Simultaneous Equations and Instrumental Variables

Nirav Mehta
(Based on Roy Allen's materials)

Instrumental Variables

We study a new tool that can be used to estimate parameters of economic models in three cases where OLS can fail.

Common theme: in the regression we would otherwise run, the regressor is correlated with the unobservables, so the zero conditional mean (SLR.4/MLR.4) condition fails.

Instrumental Variables

We study a new tool that can be used to estimate parameters of economic models in three cases where OLS can fail.

Common theme: in the regression we would otherwise run, the regressor is correlated with the unobservables, so the zero conditional mean (SLR.4/MLR.4) condition fails.

- Simultaneous equations
  - Price and quantity are determined jointly in competitive markets.

Instrumental Variables

We study a new tool that can be used to estimate parameters of economic models in three cases where OLS can fail.

Common theme: in the regression we would otherwise run, the regressor is correlated with the unobservables, so the zero conditional mean (SLR.4/MLR.4) condition fails.

- Simultaneous equations
    - Price and quantity are determined jointly in competitive markets.
- Measurement error in a covariate.

Instrumental Variables

We study a new tool that can be used to estimate parameters of economic models in three cases where OLS can fail.

Common theme: in the regression we would otherwise run, the regressor is correlated with the unobservables, so the zero conditional mean (SLR.4/MLR.4) condition fails.

- Simultaneous equations
  - Price and quantity are determined jointly in competitive markets.
- Measurement error in a covariate.
- Endogeneity from omitted variables.
  - Example: education may be correlated with ability.

Instrumental Variables

We study a new tool that can be used to estimate parameters of economic models in three cases where OLS can fail.

Common theme: in the regression we would otherwise run, the regressor is correlated with the unobservables, so the zero conditional mean (SLR.4/MLR.4) condition fails.

- Simultaneous equations
  - Price and quantity are determined jointly in competitive markets.
- Measurement error in a covariate.
- Endogeneity from omitted variables.
  - Example: education may be correlated with ability.

Simultaneous Equations

Some economic models describe how *several* variables are determined.

In competitive markets, we believe

- Supply depends on prices.
- Demand depends on prices.
- Markets clear (supply = demand).

Simultaneous Equations

Some economic models describe how *several* variables are determined.

In competitive markets, we believe

- Supply depends on prices.
- Demand depends on prices.
- Markets clear (supply = demand).

Supply Curve:

$$Q_S = \tilde{\gamma}_0 + \tilde{\gamma}_1 P + \varepsilon_S.$$

Demand Curve:

$$Q_D = \tilde{\delta}_0 + \tilde{\delta}_1 P + \varepsilon_D.$$

Markets Clear:

$$Q_S = Q_D.$$

Simultaneous Equations

Suppose observation i is generated according to the model. In other words, these equations hold

$$Q_i = \tilde{\gamma}_0 + \tilde{\gamma}_1 P_i + u_{i,S}.$$

$$Q_i = \tilde{\delta}_0 + \tilde{\delta}_1 P_i + u_{i,D}.$$

Simultaneous Equations

Suppose observation i is generated according to the model. In other words, these equations hold

$$Q_i = \tilde{\gamma}_0 + \tilde{\gamma}_1 P_i + u_{i,S}.$$

$$Q_i = \tilde{\delta}_0 + \tilde{\delta}_1 P_i + u_{i,D}.$$

⚠ Regressing Q on P will not consistently estimate $\tilde{\gamma}_1$ or $\tilde{\delta}_1$.

- We will analyze why and develop a new tool to estimate the supply and demand curves.

Simultaneous Equations

Set supply equal to demand,

$$\tilde{\gamma}_0 + \tilde{\gamma}_1 P_i + u_{i,S} = \tilde{\delta}_0 + \tilde{\delta}_1 P_i + u_{i,D}.$$

Simultaneous Equations

Set supply equal to demand,

$$\tilde{\gamma}_0 + \tilde{\gamma}_1 P_i + u_{i,S} = \tilde{\delta}_0 + \tilde{\delta}_1 P_i + u_{i,D}.$$

Solve to get

$$P_i = \frac{\tilde{\delta}_0 - \tilde{\gamma}_0}{\tilde{\gamma}_1 - \tilde{\delta}_1} + \frac{u_{i,D} - u_{i,S}}{\tilde{\gamma}_1 - \tilde{\delta}_1}$$

⚠ P_i and the unobservables are necessarily correlated.

- With this setup, P_i is actually a function of the unobservables (it is endogenous).

Simultaneous Equations

Set supply equal to demand,

$$\tilde{\gamma}_0 + \tilde{\gamma}_1 P_i + u_{i,S} = \tilde{\delta}_0 + \tilde{\delta}_1 P_i + u_{i,D}.$$

Solve to get

$$P_i = \frac{\tilde{\delta}_0 - \tilde{\gamma}_0}{\tilde{\gamma}_1 - \tilde{\delta}_1} + \frac{u_{i,D} - u_{i,S}}{\tilde{\gamma}_1 - \tilde{\delta}_1}$$

⚠ P_i and the unobservables are necessarily correlated.

- With this setup, P_i is actually a function of the unobservables (it is endogenous).
- We use some additional **exogenous** variables that shift the supply or demand curve.
- By *exogenous*, we mean not systematically related to the unobservables. In the IV setup, we will formalize this with covariance restrictions.

Simultaneous Equations

We assume there is a variable that shifts the *supply* curve (Z_1).

This will be key to estimating the *demand* curve separately.

