Standard Regression Model-based Ratio-synthetic Estimators Assuming Unequal or Equal Unit Error Variances and Their Use in Survey Practice

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

Abstract

Assuming sample survey framework of two domains, domain of interest Ui and complementary domain Uc in sample design strata, unequal unit error variances proportional to auxiliary variable values, an assumption appropriate in household surveys, ratio-synthetic estimator was proved to be more efficient than mixed model-based BLUP estimator. Two components of approximate efficiency of ratio-synthetic estimator were derived, assuming known domain population totals or quantitative auxiliary variable totals (Ghangurde, P. D. (2014)). In this paper unit error variances are assumed to be equal, an assumption appropriate in most other sample surveys. Approximate efficiency of ratio-synthetic estimator is derived by unconditional analysis. The results in the case of sample surveys under unified model are similar to those in earlier paper. Approximate efficiencies under two models are compared assuming that auxiliary variable has lognormal distribution. Ratio-synthetic is more efficient than BLUP in sample surveys under standard regression model assuming unequal or equal unit error variances; it is simple to use as compared to BLUP. Some applications in survey practice and methods to estimate domain totals and means are reviewed.

Key Words: Efficiency, ratio-synthetic, model-based, BLUP, survey, practice

1. Introduction

Standard regression mixed model-based BLUP estimator is an extension of regression-synthetic estimator, obtained by addition of estimator of small area domain effect ηi in the mixed model for domain of interest Ui. Thus BLUP estimator of domain mean of y-variable $\bar{\mu}i$ based on vector of p (≥ 1) domain means of auxiliary variable values $\bar{X}i.$ and p regression coefficients βk is:

$$\hat{\bar{\mu}}i = \bar{X}i.' \hat{\beta} + \hat{\eta}i \; ; \; i = 1,2,....,m \; , \qquad (1.1)$$

where $\hat{\beta}$ is vector of estimators βk; k = 1,2....,p in the mixed model and $\hat{\eta}i$ is estimator of ηi. In the framework of Type-B model (see pages 135–137; Rao, J.N.K.(2003)) there are m (≥ 2) domains called small areas unlike in the

---

Prepared for presentation at the meetings of the Survey Research Methods Section of the ASA, July 30 – August 4, 2016, Chicago, USA.

framework of domain estimation in sampling theory, where domain of interest Ui and complementary domain Uc are two domains within each stratum or strata enclosing Ui (Hartley, H. O. (1959); pages 34-38;Cochran,W.G. (1977)). The mixed model-based BLUP estimator has been used in survey practice since 1990's; however, results on relationship between standard regression model-based domain estimator, synthetic estimator and ratio-synthetic estimator derived from it were first proved by the author (Ghangurde, P. D. (2012) and (2013)). Under standard regression model ratio-synthetic estimator was obtained from synthetic estimator by a simple substitution. It was also proved that under the model ratio-synthetic estimation makes the same well-known assumption as made in synthetic estimation.

Assuming a random sample of size n units from a stratum of sample design of a household survey, ni (> 0) sample units from Ui and (n-ni) sample units from

Uc, known domain mean $\bar{X}i.$ of one auxiliary variable, efficiency of ratio-synthetic as compared to BLUP estimator was first derived by the author assuming model with unequal unit error variances. The approximate efficiency is determined by ratios of harmonic means of auxiliary variable values of sample units to sample sizes and by ratio of total sample size to that in domain Ui (Ghangurde, P. D. (2014)).

For high values of ratios $\bar{X}ih/ ni$ and $\bar{X}ch/(n-ni)$, where $\bar{X}ih$ and $\bar{X}ch$ are harmonic means of x-values of sample units in Ui and Uc respectively, m.s.e of BLUP is close to variance of ratio-synthetic estimator but is still greater than the latter. For low values of these ratios of harmonic means, as in sample surveys, ratio-synthetic estimator is far more efficient than BLUP estimator (Ghangurde, P. D. (2014)). This also holds in other sample surveys of households (e.g.Survey of Household Spending and LFS sample design-based surveys). These household surveys use the same estimation and variance estimation methodology as used in the LFS.

Standard regression model with equal unit error variances, appropriate in model -based domain estimation in most other surveys, is introduced in Section 2. In Section 3 approximate efficiency of ratio-synthetic estimator in sample surveys derived under unified model has the same expression as in Ghangurde, P. D. (2014), except that the second term is multiplied by square of inverse sampling ratio. When the auxiliary variable is quantitative the efficiency is obtained by assuming lognormal distribution and substituting its moments as approximations for means and means of squares of sample x-values from two domains. The differences in empirical results under these two models can be attributed to assumptions about errors made in the models . The use of lognormal as a distribution of auxiliary variable in model-based domain estimation gives these remarkable empirical results.

In view of efficiency of ratio-synthetic estimator as compared to BLUP estimator under standard regression model with unequal and also equal unit

error variances in sample surveys use of ratio-synthetic estimation in sample surveys is appropriate. In Section 4 its use in survey practice and methods involved are reviewed. In Section 5 some concluding remarks are given.

