An Extension of Generalized Regression Estimator to Dual Frame Surveys

A.C. Singh and S. Wu, RTI International
A.C. Singh, RTI, 3040 Cornwallis Road, Research Triangle Park, NC 27709 asingh@rti.org

Abstract

Estimation from multiframe survey data is essentially a problem of combining domain estimates such that the new auxiliary information (for key study variables, z) obtained from several estimates for overlapping domains is used in addition to the usual auxiliary information for socio-demographic and geographic variables (x) used in poststratification. A natural approach is to use optimal regression, but as in the case of single frame surveys, it may be unstable due to insufficient degrees of freedom available for estimating regression coefficients when the number of z-variables is large while estimating for all z in a multivariate sense. As an alternative, an extension of the generalized regression (GREG or GR) estimator can be used which is suboptimal but has stable regression coefficient estimates in the case of multivariate z, and has a convenient calibration form involving final weights. The main problem in GR formulation is how to take account of possibly different designs from multiple frames. An earlier attempt (MR-dualframe of Singh and Wu, 1996), based on the modified regression (MR) methodology of Singh (1994,1996), was made using the relative effective sample size (based on design effect) as the scaling factor in the GREG covariance matrix, but only with partial success. In this paper, we present an enhancement of MR-dualframe which takes account of different designs, allows for range-restricted weights as in single frame surveys, provides calibration weights, has built-in bias-correction due to difference in survey mode effects and can be applied to dependent samples. Monte Carlo simulation results on relative performance of a few dual frame estimators are also presented.

Key Words: Calibration weights, Combining estimates, Mode effects, Optimal regression

1. Introduction

We consider the problem of efficient estimation by combining information in samples from overlapping frames which together cover the target population. Typically, in practice, a dual frame problem arises when one frame is complete but expensive to sample, while the other frame is incomplete but cheaper to sample. In this paper, we introduce a methodology for estimating parameters and variances of their estimates for dual frames although generalization to multi-frame problem is fairly straightforward. While we are mainly concerned with efficiency at the estimation stage, it should be emphasized that this is not a substitute for efficiency considerations at the design stage while dealing with cost-variance issues in sample allocation in anticipation of dual frame estimation. An important aspect that differentiates dual frame from single frame problem is due to the possibility of different sampling designs (simple or complex) for different frames. For a good review, see Skinner and Rao (1996).

The pioneering work in the area of multiple frames is due to Hartley (1962, 1974). Later, important contributions were made by Lund (1968), Fuller and Burmeister (1972), Bankier (1986), Kalton and Anderson (1986), Skinner (1991), Skinner and Rao (1996), Singh and Wu (1996), and Lohr and Rao (2000). We will now define the estimation problem in the case of two frames A and B. Let s_A, s_B be two independent samples drawn respectively from A and B under designs p_{sA} and p_{sB}. For simplicity, we will allow for the possibility of duplicate units in the two samples, and assume that the resulting loss of efficiency is negligible. Let N_A, N_B denote the population sizes and n_A, n_B denote the corresponding sample sizes. We will make the usual assumption that the population counts N_A, N_B are known. In addition, population totals (frame-specific or not) may be available in practice for several auxiliary variables (x) such as demographic and geographic counting variables; the term frame-specific signifies that the auxiliary information is separate for frames A and B. Let domains $A \cap B^c, A^c \cap B$, and $A \cap B$ be denoted respectively by *a, b*, and *c*. Then, for the study variable *y*, parameters of interest are the population total T_{yd} and average A_{yd} for each domain *d*. Denoting by N_c the unknown population count of domain *c*, the component parameters of the population total of y for the combined frame are given by

$$T_y = T_{ya} + T_{yb} + T_{yc} = (N_A - N_c)A_{ya} + (N_B - N_c)A_{yb} + N_c A_{yc} \quad (1.1a)$$

where $A_{ya} = T_{ya}/N_a$, $A_{yb} = T_{yb}/N_b$,
$A_{yc} = T_{yc}/N_c$, $N_a = N_A - N_c$, $N_b = N_B - N_c$. (1.1b)

Let $\hat{T}_{ya}^{GR}$, $\hat{T}_{yb}^{GR}$ denote the usual generalized regression (GR) estimates of T_{ya} and T_{yb} respectively from samples s_A and s_B as defined by Särndal (1980). In particular, they could be simply ratio-adjusted Horvitz-Thompson estimates when N_A, N_B are the only available auxiliary information. For the common (or overlapping) domain *c*, let $\hat{T}_{ycA}^{GR}$ and $\hat{T}_{ycB}^{GR}$ denote the two estimates from s_A and s_B respectively.