Simultaneous Equations

We assume there is a variable that shifts the *supply* curve (Z_1).

This will be key to estimating the *demand* curve separately.

Supply Curve:

$$Q_S = \tilde{\gamma}_0 + \tilde{\gamma}_1 P + \tilde{\gamma}_2 Z_1 + \varepsilon_S.$$

Demand Curve:

$$Q_D = \tilde{\delta}_0 + \tilde{\delta}_1 P + \varepsilon_D.$$

Markets Clear:

$$Q_S = Q_D.$$

Simultaneous Equations

Assume data are generated according to the model and satisfy

$$Q_i = \tilde{\gamma}_0 + \tilde{\gamma}_1 P_i + \tilde{\gamma}_2 Z_{i,1} + u_{i,S}$$
$$Q_i = \tilde{\delta}_0 + \tilde{\delta}_1 P_i + u_{i,D}$$

- We will show how a regression can be used to estimate $\tilde{\delta}_1$, i.e. the slope of the demand curve.
- This requires several steps and a lot of algebra (sorry).

Simultaneous Equations

Set supply equal to demand

$$\tilde{\gamma}_0 + \tilde{\gamma}_1 P_i + \tilde{\gamma}_2 Z_{i,1} + u_{i,S} = \tilde{\delta}_0 + \tilde{\delta}_1 P_i + u_{i,D}.$$

Simultaneous Equations

Set supply equal to demand

$$\tilde{\gamma}_0 + \tilde{\gamma}_1 P_i + \tilde{\gamma}_2 Z_{i,1} + u_{i,S} = \tilde{\delta}_0 + \tilde{\delta}_1 P_i + u_{i,D}.$$

We can solve for P_i to get

$$\begin{aligned} P_i &= \frac{1}{\tilde{\gamma}_1 - \tilde{\delta}_1} \left[\tilde{\delta}_0 - \tilde{\gamma}_0 - \tilde{\gamma}_2 Z_{i,1} + u_{i,D} - u_{i,S} \right] \\ &= \left(\frac{\tilde{\delta}_0 - \tilde{\gamma}_0}{\tilde{\gamma}_1 - \tilde{\delta}_1} \right) - \left(\frac{\tilde{\gamma}_2}{\tilde{\gamma}_1 - \tilde{\delta}_1} \right) Z_{i,1} + \left(\frac{u_{i,D} - u_{i,S}}{\tilde{\gamma}_1 - \tilde{\delta}_1} \right). \end{aligned}$$

Simultaneous Equations

Set supply equal to demand

$$\tilde{\gamma}_0 + \tilde{\gamma}_1 P_i + \tilde{\gamma}_2 Z_{i,1} + u_{i,S} = \tilde{\delta}_0 + \tilde{\delta}_1 P_i + u_{i,D}.$$

We can solve for P_i to get

$$\begin{aligned} P_i &= \frac{1}{\tilde{\gamma}_1 - \tilde{\delta}_1}\left[\tilde{\delta}_0 - \tilde{\gamma}_0 - \tilde{\gamma}_2 Z_{i,1} + u_{i,D} - u_{i,S}\right] \\ &= \left(\frac{\tilde{\delta}_0 - \tilde{\gamma}_0}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right) - \left(\frac{\tilde{\gamma}_2}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right) Z_{i,1} + \left(\frac{u_{i,D} - u_{i,S}}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right). \end{aligned}$$

We can rewrite this as...

$$P_i = \beta_0 + \beta_1 Z_{i,1} + e_i.$$

Simultaneous Equations

Key equations:

$$P_i = \left(\frac{\tilde{\delta}_0 - \tilde{\gamma}_0}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right) - \left(\frac{\tilde{\gamma}_2}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right) Z_{i,1} + \left(\frac{u_{i,D} - u_{i,S}}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right)$$
$$P_i = \beta_0 + \beta_1 Z_{i,1} + e_i.$$

Simultaneous Equations

Key equations:

$$P_i = \left(\frac{\tilde{\delta}_0 - \tilde{\gamma}_0}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right) - \left(\frac{\tilde{\gamma}_2}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right) Z_{i,1} + \left(\frac{u_{i,D} - u_{i,S}}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right)$$
$$P_i = \beta_0 + \beta_1 Z_{i,1} + e_i.$$

We can estimate β_1 by regressing P on Z_1 if we assume $\mathbb{E}[e_i \mid Z_{i,1}] = 0$.

Simultaneous Equations

Key equations:

$$P_i = \left(\frac{\tilde{\delta}_0 - \tilde{\gamma}_0}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right) - \left(\frac{\tilde{\gamma}_2}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right) Z_{i,1} + \left(\frac{u_{i,D} - u_{i,S}}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right)$$

$$P_i = \beta_0 + \beta_1 Z_{i,1} + e_i.$$

We can estimate β_1 by regressing P on Z_1 if we assume $\mathbb{E}[e_i \mid Z_{i,1}] = 0$.

- An estimate of β_1 gives us an estimate of $\left(-\frac{\tilde{\gamma}_2}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right)$

Simultaneous Equations

Key equations:

$$P_i = \left(\frac{\tilde{\delta}_0 - \tilde{\gamma}_0}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right) - \left(\frac{\tilde{\gamma}_2}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right) Z_{i,1} + \left(\frac{u_{i,D} - u_{i,S}}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right)$$

$$P_i = \beta_0 + \beta_1 Z_{i,1} + e_i.$$

We can estimate β_1 by regressing P on Z_1 if we assume $\mathbb{E}[e_i \mid Z_{i,1}] = 0$.

