Small area estimation in household surveys when auxiliary variable totals are known

P. D. Ghangurde, Consultant
Ottawa, Ontario, Canada

1.Introduction

Assuming standard regression model for a small area domain and its complement with known auxiliary variable population totals BLUE of domain total of an estimation variable is derived as regression-synthetic estimator, with regression coefficients estimated from the model. In household surveys synthetic estimation is done within post-strata (e. g. age-gender groups) which are known to be more homogeneous than geographic strata. Thus addition of a random small area domain effect in the model, although of theoretical interest, is not of practical use in household surveys.

When there is one auxiliary variable, the regression-synthetic estimator obtained from the model reduces to synthetic estimator with ratio-adjustment, assuming that in each post-stratum the ratio of of estimation variable total to auxiliary variable total in the domain is equal to the ratio in the post-stratum. The assumption is made explicitly in synthetic estimation but is implicit in ratio-synthetic estimation. The synthetic estimator with ratio-adjustment reduces to ratio-synthetic estimator with auxiliary variable based on administrative data instead of population. This reduction involves a simple substitution for synthetic weights in terms of administrative data instead of population counts, which can become outdated in inter-census period.

Synthetic estimator with ratio-adjustment in age-gender post-strata was defined for Canadian Labour Force Survey (LFS) and evaluated using census data and super-population models by Ghangurde and Singh(1977). Gonzalez and Waksberg(1973) and Gonzalez and Hoza (1978) evaluated synthetic estimates without ratio-adjustment for the Current Population Survey (CPS) using mean square error criterion.

In section 2 the standard regression model for domain estimation is introduced and the BLUE of domain total of an estimation variable is derived. In the case of one auxiliary variable the estimator is extended to post-strata and reduced to synthetic estimator with ratio-adjustment and then to ratio-synthetic estimator with auxiliary variable based on administrative data. In section 3 practical issues in ratio-synthetic estimation such as effect of inaccurate data on domain totals, estimates and predicted values of estimation variables are reviewed. Sources of administrative data for estimation of synthetic weights in inter-census period are reviewed and the use of synthetic estimation with ratio-adjustment using synthetic weights based on administrative data as an alternative to ratio-synthetic estimation based on administrative data is suggested.

2.The model: domain, synthetic and ratio-synthetic estimators

Let population U consist of a small area domain Ui and its complement Uc = U - Ui. A typical domain Ui cuts across several strata and is comprised of domains within these strata. The model is for domain Ui and its complement Uc in a stratum called population. It gives theoretical results on optimality of regression coefficients and estimators of domain total. In practice the estimators have to be extended to strata of sample design of a survey.

Prepared for presentation at the meetings of Survey Research Methods Section of the American Statistical Association, San Diego, July 28 – August 2, 2012

Let U consist of N elements. A sample s of size n is drawn from U by simple random sampling. Let si represent ni (> 0) sample elements from Ni elements in Ui and sc represent (n – ni) sample elements from (N – Ni) elements from Uc . It is assumed that Ni is large and ni/Ni is small. Consider y_{ij}, y-value of jth element in Ui and y_{cj}, y-value of jth element in Uc. , assumed to be related to p auxiliary variable values, $\chi'_{ij} = (\chi_{ij1}, \ldots, \chi_{ijp})$ and $\chi'_{cj} = (\chi_{cj1}, \ldots, \chi_{cjp})$ respectively, χ_{ijk} and χ_{cjk} being x-value of kth auxiliary variable for jth element in Ui and Uc respectively. The model for domain estimation is:

$$y_{ij} = \chi'_{ij}\beta + \varepsilon_{ij}, \quad j \in U_i,$$
$$y_{cj} = \chi'_{cj}\beta + \varepsilon_{cj}, \quad j \in U_c, \qquad (2.1)$$

where $\beta = (\beta_1, \ldots, \beta_p)'$ is a column vector of regression coefficients, ε_{ij} and ε_{cj} are errors, $\varepsilon_{ij} = K_{ij}\bar{\varepsilon}_{ij}$ and $\varepsilon_{cj} = K_{cj}\bar{\varepsilon}_{cj}$, K_{ij} and K_{cj} are known constants, $\bar{\varepsilon}_{ij}$ and $\bar{\varepsilon}_{cj}$ are i.i.d. random variables, $E(\bar{\varepsilon}_{ij}) = E(\bar{\varepsilon}_{cj}) = 0$, $V(\bar{\varepsilon}_{ij}) = V(\bar{\varepsilon}_{cj}) = \sigma^2$. Thus $V(\varepsilon_{ij}) = K_{ij}^2\sigma^2$, $j \in U_i$ and $V(\varepsilon_{cj}) = K_{cj}^2\sigma^2$, $j \in U_c$.

