Lab_5

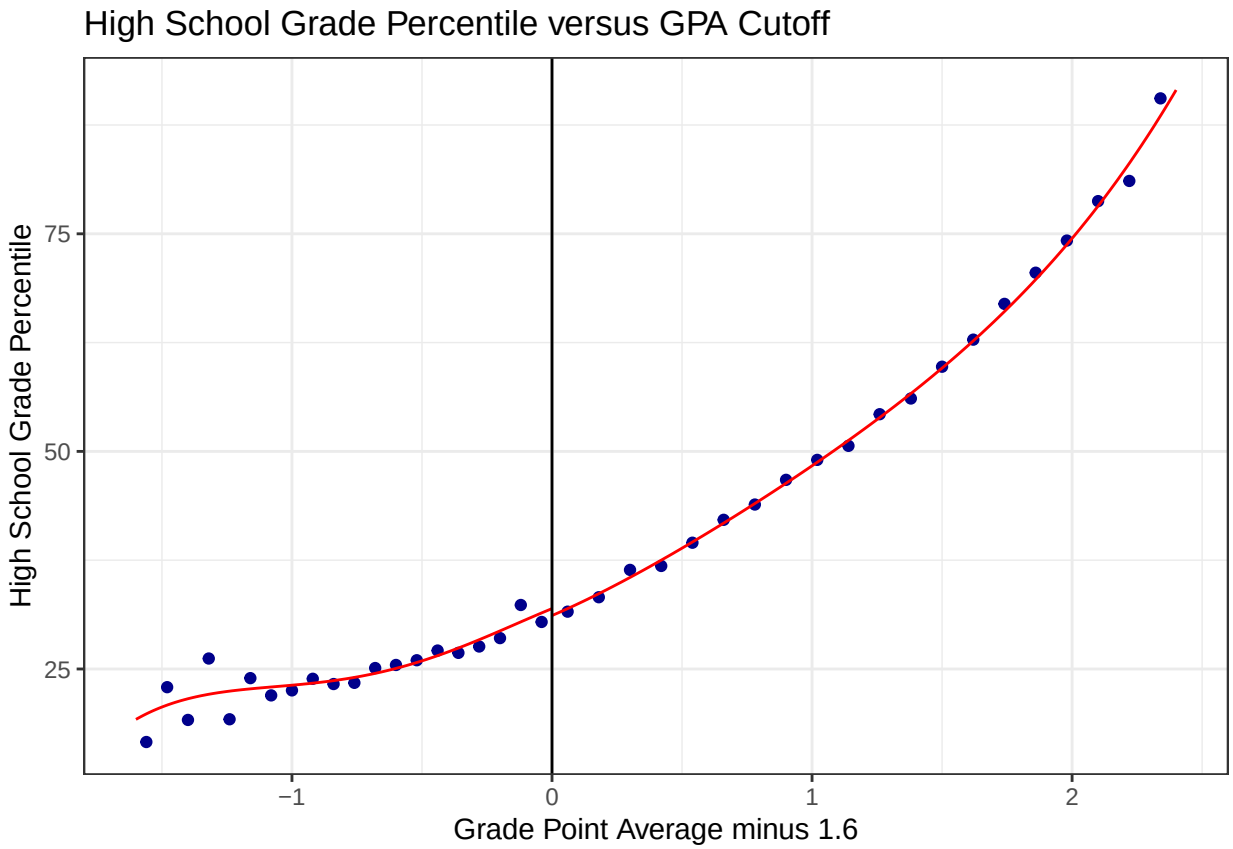
2023-03-25

Question 1 We don't want to all students who are on probation with all students who are not on probation to evaluate this University's program because we are trying evaluate the potential causal impact of the probation policy on educational outcomes. To do so, we are employing a regression discontuinity design that examines educational outcomes of students who did and did not receive probation but are otherwise roughly the same in other characteristics. This requires us to only look at observations right around (ie, right above and below) the cutoff of the 1.60 GPA threshold that triggers a student receiving academic probation.

