The Mission to Find the Conditions for Admission

Group C: Jonah Botvinick-Grenhouse, Grace Cho, Maya Mizrahi, Kitty Girjau

5-13-19

Abstract

In our mission to discover what factors influence the criteria for acceptance into American colleges and universities, we evaluated the ways in which a school's average cost of attendance, average SAT score, regional location, and institution type predicted admission rates. With data obtained from College Scorecard, we used R to compile a random sample of 200 colleges and universities in the United States in 2010, from which we were able to obtain values for our predictor variables. Using this data, we viewed individual distributions of our quantitative variables, and considered the ways in which they varied when split into categorical factors. Fortunately, we found that our quantitative variables were normally distributed. We then proceeded to conduct simple linear regression and multiple linear regression, to attempt to discover the degree to which our predictor variables could account for the variability in admission rates of various colleges and universities. After eliminating terms that were not statistically significant from our final multiple regression model, we were able to account for about 21% of the variability in admission rate by knowing the values of our predictor variables. Although this model is helpful in predicting admission rates, there are clearly other factors that are influencing college admissions, such as an applicant's GPA or extracurricular involvements.

Introduction

Have you ever wondered exactly what got you into Amherst? Well, that's what we're trying to find out. Our project, "The Mission to Find the Conditions for Admissions," looks to discover what factors most strongly impact college admissions rates. Using four aformentioned predictor variables, we attempted to see how each of these predictor variables affected admission rates in both simple linear regression models and multiple linear regression models. This topic intrested us, as college admissions has not only played a big role in our lives, but has also been a large topic of discussion after a recent epidemic of college admission scandals.

After checking the conditions for regression, we continued to conduct simple linear regression on admission rate, as predicted by average SAT score, and later by average cost of attendance. The latter of these two graphs produced an R-squared value of around zero and a high p-value, so our initial model did not produce a meaningful correlation between average cost of attendance and admission rate, which is rather interesting. Moreover, the regression of admission rate, as predicted by average SAT score, produced an R-squared value of about 6%, with statistically significant results. Although this is a small portion of variability that we can account for in admissions, it was a good start that paved our way to analyzing various multiple regression models.

In conducting multiple regression, there were a multitude of possible combinations for graphs we could analyze, seeing as we had so many predictor variables. With the overwhelming amount of possibilities, we began with a model that included every possible predictor and every possible interaction term. From there, we removed the statistically insignificant interactions and predictors. Interestingly enough, we found that region was not a meaningful predictor, as it only accounted for about .3% of the variability in admissions after taking every other variable and interaction term into account. Moreover, in this final model average cost of attendance did turn out to yield statistically significant results, which was surprising given our statistically insignificant simple linear regression model with cost of attendance as a predictor. We will further analyze our multiple regression models when we discuss diagnostics, as there are several graphs worth spending time on. Ultimately, we were able to account for about 21% of the variability in admisisons by knowing the values of our previously stated predictors. Our research helped us gain further insight into the world of college admissions, but we realized that there are many more factors that influence college admissions, several of which cannot be quantified. For this reason, our model has clear limitations and realistically cannot be used to determine the true admission rate of a college or university. Even though there is much work left to be done, our group has taken important initial strides in creating a model that accurately predicts admission rates of schools in the United States.

Data:

In discussing how we worked with our data, we will first list the variables that we used in our models, the ranges that they took in our sample, and their purpose in our project. Then, we will walk through our process of compiling data, as well as how we aimed to use the data in our analysis of college admission rates.

Variables:

Admission Rate (ADM_RATE)

This is a quantitative response variable which is the average admission rate for each institution in the year 2010, given in terms of percentage, which in our sample takes values between 10-100%.

Cost of Attendance (COSTT4_A)

Cost of attendance is a quantitative predictor variable which was the average annual cost in dollars per year of attending the each institution. This number is the average cost of attendance for students currently enrolled in each institution, taking into account tuition, fees, books, supplies, and living expenses for full-time, first-time, degree/certificate-seeking undergraduates. This number is specifically for students who receive Title IV aid, which is federal grants, loans, and work-study programs, and the number in particularly for academic year institutions. In our sample, values range from $8391 to $53660, and this number was specifically for the year 2010.

Average SAT Score (SAT_AVG_ALL)

Average SAT score is a quantitative predictor variable. It is the average SAT score of all students currently enrolled in each institution on all campuses in the year 2010, rolled up to the six-digit OPE ID. Our sample contains values between 780 and 1500.

Region (REGION):

Region is a categorical predictor variable which identifies regions within the United states, which we separated into ten different categories: 0) U.S. Service Schools (IPEDS), 1) New England (CT, ME, MA, NH, RI, VT), 2) Mid East (DE, DC, MD, NJ, NY, PA), 3)Great Lakes (IL, IN, MI, OH, WI), 4) Plains (IA, KS, MN, MO, NE, ND, SD), 5) Southeast (AL, AR, FL, GA, KY, LA, MS, NC,SC, TN, VA, WV) , 6) Southwest (AZ, NM, OK, TX), 7) Rocky Mountains (CO, ID, MT, UT, WY), 8) Far West (AK, CA, HI, NV, OR, WA), and 9) Outlying Areas (AS, FM, GU, MH, MP, PR, PW, VI).

Type of Institution (CONTROL):

The type of institution is a categorical predictor variable. There were three types of institutions which were present in the data set. 1 was used to denote public institutions, 2 was used for private nonprofit institutions, and 3 was used for private for profit institutions.

Compiling Data:

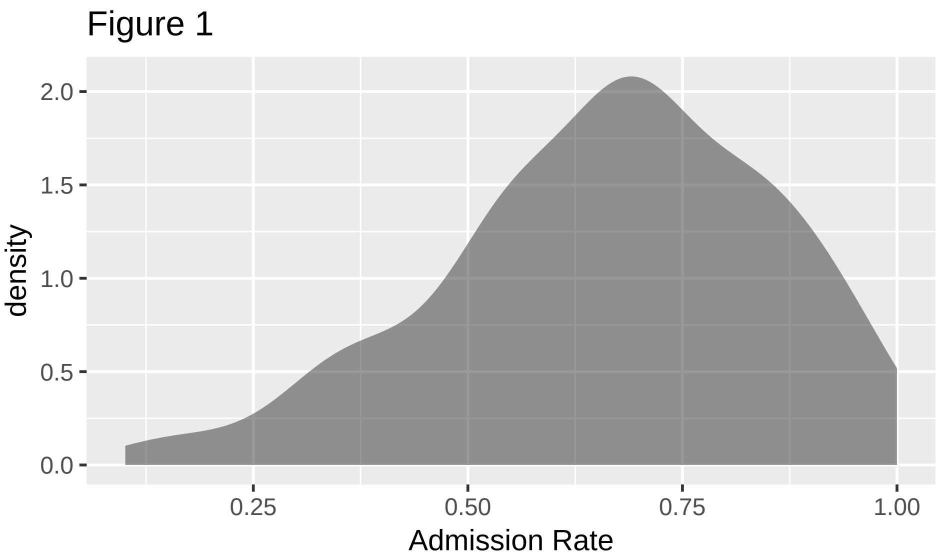
In compiling our data, we used a data set from CollegeScorecard (https://collegescorecard.ed.gov/) with over 1800 colleges and universities in the United States from 2010 to select a random sample of 200 schools. We used R to compile such a sample and then removed every variable from the data set, aside from the ones of interest. We predicted that college admissions rates would decrease as regional population density increases and SAT and average cost of enrollment increase, and that admission rates would differ depending on the type of institution.

We will now begin our univariate analysis of the data, starting by loading the random sample of 200 colleges into R.

```r
finalsample2 <- read_csv('finalsample2.csv')
```

We began our analysis by examining the distributions of the variables with which we are working. First, we looked at admission rate (Figure 1) and found that the data was relatively unimodal and symmetric with a slight left skew. Likely, this skew is due to a few outlying schools that are unusually selective. The median of the distribution was an admission rate of 68% with an IQR of about 27%.

```r
gf_density(~ADM_RATE, data = finalsample2, xlab = "Admission Rate") + labs(title = "Figure 1")
```



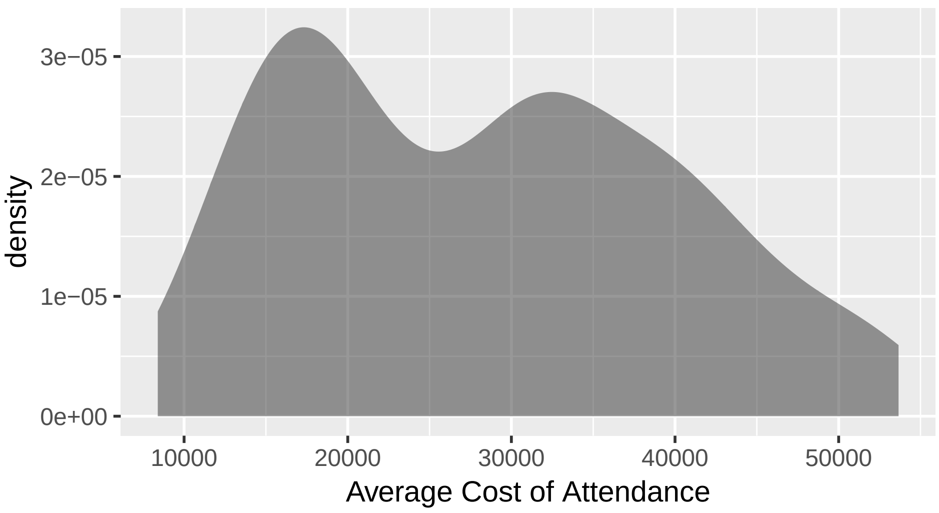
```r
favstats(~ADM_RATE, data = finalsample2)
```

```
##     min       Q1  median      Q3 max      mean        sd   n missing
##  0.1008 0.540425 0.68165 0.81085   1 0.6632095 0.1911175 200       0
```

Next, we proceeded to examine the average cost of attendance, which took on a bimodal distribution (Figure 2). This is likely due to the fact that our data was sampled from both public and private institutions. The mode most likely generated by public schools was centered at about 17k dollars, while the mode corresponding to the private institutions was centered at about 33k dollars. The public school mode has a higher frequency, as there were more public schools in our random sample than private schools.

```r
gf_density(~COSTT4_A, data = finalsample2, xlab = "Average Cost of Attendance") + labs(title = "Figure
```

Figure 2



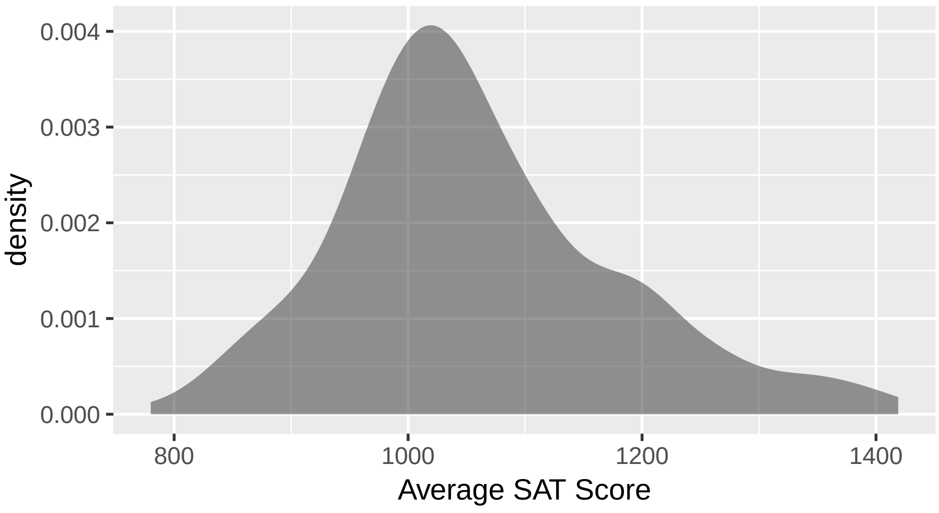
```
favstats(~COSTT4_A, data = finalsample2)
```

```
##   min       Q1  median      Q3   max     mean       sd   n missing
##  8391 17856.25 28847.5 37432.5 53660 28462.35 11713.66 200       0
```

