REGRESSION ESTIMATOR AFTER A PRELIMINARY TEST OF SIGNIFICANCE FOR CORRELATION COEFFICIENT

By
S. S. Alam
Indian Institute of Technoloy, Kharagpur
(Received : September, 1974)

1. Introductio n

The implications of the test-estimation procedure for a linear regression model

$$Y=\beta_1 x_1+\beta_2 x_2$$

was first pointed out by Bancroft ([2] [3]). The bias due to the use of a preliminary t-test of significance of β_2 for the estimation of β_1, was studied. For the same model a mean square error (MSE) criterion has been proposed by Wallace [7] for making a choice between the ordinary least square estimator b_1 and the regression least square estimator $\hat{\beta}_1$ in the context of high intercorrelation between the regressor variables. The bias and the MSE expressions for the estimator of β_1 based on the MSE test to decide whether or not to include x_2, have been obtained by Toro [6]. Ashar [1] has pointed out that in case of serious collinearity, the usual least square estimator of β_1, although still unbiased, becomes less and less reliable, its variance fast approaching infinity as ρ^2 approaches one. The problem of estimation of the location parameter in the linear regression model

$$Y=\alpha+\beta x$$

by using a preliminary test of significance for β, has been considered by Saleh [5].

In this paper a sometimes-pool and sometimes-regression estimation procedure has been proposed for the estimation of the population mean of a variable which might be correlated with another variable. The proposed estimator is unbiased. The formula for the mean square error has been derived. The relative efficiency of the estimator to the ordinary regression estimator has been examined for a number of cases.

2. Estimation Procedure

Consider a first stage random sample (x_{1i}, Y_{1i} ; $i=1, 2, \ldots\ldots, n$) of size n on x and y which are jointly normally distributed with unknown means μ_x and μ_y, variances σ_x^2 and σ_y^2 respectively and correlation coefficient ρ. If the population correlation coefficient is not very small and the cost of observing y is higher than that of observing x, a second stage sample (x_{2j} ; $j=1, 2\ldots\ldots, n_1$) of size n_1, on x may be taken and the regression estimator

$$\hat{\mu}_r=\bar{y}_1+b(\bar{x}-\bar{x}_1) \qquad \ldots(2.1)$$

where

$$\bar{x}_1=\frac{1}{n}\sum_{i=1}^{n} x_{1i},\ \bar{y}=\frac{1}{n}\sum_{i=1}^{n} y_{1i},$$

$$\bar{x}_2=\frac{1}{n_1}\sum_{j=1}^{n_1} x_{2j},\ \bar{x}=\frac{n\bar{x}_1+n_1\bar{x}_2}{n+n_1},$$

$$s_x^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2,$$

$$s_y^2=\frac{1}{n-1}\sum_{i=1}^{n}(y_{1i}-\bar{y}_1)^2,$$

and

$$\gamma=\frac{\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)(y_{1i}-\bar{y}_1)}{\left[\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2\sum_{i=1}^{n}(y_{1i}-\bar{y}_1)^2\right]^{1/2}}$$

may be used for estimating μ_y. But, if x and y are not corrrelated then the regression estimator $\hat{\mu}_Y$ is not a desirable estimator. In this case a sample on y alone should be used for the estimation of μ_y. Again if it is suspected but not known for certain that $\rho=0$, neither the regression estimator $\hat{\mu}_\gamma$ nor an estimator based on a sample of y alone can be used indiscriminately. For such a situation a two-stage sometimes-pool and sometimes-regression estimator for μ_y may be used.

A first-stage sample (x_{1i}, y_{1i} ; $i=1,2,\ldots\ldots, n$) of size n on x and y is observed and a preliminary test for the null hypothesis H_o: ($\rho=0$) against the alternative H_1: ($\rho\neq0$) is performed with the critical region $\gamma^2 \geqslant \gamma_\alpha^2$, where γ is the sample correlation coefficient and γ_α is the upper ($\alpha/2$) 100% probability point of the distribution of γ when $\rho=0$. If the hypothesis H_o : ($\rho=0$) is rejected, a second stage sample (x_{2j} ; $j=1, 2,\ldots\ldots, n_1$) of size n_1 is observed on x alone and the regression estimator $\hat{\mu}_\gamma$ is used, otherwise a second stage sample (y_{2j} ; $j=1, 2, \ldots\ldots, n_2$) of size n_2 is observed on y alone and the pooled mean $\hat{\mu}_p$ of y, where

$$\hat{\mu}_p=\frac{n\bar{y}_1+n_2\bar{y}_2}{n+n_2} \qquad \ldots(2.2)$$

and

$$\bar{y}_2=\frac{1}{n_2}\sum_{j=1}^{n_2} y_{2j},$$

is used as the estimator of μ_y. Hence the sometimes-pool and sometimes-regression estimator for μ_y is

$$\hat{\mu}_y=\begin{cases}\hat{\mu}_r, & \text{if } \gamma^2 \geqslant \gamma_\alpha^2\\ \hat{\mu}_p, & \text{if } \gamma^2 < \gamma_\alpha^2\end{cases} \qquad \ldots(2.3)$$

