Core Econometrics III: Problem Set 3

Owen Jetton

05/29/2022

Question 01

```
allsites = read_dta("allsites.dta")
```

From looking at the data set, I can see that there are an extensive number of variables that include information on the properties of houses in each tract, the demographics of the area, as well as historical data for each observation. Something that sticks out to me is that not every observation has its state listed, but all of them seem to have the statefips code.

Question 02

```
# Regressing 2000 log median house value for a Census tract on the indicator for whether the tract was
reg_2 = feols(data = allsites, lnmdvalhs0 ~ npl2000)
```

```
table2 = etable(reg_2,
                style.tex = style.tex("aer"),
                tex = TRUE)
```

	Log Median House Price
	(1)
(Intercept)	11.72***
	(0.0031)
npl2000	-0.0213
	(0.0202)
Observations	42,974
R^2	2.59×10^{-5}
Adjusted R^2	2.64×10^{-6}

The coefficient on the indicator for whether the tract was listed on the National Priorities List before 2000 is -0.0213, meaning that if the tract was listed, then the log median house value for that tract will decrease by -0.0213. However, this estimate is not statistically significant.

Question 03

```
reg_3 = feols(data = allsites, lnmdvalhs0 ~ npl2000, cluster = "statefips")
```

```
table3 = etable(reg_3,
                style.tex = style.tex("aer"),
                tex = TRUE)
```

	Log Median House Price
	(1)
(Intercept)	11.72***
	(0.0924)
npl2000	-0.0213
	(0.0488)
Observations	42,974
R^2	2.59×10^{-5}
Adjusted R^2	2.64×10^{-6}

Clustering by statefips code (which clusters by state) yields larger standard errors. Since I believe it is likely that there is correlation among the dummy variable (npl2000) within states, I believe that clustering by state should affect our standard errors.

Question 04

```
# Control for 1980 housing values
reg_4.1 = feols(data = allsites, lnmdvalhs0 ~ npl2000 + lnmeanhs8)

# Also control for economic and demographic variables. (Report which variables you included.)
reg_4.2 = feols(data = allsites, lnmdvalhs0 ~ npl2000 + lnmeanhs8 + pop_den8 + shrblk8 + old8 + ffh8 +

# Also add state fixed effects.
reg_4.3 = feols(data = allsites, lnmdvalhs0 ~ npl2000 + lnmeanhs8 + pop_den8 + shrblk8 + old8 + ffh8 +
```

```
table4 = etable(reg_4.1, reg_4.2, reg_4.3,
                  style.tex = style.tex("aer"),
                  tex = TRUE)
```

	Log Median House Price		
	(1)	(2)	(3)
(Intercept)	2.404***	3.360***	3.360***
	(0.0384)	(0.0466)	(0.7878)
npl2000	0.0400***	0.1216***	0.1216***
	(0.0131)	(0.0119)	(0.0170)
lnmeanhs8	0.8557***	0.7212***	0.7212***
	(0.0035)	(0.0043)	(0.0731)
pop_den8		1.47×10^{-5}***	1.47×10^{-5}***
		(1.96×10^{-7})	(1.12×10^{-6})
shrblk8		-0.1992***	-0.1992***
		(0.0112)	(0.0677)
old8		0.3066***	0.3066
		(0.0260)	(0.1919)
ffh8		0.1623***	0.1623
		(0.0237)	(0.1606)
unemprt8		-0.1959***	-0.1959
		(0.0549)	(0.6836)
povrat8		0.2213***	0.2213
		(0.0303)	(0.2531)
avhhin8		1.72×10^{-5}***	1.72×10^{-5}***
		(2.87×10^{-7})	(2.12×10^{-6})
Standard-Errors	IID		statefips
Observations	42,974	42,974	42,974
R^2	0.57916	0.65512	0.65512
Adjusted R^2	0.57914	0.65505	0.65505

Economic & demographic variables added for (2): population of tract in 1980 (pop_den8), share of black population (shrblk8), share of hispanic population (shrhsk8), share of population over 65 in 1980 (old8), share of female heads of household (ffh8), population unemployed over 16 years old in 1980 (unemprt8), population below poverty line in 1980 (povrat8), and average household income 1980 (avhhin8).

With these controls added, the coefficient on the variable npl2000 is positive and statistically significant. Meaning that if the tract was listed on the National Priorities List before 2000, that increases the price of the average house in that tract.

Question 05

In order for the coefficient on the NPL indicator in the regressions thus far to be unbiased, we must assume that the disturbance terms are uncorrelated with all of the covariates and with each other, and that the true model is linear in these covariates.

