JOURNAL OF ECONOMIC SURVEYS

doi: 10.1111/joes.12245

BALANCED VARIABLE ADDITION IN LINEAR MODELS

Giuseppe De Luca
Università di Palermo

Jan R. Magnus
Vrije Universiteit Amsterdam and Tinbergen Institute

Franco Peracchi*
Georgetown University, EIEF and University of Rome Tor Vergata

Abstract. This paper studies what happens when we move from a short regression to a long regression in a setting where both regressions are subject to misspecification. In this setup, the least-squares estimator in the long regression may have larger inconsistency than the least-squares estimator in the short regression. We provide a simple interpretation for the comparison of the inconsistencies and study under which conditions the additional regressors in the long regression represent a "balanced addition" to the short regression.

Keywords. Bias amplification; Inconsistency; Least-squares estimators; Mean squared error; Omitted variables; Proxy variables

1. Introduction

Ludwig van Beethoven composed nine symphonies. Suppose a 10th 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, 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 linear regression model and omitted variables but, given the connection between omitted variables and many other forms of misspecification, our analysis extends to

*Corresponding author contact email: fp211@georgetown.edu; Tel: +1 202 687 6131.

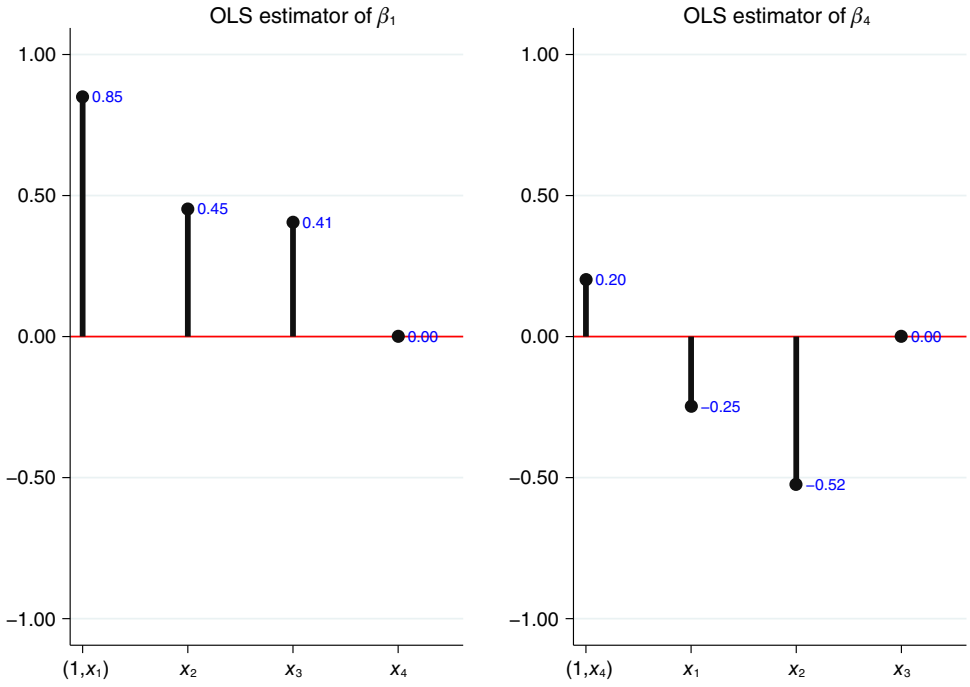

Figure 1. Bias of the OLS Estimators of β_1 and β_4 by Adding Regressors. [Colour figure can be viewed at wileyonlinelibrary.com]

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:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$$

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. (The underlying data are available upon request.) 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 x_1, x_2, and x_3. In the right panel of Figure 1, the bias no longer decreases monotonically to zero. 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 usually mentioned in textbooks, a notable exception being Hansen (2017, p. 39). 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 the bias 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 message from the textbook analysis is that adding variables to the model always decreases the bias of the OLS estimator of interest but increases its sampling variance. For example, Wassermann (2010, p. 218) argues: "As you add more variables to a regression, the bias of the predictions decreases and the variance increases." 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 stochastic regressors or heteroskedastic errors (which we shall also consider), the conclusions are more nuanced because the inclusion of additional variables may also decrease the sampling variance. This bias-variance 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.

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 the 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 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. Section 2 presents our setup. Section 3 presents and compares the bias and inconsistency of our OLS estimators. Unlike Clarke (2005) and Hansen (2017), we provide a simple interpretation for the comparison of the inconsistencies and study under which conditions the additional regressors in the long regression represent a balanced addition to the short regression. Section 4 discusses the relationship between our results and two strands of literature: one which considers the potential bias-reducing role of proxy variables and one which studies the sensitivity of conventional methods for estimating causal parameters to the choice of conditioning variables. Section 5 discusses the variance and mean squared error (MSE) of our estimators, both under fixed and stochastic regressors. Finally, Section 6 concludes.

2. Setup

Suppose the data are generated by the process

$$y = X_1\beta_1 + X_2\beta_2 + \delta + \epsilon \tag{1}$$

where the vector y $(n \times 1)$ contains the observations on the outcome of interest, X_1 $(n \times k_1)$ and X_2 $(n \times k_2)$ are (possibly random) matrices of regressors with $k_1 \geq 1$, $k_2 \geq 1$, and $k = k_1 + k_2 < n$, β_1 and

β_2 are unknown parameter vectors, δ is a (possibly random) vector representing misspecification, and ϵ is a vector of random errors satisfying

$$E(\epsilon \mid X_1, X_2, \delta) = 0 \tag{2}$$

We assume that the matrix $X = [X_1 : X_2]$ has full column-rank k (with probability 1). We do not assume normality, but we do assume that all random variables that appear in the DGP (1) have finite second moments.

Given (1), the moment condition (2) implies that

$$E(y \mid X_1, X_2, \delta) = X_1\beta_1 + X_2\beta_2 + \delta$$

and this allows us to interpret β_1, our parameter of primary interest, as the causal effect of the variables in X_1 on the outcome y (see, e.g., Angrist and Pischke, 2009, 2015). A special case of interest is when X_1 consists of an intercept and a binary treatment indicator, as in the "modern econometric paradigm" emphasized by Angrist and Pischke (2017).

The DGP (1) encompasses a wide range of misspecification problems, such as incorrect choice of functional form, measurement errors in the regressors, simultaneity, unobserved heterogeneity, censoring, and sample selection.

Since the DGP is not known, δ is excluded from any model used for estimation purposes. We consider two regression models, labeled "long" and "short," respectively. The long regression model (indexed by u for "unrestricted") is given by

$$y = X_1\beta_{1u} + X_2\beta_{2u} + \epsilon_u \tag{3}$$

and the short regression model (indexed by r for "restricted") by

$$y = X_1\beta_{1r} + \epsilon_r \tag{4}$$

In general, neither the long nor the short regression has a causal interpretation. In the special case where $X'\delta = 0$, we have $\beta_{1u} = \beta_1$ and $\beta_{2u} = \beta_2$. This is the textbook case where the misspecification δ does not affect the estimation of β_1 and β_2 in the long model. However, if $X'\delta \neq 0$ then both the long and the short 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).

