Econ 422 – Lecture Notes
Part VI

(These notes are slightly modified versions of lecture notes provided by Stock and Watson, 2007. They are for instructional purposes only and are not to be distributed outside of the classroom.)

Instrumental Variables Regression

Three important threats to internal validity are:

- omitted variable bias from a variable that is correlated with X but is unobserved, so cannot be included in the regression;
- simultaneous causality bias (X causes Y, Y causes X);
- errors-in-variables bias (X is measured with error)

Instrumental variables regression can eliminate bias when $E(u|X) \neq 0$ – using an *instrumental variable*, Z

IV Regression with One Regressor and One Instrument

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- IV regression breaks X into two parts: a part that might be correlated with u, and a part that is not. By isolating the part that is not correlated with u, it is possible to estimate β_1.
- This is done using an ***instrumental variable***, Z_i, which is uncorrelated with u_i.
- The instrumental variable detects movements in X_i that are uncorrelated with u_i, and uses these to estimate β_1.

Terminology: endogeneity and exogeneity

An ***endogenous*** variable is one that is correlated with *u*

An ***exogenous*** variable is one that is uncorrelated with *u*

Historical note: "Endogenous" literally means "determined within the system," that is, a variable that is jointly determined with Y, that is, a variable subject to simultaneous causality. However, this definition is narrow and IV regression can be used to address OV bias and errors-in-variable bias, not just to simultaneous causality bias.

Two conditions for a valid instrument

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

For an instrumental variable (an "***instrument***") Z to be valid, it must satisfy two conditions:

1. ***Instrument relevance***: $\text{corr}(Z_i, X_i) \neq 0$
2. ***Instrument exogeneity***: $\text{corr}(Z_i, u_i) = 0$

Suppose for now that you have such a Z_i (we'll discuss how to find instrumental variables later).

How can you use Z_i to estimate β_1?

The IV Estimator, one *X* and one *Z*

Explanation #1: Two Stage Least Squares (TSLS)

As it sounds, TSLS has two stages – two regressions:

(1) First isolates the part of *X* that is uncorrelated with *u*:
regress *X* on *Z* using OLS

$$X_i = \pi_0 + \pi_1 Z_i + v_i \qquad (1)$$

- Because Z_i is uncorrelated with u_i, $\pi_0 + \pi_1 Z_i$ is uncorrelated with u_i. We don't know π_0 or π_1 but we have estimated them, so…
- Compute the predicted values of X_i, $\hat{X}_i$, where $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$, $i = 1,\ldots,n$.

Two Stage Least Squares, ctd.

(2) Replace X_i by $\hat{X}_i$ in the regression of interest:

regress Y on $\hat{X}_i$ using OLS:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i \qquad (2)$$

- **Because $\hat{X}_i$ is uncorrelated with u_i (if n is large), the first least squares assumption holds (if n is large)**
- Thus β_1 can be estimated by OLS using regression (2)
- This argument relies on large samples (so π_0 and π_1 are well estimated using regression (1))
- This the resulting estimator is called the *Two Stage Least Squares (TSLS)* estimator, $\hat{\beta}_1^{TSLS}$.

Two Stage Least Squares, ctd.

Suppose you have a valid instrument, Z_i.

Stage 1: Regress X_i on Z_i, obtain the predicted values $\hat{X}_i$

Stage 2: Regress Y_i on $\hat{X}_i$; the coefficient on $\hat{X}_i$ is the TSLS estimator, $\hat{\beta}_1^{TSLS}$.

$\hat{\beta}_1^{TSLS}$ is a consistent estimator of β_1.

The IV Estimator, one X and one Z, ctd.

Explanation #2: a little algebra…

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

Thus,

$$\begin{aligned}
\text{cov}(Y_i,Z_i) &= \text{cov}(\beta_0 + \beta_1 X_i + u_i, Z_i) \\
&= \text{cov}(\beta_0,Z_i) + \text{cov}(\beta_1 X_i,Z_i) + \text{cov}(u_i,Z_i) \\
&= \quad 0 \quad + \text{cov}(\beta_1 X_i,Z_i) + \quad 0 \\
&= \beta_1 \text{cov}(X_i,Z_i)
\end{aligned}$$

where $\text{cov}(u_i,Z_i) = 0$ (instrument exogeneity); thus

$$\beta_1 = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(X_i, Z_i)}$$

The IV Estimator, one X and one Z, ctd.

$$\beta_1 = \frac{\text{cov}(Y_i, Z_i)}{\text{cov}(X_i, Z_i)}$$

The IV estimator replaces these population covariances with sample covariances:

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}},$$

s_{YZ} and s_{XZ} are the sample covariances. This is the TSLS estimator – just a different derivation!

