Notes on Bias in Estimators for Simultaneous Equation Models

Jinyong Hahn and Jerry Hausman[1]

June, 2001

We begin with the simplest model specification with one right hand side (RHS) jointly endogenous variable so that the left hand side variable (LHS) depends only on the single jointly endogenous RHS variable. This model specification accounts for other RHS predetermined (or exogenous) variables, which have been "partialled out" of the specification.[2] We will assume that

(1) $$y_1 = by_2 + e_1 = bzp_2 + v_1$$

(2) $$y_2 = zp_2 + v_2,$$

where $\dim(p_2) = K$. Thus, the matrix z is the matrix of all predetermined variables, and equation (2) is the reduced form equation for y_2 with coefficient vector p_2. We also assume homoscedasticity:

(3) $$\begin{pmatrix} v_{1i} \\ v_{2i} \end{pmatrix} \sim N(0,\Omega) \sim N\left(0, \begin{bmatrix} w_{11} & w_{12} \\ w_{12} & w_{22} \end{bmatrix}\right).$$

We use the following notation:

$$y \equiv \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad z \equiv \begin{pmatrix} z_1' \\ \vdots \\ z_n' \end{pmatrix}, \quad s_e^2 \equiv \mathrm{Var}(e_{1i}), \quad s_{ev_2} \equiv \mathrm{Cov}(e_{1i}, v_{2i}), \quad s_{ev_1} \equiv \mathrm{Cov}(e_{1i}, v_{1i}).$$

I. Bias in 2SLS and OLS

A common finding in empirical research is that when 2SLS is used the coefficient estimate increases in magnitude from the OLS estimate. However, in finite samples under certain situations even when 2SLS is used on equation (1), bias remains because an

---

[1] Brown University and MIT. Please send comments to jhausman@mit.edu

estimate of p_2 from equation (2) is used, since the true parameters are unknown. We now demonstrate how this result occurs.

Suppose that zp_2 is measured without error. Then, OLS of y_1 on zp_2 would be unbiased. Instead, zp_2 must be estimated, *i.e.*, we have to rely on 2SLS. Let $\hat{p}_2$ denote the first stage OLS estimator. We have

$$(4) \qquad b_{2SLS} - b = \frac{\sum_{i=1}^{n}\left(v_{1i} - bz_i' \cdot (\hat{p}_2 - p_2)\right) \cdot z_i'\hat{p}_2}{\sum_{i=1}^{n}(z_i'\hat{p}_2)^2} = \frac{\sum_{i=1}^{n}\left(v_{1i} - bz_i' \cdot (\hat{p}_2 - p_2)\right) \cdot z_i'\hat{p}_2}{R_f^2 \cdot \sum_{i=1}^{n} y_{2i}^2},$$

where R_f^2 is the R^2 in the first stage regression to obtain $\hat{p}_2$. It can be shown that:

$$E\left[\sum_{i=1}^{n}\left(v_{1i} - bz_i' \cdot (\hat{p}_2 - p_2)\right) \cdot z_i'\hat{p}_2\right] = K \cdot s_{ev_2}, \text{ and } E\left[\sum_{i=1}^{n}(z_i'\hat{p}_2)^2\right] = n \cdot p_2' R p_2 + K \cdot s_{v_2 v_2},$$

where $R \equiv E[z'z/n]$. Here, $K \cdot s_{v_2 v_2}$ is the expectation of the term $v_2' P_z v_2$, which is $s_{v_2 v_2}$ times a χ^2- random variable with expectation equal to the dimension of the projection matrix P_z.

Therefore, we expect bias approximately equal to

$$(5) \qquad E[b_{2SLS}] - b \approx \frac{K \cdot s_{ev_2}}{R_f^2} \frac{1}{\sum_{i=1}^{n} y_{2i}^2} \approx \frac{K s_{eu_2}}{n \cdot p_2' R p_2 + K \cdot s_{v_2 v_2}}$$

Equation (5) indicates that the bias is monotonically increasing in s_{ev_2} and K, but monotonically decreasing in R_f^2. Conventional asymptotics, which lets $n \to \infty$ keeping DGP fixed, ignores the influence of s_{ev_2}, K, and R_f^2.

For comparison purposes we calculate the bias of OLS. We find approximately that

$$(6) \qquad E[b_{OLS}] - b \approx \frac{\mathrm{Cov}(y_2, e)}{\mathrm{Var}(y_2)} \approx \frac{s_{eu_2}}{p_2' R p_2 + s_{v_2 v_2}}$$

[2] See Hahn and Hausman (1999), available at http://web.mit.edu/jhausman/www/, for more complete

Note that in equation (5) the denominator typically becomes large as the sample size *n* becomes large so that the bias of 2SLS decreases. However, in the OLS bias equation (6) the denominator does not change size as *n* increases so that the bias does not decrease. Thus, 2SLS is consistent and OLS is inconsistent, as is well known.

II. <u>No Identification</u>

We now use equation (5) to explore what happens in the unidentified situation of $p_2 = 0$. The denominator of equation (5) becomes $K \cdot s_{v_2 v_2}$. Thus, when $p_2 = 0$, equation (5) predicts the bias of the 2SLS estimator to be approximately

$$(7) \qquad E[b_{2SLS}] - b \approx \frac{s_{ev_2}}{s_{v_2 v_2}}$$

In large samples the result holds in the limit without the necessity of assuming that the stochastic disturbances are normal. Note that the bias does not decrease here as *n* becomes large as it did in the last section. This result is expected because without identification we cannot find a consistent estimator of beta.

