Instrumental Variables

EC420 MSU Online

Justin Kirkpatrick
Last updated June 08, 2021

Lectures:

(1) Where We're At and Motivation

(2) Motivation: Causal Interpretation

(3) Introducing Instrumental Variables

(4) Valid Instruments

(5) More on Instruments

(6) Two-Stage Least Squares (2SLS)

(7) Testing IV Assumptions

(8) IV in R

(9) Simultaneous Equations

(10) Simultaneity Bias

(11) Identification in Simultaneous Equations

Where We're At and Motivation

top

So, while we are no longer focusing on the *mechanics* of estimating $\hat{\beta}$, we are focusing on the assumptions and properties.

Any time we see a regression without the "hat" on $\hat{\beta}$, we will think

> "a-ha, I know how to estimate β, and under which assumptions it is unbiased!"

Parameter of interest

We focus on assumptions because they tell us whether or not we have a good estimate of the "parameter of interest"

D is our "variable of interest".

The β on D is our "parameter of interest"

Interpreting regressions

Last week, we worked a lot with $E[Y|D]$.

I really want to remind everyone that *this is the same as a regression we know and love*:

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

$$E[Y|D=1] = E[\beta_0 + \beta_1 \times 1 + u] = \beta_0 + \beta_1 + 0$$

$$E[Y|D=0] = E[\beta_0 + \beta_1 \times 0 + u] = \beta_0 + 0$$

Which means that:

$$E[Y|D=1] - E[Y|D=0] = (\beta_0 + \beta_1) - (\beta_0) = \beta_1$$

And last week we learned that $E[Y|D=1] - E[Y|D=0]$ is $SATE$ + selection bias

$E[u|D] = 0$ implies $E[Y_0|D = 1] = E[Y_0]$

- Which means **no selection bias.**
- The selection bias problem **is** a violation of MLR.4

When we worry that the error term has something in it that is:

- Correlated with D
- And affects y

We are worried that:

1. $E[u|D] \neq 0$ (u changes with D)
2. $E[Y_0|D = 1] \neq E[Y_0|D = 0]$ (selection bias)

Anything that is related to D and the potential outcomes of Y that is "in the error term" (not in the regression) poses a problem.

Selection bias is a type of MLR4 violation

We can also add other controls in this regression:

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

If D is binary then we can still write:

$$\beta_1 = E[Y|D = 1, X] - E[Y|D = 0, X]$$

We just add in the other X's in this notation.

We could calculate, say, $\hat{E}[Y|D = 1, X = 1]$ by taking all the observations where $D = 1$ and $X = 1$ and taking the mean of Y.

It's harder when X is continuous.

- We'd either have to condition on every possible value of X or we'd only be able to use the β_1 from the regression

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

Motivation and Causal Interpretation

top

What *are* our parameters of interest anyways?

Depends on what we're trying to explain:

- Are we testing an economic theory?
- Then our model tells us what the parameter of interest may be:

> We are studying the effect of raising the minimum wage. A simple model of production and wages tells us that wage, w, is equal to the *marginal product of labor*. A worker whose marginal production is less than the minimum wage would not be hired. Thus, if we look at employment change when the minimum wage changes, then we should be able to *test this model*.

- We are interested in **employment** as the outcome, y
- We are interested in **the minimum wage** as the variable of interest
  - That is what is changing employment, y

What *are* our parameters of interest anyways?

To set up our regression, we have to know what the "unit of observation" may be. Here, let's say we have county-level monthly variation in minimum wage (panel data).

- This just means that minimum wage is set at the county level, i and it *can* change at the month level, t.

$$employment_{it} = \beta_0 + \beta_1 minwage_{it} + \Phi_i + \Gamma_t + u$$

Φ_i is a shorthand way of writing "a fixed effect for every county i". Similar for Γ_t.

Our model tells us that $employment_{it}$ will be lower (if $\beta_1 < 0$) in months t and counties i where $minwage_{it}$ is higher.

This means the model tells us something about $\frac{\partial employment}{\partial minwage} = \beta_1$.

β_1 is our parameter of interest.

The equation

$$employment_{it} = \beta_0 + \beta_1 minwage_{it} + \Phi_i + \Gamma_t + u$$

has two-way fixed effects. In R code:

```
myModel <- lm(employment ~ minwage + as.factor(individual) + as.factor(time), data = myData)
coeftest(myModel, vcov = vcovHC(myModel, 'HC1')) # for robust se
```

Just to make sure the notation Φ_i and γ_t doesn't take you by surprise. It looks a lot more familiar in R code.

What *are* our parameters of interest anyways?

The test of $\beta_1 = 0$ tells us whether or not the data rejects the simple *model of production and wages* we constructed **if** we can say it is unbiased.

That's why it's our "parameter of interest".

Controls (also called Covariates)

Notice that we had other controls: fixed effects for county i in Φ_i and fixed effects for month t in Γ_t

- These were *not* the parameter of interest. They were **controls** since there may be some unobserved things about each county or time period that would also affect $employment_{it}$.
- **Controls** or **covariates** are anything that is observable or can be included in a regression, like Γ_t, that vary with the variable of interest ($minwage_{it}$) and the outcome $employment_{it}$.

Statistical controls

Anything we include in our regression will be a "statistical control" (the x_j's).

- They will help to explain our outcome variable, y
- Each control variable is used in partialling out
- The prediction error is minimized ("least" in OLS)

Let's think about the "treatment effects" framework

- The variable of interest is the "treatment"
  - $minwage_{it}$

Recall our ealier lectures on ATE:

- Potential outcomes (Y_{i0}, Y_{i1})
  - We only observe one for each i

And we get *selection bias* when:

- We try to compare $E[Y_1|D=1]$ to $E[Y_0|D=0]$
- Selection bias:

$$E[Y_1|D=1] - E[Y_0|D=0] = \underbrace{E[(Y_1 - Y_0)|D=1]}_{\text{SATE}} + \underbrace{E[Y_0|D=1] - E[Y_0|D=0]}_{\text{selection bias}}$$

- Which occurs when people have treatment selected based on (Y_{0i}, Y_{1i})

When our statistical controls also control for selection,

- Then we are doing **program evaluation** with a **treatment effects** model
- And, most important, our results have a **causal** interpretation

Random assignment helps with selection bias

- Remember when we talked about *random assignment*?
- If we can **randomly assign treatment** we don't have to worry about selection bias

Similarly, if *conditional on our controls, treatment is as good as randomly assigned*, and we have controlled for all other confounders, then we have a **causal interpretation**.

