An Estimation Method for Matrix Survey Sampling

Takis Merkouris[*]

Abstract

Matrix sampling, sometimes referred to as a split-questionnaire, is a sampling design that involves dividing a questionnaire into subsets of questions, possibly overlapping, and then administering these subsets to different subsamples of an initial sample. This design reduces the data collection costs and addresses concerns related to response burden and data quality, but also reduces the number of sample units that are asked each question. For matrix survey sampling with overlapping subsets of questions, we propose an estimation method that exploits correlations among variables surveyed in the various subsamples in order to improve the precision of the survey estimates. The proposed method uses a suitable calibration scheme, which is equivalent to a generalized regression procedure based on the principle of best linear unbiased estimation. The method is computationally very convenient, and facilitates variance estimation.

Key Words: Split-questionnaire, calibration, generalized regression estimator, best linear unbiased estimator, composite estimator

1. Introduction

Matrix sampling is a sampling design in which a large questionnaire is divided into subsets of questions, possibly overlapping, and these subsets are then administered to different subsamples of an initial sample. In its various forms this design may serve a variety of purposes, such as reducing the length of the survey process and addressing concerns related to response burden and data quality associated with a large questionnaire. There is a long history of use of matrix sampling in other fields, primarily in educational statistics, but there is a paucity of related research or practice in survey sampling. A review of matrix sampling and a discussion of the issues arising in its implementation in surveys is given in Gonzalez and Eltinge (2007). For recent work on estimation for matrix survey sampling, see Gonzalez and Eltinge (2008), and Chipperfield and Steel (2009).

Recent uses of matrix sampling in various statistical agencies (e.g., US Bureau of Labor Statistics, British Office of National Statistics, Australian Bureau of Statistics), marking current trends and offering an outlook for future survey practice, relate to the integration of a number of existing independent household surveys for the additional benefit of streamlined survey operations, harmonized survey content and data consistency. In this non-ordinary matrix sampling setting, the distinct surveys may use subsamples of a large master sample or independent (and non-overlapping) samples from the same population. It is to be noted that the advantages of matrix sampling are not always contingent on using subsamples (necessarily dependent) of an initial sample. On the contrary, it may be more practical in certain situations to use independent samples.

We consider four basic matrix sampling designs, varied in the number of subsamples and the questions administered to each subsample:

a. Different (non-overlapping) sets of questions are administered to different (disjoint) subsamples.

b. An additional common set of questions is administered to each subsample of design (a). There are several reasons for surveying a core set of variables in all subsamples:

[*] Athens University of Economics and Business, Patision 76, 10434 Athens, Greece

Special interest in some of those variables, and required high precision for related estimates; some variables may define subpopulations, and be used in cross-tabulations of survey results; the correlation of these variables may be used for various purposes, the most important of which, in the context of this paper, is to enhance the precision of estimates for all variables.

c. This is a variant of design (a), which involves an additional subsample receiving the full questionnaire. The prime motivation for this scheme is to enhance the analytic capacity of the survey, by having responses to all questions from the units of the additional sample, and to aid the the treatment of missing values.

d. This is a variant of design (c), in which the additional common set of questions is administered to all subsamples. This design accommodates all survey requirements, embodying all features of the previous designs.

In this paper we address the estimation problem in matrix sampling. A serious trade-off in splitting a questionnaire is the reduction of the size of the sample that is available for each of the survey variables, which implies loss of precision of survey estimates. For matrix sampling designs (b), (c) and (d), involving overlapping subsets of questions, the dual estimation task is to combine data on common variables from different subsamples for improved estimation, and to exploit correlations among variables surveyed in different subsamples for more efficient estimation for all variables. We propose an estimation method that uses a suitable calibration scheme for the sampling weights of the combined sample, which is equivalent to a generalized regression procedure based on the principle of best linear unbiased estimation. For simplicity of the exposition we deal with a matrix sampling setting in which instead of subsamples of an initial sample we have independent but non-overlapping samples from the same population. In this context, the problem of combining data from the two independent samples in case (b) has been dealt with in the literature; see, for example, Renssen and Nieuwenbroek(1997), Merkouris (2004, 2010) and Wu (2004). In the following Sections 2 and 3 we describe the proposed method for design (c). We conclude with a discussion in Section 4.

2. Best Linear Unbiased Estimation

A general estimation method for matrix sampling is illustrated for design (c) in the simple case involving three independent samples S_1, S_2 and S_3, representing the population U, with vectors of variables $\mathbf{x}$ and $\mathbf{y}$ surveyed in S_1 and S_2, respectively, and both vectors surveyed in S_3.

