Introductory Econometrics I

Endogeneity and Instrumental Variables

Yingjie Feng

School of Economics and Management

Tsinghua University

May 19, 2024

Review: Assumptions for OLS

- Recall the classical linear model (CLM) assumptions for OLS regression:
  - MLR.1: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$
  - MLR.2: random sampling from the population
  - MLR.3: no perfect collinearity in the sample
  - MLR.4: $\mathbb{E}[u|x_1, ..., x_k] = \mathbb{E}[u] = 0$ (**exogenous explanatory variables**)
  - MLR.5: $\mathbb{V}[u|x_1, ..., x_k] = \mathbb{V}[u] = \sigma^2$ (homoskedasticity)
  - MLR.6: $u|x_1, \cdots, x_k \sim \mathsf{Normal}(0, \sigma^2)$
- We have relaxed MLR.6, MLR.5 and MLR.2 (the independence part)
- MLR.3 is mild (just used to guarantee the existence of OLS estimators)
- MLR.1 is a functional form assumption, which will be relaxed as well
- MLR.4 is the key identifying assumption (determines what you really get!)

Endogeneity

- The violation of MLR.4 is usually referred to as **endogeneity** in economics.
- We know if MLR.4 fails, OLS estimators are no longer unbiased/consistent.
- In today's class we first discuss the source of endogeneity, and then provide solutions to it.
- In particular, we will discuss the instrumental variables approach, which has been a standard method in economists' toolkit.

Outline

Outline

Omitted Variables

- Recall omitting *relevant* variables may make OLS estimators biased
  - For example, if $\mathbb{E}[x_2|x_1] \neq 0$,

$$y = \beta_0 + \beta_1 x_1 + \underbrace{\beta_2 x_2 + u}_{\downarrow}, \quad \mathbb{E}[u|x_1, x_2] = 0$$

$$y = \beta_0 + \beta_1 x_1 + \quad e, \qquad \mathbb{E}[e|x_1] \neq 0$$

  - In this case,

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\delta}_1 \xrightarrow{p} \beta_1 + \beta_2 \cdot \delta_1$$

$$x_2 = \delta_0 + \delta_1 x_1 + v.$$

    - ★ $\hat{\beta}_j$: estimate of β_j from regressing y on x_1 and x_2
    - ★ $\tilde{\beta}_j$: estimate of β_j from regressing y on x_1 only
    - ★ $\tilde{\delta}_1$: estimate of the slope on x_1 from regressing x_2 on x_1
  - Bias arises when $\beta_2 \neq 0$ and $\tilde{\delta}_1 \neq 0$

Outline

Measurement Error

- Another potential source of endogeneity is **measurement error**, which can be present in the dependent variable or one or more explanatory variables.

- In practice, we are often unable to collect data on the economic variables we really need. Instead, we only have some **imprecise** measurements.

- Measurement error may or may not lead to endogeneity, depending on how we think about the relation between the measurement error and explanatory variables in the model.

Measurement Error in the Dependent Variable

- Suppose the population model of interest is

$$y^* = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta x_k + u$$

  - We collect a random sample $\{x_{i1}, ..., x_{ik}\}_{i=1}^n$
  - But we do not directly observe y^*.

- Example: suppose y^* is actual family saving for a year. But what we can observe is y, the amount reported by the family.
  - Our random sample is $\{y_i, x_{i1}, x_{i2}, ..., x_{ik}\}_{i=1}^n$
  - It is **infeasible** to regress y_i^* on $x_{i1}, x_{i2}, ..., x_{ik}$
  - But we can regress y_i on $x_{i1}, x_{i2}, ..., x_{ik}$.

- **Question:** Suppose the model with y^* satisfies Assumptions MLR.1-MLR.4. When does mismeasured dependent variable not cause bias in OLS?

Measurement Error in the Dependent Variable

- **Measurement error** $e_0 = y - y^*$
- Since we can write $y = y^* + e_0$, we have

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u + e_0$$

- We have a composite error: $\tilde{u} = u + e_0$.
- Assumption MLR.4 for the **original** model: $\mathbb{E}[u|x_1, \cdots, x_k] = 0$
- OLS estimates are unbiased (and consistent) if $\mathbb{E}[\tilde{u}|x_1, \cdots, x_k] = 0$
- So the key requirement is

$$\mathbb{E}[e_0|x_1, \cdots, x_k] = 0$$

  - The measurement error in y is uncorrelated with all the explanatory variables

Measurement Error in the Dependent Variable

- The traditional assumptions are
    1. $\mathbb{E}[e_0] = 0$ (almost for free)
    2. e_0 is statistically independent of $(x_1, x_2, ..., x_k)$
- **Consequence:** OLS estimators are unbiased and consistent; OLS inference (t, F tests, etc.) is valid.
- Usually, it is also assumed that $\mathbb{C}ov[u, e_0] = 0$ and thus

$$\mathbb{V}[u + e_0] = \sigma_u^2 + \sigma_0^2 > \sigma_u^2$$

$\tilde{u}$

- **Consequence**: measurement error in y increases the error variance and thus the variance of OLS estimators.
- **Key message**: If the measurement error in y is some random reporting error independent of explanatory variables, OLS still has good properties

Measurement Error in the Dependent Variable

- However, OLS estimates are biased if the measurement error in y is systematically related to one or more explanatory variables.
- An example:

$$rd^* = \beta_0 + \beta_1 hightech + u$$

rd^*: actual R&D expenditure; $hightech$: a binary variable for high-tech firm identification.

