THE CHOICE OF AUXILIARY VARIABLES IN MULTIVARIATE RATIO AND REGRESSION ESTIMATORS

G. W. Lynch, University of Ottawa

ABSTRACT

Previously, the author has developed several multivariate ratio and regression estimators for use in small and moderate size sample surveys. In this study, we examine the conditions under which these estimators are the most precise. Our results indicate that: (1) when all correlations are large, the estimator attributed to Olkin is the most precise; (2) if all correlations are fairly large, or if the correlations between the auxiliary variables are small and those between the auxiliary variables and the characteristic of interest, y, are not small, then the geometric ratio estimator appears to be the most precise; (3) when all correlations are small, then the mean estimator ($\bar{y}$) which does not utilize information from the auxiliary variables is the estimator of choice. These results are in agreement with those obtained through Monte Carlo simulations of five real populations.

1. Introduction

Several authors, including Olkin [3], Raj [4], Rao and Mudholkar [5], Shukla [6], Singh [7], and Srivasta [8, 9, 10] , have proposed ratio-type estimators which utilize data from several auxiliary variables. These estimators involve the use of unknown weights which have to be estimated and assume knowledge of the population means of the auxiliary characteristics used. These estimators do not appear, from the points of view of users, to be satisfactory. Sukhatme and Chand [11], proposed a ratio-type multivariate estimator which does not involve unknown weights and which, at most, assumes knowledge of the population means of the auxiliary characteristic least correlated with the characteristic of interest. They used a multiphase sampling plan where information from one auxiliary variable is gathered at each phase and used in the final estimation process. This estimator will not be further considered since this paper is concerned with one stage or one phase sampling procedures.

In this paper, we shall present several multivariate ratio and regression estimators which had been previously developed by the author [2]. These estimators only assume that the sample means of the auxiliary variables are known and that a simple random sample (one stage or one phase) is drawn. Thus, we believe that these estimators are useful in small and moderate size sample surveys. We shall compare their approximate variances so that judgements can be made concerning the best choice of estimators in finite population sample surveys.

2. Notation

We shall specify that the population is finite and contains N units. For the i $\underline{th}$ unit, the vector $(y_i, x_{1i}, x_{2i}, \ldots, x_{pi})$ gives the $(p+1)$ variate vector - each of which are assumed to be non-negative. This requirement of non-negativity is not essential, but is the typical situation in practice and simplifies the late manipulations.

The N units of the finite population are:

$Y_1, Y_2, \ldots, Y_N$ with $\bar{Y}$ unknown and to be estimated,

$X_{11}, X_{12}, \ldots, X_{1N}$ with $\bar{X}_1$ known and positive,

$X_{21}, X_{22}, \ldots, X_{2N}$ with $\bar{X}_2$ known and positive,

. . . .

$X_{p1}, X_{p2}, \ldots, X_{pN}$ with $\bar{X}_p$ known and positive,

Suppose that a simple random sample of size n is observed from the population and the population mean, $\bar{Y}$, is to be estimated.

3. Multivariate Ratio and Regression Estimators and their Approximate Variance

Since the most frequent applications for these methods are likely to be when there are two auxiliary variates (x-variates), the methods are described for this case. However, parts of these developments can be easily be extended to the more general case.

In the approximation of the various functions, the complicated form of the terms of the order of n^{-2} makes it difficult to consider them. Accordingly, only terms of the order n^{-1} will be considered.

Mean:

$$\bar{y} = \Sigma y_i/n \qquad (3.1.1)$$

$$V(\bar{y}) = \frac{1-f}{n} S_y^2 , \qquad (3.1.2)$$

where $S_y^2 = \Sigma(Y_i - \bar{Y})^2 / (N - 1)$

Univariate Ratio with x_i:

$$\tilde{Y}_{R_i} = \bar{y}\,\bar{X}_i/\bar{x}_i \qquad (3.2.1)$$

$$V(\tilde{Y}_{R_i}) = \frac{1-f}{n} \bar{Y}^2 (C_y^2 - 2C_{yx_i} + C_{x_i}^2) \qquad (3.2.2)$$

where $\bar{x}_i$ is the sample mean of the variate, X_i, and C's are the usual coefficients of variation.

Olkin:

$$\tilde{Y}_{OLK} = w_1 r_1 \bar{X}_1 + w_2 r_2 \bar{X}_2, \qquad (3.3.1)$$

where $r_i = \bar{y}/\bar{x}_i$, $\bar{y}$ and $\bar{x}_i$ are the usual sample means, and $w_1 + w_2 = 1$ are chosen so that the precision is minimized.