Let's take our minimum wage example:

$$employment_{it} = \beta_0 + \beta_1 minwage_{it} + \Phi_i + \Gamma_t + u$$

- We might think minimum wage is not randomly assigned (thus, potential for selection bias).
- Maybe we think that counties with growing tech companies are more likely to increase the minimum wage (higher average wages, income inequality)
- They are also more likely to have higher employment anyways
  - $(Y_{0i}, Y_{1i}) \not\perp D$ and $Y_{0i}|D = 1 > Y_{0i}|D = 0$

So what's an econometrician to do?

- Control for $tech_{it}$, the share of employees in county i at time t in tech.

$$employment_{it} = \beta_0 + \beta_1 minwage_{it} + \beta_2 tech_{it} + \Phi_i + \Gamma_t + u$$

Conditions for Causality summarized

See W3.7(e) and 7.6

1. We have a treatment, D: Note Wooldridge uses w
2. We may have controls, x
3. D is independent of (Y_{0i}, Y_{1i}) conditional on x
    - "Treatment ignorability" or "unconfoundedness of assignment"
    - $(Y_{0i}, Y_{1i})|X \perp D$

Number 3 is always true when treatment is randomly assigned, but random assignment almost never holds in economics.

- This is our new focus - how to get **as-good-as-random assignment** so that we have a causal interpretation of the coefficient of interest.

What about our minimum wage. Was that "causal"?

$$employment_{it} = \beta_0 + \beta_1 minwage_{it} + \beta_2 tech_{it} + \Phi_i + \Gamma_t + u$$

- Why else would a county i would "select into" a higher minimum wage?
- What "controls" do we have to address those?
- Do they make assignment to treatment (higher $minwage$) "as good as randomly assigned?

Let's discuss in class

Selection on Observables

When we can name the things that:

1. affect y,
2. may drive selection,

Then we have **selection on observables**.

We have an easy solution for selection on observables:

Just include them in the regression.

Selection on Unobservables

What if selection into treatment is affected by something *unobserved*?

- We have controls for Φ_i in our regression.
- It's just a dummy for each county i in our data.
- This will control for *anything* that is always present in a county i
  - Maybe it's a county with a great University like MSU - they tend to have higher employment.
  - Φ_{Ingham} will account for this, as long as it's always true over every time period.
  - Would Φ_i account for a *growing* tech presence in a county?
- Remember our two-way fixed effects example?

When there are unobserved things that

1. affect y, and
2. may drive selection, we say we have **selection on unobservables**, which implies selection bias.

We do not have an easy solution to selection on unobservables

If they are common to all i, then a fixed effect like Φ_i (dummy or categorical variables) controls for them, which is helpful.

If they vary over time within an individual and are unobserved, then we have a problem.

We can move now into one potential solution: instrumental variables

Introducing Instrumental Variables

top

It's all about causality...

- Our treatment, D (or w in Wooldridge), may or may not be independent of (Y_{0i}, Y_{1i})
- In our employment/minimum wage example, a higher minimum wage may be:
  - Randomly assigned (**super!**)
  - The result of something else that might affect $employment$
  - For instance, some county i might have expectations of a strong local economy coming
  - Maybe they have a new factory about to open. Or lots of (unobserved) tech workers.
  - Because they think the economy will be strong, they decide it's a good time to increase the minimum wage.
  - Due to the new factory, employment goes up
  - At the same time, and also due to the new factory, *minimum wage* goes up
  - And we have selection bias!

How can we get "as good as randomly assigned treatment conditional on the x's"

- Well, we could include a binary for the presence of the factory.
- But that's not in our data.
  - We don't even know about it!
  - But we'd still be worried about it.

An "instrument" is something that:

- Induces variation in the variable of interest (w in Wooldridge, D in MM)
- Is as "good as randomly assigned" conditional on other x's
- But does **not change the outcome, y, except through it's affect on the variable of interest**.

Let's call the instrument Z, the variable of interest D, and the outcome Y.

Let's think about the concept:

What if we could play God and manipulate something that makes the minimum wage increase?

Of course, since we're playing God, we can randomly choose which counties get the thing that makes minimum wage increase.

Then, we wouldn't have to worry much about selection, *especially* for those counties we manipulated.

It would sort of solve our problem.

You know, if we could play God

Valid Instruments

top

Let's call the instrument Z, the variable of interest D, and the outcome Y.

There are three parts to a valid instrument. Let's discuss them first, then see why we need them.

This is like back in multiple regression when we made some assumptions, put them together, got an estimator, and then found that those assumptions gave us an unbiased estimator.

First, Z must have a causal effect on D

The instrument, Z, has to have a casual effect on the variable of interest, D.

- In our example, this means Z has to *cause a change in minimum wage*
- This is the *relevant first stage* condition

Second, Z must be as good as randomly assigned

The instrument, Z, cannot be determined by the omitted variable / selection bias we're trying to get out.

- In our example, this means Z cannot be the result of the (unobserved) expected factory opening and improved economic conditions.
- This is the *independence assumption*

Third, Z must affect the outcome only through the variable of interest

The instrument, Z, cannot have a direct affect on the outcome

- In our example, Z would only affect $minwage$, and would not affect $employment$
- This is the *exclusion restriction*.

These three conditions are necessary to justify any instrument.

Let's draw this out

Let's switch examples to the one in MM:

> MM is interested in the effect of a charter school, KIPP, on test scores. The outcome, y is a student's test score. The treatment, D, is attending a KIPP charter school. We are worried that a simple regression of $testscore_i = \beta_0 + \beta_1 KIPP_i + u_i$ will be biased. Why? Because kids who opt to attend a charter school are not like those who don't - foremost, because by definition the attendees have parents who are more likely to be involved in their education. Parental involvement *also* increases test scores, so we have classic *selection bias*.

- y = Test score (continuous)
- D = KIPP attendance, a charter school: $D = 1$ means attending.
- Z = The instrument...

MM uses the charter school offer lottery

In the case of KIPP, the district holds a lottery for *eligibility* for a KIPP transfer. Out of all parents who enter the lottery, only *some randomly selected group of them* get the option to enroll.

Let's say $Z = 1$ if the student wins the lottery; $Z = 0$ otherwise.

