Pset3

Becky Cardinali, Yuen Ho, Sara Johns, and Jacob Lefler

12/4/2020

Question 1: OLS Regressions

Question 2: Regression Discontinuity Design

(a) Consider the HRS score as the running variable for an RD research design. What assumptions are needed on the HRS score? How do each of the below "facts" impact the appropriateness of these assumptions?

In order for regression discontinuity to be a valid research design, we need to assume that the potential outcomes $Y_i(0)$ and $Y_i(1)$ (housing prices) are smooth functions of the running variable X_i (the HRS score) as it crosses the threshold c (28.5). In other words, $E[Y_i(0)|X_i = x]$ and $E[Y_i(1)|X_i = x]$ are continuous in x. If there is imperfect compliance, that is if the probability of treatment increases, but by less than 100 pp, when the running variable crosses the threshold, then we need to use a fuzzy RD design. In this case, we need to make an additional monotonicity assumption that $D_i(x^*)$ is non-increasing in x^* at $x^* = c$, that is we need to assume there are no "defiers."

Importantly, our first assumption is violated if there is manipulation based on the HRS score. In other words, if individuals understand the assignment mechanism and can manipulate the HRS score to place a census block just above (or below) the threshold, then there is selection into treatment so census tracts just above and below the threshold are no longer comparable. Thus we need to assume that individuals cannot game the assignment mechanism in order for this to be a valid research design. Relatedly, we also need to assume that covariates are smooth at the threshold, that is that covariates are balanced above and below the threshold. If this is not true, then we have selection into treatment and observations just above and below the threshold are again not comparable.

```
(i) The EPA assertion that "the 28.5" cutoff was selected because it produced a manageable number
of sites."
```

This fact makes it more likely that our assumptions hold, because the threshold was not selected based on specific site characteristics, which would have potentially made covariates imbalanced across the threshold. For example, if instead the "28.5" cutoff was selected because a HRS rating of 28.5 or higher is especially (disproportionately) dangerous for human health, then our first assumption will no longer hold because houses close to sites above this threshold may benefit disproportionately from treatment.

```
(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 score.
```

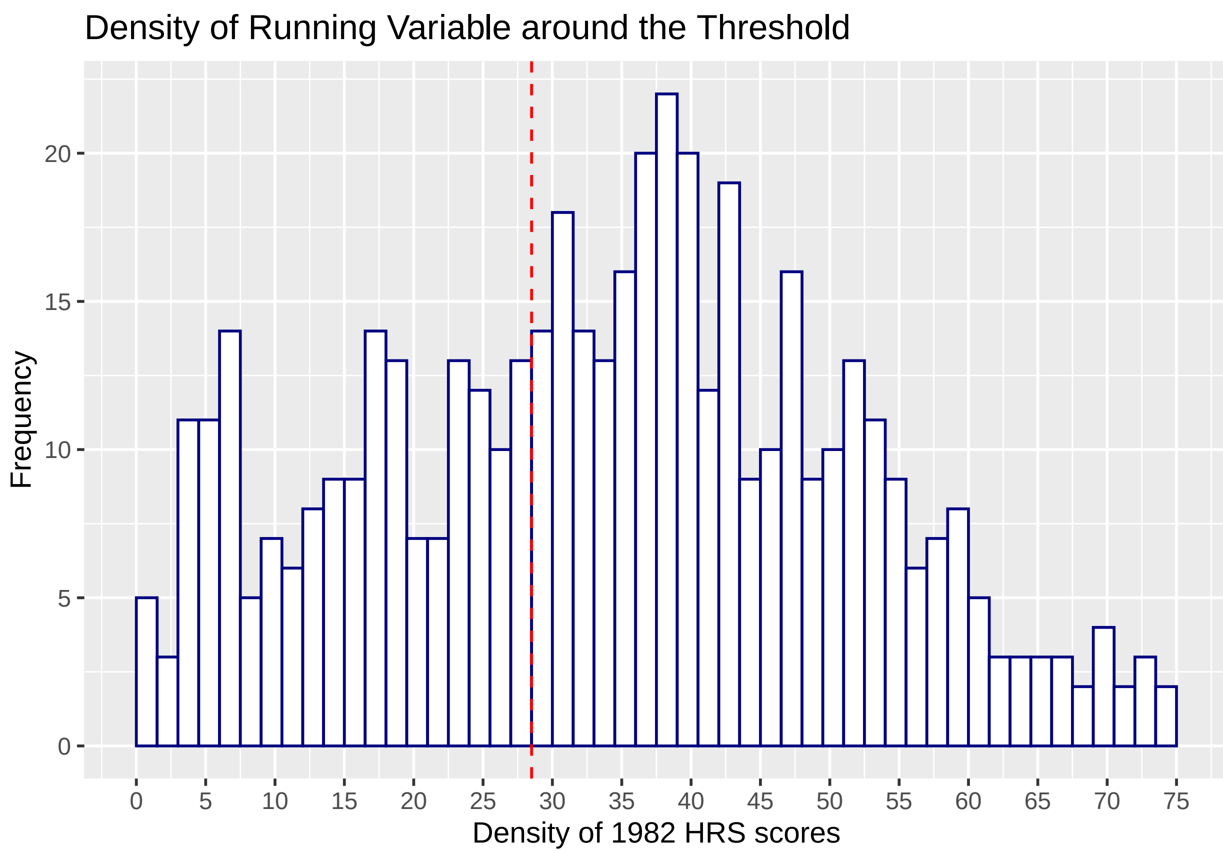
This fact makes it more likely that our assumptions hold. In particular, if none of the individuals involved knew the cutoff threshold score, it is less likely they were able to manipulate the test results to make certain census tracts be above (or below) the threshold. Even if individuals had an incentive to cheat, without knowing the assignment mechanism they would not have been able to game the system effectively.

```
(iii) EPA documentation emphasizes that the HRS test is an imperfect scoring measure
```