Now, a naive *GR*-estimator of T_y can be defined as

$$\hat{T}_{y,naive}^{GR} = \hat{T}_{y\,a}^{GR} + \hat{T}_{y\,b}^{GR} + \left(\hat{T}_{y\,c\,A}^{GR} + \hat{T}_{y\,c\,B}^{GR}\right)/2\,, \quad (1.2)$$

where the component estimates are equally weighted and thus are not combined optimally. The methods proposed in the literature use, however, some form of optimality considerations to obtain a composite estimator. They can be classified as either under separate frame or combined frame approaches. In the separate frame approach, using GR, sampling weights are first adjusted separately for each of the samples s_A and s_B in light of the usual auxiliary information on x to get estimates of domain totals in the linear parametrization (1.1a) of T_y. Then unlike the above naïve GR-estimator (1.2), efficient estimates for the overlapping domain are obtained via optimal regression for combining the two estimates as well as for incorporating extra auxiliary information contained in two estimates of the overlapping domain for a key set of other correlated study variables, z. This can be done as a problem in weight calibration by suitably defining optimal regression for complex designs under the assumption of with replacement PSUs as shown by Rao (1994). The methods of Hartley (H), and Fuller-Burmeister (FB) fall under this category. The method of modified regression (MR-dualframe) of Singh-Wu also falls under this category although it uses suboptimal regression. Note that in computing expansion estimates under the separate frame approach, the two estimates for the common domain are averaged using either optimal or suboptimal regression.

On the other hand, in the combined frame approach, sampling weights for the combined sample, *s*, are adjusted so that it can be visualized as a single sample. These weights could be further calibrated in light of the auxiliary information. Finally, the expansion estimates are computed from the calibrated weights in the usual way. Under this approach, T_y-estimation problem can be addressed either through the linear parametrization (1.1a) in terms of domain totals or through the nonlinear parametrization (1.1b) in terms of domain means. In the nonlinear case, the estimates are nonlinear functions (e.g. ratio-cum-product) of usual regression (i.e. GR) estimates. The methods of Bankier (B), and Kalton and Anderson (KA) take the linear parametrization for the T_y-estimation problem, and consider the combined sample $s = s_A \cup s_B$ as coming from a single frame, and assign inclusion probabilities to units in the three domains a, b, and c. Once this is done, usual expansion methods can be applied for estimating T_y. The methods of Lund (L), other one due to Fuller-Burmeister (FB*) when the design is restricted to simple random samples, and the method of Skinner-Rao (SR)) which is applicable to complex designs, use the nonlinear parametrization via domain means, and can be classified under the combined frame approach.

Before considering limitations of the existing methods, it is useful to list desirable goals for a dual frame composite estimator:

(i) it can be expressed as a calibration estimator,
(ii) the adjusted weights should satisfy the usual x-controls, and the weight adjustment factors should satisfy range restrictions such as being nonnegative,
(iii) it should take into account of difference in complexity of the sample designs from the two frames,
(iv) it should allow for possible dependence of the two samples,
(v) it should be multivariate in nature, i.e., composite estimates for several y-variables can be generated from the same set of calibration weights which may depend on a common set of z-variables,
(vi) it should correct for possible bias due to difference in survey mode from the two frames,
(vii) it should be applicable to three or more frames, and finally
(viii) it should provide significant gains in efficiency.

If there was no bias due to mode effect, then the method based on optimal regression can be used to meet most of the goals above except possibly the final one about efficiency which is after all the ultimate goal of any composite estimation. The reason for this is that optimal regression may lead to instability of estimates in the presence of several auxiliary variables (re: Goal (v)) due to inadequate degrees of freedom available for estimating covariance matrices (see, e.g., Rao, 1994). The MR-dualframe method was proposed to serve as a GR-type alternative to the optimal regression (i.e., optimal for simple random samples but expected to be robust for complex designs) to overcome the problem of insufficient degrees of freedom. The main differences between MR-dualframe and the usual GR method for single frames are that general predictors (in the form of difference of two estimates for the overlapping frame), and relative measures of the effective sample size as a scaling factor in the GR-covariance matrix are incorporated. However, there are two main limitations: lack of an alternative factor as a measure of the effective sample size without relying on the design effect which is inadequate for the multivariate case, and lack of a more objective factor than using 1/2 for combining estimates for the overlapping frame. The main purpose of this paper is to propose an objective and efficient choice of the two factors (η_A, ζ_A) defining respectively the scaling factor in the GR-covariance matrix and the combining factor for frame A, the corresponding factors for frame B are defined by the complements, $\eta_B = 1-\eta_A, \zeta_B = 1-\zeta_A$. The enhanced version of MR-dualframe is termed in this paper as the dual frame calibration (DFC) method. Besides optimal regression and MR-

dualframe, other existing methods mentioned above also have various limitations with respect to the eight goals listed above, see Singh and Wu (1996). In Section 2, a review of MR-dualframe and motivation for the proposed enhancements are provided. The proposed DFC method is described in Section 3 followed by its properties in Section 4. Empirical evaluation of some methods based on a Monte Carlo study is presented in Section 5 followed by concluding remarks in Section 6.