Question 2 The running variable is the student's GPA. The cutoff is whether that GPA is below or above 1.6.

```
#Question 3: Binned scatter plots and histograms for 2-3 predetermined characteristics compared with GP

#Create GPA cutoff variable and filtered dataset
df <- df |>
      mutate(gpa_cutoff = GPA - 1.6)

df_narrow = df |>
            filter(gpa_cutoff <= 1.2 & gpa_cutoff >= -1.2)
view(df_narrow)

#Binned scatterplot for high school grade percentile
binned1 <- rdplot(x = df$gpa_cutoff, #outcome variable
                  y = df$hsgrade_pct, #running variable
                  c = 0,
                  nbins = c(20, 20),
                  binselect = "es",
                  x.label = "Grade Point Average minus 1.6",
                  y.label = "High School Grade Percentile",
                  title = "High School Grade Percentile versus GPA Cutoff")
```

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

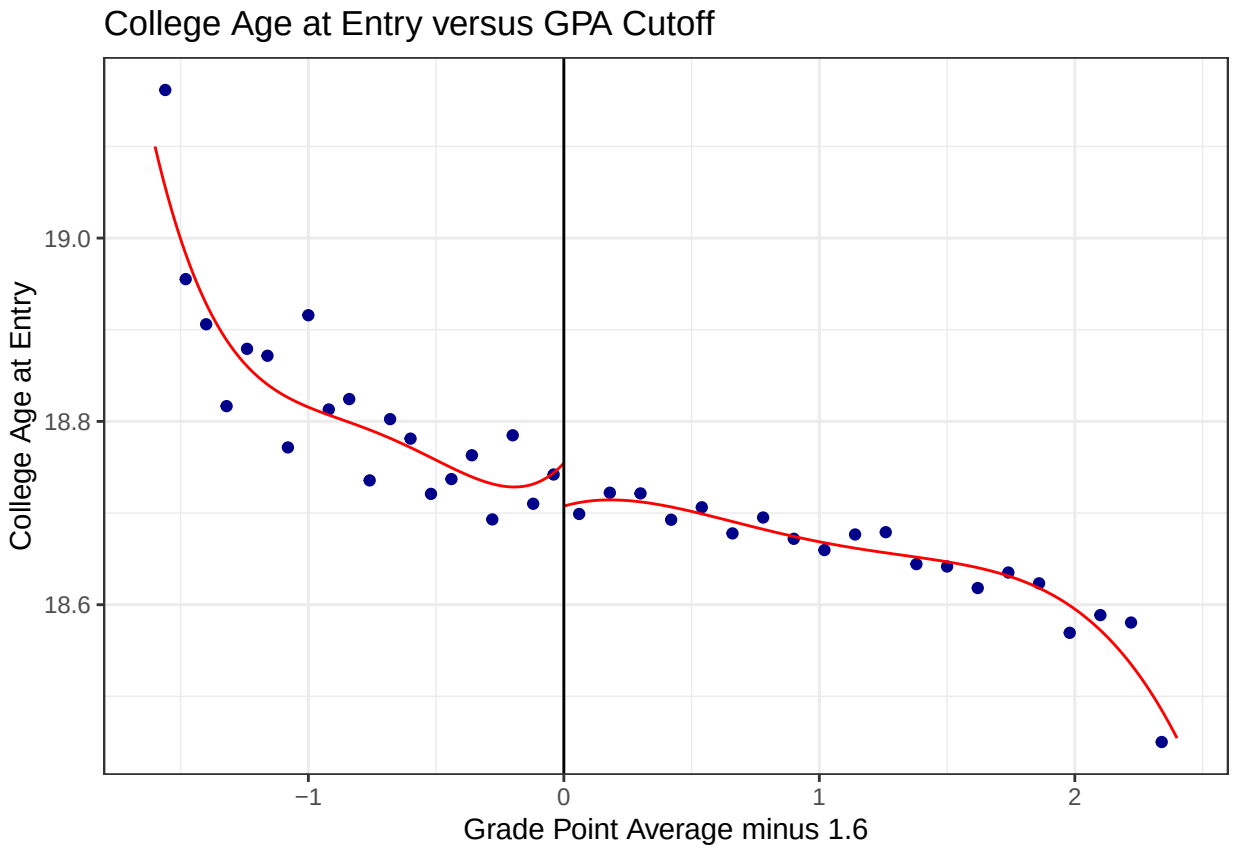



```
binned1
```

```
## Call: rdplot
##
## Number of Obs.                44362
## Kernel                      Uniform
##
## Number of Obs.                 7151           37211
## Eff. Number of Obs.            7151           37211
## Order poly. fit (p)               4               4
## BW poly. fit (h)              1.600           2.400
## Number of bins scale          1.000           1.000
```

```
#Binned scatterplot for age at entry
binned2 <- rdplot(x = df$gpa_cutoff, #outcome variable
                  y = df$age_at_entry,
                  c = 0,
                  nbins = c(20, 20),
                  binselect = "es",
                  x.label = "Grade Point Average minus 1.6",
                  y.label = "College Age at Entry",
                  title = "College Age at Entry versus GPA Cutoff" )
```