For this estimator, Cochran [1] has given that

$$V(\tilde{Y}_{OLK}) = (V_{11} V_{22} - V_{12}^2)/(V_{11} - 2V_{12} + V_{22}) \qquad (3.3.2)$$

where $V_{ij} = \frac{1-f}{n} \bar{Y}^2 (C_y^2 - C_{yx_i} - C_{yx_j} + C_{x_i x_j})$

$V(\tilde{Y}_{OLK})$ can be rewritten in the form

$$V(\tilde{Y}_{OLK}) = \frac{V(\tilde{Y}_{R_1}) \, V(\tilde{Y}_{R_2}) - \text{cov}^2 (\tilde{Y}_{R_1}, \tilde{Y}_{R_2})}{\frac{1-f}{n} (C_{x_1}^2 - 2C_{x_1 x_2} + C_{x_2}^2)} \qquad (3.3.3)$$

Average Ratio:

When $w_1 = w_2 = \frac{1}{2}$, the Olkin estimator becomes

$$\tilde{Y}_{MAVG} = \frac{1}{2} \bar{y} \left(\frac{\bar{X}_1}{\bar{x}_1} + \frac{\bar{X}_2}{\bar{x}_2}\right) \qquad (3.4.1)$$

$$V(\tilde{Y}_{MAVG}) = \frac{1-f}{4n} \bar{Y}^2 (4C_y^2 - 4C_{yx_1} - 4C_{yx_2} + C_{x_1}^2 + 2C_{x_1 x_2} + C_{x_2}^2) \qquad (3.4.2)$$

$$= \frac{V(\tilde{Y}_{R_1}) + V(\tilde{Y}_{R_2})}{2} - \frac{1-f}{4n} \bar{Y}^2 (C_{x_1}^2 - 2C_{x_1 x_2} + C_{x_2}^2) \qquad (3.4.3)$$

$$= V(\bar{y}) + \left(\frac{1-f}{4n}\right) \bar{Y}^2 (C_{x_1}^2 + 2C_{x_1 x_2} + C_{x_2}^2 - 4C_{yx_1} - 4C_{yx_2}) \qquad (3.4.4)$$

Geometric:

Another alternative estimator to be considered is

$$\tilde{Y}_{MG} = \bar{y} \sqrt{\left(\frac{\bar{X}_1 \bar{X}_2}{\bar{x}_1 \bar{x}_2}\right)} \qquad (3.5.1)$$

$$V(\tilde{Y}_{MG}) = V(\tilde{Y}_{MAVG}) \qquad (3.5.2)$$

Linear Regression:

$$\tilde{Y}_{MLR} = \bar{y} + b_1 (\bar{X}_1 - \bar{x}_1) + b_2 (\bar{X}_2 - \bar{x}_2) \qquad (3.6.1)$$

$$V(\tilde{Y}_{MLR}) = \frac{1-f}{n} \left[S_y^2 - b_1^2 S_{x_1}^2 - b_2^2 S_{x_2}^2 - b_1 b_2 S_{x_1 x_2} \right] \qquad (3.6.2)$$

$$= \frac{1-f}{n} S_y^2 \left[1 - (\rho_{yx_1}^2 + \rho_{yx_2}^2 - 2\rho_{yx_1} \rho_{yx_2} \rho_{x_1 x_2}) / (1 - \rho_{x_1 x_2}^2) \right] \qquad (3.6.3)$$

where b_1 and b_2 are found by the usual least-square methods, and the ρ's are the usual correlation coefficients.

4. Comparisons of Variances

Cochran [1], in theorem 6.3, clearly states the conditions under which it would be advantageous to employ the univariate ratio estimator over the mean estimator. This theorem shows that the issue depends on the correlations between y and x and on the coefficients of variations of these two variates. Thus, (1) if the auxiliary variable, x, has a coefficient of variation which is more than twice as large as that of y, then this auxiliary variate should not be used in the estimation process since the sample man would be more precise; (2) if the coefficients of variation of y and x are approximately equal (as in the case when x_i is the value of y_i at some previous time) and the correlation between y and x is greater than 0.5, then the univariate ratio estimator is the more precise. Of course, this theorem only applies to those samples which are large enough for the approximate formula for $V(\tilde{Y}_R)$ to be valid.

