Econ 771 – Module 3 Empirical Exercise

November 4

Name: Amy Lim

1 Overview

In this assignment, we're going to work through some applied issues related to regression discontinuity designs. We'll cover the basics of strict and fuzzy RD, and we'll work through standard specification tests. We'll also introduce some more technical aspects of bin and bandwidth selection.

2 Data

The data for this assignment comes from the AEJ: Policy website, where Keith Ericson's complete dataset is available. For this assignment, I downloaded all of the data from our class OneDrive folder.

3 Data Cleaning

This assignment was a lot more straight forward given that the author provides Stata code for replication. I used a lot of it.

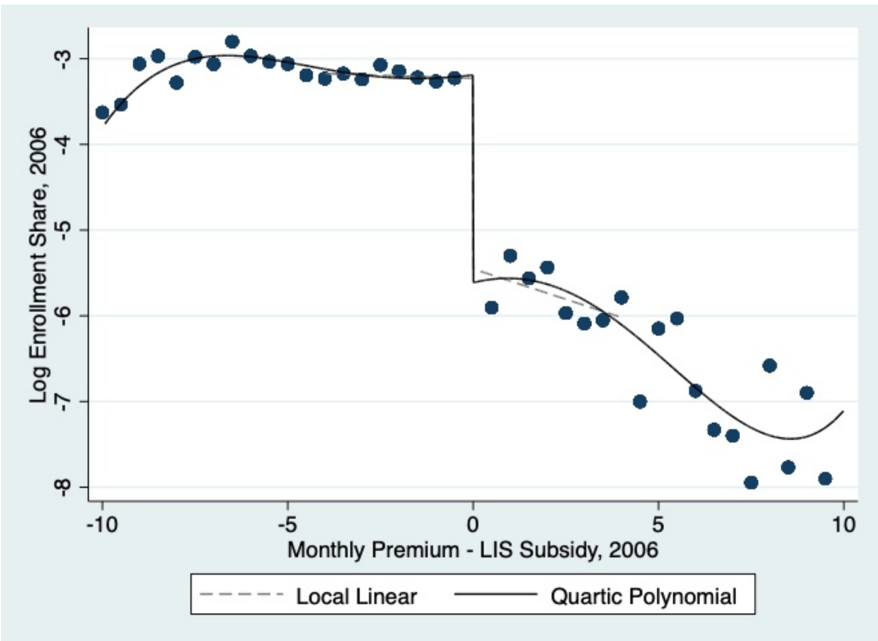
4 Questions

1. Recreate the table of descriptive statistics (Table 1) from Ericson (2014).

Table 1: Descriptive Statistics of Medicare Part D Plans

	2006	2007	2008	2009	2010
Mean monthly premium	\$37	\$40	\$36	\$30	\$33
	(13)	(17)	(20)	(5)	(9)
Mean deductible	\$92	\$114	\$146	\$253	\$118
	(116)	(128)	(125)	(102)	(139)
Fraction enhanced benefit	0.43	0.43	0.58	0.3	0.69
Fraction of plans offered by firms already offering a plan ..					
... in the United States	0.00	0.76	0.98	1.00	0.97
... in the same State	0.00	0.53	0.91	0.68	0.86
Number of unique firms	51	38	16	5	6
Number of plans	1,429	658	202	68	107

2. Recreate Figure 3 from Ericson (2014).



3. Calonico, Cattaneo, and Titiunik (2015) discuss the appropriate partition size for binned scatterplots such as that in Figure 3 of Ericson (2014). More formally, denote by $P_{-,n} = \{P_{-j} : j = 1, 2, \ldots J_{-,n}\}$ and $P_{+,n} = \{P_{+j} : j = 1, 2, \ldots J_{+,n}\}$the partitions of the support of the running variable x_i on the left and right (respectively) of the cutoff, $\bar{x}.P_{-,j}$ and $P_{+,j}$,n denote the actual supports for each j partition of size $P_{-,n}$ and $P_{+,n}$, such that $[x_l, \bar{x}) = \bigcup_{j=1}^{J_{-n}} P_{-j}$ and $(\bar{x}, x_u] = \bigcup_{j=1}^{J_{+n}} P_{+,j}$.

