE's for the larger sample size $n = 1000$.

.

Table 1.7. Variance Estimation Simulation Results: Relative Bias, Empirical CV, Relative RMSE, 95% CI Coverage, and Average CI Width

n=100															
	Relative Bias (%)			**Empirical CV (%)**			**RMSE**			**95% CI Coverage**			**Average CI Width**		
HT Variance	1.0			80.1			504.1			95.9			2540.2		
Model-based Estimators	**M1***	**M2****	**M3*****	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**
Bootstrap	210.2	1551.4	3095.6	$1.4x10^6$	$8.0x10^6$	$1.4x10^7$	$1.5x10^4$	$5.7x10^7$	$6.8x10^8$	31.0	67.4	86.0	3133.3	$1.2x10^4$	$8.0x10^4$
Sandwich	0.6	-0.1	-1.0	112.1	60.1	1.3	665.2	441.9	4720.1	32.5	37.6	32.8	2450.3	2540.5	2540.5
Prop. Robust	0.1	2.2	-0.5	22.7	82.1	18.7	125.6	1685.9	2590.4	20.8	89.2	98.9	2118.3	4983.4	4983.4
MSE	8.9	3.6	-0.9	430.5	376.0	45.3	5162.0	3701.6	4840.2	98.2	88.0	67.4	6319.5	5432.8	5432.8
Design MSE	0.6	-0.1	-1.0	112.1	60.1	1.3	665.2	441.9	4720.1	32.5	37.6	32.8	2540.2	2540.2	2540.2
n=500															
	Relative Bias (%)			**Empirical CV (%)**			**RMSE**			**95% CI Coverage**			**Average CI Width**		
HT Variance	0.3			42.1			129.3			96.9			1129.3		
Model-based Estimators	**M1***	**M2****	**M3*****	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**
Bootstrap	0.0	0.6	$1.8x10^6$	15.8	192.4	$1.7x10^6$	40.0	6674.3	$1.1x10^9$	0.0	2.0	53.2	988.0	1590.2	9556.1
Sandwich	0.5	0.1	-0.4	43.5	23.1	7.3	171.6	93.5	239.1	0.0	0.0	3.8	1214.5	1445.6	1943.2
Prop. Robust	0.6	0.2	-0.4	47.2	25.5	8.7	196.6	109.7	231.6	0.0	0.0	3.8	1252.2	1475.2	1965.7
MSE	38.9	22.3	6.7	750.2	850.1	663.1	$8.0x10^4$	8258.4	5905.0	100.0	100.0	84.8	6232.5	6443.2	6066.5
Design MSE	0.3	-0.3	-0.8	43.6	23.1	7.1	136.5	131.0	494.9	0.0	0.0	0.1	1129.3	1129.4	1129.7
n=1000															
	Relative Bias (%)			**Empirical CV (%)**			**RMSE**			**95% CI Coverage**			**Average CI Width**		
HT Variance	0.4			29.4			86.7			99.4			779.3		
Model-based Estimators	**M1***	**M2****	**M3*****	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**	**M1**	**M2**	**M3**
Bootstrap	0.2	0.7	9.7	18.2	35.5	$4.9x10^4$	48.3	170.1	$1.8x10^5$	0.0	0.0	11.2	690.1	1100.8	2396.7
Sandwich	8.8	17.9	19.2	46.5	74.2	84.3	1399.4	3868.8	6938.9	0.0	76.0	100.0	1944.1	3683.5	6365.2
Prop. Robust	6.4	15.3	17.1	32.0	66.8	78.8	1023.1	3320.0	6188.5	0.0	54.5	100.0	1693.3	3427.5	6028.8
MSE	107.0	75.4	42.5	11.0	114.8	989.3	$1.7x10^4$	$1.7x10^4$	$1.6x10^5$	100.0	100.0	100.0	6436.4	7394.5	9285.1
Design MSE	0.6	-0.1	-0.7	32.8	17.5	6.3	107.0	49.6	251.1	0.0	0.0	0.0	779.5	780.1	784.4

* M1: $w_i^{-1} = \beta_0 + \beta_1 y_i + \varepsilon_i$, **M2: $w_i^{-1} = \beta_0 + \beta_1 y_i + y_i \varepsilon_i$, ***M3: $w_i^{-1} = \beta_0 + \beta_1 y_i + y_i^2 \varepsilon_i$.

1.5. Discussion and Limitations

In general, the evaluation study results indicate that the model-based weight trimming method is very sensitive to specification of the weights model. It also appears very difficult to obtain a weights model that works for "all," or different kinds, of survey variables, as shown in Sec. 1.4.1, 1.4.2, and 1.4.3. This method also involves smoothing all of the HT weights; typically in a design-based setting a small set of unusually large and influential weights is trimmed, with minimal impact on the weights of non-trimmed cases. Future work could consider if this is a beneficial weight trimming method when combined with calibration to population totals to ensure better design-based properties. This is the conventional approach in SUDAAN's weighting and trimming procedures, where trimming is done prior to calibration adjustments (RTI 2010). Another extension would be to consider applying this approach to other estimators (like means or model parameter estimates), or other sample designs like cluster sampling.

Also, while it appears that the NP methods are comparable to Beaumont's, further steps would include whether or not a more formal bias correction factor needs to be developed. The gains in the NP methods were minimal in the replication of Beaumont's simulation study. However, as expected, the NP estimators produce totals with lower bias in the presence of outlying weights, such as those that mimic differential nonresponse adjustments in Sec. 1.4.3. However, in Sec. 1.4.1 the NP estimators did not consistently outperform the Beaumont estimators in all scenarios; the inefficiency in these estimators also drove the RMSE measures.

In Sec. 1.4.4, I demonstrate empirically that the Beaumont-proposed variance and MSE estimators also depend heavily on the extent to which the weights model holds.

And the model misspecification bias does not decrease as the sample size increases since there is no finite population correction factor in this variance. In Sec. 1.3.1, I propose a more robust variance estimator than Beaumont's model-based variance estimator and evaluate it empirically in Sec. 1.4.4. While the proposed variance estimator performed similar to the Beaumont and other alternative estimators, inferential results related to confidence intervals are sensitive to the bias in the estimated totals. A variance estimator that is robust to misspecification of the variance component in the weights model will not overcome this bias.

Generally, in the Beaumont replicated simulation in Sec. 1.4.1, the fifth-order polynomial term with stepwise selection had the lowest RMSE's across the variables and weights model, with the exception of y_3 under the exponential model. This suggests that some type of robust polynomial model, such as a penalized spline (e.g., Breidt *et al.* 2005), may be an appropriate extension to this method. Another potential extension is to consider modeling weights that have additional adjustments, such as nonresponse and poststratification. For example, in Sec. 1.4.3, a pps sample was selected so that weights vary, then a nonresponse adjustment is made to subsets of units. This evaluation study suggests that the NP methods with an Exponential weights model are more appropriate for modeling weights obtained using a propensity-model based weight with a PS adjustment, but more explicit models to account for other types of nonresponse adjustments can be developed.

In other (omitted) results, it was found empirically that using Beaumont's pseudopopulation with y being linearly related to z, selecting a sampled with probabilities proportional to z, and fitting an inverse weights model leads to a Beaumont estimator that is exactly equal to the HT. This is another scenario in which this method does not apply. Other situations where this method does not apply, e.g., self-weighting designs, simple random sampling, etc., were discussed earlier, at the end of Sec. 1.2.2. Last, a practical implication is that the linear and inverse weights models can potentially produce negative weights. For the linear model, small weights can be predicted as negative. However, with the inverse model, units with small π_i can be predicted to be negative and close to zero, which produces very large negative weights. If this occurs for a particular sample, it can lead to seriously biased estimates of totals if the π_i estimated from the model is not bounded below by zero.

Paper 2: Using Bayesian Priors in Model-Assisted Penalized Spline Estimators

Abstract: Penalized (*p*-) spline models have been used to produce more robust estimators of the population total. Breidt *et al.* (2005) use a calibration approach to produce a model-assisted *p*-spline estimator that shares many properties with the generalized regression estimator. I propose extending the Breidt *et al.* model using prior distributions for the unknown model parameters, such as those used in Zheng and Little (2003). In this paper, I evaluate the proposed total against conventional alternatives using linear and nonlinear data, and compare model-based, Taylor series approximation, and jackknife replication variance estimators for it. Results indicate that the proposed estimator can produce totals with lower mean square errors, but they are sensitive to the number of terms used in the model.

2.1. Introduction, Research Plan, and Research Hypotheses

2.1.1. Introduction

Alternative to the existing design-based weight trimming methods described in Sec. 1.1.1, model-based methods to adjust survey weights have also been developed. In one approach, Bayesian methods that pool or group data together have been recently proposed for weight trimming. Unlike the design-based methods, the Bayesian methods involve "smoothing over" large varying weights rather than truncating them and redistributing the value. There are two complementary approaches: "weight pooling" and "weight smoothing." While both use models that appear similar, weight pooling is the Bayesian extension of design-based trimming and weight smoothing is the Bayesian extension of classical random effect smoothing. In weight pooling models, cases are grouped into strata, some of which are collapsed into groups, and the group weight replaces the original weights. The strata here can be defined by the size of the original weight (based on the probability of selection) such that weights are closest in value within each stratum. In weight smoothing, a model that treats the group means as random effects smoothes the survey response values. The strata here can be defined by the size of the original base weight or the weight after adjustments for nonresponse and poststratification. In this approach, the survey response variable means are smoothed, not the weights, but the influence of large weights on the estimated mean and its variance is reduced under the smoothing model. In both weight trimming and weight smoothing, Bayesian models are used to average the means across all possible trimming points, which are obtained by varying the cut point for smoothing. This makes it possible for both methods to produce variable-dependent weights. In addition, these methods have been developed from a very theoretical framework, for specific inferences, and may be difficult to apply and validate in practice. These methods has also been primarily applied to non-informative sampling designs, which limits its application to complex surveys. For example, once the weights are used to determine the cutpoint stratum, they are not used further in estimation. However, the related literature has demonstrated that they are capable of producing estimators with overall lower mean square errors and that particular choices of priors can make the models robust to misspecification.

The other model-based approach, the superpopulation model approach, involves using the sample-based information and external auxiliary information to predict the total of units in the population but not in the sample. The related theory (e.g., Valliant *et. al* 2000) has shown that, when the model is correct, the Best Linear Unbiased

Predictor (BLUP) is the best estimator for the finite population total. Several options have also been proposed to adjust the BLUP estimator of the total for model misspecification using various robust regression methods including the *p*-spline estimator. Most of these estimators use a form of the BLUP plus some residual-based adjustment. There exist underlying case weights associated with these estimators, even if they are only implicitly defined. I cite several examples of these and show that the model-assisted p-spline estimator was derived from a model-based difference estimator.

Breidt *et al.* (2005), Claeskens *et al*. (2009), and Breidt and Opsomer (2000) use a local polynomial penalized (*p*)-spline calibration model to produce a model-assisted *p*-spline estimator that shares many properties with the generalized regression estimator (Särndal *et al*. 1992). They demonstrate theoretically and empirically that their *p*-spline model produces a more robust GREG estimator (robust to model misspecification) than one obtained under a linear model, without much loss of efficiency. Breidt *et al.* (2005) also proposed using survey weighted least squares and restricted maximum likelihood (REML) methods to estimate the unknown variance component parameters of the models. I propose extending their model using prior distributions for the unknown model parameters. In particular, using conventional priors (Gelman 2006) for the unknown variance components can guarantee avoidance of the negative estimates that can occur when using REML-based methods in practice.

2.1.2. Research Plan and Hypotheses

My goal is to produce an estimator for the finite population total with lower mean square errors and evaluate it against the separate weight pooling, weight smoothing, and *p*-spline model-based estimators, and some conventional design-based alternatives. To do this, I first propose a relatively simple version of the model, develop it theoretically, and evaluate it in a simulation. I also develop a design-based resampling variance estimator for this total by extending the approach in Zheng and Little (2005). I then evaluate the proposed estimator against existing alternative methods in a simulation of single-stage sample designs in Sec. 2.4. Since the proposed method is computer-intensive, I explore replicate group-based variance estimation methods, rather than "delete-a-unit" approaches. Then I use a simulation study to gauge the variance estimators' performance.

Generally, the proposed estimators can potentially have higher efficiency (lower variance) and are more robust (i.e., have lower bias) than the existing alternative methods. This produces estimators of totals with overall smaller MSE's. They are expected to have higher efficiency than the design-based estimators, but will incur a bias when the model does not hold. However, since the model-based estimators can be viewed as calibration estimators and are thus asymptotically design-unbiased when the model does not hold (Breidt *et. al* 2005), I expect that they will have a small amount of bias and not much loss in efficiency. Producing non-negative variance component estimates should also improve the MSE of the estimated totals by eliminating the possibility of egregious results.

2.2. Literature Review

This section mixes summaries of existing methods of estimation and examples of the approaches proposed in the related Bayesian, the alternative BLUP and robust superpopulation model-based estimators, and penalized spline modeling literature. After introducing the model prediction approach, I provide some simple BLUP-based examples and existing robust methods proposed to estimate a total, including the penalized spline estimator.

2.2.1. Bayesian Methods

While Bayesian inference for finite populations is not new (e.g., Basu 1971; Ericsson 1969, 1988; Ghosh and Meeden 1997; Rubin 1983, 1987; Scott 1977), Bayesian model-based approaches related to weight trimming have been recently developed. The general Bayesian inference approach first specifies a model for the population values $\mathbf{Y}$ as a function of some unknown parameter $\boldsymbol{\theta}$, denoted $p(\mathbf{Y}|\boldsymbol{\theta})$. We denote $\mathbf{X}$ as the matrix of covariates and $\mathbf{I}$ as the vector of sample inclusion indicators. To make all inferences for the finite population quantities, we use the posterior predictive distribution $p(\mathbf{y}_r|\mathbf{y}_s,\mathbf{I})$, where $\mathbf{y}_r$ are the $N-n$ non-sampled units of $\mathbf{Y}$ and $\mathbf{y}_s$ the n sampled values (Little 2004). The distribution $p(\mathbf{Y}|\boldsymbol{\theta})$ is combined with a prior for $\boldsymbol{\theta}$, denoted by $p(\boldsymbol{\theta})$, to produce the posterior distribution. From Bayes' theorem, the posterior predictive distribution of $\mathbf{y}_r$ is

$$p(\mathbf{y}_r|\mathbf{y}_s) \propto \int p(\mathbf{y}_r|\mathbf{y}_s,\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{y}_s)\, d\boldsymbol{\theta}. \tag{2.1}$$

where $p(\boldsymbol{\theta}|\mathbf{y}_s) = \frac{p(\boldsymbol{\theta},\mathbf{y}_s)}{p(\mathbf{y}_s)} = \frac{p(\boldsymbol{\theta})\, p(\mathbf{y}_s|\boldsymbol{\theta})}{p(\mathbf{y}_s)}$ is the posterior distribution of the model parameters, $p(\mathbf{y}_s|\boldsymbol{\theta})$ the likelihood (as a function of $\boldsymbol{\theta}$), and $p(\mathbf{y}_s)$ a normalizing constant. The distribution in (2.1) is used to make all inferences about the non-sampled population values $\mathbf{y}_r$. To make inference to the population total $T = \sum_{i=1}^{N} y_i$, we use the posterior distribution $p\left(T \middle| \mathbf{y}_s\right)$.

Elliott and Little (2000) and Lazzeroni and Little (1993, 1998) propose using the Bayesian model-based framework to pool or collapse strata when estimating finite population means under post-stratification adjustments applied within strata. They first establish that a model is assumed under various methods that pool data at either the weight trimming (related to weight pooling) or estimation (weight smoothing) stages. In both the weight pooling and weight smoothing approaches, strata are first created using the size of the weights. These strata may either be formal strata from a disproportional stratified sample design ("inclusion strata") or "pseudo-strata" based on collapsed/pooled weights created from the selection probabilities, poststratification, and/or nonresponse adjustments. These inclusion strata are ordered by the inverse of the probability of selection, the strata above a predetermined boundary (the "cutoff" or "cutpoint stratum") are identified, and data above the cutpoint are smoothed.

To obtain the final estimate of the finite population mean in both approaches, estimates of means are calculated for each possible smoothing scenario; the key distinction is that weights are smoothed in weight pooling while the means are smoothed in weight smoothing. The final estimate is a weighted average across the means for all possible pooling scenarios, where each mean estimate produced under a trimming scenario is "weighted" by the probability that the associated trimming scenario is "correct." Since the probability that the trimming scenario is correct is calculated using the posterior probability of each smoothing cut point, conditional on the observed data and proposed Bayesian model, this method becomes variable-dependent. Although the

Bayesian models look similar, the two approaches are different and their technical details are discussed separately.

Weight Pooling Details

After dividing the sample into "strata," as defined above, and sorting the strata by size of weights w_h, the untrimmed (or "fully weighted") sample-based estimate for a mean under stratified simple random sampling is given by

$$\bar{y}_w = \frac{\sum_{h=1}^{H}\sum_{i\in h} w_h y_{hi}}{\sum_{h=1}^{H}\sum_{i\in h} w_h} = \sum_{h=1}^{H}\frac{N_h\bar{y}_h}{N}, \tag{2.2}$$

where $w_h = N_h/n_h$. Elliott and Little (2000) show that when the weights for all units within a set of strata (separated by a "cut point," denoted by l) are trimmed to the predetermined cutoff w_0, estimate (2.8) can be written as

$$\bar{y}_t = \sum_{h=1}^{l-1}\frac{\gamma N_h\bar{y}_h}{N} + \sum_{h=l}^{H}\frac{w_0 n_h\bar{y}^*}{N}, \tag{2.3}$$

where $\gamma = \frac{N - w_0\sum_{h=l}^{H} n_h}{\sum_{h=1}^{l-1} N_h}$ is the amount of "excess weight" (the weight above the cutpoint) absorbed into the non-trimmed cases and $\bar{y}^* = \frac{\sum_{h=l}^{H} n_h\bar{y}_h}{\sum_{h=1}^{l-1} n_h}$. They also show that choosing $w_0 = \sum_{h=l}^{H} N_h \Big/ \sum_{h=l}^{H} n_h$, which gives $\gamma = 1$, makes the trimmed estimator (2.3) correspond to a model-based estimator from a model that assumes distinct stratum means (μ_h) for smaller weight strata and a common mean (μ_l) for larger weight strata:

$$
\begin{aligned}
&y_{hi}|\mu_i \overset{\text{ind}}{\sim} N\left(\mu_h, \sigma^2\right), h < l \\
&y_{hi}|\mu_h \overset{\text{ind}}{\sim} N\left(\mu_l, \sigma^2\right), h \geq l\,. \\
&\mu_h, \mu_l \propto \text{constant}
\end{aligned}
\tag{2.4}
$$

They extend model (2.4) to include a noninformative prior for the weight pooling stratum, denoted $P(L = l) = H^{-1}$, and recognize that this model is a special case of a Bayesian variable selection problem (see their references): $\mathbf{y}|\boldsymbol{\beta}_l, l, \sigma^2 \overset{\text{ind}}{\sim} N\left(\mathbf{Z}_l^T \boldsymbol{\beta}_l, \Sigma\right)$, where $\mathbf{Z}_l^T$ is a $n \times l$ matrix with an intercept and dummy variables for each of the first $l-1$ strata, the parameters $\mu_1 = \beta_0, \ldots, \mu_l = \beta_0 + \beta_{l-1}$ in $\boldsymbol{\beta}_l$ are the model parameters associated with each smoothing scenario, and Σ is σ^2 times an identity matrix. That is, each smoothing scenario corresponds to a dummy variable parameter (1 for smoothing the weights within all strata above the cut point l under the pooling scenario, 0 for not smoothing) in $\mathbf{Z}_l^T$. Elliott and Little (2000) incorporate additional priors for the unknown parameters σ^2 and $\boldsymbol{\beta}_l$ from the Bayesian variable selection literature and propose

$$
\begin{aligned}
&y_{hi}|\mu_i \overset{\text{ind}}{\sim} N\left(\mu_h, \sigma^2\right), h < l \\
&y_{hi}|\mu_h \overset{\text{ind}}{\sim} N\left(\mu_l, \sigma^2\right), h \geq l \\
&P(L = l) = H^{-1} \qquad , \\
&P\left(\sigma^2 \middle| L = l\right) \propto \sigma^{-(l+1/2)} \\
&P\left(\boldsymbol{\beta}_l | \sigma^2, L = l\right) \propto (2\pi)^{-l}
\end{aligned}
\tag{2.5}
$$

where $\mu_1 = \beta_0, \ldots, \mu_l = \beta_0 + \beta_{l-1}$. Since the probability that the trimming scenario is correct is calculated using the posterior probability of each cut point l, conditional on the observed data and the hierarchical Bayesian model (2.5), this method becomes variable-dependent. That is, pooling scenarios that are identified as the "most correct" for one variable may not be for others. Also, assuming the prior (2.5) produces a posterior distribution that does not have a closed-form estimate like (2.4); while the model is more flexible, Elliott and Little found it can be susceptible to "over-pooling." Elliott (2008) used Bayesian analytic methods, such as data-based priors (Bayes Factors) and pooling conterminous strata, which improved robustness of the model and reduced the over-pooling.

Weight Smoothing Details

For their weight smoothing model, Lazzeroni and Little (1993, 1998) assume that both the survey response variables and their strata means follow Normal distributions:

$$\begin{aligned} Y_{hi} \mid \mu_h &\overset{\text{ind}}{\sim} N\left(\mu_h, \sigma^2\right) \\ \boldsymbol{\mu} &\overset{\text{ind}}{\sim} N_H\left(\mathbf{x}^T \boldsymbol{\beta}, \mathbf{D}\right) \end{aligned}, \qquad (2.6)$$

where $\boldsymbol{\mu}$ is the vector of stratum means, $\boldsymbol{\beta}$ is a vector of unknown parameters, $\mathbf{D}$ a covariance matrix, and H the total number of strata. Under model (2.6), each data value is normally distributed around the true stratum mean μ_h with constant variance σ^2. Since each μ_h is unknown, the model assumes each stratum mean follows a Normal distribution with a mean that is a linear combination of a vector of known covariates $\mathbf{x}$ and jointly follow an *H*-multivariate Normal distribution. They use this model to predict post-strata means. For $\mathbf{y}_r$, observations not in the sample, Y_{hi} is estimated by $\hat{\mu}_h$, the expected value of Y_{hi} given the data. The estimated finite population mean is the mean of the posterior distribution obtained assuming prior (2.6). It can be written as:

$$\bar{y}_{wt} = \frac{1}{N}\sum_{h=1}^{H}\left[n_h\bar{y}_h + \left(N_h - n_h\right)\hat{\mu}_h\right]. \tag{2.7}$$

The $\hat{\mu}_h$ term in (2.7) is an estimate of $\bar{y}_h$ that is smoothed toward $\mathbf{x}^T\boldsymbol{\beta}$. The (2.7) mean has a lower variance than the fully weighted poststratified mean (in (2.2), with no trimming) and the weights have less influence since we borrow information from $\hat{\mu}_h$, the means that are predicted using $\mathbf{x}^T\boldsymbol{\beta}$. In large samples, estimator (2.7) behaves like estimator (2.2) (as the $\hat{\mu}_h$ term in (2.7) tends to $\bar{y}_h$), but it smoothes stratum means toward $\mathbf{x}^T\boldsymbol{\beta}$ when the sample size is small.

Lazzeroni and Little (1998) use linear and exchangeable random effects models to estimate the parameter $\boldsymbol{\beta}$. They also consider the groups ("strata") used for establishing trimming levels (denoted by h) as being fixed. Elliott and Little (2000) extend model (2.7) and relax the assumption that h is fixed. They create the strata under particular fixed pooling patterns and use non-informative priors (in model (2.6)) for the unknown model (2.7) parameters. Estimates of smoothed means are calculated for each possible smoothing scenario. Like the weight pooling method, the final estimate is a weighted average across means for all possible pooling scenarios, where each mean estimate is "weighted" by the probability than the smoothing scenario is "correct." Using their proposed prior produces a posterior distribution that does not have a closed-form estimate like (2.7), but the model is more flexible. Elliott (2007) extends weight smoothing models to estimate the parameters in linear and generalized linear models. Elliott (2008) extends model (2.5) for linear regression, allows pooling all conterminous strata (which extends model-robustness and prevents over-pooling), and uses a fractional Bayes factor prior to compare two weight smoothing models (which increases efficiency). Elliott (2009) extends this method, using Laplace approximations to draw from the posterior distribution and estimate generalized linear model parameters. Elliott and Little (2000) also show that a semi-parametric penalized spline produces estimators of means that are more robust under model misspecification related to the necessity of pooling.

2.2.2. Superpopulation Model Prediction Approaches

Here I describe the other model-based approach in survey inference, other than the Bayesian approach described in Sec. 2.2.1. This approach involves assuming that the population survey response variables $\mathbf{Y}$ are a random sample from a larger (“super”) population and assigned a probability distribution $P\left(\mathbf{Y}|\boldsymbol{\theta}\right)$ with parameters $\boldsymbol{\theta}$. Typically the Best Linear Unbiased Prediction (BLUP, e.g., Royall 1976) method is used to estimate the model parameters. The theoretical justification for this is described next.

The BLUP Estimator and Associated Case Weights

Here, for observation i, we assume that the population values of $\mathbf{Y}$ follow the model

$$E_M\left(\mathbf{Y}_i|\mathbf{x}_i\right)=\mathbf{x}_i^T\boldsymbol{\beta},\ Var_M\left(\mathbf{Y}_i|\mathbf{x}_i\right)=\sigma^2 D_i, \tag{2.8}$$

where $\mathbf{x}_i$ denotes a p-vector of benchmark auxiliary variables for unit i, which is known for all population units. A full model-based approach uses the BLUP method to estimate the parameter $\boldsymbol{\beta}$ (Royall 1976). The BLUP-based estimator of a finite population total is the sum of the observed sample units’ total plus the sum of predicted values for the non-sample units, denoted by r:

$$\hat{T}_{BLUP} = \sum_{i \in s} y_i + \sum_{i \in r} \mathbf{x}_i^T \hat{\boldsymbol{\beta}} . \tag{2.9}$$

Valliant *et al.* (2000) demonstrate that, when the corresponding model holds, estimator (2.9) is the best estimator of the total. However, when the model does not hold, model-misspecification related bias is introduced. Note that estimator (2.9) is also variable-specific; a separate model must be formulated for each y-variable of interest. In addition, the resulting BLUP-based weights for a particular sample unit can be negative or less than one, which is undesirable from a design-based perspective.

When x_i is a scalar, every design- and model-based estimator can be written in a form resembling (3.11) using the following expression (p. 26 in Valliant *et al.* 2000):

$$\hat{T} = \sum_{i \in s} y_i + \left[\frac{\hat{T} - \sum_{i \in s} y_i}{\sum_{i \in r} x_i} \right] \sum_{i \in r} x_i . \tag{2.10}$$

Alternatively, we can write (2.10) as

$$\begin{aligned} \hat{T} &= \sum_{i \in s} y_i + \sum_{i \in r} \hat{y}_i \\ &= \sum_{i \in s} y_i + \sum_{i \in s} \left(w_i - 1 \right) y_i \end{aligned} , \tag{2.11}$$

where the component $\sum_{i \in s} \left(w_i - 1 \right) y_i$ is an estimator of the term $\sum_{i \in r} y_i$. All subsequent estimators can be written in these forms. Both (2.10) and (2.11) can be used to explicitly define the associated case weights. In general, for $\mathbf{Y} = \left(y_1, \ldots, y_N \right)^T$ denoting the vector of population y-values, since the total is a linear combination of $\mathbf{Y}$, we can write $T = \boldsymbol{\gamma}^T \mathbf{Y}$, where $\gamma_i = 1$. We partition both components into the sample and non-sample values, $\mathbf{Y} = \left(\mathbf{y}_s^T , \mathbf{y}_r^T \right)$ and $\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}_s^T , \boldsymbol{\gamma}_r^T \right)$. The population total is then

$T = \boldsymbol{\gamma}_s^T \mathbf{y}_s + \boldsymbol{\gamma}_r^T \mathbf{y}_r$. If we denote the linear estimator as $\hat{T} = \mathbf{g}_s^T \mathbf{y}_s$, where $\mathbf{g} = (g_1, \ldots, g_n)^T$ is a n-vector of coefficients, the estimator error in $\hat{T}$ is

$$\begin{aligned} \hat{T} - T &= \mathbf{g}_s^T \mathbf{y}_s - \left(\boldsymbol{\gamma}_s^T \mathbf{y}_s + \boldsymbol{\gamma}_r^T \mathbf{y}_r \right) \\ &= \mathbf{a}^T \mathbf{y}_s - \boldsymbol{\gamma}_r^T \mathbf{y}_r \end{aligned}, \tag{2.12}$$

where $\mathbf{a} = \mathbf{g}_S \boldsymbol{\gamma}_S$. The term $\mathbf{a}^T \mathbf{y}_s$ in (2.12) is known from the sample, but $\boldsymbol{\gamma}_r^T \mathbf{y}_r$ must be predicted using the model parameters estimated from the sample and the $\mathbf{x}$-values in the population that are not in the sample. Thus, we similarly partition the population covariates $\mathbf{x} = \begin{bmatrix} \mathbf{X}_s \\ \mathbf{X}_r \end{bmatrix}$, where $\mathbf{X}_s$ is the $n \times p$ matrix and $\mathbf{X}_r$ is $(N-n) \times p$, and variances $\mathbf{V} = \begin{bmatrix} \mathbf{V}_{ss} & \mathbf{V}_{sr} \\ \mathbf{V}_{rs} & \mathbf{V}_{rr} \end{bmatrix}$, where $\mathbf{V}_{ss}$ is $n \times n$, $\mathbf{V}_{rr}$ is $(N-n) \times (N-n)$, $\mathbf{V}_{sr}$ is $n \times (N-n)$, and $\mathbf{V}_{rs} = \mathbf{V}_{sr}^T$. Then, under the general prediction theorem (Thm. 2.2.1 in Valliant *et al.* 2000), the optimal estimator of a total is

$$\hat{T}_{opt} = \boldsymbol{\gamma}_s^T \mathbf{y}_s + \boldsymbol{\gamma}_r^T \left(\mathbf{X}_r^T \hat{\boldsymbol{\beta}} + \mathbf{V}_{rs} \mathbf{V}_{ss}^{-1} \left(\mathbf{y}_s - \mathbf{X}_s^T \hat{\boldsymbol{\beta}} \right) \right), \tag{2.13}$$

where $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}_s^T \mathbf{V}_{ss}^{-1} \mathbf{X}_s \right)^{-1} \mathbf{X}_s^T \mathbf{V}_{ss}^{-1} \mathbf{y}_s$. Using Lagrange multipliers, Valliant *et al.* (2000) obtain the optimal value of $\mathbf{a}$ as

$$\mathbf{a}_{opt} = \mathbf{V}_{ss}^{-1} \left[\mathbf{V}_{sr} + \mathbf{X}_s \mathbf{A}_s^{-1} \left(\mathbf{X}_r^T - \mathbf{X}_s^T \mathbf{V}_{ss}^{-1} \mathbf{V}_{sr} \right) \right] \boldsymbol{\gamma}_r, \tag{2.14}$$

where $\mathbf{A}_s = \mathbf{X}_s^T \mathbf{V}_{ss}^{-1} \mathbf{X}_s$. This leads to the optimal vector of BLUP coefficients $\mathbf{g}_s = \mathbf{a}_{opt} + \boldsymbol{\gamma}_s$, where the ith component is the "weight" on sample unit i (Valliant *et. al* 2000; Valliant 2009). This weight depends on the regression component of the model

$E_M(\mathbf{Y})$, the variance $Var_M(\mathbf{Y})$, and how the sample and non-sample units are designated. In general, the case weights for a total are

$$\mathbf{g}_s = \mathbf{V}_{ss}^{-1}\left[\mathbf{V}_{sr} + \mathbf{X}_s\mathbf{A}_s^{-1}\left(\mathbf{X}_r^T - \mathbf{X}_s^T\mathbf{V}_{ss}^{-1}\mathbf{V}_{sr}\right)\right]\mathbf{1}_r + \mathbf{1}_s, \tag{2.15}$$

where $\mathbf{1}_s, \mathbf{1}_r$ are $n \times 1$ and $(N-n)\times 1$ vectors of units with elements that are all 1's, respectively. For the total under the general linear model with constant variance, (2.15) reduces to

$$\mathbf{g}_s = \mathbf{X}_s\left(\mathbf{X}_s^T\mathbf{X}_s\right)^{-1}\mathbf{X}_r^T\mathbf{1}_r + \mathbf{1}_s. \tag{2.16}$$

In the remainder of this section, I provide examples of simple BLUP-based estimators of totals and the associated case weights, as well as more robust alternatives that have been proposed in the related literature.

Simple Model-based Weight Examples

Example 2.1. *HT Estimator, simple random sampling*. In simple random sampling, where $w_i = N/n$ and the model is $y_i | X_i = \mu + e_i,\ e_i \overset{\text{ind}}{\sim} \left(0, \sigma^2\right)$, $\hat{T}_{HT} = \sum_{i\in s} y_i + (N-n)\bar{y}_s$. In this case, every unit in the population but not in the sample is predicted with the same value, the sample mean $\bar{y}_s = \frac{1}{n}\sum_{i\in s} y_i$. We also see how the HT estimator does not incorporate any auxiliary information when formulated this way, which corresponds to a very simple model.
■

Example 2.2. *Ratio Estimator, simple random sampling*. For a single auxiliary variable X_i, suppose the true model is the ratio model, $y_i | X_i = X_i\beta + e_i, e_i \sim \left(0, X_i\sigma^2\right)$, or a regression through the origin with a variance proportional to X_i. The optimal estimator associated with this model is the ratio estimator $\hat{T}_R = \frac{N\bar{y}\bar{X}}{\bar{x}}$ for $\bar{y}, \bar{x}$ denoting the sample means and $\bar{X}$ the population mean. The ratio estimator has the equivalent form to (2.11) as $\hat{T}_R = \sum_{i\in s} y_i + \sum_{i\in r} \hat{\beta} X_i = \frac{\sum_{i\in U} X_i \sum_{i\in s} y_i}{\sum_{i\in s} X_i} = \frac{N\bar{X}_U \bar{y}_s}{\bar{X}_s}$, where $\bar{X}_s = \frac{1}{n}\sum_{i\in s} X_i$, $\bar{X}_U = \frac{1}{N}\sum_{i\in U} X_i$, and the weights are the same for all units, i.e., $w_i = N\bar{X}_U / n\bar{X}_s$.

■

Example 2.3. *Simple Linear Regression Estimator, simple random sampling*. Here the model for the regression estimator is $y_i | X_i = \beta_0 + \beta_1 X_i + e_i,\ e_i \overset{\text{ind}}{\sim} \left(0, \sigma^2\right)$, $\hat{T}_{REG} = \sum_{i\in s} y_i + N\hat{\beta}_1\left(\bar{X}_U - \bar{X}_s\right)$, where $\hat{\beta}_1 = \sum_{i\in s}\left(y_i - \bar{y}_s\right)\left(X_i - \bar{X}_s\right) \Big/ \sum_{i\in s}\left(X_i - \bar{X}_s\right)^2$. The weights here are equivalent to (2.16), where $\mathbf{X}_s = \left[\mathbf{1}_s \quad X_s\right]$, $\mathbf{1}_s$ is a $n\times 1$ vector of 1's, $X_s = \left(X_1, \ldots, X_n\right)^T$, $\mathbf{X}_r = \left[\mathbf{1}_r \quad X_r\right]$, $\mathbf{1}_r$ is a $\left(N-n\right)\times 1$ vector of 1's, and $X_r = \left(X_1, \ldots, X_{N-n}\right)^T$.

■

Robust Model-based Weight Examples

Since the efficiency of the simple methods in Ex. 2.1-2.3 depend on how well the associated model holds, these methods can be susceptible to model misspecification. When comparing a set of candidate weights to a preferable set of weights, the difference in the estimated totals under the "preferable" model attributed to model misspecifcation is a measure of design-based inefficiency or model bias. To overcome the bias, the superpopulation literature has developed a few robust alternatives, with examples given here. Generally, each approach involves using the preferable alternative model to produce an adjustment factor that is added to the BLUP.

Example 2.4. *Dorfman's Kernel Regression Method.* Dorfman (2000) proposed outlier-robust estimation of $\boldsymbol{\beta}$ in the BLUP estimator in (2.13) using kernel regression smoothing. Suppose that the true model is $y_i | x_i = m\left(x_i\right) + v_i e_i$, where $m\left(x_i\right)$ is a smooth and (at least) twice-differentiable function. For $j \in r$, he proposed estimating $m\left(x_j\right)$ with $\hat{m}\left(x_j\right) = \sum_{i\in s} w_{ij} y_i$, where $\sum_{i\in s} w_{ij} = 1$ and larger w_{ij}'s imply that $x_i, i \in s$ is closer to $x_j, j \in r$. He proposed to use kernel regression smoothing to produce the weights $w_{ij} = \dfrac{K_b\left(x_i - x_j\right)}{\sum_{i\in s} K_b\left(x_i - x_j\right)}$, where $K(u)$ denotes a density function that is symmetric around zero, from which a family of densities is produced from using the scale transformation $K_b(u) = b^{-1}K(u/b)$, and the scale b is referred to as a "bandwidth." His nonparametric estimator for the finite population total is thus given by

$$\begin{aligned}\hat{T}_D &= \sum_{i\in s} y_i + \sum_{i\in r} \hat{m}(x_i) \\ &= \sum_{i\in s} y_i + \sum_{i\in s}\sum_{j\in r} w_{ij} y_i , \\ &= \sum_{i\in s}(1 + w_{i+}) y_i\end{aligned} \tag{2.17}$$

where $w_{i+} = \sum_{j\in r} w_{ij}$. Here the case weights are $w_i = 1 + w_{i+}$. While Dorfman found this estimator can produce totals with lower MSE's, it was sensitive to the choice of bandwidth. When x_i is categorical, this method is not appropriate since there is not a range of x_i over which to smooth.

■

Example 2.5. *Chambers et al.'s NP Calibration Method.* Chambers *et al.* (1993) proposed an alternative to Dorfman's kernel regression approach (Ex. 2.4) that applies a model-bias correction factor to linear regression case weights. This bias correction factor is produced using a nonparametric smoothing of the linear model residuals against frame variables known for all population units is applied to the BLUP estimator (2.13). Suppose that the true model is $y_i|\mathbf{x}_i = m(\mathbf{x}_i) + v_i e_i$, with working model variance $Var(y_i|\mathbf{x}_i) = \sigma^2 D_i$, where D_i is a measure of size for population unit i. If the BLUP estimator (2.13) was used to estimate the finite population total, then the model bias in the total is $E_M\left(\hat{T}_{BLUP} - T\right) = \sum_{i\in r}\delta(\mathbf{x}_i)$, where $\delta(\mathbf{x}_i) = \mathbf{x}_i^T E_M\left(\hat{\boldsymbol{\beta}}\right) - m(\mathbf{x}_i)$. Since the residual $\hat{e}_i = y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $-\delta(\mathbf{X}_i)$, they used sample-based residuals to estimate the nonsample $\delta(\mathbf{x}_i)$ values. This produced the *nonparametric calibration estimator* for the finite population, given by

$$\begin{aligned}\hat{T}_{C1} &= \sum_{i\in s} y_i + \sum_{i\in r}\mathbf{x}_i^T\hat{\boldsymbol{\beta}} - \sum_{i\in s}\hat{e}_i \\ &= \hat{T}_{BLUP} + \sum_{i\in s}\hat{\delta}(\mathbf{x}_i)\end{aligned} , \tag{2.18}$$

Here, the associated case weights are $\mathbf{g}_s = \mathbf{X}_s\left(\mathbf{X}_s^T\mathbf{X}_s\right)^{-1}\mathbf{X}_r^T\mathbf{1}_r + \mathbf{1}_s + \mathbf{m}_s$, where $\mathbf{m}_s$ contains the residual-based estimates of $\delta(\mathbf{x}_i)$.

■

Example 2.6. *Chambers' Ridge Regression Method.* Chambers (1996) proposed an alternative outlier-robust estimation of $\boldsymbol{\beta}$ in the BLUP estimator in (2.13) using a GREG-type approach. He proposes to find the sets of weights $\mathbf{w}$ that minimize a $\boldsymbol{\lambda}$-scaled, cost-ridged loss function

$$Q_{\lambda}\left(\boldsymbol{\Omega},\mathbf{g},\mathbf{c}\right)=\sum_{i\in s}\Omega_i\left(\frac{\left(w_i-g_i\right)^2}{g_i}\right)+\frac{1}{\lambda}\sum_{j=1}^{p}c_j\left(\hat{\mathbf{T}}_{x_j\mathbf{w}}-\mathbf{T}_{x_j}\right)^2, \tag{2.19}$$

where Ω_i and g_i are both pre-specified constants (e.g., $\Omega_i=D_i$ and $g_i=1$ for the BLUP), w_i is the original weight, and $\mathbf{c}=diag\left(c_j\right), j=1,\ldots,p$ is a vector of prespecified non-negative constants representing the cost of the case weighted estimator not satisfying the calibration constraint $\hat{\mathbf{T}}_{x_j\mathbf{w}}-\mathbf{T}_{x_j}$, where $\hat{\mathbf{T}}_{x_j\mathbf{w}}=\sum_{i\in s}w_i x_{ij}$, $\hat{\mathbf{T}}_{x_j}=\sum_{i\in U}x_{ij}$ is the population total of variable j, and λ is a user-specified scale function. Minimizing (2.19) produces the ridge-regression weights:

$$\mathbf{w}_{\lambda}=\mathbf{g}_s+\mathbf{A}_s^{-1}\mathbf{X}_s\left(\lambda\mathbf{c}^{-1}+\mathbf{X}_s^T\mathbf{A}_s^{-1}\mathbf{X}_s\right)^{-1}\left(\hat{\mathbf{T}}_{x_j\mathbf{w}}-\mathbf{T}_{x_j}\right), \tag{2.20}$$

where $\mathbf{T}_x$ is the vector of population totals, $\mathbf{X}_s$ is the vector of sample totals, and $\mathbf{A}_s$ is the diagonal variance matrix with ith diagonal element $\Omega_i g_i^{-1}$ (e.g., $diag\left(\mathbf{A}_s\right)=D_i$ for Chambers' BLUP). Chambers showed how $\lambda=0$ reduces expression (2.20) to the calibration weights and $\lambda>0$ produces weights that produce estimators that are biased, but have lower variance. The population total is estimated by

$$\hat{T}_{C2}=\sum_{i\in s}y_i+\sum_{i\in r}\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{\lambda}, \tag{2.21}$$

where $\hat{\boldsymbol{\beta}}_{\lambda}$ is the ridge-weighted estimator of $\boldsymbol{\beta}$ using the weights (2.20) and the linear model $E_M\left(\mathbf{Y}\right)=\mathbf{x}^T\boldsymbol{\beta}$.

■

Example 2.7. *Chambers' NP Bias Correction Ridge Regression Method.* Chambers (1996) also proposed a nonparametric approach to obtaining the ridge regression weights. His NP version of the weights (2.22) is

$$\mathbf{w}_{\lambda,m}=\mathbf{1}_s+\mathbf{m}_s+\mathbf{A}_s^{-1}\mathbf{X}_s\left(\lambda\mathbf{c}^{-1}+\mathbf{X}_s^T\mathbf{A}_s^{-1}\mathbf{X}_s\right)^{-1}\left(\mathbf{T}_x-\mathbf{X}_s^T\mathbf{1}_s-\mathbf{X}_s^T\mathbf{m}_s\right), \tag{2.23}$$

where $\mathbf{1}_s$ is a vector of 1's of length n and $\mathbf{m}_s$ the NP-corrected weights (e.g., the kernel-smoothing weights in Ex. 2.4). The (2.23) weights depend on the choice of λ and choices related to how the NP weights are constructed. For example, using the kernel smoothing-based weights (Ex. 2.4), the weights in (2.23) depend on the bandwidths of the kernel smoother, which is a separate choice from determining λ. He also recommends choosing λ such that all weights $\mathbf{w}_{\lambda,m} \geq 1$. Assuming that the ridge estimator model is correct, but the BLUP model was used to estimate the total, he incorporates a bias correction factor into his ridge estimator:

$$\begin{aligned}\hat{T}_{C3} &= \sum_{i \in s} y_i + \sum_{i \in r} \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\lambda + \sum_{i \in s} m_i \left(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\lambda \right) \\ &= \hat{T}_{C2} + \sum_{i \in s} m_i \left(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\lambda \right)\end{aligned}, \qquad (2.24)$$

where $\hat{\boldsymbol{\beta}}_\lambda$ is the ridge-weighted estimator $\boldsymbol{\beta}$ (estimated using the weights (2.23) and linear model $\mathbf{Y} = \mathbf{x}^T \boldsymbol{\beta}$) and $m_i \left(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\lambda \right)$ are the nonparametric predicted weights obtained by summing the contributions of unit i to the NP prediction of the linear model residual, calculated for all $N - n$ units in r. The case weights applied to each y_i are (2.23).

■

Example 2.8. *Firth and Bennett's Method.* Firth and Bennett (1998) produce a similar bias-correction factor to Chambers *et al.* (1993, see Ex. 3.5) for a difference estimator (Cassell *et al.* 1976) as follows:

$$\hat{T}_D = \hat{T}_{BLUP} + \sum_{i \in s} (w_i - 1) \left(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}} \right). \qquad (2.25)$$

The associated weights are $\mathbf{g}_s = \mathbf{V}_{ss}^{-1} \left[\mathbf{V}_{sr} + \mathbf{X}_s \mathbf{A}_s^{-1} \left(\mathbf{X}_r^T - \mathbf{X}_s^T \mathbf{V}_{ss}^{-1} \mathbf{V}_{sr} \right) \right] \mathbf{1}_r + \mathbf{1}_s + \mathbf{m}_s$, where $\mathbf{m}_s$ contains the residual-based estimates of $(w_i - 1)\left(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}} \right)$, and w_i are the original BLUP weights. Firth and Bennett also define the *internal bias calibration* property to hold when $\sum_{i \in s} w_i \left(y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}} \right) = 0$, for all s under the given sample design. They also provide examples of when this property holds, e.g., using generalized linear models with a canonical link function to predict $\mathbf{x}_i^T \hat{\boldsymbol{\beta}}$ and incorporating the survey weights in the estimating equations for the model parameters or the regression model.

■

As shown in the preceding examples, the superpopulation inference approach can indirectly induce weight trimming by producing estimators that take advantage of an underlying model relationship, but they have not been directly developed specifically for this purpose. The main advantage to the model-based approach is that when the underlying superpopulation model holds, estimators of totals have a lower MSE due to a decrease in variance and no bias. Examples 2.4 through 2.8 illustrate solutions to problems of model misspecification and robustness to extreme values. However, this has not been developed for weight trimming models. When the assumed underlying model does not hold, the bias of the estimates increases and can offset the MSE gains achieved by having lower variances. It is also necessary to postulate and validate a model for each variable of interest, which leads to variable-specific estimators. This can be practically inconvenient when analyzing many variables.

