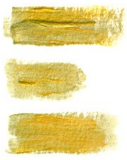
EIEF WORKING PAPER SERIES



EINAUDI INSTITUTE FOR ECONOMICS AND FINANCE

EIEF Working Paper 05/15
May 2015

On the ambiguous consequences of omitting variables

by
Giuseppe De Luca
(University of Palermo)
Jan R. Magnus
(VU University Amsterdam
and Tinbergen Institute)
Franco Peracchi
(University of Rome "Tor Vergata" and EIEF)

On the ambiguous consequences of omitting variables*

Giuseppe De Luca
University of Palermo, Italy

Jan R. Magnus
VU University Amsterdam and Tinbergen Institute, The Netherlands

Franco Peracchi
University of Rome Tor Vergata and EIEF, Italy

May 21, 2015

* We are grateful to Eveline de Jong for providing the example in the introduction and to Ed Leamer for useful suggestions.

Abstract: This paper studies what happens when we move from a short regression to a long regression (or vice versa), when the long regression is shorter than the data-generation process. In the special case where the long regression equals the data-generation process, the least-squares estimators have smaller bias (in fact zero bias) but larger variances in the long regression than in the short regression. But if the long regression is also misspecified, the bias may not be smaller. We provide bias and mean squared error comparisons and study the dependence of the differences on the misspecification parameter.

Keywords: Omitted variables; Misspecification; Least-squares estimators; Bias; Mean squared error

JEL Classification: C13; C51; C52

Short title: Omitting variables

Corresponding author:
Giuseppe De Luca
University of Palermo
Department SEAS
Palermo
Italy
E-mail: giuseppe.deluca@unipa.it

1 Introduction

Ludwig van Beethoven composed nine symphonies. Suppose a tenth symphony is discovered. There is no full score, only three parts are available: first violin, cello, and clarinet. This version is recorded and creates a big hit. Of course everybody realizes that many instruments are missing — still, it seems one gets a good idea of Beethoven's tenth. Now the trumpet part is discovered and a new recording is made. The new recording is received less enthusiastically than the first recording and music experts claim that adding the trumpet moves us *away* from how the real symphony should sound.

This creates a puzzle and a debate among scientists of various disciplines. How is it possible that getting closer to the true instrumentation does not get us closer to the true sound? Of course, adding *all* instruments to the score creates the true sound, but it seems that adding only *some* of them may not lead to an improvement. An addition in itself is not necessarily an improvement, it must be a 'balanced addition'.

What does this mean: a 'balanced addition'? The current paper contains our attempt to answer this question. We do so in the context of the standard linear regression model and omitted variables but, given the connection between omitted variables and many other forms of misspecification, our analysis extends to a variety of problems, such as the choice of suitable functional forms (polynomial terms, interactions, lag lengths, etc.), errors in variables, simultaneity, unobserved heterogeneity, censoring, and sample selection.

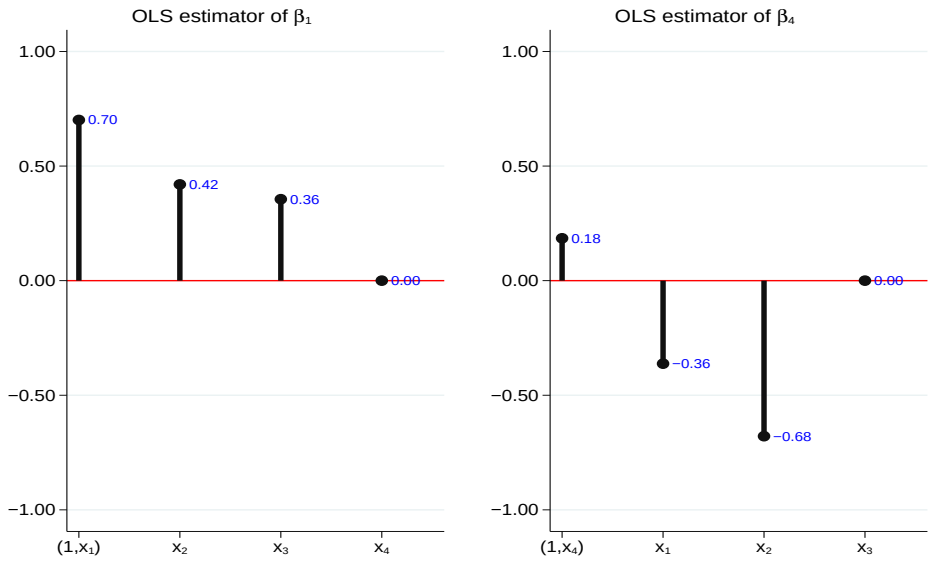
To illustrate the issue, consider a data-generation process (DGP) containing a constant term and four regressors:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon. \tag{1}$$

Our interest is in estimating β_1, and we are particularly concerned with estimation bias. In our first model, the only regressors are the constant term and x_1. In the second model we add x_2, in the third we add x_3, and finally, in the fourth model, we add x_4. These four models give us four different ordinary least-squares (OLS) estimators of β_1, each with its own bias, and we know that the bias in the last model equals zero. The left panel of Figure 1 shows that the bias in estimating β_1 decreases monotonically to zero as more regressors are sequentially added to the basic model. (If we change the order in which x_2, x_3, and x_4 are added to the basic model, then this is still true in this case.) It would therefore seem that adding more regressors (getting closer to the truth) always decreases the bias in estimating β_1.

But now consider estimating β_4. We start with the constant term and x_4 as the only regressors (our basic model) and then we sequentially add

Figure 1: Bias of the OLS estimators of β_1 and β_4 by adding regressors



x_1, x_2, and x_3. In the right panel of Figure 1 the bias no longer decreases monotonically to zero, and again this result does not depend on the order in which x_1, x_2, and x_3 are added to the basic model. Apparently, adding variables to our model does not necessarily decrease the size of the bias even when these variables belong to the DGP.