Question 06

```
allcovariates = read_dta("allcovariates.dta")

# listed in NPL
reg_6.1 = feols(data = allcovariates %>% filter(npl2000 == 1), mdvalhs0 ~ mdvalhs9 + pop_den8 + shrblk8


# Not listed in NPL
reg_6.2 = feols(data = allcovariates %>% filter(npl2000 == 0), mdvalhs0 ~ mdvalhs9 + pop_den8 + shrblk8
```

```r
table6 = etable(reg_6.1, reg_6.2,
                style.tex = style.tex("aer"),
                tex = TRUE)
```

	Median House Price 2000	
	(NPL)	(Not in NPL)
(Intercept)	-1,556.5	-24,386.6***
	(9,414.7)	(1,438.4)
mdvalhs9	0.6194***	0.8851***
	(0.0216)	(0.0042)
pop_den8	1.944**	0.8983***
	(0.7636)	(0.0357)
shrblk8	-32,038.6***	-39,205.7***
	(12,337.1)	(1,859.7)
ffh8	19,516.4	11,430.9***
	(23,182.1)	(3,947.9)
unemprt8	-90,952.9*	-54,410.7***
	(48,445.4)	(9,034.9)
povrat8	31,567.7	87,438.8***
	(34,828.5)	(5,065.8)
avhhin8	3.866***	3.656***
	(0.3518)	(0.0495)
Observations	985	47,260
R^2	0.64648	0.69613
Adjusted R^2	0.64395	0.69608

There are a few noticeable differences in the coefficients between these regressions. A few examples are that an increase in the median price of a house in a tract in 1990 (mdvalhs9) has a greater effect on the price in 2000 for tracts *not* listed in the NPL than those listed in it, unemployment rate has a greater negative effect on the price of houses in tracts listed in NPL, and surprisingly, the poverty rate does not have significant effect on the price of houses in tracts listed in the NPL but has positive effect on houses that were not listed. The covariates do not appear to be balanced across these two groups.

Note that this uses the median household value instead of the log median household value as in previous regressions.

Question 07

```r
# Getting the data
sitecovariates = read_dta("sitecovariates.dta")

# NPL-listed prior to 2000 vs. not NPL-listed prior to 2000 (what we've been doing so far)
reg_7.1t = feols(data = sitecovariates %>% filter(npl2000 == 1),
                 mdvalhs0 ~ mdvalhs9 + pop_den8 + shrblk8 + ffh8 + unemprt8 + povrat8 + avhhin8)

reg_7.1c = feols(data = sitecovariates %>% filter(npl2000 == 0),
                 mdvalhs0 ~ mdvalhs9 + pop_den8 + shrblk8 + ffh8 + unemprt8 + povrat8 + avhhin8)

# HRS score above 28.5 vs. HRS score below 28.5
reg_7.2t = feols(data = sitecovariates %>% filter(hrs_82 > 28.5),
                 mdvalhs0 ~ mdvalhs9 + pop_den8 + shrblk8 + ffh8 + unemprt8 + povrat8 + avhhin8)
```

```
reg_7.2c = feols(data = sitecovariates %>% filter(hrs_82 < 28.5),
                 mdvalhs0 ~ mdvalhs9 + pop_den8 + shrblk8 + ffh8 + unemprt8 + povrat8 + avhhin8)

# HRS score in [16.5, 28.5) vs. HRS score in [28.5, 40.5]
reg_7.3t = feols(data = sitecovariates %>% filter(hrs_82 > 16.5 & hrs_82 < 28.5),
                 mdvalhs0 ~ mdvalhs9 + pop_den8 + shrblk8 + ffh8 + unemprt8 + povrat8 + avhhin8)

reg_7.3c = feols(data = sitecovariates %>% filter(hrs_82 >= 28.5 & hrs_82 < 40.5),
                 mdvalhs0 ~ mdvalhs9 + pop_den8 + shrblk8 + ffh8 + unemprt8 + povrat8 + avhhin8)
```

```
table7.1 = etable(reg_7.1t, reg_7.1c,
                  style.tex = style.tex("aer"),
                  tex = TRUE)
```