Finally, we examined the distribution of SAT scores among the schools in our sample (Figure 3). The SAT exam is designed to have a normal distribution across the nation, so the fact that our sample of average SAT scores from 200 colleges also follows a normal distribution is unsurprising. The Central Limit Theorem also predicts this to be the case, as we are resampling means of a quantitative variable. The mean SAT score from our sample was 1060 with a standard deviation of about 120 points, which is well within one standard deviation of the mean SAT score from 2010.

```
gf_density(~SAT_AVG_ALL, data = finalsample2, xlab = "Average SAT Score") + labs(title = "Figure 3")
```

Figure 3



```
favstats(~SAT_AVG_ALL, data = finalsample2)
```

```
##  min  Q1 median      Q3  max     mean       sd   n missing
##  780 979   1038 1124.25 1419 1059.015 119.9774 200       0
```

Results

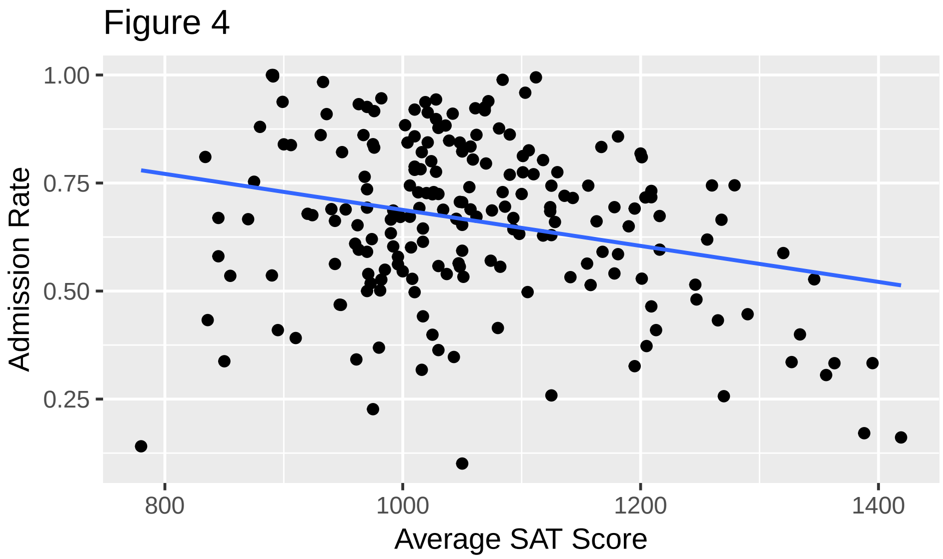
To familiarize yourself with our final multiple regression model, refer to the diagnostics section of the report. Here we will give you the main take aways from that model and their implications for our project's goals. Our final multiple regression model allowed us to account for 21% of the variability in admission rates by knowing a school's average cost of attendance, its average SAT score, and its type. With a p-value of 4.2e-9, we reject the null hypothesis that there is no correlation between these factors and admission rate. It is important to note, though, that we removed region as a predictor from our final multiple regression, as it yielded statistically insignificant results. Doing so only resulted in a difference in R-squared of about .3% and lowered the p-value of our model significantly, so we preferred the multiple regression model without region. When we began our analysis with simple linear regression, we were unable to find a meaningful correlation between admission rate and average cost of attendance. With an R-squared value of essentially zero and a p-value above .05, we thought that cost of attendance would not play an important role in our multiple regression model. However, cost of attendance did turn out to be a statistically significant predictor, with a p-value of order less than 10^-5, so the only initial predictor that we ended up removing from our final multiple regression was region.

Now, let's put this final multiple regression model into context by discussing some of the coefficients. Given that region was an indicator with 9 levels, excluding it as an explanatory variable simplifies our regression equation significantly, which is another advantage of this final model. The coefficients for the explanatory variables we kept are quite small, of orders as low as 10^-5, meaning that, after accounting for the effects of all other variables, no single predictor has a very big influence on admission rates. For example, if all other variables are accounted for and the average SAT score of a school increased by about 100 points, we would expect that school's admission rate to change by anywhere between 7% and 19%, depending on the school type. We would find a 7% change for a public school, while both types of private schools would experience a change closer to 19%. With our base level for school type being public, the coefficients for the interaction terms SAT*school-type are positive, meaning that at private schools (both for profit and not for profit), better SAT scores are generally associated with a higher admission rate. Overall, the model's p-value tells us that our results are statistically significant, but an analysis of the coefficients in our regression equation suggests that the difference these predictors make in admission rates is not very practically significant, especially with a relatively large residual standard error of 17%. So while the explanatory variables we have picked are reasonably associated with admission rates and can account for some variability therein, it is probably not worth it to use our model for any practical purposes such as estimating one's chance to be accepted in a specific school.

Diagnostics:

We will now continue our analysis of admission rates with a bivariate analysis of our quantitative variables. We will first use simple linear regression models to determine whether it would be feasible to run a multiple regression model with our set of predictor variables. Predicting admission rate by average SAT score yielded a moderately linear scatter plot with moderate negative correlation (Figure 4). The R-squared value for this regression was .07, which means we were able to account for about 7% of the variability in college admission rates, just by knowing the average SAT score of a school. Moreover, the residual standard error was .18 on 198 degrees of freedom, while the p-value was .0002, which means that these results are statistically significant. The histogram of the residuals is relatively normal (Figure 5), although it does have a slight left skew. The fitted scatterplot of the residuals also seems to show close to no pattern (Figure 6), only with slight sparicty of points on the left side of the graph. Finally, the qq plot is linear (Figure 7), so running simple linear regression on this model was appropriate. It is worth noting that there was one outlier in our model, which can be found at the bottom left corner of the regression plot. Victory University turns out to have an average SAT score below 800, but an admission rate less than 25%, which makes the slope of the regression slightly less negative than it would have been otherwise, as this is a high leverage point. The coefficient of the slope tells us that for every 100 point decrease in the average SAT score of a school, our model would predict that its admission rate would lower by about 4%.

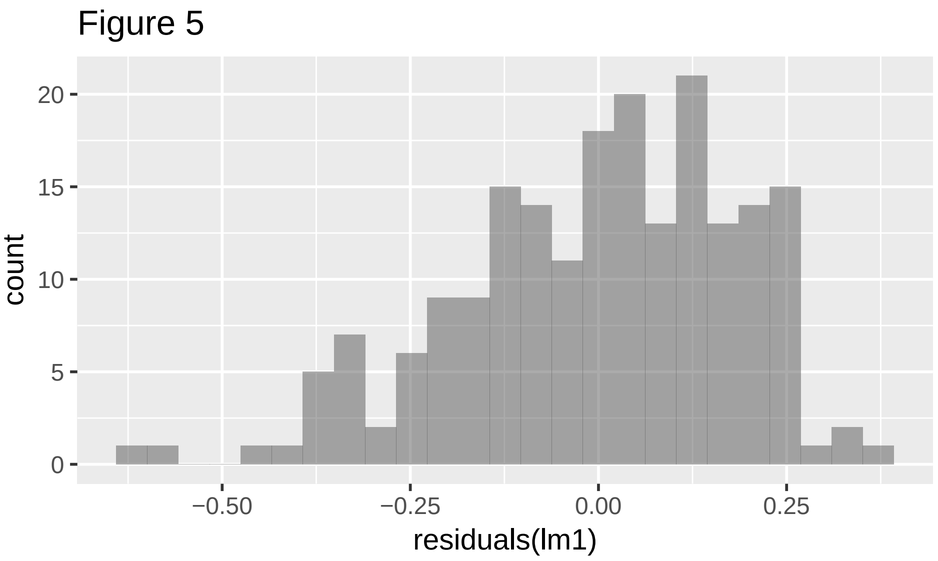
```
gf_point(ADM_RATE ~ SAT_AVG_ALL, data = finalsample2, xlab = "Average SAT Score", ylab = "Admission Rat
  gf_lm()
```



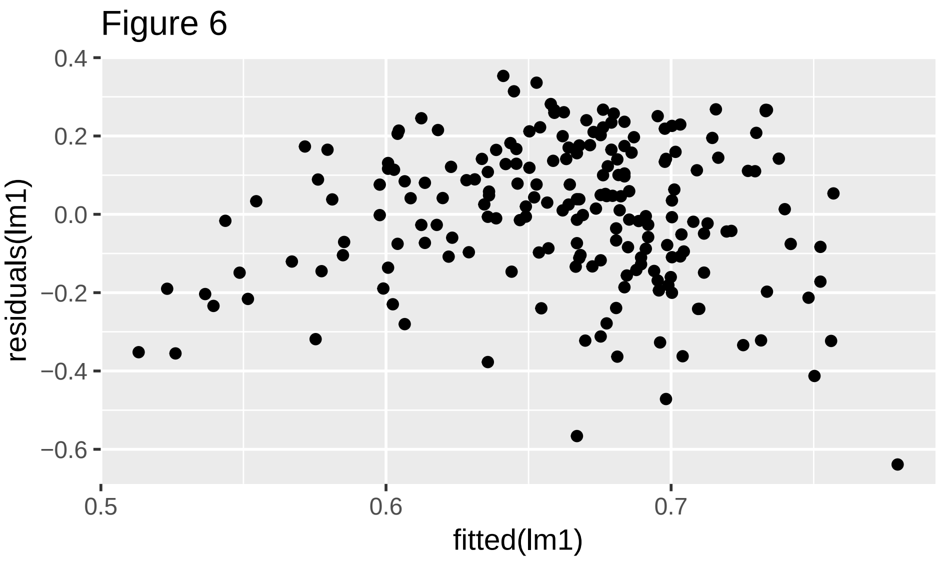
```
lm1 <- lm(ADM_RATE ~ SAT_AVG_ALL, data = finalsample2)
summary(lm1)
```

```
##
## Call:
## lm(formula = ADM_RATE ~ SAT_AVG_ALL, data = finalsample2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6388 -0.1129  0.0225  0.1412  0.3534
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.1043886  0.1164506   9.484  < 2e-16 ***
## SAT_AVG_ALL -0.0004166  0.0001093  -3.813 0.000184 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1849 on 198 degrees of freedom
## Multiple R-squared:  0.06839,    Adjusted R-squared:  0.06369
## F-statistic: 14.54 on 1 and 198 DF,  p-value: 0.0001836
```

```
gf_histogram(~ residuals(lm1), title = "Figure 5")
```

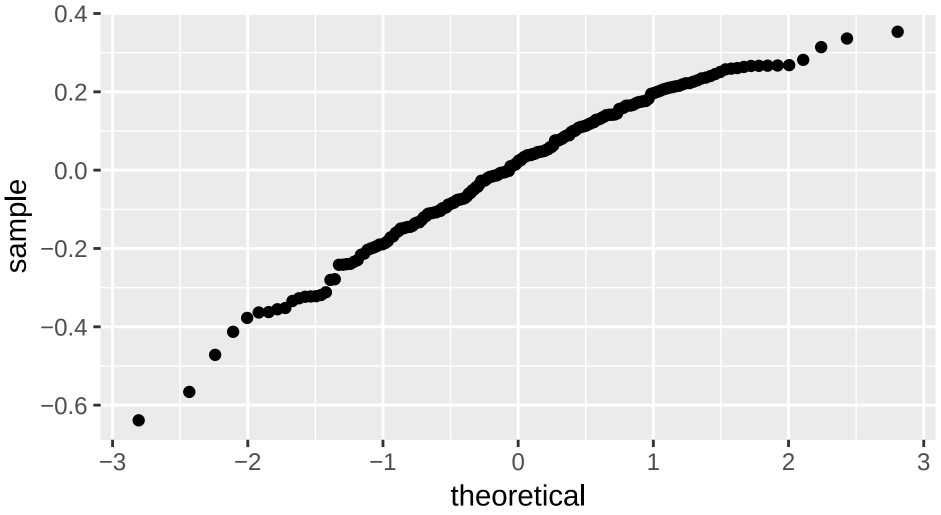


```r
gf_point(residuals(lm1) ~ fitted(lm1), title = "Figure 6")
```



```r
gf_qq(~resid(lm1), title = "Figure 7")
```