Consistency of the TSLS estimator

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$$

The sample covariances are consistent: $s_{YZ} \xrightarrow{p} \text{cov}(Y,Z)$ and $s_{XZ} \xrightarrow{p} \text{cov}(X,Z)$. Thus,

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}} \xrightarrow{p} \frac{\text{cov}(Y,Z)}{\text{cov}(X,Z)} = \beta_1$$

- The instrument relevance condition, $\text{cov}(X,Z) \neq 0$, ensures that you don't divide by zero.

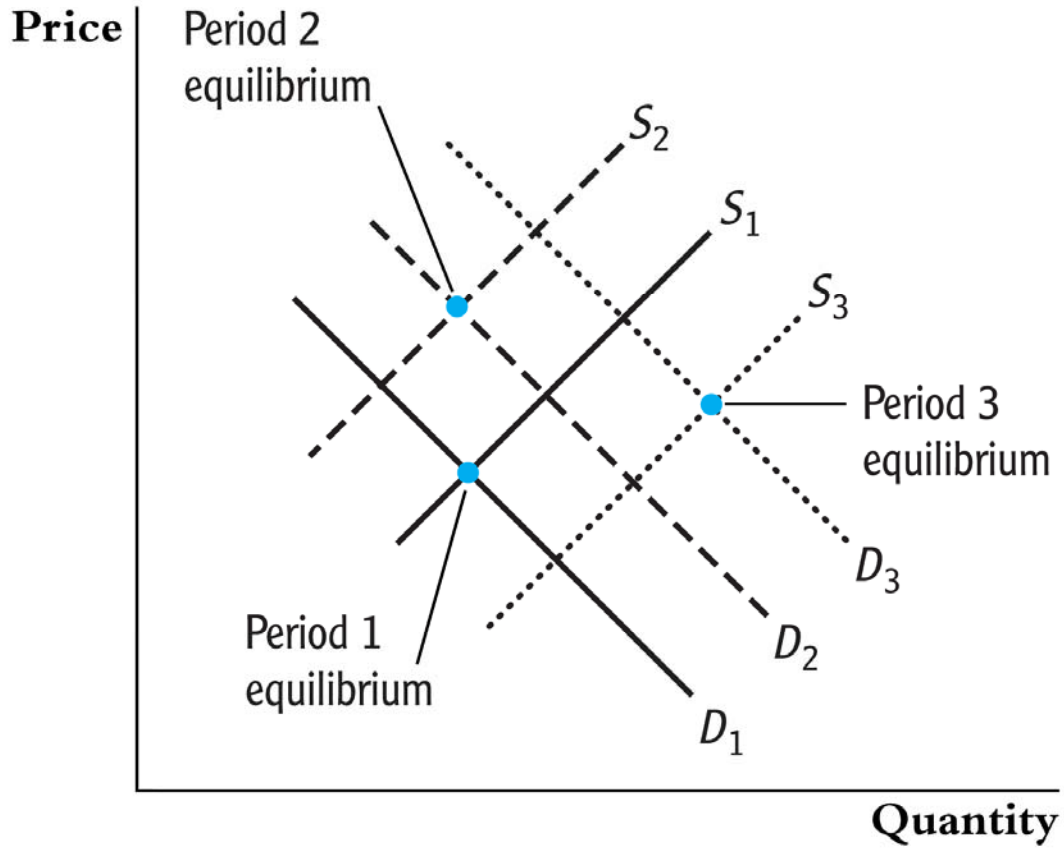
Example #1: Supply and demand for butter

IV regression was originally developed to estimate demand elasticities for agricultural goods, for example butter:

$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i$$

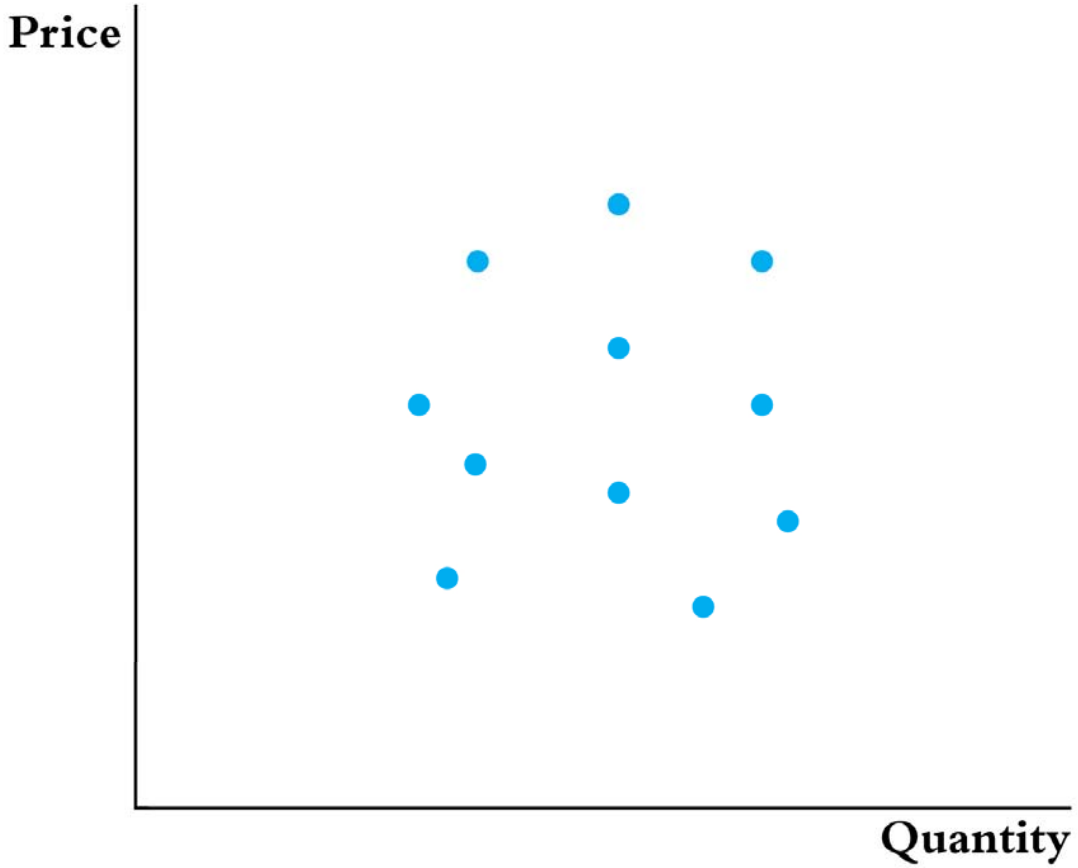
- β_1 = price elasticity of butter = percent change in quantity for a 1% change in price (recall log-log specification discussion)
- Data: observations on price and quantity of butter for different years
- The OLS regression of $\ln(Q_i^{butter})$ on $\ln(P_i^{butter})$ suffers from simultaneous causality bias (*why*?)

Simultaneous causality bias in the OLS regression of $\ln(Q_i^{butter})$ on $\ln(P_i^{butter})$ arises because price and quantity are determined by the interaction of demand *and* supply



