Phase 3: Predicting Student Debt

Sierra Allred

July 27, 2016

Many college students today incur vast amounts of debt, sometimes much more than they bargained for when they entered school. It would be useful to have a model that predicted the amount of debt a student would acquire based on a number of factors. The goal of this phase of my project was to create a linear model to predict student debt based on the total cost of a school, the cost of living (COL) in that area, average SAT score, median ACT score, and average earnings 10 years after graduation.

1 Data

The data I am using for my analysis is taken from the College Scorecard Dataset and AreaVibes.com. The Scorecard data includes student debt after graduation (DEBT), average earnings 10 years after graduation (EARN), average SAT score (SAT), median ACT score (ACT), total cost (COST), and whether a school was public or private non-profit. The AreaVibes data includes the cost of living in the city each school is located in.

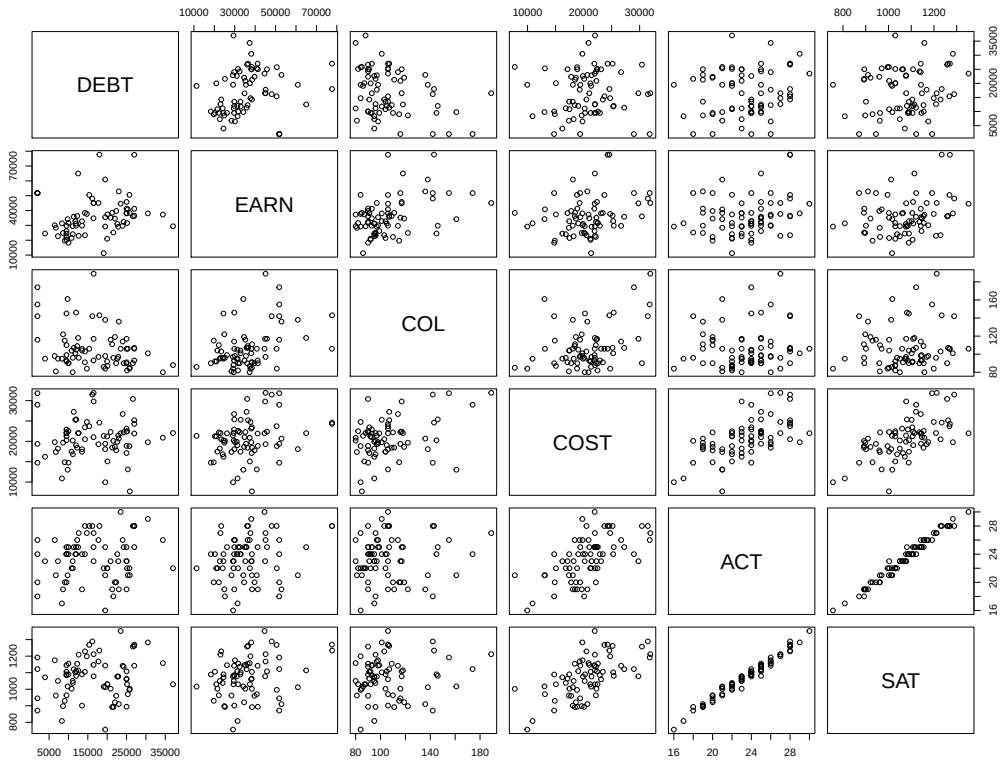
Unfortunately there is a lot of missing data about many of the schools in the Scorecard dataset, thus there are only 246 complete data points to work with. After some initial analysis I decided to control for the type of institution by creating separate models for public and private institutions. I found that private institutions are on average more expensive than public institutions, and I do not want that to affect the result if I do not separate them. 173 schools fall into the private category, and 73 are public.

2 Hypothesis

I predict that there will be certain factors that show a strong correlation with debt, such as total cost, test scores, and cost of living. However, because debt is largely based on factors in the personal life of may students, I do not expect to find a model that predicts debt very accurately. I foresee a lot of correlation between the factors other than debt, and am prepared to remove them if the correlation is too significant.

3 Predicting Student Debt - Public

The first thing I did was get to know the data by finding and comparing the correlation between each data set and plotting them against each other.



The figure above shows the correlation matrix between all factors. ACT and SAT test scores are very correlated which told me that schools have a good scale when comparing SAT and ACT scores. Because of this I had to be careful not to use a model with both factors. There is also quite a bit of correlation between cost and test scores. It is important to note that there is not a high or noticeable correlation between debt and any other factor.

