Supplement-Sample Integration for Prediction of Remainder for Enhanced GREG

Avinash C. Singh
Division of Survey and Data Sciences
American Institutes for Research, Rockville, MD 20852
asingh@air.org

Abstract

With the growing demand for fit-for-purpose surveys to save cost and time by not using rigorous data collection methods for producing special subpopulation estimates, old issues of whether suitable inferences can be made from purposive or nonprobability samples are again at the forefront. For purposive samples, design-based methods are clearly not suitable. There is, however, the possibility of using model-based methods, but they assume the design to be non-informative (i.e., the model is assumed to hold for the sample) so that an optimal prediction of the unseen total (or the remainder) can be made from the seen under the model. This assumption is in general not tenable in practice. In addition to this concern, another concern with any model-based method is that it is subject to potential misspecification of the model mean resulting in bias even if the design is non-informative. To overcome these concerns, an alternative approach termed "model-over-design" (MOD) integration for a simplified problem is proposed under the joint design-model randomization when the purposive sample is available as a supplement to the core probability sample, although in practice it sometimes could be larger than the core sample. A design-based estimate such as a generalized regression estimator for the population total is first constructed using the probability sample, which uses the synthetic estimator based on the systematic part of the model mean containing fixed parameters, and then corrects it for the total model error corresponding to the random part of the model. Next, the above model-error correction is improved by using a model-based estimator from the additional seen observations in the purposive sample. We remark that while the initial probability sample is used for both estimation of model parameters to obtain a synthetic estimator and for estimation or prediction of the total model-error, the purposive supplement is only used to improve the model-error correction from the additional seen units. Under regularity conditions, the resulting estimator is consistent and its mean squared error can be estimated using Taylor linearization under the joint randomization of man-made probability sample design, nature-made purposive sample design, and the model for the finite population. Potential applications to NORC's AmeriSpeak Panel survey with opt-in or nonprobability supplements are briefly described. Considerations of MOD integration also lead to potential solutions to the problems of making inferences from a single purposive sample or from a single probability sample with high nonresponse.

Key Words: Fit-for-purpose samples; Informative designs; Joint design-model-based inference; Nonprobability or purposive samples; Probability samples; Selection bias

1. Introduction

There is a resurgence of interest and controversy among practitioners in the feasibility of making valid inferences from purposive or nonprobability samples in the 21st century, even though a similar controversy in the early 20th century was addressed in the fundamental paper by Neyman (1934), who emphasized the need of probability samples and randomization-based inference in survey sampling, and in the contributions to the theory of probability-based survey sampling in the early books by Hansen, Hurvitz, and Madow (1953) and Cochran (1953). The main reason for such a renewed interest in purposive samples is the desire to obtain more precise estimators than the commonly used design-based estimators, such as a generalized regression (GREG) estimator, when dealing with lower-level geographies or small subpopulations. This is a very practical problem that arises in using low-cost big data (such as administrative data, registries, and other extant data) and data from fit-for-purpose surveys that do not adhere to rigorous probability sampling protocols in design and data collection as an alternative to the costly option of increasing the sample size of traditional probability surveys.

In this paper, use of the term "purposive sample" is preferred over the term "nonprobability sample" because the nonprobability sample (to be denoted by s^*) can be perceived as a conceptual nature-made probability sample (in contrast to the man-made sample design) with unknown selection probabilities π_k^*'s for units k in the target universe U; here π_k^* can be 0 for units omitted on purpose leading to undercoverage of U, and is likely to be strictly less than 1 in the case of self-selection due to unit nonresponse, which is indistinguishable from noncontact with the sampling unit. The recent American Association for Public Opinion Research (AAPOR) Task Report (Baker et al., 2013) shows clearly the conundrum in using purposive samples for the following reasons. On the one hand, use of purposive samples is rather attractive as it promotes use of low-cost extant data or other data such as internet opt-in panel data to obtain more detailed information about small subpopulations and specialized domains. On the other hand, there is the conceptual problem in its representativeness of the target universe resulting in biased estimates, and the lack of any reasonable randomization framework for measuring precision of resulting estimates without making strong untestable assumptions. In this paper, we attempt to provide a solution by first reviewing the assumptions underlying the two basic principled approaches to inference from probability samples in surveys—design-based using the probability sample s given the target universe U (Hansen and Hurvitz, 1943; Narain, 1951; Horvitz and Thompson, 1952; Särndal, 1980) and model-based given the particular probability sample (s) as in Royall (1970, 1976) and Valliant, Dorfman, and Royall (2000). We then propose a solution for a simplified problem in which the purposive sample (s^*) serves as a supplement to the core probability sample s rather than the problem of making inference from s^* alone.

The main contribution of the proposed approach for integrating s and s^* can be summarized as follows. For efficient estimation, models are often used to incorporate auxiliary information from multiple sources such as administrative data, censuses, and related sample surveys. In the context of using models for estimating finite population quantities such as totals, models refer to superpopulation models governing selection of the finite population under consideration. As is commonly done in a model-based approach, the assumed model for the finite population is taken as a linear regression model. Now with s^*, there are obvious concerns about representativeness and selection bias since the underlying nature-made random mechanism for s^* is unknown. To address these concerns and in the interest of avoiding strong model assumptions and possible bias due to model misspecification, we first propose to use only the probability sample (s) to estimate fixed model parameters. The estimated regression parameters are then used to obtain the synthetic estimator; i.e., total of the systematic part of the model for the study variable Next, given the regression parameters, instead of simply using the (weighted) observed model errors from s for estimating the total model error (this is the random part of the model) as in the case of the commonly used GREG estimator of Särndal (1980), we propose to combine it with the additional (unweighted) observed model errors provided by s^*. The underlying premise is that although s^* may not be deemed fit for estimating fixed model parameters due to its selection bias, it does provide valid information about model errors from additional observed units, which can be beneficially used for efficiency gains (i.e., variance reduction) under a suitable joint randomization framework for the man-made probability sample design (π), nature-made purposive sample (π^*), and the postulated model (ξ) for the finite population.

Thus, the proposed approach starts with a design-based estimator (such as GREG) using the core probability sample s, which for large samples has the desirable asymptotic design consistency (ADC) property for robustness against possible model misspecifications. It then improves its efficiency without increasing the sample size by integrating the model-based estimator of the total model error from the purposive supplementary sample s^* under the joint randomization. It relies only on s (and not on s^*) for any adjustments for biases due to noncoverage or nonresponse but takes advantage of s^* for variance efficiency. This approach, termed in this paper as the "model-over-design integration" or MOD-I, builds model-based enhancements over the design-based approach. The term "integration" signifies that it uses ideas from both design-based and model-based approaches. It uses a nonoptimal combination, on purpose, of the two estimates of the total model error so that it can be robust to model misspecification by maintaining the ADC property of the basic design-based estimator GREG. In other words, the main reason for the preference of a nonoptimal combination is to avoid overshrinkage of the design-based estimator of the total model error to the model-based estimator, which is based on somewhat tenuous assumptions. Overshrinkage could happen because the model-based estimator of the total model error from s^* tends to have much smaller variance than the design-based estimator from s due to the absence of design weights. Besides, the nonoptimal combination allows for the new estimator to have an expansion form involving one set of final weights that can be used for other study variables as well.

We remark that although the MOD-I method does not provide a solution to the original inference problem from a single purposive sample s^*, it does provide a solution to a simplified version of the original problem by assuming that s^* is available as a supplement to s even though in practice it could sometimes be larger than s. For the simplified problem, there are other methods proposed in the literature that blend s^* and s. Elliott (2009) provides an innovative approach using propensity score modeling to obtain pseudo-weights for $s^*$where s is used as the control group and s^* as the treatment group. Another innovative approach is due to DiSogra et al. (2011), who use a dual frame approach and sampling weight calibration methods where an initial weight of 1 is assigned to s^*. Although these are among the few serious attempts to address the challenging but important practical problem of blending s and s^*, the underlying assumptions seem difficult to justify. In all these papers and as is the case in this paper, s^* is conceptually treated as a probability sample with an unknown random selection mechanism.

The organization of this paper is as follows. Section 2 provides background and motivation of the proposed approach. In particular, the two basic approaches of design-based and model-based for estimation in survey sampling with a single probability sample are first reviewed in detail in order to motivate the proposed approach and consider some variants used later on for integration with the purposive sample. To this end, we make two basic assumptions (C1 and C2) for unbiased point estimation, a third assumption C3 for variance estimation, and a fourth C4 for a simplified variance estimation along with the regularity conditions needed for the asymptotic behavior of Horvitz-Thompson type estimators in survey sampling (see Fuller, 2009; Section 1.3). For asymptotics, we assume the probability sample size n and the population size N go to ∞ such that n/N goes to 0, but the purposive sample size n^* remains bounded. Note that n^* is random in general. The assumptions C1-C4 are:

> *C1: The model mean is correctly specified, but other aspects such as the model covariance structure may not be.*
> *C2: Given covariates, the model errors ε_k's are uncorrelated with the conceptual selection probabilities π_k^* of units in the target universe U that could be selected in the purposive sample. (A similar assumption for the probability sample design π can be made unless the model is enlarged to include π_k's as a new covariate.)*
> *C3: The model covariance structure is correctly specified.*
> *C4: Given covariates, the products $(\varepsilon_k\varepsilon_l's)$of model errors corresponding to pairs of units in U are uncorrelated with the corresponding conceptual joint selection probabilities π_{kl}^* in the purposive sample. A similar assumption for π is made.*

Assumptions C1 and C3 are the usual first two moment assumptions to specify a semiparametric model for the finite population. Assumption C2 is much weaker than the non-informative design assumption. It may be deemed to be satisfied in general because the design π^* for s^* is nature-made. Therefore, π_k^*'s are expected to be functions of unit covariates or unit profile, and not as complex as in the case of the man-made design π for s. Thus, C2 would be valid if the model already includes suitable covariates that are expected to govern nature's random mechanism for selection of s^*. It is probably reasonable to expect that covariates that are good predictors of the study variable y are also good predictors of the inclusion probabilities π_k^* under the nature-made design. Assumptions C4 along with C2 capture the essence of non-informative designs in order to study the first and second order properties of estimators under the model.

In Section 3, we consider how the two estimates (one each from s and s^*) of the total model error can be combined under the joint randomization of the superpopulation model (ξ), the known probability sample design π for s, and the unknown random design π^* for s^*. Note that under this joint framework, the two estimates can be made (approximately) unbiased for the total model error—the common finite population parameter, which makes it convenient to compare the new estimator in terms of variance efficiency without the burden of bias considerations. The appendix shows how suitable variance estimates of all estimators considered can be obtained under the joint random mechanism. (Incidentally, since the population total parameter is random, it is customary to use the term "mean squared error (MSE) even when the estimator is unbiased. However, we prefer to use the term "variance" to distinguish it from MSE when the estimator may be biased.) Analogous to GREG, the proposed estimator can be expressed in an expansion form due to the use of a nonoptimal combination, and the original auxiliary control totals for GREG continue to be satisfied by the new set of weights. However, unlike the case of dual frame samples, the final estimator is not a calibration estimator in the strict sense because there are no suitable initial weights that can be attached to the purposive sample.

The problem of subpopulation or domain estimation is considered in Section 4. Here MOD-I is especially expected to be useful because GREG may not be reliable due to insufficient domain sample size but can be made so in combination with a purposive sample from the domain of interest. The question of bias-variance trade-off in MOD-I is considered in Section 5. An interesting finding is that if C2 is not satisfied, the contribution of sample selection bias in s^* is relatively negligible in the model-based estimator of the total model error and the corresponding MSE estimator. However, this is not the case if C1 does not hold. In this context, limiting shrinkage of the design-based estimator toward the model-based estimator of the total model error becomes crucial. Finally, Section 6 contains summary and remarks about how MOD-I considerations can lead to potential solutions to the problems of making inferences from a single purposive sample or from a probability sample with high nonresponse.