2. The Model and Model-based Estimators

We assume a population U of N units. A sample of size n units is drawn from U by simple random sampling without replacement. Let ni (>0) be number of sample units from Ni units in Ui and (n-ni) be number of sample units from (N-Ni) units in Uc. The mixed model for domain estimation assuming one auxiliary variable is defined as :

$$Yij = \text{Ƙ}ij\ \beta + \eta i + \text{Є} ij;\ j = 1,....,ni;\ j \varepsilon Ui;$$

$$Ycj = \text{Ƙ}cj\ \beta + \eta c + \text{Є}cj;\ j = (ni+1),....,n;\ j \varepsilon Uc, \quad (2.1)$$

where Yij and Ycj are y-values of jth sample unit from Ui and Uc and Ƙ ij and Ƙcj are x-values of jth sample unit from Ui and Uc respectively. We assume $Yij \geq 0$; $Ycj \geq 0$; $\text{Ƙ}ij > 0$; and $\text{Ƙ}cj > 0$. It includes the case $\text{Ƙ}ij = \text{Ƙ}cj = 1$, which implies that jth units in Ui and Uc are in the sample. The assumptions $\text{Ƙ}ij = 1$ and $\text{Ƙ}cj = 1$ under unified model (Ghangurde, P. D. (2014)) make analysis of efficiency possible for surveys in which unit error variances are equal, following the same approach as in LFS sample design-based surveys in which unit error variances are unequal. The estimator of regression coefficient β and its variance are derived in this paper assuming that sampling errors Єij and Єcj are independent of random domain effects ηi and ηc. We assume that

$E(\eta i) = E(\eta c) = 0$; $V(\eta i) = \sigma i^2$; $V(\eta c) = \sigma c^2$; $E(\text{Є}ij) = E(\text{Є}cj) = 0$; $V(\text{Є}ij) = V(\text{Є}cj) = \sigma e^2$. This was also the model for domain estimation in evaluation of efficiency of ratio-synthetic estimator in sample surveys combining time-series and cross-sectional data (see Ghangurde, P. D. (2015)). We now derive estimator $\hat{\beta}$ and its variance $V(\hat{\beta})$ under the standard model (2.1) assuming equal unit error variances and $\eta i = \eta c = 0$. Thus

$$\hat{\beta} = [\ Xi'\ Vi^{-1}\ Xi + Xc'\ Vc^{-1}\ Xc\ ]\ [\ Xi'\ Vi^{-1}\ Yi + Xc'\ Vc^{-1}\ Yc\ ], \quad (2.2)$$

where

$$Yi = [\ \text{row}\ (Yij)\ ]';\ 1 \leq j \leq ni \qquad Yc = [\ \text{row}\ (Ycj)\ ]';\ (ni+1) \leq j \leq n$$

$$X_i = [\text{row } (x_{ij})]; \quad 1 \le j \le n_i \qquad X_c = [\text{row } (x_{cj})]; \quad (n_i+1) \le j \le n$$

$$V_i = \sigma_e^2 [I_{n_i}]; \quad V_c = \sigma_e^2 [I_{(n-n_i)}].$$

Also,

$$V(\hat{\beta}) = [X_i' V_i^{-1} X_i + X_c' V_c^{-1} X_c]^{-1}. \qquad (2.3)$$

Thus

$$\hat{\beta} = [\Sigma_{j=1}^{n_i} Y_{ij} x_{ij} + \Sigma_{j=(n_i+1)}^{n} Y_{cj} x_{cj}] / [\Sigma_{j=1}^{n_i} x_{ij} + \Sigma_{j=(n_i+1)}^{n} x_{cj}] \qquad (2.4)$$

$$V(\hat{\beta}) = \sigma_e^2 [\Sigma_{j=1}^{n_i} x_{ij} + \Sigma_{j=(n_i+1)}^{n} x_{cj}]^{-1}, \qquad (2.5)$$

assuming

$x_{ij} > 0; j=1,...,n_i; \; j \in U_i; \quad x_{cj} > 0; j = (n_i+1),....,n; \; j \in U_c.$

The ratio-synthetic estimator assuming one auxiliary variable and known domain mean $\bar{X}_i$. is:

$$\hat{\bar{\mu}}_i = \bar{X}_i \hat{\beta} \qquad (2.6)$$

and its variance under standard regression model (2.1) assuming $\eta_i = \eta_c = 0$ is:

$$\bar{X}_i^2 V_0(\hat{\beta}).$$

Under the mixed model (2.1) assuming N_i is large and n_i/N_i is small the BLUP estimator is given by

$$\hat{\bar{\mu}}_i = \bar{X}_i \hat{\beta} + \hat{\eta}_i, \qquad (2.7)$$

where $\hat{\beta}$ is estimator of β and $\hat{\eta}_i$ is estimator of η_i based on the mixed model (2.1).

We derive below m.s. error of $\hat{\bar{\mu}}_i$ under the mixed model and then derive efficiency of ratio-synthetic estimator under the model (2.1) assuming $\eta_i = \eta_c = 0$.