The domain totals $X_{i.k} = \sum_{j=1}^{N_i} \chi_{ijk}$ and $X_{c.k} = \sum_{j=1}^{N-N_i} \chi_{cjk}$ for p auxiliary variables are not known, but can be estimated from x-values of sample elements in domains using design weights. However, auxiliary variable population totals $X_{..k} = X_{i.k} + X_{c.k}$; k = 1,2,. . . ,p are assumed to be known. Let $X'_{..} = (X_{..1}, \ldots, X_{..p})$ and $\bar{X}'_{..} = (\bar{X}_{..1}, \ldots, \bar{X}_{..p})$ be row vectors of population totals and means respectively. The domain of interest is Ui ; hence for domain estimation $y_{cj} = 0$, $j \in s_c$, i .e. elements in the sample from the domain Uc are assumed to have zero y-values.

Domain Ui mean $\bar{\mu}_i = \sum_{j=1}^{N_i} y_{ij} / N_i$. Under model (2.1) by summation of all rows we have

$$\bar{\mu}_i = \frac{1}{N_i} X'_{..}\beta \quad . \qquad (2.2)$$

We assume that sample values obey the assumed model (2.1), which in matrix notation is

$$\begin{bmatrix} Y_i \\ Y_c \end{bmatrix} = \begin{bmatrix} X_i \\ X_c \end{bmatrix} \beta + \begin{bmatrix} \epsilon_i \\ \epsilon_c \end{bmatrix}, \qquad (2.3)$$

where Y_i is nix1 and Y_c is (n-ni)x1 column vector of y-values, ϵ_i is nix1 and ϵ_c is (n-ni)x1 column vector of sampling errors, X_i is nixp and X_c is (n-ni)xp matrix of x-values. The BLUE of $\bar{\mu}_i$ is given by

$$\hat{\bar{\mu}}_i = \frac{1}{N_i} X'_{..}\hat{\beta}, \qquad (2.4)$$

where $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$, the BLUE of column vector β under model (2.3) is given by

$$\hat{\beta} = [X' V X]^{-1} [X' V Y], \quad (2.5)$$

where $X' = [X'i, X'c]$, $Y' = [Y' i, Y'c]$, $V = \begin{vmatrix} Vi & 0nix(n-ni) \\ 0(n-ni)xni & Vc \end{vmatrix}$,

$$Vi = V(\epsilon i) = \sigma^2 \, diag(Kij)^{-1}, \quad 1 \le j \le ni \quad \text{and} \quad Vc = V(\epsilon c) = \sigma^2 \, diag(Kcj)^{-1}, \quad ni+1 \le j \le n .$$

Since Yc is assumed to be a column vector of zero y-values in estimation for domain Ui ,

$$\hat{\beta} = [X'i \, Vi^{-1} \, Xi + X'c \, Vc^{-1} \, Xc]^{-1} [X'i \, Vi^{-1} \, Yi]. \quad (2.6)$$

In household surveys our interest is in the case of one auxiliary variable. By suppressing subscript k $X'i = (\chi i1, \ldots, \chi ini)$ and. $X'c = (\chi cni+1, \ldots \chi cn)$ are row vectors. of ni and (n-ni) elements.

Let $Kij = \chi ij^{1/2}$ then $Kij^{-1} = \chi ij^{-1/2}$; $j=1, \ldots, ni$; let $Kcj = \chi cj$, then $Kcj^{-1} = \chi cj^{-1}$; $j = ni+1, \ldots, n$.
This assumption is appropriate in the case of pps sampling used in household surveys. Thus,

$Vi^{-1} = \sigma^{-2} \, diag(\chi ij^{-1})$, $1 \le j \le ni$ and $Vc^{-1} = \sigma^{-2} \, diag(\chi cj^{-1})$, $ni+1 \le j \le n$. After substitution for Vi and Vc

$$X'i \, Vi^{-1} \, Xi = \sigma^{-2} \sum_{j=1}^{ni} \chi ij^{-1}, \quad X'c \, Vc^{-1} \, Xc = \sigma^{-2} \sum_{j=ni+1}^{n} \chi cj^{-1} \quad \text{and} \quad X'i \, Vi^{-1} \, Yi = \sigma^{-2} \sum_{j=1}^{ni} yij .$$

Hence,

$$\hat{\beta} = \sum_{j=1}^{ni} yij \Big/ \left[\sum_{j=1}^{ni} \chi ij^{-1} + \sum_{j=ni+1}^{n} \chi cj^{-1} \right] \quad (2.7)$$

