Regression II
Vanderbilt University
Prof. Sean P. Corcoran

Due December 8, 2022 (11:59 pm)
56 total points

Problem Set 6

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. (**50 points**)

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* will be 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.

(a) Conduct a regression discontinuity analysis that includes the following elements. Before doing so, write down the assumptions that need to hold for an RD to produce the causal effect of incumbency. (**40 points**)

- Two main outcome variables: *difdemsharenext*, the difference between the Democratic vote share and the largest vote share of the other parties in the next election, and *demwinnext*, an indicator variable equal to one if a Democrat won the next election.

- Scatterplot and `binscatter` showing the relationship between *demsharenext* and the running variable. Hint: it may help visually to focus on observations in which $abs(difdemshare) < 0.25$, and to increase the number of bins in `binscatter`.

- Parametric RD models assuming a linear relationship with the running variable, then a quadratic, then a quartic (i.e., up to the fourth power). In all three cases allow the relationship to differ on each side of the cutoff, and allow for clustering in the standard error calculation (using the Congressional district id as the clustering variable). Repeat the same models but include covariate controls: *demofficeexp* and *othofficeexp* (measures of the Democrat's and opposition's experience in office).

- Non-parametric RD estimates using `rd`, using the default bandwidth. Again use the clustered standard errors.

- For your quartic model, create a scatterplot that includes the fitted model on each side of the cutofff. (Do this only for the continuous outcome *difdemsharenext*).

- A histogram for the running variable and a McCrary test to look for manipulation at the cutpoint (Russian hackers?). Here again, for your histogram it may help to focus on observations where $abs(difdemshare) < 0.25$.

- A validity check in which you use *demshareprev* and *demwinprev* as the outcome variables. What does this accomplish?

Write up your findings, interpreting and comparing your point estimates across the different models.

(b) Test for continuity in the relationship between *difdemsharenext* and the running variable by creating 9 dummy variables equal to one if x (the running variable) is greater than the 1st decile of x, greater than the 2nd decile of x, and so on. Then, estimate an OLS regression of *difdemshare* with a quartic in x and these nine dummy variables. (Also include the original indicator of a Democratic win, since we know there is a discontinuity there). Conduct a joint F-test for the significance of these nine dummies, and interpret. (**5 points**)

(c) Let's demonstrate to ourselves what `rd` is doing behind the scenes. First, use `rd` to get a non-parametric estimate of the effect of incumbency on *difdemsharenext*. Specifically set the bandwidth to be 0.275. Note the point estimate. Then try the following syntax. (**5 points**)

```
lpoly difdemsharenext difdemshare if difdemshare < 0, deg(1) ker(tri) bwidth(0.275) ///
     gen(L) at(difdemshare) graph
```

```
lpoly difdemsharenext difdemshare if difdemshare >= 0, deg(1) ker(tri) bwidth(0.275) ///
     gen(R) at(difdemshare) graph
gen diff = R - L
sum diff if difdemshare==0
drop R L diff
```

`lpoly` with `degree(1)` fits a local linear regression in the specified bandwidth, using the specified kernel weights (here, triangle). Here you are doing this on both the left and right-hand side of the cutpoint. `lpoly` gives you predicted values from this fitted model, which you are storing here and then calculating the gap at the cutpoint. Compare what you find here to your `rd` point estimate.

Finally, try the syntax below. How does the OLS point estimate below compare to what you found using `rd` and `lpoly`?

```
gen kwt=max(0,0.275-abs(difdemshare))
gen win=difdemshare>0
/* see the triangle kernel */
scatter kwt difdemshare if abs(difdemshare)<=0.275
reg difdemsharenext difdemsahre win [pw=kwt]
```

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 from the in-class exercise, or any other valid method. **(6 points)**