CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Financial Econometrics

Lecture VI:

Endogeneity, Instrumental Variable, and 2SLS estimator

Narapong Srivisal, Ph.D.

Outline

Causes of

try to put a lot of controls to the model as regressors so that only noises remain in error term

Endogeneity

- Omitted Variable Bias left some random variables that can explain y
  ↳ in the error term → endogeneity problem
- Measurement Errors
- Simultaneous Equations
- Reversed causality → easily correct

Resolutions for Endogeneity Problem

- Instrument
- Instrumental Variable Estimator
- 2SLS Estimator
- Over-identifying Test

Endogeneity

- Suppose that we want to study *ceteris paribus* effects on Y by using the linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + U$$

- We call the regressor X_j **exogenous** (variable) **if** $\boldsymbol{Cov(X_j, U) = 0}$
- We call the regressor X_j **endogenous if** $\boldsymbol{Cov(X_j, U) \neq 0}$
- We have an **endogeneity problem** if there is an endogenous regressor. (at least 1 random variable x is endogenous)
- This problem results in the OLS estimator being inconsistent and biased.

$$y = \beta_0 + \beta_1 x + (\underbrace{\gamma w + u}_{\text{error}})$$

Suppose we run reg y x

if w and x are correlated → endogeneity problem!

* $cov(x,x) = var(x)$

$$\hat{\beta}^{OLS} = \frac{\widehat{cov}(y,x)}{\widehat{var}(x)} \xrightarrow[p]{WLLN} \frac{cov(y,x)}{var(x)}$$

$y \to \beta_0 + \beta_1 x + \gamma w + u$

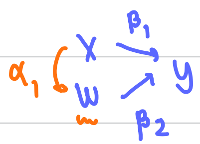
$cov(\beta_1 x, x) = \beta_1 * var(x)$

$$\beta_1 + \frac{cov(\gamma w, x) + \cancel{cov(u,x)}^{\,0}}{var(x)}$$

$$\hat{\beta}_1^{OLS} \to \beta_1 + \frac{\gamma\, cov(w,x)}{var(x)}$$

can omit this iff $cov(w,x) = 0$

$$y = \beta_0 + \beta_1 x + \beta_2 w + u \quad ; \quad cov(x,u) = 0, \quad cov(w,u) = 0$$



diff

$w = \alpha_0 + \alpha_1 x + \varepsilon$ when x moves 1 unit, w moves α_1

$y = \gamma_0 + \gamma_1 x + \eta$ → when x↑ 1 unit, how much y↑ if not control for w? not get pure effect of x

$\beta_1 + \alpha_1 \cdot \beta_2$ get only correlation indicator not pure effect

β_1: ceteris paribus effect of x on y

$\alpha_1 \cdot \beta_2$: contaminated effect from w

MSF Chula

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Omitted Variable Bias

- Suppose that we want to study the effect of X on Y, and the true underlying model is

$$Y = X'\beta + \boldsymbol{\gamma W} + \boldsymbol{U}$$

- If we omit the regressor W in our OLS estimation, the OLS estimator for β is then

$$\hat{\beta}_{omitted} = \left(\sum_{i=1}^{n} X_i X_i{}'\right)^{-1} \left(\sum_{i=1}^{n} X_i Y_i\right)$$

- By the WLLN and CMT, suppose $E[XU] = 0$, this converge in probability to

$$E[XX']^{-1}E[X\boldsymbol{Y}] = E[XX']^{-1}E[X(X'\beta + \boldsymbol{\gamma W} + \boldsymbol{U})]$$

$$= \beta + \gamma E[XX']^{-1}E[\boldsymbol{XW}]$$

- So, our **OLS will be consistent only if $E[XW] = 0$**, that is

$$E[W] = E[X_1 W] = E[X_2 W] = \cdots = E[X_k W] = 0$$

- Or only if the omitted variable W is uncorrelated with each of the other regressors
- Similarly for unbiasedness,

$$E\left[\hat{\beta} \middle| X_1, \ldots, X_n\right] = E\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i{}'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i (X_i'\beta + \boldsymbol{\gamma W} + \boldsymbol{U_i})\right)|X\right]$$

$$= \beta + \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i{}'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} X_i (\boldsymbol{\gamma E[W_i|X] + E[U_i|X]})$$

- So, we **need a stronger condition that $E[W|X] = 0$ in addition to $E[U|X] = 0$ to get unbiased OLS estimator** in this case.

Correlation vs Causation Revisit

- Suppose that we want to study the effect of education on wage, and the true underlying model is

$$\ln(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + U$$

where, supposedly, the correlations between these two regressors and the error term are zero.