2. Background and Motivation

As mentioned in the introduction, purposive surveys such as fit-for-purpose surveys being in demand by users for time and cost efficiency do not follow a rigorous probability sampling design protocol. It is therefore difficult to obtain theoretically justifiable point estimates and their standard errors from such survey data without making strong modeling assumptions. However, with purposive supplements to a core probability sample, it is possible to make suitable inferences about the population under consideration. The proposed method is motivated from the two basic approaches to estimation that form the foundation of survey sampling inference. These are design-based and model-based approaches. In the design-based approach, one relies on the likely behavior of sample estimates under the man-made random mechanism π of probability sampling from a given finite target population *U*. On the other hand, in the model-based approach, given a sample, one relies on the likely behavior of sample estimates under the nature-made random mechanism ξ governing the creation of the target population from a conceptual infinite universe or a superpopulation.

A commonly used design-based method GREG for estimating population totals consists of first obtaining an estimate of the fixed part under a model (i.e., the synthetic part) and then correcting it by adding an estimate of the random part (i.e., model error) given by a weighted estimator from observed errors in the sample. The model postulated here is a regression model for predicting the outcome of interest by auxiliary variables. The synthetic estimator of the population total is simply the sum of model predictions based on the systematic part for each individual in the population. The synthetic predictions require known values of auxiliaries and estimates of regression coefficients in the model mean function. The regression coefficients are estimated by solving weighted estimates of census estimating functions (EFs), where weights refer to inverse of individual selection probabilities in the sample, and census EFs are usual quasi-likelihood EFs when the sample is the full finite population. The resulting estimator (GREG) is natural to consider in connection with a model-based estimator (to be denoted by PRED as in Brewer, 2002, signifying the prediction approach of Royall, 1970, 1976) because both use models to start with. Here, unlike mainstream statistics, the parameters of interest are not model parameters, but the finite population totals involving fixed and random effects or model errors. In the following, we first review GREG followed by PRED in some detail for a single probability sample because it lays down the necessary theoretical foundation for the proposed estimator. For interesting comparisons of design-based and model-based approaches, see Hansen et al. (1983) and Little (2004).

2.1 Design-based Approach

Specifically, consider a linear model ξ for y_k with covariates $(x_{ik})_{1\le i\le p}$ for the kth unit, $1 \le k \le N$, given by

$$\xi: y_k = x_k'\beta + \varepsilon_k\ ,\ \varepsilon_k \sim_{ind} (0, \sigma_\varepsilon^2 c_k) \tag{1}$$

where β is a p-vector of regression coefficients, c_k's are known constants and N is the finite population size. We will assume for convenience that β is known initially, but later on we will substitute it with a design-weighted estimator as in GREG based on the probability sample s of size n under design π. The synthetic estimator of T_y is then given by

$$t_{y,syn}(\beta) = T_x'\beta \tag{2}$$

where $T_x = \sum_U x_k$ and the finite population total parameter T_y is similarly $\sum_U y_k$; the summation notations $\sum_U y_k$ and $\sum_{k\in U} y_k$ will be used interchangeably. The design bias of the synthetic estimator is $T_x'\beta - T_y$ or $-\sum_U \varepsilon_k$ where $y_k - x_k'\beta = \varepsilon_k$. The GREG estimator corrects this bias (which is simply minus the total model error) by using a design-based unbiased estimator such as Horvitz-Thompson, or HT for short. It is given by

$$\textbf{GREG: } t_{y,grg}(\beta) = T_x'\beta + \sum_{k\in s} \varepsilon_k w_k \tag{3a}$$

$$= t_{yw} + \beta'(T_x - t_{xw}) \tag{3b}$$

where the design weight $w_k = \pi_k^{-1}$, π_k is the sample inclusion probability of unit k, and t_{yw}, for example, is $\sum_s y_k w_k$, the HT-estimator. With known β, $t_{y,grg}(\beta)$ as an estimate of T_y is design, or π −unbiased, and is also π −consistent (or ADC) as n, N get large under general regularity conditions (see the asymptotic framework of Isaki and Fuller, 1982; the book by Fuller, 2009; Section 1.3; and also Kott, 2009); i.e., with high π −probability, it is close to the true value T_y. Here and in what follows, all the asymptotic properties are with respect to the mean estimator (such as $N^{-1}t_{y,grg}$) of the population mean $N^{-1}T_y$. It is interesting and important to remark that even if the model mean function is misspecified, the GREG estimator remains ADC; i.e., under π −randomization, $N^{-1}\left(t_{y,grg}(\beta) - T_y\right) = N^{-1}(\sum_s \varepsilon_k w_k - \sum_U \varepsilon_k) = o_p(1)$. This robustness property of GREG is desirable in practice because, as is well known, no model is perfect. Note that the model does play an important role in GREG for improving its relative efficiency over HT estimators. However, the model's validity is not vital for its ADC property, and hence GREG is also referred to as model-assisted. For the proposed method for combining s and s^*, we also strive for the ADC property analogous to GREG.

In practice, the regression parameters are replaced by weighted estimators motivated by census EFs, where all the population totals are replaced by HT estimators to obtain sample EFs; see Binder (1983) and also Särndal (1980). In particular, the census EFs for β are given by

$$\sum_{k\in U} x_k\,(y_k - x_k'\beta)/c_k = 0 \tag{4a}$$

and the corresponding sample EFs are given by

$$\sum_{k\in s} x_k\,(y_k - x_k'\beta)w_k/c_k = 0 \tag{4b}$$

It is easily seen that

$$\hat{\beta}_w = (\textstyle\sum_s x_k\, x_k' w_k/c_k)^{-1}(\sum_s x_k\, y_k w_k/c_k) = (X'C^{-1}WX)^{-1}X'C^{-1}Wy \tag{5}$$

where $W = diag(w_k)_{1\le k\le n}$, $C = diag(c_k)_{1\le k\le n}$, and X is the $n \times p$ matrix of the sample covariate values x_k's. Here, the main reason for using sampling weights in (4b) is to make the sample EF $\pi\xi$ −unbiased for 0, because C2 for π may not be satisfied. The estimator $\hat{\beta}_w$ is optimal under the joint $\pi\xi$ −randomization as defined by Godambe and Thompson (1986). However, under π −randomization given ξ, it is not optimal in the usual sense; i.e., the regression coefficient $\hat{\beta}_w$ does not correspond to optimal regression in the sense of minimizing the $\pi|\xi$ −variance of the regression estimator about T_y. Although, GREG with $\hat{\beta}_w$ (to be denoted by $t_{y,grg}$ instead of $t_{y,grg}\left(\hat{\beta}_w\right)$) is no longer unbiased, it remains asymptotically design unbiased as well as ADC under general conditions (see Robinson and Särndal, 1983). It is also in general more efficient than the HT estimator in view of the observation that the model residuals $e_{k,grg} = y_k - x_k'\hat{\beta}_w$ tend to be less variable than y_k's, and $t_{y,grg}$ yields perfect estimates (i.e., with no error) of totals of covariates x_k's when y_k is replaced by x_k's. (Note that if β is not estimated, the residual $y_k - x_k'\beta$ can be denoted by $e_k(\beta)$, which will be identical to the model error ε_k if the model mean is not misspecified. This distinction is useful in discussion on bias-robustness in Section 6.) The above property of GREG being perfect for estimating T_x is easily seen from the calibration form of the GREG estimator as introduced by Deville and Särndal (1992) and is given by

$$t_{y,grg} = \textstyle\sum_{k\in s} y_k w_k a_{k,grg}\,, \quad a_{k,grg} = 1 + x_k' c_k^{-1}\hat{\eta}_{grg} \tag{6}$$

where $\hat{\eta}_{grg} = (X'C^{-1}WX)^{-1}(T_x - t_{xw})$. Observe that the sample x_k-values inflated or deflated by the weight adjustments $(a_{k,grg})_{1\le k\le n}$ satisfy the auxiliary control totals T_x exactly. Moreover, denoting the predicted value $x_k'\hat{\beta}_{gr}$ by $\hat{y}_k$, the weighted estimator $\sum_s \hat{y}_k w_k$ using predicted values matches exactly with the direct estimator $\sum_s y_k w_k$ whenever the unit vector $1_{n\times 1}$ is in the column space of $C^{-1}X$—an important special case being when c_k is one of the $x_k{}'s$ (see Appendix A1). Equivalently, the weighted sum of residuals $\sum_s e_{k,grg} w_k$ becomes zero under the above condition on covariates. The built-in benchmarking property of GREG residuals to sum to zero when the unit vector is in the column space of X (commonly satisfied in practice) is attractive for robustification to possible model misspecifications. With respect to the precision of GREG, the $\pi\xi$ −variance of $t_{y,grg}$ about T_y can be approximated well for large samples by the $\pi|\xi$ −variance using the Taylor linearization or delta method (see Appendix A2).

2.2 Model-based Approach

So far, we considered a design-based estimator GREG, which for large samples has desirable properties of ADC in that it remains close to the true population total with high probability and is robust to model misspecification in that it remains ADC even if the model is misspecified. The alternative model-based estimator PRED (defined below) uses an unweighted estimator of regression coefficients in the model for corresponding predictions of the systematic part in the model mean function for each individual in order to construct a synthetic estimator of the population total. Analogous to GREG, it then corrects it by adding an estimate of the total model error by using an unweighted estimator from observed errors in the sample—it only corrects the total model error corresponding to the seen units in s. Thus, unlike the design-based estimator GREG, the model-based estimator PRED does not rely on sampling weights because it considers the likely behavior of the estimate given a particular observed sample.

PRED: We now consider in some detail the model-based estimator PRED proposed by Royall (1970, 1976), which uses the prediction approach for estimating model errors under ξ given π; i.e., given the sample s. The formulation of the PRED estimator will be useful for integrating information about the additional seen units from s^* because the observed sample under the model-based approach is not required to have a known probability sample design. Given β, the PRED estimator of T_y is given by

$$t_{y,prd}(\beta) = \sum_{k\in s} y_k + \sum_{k\in U\setminus s}(x_k'\beta + 0) \tag{7}$$

where the first sum on the right is the sum of the observed y −values from the seen units, and the second sum is the predicted value under the model for the remainder or unseen units; i.e., the set $U \setminus s$ of units from the population U that were not selected in s. The $x_k'\beta$ term in the second sum on the right is the predictor of the fixed part (or the model mean) in the unknown y_k under the model, and 0 signifies the best linear unbiased predictor (BLUP) of the model error ε_k for the unseen because all the error terms are uncorrelated. If the error terms ε_k's were correlated, then BLUP of ε_k for the unseen could have been improved by using the observed values of ε_k's for the seen units in the sample. The estimator $t_{y,prd}(\beta)$ can alternatively be expressed as

$$t_{y,prd}(\beta) = (T_x - \sum_{k\in s} x_k)'\beta + \sum_{k\in s} y_k \tag{8a}$$
$$= T_x'\beta + \sum_{k\in s} \varepsilon_k \tag{8b}$$

which looks very similar to the expression (3a) for GREG except that the predictions for model errors in the sample are not weighted. Note that in the case of GREG, the predicted value of the remainder is taken as $\sum_{U\setminus s} x_k'\beta + (\sum_s \varepsilon_k w_k - \sum_s \varepsilon_k)$. The weighted sum of model errors, or residuals $\sum_s \varepsilon_k w_k$ used in GREG under π −randomization, provides an unbiased adjustment (through the commonly used HT estimator) for the design bias $(-\sum_U \varepsilon_k)$ in the synthetic estimator $T_x'\beta$, while the unweighted sum $\sum_s \varepsilon_k$ used in PRED under ξ −randomization provides an unbiased prediction (optimal under the model) of the total model error $\sum_U \varepsilon_k$.

