Fall 2022
SS 340
Loujaina Abdelwahed

You may work with other students, but the write-ups must be unique. Please save your answers in .docx format and submit it on MS Teams

Part 1

In the first part, we will use *simulated data* to see how an instrumental variable z can solve the problem of a correlation between error term ε and an explanatory variable x.

I created the simulated data for you, but feel free to create your own if you want/ have time --- just for practice. The simulated data is attached in the data file called "Simulated.dta"

Here are the exact steps I followed:

- I drew 10,000 observations for the explanatory variable $x1$ and the error $\varepsilon 1$ from a (2-dimensional) multivariate normal with mean vector [10, 0] and the identity matrix as the covariance matrix.
- I called my variables here X1 and E1.
- I then created a dependent variable called Y1, that is equal to $2+3x_1+\varepsilon$
- So now, you KNOW for sure the true values of the coefficients $\beta 0$ and $\beta 1$ in a regression $y=\beta 0+\beta 1x+\varepsilon$

What you have to do:

a. Make a scatter plot of the data (X1 and Y1) as well as a regression line
b. To see if OLS works well for our simulated data, fit an OLS regression of y_1 on x_1. Show me the coefficients you estimate. Compare the estimated coefficients to the true values of $\beta 0$,$\beta 1$

Now, I will generate new variables, X2 and E2. The only thing that I really change here is the covariance matrix, whereby I now make the covariance between X2 and E2 = 0.8. Now X2 and E2 are correlated --- In other words, I am intentionally violating the conditional independence assumption.

I also create the dependent variable called Y2, where $y_2=2+3x_2+\varepsilon$

c. Fit an OLS regression of y_2 on x_2. Show me the coefficients you estimate. Compare the estimated coefficients to the true values of $\beta 0$,$\beta 1$

As you now hopefully will have noticed, OLS does not yield consistent estimates for the true relationship anymore. However, an instrumental variable will help us to receive consistent estimates.

Fall 2022
SS 340
Loujaina Abdelwahed

I created an instrument z , where cov(x,z)=0.3 and cov(z,ε)=0.

d. Now, estimate the equation with 2SLS "by hand" using the following steps:
   - Regress the endogenous variable X2 on the instrument Z and make a prediction. Save the predicted values in a variable called x2_hat. This is the first stage.
   - Regress Y2 on x2_hat. This is the second stage. Compare the results to the OLS results from before. Are we getting closer to the true estimate?

e. Use ivregress in STATA or statsmodels IV2SLS function in Python to estimate the model. Compare the results to what you estimated by hand.

Part 2

In this part of the assignment, I created the dataset. The information here is not exactly real. Rather, the constructed dataset will enable you to think precisely about the sources of bias because we know what the true causal effect is.

The question we try to answer: "what is the impact of taking an AP course in high school on future earnings?" The dataset is constructed such that the causal effect of taking an AP course is $50 higher wages… You know the true causal estimate now. Therefore, by using different comparisons, we can really understand how these different comparisons generated biases.

The dataset includes the following:

schlid: school ID. The data includes 100 schools labeled 1 – 100

year: year in which the school is observed. For each of the 100 schools we have data for 10 years

studid: student ID. For each school, in each year, we have 500 students

schleff: school specific effects. The data is constructed such that larger schools perform better

ability: student-specific measure of ability, which is generated as a random variable that is a function of the school effects. Therefore, students in better schools are expected to have higher ability on average

motive: student-specific measure of motivation, which is a function of school effects

ap_schl: indicator variable for whether the student is in a school that offers AP. The top 20 schools offer AP. The rest do not

ap: a dummy variable for whether the student takes AP classes or not.

wage: student future wage, which is a function of ability, motivation, some random effects, and 50* AP. Therefore, the dataset is construct such that a student taking an AP class has 50$ more in wages.

Fall 2022
SS 340
Loujaina Abdelwahed

For now, do not worry about any variable that has number 2 in its name. Just act as if they do not exist. We will deal with them later.

Below is the summary statistics of the dataset:

	Count	Mean	SD	Min	Max
Motivation	500,000	10.15	2.02	-2.18	23.87
School Effects	500,000	15.15	8.66	0.30	30
Ability	500,000	10.15	2.02	-2.83	25.21
Wage	500,000	264.17	46.69	48.88	559.52
Student took AP class (1= Yes)	500,000	0.10	0.30	0	1
School offer AP (1= Yes)	500,000	0.20	0.40	0	1

Please bear in mind the following assumptions when you think about the questions:

- Ability is an observable characteristic that you can control for
- Motivation is an unobservable characteristic that you cannot control for. Here you have it in in the data, but in a real problem, you will not have it
- On top of motivation, each student has some specific factors that you do not see in the dataset. The combination of ability, motivation, taking AP class, and student specific characteristics is what determines the wages
- Schools have unobserved factors as well
- You know what exactly the causal effect of taking an AP class on wages is. An AP class increases wages by 50$. This is how the data is constructed.

You have a three-level panel here. School, student, year. So you can denote wages as: $wage_{ist}$ which stands for wages of student i, who went to school s, in time t. Think about each of your variables carefully, what is measured at the school-time level, and what is measured at the student-school-time level.

Below are the questions of the assignment:

1. Run a linear regression that corresponds to each of the following comparisons. In your written response, write out the equation you estimate, use Stata or Python to estimate it. You do not have to construct any tables here. Just focus on the coefficient of interest, which is the AP coefficients. Discuss if the estimates from each of these comparisons is biased or not, and explain conceptually why estimates are biased. Recall that based on part 1, you know that the true estimate is $50 and you know the exact data construction.
    a. Compare wages for students who take AP courses to students who don't.
    b. Compare wages for students who take AP courses to students with similar observables who don't.

c. Compare wages for students who take AP courses to other students at their school who don't take AP courses.

2. Briefly describe the means comparison corresponding to each of the following regressions. If the model is not identified, explain why. This question requires no coding AT ALL (I want you to think intuitively about the fixed effects here). You can try running these regressions and figuring out if the models are identified or not. But you will be graded on your explanations, not on your results. So please do not use any numbers in your write-up
   a. Reg wage= AP
   b. Reg wage=AP + ability
   c. Reg wage= AP + school FE
   d. Reg wage=AP + school FE + Time FE
   e. Reg wage =AP + individual fixed effect
   f. Ivregress wage = (AP = APoffered)
   g. Ivregress wage = (AP = APoffered) + School FE

3. Now we are going to tweak the data a little bit to introduce a treatment, which is new schools offering AP classes.
   As of now, only 20 out of our 100 schools offer AP. They all offer the AP for the whole sample period (10 years).
   I tweak the data generating process as follows:
      a. In the 5^{th} year of the data, 20 additional schools begin offering AP courses.
      b. I updated the wage construction variable to reflect the changes to AP.

These are the variables ap_schl2, ap2, and wage2

Before you start working, we want to reduce the size of the dataset a bit to make your life easier when you do DID. We want to have only 2 schools: one with AP and one without. Here is what you are going to do:

Keep only schools with ID 1 and 75. School 1 is the bad quality school that never offers AP and school 75 is one of the schools that did not offer AP before year 5, but started to offer AP from year 5.

So now, your dataset has only 2 schools: school number 1 and school number 75. Make sure that the ap_schl2 variable reflects the change in the status of the second school in terms of offering AP.

Use your smaller dataset to implement a DID analysis.

a. Run a regression to calculate the DID estimate and its standard error.
b. Create a figure illustrating the DID analysis and also showing whether there is empirical evidence in favor of the common trend assumption.
c. Based on your constructed data, explain whether the common trend assumption actually holds.