- Then, β_1 captures the ceteris paribus effect of education on wage.
- Suppose that education and experience are correlated, that is $\alpha_1 \neq 0$ in the regression model:

$$exper = \alpha_0 + \alpha_1 educ + \epsilon$$

- If we omit experience in the regression, **OLS will not consistently estimate the effect (β_1)** but how $\ln(wage)$ is correlated with education, which **include both the effect of education and the effect of experience**:

MSF Chula

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

AACSB ACCREDITED | EFMD EQUIS ACCREDITED | EFMD EPAS ACCREDITED

```
. * Interpretation of correlation vs causation
. *---------------------------------------------
. reg lwage educ, nohead
```

without exper

```
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0827444   .0075667    10.94   0.000     .0678796    .0976092
       _cons |   .5837727   .0973358     6.00   0.000     .3925563    .7749891
------------------------------------------------------------------------------
```

γ_1 (educ)

```
. reg lwage educ exper, nohead
------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .0979356   .0076224    12.85   0.000     .0829613    .1129099
       exper |   .0103469   .0015551     6.65   0.000     .0072919     .013402
       _cons |   .2168544    .108595     2.00   0.046     .0035183    .4301904
------------------------------------------------------------------------------
```

$\hat{\beta}_1$ (educ): effect of educ on wage

$\hat{\beta}_2$ (exper): effect of exper

```
. reg exper educ, nohead
------------------------------------------------------------------------------
       exper |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -1.468182   .2042881    -7.19   0.000    -1.869507   -1.066858
       _cons |    35.4615   2.627905    13.49   0.000     30.29898    40.62402
------------------------------------------------------------------------------
```

$\hat{\alpha}_1$ (educ)

Remark: .0827444 = .0979356 + .0103469*(-1.468182)

$$\underbrace{\hat{\gamma}_1}_{\text{total movement}} = \underbrace{\hat{\beta}_1}_{\text{direct effect}} + \underbrace{\hat{\beta}_2 \hat{\alpha}_1}_{\text{contaminated effect from exper}}$$

Measurement Errors

Suppose we want to estimate the impact of X on Y according to the linear regression model

$$Y = \beta_0 + \beta_1 X + U$$

Measurement Error in a Regressor

- We **cannot measure X precisely but with an error**. That is we can only observe and use $\tilde{X}$ as the regressor, where

$$\tilde{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{\epsilon}; \;\; E[\epsilon|X] = 0; \;\; E[\epsilon|Y] = 0; \;\; E[\epsilon U] = 0$$

- So, in term of $\tilde{X}$, your true model is

$$Y = \beta_0 + \beta_1(\tilde{\boldsymbol{X}} - \boldsymbol{\epsilon}) + U = \beta_0 + \beta_1 \tilde{X} + \underbrace{(\boldsymbol{U} - \boldsymbol{\beta_1 \epsilon})}_{\substack{\textbf{error term} \\ \textbf{if regressing } Y \textbf{ on} \\ \tilde{X} \textbf{ instead of } X}}$$

$wage = \beta_0 + \beta_1 edu + controls + (A + \varepsilon)$

left in error term bcs it cannot be captured
└ cannot be separate from ε

Innate capabilities of people → cause endogeneity
⇓
solved by use other than OLS

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

brokerage account size

income
(cannot be observed precisely
(only observe

reported income = income + noise

measurement error → cannot really regress y on x

$\tilde{x} = x + \varepsilon$

observed, not, not

can only reg y $\tilde{x}$

try to write in form of y & $\tilde{x}$;

$y = \beta_0 + \beta_1(\tilde{x} - \varepsilon) + u$

$y = \beta_0 + \beta_1 \tilde{x} + (u - \beta_1 \varepsilon)$

↳ $cov(\tilde{x}, \varepsilon) \neq 0$!!! ⇒ endogeneity problem

impact of reported income on y ⇒ cheating
not income

Measurement error in x

reg y $\tilde{x}$

$$\hat{\beta}_1^{OLS} = \frac{\widehat{cov}(y, \tilde{x})}{\widehat{var}(\tilde{x})} \xrightarrow{\text{converge to}} \frac{cov(y, \tilde{x})}{var(\tilde{x})}$$

$y = \beta_0 + \beta_1 \tilde{x} + u - \beta_1 \varepsilon$

not corr with any

if not corr with $\tilde{x}$

$$\frac{\beta_1 \cancel{var(\tilde{x})} - \beta_1 cov(\tilde{x}, \varepsilon)}{\cancel{var(\tilde{x})}}$$

$$\beta_1 - \beta_1 \left(\frac{cov(\underbrace{\tilde{x}}_{x+\varepsilon}, \varepsilon)}{var(\underbrace{\tilde{x}}_{x+\varepsilon})} \right)$$

if $cov(x, \varepsilon) = 0$ what remains:

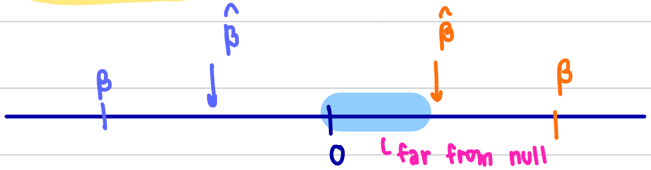
$\longrightarrow var(\varepsilon)$

$\longrightarrow var(x) + var(\varepsilon)$

$$\beta_1 - \beta_1 \left(\frac{var(\varepsilon)}{var(x) + var(\varepsilon)} \right) \rightarrow \beta_1 \left(1 - \frac{var(\varepsilon)}{var(x) + var(\varepsilon)} \right) = \beta_1 \left(\frac{var(x) + \cancel{var(\varepsilon)}}{var(x) + var(\varepsilon)} - \frac{\cancel{var(\varepsilon)}}{var(x) + var(\varepsilon)} \right)$$

$$= \beta_1 \left(\frac{var(x)}{var(x) + var(\varepsilon)} \right) \quad \begin{matrix} > 0 \\ < 0 \end{matrix}$$

converge to just a fraction of β



"Biased toward zero" $\rightarrow$ estimate is closer to zero that than population β what it is suppose to be

want to do inference

$H_0: \beta = 0$ vs $H_1: \beta \neq 0$

reject H_0 with $\hat{\beta}^{OLS}$ $\rightarrow$ don't have to move to use diff estimator

(safely say that we reject this in population as well as we know the direction in biase

safely make conclusion that $\beta \neq 0$

If not reject H_0 with $\hat{\beta}^{OLS}$ $\rightarrow$ have to use new estimator — not easy!

cannot make any conclusion whether we can reject or not reject

- Since we use $\tilde{X}$ instead of X as the regressor, our OLS is

$$\hat{\beta}_1 = \frac{\hat{\sigma}_{\tilde{X}Y}}{\hat{\sigma}^2_{\tilde{X}}} \xrightarrow{P} \frac{cov(\tilde{X},\ Y)}{Var(\tilde{X})} = \frac{cov\left(\tilde{X},\ \beta_0+\beta_1\tilde{X}+(U-\beta_1\epsilon)\right)}{Var(\tilde{X})}$$

$$= \beta_1 + \frac{cov(X+\epsilon, U-\beta_1\epsilon)}{Var(X+\epsilon)}$$

$$= \beta_1 - \beta_1\left(\frac{Var(\epsilon)}{Var(X)+Var(\epsilon)}\right)$$

$$= \beta_1\left(\frac{Var(X)}{Var(X)+Var(\epsilon)}\right)$$

→ 1
→ not have much biase ↗ get closer to β_1
↳ if this is close to zero
↳ if large → get whole thing close to zero

If data is noisy
↓
closer to zero
↓
cannot detect any effect of x on y

- Notice that the parenthesis is always greater than zero and less than one. So, our OLS estimator is inconsistent and **biased toward zero**.
- If the error term has a larger variance relative to the correct measure of regressor X, the estimate tend to get closer to zero

Measurement Error in the Regressand

- We cannot measure Y precisely but observe $\tilde{Y}$ for use as the regressand, where

$$\tilde{\boldsymbol{Y}} = \boldsymbol{Y} + \boldsymbol{\eta};\;\; E[\eta|X] = 0;\;\; E[\eta|Y] = 0;\;\; E[\eta U] = 0$$

- So, in term of $\tilde{Y}$, our true model is

$$\tilde{Y} = \beta_0 + \beta_1 X + \underbrace{(\boldsymbol{U} - \boldsymbol{\eta})}_{\substack{\text{error term} \\ \text{if regressing} \\ \tilde{Y} \text{ instead of } Y \text{ on } X}}$$

- Then, our OLS estimator is

$$\hat{\beta}_1 = \frac{\hat{\sigma}_{X\tilde{Y}}}{\hat{\sigma}_X^2} \xrightarrow{P} \frac{cov(X,\tilde{Y})}{Var(X)} = \frac{cov\big(X,\beta_0+\beta_1 X+(U-\eta)\big)}{Var(X)}$$

- So, in this case, our **OLS is still consistent under the regular assumption that $\boldsymbol{E[XU] = 0}$.**

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

port size — Income

If this is noisy problem ← cannot be observed

$\tilde{y} = y + \eta$ ← measurement error in y

(reported)

Need to run reg $\tilde{y}$ x

$\tilde{y} - \eta = \beta_0 + \beta_1 x + u$

$\tilde{y} = \beta_0 + \beta_1 x + \underbrace{u + \eta}_{\text{error term}}$

typically not correlated bcs of diff r.v.

