Final Project

Zachary Stept

2020-05-12

Question of Interest

The question that I am proposing is if there is a relationship between acceptance rates and graduation rates at four yeor colleges. I will also be looking to see if there is a difference when it comes to private versus public school. What sparked my interest lately has been making schools test optional due to COVID-19. With this change in admissions policy, I think that it must affect how colleges provide an education. Colleges with a higher acceptance rates tend to be perceived as a school that is not as good as a school with a lower acceptance rate. This means that higher acceptance rated schools tend to have a weaker education than lower acceptance rated schools (i.e. University of Pennsylvania is better than George Mason University). Also, people tend to say that private schools have a better education because they receive more funding and are smaller than public schools. In order to analyze this question I will be using the ADM_RATE, C150_4, and CONTROL variables in a linear model to see what the relationship is between these three variables.

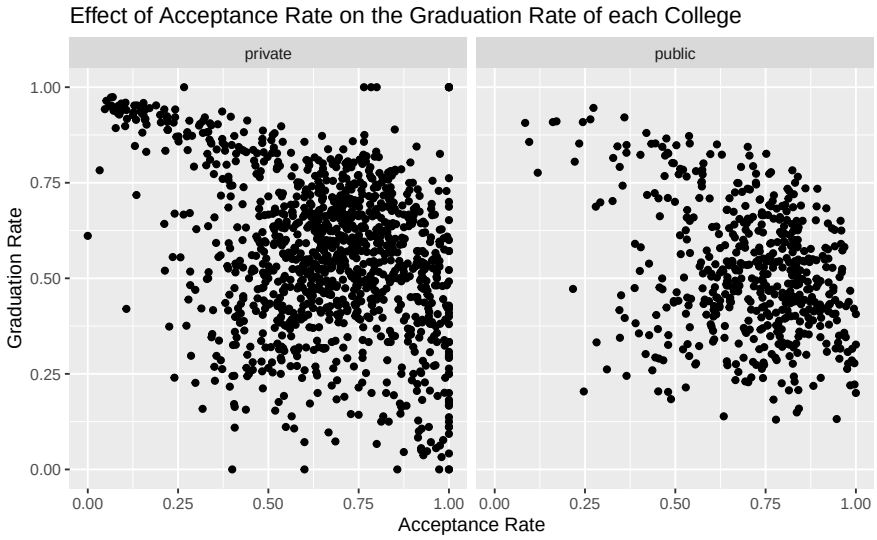
Preprocessing

```
college_reduced <- college %>%
  select(ADM_RATE, C150_4, CONTROL) %>%
  rename(acceptance_rate = ADM_RATE, graduation_rate = C150_4, school_type = CONTROL) %>%
  mutate(school_type = recode(school_type, '1' = "public", '2' = "private", '3' = "private")) %
  na.omit()
```

In order to minimize the size of the dataset, I filtered my two columns that are my explanatory and response variables. Then, I renamed the variables to their full length names to better understand what the variables are. Finally, I omitted all rows that have NA because they provide no information that helps me to answer my question.

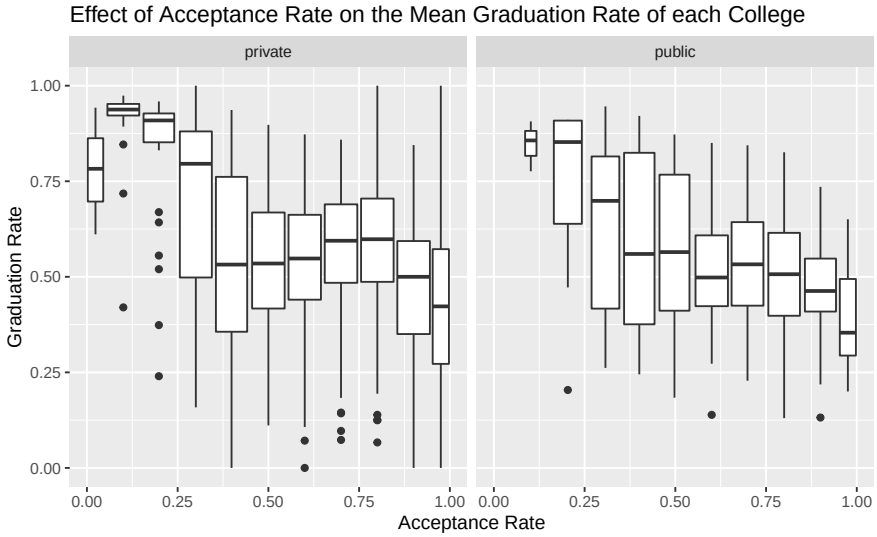
Visualization

```
ggplot(data = college_reduced) +
  geom_point(mapping = aes(x = acceptance_rate, y = graduation_rate)) +
  facet_wrap(~ school_type) +
  labs(x = "Acceptance Rate", y = "Graduation Rate", title = "Effect of Acceptance Rate on the
```