The estimator of small area domain effect ηi for the case of equal unit variances is

$$\hat{\eta i} = \xi i (\bar{Y}ia - \bar{X}ia \hat{\beta}). \qquad (2.8)$$

This is the case $kij = 1$ (page 135-36; Rao,J.N.K. (2003)). The derivation of $\hat{\beta}$ and $V(\hat{\beta})$ is given later. Thus in (2.8)

$$\bar{Y}ia = [\sum_{j=1}^{ni} Yij]/ni \text{ ; since } aij = 1 \text{ in the model with equal unit error variances,}$$

$$ai. = \sum_{j=1}^{ni} aij = ni. \quad \text{Similarly, } \bar{X}ia = [\sum_{j=1}^{ni} Xij]/ ni \text{ and } \bar{X}ca = [\sum_{j=(ni+1)}^{n} Xcj] / (n- ni) .$$

The optimal weights derived for the two domains in the mixed model (2.1) are:

$$\xi i = \sigma i^2 /[\sigma i^2 + \sigma e^2 / ni] = 1/ [1 + (\sigma e^2 / \sigma i^2) /ni] ; \qquad (2.9)$$

$$\xi c = 1/ [1 + (\sigma e^2 / \sigma c^2) / (n-ni)] . \qquad (2.10)$$

Substituting for ηi from (2.8) in (2.7)

$$\hat{\bar{\mu}i} = (\bar{X}i. - \xi i \bar{X}ia) \hat{\beta} + \xi i \bar{Y}ia . \qquad (2.11)$$

Conditioning on sample x-values $E[Yij - E (Yij)] = E[\eta i + \epsilon ij] = 0$.

Using the expression for $\hat{\beta}$ given later the two terms in (2.11) can be proved to have zero covariance.

Also, $V[Yij] = E[\eta i + \epsilon ij]^2 = [\sigma i^2 + \sigma e^2] = \sigma e^2 [1 + (\sigma i^2 / \sigma e^2)]$.

Using the last result we have

$$MSE(\hat{\bar{\mu}i}) = (\bar{X}i. - \xi i \bar{X}ia)^2 V_m(\hat{\beta}) + \xi i^2 \sigma e^2 [1 + (\sigma i^2 / \sigma e^2)] / ni. \qquad (2.12)$$

The term (7.1.13), page 176 ; Rao, J.N.K and Molina, I.(2015) is incorrect.

$V_m(\hat{\beta})$, variance of $\hat{\beta}$ under the mixed model (2.1), is given by

$$V_m(\hat{\beta}) = [Xi' Vi^{-1} Xi + Xc' Vc^{-1} Xc]^{-1} , \qquad (2.13)$$

where

$$V_i = \sigma_e^2 [\, I_{ni} + (\sigma_i^2/\sigma_e^2)\, 1_{ni} 1_{ni}' \,];$$

$$V_c = \sigma_e^2 [\, I_{(n-ni)} + (\sigma_c^2/\sigma_e^2)\, 1_{(n-ni)} 1_{(n-ni)}' \,],$$

where 1_{ni} and $1_{(n-ni)}$ are column vectors of ones. Using a result on matrix inversion :

$$[A + u v'\,]^{-1} = A^{-1} - [A^{-1} u v' A^{-1}\,] / [\,1 + v' A^{-1} u\,],$$

$$V_i^{-1} = \sigma_e^{-2} [\, I_{ni} - \xi_i [\, 1_{ni} 1_{ni}' \,] / n_i \,];$$

$$V_c^{-1} = \sigma_e^{-2} [\, I_{(n-ni)} - \xi_c [\, 1_{(n-ni)} 1_{(n-ni)}' \,] / (n-n_i) \,].$$

We derive $\hat{\beta}$ under the mixed model (2.1) and prove $E(\hat{\beta}) = \beta$, as done in Ghangurde,P.D.(2014).

$$\hat{\beta} = [\, X_i' V_i^{-1} X_i + X_c' V_c^{-1} X_c \,]^{-1} [\, X_i' V_i^{-1} Y_i + X_c' V_c^{-1} Y_c \,]. \quad (2.14)$$

Substituting for V_i^{-1} and V_c^{-1} in $\hat{\beta}$ we have

$$\hat{\beta} = [X_i' X_i - \xi_i X_i' 1_{ni} 1_{ni}' X_i / n_i + X_c' X_c - \xi_c X_c' 1_{(n-ni)} 1_{(n-ni)}' X_c / (n-n_i)]^{-1}$$

$$[\, X_i' Y_i - \xi_i X_i' 1_{ni} 1_{ni}' Y_i / n_i + X_c' Y_c - \xi_c X_c' 1_{(n-ni)} 1_{(n-ni)}' Y_c / (n-n_i) \,]$$

$$= [\, \sum_{j=1}^{n_i} \bar{X}_{ij} - \xi_i n_i \bar{X}_{ia} + \sum_{j=(n_i+1)}^{n} \bar{X}_{cj} - \xi_c (n-n_i) \bar{X}_{ca} \,] \text{ multiplied by}$$

$$[\, \sum_{j=1}^{n_i} \bar{X}_{ij} \bar{Y}_{ij} - \xi_i n_i \bar{X}_{ia} \bar{Y}_{ia} + \sum_{j=(n_i+1)}^{n} \bar{X}_{cj} \bar{Y}_{cj} - \xi_c (n-n_i) \bar{X}_{ca} \bar{Y}_{ca} \,]. \quad (2.15)$$

