Data Analytics for Business

Homework 4 – Part 1

Instrumental Variable

Return on education has been a question attracting constant research interest with competing research findings. The general notion is that better-educated workers earn higher wages. Nonetheless, length of education is a choice variable of each individual and is affected by lots of unobservable factors that may also affect the earnings of individuals. This constitutes a classical example of endogeneity caused by omitted variables. Simply regressing earnings on years of education will lead to biased estimation of return on education. However, the direction of the bias could be hard to determine and may result in counterintuitive findings.

In this exercise, let’s replicate (in a simplified version) a real research study on return on education based on a real data set by using instrumental variables to correct the potential bias caused by the endogeneity issue, and see how return on education could have been misleadingly affected.

The data file is "**Education_data.csv**", which has the following variables:

Data Description

Column Name	Variable Description
id	The id of individual survey participant
wage	Hourly wage, in cents
educ	Years of schooling
exper	Years of post-school experience (calculated as age-educ-6)
nearc4	Whether lives near a 4-year college or not
fatheduc	Father’s years of schooling
motheduc	Mother’s years of schooling

Import that data into R and perform the following analysis.

Tasks:

1. Consider the following linear regression model:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u$$

   First run a simple OLS and report the estimation results.

2. Since the variable of interest $educ$ is likely to be endogenous, we need to find some good instrument(s) for it. Among the possible candidates: *nearc4*, *fatheduc*, and *motheduc*, which one(s) do you think could be the most appropriate IV(s)?

3. Let's use *nearc4* as the instrumental variable. We will first perform the two-stage least square manually. First, perform the first-stage regression. Is the endogenous variable $educ$ is correlated with the proposed IV *nearc4*?

4. Then perform the second-stage regression, and obtain the 2SLS estimation results.

5. Next, use the **ivreg()** function in the **AER** package, and verify that all coefficient estimates are the same as your manual 2SLS results.

6. Now compare the estimated coefficients from the OLS and the 2SLS. How did OLS bias the estimated return on education? Is the direction somewhat counterintuitive? Can you come up with some reasoning to justify the finding?