

An Introduction to Instrumental Variables Using Simulations

Aratrika Patnaik

Isha BK*

Abstract

Data often show that two variables move together, but this does not prove that one causes the other. This tutorial introduces instrumental variables (IV) to readers new to econometrics. It explains why ordinary least squares (OLS) gives a biased estimate when a hidden factor drives both the explanatory variable and the outcome, and how IV removes this bias. A simple simulation, in which the true effect is known, makes these ideas concrete. It shows that OLS stays biased however much data we collect, that IV recovers the true effect when the sample is large and the instrument is strong, and that IV pays for this with less precise estimates. These results are well known. Our aim is to make them easy to see.

Keywords: instrumental variables, endogeneity, two-stage least squares, Monte Carlo simulation

1. Introduction

Many questions in economics are about cause and effect. Does more schooling raise wages? People with more schooling do earn more on average, but part of this gap may come from ability rather than schooling itself. A more able person may stay in school longer and also earn more at work. If ability is not measured, a simple comparison mixes the two effects and overstates the value of schooling.

Economists call this problem *endogeneity*, and instrumental variables (IV) is one of the main tools for dealing with it. This tutorial explains the method for readers who are new to econometrics. We assume no background beyond the idea of fitting a line through data. We do not offer new results. Instead, we use a computer simulation to show the main ideas at work. Because we generate the data ourselves, we know the true effect and can see directly which method recovers it.

The simulation illustrates three standard lessons. First, ordinary least squares (OLS) is biased when a hidden factor drives both variables, and more data does not help. Second, IV removes most of this bias, but only when the sample is large enough and the instrument is strong enough. Third, IV estimates are less precise than OLS estimates. Readers who want a fuller treatment can turn to Angrist and Pischke [1]. Section 2 sets out the problem and the method, Section 3 describes the simulation, Section 4 walks through the results, and Section 5 concludes.

*This tutorial was written under the guidance of Dr. Parush Arora, Ashoka University.

2. Endogeneity and Instrumental Variables

OLS estimates the effect β of an explanatory variable X on an outcome Y by fitting

$$Y_i = \alpha + \beta X_i + e_i, \tag{1}$$

where i indexes individuals and the error term e_i collects everything else that affects Y . OLS recovers β only if X is unrelated to e_i . When it is related, X is *endogenous* and the OLS estimate is *biased*, meaning it is wrong on average however much data we collect. Endogeneity usually comes from an omitted variable that affects both X and Y , such as ability in the schooling example, or from reverse causality, where Y also affects X . The bias grows with how strongly the hidden factor affects X and Y , and in bad cases it can hide a real effect or flip the sign of the estimate.

An *instrument* Z is a variable that shifts X but affects Y only through X . A well-known example is the distance from a student's home to the nearest college [2]. Students who live near a college tend to get more schooling, because attending is cheaper. If distance has no other link to wages, any wage gap between near and far students must come through schooling. The IV estimate is then a ratio,

$$\hat{\beta}_{IV} = \frac{\text{effect of } Z \text{ on } Y}{\text{effect of } Z \text{ on } X}. \tag{2}$$

For illustration, suppose living near a college raises schooling by half a year and monthly wages by 500 rupees. Each extra year of schooling is then worth $500/0.5 = 1000$ rupees per month. Ability plays no part in this calculation, because distance is unrelated to ability.

A valid instrument must meet two conditions. The first is *relevance*, meaning Z must actually move X . This can be checked with the first-stage F -statistic, and a common rule of thumb asks for a value above 10 [3]. An instrument that barely moves X is called *weak*, and weak instruments give noisy estimates that drift back toward the OLS answer. The second is *exogeneity*, meaning Z must be unrelated to the error term. This cannot be tested, because the error term is unobserved, so it must be argued from knowledge of the setting. If families who value education choose to live near colleges, for instance, distance would be linked to wages in other ways and the condition would fail.

