What Makes Colleges So Expensive?

Exploratory Analysis of U.S. College Admissions Data in 2013

Julius Lin

5/7/2022

Contents

Executive Summary

- **The defining factor on tuition is private or public ownership.** The defining factor on tuition is public or private ownership. Private schools are much more expensive and more selective in admissions, and they also produce proportionately higher graduation rate to tuition than their public counterparts.
- **The more ethnically diverse schools also tend to be more expensive.** The percentage of underrepresented students is positively associated with tuition, which effectively means that underrepresented students disproportionately pay more for colleges where they are more represented.
- **Overall candidate quality is a strong predictor of tuition.** Median SAT score among applicants, which is positively associated with tuition, is also a strong predictor of colleges' selectivity, at least from a certain threshold onward (1200+), despite claims of holistic admissions.
- **Urban campuses are disproportionately more expensive.** Colleges in larger cities are disproportionately more expensive, though this effect is much less pronounced among mid-sized and small cities.

Data Overview

Using a data set of admissions data by college in 2013 from Kaggle, 1155 entries (colleges) of data are imported, with 40 different parameters.

```
colnames(d)
```

```
##  [1] "Highest.degree.offered"
##  [2] "Applicants.total"
##  [3] "Admissions.total"
##  [4] "Enrolled.total"
##  [5] "Percent.of.freshmen.submitting.SAT.scores"
##  [6] "Percent.of.freshmen.submitting.ACT.scores"
##  [7] "SAT.Critical.Reading.25th.percentile.score"
##  [8] "SAT.Critical.Reading.75th.percentile.score"
##  [9] "SAT.Math.25th.percentile.score"
## [10] "SAT.Math.75th.percentile.score"
## [11] "Tuition.and.fees..2013.14"
## [12] "State.abbreviation"
## [13] "Control.of.institution"
## [14] "Historically.Black.College.or.University"
## [15] "Degree.of.urbanization..Urban.centric.locale."
## [16] "Total..enrollment"
## [17] "Undergraduate.enrollment"
## [18] "Graduate.enrollment"
## [19] "Percent.of.total.enrollment.that.are.White"
## [20] "Percent.of.total.enrollment.that.are.Nonresident.Alien"
## [21] "Percent.of.total.enrollment.that.are.women"
## [22] "Graduation.rate...Bachelor.degree.within.6.years..total"
## [23] "Percent.of.freshmen.receiving.any.financial.aid"
```

After some data cleaning, the following variables are produced for investigation.

```
dim(df)
```

```
## [1] 1155   18
```

```
colnames(df)
```

```
##  [1] "State"               "Ownership"           "Tuition"
##  [4] "HighestDegree"       "HBCU"                "Locale"
##  [7] "LocaleSize"          "TotalEnroll"         "EnrollRate"
## [10] "AdmitRate"           "TestRate"            "MidSAT"
## [13] "UndergradPct"        "FinancialAidPct"     "GraduationRate"
## [16] "NonresidentAlienPct" "WomenPct"            "NonWhitePct"
```

The categorical variables include:

1. `State` is the name of the state the college is in.
2. `Ownership` describes the nature of the college's ownership, either public or private.
3. `HighestDegree` describes the highest degree that the college offers, Doctor's, Master's, or Bachelor's.

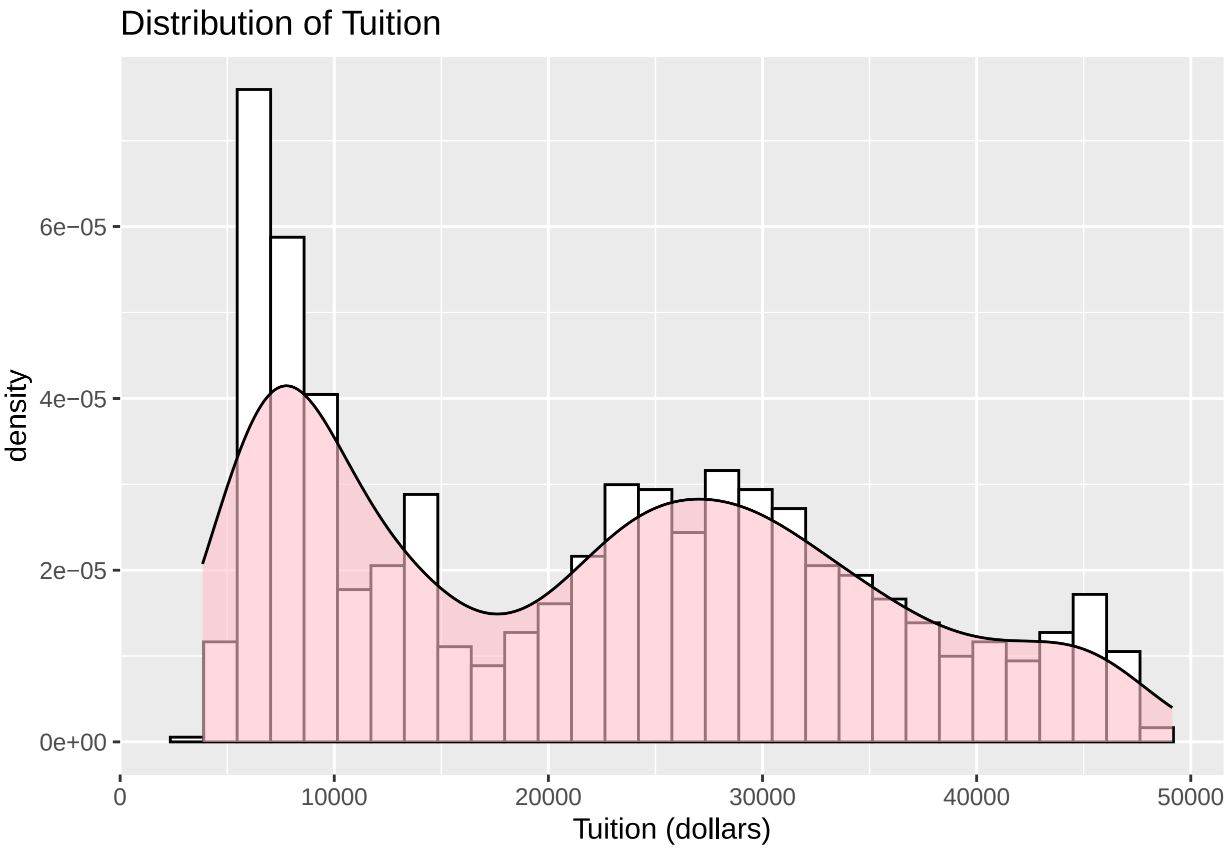
4. `HBCU` describes whether or not the college is a Historically Black College or University.
5. `Locale` describes the type of environment the college is situated in, city, suburban, rural, or town.
6. `LocaleSize` is a sub-category under `Locale`: cities and suburbs are categorized into small, midsize, and large; rurals and towns are categorized into fringe, remote, and distant.

The continuous variables include:

1. `Tuition` is the tuition in the academic year of 2013-2014 in dollars.
2. `TotalEnroll` is the total number of currently enrolled students in the college.
3. `EnrollRate` is the percentage of admitted applicants who actually chose to enroll.
4. `AdmitRate` is the percentage of applicants who were admitted.
5. `TestRate` is the percentage of applicants who submitted standardized testing; this is an underestimate.
6. `MidSAT` is an estimate of the median SAT of applicants (out of 1600).
7. `UndergradPct` is the percentage of undergraduate students in the entire student body.
8. `FinancialAidPct` is the percentage of students who are receiving financial aid.
9. `GraduationRate` is the rate of graduation within 6 years.
10. `NonresidentAlienPct` is the percentage of international students in the student body.
11. `WomenPct` is the percentage of women in the student body.
12. `NonWhitePct` is the percentage of non-White students in the student body.

Data Analysis

```r
ggplot(aes(x = Tuition), data = df) +
  geom_histogram(aes(y=after_stat(density)), color = "black", fill = "white", bins=30) +
  geom_density(alpha = .6, fill = "pink") +
  ggtitle("Distribution of Tuition") + xlab ("Tuition (dollars)")
```



