Efficiency of Standard Regression Model-based Ratio-synthetic Estimators in Sample Surveys Combining Time Series and Cross-sectional Data

P. D. Ghangurde
Consultant, 1370 Plante Drive, Ottawa, ON K1V9G3

Abstract
Efficiency of ratio-synthetic estimator compared to BLUP estimator , both based on cross-sectional data, depends on ratios of harmonic means of auxiliary variable values of sample units to sample sizes in respective domains, domain of interest Ui and complementary domain Uc. Assuming equal unit error variances in model-based domain estimation BLUP estimator of domain mean or total is less efficient when compared to ratio-synthetic estimator (Ghangurde, P. D. ; JSM (2014)). In this paper efficiencies of ratio-synthetic estimators based on time series data as compared to the same based on cross-sectional data are analyzed for various values of correlation coefficient in AR(1) model for errors in the regression model assuming 5 and 10 time-points or occasions with diagonal covariance matrix for sampling errors. These efficiencies are not affected by small area domain effects in the mixed model assumed to derive BLUP estimator. Empirical results on gains in efficiencies of ratio-synthetic are compared to those of BLUP estimator(pages 159-62; Rao, J.N.K. (2003)). Empirical results can be similarly obtained for ratio-synthetic estimators in household surveys with rotation sample designs.

Key Words: Efficiency, ratio-synthetic, model-based, time-series, cross-sectional, data

1. Introduction

In BLUP extension of ratio-synthetic estimator $\bar{X}i.\ \hat{\beta}$, $\bar{X}i.$ is known mean of auxiliary variable (e.g. buiness or household income) values of units in the domain of interest Ui and $\hat{\beta}$ is estimator of β, regression coefficient in the mixed model. In household surveys the auxiliary variable is population of areal units such as primary sampling units or clusters of households within strata. Evaluation of variance of ratio-synthetic as compared to m.s.e. of BLUP estimator was done for several sample sizes and harmonic means in the domain of interest Ui and complementary domain Uc, unit error variances and variances of domain effects in the mixed model. The evaluation has provided new theoretical and empirical results on efficiencies of BLUP and

Prepared for presentation at the meetings of the Survey Research Methods section of the ASA August 8 - 13, 2015, Seattle, USA.

ratio-synthetic estimators assuming sample survey framework. The important conclusion from the study is that in domain estimation in sample surveys assuming two domains Ui and Uc, equal unit error variances in the model ratio-synthetic estimator is more efficient than BLUP estimator based on the mixed model. In ratio-synthetic estimator $\bar{X}i.\ \hat{\beta}$, $\hat{\beta}$ and its variance are based on standard regression model. Alternative methods for obtaining total population in the domain of interest Ui in inter-census period have also been briefly reviewed in a paper on the study (Ghangurde, P. D.; JSM (2014)).

Whatever methodology is used in practice to obtain auxiliary variable and population totals the theoretical and empirical results on efficiency comparison of ratio-synthetic and BLUP estimators referred to above hold good, since the totals are assumed to be known.

Efficiencies of ratio-synthetic estimators are not affected by small area domain effects in the mixed model assumed in the case of BLUP estimators (pages 135–38; Rao, J.N.K. (2003)). In this paper efficiencies of ratio-synthetic estimators, based on time-series data vs cross-sectional data, have been analyzed by taking into consideration efficiencies of ratio-synthetic as compared to BLUP estimators both based on cross-sectional data (Table 3; Ghangurde, P. D; JSM(2014)) .

2. Ratio-synthetic Estimation Using Time-series vs Cross-sectional Data

Let population U consist of N units; a sample of n units is drawn from U by simple random sampling without replacement. Let ni (> 0) and nc=(n – ni) be units in the sample from domains Ui and Uc respectively. Let Ni and Nc = N – Ni be total number of units in domains Ui and Uc respectively.

We assume that ni sample units from Ui and (n – ni) sample units from Uc satisfy the model :

$$Yijt = \chi ijt\ \beta + \epsilon ijt\ ;\ t = 0,1,\ldots,T;\ j = 1,\ldots, ni;\ j\ \varepsilon\ Ui;$$

$$Ycjt = \chi cjt\ \beta + \epsilon cjt\ ;\ t = 0,1,\ldots,T;\ j = (ni+1),\ldots, n\ ;\ j\ \varepsilon\ Uc, \quad (2.1)$$

where β is regression coefficient, Єijt and Єcjt are i. i. d. errors; χijt and χcjt are x-values of jth sample unit for t = 0,1,...., T; Yijt and Ycjt are y-values of jth sample unit for t = 0,1, …., T in Ui and Uc respectively. We assume Yijt ≥ 0 and Ycjt ≥ 0 and χijt > 0 and χcjt > 0, which includes the case χijt = χcjt = 1 for sample surveys. We assume that E(Єijt) = E(Єcjt) = 0; V(Єijt) = V(Єcjt)= σ^2 . Thus equal unit error variances are assumed in the model (2.1) .