The assumptions for an instrument:

- Is it a *relevant first stage*
  - That is, does $Z = 1$, winning the lottery, increase the probability that $D = 1$, the student enrolls?
  - Answer:

- Is the lottery *independent of the omitted variable/selection bias*?
  - That is, is winning the lottery, Z, unassociated with parental involvement in the child's education?
  - Answer:

- Does winning the lottery meet the *Exclusion restriction*
  - That is, does the instrument affect scores *only* through it's affect on attending KIPP?
  - Answer:

How do they work?

From MM:

The effect of *winning the lottery*, $Z = 1$, on test scores y is:

{Effect of winning lottery on test scores} =
{Effect of attending KIPP on test scores} $\times$ *{Effect of winning on attending KIPP}*

$$\frac{dY}{dZ} = \frac{dY}{dD} \times \frac{dD}{dZ}$$

- The middle term is what we're interested in, but we observe only the first and last

Let's re-write:

{Effect of attending KIPP on test scores} =
{Effect of winning lottery on test scores}/{Effect of winning on attending KIPP}

Using expectations:

{Effect of attending KIPP on test scores} is $E[Y|D=1] - E[Y|D=0]$
{Effect of winning lottery on test scores} is $E[Y|Z=1] - E[Y|Z=0]$
{Effect of winning on attending KIPP} is $E[D|Z=1] - E[D|Z=0]$

Then:

$$E[Y|D=1] - E[Y|D=0] = \frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]}$$

We observe everything on the right. **This is our IV estimator** for the relationship on the left.

- The numerator isn't biased because Z is random
- The denominator might not be biased, it depends on whether or not Z is as-good-as-randomly assigned to people who will take D if they win the lottery.
- Since both are unbiased (not associated with anything that affects (Y_{01}, Y_{1i})), then *we do not have to worry about selection bias*

The relationship in the previous page really relies on people who enroll in KIPP if they win the lottery

So, we add some assumptions:

- There are some people who are *compliers*: those who would enroll in KIPP if they win the lottery, but wont otherwise.
  - The estimator, above, estimates these people's treatment effect.
- There are **no defiers**
  - A *defier* is someone who enrolls in KIPP if they *lose* the lottery, and do not enroll if they win
  - Strange, right? Usually, we can safely make this assumption
- *Always-takers* and *Never-takers* are OK
  - These are people who always enroll in KIPP, regardless of winning the lottery
  - Never-takers are the opposite: they never enroll in KIPP even if they win.

LATE theorem:

- If there are *no* defiers, and if the three assumptions (relevant first stage, independence of Z, exclusion restriction), then the IV estimate:

$$\frac{E[Y|Z=1] - E[Y|Z=0]}{E[D|Z=1] - E[D|Z=0]}$$

  Is the $LATE$ - the **local average treatment effect**

- It is the ATE (hooray!) for the *compliers*, the people who can be affected by the instrument
- It has a causal interpretation.
- It can be used in treatment effects and program evaluation.

Next lecture:

We learned that we could estimate things like $E[Y|Z=1] - E[Y|Z=0]$ as β_1 in a regression like:

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

and $E[D|Z=1] - E[D|Z=0]$ as γ_1 in:

$$D = \gamma_0 + \gamma_1 Z + v$$

We use these to build our two-stage least-squares estimator (2SLS) for the IV.

More on Instruments

top

Recall last time we:

- Introduced the Instrumental Variable (IV)
  - When our variable of interest, D...
  - is correlated with something in the error and thus is biased.
  - The IV gives us independent (or *as-good-as-random*) variation in D.
  - Which lets us *identify* the effect of D on Y

- IV Requirements
  - 3 conditions for a valid IV

When we want an *unbiased* estimate of:

$$ATE = E[Y_1 - Y_0]$$

But all we have is observed data:

$$E[Y_1|D = 1] - E[Y_0|D = 0]$$

And we're worried about selection bias if we use our observed data because D is

- Possibly correlated with something in the error term
- The potential outcomes $(Y_{i0}, Y_{i1}) \not\perp D$
  - Both of these mean the same thing / are the same problem

We introduced the term *endogenous*

- "Endogenous" means "determined within the system"
- When treatment may be affected by something within our model, we say it may be "endogenous"
  - When $(Y_{i0}, Y_{i1}) \not\perp D$, then D is endogenous
  - D is determined in part by Y_{i0} or Y_{i1} - *within the system*

And the term *exogenous*

- "Exogenous" means "determined outside the system"
- Nothing in our regression helped determine D
  - Not Y_{i1} or Y_{i0} or anything in u

Conditionally exogenous is fine:

- D could be conditionally exogenous, conditional on some controls X

And note that:

Any x can be endogenous or exogenous. This isn't limited to D. Whether or not endogeneity is a problem depends on context.

Reminder:

$E[Y|D=1] - E[Y|D=0]$ is the same as β_1 in:

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

- D is our variable of interest, our *treatment*
- Y is our outcome

But we are afraid that something in u is correlated with getting treatment, D

- $E[u|D] \neq 0$
- D is endogenous
- If we know $E[u|D] = 0$, then our comparison of means is unbiased (with some exceptions)

Then, to get an unbiased estimate of β_1 we need an instrument, Z that:

- Determines or affects D
- Is "as good as (conditionally) randomly assigned" (is conditionally exogenous)
- But does not correlate with u (it only affects y through D - same thing)
  - When its uncorrelated, then $E[Y_0|Z=1] - E[Y_0|Z=0]$ is 0, and no selection bias
  - Except this just tells us about the potential outcomes over Z, not D, our variable of interest.

Recall our KIPP example:

We were interested in calculating the *{Effect of attending KIPP on test scores}*

This would be:

$$ATE = E[TestScore_{i1} - TestScore_{i0}]$$

We could calculate:

$$E[TestScore|KIPP == 1] - E[TestScore|KIPP == 0]$$

But, in our selection bias section, we learned that this is not going to get us the actual effect of KIPP because of self-selection into KIPP.