We denote by $\mathbf{w}_i$ the vector of design weights for sample S_i, $i = 1, 2, 3$, and by $\mathbf{X}_i$ and $\mathbf{Y}_i$ the sample matrices for $\mathbf{x}$ and $\mathbf{y}$ — the subscripts indicating the sample. We obtain simple Horwitz-Thompson(H-T) estimates $\hat{\mathbf{X}}_1 (= \mathbf{X}_1' \mathbf{w}_1)$ and $\hat{\mathbf{X}}_3$ of the population total $\mathbf{t}_{\mathbf{x}}$, for $\mathbf{x}$, based on S_1 and S_3, respectively, and estimates $\hat{\mathbf{Y}}_2$ and $\hat{\mathbf{Y}}_3$ of the total $\mathbf{t}_{\mathbf{y}}$, for $\mathbf{y}$, based on S_2 and S_3. For more efficient estimation of the totals $\mathbf{t}_{\mathbf{x}}$ and $\mathbf{t}_{\mathbf{y}}$, we seek composite estimates $\hat{\mathbf{X}}^c$ and $\hat{\mathbf{Y}}^c$, respectively, that incorporate all the available information on $\mathbf{x}$ and $\mathbf{y}$ in the three samples. Such composite estimates that are best linear unbiased estimates (BLUE), i.e., minimum-variance linear unbiased combinations of the four estimates $\hat{\mathbf{X}}_1$, $\hat{\mathbf{Y}}_2$, $\hat{\mathbf{X}}_3$ and $\hat{\mathbf{Y}}_3$, are given in matrix form by

$$\begin{pmatrix} \hat{\mathbf{X}}^c \\ \hat{\mathbf{Y}}^c \end{pmatrix} = \boldsymbol{\mathcal{P}}(\hat{\mathbf{X}}_1, \hat{\mathbf{Y}}_2, \hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3)', \tag{1}$$

where $\boldsymbol{\mathcal{P}} = (\mathbf{W}'\mathbf{V}^{-1}\mathbf{W})^{-1}\mathbf{W}'\mathbf{V}^{-1}$, $\boldsymbol{\mathcal{P}}$ the matrix $\mathbf{W}$ expressing the condition of unbiasedness is such that $\boldsymbol{\mathcal{P}}\mathbf{W} = \mathbf{I}$, and $\mathbf{V}$ is the variance-covariance matrix of $(\hat{\mathbf{X}}_1, \hat{\mathbf{Y}}_2, \hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3)'$. This estimation method was proposed by Chipperfield and Steel (2009), who provided analytic expressions of the BLUE for scalars x and y assuming simple random sampling and known $\mathbf{V}$. They adopted the form of the BLUE found in the literature on composite estimation in a different context of survey sampling; see Fuller (1990), Wolter (1979) and Jones (1980). In general, computation of the BLUE is not practical, as only an approximation of the true $\boldsymbol{\mathcal{P}}$ would be conceivable (based on an estimated $\mathbf{V}$) but exceedingly difficult, especially if the number of variables or the number of samples is large. It would be hopeless if domain estimates were also of interest.

A simpler formulation of this estimation procedure is as follows. First, we express the composite estimates in (1) explicitly as linear combinations of the H-T estimates $\hat{\mathbf{X}}_1$, $\hat{\mathbf{Y}}_2$, $\hat{\mathbf{X}}_3$ and $\hat{\mathbf{Y}}_3$, i.e.,

$$\begin{aligned}\hat{\mathbf{X}}^c &= \mathbf{B}_{1\mathbf{x}}\hat{\mathbf{X}}_1 + \mathbf{B}_{2\mathbf{x}}\hat{\mathbf{Y}}_2 + \mathbf{B}_{3\mathbf{x}}\hat{\mathbf{X}}_3 + \mathbf{B}_{4\mathbf{x}}\hat{\mathbf{Y}}_3\\ \hat{\mathbf{Y}}^c &= \mathbf{B}_{1\mathbf{y}}\hat{\mathbf{X}}_1 + \mathbf{B}_{2\mathbf{y}}\hat{\mathbf{Y}}_2 + \mathbf{B}_{3\mathbf{y}}\hat{\mathbf{X}}_3 + \mathbf{B}_{4\mathbf{y}}\hat{\mathbf{Y}}_3.\end{aligned}$$