NPL-listed prior to 2000 vs. not NPL-listed prior to 2000

	Median House Price 2000	
	(NPL)	(Not in NPL)
(Intercept)	-58,676.1***	-42,678.1*
	(16,128.1)	(22,990.5)
mdvalhs9	0.4666***	0.6086***
	(0.0332)	(0.0762)
pop_den8	-3.118**	-1.135
	(1.288)	(0.9080)
shrblk8	-52,763.2***	-14,654.6
	(16,648.4)	(23,067.7)
ffh8	90,555.4***	76,818.5**
	(27,348.8)	(35,702.7)
unemprt8	7,592.8	-82,864.1
	(66,822.5)	(81,006.9)
povrat8	73,012.0	-13,910.8
	(50,185.2)	(61,431.0)
avhhin8	6.395***	5.522***
	(0.5895)	(0.9681)
Observations	332	155
R^2	0.67576	0.59398
Adjusted R^2	0.66876	0.57465

```
table7.2 = etable(reg_7.2t, reg_7.2c,
                  style.tex = style.tex("aer"),
                  tex = TRUE)
```

HRS score above 28.5 vs. HRS score below 28.5

	Median House Price 2000	
	(HRS greater than 28.5)	(HRS less than 28.5)
(Intercept)	-63,859.0***	-38,573.8*
	(16,527.5)	(21,624.4)
mdvalhs9	0.4357***	0.6084***
	(0.0362)	(0.0564)
pop_den8	-2.873**	-1.173
	(1.312)	(0.8728)
shrblk8	-54,608.3***	-13,462.1
	(17,826.8)	(20,778.9)
ffh8	88,876.4***	82,186.3**
	(28,090.9)	(32,986.7)
unemprt8	34,039.4	-89,560.6
	(70,377.5)	(72,004.5)
povrat8	78,009.8	-25,505.4
	(51,329.3)	(58,633.7)
avhhin8	6.713***	5.294***
	(0.6127)	(0.8626)
Observations	306	181
R^2	0.67106	0.63012
Adjusted R^2	0.66333	0.61516

```
table7.3 = etable(reg_7.3t, reg_7.3c,
                  style.tex = style.tex("aer"),
                  tex = TRUE)
```

HRS score in [16.5, 28.5) vs. HRS score in [28.5, 40.5]

	Median House Price 2000	
	HRS score in [16.5, 28.5)	HRS score in [28.5, 40.5]
(Intercept)	-49,993.4	-36,618.6*
	(40,170.1)	(21,836.1)
mdvalhs9	0.5281***	0.3664***
	(0.0672)	(0.0420)
pop_den8	-0.8244	-0.1053
	(1.503)	(1.538)
shrblk8	-26,492.9	-32,381.9
	(32,495.3)	(24,387.6)
ffh8	45,821.1	-17,684.1
	(68,324.4)	(44,203.6)
unemprt8	-48,440.3	-25,309.1
	(115,905.5)	(90,746.4)
povrat8	42,786.8	100,972.0
	(89,581.2)	(61,047.3)
avhhin8	5.944***	6.056***
	(1.409)	(0.7758)
Observations	90	137
R^2	0.63726	0.71445
Adjusted R^2	0.60630	0.69896

The effect of the median house price in the tract in 1990 are greater for the tracts not in the 2000 NPL listing, just like in question **6**. The unemployment and poverty rates are never statistically significant. In groups with larger HRS scores in 1982 for the second and third tables, the effects of the 1990 median price on the 2000 price is *lower* than in groups with smaller HRS scores in 1982.

Question 08

In order to use HRS score as an instrument for NPL listing, we must assume that the HRS score is uncorrelated with all disturbances that are correlated with both the NPL listing & the median household price in 2000. Another assumption is that the covariance between the HRS score and the NPL listing is not zero. Our final assumption is that HRS score is *not* correlated with our outcome variable (median price in 2000).

Question 09

If we are to use a regression discontinuity to estimate the effect of NPL listing where HRS is the running variable with a cutoff at 28.5, then we must assume that NPL listing is determined by HRS score crossing the cutoff threshold. We also must assume that the median price of houses in treated and untreated (in NPL or not) tracts are continuous in HRS score at the cutoff, *and* that all included covariates are continuous in HRS score at the cutoff.

Question 10

Consider the following three "facts":
i.) The EPA states that the 28.5 cutoff was selected because it produced a manageable number of sites.
ii.) None of the individuals involved in identifying the site, testing the level of pollution, or running the 1982 HRS test knew the cutoff/threshold.
iii.) EPA documentation emphasizes that the HRS test is an imperfect scoring measure.

Our assumption for IV that "*HRS score is uncorrelated with all disturbances that are correlated with both the NPL listing & the median household price in 2000*" is weakened by fact (iii.) as HRS being an imperfect scoring measure could mean many of the (assumed) independent disturbances may not be independent. This could make IV less valid.
Fact (ii.) strengthens our assumption for RD that "*the median price of houses in treated and untreated (in NPL or not) tracts are continuous in HRS score at the cutoff*" because without the cutoff being known, sorting is unlikely.

