Estimation from Purposive Samples with the Aid of Probability Supplements but without Data on the Study Variable

A.C. Singh[1,2], V. Beresovsky[2], and C. Ye[1]
[1] Survey and Data Sciences, American Institutes for Research, Rockville, MD 20852
[2]Division of Research and Methodology, National Center for Health Statistics, Hyattsville, MD 20782
asingh@air.org, vberesovsky@cdc.gov, cye@air.org

Abstract

Consider estimation of the population total T_y for an outcome or study variable y from a low-budget purposive sample $\tilde{s}$ with the aid of an ongoing high-budget reference probability sample s^* with no data on y but data on common auxiliary variables or covariates x. Using Royall's model-based approach, a prediction estimator can be constructed from $\tilde{s}$ with known totals of x or their estimates from s^* under the postulated assumption of model holding for $\tilde{s}$ and its complementary part; i.e., nonselected units. Using Särndal's design-based approach, GREG (generalized regression) can be constructed from $\tilde{s}$ after estimating the sample inclusion propensities using the calibration approach under the postulated assumption of model holding for $\tilde{s}$ and its complement. By treating the problem as a complete missing data problem for s^*, a new estimator *i*GREG (*i* for imputation) can be constructed from s^*after imputing y for all units in s^* by using $\tilde{s}$ as the donor dataset under a model whose validity can be partially tested using x observed in both samples. Analogous to the quasi-design based approach in probability samples with nonresponse, we start with a design-based approach using the reference sample s^*, but build over it by integrating $y-$information from the purposive sample $\tilde{s}$ under an imputation model. This approach is termed model-over-design (MOD) integration following Singh (2015). The information on the differences between imputed and observed values of x provide extra covariates with the constraints of zero control totals to reduce the imputation bias via weight calibration. Variance estimates for *i*GREG can be obtained by extending results under the reverse framework for nonresponse imputation in probability samples (where the respondent subsample serves as the donor dataset) to the case of complete missingness by design where an external dataset ($\tilde{s}$) serves as the donor dataset. Limited simulation results are presented for illustration.

Key Words: Prediction Estimator; GREG with Estimated Sample Inclusion Propensity; GREG with Imputed Outcome Variable; MAR-type Assumption for Non-selected Units in Purposive Samples

1. Introduction

Nonprobability samples are gaining momentum as an alternative to traditional probability samples due to their efficiencies in cost, time, and their ability to serve specific purposes such as targeting special domains (or subpopulations) of interest, or collecting more detailed information that may not be practical to collect in large scale probability surveys. The goal is to make valid inferences from such samples for a target population. However,

as was concluded in the AAPOR panel task report on nonprobability samples (Baker et al., 2013), there might be ways to come up with seemingly reasonable point estimates, but without a suitable framework for measuring the margin of error (MOE or half-width of the confidence interval), it is difficult to build user confidence in any estimation methodology. The crux of the problem with nonprobability samples is providing reliable MOE for a given method which typically involves variance estimation. This problem is addressed in this paper using some key ideas of variance estimation in the presence of nonresponse for probability survey sampling.

We consider the following formulation of the problem. There are two samples: $\tilde{s}$ (purposive or nonprobability) supplemented with s^*(reference or probability) from populations $\tilde{U}$ and U^* respectively where both have information about common auxiliary variables or covariates x's but information about the study variable y is only collected in $\tilde{s}$ and not in s^*. Without loss of generality, we will assume that the two populations $\tilde{U}$ and $U^*$are identical although $\tilde{U}$, in general, would be a subset of U^* in which case U^* could be replaced by its subset $\tilde{U}$; i.e., the target parameter could be redefined. The sample s^* is a man-made probability sample and so the selection probabilities under the design π^* are known whereas the sample $\tilde{s}$ is a nonprobablity sample which can be termed purposive in that the selection of units satisfying eligibility criteria is based on considerations of convenience for cost and time efficiency rather than a rigorous protocol for sample representativeness. It is, however, assumed that the purposive sample and population distributions of the covariates have a common support; i.e., the $x-$values in $\tilde{s}$ are well dispersed. Conceptually, the sample $\tilde{s}$ can also be deemed as a probability sample with unknown selection probabilities under the nature-made design $\tilde{\pi}$ although, for simplicity, it is referred to as a nonprobability sample to distinguish it from traditional probability samples. Some examples of pairs of s^* and $\tilde{s}$ respectively are surveys on Youth Risk Behavior Surveillance System paired with a purposive sample on other risk behavior, National Health Interview Survey paired with an opt-in internet panel sample on detailed health characteristics, and National Household Education Survey paired with randomized trials on innovative education programs.

It is of interest to estimate the population total T_y of y based on s^* and $\tilde{s}$. There exist methods that rely on rather strong assumptions to make inference from $\tilde{s}$ about U^*such as the prediction method of Royall (1970, 1976) based on a model for the study variable y as a function of x's, or the propensity score model-based method considered by Valliant and Dever (2011) and Elliott (2009) for inference from $\tilde{s}$; see Elliott and Valliant (2017) for a good review. The latter method is also known as propensity score weighting in randomized trials; see Stuart et al. (2011). Here, propensity refers to the probability $\tilde{\pi}_k$ of the event for a unit k from U^* to be selected in the purposive sample $\tilde{s}$, and is typically modeled as a logit function of x's. For the above methods, the supplementary sample s^* is not used directly but could be used indirectly to provide estimated totals T_z^* for other outcome variable z observed in both samples and deemed to be good predictors in the model for the outcome variable or the propensity score; see next section for details. Besides the first order (i.e., the regression mean function) assumptions needed to obtain point estimates, which although untestable may be deemed plausible, above methods also require second order assumptions for variance estimation which are much stronger and highly speculative in nature. Regardless of their plausibility, such assumptions are needed for computing MOE for point estimators.

Clearly, it is best to minimize assumptions that are difficult to test, and therefore, it would be preferable to rely only on first order assumptions in models for the study or outcome

variable or the propensity score as they seem more plausible in any given application. In fact, in surveys with probability samples, either outcome variable model for imputation for missing data or propensity score model for nonresponse adjustment to sampling weights is commonly used under only first order assumptions in addition to the assumption of census nonresponse (i.e., the reverse framework in which for a given survey, observation units at the population level itself can be designated as respondent or nonrespondent regardless of whether they get selected in the sample or not) and small sampling fractions; see Shao and Steel (1999). The resulting estimates are quasi-design based and their properties are derived under the joint design and model random mechanism. The proposed method takes advantage of these key ideas in survey sampling, and generalizes them to the present problem by transforming it to a missing data problem for the supplementary probability sample s^* which does not collect any information about y; i.e., there is complete nonresponse by design, and where the purposive sample $\tilde{s}$ is used as a donor dataset for imputation. However, it is different from the usual quasi-design based approach in the presence of item or unit nonresponse because there is no respondent subsample of s^* to act as a donor dataset. Thus, for the proposed method, instead of making a long leap from $\tilde{s}$ to U^*, we make a short leap from $\tilde{s}$ to s^* to impute all $y-$values under only first order model assumptions, and then s^*, being a probability sample, provides a solid design-based foundation to make inference about U^*using commonly used estimators such as generalized regression (GREG) of Särndal (1980). Following Singh (2015), this approach is termed model-over-design (MOD) integration signifying the use of design-based approach to minimize assumptions by using $s^*$and not $\tilde{s}$ as the base for inference, and then integrating in it the $y-$information under an imputation model based on $\tilde{s}$. It may be remarked that there is a price paid in terms of loss of precision when inferring about U^* from $\tilde{s}$ by way of s^* as opposed to inferring directly from $\tilde{s}$ if models under strong assumptions were indeed true, but it is probably worth it in view of much weaker and plausible assumptions.

Using a generalization of the reverse framework considered by Shao and Steel (1999), it is observed that variance estimates under the MOD integration approach can be obtained without any second order assumptions for the imputation model, and without any assumptions on the second order inclusion probabilities under $\tilde{\pi}$. In fact, reliance on the first order inclusion probabilities for $\tilde{\pi}$ estimated under a propensity score model can also be made weaker. Note that for variance estimation in the presence of imputed data for nonrespondents under probability samples using the usual quasi-design-based reverse framework, the total variance of the imputed estimator about the population total T_y has two parts: first is the design-based variance of the imputed estimator about a census-estimated population total with estimates of mean values for the nonrespondents; i.e., $\sum_U y'_k$ $(= \sum_{U_r} y_k + \sum_{U_{nr}} \mu_{kN})$ where U is the target population, U_r is the subpopulation of respondents, U_{nr} is the subpopulation of nonrespondents, and where μ_{kN}'s are finite population quantities defined by the estimate for the mean of y_k given covariates under first order assumptions of the outcome regression model at the census level; i.e., if the sample itself were the census; while the second part is the model-based variance of $\sum_U y'_k$ about the target parameter T_y $(= \sum_U y_k)$. For example, μ_{kN} can be $x'_{k+}\gamma_N$ under ξ of (1) where γ_N is an estimate of γ based on the complete finite population U; see Section 2. The second part of the variance is negligible in general for small sampling fractions, while the first part can be estimated under quasi-design-based approaches for two phase samples where the first phase refers to data collection under the actual design of the probability sample which could be multi-stage, and the second phase refers to the use of additional information in estimation of μ_{kN}'s for nonrespondents via deterministic imputation which

might also entail a model $\tilde{\psi}$ for random imputation given the first phase respondent subsample information. Here, the estimated μ_{kN}'s are typically nonlinear and are linearized before applying Taylor variance estimation techniques for totals.

