Problem Set 4

Goal: Explore Regression Discontinuity & IV
Application: Duflo et al. 2011

In this problem set, we will look at Duflo et al. (2011). This paper looks at the effect of tracking on educational attainment. "Tracking" means that high-performing and low-performing students are separated into two sections. The authors run experiment in which they randomly assign 60 out of 120 primary schools in rural Kenya to tracking. In Table 5, the authors explore the effects of being in the bottom section of a tracking school. The problem set will focus mostly on replicating this table. Some of your estimates will not be exactly the same as those in Table 5 of the paper, but they should be very close.

The name of the dataset is `forpeerpapers.dta`. `tracking` is a dummy variable that denotes whether the student is a tracking school. `bottomhalf` denotes the dummy variable B_{ij} which is 1 if the student was assigned to the bottom section and 0 otherwise, `percentile` denotes P_{ij} which is the student's test score percentile at baseline, `etpteacher` denotes the type of teacher, `girl` denotes gender, `agetest` denotes age, and `stdR_totalscore` is standardized test scores.

Part 1. Introduction

1. Read the introduction of the paper. List three channels through which tracking might affect students, according to the authors.

Part 2. Regression Discontinuity

In Panel A of Table 5, the authors use a regression discontinuity approach to explore the effects of being in the bottom section of a tracking school on test scores.

1. Let's start with Specification 1. This is a classic regression discontinuity, in which the authors estimate the below equation, where B_{ij} is a dummy for whether or not individual i in school j is in the bottom section, P_{ij} is the student's test score percentile at baseline, and X_{ij} is a vector of controls (including a constant, type of teacher, gender, and age at time of test).

$$y_{ij} = \delta B_{ij} + \lambda_1 P_{ij} + \lambda_2 P_{ij}^2 + \lambda_3 P_{ij}^3 + X_{ij}\beta + \epsilon_{ij} \tag{1}$$

   (a) Explain why δ is a plausible causal estimate of the effect of being in the bottom section on educational attainment

   (b) Load the data and create a new data frame that includes tracking schools only (tracking==1). You will use this data frame for the rest of the problem set.

   (c) Create variables P^2 and P^3 and estimate Equation 1 to replicate the coefficient in Column 1.

   (d) Estimate the standard error of $\hat{\delta}$, using bootstrap and clustering at the school level.[1] (*Hint*: Code for doing this using loops has been provided for you on moodle,

[1] When estimating regression discontinuity models, we cannot use conventional standard errors – see Lee and Card 2007 – so the boostrap is a good alternative. Moreover, since the treatment was randomized at the school level and it is plausible to assume that students' shocks are correlated within school, we need to cluster our standard errors at the school level.

since this is a hard problem. You can also try using the package `multiwayvcov` and the function `cluster.boot`.)

(e) Explain why school fixed effects might be especially important in this context (keeping in mind your answer to (a)). Replicate the coefficient in Column 2 by adding school fixed effects to the model in Equation 1. Report your results omitting the coefficients on the school dummies.

(f) Compute the cluster standard error for $\hat{\delta}$ using the bootstrap. Note that, when running this regression on the bootstrapped sample you must use the new school indexes (`df_boot$new` in my code) to generate your school fixed effects.

2. Now let's move on to Specification 2. Here, the authors estimate something similar to Equation 1, except this time, they "estimate a second order polynomial separately on each side of the discontinuity" (p. 1754). This amounts to estimating the following equation (adding the terms in red to Equation 1):

$$y_{ij} = \delta B_{ij} + \lambda_1 P_{ij} + \lambda_2 P_{ij}^2 + \phi_1 P_{ij} * B_{ij} + \phi_2 P_{ij}^2 * B_{ij} + X_{ij}\beta + \epsilon_{ij} \tag{2}$$

(a) Explain why this might change our estimated effect of being in the bottom section.

(b) Create the necessary variables and estimate the model in Equation 2.

(c) Replicate the coefficient in Column 3 by predicting the effect of being in the bottom section for a student at the 50th percentile.

(d) Compute clustered standard errors using the bootstrap by adapting the code provided.

Part 3. Instrumental Variables

In Panels B and C of Table 5, the authors examine one channel by which being assigned to the bottom section might affect attainment: low peer quality. To do this, they use instrumental variables. The first stage will regress mean peer score ($\bar{y}_{-ij}$) on being in the bottom half, which is plausibly exogenous when we include controls for the flexible polynomial of baseline attainment, for the reasons discussed in 1a.

$$\bar{y}_{-ij} = \delta B_{ij} + \lambda_1 P_{ij} + \lambda_2 P_{ij}^2 + \lambda_3 P_{ij}^3 + X_{ij}\beta + \epsilon_{ij} \tag{3}$$

The second stage will regress endline scores on mean peer scores, using the predicted values of mean peer score from the first stage regression.

1. First we will replicate the first stage regression in Panel C Column 1

(a) Create a new data frame that contains the subset of observations for which *none* of the variables in the model are missing. (*Hint:* Use `na.omit` and `subset` with the option *select*.)

(b) Estimate the first stage by estimating Equation 3 with `rMEANstream_std_total` as your outcome variable.

(c) Compute clustered standard errors using the bootstrap by adapting the code provided

2. Now let's replicate the second stage regression in Panel B Column 1.

   (a) First, do this "manually" by estimating the following:

$$y_{ij} = \delta \hat{y}_{-ij} + \lambda_1 P_{ij} + \lambda_2 P_{ij}^2 + \lambda_3 P_{ij}^3 + X_{ij}\beta + U_{ij} \tag{4}$$

   where $\hat{y}_{-ij}$ are the predicted values of $\bar{y}_{-ij}$ from the first stage regression. To estimate $\hat{y}_{-ij}$ in R, you can use `firststage$fitted.values`, where `firststage` is the object that stores the first stage regression.

   (b) Second, do this using `ivreg` from the AER package. Type `?ivreg` in the console to look up the syntax. You should replicate your coefficient from (a) exactly (but not the standard error).

   (c) Compute clustered standard errors using the bootstrap by adapting the code provided

3. When using an instrumental variable, one of the identification assumptions is that the instrument is exogenous: $Cov(B_{ij}, U_{ij}) = 0$. In plain English, this assumption imposes that being assigned to the bottom session only impacts your tests scores by changing the quality of your peers (y_{-ij}). As consequence, the exogeneity assumption implies that that there is no direct impact of being assigned to the bottom session on your own test scores. Do you believe this assumption is plausible? Think about what unobservables variables are inside the term U_{ij} or about possible direct effects of being assigned to a bottom session.