But we know that:

{Effect of winning lottery on test scores} =
{Effect of winning on attending KIPP} $\times$ **{Effect of attending KIPP on test scores}**

Written another way that may make more sense:

$$\underbrace{\frac{\text{Change in test scores}}{\text{Change in lottery win}}}_{\text{Observed and Unbiased}} = \underbrace{\frac{\text{Change in test scores}}{\text{Change in attendance}}}_{\text{Thing we're trying to measure without bias}} \times \underbrace{\frac{\text{Change in attendance}}{\text{Change in lottery win}}}_{\text{Observed and Unbiased}}$$

But what is that first term? It is:

$$E[TestScore|Lottery == 1] - E[TestScore|Lottery == 0]$$

We can take the sample analog of E and calculate the mean of $TestScore$ for those who won the lottery and those who didn't (ignoring attendance). This is the same as a regression:

$$TestScore = \phi_0 + \phi_1 Lottery + v$$

To clarify that point:

$$y = \phi_0 + \phi_1 Z + v$$

- $y = TestScore$
- $Z = Lottery$
- $D = KIPP$ (attendance)

Is just different notation for:

$$TestScore = \phi_0 + \phi_1 Lottery + v$$

And:

$$\begin{aligned}\beta_1 &= E[TestScore|Lottery == 1] - E[TestScore|Lottery == 0] \\ &= E[y|Z == 1] - E[y|Z == 0]\end{aligned}$$

When Z is as good as randomly assigned, which we'll get to shortly.

We have unbiased estimate of:

$$E[y|Z == 1] - E[y|Z == 0]$$

Which is the same as ϕ_1 in:

$$y = \phi_0 + \phi_1 Z + v$$

And we have **unbiased**

$$E[D|Z == 1] - E[D|Z == 0]$$

Which we can also get from γ_1 in a regression:

$$D = \gamma_0 + \gamma_1 Z + w$$

Because Z is *as good as randomly assigned* (more on that in a second)

If we take ϕ_1 as $\frac{\text{Change in test scores}}{\text{Change in lottery win}}$:

$$y = \phi_0 + \phi_1 Z + v$$

And if we take γ_1 as $\frac{\text{Change in attendance}}{\text{Change in lottery win}}$

$$D = \gamma_0 + \gamma_1 Z + w$$

Then we can calculate

$$\frac{\text{Change in test scores}}{\text{Change in attendance}} = \frac{\frac{\text{Change in test scores}}{\text{Change in lottery win}}}{\frac{\text{Change in attendance}}{\text{Change in lottery win}}}$$

as:

$$\beta_1^{IV} = \frac{\phi_1}{\gamma_1}$$

IV requires that Z is "as good as randomly assigned, conditional on x 's"

What does "as good as randomly assigned" mean?

- In our KIPP example, Z was randomly assigned because it was the result of a lottery.
- But can we have a Z that is not totally randomly assigned?

Yes, we can

- As long as we *control* for all the things that might not be random.
- That is, once we have statistical controls in our regression that explain part of how Z affects D, the rest of the variation in Z is uncorrelated with anything else in our regression.
- Example on next slide:

As good as random

Papers by the MM authors looked at the effect of military service on lifetime earnings. Since "people who may be inclined to join the army" may have some unobserved characteristics that might also affect military service (e.g. selection bias, $E[u|X] \neq 0$), the authors used people's draft numbers during Vietnam to *instrument* for likelihood of serving in the military.

Draft numbers were randomly assigned. A lower draft number meant you were more likely to be drafted. Some people with low draft numbers joined right away (you got to choose your assignment if you enlisted voluntarily); some always joined (always-takers!).

If drafts were done within age groups - that is, the draft board worked it's way up the numbers until each age group's quota was filled - then your age plus your number would also predict your probability of being drafted. So, your draft number was random *conditional on* your age.

SO...

$$Drafted = \gamma_0 + \gamma_1 LowDraftNumber + \gamma_2 AgeGroup + w$$

Conditional on age, draft number was as good as randomly assigned.

- D is $Drafted$
- Z is $LowDraftNumber$
- $AgeGroup$ is x's.

And "conditional on x's, Z is as good as randomly assigned"

Two Stage Least Squares (2SLS)

top

Define "Identification"

The term **identification** is used a lot in econometrics. With our β^{IV}, we would say we have *identified* β_1.

identification means we can write the parameter of interest, β, in terms of population moments.

- $\frac{\phi}{\gamma}$ is in terms of population moments because it is $\frac{\frac{Cov(Y,Z)}{Var(Z)}}{\frac{Cov(D,Z)}{Var(Z)}}$
- *identification* is a population-moments concept. It is a statement about the population parameter, β.
- Of course, if the population moment isn't identified, then our sample analog, $\frac{\hat{\phi}}{\hat{\gamma}}$ is useless.

On that last slide:

$$\beta_1^{IV} = \frac{\frac{Cov(Y,Z)}{Var(Z)}}{\frac{Cov(D,Z)}{Var(Z)}}$$

If we cancel things out...

$$\beta_1^{IV} = \frac{Cov(Y,Z)}{Cov(D,Z)}$$

Remember our first requirement: *relevant first stage*

If Z has no effect on D, what is $Cov(D,Z)$?

So we see where that comes from! We'll point out the other two IV requirements when we run across them

$\frac{\phi}{\gamma}$ is not useful when you have multiple x's

We could get $\hat{\beta}^{IV}$ from ϕ and γ, but if we have multiple x's, it's difficult. So we have a different *method* of estimating β^{IV}:

Two Stage Least Squares (2SLS)

2SLS, the first stage

First Stage: regress D on Z, X:

$$D = \gamma_0 + \gamma_1 Z + \gamma_2 x_1 + \cdots + w$$

Since w is uncorrelated with things correlated with Z, γ_1 is unbiased.

Next, generate $\hat{D}$:

Take $\hat{\phi}$, your estimates and get the predicted value of D, $\hat{D}$

$$\hat{D} = \hat{\gamma}_0 + \hat{\gamma}_1 Z + \hat{\gamma}_2 x_1 + \cdots$$

This has a "partialling out" flavor

- w contains all the variation in D that is not explained by Z.
- And likewise $\hat{D}$ contains all the variation in D that **is** explained by Z (w is not in $\hat{D}$)

Remember

- IMPORTANT: our problem in the first place was that D was correlated with something unobserved in the error.
- But $\hat{D}$ is entirely from Z and other exogenous statistical controls, and Z is not correlated with this unobserved problem by *the exclusion restriction*

2SLS, the second stage

For our second stage:

$$Y = \beta_0 + \beta_1 \hat{D} + \beta_2 x_2 + \cdots + u$$

And the estimate of β_1 here is our IV estimate: $\hat{\beta}_1^{IV}$. That is our coefficient of interest!

β_1 is unbiased!

We estimated β_1 using *only* variation in D associated with Z, and Z is (assumed to be) uncorrelated with u, solving our initial problem.

What if we have multiple endogenous variables?