  - But suppose we only have *self-reported* R&D expenditure rd
  - "High-tech" firms are more likely to report R&D expenditure above the actual value (to justify their high-tech identification?), $hightech$ and $e_0 = rd - rd^*$ are positively correlated
  - The regression will overestimate the effect of the "high-tech firm designation" program.

Measurement Error in the Independent Variable

- Measurement in **explanatory variables** is usually viewed as more of a problem than measurement error in y, but this hinges on a particular view of measurement error.
- Consider the case of simple regression:

$x_1^* = x_1 - e_1$

$$y = \beta_0 + \beta_1 x_1^* + u, \qquad x_1 = x_1^* + e_1$$

  x_1 is a measurement of x_1^*. Assume $\mathbb{E}[e_1] = 0$ (almost for free).
- Write an equation that contains the observed measure:

$$y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$$

  - We still assume $\mathbb{C}ov[u, x_1] = 0$
  - But what about e_1?

Measurement Error in the Independent Variable

$$y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$$

- The **classical errors-in-variables (CEV)** assumption is $\mathbb{C}ov[x_1^*, e_1] = 0$.
  - e_1 is random reporting error uncorrelated with the true explanatory variable.
- The CEV assumption implies

$$\mathbb{C}ov[x_1, e_1] = \mathbb{C}ov[x_1^* + e_1, e_1] = \mathbb{V}[e_1] = \sigma_{e_1}^2$$

$$\mathbb{V}[x_1] = \mathbb{V}[x_1^* + e_1] = \mathbb{V}[x_1^*] + \mathbb{V}[e_1] = \sigma_{x_1^*}^2 + \sigma_{e_1}^2$$

- **Key result**:

$$\mathbb{C}ov[x_1, u - \beta_1 e_1] = \mathbb{C}ov[x_1, u] - \beta_1 \mathbb{C}ov[x_1, e_1] = -\beta_1 \sigma_{e_1}^2$$

- OLS estimator is **inconsistent** under the CEV assumption.

Measurement Error in the Independent Variable

- The OLS estimator is inconsistent under the CEV assumption:

$$
\begin{aligned}
\text{plim}(\hat{\beta}_1) &= \beta_1 + \frac{\mathbb{C}ov[x_1, u - \beta_1 e_1]}{\mathbb{V}[x_1]} \\
&= \beta_1 - \frac{\beta_1 \sigma^2_{e_1}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}} = \beta_1 \left(1 - \frac{\sigma^2_{e_1}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}} \right) \\
&= \beta_1 \left(\frac{\sigma^2_{x_1^*}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}} \right) = \beta_1 \left(\frac{\sigma^2_{x_1^*}}{\sigma^2_{x_1}} \right)
\end{aligned}
$$

- The term multiplying β_1,

$$
\frac{\sigma^2_{x_1^*}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}} = \frac{\sigma^2_{x_1^*}}{\sigma^2_{x_1}} < 1
$$

  unless there is no measurement error ($\sigma^2_{e_1} = 0$).

- Important conclusion: $|\text{plim}(\hat{\beta}_1)| < |\beta_1|$.
  - It is called **attenuation bias**: the estimator is systematically too close to zero compared with β_1.

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(V_i - \bar{V})}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

OLS

$$Y = \beta_0 + \beta_1 X + V.$$

$$\xrightarrow{p.} \frac{\text{cov}[X, V]}{V[X]}.$$

Measurement Error in the Independent Variable

- **Reminder:** Attenuation bias calculation influenced much empirical work, but one should understand it depends critically on the CEV assumption

$$\mathbb{C}ov[x_1^*, e_1] = 0.$$

- If the other extreme holds, i.e., $\mathbb{C}ov[x_1, e_1] = 0$, there is **no** attenuation bias.
- Intermediate cases are possible.
  - Wooldridge example: how many times did you smoke in the last 30 days?

$$smoked = smoked^* + e_1$$

  - At least, it is likely $e_1 = 0$ if $smoked^* = 0$ while $e_1 \neq 0$ if $smoked^* \neq 0$

Outline

Sample Selection

- Another important source of endogeneity is sample selection
- The sample may be **not representative** of the population. Some groups may be sampled more (or less) frequently than dictated by their population representation.
  - For example, too many low-income families and too few high-income families
- Sample selection may or may not lead to endogeneity
  1. Exogenous sampling: sampling is based on the values of explanatory variables
     - OLS estimators are still unbiased and consistent
  2. Endogenous sampling: sampling is systematically related to dependent variable y or error term u
     - OLS estimators may be biased and inconsistent

Endogenous Sampling

- A typical example:

$$lwage^o = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u,$$

  $lwage^o$ is log of of "wage offer," the wage a woman could work for. But we observe $lwage^o$ **only if she is working**.

