Outline Chapter 15

1. **The identification problem—when do we need instruments**

2. **What makes a good instrument—conditions**

3. **The IV/2SLS estimator—single variable case**
    a. **How IV estimator is constructed**
    b. **Proof that is consistent**
    c. **How 2SLS estimator is constructed**
    d. **Proof is same as IV estimator with single var**

4. **Comparison of OLS and IV estimators**
    a. **Comparing bias when have weak instruments**
    b. **Comparing standard errors**

5. **Multiple variable case**
    a. **Multiple exogenous vars**
    b. **Multiple instruments**
    c. **Multiple endogenous vars and multiple instruments**

6. **Testing for endogeneity**
    a. **Endogeneity of X: Hausman test**
    b. **Endogeneity of Z: overid tests**

Primary Concerns in Estimation:

1. Biased coefficients—incorrect magnitude/sign
2. Biased standard errors—efficiency, incorrect inferences

Sources of biased coefficients

1. Mis-measured X / errors-in-variables—attenuation bias (bias to zero)

2. Omitted Variables (Z→X, Z→Y and therefore if omit Z is in error)

3. Reverse causation (X→Y, Y→X)

Chapter 13/14: panel models one way to deal with time invariant forms of omitted variables.
Chapter 15: another method for dealing with omitted variables – instrumental variables (IV). IV can be used to solve error-in-variables and simultaneous causality problems as well as omitted variables.

The basic idea:
If *x* is correlated with *u*, we can think about decomposing *x* into two components,

(1) the part that is uncorrelated with *u* and
(2) the part that is correlated with *u*.

If we can find information that allows us to isolate the first part we can use that part of the variation in *x* to consistently estimate β_1

In Chapter 13/14, relied on assumption that it is often the fixed (time invariant) part of X that is correlated with u. Estimating with dummy variables removed that variation.

The Basic Model

$y_i = \beta_0 + \beta_1 x_i + u_i$ and (y_i, x_i, z_i) i= 1,..., n

where *i* denotes entities, *y* is the dependent variable, and *x* is an explanatory variable for each entity and *z* is an instrument.

If $Cov(x_i, u_i) \neq 0$ the OLS estimator is inconsistent.

IV uses an additional variable *z* to isolate the part of *x* that is uncorrelated with *u*.

Conditions for Valid Instruments

(1) Instrument Relevance $Cov(z_i, x_i) \neq 0$

(2) Instrument Exogeneity $Cov(z_i, u_i) = 0$

Together these imply that Z only affects Y through X

Note: We *can* test whether Cov(zi, xi) ≠ 0 (How?)
We usually *cannot* test whether Cov(zi, ui) = 0 (Why not?)

Identification—Construction of IV estimator in single variable case

Identification of a parameter in this context means that we can write β_1 in terms of population moments (parameters) that can be estimated using sample data.

From

$y_i = \beta_0 + \beta_1 x_i + u_i$ and (y_i, x_i, z_i) i= 1,..., n

Recall that $\beta_1 = Cov(y_i, x_i)/Var(x_i) = \sigma_{xy}/\sigma_x^2$

We can write this in terms of how vary with z:

$Cov(y_i, z_i) = \beta_1 Cov(z_i, x_i) + Cov(z_i, u_i)$

$Cov(z_i, u_i)=0$ so

$\beta_1 = Cov(y_i, z_i) / Cov(z_i, x_i) = \sigma_{zy}/\sigma_{zx}$

Sample analog: $\hat{\beta}^{IV}_1 = \frac{1/n-1 \sum (z_i - \bar{z})(y_i - \bar{y})}{1/n-1 \sum (z_i - \bar{z})(x_i - \bar{x})}$

Again note: if Z=X, then get OLS estimator

In matrix notation:

$$\hat{\beta}_1^{IV} = (Z'X)^{-1}(Z'y)$$

Large Sample Properties—Is this a consistent estimator of beta?

$$\hat{\beta}_1^{IV} = \frac{\hat{\sigma}_{zy}}{\hat{\sigma}_{zx}} \overset{p?}{\rightarrow} \beta_1$$