This simple fact is not mentioned in textbooks, at least not in the textbooks we consulted (e.g. Maddala 1992; Davidson and MacKinnon 2004; Cameron and Trivedi 2005; Angrist and Pischke 2009, 2015; Greene 2011; Wooldridge 2012). The usual story is that the 'long' regression (where model and DGP coincide) yields unbiased estimators, and that the 'short' regression (where one or more of the relevant regressors are omitted) yields biased estimators. The size of this 'omitted variable bias' depends on the size of the parameters associated with the omitted regressors and the correlation between included and omitted regressors, so that it will be small if and only if the omitted regressors are either relatively 'unimportant' (i.e. their parameters are relatively small) or almost uncorrelated with the included regressors. As this bias does not vanish asymptotically, the OLS estimator from the short regression is also inconsistent.

The implicit message from the textbook analysis is that adding variables to the model always decreases the bias of the OLS estimator of interest. In

finite sample setups with fixed regressors and homoskedastic errors, the inclusion of additional variables necessarily increases the sampling variance, giving rise to a bias-precision trade-off. Since this increase in variance does *not* depend on the size of the omitted parameters, it is advantageous to delete 'unimportant' regressors, even when we know for certain that they belong to the DGP, because the small increase in bias will be more than offset by the decrease in variance. (In the case of either stochastic regressors or heteroskedastic errors, the conclusions are more nuanced because the inclusion of additional variables may also decrease the sampling variance.) This trade-off is a typical finite-sample problem. In large samples the bias dominates the variance, so it is advisable to avoid misspecification at all cost; see, for example, Davidson and MacKinnon (2004, p. 116).

This is the textbook story and it is correct, but only if we compare the smaller model with the full DGP, not if we compare a small model with a larger model which is still smaller than the DGP, as demonstrated by Figure 1. Since, in practice, *any* model is likely to be smaller than the DGP, we can never be certain that the bias of our OLS estimator decreases when we add more variables.

Kevin Clarke seems to have been the first to analyze this somewhat counterintuitive situation, but his 2005 paper was published in a journal not typically read by econometricians and it went largely unnoticed. Clarke criticized the use of 'bloated specifications' based on the 'key underlying assumption [...] that the danger posed by omitted variable bias can be ameliorated by the inclusion of relevant control variables' when, in fact, 'the inclusion of additional control variables may increase or decrease the bias, and we cannot know for sure which is the case in any particular situation'. Although important, his study relies on a simplified DGP, does not provide analytical conditions to interpret bias comparisons of OLS estimators from models with different sets of regressors, and his conclusions about the 'phantom menace' are based on the results of a simple Monte Carlo experiment.

The aim of this paper is to analyze the issue in greater detail and discuss its consequences. Unlike Clarke (2005), we provide analytical conditions on bias and mean squared error (MSE) comparisons of OLS estimators from models with different sets of regressors in a setting where 'long' and 'short' models are both subject to general forms of misspecification. In addition, we analyze properties of the residuals, the OLS estimators of the error variance, and the usual F-test for misspecification.

The plan of the paper is as follows. In Section 2 we present the setup. The bias and mean squared error of the estimators are presented and compared in Sections 3 and 4. We discuss residuals, the estimation of the error variance, and the usual F-test for misspecification in Section 5. Section 6 concludes. A

data appendix contains the data underlying the little experiment presented above and a mathematical appendix provides proofs of our propositions.

2 Setup

Our data are assumed to be generated by the process

$$y = X_1\beta_1 + X_2\beta_2 + \delta + \epsilon, \tag{2}$$

where y is the $n \times 1$ vector containing the observations on the outcome of interest, X_1 $(n \times k_1)$ and X_2 $(n \times k_2)$ are matrices of regressors, β_1 and β_2 are unknown parameter vectors, δ is a vector representing misspecification, and ϵ is a vector of random disturbances. For example, if the assumed model for y includes X_1 and X_2 but omits a set of relevant regressors X_3, then $\delta = X_3\beta_3$. As another example, if $y = X_1\beta_1 + X_2^*\beta_2 + \epsilon$, but X_2^* is unobservable and we only have available a set $X_2 = X_2^* + U$ of 'proxy' variables (McCallum 1972), then $\delta = -U\beta_2$.

We let $k_1 \geq 1$, $k_2 \geq 1$, and $k = k_1 + k_2 < n$, and assume that the matrix $X = (X_1 : X_2)$ has full column-rank k. We explicitly exclude the trivial case where $X_1'X_2 = 0$. For simplicity we also assume that the regressors and the misspecification vector δ are all nonrandom. In this simple setting, the randomness in y is caused exclusively by ϵ, which has mean and variance

$$\mathrm{E}(\epsilon) = 0, \qquad \mathrm{var}(\epsilon) = \sigma^2 I_n, \tag{3}$$

respectively. Equations (2) and (3) define the DGP in the case of fixed regressors. The specification of the DGP could be extended to cover the case of stochastic regressors and heteroskedastic errors. Such generalizations are ignored here because they would not change the biases of our OLS estimators of β_1 and β_2. However, as discussed in Section 4, they could affect their sampling variances and thus the MSE comparisons.

Since the DGP is not known, δ is excluded from any model used for estimation purposes. We consider two models:

$$y = X_1\beta_1 + \epsilon, \tag{4}$$

and

$$y = X_1\beta_1 + X_2\beta_2 + \epsilon, \tag{5}$$

which we call the 'short' and the 'long' model, respectively. If δ is zero, then the long model coincides with the DGP. This is the textbook case. If δ is not zero, then both models are underspecified, as in the so-called $\mathcal{M}$-open perspective adopted in the Bayesian literature on model selection and

model averaging; see, for example, Bernardo and Smith (1994), Hoeting et al. (1999), and Clyde and Iversen (2013). If all or some components of β_2 are zero, then the short model includes parameters (namely the zero components of β_2) which are absent from the DGP, so it is underspecified and overspecified at the same time.

Letting $M_1 = I_n - X_1(X_1'X_1)^{-1}X_1'$, the usual symmetric idempotent matrix of rank $n - k_1$, and defining

$$A = X_1(X_1'X_1)^{-1}, \qquad B = M_1X_2(X_2'M_1X_2)^{-1}, \tag{6}$$

we can write the restricted OLS estimator of β_1 in the short model (4) as

$$\hat{\beta}_{1r} = A'y, \tag{7}$$

and the unrestricted OLS estimators of β_1 and β_2 in the long model (5) as

$$\hat{\beta}_{1u} = A'y - A'X_2B'y, \qquad \hat{\beta}_{2u} = B'y. \tag{8}$$

Note that we have excluded the case where $X_1'X_2 = 0$. The reason is now clear: if $X_1'X_2 = 0$ then $\hat{\beta}_{1r} = \hat{\beta}_{1u}$, and a comparison is meaningless.