The sometimes-pool and sometimes-regression estimator is unbiased. For,

$$\begin{aligned}E(\hat{\mu}_y)&=E(\hat{\mu}_r \mid \gamma^2\geqslant\gamma_\alpha^2)\,\rho+E(\hat{\mu}_p \mid \gamma^2<\gamma_\alpha^2)(1-\rho)\\&=\mu_y\,P+\mu_y(1-P)\\&=\mu_y\end{aligned}$$

where

$$P=\rho(\gamma^2 \geqslant \gamma_\alpha^2).$$

3. Mean Square Error

The mean square error MSE $(\hat{\mu}_y)$ of the estimator $\hat{\mu}_y$ is

$$\text{MSE}(\hat{\mu}_y) = E[\{(\bar{y}_1 - \mu_y) + b(\bar{x} - \bar{x}_1)\}^2 \mid \gamma^2 \geqslant \gamma_\alpha{}^2]\rho + [(\bar{y}_p - \mu_y)^2 \mid \gamma^2 < \gamma_\alpha{}^2]\ (1-\rho)$$

$$= \frac{\sigma_y{}^2}{n} P - 2m\rho\, \frac{\sigma_x\, \sigma_y}{n} E(b \mid \gamma^2 \geqslant \gamma_\alpha{}^2)\, P$$

$$+ \frac{m}{n}\, \sigma_x{}^2\, E(b^2 \mid \gamma^2 \geqslant \gamma_\alpha{}^2)\, P + \frac{\sigma_y{}^2}{n+n_2}(1-P) \quad \text{...(3.1)}$$

where

$$m = n_1/(n+n_1).$$

The conditional expectations required in equation (3.1) are obtained by considering the joint density function $f(\gamma, \nu)$ of the correlation coefficient γ and the variance ratio $\nu = s_y{}^2/s_x{}^2$ (cf. Kendall and Stuart, [4].) The joint density is given by

$$f(\gamma, \nu) = \frac{2^{n-3}(n-2)}{\pi} \left[R(1-\rho^2) \right]^{\frac{n-1}{2}} (1-\gamma^2)^{\frac{n-4}{2}}$$

$$\nu^{\frac{n-3}{2}} [\nu + R - 2\rho\gamma\sqrt{R\nu}]^{-(n-1)} \quad \text{...(3.2)}$$

where

$$R = \frac{\sigma_y{}^2}{\sigma^{x_2}}.$$

The mean square error MSE $(\hat{\mu}_y)$ is finally obtained as

$$\text{MSE}\ (\hat{\mu}_y) = \sigma_y{}^2 \left[\frac{P}{n} + \frac{(1-P)}{n+n_2} - \frac{m(1-\rho^2)^{\frac{n-1}{2}}}{n(n-3)} \right.$$

$$\left. \sum_{i=0}^{\infty} (a_i - b_i)\, I_{r_\alpha^2} \left(\frac{n}{2} - 1,\ i + \frac{3}{2} \right) \right] \quad \text{...(3,3)}$$

where

$$a_i = \frac{\rho^2(n+2i-3)}{2i}\, a_{i-1},$$

$$a_o = 1,$$

$$b_i = \frac{\rho^2(n+2i-5)\,(2i+1)}{(2i)\,(2i-1)}\, b_{i-1},$$

$$b_o = 1,$$

$$\bar{\gamma}_\alpha{}^2 = 1 - y_\alpha{}^2,$$

$$I_x(p, q) = \frac{1}{\beta(p, q)} \int_0^x t^{p-1}\,(1-t)^{q-1}\, dt,$$

$$\beta(p, q) = \int_0^1 t^{p-1}\,(1-t)^{q-1}\, dt.$$

The mean squre error MSE $(\hat{\mu}_Y)$ of the usual regression estimator $\hat{\mu}_Y$, without any preliminary test is

$$\text{MSE}(\hat{\mu}_Y) = \frac{\sigma_y{}^2}{n}\left[\, 1 - m\rho^2 - m(1-\rho^2)/(n-3) \,\right] \ldots \quad (3.4)$$

The performance of the proposed test-estimation procedure may be best examined by comparing its MSE with the MSE of the usual regression estimator, or rather computing the relative efficiency

$$e(\hat{\mu}_y) = \text{MSE}(\hat{\mu}_Y)/\text{MSE}(\hat{\mu}_y)$$

of the sometimes-pool and sometimes-regression estimator $\hat{\mu}_y$ to the usual regression estimator $\hat{\mu}_Y$ for different values of the parameters. To have a meaningful comparison we consider that both the estimators have the same first-stage sample size n and the second-stage size n_1 on x alone. The second-stage sample of size n_2 on y alone is chosen in such a way that costs of sampling foı both the estimators are equal.