In practice IV is computed by two-stage least squares (2SLS). The first stage regresses X on Z and saves the predicted values $\hat{X}_i$, which contain only the part of schooling driven by distance. The second stage regresses Y on $\hat{X}_i$, and the resulting slope is the IV estimate. With a single instrument this equals the ratio in equation (2) [1].

3. Simulation Design

With real data the true β is never known, so we cannot tell whether an estimate is right. We therefore generate data with a known β and compare the two methods. Schooling and wages are

generated as

$$X_i = \pi Z_i + \gamma U_i + \eta_i, \quad \eta_i \sim \mathcal{N}(0, 1), \quad (3)$$

$$Y_i = \beta X_i + \delta U_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad (4)$$

where X_i is years of schooling, Y_i is the wage, Z_i is distance to college, and U_i is a hidden factor such as ability. The parameter π sets the strength of the instrument, γ and δ set how strongly U affects schooling and wages, and σ sets the noise in wages. Because U enters both equations, OLS is biased whenever γ and δ are not zero. The instrument is drawn separately from U and is left out of the wage equation, so it meets both conditions by construction.

For each setting we draw 100 samples, estimate β by OLS and by IV in each, and record the *bias*, which is the average estimate minus the true β . We also record the variance of the estimates across samples, the mean squared error, and the first-stage F -statistic. We change one setting at a time, namely the sample size n , the instrument strength π , the degree of endogeneity, and the noise level σ .

4. Results

Figure 1 reports the bias of each method. Panels (a) and (b) show that sample size and instrument strength matter a great deal for IV but not at all for OLS. OLS bias stays near 2 whatever the sample size, so more data does not fix it. IV bias falls from about 1.8 with 100 observations to about 0.4 with 1,000 and is close to zero at 5,000 and above. It falls in the same way as the instrument gets stronger, from about 1.0 when $\pi = 0.01$ to almost zero when $\pi = 0.1$. A small sample or a weak instrument leaves IV biased toward the OLS answer even when the instrument is valid.

Panels (c) and (d) make the test harder by holding the sample size and instrument strength low. As endogeneity doubles, OLS bias roughly doubles too, rising from about 1 to about 8, while IV bias stays below about 0.3. Raising the noise level leaves OLS bias near 2 and IV bias below about 0.2. Noise makes wages more random, but it is unrelated to schooling and to the instrument, so it adds no bias to either method. The OLS bias in panel (d) comes from the hidden factor, not from the noise.

IV has a cost. OLS uses all of the variation in schooling, while IV uses only the part driven by the instrument, which is a much smaller slice. IV estimates therefore vary more from sample to sample. OLS is precise but wrong, and IV is right on average but less precise. A larger sample and a stronger instrument narrow this gap.

These results come with limits. With 100 repetitions some numbers move around by chance. The model has one hidden factor and normal errors, while real data have many hidden factors and messier shocks. Most importantly, our instrument is valid by construction. In real research exogeneity cannot be tested, so the simulation shows how IV behaves when its conditions hold, not how it performs on any real data set.

