The Economic Journal of Nepal Vol. 8, No. 2, Issue 30, April-June 1985 © EIC-TU

Nature of Bias and Precision in Regression Analysis

GOVINDA P. KOIRALA*

Introduction

When the name of the game *tug of war* is heard it is not unusual to jump quickly to imagine a strong rope being pulled by two groups of players–whichever group is stronger that pulls the rope towards it and gets the title of the winner. However, it is not the physically played game, a real tug of war, discussed here.

In a regression analysis one tries to explain the nature of a dependent variable the way it behaves when other explanatory variables change in terms of their magnitude and directions. But those explanatory variables that affect our dependent variable are usually unknown to any regression analyst. He usually lists a few variables that might seem to affect the dependent variable. He may list those variables seemingly having significant effect but actually they do not have or, he may ignore those variables (knowingly or unknowlingly) which have a significant effect on the dependent variable. Sometimes he can not include it in the analysis because the data are not available or because it is not quantifiable or whatever the reason is.

If a variable is known to be affected by a number of other variables, regression analyst tries to find out the nature of relationship and provide the estimates of the parameters of the relationship. The analyst draws upon a theory relating to the determination of the dependent variable and attempts to specify the independent variables. Suppose it is known that Y (a dependent variable) depends on k explanatory variables X_1, X_2,.....X_k; not all the coefficients of these variables can be estimated with any reasonable precision. Hence one has to develop a procedure to consider which variables to include and to exclude.

In doing this, it is possible to make two sorts of errors. Firstly, the analyst may fail to include an important variable in his model either by his ignorance or by overlooking the factor determining the dependent variable.

Secondly, he may think a variable important to be included in his model. Actually

* *Mr. Koirala is a lecturer in Statistics at Tribhuvan University, Kirtipur.*

it may just be an illustration and the variable has no significant effect in determining the dependent variable.

One of the error leads to biased estimate keeping the estimate under reasonable precision while the other leads to unbiased estimate but with reduced precision. And, thus, the tug of war develops between bais and precision.

Case of too many Variables

Consider the standard linear regression model,in which,suppose the true relationship (but unknown) is

$$Y = B_0 + \beta_1 X_1 + \beta_2 X_2 + U \quad \ldots \text{(i)}$$

In all the following discussions, unless otherwise mentioned, the error term, U is assumed to satisfy all the usual asumptions made for obtaining the best linear unbiased estimates while applying the least square techniques to estimate the parameters.

Taking (i) in deviation form we get:

$$Y = \beta_1 X_1 + \beta_2 X_2 + U \quad \ldots \text{(ii)}$$

However, by the ignorance or the unavailability of observations for X_2, the analyst estimates the following regression model:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \varepsilon \quad \ldots \text{(iii)}$$

In deviation from, $y = \hat{\beta}_1 x_1 + \varepsilon \quad \ldots \text{(iv)}$

In this case, the estimated error term has to capture all the influences made by the excluded variable X_2.

Here, β_1 is obtained by OLS technique, so that,

$$\hat{\beta}_1 = \frac{\sum x_1 y}{\sum x_1^2}$$

Since the truth is (ii), β_1 can be writen as

$$\beta_1 = \frac{\sum x_1(\beta_1 x_1 + \beta x_2 + u)}{\sum x_1^2}$$

$$= \beta_1 + \beta_2 \frac{\sum x_1 x}{\sum x_1^2} + \frac{\sum x_1 u}{\sum x_1^2} \quad \ldots \text{(v)}$$

Therefore,

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \frac{\sum x_1 x_2}{\sum x_1^2}$$ Since from the standard assumption cov(xy u) = 0

Thus, the bias is

$$E(\hat{\beta}_1) - \beta_1 = \beta_2 \frac{\sum x_1 x_2}{\sum x_1^2} \quad \ldots \text{(vi)}$$

This bias depends on two terms, the regression coefficient of the left-out variable in the true relation β_2 and the movements of the left-out variable with the included variable, $\left(\frac{\sum x_1 x_2}{\sum x_1^2}\right)$.

Thus the bias will not arise only when either of the two situations appears. First, when β_2 is zero, this indicates that X_2 is not an explanatory variable for Y, which, from our assumption of true relationship given by (ii), is not the case. Secondly, when the correlation between X_1 and X_2 is zero. This, in practical situations, generally does not occur.

