Model Specification in Instrumental-Variables Regression: Simulation Results

Thad Dunning

Department of Political Science, Yale University

Box 208301, New Haven, CT 06520

email: thad.dunning@yale.edu

This document reports on simulation results referenced in the article "Model Specification in Instrumental-Variables Regression," published in *Political Analysis*.

1 Simulation results

Each simulation draws 1000 independent samples of size $n = 250$.[1] As in section 4 of the paper, the true data-generating process is

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i, \tag{1}$$

for each observation i, where β_1 and β_2 are parameters. The subjects are i.i.d., and $\mathrm{E}(\epsilon_i) = \mathrm{E}(X_{1i}) = \mathrm{E}(X_{2i}) = 0$. In addition, $(X_{1i}, X_{2i}, \epsilon_i)$ is trivariate normal and has the following variance-covariance matrix:

$$\begin{pmatrix} 1.0 & d & c \\ d & 1.0 & 0.0 \\ c & 0.0 & 1.0 \end{pmatrix} \tag{2}$$

For instance, $\mathrm{Var}(X_{1i}) = \mathrm{Var}(X_{2i}) = \mathrm{Var}(\epsilon_i) = 1.0$. The parameters c and d represent $\mathrm{Cov}(X_{1i}, \epsilon_i)$ and $\mathrm{Cov}(X_{1i}, X_{2i})$, respectively, with $-1 \leq c, d \leq 1$. Finally, note that

$$X_{2i} \perp\!\!\!\perp \epsilon_i, \tag{3}$$

so X_{2i} is exogenous; the independence of X_{2i} and ϵ_i follows from $\mathrm{Cov}(X_{2i}, \epsilon_i) = 0$ and joint normality.

Suppose we assume that data were generated according to

$$Y_i = \beta(X_{\mathrm{T}i}) + \epsilon_i, \tag{4}$$

where $X_{\mathrm{T}i} \equiv X_{1i} + X_{2i}$. The tables below report what IVLS estimates under different assumptions about the correlation between X_{1i} and X_{2i}. I consider two different cases: (i) $\beta_1 \neq \beta_2$, so equation

[1] All reported simulations were conducted in Intercooled Stata 9.2. The .do files are available upon request from the author.

(4) is misspecified, as in section 4 of the text; and (*ii*) $\beta_1 = \beta_2$, so the assumed model is correct. It may be instructive to compare IVLS estimates of equation (4) to OLS estimates, so I report the latter as well.

1.1 Case 1: X_{1i} and X_{2i} are Independent

In the two simulations reported in Table 1, β_1 is held constant at 1.0; $\beta_2 = 1.0$ in one simulation, and $\beta_2 = 2.0$ in the other. Here, $c = 0.3$, so X_{1i} is endogenous. Also, $d = 0$, so X_{1i} and X_{2i} are independent. Thus, the structure of the simulation is like the example on political attitudes and lottery winnings in the paper.

The first two columns of Table 1 report $\bar{\beta}_{\text{IVLS}}$ (the average of $\hat{\beta}_{\text{IVLS}}$ over the 1000 replications) and sd_{IVLS} (the standard deviation of $\hat{\beta}_{\text{IVLS}}$ over the 1000 replicates) for each of the two true values of β_2. The final two columns report the analogous quantities for the OLS estimator, $\bar{\beta}_{\text{OLS}}$ and sd_{OLS}.

There are two key results of interest. First, as shown analytically in section 4 of the text, IVLS estimates β_2 (and not β_1 or some mixture of β_1 and β_2). For instance, when β_2 is set at 2.0 (first row), $\bar{\beta}_{\text{IVLS}} = 2.0028$ (and $\text{sd}_{\text{IVLS}} = 0.0749$). When β_2 is set at 1.0 (second row), $\bar{\beta}_{\text{IVLS}} = 1.0010$ (and sd_{IVLS} is 0.0650).

Second, OLS estimates an average of β_1 and β_2, weighted by the correlation between X_{1i} and the error term. In the first simulation, with $\beta_1 = 1.0$ and $\beta_2 = 2.0$, $\bar{\beta}_{\text{OLS}} = 1.6518$, with $\text{sd}_{\text{OLS}} = 0.0478$ (first row, Table 1). Notice that OLS here comes closer to the true value of β_1 than IVLS, since it gives an estimate that lies between the true values of β_1 and β_2. However, the estimate is pulled above the average of the true coefficients (that is, above $\frac{\beta_1+\beta_2}{2} = 1.5$), due to the positive correlation between X_{Ti} and the error term. In the second simulation, with $\beta_1 = \beta_2 = 1$, $\bar{\beta}_{\text{OLS}} = 1.1501$ and $\text{sd}_{\text{OLS}} = 0.0437$; here, the endogeneity bias pulls OLS away from the true value of β_1 (as well as away from β_2).

Table 1: Simulation results. Investigating the IVLS and OLS estimators when X_{1i} and $X_2 i$ are independent. Here, $\beta_1 = 1.0$, $d = 0.3$, and $c = 0$.

	$\bar{\beta}_{\text{IVLS}}$	sd_{IVLS}	$\bar{\beta}_{\text{OLS}}$	sd_{OLS}
$\beta_2 = 2.0$	2.0028	0.0749	1.6518	0.0478
$\beta_2 = 1.0$	1.0010	0.0650	1.1501	0.0437

1.2 Case 2: X_{1i} and X_{2i} Are Correlated

I now investigate the performance of the estimators when the components of $X_{\text{T}i}$ are correlated, rather than independent as in the simulations above. Here, Corr(X_{1i}, X_{2i})=0.4. Other parameters are as in Table 1.

As per the analytic results, Table 2 suggests that when the components of $X_{\text{T}i}$ are correlated, $\hat{\beta}_{\text{IVLS}}$ estimates a mixture of β_1 and β_2. For instance, in the first simulation, $\bar{\beta}_{IVLS} = 1.7150$. So does OLS, though the positive relationship between X_{1i} and the error term induces an upward bias in the estimates. Other simulations were run, with similar results.

In short, the simulations give the same message as the analytic results in Section 4. When the true data-generating process involves different coefficients for different components of the treatment variable X_i, and we have assume that these components have the same coefficients, IVLS and OLS both estimate a data-dependent mixture of the structural parameters. This mixture may not be the quantity of interest. The simulation results, like the analytic discussion in the text, therefore underscore the key role played by model specification: exogeneity of the instruments, given the model, is necessary but not sufficient for valid application of IVLS or any other estimation strategy. The underlying model must be correct.

Table 2: Simulation results. Investigating the IVLS and OLS estimators when X_{1i} and X_{2i} are correlated at 0.4. As in Table 1, $\beta_1 = 1.0$, $d = 0.3$, and $c = 0$.

	$\bar{\beta}_{\text{IVLS}}$	sd_{IVLS}	$\bar{\beta}_{\text{OLS}}$	sd_{OLS}
$\beta_2 = 2.0$	1.7150	0.0497	1.5761	0.0493
$\beta_2 = 1.0$	0.9998	0.0456	1.0765	0.0436