Letting $M_1 = I_n - X_1(X_1'X_1)^{-1}X_1'$, the usual symmetric idempotent matrix of rank $n - k_1$, we can write the unrestricted OLS estimators of β_{1u} and β_{2u} in the long model (3) as

$$\hat{\beta}_{1u} = (X_1'X_1)^{-1}X_1'y - (X_1'X_1)^{-1}X_1'X_2(X_2'M_1X_2)^{-1}X_2'M_1y$$

and

$$\hat{\beta}_{2u} = (X_2'M_1X_2)^{-1}X_2'M_1y$$

and the restricted OLS estimator of β_{1r} in the short model (4) as

$$\hat{\beta}_{1r} = (X_1'X_1)^{-1}X_1'y$$

This implies that

$$\hat{\beta}_{1u} - \beta_1 = (X_1'X_1)^{-1}\left[X_1'(\delta + \epsilon) - (X_1'X_2)(X_2'M_1X_2)^{-1}X_2'M_1(\delta + \epsilon)\right] \tag{5}$$

and

$$\hat{\beta}_{1r} - \beta_1 = (X_1'X_1)^{-1}[(X_1'X_2)\beta_2 + X_1'(\delta + \epsilon)] \tag{6}$$

Note that in the special case $X_1'X_2 = 0$ we have $\hat{\beta}_{1r} = \hat{\beta}_{1u}$, and a comparison is meaningless.

With a view toward an asymptotic comparison, we write

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \frac{1}{n}\begin{pmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{pmatrix}$$

and

$$q = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix} = \frac{1}{n}\begin{pmatrix} X_1' \\ X_2' \end{pmatrix}(\delta + \epsilon) \tag{7}$$

Note that the matrices Σ_{ij} $(i, j = 1, 2)$ and the vectors q_i $(i = 1, 2)$ depend on n but we shall not make this dependence explicit in the notation, unless there is possibility for confusion. Letting

$$\Sigma^{-1} = \begin{pmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{pmatrix}$$

where

$$\Sigma^{11} = \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}\left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1}\Sigma_{21}\Sigma_{11}^{-1} \tag{8}$$

$$\Sigma^{12} = -\Sigma_{11}^{-1}\Sigma_{12}\left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1}$$

and

$$\Sigma^{22} = \left(\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)^{-1}$$

we obtain from (5) and (6):

$$\hat{\beta}_{1u} - \beta_1 = \Sigma^{11}q_1 + \Sigma^{12}q_2 \tag{9}$$

and

$$\hat{\beta}_{1r} - \beta_1 = \Sigma_{11}^{-1}q_1 + \Sigma_{11}^{-1}\Sigma_{12}\beta_2 \tag{10}$$

Closer inspection of (9) and (10) reveals that the bias and variance of the unrestricted estimator $\hat{\beta}_{1u}$ does not depend on β_2, but that the bias of the restricted estimator $\hat{\beta}_{1r}$ does depend on β_2. The variance of $\hat{\beta}_{1r}$ also depends on β_2, but only in finite samples. Asymptotically, the term $\Sigma_{11}^{-1}\Sigma_{12}$ converges to a nonrandom matrix and therefore var$(\hat{\beta}_{1r})$ is asymptotically equal to the variance of $\Sigma_{11}^{-1}q_1$, which does not depend on β_2.

In the next three sections, we study the bias, inconsistency, variance, and MSE of these two estimators of β_1, in particular their dependence on δ.

3. Bias and Inconsistency

Since the bias dominates the variance in sufficiently large samples, we concentrate on the bias first. The following result gives the biases of $\hat{\beta}_{1u}$ and $\hat{\beta}_{1r}$ when X and δ are fixed (nonrandom).

Proposition 1. *Suppose that X and δ are fixed (nonrandom). Then, under* (1) *and* (2),

$$b_{1u} = E(\hat{\beta}_{1u} - \beta_1) = \Sigma^{11}(X_1'\delta/n) + \Sigma^{12}(X_2'\delta/n)$$

and

$$b_{1r} = E(\hat{\beta}_{1r} - \beta_1) = \Sigma_{11}^{-1}(X_1'\delta/n) + \Sigma_{11}^{-1}\Sigma_{12}\beta_2$$

Proof. This follows directly from (9) and (10), using the fact that $E(X'\epsilon) = 0$. ■

The next result gives the inconsistencies of $\hat{\beta}_{1u}$ and $\hat{\beta}_{1r}$ when X and δ are random.

Proposition 2. *Suppose that X and δ are random, and that*

$$\text{plim}\Sigma = \bar{\Sigma} = \begin{pmatrix} \bar{\Sigma}_{11} & \bar{\Sigma}_{12} \\ \bar{\Sigma}_{21} & \bar{\Sigma}_{22} \end{pmatrix}, \qquad \text{plim}\frac{1}{n}\begin{pmatrix} X_1'\delta \\ X_2'\delta \end{pmatrix} = \bar{q} = \begin{pmatrix} \bar{q}_1 \\ \bar{q}_2 \end{pmatrix},$$

respectively, a finite nonsingular k × k matrix and a finite k × 1 vector. Then, under (1) *and* (2)*,*

$$\bar{b}_{1u} = \text{plim}(\hat{\beta}_{1u} - \beta_1) = \bar{\Sigma}^{11}\bar{q}_1 + \bar{\Sigma}^{12}\bar{q}_2$$

and

$$\bar{b}_{1r} = \text{plim}(\hat{\beta}_{1r} - \beta_1) = \bar{\Sigma}_{11}^{-1}\bar{q}_1 + \bar{\Sigma}_{11}^{-1}\bar{\Sigma}_{12}\beta_2$$

Proof. This also follows from (9) and (10), now using the limiting assumptions stated in the proposition. ■

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

$$\bar{b}_{1u} = 0, \qquad \bar{b}_{1r} = \bar{\Sigma}_{11}^{-1}\bar{\Sigma}_{12}\beta_2$$

so the unrestricted estimator $\hat{\beta}_{1u}$ is consistent, while the restricted estimator is consistent only if β_2 lies in the null space of the matrix $\bar{\Sigma}_{12}$. Another special case arises when

$$\text{plim}\frac{1}{n}X_1'(X_2\beta_2 + \delta) = 0$$

in other words, when X_1 is orthogonal to the misspecification error $X_2\beta_2 + \delta$ in the short model. Then,

$$\bar{b}_{1u} = \bar{\Sigma}^{12}(\bar{\Sigma}_{22}\beta_2 + \bar{q}_2), \qquad \bar{b}_{1r} = 0$$

so the restricted estimator is consistent, while the unrestricted estimator is generally inconsistent.