In the next two sections we shall compare the bias, variance, and MSE of these two estimators of β_1, in particular their dependence on δ. We shall say that X_2 represents a 'balanced addition' to X_1 if either the bias or the MSE of the unrestricted estimator of β_1 is smaller than the bias or the MSE of the restricted estimator.

3 Bias

Since the bias dominates the variance in sufficiently large samples, we first consider the two biases $b_r = \mathrm{E}(\hat{\beta}_{1r} - \beta_1)$ and $b_u = \mathrm{E}(\hat{\beta}_{1u} - \beta_1)$. Their difference $b_r - b_u$ plays an important role in econometrics, as it helps understand the relationship between estimators in models with different sets of control variables (Angrist and Pischke 2009, 2015), and represents the basis for a variety of specification tests (see, e.g., Hausman 1978).

PROPOSITION **3.1** *Under the DGP given by (2) and (3), the biases of the restricted estimator $\hat{\beta}_{1r}$ and the unrestricted estimator $\hat{\beta}_{1u}$ are*

$$b_r = A'\delta + A'X_2\beta_2, \qquad b_u = A'\delta - A'X_2B'\delta,$$

respectively.

Except for a few special cases, it is not clear *a priori* which of the two biases is larger. In the textbook case, where $\delta = 0$, we have

$$b_r = A'X_2\beta_2, \qquad b_u = 0,$$

so the unrestricted estimator $\hat{\beta}_{1u}$ is unbiased, while the restricted estimator is unbiased only if β_2 lies in the null space of the matrix $X_1'X_2$.

A second special case arises when $\delta \neq 0$ and X_1 is orthogonal to the misspecification vector $\zeta = X_2\beta_2 + \delta$ in the short model (4). Then,

$$b_r = 0, \qquad b_u = -A'X_2B'\zeta,$$

so the restricted estimator is unbiased, while the unrestricted estimator is not.

A third example is the 'proxy' setup of McCallum (1972), where $X_2 = X_2^* + U$ and $\delta = -U\beta_2$. In this case

$$b_r = A'(X_2 - U)\beta_2, \qquad b_u = A'(X_2B' - I_n)U\beta_2.$$

After imposing the orthogonality restrictions $X_1'U = 0$ and $X_2^{*'}U = 0$, the two biases become $b_r = A'X_2^*\beta_2$ and $b_u = A'X_2^*C\beta_2$, where

$$C = (X_2^{*\prime}M_1X_2^* + U'U)^{-1}U'U.$$

In the special case when $k_2 = 1$, the matrix C reduces to a scalar between zero and one, so that $\hat{\beta}_{1u}$ always has a smaller bias than $\hat{\beta}_{1r}$. This result does not, however, extend to more general settings where the measurement error U is correlated with X_1 (Frost 1979), or where some additional variable in either X_1 or X_2 is also measured with error (Barnow 1976; Garber and Klepper 1980; Bekker and Wansbeek 1996).

In terms of the misspecification vector $\zeta = X_2\beta_2 + \delta$, we can rewrite the biases of $\hat{\beta}_{1r}$ and $\hat{\beta}_{1u}$ from Proposition 3.1 as

$$b_r = A'\zeta, \qquad b_u = A'(I_n - X_2B')\zeta. \tag{9}$$

Based on the bias criterion, we say that X_2 represents a 'balanced addition' to model (4) if $b_u'b_u \leq b_r'b_r$. We can also write this inequality as

$$\frac{b_u'b_u}{b_r'b_r} = \frac{\eta'AA'\eta}{\zeta'AA'\zeta} \leq 1,$$

where $\eta = \zeta - X_2B'\zeta$ is the vector of 'partial residuals' (Larsen and McCleary 1972) in the regression of ζ on X_1 and X_2. The inequality trivially holds when $\delta = 0$.

4 Mean squared error

In finite samples both bias and variance matter. The biases of $\hat{\beta}_{1r}$ and $\hat{\beta}_{1u}$, given in Proposition 3.1, depend on δ. Their variances are

$$\operatorname{var}(\hat{\beta}_{1r}) = \sigma^2 A'A, \qquad \operatorname{var}(\hat{\beta}_{1u}) = \sigma^2 A' \left(I_n + X_2 B' B X_2'\right) A,$$

respectively, and these do not depend on δ. Hence, with or without misspecification, the restricted estimator $\hat{\beta}_{1r}$ is more precise (has smaller variance) than the unrestricted estimator $\hat{\beta}_{1u}$. (Recall that we have excluded the case $X_1'X_2 = 0$.) Note, however, that this result does not necessarily extend to the case of stochastic regressors (Kinal and Lahiri 1983; Teräsvirta 1987) or heteroskedastic errors (Hansen 2015, p. 176), where the inclusion of additional variables in the regression may decrease the variance.

Combining variance and bias into MSE matrices gives

$$\operatorname{MSE}(\hat{\beta}_{1r}) = \sigma^2 A'A + A'(X_2\beta_2 + \delta)(X_2\beta_2 + \delta)'A$$

and

$$\operatorname{MSE}(\hat{\beta}_{1u}) = \sigma^2 A' \left(I_n + X_2 B' B X_2'\right) A + A'(I_n - X_2 B')\delta\delta'(I_n - B X_2')A.$$

Naturally, we want to know under which conditions one MSE matrix is larger than the other, and the role of δ in this comparison. We first compare the trace of the two matrices, then we compare the matrices themselves. From the previous two MSE expressions we see that $\operatorname{tr}[\operatorname{MSE}(\hat{\beta}_{1u})] \leq \operatorname{tr}[\operatorname{MSE}(\hat{\beta}_{1r})]$ if and only if

$$\frac{b_u'b_u - b_r'b_r}{\sigma^2} \leq -\operatorname{tr}[(X_1'X_1)^{-1}X_1'X_2(X_2'M_1X_2)^{-1}X_2'X_1(X_1'X_1)^{-1}],$$

which shows that we should choose the restricted estimator more frequently when we consider both bias and variance than if we only consider the bias. Parsimonious modeling is thus a greater virtue than previously thought, a result that we shall see and emphasize again later.