I then began with a model that included all the given factors used to predict debt. A summary of this model showed that earnings was significant at alpha = .01 and COL was significant at alpha = .001, with no other factors showing significance. The adjusted R^2 for the model was .1697, not satisfactory for making predictions. This was expected however, and I ran the model through the stepAIC command in R. The best model it suggested to predict debt was

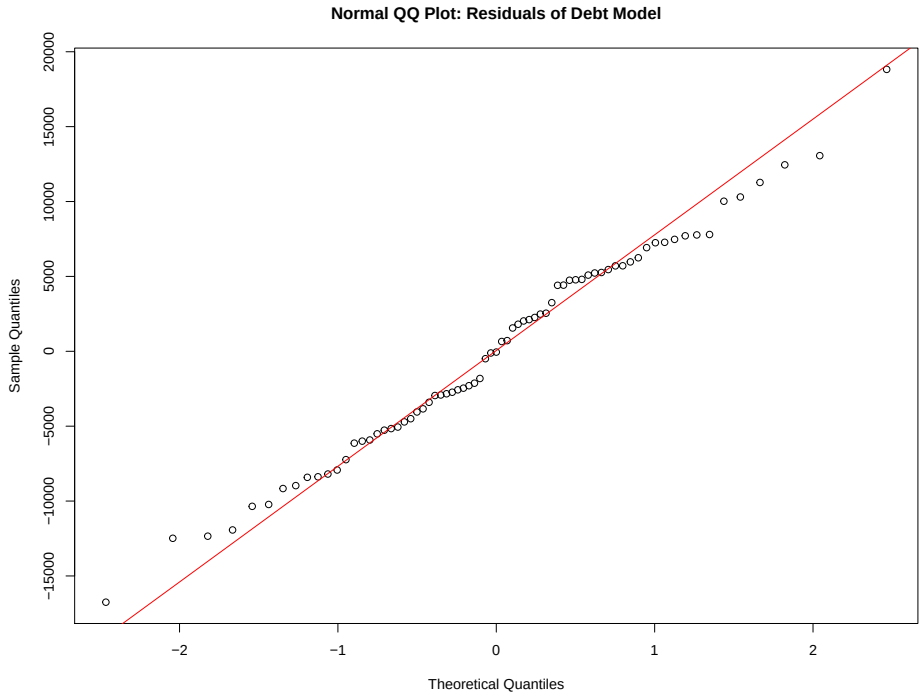
$$predictedDebt = 0.2308 * (EARN) - 164.0901 * (COL) + 25832.6869$$

with an AIC (Akaike information criterion) of 1508. I then used this model to predict the values for debt.

