ESTIMATION FOR SMALL AREAS IN HOUSEHOLD SURVEYS

P.D. Ghangurde and G.B. Gray, Statistics Canada

Abstract

In many socio-economic surveys the objective is estimation of total or proportion of persons with a particular attribute. Multi-stage area samples are drawn from geographic strata and population within areal units is used as an auxiliary variable in ratio estimation. The auxiliary variable totals are available as population projections based on the last census, for large administrative areas. However, for small areas population changes are significantly affected by non-demographic factors and hence projections with high enough reliability are not available for small areas. In such situations, the efficiency of design-based estimates for small areas can be improved by a ratio adjustment based on auxiliary variable total for a large area. An inequality on efficiency of ratio adjusted estimate as compared to the design-based estimate is established and bias and efficiency of the ratio adjusted estimate is investigated.

1. INTRODUCTION

In continuous household surveys conducted with the main objective of estimation of total number of persons with a particular attribute (e.g. labour force status, health status) population count within areal units is used as an auxiliary variable in ratio estimation. The auxiliary variable totals are available as population projections, based on the last census, for large administrative areas like provinces, regions, etc. However, for small areas of the size of a group of counties, population changes are significantly affected by non-demographic factors like migrations and hence projections with high enough reliability are not available for such areas (Ericksen (1974)). The problem and the proposed estimate can be explained by the following example.

The Canadian Labour Force Survey is a monthly survey in which dwelling is the final stage sampling unit. Each province is divided into a number of economic regions which usually consist of a group of contiguous counties with similar economic structure. One or more strata are formed in these economic regions and multi-stage area samples, with probabilities proportional to population as of the last census, are drawn independently from these strata in two stages in urban strata and in three or four stages in rural strata. Since important labour force characteristics like the count of 'employed', 'unemployed' are known to be correlated to population count, it is used as a size measure in sample selection and as an auxiliary variable in ratio estimation. Further details on the sample design and estimation are given in Platek and Singh (1976); Fellegi, Gray and Platek (1967).

Let a province be divided into L strata and let multi-stage area samples be drawn from these strata. Let 'a' and 'b' be two sets of strata (b⊂a); the set a is the province and set b is the subprovincial area for which estimate of $Y_{b.}$, the characteristic total in set b, is required. The projected population total for set a, $X_{a.}$, is available but not $X_{b.}$. However, an unbiased estimate $\hat{Y}_{b.}$, based on the sample design within strata can be improved by ratio adjustment by using projected population for 'a' to obtain a more efficient estimate of $Y_{b.}$. The ratio estimate is given by

$$\hat{\hat{Y}}_{b.} = \frac{\hat{Y}_{b.}}{\hat{X}_{a.}} X_{a.} \qquad (1.1)$$

where $\hat{X}_{a.}$ is an unbiased estimate of $X_{a.}$, based on the sample design. The adjustment of the unbiased estimate $\hat{Y}_{b.}$ for the small area by multiplication by the factor $X_{a.}/\hat{X}_{a.}$ is usually called a ratio adjustment and is used in a number of periodic household surveys at a level at which estimates are required. The objective of this paper is to investigate bias and efficiency of the estimate obtained by the ratio adjustment. It may be noted that small area b is composed of complete strata.

In section 2, we give formulae for bias and variance of the estimate $\hat{\hat{Y}}_{b.}$ and obtain a result on the efficiency as compared to $\hat{Y}_{b.}$. In section 3, we give some empirical results on efficiency based on the Canadian Labour Force Survey.

2. BIAS AND VARIANCE

The use of truncated Taylor series approximation for obtaining expressions for bias and variance of ratio estimates is well known. The method, also known as linearization method, was used by Keyfitz (1957) to obtain variances for specific designs and was extended by Woodruff (1971) for application to complex designs. The assumption made in the following derivations is that the sample size within a set of strata 'a', is large enough for the approximation to be valid. Let

$$\hat{R}_{(b,a)} = \frac{\hat{Y}_{b.}}{\hat{X}_{a.}} \qquad (2.1)$$

be the ratio estimate of $R_{(b,a)} = Y_{b.}/X_{a.}$. The two subscripts of R indicate the set of strata on which the estimate of characteristic and population totals are based. By second order Taylor series approximation to $1/\hat{X}_{a.}$, the relative bias is given by

$$\frac{E[\hat{R}_{(b,a)} - R_{(b,a)}]}{R_{(b,a)}} = \left[-\frac{Cov(\hat{Y}_{b.}, \hat{X}_{a.})}{Y_{b.} X_{a.}} + \frac{V(\hat{X}_{a.})}{X_{a.}^2}\right]$$

$$= \frac{V(\hat{X}_{a.})}{X_{a.}^2}\left[-\frac{Cov(\hat{Y}_{b.}, \hat{X}_{b.})}{V(\hat{X}_{a.})}\frac{X_{a.}}{Y_{b.}} + 1\right]. \qquad (2.2)$$