The condition of unbiasedness, $E(\hat{\mathbf{X}}^c) = \mathbf{t_x}$ and $E(\hat{\mathbf{Y}}^c) = \mathbf{t_y}$, implies that $\mathbf{B}_{3\mathbf{x}} = \mathbf{I} - \mathbf{B}_{1\mathbf{x}}$, $\mathbf{B}_{4\mathbf{x}} = -\mathbf{B}_{2\mathbf{x}}$ and $\mathbf{B}_{4\mathbf{y}} = \mathbf{I} - \mathbf{B}_{2\mathbf{y}}$, $\mathbf{B}_{3\mathbf{y}} = -\mathbf{B}_{1\mathbf{y}}$. Thus, $\boldsymbol{\mathcal{P}}$ and $\mathbf{W}$ can be expressed as

$$\boldsymbol{\mathcal{P}} = \begin{pmatrix} \mathbf{B}_{1\mathbf{x}} & \mathbf{B}_{2\mathbf{x}} & \mathbf{I} - \mathbf{B}_{1\mathbf{x}} & -\mathbf{B}_{2\mathbf{x}} \\ \mathbf{B}_{1\mathbf{y}} & \mathbf{B}_{2\mathbf{y}} & -\mathbf{B}_{1\mathbf{y}} & \mathbf{I} - \mathbf{B}_{2\mathbf{y}} \end{pmatrix}, \qquad \mathbf{W} = \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} & \mathbf{I} \end{pmatrix},$$

respectively, and the two composite estimates have necessarily the regression form

$$\begin{aligned}\hat{\mathbf{X}}^c &= \hat{\mathbf{X}}_3 + \mathbf{B}_{1\mathbf{x}}(\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3) + \mathbf{B}_{2\mathbf{x}}(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3)\\ \hat{\mathbf{Y}}^c &= \hat{\mathbf{Y}}_3 + \mathbf{B}_{1\mathbf{y}}(\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3) + \mathbf{B}_{2\mathbf{y}}(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3).\end{aligned} \tag{2}$$

Then writing $\boldsymbol{\mathcal{P}} = (\boldsymbol{\mathcal{B}},\ \mathbf{I} - \boldsymbol{\mathcal{B}})$, in obvious notation for $\boldsymbol{\mathcal{B}}$, we can write (1) as

$$\begin{pmatrix} \hat{\mathbf{X}}^c \\ \hat{\mathbf{Y}}^c \end{pmatrix} = \boldsymbol{\mathcal{B}} \begin{pmatrix} \hat{\mathbf{X}}_1 \\ \hat{\mathbf{Y}}_2 \end{pmatrix} + (\mathbf{I} - \boldsymbol{\mathcal{B}}) \begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \end{pmatrix} + \boldsymbol{\mathcal{B}} \begin{pmatrix} \hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3 \end{pmatrix}, \tag{3}$$

the right-hand side of (3) being the matrix form of (2). The problem of finding the optimal (variance-minimizing) $\boldsymbol{\mathcal{P}}$ of the BLUE in (1) reduces then to that of finding the optimal matrix $\boldsymbol{\mathcal{B}}$ in (3). An estimated optimal $\hat{\boldsymbol{\mathcal{B}}}^o$ is given by

$$\hat{\boldsymbol{\mathcal{B}}}^o = -\widehat{\text{Cov}}\left(\begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \end{pmatrix}, \begin{pmatrix} \hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3 \end{pmatrix} \right) \widehat{\text{Var}}^{-1} \begin{pmatrix} \hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3 \end{pmatrix},$$

and because of the assumed independence of the three samples it reduces to

$$\hat{\boldsymbol{\mathcal{B}}}^o = \widehat{\text{Var}} \begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \end{pmatrix} \left[\widehat{\text{Var}} \begin{pmatrix} \hat{\mathbf{X}}_1 \\ \hat{\mathbf{Y}}_2 \end{pmatrix} + \widehat{\text{Var}} \begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \end{pmatrix} \right]^{-1}.$$

With such optimal $\hat{\boldsymbol{\mathcal{B}}}^o$, the composite estimator (3) is an optimal regression estimator, and thus the BLUE in (1) with $\hat{\boldsymbol{\mathcal{P}}} = (\mathbf{W}'\hat{\mathbf{V}}^{-1}\mathbf{W})^{-1}\mathbf{W}'\hat{\mathbf{V}}^{-1}$ (involving the estimated $\hat{\mathbf{V}}$, and satisfying $\hat{\boldsymbol{\mathcal{P}}} = (\hat{\boldsymbol{\mathcal{B}}}^o, I - \hat{\boldsymbol{\mathcal{B}}}^o))$ is an optimal regression estimator.