For the present problem with purposive samples, U_r is empty and μ_{kN}'s are estimated from an independent sample $\tilde{s}$ and not the respondent subsample. Now, as discussed in Sections 2 and 3, the usual reverse framework under the quasi-design-based approach can be generalized to make it applicable to our problem under the MOD-integration approach if we treat $\tilde{s}$ as collecting extra information in a second phase to provide estimates or imputed values for μ_{kN}'s which in the case of random imputation will require additional information under a model $\tilde{\psi}$. The above generalization, depending on the imputation method and assumptions made, might require estimation of first order inclusion probabilities under unknown $\tilde{\pi}$. These probabilities are, in general, difficult to estimate reliably but can be done somewhat robustly using calibration equations with good covariates as described in Section 2. Also, need for second order inclusion probabilities under $\tilde{\pi}$ for the second phase variance estimation could be avoided which are known to be even more difficult to estimate. To this end, as explained in Sections 2 and 3, besides the usual assumption of WRPSU (with replacement primary sampling unit) selection in the first phase sample (i.e., s^*), if additional assumptions of conditional unbiasedness and conditional independence of imputed values about μ_{kN}'s can be made given s^*, then standard variance estimates for single phase designs become applicable which do not require explicitly the second phase variance estimate; see Singh (2008). The above additional assumptions can be approximately satisfied by using non-parametric imputation methods (such as nearest neighborhood means for imputing μ_{kN}'s under an unspecified model ξ^*) which are also expected to be robust to model mis-specification for estimating inclusion probabilities $\tilde{\pi}$ for units in $\tilde{s}$. Thus, the proposed variance estimators do not require strong assumptions about the propensity score model for $\tilde{\pi}$ and the nonparametric imputation model ξ^*.

In any model for imputation, there is great concern about model misspecifications. The proposed method further alleviates this problem by including extra covariates in the GREG estimator from s^* as follows. Observe that with the imputed y-values (y^I) from $\tilde{s}$, the usual GREG with key auxiliary control totals for x's can be applied on s^* to obtain point estimates, to be termed *i*GREG where *i* denotes imputation. However, if covariates x's (as well as other outcome variables z's known to be correlated with y) with known values in s^* are also imputed alongside y, then additional covariates based on differences $x^I - x$ and $z^I - z$ with corresponding constraints of zero control totals can be used to improve GREG with regard to possible imputation bias due to model mispecifications; see Singh, Iannacchione, and Dever (2003) on the use of zero controls for reducing mode effects. This way, the GREG-calibrated weights will yield estimates for T_x based on imputed values x^I that match exactly with known x -control totals T_x or estimated controls T_z^* in the case of $z-$variables obtained from s^*. This is a form of benchmarking and the resulting estimate denoted *ix*GREG (*x* for extra covariates with zero controls) is expected to reduce imputation bias and be robust to imputation model misspecification. In addition to such robustification of *i*GREG with respect to the imputation model, doubly robust imputation methods involving imputation class based on propensity score model and outcome regression model (parametric or nonparametric) for imputation within the imputation class (see, e.g., Haziza and Beaumont, 2007) can be used to obtain y^I for s^*.

The organization of this paper is as follows. Section 2 contains background review of prediction method and propensity score weighting method based primarily on $\tilde{s}$ while s^*

provides supplementary controls T_z^*, and motivation for the proposed method *i*GREG under the MOD-integration approach based primarily on s^* while $\tilde{s}$ provides imputed y-values. In Section 3, the proposed class of estimators *i*GREG is described. For imputation, both random imputation (hot deck) and deterministic imputation (neighborhood mean) using nearest neighbors defined by the distance metric under predictive mean matching (PMM, Little, 1988) are used. In the interest of double robustness, alternative versions of *i*GREG are defined where PMM is used within imputation classes defined by quantiles of the propensity score for all units in the combined sample $\tilde{s} \cup s^*$. Here, assuming that $\tilde{s}$ and s^* do not overlap, propensity refers to the probability of the event for a unit from $\tilde{s} \cup s^*$ to be selected in $\tilde{s}$; i.e., conditional on $\tilde{s} \cup s^*$. Note that $\tilde{s}$ plays the role of the respondent subsample while s^* as the nonrespondent subsample for which y-values are missing. The above propensity is different from the propensity used in propensity score weighting mentioned earlier. Variance estimates are provided in Section 4, and empirical results based on a limited simulation study in Section 5. Finally, Section 6 contains concluding remarks.

2. Background Review and Motivation for the Proposed Class of Methods

We first consider methods based primarily on $\tilde{s}$. There are two methods as mentioned in the introduction.

2.1 PRED ($\tilde{s}$)—Prediction Estimator based on $\tilde{s}$

It is the prediction estimation method of Royall (1970, 1976) where y is modeled given covariates observed in $\tilde{s}$. Let x_{i+} denote the vector of two types of covariates x_i and z_i for the *i*th unit in $\tilde{s}$ deemed to be well correlated with the outcome variable y_i where x_i are usual covariates (including the constant covariate of 1 corresponding to the intercept) with known totals T_x and z_i are other outcome variables with unknown totals T_z whose estimates T_z^* can be obtained from s* after weight calibration to control totals T_x; i.e., $T_z^* = \sum_{k\in s^*} z_k w_{k(1)}^*$. Here, $w_{k(1)}^*$ are GREG-calibrated weights obtained from the initial weights w_k^* and control totals T_x which include the population count *N* corresponding to the constant covariate of 1. Now, invoke a super-population model ξ with an intercept:

$$\xi\colon\ y_i = \mu_i + \varepsilon_i = x_{i+}'\gamma + \ \varepsilon_i\ ,\ \varepsilon_i \sim_{ind} (0,\ \sigma_\varepsilon^2) \qquad (1)$$

where the first order parameters are γ-- the fixed regression coefficients, and the second order parameters refer to the error variance σ_ε^2 and zero covariances of model errors. The postulated model ξ is at best a plausible model which may be misspecified. However, the unbiasedness property of estimators requires that at least the mean function is correctly specified. Let $\tilde{\gamma}_u$ denote the ordinary least squares estimator of γ where the symbol ~ signifies that it is based on the purposive sample $\tilde{s}$ and the subscript u signifies that it is unweighted; i.e., the sampling weights are not used which, in fact, are not known. The estimator is given by

$$t_{y,PRED(\tilde{s})} = T_{x+}'\tilde{\gamma}_u + \ \sum_{i\in\tilde{s}}(y_i - x_{i+}'\tilde{\gamma}_u) \qquad (2)$$

$$= T_{x+}'\tilde{\gamma}_u + 0$$

The total residual term on the right hand side of (2) is zero due to the presence of the intercept in the model. Variance of $t_{y,PRED(\tilde{s})}$ can be estimated under the model (1) and

requires estimating σ_ε^2 for which the usual estimate based on residuals can be plugged in. The main limitation of $PRED(\tilde{s})$ is that the model for the mean of y given x_{i+} under ξ is assumed to hold for $\tilde{s}$ and its complement. This may not be true due to potential selection bias in $\tilde{s}$. Besides the mean function may itself be misspecified which is clearly untestable due to unknown y-values precisely for units that are nonobserved in the sample and for whom we wish to predict. The above limitation is for point estimation. The second limitation is that even if the mean function is correctly specified and there is no selection bias, the variance estimate is obtained under the assumed covariance structure of ξ which is rather tenuous. The problem of variance estimation gets worse in the presence of selection bias.

2.2 cPROP ($\tilde{s}$)—Calibration Propensity Estimator based on $\tilde{s}$

This is the propensity score weighting method. Consider a logit linear model for the propensity $\tilde{\pi}_i$ of unit i being in $\tilde{s}$:

$$\text{logit}\,\tilde{\pi}_i = x'_{i+}\lambda \tag{3}$$

The propensity $\tilde{\pi}_i$ of unit i represents the probability of the compound phenomenon comprising observation unit recruitment, cooperation, and completion. The model parameters λ can be estimated by using a calibration method (Folsom and Singh, 2000) such that the control totals T_{x+} are satisfied by new weights $\tilde{w}_i(=\tilde{\pi}_i^{-1})$ obtained by adjusting the initial weights of 1. That is, the estimated parameters $\tilde{\lambda}$ satisfy the estimating equations

$$\sum_{i\in\tilde{s}} x_{i+}\tilde{w}_i = T_{x+} \tag{4}$$

and

$$\tilde{w}_i = \tilde{\pi}_i^{-1} = \left(1 + \exp\left(-x'_{i+}\tilde{\lambda}\right)\right) \tag{5}$$

The propensity score weighting estimator is now given by an expansion estimator like Horvitz-Thompson,

$$t_{y,\text{cPROP}(\tilde{s})} = \sum_{i\in\tilde{s}} y_i\tilde{w}_i = \tilde{y}'\tilde{w} \tag{6}$$

The main limitation of cPROP($\tilde{s}$) is the use of the untestable Missing-at-Random (MAR)-type assumption for nonselected units; i.e., given the covariates, the selection probability $\tilde{\pi}_i$ does not depend on the outcome variable y. This is similar to MAR for bias adjustment for nonresponse in probability surveys except that it is compounded by the fact that it also includes the event of selection of units besides the event of response. However, some robustification to model misspecification is provided by weight calibration constraints. The other main limitation is that the variance estimation requires the WRPSU-type assumption in the absence of second order selection probabilities where PSU refers to the ultimate cluster in the sampling design. This assumption is also purely speculative because the sampling design underlying $\tilde{s}$ is unknown and so the concept of ultimate cluster is hypothetical.