The interpretation of (2.2) seems difficult in the case of multi-stage unequal probability sampling. For single stage designs with simple random sampling of clusters within strata (2.2) can have simple interpretation. The relative bias is given by

$$\frac{E(\hat{R}_{(b,a)} - R_{(b,a)})}{R_{(b,a)}} \doteq \frac{V(\hat{X}_{a.})}{X_{a.}^2}\left[1 - \frac{\beta_{(y,x)}}{R_{(b,b)}} \cdot \frac{V(\hat{X}_{b.})}{V(\hat{X}_{a.})} \cdot \frac{X_{a.}}{X_{b.}}\right], \qquad (2.3)$$

where $\beta_{y,x} = \mathrm{Cov}(\hat{Y}_{b.}, \hat{X}_{b.})/V(\hat{X}_{b.})$ is regression coefficient of characteristic total on population within clusters in set b. It is assumed that the finite population of clusters within each stratum is drawn as a simple random sample from an infinite population and that the line or regression passes through the origin (which is a realistic assumption for the characteristics considered) and is the same for all strata in b. Under these assumptions $\beta_{(y,x)} = R_{(b,b)}$. If further the coefficient of variation for set a and b is approximately equal, i.e.,

$$[V(\hat{X}_{b.})/V(\hat{X}_{a.})](X_{a.}^2/X_{b.}^2) \doteq 1, \text{ we have}$$

$$\frac{E[\hat{R}_{(b,a)} - R_{(b,a)}]}{R_{(b,a)}} \doteq \frac{V(\hat{X}_{a.})}{X_{a}^2} \cdot \frac{X_{c.}}{X_{a.}}, \qquad (2.4)$$

where $c = a-b$. From (2.3), it can be seen that the relative bias of $\hat{\hat{Y}}_{b.}$ is of the order $1/n$, where n is the total sample size in set a. For derivation of variance of $\hat{\hat{Y}}_{b.}$ we have

$$\hat{\hat{Y}}_{b.} - Y_{b.} = [\hat{R}_{(b,a)} - R_{(b,a)}]\, X_{a.} \doteq [\hat{Y}_{b.} - R_{(b,a)}\, \hat{X}_{a.}], \qquad (2.5)$$

by first order Taylor series approximation. Further

$$\hat{\hat{Y}}_{b.} - Y_{b.} \doteq [\hat{Y}_{b.} - R_{(b,a)}\, \hat{X}_{b.}] - R_{(b,a)}\, \hat{X}_{c.}.$$

Since sampling is done independently within each stratum,

$$V(\hat{\hat{Y}}_{b.}) = V[\hat{Y}_{b.} - R_{(b,a)}\, \hat{X}_{b.}] + V[R_{(b,a)}\, \hat{X}_{c.}]. \qquad (2.6)$$

The second term in (2.6) appears if c is non-null, i.e. bca. It can be seen that $\hat{\hat{Y}}_{b.}$ is more efficient than $\hat{Y}_{b.}$ if

$$\mathrm{Cov}(\hat{Y}_{b.}, \hat{X}_{b.}) > \frac{1}{2} R_{(b,a)} [V(\hat{X}_{b.}) + V(\hat{X}_{c.})] \quad \text{i.e.} \qquad (2.7)$$

$$\delta(\hat{Y}_{b.}, \hat{X}_{b.}) > \frac{1}{2} R_{(b,a)} \sqrt{\frac{V(\hat{X}_{b.})}{V(\hat{Y}_{b.})}} + \frac{1}{2} R_{(b,a)} \frac{V(\hat{X}_{c.})}{\sqrt{V(\hat{Y}_{b.})\, Y(\hat{X}_{b.})}} \qquad (2.8)$$

where δ is correlation coefficient defined as $\delta(\hat{Y}_{b.}, \hat{X}_{b.}) = \mathrm{Cov}(\hat{Y}_{b.}, \hat{X}_{b.})/\sqrt{V(\hat{Y}_{b.})V(\hat{X}_{b.})}$.