2. Review of MR-dualframe and motivation

MR-dualframe uses the modified regression methodology of Singh (1994, 1996) which was inspired by the contributions in survey statistics of Fuller (1975) and Särndal (1980) on regression methods, and of Rao and Scott (1981) on pseudo-maximum likelihood estimation, and the contributions in classical statistics by Liang and Zeger (1986) on generalized estimating equations, and of Godambe and Thompson (1989) on optimal estimating functions. The MR-methodology, based on the idea of finite population (semiparametric) modelling with a working covariance structure within the estimating function framework, encompasses the GR-methodology of Särndal (1980) (see also Fuller, 1975), based on the idea of superpopulation modelling within the model-assisted framework.

2.1 MR-dualframe for completely overlapping case

Suppose, for convenience, the two frames A and B overlap completely and the samples are independent. In other words, we have two independent samples from a single frame. Thus, the problem reduces to the familiar problem of combining two (approximately) unbiased and independent estimates $\hat{T}_{zA}^{GR}$ and $\hat{T}_{zB}^{GR}$ of the same parameter T_z. Such problems often arise in practice with rotating panel surveys. Besides combining the two estimates, it is desirable in the interest of efficiency gains to combine correlated auxiliary information in the form of predictor zero functions for other selected study variables (z) via regression. Also, it is important to continue to satisfy the usual x-controls used in GR-estimator by the composite estimator if z is replaced by x. One way to do this is to use partial regression with new z-predictors after GR is performed using the x-predictors. A simpler alternative is to perform a multiple regression on all the usual predictor zero functions based on the design weights denoted by $\hat{T}_{\mathrm{xA}}^{\mathrm{HT}} - T_{\mathrm{xA}}$ (used in $\hat{T}_{zA}^{GR}$), $\hat{T}_{\mathrm{xB}}^{\mathrm{HT}} - T_{\mathrm{xB}}$ (used in $\hat{T}_{zB}^{GR}$), and the additional predictors, $\hat{T}_{zA}^{HT} - \hat{T}_{zB}^{HT}$, due to overlapping frames. Here, $\mathrm{x_A}$, $\mathrm{x_B}$ denote respectively the vector of auxiliary variables specific to frames A and B (which are identical for the present case) and HT signifies the Horvitz-Thompson estimator. Now, instead of optimal regression, the MR-estimator, $\hat{T}_z^{MR}$, uses a suboptimal regression under a working covariance matrix. This is done under a finite population (semiparametric) common mean model for the elementary estimates, consisting of four types of estimates of T_z:

(i) $\hat{T}_{z,naive}^{HT} := (\hat{T}_{zA}^{HT} + \hat{T}_{zB}^{HT})/2$, (ii) $\hat{T}_{z,naive}^{HT} + (\hat{T}_{\mathrm{xA}}^{\mathrm{HT}} - T_{\mathrm{xA}})$, (iii) $\hat{T}_{z,naive}^{HT} + (\hat{T}_{\mathrm{xB}}^{\mathrm{HT}} - T_{\mathrm{xB}})$, (iv) $\hat{T}_{z,naive}^{HT} + (\hat{T}_{\mathrm{zA}}^{\mathrm{HT}} - \hat{T}_{\mathrm{zB}}^{\mathrm{HT}})$.

In other words, each predictor zero function gives rise to a new estimator of T_z by adding it to $\hat{T}_{z,naive}^{HT}$; the predictor is of course assumed to be correlated with $\hat{T}_{z,naive}^{HT}$ in order to be useful. In the above model, elementary estimates can be visualized as working sufficient statistics for T_z as they represent a condensed form of raw survey data before being modelled. When the designs p_{sA} and $\mathrm{p_{sB}}$ are identical, MR can be defined in a manner similar to GR as follows; expressions for $\hat{T}_{zA}^{GR}$ and $\hat{T}_{zB}^{GR}$ are also given for comparison purposes.

$$\hat{T}_{zA}^{GR} = \hat{T}_{zA}^{HT} + (z_A'\Gamma_A X_A)(X'_A\Gamma_A X_A)^{-1}(T_{\mathrm{xA}} - X'_A\Gamma_A 1_A)$$

$$\hat{T}_{zB}^{GR} = \hat{T}_{zB}^{HT} + (z_B'\Gamma_B X_B)(X'_B\Gamma_B X_B)^{-1}(T_{\mathrm{xB}} - X'_B\Gamma_B 1_B)$$

$$\hat{T}_z^{MR} = \frac{1}{2}\left[(\hat{T}_{zA}^{HT} + \hat{T}_{zB}^{HT}) + (z'\Gamma X)(X'\Gamma X)^{-1}(T_{\mathrm{x}} - X'\Gamma 1)\right] \qquad (2.2)$$