Wooldridge uses y_1 for the outcome, and $y_2, \cdots$ on the right hand side for an endogenous variable that we need to instrument. Wooldridge uses z for all exogenous variables, instruments or not. I prefer to use x since that leaves z for instruments only

In the regression:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3 + \beta_3 x_1 + \beta_4 x_2 + u$$

The problems are y_2 and y_3. Since there are two, **we need two valid instruments**, z_1 and z_2.

In our first stage, we also include the x's from the main regression:

$$y_2 = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 x_1 + \gamma_4 x_2 + v$$

Gives us $\hat{y}_2$, while:

$$y_3 = \kappa_0 + \kappa_1 z_1 + \kappa_2 z_2 + \kappa_3 x_1 + \kappa_4 x_2 + w$$

Which gives us $\hat{y}_3$. Finally, we get a second stage β^{IV}:

$$y_1 = \beta_0 + \beta_1 \hat{y}_2 + \beta_2 \hat{y}_3 + \beta_3 x_1 + \beta_3 x_2 + u$$

Those are the same x's in each regression, no hats.

A couple notes:

- The instruments never appear in the second stage
  - Because $\hat{y}_2$ and $\hat{y}_3$ are perfectly colinear with Z
- The exogenous statistical controls, x_1, x_2 appear in both stages.
  - In fact, if they are used in the first, then left out of the second, the result can be biased.

We have to have at least one exogenous instrument z for every endogenous variable.

When we have 2 endogenous variables, we need 2 instruments with two first-stage regressions that both meet the *relevant first stage* requirement.

We can have more than one instrument per endogenous variable!

We say we are *overidentified* when this is the case. We can test this (Hausman Test - W15.5)

Overidentification is not necessarily a bad thing, despite the name.

Testing IV assumptions

top

First, Z must have a causal effect on D

The instrument, Z, has to have a casual effect on the variable of interest, D.

- In our example, this means Z has to *change enrollment* (in expectation)
- This is the *relevant first stage* requirement or condition

Testing this requirement

- We need a test that tells us if all the variables in our model of $D = \gamma_0 + \gamma_1 Z + w$ are any better at predicting D than just guessing $\gamma_0 = \bar{D}$.
- Sound familiar?

Relevant First Stage test

- In a single variable regression (one instrument), then we can just test if $\gamma_1 = 0$.
- If we have two instruments or more, then we use the F test for:
  - $D = \gamma_0 + \gamma_1 Z_1 + \gamma_2 Z_2 + w$.
- Rule of thumb: F-stat needs to be > 10 (Stock and Yogo)

Second, Z must be as good as randomly assigned

The instrument, Z, cannot be determined by the omitted variable / selection bias we're trying to get out.

- This is the *independence requirement*

This one, we can't test for.

This is for the same reason that we can't just test for $E[u|X] = 0$, there might still be something out there that is correlated with Z and is also correlted with D, which means Z is not randomly assigned.

Instead, we have to make a compelling argument for this to be true.

Third, Z must affect the outcome only through the variable of interest

The instrument, Z, cannot have a direct affect on the outcome

- This is the *exclusion restriction*.

This one, we can't directly test for either

We have to make a compelling argument for this to be true.

If we have multiple instruments, we can sort of test this, though...

If we have many instruments available (more than we have endogenous variables), we can choose which one is best

We do this by comparing the β^{IV} estimated with one instrument, then the other. They *should* give the same result if they both meet the requirements, right? If they're different (statistically speaking), then we have a problem. One (or both) doesn't meet the *exclusion restriction*.

Don't assume failing-to-reject this test means the instruments meet the *exclusion restriction*.

The overidentification (Hausman) Test

First, take the $\hat{u}$ from your second stage

Remember, this is $y - \hat{y}$ where $\hat{y}$ is from the 2nd stage.

If all of your instruments meet the *exclusion restriction*, then they should not be correlated with this residual, $\hat{u}$.

$$\hat{u} = \eta_0 + \eta_1 z_1 + \eta_2 z_2 + \eta_3 x_1 + \cdots + \varepsilon$$

The R^2 tells us whether or not all of the (hopefully exogenous) z's and x's on the RHS are unrelated to $\hat{u}$.

$nR^2 \sim \chi^2_q$ where q is the number of instruments minus the number of endogenous variables.

What about R^2 of the second stage?

$$y = \beta_0 + \beta_1 \hat{D} + \beta_2 x_1 + \cdots + u$$

As Wooldridge states, R^2 from IV is not very useful; we aren't trying to explain more variation, we're trying to use a specific subset of the variation to get at a causal estimate.

We ignore it.

The whole point of this was to do inference on an *unbiased* estimate of $\hat{\beta}$, which is $\hat{\beta}^{IV}$.

But we used two stages, so what is the $se(\hat{\beta^{IV}})$? We can't use just the 2nd stage, right?

- (No, we can't)

First, we assume:

Something-like-MLR5: $Var(u|z) = \sigma^2$

- Same as before, but z instead of x
- Still homoskedasticity
- Still have a robust version

For one endogenous x, and one instrument z

$$se(\hat{\beta}_1^{IV}) = \frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2}$$

- $n\sigma_x^2$ looks a lot like SST_x since $\hat{\sigma}_x^2 = \frac{1}{N-1}\sum(x_i - \bar{x})^2$

What is $\rho_{x,z}$?

It is the *correlation coefficient* of x and z:

$$\rho_{x,z} = \frac{Cov(X,Z)}{\sqrt{Var(X)}\sqrt{Var(Z)}}$$

Another way of writing $se(\hat{\beta}_1^{IV})$ is:

$$se(\hat{\beta}_1^{IV}) = \frac{\hat{\sigma^2}}{SST_x R^2_{x,z}}$$

Where $R^2_{x,z}$ is the R^2 in the **first-stage** regression.

- If that $R^2_{x,z}$ is exactly 1, then this is the same as $\hat{\beta}^{OLS}$.
  - What would it mean if $R^2_{x,z} = 1$?
- If $R^2_{x,z}$ is small, then what happens?

IV in R

top

The AER package in R has an excellent function

```
install.packages('AER')
```

```
ivreg(Y ~ D + X1 | X1 + Z1 + Z2, data=df)
```

- The first part of the forumla is just like an OLS regression
- Here, I'm following lecture notation:
  - Y is the outcome of interest.
  - $D$is the endogenous variables of interest
  - $X1$ is an exogenous statistical control
  - $Z1$ and $Z2$ are the instruments

Note that this has more than one instrument (Z1, Z2)

Note that X1 *instruments for itself*. R will instrument everything to the left of the "|" with everything to the right of the "|"

In our KIPP example:

```
IVmodel = ivreg(TestScores ~ KIPP + year | year + Lottery, data=df)
```

I've added *year* as an exogenous variable. Time is always exogenous. All instruments will be used to instrument $KIPP$ and $year$.