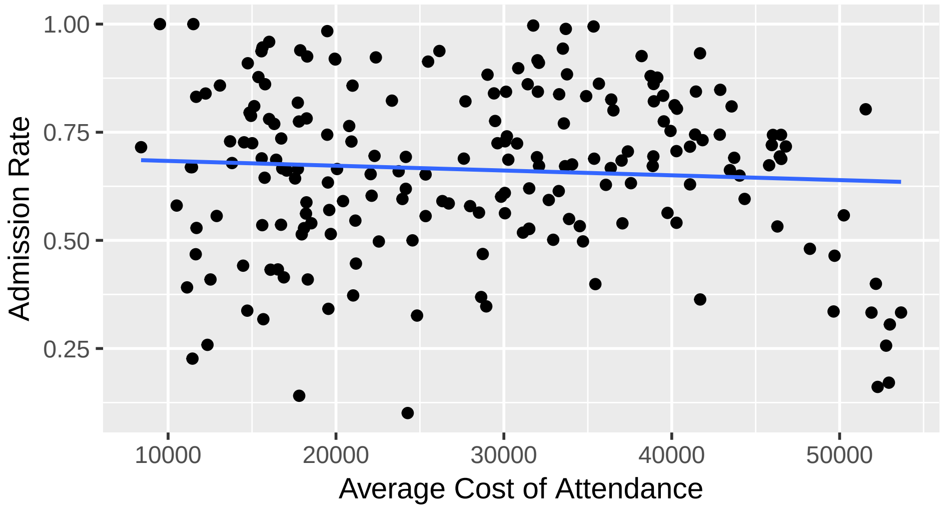
Figure 7



Next, we ran a linear regression predicting admission rate by average cost of attendance (Figure 8), our second quantitative predictor. The scatterplot showed essentially no pattern, the slope of the regression line is essentially zero, and the p-value is .33, far greater than alpha (0.05), indicating that our results are not statistically significant. The histogram of the residuals was relatively normal and the qq plot was fairly linear (Figures 9 and 11). However, the fitted scatterplot of the residuals was slighly heteroscedastic, but not overwhelmingly so (Figure 10). The diagnostics tell us that regression would have been appropriate to run, but with such a high p-value and an R-squared value close to zero, we were unable to find a meaningful correlation between admission rate and average cost of attendance. However, the validity of these diagnostics does tell us that we will eventually be able to run multiple regression with average cost of attendance as a predictor.

```
gf_point(ADM_RATE ~ COSTT4_A, data = finalsample2, xlab = "Average Cost of Attendance", ylab = "Admissi
  gf_lm()
```

Figure 8

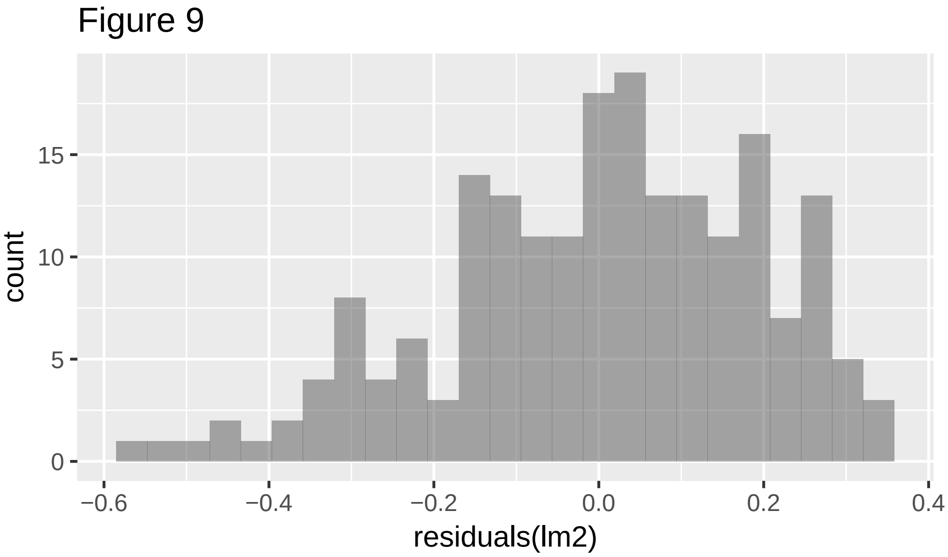


```
lm2 <- lm(ADM_RATE ~ COSTT4_A, data = finalsample2)
summary(lm2)
```

```
##
## Call:
```

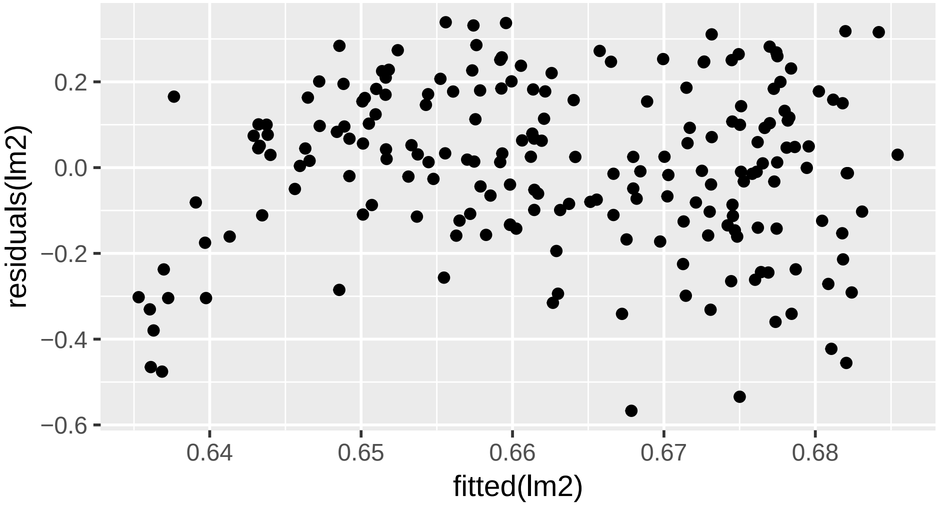
```
## lm(formula = ADM_RATE ~ COSTT4_A, data = finalsample2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5671 -0.1166  0.0194  0.1542  0.3389
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.947e-01  3.559e-02  19.519   <2e-16 ***
## COSTT4_A    -1.107e-06  1.157e-06  -0.957     0.34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1912 on 198 degrees of freedom
## Multiple R-squared:  0.004606,   Adjusted R-squared:  -0.000421
## F-statistic: 0.9163 on 1 and 198 DF,  p-value: 0.3396
```

```
gf_histogram(~ residuals(lm2), title = "Figure 9")
```



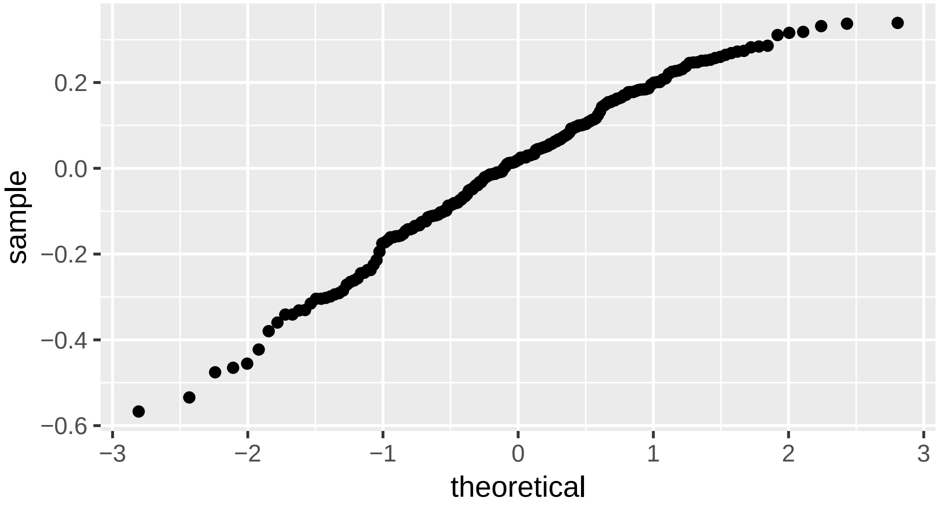
```
gf_point(residuals(lm2) ~ fitted(lm2), title = "Figure 10")
```

Figure 10



```
gf_qq(~resid(lm2), title = "Figure 11")
```

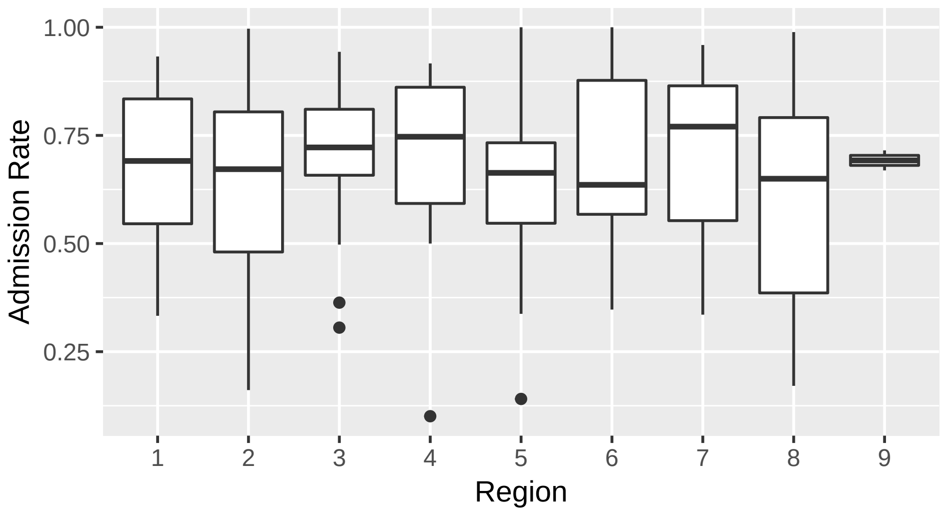
Figure 11



We then moved onto a categorical analysis of admission rate, classifying it in terms of both region and school type. Admission rate turned out to vary little by region (Figure 12). Moreover, nonprofit private schools and public schools had similar admission rates, while private schools for profit generally had lower admission rates (Figure 13).

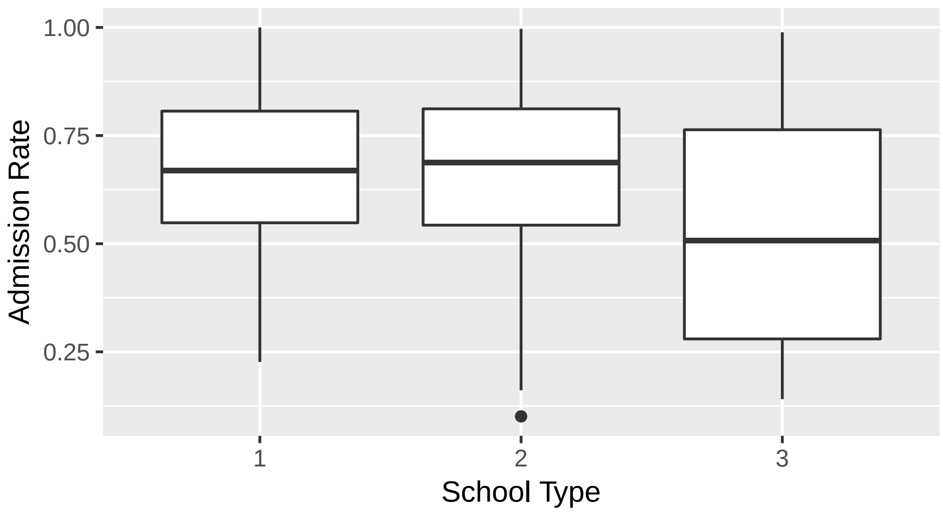
```
finalsample2 <- finalsample2 %>%
  mutate(region = as.character(REGION), control = as.character(CONTROL))

gf_boxplot(ADM_RATE ~ region, data = finalsample2, xlab = "Region", ylab = "Admission Rate") + labs(titl
```

Figure 12



```
gf_boxplot(ADM_RATE ~ control, data = finalsample2, xlab = "School Type", ylab = "Admission Rate") + lab
```

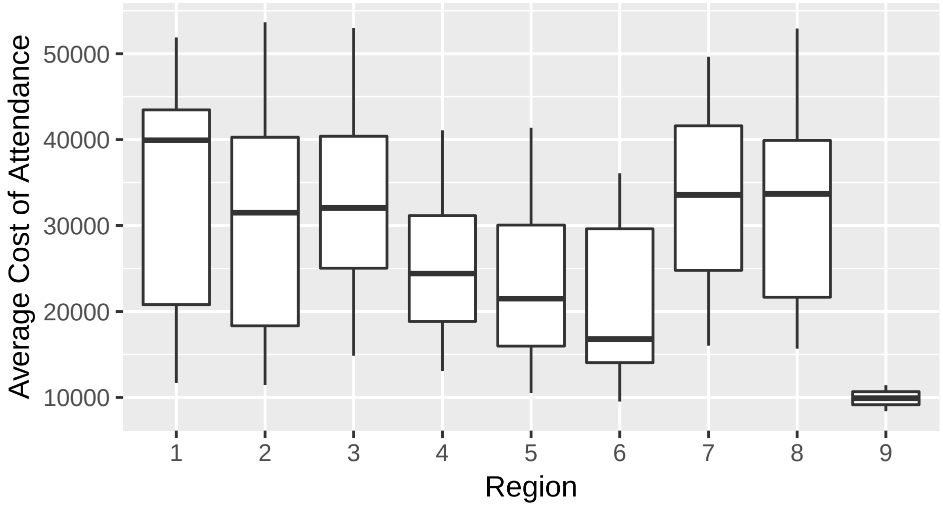
Figure 13



Our boxplots of average cost of attendance by region showed that New England schools were most expensive (Figure 14), while the boxplots of average cost of attendance by school type showed that both types of private schools are generally more expensive than public schools (Figure 15).

