A Model-Over-Design Integration for Estimation from Purposive Supplements to Probability Samples

Avinash C. Singh, NORC at the University of Chicago, Chicago, IL 60603
singh-avi@norc.org

Abstract

For purposive samples, design-based methods are clearly not suitable. There is the possibility of using model-based methods but there are concerns about the design being informative and potential misspecification of the model mean. 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. A design-based estimate such as GREG 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 another estimator from the additional seen observations in the purposive sample. 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 used only to improve the model-error correction from the additional seen units. The MSE of the resulting estimator can be estimated under the joint randomization of man-made probability sample design, nature-made purposive sample design, and the model for the finite population.

Key Words: Fit-For-Purpose Samples; Informative Designs; Joint design-model-based inference; Non-probability 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, 1st ed.). 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 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.

The recent AAPOR Task Report (Baker et al., 2013) shows clearly the conundrum in using purposive samples for the following reasons. First we note that the term 'purposive sample' used in this paper 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 (instead of 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 due to unit nonresponse which is difficult to distinguish from self-selection. 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 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 super-population models governing selection of the finite population under consideration. As is commonly done in model-based approach, a linear regression model is first assumed for the finite population. 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, we 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 mean. Next, given the regression parameters, instead of simply using the observed model errors from s for estimating the total model error (this is the random part of the model mean) as in the case of commonly used GREG estimator of Särndal (1980), we propose to combine it with the additional observed model errors provided by s^*. The underlying premise is that although s^* may not be deemed fit for estimating model parameters, 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 misspecification. 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 (MOD) integration, 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. 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 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. In addition, 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 integration 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. 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 reviewed in detail in order to motivate the proposed approach and consider some variations used later on for integration with the purposive sample. To this end, we make two basic assumptions which are quite natural for the problem at hand.

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 are uncorrelated with the conceptual selection probabilities π_k^* of units that could be selected in the purposive sample.

The assumption C2 may be deemed to be satisfied in general because of the nature-made design π^* for s^* as π_k^*'s are expected to be functions of unit covariates or its 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 includes suitable covariates that are expected to govern nature's random mechanism for selection of s^*. 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. 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. A modification of MOD-Integration is considered in Section 4 to deal with subpopulation or domain estimation where GREG may not be reliable due to insufficient sample size but can be made so in combination with a purposive sample from the domain of interest.

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 deign 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. 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 super-population.

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 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, 2000, signifying the prediction approach of Royall, 1970, 1976) because both use models to start with, although unlike mainstream statistics, the parameters of interest are not model parameters but the finite population totals involving random effects (or model errors) also. In the following, we first review GREG followed by PRED in some details 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 *k*th unit, $1 \le k \le N$, given by

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

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

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

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 $\varepsilon_k = y_k - x_k'\beta$. 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

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

$= t_{yw} + \beta'(T_x - t_{xw})$ (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—asymptotically design consistent) as n, N get large under general regularity conditions (see the asymptotic framework of Isaki and Fuller, 1982, and the book by Fuller, 2009, Section 1.3); 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, $N^{-1}\left(t_{y,grg}(\beta) - T_y\right) = N^{-1}(\sum_{k\in s} \varepsilon_k w_k - \sum_{k\in U} \varepsilon_k) = o_p(1)$. 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 efficiency of HT estimators. However, its 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

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

and the corresponding sample EFs are given by

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

It is easily seen that

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

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 $\pi|\xi$ −MSE of the regression estimator about T_y where MSE denotes mean square error. Although, GREG with $\hat{\beta}_w$ (to be denoted by $t_{y,grg}$ instead of $t_{y,grg}(\hat{\beta}_w)$) is no longer unbiased, it remains asymptotically design unbiased (ADU) 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 observations that the model residuals $e_k(\hat{\beta}_w) = y_k - x_k'\hat{\beta}_w$ 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. This is easily seen from the calibration form of the GREG estimator as introduced by Deville and Särndal (1992) and is given by