- An estimate of β_1 gives us an estimate of $\left(-\frac{\tilde{\gamma}_2}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right)$
- We can learn *something* about the parameters of the supply and demand system!

Simultaneous Equations

Key equations:

$$P_i = \left(\frac{\tilde{\delta}_0 - \tilde{\gamma}_0}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right) - \left(\frac{\tilde{\gamma}_2}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right) Z_{i,1} + \left(\frac{u_{i,D} - u_{i,S}}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right)$$

$$P_i = \beta_0 + \beta_1 Z_{i,1} + e_i.$$

We can estimate β_1 by regressing P on Z_1 if we assume $\mathbb{E}[e_i \mid Z_{i,1}] = 0$.

- An estimate of β_1 gives us an estimate of $\left(-\frac{\tilde{\gamma}_2}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right)$
- We can learn *something* about the parameters of the supply and demand system!

⚠ We *need* $\tilde{\gamma}_2 \neq 0$. The shifter Z_1 needs to actually shift the supply curve.

Simultaneous Equations

We can plug this equation for P_i back into the demand equation to motivate a different regression.

$$
\begin{aligned}
Q_i &= \tilde{\delta}_0 + \tilde{\delta}_1 \frac{1}{\tilde{\gamma}_1 - \tilde{\delta}_1} \left[\tilde{\delta}_0 - \tilde{\gamma}_0 - \tilde{\gamma}_2 Z_{i,1} + u_{i,D} - u_{i,S} \right] + u_{i,D} \\
&= \alpha_0 + \alpha_1 Z_{i,1} + r_i.
\end{aligned}
$$

We can estimate α_1 by regressing Q on Z_1 if we assume $\mathbb{E}[r_i \mid Z_1] = 0$.

- An estimate of α_1 gives us an estimate of $\left(\frac{-\tilde{\delta}_1 \tilde{\gamma}_2}{\tilde{\gamma}_1 - \tilde{\delta}_1} \right)$
- We can learn something else about the parameters of the supply and demand system!

Simultaneous Equations

- An estimate of β_1 gives us an estimate of $\left(-\frac{\tilde{\gamma}_2}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right)$

Simultaneous Equations

- An estimate of β_1 gives us an estimate of $\left(-\frac{\tilde{\gamma}_2}{\tilde{\gamma}_1-\tilde{\delta}_1}\right)$
- An estimate of α_1 gives us an estimate of $\left(\frac{-\tilde{\delta}_1\tilde{\gamma}_2}{\tilde{\gamma}_1-\tilde{\delta}_1}\right)$

Simultaneous Equations

- An estimate of β_1 gives us an estimate of $\left(-\frac{\tilde{\gamma}_2}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right)$
- An estimate of α_1 gives us an estimate of $\left(\frac{-\tilde{\delta}_1\tilde{\gamma}_2}{\tilde{\gamma}_1 - \tilde{\delta}_1}\right)$

We can estimate $\tilde{\delta}_1$, the slope of the demand curve.

- We can estimate it by the *ratio* of two regression coefficients...
- $\hat{\tilde{\delta}}_1 = \frac{\hat{\alpha}_1}{\hat{\beta}_1}$
  - $\hat{\beta}_1$ was from a regression of P on Z_1.
  - $\hat{\alpha}_1$ was from a regression of Q on Z_1.

Simultaneous Equations

Recap

Simultaneous Equations

Recap

Prices depend on the unobservables.

- Regressing quantity on price will not give us a good estimator for *either* the supply or demand curve.

Simultaneous Equations

Recap

Prices depend on the unobservables.

- Regressing quantity on price will not give us a good estimator for *either* the supply or demand curve.

Given a supply shifter (Z_1), we can estimate the slope of the demand curve in a two-step process.

- Regress P on Z_1.
- Regress Q on Z_1.
- Divide the regression coefficients to obtain an estimate of the slope of the demand curve.

Simultaneous Equations

Recap

Prices depend on the unobservables.

- Regressing quantity on price will not give us a good estimator for *either* the supply or demand curve.

Given a supply shifter (Z_1), we can estimate the slope of the demand curve in a two-step process.

- Regress P on Z_1.
- Regress Q on Z_1.
- Divide the regression coefficients to obtain an estimate of the slope of the demand curve.

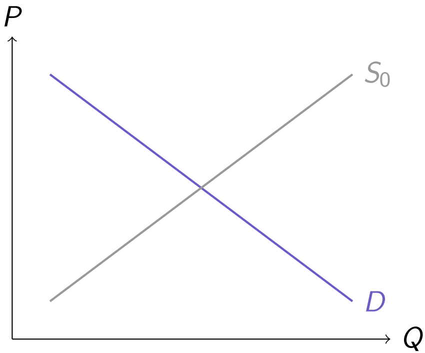
This technique is a form of *instrumental variables* (IV).

Simultaneous Equations

The presence of Z_1 (a supply shifter) was key to estimating the demand curve.

The same technique applies if we have a demand shifter. We can then estimate the supply curve.

