Some Notes on Misspecification in Multiple Regressions

POTLURI RAO*
University of Washington

In empirical research one often faces the problem of estimating a misspecified model. Misspecification can arise either because of omission of a variable specified by the truth, the case of the *left out variable*, or because of inclusion of a variable not specified by the truth, the case of the *irrelevant variable*. Misspecification is usually interpreted as a case of left out variables, and many researchers are concerned only with the bias resulting from it, the specification bias. Researchers seldom pay attention to the other aspects of misspecification. In particular, little note is made of the consequences of irrelevant variables, or of the effects of misspecification on the variance and mean square error of the regression estimates. This is mainly because these results are not readily available.[1] In view of the importance of these aspects of misspecification in empirical research, some major results of misspecification are presented in this paper with simple proofs.

We shall consider the classical linear regression model where all the independent variables are nonstochastic and the error terms are homoscedastic and serially independent. Let the two regression equations, of which only one is the truth, be:

$$y_t = \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_k x_{kt} + \eta_t \tag{1}$$

$$y_t = \alpha_1 x_{1t} + \alpha_2 x_{2t} + \cdots + \alpha_k x_{kt} + \alpha_{k+1} x_{k+1t} + \epsilon_t$$

$$t = 1, 2, \ldots, T. \tag{2}$$

When equation (1) is the truth, equation (2) is a misspecified model because of the presence of the irrelevant variable x_{k+1}. When equation (2) is the truth, equation (1) is a misspecified model because of the left out variable x_{k+1}. These two equations may be written in the matrix form as:

$$Y = X\beta + \eta \tag{3}$$

$$Y = \bar{X}\alpha + \epsilon \tag{4}$$

where Y is a vector of observations on the dependent variable, and X and $\bar{X}$ are matrices of independent variables in the equations (1) and (2) respectively. Without any loss of generality the equation (3) may be rewritten as:

$$Y = [XX_{k+1}]\begin{bmatrix}\beta\\ 0\end{bmatrix} + \eta \tag{5}$$

where X_{k+1} is a vector of observations on the independent variable x_{k+1}.

Theorem 1: *In the classical linear regression model, omission of a variable specified by the truth introduces bias in all the least squares estimates.*

This theorem is well known. (For example, see Griliches [1].) For consistency of notation we shall restate the proof. Let (2) be the truth, and let a misspecified model, given by equation (1), be estimated. The least squares estimates of the β's in equation (1) are given by:

$$\hat{\beta} = (X'X)^{-1}X'Y. \tag{6}$$

Since equation (2) is the truth we may rewrite equation (6) as:

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'(\bar{X}\alpha + \epsilon)\\ &= (X'X)^{-1}X'\bar{X}\alpha + (X'X)^{-1}X'\epsilon.\end{aligned} \tag{7}$$

Since all the independent variables are nonstochastic we have:

$$E(\hat{\beta}) = (X'X)^{-1}X'\bar{X}\alpha. \tag{8}$$

Equation (8) may be rewritten as:

$$\begin{aligned}E(\hat{\beta}) &= (X'X)^{-1}X'(XX_{k+1})\alpha\\ &= (\alpha_1\alpha_2 \ldots \alpha_k)' + \alpha_{k+1}(X'X)^{-1}X'X_{k+1}.\end{aligned} \tag{9}$$

To facilitate the interpretation of the results we shall introduce an auxiliary regression equation in Yule's notation as[2]:

$$x_{k+1} = b_{k+1,1.23\ldots k}x_1 + b_{k+1,2.13\ldots k}x_2 + \cdots + b_{k+1,k.12\ldots k-1}x_k + e_{k+1} \tag{10}$$

where the b's are the ordinary least squares estimates,[3] and e_{k+1} is the residual. The auxiliary regression equation (10) is introduced only as an algebraic convenience and need not have any causal interpretation.

The expected value of the regression coefficient of the independent variable x_1 may be written as:

$$E(\hat{\beta}_1) = \alpha_1 + \alpha_{k+1} \cdot b_{k+1,1.23\ldots k}. \tag{11}$$

The bias in the regression coefficient ($\hat{\beta}_1$) is proportional to the auxiliary regression coefficient $b_{k+1,1.23\ldots k}$. When

* The author is grateful to Professors Zvi Griliches and T. Dudley Wallace for their helpful comments.

[1] The problems of mean square error were investigated by Wallace [3], Toro and Wallace [2], and Wallace and Toro [4]. The implications of their results for misspecification problems failed to draw wide attention of the profession, perhaps because their analysis did not exploit its similarity with the well known misspecification results.