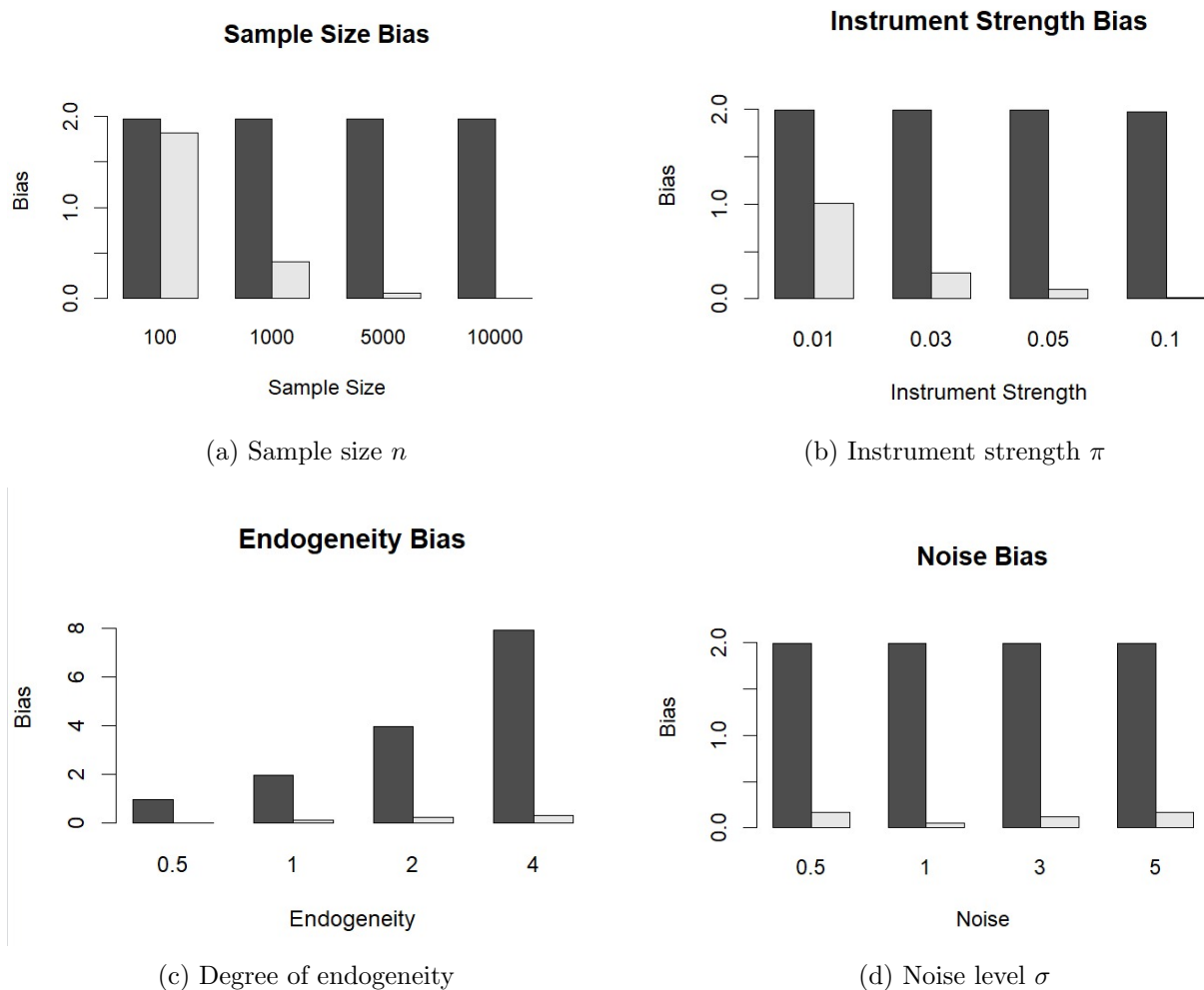


Figure 1: Bias of OLS (dark bars) and IV (light bars) across 100 simulated samples. Each panel varies one setting and holds the others fixed. In panels (c) and (d), sample size and instrument strength are held near the smallest values at which IV performs well.

5. Conclusion

This tutorial used a simple simulation to show how instrumental variables work. When a hidden factor affects both the explanatory variable and the outcome, OLS gives biased estimates and more data does not help. IV removes this bias by using only the variation that comes from an instrument. In the simulation, IV stayed close to the true effect when the sample was large and the instrument was strong, at the cost of lower precision.

The main lesson for applied work is that IV is only as good as its instrument. Relevance can be checked with the first-stage F -statistic, but exogeneity must be defended by reasoning about the setting. A perfect instrument is easy to build in a simulation and hard to find in real life. Readers can extend the exercise by giving the instrument a direct effect on wages, which breaks exogeneity, and seeing how far the IV estimate moves from the truth.

References

- [1] Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- [2] Card, D. (1995). Using geographic variation in college proximity to estimate the return to schooling. In L. N. Christofides, E. K. Grant, and R. Swidinsky (Eds.), *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, pp. 201–222. University of Toronto Press.
- [3] Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3), 557–586.