College Proximity IV Analysis

Paul Laskowski and Jeffrey Yau

Introduction In a seminal paper, David Card attempts to measure returns to education using proximity to a 4-year college as an instrumental variable. You will perform a simplified version of his analysis in this exercise.

Load the dataset in the file card.RData. We assume a causal population model,

$$lwage = \beta_0 + \beta_1 educ + u$$

Question 1

Explain why educ is unlikely to be exogenous in this model. List a few possible omitted variables that could create bias.

Question 2

Run an OLS regression to estimate the population model above. What slope coefficient do you get? Explain whether you think endogeneity bias is driving your slope coefficient up or down and why.

Question 3

The variable nearc4 is a dummy variable indicating whether an individual lives near a 4-year college. Explain whether you believe this variable is a valid instrument for education. Test any requirements that you can test using the dataset, and argue one way or the other for requirements that you can't test directly.

Question 4

Estimate the IV regression model, including the correct standard errors.

Question 5

Considering the LATE theorem, what is the subpopulation of individuals for which the IV model estimates the average return to education?

Question 6

Which method of estimate, ols or two-stage least squares, yielded the larger estimate of return to education? Provide a possible reason for this.

Fertility IV Analysis

Paul Laskowski and Jeffrey Yau

Introduction

This execise is adapted from Wooldridge Exercise 15.C.2.

The file Fertil2.RData contains data from Botswana's 1988 Demographic and Health Survey. You will use it to estimate the effect education has on the number of children a woman has.

Note that the number of children is a count variable which only takes on a few values. Because of this, more advanced techniques (e.g. Poisson regression) may be more appropriate for this dataset. Nonetheless, we will use it to demonstrate IV estimation.

We assume a causal population model,

$$children = \beta_0 + \beta_1 educ + u$$

Question 1

Explain why educ is unlikely to be exogenous in this model. List a few possible omitted variables that could create bias.

Question 2

Run an OLS regression to estimate the population model above. What slope coefficient do you get? Explain whether you think endogeneity bias is driving your slope coefficient up or down and why.

Question 3

The variable frsthalf is a dummy variable indicating whether an individual was born in the first 6 months of the year. Explain whether you believe this variable is a valid instrument for education. Test any requirements that you can test using the dataset, and argue one way or the other for requirements that you can't test directly.

Question 4

Estimate the IV regression model, including the correct standard errors.

Question 5

Which method of estimate, ols or two-stage least squares, yields the larger estimate of return to education? Provide a possible reason for this.