[2] In this notation the first subscript denotes the dependent variable, second subscript denotes the independent variable corresponding to the b, and the rest of the subscripts separated from the first two by a period (.) denote the other independent variables present in the regression equation. Since the values of b's change with the independent variables present in a regression all the subscripts are relevant in identifying the b's. For further details on the notation see Yule and Kendall [5], p. 284.

[3] Note that these b's are not used as statistical estimates of any parameters. They are used as algebraic equivalents of the expressions one would obtain if equation (10) were estimated by least squares.

the auxiliary regression coefficient is zero, then, of course, the bias is zero; but this case is rare in empirical work.[4] Since the choice of the first independent variable is arbitrary, the result in equation (11) holds for all the independent variables.

Theorem 2: *In the classical linear regression model, omission of a variable specified by the truth decreases the variance of all the least squares estimates.*

Let the truth be given by equation (2) and the misspecified model be equation (1), so that the left out variable is x_{k+1}. The least squares estimates of the β's are given by equation (6). The variance of the estimate vector $\hat{\beta}$ is:

$$V(\hat{\beta}) = E[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))'] = E[(X'X)^{-1}X'\epsilon\epsilon'X(X'X)^{-1}] = \sigma_\epsilon^2(X'X)^{-1} \tag{12}$$

Let us consider the variance of the regression coefficient corresponding to the independent variable x_1. The variance of the regression coefficients may be rewritten in the partitioned matrix form as:

$$V(\hat{\beta}) = \sigma_\epsilon^2 \begin{bmatrix} X_1'X_1 & X_1'Z \\ Z'X_1 & Z'Z \end{bmatrix}^{-1} \tag{13}$$

where X_1 is a vector of observations on the independent variable x_1, and Z is a matrix of observations on the rest of the independent variables in equation (1). Using the partitioned matrix inversion rule we have:

$$V(\hat{\beta}) = \sigma_\epsilon^2 \begin{bmatrix} (X_1'X_1 - X_1'Z(Z'Z)^{-1}Z'X_1)^{-1} & A \\ B & C \end{bmatrix} \tag{14}$$

where A and B are vectors, and C is a matrix. The variance of the regression coefficient $\hat{\beta}_1$ is therefore:

$$V(\hat{\beta}_1) = \sigma_\epsilon^2(X_1'X_1 - X_1'Z(Z'Z)^{-1}Z'X_1)^{-1} \tag{15}$$

To be able to interpret the expression given in equation (15), consider the following auxiliary regression equation:

$$x_1 = b_{12.34\ldots k}x_2 + b_{13.24\ldots k}x_3 + \cdots + b_{1k.23\ldots k-1}x_k + e_1 \tag{16}$$

where the b's are the least squares estimates, and e_1 is the residual. The residual sum of squares in the auxiliary regression equation (Σe_1^2) may be written in Yule's notation as:

$$S^2_{1.23\ldots k} = \Sigma e_1^2 = X_1'X_1 - X_1'Z(Z'Z)^{-1}Z'X_1. \tag{17}$$

Hence the variance of $\hat{\beta}_1$ may be rewritten as:

$$V(\hat{\beta}_1) = \sigma_\epsilon^2/S^2_{1.23\ldots k}. \tag{18}$$

It is well known that the variance of the least squares estimates of the true equation (2) are:

$$V(\hat{\alpha}) = \sigma_\epsilon^2(\bar{X}'\bar{X})^{-1}. \tag{19}$$

[4] In many practical situations the extent of bias may be smaller than the rounding error in truncating the decimal places. In empirical research what is relevant is the extent of bias and not its mere presence.

By using the above analysis we may write the variance of $\hat{\alpha}_1$ as:

$$V(\hat{\alpha}_1) = \sigma_\epsilon^2/S^2_{1.23\ldots k,k+1} \tag{20}$$

where $S^2_{1.23\ldots k,k+1}$ is the residual sum of squares in an auxiliary regression with x_1 as the dependent variable and $(x_2, \ldots, x_k, x_{k+1})$ as the independent variables. Since the residual sum of squares cannot increase by adding an independent variable to a regression equation, it follows that the residual sum of squares $S^2_{1.23\ldots k,k+1}$ cannot be larger than $S^2_{1.23\ldots k}$. Hence we have the inequality:

$$V(\hat{\beta}_1) \leq V(\hat{\alpha}_1). \tag{21}$$

This inequality becomes an equality only when the partial relation between the independent variable x_1 and the left out variable x_{k+1} holding $(x_2, \ldots, x_k)$ constant is zero; this case seldom arises in empirical work. Since the choice of x_1 is arbitrary, the inequality (21) holds for all coefficients.