(a) Demand and supply in three time periods

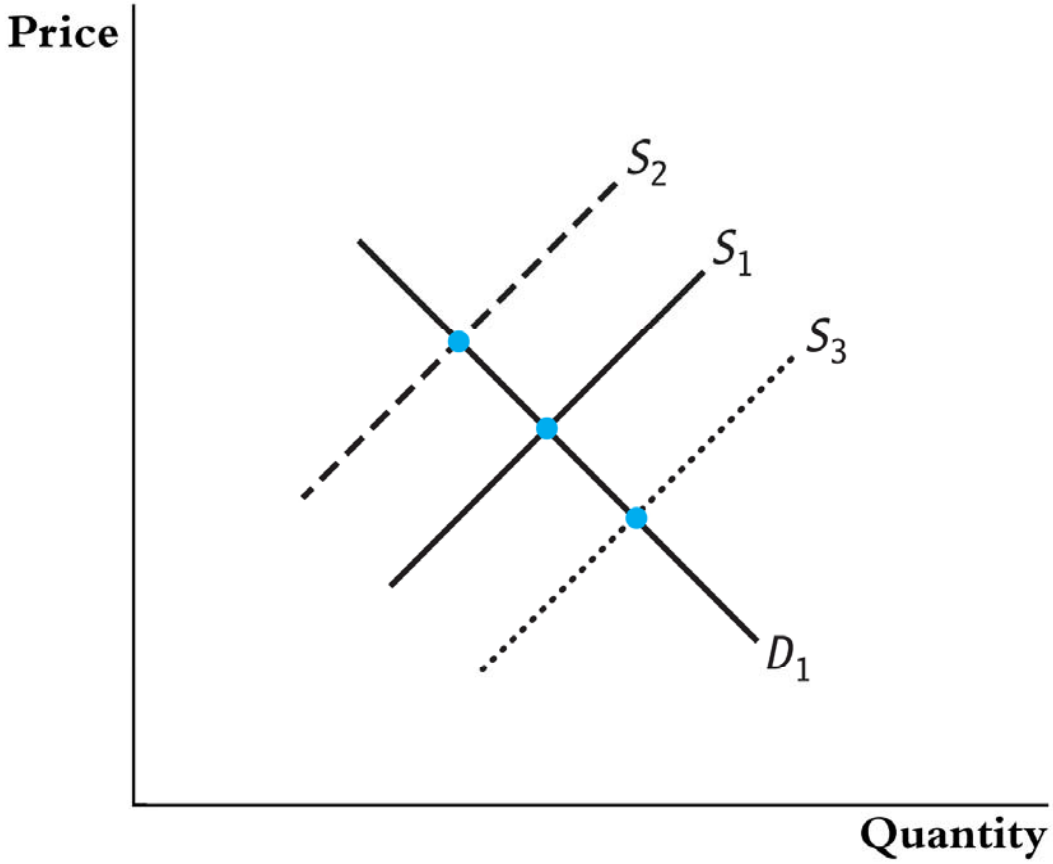
This interaction of demand and supply produces…



(b) Equilibrium price and quantity for 11 time periods

Would a regression using these data produce the demand curve?

But…what would you get if only supply shifted?



(c) Equilibrium price and quantity when only the supply curve shifts

- TSLS estimates the demand curve by isolating shifts in price and quantity that arise from shifts in supply.
- *Z* is a variable that shifts supply but not demand.

TSLS in the supply-demand example:

$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i$$

Let Z = rainfall in dairy-producing regions.

Is Z a valid instrument?

(1) Exogenous? corr($rain_i$,u_i) = 0?

Plausibly: whether it rains in dairy-producing regions shouldn't affect demand

(2) Relevant? corr($rain_i$,$\ln(P_i^{butter})$) ≠ 0?

Plausibly: insufficient rainfall means less grazing means less butter

TSLS in the supply-demand example, ctd.

$$\ln(Q_i^{butter}) = \beta_0 + \beta_1 \ln(P_i^{butter}) + u_i$$

$Z_i = rain_i$ = rainfall in dairy-producing regions.

Stage 1: regress $\ln(P_i^{butter})$ on *rain*, get $\widehat{\ln(P_i^{butter})}$

$\widehat{\ln(P_i^{butter})}$ isolates changes in log price that arise from supply (part of supply, at least)

Stage 2: regress $\ln(Q_i^{butter})$ on $\widehat{\ln(P_i^{butter})}$

The regression counterpart of using shifts in the supply curve to trace out the demand curve.

Example #2: Test scores and class size

- The California regressions still could have OV bias (e.g. parental involvement).
- This bias could be eliminated by using IV regression (TSLS).
- IV regression requires a valid instrument, that is, an instrument that is:
  (1) relevant: corr(Z_i,STR_i) $\neq 0$
  (2) exogenous: corr(Z_i,u_i) = 0

Example #2: Test scores and class size, ctd.