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

where $\hat{\eta}_{grg} = (X'C^{-1}WX)^{-1}(T_x - t_{xw})$, $W = diag(w_k)_{1\le k\le n}$, $C = diag(c_k)_{1\le k\le n}$ and X is the $n \times p$ matrix of the sample covariate values x_k's. Observe that the sample x_k-values inflated by the adjusted weights $(w_k a_{k,grg})_{1\le k\le n}$ 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 $\sum_s e_k(\hat{\beta}_w)w_k$ becomes zero under the above condition on covariates. This built-in benchmarking property of GREG residuals to sum to zero is attractive in practice for robustification to possible model misspecifications. Incidentally, the $\pi|\xi$ −MSE of $t_{y,grg}$ about T_y can be approximated well for large samples using the Taylor 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 model error by using an unweighted estimator from observed errors in the sample. However, 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

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

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 model mean predictor of 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

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

$$= T_x'\beta + \sum_{k\in s} \varepsilon_k \qquad (8b)$$

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 (and has optimality of the HT estimator) for the design bias $(-\sum_U \varepsilon_k)$ in the synthetic estimator $T_x'\beta$, while the unweighted sum $\sum_{k\in s} \varepsilon_k$ 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. Under PRED, however, the regression parameters are estimated by

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

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 MSE) if the model holds for the sample.

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

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

where $\hat{\eta}_{prd} = (X'C^{-1}X)^{-1}(T_x - t_{xu})$, 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. In addition, 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 although they appear to be so because $t_{y,prd}$ does reproduce true totals T_x when y is replaced by x. A $\xi|\pi$ −MSE 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 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. However, this assumption 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 multi-stage 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, there is model-bias (also known as the selection bias) in the model-based estimator even though for units in the finite population U, the model

mean is correctly specified. 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. (Note 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 again can be estimated by *n*.) Besides the above problem, even if the design is noninformative for the selected model, there may be model-bias due to misspecification of the model mean. The above two concerns (model bias 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 assumptions 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 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 preserve the ADC property of GREG while integrating it with the model-based estimator PRED in order 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 (MOD) integration 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 Efficient 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 $\hat{\beta}_u$--the estimator $\hat{\beta}_w$ is preferable to $\hat{\beta}_u$ for reasons mentioned earlier. The PREG estimator (to be denoted by $t_{y,prg}$) is defined as

PREG: $$t_{y,prg} = T_x'\hat{\beta}_w + \sum_{k\in s} e_k(\hat{\beta}_w) \qquad (11)$$

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

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

where $\hat{\eta}_{prg} = (X'C^{-1}WX)^{-1}(T_x - t_{xu})$. An estimator of the MSE 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. Even if the new estimator were less efficient than PREG, it would be preferred to PREG due to tenuous model assumptions underlying 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 is satisfied for the sample s ; i.e.,

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

$$E_\xi((\sum_{\mathrm{U}} \varepsilon_k\, \pi_k - \sum_U \varepsilon_k\,|x_k, 1 \le k \le N) = 0. \qquad (13)$$

where $\sum_{\mathrm{U}} \varepsilon_k\, 1_{k\in s} = \sum_s \varepsilon_k$. The above condition holds if π_k's are functions of x_k's which is unlikely but can be easily satisfied by enlarging the model to include $\pi_k's$ as 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, the assumption of π_k being a function of x_k's (after enlargement if necessary) is much weaker than the assumption of noninformative designs. Although, to satisfy it, introduction of π_k (a design-specific feature) as a covariate may seem somewhat an artifact for a specific purpose because the sampling design refers to the finite population and not to the super-population, it may nevertheless serve as a good covariate on 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 Supplementary-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

S-PREG: $$t_{y,spg} = T_x'\hat{\beta}_w + \sum_{k\in s^*} e_k(\hat{\beta}_w) \qquad (14)$$

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

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

where $\hat{\eta}_{spg} = (X'C^{-1}WX)^{-1}(T_x - t_{xu^*})$. An estimator of the MSE 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 and s^* as a supplement; i.e., how to integrate the two estimators of the total model error from GREG and SPREG 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 is not likely to be available for all units.