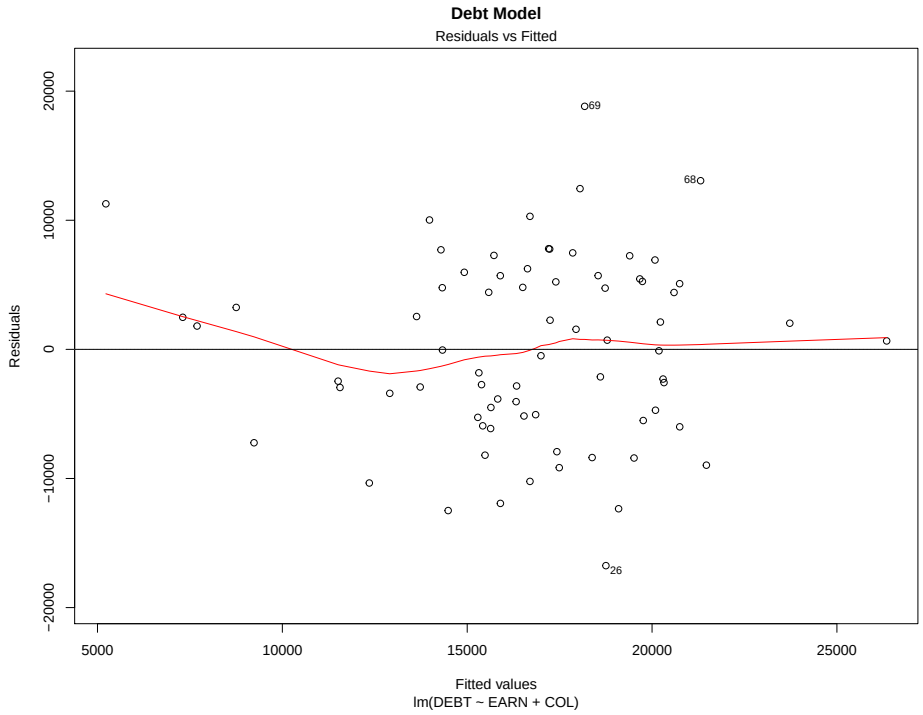
Debt Model Assumptions

Before using a multiple regression model to predict debt, I needed to check that the data fulfilled all the model assumptions. Namely, I needed to ensure that the residuals of the model were normally distributed, that there was homoscedasticity or constant variance of the errors, a lack of multicollinearity or highly correlated independent variables, an absence of significant outliers, and that the observations were independent.

First, to check the normality of our residuals, I used a normal QQ-plot. The data points falling nicely along the 45 line, implies that the residuals are normal. I also ran the Shapiro-Wilks test that gave a p-value of 0.7535, also strongly implying normality.



Second, to test for homoscedasticity or constant variance of the errors, I plotted the standardized residuals as a function of the standardized predicted values. The data points appear to be spread out fairly evenly, with no evident widening or narrowing pattern, implying that the variance of the errors are constant.



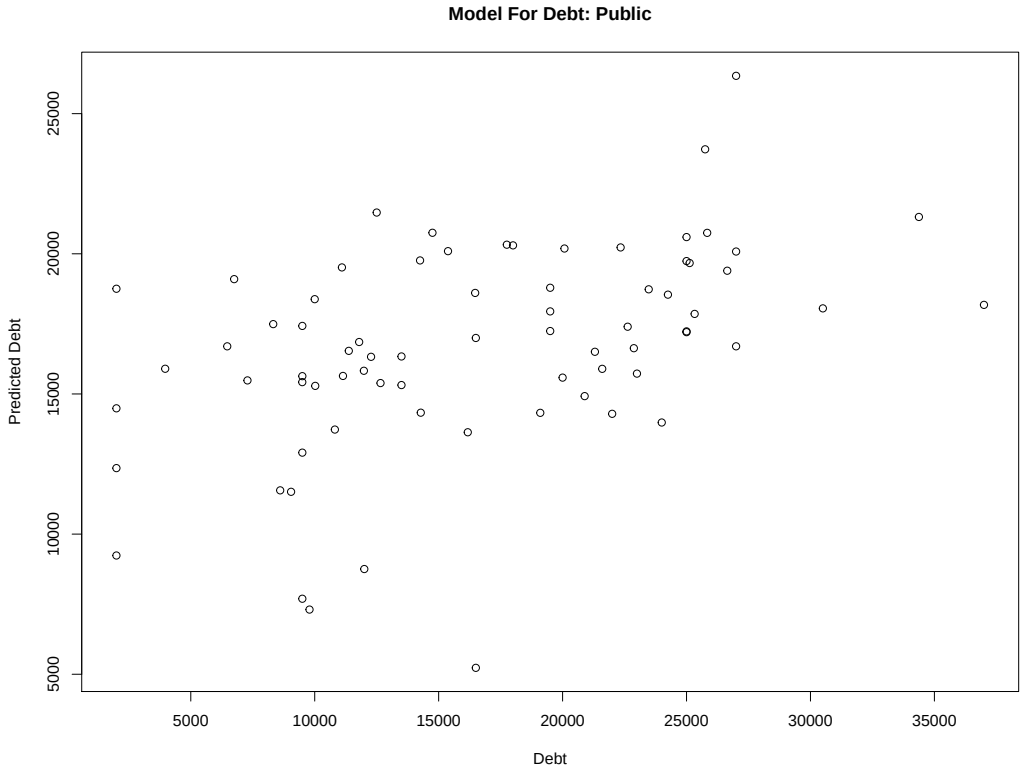
Third, to test for a lack of multicollinearity or significant correlation between independent variables, I checked the correlation between our two independent variables in the model, COL and earnings. The correlation coefficient was 0.4064985, which is not significant.

Fourth, the data contained no outliers that were affecting the fit of the model.

Finally, independence of the observations is dependent on the manner in which the data was gathered. I had no control over data collection so I assumed appropriate sampling techniques were employed that ensured the independence of the observations.

All of the assumptions of my model were met and I decided to proceed.

I plotted the predicted values given by our new debt model against the actual debt values to get the plot you see below.