Question 11

```
miledata = read_dta("2miledata.dta")

# First Stage
reg_11.1 = feols(data = miledata, cluster = "statefips",
                 npl2000 ~ hrs_82 + mdvalhs9 + pop_den8_nbr + shrblk8_nbr + ffh8_nbr + unemprt8_nbr + po

  # add fitted values to data
miledata %<>% mutate(npl_fit = fitted(reg_11.1))


# Second Stage
reg_11.2 = feols(data = miledata, cluster = "statefips",
                 mdvalhs0 ~ npl_fit + mdvalhs9 + pop_den8_nbr + shrblk8_nbr + ffh8_nbr + unemprt8_nbr +


# Reduced Form
```

```
reg_11.3 = feols(data = miledata, cluster = "statefips",
                 mdvalhs0 ~ hrs_82 + mdvalhs9 + pop_den8_nbr + shrblk8_nbr + ffh8_nbr + unemprt8_nbr + p
```

Note that these regressions (and all remaining regressions) include the following covariates: median house price 1990 (mdvalhs9), population of tract in 1980 (pop_den8_nbr), share of black population (shrblk8_nbr), share of hispanic population (shrhsk8_nbr), share of female heads of household (ffh8_nbr), population unemployed over 16 years old in 1980 (unemprt8_nbr), population below poverty line in 1980 (povrat8), and average household income 1980 (avhhin8_nbr). The instructions never said to do otherwise

First Stage
This stage uses a regression with NPL listings in 2000 as our outcome, using our instrument HRS score in the regression to get estimates of NPL listings.

The coefficient on HRS score is $\gamma = 0.01989462$

Second Stage
Using the fitted values from the first stage regression in place of NPL listings, we estimate our second stage regression to get an unbiased estimate of the effect of NPL listings on median household price in 2000.

The coefficient on the fitted values of NPL listings is $\beta = 5807.319$

Reduced Form
The reduced form regression uses the outcome variable from the second stage and the input variables from the first stage, directly regressing HRS score on median household prices in 2000.

The coefficient on HRS score is $\pi = 115.5344$

Notice that $\beta = \frac{\pi}{\gamma} \Rightarrow 5807.317 = \frac{115.5344}{0.01989462}$

Question 12

```
# First Stage w/ state fixed effects
reg_12.1 = feols(data = miledata, cluster = "statefips",
                 npl2000 ~ hrs_82 + mdvalhs9 + pop_den8_nbr + shrblk8_nbr + ffh8_nbr + unemprt8_nbr + po


# Reduced Form w/ state fixed effects
reg_12.red = feols(data = miledata, cluster = "statefips",
                 mdvalhs0 ~ hrs_82 + mdvalhs9 + pop_den8_nbr + shrblk8_nbr + ffh8_nbr + unemprt8_nbr + p
```

The coefficient on the NPL listing from the first stage regression is $\gamma' = 0.01959304$ which is very close to our estimate from part **11**, meaning the addition of state fixed effects to the regression did not change our first stage by much. However, the coefficient on HRS score from the reduced form regression is now $\pi' = 3.668719$ which is very different from our estimate in question **11**.

This tells us that our IV estimate is: $\beta' = \frac{\pi'}{\gamma'} = \frac{3.668719}{0.01959304} = 187.246$

The large reduction in the estimate from the reduced form regression (and β') tells us that without controlling for state fixed effects, the estimate of HRS score on median house price was likely biased up.

Question 13

```
# Changing instrument to n indicator for whether HRS score is above 28.5 (no fixed effects)
miledata %<>% mutate(hrs_ind = ifelse(hrs_82 > 28.5, 1, 0))


# First Stage
```

```r
reg_13.1 = feols(data = miledata, cluster = "statefips",
                 npl2000 ~ hrs_ind + mdvalhs9 + pop_den8_nbr + shrblk8_nbr + ffh8_nbr + unemprt8_nbr + p

  # add fitted values to data
miledata %<>% mutate(npl_fit_13 = fitted(reg_13.1))


# Second Stage
reg_13.2 = feols(data = miledata, cluster = "statefips",
                 mdvalhs0 ~ npl_fit_13 + mdvalhs9 + pop_den8_nbr + shrblk8_nbr + ffh8_nbr + unemprt8_nbr


# Reduced Form
reg_13.3 = feols(data = miledata, cluster = "statefips",
                 mdvalhs0 ~ hrs_ind + mdvalhs9 + pop_den8_nbr + shrblk8_nbr + ffh8_nbr + unemprt8_nbr +
```