- This might cause a sample selection problem.
  - We only have a "subsample" of women who are in the labor force.
  - But women self-select into the labor force.
  - Decision to work correlated with u: more "motivated" women are more productive and more likely to be in the labor force

Outline

1. Source of Endogeneity
   - Omitted Variables
   - Measurement Error
   - Sample Selection*
   - Simultaneity*
2. Instrumental Variables
   - IV Estimation of Simple Regression Model
   - IV Estimation of Multiple Regression Model
   - Multiple Instruments: 2SLS
   - IV Solution to Error-in-Variables Problems
   - Testing Whether a Variable is Endogenous
   - Testing Overidentification Restrictions

Simultaneity Bias

- Another source of endogeneity is simultaneity (for example, supply and demand curves)

$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1$$

$$y_2 = \alpha_2 y_1 + \beta_2 z_2 + u_2$$

- Then, we can write

$$y_2 = \alpha_2(\alpha_1 y_2 + \beta_1 z_1 + u_1) + \beta_2 z_2 + u_2$$

$$(1 - \alpha_1\alpha_2) y_2 = \alpha_2\beta_1 z_1 + \beta_2 z_2 + (\alpha_2 u_1 + u_2)$$

- So in general (when $\alpha_2 \neq 0$ and $\alpha_1\alpha_2 \neq 1$), $\mathbb{C}ov[y_2, u_1] \neq 0$
- See Wooldridge, Chapter 16

Outline

Outline

Instrumental Variables

- Now we consider one solution for endogeneity (not the only one): instrumental variables
- Consider a simple regression model:

$\mathbb{E}[xu] \neq 0$

$$y = \beta_0 + \beta_1 x + u$$

- An instrumental variable z needs to satisfy two restrictions:
  1. z is **exogenous** to the equation (also called exclusion restriction):

  $$\mathbb{C}ov(z, u) = 0$$

     ★ In general, we cannot test it since u is not observed.
  2. z is **relevant** for explaining x:

  $$\mathbb{C}ov(z, x) \neq 0$$

     ★ We can test it since both x and z are observable.

Instrumental Variables

- Consider a simple regression model:

$$y = \beta_0 + \beta_1 x + u, \quad \mathbb{C}ov(z,u) = 0, \quad \mathbb{C}ov(z,x) \neq 0$$

- How can we use a variable z satisfying these two requirements?
- Take the covariance of z with both sides of the equation:

$$\mathbb{C}ov(z,y) = \beta_1 \mathbb{C}ov(z,x) + \mathbb{C}ov(z,u).$$

- By $\mathbb{C}ov(z,u) = 0$ (exogeneity),

$$\mathbb{C}ov(z,y) = \beta_1 \mathbb{C}ov(z,x).$$

- Next, by $\mathbb{C}ov(z,x) \neq 0$ (relevance),

$$\beta_1 = \frac{\mathbb{C}ov(z,y)}{\mathbb{C}ov(z,x)}$$

- We have written β_1 as two population moments in observable variables!

Instrumental Variables

$$\beta_1 = \frac{\mathbb{C}ov(z, y)}{\mathbb{C}ov(z, x)}$$

- This motivates the **IV estimator** of β_1 (method of moment):

$\beta_1(x_i - \bar{x}) + u_i - \bar{u}$

$$\hat{\beta}_1^{IV} = \frac{n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$$

- IV estimator is consistent, but not unbiased.
  - Bias can be large if the correlation between z and x is "small".
- The variance of the IV estimator can be large. Under homoskedasticity of u $(\mathbb{V}[u|z] = \mathbb{V}[u])$

$$\mathbb{V}[\hat{\beta}_1^{IV}] \approx \frac{\sigma_u^2}{n\sigma_x^2\rho_{x,z}^2}$$

$\sigma_u^2 = \mathbb{V}[u]$, $\sigma_x^2 = \mathbb{V}[x]$ and $\rho_{x,z} = Corr(x, z)$.

$$\hat{\beta}_1^{IV} - \beta_1 = \frac{\frac{1}{n}\sum (z_i - \bar{z})(u_i - \bar{u})}{\frac{1}{n}\sum (z_i - \bar{z})(x_i - \bar{x})}$$

$$\xrightarrow{p} \frac{\text{cov}[z_i, u_i] \to = 0.}{\text{cov}[z_i, x_i]}$$

$$V[\hat{\beta}_1^{IV}] = V\left[\frac{\frac{1}{n}\sum (z_i - \bar{z})(u_i - \cancel{\bar{u}})}{\frac{1}{n}\sum (z_i - \bar{z})(x_i - \bar{x})}\right] = \frac{\frac{1}{n^2} V\left[\sum (z_i - \bar{z}) u_i\right].}{\left(\frac{1}{n}\sum (z_i - \bar{z})(x_i - \bar{x})\right)^2}$$

