Misspecified sampling weights in weight-smoothing methods

Xia Li[*] Eric V. Slud [†]

Abstract

Modifications of weights due to calibration, trimming, sometimes in multiple stages, are very common in survey analysis. It is typical to work with modified as opposed to design/inverse-inclusion-probability weights, especially in publicly released survey data. Various weight-smoothing methods (Pfeffermann and Sverchkov 1999; Zheng and Little 2003; Beaumont 2008) have been proposed to improve the efficiency of the Horvitz-Thompson (HT) and Generalized Regression (GREG) estimators of survey totals. These methods depend on correctness of model relationships between the survey attribute Y, covariate X, and Y and the given weights w. There has been little systematic study about the impact of treating modified weights as design weights, when the model assumptions connecting x, y, w might also be misspecified. It is, therefore, important to evaluate the performances of these three methods under different modified weights and misspecified models. In this simulation study, we generate finite frame populations from superpopulation models, simulate misspecified models, and quantify mis-calibrations of the weights to compare GREG and HT results with estimators based on the three weight-smoothing methods.

Keywords: sampling theory, superpopulation models, design-based estimates, model-based estimates.

1 Introduction

We consider probability sampling designs where a random sample S with elements Y_i is drawn from the finite population according to the inclusion probabilities π_i, $i = 1, \ldots, N$. The main interest is to estimate the population total of the outcome variable Y, defined as $t_Y = \sum_{i=1}^{N} Y_i$. The HT estimator is a design-unbiased estimator of t_Y with unequal probabilities of inclusion π_i, defined as

$$\hat{t}_Y^{HT} = \sum_{i \in S} \frac{Y_i}{\pi_i}. \tag{1}$$

This universal unbiased property does not depend on any correctness of distributional model assumption of survey measurement Y_i and covariates X_i. When the outcome variable and the inclusion probability are weakly related, the HT estimator could be very inefficient. Generalized regression estimator (GREG), given by Eq. (2) is also design-consistent utilizing the association between covariate X_i and outcome Y_i when the total $\sum_{i=1}^{N} X_i$ is known. Here $\hat{Y}_i$ is based on a simple weighted linear regression of Y_i on covariates.

$$\hat{t}_Y^{GREG} = t'_{\mathbf{X}} \hat{\beta}^{PS}, \quad \mathbf{X}_i = (1, X'_i)'. \tag{2}$$

A concern that often is ignored is that the survey professionals often do modify the sample weights in multiple stages. Common modification procedures include calibration, trimming,

[*] The Applied Mathematics & Statistics, and Scientific Computation Programj, University of Maryland, College Park, USA

[†] Department of Mathematics, University of Maryland, College Park, USA

etc. Before any smoothing method is applied, the weights $\{d_i\}_{i=1}^N$ have been modified through calibration already, sometimes in multiple steps. When releasing the final data sets to the public and making microdata available to users (e.g. applied statisticians and epidemiologists), all the black-box style of massaging the sampling weights are masked to public users and we all treat the modified weights as if they were the true sampling weights. Misspecification happens when we calibrate on wrong totals or on wrong covariates. It will be different from (Ybarra and Lohr 2008) who discussed measurement error in auxiliary information in small area estimation models.

1.1 Brief Review of Weight-Smoothing Methods Considered in Simulation

Different smoothing methods have been proposed to improve the efficiency by modifying the survey weights (Pfeffermann and Sverchkov 1999; Zheng and Little 2003; Beaumont 2008). Pfeffermann and Sverchkov (1999) considered smoothing the weights by a function of covariates and then applying the weighted least squares. We consider only the semi-parametric estimation, denoted by $\hat{t}_Y^{PS} = t_{\mathbf{X}}\hat{\beta}^{PS}$. The estimated coefficient vector has the form

$$\begin{gathered}\hat{\beta}^{PS} = \left(\sum_{i\in S} q_i \mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\sum_{i\in S} q_i\mathbf{X}_i Y_i\right) \\ \text{where } q_i = \frac{d_i}{\hat{\mathrm{E}}(d_i|X_i, I_i = 1)}\end{gathered} \tag{3}$$

and $\hat{\mathrm{E}}(d_i|X_i, I_i = 1)$ is obtained by regressing d_i against X_i. Here $d_i = 1/\pi_i$ denotes the working weights.

