HKS SUP-135 Lab 5: Evaluating Education Policy using Regression Discontinuity Design

Matt Khinda

3/24/2023

Question 1: Limitations of Direct Comparison

A regression discontinuity design relies on the identification assumption that the sample observations on either side of the threshold are nearly identical and thus directly comparable. This is a plausible assumption when looking at those just above and just below the cut-off, however the further from the threshold one looks the less this assumption is likely to hold. In this study of probation, for example, we can reasonably assume that a student who earned a 1.59 GPA is not fundamentally different to one who earned a 1.61 GPA, while students who earned a 0.6 GPA are possibly quite different from those who earned a 2.6 GPA in terms of academic motivation or other potential confounding variables. Direct comparison of all students above to all students below the threshold would include far too many of these dissimilar students for the comparison to meaningfully attribute any kind of causal effect to the use of an academic probation policy.

Question 2: Running Variable

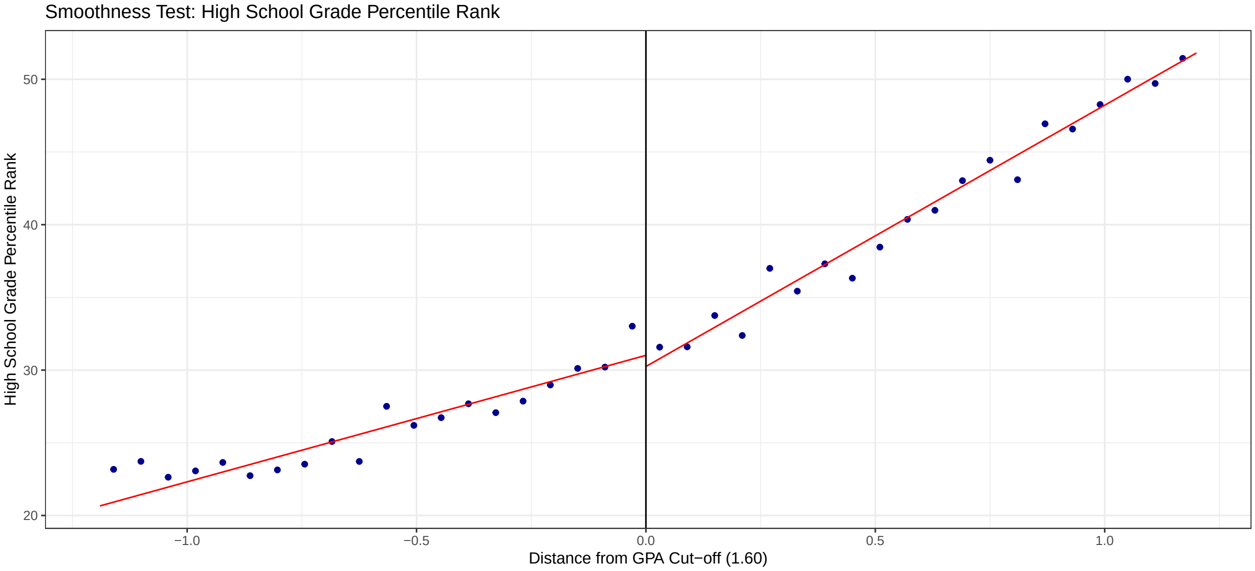
In this research design, the running variable is GPA which ultimately determines whether or not a student ends up on probation (treatment group) or not (control group).

Question 3: Predetermined Characteristics Plots

```r
## Calculate distance from threshold
probation$dist_from_cut <- probation$GPA - 1.6

## Create boolean above threshold variable
probation$above_threshold <- ifelse(probation$dist_from_cut >= 0, 1, 0)

## Subset dataset for bandwidth
probation_narrow <- subset(probation, dist_from_cut<=1.2 & dist_from_cut>=-1.2)

## Plot discontinuity of high-school grade percentile rank
rdplot(y = probation_narrow$hsgrade_pct,
       x = probation_narrow$dist_from_cut,
       c = 0,
       p = 1,
       nbins = 20,
       x.label = "Distance from GPA Cut-off (1.60)",
       y.label = "High School Grade Percentile Rank",
       title = "Smoothness Test: High School Grade Percentile Rank"
       )
```

