Econometrics I

Lecture 8: Instrumental Variables and GMM

Paul T. Scott
NYU Stern

Fall 2018

Preliminaries

- Econometrics II next semester with Chris Conlon
- Revised syllabus

Basic Idea

- Basic Idea of Instrumental Variable (IV):
  - What if we have a variable that is correlated with X *but not with* Y
  - Then any changes in Y caused by that variable will reflect causal changes by X
- Equation of interest:

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

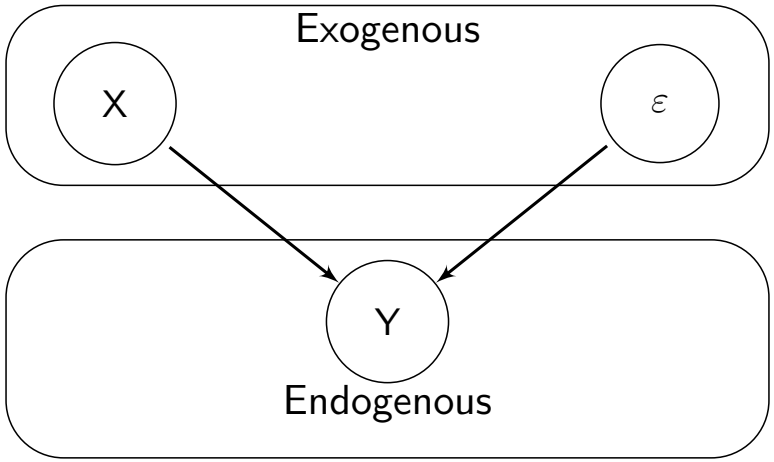
- IVs break X into two pieces that are themselves uncorrelated:

$$X_i = \gamma_0 + \gamma_1 Z_i + \eta_i$$

  - A piece that is not correlated with ε ($Cov(Z, \varepsilon) = 0$)
  - A piece that is correlated with ε ($Cov(\eta, \varepsilon) \neq 0$) – source of the endogeneity problem
  - Finally, $Cov(Z, \eta) = 0$
- Z_i is an **instrumental variable**

Terminology Review

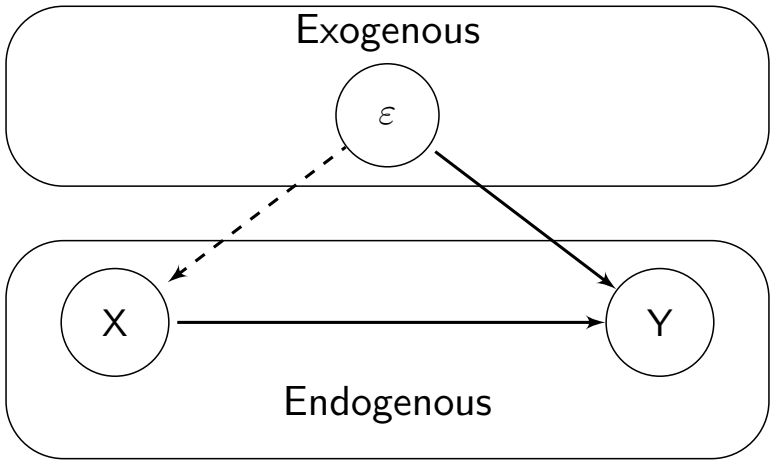
A "Good" Regression:



- **Exogenous Variables:** Variables in the data that do not cause each other
  - ε is always exogenous, so exogenous also just means variables not correlated with ε
- **Endogenous Variables:** Variables that are determined by exogenous variables in the model
  - ε is always in Y so Y is always endogenous

Omitted Variable Bias with Pictures

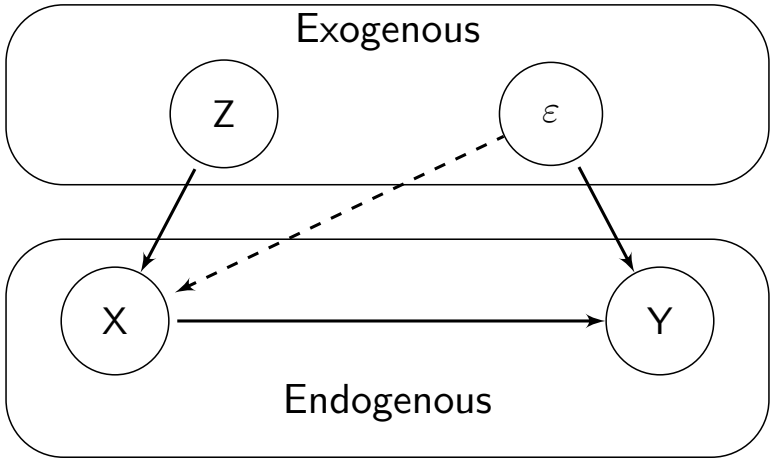
Selection/OVB:



- In this picture X is **endogenous** because ε now causes X as well
- What if there is another exogenous variable that *does not* directly cause Y?

Omitted Variable Bias with Pictures

Instrumental Variable:



- Z causes Y *only indirectly* through X
- We can estimate "causal effect" of Z on Y and this MUST be the causal effect of Z on X and the causal effect of X on Y
  - Mathematically we need to split effect of on Z on Y into effect of Z on X and X on Y
  - Related concept in statistics: Path Analysis

Formal Definition of an Instrumental Variable

Model:

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

- We call a variable Z a a valid **instrumental variable** if the following two conditions hold:

1. **Relevance:** $Cov(X, Z) \neq 0$
   - An arrow from Z to X in the pictures
2. **Exogeneity:** $Cov(\varepsilon, Z) = 0$
   - No arrow from Z to Y or Z to ε in the picture

Key Assumption #1 of Instrumental Variables

- **Relevance:** $Cov(X, Z) \neq 0$
- This assumption just means that X and Z are correlated
- We observe both X and Z, so can easily test this assumption by regressing:

$$X_i = \beta_0 + \beta_1 Z_i + \varepsilon_i$$

- If $\beta_1 \neq 0$ in regression, we say instrument is relevant

Key Assumptions #2 of Instrumental Variables

- **Exogeneity:** $Cov(\varepsilon, Z) = 0$
- This assumption means that Z and ε cannot be correlated
- We do not observe ε, so we **cannot** test this assumption
- In general, we need to "defend" this assumption by telling a story about why Z and ε are unlikely to be correlated
- Much like parallel trends or strict exogeneity, these crucial identifying assumptions cannot be tested on their own. However, we can test them against each other.

Best Defense of Exogeneity IV Assumption: Randomized Experiment

- Back to the class size example:

$$score_i = \beta_0 + \beta_1 CS_i + \varepsilon_i$$

- where:
  - $score_i$: Test score of student i
  - CS_i: Class size of student i
- Suppose that we use a coin flip that sends kids that get a "head" to a small class and kids getting a "tails" to big class
  - This is our randomized experiment!
- Let's call the coin flip our instrument Z (where $Z_i = 1$ if heads, $Z_i = 0$ if tails)

Best Defense of Exogeneity IV Assumption: Randomized Experiment

$$score_i = \beta_0 + \beta_1 CS_i + \varepsilon_i$$

- Is Z (our coin flip) a good instrument?
- **Relevance:** $Cov(CS, Z) \neq 0$? Yes, if $Z_i = 1$ kid gets small class, if $Z_i = 0$ kid gets big class
  - So the regression $CS_i = \beta_0 + \beta_1 Z_i + \varepsilon_i$ will estimate that $\beta_1 < 0$
- **Exogeneity:** $Cov(\varepsilon, Z) \neq 0$? Untestable – so need to tell a story
  - Story: Exogeneity holds because coin flip is random and does not depend on any student or parent characteristics that would affect test scores. Therefore, there is nothing related to Z (besides X) that is also related to test scores, so ε and Z must be uncorrelated.
    This is a good story! It's the "magic" of randomization.