Zheng and Little (2003) considered smoothing the outcome variable by modeling Y against the inclusion probability using a p-spline function, defined in Eq. (4). One has to decide how delicate the p-spline model is by choosing the degree p, number of knots m, and the exponent k. Usually $k = 0, 1/2$ or 1. Zheng & Little suggested using random-effect terms as coefficients $\beta_{p+1}, \ldots, \beta_{p+m}$. To simplify, fixed effects were considered in this study.

$$\begin{gathered}Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j \pi_i^j + \sum_{l=1}^{m} (\pi_i - \kappa_l)_+^p + \varepsilon_i \\ \text{where } \varepsilon_i \sim \text{ iid } N(0, \pi_i^{2k}\sigma_\varepsilon^2)\end{gathered} \tag{4}$$

Assuming π_i is only known for the sample S, the estimated population total is given by

$$\hat{t}_Y^{ZL1} = \sum_{i\in S} \hat{Y}_i/\pi_i. \tag{5}$$

If π_i are known for the whole population, the estimated population total could be given by

$$\hat{t}_Y^{ZL2} = \sum_{i\in S} Y_i + \sum_{i\notin S} \hat{Y}_i/\pi_i. \tag{6}$$

In both Eq. (5) and (6), $\hat{Y}_i = \hat{\mathrm{E}}(Y_i|\pi_i)$ based on Eq. (4).

Beaumont (2008) dealt with the inefficiency of the HT estimator by smoothing weights against the outcome variable. Estimated smoothed weights $\hat{d}_i$ could be obtained by regressing $\log(d_i - 1)$ on vector h_i which is a known function of Y_i. Then the smoothed estimator of population total could be given by

$$\hat{t}_Y^{B} = \sum_{i\in S} \hat{d}_i Y_i. \tag{7}$$

2 Probabilistic Models of Weight Misspecification

From now on let d_i be the design weight or original weight for unit i in the finite population, which is the weight before any modification. Then $\pi_i^0 = P(i \in S) = 1/d_i$ is the inclusion probability. Let w_i be the weight for unit i after some modification procedures including weight trimming, calibration, etc. Let $\pi_i^F = 1/w_i$ represent the reciprocal of the modified weight. In general,

$$d_i \neq w_i, \tag{8}$$

or equivalently $\pi_i^F \neq \pi_i^0$. One might think that π_i^F could be viewed as a function of X_i and thus can be written as $P(I_i = 1|X_i)$. Then if the propensity model were correct and known, this conditional propensity factor would cancel out of the expression for the expected weighted total and still the resulting estimator would be design unbiased. But if this conditional probability is misspecified, then w_i and π_i^0 would not cancel out when taking conditional expectation of the HT estimator. To develop a simulation structure and study the possible consequences of (8), we introduce first, as a misspecification model in Section 2.1, a random variable η accounting for the random modification processes, satisfying that

$$w_i = d_i\eta_i, \text{ where}\, \eta_i \sim iid\, F_\eta,\ \mathrm{E}(\eta_i) = 1$$

so that we still have unbiased HT estimator

$$\mathrm{E}\left(\hat{t}_Y^{HT}\right) = \mathrm{E}_p\left(\mathrm{E}_d\left(\sum_{i=1}^{N} I_i w_i Y_i \,|\, \mathcal{F}\right)\right) = \sum_{i=1}^{N} \mathrm{E}_p(Y_i) \tag{9}$$

where $\mathcal{F}$ is defined as the entire finite population of $\{Y_1, \ldots, Y_N\}$ of attributes, $\mathrm{E}_d(\cdot)$ denotes the average over all samples possible under the design for the finite population $\mathcal{F}$ while $\mathrm{E}_p(\cdot)$ denotes the expectation with respect to the superpopulation model, following Isaki & Fuller's notation (Isaki and Fuller 1982).