given $\{x_i, z_i\}_{i=1}^{n}$

$$= \frac{\frac{1}{n^2} \sum_{i=1}^{n} (z_i - \bar{z})^2 \, V[u_i] \to \sigma_u^2}{\left(\frac{1}{n}\sum_{i=1}^{n} (z_i - \bar{z})(x_i - \bar{x})\right)^2 \xrightarrow{p} \text{cov}[x, z]^2.}$$

$$\xrightarrow{p} \frac{1}{n} V[z].$$

$$\rho_{x,z} = \frac{\text{cov}[x, z]}{\sqrt{V[x] V[z]}}$$

so $$V[\hat{\beta}_1^{IV}] \xrightarrow{p} \frac{\sigma_u^2}{n \frac{\text{cov}[x,z]^2}{V[z]}} = \frac{\sigma_u^2}{n \sigma_x^2 \rho_{x,z}^2}$$

Instrumental Variables

- Compare with OLS variance (when OLS is consistent):

$$\mathbb{V}[\hat{\beta}_1^{IV}] \approx \frac{\sigma_u^2}{n\sigma_x^2 \rho_{x,z}^2} \quad \geq \quad \mathbb{V}[\hat{\beta}_1^{OLS}] \approx \frac{\sigma_u^2}{n\sigma_x^2}$$

- A rough rule of thumb: under homoskedasticity,

$$se(\hat{\beta}_1^{IV}) \approx \frac{1}{r_{xz}} \times se(\hat{\beta}_1^{OLS})$$

  where r_{xz} is the sample correlation between x_i and z_i.
    - Think of this factor as the cost of doing IV when we could use OLS (If OLS is inconsistent, the variance comparison makes little sense)
    - Variance of IV estimator could be large if x and z are "weakly" correlated
- Heteroskedasticity/cluster/serial correlation robust inference is also possible
- **No restrictions** on the nature of x_i or z_i (e.g., each could be binary, or just one of them)

Instrumental Variables

- In Stata the command is `ivreg` (or `ivregress 2sls`):
  - `ivreg y (x = z)`
  - `ivreg y (x = z), robust`
- To proceed with IV, first demonstrate that z_i helps predict x_i (and in the direction suggested by economics or common sense).
  - Easiest way: regress x_i on z_i and do a robust t test.

reg x z

- Some research on "weak instruments" says that, in this simple case, t statistic should be at least $3.2 \approx \sqrt{10}$– much higher than the standard 5%-level critical value.
  - You may often hear people say the F stat of regressing x on z should be greater than 10 (just a rule-of-thumb)
  - Many other formal statistical tests for weak instruments (i.e., whether relevance requirement is satisfied)

Instrumental Variables

- Relevance requirement is important, but at least we can test for it.
- By contrast, the exogeneity requirement is more of an issue. You have to justify it using a "story"
- It is related to the question "where instrumental variables come from"?
  - Randomized eligibility works well as an IV for participation in a program. So $x_i = 1$ if person actually participates. $z_i = 1$ if the person was made eligible.
  - Caution: The fact that a variable is randomized does not always make it exogenous to a model. Economic agents can change their behavior!
  - Example: Angrist (1990, *American Economic Review*), Vietnam draft

    y: earnings; x: Vietnam veteran status; z: draft eligibility

    - $z = 1$ if lottery num. $\leq$ cutoff; $z = 0$ if lottery num. $>$ cutoff
    - $x = 1$ if the person is veteran; $x = 0$ if otherwise

Instrumental Variables: Examples

- Example: Angrist and Evans (1998, *American Economic Review*). Weekly hours equation

$$hours = \beta_0 + \beta_1 kids + u$$

  for women with at least two children (so $kids \geq 2$). One proposed IV is $samesex$, equal to one if the first two children have the same gender.

- Even if gender is exogenous, the family's budget constraint is subsequently affected. (Kids of the same gender can more easily share a room, clothes, and toys.)

- We will illustrate using a (small!) subset of data from Angrist and Evans (`labsup.dta`). Note how large the sample size is, yet IV estimator is barely statistically significant.

Instrumental Variables: Examples

- Describe the dataset

```
. des hours kid samesex

              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------
hours           byte    %8.0g                 hours of work per week, mom
kids            byte    %8.0g                 number of kids
samesex         byte    %8.0g                 first two kids are of same sex

. sum hours kid samesex

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       hours |     31,857    21.22011    19.49892          0         99
        kids |     31,857    2.752237    .9771916          2         12
     samesex |     31,857     .502778    .5000001          0          1
```

Instrumental Variables: Examples

- OLS regression

```
. reg hours kids, robust

Linear regression                               Number of obs     =     31,857
                                                F(1, 31855)       =     585.25
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0178
                                                Root MSE          =     19.325

------------------------------------------------------------------------------
             |               Robust
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -2.664309   .1101318   -24.19   0.000    -2.880171   -2.448446
       _cons |   28.55292   .3200455    89.22   0.000     27.92562    29.18022
------------------------------------------------------------------------------
```