Writing the variance of an H-T estimator as a quadratic form in the associated variable and with matrix $\mathbf{\Lambda}^0 = \{(\pi_{kl} - \pi_k\pi_l)/\pi_k\pi_l\pi_{kl}\}$ (in terms of first-and-second order probabilities of selection), and using some matrix algebra, it can be shown that

$$\hat{\boldsymbol{\mathcal{B}}}^o = (\boldsymbol{\mathcal{X}}_3'\mathbf{\Lambda}^0\boldsymbol{\mathcal{X}}_3)(\boldsymbol{\mathcal{X}}'\mathbf{\Lambda}^0\boldsymbol{\mathcal{X}})^{-1} = (\boldsymbol{\mathcal{X}}_3'\mathbf{\Lambda}^0\boldsymbol{\mathcal{X}}_3)(\boldsymbol{\mathcal{X}}_3'\mathbf{\Lambda}^0\boldsymbol{\mathcal{X}}_3 + \boldsymbol{\mathcal{X}}_{12}'\mathbf{\Lambda}^0\boldsymbol{\mathcal{X}}_{12})^{-1},$$

where

$$\boldsymbol{\mathcal{X}} = \begin{pmatrix} -\mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & -\mathbf{Y}_2 \\ \mathbf{X}_3 & \mathbf{Y}_3 \end{pmatrix}$$

is the design matrix corresponding to the regression setup of the regression estimator (3), $\boldsymbol{\mathcal{X}}_3$ is the matrix $\boldsymbol{\mathcal{X}}$ with the first two rows set equal to zero, $\boldsymbol{\mathcal{X}}_{12}$ is the matrix $\boldsymbol{\mathcal{X}}$ with the third row set equal to zero, and $\mathbf{\Lambda}^0$ is the block-diagonal matrix $\text{diag}(\mathbf{\Lambda}_i^0)$ with the matrix $\mathbf{\Lambda}_i^0$ associated with the sample S_i.

Although exact calculation of $\hat{\boldsymbol{\mathcal{B}}}^o$ is theoretically and computationally easier than calculation of $\hat{\boldsymbol{\mathcal{P}}}$, it is still a formidable task not least because the probabilities π_{kl} are not known for most sampling designs. The alternative approach of variance estimation by replication methods is also not practical. An approximately optimal composite regression estimator is developed in the next section.

3. Composite Generalized Regression Estimation

A computationally very convenient, but generally suboptimal, version of $\hat{\boldsymbol{\mathcal{B}}}^o$ is obtained by replacing the matrices $\mathbf{\Lambda}_i^0$ with the diagonal matrices $\mathbf{\Lambda}_i$, whose entries are the sampling weights of S_i. This gives the multivariate composite generalized regression estimator (CGREG)

$$\begin{pmatrix} \hat{\mathbf{X}}^{CGR} \\ \hat{\mathbf{Y}}^{CGR} \end{pmatrix} = \hat{\boldsymbol{\mathcal{B}}} \begin{pmatrix} \hat{\mathbf{X}}_1 \\ \hat{\mathbf{Y}}_2 \end{pmatrix} + (\mathbf{I} - \hat{\boldsymbol{\mathcal{B}}}) \begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \end{pmatrix} = \begin{pmatrix} \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_3 \end{pmatrix} + \hat{\boldsymbol{\mathcal{B}}} \begin{pmatrix} \hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3 \\ \hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3 \end{pmatrix}, \quad (4)$$

where $\hat{\boldsymbol{\mathcal{B}}} = (\boldsymbol{\mathcal{X}}_3'\mathbf{\Lambda}\boldsymbol{\mathcal{X}}_3)(\boldsymbol{\mathcal{X}}'\mathbf{\Lambda}\boldsymbol{\mathcal{X}})^{-1}$, and $\mathbf{\Lambda} = \text{diag}(\mathbf{\Lambda}_i)$.

The generalized regression procedure leading to the estimator (4) is a special calibration procedure, involving the combined sample $S = S_1 \cup S_2 \cup S_3$, whereby a vector of calibrated weights $\mathbf{c}$ is constructed to satisfy the constraints $\hat{\mathbf{X}}_1^{CGR} = \hat{\mathbf{X}}_3^{CGR}$ and $\hat{\mathbf{Y}}_2^{CGR} = \hat{\mathbf{Y}}_3^{CGR}$ while minimizing the generalized least-squares distance $(\mathbf{c} - \mathbf{w})'\mathbf{\Lambda}^{-1}(\mathbf{c} - \mathbf{w})$ between $\mathbf{c}$ and the vector $\mathbf{w}$ of the survey weights of S. This vector $\mathbf{c}$ is given by