Conditioning on sample x-values it can be seen that $E(\hat{\beta}) = \beta$ and that covariance between two terms in (2.11) is zero. Also,

$$V(\hat{\beta}) = \sigma_e^2 [\, X_i' X_i - \xi_i X_i' 1_{ni} 1_{ni}' X_i / n_i + X_c' X_c - \xi_c X_c' 1_{(n-ni)} 1_{(n-ni)}' X_c / (n-n_i) \,]^{-1}$$

$$= \sigma_e^2 [\, \sum_{j=1}^{n_i} \bar{X}_{ij} - \xi_i n_i \bar{X}_{ia} + \sum_{j=(n_i+1)}^{n} \bar{X}_{cj} - \xi_c (n-n_i) \bar{X}_{ca} \,]^{-1}. \quad (2.16)$$

3. Efficiency of Ratio-synthetic Estimators Assuming Equal (and Unequal) Unit Error Variances

The efficiency defined as in Ghangurde, P. D (2014) assuming known $\bar{X}i.$, is:

$$MSE(\hat{\bar{\mu}}i)/\bar{X}i.\,V_0(\hat{\beta}) = [1 - \xi i\,(\bar{X}ia/\bar{X}i.)]^2 \frac{V_m(\hat{\beta})}{V_0(\hat{\beta})} + \frac{\xi i^2\,\sigma e^2\,[\,1 + (\sigma i/\sigma e)^2\,]}{\bar{X}i.^2\,V_0(\hat{\beta})\,ni}. \qquad (3.1)$$

Given a partition of population U into two non-empty domains Ui and Uc expectation of sample mean $\bar{X}ia$ is domain mean $\bar{X}i.$; thus $\bar{X}ia / \bar{X}i. = 1$ approximately. Substituting for $V_m(\hat{\beta})$ and $V_0(\hat{\beta})$ approximate efficiency is :

$$[1 - \xi i]^2 \frac{[\sum_{j=1}^{ni} Xij^2 + \sum_{j=(ni+1)}^{n} Xcj^2]}{[\sum_{j=1}^{ni} Xij^2 - \xi i\,ni\,\bar{X}ia^2 + \sum_{j=(ni+1)}^{n} Xcj^2 - \xi c\,(n-ni)\,\bar{X}ca^2]} \qquad (3.2)$$

$$+ \xi i^2 \frac{[\sum_{j=1}^{ni} Xij/\bar{X}i.^2 + \sum_{j=(ni+1)}^{n} Xcj/\bar{X}i.^2]}{ni}\,[\,1 + (\sigma i/\sigma e)^2\,]. \qquad (3.3)$$

We call $[1 - \xi i]^2$ and ξi^2 as weights and weighted components as terms. In case of surveys of incomes of households or businesses when interest is in fitting regression models to estimate expenditures on various categories from income (e.g.expenditures on food from household income or employee wages and salaries from business income) the auxiliary variable distribution is uni-modal and positively skew. In Economics and Business two or three parameter lognormal distributions have been studied as models for household and business income. The sample of n units in model (2.1) can be assumed to be drawn from

lognormal distribution with parameters μ and σ^2. This lognormal distribution has rth moment about the origin $E(x^r) = \exp(r\mu + [r^2\sigma^2]/2)$. For some μ and $\sigma^2 = 1$, $E(x) = e^{\mu + ½}$ and $E(x^2) = e^{2\mu + 2}$ (see pages 3 - 9; Crow, E.L. and Shimizu, K. (1988)). Also, $E(n_i \bar{X}_{ia}) = n_i e^{\mu + ½}$; $E([n-n_i]\bar{X}_{ca}) = [n-n_i] e^{\mu + ½}$; $E[\sum_{j=1}^{n_i} \chi_{ij}] = e^{2\mu + 2} n_i$; $E[\sum_{j=(n_i+1)} \chi_{cj}] = e^{2\mu + 2}(n - n_i)$, where χ_{ij} and χ_{cj} are x-values of sample units from U_i and U_c respectively.

Substituting $\bar{X}_{ia}^2$ for $\bar{X}_{i.}^2$ and $e^{2\mu + 1}$ for $\bar{X}_{ia}^2$ approximate efficiency by unconditional inference is:

$$[1 - \xi_i]^2 \frac{e^{2\mu}\,[e^2 n_i + e^2 (n - n_i)]}{e^{2\mu}\,[e^2 n_i - \xi_i n_i e^2] + e^{2\mu}\,[e^2 (n - n_i) - \xi_c (n - n_i) e^2]}$$

$$+ \frac{\xi_i^2\, e^{2\mu}\,[e^2 n_i + e^2 (n - n_i)]}{n_i\, e^{2\mu + 1}} [1 + (\sigma_i / \sigma_e)^2].$$

