POLS0012 Causal Analysis: Tutorial Exercise 8

For this tutorial, we will analyze Erik Meyersson's 2014 paper, "Islamic Rule and the Empowerment of the Poor and Pious." He asks whether political control by Islamic religious political parties leads to a decrease in women's rights, particularly female education rates. This is difficult to answer using a standard regression, due to selection bias: places that are already favor more traditional gender roles may also be more likely to elect Islamic parties. He looks at Turkey in 1994, where an Islamic party gained political control in many municipalities, and a number of the election results were very close. He uses a regression discontinuity analysis of Islamic control on the rate of secondary school completion by girls, focusing on the Local Average Treatment Effect in areas where Islamic parties barely won or lost their elections.

You'll need to install the `rdd` package for this tutorial. The dataset is called "islamic.csv" and contains the following variables:

- *iwm94*: margin of Islamic party win or loss in 1994, pp: 0 indicates an exact tie
- *hischshr1520f*: secondary school completion rates for ages 15-20 females
- *lpop1994*: log of the locality population in 1994
- *sexr*: sex ratio in locality
- *lareapre*: log of locality area

a) Create a treatment variable, `islamicwin`, that indicates whether or not the Islamic party won the 1994 election.
Code Hint: A margin of greater than zero means the Islamic party won

b) Calculate the difference in means in secondary school completion rates for women between regions where Islamic parties won and lost in 1994. Do you think this is a credible estimate of the causal effect of Islamic party control? Why or why not?
Code hint: use the option `na.rm=T` so that R knows to ignore missing data

c) Now we'll start regression discontinuity analysis. First, select optimal bandwidths for testing female high school completion rates using the Imbens-Kalyanaram procedure. Explain what the bandwidth means in this case
Code Hints: Use the `IKbandwidth` function from the `rdd` package. Read the help file to see what the function requires

d) Use your answer to (c) to create a new dataset containing only data within the optimal bandwidth

e) Find the Local Average Treatment Effect of Islamic party control on women's secondary school education at the threshold, using the dataset you created in (d) and a simple linear regression that includes the treatment and running variable. How credible do you think this result is?

f) Use RD estimation to find the Local Average Treatment Effect of Islamic party control on men's and women's secondary school education at the threshold, using local linear regression estimated with the `RDestimate` function from the `rdd` package. Does the estimate differ from part (e)?
Code Hints: Your code should be of the form `RDestimate(outcome~running.variable,`bw = bandwidth, data=)

g) Plot the relationship between the running variable and outcome using local linear regressions. Use your plot to explain why your results do or do not differ strongly between (e) and (f)
Code Hints: You can just use the `plot()` function directly on the object you created in (f). Use the `range` argument to control the x axis

h) Perform placebo tests to check that the relationship between the running variable and outcome is not fundamentally discontinuous, by estimating RD estimates at placebo cutoffs of -0.1, -0.05, 0.05 and 0.1. What do you conclude?

i) Perform a robustness check for local randomisation at the threshold by estimating RD estimates in the same way as (g) for the three background covariates *sexr*, *lop1994* and *lareapre*. What do you conclude?

j) Perform a McCrary test: another way to check for sorting at the theshold. Plot and interpret the results.
Code Hints: Use the `DCdensity` function in the `rdd` package with the option `verbose=TRUE`

k) **[Challenging Question]** Examine the sensitivity of the main RD result to the choice of bandwidth by calculating and plotting RD estimates and their associated 95% confidence intervals for a range of bandwidths from 0.05 to 0.6. To what extent do the results depend on the choice of bandwidth?
Hints: Begin by creating a vector of thresholds such as `thresholds <- seq(from=0.05,to=0.6, by=0.005)`. Then use a for loop. You can extract the estimate and standard error from an RD estimate named `rdest` with the code `rdest$est[1]` and `rdest$se[1]`