2.1 Three Misspecification Models

There are three models of misspecification that are considered in this study. The first is that the misspecification is purely random, independent of (Y_i, X_i, X_i^*, d_i), where X_i^* are additional covariates that will be introduced below as observable surrogates for an unobservable X_i. This represents the case when the modification procedures mainly bring in purely random noise which does not depend on any covariate or outcome variable. In this case, the anticipated variance of HT using misspecified weights will be inflated by a multiplied factor $1 + \mathrm{V}(\eta)$. It is expected that the more noisy the modification process is, the more the variance of HT would be enlarged.

$$\mathrm{E}\left[\mathrm{V}\left(\hat{t}_Y^{HT}|\mathcal{F}\right)\right] = \mathrm{E}\left[\mathrm{V}\left(\sum_{i=1}^{N} I_i d_i \eta_i Y_i \Big| \mathcal{F}\right)\right] = \left(\sum_{i=1}^{N}(d_i - 1)\,\mathrm{E}(Y_i^2)\right) \cdot (1 + \mathrm{V}(\eta_i)) \tag{10}$$

The second model is that the misspecification is related to the covariates X_i through

$$\eta_i = \exp\{b(X_i) + a(X_i)\zeta_i\} \tag{11}$$

where ζ_i iid $\sim F_\zeta$ is independent of (Y_i, X_i, X_i^*, d_i) satisfying

$$\exp\{b(X_i)\} = \{m_\zeta(a(X_i))\}^{-1}.$$

Here $m_\zeta(\cdot)$ denotes the moment generating function of ζ_i. The misspecification model (11) still results in unbiased HT estimates, since $E(\eta_i \,|\, X_i, Y_i) = 1$.

The third model arises in any setting where conditional inclusion probabilities (propensities) $P(i \in S \,|\, X_i)$ models could be correctly specified, but X_i is for some reason not properly observable and is replaced by a surrogate (a "wrong covariate") X_i^* which is conditionally independent of Y_i given X_i. In this study, we used (11) and replaced X_i^* with X_i.

2.2 Quantifying Misspecification

In the literature, there are several choices to measure the distances of two sets of sampling weights (Deville and Särndal 1992; Deville et al. 1993). There are also other natural metrics to measure the magnitude of modification such as $\sum_{i=1}^{N} |d_i - w_i|$. In single modification stage, a practitioner may avoid large changes in single unit weights by requiring that the proportion of the relative changes exceeding certain value among sample would be bounded, that is

$$\frac{1}{|S|}\sum_{i\in S}\mathbf{I}\left\{\left|\frac{w_i - d_i}{d_i}\right| \geq q\right\} \leq K.$$

In this study, $q = .2$ and $K = .5$ will be considered as well-controlled modifications, indicating that no more than half of the units would have relative weight changes exceeding 20%. $q = K = .5$ would be considered as a mild level of modification, with no more than half of the relative weight changes exceeding 50%.

3 Simulation Study

3.1 Superpopulation Parameters

In this simulation study, we are interested in a scalar outcome variable denoted by Y. Along with Y, we will also collect some covariates X and X^*. To simplify, let the dimension of X and X^* to be 1 in this study. $\{(Y_i, X_i, X_i^*, \pi_i^0)\}_{i=1}^N$ are independent with identical distribution $\mathcal{G}$, where N is the finite population size. The finite population, denoted by $\mathcal{F}$, has five strata with equal stratum sizes, where the stratum label for unit i is denoted by $L_i \in \{1, 2, 3, 4, 5\}$.

We assume only X_i plays a role in Y_i, that is, Y_i is conditionally independent of all other covariates and π_i^0 given X_i. Let β_k represent the coefficient of X_i among the k^{th} stratum. The conditional expectation of Y_i given X_i in stratum k is given by

$$\mathrm{E}(Y_i|X_i) = \beta_0 + \beta_k X_i, \quad \text{if} \quad L_i = k. \tag{12}$$

That is, X_i is the covariate with predictive value for Y_i, with Y_i and X_i^* conditionally independent given X_i. One possible situation is that X_i^* is an observable covariate related to X_i but related to Y_i only through X_i. One might calibrate on X_i^*, the "wrong" variable by mistake, where the sample values of X_i^* and $\sum_{i=1}^N X_i^*$ are also known. However, this surrogate variable X_i^* is not the right predictive variable for a properly specified regression (12). Jointly, the vector (X_i, X_i^*) is assumed here to follow

$$\begin{pmatrix} X_i \\ X_i^* \end{pmatrix} = .5 + \begin{pmatrix} |\tilde{X}_i|^\gamma \\ |\tilde{X}_i^*|^\gamma \end{pmatrix}, \text{ where } \begin{pmatrix} \tilde{X}_i \\ \tilde{X}_i^* \end{pmatrix} \sim \mathcal{N}_2\left(\mu_k, \Sigma_k\right), \; L_i = k \tag{13}$$

Since β_k varies across different strata, so the linear relationship of Y and X varies by stratum. Three different sets of values of β_k are used to specify different frame populations, as shown in Fig. 1. Under finite population 1, Y and X were weakly related in population level; under finite population 2, Y and X were linearly correlated overall; under finite population 3, Y displayed a quadratic dependence on X, therefore a linear model on X only would be misspecified.