$$\mathbf{c} = \mathbf{w} + \mathbf{\Lambda}\boldsymbol{\mathcal{X}}(\boldsymbol{\mathcal{X}}'\mathbf{\Lambda}\boldsymbol{\mathcal{X}})^{-1}(\mathbf{0} - \boldsymbol{\mathcal{X}}'\mathbf{w}),$$

and expression (4) is then obtained simply as $\boldsymbol{\mathcal{X}}_3'\mathbf{c}$. Note that using the calibrated weights of sample S_3 only, we obtain the composite estimators in (4) directly in the simple linear forms

$$\hat{\mathbf{X}}^{CGR} = \mathbf{X}_3'\mathbf{c}_3 = \sum_{S_3} c_k\mathbf{x}_k; \qquad \hat{\mathbf{Y}}^{CGR} = \mathbf{Y}_3'\mathbf{c}_3 = \sum_{S_3} c_k\mathbf{y}_k.$$

Yet, a decomposition of the vector $\mathbf{c}$ (Merkouris 2004) gives an analytical expression of (4) of the form (2), which sheds light onto the structure of the CGREG estimator. Thus, if we write $\boldsymbol{\mathcal{X}} = (\boldsymbol{X}, \mathbf{\Psi})$, then

$$\mathbf{c} = \mathbf{w} + \mathbf{L}_{\mathbf{\Psi}}\boldsymbol{X}(\boldsymbol{X}'\mathbf{L}_{\mathbf{\Psi}}\boldsymbol{X})^{-1}[\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3] + \mathbf{L}_{\boldsymbol{X}}\mathbf{\Psi}(\mathbf{\Psi}'\mathbf{L}_{\boldsymbol{X}}\mathbf{\Psi})^{-1}[\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3], \quad (5)$$

where $\mathbf{L}_{\boldsymbol{X}} = \mathbf{\Lambda}(\mathbf{I} - \mathbf{P}_{\boldsymbol{X}})$ with $\mathbf{P}_{\boldsymbol{X}} = \boldsymbol{X}(\boldsymbol{X}'\mathbf{\Lambda}\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbf{\Lambda}$, and $\mathbf{L}_{\mathbf{\Psi}} = \mathbf{\Lambda}(\mathbf{I} - \mathbf{P}_{\mathbf{\Psi}})$ with $\mathbf{P}_{\mathbf{\Psi}} = \mathbf{\Psi}(\mathbf{\Psi}'\mathbf{\Lambda}\mathbf{\Psi})^{-1}\mathbf{\Psi}'\mathbf{\Lambda}$. It follows that

$$\begin{aligned} \hat{\mathbf{X}}^{CGR} = \mathbf{X}_3'\mathbf{c}_3 &= \hat{\mathbf{X}}_3 + \hat{\mathbf{B}}_{1\mathbf{x}}(\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3) + \hat{\mathbf{B}}_{2\mathbf{x}}(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3) & (6) \\ &= \hat{\mathbf{B}}_{1\mathbf{x}}\hat{\mathbf{X}}_1 + (\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}})\hat{\mathbf{X}}_3 + \hat{\mathbf{B}}_{2\mathbf{x}}(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3), & (7) \end{aligned}$$

in obvious notation for $\hat{\mathbf{B}}_{1\mathbf{x}}$ and $\hat{\mathbf{B}}_{2\mathbf{x}}$. Similar is the expression for $\hat{\mathbf{Y}}^{CGR}$. It is seen from (7) that the CGREG estimator $\hat{\mathbf{X}}^{CGR}$ derives its efficiency from combining the two elementary estimators $\hat{\mathbf{X}}_1$ and $\hat{\mathbf{X}}_3$ (pooling information from samples S_1 and S_2) and from borrowing strength from sample S_2 through the correlation between $\mathbf{x}$ and $\mathbf{y}$. The vector $\mathbf{c}$ can be expressed as

$$\mathbf{c} = \mathbf{c}_{_\Psi} + \mathbf{L}_{\boldsymbol{\Psi}}\boldsymbol{X}(\boldsymbol{X}^{'}\mathbf{L}_{\boldsymbol{\Psi}}\boldsymbol{X})^{-1}[\mathbf{0} - \boldsymbol{X}^{'}\mathbf{c}_{_\Psi}],$$

where the vector

$$\mathbf{c}_{_\Psi} = \mathbf{w} + \boldsymbol{\Lambda}\boldsymbol{\Psi}(\boldsymbol{\Psi}^{'}\boldsymbol{\Lambda}\boldsymbol{\Psi})^{-1}[\mathbf{0} - \boldsymbol{\Psi}^{'}\mathbf{w}]$$

