Problem Set 2

Instructions: Be sure to hand-in your computer code with sufficient annotation.

Note: For computational questions, your parameter estimates will vary with your random number draws. If your answers are more than 25 percent different than the known corresponding estimand values, it likely reflects some mistake. You can check your work/coding by increasing the sample size to see if your estimates start to converge to the estimands.

Part 1: Analytic Exercises

1) Assume a returns to schooling model with just 1 unobserved variable ability A distributed $U[0,1]$. Our potential outcomes model for earnings by schooling (treated) or not is

$$Y_1 = 1 + 0.5A$$

$$Y_0 = A$$

Treatment/schooling is determined by

$$D = 1\{-0.5 + A > 0\}$$

a) What is the ATE?

b) What fraction of the population takes the treatment ($pr(D = 1)$)?

c) What is the maximum treatment effect? What is the minimum treatment effect?

d) If A were instead distributed Normal, $A \sim N(0,1)$, what would be the minimum and maximum treatment effects? (Assume $A \sim N(0,1)$ for this part only. For all other parts continue to assume the model above with $A \sim U[0,1]$.)

e) Compute the ATET and ATEU.

f) Why is the ATEU > ATET? Explain in words.

g) Compute the OLS estimand for the effect of Y on D: $\beta(OLS) = E(Y|D=1) - E(Y|D=0)$.

h) Explain in words why the OLS estimand is biased upward for the ATE?

2) Assume the potential outcomes model with $V = \delta_0 + \delta_1 Z + U_V$, for instrument $Z \in \{0,1\}$.

a) Prove that this model implies the Angrist and Imbens monotonicity assumption. (Note: I have omitted i subscripts, but δ_0 and δ_1 are homogeneous parameters.)

b) In this model, write a new function for V such that monotonicity does not hold.

3) Assume U_V is distributed Uniform $[-2,2]$, and $V = Z + U_V$ with $Z \in \{0,1\}$:

a) Show the range of U_V values for the complier, defier, always taker, and never taker groups.

b) Compute the fraction of the population in each group.

4) Assume there are 2 types in the population. Type 1 has treatment effect $\Delta = 2$ and Type 2 has $\Delta = -1$. 30 percent of the population is Type 1, 70 percent Type 2. Type 1s have utility given by $V = Z + U_V$, with $U_V \sim U[-1,1]$ and Type 2s have utility given by $V = 2Z + U_V$, with $U_V \sim U[-1,1]$. Let the instrument $Z \in \{0,1\}$ and $Pr(Z=1) = 0.5$.

a) Compute the ATE.

b) Compute $Pr(D=1|Z=1)$ and $Pr(D=1|Z=0)$.

c) Compute the LATE. (Hint: Although this should be computed analytically, feel free to use STATA to program this scenario using random number draws–this will help to check your math.)

Part 2: Monte Carlo Exercises

Question 1

Assume log hourly earnings for individual i takes this form:

$$\ln w_i = 1 + 0.05 s_i + 0.1 a_i + \epsilon_i,$$

where s_i is units of observed schooling and a_i is unobserved ability. $\epsilon_i \sim N(0, 0.5)$. $a_i \sim N(0, 4)$

Units of schooling for each individual i are

$$s_i = 3a_i + z_{i1} + z_{i2} + \eta_i,$$

where $\eta_i \sim N(0, 1)$. z_{i1} and z_{i2} reflect the cost of schooling for individual i. $z_{i1} \sim N(0, 0.1)$. $z_{i2} \sim N(0, 25)$. There is another variable z_{3i} which is distributed Uniform $[0, 1]$. z_{1i}, z_{2i}, z_{3i} are all independent of each other, and independent of η_i, a_i, ϵ_i.

a) Write a computer program that simulates $N = 2,000$ observations from this population. An observation is a log wage ($\ln w_i$), an observed level of schooling (s_i), and instruments z_{i1}, z_{i2}, z_{i3}. We do not observe a_i.

b) Compute the OLS estimates for β_1 and β_2 for this model:

$$\ln w_i = \beta_1 + \beta_2 s_i + u_i.$$

c) We believe s_i is an endogenous variable. Consider three different instruments for s_i, z_{1i}, z_{2i}, z_{3i}. Discuss whether each of these instruments is valid. Calculate the correlation coefficient between each instrument and the endogenous variable s_i. Which instruments would you consider "weak"?

d) Compute the 2SLS estimate for the β_1 and β_2 parameters for all possible combinations of the instruments (7 total: each separately, the three combinations of two instruments, and using all three). For each combination of instruments, report the the estimates, standard errors, and first stage F-statistic in one concise table.

e) Discuss the relationship between the OLS estimates and the IV estimates. How does the strength of the instrument affect the 2SLS estimates?

f) Repeat a)-d) using a dataset of 500,000 observations. Report the results in a separate column next to the column for the results using the $N = 2,000$ observation data set.

Question 2

Consider the following potential outcomes model:

$$Y_0 = 1 + U_0$$

$$Y_1 = 4 + U_1$$

$$V = -1 + 2Z + U_V$$

with $(U_0, U_1, U_V) \sim N(0, \Sigma)$, and Σ corresponding to the following variance-covariance elements: $V(U_0) = 1$, $V(U_1) = 1$, $V(U_V) = 1$, $Cov(U_0, U_1) = 0.5$, $Cov(U_0, U_V) = 0.3$, and $Cov(U_1, U_V) = 0.7$. Z is a valid instrument with $Z \in \{0, 1\}$, and $Pr(Z = 1) = 0.3$.

a) Write computer code to simulate this model for 10,000 draws from the (U_0, U_1, U_V, Z) distribution.

b) Compute the following: i) ATE, ii) ATET, iii) ATEU, iv) OLS/Naive estimator, v) Direct/Reduced Form/ITT estimator, vi) IV estimator.

c) What fraction of the sample is a complier?

d) Re-compute the model and answer Parts b) and c) with $V = -1 + 3Z + U_V$.