Advance Access publication June 25, 2008

Political Analysis (2008) 16:345–349
doi:10.1093/pan/mpn003

Wouldn't It Be Nice . . .? The Automatic Unbiasedness of OLS (and GLS)

Robert C. Luskin
Department of Government, University of Texas, Austin, TX 78712
e-mail:rcluskin@stanford.edu

In a recent issue of this journal, Larocca (2005) makes two notable claims about the best linear unbiasedness of ordinary least squares (OLS) estimation of the linear regression model. The first, drawn from McElroy (1967), is that OLS remains best linear unbiased in the face of a particular kind of autocorrelation (constant for all pairs of observations). The second, much larger and more heterodox, is that the disturbance need not be assumed uncorrelated with the regressors for OLS to be best linear unbiased. The assumption is unnecessary, Larocca says, because ''orthogonality [of disturbance and regressors] is a property of all OLS estimates'' (p. 192). Of course OLS's being best linear unbiased still requires that the disturbance be homoskedastic and (McElroy's loophole aside) non-autocorrelated, but Larocca also adds that the same automatic orthogonality obtains for generalized least squares (GLS), which is also therefore best linear unbiased, when the disturbance is heteroskedastic or autocorrelated.

That would indeed be nice. No longer would anyone have to worry about omitted, endogenous, or imperfectly measured regressors, all producing correlation between one or more regressors and the disturbance. No longer would there be frequent need to move beyond single equations to take account of endogeneity or measurement error—nor, given linearity in the parameters, beyond OLS or GLS for estimation.

Unfortunately, however, this second claim is false. The issue on which it turns is unbiasedness, an obviously necessary condition for best linear unbiasedness. Only an unbiased estimator, by definition, can be best linear unbiased. But the conclusions familiar from the econometrics texts still apply: neither OLS nor GLS is unbiased nor therefore best linear unbiased, when the disturbance is correlated with one or more of the regressors. Neither is even consistent.

To fix notation, let $\mathbf{y}$ be an $n \times 1$ vector of the n observations on the dependent variable, $\mathbf{X}$ be an $n \times (K + 1)$ matrix of the n observations on the K regressors, led by an initial column of 1's to allow for an intercept; $\boldsymbol{\beta}$ be the corresponding $(K + 1) \times 1$ vector of coefficients, and $\mathbf{u}$ be the $n \times 1$ vector of the n observations on the disturbance, so that the regression model becomes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \tag{1}$$

Author's note: I wish to thank Christopher Achen, John Bullock, James Granato, Susan Holmes, and Mathieu Turgeon for comments. I alone am responsible for any errors.

© The Author 2008. Published by Oxford University Press on behalf of the Society for Political Methodology.
All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

https://doi.org/10.1093/pan/mpn003 Published online by Cambridge University Press

Let $\boldsymbol{\beta}$ be estimated by the $(K+1) \times 1$ vector $\hat{\boldsymbol{\beta}}$, engendering the predicted score and residual vectors $\hat{\mathbf{y}} \equiv \mathbf{X}\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{u}} \equiv \mathbf{y} = \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. Denote the scalar contents of these matrices and vectors by y_i, x_{ik}, β_k, u_i, $\hat{y}_i$, and $\hat{u}_i$, where the i and k subscripts distinguish observations and variables ($i = 1, 2, \ldots, n$; $k = 1, 2, \ldots, K$). This differs only minorly from Larocca's s notation.[1]

The key assumption for unbiasedness ($E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$) concerns the relationship between the disturbance and the regressors. The stronger version (setting aside the simplifying but generally unrealistic possibility that $\mathbf{X}$ is fixed) is that $\mathbf{u}$ is independent of $\mathbf{X}$. This plays the pivotal role in the proof of unbiasedness, although it is actually enough, as Larocca effectively notes, to assume *mean* independence: that $E(\mathbf{u}|\mathbf{X}) = E(\mathbf{u})$. OLS would still be unbiased if some *other* feature of $\mathbf{u}$'s distribution depended on $\mathbf{X}$ (although it would not be best linear unbiased if that other feature were the variance). For economy's sake, I shall write simply of "independence." A weaker version of this assumption is that $\mathbf{u}$ is merely uncorrelated (orthogonal) with $\mathbf{X}$, expressible as $\mathrm{E}(\mathbf{X}'\mathbf{u}) = \mathrm{plim}\,(1/n)\,\mathbf{X}'\mathbf{u} = 0$.[2] This, however, does not establish unbiasedness, only consistency (plim $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$), as shown below.