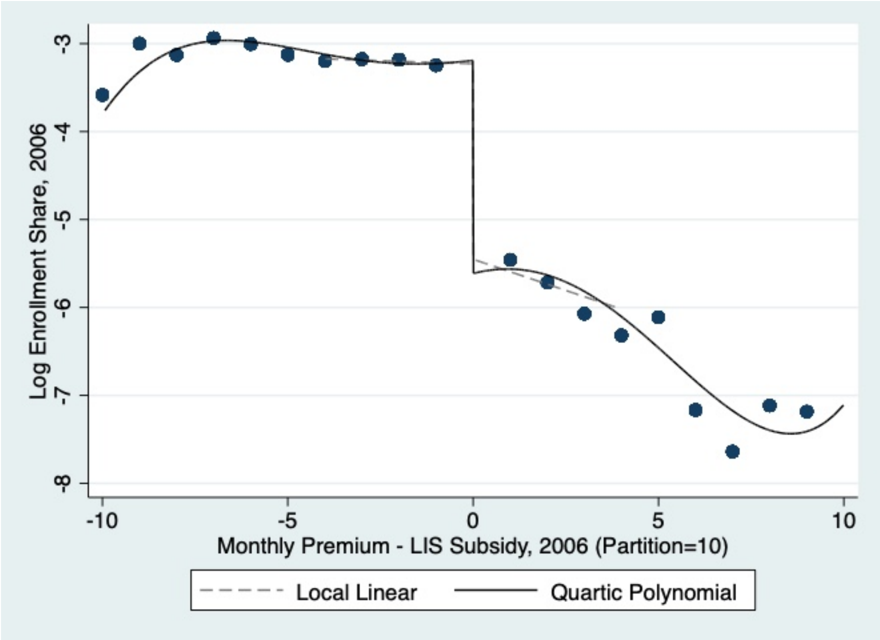
Individual bins are denoted by $P_{-,j}$ and $P_{+,j}$. With this notation in hand, we can write the partitions $P_{-,n}$ and $P_{+,n}$ with equally-spaced bins as:

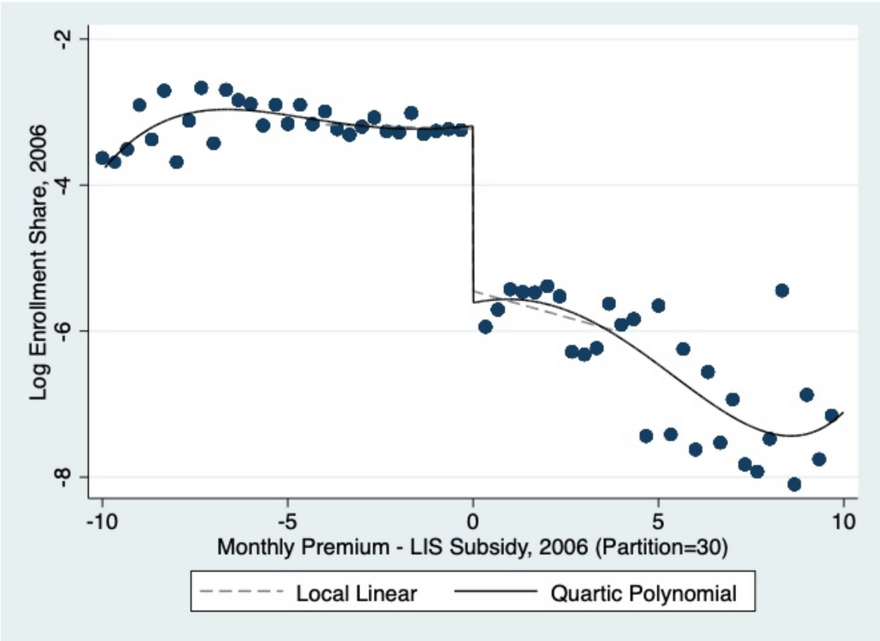
$$p_{-j} = x_l + j \times \frac{\bar{x} - x_l}{J_{-,n}}$$

and

$$p_{+j} = \bar{x} + j \times \frac{x_u - \bar{x}}{J_{+,n}}$$

Recreate Figure 3 from Ericson (2014) using $J_{-,n} = J_{+,n} = 10$ and $J_{-,n} = J_{+,n} = 30$. Discuss your results and compare them to your figure in Part 2.