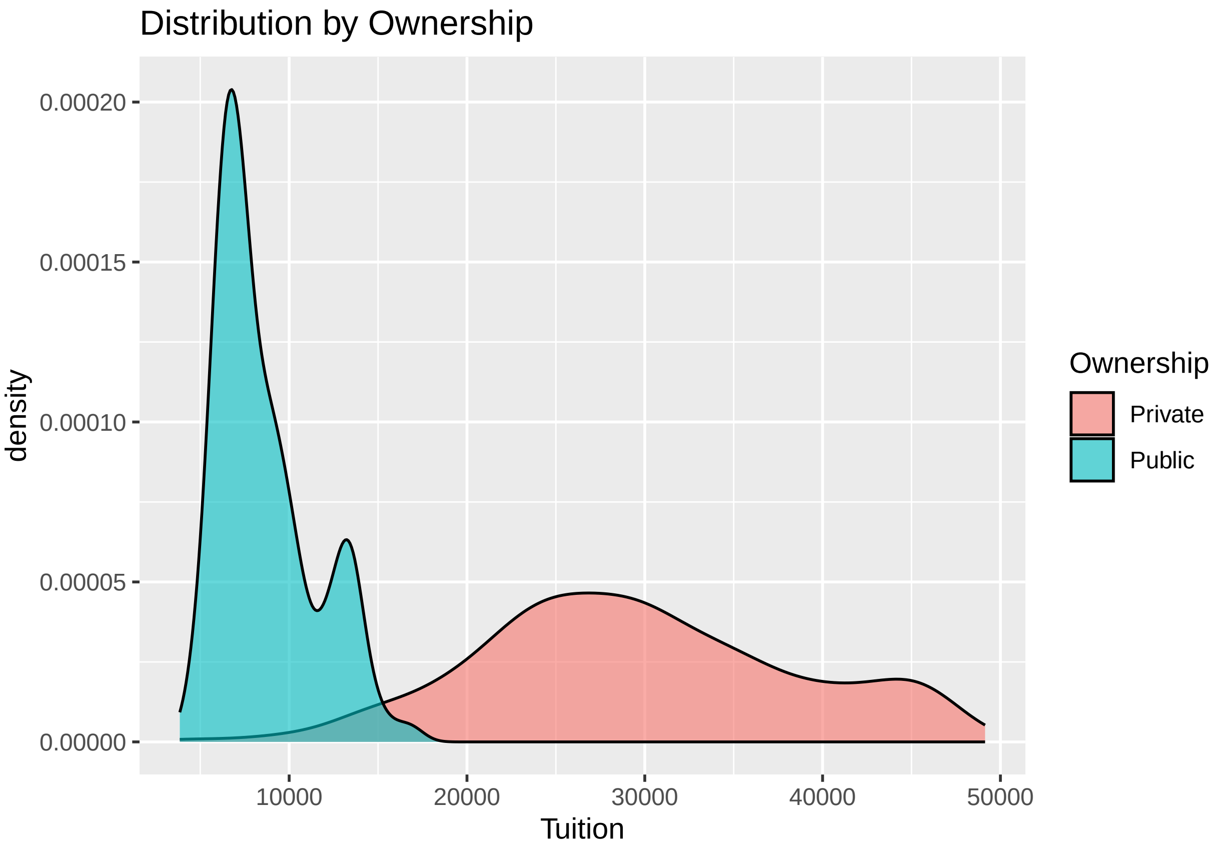
The histogram above shows the overall distribution of tuition. The two peaks suggest there might be two main categories. Based on our prior knowledge of college tuition, it is suspected that the two peaks reflect private schools and public schools.

```
t.test(df[df$Ownership == "Public",]$Tuition,
       df[df$Ownership == "Private",]$Tuition)
```

```
##
##  Welch Two Sample t-test
##
## data:  df[df$Ownership == "Public", ]$Tuition and df[df$Ownership == "Private", ]$Tuition
## t = -59.645, df = 920.77, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -21736.53 -20351.68
## sample estimates:
## mean of x mean of y
##  8623.781 29667.886
```

With a t-test ($p < 2.2e-16$), it can be seen that the tuition of public schools is lower than their private counterparts with high statistical significance. This prompts us to create a new histogram, this time separating the two:

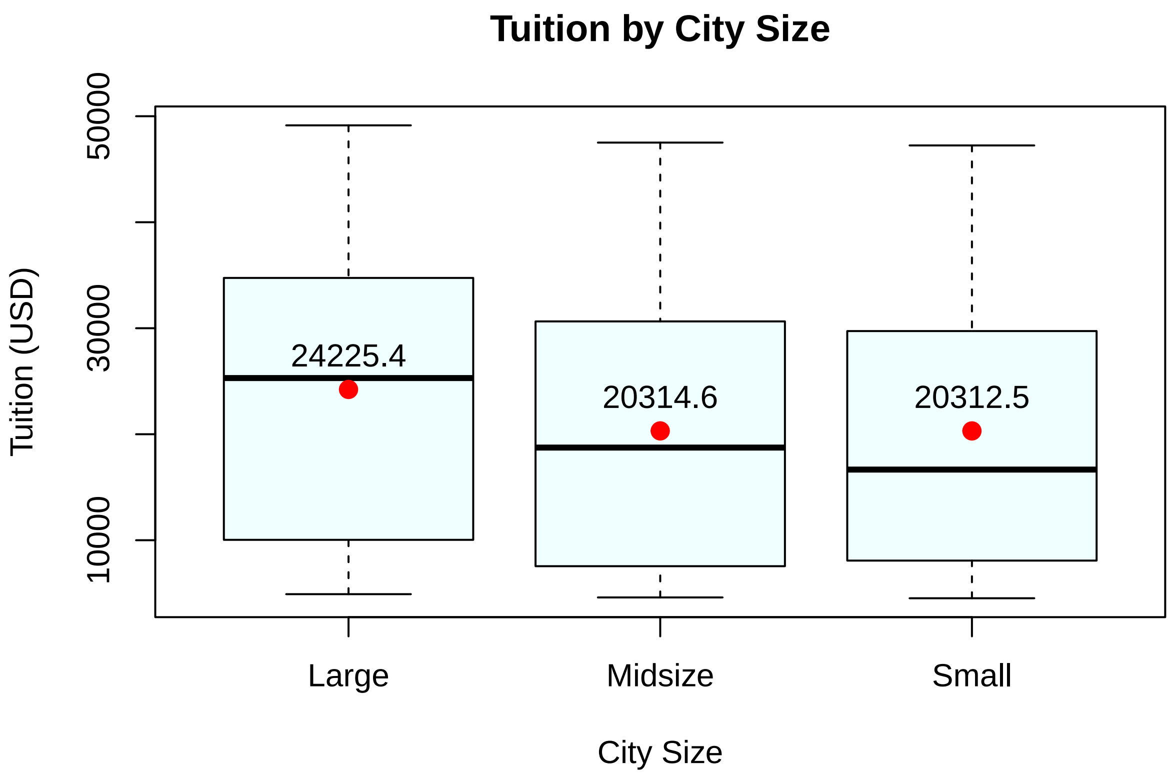
```
ggplot(aes(x = Tuition, fill = Ownership), data = df) +
  geom_density(alpha = .6) +
  ggtitle("Distribution by Ownership")
```



This new graph clearly illustrates that the two peaks can be explained by the public-private division. Moreover, whereas private schools spread over a large range of tuition, public schools are relatively low-cost most of the time.

Could geography impact tuition? Are universities in larger cities more expensive than in smaller cities? Zooming on to city universities only, we first create a boxplot to see if there is any observable difference in tuition by city sizes.

```
dftemp <- df[df$Locale == "City",]
boxplot(Tuition ~ LocaleSize,
        main = "Tuition by City Size",
        xlab = "City Size", ylab = "Tuition (USD)", col = "azure",
        data = dftemp)
mean_rate <- tapply(dftemp$Tuition, dftemp$LocaleSize, mean)
points(mean_rate, col = "red", pch = 19, cex = 1.2)
text(x = c(1:3), y = mean_rate + 3200, labels = round(mean_rate, 1))
```



The boxplot shows that universities in large cities are the most expensive, with both its mean and median being the highest among the three categories. It is less clear, however, whether there is a real difference between those in midsize and small cities, which can be resolved by performing one-way ANOVA on tuition by city size. But first, is the assumption of ANOVA is reasonably met?

```
print("S.D. by Grade")
```

```
## [1] "S.D. by Grade"
```

```
(sds <- tapply(dftemp$Tuition, dftemp$LocaleSize, sd))
```

```
##    Large  Midsize    Small 
## 13145.53 13022.14 12779.64
```

```
print("Ratio of Max/Min Sample S.D.")
```

```
## [1] "Ratio of Max/Min Sample S.D."
```

```
round(max(sds)/min(sds), 2)
```

```
## [1] 1.03
```

One-way ANOVA

As can be seen, the ratio between maximum S.D. and minimum S.D. is 1.03, which is far less than 2. Therefore, the equal variance assumption of ANOVA is reasonably met. Therefore, an ANOVA model is built as follows:

```
aov1 <- aov(dftemp$Tuition ~ dftemp$LocaleSize)
print("Summary Information for the Model")
```

```
## [1] "Summary Information for the Model"
```

```
summary(aov1)
```