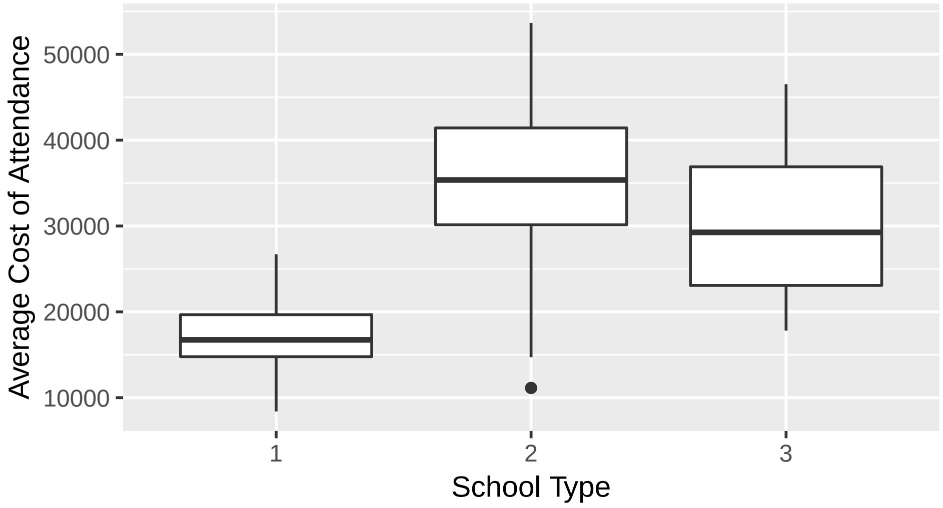
```
gf_boxplot(COSTT4_A ~ region, data = finalsample2, ylab = "Average Cost of Attendance", xlab = "Region"
```

Figure 14



```
gf_boxplot(COSTT4_A ~ control, data = finalsample2, xlab = "School Type", ylab = "Average Cost of Attend
```

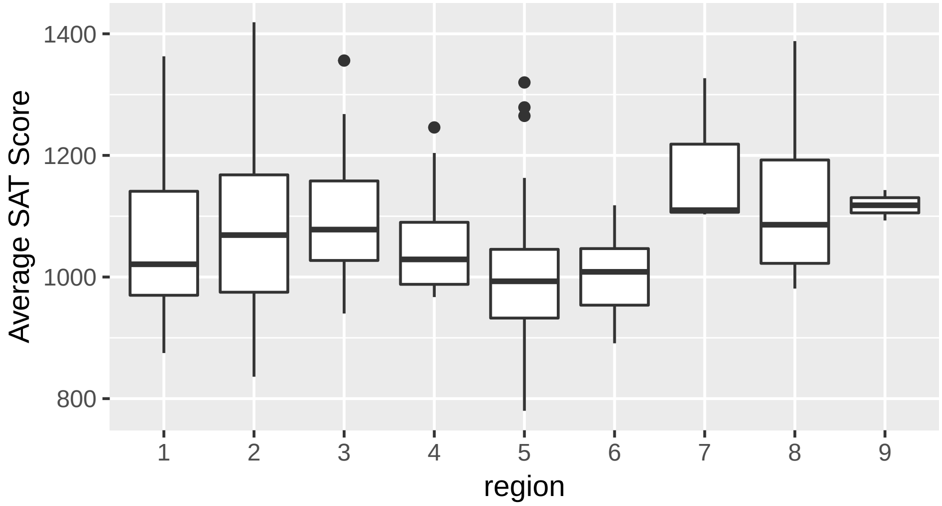
Figure 15



Finally, we found that average SAT scores were relatively uniform over school type (Figure 17). The Southwest, Rocky Mountains, and Far West did tend to have a higher median of average SAT scores than the rest of the other regions, though (Figure 16).

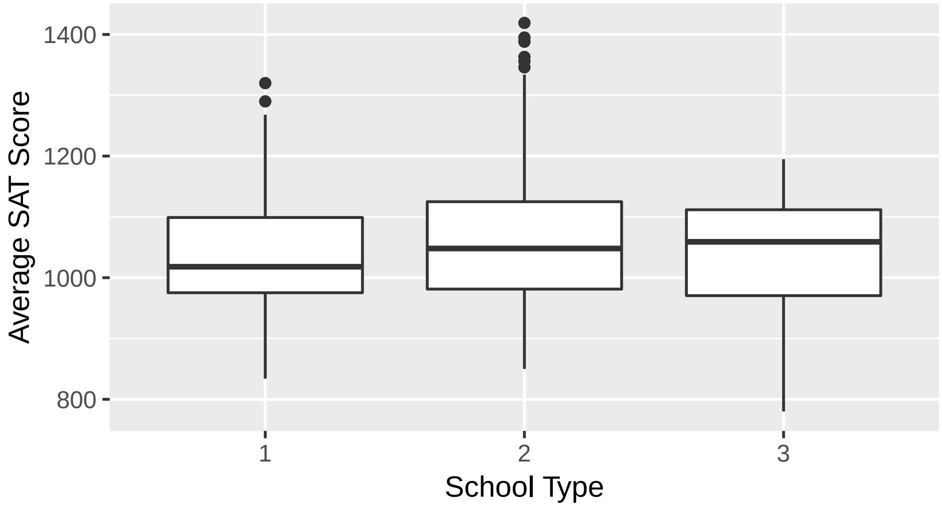
```
gf_boxplot(SAT_AVG_ALL ~ region, data = finalsample2, xlab = "region", ylab = "Average SAT Score") + lal
```

Figure 16



```
gf_boxplot(SAT_AVG_ALL ~ control, data = finalsample2, xlab = "School Type", ylab = "Average SAT Score"
```

Figure 17



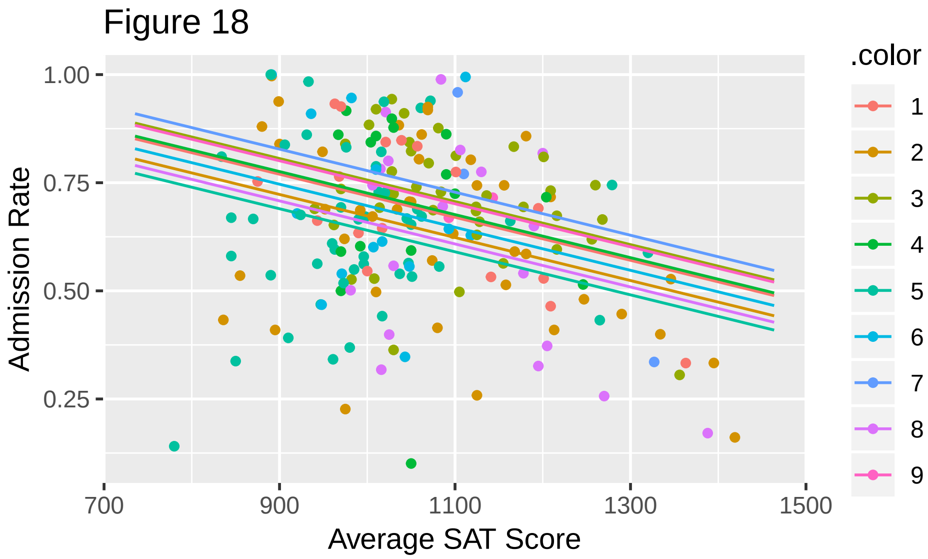
We then moved onto creating various multiple regression models, as a way of better predicting admission rate. Ultimately we ran a "kitchen-sink" multiple regression model with every possible combination of predictors and interaction terms. If there was a statistically significant interaction term between a quantitative variable and a categorical variable, we plotted a graph of admission rate, as predicted by those two variables, with an interaction term. If an interaction term was not statistically significant, we used a parallel slopes model instead.

We began by predicting admission rate by average SAT and region and observed the same negative correlation between admission rate and average SAT for all regions, just as we did during our simple linear regression (Figure 18). Amazingly, we found that our R-squared value began to grow when we started to differentiate by region, reaching a value of around 9%. Moreover, the p-value for this model is .003, indicating that our results are statistically significant. According to the model, we can claim that if a school in the Southwest has the same average SAT score as a school in the Rocky Mountains, the school in the Southwest will have an SAT that is on average lower by about 12.5%. Finally, the diagnostic plots for this model tell us once again that regression was appropriate, as the histogram of the residuals is mostly normal, with a slight left skew, the scatterplot of the residuals mostly shows no pattern, and the qq plot is linear (Figures 19, 20, and 21).

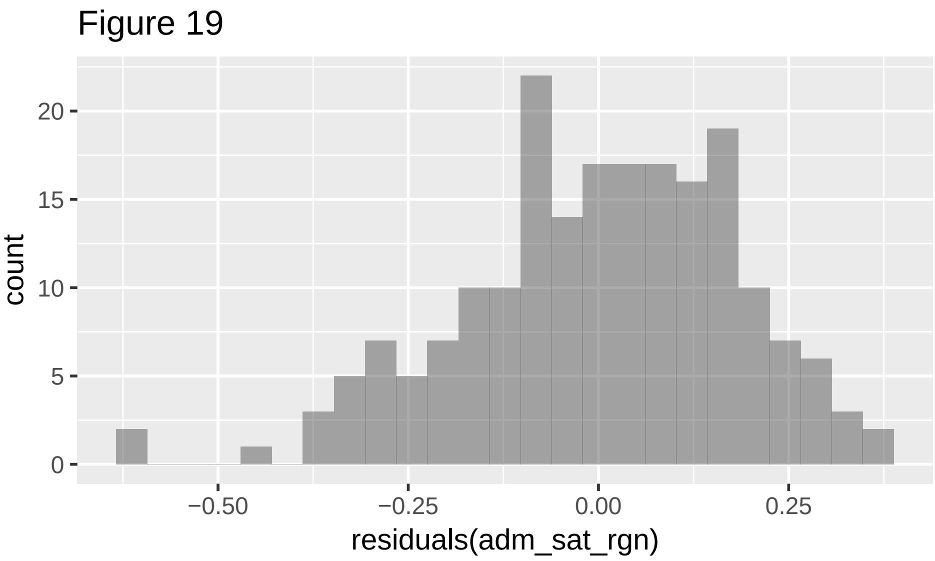
```
adm_sat_rgn <- lm(ADM_RATE ~ SAT_AVG_ALL + region, data = finalsample2)
msummary(adm_sat_rgn)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.2175010  0.1323356   9.200  < 2e-16 ***
## SAT_AVG_ALL -0.0004977  0.0001174  -4.241 3.48e-05 ***
## region2     -0.0467569  0.0529893  -0.882    0.379
## region3      0.0363298  0.0532802   0.682    0.496
## region4      0.0061931  0.0639395   0.097    0.923
## region5     -0.0798043  0.0523869  -1.523    0.129
## region6     -0.0230664  0.0665453  -0.347    0.729
## region7      0.0580353  0.1157523   0.501    0.617
## region8     -0.0616292  0.0615596  -1.001    0.318
## region9      0.0311305  0.1373416   0.227    0.821
##
## Residual standard error: 0.1835 on 190 degrees of freedom
## Multiple R-squared:  0.1197, Adjusted R-squared:  0.078
## F-statistic: 2.871 on 9 and 190 DF,  p-value: 0.003368
```

