Empirical Assignment: Regression Discontinuity

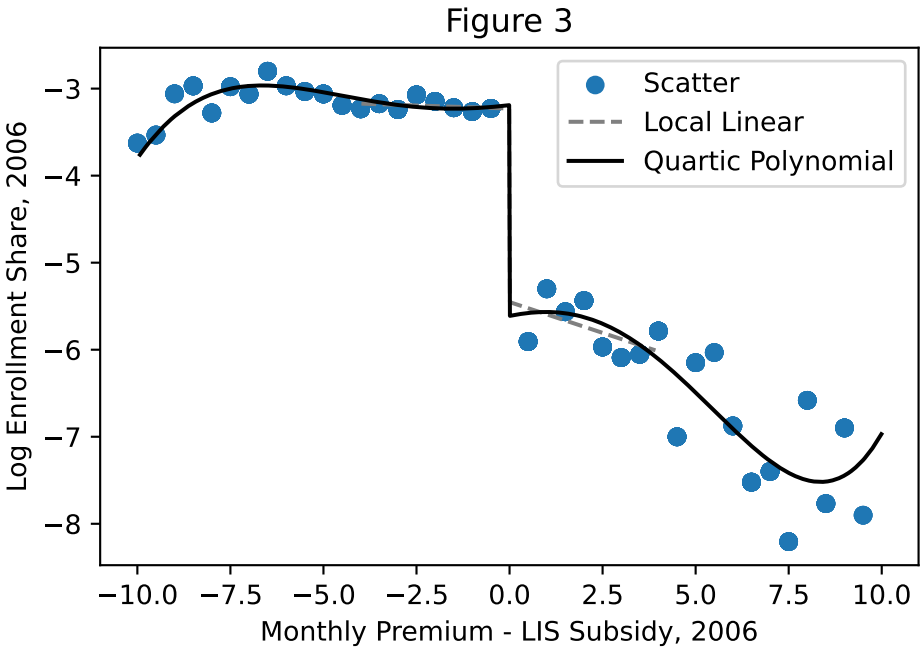
Michaela Philip

All data from Keith M. Marzilli Ericson (2014)

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

	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	0.03	0.69
Fraction of US firms	0.0	0.76	0.98	1.0	0.97
Fraction of state firms	0.0	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).



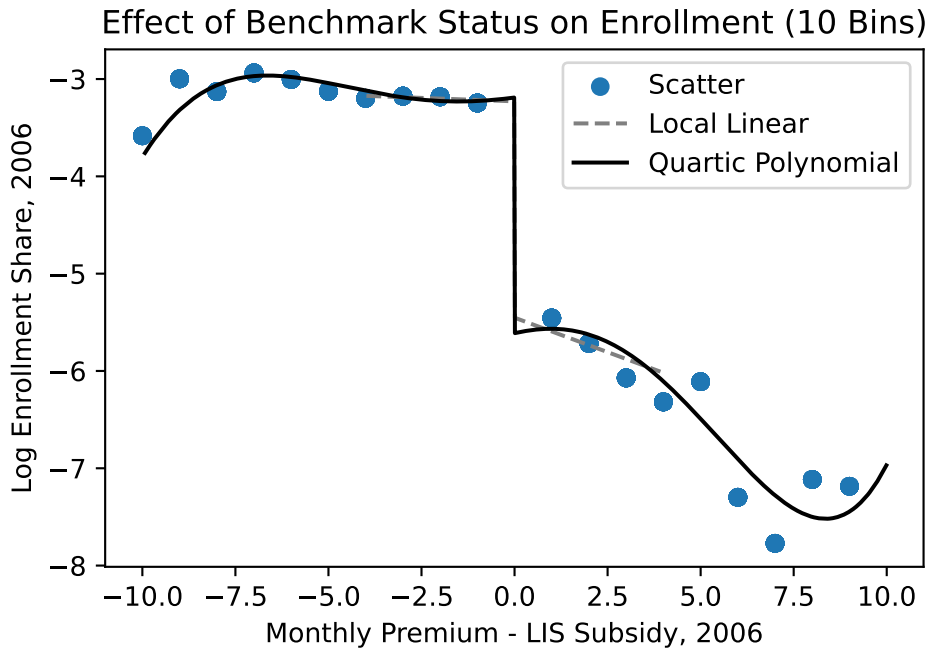
3. Calonico, Cattaneo, and Titiunik (2015) discuss the appropriate partition size for binned scatterplots such as that in Figure 3 of Ericson (2014). More formally, denote by $\mathbb{P}_{-,n} = \{P_{-,j} : j = 1, 2, ...J_{-,n}\}$ and $\mathbb{P}_{+,n} = \{P_{+,j} : j = 1, 2, ...J_{+,n}\}$ the partitions of the support of the running variable x_i on the left and right (respectively) of the cutoff, $\bar{x}$. $P_{-,j}$ and $P_{+,n}$ denote the actual supports for each j partition of size $J_{-,n}$ and $J_{+,n}$ such that $[x_l, \bar{x}) = \cup_{j=1}^{J_{-,n}} P_{-,j}$ and $(\bar{x}, x_u] = \cup_{j=1}^{J_{+,n}} P_{+,j}$. Individual bins are denoted by $p_{-,j}$ and $p_{+,j}$. With this notation in hand, we can write the partitions $J_{-,n}$ and $J_{+,n}$ with equally-spaced bins as