where $\hat{T}_{zA}^{HT} = \sum_{k\in s_A} z_k h_{kA}$, $\hat{T}_{zB}^{HT} = \sum_{k\in s_B} z_k h_{kB}$, $\mathrm{z_A} = vec(z_k : k \in s_A), \mathrm{z_B} = vec(z_k : k \in s_B)$, the design weights are $\mathrm{h_A} = vec(h_k : k \in s_A)$, $\mathrm{h_B} = vec(h_k : k \in s_B)$, $\Gamma_A = diag(h_A), \Gamma_B = diag(h_B)$, T_{xA}, T_{xB} are auxiliary control totals, $1_A, 1_B, 1$ are vectors of 1s of dimensions $n_A, n_B, and\ n = n_A + n_B$ respectively, $z' = (z'_A, z'_B)$, Γ = block diag (Γ_A, Γ_B), $T'_x = (T'_{xA}, T'_{xB}, 0)$ and

$$X := \begin{pmatrix} X_A^* \\ X_B^* \end{pmatrix} = \begin{pmatrix} X_A & O & z_A \\ O & X_B & -z_B \end{pmatrix} \qquad (2.3)$$

Note that the X and Γ matrices for MR are simply enlarged versions of the corresponding matrices for GR. Also the covariance matrices used in the regression are suboptimal in that for simple random samples, they reduce to being optimal provided the initial design weights sum to the population count for each frame., see Singh and Sarndal (2003). If the two designs p_{sA} and p_{sB} are different, then one could use a relative measure of inverse effective sample size for a chosen variable (using the design effect as in Skinner-Rao) to give

differential weights to the two matrices Γ_A, Γ_B used in the working covariance structure for MR. Denoting by $\lambda_A^{-1}, \lambda_B^{-1} = 1 - \lambda_A^{-1}$ the relative measures of the effective sample size for designs p_{sA}, p_{sB}, the working covariance matrix for the case of different designs is modified by replacing Γ by $\Lambda\Gamma$ defined as

$$\Lambda\Gamma = \text{block diag}(\lambda_A\Gamma_A, \lambda_B\Gamma_B) \qquad (2.4)$$

Clearly $\lambda_A = \lambda_B = 1$ if the two designs are identical. Also notice that the final calibrated weights $w\{=(w_A', w_B')'\}$, through which $\hat{T}_z^{MR}$ can be represented as an expansion estimator (i.e. $\hat{T}_z^{MR} = z'w/2$) except that estimates for the common domain are averaged, can be obtained as

$$w = h + \Lambda\Gamma X(X'\Lambda\Gamma X)^{-1}(T_x - X'\Gamma 1), \qquad (2.5)$$

where $h=(h_A', h_B')$. Now, in the realistic situation where the two frames overlap only partially, we get three domains a, b, c, and the additional predictors for MR are generated from two estimates for c corresponding to several study variables. This does not pose any new problems as the estimator $\hat{T}_z^{MR}$ of (2.2) can be easily modified by redefining the part of X corresponding to additional predictors, as shown in the next subsection.

2.2 MR-dualframe for partially overlapping frames

From the samples in the common domain c, we get additional predictors such as $\hat{T}_{zcA}^{HT} - \hat{T}_{zcB}^{HT}$ corresponding to each selected z, and of course $\hat{N}_{cA}^{HT} - \hat{N}_{cB}^{HT}$ for the counting variable. The usual predictors are $\hat{T}_{xA}^{HT} - T_{xA}$ and $\hat{T}_{xB}^{HT} - T_{xB}$ respectively for the two frames. In practice one may also have some predictors for the combined frame which can be expressed as $\hat{T}_{x,\,naive}^{HT} - T_x$. Now $\hat{T}_z^{MR}$ combines above pieces of information (available in the form of predictors) with the domain estimates $\hat{T}_{za}^{HT}, \hat{T}_{zb}^{HT}, \hat{T}_{zcA}^{HT}$, and $\hat{T}_{zcB}^{HT}$ to get $\hat{T}_z^{MR}$. Note that the proposed method also allows for using other correlated study variables (z) in estimating the total for any y. To define $\hat{T}_y^{MR}$ for any study variable, we start with $\hat{T}_{x,\,naive}^{HT}$ and then all we need is to modify the matrix X in (2.5) appropriately to get w so that $\hat{T}_y^{MR}$ can be represented as

$$\begin{aligned}\hat{T}_y^{MR} &= \hat{T}_{ya}^{MR} + \hat{T}_{yb}^{MR} + \hat{T}_{yc}^{MR} \\ &= y_a'w_a + y_b'w_b + \left(y_{cA}'w_{cA} + y_{cB}'w_{cB}\right)/2 \end{aligned} \qquad (2.6)$$

where

$$y' = (y_a', y_{cA}', y_{cB}', y_b'), \quad w' = (w_a', w_{cA}', w_{cB}', w_b') \ . \qquad (2.7)$$

Now, the matrix X is $n \times q$ where $n = n_A + n_B$, and q is the total number of predictors. The number q is, in general, equal to $q_1 + q_2 + q_3 + q_4$ where q_1 is the number of frame A-specific predictors, q_2 is the number of frame B-specific predictors (these are the usual predictors for GR), q_3 is the number of predictors for the combined frame, and q_4 is the number of predictors chosen for the common domain c. As in (2.3), the matrix X can be horizontally partitioned into a $n_A \times q$ matrix X_A^* and a $n_B \times q$ matrix X_B^*. These matrices are defined as follows in four parts:

(i) the first q_1 columns of the matrix X_A^* represent n_A observations on the usual predictors for frame A,

(ii) the next q_2 columns are zeros because they correspond to frame B predictors,

(iii) the next q_3 columns contain for each x for the combined frame either x_k or $x_k/2$ depending on whether k is in a or c, and finally

(iv) the last q_4 columns contain for each chosen variable y for the common domain either $+z_k$ or 0 depending on whether k is in c or not.

