Jour. Ind. Soc. Ag. Statistics
44 (3), 1992 : 227-235

The Effect of Two-Stage Sampling on O L S Estimators

R.C. Agarwal and O.P. Kathuria
IASRI, New Delhi – 110 012
(Received : February, 1989)

Summary

Two-stage sampling is commonly adopted by selecting a sample of clusters at the first stage and then a sample of individuals within selected clusters at the second stage. A regression model under some suitable conditions which reflects the effect of clustering on the character under study has been considered. The *misspecification effect* described by Scott and Holt [3] has been examined under some suitable cost function and certain other conditions.

Key words : Two-stage sampling, OLS estimators, regression model, misspecification effect.

Introduction

The usual approach of estimating the regression parameter 'B' of a finite population can not be readily applied to situations where the population may consist of hierarchy of sample units such that different stages of sampling units exhibit different degrees of variability in the population (Chaudhuri, [1]). In two-stage sampling involving selection of a sample of clusters in the first stage and then a sample of individuals from the selected clusters at the second stage, clusters used in sampling designs almost always exhibit some degree of homogeneity with respect to the variables under study and this homogeneity has also been found to occur with regression residuals (Scott and Holt, [3]). The consequence of this homogeneity is that the units within a selected cluster are not independent of each other. So, it is desirable to study this effect on the estimators using regression models.

2. *Regression model under Two-stage sampling*

The regression model

$$y_{ij} = \beta_0 + \beta_x x_{ij} + \beta_z z_{ij} + \alpha_i + e_{ij} \qquad (1)$$

$$i = 1, 2, \ldots, a$$
$$j = 1, 2, \ldots, n_i$$

may be considered to reflect the effect of clustering on the character under study where the subscript (i, j) refers to individual j in cluster i and α_i is the effect of i^{th} cluster. x's and z's are the two explanatory variables, β_x and β_z are the regression coefficients of x and z respectively and β_o is the average effect. The following assumptions are considered in the model:

(i) α_i is the assumed to be random effect with $E(\alpha_i) = 0$ and $V(\alpha_i) = \sigma_\alpha^2$

(ii) $E(e_{ij}) = 0$ and $V(e_{ij}) = \sigma_e^2$
$COV(e_i, e_k) = 0$ for $i \neq k$

(iii) α_i's and e_{ij}'s are uncorrelated with the x's and z's.

(iv) The random factor α_i and the error term e_{ij}'s are also uncorrelated.

3. *OLS Estimators of the Regression Coefficients and their Variance:*

Consider the model

$$y_{ij} = \beta_0 + \beta_x x_{ij} + \beta_z z_{ij} + e_{ij} \qquad (2)$$

then it can be proved that the least square estimators for β_x and β_z are:

$$b_x = \frac{S_{zz} \sum_i \sum_j (x_{ij} - \bar{x})\, y_{ij} - S_{xz} \sum_i \sum_j (z_{ij} - \bar{z})\, y_{ij}}{\Delta} \qquad (3)$$

and

$$b_z = \frac{S_{xx} \sum_i \sum_j (z_{ij} - \bar{z})\, y_{ij} - S_{xz} \sum_i \sum_j (x_{ij} - \bar{x})\, y_{ij}}{\Delta} \qquad (4)$$

where,

$$S_{zz} = \sum_i \sum_j (z_{ij} - \bar{z})^2$$

$$S_{xz} = \sum_i \sum_j (x_{ij} - \bar{x})(z_{ij} - \bar{z})$$

and $$\Delta = \sum_i \sum_j (x_{ij} - \bar{x})^2 \sum_i \sum_j (z_{ij} - \bar{z})^2 - \left[\sum_i \sum_j (x_{ij} - \bar{x})(z_{ij} - \bar{z})\right]^2$$