Randomized Experiment as an IV

- So a randomized experiment can be treated as an IV, but there is a big difference in how we evaluate the assumption for an experimental or non-experimental setting
- When running an experiment, the exogeneity assumption is justified if the randomization was implememented in way that was not correlated with anything else that could influence outcomes. Attrition and manipulation can undermine exogeneity in an experimental context if they create a correlation between treatment status and other factors that might influence outcomes. Thus, exogeneity amounts to an assumption about how the randomization was implemented.
- In contrast, the IV exogeneity assumption in a non-experimental context is a broad and vague assumption about how the world works; it's much more difficult to interpret and evaluate. The same is true of strict exogeneity or parallel trends.

IV Estimation: Example with a 0/1 Z

Let's begin with the simplest 0/1 model (this time for Z):

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

- Now X is endogenous but we have a 0/1 instrument, Z
  - Because Z is 0/1 we can focus on group means.
- What assumptions do we need to make?
  - $Cov(X, Z) \neq 0$
  - $Cov(Z, \varepsilon) = 0$
- The intuition from above: solve for the "indirect" effect of Z on Y

The Relationship Between Y on Z

What is the relationship between Y and Z?

$$\begin{aligned} E(Y|Z) =& E\left(\beta_0 + \beta_1 X + \varepsilon | Z\right) \\ =& \beta_0 + \beta_1 E(X|Z) + \underbrace{E(\varepsilon|Z)}_{=0} \\ =& \beta_0 + \beta_1 E(X|Z) \end{aligned}$$

Since Z is 0/1...

The Relationship Between Y on Z

What is the relationship between Y and Z?

$$\begin{aligned} E(Y|Z) =& E(\beta_0 + \beta_1 X + \varepsilon|Z) \\ =& \beta_0 + \beta_1 E(X|Z) + \underbrace{E(\varepsilon|Z)}_{=0} \\ =& \beta_0 + \beta_1 E(X|Z) \end{aligned}$$

Since Z is 0/1...

$$\begin{aligned} E(Y|Z=1) =& \beta_0 + \beta_1 E(X|Z=1) \\ E(Y|Z=0) =& \beta_0 + \beta_1 E(X|Z=0) \end{aligned}$$

Rearranging...

The Relationship Between Y on Z

What is the relationship between Y and Z?

$$\begin{aligned} E(Y|Z) =& E(\beta_0 + \beta_1 X + \varepsilon|Z) \\ =& \beta_0 + \beta_1 E(X|Z) + \underbrace{E(\varepsilon|Z)}_{=0} \\ =& \beta_0 + \beta_1 E(X|Z) \end{aligned}$$

Since Z is 0/1...

$$\begin{aligned} E(Y|Z=1) =& \beta_0 + \beta_1 E(X|Z=1) \\ E(Y|Z=0) =& \beta_0 + \beta_1 E(X|Z=0) \end{aligned}$$

Rearranging...

$$E(Y|Z=1) - E(Y|Z=0) = \beta_1 \times (E(X|Z=1) - E(X|Z=0))$$

The IV Regression for a 0/1 Z

From the previous slide we have:

$$\beta_1 = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(X|Z=1) - E(X|Z=0)}$$

The IV Regression for a 0/1 Z

From the previous slide we have:

$$\beta_1 = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(X|Z=1) - E(X|Z=0)}$$

From the LLN we can construct the following estimator:

$$\hat{\beta}_1 = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{X}_1 - \bar{X}_0}$$

- The subscript refers to $Z = 1$ or $Z = 0$
- We call this the IV estimator estimator
- **Identification** refers to finding population moments with sample analogs that solve β_1 (LLN implies it will work)

Interpreting the Estimator for 0/1 Z

Estimator:

$$\hat{\beta}_1^{IV} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{X}_1 - \bar{X}_0}$$

- For a 0/1 X we look at the change in Y over the change in X:

$$\hat{\beta}_1^{OLS} = \frac{\bar{Y}_{X=1} - \bar{Y}_{X=0}}{\bar{X}_{X=1} - \bar{X}_{X=0}} = \Delta \bar{Y}$$

  - For a 0/1 X, $\Delta \bar{X} = 1$
  - Basically looking at how much Y changes for a change in X
- Intuition for a 0/1 Z
  - The effect of X on Y is still the change in Y over a change in X
  - Use the change in Y *given* Z to shut down the effects of ε
  - Use the change in X *given* Z to get the right scaling

The IV Regression with a Continuous Z

Same model as before:

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

- Now assume that Z is continuous
- How to capture the indirect effect of Z?
  - Intuition from OLS: use the covariance between Z and Y as a measure of the relationship:

$$\begin{aligned} Cov(Y, Z) &= Cov(\beta_0 + \beta_1 X + \varepsilon, Z) \\ &= \beta_1 Cov(X, Z) \end{aligned}$$

  - Rearranging:

$$\beta_1 = \frac{Cov(Y, Z)}{Cov(X, Z)}$$

- **Instrumental Variables Estimator:**

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

IV Intuition

- I will formalize the idea and the math behind IVs with a simple example
- Suppose that we are interested in investigating the effect of studying on grades:

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

- where
  - Y_i: is GPA
  - X_i: is study time (hours per day)
- We expect that our OLS estimator $\hat{\beta}_1$ will be severely biased here (why?)

IV Model

- To get rid of OVB, we shall use an IV
- One possible IV: whether your roommate has a N64 (or playstation/whatever video game console kids use today)
- Important feature: Roommates are randomly assigned in college
  - At least at Berea College (Kentucky) where this example comes from (see Stinebrickner and Stinebrickner (2008))

Instrument Validity

- First thing for any instrument is to think of validity. Two conditions:
  - Instrument is $N64$, which is indicator variable that your roommate has N64

1. Relevance: $Cov(Z, X) \neq 0$; here $Cov(N64, study) \neq 0$
   - Seems likely to hold as everyone prefers playing Mario Kart to studying
   - Testable
2. Exogeneity: $Cov(Z, \varepsilon) = 0$; here $Cov(N64, \varepsilon) = 0$
   - Untestable

Thinking About Exogeneity

- Exogeneity: $Cov(Z, \varepsilon) = 0$
- "Storytime" should talk about how two things hold
  1. People do not select into Z in some manner that is likely to be correlated with Y
     - Concern: People that pick roommates that are "fun" and have N64s are likely people who do not care too much about grades
  2. Z only affects Y through X
     - Concern: Having a roommate with a video game affects your GPA through other means than affecting your study hours
- Either story "invalidates" the instrument (but in different ways)

Thinking About Exogeneity

1. Selection story: there is an omitted variable related to both Z and Y
   - Example: omitted variable is effort because students that really put in a lot of effort make sure they do not get a roommate with N64
2. Other channel story:
   - Possible Story 1: Mario Kart helps me grasp physics, so I ace my physics exam
     - so Z directly affects Y *independent* of X
   - Possible Story 2: Other people hang out in our room due to our N64, making it really loud and so I cannot study effectively
     - so Z affects Y through *another* X

IV Directly into Model

- Suppose that our IV assumptions hold
- Then we can just directly replace X with our IV in our model:

$$Y_i = \pi_0 + \pi_1 Z_i + \varepsilon_i$$

- where
  - Y_i: is GPA
  - Z_i: is whether roomate has N64
- **Interpretation of $\hat{\beta}_1$:** Having a roommate with a N64 **causes** a $\hat{\beta}_1$ decrease in GPA
  - Causal effect because validity of instrument, $cov(Z, \varepsilon) = 0$, implies OLS is consistent for above equation. Of course we could question the instrument's validity.
- But we want the effect of *studying* on GPA!