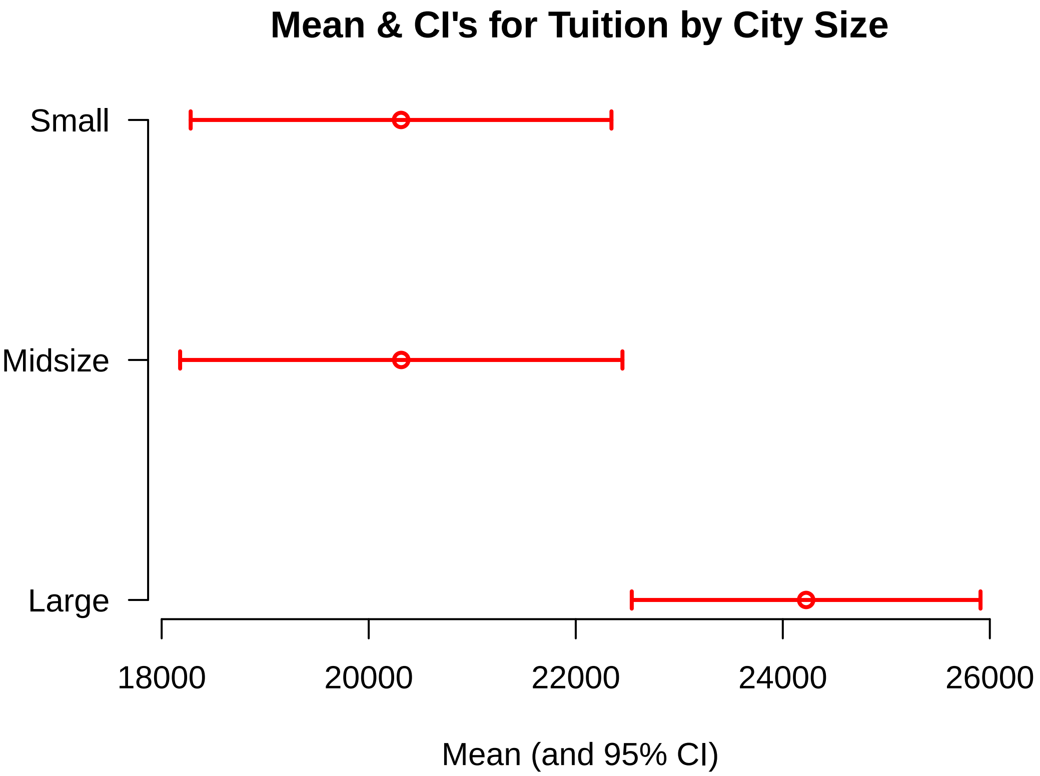
```
##                    Df    Sum Sq   Mean Sq F value  Pr(>F)
## dftemp$LocaleSize   2 1.995e+09 997575179   5.899 0.00293 **
## Residuals         528 8.929e+10 169116012
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The summary information shows that mean tuition rates across city sizes are statistically significantly different, with a $p = 0.00293 < 0.05$. With this knowledge, tuition is fit based on city size as a regression model without an intercept. The graph below shows the 95% confidence interval of tuition by city size.

```
lm1 <- lm(dftemp$Tuition ~ dftemp$LocaleSize -1)
summary(lm1)
```

```
##
## Call:
## lm(formula = dftemp$Tuition ~ dftemp$LocaleSize - 1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19314.4 -12783.1   -287.4  10263.0  27195.4
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## dftemp$LocaleSizeLarge    24225.4      857.5   28.25   <2e-16 ***
## dftemp$LocaleSizeMidsize  20314.6     1087.5   18.68   <2e-16 ***
## dftemp$LocaleSizeSmall    20312.5     1034.6   19.63   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 13000 on 528 degrees of freedom
## Multiple R-squared:  0.7438, Adjusted R-squared:  0.7423
## F-statistic: 510.9 on 3 and 528 DF,  p-value: < 2.2e-16
```

```
CIs <- confint(lm1)
coefs <- coef(lm1)
par(mar=c(5,8,4,2))
plotCI(coefs, 1:(length(coefs)), ui = CIs[,2], li = CIs[,1], axes = FALSE,
       err = "x", ylab = "", xlab = "Mean (and 95% CI)",
       main = "Mean & CI's for Tuition by City Size", lwd = 2, col = "red")
axis(side = 1)
axis(side = 2, at = 1:(length(coefs)),
     label = levels(as.factor(dftemp$LocaleSize)), las = 2)
```

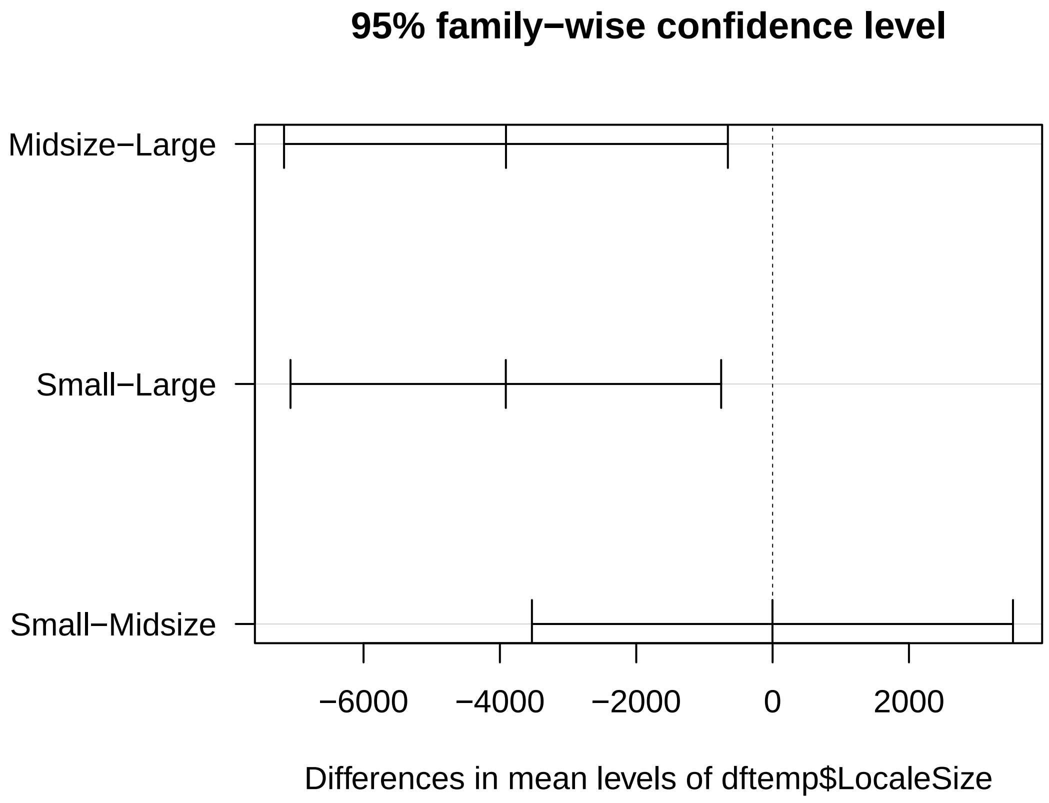


The Tukey confidence intervals show that two pairs of city sizes are statistically significantly different at 95% confidence level: Midsize-Large and Small-Large. Small-Midsize contains 0 in its confidence interval, and they are barely different from each other. We can conclude, therefore, that the tuition of colleges in large city is higher than their counterparts in midsize and small cities. This is plausible because living in a large city usually entails a higher cost of living, such as food and rent. However, there is insufficient evidence to conclude that colleges in mid-size cities are more expensive than in small cities. It is possible that cost of living does not increase significantly until city size reaches a certain threshold.