2.2.3. Penalized Spline Estimation

Recent survey methodology research has focused on a class of estimators based on penalized (*p*-) spline regression to estimate finite population parameters (Zheng and Little 2003, 2005, Breidt *et al.* 2005; Krivobokova *et al*. 2008; Claeskens *et. al* 2009). Separately, Eilers and Marx (1996) introduced penalized spline estimators; Ruppert *et al*. (2003), Ruppert and Carroll (2000), and Wand (2003) developed them further theoretically. Breidt *et al.* (2005) develop a model-assisted *p*-spline estimator similar to the GREG estimator. In application, they showed their *p*-spline estimator is more efficient than parametric GREG estimators when the parametric model is misspecified, but the *p*-spline estimator is approximately as efficient when the parametric specification is correct. However, this method applies for quantitative covariates.

Breidt *et al*. (2005) convert the Ruppert *et al.* (2003) model into finite population sampling by assuming that quantitative auxiliary variables x_i are available and known for all population units. The details related to Ex. 3.10 are described here. They propose the following superpopulation regression model:

$$y_i = m(x_i) + \varepsilon_i,\ \varepsilon_i \overset{\text{ind}}{\sim} N\left(0, v(x_i)\right). \tag{2.26}$$

Treating $\{(x_i, y_i) : i \in U\}$ as one realization from model (2.26), the *p-spline function* using a linear combination of truncated polynomials is

$$m(x, \boldsymbol{\beta}) = \beta_0 + \beta_1 x + \cdots \beta_p x^p + \sum_{q=1}^{Q} \beta_{q+p} \left(x - \kappa_q\right)_+^p, i = 1, \ldots, N, \tag{2.27}$$

where the constants $\kappa_1 < \ldots < \kappa_L$ are fixed "knots," and the term $(u)_+^p = u^p$ if $u > 0$ and zero, otherwise, p is the degree of the spline, and $\boldsymbol{\beta} = \left(\beta_0, \ldots, \beta_{p+Q}\right)^T$ is the coefficient vector. The splines here are piecewise polynomial functions that are smooth to a certain degree, and can be expressed as a linear combination of a set of basis functions defined with respect to the number of knots. The truncated polynomial version shown in (2.27) is often chosen for its simplicity over other alternatives (e.g., *B*-splines, as used in Eilers and Marx 1996). Zheng and Little (2003) adjusted the superpopulation model (2.26) to produce a *p*-spline estimator that accounts for the effect of non-ignorable design weights:

$$y_i = m(\pi_i, \boldsymbol{\beta}) + \varepsilon_i,\ \varepsilon_i \overset{\text{ind}}{\sim} N\left(0, \pi_i^{2k} \sigma^2\right), \tag{2.28}$$

where the constant $k \geq 0$ reflects knowledge of the error variance heteroskedasticity and

$$m(\pi_i, \boldsymbol{\beta}) = \beta_0 + \sum_{j=1}^{p} \beta_j \pi_i^j + \sum_{q=1}^{Q} \beta_{q+p} \left(\pi_i - \kappa_q\right)_+^p, i = 1, \ldots, N \tag{2.29}$$

is the spline function. Ruppert (2002) and Ruppert *et al.* (2003) recommend using a relative large number of knots (15 to 30) at pre-specified locations, such that the smoothing is achieved by treating the parameters $\beta_{p+1}, \ldots, \beta_{p+Q}$ as random effects centered at zero. Otherwise, using a least-squares approach to estimate $\beta_{p+1}, \ldots, \beta_{p+Q}$ can result in over-fitting the model.

While knot selection methods exist (Friedman and Silverman 1989; Friedman 1991; Green 1995; Stone *et al.* 1997; Denison *et al.* 1998), in *penalized* (*p*)*-spline regression* the number of knots is large, but their influence is bounded using a constraint on the Q spline coefficients. One such constraint with the truncated polynomial model is to bound $\sum_{q=1}^{Q} \beta_{q+p}^2$ by some constant, while leaving the polynomial coefficients $\beta_0, \ldots, \beta_p$ unconstrained. This smoothes the $\beta_{p+1}, \ldots, \beta_{p+Q}$ estimates toward zero. Adding the constraint as a Lagrange multiplier, denoted by α, in the least squares equation gives

$$\hat{\boldsymbol{\beta}} = \arg_{\boldsymbol{\beta}} \min \sum_{i \in U} \left(y_i - m\left(\mathbf{x}_i, \boldsymbol{\beta}\right)\right)^2 + \alpha \sum_{q=1}^{Q} \beta_{q+p}^2 \tag{2.30}$$

for a fixed constant $\alpha \geq 0$. The smoothing of the resulting fit depends on α; larger values produce smoother fits. Zheng and Little (2003) recognized that treating the $\beta_{p+1}, \ldots, \beta_{p+Q}$ as random effects, given the penalty $\alpha \sum_{q=1}^{Q} \beta_{q+p}^2$, is equivalent to using a multivariate normal prior $\beta_{p+q}, \ldots, \beta_{p+Q} \sim N_L\left(0, \tau^2 \mathbf{I}_Q\right)$, where $\tau^2 = \sigma^2 / \alpha$ is an additional parameter estimated from the data and $\mathbf{I}_Q$ is a $Q \times Q$ identity matrix. For $m_i = m\left(\mathbf{x}_i, \boldsymbol{\beta}_U\right), i \in U$ denoting the *p*-spline fit obtained from the hypothetical population fit at $\mathbf{x}_i$, Breidt *et al*. (2005) incorporate m_i into survey estimation by using a difference estimator

$$\sum_{i \in U} m_i + \sum_{i \in s} \frac{y_i - m_i}{\pi_i}. \tag{2.31}$$

Given a sample, m_i in (2.31) can be estimated using a sample-based estimator $\hat{m}_i$. For $\mathbf{W} = diag\left(1/\pi_i\right), i \in U$ and $\mathbf{W}_s = diag\left(1/\pi_i\right), i \in s$ as the matrices of the HT weights in the population and sample, for fixed α, the π-weighted estimator for the *p*-spline model coefficients is

$$\begin{aligned}\hat{\boldsymbol{\beta}}_\pi &= \left(\mathbf{x}_s^T \mathbf{W}_s \mathbf{x}_s + \mathbf{D}_\alpha\right)^{-1} \mathbf{x}_s^T \mathbf{W}_s \mathbf{y}_s \\ &= \mathbf{G}_\alpha \mathbf{y}_s\end{aligned}, \tag{2.32}$$

such that $\hat{m}_i = m\left(\mathbf{x}_i, \hat{\boldsymbol{\beta}}_\pi\right)$. The model-assisted estimator for the finite population total is

$$\begin{aligned}\hat{T}_{mpsp} &= \sum_{i\in U} \hat{m}_i + \sum_{i\in s} \frac{y_i - \hat{m}_i}{\pi_i} \\ &\doteq \sum_{i\in s} \left[\frac{1}{\pi_i} + \sum_{j\in U} \left(1 - \frac{I_j}{\pi_j}\right) \mathbf{x}_j^T \mathbf{G}_\alpha e_i\right] y_i, \\ &= \sum_{i\in s} w_i^* y_i\end{aligned} \tag{2.33}$$

where $I_j = 1$ if $j \in s$ and zero otherwise and $e_i = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\pi$. From (2.33), their *p*-spline estimator is a linear estimator. The case weights here are w_i^* in (2.33). Chambers' ridge regression estimator in (2.21) has a similar form, with ridge matrix $diag\left(\alpha_1, \ldots, \alpha_p\right)$, where $\alpha_i = 0$ for covariates corresponding to the calibration constraints that must be met. Breidt *et. al* (2005) also showed that this estimator shares many of the desirable properties of the GREG estimator. However, since it uses a more flexible model, the *p*-spline estimator (2.33) had improved efficiency over the GREG when the linear model did not hold.

An appealing property of the penalized spline estimator is that it can be rewritten into a mixed-model format. For example, the Zheng and Little model in (2.28) and (2.29) can be rewritten as:

$$\mathbf{y}_s = \mathbf{x}\mathbf{B}_1 + \mathbf{z}\mathbf{B}_2 + \boldsymbol{\varepsilon}, \tag{2.34}$$

where $\mathbf{y}_s = \left(y_1, \ldots, y_n\right)^T$, $\mathbf{B}_1 = \left(\beta_1, \ldots, \beta_p\right)^T$, $\mathbf{B}_2 = \left(\beta_{p+1}, \ldots, \beta_{p+Q}\right)^T \sim N_Q\left(0, \tau^2 \mathbf{I}_Q\right)$,

$$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T \overset{\text{ind}}{\sim} N_n\left(0, \sigma^2 diag\left(\pi_1^{2k}, \ldots, \pi_n^{2k}\right)\right),$$

$$\mathbf{x} = \begin{pmatrix} 1 & \pi_1 & \cdots & \pi_1^p \\ 1 & \pi_2 & \cdots & \pi_2^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \pi_n & \cdots & \pi_n^p \end{pmatrix}, \ \mathbf{z} = \begin{pmatrix} (\pi_1 - \kappa_1)_+^p & \cdots & (\pi_1 - \kappa_Q)_+^p \\ \vdots & \ddots & \vdots \\ (\pi_n - \kappa_1)_+^p & \cdots & (\pi_n - \kappa_Q)_+^p \end{pmatrix}, \text{ and } q = 1, \ldots, Q$$

For $k \neq 0$ and Normal errors, the maximum likelihood estimator of $\mathbf{B} = (\beta_0, \ldots, \beta_{p+Q})^T$ is

$$\mathbf{B}_{ML} \big| \alpha = \sigma^2 / \tau^2 = \left[\mathbf{\Pi}^T \mathbf{V}_s \mathbf{\Pi} + \mathbf{D}_\alpha\right]^{-1} \mathbf{\Pi}^T \mathbf{V}_s \mathbf{y}_s, \qquad (2.35)$$

where the ith row of the matrix $\mathbf{\Pi}$ is $\mathbf{\Pi}_i = \left(1, \pi_i, \ldots, \pi_i^p, (\pi_i - \kappa_1)_+^p, \ldots, (\pi_i - \kappa_Q)_+^p\right)$; $\mathbf{D}_\alpha$ is a diagonal matrix with the first $p+1$ elements being zero and remaining Q elements all equal to the penalty $\alpha = \sigma^2 / \tau^2$, the value chosen to maximize the likelihood of the GLM model; and $\mathbf{V}_s = diag\left(\pi_1^{2k}, \pi_2^{2k}, \ldots, \pi_n^{2k}\right)$ denotes the variance-covariance matrix specified under the model.

If the component $\hat{\alpha}$ is fixed, then the *p*-spline estimator is equivalent to Chambers' ridge regression estimator (ex. 2.6). However, since the variance components σ^2 and τ^2 are unknown, Breidt *et al.* (2005) propose using Ruppert and Carroll's (2000) "data-driven" penalty obtained using the GLM formulation of the model and REML (Patterson and Thompson 1971; Harville 1977; Searle *et al*. 1992) to estimate $\mathbf{D}_\alpha$ with $\hat{\mathbf{D}}_\alpha$, whose last Q elements are all $\hat{\alpha}_{REML} = \hat{\sigma}_{REML}^2 / \hat{\tau}_{REML}^2$. However, in many applications (e.g., Milliken and Johnson 1992), REML-based estimators can produce negative estimates of variance components. Often conventional software will truncate the variance estimate to zero; here this corresponds to an estimate of either $\hat{\alpha}_{REML} = 0$ or $\hat{\alpha}_{REML} = \infty$. If $\hat{\alpha}_{REML} = 0$ (i.e., if $\hat{\sigma}^2_{REML} = 0$), then the spline coefficients $\beta_{p+1},\ldots,\beta_{p+Q}$ are all exactly zero. However, if $\hat{\alpha}_{REML} = \infty$ (i.e., if $\hat{\tau}^2_{REML} = 0$), then they are all undefined.

Accounting for the sample design features, such as stratification, clustering, and weighting, can increase model robustness. Zheng and Little's (2003) *p*-spline model produced estimates of the finite population total that had negligible bias and improved efficiency over the HT and GREG estimators. In addition to proposing new estimators, I compare design-based, Bayesian model-based, and model-assisted estimators against the proposed alternative estimators in single-stage sample designs in Sec. 2.4.1

Zheng and Little (2005) extended their approach to use model-based, jackknife, and balanced repeated replicate variance estimation methods for the *p*-spline estimators. This improved inferential results, such as confidence interval coverage. In Sec. 2.3.2, I adopt variance estimators proposed in the related literature for the estimator in Sec. 2.3.1.

2.2.4. Summary

The Bayesian trimming procedures are a theoretical breakthrough for weight trimming. They lay the foundation for particular forms of estimation. They account for effects in the realized survey response variable values by taking posterior distribution estimates that are conditional on the observed sample data. Their main advantage here is that, when the underlying model holds, the resulting trimmed weights produce point estimates with lower MSE due to a decrease in variance that is larger than the increase in squared bias (Little 2004). Elliott and Little (2000) and Elliott (2007, 2008, 2009) demonstrate empirically with simulations and case studies that their methods can potentially increase efficiency and decrease the MSE. The use of *p*-spline estimation has also produced results that are more robust when the Bayesian model does not hold, without much loss of efficiency.

One drawback to the Bayesian method is that the smoothing occurs for a particular set of circumstances: weighting adjustments performed within design strata, under noninformative and equal probability sampling, and for one estimation purpose (e.g., means, regression coefficients, etc.) of a small number of (often one) variables of interest. Also, for both model-based methods, it also is necessary to propose and validate a model for each variable of interest, which may then lead to variable-specific sets of weights. Although these model-based approaches may be appealing from the viewpoint of statistical efficiency, they may be practically inconvenient when there are many variables of interest. The presence of variable-dependent weights on a public use file is potentially confusing to data users, particularly when they conduct multivariate analyses on the data. However, the MSE-minimization benefit of the Bayesian method may outweigh these practical limitations. The simplicity and flexibility of penalized splines can improve the model robustness and may reduce the need for variable-dependent weights. This method also requires the availability of quantitative auxiliary information; methods for categorical and binary covariates have not been examined.

2.3. Proposed Model and Estimation Methods

2.3.1. Using Priors in the Breidt et al. Model

Here, I propose a *p*-spline model that is a modification of Breidt *et al*.'s model (2.27):

$$y_i = m(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i,\ \varepsilon_i \overset{\text{ind}}{\sim} N\left(0, x_i^k \sigma^2\right), \tag{2.36}$$

where $\boldsymbol{\beta}$ is the slope coefficient vector, $\mathbf{x}_i$ is a matrix of unit-level covariates (e.g., design variables, auxiliary frame variables, or variables used to create the weights, like those used in poststratification or nonresponse adjustments) and σ^2 is a variance component. In (2.36), the estimate of y_i is the $\hat{m}_i$, the estimated value of y_i under the model and conditional on the data. From (2.36), the following spline function is fit:

$$\begin{aligned} m(\pi_i, \boldsymbol{\beta}) &= \beta_0 + \sum_{j=1}^{p} \beta_j x_i^j + \sum_{q=1}^{Q} \beta_{(l+q)} \left(x_i - \kappa_q\right)_+^p, \\ &= \beta_0 + \sum_{j=1}^{p} \beta_j x_i^j + \sum_{q=1}^{Q} \beta_{(l+q)} z_i \end{aligned} \tag{2.37}$$

where $z_i = \left(x_i - \kappa_q\right)_+^p = \left(x_i - \kappa_q\right)^p$ if $x_i - \kappa_q > 0$, and zero otherwise. Model (2.36) can be rewritten in the GLM form as follows:

$$\mathbf{y} = \mathbf{x}^T \mathbf{B}_1 + \mathbf{z}^T \mathbf{B}_2 + \mathbf{e}, \tag{2.38}$$

where $\mathbf{y} = \left(y_1, \ldots, y_n\right)^T$, $\mathbf{B}_1 = \left(\beta_1, \ldots, \beta_p\right)^T$, $\mathbf{B}_2 = \left(\beta_{(p+1)}, \ldots, \beta_{(p+Q)}\right)^T$, $\mathbf{e} = \left(e_1, \ldots, e_n\right)^T \overset{\text{ind}}{\sim} N_n\left(0, \sigma^2 \mathbf{V}_e\right)$, $\mathbf{V}_e = diag\left(x_1^{2k}, \ldots, x_n^{2k}\right)$, p is the degree of the truncated polynomial (the number of fixed effects, including the intercept), Q is the total number of knots (and the number of random effects), $\mathbf{x}$ is a $n \times (p+1)$ vector, where the i row is $\begin{bmatrix} 1 & x_i & \cdots & x_i^p \end{bmatrix}$, and $\mathbf{z}$ is a $n \times Q$ vector, where the i row is $\begin{bmatrix} (x_i - \kappa_1)_+^p & \cdots & \left(x_i - \kappa_Q\right)_+^p \end{bmatrix}$. One choice for knot κ_q is the sample quantile of x_i corresponding to probability $\frac{\kappa_q}{Q+1}$.

The following conventional priors can be used for the additional unknown parameters (Crainiceau *et al.* 2004; Gelman 2006):

$$\begin{aligned}
&\mathbf{B}_1 \overset{\text{ind}}{\sim} N\left(0, A\right), A \overset{\text{ind}}{\sim} Uniform\left(0, U_1\right) \\
&\mathbf{B}_2 \overset{\text{ind}}{\sim} N_Q\left(0, \tau^2 \mathbf{I}_Q\right), \tau^2 \overset{\text{ind}}{\sim} Uniform\left(0, U_2\right) \\
&\mathbf{e} \overset{\text{ind}}{\sim} N_n\left(0, \sigma^2 \mathbf{V}_e\right), \mathbf{V}_e = diag\left(\mathbf{x}_i^{2k}\right), \sigma^2 \overset{\text{ind}}{\sim} Uniform\left(0, U_3\right)
\end{aligned} , \tag{2.39}$$

where U_1, U_2, U_3 are appropriate upper boundaries for the variance components (Gelman 2006) The *p*-spline estimator for the finite population total is

$$\begin{aligned}
\hat{T}_{psp} &= \sum\nolimits_{i \in U} \hat{m}_i + \sum\nolimits_{i \in s} \frac{y_i - \hat{m}_i}{\pi_i} \\
&= \sum\nolimits_{i \in s} \frac{y_i}{\pi_i} + \sum\nolimits_{i \in U} \hat{m}_i - \sum\nolimits_{i \in s} \frac{\hat{m}_i}{\pi_i} \\
&= \hat{T}_{HT} + \sum\nolimits_{i \in U} \left(\hat{\mathbf{B}}_1^T \mathbf{x}_i + \hat{\mathbf{B}}_2^T \mathbf{z}_i\right) - \sum\nolimits_{i \in s} \frac{\left(\hat{\mathbf{B}}_1^T \mathbf{x}_i + \hat{\mathbf{B}}_2^T \mathbf{z}_i\right)}{\pi_i} \\
&\equiv \hat{T}_{HT} + \hat{\mathbf{B}}_1^T \left(T_{\mathbf{X}} - \hat{T}_{HT\mathbf{X}}\right) + \hat{\mathbf{B}}_2^T \left(T_{\mathbf{Z}} - \hat{T}_{HT\mathbf{Z}}\right)
\end{aligned} , \tag{2.40}$$

where r denotes the $N - n$ units in the population but not the sample, $\hat{T}_{HT}$ is the total estimated using the base (HT) weights, $\left(T_{\mathbf{X}} - \hat{T}_{HT\mathbf{X}}\right)$, $\left(T_{\mathbf{Z}} - \hat{T}_{HT\mathbf{Z}}\right)$ are the differences in the known population totals of the polynomial and spline components in $\mathbf{x}, \mathbf{z}$ and the estimated totals using the base weights, and $\hat{\mathbf{B}}_1, \hat{\mathbf{B}}_2$ are the estimates of the model parameters. A conventional approach like weighted least squares could be used to estimate $\mathbf{B}_1, \mathbf{B}_2$ and residual maximum likelihood (REML) could be used to obtain an estimate of the model variances components. For example, if the model variances are fixed, then the posterior distribution of $\mathbf{B}_1, \mathbf{B}_2$ (Krivobokova *et al.* 2008) is

$$\begin{bmatrix}\mathbf{B}_1 & \mathbf{B}_2\end{bmatrix} \sim MVN_{Q+p+1}\left(\left(\mathbf{c}^T\mathbf{c}+\mathbf{D}_\alpha\right)^{-1}\mathbf{c}^T\mathbf{y}_s, \sigma^2\left(\mathbf{c}^T\mathbf{c}+\mathbf{D}_\alpha\right)^{-1}\right), \tag{2.41}$$

where $\mathbf{c}=\begin{bmatrix}\mathbf{x} & \mathbf{z}\end{bmatrix}$, $\mathbf{D}_\alpha = diag\left(0,\ldots,0,\sigma^2/\tau^2,\ldots,\sigma^2/\tau^2\right)$ is a diagonal matrix with $p+1$ rows of zeroes and Q rows with the penalty $\alpha=\sigma^2/\tau^2$. A more sophisticated approach is to incorporate the priors such as those in (2.39). The posterior distributions of the variance component τ conditional on the data, flat Normal priors for $\mathbf{B}_1,\mathbf{B}_2$, and $IG(0.001, 0.001)$ and uniform priors for τ^2 are (using analogous results for the Zheng and Little estimator in Appendix b of Chen *et al.* 2010)

$$\tau\left|\mathbf{B}_1,\mathbf{B}_2\right. \sim IG\left(0.001+\frac{Q}{2}, 0.001+\frac{\left\|\mathbf{B}_2\right\|^2}{2}\right), \tag{2.42}$$

and

$$\tau\left|\mathbf{B}_1,\mathbf{B}_2\right. \sim IG\left(\frac{Q-1}{2}, \frac{\left\|\mathbf{B}_2\right\|^2}{2}\right), \tag{2.43}$$

respectively.

In summary, I propose extending the Breidt *et al*. model by incorporating priors for the unknown model parameters. I propose this to produce a more efficient yet robust estimator of the finite population total. Incorporating the covariates and HT weights in model (2.39) should produce improved estimates of the total. Either prior distributions or a REML-type approach can be used to estimate the unknown parameters in (2.39). However, following Zheng and Little (2003), incorporating priors can guarantee non-negative variance component estimates. Here I propose simple priors in (2.39), focus on the simple model, and evaluate its potential performance in a simulation study against the corresponding simple (but separate) weight smoothing and weight pooling models.

2.3.2. Variance Estimation

The GREG AV (Linearization Variance Estimator)

Since the model-assisted *p*-spline estimator falls into the general class of calibration estimators, the anticipated variance of the GREG estimator can be used to approximate the variance (Särndal *et al*. 1992). A linearization of the GREG estimator (Exp. 6.6.9 in Särndal *et al*. 1999) for the proposed *p*-spline estimator is

$$\begin{aligned}
\hat{T}_{GREG} &\doteq \hat{T}_{HT} + \left(\mathbf{T}_{\mathbf{X}} - \hat{\mathbf{T}}_{HT\mathbf{X}}\right)^T \mathbf{B}_1 + \left(\mathbf{T}_{\mathbf{Z}} - \hat{\mathbf{T}}_{HT\mathbf{Z}}\right)^T \mathbf{B}_2 \\
&= \sum\nolimits_{i \in s} \frac{y_i}{\pi_i} + \mathbf{T}_{\mathbf{X}}^T \mathbf{B}_1 + \mathbf{T}_{\mathbf{Z}}^T \mathbf{B}_2 - \sum\nolimits_{i \in s} \frac{\left(\mathbf{x}_i^T \mathbf{B}_1 + \mathbf{z}_i^T \mathbf{B}_2\right)}{\pi_i} \\
&= \mathbf{T}_{\mathbf{X}}^T \mathbf{B}_1 + \mathbf{T}_{\mathbf{Z}}^T \mathbf{B}_2 + \sum\nolimits_{i \in s} \frac{e_i}{\pi_i} \\
&= \mathbf{T}_{\mathbf{X}}^T \mathbf{B}_1 + \mathbf{T}_{\mathbf{Z}}^T \mathbf{B}_2 + \tilde{e}_U
\end{aligned} \tag{2.44}$$

where $\mathbf{T}_{\mathbf{X}}, \mathbf{T}_{\mathbf{Z}}$ is the known population totals of $\mathbf{x}$ and $\mathbf{z}$, $\hat{\mathbf{T}}_{HT\mathbf{X}}, \hat{\mathbf{T}}_{HT\mathbf{Z}}$ are the vector of HT estimators, $\mathbf{B}_1, \mathbf{B}_2$ is the population coefficients, $e_i = y_i - \mathbf{x}_i^T \mathbf{B}_1 - \mathbf{z}_i^T \mathbf{B}_2$ is the residual, $\mathbf{x}_i^T, \mathbf{z}_i^T$ are row vectors of the (fixed) polynomial and (random) spline components, respectively, $\tilde{e}_U = \sum\nolimits_{i \in s} \frac{e_i}{\pi_i}$ is the sum of the weighted unit-level residuals, and π_i is the probability of selection. The last (2.44) component has design-expectation

$$\begin{aligned}
E_\pi\left(\sum\nolimits_{i \in s} \frac{e_i}{\pi_i}\right) &= \frac{1}{n} \sum\nolimits_{i \in U} \frac{E_\pi(\delta_i) e_i}{\pi_i} \\
&= \sum\nolimits_{i \in U} e_i \qquad , \\
&= E_U
\end{aligned} \tag{2.45}$$

where $E_U = \sum\nolimits_{i \in U} e_i$. From (2.44), $\hat{T}_{GREG} - \mathbf{T}_{\mathbf{X}}^T \mathbf{B}_U - \mathbf{T}_{\mathbf{Z}}^T \mathbf{B}_2 \doteq \sum\nolimits_{i \in s} \frac{e_i}{\pi_i}$, with design-variance (Exp. 6.6.3 in Särndal *et. al* 1992)

$$Var_\pi\left(\hat{T}_{GREG} - \mathbf{T}_\mathbf{X}^T\mathbf{B}_U - \mathbf{T}_\mathbf{Z}^T\mathbf{B}_2\right) = Var_\pi\left(\sum_{i\in s}\frac{e_i}{\pi_i}\right) \\ = \sum_{i\in U}\sum_{j\in U}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_i\pi_j}\right)\frac{e_ie_j}{\pi_i\pi_j} . \qquad (2.46)$$

For the p-spline estimator, the linearization-based GREG AV uses the residual $e_i = y_i - \left(\mathbf{x}_i^T\hat{\mathbf{B}}_1 + \mathbf{z}_i^T\hat{\mathbf{B}}_2\right)$ in (2.46). We estimate (2.46) with

$$var_\pi\left(\hat{T}_{GREG} - \mathbf{T}_\mathbf{X}^T\mathbf{B}_U - \mathbf{T}_\mathbf{Z}^T\mathbf{B}_2\right) = \sum_{i\in s}\sum_{j\in s}\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_i\pi_j}\right)\left(\frac{g_ie_i}{\pi_i}\frac{g_je_j}{\pi_j}\right), \qquad (2.47)$$

where the “g-weight” is $g_i = \frac{1}{\pi_i} + \sum_{j\in U}\left(1 - \frac{I_j}{\pi_j}\right)\mathbf{x}_j^T\mathbf{G}_\alpha e_i$.

Alternatively, we can estimate the GREG AV (2.46) with

$$var_\pi\left(\hat{T}_{GREG} - \mathbf{T}_\mathbf{X}^T\mathbf{B}_U - \mathbf{T}_\mathbf{Z}^T\mathbf{B}_2\right) \doteq \left(1 - \frac{n}{N}\right)\sum_{i\in s}\frac{g_i^2e_i^2}{\pi_i} . \qquad (2.48)$$

The variance estimator (2.48) uses a with-replacement variance and a finite population correction adjustment factor $1 - \frac{n}{N}$ to approximately account for without-replacement sampling (see, e.g., Valliant 2002).

The Delete-a-Group Jackknife Variance Estimator

As introduced in Quenoulle (1949; 1956), the jackknife method has been used in both finite and infinite population inference (Shao and Wu 1989). In the “delete-a-group” jackknife, the data are first grouped in some way. The most common method of assigning group membership of the n to the G groups is completely random. The group jackknife has also been shown to perform best (i.e., have minimum bias) when the groups have equal size (Valliant *et. al* 2008). When all units within a particular group g are "dropped," their weights are set equal to zero, then the weights for all the other units within the same stratum are adjusted. Weights within the groups are unchanged and the sample-based estimate is computed, denoted $\hat{T}_g = G\hat{T} + (G-1)\hat{T}_{(g)}$, where $\hat{T}_{(g)}$ is the total computed from the sample of reduced $\frac{n(G-1)}{G}$ units. This process is continued through all of the groups and G estimates are obtained. One version of the jackknife variance estimator (Rust 1985) is the variance of the replicate group totals across groups:

$$Var_J\left(\hat{T}\right) = \frac{G}{G-1}\sum_{g=1}^{G}\left(\hat{T}_{(g)} - \hat{T}\right)^2 . \tag{2.49}$$

For the *p*-spline estimator, Zheng and Little (2005) advocate not estimating the parameter α within each replicate, but using the full-sample based estimate and replicate-based estimates of the other model parameters. If viewing the component $\hat{\alpha}$ as fixed, as described earlier, then the *p*-spline estimator is equivalent to Chambers' ridge regression estimator (Ex. 2.6). Specifically, they advocate not computing $\mathbf{D}_{\hat{\alpha}(g)}$ for each replicate if this is "burdensome," provided that the overall sample size is very large and the portion of units being omitted is not large. In this case, the full-sample estimate $\mathbf{D}_{\hat{\alpha}}$ can be used, which does not greatly impact the jackknife consistency theory.

Zheng and Little (2005) also prove that the delete-one-unit jackknife variance estimation method is appropriate for the *p*-spline regression whenever the jackknife variance estimator is appropriate for simple linear regression of splines. If the *p*-spline is a low-dimensional smoother, then the dimension of the design matrix $\mathbf{W}^{1/2}\mathbf{X}$ is small relative to the sample size; under these conditions, they show that the delete-one-unit jackknife variance estimator for estimating the variance of the *p*-spline regression estimator has asymptotic properties that are close to the jackknife variance estimator for a linear spline regression estimator.

Here, related theory (Lemma 4.3.3, p.166 in Wolter 2007) states that the delete-one-group jackknife variance estimator is consistent in pps samples when equal-sized groups are used. This is used in the variance estimation evaluation in Sec. 2.4.3. For stratified sampling, Wolter (2007) also provides the theory and necessary conditions for the jackknife to be consistent in infinite (Thm. 4.2.2) and unbiased in finite (p. 176) populations, as well as general results for nonlinear estimators (Sec. 4.4). This is relevant since the model-assisted *p*-spline estimators fall under the general class of nonlinear estimators. Related theory for the jackknife can be found in Rao and Wu (1988) and Krewski and Rao (1981).

Model-based Variance Estimator

The model (2.36) can be used to estimate the variance of the total (2.40). Following Zheng and Little (2005), the Empirical Bayes (EB) posterior variance of $\mathbf{B} = \left[\mathbf{B}_1 \quad \mathbf{B}_2\right]$ conditional on $\hat{\sigma}^2$ and $\hat{\alpha} = \hat{\sigma}^2 / \hat{\tau}^2$, is given by

$$Var\left(\mathbf{B} \middle| \hat{\alpha}, \mathbf{x}, \mathbf{z}, \mathbf{y}\right) = \hat{\sigma}^2 \left(\mathbf{c}_s^T \mathbf{c}_s + \hat{\mathbf{D}}_\alpha\right)^{-1}, \tag{2.50}$$

where $\mathbf{c}_s = \left[\mathbf{x}_s \quad \mathbf{z}_s\right]$ contains the values of $\mathbf{x}$ and $\mathbf{z}$ in the sample and $\hat{\sigma}^2$ is the posterior estimate of σ^2.

To estimate the variance of the total, following the linearization approach for the GREG AV from (2.44) and using the variance of the model parameters in (2.50), we have

$$\begin{aligned}\hat{T}_{psp} &= \hat{T}_{HT} + \left(\mathbf{T_X} - \hat{\mathbf{T}}_{HT\mathbf{X}}\right)^T \hat{\mathbf{B}}_1 + \left(\mathbf{T_Z} - \hat{\mathbf{T}}_{HT\mathbf{Z}}\right)^T \hat{\mathbf{B}}_2 \\ &\doteq \hat{T}_{HT} + \left(\mathbf{T_X} - \hat{\mathbf{T}}_{HT\mathbf{X}}\right)^T \mathbf{B}_1 + \left(\mathbf{T_Z} - \hat{\mathbf{T}}_{HT\mathbf{Z}}\right)^T \mathbf{B}_2 \\ &= \sum_{i\in s} \frac{y_i}{\pi_i} + \mathbf{T}_\mathbf{X}^T \mathbf{B}_1 + \mathbf{T}_\mathbf{Z}^T \mathbf{B}_2 - \sum_{i\in s} \frac{\left(\mathbf{x}_i^T \mathbf{B}_1 + \mathbf{z}_i^T \mathbf{B}_2\right)}{\pi_i}\end{aligned} \quad . \tag{2.51}$$

Therefore,

$$\begin{aligned}\hat{T}_{psp} - \mathbf{T}_\mathbf{X}^T \mathbf{B}_1 - \mathbf{T}_\mathbf{Z}^T \mathbf{B}_2 &\doteq \sum_{i\in s} \frac{y_i}{\pi_i} - \sum_{i\in s} \frac{\left(\mathbf{x}_i^T \mathbf{B}_1 + \mathbf{z}_i^T \mathbf{B}_2\right)}{\pi_i} \\ &= \sum_{i\in s} \frac{1}{\pi_i}\left(y_i - \mathbf{x}_i^T \mathbf{B}_1 - \mathbf{z}_i^T \mathbf{B}_2\right) \\ &= \sum_{i\in s} \frac{e_i}{\pi_i}\end{aligned} \quad . \tag{2.52}$$

If we rewrite the model coefficient estimator $\hat{\mathbf{B}} = \left[\hat{\mathbf{B}}_1 \quad \hat{\mathbf{B}}_2\right]$ in (2.32) as

$$\begin{aligned}\hat{\mathbf{B}} &= \left(\mathbf{c}_s^T \mathbf{c}_s + \hat{\mathbf{D}}_\alpha\right)^{-1} \mathbf{c}_s^T \mathbf{y}_s \\ &= \mathbf{A}_s^{-1} \mathbf{c}_s^T \mathbf{y}_s\end{aligned} , \tag{2.53}$$

where $\mathbf{A}_s = \mathbf{c}_s^T \mathbf{c}_s + \hat{\mathbf{D}}_\alpha$, then the estimated total can be written as

$$\hat{T}_{psp} = \left(\mathbf{d}^T + \mathbf{A}_s^{-1} \mathbf{c}_s^T\right) \mathbf{y}_s , \tag{2.54}$$

where $\mathbf{d}^T = \left(\pi_1^{-1}, \ldots, \pi_n^{-1}\right)$. The variance of the total is then

$$Var_M\left(\hat{T}_{psp}\right) = \left(\mathbf{d}^T + \mathbf{A}_s^{-1} \mathbf{c}_s^T\right) Var_M\left(\mathbf{y}_s\right)\left(\mathbf{d} + \mathbf{c}_s^{-1} \mathbf{A}_s\right), \tag{2.55}$$

We can estimate the variance (2.55) with

$$\begin{aligned}var_M\left(\hat{T}_{psp}\right) &= \left(\mathbf{d}^T + \mathbf{A}_s^{-1} \mathbf{c}_s^T\right) var_M\left(\mathbf{y}_s\right)\left(\mathbf{d} + \mathbf{c}_s^{-1} \mathbf{A}_s\right) \\ &= \left(\mathbf{d}^T + \mathbf{A}_s^{-1} \mathbf{c}_s^T\right) \mathbf{e}\mathbf{e}^T \left(\mathbf{d} + \mathbf{c}_s^{-1} \mathbf{A}_s\right)\end{aligned} , \tag{2.56}$$

where $\mathbf{e} = (\hat{e}_1, \ldots, \hat{e}_n)^T = \left(y_1 - \mathbf{x}_1^T \hat{\mathbf{B}}_1 - \mathbf{z}_1^T \hat{\mathbf{B}}_2, \ldots, y_n - \mathbf{x}_n^T \hat{\mathbf{B}}_1 - \mathbf{z}_n^T \hat{\mathbf{B}}_2\right)^T$ are the residuals. Since the EB posterior variance does not account for variability incurred due to using estimates of σ^2 and $\alpha = \sigma^2 / \tau^2$, it can underestimate the variance of $\hat{T}_{psp}$. However, Ruppert and Carroll (2000) demonstrate that this does not seriously bias the variance estimation.

2.4. Evaluation Studies

2.4.1. Alternative Estimators of Totals

This section contains a simulation study to illustrate how the proposed *p*-spline smoothing model performs against some design-based and model-assisted alternatives, as well as Zheng and Little's and Breidt *et. al*'s model.

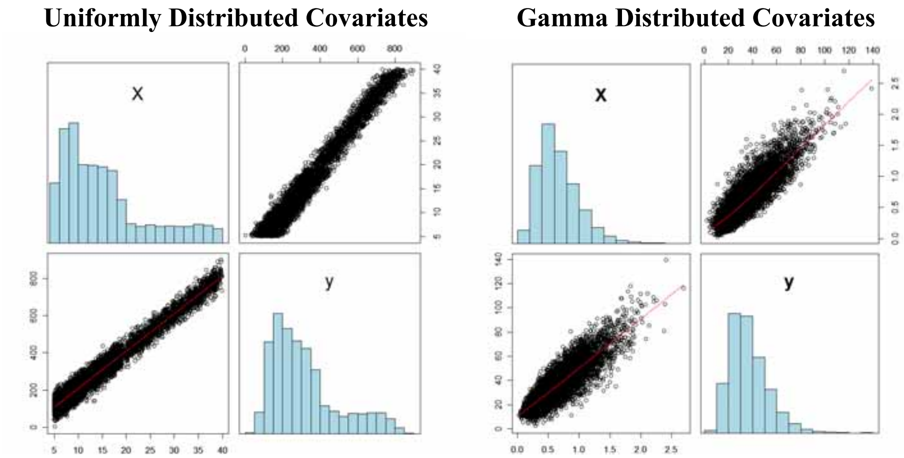
This simulation study has three factors, the covariate structure in the pseudo population (with two levels), the design used to select the sample and create the weighting strata (two), and estimator of the total (seven). First, data generated for this simulation study is a smaller version of that done by Elliott and Little (2000). Here, a population of 8,300 units (with $N_h = (800, 1000, 1500, 2000, 3000)$) is generated from the model $y_{hi} = \alpha + \beta X_{hi} + e_i$, where $e_{hi} \overset{\text{ind}}{\sim} N(0,10)$ and the X_{hi}'s follow one of two patterns:

- Uniform within strata: $X_{1i} \overset{\text{ind}}{\sim} Unif(20,40), i = 1, \ldots, 800$;
$X_{2i} \overset{\text{ind}}{\sim} Unif(19,39), i = 1, \ldots, 1000$; $X_{3i} \overset{\text{ind}}{\sim} Unif(15,19), i = 1, \ldots, 1500$;
$X_{4i} \overset{\text{ind}}{\sim} Unif(10,15), i = 1, \ldots, 2000$; $X_{5i} \overset{\text{ind}}{\sim} Unif(5,10), i = 1, \ldots, 3000$.

- Common gamma distribution across strata: $X_{hi} \overset{\text{ind}}{\sim} Gamma(3,4), h = 1, \ldots, 5$.

The first population corresponds to a situation when trimming the weights is sensible, while the second corresponds to a situation when it is not as necessary. The population plots are given in Figure 2.1.

Figure 2.1. Population Plots and Loess Lines for Simulation Comparing Alternative Totals



Two hundred samples of size 350were drawn without replacement and the alternative models to estimate the means were applied for each of the possible pooling patterns. For the sample design factor, two types of samples are drawn from each pseudopopulation. First, stratified simple random sampling with five fixed strata is used with $n_h = (90, 80, 70, 60, 50)$ and the weight pooling cutpoint being fixed at the upper two strata. Second, probability proportional to X sampling is used, and an adhoc design-based trimming method (where weights exceeding the 95th quantile of the weights are trimmed to this value). In both sample designs, the excess weight is redistributed equally to the non-trimmed weights. For the third factor, seven alternative estimators of the finite population total are compared:

- The un-trimmed Horvitz-Thompson (HT) estimator: $\hat{T}_{HT} = \sum_{i \in s} w_i y_i$, where $w_i = N_h / n_h$ for stratified simple random sampling and $w_i = \sum_{i \in U} x_i / n x_i$ for pps;

- The fully-trimmed estimator: $\hat{T}_{FT} = \frac{N}{n}\sum_{i \in s} y_i$;

- A design-based trimming estimator (i.e., the weight pooled estimator with fixed cutpoint stratum): $\hat{T}_{WP} = \sum_{h<l}\sum_{i \in s_h} w_i y_i + \sum_{h \geq l} \bar{w} \sum_{i \in s_h} y_i$, where $\bar{w} = \left(N_l + N_{l+1}\right) / \left(n_l + n_{l+1}\right)$ is the combined stratum weight in the upper two weight strata for stratified sampling and the 95th weight quantile for pps sampling (in both designs, the weight above the cutoff was redistributed to non-trimmed weights);

- The calibration estimator with a common linear, quadratic, and cubic polynomial model fit across all strata;

- The Breidt *et al.* model-assisted estimator, with quadratic and cubic (second- and third-degree) *p*-splines, $Q = 5,10,15$ knots as the appropriate sample quantiles of x_{hi}. The estimated total has components given in (2.33), is ;

- Zheng and Little's estimator (2.40), with quadratic and cubic (second- and third-degree) *p*-splines, $Q = 5,10,15$ knots as the appropriate sample quantiles of π_i ., and the following priors for the unknown model parameters:

$$\begin{aligned}
&\mathbf{B}_1 \overset{\text{ind}}{\sim} flat(\ ), \\
&\mathbf{B}_2 \overset{\text{ind}}{\sim} N_Q\left(0, \tau^2 \mathbf{I}_Q\right), \tau^2 \overset{\text{ind}}{\sim} Unif\left(0,100000\right), \\
&\mathbf{e} \overset{\text{ind}}{\sim} N\left(0, \sigma^2 \mathbf{I}_n\right), \sigma^2 \overset{\text{ind}}{\sim} Unif\left(0,100000\right)
\end{aligned} \tag{2.57}$$

where *flat( )* denotes a noninformative uniform prior. I experimented with also using the inverse gamma (IG) prior for the variance components in (2.57). However, the IG prior results were omitted due to convergence-related problems in the MCMC samples. The total is then $\hat{T}_{ZL} = \sum_{i \in s} y_i + \mathbf{x}_r^T \hat{\mathbf{B}}_1 + \mathbf{z}_r^T \hat{\mathbf{B}}_2$, where $\mathbf{x}_r$ and $\mathbf{z}_r$ are the non-sample components of $\mathbf{x}$ and $\mathbf{z}$ in (2.34);

- The proposed *p*-spline estimator with quadratic (second-degree), with $Q = 5,10,15$ knots being the sample quantiles of x_i, and the same priors as those given in (2.57). The model simplifies to

$$y_{hi} = \beta_0 + \beta_1 x_{hi} + \beta_2 x_{hi}^2 + \sum_{q=1}^{Q} \beta_{(2+q)} \left(x_{hi} - \kappa_q\right)_+^2 + e_{hi} . \tag{2.58}$$

The estimated total is $\hat{T}_{psp} = \hat{T}_{HT} + \left(T_{\mathbf{X}} - \hat{T}_{HT\mathbf{X}}\right)^T \hat{\mathbf{B}}_1 + \left(T_{\mathbf{Z}} - \hat{T}_{HT\mathbf{Z}}\right)^T \hat{\mathbf{B}}_2$, where the components are given in (2.40). I also experimented with the cubic *p*-spline version of (2.58) using the model $y_{hi} = \beta_0 + \beta_1 x_{hi} + \beta_2 x_{hi}^2 + \beta_3 x_{hi}^3 + \sum_{q=1}^{Q} \beta_{(3+q)} \left(x_{hi} - \kappa_q\right)_+^3 + e_{hi}$. However, some of the results were unstable, and thus were also omitted. Despite this, the simulation evaluation here is a reasonable comparison of existing alternatives against the proposed estimator.

For the pps and stratified samples, one model is fit to all the data for all alternatives. For the Zheng and Little estimators, 250 samples using 200 MCMC samples for each estimator were drawn, with the first fifty samples being the burn-in, and thus disregarded in estimation of the model parameters. For the proposed estimators, fifty samples using 10,500 MCMC samples for each estimator were drawn, with the first 500 samples for the burn-in. The model convergence of three MCMC chains was assessed by examining plots of the generated posterior distribution data. For example, Figure 2.2 (below and on the following page) shows an example of the first sample results for the τ^2 and σ^2 variance components of the three chains for the proposed quadratic model with 10 knots, respectively.