is generated by calibration of the design weights involving only $\boldsymbol{\Psi}$. Then, the estimator $\hat{\mathbf{X}}^{CGR}$ takes the forms

$$\begin{aligned}
\hat{\mathbf{X}}^{CGR} &= \mathbf{X}_3^{'}\mathbf{c}_{_{3\Psi}} + \mathbf{X}'_3\mathbf{L}_{\boldsymbol{\Psi}}\boldsymbol{X}(\boldsymbol{X}^{'}\mathbf{L}_{\boldsymbol{\Psi}}\boldsymbol{X})^{-1}[\mathbf{X}_1^{'}\mathbf{c}_{_{1\Psi}} - \mathbf{X}_3^{'}\mathbf{c}_{_{3\Psi}}] && (8)\\
&= \hat{\mathbf{X}}_3^{GR} + \hat{\mathbf{B}}_{1\mathbf{x}}[\hat{\mathbf{X}}_1 - \hat{\mathbf{X}}_3^{GR}] && (9)\\
&= \hat{\mathbf{B}}_{1\mathbf{x}}\hat{\mathbf{X}}_1 + (\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}})\hat{\mathbf{X}}_3^{GR}, && (10)
\end{aligned}$$

where $\hat{\mathbf{X}}_3^{GR}$ is the generalized regression (GREG) estimator $\hat{\mathbf{X}}_3 + \mathbf{X}_3^{'}\boldsymbol{\Lambda}\boldsymbol{\Psi}(\boldsymbol{\Psi}^{'}\boldsymbol{\Lambda}\boldsymbol{\Psi})^{-1}(\hat{\mathbf{Y}}_2 - \hat{\mathbf{Y}}_3)$, incorporating the regression effect of the last term in (7). The matrix regression coefficient $\hat{\mathbf{B}}_{1\mathbf{x}}$ is written explicitly as $\hat{\mathbf{B}}_{1\mathbf{x}} = \mathbf{X}_3^{'}\mathbf{L}_{\boldsymbol{\Psi}}\boldsymbol{X}(\mathbf{X}_1^{'}\boldsymbol{\Lambda}_1\mathbf{X}_1 + \mathbf{X}_3^{'}\mathbf{L}_{\boldsymbol{\Psi}}\boldsymbol{X})^{-1}$, where $\mathbf{X}_3^{'}\mathbf{L}_{\boldsymbol{\Psi}}\boldsymbol{X} = \mathbf{X}_3^{'}\boldsymbol{\Lambda}_3\mathbf{X}_3 - \mathbf{X}_3^{'}\boldsymbol{\Lambda}_3\mathbf{Y}_3(\mathbf{Y}_2^{'}\boldsymbol{\Lambda}_2\mathbf{Y}_2 + \mathbf{Y}_3^{'}\boldsymbol{\Lambda}_3\mathbf{Y}_3)^{-1}\mathbf{Y}_3^{'}\boldsymbol{\Lambda}_3\mathbf{X}_3$. If $\mathbf{x}$ and $\mathbf{y}$ were orthogonal (uncorrelated), or if information on $\mathbf{y}$ was not used in the estimation of $\mathbf{t}_{\mathbf{x}}$, then it would be $\hat{\mathbf{X}}_3^{GR} = \hat{\mathbf{X}}_3$ and $\hat{\mathbf{B}}_{1\mathbf{x}} = \mathbf{X}_3^{'}\boldsymbol{\Lambda}_3\mathbf{X}_3(\mathbf{X}_1^{'}\boldsymbol{\Lambda}_1\mathbf{X}_1 + \mathbf{X}_3^{'}\boldsymbol{\Lambda}_3\mathbf{X}_3)^{-1}$ and $\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}} = \mathbf{X}_1^{'}\boldsymbol{\Lambda}_1\mathbf{X}_1(\mathbf{X}_1^{'}\boldsymbol{\Lambda}_1\mathbf{X}_1 + \mathbf{X}_3^{'}\boldsymbol{\Lambda}_3\mathbf{X}_3)^{-1}$. But the GREG estimator $\hat{\mathbf{X}}_3^{GR}$ is more efficient than the H-T estimator $\hat{\mathbf{X}}_3$, and since $\mathbf{X}_1^{'}\boldsymbol{\Lambda}_1\mathbf{X}_1 + \mathbf{X}_3^{'}\mathbf{L}_{\boldsymbol{\Psi}}\boldsymbol{X} < \mathbf{X}_1^{'}\boldsymbol{\Lambda}_1\mathbf{X}_1 + \mathbf{X}_3^{'}\boldsymbol{\Lambda}_3\mathbf{X}_3$ (in the partial order of non-negative definite matrices), it is clear that more weight is given to $\hat{\mathbf{X}}_3^{GR}$ in (10), through $\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}} = \mathbf{X}_1^{'}\boldsymbol{\Lambda}_1\mathbf{X}_1(\mathbf{X}_1^{'}\boldsymbol{\Lambda}_1\mathbf{X}_1 + \mathbf{X}_3^{'}\mathbf{L}_{\boldsymbol{\Psi}}\boldsymbol{X})^{-1}$, than would have been given to the component estimator $\hat{\mathbf{X}}_3$ in the simple composite estimator involving only information on $\mathbf{x}$. This suggests that the CGREG estimator in (10), incorporating information from sample S_2, is a more efficient estimator.