In order to compare the full MSE matrices we define the $k_1 \times (k_2 + 1)$ matrix

$$Q_1 = \left[\sigma A'X_2(X_2'M_1X_2)^{-1/2} : A'(I_n - X_2B')\delta\right] \tag{10}$$

and the $(k_2 + 1)$-vector

$$\theta = \begin{pmatrix} \theta_1 \\ 1 \end{pmatrix}, \qquad \theta_1 = (X_2'M_1X_2)^{1/2}(\beta_2 + B'\delta)/\sigma, \tag{11}$$

and note that $Q_1\theta = b_r$, the bias of the restricted estimator. The MSE difference then takes the form

$$\Delta_1 = \text{MSE}(\hat{\beta}_{1u}) - \text{MSE}(\hat{\beta}_{1r}) = Q_1(I_{k_2+1} - \theta\theta')Q_1'. \tag{12}$$

The rank of Q_1 can only take one of two values, as follows.

PROPOSITION **4.1** *Let* $r = \text{rank}(X_1'X_2) \geq 1$. *Then the rank of* Q_1 *is*

$$\text{rank}(Q_1) = \begin{cases} r, & \text{if } \delta = M_1\delta_1 + X_2\delta_2 \text{ for some } \delta_1 \text{ and } \delta_2, \\ r+1, & \text{otherwise.} \end{cases}$$

In what follows we shall carefully distinguish between these two cases. We first consider the case where $\text{rank}(Q_1) = r + 1$ and introduce the symbol $^+$ to denote the Moore-Penrose inverse of a matrix.

PROPOSITION **4.2** *If* δ *is not a linear combination of the columns of* M_1 *and* X_2, *then* $\text{rank}(Q_1) = r + 1$ *and, letting*

$$q(\delta) = \theta'Q_1'(Q_1Q_1')^+Q_1\theta,$$

we obtain

$$\begin{aligned} \Delta_1 \geq 0 &\iff q(\delta) \leq 1, \\ \Delta_1 > 0 &\iff q(\delta) < 1 \text{ and } r = k_1 - 1, \end{aligned}$$

while Δ_1 *is never negative (semi)definite.*

Proposition 4.2 tells us that, in the situation where δ does not depend linearly on M_1 and X_2, the unrestricted estimator $\hat{\beta}_{1u}$ never dominates the restricted estimator $\hat{\beta}_{1r}$, and that the restricted estimator dominates the unrestricted estimator if and only if the quadratic form $q(\delta)$ is smaller than or equal to one. When does this happen? We can write

$$q(\delta) = \frac{\theta'Q_1'(Q_1Q_1')^+Q_1\theta}{\theta'\theta} \cdot \theta'\theta$$

and note that $\theta'Q_1'(Q_1Q_1')^+Q_1\theta/\theta'\theta \leq 1$ and $\theta'\theta = 1 + \theta_1'\theta_1 \geq 1$, where the first inequality follows from the fact that $Q_1'(Q_1Q_1')^+Q_1$ is symmetric and idempotent, so its eigenvalues are only zero and one. This tells us that, in general, it is not clear whether q is larger or smaller than one.

What can we say about $q(\delta)$ in the neighborhood of $\delta = 0$? In other words, when we move from no specification to a small amount of misspecification,

how sensitive is q to such a small change? Such questions are typically answered by computing the local sensitivity, that is, the derivative of $q(\delta)$ at $\delta = 0$ (Magnus and Vasnev 2007). The function q is defined for all δ, whether or not δ can be written as a linear combination of the columns of M_1 and X_2. In particular, $q(0)$ is defined, but local sensitivity is not, because the function q is not even continuous at $\delta = 0$. This follows because $\text{rank}(Q_1) = r$ when $\delta = 0$, but $\text{rank}(Q_1) = r + 1$ when δ does not lie in the space spanned by the columns of M_1 and X_2, however close to zero it is. Hence, there is a discontinuity in rank at $\delta = 0$. It then follows from Magnus and Neudecker (1999, Section 8.5) that Q_1^+ is discontinuous at $\delta = 0$ unless Q_1 has full column- or row-rank. What this means is that a small perturbation of δ may have a large effect on $q(\delta)$.

We note one case of special interest. When $r = k_2$ then Q_1 has full column-rank, so $q(\delta) = \theta'\theta = 1 + \theta_1'\theta_1$. Hence,

$$\Delta_1 \geq 0 \iff X_2'M_1(X_2\beta_2 + \delta) = 0,$$

while Δ_1 is never positive definite.

Let us next consider the case where δ lies in the space spanned by the columns of M_1 and X_2, so that we can write $\delta = M_1\delta_1 + X_2\delta_2$. The DGP (2) then takes the form

$$y = X_1\beta_1 + X_2(\beta_2 + \delta_2) + M_1\delta_1 + \epsilon,$$

from which it follows that there is no loss in generality by setting $\delta_2 = 0$. The condition $\delta = M_1\delta_1 + X_2\delta_2$ then reduces to $X_1'\delta = 0$, and the misspecification δ affects neither the bias nor the variance of the restricted estimator $\hat{\beta}_{1r}$, although it does affect the bias (but not the variance) of the unrestricted estimator $\hat{\beta}_{1u}$, unless X_2 is also orthogonal to δ.