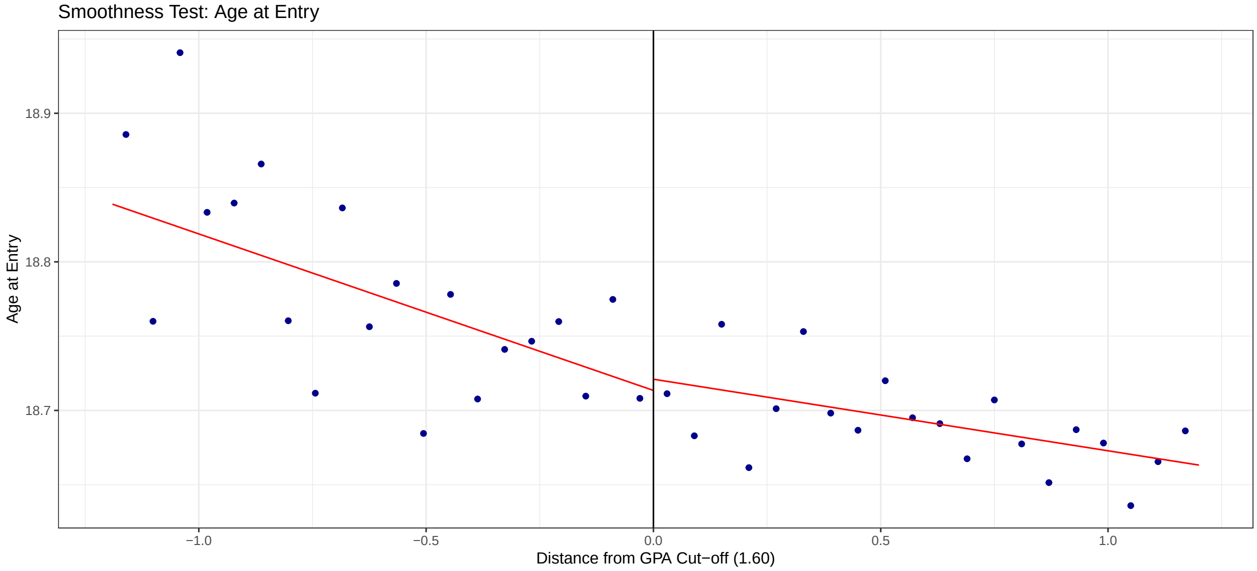
3a: Binned Scatter Plots

```
## [1] "Mass points detected in the running variable."
```



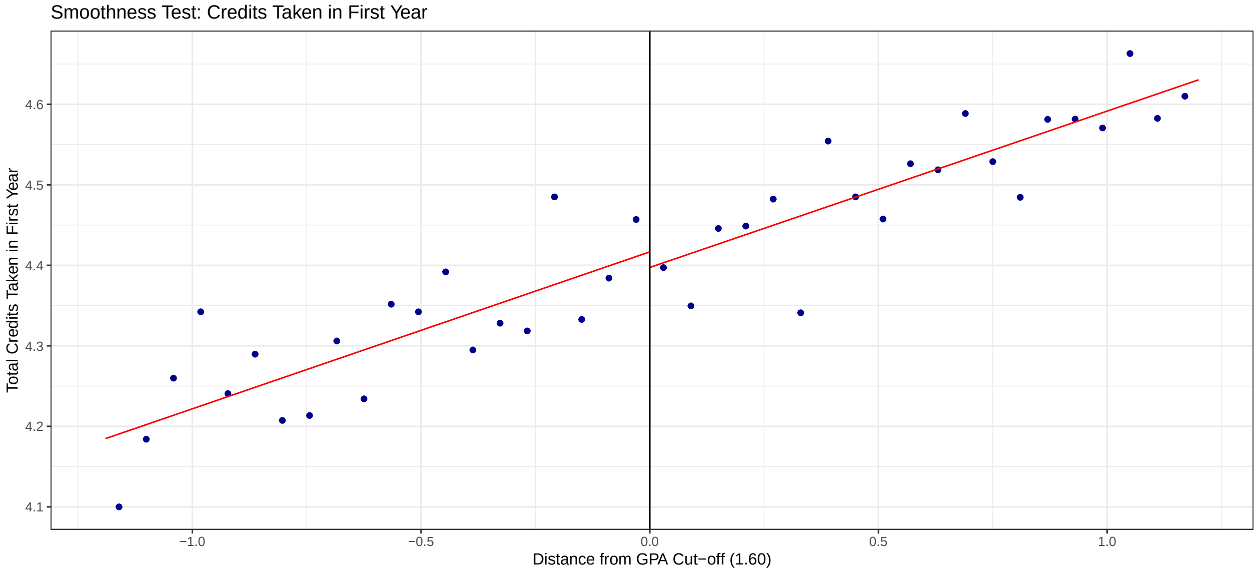
```
## Plot discontinuity of age at entry
rdplot(y = probation_narrow$age_at_entry,
       x = probation_narrow$dist_from_cut,
       c = 0,
       p = 1,
       nbins = 20,
       x.label = "Distance from GPA Cut-off (1.60)",
       y.label = "Age at Entry",
       title = "Smoothness Test: Age at Entry"
       )
```

```
## [1] "Mass points detected in the running variable."
```



```r
## Plot discontinuity of age at entry
rdplot(y = probation_narrow$totcredits_year1,
       x = probation_narrow$dist_from_cut,
       c = 0,
       p = 1,
       nbins = 20,
       x.label = "Distance from GPA Cut-off (1.60)",
       y.label = "Total Credits Taken in First Year",
       title = "Smoothness Test: Credits Taken in First Year"
       )
```