Here is a (hypothetical) instrument:

- some districts, randomly hit by an earthquake, "double up" classrooms:

$$Z_i = Quake_i = 1 \text{ if hit by quake, } = 0 \text{ otherwise}$$

- *Do the two conditions for a valid instrument hold*?
- The earthquake makes it *as if* the districts were in a random assignment experiment. Thus the variation in *STR* arising from the earthquake is exogenous.
- The first stage of TSLS regresses *STR* against *Quake*, thereby isolating the part of *STR* that is exogenous (the part that is "as if" randomly assigned)

We'll go through other examples later…

Inference using TSLS

- In large samples, the sampling distribution of the TSLS estimator is normal
- Inference (hypothesis tests, confidence intervals) proceeds in the usual way, e.g. $\pm 1.96SE$
- The idea behind the large-sample normal distribution of the TSLS estimator is that – like all the other estimators we have considered – it involves an average of mean zero i.i.d. random variables, to which we can apply the CLT.
- Here is a sketch of the math ...

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})(Z_i - \bar{Z})}{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}$$

$$= \frac{\sum_{i=1}^{n} Y_i(Z_i - \bar{Z})}{\sum_{i=1}^{n} X_i(Z_i - \bar{Z})}$$

Substitute in $Y_i = \beta_0 + \beta_1 X_i + u_i$ and simplify:

$$\hat{\beta}_1^{TSLS} = \frac{\beta_1 \sum_{i=1}^{n} X_i(Z_i - \bar{Z}) + \sum_{i=1}^{n} u_i(Z_i - \bar{Z})}{\sum_{i=1}^{n} X_i(Z_i - \bar{Z})}$$

so…

$$\hat{\beta}_1^{TSLS} = \beta_1 + \frac{\sum_{i=1}^{n} u_i (Z_i - \bar{Z})}{\sum_{i=1}^{n} X_i (Z_i - \bar{Z})}.$$

so $$\hat{\beta}_1^{TSLS} - \beta_1 = \frac{\sum_{i=1}^{n} u_i (Z_i - \bar{Z})}{\sum_{i=1}^{n} X_i (Z_i - \bar{Z})}$$

Multiply through by $\sqrt{n}$:

$$\sqrt{n}(\hat{\beta}_1^{TSLS} - \beta_1) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Z_i - \bar{Z}) u_i}{\frac{1}{n} \sum_{i=1}^{n} X_i (Z_i - \bar{Z})}$$

$$\sqrt{n}(\hat{\beta}_1^{TSLS} - \beta_1) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Z_i - \bar{Z})u_i}{\frac{1}{n}\sum_{i=1}^{n}X_i(Z_i - \bar{Z})}$$

- $\frac{1}{n}\sum_{i=1}^{n}X_i(Z_i - \bar{Z}) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z}) \xrightarrow{p} \text{cov}(X,Z) \neq 0$
- $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Z_i - \bar{Z})u_i$ is dist'd $N(0,\text{var}[(Z-\mu_Z)u])$ (CLT)

so: $\hat{\beta}_1^{TSLS}$ is approx. distributed $N(\beta_1, \sigma^2_{\hat{\beta}_1^{TSLS}})$,

where $\sigma^2_{\hat{\beta}_1^{TSLS}} = \frac{1}{n}\frac{\text{var}[(Z_i - \mu_Z)u_i]}{[\text{cov}(Z_i, X_i)]^2}$.

where $\text{cov}(X,Z) \neq 0$ because the instrument is relevant

Inference using TSLS, ctd.

$$\hat{\beta}_1^{TSLS} \text{ is approx. distributed } N(\beta_1, \sigma^2_{\hat{\beta}_1^{TSLS}}),$$

- Statistical inference proceeds in the usual way.
- The justification is (as usual) based on large samples
- This all assumes that the instruments are valid – we'll discuss what happens if they aren't valid shortly.
- ***Important note on standard errors***:
  - The OLS standard errors from the second stage regression aren't right – they don't take into account the estimation in the first stage ($\hat{X}_i$ is estimated).
  - Instead, use a single specialized command that computes the TSLS estimator and the correct *SE*s.
  - as usual, use heteroskedasticity-robust *SE*s

Summary of IV Regression with a Single X and Z

- A valid instrument Z must satisfy two conditions:
  (1) *relevance*: $\text{corr}(Z_i, X_i) \neq 0$
  (2) *exogeneity*: $\text{corr}(Z_i, u_i) = 0$
- TSLS proceeds by first regressing X on Z to get $\hat{X}$, then regressing Y on $\hat{X}$.
- The key idea is that the first stage isolates part of the variation in X that is uncorrelated with u
- If the instrument is valid, then the large-sample sampling distribution of the TSLS estimator is normal, so inference proceeds as usual.