The matrix X_B^* is similarly defined, the main difference being in the last q_4 columns which contain either $-z_k$ or 0 depending on whether k is in c or not. This completes the description of the proposed method. Note that the control totals T_x^* corresponding to predictors $\hat{T}_{zcA}^{HT} - \hat{T}_{zcB}^{HT}$ for the common domain c will be simply zeros. Moreover, when y is replaced by z, we will have $\hat{\theta}_{ycA}^{MR} = \hat{\theta}_{ycB}^{MR} = \hat{\theta}_{yc}^{MR}$.

2.3 Motivation for enhancing MR-dualframe

One of the main limitations of $\hat{T}_y^{MR}$ is the heuristic use of the combining factor of 1/2 in $\hat{T}_{x,\,naive}^{HT}$ in the GR-framework. The other main limitation is the use of the scaling factors $\lambda_A^{-1}, \lambda_B^{-1} = 1 - \lambda_A^{-1}$ as the relative effective sample sizes obtained via design effect. Again, this is only heuristic because design effect is not applicable to covariance terms in the GR-covariance matrix. To get a better understanding, we consider optimal regression under SRS for combining

estimates from a common frame. An optimal linear combination to minimize the variance is given by

$$\hat{T}_{y,opt} = \alpha_{opt}\,\hat{T}_{y(A)} + (1-\alpha_{opt})\hat{T}_{y(B)} \quad (2.8a)$$

where,

$$\alpha_{opt} = V\left(\hat{T}_{y(B)}\right)\left[V\left(\hat{T}_{y(A)}\right)+V\left(\hat{T}_{y(B)}\right)\right]^{-1} \quad (2.8b)$$

For SRS, we have

$$\hat{T}_{y(A)} = \sum_{s_A} y_{kA}\, d_{kA}$$

where, $d_{ka} = N/n_A$ and the usual variance estimate is

$$\hat{V}\left(\hat{T}_{y(A)}\right) = \left(1-\frac{n_A}{N}\right)\left(\frac{N}{n_A-1}\right)\sum_{s_A} y_{kA}\, d_{kA}\,(y_{kA}-\bar{y}_A) \quad (2.8c)$$

Similarly, $\hat{V}_{y(B)}$ can be defined. Thus

$$1-\hat{\alpha}_{opt} = \frac{\sum_{s_A} y_{kA} d_{kA}(y_{kA}-\bar{y}_A)\left(1-\frac{n_A}{N}\right)\left(\frac{N}{n_A-1}\right)}{\hat{V}\left(\hat{T}_{y(A)}\right)+\hat{V}\left(\hat{T}_{y(B)}\right)} \quad (2.8d)$$

Now, rewriting $\hat{T}_{y,opt}$ as

$$\hat{T}_{y,opt} = \hat{T}_{y(A)} + (1-\hat{\alpha}_{opt})\left(\hat{T}_{y(B)}-\hat{T}_{y(A)}\right) \quad (2.9a)$$

wehave, $\hat{T}_{y,opt} =$

$$\begin{aligned}
&\sum_{s_A} y_{kA} d_{kA}\left[1+c_{kA}(y_{kA}-\bar{y}_A)\hat{V}^{-1}\left(0-\left(\hat{T}_{y(A)}-\hat{T}_{y(B)}\right)\right)\right] \\
&\approx \sum_{s_A} y_{kA} d_{kA}[1+(n_A/n)^{-1}(\hat{\lambda}_{yA}/n)(y_{kA}-\bar{y}_A)] \\
&= \sum_{s_A} y_{kA} d_{kA} a_{kA} = \sum_{s_A} y_{kA} w_{kA}
\end{aligned} \quad (2.9b)$$

where,

$$c_{kA} = \left(1-\frac{n_A}{N}\right)\frac{N}{n_A-1} \approx \frac{N}{n_A},$$

ignoring the finite population correction (fpc);

$$\hat{\lambda}_{yA} = N\left[\hat{V}\left(\hat{T}_{y(A)}\right)+\hat{V}\left(\hat{T}_{y(B)}\right)\right]^{-1}\left(0-\left(\hat{T}_{y(A)}-\hat{T}_{y(B)}\right)\right);$$

and

$$a_{kA} = 1+(n_A/n)^{-1}(\hat{\lambda}_{yA}/n)(y_{kA}-\bar{y}_A).$$

This formula is similar to a linear regression calibration estimator. The factor $\hat{\lambda}_{yA}/n$ can be seen to be $O_p(n^{-1/2})$ under usual conditions. Note that unlike the usual regression estimation for single frame surveys with auxiliary x-variables, here the λ-parameter in the adjustment factor is scaled by the inverse of the relative effective sample size n_A/n. It may be instructive to note that the larger sample, as expected, tends to have smaller adjustments (i.e., the factors are closer to 1). In other words, weights for each sample are differentially adjusted, the sample with higher relative sample size is relatively less adjusted such that the two estimates become identical.