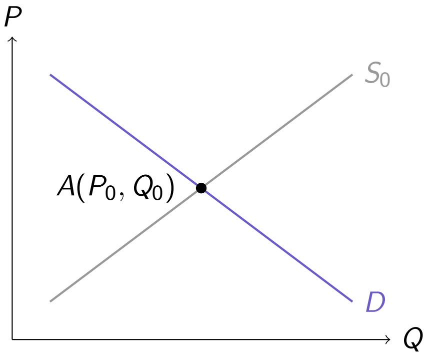
Simultaneous Equations: Graphical Intuition



Supply: $Q_S = \tilde{\gamma}_0 + \tilde{\gamma}_1 P + \tilde{\gamma}_2 Z_1 + \varepsilon_S$

Demand: $Q_D = \tilde{\delta}_0 + \tilde{\delta}_1 P + \varepsilon_D$

Simultaneous Equations: Graphical Intuition

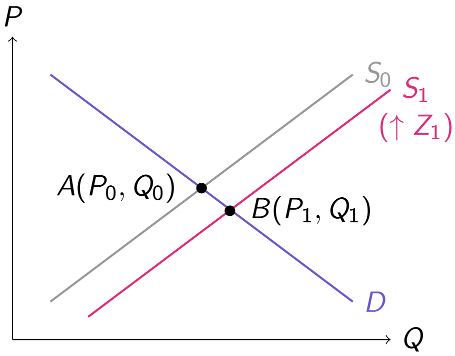


Supply: $Q_S = \tilde{\gamma}_0 + \tilde{\gamma}_1 P + \tilde{\gamma}_2 Z_1 + \varepsilon_S$

Demand: $Q_D = \tilde{\delta}_0 + \tilde{\delta}_1 P + \varepsilon_D$

Equilibrium: $Q_S = Q_D$

Simultaneous Equations: Graphical Intuition



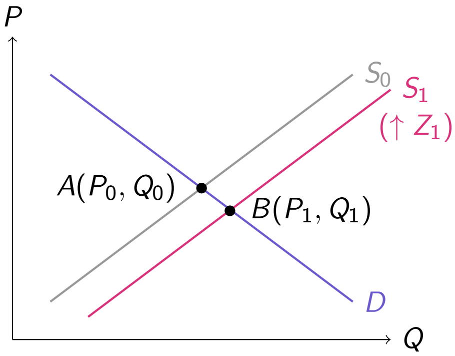
Supply: $Q_S = \tilde{\gamma}_0 + \tilde{\gamma}_1 P + \tilde{\gamma}_2 Z_1 + \varepsilon_S$

Demand: $Q_D = \tilde{\delta}_0 + \tilde{\delta}_1 P + \varepsilon_D$

Equilibrium: $Q_S = Q_D$

First reg.: $P_i = \beta_0 + \beta_1 Z_{i,1} + e_i$

Simultaneous Equations: Graphical Intuition



Supply: $Q_S = \tilde{\gamma}_0 + \tilde{\gamma}_1 P + \tilde{\gamma}_2 Z_1 + \varepsilon_S$

Demand: $Q_D = \tilde{\delta}_0 + \tilde{\delta}_1 P + \varepsilon_D$

Equilibrium: $Q_S = Q_D$

First reg.: $P_i = \beta_0 + \beta_1 Z_{i,1} + e_i$

Second reg.: $Q_i = \alpha_0 + \alpha_1 Z_{i,1} + r_i$

Slope of demand: $\tilde{\delta}_1 = \alpha_1 / \beta_1$

Instrumental Variables

The previous example has illustrated one use of a technique called *instrumental variables* (IV).

Instrumental Variables

The previous example has illustrated one use of a technique called *instrumental variables* (IV).

Basic idea: prices and quantities were jointly determined, but we can use an exogenous shifter (Z_1) to estimate parameters of the original economic model.

Instrumental Variables

The previous example has illustrated one use of a technique called *instrumental variables* (IV).

Basic idea: prices and quantities were jointly determined, but we can use an exogenous shifter (Z_1) to estimate parameters of the original economic model.

Z_1 is an example of an instrument. We can use such variables to address endogeneity or omitted-variable problems.

Instrumental Variables

As usual, begin with the economic model so we're clear about what we're interested in:

$$Y = \tilde{\beta}_0 + \tilde{\beta}_1 X + \varepsilon.$$

We are interested in $\tilde{\beta}_1$. This tells us how Y changes if X changes, *all else equal*.

Instrumental Variables

The data are generated according to

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

We maintain that $\tilde{\beta}_1 = \beta_1$, *but* do not assume that $\mathbb{E}[u \mid X] = 0$.

- This is a violation of SLR.4, so a regression of Y on X may not produce a consistent estimator of $\tilde{\beta}_1$.
- The issue is that X may be systematically related to the unobservables.
- This can happen because of omitted variables, simultaneity, or measurement error.

Instrumental Variables

We assume we observe an instrument Z_i that has these important properties:

- Exogeneity: $\text{Cov}(Z, u) = 0$.
- Relevance: $\text{Cov}(Z, X) \neq 0$.

Instrumental Variables

We assume we observe an instrument Z_i that has these important properties:

- Exogeneity: $\text{Cov}(Z, u) = 0$.
- Relevance: $\text{Cov}(Z, X) \neq 0$.

We need that Z is *not* systematically related to the unobservables, but *is* related to X itself.

We call Z an *instrument* for X.

In the supply/demand example, the supply shifter was an instrument for price.

Instrumental Variables

Given these assumptions, we use data on Y_i, X_i, Z_i to construct an *instrumental variables* (IV) estimator for $\tilde{\beta}_1$.

