Jour. Ind. Soc. Ag. Statistics
47(2), 1995 : 176–183

The Effect of Survey Design on Regression Analysis : An Empirical Investigation

R.C. Agrawal and O.P. Kathuria*
National Bureau of Plant Genetic Resources, New Delhi-110012
(Received : September, 1993)

SUMMARY

An empirical investigation is carried out to study the effect of survey design on regression analysis. Under three different situations A, B and C, six different sample designs have been considered for the study. Under situation A, the complete population has been taken into consideration with X_1 as dependent, X_2 as independent variable and X_3 as design variable. In situation B and C, two phase sampling has been adopted; X_3 and X_1 have been used as design variables respectively. The bias of OLS estimator and mean square errors of other estimators have been compared under different sampling designs for the three situations.

Key Words : OLS estimator, Double sampling, Design variable, Regression analysis.

Introduction

In the complex survey design, the data is often analysed using regression techniques without further regard to the sample design (Nathan and Holt [4], Holt *et al.* [2]. The algebraic comparison of different estimators proposed by Nathan and Holt [4] in case of most of the sampling designs is difficult to put in practice. So, an empirical investigation is adopted for this comparison. For this purpose, data of a pilot sample survey for estimation of inland fishery resources and catch in a region of West Bengal, India is used.

2. *Description of the investigation*

The data from 1350 ponds obtained from a study conducted in India for developing sampling methodology for estimating the extent of area under ponds and catch of fish from them has been considered as the complete population (Kathuria et al [3]). For each pond, the observations on fish catch in Kg. (X_1), total quantity of fish seed used in Kg. (X_2) and the area of the pond in acres (X_3) have been taken from the survey. Variable X_3 is treated as the design variable.

The population of 1350 ponds has been divided into 5 strata, on the basis of the values of X_3 with stratum sizes 119, 516, 482, 153 and 80.

* Indian Agricultural Statistics Research Institute, New Delhi-110012

The following six survey designs have been considered :

(a) Simple random sampling.

(b) Stratified sampling with proportional allocation.

(c) Stratified sampling with equal sample sizes.

(d) Stratified sampling with sample sizes in U-shape.

(e) Stratified sampling with sample sizes in increasing or decreasing order.

(f) Probability proportional to size and with replacement.

These survey designs have been used under three different situations :

(A) The population of 1350 ponds is considered as the complete population and X_3 is used as the design variable.

(B) A large sample of size 800 has been selected from the population of 1350 ponds. This has been done to see whether double sampling as a method of design could be adopted for estimating regression coefficients when the design variable X_3 is already available from the survey or can be measured cheaply.

(C) A special situation is when X_1 is measured at the first phase and so the dependent variable is used as the design variable (Nathan and Holt, [4], p.385). For this situation also, a random sample of size 800 has been drawn for X_1 at the first phase.

From the population of 1350 ponds, the following parameter values have been obtained :

$$\rho_{12} = 0.32592 \qquad \rho_{13} = 0.13233$$

$$\rho_{23} = 0.38352 \qquad \rho_{13.2} = 0.00839$$

$$\beta_{12} = 2.84782 \qquad \sigma_1 = 114.29210$$

$$\sigma_3 = 0.42250 \qquad \sigma_2 = 13.08020$$

These values of the parametes have been calculated from the complete population and have been assumed to be known for three different situations. For any particuar design, 100 repeated samples have been drawn and the design dependent parameters are estimated. Finally the estimates of the popultion parameters and the design dependent parameters have been combined to estimate the moments of the various estimators. In this way the bias and the mean square error of ordinary least squares estimator and variance of unbiased estimators under three different situations using various survey designs defined earlier have been calculated.

Situation A :

N = 1350, n = 140, X_3 is design varibale

Strata (On the basis of X_3 variable)

S.No.	Interval (in acres)	Size
1	0.00–0.25	119
2	0.25-0.50	516
3	0.50–1.00	482
4	1.00–1.50	153
5	1.50–3.25	80
		1350

Conversion : one acre = 0.405 ha

Survey design :

(a) Simple random sampling.

(b) Stratified sampling with proportional allocation i.e. with sample sizes (12, 53, 49, 15, 11).