```
TukeyHSD(aov1, conf.level = .05)
```

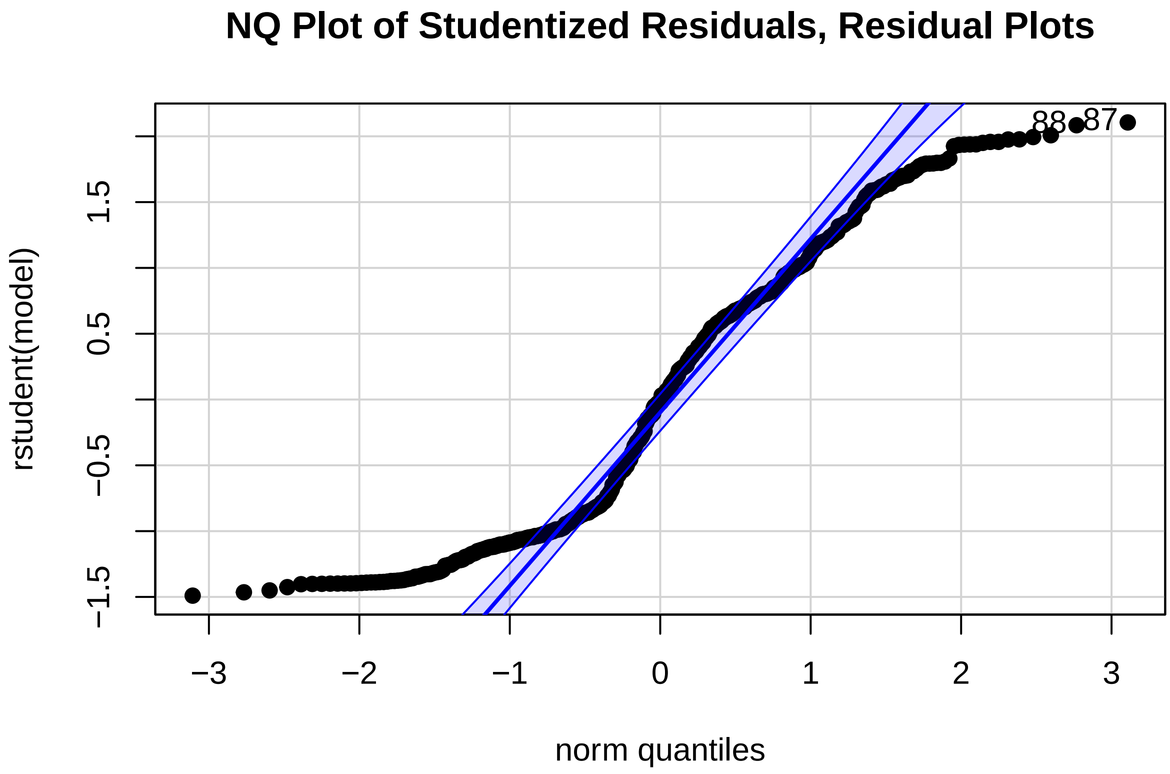
```
##   Tukey multiple comparisons of means
##     5% family-wise confidence level
##
## Fit: aov(formula = dftemp$Tuition ~ dftemp$LocaleSize)
##
## $`dftemp$LocaleSize`
##                       diff        lwr        upr     p adj
## Midsize-Large -3910.803709 -4333.2813 -3488.3261 0.0136270
## Small-Large   -3912.894441 -4322.8197 -3502.9692 0.0104454
## Small-Midsize    -2.090732  -459.9879   455.8064 0.9999989
```

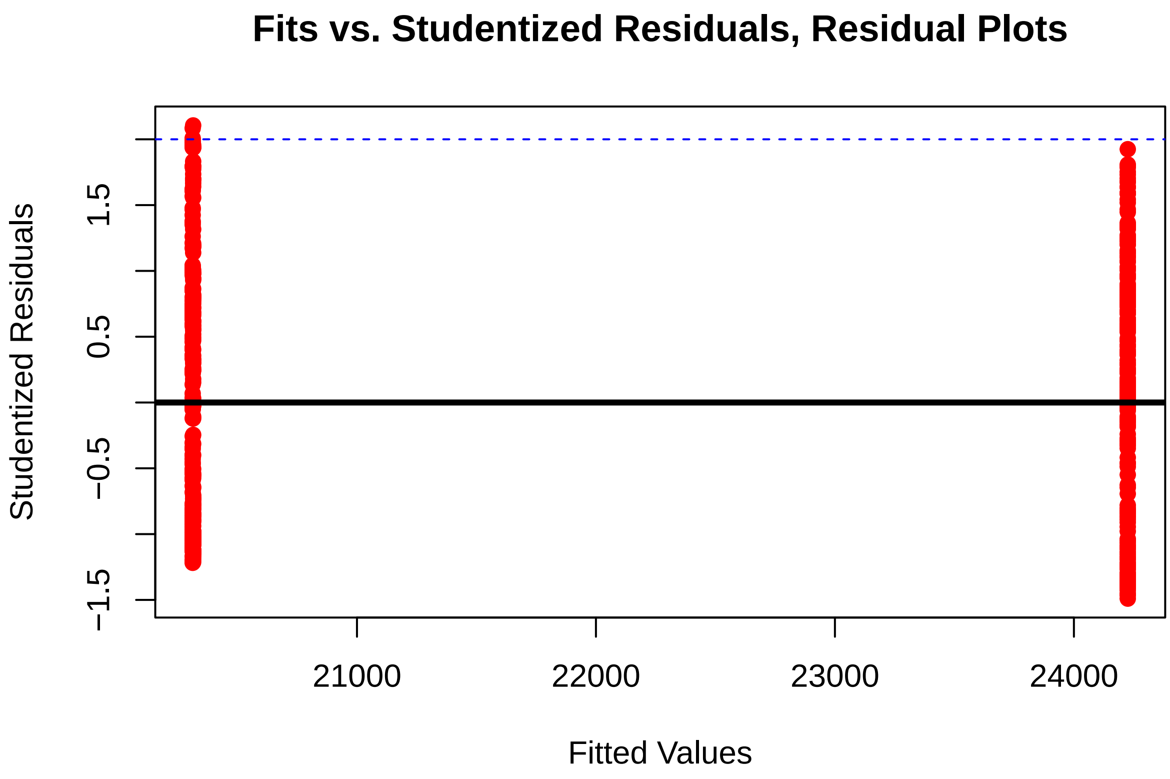
```
par(mar=c(5, 11, 4, 1))
plot(TukeyHSD(aov1), las = 1, cex.lab = 0.5)
```



Finally, is this model a good fit?

```
myResPlots2(aov1)
```





Even though the residual plot only shows two vertical lines, this is not evidence for heteroskadacity. As can be seen from the boxplot before, the tuition of midsize and large cities are in fact so close to each other that on the residual plot that they appear to be the same vertical line. Had the difference between them been larger, there would have been three vertical lines. Otherwise the residual plot shows no sign of heteroskadacity. Most data points also fall within -2 < t < 2, so there are no significant outliers. The normal quantile plot shows that most data points in the middle section of the plot around 0 fall within the blue region, though many points at the two tails are not. This shows that the residuals are fairly normally distributed except at the two tails. Overall, this model is a decent fit.

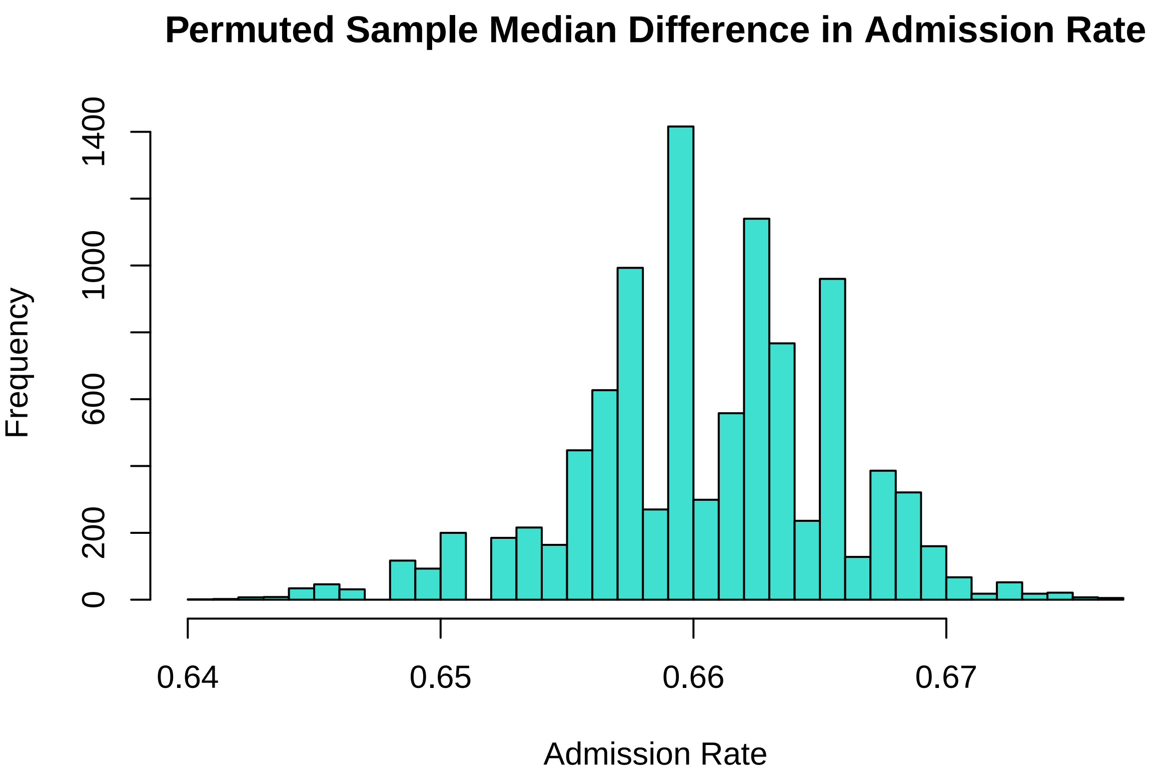
Permutation Test

Does ownership have an impact on admission rate? A permutation test is conducted to see if there is a real difference in median admission rate between public schools and private schools by taking a sample of 10,000.

```
fakeowner <- sample(df$Ownership)
(actualdiff <- median(df$AdmitRate[df$Ownership == "Private"])
  - median(df$AdmitRate[df$Ownership == "Public"]))
```

```
## [1] -0.03637797
```

```
N <- 10000
diffvals <- rep(NA, N)
for (i in 1:N) {
  fakeowner <- sample(df$Ownership)  # default is replace = FALSE
  diffvals[i] <- median(df$AdmitRate[fakeowner == "Private"])
  - median(df$AdmitRate[fakeowner == "Public"])
}
hist(diffvals, col = "turquoise",
     main = "Permuted Sample Median Difference in Admission Rate",
     xlab = "Admission Rate", breaks = 50)
abline(v = actualdiff, col = "red", lwd = 3)
text(actualdiff - 0.0022, 500, paste("Actual Difference =",
                                      round(actualdiff, 2)), srt = 90)
```