When auxiliary variable is quantitative (counts or continuous) unit variances are unequal and proportional to their values, which was the assumption made in derivation of model-based domain and ratio-synthetic estimators (Ghangurde P. D.; JSM (2012)).

The model (2.1) when $t = 0$ is the model for cross-sectional data. Given $\bar{X}_{i.}$ as known mean of auxiliary variable values for U_i the ratio-synthetic estimator is defined as :

$$\bar{X}_{i.}\ \hat{\beta}\ , \tag{2.2}$$

where $\hat{\beta}$ is estimator of regression coefficient β under model (2.1) based on ni and (n-ni) sample units from U_i and U_c respectively for cross-sectional data. It is given by

$$\hat{\beta} = [\ X_i' V_i^{-1} X_i + X_c' V_c^{-1} X_c\ ]^{-1}\ [\ X_i' V_i^{-1} Y_i + X_c' V_c^{-1} Y_c\ ]\ . \tag{2.3}$$

The variance of $\hat{\beta}$ based on cross-sectional data is given by

$$V_0(\hat{\beta}) = [\ X_i' V_i^{-1} X_i + X_c' V_c^{-1} X_c\ ]^{-1}\ , \tag{2.4}$$

where $V_i^{-1} = \sigma^{-2} [\ I_{ni}\ ]$; $V_c^{-1} = \sigma^{-2} [\ I_{(n-ni)}\ ]$;

$$X_i' = [\ \underset{1 \le j \le ni}{\text{row}} (\chi_{ij0})\ ]; \quad X_c' = [\ \underset{(ni+1) \le j \le n}{\text{row}} (\chi_{cj0})\ ]\ . \tag{2.5}$$

Thus

$$V_0(\hat{\beta}) = \sigma^2 [\ \sum_{j=1}^{ni} \chi_{ij0}^2 + \sum_{j=(ni+1)}^{n} \chi_{cj0}^2\ ]^{-1} \tag{2.6}$$

is the variance of $\hat{\beta}$ for cross-sectional data assuming auxiliary variable values greater than zero. The estimator of regression coefficient is given by

$$\hat{\beta} = [\ \sum_{j=1}^{ni} Y_{ij0}\, \chi_{ij0} + \sum_{j=(ni+1)}^{n} Y_{cj0}\, \chi_{cj0}\ ]\ /\ [\ \sum_{j=1}^{ni} \chi_{ij0}^2 + \sum_{j=(ni+1)}^{n} \chi_{cj0}^2\ ]. \tag{2.7}$$

However, for sample surveys $\chi ij0 = \chi cj0 = 1$ and $V(\hat{\beta}_0) = \sigma^2 / n$. This is the case of interest for evaluation of efficiency of ratio-synthetic estimator which uses time-series data in AR(1) model for sampling errors in model (2.1) as compared to that based on cross-sectional data.

For time-series data of T time points we will assume a first order auto-regressive process for sampling errors Єijt, Єcjt ; t=1,....., T. The estimator $\hat{\beta}$ of β for occasion t = 0 is based on the same sample units. Assuming values of correlation coefficient ρ , | ρ | < 1 , in the AR(1) model for sampling errors in the standard regression model reduces variance of $\hat{\beta}$ and increases efficiency. The AR(1) model is an extension for sample units in Ui and Uc in sample survey framework of model (5.4.8) (see page 83; Rao, J.N.K.(2003)) and is as follows:

$$Єijt = \rho\ Єij,t\text{-}1 + eijt;\ t = 1 .....,T ;\ j=1,.....,ni ;\ j \ \varepsilon\ Ui ;$$

$$Єcjt = \rho\ Єcj,t\text{-}1 + ecjt;\ t = 1,.....,\ T;\ j=(ni+1),......,n;\ j\ \varepsilon\ Uc , \quad (2.8)$$

where E(eijt) = 0; E(ecjt) = 0; V(eijt) = V(ecjt) = σ^2. The covariance matrices of **Єi** = (Єij1,..., ЄijT) and **Єc** = (Єcj1,..., ЄcjT) are the same for each sample unit in Ui and Uc due to the same ρ assumed in Ui and Uc. Thus for ni and (n-ni) sample units in Ui and Uc respectively we have block diagonal covariance matrices $\sigma^2 \Lambda i = \sigma^2 \Lambda c$, where $\Lambda i = \Lambda c = \Lambda$.