Figure 2.2. One-Sample Model Convergence for τ^2 and σ^2-Parameter, Proposed Quadratic Model, 10 Knots: Posterior Density Plots

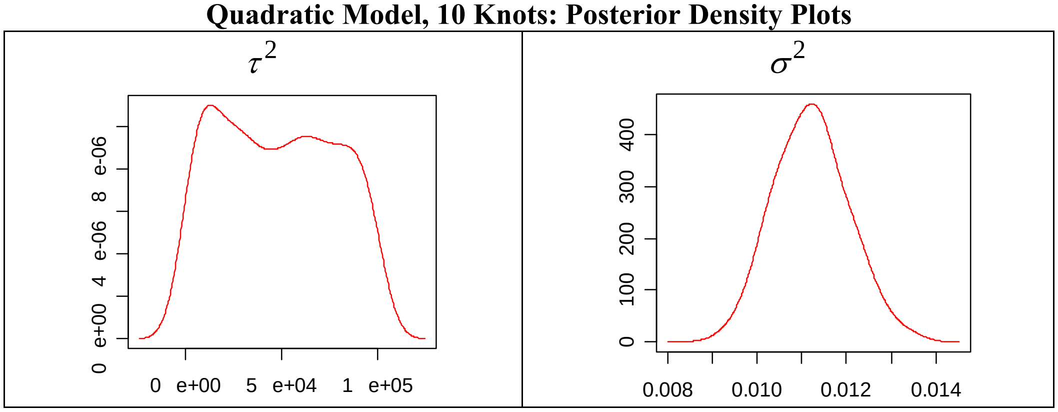
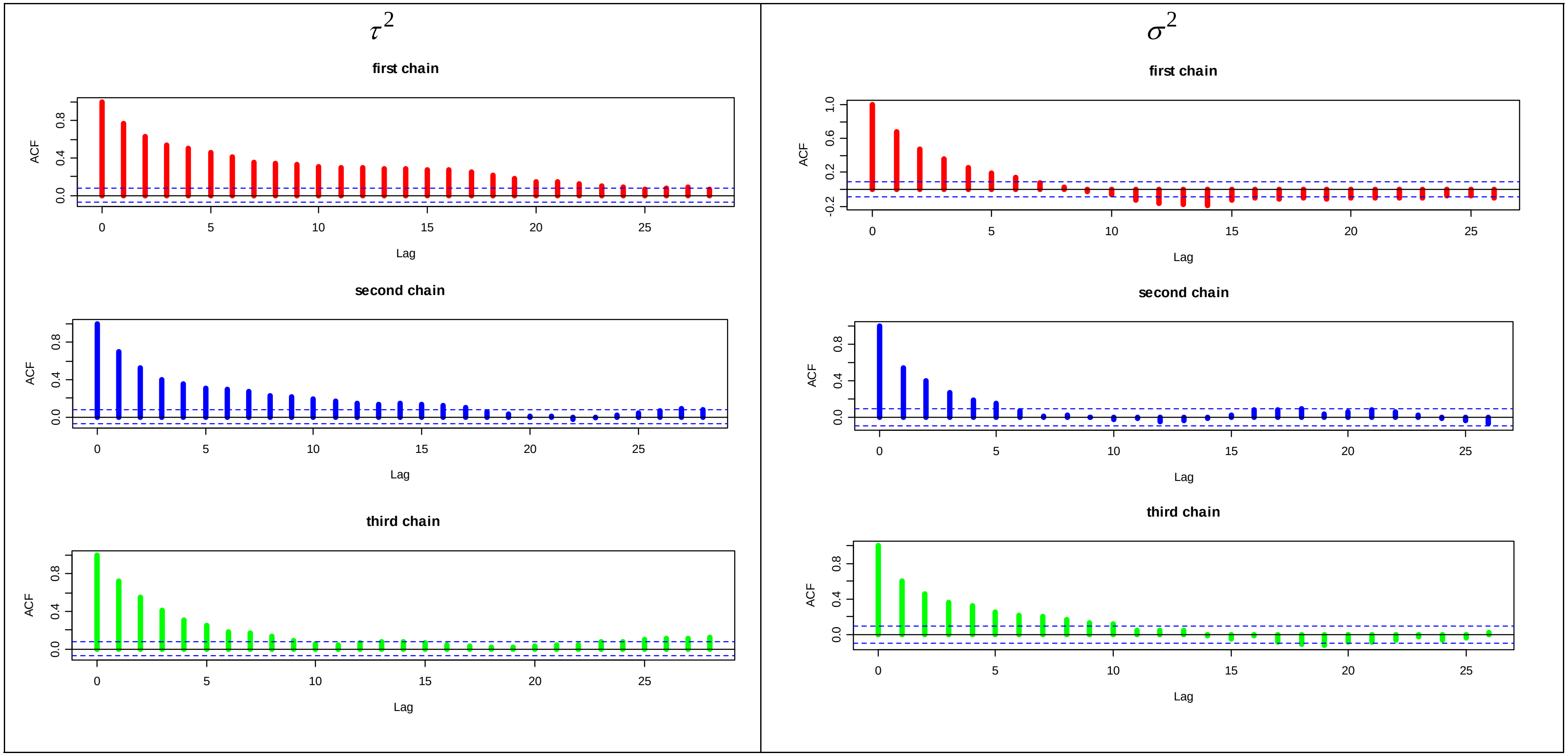


Figure 2.2. One-Sample Model Convergence for τ^2 and σ^2-Parameter, Proposed Quadratic Model, 10 Knots, cont'd: AutoCorrelation Plots



From Figure 2.2, we see first that from the posterior density plots, both estimates of the variance components (the means of the densities) are positive. In this particular sample, $\hat{\sigma}^2 = 0.011$ and $\hat{\tau}^2 = 47{,}550$, such that the penalty applied to the knots is very large ($\hat{\alpha} = \hat{\sigma}^2 / \hat{\tau}^2 \approx 2.36 \times 10^{-7}$ is small in magnitude, but this value means there is a lot of smoothing on the spline component coefficients). The shape of the τ^2 posterior distribution more closely resembles a uniform distribution, the prior assumed for this parameter, while the posterior for σ^2 is more symmetric. This suggests that more data is required to estimate τ^2. For the autocorrelation plots, we should see low autocorrelations as the number of lags increases for each chain, i.e., autocorrelations that fall within the dotted 95% confidence lines. This occurs more quickly for σ^2 than τ^2, where the autocorrelations fall between the boundaries after 20 lags.

I use five summary measures to compare the alternative totals:

- *Relative bias*: the percentage of the average distance between $\hat{T}_b$, an alternative estimator for the total of y obtained on iteration $b = 1, \ldots, B$ $B = 50$ or 200, and population total T, relative to the total: $RelBias\left(\hat{T}\right) = 100 \times \left(BT\right)^{-1} \sum_{b=1}^{B}\left(\hat{T}_b - T\right)$,

- *Variance Ratio*: the ratio of the empirical variance of an alternative total to that of the untrimmed HT estimator: $VarRatio\left(\hat{T}\right) = \dfrac{\sum_{b=1}^{B}\left(\hat{T}_b - \hat{\bar{T}}\right)^2}{\sum_{b=1}^{B}\left(\hat{T}_{HTb} - \hat{\bar{T}}_{HT}\right)^2}$, for $\hat{\bar{T}} = B^{-1}\sum_{b=1}^{B}\hat{T}_b$ and $\hat{\bar{T}}_{HT} = B^1 \sum_{b=1}^{B}\hat{T}_{HTb}$.

- *Empirical RMSE*: the mean square error of the alternative estimator, relative to that of the fully weighted estimator: $RelRMSE\left(\hat{T}\right) = \dfrac{\sum_{b=1}^{B}\left(\hat{T}_b - T\right)^2}{\sum_{b=1}^{B}\left(\hat{T}_{HTb} - T\right)^2}$.

- *95% CI Coverage rate*: the percentage of the 200 simulated confidence intervals that contain the true population total: $\left|\hat{T}-T\right| \Big/ \sqrt{B^{-1}\sum_{b=1}^{B}\left(\hat{T}_b-\hat{\bar{T}}\right)^2} \le z_{\alpha/2} = 1.96$. Note that this uses the empirical variance across the simulations in the denominator, not an estimate from each sample.

- *Average CI width*: the average width of the 95% CI's (note this also uses the empirical variance, not a variance estimator): $2(1.96)B^{-1}\sum_{b=1}^{B}\sqrt{B^{-1}\sum_{b=1}^{B}\left(\hat{T}_b-\hat{\bar{T}}\right)^2}$.

The evaluation measures are summarized for each sample design and estimator combination for the population with covariates that are uniformly distributed within each stratum in Tables 2.1 and 2.2.

Table 2.1. Simulation Results for Population with Uniformly Distributed Covariates, Stratified SRS Samples (T = 2,609,793)

Estimator	*Relative Bias (%)*	*Variance Ratio*	*RelRMSE*	*CI Coverage*	*Ave. CI Width*
Design-based Estimators (200 samples)					
Fully weighted (HT)	-0.05	1.00	1.00	95.5	99,982
Fully smoothed (FS)	36.72	2.00	1,409.24	0.0	141,564
Trimmed/Weight pooling (WP)	2.78	0.98	9.02	16.5	99,101
Model-Assisted Estimators (200 samples)					
Calibration					
Linear (CalL)	0.04	0.42	0.42	95.5	71,656
Quadratic (CalQ)	0.04	0.42	0.42	95.5	71,888
Cubic (CalC)	0.03	0.42	0.42	95.5	71,571
Breidt *et al.* model (REML)					
Quadratic model: 5 knots (Q5)	0.03	0.43	0.43	95.0	72,043
10 knots (Q10)	0.04	0.44	0.44	95.0	72,897
15 knots (Q15)	0.04	0.44	0.44	95.0	73,563
Cubic model: 5 knots(C5)	0.03	0.43	0.43	95.5	72,423
10 knots (C10)	0.03	0.44	0.44	95.5	73,180
15 knots (C15)	0.04	0.45	0.45	94.5	73,922
Zheng and Little Estimators (200 samples)					
Uniform prior on variance components					
Quadratic model: 5 knots (Q5)	-1.66	0.95	3.84	60.5	97,366
10 knots (Q10)	-0.84	1.00	1.73	88.0	99,870
15 knots (Q15)	-1.35	0.90	2.80	72.5	94,714
Cubic model: 5 knots(C5)	-0.18	3.73	3.75	96.0	193,014
10 knots (C10)	0.49	2.30	2.54	93.5	151,589
15 knots (C15)	0.98	8.17	9.15	93.5	285,801
Proposed P-spline Estimators (50 samples)					
Uniform prior on variance components					
Quadratic model: 5 knots (Q5)	-0.94	81.10	80.43	94.0	942,556
10 knots (Q10)	0.00	0.43	0.42	98.0	68,413
15 knots (Q15)	0.09	0.42	0.42	92.0	68,136

Table 2.2. Simulation Results for Population with Uniformly Distributed Covariates, *ppswor* Samples ($T = 2{,}609{,}793$)

Estimator	*Relative Bias (%)*	*Variance Ratio*	*RelRMSE*	*CI Coverage*	*Ave. CI Width*
Design-based Estimators (200 samples)					
Fully weighted (HT)	-0.02	1.00	1.00	95.5	62,020
Fully smoothed (FS)	32.37	4.70	2,855.19	0.0	134,471
Trimmed/Weight pooling (WP)	1.62	1.32	8.92	32.5	71,343
Model-Assisted Estimators (200 samples)					
Calibration					
Linear (CalL)	-0.00	1.00	1.00	95.0	64,521
Quadratic (CalQ)	-0.00	0.97	0.97	95.0	63,724
Cubic (CalC)	-0.00	0.98	0.98	95.0	63,960
Breidt *et al.* model (REML)					
Quadratic model: 5 knots (Q5)	0.00	0.99	0.99	95.5	64,145
10 knots (Q10)	-0.01	0.98	0.98	95.5	63,802
15 knots (Q15)	0.00	0.99	0.99	95.0	64,283
Cubic model: 5 knots(C5)	0.00	0.99	0.99	95.5	64,117
10 knots (C10)	-0.01	0.97	0.97	95.0	63,748
15 knots (C15)	0.00	0.99	0.99	94.5	64,179
Zheng and Little Estimators (200 samples)					
Uniform prior on variance components					
Quadratic model: 5 knots (Q5)	-0.15	0.95	1.01	93.0	60,397
10 knots (Q10)	0.04	0.96	0.97	96.0	60,881
15 knots (Q15)	-0.09	0.92	0.94	93.5	59,459
Cubic model: 5 knots(C5)	-0.10	1.69	1.72	96.0	80,705
10 knots (C10)	0.10	1.19	1.21	95.5	67,545
15 knots (C15)	0.30	2.37	2.61	95.5	95,508
Proposed P-spline Estimators (50 samples)					
Uniform prior on variance components					
Quadratic model: 5 knots (Q5)	-1.44	580.22	583.79	90.0	1,229,573
10 knots (Q10)	0.19	2.11	2.24	96.7	74,161
15 knots (Q15)	0.09	2.84	2.85	96.7	86,078

For the population with x_i-values that are uniformly distributed with different means across strata, the untrimmed HT estimator is unbiased across the 200 samples and has nominal 95 percent coverage. The fully smoothed estimator, where all sample units are given the common weight 8,300/350, is severely positively biased, which drives the root mean square error to be among the largest among the alternative estimators. The trimmed weight estimator is biased, but more efficient than the untrimmed HT. For the model-assisted estimators, the calibration estimators are essentially unbiased and have the lowest empirical variance, producing significantly lower –by more than half– the

RMSE's relative to the HT estimator. The proposed *p*-spline estimator with uniform priors also performed relatively well, very comparable to the model-assisted estimators. The estimator with 10 and 15 knots were among the estimators with the lowest RMSE's in the stratified samples. Generally, estimators with larger bias have lower (including zero) confidence interval coverage, while the more inefficient estimators produced confidence intervals with larger average width.

However, while the proposed estimators are comparable to the other alternatives, they appear to be sensitive to the number of model components. In results that are not shown, we fitted *p*-spline models with cubic polynomials and fit all of the proposed models within strata. The bias and variances increased dramatically as the number of model terms –both number of knots and polynomial degree– increased. The Zheng and Little estimator using the probabilities of selection as covariates did not suffer as much from this "curse of dimensionality." I hypothesize this did not occur since the Zheng and Little model covariates, the probabilities of selection, are bounded within (0,1). This bounding leads to generally more stable knot components in the models. However, the slight inefficiency in the Zheng-Little cubic model totals in Tables 2.1 and 2.2 may be caused by the smaller number of MCMC samples (200).

Results for the population with x_i-values following a common Gamma distribution across strata are shown in Tables 2.3 and 2.4. For these results, again the untrimmed HT estimator is unbiased across the 200 samples and has nominal 95 percent coverage. The fully smoothed estimator is still biased, but here it is negative biased. The trimmed weight estimator is also biased but more efficient than the untrimmed HT. The model-assisted calibration estimators performed the best, while the model-assisted *p*-

splines with lower ordered models performed adequately. The quadratic Zheng and Little estimators are slightly more biased, but slightly more efficient than the HT estimator; the cubic model totals are less biased but more inefficient. The calibration model-assisted estimators here also generally performed the best, but the proposed *p*-spline estimator again is very comparable. In this data, here there is less of a difference between the alternatives. As in the first population, biased estimators have lower confidence interval coverage, while inefficient estimators had larger average CI width.

Table 2.3. Simulation Results for Population with Gamma Distributed Covariates, Stratified SRS Samples ($T = 303{,}439$)

Estimator	*Relative Bias (%)*	*Variance Ratio*	*RelRMSE*	*CI Coverage*	*Ave. CI Width*
Design-based Estimators (200 samples)					
Fully weighted (HT)	-0.22	1.00	1.00	95.0	24,520
Fully smoothed (FS)	-20.61	0.30	100.00	0.0	13,366
Trimmed/Weight pooling (WP)	-4.03	0.70	4.51	34.5	20,504
Model-Assisted Estimators (200 samples)					
Calibration					
Linear (CalL)	-0.04	0.48	0.48	94.5	19,221
Quadratic (CalQ)	-0.03	0.47	0.47	93.5	18,991
Cubic (CalC)	0.01	0.48	0.48	94.5	19,169
Breidt *et al*. model (REML)					
Quadratic model: 5 knots (Q5)	0.06	0.53	0.53	94.5	20,166
10 knots (Q10)	0.06	0.63	0.64	94.5	22,111
15 knots (Q15)	0.14	0.87	0.87	95.5	25,875
Cubic model: 5 knots(C5)	0.06	0.53	0.53	94.5	20,193
10 knots (C10)	0.07	0.67	0.67	95.5	22,706
15 knots (C15)	0.16	0.97	0.97	95.5	27,276
Zheng and Little Estimators (200 samples)					
Uniform prior on variance components					
Quadratic model: 5 knots (Q5)	-2.00	0.86	1.79	81.0	22,767
10 knots (Q10)	-1.33	0.88	1.28	90.0	22,960
15 knots (Q15)	-1.75	0.86	1.57	87.5	22,750
Cubic model: 5 knots(C5)	-0.71	1.33	1.44	95.5	28,310
10 knots (C10)	-0.14	1.17	1.16	95.5	22,558
15 knots (C15)	0.17	1.70	1.69	93.5	31,967
Proposed P-spline Estimators (50 samples)					
Uniform prior on variance components					
Quadratic model: 5 knots (Q5)	-0.02	0.48	0.48	96.0	17,303
10 knots (Q10)	-0.02	0.48	0.48	96.0	17,301
15 knots (Q15)	-0.02	0.48	0.48	96.0	17,309

Table 2.4. Simulation Results for Population with Gamma Distributed Covariates, *ppswor* Samples ($T = 303{,}439$)

Estimator	*Relative Bias (%)*	*Variance Ratio*	*RelRMSE*	*CI Coverage*	*Ave. CI Width*
Design-based Estimators (200 samples)					
Fully weighted (HT)	-0.01	1.00	1.00	95.5	15,015
Fully smoothed (FS)	18.59	1.20	220.15	0.0	16,548
Trimmed/Weight pooling (WP)	2.85	1.30	6.43	49.5	17,129
Model-Assisted Estimators (200 samples)					
Calibration					
Linear (CalL)	0.11	0.99	0.99	95.0	14,707
Quadratic (CalQ)	0.11	0.99	0.99	95.0	14,695
Cubic (CalC)	0.10	0.98	0.98	95.0	14,642
Breidt *et al*. model (REML)					
Quadratic model: 5 knots (Q5)	0.11	0.98	0.98	95.0	14,622
10 knots (Q10)	0.12	0.96	0.96	95.0	14,280
15 knots (Q15)	0.11	0.99	0.99	95.0	14,706
Cubic model: 5 knots(C5)	0.10	0.96	0.96	95.0	14,495
10 knots (C10)	0.11	1.00	1.00	94.5	14,789
15 knots (C15)	0.12	1.00	1.01	95.0	14,817
Zheng and Little Estimators (200 samples)					
Uniform prior on variance components					
Quadratic model: 5 knots (Q5)	-0.52	1.06	1.23	94.0	15,473
10 knots (Q10)	0.16	1.08	1.09	94.0	15,576
15 knots (Q15)	-0.27	1.07	1.11	94.5	15,496
Cubic model: 5 knots(C5)	-0.28	1.09	1.14	93.5	15,699
10 knots (C10)	0.16	1.09	1.10	94.0	15,660
15 knots (C15)	0.76	1.64	2.00	90.5	19,207
Proposed P-spline Estimators (50 samples)					
Uniform prior on variance components					
Quadratic model: 5 knots (Q5)	0.04	0.96	0.97	96.0	15,997
10 knots (Q10)	0.04	0.96	0.96	96.0	15,992
15 knots (Q15)	0.04	0.98	0.96	96.0	15,980

Figures 2.3 through 2.6 show boxplot distributions of the alternative totals estimated from the stratified and pps samples drawn from each population. In each plot, the true population total is shown with a vertical line. In the figures, we see how the fully smoothed and weight smoothing totals are biased, while the proposed estimators with five knots have larger variances (note: the cubic Breidt *et. al* model results are omitted). Generally, all of the estimators have lower bias and variance for the population with the Gamma distributed covariates.

Figure 2.3. Boxplots of Estimated Totals, Population with Uniformly Distributed Covariates, Stratified Samples

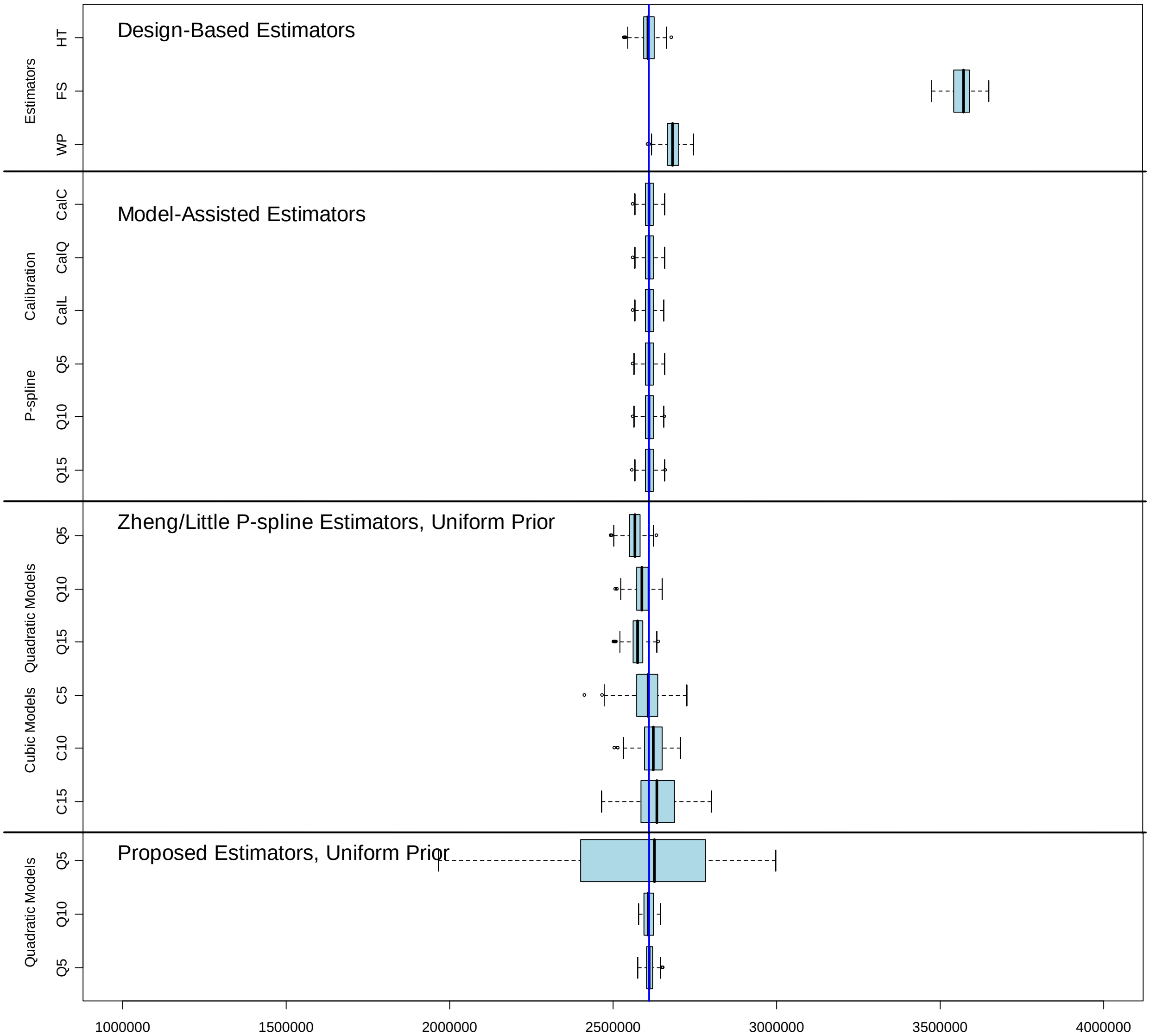


Figure 2.4. Boxplots of Estimated Totals, Population with Uniformly Distributed Covariates, *ppswor* Samples

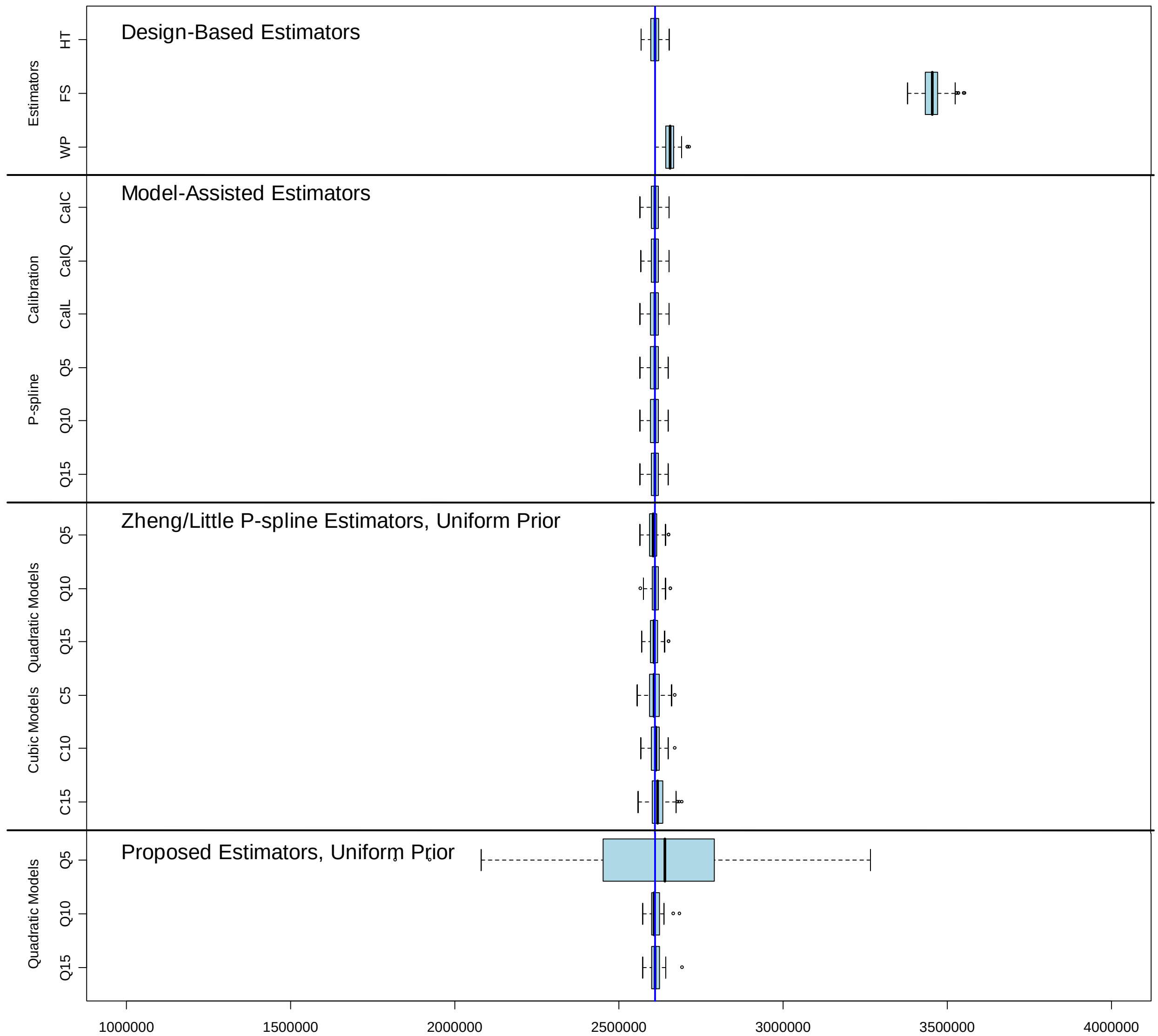


Figure 2.5. Boxplots of Estimated Totals, Population with Gamma Distributed Covariates, Stratified Samples

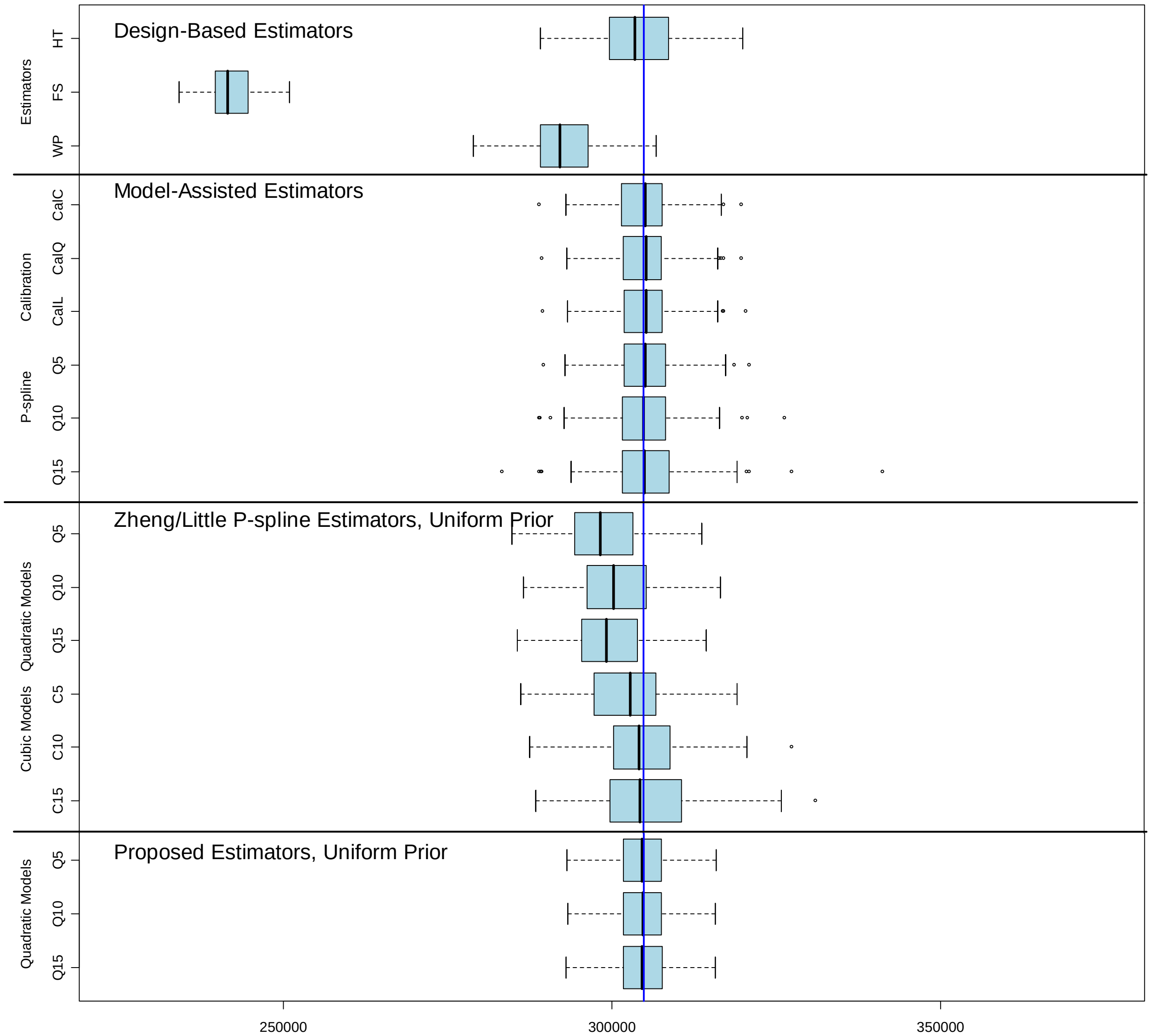
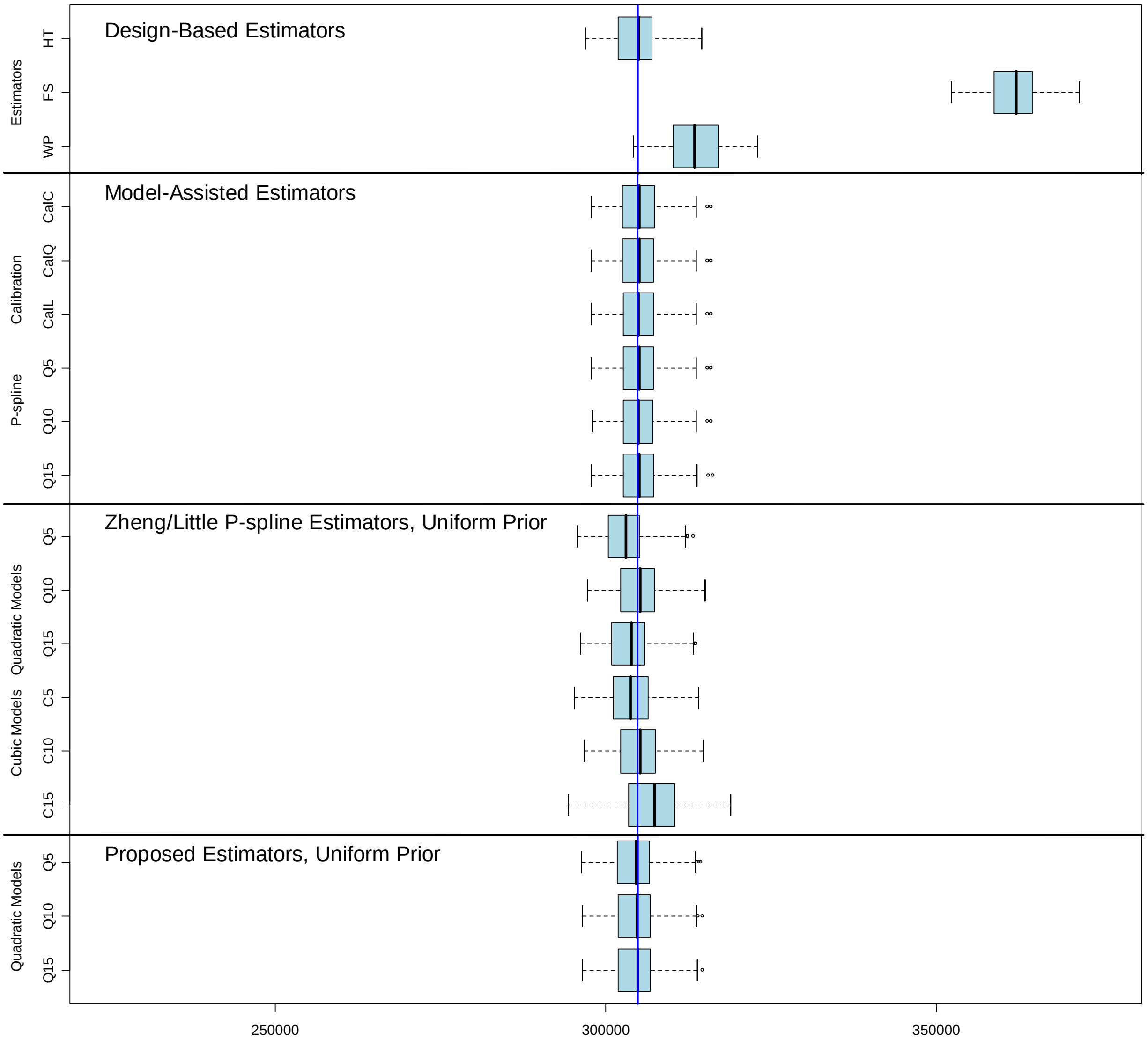


Figure 2.6. Boxplots of Estimated Totals, Population with Gamma Distributed Covariates, *ppswor* Samples

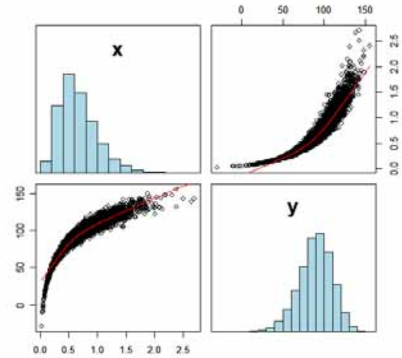


2.4.2. One-Sample Illustration of P-spline Models in Nonlinear Data

While the Sec. 2.4.1 simulations demonstrated that the Breidt *et. al* and proposed *p*-spline estimators produced comparable totals to the GREG estimators, the underlying population data followed linear patterns. Here the *p*-spline estimators are applied to nonlinear data, to demonstrate their potentially superior performance. The pseudopopulation data in Sec. 2.4.1 is altered to fit the following mixed model: $y_i = 100 + 0.95\left(10 + 40\log\left(X\right)\right) + 0.5\sqrt{X} + e_i$, where $X \sim Gamma\left(4,6\right)$ and

$e_i \sim N(0,10)$, $i = 1,\ldots,8300$. The pseudopopulation plot is shown in Figure 2.7.

Figure 2.7. Population Plots and Loess Lines for One-Sample Example of P-spline Models



One sample of size 250 was drawn from this population using probability proportional to size of X. The sample data, along with the prediction lines produced from several alternative models, are shown in Figure 2.8 below and on the following pages.

Figure 2.8. Sample Plots of P-spline Model Examples, *ppswor* Sample

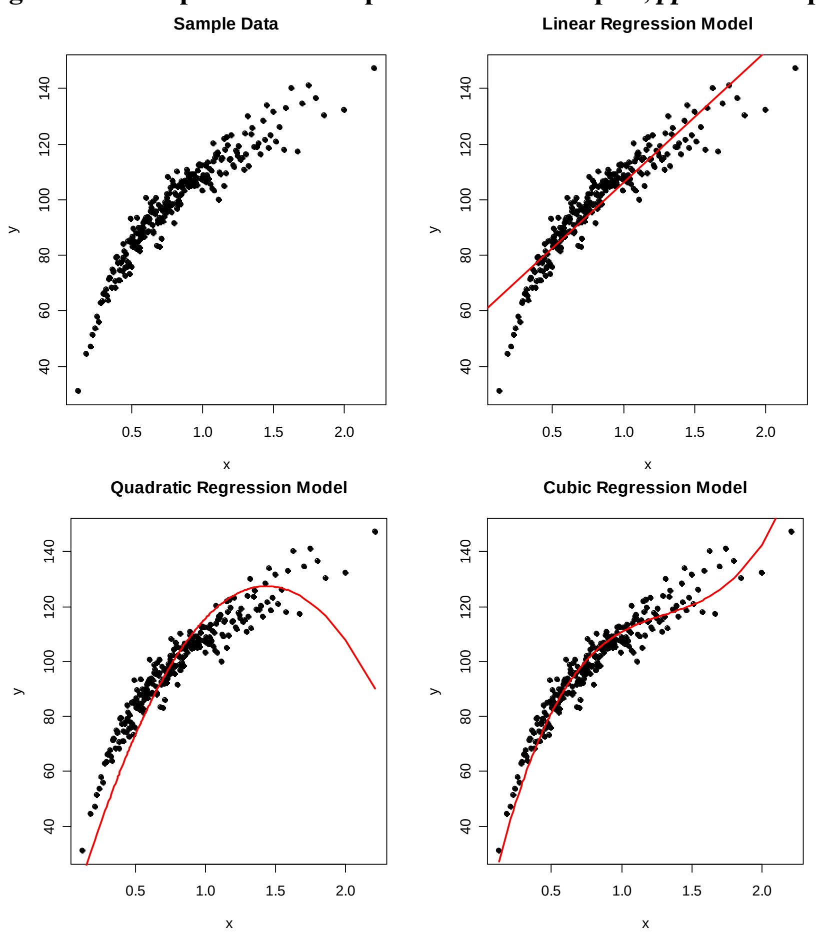


Figure 2.8. Sample Plots of P-spline Model Examples, *ppswor* Sample, cont'd

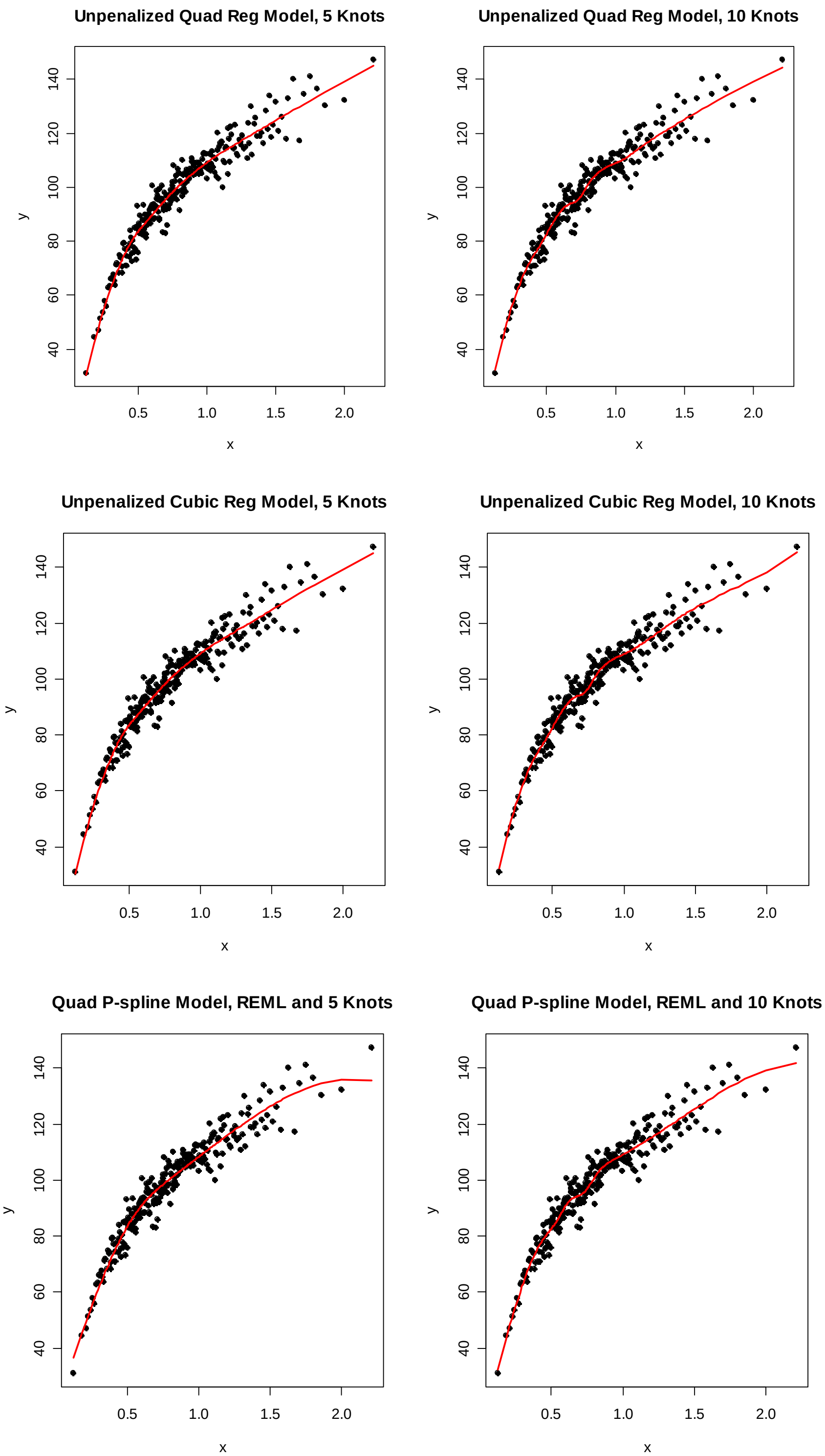
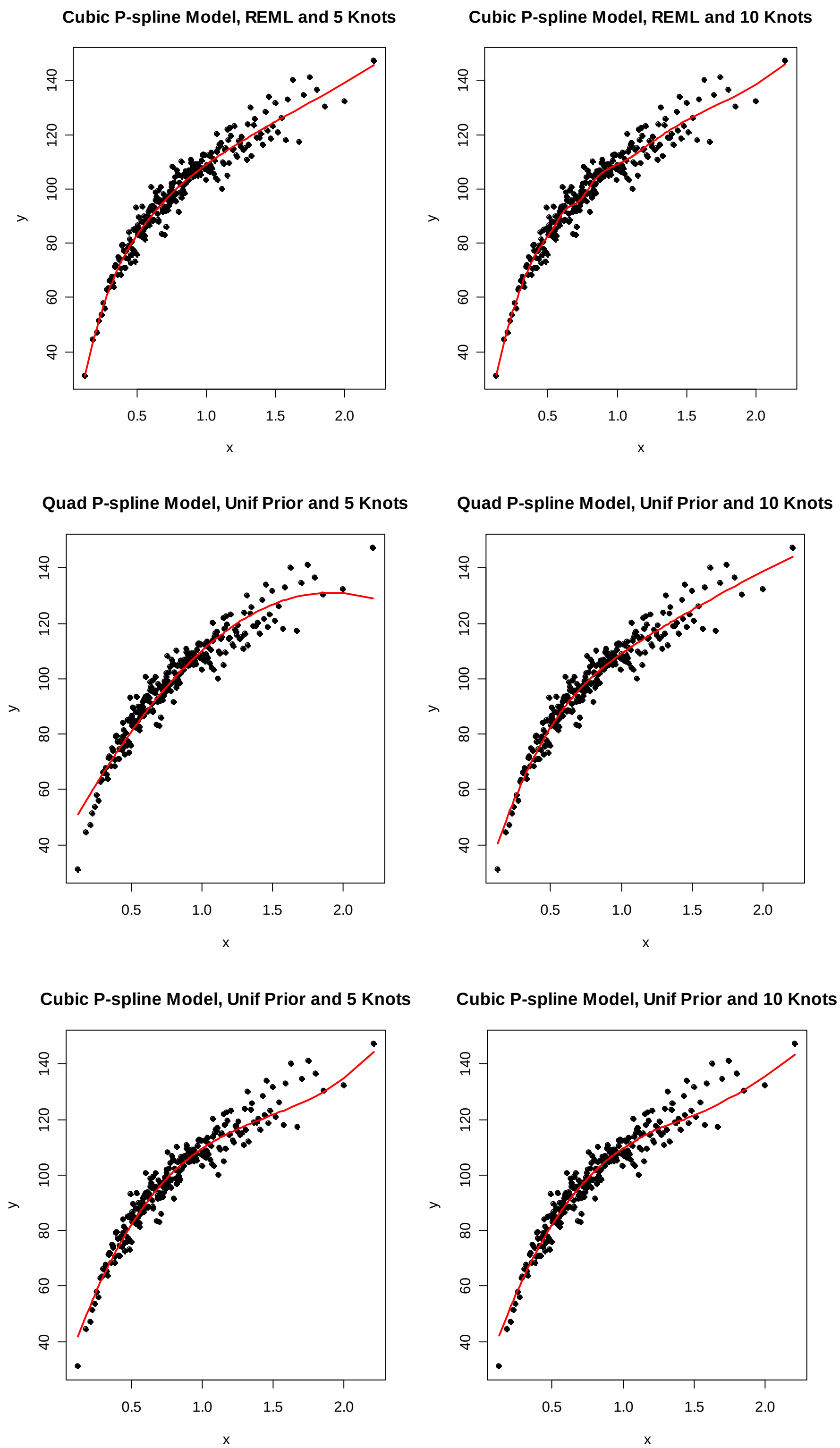


Figure 2.8. Sample Plots of P-spline Model Examples, *ppswor* Sample, cont'd.

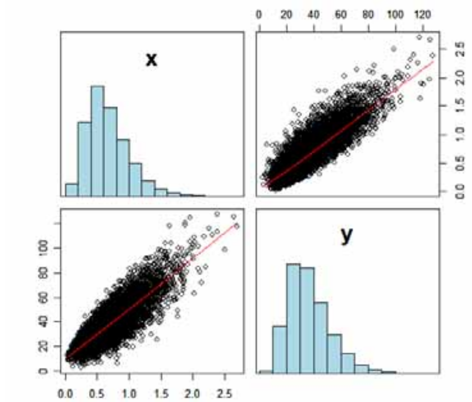


From Figure 2.8, we first see that, for this nonlinear data, the linear model between y_i, x_i is not appropriate. For smaller values of x_i, predicted values of y_i are too large, while larger y_i, x_i values have y_i, x_i predictions that are too small. The model fitting y_i, x_i, x_i^2 does not perform much better; smaller and larger values of x_i have predicted y_i's that are too small. The cubic model has a better fit, with the exception of larger x_i values due to a heteroscedastic variance. The regression spline models (the *p*-spline models with no penalty on the knots, titled "unpenalized quad/cubic reg models" in Figure 2.8) have a better fit to the data, although the models with 10 knots contribute more "wigglyness" to the prediction line than 5 knots. In contrast, the REML-based penalized model and the proposed model using the Uniform prior for the variance components (that determines the penalty) fit the data well but are much smoother. For this sample data, the models with 10 knots fit the data more appropriately than the models using 5 knots. Generally, this example illustrates that the *p*-spline estimators, including the proposed estimator using the Bayesian priors for unknown model parameters, have a better fit to the nonlinear data. This should extend to superior performance in estimation of totals, moreso than the benefits noted in the Sec. 2.4.1 evaluation.

2.4.3. Variance Estimators

This section contains an evaluation study related to estimating the variance of some *p*-spline estimators examined in Sec. 2.4.1. This simulation study uses the pseudopopulation from Sec. 2.4.1 with the gamma-distributed covariates, and has two factors. Here however, the population size was increased to 10,000 units and only the *ppswor* sampling was used to select four hundred samples of size $n = 250$ from the population. The population plots are given in Figure 2.9 on the following page.