Let C_{xy} be the cost of observing a pair (x, y), C_x be the cost of observing one unit of x alone and C_y be the cost of observing one unit of y alone. The cost of observing n pairs of (x, y) in the first-stage and n_1 of x in the second-stage for a regression estimator is

$$C_R = nC_{xy} + n_1 C_x.$$

Again for the sometimes-pool and sometimes-regression estimator the expected cost of observing the sample is

$$C = nC_{xy} + n_1 C_x \, P + n_2 \, C_y \, (1-P)$$
$$= nC_{xy} + n_2 C_y + P(n_1 C_x - n_2 C_y) \qquad .. (3.5)$$

where

$$P = P(\Upsilon^2 \geqslant \Upsilon_\alpha^2).$$

Thus to make C and C_R equal, let

$$n_1 C_x = n_2 C_y$$

or

$$n_2 = R_C \, n_1 \qquad ...(3.6)$$

where

$$R_C = C_x / C_y.$$

Hence for given n, n_1 and the cost ratio R_C, the second-stage sample size n_2 on y alone is obtained by using (3.6).

For given n, n_1 and the cost ratio R_C, the relative efficiency depends on the probability level of significance α, of the preliminary test and the population correlation coefficient ρ. Let the relative efficiency be denoted by $e(\alpha, \rho/n, n_1, R_C)$ or simply by e. It may be noted that the relative efficiencies at

$$\alpha = 0 \text{ and } \alpha = 1 \textit{ viz.}, \; e(0, \rho/n, n_1, R_C) \text{ and } e(1, \rho/n, n_1, R_C)$$

give respectively the relative efficiencies for the extreme cases *viz*, the always pool estimator $\hat{\mu}_p$ and the regression estimator $\hat{\mu}_Y$. Obviously $e(1, \rho/n, n_1 R_C) = 1$ for all values of ρ, n, n_1 and R_C. It may be further noted that

$$e(\alpha, \rho/n, n_1, R_C) = e(\alpha, -\rho/n, n_1, R_C)$$

Hence it is sufficient to study the relative efficiency for only positive values of ρ.

The relative efficiency has been computed for different combinations of n, n_1, R_C, α, ρ and $\sigma_y = 1$. The behaviour of the relative efficiency with respect to variation in ρ has been shown in figures 1 and 2 for two different situations.

For low values of α, the relative efficiency is maximum at $\rho=0$ having e greater than one and decreases, taking values less than one, as $|\rho|$ increases. For some moderate values of α, the efficiency is greater than one when $\rho=0$, increases slowly to a maximum value and then decreases taking values less than one as $|\rho|$ increases. Again for high values of α the efficiency curve is close to the line $e=1$ and intersect at some moderate value of $|\rho|$. Thus, for low values of $|\rho|$ the relative efficiency is always greater than one and it increases with the decrease in the probability level of significance α. For moderate values of $|\rho|$, particularly when R_C is very small, the relative efficiency is greater than one and decreases with the decrease in α. Thus there are situations (*e.g.*, $n=10$, $n_1=20$, $R_C=0.0$, $\alpha=.20$) where the proposed estimator may be preferred to either the regression or the always-pool estimator.

When the cost ratio $R_C=0$, the second-stage sample size has no remarkable effect on the efficiency. For larger values of the cost ratio R_C, the efficiency increases with the increase in the second-stage sample size n_1. But reverse is the case for the first-stage sample size n, the efficiency increases with the decrease in the first-stage sample size.

SUMMARY

On the basis of the outcome of a preliminary test for the significance of the correlation coefficient between two normal variables a sometimes-pool and sometimes-regression estimator has been proposed for the mean of the first variable which might be correlated with the second variable. The proposed estimator is unbiased. The formula for the mean square error has been derived and the relative efficiency of this estimator to the ordinary regression estimator has been examined for a number of cases.

ACKNOWLEDGEMENT

The author wishes to express his sincere thanks to Dr. G.P. Bhattacharjee for his constant help and guidance during the course of preparation of this paper.

References

[1] Ashar, V.G. (1968) : On the use of preliminary tests in analysis. *Presented at the annual meetings, Amer. Statist Assoc.*

[2] Bancroft, T.A. (1944) : On biases in estimation due to the use of preliminary tests of significance. *Ann. Math. Statist.* 15, 190–204.

[3] Bancrof, T.A. (1950) : Bias due to the omission of independent variables in ordinary multiple regression analysis (abstract), *Ann. Math. Statist.*, 21, 142.

[4] Kendall, M.G. and Stuart, A. (1963) : *The advanced theory of statistics*, Vol. I, Charles Griffin and Company Ltd., London.

[5] Saleh, A.K., Md Ehsanes (1971) : A class of estimates of location parameter after a preliminary test on regression (abstract). *Ann. Math. Statist.* 41, 1783-1784.

[6] Toro, C.E. (1968) : Multicollinearity and the mean square error criterion in multiple regression: A test and some sequential estimator comparisons. *Unpublished Ph. D. thesis, N.C. State University, Raleigh.*

[7] Wallace, T.D. (1964) : Efficiencies for stepwise regressions. *J. Amer. Statist. Assoc.* 59, 1179—1182.