Problem Set 3 and 4 - RDD and Panel Data.

Due: Dec 15th at 17:00 by email to Carrie Huffaker (carrie.huffaker@phd.unibocconi.it)

December 1, 2015

1 Problem Set 3: RDD[1]

In this assignment we use data on Italian Municipalities from 1993 to 2001 to study the effect of a wage increase on political selection. We study the effect for elected majors only, using the dataset SIN19932001.dta posted online.

1. According to the Italian legislation the wage of the major increases by 33% with the size of the resident population at a threshold of 5000 inhabitants. How would you exploit this fact using an RDD design to estimate the effect of a wage increase on the years of education of the major? Would you implement a fuzzy or a sharp design?

2. We want to test formally the assumption that a higher wage attracts more citizens with high opportunity cost into politics, that is, more skilled individuals with high alternative remunerations in the private sector. Define X_i as the general characteristics of town i, Y_i as some education indicator, P_i as the population size, W_i as the wage paid to the major. The wage sharply increase at the population threshold P_c. So if $P_i > P_c$ then $W_i = W_h$ otherwise if $P_i < P_c$ then $W_i = W_l < W_h$. Using the potential outcome framework, state the assumptions needed to get the estimand of interest and how you identify the effect of the wage increase on political selection.

3. According to Ade and Freier (2011) when dealing with population thresholds we should verify three fundamental facts to be sure our estimation strategy holds. What checks would you do to provide evidence that these three assumptions are satisfied in this problem?

   (a) The population threshold only defines the treatment considered and NOT additional simultaneous exogenous co-treatments

   (b) No additional endogenous choices on other institutions are taken simultaneously

   (c) No manipulation and precise control over population measure

[1] I thank Emanuele Dicardo for preparing this problem set as part of his in-depth assignment.

4. Graph the years of education of the winning major in a municipality against the size of the population in that city. In the dataset SIN19932001.dta, the former can be determined by looking at the variable "years_school"; the latter can be determined from the variable "pop_census." Note: you do not simply want to graph years_school against pop_census; instead, generate "bins" that correspond to small categories of pop_census and within each bin (which will contain the size of municipalities with very similar pop_census) calculate the mean value of the schooling years of the majors. You can use the variable "pop_census_bin" already generated for you.

5. Does a simple scatter plot give a clear picture of the situation? Do you identify a clear jump at the threshold?

6. How does the graph change when you add to the same plot local linear (LLR) and weighted (LWR) regressions to have a clearer picture? You need to use the twoway command. For the weighted regression you should simply add the command "lowess" and you have to do it twice. First to the left of the 5000 threshold, then to the right. To add the linear regression fit, instead, you first need to regress years_school on pop_census on both sides of the threshold and then predict the values of the dependent variable. Finally you can add to the plot the predicted y values adding "line" to the twoway command. Since we are dealing with a LLR, you should perform this regression on a smaller window around the threshold; try with different bandwidths of inhabitants to the left and right of the threshold and choose the window you think is more appropriate. You can try with all the sample, half of it (N=1100) and one fourth (N=550), just present one. What are your findings? Do the weighted and local linear regression display a jump at the threshold?

7. Provide evidence that the regression discontinuity design is valid in this case. Recall that the assumption behind RDD is that on either side of the threshold, everything else should be very similar. That is, only one thing should change sharply at the discontinuity, but nothing else should. Any other change should be gradual. You may do this with graphs or tables. First check for individual characteristic such as gender and age. We are comparing municipalities close to the threshold and not individuals. We are looking for a council effect of the wage increase, so that we should find the effect of a wage increase on the education level of all the members of a town council. However we are using major's education only for simplicity (and because taking into account council composition results do not change very much). For this reason you should also check for municipalities characteristics: area extension (extension), altitude (alt_center) and north south imbalances (NORTH). To do so note that municipality features are unique for each city, so that you have to leave only one observation for each city in the dataset when dealing with extension, alt_center, NORTH. To help you with this you can sort the dataset using the variable

id_city. Then one strategy is to enumerate observations of the same city and then leave for the analysis only one observation per municipality.

8. Perform the regression-version of the graph you presented in e). Write down the regression you estimate and the coefficient that represents the incumbency advantage effect. Try first to redo a simple LLR with different bandwidths as you did with the graph and then you can add covariates. Also try adding polynomials and interactions and see what you get. Do you have any insight from the polynomial regression about the bandwidth?

9. Estimating the treatment effect using Cattaneo et al., IK and Nichols commands. These methods typically do not work well with population thresholds because their algorithms choose a too small bandwidth compared to the size of the population. What are your findings? What could be the problem of the algorithms in your opinion? Which one works, which does not? Why does this happen in your opinion?

2 Problem Set 4. Panel Data

We estimate the effect of having a bank account for beneficiaries of social programs in Mexico. The original dataset was a panel for 5,768 households in 25 of the 32 states of Mexico. The survey is representative at the national level, with a total of four rounds from 2004 to 2007 (with questions referring to the previous year 2003 to 2006).

Among sampled households there was a group that participated in social programs such as Oportunidades (Progresa), which had accounts opened by the Federal Government in order to deposit the transfer. The authorities of the social programs claimed that they opened accounts to all beneficiaries that were living in a locality that was relatively close to a branch of the National Public Bank, the remaining beneficiaries kept on receiving the benefits in cash. The variable "treated" indicates if the household had an account opened by the program. It is equal to 1 from the first round the account was opened.