(c) Stratified sampling with equal sample sizes (28, 28, 28, 28, 28).

(d) Stratified sampling with sample sizes in U-shape (40, 25, 10, 25, 40).

(e) Stratified sampling with sample sizes (70, 35, 20, 10, 5).

(f) Probability proportional to size.

Situtaion B :

Double sampling, X_3 is design variable. Sample size for X_3 variable at first phase n′ = 800 and n = 83.

Strata (On the basis of X_3 variable)

S.No.	Interval (in acres)	Size
1	0.00–0.25	75
2	0.25-0.50	314
3	0.50–1.00	253
4	1.00–1.50	107
5	1.50–3.25	51
		800

Survey design :

(a) Simple random sampling.

(b) Stratified sampling with proportional allocation i.e. with sample sizes (8, 32, 26, 11, 6).

(c) Stratified sampling with equal sample sizes (16, 16, 17, 17, 17).

(d) Stratified sampling with sample sizes (8, 20, 27, 20, 8).

(e) Stratified sampling with sample sizes (8, 13, 17, 20, 25).

(f) Probability proportional to size.

Situation C :

Double sampling, X_1 is used as the design variable. Sample size for X_1 variable at first phase $n' = 800$ and $n = 83$.

Strata (On the basis of X_3 variable)

S.No.	Interval (in Kg)	Size
1	0–50	170
2	50-100	256
3	100–200	190
4	200–300	124
5	300–400	60
		800

Survey design

(a) Simple random sampling.

(b) Stratified sampling with proportional allocation i.e. with sample sizes (17, 25, 20, 13, 8).

(c) Stratified sampling with equal sample sizes (16, 16, 17, 17, 17).

(d) Stratified sampling with sample sizes (8, 20, 27, 20, 8).

(e) Stratified sampling with sample sizes (8, 13, 17, 20, 25).

(f) Probability proportional to size.

3. *Methodology and Formulae used*

We consider the simple situation described by Nathan and Holt (1980). Suppose, "design" variable X_3 is known at the design stage for each member of the finite population. After sampling, observations are made on X_1, the

dependent variable, and on X_2 the independent variable in the regression analysis.

Consider the regression model,

$$E(X_1 \mid X_2) = \mu_1 + \beta_{12}(X_2 - \mu_2) \tag{1}$$

In the above model, X_3 variable is used at the design stage but not explicity in the regression model. In other words, this model has been considered under the situation where the design variable X_3 is not a cause of the dependent variable.

We consider a finite population of size N selected from a superpopulation such that the observed values of X_3 are independently and indentically distributed with mean μ_3 and variance σ_3^2. The assumption of the survey design is that X_3 is related in some way to X_1 and or X_2. The following are the assumptions for the model considered :

$$\begin{aligned} X_{1\alpha} &= \mu_1 + \beta_{13}(X_{3\alpha} - \mu_3) + e_{1\alpha} \\ X_{2\alpha} &= \mu_2 + \beta_{23}(X_{3\alpha} - \mu_3) + e_{2\alpha} \\ e_{1\alpha} &= \beta_{12.3}\, e_{2\alpha} + \eta_{1\alpha} \end{aligned} \tag{2}$$

where, $E(e_{2\alpha} \mid X_{3\alpha}) = E(\eta_{1\alpha} \mid X_{3\alpha}) = E(e_{2\alpha}\, \eta_{1\alpha} \mid X_{3\alpha}) = 0$

and $E(e_{2\alpha}^2 \mid X_{3\alpha}) = \sigma_{2.3}^2$, $E(\eta_{1\alpha}^2 \mid X_{3\alpha}) = \sigma_{1.23}^2$

These conditions are equivalent to the conditional expectations of X_1 and X_2 given X_3 being linear in X_3 and the conditional covariance matrix of X_1 and X_2 given X_3, not depending on X_3.

