Econometrica, Vol. 40, No. 4 (July, 1972)

RELATIVE ASYMPTOTIC BIAS FROM ERRORS OF OMISSION AND MEASUREMENT

BY B. T. MCCALLUM[1]

IN APPLIED ECONOMETRICS the researcher all too often discovers that observations are not obtainable on some variable that, according to economic theory, should play an important role in the relationship under investigation.[2] Consider the case in which the object is to estimate the value of some parameter other than the one attached to the missing variable. Typically the researcher will know of another variable, upon which observations *are* obtainable, that is highly correlated with the missing variable and is thus available as a "proxy." The question then naturally arises: should he use the proxy in place of the missing variable, or simply omit the latter from the equation estimated? In either case, of course, his estimators will be both biased and inconsistent. But which course of action will produce estimators with less bias or inconsistency? Intuition suggests that misestimation due to imperfect measurement—i.e., use of the proxy— will be less serious than misestimation due to no measurement at all, except in cases in which the measurement errors are large (in some sense[3]). It turns out, however, that under fairly general (and "classical") conditions, the qualifying clause in the preceding statement is unnecessary, at least for large samples. More precisely, if the measurement errors occasioned by use of the proxy are random and independent of the true regressor values, then the asymptotic bias[4] will *in all cases* be smaller if the proxy is used than if the missing variable is simply omitted. Proof of this proposition follows.

Let us begin with the two-regressor case, writing the relation of interest as

$$(1)\qquad y = \beta x + \gamma z + u,$$

where the $T \times 1$ disturbance vector[5] u has independent elements, each of which has mean 0 and variance σ_u^2. Regressors are assumed to be independent of current and previous disturbances.[6] The parameters are β and γ, the former being the object of estimation. We suppose that x is observed with full accuracy but that z is unobservable, so that the researcher must either omit this variable or else use the proxy

$$(2)\qquad p = z + e,$$

where e is a $T \times 1$ stochastic vector representing observation error. The elements of e, which have mean 0 and variance σ_e^2, are assumed independent. Also, as in the classical case, e is taken to be independent of x, z, and u.[7] Let σ_{vw} denote plim $T^{-1}v'w$ for any $T \times 1$ vectors v and w. Then our assumptions imply $\sigma_{ee} = \sigma_e^2$ and $\sigma_{xu} = \sigma_{zu} = \sigma_{xe} = \sigma_{ze} = \sigma_{ue} = 0$. Finally we assume that the regressors are "well behaved" in the limit.[8]

[1] I am indebted to Professor Zvi Griliches for suggestions that led to considerable streamlining.

[2] This relationship is here taken to be well represented by a single equation linear regression model. The unobserved variable, we also assume, is not orthogonal to the other regressors.

[3] Such as the variance of the error variable being large relative to the variance of the variable that is imperfectly measured.

[4] Under our assumption, the asymptotic bias of each estimator will coincide with its inconsistency.

[5] Which may include errors in measuring the dependent variable.

[6] Thus we permit x to be the dependent variable, lagged. Our notation is intended to be appropriate for stochastic and/or fixed regressors.

[7] See Kendall and Stuart [3, pp. 375–9] for the classical model, which also typically assumes normality. Note how unrestrictive is the assumption $Ee = 0$ [3, p. 376].

[8] See Christ [1, p. 354]. Besides non-singularity of the limiting regressor second moment matrix, this assumption implies that plim $T^{-1}x'z = \lim_{T\to\infty} ET^{-1}x'z$, etc. These equalities assure that the asymptotic bias and inconsistency of our estimators will coincide.

