Specification errors in linear regression models [*]

Jean-Marie Dufour [†]
McGill University

First version: February 2002
Revised: December 2011
This version: December 2011
Compiled: December 9, 2011, 22:33

[*]This work was supported by the William Dow Chair in Political Economy (McGill University), the Bank of Canada (Research Fellowship), a Guggenheim Fellowship, a Konrad-Adenauer Fellowship (Alexander-von-Humboldt Foundation, Germany), the Canadian Network of Centres of Excellence [program on *Mathematics of Information Technology and Complex Systems* (MITACS)], the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, and the Fonds de recherche sur la société et la culture (Québec).

[†] William Dow Professor of Economics, McGill University, Centre interuniversitaire de recherche en analyse des organisations (CIRANO), and Centre interuniversitaire de recherche en économie quantitative (CIREQ). Mailing address: Department of Economics, McGill University, Leacock Building, Room 519, 855 Sherbrooke Street West, Montréal, Québec H3A 2T7, Canada. TEL: (1) 514 398 8879; FAX: (1) 514 398 4938; e-mail: jean-marie.dufour@mcgill.ca . Web page: http://www.jeanmariedufour.com

Contents

1. Classification of specification errors

The classical linear model is defined by the following assumptions.

1.1 Assumption $y = X\beta + \varepsilon$
where y is a $T \times 1$ vector of observations on a dependent variable ,
X is a $T \times k$ matrix of observations on explanatory variables,
β is a $k \times 1$ vector of fixed parameters,
ε is a $T \times 1$ vector of random disturbances.

1.2 Assumption $E(\varepsilon) = 0$.

1.3 Assumption $\mathsf{E}\left(\varepsilon\varepsilon'\right) = \sigma^2 I_T$.

1.4 Assumption *X is fixed (non-stochastic).*

1.5 Assumption $\mathrm{rank}\,(X) = k < T$.

To these, the assumption of error normality is often added.

1.6 Assumption *ε follows a multinormal distribution.*

An important problem in econometrics consists in studying what happens when one or several of these assumptions fail. Important failures include the following ones.

1. Incorrect regression – The linear regression model entails that

$$\mathsf{E}(y) = X\beta = x_1\beta_1 + x_2\beta_2 + \cdots + x_k\beta_k\,. \tag{1.1}$$

   Here we suppose that:
   a) $x_1, x_2, \ldots, x_k$ are the correct explanatory variables;
   b) the relationship is linear.
   Suppose now we estimate

$$y = Z\gamma + \varepsilon \tag{1.2}$$

   instead of (1.1). What are the consequences on estimation and statistical inference?

2. The error vector ε does not have mean zero:

$$\mathsf{E}(\varepsilon) \neq 0\,. \tag{1.3}$$

3. Incorrect covariance matrix – The covariance matrix of ε is not equal to $\sigma^2 I_T$:

$$\mathsf{E}\left[\varepsilon\varepsilon'\right] = \Omega \tag{1.4}$$

   where Ω is a positive semidefinite matrix.

4. Non-normality – The error vector ε does not follow a multinormal distribution.
5. Multicollinearity – The matrix X does not have full column rank:

$$\operatorname{rank}(X) < k\,. \tag{1.5}$$

6. Stochastic regressors – The matrix X is not fixed.

2. Incorrect regression function

2.1. The problem of incorrect regression function

Let us suppose the "true model" is:

$$y = X\beta + \varepsilon \tag{2.1}$$

where the assumptions of the classical linear model (**1.1**) - (**1.5**) are satisfied. However, we estimate instead the model

$$y = Z\gamma + \varepsilon \tag{2.2}$$

where Z is a $T \times G$ fixed matrix. Then the least squares estimator of γ is:

$$\begin{aligned} \hat{\gamma} &= \left(Z'Z\right)^{-1} Z'y \\ &= (Z'Z)^{-1}Z'(X\beta + \varepsilon) \\ &= (Z'Z)^{-1}Z'X\beta + (Z'Z)^{-1}Z'\varepsilon \end{aligned} \tag{2.3}$$

and the expected value of $\hat{\gamma}$ is

$$\mathsf{E}(\hat{\gamma}) = (Z'Z)^{-1}ZX\beta = P\beta \tag{2.4}$$

where $P = (Z'Z)^{-1}Z'X$. In general,

$$\mathsf{E}(\hat{\gamma}) \neq \beta \tag{2.5}$$

so that $\hat{\gamma}$ is not an unbiased estimator of β. Usual tests and confidence intervals are not valid.