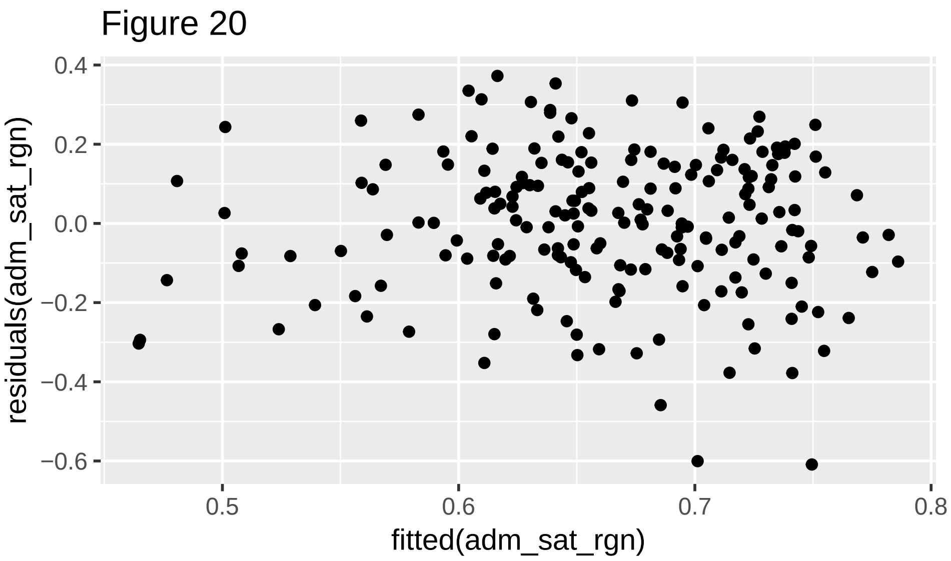
```
plotModel(adm_sat_rgn) + labs(x = "Average SAT Score", y = "Admission Rate", title = "Figure 18")
```



```
gf_histogram(~ residuals(adm_sat_rgn), title = "Figure 19")
```

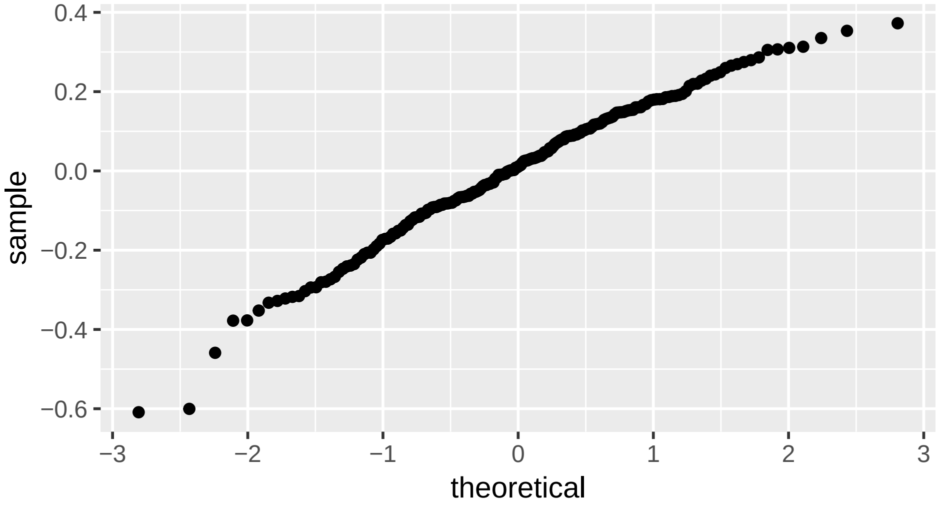


```r
gf_point(residuals(adm_sat_rgn) ~ fitted(adm_sat_rgn),  title = "Figure 20")
```



```r
gf_qq(~resid(adm_sat_rgn), title = "Figure 21")
```

Figure 21

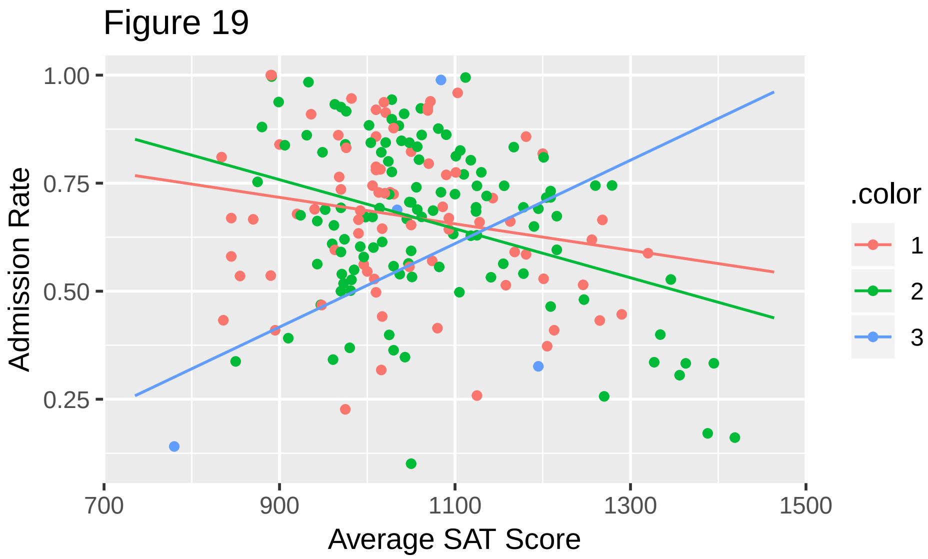


Next, we looked at a graph of admission rate, as predicted by average SAT score and school type (Figure 19). In the "kitchen-sink" model, this particular interaction term was statistically significant, so we chose to display this graphic without parallel slopes, which yields surprising results. We see that the both public schools and non-profit private schools follow the expected behavior, as their admission rates vary inversely with their respective average SAT scores. However, the opposite seems to be true for private schools for profit, as we see a positive correlation between average SAT and admission rate for color 3. Although this is unexpected, we must consider some factors at play before jumping to conclusions. First of all, we have very limited data on private schools that are for profit, as they make up a relatively small proportion of all colleges and universities in the United States. If we had more data, it is likely that we would obtain different results, which is something that our group could potentially look into in the future. To accomplish this, we could collect a stratified sample of types of schools, rather than a random sample of all types, to ensure we have enough data on every type of school. That being the case, though, we can still postulate why private schools for profit display such an unexpected behavior. Typically private schools for profit, such as performing arts academies, or military colleges have atypical admission criteria, which may result in admission rates not being as heavily dependent on SAT scores. Realistically, if we had more data points, we might have found no relationship between acceptance rate and average SAT for these schools, rather than a positive correlation. This could be one reason for the unexpected trend seen in the graph. Moreover, the R-squared value is almost 9%, so we can account for about 9% of the variability in admission rates by knowing the values of a school's average SAT, as well as its institution type. Moreover, the p-value is far below the alpha level of .05, so these results are statistically significant. Our residual standard error is 18%. The histogram of the residuals is fairly normal, but the scatterplot of the fitted residuals shows a slight pattern of sparcity on the left side of the graph (Figures 20 and 21). There is one heavy outlier, but without it there is almost no pattern. Moreover, the qq plot is linear, so our regression was valid (Figure 22).

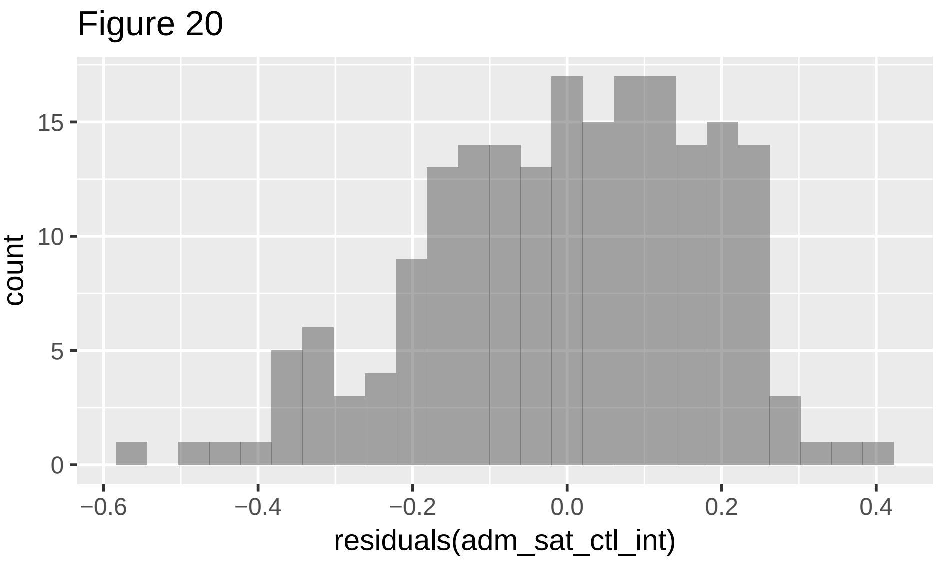
```
adm_sat_ctl_int <- lm(ADM_RATE ~ SAT_AVG_ALL + control + SAT_AVG_ALL*control, data = finalsample2)
msummary(adm_sat_ctl_int)
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.9928437  0.1905328   5.211 4.78e-07 ***
## SAT_AVG_ALL            -0.0003064  0.0001819  -1.684   0.0938 .
## control2                0.2758647  0.2426800   1.137   0.2570
## control3               -1.4446571  0.6494059  -2.225   0.0273 *
## SAT_AVG_ALL:control2 -0.0002610  0.0002291  -1.139   0.2561
## SAT_AVG_ALL:control3  0.0012718  0.0006271   2.028   0.0439 *
##
## Residual standard error: 0.1825 on 194 degrees of freedom
## Multiple R-squared:  0.1111, Adjusted R-squared:  0.08817
## F-statistic: 4.848 on 5 and 194 DF,  p-value: 0.000335
```

```
plotModel(adm_sat_ctl_int) + labs(x = "Average SAT Score", y = "Admission Rate", title = "Figure 19")
```

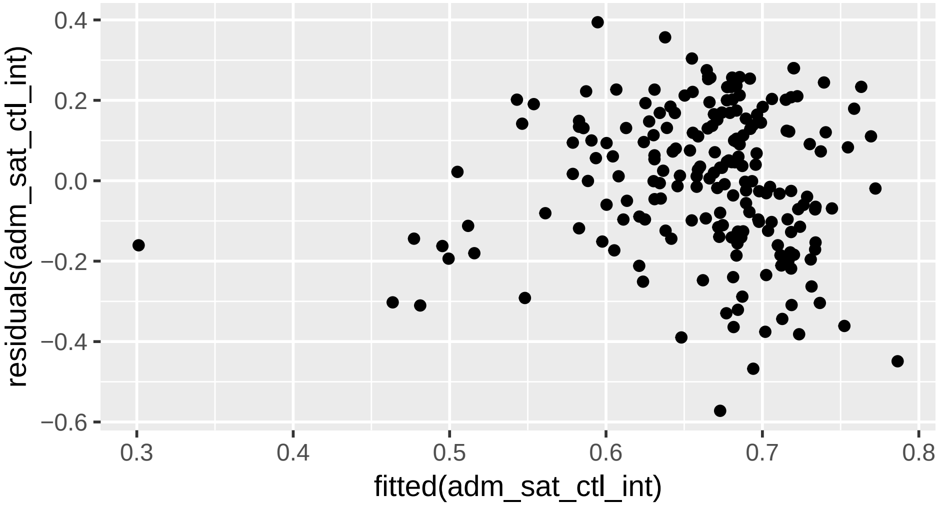


```
gf_histogram(~ residuals(adm_sat_ctl_int), title = "Figure 20")
```



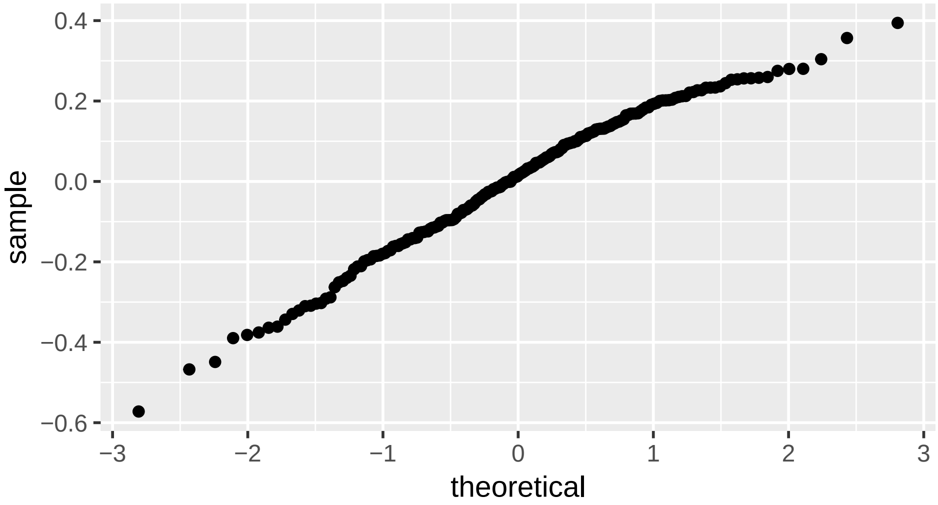
```
gf_point(residuals(adm_sat_ctl_int) ~ fitted(adm_sat_ctl_int), title = "Figure 21")
```