```
paste0("p-value = ", mean(abs(diffvals) >= abs(actualdiff)))
```

```
## [1] "p-value = 1"
```

The permutation test yields p = 0.0042 < 0.05, which means that the admission rates between public and private schools are statistically significantly different. This also means that about 0.42 percent of the 10000 simulations produced a difference that was more extreme than the actual difference in medians. Therefore, there is sufficient evidence to reject the null hypothesis that there is no true difference in median admission rate between between public and private schools. It's now clear that private schools have a significantly lower admission rate, i.e. that they are more selective.

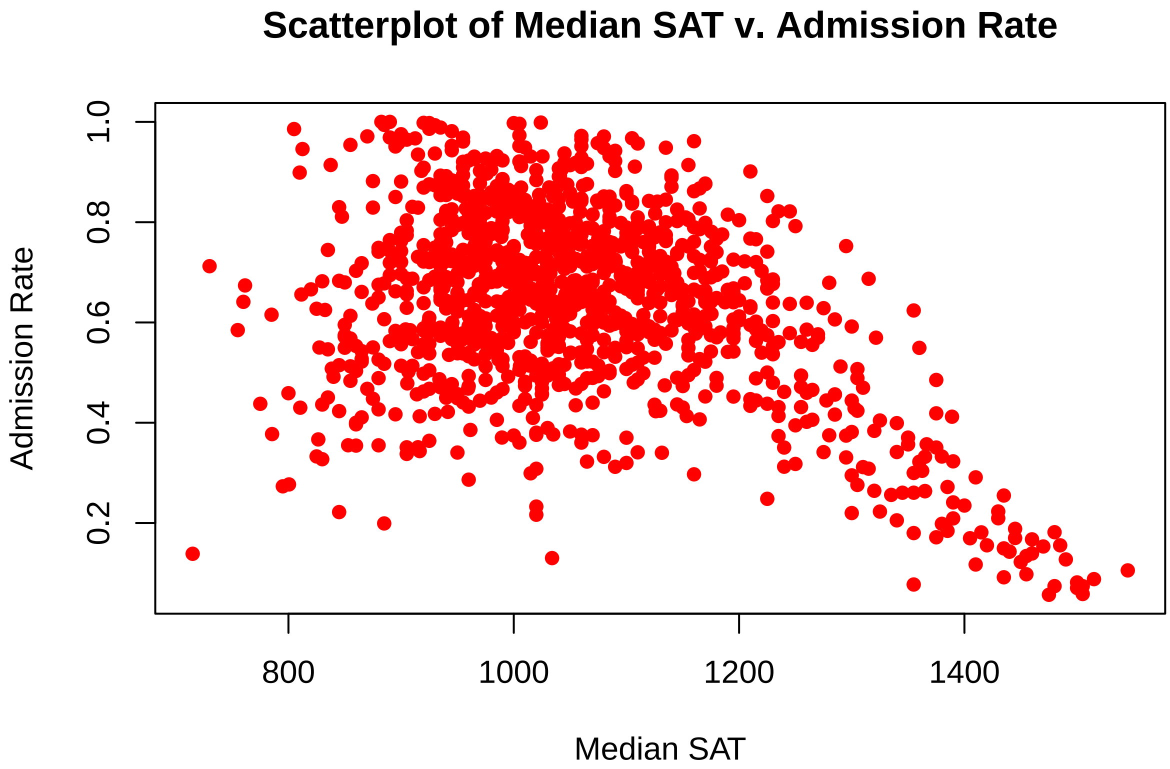
Correlation Test

What other factors might predict admission rate? Do more selective colleges require higher SAT scores? Correlation test shows that the answer is yes. The result shows that the correlation between admission rate and median SAT score is highly statistically significant with p < 2.2e-16. They have a fairly strong negative linear relationship with R = -0.42, which means that the higher the median SAT score of applicants, the lower the admission rate.

```
cor.test(df$AdmitRate, df$MidSAT)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df$AdmitRate and df$MidSAT
## t = -15.835, df = 1153, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4689005 -0.3740881
## sample estimates:
##       cor
## -0.42265
```

```
plot(AdmitRate ~ MidSAT, main = "Scatterplot of Median SAT v. Admission Rate",
     xlab = "Median SAT", ylab = "Admission Rate", pch = 16, col = "red",
     data = df)
```



However, from observing the graph, although the data seems to have a strong linear correlation after around 1200, before 1200 the data seems to be fairly randomly distributed. This is evidence for the existence of two groups in our data, so it would be unfair to mix them together. An indicator variable is created to show whether or not median SAT is above or below 1200:

```
df$above1200 <- ifelse(df$MidSAT > 1200, ">1200", "<=1200")
lm_ancova1 <- lm(AdmitRate ~ MidSAT + above1200 + MidSAT*above1200, data = df)
summary(lm_ancova1)
```

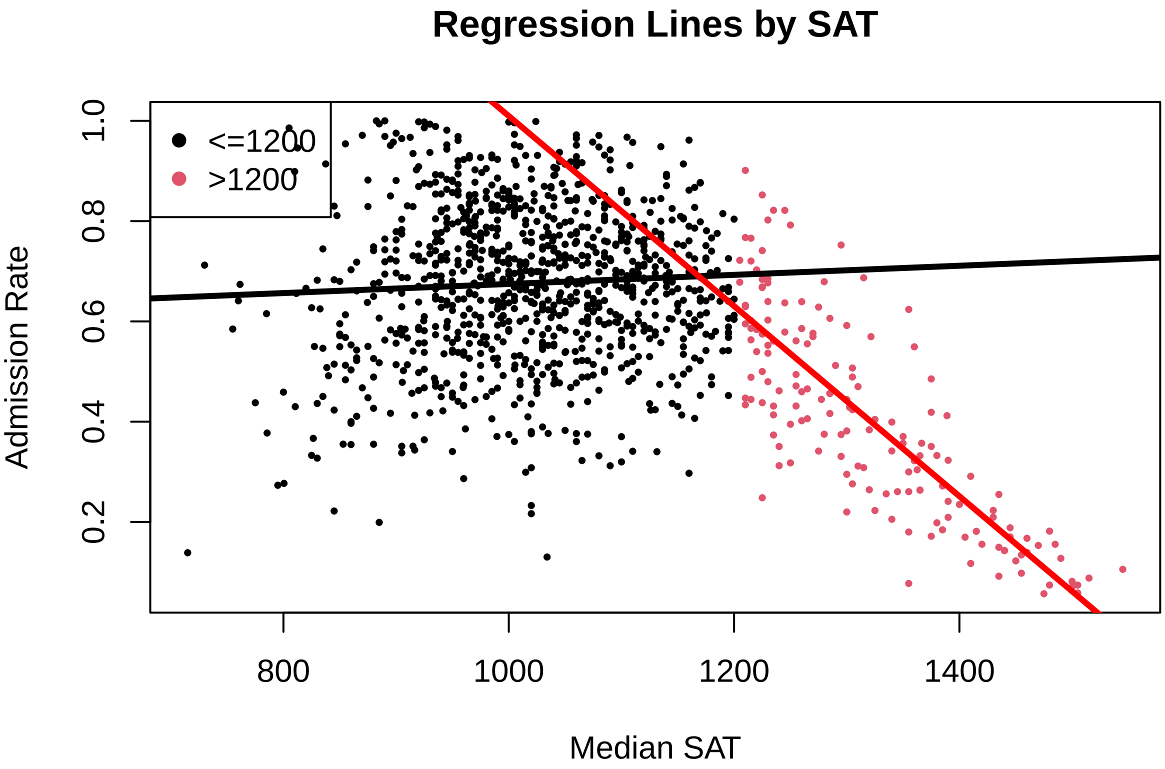
```
##
```

```
## Call:
## lm(formula = AdmitRate ~ MidSAT + above1200 + MidSAT * above1200,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.54770 -0.09732  0.00351  0.09997  0.33599
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.838e-01  5.282e-02  11.053   <2e-16 ***
## MidSAT                 9.088e-05  5.167e-05   1.759   0.0789 .
## above1200>1200         2.322e+00  1.856e-01  12.507   <2e-16 ***
## MidSAT:above1200>1200 -1.987e-03  1.444e-04 -13.760   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1501 on 1151 degrees of freedom
## Multiple R-squared:  0.3564, Adjusted R-squared:  0.3547
## F-statistic: 212.5 on 3 and 1151 DF,  p-value: < 2.2e-16
```

```
Anova(lm_ancova1, type = 'III')
```

```
## Anova Table (Type III tests)
##
## Response: AdmitRate
##                   Sum Sq   Df  F value  Pr(>F)
## (Intercept)       2.7517    1 122.1731 < 2e-16 ***
## MidSAT            0.0697    1   3.0931 0.07889 .
## above1200         3.5230    1 156.4180 < 2e-16 ***
## MidSAT:above1200  4.2646    1 189.3467 < 2e-16 ***
## Residuals        25.9237 1151
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coef <- coef(lm_ancova1)
plot(AdmitRate ~ MidSAT, col = factor(above1200), pch = 16, cex = .5,
     main = "Regression Lines by SAT",
     xlab = "Median SAT", ylab = "Admission Rate", data = df)
legend("topleft", col = 1:2, legend = levels(factor(df$above1200)), pch = 16)
abline(a = coef[1], b = coef[2], col = "black", lwd = 3)
abline(a = coef[1] + coef[3], b = coef[2] + coef[4], lwd = 3, col = "red")
```



This new graph shows that, in fact, for SAT > 1200, there is an even stronger linear, negative correlation between Median SAT and Admission Rate, whereas for SAT <= 1200, this correlation is virtually non-existent. This might be explained by the fact that the colleges consider an SAT score of 1200 or above to be a competitive score, and above this threshold colleges highly value SAT score. By contrast, for an SAT score below 1200, the colleges might basically make no distinction between students on account of their SAT score anymore. This result shows that SAT score, at for those least above 1200, is in fact a strong predictor of the school's selectivity.

Multiple Regression

A more systematic understanding of factors that influence tuition can be gained by building a multiple regression model. Since there are many variables in this model, a best subsets regression according to BIC is conducted. The following variables, deemed most relevant, are considered: total enrollment, admissions rate, test rate, median SAT, percentage of students receiving financial aid, graduation rate within 6 years, percentage of women, and percentage of non-White students. The motivation for using these variables is that, whereas factors like total enrollment, admission rate, test rate, median SAT, and graduation rate tell us about the size and competitiveness of the colleges, other factors like percentage of students receiving financial aid, of women, and of minority students tell us about the demographics of the colleges.