To gain further insight into the properties of the two estimators of β_1, we write their inconsistencies as

$$\bar{b}_{1u} = \tau_1, \qquad \bar{b}_{1r} = \tau_1 + \Psi(\beta_2 + \tau_2) \tag{11}$$

where $\Psi = \bar{\Sigma}_{11}^{-1}\bar{\Sigma}_{12}$ is the $k_1 \times k_2$ matrix of (population) coefficients in the linear projection of X_2 on X_1, and

$$\tau_1 = \bar{\Sigma}^{11}\bar{q}_1 + \bar{\Sigma}^{12}\bar{q}_2, \qquad \tau_2 = \bar{\Sigma}^{21}\bar{q}_1 + \bar{\Sigma}^{22}\bar{q}_2$$

are, respectively, the $k_1 \times 1$ and $k_2 \times 1$ vectors of (population) coefficients in the linear projection of δ on X_1 and X_2. This linear projection may be replaced by the conditional mean of δ given X_1 and X_2 when the latter is linear in parameters, for example, when X_1 and X_2 are both discrete. Figure 2 presents a path diagram of the DGP based on the coefficients of these auxiliary regressions. The diagram makes it clear that the inconsistencies of the two estimators depend on the two indirect effects of X_1 on y: one operating through X_2, namely, $\Psi(\beta_2 + \tau_2)$; and one operating through δ, namely, τ_1. The first affects only $\bar{b}_{1r}$, while the second affects both $\bar{b}_{1r}$ and $\bar{b}_{1u}$. Notice that the expression for $\bar{b}_{1r}$ in (11) generalizes the classical omitted variables bias formula to the case when the long regression is smaller than the DGP.

To illustrate, consider the important special case when $k_1 = 1$, so the matrix Ψ reduces to a $1 \times k_2$ vector denoted by ψ'. In this case,

$$\bar{b}_{1u} = \tau_1, \qquad \bar{b}_{1r} = \tau_1 + \psi'(\beta_2 + \tau_2)$$

If the unrestricted estimator is inconsistent for β_1 ($\tau_1 \neq 0$), then we can write the ratio of the inconsistencies of the two estimators as

$$\frac{\bar{b}_{1r}}{\bar{b}_{1u}} = 1 + \frac{\psi'(\beta_2 + \tau_2)}{\tau_1} \tag{12}$$

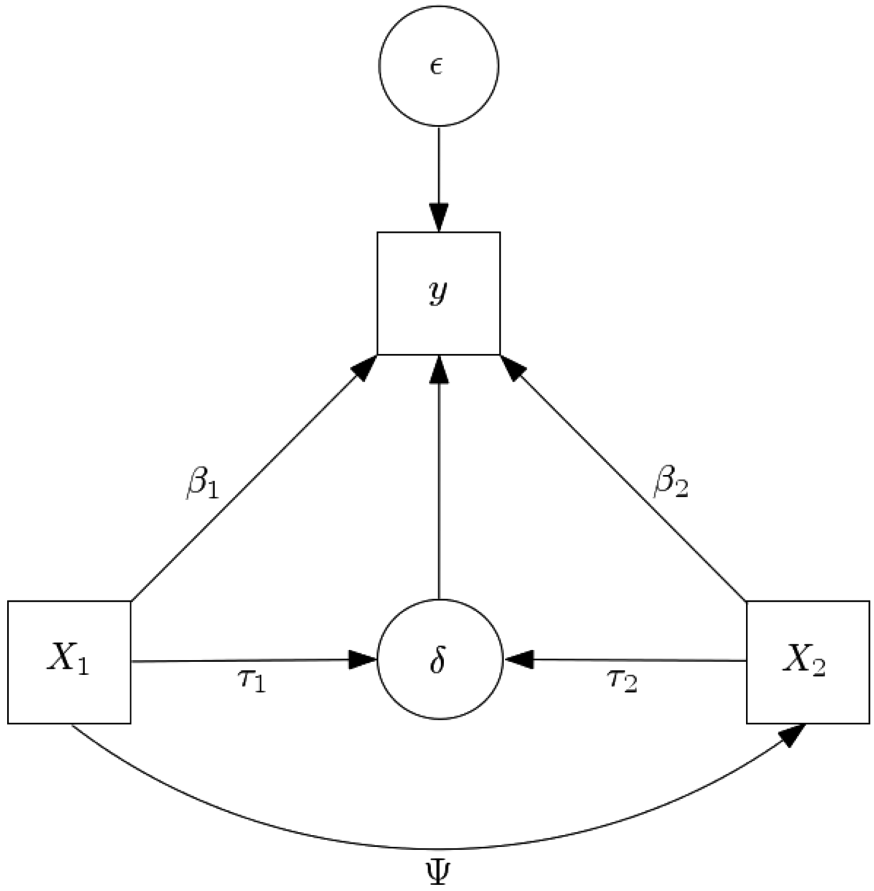

Figure 2. Path Diagram of the DGP.

This ratio is a function of $\psi'(\beta_2 + \tau_2)$ and τ_1 only, and depending on their values it can be either positive or negative, and either greater or smaller than one in absolute value. We distinguish four cases:

(a) If $\psi'(\beta_2 + \tau_2)$ and τ_1 have the same sign, then $\bar{b}_{1r}/\bar{b}_{1u} > 1$. In this case, the two estimators have inconsistencies of the same sign, but the inconsistency of $\hat{\beta}_{1r}$ is larger than that of $\hat{\beta}_{1u}$ (balanced addition).
(b) If $\psi'(\beta_2 + \tau_2)$ and τ_1 have opposite signs and $|\psi'(\beta_2 + \tau_2)| < |\tau_1|$, then $0 < \bar{b}_{1r}/\bar{b}_{1u} < 1$. In this case, the two estimators have inconsistencies of the same sign, but the inconsistency of $\hat{\beta}_{1r}$ is smaller than that of $\hat{\beta}_{1u}$ (not a balanced addition).
(c) If $\psi'(\beta_2 + \tau_2)$ and τ_1 have opposite signs and $|\tau_1| < |\psi'(\beta_2 + \tau_2)| < 2|\tau_1|$, then $-1 < \bar{b}_{1r}/\bar{b}_{1u} < 0$. In this case, the two estimators have inconsistencies of opposite sign, but the inconsistency of $\hat{\beta}_{1r}$ is smaller than that of $\hat{\beta}_{1u}$ (not a balanced addition).
(d) If $\psi'(\beta_2 + \tau_2)$ and τ_1 have opposite signs and $|\psi'(\beta_2 + \tau_2)| > 2|\tau_1|$, then $\bar{b}_{1r}/\bar{b}_{1u} < -1$. In this case, the two estimators have inconsistencies of opposite sign, but the inconsistency of $\hat{\beta}_{1r}$ is larger than that of $\hat{\beta}_{1u}$ (balanced addition).