Instrumental Variables: Examples

- Check that *samesex* is relevant for *kids*:

```
. reg kids samesex, robust

Linear regression                               Number of obs     =     31,857
                                                F(1, 31855)       =      40.90
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0013
                                                Root MSE          =     .97658
```

kids	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
samesex	.0699933	.0109439	6.40	0.000	.0485429	.0914437
_cons	2.717045	.007806	348.07	0.000	2.701745	2.732346

Instrumental Variables: Examples

- IV with heteroskedasticity-robust standard errors:

```
. ivreg hours (kids = samesex), robust

Instrumental variables (2SLS) regression        Number of obs     =     31,857
                                                F(1, 31855)       =       3.19
                                                Prob > F          =     0.0743
                                                R-squared         =          .
                                                Root MSE          =     19.534

------------------------------------------------------------------------------
             |               Robust
       hours |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |   -5.58186   3.127136    -1.78   0.074    -11.71117    .5474471
       _cons |   36.58271   8.606509     4.25   0.000     19.71362    53.45179
------------------------------------------------------------------------------
Instrumented:  kids
Instruments:   samesex
------------------------------------------------------------------------------
```

More than twice as large in magnitude, but 95% CI actually contains zero

Instrumental Variables: Examples

- Correlation between *kids* and *samesex* is small:

```
. corr kids samesex
(obs=31,857)

             |     kids  samesex
-------------+------------------
        kids |   1.0000
     samesex |   0.0358   1.0000
```

- Ratio of IV s.e. to OLS s.e. is $3.127/0.110 \approx 28.4$
- Ratio $\frac{1}{r_{xz}}$ from rule-of-thumb: $1/0.0358 \approx 27.9$
- In this example, there is no way to test whether *samesex* is exogenous. We must **assume** it in order to trust IV estimators to be consistent.

Instrumental Variables

- The point estimates from OLS and IV are very different in this case.
- Do not ignore the possibility that the instrument is somewhat endogenous:

$$\text{plim}(\hat{\beta}_1^{OLS}) = \beta_1 + \frac{\sigma_u}{\sigma_x} \cdot Corr(x, u)$$

$$\text{plim}(\hat{\beta}_1^{IV}) = \beta_1 + \frac{\sigma_u}{\sigma_x} \cdot \frac{Corr(z, u)}{Corr(z, x)}$$

- So even if $Corr(z, u) < Corr(x, u)$, the bias in IV can be larger because $Corr(z, u)$ is blown up by $\frac{1}{Corr(z,x)}$
  - Small $Corr(z, x)$ is not unusual.
  - Angrist and Evans example: the correlation was less than .04

Outline

IV Estimation of Multiple Regression Model

- Sometimes an instrument is exogenous only when other factors have been controlled for. For example,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \quad \mathbb{E}[u] = 0,\ \mathbb{C}ov[u, x_1] \neq 0,\ \mathbb{C}ov[u, x_2] = 0$$

- x_2 is exogenous but it is not enough for identifying the model
  - We have three parameters β_0, β_1, β_2, but only two moment conditions $\mathbb{E}[u] = 0$ and $\mathbb{E}[x_2 u] = 0$.
- So we need at least one IV for x_1, say, z_1:

$$\mathbb{C}ov[z_1, u] = 0 \quad \textbf{(exogeneity)}$$

- If x_2 is not controlled for, z_1 could be an invalid IV due to correlation between x_2 and z_1.

IV Estimation of Multiple Regression Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

- Three moment conditions are available

$$\mathbb{E}[u] = \mathbb{E}[y - \beta_0 - \beta_1 x_1 - \beta_2 x_2] = 0$$

$$\mathbb{E}[z_1 u] = \mathbb{E}[z_1(y - \beta_0 - \beta_1 x_1 - \beta_2 x_2)] = 0$$

$$\mathbb{E}[x_2 u] = \mathbb{E}[x_2(y - \beta_0 - \beta_1 x_1 - \beta_2 x_2] = 0$$

- Method of moment estimation:

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) = 0$$

$$\sum_{i=1}^{n} z_{i1}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) = 0$$

$$\sum_{i=1}^{n} x_{i2}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) = 0$$

IV Estimation of Multiple Regression Model

- However, do not forget the **relevance requirement**: as a valid IV, z_1 must be correlated with x_1
- But the presence of the control variable x_2 complicates the requirement
- Intuitively, we need z_1 to be *partially* correlated with x_1. It is easiest to test with the regression

$$x_1 = \pi_0 + \pi_1 z_1 + \pi_2 x_2 + v$$

- We want to reject $H_0 : \pi_1 = 0$ with high confidence (at small level)
- We can add more exogenous explanatory variables to the regression model. The analysis is the same.