Two Stage Least Squares

- A popular IV estimator is Two Stage Least Squares (2SLS)
- This effectively runs regression $Y_i = \beta_0 + \beta_1 Z_i + \varepsilon_i$, but *scales* β_1 by how much Z affects X
  - If Z affects X at one-to-one rate, β_1 in above regression will capture effect of X
  - If Z affects X only a little, we need to scale β_1 up a lot to say what effect of X on Y is
- We therefore estimate in two steps:
  1. Regress X on Z to capture effect of Z on X
  2. Regress fitted values of step 1 to capture effect of X on Y
- Under IV assumptions, this gives us **causal** effect of X on Y

2SLS Implementation

- Step 1: regress

$$X_i = \pi_0 + \pi_1 Z_i + \eta_i$$

- Step 2: take your predicted values from Step 1, $\hat{X}_i$, and regress

$$Y_i = \gamma_0 + \gamma_1 \hat{X}_i + \varepsilon_i$$

- Under IV assumptions, step 1 finds the "good part" of X (that is not related to ε), and then step 2 takes that "good part" of X as a regression to find the causal relationship between X and Y

2SLS Estimator

- Why is $\hat{\beta}_1^{2SLS} = \frac{Cov(Y,Z)}{Cov(X,Z)}$?
- True model:

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

- Take covariance of both sides with respect to Z:

$$Cov(Y_i, Z_i) = Cov(\beta_0 + \beta_1 X_i + \varepsilon_i, Z_i)$$
$$Cov(Y_i, Z_i) = Cov(\beta_0, Z_i) + Cov(\beta_1 X_i, Z_i) + Cov(\varepsilon_i, Z_i)$$
$$Cov(Y_i, Z_i) = 0 + \beta_1 Cov(X_i, Z_i) + 0 \quad \textit{since } Cov(\varepsilon_i, Z_i) \textit{ by assumption}$$
$$\implies \beta_1 = \frac{Cov(Y, Z)}{Cov(X, Z)}$$

Single Variable IV in R: Manually

- Data:
  - Y_i: is GPA
  - X_i: is study time (hours per day)
  - Z_i: is indicator for roomate having N64
- Running 2SLS by doing both steps manually:
- Step 1: Regress $X_i = \pi_0 + \pi_1 Z_i + v_i$ and get predicted study time, $\hat{X}_i$
  - Data\$xHat = predict(lm(X∼Z, data=Data))
- Step 2: Regress $y_i = \gamma_0 + \gamma_1 \hat{X}_i + \varepsilon_i$
  - m3 = lm(Y∼xHat, data=Data)
  - coeftest(m3, vcov = vcovHC(m3, type = "HC1"))
    - Note: Standard errors will be wrong!
  - `predict` gives the fitted values of X from the first stage
  - Then `lm` can be used to get the second stage
  - Problem: the standard errors (even robust) will be wrong

Doing Single Variable IV in R: All at once

- Need a new package called AER for this
  - install.packages("AER")
  - library(AER)
- R can run 2SLS all at once:
- m4 = ivreg(Y $\sim$ X | Z, data=Data)
- Use new standard error command to get "right" standard errors:
  - coeftest(m4, vcov=sandwich)

IV Setup with Matrix Notation

- Consider a multiple regression equation

$$\mathbf{y} = \beta\mathbf{X} + \varepsilon$$

  The IV assumptions are now that
  - $Cov(\mathbf{z}_i, \varepsilon_i) = \mathbf{0}$
  - $Cov(\mathbf{z}_i, \mathbf{x}_i)$ has *full rank*, i.e., rank K, where K is the number of regressors in $\mathbf{x}$
- We can have multiple regressors as well as instruments. Some regressors can also be instruments; if $\mathbf{z}_i = \mathbf{x}_i$ then we can just use OLS.
- Regressors that are instruments are **exogenous regressors**. Regressors that are not instruments are **endogenous regressors**.
- For the rank assumption, we need at least one instrument for each endogenous regressor.

2SLS with Matrix Notation

- The 2SLS formula is now

$$\mathbf{b}_{2SLS} = \left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)^{-1}\hat{\mathbf{X}}'\mathbf{y}$$
$$\text{with } \hat{\mathbf{X}} = \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}$$

- With homoscedasticity, the asymptotic variance of the estimator is

$$Var\left(\mathbf{b}_{2SLS}\right) = \frac{s^2}{n}\left[\mathbf{Z}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\right]^{-1}$$

- Note: this (asymptotically correct) estimate of the variance is not equal to the (incorrect) formula we would get from naïvely applying OLS formulas to the second-stage regression.

Doing General IV Case in R

- IV w/ controls
- m5 = ivreg(workedm $\sim$ morekids + boy1st + boy2nd + black + hispan + othrace + agem1 + agefstm | samesex + black + hispan + othrace + agem1 + agefstm+ boy1st + boy2nd, data=babyData)
- Standard errors:
- coeftest(m5, vcov=sandwich)
- Still use the pipe to put in the first stage
- Need to include all exogenous variables on BOTH sides of the pipe!
- Inference does not change at all

Comparing 2SLS Variance to OLS Variance

under Homoscedasticity

For the single-variable case, we can show that

$$Var(\hat{\beta}_{2SLS}) = \frac{1}{N} \frac{\sigma^2}{\sigma_X^2 R_{XZ}^2}$$

- R_{XZ}^2 is the R^2 from the first stage regression of X on Z
- OLS variance under homoscedasticity: $Var(\hat{\beta}^{OLS}) = \frac{1}{N} \frac{\sigma^2}{\sigma_X^2}$
  - Since $R_{XZ}^2 \leq 1$, IV variance is larger
  - The higher first stage R_{XZ}^2, the lower the variance. If X and Z are collinear, variance is equivalent to OLS.
  - If $R_{XZ}^2 \approx 0$ then variance blows up!

Unbiasedness "Proof"

$$\mathbb{E}[\hat{\beta}_1^{2SLS}] = \mathbb{E}\left[\frac{\sum_{i=1}^{N}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})(z_i - \bar{z})}\right]$$

$$= \mathbb{E}\left[\beta_1 + \frac{\sum_{i=1}^{N}(z_i - \bar{z})\varepsilon_i}{\sum_{i=1}^{N}(x_i - \bar{x})(z_i - \bar{z})}\right]$$

$$= \beta_1 + \frac{\sum_{i=1}^{N}(z_i - \bar{z})\mathbb{E}[\varepsilon_i]}{\sum_{i=1}^{N}(x_i - \bar{x})(z_i - \bar{z})} = \beta_1 + 0$$

Checking The IV Assumptions I

- The two major assumptions:
  1. Relevance: $Cov(X, Z) \neq 0$, or that $E\left[\mathbf{z}_i \mathbf{x}_i'\right]$ has full rank
  2. Exogeneity: All instruments must be uncorrelated with ε
- First, let's talk about the relevance assumption.
- It's easy to check that the sample analog to $Cov(X, Z) \neq 0$, or the analogous matrix contidition.
- The more practially relevant concern is that the covariance will be *close* to zero. This is called **weak instruments**.
- How do we test and deal with weak instruments in practice?

Weak Instruments I

- However, in finite samples there's a possibility that the sample covariance

$$\sum_{i=1}^{N}(x_i - \bar{x})(z_i - \bar{z})$$

  is close to zero, especially if the true covariance is relatively small.

- Dividing by zero is *bad*.

- Note we don't have the same problem in OLS as long as we have variation in the regressors.