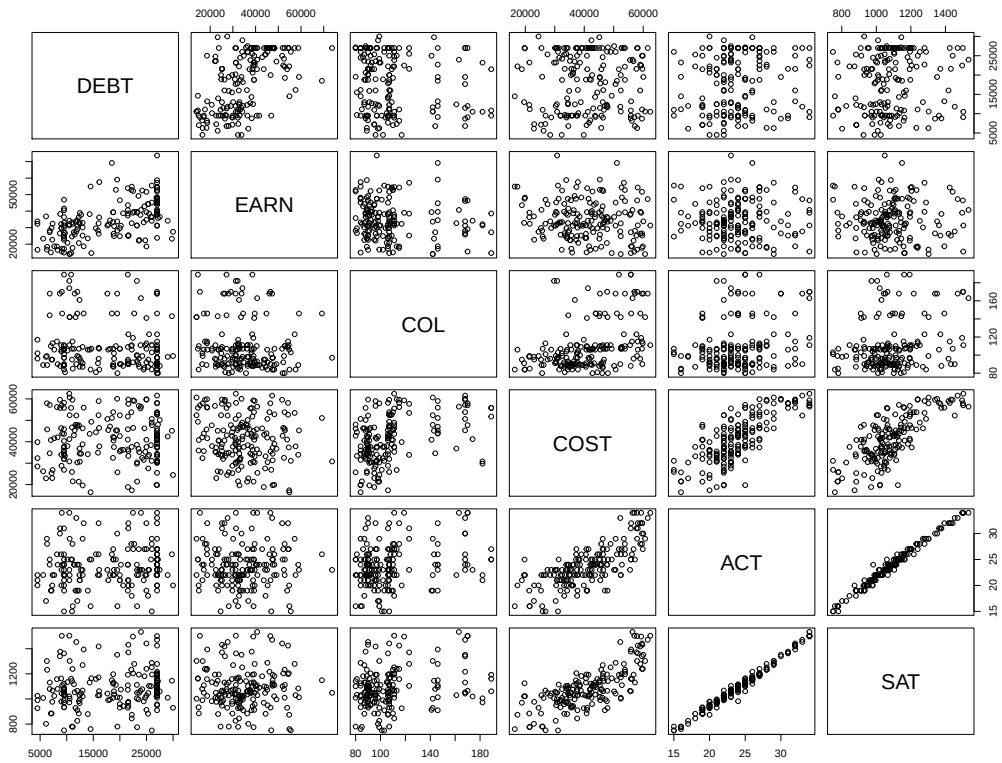


The correlation between the predicted values and the actual debt was $R^2 = .459$, which is not very significant. I looked at a summary of the model and found that the adjusted R^2 was .1997, also not significant. The factor EARN was significant at alpha= .01 with a p-value of 0.0026, and COL was significant at alpha= .001 with a p-value of 0.00015.

After deciding that this model was not satisfactory, I continued by transforming different factors in the data based on their QQ-plots and histograms. I created many different models and was unable to find any significant results.

4 Predicting Student Debt - Private

I used the same methods when finding a debt model for private schools as I did for public schools. Below is the correlation matrix for variables related to private schools.



ACT and SAT test scores are very correlated like they were for public schools. There is also an even higher correlation between cost and test scores. With an R^2 value of .55 debt and earnings have a noticeable correlation.

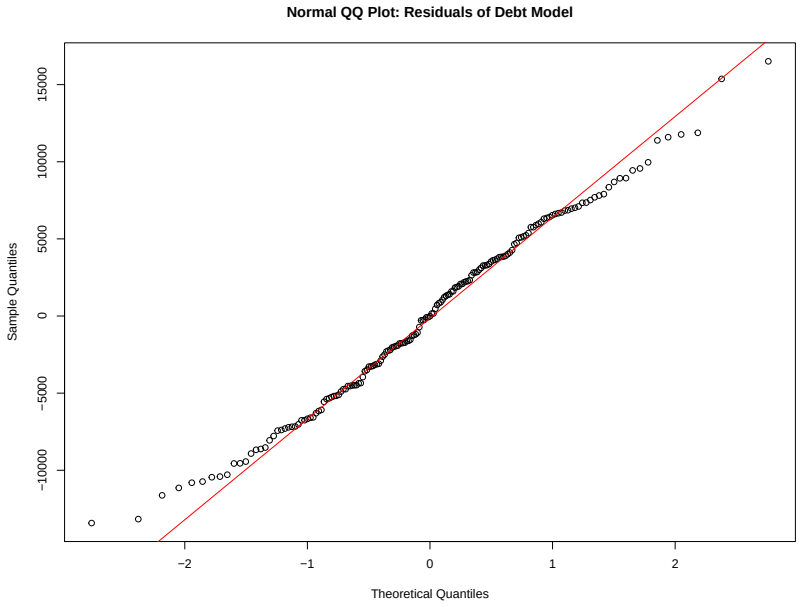
I started with model that included all the given factors. A summary of this model showed that earnings was significant at alpha = .001 and COL was significant at alpha = .05, with no other factors showing significance. The adjusted R^2 for the model was .3277, not satisfactory for making predictions. The best model the stepAIC command suggested to predict debt was