Instrumental Variables

Given these assumptions, we use data on Y_i, X_i, Z_i to construct an *instrumental variables* (IV) estimator for $\tilde{\beta}_1$.

$$\hat{\beta}_1^{IV} = \frac{\sum_{i=1}^{n}(Z_i - \overline{Z})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(Z_i - \overline{Z})(X_i - \overline{X})}$$

⚠ Does this look familiar at all?

Instrumental Variables

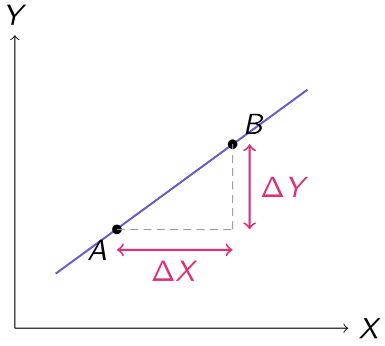
Given these assumptions, we use data on Y_i, X_i, Z_i to construct an *instrumental variables* (IV) estimator for $\tilde{\beta}_1$.

$$\hat{\beta}_1^{IV} = \frac{\sum_{i=1}^{n}(Z_i - \overline{Z})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(Z_i - \overline{Z})(X_i - \overline{X})}$$

⚠ Does this look familiar at all?

- *In this simple setup*, we can obtain $\hat{\beta}_1^{IV}$ by two regressions.
- First regress X on Z, then regress Y on Z. Take the ratio of the regression coefficients.
- We did this in the price/quantity example earlier.

Instrumental Variables: Rise Over Run



When Z changes from z_0 to z_1, X changes by ΔX, and that induced change in X moves Y by ΔY.

$$\Delta Z = z_1 - z_0$$

- "First stage": $\Delta X / \Delta Z$
- "Second stage": $\Delta Y / \Delta Z$

So the IV slope is the rise from the second regression divided by the run from the first stage:

$$\hat{\beta}_1^{IV} \approx \frac{\Delta Y / \Delta Z}{\Delta X / \Delta Z} = \frac{\Delta Y}{\Delta X}.$$

Instrumental Variables

What is $\hat{\beta}_1^{IV}$ estimating under our assumptions?

- Exogeneity: $\text{Cov}(Z, u) = 0$.
- Relevance: $\text{Cov}(Z, X) \neq 0$.

Instrumental Variables

What is $\hat{\beta}_1^{IV}$ estimating under our assumptions?

- Exogeneity: $\text{Cov}(Z, u) = 0$.
- Relevance: $\text{Cov}(Z, X) \neq 0$.

We get

$$\text{Cov}(Z, Y) = \beta_1 \,\text{Cov}(Z, X) + \text{Cov}(Z, u).$$

Rearranging and applying our assumptions,

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

Instrumental Variables

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

Instrumental Variables

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

Since we assumed $\beta_1 = \tilde{\beta}_1$, we obtain

$$\hat{\beta}_1^{IV} \to \frac{\mathrm{Cov}(Z, Y)}{\mathrm{Cov}(Z, X)} = \tilde{\beta}_1.$$

In a large sample, the IV estimator is close to the parameter we want to know, $\tilde{\beta}_1$.

Measurement Error

We will work through another example, in which *measurement error* in a regressor leads to inconsistency of the OLS estimator.

In other words, in a large sample the OLS estimator typically will not be close to the parameter we are interested in.

Measurement Error

We will work through another example, in which *measurement error* in a regressor leads to inconsistency of the OLS estimator.

In other words, in a large sample the OLS estimator typically will not be close to the parameter we are interested in.

We will then show that instrumental variables (IV) can also help address this problem.

Measurement Error

Suppose as before we are interested in an economic model:

$$Y = \tilde{\beta}_0 + \tilde{\beta}_1 X^* + \varepsilon.$$

Measurement Error

Suppose as before we are interested in an economic model:

$$Y = \tilde{\beta}_0 + \tilde{\beta}_1 X^* + \varepsilon.$$

Now we describe how the data are generated:

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

and $\beta_0 = \tilde{\beta}_0$, $\beta_1 = \tilde{\beta}_1$, $\mathbb{E}[u \mid X^*] = 0$.

Measurement Error

Suppose as before we are interested in an economic model:

$$Y = \tilde{\beta}_0 + \tilde{\beta}_1 X^* + \varepsilon.$$

Now we describe how the data are generated:

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

and $\beta_0 = \tilde{\beta}_0$, $\beta_1 = \tilde{\beta}_1$, $\mathbb{E}[u \mid X^*] = 0$.

⚠ We do not observe X^* directly. We observe X^* measured with error,

$$X_i = X_i^* + e_i.$$

Measurement Error

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

and $\beta_0 = \tilde{\beta}_0$, $\beta_1 = \tilde{\beta}_1$, $\mathbb{E}[u \mid X^*] = 0$.

Measurement Error

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

and $\beta_0 = \tilde{\beta}_0$, $\beta_1 = \tilde{\beta}_1$, $\mathbb{E}[u \mid X^*] = 0$.

⚠ We do not observe X^* directly. We observe X^* measured with error,

$$X_i = X_i^* + e_i.$$

What do we get when we regress Y on X? How does it relate to $\tilde{\beta}_1$?

Measurement Error

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

and $\beta_0 = \tilde{\beta}_0$, $\beta_1 = \tilde{\beta}_1$, $\mathbb{E}[u \mid X^*] = 0$.

⚠ We do not observe X^* directly. We observe X^* measured with error,

$$X_i = X_i^* + e_i.$$

What do we get when we regress Y on X? How does it relate to $\tilde{\beta}_1$?