PROPOSITION **4.3** *If* $X_1'\delta = 0$, *then* $\text{rank}(Q_1) = r$ *and, letting*

$$\omega(\delta) = \beta_2'X_2'X_1\left(X_1'X_2V_1X_2'X_1\right)^+X_1'X_2\beta_2$$

with

$$V_1 = \sigma^2(X_2'M_1X_2)^{-1} + B'\delta\delta'B,$$

we obtain

$$\begin{aligned}
\Delta_1 \geq 0 &\iff \omega(\delta) \leq 1,\\
\Delta_1 > 0 &\iff \omega(\delta) < 1 \text{ and } r = k_1,\\
\Delta_1 \leq 0 &\iff \omega(\delta) \geq 1 \text{ and } r = 1,\\
\Delta_1 < 0 &\iff \omega(\delta) > 1 \text{ and } k_1 = 1.
\end{aligned}$$

Again we should ask when it is true that $\omega(\delta) \leq 1$. First notice that V_1 is nonsingular and that its inverse is given by

$$V_1^{-1} = \frac{1}{\sigma^2}\left[X_2'M_1X_2 - \frac{X_2'M_1\delta\delta'M_1X_2/\sigma^2}{1+\delta'M_1X_2(X_2'M_1X_2)^{-1}X_2'M_1\delta/\sigma^2}\right].$$

Next, letting $W = X_1'X_2V_1^{1/2}$, we may express $\omega(\delta)$ as

$$\omega(\delta) = \beta_2'V_1^{-1/2}W'(WW')^{+}WV_1^{-1/2}\beta_2.$$

A sufficient condition for $\omega(\delta) \leq 1$ is therefore $\beta_2'V_1^{-1}\beta_2 \leq 1$, but this condition is, in general, not necessary. Using the expression for V_1^{-1} we find

$$\beta_2'V_1^{-1}\beta_2 \leq 1 \iff \lambda \leq 1 + \lambda_\delta,$$

where

$$\lambda = \frac{\beta_2'X_2'M_1X_2\beta_2}{\sigma^2}, \qquad \lambda_\delta = \frac{\delta'M_1X_2\beta_2\beta_2'X_2'M_1\delta/\sigma^2}{\sigma^2[1+\delta'M_1X_2(X_2'M_1X_2)^{-1}X_2'M_1\delta/\sigma^2]}. \quad (13)$$

We note that λ is the noncentrality parameter in the distribution of the classical F-statistic for testing the hypothesis that $\beta_2 = 0$ in the long model (5) when the errors are normal. The condition $\lambda \leq 1$ is well-known as the condition under which the *complete* restricted estimator $(\hat{\beta}_{1r}, 0)$ has smaller MSE than the *complete* unrestricted estimator $(\hat{\beta}_{1u}, \hat{\beta}_{2u})$ in the absence of misspecification; see Toro-Vizcarrondo and Wallace (1968, Equation (19)).

In contrast to the setup in Proposition 4.2 the function ω is now differentiable at $\delta = 0$. This is because ω depends on δ only through $M_1\delta$. The derivative of ω at $\delta = 0$ vanishes, but the second derivative is nonzero, in fact negative semidefinite. Hence, ω achieves a maximum at $\delta = 0$. We can see this also by noting that the fact that

$$X_1'X_2V_1X_2'X_1 \geq \sigma^2X_1'X_2(X_2'M_1X_2)^{-1}X_2'X_1$$

implies that

$$(X_1'X_2V_1X_2'X_1)^{+} \leq \frac{[X_1'X_2(X_2'M_1X_2)^{-1}X_2'X_1]^{+}}{\sigma^2},$$

because the rank of the matrix $X_1'X_2V_1X_2'X_1$ does not depend on δ; see Milliken and Akdeniz (1977) and Magnus and Neudecker (1999, Miscellaneous Exercise No. 13). This means that ignoring misspecification favors the unrestricted estimator, and that we should therefore be even more parsimonious in our modeling than common practice prescribes.

We consider two special cases. First, when $r = k_2$ then the sufficient condition $\beta_2' V_1^{-1} \beta_2 \leq 1$ is also necessary, and hence

$$\begin{aligned}
\Delta_1 \geq 0 &\iff \lambda \leq 1 + \lambda_\delta, \\
\Delta_1 > 0 &\iff \lambda < 1 + \lambda_\delta \text{ and } k_1 = k_2, \\
\Delta_1 \leq 0 &\iff \lambda \geq 1 + \lambda_\delta \text{ and } k_2 = 1, \\
\Delta_1 < 0 &\iff \lambda > 1 + \lambda_\delta \text{ and } k_1 = k_2 = 1.
\end{aligned}$$

When, in addition, $X_2'\delta = 0$, then $\lambda_\delta = 0$ and we find that the restricted estimator $\hat{\beta}_{1r}$ dominates the unrestricted estimator $\hat{\beta}_{1u}$ if and only if $\lambda \leq 1$.

Second, when, in addition to $X_1'\delta = 0$, also $X_2'\delta = 0$ (which is less restrictive than the textbook case $\delta = 0$), we find $\omega(\delta) = \omega_0$, where

$$\omega_0 = \frac{\beta_2' X_2' X_1 \left[X_1' X_2 (X_2' M_1 X_2)^{-1} X_2' X_1\right]^+ X_1' X_2 \beta_2}{\sigma^2}. \tag{14}$$

This special case was, in essence, first derived by Magnus and Durbin (1999, Theorem 1) and is required in the 'equivalence theorem' which motivates the class of weighted-average least squares (WALS) estimators; see Magnus and De Luca (2014) for a survey. Of course, when also $r = k_2$ then we find again that $\omega(\delta) = \lambda$.

In this second special case, where δ is orthogonal to both X_1 and X_2, the unrestricted estimator $\hat{\beta}_{1u}$ is unbiased. Thus, $Q_1 Q_1' = \text{var}(\hat{\beta}_{1u}) - \text{var}(\hat{\beta}_{1r})$ and ω_0 corresponds, in essence, to the noncentrality parameter in the distribution of the Hausman statistic for testing the hypothesis $\beta_2 = 0$ in model (5) with normal errors; see Holly (1982, p. 754).