```
college_lm <- df[,c("Tuition", "TotalEnroll", "AdmitRate", "TestRate", "MidSAT",
                    "FinancialAidPct", "GraduationRate", "NonresidentAlienPct",
                    "WomenPct", "NonWhitePct")]
mod1 <- regsubsets(Tuition ~ ., data = college_lm, nvmax = 10)
mod1sum <- summary(mod1)
modnum <- which.min(mod1sum$bic)
college_lmtemp <- college_lm[,mod1sum$which[which.min(mod1sum$bic), ]]
college_lmfinal <- college_lm[, mod1sum$which[9, ]]
modfin <- lm(Tuition ~ ., data = college_lmfinal)
summary(modfin)
```

```
##
## Call:
## lm(formula = Tuition ~ ., data = college_lmfinal)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -27887  -5154    358   5108  34036
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -4.333e+04  5.155e+03  -8.405  < 2e-16 ***
## TotalEnroll         -5.404e-01  2.580e-02 -20.946  < 2e-16 ***
## AdmitRate           -1.803e+03  1.499e+03  -1.203    0.229
## TestRate            -1.104e+04  1.868e+03  -5.910 4.50e-09 ***
## MidSAT               2.303e+01  3.514e+00   6.555 8.41e-11 ***
## FinancialAidPct      2.605e+04  2.480e+03  10.506  < 2e-16 ***
## GraduationRate       3.728e+04  2.330e+03  16.000  < 2e-16 ***
## NonresidentAlienPct  2.878e+04  5.350e+03   5.381 9.00e-08 ***
## WomenPct             1.184e+04  2.057e+03   5.756 1.10e-08 ***
## NonWhitePct          6.506e+03  1.340e+03   4.857 1.36e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7608 on 1145 degrees of freedom
## Multiple R-squared:  0.6295, Adjusted R-squared:  0.6266
## F-statistic: 216.2 on 9 and 1145 DF,  p-value: < 2.2e-16
```

The model as a whole is very significant, with a p-value < 2.2e-16. The model also has a fairly large R-squared = 0.6295, which means that about 63% of variations in tuition can be explained by the regression model. All but one predictor, admission rate, are significant.

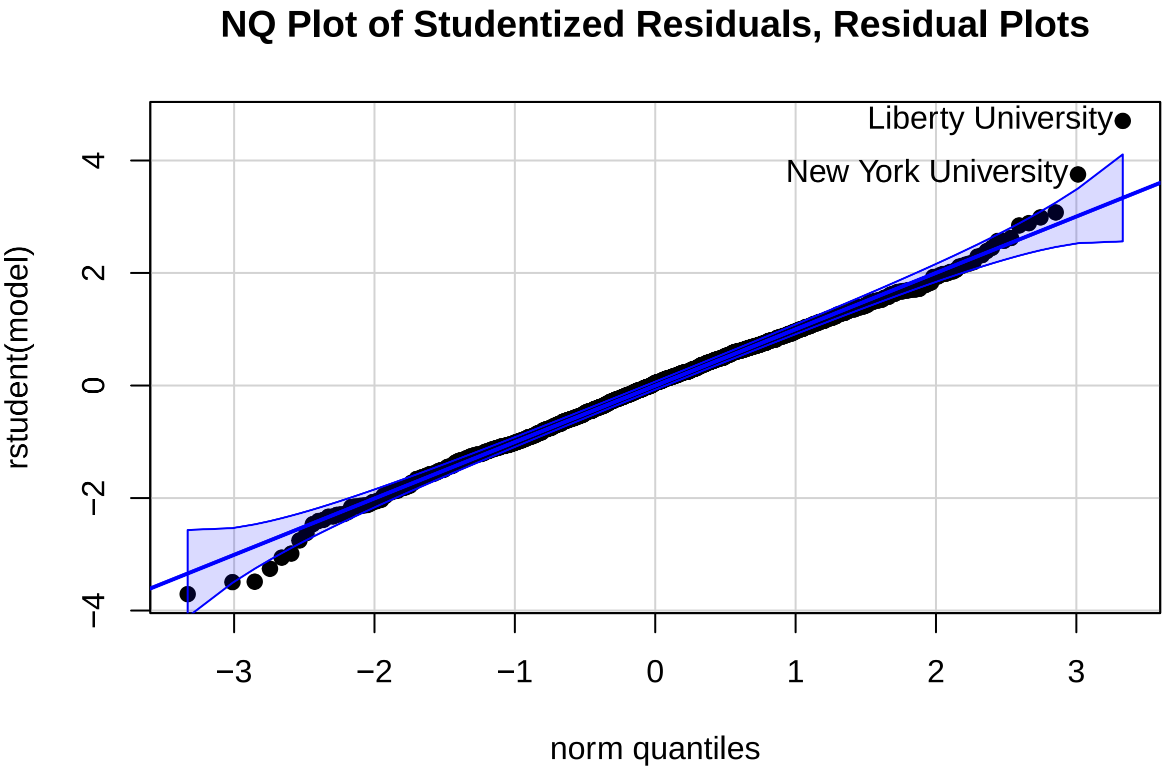
Tuition is negatively associated with total enrollment, which means that the larger the university, the lower the tuition. This makes sense because there might be an economy of scale with student body size, where more enrolled students make it possible for colleges to charge less on each individual one; large school size is also associated with public universities, a factor known to entail lower tuition.

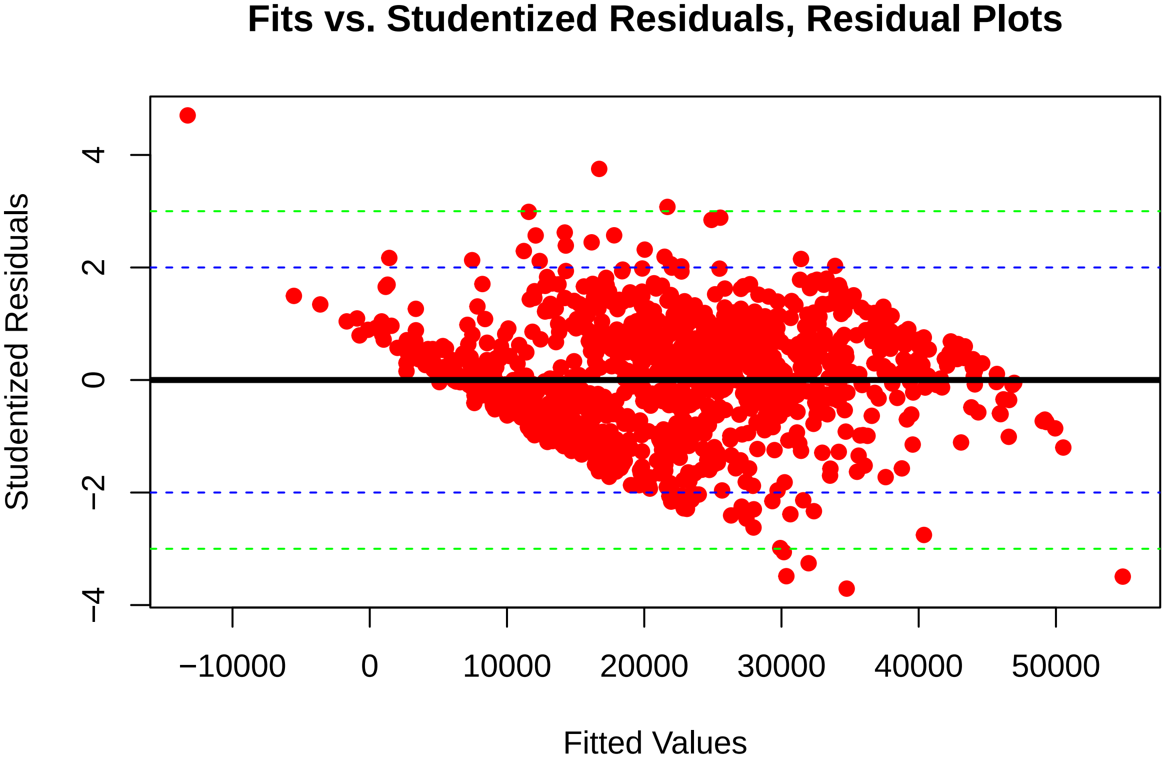
Tuition is also negatively associated with testing rate, which means that the more students who submitted their test scores, the less they are charged. This might be explained as a meritocratic tendency in colleges, especially for public schools, which look at test scores as a primary source of admission, so they are less expensive when more students submit test scores. All other predictors are positively associated with tuition. The fact that the higher the SAT, the higher the tuition may be explained by the fact that the most competitive colleges have the highest SAT score, which tend to be private schools that charge more.

Most interestingly, all four indicators of diversity – percentage of financial aid receivers, of foreign students, of women, and of non-White students – are all associated positively with tuition. This means that, in effect, **the more diverse, the more expensive.** This might be explained by competitive private colleges who tend to enforce affirmative action policies to create diversity. On the other hand, this also means that for minority students, their college experience is more expensive than their non-minority counterparts. This might reflect a systematic disparity that need further examination.

Is this model a good fit?