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

For MOD integration, the conditions C1 for ξ and C2 for π^* are assumed to hold as mentioned earlier. 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. The sampling designs for $s^*$and s are assumed to be independent and, 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 new

estimator under MOD-Integration, termed Supplementary-sample for Integrated PREG (SI-PREG for short and denoted by $t_{y,sig}$) as follows.

$$\textbf{SI-PREG: } t_{y,sig} = T_x'\hat{\beta}_w + (1-\lambda^*)\sum_s e_k(\hat{\beta}_w)w_k + \lambda^*\sum_{s^*} e_k(\hat{\beta}_w) \quad (16a)$$

$$= T_x'\hat{\beta}_w + \sum_s e_k(\hat{\beta}_w)w_k + \lambda^*\left(\sum_{s^*} e_k(\hat{\beta}_w) - \sum_s e_k(\hat{\beta}_w)w_k\right) \quad (16b)$$

where the coefficient λ^* is obtained in a nonoptimal manner for stability and for obtaining an expansion form of the estimator. (Incidentally, an optimal choice of λ^* can be obtained by minimizing the MSE of $t_{y,sig} - T_y$; i.e., it 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 MSE and mean cross-product error (Isaki and Fuller, 1982) about T_y under the joint $\pi^*\pi\xi$ −randomization. This integration of the two estimators is nonoptimal because λ^* 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^* can be obtained as

$$\lambda^* = \tilde{v}_{grg}/(\tilde{v}_{grg} + \tilde{v}_{spg}) \quad (17)$$

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

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

where $\hat{\sigma}^2_{\varepsilon w} = \sum_s e_k(\hat{\beta}_w)^2 w_k c_k^{-1}/\sum_s w_k$. The anticipated mean cross-product error of GREG and S-PREG given β is zero because of unbiasedness of GREG and independence of $s^*$and s. Although the factor $\hat{\sigma}^2_{\varepsilon w}$ is common in the numerator and the denominator of λ^*, we do not cancel it out as its presence in the numerator allows for an expansion form of the estimator SI-PEG somewhat analogous to a calibration estimator. To see this, observe that the numerator of $\hat{\sigma}^2_{\varepsilon w}$ can be alternatively expressed as $\sum_s y_k\, e_k(\hat{\beta}_w)w_k c_k^{-1}$ because

$$\sum_s e_k(\hat{\beta}_w)^2 w_k c_k^{-1} = \sum_s (y_k - x_k'\hat{\beta}_w)\, e_k(\hat{\beta}_w)w_k c_k^{-1}$$

$$= \sum_s y_k\, e_k(\hat{\beta}_w)w_k c_k^{-1} - \hat{\beta}_w'\sum_s x_k\, e_k(\hat{\beta}_w)w_k c_k^{-1} \quad (19)$$

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^* to obtain an expansion form of SI-PEG as shown below.

$$t_{y,sig} = \sum_s y_k\, w_k a_{k,sig}, \quad a_{k,sig} = a_{k,grg} + e_k(\hat{\beta}_w)c_k^{-1}(\sum_s w_k)^{-1}\hat{\zeta}_{sig}, \quad (20a)$$

$$\hat{\zeta}_{sig} = (\sum_s c_k w_k(w_k - 1) + \sum_{s^*} c_k) \times \left(\hat{\sigma}^2_{\varepsilon w}\left(\sum_s c_k w_k^2 - \sum_{s^*} c_k\right)\right)^{-1}\left(\sum_{s^*} e_k(\hat{\beta}_w) - \sum_s e_k(\hat{\beta}_w)w_k\right) \quad (20b)$$

and $a_{k,grg}$ is given by (6). We remark that the new set of adjusted weights $w_k a_{k,sig}$'s continue to satisfy the GREG calibration controls because $\sum_s y_k\, w_k e_k(\hat{\beta}_w)c_k^{-1}$ is zero when y_k is replaced by one of the covariates from x_k, and therefore, the contribution from the adjustment in $a_{k,sig}$ to $a_{k,grg}$ is zero. An estimate of MSE of $t_{y,sig}$ about T_y under the joint $\pi^*\pi\xi$ −randomization is given in Appendix A7.