Work with numerator:

$$\sigma_{zy} = 1/_{n-1} \sum (z_i - \bar{z})(y_i - \bar{y})$$

$$= 1/_{n-1} \sum (z_i - \bar{z})(\beta_1 (x_i - \bar{x}) + (u_i - \bar{u}))$$

$$= \frac{\beta_1 \sum (z_i - \bar{z})(x_i - \bar{x})}{n-1} + \frac{\sum (z_i - \bar{z})(u_i - \bar{u})}{n-1}$$

$$= \frac{\beta_1 (n-1)\hat{\sigma}_{zx}}{n-1} + \frac{\sum (z_i - \bar{z})(u_i)}{n-1}$$

$$= \beta_1 \hat{\sigma}_{zx} + \frac{\sum (z_i - \bar{z})(u_i)}{n-1}$$

So

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

$$= \beta_1 + \frac{\frac{1}{n}\sum (z_i - \bar{z})(u_i)}{\frac{1}{n}\sum (z_i - \bar{z})(x_i - \bar{x})}$$

Now apply LLN:

$$\hat{\beta}^{IV} \overset{p}{\rightarrow} \beta_1 + \frac{\frac{1}{n}\sum (z_i - \mu_z)(u_i)}{\sigma_{zx}}$$

As N gets large, second term gets small assuming that Cov(z,u) = 0 → consistent estimator

In small samples, in practice usually biased. This is because if x is correlated with u, in practice is very rarely case that z and u have exactly zero correlation. Underlines importance of large n with IV for consistency.

The Two Stage Least Squares Estimator

Assumptions:

1. Linear in parameters $y_i = \beta_0 + \beta_1 x_i + u_i$
2. (y_i, x_i, z_i) are iid draws—random sampling
3. No perfect collinearity—rank condition
4. $E(u_i) = 0$ and $Cov(z_i, u_i) = 0$—Exogenous IVs

→1-4 give us consistency

5. $E(u^2_i|z_i) = \sigma^2$ Homoskedasticy →Efficiency.

If the assumption are satisfied, β_1 can be estimated using a particular IV estimator called two stage least squares (2SLS or TSLS).

2SLS

First stage:

$$x_i = \pi_0 + \pi_1 z_i + v_i \rightarrow \hat{x}_i = \hat{\pi}_0 + \hat{\pi}_1 z_i$$

z_i is exogenous→ $\pi_0 + \pi_1 z_i$ represents the part of x_i that can be predicted by z_i → this part is therefore also exogenous

The other part of x_i is the v_i → this is the part that must be related to u_i

So 2SLS uses the exogenous part and disregards the v_i

Second stage:

$y_i = \beta_0 + \beta_1 \hat{x}_i + u_i$ This gives us the 2SLS estimates of β_0 and β_1

Consistency of 2SLS

Is the same as the formula for the IV estimator we already wrote down and proved was consistent? Does $\hat{\beta}^{IV} = \hat{\beta}^{TSLS}$?

Does this equal

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

Does it converge in probability to σ_{zy}/σ_{zx}?

$$\hat{\beta}_1^{TSLS} = \frac{1/n \sum (\hat{x}_i - \bar{\hat{x}})(y_i)}{1/n \sum (\hat{x}_i - \bar{\hat{x}})^2}$$

$\hat{x}_i = \hat{\pi}_0 + \hat{\pi}_1 z_i$

Work with numerator:

$$\text{cov}(\hat{x}_i, y_i) = \hat{\sigma}_{\hat{x}y} = 1/n \sum_i (\hat{x}_i - \bar{\hat{x}}) y_i = \sum_i \hat{\pi}_0 + \hat{\pi}_1 z_i - \hat{\pi}_0 - \hat{\pi}_1 \bar{z}) y_i = \sum \hat{\pi}_1 (z_i - \bar{z}) y_i = \hat{\pi}_1 \hat{\sigma}_{zy}$$

Work with denominator:

$$\text{var}(\hat{x}_i) = \hat{\sigma}^2{}_{\hat{x}} = 1/n \sum_i (\hat{x}_i - \bar{\hat{x}})^2 = \sum_i \hat{\pi}_0 + \hat{\pi}_1 z_i - \hat{\pi}_0 - \hat{\pi}_1 \bar{z})^2 = \sum \hat{\pi}_1^2 (z_i - \bar{z})^2 = \hat{\pi}_1^2 \hat{\sigma}_z^2$$