The (t, s) th element of Λ is given by

$$\rho^{|t-s|} / (1 - \rho^2) ; \quad t = 1,......,T \ ;s = 1,.....,T . \qquad (2.9)$$

Thus in the case of time-series of T time-points t = 1,......, T variance of $\hat{\beta}$, as extension of (2.4) under model (2.1) with t = 0 to AR(1) model (2.8), is :

$$V(\hat{\beta}_T) = [\ Xi' \operatorname{diag}_{1 \le j \le ni} [\ \boldsymbol{\chi ij}' Vij^{-1}\ \boldsymbol{\chi ij}\ ]\ Xi + Xc' \operatorname{diag}_{(ni+1) \le j \le n} [\ \boldsymbol{\chi cj}' \ Vcj^{-1}\ \boldsymbol{\chi cj}]\ Xc\ ]^{-1}, \quad (2.10)$$

where $Xi'=[row_{1 \le j \le ni}(\chi ij0)]$; $Xc' = [\ row_{(ni+1) \le j \le n}(\chi cj0)]$; $\boldsymbol{\chi ij}' = [row_{1 \le t \le T} (\chi ijt)\ ]$; $\boldsymbol{\chi cj}' =[row_{1 \le t \le T} (\chi cjt)]$;

$$V_{ij} = \sigma^2 \,[\, I + \Lambda \,] \,; \qquad V_{cj} = \sigma^2 \,[\, I + \Lambda \,] \,, \tag{2.11}$$

where Vij and Vcj have been defined for AR(1) model (2.8) for sampling errors and I is T x T identity matrix. Let variances

$\boldsymbol{\chi_{ij}}' V_{ij}^{-1} \boldsymbol{\chi_{ij}} = \sigma^{-2} S(i,j,T,\rho)$ and $\boldsymbol{\chi_{cj}}' V_{cj}^{-1} \boldsymbol{\chi_{cj}} = \sigma^{-2} S(c,j,T,\rho)$, which are scalars based on data of T time-points, $\boldsymbol{\chi_{ij}}$ and $\boldsymbol{\chi_{cj}}$. The notation $\sigma^{-2} S(i,j,T,\rho)$ and $\sigma^{-2} S(c,j,T,\rho)$ for these variances indicates that these are obtained for given T and ρ for j th sample unit in Ui and Uc respectively. Thus

$$V_T(\hat{\beta}) = [\, W_i' X_i + W_c' X_c \,]^{-1},$$

where $W_i' = \sigma^{-2} [\, \mathrm{row}(\chi_{ij0} S(i,j,T,\rho)) \,]_{1 \le j \le n_i}$; $W_c' = \sigma^{-2} [\mathrm{row}(\chi_{cj0} S(c,j,T,\rho))]_{(n_i+1) \le j \le n}$.

Hence

$$V_T(\hat{\beta}) = \sigma^2 \Big[\sum_{j=1}^{n_i} \chi_{ij0}^2 \, S(i,j,T,\rho) + \sum_{j=(n_i+1)}^{n} \chi_{cj0}^2 \, S(c,j,T,\rho) \Big]^{-1}. \tag{2.12}$$

3. Efficiency of Ratio-synthetic Estimator under AR(1) Model for Time-series Data as Compared to Cross-sectional Data

The efficiency of ratio-synthetic estimator $\bar{X}_{i.} \hat{\beta}$ based on time-series data as compared to that based on cross-sectional data for t = 0, assuming known $\bar{X}_{i.}$, is defined as:

$$\text{Efficiency} = V_0(\hat{\beta}) / V_T(\hat{\beta}) \tag{3.1}$$

$$= \frac{\Big[\sum_{j=1}^{n_i} \chi_{ij0}^2 \, S(i,j,T,\rho) + \sum_{j=(n_i+1)}^{n} \chi_{cj0}^2 \, S(c,j,T,\rho) \Big]}{\Big[\sum_{j=1}^{n_i} \chi_{ij0}^2 + \sum_{j=(n_i+1)}^{n} \chi_{cj0}^2 \Big]} . \tag{3.2}$$

If error variances are unequal with $V(\epsilon_{ijt}) = \chi_{ijt}\sigma^2$ and $V(\epsilon_{cjt}) = \chi_{cjt}\sigma^2$; $t=1,....,T$ for jth sample units in U_i and U_c respectively, $S(i,j,T,\rho)$ and $S(c,j,T,\rho)$ will have to be redefined for appropriate time-series model. The case of unequal unit error variances has not been considered in this paper.

In the case of sample surveys $\chi_{ij0} = 1$; $j = 1,...,n_i$; $j \varepsilon U_i$ and $\chi_{cj0} = 1$; $j = (n_i+1),....,n$; $j \varepsilon U_c$ under unified model (see Ghangurde, P.D.; JSM (2014)).