If η is not related to x

$cov(x, \eta) = 0 \Rightarrow$ no endogeneity problem

Reversed causality

$\downarrow y^{\uparrow} = \beta_0 + \beta_1 x + u^{\uparrow}\downarrow$

reg x y ⇒ cofficient will not converge to $\frac{1}{\beta_1}$ ⇒ cannot make conclusion that $\beta = \frac{10}{7}$

$0.7 = \frac{7}{10}$

rewrite
x on LHS
y on RHS

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

$x = \frac{-\beta_0}{\beta_1} + \left(\frac{1}{\beta_1}\right) y - \frac{u}{\beta_1}$

u and y are correlated → when u↑, y↑

endogeneity problem

↓ just use y as regresand

may happen with

Endogeneity
- Omitted Variable Bias → control as another regressor in model
- Measurement Errors
  - in x: biased toward zero
  - in y: not a problem
- Simultaneous Equations → happen at the same time ; x & y satisfy various equations at the same time

Reversed causality problem ⇒ switch back
⇓
require instrument to solve this

Simultaneous Equations

- Suppose we have two equations:

$$\boldsymbol{Y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 X} + \boldsymbol{U} \text{ and } \boldsymbol{Y} = \boldsymbol{\gamma_0} + \boldsymbol{\gamma_1 X} + \boldsymbol{\eta}$$

- If **both equations hold simultaneously**, we can solve for X from the equation: $\beta_0 + \beta_1 X + U = Y = \gamma_0 + \gamma_1 X + \eta$ and get

$$X = -\frac{\beta_0 - \gamma_0}{\beta_1 - \gamma_1} + \frac{\eta}{\beta_1 - \gamma_1} - \frac{U}{\beta_1 - \gamma_1}$$

- Then,

$$cov(X, \eta) = cov\left(\frac{\eta}{\beta_1 - \gamma_1}, \eta\right) = \frac{Var(\eta)}{\beta_1 - \gamma_1} \neq 0$$

$$cov(X, U) = cov\left(-\frac{U}{\beta_1 - \gamma_1}, U\right) = -\frac{Var(U)}{\beta_1 - \gamma_1} \neq 0$$

- Therefore, if we only observe (X, Y) when both equations hold simultaneously, OLS will neither be a consistent estimator for the regression $Y = \beta_0 + \beta_1 X + U$ nor $Y = \gamma_0 + \gamma_1 X + \eta$.

- This problem is often present when estimating demand or supply curves, If you use transaction data ⇒ $Q^{supply} = Q^{demand}$ ⇒ Q & P satisfy both equations at the same time

$$Q^{Demand} = \beta_0 + \overset{-}{\beta_1} P + U$$

$$Q^{Supply} = \gamma_0 + \underset{+}{\gamma_1} P + \eta$$

- We would expect $\beta_1 < 0$ and $\gamma_1 > 0$.
- However, available data are typically transactions occurred at market equilibria where price equates quantity demanded and supplied:

$$Q^{Demand} = Q^{equilibrium} = Q^{Supply}$$

$$P^{Demand} = P^{equilibrium} = P^{Supply}$$

- With equilibrium transaction data only, we would see the estimated coefficient of price between β_1 and γ_1.

correlated   correlated

$$q^{demand} = \beta_0 + \beta_1 X + U = \gamma_0 + \gamma_1 X + \eta = q^{supply}$$

$$\Downarrow$$

$$(\beta_1 - \gamma_1) X = -\beta_0 + \gamma_0 + \eta - U$$

$$X = -\left(\frac{\beta_0 - \gamma_0}{\beta_1 - \gamma_1}\right) + \frac{1}{\beta_1 - \gamma_1}(\eta - U)$$

cannot use OLS estimator!

When people use more credit card → people will use less cash

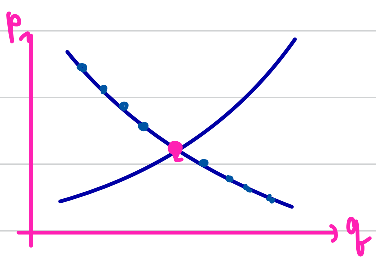
1st equation   E.g. Cash holding in economy $= \beta_0 + \beta_1$ card usage $+$ controls $+ U$

(−)

can also intepret as causation as well !!!

no cash → force to use credit card

2nd equation : card $= \gamma_0 + \gamma_1$ cash $+$ controls $+ \eta$



1st solution : move from transaction data, to survey data e.g. find willingness to pay for a given price

very carefully decide this, not straight forward

Resolution for Endogeneity

- Recall that we use the moment condition $E[U] = E[XU] = 0$ to solve for β coefficients.
- and use the sample-counterpart conditions from FOCs $\sum_{i=1}^{n} \widehat{U}_i = \sum_{i=1}^{n}(X_i \widehat{U}_i) = 0$ to solve for the OLS estimator.
- E.g. for bivariate case: $\beta_1 = \frac{cov(X,Y)}{Var(X)} \Rightarrow \hat{\beta}_1^{OLS} = \frac{\hat{\sigma}_{XY}}{\hat{\sigma}_X^2}$
- With the endogeneity problem, we do not have $E[XU] = 0$.
- A resolution for this is to **find another moment condition to solve for the β and construct a consistent estimator from a sample counterpart of the new solution**.