```
myResPlots2(modfin)
```





The normal quantile plot shows that the residuals are normally distributed. Almost all data points fall within the blue region, except several universities at the two tails: especially Liberty University, which is a for-profit institution (that also happens to advertise itself as "Flexible & Affordable"). Another outlier is New York University, which similarly has a reputation of disproportionately high tuition. The fitted plot shows no evidence of heteroskadacity. The straight edge of the data points is a result of truncation, as tuition is cut off at 0: there is no such thing as negative tuition. Otherwise there is no observable pattern. This model is thus a good fit.

ANCOVA

The regression model shows that graduation rate is a strong predictor of tuition. Does this relationship differ between public and private universities?

```
lm_ancova <- lm(Tuition ~ GraduationRate + Ownership + Ownership*GraduationRate,
                data = df)
summary(lm_ancova)
```

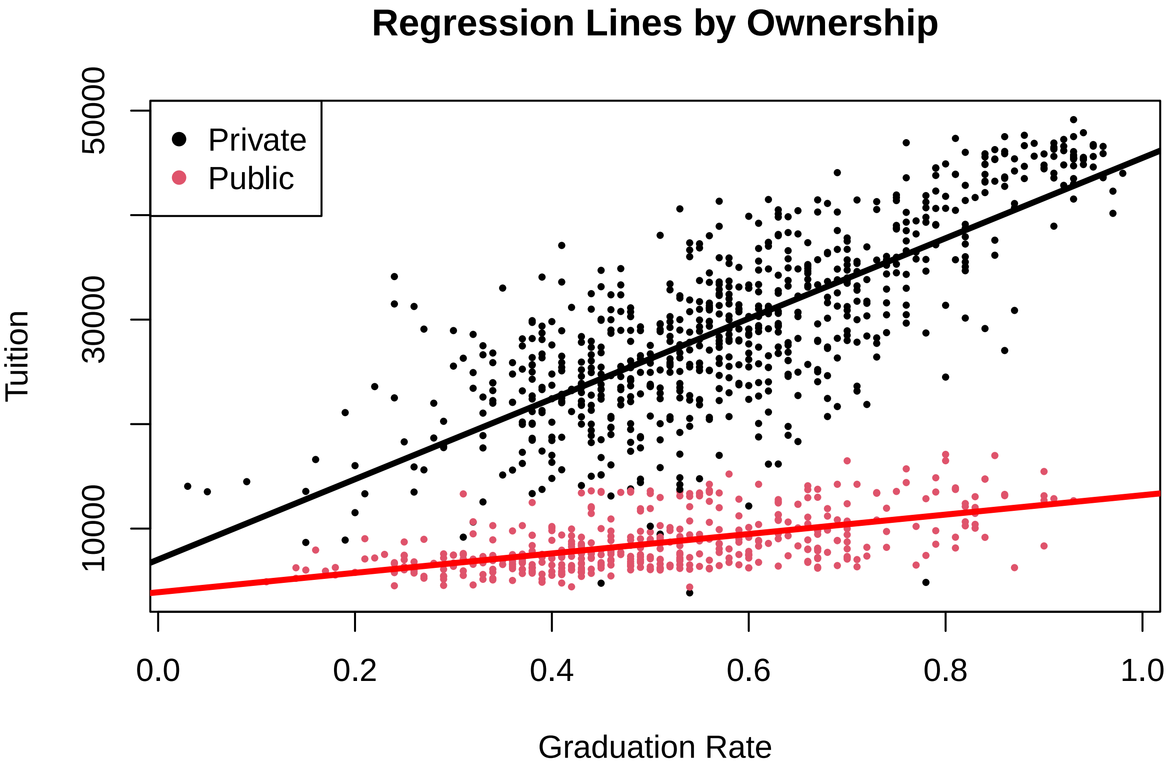
```
##
## Call:
## lm(formula = Tuition ~ GraduationRate + Ownership + Ownership *
##     GraduationRate, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32170  -2117    -51   2819  17858
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      7036.8      607.8  11.578  < 2e-16 ***
## GraduationRate                  38439.6      989.3  38.855  < 2e-16 ***
## OwnershipPublic                 -3130.4      941.3  -3.326  0.00091 ***
## GraduationRate:OwnershipPublic -29147.4     1671.0 -17.443  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4647 on 1151 degrees of freedom
## Multiple R-squared:  0.8611, Adjusted R-squared:  0.8607
## F-statistic:  2378 on 3 and 1151 DF,  p-value: < 2.2e-16
```

```
Anova(lm_ancova, type = 'III')
```

```
## Anova Table (Type III tests)
##
## Response: Tuition
##                              Sum Sq   Df F value    Pr(>F)
## (Intercept)              2.8947e+09    1  134.06 < 2.2e-16 ***
## GraduationRate           3.2599e+10    1 1509.69 < 2.2e-16 ***
## Ownership                2.3881e+08    1   11.06 0.0009101 ***
## GraduationRate:Ownership 6.5702e+09    1  304.27 < 2.2e-16 ***
## Residuals                2.4854e+10 1151
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
coef <- coef(lm_ancova)
plot(Tuition ~ GraduationRate, col = factor(Ownership), pch = 16, cex = .5,
     main = "Regression Lines by Ownership",
     xlab = "Graduation Rate", ylab = "Tuition", data = df)
```

```
legend("topleft", col = 1:2, legend = levels(factor(df$Ownership)), pch = 16)
abline(a = coef[1], b = coef[2], col = "black", lwd = 3)
abline(a = coef[1] + coef[3], b = coef[2] + coef[4], lwd = 3, col = "red")
```



It is clear from this graph that, even though for both public and private schools tuition increases as its graduation rate increases, the rates of increase are very different. The slope for private school is much steeper than that for their public counterpart. What this means in practice is that private schools charge students much more for every unit of increase in graduation rate. This should not be surprising because many private schools function more like businesses, and the graduation certificate is essentially the product for the business. As a result, the more one pays into the private universities, the more they are inclined to let them graduate. By contrast, lacking such financial incentive, in public schools this relationship is much weaker: Tuition increases slowly even as graduation rate increases.

Web Scraping

Finally, potential external factors are examined, and web scraping is used to supplement data about median income of states (National Center for Education Statistics).

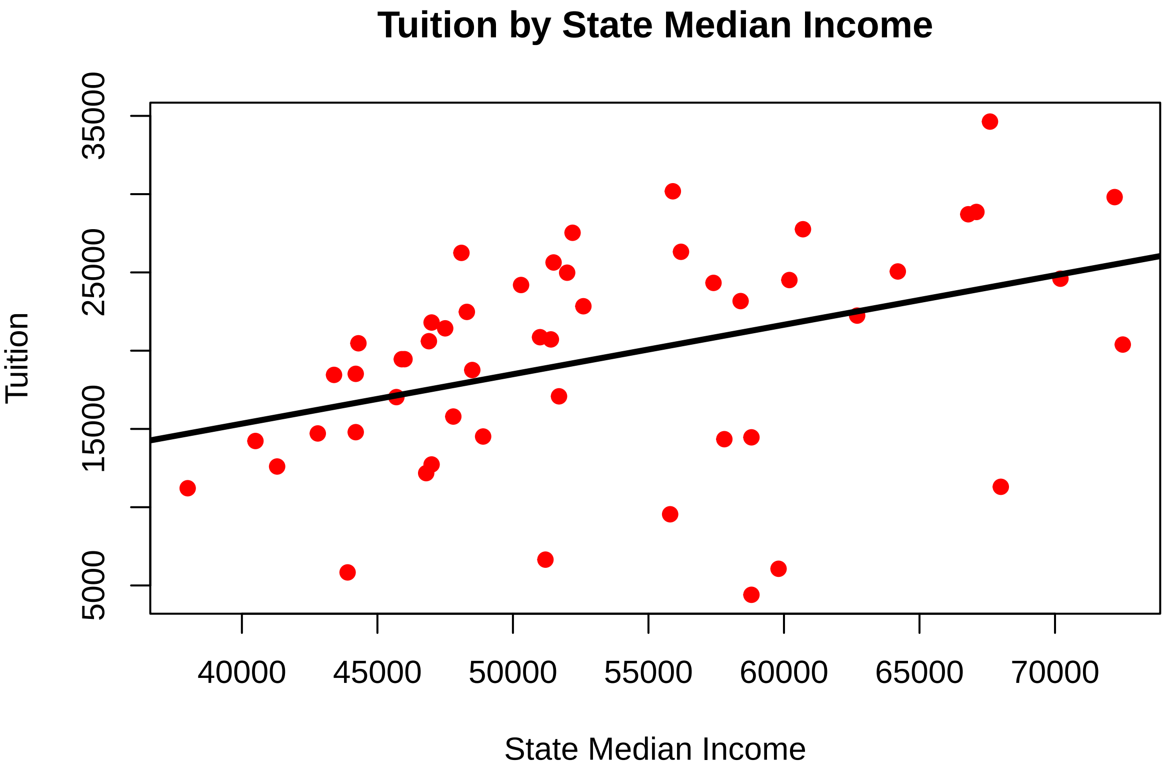
```
url <- "https://nces.ed.gov/programs/digest/d14/tables/dt14_102.30.asp"
webpage <- read_html(url)
median.incHTML <- html_nodes(webpage, 'td.TblCls005 , .TblCls011')
median.incHTML <- gsub("</td>", "", median.incHTML)
median.incHTML <- gsub("<td class=\"TblCls011\">", "", median.incHTML)
median.incHTML <- gsub("<td class=\"TblCls005\">", "", median.incHTML)
median.inc <- median.incHTML[-(1:9)*6]
median.inc <- as.numeric(gsub(",", "", median.inc))
```

After data cleaning, a correlation between tuition and state median income is examined, grouping colleges by state.