This scatter plot shows the acceptance rate and graduation rate of each college in their respective school type category. I used this graph because it shows the whole picture of where each college stands on acceptance rate and graduation rate. It then helps to compare that college to the ather colleges plotted. In this graph, we can see that their is a relationship where the graduation rate is decreasing as the acceptance rate is increasing. You can also tell that there is a more defined decreasing curve for public schools while private schools have a cluster in the center. This graph proves that the higher the acceptance rate, the lower the graduation rate, but only for public schools.

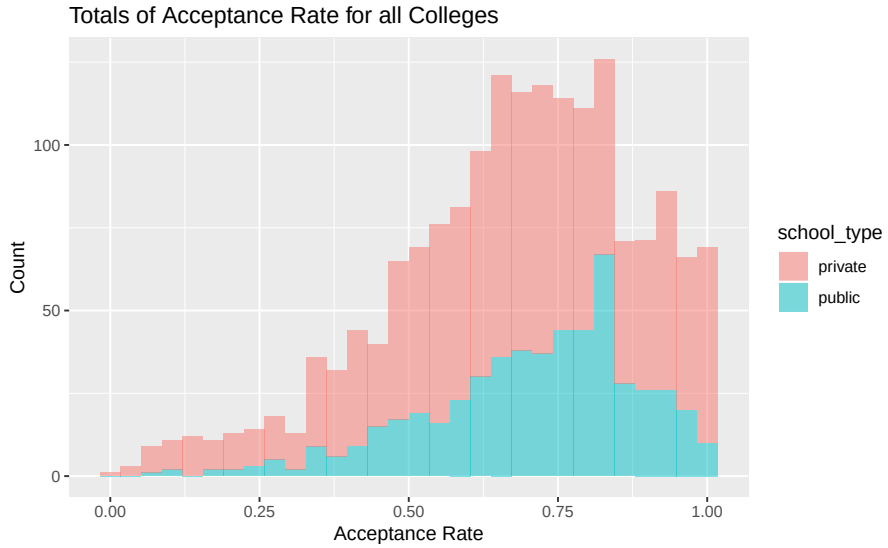
```
ggplot(data = college_reduced, mapping = aes(x = acceptance_rate, y = graduation_rate)) +
  geom_boxplot(mapping = aes(group = cut_width(acceptance_rate, 0.1))) +
  facet_wrap(~ school_type) +
  labs(x = "Acceptance Rate", y = "Graduation Rate", title = "Effect of Acceptance Rate on the
```



In this graph, we plotted 11 boxplots by splitting acceptance rate using a width of .1 and plotting the boxplots in their respective school type category. I used this graph because it shows a trend in the mean graduation rate for the mean acceptance rate which helps to visualize the curve of the data. In this graph, we can observe that the mean graduation rate decreases as the acceptance rate increases. This graph proves that the higher the acceptance rate, the lower the graduation rate.

```
ggplot(data = college_reduced) +
  geom_histogram(mapping = aes(x = acceptance_rate, fill = school_type), alpha = .5) +
  labs(x = "Acceptance Rate", y = "Count", title = "Totals of Acceptance Rate for all Colleges"
```

