Regression II
Vanderbilt University
Prof. Sean P. Corcoran

Due December 5, 2023
60 total points

Problem Set 7

Instructions: Answer the following questions in a Stata do-file. Submit your problem set as do-file and/or a PDF via email to `sean.corcoran@vanderbilt.edu`. Use your last name and problem set number as the filename. Working together is encouraged, but all submitted work should be that of the individual student.

Question 1. This problem will replicate some of the results in Lee (2008), one of the most influential studies using regression discontinuity. Lee analyzed 50 years of election results for the U.S. House of Representatives to determine whether incumbent parties are more likely to win elections. The theory is that, once in office, the party can garner resources to give them an edge in the next election. This question is a difficult one to answer causally, since all is not typically held equal. Incumbents arguably have qualities that made them appealing candidates to begin with, and these qualities make them more likely to win in subsequent elections. To address this, Lee proposed comparing elections in which a party won (vs. lost) by a small margin. Due to idiosyncratic variation in turnout and other factors, some parties are likely to win (or lose) simply due to chance (ask Al Gore). As long as party preferences are continuous through the threshold for victory (e.g., 50%) one wouldn't expect a candidate who barely won an election to have an edge in the next election, unless there is an incumbency advantage.

For this problem you will need the dataset on Github called *Lee_2008_for_RD.dta*. Each observation is a Congressional district election between 1948 and 1998. Due to redistricting every 10 years, elections are not included if their boundaries changed from election $t-1$ to t or from t to $t+1$. The running variable is *difdemshare*, the difference between the Democratic candidate's vote share and the largest vote share of the other parties. If the Democrat won, *difdemshare* is greater than zero. Here we will focus on Democratic incumbents, but the results would look the same if we flipped things around and looked at Republican incumbents instead.

Conduct a regression discontinuity analysis to estimate the effect of Democratic incumbency in year t on two outcomes: *difdemsharenext*, the difference between the Democratic vote share and the largest vote share of the other parties in the next election (year $t+1$), and *demwinnext*, a binary variable equal to 1 if a Democrat won the next election and 0 otherwise. Your analysis should include the following elements: (**55 points**)

(a) Write down the assumptions that should hold in order for your RD estimate to be considered the causal effect of incumbency. (**5 points**)

(b) A scatterplot and RD plot showing the relationship between *demsharenext* and the running variable across the full range of data. (If it helps visually, you can also show the scatter and RD plots for observations closer to the cutpoint, e.g., $abs(difdemshare) < 0.25$). Is there visual evidence of a discontinuity? (**5 points**)

(c) Parametric RD models using OLS for each outcome assuming a linear relationship with the running variable, then a quadratic ($p = 2$), and then a quartic ($p = 4$). In each case allow the slope coefficients to differ on each side of the cutoff. Repeat these models but include two covariates in the regression: *demofficeexp* and *othofficeexp* (measures of the Democrat's and opposition's experience in office). You may want to collect your regression results into one or more tables for easy comparison. (There will be a total of 12 regressions). What do these regressions show? Do the differing polynomial orders lead to different conclusions? (**10 points**)

(d) Non-parametric RD models for each outcome using a local linear regression, the MSE-optimal bandwidth, and triangular kernel. Repeat, including the two covariates listed in part (c). (There should be 4 regressions for this part). What do these regressions show? Do the conclusions differ from part (c)? (**10 points**)

(e) An RD plot for each outcome based on the optimal bandwidths used in part (d) for the models without covariates. (**5 points**)

(f) A density test for manipulation around the cutoff. Provide the density plot and report the p-value of the test (and conclusion). Is manipulation theoretically plausible in this case? Why or why not? (**5 points**)

(g) As a validity check, repeat part (d)—without covariates—in which you use *demshareprev* and *demwinprev* as the outcome variables. (These represent the Democratic vote share and a Democratic win in the *previous* election, $t - 1$). What does this accomplish and what do you find? (**5 points**)

(h) One would not expect there to be a discontinuity in the covariates used in parts (c)-(d) at the cutpoint. Repeat part (d)—without covariates—in which you use *demofficeexp* and *othofficeexp* as the outcome variables. What do you find? (**5 points**)

(i) Finally, conduct some tests for discontinuities elsewhere the distribution of *difdemsharenext*. I suggest looking at "fake" cutpoints equal to the 1st, 2nd, 3rd, 4th, etc., deciles of the *difdemsharenext* distribution. Since there is a known "real" cutpoint at 0, limit these analyses to either values below 0 (Republican win) or above 0 (Democratic win), depending on where your "fake" cutpoint sits. Summarize what you find. (**5 points**)

Question 2. Consider the sharp RD model in which the running variable (x_i) is allowed to have a linear relationship with the outcome (Y_i) that varies on either side of the cutoff (c). Let the treatment status variable $D_i = 1$ whenever $x_i > c$.

$$Y_i = \pi_0 + \pi_1 x_i + \pi_2 D_i + \pi_3 (D_i \times x_i) + v_i$$

Suppose that the running variable x_i is *not* centered at c. (That is, we do not first subtract off c from x_i). Show that π_2 in this case is *not* the impact of the treatment at the threshold c. You can show this however you like: algebraically, using the simulated data, as in the in-class exercise, or any other valid method. **(5 points)**