We can still use our robust standard errors

```
coeftest(IVmodel, vcov = vcovHC(IVmodel, 'HC1')
```

Simultaneous Equations

top

We think of the variables in our data as being either "endogenous" or "exogenous"

This tells us whether or not we should be worried about correlation with u.

Exogenous

Exogenous means "determined outside the system".

- Things like *rainfall* in ag production and *winning the KIPP lottery* are *exogenous*
  - There is usually nothing *inside* the system that helps determine them.
  - Although...we could think of times that even rainfall is endogenous.
  - What about a model that includes the selection of land for farming?

Simultaneity

Simultaneity occurs when the *dependent variable*, the y, the left-hand-side, is determined jointly with one or more right-hand-side variables.

- Of course, it's always the case that the dependent variable y is *determined* by one or more right hand side explanatory variables.
- $y = \beta_0 + \beta_1 x_1 + u$ shows this.
- But *simultaneity* is unique in that x_1 itself is *jointly determined* with y.

An example of a county-level labor supply function

$$h_s = \alpha_1 w + \beta_1 z_1 + u_1$$

- h_s is the hours supplied each week by workers in the county
- w is the wage
- z_1 is anything that affects hours supplied
- u_1 is the error term for hours supplied

This equation stands on its own

- It has a causal interpretation (if α_1 can be estimated without bias)
- It is derived from economic theory (higher wages cause people to substitute out of leisure and into labor)

$\Rightarrow$ **So we call this a structural equation**

$$h_s = \alpha_1 w + \beta_1 z_1 + u_1$$

It suffers from simultaneity because:

- A county's w will be determined, in part, by h_s, the supply.
- Wage is determined jointly by the interaction of h_s, w, and h_d, the hours demanded.
- Thus, **simultaneity**.

The "link" between h_s and h_d is the equilibrium

- $h_s = h_d = h$. Since this happens in every county, we use h_i.
- We only observe this equilibrium, but we might want to know about the values of α_1 and α_2

So we can take our two equations:

$$h_s = \alpha_1 w + \beta_1 z_1 + u_1$$

$$h_d = \alpha_2 w + \beta_2 z_2 + u_2$$

And impose the equilibrium condition: for every i, $h_s = h_d = h_i$

$$h_i = \alpha_1 w_i + \beta_1 z_{1i} + u_{1i}$$

$$h_i = \alpha_2 w_i + \beta_2 z_{2i} + u_{2i}$$

In this simultaneous system of equations:

$$h_i = \alpha_1 w_i + \beta_1 z_{1i} + u_{1i}$$

$$h_i = \alpha_2 w_i + \beta_2 z_{2i} + u_{2i}$$

h_i and w_i are the endogenous variables. Why?

Because, given $z_{1i}, z_{2i}, u_{1i}, u_{2i}$, then h_i and w_i are completely determined

- with a few assumptions about α_1 and α_2

The dependent variable and one or more explanatory variables are jointly determined within the system.

This happens often in economics

We have many parties interacting with each other, and equilibriums are the outcomes of those interactions.

Think of *marginal analysis* - how we think of a seller setting a price in a market. It's a lot of expectations about interactions.

Back to the simultaneous system of equations:

$$h_i = \alpha_1 w_i + \beta_1 z_{1i} + u_{1i}$$

$$h_i = \alpha_2 w_i + \beta_2 z_{2i} + u_{2i}$$

Note that the z_{1i} and z_{2i} are different variables, while w_i is the same in both equations.

- u_{1i} and u_{2i} are different as well. And uncorrelated with each other.
- We refer to the u_{1i} and u_{21} as the *structural errors*.

Example W 16.1

$$murdpc = \alpha_1 polpc + \beta_{10} + \beta_{11} incpc + u_1$$

$$polpc = \alpha_2 murdpc + \beta_{20} + other$$

- $murdpc$ is murders per capita
- $incpc$ is income per capita, which shifts murder rates
- β_{10} is the intercept for equation 1
- $polpc$ is police per capita
- β_{20} is the intercept for equation 2

Is this simultaneous?

- Yes. Just as hours supplied, hours demanded, and wage are jointly determined, $murdpc$ and $polpc$ are jointly determined.
- The city chooses $polpc$ based, in part, on $murdpc$, while murderers choose $murdpc$ based, in part, on $polpc$.
- Even though we're interested in α_1, we need to understand the second equation to avoid bias.

Simultaneity Bias

top

Simultaneity bias

We can formally show the bias in simultaneous equations. Remember, bias occurs when an explanatory variable is correlated with u (and thus $E[u|x] \neq 0$)

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

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

y_1, y_2 could be $murdpc$ and $polpc$ from the previous section.

But estimating either equation by OLS would result in a biased α. So we can't do it.

- Specifically, we are in trouble on the first if y_2 is correlated with u_1;
- And if y_1 is correlated with u_2 for the second.
- Let's see why this is true...

To see bias, substitute the first equation into the second

$$y_2 = \alpha_2 \underbrace{(\alpha_1 y_2 + \beta_1 z_1 + u_1)}_{y_1} + \beta_2 z_2 + u_2$$

$$(1 - \alpha_2 \alpha_1) y_2 = \beta_2 z_2 + \alpha_2 \beta_1 z_1 + \underbrace{\alpha_2 u_1}_{uh-oh} + u_2$$

This shows that y_2 is correlated with u_1. Estimating the y_2 equation from the previous slide gives a biased α_2. This is simultaneity bias.

Identification in Simultaneous Equations

We started with:

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

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

We plugged in y_1 as a function of y_2, z_1, u_1 to estimate α_2

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

and can do the same for y_2

$$(1 - \alpha_2\alpha_1)y_1 = \beta_1 z_1 + \alpha_1\beta_2 z_2 + \alpha_1 u_2 + u_1$$

In the y_2 equation, Divide both sides by $(1 - \alpha_2\alpha_1)$:

$$y_2 = \frac{\beta_2}{(1-\alpha_2\alpha_1)}z_2 + \frac{\alpha_2\beta_1}{(1-\alpha_2\alpha_1)}z_1 + \underbrace{\frac{\alpha_2}{(1-\alpha_2\alpha_1)}u_1 + \frac{1}{(1-\alpha_2\alpha_1)}u_2}_{v_2}$$

Which gives us:

$$y_2 = \pi_{21}z_1 + \pi_{22}z_2 + v_2$$

This is called the *reduced form* equation for y_2

- We *can* estimate π_{21} and π_{22}, but the coefficients lose their structural interpretation.
- Our estimation of π_{21} and π_{22} is unbiased - z's are exogenous.
- π_{21} and π_{22} are functions of the structural parameters $\{\alpha_1, \alpha_2, \beta_1, \beta_2\}$

When can we identify α_1 and α_2?