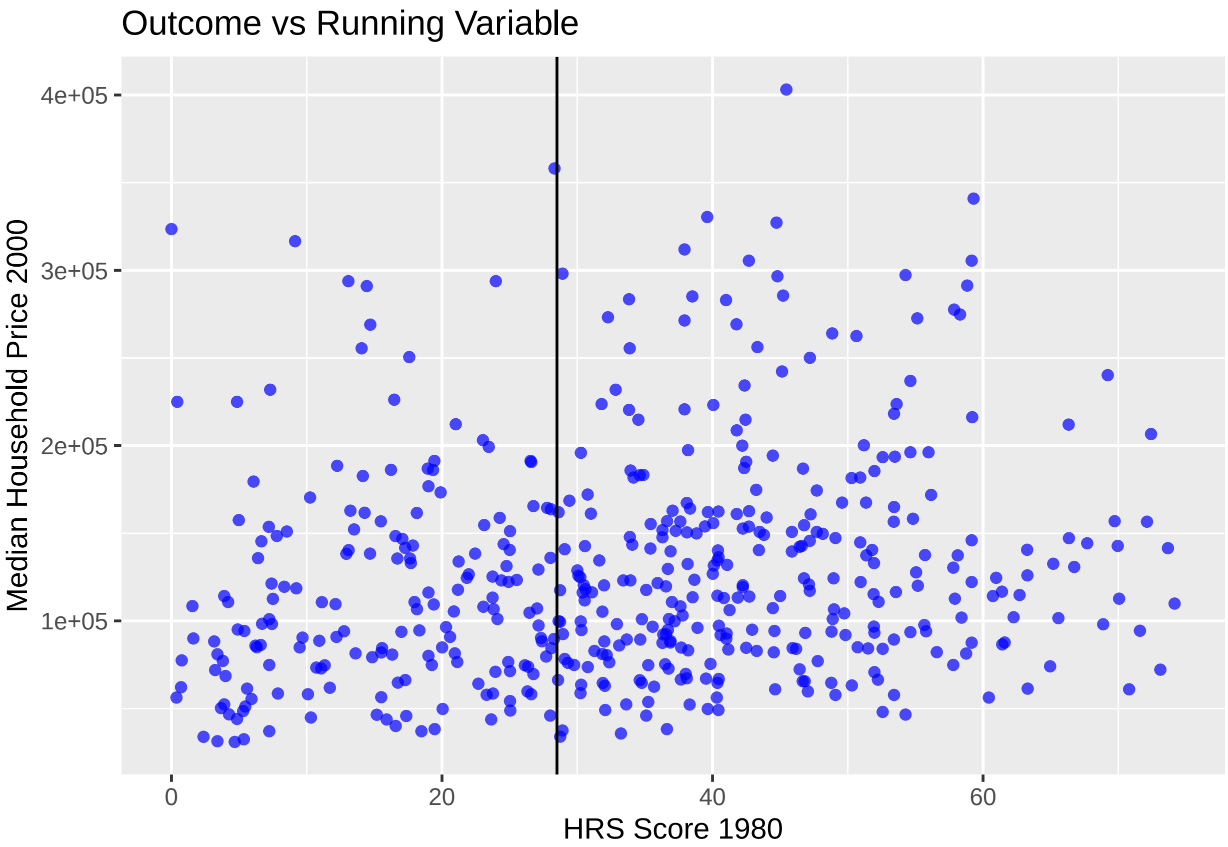
By changing our instrument from the HRS score to an indicator variable on if HRS score is above 28.5, we do see a few changes in our coefficient estimates:
- The estimate from the first stage is $\gamma'' = 0.8123954$ which is a smaller estimate than in question **11**.
- The estimate from the second stage (the IV estimate) is $\beta'' = 6085.865$ which is larger than our estimate in question **11**.
- The estimate from the reduced form is $\pi'' = 4944.129$ which is *much* larger than the estimate in question **11**.