So,

$$\hat{\beta}_1^{TSLS} = \frac{\hat{\sigma}_{zy}}{\hat{\pi}_1 \hat{\sigma}_z^2} = \frac{\hat{\sigma}_{zy}}{\hat{\sigma}_{zx}} \text{ since } \hat{\pi}_1 = \frac{\hat{\sigma}_{zx}}{\hat{\sigma}_z^2}$$

With a single var these two estimators are identical

Comparison of Bias in OLS and 2SLS & Importance of Testing For Instrument Relevance

Again, recall 2 conditions for a valid instrument

1. Instrument Relevance Cov(zi, xi) $\neq 0$
2. Instrument Exogeneity Cov(zi, ui) = 0

The Weak Instruments Problem

Question: What if (1) x and z are weakly correlated and (2) z and u are weakly correlated? Is using the instrument better than OLS or not?

Result: Weak correlation between *x* and *z* can lead to large asymptotic bias even if *z* and *u* are only moderately correlated.

Go back to Consistency proof. At end showed

$$\hat{\beta}^{TSLS} \xrightarrow{p} \beta_1 + \frac{\frac{1}{n}\sum (z_i - \mu_z)(u_i)}{\sigma_{zx}} = \beta_1 + \frac{Cov(zi, ui)}{Cov(zi, xi)}$$

$$= \beta_1 + \frac{Corr(zi, ui)}{Corr(zi, xi)} . \frac{\sigma_u}{\sigma_x}$$

because Corr = Cov/stddev

Asymptotic bias will be big when
(1)zi and ui are highly correlated,
(2) zi and ui are not very correlated, but zi and xi are not very correlated either

Recall that showed for OLS:

$$\hat{\beta}^{OLS} \xrightarrow{p} \beta_1 + Corr(x_i, u_i) . \frac{\sigma_u}{\sigma_x}$$

So which one will be less biased depends on the relative magnitude of these correlations.

Rule of Thumb Test: If F statistic for z vars (test that coeffs on zs are all equal to zero in first stage where regress x on the z's) is less than 10, you have weak instruments

More complex forms of this test when have multiple instruments, multiple endogenous variables (Bound, Jaeger Baker; Shea; Anderson test stats)

2SLS Estimation in Multiple Variable Case—One Endogenous Explanatory Variable, 1 instrument, multiple exogenous vars

---

Digression:

How do we come up with estimators?

- One method is least squares method—OLS came from minimizing sum of squared errors
- Another method is known as "method of moments". A different class of estimators. These come from matching up sample statistics (functions of data) to some function of population parameters. Turns out (surprise) OLS is also a MOM estimator (covariance/variance)

  Going to show a MOM estimator here
- Later, we'll also describe maximum likelihood estimators. There pick estimators by choosing parameters that maximize likelihood of drawing our particular sample. Will do later with binary dependent variable models.

Notation: Here use y for the dependent var, x for endogenous independent var, w for exogenous independent vars, and z for the instruments. Wooldridge uses y for the dependent var and all endogenous independent vars, z for exogenous independent vars and all instruments.

Again: endogenous means correlated with error

Suppose model is $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 w_{1i} + u_i$

This is the **structural** model: β_1 represents the causal effect of x_1 on y

If x1 is endogenous (correlated with u), then all coefficients will be biased.

Recall need an instrument z that is both **exogenous** and **relevant**

How do we express these conditions in the multiple variable case?

1st condition: Instrument Exogeneity: Need instrument for z for x1 where
E(u) =0, Cov(w1, u) = 0 and cov(z, x1) = 0

Express as "Moment conditions"—get estimators for β_0 ,β_1, β_2 by solving

$$\sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 w_{i1}) = 0$$
$$\sum_i w_{1i} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 w_{i1}) = 0$$
$$\sum_i z_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 w_{i1}) = 0$$

2nd condition: Instrument Relevance**:** Need z to be correlated with x1, though now we have to also take w1 into account as well.
Easiest to write this relevance condition down by writing the **reduced form:**

$x_i = \pi_0 + \pi_1 w_{1i} + \pi_2 z_i + v_i$ --Endogenous variables as functions of ONLY exogenous variables

Need $\pi 2 \neq 0$

Note that this reduced form is also the 1st stage. All of this generalizes easily if have multiple ws.