. $E(x^{-1}) = e^{\mu - ½}$; $E(x^{-2}) = e^{2\mu - 2}$; $E(\bar{X}_{ih}) = E(\bar{X}_{ch}) = e^{\mu - ½}$ and $E(\bar{X}_{ih2}) = e^{2\mu - 2}$. These will be used in the model with unequal unit error variances. After simplification approximate efficiency under standard regression model with equal unit error variances assuming lognormal distribution with unspecified μ and $\sigma^2 = 1$ is :

$$[1 - \xi_i]^2 \frac{[n / n_i]}{[1 - (\xi_i / e) + [(n / n_i) - 1][1 - (\xi_c / e)]} + \xi_i^2 e\, [n / n_i] [1 + (\sigma_i / \sigma_e)^2]. \qquad (3.4)$$

The optimal weights ξ_i and ξ_c in the first term of (3.4) are reduced by factor $(1 / e)$ and the second term is increased by factor e as compared to the same terms in approximate efficiency (3.5) below under standard regression model with unequal unit error variances. The efficiencies under

two models differ due to different error assumptions. The lognormal with $\sigma^2 = 1$ has unimodal longtailed skew distribution appropriate for household or business income. The efficiency (3.5) depends on parameter μ since ξiu and ξcu depend on $\bar{X}ih$ and $\bar{X}ch$ and second term depends on $\bar{X}ih / \bar{X}ih2$. The approximate efficiency under standard regression model with unequal unit error variances is obtained by substituting for $\bar{X}ih$, $\bar{X}ch$ and $\bar{X}ih2$, the last one being harmonic mean of squares of x-values of sample units from Ui, their expectations assuming lognormal with $\mu = 11.0$ and $\sigma^2 = 1$. Since for these values $e^{\mu} = 60{,}000$ it is realistic as distribution of household incomes and appropriate for computation of terms in (3.5). In Ghangurde, P. D.(2014) values were assumed for $\bar{X}ih$ and $\bar{X}ch$ to evaluate efficiencies. Now $\bar{X}ih / \bar{X}ih2 = E(\bar{X}ih) / E(\bar{X}ih2) = e^{-9.5} = 0.000075$, giving approximate efficiency:

$$[1 - \xi iu]^2 \frac{[n / ni]}{[1- \xi iu] + [(n/ni) - 1][1 - \xi cu]} + \xi iu^2 [n/ni] [1 +(\sigma i/ \sigma e)^2 (\bar{X}ih/\bar{X}ih2)], \quad (3.5)$$

where optimal weights ξiu and ξic are approximated assuming the lognormal; $\xi iu = 1 / [1 + (\sigma e/\sigma i)^2 (e^{10.5} / ni)]$ and $\xi cu = 1 / [1+ (\sigma e / \sigma c)^2 (e^{10.5} /(n – ni))]$ and $\sigma e^2 = 5\sigma i^2 = 5\sigma c^2$; $(\sigma i / \sigma e)^2 = 0.2$. Thus second term of (3.5) reduces to $\xi iu^2 [n/ni] [1.000015]$. Due to approximation of optimal weights ξiu, ξcu and $\bar{X}ih / \bar{X}ih2$ in (3.5) the differences in efficiencies based on two models can be attributed to differences in the assumptions about errors in these two models. The efficiencies based on (3.4) and (3.5) are presented in Tables 1 and 3.

We now consider efficiencies in sample surveys under unified model defined in Ghangurde. P. D.(2014). Assuming $\alpha ij = \alpha cj = 1$ for units in sample and zero outside sample term (3.2) is the same under two models. Also, in sample surveys optimal weights ξi and ξc are the same under two models.

Since Ui is a random domain from population U, $E[\bar{X}i.] = \bar{X} = [n / N]$ and second term (3.3) is $\xi i^2 [N / n]^2[n/ni]^2 [1+ (\sigma i / \sigma e)^2]$ by unconditional inference. We have assumed $\sigma e^2 = 5\ \sigma i^2$ giving approximate efficiency in sample surveys:

$$[\,1 - \xi i\,]^2 \frac{n/ni}{[1 - \xi i\,] + [1 - \xi c]\,[(n/ni) - 1\,]} + \xi i^2\,[\,N/n]\,[\,n/ni\,]\,[\,1.2^2\,]\,. \qquad (3.6)$$

The second term in the model with unequal unit error variances, assuming $\aleph ij = \aleph cj = 1$ in sample surveys is $\xi i^2\,[\,n/ni]\,[\,1.2^2\,]$; see Ghangurde, P. D.(2014).

In (3.6) second term is greater due to factor $[N/n]^2$ making ratio-synthetic estimator more efficient than BLUP in sample surveys under model with equal unit error variances (see Tables 2 and 4) .

In household surveys the use of ratio-synthetic estimator in survey practice assumes known domain totals of households and persons. In other sample surveys, where regression models are fitted for estimation, domain means of quantitative auxiliary variable (e.g. household income) are assumed to be known, which is the main issue in use of ratio-synthetic estimators in survey practice. In ratio-synthetic and BLUP estimators use of administrative data-based domain totals or means of auxiliary variable has been suggested in the literature on small area estimation. However, appropriate changes in processing of administrative data and other issues have not been addressed (see Ghangurde P.D.(2013); Rao J.N.K. and Molina, I. (2015)).

