Problem Set 3
Quantitative and Statistical Methods II

General Instructions

- The problem set is due by February 26th at noon;
- You should send it by email *bruno.conte@barcelonagse.eu*; it must include your *unique* Stata code, *a unique* log file, and your answer sheet. Tidiness is appreciated – e.g. material correcly labeled and organized in a zip file;
- Working in teams is allowed and strongly recommended (keeping the same groups as in the presentations is encouraged);
- The datasets needed to solve the computational questions are uploaded in the Classroom material.

Part 1: Dell, M., 2010. The persistent effects of Peru's mining mita. *Econometrica*.

1. Explain what the "mita" system was and how it creates a discontinuity in the paper's context. Is that a sharp or fuzzy RD design, and why? Explain what "sharpness" and "fuzziness" mean in this context.
2. Recall the two conditions that characterise an RD setup, discontinuity of treatment on Z_i but continuity of potential outcomes on Z_i in a neighbourhood of z_0, respectively formalised as

$$\lim_{z \to z_0^+} P(D_i = 1|Z_i = z) \neq \lim_{z \to z_0^-} P(D_i = 1|Z_i = z), \tag{1}$$

$$\lim_{z \to z_0^+} P(Y_{ij} \leq r|Z_i = z) = \lim_{z \to z_0^-} P(Y_{ij} \leq r|Z_i = z) \quad (j = 0, 1). \tag{2}$$

   (a) Show analytically how condition (2) above allows you to get

$$\lim_{z \to z_0^+} \mathbb{E}[Y_{ij}|Z_i = z] = \lim_{z \to z_0^-} \mathbb{E}[Y_{ij}|Z_i = z] \quad \forall j. \tag{3}$$

   [Hint: use the fact, as shown in class, that $P(Y_{ij} \leq r|z_i = z) = F_{Y_{ij}}(r|z_i = z), \quad \forall j$]

   (b) Suppose we are working on a *homogeneous* effects, *sharp* RD framework, so that $Y_i = \alpha D_i + Y_{0i}$. Use your previous findings to demonstrate formally that α can be identified as

$$\alpha = \lim_{z \to z_0^+} \mathbb{E}[Y_i|Z_i = z] - \lim_{z \to z_0^-} \mathbb{E}[Y_i|Z_i = z]. \tag{4}$$

3. Explain why Table 1 provides empirical evidence for the condition (3) to hold in the context of the paper. Why is that condition crutial for interpreting RD estimates as causal?

4. Consider now equation (1) of the paper.

   (a) What does $f(\text{geographic location}_d)$ stand for, and why is so important for the estimation of α_{RD}?

   (b) Table 2 describes the estimations results of that equation. What is the cutoff z_0 in the specification of Panel C? That is, what is z? How would you define D_i as a function of it?

   (c) Use `delldata_consumption.dta` and `delldata_childstunt.dta` to replicate the results of Panel C. Make sure you specify $f(\text{geographic location}_d)$ and D_i as discussed in (4a) and (4b), and that you check the table's notes to understand which is the specification used. *[Hint: interpret, for this application, the "euclidian distance" as its absolute value!]*

5. Suppose we are interested in learning more about *fuzzy* RD frameworks with *heterogeneous* effects. Which is the difference with respect to the relation between D_i and Z_i that makes α_{RD} different from what you obtained in point (2b)?

Part 2: Angrist, J.D. and Lavy, V., 1999. Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*.

6. Explain what the Maimonides' rule is and how it creates a discontinuity in class size assignment to children. Is that a sharp or fuzzy RD, and why?

7. The data file `angrist1999.dta` contains the data used in that paper. Use it for the following points:

   (a) Reproduce Figure 2-Panel A, but using math scores instead of reading scores. Which preliminary conclusions about the relationship between class size and school performance?

   (b) Now run an OLS regression of "average math score" on "average class size", controlling for "percentage of disadvantaged kids" and "enrollment". Is the estimate for "average class size" causal? Is its sign consistent with your discussion in question (7a)?

   (c) Implement an IV estimation of "average math test scores" on "average class size" by instrumenting it with the "Maimonides Rule". Control for "percentage of disadvantaged kids" and "enrollment". Comment the estimate for "class size" – is it causal, and how it compares to what was found in (7b)? If causal, which assumptions we need to make for that to hold?

8. Let us now give a LATE-type of interpretation to our problem. Disregard enrollment cohorts with more than 50 students. Suppose we are interested in looking at the local effect in the neighbourhood of 5 students around the class size cutoff (i.e. $z_0 = 40$, $e = h = 5$).

   (a) Which is the extra assumption needed for this estimation, and how is it formally stated? How would it be interpreted in the current context?

   (b) Estimate a LATE-like α_{RD} for math scores as the outcome variable. That is, estimate[1]

$$\hat{\alpha}_{RD} = \frac{\mathbb{E}[Y_i|W_i = 1, S_i = 1] - \mathbb{E}[Y_i|W_i = 0, S_i = 1]}{\mathbb{E}[D_i|W_i = 1, S_i = 1] - \mathbb{E}[D_i|W_i = 0, S_i = 1]} \tag{5}$$

[1] You will be looking at the cohorts with $z_0 \pm h$ pupils. S_i is a dummy for belonging to such group, and W_i a dummy for belonging to an enrollment cohort above the class size cutoff.