2SLS Estimation in Multiple Variable Case—Single Endogenous Explanatory Variable, Multiple Instruments

First stage:

$$x_{1i} = \pi_0 + \gamma_1 w_{1i} + \ldots + \gamma_k w_{ki} + \pi_1 z_{1i} + \ldots + \pi_j z_{ji} + v_i \rightarrow \hat{x}_i$$

This is also called the "Reduced Form"

Run a regression of endogenous independent var on ALL exogenous vars (instruments and other exogenous vars in model)

Need to have at least one $\pi \neq 0$

If we have just one instrument (as before), we say that the model is "**just identified**" or "exactly identified"

If we have more than one instrument, we say the model is "**overidentified**"

Second stage:

$$y_i = \beta_0 + \beta_1 \hat{x}_i + \beta_2 w_{1i} + \ldots + \beta_{k+1} w_{ki} + u_i$$

2SLS Estimation in Multiple Variable Case—Multiple Endogenous Explanatory Variable, Multiple Instruments

Model is $y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \beta_{k+1} w_{1i} + u_i$
Again, generalized easily. Now have multiple equations to estimate in the first stage—1 for each of the Xs. In second stage will plug the predicted values for each into equation for Y.

How many instruments do we need? At least one for every endogenous variable. Order condition. Again, if have more instruments than Xs, have an **overidentified** model.

STATA note:

```
ivreg yvar indepvar (xvars = ivvars)
```

will report just the second stage estimates.

```
ivreg yvar indepvar (xvars = ivvars), first
```

will report all the first stage results as well

Comparing OLS and IV Estimators

1. **Consistency:**

   Recall that OLS is biased if x is correlated with u. IV is biased if z is correlated with u, and even if that correlation is small, bias may be larger than OLS bias if z is weakly correlation with x.

2. **Comparison of OLS and 2SLS standard errors**

Recall sample analog for OLS estimator standard error:

$$se(\hat{\beta}_1) = \sqrt{\frac{1/n-2\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}x_i^2}}$$

2SLS: $\text{var}(\hat{\beta}_1) = \frac{\sigma_u^2}{n\sigma_x^2\rho_{xz}^2}$

Sample analog:

$$se(\hat{\beta}_1) = \sqrt{\frac{1/n-2\sum_{i=1}^{n}\hat{u}_i^2}{n\left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right)R_{xz}^2}} = \sqrt{\frac{1/n-2\sum_{i=1}^{n}\hat{u}_i^2}{\left(\sum_{i=1}^{n}x_i^2\right)R_{xz}^2}}$$, R^2_{xz} from regression of x_i on z_i

Don't need to know formula, do need to know
Implications:

- Like OLS estimator, variance gets smaller as n grows
- R^2 is less than 1→2SLS/IV se will be larger than OLS ones (so want to use OLS unless OLS bias is smaller than IV bias)
- when R^2_{xz} is small (weak instrument) SE will be particularly large
- when z=x→ R^2_{xz}=1→OLS variance

Testing For Endogeneity

Checking if x is endogenous; Checking if z is endogenous

Again, recall 2 conditions for a valid instrument

1. Instrument Relevance $\quad Cov(z_i, x_i) \neq 0$
2. Instrument Exogeneity $\quad Cov(z_i, u_i) = 0$

Exogeneity hard to satisfy. Can we check if z is truly exogenous?

Also recall that standard errors for IV are larger than OLS—do we even need IV estimates? Is x really endogenous or not?

Testing Endogeneity of Xs—The Hausman test

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 w_{1i} + u_i$$

Suppose are not sure if Cov(xi ,ui) = 0 , but do have a valid instrument for x.

Like Hausman test for FE/RE, here have 2 estimators:

Null: Cov(xi ,ui) = 0
OLS: Unbiased, efficient under null, Biased until alternative
TSLS: Unbiased in either case, but under null is less efficient

Hausman (1978) Is the difference between OLS and IV estimates statistically significant? If it is, then x must be endogenous.