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

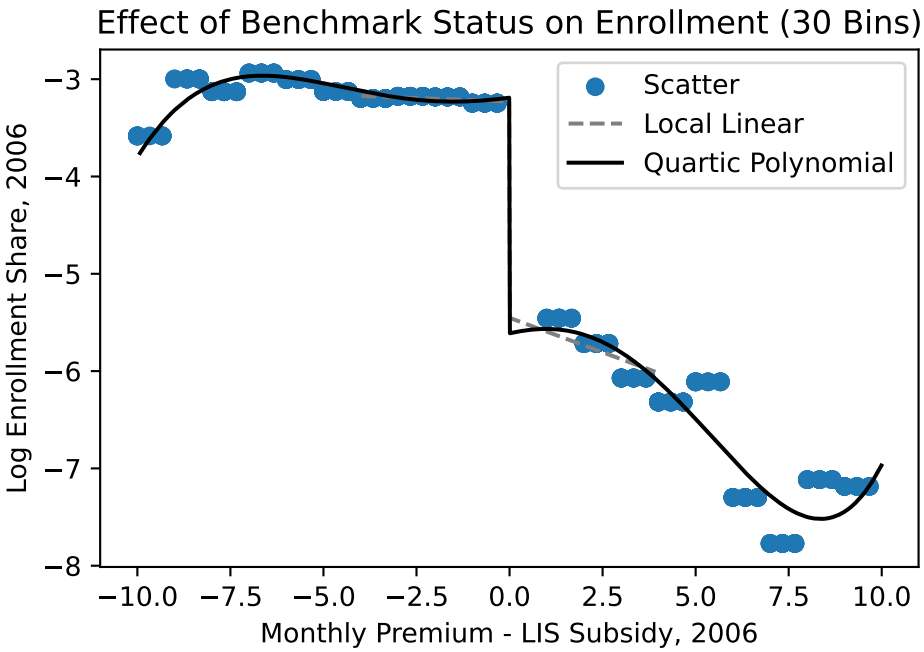
and

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

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.



Visually, using 10 bins seems to fit the data pretty well. The fit in the middle is very similar to Ericson's Figure 3, and although the two extremes aren't perfect they are not worse than Ericson's figure and may fit a bit better.



Using 30 bins for this graph seems to mostly have included a lot of noise - the fit is not significantly better but the graph looks much more busy.

