Lab Exercise: Discontinuities

This week we conduct the same simulation as usual, but this time using a regression discontinuity to estimate the effect.

1. First, let's generate an 'income' variable for 50,000 people. The data should be drawn randomly from the normal distribution with mean 500 and standard deviation 50.
2. Now let's simulate potential outcomes (let's say the outcome is 'attitude to redistribution') for each person that depends on their income. Assume:

$$y_0 = N(10, 2) + \frac{income}{100} + (\frac{income}{500})^2$$

$$y_1 = y_0 + 2$$

So there is a constant treatment effect of 2.

3. Actual treatment assignment is not random but deterministic: Imagine it is a poverty-relief program that you receive only if your income is less than 500.
4. Now calculate the observed outcome based on the potential outcomes and actual treatment status.
5. Now let's calculate the 'naive' Average Treatment Effect by running a simple OLS regression of the observed outcomes on treatment. What is your estimate of the average treatment effect? How does this compare to the treatment effect we specified earlier?
6. Now apply the 'full bandwidth' regression discontinuity method. This just means controlling for the running variable, which in this case is income. Interpret the results. How do they compare to the treatment effect we specified?
7. Let's try and make a regression discontinuity plot to understand what's going on better. First, plot all the data, with the running variable (which one is that?) on the x-axis against the observable outcome variable on the y-axis. What can you see in the graph?
8. The graph in Q7 is difficult to interpret so most regression discontinuity plots 'bin' the data into groups to more easily see the pattern. Bin the income data into 20 groups and plot this against the average observed outcome in each bin.
9. An easy way to make a nice regression discontinuity plot is to use the *rdplot* command in the *rdrobust* package.
10. An alternative way of estimating the treatment effect is to perform a simple difference-in-means between the values just to the left and just to the right of the cutoff. Using the data between 480 and 520 on the income scale, perform a difference-in-means test for the efffect of treatment on the outcome. How does this compare to your regression discontinuity estimate?
11. A third method of performing a regression discontinuity is to perform the regression approach only on a narrow 'bandwidth' close to the cutoff. Subset the data again to between 480 and 520 on the income scale and perform a regression discontinuity analysis. How do the results compare to your full-bandwidth regression discontinuity in Q6 and the difference-in-means estimate in Q10?
12. One way of easily picking an 'optimal' bandwidth is to use the automatic process in the *rdrobust* command of the *rdrobust* package. This method also provides more accurate standard errors. Apply the method and interpret the results. (You can choose whether to use a linear or quadratic regression).

13. Recall that in our original specification of the potential outcomes we made them depend on $income^2$. Do we get a better estimate of the treatment effect if we include a quadratic term in our optimal-bandwidth regression discontinuity?

14. Just to check our assumptions: We know there is no sorting around the cutoff in our model because we specified the treatment to be precisely based on the income cutoff. But anyway let's run the standard test for sorting - the McCrary density test using the *rddensity* package. Also make a nice graph with the *rdplotdensity* command.

15. Now let's change how our potential outcomes are defined. Our y_0 stays the same, but this time let's abandon our constant treatment effect and assume the treatment effect itself varies depending on income:

$$y_0 = N(10, 2) + \frac{income}{100} + (\frac{income}{500})^2$$

For those with income between 490 and 510, the treatment effect is:

$$y_1 = y_0 + 10$$

For everyone else, the treatment effect is:

$$y_1 = y_0 + 3$$

16. Calculate the observed outcomes again and run the optimal-bandwidth regression discontinuity analysis using *rdrobust*. How do you interpret the results? What treatment effect is estimated here?