Weak Instruments II

- There is **big** problem with IV estimation: 2SLS performs **terribly** in small samples
- Intuition: Instrumental variables relies on $Corr(X, Z) \neq 0$
  - If $Cov(X, Z) = 0$ then IV estimator is infinite!
- In practice, very rarely will $Cov(X, Z) = 0$ in data (just randomness)
  - Hence, IV will usually "work" with any choice of X and Z
  - Realistically: $Cov(X, Z)$ is small if Z only explains a small portion of X
- If $Cov(X, Z)$ is small we say that Z is a weak instrument

Weak Instruments and Small Sample Bias

Suppose that Z is *a bad instrument* in that $Cov(Z, \varepsilon)$ are correlated

- Similar to OVB

$$\hat{\beta}_1^{IV} \to \beta_1 + \frac{Cov(Z, \varepsilon)}{Cov(Z, X)} \times \frac{\sigma_\varepsilon}{\sigma_X}$$

- Compare to OLS case (biased because of OVB):

$$\hat{\beta}_1^{OLS} \to \beta_1 + Cov(X, \varepsilon) \times \frac{\sigma_\varepsilon}{\sigma_X}$$

- Which estimator is better here?
  - IV's advantage is $Corr(Z, \varepsilon)$ should be smaller than $Corr(X, \varepsilon)$
  - But if instrument is weak $Corr(Z, X) \approx 0$, then bias blows up even if $Corr(Z, \varepsilon)$ small

- Danger of weak instruments:
  - Very small bias in Z can be greatly magnified if instruments are weak
  - Amplified bias may have different sign than OLS

Weak Instruments and Inference

To do inference we rely on the CLT:

$$\hat{\beta}^{IV} \sim \mathcal{N}\left(\beta, Var\left(\hat{\beta}^{IV}\right)\right)$$

- First issue: the Normal Approximation fails unless N is huge! (the CLT does not work well)
  - Intuition: $\hat{\beta}^{IV} = Cov(Y, Z)/Cov(X, Z)$.
  - If $Cov(X, Z) \approx 0$ then small changes in estimated covariance can have huge effects on estimated coefficients
- Second issue: even if the normal approximation works the variance explodes!
  - Intuition from homoscedastic case: $Var(\hat{\beta}^{IV})) = \sigma^2/(N\sigma_X^2 R_{XZ}^2)$
  - If the R^2 of X on Z is very small, then variance is very large

How To Deal With Weak Instruments?

What are we to do if instruments are weak?

- Easiest: get more instruments or better instruments
- Harder: adjust how to do inference
  - Turns out that if instruments are weak, asymptotic distribution of $\hat{\beta}$ will be the *ratio* of two *correlated* normal random variables
  - Very advanced, we will not pursue this (uses LIML)
- The good news: We can test $Cov(X, Z)$ by regressing $x_i = \pi_0 + \pi_1 Z_i + \eta_i$
  - Since Z and X are data, we can just see if $\hat{\pi}_1$ is far from zero
- **Rule of Thumb:** Only use instrument if t-stat of null hypothesis that $\pi_1 = 0$ is ≥ 10 (much bigger than stat significance of 1.96!)
  - With multiple instrument, "rule" changes to F-stat≥ 10 for joint test that first-stage coefficients on all instruments are zero.

Weak Instruments

The model:

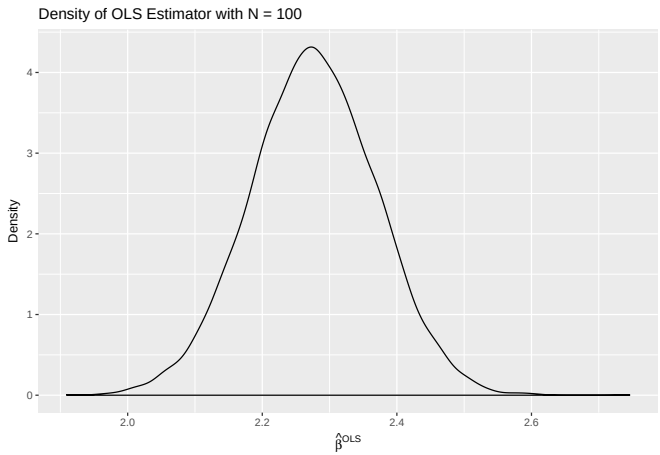
$$\begin{aligned} Y_i =& \beta_0 + \beta_1 X_i + \varepsilon_i \\ X_i =& \pi_0 + \pi_1 Z_i^{(1)} + \pi_2 Z_i^{(2)} + ... + \pi_m Z_i^{(m)} + \eta_i \end{aligned}$$

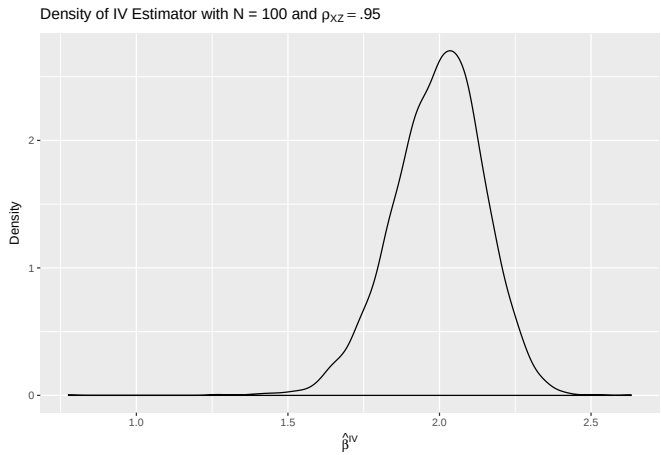
- $\pi \approx 0 \Rightarrow$ because of *sampling error* it is likely that $\hat{\pi} < 0$ even if $\pi > 0$
- If $\hat{X}$ is the wrong *sign* or close to *zero* then it is very likely that $\hat{\beta}^{IV}$ will behave poorly
- **NB:** As $N \to \infty$, $\hat{\pi} \to \pi$ *exactly*—so weak instruments is a *small sample problem*
  - Basically, if $\pi \approx 0$ then CLT will be a bad approximation
  - If $\pi = 0$ then things break even as $N \to \infty$

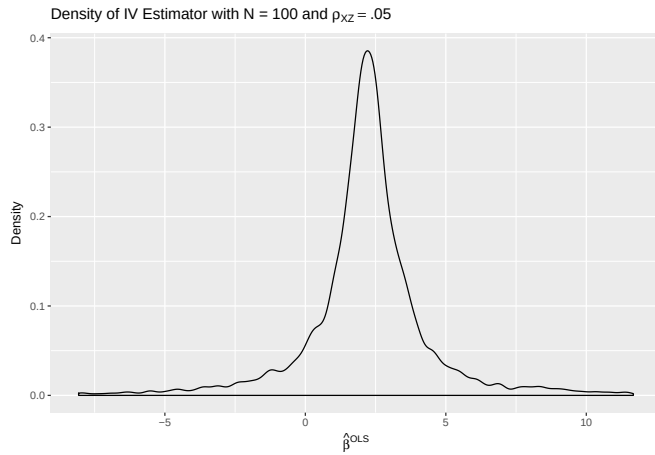
The Sampling Distribution of $\hat{\beta}$