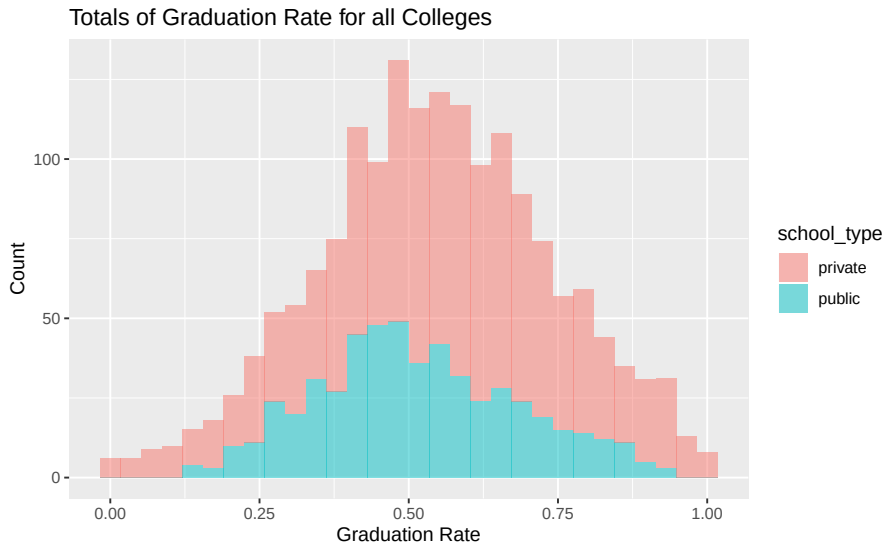
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



In this histogram, we count the number of each acceptance rate and graph them in order to see the distribution. We also stack the school type on top of each other to see the difference in count. In this histogram, we can tell that there are more private school than public schools in the dataset as well as that they are both left skewed graphs.

```
ggplot(data = college_reduced) +
  geom_histogram(mapping = aes(x = graduation_rate, fill = school_type), alpha = .5) +
  labs(x = "Graduation Rate", y = "Count", title = "Totals of Graduation Rate for all Colleges"
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



In this histogram, we count the number of each graduation rate and graph them in order to see the

distribution. We also stack the school type on top of each other to see the difference in count. In this histogram, we can tell that there are more private school than public schools in the dataset as well as that they are both symmetric graphs.

Summary Statistics

```
college_reduced %>%
  group_by(school_type) %>%
  summarize(
    count = n(),
    mean = mean(acceptance_rate),
    median = median(acceptance_rate),
    sd = sd(acceptance_rate),
    iqr = IQR(acceptance_rate),
    min = min(acceptance_rate),
    max = max(acceptance_rate)
  )
```

school_type	count	mean	median	sd	iqr	min	max
private	1178	0.6618590	0.67935	0.2166689	0.28935	0.0000	1
public	537	0.7076007	0.74000	0.1766512	0.22920	0.0844	1

We can get the summary statistics of acceptance rates by using the functions of each summary statistic. We can see that public schools have a higher mean, median, and minimum acceptance rates while private schools have a higher standard deviation and interquartile range for acceptance rates.

```
college_reduced %>%
  group_by(school_type) %>%
  summarize(
    count = n(),
    mean = mean(graduation_rate),
    median = median(graduation_rate),
    sd = sd(graduation_rate),
    iqr = IQR(graduation_rate),
    min = min(graduation_rate),
    max = max(graduation_rate)
  )
```

school_type	count	mean	median	sd	iqr	min	max
private	1178	0.5573767	0.5652	0.2052233	0.268475	0.0000	1.0000
public	537	0.5144555	0.4992	0.1687654	0.233500	0.1302	0.9458

We can get the summary statistics of graduation rates by using the functions of each summary statistic. We can see that public schools have a higher mimimum graduation rates while private schools have a higher mean, median, standard deviation, interquartile range, and maximum for

graduation rates.

Data Analysis

```
college_model_1 <- lm(graduation_rate ~ acceptance_rate + school_type, data = college_reduced)
college_model_1 %>%
  tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.7703112	0.0152464	50.524184	0.0000000
acceptance_rate	-0.3217217	0.0215797	-14.908507	0.0000000
school_typepublic	-0.0282051	0.0095839	-2.942964	0.0032946

```
college_model_1 %>%
  glance()
```

r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df.re
0.1240885	0.1230652	0.183086	121.2676	0	3	479.7475	-951.495	-929.7064	57.38707	

```
college_model_2 <- lm(graduation_rate ~ acceptance_rate * school_type, data = college_reduced)
college_model_2 %>%
  tidy()
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.7754449	0.0171553	45.2015897	0.0000000
acceptance_rate	-0.3294783	0.0246345	-13.3746942	0.0000000
school_typepublic	-0.0514703	0.0368852	-1.3954201	0.1630703
acceptance_rate:school_typepublic	0.0333804	0.0511038	0.6531882	0.5137227