5 Residuals, estimation of σ^2, and testing

An estimate of the error variance σ^2 is needed in order to assess the sampling variability of the OLS estimators. This estimate is typically constructed by suitably rescaling the sum of the squared residuals. The residuals in the short model (4) and the long model (5) are given by

$$\hat{\epsilon}_r = M_1 y, \qquad \hat{\epsilon}_u = M y,$$

respectively, where $M = I_n - X(X'X)^{-1}X'$ is the usual symmetric idempotent matrix of rank $n - k$. By the properties of M_1 and M, we have

$$\hat{\epsilon}_r' \hat{\epsilon}_r = y' M_1 y, \qquad \hat{\epsilon}_u' \hat{\epsilon}_u = y' M y,$$

and the restricted and unrestricted OLS estimators of σ^2 are, respectively,

$$s_r^2 = \frac{y'M_1y}{n-k_1}, \qquad s_u^2 = \frac{y'My}{n-k}.$$

Under the additional assumption that the regression errors in (2) are normally distributed, the distributions of s_r^2 and s_u^2 are noncentral chi-squared:

$$\frac{(n-k_1)s_r^2}{\sigma^2} \sim \chi^2(n-k_1, \lambda_r), \qquad \frac{(n-k)s_u^2}{\sigma^2} \sim \chi^2(n-k, \lambda_u),$$

where

$$\lambda_r = \frac{\zeta'M_1\zeta}{\sigma^2}, \qquad \lambda_u = \frac{\zeta'M\zeta}{\sigma^2} = \frac{\delta'M\delta}{\sigma^2}$$

are the noncentrality parameters. This implies that

$$\mathrm{E}(s_r^2) = \sigma^2\left(1 + \frac{\lambda_r}{n-k_1}\right), \qquad \mathrm{E}(s_u^2) = \sigma^2\left(1 + \frac{\lambda_u}{n-k}\right).$$

The textbook case where $\delta = 0$ gives $\lambda_u = 0$ and $\zeta = X_2\beta_2$, so that $\lambda_r = \lambda$ as defined in (13). In this case the restricted estimator s_r^2 is biased, while the unrestricted estimator s_u^2 is unbiased. If $\delta \neq 0$, then both estimators of σ^2 are biased upward and is not clear *a priori* which of the two biases is larger. In fact, $\mathrm{bias}(s_u^2) \leq \mathrm{bias}(s_r^2)$ if and only if

$$\frac{\lambda_u}{\lambda_r} \leq \frac{n-k}{n-k_1}.$$

This inequality holds when $\delta = 0$, in which case $\lambda_u = 0$, but when $\delta \neq 0$ we may have $(n-k)/(n-k_1) \leq \lambda_u/\lambda_r \leq 1$, which implies that $\mathrm{bias}(s_u^2) \geq \mathrm{bias}(s_r^2)$. Also notice that, unless $n \to \infty$, the two biases in estimating σ^2 do not vanish even when X_1, X_2, and δ are orthogonal. Although $\hat{\beta}_{1r}$ and $\hat{\beta}_{1u}$ are unbiased in this limiting case, inference about β_1 based on classical test statistics and confidence intervals will be incorrect because actual coverage levels are higher than nominal.

Next we consider the classical F-statistic

$$F = \frac{(\hat{\epsilon}_r'\hat{\epsilon}_r - \hat{\epsilon}_u'\hat{\epsilon}_u)/k_2}{\hat{\epsilon}_u'\hat{\epsilon}_u/(n-k)}$$

for testing the correct specification of the short model (4). In the textbook case where $\delta = 0$, this statistic follows a noncentral F-distribution with $(k_2, n-k)$ degrees of freedom and noncentrality parameter λ as defined

in (13). In our more general setup where δ can be different from zero, the basic decomposition

$$M_1 = M + M_1 X_2 (X_2' M_1 X_2)^{-1} X_2' M_1$$

implies that

$$\hat{\epsilon}_r' \hat{\epsilon}_r - \hat{\epsilon}_u' \hat{\epsilon}_u = y' M_1 X_2 (X_2' M_1 X_2)^{-1} X_2' M_1 y.$$

The numerator and the denominator of the F-statistic are still independent as $(M_1 - M)M = 0$, but now *both* follow noncentral chi-squared distributions

$$\frac{\hat{\epsilon}_r' \hat{\epsilon}_r - \hat{\epsilon}_u' \hat{\epsilon}_u}{\sigma^2} \sim \chi^2(k_2, \lambda_r - \lambda_u), \qquad \frac{\hat{\epsilon}_u' \hat{\epsilon}_u}{\sigma^2} \sim \chi^2(n-k, \lambda_u).$$

The distribution of the F-statistic is therefore a *doubly* noncentral F with $(k_2, n-k)$ degrees of freedom and noncentrality parameters $(\lambda_r - \lambda_u, \lambda_u)$. We refer the reader to Johnson, Kotz, and Balakrishnan (1995, Chapter 30) for a discussion on the properties of this distribution.

In our framework, testing the correct specification of the short model (4) amounts to testing the null hypothesis that $\zeta = 0$. Under this null, the two noncentrality parameters $\lambda_r - \lambda_u$ and λ_u are both zero and the distribution of the F-statistic is a central F with $(k_2, n-k)$ degrees of freedom. In contrast, testing the correct specification of the long model (5) amounts to testing the null hypothesis that $\delta = 0$. Under this null, $\lambda_u = 0$ and $\lambda_r - \lambda_u = \lambda$ defined in (13). The distribution of the F-statistic is then a *singly* noncentral F with $(k_2, n-k)$ degrees of freedom and noncentrality parameter λ.

6 Conclusions

It is not generally true that adding variables to a linear regression model reduces the bias of the parameters of interest. This *is* true when we compare a short model with a long model which coincides with the DGP, but it is not necessarily true when both the short and the long model are underspecified, as is the common situation. In this more common situation the strategy of adding variables may increase both the bias and the variance of the OLS estimators. The consequences of adding or omitting variables are ambiguous.

We have analyzed this ambiguity by providing exact expressions for the bias and MSE comparisons of the OLS estimators from two misspecified models with different sets of regressors, and shown that MSE comparisons are particularly sensitive to small perturbations of the misspecification vector δ in a neighborhood of $\delta = 0$. We related our conditions for bias and MSE dominance to previous findings in the literature, and we discussed the local

sensitivity of MSE comparisons to small perturbations in misspecification, the estimation of the error variance, and the usual F-test for misspecification.

Throughout we have focused on a simple DGP with fixed regressors and homoskedastic errors, and our results on MSE comparisons do not directly extend to frameworks with either stochastic regressors or heteroskedastic errors, where the inclusion of additional variables may decrease the variance of the corresponding OLS estimator. The extension to stochastic regressors, which could be developed by generalizing the setup in Kinal and Lahiri (1983), would be particularly useful because it would provide insight in a whole class of important problems where misspecification occurs in the form of a random variable, e.g. errors in variables, simultaneity, and sample selection.