The inequality (2.8) is satisfied if c is small compared to b and correlation δ is large enough. The variance estimate of $\hat{\hat{Y}}_{b.}$ can be obtained from (2.6) by using $\hat{R}_{(b,a)}$ as an approximation to $R_{(b,a)}$, whenever $R_{(b,a)}$ appears in the estimate.

In the following application to the Canadian Labour Force Survey, the homogeneity of ratios within age-sex groups is exploited in the estimation by post-stratification of the sample and a separate ratio estimate is defined as

$$\tilde{Y}_{b.} = \sum_{j=1}^{K} \frac{\hat{Y}_{bj}}{\hat{X}_{aj}} X_{aj} = \sum_{j=1}^{K} \hat{R}_{(b,a)j}\, X_{aj}, \qquad (2.9)$$

where X_{aj} is projected population in set a and age-sex group j, $\hat{X}_{aj}$ the corresponding unbiased estimate and $\hat{Y}_{bj}$, the unbiased estimate and Y_{bj}, characteristic total in set b and jth age-sex group, $j = 1, 2, \ldots, K$. The derivation of bias and variance of $\tilde{Y}_{b.}$ can be done on the same lines as that for $\hat{\hat{Y}}_{b.}$; however, interpretation of the expressions for bias becomes difficult. The extension of (2.7) to $\tilde{Y}_{b.}$ is given by

$$\sum_{j=1}^{K} R_{(b,a)j}\, \delta(\hat{Y}_{bj}, \hat{X}_{bj}) \sqrt{V(\hat{X}_{bj}) \cdot V(\hat{Y}_{bj})} > \frac{1}{2} V\Big(\sum_{j=1}^{K} R_{(b,a)j}\, \hat{X}_{bj}\Big) + \frac{1}{2} V\Big(\sum_{j=1}^{K} R_{(b,a)j}\, \hat{X}_{cj}\Big), \qquad (2.10)$$

where symbols have obvious meanings. The inequality (2.10) cannot be simplified to an inequality with simple interpretation as in (2.8).

3. APPLICATION

Table I shows estimated $CV(\hat{Y}_{a.})$ (coefficient of variation of $\hat{Y}_{a.}$), $CV(\hat{\hat{Y}}_{a.})$, $CV(\tilde{Y}_{a.})$, $\delta(\hat{Y}_{a.}, \hat{X}_{a.})$ and $\frac{1}{2} \cdot CV(\hat{X}_{a.})/CV(\hat{Y}_{a.})$ for five important labour force characteristics. Table II shows estimated $CV(\tilde{Y}_{b.})$ and efficiency defined as the ratio of $CV(\hat{Y}_{b.})$ to $CV(\tilde{Y}_{b.})$ for five characteristics. The data of September 1975 survey in the province of Ontario, which is divided into ten economic regions are used in these calculations. The set b is successively increased from one economic region to the whole province (i.e. b=a).

TABLE I

Estimates of CVs, Correlations and Ratios
for Five Characteristics

Estimate	Employed	Unemployed	Employed Agric.	Employed Non-Agric.	In Labour Force
$CV(\hat{Y}_{a.})$	2.20	5.05	7.47	2.27	2.17
$CV(\hat{\hat{Y}}_{a.})$	0.99	5.40	9.16	1.04	0.91
$CV(\tilde{Y}_{a.})$	0.61	4.75	7.74	0.70	0.58
$\delta(\hat{Y}_{a.},\hat{X}_{a.})$	0.895	0.493	0.022	0.895	0.918
$\frac{1}{2}\frac{CV(\hat{X}_{a.})}{CV(\hat{Y}_{a.})}$	0.421	0.154	0.104	0.406	0.417

It can be seen that $\tilde{Y}_{b.}$ is more efficient than $\hat{Y}_{b.}$ for 'employed, 'employed non-agriculture' and 'in labour force' for all sets bca. The correlation at level a for these characteristics is high and is expected to be about the same for bca. The characteristic 'unemployed' has low correlation and this can explain the loss in efficiency for set b composed of one economic region. For the characteristic 'employed agriculture' with correlation of .022 at level a, there is loss in efficiency due to ratio estimation even for b=a, since the inequality (2.8) is not satisfied. The efficiency of $\tilde{Y}_{b.}$ over $\hat{Y}_{b.}$ generally decreases as set b is decreased. The combined ratio estimate $\hat{\hat{Y}}_{a.}$, has larger coefficient of variation than separate ratio estimate $\tilde{Y}_{a.}$ for all characteristics, due to homogeneity of ratios within post-strata.