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

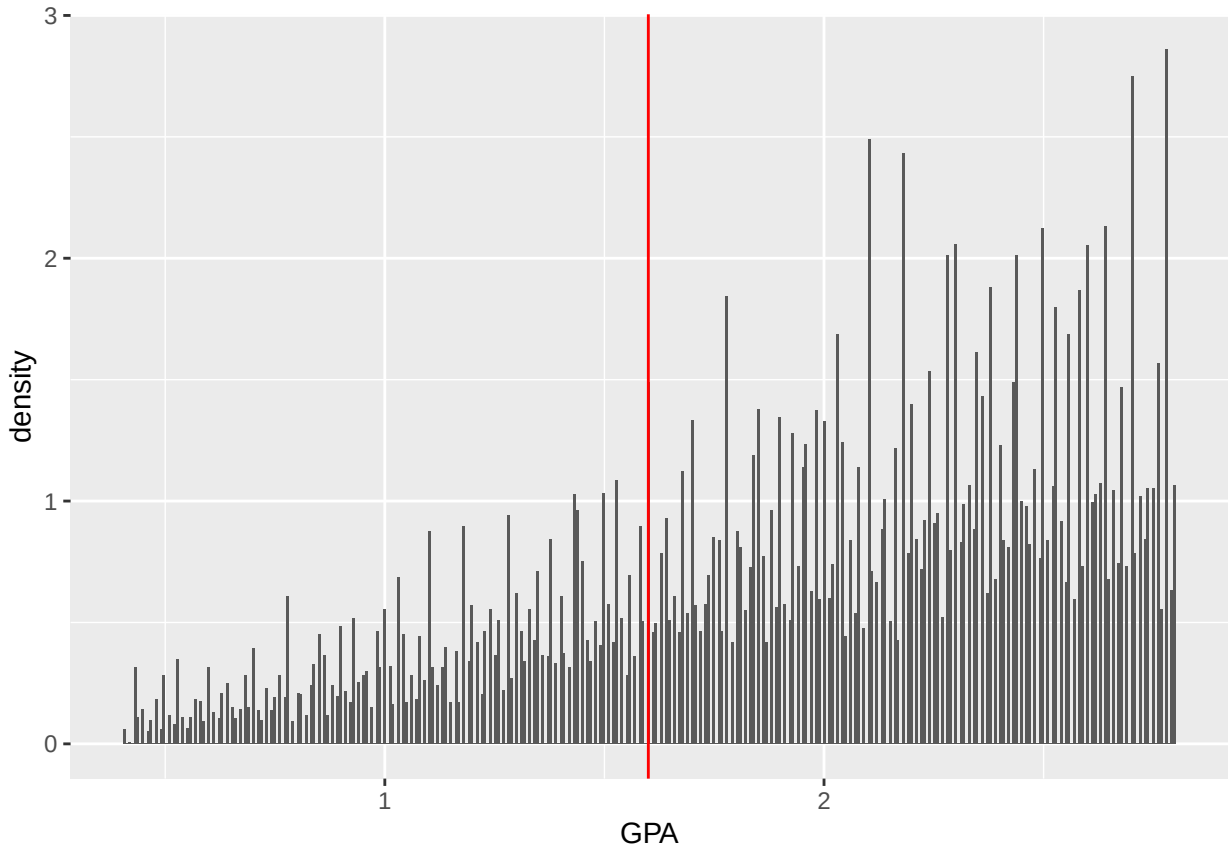



```
binned2
```

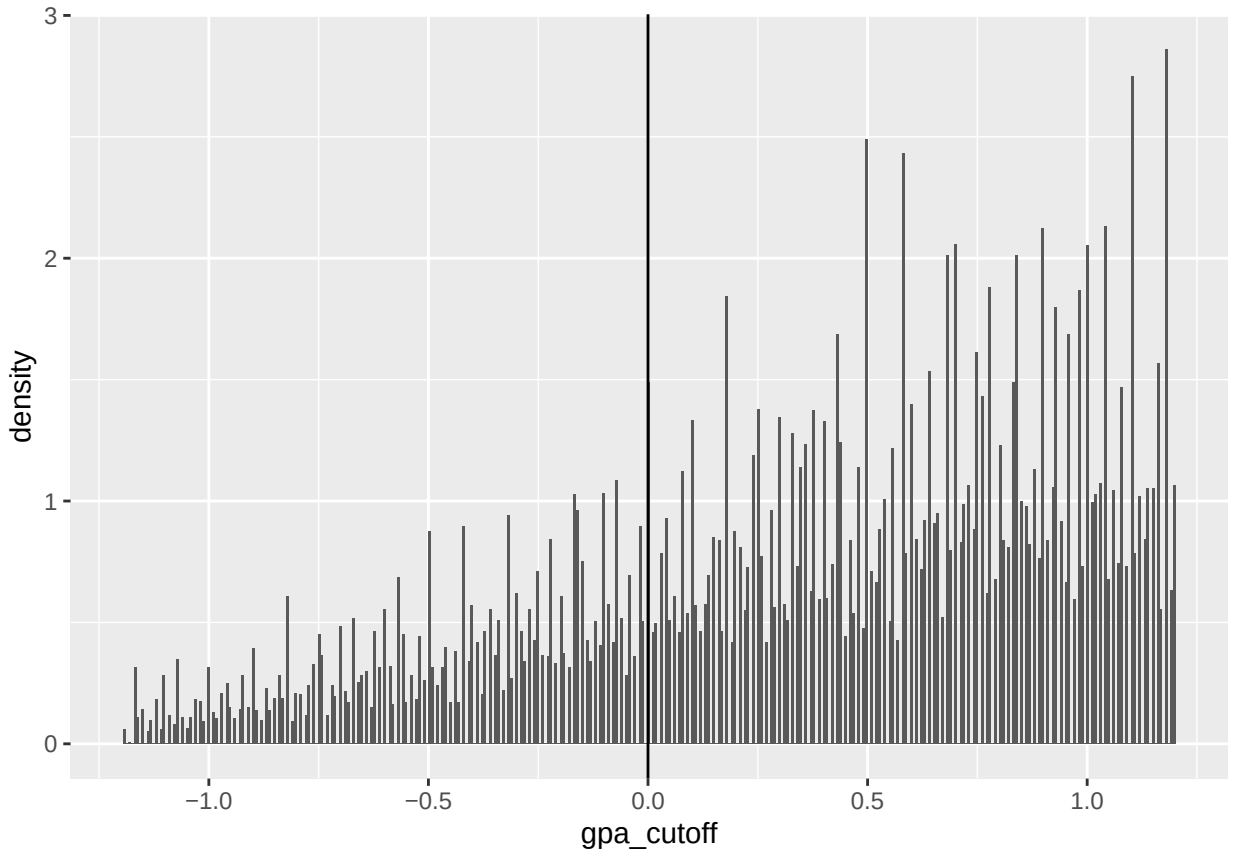
```
## Call: rdplot
##
## Number of Obs.                44362
## Kernel                      Uniform
##
## Number of Obs.                 7151          37211
## Eff. Number of Obs.            7151          37211
## Order poly. fit (p)               4              4
## BW poly. fit (h)              1.600          2.400
## Number of bins scale          1.000          1.000
```

```
#Histograms to check for a spike in density near threshold

histo <- ggplot(data = df_narrow, aes(x = GPA, y= ..density..)) +
         geom_histogram(bins = 400) +
         geom_vline(xintercept = 1.6, color = 'red')

histocenter <- ggplot(df_narrow, aes(x = gpa_cutoff, y = ..density..))+
               geom_histogram(bins = 400) + geom_vline(xintercept = 0)

histo
```

```
## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
```