Theorem 3: *In the classical linear regression model, discarding an independent variable whose parameter value is smaller (in magnitude) than the theoretical standard deviation of its estimate (from given data) will decrease the mean square error of all the least squares estimates.*[5]

We have shown that when equation (2) is the truth, and equation (1) is estimated, the regression coefficients are biased but have smaller variance when compared to the corresponding estimates from the true model. The Mean Square Error (MSE) may, however, increase or decrease depending on whether the gain in variance is compensated by the loss in the bias or not. The mean square errors in the two cases are:

$$\text{MSE}(\hat{\beta}_1) = \alpha^2_{k+1} \cdot b^2_{k+1,1.23\ldots k} + \sigma_\epsilon^2/S^2_{1.23\ldots k} \tag{22}$$

$$\text{MSE}(\hat{\alpha}_1) = \sigma_\epsilon^2/S^2_{1.23\ldots k,k+1}. \tag{23}$$

To simplify the algebra we shall use the following properties of least squares estimates due to Yule and Kendall [5][6]

$$S^2_{1.23\ldots k,k+1} = S^2_{1.23\ldots k}(1 - r^2_{1,k+1.23\ldots k}) \tag{24}$$

$$b_{k+1,1.23\ldots k} = r_{1,k+1.23\ldots k} \cdot \frac{S_{k+1.23\ldots k}}{S_{1.23\ldots k}} \tag{25}$$

where $r_{1,k+1.23\ldots k}$ is the partial correlation between the variables x_1 and x_{k+1} keeping all the other independent variables $(x_2, x_3, \ldots, x_k)$ constant.[7] Using the relations (24) and (25), we may rewrite the mean square errors of the estimates $\hat{\beta}_1$ and $\hat{\alpha}_1$ as:

$$\text{MSE}(\hat{\beta}_1) = \alpha^2_{k+1} \cdot r^2_{1,k+1.23\ldots k} \cdot \frac{S^2_{k+1.23\ldots k}}{S^2_{1.23\ldots k}} + \frac{\sigma_\epsilon^2}{S^2_{1.23\ldots k}} \tag{26}$$

$$\text{MSE}(\hat{\alpha}_1) = \frac{\sigma_\epsilon^2}{S^2_{1.23\ldots k}(1 - r^2_{1,k+1.23\ldots k})} \tag{27}$$

The conditions under which the mean square error of

[5] Wallace [3] proved this Theorem in a case of only two independent variables. His analysis does not lead to generalization.

[6] See Yule and Kendall [5], p. 287.

[7] Note that $r_{1,k+1.23\ldots k} = r_{k+1,1.23\ldots k}$

$\hat{\beta}_1$ is smaller than that of $\hat{\alpha}_1$ may be obtained as:

$$\alpha^2_{k+1} \cdot r^2_{1,k+1.23\ldots k} \cdot \frac{S^2_{k+1.23\ldots k}}{S^2_{1.23\ldots k}} + \frac{\sigma_\epsilon{}^2}{S^2_{1.23\ldots k}} \leq \frac{\sigma_\epsilon{}^2}{S^2_{1.23\ldots k}(1 - r^2_{1,k+1.23\ldots k})} \quad (28)$$

By rearrangement of terms:

$$\alpha^2_{k+1} \cdot r^2_{1,k+1.23\ldots k} \cdot \frac{S^2_{k+1.23\ldots k}}{S^2_{1.23\ldots k}} \leq \frac{\sigma_\epsilon{}^2}{S^2_{1.23\ldots k}} \left(\frac{1}{1 - r^2_{1,k+1.23\ldots k}} - 1 \right) \quad (29)$$

$$\alpha^2_{k+1} \cdot S^2_{k+1.23\ldots k}(1 - r^2_{1,k+1.23\ldots k}) \leq \sigma_\epsilon{}^2 \quad (30)$$

Since

$$S^2_{k+1.123\ldots k} = S^2_{k+1.23\ldots k}(1 - r^2_{1,k+1.23\ldots k}) \quad (31)$$

we may rewrite equation (30) as:

$$\alpha^2_{k+1} \leq \frac{\sigma_\epsilon{}^2}{S^2_{k+1.123\ldots k}} \quad (32)$$

But the variance of the estimate $\hat{\alpha}_{k+1}$ in the true equation (2) is given by

$$V(\hat{\alpha}_{k+1}) = \sigma_\epsilon{}^2 / S^2_{1+1.123\ldots k}. \quad (33)$$