2.2. Estimation of regression coefficients

2.2.1. The case of one missing explanatory variable

We will now study the case where a variable has been left out of the regression. Let

$$X \;=\left[X_1 \,\vdots\, x_k\right] \tag{2.1}$$

and

$$y = X_1\beta_1 + x_k\beta_k + \varepsilon \tag{2.2}$$

where X_1 is a $T \times (k-1)$ fixed matrix and x_k is a $T \times 1$ fixed vector. Instead of (2.2), we estimate:

$$y = X_1 \gamma + \varepsilon \,. \tag{2.3}$$

This corresponds to the case where $Z = X_1$ and

$$\begin{aligned} P &= (X_1'X_1)^{-1}X_1'X \\ &= (X_1'X_1)^{-1}X_1'\left[X_1 \vdots x_k\right] \\ &= \left[\left(X_1'X_1\right)^{-1}X_1'X_1 \vdots \left(X_1'X_1\right)^{-1}X_1'x_k\right] \\ &= \left[I_{k-1} \vdots \hat{\delta}_k\right] \end{aligned} \tag{2.4}$$

where

$$\hat{\delta}_k = (X_1'X_1)^{-1}X_1'x_k \,. \tag{2.5}$$

Note $\hat{\delta}_k$ is the regression coefficient vector obtained by regressing x_k on X_1 :

$$(x_k - X_1\hat{\delta}_k)'(x_k - X_1\hat{\delta}_k) = \min_{\delta_k}(x_k - X_1\delta_k)'(x_k - X_1\delta_k) \tag{2.6}$$

so $X_1\hat{\delta}_k$ provides the best linear approximation (in the least square sense) of the missing variable x_k based on the non-missing variables X_1. Thus

$$\begin{aligned} \mathsf{E}(\hat{\gamma}) &= P\beta && (2.7) \\ &= \left[I_{k-1} \vdots x_k\right]\begin{pmatrix} \beta_1 \\ \beta_k \end{pmatrix} && (2.8) \\ &= \beta_1 + \hat{\delta}_k\beta_k && (2.9) \end{aligned}$$

so the bias of $\hat{\gamma}$

$$\mathsf{E}(\hat{\gamma}) - \beta_1 = \hat{\delta}_k\beta_k \tag{2.10}$$

is determined by β_k and $\hat{\delta}_k$. We see easily that $\hat{\gamma}$ is unbiased for β_1 in two cases:

1. x_k does not belong to the regression: $\beta_k = 0$;
2. x_k is orthogonal with all the other regressors (every column of X_1): $X_1'x_k = 0$.

It is also interesting to look at the effect of excluding a regressor on

$$\hat{y} = X_1\hat{\gamma} \tag{2.11}$$

which can be interpreted as an estimator of estimator of $E(y)$, the mean of y. We have:

$$\begin{aligned} \mathsf{E}(\hat{y}) &= X_1\mathsf{E}(\hat{\gamma}) = X_1\left[\beta_1 + \hat{\delta}_k\beta_k\right] \\ &= X_1\beta_1 + (X_1\hat{\delta}_k)\beta_k. \end{aligned} \tag{2.12}$$

Since $\mathsf{E}(y) = X_1\beta_1 + x_k\beta_k$, it is clear that $\mathsf{E}(\hat{y}) \neq \mathsf{E}(y)$ even if $X_1'x_k = 0$, unless very special conditions hold. We have:

$$\begin{aligned} \mathsf{E}(\hat{y}) &= \mathsf{E}(y) \Leftrightarrow (X_1\hat{\delta}_k)\beta_k \qquad (2.13) \\ &\Leftrightarrow X_1\hat{\delta}_k = x_k \text{ or } \beta_k = 0\,. \qquad (2.14) \end{aligned}$$

In other words, $\mathsf{E}(\hat{y}) = \mathsf{E}(y)$ if and only if $\beta_k = 0$ (x_k is not a missing explanatory variable) or x_k is linear combination of the columns of X_1 (x_k is a redundant explanatory variable).