Of course, unbiasedness and consistency, not to mention the other finite-sample and asymptotic properties classically claimed by OLS, require additional assumptions. For unbiasedness, we also assume that $E(\mathbf{u}) = \mathbf{0}$, where $\mathbf{0}$ is an $n \times 1$ vector of 0's, and that $\rho(\mathbf{X}) = K + 1$, where ρ denotes rank (implying that $\mathbf{X}'\mathbf{X}$ is nonsingular). For consistency, we similarly assume that plim $((1/n)\mathbf{X}'\mathbf{X})$ is finite and nonsingular.[3] For best linear unbiasedness, we assume that $E(\mathbf{u}\mathbf{u}') = \sigma^2\mathbf{I}$, where the variance σ^2 is a nonzero constant, and $\mathbf{I}$ is the $n \times n$ identity matrix. Other assumptions underpin other properties. It is unnecessary, for present purposes, to review them all.

Larocca's claim is that OLS is unbiased and thus capable of being best linear unbiased even when the regressors are correlated with the disturbance. This rests on two errors, individually fatal. The first concerns the consequences of falling back, in our assumptions, from independence to orthogonality. As already noted, independence (accompanied by the additional assumptions noted) implies unbiasedness, but orthogonality (also accompanied by the additional assumptions noted) implies only consistency.

This can be seen from the matrix version of Larocca's equation (2), the formula for OLS:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} = \boldsymbol{\beta} + ((1/n)\mathbf{X}'\mathbf{X})^{-1}(1/n)\mathbf{X}'\mathbf{u},$$

which leads straightforwardly to

$$E(\hat{\boldsymbol{\beta}}|\mathrm{X}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u}|\mathbf{X}), \tag{2}$$

and to

$$\mathrm{plim}\,\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathrm{plim}((1/n)\mathbf{X}'\mathbf{X})^{-1}\mathrm{plim}(1/n)\mathbf{X}'\mathbf{u}, \tag{3}$$

since the probability limit of the product of two (matrices of) random variables is the product of their probability limits, but *not* to

[1] He reverses the order of the subscripts on the x's, uses b's rather than β's, and leaves the column of 1's and the intercept implicit, thus numbering the columns of $\mathbf{X}$ and rows of $\boldsymbol{\beta}$ only from 1 to K.

[2] The first $(K + 1) \times 1$ vector is n times the second; the assumption is that they are both zero.

[3] The assumptions about $\mathbf{X}'\mathbf{X}$ and plim $((1/n)\mathbf{X}'\mathbf{X})$ are actually necessary for identification, without which no estimator can be unbiased or consistent.

https://doi.org/10.1093/pan/mpn003 Published online by Cambridge University Press

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + E((\mathbf{X}'\mathbf{X})^{-1})E(\mathbf{X}'\mathbf{u}),$$

since the expectation of the product of two (matrices of) random variables is *not* generally the product of their expectations.

Now, if we assume that that $E(\mathbf{u}|\mathbf{X}) = E(\mathbf{u}) = \mathbf{0}$, it follows directly from equation (2) that

$$E(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta},$$

and thus that

$$E(\hat{\boldsymbol{\beta}}) = E(E(\hat{\boldsymbol{\beta}}|\mathbf{X})) = \boldsymbol{\beta}.$$

The independence of $\mathbf{u}$ and $\mathbf{X}$ implies unbiasedness. If instead we assume only that plim $(1/n)\ \mathbf{X}'\mathbf{u} = 0$ (and that plim $((1/n)\mathbf{X}'\mathbf{X})$ is finite and nonsingular), it follows directly from equation (3) that plim $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$. The orthogonality of $\mathbf{u}$ and $\mathbf{X}$ implies consistency.