In addition, we assume conditional independence for different units i.e. $(\eta_{1\alpha}, e_{2\alpha})$ $(\eta_{1\beta}, e_{2\beta})$ are conditionally independent given $X_3' = (X_{31}, X_{32}, ..., X_{3N})$. A sample 'S' is selected from the finite population by any sample design (including purposive designs) of fixed size n. The design may be based on the known population values of X_3. The parameter of interest is the superpopulation regression coefficient β_{12} of X_1 on X_2 which can be defined as $\beta_{12} = \rho_{12}\, \sigma_1/\sigma_2$

We further define the following statistics based on the entire finite population :

$$\begin{aligned} \hat{\mu}_3 &= (\Sigma_{\alpha=1}^{N} X_{3\alpha})/N \\ \delta_3^2 &= \Sigma_{\alpha=1}^{N} (X_{3\alpha} - \hat{\mu}_3)^2/(N-1) \end{aligned} \tag{3}$$

The sample statistics $\bar{x}_i$, s_i^2, s_{ij}, $s_{ij.k}$ (i, j, k = 1, 2, 3) etc. are defined in the usual way, analogous to the corresponding distribution parameters, μ_i, σ_i^2, σ_{ij}, $\sigma_{ij.k}$ which appear in the assumptions defined earlier.

We have used Ordinary Least-Square (OLS) estimator (b_{12}) and an alternative estimator ($\hat{\beta}_{12}$) defined by Nathan and Holt (1980).

$$b_{12} = S_{12}/S_2^2 \tag{4}$$

$$\hat{\beta}_{12} = \frac{S_{12} + (S_{13}\, S_{23}/S_3^2)\,(\hat{\sigma}_3^2)/S_3^2 - 1)}{S_2^2 + (S_{23}^2/S_3^2)\,(\hat{\sigma}_3^2/S_3^2 - 1)} \tag{5}$$

The maximum likelihood estimator $\hat{\beta}_{12}$, under a trinormal distribution for (X1, X2, X3) was originally due to Pearson [5] and is proposed by Demets and halperin [1] to provide an asymptotically unbiased estimator of β_{12} to $O(n^{-1})$.

The weighted estimators of b_{12} and β_{12} where the weights are the inverse of the sample inclusion probabilities π_α, have also been utilized for comparison purposes. The weighed estimators can be written as –

$$b_{12}^* = S_{12}^*/S_2^{*2} \tag{6}$$

and

$$\hat{\beta}_{12}^* = \frac{S_{12}^* + (S_{13}^*\, S_{23}^*/S_3^{*2})\,(\hat{\sigma}_3^2)/S_3^{*2} - 1)}{S_2^{*2} + (S_{23}^{*2}/S_3^{*2})\,(\hat{\sigma}_3^2/-1)} \tag{7}$$

where,

$$\bar{X}_i = \Sigma_{\alpha \in S}\, Xi_{i\alpha}/N\pi_\alpha \;(i = 1, 2, 3)$$

$$S_{ij}^* = \Sigma_{\alpha \in S}\, \frac{X_{i\alpha}\, X_{j\alpha}}{N\,\pi_\alpha} - \frac{\bar{X}_i\, \bar{X}_j}{(\Sigma_{\alpha \in S}{}^1/N\pi_\alpha}$$

$$S_i^{**} - S_{ii}^* \;(i, j = 1, 2, 3)$$

and $\pi_\alpha = \text{prob.}\,(\alpha \in S \mid X3) > 0\,;\, (\alpha = 1, 2, \ldots N).$

The variance and mean square error (mse) of these four estimators have been compared under different survey designs considered. Since the variance expressions of b_{12} and β_{12} depends upon Q where $Q = \dfrac{E(s_3^2)}{\sigma_3^2}$, this term has also been calculated for comparision purposes.

4. *Results and discussion :*

Table 1. *Bias and Mean squre error of OLS estimator and variance of unbiased estimators*