The development of the multivariate linear regression estimate specified that the regression plane is approximately linear. Thus, for large n, it seems plausible that, among all linear estimators, the multivariate linear regression estimator would yield the smallest variance.

It is also quite clear that the multivariate linear regression estimator can be more precise than the mean estimator when ρ_{yx}, ρ_{yx_2}, and $\rho_{x_1 x_2}$ are positive since, from equations (3.1.2) and (3.6.3), we have that

$$V(\bar{y}) - V(\tilde{Y}_{MLR}) \geq \frac{1-f}{n} S_y^2 (\rho_{yx_1}^2 - 2\rho_{yx_1} \rho_{yx_2} \rho_{x_1 x_2} + \rho_{yx_2}^2) / (1 - \rho_{x_1 x_2}^2)$$

$$\geq \frac{1-f}{n} S_y^2 (\rho_{yx_1}^2 - 2\rho_{yx_1} \rho_{yx_2} + \rho_{yx_2}^2) / (1 - \rho_{x_1 x_2}^2), \text{ if } \rho_{x_1 x_2} > 0$$

$$= \frac{1-f}{n} S_y^2 (\rho_{yx_1} - \rho_{yx_2})^2 / (1 - \rho_{x_1 x_2}^2)$$

which is positive for ρ_{yx_1} and ρ_{yx_2} positive. In most applications, ρ_{yx_1}, ρ_{yx_2}, and $\rho_{x_1x_2}$ will be positive. Under these circumstances, the largest gains will be achieved when $\rho_{x_1x_2}$ is large and the correlation of y with x_1 is as different from that with x_2 as possible.

Olkin [3] gives the theorem which states that, if $p \geq q$, then $V(\tilde{Y}_{OLK}|p) \leq V(\tilde{Y}_{OLK}|q)$ where p and q are the number of auxiliary variables used in the estimation process. The practical relevance of this theorem is that it suggests the conditions under which the Olkin estimator (using optimal weights) is likely to be superior to the univariate ratio estimators which can be derived from the Olkin estimator setting $w_1 = 1$, $w_2 = 0$. Thus in trying to decide which estimator to use, plot the graph of y_i against x_{1i} and x_{2i}. If the relationships are roughly linear and contain the origin, then the Olkin estimator may be more precise than either of the univariate ratio estimators.

Olkin [3] also shows that if all of the coefficients of variation are approximately equal and if the correlations between all of the variables are also approximately equal, then the use of the auxiliary variables in estimation may result in increased precision over the mean estimator if $\rho > 1/(p+1)$. Thus, the Olkin estimator may be more precise than the mean estimator if the coefficients of variation of the axuiliary variables are approximatley equal and their correlations with y_i greater than $1/(p+1)$.

The Olkin estimator was constructed to minimize the variance of the expression:

$$\tilde{Y}_{OLK} = w_1 r_1 \bar{X}_1 + w_2 r_2 \bar{X}_2$$

subject to the restriction that $w_1 + w_2 = 1$. Clearly, this minimization is over a wider class than the class of linear unbiased estimators. Thus, the Olkin estimator would seem to have the potential to be more precise than the multivariate linear regression estimators. However, in view of the amount of computations involved, this estimator may be most useful "in small surveys of a specialized nature". [1]

Of course, the above discussion tacitly assumes that the correlations are fairly large since, if the correlations between y and the auxiliary variables, x_i, are small: (a) it is well known that the mean estimator may be more precise than the univariate ratio estimators; (b) we can rewrite the formula for the variance of the Olkin estimator, equation (3.3.3), in the form:

$$V(\tilde{Y}_{OLK}) \doteq \left[V(\tilde{Y}_{R_1})\, V(\tilde{Y}_{R_2}) - \text{cov}^2(\tilde{Y}_{R_1}, \tilde{Y}_{R_2})\right] / \left(\frac{1-f}{n}\right)(C^2_{x_1} - 2C_{x_1x_2} + C^2_{x_2})\,\bar{Y}^2$$

and the numerator is large in relationship to the denominator for the small correlations, ρ_{yx_1}, ρ_{yx_1}, ρ_{yx_2} and $\rho_{x_1x_2}$; and (c) we have stated earlier that the multivariate linear regression estimator was designed to take advantage of the correlations between y and the auxiliary variates.