Instrument

need to find z → not recommend

- A valid **instrument** or **instrumental variable**, Z, is a random variable satisfying:
  - **Instrumental exogeneity:** it is uncorrelated with the error term

    $cov(Z, U) = 0$ **or** $E[ZU] = 0$

    relevance to x
  - **Instrumental relevance:** it is relevance by still correlated with the endogenous variable X

    $cov(Z, X) \neq 0$
- We can use a valid instrument to help estimate a model with endogeneity problem.

OLS estimator

start with assumption

- $y = \beta x + u$
- $E[u] = 0$ for all intepretation
- $\cancel{E[XU] = 0}$ X if endogeneity

$E[\ ] = 0$

solve for $\beta \Rightarrow$

$\cancel{\beta_0 = E[y] - \beta_1 E[x]}$ ($\bar{y}$, $\hat{\beta}_1$, $\bar{x}$) use sample counterpart converge to population

$$\cancel{\beta_1 = \frac{cov(y,x)}{var(x)}}$$

($c\hat{ov}$, $v\hat{ar}$) formula change

MSF Chula

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

Instrumental Variable Estimator

A resolution to the endogeneity problem is to use the moment conditions related to instruments to solve for the coefficient parameter β's and find a sample counterpart as the estimator.

Bivariate Model with One Endogenous Variable and One Instrument

- Use $E[U]=0$ and $[ZU]=0$ instead of $E[XU]=0$ to solve for β_0 and β_1:

$$\ast\ E[U]=0 \Rightarrow E[Y-(\beta_0+\beta_1 X)]=0 \Rightarrow \beta_0 = E[Y]-\beta_1 E[X]$$

$$\ast\ E[ZU]=0 \Rightarrow 0 = E[Z(Y-\beta_0-\beta_1 X)]$$

use this to solve for β $= E[Z(Y-E[Y]+\beta_1 E[X]-\beta_1 X)]$

$$\Rightarrow \beta_1 = \frac{E[Z(Y-E[Y])]}{E[Z(X-E[X])]} = \frac{cov(Z,Y)}{cov(Z,X)}$$

still need $cov(z,x) \neq 0$ if not, we cannot find β

z : instrumental variable of estimator

- Hence, we can get for the sample counterpart,

$$\hat{\beta}_1^{IV} = \frac{\hat{\sigma}_{ZY}}{\hat{\sigma}_{ZX}} \xrightarrow[\text{WLLN}]{\text{converge to}} \frac{cov(z,y)}{cov(z,x)}$$

& continuous mapping theorem

$$\hat{\beta}_0^{IV} = \bar{Y}_n - \left(\frac{\hat{\sigma}_{ZY}}{\hat{\sigma}_{ZX}}\right)\bar{X}_n$$

- This is called the **instrumental variable (IV) estimator**.

- Under the regular conditions for WLLN, the IV estimator for β_1 and β_0 is consistent because

$$\hat{\beta}_1^{IV} = \frac{\hat{\sigma}_{ZY}}{\hat{\sigma}_{ZX}} \xrightarrow{P} \frac{cov(Z,Y)}{cov(Z,X)} = \beta_1$$

$$\hat{\beta}_0^{IV} = \bar{Y}_n - \left(\frac{\hat{\sigma}_{ZY}}{\hat{\sigma}_{ZX}}\right)\bar{X}_n \xrightarrow{P} E[Y] - \beta_1 E[X] = \beta_0$$

- However, the IV estimator is typically biased why it's hard to find z?

e.g. $\ln[wage] = \beta_0 + \beta_1 edu + controls + (A + U)$

A: innate capability that correlate with edu

* need to find z that correlated to edu but not A

z: the distance btw
- birthplace — genius can be borned anywhere
- the near river — good education places near river

⊙.⊙!

* Recommend to review a lot of literature
  - face endogeneity → what is the cause?
  - replicate from other literature

eg. $\ln q^D = \beta_0 + \beta_1 \ln p + controls + U$

controls: budget, lifestyle — can correlate to controls

U: · behaviors · attitude

z: tax, fees, cost of production

this may not be instrument
e.g. เห็นเครื่องผลิตรักนักแล้วอยากซื้อ >_<

1) Think what's in U?