- Our problem here is endogeneity, so we need an instrument.
- Something that shifts y_2 but is not correlated with u_1 (exclusion restriction)
- Do we have something in y_2?

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

Yes. We have z_2, which is exogenous by definition.

It can shift y_2, and is not correlated with u_1. It does not shift y_1 except through y_2 because it is not in the equation for y_1.

Similarly, we can use z_1 to shift y_1.

And both equations can be identified because we have *one exogenous shifter for each endogenous variable in each equation*.

The Rank Condition

In a two-equation system, we can only identify an equation with an endogenous variable if the *other* equation has one or more exogenous variable that does not enter the first equation.

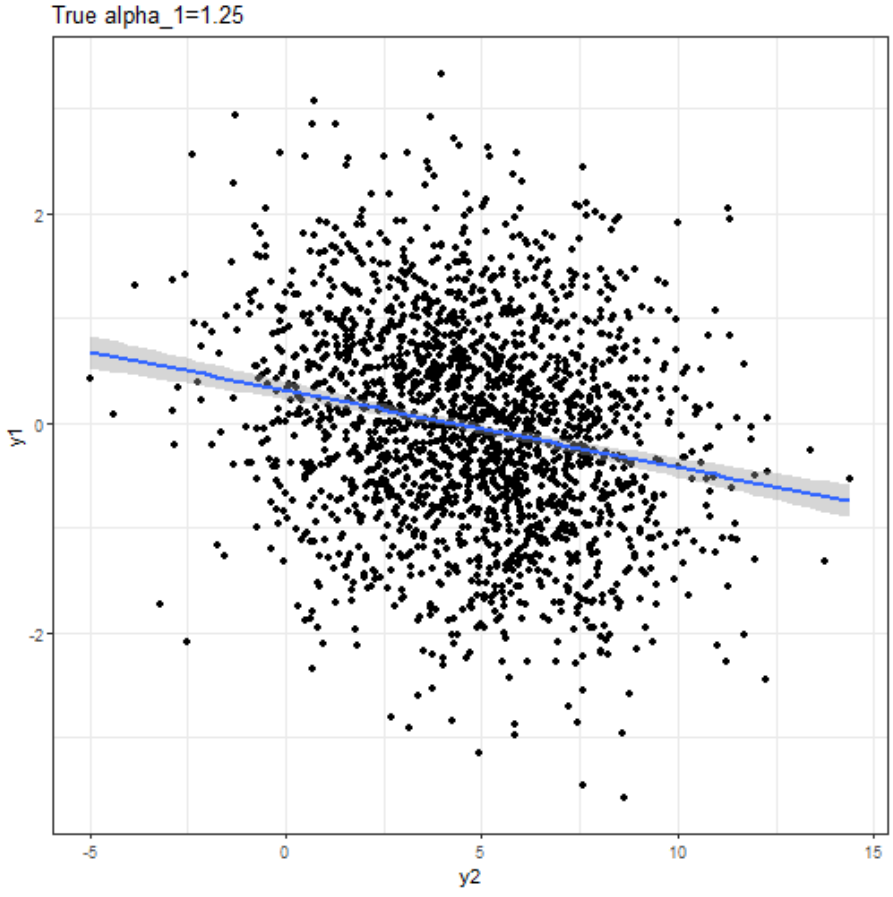
- The instrument must have a non-zero population coefficient

Our two equation system, again (with the problems in red and blue:

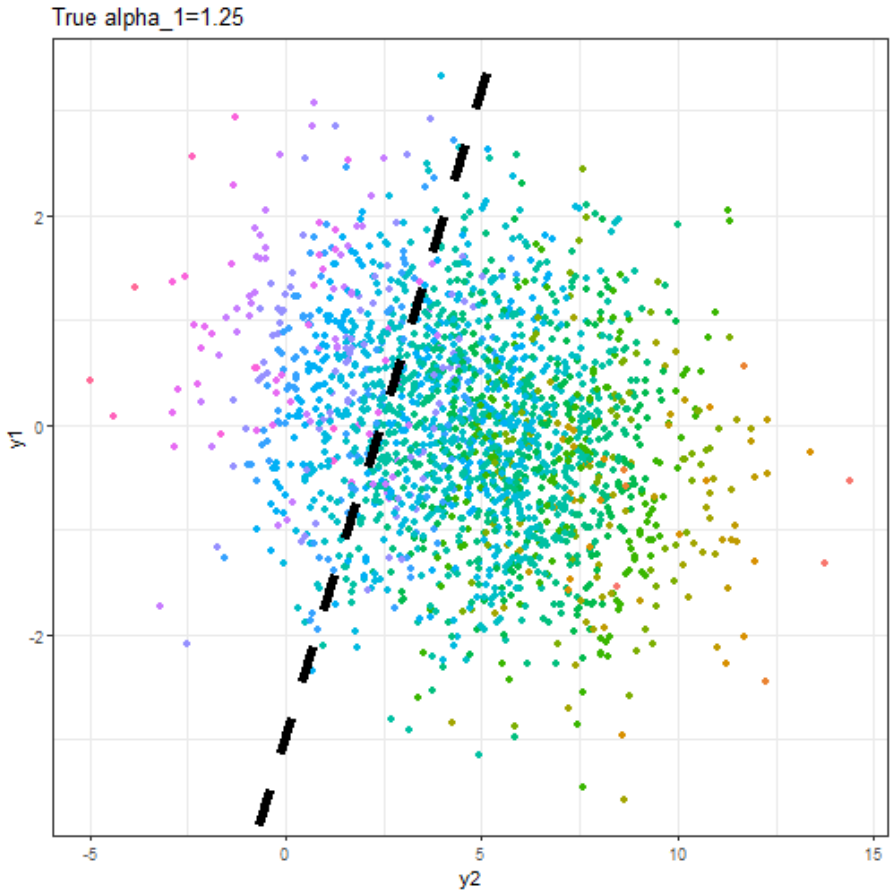
$$y_1 = \alpha_1 y_2 + \beta_1 z_1 + u_1$$

