STATA Exercise 3: Regression Discontinuity Design and the Effect of Social Assistance Benefits on Employment

April 7, 2019

In this exercise, we will analyze the effect of the benefit level on employment. Between 1969 and 1989, the social assistance benefits in the Canadian province of Quebec was much lower for recipients without children aged below 30 years. Before the benefit levels were equalized for recipients (without children) younger and older than 30 years in 1989, the benefit level was \$185 per month for individuals aged below 30 years, whereas the benefit level was \$507 for individuals aged above.[1] To investigate this question, we use data from the 1986 and 1991 Canadian Censuses. However, we do not have access to individual data, but can only work with data averaged on years of age for respectively Quebec and the rest of Canada in 1986 and 1991. Since only the ages $20 - 39$ are relevant for the policy, the data set contains $20 * 2 * 2 = 80$ observations. For each cell we have the number of observations (*nobs*), but unfortunately not the standard errors on the mean employment (*empc*) for each cell. Therefore, we weight each regression with the number of cell observations. The data set is called *LemieuxMilligan.dta*.

NB! For the first nine questions you should only use observations for Quebec (*quebec=1*) for 1986 (*year=1986*).

1. Use STATA's `twoway` command to make a `scatter` plot between *empc* and *age*. Use the option `xline(30)` to obtain a vertical line at the age of 30. Remember to only use observations for Quebec for 1986.[2]

2. Estimate the simplest linear regression discontinuity model

$$Empc_i = \beta_0 + \beta_1 age_i + \beta_3 treat_i + u_i \tag{1}$$

[1] The policy, we consider, has been considered in Fortin, B., G. Lacroix and S. Drolet (2004), "Welfare Benefits and the Duration of Welfare Spells: Evidence from a Natural Experiment in Canada", *Journal of Public Economics*, Vol. 88, pp. 1495–1520, and Lemieux, T. and K. Milligan (2008), "Incentive Effects of Social Assistance: A Regression Discontinuity Approach", *Journal of Econometrics*, Vol. 142, pp. 807-828.

[2] We could have used `binscatter` for this exercise, but due to the discreteness of the data, we will instead use `twoway scatter`.

where $treat_i = \begin{cases} 0 & \text{if } age_i < 30 \\ 1 & \text{if } age_i \geq 30 \end{cases}$. Due to the non-linearity of the running variable, we will only use observations for age 25 and above. Remember to use the number of individuals at each age (*nobs*) in the weight option. How should we interpret the estimate?

3. Extend the simple linear model in equation (1) by adding a) a quadratic function of the running variable and b) quadratic and cubic functions of the running variable.

4. Estimate the simple linear model in equation (1) as well as the quadratic and cubic versions using a donut regression discontinuity. Irrespective of whether you find different estimates using donut regression discontinuity, continue in the subsequent questions without using the donut RD.

5. On the basis of the plot above, we would like to take account of the slopes being different to the left and right of the cut-off. Therefore, estimate the following regression

$$Empc_i = \beta_0 + \beta_1 age_i + \beta_2 age_i * treat_i + \beta_3 treat_i + u_i$$

Give an graphical argument for what goes wrong with this formulation.

6. Rather than using the wrong specification above, estimate the following regression

$$Empc_i = \beta_0 + \beta_1 (age_i - 30) + \beta_2 (age_i - 30) * treat_i + \beta_3 treat_i + u_i \quad (2)$$

7. Use STATA's twoway command to make a scatter plot between *empc* and *age*, where you insert a vertical xline at age 30 as well as separate regression lines to the left and right of the cutoff. [Hint: Use STATA's post estimation command, `predict`, to predict $\widehat{Empc}$ and use this new variable to construct the regression lines].

8. Extend the linear model in equation (2) by adding a quadratic function of the running variable. Estimate this model for i) individuals aged 25-39 in Quebec in 1986 and ii) individuals aged 20-39 in Quebec in 1986. Again, construct a scatter plot between *empc* and *age*, where you insert a vertical `xline` at age 30 as well as separate regression lines to the left and right of the cutoff. Comment on your results.

9. We would like to show that there is no manipulation of the running variable, age. Explain what data you ideally would like to use for the test. In absence of better data, use the number of observations (*nobs*) and construct a graph, which can show us whether manipulation of the running variable seems to be a concern.

10. For each year between 27 and 33 use placebo cutoffs for the simple linear model in equation (1) to test whether other thresholds between ages of 27 and 33 are significant.

11. Use the simple linear model in equation (1) and construct falsification tests using that there is no discontinuity in the age for Quebec in 1991 and for the rest of Canada in both 1986 and 1991.

12. How could we estimate the effect of receiving lower benefit level using another estimation strategy than regression discontinuity design? Do you obtain similar results. What are the identifying assumptions here?