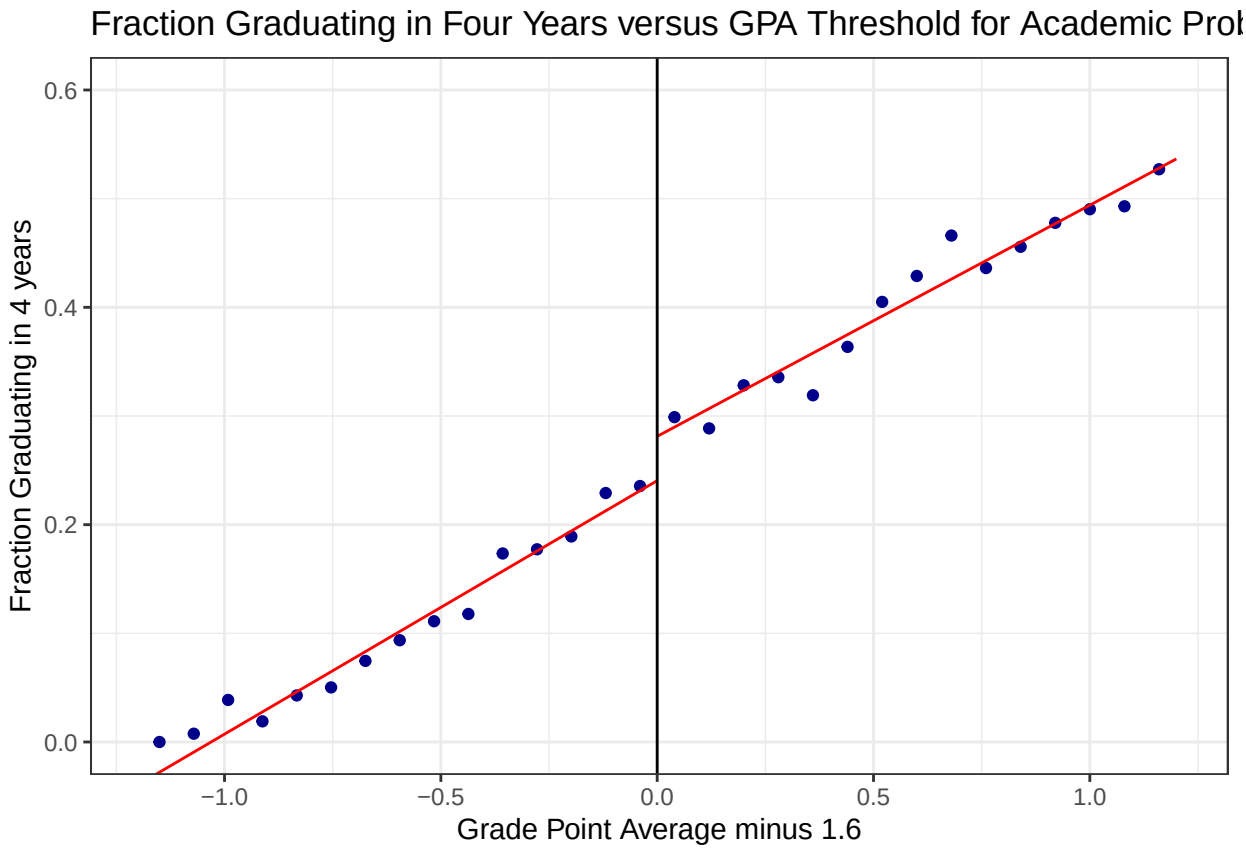

```
histocenter
```




```
#There doesn't appear to be a spike in density along the cutoff, and
#predetermined characteristics appear to be similar on either side of the threshold.
```

```
#Graphing binned scatter plot of GPA versus on-time graduation

fig1 <- rdplot(y = df_narrow$gradin4,
       x = df_narrow$gpa_cutoff,
       c = 0,
       p = 1,
       nbins = c(15, 15),
       binselect = "es",
       y.lim = c(0, 0.6),
       x.label = "Grade Point Average minus 1.6",
       y.label = "Fraction Graduating in 4 years",
       title = "Fraction Graduating in Four Years versus GPA Threshold for Academic Probation")
```