Inclusion probability π_i^0 for unit i in stratum k is proportional to the size variable V_i where $V_i = plogis(\beta_V + \beta_{VZ} Z_i)$, $Z_i \sim N(\beta_Z + \beta_{ZX} X_i + \beta_{ZY} Y_i, \sigma_Z^2)$. One set of values of $(\beta_V, \beta_{VZ}, \beta_Z, \beta_{ZX}, \beta_{ZY}, \sigma_Z^2)$ is used for all three finite populations. But (π^0, Y) and (Y, d) display different relations according to different finite population settings, as shown in Fig. 2 and 3.

For each set of finite frame population parameters, we first generated $B = 25$ finite populations of size $N = 10000$ with values $(Y_i, X_i, X_i^*, d_i, \pi_i^0, w_i^{(1)}, w_i^{(2)}, w_i^{(3)}, w_i^{(4)})$, $i = 1, \ldots, N$, where

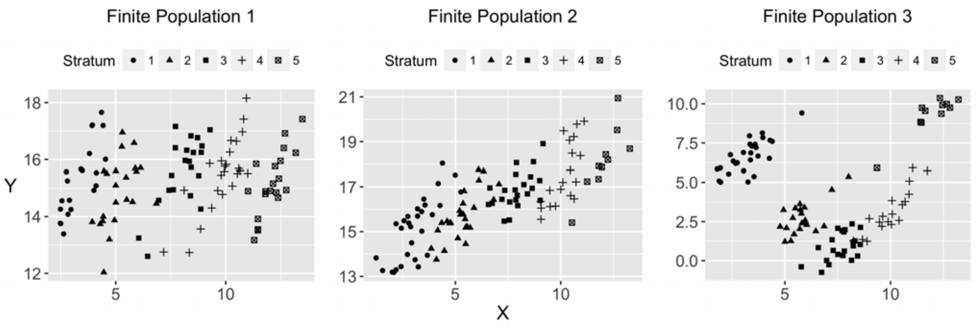

Figure 1: Three simulated populations (N=10,000) X-axis, X_i; Y-axis, Y_i

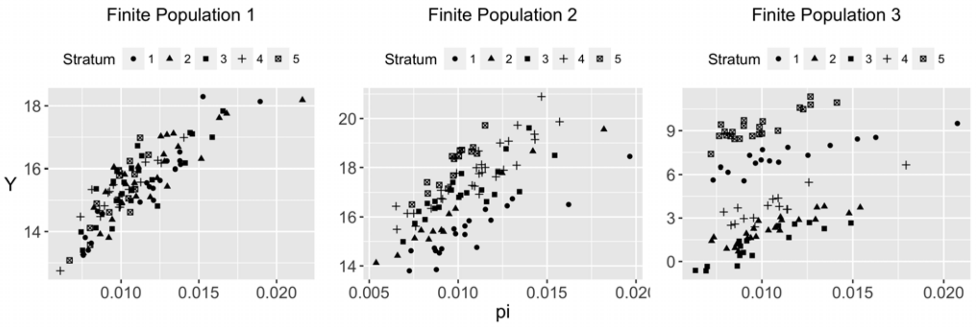

Figure 2: Three simulated populations (N=10,000) X-axis, π_i; Y-axis, Y_i

$w_i^{(k)}$, $k = 1, 2, 3, 4$ represented the misspecified weights. Independent samples were drawn by Poisson sampling, where the sample indicator I_i followed a Bernoulli(π_i^0) distribution. The expected sample size was $n = 100$ where the sample size for each stratum were calculated by using Neyman allocation. For each sample, a set of six estimators were calculated and compared under different misspecification situations as follows

