PhD Econometrics (ECON50580)
Problem Set 2: Instrumental Variables

This problem set will be graded. Rules:

- You can work in groups of 1-5 students
- Submit your solutions in 1 pdf
- The code should be in the appendix
- Results should be presented graphically or in tables. No screenshots from statistical software!
- Use soft-coding in your code.
- For empirical exercises, show evidence that you used version control
- Submit via Brighspace

Submission deadline: Wednesday, 10 February, 3:59:59pm.

---

1 Theory

a) Suppose you have a binary instrumental variable z. Consider the regression model

$$y = \beta_0 + \beta_1 x + u$$

Show that the IV estimator for β_1 is

$$\widehat{\beta_1} = \frac{\bar{y_1} - \bar{y_0}}{\bar{x_1} - \bar{x_0}},$$

where $\bar{y_1}$ is the mean of y if $z = 1$ and $\bar{y_0}$ the mean of y when $z = 0$, and analogously for $\bar{x_1}$ and $\bar{x_0}$.

b) In the lecture we have proven that in absence of defiers the IV estimand for an outcome Y_i, a binary treatment D_i and a binary instrument Z_i is

$$\beta^{IV*} = E(Y_i|D_i = 1) - E(Y_i|D_i = 0)) = E(Y_{i1} - Y_{i0}|complier). \quad (1)$$

Now suppose the share of defiers is $0 < \pi_D < 1$.

i) Derive the IV estimand for this case.
ii) Assume that $\pi_D = a \times \pi_C$ with $0 < a < 1$ and $\beta^{IV*} > 0$. Discuss whether additional, plausible assumptions on $E(Y_{i1} - Y_{i0}|defier)$ allow you to recover a lower bound for β^{IV*}.

---

2 Empirical Application

The empirical application is based on a cross-sectional dataset *assign2.dta*, which contains the following variables:

- *age*: age of surveyed individual
- *logearn*: log annual earnings
- *yob*: year of birth

- *schooling*: age at which the person left school.

a) The goal is to estimate the returns to education. For this purpose, estimate an OLS regression of logearn on schooling, controlling for fourth-order polynomials in age and year of birth. Interpret the coefficient of schooling.

A common way to obtain causal estimates is to use changes in compulsory schooling laws for identification. In this case, birth cohorts born before 1933 (yob<33) had to go to school until they were 14 years old, whereas compulsory schooling age was raised to 15 years for all cohorts born from 1933 onwards. This change in compulsory schooling laws can be used as an instrumental variable for the actual duration of schooling. The instrument is a dummy LAW that equals unity if a person is born 1933 or later and zero otherwise.

b) Discuss this instrument in theory, assuming that schooling S_i is related to the instrument Z_i by the latent assignment mechanism

$$S_i = \mathbf{1}(\gamma_0 + \gamma_1 Z_i > \eta_i), \text{ with } E(Z_i \eta_i) = 0.$$

The random variable η_i represents the individual resistance to treatment. Why could there be a first stage? Under what condition is this instrument valid? What are potential threats to validity? Furthermore, explain who are the compliers, always-takers and never-takers in this case. Would the IV estimate correspond to the average treatment effect (why or why not)?

c) As in any good empirical project, begin with a graphical inspection of the relationships of interest. This is best done through binscatters. Produce and discuss the graphs listed below. In all graphs, include a vertical line at *yob=33*.

- Plot the probability that a person leaves school before age 15 against the year of birth.
- Binscatter of schooling and year of birth
- Binscatter of log earnings and year of birth.

d) Calculate the Wald estimator (without controls) "by hand", i.e. based on conditional averages. Compare your results to those of a 2SLS estimation based on an inbuilt command (e.g. *ivregress* in Stata or *ivregress* in R). Interpret your results and compare them to the OLS results in a).

e) Now we estimate the resturns to education with an instrumental variables estimator using the same controls as in a). Do the following:

- Estimate the first stage and reduced form and interpret the results. Compute the IV estimate from these results.
- Separately estimate the first and second stage.
- Estimate the model with an inbuilt 2SLS command. Compare the coefficient and standard error of *schooling* to those obtained in the previous regressions in e).
- Comment on the difference between the OLS and IV results.
- Comment on the difference between the IV results and the estimates obtained in d).
- Comment on the strength of the first stage using appropriate techniques