Applying these OLS estimators to the model (Eq–1) having one additional effect α_i which is known as the cluster effect, the variance of estimates of regression coefficients can be written as,

$$V(b_x) = \frac{\sum_i \sum_j K_{ij}^2 V(y_{ij}) + \sum_i \sum_{j \neq j'} K_{ij} K_{ij'} \text{Cov}(y_{ij}, y_{ij'})}{\Delta^2}$$

where,

$$K_{ij} = \sum_i \sum_j \left[S_{zz}(x_{ij} - \bar{x}) - S_{xz}(z_{ij} - \bar{z}) y_{ij}\right]$$

Since, $V(y_{ij}) = \sigma_\alpha^2 + \sigma_e^2$
and $\text{Cov}(y_{ij}, y_{i'j'}) = \sigma_\alpha^2$ for $j \neq j'$, this can be written as

$$V(b_x) = \frac{\sum_i \sum_j K_{ij}^2 (\sigma_\alpha^2 + \sigma_e^2) + \sum_i \sum_{j \neq j'} K_{ij} K_{ij'} \sigma_\alpha^2}{\Delta^2}$$

$$= \frac{\sigma_\alpha^2 \left[\sum_i \left(\sum_j K_{ij}\right)^2\right] + \sigma_e^2 \left[\sum_i \sum_j K_{ij}^2\right]}{\Delta^2}$$

$$= \frac{\sigma_\alpha^2 \left[\sum_i n_i^2 \left\{S_{zz}(x_i - \bar{x}) - S_{xz}(z_i - \bar{z})\right\}^2\right]}{\Delta^2}$$

$$+ \frac{\sigma_e^2 \left[\sum_i \sum_j \left\{S_{zz}(x_{ij} - \bar{x}) - S_{xz}(z_{ij} - \bar{z})\right\}^2\right]}{\Delta^2} \qquad (5)$$

Assuming that same sample size is selected from each cluster i.e. $n_i = \bar{n}$, let

$\eta^2(x)$ be the proportion of the variance in x explained by the clusters i.e.,

$$\eta^2_{(x)} = \frac{\bar{n}\sum_i (x_i - \bar{x})^2}{\sum_i \sum_j (x_{ij} - \bar{x})^2}$$

Similarly, define the terms $\eta^2(z)$ and $\eta(xz)$.

If 'a' is the number of clusters selected at first stage then from Eq–5,

$$V(b_x) = A\left[\frac{\sigma_\alpha^2 \eta^2}{a} + \frac{\sigma_e^2}{\bar{n}\, a}\right] \qquad (6)$$

where $A = \dfrac{\sigma_z^2}{\sigma_x^2 \sigma_z^2 - \sigma_{xz}^2}$ and $\eta^2_{(x)} = \eta^2_{(z)} = \eta_{(xz)} = \eta^2$

An estimator of $V(b_x)$ may be obtained by substituting estimates $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_e^2$ for σ_α^2 and σ_e^2 in the expression for $V(b_x)$ (Kalton, [2]).

The quantity σ_e^2 may be estimated by the residual mean square from one way analysis of variance of the y values in the clusters, i.e. by,

$$\hat{\sigma}_e^2 = \sum_i \sum_j \frac{(y_{ij} - \bar{y}_i)^2}{n - a} \qquad (7)$$

and σ_α^2 may be estimated by

$$\hat{\sigma}_\alpha^2 = \frac{\sum_i \sum_j (y_{ij} - b_o - b_x x_{ij} - b_z z_{ij})^2 - (n-2)\sigma_e^2}{\lambda(a-2)} \qquad (8)$$

where $\sum_i \sum_j (y_{ij} - b_o - b_x x_{ij} - b_z z_{ij})^2$ is the residual sum of squares

from the regression of y on x, b_o is the sample estimate of the intercept β_o and

$$\lambda = \frac{n^2 - \sum_i n_i^2}{n\,(a-2)}$$ [Snedecor and Cochran, [4] Sec. 13.7]

3. *Optimum Sub-sample Size* $\bar{n}$

Assume a simple cost model of the form,

$$C = a\,c_a + n\,c$$

where c_a is the cost of including a cluster in the sample, c is the cost of including an individual and $n = \bar{n}\,a$ is the total sample size.