Figure 21



```
gf_qq(~resid(adm_sat_ctl_int), title = "Figure 22")
```

Figure 22

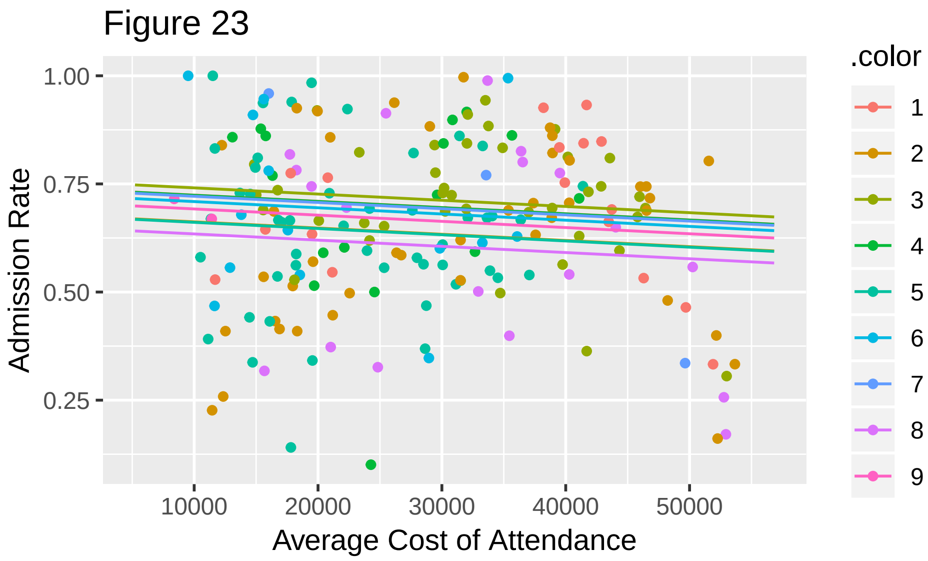


Next, we considered admission rate as predicted by region and average cost of attendance (Figure 23). Using region as an indicator variable, we obtained 9 parallel regression lines in our graph, each with a slightly negative slope, almost close to 0. This suggests that there is a very weak relationship, if any at all, between the variables. It seems that an increase in the cost of attendance could be associated with a decrease in admission rate, but this could simply be due to chance. Looking at the p-value, we see it is quite high, at 0.4903, meaning that our results in this particular regression are not statistically significant. The adjusted R-squared is also approximately 0, meaning that pretty much none of the variability in admission rates can be accounted for by the variation in cost of attendance and region. Overall, this model does not yield significant results, although our regression has been performed appropriately: a histogram of the residuals is approximately normal, although slightly skewed to the left, the scatterplot of the residuals shows no pattern whatsoever, and the normal probability plot is relatively straight (Figures 24, 25, and 26).

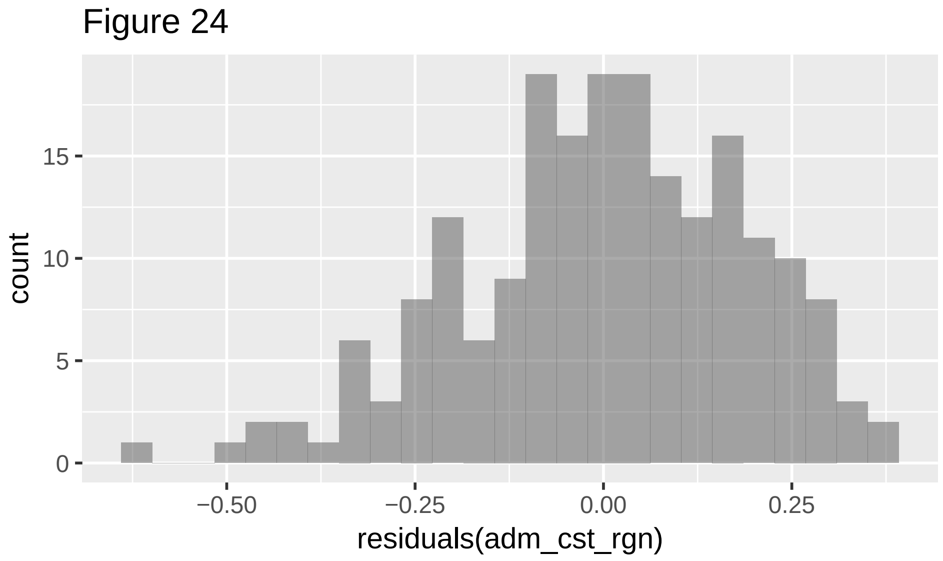
```
adm_cst_rgn <- lm(ADM_RATE ~ COSTT4_A + region, data = finalsample2)
msummary(adm_cst_rgn)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.384e-01  6.405e-02  11.529   <2e-16 ***
## COSTT4_A    -1.435e-06  1.282e-06  -1.119    0.264
## region2     -6.179e-02  5.539e-02  -1.116    0.266
## region3      1.678e-02  5.546e-02   0.302    0.763
## region4     -3.575e-04  6.769e-02  -0.005    0.996
## region5     -6.294e-02  5.598e-02  -1.124    0.262
## region6     -1.495e-02  7.118e-02  -0.210    0.834
## region7     -2.665e-03  1.199e-01  -0.022    0.982
## region8     -8.974e-02  6.393e-02  -1.404    0.162
## region9     -3.196e-02  1.465e-01  -0.218    0.828
##
## Residual standard error: 0.1914 on 190 degrees of freedom
## Multiple R-squared:  0.0427, Adjusted R-squared:  -0.002648
## F-statistic: 0.9416 on 9 and 190 DF,  p-value: 0.4903
```

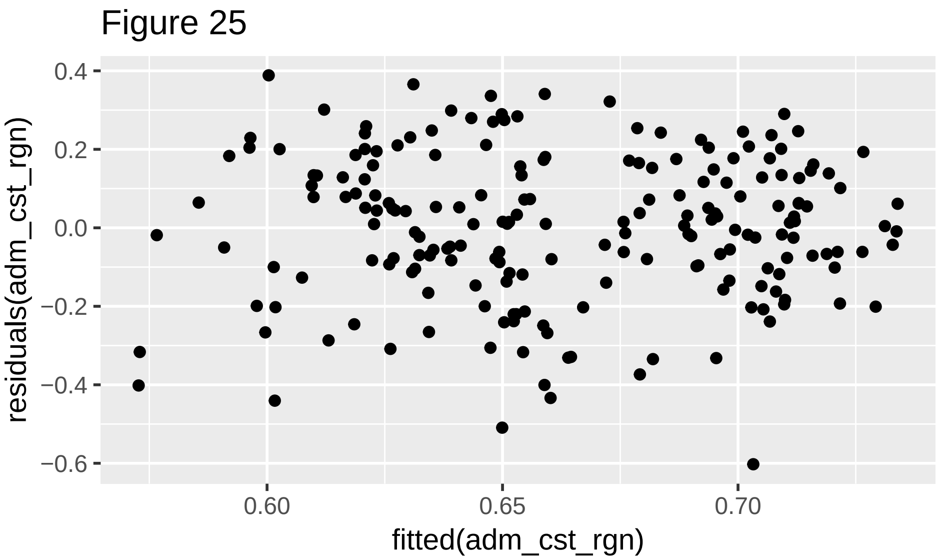
```
plotModel(adm_cst_rgn) + labs(x = "Average Cost of Attendance", y = "Admission Rate", title = "Figure 23
```



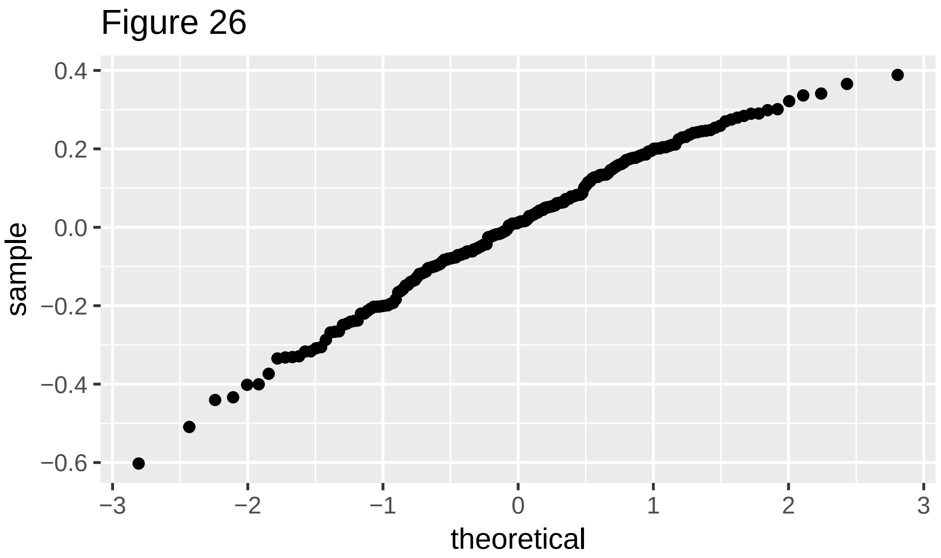
```
gf_histogram(~ residuals(adm_cst_rgn), title = "Figure 24")
```



```
gf_point(residuals(adm_cst_rgn) ~ fitted(adm_cst_rgn), title = "Figure 25")
```



```
gf_qq(~resid(adm_cst_rgn), title = "Figure 26")
```

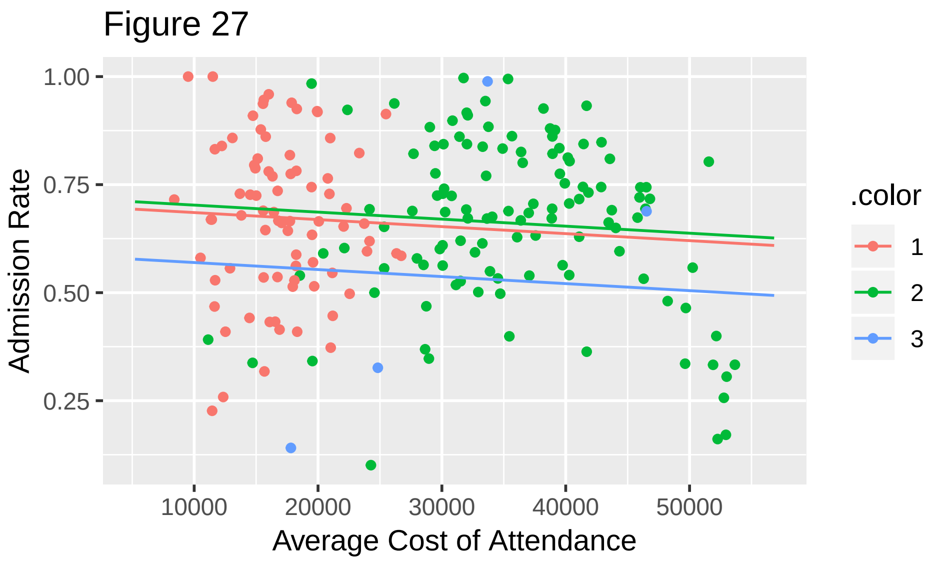


A regression of admission rate against cost of attendance and institution type also yields statistically insignificant results (Figure 27). With institution type as an indicator variable, we obtain 3 parallel regression lines with only a slight negative slope which again suggests a decrease in admission rate as cost of attendance increases. However, the slope is so close to 0 that it might not reflect an actual association between the variables. A quick look at the high p-value of 0.4254 confirms our concern that these results are again not statistically significant. Our adjusted R-squared is again around 0, so the variation in cost of attendance and school type has failed to account for any variability in cost of attendance. Moreover, one should be wary about using this graph to predict admission rates for other reasons. There is clear extrapolation occuring within the graph, as public schools are all focused on the left side, private schools are all focused on the right side, but their respective regression lines span both sides of the graph. However, the diagnostic plots do meet the conditions for regression with the exception of the major outliers in the fitted scatterplot of the residuals (Figures 28, 29, and 30). For several reasons, this particular multiple regression is unhelpful in predicting admission rates.