Compared to the textbook case where the long model and the DGP coincide, our findings tilt the bias-precision trade-off between the restricted and unrestricted OLS estimators in favor of the restricted estimators. This emphasizes again the importance of model parsimony and the importance of recognizing models as approximations of an unknown DGP. The first issue was championed by Einstein whose words 'As simple as possible, but not simpler' may be difficult to apply, but remain excellent advice. The second issue has implications for model-building strategies, as pointed out by Hansen (2005) and others, emphasizing the dangers of ignoring the impact of model selection on inference. The development of model-averaging techniques in an $\mathcal{M}$-open framework, where the DGP is not included in the assumed set of models, is therefore one of several challenging lines for future research.

Appendix A: Data

Table 1 presents the data underlying the example in the introduction. These data have been obtained by repeatedly drawing thirty observations from a multivariate normal distribution with zero means, unit variances, and correlation matrix

$$R = \begin{bmatrix} 1 & -0.25 & 0.20 & 0.35 \\ -0.25 & 1 & -0.20 & -0.30 \\ 0.20 & -0.20 & 1 & -0.40 \\ 0.35 & -0.30 & -0.40 & 1 \end{bmatrix}$$

until the absolute value of the differences between sample and theoretical correlations did not exceed 0.05. In the chosen pseudo-random sample, the regressors x_1, x_2, x_3, and x_4 have means

$$m = \begin{pmatrix} -0.1069 & 0.0016 & -0.3261 & 0.1297 \end{pmatrix},$$

Table 1: Data underlying the example in the introduction.

obs	x_1	x_2	x_3	x_4
1	0.8163	0.9657	−0.3117	0.1100
2	−2.2118	1.2288	0.7152	−2.1125
3	−0.0853	0.2224	−1.1263	1.4286
4	−0.6089	1.5049	−0.6765	−1.6555
5	−1.7725	0.0492	−1.5358	1.2491
6	−0.6467	0.4116	−0.0659	0.8478
7	−0.0896	0.5599	−1.3879	0.0872
8	0.9110	0.0359	0.7732	−1.1832
9	−1.1612	0.5058	−1.5293	0.2883
10	0.0072	−1.4850	0.2864	1.3529
11	0.9896	−0.4203	0.5608	1.0412
12	0.7906	0.5619	0.6586	−0.4542
13	0.3187	1.1368	−1.4688	−0.0096
14	0.0649	1.2908	0.1675	0.0866
15	−1.5940	−1.2709	−1.2838	0.4529
16	−0.4108	−1.9377	2.1200	−1.3335
17	0.6693	0.2766	−0.5916	−0.2905
18	−0.2095	−0.4355	−0.4012	−0.4672
19	0.6368	−1.5590	−0.7939	1.7599
20	−0.7128	0.1294	−0.1938	−1.3237
21	1.4261	−1.8963	−0.6207	2.0997
22	1.2016	−0.7201	0.6938	0.5522
23	−3.3882	0.8007	−1.4663	−0.2741
24	0.0546	−0.9427	0.7044	−0.9465
25	1.5487	0.6487	−2.8828	1.6191
26	−0.4126	1.4388	1.2894	0.1542
27	−0.9765	−0.4683	−0.6976	−0.8480
28	−0.3026	0.2605	−2.1524	1.3546
29	0.3177	−0.6855	0.3742	−0.0908
30	1.6229	−0.1594	1.0584	0.3972

standard deviations

$$s = \begin{pmatrix} 1.1518 & 0.9853 & 1.1230 & 1.0801 \end{pmatrix},$$

and correlation matrix

$$R = \begin{bmatrix} 1 & -0.2237 & 0.1930 & 0.3429 \\ -0.2237 & 1 & -0.2223 & -0.2973 \\ 0.1930 & -0.2223 & 1 & -0.4348 \\ 0.3429 & -0.2973 & -0.4348 & 1 \end{bmatrix}.$$

The DGP underlying Figure 1 includes a constant term and the four regressors in Table 1 with associated parameters equal to $\beta_0 = 1$, $\beta_1 = 1$, $\beta_2 = -1$, $\beta_3 = 1$, and $\beta_4 = 1$.

Appendix B: Proofs

In this appendix we present proofs of the propositions in the paper.

Proof of Proposition 3.1. From (7) we have

$$b_r = \mathrm{E}(\hat{\beta}_{1r}) - \beta_1 = A'(\mathrm{E}\, y) - \beta_1$$

and, from (8),

$$b_u = \mathrm{E}(\hat{\beta}_{1u}) - \beta_1 = (A' - A'X_2B')(\mathrm{E}\, y) - \beta_1.$$

The results then follow from the fact that $\mathrm{E}\, y = X_1\beta_1 + X_2\beta_2 + \delta$. ‖

Proof of Proposition 4.1. Since $\mathrm{rank}(A'X_2) = \mathrm{rank}(X_1'X_2) = r$ we have $\mathrm{rank}(Q_1) = r$ if and only if

$$A'(I_n - X_2B')\delta = A'X_2(X_2'M_1X_2)^{-1/2}\mu$$

for some μ. This occurs if and only if

$$(I_n - X_2B')\delta - X_2(X_2'M_1X_2)^{-1/2}\mu = M_1\delta_1$$

for some δ_1 and μ, and hence if and only if $\delta = M_1\delta_1 + X_2\delta_2$ for some δ_1 and δ_2. ‖

To prove Propositions 4.2 and 4.3 we shall need the following lemma.