In the discussion so far, the parameters β were assumed to be known. In practice, they are unknown and are estimated differently in PRED from GREG. Under GREG, $\hat{\beta}_w$ is based on weighted sample EFs, which, in turn, give rise to several desirable properties as mentioned earlier, including the ADC of GREG when C1 holds but C2 may not. Under PRED, however, the regression parameters are estimated by

$$\hat{\beta}_u = (\sum_s x_k x_k'/c_k)^{-1}(\sum_s x_k y_k/c_k) = (X'C^{-1}X)^{-1}X'C^{-1}y \tag{9}$$

which is derived from best linear unbiased EFs under the model and does not involve design weights. Under ξ −randomiztion given π and general regularity conditions, the PRED estimator with $\hat{\beta}_u$ (to be denoted by $t_{y,prd}$) has desirable properties in that it is unbiased, consistent, and optimal (in the sense of minimum variance) if the model holds for the sample.

Interestingly, analogous to the GREG expression (6), PRED can also be expressed as an expansion estimator with adjustment factors $a_{k,prd}$ but without design weights w_k. We have,

$$t_{y,prd} = \sum_{k\in s} y_k a_{k,prd}\ , \quad a_{k,prd} = 1 + x_k' c_k^{-1} \hat{\eta}_{prd} \tag{10}$$

where $\hat{\eta}_{prd} = (X'C^{-1}X)^{-1}(T_x - t_{xu})$, and t_{xu} is the unweighted sample sum $\sum_s x_k$. In general, if the variance of the model error is heteroscedastic, the adjustment factor $a_{k,prd}$ depends on it because $\hat{\beta}_u$ does. Therefore, unlike GREG, the weight adjustment factor may vary with the outcome variable y. A useful way to interpret (10) is in the sense of calibrating the initial weights of 1 in s by the adjustment factor $a_{k,prd}$ such that exact totals T_x are reproduced when y is replaced by x. However, the expression (10) of $t_{y,prd}$ is not strictly a calibration estimator in the sense of Deville and Särndal (1992) because $(T_x - t_{xu})$, the difference of the vector of population totals and the corresponding sample sums on which the weight adjustment factor depends, is not a zero function vector; i.e., its expectation is not zero under $\xi|\pi$ −randomization. This implies that the known totals T_x are not truly calibration control totals. A $\xi|\pi$ −variance estimate of $t_{y,prd}$ about T_y is provided in Appendix A3.

The fundamental assumptions underlying the model-based approach are that the model is correctly specified for the population (C1 and C3 corresponding to the first two moments are sufficient for our purpose), and the sampling design is non-informative for the model. Here the randomization is with respect to the ξ −distribution conditional on the sample design π. The non-informative design assumption requires that the joint distribution of the outcome variable in the population given the auxiliaries does not depend on the random variables indicating inclusion or exclusion of population units in the sample. In fact, it is sufficient to assume C2 and C4 for our purpose. However, even the weaker set of assumptions is quite strong and is generally not expected to be satisfied in practice because it is not feasible to include all key design variables (that govern inclusion of units in the sample) in the model as auxiliaries that are deemed to be correlated with the study variable. The main reason is that the man-made sampling design can be quite complex in that, besides stratification and disproportionate sample allocation, samples within strata may be drawn in stages with varying selection probabilities of clusters of units at any given stage depending on the size variable in the interest of over- or under-sampling of special domains. Even in situations where important design variables could be included in the model, the covariate totals needed for prediction with linear models might not be available for design variables; e.g., such totals are usually not known for non-selected clusters in multistage designs. Besides, if the model of interest is nonlinear, as is often the case with discrete variables, use of model-based prediction requires even more detailed information such as the unit-level information for all the covariates in the population. This problem does not arise with GREG because the role of model is secondary, and therefore, even for discrete variables, one can use linear models, although it is not strictly correct because the range restrictions on model means and errors imposed by nonlinear models are not satisfied.

If the design is informative due to C2 not being satisfied, there is design bias (also known as selection bias) in the model-based estimator even though for units in the finite population U, the model mean is correctly specified; i.e., C1 holds. It is possible to correct this problem by including π_k as a covariate in the ξ −model, but still the model may not hold for s because the model covariance structure for the sampled units may not be correctly specified. Incidentally, with the inclusion of the new covariate π_k, the model (1) changes with new β −parameters and the model errors ε's, but this change does not invalidate the original model because the old model mean is marginal of the new model mean. (Note also that with π_k as a covariate, we don't need to know these for all units in U for computing the synthetic estimator under PRED as it is sufficient to know $\sum_U \pi_k$, which is n for fixed sample design or $E_\pi(n)$ for random sample designs and which can be estimated by n.) Besides the above problem of selection bias, there may be model bias due to misspecification of the model mean. The above two concerns (biases due to informativeness of the design and due to model misspecification) for probability samples get magnified with purposive samples because the underlying conceptual sampling design (π^*) for the purposive sample is not even known. Nevertheless, a good understanding of the implications of model and design assumptions on model-based estimators is important for finding

a suitable solution to the problem of integrating $s^*$with s. The main reason for this is that the model-based methods do not inherently require knowledge of the underlying probability sample design.

2.3 Motivation for Integration of Design-based and Model-based Approaches

In view of the desirable ADC property of GREG making it robust to model misspecification, our goal is to preserve the ADC property of GREG while integrating it with the model-based estimator PRED. The ultimate goal is to increase its efficiency for population total estimation in general and for subpopulation or domain estimation in particular, which suffer from the problem of insufficient number of observations. With this in mind, from expressions (3a) and (8b) for GREG and PRED, respectively, it is observed that if common values of the β −parameters are used in both estimators, then the synthetic estimates for the two become identical, but we have two different estimates of the same total model error. So it may be possible to improve the prediction of the total model error $\sum_U \varepsilon_k$ by combining the two estimates under $\pi\xi$ −randomization. This is the underlying premise of the proposed integration of ideas from design-based and model-based approaches, which is quite different from the usual combination of two estimators under either a design-based $(\pi|\xi)$ or a model-based approach $(\xi|\pi)$. It is introduced in the next section and termed "model-over-design integration" (MOD-I) because it starts with GREG—a design-based estimator as the basic estimator and then improves its prediction of the random part by bringing over the PRED-type estimator of the random part.

With the above motivation, we first construct a new estimator termed "prediction of remainder for enhancing generalized regression" (PREG for short), which uses the design-based synthetic estimator of GREG, but the model-based estimator of the total model error from PRED modified by using $\hat{\beta}_w$ in place of $\hat{\beta}_u$. Note that the estimator $\hat{\beta}_w$ is preferable to $\hat{\beta}_u$ for reasons mentioned earlier. Thus, the PREG estimator (to be denoted by $t_{y,prg}$) is defined as

$$\textbf{PREG:} \qquad t_{y,prg} = T_x'\hat{\beta}_w + \sum_{k\in s} e_{k,grg} \tag{11}$$

Clearly, the only difference between GREG and PREG is that PREG uses unweighted residuals. Analogous to (6), the expansion form of PREG is given by

$$t_{y,prg} = \sum_{k\in s} y_k w_k a_{k,prg}, \quad a_{k,prg} = \pi_k + x_k' c_k^{-1}\hat{\eta}_{prg} \tag{12}$$

where $\hat{\eta}_{prg} = (X'C^{-1}WX)^{-1}(T_x - t_{xu})$. An estimator of the variance of $t_{y,prg}$ about T_y is given in Appendix A4.

Having now PREG in addition to GREG, it is natural to ask how to combine the two estimates of the total model error to obtain a new estimate that is more efficient than GREG. Here, we prefer a nonoptimal combination that diminishes the influence of PREG in order to avoid potential biases of PREG. To this end, we first assume C1; i.e., while the full model with the mean and covariance structure could be misspecified, the model mean is at least correctly specified. Specifically, $E_\xi((y_k - x_k'\beta)|x_k) = 0$, so that $\sum_s \varepsilon_k$ has a chance to be unbiased for $\sum_U \varepsilon_k$ under the joint $\pi\xi$ −randomization. In other words, we want $E_{\pi\xi}((\sum_s \varepsilon_k - \sum_U \varepsilon_k)\,|x_k, 1 \le k \le N) = 0$. However, this may not be true unless C2 for π is satisfied for the sample; i.e.,

$$E_{\pi\xi}\big((\textstyle\sum_{\mathrm{U}} \varepsilon_k\, 1_{k\in s} - \sum_U \varepsilon_k\,|x_k, 1 \le k \le N)\big) = E_\xi((\textstyle\sum_{\mathrm{U}} \varepsilon_k\,\pi_k - \sum_U \varepsilon_k\,|x_k, 1 \le k \le N) = 0 \tag{13a}$$

where $\sum_{\mathrm{U}} \varepsilon_k\, 1_{k\in s} = \sum_s \varepsilon_k$. The above condition holds if $E_{\pi|\xi}(1_{k\in s}|\varepsilon_k, x_k)$ does not depend on y_k through ε_k; i.e., given x_k, the selection probability π_k does not depend on ε_k. In other words,

$$E_\xi(\varepsilon_k\pi_k|x_k) = E_\xi(\varepsilon_k|x_k)E_\xi(\pi_k|x_k) \tag{13b}$$

so that C2 holds for π. This will be the case if π_k's are functions of x_k's, which is unlikely but can be easily satisfied by enlarging the model to include π_k's as values of an extra covariate. Now, with the enlarged model, both $\sum_s \varepsilon_k$ and $\sum_s \varepsilon_k w_k$ are $\pi\xi$ −unbiased for $\sum_U \varepsilon_k$, and therefore, it makes it possible to combine the two under a common randomization scheme without the burden of accounting for bias. Incidentally, to satisfy C2, introduction of π_k (a design-specific feature) as a covariate may seem somewhat an artifact to reach a specific goal because the sampling

design refers to the finite population and not to the superpopulation, although it may nevertheless serve as a good covariate in its own right.

Above considerations will also pave the way for using s^* in improving estimators from s because the unbiasedness of model-based estimators does not require knowledge of the random mechanism under a probability sample as long as C2 holds. In fact, as mentioned in the introduction, C2 is likely to hold for s^* without introducing π_k^*'s in the model as another covariate because the nature-made design π^* is not expected to be as complex as the man-made design π. This anticipated property of π^* is the basis for defining another estimator termed "supplement-sample for PREG estimation" (S-PREG for short and denoted by $t_{y,spg}$) needed for MOD-integration of s^* and s, and is given by

$$\textbf{S-PREG:} \qquad t_{y,spg} = T_x'\hat{\beta}_w + \sum_{k\in s^*} e_{k,grg} \tag{14}$$

Letting $t_{xu^*} = \sum_{s^*} x_k$, the expansion form of the S-PREG estimator is given by

$$t_{y,spg} = \sum_s y_k w_k a_{k,spg} + \sum_{s^*} y_k, \quad a_{k,spg} = x_k' c_k^{-1}\hat{\eta}_{spg} \tag{15}$$

where $\hat{\eta}_{spg} = (X'C^{-1}WX)^{-1}(T_x - t_{xu^*})$. An estimator of the variance of $t_{y,spg}$ about T_y under the joint $\pi^*\pi\xi$ −randomization is given in Appendix A5. In the next section, we consider the problem of integrating two samples—s with the supplement s^*; i.e., how to integrate the two estimators of the total model error from GREG and S-PREG for improving the GREG efficiency. With s and s^*, it is tempting to combine the three estimators of the total model error corresponding to GREG, PREG, and S-PREG, respectively, but $\pi\xi$ −unbiasedness of PREG requires enlarging the model in order to satisfy C2 for π, which, in turn, requires knowledge of π_k's for units in s^*, and this may not be available for all units (see Section 6 for more comments). A summary of all estimators (new and old) considered in this paper is presented in Table 1.