The calibration method for fitting the propensity model has the advantage that it does not require any unit level information about nonselected units as is needed in the alternative quasi-likelihood method commonly used for fitting response propensity models in probability surveys. Valliant and Dever (2011) used the quasi-likelihood method for fitting the propensity model for a unit in U^* to be selected in $\tilde{s}$ and took advantage of s^* with information about x_{k+} and weights w_k^* for units $k \in s^*$ to estimate the population quantities

in the quasi-likelihood estimating function for λ based on U^*. Here, it is assumed that there is no overlap between s* and $\tilde{s}$ but sampling weights of s* are not adjusted to reflect the subpopulation $U^* \setminus \tilde{s}$ because its effect is expected to be negligible. Elliott (2009) and Robbins et al. (2017) used a different approach using Bayes theorem for estimating the propensity $\tilde{\pi}_i$ for which we describe an alternative simplified derivation as follows.

First a logit model for a different propensity φ_j for a unit j in the combined sample $\tilde{s} \cup s^*$to be selected in $\tilde{s}$ is fit using the quasi-likelihood approach. It is assumed that the two samples $\tilde{s}$ and s^* do not overlap or they can be made so by dropping common units from $\tilde{s}$. This model is fit conditional on $\tilde{s} \cup s^*$ and so there is no need for sampling weights and, in fact, they are not even known for $\tilde{s}$. Now for a unit $i \in \tilde{s}$, if we knew the selection probability π_i^* for it to be in s^*, then the required propensity $\tilde{\pi}_i$ can be obtained as $\pi_i^*(\varphi_i/(1-\varphi_i))$. To see this, note that

$$\begin{aligned} \Pr(j \in \tilde{s}) &= \Pr(j \in \tilde{s}\ and\ j \in \tilde{s} \cup s^*) \\ &= \Pr(j \in \tilde{s} \cup s^*) \times \Pr(j \in \tilde{s} | j \in \tilde{s} \cup s^*) \end{aligned} \quad (7)$$

and similarly for $\Pr(j \in s^*)$. The desired result follows from the observation that the ratio $\Pr(j \in \tilde{s}) \,/\, \Pr(j \in s^*)$ can be expressed as $\varphi_j/(1-\varphi_j)$ since $\Pr(j \in \tilde{s} | j \in \tilde{s} \cup s^*) = \varphi_j$, and $\Pr(j \in s^* | j \in \tilde{s} \cup s^*) = 1 - \varphi_j$. It may be remarked that all the above probabilities are conditional on the covariates x_{j+}. In practice, π_i^* for units $i \in \tilde{s}$ are not known and another logit model can be fit to observed π_k^* in s^* to approximate it as a function of covariates. This model can then be used to estimate π_i^*. Note that the additional model for π_k^* is not for the same purpose as the propensity score modeling, and will not be needed if the sample design π^* is based on covariates known for $\tilde{s}$ although this is unlikely in practice. The proposed calibration method based on (4) above offers a simpler alternative to above methods based on quasi-likelihood.

2.3 Motivation for the Proposed Class of Estimators

With the estimated sampling weights $\tilde{w}_i$ for $i \in \tilde{s}$ using propensity modeling, an alternative estimator *GREG* ($\tilde{s}$) of Särndal (1980) based on $\tilde{s}$ can be easily defined as

$$\begin{aligned} t_{y,GREG(\tilde{s})} &= T'_{x+}\tilde{\gamma}_w + \sum_{i \in \tilde{s}} (y_i - x'_{i+}\tilde{\gamma}_w)\, \tilde{w}_i \\ &= T'_{x+}\tilde{\gamma}_w + 0 \end{aligned} \quad (8)$$

which in contrast to *PRED* ($\tilde{s}$) of (2) uses a weighted estimate $\tilde{\gamma}_w$ of regression parameters γ based on estimated weights $\tilde{w}_i$, and the second term on the right hand side of (8) is the sum of weighted residuals which reduces to zero due to presence of the intercept in the model. There is an interesting and desirable property of *cPROP*($\tilde{s}$) in that it is equal to *GREG* ($\tilde{s}$) whenever the unit vector (denote by 1) is in the column space of covariates x_+; i.e., $1 = \tilde{X}_+\alpha$ for some real vector α and where $\tilde{X}_+$denotes the matrix of covariate values with rows given by x'_{i+}, which, in fact, is the case for models with an intercept. This suggests a robustness property for *cPROP*($\tilde{s}$) in that two completely different models (propensity score and outcome regression) give rise to the same estimator. To see this, note that

$$\begin{aligned} t_{y,\mathrm{cPROP}(\tilde{s})} &= \tilde{y}'\tilde{w} = \tilde{y}'\tilde{W}1 = \tilde{y}'\tilde{W}\tilde{X}_+\alpha \\ &= (\tilde{y}'\tilde{W}\tilde{X}_+)\left(\tilde{X}'_+\tilde{W}\tilde{X}_+\right)^{-1}\left(\tilde{X}'_+\tilde{W}\tilde{X}_+\right)\alpha \end{aligned}$$

$$= \tilde{\gamma}_w'\left(\tilde{X}_+'\tilde{W}\tilde{X}_+\right)\alpha = \tilde{\gamma}_w'\tilde{X}_+'\tilde{w}$$

$$= \tilde{\gamma}_w' T_{x+} + 0 = t_{y,GREG(\tilde{s})} \qquad (9)$$

because the weighted sum of residuals in *GREG* ($\tilde{s}$) defined by (8) is zero due to presence of the intercept term in the model. The above property is not shared by *PRED* ($\tilde{s}$) because it uses the unweighted regression parameter estimates $\tilde{\gamma}_u$ instead of the weighted estimates $\tilde{\gamma}_w$. Thus, *PRED* ($\tilde{s}$) is subject to design bias if the model ξ does not hold for $\tilde{s}$ and its complement; i.e., if the unknown design $\tilde{\pi}$ is nonignorable for the model given covariates x_{i+}.

Now, suppose the imputed values y_k^I are given by the weighted predictive means (PM), $x_{k+}'\tilde{\gamma}_w$ for all $k \in$ s*, and then a GREG estimator is computed. Denote it by *iGREG (PM)*. Somewhat surprisingly, it is observed that the two *GREG* point estimates *iGREG* (PM) and *GREG* (s̃) based respectively on different samples s* and s̃ do, in fact, coincide. This can be explained as follows. First note that the GREG calibration for w_k^* −weights in s* does not change whether T_{x+} with the extra control T_z^* is used or only T_x (including the control *N* but without the extra control T_z^*) because T_z^* does not add any new information. It is so because T_z^* is itself obtained as a calibration estimator using calibrated weights $w_{k(1)}^*$ computed from the initial w_k^* −weights under controls T_x. However, the corresponding regression coefficients β_w^* and γ_w^* (corresponding to the case of extra control) are different as they have different dimensions. Now the prediction or the first part of GREG is identical for both estimators because with imputed values y_k^I as $x_{k+}'\tilde{\gamma}_w$, the weighted regression coefficient estimator γ_w^* from s* coincides with the weighted regression coefficient estimator $\tilde{\gamma}_w$ obtained from s̃. That is,

$$\gamma_w^* = (X_+^{*\prime} W^* X_+^*)^{-1}(X_+^{*\prime} W^* y^I)$$

$$= (X_+^{*\prime} W^* X_+^*)^{-1}(X_+^{*\prime} W^* X_+^*)\tilde{\gamma}_w = \tilde{\gamma}_w. \qquad (10)$$

For the second part in *GREG*; i.e., the weighted sum of residuals, it is already shown to be zero for *GREG* (s̃), and is also zero for *iGREG* (PM) as $\sum_{k\in s^*}(y_k^I - x_{k+}'\gamma_w^*)w_k^*$ reduces to zero because y_k^I defined by $x_{k+}'\tilde{\gamma}_w$ equals $x_{k+}'\gamma_w^*$ in view of (10). However, the two estimators *iGREG (PM)* and *GREG* ($\tilde{s}$) can have different variances depending upon the choice of the underlying random mechanism. In our case, *iGREG (PM)* is driven jointly by the design π^* for s*, $\tilde{\pi}$ for s̃ and the outcome regression model ξ, and *GREG* ($\tilde{s}$) is driven by the design $\tilde{\pi}$ based on the propensity score model.

Cleary, *iGREG (PM)* is a candidate for the proposed class of estimators under MOD-Integration. However, the main question is whether a reliable estimate of its variance can be obtained without making strong second order assumptions for the imputation model ξ or the propensity score model $\tilde{\pi}$. The answer is no in general but it might work well for very large $\tilde{s}$ relative to s^*. To see this, consider the analogy with imputation for nonrespondents in probability surveys where respondent subsamples serve as the donor dataset. In such surveys with nonresponse, imputed values can be viewed as providing the second phase information and variance can be estimated using standard single phase formulas (i.e, WRPSU-type) under the reverse framework as mentioned in the introduction if imputed values are conditionally unbiased and conditionally independent across units with missing values where conditional refers to the first phase sample. It is shown in Singh

(2008) that the usual requirements of invariance and independence (Särndal, Swensson, and Wretman, 1992, pp. 134) for variance estimation in (single phase) two stage sampling can be simplified and replaced by weaker conditions of conditional unbiasedness and conditional independence given the first phase sample, and thus can be used to obtain two phase variance estimates via single phase methods; see Appendix A.5 for more details.