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

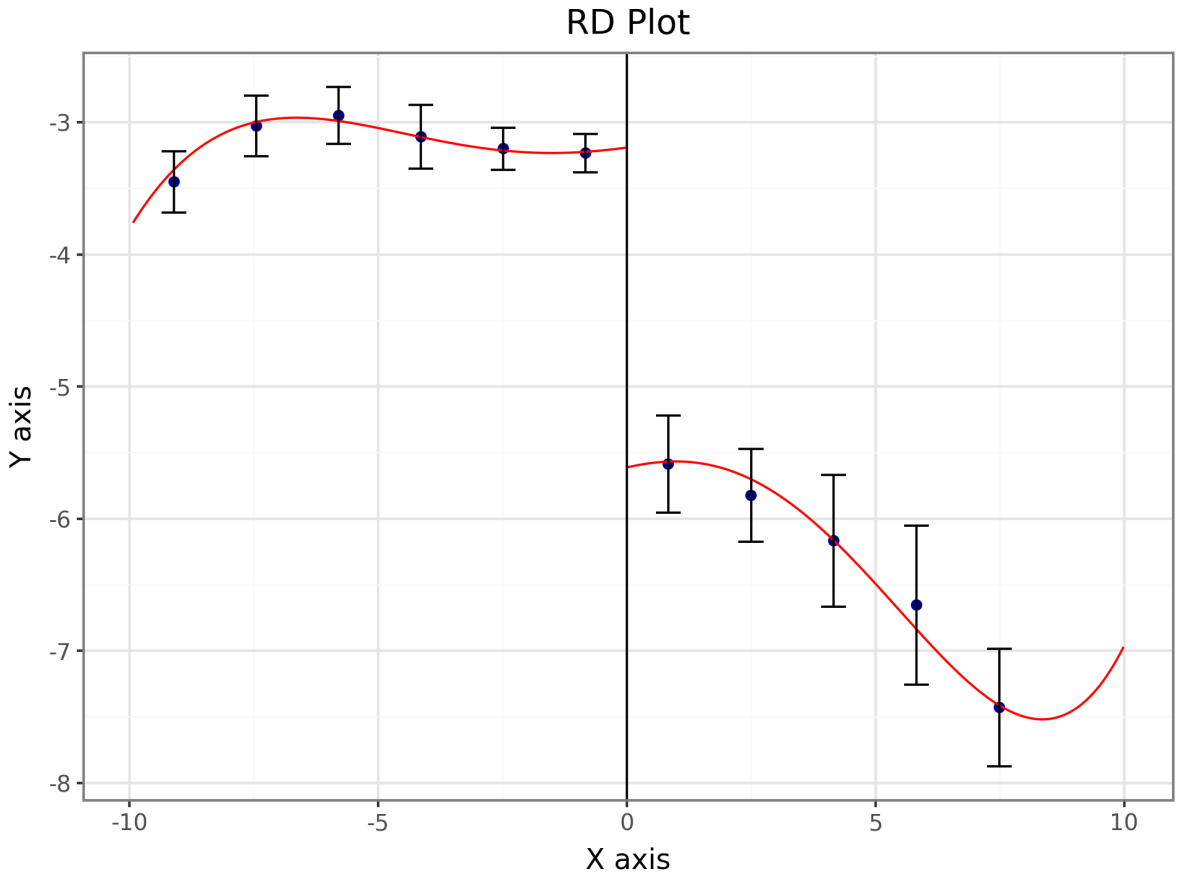
$$J_{ES,-,n} = [\frac{V_-}{\mathbb{V}_{ES,-}} \frac{n}{\log(n)^2}]$$