3. MOD-Integration of a Purposive Supplement to a Probability Sample

For MOD-I, the conditions C1 and C3 for ξ and C2 and C4 for π^* are assumed to hold as mentioned in the introduction. The validity of C2, unlike the case of the probability sample s, seems quite plausible because the individual characteristics that govern the nature-made design π^* for self- or purposive selection of an individual from U may be known to the analyst, and are likely to be included as covariates in the model because they typically will be deemed to be correlated with the outcome variables of interest. In Section 5, the impact on bias and variance due to departures from the above conditions is considered. The sampling designs for $s^*$and s are assumed to be independent, and, therefore, in general, there may be an overlap between the two. The new predictor $\sum_{s^*}\varepsilon_k$ used in S-PREG of the total model error based on the new seen units in s^* can be used to improve the total model error prediction from GREG; this time, however, under the joint $\pi^*\pi\xi$ −randomization. We can now define the proposed estimator under MOD-I, termed "supplement-sample for integrated PREG" (SI-PREG for short and denoted by $t_{y,sig}$), as follows:

$$\textbf{SI-PREG:} \quad t_{y,sig} = (1-\lambda_{sig})t_{y,grg} + \lambda_{sig}t_{y,spg} \tag{16a}$$

$$= T_x'\hat{\beta}_w + \sum_s e_{k,grg}w_k + \lambda_{sig}\left(\sum_{s^*} e_{k,grg} - \sum_s e_{k,grg}w_k\right) \tag{16b}$$

where the coefficient λ_{sig} is obtained in a nonoptimal manner for stability and for obtaining an expansion form of the estimator. (Incidentally, an optimal choice of λ_{sig} can be obtained by minimizing the variance of $t_{y,sig} - T_y$, which is given by minus the optimal regression coefficient of $(\sum_s \varepsilon_k w_k - \sum_U \varepsilon_k)$ on $(\sum_{s^*}\varepsilon_k - \sum_s \varepsilon_k w_k)$.) We remark that for estimation of the total $\sum_U \varepsilon_k$ through regression, the estimator $\hat{\beta}_w$ can be treated as fixed because the fixed parameters β and random parameters ε_k's are distinct. For nonoptimal regression in SI-PREG, we use anticipated variances and covariances (Isaki and Fuller, 1982) about T_y under the joint $\pi^*\pi\xi$ −randomization. Thus, this integration of the two estimators is nonoptimal because λ_{sig} is obtained under the working assumption that the model holds for both samples. This is analogous to the assumption used in an alternate derivation of GREG using nonoptimal regression (weighted SRS-type variances and covariances) of t_{yw} on $(T_x - t_{xw})$ in estimating β by $\hat{\beta}_w$ (see Singh, 1996). Thus, the coefficient λ_{sig} can be obtained as

$$\lambda_{sig} = \breve{v}_{grg}/(\breve{v}_{grg} + \breve{v}_{spg}) \tag{17}$$

where $\breve{v}_{grg}$ denotes a working variance estimate of GREG, assuming β is given and later substituted by $\hat{\beta}_w$, and $\breve{v}_{spg}$ is defined similarly. We have from Appendix A6,

$$\breve{v}_{grg} = \hat{\sigma}^2_{\varepsilon w} \sum_s w_k (w_k - 1) c_k \, , \breve{v}_{spg} = \hat{\sigma}^2_{\varepsilon w} (\sum_s c_k w_k - \sum_{s^*} c_k) \, , \tag{18}$$

where $\hat{\sigma}^2_{\varepsilon w} = \sum_s e^2_{k,grg} w_k c_k^{-1} / \sum_s w_k$. The anticipated covariance of GREG and S-PREG given β is zero because of the unbiasedness of GREG and independence of $s^*$and s. The expansion form of the SI-PREG estimator is given by

$$t_{y,sig} = \sum_s y_k \, w_k a_{k,sig} + \lambda_{sig} \sum_{s^*} y_k, \quad a_{k,sig} = \big(1 - \lambda_{sig}\big) a_{k,grg} + \lambda_{sig} a_{k,spg} \tag{19}$$

where λ_{sig} is $\sum_s c_k w_k (w_k - 1) / (\sum_s c_k w_k^2 - \sum_{s^*} c_k)$, and $a_{k,grg}$ and $a_{k,spg}$ are given by (6) and (15), respectively. The new set of adjusted weights given by (19) continue to satisfy the GREG calibration controls because the corresponding adjusted weights for GREG and S-PREG satisfy the controls in view of the fact that residuals $e_{k,grg}$ become zero when y_k is replaced by one of the covariates from x_k. We remark that the final weights $w_k a_{k,sig}$'s are only defined for the sample s and not for both samples, unlike the usual case of combining two probability samples because the second sample s^* being purposive has no initial weights for adjustment. Therefore, the SI-PREG is not a true calibration estimator in the sense of Deville and Särndal (1992). An estimate of the variance of $t_{y,sig}$ about T_y under the joint $\pi^* \pi \xi$ −randomization is given in Appendix A7.

The above expansion form of SI-PREG is convenient for the univariate case; i.e., when there is only one new predictor $(\sum_{s^*} \varepsilon_k - \sum_s \varepsilon_k w_k)$ corresponding to the study variable y. However, for the multivariate extension of SI-PREG when y is multivariate—i.e., for the case of several key study variables—it is of interest to produce one set of final adjusted weights. Now we have a vector of new predictors of the form $(\sum_{s^*} \varepsilon_k - \sum_s \varepsilon_k w_k)$ corresponding to each element of y. A new SI-PREG estimator can be constructed using all the extra predictors for further gains in efficiency. The regression coefficient λ_{sig} for the nonoptimal combination in the multivariate case will now be replaced by a matrix, each row of which consists of non-diagonal elements as covariances with the other study variables corresponding to each of the study variables and from which the value y_k of the study variable of interest can be factored out. Thus, unlike (19) where λ_{sig} is not used for factoring out y_k, here we take the standard calibration approach in constructing a new set of final weights that can be used for all study variables besides the key variables already used in defining new predictors of the total model error.

In contrast to (19), the above alternative way of constructing the final set of expansion weights amenable to the multivariate case is now shown for the univariate case for simplicity. Here, even though the factor $\hat{\sigma}^2_{\varepsilon w}$ is common in the numerator and the denominator of λ_{sig}, we do not cancel it out as its presence in the numerator allows for an expansion form of the estimator SI-PREG, somewhat analogous to a calibration estimator. To see this, observe that the numerator of $\hat{\sigma}^2_{\varepsilon w}$ can be alternatively expressed as $\sum_s y_k \, e_{k,grg} w_k c_k^{-1}$ because

$$\begin{aligned} \sum_s e^2_{k,grg} w_k c_k^{-1} &= \sum_s \big(y_k - x_k' \hat{\beta}_w\big) e_{k,grg} w_k c_k^{-1} \\ &= \sum_s y_k \, e_{k,grg} w_k c_k^{-1} - \hat{\beta}_w' \sum_s x_k \, e_{k,grg} w_k c_k^{-1} \end{aligned} \tag{20a}$$

and the last term with the negative sign is zero as the EFs for β evaluated at $\hat{\beta}_w$ are zeros. Therefore, the value y_k of the study variable of interest can be factored out from the regression coefficient λ_{sig} to obtain an expansion form of SI-PREG with a different set of adjustment factors $\tilde{a}_{k,sig}$, as shown below.

$$t_{y,sig} = \sum_s y_k \, w_k \tilde{a}_{k,sig}, \quad \tilde{a}_{k,sig} = a_{k,grg} + e_{k,grg} c_k^{-1} (\textstyle\sum_s w_k)^{-1} \hat{\zeta}_{sig} \tag{20b}$$

$$\hat{\zeta}_{sig} = \lambda_{sig} \, \hat{\sigma}^{-2}_{\varepsilon w} \big(\textstyle\sum_{s^*} e_{k,grg} - \sum_s e_{k,grg} w_k\big) \tag{20c}$$

We remark that, as desired, the new set of adjusted weights $w_k \tilde{a}_{k,sig}$'s continue to satisfy the GREG calibration controls because $\sum_s y_k \, w_k e_{k,grg} c_k^{-1}$ is zero when y_k (not in $e_{k,grg}$ though) is replaced by one of the covariates from x_k, and therefore, the contribution from the adjustment in $\tilde{a}_{k,sig}$ beyond $a_{k,grg}$ is zero. Here the adjusted weights are only defined for the sample s. Extra information from the second sample s^* is used in the form of the predictor

$(\sum_{s^*} \varepsilon_k - \sum_s \varepsilon_k w_k)$ for regression analogous to the predictor $(T_x - t_{xw})$ in GREG, and appears in the adjustment factor $\tilde{a}_{k,sig}$.

We also note that the coefficient λ_{sig} is expected to be between 0 and 1 because $\sum_s c_k w_k$ estimates $\sum_U c_k$, which is larger than $\sum_{s^*} c_k$. This property of a convex combination is attractive for ease in interpretation. Thus, λ_{sig} behaves like a shrinkage factor in that high values of λ_{sig} imply that the design-based predictor $\sum_s \varepsilon_k w_k$ is shrunk more to the model-based predictor $\sum_{s^*} \varepsilon_k$. In practice, it may be preferable to have λ_{sig} not more than 1/2 so that GREG can dominate over S-PREG in the SI-PREG formulation in the interest of robustness to model misspecifications. However, under general conditions, we have $\breve{v}_{grg} = O_p(N^2/n)$, and $\breve{v}_{spg} = O_p(N)$, which imply that λ_{sig} will tend to be close to 1 because $\breve{v}_{spg}$ is of much lower order than $\breve{v}_{grg}$. The practical implication of this is clearly not desirable even though S-PREG tends to be more efficient than GREG if C1-C4 hold (see Section 5). It is probably better to have only moderate gains in efficiency over GREG in the interest of robustness to model misspecifications and selection bias.

With the above observation in mind and in the spirit of working variances and covariances used in the specification of λ_{sig} to achieve a certain objective, we first inflate $\breve{v}_{spg}$ by N/n^γ $(0 < \gamma < 1$; e.g., $\gamma = 1/2)$ so that the product is $O_p(N^2/n^\gamma)$, with the order being larger than the order of $\breve{v}_{grg}$. Next we introduce a constraining factor ψ (between 0 and 1 but bounded away from 0; e.g., greater than .01), choice of which is based on other practical considerations mentioned below. This way, $\lambda_{sig} \to 0$ as $n, N \to \infty$, which will imply ADC of the new estimator. Therefore, as a modification to SI-PREG, we define another estimator termed SI-PREG-constrained (or SI-PREG(c) for short and denoted by $t_{y,sig(c)}$) as follows:

$$\textbf{SI-PREG(c):}\ t_{y,sig(c)} = \big(1 - \lambda_{sig(c)}\big) t_{y,grg} + \lambda_{sig(c)} t_{y,spg} \tag{21a}$$

$$= T_x' \hat{\beta}_w + \sum_s e_{k,grg} w_k + \lambda_{sig(c)} \big(\sum_{s^*} e_{k,grg} - \sum_s e_{k,grg} w_k\big) \tag{21b}$$

where the specification of $\lambda_{sig(c)}$ is quite similar to that of λ_{sig} by (17), except that $\breve{v}_{spg}$ in the denominator is multiplied by a constraining factor $\psi_{sig(c)}(N/n^\gamma)$. That is,

$$\lambda_{sig} = \breve{v}_{grg} / \big(\breve{v}_{grg} + \psi_{sig(c)} (N/n^\gamma)\, \breve{v}_{spg}\big), \tag{22}$$

where $0 < \gamma < 1$, and $\psi_{sig(c)}$ is set between 0 and 1 but bounded away from 0 such that SI-PREG(c) has a reasonable improvement in precision over GREG but the point estimate itself is not too far off from GREG. For this purpose, we follow Efron and Morris's (1972) suggestion on limiting over-shrinkage of empirical Bayes estimators in small area estimation as a guideline. In particular, we propose to choose $\psi_{sig(c)}$ such that it does not make $t_{y,sig(c)}$ lie outside the interval defined by boundaries $t_{y,grg} \pm \tilde{\breve{v}}_{grg}^{1/2}$, where $\tilde{\breve{v}}_{grg}$ denotes the variance estimate under the model as given in A2. Note that we need $\psi_{sig(c)}$ away from 0 in order to keep SI-PREG(c) not too far from GREG. Now, the expansion form of $t_{y,sig(c)}$ is similar to $t_{y,sig}$ except that in (19), $a_{k,sig}$ is replaced by $a_{k,sig(c)}$ defined in an analogous manner. An estimate of the variance of $t_{y,sig(c)}$ about T_y under the joint $\pi^*\pi\xi$ −randomization can be obtained as in Appendix A7 after λ_{sig} is substituted by $\lambda_{sig(c)}$.