2) try to think of x-related that avoid such factors in U.

we may have many instruments for one endogenous x

Bivariate Model with One Endogenous Variable but Several Instruments

- Suppose that we have m valid instruments $Z_1, Z_2, \dots, Z_m$
- Then, $cov(Z_1, U) = 0, cov(Z_2, U) = \cdots = cov(Z_m, U) = 0,$

$$cov(Z_1, X) \neq 0, cov(Z_2, X) \neq 0, \dots, cov(Z_m, X) \neq 0$$

- We can use either each of $Z_1, Z_2, \dots, Z_m$ or their linear combinations as an instrument in the IV estimator function.
- Caution: if using a linear combination $\alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + \cdots + \alpha_m Z_m$, we need to check that

$$\alpha_1 cov(Z_1, X) + \alpha_2 cov(Z_2, X) + \cdots + \alpha_m cov(Z_m, X) \neq 0$$

most correlate with x

- **The most efficient way** is to use the best linear projection of X given $Z_1, Z_2, \dots, Z_m$ (the second interpretation) as the instrument:

$$\boxed{X = \underbrace{\boldsymbol{\pi_0 + \pi_1 Z_1 + \pi_2 Z_2 + \cdots + \pi_m Z_m}}_{\text{BLP}[X|Z_1, \dots, Z_m]}} + \eta$$

want to find best linear approximation of x

Linear combinations :

e.g. 2 valid instruments ; $cov(z_1, x) = \frac{1}{2}$, $cov(z_2, x) = \frac{1}{3}$ $\Big\}$ $\underbrace{\alpha_1 z_1 + \alpha_2 z_2}$ $= \bar{z}$

if it correlate with x → can use it

$cov(\bar{z}, u) = \alpha_1 \cancel{cov(z_1, u)}^{0} + \alpha_2 \cancel{cov(z_2, u)}^{0}$

- $cov(\bar{z}, x) = \underset{-2}{\alpha_1} \underset{\frac{1}{2}}{cov(z_1, x)} + \underset{+3}{\alpha_2} \underset{\frac{1}{3}}{cov(z_2, x)}$

$\hat{\beta}^{IV}$ → consistent
↘ BIASED → may not equal to population] need large enough sample size (ไม่มันเวอ biased

---

Weak instrument → $\underbrace{cov(x, z)}$ low → biased & slow consistent

want this to be high

- However, since we don't see the whole population of $X, Z_1, \ldots, Z_m$, we don't know $\pi_0, \pi_1, \ldots, \pi_m$ but can use the OLS to consistently estimate them.
- In other words, **use $\widehat{X}^{OLS} = \widehat{\pi}_0 + \widehat{\pi}_1 Z_1 + \cdots + \widehat{\pi}_m Z_m$ as the instrument**.
- Then, our IV estimator is

$$\hat{\beta}_0^{IV} = \bar{Y}_n - \hat{\beta}_1^{IV} \bar{X}_n; \quad \hat{\beta}_1^{IV} = \frac{cov(\hat{X}, Y)}{cov(\hat{X}, X)} = \frac{\hat{\sigma}_{\hat{X}Y}}{\hat{\sigma}_{\hat{X}X}}$$

$\frac{\hat{\sigma}_{\hat{x}y}}{Var(\hat{x})}$

$\hat{\beta}$ reg y $\hat{x}$

if you $\hat{x}$ as instrument

- Remarks, by the properties of OLS,

$$\sum_{i=1}^{n} \hat{\eta}_i = \sum_{i=1}^{n} (Z_1 \hat{\eta}_i) = \cdots = \sum_{i=1}^{n} (Z_m \hat{\eta}_i) = 0 \Rightarrow \sum_{i=1}^{n} (\hat{X}_i \hat{\eta}_i) = 0$$

- This implies that the sample covariance of $\hat{X}$ and $\hat{\eta}$ is zero.
- Therefore,

$$\hat{\sigma}_{\hat{X}X} = \widehat{cov}(\hat{X}, X) = \widehat{cov}(\hat{X}, \hat{X} + \hat{\eta}) = \widehat{Var}(\hat{X}) + \underbrace{\widehat{cov}(\hat{X}, \hat{\eta})}_{=0} = \hat{\sigma}_{\hat{X}}^2$$

by property of ols estimator

2SLS or TSLS Estimator

no one do this → ask STATA to do this

- Because $\hat{\sigma}_{\hat{X}X} = \hat{\sigma}^2_{\hat{X}}$, the IV estimator is equal to

$$\hat{\beta}_1^{IV} = \frac{\hat{\sigma}_{\hat{X}Y}}{\hat{\sigma}_{\hat{X}X}} = \frac{\hat{\sigma}_{\hat{X}Y}}{\hat{\sigma}^2_{\hat{X}}}$$

2 steps least square

prove this is the same

- which is the OLS estimator of β_1 of the regression $Y = \beta_0 + \beta_1 \hat{X} + U$.
- This means that we can also get the estimate by doing the 2 steps of the OLS estimation method:
    i. **Regress X on $Z_1, \dots, Z_m$ and get the fitted value $\hat{X}$**
    ii. **Regress Y on $\hat{X}$ to get consistent estimates for β_0, β_1**
- Hence, people call this method as **Two-Step Least Squares (TSLS or 2SLS) estimator**, which gives exactly the same estimated values as the IV estimator with BLS$[X|Z_1, \dots, Z_m]$ as the instrument.