From the formulae, (3.4.3) and (3.5.2) for the approximations (to the order n^{-1}) to the variance of the geometric and average ratio estimators, it is clear that when $C^2_{x_1} \doteq C_{x_1x_2} \doteq C^2_{x_2}$, then $V(\tilde{Y}_{MG} = V(\tilde{Y}_{MAVG})$ can be greater than either $V(\tilde{Y}_{R_1})$ or $V(\tilde{Y}_{R_2})$ - but not both.

In summary to this point, the suggestions are (1) when the correlations between y and the auxiliary variates are large and the coefficients of variation are approximately equal, the Olkin and the multivariate linear regression estimators would seem to be the most precise estimators when n is large enough for the variance formulae to be valid. (2) If the correlations are large and the regression plane is approximately linear and contains the origin, use the Olkin estimator unless the calculations are prohibitive; if the regression plane does not contain the origin, use the multivariate linear regression estimator. (3) When the correlations between y and x_1, and y and x_2 are large but $\rho_{x_1x_2}$ is small ($C^2_{x_1} \doteq C^2_{x_2} \neq C_{x_1x_2}$), then use the geometric or separate ratio estimators. Now it remains to decide which estimator to use when the correlations are small or moderate.

It is clear that the univariate ratio, the Olkin and the multivariate linear regression estimators rely heavily on the correlations between y and the auxiliary variates. Thus we shall now restrict our attention to comparisons between the geometric (and the separate ratio) estimator and the mean estimator.

We can rewrite $V(\tilde{Y}_{MG}) = V(\tilde{Y}_{MAVG})$ in the form:

$$V(\tilde{Y}_{MG}) = V(\tilde{Y}_{MAVG}) \doteq \frac{1-f}{n} S^2_y + \frac{1-f}{4n}\bar{Y}^2 (C^2_{x_1} + 2C_{x_1x_2} + C^2_{x_2} - 4C_{yx_1} - 4C_{yx_2})$$

$$= V(\bar{y}) + \frac{1-f}{n}\bar{Y}^2 (C^2_{x_1} + 2C_{x_1x_2} + C^2_{x_2} - 4C_{yx_1} - 4C_{yx_2}) \qquad (4.1)$$

Thus, if all of the correlations, $\rho_{x_1x_2}$, ρ_{yx_1} and ρ_{yx_2}, are small but the coefficients of variation, $C^2_{x_1}$, $C^2_{x_2}$ and C^2_y are not too small (thereby implying that $C_{x_1x_2}$, C_{yx_1} and C_{yx_2} are small), then $V(\tilde{Y}_{MG}) = V(\tilde{Y}_{MAVG}) \geq V(\bar{y})$. If the coefficients of variation of y and the auxiliary variates are approximately equal, then equation (4.1) becomes:

$$V(\tilde{Y}_{MG}) = V(\tilde{Y}_{MAVG}) = V(\bar{y}) + \frac{1-f}{4n} \bar{Y}^2$$

$$(2 + 2\rho_{x_1x_2} - 4\rho_{yx_1} - 4\rho_{yx_2}) C_y^2$$

$$\Rightarrow \quad V(Y_{MG}) = V(Y_{MAVG}) \leq V(\bar{y}) \text{ when } (1 + \rho_{x_1x_2}) \leq$$

$$2(\rho_{yx_1} + \rho_{yx_2}).$$

Then the geometric and separate ratio estimators would seem to be the estimators of choice when $\rho_{x_1x_2}$ is small or negative and ρ_{yx_1} and ρ_{yx_2} are of moderate size - this size depending on the value of $\rho_{x_1x_2}$. Of course, if ρ_{yx_1} and ρ_{yx_2} are both greater than 0.5, then these two estimators appear to be more precise than the mean estimator. Lastly $C_{x_1x_2}$ is very small but $C^2_{x_1}$ and $C^2_{x_2}$ are not too small (thus $\rho_{x_1x_2}$ is small), equation (4.1) reduces to:

$$V(\tilde{Y}_{MG}) = V(\tilde{Y}_{MAVG}) = V(\bar{y}) + \frac{1-f}{4n} \bar{Y}^2$$

$$(C^2_{x_1} \quad C^2_{x_2} - 4C_{yx_1} - 4C_{yx_2})$$

$$= V(\bar{y}) + \frac{1-f}{4n} \bar{Y}^2 \quad (C^2_{x_1} \quad C^2_{x_2} -$$

$$4\rho_{yx_1} C_y C_{x_1} - 4\rho_{yx_2} C_y C_{x_2}).$$

Thus, if $4\rho_{yx_1} C_y > C_{x_1}$ and $4\rho_{yx_2} C_y > C_{x_2}$, then the geometric and separate ratio estimators would seem to be the most precise.