$$Y = 1 + 2X + \varepsilon$$

Figure: Distribution of $\hat{\beta}$ for OLS, strong IV, weak IV with $N = 100$





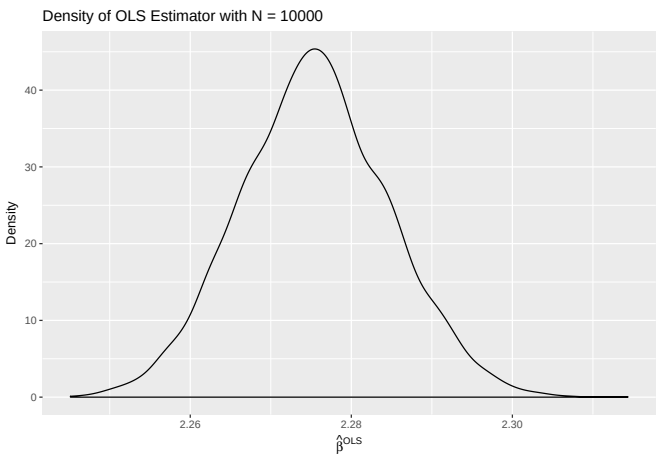


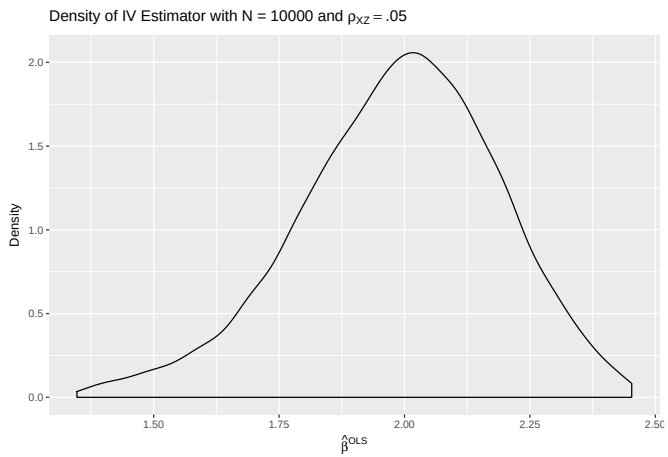
- OLS is biased with $\mu = 2.27$ and $\sigma^2 = .008$
- Strong IV is unbiased (asymptotically) with $\mu = 1.99$ and $\sigma^2 = .024$
- Weak IV is unbiased (asymptotically) with $\mu = 2.05$ but $\sigma^2 = \mathbf{1653.11}$

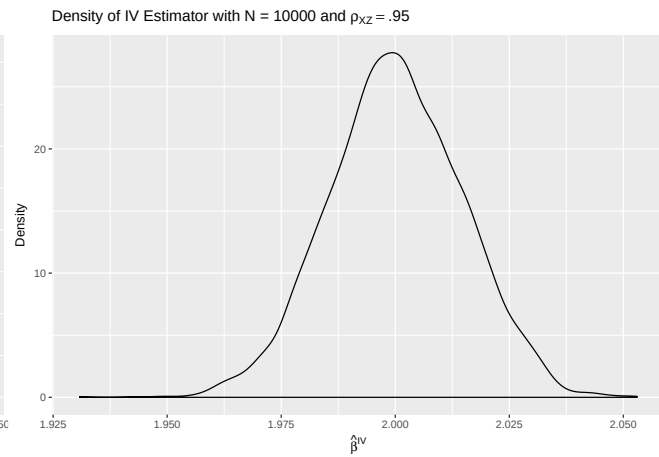
The Sampling Distribution of $\hat{\beta}$

$$Y = 1 + 2X + \varepsilon$$

Figure: Distribution of $\hat{\beta}$ for OLS, weak IV, strong IV with $N = 10000$







- OLS remains biased with $\mu = 2.28$ and $\sigma^2 = .000085$
- Strong IV is consistent with $\mu = 1.99$ and $\sigma^2 = .00021$
- Weak IV is consistent with $\mu = 2.05$ and even with $N = 10k$, $\sigma^2 = .05$
  - Still 100x larger variance than strong instruments!

What is the Danger of the CLT Failing?

- Clearly with weak IV, CLT fails and behaves poorly even if N is large
- Who cares if $E(\hat{\beta}) = \beta$?
- Remember that for testing we want to reject the null hypothesis incorrectly 5% of the time
  - When the CLT fails, our trusty 1.96 critical value is way off base
  - Example rejection probabilities for test that $\hat{\beta} = 2$

Hypothetical Situation	$P(\lvert t \rvert > 1.96)$
Weak IV, $N = 100$	0.3%
Strong IV, $N = 100$	4.2%
Weak IV, $N = 10000$	3.2%
Strong IV, $N = 10000$	5.1%

  - Goal should be 5% but for weak IV with N small, almost never reject; still only 3% even for N large—this is a bad test!

Checking The IV Assumptions II

- The two major assumptions:
  1. Relevance: $Cov(X, Z) \neq 0$, or that $E\left[\mathbf{z}_i \mathbf{x}_i'\right]$ has full rank
  2. Exogeneity: All instruments must be uncorrelated with ε
- Now, let's talk about the exogeneity assumption.
- With one IV, it's impossible to test exogeneity, hence the need to evaluate exogeneity by thinking about whether some unobserved variables might be correlated with the instrument.
- With extra instruments, we can test the IVs against each other. Similarly, one IV allows us to test the validity of OLS.
- In all cases, the most we can do is test one exogeneity assumption against another, we can never test the validity of an exogeneity assumption on its own.

Overidentification

- When the number of regressors equals the number of instruments, the model is **just identified**. When there are more instruments than regressors, the model is **overidentified**.
- For a just identified model, the 2SLS estimator simplifies to

$$\mathbf{b}_{IV} = \left(\mathbf{Z}'\mathbf{X}'\right)^{-1}\mathbf{Z}'\mathbf{y},$$

  which is often called simply the **IV estimator**.
- (Derivation on board)

A Note on Moments I

- Recall the property of the OLS estimator:

$$\mathbf{X}'\mathbf{e}_{OLS} = \mathbf{0}$$

where

$$\mathbf{e}_{OLS} = (\mathbf{y} - \mathbf{X}\mathbf{b}_{OLS})$$

- Similarly, for the IV estimator,

$$\mathbf{Z}'\mathbf{e}_{IV} = \mathbf{0}$$

where

$$\mathbf{e}_{IV} = (\mathbf{y} - \mathbf{X}\mathbf{b}_{IV})$$

- (Derivation on board)

A Note on Moments II

$$\mathbf{X}'\mathbf{e}_{OLS} = \mathbf{0}$$
$$\mathbf{Z}'\mathbf{e}_{IV} = \mathbf{0}$$

- A consequence of this is that we can't test the assumption that the instrument(s) and error terms are uncorrelated by looking at the correlation between the instrument(s) and residuals.
- Similarly, the residuals from OLS don't allow us to test whether the error terms are correlated with the regressors.
- However, the residuals from IV can be used to test the OLS exogeneity assumption. Similarly, if we have more instruments than we need (overidentification), we can test the instruments against each other.

Wu-Hausman Test

- Is OLS biased? Does the use of instrumental variables really make a difference in results?
- Intuition: we can compare the OLS and IV coefficient estimates, see if they're different. If they differ and we believe in the validity of the IV, that's evidence that OLS is biased because of an endogenous regressor.
- The Wu-Hausman test implements this idea formally. It can be derived as a Wald test using the IV and OLS regression results, and it can also be implemented using the following regression:

$$\mathbf{y} = \mathbf{X}\beta + \hat{\mathbf{X}}^*\gamma + \varepsilon^*$$

  where $\mathbf{X}^*$ are the fitted values of $\mathbf{X}$ from the first-stage regression of $\mathbf{X}$ on $\mathbf{Z}$.
- Testing $\gamma = 0$ is one way of implementing the Hausman test. It can also be implemented using the AER package.