For given A, the optimum choice of $\bar{n}$ that minimizes $V(b_x)$ (Eq-6) for fixed total cost C may be obtained by using Cauchy- Schwartz inequality. This gives,

$$\bar{n}_{opt.} = \left(\frac{C_a}{c} \cdot \frac{\sigma_e}{\sigma_\alpha\,\eta}\right)^{1/2} \qquad (9)$$

Defining the intra-class correlation coefficient for the clusters as the proportion of the variance of y_{ij} conditional on x_i that is accounted for by the cluster effect i.e. if ρ is the intra-class correlation then,

$$\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_e^2} = \frac{\sigma_\alpha^2}{\sigma^2} \qquad \text{where } \sigma^2 = \sigma_\alpha^2 + \sigma_e^2 \qquad (10)$$

Now $V(b_x)$ can be expressed in terms of ρ as

$$V(b_x) = A\left[\frac{\rho\,\sigma^2\eta^2}{a} + (1-\rho)\,\frac{\sigma^2}{n}\right]$$

$$= \frac{A\sigma^2}{n\left[1 + \rho(\bar{n}\eta^2 - 1)\right]} \qquad (11)$$

Hence, optimum sample size for each cluster can be written as

$$\bar{n}_{opt.} = \left[\frac{c_a}{c} \times \frac{1-\rho}{\eta^2 \rho} \right]^{1/2}$$

The results obtained can also be applied as an approximation to situations where the sub-sample size varies to a small extent between clusters (in which case $\bar{n}$ represents the average sub-sample size).

4. *Misspecification Effect*

If the cluster effect α_i in the model is not considered then $V(b_x)$ (Eq-11) can be written as,

$$V(b_x) = A \frac{\sigma^2}{n} \qquad (12)$$

Hence, ignoring the cluster effect α_i for the i^{th} cluster underestimates the variance by a factor $\left[1 + \rho (\bar{n}\eta^2 - 1) \right]$.

Scott and Holt [3] have coined this effect as the *misspecification effect* since it represents the effect of wrongly omitting the clustering effect from the model.

In the table, misspecification effect and the optimum sample size have been worked out for some hypothetical cost ratios. Since the misspecification effect changes with the change in Intra-class correlation coefficient ρ and η^2, these cost ratios have been considered over a wide range of ρ(0.1 to 0.9) and η^2(0.5 to 0.9).

It is clear from the table that for given cost ratio as η^2 increases, optimum sample size either decreases or remains same. Further, at low cost ratio and for different values of η^2's optimum sample sizes are almost same. For different η^2's as cost ratio increases, the difference between optimum sample sizes increases. At higher values of intra-class correlation coefficient, the optimum sample size reduces to a great extent (even upto 1).

It is further observed that for any given cost ratio, the misspecification effect increases as the value of η^2 increases. Because the optimum sample size reduces as cost ratio increases and misspecification effect depends largely upon optimum sample size, so from the table it is evident that sample size inversely affects the misspecification effect.

Table. Optimum sample size and misspecification effect for some hypothetical cost ratios