The coefficient $\lambda^*$is not based on any variance optimality considerations. Following Singh et al. (2013), design adjustment factors γ_{grg} and γ_{prg*} between 0 and 1, $\gamma_{grg} + \gamma_{spg} = 1$, could be introduced in the definition of λ^* by transforming $\tilde{v}_{grg}$ to $\gamma_{grg}\tilde{v}_{grg}$ and $\tilde{v}_{spg}$ to $\gamma_{spg}\tilde{v}_{prg*}$ such that the unequal weighting effect

$(n/N^2)\sum_s\left(w_k a_{k,sig}\right)^2$ is minimized; here $\sum_s w_k a_{k,sig} = N$. We also remark that the final weights $w_k a_{k,sig}$'s are only defined for the sample s and not both samples unlike the usual case in 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). 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 $a_{k,sig}$.

A multivariate extension of SI-PREG can be easily made. That is, with several key study variables of interest; i.e., when y is multivariate, there is now 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^* in the multivariate case will now be replaced by a matrix, each row of which corresponds to the corresponding study variable. This way, a new set of final weights can be constructed which can be used for all study variables besides the key variables already used in defining new predictors of the total model error.

Observe that the coefficient λ^* reduces to $\sum_s c_k w_k(w_k - 1)$ divided by $\left(\sum_s c_k w_k^2 - \sum_{s^*} c_k\right)$ which 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^* behaves like a shrinkage factor in that high values of λ^* 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^* 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 misspecification. However, under general conditions, we have $\tilde{v}_{grg} = O_p(N^2/n)$, and $\tilde{v}_{spg} = O_p(N)$ which imply that λ^* will tend to be close to 1 because $\tilde{v}_{spg}$ is of much lower order than $\tilde{v}_{grg}$. The practical implication of this is clearly not desirable even though S-PREG would be more efficient than GREG if C1 and C2 truly hold. It is probably better to have only moderate gains in efficiency over GREG in the interest of robustness to model misspecifications. With this in mind, in the spirit of working variances and covariances used in the specification of λ^* to achieve certain goals, we introduce a constraining factor ψ^* based on a priori considerations such that $\lambda^* \to 0$ as $n, N \to \infty$, but n^* remains bounded 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.

SI-PREG^c: $t_{y,sig^c} = T_x'\hat{\beta}_w + (1-\lambda^{*c})\sum_s e_k(\hat{\beta}_w)w_k + \lambda^{*c}\sum_{s^*} e_k(\hat{\beta}_w)$ (21)

where the specification of λ^{*c} is quite similar to that of λ^* by (17) except that $\tilde{v}_{spg}$ in the denominator is multiplied by a constraining factor ψ^* defined below. Letting $\lambda^{*c}(N/n^*)$ denote the value of λ^{*c} when $\psi^* = N/n^*$, we have

$$\psi^* = \begin{cases} N/n^* & if\ \lambda^{*c}(N/n^*) \le 1/2 \\ \tilde{v}_{grg}/\tilde{v}_{spg} & if\ \lambda^{*c}(N/n^*) > 1/2 \end{cases} \quad (22)$$

In other words, λ^{*c} is constrained to be at or below ½ by choosing ψ^* suitably. The choice of λ^{*c} can be improved further by using the design adjustment factors γ_{grg} and γ_{spg} as mentioned earlier in the case of SI-PREG before constraining by ψ^*. The expansion form of $t_{y,cpg*}$ is similar to t_{y,sig^c} except that in (20a), $a_{k,sig}$ is replaced by a_{k,sig^c} defined in an analogous manner and $\hat{\zeta}_{sig}$ replaced by $\hat{\zeta}_{sig^c}$ given by a slightly modified version of (20b) in which the middle term on the right is replaced by $\left(\hat{\sigma}_{\varepsilon w}^2\left(\sum_s c_k w_k(w_k - 1) + \psi^*\left(\sum_s c_k w_k - \sum_{s^*} c_k\right)\right)\right)^{-1}$. An estimate of MSE of t_{y,sig^c} about T_y under the joint $\pi^*\pi\xi$ −randomization is given as in Appendix A7 except that λ^* is substituted by λ^{*c}.