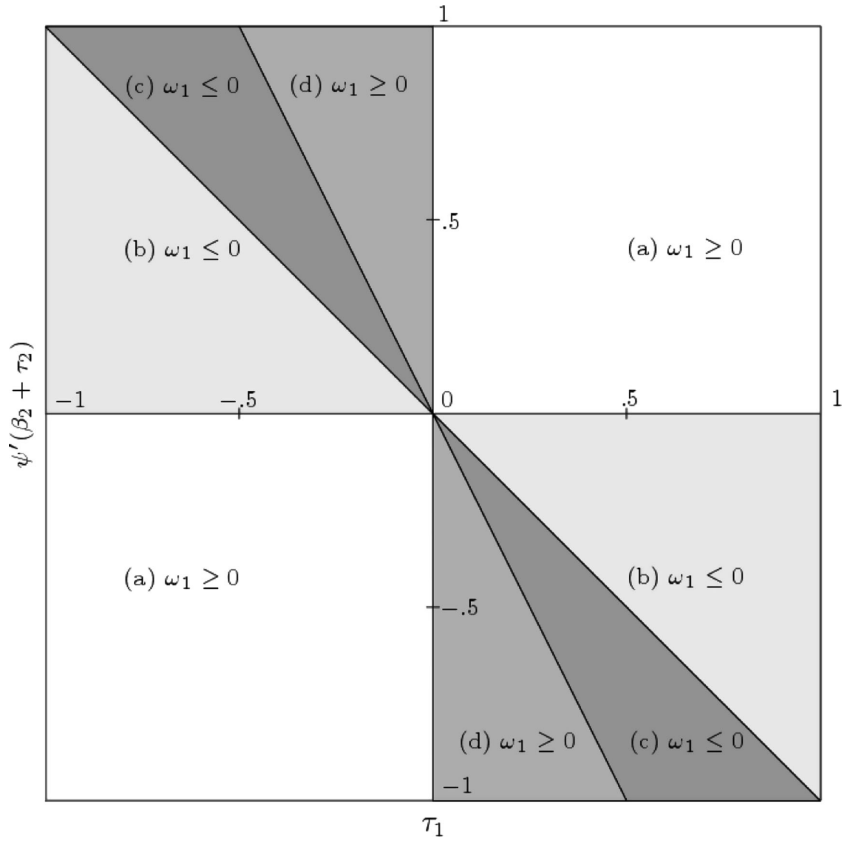

Figure 3. Values of ω_1 as a Function of τ_1 and $\psi'(\beta_2 + \tau_2)$.

We plot the four regions in Figure 3. Notice that $|\bar{b}_{1r}/\bar{b}_{1u}| = 1$ along the two lines $\tau_1 = -\psi'(\beta_2 + \tau_2)/2$ and $\psi'(\beta_2 + \tau_2) = 0$. Also notice that $\psi'(\beta_2 + \tau_2)$ and τ_1 must have opposite signs for $|\bar{b}_{1r}/\bar{b}_{1u}| < 1$, but this condition is only necessary, not sufficient. Since $\psi'(\beta_2 + \tau_2)$ can be estimated consistently by the difference $\hat{\beta}_{1r} - \hat{\beta}_{1u}$, we can assess the relative inconsistency of the two estimators of β_1 if we have prior information on the sign, and possibly the magnitude, of τ_1. For an applied researcher, a possible use of these results is in terms of sensitivity analysis. Given a consistent estimate of $\psi'(\beta_2 + \tau_2)$, the researcher may use (12) and the conditions underlying the four regions (a)–(d) to identify what magnitude and direction of τ_1 would compel her to use the restricted model for inference. It is not hard to imagine situations in which prior knowledge about the phenomenon of interest is strong enough for this to be a useful exercise.

In the above discussion (for $k_1 = 1$), we have implicitly defined the term "balanced addition" as the case where $|\bar{b}_{1u}| < |\bar{b}_{1r}|$, that is, the case where adding X_2 to the regression diminishes the bias (in absolute value). If there is more than one focus regressor in the model (i.e., $k_1 > 1$), we say that X_2 represents a balanced addition to X_1 (in terms of inconsistencies) if

$$\omega_1 = \bar{b}'_{1r}\bar{b}_{1r} - \bar{b}'_{1u}\bar{b}_{1u} \geq 0$$

We must be careful in the interpretation of ω_1 because, unlike (12), it is not invariant to scale transformations of the variables in X_1. Our results remain of course valid if we study the X_1 regressors one-by-one: for the jth component of the vector β_1 we simply replace ψ' in (12) with the jth row of the matrix Ψ. For a joint bias comparison, we need to make sure that the coefficients in β_1 are of the same order of magnitude.

4. Discussion

We now use the setup and the results in Sections 2 and 3 to connect two strands of the literature: one which considers the potential bias-reducing role of proxy variables, and one which studies bias amplification in observational studies.

The seminal paper of McCallum (1972) assumes that $y = \beta x + \gamma z + u$, where x is observable and z is unobservable but may be proxied by $p = z + e$ with $E(e) = 0$. To estimate β, the parameter of primary interest, one may choose between the short regression of y on x and the long regression of y on x and p. In the notation of Section 2, $X_1 = x$, $\beta_1 = \beta$, $X_2 = p$, $\beta_2 = 0$, and $\delta = \gamma z$. It then follows from our results in Section 3 that the inconsistencies of the unrestricted and restricted estimators of β_1 are

$$\bar{b}_{1u} = \tau_1, \qquad \bar{b}_{1r} = \tau_1 + \psi\tau_2$$

where τ_1 and τ_2 are the population slope coefficients in the linear regression of δ on X_1 and X_2, and ψ is the population slope coefficient in the linear regression of X_2 on X_1. If the proxy error e is uncorrelated with x and z, then

$$\psi\frac{\tau_2}{\tau_1} = \left(1 - \rho_{xz}^2\right)\frac{\sigma_z^2}{\sigma_e^2} > 0$$

where ρ_{xz} denotes the correlation between x and z, and σ_z^2 and σ_e^2 denote the variances of z and e. Thus, $\hat{\beta}_{1u}$ always has a smaller inconsistency than $\hat{\beta}_{1r}$. This result is special and does not extend to more realistic settings in which either X_1 and e are correlated (Frost, 1979) or more than one regressor is measured with error (Barnow, 1976; Garber and Klepper, 1980; Bekker and Wansbeek, 1996).

Wooldridge (2002, section 9.2) and Angrist and Pischke (2009, pp. 66–68) provide examples of the proxy variables setup in the specific context of estimating the returns to education from an income-generating process that depends on education, experience, and ability. Without loss of generality, we can write their DGP as

$$y = \beta E + \delta + \epsilon$$

where y is the logarithm of earnings, E is education, $\delta = \gamma A$ depends linearly on unobservable innate ability A and ϵ is mean independent of E and A. As in McCallum (1972), the issue is whether to augment the short regression of y on E with a proxy P for A (e.g., IQ scores). In our notation, $X_1 = E$, $\beta_1 = \beta$ is the parameter of primary interest, $X_2 = P$ and $\beta_2 = 0$. From our results in Section 3,

$$\bar{b}_{1u} = \alpha_1\gamma, \qquad \bar{b}_{1r} = \alpha_1\gamma + \psi\alpha_2\gamma$$

where α_1 and α_2 are the population slope coefficients in the linear regression of A on E and P, and ψ is the population slope coefficient in the linear regression of P on E.