$\hat{\beta}$ is the BLUE of β under model (2.1); but $\hat{\beta}$ is not design-consistent. Multiplying sample y-values in Ui and sample x-values in Ui and Uc by design-weights

$$\hat{\beta} = \hat{Y}i. / [\hat{X}i. + \hat{X}c.] = \hat{Y}i. / \hat{X}.. , \quad (2.8)$$

where $\hat{Y}i.$ is design-weighted estimator of y-total for domain Ui and $\hat{X}..$ is design-weighted estimator of x-total for population U. By substitution in (2.4) the estimator of domain y-total, μi, noting that $\mu i = Ni \, \bar{\mu}i$ is

$$\hat{\mu}i = \frac{\hat{Y}i.}{\hat{X}..} X.. . \quad (2.9)$$

By an extension to post-strata, assuming that sample elements in each post-stratum g, $nig > 0$, the post-stratified estimator of domain Ui y-total is obtained as

$$\hat{\mu}i = \sum_g \frac{\hat{Yig}}{\hat{X.g}} X.g , \qquad (2.10)$$

where $X.g$ is known x-total in the post-stratum g, $\hat{X.g} / X.g$ is ratio-adjustment factor in g and $\hat{Yig}$ is design-weighted estimator of y-total for domain Ui in g. The condition $nig > 0$ assures that $\hat{\mu}i$ in (2.10) is p-unbiased.

In household surveys sample design weights for elements within strata differ. The standard model can be extended to weighted data as aggregated model by following a method used for type-B model (pages 148-49, Rao (2003)) ; the estimators (2.9) and (2.10) with weighted y-values and x-values can be proved to be the BLUE of μi under the aggregated $\hat{}$ model for weighted data. We have not done the extension of the model to weighted data to obtain β as BLUE for weighted data for household surveys. However, starting with estimator of domain total (2.10) based on design weights, the results below on relationship between domain and synthetic estimator with ratio-adjustment and reduction of the latter to ratio-synthetic estimator based on administrative data are true for design-weighted estimators in household surveys.

Due to area sample design of household surveys $\hat{Yig}$ has high variance; when there are no sample households in the domain $\hat{Yig}$ is equal to zero. Hence, synthetic estimates which are biased but have low variance were evaluated and used in survey practice (Gonzalez and Hoza (1978) ; Gonzalez and Waksberg (1973)). In synthetic estimation it is assumed that $Yig = Y.g\, Wig$, where $Wig = X ig / X.g$, proportion of population of post-stratum g in domain Ui ; Wig can be called as synthetic weight. The assumption is the same as $Yig / Xig = Y.g / X.g$ i.e. in each post-stratum g the ratio in the domain Ui is equal to the ratio in the post-stratum. Substituting $\hat{Yig} = \hat{Y.g}\, Wig$ in (2.10) we have synthetic estimator with ratio-adjustment as

$$\hat{Yis} = \sum_g \frac{\hat{Y.g}}{\hat{X.g}} Wig\, X.g . \qquad (2.11)$$

In practice a domain can cut across several strata. The assumption that $Yig / Xig = Y.g / X.g$ is extended to these strata by defining these two as combined ratios of y-total to x-total for persons of group g in households in these domains and strata respectively. Also, X.g can be defined as x-total for persons of group g in households in these strata and $\hat{Y.g} / \hat{X.g}$ can be defined as estimated ratio based on persons of group g in sample households in these strata.

In the post-census period, synthetic weights Wig can become inaccurate in strata of the design with uneven population growth. In these areas bias of $\hat{Yis}$ has an extra component due to uneven population growth, in addition to p-bias, which occurs if the assumption about ratios is not true. Ghangurde and Singh (1977) have derived the expression for the component of bias due to uneven population growth, developed methodology of evaluation of the component of bias and obtained empirical results.

It should be noted that the ratio-adjustment based on population projections has been an important step in obtaining estimates for household surveys. In Canada, in inter-census period population projections are available for provinces by age-gender groups. The ratio-adjustment factor $X.g/\hat{X}.g$ at province level is used in the LFS estimates for provinces and also for sub-provincial areas composed of group of strata. The ratio-adjustment reduces variance of design-weighted estimates by taking advantage of correlation between a characteristic (e.g. "employed", "in labour force" and "unemployed") and population (Ghangurde and Gray (1981)). The ratio-adjustment also corrects these estimates for under-coverage bias , called "slippage", due to dwellings missed in the lists created by survey interviewers for sub-sampling of dwellings in the areas of population growth.