Figure 2.9. Population Plots and Loess Lines for Simulation Comparing Alternative Variances



Only one version of the proposed estimator of the total is used: using the uniform prior on the variance components, with a quadratic (second-degree) polynomial model and $Q = 10$ knots from the estimators examined in Sec. 2.4.1. This estimator is used here since it had one of the lowest relative bias and the lowest RMSE for the *ppswor* samples drawn from this population (see Table 2.4). The only simulation factor here is the variance estimator; five alternatives (including three versions of the grouped jackknife) are compared:

- The with-replacement Taylor series-based variance estimator from the GREG AV:

  $$var_{\pi}\left(\hat{T}_{GREG} - \mathbf{T}_x^T \mathbf{B}_U\right) \doteq \left(1 - \frac{n}{N}\right) \sum_{i \in s} \frac{g_i^2 e_i^2}{\pi_i};$$

- The model-based variance: $var_M\left(\hat{T}_{psp}\right) = \left(\mathbf{d}^T + \mathbf{A}_s^{-1} \mathbf{c}_s^T\right) \mathbf{e}\mathbf{e}^T \left(\mathbf{d} + \mathbf{c}_s^{-1} \mathbf{A}_s\right)$.

- The delete-one-group jackknife variance estimator, using 10, 25, and 50 equal-sized groups: $var_J\left(\hat{T}\right) = \frac{G-1}{G} \sum_{g=1}^{G} \left(\hat{T}_{(g)} - \hat{T}\right)^2$.

I use five evaluation measures to compare the above alternatives produced for $B = 400$ simulation samples:

- *Relative bias*: the average distance between the variance estimator $var_b\left(\hat{T}_{psp}\right)$ and empirical variance of $\hat{T}_{psp}$, denoted $v\left(\hat{T}_{psp}\right) = B^{-1}\sum_{b=1}^{B}\left(\hat{T}_{psp} - B^{-1}\sum_{b=1}^{B}\hat{T}_{psp}\right)^2$:
$$RB\left(var\left(\hat{T}_{psp}\right)\right) = \left[B^{-1}\sum_{b=1}^{B}\left(var_b\left(\hat{T}_{psp}\right) - v\left(\hat{T}_{psp}\right)\right)\right] \Big/ v\left(\hat{T}_{psp}\right),$$

- *Empirical CV*: the standard error of the variance estimator, expressed as a percentage of the empirical variance:
$$CV\left(var\left(\hat{T}_{psp}\right)\right) = \sqrt{B^{-1}\sum_{b=1}^{B}\left(var_b\left(\hat{T}_{psp}\right) - B^{-1}\sum_{b=1}^{B}var_b\left(\hat{T}_{psp}\right)\right)^2} \Big/ v\left(\hat{T}_{psp}\right);$$

- *Empirical RelRMSE*: the mean square error of the variance estimator, expressed as a percentage of the empirical variance:
$$RelRMSE\left(var\left(\hat{T}_{psp}\right)\right) = \sqrt{B^{-1}\sum_{b=1}^{B}\left(var_b\left(\hat{T}_{psp}\right) - v\left(\hat{T}_{psp}\right)\right)^2} \Big/ v\left(\hat{T}_{psp}\right);$$

- *95% CI Coverage rate*: the percentage of the 400 simulated confidence intervals that contain the true population total: $\left|\hat{T}_{psp} - T\right| \Big/ \sqrt{var\left(\hat{T}_{psp}\right)} \le z_{\alpha/2} = 1.96$;

- *Average CI width*: the average width of the 95% confidence intervals: $2B^{-1}\sum_{b=1}^{B}1.96\sqrt{var_b\left(\hat{T}_{psp}\right)}$.

The evaluation measures are summarized for each variance estimator in Table 2.5.

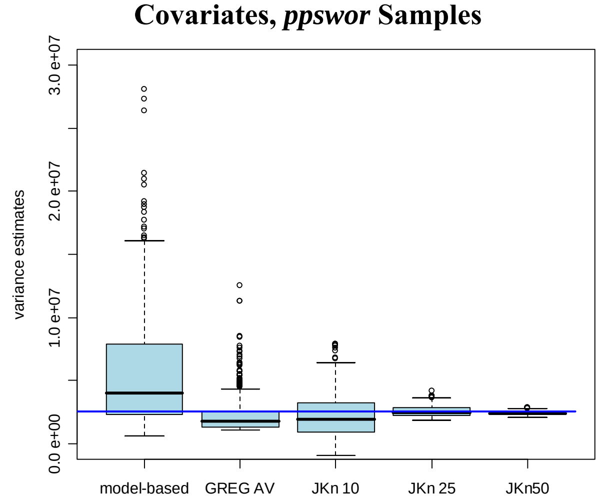
Table 2.5. Variance Estimation Simulation Results for Population with Gamma Distributed Covariates, *ppswor* Samples, and Estimator with Uniform Prior, 10 Knots ($T = 365,862$)

Variance Estimator	*Relative Bias (%)*	*Empirical CV (%)*	*RelRMSE*	*95% CI coverage*	*Ave. 95% CI Width*
Model-based	2.54	6.08	9.08	98.4	43,377.82
Taylor series/GREG AV	-1.39	3.06	4.36	90.7	3,977.06
Delete-one-group jackknife					
10 groups	-1.28	2.91	3.35	92.0	7,352.16
25 groups	-0.79	1.71	1.88	94.0	4,353.03
50 groups	-0.70	0.32	0.59	95.5	3,029.28

From Table 2.5, we see that all of the alternative variance estimators for $\hat{T}_{psp}$ have a low amount of bias; the largest at 2.5 percent is the model-based variance estimator. The other alternatives were slightly negatively biased, with the bias in the jackknife variance estimators decreasing as the number of replicate groups increased. Despite only using $B = 400$ samples, the empirical CV's of the variance estimators are all less than ten percent, with the model-based variance being the highest. From these, the model-based variance has the highest root mean square error relative to the empirical variance. All of the relative RMSE's exceeded one except the jackknife with fifty replicate groups, by varying degrees. The 95 percent CI coverage rates and average CI widths allow me to evaluate how well the alternative variance estimators perform inferentially with respect to $\hat{T}_{psp}$; the coverage is closest to the nominal rate and the average CI width was smallest for the jackknife variance with 50 replicate groups.

Figure 2.9 contains the boxplot distributions of the alternative variance estimates across the 400 simulation samples. The horizontal line in the variance plot is the empirical variance of the proposed *p*-spline total.

Figure 2.9. Boxplots of Estimated Variances, Population with Gamma Distributed Covariates, *ppswor* Samples



The boxplots reiterate the summary measures from Table 2.7: the model-based variance is positively biased and has the highest variance, the GREG AV and jackknife variance estimators are slightly negatively biased, and the variance of the jackknife variances decreases as the number of groups increases. The model-based variance also had some extreme outliers, which contributed to its slight positive bias, large variance, and large CI width in Table 2.5.

2.5. Discussion and Limitations

In this paper, I propose a modification of the Breidt *et al*. model-assisted penalized spline estimator. I also compare the proposed estimator against design-based, Bayesian model-based, and model-assisted estimators in a simulation evaluation. The proposed estimator performed well against the other methods in the Sec. 2.4 evaluation studies, as long as a sufficient number of knots were used in the *p*-spline. In particular, the proposed estimator had one of the lowest root mean square errors within the stratified samples. While the gains in precision and MSE for the proposed estimators were also comparable in the *ppswor* samples, generally the gains in all alternative estimators were not as large here compared to the stratified sample results. In both sample designs, totals estimated using the quadratic model with 10 and 15 knots were more efficient than using the same model with 5 knots. Unlike the quadratic model, in results not presented here, the cubic model for the proposed estimator generated an extreme range of totals, including negative ones. Further investigation as to why this occurred is needed.

In this paper, I also compare some alternative variance estimators for the proposed total. I demonstrate that the resampling-type jackknife variance estimator is accurate for estimating the variance of the proposed total. In my empirical application, the model-based variance estimator was overly conservative.

One of the main limitations of my proposed method is one central to both the model-based and model-assisted approaches: we must have good auxiliary information related to our survey variable of interest to produce an improved estimator of the finite population total. This information must be available for every unit in the population. This is not restrictive in some surveys, e.g., establishment surveys, where unit-level frame variables are available. If such information is available, then the *p*-spline model has potential to be more robust and efficient than alternative weight trimming and smoothing approaches.

Practically, it is noteworthy that both the Zheng and Little and proposed estimator are very computer-intensive. This problem increased for smaller sample sizes, e.g., the omitted results when fitting the models within strata. Also, the estimated totals can be sensitive to the number of model terms, i.e. the polynomial degree and the number of knots used.

Paper 3: Diagnostic Measures for Changes in Survey Weights

Abstract: Here I propose two different diagnostic measures to gauge the impact of various adjustments on weights: (1) a model-based extension of the design-effect measures for a summary-level diagnostic for different variables of interest, in single-stage and cluster sampling and under calibration weight adjustments; and (2) unit-level diagnostics to flag individual cases within a given sample that are more or less influential after their weights are adjusted. The proposed methods are illustrated using complex sample case studies.

3.1. Introduction and Research Plan

There are several approaches to adjusting weights, trimming weights, and bounding weights, as discussed and illustrated in Papers 1 and 2. Different approaches have also been developed to summarize the impact of differential weighting. The most popular measure is Kish's (1965, 1992) design-based design effect. Gabler *et al.* (1999) showed that, for cluster sampling, this estimator is a special form of a design effect produced using variances from random effects models, with and without intra-class correlations. Spencer (2000) proposed a simple model-based approach that depends on a single covariate to estimate the impact on variance of using variable weights.

However, currently these approaches do not provide a summary measure of the impact of weighting changes and adjustments on sample-based inference. While Kish-based design effects attempt to measure the impact of variable weights, they hold only under special circumstances, do not account for alternative variables of interest, and can incorrectly measure the impact of differential weighting in some circumstances.

Spencer's approach holds for a very simple estimator of the total, that uses base weights with no adjustments under with-replacement single-stage sampling.

More specifically, the Kish and Spencer measures may not accurately produce design effects for unequal weighting induced by calibration adjustments, which are often applied to reduce variances and correct for undercoverage and/or nonresponse (e.g., Särndal and Lundström 2005). When the calibration covariates are correlated with the coverage/response mechanism, calibration weights can improve the MSE of an estimator. However, in many applications, calibration produces weights that are more variable than the base weights or category-based nonresponse or postratification adjustments, since calibration involves unit-level adjustments. Thus, an ideal measure of the impact of calibration weights also incorporates not only the correlation between y and the weights, but also y and the calibration covariates $\mathbf{x}$.

I propose extending these existing design effect approaches as follows:

- Produce new variable-specific design-effect measures that summarize the impact of calibration weight adjustments before and after they are applied to survey weights. Specifically, I propose a new summary measure that incorporates the survey variable like Spencer's model that uses a generalized regression variance to incorporate multiple calibration covariates.

- Develop the estimators for the proposed design effect for single-stage and cluster sampling.

- Apply the estimators in case studies involving complex survey data and demonstrate empirically how the proposed estimator outperforms the existing methods in the presence of calibration weights.

In addition, there is limited research and methods to identify particular sample units' weights that have undue influence on the sample based estimator. On this topic, I propose to accomplish the following:

- Assess the influence of alternative case weights when estimating a total. I use statistical distance-based functions from related statistical literature to identify particular sample units with weights that are more or less influential on the sample-based total. The goal is to identify practical metrics to assist survey methodologists in determining whether or not a particular weight should be trimmed.

- Illustrate the adopted methods on a case study of complex survey data.

Both of these extensions lead toward the general goal of producing practical metrics that quantify and gauge the impact of weights on sample-based estimation. The proposed design effects account for unequal weight adjustments in the larger class of calibration estimators used in single-stage and cluster samples. The case-level influence measures identify particular sample units whose weights drive a particular survey's estimates.

3.2. Literature Review: Design Effect Measures for Differential Weighting Effects

This section describes existing design-effect measures for differential weights.

3.2.1. Kish's "Haphazard-Sampling" Design-Effect Measure for Single-Stage Samples

Kish (1965, 1990) proposed the "design effect due to weighting" as a measure to quantify the variability within a given set of weights. For $\mathbf{w} = \left(w_1, \ldots, w_n\right)^T$, in simple random sampling, this measure is

$$\begin{aligned} deff_K\left(\mathbf{w}\right) &= 1 + \left[CV\left(\mathbf{w}\right)\right]^2 \\ &= 1 + \frac{1}{n}\sum_{i \in s} \frac{\left(w_i - \bar{w}\right)^2}{\bar{w}^2} \\ &= \frac{n\sum_{i \in s} w_i^2}{\left[\sum_{i \in s} w_i\right]^2} \\ &= \frac{n^{-1}\sum_{i \in s} w_i^2}{\bar{w}^2} \end{aligned} , \qquad (3.1)$$

where $\bar{w} = n^{-1}\sum_{i \in s} w_i$. This is the most commonly used measure of an unequal weighting effect. Expression (3.1) is actually the ratio of the variance of the weighted survey mean under disproportionate stratified sampling to the variance under proportionate stratified sampling when all stratum unit variances are equal (Kish 1992).

3.2.2. Design Effect Measures for Cluster Sampling

Kish (1987) proposed a similar measure for cluster sampling. Some alternative notation is first needed. We consider that a finite population of M elements is partitioned into N clusters, each of size M_i, and denoted by $U = \{(i, j): i = 1, \ldots, N, j = 1, \ldots, M_i\}$. We select an equal-probability sample s' of n clusters using two-stage sampling from U and obtain a set of $s = \{(i, j): i = 1, \ldots, n, j = 1, \ldots, m_i\}$ respondents. Further, assume that there are G unique weights in s such that the m_{ig} elements within each cluster i have the same weight, denoted by $w_{ig} = w_g$ for $g = 1, \ldots, G$, m_g is the number of elements within weighting class g and $m = \sum_{g=1}^{G} m_g$ is the total number of elements in the sample. We estimate the population mean $\bar{Y} = T/M$ using the weighted sample mean $\bar{y}_w = \hat{T}_{HT} / \hat{M}_{HT} = \sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij} y_{ij} \Big/ \sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}$. Kish's (1987) decomposition model for $\bar{y}_w$ assumes that the G weighting classes are randomly ("haphazardly") formed with respect to y_{ij}, assuming that the y_{ij} have with a common variance and that s' is an epsem sample in which the variation among the m_i's within s is not significant. The resulting design effect is

$$deff\left(\bar{y}_w\right) = deff_w\left(\bar{y}_w\right) \times deff_c\left(\bar{y}_w\right)$$
$$\equiv \frac{m\sum_{g=1}^{G} w_g^2 m_g}{\left[\sum_{g=1}^{G} w_g m_g\right]^2} \times \left[1 + \left(\bar{m} - 1\right)\rho_c\right], \tag{3.2}$$

where $\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i$ is the average cluster size and ρ_c is the measure of intra-cluster homogeneity. The first component in (3.2) is the cluster-sample equivalent of (3.1), and can be written in a similar form to (3.1), using the squared CV of the weights. The second (3.2) component is the standard design effect due to the cluster sampling (e.g., Kish 1965). Expression (3.2) may not hold if there is variation in the m_{ig} across clusters (Park 2004) or moderate correlation between the survey characteristic and weights (Park and Lee 2004).

Gabler *et al*. (1999) used a model to justify measure (3.2) that assumes y_{ij} is a realization from a one-way random effects model (i.e., a one-way ANOVA-type model with only a random cluster-level intercept term plus an error) that assumes the following covariance structure:

$$Cov_{M1}\left(y_{ij}, y_{i'j'}\right) = \begin{cases} \sigma^2 & i = i', j = j' \\ \rho_e\sigma^2 & i = i', j \neq j \\ 0 & i \neq i' \end{cases}. \tag{3.3}$$

If the units are uncorrelated, then (3.3) reduces to $Cov_{M2}\left(y_{ij}, y_{i'j'}\right) = \sigma^2$ for $i = i', j = j'$ and 0 otherwise. More general models can be found in Rao and Kleffe (1988, p. 62). Under this model, Gabler *et al.* (1999) take the ratio of the model-based variance of the weighted survey mean under model M1 with covariance structure (3.3) to the variance under the uncorrelated errors version in model M2 and derive

$$deff\left(\bar{y}_w\right) = deff_w\left(\bar{y}_w\right) \times deff_{g_1}\left(\bar{y}_w\right)$$

$$\equiv \frac{m\sum_{g=1}^{G} w_g^2 m_g}{\left[\sum_{g=1}^{G} w_g m_g\right]^2} \times \left[1 + \left(\bar{m}_{g_1} - 1\right)\rho_e\right], \tag{3.4}$$

where $\bar{m}_{g_1} = \dfrac{\sum_{i=1}^{n}\left(\sum_{g=1}^{G} w_g m_{ig}\right)^2}{\sum_{g=1}^{G} w_g^2 m_g}$. They also established an upper bound for (3.4):

$$UB\left[deff\left(\bar{y}_w\right)\right] = deff_w\left(\bar{y}_w\right) \times deff_{g_2}\left(\bar{y}_w\right)$$

$$\equiv \frac{m\sum_{g=1}^{G} w_g^2 m_g}{\left[\sum_{g=1}^{G} w_g m_g\right]^2} \times \left[1 + \left(\bar{m}_{g_2} - 1\right)\rho_e\right], \tag{3.5}$$

where $\bar{m}_{g_2} = \dfrac{\sum_{i=1}^{n} m_i \sum_{g=1}^{G} w_g^2 m_{ig}}{\sum_{i=1}^{n}\sum_{g=1}^{G} w_g^2 m_{ig}}$ is a weighted average of cluster sizes. Note that here ρ_e is actually a model parameter (see Ch. 8 in Valliant *et al*. 2000). It can be estimated using an analysis of variance (ANOVA) estimator

$$\hat{\rho}_{ANOVA} = \frac{MSB - MSW}{MSB + \left(K - 1\right)MSW}, \tag{3.6}$$

where $MSB = \dfrac{1}{n-1}\sum_{i=1}^{n} m_i\left(\bar{y}_i - \bar{y}\right)^2$ is the "between-cluster" mean square error, $MSW = \dfrac{1}{n-I}\sum_{i=1}^{I}\sum_{j=1}^{m_i}\left(y_{ij} - \bar{y}_i\right)^2$ is the "within-cluster" mean square error, I is the number of clusters, and $K = \dfrac{n - \frac{1}{n}\sum_{i=1}^{n} m_i^2}{n-1}$. With this estimator, $\hat{\rho}_{ANOVA} \geq -\left(\bar{M} - 1\right)^{-1}$, where $\bar{M} = \dfrac{1}{N}\sum_{i=1}^{N} M_i$ is the average cluster size. Park (2004) further extends this approach to three-stage sampling, assuming that a systematic sampling is used in the first stage to select the clusters. Gabler *et al.* (2005) provide examples of special cases of (3.4), such as equal sampling/coverage/response rates across domains. More sophisticated ways, like REML (e.g., Searle 1977), have also been developed for estimating the MSE components.

3.2.3. Spencer's Model-based Measure for PPSWR Sampling

Spencer (2000) derives a design-effect measure to more fully account for inefficiency in variable weights that are correlated with the survey variable of interest. Suppose that p_i is the one-draw probability of selecting unit i, which is correlated with y_i and that a linear model holds for y_i: $y_i = A + Bp_i + e_i$, where e_i is not a model error; it is defined to be $e_i = y_i - A - Bp_i$. A particular case of this would be $p_i \propto x_i$, where x_i is a measure of size associated with unit i. If the entire finite population were available, then the ordinary least squares model fit is $y_i = \alpha + \beta p_i + e_i$. The estimates of α, β are $\alpha = \bar{Y} - \beta\bar{P}$ and $\beta = \frac{\sum_{i \in U}\left(y_i - \bar{Y}\right)\left(p_i - \bar{P}\right)}{\sum_{i \in U}\left(p_i - \bar{P}\right)^2}$, where $\bar{Y}, \bar{P}$ are the finite population means for y_i and p_i. The finite population variance is $\sigma_e^2 = \frac{\left(1 - \rho_{yp}^2\right)}{N}\sum_{i \in U}\left(y_i - \bar{Y}\right)^2 = \left(1 - \rho_{yp}^2\right)\sigma_y^2$, where ρ_{yp} is the finite population correlation between y_i and p_i. The weight under this *ppswr* sampling is $w_i = \left(np_i\right)^{-1}$. The estimated total is $\hat{T}_{pwr} = \sum_{i \in s} w_i y_i$, with design-variance $Var\left(\hat{T}_{pwr}\right) = \frac{1}{n}\sum_{i \in U} p_i \left(\frac{y_i}{p_i} - T\right)^2$ in single-stage sampling. Spencer substituted the model-based values for y_i into the variance and took its ratio to the variance of the estimated total using simple random sampling to produce the following design effect for unequal weighting (see Appendix 8 for derivation):

$$\begin{aligned} deff_S &= \frac{Var\left(\hat{T}_{pwr}\right)}{Var_{srs}\left(\hat{T}_{pwr}\right)} \\ &= \frac{\alpha^2}{\sigma_y^2}\left(\frac{n\bar{W}}{N}-1\right)+\frac{n\bar{W}}{N}\left(1-\rho_{yp}^2\right)+\frac{n\rho_{e^2w}\sigma_{e^2}\sigma_w}{N\sigma_y^2}+\frac{2\alpha n\rho_{ew}\sigma_e\sigma_w}{N\sigma_y^2} \end{aligned} . \tag{3.7}$$

Assuming that the correlations in the last two terms of (3.7) are negligible, Spencer approximates (3.7) with

$$deff_S \approx \left(1-\rho_{yp}^2\right)\frac{n\bar{W}}{N}+\left(\frac{\alpha}{\sigma_y}\right)^2\left(\frac{n\bar{W}}{N}-1\right), \tag{3.8}$$

where $\bar{W}=\frac{1}{N}\sum_{i\in U}w_i=\frac{1}{nN}\sum_{i\in U}\frac{1}{p_i}$ is the average weight in the population (see Appendix 8 for derivation). Spencer proposed estimating measure (3.7) with

$$\begin{aligned} \widehat{deff}_S &= \left(1-R_{yp}^2\right)\left[1+\left[CV(\mathbf{w})\right]^2\right]+\left(\frac{\hat{\alpha}}{\hat{\sigma}_y}\right)^2\left[CV(\mathbf{w})\right]^2 \\ &= \left(1-R_{yp}^2\right)deff_K(\mathbf{w})+\left(\frac{\hat{\alpha}}{\hat{\sigma}_y}\right)^2\left(deff_K(\mathbf{w})-1\right) \end{aligned} , \tag{3.9}$$

where $R_{yp}^2, \hat{\alpha}$ are the R-squared and estimated intercept from fitting the model $y_i=\alpha+\beta p_i+e_i$ with survey weighted least squares, and $\hat{\sigma}_y^2=\frac{\sum_{i\in s}w_i\left(y_i-\bar{y}_w\right)^2}{\sum_{i\in s}w_i}$ is the estimated population unit variance (see Appendix 8). When ρ_{yp} is zero and σ_y is large, measure (3.9) is approximately equivalent to Kish's measure (3.1). However, Spencer's method incorporates the survey variable y_i, unlike (3.1), and implicitly reflects the dependence of y_i on the selection probabilities p_i. We can explicitly see this by noting when N is large, $\alpha = \bar{Y} - \beta N^{-1} \approx \bar{Y}$, and (3.8) can be written as

$$\begin{aligned} deff_S &\approx \left(1-\rho_{yp}^2\right)\frac{n\bar{W}}{N} + \left(\frac{\bar{Y}}{\sigma_y}\right)^2 \left(\frac{n\bar{W}}{N} - 1\right) \\ &= \left(1-\rho_{yp}^2\right)\frac{n\bar{W}}{N} + \frac{1}{CV_{\bar{Y}}^2}\left(\frac{n\bar{W}}{N} - 1\right) \end{aligned}, \tag{3.10}$$

where $CV_{\bar{Y}}^2$ is the population-level unit coefficient of variation. We estimate (3.10) with

$$\widehat{deff}_S = \left(1-R_{yp}^2\right)\left[1+\left[CV\left(\mathbf{w}\right)\right]^2\right] + \frac{1}{\widehat{CV}_{\bar{y}_w}^2}\left[CV\left(\mathbf{w}\right)\right]^2, \tag{3.11}$$

where $\widehat{CV}_{\bar{y}_w}^2 = \hat{\sigma}_y^2 \big/ \hat{\bar{y}}_w^2$ (not the standard CV estimate produced in conventional survey software).

3.2.4. Summary

In general, the design effect measures currently provide the best comprehensive measures to summarize the impact of variable weights within a given survey. While the measures are generally proposed for a survey mean, they can often also be used for totals (see Exp. 8.7.7 in Särndal *et. al* 1992). However, each existing measure has associated limitations, and there is a lack of empirical applications in the literature proposing the methods, e.g., when particular approximations hold empirically and when they do not. These limitations are discussed in the remainder of this section.

The Kish summary measure is the most widely used summary measure to gauge the impact of variable weights. It requires only the values of the survey weights, thus is very simple to compute from a given sample. However, measure (3.1) can easily be misinterpreted as the measure of the increase in the variance of an estimator due to unequal weights, but clearly it does not involve a survey variable of interest. The weights w_i as well as the product $w_i y_i$ can both contribute to increased variability; Kish's measure does not account for the latter. Kish (1992) also indicates that differential weights can be much more efficient than equal weights in particular cases, such as establishment surveys, where the variances differ across strata, household surveys with oversampled subgroups to meet target sample sizes, or samples that have differential nonresponse across subgroups, such that the nonresponse adjustments produce variable weights. Measure (3.1) could produce misleading results if used to measure variability in the weights in these circumstances. These examples are not cases of the "haphazardly" formed weighting class cells that Kish's design effect (3.1) is designed to measure the impact of.

The Kish design effect, and Gabler *et. al* model-based equivalent, for cluster sampling is more restrictive than the equivalent measure for single-stage sampling. It only holds under a particular form of weighting adjustments, where the survey data are grouped into the G groups and each unit within a group is assigned a common weight. This design effect only holds under special cases of cell-based weighting adjustments, like poststratification, and not other weighting adjustments where the individual units are allowed to have differing weights. It is not clear how the existing design effect would be modified to account for variability in weights under different adjustments; application of this design effect under other types of weighting adjustments would be ad hoc at best.

Spencer's design effect addresses a limitation in the Kish measure by incorporating a correlation between the survey variable of interest and the weights.

However it only holds for estimating a total under the particular combination of single-stage, with-replacement probability-proportional to size sampling, and the PWR estimator. Spencer also does not provide any theoretical recommendations or empirical evidence that the correlation terms in (3.7) are negligible, in which case the approximation in (3.8) is appropriate.

To address some weaknesses in the existing design-effect measures, Section 3.3.1 describes my proposed method to extend Spencer's measure to the calibration estimator in single-stage sampling. This accounts for measuring the variation in single-stage sample weights that fall under the general category of calibration. This incorporates more forms of commonly used weighting adjustment methods. Section 3.3.2 extends this measure to cluster sampling. In Sec. 3.3.3, I use a heuristic approach by proposing to use some nonparametric measures proposed in the statistical literature, and apply them in an empirical case study in Sec. 3.4.3.

3.3. Proposed Methods

3.3.1. Design-Effect Measure for Single-Stage Sampling

Here I propose to extend Spencer's (2000) approach in single-stage sampling to produce a new weighting design effect measure for a calibration estimator. Spencer's approach produces a variable-level design-effect measure that incorporates auxiliary information only in p_i. However, here the proposed design effect estimates the joint effect of the sample design and calibration estimator weights, which covers a range of more commonly used estimators, including poststratification, raking, and the GREG estimator. While Spencer's model assumed $y_i = A + Bp_i + e_i$, here the assumed model is

$y_i = \alpha + \mathbf{x}_i^T \boldsymbol{\beta} + e_i = \dot{\mathbf{x}}_i^T \mathbf{B}_U + e_i$, where $\dot{\mathbf{x}}_i = \begin{bmatrix} 1 & \mathbf{x}_i \end{bmatrix}$ and $\mathbf{B}_U = \begin{bmatrix} \alpha & \boldsymbol{\beta} \end{bmatrix}$. Again, the term $y_i = \alpha + \mathbf{x}_i^T \boldsymbol{\beta} + e_i = \dot{\mathbf{x}}_i^T \mathbf{B}_U + e_i$ is not a model error, just a term that is equivalent to $e_i = y_i - \alpha + \mathbf{x}_i^T \boldsymbol{\beta} = y_i - \dot{\mathbf{x}}_i^T \mathbf{B}_U$. I reformulate the model as $y_i - \mathbf{x}_i^T \boldsymbol{\beta} = \alpha + e_i$ to simplify the results and implicitly account for a number of correlations in the model components, namely between y and $\mathbf{x}$. Note that using the model $y_i = A + \mathbf{x}_i^T \mathbf{B} + Cp_i + e_i$, which might seem to be the natural extension of Spencer's formulation, will produce non-estimable parameters (due to singular matrices).

A linearization of the GREG estimator (Exp. 6.6.9 in Särndal *et al*. 1992) is

$$\begin{aligned}
\hat{T}_{GREG} &\doteq \hat{T}_{HT} + \left(\mathbf{T}_x - \hat{\mathbf{T}}_{HTx}\right)^T \mathbf{B}_U \\
&= \sum_{i \in s} \frac{y_i}{\pi_i} + \mathbf{T}_x^T \mathbf{B}_U - \sum_{i \in s} \frac{\dot{\mathbf{x}}_i^T \mathbf{B}_U}{\pi_i} \\
&= \mathbf{T}_x^T \mathbf{B}_U + \sum_{i \in s} \sum_{j \in s_i} \left(\frac{y_i}{\pi_i} - \frac{\dot{\mathbf{x}}_i^T \mathbf{B}_U}{\pi_i} \right) \\
&= \mathbf{T}_x^T \mathbf{B}_U + \sum_{i \in s} \frac{e_i}{\pi_i} \\
&= \mathbf{T}_x^T \mathbf{B}_U + \tilde{e}_U
\end{aligned} \tag{3.12}$$

where $\mathbf{T}_x$ is the known population total of $\mathbf{x}$, $\hat{\mathbf{T}}_{HTx}$ is the vector of HT estimators, $\mathbf{B}_U$ is the population coefficients, $e_i = y_i - \dot{\mathbf{x}}_i^T \mathbf{B}_U$ is the calibration residual, $\dot{\mathbf{x}}_i^T$ is a row vector including the intercept, $\tilde{e}_U = \sum_{i \in s} \frac{e_i}{\pi_i}$ is the sum of the weighted unit-level residuals, and π_i is the overall probability of selection. If we assume that with-replacement sampling was used, then $\pi_i = np_i$ and (3.12) becomes

$$\hat{T}_{GREG} \doteq \mathbf{T}_x^T \mathbf{B}_U + \frac{1}{n}\sum_{i\in s}\frac{e_i}{p_i} . \tag{3.13}$$

The second component in (3.13) has design-expectation

$$\begin{aligned} E_\pi\left(\frac{1}{n}\sum_{i\in s}\frac{e_i}{p_i}\right) &= \frac{1}{n}\sum_{i\in U}\frac{E_\pi(\delta_i)e_i}{p_i} \\ &= \frac{1}{n}\sum_{i\in U}\frac{(np_i)e_i}{p_i} \quad , \\ &= \sum_{i\in U} e_i \\ &= E_U \end{aligned} \tag{3.14}$$

where $E_U = \sum_{i\in U} e_i$. From (3.13), $\hat{T}_{GREG} - \mathbf{T}_x^T \mathbf{B}_U \doteq \frac{1}{n}\sum_{i\in s}\frac{e_i}{p_i}$, with design-variance

$$\begin{aligned} Var_\pi\left(\hat{T}_{GREG} - \mathbf{T}_x^T \mathbf{B}_U\right) &= Var_\pi\left(\frac{1}{n}\sum_{i\in s}\frac{e_i}{p_i}\right) \\ &= \frac{1}{n}\sum_{i\in U} p_i\left(\frac{e_i}{p_i} - E_U\right)^2 . \end{aligned} \tag{3.15}$$

We can follow Spencer's approach and use a model-based plug-in to variance (3.15) to formulate a design-effect measure. However, here we substitute in the model-based equivalent to e_i, not y_i as Spencer does. This measure captures the combined effect of unequal weighting from the sample design and calibration weights, since the variance (3.15) can be used for all calibration estimators (Särndal *et. al* 1992). Substituting the GREG-based residuals into the variance and taking its ratio to the variance of the pwr-estimator in simple random sampling with replacement, $Var_{srs}\left(\hat{T}_{pwr}\right) = \frac{N^2\sigma_y^2}{n}$, where $\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2$, produces the approximate design effect due to unequal calibration weighting.

We can simplify things greatly by reformulating our model as $u_i = \alpha + e_i$, where $u_i = y_i - \mathbf{x}_i^T \boldsymbol{\beta}$. The resulting design effect (see Appendix 9) is

$$
\begin{aligned}
deff_{S*} &= \frac{Var\left(\hat{T}_{GREG}\right)}{Var_{srs}\left(\hat{T}_{pwr}\right)} \\
&= \frac{n\bar{W}}{N}\left(\frac{\sigma_u^2}{\sigma_y^2}\right) + \frac{\left(\bar{U}-\alpha\right)^2}{\sigma_y^2}\left(\frac{n\bar{W}}{N}-1\right) + \frac{n\sigma_w}{N\sigma_y^2}\left[\rho_{u^2w}\sigma_{u^2} - 2\alpha\rho_{uw}\sigma_u\right]
\end{aligned}, \quad (3.16)
$$

where $\alpha = \bar{U} - \gamma N^{-1}$, $\bar{U} = \frac{1}{N}\sum_{i=1}^N u_i = \frac{1}{N}\sum_{i=1}^N \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta}\right)$, $\sigma_u^2 = \frac{1}{N}\sum_{i=1}^N \left(u_i - \bar{U}\right)^2$, and $\sigma_y^2 = \frac{1}{N}\sum_{i=1}^N \left(y_i - \bar{Y}\right)^2$. Under our model $u_i = A + e_i$, $\alpha = \bar{U}$ and (3.16) becomes

$$deff_{S*} = \frac{n\bar{W}}{N}\left(\frac{\sigma_u^2}{\sigma_y^2}\right) + \frac{n\sigma_w}{N\sigma_y^2}\left[\rho_{u^2w}\sigma_{u^2} - 2\alpha\rho_{uw}\sigma_u\right]. \quad (3.17)$$

To estimate (3.17), we use the following (Appendix 9):

$$\widehat{deff_{S*}} \approx \left(1 + \left[CV\left(\mathbf{w}\right)\right]^2\right)\frac{\hat{\sigma}_u^2}{\hat{\sigma}_y^2} + \frac{n\hat{\sigma}_w}{N\hat{\sigma}_y^2}\left[\hat{\rho}_{u^2w}\hat{\sigma}_{u^2} - 2\hat{\alpha}\hat{\rho}_{uw}\hat{\sigma}_u\right], \quad (3.18)$$

where the model parameter estimate $\hat{\alpha}$ is obtained using survey-weighted least squares,

$$\hat{\sigma}_y^2 = \frac{\sum_{i\in s} w_i \left(y_i - \bar{y}_w\right)^2}{\sum_{i\in s} w_i}, \; \hat{\bar{y}}_w = \frac{\sum_{i\in s} w_i y_i}{\sum_{i\in s} w_i}, \; \hat{\sigma}_u^2 = \frac{\sum_{i\in s} w_i \left(\hat{u}_i - \bar{u}_w\right)^2}{\sum_{i\in s} w_i}, \; \hat{\bar{u}}_w = \frac{\sum_{i\in s} w_i \hat{u}_i}{\sum_{i\in s} w_i}$$

and $\hat{u}_i = y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$.

Following Spencer, if the correlations in (3.17) are negligible, then expression (3.17) is the following:

$$deff_{S*} \approx \frac{n\bar{W}}{N}\left(\frac{\sigma_u^2}{\sigma_y^2}\right), \quad (3.19)$$

which we estimate with

$$\widehat{deff_{S*}} \approx \frac{\hat{\sigma}_u^2}{\hat{\sigma}_y^2}\left(1+\left[CV\left(\mathbf{w}\right)\right]^2\right). \tag{3.20}$$

Note that without calibration, we have $\hat{u}_i = y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}} \approx y_i$, and $\sigma_u^2 \approx \sigma_y^2$, in which case the design effect approximation in (3.19) becomes $deff_{S*} \approx \frac{n\bar{W}}{N}$, which we estimate with Kish's measure $\widehat{deff_K} \approx 1+\left[CV\left(\mathbf{w}\right)\right]^2$. However, when the relationship between the calibration covariates $\mathbf{x}$ and y is stronger, we should expect the variance σ_u^2 to be smaller than σ_y^2. In this case, measure (3.20) is smaller than Kish's estimate using only the weights. Variable weights produced from calibration adjustments are thus not as "penalized" (shown by overly high design effects) as they would be using the Kish and Spencer measures. However, if we have "ineffective" calibration, or a weak relationship between $\mathbf{x}$ and y, then σ_u^2 can be greater than σ_y^2, producing a design effect greater than one. The Spencer measure only accounts for an indirect relationship between $\mathbf{x}$ and y if there was only one $\mathbf{x}$ and it was used to produce p_i. This is illustrated in the Sec. 3.4.1 case study that mimics establishment-type data. On a practical note, calibration weights fit within the models should all be positive.

3.3.2. Design-Effect Measure for Cluster Sampling

Here, the method used in Sec. 3.3.1 is extended to cluster sampling. A two-stage sample of clusters and units within clusters is assumed. For cluster sampling, we start with N clusters in the population, with M_i elements within cluster i. For $\dot{\mathbf{x}}_{ij} = \begin{bmatrix}1 & \mathbf{x}_{ij}\end{bmatrix}$,

$\mathbf{B}_U = \left(\alpha_U \quad \boldsymbol{\beta}\right)^T$, and $e_{ij} = y_{ij} - \alpha_i - \mathbf{x}_{ij}^T\boldsymbol{\beta} = y_{ij} - \dot{\mathbf{x}}_{ij}^T\mathbf{B}_U$. The GREG estimator here is given by

$$\begin{aligned}
\hat{T}_{GREG} &= \hat{T}_{HT} + \left(\mathbf{T}_x - \hat{\mathbf{T}}_{HTx}\right)^T \hat{\mathbf{B}} \\
&\doteq \hat{T}_{HT} + \left(\mathbf{T}_x - \hat{\mathbf{T}}_{HTx}\right)^T \mathbf{B}_U \\
&= \sum_{i\in s}\sum_{j\in s_i}\left(\frac{y_{ij}}{\pi_{ij}} - \frac{\dot{\mathbf{x}}_{ij}^T\mathbf{B}_U}{\pi_{ij}}\right) + \mathbf{T}_x^T\mathbf{B}_U\,, \\
&= \sum_{i\in s}\sum_{j\in s_i}\frac{e_{ij}}{\pi_{ij}} + \mathbf{T}_x^T\mathbf{B}_U
\end{aligned} \tag{3.21}$$

where $\mathbf{T}_x$ is the known population total of $\mathbf{x}$, $\hat{\mathbf{T}}_{HTx}$ is the vector of HT estimators, $\mathbf{B}_U$ is the population coefficients, $e_{ij} = y_{ij} - \mathbf{x}_{ij}^T\boldsymbol{\beta}$ is the "within-cluster" residual, $\mathbf{x}_{ij}^T$ is a row vector, and the overall probability of selection is the product of the first- and second-stage selection probabilities: $\pi_{ij} = \pi_i\pi_{j|i}$. From (3.21), $\hat{T}_{GREG} - \mathbf{T}_x^T\mathbf{B}_U \doteq \sum_{i\in s}\sum_{j\in s_i}\frac{e_{ij}}{\pi_{ij}}$. Assuming that we have probability-with-replacement (pwr) sampling of clusters, the probability of selection for clusters is approximately $\pi_i = 1-\left(1-p_i\right)^n \doteq np_i$ (if p_i is not too large), where p_i is the one-draw selection probability. Suppose that simple random sampling is used to select elements within each cluster, such that the second-stage selection probability is $\pi_{j|i} = \frac{m_i}{M_i}$ for element j in cluster i. Then the overall selection probability is approximately $\pi_{ij} = \pi_i\pi_{j|i} \doteq \frac{np_im_i}{M_i}$ and expression (3.21) becomes

$$\hat{T}_{GREG} - \mathbf{T}_x^T \mathbf{B}_U \doteq \sum_{i \in s} \sum_{j \in s_i} \frac{M_i e_{ij}}{n p_i m_i}, \\ = \sum_{i \in s} w_i \hat{T}_{ei} \tag{3.22}$$

where $w_i = (np_i)^{-1}$ and $\hat{T}_{ei} = \frac{M_i}{m_i} \sum_{j \in s_i} e_{ij}$. The first component in (3.22) has design-expectation

$$\begin{aligned} E_\pi\left(\frac{1}{n}\sum_{i \in s}\sum_{j \in s_i}\frac{M_i e_{ij}}{p_i m_i}\right) &= E_\pi\left(\frac{1}{n}\sum_{i \in U}\sum_{j \in U_i}\frac{\delta_{ij} M_i e_{ij}}{p_i m_i}\right) \\ &= \frac{1}{n}\sum_{i \in U}\sum_{j \in U_i}\frac{E_\pi\left(\delta_{ij}\right) M_i e_{ij}}{p_i m_i} \\ &= \frac{1}{n}\sum_{i \in U}\sum_{j \in U_i}\frac{\left(np_i\, m_i / M_i\right) M_i e_{ij}}{p_i m_i}, \\ &= \sum_{i \in U}\sum_{j \in U_i} e_{ij} \\ &= \sum_{i \in U} E_{i+} \\ &= E_U \end{aligned} \tag{3.23}$$

where $E_U = \sum_{i \in U}\sum_{j \in U_i} e_{ij}$ and $E_{i+} = \sum_{j \in U_i} e_{ij}$. Expression (3.22) has the approximate design-variance

$$\begin{aligned} Var\left(\hat{T}_{GREG}\right) &= \frac{1}{n}\sum_{i=1}^{N} p_i \left(\frac{e_{Ui+}}{p_i} - E_{U+}\right)^2 + \sum_{i=1}^{N}\frac{M_i^2}{np_i m_i}\left(1 - \frac{m_i}{M_i}\right)\frac{1}{M_i - 1}\sum_{j=1}^{M_i}\left(e_{ij} - \bar{e}_{U_i}\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^{N}\left(\frac{e_{Ui+}^2}{p_i} - E_{U+}^2\right) + \sum_{i=1}^{N}\frac{M_i^2}{np_i m_i}\left(1 - \frac{m_i}{M_i}\right) S_{U_i}^2 \end{aligned},$$

(3.24)

where $\bar{e}_{U_i} = \frac{1}{M_i}\sum_{j=1}^{M_i} e_{ij}$, $e_{U_i} = \sum_{j=1}^{M_i} e_{ij}$, and $E_{U+} = \sum_{i=1}^{N} e_{ij}$. Suppose that the second-stage sampling fraction is negligible, i.e., $\frac{m_i}{M_i} \approx 0$. We can follow the approach used in Sec. 3.3.1 and use variance (3.24) to formulate a design-effect measure to capture the effect of unequal weighting from the calibration weight adjustments used in cluster sampling. To match the theoretical variance formulation in (3.24), we fit the model $y_{ij} = A + \mathbf{x}_{ij}^T \mathbf{B} + e_{ij}$, where $A, \mathbf{B}$ are the finite population model parameters. In the ordinary least squares fit, $y_{ij} = \alpha_U + \mathbf{x}_{ij}^T \boldsymbol{\beta} + e_{ij}$, where $\alpha_U = \bar{Y}_U - \bar{\mathbf{x}}_U^T \boldsymbol{\beta}$, where $\bar{Y}_U = \sum_{i \in U} \sum_{j \in U_i} Y_{ij} \Big/ \sum_{i \in U} M_i$.

Similar to 3.3.1, we reformulate the model as $u_{ij} = y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta}$, such that $e_{ij} = u_{ij} - \alpha_U$ and incorporate $\alpha_i = \alpha M_i$ as the cluster-level (random) intercept. The model with only the intercept α_i and error term is equivalent to Gabler *et. al*'s (1999) random effects model described in Sec. 1.3. Substituting the GREG-based residuals into the two (3.24) variance components and taking its ratio to the pwr-variance under simple random sampling, assuming that the M_i are large enough such that $\frac{M_i}{M_i - 1} \approx 1$ and the within-cluster sampling fractions are negligible, we obtain the following approximate design effect (see Appendix 10 for details):

$$deff_C \approx \frac{n\bar{W}}{N}\left(\frac{\sigma_{u_+}^2 + \sigma_\alpha^2 - 2\overline{U\alpha}}{\sigma_y^2}\right) + \left(\frac{n\bar{W}}{N} - 1\right)\left(\frac{\bar{U}^2 + \bar{\alpha}^2}{\sigma_y^2}\right) + \frac{2\bar{U}\bar{\alpha}}{\sigma_y^2}$$
$$+ \frac{n\sigma_w\left(\rho_{u_+^2 w}\sigma_{u_+^2} + \rho_{\alpha^2 w}\sigma_{\alpha^2} - 2\rho_{u_+\alpha, w}\sigma_{u_+\alpha}\right)}{N\sigma_y^2} + \frac{n}{N^2\sigma_y^2}\sum_{i=1}^{N}\frac{M_i^2 w_i S_{U_{ui}}^2}{m_i}, \quad (3.25)$$

which we estimate with

$$\widehat{deff_C} \approx \left[1+\left(CV\left(\mathbf{w}\right)\right)^2\right]\left(\frac{\hat{\sigma}_{u_+}^2+\hat{\sigma}_{\alpha}^2-2\widehat{\overline{u_w\alpha_w}}}{\hat{\sigma}_y^2}\right)+\left(CV\left(\mathbf{w}\right)\right)^2\left(\frac{\hat{\bar{u}}_w^2+\hat{\bar{\alpha}}_w^2}{\hat{\sigma}_y^2}\right)+\frac{2\hat{\bar{u}}_w\hat{\bar{\alpha}}_w}{N\hat{\sigma}_y^2}$$
$$+\frac{n\hat{\sigma}_w\left(\hat{\rho}_{u_+^2w}\hat{\sigma}_{u_+^2}+\hat{\rho}_{\alpha^2w}\hat{\sigma}_{\alpha^2}-2\hat{\rho}_{u_+\alpha,w}\hat{\sigma}_{u_+\alpha}\right)}{N\hat{\sigma}_y^2}+\frac{n}{N^2\hat{\sigma}_y^2}\sum_{i=1}^{n}\frac{M_i^2w_i\hat{S}_{U_{ui}}^2}{m_i}, \quad (3.26)$$

where the model parameter estimates $\hat{\alpha}_i$ are obtained using survey-weighted least squares, $\hat{\sigma}_\alpha^2=\frac{\sum_{i\in s}w_i\left(\hat{\alpha}_i-\hat{\bar{\alpha}}_w\right)^2}{\sum_{i\in s}w_i}$, $\hat{\bar{\alpha}}_w=\frac{\sum_{i\in s}w_i\hat{\alpha}_i}{\sum_{i\in s}w_i}$, $\hat{\sigma}_{u_+}^2=\frac{\sum_{i\in s}w_i\left(\hat{u}_i-\hat{\bar{u}}_w\right)^2}{\sum_{i\in s}w_i}$,

$\hat{\sigma}_y^2=\frac{\sum_{i\in s}\sum_{j\in s_i}w_{ij}\left(y_{ij}-\hat{\bar{y}}_w\right)^2}{\sum_{i\in s}\sum_{j\in s_i}w_{ij}}$, $\hat{\bar{y}}_w=\frac{\sum_{i\in s}\sum_{j\in s_i}w_{ij}y_{ij}}{\sum_{i\in s}\sum_{j\in s_i}w_{ij}}$, $\hat{S}_{U_{ui}}^2=\frac{\sum_{j\in s_i}\left(\hat{u}_{ij}-\hat{\bar{u}}_i\right)^2}{m_i-1}$,

$\widehat{\overline{u_w\alpha_w}}=\frac{\sum_{i=1}^{n}w_i\hat{u}_{i+}\hat{\alpha}_i}{\sum_{i=1}^{n}w_i}$, $\hat{\bar{u}}_w=\frac{\sum_{i\in s}w_i\hat{u}_{i+}}{\sum_{i\in s}w_i}$, $\hat{\bar{u}}_i=\frac{1}{m_i}\sum_{j\in s_i}\hat{u}_{ij}$, and

$\hat{u}_{i+}=y_{i+}-\mathbf{x}_{i+}^T\hat{\boldsymbol{\beta}}=\sum_{j\in s_i}\hat{u}_{ij}=\sum_{j\in s_i}\left(y_{ij}-\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}\right)$.

