Copenhagen Business School | November 2022 | Applied Econometrics for Researchers

Workshop 6 – Endogeneity and Instrumental Variables

Motivation: This workshop investigates the economic returns to education using a sample of American working women. We will focus on the potential endogeneity of education – namely its consequences when using OLS estimators, how to address the issue, and how to conduct the key tests covered in the lecture.

We will consider a sub-sample of 411 women who were participating in the labor market by the time of the data collection. These data were used in a number of examples in Wooldridge, “Introductory Econometrics: A Modern Approach” (see the examples in Chapter 15).

1. **Data:**

   Use the data file mroz.dta. Let’s assume sample selection bias is not an issue in these data, so please drop the observations for which *inlf* = 0 (i.e., women who are not working, and for whom the wages are not observed) and focus on working women only.[1]

2. **Variables:**

   We will use the following variables to estimate the structural equation for wages: *lwage, educ, exper, expersq.* They refer to the 411 working women in the sample. As instrumental variables we will consider the length of education of their mother (*motheduc*), father (*fatheduc*) and husband (*huseduc*). Make sure that you know what these variables measure/how they are constructed.

   - Examine the descriptive statistics of these variables and comment on their main characteristics. *Hint: use the summarize, tab and hist commands to get an overview of the variables.*

   - Do some initial checks to test the relevance of *motheduc*, *fatheduc*, and *huseduc* as possible instruments for educ. (*Hint: check the pairwise correlations, and the*

[1] We have used this data for an example in the Lecture on Heckman models and sample selection. In that exercise, we have concluded that there is no sample selection bias in this dataset.

joint significance of their coefficients in a linear regression of educ on all the exogeneous variables.)

3. Wage regressions neglecting and addressing endogeneity in *education*

a. Estimate an OLS regression of *lwage* on *educ*, *exper*, and *expersq*. How would you interpret the coefficients? What is the potential cause of endogeneity in the *educ* variable?

b. Use the IV approach to estimate the returns to education, using one instrument (*fatheduc*). Estimate manually the reduced form equation, obtain the predicted values of *educ* and use them as a regressor in the structural equation for wage. Repeat the procedure using the *ivreg2* command (*Note: remember to install it first*).

c. Compare the coefficients of OLS and IV estimators using a Hausman test. What do you conclude? Do the results hint that endogeneity is indeed a problem?

d. Test for the presence of endogeneity using the steps of slide 20, i.e., obtain the residuals of the reduced-form equation, and include them as an additional regressor in the structural equation for wages. What do you conclude?

e. Run the post-estimation command "*ivendog*" as an alternative test for endogeneity in education. (note: you need to install this command first; type *ssc install ivendog*, or *findit ivendog* and then follow the instructions). Are your conclusions consistent with the previous tests?

f. Use both the mother's and the father's education as instruments for education. What is the estimate of the returns to education now? Given that now you have an overidentified model (i.e., two instruments for one endogenous variable), run the test for overidentifying restrictions. What do you conclude? (Note: do it first in steps, using the procedure explained in slide 22. Then confirm your results using the "*overid*" post-estimation command, after running the 2SLS regression with both instruments).

g. Now use all the three instruments to estimate the effect of education on wages. Repeat the test for overidentifying restrictions. What do you conclude regarding the correlation between the structural error term and the instruments?

h. Given the results you have obtained for different models and from different tests, what would be your preferred model to estimate the returns to education. Why?