Whether this fact violates our assumptions depends on the type of error associated with the HRS test. If this is classical measurement error then it should not affect our assumptions. However, if the error is correlated with our covariates or with our outcome variable (housing prices) then this would violate our first assumption.

(b) Create a histogram of the distribution of the 1982 HRS scores by dividing the HRS score into non-overlapping bins. Include a vertical line at 28.5. Next run local linear regressions on either side of 28.5 using the midpoints of the bins as the data. What do you conclude?

```
## histogram of the density of 1982 HRS scores
ggplot(data, aes(x = hrs_82)) +
  geom_histogram(binwidth = 1.5, boundary = 0, closed = "left", col = "navy", fill = "white") +
  geom_vline(xintercept = 28.5, linetype = "dashed", color = "red") +
  theme_gray() +
  scale_x_continuous(breaks = seq(0,75,5)) +
  xlab("Density of 1982 HRS scores") +
  ylab("Frequency") +
  ggtitle("Density of Running Variable around the Threshold")
```




```
## Run local linear regressions on either side of threshold, using the midpoints of the bins as the data

range(data$hrs_82)# between 0 and 74.16
```

```
## [1]  0.00 74.16
```

```
h = 1.5 #set bandwidth
bins = seq(from = 0, to = 75, by = h) # set cutoffs for bins
length(bins)
```

```
## [1] 51
```

```
# returns the bin index for each observation
data$hrs_82_bin <- cut(data$hrs_82, breaks = bins, right = FALSE)

# calculate the midpoint of each bin
bins.midpoint = (bins[-1] + bins[-(length(bins))])/2

# assign a bin midpoint to each observation
data$hrs_82_binmid = bins.midpoint[ data$hrs_82_bin ]
```

```r
# generate average of the treatment variable (NPL assignment) for each bin
npl2000_bin =tapply(data$npl2000, data$hrs_82_bin, mean)

# generate average outcome in each bin
lnmdvalhs0_nbr_bin = tapply(data$lnmdvalhs0_nbr, data$hrs_82_bin, mean)

# regression fitted on data below the cutoff
below_lm <- lm(lnmdvalhs0_nbr_bin ~ bins.midpoint, data.frame(lnmdvalhs0_nbr_bin, bins.midpoint)[1:19,])
summary(below_lm)
```

```
##
## Call:
## lm(formula = lnmdvalhs0_nbr_bin ~ bins.midpoint, data = data.frame(lnmdvalhs0_nbr_bin,
##     bins.midpoint)[1:19, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32629 -0.08308  0.03012  0.08214  0.30266
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.486342   0.080292 143.057   <2e-16 ***
## bins.midpoint  0.007585   0.004881   1.554    0.139
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1748 on 17 degrees of freedom
## Multiple R-squared:  0.1244, Adjusted R-squared:  0.07284
## F-statistic: 2.414 on 1 and 17 DF,  p-value: 0.1387
```

```r
# regression fitted on data above the cutoff
above_lm <- lm(lnmdvalhs0_nbr_bin ~ bins.midpoint, data.frame(lnmdvalhs0_nbr_bin, bins.midpoint)[20:50,])
summary(above_lm)
```

```
##
## Call:
## lm(formula = lnmdvalhs0_nbr_bin ~ bins.midpoint, data = data.frame(lnmdvalhs0_nbr_bin,
##     bins.midpoint)[20:50, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.44550 -0.10299 -0.02503  0.11681  0.31197
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.657923   0.124356  93.746   <2e-16 ***
## bins.midpoint  0.001049   0.002326   0.451    0.655
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1738 on 29 degrees of freedom
## Multiple R-squared:  0.006959,   Adjusted R-squared:  -0.02728
## F-statistic: 0.2032 on 1 and 29 DF,  p-value: 0.6555
```

We estimate the treatment effect $\hat{\tau} = \hat{\alpha}_r - \hat{\alpha}_l =$ 11.658 - 11.486 = 0.172, the difference in the estimated intercepts from our local linear regressions above and below the threshold. This is suggestive evidence that being above the threshold, and therefore being more likely to be placed on the NPL, is associated with higher mean housing prices in 2000. Of course, a drawback of fitting separate local linear regressions on either side of the threshold is that we cannot conduct statistical inference on our estimated treatment effect.

Question 3: First Stage of RD Design

Question 4: Second Stage of RD Design

Question 5: Conclusion