```
college_model_2 %>%
  glance()
```

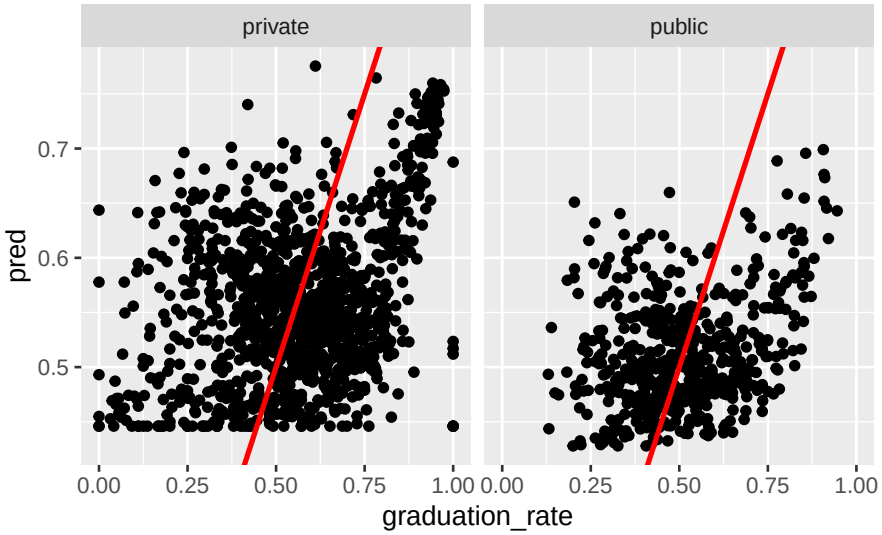
r.squared	adj.r.squared	sigma	statistic	p.value	df	logLik	AIC	BIC	deviance	df
0.1243068	0.1227714	0.1831167	80.96023	0	4	479.9613	-949.9226	-922.6868	57.37276	

I created two model to determine whether the model is better with interacting variables or predictor variables. From looking at the r squared values for both the model using glance(), we can see that the second model is better when it comes to capturing the variability of the response variable because it is larger than the value for the first model. However, since the value is not near 1, it means that it is not a good model. The tidy() function function tells us the intercept and slope of each variable. From looking at the slopes, we can tell that they are decreasing which is what we observed for our plots/graphs in the visualization section.

```
college_reduced_df <- college_reduced %>%
  add_predictions(college_model_2) %>%
  add_residuals(college_model_2)
```

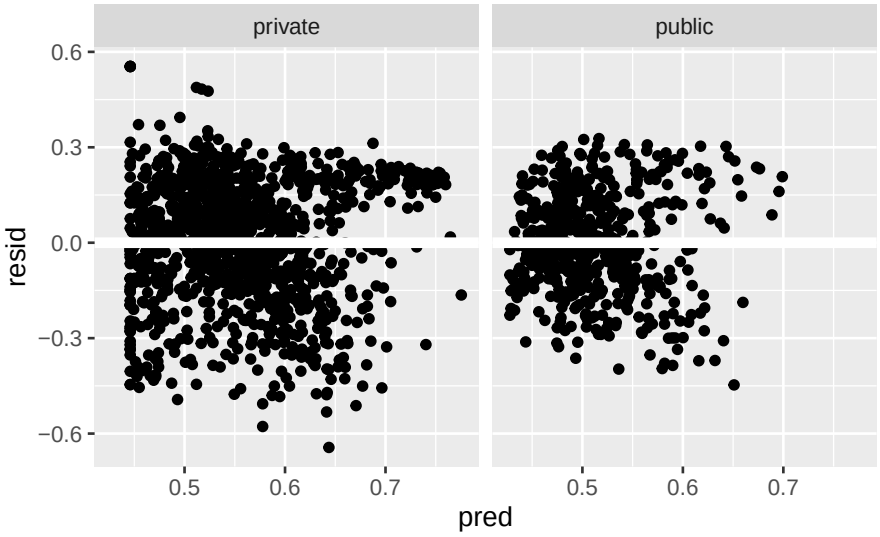
Since the second model is the one I will be using to analyze, we create a predictions and residuals for the following graphs.