But, contrary to Larocca, assuming that $E(\mathbf{X}'\mathbf{u}) = 0$ does *not* reduce $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + E((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u})$ to $\boldsymbol{\beta}$. The orthogonality of $\mathbf{u}$ and $\mathbf{X}$ does not imply unbiasedness. Since independence implies orthogonality, and nonorthogonality therefore implies dependence, it is *necessary*, just not sufficient. Larocca's position—that it is unnecessary but sufficient—is exactly wrong. So even if OLS made $\mathbf{u}$ automatically orthogonal with $\mathbf{X}$, that would not be enough to make OLS unbiased. Consistent yes, unbiased no.

That takes us to the second error, which is that OLS does not in fact make $\mathbf{u}$ automatically orthogonal with $\mathbf{X}$. It is true, as Larocca (p. 192) says, that "orthorgonality is a property of all OLS estimates." But the intrinsic orthogonality he is thinking of is of $\mathbf{X}$ with $\hat{\mathbf{u}}$, not $\mathbf{u}$. Given OLS, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, so $\hat{\mathbf{u}} \equiv \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, and

$$\mathbf{X}'\hat{\mathbf{u}} = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{0}. \tag{4}$$

Since the first element of the $(K + 1) \times 1$ vector $\mathbf{X}'\hat{\mathbf{u}}$ is $\sum \hat{u}_i$ (n times sample mean of $\hat{u}$), and the remaining K elements are $\sum x_{ik}\hat{u}_i$, where the summation is from $i = 1$ to n, and $k = 1, 2, \ldots, K$, (4) says that $\hat{u}$ necessarily has a sample mean of zero and is necessarily uncorrelated in the sample with all the x's.[4] Given OLS, this is true by construction in any sample (and thus in the population). But only of $\hat{u}$, not u. We do *assume* u to be uncorrelated with the x_k, indeed typically to be independent of them, but the point is that we have to assume it. That $\mathbf{X}'\hat{\mathbf{u}}$ automatically $= \mathbf{0}$ does absolutely nothing to establish that $E(\mathbf{X}'\mathbf{u}) = 0$.[5]

The story for GLS is essentially the same, with a couple of additional fillips. One is that GLS, in practice, is usually "feasible GLS," for which unbiasedness, and thus best linear unbiasedness, is out of the question. The difference is that in GLS, the variance-covariance matrix $\boldsymbol{\Omega} \equiv E(\mathbf{u}\mathbf{u}')$ is known, whereas in feasible GLS it must be (consistently) estimated. The GLS estimator is $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}$, the feasible GLS estimator $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\boldsymbol{\Omega}}^{-1}\mathbf{y}$, where $\hat{\boldsymbol{\Omega}}$ is a consistent estimator of $\boldsymbol{\Omega}$. If $\mathbf{u}$ is independent of $\mathbf{X}$, GLS is unbiased. Feasible GLS, however, is only consistent. But let us assume,

[4]Given that the sample mean of $\hat{u}$ is 0, $\sum x_{ik}\hat{u}_i$ is n times the sample covariance between x_k and $\hat{u}$.

[5]Of course, even assuming $E(\mathbf{X}'\mathbf{u}) = \mathbf{0}$, $\mathbf{X}'\mathbf{u} \neq \mathbf{0}$. That is, the sample (as opposed to the population) mean of the disturbance (as opposed to the residual) and its sample (as opposed to population) covariances with the regressors will not be zero. If the assumption is correct (and the sample not tiny), they will generally be close but will never be zero exactly (assuming a continuous disturbance).

https://doi.org/10.1093/pan/mpn003 Published online by Cambridge University Press

in most cases unrealistically, that $\boldsymbol{\Omega}$ is known. The other fillip is that even the *residual* $\hat{\mathbf{u}} \equiv \hat{\mathbf{u}} \equiv \mathbf{y} - \mathbf{X}\left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}$ is not automatically uncorrelated with $\mathbf{X}$. Larocca's algebra relates $\hat{\mathbf{u}}$ (which he continues to confuse with $\mathbf{u}$) to $\mathbf{X}\left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}$, not $\mathbf{X}$. Neither point affects the bottom line: The supposed orthogonality (which in this case does not obtain at all) would still be with $\hat{\mathbf{u}}$, not $\mathbf{u}$, and, even if it were with $\mathbf{u}$, would still not imply unbiasedness.