$$predictedDebt = .3493 * (EARN) - 34.8309 * (COL) + 7.6695 * (SAT) + 1336.1831$$

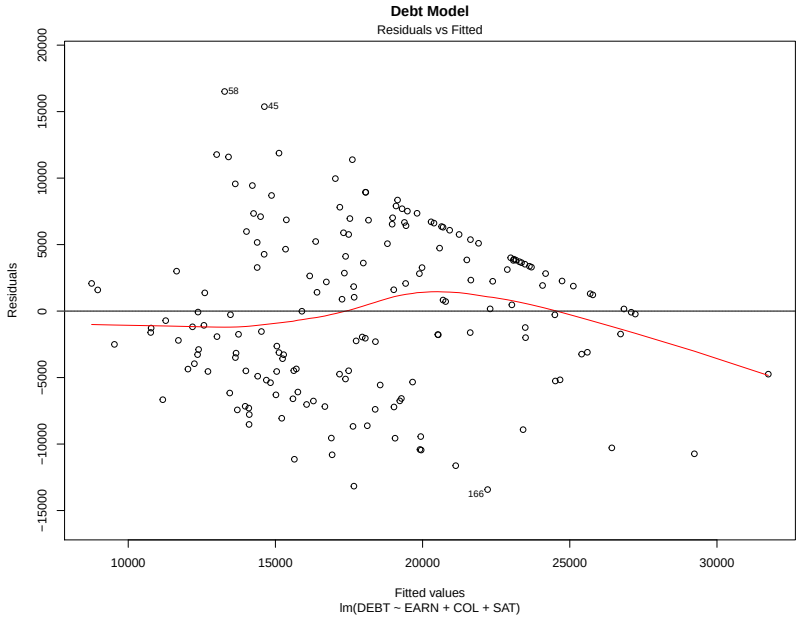
with an AIC of 3514. I chose to proceed with this model.

Debt Model Assumptions

Normality: The QQ-plot implies that the residuals of my model are normal. The Shapiro-Wilks test that gave a p-value of 0.2766, also implying normality.



Homoscedasticity: The plot of the standardized residuals shows a slight narrowing pattern, implying that the variance of the errors might not be constant. I decided to proceed with caution.



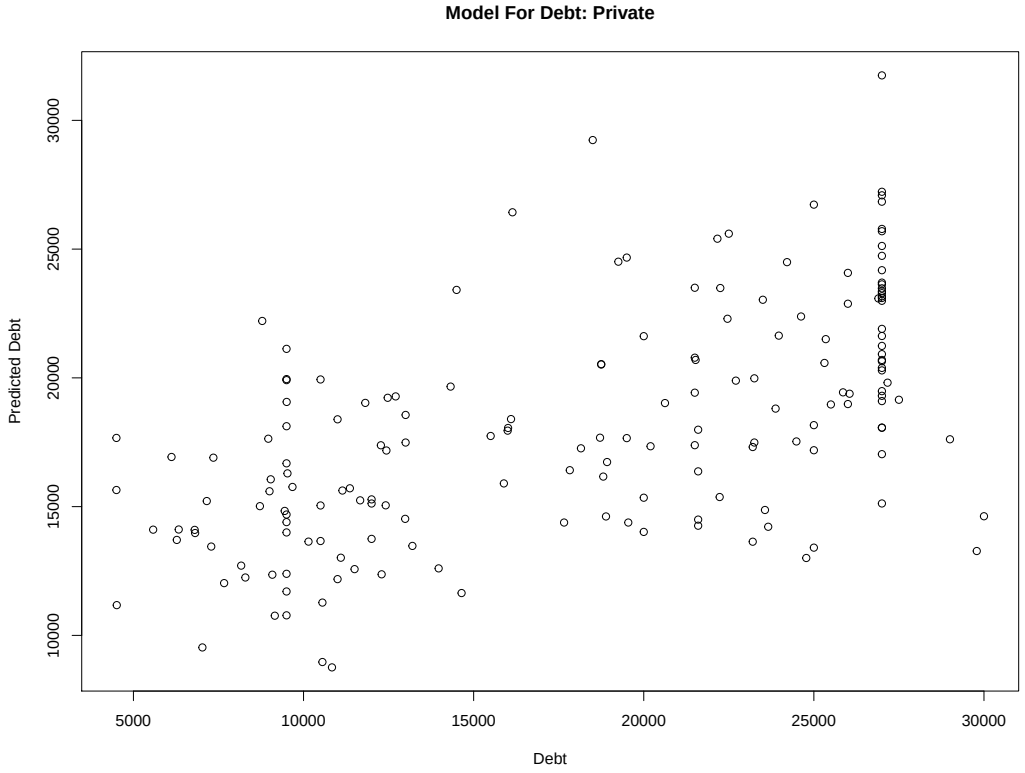
Correlation between variables:
For EARN and COL, R^2 = -.115
For COL and SAT, R^2 = 0.26
For EARN and SAT, R^2 = -0.015
None of the correlation coefficients are significant.