- We will show that, typically, the OLS estimator does *not* consistently estimate $\tilde{\beta}_1$.
- We will also show how instrumental variables (IV) can be used to consistently estimate $\tilde{\beta}_1$.

Measurement Error

To study what the regression of Y on X is estimating, we write,

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i^* + u_i \\ &= \beta_0 + \beta_1 (X_i - e_i) + u_i \\ &= \beta_0 + \beta_1 X_i + (-\beta_1 e_i + u_i) \\ &= \beta_0 + \beta_1 X_i + u_i'. \end{aligned}$$

What is a tool we can use to understand what $\hat{\beta}_1^{OLS}$ is estimating here?

- Earlier we provided different characterizations of *what* $\hat{\beta}_1^{OLS}$ is estimating in a large sample. Is there a useful one for this?

Measurement Error

In a large sample, the OLS estimator is consistently estimating

$$\hat{\beta}_1^{OLS} \rightarrow \tilde{\beta}_1 + \frac{\text{Cov}(X, u')}{\text{Var}(X)}$$

We can use this to understand the bias of the OLS estimator.

Measurement Error

In a large sample, the OLS estimator is consistently estimating

$$\hat{\beta}_1^{OLS} \rightarrow \tilde{\beta}_1 + \frac{\text{Cov}(X, u')}{\text{Var}(X)}$$

We can use this to understand the bias of the OLS estimator.

Remember

$$X_i = X_i^* + e_i$$
$$u_i' = -\tilde{\beta}_1 e_i + u_i$$

Measurement Error

$$
\begin{aligned}
\text{Cov}(X, u') &= \text{Cov}(X^* + e, -\tilde{\beta}_1 e + u) \\
&= \text{Cov}(X^*, -\tilde{\beta}_1 e + u) + \text{Cov}(e, -\tilde{\beta}_1 e + u) \\
&= \left[\text{Cov}(X^*, -\tilde{\beta}_1 e) + \text{Cov}(X^*, u)\right] \\
&\quad + \left[\text{Cov}(e, -\tilde{\beta}_1 e) + \text{Cov}(e, u)\right].
\end{aligned}
$$

Measurement Error

$$
\begin{aligned}
\text{Cov}(X, u') &= \text{Cov}(X^* + e, -\tilde{\beta}_1 e + u) \\
&= \text{Cov}(X^*, -\tilde{\beta}_1 e + u) + \text{Cov}(e, -\tilde{\beta}_1 e + u) \\
&= \left[\text{Cov}(X^*, -\tilde{\beta}_1 e) + \text{Cov}(X^*, u)\right] \\
&\qquad + \left[\text{Cov}(e, -\tilde{\beta}_1 e) + \text{Cov}(e, u)\right].
\end{aligned}
$$

We can plausibly assume three of these terms are 0:

- $\text{Cov}(X^*, e) = 0$: The measurement error (e_i) is not systematically related to X_i^*. This implies $\text{Cov}(X^*, -\tilde{\beta}_1 e) = 0$.
- $\text{Cov}(X^*, u) = 0$: X^* itself is exogenous. (u was the unobservable in the ideal equation $Y_i = \beta_0 + \beta_1 X_i^* + u_i$.)
- $\text{Cov}(e, u) = 0$. The measurement error is uncorrelated with the unobservables u.

Measurement Error

With these assumptions,

$$\text{Cov}(X, u') = \text{Cov}(e, -\tilde{\beta}_1 e) = -\tilde{\beta}_1 \text{Var}(e).$$

- Whenever there is measurement error ($\text{Var}(e) \neq 0$) and X^* actually has a causal effect on Y ($\tilde{\beta}_1 \neq 0$), the OLS estimator is biased and inconsistent.

Measurement Error

In a large sample, the OLS estimator is consistently estimating

$$\hat{\beta}_1^{OLS} \rightarrow \tilde{\beta}_1 + \frac{\text{Cov}(X, u')}{\text{Var}(X)}$$

We can learn more about the bias, since we can also calculate $\text{Var}(X)$.

$$\text{Var}(X) = \text{Var}(X^*) + \text{Var}(e) + 2\,\text{Cov}(X^*, e).$$

We assumed $\text{Cov}(X^*, e) = 0$.

Measurement Error

In a large sample, the OLS estimator is consistently estimating

$$\hat{\beta}_1^{OLS} \to \tilde{\beta}_1 + \frac{\text{Cov}(X, u')}{\text{Var}(X)}$$

We can learn more about the bias, since we can also calculate $\text{Var}(X)$.

$$\text{Var}(X) = \text{Var}(X^*) + \text{Var}(e) + 2\,\text{Cov}(X^*, e).$$

We assumed $\text{Cov}(X^*, e) = 0$.

Measurement Error

In a large sample, the OLS estimator is consistently estimating

$$\hat{\beta}_1^{OLS} \rightarrow \tilde{\beta}_1 + \frac{\text{Cov}(X, u')}{\text{Var}(X)}$$

We can learn more about the bias, since we can also calculate $\text{Var}(X)$.

$$\text{Var}(X) = \text{Var}(X^*) + \text{Var}(e) + 2\,\text{Cov}(X^*, e).$$

We assumed $\text{Cov}(X^*, e) = 0$. So...

$$\hat{\beta}_1^{OLS} \rightarrow \tilde{\beta}_1 + \frac{-\tilde{\beta}_1 \text{Var}(e)}{\text{Var}(X^*) + \text{Var}(e)}$$

Measurement Error

$$\hat{\beta}_1^{OLS} \rightarrow \tilde{\beta}_1 + \frac{-\tilde{\beta}_1 \text{Var}(e)}{\text{Var}(X^*) + \text{Var}(e)}$$

We can rearrange,

$$\begin{aligned}\hat{\beta}_1^{OLS} \rightarrow \tilde{\beta}_1 &\left(1 - \frac{\text{Var}(e)}{\text{Var}(X^*) + \text{Var}(e)}\right) \\ &= \tilde{\beta}_1 \left(\frac{\text{Var}(X^*)}{\text{Var}(X^*) + \text{Var}(e)}\right)\end{aligned}$$

⚠ Can we interpret this? Is there a lesson here?