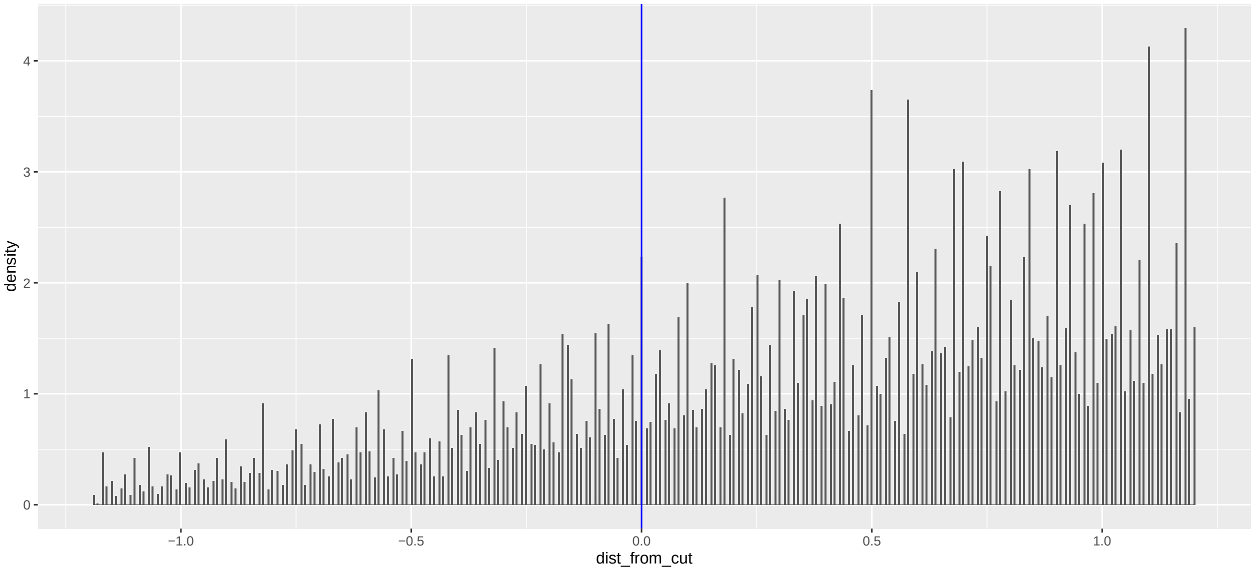
```
## [1] "Mass points detected in the running variable."
```



As shown in the three plots above, the predetermined characteristics of high school grade percentile (hsgrade_pct), age at entry (age_at_entry), and total credits attempted in the first year (totcredits_year1) exhibit low discontinuity at the threshold. This is a strong validation of the assumption underlying the regression discontinuity design.

```r
ggplot(probation_narrow) +
  geom_histogram(aes(x = dist_from_cut, y = ..density..), bins = 600) +
  geom_vline(xintercept = 0, color = "blue")
```