If auxiliary variable in the estimator $\hat{Y}is$ is based on administrative data instead of population, $Wig = Xig / X.g$ will be known in the inter-census period. For example, for estimation variable "unemployed" in the LFS auxiliary variable can be "unemployed persons receiving employment insurance benefits" from a government agency. Since both Xig and X.g, counts of recipients of benefits of certain age-gender group in the domain Ui and the population (i.e. group of strata which comprise the domain) respectively, can be obtained from the agency providing the data Wig will be known. In practice there are several issues due to which data obtained for Xig and estimates computed could be inaccurate (see Section 3). By substituting $Wig\ X.g = Xig$ in the synthetic estimator $\hat{Y}is$ it reduces to ratio-synthetic estimator

$$\hat{Y}is/R = \sum_g \frac{\hat{Y}.g}{\hat{X}.g} Xig . \qquad (2.12)$$

(4.2.5, page 47, Rao (2003)). Apart from being a reliable source of administrative data which is related to the estimation variable the auxiliary variable should have good correlation with the estimation variable to maintain standard error of these predicted values, based on ratio estimation.

The above derivation shows that the basic assumption in synthetic estimation with ratio-adjustment and ratio-synthetic estimation is the same. Also, the ratio-synthetic estimator has a form which is convenient for use of administrative data in small area domain means. These data are available for small area domains in inter-census period, unlike population counts by age-gender groups.

The ratio-synthetic estimator (2.12) can be derived from the model and made design-weighted. Let $\bar{X}'i. = (\bar{X}i.1, . . , \bar{X}i.p)$ be row vector of known auxiliary variable domain means. Following the same derivation as before it can be seen that the estimator of domain Ui mean $\bar{\mu}i = \bar{X}'i.\beta$ is given by

$$\hat{\bar{\mu}}i = \bar{X}'i.\ \hat{\beta}, \qquad (2.13)$$

where

$$\hat{\beta} = [\, X'i\ Vi\ Xi + X'c\ Vc\ Xc\,]^{-1}\ [\, X'i\ Vi\ Yi + X'c\ Vc\ Yc\,]^{-1}. \qquad (2.14)$$

The difference between expressions for $\hat{\beta}$ in (2.6) and (2.14) is due to the assumption $Ycj = 0$, $j \varepsilon sc$, made for domain estimation in (2.6). For $p = 1$ domain estimator was reduced to synthetic estimator with ratio-adjustment and then to ratio-synthetic estimator, assuming that in each post-stratum g $Yig/Xig = Y.g/X.g$. When Ycj , $j \varepsilon sc$, are not assumed to be zero, domain estimator reduces to ratio-synthetic estimator when $p = 1$. Assuming the same values of Kij; $j = 1,. . . . ni$ and Kcj, $j = ni+1, ... ,n$, with population as the only auxiliary variable for the case $p = 1$ and following the same derivation as in (2.8) for domain estimation we have

$$\hat{\beta} = [\sum_{j=1}^{ni} Yij + \sum_{j=ni+1}^{n} Ycj\ ] \ / \ [\sum_{j=1}^{ni} \chi ij + \sum_{j=ni+1}^{n} \chi cj\ ], \qquad (2.15)$$

the BLUE of β under the model (2.1) ; but $\hat{\beta}$ is not design-consistent . Multiplying sample y-values and x-values by design-weights

$$\hat{\beta} = \frac{\hat{Yi}.+ \hat{Yc}.}{\hat{Xi}.+ \hat{Xc}.} = \frac{\hat{Y}..}{\hat{X}..}, \qquad (2.16)$$

which is design-consistent $\hat{}$ but not BLUE. We have not done extension of the model for weighted data to obtain $\hat{\beta}$ as BLUE for weighted data.

Substituting for $\hat{\beta}$ in (2.13) for the case $p = 1$ we get $\hat{\mu} i = Ni\ \hat{\bar{\mu}}i = [\ Y../\ X..]\ Xi.$ as design-weighted estimator of μi from the model. Extending to post-strata we obtain design-weighted ratio-synthetic estimator directly from the model as

$$\hat{Yis/R} = \sum_{g} \frac{\hat{Y.g}}{\hat{X.g}}\ Xig\ . \qquad (2.17)$$

The expression for $\hat{\beta}$ in (2.14) is simpler than that from the type-B model with small area domain effect assumed equal to zero (pages 134-37, Rao (2003)) due to the assumption of two domains and samples drawn independently from strata.