The higher the correlation between X_1 and X_2, the greater will be the bias. Also, if X_1 and X_2 are correlated (may not be perfect, in which case, problem of multicollinearity appears and only one variable will be sufficient to describe the variability in Y), $\mathcal{E}$ and X_1 will also be correlated. This makes E ($\mathcal{E}$, X_2) $\neq$ O. This again is a violation of the standard assumpion. This makes biased and inconsistent estimates for both β_0 and β_1.

Case of Irrelevant Variable

Consider the standard linear regression model; in which the true relationship is

$$Y=\alpha_0+\alpha_1 X_1 +U \qquad \ldots (1)$$

Taking in deviation form, (1) reduces to $y=\alpha_1 x_1+u$ ... (2)

However, the analyst thinks X_2 as another variable to influence Y and includes in his analysis and estimates the following regression.

$$Y=\hat{\alpha}_0+\hat{\alpha}_1 X_1+\hat{\alpha}_2 X_2+\mathcal{E} \qquad \ldots (3)$$

In the deviation form,

$$Y=\hat{\alpha}_1 X_1=\hat{\alpha}_2 X_2+\mathcal{E} \qquad \ldots (4)$$

Here, $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are the least square estimates. Actually X_2 was not necessary to include. If this were the case $\hat{\alpha}_2$ should be obtained not significantly different from zero.

There is nothing wrong in having a regression coefficient whose value is zero. This will simply mean that the variable with zero regression coefficient would have no influence on the dependent variable.

Since $\hat{\alpha}_1$ and $\hat{\alpha}_2$ are the least square estimates, by the OLS technique $\hat{\alpha}_1$ comes to be:

$$\hat{\alpha}=\frac{\sum yx_1 \sum x_2^2-\sum yx_2 \sum x_1x_2}{\sum x_1^2 \sum x_2^2-\sum x_1x_2 \sum x_1x_2} \qquad \ldots (5)$$

and

$$\hat{\alpha}=\frac{\sum yx_2 \sum x_1^2-\sum yx_1 \sum x_1x_2}{\sum x_1^2 \sum x_2^2-\sum x_1x_2 \sum x_1x_2} \qquad \ldots (6)$$

Since the true relation is (2), $\hat{\alpha}_1$ can be written as,

$$\hat{\alpha}=\frac{\sum (\alpha_1x_1+u)x_1 \sum x_2^2-\sum (\alpha_1x_1+u)x_2 \sum x_1x_2}{\sum x_1^2 \sum x_2^2-\sum x_1x_2 \sum x_1x_2}$$

$$=\alpha_1+\frac{\sum ux_1 \sum x_2^2-\sum ux_2 \sum x_1x_2}{\sum x_1^2 \sum x_2^2-\sum x_1x_2 \sum x_1x_2} \qquad \ldots (7)$$

Therefore,

$E(\hat{\alpha_1}) = \alpha_1$; Since the standard assumption made for the error term is $\text{cov}(x_1, u) = \text{cov}(x_2, u) = 0$.

Thus, the bias $E(\hat{\alpha_1}) - \alpha_1 = 0$. ... (8)

Similarly, it can be shown that $E(\alpha_2) = 0$ ... (9)

The true value of the parameter α_2 is also zero. Because, for analysis, the true relation (2) can be written as :

$$y = \beta_1 x_1 + 0x_2 + u$$

Thus the bias, $E(\hat{\alpha_2}) - \alpha_2 = 0$. ... (10)

The excercise above confirms that we will get unbiased estimates of the parameters even if we include irrelevant variable in any regression analysis, through their zero valued parameter, then, the regression analyst may attempt to include as many variables in regression analysis as he thinks of. Because, irrelevant variables will automatically vanish through their zero valued parameters.

But, if we calculate the variance of the estimated parameter $\hat{\alpha}_1$ we get :

$$\text{var}(\hat{\alpha_1}) = E(\hat{\alpha_1} - E(\hat{\alpha_2})^2$$

$$= E(\hat{\alpha_1} - \alpha_1)^2$$

$$= E\left[\frac{\sum ux_1 \sum x_2^2 - \sum ux_2 \sum x_1 x_2}{\sum x_1^2 \sum x_2^2 - \sum x_1 x_2 . \sum x_1 x_2}\right], \text{ from (7).}$$

Expanding and, using the standard assumptions of OLS technique, we get :

$$\text{var}(\hat{\alpha_1}) = \frac{\sigma^2}{\sum x_1^2 (1-r^2)} \quad \ldots (11)$$

where σ^2 is variance of the error term and r correlation coefficient between X_1 & X_2.