Question 14

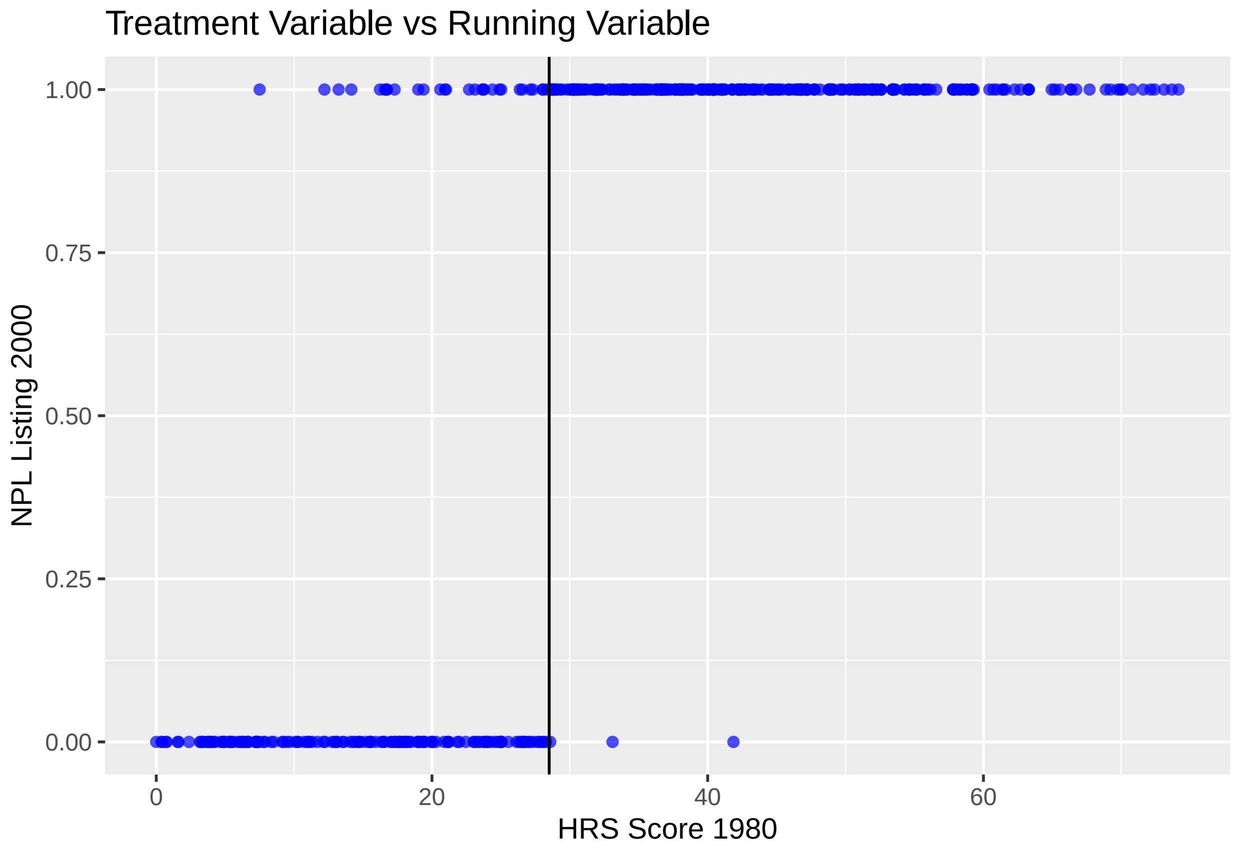
Since our estimates of coefficients **11** and **13** aren't close matches, this implies that the HRS score of a tract crossing the threshold score of 28.5 does not always mean that the tract was listed on the NPL. This means that if we were to estimate the effect using a regression discontinuity, we would use a fuzzy RD.