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




```
fig1
```

```
## Call: rdplot
##
## Number of Obs.                17670
## Kernel                      Uniform
##
## Number of Obs.                 4486           13184
## Eff. Number of Obs.            4486           13184
## Order poly. fit (p)               1               1
## BW poly. fit (h)              1.190           1.200
## Number of bins scale          1.000           1.000
```

```
#Question 5A: Regression of on-time grad on GPA restricted to data left of the threshold and predicted

#regression
df_left <- df_narrow |>
          filter(GPA < 1.6 & GPA >= 0.4)

reg1 <- lm(gradin4 ~ GPA, data = df_left)
summary(reg1)
```

```
##
## Call:
## lm(formula = gradin4 ~ GPA, data = df_left)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.23815 -0.18683 -0.12385 -0.03055  1.01610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.13273    0.01915   -6.93 4.82e-12 ***
## GPA          0.23326    0.01601   14.57  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3357 on 4484 degrees of freedom
##   (1881 observations deleted due to missingness)
## Multiple R-squared:  0.04521,    Adjusted R-squared:  0.045
## F-statistic: 212.3 on 1 and 4484 DF,  p-value: < 2.2e-16
```

```
#prediction
pred1 = reg1$coefficients[1] + reg1$coefficients[2]*1.6
pred1
```

```
## (Intercept)
##   0.2404832
```

```
#Question 5B: Regression of on-time grad on GPA restricted to data right of the threshold and predicted
df_right <- df_narrow |>
          filter(GPA >= 1.6 & GPA <= 2.8)

reg2 <- lm(gradin4 ~ GPA, data = df_right)
summary(reg2)
```

```
##
## Call:
## lm(formula = gradin4 ~ GPA, data = df_right)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5366 -0.4281 -0.3196  0.5400  0.7187
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.05917    0.02826  -2.094   0.0363 *
## GPA          0.21278    0.01243  17.124   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4881 on 13182 degrees of freedom
##   (6001 observations deleted due to missingness)
## Multiple R-squared:  0.02176,    Adjusted R-squared:  0.02169
## F-statistic: 293.2 on 1 and 13182 DF,  p-value: < 2.2e-16
```

```
#prediction
pred2 = reg2$coefficients[1] + reg2$coefficients[2]*1.6
pred2
```

```
## (Intercept)
##   0.2812735
```

```
#Question 5C: Calculating the difference between the two predicted values
pred1 - pred2
```

```
## (Intercept)
## -0.04079028
```

```
#Question 6: estimatign regression discontinuity through multivariate regression

#Creating above indicator
df_narrow <- df_narrow |>
             mutate(above = ifelse(GPA >= 1.6, 1, 0))
view(df_narrow)

#Creating interaction variable
df_narrow <- df_narrow |>
             mutate(interaction = above*gpa_cutoff)

#Multilinear regression
reg3 <- lm(gradin4 ~ above + gpa_cutoff + interaction, data = df_narrow)
coeftest(reg3, vcov = vcovHC(reg3, type = "HC1"))
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  0.2404832  0.0099183 24.2465 < 2.2e-16 ***
## above        0.0407903  0.0132494  3.0786  0.002083 **
## gpa_cutoff   0.2332585  0.0139726 16.6940 < 2.2e-16 ***
## interaction -0.0204812  0.0185736 -1.1027  0.270169
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#coefficient on above matches difference calculated in question 5c!
```

Question 7 The p-value of the coefficient on above is less than 0.01, indicating it is statistically significant at the 99% level.

Question 8 The college's academic probation program is not very successful, judging from these analyses. Students right above the probation threshold actually had a higher rate of graduating on time about 4 percentage points more likely to do so than students just below the threshold. There is a chance that the difference in graduation on time rates may be due to chance, as evidenced by the lower signficance level that the relevant regression coefficient displayed to the other coefficients.