MSF Chula

CHULALONGKORN BUSINESS SCHOOL FLAGSHIP FOR LIFE

AACSB ACCREDITED · EFMD EQUIS ACCREDITED · EFMD EPAS ACCREDITED

Multivariate Model
One Endogenous Variable, One Instrument

- Suppose that we have a multivariate regression with X_1 as the only endogenous regressor, and the other regressors are exogenous:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + U$$

- Again, two conditions are required for Z to be a valid instrument:
  - **Instrumental exogeneity:** $Cov(Z, U) = 0$
  - **Instrumental relevance:** $Cov(Z, X_1 | X_2, \dots, X_k) \neq 0$
- The instrumental relevance says that Z is required to be correlated with X_1 after controlling for the other exogenous regressors.
- This means π_1 in the following best linear projection cannot be zero:

$$X_1 = \pi_0 + \pi_1 Z + \pi_2 X_2 + \cdots + \pi_k X_k + \eta$$

- So, for this model we have $k+1$ moment conditions to solve for β:

$$E[U] = E[ZU] = E[X_2 U] = \cdots = E[X_k U] = 0$$

- Or in matrix notation $E[WU] = 0$, where $W = (1, Z, X_2, \dots, X_k)$.
- As we did in case of the OLS before, we can solve for β

$$E[WU] = 0 \Rightarrow E[W(Y - X'\beta)] = 0 \Rightarrow E[WY] - E[WX']\beta = 0$$

$$\beta = E[WX']^{-1} E[WY]$$

- And the instrumental variable estimator is the sample counterpart:

$$\hat{\beta}^{IV} = \left(\frac{1}{n}\sum_{i=1}^{n} W_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} W_i Y_i\right)$$

- Similar to the bivariate case, if we use $\hat{X}_1 = \hat{\pi}_0 + \hat{\pi}_1 Z + \hat{\pi}_2 X_2 + \cdots + \hat{\pi}_k X_k$ as the instrument instead, $W = (1, \hat{X}_1, X_2, \dots, X_k)$

$$\hat{\beta}^{IV} = \left(\frac{1}{n}\sum_{i=1}^{n} W_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} W_i Y_i\right) = \left(\frac{1}{n}\sum_{i=1}^{n} W_i W_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} W_i Y_i\right) = \hat{\beta}^{2SLS}$$

Multivariate Model with Several Instruments

- First, suppose that we have $m > 1$ instruments for a single endogenous regressor X_1. Again, we can construct an instrument $\hat{X}_1$ from the OLS fitted value of

$$X_1 = \pi_0 + \gamma_1 Z_1 + \cdots + \gamma_m Z_m + \pi_2 X_2 + \cdots + \pi_k X_k + \eta$$

- Next, if the model have several, say q, endogenous regressors:

$$Y = \beta_0 + \underbrace{\beta_1 X_1 + \cdots + \beta_q X_q}_{\text{endogenous terms}} + \underbrace{\beta_{q+1} X_{q+1} + \cdots + \beta_k X_k}_{\text{exogenous terms}} + U$$

- For this case, **we need to have $m \geq q$ instruments** and find q fitted values, for $1 \leq j \leq q$:

$$X_j = \pi_{j0} + \gamma_{j1} Z_1 + \cdots + \gamma_{jm} Z_m + \pi_{j(q+1)} X_{q+1} + \cdots + \pi_k X_k + \eta$$

MSF Chula

CHULALONGKORN BUSINESS SCHOOL
FLAGSHIP FOR LIFE

AACSB ACCREDITED · EFMD EQUIS ACCREDITED · EFMD EPAS ACCREDITED

easier to do this — no need to do 2 steps

exogenous regressor

y

endo

family income

instruments

```
. webuse hsng2, clear
(1980 Census housing data)

. ivregress 2sls rent i.region pcturban popden (hsngval = faminc hsng)
------------------------------------------------------------------------------
        rent |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     hsngval |   .0041469   .0009073     4.57   0.000     .0023686    .0059252
             |
      region |
     N Cntrl |    -15.666   18.82128    -0.83   0.405    -52.55502    21.22303
       South |  -5.606589    17.2554    -0.32   0.745    -39.42656    28.21338
        West |  -70.46023   28.48546    -2.47   0.013    -126.2907   -14.62974
             |
    pcturban |  -.1464167   .5531242    -0.26   0.791     -1.23052    .9376869
      popden |  -.0057151   .0038934    -1.47   0.142     -.013346    .0019158
       _cons |   76.19836   32.39852     2.35   0.019     12.69843    139.6983
------------------------------------------------------------------------------
Instrumented:  hsngval
Instruments:   2.region 3.region 4.region pcturban popden faminc hsng
```

$\hat{\beta}$ (Coef.)