Assuming that the correlations in (3.25) are negligible or the clusters were selected with equal probabilities, the design effect is approximately

$$deff_C \approx \frac{n\bar{W}}{N}\left(\frac{\sigma_{u_+}^2+\sigma_\alpha^2-2\overline{U\alpha}}{\sigma_y^2}\right)+\left(\frac{n\bar{W}}{N}-1\right)\left(\frac{\bar{U}^2+\bar{\alpha}^2}{\sigma_y^2}\right)+\frac{2\bar{U}\bar{\alpha}}{\sigma_y^2}$$
$$+\frac{n}{N^2\sigma_y^2}\sum_{i=1}^{N}\frac{M_i^2w_iS_{U_{ui}}^2}{m_i}, \quad (3.27)$$

which we estimate with

$$\widehat{deff_C} \approx \left[1+\left(CV\left(\mathbf{w}\right)\right)^2\right]\left(\frac{\hat{\sigma}_{u_+}^2+\hat{\sigma}_{\alpha}^2-2\widehat{u_w\alpha_w}}{\hat{\sigma}_y^2}\right)+\left(CV\left(\mathbf{w}\right)\right)^2\left(\frac{\hat{\bar{u}}_w^2+\hat{\bar{\alpha}}_w^2}{\hat{\sigma}_y^2}\right)+\frac{2\hat{\bar{u}}_w\hat{\bar{\alpha}}_w}{N\hat{\sigma}_y^2} \\ +\frac{n}{N^2\hat{\sigma}_y^2}\sum_{i=1}^{n}\frac{M_i^2 w_i \hat{S}_{U_{ui}}^2}{m_i} . \tag{3.28}$$

Assuming that M_i are close enough such that $M_i \approx \bar{M}$ and $\alpha_i = M_i\alpha \doteq \bar{M}\alpha$ and $\sigma_{\alpha}^2 = \sigma_{\alpha^2} = 0$, then (3.25) becomes

$$deff_C \approx \frac{n\bar{W}}{N}\left(\frac{\sigma_{u_+}^2}{\sigma_y^2}\right)+\left(\frac{n\bar{W}}{N}-1\right)\frac{\left(\bar{U}-\bar{M}\alpha\right)^2}{\sigma_y^2}+ \\ +\frac{n\sigma_w\left(\rho_{u_+^2 w}\sigma_{u_+^2}-2\bar{M}\alpha\sigma_u\rho_{u_+\alpha,w}\right)}{N\sigma_y^2}+\frac{n\bar{M}^2}{N^2\sigma_y^2}\sum_{i=1}^{N}\frac{w_i S_{U_{ui}}^2}{m_i} . \tag{3.29}$$

Measure (3.29) can be estimated using

$$\widehat{deff_C} \approx \left[1+\left(CV\left(\mathbf{w}\right)\right)^2\right]\frac{\hat{\sigma}_{u_+}^2}{\sigma_y^2}+\left(CV\left(\mathbf{w}\right)\right)^2\frac{\left(\hat{\bar{u}}_w-\bar{M}\hat{\alpha}_w\right)^2}{\hat{\sigma}_y^2} \\ +\frac{n\hat{\sigma}_w\left(\hat{\rho}_{u_+^2 w}\hat{\sigma}_{u_+^2}-2\bar{M}\hat{\alpha}_w\hat{\rho}_{u_+\alpha,w}\hat{\sigma}_u\right)}{N\hat{\sigma}_y^2}+\frac{n\bar{M}^2}{N^2\hat{\sigma}_y^2}\sum_{i=1}^{N}\frac{w_i\hat{S}_{U_{ui}}^2}{m_i} . \tag{3.30}$$

The Kish measure is also a special case of (3.30), when there are no cluster-level effects. That is, if $\alpha_i = \alpha$ for all i and we have no auxiliary information in $\dot{\mathbf{x}}$, and no cluster sampling, i.e., $\bar{U} \approx \bar{Y}, \sigma_{u_+}^2 \approx \sigma_y^2, \sigma_{\alpha^2} = \sigma_{\alpha}^2 = 0$, and N is large such that $\bar{M}\alpha = \alpha \approx \bar{Y}$, then (3.27) reduces to $deff_C \approx \frac{n\bar{W}}{N}$. In other words, with no correlations, large N, no calibration, and no cluster-level effects, we derive Kish's measure for unit-level sampling.

3.3.3. One Example of Unit-Level Diagnostics: Cook's Distance Measure

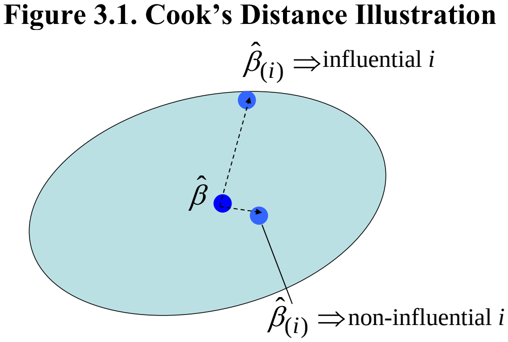
The model-based approach described in Sec. 2.2.2 can also be used to express the case weights associated with the estimators in Ex. 3.1-3.13 in a general form. This leads to developing distance-based functions to identify particular sample units who have more or less influence on the sample-based total. I borrow one method proposed in the statistical regression diagnostics literature (Cook 1977; 1979) to develop a practical metric to assist survey methodologists in determining whether or not a particular case weight should be trimmed, or at least examined carefully.

Cook's distance (Cook 1977; 1979) measures the influence of a particular unit i on estimating the regression coefficient β:

$$CD_i = \left(\hat{\beta}_{(i)} - \hat{\beta}\right)^T \left[Var\left(\hat{\beta}\right)\right]^{-1} \left(\hat{\beta}_{(i)} - \hat{\beta}\right), \tag{3.31}$$

where $\hat{\beta}$ is the estimate of β from the full sample, $\hat{\beta}_{(i)}$ is the estimate of β when deleting unit i, and $Var\left(\hat{\beta}\right)$ is the appropriate variance-covariance matrix of $\hat{\beta}$. The idea is to form a confidence ellipsoid for β and identify any individual points that move $\hat{\beta}$ closer towards the edge of the ellipse, as shown in Figure 3.1.

Figure 3.1. Cook's Distance Illustration



Extending this to weight trimming, we can form the Cook's distance as follows:

$$CD_i\left(\hat{t}\right) = \left(\hat{t}_{(i)} - \hat{t}\right)^T \left[Var\left(\hat{t}\right)\right]^{-1} \left(\hat{t}_{(i)} - \hat{t}\right), \tag{3.32}$$

where $\hat{t}$ is the estimate of the finite population from the full sample before adjusting the weights (e.g., trimming and redistributing the weight), $\hat{t}_{(i)}$ is the estimate of T when removing unit i from the sample and adjusting the other sample units' weights to compensate for its absence, and $Var\left(\hat{t}\right)$ is an appropriate variance-covariance matrix.

3.4. Evaluation Case Studies

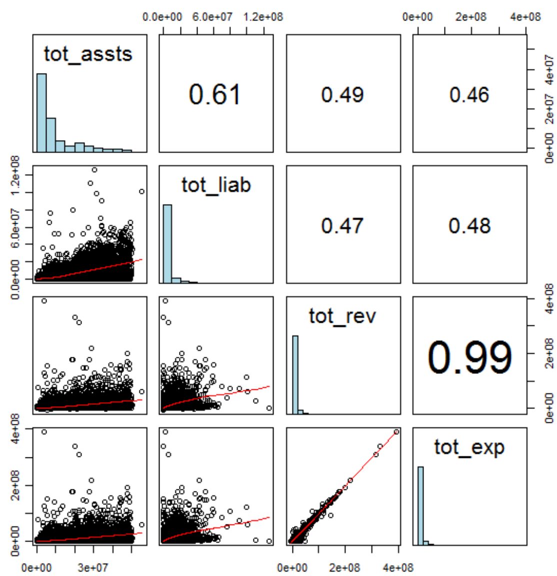
3.4.1. Single-Stage Design Effects Using Establishment Data

Here a sample dataset of tax return data is used to mimic an establishment survey setup. The data come from the Tax Year 2007 SOI Form 990 Exempt Organization (EO) sample. This is a stratified Bernoulli sample of 22,430 EO tax returns selected from 428,719 filed to and processed by the IRS between December 2007 and November 2010. This sample dataset, along with the population frame data, is free and electronically available online (Statistics of Income 2011). These data make a candidate "establishment-type" example dataset for estimating design effects. Since the means and variances of the variables are different across strata, Kish's design effect may not apply.

The SOI EO sample dataset is used here as a pseudopopulation for illustration purposes. Four variables of interest are used: Total Assets, Total Liabilities, Total Revenue, and Total Expenses. Returns that were sampled with certainty and having "very small" assets (defined by having Total Assets less than $1,000,000, including zero) were removed. This resulted in a pseudopopulation of 8,914 units.

Figure 3.2 shows a pairwise plot of the pseudpopulation, including plots of the variable values against each other in the lower left panels, histograms on the diagonal panels, and the correlations among the variables in the upper right panels. This plot mimics establishment-type data patterns. First, from the diagonal panels, we see that the variables of interest are all highly skewed. Second, from the lower left panels, there exists a range of different relationships among them. The Total Assets variable is less related to Total Revenue, and Total Expenses (despite relatively high correlations of 0.46-0.48), while Total Revenue and Total Expenses are highly correlated.

Figure 3.2. Pseudopopulation Values and Loess Lines for Single-Stage Design Effect Evaluation



Three *ppswr* samples were selected ($n = 100;500;1000$) from the pseudopopulation using the square root of Total Assets. The HT weights were then calibrated using the "Linear" method in the calibrate function in the survey package for R (corresponding to a GREG estimator, Lumley 2010) to match the totals of Total Assets and Total Revenue. The analysis variables are thus Total Liabilities and Total Expenses. Figures 3.3 and 3.4

show boxplots and plots of the sample weights before (labeled “HT wt” in Fig. 3.3) and after (“cal wt”) these adjustments.

Figure 3.3. Boxplots of *ppswr* Sample Weights Before and After Calibration Adjustments

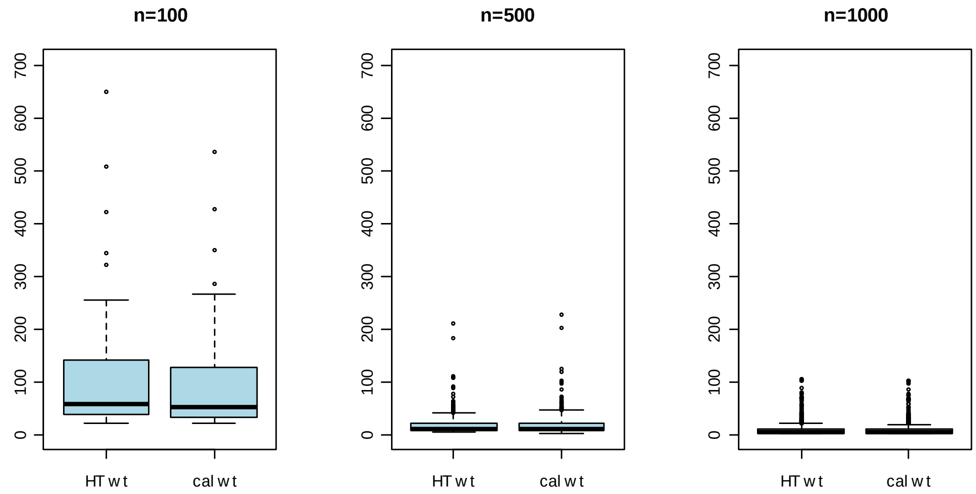
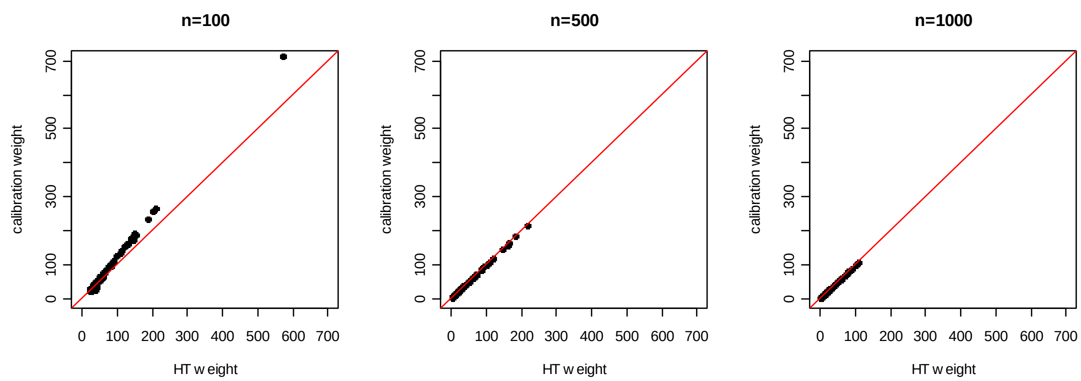


Figure 3.4. Plots of *ppswr* Sample Weights Before and After Calibration Adjustments



Seven estimates of the design effects are considered, with results shown in Table 3.1:

- The Kish measure (3.1);
- Three Spencer measures: the exact measure that estimates (3.7), the approximation (3.9) assuming zero correlation terms, the large-population approximation (3.11)
- Two proposed measures: the exact proposed single-stage design effect (3.18) and the zero-correlation approximation (3.20).

Table 3.1. Single-Stage Sample Design Effect Estimates of *ppswr* Samples Drawn from the SOI 2007 Pseudopopulation EO Data

	Variable of Interest					
	Total Liabilities (weakly correlated with $\mathbf{x}$)			Total Expenses (strongly correlated with $\mathbf{x}$)		
Design Effect Estimates	$n=100$	$n=500$	$n=1000$	$n=100$	$n=500$	$n=1000$
Standard design effects						
Before calibration*	1.31	0.88	1.19	0.92	1.08	1.24
After calibration**	0.56	0.61	0.82	0.03	0.01	0.02
Kish	2.99	2.66	2.49	2.99	2.66	2.49
Spencer						
Exact	0.63	0.37	0.52	0.47	0.66	0.64
Zero-corr. approx.	2.54	1.81	1.76	2.64	2.29	2.06
Large-*N* approx.	3.10	2.34	2.25	3.05	2.78	2.52
Proposed						
Exact	0.48	0.57	0.70	0.03	0.01	0.02
Zero-corr. approx.	1.86	2.07	2.04	0.12	0.05	0.06

* $Var_{\pi}\left(\hat{T}_{\pi}\right)/Var_{srs}\left(\hat{T}_{\pi}\right)$; ** $Var_{\pi}\left(\hat{T}_{GREG}\right)/Var_{srs}\left(\hat{T}_{\pi}\right)$; both measures calculated with R's svytotal function.

Several results are clear from Table 3.1. For this pseudopopulation and *ppswr* samples, the Kish measure is consistently above one for all sample sizes. This measure also does not depend on the variable of interest, and the estimates exceeding two implies the *ppswr* sample design and calibration weighting is inefficient. However, the standard design effects were all well less than one and both Total Liabilities and Total Expenses are positively correlated with the calibration variable Total Revenue (see Fig. 3.2). This conflicts with the Kish measure implications. For both variables, the Spencer measures are all lower than the Kish measures, since they take into account the moderate correlation with the Total Assets variable (which was used to select the *ppswr* samples). However, the exact Spencer design effect estimates that account for the correlations in the weights and errors are all less than one, while the approximations are greater than one. This indicates for this *ppswr* sampling and calibration weighting, the Spencer correlations are not negligible. The same pattern occurred for the proposed design effect for Total Liabilities; the exact measure here was much smaller, and less than one, for all sample sizes. This occurred since this variable had approximately the same correlation (0.48) with the calibration variable Total Revenue (approximately 0.47). However, Total Expenses is highly correlated with the second calibration variable Total Revenue (0.99, see Fig. 3.2), so the proposed design effects are much smaller and closer to each other in value. This implies that the zero-approximation design effect is appropriate when the correlation between the survey and auxiliary variables is extremely strong; otherwise the exact estimate should be used. Figures 3.5 and 3.6 show boxplots of u_i and y_i for each variable and sample size.

Figure 3.5. Boxplots of y_i and u_i-values from *ppswr* Samples from the 2007 SOI EO Data, Total Liabilities Variable

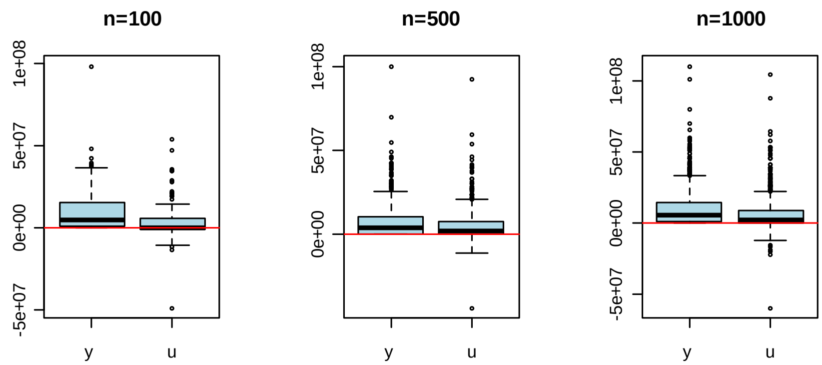
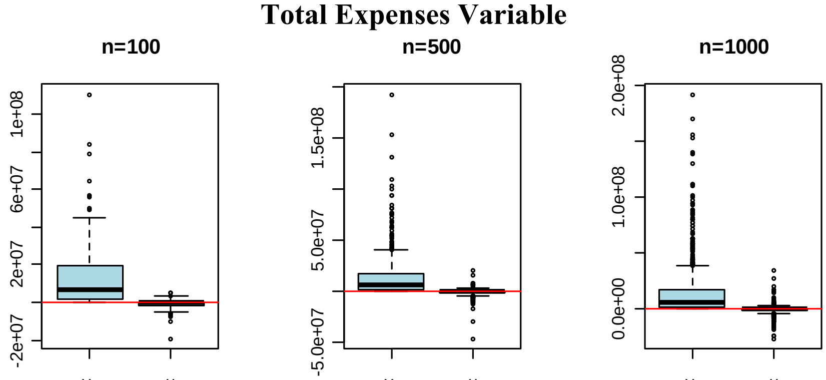


Figure 3.6. Boxplots of y_i and u_i-values from *ppswr* Samples from the 2007 SOI EO Data, Total Expenses Variable



We see that, particularly for the Total Expenses variable, the u_i-values have lower ranges of values and less variation than y_i. This occurs since the Total Expenses variable is highly correlated with the calibration variable Total Revenue (see Figure 3.2). This is why the proposed design effect measures are so much smaller for Total Expenses.

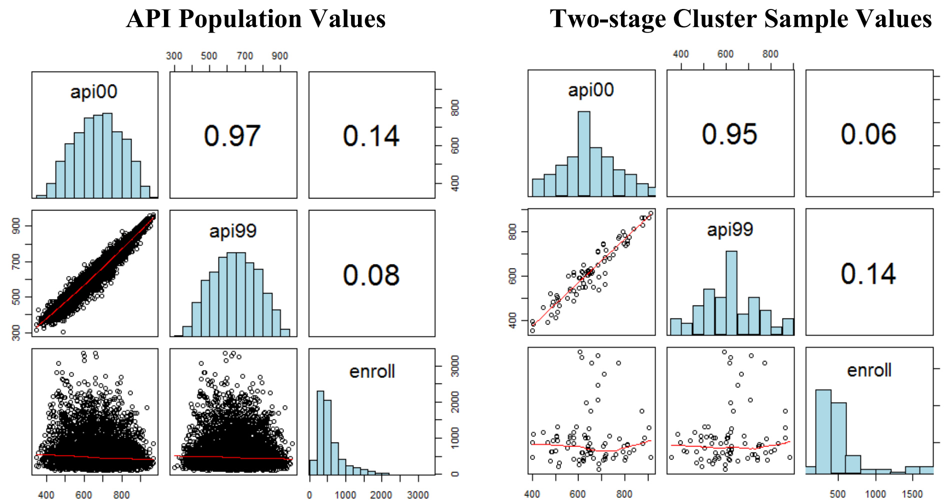
3.4.2. Cluster Sampling Design Effects Using California Education Academic Performance Indicator Data

To illustrate the design effect measures proposed in Sec. 3.2.1, a two-stage sample of children selected within schools from the California Academic Performance Index (API) dataset is used. This dataset is well-documented for the R survey procedures (e.g., Lumley 2010), including the procedures for calibrating two-stage cluster sample weights for schools selected within school districts. Three variables are used: student enrollment and the school's API score for 1999 and 2000. Five-hundred and eighty-nine observations in R's apipop dataset were removed due to missing enrollment values and one outlier district (with 552 schools, creating a dataset with 741 districts and 5,605 schools). To avoid variance estimation complications, for the 182 districts with only one school, one school from another district was randomly sampled using with-replacement simple random sampling and placed within each one-school district. This increased the pseduopopulation dataset to 741 districts (clusters) and 5,787 schools (elements). Forty clusters were selected from the apipop dataset using probability proportional to the number of schools within each district, then two schools were selected from each cluster using simple random sampling without replacement. This resulted in a sample of forty clusters and eighty elements.

Figure 3.7 on the following page shows plots of the API population and two-stage cluster sample values. There are two analysis variables: the API score in 2000 and number of students enrolled; the API score in 1999 is used as the calibration covariate.

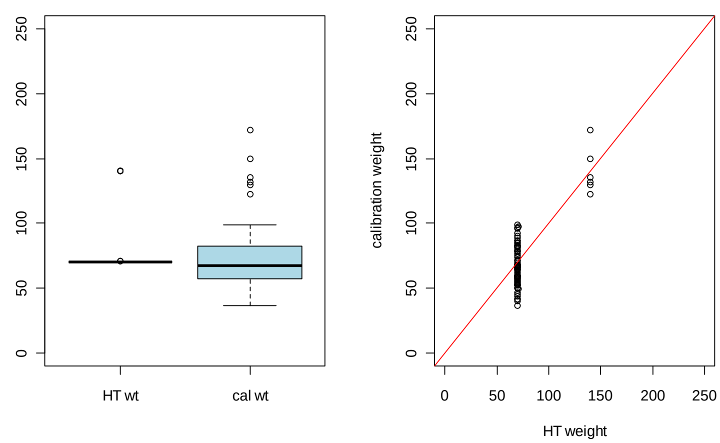
From Figure 3.7, we see that API99 is highly correlated with API00 in both the population and sample, while enrollment is not.

Figure 3.7. API-Population and Two-stage Cluster Sample Values and Loess Lines for Cluster-Level Design Effect Evaluation



After the sample was drawn, the two-stage cluster base weights were calibrated to match the population total number of schools and total API score from 1999. Figure 3.8 shows plots of the weights before (labeled "HTwt") and after ("cal wt") these adjustments.

Figure 3.8. API Two-stage Cluster Sample Weights Before and After Calibration Adjustments



First we see that base weights did not vary much due to the two-stage cluster sample selection method used; the calibration weights are much more varied. The estimated population totals using the cluster-level *ppswr* sample weights for the population size

(6,025) was relatively close to the actual population size (5,787). The base-weighted estimate of total API99 (3,718,984) was very close to actual total (3,718,033), but the varying API99 amounts across the schools produced calibration weights that were more variable than the base weights.

Five estimates of the design effects are considered: the Kish measure (3.1) ignoring the clustering, an ad hoc version of the Kish design effect (3.2) using the Kish measure (3.1) for the first component (since cell-based weighting adjustments were not used), and the three proposed design effect measures--the exact formulation estimate from (3.26), the zero-correlation approximation from (3.28), and approximation when the cluster effects are negligible (3.30). These design effect results are shown in Table 3.2 for each variable of interest.

Table 3.2. Cluster Sample Design Effect Estimates from a Two-Stage Cluster Sample Drawn from the 1999-2000 California Educational Performance Index Data

	Variable of Interest	
Design Effect Estimates	School Enrollment Size	2000 API Test Score
Standard design effects		
Before calibration*	2.90	3.07
After calibration**	3.11	0.45
Kish Methods		
No cluster approx.	3.51	3.51
Ad hoc approx.	10.92	1.60
Proposed Methods		
Exact	1.95	0.12
Zero-corr. approx.	1.95	0.12
Equal cluster size approx.	290.46	17.07

* $Var_{\pi}\left(\hat{T}_{\pi}\right) / Var_{srs}\left(\hat{T}_{\pi}\right)$; ** $Var_{\pi}\left(\hat{T}_{GREG}\right) / Var_{srs}\left(\hat{T}_{\pi}\right)$; both measures calculated with R's svytotal function.

From Table 3.2, we see that the clustering has an effect on the sample-based totals, with standard design effects before calibration exceeding two for both variables. However, calibrating the weights to the total number of schools and total API99 drastically reduces the sample design effect only for the API 2000 score variable (from 3.07 to 0.45). This is expected due to its high correlation with API99 (0.97, see Fig. 3.7). However, the calibration weights do not have this effect for school enrollment, which is weakly correlated with API99 (0.14 in the population). As a result, the standard design effect after calibration is higher than one.

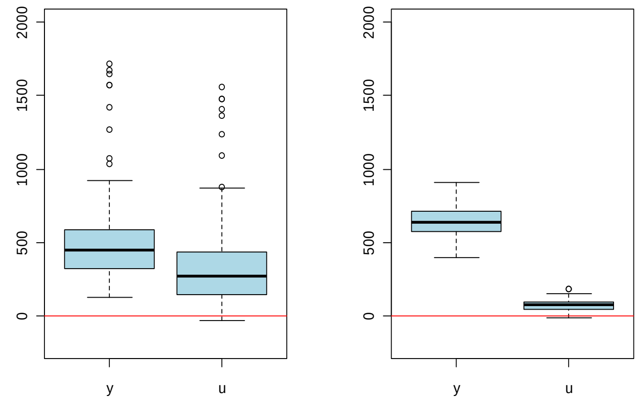
When the Kish approximation methods were applied to the calibration weights, the unit-level sampling measure, labeled “No cluster approx.” in Table 3.2, also exceeds one. This implies that the combination of the cluster sampling and calibration weighting increases the variance of both variables. However, the ad hoc Kish approximation (equal to the product of the “No cluster approx” and the “standard design effect after calibration” design effects) indicates that this sample strategy is not optimal for estimating the total number of student enrollment, but decreases by more than half for the API00 score. The Kish design effects are misleading when compared to the standard (directly computed) ones, which imply the calibration weight adjustment improves the API00 estimate.

For the proposed measures, the exact and zero-correlation design effects are both less than one for API00. However, the equal cluster-size approximation is unusually large. This discrepancy indicates that the variation in cluster sizes cannot be ignored. The proposed design effects for student enrollment also mirror the standard design effects for this variable in that all exceed one. Again, here the equal cluster-size approximation produces an unreasonably large design effect measure. The empirical correlation components in the exact proposed measure are zero for both variables within two decimal points, making the proposed zero-correlation appear exactly equal to the exact formulation for this sample.

Figure 3.9 on shows boxplots of the y_{ij} and u_{ij}-values for each variable.

Figure 3.9. Boxplots of y_{ij} and u_{ij}-values from Two-Stage Cluster Sample from the 1999-2000 California Educational Performance Index Data, by Variable

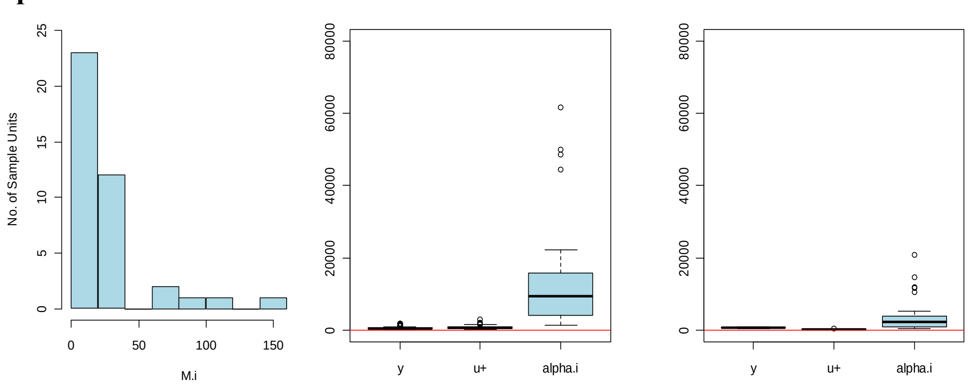
Student Enrollment Size **2000 API Test Score**



Like the single-stage design effect evaluation, here when y_{ij} and $\mathbf{x}_{ij}$ are strongly correlated (in API 2000), the $\hat{u}_{ij}$ values are small and less variable than y_{ij}. However, when the relationship between the survey and calibration variables is weaker (student enrollment), the $\hat{u}_{ij}$ values are smaller but not less variable than y_{ij}. This produced mean and variance components related to $\hat{u}_{ij}$ that are larger for student enrollment than API00, and thus larger design effect estimates. This also holds at the cluster-level; Figure 3.10 shows boxplots of the y_{ij} values against $\hat{u}_{i+}$ and $\hat{\alpha}_i = M_i\hat{\alpha}$.

Figure 3.10. Plots of M_i and Boxplots of y_{ij}, u_{i+}, and $\hat{\alpha}_i$-values from Two-Stage Cluster Sample from the 1999-2000 California Educational Performance Index Data, by Variable

M_i, Population Cluster Size **Student Enrollment Size** **2000 API Test Score**



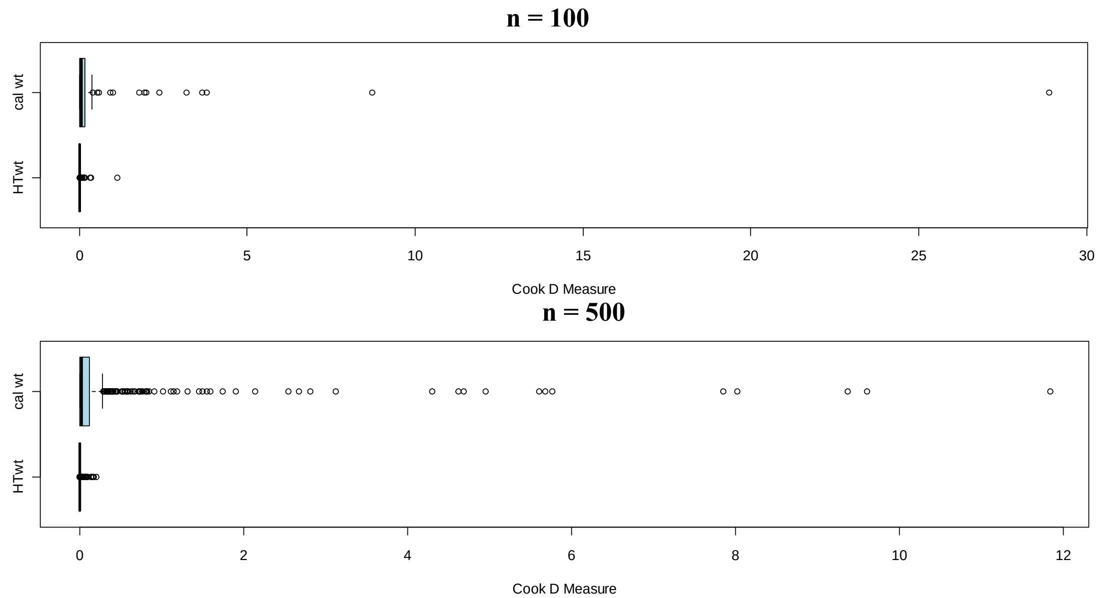
In Figure 3.10, we see that the M_i vary, which means that the $\hat{\alpha}_i$'s also vary. For this sample, the equal-cluster size approximation for both variables is not appropriate. Again, the $\hat{u}_{i+}$ and $\hat{\alpha}_i$ values are much larger and more variable than y_{ij} for the student enrollment variable than API00. This produced mean and variance components related to $\hat{u}_{i+}$ and $\hat{\alpha}_i$ being higher, producing larger proposed design effects than those for the API00 variable. However, this is reasonable since the standard design effects after the calibration weight adjustments were applied indicate that the calibration was more effective for API00 than student enrollment.

3.4.3. Cook's Distance Measure Example

Here the examples provided in Section 3.3.3 are illustrated using the 2007 SOI tax-exempt data used in Sec 3.4.1. In this evaluation, the Cook's Distance measure (3.32) is used to flag particular sample units with values that have the most influence on the estimated totals of the Total Liabilities and Total Expenses variables. This measure is produced for weight-trimming and redistribution adjustments performed on two types of weights: base survey weights and calibration weights (calibrating to the population size and sum of Total Revenue). Samples of $n=100$ and 500 units were drawn from the pseudopopulation data using probability proportional to the size of the square root of Total Assets. Each unit was omitted from the sample, its weight was equally redistributed to the other sample units' weights, the total $\hat{t}_{(i)}$ was estimated, then the Cook's D measure was calculated.

Figure 3.11 on the following page shows the boxplot distributions of the Cook's D measures (note a difference in scale between the sample sizes).

Figure 3.11. Boxplots of Cook's Distance Measures for HT and Calibration Weights, *ppswr* Samples from the 2007 SOI EO Data



Since larger values of Cook's D measures indicate that trimming a particular sample unit's weight has greater impact on the estimated total, cases with the largest absolute values of the Cook's D measures are the most likely candidates for data edit and review. Thus, seeing how skewed this measure is in Fig. 3.11, examining the largest values is a sensible approach. A univariate measure was calculated for Total Expenses. The sample data and largest Cook's D measures are shown in Tables 3.3 and 3.4 (five values for $n = 100$ and ten for $n = 500$).

Table 3.3. Sample Data with Largest Five Values of Cook's D Measures for Total Expenses, *ppswr* Samples from the 2007 SOI EO Data, n=100

Base Weights				*Calibration Weights*			
Case ID	**Weight**	y_i **-Value**	**Cook's D**	**Case ID**	**Weight**	y_i **-Value**	**Cook's D**
2330	877.17	422,895	1.14	2330	1,013.34	422,895	28.89
3203	504.75	1,140,408	0.35	3203	581.76	1,140,408	8.72
19225	39.09	219,296,760	0.33	2080	358.71	25,140	3.79
6155	113.06	58,715,572	0.16	3604	396.34	1,906,734	3.67
2080	310.56	25,140	0.15	6155	137.70	58,715,572	3.18

Table 3.4. Sample Data with Largest Ten Values of Cook's D Measures for Total Expenses, *ppswr* Samples from the 2007 SOI EO Data, n=500

Base Weights				*Calibration Weights*			
Case ID	**Weight**	y_i **-Value**	**Cook's D**	**Case ID**	**Weight**	y_i **-Value**	**Cook's D**
10543	208.11	2,427,593	0.20	10543	197.64	2,427,593	11.84
886	165.04	926,507	0.17	886	156.61	926,507	9.60
7574	163.19	938,631	0.17	7574	154.83	938,631	9.38
8317	145.36	450,943	0.14	8317	137.88	450,943	8.02
14391	146.00	946,091	0.14	14391	138.50	642,091	7.86
12883	166.67	3,843,079	0.10	12883	158.36	3,843,079	5.77
10127	120.88	295,769	0.10	10127	114.66	295,769	5.68
12152	47.70	45,559,752	0.10	12152	46.51	45,559,752	5.61
9630	112.08	569,976	0.08	9630	106.33	569,976	4.95
9114	118.53	1,348,550	0.08	9114	112.52	1,348,550	4.68

Within each sample, we see that many of the same cases are flagged as being the most influential units when trimming their weights and estimating the sample-based total. Cases are identified as "influential" by having a large combination of the survey value and the weight, which includes values with large weights and moderate y_i-values (such as cases 2330 and 2080). Interestingly, for the sample of size 100, the calibration weighting has controlled one of the influential observations (case ID 19225), but makes another observation influential that was not influential when combined with its base weights (case ID 3604). The first observation has a larger Cook D measure since the weight is increased and the calibration variance was lower than that of the HT estimator. For the sample of size 500, while we see that the calibration weights are smaller than the base weights, the same set of sample cases remain the most influential for both sets of weights (and in the same order). However, all of these observations have higher Cook's D measures using the calibration weights since the variance of the total was lower than the HT estimator variance.

When large Cook's D measures are produced, this can indicate that the associated case is influential on estimating the total due to a large weight (e.g., case 2330 in Table

3.3), large survey value (case 19335 in Table 3.3), or a large combination of both (case 10463 in Table 3.4). Thus, large Cook's D values are indicators for further investigation into particular cases to determine what caused the measure to be large. If the survey value is large, then data editing techniques are more applicable; if the weight or the product of the weight and the survey variable are large, then trimming the weights may be a more appropriate solution. Note that these effects can vary by each survey variable; a multivariate extension may be simpler than producing several univariate measures to investigate independently.

3.5. Discussion and Limitations

For this paper, I propose new diagnostic measures that attempt to gauge the impact of weighting adjustments on a sample-based total in both single-stage and cluster sampling. In the design effect evaluations, the existing Kish design effect measures produced misleading results of design effects that were too high, particularly for the single-stage case study in Sec. 3.4.1. The empirical results also demonstrate that the correlation components in Spencer's design effect were not always negligible for the data examined.

However, the proposed design effect gauges the impact of variable calibration weights by using the GREG AV variance approximation for the class of calibration estimators. As demonstrated empirically in Sec. 3.4.1 and 3.4.2, the proposed design effects do not penalize unequal weights when the relationship between the survey variable and calibration covariate is strong. However, high correlations between survey and auxiliary variables may be unattainable for some surveys that lack auxiliary information further than population counts to use in poststratification (e.g., many household surveys).

The proposed design effects also do not incorporate additional weighting adjustments beyond calibration, such as trimming outlying weights. Additional modifications would incorporate a mean square error rather than a variance, which is more difficult to estimate from one given sample. It is also noteworthy that both the Spencer and exact design effect estimates can be negative due to negative correlations; in these cases the large-N or zero-correlation approximations should be considered. In addition, calibration weights that are negative should be bounded in order to produce the design effect estimates; several methods (see Sec. 1.1 for examples) to do this exist.

The case study evaluation in Sec. 3.4.3 demonstrates that the proposed unit-level diagnostic measures have promise for use in samples. They "successfully" identified pre-identified influential observations in the sample, produced by large weights, large survey values, or a combination of both. However, the most severe limitation with the proposed case-level measures is absence of theoretical properties under finite population sampling.

Future Work

Since the approaches in Papers 1 and 2 are model-based, they share similar extensions for future consideration. While the weight smoothing methods examined in Paper 1 have serious defects, there may exist certain circumstances in which this method is appropriate. Future work could include researching these circumstances. Once they are established, more complex models, such as those for cluster sampling, can be developed. Extensions to the model-assisted *p*-spline approach are also somewhat limited in that this method is limited to survey designs in which quantitative calibration covariates are readily available, excluding common household surveys; however the models proposed here can also be extended to cluster sampling.

Future considerations for the proposed design effect measures include extending the cluster-level measures to some sample selection method other than simple random sampling within a cluster. This would produce a more complicated form of the second variance component. However, the current design effect can be used for equal and unequal sampling of the PSU's; the "exact" estimator holds for unequal probability sampling, while the "zero-correlation" approximation can be used in equal probability sampling of the clusters.

Also, there are additional unit-level diagnostics that can be considered. For example, we can extend the Cook's Distance idea. If we identify a particular set of weights w_i^* that is "preferable" to the original, unadjusted weights w_i, then we can form

$$CD2_i\left(\hat{T}_1,\hat{T}_2\right)=\left(\hat{T}_1-\hat{T}_2\right)^T\left[Var\left(\hat{T}_1-\hat{T}_2\right)\right]^{-1}\left(\hat{T}_1-\hat{T}_2\right), \quad (1)$$

where $\hat{T}_1 = \sum_{i \in s} w_i y_i$, $\hat{T}_2 = \sum_{i \in s} w_i^* y_i$ are vectors of the estimated totals using the unadjusted and adjusted weights, respectively. In (1), it will be important to incorporate the covariance between $\hat{T}_1$ and $\hat{T}_2$, since this is expected to be high due to common y_i. This could also be treated as a hypothesis test (e.g., Pfeffermann 1993), since we can write $\hat{T}_1 - \hat{T}_2 = \sum_{i \in s} \left(w_i - w_i^* \right) y_i$. Under certain properties, $\hat{T}_1 - \hat{T}_2$ is asymptotically normal, even if the weights are complex (e.g., if the weights are the product of separate adjustment factors like nonresponse or poststratification).

It is also possible to borrow other methods proposed in the statistical literature to develop practical metrics to assist survey methodologists in determining whether or not a particular case weight should be trimmed, or at least examined carefully. Examples include distributional-based summary measures like the Kolmogorov-Smirnov test, the Cramér-von Mises test, and the Anderson-Darling Test. Each is described next.

The Kolmogorov-Smirnov (KS; Kolmogorov 1933; Smirnov 1948) statistic is a nonparametric method developed to find a confidence band for the distribution of a continuous random variable. The sign test and rank sum tests can be used for discrete variables (Mann and Whitney 1947; Dixon and Massey 1966). There is a one- and two-sample version of KS test. The one-sample test involves gauging whether a particular set of weights follows a prescribed statistical distribution. This can gauge the usefulness of ad hoc weight trimming methods that use statistical distribution values for cutoffs (described in Sec. 1.2.1). The two-sample test involves testing whether two sets of weights follow a common distribution. This test can be used as a summary measure to gauge the impact of weighting adjustments before and after they are applied, or compare a candidate weights against a more "preferable" set of weights.

For the one-sample KS statistic, if the weights $w_1^T, w_2^T, \ldots, w_n^T$ are viewed as a random sample from the population with cumulative distribution function $F(w)$ and $w_1, w_2, \ldots, w_n$ denote the ordered sample weights. The ordered weights are used to construct upper and lower step functions, where $F(w)$ is contained between them with a specified probability. The sample distribution function is the following step function:

$$S_n(w) = \begin{cases} 0, & \text{if } w < w_1 \\ \dfrac{k}{n}, & \text{if } w_k \le w \le w_{k+1} \\ 1, & \text{if } w \ge w_n \end{cases}. \tag{2}$$

If the function $F(w)$ is known, then it is possible to calculate $\left|F(w) - S_n(w)\right|$ for any desired value of w. It is also possible to calculate

$$D_n = \max_w \left|F(w) - S_n(w)\right|, \tag{3}$$

the maximum vertical distance between $F(w)$ and $S_n(w)$ over the range of possible w-values. Since $S_n(w)$ varies by sample, D_n is a random variable. However, since the distribution of D_n does not depend on $F(w)$, D_n can be used as a nonparametric variable for constructing a confidence band for $F(w)$. Combinatorial methods can be used to find the distribution of D_n for a particular n (Pollard 2002). Examples of critical values from this distribution for particular values of α are given in Hoel (Table

VIII 1962). In general, denote D_n^α as the α-level critical value, i.e., satisfying $P\left(D_n \le D_n^\alpha\right) = 1-\alpha$. From this, the following equalities hold:

$$\begin{aligned} P\left(D_n \le D_n^\alpha\right) &= 1-\alpha \\ &= P\left\{\max_w \left|F(w) - S_n(w)\right| \le D_n^\alpha\right\} \\ &= P\left\{\left|F(w) - S_n(w)\right| \le D_n^\alpha \text{ for all } w\right\} \\ &= P\left\{S_n(w) - D_n^\alpha \le F(w) \le S_n(w) + D_n^\alpha \text{ for all } w\right\} \end{aligned} . \tag{3}$$

The last equality in (3) shows how the two step functions, $S_n(w) - D_n^\alpha$ and $S_n(w) + D_n^\alpha$, produce a $1-\alpha$ level confidence interval for the unknown distribution $F(w)$.

The two-sample KS test is a variation of the one-sample and a generalization of the two-sample *t*-test. Instead of comparing the empirical weights distribution function to some theoretical distribution function, we compare between two empirical distribution functions and formally test whether or not the samples come from a common distribution. This test is more appropriate in gauging the impact of weighting adjustments before and after they are applied to a particular set of weights.

For two sets of n weights, denoted w_1 and w_2, the two-sample KS statistic is

$$D_n = \max\left|F(w_1) - F(w_2)\right|, \tag{4}$$

where $F(w_1)$ and $F(w_2)$ are the empirical distribution functions of the two sets of weights. Again, critical values for D_n can be obtained in tables or conventional software (e.g., R). This test can also be used to compare between alternative case weights. Here w_1 is a "preferable" set of weights, such as the Breidt *et. al* (2005) robust calibration weights, while $F(w_2)$ is the distribution of some candidate weights, e.g., some other nonresponse-adjusted calibrated set of weights that are simpler to produce. One approach would be to compare the distributions of the weights before and after calibration adjustments and use a formal hypothesis test of whether or not the distributions of the two sets of weights are similar, rejecting a null hypothesis of similarity at level α if the test statistic $\sqrt{\frac{n^2}{2n}}D_n$ (Stephens 1979; Marsaglia *et al*. 2003), exceeds the associated critical value of the D_n distribution. The values of $\left|F\left(w_1\right)-F\left(w_2\right)\right|$ can also be viewed simply as descriptive statistics for comparing two sets of weights.