Cost ratio c_a/c	Values of η^2									
	0.5	0.6	0.7	0.8	0.9	0.5	0.6	0.7	0.8	0.9
	ρ= 0.1					ρ= 0.2				
9	13[@] 1.55[!]	12 1.62	11 1.67	10 1.70	9 1.71	8 1.60	8 1.76	7 1.76	7 1.92	6 1.88
12	15 1.65	13 1.68	12 1.74	12 1.86	11 1.89	10 1.80	9 1.88	8 1.92	8 2.08	7 2.06
18	18 1.80	16 1.86	15 1.95	14 2.02	13 2.07	12 2.00	11 2.12	10 2.20	9 2.24	9 2.24
27	22 2.00	20 2.10	19 2.23	17 2.26	16 2.34	15 2.30	13 2.36	12 2.48	12 2.72	11 2.78
39	26 2.22	24 2.34	22 2.24	21 2.58	20 2.70	18 2.60	16 2.72	15 2.90	14 3.04	13 3.14
54	31 2.45	28 2.58	26 2.72	25 2.90	23 2.97	21 2.90	19 3.08	18 3.32	16 3.36	15 3.50
	ρ= 0.5					ρ= 0.6				
9	4 1.50	4 1.70	4 1.90	3 1.70	3 1.85	3 1.30	3 1.48	3 1.66	3 1.84	3 2.02
12	5 1.75	4 1.70	4 1.90	4 2.10	4 2.30	4 1.60	4 1.84	3 1.66	3 1.84	3 2.02
18	6 2.00	5 2.00	5 2.25	5 2.50	4 2.30	5 1.90	4 1.84	4 2.08	4 2.32	4 2.56
27	7 2.25	7 2.60	6 2.60	6 2.90	5 2.75	6 2.20	5 2.20	5 2.50	5 2.80	4 2.56
39	9 2.75	8 2.90	7 2.95	7 3.30	7 3.65	7 2.50	7 2.92	6 2.92	6 3.28	5 3.10
54	10 3.00	9 3.20	9 3.65	8 3.70	8 4.10	8 2.80	8 3.28	7 3.34	7 3.76	6 3.64

[@] Optimum sample size
[!] Misspecification effect

Cost ratio C_a/C	Values of η^2									
	0.5	0.6	0.7	0.8	0.9	0.5	0.6	0.7	0.8	0.9
	$\rho = 0.7$					$\rho = 0.8$				
9	3 1.35	3 1.56	2 1.28	2 1.42	2 1.56	2 1.00	2 1.16	2 1.32	2 1.48	2 1.64
12	3 1.35	3 1.56	3 1.77	3 1.98	2 1.56	2 1.00	2 1.16	2 1.32	2 1.48	2 1.64
18	4 1.70	4 1.98	3 1.77	3 1.98	3 2.19	3 1.40	3 1.64	3 1.88	2 1.48	2 1.64
27	5 2.05	4 1.98	4 2.26	4 2.54	4 2.82	4 1.80	3 1.64	3 1.88	3 2.12	3 2.36
39	6 2.40	5 2.40	5 2.75	5 3.10	4 2.82	4 1.80	4 2.12	4 2.44	3 2.12	3 2.36
54	7 2.75	6 2.82	6 3.24	5 3.10	5 3.45	5 2.20	5 2.60	4 2.44	4 2.76	4 3.08
	$\rho = 0.9$									
9	1 0.55	1 0.64	1 0.73	1 0.82	1 0.91					
12	2 1.00	1 0.64	1 0.73	1 1.82	1 0.91					
18	2 1.00	2 1.18	2 1.36	2 1.54	1 0.91					
27	2 1.00	2 1.18	2 1.36	2 1.54	2 1.72					
39	3 1.45	3 1.72	2 1.36	2 1.54	2 1.72					
54	3 1.45	3 1.72	3 1.99	3 2.26	3 2.53					

It is concluded that at high values of intra-class correlation coefficient (ρ) when the proportion of variance in either of the variables (i.e. x, z or xz) explained by clusters is not very high, the inclusion of cluster effect in the regression model is not of much interest. Even at the low values of η^2 if optimum sample size is reasonably small, one should take a serious concern about the inclusion of cluster effect in the regression model.

REFERENCES

[1] Chaudhuri Arijit, 1987. Analysis of data with complex survey designs: a review. Paper presented in National Symposium on Sample Surveys in Indian Agriculture-Problems and Prospects, 17-19 Sept., 1987, Technical Session III, I.A.S.R.I., New Delhi.

[2] Kalton, G., 1983. Estimating regression coefficients from clustered samples: Sampling errors and optimum sample allocation. NASA Contractor Report, 166117.

[3] Scott and Holt D., 1982. The effect of Two-Stage Sampling on OLS Methods: *Journal of the American Statistical Association*, **77**, 848-854.

[4] Snedecor, G.W. and Cochran W.G., 1980. Statistical Methods 7thed. Iowa State University Press, Ames, Iowa.