Question 15

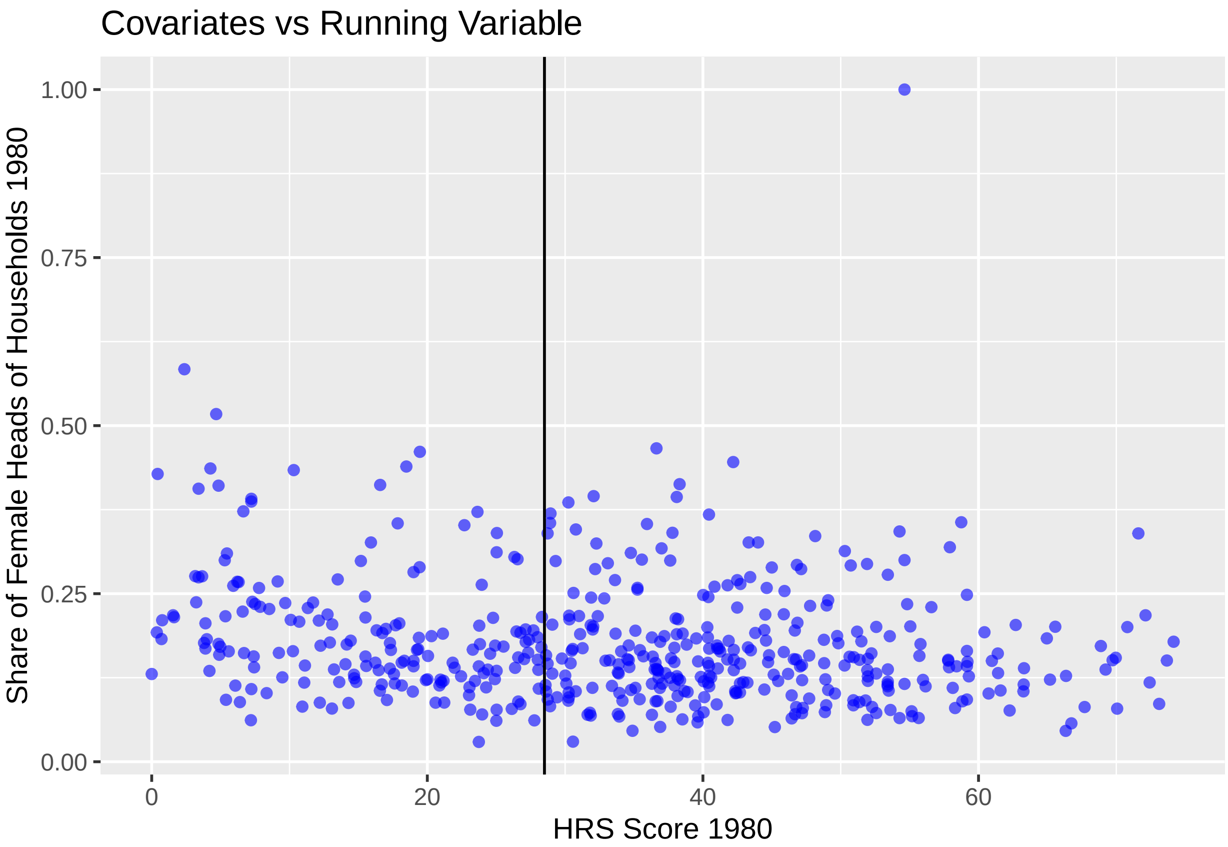
```r
# The outcome variable vs. the running variable
ggplot(data = miledata, aes(y = mdvalhs0, x = hrs_82)) +
  geom_point(color = "blue", alpha = 0.7) +
  geom_vline(xintercept = 28.5) +
  labs(title = "Outcome vs Running Variable",
       y = "Median Household Price 2000",
       x = "HRS Score 1980")
```