```
ggplot(data = college_reduced_df) +
  geom_point(mapping = aes(graduation_rate, pred)) +
  geom_abline(
    slope = 1,
    intercept = 0,
    color = "red",
    size = 1
  )  +
  facet_wrap(~ school_type)
```



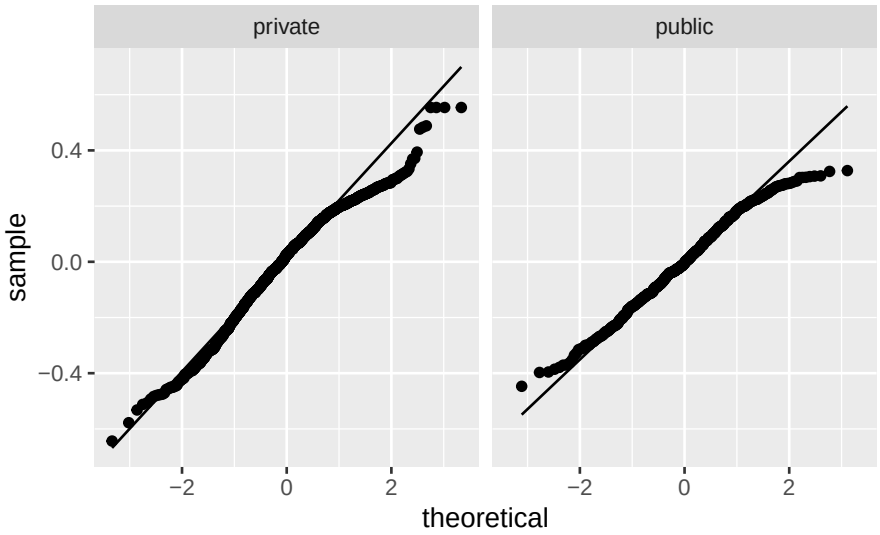
These graphs test the linearilty for the relationship between the explanatory variable and the response variable. This model fails the test because there is no linear relationship present on either graphs between the two variables. Instead, there is just one big clump of data points with others scattered around for both graphs.

```
ggplot(data = college_reduced_df) +
  geom_point(aes(pred, resid)) +
  geom_ref_line(h = 0) +
  facet_wrap(~ school_type)
```



These graphs test the constant variablity for the variablity of points around the model line. This model fails the test for private schools, but not for public schools. It fails for private schools because there is no even spread of data points above and below the model line. Instead, there is a large clump of data points above the line and more scattered data points below the line. Also, the data points on the bottom are further from the model line compared to the data points above the line. However, the public schools is pretty evenly spread on both the top and bottom of the model line.

```
ggplot(data = college_reduced_df) +
  geom_qq(mapping = aes(sample = resid)) +
  geom_qq_line(mapping = aes(sample = resid)) +
  facet_wrap(~ school_type)
```



These graphs test the nearly normal residuals to determine a bell shaped curve. This model passes the test because most of the data points are on the reference lines. Even though, the data points fall off the refernce line towards the top right, the rest of the data poinst make that a minor issue.

Conclusion

After analyzing this model, I can conclude that the acceptance rates of schools affect the grauduation rates of those schools because it was able to pass two of the three tests. This relationship between acceptance rates and graduation rates is more apparent in four year public schools. However, this model is not strong enough to fully prove the relationship because of its extremely low r squared value. If you were just looking at the visualization section, you can tell that the relationship is still more adherent in public schools because there is a smaller range between accpentance rates and graudation rates as well as less outliers in the boxplots. Also, there is a more defined curve in the scatter plot. From looking at the summary statistics section, you can tell that there is less variation in public schools because its standard deviation is smaller than private schools which means its data points are closer to the mean. It also means there are less variance in the outcome of the data making the trends more distinct. To conlcude, four year public colleges with higher acceptance rates tend to have a lower graudation rate. If you are looking at colleges it is best to go to a four year public school with a low acceptance rate or you can take a risk with a four year private school.