Now, in the case of our problem, there is complete nonresponse in s^* and a separate independent sample $\tilde{s}$ serves as the donor dataset. Despite this difference, the simplified variance estimation under the reverse framework can be generalized for our purposes if imputation methods are chosen appropriately. Observe that the regression coefficients $\tilde{\gamma}_w$ are common in PM imputation for all units in s^*, and therefore, the imputed values do not satisfy the conditional independence assumption given s^*. However, they may satisfy it approximately if $\tilde{s}$ is much larger than s^* in which case, under general regularity conditions, $\tilde{\gamma}_w$ is consistent for γ_N under the distribution $\tilde{\pi}|\xi$ and can be treated asymptotically as fixed; i.e., $\tilde{\gamma}_w$ is very close to γ_N with high probability. Then the covariances between y_k^I's about their mean values μ_{kN} (= $x'_{k+}\gamma_N$) would be negligible under $\tilde{\pi}|\xi$. As a consequence, standard variance estimation methods in survey sampling adjusted for imputation would be applicable to *iGREG (PM*) under the joint distribution $\pi^*\tilde{\pi}\xi$. However, the assumption of negligible covariances for applicability of standard methods is not likely to be tenable in practice and some adjustments are required; see Subsection A.5. Moreover, the imputed values in *iGREG (PM*) do depend strongly on the model ξ and are not robust to model mispecifications. These considerations lead to semiparametric imputation methods such as predictive mean matching (PMM) imputation by random hot deck within PM neighborhood (PMN) or deterministic using PMN means as proposed in the next section. They are semiparametric in nature because although a parametric model ξ is used for PMM for computing the distance metric, a nonparametric model ξ^* with an unspecified functional form of the outcome mean given the covariates is used for imputation via nearest neighbors.

3. Proposed Class of Estimators under MOD-Integration: *iGREG*

Based on the motivation in the previous section, the proposed estimators begin with a design-based approach using the reference sample s^*, and then build over it by integrating information about $y-$values from the purposive sample $\tilde{s}$ using an imputation model. That is, under the model ξ, impute y_k for each unit $k \in s^*$ from $\tilde{s}$ (treated as a donor dataset), and then define

$$t_{y,iGREG(s^*)} = (y^I)' w^*_{(1)} \qquad (11)$$

where $w^*_{(1)}$ are the calibrated weights for s* satisfying the controls T_x.

To define various members of the class *iGREG (s*)* based on PMM for different y^I, first fit the imputation model ξ using $\tilde{s}$ and the estimated weights $\tilde{w}$ obtained under the propensity score model and denote the estimated predictive means $x'_{i+}\tilde{\gamma}_w$ as PM_i for $i \in \tilde{s}$ and $x'_{k+}\tilde{\gamma}_w$ as PM_k for $k \in s^*$. For each unit $k \in s^*$, use the above predictive means in a distance metric to find K_0 (10 to 20, for example) nearest neighbors from $\tilde{s}$ where the distance metric between units *i* and *k* is defined by $|PM_i - PM_k|$.

For predictive mean neighborhood (PMN) random hot deck imputation, draw a unit at random from K_0 neighbors using weighted hot deck; i.e., with estimated $\tilde{w}_i$ as size measures for the probability proportional to size (PPS) sampling. This random draw mechanism is denoted by $\tilde{\psi}$. Also denote the resulting random imputation by $y^I_{k,PMN-r}$ where r signifies random. Similarly, obtain $x^I_{k,PMN-r}$ and $z^I_{k,PMN-r}$ for known covariate values from the same donor for imputation model diagnostics and bias correction. Observe that for the above PMN-r method, unlike the PM imputation method described in the previous section, the resulting imputed values across units in s^* are only weakly dependent on the common regression coefficient estimator $\tilde{\gamma}_w$ because it is used only in the distance metric to find neighbors. This semiparametric method is useful for making PMN-r robust to misspecifications of the imputation model. To see that the PMN-r imputations y^I_k satisfy approximately the assumptions of conditional unbiasedness for μ_{kN} (only conceptual under ξ^* for the census data) and independence given s^*, suppose for each unit $k \in s^*$, the random draw from the neighborhood PMN_k gives rise to the imputed value y^*_i corresponding to a unit $i \in PMN_k$. We have the identity,

$$y^I_k - \mu_{kN} = y^*_i - \mu_{kN} = (y^*_i - \bar{y}^*_k) + (\bar{y}^*_k - \mu_{kN}) \qquad (12)$$

where $\bar{y}^*_k$ is the weighted average of the donors in the neighborhood PMN_k. Since the random draws are independent from different PMNs, the first terms $(y^*_i - \bar{y}^*_k)$ across units $k \in s^*$ on the right hand side of (12) are conditionally independent under $\tilde{\psi}$ given $\pi^*\tilde{\pi}\xi^*$. The second terms $(\bar{y}^*_k - \mu_{kN})$ across units $k \in s^*$ can be deemed to be approximately conditionally independent under $\tilde{\pi}$ given $\pi^*\xi^*$ assuming that the neighborhood size is much smaller than the size of $\tilde{s}$, and $\tilde{\gamma}_w$ used in the distance metric is a consistent estimate of γ_N for large $\tilde{s}$. Moreover, the approximate conditional unbiasedness of y^I_k about μ_{kN} also holds because the mean of $(y^*_i - \bar{y}^*_k)$ is zero due to random draws, and the mean of $\bar{y}^*_k$ can be assumed to be asymptotically unbiased for μ_{kN} by construction of the neighborhood based on nearest neighbors. The resulting estimator is denoted by *iGREG(PMN-r)* for which the standard variance estimators adjusted for nonresponse in survey sampling become approximately valid. That is, the contribution to the total variance due to imputation in the second phase is embedded in the first phase variance between PSU estimates under the WRPSU-type assumption; see Subsection A.5. Here, PSU denotes the generic ultimate cluster which could be elementary units if there is no clustering in the first phase sampling design for s^*. In practice, however, unless $\tilde{s}$ is very large relative to s^*, PMNs are likely to have some donor units common for different units $k \in s^*$, which would lead to nonnegligible covariances between PMN means. In such cases, a standard but conservative variance estimate can be obtained; see Subsection 4.3.

The PMN-deterministic imputation denoted by $y^I_{k,PMN-d}$ where d signifies deterministic, is obtained by the weighted average of the $y-$values in the neighborhood PMN_k. Similarly, for the covariates, we can obtain $x^I_{k,PMN-d}$ and $z^I_{k,PMN-d}$ for all $k \in s^*$. Following the argument above for PMN-r, standard variance estimation formulas can be used for PMN-d methods. The resulting estimator is denoted by *iGREG (PMN-d).*

In light of the desirable double robustness property for imputation mentioned in the introduction when both outcome regression and propensity score models are used, we can easily define both random hot deck and deterministic imputations for PMN within class

(PMNC); see e.g., Haziza and Beaumont (2007). To this end, we first fit the propensity model for a unit $j \in (\tilde{s} \cup s^*)$ to be in $\tilde{s}$ using the usual quasi-likelihood as described in Subsection 2.2, and then partition the combined sample $\tilde{s} \cup s^*$into five or so equal parts using quantiles of the propensity score distribution. Now, depending on which of the five classes each $k \in s^*$ belongs based on its propensity score, PMN-r and PMN-d imputations are performed within each class. Here, the random hot deck and the PMN means are unweighted within the neighborhood because the donors belong to the same propensity class with approximately equal propensity scores. This feature of being able to use unweighted imputation is conducive for efficiency of estimators as they are not subject to instability in estimation of propensity scores. Moreover, the resulting estimators, to be denoted by $iGREG(PMNC$-r) and $iGREG(PMNC$-d), become more robust.

Finally, imputation model diagnostics can be developed using observed differences $x_{k+}^I - x_{k+}$ where x_{k+}, although known in s^*, is also imputed from the same donor for diagnostic purposes. Moreover, the estimator *iGREG* can be corrected for bias by using extra covariates $x_{k+}^I - x_{k+}$ with zero control totals. It is also possible to control the total observed differences over subsets of x_+ defined, for example, by quartiles. The bias corrected *iGREG* estimators are denoted by *ixGREG(PMN-r)* and *ixGREG(PMNC-r)* where '*x*' signifies the extension of covariates with zero controls. Similarly, *ixGREG* estimators with deterministic imputations can be defined as before.

4. Variance Estimation

For variance estimation, it is convenient to express all estimators as calibrated expansion estimators based on s^* using weights $w_{k(1)}^*$ where y −values are replaced by corresponding imputed values. Interestingly, *PRED($\tilde{s}$)* can also be expressed as a calibration estimator. Note that the calibrated weights $w_{k(1)}^*$ can be obtained from the initial weights w_k^* after multiplying them by the GREG-calibration adjustment factors (denote by g_k's) such that the calibration controls T_x (including the population count *N* but not T_z^*) are satisfied. By expressing all estimators as calibrated expansion estimators, it makes it possible for a meaningful comparison of all variance estimates under the common joint randomization of $\pi^*\tilde{\pi}\xi$.