We now compare this 2SLS bias with the bias of OLS on equation (1) again where no identification exists so that $p_2 = 0$. We use equation (6). When $p_2 = 0$, the denominator is equal to $s_{v_2 v_2}$. Thus, we find that the bias of OLS is the same as the bias of 2SLS in the unidentified case of $p_2 = 0$:

$$(8) \qquad E[b_{OLS}] - b \approx \frac{s_{ev_2}}{s_{v_2 v_2}}$$

---

discussion of models with other RHS predetermined variables.

See Phillips (1989) for related results.

III. Local Non-Identification

We now consider what happens when we are close to being unidentified so that $p_2 = a/\sqrt{n}$, where the vector a has dimension *K*. Thus, the reduced form coefficients are "local to zero". Stock and Staiger (1997) refer to this situation as "weak instruments". We disagree somewhat with this terminology because the result of badly biased IV estimators also depends on the value of covariance term in the numerator of equation (5) as we discuss in Hahn and Hausman (1999).

With $p_2 = a/\sqrt{n}$, equation (5) predicts the bias of 2SLS to be

(9) $$E[b_{2SLS}] - b \approx \frac{K s_{ev_2}}{a'Ra + K s_{v_2 v_2}} = \frac{s_{ev_2}}{\frac{1}{K} a'Ra + s_{v_2 v_2}}$$

Equation (9) is an approximation to the asymptotic bias of 2SLS under the asymptotics where $p_2 = a/\sqrt{n}$. When *K* is sufficiently large, the difference between equation (9) and the asymptotic bias is negligible. See Chao and Swanson (2000, Theorem 3.1 (c)). On the other hand, equation (6) predicts the approximate bias for OLS to be:

(10) $$E[b_{OLS}] - b \approx \frac{s_{eu_2}}{\frac{1}{n} a'Ra + s_{v_2 v_2}}$$

Comparison of (9) and (10) suggest that the bias of 2SLS is smaller than OLS as long as *K*<*n*, a condition which will always be satisfied in practice.

We have considered three asymptotic approximation: (i) $p_2 \neq 0$ and fixed; (ii) $p_2 = a/\sqrt{n}$, $a \neq 0$; (iii) $p_2 = 0$. For the first two cases, our approximate bias formulae predict that 2SLS has less bias than OLS. For the last case, our formulae predict that 2SLS has approximately equal bias as OLS.

IV. Bias Corrected 2SLS

We can also use equation (5) to construct an approximately unbiased 2SLS estimator. While it first appears that we have only one equation (moment) and two unknowns in b and s_{ev_2}, it turns out that this second parameter is a function of beta:

$$s_{ev_2} = E\left[\frac{1}{N-K}(y_2'Q_z)(y_1 - y_2 b)\right] \quad \text{where} \quad Q_z \equiv I - P_z.$$

Now we can solve for b which is a linear equation. The derivation is:

$$E\left[\hat{b} - b\right] \approx \frac{K s_{ev_2}}{y_2' P_z y_2} \approx E\left[\frac{M}{N-K}(y_2'Q_Z)(y_1 - y_2 b)\right] = E\left[dq'(y_1 - y_2 b)\right]$$

where $M = K / y_2'P_z y_2$, $d = M/(N-K)$ and $q' = y_2'Q_z$. Thus we can solve for beta to find a bias corrected estimator b_{BC}:

$$\hat{b}_{BC} = \left(\hat{b} - dq'y_1\right)/\left(1 - dq'y_2\right)$$

If we now consider the (approximate) bias of the estimator we find it to be zero by construction. Thus, the estimator is approximately unbiased as claimed.

This estimator turns out to be the same as Nagar’s estimator (EMA 1959), which was derived in a considerably more difficult manner using a higher order expansion approach. This equivalence can be seen from:

$$\hat{b}_{BC} = \hat{b}_N = \frac{\dfrac{y_2'P_z y_1}{y_2'P_z y_2} - \dfrac{K}{N-K}\dfrac{y_1'Q_z y_2}{y_2'P_z y_2}}{1 - \dfrac{K}{N-K}\dfrac{y_2'Q_z y_2}{y_2'P_z y_2}} = \frac{y_2'P_z y_1 - \dfrac{K}{N-K} y_1'Q_z y_2}{y_2'P_z y_2 - \dfrac{K}{N-K} y_2'Q_z y_2}$$

Unfortunately, the estimator has no moments, and performs poorly when the model is nearly non-identified.[3] This poor performance follows by noting that the denominator is zero when $p_2 = 0.$ The Nagar estimator "blows up" in this situation in contrast to the 2SLS estimator, which is inconsistent but has its moments existing. For near non-identification, the Nagar estimator similarly works poorly because the non-existence of moments from the denominator being near zero leads to poor results in many situations. Hahn, Hausman, and Kuersteiner (2001) give Monte Carlo results that demonstrate the poor performance of the Nagar estimator in this situation. Thus, the Nagar estimator is not very useful in the situation where 2SLS has substantial bias. Hahn, Hausman, and Kuersteiner (2001) explore alternative estimators to use in this situation.

---

[3] See Sawa (1972).

References

CHAO, J., AND N. SWANSON (2000): “Bias and MSE of the IV Estimator under Weak Identification”, *unpublished working paper*.

HAHN, J., AND J. HAUSMAN (1999): “A New Specification Test for the Validity of Instrumental Variables”, forthcoming in *Econometrica*.

HAHN, J., J. HAUSMAN, AND G. KUERSTEINER (2001): “Higher Order MSE of Jackknife 2SLS”, *unpublished working paper*.

NAGAR, A.L. (1959): “The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations,” *Econometrica*, 27, 575 – 595.

PHILLIPS, P.C.B. (1989): “Partially Identified Econometric Models”, *Econometric Theory*, 5, 181-240.

SAWA, T. (1972): “Finite Sample Properties of the k-Class Estimators,” *Econometrica*, 40, 653-680.