Problem Set 2 - Econometrics

Giacomo Lo Conte, Kun Wu, Francesca Eustacchi, Neeharika Kakunuri, Noor Ahmed Khoso

February 1, 2022

1. Theory

a. Consider a case where Z is a binary instrument, T is a discrete multi-valued treatment ($TZ \in 0, 1, 2, ..., J$), $T = ZT_1 + (1Z)T_0$ is the observed level of treatment, and Y is the outcome. Assume that the random variables $T_0, T_1, Y_0, Y_1, ..., Y_J$ are jointly independent of J and that the level of treatment is higher when $Z = 1$ than when $Z = 0$, i.e. $T_1T_0 \geq 0$. Further, assume that $Pr(T_1 \geq j \geq T_0) > 0$ for at least one $j \in 0, 1, 2, ...J$, which means that the instrument on average affects the level of treatment. Show that under these assumptions the Wald estimator is a weighted average of the causal response of a unit change in treatment,

$$\frac{E[Y|Z=1] - E[Y|Z=0]}{E[T|Z=1] - E[T|Z=0]} = \Sigma^J j_{=1} w_j E[Y_j - Y_{j1}|T_1 \geq j > T_0], \tag{1}$$

where the weight are

$$w_j = \frac{Pr(T_1 \geq j > T_0)}{\Sigma^J{}_{i=1} Pr(T_1 \geq i > T_0)} \tag{2}$$

b) Provide an interpretation of both factors on the RHS if Equation (1), i.e . $E[Y_jY_j1|T1 \geq j > T_0]$ and j. What does the presence of both factors teach us about the local average treatment effect?

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

$$\beta^I V = E(Y_i|D_i = 1)E(Y_i|D_i = 0)) = E(Y_{i1}Y_{i0}|complier) \tag{3}$$

Now suppose the share of defiers is $0 < \phi D < 1$. i) Derive the IV estimate for this case. ii) Assume that $\phi D = a\phi C with 0 < a < 1 and \beta^I V > 0$. Discuss whether additional, plausible assumptions on $E(Y_{i1}Y_{i0}|defier)$ allow you to recover a lower bound for $\beta^I V$.

2. Simulation Exercise

b) Weaker instruments Now repeat the analysis from a), but instead assume that the correlation between the instrument z and the regressor x is 0.15. Discuss the difference in sampling distributions between OLS and IV and, in addition, discuss the difference between sampling distributions based on weak and strong IVs.**An important property of the IV estimator is that it is biased in small samples but consistent. For this reason one should never write that an IV estimator is used to obtain unbiased estimates. We want to better understand the small sample properties through simulations of the sampling distribution of IV and OLS estimators. In all simulations, let x, y, z, u and be random variables and assume the data-generating process**

$$y = \alpha + \beta x + \epsilon \tag{4}$$

$$x = \gamma_0 + \gamma_1 z + u \tag{5}$$

Set the parameter = 1 and, unless required otherwise, = 0 = 0. Moreover, construct such that it's Pearson correlation with x is 0.4.

a) Sampling Distribution under strong instruments

Construct the instrument z such that its Pearson correlation with $xis0.5$ while its correlation with ϵ is zero. Consider four different sample sizes: $N = 50, N = 100, N = 250andN = 1000$. For each sample size, run at least $10,000$ simulations whereby you estimate β^OLS and β^IV (choose fewer replications if you have problems with computing power). In the same graph, plot the sampling distributions for OLS and IV for all four sample sizes. Discuss the difference in shape of the sampling distributions between OLS and IV.

b) Weaker instruments

b) Weaker instruments Now repeat the analysis from a), but instead assume that the correlation between the instrument z and the regressor x is 0.15. Discuss the difference in sampling distributions between OLS and IV and, in addition, discuss the difference between sampling distributions based on weak and strong IVs.

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 coeffcient 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 Si is related to the instrument Zi by the latent assignment mechanism

$$S_i = 1(\gamma_0 + \gamma_1 Z_i > \eta_i), with E(Z_i \eta_i) = 0$$

The random variable 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? 2 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 *ivreg* in R). Interpret your results and compare them to the OLS results in a).