Apparently recognizing the implausibility of OLS's always being unbiased and thus, given a homoskedastic and nonautocorrelated disturbance, best linear unbiased, Larocca tries to uncouple the issues of specification and bias, at least with respect to the omission of relevant explanatory variables. Acknowledging that such omissions "will introduce bias in OLS estimates," he writes (p. 193) that "this is a question of model specification and not technically a problem that concerns the applicability of the Gauss-Markov theorem [establishing best linear unbiasedness]."

To see the problem with this stance, suppose that our model is that of (1), while the true model is instead

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \gamma\mathbf{Z} + \mathbf{e},$$

where $\mathbf{Z}$ is an $n \times 1$ vector of observations on the omitted variable z, γ is a scalar coefficient, and $\mathbf{e}$ is an $n \times 1$ vector of observations on the true disturbance e, which has a mean of 0 and is independent of $\mathbf{X}$ and $\mathbf{Z}$. Our disturbance $\mathbf{u}$, which then = $\gamma\mathbf{Z} + \mathbf{e}$, must then be correlated with $\mathbf{X}$, so long as z is correlated with one or more of the variables in $\mathbf{X}$ (part of what makes z a "relevant" omission).[6]

Larocca says that the OLS $\hat{\boldsymbol{\beta}}$ is nevertheless an unbiased estimator of $\boldsymbol{\beta}$ in (1)—that it is in fact always an unbiased estimator, no matter what our equation. The problem, on his account, is simply that (1) is the wrong equation. That *is* part of the story: (1), in this set-up, is the wrong equation. But what makes it wrong, its omission of z, necessarily produces a correlation between $\mathbf{u}$ and $\mathbf{X}$. And that, in turn, as we have seen, does bias the OLS $\hat{\boldsymbol{\beta}}$ as an estimator of $\boldsymbol{\beta}$ in (1).

The story, in short, remains the one told by the econometrics texts. Anything inducing a correlation between $\mathbf{X}$ and $\mathbf{u}$—omitted variables, endogeneity, measurement error—will bias OLS (and GLS). It will also render them inconsistent. That will in turn frequently motivate more elaborate modeling and/or alternative estimation, in the interest of reducing the bias and inconsistency. The model may add regressors or (structural or epistemic) equations. The estimation may switch to an appropriate form of instrumental variables (including two- or three-stage least squares) or maximum likelihood, for example.

Of course disturbances completely uncorrelated with the regressors exist only in our dreams. There will always be some relevant omissions, frequently some endogeneity, always some measurement error—and thus always some correlation—no matter what we do. The question, in practice, is one of degree. Small enough correlations do not leave OLS and GLS very biased or inconsistent. But larger ones do urge more elaborate models and/or alternative estimators.

[6]Note that $\mathbf{u} = \gamma(\mathbf{Z} - E(\mathbf{Z})) + \mathbf{e}$, since $E(\mathbf{u}) = E(\mathbf{e}) = \mathbf{0}$. Thus $E(\mathbf{X}'\mathbf{u}) = \gamma E(\mathbf{X}'(\mathbf{Z} - E(\mathbf{Z}))$, given $E(\mathbf{X}'\mathbf{e}) = \mathbf{0}$. $E(\mathbf{X}'(\mathbf{Z} - E(\mathbf{Z}))$ is an $n \times 1$ vector containing a 0 followed by the covariances of z with each of the x's. If any one of those covariances is nonzero, $E(\mathbf{X}'\mathbf{u}) \neq \mathbf{0}$.

https://doi.org/10.1093/pan/mpn003 Published online by Cambridge University Press

References

Larocca, Roger. 2005. Reconciling conflicting Gauss-Markov conditions in the classical linear regression model. *Political Analysis* 13(2):188–207.

McElroy, FW. 1967. A necessary and sufficient condition that ordinary least-squares estimators be best linear unbiased. *Journal of the American Statistical Association* 62:1302–04.

https://doi.org/10.1093/pan/mpn003 Published online by Cambridge University Press