Also, $\chi_{ijt} = \chi_{cjt} = 1$; $t=1,...,T$; $j \varepsilon U_i$ and $j \varepsilon U_c$; $\boldsymbol{\chi_{ij}' = \chi_{cj}' = 1'}$ are row vectors and $\boldsymbol{\chi_{ij} = \chi_{cj} = 1}$ are column vectors of T elements equal to 1.

Thus $S(i,j,T,\rho) = \sigma^2 [1' V_{ij}^{-1} 1]$ and $S(i,j,T,\rho) = \sigma^2 [1' V_{cj}^{-1} 1]$. Substituting these values in (3.2) efficiency of ratio-synthetic estimator for time series as compared to cross-sectional data is:

$$\text{Efficiency} = \sigma^2 \left[\, n_i [1' V_{ij}^{-1} 1] + (n - n_i)[1' V_{cj}^{-1} 1] \,\right] / [n_i + (n - n_i)], \quad (3.3)$$

where $V_{ij} = \sigma^2 [I + \Lambda]$ and $V_{cj} = \sigma^2 [I + \Lambda]$ for any j th sample unit in U_i and U_c. Thus (3.3) reduces as

$$\text{Efficiency} = \text{Total of values of terms in } [I + \Lambda]^{-1}, \quad (3.4)$$

where I is TxT identity matrix and Λ is as defined in (2.9) .

For $T = 5$ and $T = 10$ substituting $\rho = 0.3, 0.5, 0.7$ and 0.9 efficiencies were obtained. These are presented in Table 1 below. The efficiencies are greater than those for BLUP estimator (pages 159 – 62; Rao, J. N. K.(2003)).

Table 1: Efficiencies of Ratio-synthetic Estimator under AR(1) Model in Sample Surveys for Time-series as Compared to Cross-sectional Data

ρ	**T = 5**	**T = 10**
0.3	**1.81**	**3.45**
0.5	**1.28**	**2.26**
0.7	**0.72**	**1.14**
0.9	**0.22**	**0.26**

The domain effects assumed in the model to derive BLUP estimator reduce efficiencies of BLUP estimator as compared to ratio-synthetic estimator in sample surveys assuming $\sigma^2 = 5\,\sigma_i^2 = 5\,\sigma_c^2$. For sample size n = 100 , sample sizes in Ui, ni = 2 to 80, efficiency of ratio-synthetic as compared to BLUP estimator varies from 13.15 to 1.37 (see Table 3; Ghangurde; P. D.;JSM (2014)). If we assume $\sigma^2 = \sigma_i^2 = \sigma_c^2$ ratio-synthetic estimator for cross-sectional data is expected to be even more efficient. There are no studies on efficiency of ratio-synthetic estimator as compared to BLUP estimator in Rao, J.N.K.(2003).

The efficiencies are greater for T = 10 than for T = 5. The efficiencies decrease for increasing values of ρ and are less than 1 for T = 5 and $\rho = 0.7$ and 0.9 and T = 10 and $\rho = 0.9$.

Due to sample survey framework assumed in the models (2.1) and (2.8) efficiencies can be obtained for assumed values of ρ without simulation.

4. Concluding Remarks

The covariance matrix for errors in the standard regression model is diagonal due to assumption about errors made in model-based domain, synthetic, ratio-synthetic and BLUP estimation as in previous papers. By assuming sample survey framework the model for domain estimation gives basic theoretical and empirical results on relationship between estimators used in sample surveys (see Ghangurde, P. D. ; JSM (2012) and JSM (2014)).

In sample surveys efficiency is obtained by substituting values $\chi_{ijt} = \chi_{cjt} = 1$; $t=1,\ldots,T$; $j \in U_i$ and $j \in U_c$ in (3.2) . Thus we have simpler formula for efficiency (3.4). For rotation sample designs used in household surveys empirical values in covariance matrix with correlated errors can be used in place of matrix $[I + \Lambda]$ in (3.4) to obtain efficiencies. The AR(1) model-based covariance matrix will not be needed.

References

[1] Ghangurde, P. D. (2012), Small area estimation in household surveys when auxiliary variable totals are known, Presented at the Joint Statistical Meetings of the ASA, San Diego, July 28 – August 2, 2012 .

[2] Ghangurde, P. D. (2014), Evaluation of efficiency of standard regression mixed model-based BLUP estimators in household surveys assuming unequal error variances, Presented at the Joint Statistical Meetings of the ASA, Boston, August, 3 - 8, 2014.

[3] Rao, J. N. K. (2003), Small Area Estimation, Wiley-Interscience .