LEMMA **B.1** *Let $Q \neq 0$ be an $m \times n$ matrix ($m \geq 1$, $n \geq 1$) and let θ be an $n \times 1$ vector. Define the $m \times m$ matrix $\Delta = Q(I_n - \theta\theta')Q'$. Then,*

$$\text{rank}(\Delta) = \begin{cases} \text{rank}(Q) - 1, & \text{if } \theta'Q'(QQ')^+Q\theta = 1, \\ \text{rank}(Q), & \text{otherwise,} \end{cases}$$

where A^+ denotes the Moore-Penrose inverse of A. Further,

$$\begin{aligned} \Delta \geq 0 &\iff \theta'Q'(QQ')^+Q\theta \leq 1, \\ \Delta > 0 &\iff \theta'Q'(QQ')^{-1}Q\theta < 1 \text{ and } \text{rank}(Q) = m, \\ \Delta \leq 0 &\iff \theta'Q'(QQ')^+Q\theta \geq 1 \text{ and } \text{rank}(Q) = 1, \\ \Delta < 0 &\iff \theta'Q'(QQ')^{-1}Q\theta > 1 \text{ and } m = 1. \end{aligned}$$

Proof. Let $A = QQ'$ and $a = Q\theta$. Note that $Q'(QQ')^-Q$ is unique, hence equal to $Q'(QQ')^+Q$, and that $\text{rank}(A : a) = \text{rank}(A)$. The results about the rank and the semidefiniteness then follow from Lemma A1 in Magnus and Durbin (1999). The statements about $\Delta > 0$ and $\Delta < 0$ follow by adding the requirement that Δ is nonsingular. ‖

Proof of Proposition 4.2. If $\delta \neq M_1\delta_1 + X_2\delta_2$, then $\text{rank}(Q_1) = r + 1$ because of Proposition 4.1. The result then follows from Lemma B.1. ‖

Proof of Proposition 4.3. If $X_1'\delta = 0$, then $\text{rank}(Q_1) = r$ because of Proposition 4.1. Also,

$$Q_1 = A'X_2 \left[\sigma(X_2'M_1X_2)^{-1/2} : -B'\delta\right], \qquad Q_1\theta = A'X_2\beta_2,$$

so that

$$\Delta_1 = Q_1Q_1' - Q_1\theta\theta'Q_1' = A'X_2(V_1 - \beta_2\beta_2')X_2'A.$$

The matrix Δ_1 is positive (negative) (semi)definite if and only if the matrix

$$X_1'X_2(V_1 - \beta_2\beta_2')X_2'X_1$$

is positive (negative) (semi)definite, and the result follows again from Lemma B.1. ‖

References

Angrist, J. A., and Pischke, J.-S. (2009) *Mostly Harmless Econometrics*. Princeton University Press, Princeton.

Angrist, J. A., and Pischke, J.-S. (2015) *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press, Princeton.

Barnow, B. S. (1976) The use of proxy variables when one or two independent variables are measured with error. *American Statistician* 30: 119–121.

Bekker, P. A. and Wansbeek, T. J. (1996) Proxy versus omitted variables in regression analysis. *Linear Algebra and Its Applications* 237: 301–312.

Bernardo, J. M., and Smith, A. F. M. (1994) *Bayesian Theory*. Wiley, New York.

Cameron, A. C., and Trivedi, P. K. (2005) *Microeconometrics: Methods and Applications*. Cambridge University Press, New York.

Clarke, K. A. (2005) The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science* 22: 341–352.

Clyde, M. A., and Iversen, E. S. (2013) Bayesian model averaging in the M-open framework. In P. Damien, P. Dellaportas, N. G. Polson and D. A. Stephens (eds.) *Bayesian Theory and Applications* (p. 483–498). Oxford University Press, Oxford.

Davidson, R., and MacKinnon, J. G. (2004) *Econometric Theory and Methods*. Oxford University Press, New York.

Frost, P. A. (1979) Proxy variables and specification bias. *The Review of Economics and Statistics* 61: 323–325.

Garber, S., and Klepper, S. (1980) Extending the classical normal errors-in-variables model. *Econometrica* 48: 1541–1546.

Greene, W. H. (2011) *Econometric Analysis* (7th ed). Prentice Hall, New York.

Hansen, B. E. (2005) Challenges for econometric model selection. *Econometric Theory* 21: 60–68.

Hansen, B. E. (2015) *Econometrics*. Unpublished manuscript (http://www.ssc.wisc.edu/∼bhansen/).

Hausman, J. A. (1978) Specification tests in econometrics. *Econometrica* 46: 1251–1271.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999) Bayesian model averaging: A tutorial. *Statistical Science* 14: 382–417.

Holly, A. (1982) A remark on Hausman's specification test. *Econometrica* 50: 749–760.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995) *Continuous Univariate Distributions* (Volume 2, 2nd ed). John Wiley & Sons, New York.

Kinal, T., and Lahiri, K. (1983) Specification error analysis with stochastic regressors. *Econometrica* 51: 1209–1219.

Larsen, W. A., and McCleary, S. J. (1972) The use of partial residual plots in regression analysis. *Technometrics* 14: 781–790.

Maddala, G. S. (1992) *Introduction to Econometrics* (2nd ed). Macmillan Publishing Company, New York.

Magnus, J. R., and De Luca, G. (2014) Weighted-average least squares (WALS): A survey. *Journal of Economic Surveys* doi: 10.1111/joes.12094.

Magnus, J. R., and Durbin, J. (1999) Estimation of regression coefficients of interest when other regression coefficients are of no interest. *Econometrica* 67: 639–643.

Magnus, J. R., and H. Neudecker (1999) *Matrix Differential Calculus with Applications in Statistics and Econometrics* (2nd ed). John Wiley & Sons, England.

Magnus, J. R., and Vasnev, A. L. (2007) Local sensitivity and diagnostic tests. *Econometrics Journal* 10: 166–192.

McCallum, B. T. (1972) Relative asymptotic bias from errors of omission and measurement. *Econometrica* 40: 757–758.

Milliken, G. A., and Akdeniz, F. (1977) A theorem on the difference of the generalized inverses of two nonnegative matrices. *Communications in Statistics—Theory and Methods* A6: 73–79.

Teräsvirta, T. (1987) Usefulness of proxy variables in linear models with stochastic regressors. *Journal of Econometrics* 36: 377–382.

Toro-Vizcarrondo, C., and Wallace, T. D. (1968) A test of the mean square error criterion for restrictions in linear regression. *Journal of the American Statistical Association* 63: 558–572.

Wooldridge, J. M. (2012) *Introductory Econometrics: A Modern Approach* (5th ed). South-Western Cengage Learning, Mason.