Econ 322: Econometrics

5th Assignment: Instrument Variable

Due Date: 23:00, 12/10/2023, EST

Read the following instructions carefully.

- You can submit your reports unlimited times before the deadline. Do wait until the last minute and tell me that you have an internet problem.
- Report the required results in a word or PDF document and submit it **together with your code**. In the report, make your answers concise and to the point. My TA has a lot of submissions to grade. So make sure that your answers are easy to spot, in case my TA may miss them.
- Name your submitted files in the following format: initial first_full last name_A5 (so that we know this is your answer to the second assignment).
- You can refer to the sample codes I gave in my lecture notes and the uploaded Python codes.

This assignment is designed to help you understand more about instrument variables. It is very helpful in preparing for the final exam, which will be heavily related to the IV method.

In the dataset **fertility.csv**, you will find the dataset from Botswana's (a country in Southern Africa) 1988 Demographic and Health Survey. We are going to use this dataset to explore factors that might affect the age of a woman when she has her first child in Bostswana.

We will consider a linear regression of *agefbrth* (age at first birth) on *ceb* (children ever born), *monthfm* (month of first marriage), *idlnchld* ('ideal' number of children), and *educ* (years of education).

Q1. (30 points) Report the estimated slop for *educ*.

It is reasonable to believe that the most endogenous variable is education. If a child is born at a young age, there is less time for education, and it is impossible to determine which is the causal variable.

Consider *electricity* (=1 if has electricity) as an instrumental variable. There is no reason to believe that errors in age of birth and electricity are directly related to each other. However, education and electricity are probably related because places that have electricity

are probably more developed and thus more likely to have a school. So, electricity is related to age of first birth only via education.

Q2. (20 points) Test for the relevancy of electricity as an instrumental variable: run relevancy equation where exogenous variables and instrument predict the endogenous variable. Then test whether the coefficient on the instrument is 0. There is only one instrument, hence this will be a t test.

(a) Report the formula that you are going to use for the testing regression model (in the format as depen_var $\sim$ regressor1 + regressor2).

(b) Report the robust t statistics (using "HC3") and the corresponding p-value.

Q3. (20 points) We also want to make sure that the variable *educ* is indeed endogenous. Following the procedure outlined in the lecture note (also refer to the sample code), run an OLS to test for endogeneity.

(a) Report the formula that you are going to use for the testing regression model.

(b) Report the robust t statistics (using "HC3") and the corresponding p-value.

Q3. (10 points) Now use *electricity* as the instrument to run the IV (2SLS) regression. Report the estimated slope for *educ*.

Q4. (20 points) Now we instrument for education using more than one instrumental variable. Living in an urban area should not be related to differences in the age of first birth, however, it will affect educational attainment. Again, more developed areas should (presumably) have better access to schools and education. More specically, we include *urban* (=1 if live in urban area) as another instrucment variable.

(a) Report the estimated slope of *educ* in this case.

(b) Perform the over-identification test (using robust standard errors) and report the statistics and the p-value. Is this model over-identified or not?