Outliers: There were no outliers.

Independence of Observations: I assumed appropriate sampling techniques were employed.

All of the assumptions of my model were met and I decided to proceed.

Below is a plot of the predicted values against the actual debt values.



The correlation between the predicted values and the actual debt was not significant at $R^2 = 0.584$. The adjusted R^2 for the model was also not significant at 0.33, although it was better than the model for public schools. EARN was significant at alpha= 0.001 with a p-value close to 0, SAT was significant at alpha= 0.0111, and COL was barely significant at alpha= 0.1 with a p-value of 0.0567.

I concluded that the model I created could not be used for prediction. I used transformative techniques to create a few more models and did not find any significant results.

5 Interpretation and Future Work

At first I was surprised not to find a high correlation between debt and many of the factors I used. My reasoning at the beginning of my investigation was that the greater the cost of attendance and the cost of living at a school, the higher the debt afterwards. I also thought that higher earnings after college would mean higher debt because the student pursued a longer or harder degree that took a lot more time or tuition. This seemed to be somewhat true for private schools, where I found a fairly high correlation between debt and earnings compared to all other factors.

After my analysis I realized why my results were not as significant as I had hoped. Debt is not as affected by the money required to attend a school as it is by the financial situation of the student. If I were to continue this project I would like to see what information I could gather about individuals financial situation before college as opposed to after, and add in information about the scholarships and financial aid received, to see how debt is affected by these factors. I would also like to find more complete data so I could have a larger sample size to base my model on.

I would probably have had better results if I had chosen to predict the cost of a school, because schools that have a higher total cost tend to have higher test scores, are in wealthier areas with a higher cost of living and might even produce students with higher earnings. Students may not care as much about the average debt of students at the schools they are looking at as they care about how much it will cost them to go there because of their individual financial situation. Because of this, a model for cost could be more useful than a model for debt. This is an area I would be interested in pursuing given more time.

I had hoped to see how the cost of living data and other variables, especially earnings after graduation, but found no significant correlations. One thing that might have affected this was that many students aren't from the same city where the school is located, and things that happened before school, such as test scores, would not be affected. For earnings after graduation there is a 10 year gap, and it is likely that students move to new cities after they graduate.

6 Conclusion

I have learned that it is not uncommon to look for relationships in the data and find that they do not exist. The level at which the data was transformed and combed through hardly changed the strength of my prediction model. It was clear that a lot of my data was uncorrelated. When testing models I found that it was helpful to follow this process:

1. Proceed if the model assumptions are met.
2. Look for relevant relationships between the given variables.
3. If none exist, consider data transformations.
4. If the relationships continue to be non-existent or unclear

a. Consider starting over (trying new transformations)
b. Give up, (no relationship truly exists)

When it came to predicting student debt there came a point that I had to accept the best models I found, although they were not accurate. This led me to understand that there may be other factors that would be more useful at predicting student debt in the future. Students debt is an increasing problem in the United States, and through this process I gained an understanding of how complicated that problem can be. The cost of institutions were not at all related to the average debt students had after graduation, which means that the instinctive solution of simply reducing the cost of schools may not be the most effective solution by itself. Another option, and one that this project has tried to achieve, is to help students make educated decisions about the schools they choose.