4. An Enhancement of MOD-Integration for Domain Estimation

The method of MOD-Integration is expected to be especially useful in estimation for small or specialized domains which may not be well represented in the full sample, and hence the need for a purposive supplement with a marginal cost. Domains in practice are like socio-demographic subgroups which 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 estimation using GREG is 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. Similarly, SI-PREG for domain estimation can be easily defined by modifying (20a,b) suitably in order to improve precision of GREG estimators for domains. However, precision of SI-PREG for domains obtained in the above standard manner could be further improved if we use full sample (i.e., combined sample over all domains) to estimate fixed parameters $\beta, \sigma_\varepsilon^2$ and λ^*. In other words, for these parameters, we use the same estimators as in the regular SI-PREG estimators for population totals and not subpopulations or domains, but everywhere else we multiply y_k, x_k, and as a result e_k by $1_{k\in U_d}$ to get their contributions for the domain of interest. Thus, for SI-PREG of domains, the effective domain sample size remains the same based on the combined s and s^* but the resulting estimators are expected to be more stable (and hence more precise) due to less variability in the estimates of fixed parameters $\beta, \sigma_\varepsilon^2$ and λ^* 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 (Singh and Mian, 1995, and 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.) The price for obtaining more stable domain level GREG in the above manner is more work 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. The proposed enhancement of SI-PREG for domains starts with the enhanced GREG for domains and improves it further by integrating with domain-specific purposive samples. We now define domain-specific estimators $GREG^d$, $S\text{-}PREG^d$ in order to define $SI\text{-}PREG^d$ denoted respectively by t_{yd,grg^d}, t_{yd,spg^d}, and t_{y,sig^d} as follows.

$$\boldsymbol{GREG^d}\text{:} t_{yd,grg^d} = \sum_{k\in s} y_k w_k a_{kd,grg^d}\,.\, a_{kd,grg^d} = 1_{k\in U_d} + x_k' c_k^{-1} \hat{\eta}_{d,grg^d} \quad (23)$$

where $\hat{\eta}_{d,grg^d} = (X'WC^{-1}X)^{-1}(T_{xd} - t_{xdw})$. 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 $\sum_s e_k(\hat{\beta}_w) w_k 1_{k\in U_d}$ is no longer zero.

$$\boldsymbol{S\text{-}PREG^d}\text{:} t_{yd,spg^d} = \sum_s y_k w_k a_{kd,spg^d} + \sum_{s^*} y_k 1_{k\in U_d},$$
$$a_{kd,spg^d} = x_k' c_k^{-1} \hat{\eta}_{d,spg^d},\ \ \hat{\eta}_{d,spg^d} = (X'WC^{-1}X)^{-1}(T_{xd} - t_{xdu^*}). \quad (24)$$

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

$$\boldsymbol{SI\text{-}PREG^d}\text{:} \quad t_{yd,sig^d} = \sum_s y_k\, w_k a_{kd,sig^d}, \quad (25)$$
$$a_{kd,sig^d} = a_{kd,grg^d} + e_k(\hat{\beta}_w) c_k^{-1} (\textstyle\sum_s w_k)^{-1} \hat{\zeta}_{d,sig^d},$$
$$\hat{\zeta}_{d,sig^d} = \hat{\sigma}_{\varepsilon w}^{-2} \lambda^* \big(\textstyle\sum_{s^*} e_k\,(\hat{\beta}_w) 1_{k\in U_d} - \sum_s e_k(\hat{\beta}_w) w_k 1_{k\in U_d}\big).$$

Note that the domain level control totals T_{xd} continue to be satisfied by the *SI-PREG*^{d} expansion weights. The *SI-PREG*^{d}constrained (denoted by *SI-PREG*^{cd}) estimator can be defined in an analogous manner by replacing λ^* by λ^{*c}, common for all domains. Here we may want to relax the constraining factor ψ^* so that λ^{*c} is at most 2/3, for example. Similarly the design adjustment factors can be introduced. Estimates of MSE of the above estimators about T_{yd} can easily be obtained from previous formulas for full population level estimators by replacing e_k by $e_k 1_{k\in U_d}$.