4. An Enhancement of MOD-Integration for Domain Estimation

The method of MOD-I is expected to be especially useful in estimation for small or specialized domains that may not be well represented in the full sample, and hence the need for a purposive supplement with only a marginal additional cost. A common example of domains in practice is given by socio-demographic subgroups that partition the total population *U* into nonoverlapping subpopulations but are not strata, and therefore the sample size for each domain is random. The standard domain estimators using GREG are defined by replacing y_k in (6) by $y_k 1_{k \in U_d}$ where U_d denotes the *d*th domain, $1 \le d \le D$, and *D* being the total number of domains. Now, in order to improve precision of domain-level GREG, we can easily obtain domain-level SI-PREG by modifying (16) and (19) suitably. However, precision of such domain-level SI-PREG estimators obtained using the standard theory of domain estimation could be improved if we use full sample (i.e., combined sample over all domains) to estimate fixed parameters $\beta, \sigma_\varepsilon^2$, and λ_{sig} rather than separately for each domain. In other words, for these parameters, we use the same estimators as in the case of regular

SI-PREG estimators for population totals and not subpopulations or domains, but everywhere else we multiply y_k, x_k (therefore, e_k) by $1_{k\in U_d}$ to get their contributions only for the domain of interest. It follows that for SI-PREG of domains, although the effective domain sample size based on the combined s and s^* remains the same, we could make the resulting estimators more stable (and hence more precise) due to less variability in the estimates of fixed parameters $\beta, \sigma_\varepsilon^2$, and λ_{sig} needed for their computation.

The above enhancement of MOD integration is along the lines of enhancing stability of GREG estimators for domains in the context of small area estimation where the full sample estimator $\hat{\beta}_w$ is used for regression parameters (see e.g., Singh and Mian, 1995; Rao, 2003; Section 2.5), but domain-level auxiliary totals T_{xd} and the domain-level HT-estimator t_{xdw} in the calibration form (6) are used to obtain $t_{yd,grg}$; i.e., GREG for domain d. (Here for some x-variables, T_{xd} could be at the population and not subpopulation level.) This increases the computational burden for obtaining more stable domain-level GREG estimators in the above manner because the GREG calibration weights will need to be computed now for each domain separately, unlike the customary GREG with one set of final weights for all study variables. Thus, the proposed enhancement of SI-PREG for domains starts with the enhanced GREG for domains and improves it further by integrating it with domain-specific purposive samples. We can now define domain-specific estimators GREG(d) and S-PREG(d) in order to define SI-PREG(d) denoted respectively by $t_{y,grg(d)}$, $t_{y,spg(d)}$, and $t_{y,sig(d)}$ as follows:

$$\textbf{GREG(d):}\ t_{y,grg(d)} = \sum_{k\in s} y_k w_k a_{k,grg(d)}\ ,\ a_{k,grg(d)} = 1_{k\in U_d} + x_k' c_k^{-1} \hat{\eta}_{grg(d)} \tag{23}$$

where $\hat{\eta}_{grg(d)} = (X'WC^{-1}X)^{-1}(T_{xd} - t_{xdw})$. Note that the GREG(d) calibration weights satisfy the domain-specific control totals T_{xd}. Moreover, unlike the usual GREG for domains, even if $1_{n\times 1}$ is in the column space of $C^{-1}X$, the weighted sum of residuals $\sum_s e_{k,grg} w_k 1_{k\in U_d}$ is no longer zero.

$$\begin{aligned}\textbf{S-PREG(d):}\ & t_{y,spg(d)} = \sum_s y_k w_k a_{k,spg(d)} + \sum_{s^*} y_k 1_{k\in U_d},\\ & a_{k,spg(d)} = x_k' c_k^{-1} \hat{\eta}_{spg(d)},\ \ \hat{\eta}_{spg(d)} = (X'WC^{-1}X)^{-1}(T_{xd} - t_{xdu^*}).\end{aligned} \tag{24}$$

The t_{xdu^*} estimator is defined analogous to t_{xu^*} except that it uses the domain subsample.

$$\begin{aligned}\textbf{SI-PREG(d):}\quad & t_{y,sig(d)} = \sum_s y_k\, w_k a_{k,sig(d)} + \lambda_{sig} \sum_{s^*} y_k 1_{k\in U_d}\\ & a_{k,sig(d)} = (1-\lambda_{sig}) a_{k,grg(d)} + \lambda_{sig} a_{k,spg(d)}.\end{aligned} \tag{25}$$

Note that the domain-level control totals T_{xd} continue to be satisfied by the SI-PREG(d) expansion weights as desired. The SI-PREG(d)-constrained (denoted by SI-PREG(dc)) estimator can be defined in an analogous manner by replacing λ_{sig} by $\lambda_{sig(c)}$, common for all domains. Estimates of variance of the above estimators about T_{yd} can be easily obtained from previous formulas for full population-level estimators by replacing $e_{k,grg}$ by $e_{k,grg} 1_{k\in U_d}$ but retaining full sample estimates for $\beta, \sigma_\varepsilon^2$, and λ_{sig}.

5. Bias and Variance Trade-Off in MOD-Integration Methods

Under the MOD-I approach, the bias and variance of SI-PREG about T_y depends on the bias and variance of GREG and S-PREG as estimators of T_y. While GREG is ADC and asymptotically design unbiased of T_y under $\pi|\xi$ and hence under $\pi^*\pi\xi$ without requiring any extra conditions, we do need C1 and C2 for asymptotic unbiasedness of S-PREG about T_y under $\pi^*\pi\xi$. Although GREG's asymptotic unbiasedness is robust to departures from C1 and C2, S-PREG is not, but it tends to be more precise than GREG under C1 and C2. This can be explained using simplified expressions using the concept of anticipated variances and covariances under the additional assumptions of C3 and C4 (see (A2.4) and (A5.3) for anticipated variance expressions of GREG and S-PREG, respectively, and Appendix A8). As noted in Section 3, this is the property from the efficiency perspective that lends support to allowing GREG to shrink more toward S-PREG by letting λ_{sig} be close to 1 in the definition of SI-PREG. However, for fear of biases (model misspecification or sample selection) that arise if C1 or C2 does not hold, it is preferable to limit the shrinkage factor λ_{sig}. In this section, we consider the order of magnitude of the relative bias squared (i.e., bias squared divided by variance) in order to check the seriousness of the impact of bias on SI-PREG under the following two scenarios.

Under scenario one where C1 holds but not C2, it follows from (A5.1) that for S-PREG,

$$t_{y,spg} - T_y \approx \left(\sum_s \varepsilon_k a_k(\eta_{spg})\ w_k + \sum_{s^*} \varepsilon_k - \sum_U \varepsilon_k \left(a_k(\eta_{spg})\ + \pi_k^*\right)\right) + \sum_U \varepsilon_k \left(a_k(\eta_{spg})\ + \pi_k^* - 1\right), \quad (26)$$

which implies that the asymptotic bias $E_\xi\left(\sum_U \varepsilon_k \left(a_k(\eta_{spg})\ + \pi_k^* - 1\right)\right)$ under $\pi^*\pi\xi$ is at most $O(\sqrt{N})$ because $E_\xi\left(\sum_U \varepsilon_k \left(a_k(\eta_{spg})\ - 1\right)\right) = 0$ as ε_k has mean 0 , and $E_\xi(\sum_U \varepsilon_k \pi_k^*) \leq O(\sqrt{N})$ (see Appendix A9) since $\varepsilon_k \pi_k^*$ does not have mean 0 if C2 fails. The relative bias square is $O(n/N)$ because variance of $t_{y,spg}$ is $O(N^2/n)$. In fact, under an additional mild assumption, $E_\xi(\sum_U \varepsilon_k \pi_k^*)$ is only $O(1)$, in which case the relative bias square is of even lower order. Thus, S-PREG is bias-robust to departures from C2 since the relative bias squared goes to zero as $n/N \to 0$ under the given asymptotic framework, where $n, N \to \infty$, and n^* remains bounded. It follows that SI-PREG is even less affected by the above bias due to the introduction of the shrinkage factor λ_{sig}.

Under scenario two where C1 does not hold, the bias problem gets more serious because the relative bias squared does not go to zero. Note that the question of the validity of C2 does not even arise if C1 does not hold. When the postulated model mean is not correctly specified, the limit in probability under $\pi\xi$ of $\hat{\beta}_w$ is no longer β but some other value to be denoted by $\tilde{\beta}$. Also let $\tilde{\varepsilon}_k$ denote the new residual $(y_k - x_k'\tilde{\beta})$, which is not the true residual $(y_k - \mu_k)$ $(= \varepsilon_k)$ where μ_k is the unknown mean under the true model ξ. We now have

$$t_{y,spg} - T_y \approx \left(\sum_s \tilde{\varepsilon}_k a_k(\eta_{spg})\ w_k + \sum_{s^*} \tilde{\varepsilon}_k - \sum_U \tilde{\varepsilon}_k \left(a_k(\eta_{spg})\ + \pi_k^*\right)\right) + \sum_U \tilde{\varepsilon}_k \left(a_k(\eta_{spg})\ + \pi_k^* - 1\right), \quad (27)$$

which implies that the bias is $O(N)$ because even $E_\xi\left(\sum_U \tilde{\varepsilon}_k \left(a_k(\eta_{spg})\ - 1\right)\right) \neq 0$. It follows that the point estimator S-PREG is not robust to departures from C1 because the relative bias squared is now $O(n)$.

The above bias-variance trade-off analysis shows that the violation of C2 is of little consequence for SI-PREG compared to the violation of C1. In practice, in the absence of any substantive evidence about the validity of C1, it is probably safe to limit the contribution of S-PREG in SI-PREG by constraining λ_{sig} to $\lambda_{sig(c)}$ as proposed in Section 3. It is also important to develop diagnostics for checking model validity analogous to small area estimation using internal and external evaluation (see e.g., Rao, 2003; Section 7.1.4). However, diagnostics for unit-level models as is the case here that take the complex design into account need to be further developed; see Graubard and Korn (2009) for some innovative ideas.