If variance of $\bar{\alpha_1}$ were obtained from the true relation we would have obtained it as : $\text{var}(\bar{\alpha_1}) = \frac{\sigma^2}{\sum x_1^2}$ ... (12)

Comparing (11) and (12) and Since r^2 is always a positive fraction lying between O and 1; we get, $\text{var}(\hat{\alpha_1}) > \text{Var}(\bar{\alpha_1})$

This means that we will never get a more precise estimate by including an irrelevant variable in the regression analysis. As the correlation between X_1 and X_2 increases more and more, the higher and higher will be the loss in precision in the estimates.

An illustration

Consider a regression analysis made for acrerge response of price of sugarcane in Nepal. It seems very logical to think that the last year's price does have a significant effect

on pressent sugarcane growing area, and so does the last year's sugarcane grown area. To capture the technology changes, time factor also may be included in the model. Thus, the, following model is observed.

(A) $$X_t = 6269.32 + 2.34p_{t-1} + 0.49X_{t-1} + 618.31T_t \qquad R^2 = 0.9006$$
$$(1.32) \quad (0.098) \quad (1.505) \quad (1.176)$$

From the conventional test using t-statistics for the estimates (given in the paranthesis), one might jump to the conclusion that none of the mentioned variables had significant impact on acreage response while the model appears to be quite fit as suggested by the R^2 statistic. The values, high R^2 and insignificant 't', are suggesting to suspect the multicollinearity among the variables. Here the guilty variables seems to be the time trend. However, it is not. We have included it in the model, along with the price variable which in general are moving together. Hence, the real culprit is not the trend variable (T) but the multicollinearity. In the pressure of multicollinearity precise estimates cannot be expected.

Now, dropping time trend and estimating parameters, the following equation is obtained:

(B) $$X_t = 1554.72 + 21.52P_{t-1} + 0.82X_{t-1} \qquad R^2 = 0.8809$$
$$(0.65) \quad (1.22) \quad (4.69)$$

Looking at this relation with the conventional 't'-test on the back of his mind once again, the analyst might jump to conclude that the price is irrelevant in the model, and does not affect the sugarcane growing area for the next year. But again, this is not the case. Here also the real culprit is neither the price nor the previous year's area but it is the dominating relation between last year's area and this year's area.

Even though all the variables mentioned earlier affect the dependant variable, we can not include them in the analysis. Finally, when the dominating variable X_{t-1} (last year's sugarcane grown area) is dropped, the relation appears to be:

(C) $$X_t = 9575.35 + 78.08\ P_{t-1} \qquad R^2 = 0.5533$$
$$(3.133) \quad (3.34)$$

Conclusion

Above discussion left us in mist of confusion to assume any one as the real winner. Neither precision nor the bias is winner. It is up to the analyst to anounce the winning group for any particular situation. If the analyst is heading for the test of significance of the estimate itself, he should really think of precision in his favor but if he wants a forecast of the dependant variable, he should think of unbiased estimate of the parameters. This the real winner will be the analyst himself in this tug of war between bias and precision, whoever he wants he can announce the winner.

If we refer the illustration, we find that the bias and precision are fighting each

other. But we can now choose which way to go. If we are interested in the acreage response of price, relation (c) with significant t-value is preferred, and precision has become the focus. But if we want acreage to forecast for the future, relation (A) with high R^2 is preferred and unbiasedness becomes our target.

Selected References

Desai, Meghnad (1979) : *Applied Econometrics* (New Delhi : Tata-McGraw Hill Publishing Company Ltd)

Gujrati, Damodar (1978) : *Basic Econometrics* 2nd Ed. (Tokyo : Mc Graw Hill Book Company).

Johhston, J. (1978) : *Econometric Methods* (Tokyo : Mc Graw Hill-Kogakusha Ltd).

Kelegian, H. H. and Oates, W.E. (1974) : *Introduction to Econometrics-Principle and Application* (New York: Harper-Row Publishers).

Koutsoyiannis, A. (1978) : *Theory of Econometrics*, 2nd Ed. (London : The Macmillan Press Ltd.).

Maddala, C. S. (1977) : *Econometrics* (Tokyo : Mc Graw Hill—Kogakusha Ltd).

Pindyck, R. S. and Rubinfeld, D. L. (1976): *Econometric Model and Economic Forecasts* (Tokyo: Mc Graw Hill-Kogakusha Ltd).

Rao, Potlure and Miller, Roger Leroy (1972) : *Applied Econometrics* (New Delhi : Prentice-Hall of India Private Limited).

Wonnacott, R. J. and Wonnacott, T. H. (1970) : *Econometrics* (John Wiley and Sons. Inc.).