Lab II – Endogeneity & Instrumental Variables

Graduate Econometrics, Master-level (November 14, 2023)

Teacher: Yannik Obelöer (Yannik.obeloer@economics.gu.se)

Note: The deadline for this problem set is: **16th Nov at 23:59**. Please follow the instruction on submission on Canvas. The maximum mark for this assignment is 100. The points per question are indicated in brackets. The final grade for this assignment is calculated by your total points divided by 10.

You must upload:

1. Pdf file with key code excerpts and your answers in sentences to problem set questions
2. Do file (Stata) or your script
3. Log file in txt format

Part I - Endogeneity

Background and Dataset: From Canvas download the file `exam_results.dta`. The dataset contains information about students performing in an econometrics exam in their last year of their master's degree. Their study hours and attendance to lectures have been closely observed by students in the year below.[1] As the first-year students with a decent background in econometrics, we are now curious if studying is worth it. How much do I really gain on average by putting in that one extra hour in the library?

Questions:

1. Get acquainted with the dataset and produce some summary statistics. Do you see anything that surprises you? (4 points)
2. Make a scatterplot of the test score and hours studied and calculate the correlation between them. (4 points)
3. Estimate an OLS regression where test score is the dependent variable, and independent variable is hours studied:

$$test\ score\ = \beta_0 + \beta_1 hours\ studied + u_i$$

   And interpret the coefficients (6 points)
4. State the null hypothesis that you want to test to answer our research question. Test the hypothesis using 1%, 5% and 10% significance levels, and interpret the results. (4 points)
5. Plot the residuals of Model 1 against educ. According to the plot, do you think that homoscedasticity assumption is satisfied? Explain and adapt your regression if necessary. (3 points)
6. We hear that some spotters in the lab recording the hours studied fell asleep during our data collection. You decide that you rather want to categorize students into three categories than rely

[1] Don't worry: these people do not exist, but we pretend that they do for this exercise. Do not replicate this set-up for your own and your fellow students' sake.

on the precise estimates. Use: 1. Heavy studying 2. Normal studying and 3. Little studying. Create this categorical variable (at levels you deem reasonable) and run the regression above but with the categories as the independent variables. Carefully interpret your new results. (5 points)

7. You are very happy about your results and plan your study plan accordingly. Your friend raises the concern that what you are measuring is not a causal estimate. List some reasons why your friend might be right in this case and name the source of the endogeneity with the appropriate name. (8 points)
8. You realize that you did not only keep records of the time students spent studying, but also if they were present at most of the lectures. Estimate and interpret carefully: (5 points)

$$test\ score\ = \beta_0 + \beta_1 hours\ studied + \beta_2 most\ lectures + u_i$$

9. Are potential issues of endogeneity now solved? Why / why not? If they are not solved with the inclusion of this variable, come up with a hypothetical setting / experiment where you could produce a causal estimator to our research question. It must not be actually feasible. (7 points)
10. Are there potential issues that could still threaten you're the validity in case you decide to do an experiment? (4 points)

Part II – Instrumental Variables

Background and Dataset: For the IV section we will use the set-up of an actual study, please download `colonial_legacy.dta`. It is a dataset inspired by "Lowes, Sara, and Eduardo Montero. 2021. "The Legacy of Colonial Medicine in Central Africa." *American Economic Review*, 111 (4): 1284-1314."[2] From their abstract:

"*Between 1921 and 1956, French colonial governments organized medical campaigns to treat and prevent sleeping sickness* [a potentially lethal disease cause by bites of the Tsetse fly]. *Villagers were forcibly examined and injected with medications with severe, sometimes fatal, side effects. We digitized 30 years of archival records to document the locations of campaign visits at a granular geographic level for five central African countries. […]"* The authors test the sustained effects of these horrific campaigns on today's vaccination rates and trust in medicine. We will focus just on the vaccination index.

To control for geographic and individual differences, we include for a group of variables as controls. These include baseline, geographical, and institutional characteristics. They and their coefficients can be disregarded in your answers. In other words, for the sake of this exercise, assume that the addition of these controls rules out any geographic or baseline differences. Please add this list to all your regressions:

```
local controls child_age_cont child_age_cont2 b4 hv007 hv025 elev LATNUM
LONGNUM mean_temp mean_rain land_suit malaria_ecology tsi_grid_tsi
atlantic_all_years dist_missions
```

[2] We will be using a dataset similar to the one used in the actual paper, but due to data availability and privacy issues, the actual datapoints and precise results do not mirror those in the paper.

either as a local (by adding `controls'; see Stata lecture) or by copy pasting this list (excluding the words "local controls") into each regression you are running. We will also use cluster-robust standard errors at the survey cluster level. Just add ", cluster(cluster_id)" at the end of all your regressions – for the sake of the problem set you don't need to worry about this for your answers.

Questions:

11. First estimate the naïve OLS model:

$$Vaccination\ index_i = \beta_0 + \beta_1 Times\ Visit\ Prospect_i + \boldsymbol{X'_{rti}B} + \boldsymbol{u_{rti}}$$

    Where $\boldsymbol{X'_{rti}B}$ is a vector of the control variables we defined above. Interpret β_1. (4 points)
12. Explain why the OLS model is likely to not show the causal effect of colonial wrongdoings on modern day medical interventions. (6 points)

The authors propose an IV strategy using suitability for cassava (a new world staple food also known as manioc) as their instrument. Specifically, they use the log soil suitability for cassava relative to the log soil suitability for millet (relative_suitability in the dataset). Due to the way cassava is farmed (processing is done near water and less land must be cleared) there is more interaction with the Tsetse fly – the transmitter of sleeping sickness.

13. Draw a directed acyclic graph (fancy name for the graph with arrows on lecture slide 6) of this set-up and discuss briefly. You can draw this for instance using PowerPoint or Excel and just take a screenshot. (5 points)
14. What are the crucial assumptions for IVs in general and what does this mean in our specific case? (10 points)
15. What could be potential threats to the validity? Come up with some ways to potentially test the validity of the instrument (again, you can abstract from issues such as data availability – you can be creative in this section). You do not have to do the testing, just describe how one could test this. (6 points)
16. Specify the first stage equation and estimate it (again, include all controls but disregard their coefficients in your interpretation). Is this instrument a "weak instrument"? How would that be an issue? Elaborate. (6 points)
17. Specify the reduced form equation and estimate it. (3 points)
18. Use a 2SLS set-up to run the instrumental variable regression. You can use the Stata command ivreg2. (3 points)
19. Carefully interpret the IV coefficients. (6 points)