6. Summary and Remarks

In this paper, before dealing with the problem of integrating a purposive supplement s^* to a probability sample s, a new approach termed MOD-I was first proposed in the case of a single probability sample s for integrating the two traditional approaches to survey sampling—design-based and model-based. Under the joint $\pi\xi$ −randomization, MOD-I starts with a design-based estimator GREG (which being model-assisted is conducive for integration) and then borrows the prediction idea of the model-based estimator PRED to create a new estimator PREG, which is made up of the synthetic part from GREG and the random part from PRED but with unweighted GREG residuals. The new estimator PREG as an alternative to GREG turns out to be the key for the proposed MOD integration and various MOD-I estimators summarized in Table 1.

Conditions C1 (for validity of the model mean) and C2 for the π −design (for lack of correlation between the model error and the selection probability π_k given the auxiliaries) are needed for approximate $\pi\xi$ −unbiasedness of PREG. Now, without having the burden of accounting for bias, the two estimators GREG (which does not require C1 and C2 for its approximate unbiasedness) and PREG can be integrated to obtain a more efficient estimator that can be termed "integrated-PREG" or I-PREG. The term integration in MOD-I is used to distinguish from the customary term of composition of two estimators because it deals with two completely different random mechanisms—one of the estimators is design-based and the other model-based. In addition, unlike the usual composition, the two estimators GREG and PREG have common synthetic parts but different estimates of the random part or the total model error. Thus, MOD-I starts with GREG and then improves its prediction of the remainder by using PREG; hence the use of

the term "model-over-design." Although the estimators PREG and I-PREG are quite important in their own right for improving GREG estimation from a single sample s, and for motivating the new estimators S-PREG and SI-PREG for integrating s^* with s, they were not considered in this paper for integrating with the supplement purposive sample. The reason for this was the potential need to enlarge the model with the selection probabilities π_k's as a new covariate in order to satisfy C2, which, in turn, would require knowledge of π_k's for s^*, and this may not be available in practice for all units.

The question of bias-variance trade-off in SI-PREG was also considered. An interesting finding was that unlike PREG, the estimator S-PREG, which uses prediction of model errors based only on s^*, is robust to failure of C2 for π^* because of bias being negligible under general regularity conditions. In fact, C2 might be deemed to be satisfied for π^*due to the design being nature-made unlike the man-made design π. However, in contrast to the robustness of GREG, failure of C1 introduces a serious bias in S-PREG and hence the need of SI-PREG and SI-PREG(c) arises in order to dampen the impact of bias. In practice, for the goal of increasing efficiency of GREG using the supplement s^*, we clearly need to rely on C1 and be willing to trade some bias with higher precision as in small area estimation. However, suitable model diagnostics (see, e.g., Graubard and Korn, 2009) should be performed and measure should be taken to limit the risk of bias by using SI-PREG(c) which is robust by construction for large n as it shares the ADC property of GREG. In approximating the asymptotic variance of SI-PREG, it was observed that all the parts of the contributions of the s^* sample involving the design π^*can be either neglected relative to other higher-order terms or can be approximated in a conservative sense. This turns out to be fortunate for our application due to π^*being unknown.

In situations where π_k's happen to be available for s^*, and hence can be used as an extra covariate in the model, SI-PREG can be further improved by using another estimator to be termed "total-sample integrated PREG" (or TI-PREG and denoted by $t_{y,tig}$), which can be defined as follows. In this regard, we first define the total-sample PREG (or T-PREG denoted by $t_{y,tpg}$) analogous to S-PREG.

$$\textbf{T-PREG:}\quad t_{y,tpg} = T_x'\hat{\beta}_w + \sum_{k\in s\cup s^*} e_{k,grg} \tag{28}$$

The expansion form of the T-PREG estimator can be obtained as in (15) for S-PREG and its variance as in A5. The MOD-I version of T-PREG is given by

$$\textbf{TI-PREG:}\quad t_{y,tig} = \left(1-\lambda_{tig}\right)t_{y,grg} + \lambda_{tig}t_{y,tpg} \tag{29}$$

where λ_{tig} is defined similar to (17). The expansion form of TI-PREG can be easily obtained along the lines of (19) for SI-PREG and its variance as in A7. The constrained version TI-PREG(c) can also be defined in a manner similar to SI-PREG(c). Note that in the interest of reducing bias due to failure of C2 for π, we could introduce π_k's in the model as a covariate and then use $\hat{\beta}_u$ instead of $\hat{\beta}_w$ in defining T-PREG. However, use of $\hat{\beta}_w$ allows for a common β −estimator for GREG and T-PREG needed for defining TI-PREG. Now for the case of a single sample s, the new estimator I-PREG (which can be denoted by $t_{y,ipg}$) mentioned above can also be easily defined like SI-PREG, except that $t_{y,spg}$ is replaced by $t_{y,prg}$ and λ_{sig} by its natural analogue λ_{ipg}.

It was observed that the MOD-I estimators have several useful features such as they continue to satisfy GREG calibration controls; have expansion forms, although, strictly speaking, are not quite calibration estimators; can have a multivariate form with several new predictors corresponding to each key study variable; and have linearized variance estimators under $\pi^*\pi\xi$ −randomization. For the important practical application in domain estimation, which is where the integration of s and s^* is likely to be most needed, domain-level SI-PREG(d) were defined along the same lines as GREG(d), which is different from the usual GREG due to the use of full sample estimates of regression parameters, common for all domains, in the interest of stability. The SI-PREG method is expected to have immediate applications to the NORC AmeriSpeak initiative, which uses a household panel selected by probability sampling from NORC's National Sample Frame—a large, nationally representative, first phase sample based on a stratified multistage unequal probability design. AmeriSpeak households are invited to participate in research studies approximately two to three times a month. The probability sample of households in the panel is supplemented by nonprobability samples for studies targeting low-incidence subpopulations (see www.amerispeak.org for more information).

It may be of interest to note that although the MOD-I methodology developed in this paper is for integrating a purposive supplement s^* with a core probability sample s, it does suggest a new approach to dealing with a purposive sample s^* alone without the benefit of having a core sample s providing information about the same set of study or outcome variables as in s^*. Suppose there is an alternative probability sample $\tilde{s}$ representative of the population or subpopulation of interest for the purposive sample with information about the auxiliary variables common with s^* but not about the outcome variables of interest. In such situations, a common approach is to attach sampling weights from $\tilde{s}$ to the purposive sample by matching methods based on propensity scores, where propensity refers to the probability that an individual in the population can be selected in s^*(treatment group) in contrast to $\tilde{s}$ (control group). As an alternative, S-PREG developed for MOD-I could be used for point estimation that requires only sample weighted estimates of regression parameters from $\tilde{s}$ and GREG residuals from s^*. Therefore, we need a suitable value of $\hat{\beta}_w$ from $\tilde{s}$. To this end, we propose a heuristic solution, which, however, requires further investigation. First, we obtain $\hat{\beta}_{u^*}$ from $s^*$and then compute prediction scores (these are just predictive means $x_k'\hat{\beta}_{u^*}$) for all units in $\tilde{s}$ and s^*. Now impute the y-values for all units in $\tilde{s}$ using a method such as predictive mean matching based on prediction scores to find the $y-$value of the matched unit in s^* serving as the donor dataset, which in practice is typically large. The estimate $\hat{\beta}_{u^*}$ based on s^* is likely to be biased because the model mean may not hold for s^*, but they may be deemed to be adequate as an initial estimator. Now find a revised estimate $\hat{\beta}_w$ of β from $\tilde{s}$ using imputed $y-$values. Next, compute a revised set of prediction scores for all units in $\tilde{s}$ and s^* using the current $\hat{\beta}_w$, and repeat the imputation and estimation cycle until convergence. For variance estimation, although the usual linearization method seems intractable, we can use a suitable replication method to create replicate subsamples of $\tilde{s}$, and then find the corresponding replicate $\hat{\beta}_w$ for each subsample. A variance estimate can now be obtained easily from S-PREG replicate estimates.

Finally, we remark that for probability surveys with high nonresponse, the important robustness feature of GREG to misspecifications of the (study or outcome) model becomes questionable due to its strong dependence on the response model for weight adjustments of the respondent subsample. The reason for this is that, like the outcome model ξ, the response model could also be misspecified. If the model ξ covariates are identical to the covariates in the response model, and if the response model can be approximated well by a linear model, although different from the linear model ξ, then the weight adjustment in GREG can be interpreted as a nonresponse adjustment (Folsom and Singh, 2000; Kott, 2006). Moreover, GREG has double protection against bias in that it is approximately unbiased if either outcome model mean does not hold but the response model mean holds or vice-versa (see Kim and Park, 2006; Kott and Liao, 2015). However, GREG may not have adequate precision. On the other hand, if C1 and C2 (after introducing π_k as a new covariate) hold, then PRED is attractive due to its unbiasedness because it tends to be more efficient and does not, in principle, require selection probabilities under a response model. Therefore, with PRED or its related version PREG, we can work directly with the respondent subsample without requiring any modeling for nonresponse. However, if C1 is misspecified, it could be seriously biased. In such situations, I-PREG(c) may provide a useful efficient alternative to GREG and a somewhat robust alternative to PREG. This estimator also needs to be investigated further.

Acknowledgment

I would like to thank Kirk Wolter for organizing and leading an NORC brainstorming day in the fall of 2012 on "Probability and Nonprobability Surveys Fusion" under a strategic initiative that provided the necessary impetus for this research. I would also like to thank Dan Kasprzyk for his valuable support and encouragement, and Mike Dennis and Vicki Pineau for discussions on potential practical applications. This research was partially funded by an NORC CESR grant and the NORC Working Paper Series programs. A short version of this paper is in the 2015 AAPOR-SRMS proceedings.

Table 1: Summary of Estimators – Old and New
(design-based, model-based, MOD-based and MOD-integration-based)

Notation	Acronym and Full Form	Description
	Old Estimators	
$t_{y,grg}$	GREG—Generalized Regression	Design-based, robust to model misspecification
$t_{y,prd}$	PRED—Prediction Approach	Model-based, optimal under certain assumptions, not robust to model failures
	MOD-Estimators (without Integrated Prediction of the Total Model Error)	
$t_{y,prg}$	PREG—Prediction of Remainder for Efficient Generalized Regression	Uses estimation of the synthetic part as in GREG, but estimation of the random part as in PRED
$t_{y,spg}$	S-PREG—Supplement Sample PREG	Uses estimation of the synthetic part as in GREG, but estimation of the random part from the supplement sample similar to PRED
$t_{y,tpg}$	T-PREG—Total Sample PREG	Uses estimation of the synthetic part as in GREG but estimation of the random part from the total (core and the supplement) sample similar to PRED
	MOD-I Estimators (with Integrated Prediction of the Total Model Error)	
$t_{y,ipg}$	I-PREG—Integrated PREG	Uses an integration factor for a convex linear combination of GREG and PREG
$t_{y,ipg(c)}$	I-PREG(c)—I-PREG Constrained	Constrains the integration factor so that I-PREG lies within one standard error of GREG
$t_{y,sig}$	SI-PREG—Supplement Sample-Based I-PREG	Integrates GREG and S-PREG
$t_{y,sig(c)}$	SI-PREG(c)—SI-PREG Constrained	Constrains so that SI-PREG lies within one standard error of GREG
$t_{y,tig}$	TI-PREG—Total Sample-Based I-PREG	Integrates GREG and T-PREG
$t_{y,tig(c)}$	TI-PREG(c)—TI-PREG Constrained	Constrains so that TI-PREG lies within one standard error of GREG

Note: For domain estimation, use a natural extension (d) of the above notation to obtain, for example, $t_{y,grg(d)}$, $t_{y,spg(d)}$, $t_{y,sig(d)}$, and $t_{y,sig(dc)}$. Estimates for fixed parameters (regression coefficients, model error variance, and the integration factor) are chosen to be common for all domains in the interest of stability of domain-level estimators.