Wooldridge (2002) considers two cases. In the "good proxy" case, E and A are uncorrelated, so $\alpha_1 = 0$ and the unrestricted estimator is consistent. In the "bad proxy" case, $\alpha_1 > 0$ and both estimators are inconsistent. As for the comparison of the two biases, everything depends on the sign and magnitude of α_2/α_1 and ψ. Wooldridge (2002) argues that: "we could still be getting an upward bias in the return to education by using P as a proxy for A if P is not a good proxy. But we can reasonably hope that this bias is smaller than if we ignored the problem of omitted ability entirely." In fact, $\psi\alpha_2\gamma$ can be estimated

consistently by $\widehat{\beta}_{1r} - \widehat{\beta}_{1u}$. Table 9.2 in Wooldridge (2002) shows that $\widehat{\beta}_{1r} - \widehat{\beta}_{1u} = 0.065 - 0.054 > 0$. Thus, in this example, augmenting the short regression with P is likely to represent a balanced addition. Since inconsistency is unavoidable when $\alpha_1 > 0$, the "bad proxy" P is not so bad.

The literature also offers examples where the unrestricted estimator does not do so well. One such case is the measurement error model of Griliches and Mason (1972) in which

$$y = \beta_1(S + Q) + \beta_2 A + \epsilon$$

where S is the observable quantity of schooling, Q is the unobservable quality of schooling and A is an error-free measure of ability. In this case, ability is observed without error but there is misspecification because Q is unobservable ($\delta = \beta_1 Q$). The issue is again whether adding A to the short regression of y on S reduces the bias in estimating β_1. Assuming that Q is uncorrelated with S implies that $\tau_1 + \psi\tau_2 = 0$, so

$$\bar{b}_{1u} = \tau_1 = -\psi\tau_2, \qquad \bar{b}_{1r} = \tau_1 + \psi(\beta_2 + \tau_2) = \psi\beta_2$$

Since ψ, τ_2, and β_2 are plausibly all positive, we have that $\bar{b}_{1u} < 0$ and $\bar{b}_{1r} > 0$. Thus, the error-free measure of ability represents a balanced addition to the short regression only if $\beta_2 > \tau_2$.

Heckman and Navarro-Lozano (2004) discuss the issue of how to choose the conditioning variables in order to identify various treatment parameters via matching methods. They consider a model in which the observed outcome is $y = y_1 d + y_0(1 - d)$, where $y_1 = \mu_1 + u_1$ and $y_0 = \mu_0 + u_0$ are the potential outcomes, μ_1 and μ_0 may depend on a vector x of exogenous regressors, $d = 1\{z'\gamma + u_V\}$ is a binary treatment indicator, z is a vector of exogenous regressors, and u_V, u_1, and u_0 are unobservable components that are correlated because they are driven by the same factors f_1 and f_2. Letting $u_V = \alpha_{V1} f_1 + \alpha_{V2} f_2 + \epsilon_V$, $u_1 = \alpha_{11} f_1 + \alpha_{12} f_2 + \epsilon_1$, and $u_0 = \alpha_{01} f_1 + \alpha_{02} f_2 + \epsilon_0$, where ϵ_V, ϵ_1, and ϵ_0 are independent of each other, gives the model

$$y = \mu_0 + \beta d + \alpha_{01} f_1 + \alpha_{02} f_2 + [(\alpha_{11} - \alpha_{01}) f_1 + (\alpha_{12} - \alpha_{02}) f_2]\, d + \epsilon$$

where $\beta = \mu_1 - \mu_0$ is the treatment parameter of interest and $\epsilon = \epsilon_0 + (\epsilon_1 - \epsilon_0)d$. Assuming that $\mu_0 = 0$ and ignoring the exogenous regressors, we can write their model in the form (1), where X_1 contains the observations on d, X_2 contains the observations on f_2 and $f_2 d$, and $\delta = \alpha_{01} f_1 + (\alpha_{11} - \alpha_{01}) f_1 d$. They show that adding f_2, but not f_1, to the exogenous regressors in x and z is not guaranteed to reduce the bias in estimating β. Wooldridge (2005) makes a similar point.

In a recent paper, Clarke *et al.* (2017) consider a special case of Heckman and Navarro-Lozano's model by restricting X_1 and X_2 to be binary. The results for this case follow immediately from our results in Section 3 after replacing the linear projection of δ on X_1 and X_2 with its linear-in-parameters conditional mean.

Pearl (2010, 2011) studies the bias-amplifying effect of adding covariates in an underspecified structural model of the form

$$y = c_0 x + c_4 z + c_2 u + \epsilon$$
$$x = c_1 u + c_3 z + \zeta^*$$

where x is a treatment variable, c_0 is the causal effect of interest, z is an "imperfect instrument" (i.e., z is uncorrelated with u and it affects y both directly and indirectly through x), and u is an unobserved confounder. It is assumed that x, u, and z each have zero mean and unit variance, and that ϵ and ζ^* satisfy the usual restrictions $E(\epsilon \mid x, z, u) = 0$ and $E(u\zeta^*) = E(z\zeta^*) = 0$. In the notation of Section 2, $X_1 = x$, $X_2 = z$, $\delta = c_2 u$, $\beta_1 = c_0$, and $\beta_2 = c_4$. The lack of correlation between z and u implies that $\bar{q}_2 = 0$, so

$$\bar{q}_1 = c_1 c_2, \qquad \bar{\Sigma}^{11} = \bar{\Sigma}^{22} = \frac{1}{1 - c_3^2}, \qquad \bar{\Sigma}^{12} = -\frac{c_3}{1 - c_3^2}$$

from which we obtain

$$\gamma = c_3, \qquad \tau_1 = \frac{c_1 c_2}{1 - c_3^2}, \qquad \tau_2 = -\frac{c_1 c_2 c_3}{1 - c_3^2}$$

and therefore

$$\bar{b}_{1u} = \frac{c_1 c_2}{1 - c_3^2}, \qquad \bar{b}_{1r} = c_1 c_2 + c_3 c_4$$

When the instrument is "perfect" (i.e., $c_4 = 0$), we have $|b_{1r}| \leq |b_{1u}|$, with strict inequality whenever $|b_{1r}| > 0$ and $c_3 \neq 0$. Thus, regardless of the signs of c_1 and c_2, augmenting the short regression of y and x with a perfect instrument amplifies the bias of the restricted estimator by a factor $1/(1 - c_3^2) > 1$. On the other hand, when z is an imperfect instrument, augmenting the short regression with z may or may not represent a balanced addition.

Finally, Oster (2016) considers a special case of our DGP (1) with X_1 scalar, X_2 and δ orthogonal, and X_1, X_2, and δ satisfying the condition

$$\varphi \, \frac{\beta_2' \Sigma_{21}}{\beta_2' \Sigma_{22} \beta_2} = \frac{\sigma_{1\delta}}{\sigma_{\delta\delta}} \tag{13}$$

for some $\varphi \geq 1$. These assumptions generalize similar assumptions introduced by Altonji *et al.* (2005), who consider the case $\varphi = 1$, but they are quite restrictive. The "proportional selection relationship" (13) is also not easy to interpret. The main result in Oster (2016) is an expression for the difference of the inconsistencies $\bar{b}_{1r} - \bar{b}_{1u}$, and for the differences of the population regression R-squares from the DGP (R_*^2), the long regression (3) and the short regression (4), in terms of the inconsistency $\bar{b}_{1u} = \tau_1$ of the unrestricted estimator of the focus parameter β_1. This expression is essentially a restatement of our result (12) under more restrictive conditions and, to be useful in practice (e.g., for sensitivity analysis), it requires preliminary information on the proportionality factor φ and the population regression R-squared R_*^2.