If $\boldsymbol{\Lambda}_i$, $i = 1, 2$, is replaced by $\boldsymbol{\Lambda}_i^0$, in which case the estimator $\hat{\mathbf{X}}^{CGR}$ becomes the optimal composite regression estimator, we get $\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}}^{o} = \widehat{\mathrm{Var}}(\hat{\mathbf{X}}_1)[\widehat{\mathrm{Var}}(\hat{\mathbf{X}}_1) + \widehat{\mathrm{Var}}(\hat{\mathbf{X}}_3) - \widehat{\mathrm{Cov}}(\hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3)[\widehat{\mathrm{Var}}(\hat{\mathbf{Y}}_2) + \widehat{\mathrm{Var}}(\hat{\mathbf{Y}}_3)]^{-1}\widehat{\mathrm{Cov}}^{'}(\hat{\mathbf{X}}_3, \hat{\mathbf{Y}}_3)]^{-1}$. It is clear then that the stronger the correlation between $\mathbf{x}$ and $\mathbf{y}$ the larger the $\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}}^{o}$ and more weight is given to component $\hat{\mathbf{X}}_3^{GR}$. In this connection, it can been shown that the approximate (large sample) variance of the optimal composite regression estimator is $\widehat{\mathrm{Var}}(\hat{\mathbf{X}}_1)\hat{\mathbf{B}}_{1\mathbf{x}}^{o}$, and thus the stronger the correlation between $\mathbf{x}$ and $\mathbf{y}$ the smaller becomes this variance.

In contrast with $\hat{\mathbf{B}}_{1\mathbf{x}}^{o}$, the suboptimal regression coefficient $\hat{\mathbf{B}}_{1\mathbf{x}}$ gives a CGREG estimator $\hat{\mathbf{X}}^{CGR}$ which in general is somewhat less efficient than the optimal composite regression estimator. This observation extends to the overall regression coefficient $\hat{\boldsymbol{\mathcal{B}}}$ vis a vis the optimal coefficient $\hat{\boldsymbol{\mathcal{B}}}^{o}$ defined above. The matrix $\hat{\boldsymbol{\mathcal{B}}}$ is, nevertheless, optimal in the sense of minimizing the quadratic form in the sample residuals corresponding to the regression setup leading to the CGREG estimator (4). For certain sampling designs, $\hat{\boldsymbol{\mathcal{B}}} = \hat{\boldsymbol{\mathcal{B}}}^{o}$, and the CGREG estimator is optimal. For instance, this is true when the design for all three samples is Poisson and the kth entry of the matrix $\boldsymbol{\Lambda}_i$ is divided by $q_{ik} = \pi_{ik}/(1 - \pi_{ik})$. Other such designs include simple random sampling without replacement (SRRWOR) and stratified (SRSWOR) with proper adjustments q_{ik} to the entries of $\boldsymbol{\Lambda}_i$ and with an intercept included in the design matrix $\boldsymbol{\mathcal{X}}$ — see remark in the next paragraph. This property is based on arguments found in Merkouris (2010). For general designs, the value of $q_{ik} = n_i$ should be used in the adjustment of the entries of $\boldsymbol{\Lambda}_i$ to take into account the differential in effective sample sizes n_i.