IV Estimation: Examples

- One example (CARD.DTA): return to education

```
. ivreg lwage exper expersq black smsa south smsa66 reg662-reg669 (educ = nearc4), robust

Instrumental variables (2SLS) regression          Number of obs     =      3,010
                                                  F(15, 2994)       =      55.76
                                                  Prob > F          =     0.0000
                                                  R-squared         =     0.2382
                                                  Root MSE          =     .38833
```

lwage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.1315038	.0541436	2.43	0.015	.0253414	.2376663
exper	.1082711	.0234089	4.63	0.000	.062372	.1541702
expersq	-.0023349	.0003488	-6.69	0.000	-.0030188	-.0016511
black	-.1467757	.0525019	-2.80	0.005	-.2497193	-.0438322
smsa	.1118083	.0311448	3.59	0.000	.0507409	.1728757
south	-.1446715	.0291429	-4.96	0.000	-.2018136	-.0875294
smsa66	.0185311	.0205651	0.90	0.368	-.021792	.0588542

Outline

Multiple Instruments

$z = c_1 z_1 + c_2 z_2.$

$cov[z, u] = c_1 cov[z_1, u] + c_2 cov[z_2, u] = 0.$

- Sometimes we have more instruments than necessary, e.g.,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \quad \mathbb{C}ov[u, x_1] \neq 0, \quad \mathbb{C}ov[u, x_2] = 0$$

$$\mathbb{C}ov[z_1, u] = \mathbb{C}ov[z_2, u] = 0, \quad \mathbb{C}ov[z_1, x_1] \neq 0, \quad \mathbb{C}ov[z_2, x_1] \neq 0$$

- Of course, we can use each of z_1 and z_2 as an IV and get two IV estimators, but in general, neither of them is efficient
  - In fact, we can use any linear combination of z_1 and z_2 as an IV as long as it is still correlated with x_1
- So why not regress x_1 on z_1 and z_2?

$$x_1 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 x_2 + v,$$

$$\mathbb{E}[v] = 0, \quad \mathbb{C}ov[z_1, v] = \mathbb{C}ov[z_2, v] = \mathbb{C}ov[x_2, v] = 0$$

  - $x_1^* = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 x_2$ is the best linear prediction of x_1 based on z_1, z_2 and x_2!

$\min_{c_0, c_1, c_2, c_3} \mathbb{E}[(x_1 - (c_0 + c_1 z_1 + c_2 z_2 + c_3 x_2))^2]$

Two Stage Least Squares

$$x_1 = x_1^* + v = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 x_2 + v$$

- Intuition: This equation breaks the endogenous x_1 into two pieces:
  1. "Exogenous" component x_1^*: it is based on exogenous variables z_1, z_2 and x_2
  2. "Endogenous" component v: it must be correlated with u
- Use x_1^* as IV: moment conditions

$$\mathbb{E}[u] = \mathbb{E}[y - \beta_0 - \beta_1 x_1 - \beta_2 x_2] = 0$$

$$\mathbb{E}[x_1^* u] = \mathbb{E}[x_1^*(y - \beta_0 - \beta_1 x_1 - \beta_2 x_2)] = 0$$

$$\mathbb{E}[x_2 u] = \mathbb{E}[x_2(y - \beta_0 - \beta_1 x_1 - \beta_2 x_2] = 0$$

- x_1^* is not observed, but we can regress x_1 on z_1, z_2, x_2 to "estimate" it by the fitted value

Two Stage Least Squares

$$x_1 = x_1^* + v = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 x_2 + v$$

- Intuition: This equation breaks the endogenous x_1 into two pieces:
  1. "Exogenous" component x_1^*: it is based on exogenous variables z_1, z_2 and x_2
  2. "Endogenous" component v: it must be correlated with u
- 1st stage: reg x_1 on z_1, z_2 and x_2 and obtain

$$\hat{x}_1 = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 x_2$$

- 2nd stage: use $\hat{x}_1$ as an IV for x_1

$$\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) = 0$$

$$\sum_{i=1}^{n} \hat{x}_{i1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) = 0$$

$$\sum_{i=1}^{n} x_{i2} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) = 0$$

Two Stage Least Squares

$$x_1 = x_1^* + v = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 x_2 + v$$

- Intuition: This equation breaks the endogenous x_1 into two pieces:
  1. "Exogenous" component x_1^*: it is based on exogenous variables z_1, z_2 and x_2
  2. "Endogenous" component v: it must be correlated with u
- Algebraically, the second stage is equivalent to

$$\text{reg } y \text{ on } \hat{x}_1, x_2$$

- In Stata: still use `ivreg`. For example,

```
ivreg y (x1 = z1 z2) x2
```

Two Stage Least Squares: Remarks

Point estimates $\hat{\beta}$ is right, → s.e., CI, tests, etc. are wrong.

- Do not compute 2SLS manually: inference would be problematic
- Tests after 2SLS are not based on the conventional formula for OLS. But in practice most software will do it correctly and automatically if you have implemented 2SLS
- Include x_2 (more generally, all other exogenous explanatory variables) in the first stage regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \quad x_1 = \hat{x}_1 + \hat{v}$$
$$\rightarrow \quad y = \beta_0 + \beta_1 \hat{x}_1 + \beta_2 x_2 + u + \beta_1 \hat{v}$$

  - Intuitively, we want to guarantee $\hat{v}$ is uncorrelated with x_2.