Since ratio-synthetic estimator is more efficient than BLUP under unified model for sample surveys assuming equal or unequal unit error variances, it should be used in survey practice. The use of estimates of household and population totals for domains based on counts of households in sample clusters and counts of persons in households, obtained in the first month in survey data collection, is outlined in brief in Section 4. This can be expanded based on survey practice.

4. Ratio-synthetic Estimation in Survey Practice

In household surveys with area sample designs clusters of households are sampling units within strata and households within clusters are the last stage sampling units. Weighted household and population totals within sample clusters in domains within strata can be used in ratio-synthetic estimation to obtain estimated domain totals Xi. for households and persons. There are about 200 households per cluster in the LFS (see Methodology of the Canadian LFS; Statistics Canada (2008)). Cluster weights are multiplied by inverse sampling ratios within individual clusters to obtain household weights. In

strata with extreme growth or decline of population clusters may be re-designed and households listed. For details on procedure of household listing see Guide to the Labour Force Survey; Statistics Canada (2012).

In the LFS listing of households in clusters in rotation groups to be introduced in the survey and updating household weights is part of household survey operations. In domains of interest within strata listing of households in clusters not in the survey sample can be done by survey operations staff as an extension of sample cluster listing done a few months in advance of the first survey month when sample clusters in a rotation group are introduced. In the case of LFS a much better option would be matching clusters with areal units in the last Census of Population to obtain household counts for clusters in domains and updating these at the time of introduction of rotation groups in the sample by using Address Register (AR). The advantage of obtaining household totals for domains as part of ongoing survey operations or using AR to update cluster lists is that the statistical agency would be responsible for these activities and can ensure data quality.

Methods can be developed for multipliers to obtain population counts for domains in rural areas, suburbs, urban core and fringe areas from counts of LFS sample household members collected in the first month that rotation groups are in the survey. These methods can be refined over time and multipliers can be verified. These multipliers can be applied to weighted household counts in sample clusters to obtain estimated population in various areas identified within domain Ui enclosed by strata .

This approach would be more objective than using administrative data for related auxiliary variables from other government agencies without studies on changes in processing data to obtain estimates and whether increased correlations of these x-variables with y-variable are enough to offset substantial reduction in number of matched administrative records available (e.g increased correlation of unemployed with unemployment insurance receipients vs substantial reduction in persons eligible for unemployed insurance benefits). These issues with the use of administrative data have not been addressed in Rao, J.N.K.and Molina, I. (2015).There are other issues in the area of data quality control in agencies providing data (Section 4; Practical issues; Ghangurde, P. D. (2013)).

Domain totals Xi. are usually not known in inter-census period. Ratio-synthetic estimation transforms the problem of updating synthetic weights in inter-census period to much simple one of estimation of domain totals for households and persons. Domains are parts of geographic strata. Ratio-synthetic estimation would be done within individual strata which enclose parts of domain of interest Ui. Estimation and variance estimation of other strata is not changed; also Xi. are not needed for these strata, since there is no domain estimation in these strata. Household and population totals for these strata without domains would not need updating until re-design.

Thus estimation of domain mean $\bar{\mu}i$, based on weighted data in a stratified sample design, would reduce to identifying strata enclosing parts of domain and obtaining estimates $\hat{\beta}$ for the strata multiplied by estimated domain totals Xi. based on the survey estimation methodology. The ratio-adjustment of sub-provincial estimates is a required step in household surveys based on LFS area sample design (see Ghangurde, P.D. and Gray, G.B. (1981)). Thus ratio-synthetic estimation would be more efficient and simple to implement as compared to BLUP as an integral part of estimation and variance estimation system for sub-provincial areas of household surveys.

Ratio-synthetc estimation can be extended to time-series data in household surveys with rotation sample designs; it is simpler than extention of BLUP . In a study on efficiency of ratio-synthetic estimators under standard regression model assuming equal unit error variances and 5 and 10 time-points or occasions AR(1) model was used (Ghangurde, P. D.(2015)). Time series models are useful in theoretical studies of ratio-synthetic and BLUP estimators. In practice, correlations in LFS survey estimates for characteristics over survey months can be based on matched households or persons in each rotation group; these correlations are higher for characteristics employed and in-labour force than for unemployed. Ratio-synthetic estimators for a domain total or mean and their standard errors can be obtained by using matched records over the last five months. The estimators can be combined for six rotation groups with proportion of matched households used as weights for estimates of each group providing efficiency gains in ratio-synthetic estimation in the LFS based on time-series of survey data of six months as compared to that for the current month.

The use of time-series models assuming equal or unequal unit error variances is not needed in the case of LFS, since correlations in survey data for a characteristic can be estimated from matched sample records of households and persons using survey variance estimation system. The same is true for other LFS area sample design-based surveys.