Checking Instrument Validity II

Exogeneity Assumption:

$$Corr(Z_i, \varepsilon_i) \quad \text{for all } Z$$

- If exogeneity violated then $\hat{X}$ from the first stage will still be correlated with $\varepsilon \Rightarrow$ omitted variables bias
- Usual case: need to *assume* exogeneity
- With multiple instruments can [with some limitations] test exogeneity
  - Intuition: if $Z^{(1)}$ and $Z^{(2)}$ are both valid instruments then using them separately should produce the same value of $\hat{\beta}$
  - If $\hat{\beta}$ different for different Z's then one Z must be bad

Operationalizing the Intuition

Defining the J Statistic

- Procedure to Test Overidentifying Restrictions
  1. Estimate the model using 2SLS
  2. Compute $\hat{\varepsilon}_i$ with X_i (not $\hat{X}_i$):

$$\hat{\varepsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i^{(1)} - ... \hat{\beta}_m X_i^{(k)}$$

  3. Regress $\hat{\varepsilon}_i$ on all Z's and exogenous variables:

$$\hat{\varepsilon}_i = \alpha_0 + \alpha_1 Z_i^{(1)} + ... + \alpha_m Z_i^{(m)} + Exogenous + \varepsilon_i$$

  4. Compute the F statistic for $\alpha_1 = ... = \alpha_m = 0$ but do *not* do the standard F test (degrees of freedom would be wrong)
  5. Compute $J = mF$, where m is the number of instruments

Operationalizing the Intuition

Performing the J Test

- Under the null hypothesis, $J \sim \chi^2_{m-k}$, where k is the number of regressors.
  - Notice: We use $m-k$ degrees of freedom, not m
  - Why? Because $\hat{\varepsilon}$ is a function of $\hat{\beta}$, which has k parameters
  - If at least one $\alpha \neq 0$ then at least one instrument is endogenous
- The J-test:
  - Calculate the J statistic
  - Compare the J statistic to the critical value for a χ^2_{m-k} distribution
  - Instruments are not exogenous if the test is rejected (so we want the null to not be rejected)
- This is know as Sargan's J test, and it can be implemented using the AER package.
- We can only do this if over-identified
  - If exactly identified, then $Cov(\hat{\varepsilon}, Z) = 0$ automatically

Interpreting J Test Results

- What happens if we reject the null?
  - The α that is non-zero does not necessarily indicate the problematic instrument —$\hat{\varepsilon}$ is itself not unbiased if IV assumptions violated
  - Need to think hard about which instruments may be invalid
- What happens if we do not reject the null?
  - None of your instruments contradict each other by predicting very different $\hat{\varepsilon}$'s
  - Does NOT imply that instruments must be valid—could all be invalid in the same way

Simultaneous Equations Models

New Model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$X_i = \alpha_0 + \alpha_1 Y_i + \nu_i$$

- In many economic environments, variables both determine each other "in equilibrium" or are determined together "jointly"
- A model where X and Y "cause" each other is called a **simultaneous equations model**
- Examples:
  - Angrist & Evans is an example of this: labor supply and fertility decisions are determined jointly, there is no one-way channel
  - Supply and demand is the canonical example: supply equations and demand equations are of interest but in equilibrium only observe $S = D$
  - Also common in macroeconomic environments: interest rates, unemployment rates, etc.

Simultaneity Bias

Direct Effect of X on Y:

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

First let's solve for X and Y *jointly*

$$\begin{aligned} X_i &= \alpha_0 + \alpha_1 Y_i + \nu_i \\ &= \alpha_0 + \alpha_1(\beta_0 + \beta_1 X_i + \varepsilon_i) + \nu_i \\ &= \alpha_0 + \beta_0\alpha_1 + \alpha_1\beta_1 X_i + \alpha_1\varepsilon_i + \nu_i \end{aligned}$$

Which implies...

$$X_i = \frac{\alpha_0 + \alpha_1\beta_0 + \nu_i}{1 - \alpha_1\beta_1} + \underbrace{\frac{\alpha_1}{1 - \alpha_1\beta_1} \times \varepsilon_i}_{OVB}$$

Simultaneity Bias

Direct Effect of X on Y:

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

What happens with OLS? Think about the OVB equation:

$$\begin{aligned}
\hat{\beta}_1^{OLS} &\rightarrow \beta_1 + \frac{Cov(X, \varepsilon)}{Var(X)} \\
&= \beta_1 + \frac{Cov\left(\frac{\alpha_0 + \alpha_1\beta_0 + \nu_i}{1 - \alpha_1\beta_1} + \frac{\alpha_1}{1 - \alpha_1\beta_1} \times \varepsilon_i, \varepsilon_i\right)}{Var\left(\frac{\alpha_0 + \alpha_1\beta_0 + \nu_i}{1 - \alpha_1\beta_1} + \frac{\alpha_1}{1 - \alpha_1\beta_1} \times \varepsilon_i\right)}
\end{aligned}$$

Simultaneity Bias

Direct Effect of X on Y:

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

What happens with OLS? Think about the OVB equation:

$$\begin{aligned}
\hat{\beta}_1^{OLS} &\rightarrow \beta_1 + \frac{Cov(X, \varepsilon)}{Var(X)} \\
&= \beta_1 + \frac{Cov\left(\frac{\alpha_0+\alpha_1\beta_0+\nu_i}{1-\alpha_1\beta_1} + \frac{\alpha_1}{1-\alpha_1\beta_1} \times \varepsilon_i, \varepsilon_i\right)}{Var\left(\frac{\alpha_0+\alpha_1\beta_0+\nu_i}{1-\alpha_1\beta_1} + \frac{\alpha_1}{1-\alpha_1\beta_1} \times \varepsilon_i\right)} \\
&= \beta_1 + \frac{\frac{\sigma_{\varepsilon\nu}+\alpha_1\sigma_\varepsilon^2}{1-\alpha_1\beta_1}}{\frac{\sigma_\nu^2+\alpha_1^2\sigma_\varepsilon^2+2\alpha_1\sigma_{\varepsilon\nu}}{(1-\alpha_1\beta_1)^2}}
\end{aligned}$$

Simultaneity Bias

Direct Effect of X on Y:

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

What happens with OLS? Think about the OVB equation:

$$\begin{aligned}
\hat{\beta}_1^{OLS} &\rightarrow \beta_1 + \frac{Cov(X, \varepsilon)}{Var(X)} \\
&= \beta_1 + \frac{Cov\left(\frac{\alpha_0+\alpha_1\beta_0+\nu_i}{1-\alpha_1\beta_1} + \frac{\alpha_1}{1-\alpha_1\beta_1} \times \varepsilon_i, \varepsilon_i\right)}{Var\left(\frac{\alpha_0+\alpha_1\beta_0+\nu_i}{1-\alpha_1\beta_1} + \frac{\alpha_1}{1-\alpha_1\beta_1} \times \varepsilon_i\right)} \\
&= \beta_1 + \frac{\frac{\sigma_{\varepsilon\nu}+\alpha_1\sigma_\varepsilon^2}{1-\alpha_1\beta_1}}{\frac{\sigma_\nu^2+\alpha_1^2\sigma_\varepsilon^2+2\alpha_1\sigma_{\varepsilon\nu}}{(1-\alpha_1\beta_1)^2}} \\
&= \beta_1 + (1-\alpha_1\beta_1) \times \frac{\alpha_1\sigma_\varepsilon^2 + \sigma_{\varepsilon\nu}}{\alpha_1^2\sigma_\varepsilon^2 + \sigma_\nu^2 + 2\alpha_1\sigma_{\varepsilon\nu}}
\end{aligned}$$

Understanding Simultaneity Bias

OLS with Simultaneity:

$$\hat{\beta}_1^{OLS} \to \beta_1 + (1 - \alpha_1\beta_1) \times \frac{\alpha_1\sigma_\varepsilon^2 + \sigma_{\varepsilon\nu}}{\alpha_1^2\sigma_\varepsilon^2 + \sigma_\nu^2 + 2\alpha_1\sigma_{\varepsilon\nu}}$$

Consider the special case that ε and ν are independent:

$$\hat{\beta}_1^{OLS} = \beta_1 \frac{\sigma_\nu^2}{\sigma_\nu^2 + \alpha_1^2\sigma_\varepsilon^2} + \frac{1}{\alpha_1} \frac{\alpha_1^2\sigma_\varepsilon^2}{\sigma_\nu^2 + \alpha_1^2\sigma_\varepsilon^2}$$

- OLS becomes a weighted average of two pieces:
  1. β_1: the effect of X on Y
  2. $1/\alpha_1$: the feedback loop from Y to X back to Y
- This is **simultaneity bias**

IV Can Solve Simultaneity Bias

Why cover simultaneity bias *now*?

IV Can Solve Simultaneity Bias

Why cover simultaneity bias *now*?