To summarize, when n is large enough for the variance formulae to be valid, the following are suggested:

(1) If y and x_1, and y and x_2, and x_1 and x_1 are all highly correlated and the coefficients of variation are also large, use the Olkin estimator when the regression plane is approximately linear and contains the origin unless the computations are prohibitive - if the computations for the Olkin estimator are too lengthy, then use the multivariate linear regression estimator; when the regression plane is approximately linear but does not contain the origin, and the correlations and coefficients of variation are large, the multivariate linear regression estimator would seem to be the estimator of choice.

(2) If all correlations are fairly large and the coefficients of variation are small, the geometric or separate ratio estimators may be the most precise.

(3) If the correlation between x_1 and x_2 is small or negative and the remaining correlations are not very small, the geometric or separate ratio estimators again may yield the smallest variances.

(4) If all correlations are small, use the mean estimator.

In a separate work, Lynch [2] used Monte Carlo simulations on five natural populations to ascertain whether the suggestions stated above for sample sizes large enough for the approximate variances to be valid, could be substantiated for small or moderate n. In that work, the author used populations where the coefficients of variations varied from 0.1 to 1.4, the correlation from about 0.1 to about 1.0 and population sizes from 33 to 350. His results indicated agreement, in general, with the above findings.

5. Summary and Conclusions

Several methods of utilizing auxiliary variables in the estimation of some characteristic, y, have been proposed and considered. These methods are generally referred to as ratio and regression estimators. In addition to improving the efficiency of the estimation procedure, one of the aims of these procedures may be to assist in the choice of the structure. In this work, we examine the conditions under which each of the given estimators are the most precise.

Our results indicate that: (1) when all correlations are large, the estimator attributed to Olkin is the most precise; (2) if all correlations are fairly large, or if the correlations between the auxiliary variables are small and all other correlations are not small, then the geometric ratio estimator appears to be the most precise; (3) when all correlations are small, the mean estimator ($\bar{y}$) which does not utilize information from any auxiliary variables is the estimator of choice.

Acknowledgements

The author is indebted to Professor D.J. Thompson for suggesting the topic and guidance throughout. Much of this work was carried out while the author was a Graduate Student at the University of Washington in Seattle and was a National (Canada) Student Health Fellow. Currently, the author is supported by the Grant, RD-10, of the Government of Ontario, Canada.

References

1. Cochran, W.G. (1963). *Sampling Techniques*, 2nd Edition, New York, Wiley Publications.

2. Lynch, G.W. (1976). "Some Multivariate Ratio and Regression Estimators in Finite Population Survey Sampling: Developments and Peformances". Unpublished Ph.D. dissertation, University of Washington, Seattle.

3. Olkin, I. (1958). "Multivariate Ratio Estimation for Finite Populations". *Biometrika*, 45: 154-165.

4. Raj, D. (1965). "On a Method of Using Multi-auxiliary Information Sample Surveys". *JASA*, 60: 270-77.

5. Rao, P.S.R.S. and Mudholkar, G.D. (1967). "Generalized Multivariate Estimator for the Mean of Finite Populations". *JASA*, 62: 1009-1012.

6. Shukla, G.K. (1966). "An Alternative Multivariate Ratio Estimator for Finite Populations". *Calcutta Statistical Association Bulletin*, 15: 127-134.

7. Singh, M.P. (1967). "Ratio Cum Product Method of Estimation". *Metrika*, 12: 34-42.

8. Srivasta, S.K. (1965). "An Estimation of the Mean of a Finite Population Using Several Auxiliary Variables". *Journal of the Indian Statistical Association*, 3: 189-194.

9. Srivasta, S.K. (1967). "An Estimator Using Auxiliary Information in Sample Surveys". *Calcutta Statistical Association Bulletin*, 16: 121-132.

10. Srivasta, S.K. (1971). "A Generalized Estimator for the Mean of a Finite Population Using Multi-auxiliary Information". *JASA*, 66: 404-407.

11. Sukhatme, B.V. and Chand, L. (1978). "Multivariate Ratio-type Estimators". Proceedings of the Social Statistics Section of the American Statistical Association Meetings, Chicago, Illinois. August 15-18, 1977, 927-931.