Instrumental Variables

Ethan Ligon

April 10, 2023

Roadmap

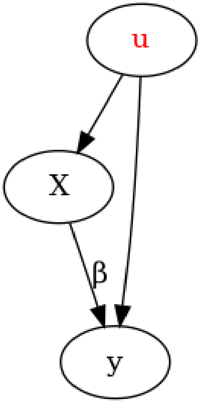
Here are some different forms of "estimating equations" we'll encounter, built around moment restrictions:

	Parametric	Non-Parametric
Separable	$\mathbb{E}[y - f(X, \beta) \mid Z] = 0$	$\mathbb{E}[y - g(X) \mid Z] = 0$
Non-separable	$\mathbb{E}[f(y, X, \beta, \epsilon) \mid Z] = 0$	$\mathbb{E}[g(y, X, \epsilon) \mid Z] = 0$

All of these can be thought of as estimation problems involving instrumental variables. It's also worth noting that there's a practical sense in which they all involve an *infinite* number of instruments.

Linear Models with Endogenous RHS Variables

We earlier considered the canonical demand & supply model, which features the equation $q = \mu + \alpha p(u, v) + u$; in this model the RHS variable p is a function of both demand and supply shocks (u, v).



Model Equation

$$y = \boldsymbol{X}\beta + u$$

Regression Equation

$$\boldsymbol{y} = \boldsymbol{X}b + \boldsymbol{e}$$

Identification via Instrumental Variables

Wright's solution to the identification failure of the demand and supply model with the linear regression model was to find an *instrument* $\boldsymbol{Z}$ that he thought was correlated with supply shocks, but not with demand. (The term "instrument" is apparently due to Frisch.)

Requirements for "valid" instruments

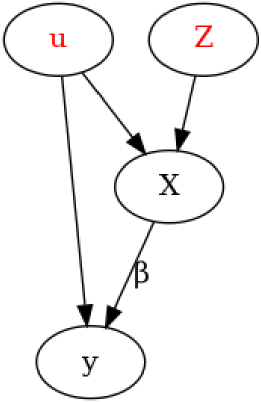
There are two requirements in the linear case:

Orthogonality $\mathbb{E}\boldsymbol{Z}^\top\boldsymbol{u} = 0$;

Relevance $\mathbb{E}\boldsymbol{Z}^\top\boldsymbol{X} = \boldsymbol{Q}$, where $\boldsymbol{Q}$ has full column rank.

Linear Models with Endogenous RHS Variables

We earlier considered the canonical demand & supply model, which features the equation $q = \mu + \alpha p(u, v) + u$; in this model the RHS variable p is a function of both demand and supply shocks (u, v).



Model Equation

$$y = \boldsymbol{X}\beta + u$$

$$\mathbb{E}(u|Z) = 0$$

IV Regression

$$\boldsymbol{y} = \boldsymbol{X}b + \boldsymbol{e}$$

Special Case of General Linear Model

With these two assumptions, the linear IV estimator becomes a special case of the general linear model, with $\boldsymbol{T} = (\boldsymbol{Z}^\top \boldsymbol{X})^+ \boldsymbol{Z}^\top$, which we've already solved. However, we'd like to consider the limits of the orthogonality & relevance requirements:

Weak instruments

The matrix $\boldsymbol{Q} = \boldsymbol{Z}^\top \boldsymbol{X}$ may formally satisfy the relevance (rank) condition, but still be *nearly* rank-deficient.

"Plausibly exogenous"

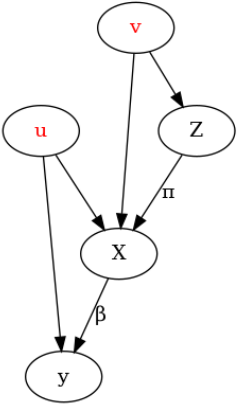
What if there's linear dependence between our instruments and e, but this dependence is small (in some sense)?

Inference in Finite Samples

In finite samples the distribution of the IV estimator can be very poorly behaved.

Expanded Linear Specification

Consider the following expanded specification; though it *looks* more complicated it's actually a special case of our earlier linear IV model.



Model Equation

$$y = \boldsymbol{X}\beta + \boldsymbol{Z}\gamma + u$$
$$\boldsymbol{X} = \boldsymbol{Z}\pi + v$$

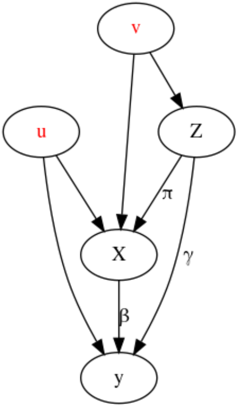
IV Regression

$$\boldsymbol{Z}^\top \boldsymbol{y} = \boldsymbol{Z}^\top \boldsymbol{X} b + \boldsymbol{Z}^\top \boldsymbol{e}$$

Weak Instruments

This can happen either because the covariance between variables in X and Z is small, or because the number of variables in Z is large.

Case in which we allow structured violations of $\mathbb{E}(u|Z) = 0$; leads to “set identification.” (No longer point identified.)



Model Equation

$$y = \boldsymbol{X}\beta + \boldsymbol{Z}\gamma + u$$
$$\boldsymbol{X} = \boldsymbol{Z}\pi + v$$
$$\gamma \in \Gamma$$

Inference in Finite Samples

A leading empirical case is just-identified linear IV. Though asymptotically this estimator is $\sqrt{N}$ consistent and asymptotically normal, it may have *terrible* properties in finite samples. In fact, not only is the estimator not-unbiased; in the normal homoskedastic case its expectation doesn't even exist!

Handwaving and the plim

In the standard case with x and z both mean zero scalar random variables we obtain

$$\operatorname{plim} b_{IV} = \beta + \frac{\mathbb{E} zu}{\mathbb{E} zx}.$$

Overidentification

Existence of Moments

In a result due to Kinal (1980), if we have ℓ instruments and k parameters, with disturbance normally distributed:
The number of moments of b_{IV} which *exist* is $\ell - k$.

Overidentified Case

In the overidentified case, typically $Z^\top e \neq \mathbf{0}$ with probability one, even when the model assumption that $\mathbb{E}(u|Z) = 0$ is satisfied.
This is due to a combination of causes:

- Sampling variation; and
- We're effectively trying to solve ℓ equations using $k < \ell$ unknowns.

Pitfalls for the Unwary: Two stage least squares

The interpretation of linear IV as "Two Stage Least Squares" provides an intuition about how things work, but this intuition is misleading in important ways.

1. Don't take "two stage" least squares literally in implementation.
2. In the IV estimator your full matrix of "Instruments" should include *all* of the variables that you think satisfy $\mathbb{E}(Z^\top u) = 0$.
3. Don't try to develop a "structural" first stage & use IV. (Hausman's "Forbidden Regression".) To use this extra information use simultaneous equations instead.
4. You may have more data on (X, Z) than on y. Don't use it!