(Wooldridge method) Construction of test statistic:

(1) Estimate **reduced form** for x by regressing it on all exogenous vars (all ws and zs)→ get residuals $\hat{v}_i$

Since all zs are uncorrelated with ui, then testing if x is uncorrelated with ui is equivalent to testing if vi are uncorrelated with ui

(2) Add residuals to **structural** equation:

(3) Regress y on x, w, and $\hat{v}_i$
Test if coefficient on $\hat{v}_i$ =0. If not, then x is endogenous

Note that the coeff on x in this regression will be same as coeff in a TSLS regression

Multiple variables→ do an F-test

Alternative construction-- just like Hausman test with FE vs RE:

$$(\hat{\beta}_{IV} - \hat{\beta}_{OLS})'[V(\hat{\beta}_{IV}) - V(\hat{\beta}_{OLS})]^{-1}(\hat{\beta}_{IV} - \hat{\beta}_{OLS}) \sim \chi^2_k$$

Where k is # of parameters in structural model. Also t test version just like before as well when have single x and single z

STATA—just like Hausman test with FE vs RE:

--Run OLS model
--estimates store betaols
--Run IV model
--estimates store betaiv
--Hausman betaiv betaols
--Null: betaiv = betaols If reject, use IV estimates

When will Hausman test fail to reject? (Prefer OLS)

- x really is exogenous
- var-cov matrix is large—large IV standard errors because z is weakly correlated with x
- Have a bad instrument—z is not really exogenous and so IV and OLS are the same because BOTH of them are biased in same direction

Last 2 reasons why need theory and not just this test.

Testing Endogeneity of Zs—The Overidentification test

If our model is exactly identified (exactly same number of zs as endogenous xs), can't test whether z are exogenous or not. Why not?

But if overidentified (extra zs), turns out can test. Another LM test.

Null: All zs are uncorrelated with ui
Alternative: At least one is correlated with ui

Construction of test statistic:

(1) Estimate under the null: Compute 2SLS estimates for structural equation→ get residuals $\hat{u}_i$

(2) Regress $\hat{u}_i$ on all ws, zs. None of these should be correlated with ui if null is true.

(3) Calculate nR^2 from that regression ~ $\chi 2$ with # of overidentifying parameters (# of Zs - # of endogenous Xs) If nR^2 is big, then we reject the null→at least one of our instrument is not valid.

Note that is model is exactly identified, the R^2 will be nearly zero.

Alterative construction: Check whether each estimator using just identified first stages (using subsets of Z's one at a time)—compare alternative estimators to see if are equal with these subsets of Z

Caveats for Over id tests:
Null: All zs are uncorrelated with ui

- Potential for Type II error: If IV estimates are imprecise (big standard errors)→ get low test statistics and will fail to reject null that estimates using alternative Z's are the same. However, may not be that the Z's are exogenous, just that they are weakly correlated with x (see formula for standard error)

- Potential for Type I error: Alternatively, may have very precisely estimated IVs. However, the implied "treatment" from the IVs may have different effects. In other words, IV estimates pick up effect of x on that marginal population (the population whose choice of x is affected by z). An alternative estimator using a different z may affect a different marginal population. If those populations have different (heterogenous) effects of x on y, the IV estimates will be very different. This is NOT because the IVs are invalid/endogenous, but because the treatment effects implied by each IV are different.

Other sections we are skipping:

--Interpretation of R2
--Autocorrelation/Heteroskedasticity/Panel data
If your project involves this type of data, read it!

<u>Summary of all of these tests:</u>

1. Are the x's exogenous? (Correlated with errors?)
   - ➔ Hausman test. Need a valid instrument to perform it
   - ➔ Back it up with theory

2. Is z relevant? Is it corr with the xs?
   - ➔ Simplest version--F test of joint significance of z's in the first stage—want F>10

3. Is z exogenous? Is it correlated with the errors in the structural model?
   - →Over id test. Need multiple instruments to perform it
   - →Back it up with theory

Applications of IV: Errors-in-Variables

--Have already mentioned that can use IV if have omitted vars that are correlated with Xs. Next chapter will show how IV works when reverse causation.

Final reason why coeff on X might be biased? Measurement error in X

- Solution 1: Use another measure as an instrument:
  Suppose that have 2 "bad" measures of true x*: x1 and x2

  Can use one as an instrument for other. Recall first stage will only pick up the part of x1 that is related to x2. If the measurement error in the two vars is independent, only the "true" part will remain.

- Solution 2: Use another exogenous var as an instrument.