Econometrics

Problem Set 1 (due 09/14 at 11:59 p.m.)

Note: This problem set is to be done jointly with your reading of Chapters 1, 2, 4.1 (IV and Causality) and 4.4 (IV with Heterogeneous Potential Outcomes) of the book "Most Harmless Econometrics", by Angrist and Pischke.

1 Conceptual

For each question below, explain how you would implement an experiment to answer this question. What problems do you expect to encounter? (This question is vague on purpose, make sure to try to get more precise as you answer each question. It is OK to consider an experiment that is not actually feasible to implement, as long as you describe the most feasible experiment you can conceive.)

1. What is the effect of whether the child watches TV on their math skill.
2. Is working around retirement age good for the person's health?
3. Is the racial wage gap in part due to discrimination against racial minorities? (To keep things simple: consider only two groups: racial minorities vs. racial majorities or Blacks vs. Whites.)
4. What is the effect of wether the mother receives welfare money support while the child is young on the child's future earnings (by age 40)?

Note: You will notice that in all questions above the treatment variable is binary, and yet there is some loss of information in pretending the treatment variable of interest is binary. We will discuss a lot about how to improve on that in future classes.

2 Coding

1. Create a new dataset in either Stata or R with the following features:

   (a) Set your random generator seed as 1 (this is for your results to always be replicable).

   (b) $N = 10,000$ observations.

   (c) Draw ϵ_i^D and ϵ_i^Y independently from each other, each distributed as a $\mathcal{N}(0, 1)$.

   (d) Draw $U_i \sim \mathcal{N}(0, 0.5)$. Note: standard deviation= 0.5, not variance.

   (e) Create $Z_i = \mathbf{1}(z_i > 0.5)$, where z_i is randomly drawn from a uniform distribution between 0 and 1. $\mathbf{1}(\cdot)$ is an indicator function that takes the value 1 if the expression in parenthesis is true, and 0 otherwise.

   (f) Create $D_i = \mathbf{1}(\alpha_0 + \alpha_Z Z_i + \alpha_U U_i + \epsilon_i^D{>}0)$, with $\alpha_0 = -4$, $\alpha_Z = 5$ and $\alpha_U = 4$.

   (g) Create $Y_i = \beta_0 + \beta_D D_i + \underbrace{\beta_Z Z + \beta_U U_i + \epsilon_i^Y}_{\varepsilon_i}$, with $\beta_0 = 3$, $\beta_D = 2$, $\beta_Z = 0$ and $\beta_U = 6$.

2. Run an OLS regression of Y on D (with a constant).

3. Run an IV regression of Y on D (with a constant) using Z as the IV for D.

4. Run an OLS regression of D on Z (with a constant), get this regression's predicted value and residual, and call them respectively $\hat{D}$ $(= L(D|Z))$. You can program it as "Dhat") and $\tilde{D}$ $(= D - L(D|Z))$. You can program it as "Dres"). Then, run two new OLS regressions:

(a) regress Y on $\hat{D}$ (with a constant);

(b) regress Y on D and $\tilde{D}$ (with a constant).

Explain why the coefficient of $\hat{D}$ in regression (a) and the coefficient of D in regression (b) are the same (down to several decimals). Explain why both are also the same as the coefficient of D in the IV regression in the previous item. What is the intuitive interpretation of the coefficient of $\tilde{D}$ in the second regression? Optional: explain intuitively why the standard errors of the three estimates we just mentioned (the ones that are identical) are so different across these three regressions.

5. Change the data generating process (d.g.p.) so that $\beta_Z = 1$ and re-do item 3. Next, change the d.g.p. so that $\beta_Z = -1$ and re-do item 3. Explain why the IV estimators of β_D in these two cases are biased. Explain intuitively why the bias of the estimator of β_D changes sign depending on the sign of β_Z. Optional: explain the intuition for why the bias of the estimator of the coefficient of D increases when we go from $\beta_Z = 0$ to $\beta_Z = 1$.

6. Re-do item 3 for $\beta_Z = 0$ and $\beta_Z = 1$ when $\alpha_Z = 10$. Compare how the bias of the coefficient of D changes between the baseline case ($\alpha_Z = 5$) and the new case ($\alpha_Z = 10$). Explain the intuition for these results. Optional: explain the intuition for why the bias of the estimator of the coefficient of D increases further when we go from $\beta_Z = 0$ to $\beta_Z = 1$ when $\alpha_Z = 5$ than when $\alpha_Z = 10$.

7. Go back to the baseline d.g.p. where $\beta_Z = 0$ and $\alpha_Z = 5$. By simply looking at the equation describing D shown above, explain why we are sure that there are no defiers in this example. (Remember: Defiers in this case are those observations where D_i would have been 0 if $Z_i = 1$ but D_i would have been 1 if $Z_i = 0$.)

8. Go back to the baseline d.g.p. where $\beta_Z = 0$ and $\alpha_Z = 5$. Calculate the proportion of compliers, always-takers, and never-takers. (Remember: Compliers in this case are

those observations where D_i would have been 1 if $Z_i = 1$ but D_i would have been 0 if $Z_i = 0$. Always-takers are those observations where D_i would have been 1 irrespective of the value of Z_i, and never-takers are those observations where D_i would have been 0 irrespective of the value of Z_i.)

9. Change the d.g.p. so that $\alpha_Z = 10$. Calculate the proportion of compliers, always-takers and never-takers, and compare these proportions with the proportions in the previous item.

10. How does the correlation between D and Z vary with the proportion of compliers when we compare the cases $\alpha_Z = 5$ and $\alpha_Z = 10$? Explain this relationship intuitively.

11. Irrespective of whether $\alpha_Z = 5$ or $\alpha_Z = 10$, there is no external validity problem (e.g., $LATE \neq ATE$) in this case. Show that you can justify this assertion directly from item 1 alone.

12. Set your random generator seed as 2, then 3, then 4. For each of these seeds, run the whole code again, and check that all questions above are answered in the same way, but note that numbers changed a bit. (No need to put these results in your answer. We just want you to do this exercise and note the similarity of your answers, but also note how the results change because of random sampling. We want you to get an intuitive feel about how random sampling makes things a bit different, but in ideal scenarios – as this one – random sampling is very well behaved, so nothing major should be happening.)

13. Re-create the analogous dgp but setting $N = 500$ instead of $N = 10,000$. Compare results for seeds 1, 2, 3 and 4, checking that they change more across seeds now relative to the previous item. This happens because the conditions are not as ideal as before, since N is much lower. (Again, no need to show results here, this is just for you all to use this as a sandbox and get some intuition about the role of random sampling. You

will see that if you try much lower N the results will start to get crazy.)