Fall 2021
SS 340
Loujaina Abdelwahed

Submit your work at the beginning of class on Monday (Nov 22). 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 $\beta 0,\beta 1$

Fall 2021
SS 340
Loujaina Abdelwahed

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.

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 the second part, we real data. Import the provided earnings dataset *schooling_earnings.csv* and get familiar with it. Here is a description of the main variables that are relevant to the analysis:

Relevant variables:

- Log of annual earnings: 'log_earnings'
- Years of schooling: 'yrsed'
- Distance to nearest college: 'dist'
- Father's college degree: 'dadcoll'
- Mother's college degree: 'momcoll'

a. We want to find out what the causal impact of one more year of schooling is on wages. Therefore, run a regression of log-earnings on years of schooling. Why can the estimated coefficient not be expected to measure a causal effect? Please explain thoroughly.

b. Use the distance of the parents' house to the nearest college (dist) as an instrument for years of schooling to estimate the (potentially causal) effect of schooling on wages. Do you believe this instrument to be good? Why or why not?

c. Now include dummies as covariates that indicate the parents' college degrees. First, fit an OLS regression. Then perform a 2SLS regression. Compare the results.

d. Fit the first stage regressions for the models with and without parental control variables. Comment on how the F statistic changes and discuss if the instrument is weak or not.