Another potential diagnostic to gauge the impact of weighting adjustments is the Cramér-von-Mises test. Anderson (1962) generalized this test for two samples. Here, for $w_1, w_2, \ldots, w_n$ and $w_1^*, w_2^*, \ldots, w_n^*$ denoting two sets of ordered weights, and $r_1, r_2, \ldots, r_n$ the ranks of the weights $w_1, w_2, \ldots, w_n$ when combined and $s_1, s_2, \ldots, s_n$ the ranks of the weights $w_1^*, w_2^*, \ldots, w_n^*$ when combined. To test the hypothesis that $w_1, w_2, \ldots, w_n$ and $w_1^*, w_2^*, \ldots, w_n^*$ are equivalent, Anderson (1962) developed the test statistic

$$T = \frac{n\sum_{i=1}^{n}\left(r_i - i\right)^2 + n\sum_{i=1}^{n}\left(s_j - j\right)^2}{2n^3} - \frac{4n^2-1}{12n}, \qquad (5)$$

which is compared to a predetermined critical value from an $F\left(n,n\right)$ distribution. Expression (5) assumes that there are no ties in the ranks, but alternative methods (e.g., using “mid-ranks,” e.g., Ruymgaart 1980; Stephens 1986) have been developed.

The Cramér-von-Mises test is a special case of the Anderson-Darling test statistic (Anderson 1962; Darling 1952). The two-sample test statistic is

$$A = \frac{n^2}{2n} \int_{-\infty}^{\infty} \frac{\left(F\left(w_1\right) - F\left(w_2\right)\right)^2}{F\left(w\right)\left(1 - F\left(w\right)\right)} dF\left(w\right), \tag{6}$$

where $F\left(w\right) = \frac{F\left(w_1\right) + F\left(w_2\right)}{2}$ is the empirical distribution function of the pooled samples. For two samples, measure (6) can be used to test the hypothesis of the weights following the same distribution without actually specifying the common distribution (Scholz and Stephens 1987).

For all of these tests, for two sets of candidate weights, it would also be sensible to compare the weighted empirical distribution functions for different survey variables and compare differences between them. This type of comparison has some promise in identifying whether two sets of weights lead to noticeably different estimates of extreme quantiles.

Appendices

Appendix 1: HT and Beaumont Estimator Expectation and Consistency Properties

Beaumont (2008) derived some of the theory contained in this Appendix. However, it is detailed here to illustrate how the theory differs from a conventional model-based approach (i.e., one that posits a model for the survey response variable, not the weights). Additional theory that is developed and presented here that Beaumont did not derive is also identified as such.

By definition, $\hat{T}_{HT} = \sum_{i \in s} w_i y_i$, where $w_i = \pi_i^{-1}$, is the HT estimator. An estimator proposed to reduce the variability in the w_i's replaces them with their conditional expected value (conditional on the sample and a model for the weights):

$$\begin{aligned} \tilde{T}_B &= E_M\left(\hat{T}_{HT} \middle| \mathbf{I}, \mathbf{Y}\right) \\ &= E_M\left(\sum_{i \in s} w_i y_i \middle| \mathbf{I}, \mathbf{Y}\right) \\ &= \sum_{i \in s} E_M\left(w_i y_i \middle| \mathbf{I}, \mathbf{Y}\right), \\ &= \sum_{i \in s} E_M\left(w_i \middle| \mathbf{I}, \mathbf{Y}\right) y_i \\ &= \sum_{i \in s} \tilde{w}_i y_i \end{aligned} \quad \text{(A.1)}$$

where $\mathbf{I} = \left(I_1, \ldots, I_N\right)^T$ is the vector of sample inclusion indicators and $\mathbf{Y} = \left(Y_1, \ldots, Y_N\right)^T$ are the values of the survey response variable y_i. Since $\tilde{w}_i = E_M\left(w_i \middle| \mathbf{I}, \mathbf{Y}\right)$ is unknown, we estimate it with $\hat{w}_i$. The estimator for the finite population total is then $\hat{T}_B = \sum_{i \in s} \hat{w}_i y_i$. If our model for the weights, denoted by M, is correct, then we have

$$
\begin{aligned}
E_M\left(\hat{T}_B \middle| \mathbf{Y}\right) &= E_M\left(\sum_{i\in s}\hat{w}_i y_i \middle| \mathbf{Y}\right) \\
&= \sum_{i\in s} E_M\left(\hat{w}_i y_i \middle| \mathbf{Y}\right) \\
&= \sum_{i\in s} E_M\left(\hat{w}_i \middle| \mathbf{Y}\right) y_i \;. \\
&= \sum_{i\in s} \tilde{w}_i y_i \\
&= \tilde{T}_B
\end{aligned}
\tag{A.2}
$$

Generalized design-based inference is defined as "any inference that is conditional on $\mathbf{Y}$ but not $\mathbf{I}$." Probability sampling is assumed, such that $p\left(\mathbf{I}\middle|\mathbf{Z},\mathbf{Y}\right) = p\left(\mathbf{I}\middle|\mathbf{Z}\right)$. For inferential purposes, we also consider $\mathbf{Z} = \left(Z_1,\ldots,Z_N\right)^T$, the vector of design-variables. Beaumont (2008) takes expectations with respect to the joint distribution of $\mathbf{Z}$ and $\mathbf{I}$, conditional on $\mathbf{Y}$, denoted by $F_{\mathbf{Z},\mathbf{I}|\mathbf{Y}}$. Estimators are evaluated with respect to the sample design and the model for the weights, denoted by $E_F = E_M\left[E_\pi\left(\bullet\right)\right]$ or $E_F = E_F\left[E_M\left(\bullet\right)\right]$. Under this approach, several properties hold.

<u>Property 1: The HT estimator is always unbiased across the model and designs</u>

Assuming that $E_\pi\left[I_i\middle|\mathbf{Y}\right] = w_i^{-1}$, we have

$$
\begin{aligned}
E_F\left(\hat{T}_{HT}\right) &= E_\pi\left[E_M\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right] \\
&= E_\pi\left[E_M\left(\sum_{i\in s} w_i y_i\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right] \\
&= E_\pi\left[\sum_{i\in s} E_M\left(w_i\middle|\mathbf{Z},\mathbf{Y}\right) y_i\middle|\mathbf{Y}\right] \\
&= E_\pi\left[\sum_{i\in s} w_i y_i\middle|\mathbf{Y}\right] \\
&= T
\end{aligned}
\tag{A.3}
$$

This property holds in "both directions" of expectation:

$$
\begin{aligned}
E_M\left[E_\pi\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right] &= E_M\left[E_\pi\left(\sum_{i\in s} w_i y_i\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right] \\
&= E_M\left[\sum_{i\in U} E_\pi\left(I_i\middle|\mathbf{Z},\mathbf{Y}\right) w_i y_i\middle|\mathbf{Y}\right] \\
&= E_M\left[\sum_{i\in U} \pi_i w_i y_i\middle|\mathbf{Y}\right] \\
&= \sum_{i\in U} E_M\left(y_i\middle|\mathbf{Y}\right) \\
&= T
\end{aligned} \qquad . \qquad \text{(A.4)}
$$

<u>Property 2: If the model for the weights is right, then the smoothed HT estimator is unbiased</u>

Similar to the HT estimator proofs, for $\hat{T}_B = \sum_{i\in s} \hat{w}_i y_i$, we have

$$
\begin{aligned}
E_F\left(\hat{T}_B\right) &= E_\pi\left[E_M\left(\hat{T}_B\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right] \\
&= E_\pi\left[E_M\left(\sum_{i\in s} \hat{w}_i y_i\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right] \\
&= E_\pi\left[\sum_{i\in s} \tilde{w}_i y_i\middle|\mathbf{Y}\right] \text{ if } M \text{ is correct} \\
&= E_\pi\left[E_M\left(\sum_{i\in s} w_i y_i\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right] \text{by definition of } \tilde{T}_B \\
&= E_M\left[E_\pi\left[\sum_{i\in s} w_i y_i\middle|\mathbf{Y}\right]\middle|\mathbf{Z},\mathbf{Y}\right] \\
&= E_M\left[T\middle|\mathbf{Z},\mathbf{Y}\right] \\
&= T
\end{aligned} \qquad \text{(A.5)}
$$

The result is a consequence of the fact that, under the model for the weights, $E_M\left(\tilde{w}_i\right) = w_i$ .We can also reach this result as follows (as Beaumont does, in a convoluted sort of way):

$$
\begin{aligned}
E_\pi\left[E_M\left(\hat{T}_B \middle| \mathbf{I},\mathbf{Y}\right)\middle|\mathbf{Y}\right] &= E_\pi\left[E_M\left(\tilde{T}_B \middle| \mathbf{Y}\right)\right] \text{ if } M \text{ is correct} \\
&= E_\pi\left[E_M\left(\textstyle\sum_{i\in s} w_i y_i \middle| \mathbf{I},\mathbf{Y}\right)\middle|\mathbf{Y}\right] \text{ by definition of } \tilde{T}_B \\
&= E_\pi\left[\underbrace{\textstyle\sum_{i\in s} E_M\left(w_i y_i I_i \middle| \mathbf{I},\mathbf{Y}\right)}_{I_i=1} + \underbrace{\textstyle\sum_{i\in r} E_M\left(w_i y_i I_i \middle| \mathbf{I},\mathbf{Y}\right)}_{I_i=0}\middle|\mathbf{Y}\right] \\
&= E_\pi\left[\textstyle\sum_{i\in s} E_M\left(w_i \middle| \mathbf{I},\mathbf{Y}\right) y_i + (0)\middle|\mathbf{Y}\right] \qquad \text{(A.6)} \\
&= E_\pi\left[\textstyle\sum_{i\in s} E_M\left(w_i \middle| \mathbf{I},\mathbf{Y}\right) y_i \middle|\mathbf{Y}\right] \\
&= E_\pi\left[\hat{T}_{HT} \middle| \mathbf{Y}\right] \\
&= T
\end{aligned}
$$

<u>Property 3: If the model for the weights does not hold, then the Beaumont estimator is not unbiased</u>

While Beaumont presented the unbiasedness proof, he did not indicate that when the model is wrong, his estimator is not unbiased. That proof, as well as two examples under specific models for the weights, is detailed here. Similar to the Property 2 proof, for $\hat{T}_B = \sum_{i\in s} \hat{w}_i y_i$, we have

$$
\begin{aligned}
E_F\left(\hat{T}_B\right) &= E_\pi\left[E_M\left(\hat{T}_B \middle| \mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right] \\
&= E_\pi\left[E_M\left(\textstyle\sum_{i\in s} \hat{w}_i y_i \middle| \mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right] \\
&= E_\pi\left[\textstyle\sum_{i\in s} E_M\left(\hat{w}_i \middle| \mathbf{Z},\mathbf{Y}\right) y_i \middle|\mathbf{Y}\right] \text{ if } M \text{ is wrong} \\
&= \textstyle\sum_{i\in U} E_M\left(\hat{w}_i \middle| \mathbf{Z},\mathbf{Y}\right) y_i E_\pi\left(I_i \middle| \mathbf{Y}\right) \qquad \text{(A.7)} \\
&= \textstyle\sum_{i\in U} E_M\left(\hat{w}_i \middle| \mathbf{Z},\mathbf{Y}\right) y_i \pi_i \\
&\neq \textstyle\sum_{i\in U} y_i \\
&= T
\end{aligned}
$$

The extent to the bias in (A.7) depends on how far the expected value $E_M\left(\hat{w}_i \middle| \mathbf{Z},\mathbf{Y}\right)$ is from $w_i = \pi_i^{-1}$:

$$\begin{aligned} E_F\left(\hat{T}_B - T\right) &= \sum\nolimits_{i\in U} E_M\left(\hat{w}_i \middle| \mathbf{Z},\mathbf{Y}\right) y_i \pi_i - \sum\nolimits_{i\in U} y_i \\ &= \sum\nolimits_{i\in U} \left[E_M\left(\hat{w}_i \middle| \mathbf{Z},\mathbf{Y}\right) - w_i \right] \pi_i y_i \qquad . \\ &= \sum\nolimits_{i\in U} \left(\tilde{w}_i \pi_i - 1\right) y_i \end{aligned} \tag{A.8}$$

Since the Beaumont estimator involves replacing weights with their predicted means, (A.8) can be derived for special circumstances. Beaumont did not derive this theoretical result. Two examples follow.

Ex. A1. Suppose that the exponential weights model is correct, but the linear model is used. Then the bias in the Beaumont estimator is

$$\begin{aligned} E_F\left(\hat{T}_B - \tilde{T}_B\right) &= \sum\nolimits_{i\in U} \hat{w}_i y_i \pi_i - \sum\nolimits_{i\in U} E_M\left(\hat{w}_i \middle| \mathbf{Z},\mathbf{Y}\right) y_i \pi_i \\ &= \sum\nolimits_{i\in U} \mathbf{H}_i^T \hat{\boldsymbol{\beta}} y_i \pi_i - \sum\nolimits_{i\in U} \left(1 + \exp\left(\mathbf{H}_i^T \boldsymbol{\beta}\right)\right) y_i \pi_i \qquad . \\ &= \sum\nolimits_{i\in U} \left(\mathbf{H}_i^T \hat{\boldsymbol{\beta}} - 1 - \exp\left(\mathbf{H}_i^T \boldsymbol{\beta}\right)\right) y_i \pi_i \end{aligned} \tag{A.9}$$

If the exponential can be approximated by the first two terms in a MacLaurin series, i.e., $\exp\left(\mathbf{H}_i^T \boldsymbol{\beta}\right) \doteq 1 + \mathbf{H}_i^T \boldsymbol{\beta}$, then $\hat{T}_B$ is approximately unbiased.

Ex. A2. Suppose that the inverse weights model is correct, but the linear model is used. Then the bias in the Beaumont estimator is

$$
\begin{aligned}
E_F\left(\hat{T}_B - \tilde{T}_B\right) &= \sum_{i\in U} \hat{w}_i y_i \pi_i - \sum_{i\in U} E_M\left(\hat{w}_i \middle| \mathbf{Z},\mathbf{Y}\right) y_i \pi_i \\
&= \sum_{i\in U} \mathbf{H}_i^T \hat{\boldsymbol{\beta}} y_i \pi_i - \sum_{i\in U}\left(\mathbf{H}_i^T \boldsymbol{\beta}\right)^{-1} y_i \pi_i \\
&= \sum_{i\in U}\left(\mathbf{H}_i^T \hat{\boldsymbol{\beta}} - \left(\mathbf{H}_i^T \boldsymbol{\beta}\right)^{-1}\right) y_i \pi_i
\end{aligned}
\quad . \qquad \text{(A.10)}
$$

<u>Property 4: Consistency of the HT Estimator</u>

Beaumont (2008) provided the theory that the HT and Beaumont estimators are consistent. However, it is not clear until examining the details of the theory that the latter holds only if the weights model is correct. This section thus contains the details of booth proofs.

In order to establish the consistency of the HT estimator under the weights model and the sample design, we need to make the following assumption:

Assumption 1. $E_\pi\left[Var_M\left(\hat{T}_{HT} \middle| \mathbf{Z},\mathbf{Y}\right) \middle| \mathbf{Y}\right] = O\left(N^2/n\right)$.

Assumption 1 thus implicitly implies that, for totals we have $N^{-1}\left(\hat{T}_{HT} - T\right) \rightarrow 0$ as $n \rightarrow \infty$, which implies $E_\pi\left[Var_M\left(\hat{T}_{HT} \middle| \mathbf{Z},\mathbf{Y}\right) \middle| \mathbf{Y}\right] = O(N)$. Under Assumption 1, we have

$$
\begin{aligned}
Var_F\left(\hat{T}_{HT} \middle| \mathbf{Y}\right) &= E_\pi\left[Var_M\left(\hat{T}_{HT} \middle| \mathbf{Z},\mathbf{Y}\right) \middle| \mathbf{Y}\right] + Var_\pi\left[E_M\left(\hat{T}_{HT} \middle| \mathbf{Z},\mathbf{Y}\right) \middle| \mathbf{Y}\right] \\
&= E_\pi\left[Var_M\left(\hat{T}_{HT} \middle| \mathbf{Z},\mathbf{Y}\right) \middle| \mathbf{Y}\right] + Var_\pi\left[\tilde{T}_B \middle| \mathbf{Y}\right] \\
&= O\left(N^2/n\right) \quad \text{under Assumption 1}
\end{aligned}
\quad . \qquad \text{(A.11)}
$$

Note: result (A.11) only holds assuming that $Var_{\pi}\left[\tilde{T}_B \middle| \mathbf{Y}\right]$ is also $O\left(N^2/n\right)$. Since $\hat{T}_{HT}$ was proved to be unbiased in (A.3) and (A.4), using the bounded variance theorem (e.g., Thm. 6.2.1 in Wolter 1985), we have

$$\begin{aligned} \hat{T}_{HT} - T &= N\left[\bar{y}_{HT} - \bar{Y}\right] \\ &= O_p\left(N/\sqrt{n}\right) \end{aligned} . \tag{A.12}$$

Property 5: Consistency of the Beaumont Estimator When the Model is Right

It was shown in (A.5) and (A.6)and (A6) that $E_F\left(\tilde{T}_B \middle| \mathbf{Y}\right) = T$. From this,

$$\begin{aligned} Var_F\left(\tilde{T}_B \middle| \mathbf{Y}\right) &= E_{\pi}\left[Var_M\left(\tilde{T}_B \middle| \mathbf{Z},\mathbf{Y}\right) \middle| \mathbf{Y}\right] + Var_{\pi}\left[E_M\left(\tilde{T}_B \middle| \mathbf{Z},\mathbf{Y}\right) \middle| \mathbf{Y}\right] \\ &= E_{\pi}\left[Var_M\left(\tilde{T}_B \middle| \mathbf{Z},\mathbf{Y}\right) \middle| \mathbf{Y}\right] + Var_{\pi}\left[\hat{T}_{HT} \middle| \mathbf{Y}\right] \\ &= O\left(N^2/n\right) \text{ under Assumption 1} \end{aligned} . \tag{A.13}$$

Similarly,

$$\tilde{T}_B - T = O_p\left(N/\sqrt{n}\right). \tag{A.14}$$

For totals, assuming that T/N converges to some constant, result (A.14) implies $\frac{1}{N}\left(\tilde{T}_B - T\right) \xrightarrow[n\to\infty]{p} 0$, i.e., $\tilde{T}_B$ is consistent for T. By Beaumont (p. 544) "In practice, we expect that [$\hat{T}_B$] inherits properties of [$\tilde{T}_B$]." However, if the model for the weights is wrong, there is no guarantee that this happens. Assuming that the weights model is correct, then properties (A.11) and (A.12) also hold:

$$Var_F\left(\hat{T}_B \middle| \mathbf{Y}\right) = O\left(N^2/n\right). \tag{A.15}$$

and

$$\hat{T}_B - T = O_p\left(N/\sqrt{n}\right). \tag{A.16}$$

Appendix 2: Additional Properties of Beaumont Estimator Under Linear Weights Model

Property 6: Properties of $\boldsymbol{\beta}$ under the Model

Beaumont (2008) provided additional theory had holds under a linear model for the weights. Since this model is posited for the proposed variance estimators, the details of (A.13) under Model 1 are given here. It is simpler to use matrix notation. Under Model 1, $E\left(w_i \middle| \mathbf{I}, \mathbf{Y}\right) = \mathbf{H}_i^T \boldsymbol{\beta} + v_i^{1/2} \boldsymbol{\varepsilon}$, where $\mathbf{H}_i = \mathbf{H}_i\left(\mathbf{y}_i\right)$ is a vector of the specified function of different y-values for unit i and $\boldsymbol{\varepsilon} \sim \left(0, \sigma^2\right)$ are independent. The predicted weights are $\hat{w}_i = \mathbf{H}_i^T \hat{\boldsymbol{\beta}}$, such that $\hat{T}_B = \hat{\mathbf{w}}^T \mathbf{y}_s$, where $\hat{\mathbf{w}} = \mathbf{H}^T \hat{\boldsymbol{\beta}} = \left(\hat{w}_1 \ldots \hat{w}_n\right)^T$ is the vector of predicted weights, and $\mathbf{H} = \left[\mathbf{H}_1^T \boldsymbol{\beta} \quad \mathbf{H}_2^T \boldsymbol{\beta} \quad \cdots \quad \mathbf{H}_n^T \boldsymbol{\beta}\right]^T$ is the $n \times p$ matrix with rows of the vector $\mathbf{H}_i$. We then denote

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\mathbf{H}^T \mathbf{V}^{-1} \mathbf{H}\right)^{-1} \mathbf{H}^T \mathbf{V}^{-1} \mathbf{w} \\ &= \mathbf{A}^{-1} \mathbf{H}^T \mathbf{V}^{-1} \mathbf{w} \end{aligned}. \tag{A.17}$$

where $\mathbf{V} = diag\left(v_i\right)$ is the variance matrix specified under the model for the weights. Under the model, $\hat{\boldsymbol{\beta}}$ is the generalized least squares estimator of $\boldsymbol{\beta}$ and is thus unbiased.

The estimator $\hat{\boldsymbol{\beta}}$ has variance

$$Var_M\left(\hat{\boldsymbol{\beta}}\right)=\sigma^2\mathbf{A}^{-1}, \tag{A.18}$$

where $\mathbf{A}=\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H}$.

<u>Property 7: Variance of $\hat{T}_{HT}$ and $\hat{T}_B$ under the Model</u>

The corresponding versions of (A.11) and (A.13) are:

$$\begin{aligned}Var_F\left(\hat{T}_{HT}\middle|\mathbf{Y}\right)&=E_\pi\left[Var_M\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right]+Var_\pi\left[\tilde{T}_B\middle|\mathbf{Y}\right]\\&=E_\pi\left[\sum\nolimits_{i\in s}Var_M\left(w_iy_i\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right]+Var_\pi\left[\tilde{T}_B\middle|\mathbf{Y}\right].\\&=E_\pi\left[\sigma^2\sum\nolimits_{i\in s}v_iy_i^2\middle|\mathbf{Y}\right]+Var_\pi\left[\tilde{T}_B\middle|\mathbf{Y}\right]\end{aligned} \tag{A.19}$$

and

$$\begin{aligned}Var_F\left(\hat{T}_B\middle|\mathbf{Y}\right)&=E_\pi\left[Var_M\left(\hat{T}_B\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right]+Var_\pi\left[\tilde{T}_B\middle|\mathbf{Y}\right]\\&=E_\pi\left[Var_M\left(\sum\nolimits_{i\in s}\hat{w}_iy_i\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right]+Var_\pi\left[\tilde{T}_B\middle|\mathbf{Y}\right]\\&=E_\pi\left[Var_M\left(\sum\nolimits_{i\in s}y_i\mathbf{H}_i\,\hat{\boldsymbol{\beta}}\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right]+Var_\pi\left[\tilde{T}_B\middle|\mathbf{Y}\right]\\&=E_\pi\left[\sum\nolimits_{i\in s}y_i\mathbf{H}_i\sigma^2\left[\sum\nolimits_{i\in s}\frac{\mathbf{H}_i\mathbf{H}_i^T}{v_i}\right]^{-1}\sum\nolimits_{i\in s}\mathbf{H}_i^Ty_i\middle|\mathbf{Y}\right]+Var_\pi\left[\tilde{T}_B\middle|\mathbf{Y}\right]\\&=\sigma^2\sum\nolimits_{i\in s}y_i\mathbf{H}_i\left[\sum\nolimits_{i\in s}\frac{\mathbf{H}_i\mathbf{H}_i^T}{v_i}\right]^{-1}\sum\nolimits_{i\in s}\mathbf{H}_i^Ty_i+Var_\pi\left[\tilde{T}_B\middle|\mathbf{Y}\right]\end{aligned}. \tag{A.20}$$

Expressions (A.19) and (A.20) can more easily be derived using matrix notation, which Beaumont (2008) does not use. Since this notation is used for the proposed variance estimator, I prove the preceding results using the more flexible matrix notation. First, Beaumont drops the $Var_{\pi}\left[\tilde{T}_B \middle| \mathbf{Y}\right]$ term and obtains

$$\begin{aligned} Var_B\left(\hat{T}_{HT} \middle| \mathbf{Y}\right) &= Var_M\left(\hat{T}_{HT} \middle| \mathbf{Z}, \mathbf{Y}\right) \\ &= Var_M\left(\mathbf{w}^T \mathbf{y}_s \middle| \mathbf{Z}, \mathbf{Y}\right) \\ &= \mathbf{y}_s^T Var_M\left(\mathbf{w} \middle| \mathbf{Z}, \mathbf{Y}\right) \mathbf{y}_s \ . \\ &= \mathbf{y}_s^T \left(\sigma^2 \mathbf{V}\right) \mathbf{y}_s \\ &= \sigma^2 \mathbf{y}_s^T \mathbf{V} \mathbf{y}_s \end{aligned} \tag{A.21}$$

Since $\sigma^2 \mathbf{y}_s^T \mathbf{V} \mathbf{y}_s = \sigma^2 \sum_{i \in s} v_i y_i^2$, we get expression (A.19). From (A.21), the equivalent proof using my notation is

$$\begin{aligned} Var_F\left(\hat{T}_{HT} \middle| \mathbf{Y}\right) &= E_{\pi}\left[Var_M\left(\hat{T}_{HT} \middle| \mathbf{Z}, \mathbf{Y}\right) \middle| \mathbf{Y}\right] + Var_{\pi}\left[\tilde{T}_B \middle| \mathbf{Y}\right] \\ &= E_{\pi}\left[\sigma^2 \mathbf{y}_s^T \mathbf{V} \mathbf{y}_s \middle| \mathbf{Y}\right] + Var_{\pi}\left[\tilde{T}_B \middle| \mathbf{Y}\right] \\ &= \sigma^2 \sum_{i \in U} v_i E_{\pi}\left(I_i\right) y_i^2 + Var_{\pi}\left[\tilde{T}_B \middle| \mathbf{Y}\right] \qquad , \\ &= \sigma^2 \sum_{i \in U} v_i \pi_i y_i^2 + Var_{\pi}\left[\tilde{T}_B \middle| \mathbf{Y}\right] \\ &= \sigma^2 \mathbf{y}_U^T \left(\mathbf{V} \bullet diag\left(\mathbf{\Pi}\right)\right) \mathbf{y}_U + Var_{\pi}\left[\tilde{T}_B \middle| \mathbf{Y}\right] \end{aligned} \tag{A.22}$$

where $\mathbf{\Pi} = \left(\pi_{ij}\right)$ is the $N \times N$ matrix of the selection probabilities (with diagonal elements the first-order probabilities) and $\bullet$ denotes a Hadamard product. Second, for (A.20), here the proof is

$$
\begin{aligned}
Var_M\left(\hat{T}_B \middle| \mathbf{Z},\mathbf{Y}\right) &= Var_M\left(\hat{\mathbf{w}}^T\mathbf{y}_s \middle| \mathbf{Z},\mathbf{Y}\right) \\
&= \mathbf{y}_s^T Var_M\left(\mathbf{H}^T\hat{\boldsymbol{\beta}} \middle| \mathbf{Z},\mathbf{Y}\right)\mathbf{y}_s \\
&= \mathbf{y}_s^T \mathbf{H} Var_M\left(\hat{\boldsymbol{\beta}} \middle| \mathbf{Z},\mathbf{Y}\right)\mathbf{H}^T\mathbf{y}_s \\
&= \mathbf{y}_s^T \mathbf{H}\left(\sigma^2\mathbf{A}^{-1}\right)\mathbf{H}^T\mathbf{y}_s \\
&= \sigma^2\mathbf{y}_s^T\mathbf{D}\mathbf{y}_s \\
&= \sigma^2\sum_{i\in s}\sum_{j\in s} D_{ij}y_iy_j
\end{aligned}, \qquad \text{(A.23)}
$$

where $\mathbf{D}=\mathbf{H}\mathbf{A}^{-1}\mathbf{H}^T$ has elements $D_{ij}, i=1,\ldots,n,\ j=1,\ldots,n$. From (A.22), we have

$$
\begin{aligned}
Var_F\left(\hat{T}_B \middle| \mathbf{Y}\right) &= E_\pi\left[Var_M\left(\hat{T}_B \middle| \mathbf{Z},\mathbf{Y}\right)\middle| \mathbf{Y}\right] + Var_\pi\left[\tilde{T}_B \middle| \mathbf{Y}\right] \\
&= E_\pi\left[\sigma^2\sum_{i\in s}\sum_{j\in s} D_{ij}y_iy_j\right] + Var_\pi\left[\tilde{T}_B \middle| \mathbf{Y}\right] \\
&= \sigma^2\sum_{i\in U}\sum_{j\in U} E_\pi\left(I_{ij}\right)D_{ij}y_iy_j + Var_\pi\left[\tilde{T}_B \middle| \mathbf{Y}\right] \\
&= \sigma^2\sum_{i\in U}\sum_{j\in U} \pi_{ij}D_{ij}y_iy_j + Var_\pi\left[\tilde{T}_B \middle| \mathbf{Y}\right] \\
&= \sigma^2\mathbf{y}_U^T\left(\mathbf{\Pi}\bullet\mathbf{D}\right)\mathbf{y}_U + Var_\pi\left[\tilde{T}_B \middle| \mathbf{Y}\right]
\end{aligned} \qquad \text{(A.24)}
$$

where $\mathbf{\Pi}$ is the $N \times N$ matrix of the selection probabilities, defined above, and $\bullet$ denotes a Hadamard product.

Property 8: Conditional on the Weights Model, Beaumont's Estimator is always as efficient or more efficient than the HT estimator under the model

Beaumont (2008) proved that, under a linear weights model, the variance of the HT estimator is an upper bound for the variance of his estimator. However, this proof is dependent upon the fact that the weights model is correctly specified. In order to demonstrate the bias of the variance estimator under a weights model whose variance component is misspecified, I first detail Beaumont's result under a correct model.

From (A.19) and (A.20), we have

$$
\begin{aligned}
Var_M\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)-Var_M\left(\hat{T}_B\middle|\mathbf{Z},\mathbf{Y}\right) &= \sigma^2\sum\nolimits_{i\in s} v_i y_i^2 + Var_\pi\left[\tilde{T}_B\middle|\mathbf{Y}\right] \\
&= \sigma^2\left[\sum\nolimits_{i\in s} v_i y_i^2 - \sum\nolimits_{i\in s} y_i \mathbf{H}_i^T\left[\sum\nolimits_{i\in s}\frac{\mathbf{H}_i\mathbf{H}_i^T}{v_i}\right]^{-1}\sum\nolimits_{i\in s}\mathbf{H}_i^T y_i\right] \\
&= \sigma^2\left[\sum\nolimits_{i\in s} v_i y_i\left(y_i-\frac{\mathbf{H}_i^T\hat{\Omega}}{v_i}\right)\right] \\
&= \sigma^2\left[\sum\nolimits_{i\in s} v_i\left(y_i-\frac{\mathbf{H}_i^T\hat{\Omega}}{v_i}\right)^2\right]
\end{aligned}
\tag{A.25}
$$

where $\hat{\Omega}=\left[\sum\nolimits_{i\in s}\frac{\mathbf{H}_i\mathbf{H}_i^T}{v_i}\right]^{-1}\sum\nolimits_{i\in s}\mathbf{H}_i^T y_i$. The last line above holds since

.

$$
\begin{aligned}
\sum_{i\in s} v_i y_i \left(y_i - \frac{\mathbf{H}_i^T \hat{\Omega}}{v_i} \right) &= \sum_{i\in s} v_i \left(y_i - \frac{\mathbf{H}_i^T \hat{\Omega}}{v_i} + \frac{\mathbf{H}_i^T \hat{\Omega}}{v_i} \right)\left(y_i - \frac{\mathbf{H}_i^T \hat{\Omega}}{v_i} \right) \\
&= \sum_{i\in s} v_i \left(y_i - \frac{\mathbf{H}_i^T \hat{\Omega}}{v_i} \right)^2 + \mathbf{H}_i^T \hat{\Omega} \left(y_i - \frac{\mathbf{H}_i^T \hat{\Omega}}{v_i} \right) \\
&= \sum_{i\in s} v_i \left(y_i - \frac{\mathbf{H}_i^T \hat{\Omega}}{v_i} \right)^2 + \sum_{i\in s} \hat{\Omega}^T \mathbf{H}_i^T \left(y_i - \frac{\mathbf{H}_i^T \hat{\Omega}}{v_i} \right) \\
&= \sum_{i\in s} v_i \left(y_i - \frac{\mathbf{H}_i^T \hat{\Omega}}{v_i} \right)^2 + \hat{\Omega}^T \sum_{i\in s} \mathbf{H}_i y_i - \hat{\Omega}^T \sum_{i\in s} \frac{\mathbf{H}_i \mathbf{H}_i^T}{v_i} \hat{\Omega} \\
&= \sum_{i\in s} v_i \left(y_i - \frac{\mathbf{H}_i^T \hat{\Omega}}{v_i} \right)^2 + \hat{\Omega}^T \sum_{i\in s} \mathbf{H}_i y_i \\
&\quad - \left(\sum_{i\in s} \mathbf{H}_i y_i \right)\left(\sum_{i\in s} \frac{\mathbf{H}_i \mathbf{H}_i^T}{v_i} \right)^{-1} \left(\sum_{i\in s} \frac{\mathbf{H}_i \mathbf{H}_i^T}{v_i} \right)\left(\sum_{i\in s} \frac{\mathbf{H}_i \mathbf{H}_i^T}{v_i} \right)^{-1} \sum_{i\in s} \mathbf{H}_i y_i \\
&= \sum_{i\in s} v_i \left(y_i - \frac{\mathbf{H}_i^T \hat{\Omega}}{v_i} \right)^2 + \hat{\Omega}^T \sum_{i\in s} \mathbf{H}_i y_i - \left(\sum_{i\in s} \mathbf{H}_i y_i \right)\left(\sum_{i\in s} \frac{\mathbf{H}_i \mathbf{H}_i^T}{v_i} \right)^{-1} \sum_{i\in s} \mathbf{H}_i y_i \\
&= \sum_{i\in s} v_i \left(y_i - \frac{\mathbf{H}_i^T \hat{\Omega}}{v_i} \right)^2 + \hat{\Omega}^T \sum_{i\in s} \mathbf{H}_i y_i - \left(\sum_{i\in s} \mathbf{H}_i y_i \right)\hat{\Omega} \\
&= \sum_{i\in s} v_i \left(y_i - \frac{\mathbf{H}_i^T \hat{\Omega}}{v_i} \right)^2
\end{aligned}
\qquad \text{(A.26)}
$$

It is also relatively simple to prove that, under the linear model, the HT theoretical variance is an upper bound for the Beaumont estimator variance (with respect to the model). While all the terms in (A.25) are positive, unlike Beaumont, I demonstrate here that again the Property 7 proof is simpler to do via matrix notation and utilize a known theorem for positive definite matrices to prove that the conditional variance of the HT estimator under the weights model is an upper bound for the conditional variance of the Beaumont estimator. I then extend this theory for the "total variances," i.e., the variances with respect to both the sample design and the weights model under Property 8.

Since we can write

$$Var_M\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)-Var_M\left(\hat{T}_B\middle|\mathbf{Z},\mathbf{Y}\right)=\sigma^2\mathbf{y}_s^T\mathbf{V}\mathbf{y}_s-\sigma^2\mathbf{y}_s^T\mathbf{D}\mathbf{y}_s \\ =\sigma^2\mathbf{y}_s^T\left(\mathbf{V}-\mathbf{D}\right)\mathbf{y}_s, \quad \text{(A.27)}$$

we thus have $Var_M\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)\geq Var_M\left(\hat{T}_B\middle|\mathbf{Z},\mathbf{Y}\right)$ if $\mathbf{y}_s^T\left(\mathbf{V}-\mathbf{D}\right)\mathbf{y}_s\geq\mathbf{0}$. Since $\mathbf{y}_s^T\left(\mathbf{V}-\mathbf{D}\right)\mathbf{y}_s\geq\mathbf{0}$ is a quadratic form, if $\mathbf{V}-\mathbf{D}$ is invertible, then $\mathbf{y}_s^T\left(\mathbf{V}-\mathbf{D}\right)\mathbf{y}_s$ is a positive definite quadratic form and thus $\mathbf{y}_s^T\left(\mathbf{V}-\mathbf{D}\right)\mathbf{y}_s\geq\mathbf{0}$. To prove this, we can use the formula for the inverse of the sums of matrices (Theorem 9.5.16 on p. 315 of Valliant *et al*. 2000) to rewrite $\left(\mathbf{V}-\mathbf{D}\right)^{-1}$:

Theorem 1. For matrices $\mathbf{B},\mathbf{C},\mathbf{D},\mathbf{E}$, $\left(\mathbf{B}+\mathbf{DCE}\right)^{-1}=\mathbf{B}^{-1}-\mathbf{B}^{-1}\mathbf{D}\left(\mathbf{C}^{-1}+\mathbf{EB}^{-1}\mathbf{D}\right)^{-1}\mathbf{EB}^{-1}$.

For $\mathbf{A}=\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H},\mathbf{B}=\mathbf{V},\mathbf{C}=\mathbf{A}^{-1},\mathbf{D}=-\mathbf{H},\mathbf{E}=\mathbf{H}^T$, we have

$$\begin{aligned}\left(\mathbf{V}-\mathbf{D}\right)^{-1}&=\left(\mathbf{V}-\mathbf{HA}^{-1}\mathbf{H}^T\right)^{-1}\\&=\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{H}\left(\left(\mathbf{A}^{-1}\right)^{-1}+\mathbf{H}^T\mathbf{V}^{-1}\left(-\mathbf{H}\right)\right)^{-1}\mathbf{H}^T\mathbf{V}^{-1}\\&=\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{H}\left(\mathbf{A}-\mathbf{H}^T\mathbf{V}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{V}^{-1}\\&=\mathbf{V}^{-1}-\mathbf{V}^{-1}\mathbf{H}\left(\mathbf{A}-\mathbf{A}\right)^{-1}\mathbf{H}^T\mathbf{V}^{-1}\\&=\mathbf{V}^{-1}\end{aligned} \quad \text{(A.28)}$$

Since $\mathbf{V}^{-1}$ is obviously invertible, then $\left(\mathbf{V}-\mathbf{D}\right)^{-1}$ exists, $\mathbf{y}_s^T\left(\mathbf{V}-\mathbf{D}\right)\mathbf{y}_s\geq\mathbf{0}$, and $Var_M\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)\geq Var_M\left(\hat{T}_B\middle|\mathbf{Z},\mathbf{Y}\right)$.

Property 9: Beaumont's Estimator is always as efficient or more efficient than the HT estimator under the model and sample design

Again, for (A.19) and (A.20), Beaumont seems to be ignoring the E_π expectation and considers the difference in the theoretical variances with respect only to the model under the one realized sample. However, unlike Beaumont's approach of comparing the variances conditional on the weights model, it is more comprehensive to consider a similar comparison between expressions (A.22)and(A.24):

$$\begin{aligned} Var_F\left(\hat{T}_{HT}\middle|\mathbf{Y}\right) - Var_F\left(\hat{T}_B\middle|\mathbf{Y}\right) &= E_\pi\left[Var_M\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right] - E_\pi\left[Var_M\left(\hat{T}_B\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right] \\ &= \sigma^2\mathbf{y}_U^T\left(diag\left(\mathbf{\Pi}\right)\bullet\mathbf{V}\right)\mathbf{y}_U - \sigma^2\mathbf{y}_U^T\left(\mathbf{\Pi}\bullet\mathbf{D}\right)\mathbf{y}_U \\ &= \sigma^2\mathbf{y}_U^T\left(\mathbf{\Pi}\bullet\mathbf{V} - \mathbf{\Pi}\bullet\mathbf{D}\right)\mathbf{y}_U \\ &= \sigma^2\mathbf{y}_U^T\mathbf{\Pi}\bullet\left(\mathbf{V}-\mathbf{D}\right)\mathbf{y}_U \end{aligned} \tag{A.29}$$

A similar comparison to that used in (A.28) can verify here that $Var_F\left(\hat{T}_{HT}\middle|\mathbf{Y}\right) \geq Var_F\left(\hat{T}_B\middle|\mathbf{Y}\right)$. Intuitively this should hold since by (A.28), the HT estimator has a conditional variance that is lower than that of the Beaumont estimator, which we should expect to hold when averaged across all possible samples. By the properties of Hadamard products, $\left[\mathbf{\Pi}\bullet\left(\mathbf{V}-\mathbf{D}\right)\right]^{-1} = \tilde{\mathbf{\Pi}}\left(\mathbf{V}-\mathbf{D}\right)^{-1}$, where $\tilde{\mathbf{\Pi}}$ is an $n \times n$ matrix with elements $1/\pi_{ij}$. Thus, $\mathbf{\Pi}\bullet\left(\mathbf{V}-\mathbf{D}\right)$ is positive definite and expression (A.29) is positive.

Appendix 3: Approximate Theoretical Variance of Beaumont Estimator, When the Weights Model Holds

For Beaumont's proposed variance estimation, we start with the following assumption:

$$\textit{Assumption 2. } E_\pi\left(\hat{T}_B \middle| \mathbf{I},\mathbf{Y}\right)=\tilde{T}_B+o_p\left(N/\sqrt{n}\right). \tag{A.30}$$

Under Assumption 2, $E_M\left(\hat{T}_B \middle| \mathbf{I},\mathbf{Y}\right)\approx\tilde{T}_B$, since $\hat{T}_B-T=O_p\left(N/\sqrt{n}\right)$. Note that Assumption 2 only holds if $\mathbf{y}_s$ is bounded and $E_M\left(\hat{w}_i \middle| \mathbf{I},\mathbf{Y}\right)=\tilde{w}_i+O_p\left(N/n^{3/2}\right)$ (not $\frac{N}{n}o_p\left(n^{-1/2}\right)$ as in Beaumont 2008). Equality holds under the linear model, i.e., $E_M\left(\hat{w}_i \middle| \mathbf{I},\mathbf{Y}\right)=\tilde{w}_i$. Also, since $Var_\pi\left[E_M\left(\hat{T}_B \middle| \mathbf{I},\mathbf{Y}\right)\middle|\mathbf{Y}\right]\approx Var_\pi\left[\tilde{T}_B \middle| \mathbf{Y}\right]$, we approximate $Var_F\left(\hat{T}_B \middle| \mathbf{Y}\right)$ with

$$\begin{aligned} Var_F\left(\hat{T}_B \middle| \mathbf{Y}\right) &= E_\pi\left[Var_M\left(\hat{T}_B \middle| \mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right]+Var_\pi\left[E_M\left(\hat{T}_B \middle| \mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right] \\ &\approx E_\pi\left[Var_M\left(\hat{T}_B \middle| \mathbf{I},\mathbf{Y}\right)\middle|\mathbf{Y}\right]+Var_\pi\left[\tilde{T}_B \middle| \mathbf{Y}\right] \end{aligned}. \tag{A.31}$$

We obtain a formula for $Var_F\left[\tilde{T}_B \middle| \mathbf{Y}\right]$ as follows. From (A.11) we have

$$Var_F\left(\hat{T}_{HT} \middle| \mathbf{Y}\right)=Var_F\left[\tilde{T}_B \middle| \mathbf{Y}\right]+E_\pi\left[Var_M\left(\hat{T}_{HT} \middle| \mathbf{I},\mathbf{Y}\right)\middle|\mathbf{Y}\right]. \tag{A.32}$$

Using the conditional variance formula, is it also true that

$$Var_F\left(\hat{T}_{HT}\middle|\mathbf{Y}\right)=E_\pi\left[Var_M\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right]+Var_\pi\left[E_M\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right]$$
$$=E_\pi\left[Var_M\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right], \qquad \text{(A.32)}$$

where the last line follows from the fact that $Var_\pi\left[E_M\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right]=0$. From (A.32) and (A.32), we have

$$Var_F\left[\tilde{T}_B\middle|\mathbf{Y}\right]=E_\pi\left[Var_M\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right]-E_\pi\left[Var_M\left(\hat{T}_{HT}\middle|\mathbf{I},\mathbf{Y}\right)\middle|\mathbf{Y}\right]. \qquad \text{(A.33)}$$

Consequently, we can approximate $Var_F\left(\hat{T}_B\middle|\mathbf{Y}\right)$ in (A.31) with

$$Var_F\left(\hat{T}_B\middle|\mathbf{Y}\right)\approx E_\pi\left[Var_M\left(\hat{T}_B\middle|\mathbf{I},\mathbf{Y}\right)\middle|\mathbf{Y}\right]+E_\pi\left[Var_M\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right]-E_\pi\left[Var_M\left(\hat{T}_{HT}\middle|\mathbf{I},\mathbf{Y}\right)\middle|\mathbf{Y}\right]$$
$$=E_\pi\left[Var_M\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right]+E_\pi\left[\left\{Var_M\left(\hat{T}_B\middle|\mathbf{I},\mathbf{Y}\right)-Var_M\left(\hat{T}_{HT}\middle|\mathbf{I},\mathbf{Y}\right)\right\}\middle|\mathbf{Y}\right]$$
(A.34)

Appendix 4: Beaumont Proposed Variance Estimators of Beaumont Estimator When the Weights Model Holds

To estimate the variance in (A.34), Beaumont proposes

$$var_B\left(\hat{T}_B\middle|\mathbf{Y}\right)=var_\pi\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)+\left\{var_M\left(\hat{T}_B\middle|\mathbf{I},\mathbf{Y}\right)-var_M\left(\hat{T}_{HT}\middle|\mathbf{I},\mathbf{Y}\right)\right\}, \qquad \text{(A.35)}$$

where $var_\pi\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)$ is a design-consistent variance estimator for $Var_\pi\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)$, but $var_M\left(\hat{T}_B\middle|\mathbf{I},\mathbf{Y}\right)$ and $var_M\left(\hat{T}_{HT}\middle|\mathbf{I},\mathbf{Y}\right)$ are consistent variance estimators with respect to the model M for the weights. In the last component of estimator (A.35), again the expectation with respect to the design is ignored; the estimators are conditional only on the model. For example, from the theoretical variance (A.25) under the linear model, (A.35) becomes

$$var_M\left(\hat{T}_B \middle| \mathbf{Z},\mathbf{Y}\right) - var_M\left(\hat{T}_{HT} \middle| \mathbf{Z},\mathbf{Y}\right) = -\hat{\sigma}^2\left[\sum_{i\in s} v_i y_i \left(y_i - \frac{\mathbf{H}_i^T\hat{\Omega}}{v_i}\right)\right], \qquad \text{(A.36)}$$

where $\hat{\Omega} = \left[\sum_{i\in s}\frac{\mathbf{H}_i\mathbf{H}_i^T}{v_i}\right]^{-1}\sum_{i\in s}\mathbf{H}_i^T y_i$ and $\hat{\sigma}^2$ is a model-consistent estimator of σ^2.

In particular, the first (A.35) component, $var_\pi\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)$, is not an appropriate estimator for $E_\pi\left[Var_M\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)\middle|\mathbf{Y}\right]$. A more appropriate estimator, which is described in Sec. 1.3.1, corresponds to the first component of (A.22). Beaumont did not prove that his general proposed variance estimator is always positive, but the first (A.35) component is $O\left(N^2/n\right)$ while the second component is $O(n)$, so the second component is much smaller in magnitude. Thus, for any weights model, the Beaumont variance estimator should be positive in large samples.