Even if $\hat{\gamma}$ is a biased estimator of β_1, it is possible that the mean squared error (MSE) of $\hat{\gamma}$ be smaller than the MSE of estimator $\hat{\beta}_1$ based on the complete model (2.2). The MSE of $\hat{\gamma}$ is

$$\begin{aligned} \mathsf{E}\big[(\hat{\gamma}-\beta_1)(\hat{\gamma}-\beta_1)'\big] &= \mathsf{V}(\hat{\gamma}) + [\mathsf{E}(\hat{\gamma})-\beta_1][\mathsf{E}(\hat{\gamma})-\beta_1]' \\ &= \sigma^2(X_1'X_1)^{-1} + (\hat{\delta}_k\beta_k)(\hat{\delta}_k\beta_k)' \\ &= \sigma^2(X_1'X_1)^{-1} + \beta_k^2\hat{\delta}_k\hat{\delta}_k' \qquad (2.15) \end{aligned}$$

while the MSE of $\hat{\beta}_1$ is

$$\begin{aligned} \mathsf{E}\big[(\hat{\beta}_1-\beta_1)(\hat{\beta}_1-\beta_1)'\big] &= \mathsf{V}(\hat{\beta}_1) + \Big[\mathsf{E}(\hat{\beta}_1)-\beta_1\Big]\Big[\mathsf{E}(\hat{\beta}_1)-\beta_1\Big]' \\ &= \mathsf{V}(\hat{\beta}_1) = \sigma^2(X_1'M_2X_1)^{-1} \qquad (2.16) \end{aligned}$$

where we use the fact that $\mathsf{E}(\hat{\beta}_1) = \beta_1$ and $M_2 = I_T - x_k(x_k'x_k)x_k'$. In general, either of these mean squared errors can be the smallest. More precisely, on observing that

$$\begin{aligned} X_1'X_1 - X_1'M_2X_1 &= X_1'(I-M_2)X_1 \\ &= X_1'(I-M_2)'(I-M_2)X_1 \\ &= [(I-M_2)X_1]'[(I-M_2)X_1] \qquad (2.17) \end{aligned}$$

is a positive semidefinite matrix, this entails that

$$(X_1'M_2X_1)^{-1} - (X_1'X_1)^{-1} \text{ is a positive semidefinite matrix} \qquad (2.18)$$

so that

$$\mathsf{V}(\hat{\beta}_1) - \mathsf{E}\big[(\hat{\beta}_1-\beta_1)(\hat{\beta}_1-\beta_1)'\big] = \big[\sigma^2(X_1'M_2X_1)^{-1} - \sigma^2(X_1'X_1)^{-1}\big] - \beta_k^2\hat{\delta}_k\hat{\delta}_k' \qquad (2.19)$$

is the difference between two positive semidefinite matrices, which may be positive semidefinite, negative semidefinite, or not definite at all.

2.2.2. Estimation of the mean from a misspecified regression model

Consider now the general case where Z is different X :

$$\mathsf{E}(\hat{\gamma}) = (Z'Z)^{-1}ZX\beta = P\beta \qquad (2.20)$$

where

$$\begin{aligned}
P &= (Z'Z)^{-1}Z'X \\
&= (Z'Z)^{-1}Z'[x_1, \ldots, x_k] \\
&= [\left(Z'Z\right)^{-1}Z'x_1, \ldots, (Z'Z)^{-1}Z'x_k] \\
&= \left[\hat{\delta}_1, \ldots, \hat{\delta}_k\right] \qquad (2.21)
\end{aligned}$$

and

$$\hat{\delta}_i = \left(Z'Z\right)^{-1}Z'x_i\,,\; i = 1, \ldots, k\,. \qquad (2.22)$$

The fitted values for the misspecified models

$$\hat{y} = Z\hat{\gamma} \qquad (2.23)$$

can be interpreted as estimators of $\mathsf{E}(y)$, and

$$\begin{aligned}
\mathsf{E}(\hat{y}) &= ZE(\hat{\gamma}) = Z\left[\hat{\delta}_1, \ldots, \delta_k\right]\beta \\
&= \left[Z\hat{\delta}_1, \ldots, Z\hat{\delta}_k\right]\beta \\
&= [\hat{x}_1, \ldots, \hat{x}_k]\,\beta = \hat{X}\beta \\
&= Z(Z'Z)^{-1}Z'X\beta \qquad (2.24)
\end{aligned}$$

where $\hat{x}_i = Z\hat{\delta}_i$ is a linear approximation of x_i based on Z. In general $\mathsf{E}(\hat{y}) \neq \mathsf{E}(y)$, but its MSE can be smaller than the one of $X\hat{\beta}$ based on the correctly specified model (2.2).