5. Variance and MSE

In finite samples, both bias and variance matter. When X and δ are nonrandom and the regression errors are homoskedastic, that is, $\text{var}(\epsilon) = \sigma_\epsilon^2 I_n$, the biases of $\hat{\beta}_{1r}$ and $\hat{\beta}_{1u}$ depend on δ but their variances do not. Given (7) we have $\text{var}(q) = (\sigma_\epsilon^2/n)\Sigma$, and hence

$$\text{var}\,(\hat{\beta}_{1u}) = \frac{\sigma_\epsilon^2}{n}\, \Sigma^{11}, \qquad \text{var}\,(\hat{\beta}_{1r}) = \frac{\sigma_\epsilon^2}{n}\, \Sigma_{11}^{-1} \tag{14}$$

where

$$\Sigma^{11} - \Sigma_{11}^{-1} = \Sigma_{11}^{-1} \Sigma_{12} \left(\Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \right)^{-1} \Sigma_{21} \Sigma_{11}^{-1} \geq 0$$

using (8). Thus, with or without misspecification, the restricted estimator $\hat{\beta}_{1r}$ is more precise (has smaller variance) than the unrestricted estimator $\hat{\beta}_{1u}$. This shows that we should choose the restricted estimator more frequently when we consider both bias and variance than when we only consider bias. Parsimonious modeling is thus a greater virtue than previously thought, at least with fixed regressors and homoskedastic errors.

We analyze MSE comparisons for the fixed regressor case more fully in the appendix.

These conclusions do not necessarily extend to the case of stochastic regressors (Kinal and Lahiri, 1983; Teräsvirta, 1987) or heteroskedastic errors (Hansen, 2017, pp. 209–210), for which the inclusion of additional variables in the regression may either increase or decrease the variance. In this case, we have the following result.

Proposition 3. *Assume, in addition to the assumptions in Proposition 2, that*

$$\text{var}[\sqrt{n}(q-\bar{q})] \to \bar{Q} = \begin{pmatrix} \bar{Q}_{11} & \bar{Q}_{12} \\ \bar{Q}_{21} & \bar{Q}_{22} \end{pmatrix}$$

as $n \to \infty$. *Then, the asymptotic variances of* $\hat{\beta}_{1u}$ *and* $\hat{\beta}_{1r}$ *are given by*

$$\begin{aligned} V_{1u} &= \lim \text{var}\left[\sqrt{n}\,(\hat{\beta}_{1u} - \beta_1)\right] \\ &= \bar{\Sigma}^{11}\bar{Q}_{11}\bar{\Sigma}^{11} + \bar{\Sigma}^{11}\bar{Q}_{12}\bar{\Sigma}^{21} + \bar{\Sigma}^{12}\bar{Q}_{21}\bar{\Sigma}^{11} + \bar{\Sigma}^{12}\bar{Q}_{22}\bar{\Sigma}^{21} \end{aligned}$$

and

$$V_{1r} = \lim \text{var}\left[\sqrt{n}\,(\hat{\beta}_{1r} - \beta_1)\right] = \bar{\Sigma}_{11}^{-1}\bar{Q}_{11}\bar{\Sigma}_{11}^{-1},$$

respectively.

Proof. This follows from the expressions in (9) and (10), together with the assumptions that $\text{plim}\Sigma = \bar{\Sigma}$ (Proposition 2) and the assumed limiting behavior of $q - \bar{q}$. ■

The above proposition confirms the remark following (9) and (10) at the end of Section 2 that asymptotically neither $\text{var}(\hat{\beta}_{1u})$ nor $\text{var}(\hat{\beta}_{1r})$ depends on β_2, although $\text{var}(\hat{\beta}_{1r})$ will depend on β_2 in finite samples.

From Propositions 2 and 3, we can approximate the MSEs of $\hat{\beta}_{1u}$ and $\hat{\beta}_{1r}$ by

$$\bar{b}_{1u}\bar{b}'_{1u} + V_{1u}/n, \qquad \bar{b}_{1r}\bar{b}'_{1r} + V_{1r}/n,$$

respectively, and we shall say that X_2 represents a balanced addition to X_1 (in terms of MSEs) if

$$\omega_2 = \bar{b}'_{1r}\bar{b}_{1r} - \bar{b}'_{1u}\bar{b}_{1u} + \frac{1}{n}tr(V_{1r} - V_{1u}) \geq 0$$

or, equivalently, if

$$\omega_1 \geq \frac{1}{n}tr(V_{1u} - V_{1r})$$

Now conclusions are more nuanced because the term on the right-hand side of the above inequality can be either positive or negative.

6. Conclusions

It is not generally true that adding variables to a linear regression model reduces the bias of the OLS estimator 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 need not be true when the short and the long models are both underspecified.

Our paper analyzes this situation and its implications for empirical modeling by providing a simple interpretation for the comparison of the biases of OLS estimators from two misspecified models with different sets of regressors. It also discusses variance and MSE comparisons under two alternative settings: one with fixed regressors and homoskedastic errors and one with random regressors and heteroskedastic errors.

Our results emphasize the importance of recognizing models as approximations of an unknown DGP. In practical situations where *any* estimable model is likely to be smaller than the DGP, we cannot be sure that the bias of the OLS estimator decreases when we add more variables. Parsimonious modeling is therefore a greater virtue than previously thought, especially with fixed regressors and homoskedastic errors where the inclusion of additional variables always increases the variance of the OLS estimator. In

this setting, model misspecification tilts the well-known conditions for MSE dominance in favor of the restricted estimator.

Our interpretation of the bias comparisons clarifies the prior information on the model misspecification which is needed to draw sharp conclusions on the nontrivial issue of "balanced addition." Prior information about the τ_1-vector of population coefficients can be achieved by theoretical considerations on the phenomenon under investigation, related studies based on richer sources of data, or the availability of proxy variables for the unobservable misspecification. In some circumstances, information about the signs of these population coefficients can be sufficient to assess whether an additional set of regressors can be included into the model to reduce the bias in estimating the parameters of interest. When prior information on the magnitudes of these coefficients is also needed, a sensitivity analysis may help to assess the effects of alternative assumptions about the model misspecification. Considerable theoretical work is still required to extend the existing theory about the impact of model uncertainty on inference, and we believe that the development of model-averaging techniques in an $\mathcal{M}$-open framework, where the DGP is not included in the assumed set of models, remains a challenging and important line for future research.

Acknowledgments

The authors are grateful to Eveline de Jong for providing the example in the introduction; and to Ed Leamer, seminar participants at Tilburg University (May 2016), Georgetown University (October 2016), ICEEE (January 2017), and EEA-ESEM (August 2017), the editor and two referees for constructive comments and useful suggestions. Giuseppe De Luca and Franco Peracchi also acknowledge financial support from MIUR PRIN 2015FMRE5X.

References

Altonji, J.G., Elder, T.E. and Taber, C.R. (2005) Selection on observed and unobserved variables: assessing the effectiveness of Catholic schools. *Journal of Political Economy* 113: 151–184.