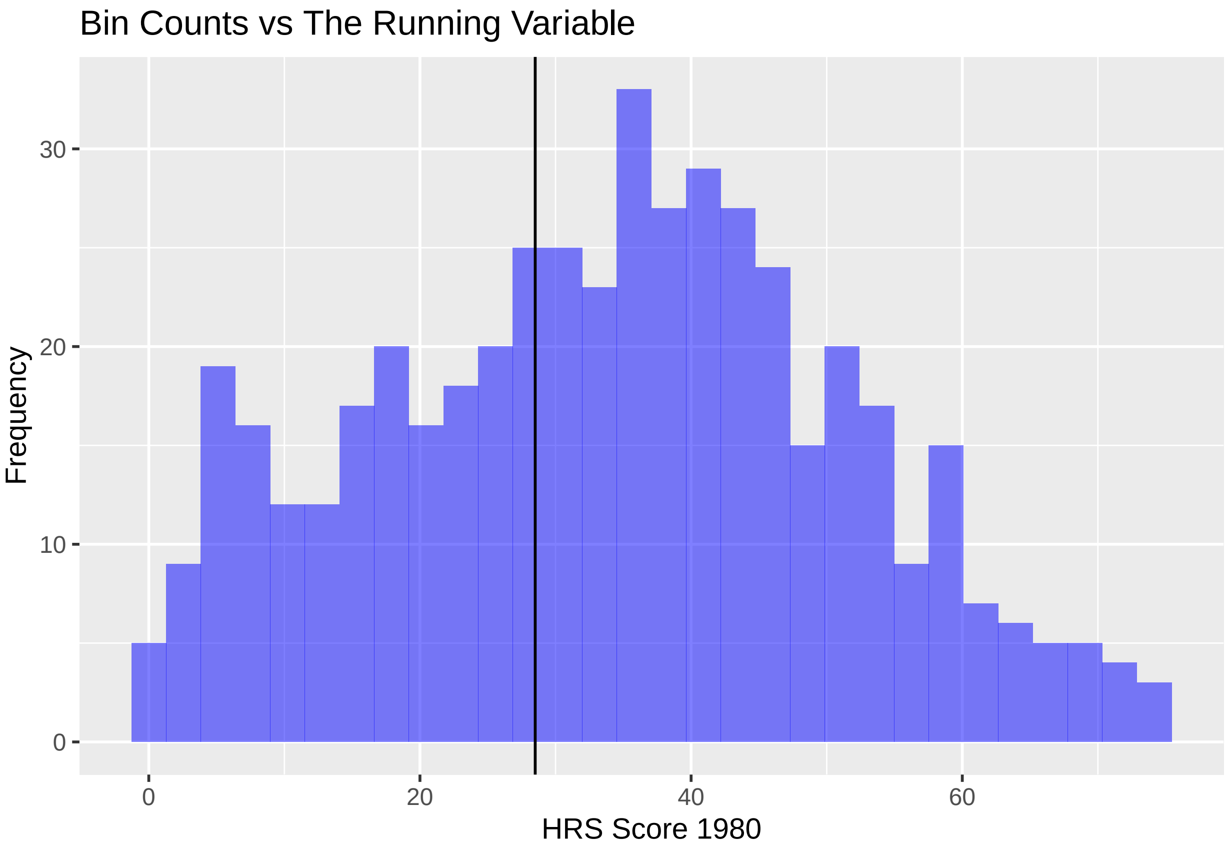
```r
# The treatment variable vs. the running variable
ggplot(data = miledata, aes(y = npl2000, x = hrs_82)) +
  geom_point(color = "blue", alpha = 0.7) +
  geom_vline(xintercept = 28.5) +
  labs(title = "Treatment Variable vs Running Variable",
       y = "NPL Listing 2000",
       x = "HRS Score 1980")
```



```
# Covariates vs. the running variable
ggplot(data = miledata, aes(y = ffh8_nbr, x = hrs_82)) +
  geom_point(color = "blue", alpha = 0.6) +
  geom_vline(xintercept = 28.5) +
  labs(title = "Covariates vs Running Variable",
       y = "Share of Female Heads of Households 1980",
       x = "HRS Score 1980")
```



```r
# Bin counts vs. the running variable
ggplot(data = miledata) +
  geom_histogram(aes(x = hrs_82), fill = "blue", alpha = 0.5, bins = 30) +
  geom_vline(xintercept = 28.5) +
  labs(title = "Bin Counts vs The Running Variable",
       y = "Frequency",
       x = "HRS Score 1980")
```



Question 16

The jump in the treatment variable around the cutoff in the running variable confirms what I said in **14** about wanting to do fuzzy RD instead of sharp RD. The histogram and the “outcome vs running variable” graphs not showing any discontinuity around the cutoff, means that our RD assumptions are satisfied. While there aren’t clear distinct trends on either side of the cutoff in the “outcome vs running variable,” I see it plausible that there could be. Therefore, with the assumptions satisfied, this does look like a plausible regression discontinuity.

Question 17

```
# limiting analysis to HRS scores between [16.5, 40.5]
miledata_17 = miledata %>% filter(hrs_82 >= 16.5 & hrs_82 <= 40.5)


# First Stage
reg_17.1 = feols(data = miledata_17, cluster = "statefips",
                 npl2000 ~ hrs_ind + mdvalhs9 + pop_den8_nbr + shrblk8_nbr + ffh8_nbr + unemprt8_nbr + p

  # add fitted values to data
miledata_17 %<>% mutate(npl_fit_17 = fitted(reg_17.1))


# Second Stage
reg_17.2 = feols(data = miledata_17, cluster = "statefips",
                 mdvalhs0 ~ npl_fit_17 + mdvalhs9 + pop_den8_nbr + shrblk8_nbr + ffh8_nbr + unemprt8_nbr
```

Our regression discontinuity estimate is $\hat{\beta} = 3197.198$ Using the smaller bandwidth of HRS scores in [16.5, 40.5], we find a tighter estimate than the 2SLS estimates found in questions 11 and 13.

Question 18

I believe the estimate from **17** to be the most credible. It was wise to use a two-stage least squares, to avoid bias arising from disturbances correlated with the NPL listing and median house price, and as shown above, HRS score is a fair instrument. Additionally, using a tighter bandwidth to not let observations farther from the cut-off point bias our estimate makes me believe it is more credible of an estimate than the regression in **13**. Therefore, the effect of being listed in the NPL 2000 listing increases the median price of a census tract by about \$3197.20