2.3. Estimation of the error variance

It is of interest to compare the estimators of the error variance derived from the misspecified and correctly specified models. The "unbiased" estimator of σ^2 based on model (2.2) is

$$s_Z^2 = y'M_Zy \big/ (T - G) \qquad (2.25)$$

where $M_Z = I_T - Z(Z'Z)^{-1}Z'$.Using

$$y = X\beta + \varepsilon \qquad (2.26)$$

we see that

$$\begin{aligned}
y'M_Zy &= (X\beta + \varepsilon)'M_Z(X\beta + \varepsilon) \\
&= \beta'X'M_ZX\beta + \varepsilon'M_Z\varepsilon + 2\beta'X'M_Z\varepsilon \qquad (2.27)
\end{aligned}$$

hence

$$\begin{aligned}
\mathsf{E}(y'M_Zy) &= \beta'X'M_ZX\beta + \mathsf{E}(\varepsilon'M_Z\varepsilon) \\
&= \beta'X'M_ZX\beta + \mathsf{E}\left[\mathrm{tr}(\varepsilon'M_Z\varepsilon)\right]
\end{aligned}$$

$$
\begin{aligned}
&= \beta' X' M_Z X \beta + \mathsf{E}\left[\operatorname{tr}(M_Z \varepsilon \varepsilon')\right] \\
&= \beta' X' M_Z X \beta + \operatorname{tr}[M_Z \mathsf{E}(\varepsilon \varepsilon')] \\
&= \beta' X' M_Z X \beta + \sigma^2 \operatorname{tr}(M_Z) \\
&= \beta' X' M_Z X \beta + \sigma^2 (T - G) \qquad (2.28)
\end{aligned}
$$

and

$$
\mathsf{E}(s_Z^2) = \sigma^2 + \frac{1}{T-G} \beta' X' M_Z X \beta \geq \sigma^2 = \mathsf{E}(s_X^2) \qquad (2.29)
$$

where

$$
s_X^2 = \frac{1}{T-k} y' M_X y \,. \qquad (2.30)
$$

If we compare two linear regression models, one of which is the correct one, the expected value of the "unbiased estimator" of the error variance is never the largest for the correct model. This provides a justification for selecting the model which yields the smallest estimated error variance (or the largest $\bar{R}^2$).

3. Error mean different from zero

In the model

$$
y = X\beta + \varepsilon \qquad (3.31)
$$

consider now the case where the assumption $\mathsf{E}(\varepsilon)$ is relaxed:

$$
\mathsf{E}(\varepsilon) = \xi \,, \qquad (3.32)
$$

$$
\mathsf{V}(\varepsilon) = \sigma^2 I_T = \mathsf{E}\left[(\varepsilon - \xi)(\varepsilon - \xi)' \right]. \qquad (3.33)
$$

Then

$$
\begin{aligned}
\mathsf{E}(\hat{\beta}) &= \mathsf{E}[\beta + (X'X)^{-1} X' \varepsilon] \\
&= \beta + (X'X)^{-1} X' \xi \qquad (3.34)
\end{aligned}
$$

and we see that $\hat{\beta}$ is not anymore an unbiased estimator (unless $X'\xi = 0$).

Suppose now that the model contains a constant term, i.e.

$$
X = [i_T, x_2, \ldots, x_k] \qquad (3.35)
$$

where $i_T = (1, 1, \ldots, 1)'$, and the components of ε all have the same mean μ :

$$
\xi = \mu i_T \,. \qquad (3.36)
$$

We can then set the mean to zero by defining

$$
v = \varepsilon - \mu i_T \qquad (3.37)
$$

so that $\mathsf{E}(v) = 0$, and rewrite the model as follows:

$$\begin{aligned} y &= i_T\beta_1 + x_2\beta_2 + \cdots + x_k\beta_k + \varepsilon \\ &= i_T\beta_1 + x_2\beta_2 + \cdots + x_k\beta_k + \mu i_T + v \\ &= i_T\bar{\beta}_1 + x_2\beta_2 + \cdots + x_k\beta_k + v \end{aligned} \tag{3.38}$$

where $\bar{\beta}_1 = \beta_1 + \mu$. In this reparameterized model, the non-zero mean problem has disappeared. The only difference with the standard case is that the interpretation of the constant term has been modified.

This type of reformulation is not typically possible when if the model does not have a constant. This suggests it is usually a bad idea not to include a constant in a linear regression (unless very theoretical reasons are available).

4. Incorrect error covariance matrix

The assumption

$$\mathsf{E}\big(\varepsilon\varepsilon'\big) = \sigma^2 I_T \tag{4.39}$$

is typically quite restrictive and is not satisfied in many econometric models. Suppose instead that

$$\mathsf{V}(\varepsilon) = \Omega \tag{4.40}$$

where Ω is a positive definite matrix.