4. With the notation above, Calonico, Cattaneo, and Titiunik (2015) derive the optimal number of partitions for an evenly-spaced (ES) RD plot. They show that:

$$J_{ES,-,n} = \left\lceil \frac{V_-}{\mathcal{V}_{ES,-}} \frac{n}{\log(n)^2} \right\rceil$$

and

$$J_{ES,+,n} = \left\lceil \frac{V_+}{\mathcal{V}_{ES,+}} \frac{n}{\log(n)^2} \right\rceil$$

where V_- and V_+ denote the sample variance of the subsamples to the left and right of the cutoff and $V_{ES,\cdot}$ is an integrated variance term derived in the paper. Use the rdrobust package in R (or Stata or Python) to find the optimal number of bins with an evenly-spaced binning strategy. Report this bin count and recreate your binned scatterplots from parts 2 and 3 based on the optimal bin number.

Table 2: Bin results from rdplot

	To the left of c	To the right of c
Bins selected	48	42
Optimal bins	9	10

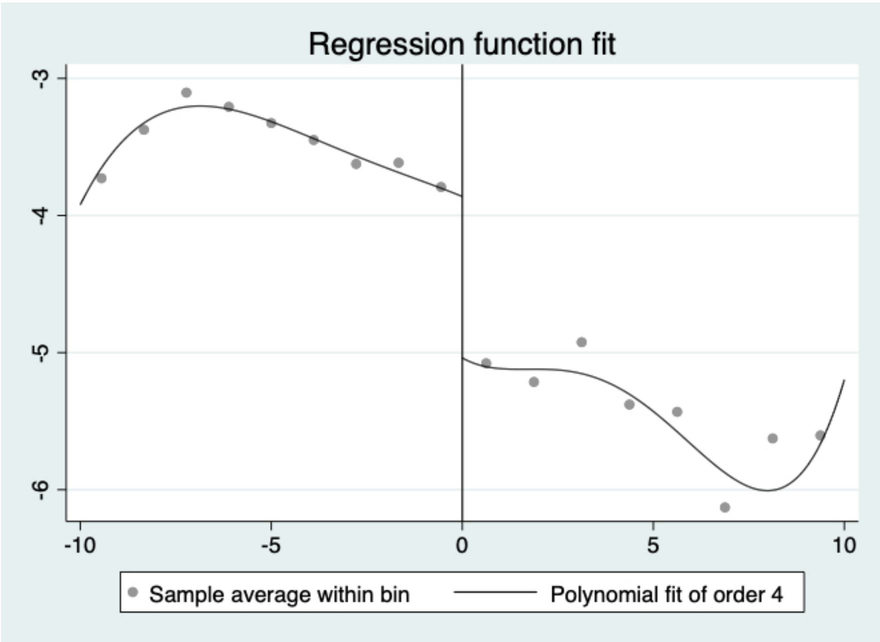


Figure 1: Replication of Figure 3 with the optimal bin count

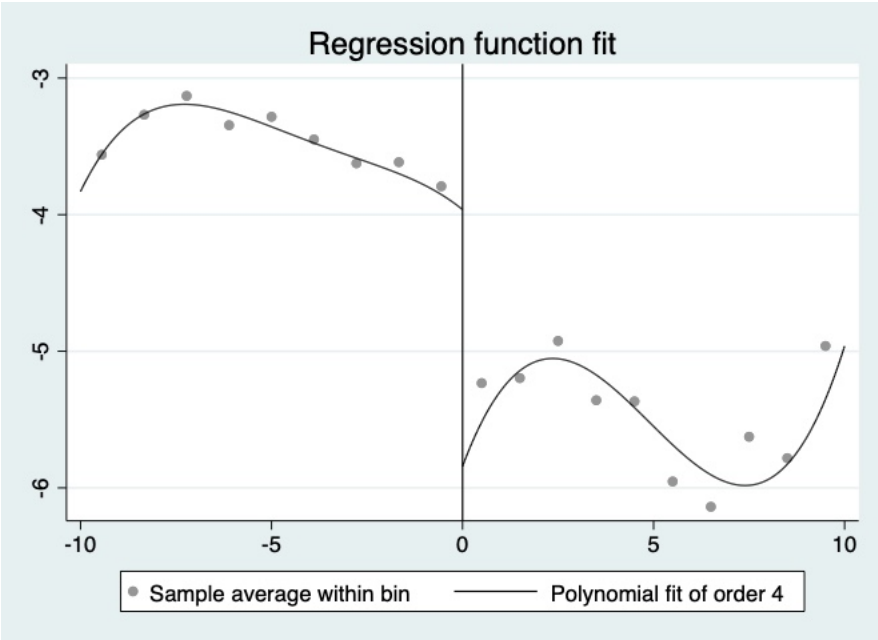


Figure 2: Replication of Figure 3 with the optimal bin count using $J_{-,n} = J_{+,n} = 10$

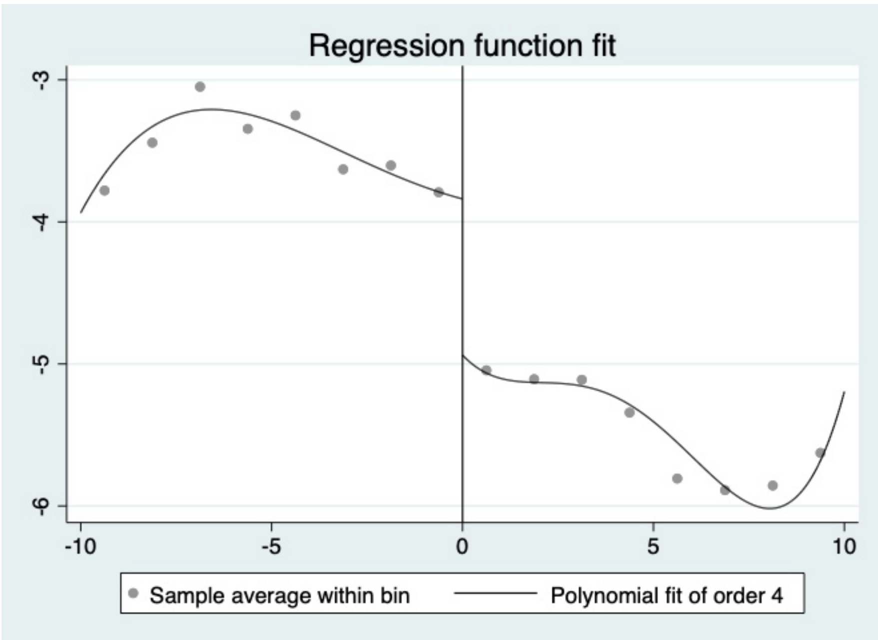


Figure 3: Replication of Figure 3 with the optimal bin count using $J_{-,n} = J_{+,n} = 30$

5. One key underlying assumption for RD design is that agents cannot precisely manipulate the running variable. While "precisely" is not very scientific, we can at least test for whether there appears to be a discrete jump in the running variable around the threshold. Evidence of such a jump may suggest that manipulation is present. Provide the results from the manipulation tests described in Cattaneo, Jansson, and Ma (2018). This test can be implemented with the rddensity package

in R, Stata, or Python.