4.1 *PRED($\tilde{s}$)*

It can be expressed as $\sum_{s^*} y_k^I w_k^* = \sum_{s^*} y_k^I w_{k(1)}^*$ where $y_k^I = x_{k+}'\tilde{\gamma}_u$, and $w_{k(1)}^* = g_k w_k^*$. That is, *PRED($\tilde{s}$)* has the form of *iGREG(PM)* but here PM is obtained without using weights $\tilde{w}$ in estimating γ which might lead to design bias as the weighted residuals for s^* do not sum to zero. Now, if y −values were known, and treating individual units *k* as PSUs, the standard variance estimate of $t_{y,GREG(s^*)}$, under the WRPSU-assumption after Taylor linearization for calibration estimation, would have been obtained as (see (A19) in the appendix)

$$v(t_{y,GREG(s^*)}) = (n^*/(n^*-1)\sum_{s^*}(e_k g_k w_k^* - \overline{egw})^2 \qquad (13)$$

where $\overline{egw}$ denotes the simple average of $e_k g_k w_k^*$ over the size n^* of s^*, and e_k's denote the GREG residuals $y_k - x_k'\beta_w^*$. Note that it is $x_k'\beta_w^*$ and not $x_{k+}'\gamma_w^*$ of (10) used here in

the residual GREG on s^* because for this calibration, there is no extra covariate. Use of g_k in (13) is known for improved finite sample properties. The formula (13) is valid when the design is single stage; else, PSUs or the ultimate clusters corresponding to multi-stage design are used under the WRPSU assumption where n^* is replaced by the number of PSUs. For imputed y_k −values (here, $y_k^I = x'_{k+}\tilde{\gamma}_u$) based essentially on the same covariates used for calibration, it seems natural to replace e_k by e_k^I defined as $(y_k^I - x'_k\beta_w^*)$ in the variance estimate (13). However, the residual definition is not quite meaningful because y_k^I is not a realized *y*- value and is itself a function of the covariates. The resulting variance estimate might be too low and in practice, in the interest of a conservative estimate, it might be better to use y_k^I in place e_k in the formula (13) without any Taylor linearization. With this substitution in (13), it follows from the appendix that under certain conditions, we can capture most of the second phase variance where the second phase refers to integrating information about y −values from $\tilde{s}$ via imputation under the random mechanism $\tilde{\pi}$ given $\pi^*\xi$. To see this, first assume y_k^I is approximately unbiased for μ_{kN} (this may not be true for *PRED(s̃)*) and then write the variance of $\sum_{s^*} y_k^I w_k^*$ about T_y as

$$V_{\pi^*\tilde{\pi}\xi}\left(\sum_{s^*} y_k^I w_k^*\right) = E_\xi V_{\pi^*\tilde{\pi}|\xi}\left(\sum_{s^*} y_k^I w_k^*\right) + V_\xi E_{\pi^*\tilde{\pi}|\xi}\left(\sum_{s^*} y_k^I w_k^*\right) \qquad (14)$$

where the second term on the right hand side is negligible relative to the first term for small sampling fractions under π^*. Moreover,

$$V_{\pi^*\tilde{\pi}|\xi}\left(\sum_{s^*} y_k^I w_k^*\right) = E_{\pi^*|\xi} V_{\tilde{\pi}|\pi^*\xi}\left(\sum_{s^*} y_k^I w_k^*\right) + V_{\pi^*|\xi} E_{\tilde{\pi}|\pi^*\xi}\left(\sum_{s^*} y_k^I w_k^*\right) \qquad (15)$$

where

$$V_{\tilde{\pi}|\pi^*\xi}\left(\sum_{s^*} y_k^I w_k^*\right) = E_{\tilde{\pi}|\pi^*\xi}\left(\sum_{s^*} x'_{k+}\tilde{\gamma}_u w_k^* - \sum_{s^*} \mu_{kN}\ w_k^*\right)^2$$

$$= E_{\tilde{\pi}|\pi^*\xi}\left(\sum_{s^*} x'_{k+}(\tilde{\gamma}_u - \gamma_N) w_k^*\right)^2 \qquad (16)$$

$$V_{\pi^*|\xi} E_{\tilde{\pi}|\pi^*\xi}\left(\sum_{s^*} y_k^I w_k^*\right) = E_{\pi^*|\xi}\left(\sum_{s^*} \mu_{kN}\ w_k^* - \sum_{U^*} \mu_{kN}\right)^2 \qquad (17)$$

For very large sample $\tilde{s}$ relative to s^*, $\tilde{\gamma}_u$ could be regarded as asymptotically fixed and very close to γ_N with high probability, and therefore, $V_{\tilde{\pi}|\pi^*\xi}\left(\sum_{s^*} y_k^I w_k^*\right)$ would be negligible. It follows from the subsection A.5 in the appendix that the condition of conditional independence is approximately satisfied by treating $\tilde{\gamma}_u$ as asymptotically fixed, and in addition, assuming approximate unbiasedness of y_k^I, a simple variance estimate similar to (13) for $t_{y,PRED(\tilde{s})}$ is obtained where e_k is replaced by y_k^I. Thus, standard methods of survey sampling could be applicable under suitable conditions. It is remarked, however, that it is probably not realistic to assume that $\tilde{\gamma}_u$ can be treated as asymptotically fixed in the above simplified variance estimate calculation for the usual $\tilde{s}$ sizes available in practice. In such cases, the resulting variance estimate would be an underestimate, and assuming approximate unbiasedness of $\tilde{\gamma}_u$, second order assumptions for the propensity score model for $\tilde{\pi}$ would be required to obtain suitable variance estimates; see A.5.

4.2 *cPROP(s̃)*

Writing it as a calibrated expansion estimator using imputed values; i.e., as $\sum_{s^*} y_k^I w_{k(1)}^*$, where y_k^I is now defined as $x'_{k+}\tilde{\gamma}_w$, it has the form of *iGREG(PM)* where PM now uses weights $\tilde{w}$ in estimating γ. The main difference between *PRED(s̃)* and *cPROP(s̃)* is the

use of weighted $\tilde{\gamma}_w$ which is expected to alleviate the problem of selection bias in $\tilde{s}$ and thus supports approximate unbiasedness of $\tilde{\gamma}_w$. Again neglecting the asymptotic variance of the second phase term due to $\tilde{\gamma}_w$ when $\tilde{s}$ is large, standard variance estimate formulas as given by (13) with e_k replaced by y_k^I can be used for the estimator $t_{y,cPROP(\tilde{s})}$. If $\tilde{s}$ were not large, additional assumptions for $\tilde{\pi}$ would be needed as mentioned earlier.

4.3 *iGREG* Estimators

The variance estimate formula (13) remains applicable, in general, to all *iGREG* and *ixGREG* estimators under much weaker assumptions after substituting e_k by suitably defined e_k^I. In particular, for *iGREG,* e_k^I is defined as $(y_k^I - x_k'\beta_w^*)$ but for *ixGREG* , it is defined as $(y_k^I - x_{k0}'\beta_{w0}^*)$ where x_{k0} denotes the original covariates x_k's extended by the new covariates $x_{k+}^I - x_{k+}$ with zero controls, and the GREG regression coefficient β_{w0}^* is defined accordingly. Observe that in the case of *PMN-d,* the term $x_{k+}'(\tilde{\gamma}_u - \gamma_N)$ on the right hand side of (16) is replaced by $\bar{y}_k^* - \mu_{kN}$ where $\bar{y}_k^*$ is defined as in (12) and μ_{kN} is the conceptual conditional mean given covariates under a nonparametric model ξ^*. Therefore, in order to be able to use formula (13) which captures the second phase variances but not covariances, it is, therefore, sufficient to ensure that $\bar{y}_k^* - \mu_{kN}$ are approximately uncorrelated over units $k \in s^*$ given $\pi^*\xi^*$; see Subsection A.5. As mentioned earlier in Section 3, this is approximately satisfied as PMNs are formed independently across different units, and any covariances between PMN means due to common $\tilde{\gamma}_w$ used in forming PMNs are expected to be negligible as the PMN size K_0 is much smaller than the size of $\tilde{s}$. Moreover, since $\tilde{\gamma}_w$ is not directly used for imputation, but only indirectly in the distance metric for computing PMs, the assumption of conditional independence (or lack of correlation) is likely to hold even when $\tilde{s}$ is not too large relative to s^*.

However, there might be considerable overlap of donors between PMNs for different units in s^* which could lead to non-negligible covariances. Ignoring these covariances is expected to lead to a liberal (i.e., biased downward) variance estimate because the covariances σ_{ij}'s are likely to be positive. This problem could be overcome by collapsing similar PSUs (within similar strata of s^* if π^* is a stratified design) to define variance-PSUs (varPSUs) such that their number is not too small such as 30. For each varPSU, donors could come from corresponding bootstrap replicates of $\tilde{s}$ by drawing subsamples of $\tilde{s}$ with replacement. This way the assumption of conditional independence of estimates from different varPSUs given s^* could be satisfied, and a conservative (i.e., biased upward) variance estimate given by the between varPSU estimate variability could be obtained. For the case of *PMN-r,* there is the additional randomization under $\tilde{\psi}$ in the second phase for random imputation so that the variance estimate is now obtained under the joint mechanism $\pi^*\tilde{\pi}\tilde{\psi}\xi^*$ which would be larger than the deterministic imputation *PMN-d* case.

5. Empirical Results

A limited simulation study was conducted to evaluate and compare various estimators. Variance estimators were not included in the study due to time constraints but will be included in further empirical work. Also, the PMNC estimators were not included. We

generated a population of 14000 retail stores somewhat similar to Hansen et al. (1983) but we introduced more study (or outcome) and related variables to avoid oversimplification. Consider study variables y_k and z_k representing total annual sales for commodities A and B for the kth store. Also, let x_k denote the store employee size and u_k the store advertisement expense. The units of measurement for various variables are appropriately transformed for the purpose of data generation.