Appendix (Technical Results)

The variance or MSE estimation results presented below are based on conditions C1-C4 and general regularity conditions for the asymptotic behavior of HT-type estimators; see, e.g., Fuller (2009, Section 1.3) and Kott (2009).

A1: $\sum_s \boldsymbol{e_k}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{w}})\boldsymbol{w_k} = \boldsymbol{0}$ if $\boldsymbol{1}_{n\times 1}$ is in the column space of $\boldsymbol{C^{-1}X}$

It follows that there exists a $p \times 1$ vector of constants τ such that $C^{-1}X\tau = 1_{n\times 1}$, which implies that $X\tau = C1_{n\times 1}$. Since $\hat{\beta}_w$ satisfies $X'C^{-1}W(y - X\beta) = 0$, we have

$$\tau'X'C^{-1}W\left(y - X\hat{\beta}_w\right) = 0 \text{ or } 1'CC^{-1}W\left(y - X\hat{\beta}_w\right) = 0. \quad \text{(A1.1)}$$

A2: $\widehat{\boldsymbol{Var}}_{\boldsymbol{\pi\xi}}(\boldsymbol{t_{y,grg}} - \boldsymbol{T_y})$ or $\boldsymbol{v_{grg}}$

By Taylor linearization of $t_{y,grg}$ about T_y under $\pi|\xi$, we have

$$t_{y,grg} - T_y \approx \sum_s \delta_{k,grg} w_k - \sum_U \varepsilon_k, \quad \delta_{k,grg} = \varepsilon_k a_k(\eta_{grg}) \quad \text{(A2.1)}$$

where $a_k(\eta_{grg})$ (=1+$x_k' c_k^{-1}\eta_{grg}$) is $a_{k,grg}$ of (6) but with $\hat{\eta}_{grg}$(= $(X'C^{-1}WX)^{-1}(T_x - t_{xw})$) replaced by the limit in probability denoted by η_{grg}, which can be interpreted as a coverage bias model parameter. It is 0 if there is no coverage bias, in which case $a_k(\eta_{grg})$ is 1. However, it helps to improve the variance estimator. We have

$$Var_{\pi\xi}\left(t_{y,grg} - T_y\right) = E_\xi V_{\pi|\xi}\left(\sum_s \delta_{k,grg} w_k\right) + E_\xi\left(\sum_U \delta_{k,grg} - \sum_U \varepsilon_k\right)^2. \quad \text{(A2.2)}$$

The first term on the right can be estimated by standard design-based methods after substitution of β and η_{grg} by $\hat{\beta}_w$ and $\hat{\eta}_{grg}$, and the second term can be estimated by $\hat{\sigma}^2_{\varepsilon w}(\sum_s\left(a_{k,grg} - 1\right)^2 w_k\, c_k)$. Thus,

$$v_{grg} = \tilde{v}_{grg} + \hat{\sigma}^2_{\varepsilon w}(\sum_s\left(a_{k,grg} - 1\right)^2 w_k\, c_k) \quad \text{(A2.3)}$$

where $\tilde{v}_{grg}$ denotes the design-based estimator $\hat{V}_{\pi|\xi}(\sum_s \delta_{k,grg} w_k)$ and approximates v_{grg} well because the second term is of a much smaller order ($O_p(N)$) than the first term ($O_p(N^2/n)$). Using the concept of anticipated variance, a simple expression $\tilde{\tilde{v}}_{grg}$ assuming ξ holds for s is obtained as

$$\tilde{\tilde{v}}_{grg} \equiv \widehat{Var}_{\xi|\pi}(t_{y,grg} - T_y) = \hat{\sigma}^2_{\varepsilon w}[\sum_s\left(w_k a_{k,grg} - 1\right)^2 c_k + \sum_s(w_k - 1)c_k] \quad \text{(A2.4)}$$

where the unknown parameters $\sigma^2_\varepsilon$and $\sum_U c_k$ in $Var_{\xi|\pi}\left(t_{y,grg} - T_y\right)$ are estimated under $\pi\xi -$ and $\pi|\xi -$ randomization, respectively, and not strictly under $\xi|\pi$. This flexibility is reasonable because $\tilde{\tilde{v}}_{grg}$ estimates $Var_{\pi\xi}\left(t_{y,grg} - T_y\right)$ under the joint randomization using the simplifying assumption of ξ holding for s.

A3: $\widehat{\boldsymbol{Var}}_{\boldsymbol{\pi\xi}}(\boldsymbol{t_{y,prd}} - \boldsymbol{T_y})$ or $\boldsymbol{v_{prd}}$

We have

$$t_{y,prd} - T_y \approx \sum_s \delta_{k,prd} - \sum_U \varepsilon_k, \quad \delta_{k,prd} = \varepsilon_k a_k(\eta_{prd}) \quad \text{(A3.1)}$$

where $a_k(\eta_{prd})$ (= $1 + x_k'\, c_k^{-1}\eta_{prd}$), and η_{prd} is (N/n) times the limit in probability of $(n/N)\hat{\eta}_{prd}$(= $(n/N)(X'C^{-1}X)^{-1}(T_x - t_{xu})$) under $\pi|\xi$. Analogous to GREG,

$$v_{prd} = \tilde{v}_{prd} + \hat{\sigma}^2_{\varepsilon w}(\sum_s\left(a_{k,prd}\pi_k - 1\right)^2 w_k\, c_k) \quad \text{(A3.2)}$$

where $\tilde{v}_{prd}$ is $\hat{V}_{\pi|\xi}(\sum_s \delta_{k,prd})$. Note that $\tilde{v}_{prd}$ like $\tilde{v}_{grg}$ is $O_p(N^2/n)$ because although it does not involve w_k's, the adjustment factor $a_k(\eta_{prd})$ itself is $O_p(N/n)$. A simplified estimate under the model is obtained as

$$\tilde{\tilde{v}}_{prd} = \hat{\sigma}^2_{\varepsilon u}[\sum_s (a_{k,prd} - 1)^2 c_k + \sum_s (w_k - 1)c_k]. \quad \text{(A3.3)}$$

A4: $\widehat{Var}_{\pi\xi}(t_{y,prg} - T_y)$) or v_{prg}

We have

$$t_{y,prg} - T_y \approx \sum_s \delta_{k,prg} w_k - \sum_U \varepsilon_k, \quad \delta_{k,prg} = \varepsilon_k a_k(\eta_{prg}) \quad \text{(A4.1)}$$

where $a_k(\eta_{prg})$ $(= \pi_k + x_k' c_k^{-1} \eta_{prg})$, and η_{prg} is the limit in probability of $\hat{\eta}_{prg} (= (X'C^{-1}WX)^{-1}(T_x - t_{xu}))$. We have

$$v_{prg} = \tilde{v}_{prg} + \hat{\sigma}^2_{\varepsilon w}(\sum_s (a_{k,prg} - 1)^2 w_k c_k) \quad \text{(A4.2)}$$

where $\tilde{v}_{grg}$ is given by $\hat{V}_{\pi|\xi}(\sum_s \delta_{k,prg} w_k)$ and a simplified expression under the model is

$$\tilde{\tilde{v}}_{prg} = \hat{\sigma}^2_{\varepsilon w}[\sum_s (a_{k,prg} w_k - 1)^2 c_k + \sum_s (w_k - 1)c_k]. \quad \text{(A4.3)}$$

A5: $\widehat{Var}_{\pi^*\pi\xi}(t_{y,spg} - T_y)$ or v_{spg}

We have

$$t_{y,spg} - T_y \approx \sum_s \delta_{k,spg} w_k + \sum_{s^*} \varepsilon_k - \sum_U \varepsilon_k, \quad \delta_{k,spg} = \varepsilon_k a_k(\eta_{spg}) \quad \text{(A5.1)}$$

where $a_k(\eta_{spg})$ $(= x_k' c_k^{-1} \eta_{spg})$, and η_{spg} is the limit in probability of $\hat{\eta}_{spg} (= (X'C^{-1}WX)^{-1}(T_x - t_{xu^*}))$. We have

$$v_{spg} = \tilde{v}_{spg} + \hat{\sigma}^2_{\varepsilon w}[\sum_s (a_{k,spg} - 1)^2 w_k c_k + 2\sum_{s^*} (a_{k,spg} - 1)c_k + \sum_{s^*} \pi_k^* c_k] \quad \text{(A5.2)}$$

where $\tilde{v}_{spg} (= \hat{V}_{\pi^*\pi|\xi}(\sum_s \delta_{k,spg} w_k + \sum_{s^*} \varepsilon_k))$ is the sum of two terms: $\hat{V}_{\pi|\xi}(\sum_s \delta_{k,prg*} w_k)$, which is obtained using standard design-based methods, and $\hat{V}_{\pi^*|\xi}(\sum_{s^*} \varepsilon_k)$, which can be approximated under the replacement PSU assumption (with elementary units as PSUs; Wolter, 2007, pp. 205) as $(n^*/(n^* - 1))\sum_{s^*} (\varepsilon_k - \bar{\varepsilon})^2$ evaluated at $\hat{\beta}_w$. In fact, the estimate v_{spg} is quite robust to departures from this assumption because the term $\hat{V}_{\pi^*|\xi}(\sum_{s^*} \varepsilon_k)$ has relatively a very small order of $O_p(E_{\pi^*}(n^*))$. The last term in (A5.2) involves unknown π_k^* but can be replaced by a conservative estimate $\sum_{s^*} c_k$. Also a simplified expression under the model is

$$\tilde{\tilde{v}}_{spg} = \hat{\sigma}^2_{\varepsilon w}[\sum_s (a_{k,spg} w_k - 1)^2 c_k + \sum_s (w_k - 1)c_k - \sum_{s^*} c_k]. \quad \text{(A5.3)}$$

A6: $\lambda_{sig} = \hat{\sigma}^2_{\varepsilon w}(\sum_s c_k w_k (w_k - 1))/[\hat{\sigma}^2_{\varepsilon w}(\sum_s c_k w_k^2 - \sum_{s^*} c_k)]$

It follows from the anticipated variance calculation in A2 that

$$\widehat{Var}_{\xi|\pi}(\sum_s \varepsilon_k w_k - \sum_U \varepsilon_k) = \hat{\sigma}^2_{\varepsilon w}[\sum_s w_k(w_k - 1)c_k] \quad \text{(A6.1)}$$

$$\widehat{Var}_{\xi|\pi^*}(\sum_{s^*} \varepsilon_k - \sum_U \varepsilon_k) = \hat{\sigma}^2_{\varepsilon w}[\sum_s w_k c_k - \sum_{s^*} c_k] \quad \text{(A6.2)}$$

and

$$E_{\xi|\pi^*\pi}(\sum_s \varepsilon_k w_k - \sum_U \varepsilon_k)(\sum_{s^*} \varepsilon_k - \sum_U \varepsilon_k) = \sigma^2_\varepsilon[\sum_{s\cap s^*} w_k c_k - \sum_s w_k c_k - \sum_{s^*} c_k + \sum_U c_k] \quad \text{(A6.3)}$$

where the last term is zero under $E_{\pi^*\pi}$ and using independence of π and π^*.