Appendix 5: Theoretical MSE of the Beaumont Estimator and Beaumont MSE Estimators When the Weights Model Does Not Hold

When the weights model does not hold, since the Beaumont estimator of the total is biased across repeated samples (see (A.7)-(A.10)) the total mean square error should be considered rather than a variance estimator:

$$\begin{aligned} MSE\left(\hat{T}_B \middle| \mathbf{Y}\right) &= E_F\left[\left(\hat{T}_B - T\right)^2 \middle| \mathbf{Y}\right] \\ &= E_\pi\left[Var_M\left(\hat{T}_B \middle| \mathbf{Z}, \mathbf{Y}\right) \middle| \mathbf{Y}\right] + B_M^2 \end{aligned} . \tag{A.37}$$

where $B_M\left[\left(\hat{T}_B - T\right) \middle| \mathbf{Z}, \mathbf{Y}\right] = E_M\left[E_\pi\left(\left(\hat{T}_B - T\right) \middle| \mathbf{Z}, \mathbf{Y}\right) \middle| \mathbf{Y}\right]$ is the bias of the estimator $\hat{T}_B$.

Beaumont MSE Estimators

Again, Beaumont proposes using a standard design-based method to estimate the variance $Var_M\left(\hat{T}_B \middle| \mathbf{Z}, \mathbf{Y}\right)$. While the design-based $\hat{T}_B - \hat{T}_{HT}$ is an unbiased estimator of the bias B_M, $\left(\hat{T}_B - \hat{T}_{HT}\right)^2$ is not an unbiased estimator of the squared bias. Thus, Beaumont proposes:

$$\hat{B}_M^2\left(\hat{T}_B \middle| \mathbf{Z}, \mathbf{Y}\right) = \max\left[0, \left(\hat{T}_B - \hat{T}_{HT}\right)^2 - var_\pi\left[\left(\hat{T}_B - \hat{T}_{HT}\right) \middle| \mathbf{Z}, \mathbf{Y}\right]\right], \tag{A.38}$$

where $var_\pi\left[\left(\hat{T}_B - \hat{T}_{HT}\right) \middle| \mathbf{Z}, \mathbf{Y}\right]$ is a design-consistent estimator of $Var_\pi\left[\left(\hat{T}_B - \hat{T}_{HT}\right) \middle| \mathbf{Z}, \mathbf{Y}\right]$. The resulting MSE estimator is given by

$$\begin{aligned} mse\left(\hat{T}_B \middle| \mathbf{Z}, \mathbf{Y}\right) &= var_M\left(\hat{T}_B \middle| \mathbf{Z}, \mathbf{Y}\right) + \hat{B}_M^2 \\ &= var_M\left(\hat{T}_B \middle| \mathbf{Z}, \mathbf{Y}\right) + \max\left[0, \left(\hat{T}_B - \hat{T}_{HT}\right)^2 - var_\pi\left[\left(\hat{T}_B - \hat{T}_{HT}\right) \middle| \mathbf{Z}, \mathbf{Y}\right]\right] \end{aligned} . \tag{A.39}$$

To ensure that $mse\left(\hat{T}_B \middle| \mathbf{Z}, \mathbf{Y}\right) \geq var_M\left(\hat{T}_{HT} \middle| \mathbf{Z}, \mathbf{Y}\right)$ in (A.39) (since theoretically in (A.25) it was shown that $Var_M\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right) \geq Var_M\left(\hat{T}_B\middle|\mathbf{Z},\mathbf{Y}\right)$), Beaumont proposes the design-based MSE estimator

$$mse_D\left(\hat{T}_B\right) = \min\left[mse\left(\hat{T}_B\middle|\mathbf{Z},\mathbf{Y}\right), var_M\left(\hat{T}_{HT}\middle|\mathbf{Z},\mathbf{Y}\right)\right]. \tag{A.40}$$

Appendix 6: Positive Bias of the Variance Estimator and Under-Estimation of the MSE When the Weights Model Does Not Hold

When the weights model does not hold, we need to consider the bias in the Beaumont model-based variance estimator. Beaumont does not incorporate this theory. Suppose that the working weights model M is used, when the true weights model is actually $\tilde{M}$, with $Var_{\tilde{M}}\left(w_i\right) = \psi_i$ denoting the model variance component. If the model M is wrong, then we have

$$\begin{aligned} E_{\tilde{M}}\left[var_B\left(\hat{T}_B\middle|\mathbf{Y}\right)\right] &\cong \sum_{i\in s} v_i y_i^2 E_{\tilde{M}}\left[\left(w_i - \hat{w}_i\right)^2\right] \\ &\cong \sum_{i\in s} v_i y_i^2 \psi_i + \sum_{i\in s} v_i y_i^2 \left[E_{\tilde{M}}\left(w_i - \hat{w}_i\right)\right]^2 \ . \\ &= Var_{\tilde{M}}\left(\hat{T}_B\middle|\mathbf{Y}\right) + \sum_{i\in s} v_i y_i^2 \left[E_{\tilde{M}}\left(w_i - \hat{w}_i\right)\right]^2 \end{aligned} \tag{A.41}$$

Both of the components in (A.41), the model-variance and the positive bias term, have the same order of magnitude, $O_p\left(N^2/n\right)$. This is the same order of magnitude as the variance component in the MSE (A.37). However, the bias component in the MSE (the second component in (A.37)) has order $O_p\left(N^2\right)$, which is higher than $O_p\left(N^2/n\right)$. This means that when the weights model is wrong the variance estimator is positively biased but still underestimates the true MSE.

Appendix 7: Special Case of Beaumont's Estimator Being Equivalent to the HT Estimator Under PPS Sampling

The Beaumont estimator has the property that, when the survey variable is linearly related to the auxiliary variable used to draw the pps sample and an inverse weights model $w_i = \mathbf{H}_i^T \boldsymbol{\beta} + e_i, e_i \sim \sigma^2 \mathbf{H}_i^2$ is used, it is equivalent to the HT estimator, i.e., $\hat{T}_B \equiv \hat{T}_{HT}$. Suppose for simplicity that we have one variable, i.e., $\mathbf{H}_i = y_i^{-1}$. That is,

$$
\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\mathbf{H}^T \mathbf{V}^{-1} \mathbf{H}\right)^{-1} \mathbf{H}^T \mathbf{V}^{-1} \mathbf{w} \\
&= \frac{1}{\sum_{i \in s} \mathbf{H}_i \frac{1}{\mathbf{H}_i^2} \mathbf{H}_i} \sum_{i \in s} \mathbf{H}_i \frac{1}{\mathbf{H}_i^2} w_i \\
&= \frac{1}{n} \sum_{i \in s} \frac{w_i}{\mathbf{H}_i} \\
&= \frac{1}{n} \sum_{i \in s} w_i y_i
\end{aligned}
\tag{A.42}
$$

From (A.42), the Beaumont estimator is

$$
\begin{aligned}
\hat{T}_B &= \sum_{i \in s} \hat{w}_i y_i \\
&= \sum_{i \in s} \mathbf{H}_i^T \hat{\boldsymbol{\beta}} y_i \\
&= \sum_{i \in s} \frac{1}{y_i} \hat{\boldsymbol{\beta}} y_i \\
&= n \hat{\boldsymbol{\beta}} \\
&= n \frac{1}{n} \sum_{i \in s} w_i y_i \\
&= \hat{T}_{HT}
\end{aligned}
\tag{A.43}
$$

In other words, the predicted weights under these circumstances are exactly equal to the HT weights, or that the inverse weights model in this situation produces the HT estimator exactly.

Appendix 8: Derivation of Spencer's Design Effect

Let y_i denote the measurement of interest, p_i the one-draw probability of selection for a sample of size n, and $w_i = (np_i)^{-1}$ is the base weight for unit i in a population of size N. Observe that the average population probability is $\bar{P} = \frac{1}{N}\sum_{i=1}^{N} p_i = \frac{1}{N}$. Consider an underlying population model defined as $y_i = A + Bp_i + \varepsilon_i$. If the entire finite population were available, the least-squares population regression line would be

$$y_i = \alpha + \beta p_i + e_i, \tag{A.44}$$

where $\alpha = \bar{Y} - \beta N^{-1}$, $\beta = \sum_{i=1}^{N}(y_i - \bar{Y})(p_i - \bar{P}) \Big/ \sum_{i=1}^{N}(p_i - \bar{P})^2$, and $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ the population mean. Denote the population variances of the y's, e's e^2, and weights as $\sigma_y^2, \sigma_e^2, \sigma_{e^2}^2, \sigma_w^2$, e.g., $\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{Y})^2$, and the finite population correlations between y and P by ρ_{yp}, e and w by ρ_{ew}, and e^2 and w by ρ_{e^2w}. For example, $\rho_{yp} = \sum_{i=1}^{N}(y_i - \bar{Y})(p_i - \bar{P}) \Big/ \sqrt{\sum_{i=1}^{N}(y_i - \bar{Y})^2 \sum_{i=1}^{N}(p_i - \bar{P})^2}$. From LS regression, $\sum_{i=1}^{N} e_i p_i = \frac{1}{N}\sum_{i=1}^{N} e_i = 0$ and $\sigma_e^2 = \left(1 - \rho_{yp}^2\right)\sigma_y^2$.

Let $\hat{T} = \sum_{i=1}^{n} w_i y_i$ denote the sample-based estimate of the population total. Its variance in within-replacement sampling is given by

$$\begin{aligned} Var\left(\hat{T}\right) &= \frac{1}{n}\sum_{i=1}^{N} p_i \left(\frac{y_i}{p_i} - T\right)^2 \\ &= \frac{1}{n}\left(\sum_{i=1}^{N} \frac{y_i^2}{p_i} - T^2\right) \end{aligned}. \tag{A.45}$$

Using the model in(A.44), we can rewrite the variance in (A.45). To do this involves several steps. First, we rewrite the population total as

$$
\begin{aligned}
T &= \sum_{i=1}^{N} y_i \\
&= \sum_{i=1}^{N} \left(\alpha + \beta p_i + e_i \right), \\
&= N\alpha + \beta \sum_{i=1}^{N} p_i \\
&= N\alpha + \beta
\end{aligned}
\tag{A.46}
$$

such that

$$
\begin{aligned}
T^2 &= \left(N\alpha + \beta \right)^2 \\
&= \left(N\alpha \right)^2 + \beta^2 + 2N\alpha\beta
\end{aligned}
.
\tag{A.47}
$$

Second, we rewrite the component $\sum_{i=1}^{N} \frac{y_i^2}{p_i}$ in (A.45) as

$$
\begin{aligned}
\sum_{i=1}^{N} \frac{y_i^2}{p_i} &= \sum_{i=1}^{N} \frac{1}{p_i} \left(\alpha + \beta p_i + e_i \right)^2 \\
&= \sum_{i=1}^{N} \frac{1}{p_i} \left(\alpha^2 + \beta^2 p_i^2 + e_i^2 + 2\alpha\beta p_i + 2\alpha e_i + 2\beta p_i e_i \right) \\
&= \alpha^2 \sum_{i=1}^{N} \frac{1}{p_i} + \beta^2 \sum_{i=1}^{N} p_i + \sum_{i=1}^{N} \frac{e_i^2}{p_i} + 2N\alpha\beta + 2\alpha \sum_{i=1}^{N} \frac{e_i}{p_i} + 2\beta \sum_{i=1}^{N} e_i \\
&= \alpha^2 \sum_{i=1}^{N} \frac{1}{p_i} + \beta^2 + \sum_{i=1}^{N} \frac{e_i^2}{p_i} + 2N\alpha\beta + 2\alpha \sum_{i=1}^{N} \frac{e_i}{p_i}
\end{aligned}
\tag{A.48}
$$

Plugging in $w_i = \left(np_i \right)^{-1}$, or $p_i = \left(nw_i \right)^{-1}$ lets us rewrite (A.48) as

$$
\begin{aligned}
\sum_{i=1}^{N} \frac{y_i^2}{p_i} &= \alpha^2 \sum_{i=1}^{N} \frac{1}{p_i} + \beta^2 + \sum_{i=1}^{N} \frac{e_i^2}{p_i} + 2N\alpha\beta + 2\alpha \sum_{i=1}^{N} \frac{e_i}{p_i} \\
&= \alpha^2 \sum_{i=1}^{N} \frac{1}{\left(nw_i \right)^{-1}} + \beta^2 + \sum_{i=1}^{N} \frac{e_i^2}{\left(nw_i \right)^{-1}} + 2N\alpha\beta + 2\alpha \sum_{i=1}^{N} \frac{e_i}{\left(nw_i \right)^{-1}} \\
&= n\alpha^2 \sum_{i=1}^{N} w_i + \beta^2 + n \sum_{i=1}^{N} w_i e_i^2 + 2N\alpha\beta + 2n\alpha \sum_{i=1}^{N} w_i e_i
\end{aligned}
\tag{A.49}
$$

Subtracting (A.47) from (A.49) gives

$$\sum_{i=1}^{N}\frac{y_i^2}{p_i}-T^2 = n\alpha^2\sum_{i=1}^{N}w_i+\beta^2+n\sum_{i=1}^{N}w_ie_i^2+2N\alpha\beta+2n\alpha\sum_{i=1}^{N}w_ie_i$$
$$-\left[\left(N\alpha\right)^2+\beta^2+2N\alpha\beta\right] \quad \text{(A.50)}$$
$$= n\alpha^2\sum_{i=1}^{N}w_i+n\sum_{i=1}^{N}w_ie_i^2+2n\alpha\sum_{i=1}^{N}w_ie_i-\left(N\alpha\right)^2$$

Dividing (A.50) by n gives

$$\frac{1}{n}\left(\sum_{i=1}^{N}\frac{y_i^2}{p_i}-T^2\right)=\alpha^2\left(\sum_{i=1}^{N}w_i-\frac{N^2}{n}\right)+\sum_{i=1}^{N}w_ie_i^2+2\alpha\sum_{i=1}^{N}w_ie_i \quad \text{(A.51)}$$

From the definition of covariance between e^2 and w, we have

$$NCov\left(w,e^2\right)=\sum_{i=1}^{N}\left(w_i-\bar{W}\right)\left(e_i^2-\overline{e^2}\right)$$
$$=\sum_{i=1}^{N}w_ie_i^2-N\bar{W}\overline{e^2} \quad , \quad \text{(A.52)}$$
$$=N\rho_{we^2}\sigma_w\sigma_{e^2}$$

where $\overline{e^2}=\frac{1}{N}\sum_{i=1}^{N}e_i^2$ which implies $\sum_{i=1}^{N}w_ie_i^2=N\rho_{e^2w}\sigma_{e^2}\sigma_w+N\bar{W}\overline{e^2}$. Similarly,

since $\sigma_e^2=\overline{e^2}$, we have

$$NCov\left(w,e\right)=\sum_{i=1}^{N}\left(w_i-\bar{W}\right)\left(e_i-\bar{e}\right)$$
$$=\sum_{i=1}^{N}w_ie_i-N\bar{W}\bar{e} \quad , \quad \text{(A.53)}$$
$$=\sum_{i=1}^{N}w_ie_i$$
$$=N\rho_{we}\sigma_w\sigma_e$$

which implies $\sum_{i=1}^{N}w_ie_i=N\rho_{ew}\sigma_e\sigma_w$. These two results means that the third and fourth terms in (A.51) can be rewritten as

$$\sum_{i=1}^{N} w_i e_i^2 = N\rho_{e^2 w}\sigma_{e^2}\sigma_w + N\bar{W}\left(\sigma_e^2 + \overline{e^2}\right)$$
$$= N\rho_{e^2 w}\sigma_{e^2}\sigma_w + N\bar{W}\sigma_e^2 \qquad \text{(A.54)}$$
$$= N\rho_{e^2 w}\sigma_{e^2}\sigma_w + N\bar{W}\left(1-\rho_{yp}^2\right)\sigma_y^2$$

and

$$\sum_{i=1}^{N} w_i e_i = N\rho_{ew}\sigma_e\sigma_w . \qquad \text{(A.55)}$$

Plugging these back into the variance (A.45) gives

$$Var\left(\hat{T}\right) = \alpha^2\left(\sum_{i=1}^{N} w_i - \frac{N^2}{n}\right) + N\rho_{e^2 w}\sigma_{e^2}\sigma_w + N\bar{W}\left(1-\rho_{yp}^2\right)\sigma_y^2 + 2\alpha N\rho_{ew}\sigma_e\sigma_w$$
$$= \alpha^2 N\left(\bar{W} - \frac{N^2}{n}\right) + N\rho_{e^2 w}\sigma_{e^2}\sigma_w + N\bar{W}\left(1-\rho_{yp}^2\right)\sigma_y^2 + 2\alpha N\rho_{\varepsilon w}\sigma_e\sigma_w .$$

(A.56)

The variance (A.45) under simple random sampling with replacement, where $w_i = \frac{N}{n}$, reduces down to

$$Var_{srs}\left(\hat{T}\right) = \frac{1}{n}\sum_{i=1}^{N}\frac{n}{N}\left(\frac{Ny_i}{n} - T\right)^2$$
$$= \frac{N}{n}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2 \quad . \qquad \text{(A.57)}$$
$$= \frac{N^2}{n}\sigma_y^2$$

Taking the ratio of (A.56) to (A.57) gives the design effect

$$deff_S = \frac{Var\left(\hat{T}\right)}{Var_{srs}\left(\hat{T}\right)}$$

$$= \frac{\alpha^2 N\left(\bar{W} - \frac{N}{n}\right) + N\rho_{e^2w}\sigma_{e^2}\sigma_w + N\bar{W}\left(1-\rho_{yp}^2\right)\sigma_y^2 + 2\alpha N\rho_{ew}\sigma_e\sigma_w}{\frac{N^2}{n}\sigma_y^2}.$$

$$= \frac{\alpha^2 n}{N\sigma_y^2}\left(\bar{W} - \frac{N}{n}\right) + \frac{n\rho_{e^2w}\sigma_{e^2}\sigma_w}{N\sigma_y^2} + \frac{n\bar{W}}{N}\left(1-\rho_{yp}^2\right) + \frac{2\alpha n\rho_{ew}\sigma_e\sigma_w}{N\sigma_y^2}$$

$$= \frac{\alpha^2}{\sigma_y^2}\left(\frac{n\bar{W}}{N} - 1\right) + \frac{n\bar{W}}{N}\left(1-\rho_{yp}^2\right) + \frac{n\rho_{e^2w}\sigma_{e^2}\sigma_w}{N\sigma_y^2} + \frac{2\alpha n\rho_{ew}\sigma_e\sigma_w}{N\sigma_y^2}$$

(A.58)

Spencer argues that if the correlations in the last two components of (A.58) are negligible, then (A.58) can be approximated by

$$deff_S \approx \frac{\alpha^2}{\sigma_y^2}\left(\frac{n\bar{W}}{N} - 1\right) + \frac{n\bar{W}}{N}\left(1-\rho_{yp}^2\right). \quad \text{(A.59)}$$

To estimate (A.59), we first start by noting from (3.1) that

$$1 + \left[CV\left(\mathbf{w}\right)\right]^2 = \frac{n^{-1}\sum_{i\in s} w_i^2}{\bar{w}^2}, \quad \text{(A.60)}$$

where $\bar{w} = \frac{1}{n}\sum_{i=1}^{n} w_i$ is the average weight. The design-expectation of the numerator in (A.60) is

$$\begin{aligned} E_\pi\left[n^{-1}\sum_{i\in s} w_i^2\right] &= n^{-1}\sum_{i\in U} E_\pi\left(\delta_i\right) w_i^2 \\ &= n^{-1}\sum_{i\in U}\left(np_i\right) w_i^2 \\ &= n^{-1}\sum_{i\in U} w_i \\ &= \frac{N\bar{W}}{n} \end{aligned}. \quad \text{(A.61)}$$

For the denominator, note that the design-expectation of $\bar{w} = \frac{1}{n}\sum_{i \in s} w_i$ is

$$E_\pi\left[\frac{1}{n}\sum_{i \in s} w_i\right] = \frac{1}{n}\sum_{i \in U} E_\pi(\delta_i) w_i$$
$$= \frac{1}{n}\sum_{i \in U}(np_i) w_i \quad . \tag{A.62}$$
$$= \frac{N}{n}$$

Thus, Spencer proposes to use $\bar{w}^2$ to estimate $\left(\frac{N}{n}\right)^2$. To see why this is reasonable, note that

$$E_\pi\left(\bar{w}^2\right) = Var_\pi(\bar{w}) + \left[E_\pi(\bar{w})\right]^2. \tag{A.63}$$

The first (A.63) term has order of magnitude $O\left(N/n^2\right)$, while the second component has order $O\left(N^2/n^2\right)$. The relative order is N, so in large populations the second (A.63) component will dominate. Thus, Spencer's approximation, which he does not discuss, is reasonable when the population size is large. However, the theory is loose; since $\bar{w}^2$ depends on the sample and population sizes, asymptotically it is not a constant. Dividing (A.61) by $\left(\frac{N}{n}\right)^2$ gives

$$\frac{\frac{N\bar{W}}{n}}{\left(\frac{N}{n}\right)^2} = \left(\frac{n}{N}\right)^2 \frac{N\bar{W}}{n} \quad . \tag{A.64}$$
$$= \frac{n\bar{W}}{N}$$

Therefore, $1+\left[CV\left(\mathbf{w}\right)\right]^2$ approximately estimates $\frac{n\bar{W}}{N}$. Using these and the R-squared value R_{yp}^2 from the model fit to estimate the correlation $\hat{\rho}_{yp}$, we have

$$\widehat{deff}_S = \left(1-R_{yp}^2\right)\left[1+\left[CV\left(\mathbf{w}\right)\right]^2\right]+\left(\frac{\hat{\alpha}}{\hat{\sigma}_y}\right)^2\left[CV\left(\mathbf{w}\right)\right]^2. \tag{A.65}$$

Note that when N is large, $\alpha = \bar{Y} - \beta N^{-1} \approx \bar{Y}$, and (A.59) can be written as

$$deff_S \approx \left(1-\rho_{yp}^2\right)\frac{n\bar{W}}{N}+\left(\frac{\bar{Y}}{\sigma_y}\right)^2\left(\frac{n\bar{W}}{N}-1\right), \tag{A.66}$$

which we estimate with

$$\widehat{deff}_S = \left(1-R_{yp}^2\right)\left[1+\left[CV\left(\mathbf{w}\right)\right]^2\right]+\left[CV_y^2\right]^{-1}\left[CV\left(\mathbf{w}\right)\right]^2, \tag{A.67}$$

where $CV_y^2 = \sigma_y^2 / \bar{Y}^2$ is the unit-level population coefficient of variation squared.

Appendix 9: Proposed Design Effect in Single-Stage Sampling

Let y_i denote the measurement of interest, $\mathbf{x}_i$ a vector of auxiliary variables, p_i the one-draw probability of selection for a sample of size n, and $w_i = \left(np_i\right)^{-1}$ is the base weight for unit i in a population of size N. Again, the average population probability is $\bar{P} = \frac{1}{N}\sum_{i=1}^{N} p_i = \frac{1}{N}$. Consider the model $y_i = A + \mathbf{x}_i^T\mathbf{B} + \varepsilon_i$. If the full finite population were available, the least-squares population regression line would be

$$y_i = \alpha + \mathbf{x}_i^T\boldsymbol{\beta} + e_i, \tag{A.68}$$

where α and $\boldsymbol{\beta}$ are the values found by fitting an ordinary least squares regression line in the full finite population. That is, $\alpha = \bar{Y} - \boldsymbol{\beta}\bar{\mathbf{X}}$, $\boldsymbol{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$, where $\mathbf{X}$ is the $N \times p$ population matrix of auxiliary variables, and $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ is the population mean. The e_i's are defined as the finite population residuals, $e_i = y_i - \alpha - \mathbf{x}_i^T\boldsymbol{\beta}$, and are not superpopulation model errors. Denote the population variance of the y's, e's, e^2, and weights as $\sigma_y^2, \sigma_e^2, \sigma_{e^2}^2, \sigma_w^2$, e.g., $\sigma_y^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2$, and the finite population correlations between the variables in the subscripts as ρ_{yp}, ρ_{ew}, and ρ_{e^2w}. The GREG theoretical design-variance in with-replacement sampling is

$$\begin{aligned} Var\left(\hat{T}_{GREG}\right) &= \frac{1}{n}\sum_{i=1}^{N} p_i \left(\frac{e_i}{p_i} - E_U\right)^2 \\ &= \frac{1}{n}\left(\sum_{i=1}^{N}\frac{e_i^2}{p_i} - E_U^2\right) \end{aligned}, \tag{A.69}$$

where $E_U = \sum_{i=1}^{N} e_i$. Using the model in (A.68) explicitly produces a design effect with several complex terms, many of which contain correlations that cannot be dropped as in Spencer's approximation. The design effect can be simplified using an alternative model formulation: $u_i = \alpha + e_i$, where $u_i = y_i - \mathbf{x}_i^T\boldsymbol{\beta}$. First, we rewrite the population total of the e_i's as

$$\begin{aligned} E_U &= \sum_{i=1}^{N} e_i \\ &= \sum_{i=1}^{N}\left(u_i - \alpha\right), \\ &= N\bar{U} - N\alpha \end{aligned} \tag{A.70}$$

where $\bar{U} = \frac{1}{N}\sum_{i=1}^{N} u_i$. From (A.70) , it follows that

$$\begin{aligned} E_U^2 &= \left(N\bar{U} - N\alpha\right)^2 \\ &= \left(N\bar{U}\right)^2 + \left(N\alpha\right)^2 - 2N^2\bar{U}\alpha \end{aligned} . \tag{A.71}$$

Second, we rewrite the component $\sum_{i=1}^{N} \frac{e_i^2}{p_i}$ as

$$\begin{aligned} \sum_{i=1}^{N} \frac{e_i^2}{p_i} &= \sum_{i=1}^{N} \frac{1}{p_i}\left(u_i - \alpha\right)^2 \\ &= \sum_{i=1}^{N} \frac{1}{p_i}\left(u_i^2 + \alpha^2 - 2\alpha u_i\right) \\ &= \sum_{i=1}^{N} \frac{u_i^2}{p_i} + \alpha^2 \sum_{i=1}^{N} \frac{1}{p_i} - 2\alpha \sum_{i=1}^{N} \frac{u_i}{p_i} \end{aligned} \tag{A.72}$$

Plugging in $w_i = \left(np_i\right)^{-1}$, or $p_i = \left(nw_i\right)^{-1}$ lets us rewrite (A.72) as

$$\begin{aligned} \sum_{i=1}^{N} \frac{e_i^2}{p_i} &= \sum_{i=1}^{N} \frac{u_i^2}{\left(nw_i\right)^{-1}} + \alpha^2 \sum_{i=1}^{N} \frac{1}{\left(nw_i\right)^{-1}} - 2\alpha \sum_{i=1}^{N} \frac{u_i}{\left(nw_i\right)^{-1}} \\ &= n\sum_{i=1}^{N} w_i u_i^2 + n\alpha^2 \sum_{i=1}^{N} w_i - 2n\alpha \sum_{i=1}^{N} w_i u_i \end{aligned} \tag{A.73}$$

Subtracting (A.71) from (A.73) gives

$$\begin{aligned} \sum_{i=1}^{N} \frac{e_i^2}{p_i} - E_U^2 &= n\sum_{i=1}^{N} w_i u_i^2 + n\alpha^2 \sum_{i=1}^{N} w_i - 2n\alpha \sum_{i=1}^{N} w_i u_i \\ &\quad - \left[\left(N\bar{U}\right)^2 + \left(N\alpha\right)^2 - 2N^2\bar{U}\alpha\right] \\ &= n\sum_{i=1}^{N} w_i u_i^2 - \left(N\bar{U}\right)^2 + n\alpha^2 \sum_{i=1}^{N} w_i - \left(N\alpha\right)^2 + 2N^2\bar{U}\alpha - 2n\alpha \sum_{i=1}^{N} w_i u_i \end{aligned} \tag{A.74}$$

Dividing (A.74) by n gives

$$\frac{1}{n}\left(\sum_{i=1}^{N}\frac{e_i^2}{p_i}-E_U^2\right)=\sum_{i=1}^{N}w_iu_i^2-\frac{\left(N\bar{U}\right)^2}{n}+\alpha^2\sum_{i=1}^{N}w_i$$
$$-\frac{\left(N\alpha\right)^2}{n}+\frac{2N^2\bar{U}\alpha}{n}-2\alpha\sum_{i=1}^{N}w_iu_i \qquad \text{(A.75)}$$
$$=\sum_{i=1}^{N}w_iu_i^2-\frac{\left(N\bar{U}\right)^2}{n}+\alpha^2\left(\sum_{i=1}^{N}w_i-\frac{N^2}{n}\right)$$
$$+\frac{2N^2\bar{U}\alpha}{n}-2\alpha\sum_{i=1}^{N}w_iu_i$$

Note that, following Spencer's approach using the covariances in (A.54) and (A.55), the first and fifth terms in (A.75) can be rewritten as

$$\sum_{i=1}^{N}w_iu_i^2=N\rho_{u^2w}\sigma_{u^2}\sigma_w+N\bar{W}\left(\sigma_u^2+\bar{U}^2\right) \qquad \text{(A.76)}$$

and

$$\sum_{i=1}^{N}w_iu_i=N\rho_{uw}\sigma_u\sigma_w+N\bar{W}\bar{U}\,. \qquad \text{(A.77)}$$

Plugging these back into the variance (A.75) gives

$$\frac{1}{n}\left(\sum_{i=1}^{N}\frac{e_i^2}{p_i}-E_U^2\right)=N\rho_{u^2w}\sigma_{u^2}\sigma_w+N\bar{W}\left(\sigma_u^2+\bar{U}^2\right)-\frac{\left(N\bar{U}\right)^2}{n}$$
$$+N\alpha^2\left(\bar{W}-\frac{N}{n}\right)+\frac{2N^2\bar{U}\alpha}{n}-2\alpha\left(N\rho_{uw}\sigma_u\sigma_w+N\bar{W}\bar{U}\right).$$

(A.78)

The variance of the *pwr*-estimator (A.69) under simple random sampling with replacement, where $p_i=N^{-1}$, reduces down to

$$
\begin{aligned}
Var_{srs}\left(\hat{T}_{pwr}\right) &= \frac{1}{n}\sum_{i=1}^{N}\frac{1}{N}\left(\frac{Ny_i}{n}-T\right)^2 \\
&= \frac{N}{n}\sum_{i=1}^{N}\left(y_i-\bar{Y}\right)^2 \\
&= \frac{N^2}{n}\sigma_y^2
\end{aligned} \qquad . \qquad \text{(A.79)}
$$

Taking the ratio of (A.78) to (A.79) gives the following design effect:

$$
\begin{aligned}
deff_{S*} &= \frac{Var\left(\hat{T}_{GREG}\right)}{Var_{srs}\left(\hat{T}_{pwr}\right)} \\
&= \frac{\left[\begin{array}{l} N\rho_{u^2w}\sigma_{u^2}\sigma_w + N\bar{W}\left(\sigma_u^2+\bar{U}^2\right) - \dfrac{\left(N\bar{U}\right)^2}{n} + N\alpha^2\left(\bar{W}-\dfrac{N}{n}\right) \\ +\dfrac{2N^2\bar{U}\alpha}{n} - 2\alpha\left(N\rho_{uw}\sigma_u\sigma_w + N\bar{W}\bar{U}\right) \end{array}\right]}{\dfrac{N^2}{n}\sigma_y^2} \\
&= \frac{n\rho_{u^2w}\sigma_{u^2}\sigma_w}{N\sigma_y^2} + \frac{n\bar{W}}{N}\left(\frac{\sigma_u^2}{\sigma_y^2}\right) + \frac{\bar{U}^2}{\sigma_y^2}\left(\frac{n\bar{W}}{N}-1\right) + \frac{n\alpha^2}{N\sigma_y^2}\left(\bar{W}-\frac{N}{n}\right) + \frac{2\bar{U}\alpha}{\sigma_y^2} \\
&\quad - \frac{2\alpha n\rho_{uw}\sigma_u\sigma_w}{N\sigma_y^2} - \frac{2\alpha n\bar{W}\bar{U}}{N\sigma_y^2} \\
&= \frac{n\rho_{u^2w}\sigma_{u^2}\sigma_w}{N\sigma_y^2} + \frac{n\bar{W}}{N}\left(\frac{\sigma_u^2}{\sigma_y^2}\right) + \frac{\bar{U}^2}{\sigma_y^2}\left(\frac{n\bar{W}}{N}-1\right) + \frac{\alpha^2}{\sigma_y^2}\left(\frac{n\bar{W}}{N}-1\right) \\
&\quad - \frac{2\alpha n\rho_{uw}\sigma_u\sigma_w}{N\sigma_y^2} - \frac{2\alpha\bar{U}}{\sigma_y^2}\left(\frac{n\bar{W}}{N}-1\right) \\
&= \frac{n\bar{W}}{N}\left(\frac{\sigma_u^2}{\sigma_y^2}\right) + \left(\frac{\bar{U}^2+\alpha^2-2\bar{U}\alpha}{\sigma_y^2}\right)\left(\frac{n\bar{W}}{N}-1\right) + \frac{n\sigma_w}{N\sigma_y^2}\left[\rho_{u^2w}\sigma_{u^2} - 2\alpha\rho_{uw}\sigma_u\right] \\
&= \frac{n\bar{W}}{N}\left(\frac{\sigma_u^2}{\sigma_y^2}\right) + \frac{\left(\bar{U}-\alpha\right)^2}{\sigma_y^2}\left(\frac{n\bar{W}}{N}-1\right) + \frac{n\sigma_w}{N\sigma_y^2}\left[\rho_{u^2w}\sigma_{u^2} - 2\alpha\rho_{uw}\sigma_u\right]
\end{aligned}
$$

(A.80)

Note that under our model $u_i = \alpha + e_i, \bar{U} = \alpha$, and (A.80) becomes

$$deff_{S*} = \frac{n\bar{W}}{N}\left(\frac{\sigma_u^2}{\sigma_y^2}\right) + \frac{n\sigma_w}{N\sigma_y^2}\left[\rho_{u^2w}\sigma_{u^2} - 2\alpha\rho_{uw}\sigma_u\right] \qquad \text{(A.81)}$$

We estimate measure (A.81) with

$$\widehat{deff_{S*}} \approx \left(1 + \left[CV\left(\mathbf{w}\right)\right]^2\right)\frac{\hat{\sigma}_u^2}{\hat{\sigma}_y^2} + \frac{n\hat{\sigma}_w}{N\hat{\sigma}_y^2}\left[\hat{\rho}_{u^2w}\hat{\sigma}_{u^2} - 2\hat{\alpha}\hat{\rho}_{uw}\hat{\sigma}_u\right] \qquad \text{(A.82)}$$

where the model parameter estimates are obtained using survey-weighted least squares,

$$\hat{\bar{y}}_w = \frac{\sum_{i\in s} w_i y_i}{\sum_{i\in s} w_i},\ \hat{\sigma}_y^2 = \frac{\sum_{i\in s} w_i\left(y_i - \hat{\bar{y}}_w\right)^2}{\sum_{i\in s} w_i},\ \hat{\sigma}_u^2 = \frac{\sum_{i\in s} w_i\left(\hat{u}_i - \hat{\bar{u}}_w\right)^2}{\sum_{i\in s} w_i},\ \hat{\bar{u}}_w = \frac{\sum_{i\in s} w_i\hat{u}_i}{\sum_{i\in s} w_i},$$

and $\hat{u}_i = y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}} = \left(\mathbf{x}^T\mathbf{x}\right)^{-1}\mathbf{x}^T\mathbf{y}$. Similar to Spencer's approach, if the correlations in the last component of (A.80) are negligible, then (A.80) can be approximated by

$$\begin{aligned} deff_{S*} &\approx \frac{n\bar{W}}{N}\left(\frac{\sigma_u^2}{\sigma_y^2}\right) + \frac{\left(\bar{U} - \alpha\right)^2}{\sigma_y^2}\left(\frac{n\bar{W}}{N} - 1\right) \\ &= \frac{n\bar{W}}{N}\left(\frac{\sigma_u^2}{\sigma_y^2}\right) \end{aligned}. \qquad \text{(A.83)}$$

Since $\bar{U} = \alpha$, measure (A.83) can be estimated using

$$\widehat{deff_{S*}} \approx \left(1 + \left[CV\left(\mathbf{w}\right)\right]^2\right)\frac{\hat{\sigma}_u^2}{\hat{\sigma}_y^2}. \qquad \text{(A.84)}$$

Note that without calibration, we have $\hat{u}_i = y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}} \approx y_i$, and $\sigma_u^2 \approx \sigma_y^2$. In that case, the design effect approximation in (A.84) becomes $deff_{S*} \approx \frac{n\bar{W}}{N}$, which we estimate with Kish's measure $\widehat{deff_K} \approx 1 + \left[CV\left(\mathbf{w}\right)\right]^2$.

Appendix 10: Proposed Design Effect in Cluster Sampling

For cluster sampling, we start with N clusters in the population, with M_i elements within cluster i. Consider the model $y_{ij} = A_U + \mathbf{x}_{ij}^T \mathbf{B} + e_{ij}$, where $A, \mathbf{B}$ are the finite population model parameters and $e_{ij} = y_{ij} - A_U - \mathbf{x}_{ij}^T \mathbf{B} = y_{ij} - \dot{\mathbf{x}}_{ij}^T \mathbf{B}$. If the full finite population were in hand, then we could fit the model by ordinary least squares to obtain $y_{ij} = \alpha_U + \mathbf{x}_{ij}^T \boldsymbol{\beta} + e_{ij} = \dot{\mathbf{x}}_{ij}^T \mathbf{B}_U + e_{ij}$, where $\dot{\mathbf{x}}_{ij} = \begin{bmatrix} 1 & \mathbf{x}_{ij} \end{bmatrix}$, $\mathbf{B}_U = \begin{pmatrix} \alpha_U & \boldsymbol{\beta} \end{pmatrix}^T$, $e_{ij} = y_{ij} - \alpha_U - \mathbf{x}_{ij}^T \boldsymbol{\beta} = y_{ij} - \dot{\mathbf{x}}_{ij}^T \mathbf{B}_U$, $\alpha = \bar{Y} - \boldsymbol{\beta}\bar{\mathbf{X}}$, $\boldsymbol{\beta} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{y}$, $\mathbf{X}$ is the $N \times p$ population matrix of auxiliary variables, and $\bar{Y}_U = \sum_{i \in U} \sum_{j \in U_i} Y_{ij} \Big/ \sum_{i \in U} M_i$ is the population mean. The GREG estimator here is given by

$$
\begin{aligned}
\hat{T}_{GREG} &= \hat{T}_{HT} + \left(\mathbf{T}_x - \hat{\mathbf{T}}_{HTx}\right)^T \hat{\mathbf{B}} \\
&\doteq \hat{T}_{HT} + \left(\mathbf{T}_x - \hat{\mathbf{T}}_{HTx}\right)^T \mathbf{B}_U \\
&= \sum_{i \in s} \sum_{j \in s_i} \left(\frac{y_{ij}}{\pi_{ij}} - \frac{\dot{\mathbf{x}}_{ij}^T}{\pi_{ij}} \mathbf{B}_U \right) + \mathbf{T}_x^T \mathbf{B}_U \\
&= \sum_{i \in s} \sum_{j \in s_i} \frac{e_{ij}}{\pi_{ij}} + \mathbf{T}_x^T \mathbf{B}_U
\end{aligned}
\qquad \text{(A.85)}
$$

From (A.85), $\hat{T}_{GREG} - \mathbf{T}_x^T \mathbf{B}_U \doteq \sum_{i \in s} \sum_{j \in s_i} \frac{e_{ij}}{\pi_{ij}}$. Assuming that we have probability-with-replacement (pwr) sampling of clusters, the probability of selection for clusters is approximately $\pi_i = 1 - \left(1 - p_i\right)^n \doteq np_i$ (if p_i is not too large), where p_i is the one-drawn selection probability. Suppose that simple random sampling is used within each cluster, such that the second-stage selection probability is $\pi_{j|i} = \dfrac{m_i}{M_i}$ for element j in cluster i.

Then the overall selection probability is approximately $\pi_{ij} = \pi_i \pi_{j|i} \doteq \dfrac{np_i m_i}{M_i}$ and expression (A.85) becomes

$$\begin{aligned} \hat{T}_{GREG} - \mathbf{T}_x^T \mathbf{B}_U &\doteq \sum\nolimits_{i \in s} \sum\nolimits_{j \in s_i} \frac{M_i e_{ij}}{np_i m_i} \\ &= \sum\nolimits_{i \in s} w_i \hat{T}_{ei} \end{aligned} \tag{A.86}$$

where $w_i = \left(np_i\right)^{-1}$ and $\hat{T}_{ei} = \dfrac{M_i}{m_i} \sum\nolimits_{j \in s_i} e_{ij}$. The approximate theoretical variance is:

$$\begin{aligned} Var\left(\hat{T}_{GREG}\right) &= \frac{1}{n} \sum\nolimits_{i=1}^{N} p_i \left(\frac{e_{Ui+}}{p_i} - E_{U+} \right)^2 + \sum\nolimits_{i=1}^{N} \frac{M_i^2}{np_i m_i} \left(1 - \frac{m_i}{M_i} \right) \frac{1}{M_i - 1} \sum\nolimits_{j=1}^{M_i} \left(e_{ij} - \bar{e}_{U_i} \right)^2 \\ &= \frac{1}{n} \sum\nolimits_{i=1}^{N} \left(\frac{e_{Ui+}^2}{p_i} - E_{U+}^2 \right) + \sum\nolimits_{i=1}^{N} \frac{M_i^2}{np_i m_i} \left(1 - \frac{m_i}{M_i} \right) S_{U_{ei}}^2 \end{aligned}$$

(A.87)

where $\bar{e}_{U_i} = \dfrac{1}{M_i} \sum\nolimits_{j=1}^{M_i} e_{ij}$, $e_{U_i} = \sum\nolimits_{j=1}^{M_i} e_{ij}$, $E_{U+} = \sum\nolimits_{i=1}^{N} e_{ij}$, and

$S_{U_{ei}}^2 = \dfrac{1}{M_i - 1} \sum\nolimits_{j=1}^{M_i} \left(e_{ij} - e_{Ui} \right)^2$.

Suppose that the second-stage sampling fraction is negligible, i.e., $\dfrac{m_i}{M_i} \approx 0$. As with the single-stage design effect, to simplify notation here we reformulate the model $y_{ij} = \alpha_U + \mathbf{x}_{ij}^T \boldsymbol{\beta} + e_{ij}$ using $u_{ij} = y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta}$, such that $e_{ij} = u_{ij} - \alpha_U$. Estimation of the two components in (A.87) is examined separately, then put together to produce the design effect.