First generate store size x-population using *Gamma* (2,5). Also generate u_k as Unif (1,7). Given x and u, generate the total annual sales (y) using Gamma such that under ξ_0

$$E_{\xi_0}(y_k|x_k, u_k) = 50 + 0.6x_k + x_k^2 + 10(u_k - 4)\ , \qquad (18)$$

and similarly,

$$E_{\xi_0}(z_k|x_k, u_k) = 25 + 0.4x_k + 5(u_k - 4), \qquad (19)$$

where ξ_0 denotes the true but unknown outcome model. For each replicate ν, generate the population $U^{(\nu)}$ and draw the reference probability sample $s^{*(\nu)}$ of size $n^* = 400$ using PPS with the size measure $x_k^0 = x_k u_k$. Next, generate a purposive sample $\tilde{s}^{(\nu)}$ of size $\tilde{n} = 800$ from $U^{(\nu)}$ using a linear model for inclusion probabilities under Poisson sampling:

$$\tilde{\pi}_k = l + (u - l)(y_k/(a + y_k)); \; l = .01, \; u = .90, \; a = 500 \qquad (20)$$

The purposive sample is also a probability sample $\tilde{s}^{(\nu)}$ from the same population $U^{(\nu)}$ but its selection probabilities are presumed unknown for construction of estimators. From the Poisson sample which has a random sample size but expected to be large, take a simple random sample (SRS) of size 800. The variable u_k is treated as unobserved. Both designs are nonignorable for the postulated model ξ (y|x, z) for prediction or imputation as given in (1). Note that to reflect reality, the true model ξ_0 is chosen to be more complex on purpose than the postulated model, but it is the assumed model under which all the estimators and their properties are studied. It is precisely for this reason that it is important to have robust estimators which are expected to give reasonable estimators even in the case of misspecified models.

Results from the above simulation study with R = 1000 replicated populations are shown in Table 1. The true parameter is taken as the population mean A_y which was obtained as 205.8. Note that the population varied for each replication but the covariates were held at the initially selected values as well as the population and sample sizes. Therefore, the parameter A_y represents the average over all population means. The weighted sample estimator of A_y from the probability sample s^*averaged over replications was observed as 205.97 which is very close to the true value as expected. For this purpose, the y-values for s^*were assumed known but were regarded as unavailable when considering proposed estimators. The unweighted estimator from the purposive sample $\tilde{s}$ averaged over replications was obtained as 326.63 which is considerably biased upwards due to much higher selection probabilities for large values. Table 1 shows four estimators *PRED* ($\tilde{s}$), *cPROP* ($\tilde{s}$), and *ixGREG(PMN-r)* along with *GREG* ($\tilde{s}$) provided as a benchmark. The estimator *GREG* ($\tilde{s}$) uses true sampling weights under $\tilde{\pi}$ although in practice they will be unknown. The columns of Table 1 show three performance characteristics, average point estimate, relative bias, and relative root mean square error. It is seen that *PRED* ($\tilde{s}$) is very

vulnerable to model misspecification and shows considerable downward bias. The two estimators *cPROP* ($\tilde{s}$) and *ixGREG(PMN-r)* perform somewhat similarly although having biases in opposite directions, but with somewhat lower absolute bias for *ixGREG(PMN-r).* However, *cPROP*($\tilde{s}$) (which is equivalent to *iGREG(PM)*) outperforms *ixGREG (PMN-r)* in terms of RRMSE. Preference of one over the other in the iGREG class of estimators in practice will depend on the performance of their variance estimators which will be investigated in future.

6. Concluding Remarks

Based on the discussion in the paper, it follows that it is generally difficult to construct a good estimator from the purposive sample $\tilde{s}$ alone due to problems of bias and invalidity of the model for the sample, and lack of any practically defensible margin of error. It might be considerably beneficial if extra information in the form of an extant reference survey data s^* were available where y is not collected but can be used to obtain extra control totals for auxiliaries. It was observed in the limited simulation study that the model-based prediction method (*PRED*) may be quite vulnerable to bias than MOD-integration methods provided by the *iGREG* class. An interesting but somewhat surprising finding was that the calibration propensity method (*cPROP*) also belongs to the *iGREG* class as it is equivalent to *iGREG(PM)* and exhibits good performance for point estimation relative to the *iGREG(PMN-r)* estimator. However, its performance with respect to variance estimation based on standard WRPSU-type formulas might not be at par due to possible underestimation based on theoretical considerations. It is planned to check for empirical evidence in this regard. It was observed that for various estimators, by viewing imputation of y in s^* as information collected in a second phase sample, standard sampling methods under the reverse framework could be used to obtain variance estimates provided they satisfy the conditions of conditional independence and unbiasedness given the first phase sample s^*.

Finally, we remark that the proposed MOD-Integration approach could also be applied to the problem of generalization of causal inference from randomized trials (Stuart et al., 2011). This problem is different from the problem considered in this paper in that there are two outcome variables of interest for each unit of the target population having a condition to be treated; i.e., two outcome measures for each unit if it were conceptually assigned separately to treatment and control groups. The randomized trial data provides two purposive samples—one for the treatment outcome and other for the control outcome. Using a suitable reference probability sample representing the target population with the condition, it is then possible to estimate population means for treatment and control outcome variables and hence the average treatment effect, for example.

Acknowledgments

The first author would like to thank Michael Kirsch and Harrison Greene of AIR and Peter Meyer, Nat Schenker, and Jennifer Parker of NCHS for the opportunity to visit NCHS on a regular part-time basis under an IPA arrangement during 2016-17. This led to several useful discussions in meetings on a related project with participants from the Division of

Research and Methodology at NCHS. An earlier version of this paper was presented in an invited session at AAPOR, 2016.

Disclaimer

The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the National Center for Health Statistics, Centers for Disease Control and Prevention.

Appendix

For the sake of completeness and convenience, we will briefly review the theory of variance estimation in survey sampling using notation common to the theory presented in this paper; for more details, see Raj (1968, Chapter 6) and Särndal, Swensson, and Wretman (1992, Chapters 4 and 9). For a probability sample s of fixed size n under the sampling design π with first order and second order inclusion probabilities π_i, π_{ij} for units $i, j \in U$ where U is the finite population of size N, we have

$$\sum_{i\in U} \pi_i = n, \sum_{j\in U, j\neq i} \pi_{ij} = (n-1)\pi_i \qquad \text{(A1)}$$

Let $\delta_{i\in s}$ denotes the indicator of the event for a unit $i \in U$ to be included in s, then $\sum_{i\in U} \delta_{i\in s} = n$, and therefore, $E_\pi\left(\sum_{i\in U} \delta_{i\in s}\right) = \sum_{i\in U} E(\delta_{i\in s}) = \sum_{i\in U} \pi_i$ obtains the first part in (A1). For the second part, write π_{ij} as π_i times the conditional probability $\pi_{j|i}$. Now, it follows from the first part of (A1) that $\sum_{j\in U\setminus\{i\}} \pi_j = n - 1$. In the following we will assume that there is no nonresponse.

A.1 Single Stage designs (without replacement)

For a single stage probability sampling design without replacement, variance of the usual expansion (Horvitz-Thompson) estimator $t_{y(\pi)} (\equiv \sum_s y_i/\pi_i)$ of the population total T_y under π is given by

$$V_\pi\left(t_{y(\pi)}\right) = \sum_{i\in U}(1-\pi_i)\frac{y_i^2}{\pi_i} + \sum_{i\in U}\sum_{j\in U, j\neq i}(\pi_{ij} - \pi_i\pi_j)\frac{y_i y_j}{\pi_i\pi_j} \qquad \text{(A2)}$$

which can also be expressed in the Sen-Yates-Grundy form for fixed n as

$$V_\pi\left(t_{y(\pi)}\right) = \frac{1}{2}\sum_{i\in U}\sum_{j\in U, j\neq i}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \qquad \text{(A3)}$$

Above follows from the observation that

$$\frac{1}{2}\sum_{i\in U}\sum_{j\in U, j\neq i}(\pi_i\pi_j - \pi_{ij})\left(\frac{y_i^2}{\pi_i^2} + \frac{y_j^2}{\pi_j^2}\right) = \sum_{i\in U}\sum_{j\in U, j\neq i}(\pi_i\pi_j - \pi_{ij})\frac{y_i^2}{\pi_i^2}$$

$$= \sum_{i\in U}\frac{y_i^2}{\pi_i}(n-\pi_i) - \sum_{i\in U}\frac{y_i^2}{\pi_i^2}\sum_{j\in U, j\neq i}\pi_{ij} = \sum_{i\in U}\frac{y_i^2}{\pi_i}(1-\pi_i) \qquad \text{(A4)}$$

An unbiased variance estimator is obtained from (A3) as

$$v_\pi\left(t_{y(\pi)}\right) = \frac{1}{2}\sum_{i\in s}\sum_{j\in s, j\neq i}(\pi_i\pi_j - \pi_{ij})\frac{1}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2 \quad \text{(A5)}$$

In the case of simple random sampling (SRS) without replacement, π_i, π_{ij} are respectively given by n/N and $n(n-1)/N(N-1)$, which from (A5) yield the familiar estimator

$$v_\pi\left(t_{y(\pi)}\right) = N^2\frac{N-n}{Nn(n-1)}\frac{1}{2n}\sum_{i\in s}\sum_{j\in s, j\neq i}(y_i - y_j)^2 = N^2\frac{N-n}{Nn(n-1)}\sum_{i\in s}(y_i - \bar{y})^2 \quad \text{(A6)}$$

in view of the wellknown identity $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{2n}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(x_i - x_j)^2$.