New Model:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \varepsilon_i \\ X_i &= \alpha_0 + \alpha_1 Z_i + \alpha_2 Y_i + \nu_i \end{aligned}$$

- If we have data on Z, we can potentially use Z as an instrument!
- What conditions would Z have to fulfill?
  1. Instrument Relevance: $\alpha_1 > 0$
  2. Instrument Exogeneity: $Cov(Z_i, \varepsilon_i) = 0$

Example: Angrist & Evans

Model:

$$\begin{aligned} Hours_i &= \beta_0 + \beta_1 Children_i + \varepsilon_i \\ Children_i &= \alpha_0 + \alpha_1 Hours_i + \alpha_2 SameSex_i + \nu_i \end{aligned}$$

- Hours and children are determined jointly
- Also $Cov(\varepsilon, \nu)$ is unlikely to be 0
  - Education can matter for both number of children and hours worked (so education is part of ε and ν)
  - Being married may matter for both
  - Quality of non-financial benefits (e.g., maternity leave)
- So formally, the instrument *SameSex* is supposed to help solve simultaneity

Example: Angrist & Evans

TABLE 8—OLS AND 2SLS ESTIMATES OF LABOR-SUPPLY MODELS USING 1990 CENSUS DATA

	All women			Married women			Husbands of married women		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Estimation method	OLS	2SLS	2SLS	OLS	2SLS	2SLS	OLS	2SLS	2SLS
Instrument for *More than 2 children*	—	*Same sex*	*Two boys, Two girls*	—	*Same sex*	*Two boys, Two girls*	—	*Same sex*	*Two boys, Two girls*
Dependent variable: *Worked for pay*	−0.155 (0.002)	−0.092 (0.024)	−0.092 (0.024) [0.743]	−0.147 (0.002)	−0.104 (0.024)	−0.104 (0.024) [0.576]	−0.102 (0.001)	0.017 (0.009)	0.017 (0.009) [0.989]

Example: Porter (1983)

- The Joint Executive Committee's (19th Century railroad cartel) main business was transporting grain from Midwest to ports on East Coast.
- Porter (1983) estimates the following simultaneous system:

$$\text{demand:} \quad \log Q_t = \alpha_0 + \alpha_1 \log P_t + \alpha_2 L_t + U_{1t}$$

$$\text{supply:} \quad \log P_t = \beta_0 + \beta_1 \log Q_t + \beta_2 S_t + \beta_3 I_t + U_{2t}$$

where

  - Q_t is the quantity (tonnage of grain) transported by the JEC
  - P_t is the price per ton of grain transported
  - L_t is a dummy for whether the Great Lakes were open to navigation
  - S_t is a vector of time dummies
  - I_t is an indicator for whether the JEC was colluding

Example: Porter (1983)

$$\text{demand:} \quad \log Q_t = \alpha_0 + \alpha_1 \log P_t + \alpha_2 L_t + U_{1t}$$

$$\text{supply:} \quad \log P_t = \beta_0 + \beta_1 \log Q_t + \beta_2 S_t + \beta_3 I_t + U_{2t}$$

- Note that L_t, a demand shifter, can be used as an instrument for supply.
- Also note that I_t, a supply shifter, can be used as an instrument for demand.
- We can apply 2SLS to each equation, but it is somewhat more asymptotically efficient (i.e., smaller standard errors) to estimate the two jointly (see **multi-equation GMM**).

Example: Porter (1983)

TABLE 3 Estimation Results*

Variable	Two Stage Least Squares (Employing *PO*) Demand	Supply
C	9.169 (.184)	−3.944 (1.760)
LAKES	−.437 (.120)	
GR	−.742 (.121)	
DM1		−.201 (.055)
DM2		−.172 (.080)
DM3		−.322 (.064)
DM4		−.208 (.170)
PO/PN		.382 (.059)
TQG		.251 (.171)

Example: Roberts and Schlenker (2013)

- How much does biofuels mandate raise price of grains? It depends on supply and demand.
- Roberts and Schlenker (2013) estimate the following supply and demand equations:

$$\text{demand:} \quad \log Q_{dt} = \alpha_0 + \alpha_1 \log P_{dt} + \varepsilon_{dt}$$

$$\text{supply:} \quad \log Q_{st} = \beta_0 + \beta_1 \log P_{st} + \beta_2 \omega_t + \varepsilon_{st}$$

where

  - Q_{st} is global grain supply (calories)
  - Q_{dt} is global grain consumption ($Q_{st} \neq Q_{dt}$ because of storage)
  - The prices are measured at different times (planting vs harvest time)
  - ω_t is a yield shock (due to weather), used as an instrument for demand
  - ω_{t-1} is used as an instrument for supply (through storage, past production matters for current prices)

Example: Roberts and Schlenker (2013)

TABLE 1—SUPPLY AND DEMAND ELASTICITY (*FAO data*)

	Instrumental variables			Three-stage least squares		
	(1a)	(1b)	(1c)	(2a)	(2b)	(2c)
Panel A. Supply equation						
Supply elast. β_s	0.102***	0.096***	0.087***	0.116***	0.112***	0.097***
	(0.025)	(0.025)	(0.020)	(0.019)	(0.020)	(0.019)
Shock ω_t	1.184***	1.229***	1.211***	1.249***	1.279***	1.241***
	(0.146)	(0.138)	(0.105)	(0.111)	(0.101)	(0.091)
First stage ω_{t-1}	−3.901***	−3.628***	−3.824***	−3.546***	−3.113***	−3.226***
	(1.145)	(0.945)	(0.910)	(0.800)	(0.704)	(0.731)
First stage ω_t	−2.918*	−2.276*	−2.372*	−2.885***	−2.350***	−2.420***
	(1.647)	(1.294)	(1.279)	(0.967)	(0.815)	(0.819)
Panel B. Demand equation						
Demand elast. β_d	−0.028	−0.055**	−0.054**	−0.034	−0.062***	−0.066***
	(0.021)	(0.024)	(0.022)	(0.023)	(0.022)	(0.021)
First stage ω_t	−5.564***	−4.655***	−4.770***	−5.354***	−4.445***	−4.332***
	(1.489)	(1.300)	(1.249)	(1.384)	(1.210)	(1.186)

Measurement Error in X

- Recall the issue of having measurement error in X:

$$\begin{aligned} Y_i =& \beta_0 + \beta_1 X_i + \varepsilon_i \\ =& \beta_0 + \beta_1 X_i^* + \underbrace{\beta_1 (X_i - X_i^*) + \varepsilon_i}_{\text{"New" Error Term: } V_i} \\ =& \beta_0 + \beta_1 X_i^* + V_i \end{aligned}$$

- X_i^* and V_i likely to be correlated because X^* is part of V
- Extra assumptions:
  - **Classical Errors-in-Variables:**

$$X_i = X_i^* + u_i$$

with u being independent of X and ε (just noise)

Measurement Error in X

- Recall that

$$\begin{aligned}\hat{\beta}^{OLS} \rightarrow & \frac{\beta_1 Cov(X, X) + 0}{Var(X + u)} \\ = & \beta_1 \times \frac{Var(X)}{Var(X + u)} \\ = & \beta_1 \times \underbrace{\frac{\sigma_X^2}{\sigma_X^2 + \sigma_u^2}}_{\text{Attenuation}}\end{aligned}$$

- Estimated coefficient converges to the truth times an attenuation term
  - Attenuation term is less than 1 $\Rightarrow$ pushes coefficient towards zero
    - **NB:** It does *not* make the term more negative—it dampens the coefficient and preserves the sign

Measurement Error and IVs

$$
\begin{aligned}
Y_i =& \beta_0 + \beta_1 X_i + \varepsilon_i \\
=& \beta_0 + \beta_1 X_i^* + \underbrace{\beta_1 (X_i - X_i^*) + \varepsilon_i}_{\text{“New” Error Term: } V_i} \\
=& \beta_0 + \beta_1 X_i^* + V_i
\end{aligned}
$$

- How do we address attenuation bias using instrumental variables?