- Analysis using (X_i, d_i) with no misspecification.
- $(X_i, w_i^{(1)})$ with purely random and well-controlled misspecification. Here $\eta_i \sim \text{iid}\, LN(-\tau^2, \tau^2)$, $\tau = \sqrt{\log(1.1)}$.
- $(X_i, w_i^{(2)})$ with purely random misspecification at mild level. Here $\eta_i \sim \text{iid}\, LN(-\tau^2, \tau^2)$, $\tau = \sqrt{\log(2)}$.
- $(X_i, w_i^{(3)})$ with misspecification related to the right covariates. Here $a(u) = (u-15)^2/100$, $\zeta_i \sim \text{iid}\, N(, 1.5^2)$.
- $(X_i^*, w_i^{(4)})$ with misspecification related to the surrogate covariates for X_i. Here again $a(u) = (u-15)^2/100$, $\zeta_i \sim \text{iid}\, N(, 1.5^2)$.

3.2 Estimators considered

- HT, Eq. (1).
- GREG, Eq. (2).
- Semi-parametric estimation proposed by Pfeffermann & Sverchkov, Eq. (3).
- Zheng & Little's method, as described in Eq. (5) and Eq. (6) with $p = 1$ and $k = 0$, using 5 knots.

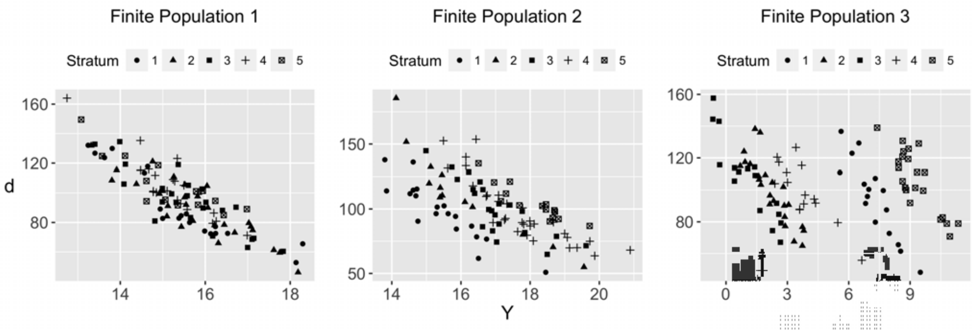

Figure 3: Three simulated populations (N=10,000) X-axis, Y_i; Y-axis, d_i

- Beaumont's method, Eq. (7) with $h_i = (1, Y_i, Y_i^2, Y_i^3)'$.

3.3 Simulation Results

Table 1: Empirical average of log of ratio of MSE with weights w_i over MSE with d_i or $N = 10,000$, $n = 100$ under purely random and strictly controlled misspecifications. HT: Horvitz-Thompson; ZL1: Zheng & Little's method assuming π_i are only known for sample S; ZL2: Zheng & Little's method assuming π_i are known for the entire population; B: Beaumont's method; PS: semiparametric method proposed by Pfeffermann & Sverchkov; GREG: generalized regression estimate

Finite Population	HT	ZL1	ZL2	B	PS	GREG
1	0.09	0.09	2.09	0.09	0.09	0.09
2	0.10	0.10	0.87	0.10	0.09	0.09
3	0.10	0.10	0.16	0.09	0.09	0.09

Summaries comparing MSE's are reported in Table 1 under purely random and well controlled misspecifications. Here all the entries in the table are the log-scaled ratio of MSE under modified weights w_i to that under design weights d_i. From the discussion before we know that the HT estimator remains unbiased. If analyzing based on w_i makes no difference from using d_i in terms of variance, then this value should be close to 0. From Eq. (10) we know that the more noisy η_i is, the larger the log-scaled ratio of MSE will be. Since Table 1 corresponds to well-controlled situation where no more than half of weight changes exceeded 20%, $\mathbb{V}(\eta)$ was expected to be small. The simulation results showed similar results and all the numbers except ZL2 were close to 0. ZL2 relies on the assumption that the weights are known for the entire population $\mathcal{F}$. It is not surprising that ZL2 performed the worst in Table 1. This indicates that when the misspecification is purely random and well-controlled, analyzing based on modified weights (except for ZL2, which extrapolates the incorrect weight model to the entire population) might not harm the results much in terms of variance.