Appendix (Technical Results)

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

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

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

A2: $\widehat{MSE}_{\pi\xi}(t_{y,grg} - T_y)$

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

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

where $a_k(\eta_{grg})$ is $a_{k,grg}$ of (6) but with $\hat{\eta}_{grg}$ 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 MSE estimator. We have

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

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 $\hat{\sigma}^2_{\varepsilon w}(\sum_s\left(a_{k,grg} - 1\right)^2 w_k\, c_k)$. The second term is much smaller order ($O_p(N)$) than the first term ($O_p(N^2/n)$) and is negligible in practice. Using the concept of anticipated variance, a simple expression assuming ξ holds for s is obtained as

$$\widehat{MSE}_{\xi|\pi}(t_{y,grg} - T_y) = \hat{\sigma}^2_{\varepsilon w}[\sum_s\left(w_k a_{k,grg} - 1\right)^2 c_k + \sum_s (w_k - 1)c_k]. \qquad \text{(A2.3)}$$

A3: $\widehat{MSE}_{\pi\xi}(t_{y,prd} - T_y)$

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

where η_{prd} is (N/n)times the limit in probability of $(n/N)\hat{\eta}_{prd}$ under $\pi|\xi$. Analogous to GREG,

$$\widehat{MSE}_{\pi\xi}(t_{y,prd} - T_y) = \hat{V}_{\pi|\xi}(\sum_s \delta_{k,prd} w_k) + \hat{\sigma}^2_{\varepsilon w}(\sum_s\left(a_{k,prd}\pi_k - 1\right)^2 w_k\, c_k). \qquad \text{(A3.2)}$$

A simplified estimate under the model is obtained as

$$\widehat{MSE}_{\xi|\pi}\left(t_{y,prd} - T_y\right) = \hat{\sigma}^2_{\varepsilon u}[\sum_s\left(a_{k,prd} - 1\right)^2 c_k + \sum_s (w_k - 1)c_k]. \qquad \text{(A3.3)}$$

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

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

$$\widehat{MSE}_{\pi\xi}(t_{y,prg} - T_y) = \hat{V}_{\pi|\xi}(\sum_s \delta_{k,prg} w_k) + \hat{\sigma}^2_{\varepsilon w}(\sum_s \left(a_{k,prg} - 1\right)^2 w_k\, c_k) \quad \text{(A4.2)}$$

$$\widehat{MSE}_{\xi|\pi}\left(t_{y,prg} - T_y\right) = \hat{\sigma}^2_{\varepsilon w}[\sum_s \left(a_{k,prg} - 1\right)^2 c_k + \sum_s (w_k - 1)c_k]. \quad \text{(A4.3)}$$

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

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

$$\widehat{MSE}_{\pi^*\pi\xi}(t_{y,spg} - T_y) = \hat{V}_{\pi^*\pi|\xi}(\sum_s \delta_{k,spg} w_k + \sum_{s^*} \varepsilon_k) + \hat{\sigma}^2_{\varepsilon w}[\sum_s \left(a_{k,spg} - 1\right)^2 w_k\, c_k + 2\sum_{s^*} \left(a_{k,spg} - 1\right) c_k + \sum_{s^*} \pi^*_k c_k] \quad \text{(A5.2)}$$

where the first term on the right 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 WRPSU assumption (with elementary units as PSUs, Wolter, 2007, pp. 205) as $(n^*/(n^* - 1)) \sum_{s^*} (\varepsilon_k - \bar{\varepsilon})^2$ evaluated at $\hat{\beta}_w$. 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

$$\widehat{MSE}_{\xi|\pi^*\pi}\left(t_{y,spg} - T_y\right) = \hat{\sigma}^2_{\varepsilon w}[\sum_s \left(a_{k,spg} w_k - 1\right)^2 c_k + \sum_s (w_k - 1)c_k - \sum_{s^*} c_k]$$