Out of the 5,768 households that were surveyed in the first round (March, 2004), 2,731 reported receiving benefits from a social program of the Federal Government in at least one of the four rounds of the survey. Since we use the variation generated by the electronic payments program for social benefits, we restrict the sample to this study group. Moreover, for our control group we only keep households that did not have and an account until 2003 (but they could have opened an account by themselves from the second round of the survey). Among treated households, we only consider as eligible those households that report having an account opened between 2004 and 2007. Therefore, we drop observations of households receiving social benefits in accounts that were opened before 2004, which reduces the sample to 1,458 households (this is the final sample posted on e-learning). The first round of the survey can be thought as a baseline for the two groups: no one had an account before the second round, and the program opened accounts for the treated group from the second round.

1. Present a "Balance Table" testing for significant average differences on observable variables at baseline for those who were ever treated vs. those who were never treated (Hint: use the variable "ever_treated"). Include both household characteristics (e.g. income from social programs, gender, region, age, occupation, household conditions) and locality characteristics (e.g. all variables beginning with "loc_" in the database represent characteristics of the locality in which the respondent was living at baseline. These locality characteristics were only measured once in the year 2,000 since they come from the Census).

2. We focus on only one outcome variable (you have more in the database if interested): Total_savings. This measures total savings in Mexican Pesos, constructed as sum of reported current savings in ROSCAs, at home, bank accounts, in kind and with friends. Also, conduct the analysis only for the sample with numrounds==4: those who answered in the four rounds (but later on we will check whether dropping the others generates bias).

   (a) For the following points you might find it useful to re-shape the data so that you have one line per household instead of several lines per household (which will be useful when you do the fixed-effects regressions). You can use something like this: reshape wide total_savigns , i(id) j(round) Note: if you tset the date, do not include j(round) option.

   (b) The outcome variable is affected by outliers, I suggest to run all the analysis using the winsorized values of ths variable (winsorized at the top 95%, that is, replace the top 5% values by the 95th percentile). You can do it in this way

      i. gen total_savingsw=total_savings
      ii. sum total_savings, det
      iii. replace total_savingsw=r(p95) if total_savings>r(p95) & total_savings<. And from now on, use the variable total_savingsw.

3. Calculate the simple difference in differences estimate by comparing the mean of total savings for treated and control individuals, before (round1) and after (average round2-round4) the program. Present the estimate and conventional standard errors. (Hint: use the command "ttest, by (ever_treated)" for the differences in each outcome before and after).

4. Run a regression for the difference in each outcome (average(Y2-Y4) – Y1) on treatment with conventional, robust and clustered (by id) standard errors. Do you find the same result as in point 3? After this go back to the original dataset (before reshaping).

5. Obtain an extended DID estimator by running a pooled OLS regression for total savings on ever_treated, one dummy equal to 1 if round 2, 3 or 4, and 0 if round 1 (create only one dummy) and the interaction of

ever_treated and this dummy. Compare your results with those in point 3.

6. Replicate the regression in 4 for total savings, by using robust standard errors, hc2 robust standard errors, hc3 robust standard errors, clustered standard errors at the household level (id) and block bootstrap (with cluster standard errors and 1000 replications). Do the results vary with the method? Given the current application, which standard errors would you choose to report if you can only report one?

7. Instead of doing pooled OLS, run a fixed effects regression with dummies for each round (2 to 4), using clustered standard errors. What do you find? [Hint: use "xtreg..., fe i(id) j(round)" command in STATA]. Present two set of standard errors: a) clustered at the household (id) level , b) clustered at the locality level (locality). [Hint 2: for this regression use the variable "treated" in the dataset, which takes the value 1 after respondent is "treated" and 0 before that]

8. What are the assumptions that you need for the fixed effects strategy to give you consistent estimates of the treatment effect? Do these assumptions make sense given the context we are studying?

9. We can weaken the assumptions if we include time-varying controls in the regressions. Repeat the regression in 6 including time-varying controls: income from social programs, log of annual expenditures, remittances and employed.

10. Would it make sense to include locality trends as control variables? We only have information for the year 2000. Let's consider, for example, the marginality index for each locality (loc_margindex). Repeat the regression in 9 including trend_margloc as a control. Do results change? How does this weaken the assumptions? We can create a linear trend for each locality in this way:

    (a) gen trend_margloc=loc_margindex if round==1

    (b) bysort id (round): replace trend_margloc = trend_margloc[_n-1]*2 if round==2

    (c) bysort id (round): replace trend_margloc = trend_margloc[_n-2]*3 if round==3

    (d) bysort id (round): replace trend_margloc = trend_margloc[_n-3]*4 if round==4

11. Attrition. For this point you need to use all the observations (not only those with numrounds==4). Using the variable "numrounds", create a variable for being an attriter (defined as not answering for at least one of the rounds of the survey). Is attrition correlated with treatment? Are

attritors different from non-attritors? Do a balance table, and run a regression interacting baseline covariates with treatment to see if they jointly predict attrition. Or interact the attrition dummy with baseline covariates and see if it predicts a baseline otcome. What types of bias can this generate if we just drop attritors and ignore the attrition problem as we did in all the previous points of the problem set.

3 CONGRATS!!! YOU HAVE FINISHED ALL THE (SEEMINGLY INFINITE) PROBLEM SETS OF MICROECONOMETRICS. ENJOY YOUR VACATIONS!!!