Measurement Error

This setup is an example of a *classical errors-in-variables* (CEV) setup.

- We assumed $\text{Cov}(X^*, e) = 0$ and $\text{Cov}(e, u) = 0$.
- We concluded that the OLS estimator is systematically closer to 0 than $\tilde{\beta}_1$.
- This is called *attenuation bias* – bias towards 0.

Measurement Error

This setup is an example of a *classical errors-in-variables* (CEV) setup.

- We assumed $\text{Cov}(X^*, e) = 0$ and $\text{Cov}(e, u) = 0$.
- We concluded that the OLS estimator is systematically closer to 0 than $\tilde{\beta}_1$.
- This is called *attenuation bias* – bias towards 0.

These assumptions may not hold for *all* forms of measurement error...
But we can develop the intuition that "typically" the OLS estimator is biased towards 0 in the presence of measurement error.

Measurement Error

We have shown that with measurement error, the OLS estimator may be biased and inconsistent.

Measurement Error

We have shown that with measurement error, the OLS estimator may be biased and inconsistent.

Under the CEV assumptions, it is biased towards 0.

Measurement Error

We have shown that with measurement error, the OLS estimator may be biased and inconsistent.

Under the CEV assumptions, it is biased towards 0.

Instrumental variables (IV) provide a solution.

- We will use an instrument Z that provides a *second* measurement of the original variable X^*.

Measurement Error

Example:

$$Q = \tilde{\beta}_0 + \tilde{\beta}_1 Inc + \varepsilon.$$

- Q is quantity of instant Ramen.
- Inc is income measured in thousands of Canadian dollars.

Measurement Error

Example:

$$Q = \tilde{\beta}_0 + \tilde{\beta}_1 Inc + \varepsilon.$$

- Q is quantity of instant Ramen.
- Inc is income measured in thousands of Canadian dollars.

We have data on

- Self-reported income ($RInc_i$).

Measurement Error

We specify a statistical model describing how the data are generated:

$$Q_i = \beta_0 + \beta_1 Inc_i + u_i$$

$$\beta_0 = \tilde{\beta}_0 \qquad \beta_1 = \tilde{\beta}_1$$

$$RInc_i = Inc_i + e_i$$

Measurement Error

We specify a statistical model describing how the data are generated:

$$Q_i = \beta_0 + \beta_1 Inc_i + u_i$$
$$\beta_0 = \tilde{\beta}_0 \qquad \beta_1 = \tilde{\beta}_1$$
$$RInc_i = Inc_i + e_i$$

So we obtain,

$$\begin{aligned} Q_i &= \beta_0 + \beta_1 RInc_i + (-\beta_1 e_i + u_i) \\ &= \beta_0 + \beta_1 RInc_i + u_i' \end{aligned}$$

As an instrument, suppose we have a second measure of income:

- Employers' reported income ($EInc_i$).

Measurement Error

$$Q_i = \beta_0 + \beta_1 RInc_i + u_i'$$

We assume that $EInc_i$ is related to $RInc_i$ but is not systematically related to unobservable heterogeneity u_i or the measurement error in $RInc_i$, denoted e_i.

Measurement Error

$$Q_i = \beta_0 + \beta_1 RInc_i + u'_i$$

We assume that $EInc_i$ is related to $RInc_i$ but is not systematically related to unobservable heterogeneity u_i or the measurement error in $RInc_i$, denoted e_i.

Formally,

- Exogeneity: $\text{Cov}(EInc, u') = 0$.
  - Here, this follows if $\text{Cov}(EInc, -\beta_1 e) = 0$ and $\text{Cov}(EInc, u) = 0$.
- $\text{Cov}(EInc, RInc) \neq 0$.
  - We assume these are two measurements of the same thing.

Measurement Error

We can use *EInc* as an instrument for *RInc*. The IV estimator is:

$$\hat{\beta}_1^{IV} = \frac{\sum_{i=1}^{n}(Q_i - \overline{Q})(EInc_i - \overline{EInc})}{\sum_{i=1}^{n}(RInc_i - \overline{RInc})(EInc_i - \overline{EInc})}$$

- As before, this can be calculated as the ratio of two regression coefficients.

Endogeneity

We showed IV can be used with simultaneous equations or measurement error.

Endogeneity

We showed IV can be used with simultaneous equations or measurement error.