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

It follows from the anticipated variance calculations in A2 that

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

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

$$\text{Est of } E_{\xi|\pi^*\pi}(\sum_s \varepsilon_k w_k - \sum_U \varepsilon_k)(\sum_{s^*} \varepsilon_k - \sum_U \varepsilon_k) = -\hat{\sigma}^2_{\varepsilon w} \sum_{s^*} c_k \quad \text{(A6.3)}$$

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

$$t_{y,sig} - T_y \approx \sum_s \delta_{k,sig} w_k + \lambda^* \sum_{s^*} \varepsilon_k - \sum_U \varepsilon_k, \quad \text{(A7.1)}$$

$$\delta_{k,sig} = \varepsilon_k[(1 - \lambda^*) a_k\left(\eta_{grg}\right) + \lambda^* a_k\left(\eta_{spg}\right)] \quad \text{(A7.2)}$$

$$\widehat{MSE}_{\pi^*\pi\xi}(t_{y,sig} - T_y) = \hat{V}_{\pi^*\pi|\xi}(\sum_s \delta_{k,sig} w_k + \lambda^* \sum_{s^*} \varepsilon_k) \quad \text{(A7.3)}$$

$$+\ \hat{\sigma}^2_{\varepsilon w} \times est\ \sum_U \left((1 - \lambda^*)\{a_k\left(\eta_{grg}\right) - 1\} + \lambda^*\{a_k\left(\eta_{spg}\right) - 1\} + \lambda^* \pi^*_k\right)^2 c_k$$

The last term involves unknown π^*_k but a conservative estimate can be used as in A5.

$$\widehat{MSE}_{\xi|\pi^*\pi}\left(t_{y,sig} - T_y\right) = \hat{\sigma}^2_{\varepsilon w}[\sum_s \left(\{(1 - \lambda^*) a_{k,grg} + \lambda^* a_{k,spg}\} w_k - 1\right)^2 c_k + ((1 - \lambda^*)^2 - 1) \sum_{s^*} c_k + \sum_s (w_k - 1) c_k] \quad \text{(A7.4)}$$

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*, Oxford University press, Inc., N.Y.

Cochran, W.G. (1953). *Sampling Techniques*. 1st ed., Wiley, NY.

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimation in survey sampling. *JASA*, 87, 376-382.
DiSogra, C., Cobb, C., Chan, E., Dennis, J.M. (2011). Calibrating nonprobaililty internet samples with probability samples using early adopter characteristics. *JSM Proceedings, SRMS*.
Elliott, M.R. (2009). Combining data from probability and nonprobability samples using pseudo weights. *Survey Practice*. 2 (6).
Fuller, W.A. (2009). *Sampling Techniques*. Wiley, N.J.
Godambe, V.P., and Thompson, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *Int. Statist. Rev*. 54, 127-138.
Hansen, M.H. and Hurvitz, W.N. 91943). 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, Wiley, NY.
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
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. Wiley, N.Y.
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. *ASA Proceedings, Section on Survey Research Methods*, pp. 120-129.
Singh, A.C. and Mian, I.U.H (1995). Generalized Sample Size Dependent Estimators for Small Areas. *ARC Proceedings*, U.S. Census Bureau, pp. 687-701.
Singh, A.C., Ganesh, N., and Lin, Y. (2013). Improved sampling weight calibration by generalized raking with optimal unbiased modification. *ASA Proceedings, Survey Research Methods Section*, 3572-3583
Valliant, R., Dorfman, A.H., and Royall, R.M. (2000). *Finite Population Sampling and Inference: A prediction Approach.* Wiley, N.Y.
Wolter, K.M. (2007). *Introduction to Variance Estimation*. 2nd ed. Springer.