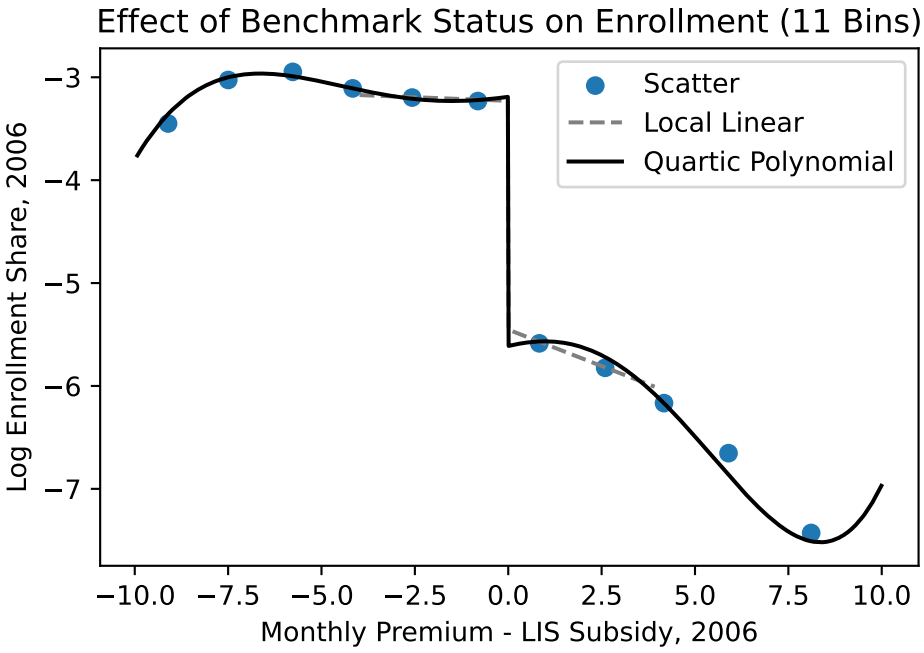
and

$$J_{ES,+,n} = [\frac{V_+}{\mathbb{V}_{ES,+}} \frac{n}{\log(n)^2}]$$

where V_- and V_+ denote the sample variance of the subsamples to the left and right of the cutoff and ν_{ES} 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.

```
C:\Users\micha\Lib\site-packages\rdrobust\rdplot.py:743: FutureWarning: Using print(plot) to
```

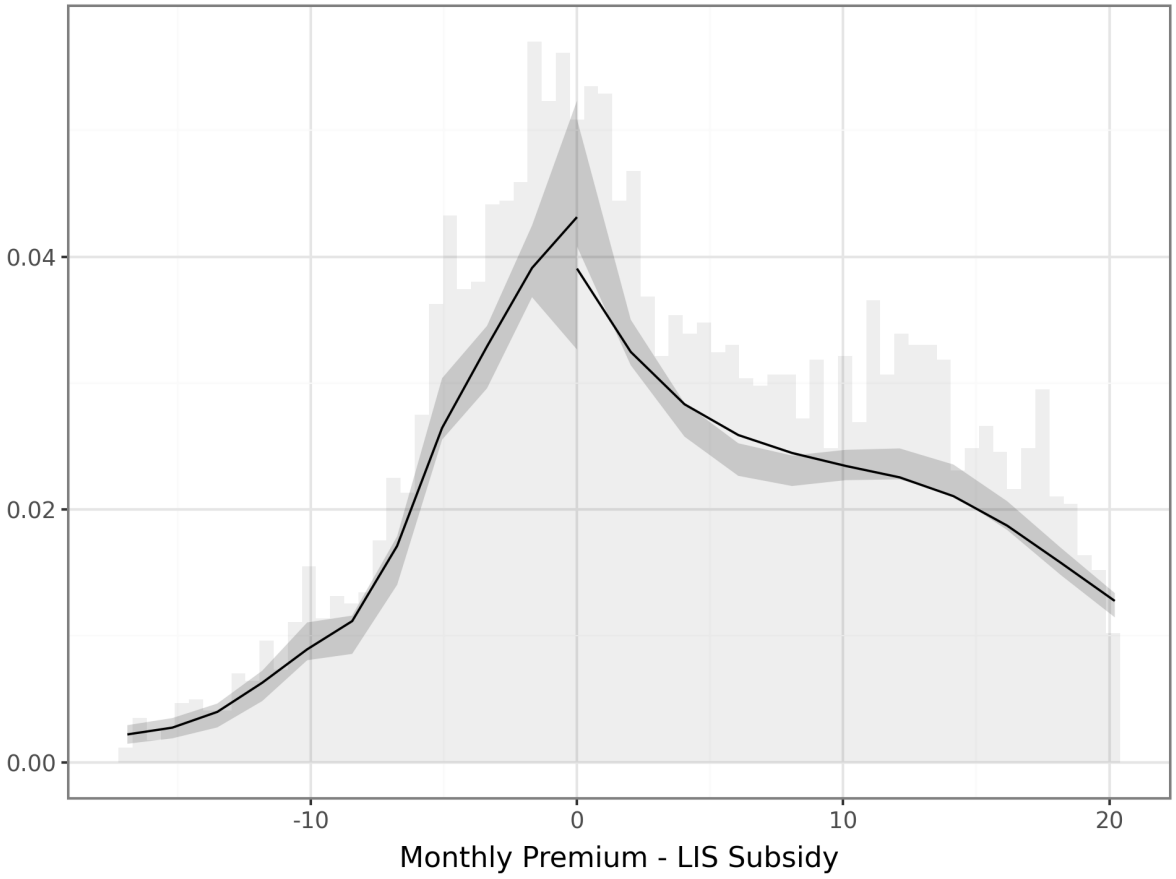





```
The optimal number of bins is 11.
```

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.

```
The p-value for the manipulation test is 2
The difference in estimated density at the cutoff is 0.0026
```