Detecting Weak IV

- Again, do not forget the relevance requirement
- In the first stage regression

$$x_1 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 x_2 + v$$

  - Test $H_0 : \pi_1 = \pi_2 = 0$ (not the overall significance test)
  - After "partialling out" x_2, z_1 and z_2 jointly help predict x_1
  - Use, for example, F test
- As mentioned before, sometimes IV is correlated with the endogenous explanatory variable, but the correlation is very weak. Then IV estimator or 2SLS could perform poorly and large sample inference could be misleading
  - Rejecting H_0 at usual 5% level is usually not good enough
  - Stock and Yogo's rule-of-thumb: $F > 10$ (do not rely on it!)

2SLS: Example

- MROZ.dta

```
. ivreg lwage (educ=motheduc fatheduc) exper expersq

Instrumental variables (2SLS) regression

      Source |       SS           df       MS      Number of obs   =       428
-------------+----------------------------------   F(3, 424)       =      8.14
       Model |  30.3074256         3  10.1024752   Prob > F        =    0.0000
    Residual |  193.020015       424  .455235885   R-squared       =    0.1357
-------------+----------------------------------   Adj R-squared   =    0.1296
       Total |  223.327441       427  .523015084   Root MSE        =    .67471

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0613966   .0314367     1.95   0.051    -.0003945    .1231878
       exper |   .0441704   .0134325     3.29   0.001     .0177679    .0705729
     expersq |   -.000899   .0004017    -2.24   0.026    -.0016885   -.0001094
       _cons |   .0481003   .4003281     0.12   0.904    -.7387744     .834975
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   exper expersq motheduc fatheduc
------------------------------------------------------------------------------
```

Multiple Endogenous Variables

① reg x_1 on IV, x_3.
reg x_2 on IV, x_3.
② reg y on $\hat{x}_1$, $\hat{x}_2$, x_3

- 2SLS can also be applied when there are more than one endogenous variables

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

$E[u] = 0$

$E[x_3 u] = 0$

  - x_3 is exogenous (uncorrelated with u)
  - x_1 and x_2 are endogenous (correlated with u)
- As before, we still apply 2SLS, using fitted values $\hat{x}_1$ and $\hat{x}_2$ in the 2nd stage
- But we have be careful
  - Necessary (order) condition: we need at least two valid IVs, say, z_1 and z_2
  - Sufficient (rank) condition: More difficult (not covered in this class)
- The basic idea: need to ensure we really have "two IVs" partially correlated with x_1 and x_2
- In Stata: e.g., `ivreg y (x1 x2 = z1 z2 z3) x3`

Outline

Use Multiple Measurements as IVs

- Recall measurement error in x_j's could lead to endogeneity

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2 + u, \quad x_1 = x_1^* + e_1$$

$$\rightarrow \quad y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + (u - \beta_1 e_1)$$

- Solution: find an IV for x_1 that is uncorrelated with u and e_1, but partially correlated with x_1
- One possibility: another noisy measurement of x_1^*

$$z_1 = x_1^* + a_1$$

  - Assume
    1. CEV conditions: $\mathbb{C}ov[a_1, x_1^*] = \mathbb{C}ov[e_1, x_1^*] = 0$, $\mathbb{C}ov[a_1, u] = \mathbb{C}ov[e_1, u] = 0$
    2. $\mathbb{C}ov[e_1, a_1] = 0$
- Since $\mathbb{C}ov[u, x_1^*] = 0$, this suffices for **exogeneity** of z_1 as an IV
- As repeated measurements, x_1 and z_1 must be correlated, and hopefully even after partialling out x_2 (**relevance**).

Use Multiple Measurements as IVs

- An example

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + abil + u, \quad abil \text{ unobserved}$$

- Instead we collect two test scores and believe

$$test_1 = \gamma_1 abil + e_1, \quad \gamma_1 > 0$$

$$test_2 = \delta_1 abil + e_2, \quad \delta_1 > 0$$

γ_1 and δ_1 do not have to be 1 as in CEV assumption

- Replace *abil* with $test_1$:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \gamma_1^{-1} test_1 + (u - \gamma_1^{-1} e_1)$$

  - *educ* is not endogenous, but $test_1$ is.
  - In general, noisy control variables could make OLS estimates of all parameters biased and inconsistent

Use Multiple Measurements as IVs

- An example

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + abil + u, \quad abil \text{ unobserved}$$

- Instead we collect two test scores and believe

$$test_1 = \gamma_1 abil + e_1, \quad \gamma_1 > 0$$

$$test_2 = \delta_1 abil + e_2, \quad \delta_1 > 0$$

γ_1 and δ_1 do not have to be 1 as in CEV assumption

- Replace *abil* with $test_1$:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \gamma_1^{-1} test_1 + (u - \gamma_1^{-1} e_1)$$

  - As we saw, if we assume e_2 is uncorrelated with e_1 and u, and e_1 is uncorrelated with *abil*, we can use $test_2$ as an IV