$se(\hat\beta)$ (Std. Err.)

$\hat\beta / se(\hat\beta)$ (z) (can be used to do inference)

exogenous variable is instrument by itself

```
. quietly reg hsngval faminc hsng i.region pcturban popden, nohead

. predict xhat, xb

. reg rent xhat i.region pcturban popden, nohead
------------------------------------------------------------------------------
        rent |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        xhat |   .0041469     .00041    10.11   0.000       .00332    .0049739
             |
      region |
     N Cntrl |    -15.666   8.505803    -1.84   0.072    -32.81958     1.48759
       South |  -5.606587   7.798145    -0.72   0.476    -21.33305    10.11987
        West |  -70.46022   12.87329    -5.47   0.000    -96.42169   -44.49876
             |
    pcturban |  -.1464166   .2499706    -0.59   0.561    -.6505303    .3576971
      popden |  -.0057151   .0017595    -3.25   0.002    -.0092635   -.0021667
       _cons |   76.19836   14.64169     5.20   0.000     46.67057    105.7261
------------------------------------------------------------------------------
```

wrong

$\hat{\beta}$

no. of
\# endogenous variable   \# instrumental variable

q m

3 cases

q $<$ m : over identified

$=$ just identified

many people use this bcs it's hard to find z

$>$ ~~under identified~~ cannot be estimated

over identifying test

But there's problem:

don't observe → cannot test, need story

$cov(u, z) = 0$

$cov(x, z) \neq 0$

test this because we observe → test: ?

MSF Chula

CHULALONGKORN BUSINESS SCHOOL FLAGSHIP FOR LIFE

Overidentifying Test

- Recall an instrumental variable, Z, must satisfy:
  - Instrumental exogeneity: $cov(Z, U) = 0$ or $E[ZU] = 0$
  - Instrumental relevance: $cov(Z, X) \neq 0$
- Note also that we observe samples of X, Y, Z but we do not observe the error term $U = Y - \beta X$ because we do not know the parameter β.
- Hence, we could check the instrumental relevance condition by estimating $cov(Z, X)$ but **we could not estimate $cov(Z, U)$** to check the instrumental exogeneity condition.
- However, if the number of candidate instruments we have is larger than the number of endogenous variables (i.e. $m > q$ or overidentifying case), we can conduct an **Over-identifying Test** to see if we have sufficient valid instruments out of the candidates.

```
. webuse hsng2, clear
(1980 Census housing data)
```

after run regression

```
. Quietly ivregress 2sls rent i.region pcturban popden ///
                            (hsngval = faminc hsng)

. estat overid

  Tests of overidentifying restrictions:

  Sargan (score) chi2(1) =  .036164  (p = 0.8492)
  Basmann chi2(1)        =    .0304  (p = 0.8616)
```

test statistics

p-value of the test

fail to reject H_0

Hansen test also popular

- Both Sargan and Basmann Tests for overidentifying have the **null hypothesis** that there are sufficient valid instruments.

null hypothesis

- Therefore, rejecting the null hypothesis (p-value < 10%, 5%, or 1%) implies that we do not have enough valid instruments.
- Note: degree of freedom of both tests equals to the number of instruments minus the number of endogenous variables in the model

Review from previous class :

$y = \beta x + u$

↳ uncorrelated : exogenous

↳ correlated : endogenous

⇑ endogeneity problem

because of

~~in y~~

biased toward zero ← in x

- omitted variable
- measurement error
- reversed causality
- simultaneous equations

1st put control

2nd can you make inference?

3rd try change $x \rightleftarrows y$

4th create new z (instrument)

↳ ① $cov(u, z) = 0$ exogeneity

② $cov(x, z) \neq 0$ relevance

↱ cannot test this bcs we cannot observe error term u

⇓ There're 3 tests

- overidentifying test # IV > # endo
  - ↳ Hansen test
  - ↳ Sangen test
  - ↳ Basmann

|| enough valid IV

H_0, H_1

⇓

we prefer not to reject H_0

$$wage = \beta_0 + \beta_1 edu + control + \overbrace{(cap + u)}^{error}$$

only 1 endogenous variable

emotional

cannot be captured but correlate to edu

not count this

IV estimator = 2SLS estimator