It is then easy to see that the estimator $\hat{\beta}$ remains unbiased:

$$\begin{aligned} \mathsf{E}(\hat{\beta}) &= \mathsf{E}[\beta + (X'X)^{-1}X'\varepsilon] \\ &= \beta + (X'X)^{-1}X'\mathsf{E}(\varepsilon) = \beta\,. \end{aligned} \tag{4.41}$$

However, the covariance matrix of $\hat{\beta}$ is modified:

$$\begin{aligned} \mathsf{V}(\hat{\beta}) &= \mathsf{E}\big[\big(X'X\big)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}\big] \\ &= \big(X'X\big)^{-1}X'\mathsf{E}(\varepsilon\varepsilon')X(X'X)^{-1} \\ &= (X'X)^{-1}X'\Omega X(X'X)^{-1} \neq \sigma^2(X'X)^{-1}\,. \end{aligned} \tag{4.42}$$

This entails that $\hat{\beta}$ is not anymore the best linear unbiased estimator of β. Further, usual formulas for computing standard errors and tests are not valid anymore.

5. Stochastic regressors

The assumption that X is fixed is also quite restrictive and implausible in most econometric regressions. However, if ε is independent of X, most results obtained from the classical linear model remain valid. This comes form the fact these hold conditionally on X.We have in this case

$$\begin{aligned} \mathsf{E}(\varepsilon|X) &= \mathsf{E}(\varepsilon) = 0\,, && (5.43)\\ \mathsf{V}(\varepsilon|X) &= \mathsf{V}(\varepsilon) = \sigma^2 I_T && (5.44) \end{aligned}$$

hence

$$\begin{aligned} \mathsf{E}(\hat{\beta}) &= \beta + \mathsf{E}\left[(X'X)^{-1}X'\varepsilon\right] \\ &= \beta + \mathsf{E}\big[\mathsf{E}\left[(X'X)^{-1}X'\varepsilon|X\right]\big] \\ &= \beta + \mathsf{E}\big[(X'X)^{-1}X'\mathsf{E}(\varepsilon|X)\big] = 0\,. && (5.45) \end{aligned}$$

Similarly usual tests and confidence intervals remain valid (under the assumption of Gaussian errors).

However, if

$$\mathsf{E}(\varepsilon|X) \neq 0 \tag{5.46}$$

we can only write

$$\begin{aligned} \mathsf{E}(\hat{\beta}) &= \beta + \mathsf{E}\left[(X'X)^{-1}X'\varepsilon\right] \\ &= \beta + \mathsf{E}\big[\mathsf{E}\left[(X'X)^{-1}X'\varepsilon|X\right]\big] \\ &= \beta + \mathsf{E}\big[(X'X)^{-1}X'\mathsf{E}(\varepsilon|X)\big] && (5.47) \end{aligned}$$

and $\hat{\beta}$ is typically biased. The Gauss-Markov theorem as well as usual tests and confidence intervals are not typically valid.

Bibliography

Aigner, D. J.(1974). MSE dominance of least squares with errors of observations, J. of Econometrics 2, 365-372. [MSE with proxy variables.]

Dhrymes,P. J. (1978). Introductory Econometrics (Springer-Verlag, N.Y.) [General sections on specification errors and errors in variables.]

Goldberger, A. S. (1961). Stepwise least squares: residual analysis and specification errors, JASA, 998-1000. [MSE for an estimator with a missing variable.]

Judge, G. G., and M. F. Bock (1978). Statistical Implications of Pre-test and Stein-rule Estimators in Econometrics (North-Holland, Amsterdam) [MSE of an estimator under constraint.]

Judge, G. G. et al (1980). The Theory and Practice of Econometrics (Wiley, N.Y.) Ch. 11, 13.2 [MSE of estimators under constraints. Discussion of proxy variables.]

Levi, M. D. (1977). Measurement errors and bounded OLS estimates, J. of Econometrics 6, 165-171.

Levi, M. D. (1973). Errors in the variables bias in the presence of correctly measured variables, Econometrica 41, 985-986.

Maddala, G. S. (1977). Econometrics (McGraw Hill, N.Y.) 155-162, 461-462, 467-468. [Discussion of proxy and missing variables.]

McCallum, B. T. (1972). Relative asymptotic bias from errors of omission and measurement, Econometrica 40, 757-758. [Discussion of proxy variables.]

Theil, H. (1971) Principles of Econometrics (Wiley, N.Y.) 540-556.

Theil, H. (1957) Specification errors and the estimation of economic relationships, Rev. of the Int. Stat. 25, 41-51. [Theorem on the bias of estimators with a missing variable.]

Wallace, T. D. (1969) Efficiencies for stepwise regression, JASA, 1179-1182. [MSE of estimators with missing variables.]

Wickens, M. R. (1972) A note on the use of proxy variables, Econometrica 40, 759-761.