In a sample survey of construction industry in Nova Scotia, designed to obtain model-based estimates of total wages and salaries using gross business income as auxiliary variable ratio-synthetic estimator was more efficient than BLUP estimator (see example (7.3.2), pages 189 – 92; Rao,J.N.K.and Molina, I(2015)). The results in the study are based on simulation from a known population; thus domain total based on this population is known. However, in survey practice total income in a domain would have to be obtained from a recent census or survey.

In the case of some sample surveys an auxiliary variable highly correlated with estimation variable can be constructed as an index based on several correlated variables for which census data are available each year. National Agricultural Statistics Service(NASS) conducts annual survey of agricultural operations for estimation of average cash rental rates based on weighted survey data from a stratified sample survey. The construction of the index involved assumptions based on census data for these correlated auxiliary variables(see Berg, E.,Cecere, W.and Ghosh, M. (2014)). The study used BLUP estimation for area level mixed model. Ratio-synthetic estimation can be used as efficient and simpler method of unit level domain estimation in a sample survey in which model with unequal unit or equal unit error variances is appropriate. The theory and empirical results to support the use of ratio-synthetic estimator instead of BLUP estimator in domain estimation at unit level are available in this paper and in Ghangurde, P. D. (2014) .

5. Concluding Remarks

The important result proved in this paper shows that under standard regression model with equal unit error variances approximate efficiency of ratio-synthetic as compared to BLUP estimator in sample surveys is even greater than that under model with unequal unit errior variances. The results on efficiencies in the case of quantitative auxiliary variable under models with equal and unequal unit error variances assuming lognormal distribution for quantitative variable are similar. Under model with equal unit error variances ratio-synthetic is more efficient than BLUP as compared to the model with unequal unit error variances in surveys in which regression models are fitted to estimate components of auxiliary variable total in a domain for several categories.

In both surveys the problem in survey practice is that of obtaining domain totals or means for the quantitative variable. Under unified model in the case of household surveys methods to estimate total households and population in domains in inter-census period were reviewed. However, in the case of other

surveys total for domains from recent census for quantitative auxiliary variable or estimate from recent survey can be used. Although ratio-synthetic estimator is more efficient and simple to use in survey practice than BLUP, for estimation of domain totals for auxiliary variables in sample surveys and domain means of quantitative auxiliary variables in surveys in which regression models are fitted for domain estimation, methods based on survey practice have to be developed.

Even so, ratio-synthetic estimation is simple and efficient method of providing survey sponsors and data analysts estimates and their standard errors for a range

of values of domain means $\bar{X}i.$ and totals $Xi.$. In most situations assuring one true value for mean or total seems not possible. We would be able to provide in household surveys estimate of persons in a domain Ui based on first month responses from households in clusters rotating in the sample. This is expected to be more accurate than estimates based on administrative data from local area jurisdictions. In case of counts of households in domains in strata, the counts based on sample clusters would be estimates. If the counts are listed by survey personnel or obtained from AR these would have bias due to under-coverage, which is called slippage in the LFS.

In practice, most surveys have stratified sample designs with different weights for sample units between strata. These weights are adjusted for different non-response rates between strata. Since lognormal distribution was used for approximations in both models results are comparable. Extension of results to weighted data for ratio-synthetic and BLUP estimators under standard regression model is only of theoretical interest. The use of ratio-synthetic estimation instead of BLUP in sample surveys is important because it is more efficient than BLUP and simple to use in survey practice. The domains of interest are likely to be enclosed by a few strata within urban or rural areas of a province. Thus estimation and variance estimation for ratio-synthetic estimates would make no assumptions than those made in methodology of estimation and variance estimation in these surveys and would be an integral part of the methodology for sub-provincial estimation in these surveys.

The assumption of lognormal distribution for auxiliary variable in obtaining efficiencies of ratio-synthetic estimators under models with unequal and equal unit error variances made it possible to attribute differences in efficiencies to different assumptions about errors. Efficiencies in the case of model with unequal unit error variances can be obtained by assuming a value of μ giving median income e^{μ} close to that based on sample surveys as done in this study.

TABLE 1

Efficiencies based on sample size n ; auxiliary variable x_i with lognormal distribution; parameters unspecified μ and $\sigma^2 = 1$

Standard regression model with equal unit error variances : σ_e^2

Sample in U n = 100 ; unit error variance $\sigma_e^2 = 5$ (σ_i^2 or σ_c^2)

Sample in Ui : ni	Optimal ξ_i	ξ_i / e	Optimal ξ_c	ξ_c / e	First term	Second term	Total Efficiency
2	0.28571	0.10513	0.95146	0.35002	O.7790	13.31	14.09
4	0.44444	0.16350	0.95050	0. 34967	O.4692	16.11	16.58
6	0.54545	0.20066	0.94949	0. 34930	O.3132	16.11	16.42
8	0.61538	0.22638	0.94845	0.34891	O.2238	15.44	15.66
10	0.66667	0.24525	0. 94737	0.34859	O.1679	14.5O	14.67
12	0.70588	0.25966	0.94623	0.34810	O.1306	13.54	13.67
50	0.90909	0.33441	0.90909	0.33443	O.0124	5.39	5.4O
60	0.92308	0.33958	0.88889	0.32700	O.0089	4.64	4.65
80	0.94118	0.34624	0.80000	0.29430	O.0052	3.61	3.62