```r
tuitionbystate <- by(df$Tuition, df$State, mean)
df_new <- data.frame(cbind(tuitionbystate, median.inc))
cor.test(df_new$tuitionbystate, df_new$median.inc)
```

```
##
##  Pearson's product-moment correlation
##
## data:  df_new$tuitionbystate and df_new$median.inc
## t = 3.0684, df = 49, p-value = 0.003501
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1415462 0.6096071
## sample estimates:
##       cor
## 0.4014694
```

```r
plot(tuitionbystate ~ median.inc, data = df_new,
     main = "Tuition by State Median Income", xlab = "State Median Income",
     ylab = "Tuition", pch = 19, col = "red") +
abline(lm(tuitionbystate ~ median.inc, data = df_new), lwd = 3)
```



```
## integer(0)
```

There is a positive linear correlation between them, with p = 0.003 < 0.05 and R = 0.40. This is plausible because the state median income is indicative of the state's price level, which in turn has an impact on college tuition.

Bootstrap

This correlation is bootstrapped using a sample of 10,000.

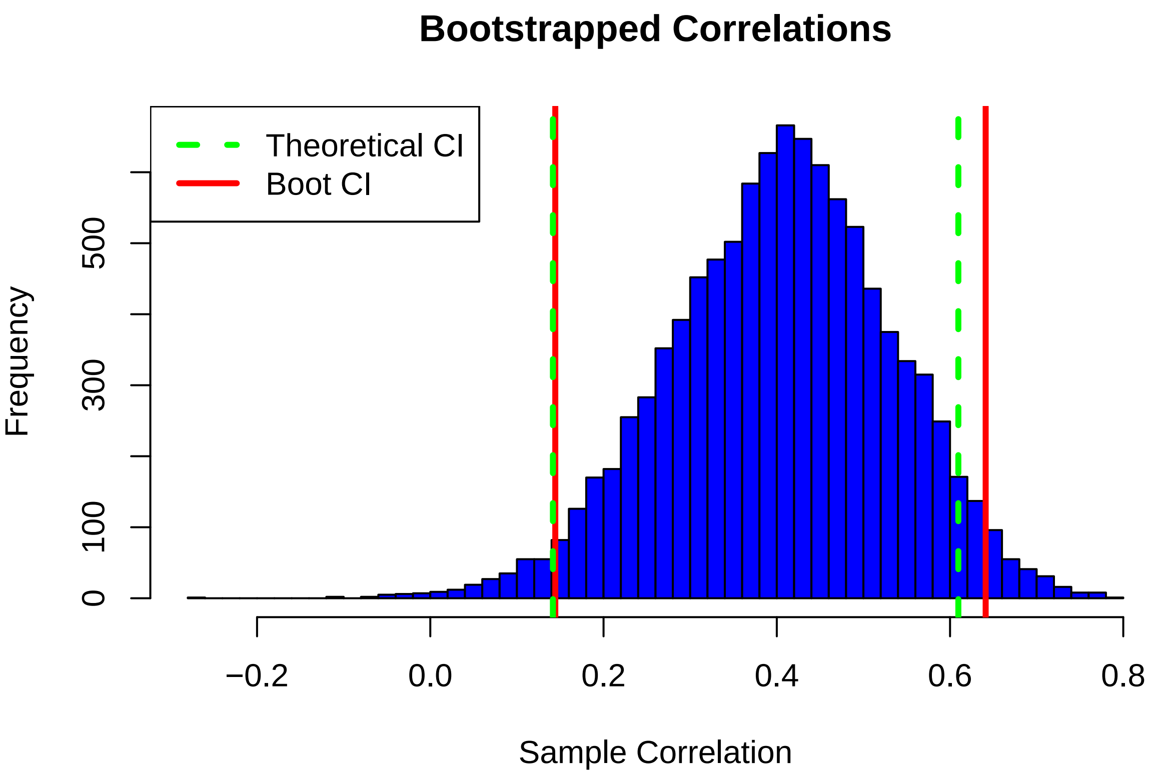
```
N <- nrow(df_new)
n_samp <- 10000
corResults <- rep(NA, n_samp)
bresults <- rep(NA, n_samp)
for(i in 1:n_samp){
  s <- sample(1:N, N, replace = T)
  fakeData <-  df_new[s, ]
  corResults[i] <- cor(fakeData[,1], fakeData[,2])
  bresults[i] <- lm(fakeData[,1] ~ fakeData[,2])$coef[2]
}
(ci_r <- quantile(corResults, c(.025, .975)))
```

```
##      2.5%     97.5%
## 0.1442876 0.6411170
```

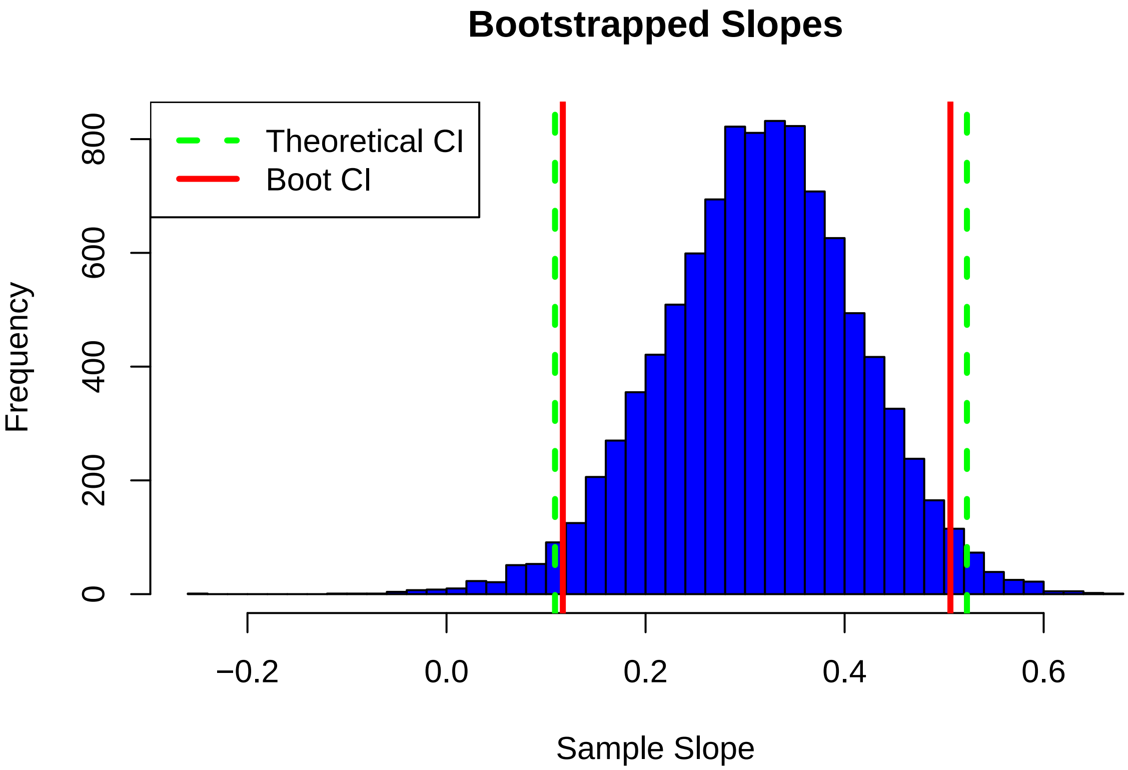
```
(ci_slope <- quantile(bresults, c(.025, .975)))
```

```
##      2.5%     97.5%
## 0.1168779 0.5062799
```

```
hist(corResults, col = "blue", main = "Bootstrapped Correlations",
     xlab = "Sample Correlation", breaks = 50)
abline(v = ci_r, lwd = 3, col = "red")
abline(v = cor.test(df_new$median.inc, df_new$tuitionbystate)$conf.int,
       lwd = 3, col = "green", lty = 2)
legend("topleft", c("Theoretical CI","Boot CI"), lwd = 3,
       col = c("green","red"), lty = c(2, 1))
```



```r
lm1 <- lm(tuitionbystate ~ median.inc, data = df_new)
hist(bresults, col = "blue", main = "Bootstrapped Slopes",
     xlab = "Sample Slope", breaks = 50)
abline(v = ci_slope, lwd = 3, col = "red")
abline(v = confint(lm1, "median.inc"), lwd = 3, col = "green", lty = 2)
legend("topleft", c("Theoretical CI","Boot CI"), lwd = 3,
       col = c("green","red"), lty = c(2, 1))
```



From the histogram it can seen that the Boot CI for correlations is wider than the theoretical CI. This might be because there are a few points that tend to inflate the correlation in our data set. By contrast, the Boot CI for slopes is narrower than the theoretical CI. This means that we are more confident about the slope of the correlation between tuition and state median income after bootstrapping. In any case, the theoretical and Boot CIs are basically identical, so it should not decrease confidence in the result too much.