As a last case, we will consider an example with omitted variables.

$$wage = \tilde{\beta}_0 + \tilde{\beta}_1 educ + \varepsilon.$$

Endogeneity

The data are generated according to

$$wage_i = \beta_0 + \beta_1 educ_i + u_i$$
$$\beta_0 = \tilde{\beta}_0 \qquad \beta_1 = \tilde{\beta}_1$$
$$u_i = \gamma_1 ability_i + u_i'$$

Endogeneity

The data are generated according to

$$wage_i = \beta_0 + \beta_1 educ_i + u_i$$
$$\beta_0 = \tilde{\beta}_0 \qquad \beta_1 = \tilde{\beta}_1$$
$$u_i = \gamma_1 ability_i + u_i'$$

We showed in the omitted variable bias example a few weeks ago that even if $\text{Cov}(educ, u') = 0$,

$$\hat{\beta}_1^{OLS} \to \tilde{\beta}_1 + \frac{\gamma_1 \,\text{Cov}(educ, ability)}{\text{Var}(educ)}.$$

Endogeneity

The data are generated according to

$$wage_i = \beta_0 + \beta_1 educ_i + u_i$$
$$\beta_0 = \tilde{\beta}_0 \qquad \beta_1 = \tilde{\beta}_1$$
$$u_i = \gamma_1 ability_i + u'_i$$

We showed in the omitted variable bias example a few weeks ago that even if $\text{Cov}(educ, u') = 0$,

$$\hat{\beta}_1^{OLS} \to \tilde{\beta}_1 + \frac{\gamma_1 \, \text{Cov}(educ, ability)}{\text{Var}(educ)}.$$

The OLS estimator may not be estimating what we want ($\tilde{\beta}_1$).

Endogeneity

One famous instrument used to address this sort of problem is proximity to school:

- $nearc4$: whether someone grew up near a four-year college.

Endogeneity

One famous instrument used to address this sort of problem is proximity to school:

- *nearc*4: whether someone grew up near a four-year college.

Idea: proximity to a university will make it **more likely** someone will go to university, but may plausibly be **unrelated** to ability.

Endogeneity

One famous instrument used to address this sort of problem is proximity to school:

- $nearc4$: whether someone grew up near a four-year college.

Idea: proximity to a university will make it **more likely** someone will go to university, but may plausibly be **unrelated** to ability.

- Relevance: We think $\text{Cov}(educ, nearc4) > 0$.
- Exogeneity: We are willing to assume $\text{Cov}(nearc4, u) = 0$.
  - This is true if $nearc4$ is uncorrelated with ability and other unobservables u'.

General IV

We have discussed a simple instrumental variables setup with one endogenous variable X and one exogenous instrument Z.

IV also works with a more general setup.

General IV

We have discussed a simple instrumental variables setup with one endogenous variable X and one exogenous instrument Z.

IV also works with a more general setup.

$$\begin{aligned} Y_i = &\beta_0 + \beta_1 X_{1,i} + \cdots + \beta_K X_{K,i} \\ &+ \gamma_1 W_{1,i} + \cdots + \gamma_L W_{L,i} + u_i. \end{aligned}$$

- We can include exogenous control variables $W_1, \ldots, W_L$. (Similar to when we discussed multiple regression.)
- We can have multiple endogenous regressors $X_1, \ldots, X_K$.
- We can have multiple instruments $Z_1, \ldots, Z_J$.

General IV

$$\begin{aligned} Y_i = &\beta_0 + \beta_1 X_{1,i} + \cdots + \beta_K X_{K,i} \\ &+ \gamma_1 W_{1,i} + \cdots + \gamma_L W_{L,i} + u_i. \end{aligned}$$

- We need instruments to deal with the fact that the X variables are endogenous.
- For J instruments $Z_{1,i}, \ldots, Z_{J,i}$, these must satisfy the exogeneity conditions
  - $\text{Cov}(Z_j, u) = 0$ for each instrument j.
- We also need to formalize the fact that the control variables (W variables) are exogenous.
  - $\text{Cov}(W_\ell, u) = 0$ for each ℓ.

General IV

$$
\begin{aligned}
Y_i = &\beta_0 + \beta_1 X_{1,i} + \cdots + \beta_K X_{K,i} \\
&+ \gamma_1 W_{1,i} + \cdots + \gamma_L W_{L,i} + u_i.
\end{aligned}
$$

We need **at least** one instrument for each endogenous regressor.

- This is called the “order condition.”

We need a stronger condition that is harder to state.

- It is called the “rank condition,” and requires the additional assumption that a certain matrix is invertible.
- When there are no control variables, one endogenous variable (X), and one instrument (Z), the rank condition is the previous relevance condition

$$
\mathrm{Cov}(X, Z) \neq 0.
$$

General IV

Key takeaways:

- IV works in a more general setup.
- Key assumptions are exogeneity conditions and relevance conditions.
- In the general setup the "relevance" condition is more complicated, and is called the "rank condition."

General IV

In applications, the most common use of IV is

- One instrument.
- One endogenous variable.
- Exogenous control variables (W variables).

General IV

⚠ Why did we often use the assumption $\mathbb{E}[u \mid X] = 0$ (SLR.4, MLR.4) when we introduced regression, but now have covariance assumptions in IV?

General IV

⚠ Why did we often use the assumption $\mathbb{E}[u \mid X] = 0$ (SLR.4, MLR.4) when we introduced regression, but now have covariance assumptions in IV?

- In regression, we often used the stronger zero conditional mean assumption $\mathbb{E}[u \mid X] = 0$.
  - For example, if $\mathbb{E}[u \mid X] = 0$ and $\mathbb{E}[u] = 0$, then $\text{Cov}(X, u) = 0$. (You do not need to prove this.)
- In instrumental variables, some variables are endogenous and some are exogenous.
- So in IV we work with covariance restrictions such as $\text{Cov}(Z, u) = 0$ for instruments and $\text{Cov}(W_\ell, u) = 0$ for exogenous controls.