A7: $\widehat{Var}_{\pi^*\pi\xi}(t_{y,sig} - T_y)$ or v_{sig}

We have

$$t_{y,sig} - T_y \approx \sum_s \delta_{k,sig} w_k + \lambda_{sig} \sum_{s^*} \varepsilon_k - \sum_U \varepsilon_k, \tag{A7.1}$$

$$\delta_{k,sig} = \varepsilon_k[(1 - \lambda_{sig})a_k(\eta_{grg}) + \lambda_{sig} a_k(\eta_{spg})] \tag{A7.2}$$

$$v_{sig} = \tilde{v}_{sig} + \hat{\sigma}^2_{\varepsilon w} \times est \sum_U \left((1 - \lambda_{sig})\{a_k(\eta_{grg}) - 1\} + \lambda_{sig}\{a_k(\eta_{spg}) - 1\} + \lambda_{sig}\pi_k^* \right)^2 c_k \tag{A7.3}$$

where $\tilde{v}_{sig} = \hat{V}_{\pi^*\pi|\xi}(\sum_s \delta_{k,sig} w_k + \lambda_{sig} \sum_{s^*} \varepsilon_k)$. The last term in (A7.3) involves unknown π_k^*, but a conservative estimate can be used as in A5. Now, a simplified estimate under the model is obtained as

$$\tilde{\tilde{v}}_{sig} = \hat{\sigma}^2_{\varepsilon w}[\sum_s(\{(1 - \lambda_{sig})a_{k,grg} + \lambda_{sig} a_{k,spg}\}w_k - 1)^2 c_k + \left((1 - \lambda_{sig})^2 - 1\right)\sum_{s^*} c_k + \sum_s (w_k - 1)c_k]. \tag{A7.4}$$

A8: $\tilde{\tilde{v}}_{spg}$ tends to be at most $\tilde{\tilde{v}}_{grg}$ if $\mathbf{1}_{n\times 1}$ is in the column space of $C^{-1}X$

The above claim will establish heuristically why S-PREG tends to be more efficient than GREG. It is possible to show in terms of simplified estimated variance expressions (A5.3) and (A2.4) when the model is assumed to hold for s and $s^*$and under the condition that $1_{n\times 1}$ is in the column space of $C^{-1}X$. We can express $a_{k,spg}$ alternatively as

$$a_{k,spg} = 1 + x_k' c_k^{-1}(X'C^{-1}WX)^{-1}\{(T_x - t_{xu^*}) - t_{xw}\}, \tag{A8.1}$$

because $x_k' c_k^{-1}(X'C^{-1}WX)^{-1}t_{xw} = 1$ if $1_{n\times 1}$ is in the column space of $C^{-1}X$. Comparing the above expression of $a_{k,spg}$ with $a_{k,grg}$ of (6), it is easily seen that the sampling weights in S-PREG are being adjusted to reduced control totals $T_x - t_{xu^*}$ after accounting for the contribution t_{xu^*} from s^* toward T_x. Therefore we expect the adjustments $a_{k,spg}$, which tend to be positive, to be less than $a_{k,grg}$ in general. This establishes feasibility of the above claim by comparing variance estimates (A5.3) and (A2.4). In fact, the leading term of order $O_p(N^2/n)$ in $\tilde{\tilde{v}}_{spg}$ of (A5.3) is $\hat{\sigma}^2_{\varepsilon w}\sum_s(w_k a_{k,spg})^2 c_k$, which can be shown to at most equal to the leading term $\hat{\sigma}^2_{\varepsilon w}\sum_s(w_k a_{k,grg})^2 c_k$ in $\tilde{\tilde{v}}_{grg}$ of (A2.4) as follows. Letting $\tilde{X}$ denote $C^{-1/2}X$, we can express $\sum_s(w_k a_{k,grg})^2 c_k$ as a sum of quadratic forms; i.e.,

$$\sum_s(w_k a_{k,grg})^2 c_k = T_x'(\tilde{X}'W\tilde{X})^{-1}(\tilde{X}'WW\tilde{X})(\tilde{X}'W\tilde{X})^{-1}T_x \tag{A8.2a}$$

$$= tr\left(W\tilde{X}(\tilde{X}'W\tilde{X})^{-1}T_x T_x'(\tilde{X}'W\tilde{X})^{-1}\tilde{X}'W\right) \tag{A8.2b}$$

$$= \sum_s u_k' T_x T_x' u_k \tag{A8.2c}$$

where u_k is the p −vector $(\tilde{X}'W\tilde{X})^{-1}\tilde{x}_k w_k$, and $\tilde{x}_k$ is the kth column of $\tilde{X}'$. Similarly, we can write

$$\sum_s(w_k a_{k,spg})^2 c_k = \sum_s u_k'(T_x - t_{xu^*})(T_x - t_{xu^*})' u_k . \tag{A8.3}$$

Now, since $0 \le (T_x - t_{xu^*}) \le T_x$ elementwise, and $\left(T_x - (T_x - t_{xu^*})\right)\left(T_x - (T_x - t_{xu^*})\right)'$ is non-negative definite, each quadratic form in the sum (A8.3) is less than or equal to the corresponding term in (A8.2c). This establishes the desired result.

Similarly, we can show why PREG tends to be more efficient than GREG. Here, we can express $a_{k,prg}$ alternatively as

$$a_{k,prg} = \pi_k + 1 + x_k' c_k^{-1}(X'C^{-1}WX)^{-1}\{(T_x - t_{xu}) - t_{xw}\} \tag{A8.4}$$

which along the lines of the argument for S-PREG also tends to be less than $\pi_k + a_{k,grg}$. Now since the contribution of the term π_k in $\tilde{\tilde{v}}_{prg}$ is relatively negligible, the desired claim for PREG seems feasible. A stronger result about the leading term $\hat{\sigma}^2_{\varepsilon w} \sum_s \left(w_k a_{k,prg}\right)^2 c_k$ in $\tilde{\tilde{v}}_{prg}$ of (A4.3) can be obtained by using an expression analogous to (A8.3) by replacing $(T_x - t_{xu^*})$ by $(T_x - t_{xu})$.

The above claim also applies to PRED though the proof is somewhat different. Given $1_{n\times 1}$ in the column space of $C^{-1}X$, we have alternate expressions

$$a_{k,prd} = x_k' c_k^{-1}(X'C^{-1}X)^{-1}T_x, \text{ and } a_{k,grg} = x_k' c_k^{-1}(X'C^{-1}WX)^{-1}T_x . \quad \text{(A8.5)}$$

It is sufficient to show that the leading term $\hat{\sigma}^2_{\varepsilon w} \sum_s \left(w_k a_{k,grg}\right)^2 c_k$ in (A2.4) of $\tilde{\tilde{v}}_{grg}$ is greater than or equal to the leading term $\hat{\sigma}^2_{\varepsilon w} \sum_s \left(a_{k,prd}\right)^2 c_k$ in (A3.3) of $\tilde{\tilde{v}}_{prd}$. We have $\sum_s \left(w_k a_{k,grg}\right)^2 c_k = T_x'\left(\tilde{X}'W\tilde{X}\right)^{-1}(\tilde{X}'WW\tilde{X})\left(\tilde{X}'W\tilde{X}\right)^{-1}T_x$, and $\sum_s \left(a_{k,prd}\right)^2 c_k = T_x'\left(\tilde{X}'\tilde{X}\right)^{-1}\left(\tilde{X}'\tilde{X}\right)\left(\tilde{X}'\tilde{X}\right)^{-1}T_x = T_x'\left(\tilde{X}'\tilde{X}\right)^{-1}T_x$. Now, using the multivariate Cauchy-Schwarz inequality, we observe that $\left((W\tilde{X})'\tilde{X}\right)^{-1}\left((W\tilde{X})'W\tilde{X}\right)\left(\tilde{X}'(W\tilde{X})\right)^{-1} - \left(\tilde{X}'\tilde{X}\right)^{-1}$ is non-negative definite. Hence, the desired result follows.

A9: $\boldsymbol{E_\xi(\sum_U \varepsilon_k \pi_k^*) \leq O(\sqrt{N})}$

By Cauchy-Schwarz, $|\sum_U \varepsilon_k \pi_k^*| \leq \sqrt{\sum_U \varepsilon_k^2}\sqrt{\sum_U \pi_k^{*2}}$, and since $\sum_U \varepsilon_k^2 = O_p(N)$, and $\sum_U \pi_k^{*2} \leq \sum_U \pi_k^* = E_{\pi^*}(n^*)$ which is bounded, we have the desired result. If we make the additional mild assumption that $(\sum_U \pi_k^* / N)^{-2} \sum_U \pi_k^{*2}/N$ is $O_p(1)$ where it is known that $\sum_U \pi_k^{*2}/N \geq (\sum_U \pi_k^*/N)^2$ by Jensen, then we have $\sum_U \pi_k^{*2} = O_p(N^{-1})$. This in turn implies that $E_\xi(\sum_U \varepsilon_k \pi_k^*)$ is only $O(1)$.

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.

Binder, D.A. (1983). On the variances of asymptotically normal estimates from complex surveys. *Int. Statist. Rev*., 51, 279-292.

Brewer, K.R.W. (2002). *Combined Survey Sampling Inference: Weighing Basu's Elephants*. London: Oxford University Press, Inc.

Cochran, W.G. (1953). *Sampling Techniques* (1st ed.). New York: Wiley.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimation in survey sampling. *JASA*, 87, 376-382.

DiSogra, C., Cobb, C., Chan, E., and Dennis, J.M. (2011). Calibrating nonprobaililty internet samples with probability samples using early adopter characteristics. In: *JSM Proceedings, Surv. Res. Meth. Sec*.

Efron, B., and Morris, C. (1972). Limiting the risk of Bayes and empirical Bayes estimators–part II: the empirical Bayes case. *JASA*, 67, 130-139.

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

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.

Fuller, W.A. (2009). *Sampling Statistics*. Hoboken, NJ: John Wiley & Sons.

Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *Int. Statist. Rev*., 54, 127-138.

Graubard, B.I., and Korn, E.L. (2009). Scatterplots with survey data. In: D. Pfeffermann and C.R. Rao (eds.), *Handbook of Statistics, 29B: Sample Surveys: Inference and Analysis*. Amsterdam: North Holland, pp. 397-419.

Hansen, M.H., and Hurvitz, W.N. (1943). On the theory of sampling from finite populations. *Ann. Math. Statist.,* 14, 333-362.

Hansen, M.H., Hurvitz, W.N., and Madow, W.G. (1953). *Sample Survey Methods and Theory*, vols. I and II. New York: John Wiley & Sons.

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.

Horvitz, D.G., and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *JASA*, 47, 663-685.
Isaki, C.T., and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *JASA*, 77, 89-96.
Kim, J.K., and Park, H. (2006). Imputation using response probability. *Can. J. Stat*., 34, 1-12.
Kott, P.S. (2009). Calibration weighting: combining probability samples and linear prediction models. In: D. Pfeffermann and C.R. Rao (eds.), *Handbook of Statistics 29B: Sample Surveys: Inference and Analysis*. Amsterdam: North Holland, pp. 55-82.
Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Surv. Methodol*., 32, 133-142.
Kott, P.S., and Liao, D. (2015). One step or two? Calibration weighting from a complete list frame with nonresponse. *Surv. Methodol*., 41, 165-181.
Little, R.J.A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *JASA,* 99, 549-556.
Narain, R.D. (1951). On sampling without replacement with varying probabilities. *Jour. Ind. Soc. Agri. Statist.*, 3, 169-175.
Neyman, J. (1934). On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *JRSS*, 97, 558-606.
Rao, J.N.K. (2003). *Small Area Estimation* (1st ed.). New York: John Wiley & Sons.
Robinson, P.M., and Särndal, C.-E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhya B,* 45, 240-248.
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.
Singh, A.C. (1996). Combining information in survey sampling by modified regression. In: *ASA Proceedings, Surv. Res. Meth. Sec.*, pp. 120-129.
Singh, A.C., and Mian, I.U.H. (1995). Generalized sample size dependent estimators for small areas. In: *ARC Proceedings,* U.S. Census Bureau, pp. 687-701.
Valliant, R., Dorfman, A.H., and Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach.* New York: John Wiley & Sons.
Wolter, K.M. (2007). *Introduction to Variance Estimation* (2nd ed.). New York: Springer.