Therefore the condition under which the mean square error of $\hat{\beta}_1$ is smaller than that of $\hat{\alpha}_1$ is given by:

$$\alpha^2_{k+1} \leq V(\hat{\alpha}_{k+1}) \quad (34)$$

or

$$|\alpha_{k+1}| \leq (V(\hat{\alpha}_{k+1}))^{1/2}. \quad (35)$$

That is, when the absolute value of the parameter α_{k+1} is smaller than the theoretical standard deviation of the estimate $\hat{\alpha}_{k+1}$, the mean square error of $\hat{\beta}_1$ would be decreased by discarding the variable x_{k+1}. Since the choice of the independent variable x_1 is arbitrary, the relation holds for all the independent variables.

Theorem 4: *In the classical linear regression model, inclusion of an irrelevant variable does not introduce bias in the least squares estimates.*

Let the truth be given by equation (1), and let the misspecified model, equation (2), with the irrelevant variable x_{k+1} be estimated. The least squares estimate of the misspecified model, equation (2), is given by:

$$\hat{\alpha} = (\bar{X}'\bar{X})^{-1}\bar{X}'Y. \quad (36)$$

Since equation (1), which may be written without loss of generality as equation (5), is the truth, we may rewrite equation (36) as:

$$\hat{\alpha} = (\bar{X}'\bar{X})^{-1}\bar{X}' \left(\bar{X} \begin{bmatrix} \beta \\ 0 \end{bmatrix} + \eta \right), \quad (37)$$

hence

$$E(\hat{\alpha}) = \begin{bmatrix} \beta \\ 0 \end{bmatrix}. \quad (38)$$

Estimation of the true equation (1) gives:

$$E(\hat{\beta}) = \beta. \quad (39)$$

Thus the estimates from the misspecified model (2), when equation (1) is the truth, are unbiased.

Theorem 5: *In the classical linear regression model, inclusion of an irrelevant variable increases the variance of all the least squares estimates.*

When equation (1) is the truth, and equation (2) is estimated, the variance of the least squares estimates, given by equation (37) is:

$$V(\hat{\alpha}) = \sigma_\eta{}^2(\bar{X}'\bar{X})^{-1}. \quad (40)$$

The variance of the least squares estimates when the true equation (1) is estimated is:

$$V(\hat{\beta}) = \sigma_\eta{}^2(X'X)^{-1}. \quad (41)$$

Using the analysis in the derivation of inequality (21), we can show that

$$V(\hat{\alpha}_1) \geq V(\hat{\beta}_1). \quad (42)$$

Even though an irrelevant variable does not introduce any bias in the regression coefficients, its presence increases the variance of all the regression coefficients.

Theorem 6: *In the classical linear regression model, inclusion of an irrelevant variable increases the mean square error of all the least squares estimates.*

Proof is obvious from theorems (4) and (5).

The results presented in this paper serve as theoretical guide lines in empirical research. Though these results do not provide practical rules on when to omit or include a variable on the basis of summary statistics, they provide some light on the consequences of omission or inclusion of a variable. For example, when a researcher is interested in using the regression estimates in decision making, he wants the least mean square error estimates rather than the best linear unbiased estimates. In such a case the researcher may "gain" by excluding a variable even though the truth specifies the variable as a part of the multiple regression. In some empirical studies researchers add variables to maximize R^2, or some other summary statistic, even though there are no theoretical reasons for their inclusion in the regression equation. The theoretical guide lines presented in this paper indicate that such a procedure may result in loss of "efficiency."

REFERENCES

[1] Griliches, Z., "Specification Bias in Estimates of Production Functions," *Journal of Farm Economics*, February 1957, 8–20.

[2] Toro-Vizcarrondo, C. and T. D. Wallace, "A Test of the Mean Square Error Criterion for Restrictions in Linear Regression," *Journal of the American Statistical Association*, June 1968, 558–72.

[3] Wallace, T. D., "Efficiencies for Stepwise Regressions," *Journal of the American Statistical Association*, December 1964, 1179–82.

[4] Wallace, T. D. and C. E. Toro-Vizcarrondo, "Tables for the Mean Square Error Test for Exact Linear Restrictions in Regression," *Journal of the American Statistical Association*, December 1969, 1649–63.

[5] Yule, G. U. and M. G. Kendall, *An Introduction to the Theory of Statistics*, Charles Griffin & Company limited (London), 1950, Fourteenth edition.