6. Recreate Panels A and B of Table 3 in Ericson (2014) using the same bandwidth of $4.00 but without any covariates.

	2006	2007	2008	2009	2010
Panel A. Local linear, bandwidth $4					
Below benchmark, 2006	2.224	0.0	0.0	0.0	0.0
	(0.255)	(0.0)	(0.0)	(0.0)	(0.0)
Below Benchmark	-0.014	-0.008	-0.014	-0.256	-0.103
	(0.051)	(0.037)	(0.022)	(0.061)	(0.044)
Above Benchmark	-0.142	-0.136	-0.076	-0.055	-0.078
	(0.114)	(0.026)	(0.014)	(0.018)	(0.016)
Observations	306.0	276.0	231.0	173.0	160.0
R^2	0.576	0.112	0.143	0.242	0.332
Panel B. Polynomial, bandwidth $4					
Below benchmark, 2006	2.311	0.0	0.0	-0.0	0.0
	(0.544)	(0.0)	(0.0)	(0.0)	(0.0)
Observations	306.0	276.0	231.0	173.0	160.0
R^2	0.583	0.119	0.215	0.342	0.398

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).

	2006	2007	2008	2009	2010
Panel A. Local linear, CE-optimal Bandwidths					
LIS Premium, 2006	-2.091	-1.142	-0.608	-0.321	-0.241
	(0.29)	(0.271)	(0.258)	(0.303)	(0.353)
Bandwidth	3.44	3.69	4.4	3.78	4.26
Panel B. Polynomial, CE-optimal Bandwidths					
LIS Premium, 2006	-2.409	-1.306	-0.634	-0.276	-0.219
	(0.361)	(0.346)	(0.317)	(0.342)	(0.384)
Bandwidth	4.29	4.74	6.2	6.4	7.83

The bandwidth at the beginning of the time frame is very close to Ericson's bandwidth of \$4. Over time, however, the optimal bandwidth grows significantly, so choosing one bandwidth for the entire analysis may not be optimal. The signs for all of my results are opposite to Ericson's which was initially alarming, but the interpretation appears to be the same as Ericson's conclusion - pricing lower in 2006 led to increasing market shares in the following years.

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.

Dep. Variable:	premium_diff	**R-squared:**	-0.1232
Estimator:	IV-2SLS	**Adj. R-squared:**	-0.1234
No. Observations:	5446	**F-statistic:**	278.56
Date:	Wed, Apr 24 2024	**P-value (F-stat)**	0.0000
Time:	19:10:47	**Distribution:**	chi2(1)
Cov. Estimator:	robust		

	Parameter	**Std. Err.**	**T-stat**	**P-value**	**Lower CI**	**Upper CI**
lnshare	1.4109	0.0845	16.690	0.0000	1.2453	1.5766

Endogenous: lnshare
Instruments: lispremium06
Robust Covariance (Heteroskedastic)
Debiased: False

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).

My dependent variable for part 8 was the difference in premium from 2006 to 2007, so we can interpret these results to say that an increased market share, as instrumented by the presence of a low-income subsidy, led to an increase in premiums from 2006 to 2007. This conclusion is consistent with the 'invest-then-harvest' theory and results from Table 4.

It doesn't seem as though different binwidths make a large difference in visual results, but using your program to choose the optimal binwidth could ensure that you provide a clear visual that fits the data while reducing noise.

Choosing different bandwidths do tend to affect the results more and, at least in this case, choosing one bandwidth for several years of data does not seem to be the best option. It seems like for contexts like healthcare, where prices can grow very rapidly, allowing for a bandwidth that changes over time may help you make the clearest comparisons.

10. Reflect on this assignment. What did you find most challenging? What did you find most surprising?

I found it surprisingly challenging to learn to use the rdplot and rddensity packages. Most packages I have used have very clear and established documentation online, so finding information about different commands or functionality is quite easy. The best way to get information on these packages, however, was directly from the creator's github. I found it a little difficult to learn to navigate the different documentation and find where the most helpful information was, but it was definitely a helpful experience to have to learn to find what I needed. Even having done it once for the first package made it easier to find what I needed for the second package, so I think it was a beneficial experience. Debugging is also always challenging and frustrating, but it is encouraging to find that I am (slowly but surely) becoming faster and more efficient at it.