Now we consider the two suggested estimators of β. Let $b = x'y/x'x$ be the least squares estimator when z is omitted. We easily find[9] the asymptotic bias of b to be

$$\text{plim}\, b - \beta = \gamma \frac{\sigma_{xz}}{\sigma_{xx}}. \tag{3}$$

The second estimator $\tilde{\beta}$ comes from least squares regression of y on x and the proxy p. Since $y = \beta x + \gamma p - \gamma e + u$, we can again use specification analysis to find the asymptotic bias that results from omission of e. Using $\sigma_{ep} = \sigma_{ee}$, $\sigma_{xp} = \sigma_{xz}$, and $\sigma_{pp} = \sigma_{zz} + \sigma_{ee}$, we obtain

$$\text{plim}\, \tilde{\beta} - \beta = -\gamma\left(\frac{0 - \sigma_{xp}\sigma_{ep}}{\sigma_{xx}\sigma_{pp} - \sigma_{xp}^2}\right) = \gamma \frac{\sigma_{xz}}{\sigma_{xx}} \frac{\sigma_{ee}}{\delta}, \tag{4}$$

where $\delta \equiv \sigma_{ee} + \sigma_{zz}(1 - \sigma_{xz}^2/\sigma_{xx}\sigma_{zz})$.[10] Now clearly the asymptotic bias of $\tilde{\beta}$ will be smaller than for b if and only if $|\sigma_{ee}/\delta| < 1.0$. But this inequality holds in all cases, since $\sigma_{xz}^2/\sigma_{xx}\sigma_{zz}$ is the square of a correlation coefficient, which implies that $\delta > \sigma_{ee} > 0$.

This result can be readily generalized to the case in which $y = X\beta + \gamma z + u$, where now β is a $K \times 1$ vector and X is a matrix of observations on $K > 1$ regressors. Define $\Sigma = \text{plim}\, T^{-1}X'X$ and $\sigma_{Xz} = \text{plim}\, T^{-1}X'z$. Then the asymptotic bias vector for $b = (X'X)^{-1}X'y$ is $\gamma\Sigma^{-1}\sigma_{Xz}$—compare equation (3). Next define $X^* = [X\ p]$. Then for $\tilde{\beta}$, the first K elements of $(X^{*\prime}X^*)^{-1}X^{*\prime}y$, the generalization of (4) is

$$\text{plim}\, \tilde{\beta} - \beta = \gamma\Sigma^{-1}\sigma_{Xz}\frac{\sigma_{ee}}{\delta}, \tag{5}$$

where now $\delta \equiv \sigma_{ee} + \sigma_{zz}[1 - (\sigma_{Xz}'\Sigma^{-1}\sigma_{Xz}/\sigma_{zz})]$. The argument is as before, with $|\sigma_{ee}/\delta| < 1.0$ necessarily because the last parenthesized expression is the square of a multiple correlation coefficient.

University of Virginia

Manuscript received April, 1971; revision received May, 1971.

REFERENCES

[1] Christ, C. F.: *Econometric Models and Methods*. New York: John Wiley and Sons, 1966.
[2] Griliches, Z.: "Specification Bias in Estimates of Production Functions," *Journal of Farm Economics*, 39 (1957), 8–20.
[3] Kendall, M. G., and A. Stuart: *The Advanced Theory of Statistics*, Vol. II. London: Charles Griffin and Co., 1961.
[4] Theil, H.: "Specification Errors and the Estimation of Economic Relationships," *Review of the International Statistical Institute*, 25 (1957), 41–51.

[9] From specification error analysis of the effect of an omitted variable [**2** and **4**].
[10] In (4) the term in parentheses is the coefficient of x in the regression of e on x and p.

[Footnotes]

[9] **Specification Bias in Estimates of Production Functions**
Zvi Griliches
Journal of Farm Economics, Vol. 39, No. 1. (Feb., 1957), pp. 8-20.
Stable URL:
http://links.jstor.org/sici?sici=1071-1031%28195702%2939%3A1%3C8%3ASBIEOP%3E2.0.CO%3B2-V

References

[2] **Specification Bias in Estimates of Production Functions**
Zvi Griliches
Journal of Farm Economics, Vol. 39, No. 1. (Feb., 1957), pp. 8-20.
Stable URL:
http://links.jstor.org/sici?sici=1071-1031%28195702%2939%3A1%3C8%3ASBIEOP%3E2.0.CO%3B2-V