The ratio-synthetic estimator (2.17) obtained from the standard regression model assuming known auxiliary variable domain means has been referred to as an estimator borrowing strength from population i.e. strata containing the domain of interest. However, the indirect derivation has shown that the basic assumption is the same as in synthetic estimation: in each post-stratum the ratio of estimation variable total to auxiliary variable total $\hat{}$ in the domain is assumed to be equal to that in strata containing the domain. An advantage of $\hat{Yis/R}$, ratio-synthetic estimator based on administrative data, over $\hat{Yis}$, synthetic estimator with ratio-adjustment, is possibly higher correlation between estimation and auxiliary variable in the ratio-synthetic estimator. . However, it needs accurate administrative data at post-stratum level.

It is interesting to note that in the derivation of BLUP estimator based on the type-B model, its extensions and estimation of model variances in Rao (2003) there is no discussion of application of the model within post-strata of household surveys. If the estimator is used within post-strata the second component due to random area effect in the BLUP estimator is very likely to have negligible contribution due to homogeneity within post-strata. When used without post-stratification in household surveys the random area effect is difficult to interpret, if there are no post-strata effects in the model.

We have presented a model for domain estimation, which shows that the estimator of y-total used in household surveys is BLUE under the model, assuming an extension of the model to design-weighted data. It reduces to synthetic estimator with ratio-adjustment, assuming that in each post-stratum the ratio of y-total to x-total in the domain is equal to the ratio in the strata containing the domain, and to ratio-synthetic estimator with auxiliary variable based on administrative data. The model easily extends to domains cutting across design strata and to post-strata used for ratio-estimation. In this respect also, it is more appropriate as a model for domain estimation than type-B model. We plan to extend it as a mixed model for unit level data using the expression for β in (2.14) modified for random area effect.

3. Practical issues

. The sample households and persons of each group g from strata needed for computation of $\hat{X}.g$ and all persons of age-gender group g from domains within strata needed for computation of Xig, for a small area estimation project have to be identified on administrative data files and their information and data for auxiliary variable copied to survey files for weighting. The administrative data files may not be organized by dwelling addresses of households or persons as record identifiers. When a domain cuts across several strata sample households in the domain can have different design weights; obtaining data for persons in households associated with correct dwelling addresses from the agency is important for accuracy of totals for domains and estimated ratios by age-gender groups. Criteria for inclusion of persons with a characteristic (e.g. poor children eligible for financial assistance, unemployed persons eligible for unemployment insurance benefits after certain period, etc.) from sample households even if well-defined and regulated by agencies providing the data can be subject to administrative delays, changes in regulations, etc. A statistical agency receiving data has to organize data quality control programs in agencies providing the data, conduct timely data quality checks to ensure that data for persons of correct age-gender groups are provided and used for computing $\hat{X}.g$ and data for persons of correct age-gender groups in the domains are provided and used in computing of Xig. Errors of omissions and incorrect classifications have to be identified, data on errors of various types have to be collected and reported to the agency.

Persons in sample households who have a characteristic e. g. “unemployed” but are reported by the agency as not eligible for unemployment insurance benefits and not included in their count can not be excluded in computing $\hat{Y}.g.$, because their exclusion would introduce bias in predicted values. The exclusion would essentially change the survey definition of “unemployed”. However, their inclusion will reduce correlation between estimation variable and auxiliary variable based on administrative

data. Any errors in computation of Xig will result in bias in the predicted value;also any decrease in correlation will increase the standard error of predicted value. Increasing the number of auxiliary variables based on several administrative data sources will add to bias of predicted value and increase its standard error due to decrease in multiple correlation of estimation variable with auxiliary variables. In practice it seems better to use one data source with auxiliary variable with good correlation with estimation variable and decrease errors over time with experience with data quality control program and thus decrease bias and standard error of predicted values.

The basic assumption made in ratio-synthetic estimation is the same as that made in synthetic estimation with ratio-adjustment. Synthetic estimation requires accurate synthetic weights in inter-census period. Estimates of Wig can be developed by using data on population growth from municipalities and provincial governments. In Canada updated lists of dwelling addresses are maintained by Canada Post. These are the source of data for a database, called Address Register, maintained by Statistics Canada. The Address Register has good potential as a source of data for estimation of synthetic weights for the LFS sample design in inter-census period.

References

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

[2] Ghangurde, P. D. and Singh, M.P. (1977), Synthetic estimation in periodic household surveys, Survey Methodology, 3, pp. 152 -181.

[3] Gonzalez, M.E. and Hoza, C.(1978), Small area estimation with application to unemployment and housing estimates, Journal of the American Statistical Association, Vol. 73, No.361, pp. 7-15.

[4] Gonzalez, M.E. and Waksberg, J. (1973), Estimation of error of synthetic estimates, Paper presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria.

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