A.2 Single Stage designs (with replacement)

If the design were with replacement with selection probability p_i for a unit $i \in U$ in any given draw, then the variance of the unbiased Hansen-Hurwitz estimator $t_{y(p)}$ $(\equiv (1/n)\sum_{i\in s} y_i/p_i)$ of T_y has a much simpler formula and is given by

$$V_p\left(t_{y(p)}\right) = \frac{1}{n}\sum_{i\in U} p_i \left(\frac{y_i}{p_i} - T_y\right)^2 \quad \text{(A7)}$$

which has an unbiased estimator given by

$$v_p\left(t_{y(p)}\right) = \frac{1}{n(n-1)}\sum_{i\in s}\left(\frac{y_i}{p_i} - t_{y(p)}\right)^2 \quad \text{(A8)}$$

Denoting np_i as π_i, the estimator $t_{y(p)}$ has the same form as $t_{y(\pi)}$ which in terms of sampling weights is given by $\sum_s y_i w_i$ where $w_i = \pi_i^{-1}$. An alternative convenient expression of its variance estimate (compare with (A5)) is given by

$$v_p\left(t_{y(p)}\right) = \frac{n}{(n-1)}\sum_{i\in s}(y_i w_i - \overline{yw})^2 \quad \text{(A9)}$$

where $\overline{yw}$ denotes the simple sample average of $y_i w_i$. For SRS with replacement, $p_i = 1/N$; and the formula (A9) reduces to the familiar formula $N^2\frac{1}{n(n-1)}\sum_{i\in s}(y_i - \bar{y})^2$ which is equal to (A6) without the finite population correction $(N-n)$/N.

A.3 Two Stage Designs (without replacement of PSUs)

Consider a two stage probability sampling design with π_1 as the first stage design for selecting a sample s_1 of n_1 primary sampling units (PSUs), and π_2 as the second stage design for selecting a sample of n_2 elementary units. Suppose the properties of invariance and independence hold for the design. Invariance means that the design π_2 is specified in advance at the design stage and does not depend on which PSUs get selected in the first stage sample s_1, and, thus, is invariant to realized samples s_1. Independence means that the selection of second stage units within PSUs is independent from PSU to PSU. Further suppose that given π_1, we have unbiased estimates t_i of totals T_i for each selected PSU i with variance σ_i^2. Now, an unbiased estimate $t_{y(\pi_1\pi_2)}$ of the population total T_y is easily obtained as

$$t_{y(\pi_1\pi_2)} = \sum_{s_1} \frac{t_i}{\pi_{1i}} \tag{A10}$$

where π_{1i} denote the first stage selection probabilities. Next, if the first stage design is without replacement of PSUs, a natural step to obtain an unbiased estimate of variance of $t_{y(\pi_1\pi_2)}$ is to consider expectation of the expression (A5) for the variance estimate $v_{\pi_1\pi_2}(t_{y(\pi_1\pi_2)})$ when y_i is replaced by t_i under the joint randomization $\pi_1\pi_2$. Let π_{1ij} denote the joint inclusion probabilities under π_1 and observe that the variance of $t_{y(\pi_1\pi_2)}$ of (A10) is given by

$$V_{\pi_1\pi_2}\left(\sum_{s_1} \frac{t_i}{\pi_{1i}}\right) = E_{\pi_1} V_{\pi_2}\left(\sum_{s_1} \frac{t_i}{\pi_{1i}}\right) + V_{\pi_1} E_{\pi_2}\left(\sum_{s_1} \frac{t_i}{\pi_{1i}}\right) \tag{A11}$$

where
$$E_{\pi_1} V_{\pi_2}\left(\sum_{s_1} \frac{t_i}{\pi_{1i}}\right) = E_{\pi_1}\left(\sum_{s_1} \frac{\sigma_i^2}{\pi_{1i}^2}\right) = \sum_U \frac{\sigma_i^2}{\pi_{1i}} \tag{A12}$$

and from (A3),

$$V_{\pi_1} E_{\pi_2}\left(\sum_{s_1} \frac{t_i}{\pi_{1i}}\right) = V_{\pi_1}\left(\sum_{s_1} \frac{T_i}{\pi_{1i}}\right) = \frac{1}{2}\sum_{i\in U}\sum_{j\in U, j\neq i}(\pi_{1i}\pi_{1j} - \pi_{1ij})\left(\frac{T_i}{\pi_i} - \frac{T_j}{\pi_j}\right)^2 \tag{A13}$$

To estimate $V_{\pi_1\pi_2}(t_{y(\pi_1\pi_2)})$ unbiasedly, first consider a provisional estimator defined as

$$\acute{v}_{\pi_1\pi_2}(t_{y(\pi_1\pi_2)}) = \frac{1}{2}\sum_{i\in s_1}\sum_{j\in s_{1,j\neq i}}(\pi_{1i}\pi_{1j} - \pi_{1ij})\frac{1}{\pi_{1ij}}\left(\frac{t_i}{\pi_{1i}} - \frac{t_j}{\pi_{1j}}\right)^2 \tag{A14}$$

We have, $E_{\pi_1}E_{\pi_2}\left(\acute{v}_{\pi_1\pi_2}(t_{y(\pi_1\pi_2)})\right)$

$$= E_{\pi_1}\frac{1}{2}\sum_{i\in s_1}\sum_{j\in s_{1,j\neq i}}(\pi_{1i}\pi_{1j} - \pi_{1ij})\frac{1}{\pi_{1ij}}\left\{\left(\frac{T_i}{\pi_{1i}} - \frac{T_j}{\pi_{1j}}\right)^2 + \left(\frac{\sigma_i^2}{\pi_{1i}^2} + \frac{\sigma_j^2}{\pi_{1j}^2}\right)\right\}$$

$$= V_{\pi_1}\left(\sum_{s_1}\frac{T_i}{\pi_{1i}}\right) + \frac{1}{2}\sum_{i\in U}\sum_{j\in U, j\neq i}(\pi_{1i}\pi_{1j} - \pi_{1ij})\left(\frac{\sigma_i^2}{\pi_{1i}^2} + \frac{\sigma_j^2}{\pi_{1j}^2}\right)$$

$$= V_{\pi_1}\left(\sum_{s_1}\frac{T_i}{\pi_{1i}}\right) + \sum_{i\in U}\frac{\sigma_i^2}{\pi_{1i}^2}\sum_{j\in U, j\neq i}(\pi_{1i}\pi_{1j} - \pi_{1ij}).$$

$$= V_{\pi_1}\left(\sum_{s_1}\frac{T_i}{\pi_{1i}}\right) + \sum_{i\in U}\frac{\sigma_i^2}{\pi_{1i}^2}\{\pi_{1i}(n - \pi_{1i}) - (n-1)\pi_{1i}\}$$

$$= V_{\pi_1\pi_2}\left(\sum_{s_1}\frac{t_i}{\pi_{1i}}\right) - \sum_{i\in U}\sigma_i^2, \tag{A15}$$

which shows that the variance estimator $\acute{v}_{\pi_1\pi_2}(t_{y(\pi_1\pi_2)})$ underestimates $v_{\pi_1\pi_2}(t_{y(\pi_1\pi_2)})$ by $\sum_{i\in U}\sigma_i^2$. So with an unbiased estimator $\hat{\sigma}_i^2$ of σ_i^2 under the second stage design, bias can be corrected by adding $\sum_{s_1}\frac{\hat{\sigma}_i^2}{\pi_{1i}}$ to $\acute{v}_{\pi_1\pi_2}(t_{y(\pi_1\pi_2)})$ to obtain an unbiased variance estimate as

$$v_{\pi_1\pi_2}(t_{y(\pi_1\pi_2)}) = \acute{v}_{\pi_1\pi_2}(t_{y(\pi_1\pi_2)}) + \sum_{s_1}\frac{\hat{\sigma}_i^2}{\pi_{1i}} \quad . \tag{A16}$$

A.4 Two Stage Designs (with replacement of PSUs)

Analogous to A.2, suppose the first stage sample of size n_1 is with replacement of PSUs with draw by draw selection probability p_{1i} (denote $n_1 p_{1i}$ by π_{1i}), and the second stage is as in A.3 under design π_2 satisfying invariance and independence. The point estimator has a form similar to (A10) given by

$$t_{y(p_1\pi_2)} = \frac{1}{n_1}\sum_{s_1}\frac{t_i}{p_{1i}} = \sum_{s_1}\frac{t_i}{\pi_{1i}} \quad . \tag{A17}$$

In this case, $E_{p_1}V_{\pi_2}\left(t_{y(p_1\pi_2)}\right) = E_{p_1}\left(\frac{1}{n_1^2}\sum_{s_1}\sum_U p_{1i}\frac{\sigma_i^2}{p_{1i}^2}\right) = \frac{1}{n_1}\sum_U\frac{\sigma_i^2}{p_{1i}}$,

and $V_{p_1}E_{\pi_2}\left(t_{y(p_1\pi_2)}\right) = V_{p_1}\left(\frac{1}{n_1}\sum_{s_1}\frac{T_i}{p_{1i}}\right) = \frac{1}{n_1}\sum_U p_{1i}\left(\frac{T_i}{p_{1i}} - T_y\right)^2$.

