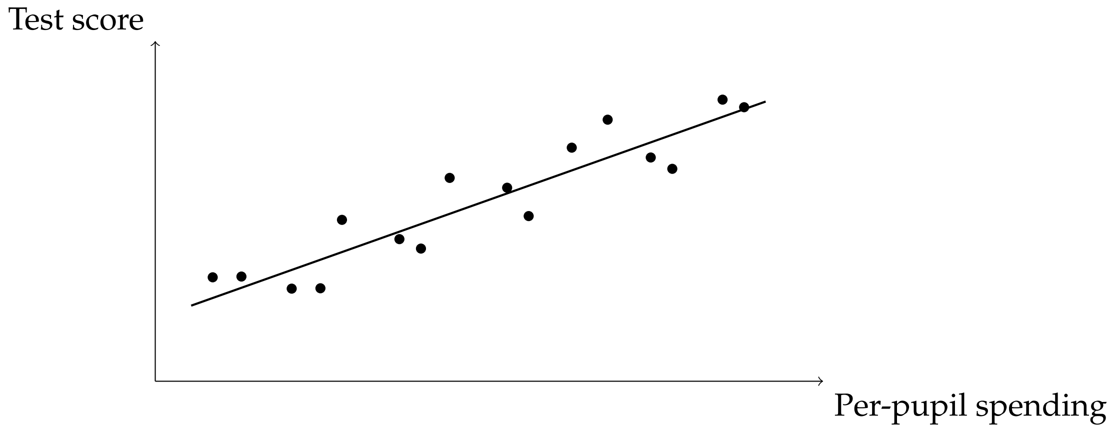
Lecture 6 – Regression and Instrumental Variables

We said that randomization gets rid of selection bias. We almost never get to randomize, so what do we do instead?

What is Regression Good For?

Regression draws the best line through a cloud of points. Almost every method in this unit uses it.



Each dot is a district. The line is the regression.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Y_i is the test score in district i, X_i is per-pupil spending, and ε_i is everything else that moves test scores. The coefficient we care about is β_1.

In this class the two most important things will be 1) understanding what an estimating equation is saying in words and 2) proper interpretation of coefficients.

Interpretation

Suppose we run this and get $\beta_1 = 3$, with spending measured in thousands of dollars. What does that tell us?

It tells us that comparing two districts that differ by $1,000 in per-pupil spending, the higher spending district scores 3 points higher.

What doesn't it tell us? It does not tell us that spending caused the difference, and it does not tell us that giving a district another $1,000 would raise its scores by 3 points. Regression tells us how two things move together. It does not tell us why.

The binary case

Often X is just an indicator for whether a student got the program. In that case β_1 is the difference in average outcomes between the two groups, which is exactly what we estimated in the RCT last week.

Controls

If you are worried that high spending districts are also higher income, you can add income to the regression.

$$Y_i = \beta_0 + \beta_1 Spending_i + \beta_2 Income_i + \varepsilon_i$$

Now β_1 compares districts that have the same income. That is all a control does – it narrows the comparison. I can't discuss the reasons in detail but if you're interested the relevant theorem is the Frisch-Waugh-Lovell.

Multiple regression changes the interpretation. For

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

A 1 unit increase in X_1 is associated with a β increase in Y while holding X_2 fixed

Omitted Variable Bias

What happens when we leave something out of our regression? It could bias our estimates of the true effect, or it might not. It depends on two things:

1. Does the omitted variable affect the outcome?
2. Is the omitted variable correlated with our treatment?

If either answer is no, leaving it out is fine. If both are yes, the estimate is biased and we can sign the bias.

Let's say you're interested in the relationship between grant aid and college graduation, and you leave out family income. Does income affect graduation? (+, richer students graduate at higher rates) Is income correlated with aid? ($-$, aid is need-based)

Positive times negative is negative, so leaving income out pushes the aid coefficient down. This is how you end up with the finding that more grant aid goes along with lower graduation.

This comes up all the time. Let's think about some examples

Y=selective college enrollment, included X=private school attendance, omitted=family income

Y=student learning, included X=teacher salary, omitted=PTA budget

Instrumental Variables

What is IV good for?

We have a variable we want to know the effect of, but it isn't random. IV allows us to get that effect if a few conditions are met.

What are the requirements for a good instrument?

1. *First Stage* The instrument is strongly correlated with treatment
2. *Independence* The instrument is as good as randomly assigned
3. *Exclusion Restriction* The instrument only affects the outcome through treatment

Only the first one is testable. The other two are arguments you make.

Who are we learning about?

Always Taker This student attends regardless of the lottery. The 4%.

Never Taker This student never attends regardless of the lottery. The 22% who win and don't enroll.

Complier This student attends only if they win. The 74%.

Defier This student attends only if they lose.

We usually assume there are no defiers, which is called *monotonicity*.

The lottery tells us nothing about always takers or never takers, since they did the same thing either way. So our estimate is the average effect among compliers, which is the *Local Average Treatment Effect (LATE)*.

Is the LATE useful? Is it about who we want to know about? Yes and no. It is the policy relevant population, but we may be interested in a broader set of students. Ask yourself every time who the compliers are.

Imperfect compliance

A charter school has more applicants than seats, so admission is decided by lottery. At KIPP, about 78% of students who are offered a seat attend, and about 4% of students who are not offered a seat attend anyway.

Is the comparison of students who attend to students who don't attend a valid one? (No) Why not? (Enrolling after winning the lottery is a choice, so we are back to selection bias)

Is the comparison of winners to losers valid? (Yes) What does it reveal? (The effect of being offered a seat) This is the *Reduced Form*, also called the *Intent to Treat (ITT)*.

Is that what we want to know? (Probably not, it is more interesting to know the effect of attending)

Scaling up

The lottery changed the probability of attending by $.78 - .04 = .74$. That is the first stage.

Suppose winning the lottery raises math scores by 7 points. That is the reduced form.

The effect of attending is then $\frac{7}{.74} \approx 9.5$ points. The lottery only worked through the 74% of students whose attendance it actually changed, so the effect on those students has to be larger than the effect on everyone.

Assessing an Instrument

Suppose we want the effect of attending college on earnings. Students choose whether to enroll, so a simple comparison has selection bias. Consider using distance from your high school to the nearest college as an instrument.

> *First Stage* – Does living closer to a college raise the chance you enroll? Which students does it move?
>
> *Independence* – Is it random who lives near a college? Did families choose to live there?
>
> *Exclusion Restriction* – Can distance affect earnings in a way other than through college attendance? Are rural labor markets the same as urban ones?
>
> *Monotonicity* – Would living closer ever make a student less likely to enroll?

Which of these worries you most? What could a researcher show you to make you more comfortable?

Wrapping Up

> A coefficient is a comparison. Practice saying it in words, including what it does not say.
>
> For omitted variable bias, sign the two relationships and multiply.
>
> Controls only fix what you can measure.
>
> IV requires a first stage, independence, and an exclusion restriction. Only the first is testable.

IV gives you the effect on compliers, not on everyone.