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

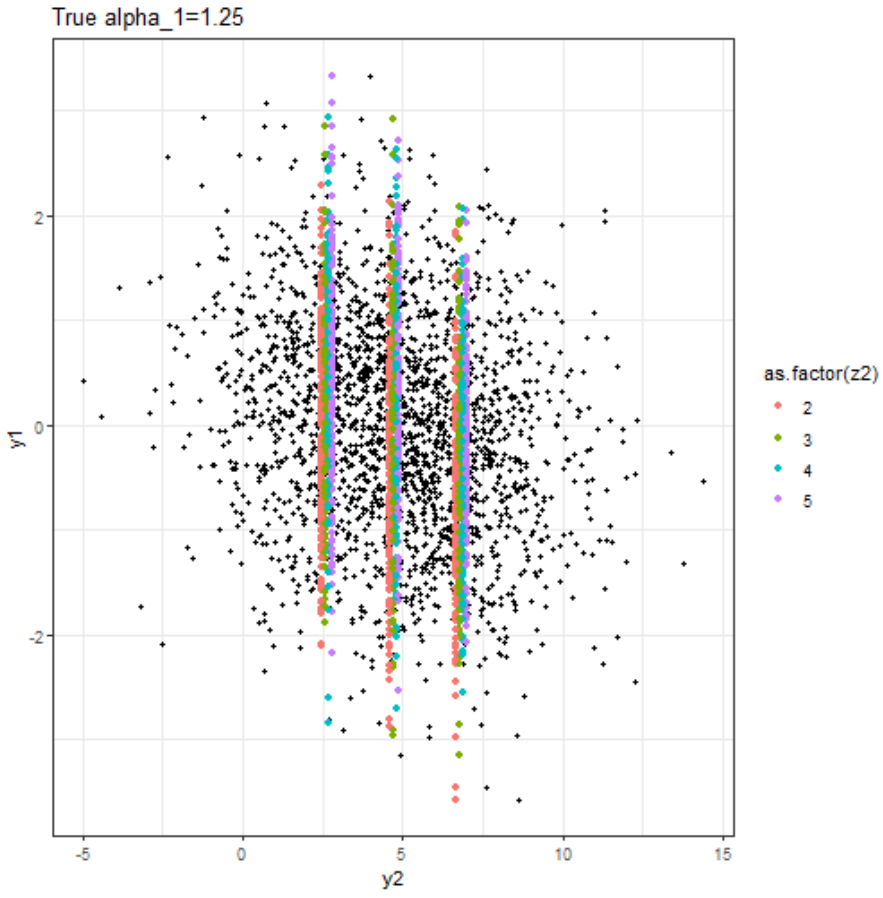
A visual example:



Since the problem is that y_2 is correlated with u_1, what if we observed u_1?



Of course, we can't control for u_1 since it is unobserved.



The colored groupings are $\hat{y}_2$. Each grouping is a different z_1. As z_2 increases, y_2 increases within each grouping. As y_2 increases, in each grouping, y_1 *increases*.

First stage

$$y_2 = \pi_{21} z_1 + \pi_{22} z_2 + v_2$$

```
##
## Call:
## lm(formula = y2 ~ z1 + z2, data = df1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6716 -1.4648 -0.0406  1.4813  7.6990
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.15937    0.20339   0.784   0.4334
## z1           2.10246    0.06017  34.941   <2e-16 ***
## z2           0.10592    0.04483   2.363   0.0182 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.201 on 1997 degrees of freedom
## Multiple R-squared:  0.381,    Adjusted R-squared:  0.3803
## F-statistic: 614.5 on 2 and 1997 DF,  p-value: < 2.2e-16
```

Second stage

$$y_1 = \alpha_1 \hat{y}_2 + \beta_1 z_1 + u$$

```
##
## Call:
## lm(formula = y1 ~ y2hat + z1, data = df1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2826 -0.6026 -0.0129  0.6212  3.0643
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.3714     0.1107  -3.354 0.000811 ***
## y2hat         1.9901     0.1811  10.991  < 2e-16 ***
## z1           -4.5455     0.3819 -11.902  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9415 on 1997 degrees of freedom
## Multiple R-squared:  0.1358,    Adjusted R-squared:  0.1349
## F-statistic: 156.9 on 2 and 1997 DF,  p-value: < 2.2e-16
```

The second-stage coefficient

We get a pretty accurate estimate for $\alpha_1 = 1.99$ from the second-stage having used z_2 to instrument for y_2.

In a panel data setting we'd have a *fixed effect* for each i:

$$y_{it1} = \alpha_1 y_{it2} + \mathbf{z_{it1}}\beta_\mathbf{1} + a_{i1} + u_{it1}$$

$$y_{it2} = \alpha_2 y_{it1} + \mathbf{z_{it2}}\beta_\mathbf{2} + a_{i2} + u_{it2}$$

a_{i1} is unobserved and potentially correlated with z_{it1}. This presents interesting problems unique to panels.

First Differencing

One way of handling an unobserved fixed effect in panel data (different from what we've learned on fixed effects) is *first differencing*.

$$y_{it1} - y_{i(t-1)1} = \alpha_1(y_{it2} - y_{i(t-1)2}) + \beta_1(z_{it1} - z_{i(t-1)1}) + a_{i1} - a_{i1} + u_{it1} - u_{i(t-$$

Which can be written using the Δ notation:

$$\Delta y_{it1} = \alpha_1 \Delta y_{it2} + \beta_1 \Delta z_{it1} + \Delta u_{it1}$$

This removes the a_{i1}, and makes it clear that we need an instrument whose *change* is

- Exogenous
- Affects only Δy_{it2} without affecting Δy_{it1} (uncorrelated with Δu_{it1}).
- And *it has to vary within each* i

Using fixed effects

A similar result happens if we include the fixed effect. The fixed effect instruments for itself, and is included as an exogenous variable.

First stage

$$y_{it2} = \pi_{21} z_{i1} + \pi_{22} z_{i2} + \gamma_i^1 + v_{i2}$$

Second stage

$$y_{it1} = \alpha_1 \hat{y}_{it2} + \beta_1 z_{i1} + \gamma_i^2 + u_{i1}$$

γ_i^1 is the fixed effect for each i in the first stage.

γ_i^2 is the fixed effect for each i in the second stage (not a squared term).

Next up

We turn to other methods of getting at causal estimates.

- Difference in Differences
- Regression Discontinuity