The three samples may collect information on some common auxiliary variables $\mathbf{z}$ for which the population totals $\mathbf{t}_{\mathbf{z}}$ are known. Then, the expression (7) of the CGREG estimator

$\hat{\mathbf{X}}^{CGR}$ may include the additional ordinary GREG terms $\hat{\mathbf{B}}_{3\mathbf{x}}(\mathbf{t}_{\mathbf{z}} - \hat{\mathbf{Z}}_1) + \hat{\mathbf{B}}_{4\mathbf{x}}(\mathbf{t}_{\mathbf{z}} - \hat{\mathbf{Z}}_2) + \hat{\mathbf{B}}_{5\mathbf{x}}(\mathbf{t}_{\mathbf{z}} - \hat{\mathbf{Z}}_3)$, where $\hat{\mathbf{Z}}_i, i = 1, 2, 3$ is the H-T estimates of $\mathbf{t}_{\mathbf{z}}$ based on $\boldsymbol{\Lambda}_i$. This estimator has improved efficiency, as it incorporates additional information, and is generated by a calibration procedure that includes the additional three constraints $\hat{\mathbf{Z}}_i^{CGR} = \mathbf{t}_{\mathbf{z}}$, and with the design matrix $\boldsymbol{\mathcal{X}}$ augmented with the block-diagonal matrix $\mathbf{Z} = \text{diag}(\mathbf{Z}_i)$. In analogy with (10), but with different matrix coefficient $\hat{\mathbf{B}}_{1\mathbf{x}}$, the composite estimator $\hat{\mathbf{X}}^{CGR}$ takes now the form

$$\hat{\mathbf{X}}^{CGR} = \hat{\mathbf{B}}_{1\mathbf{x}}\hat{\mathbf{X}}_1^{GR} + (\mathbf{I} - \hat{\mathbf{B}}_{1\mathbf{x}})\hat{\mathbf{X}}_3^{GR},$$

where $\hat{\mathbf{X}}_1^{GR}$ and $\hat{\mathbf{X}}_1^{GR}$ are GREG estimators using all information on $\mathbf{y}$ and $\mathbf{z}$ in the three samples. This very realistic sampling setting, with the three samples having some common variables with known totals, is in fact a special case of matrix sampling design (d). In the simplest case when $\mathbf{Z}_i$ is the unit column $\mathbf{1}$ (with corresponding total the size of the population), the regression setup contains an intercept and the CGREG estimator is optimal for the sampling designs mentioned in the preceding paragraph.

4. Discussion

The proposed estimation method for matrix sampling produces composite estimators of totals which are approximately (for large samples) BLUE, that is, they are approximately unbiased, as special regression estimators, and approximately of minimum variance — in certain sampling settings they are exactly BLUE.

The proposed method is computationally very convenient, requiring only a simple adaptation of the generalized regression procedure commonly used in statistical agencies. Operationally, it involves a single-step calibration of the weights of the combined sample. Estimates for all variables and at any population level can thus be obtained by using only the relevant units of sample S_3 and their calibrated weights incorporating all the available information from all three samples. Furthermore, carrying out the described calibration procedure on the combined sample greatly facilitates variance estimation by replication methods, such as the jackknife.

A generalization of the estimation method for matrix sampling with more than two sets of questions is straightforward, making more evident the operational power of the calibration procedure. The estimation method for matrix sampling scheme (d), and for dependent subsamples of an initial sample will be discussed elsewhere.

REFERENCES

Chipperfield, J.O, and Steel, D.G. (2009), "Design and Estimation for Split Questionnaire Surveys," *Journal of Official Statistics*, 25, 227–244.

Fuller, W.A. (1990), "Analysis of Repeated Surveys," *Survey Methodology*, 16, 167–180.

Gonzalez, J.M, and Eltinge, J.L. (2007), "Multiple Matrix Sampling: A review," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 3069–3075.

Gonzalez, J.M, and Eltinge, J.L. (2008), "Adaptive Matrix Sampling for the Consumer Expenditure Quarterly Interview Survey," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, 3069–3075.

Jones, R.G. (1980), "Best Linear Unbiased Estimators for Repeated Surveys," *Journal of the Royal Statistical Society*, Ser. B, 42, 221–226.

Merkouris, T. (2004), "Combining Independent Regression Estimators from Multiple Surveys," *Journal of the American Statistical Association*, 99, 1131–1139.

Merkouris, T. (2010), "Combining Information from Multiple Surveys by Using regression for More Efficient Small Domain Estimation," *Journal of the Royal Statistical Society*, Ser. B, 72, 27–48.

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, 368–375.

Wolter, K.M. (1979), “Composite estimation in Finite Populations,” *Journal of the American Statistical Association*, 74, 604–613.

Wu, C. (2004), “Combining Information from Multiple Surveys Through the Empirical Likelihood Method,” *Canadian Journal of Statistics*, 32, 15–26.