TABLE 2

Efficiencies based on sample size n in sample surveys; auxiliary variable = 1 for units in sample and = 0 for units not in sample;
Standard regression model with equal unit error variances : σ_e^2

Sample in U n = 100 ; unit error variance $\sigma_e^2 = 5$ (σ_i^2 or σ_c^2)

Inverse Sampling Ratio N/n = 10

Sample in Ui :ni	Optimal ξ_i	Optimal ξ_c	First term	Second term	Total Efficiency
2	O.28571	O.95146	8.25	490	498.25
4	O.44444	O.95050	4.43	58O	584.43
6	O.54545	O.94949	2.76	595	597.76
8	O.61538	O.94845	1.89	568	569.89
10	O.66667	O.94737	1.38	533	534.38
12	O.70588	O.94623	1.O5	498	499.05
50	0.9O9O9	O.90909	O.09	198	198.09
60	O.92308	O.88889	O.06	17O	170.06
80	O.94118	O.80000	O.04	133	133.04

TABLE 3

Efficiencies based on sample size n, with auxiliary variable with 2 lognormal distribution with parameters μ = 11.0 and σ = 1 ; Standard regression model with unequal unit error variances : $\mathfrak{X}ij\ \sigma e^2$; $\mathfrak{X}cj\ \sigma e^2$

Sample in U n = 100 ; unit error variance $\sigma e^2 = 5$ (σi^2 or σc^2)

Sample in Ui:ni	Optimal ξiu	Optimal ξcu	First term	Second term	Total Efficiency
2	O.0000110	O.0005394	1.OOO506	O.000000006	1.OOO506
4	O.0000220	O.0005285	1.OOO464	O.000000012	1.OOO464
6	O.0000330	O.0005176	1.OOO426	O.000000018	1.OOO426
8	O.0000440	O.0005066	1.OOO384	O.000000024	1.OOO384
10	O.0000551	O.0004955	1.OOO342	O.000000030	1.OOO342
12	O.0000661	O.0004845	1.OOO300	O.000000036	1.OOO300
50	O.0002754	O.0002753	O.99975O	O.000000150	O.999750
60	O.0003294	O.0002203	O.9997OO	O.000000180	O.999700
80	O.0004403	O.0001100	O.99952O	O.000000240	O.999520

TABLE 4

x-variable = 1 for units in sample; x-variable = 0 for units not in sample; Unequal unit error variances model reducing to equal in sample surveys.

Sample in U n = 100 ; unit error variance $\sigma e^2 = 5$ (σi^2 or σc^2)

$\bar{X}ih = \bar{X}ch = 1$ due to $Xij = Xcj = 1$ for sample units j

Sample in Ui : ni	Optimal Weight ξi	Optimal Weight ξc	First term	Second term	Total Efficiency
2	O.28571	O.95146	8.25	4.9O	13.15
4	O.44444	O.95050	4.43	5.8O	10.23
6	O.54545	O.94949	2.76	5.95	8.71
8	O.61538	O.94845	1.89	5.68	7.57
10	O.66667	O.94737	1.38	5.33	6.71
12	O.70588	O.94623	1.O5	4.98	6.03
50	0.9O9O9	O.90909	O.09	1.98	2.07
60	O.92308	O.88889	O.06	1.7O	1.76
80	O.94118	O.80000	O.04	1.33	1.37

References

[1] Berg, E., Cecere, W. and Ghosh, M. (2014), Small area estimation for county level farm-land cash rental rates, Journal of Survey Statistics and Methodology, Volume 2 / Number 1.

[2] Cochran, W.G. (1977), Sampling Techniques, Wiley.

[3] Crow, E. L. and Shimizu, K. (1988), Lognormal Distributions, Theory and Applications, Marcel Dekker, Inc.

[4] Ghangurde, P. D. and Gray, G.B. (1981), Estimation for small areas in household surveys, Communications in Statistics, Series A (10).

[5] Ghangurde, P. D. (2012), Small area estimation in household surveys when auxiliary variable totals are known, Presented at the Joint Statistical Meetings of the American Statistical Association, San Diego, USA.

[6] Ghangurde, P. D. (2013), Standard regression model-based small area domain estimation in household surveys, Presented at the Joint Statistical Meetings of the ASA, Montreal, Canada.

[7]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, USA.

[8] Ghangurde, P. D. (2015), Efficiency of standard regression model-based ratio-synthetic estimators in sample surveys combining time series and cross-sectional data, Presented at the Joint Statistical Meetings of the ASA, Seattle, USA.

[9] Hartley, H. O. (1959), Analytic studies of survey data, in a Volume in honor of Corrodo Gini, Instituto di Statistica, Rome, Italy.

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

[11] Rao, J.N.K. and Molina, I. (2015), Small Area Estimation, Wiley.

[12] Statistics Canada (2008), Methodology of the Canadian Labour Force survey, Catalogue No. 71-526-X.

[13] Statistics Canada (2012), Guide to the Labour Force Survey, Catalogue No. 71-543-G.