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

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

Angrist, J.A. and Pischke, J.-S. (2017) Undergraduate econometrics instruction: through our classes, darkly. *Journal of Economic Perspectives* 31(2): 125–144.

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*. New York: Wiley.

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

Clarke, K.A., Kenkel, B. and Rueda, M.R. (2017) Omitted variables, countervailing effects, and the possibility of overadjustment. *Political Science Research and Methods*. https://doi.org/10.1017/psrm.2016.46.

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* (pp. 483–498). Oxford: Oxford University Press.

Frost, P.A. (1979) Proxy variables and specification bias. *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.

Griliches, Z. and Mason, W.M. (1972) Education, income, and ability. *Journal of Political Economy* 80: S74–S103.

Hansen, B.E. (2017) Econometrics. Unpublished manuscript (http://www.ssc.wisc.edu/~bhansen/).
Hausman, J.A. (1978) Specification tests in econometrics. *Econometrica* 46: 1251–1271.
Heckman, J. and Navarro-Lozano, S. (2004) Using matching, instrumental variables, and control functions to estimate economic choice models. *Review of Economics and Statistics* 86: 30–57.
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.
Kinal, T. and Lahiri, K. (1983) Specification error analysis with stochastic regressors. *Econometrica* 51: 1209–1219.
Magnus, J.R. and De Luca, G. (2016) Weighted-average least squares (WALS): a survey. *Journal of Economic Surveys* 30: 117–148.
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 Neudecker, H. (1999) *Matrix Differential Calculus with Applications in Statistics and Econometrics* (2nd edn). Chichester, England: Wiley.
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.
Oster, E. (2016) Unobservable selection and coefficient stability: theory and evidence. *Journal of Business and Economic Statistics*, https://doi.org/10.1080/07350015.2016.1227711.
Pearl, J. (2010) On a class of bias-amplifying variables that endanger effect estimates. In P. Grunwald and P. Spirtes (eds.), *Proceedings of the Twenty-Sixth Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)* (pp. 417–424). Corvallis, OR: AUAI Press.
Pearl, J. (2011) Invited commentary: understanding bias amplification. *American Journal of Epidemiology* 174: 1223–1227.
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.
Wassermann, L. (2010) *All of Statistics. A Concise Course in Statistical Inference*. New York: Springer.
Wooldridge, J.M. (2002) *Introductory Econometrics: A Modern Approach* (2nd edn). Mason, OH: South Western Educational Publishing.
Wooldridge, J.M. (2005) Violating ignorability of treatment by controlling for too many factors. *Econometric Theory* 21: 1026–1028.

Appendix A: Mean squared error (MSE) comparisons for the fixed regressor case

For the fixed regressor case, it is possible to analyze the full MSE comparison in more detail, and we do so in this appendix. We assume that $E(\epsilon) = 0$ and that $var(\epsilon) = \sigma_\epsilon^2 I_n$. It then follows from Proposition 1 and (14), or directly from (5) and (6), that

$$MSE(\hat{\beta}_{1u}) = \sigma_\epsilon^2 \left((X_1'X_1)^{-1} + (X_1'X_1)^{-1}X_1'X_2(X_2'M_1X_2)^{-1}X_2'X_1(X_1'X_1)^{-1}\right) + b_{1u}b_{1u}'$$

and

$$MSE(\hat{\beta}_{1r}) = \sigma_\epsilon^2(X_1'X_1)^{-1} + b_{1r}b_{1r}'$$

so

$$\Delta = MSE(\hat{\beta}_{1u}) - MSE(\hat{\beta}_{1r}) = \sigma_\epsilon^2(X_1'X_1)^{-1}X_1'X_2(X_2'M_1X_2)^{-1}X_2'X_1(X_1'X_1)^{-1} + b_{1u}b_{1u}' - b_{1r}b_{1r}'$$

Next we define the $k_1 \times (k_2 + 1)$ matrix

$$G = \left[\sigma_\epsilon (X_1'X_1)^{-1}X_1'X_2(X_2'M_1X_2)^{-1/2} : b_{1u}\right]$$

and the $(k_2 + 1)$-vector

$$\theta = \begin{pmatrix} \theta_1 \\ 1 \end{pmatrix}, \qquad \theta_1 = \frac{(X_2'M_1X_2)^{1/2}\beta_2 + (X_2'M_1X_2)^{-1/2}X_2'M_1\delta}{\sigma_\epsilon}$$

The key ingredient is to note that $G\theta = b_r$, the bias of the restricted estimator. The MSE difference then takes the form

$$\Delta = MSE(\hat{\beta}_{1u}) - MSE(\hat{\beta}_{1r}) = G(I_{k_2+1} - \theta\theta')G'$$

The rank of G can only take one of two values.

Proposition A1. *Let $r = rank(X_1'X_2) \geq 1$. Then the rank of G is*

$$rank(G) = \begin{cases} r, & \text{if } \delta = M_1\gamma_1 + X_2\gamma_2 \text{ for some } \gamma_1 \text{ and } \gamma_2 \\ r+1, & \text{otherwise} \end{cases}$$

Proof. It is clear that $rank(G)$ is either r or $r + 1$, and that $rank(G) = r$ if and only if b_{1u} lies in the column space of $(X_1'X_1)^{-1}X_1'X_2(X_2'M_1X_2)^{-1/2}$, that is, if and only if

$$(X_1'X_1)^{-1}X_1'\delta - (X_1'X_1)^{-1}X_1'X_2(X_2'M_1X_2)^{-1}X_2'M_1\delta = (X_1'X_1)^{-1}X_1'X_2(X_2'M_1X_2)^{-1/2}\mu$$

for some μ. This occurs if and only if

$$\delta - X_2(X_2'M_1X_2)^{-1}X_2'M_1\delta - X_2(X_2'M_1X_2)^{-1/2}\mu = M_1\gamma_1$$

for some γ_1 and μ, and hence if and only if $\delta = M_1\gamma_1 + X_2\gamma_2$ for some γ_1 and γ_2. ■

In the next two subsections, we shall distinguish the two cases $rank(G) = r + 1$ and $rank(G) = r$. We shall need the following lemma.

Lemma A1. *Let $G \neq 0$ 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 $\Delta = G(I_n - \theta\theta')G'$. Then,*

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

where A^+ denotes the Moore–Penrose inverse of A. Furthermore,

$$\Delta \geq 0 \iff \theta'G'(GG')^+G\theta \leq 1$$

$$\Delta > 0 \iff \theta'G'(GG')^{-1}G\theta < 1 \text{ and } rank(G) = m$$

$$\Delta \leq 0 \iff \theta'G'(GG')^+G\theta \geq 1 \text{ and } rank(G) = 1$$

$$\Delta < 0 \iff \theta'G'(GG')^{-1}G\theta > 1 \text{ and } m = 1$$

Proof. Let $A = GG'$, $a = G\theta$, and notice that $rank(A : a) = rank(A)$. 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. ■

A.1 *The Case When $rank(G) = r + 1$*

Let us first consider the case in which δ does not lie in the space spanned by the columns of M_1 and X_2, so that $rank(G) = r + 1$.

