Instrumental Variables

EC420 MSU

Justin Kirkpatrick
Last updated March 16, 2021

What's behind us:

- Selection bias
  - A specific type of bias
  - Showed that $E[Y|D=1] - E[Y|D=0] = SATE$ + Selection Bias
- Randomization
  - Randomizing treatment $\Rightarrow$ no selection bias
  - $E[Y_0|D=1] = E[Y_0|D=0] = E[Y_0]$ under randomization
- D is our treatment.
  - The β on D is our parameter of interest

Where were at:

- Interpreting regressions
- Assumptions ("What's in the error")
- Applications: Program evaluation:
  - Statistical controls
  - Valid vs. Causal
- Causal and unbiased estimates of β using **Instrumental Variables**
- Continuous x continuous interactions at end
  - Circling back to something not covered before midterm

SRF vs. PRF

While we're always interested in the PRF:

$$E[Y|X] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 1(condition) + \beta_4 1(condition) x_1$$

We don't get to observe β. We are stuck with $\hat{\beta}$ and the SRF:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 1(condition) + \hat{\beta}_4 1(condition) x_1$$

And we have $y - \hat{y} = \hat{u}$

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

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

Where we're at

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

Let's say:

- x_1 is continuous,
- D is binary $\{0, 1\}$

Let's calculate some $E[y]$ (that is, $\hat{y}$):

- $E[y|x_1 = 0, D = 1]$:
- $E[y|x_1 = 2, D = 1]$:
- $E[y|x_1 = 2, D = 0]$:

Causal Interpretation and Program Evaluation

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"
  - $minwage_{it}$

Causal interpretation

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

Instrumental Variables

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

There are three parts to this.

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.

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 class:

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:

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

Next class, we use these to build our two-stage least-squares estimator (2SLS) for the IV.

Continuous interactions: a brief detour

We didn't get to cover continuous $\times$ continuous interactions so we will do that now:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \underbrace{\beta_3 x_1 x_2}_{\text{Interaction}} + u$$

- β_3 is the interaction of continuous x_1 and continuous x_2
- We know the interpretation if x_1 or x_2 is binary or categorical:
  - " β_3 shifts slope."
- What does it mean here?

Marginal effect:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \underbrace{\beta_3 x_1 x_2}_{\text{Interaction}} + u$$

- Now, it's still a slope-shifter, but it's continuous:
  - Instead of saying "the relationship (slope) between x_1 and y is β_3 more when $x_2 == 1$, *ceteris paribus*"
  - We can say "the relationship between x_1 and y is β_3 more when x_2 increases by one unit, *ceteris paribus*"
  - **And** the same for x_2 and y.

Marginal effect:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \underbrace{\beta_3 x_1 x_2}_{\text{Interaction}} + u$$

The first derivative with respect to x_1 gives us the **marginal effect** of x_1:

$$\frac{\partial y}{\partial x_1} = \beta_1 + \beta_3 x_2$$

- The left side is "the change in y per increase in x_1". It is the **slope**
- The change in y per increase in x_1 is larger when $\beta_3 x_2$ is positive!
- The same can be done with x_2:
  - $\frac{\partial y}{\partial x_2} = \beta_2 + \beta_3 x_1$

Relationship between y and x_1 changes with x_2

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

$$\frac{\partial y}{\partial x_1} = \beta_1 + \beta_3 x_2$$

As x_2 increases, the *slope* also increases $\rightarrow$

β_1	β_3	x_2	$\frac{\partial y}{\partial x_1}$
1	3	-1	-2
1	3	0	1
1	3	1	4
1	3	2	7

In R

- Use `lm(y ~ x1 + x2 + x1:x2, df)`
  - The `:` tells R to do the interaction between `x1` and `x2`, `x1 x x2`
  - If you instead use `x1*x2` you will be telling R to regress `x1, x2, x1:x2`
  - It's just a little easier to use `x1*x2` since you'll almost always need the "main" effects of `x1, x2`.

Generally, we're most interested in the sign of the coefficient (positive or negative), and if it is significant.

Continuous interactions

```
##
## Call:
## lm(formula = wage ~ educ + exper + educ:exper, data = wooldridge::wage2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -948.46 -254.53  -29.57  192.59 2150.77
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  271.934    230.227   1.181  0.23784
## educ          35.106     16.626   2.112  0.03499 *
## exper        -32.663     19.099  -1.710  0.08757 .
## educ:exper     3.904      1.462   2.670  0.00771 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 375.1 on 931 degrees of freedom
## Multiple R-squared:  0.1424,    Adjusted R-squared:  0.1397
## F-statistic: 51.54 on 3 and 931 DF,  p-value: < 2.2e-16
```

Anyone worried that $exper$ is negative?

Continuous interactions

Remember to include the interaction term. On the previous slide, if you have 12 years of education, then the marginal effect of a 1-year increase in experience is

$$\beta_{exper} + \beta_{educ:exper} \times x_{educ}$$

Using the results from the previous slide:

$$-32.663 + 12 * 3.904 = 46.84 = 16.177$$

The *marginal effect* of increasing experience (ceteris paribus) depends on the level of education

- People who have more education have larger increases in wage per 1 year of experience
- People with more education have faster wage growth!

```
##
## Call:
## lm(formula = fare ~ dist * year, data = wooldridge::airfare)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -125.92  -46.99  -15.39   40.46  228.18
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.056e+04  2.926e+03  -3.609 0.000311 ***
## dist         1.691e+00  2.514e+00   0.673 0.501163
## year         5.334e+00  1.464e+00   3.644 0.000271 ***
## dist:year   -8.081e-04  1.258e-03  -0.642 0.520671
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.34 on 4592 degrees of freedom
## Multiple R-squared:  0.3935,    Adjusted R-squared:  0.3931
## F-statistic:   993 on 3 and 4592 DF,  p-value: < 2.2e-16
```