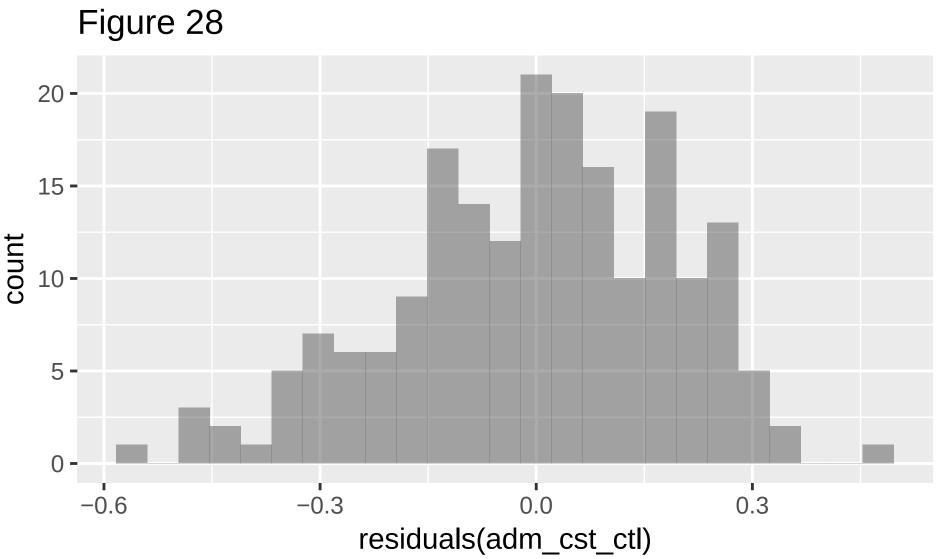
```r
adm_cst_ctl <- lm(ADM_RATE ~ COSTT4_A + control, data = finalsample2)
msummary(adm_cst_ctl)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.016e-01  3.840e-02  18.273   <2e-16 ***
## COSTT4_A    -1.625e-06  1.858e-06  -0.875    0.383
## control2     1.725e-02  4.479e-02   0.385    0.701
## control3    -1.157e-01  1.013e-01  -1.142    0.255
##
## Residual standard error: 0.1912 on 196 degrees of freedom
## Multiple R-squared:  0.01409,    Adjusted R-squared:  -0.001001
## F-statistic: 0.9337 on 3 and 196 DF,  p-value: 0.4254
```

```r
plotModel(adm_cst_ctl) + labs(x = "Average Cost of Attendance", y = "Admission Rate", title = "Figure 2
```

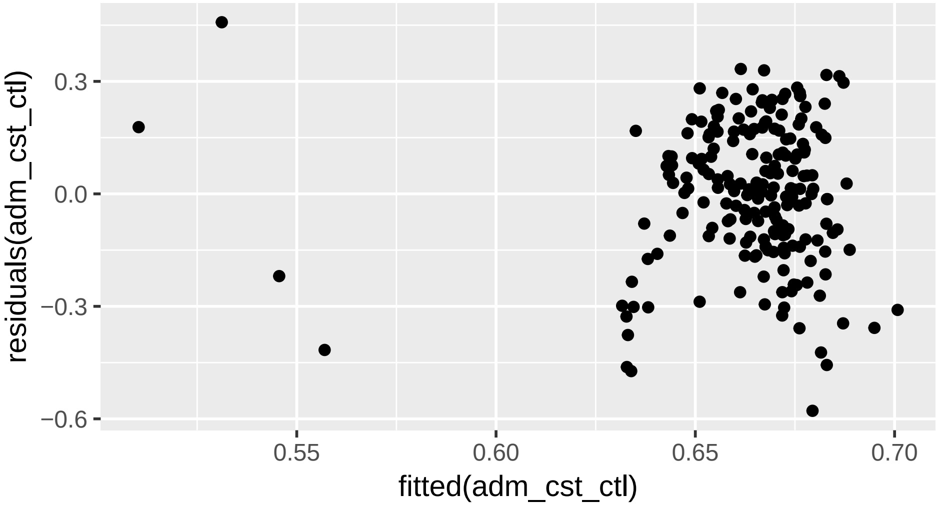


```r
gf_histogram(~ residuals(adm_cst_ctl), title = "Figure 28")
```



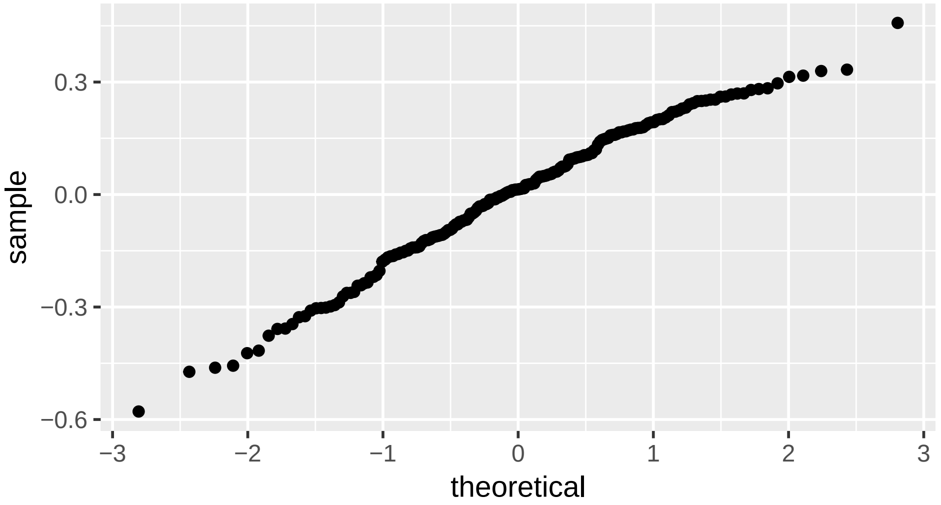
```r
gf_point(residuals(adm_cst_ctl) ~ fitted(adm_cst_ctl), title = "Figure 29")
```

Figure 29



```r
gf_qq(~resid(adm_cst_ctl), title = "Figure 30")
```

Figure 30

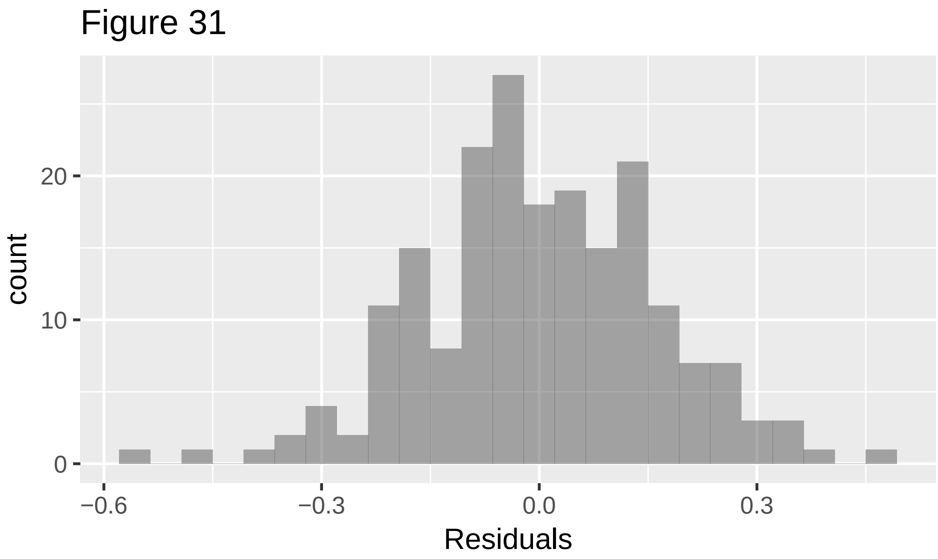


Moving on, we now incorporated all the explanatory variables in our multiple regression, as well as two interaction terms, SAT average by school type and SAT average by cost of attendance. We have tried introducing other interaction terms as well, but none other than the two aforementioned ones made much of a difference in our model, so for the sake of simplicity we decided against including them. The adjusted R-squared we get in this final regression sits at a decent 0.2134, so we accounted for around 21.34% of the variability in admissions rate with the variability in our explanatory variables. This suggests that there is more at play in determining admission rates than just the factors we considered - almost 80% of the variation has been left in the residuals, which could be reduced if we were to consider other factors that influence college admissions, such as race, GPA, extracurricular involvement, being a legacy or an international student, and more! An analysis of our residuals suggests that regression was appropriate, since the histogram is relatively normal, and the qq plot is linear (Figures 31 and 33). The scatterplot of the residuals, however, does seem more heavily clustered towards the right, but other than this, there is no pattern (Figure 32). If we were to remove outlying points, this issue would likely go away, which we would do if we had more time to conduct analysis on the data. The p-value in this model is 2.006e-07, small enough for us to conclude that we have statistically significant results.

```
adm_final_mlr_int <- lm(ADM_RATE ~ SAT_AVG_ALL + COSTT4_A + region + control + SAT_AVG_ALL*control + SA
msummary(adm_final_mlr_int)
```

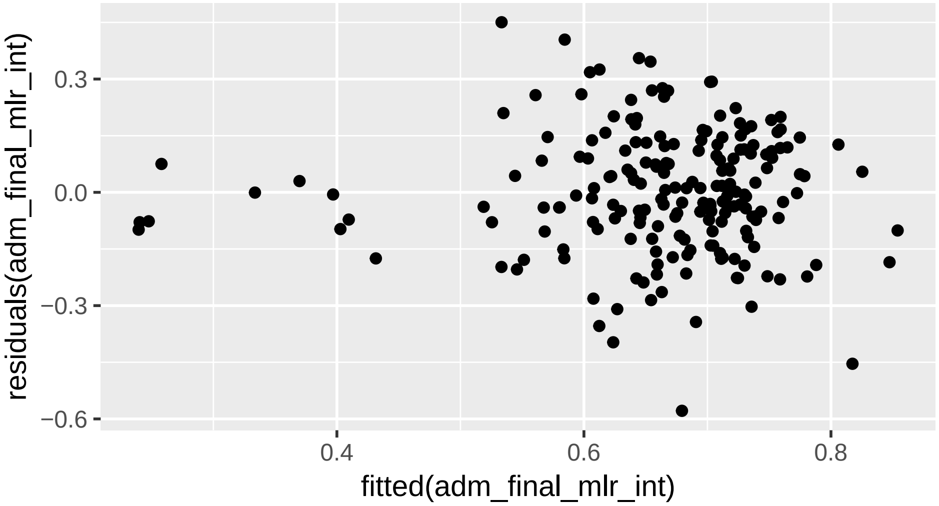
```
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)            -6.983e-02  3.060e-01  -0.228  0.81971    
## SAT_AVG_ALL             5.826e-04  2.905e-04   2.005  0.04642 *  
## COSTT4_A                6.867e-05  1.400e-05   4.905 2.04e-06 ***
## region2                -4.560e-03  4.973e-02  -0.092  0.92704    
## region3                 7.081e-02  5.091e-02   1.391  0.16592    
## region4                 7.173e-02  6.357e-02   1.128  0.26062    
## region5                 7.735e-03  5.269e-02   0.147  0.88344    
## region6                 4.955e-02  6.562e-02   0.755  0.45113    
## region7                 1.249e-01  1.083e-01   1.153  0.25036    
## region8                -3.540e-02  5.852e-02  -0.605  0.54591    
## region9                 7.925e-02  1.332e-01   0.595  0.55271    
## control2               -1.131e+00  4.221e-01  -2.679  0.00804 ** 
## control3               -1.715e+00  6.131e-01  -2.798  0.00570 ** 
## SAT_AVG_ALL:control2  9.789e-04  4.145e-04   2.362  0.01924 *  
## SAT_AVG_ALL:control3  1.482e-03  5.944e-04   2.493  0.01354 *  
## SAT_AVG_ALL:COSTT4_A -5.878e-08  1.324e-08  -4.441 1.54e-05 ***
## 
## Residual standard error: 0.1695 on 184 degrees of freedom
## Multiple R-squared:  0.2727, Adjusted R-squared:  0.2134 
## F-statistic: 4.598 on 15 and 184 DF,  p-value: 2.006e-07
```

```
gf_histogram(~ residuals(adm_final_mlr_int), xlab = "Residuals", title = "Figure 31")
```