4. CONCLUDING REMARKS

The ratio estimates ($\hat{\hat{Y}}_{b.}$ and $\tilde{Y}_{b.}$) for small areas composed of a set of strata are appropriate in situations in which population projections are not available for such areas and correlations are large enough. For the characteristics which are defined as total number or proportion of persons with a particular attribute, the relative bias of the ratio estimates is small. In household surveys, these ratio estimates have an additional advantage of reducing non-sampling bias due to coverage errors in the area frame. These coverage errors usually occur due to missed dwellings and persons. Since certain age-sex groups are more prone to under-coverage, the ratio adjustment by age-sex groups in $\tilde{Y}_{b.}$ may be more efficient than that at gross population level in $\hat{\hat{Y}}_{b.}$ for reducing the non-sampling bias (see Cochran (1975), and Fellegi (1973)).

The small areas considered in this paper are composed of complete strata and are not areal domains within strata. Under certain assumptions about the structure of the population model-based estimates have been proposed and evaluated as an alternative to design-based estimates for areal domains within strata (see e.g. Gonzales and Hoza (1978); Jones and Coopersmith (1976); Ghangurde and Singh (1978)). However, the problem considered in this paper is that of use of ratio-adjustment to improve efficiency of design-based estimates for small areas which are group of strata.

BIBLIOGRAPHY

[1] Cochran, W.G. (1963). Sampling Techniques (2nd edition). New York: John Wiley and Sons, Inc.

[2] Cochran, W.G. (1975). Two Recent Areas of Sample Survey Research. North-Holland Publishing Co.

[3] Ericksen, E.P. (1974). A regression model for estimating population changes in local areas. J. Amer. Statist. Assoc. 69, 867-75.

[4] Fellegi, I.P., Gray, G.B. and Platek, R. (1967). The new design of the Canadian Labour Force Survey. J. Amer. Statist. Assoc. 62, 421-53.

[5] Fellegi, I.P. (1973). The evaluation of the accuracy of survey results: some Canadian experiences. Int. Statist. Rev. 41, (1) 1-14.

[6] Ghangurde, P.D. and Singh, M.P. (1978). Evaluation of efficiency of synthetic estimates. Proceedings Survey Research Methods Section. Amer. Statist. Assoc.

[7] Gonzales, M.E. and Hoza, C. (1978). Small area estimation with application to unemployment and housing estimates. J. Amer. Statist. Assoc. 73, 7-15.

[8] Jones, D.H. and Coopersmith, L.A. (1976). A ratio estimator of the total of a subpopulation. Commun. Statist. Theor. Meth. A5 (3), 251-60.

[9] Keyfitz, N. (1957). Estimates of sampling variance when two units are selected from each stratum. J. Amer. Statist. Assoc. 52, 503-10.

[10] Platek, R. and Singh, M.P. (1976). Methodology of the Canadian Labour Force Survey. Technical Report. Statistics Canada.

[11] Woodruff, R.S. (1971). Simple method for approximating variance of a complicated estimate. J. Amer. Statist. Assoc. 66, 411-14.

TABLE II

Efficiency of $\tilde{\hat{Y}}_{b.}$ for Five Characteristics

Economic Regions	Employed		Unemployed		Employed Agr.		Employed Non-Agr.		In Labour Force	
	$CV(\hat{Y}_{b.})$	Eff	$CV(\hat{Y}_{b.})$	Eff	$CV(\hat{Y}_{b.})$	Eff	$CV(\hat{Y}_{b.})$	Eff	$CV(\hat{Y}_{b.})$	Eff
1	5.64	1.03	12.42	0.99	10.23	0.96	5.93	1.03	5.47	1.03
1-2	4.55	1.03	10.52	1.00	13.35	0.97	4.96	1.04	4.45	1.03
1-3	3.18	1.58	6.32	1.04	12.02	0.96	3.25	1.57	3.13	1.59
1-4	2.78	1.74	5.60	1.03	8.85	0.96	2.84	1.72	2.72	1.73
1-5	2.61	1.83	5.62	1.04	8.40	0.95	2.69	1.80	2.57	1.84
1-6	2.48	2.02	5.44	1.05	7.70	0.96	2.56	1.98	2.45	2.02
1-7	2.34	2.07	5.34	1.06	8.14	0.96	2.44	2.01	2.31	2.08
1-8	2.34	2.69	5.44	1.06	7.63	0.96	2.43	2.56	2.31	2.72
1-9	2.25	3.41	5.20	1.07	7.51	0.96	2.33	3.15	2.22	3.52
1-10	2.20	3.61	5.05	1.06	7.47	0.96	2.27	3.24	2.17	3.74