Survey Design	$E(b_{12}) - \beta_{12}$	MSE (b_{12})	$V(\hat{\beta}_{12})$	$V(b^*_{12})$	$V(\hat{\beta}^*_{12})$	Q
			Situation A			
a	0.000000	0.487550	0.487566	0.487538	0.487488	1.0
b	0.003132	0.475688	0.477120	0.464144	0.464138	1.2
c	0.020419	0.413333	0.448060	0.464144	0.464138	2.2
d	0.029660	0.381396	0.440560	0.464144	0.464138	2.9
e	–0.003505	0.500815	0.503366	0.464144	0.464138	0.8
f	0.016279	0.427990	0.452709	0.479300	0.479991	2.0
			Situation B			
a	0.000000	0.822420	0.8223950	0.822210	0.824335	1.0
b	0.001505	0.815863	0.813436	0.787738	0.787728	1.1
c	0.002520	0.703793	0.760084	0.787738	0.787728	2.2
d	0.005150	0.789950	0.798320	0.787738	0.787728	1.3
e	0.021200	0.977696	0.756826	0.787738	0.787728	2.3
f	0.014240	0.734180	0.770284	0.813036	0.813002	1.8
			Situation C			
a	0.025499	3.725208	3.247814	3.575660	3.045829	1.0
b	0.067700	4.294350	3.224890	3.113929	2.866590	1.1
c	0.584720	6.129308	3.047474	3.113929	2.866590	1.7
d	0.559320	3.750010	3.196117	3.113929	2.866590	1.2
e	0.953180	8.209517	2.980360	3.113929	2.866590	2.2
f	0.067700	4.113730	3.224860	3.423690	2.983650	1.1

From the Table 1 it can be seen that under all the three sampling situations A, B and C, the bias of OLS estimator is almost zero. Among other sampling designs, sample design (b) which is stratified sampling with proportional allocation is having minimum bias. The variance of $\hat{\beta}_{12}$ is minimum under sample design (d) in which the last stratum which is having large values of

design variable X_3 is given a greater allocation. Under sample design (a), $V(\beta_{12})$ is very high in comparison to other sample designs.

Nathan and Holt [4] have shown that if $Q = 1$, in which case the bias of b_{12} becomes of $O(n^{-1})$, then $V(b_{12}) \geq (\hat{\beta}_{12})$. This is true under all the three different situations. In situation C, when the dependent variable itself is used as the design variable, $V(\hat{\beta}_{12})$ is less than $MSE(b_{12})$ in all the three sample designs considered. Since for simple random sampling, sampling inclusion probability $\pi_\alpha = n/N$, the weighted and the unweighted estimators coincide i.e. $b_{12} = b^*_{12}$ and $\hat{\beta}_{12} - \hat{\beta}^*_{12}$, so $V(b_{12})$ and $V(b^*_{12})$, and $V(\hat{\beta}_{12})$ and $V(\hat{\beta}^*_{12})$ under all the 3 situations don't differ much from each other. Under the sample design (b) and (e) in which only few values are selected from last stratum, the weighted estimator comes out to be better than unweighted estimators.

Further weighted estimators seem relatively insensitive to the sample design. But since the weighted estimators are model free, they may be more robust to departures from the model upon which the properties of b_{12} and $\hat{\beta}_{12}$ are based. The results which hold for situation A extend to the situation B also, where double sampling has been adopted as a method of design for estimating regression coefficients when the design variable X_3 is already available from the survey or can be measured cheaply. Under the situation C, when X_1 is measured at the first phase and so the dependent variable is used as the design dependent variable itself, the bias for b_{12} is more in comparison to the situation A and B. But in this case also, the results of situation A are applicable.

REFERENCES

[1] DeMets, D. and Helperin, M., 1977. Estimation of a single regression coefficient in samples arising from a sub-sampling procedure. *Biometrics*, **33**, 47-56.

[2] Holt, D, Smith, T.M.F. and Winter, P.D., 1980. Regression analysis of data from complex surveys. *J.R. Statist. Soc. A*, **143**, 474-487.

[3] Kathuria, O.P., Bathala, H.V.L., Raheja, S.K., Gosh, K.K. and Mitra, P.M., 1984. Final Report of the Pilot Sample Survey for Estimation of Inland Fishery Resources and Catch in a Region of West Bengal. IASRI Publication, N. Delhi-12, India.

[4] Nathan, G. and Holt, D., 1980. The effect of survey design on regression analysis. *J.R. Statist. Soc. B*, **42**, 377-386.

[5] Pearson, 1902. On the Influence of Natural Selection on the variability and correlation of organs. *Phil. Trans. Roy. Soc., A*, **200**, 1-66.