Proposition A2. *If* $rank(G) = r + 1$ *and letting*

$$g(\delta) = \theta' G'(GG')^{+} G\theta$$

we obtain

$$\Delta \geq 0 \iff g(\delta) \leq 1$$

$$\Delta > 0 \iff g(\delta) < 1 \text{ and } r = k_1 - 1$$

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

Proof. If $\delta \neq M_1\gamma_1 + X_2\gamma_2$, then $rank(G) = r + 1$ because of Proposition A1. The result then follows from Lemma A1. ■

Thus, when δ does not depend linearly on M_1 and X_2, the unrestricted estimator $\hat{\beta}_{1u}$ never dominates the restricted estimator $\hat{\beta}_{1r}$. Furthermore, the restricted estimator dominates the unrestricted estimator if and only if the quadratic form $g(\delta)$ is smaller than or equal to one. When does this happen? We can write

$$g(\delta) = \frac{\theta' G'(GG')^{+} G\theta}{\theta'\theta} \cdot \theta'\theta$$

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

What can we say about $g(\delta)$ in the neighborhood of $\delta = 0$? In other words, when we move from no specification to a small amount of misspecification, how sensitive is $g(\delta)$ to such a small change? Such questions are typically answered by computing the local sensitivity, that is, the derivative of $g(\delta)$ at $\delta = 0$ (Magnus and Vasnev, 2007). The function g 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. However, local sensitivity is not defined at $\delta = 0$, because the function g is not even continuous at that point. This follows because $rank(G) = r$ when $\delta = 0$, but $rank(G) = r + 1$ 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 G^{+} is discontinuous at $\delta = 0$ unless G has full column- or row-rank. What this means is that a small perturbation of δ may have a large effect on $g(\delta)$.

In the special case where $r = k_2$, the matrix G has full column-rank, so that $g(\delta) = \theta'\theta = 1 + \theta_1'\theta_1$. In that case,

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

while Δ is never positive definite.

A.2 *The Case When* $rank(G) = r$

Let us next consider the case in which δ lies in the space spanned by the columns of M_1 and X_2, so that $\delta = M_1\gamma_1 + X_2\gamma_2$ for some γ_1 and γ_2. The DGP (1) then takes the form

$$y = X_1\beta_1 + X_2(\beta_2 + \gamma_2) + M_1\gamma_1 + \epsilon$$

which shows that β_2 and γ_2 are not separately identifiable. If we impose the restriction that $\gamma_2 = 0$, then the condition $\delta = M_1\gamma_1 + X_2\gamma_2$ 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 A3. *If $X_1'\delta = 0$ then rank$(G) = r$ and, letting*

$$h(\delta) = \beta_2' X_2' X_1 (X_1' X_2 V_1 X_2' X_1)^+ X_1' X_2 \beta_2$$

with

$$V_1 = \sigma_\epsilon^2 (X_2' M_1 X_2)^{-1} + (X_2' M_1 X_2)^{-1} X_2' M_1 \delta\delta' M_1 X_2 (X_2' M_1 X_2)^{-1}$$

we obtain

$$\begin{aligned}
\Delta \geq 0 &\iff h(\delta) \leq 1 \\
\Delta > 0 &\iff h(\delta) < 1 \text{ and } r = k_1 \\
\Delta \leq 0 &\iff h(\delta) \geq 1 \text{ and } r = 1 \\
\Delta < 0 &\iff h(\delta) > 1 \text{ and } k_1 = 1
\end{aligned}$$

Proof. If $X_1'\delta = 0$, then *rank*$(G) = r$ because of Proposition A1. Also,

$$G = (X_1' X_1)^{-1} X_1' X_2 \left[\sigma_\epsilon (X_2' M_1 X_2)^{-1/2} : -(X_2' M_1 X_2)^{-1} X_2' M_1 \delta\right]$$

and

$$G\theta = (X_1' X_1)^{-1} X_1' X_2 \beta_2$$

so that

$$\Delta_1 = GG' - G\theta\theta' G' = (X_1' X_1)^{-1} X_1' X_2 (V_1 - \beta_2 \beta_2') X_2' X_1 (X_1' X_1)^{-1}$$

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 from Lemma A1. ■

When does it happen that $h(\delta) \leq 1$? First notice that V_1 is nonsingular and that its inverse is given by

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

Next, letting $W = X_1' X_2 V_1^{1/2}$, we may express $h(\delta)$ as

$$h(\delta) = \beta_2' V_1^{-1/2} W' (W W')^+ W V_1^{-1/2} \beta_2$$

A sufficient condition for $h(\delta) \leq 1$ 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

$$\lambda = \frac{\beta_2' X_2' M_1 X_2 \beta_2}{\sigma_\epsilon^2}, \qquad \lambda_\delta = \frac{\delta' M_1 X_2 \beta_2 \beta_2' X_2' M_1 \delta / \sigma_\epsilon^2}{\sigma_\epsilon^2 \left[1 + \delta' M_1 X_2 (X_2' M_1 X_2)^{-1} X_2' M_1 \delta / \sigma_\epsilon^2\right]}$$

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 (3) when the errors are normal. The condition $\lambda \leq 1$ is well-known as the condition under which the *complete* restricted estimator $(\hat{\beta}_{1r}, 0)$ of (β_1, β_2) has smaller MSE than the *complete* unrestricted estimator $(\hat{\beta}_{1u}, \hat{\beta}_{2u})$ in the special case where $\delta = 0$; see Toro-Vizcarrondo and Wallace (1968, equation (19)).

In contrast to the setup in Proposition A2, the function h is now differentiable at $\delta = 0$. This is because h depends on δ only through $M_1\delta$. The derivative of h at $\delta = 0$ vanishes, but the second derivative is nonzero, in fact negative semidefinite. Hence, h achieves a maximum at $\delta = 0$. We can see this also by noting that the fact that

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

implies that

$$(X_1'X_2V_1X_2'X_1)^+ \leq \frac{\left[X_1'X_2(X_2'M_1X_2)^{-1}X_2'X_1\right]^+}{\sigma_\epsilon^2}$$

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, Chapter 2, Miscellaneous Exercise No. 13). We again conclude that ignoring misspecification favors the unrestricted estimator, and that we should therefore, in the presence of misspecification, 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 therefore

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

When, in addition, also $X_2'\delta = 0$, 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 δ is orthogonal to both X_1 and X_2 (which is less restrictive than the textbook case $\delta = 0$), we have $h(\delta) = h_0$, where

$$h_0 = \frac{\beta_2'X_2'X_1\left[X_1'X_2(X_2'M_1X_2)^{-1}X_2'X_1\right]^+X_1'X_2\beta_2}{\sigma_\epsilon^2}$$

This special case was 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 (2016) for a survey. Of course, when also $r = k_2$ then we find again that $h(\delta) = \lambda$.

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, $GG' = var(\hat{\beta}_{1u}) - var(\hat{\beta}_{1r})$ and h_0 corresponds to the noncentrality parameter in the distribution of the so-called Hausman statistic for testing the hypothesis $\beta_2 = 0$ in model (3) with normal errors; see Hausman (1978) and Holly (1982).