Table 3: Manipulation test from rddensity

c=0	To the left of c	To the right of c
Number of observations	1,399	1,024
Eff. number of observations	1,399	1,024
Order estimate (p)	2	2
Order bias (q)	3	3
Bandwidth estimate (q)	10	10

N=2,423

Model= unrestricted

Bandwidth method= comb

Kernel= triangular

VCE method= jackknife

Table 4: Running variable: theBinAl

Method	T	$P > \|T\|$
Robust	-2.4466	0.0144

6. Recreate Table 3 of Ericson (2014) using the same bandwidth of $4.00.

Table 5: Effect of LIS Benchmark Status in 2006 on Plan Enrollment

ln s_t	2006	2007	2008	2009	2010
Panel A. Local linear, bandwidth $4 Below Benchmark 2006	2.224***	1.332***	0.902***	0.803**	0.677
	(0.283)	(0.267)	(0.248)	(0.362)	(0.481)
Below Benchmark	-0.0141	-0.0774	-0.0731	-0.170	-0.215**
	(0.0322)	(0.0882)	(0.116)	(0.105)	(0.0878)
Above benchmark	-0.142*	-0.0331	0.0494	0.0737	0.0488
	(0.0783)	(0.110)	(0.163)	(0.170)	(0.202)
Observations	306	299	298	246	212
R^2	0.576	0.325	0.131	0.141	0.124
Panel B. Polynomial, bandwidth $4 Below Benchmark 2006	2.464***	1.364***	0.872***	0.351**	-0.277
	(0.222)	(0.321)	(0.246)	(0.324)	(0.301)
Premium-Subsidy, 2006	Quadratic	Quadratic	Quadratic	Quadratic	Quadratic
Observations	306	299	298	246	212
R^2	0.794	0.576	0.472	0.535	0.685
Panel C. Past Interactions, local linear, bandwidth $4 Below benchmark or de minimis in: 2006 and current year	2.224***	2.089***	2.377***	2.633***	2.443***
	(0.283)	(0.364)	(0.275)	(0.257)	(0.309)
2006 but not current year		0.628**	0.892**	1.068**	0.967
		(0.293)	(0.329)	(0.446)	(0.625))
Current year but not 2006		0.148	1.356***	2.107***	2.281***
		(0.290)	(0.293)	(0.242)	(0.259)
Premium—subsidy, 2006	Linear	Linear	Linear	Linear	Linear
Observations	306	299	298	246	212
R^2	0.576	0.480	0.426	0.498	0.467

7. Calonico, Cattaneo, and Farrell (2020) show that pre-existing optimal bandwidth calculations (such as those used in Ericson (2014)) are invalid for appropriate

inference. They propose an alternative method to derive minimal coverage error (CE)-optimal bandwidths. Re-estimate your RD results using the CE-optimal bandwidth (rdrobust will do this for you) and compare the bandwidth and RD estimates to that in Table 3 of Ericson (2014).

Table 6: Sharp RD estimates using local polynomial regression (rdrobust)

c=0	To the left of c	To the right of c
Number of observations	1,399	1,024
Eff. number of observations	317	315
Order estimate (p)	1	1
Order bias (q)	2	2
Bandwidth estimate (h)	1.724	1.272
Bandwidth bias (b)	4.365	3.186
Bandwidth estimate (h)	0.395	0.399
Unique observations	20	20

N=2,421

Bandwith type= two different CER-optimal bandwidth

Kernel= triangular

VCE method= NN

Table 7: Running variable: theBinAl

Method	Coefficient
Conventional	-1.0492
	(0.38085)
Robust	-
	-

8. Now let's extend the analysis in Section V of Ericson (2014) using IV. Use the presence of Part D low-income subsidy as an IV for market share to examine the effect of market share in 2006 on future premium changes.

Table 8: regression table

	(1)
Presence of Part D low-income subsidy	-0.344
	(0.294)
theBinAl	-0.181***
	(0.0449)
Constant	-4.333***
	(0.0468)
Observations	2420
FE	
r2	0.387

Standard errors in parentheses

Dependent variable: Log Market Share

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

9. Discuss your findings and compare results from different binwidths and bandwidths. Compare your results in part 8 to the invest-then-harvest estimates from Table 4 in Ericson (2014). The results are only consistent for the second year of implementation.

10. Reflect on this assignment. What did you find most challenging? What did you find most surprising? This one was okay! I think the hardest was understanding Ericson's code. He makes a lot of variables manually with few loops. I learned a lot about how to add more notes to make my code easier to understand