Table 2: Empirical average of log of ratio of MSE with weights w_i over MSE with d_i for $N = 10,000$, $n = 100$ under purely random and mild level of misspecifications

Finite Population	HT	ZL1	ZL2	B	PS	GREG
1	0.67	0.67	2.70	0.67	0.61	0.61
2	0.66	0.66	1.25	0.67	0.60	0.60
3	0.66	0.65	0.37	0.61	0.62	0.63

When the misspecifications are strengthened, by Eq. (10) we would expect that MSE would be larger even if the misspecifications were purely random. Table 2 summarizes the empirical average of log MSE ratios under purely random and mild level of misspecification. Comparing it with Table 2, we saw all the methods had increased variances. Across different methods, MSE ratios were about the same except ZL2.

Table 3: Empirical average of log of ratio of MSE with weights w_i over MSE with d_i for $N = 10,000$, $n = 100$ under mild-level misspecifications related to X

Finite Population	HT	ZL1	ZL2	B	PS	GREG
1	1.01	1.01	2.61	0.91	1.02	0.94
2	0.91	0.91	1.09	2.55	1.02	0.95
3	0.87	0.90	0.31	1.52	1.88	1.44

Table 4: Empirical average of $(\text{bias/RMSE})^2$ for $N = 10,000$, $n = 100$ under mild-level misspecifications related to X

Finite Population	HT	ZL1	ZL2	B	PS	GREG
1	0.26	0.26	0.75	0.13	0.11	0.05
2	0.24	0.24	0.65	0.43	0.09	0.05
3	0.26	0.34	0.35	0.21	0.59	0.26

Table 3 again summarizes log MSE ratio, under misspecifications related to covariates X_i at mild level. We observed increased MSE with modified weights w_i compared to MSE with design weights d_i. Since these estimators are all consistent without misspecifications, one could examine the proportion of MSE that could be explained by bias to check the consistency under w_i. From Table 4 we could see that Beaumont's method, HT and methods of Zheng & Little were not consistent under all three finite populations. Pfeffermann & Sverchkov's method and GREG borrowed the relationship of Y and X under finite population 1 and 2 and had much smaller bias. Under finite population 3 when a simple weighted regression of Y on X was misspecified, Pfeffermann & Sverchkov's method and GREG also failed to be unbiased.

Table 5: Empirical average of log of ratio of MSE with weights w_i over MSE with d_i for $N = 10,000$, $n = 100$ under mild-level misspecifications related to X^*

Finite Population	HT	ZL1	ZL2	B	PS	GREG
1	1.03	1.03	2.63	0.93	1.08	0.98
2	0.95	0.95	1.19	2.26	1.35	1.27
3	0.89	0.92	0.27	1.51	1.85	1.37

Table 5 and 6 described a similar story while this time the misspecifications were related to the surrogate covariates X_i.

4 Discussion

This study aims to develop a simulation structure and study the possible consequences of weights misspecification. By introducing different hypothetical probabilistic weight-misspecification models, we were able to compare the mean squared error and/or bias uder modified weights with that under design weights. In summary, we considered three misspecification models. The results summarized in Table 1 to 6 suggest that we may not lose design consistency when weight misspecifications are purely random. If the misspecifications are related to the covariates, the right ones or the surrogate ones, then we might expect systematic inconsistency.

The simulation results implied that defining the "correctness" of sampling weights was necessary. Analysis with modified weights may lead to biases, especially when the outcome model

Table 6: Empirical average of $(\text{bias}/\text{RMSE})^2$ for $N = 10,000$, $n = 100$ under mild-level misspecifications related to X^*

Finite Population	HT	ZL1	ZL2	B	PS	GREG
1	0.28	0.28	0.76	0.12	0.07	0.04
2	0.27	0.27	0.65	0.38	0.02	0.01
3	0.29	0.37	0.31	0.20	0.59	0.26

is misspecified. Another important issue is that theoretical analysis is needed on the impact of misspeficied weights in future studies.

References

Beaumont, J. F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, 95(3):539–553.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. *Journal of the American Statistical Association*, 88(423):1013–1020.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96.

Pfeffermann, D. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā: The Indian Journal of Statistics, Series B*, 61(1):166–186.

Ybarra, L. M. and Lohr, S. L. (2008). Small area estimation when auxiliary information is measured with error. *Biometrika*, 95(4):919–931.

Zheng, H. and Little, R. J. (2003). Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19(2):99–117.