Similarly, we can write $\hat{T}_{y,opt}$ in terms of w_{kB}. We have

$$\hat{T}_{y,opt} = \hat{T}_{y(B)} + \hat{\alpha}_{opt}\left(\hat{T}_{y(A)}-\hat{T}_{y(B)}\right) \quad (2.11a)$$

Therefore,

$$\begin{aligned}
\hat{T}_{y,opt} &\approx \sum_{s_B} y_{kB} d_{kB}[1+(n_B/n)^{-1}(\hat{\lambda}_{yB}/n)(y_{kB}-\bar{y}_B)] \\
&= \sum_{s_B} y_{kB} d_{kB} a_{kB} = \sum_{s_B} y_{kB} w_{kB}
\end{aligned} \quad (2.11b)$$

Equations (2.9b) and (2.11b) imply that the initial weights d_{kA} for s_A and d_{kB} for s_B are calibrated to w_{ka} and w_{kb}, respectively, such that estimates from each sample are identical and equal to $\hat{T}_{y,opt}$. In other words, after calibration, the difference between the two estimates becomes zero, the new control total.

3. Dual Frame Calibration (DFC) Estimator

For given scaling and combining factors (η_A, ζ_A), the DFC estimator is defined as

$$\begin{aligned}
\hat{T}_y^{DFC} &= \hat{T}_{ya}^{DFC} + \hat{T}_{yb}^{DFC} + \hat{T}_{yc}^{DFC} \\
&= y_a' w_a + y_b' w_b + \zeta_A(y_{cA}' w_{cA}) + \zeta_B(y_{cB}' w_{cB})
\end{aligned} \quad (3.1)$$

The parameter η_A is part of the weight adjustment factor. For complex designs, it is difficult in general to write the optimal linear combination in a calibration form. However, based on the motivation in the previous section, we propose the following model for the calibration adjustment factors as

$$\begin{aligned}
a_{kA} &= 1+\eta_A^{-1}(\mathbf{x}_A'\lambda_A + \mathbf{z}_A'\lambda_z), \\
a_{kB} &= 1+\eta_B^{-1}(\mathbf{x}_B'\lambda_B - \mathbf{z}_B'\lambda_z)
\end{aligned} \quad (3.2)$$

where $\boldsymbol{x}$ denotes the usual auxiliary covariates with known control totals T_x, and $\boldsymbol{z}$ is the set of key study variables (y). It may be of interest to note that the above adjustment factors can be obtained by minimizing the following distance function subject to calibration constraints:

$$\Delta(w,d) = \eta_A \sum_{s_A} d_{kA}(a_{kA}-1)^2 + \eta_B \sum_{s_B} d_{kB}(a_{kB}-1)^2 \quad (3.3)$$

Note that z appears with different signs in the above two adjustment factors because control totals for z are zero. However, the λ_z parameters are common to both adjustment factors. Also the factors (η_A, ζ_A) are chosen via a grid search such that variance of $\hat{T}_{z,comp}$ (or trace of the covariance matrix if z is multivariate) is minimized. Note that the definition of η_A comes from minimizing the variance, and so it reflects the impact of different designs via covariance matrix, but it is not the relative effective sample size as used in MR-dualframe. In the case of common frames, $\hat{T}_{z,comp}$ is same for all ζ_A, and its choice could be based on minimizing the generalized variance for a set of other y-variables.

4. Asymptotic Properties of the DFC Estimator

It can be shown that $\hat{T}_y^{DFC}$ is the solution of the estimating equation $G\Gamma_g^{-1} g = 0$ where g is the $(q+1)$-vector $(\hat{T}_{y,naive}^{HT} - T_y,\ 1'\Gamma X - T_x')'$, Γ_g is the $(q+1)\times(q+1)$ working covariance matrix of g with first row as $(y'\Gamma y, y'\Gamma X)$ and the matrix of the last q rows as $(X'\Gamma y, X'\Gamma X)$, and G is the $(q+1)\times 1$ vector $(1, 0,\ldots,0)'$. Then the estimated asymptotic variance, $\hat{V}(\hat{T}_y^{DFC})$, of $\hat{T}_y^{DFC} - T_y$ has the sandwich form,

$$\hat{V}(\hat{T}_y^{DFC}) = B'G'\Gamma_g^{-1}\hat{V}(g)\Gamma_g^{-1}G(B^{-1})', \quad (4.1)$$

where $\hat{V}(g)$ is a consistent estimate of the true covariance matrix of g, and $B = G'\Gamma_g^{-1}G$ which is a scalar in our case. Note that the vector g consists of HT-estimators for various parameters, and therefore, standard results in sampling can be used for estimating its covariance matrix.