Outline

Testing Whether a Variable is Endogenous

- As an example, consider the multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \quad \mathbb{C}ov[u, x_2] = 0$$

- We want to test for endogeneity of x_1 ($H_0 : \mathbb{C}ov[x_1, u] = 0$)
- x_2 act as its own IV. Suppose we have IVs, say z_1 and z_2, for x_1
- Then we can test for endogeneity by comparing OLS and 2SLS
  - If x_1 is exogenous, both OLS and 2SLS are consistent
  - If they are different significantly, conclude x_1 is endogenous
  - Otherwise, no strong evidence for endogeneity
  - Of course, this test works because we trust our IVs
  - In Stata: `estat endogenous` (after `ivregress 2sls`)

Outline

Testing Overidentification Restrictions

$cov[u, z_1] = 0.$

- Recall that when we only have one instrumental variable, it is impossible to test for its exogeneity
- Note $\hat{u}_i$ from 2SLS must be uncorrelated with z_i *by construction*
  - Step 1: regress endogenous x_{i1} on the IV z_{i1} and the exogenous x_{i2}

$$x_{i1} = \hat{x}_{i1} + \hat{v}_{i1} = \hat{\gamma}_0 + \hat{\gamma}_1 z_{i1} + \hat{\gamma}_2 x_{i2} + \hat{v}_{i1}$$

  - Step 2: regress y_i on $\hat{x}_{i1}$ and x_{i2}

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 \hat{x}_{i1} + \hat{\beta}_2 x_{i2} + \hat{u}_i$$

  - We know as OLS residual,

$$\frac{1}{n}\sum_{i=1}^{n} \hat{u}_i = 0, \quad \frac{1}{n}\sum_{i=1}^{n} \hat{x}_{i1}\hat{u}_i = 0, \quad \frac{1}{n}\sum_{i=1}^{n} x_{i2}\hat{u}_i = 0 \quad \Rightarrow \quad \frac{1}{n}\sum_{i=1}^{n} z_{i1}\hat{u}_i = 0$$

which holds whether z_1 is actually exogenous or not.

Testing Overidentification Restrictions

- What if we have multiple instruments, say z_1 and z_2 for one endogenous x_1?
  - Step 1: regress endogenous x_1 on the IV z_1 and the exogenous x_2

$$x_{i1} = \hat{x}_{i1} + \hat{v}_{i1} = \hat{\gamma}_0 + \hat{\gamma}_1 z_{i1} + \hat{\gamma}_2 z_{i2} + \hat{\gamma}_3 x_{i2} + \hat{v}_{i1}$$

  - Step 2: regress y on $\hat{x}_1$ and x_2

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 \hat{x}_{i1} + \hat{\beta}_2 x_{i2} + \hat{u}_i$$

  - This time we know

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{\gamma}_1 z_{i1} + \hat{\gamma}_2 z_{i2})\hat{u}_i = 0 \quad \not\Rightarrow \quad \frac{1}{n}\sum_{i=1}^{n} z_{i1}\hat{u}_i = 0, \quad \frac{1}{n}\sum_{i=1}^{n} z_{i2}\hat{u}_i = 0$$

- In general, with more instruments than needed (e.g., $k+q$ IVs and k endogenous explanatory variables), we can test overidentifying restrictions

$$H_0 : \mathbb{C}ov[u, z_1] = 0, \cdots, \mathbb{C}ov[u, z_{k+q}] = 0$$

Testing Overidentification Restrictions

$$y = \beta_0 + \underbrace{\beta_1 x_1 + \cdots + \beta_k x_k}_{\text{endogenous}} + \underbrace{\beta_{k+1} x_{k+1} + \cdots + \beta_{k+l} x_{k+l}}_{exogenous} + u$$

instruments: $z_1, z_2, \cdots, z_{k+q}$

- Overidentification test*:
  1. Estimate the model by 2SLS to get $\hat{u}_i$
  2. Regress $\hat{u}_i$ on *all exogenous variables* (IVs and exogenous explanatory variables). Obtain the R-squared R_1^2
  3. Under H_0, $nR_1^2 \overset{a}{\sim} \chi_q^2$, q is the number of IVs minus the total number of endogenous explanatory variables.
  4. If we reject, conclude that *at least some* of the IVs are not exogenous.
- In Stata, use

```
ivregress 2sls y (x1=z1 z2) x2
estat overid
```

Testing Overidentification Restrictions

- **Important**: use the overidentification test with caution!
  - Even if you fail to reject H_0, it **does not** mean the IVs are exogenous (recall we never say we accept the null hypothesis)
  - The test could have low power (fail to reject when H_1 is true)
  - When you reject H_0, there could be many other reasons (finite sample errors, functional form, etc.) for the rejection (not because IVs are really invalid)
- It is routine to report this test statistic when the model is overidentified. If the p value is too small, you should cast doubt on your estimates
- But do not rely on it too much. Use the test to help you make decisions, but you still need a "story" to justify the exogeneity