Therefore,

$$V_{p_1\pi_2}\left(t_{y(p_1\pi_2)}\right) = \frac{1}{n_1}\sum_U p_{1i}\left(\frac{T_i}{p_{1i}} - T_y\right)^2 + \frac{1}{n_1}\sum_U\frac{\sigma_i^2}{p_{1i}} \tag{A18}$$

For variance estimation, denote $w_{1i} = \pi_{1i}^{-1}$, and similar to (A9), consider

$$\acute{v}_{p_1\pi_2}\left(t_{y(p_1\pi_2)}\right) = \frac{n_1}{(n_1-1)}\sum_{s_1}(t_i w_{1i} - \overline{tw_1})^2 =$$

$$\frac{1}{n_1(n_1-1)}\sum_{s_1}\left(\frac{t_i}{p_{1i}} - t_{y(p_1\pi_2)}\right)^2 = \frac{1}{n_1(n_1-1)}\frac{1}{2n_1}\sum_{i\in s_1}\sum_{j\in s_1, j\neq i}\left(\frac{t_i}{p_{1i}} - \frac{t_j}{p_{1j}}\right)^2 \tag{A19}$$

Now,

$$E_{p_1}E_{\pi_2}\left(\acute{v}_{p_1\pi_2}\left(t_{y(p_1\pi_2)}\right)\right) = \frac{1}{n_1(n_1-1)}\frac{1}{2n_1}E_{p_1}\left[\sum_{i\in s_1}\sum_{j\in s_1, j\neq i}E_{\pi_2}\left(\frac{t_i}{p_{1i}} - \frac{t_j}{p_{1j}}\right)^2\right] =$$

$$\frac{1}{n_1(n_1-1)}\frac{1}{2n_1}E_{p_1}\left[\sum_{i\in s_1}\sum_{j\in s_1, j\neq i}\left\{\left(\frac{T_i}{p_{1i}} - \frac{T_j}{p_{1j}}\right)^2 + \left(\frac{\sigma_i^2}{p_{1i}^2} + \frac{\sigma_j^2}{p_{1j}^2}\right)\right\}\right] =$$

$$\frac{1}{n_1(n_1-1)}\frac{1}{2n_1}\sum_{i\in s_1}\sum_{j\in s_1, j\neq i}E_{p_1}\left\{\left(\frac{T_i}{p_{1i}} - \frac{T_j}{p_{1j}}\right)^2 + \left(\frac{\sigma_i^2}{p_{1i}^2} + \frac{\sigma_j^2}{p_{1j}^2}\right)\right\} =$$

$$\frac{1}{n_1(n_1-1)}\frac{1}{2n_1}\sum_{i\in s_1}\sum_{j\in s_1, j\neq i}\left\{2V_{p_1}\left(\frac{T_i}{p_{1i}}\right) + 2\sum_U\frac{\sigma_i^2}{p_{1i}}\right\} = V_{p_1\pi_2}\left(t_{y(p_1\pi_2)}\right) \tag{A20}$$

It follows that $\acute{v}_{p_1\pi_2}\left(t_{y(p_1\pi_2)}\right)$ is unbiased for $V_{p_1\pi_2}\left(t_{y(p_1\pi_2)}\right)$ unlike the case of without replacement. In other words, the second stage variability in PSU estimates gets automatically embedded in the between PSU estimate variability. This is an amazing result in survey sampling that is commonly used in practice to obtain a convenient and conservative variance estimator. It is approximately valid whenever $n_1 \ll N_1$; i.e., the first stage sampling fraction is small, and in contrast to the without replacement case, does not require knowledge of second order inclusion probabilities π_{1ij} for the first stage (which are often unknown or tedious to compute), and also does not require estimates of second stage variances σ_i^2 which may be difficult to estimate. Thus, for both first and second stage designs, only first order selection probabilities are needed for unbiased point estimation and for a simplified variance estimation.

A.5 Single and Two Phase Designs

The designs considered so far were single phase in that the data collection on sampled units was performed only once at the final stage when there are two or more stages of selection. However, if data other than what are on the sampling frame are collected at both stages, then the design becomes two phase. Such designs are useful in practice when first phase information is used for second phase selection; e.g., first phase sample s_1 is stratified before selection in the second phase. In such situations, the condition of invariance is of course not satisfied because the second phase sample s_2 depends on what is observed in the first phase. However, if s_2 is selected independently across PSUs given s_1, then the condition of conditional independence is satisfied where conditional refers to given s_1. As introduced by Singh (2008), it is sufficient to have conditional independence and unbiasedness of PSU total estimates from the second phase sample for the applicability of the simplified variance estimate $v_{p_1\pi_2}\left(t_{y(p_1\pi_2)}\right)$ as an approximation to $v_{\pi_1\pi_2}\left(t_{y(\pi_1\pi_2)}\right)$. In fact, the weaker condition of lack of correlation than the condition of independence is sufficient in practice for variance estimation. In other words, under suitable conditions, simplified single phase variance estimation can be used for two phase designs. However, if the second phase estimates t_i's are conditionally unbiased but not conditionally independent, then it is easily shown that for the without replacement case in the first phase,

$$E_{\pi_1}E_{\pi_2}\left(\acute{v}_{\pi_1\pi_2}\left(t_{y(\pi_1\pi_2)}\right)\right) = V_{\pi_1\pi_2}\left(\sum_{s_1}\frac{t_i}{\pi_{1i}}\right) - \sum_{i\in U}\sigma_i^2 - \sum_{i\in U}\sum_{j\in U, j\neq i}\sigma_{ij} \qquad \text{(A21)}$$

where σ_{ij} is the conditional covariance between t_i and t_j, and $V_{\pi_1\pi_2}\left(\sum_{s_1}\frac{t_i}{\pi_{1i}}\right)$ has an extra term $\sum_{i\in U}\sum_{j\in U, j\neq i}\pi_{ij}\frac{\sigma_{ij}}{\pi_{1i}\pi_{1j}}$. However, the bias correction now has the additional term $\left(-\sum_{i\in U}\sum_{j\in U, j\neq i}\sigma_{ij}\right)$. In the with replacement case for the first phase, we have

$$E_{p_1}E_{\pi_2}\left(\acute{v}_{p_1\pi_2}\left(t_{y(p_1\pi_2)}\right)\right) = V_{p_1\pi_2}\left(t_{y(p_1\pi_2)}\right) - \sum_{i\in U}\sum_{j\in U, j\neq i}\sigma_{ij} \qquad \text{(A22)}$$

where $V_{p_1\pi_2}\left(t_{y(p_1\pi_2)}\right)$ now has an additional term $\left(1-\frac{1}{n_1}\right)\sum_{i\in U}\sum_{j\in U, j\neq i}\sigma_{ij}$. However, there was no bias before, but now it is $\left(-\sum_{i\in U}\sum_{j\in U, j\neq i}\sigma_{ij}\right)$.

References

Baker, R., et al. (2013). Summary Report of the AAPOR Task Force on Nonprobability Sampling (with comments). *Jour. Surv. Statist. Meth.*, 1, 96-143

Elliott, M.R. (2009). Combining data from probability and nonprobability samples using pseudo weights. *Surv. Prac.*, 2 (6).

Elliott, M.R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, 32(2), 249-264

Folsom, R.E. Jr. and Singh, A.C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. In: *ASA Proceedings, Surv. Res. Meth. Sec.*, pp. 598-603.

Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983). An evaluation of model-dependent and probability sampling inferences in sample surveys. *JASA*, 78, 776-793.

Haziza, D. and Beaumont, J-F. (2007). On the construction of imputation classes in surveys. *Int. Statist. Rev*. 75, 1, 25-43.

Little, R. J.A. (1988). Missing data adjustments in large surveys. *J. Bus. Econ. Statist*., 6, 287-296.

Raj, D. (1968). *Sampling Theory.* New York: Mcgraw-Hill

Robbins, M. W., Ghosh-Dastidar, B., and Ramchand, R. (2017). Blending of probability and convenience samples as applied to a survey of military caregivers. *Ann. Appl. Statist*. (to appear).

Royall, R.M. (1970). On finite population sampling theory under certain regression models. *Biometrika*, 57, 377-387.

Royall, R.M. (1976). The linear least squares prediction approach to two-stage sampling. *JASA*, 71, 657-664.

Särndal, C.-E. (1980). On π −inverse weighting versus best linear unbiased weighting in probability sampling. *Biometrika*, 67, 639-650.

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

Singh, A.C. (2015). Purposive Supplementary-sample integration for prediction of remainder for enhanced GREG. Proceedings of the Federal Committee on Statistical Methodology, US Census Bureau.
https://fcsm.sites.usa.gov/reports/research/2015-research/

Singh, A.C. (2008). Single phase simplified variance estimation approach to two phase-stage hybrid designs. *Proceedings of the American Statistical Association, Survey Research Methods Section,* 2501-2508

Singh, A.C., Iannacchione, V.G., and Dever, J.A. (2003). "Efficient estimation for surveys with nonresponse follow-up", Proceedings of the American Statistical Association, Section on Survey Research Methods, 3920-3930.

Stuart, E.A., Cole, S.R., Bradshaw, C.P., and Leaf, P.J. (2011). The use of propensity scores to assess the generalizability of results from randomized trials. *J.R. Stat. Soc. Ser A,* 174(2), pp. 369-386

Valliant, R. and Dever, J.A. (2011). Estimating propensity adjustments for volunteer web surveys, *Sociological Methods and Research*, 40, 105-137

Table I: Comparison of Prediction, Calibration Propensity, and *iGREG* Estimators

($A_y = 205.98$, $\widehat{N}_{w^*}^{-1} t_{yw^*} = 205.97$, $\tilde{n}^{-1} t_{y\tilde{u}} = 326.63$, $R = 1000$)

Estimator	Avg PE	RB	RRMSE
GREG(s̃) **(Benchmark)**	205.81	-.0008	.016
PRED (s̃)	181.05	-.121	.124
cPROP(s̃) or iGREG (PM)-s*	201.34	-.022	.026
ixGREG (PMN-r)- s*	209.93	.019	.04