Now under the asymptotic setup of Isaki and Fuller (1982) for a sequence of finite populations and samples as $n_A,\ n_B,\ N_A,\ N_B \to \infty$ such that $n_A/n_B,\ N_A/N_B$ tend to positive constants, the HT estimators in vector g defined above are consistent estimates and follow the multivariate central limit theorem, i.e.,

$$n^{\frac{1}{2}} N^{-1}(g - 0) \to_d N_{q+1}(0, V(g)\, n/N^2). \quad (4.2)$$

Suppose also that $N^{-1}\Gamma_g$ converges in probability to a positive definite matrix where $N = N_A + N_B$. Now using the delta method, it follows that $\hat{T}_y^{DFC}$ is asymptotically design consistent. Moreover, it is asymptotically normal with mean T_y and variance given by (4.1), which can then be used for constructing confidence intervals.

5. Empirical Results

We conducted a simulation study to compare empirical MSE (denoted as EMSE) of GR(naïve), FB, and DFC estimators when the two frames are completely overlapping. Both x- and y-values for samples from frames A and B were generated according to a superpopulation model similar to Skinner-Rao which approximates asymptotically a two stage cluster sampling for frame A and a simple random sample for frame B for finite populations. Thus, the evaluation of the estimators corresponds to the infinite population case. The samples were generated as follows:

For frame A, we set $y_{ij} = \mu + bx_{ij} + \alpha_i + \varepsilon_{ij}$ where y_{ij} is the ith observation in cluster i, μ is the overall mean of y_{ij}, $x_{ij} \sim N(0,\sigma^2)$ generates the covariate values, b is the regression coefficient, $\alpha_i \sim N(0, \rho(1-b^2)\sigma^2)$ is the cluster effect, $\varepsilon_{ij} \sim N(0,(1-\rho)(1-b^2)\sigma^2)$ is the random error, and ρ is the intra-cluster correlation coefficient. All the random components are generated independently of each other. In this study, μ was set to 10, σ^2 was set at 5, b was set at .3, and ρ was chosen as .1 or .2. The population average of x-variables is zero while the average for counting variables for each frame is 1. The sample size for Frame A varied over 200, 400, and 600; and the sample size for Frame B over 300, 600, and 900. The sample size of each cluster in Frame A was set at 10.

A total of 100 replications was performed for each selected set of the design parameters. The empirical MSE (EMSE) is chosen to evaluate the performance. EMSE is defined as the average MSE of the estimators about the true value of 10 over the 100 replications. Table 1 shows the EMSEs for GR(naïve), FB, and DFC. The GR(naïve) is obtained by averaging the two GR estimates from each frame. It is seen from the evaluation results that the two methods FB and DFC have significantly smaller EMSE compared with GR, as expected. For cases considered, DFC performed better than FB, indicating the instability of the variance estimates under optimal regression.. The last column of Table 1 shows the average value of η_A obtained over the 100 runs via grid search.

It would be interesting to further increase the sample size to a

point where FB can consistently outperform DFC. It is planned to conduct a more extensive simulation study which allows for more x-and z-controls available for composite estimation.

6. Concluding Remarks

It can be summarized as follows:

(i) The DFC estimator can also be obtained in two steps, first GR for each frame to satisfy x-controls, and then the GR-weights are adjusted to satisfy z-controls. However, to continue to satisfy x-controls, the z-predictors need to be orthogonalized with respect to x-predictors as in partial regression.

(ii) To get range-restricted weights, the generalized exponential model of Folsom and Singh (2000) can be used with DFC; see Singh, Iannacchione, and Dever (2003) for an application.

(iii) It is easily seen that the DFC estimator remains applicable to dependent samples. The choice of the factors (η_A, ζ_A) is automatically adjusted to reflect dependence while minimizing the generalized variance.

(iv) In the presence of bias due to mode effects (such as under- or over-reporting), the zero functions are no longer unbiased, and the regression framework to get optimal or suboptimal estimator is not really applicable. However, using (3.2) as a model for bias correction, one can still use the same set of calibration equations to arrive at the same estimator (3.1). This observation is similar to the dual property of poststratification for both variance and bias (due to over/under coverage) reduction. Using z-controls along with x-controls, the DFC estimator implies that the difference in biases of the two estimators for z becomes zero. This implies that the bias of the DFC estimator is a compromise between biases of the individual ones. It is not bias-free but would be so as long as one of the two estimators is unbiased.