Generalized Method of Moments I

- The Generalized Method of Moments includes everything we've done so far, and more.
- **Moments** are expectations of functions of the data:

$$E\left[\mathbf{m}\left(y_i, \mathbf{x}_i; \boldsymbol{\theta}_0\right)\right] = 0,$$

where $\boldsymbol{\theta}_0$ is the true value of a vector of parameters of interest.
Typically, we consider moments that are equal to zero – if they were set equal to something else, we could just subtract that from both sides.

Generalized Method of Moments II

- A crucial ingredient for the **method of moments** is that the sample analog of a moment,

$$n^{-1} \sum_{i=1}^{n} \mathbf{m}\left(y_i, \mathbf{x}_i; \boldsymbol{\theta}\right)$$

will converge to the true analog, as long as $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ (law of large numbers)

- Furthermore, if the problem is set up well (i.e., if the model is **identified**), then

$$n^{-1} \sum_{i}^{n} \mathbf{m}\left(y_i, \mathbf{x}_i; \boldsymbol{\theta}\right) \nrightarrow 0$$

for any value of $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.

Generalized Method of Moments II

$$n^{-1}\sum_{i}^{n}\mathbf{m}\left(y_i,\mathbf{x}_i;\boldsymbol{\theta}_0\right)\to 0$$

$$\forall\boldsymbol{\theta}\neq\boldsymbol{\theta}_0:\quad n^{-1}\sum_{i}^{n}\mathbf{m}\left(y_i,\mathbf{x}_i;\boldsymbol{\theta}\right)\nrightarrow 0$$

- Thus, let's try to find a parameter vector $\hat{\boldsymbol{\theta}}$ such that

$$n^{-1}\sum_{i}^{n}m\left(y_i,\mathbf{x}_i;\hat{\boldsymbol{\theta}}\right)=0$$

- Intuitively, in a large sample, this $\hat{\boldsymbol{\theta}}$ should be close to the true parameter.

Generalized Method of Moments II

- A crucial ingredient for the **method of moments** is that the sample analog of a moment,

$$n^{-1} \sum_{i=1}^{n} \mathbf{m}\left(y_i, \mathbf{x}_i; \boldsymbol{\theta}\right)$$

will converge to the true analog, as long as $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ (law of large numbers)

- Furthermore, if the problem is set up well (i.e., if the model is **identified**), then

$$n^{-1} \sum_{i}^{n} \mathbf{m}\left(y_i, \mathbf{x}_i; \boldsymbol{\theta}\right) \nrightarrow 0$$

for any value of $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$.

OLS and IV as Method-of-Moments

- OLS can be derived as a method of moments estimator based on the moments

$$E\left[\mathbf{x}_i \varepsilon_i\right] = 0$$

or

$$E\left[\mathbf{x}_i \left(y_i - \mathbf{x}_i' \boldsymbol{\beta}\right)\right] = 0$$

- Similarly, IV estimation comes from the moments

$$E\left[\mathbf{z}_i \varepsilon_i\right] = 0$$

or

$$E\left[\mathbf{z}_i \left(y_i - \mathbf{x}_i' \boldsymbol{\beta}\right)\right] = 0$$

GMM Assumptions

1. Moments:
$$E\left[\mathbf{m}\left(y_i, \mathbf{x}_i; \boldsymbol{\theta}_0\right)\right] = 0$$
2. Assumptions on data generating process so that
$$n^{-1}\sum_{i=1}^{n}\mathbf{m}\left(y_i, \mathbf{x}_i; \boldsymbol{\theta}\right) \rightarrow E\left[m\left(y_i, \mathbf{x}_i; \boldsymbol{\theta}\right)\right]$$
i.i.d. observations $(y_i, \mathbf{x}_i)$ does the trick, but we can also make do with weaker assumptions (ergodic stationarity).
3. Identification. $E\left[\mathbf{m}\left(y_i, \mathbf{x}_i; \boldsymbol{\theta}\right)\right] \neq 0$ for all $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$
4. Some conditions to ensure asymptotics are "well-behaved". Note: for finite samples, sample moment won't be exactly zero at true parameter, but we want to ensure that if
$$n^{-1}\sum_{i}^{n} m\left(y_i, \mathbf{x}_i; \hat{\boldsymbol{\theta}}\right) = 0,$$
then $\hat{\boldsymbol{\theta}}$ is close to $\boldsymbol{\theta}_0$.

GMM Implementation

- The sample moments can be written as a function of a parameter conjecture θ:

$$\bar{\mathbf{m}}_n(\boldsymbol{\theta}) = n^{-1} \sum_i^n \mathbf{m}(y_i, \mathbf{x}_i; \boldsymbol{\theta})$$

- We define a GMM objective function using a weighting matrix $\mathbf{W}_n$, which should be positive definite:

$$q_n(\boldsymbol{\theta}) = \bar{\mathbf{m}}_n(\boldsymbol{\theta})' \mathbf{W}_n \bar{\mathbf{m}}_n(\boldsymbol{\theta})$$

- The GMM estimator is defined as follows:

$$\boldsymbol{\theta}_{GMM} = \arg\min_{\theta} q_n(\boldsymbol{\theta})$$

Weighting Matrix

$$q_n(\theta) = \bar{\mathbf{m}}_n(\theta)' \mathbf{W}_n \bar{\mathbf{m}}_n(\theta)$$

- The weighting matrix is needed in case the moments cannot be set to zero exactly. It tells us how to penalize violations of each moment.
- In the linear IV model, the weighting matrix only matters when the model is overidentified. Otherwise, we can find an estimate of $\hat{\beta}$ that sets all moments to zero exactly.
- 2SLS can handle overidentified models and uses the weighting matrix $(\mathbf{Z}'\mathbf{Z})^{-1}$.

Two-Step GMM

$$q_n(\theta) = \bar{\mathbf{m}}_n(\theta)' \mathbf{W}_n \bar{\mathbf{m}}_n(\theta)$$

- The weighting matrix will be asymptotically efficient if it converges to $\mathbf{S}^{-1}$, where
$$\mathbf{S} = E\left[\mathbf{m}(y_i, \mathbf{x}_i; \theta)\, \mathbf{m}(y_i, \mathbf{x}_i; \theta)'\right]$$
- In practice, it's common to start with either the 2SLS weighting matrix or an identity matrix. Using the estimates based on that initial weighting matrix, the $\mathbf{S}$ can be estimated, and GMM can be run again using $\hat{\mathbf{S}}^{-1}$ as the weighting matrix.

GMM Asympotics

- Given the assumptions, the GMM estimator is consistent,

$$\hat{\theta}_{GMM} \to \theta_0,$$

and its asymptotic variance is given by

$$n^{-1} \left(\mathbf{\Gamma}' \mathbf{W} \mathbf{\Gamma} \right)^{-1}$$

where $\mathbf{\Gamma}$ is the Jacobian of $E\left[\mathbf{m}\left(y_i, \mathbf{x}_i, \theta\right)\right]$, and $\mathbf{W}$ is what the weighting matrix $\mathbf{W}_n$ converges to.

- To estimate the asymptotic variance, we plug in $\mathbf{W}_n$ for $\mathbf{W}$, and we compute $\mathbf{\Gamma}$ based on the sample moments. The jth row will be

$$\frac{\partial \bar{\mathbf{m}}_j\left(\theta\right)}{\partial \theta}$$

where $\bar{\mathbf{m}}_j\left(\theta\right)$ refers to sample mean of the jth moment.

GMM Advantages

- Not only does GMM generalize things like OLS and IV it can be used for estimation in contexts where the linear regression model does not apply.
- Major example: nonlinear models.
- Econometrics II will explore more applications of the GMM framework.