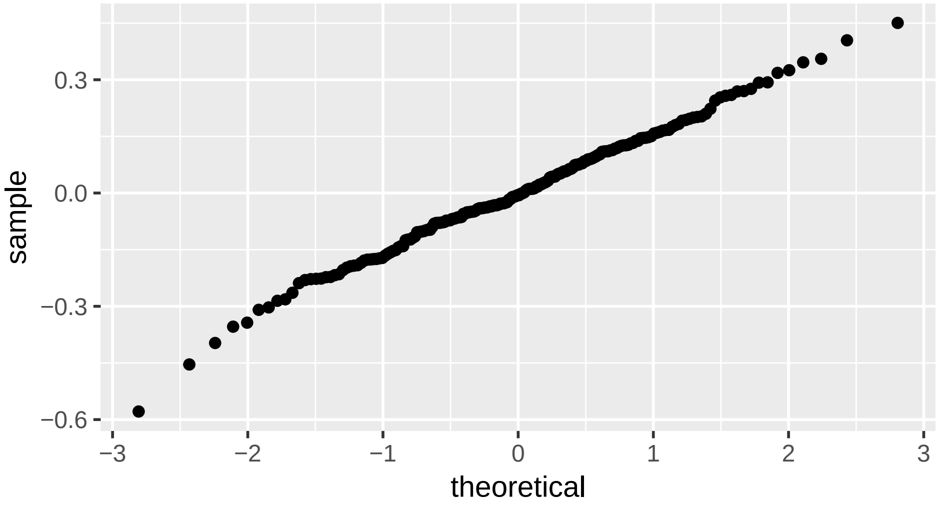
```
gf_point(residuals(adm_final_mlr_int) ~ fitted(adm_final_mlr_int), title = "Figure 32")
```

Figure 32



```
gf_qq(~resid(adm_final_mlr_int), title = "Figure 33")
```

Figure 33

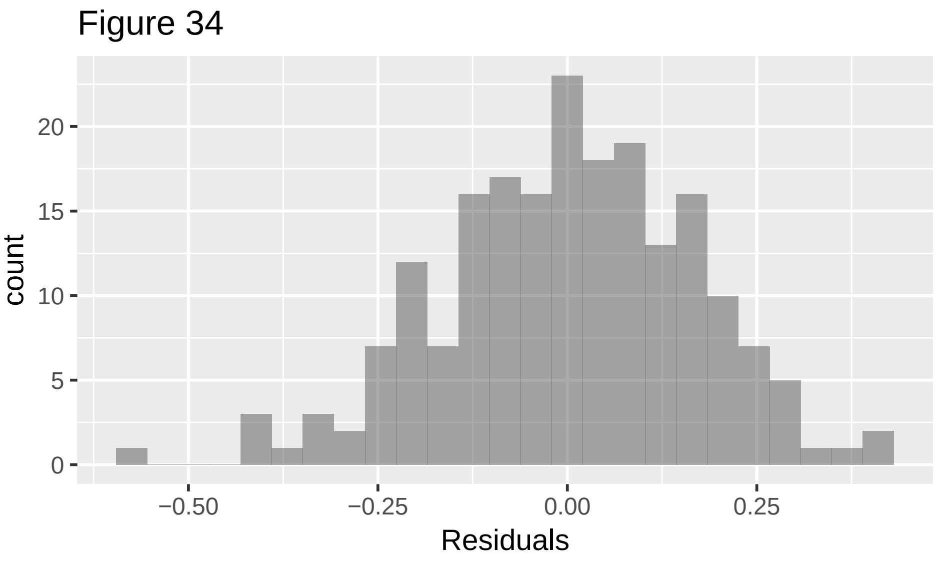


Finally, we also decided to consider a multiple regression model which excluded region as an explanatory variable, since excluding it did not seem to make much of a difference: our adjusted R-squared fell only slightly, from 0.2134 to 0.2096, so that's only 0.38% less variability accounted for. In this new model, our p-value was even lower, at 4.232e-09, so results remain statistically significant. Overall, regression was appropriate: the residual histogram is approximately normal and the qq plot is extremely linear (Figures 34 and 36). However, the scatterplot of the residuals is more clustered on the right and very sparce to the left of the graph, so a clear pattern is unfortunately visible (Figure 35). If we were to remove the points on the left from our model, then no pattern would be shown. If we had more time to work with our data, this is likely one of the next steps we would take to creating a better model of admission rates. However, for the purposes of our analysis we will use this model in its current state. This version of the multiple regression may be preferrable to the previous one, since it is a bit simpler and has less of a pattern in the residual scatterplot, has one less explanatory variable, yet still yields statistically significant results with a negligible change in the amount of variability in admission rates accounted for.

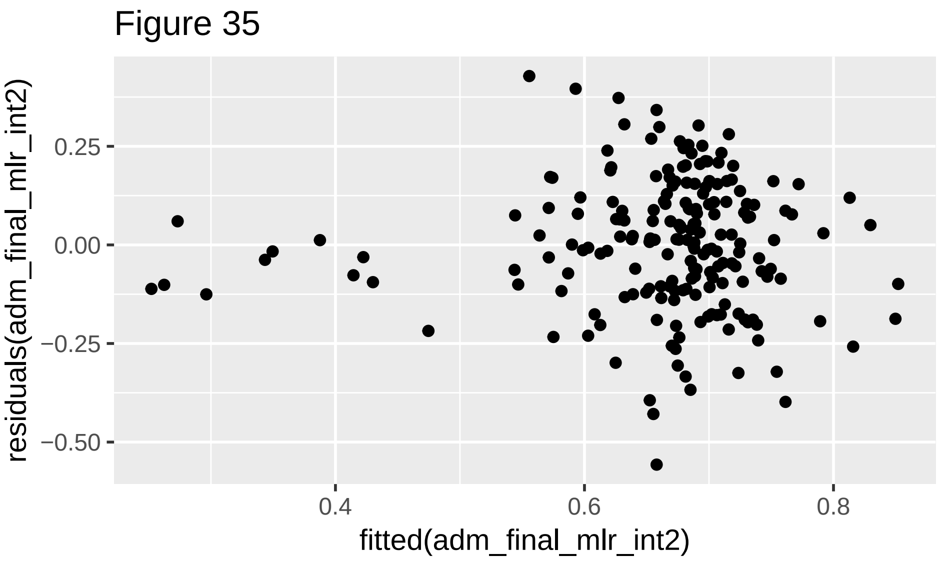
```
adm_final_mlr_int2 <- lm(ADM_RATE ~ SAT_AVG_ALL + COSTT4_A + control + SAT_AVG_ALL*control + SAT_AVG_AL
msummary(adm_final_mlr_int2)
```

```
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1.611e-01  2.886e-01  -0.558  0.57742
## SAT_AVG_ALL             7.210e-04  2.803e-04   2.572  0.01086 *
## COSTT4_A                7.304e-05  1.362e-05   5.362 2.35e-07 ***
## control2               -1.329e+00  4.143e-01  -3.208  0.00157 **
## control3               -1.646e+00  6.059e-01  -2.716  0.00721 **
## SAT_AVG_ALL:control2  1.208e-03  4.054e-04   2.980  0.00326 **
## SAT_AVG_ALL:control3  1.404e-03  5.860e-04   2.396  0.01753 *
## SAT_AVG_ALL:COSTT4_A -6.475e-08  1.293e-08  -5.008 1.24e-06 ***
##
## Residual standard error: 0.1699 on 192 degrees of freedom
## Multiple R-squared:  0.2374, Adjusted R-squared:  0.2096
## F-statistic: 8.541 on 7 and 192 DF,  p-value: 4.232e-09
```

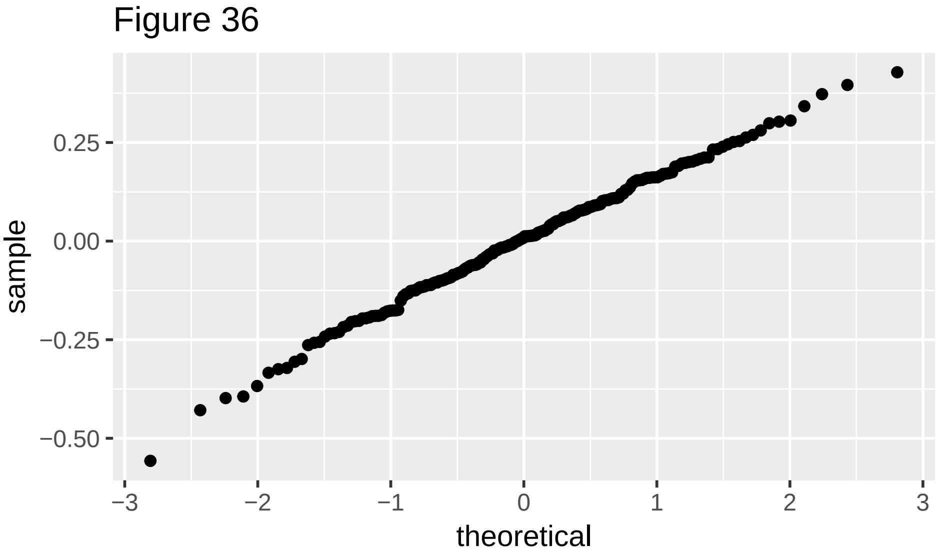
```
gf_histogram(~ residuals(adm_final_mlr_int2), xlab = "Residuals", title = "Figure 34")
```



```
gf_point(residuals(adm_final_mlr_int2) ~ fitted(adm_final_mlr_int2), title = "Figure 35")
```



```
gf_qq(~resid(adm_final_mlr_int2), title = "Figure 36")
```



Conclusion

As we added more predictors to our model, our standard error dropped, our R-squared value rose, the histogram of our residuals became more normally distributed, and our p-value approached zero. From our multiple regression model, we obtaned an adjusted R-squared value of .21, a residual standard error of .17 on 192 degrees of freedom, and a p-value of 4.2e-9. With a p-value far below alpha (0.05), we can reject our null hypothesis and conclude that our results are in fact statistically significant. However, the conditions that we examined did not encompass all of the college admissions process. While some do have a clear influence on admissions rate, like SAT score and the type of school, the degree to which these factors predicted admissions rates was not overwhelmingly strong. Ultimately, we were able to account for around 21% of the variability associated with college admissions by knowing the average cost of attendance, average SAT score, region, and the type of a given school. Likely, there are other factors that affect admission rates more strongly than our predictor variables, such as one's personal essay, GPA, or extracurricular involvements. On top of that, there is probably a significant element of luck in college admissions, so we doubt that one could create a model that perfectly predicts admissions. Since we pulled data from such a diverse body of schools, our results indicate that SAT score, cost of attendance, and school type all influence admissions to some degree, but that some do more than others. For instance, cost of attendance turned out to have much less leverage in the final multiple regression model than average SAT score, as its coefficient was an order of magnitude less than that of average SAT score.

Our study had several limitations. First of all, there were several NULL values in our data set that we had to remove fom analysis. Initially we wanted to explore how affirmative action influences college admissions rates - but we struggled finding sufficient data on admissions based on race and sex. After compiling our data we realized that in order to be able to do a statistical analysis, we had to choose predictor variables that had the most data. Another limitation was our cost of attendance variable, which only accounted for the tuition costs of students receiving Title IV aid. As a result, it is difficult to draw conclusions about tuition cost overall, as the entire population of students not on Title IV aid was unaccounted for. Moreover, it is unclear whether our final multiple regression model is completely valid, as the fitted scatterplot of the residuals is very sparce on the left side of the graph. This is likely due to the fact that some of our initial data is slighlty skewed, with a few outliers. If we had more time to conduct analysis, we could attempt to transform the data to diminish the amount of outliers and to bring the data closer to a normal distribution. We could also conduct our analysis by removing the outliers to begin with. Both procedures would result in our final residuals appearing more patternless, but seeing as that was the only issue in our final multiple regression, we are mostly content with our results.

In the future, it might be interesting to explore our initial question about the influence of race and sex on college admissions rates, which might be possible as more data continues to be collected over time. It would also be interesting to study how the variables we looked at influence admissions rates for smaller subpopulations of colleges/universities, such as Ivy League schools or elite liberal arts colleges. In conclusion, the factors that account for the variability in college admission rates are clearly very complicated and it was beyond the scope of this project to hope to detail all of them. However, we did distinguish four predictor variables that account for 21% of the variability in admissions, which is an important first step in this much more complex process.