3b: Histograms

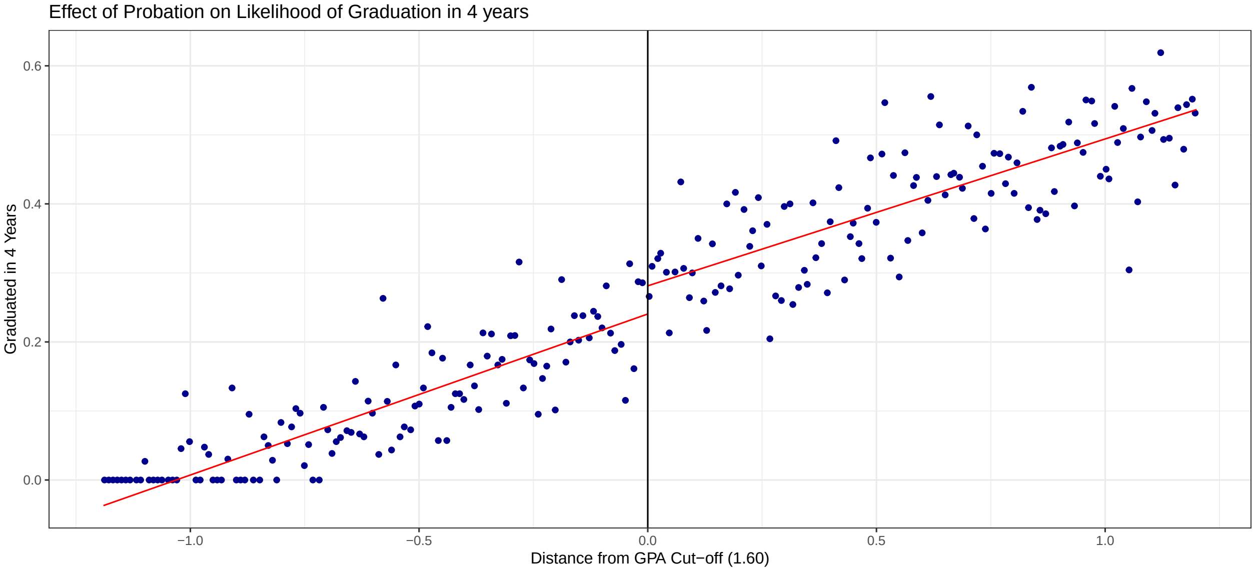


As observed in the density histogram, there is not a notable jump in density around the threshold which suggests that there is not bunching around the cut-off.

Question 4: Effect of Probation on Graduation Rates

```r
## Plot discontinuity of graduation rates
rdplot(y = probation_narrow$gradin4,
       x = probation_narrow$dist_from_cut,
       c = 0,
       p = 1,
       x.label = "Distance from GPA Cut-off (1.60)",
       y.label = "Graduated in 4 Years",
       title = "Effect of Probation on Likelihood of Graduation in 4 years"
       )
```

```
## [1] "Mass points detected in the running variable."
```



Question 5: Quantifying Discontinuity

```
# subset the data only to include those in the bandwidth and below the cut-off
probation_below <- subset(probation_narrow, above_threshold == 0)

model_1 <- lm(gradin4 ~ GPA, data = probation_below)
pred_1 <- model_1$coefficients[1] + model_1$coefficients[2]*1.6

cat("The predicted value at the threshold for the linear model based on the
    observations below the cut-off (on probation) is", pred_1)
```

5a: Predicted Value at the Threshold for Treatment Group Regression

```
## The predicted value at the threshold for the linear model based on the
##     observations below the cut-off (on probation) is 0.2404832
```

```
# subset the data only to include those in the bandwidth and above the cut-off
probation_above <- subset(probation_narrow, above_threshold == 1)

model_2 <- lm(gradin4 ~ GPA, data = probation_above)
pred_2 <- model_2$coefficients[1] + model_2$coefficients[2]*1.6

cat("The predicted value at the threshold for the linear model based on the
    observations above the cut-off (not on probation) is", pred_2)
```

5b: Predicted Value at the Threshold for Control Group Regression

```
## The predicted value at the threshold for the linear model based on the
##     observations above the cut-off (not on probation) is 0.2812735
```

```
# calculate difference in predicted values
discont <- pred_2-pred_1

cat("The difference between the predicted values at the threshold is", discont)
```

5c: Difference in Predicted Values

```
## The difference between the predicted values at the threshold is 0.04079028
```

Question 6: Multivariate Regression

```
model_3 <- lm(gradin4 ~ above_threshold + dist_from_cut +
                above_threshold*dist_from_cut, data = probation_narrow)
coeftest <- coeftest(model_3, vcov = vcovHC(model_3, type = "HC1"))
betaRD <- coeftest[2,1]

cat("The coefficient beta_rd from the multivariate regression is", betaRD, "which is
    exactly equaly to the difference in predicted values above.")
```

```
## The coefficient beta_rd from the multivariate regression is 0.04079028 which is
##     exactly equaly to the difference in predicted values above.
```

Question 7: Statistical Significance Using Standard Error

```
stdErr <- coeftest[2,2]
conf_upper <- betaRD + 1.96*stdErr
conf_lower <- betaRD - 1.96*stdErr

cat("The 95% confidence interval for the effect is between", conf_lower, "and", conf_upper, "
    which does not include 0, indicating that it is statistically significant.")
```

```
## The 95% confidence interval for the effect is between 0.0148214 and 0.06675917
##     which does not include 0, indicating that it is statistically significant.
```

Question 8: Program Effectiveness

Based on the regression discontinuity analysis, it is clear that the use of a probation program is not only ineffective at improving graduation rates within 4 years, but that it is actively detrimental to those students' graduation rates. This can be seen in the plot in Question 4 and in the values found in both Questions 5 and 6 where the effect of being above the threshold and not on probation increases the likelihood of graduation within 4 years by 4 percentage points.