(v) The zero controls used in DFC are different from the x-controls in that there is no external estimate for the z-variable provided as a control total. In fact, the main purpose of the DFC method is to provide composite estimates in the calibration form for the z-variables used in the zero controls. This should be distinguished from the use of random controls in Zieschang (1990) and Renssen and Nieuwenbroek (1997) which are first obtained for use in the calibration step. In practice, it is desirable to obtain estimates of the selected z-variables that take advantage of the correlation with other z-variables in a multivariate dual frame set-up by using a set of final calibration weights, such that these same weights are later used to produce estimates for any other y-variable. This is what DFC offers in a way that it has built-in internal consistency with estimates for the z-variables. This problem of maintaining internal consistency is an important and interesting one, and is also addressed when using the empirical likelihood method for combining information from multiple surveys, see e.g., the recent paper by C. Wu (2003). However, with empirical likelihood, it seems difficult to allow for different designs and dependence of samples from multiple frames.

Acknowledgement The first author's research was partially supported by a grant from Natural Sciences and Engineering Research Council of Canada held at Carleton University, Ottawa. Thanks are due to C. Wu for sending a preprint of his paper.

References

Bankier, M.D. (1986), "Estimators Based on Several Stratified Samples with Applications to Multiple Frame Surveys," *Journal of the American Statistical Association*, 81, 1074-1079.

Folsom, R.E. Jr. and Singh, A.C. (2000). "A Generalized Exponential Model for Sampling Weight Calibration for a Unified Approach to Nonresponse, Post-stratification, and Extreme Weight Adjustments." ASA *Proc.Surv Res. Meth. Sec.*, pp. 598-603.

Fuller, W.A. (1975). Regression analysis for sample surveys. *Sankhya Series C*., **37**, 117-132.

Fuller, W.A., and Burmeister, L.F. (1972), "Estimators for Samples Selected from Two Overlapping Frames," in *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 245-249.

Godambe, V.P. and Thompson, M.E. (1989). An extension of quasi-likelihood estimation (with discussion). *Jour. Statistical Planning and Inference*, **22**, 137-172.

Hartley, H.O. (1962), "Multiple Frame Surveys," in Proceedings of the *Social Statistics Section, American Statistical Association*, pp. 203-206.

_____ (1974), "Multiple Frame Methodology and Selected Applications," *Sankhya, Ser. C*, 36, 99-118.

Isaki, C.T. and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.*, **77**, 89-96.

Kalton, G., and Anderson, D.W. (1986), "Sampling Rare Populations," *Journal of the Royal Statistical Society, Ser. A*, 149, 65-82.

Liang, K.-Y. and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**,13-22.

Lohr, S. L. and Rao, J.N.K. (2000). Inference from dual frame surveys. *Journal of the American Statistical Association*, 95, 271-280.

Lund, R.E. (1968), "Estimators in Multiple Frame Surveys," in *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 282-288.

Rao, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*,10, 153-165.

Rao, J.N.K.,and Scott, A.J. (1981). The analysis of categorical data from complex sample surveys: chi-squared tests for goodness-of-fit and independence in two-way tables. *J. Amer. Statist. Assoc.*, **76**, 221-230.

Renssen, R.H., and Nieuwenbroek, N.J. (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92, 368-374.

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

Singh, A.C. (1994). Sampling design-based estimating functions for finite population means. Invited paper, Abstracts of the Annual meeting of the *Statistical Society of Canada*, Banff, Alberta, May 8-11, p. 48.

Singh, A.C. (1996). Modified regression for combining information in survey sampling with applications. Invited paper, *ASA Proc. Surv. Res. Meth. Sec.* 120-129.

Singh, A.C., and S. Wu (1996). Estimation for Multiframe Complex Surveys by Modified Regression. Proceedings of the Statistical Society of Canada, Survey Methods Section, pp. 69-77.

Singh, A.C. and Sarndal, C.-E. (2003) Optimal Regression, Generalized Regression, and Modified Regression, manuscript under preparation.

Singh, A.C., Iannacchione, V.G.,and Dever, J.A. (2003). Efficient estimation for surveys with nonresponse follow-up using dual frame calibration, ASA Proc., Surv. Res. Meth. Sec. (in print)

Skinner, C.J., and Rao, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 349-356.

Wu, C. (2003). Combining information from multiple surveys through empirical likelihood method. *The Canadian Journal of Statistics*, 31, (in print)

Zieschang, K.D. 91990). Sample weighting methods and estimation of totals in the consumer expenditure survey. *Journal of the American Statistical Association*,85, 986-1001.

Table 1: EMSE (x100) of Estimators

EMSEx100						
sample size (A/B)	b	ρ	GR (naive)	FB	DFC	Av η_A
200/300	0.3	0.1	2.192	1.455	1.379	0.264
400/600	0.3	0.1	1.648	0.952	0.909	0.260
600/900	0.3	0.1	1.107	0.574	0.543	0.260
200/300	0.3	0.2	2.685	1.473	1.452	0.206
400/600	0.3	0.2	2.882	0.815	0.781	0.209
600/900	0.3	0.2	2.148	0.795	0.779	0.207