First Variance Component Derivation

To derive the design effect using the variance in (A.87), we examine the separate components with respect to the model. First, for the cluster-level component, we define

$$\begin{aligned} u_{i+} &= \sum_{j \in U_i} \left(y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta} \right) \\ &= y_{i+} - \mathbf{x}_{i+}^T \boldsymbol{\beta} \end{aligned} \qquad \text{(A.88)}$$

and

$$\begin{aligned} e_{Ui+} &= \sum_{j \in U_i} e_{ij} \\ &= \sum_{j \in U_i} \left(u_{ij} - \alpha_U \right) \\ &= u_{i+} - M_i \alpha_U \\ &= u_{i+} - \alpha_i \end{aligned}, \qquad \text{(A.89)}$$

where $\alpha_i = M_i \alpha$. We rewrite the population total as

$$\begin{aligned} E_U &= \sum_{i=1}^{N} \sum_{j=1}^{M_i} e_{ij} \\ &= \sum_{i=1}^{N} e_{i+} \\ &= \sum_{i=1}^{N} \left(u_{i+} - \alpha_i \right) \\ &= N\bar{U} - N\bar{\alpha} \end{aligned}. \qquad \text{(A.90)}$$

where $\bar{U} = \frac{1}{N} \sum_{i=1}^{N} u_{i+}$, $\bar{\alpha} = \frac{1}{N} \sum_{i=1}^{N} \alpha_i = \alpha_U \bar{M}$, with $\bar{M} = \frac{1}{N} \sum_{i=1}^{N} M_i$ and $e_{i+} = \sum_{j=1}^{M_i} e_{ij}$. Consequently,

$$\begin{aligned} E_U^2 &= \left(N\bar{U} - N\bar{\alpha} \right)^2 \\ &= \left(N\bar{U} \right)^2 + \left(N\bar{\alpha} \right)^2 - 2N^2 \bar{U} \bar{\alpha} \end{aligned}. \qquad \text{(A.91)}$$

We then rewrite the component $\sum_{i=1}^{N} \frac{e_{Ui+}^2}{p_i}$ in (A.87) as

$$\begin{aligned}\sum_{i=1}^{N}\frac{e_{Ui+}^2}{p_i} &= \sum_{i=1}^{N}\frac{1}{p_i}\left(u_{i+}-\alpha_i\right)^2 \\ &= \sum_{i=1}^{N}\frac{1}{p_i}\left(u_{i+}^2+\alpha_i^2-2\alpha_i u_{i+}\right) \\ &= \sum_{i=1}^{N}\frac{u_{i+}^2}{p_i}+\sum_{i=1}^{N}\frac{\alpha_i^2}{p_i}-2\sum_{i=1}^{N}\frac{\alpha_i u_{i+}}{p_i}\end{aligned} \quad . \tag{A.92}$$

Plugging in $w_i = \left(np_i\right)^{-1}$, or $p_i = \left(nw_i\right)^{-1}$ lets us rewrite (A.92) as

$$\begin{aligned}\sum_{i=1}^{N}\frac{e_{Ui+}^2}{p_i} &= \sum_{i=1}^{N}\frac{u_{i+}^2}{\left(nw_i\right)^{-1}}+\sum_{i=1}^{N}\frac{\alpha_i^2}{\left(nw_i\right)^{-1}}-2\sum_{i=1}^{N}\frac{\alpha_i u_{i+}}{\left(nw_i\right)^{-1}} \, . \\ &= n\sum_{i=1}^{N}w_i u_{i+}^2+n\sum_{i=1}^{N}w_i\alpha_i^2-2n\sum_{i=1}^{N}w_i\alpha_i u_{i+}\end{aligned} \tag{A.93}$$

Subtracting (A.91) from (A.93) gives

$$\begin{aligned}\sum_{i=1}^{N}\frac{e_{Ui+}^2}{p_i}-E_{U+}^2 &= n\sum_{i=1}^{N}w_i u_{i+}^2+n\sum_{i=1}^{N}w_i\alpha_i^2-2n\sum_{i=1}^{N}w_i\alpha_i u_{i+} \\ &\quad -\left[\left(N\bar{U}\right)^2+\left(N\bar{\alpha}\right)^2-2N^2\bar{U}\bar{\alpha}\right] \\ &= n\sum_{i=1}^{N}w_i u_{i+}^2-\left(N\bar{U}\right)^2+n\sum_{i=1}^{N}w_i\alpha_i^2-\left(N\bar{\alpha}\right)^2 \\ &\quad +2N^2\bar{U}\bar{\alpha}-2n\sum_{i=1}^{N}w_i\alpha_i u_{i+}\end{aligned} \tag{A.94}$$

Dividing (A.94) by n gives

$$\begin{aligned}\frac{1}{n}\left(\sum_{i=1}^{N}\frac{e_{Ui+}^2}{p_i}-E_U^2\right) &= \sum_{i=1}^{N}w_i u_{i+}^2-\frac{\left(N\bar{U}\right)^2}{n}+\sum_{i=1}^{N}w_i\alpha_i^2-\frac{\left(N\bar{\alpha}\right)^2}{n} \\ &\quad +\frac{2N^2\bar{U}\bar{\alpha}}{n}-2\sum_{i=1}^{N}w_i\alpha_i u_{i+}\end{aligned} \tag{A.95}$$

First, we write the covariance as

$$
\begin{aligned}
NCov\left(w_i, u_{i+}\alpha_i\right) &\equiv \sum_{i=1}^{N}\left(w_i - \bar{W}\right)\left(u_{i+}\alpha_i - \overline{U\alpha}\right) \\
&= \sum_{i=1}^{N} w_i u_{i+}\alpha_i - \frac{1}{N}\left(\sum_{i=1}^{N} w_i\right)\left(\sum_{i=1}^{N} u_{i+}\alpha_i\right), \\
&= \sum_{i=1}^{N} w_i u_{i+}\alpha_i - \bar{W}N\overline{U\alpha}
\end{aligned}
\tag{A.96}
$$

where $\overline{U\alpha} = \frac{1}{N}\sum_{i=1}^{N} u_{i+}\alpha_i$. By definition $Cov\left(w_i, u_{i+}\alpha_i\right) \equiv \rho_{u_+\alpha,w}\sigma_{u_+\alpha}\sigma_w$, where $\rho_{u_+\alpha,w}$ is the unweighted correlation of $u_{i+}\alpha_i$ and w_i. From this and expression (A.96), we have

$$
\sum_{i=1}^{N} w_i u_{i+}\alpha_i = N\rho_{u_+\alpha,w}\sigma_{u_+\alpha}\sigma_w + N\bar{W}\overline{U\alpha}\,, \tag{A.97}
$$

Similarly, $\sum_{i=1}^{N} w_i u_{i+}^2 - N\bar{W}\overline{U^2} = N\bar{W}\sigma_{u+}^2$ and $\sum_{i=1}^{N} w_i \alpha_i^2 - N\bar{W}\overline{\alpha^2} = N\bar{W}\sigma_{\alpha}^2$, where $\overline{U^2} = \frac{1}{N}\sum_{i=1}^{N} u_{i+}^2 = \sigma_{u+}^2 + \bar{U}^2$ and $\overline{\alpha^2} = \frac{1}{N}\sum_{i=1}^{N} \alpha_i^2 = \sigma_{\alpha}^2 + \bar{\alpha}^2$. These results mean that the first, third, and sixth terms in (A.95) can be rewritten as

$$
\begin{aligned}
\sum_{i=1}^{N} w_i u_{i+}^2 &= N\rho_{u_+^2 w}\sigma_{u_+^2}\sigma_w + N\bar{W}\overline{U^2} \\
&= N\rho_{u_+^2 w}\sigma_{u_+^2}\sigma_w + N\bar{W}\left(\sigma_{u_+}^2 + \bar{U}^2\right)
\end{aligned}
\tag{A.98}
$$

and

$$
\begin{aligned}
\sum_{i=1}^{N} w_i \alpha_i^2 &= N\rho_{\alpha^2 w}\sigma_{\alpha^2}\sigma_w + N\bar{W}\overline{\alpha^2} \\
&= N\rho_{\alpha^2 w}\sigma_{\alpha^2}\sigma_w + N\bar{W}\left(\sigma_{\alpha}^2 + \bar{\alpha}^2\right)
\end{aligned}
\tag{A.99}
$$

Plugging (A.97), (A.98), and (A.99) into the variance (A.95) gives

$$\frac{1}{n}\left(\sum_{i=1}^{N}\frac{e_{Ui+}^2}{p_i}-E_U^2\right)=N\rho_{u_+^2 w}\sigma_{u_+^2}\sigma_w+N\bar{W}\left(\sigma_{u_+}^2+\bar{U}^2\right)-\frac{\left(N\bar{U}\right)^2}{n}+$$
$$+N\rho_{\alpha^2 w}\sigma_{\alpha^2}\sigma_w+N\bar{W}\left(\sigma_\alpha^2+\bar{\alpha}^2\right)-\frac{\left(N\bar{\alpha}\right)^2}{n} \quad . \tag{A.100}$$
$$+\frac{2N^2\bar{U}\bar{\alpha}}{n}-2\left(N\rho_{u_+\alpha,w}\sigma_{u_+\alpha}\sigma_w+N\bar{W}\overline{U\alpha}\right)$$

Second Variance Component Derivation

For the second component in the variance (A.87), for $u_{ij}=y_{ij}-\mathbf{x}_{ij}^T\boldsymbol{\beta}$, $e_{ij}=y_{ij}-\alpha_U-\mathbf{x}_{ij}^T\mathbf{B}=u_{ij}-\alpha_U$ and $w_{ij}=\frac{M_i}{m_i n p_i}$, we need to derive the term $\sum_{j=1}^{M_i}\left(e_{ij}-\bar{e}_{Ui}\right)^2=\sum_{j=1}^{M_i}e_{ij}^2-\sum_{j=1}^{M_i}\bar{e}_{Ui}^2$. To do this, we first rewrite the population mean $\bar{e}_{Ui}$ as

$$\begin{aligned}\bar{e}_{Ui}&=\frac{1}{M_i}\sum_{j=1}^{M_i}e_{ij}\\&=\frac{1}{M_i}\sum_{j=1}^{M_i}\left(u_{ij}-\alpha_U\right)\\&=\frac{1}{M_i}\sum_{j=1}^{M_i}u_{ij}-\frac{1}{M_i}\sum_{j=1}^{M_i}\alpha_U\\&=\bar{u}_i-\alpha_U\end{aligned} \quad . \tag{A.101}$$

From this, $\bar{e}_{Ui}^2=\left(\bar{u}_i-\alpha_U\right)^2=\bar{u}_i^2+\alpha_U^2-2\bar{u}_i\alpha_U$ and

$$\begin{aligned}e_{ij}-\bar{e}_{U_i}&=u_{ij}-\alpha_U-\bar{u}_i+\alpha_U\\&=u_{ij}-\bar{u}_i\end{aligned} \quad , \tag{A.102}$$

where $\bar{u}_i=\frac{u_{i+}}{M_i}$ is the mean of u_{ij} within cluster i. Then we have

$$\begin{aligned}\sum_{j=1}^{M_i}\left(e_{ij}-\bar{e}_{U_i}\right)^2 &= \sum_{j=1}^{M_i}\left(u_{ij}-\bar{u}_i\right)^2 \\ &\equiv \left(M_i-1\right)S_{U_{ui}}^2\end{aligned} . \tag{A.103}$$

From this, the variance component is

$$\begin{aligned}&\sum_{i=1}^{N}\frac{M_i^2}{np_i m_i}\left(1-\frac{m_i}{M_i}\right)\frac{1}{M_i-1}\sum_{j=1}^{M_i}\left(e_{ij}-\bar{e}_{Ui}\right)^2 \\ &\quad=\sum_{i=1}^{N}\frac{M_i^2}{np_i m_i}\left(1-\frac{m_i}{M_i}\right)\frac{1}{M_i-1}\sum_{j=1}^{M_i}\left(u_{ij}-\bar{u}_i\right)^2 . \\ &\quad=\sum_{i=1}^{N}\frac{M_i^2}{m_i}\left(1-\frac{m_i}{M_i}\right)w_i S_{U_{ui}}^2\end{aligned} \tag{A.104}$$

Note that, since simple random sampling within each cluster was assumed, (A.104) does not contain any differential within-cluster weights or correlations. If an alternative design was used to select units within clusters, then (A.104) will include additional related terms (such as the correlations in (A.100)).

Design Effect Derivations

Taking the ratio of (A.100) to the SRSWR variance of the PWR estimator gives the first design effect component as

$$deff_{C1} = \frac{\begin{bmatrix} N\rho_{u_+^2 w}\sigma_{u_+^2}\sigma_w + N\bar{W}\left(\sigma_{u_+}^2 + \bar{U}^2\right) - \frac{\left(N\bar{U}\right)^2}{n} + N\rho_{\alpha^2 w}\sigma_{\alpha^2}\sigma_w + \\ +N\bar{W}\left(\sigma_\alpha^2 + \bar{\alpha}^2\right) - \frac{\left(N\bar{\alpha}\right)^2}{n} + \frac{2N^2\bar{U}\bar{\alpha}}{n} - 2N\rho_{u_+\alpha,w}\sigma_{u_+\alpha}\sigma_w - 2N\bar{W}\overline{U\alpha} \end{bmatrix}}{\frac{N^2}{n}\sigma_y^2}$$

$$= \frac{n\rho_{u_+^2 w}\sigma_{u_+^2}\sigma_w}{N\sigma_y^2} + \frac{n\bar{W}}{N}\left(\frac{\sigma_{u_+}^2 + \bar{U}^2}{\sigma_y^2}\right) - \frac{\bar{U}^2}{\sigma_y^2} + \frac{n\rho_{\alpha^2 w}\sigma_{\alpha^2}\sigma_w}{N\sigma_y^2}$$

$$+ \frac{n\bar{W}}{N}\left(\frac{\sigma_\alpha^2 + \bar{\alpha}^2}{\sigma_y^2}\right) - \frac{\bar{\alpha}^2}{\sigma_y^2} + \frac{2\bar{U}\bar{\alpha}}{\sigma_y^2} - 2\frac{n\rho_{u_+\alpha,w}\sigma_{u_+\alpha}\sigma_w}{N\sigma_y^2} - 2\frac{n\bar{W}\overline{U\alpha}}{N\sigma_y^2}$$

$$= \frac{n\bar{W}}{N}\left(\frac{\sigma_{u_+}^2 + \bar{U}^2 + \sigma_\alpha^2 + \bar{\alpha}^2 - 2\overline{U\alpha}}{\sigma_y^2}\right) - \frac{\bar{U}^2}{\sigma_y^2} - \frac{\bar{\alpha}^2}{\sigma_y^2} + \frac{2\bar{U}\bar{\alpha}}{\sigma_y^2}$$

$$+ \frac{n\left(\rho_{u_+^2 w}\sigma_{u_+^2}\sigma_w + \rho_{\alpha^2 w}\sigma_{\alpha^2}\sigma_w - 2\rho_{u_+\alpha,w}\sigma_{u_+\alpha}\sigma_w\right)}{N\sigma_y^2}$$

$$= \frac{n\bar{W}}{N}\left(\frac{\sigma_{u_+}^2 + \sigma_\alpha^2 - 2\overline{U\alpha}}{\sigma_y^2}\right) + \frac{n\bar{W}}{N}\left(\frac{\bar{U}^2 + \bar{\alpha}^2}{\sigma_y^2}\right) - \frac{\bar{U}^2 + \bar{\alpha}^2}{\sigma_y^2}$$

$$+ \frac{n\sigma_w\left(\rho_{u_+^2 w}\sigma_{u_+^2} + \rho_{\alpha^2 w}\sigma_{\alpha^2} - 2\rho_{u_+\alpha,w}\sigma_{u_+\alpha}\right)}{N\sigma_y^2} + \frac{2\bar{U}\bar{\alpha}}{\sigma_y^2}$$

$$= \frac{n\bar{W}}{N}\left(\frac{\sigma_{u_+}^2 + \sigma_\alpha^2 - 2\overline{U\alpha}}{\sigma_y^2}\right) + \left(\frac{n\bar{W}}{N} - 1\right)\left(\frac{\bar{U}^2 + \bar{\alpha}^2}{\sigma_y^2}\right) + \frac{2\bar{U}\bar{\alpha}}{\sigma_y^2} \qquad . \text{(A.105)}$$

$$+ \frac{n\sigma_w\left(\rho_{u_+^2 w}\sigma_{u_+^2} + \rho_{\alpha^2 w}\sigma_{\alpha^2} - 2\rho_{u_+\alpha,w}\sigma_{u_+\alpha}\right)}{N\sigma_y^2}$$

Taking the ratio of (A.104) to the SRSWR variance of the PWR estimator and assuming that the within-cluster sampling fractions are negligible gives the second design effect component as

$$deff_{C2} = \frac{\sum_{i=1}^{N} \frac{M_i^2}{m_i}\left(1-\frac{m_i}{M_i}\right) w_i S_{U_{ui}}^2}{\frac{N^2}{n}\sigma_y^2} \doteq \frac{n}{N^2\sigma_y^2}\sum_{i=1}^{N}\frac{M_i^2 w_i S_{U_{ui}}^2}{m_i} \quad \text{(A.106)}$$

The total design effect is thus

$$\begin{aligned} deff_C &= deff_{C1} + deff_{C2} \\ &\approx \frac{n\bar{W}}{N}\left(\frac{\sigma_{u_+}^2 + \sigma_\alpha^2 - 2\overline{U\alpha}}{\sigma_y^2}\right) + \left(\frac{n\bar{W}}{N} - 1\right)\left(\frac{\bar{U}^2 + \bar{\alpha}^2}{\sigma_y^2}\right) + \frac{2\bar{U}\bar{\alpha}}{\sigma_y^2} \\ &\quad + \frac{n\sigma_w\left(\rho_{u_+^2 w}\sigma_{u_+^2} + \rho_{\alpha^2 w}\sigma_{\alpha^2} - 2\rho_{u_+\alpha,w}\sigma_{u_+\alpha}\right)}{N\sigma_y^2} + \frac{n}{N^2\sigma_y^2}\sum_{i=1}^{N}\frac{M_i^2 w_i S_{U_{ui}}^2}{m_i} \end{aligned} \quad \text{(A.107)}$$

Measure (A.107) can be estimated using

$$\begin{aligned} \widehat{deff_C} &\approx \left[1 + \left(CV(\mathbf{w})\right)^2\right]\left(\frac{\hat{\sigma}_{u_+}^2 + \hat{\sigma}_\alpha^2 - 2\widehat{u_w\alpha_w}}{\hat{\sigma}_y^2}\right) + \left(CV(\mathbf{w})\right)^2\left(\frac{\hat{\bar{u}}_w^2 + \hat{\bar{\alpha}}_w^2}{\hat{\sigma}_y^2}\right) + \frac{2\hat{\bar{u}}_w\hat{\bar{\alpha}}_w}{N\hat{\sigma}_y^2} \\ &\quad + \frac{n\hat{\sigma}_w\left(\hat{\rho}_{u_+^2 w}\hat{\sigma}_{u_+^2} + \hat{\rho}_{\alpha^2 w}\hat{\sigma}_{\alpha^2} - 2\hat{\rho}_{u_+\alpha,w}\hat{\sigma}_{u_+\alpha}\right)}{N\hat{\sigma}_y^2} + \frac{n}{N^2\hat{\sigma}_y^2}\sum_{i=1}^{n}\frac{M_i^2 w_i \hat{S}_{U_{ui}}^2}{m_i} \end{aligned} \quad \text{(A.108)}$$

where the model parameter estimates $\hat{\alpha}_i$ are obtained using survey-weighted least squares, $\hat{\sigma}_\alpha^2 = \frac{\sum_{i\in s} w_i\left(\hat{\alpha}_i - \hat{\bar{\alpha}}_w\right)^2}{\sum_{i\in s} w_i}$, $\hat{\bar{\alpha}}_w = \frac{\sum_{i\in s} w_i\hat{\alpha}_i}{\sum_{i\in s} w_i}$, $\hat{\sigma}_{u_+}^2 = \frac{\sum_{i\in s} w_i\left(\hat{u}_i - \hat{\bar{u}}_w\right)^2}{\sum_{i\in s} w_i}$,

$\hat{S}_{U_{ui}}^2 = \frac{\sum_{j\in s_i}\left(\hat{u}_{ij} - \hat{\bar{u}}_i\right)^2}{m_i - 1}$, $\widehat{u_w\alpha_w} = \frac{\sum_{i=1}^{n} w_i\hat{u}_{i+}\hat{\alpha}_i}{\sum_{i=1}^{n} w_i}$, $\hat{\bar{u}}_w = \frac{\sum_{i\in s} w_i\hat{u}_{i+}}{\sum_{i\in s} w_i}$,

$$\hat{\sigma}_y^2 = \frac{\sum_{i\in s}\sum_{j\in s_i} w_{ij}\left(y_{ij} - \hat{\bar{y}}_w\right)^2}{\sum_{i\in s}\sum_{j\in s_i} w_{ij}}, \qquad \hat{\bar{y}}_w = \frac{\sum_{i\in s}\sum_{j\in s_i} w_{ij} y_{ij}}{\sum_{i\in s}\sum_{j\in s_i} w_{ij}},$$

$\hat{u}_{i+} = y_{i+} - \mathbf{x}_{i+}^T\hat{\boldsymbol{\beta}} = \sum_{j\in s_i}\hat{u}_{ij} = \sum_{j\in s_i}\left(y_{ij} - \mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}\right)$, and $\hat{\bar{u}}_i = \frac{1}{m_i}\sum_{j\in s_i}\hat{u}_{ij}$.

Assuming that the three correlations in (A.107) are negligible gives:

$$\begin{aligned} deff_C \approx & \frac{n\bar{W}}{N}\left(\frac{\sigma_{u_+}^2 + \sigma_\alpha^2 - 2\overline{U\alpha}}{\sigma_y^2}\right) + \left(\frac{n\bar{W}}{N} - 1\right)\left(\frac{\bar{U}^2 + \bar{\alpha}^2}{\sigma_y^2}\right) + \frac{2\bar{U}\bar{\alpha}}{\sigma_y^2} \\ & + \frac{n}{N^2\sigma_y^2}\sum_{i=1}^N \frac{M_i^2 w_i S_{U_{ui}}^2}{m_i} \end{aligned} . \tag{A.109}$$

The approximate design effect measure (A.109) can be estimated using

$$\begin{aligned} \widehat{deff_C} \approx & \left[1 + \left(CV\left(\mathbf{w}\right)\right)^2\right]\left(\frac{\hat{\sigma}_{u_+}^2 + \hat{\sigma}_\alpha^2 - 2\widehat{\overline{u_w\alpha_w}}}{\hat{\sigma}_y^2}\right) + \left(CV\left(\mathbf{w}\right)\right)^2\left(\frac{\hat{\bar{u}}_w^2 + \hat{\bar{\alpha}}_w^2}{\hat{\sigma}_y^2}\right) + \frac{2\hat{\bar{u}}_w\hat{\bar{\alpha}}_w}{N\hat{\sigma}_y^2} \\ & + \frac{n}{N^2\hat{\sigma}_y^2}\sum_{i=1}^n \frac{M_i^2 w_i \hat{S}_{U_{ui}}^2}{m_i} \end{aligned} .$$

(A.110)

Assuming that M_i are close enough such that $M_i \approx \bar{M}$ and $\alpha_i = M_i\alpha \doteq \bar{M}\alpha$ and $\sigma_\alpha^2 = \sigma_{\alpha^2} = 0$, then

$$\begin{aligned} \bar{\alpha} &= \frac{1}{N}\sum_{i=1}^N \alpha_i \\ &= \frac{1}{N}\sum_{i=1}^N \bar{M}\alpha \\ &= \bar{M}\alpha \end{aligned} \tag{A.111}$$

and

$$\begin{aligned}\overline{U\alpha} &= \frac{1}{N}\sum_{i=1}^{N} u_{i+}\alpha_i \\ &= \frac{1}{N}\sum_{i=1}^{N} u_{i+}\bar{M}\alpha \\ &= \bar{M}\alpha\bar{U}\end{aligned} \tag{A.112}$$

and

$$\begin{aligned}\sigma_{u+\alpha} &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(u_{i+}\alpha_i - \overline{U\alpha}\right)^2} \\ &= \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(u_{i+}\bar{M}\alpha - \left(\bar{M}\alpha\bar{U}\right)\right)^2} \\ &= \sqrt{\frac{\left(\bar{M}\alpha\right)^2}{N}\sum_{i=1}^{N}\left(u_{i+} - \bar{U}\right)^2} \\ &= \bar{M}\alpha\sigma_u\end{aligned} \tag{A.113}$$

From (A.112) and (A.113), expression (A.107) is approximately

$$\begin{aligned}deff_C &\approx \frac{n\bar{W}}{N}\left(\frac{\sigma_{u_+}^2 - 2\bar{M}\alpha\bar{U}}{\sigma_y^2}\right) + \left(\frac{n\bar{W}}{N} - 1\right)\left(\frac{\bar{U}^2 + \left(\bar{M}\alpha\right)^2}{\sigma_y^2}\right) + \frac{2\bar{U}\bar{M}\alpha}{\sigma_y^2} \\ &\quad + \frac{n\sigma_w\left(\rho_{u_+^2 w}\sigma_{u_+^2} - 2\bar{M}\alpha\sigma_u\rho_{u_+\alpha,w}\right)}{N\sigma_y^2} + \frac{n}{N^2\sigma_y^2}\sum_{i=1}^{N}\frac{\bar{M}^2 w_i S_{U_{ui}}^2}{m_i} \\ &= \frac{n\bar{W}}{N}\left(\frac{\sigma_{u_+}^2}{\sigma_y^2}\right) + \left(\frac{n\bar{W}}{N} - 1\right)\left(\frac{\bar{U}^2 + \left(\bar{M}\alpha\right)^2 - 2\bar{M}\alpha\bar{U}}{\sigma_y^2}\right) \\ &\quad + \frac{n\sigma_w\left(\rho_{u_+^2 w}\sigma_{u_+^2} - 2\bar{M}\alpha\sigma_u\rho_{u_+\alpha,w}\right)}{N\sigma_y^2} + \frac{n}{N^2\sigma_y^2}\sum_{i=1}^{N}\frac{\bar{M}^2 w_i S_{U_{ui}}^2}{m_i} \\ &= \frac{n\bar{W}}{N}\left(\frac{\sigma_{u_+}^2}{\sigma_y^2}\right) + \left(\frac{n\bar{W}}{N} - 1\right)\frac{\left(\bar{U} - \bar{M}\alpha\right)^2}{\sigma_y^2} + \\ &\quad + \frac{n\sigma_w\left(\rho_{u_+^2 w}\sigma_{u_+^2} - 2\bar{M}\alpha\sigma_u\rho_{u_+\alpha,w}\right)}{N\sigma_y^2} + \frac{n}{N^2\sigma_y^2}\sum_{i=1}^{N}\frac{\bar{M}^2 w_i S_{U_{ui}}^2}{m_i}.\end{aligned} \tag{A.114}$$

Measure (A.109) can be estimated using

$$\widehat{deff_C} \approx \left[1+\left(CV(\mathbf{w})\right)^2\right]\frac{\hat{\sigma}_{u_+}^2}{\sigma_y^2}+\left(CV(\mathbf{w})\right)^2\frac{\left(\hat{\bar{u}}_w-\bar{M}\hat{\alpha}_w\right)^2}{\hat{\sigma}_y^2}$$
$$+\frac{n\hat{\sigma}_w\left(\hat{\rho}_{u_+^2 w}\hat{\sigma}_{u_+^2}-2\bar{M}\hat{\alpha}_w\hat{\rho}_{u_+\alpha,w}\hat{\sigma}_u\right)}{N\hat{\sigma}_y^2}+\frac{n\bar{M}^2}{N^2\hat{\sigma}_y^2}\sum_{i=1}^{N}\frac{w_i\hat{S}_{U_{ui}}^2}{m_i} \quad . \qquad \text{(A.115)}$$

When there are no correlations, no calibration (i.e., no auxiliary information in $\dot{\mathbf{x}}$) and no cluster sampling, i.e., $\bar{U} \approx \bar{Y}, \sigma_{u_+}^2 \approx \sigma_y^2$, and when N is large, $\bar{M}\alpha = \alpha \approx \bar{Y}$, we have

$$deff_{C*} \approx \frac{n\bar{W}}{N}\left(\frac{\sigma_y^2}{\sigma_y^2}\right)+\left(\frac{n\bar{W}}{N}-1\right)\frac{\left(\bar{Y}-\bar{Y}\right)^2}{\sigma_y^2}$$
$$=\frac{n\bar{W}}{N} \quad , \qquad \text{(A.116)}$$

which we estimate with Kish's measure.

References

Andrews, D. W. K. (1988), "Laws of large numbers for dependent non-identically distributed random variables," *Economic Theory*, **4**, 458-467.

Andrews, D. W. K. (1992), "Generic uniform convergence," *Economic Theory*, **8**, 183-216.

Arcones, M.A., and Yu, B. (1994), "Central limit theorems for empirical and U-processes of stationary mixing sequences," *Journal of Theoretical Probability*, **7**, 47-71.

Bardsley, P., and Chambers, R.L. (1984), "Multipurpose estimation from unbalanced samples," *Applied Statistics,* **33**, 290-299.

Beaumont, J.P. (2008), "A new approach to weighting and inference in sample surveys," *Biometrika*, **95** (3), 539-553.

Beaumont, J.P., and Alavi, A. A. (2004), "Robust Generalized Regression Estimation," *Survey Methodology*, **30** (2), 195-208.

Beaumont, J.P., and Rivest, L. P. (2009), "Dealing with outliers in survey data," in D. Pfeffermann and C. R. Rao (Eds.), *Handbook of Statistics*, *Sample Surveys: Design, Methods and Application,* **29A**, Amsterdam: Elsevier BV.

Bethlehem, J. G. (2002), "Weighting nonresponse adjustments based on auxiliary information," in R. M. Groves,, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse*. New York: Wiley, 275-287.

Bethlehem, J.G., and Keller, W. J. (1987), "Linear Weighting of Sample Survey Data", *Journal of Official Statistics*, **3**, 141-153.

Breidt, F.J., Claeskens, G., and Opsomer, J.D. (2005), "Model-assisted estimation for complex surveys using penalized splines," *Biometrika,* **92**, 831-846.

Breidt, F.J., and Opsomer, J.D. (2000), "Local polynomial regression estimators in survey sampling," *The Annals of Statistics*, **28**, 1026–1053.

Brick, M., and Montaquila, J. (2009), "Nonresponse," in D. Pfeffermann and C. R. Rao (Eds.), *Handbook of Statistics*, *Sample Surveys: Design, Methods and Application,* **29A**, Amsterdam: Elsevier BV.

Bureau of Transportation Statistics (2002), National Transportation Availability and Use Survey Public-Use File Weight Adjustments, available at: http://www.bts.gov/programs/omnibus_surveys/targeted_survey/2002_national_tr ansportation_availability_and_use_survey/public_use_data_file/html/weighting_and_calculating_variance_estimates.html.

Chambers, R.L. (1986), “Outlier Robust Finite Population Estimation,” *Journal of the American Statistical Association*, **81**, 1063-1069.

Chambers, R. L., Dorfman, A.H., and Wehrly, T.E. (1993), “Bias Robust Estimation in Finite Populations Using Nonparametric Calibration,” *Journal of the American Statistical Association*, **88**, 260-269.

Chambers, R. L. (1996), “Robust Case-Weighting for Multipurpose Establishment Surveys,” *Journal of Official Statistics*, **12** (1), 3-32.

Chen, Q., Elliott, M. R., and Little, R. J. A., (2010), “Bayesian penalized spline model-based inference for finite population proportion in unequal probability sampling,” *Survey Methodology*, **36**, 23-34.

Chowdhury, S., Khare, M., and Wolter, K. (2007), “Weight Trimming in the National Immunization Survey,” *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods, American Statistical Association*.

Čížek, P. (2004), “Asymptotics of Least Trimmed Squares Regression,” Discussion Paper, Tilburg University, Center for Economic Research, No. 2004-72.

Claeskens, G., Krivobokova, T., and Opsomer, J.D. (2009), “Asymptotic properties of penalized spline estimators,” *Biometrika*, **96**, 529-544

Cochran, R. S. (1967), “The Estimation of Domain Sizes When Frames are Interlocking,” *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods, American Statistical Association*, 332-335

Cook, D. R. (1977), "Detection of Influential Observations in Linear Regression". *Technometrics* (American Statistical Association), **19** (1), 15–18.

Cook, D. R. (1979), "Influential Observations in Linear Regression". *Journal of the American Statistical Association* (American Statistical Association), **74** (365), 169–174.

Cox, B. G., and McGrath, D. S. (1981), “An examination of the effect of sample weight truncation on the mean square error of estimates,” presented at Biometrics Society ENAR meeting, Richmond VA.

Crainiceanu, C., Ruppert, D., and Wand, M. P. (2005), “Bayesian Analysis for Penalized Spline Regression Using WinBUGS, *Journal of Statistical Software*, **14** (14).

Darling, D. A. (1957), "The Kolmogorov-Smirnov, Cramer-von MisesTests," *Annals of Mathematical Statistics,* **28**, 823-838.

Davies, P.L. (1993), "Aspects of robust linear regression", *Annals of Statistics*, **21**, 1843-1899.

Denison, D., Mallick, B., and Smith, A. (1998), "Automatic Bayesian curve fitting," *Journal of the Royal Statistical Society*, Series B, **60**, 333-350.

Dixon, W. J. and Massey, F. J. (1966), *An Introduction to Statistical Analysis*, McGraw-Hill Book Co.

Donoho D.L. (1982), "Breakdown properties of multivariate location estimators," Ph.D. qualifying paper, Harvard University.

Donoho, D. L., and Huber, P. J. (1983), "The notion of break-down." In *A. Festschrift for Erich L. Lehmann* (P.J. Bickel, K.A. Doksum, and J. L. Hodges, Jr., eds.), Wadsworth, Belmont CA, 157-184.

Dorfman, A.H. (2000), "Non-parametric Regression for Estimation of Totals" *Section on Survey Research Methods, American Statistical Association*, 47-54.

Dorfman, A. H., and Hall, P. H. (1993), "Estimators of the finite population distribution function using nonparametric regression," *Annals of Statistics*, **21**(3), 1452-1475.

Edelsbrunner, H., and Souvaine, D. L. (1990), "Computing median-of-squares regression lines and guided topological sweep," *Journal of the American Statistical Association*, **85**, 115–119.

Efron, B. (1982), "The Jackknife, the Bootstrap, and other resampling plans," CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia: Society for Industrial and Applied Mathematics (SIAM), 1982.

Eilers, P.H.C., and Marx, B.D. (1996), Flexible smoothing using B-splines and penalized likelihood (with Comments and Rejoinder). *Statistical Science* **11** (2), 89-121.

Elliott, M.R., and Little, R. J. A. (2000), "Model-based Alternatives to Trimming Survey Weights," *Journal of Official Statistics*, **16**, 191-209.

Elliott, M.R. (2007), "Bayesian Weight Trimming for Generalized Linear Regression Models," *Survey Methodology*, **33**, 23-34.

Elliott, M.R. (2008), "Model Averaging Methods for Weight Trimming," *Journal of Official Statistics*, **24**, 517-540.

Elliott, M.R. (2009), "Model Averaging Methods for Weight Trimming in Generalized Linear Regression Models," *Journal of Official Statistics*, **25**, 1-20.

Ericson, W.A. (1969), "Subjective Bayesian Models in Sampling Finite Populations," *Journal of the Royal Statistical Society, Series B*, **31**, 195-234.

Ericson, W.A. (1988), "Bayesian Inference in Finite Populations," in *Handbook of Statistics*, **6**, P.R. Krishnaiah and C. R. Rao (Eds.), Amsterdam: North-Holland, 213-246.

Firth, D. and Bennett, K.E. (1998), "Robust models in probability sampling," *Journal of the Royal Statistical Society,* Series B, **60**, 3-21.

Friedman, J.H. (1991), "Multivariate adaptive regression splines, with discussion," *The Annals of Statistics*, **19**, 1–67.

Friedman, J. H., and Silverman, B. W. (1989), "Flexible parsimonious smoothing and additive modeling (with discussion)," *Technometrics*, **31**, 3–21.

Fuller, W.A. (1991), "Sample Estimators for the Mean of Skewed Populations," *Statistica Sinica*, **1**, 137-158.

Gabler, S., Hader, S., and Lahiri, P. (1999), "A model based justification of Kish's formula of design effects for weighting and clustering," *Survey Methodology*, **25**, 105-116.

Gelfand, I. M., and Vilenkin, N. Y. (1964), *Generalized Functions, Vol. 4, Applications of Harmonic Analysis*, Academic Press, San Diego.

Gelman, A. (2006), "Prior distributions for variance parameters in hierarchical models" *Bayesian Analysis*, **3**, 515-533.

Ghosh, M., and Meeden, G. (1997), *Bayesian Methods for Finite Population Sampling*, London: Chapman & Hall.

Green, P.J. (1995), "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination," *Biometrika*, **82**, 711-732.

Griffin, R. A. (1995), "Dealing with Wide Weight Variation in Polls," *Proceedings of the Social Statistics Section*, American Statistical Association, 908-911.

Hampel, F. R. (1971), "A general qualitative definition of robustness," *Annals of Mathematical Statistics*, **42**, 1887-1896.

Hampel, F. R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986, 2005), *Robust Statistics: The Approach Based on Influence Functions*. New York: Wiley & Sons.

Hartley, H. O. (1962), "Multiple Frame Surveys," *Proceedings of the Social Statistics Section*, American Statistical Association, 203-205.

Harville, D.A., (1977), "Maximum likelihood approaches to variance components estimation and to related problems, *Journal of the American Statistical Association*, **72**, 320–340.

Hoel, P. G. (1962), *Introduction to Statistics (Third Edition)*. New York: John Wiley & Sons.

Horn, S. D., Horn, R. A., and Duncan, D. B. (1975), "Estimating heteroscedastic variances in linear models," *Journal of the American Statistical Association*, **70**, 380-385.

Horvitz, D., and Thompson, D. (1952), **"**A Generalisation of Sampling without Replacement from a Finite Universe," *Journal of the American Statistical Association,* **47***,* 663-685.

Huber, P. J. (1981, 2004), *Robust Statistics*, New York: Wiley & Sons.

Huang, Elizabeth T., and Fuller, Wayne A. (1978), "Nonnegative Regression Estimation For Sample Survey Data," *Proceedings of the Social Statistics Section, American Statistical Association,* 300-305.

Isaki, C. T., and Fuller, W. A. (1982), "Survey Design Under the Regression Superpopulation," *Journal of the American Statistical Association*, **77**, 89-96.

Johnson, E.G., *et al*. (1987), "Weighting procedures, in Implementing the new design: The NAEP 1983-1984 technical report," by A. E. Beaton. Princeton, N.J.: *National Assessment of Educational Progress*, 493-504.

Kalton, G., and Flores-Cervantes, A. (2003), "Weighting Methods," *Journal of Official Statistics*, **19** (2), 81-97

Kish, L. (1965), *Survey Sampling*, New York: John Wiley & Sons.

Kish, L. (1987), "Weighting in $deft^2$," *The Survey Statistician*, **17** (1), 26-30.

Kish, L. (1990), "Weighting: Why, When, and How?" *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods,* American Statistical Association, 121-129.

Kish, L. (1992), "Weighting for unequal Pi," *Journal of Official Statistics,* **8**, 183-200.

Kish, L. (1995), "Methods for design effects," *Journal of Official Statistics,* **11** (1), 55-77.

Kolmogorov, A. (1933), “Sulla determinazione empirica di una legge di distributzione,” *G. Inst. Ilat. Attuari*, **4**, 83.

Kott, P. (2009), “Calibration weighting: combining probability samples and linear prediction models,” in D. Pfeffermann and C. R. Rao (Eds.), *Handbook of Statistics*, *Sample Surveys: Design, Methods and Application,* **29B**, Amsterdam: Elsevier BV.

Krewski, D. and J.N.K. Rao (1981), “Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods,” *The Annals of Statistics*, **19**, 1010-1019.

Krivobokova, T., Crainiceanu, C.M., and Kauermann, G. (2008), “Fast Adaptive Penalized Splines,” *Journal of Computational and Graphical Statistics*, **17** (1), 1-20.

Kuo, L. (1988), “Classical and Prediction Approaches to Estimating Distribution Functions From Survey Data,” in *Proceedings of the Survey Research Methods Section*, American Statistical Association, 280–285.

Kuk, A.Y.C. (1988), “Estimation of distribution functions and medians under sampling with unequal probabilities,” *Biometrika*, **75**, 97-104.

Lazzeroni, L.C., and Little, R. J. A. (1993), “Models for Smoothing Post-stratification Weights,” *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods,* American Statistical Association, 761-769.

Lazzeroni, L.C., and Little, R. J. A., (1998), “Random-Effects Models for Smoothing Post-stratification Weights,” *Journal of Official Statistics,* **14**, 61-78.

Little, R.J.A. (2004), “To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling,” *Journal of the American Statistical Association*, **39**, 546-556.

Liu, B., Ferraro, D., Wilson, E., and Brick, M.J. (2004), “Trimming extreme weights in household surveys,” *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods,* American Statistical Association, 3905-3912.

MacKinnon, J. G., and White, H. (1985), “Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties,” *Journal of Econometrics*, **29**, 305-325.

Mann, H. B., and Whitney, D. R. (1947), “On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other,” *Annals of Mathematics*, **18**, 50-60.

Marsaglia, G., Tsang, W., Wang, J. (2003), “Evaluating Kolmogorov’s Distribution,” *Journal of Statistical Software*, **8** (1), 1-4.

Miller, R.G. (1974), “The jackknife – a review,” *Biometrik*a, **61**, 1-15.

Milliken, G.A., and Johnson, D.E. (1992), *Analysis of Messy Data, Volume I: Designed Experiments*. London: Chapman & Hall.

National Center for Education Statistics, (2008), “NAPE Weighting Procedures: 2003 Weighting Procedures and Variance Estimation,” available at http://nces.ed.gov/nationsreportcard/tdw/weighting/2002_2003/weighting_2003_studtrim.asp

NHANES 2009a. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data. Hyattsville, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2009, http://www.cdc.gov/nchs/nhanes.htm.

NHANES 2009b. Centers for Disease Control and Prevention (CDC). National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey 2005-2006 CPS Control Totals, MD: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2008, http://www.cdc.gov/nchs/data/nhanes/cps_totals_0506.pdf.

Park, I., and Lee, H., (2004), “Design Effects for the weight mean and total estimators under complex survey sampling,” *Survey Methodology*, **30**, 183-193.

Park, I. (2004), “Assessing complex sample designs via design effect decompositions,” *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods,* American Statistical Association, 4135-4142.

Patterson, H.D., and Thompson, R. (1971), "Recovery of inter-block information when block sizes are unequal," *Biometrika*, **58** (3), 545–554.

Pedlow, S., Porras, J., O.Muircheartaigh, C., and Shin, H. (2003), “Outlier Weight Adjustment in Reach 2010,” *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods,* American Statistical Association, 3228-3233.

Pfeffermann, D. (1993), “The Role of Sampling Weights when Modeling Survey Data,” International Statistical Review, **61**, 317-337.

Pfeffermann, D., and Rao, C. R. (Eds.), *Handbook of Statistics*, *Sample Surveys: Design, Methods and Application,* **29A-B**, Amsterdam: Elsevier BV.

Pollard, D. E., (2002), “A user’s guide to measure theory probability,” Cambridge, UK: Cambridge University Press.

Potter, F.A. (1988), “Survey of Procedures to Control Extreme Sampling Weights,” *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 453-458.

Potter, F. A. (1990), “Study of Procedures to Identify and Trim Extreme Sample Weights,” *Proceedings of the Survey Research Methods Section*, American Statistical Association, 225-230.

Quenoulle, M.H. (1949), “Approximate tests of correlation in time series,” Journal of the Royal Statistical Society, Series B, 11, 18-44.

Quenoulle, M.H. (1956), “Notes on bias in estimation,” *Biometrika*, **61**, 353-360.

Rao, C. R., and Kleff, J. (1988), *Estimation of Variance Components and Applications*. Amsterdam: North-Holland.

Rao, J.N.K., and Wu, C.F.J. (1988), “Resampling inference withcomplex survey data,” *Journal of the American Statistical Association*, **83**, 231-241.

Research Triangle Institute (2010), SUDAAN *Language Manual, Release 10.0.* Research Triangle Park, NC: Research Triangle Institute.

Reynolds, P.D., and Curtin, R.T. (2009), *Business Creation in the United States: Initial Explorations with the PSED II Data Set.* New York: Springer.

Ripley, B. D. (1994), “Neural networks and flexible regression and discrimination,” *Journal of Applied Statistics,* **21**, 39-57S.

Rousseeuw, P. J. (1984), “Least median of squares regression. *Journal of the American Statistical Association*, **79** (288), 871-880.

Rousseeuw, P. J. (1997), “Introduction to positive breakdown methods,” in G. S. Maddala and C. R. Rao (Eds.), *Handbook of Statistics*, **15**: Robust Inference, Elsevier, Amsterdam, 101-121.

Rousseeuw P. J., and Leroy, A. M. (1987, 2003), *Robust Regression and Outlier Detection*, John Wiley & Sons, New York.

Rousseeuw, P. J., and Van Driesen, K. (1999), "A Fast Algorithm for the Minimum Covariance Determinant Estimator," *Technometrics*, **41**, 212-223.

Royall, R. M. (1976), “The Linear Least-squares Prediction Approach to Two-stage Sampling,” *Journal of the American Statistical Association,* **71**, 657–664.

Rubin, D. B. (1983), “Inference and Missing Data,” *Biometrika*, **53**, 581-592.

Rubin, D. B. (1987), Multiple *Imputation for Nonresponse in Surveys*, New York: Wiley.

Ruymgaart, F.H. (1980), “A unified approach to the asymptotic distribution theory of certain midrank statistics,” In *Statistique non Parametrique Asymptotique*, J.P. Raoult (Ed.), Lecture Notes on Mathematics, No. 821, Springer, Berlin.

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

Ruppert D., and Carroll, R. (2000), “Spatially-adaptive penalties for spline fitting,” *Australian and New Zealand Journal of Statistics*, **42**, 205–223.

Ruppert, D. (2002), “Selecting the number of knots for penalized splines,” *Journal of Computational and Graphical Statistics*, **11**, 735–737.

Rust, K. F. (1985), “Variance estimation for complex estimators in sample surveys,” *Journal of Official Statistics,* **1***,* 381–397.

Salibian-Barrera, M. (2006), “The asymptotics of MM estimators for linear regression with fixed designs,” *Metrika*, **63**, 283-294.

Särndal, C. E., and Lundström, S. (2005), *Estimation in Surveys with Nonresponse*, New York: John Wiley and Sons.

Särndal, C.E., Swensson, B. and Wretman, J. (1992), *Model Assisted Survey Sampling*, Springer: Berlin, New York.

Scholz, F.W., and Stephens, M.A. (1987), “K-sample Anderson-Darling Tests,” *Journal of the American Statistical Association*, **82**, 918–924.

Scott, A. J. (1977), “Large-Sample Posterior Distributions for Finite Populations,” *The Annals of Mathematics*, **42**, 1113-1117.

Searle, S.R., Casella,G., and McCulloch, C.E. (1992), *Variance Components*, Wiley, New York.

Shao, J., and Wu, C.F.J. (1989), “A general theory for jackknife variance estimation,” *Annals of Statistics*, **17**, 1176-1197.

Sharot, T. (1986), “Weighting Survey Results,” *Journal of Market Research Society*, 269-284.

Singh, A.C., and Mohl, C.A. (1996), “Understanding Calibration Estimators in Survey Sampling.” *Survey Methodology*, **22**, 107-115.

Smirnov, N. V. (1948), “Tables for estimating the goodness of fit of empirical distributions,” *Annals of Mathematics*, **19**, 279.

Smyth, G. K., and Hawkins, D. M. (2000), “Robust frequency estimation using elemental sets,” *Journal of Computational and Graphical Statistics*, **9**, 196-214.

Spencer, B. D. (2000), “An Approximate Design Effect for Unequal Weighting When Measurements May Correlate With Selection Probabilities,” *Survey Methodology*, **26**, 137-138.

Statistics of Income (2011), “2007 Charities & Tax-Exempt Microdata Files,” available at: http://www.irs.gov/taxstats/article/0,,id=226223,00.html.

Stephens, M. A. (1979), “Test of fit for the logistic distribution based on the empirical distribution function, *Biometrika*, **66**, (3), 591-595.

Stephens, M. A. (1986), “Tests Bases on EDF Statistics,” In D’Agnostino, R. B., and Stephens, M.A., *Goodness-of-Fit Techniques*, New York: Marcel Dekker.

Stone, C.J., Hansen, M., Kooperberg, C., Truong, Y.K. (1997), “Polynomial splines and their tensor products in extended linear modeling” (with discussion). Annals of Statistics, **25**, 1371–1470.

Stromberg, A. J. (1993), “Computation of high breakdown nonlinear regression parameters,” *Journal of the American Statistical Association*, **88**, 237-244.

Sverchkov, M., and Pfeffermann, D. (2004), “Prediction of finite population totals based on the sample distribution,” *Survey Methodology*, **30**, 79–92.

Tatsuoka, K. S. and Tyler, D. E. (2000), “The uniqueness of S and M-functionals under non-elliptical distributions,” *The Annals of Statistics*, **28**, 1219-1243.

Thas, O., and Rayner, J. C. W. (2005), “Smooth tests for the zero-inflated Poisson distribution,” *Biometrika*, **61**, 808-815.

Valliant, R. (2002), “Variance Estimation for the general regression estimator,” *Survey Methodology*, **28**, 103-114.

Valliant, R. (2004), “The effect of multiple weight adjustments on variance estimation,” *Journal of Official Statistics*, **20** (1), 1-18.

Valliant, R. (2009), “Model-Based Prediction of Finite Population Totals,” in D. Pfeffermann and C. R. Rao (Eds.), *Handbook of Statistics*, *Sample Surveys: Design, Methods and Application,* **29B**, Amsterdam: Elsevier BV.

Valliant, R., Dorfman, A., and Royall, R. M. (2000), *Finite Population Sampling and Inference*, New York: Wiley & Sons.

Valliant, R., Brick, M.J., and Dever, J. (2008), "Weight Adjustments for the Grouped Jackknife Variance Estimator." *Journal of Official Statistics*, **24** (3), 469-488.

Vaart, A. W. van der (2000), *Asymptotic Statistics (Cambridge Series in Statistical and Probabilistic Mathematics)*. Cambridge, UK: Cambridge University Press.

Wand, M. P. (2003), "Smoothing and mixed models," *Computational Statistics*, **18**, 223-249.

White, H. (1980), "A heterskedasticity-consistent covariance matrix estimator and a direct test for heterskedasticity," *Econometrica*, **48**, 817-838.

Wolter, K.M. (1985; 2007), *Introduction to Variance Estimation*. Springer-Verlag, New York.

Yang, G. L., Le C., Lucien, M. (2000), *Asymptotics in Statistics: Some Basic Concepts*. Berlin: Springer.

Yohai, V.J. (1985), "High breakdown-point and high efficiency robust estimates for regression," *Technical Report No. 66*, Department of Statistics, University of Washington, Seattle.

Yohai, V. J. (1987), "High breakdown-point and high efficiency robust estimates for regression," *The Annals of Statistics*, **15**, 642-656.

Yu, B. (1994), "Rates of convergence for empirical processes of stationary mixing sequences," *The Annals of Probability*, **22** (1), 1-8.

Zaslavsky, A.M., Schenker, N., and Belin, T.R. (2001), "Downweighting influential clusters in surveys: application to the 1990 post enumeration survey," *Journal of the American Statistical Association*, **96**, 858-869.

Zinde-Walsh, V. (2002), "Asymptotic theory for some High Breakdown Point Estimators," *Working Paper*, Department of Economics, McGill University.

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

Zheng, H., and Little, R.J.A. (2005), "Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model," *Journal of Official Statistics*, **21**, 1-20.