7 Code

I have included the R code I used to find the model for public schools. The code used for private schools was very similar and uses the same functions.

```
data1 = read.csv("Public.csv", header=TRUE)
pairs(data1)

COL <- data1$COL
COST <- data1$cost
EARN <- data1$earnings
DEBT <- data1$debt
ACT <- data1$medACT
SAT <- data1$avgACT
CTRL <- data1$CONTROL

pairs(DEBT~EARN+COL+COST+ACT+SAT)
pairs(DEBT~EARN+ACT+SAT)

plot(COL, COST)
plot(COL, EARN)
abline(lm(EARN~COL), col="red")

plot(COL, DEBT)
plot(COL, ACT)
plot(COL, SAT)
plot(COL, CTRL)

plot(COST, EARN)
plot(COST, DEBT)
abline(lm(DEBT~COST), col="red")
plot(COST, ACT)
```

```
plot(COST, SAT)
plot(COST, CTRL)

plot(EARN, DEBT)
abline(lm(DEBT~EARN), col="red")
plot(EARN, ACT)
plot(EARN, SAT)
plot(EARN, CTRL)

plot(DEBT, ACT)
plot(DEBT, SAT)
plot(DEBT, CTRL)

plot(SAT, ACT)
plot(ACT, CTRL)
plot(SAT, CTRL)


?cor
cor(COL, EARN, use='pairwise') #.4
cor(COL, COST, use='pairwise') #.36
cor(COL, ACT, use='pairwise') #.085
cor(COL, SAT, use='pairwise') #.098

cor(COST, EARN, use='pairwise') #.26
cor(COST, ACT, use='pairwise') #.6
cor(COST, SAT, use='pairwise') #.6

cor(COST, DEBT, use='pairwise') #.027
cor(EARN, DEBT, use='pairwise') #.17
cor(COL, DEBT, use='pairwise') #-.31
cor(DEBT, ACT, use='pairwise') #.137
cor(DEBT, SAT, use='pairwise') #.128

cor(EARN, ACT, use='pairwise') #.23
cor(EARN, SAT, use='pairwise') #.25
cor(ACT, SAT, use='pairwise') #.987



res.lm =glm(DEBT~EARN+COL+COST+ACT+SAT) #EVERYTHING
summary(res.lm)
#EARN is significant to .01, COL to .001
#AIC = 1512.7
model1.lm =lm(DEBT~EARN+COL+COST+ACT+SAT)
model1.lm
```

```
summary(model1.lm)
#Adj R^2 .1697 F-statistic: 3.944 on 5 and 67 DF,  p-value: 0.003405
par(mfrow=c(2,2))
plot(res.lm)
par(mfrow=c(1,1))
library(MASS)
stepAIC(res.lm)
# Call:  glm(formula = DEBT ~ EARN + COL)
#
# Coefficients:
#(Intercept)          EARN           COL
# 25832.6869        0.2308     -164.0901
#
# Degrees of Freedom: 72 Total (i.e. Null);  70 Residual
# Null Deviance:          4.56e+09
# Residual Deviance: 3.598e+09  AIC: 1508

### Assumptions to Predict Debt ###
fitDebt= lm(DEBT~EARN+COL)
# check for normality
shapiro.test(residuals(fitDebt)) # Shapiro- Wilks Test
qqnorm(residuals(fitDebt), main= "Normal QQ Plot: Residuals of Debt Model")
# QQ plot
qqline(residuals(fitDebt), col= "Red")
# check for constant variance
plot(fitDebt, which= 1, main= "Debt Model") #residual plot
abline(0,0)
# check for non-correlated independent variables
cor(EARN, COL) #.4

#formula = DEBT~EARN+COL #AIC 1508
predictedDebt =  0.2308*(EARN) - 164.0901*(COL) + 25832.6869
plot(DEBT, predictedDebt, ylab = "Predicted Debt", xlab="Debt")
cor(DEBT, predictedDebt) #.459
x.lm =lm(predictedDebt~DEBT)
summary(x.lm) #Adj R^2  0.1997
abline(x.lm)
par(mfrow=c(2,2))
hist(DEBT)
qqnorm(DEBT)
hist(predictedDebt)
qqnorm(predictedDebt)
par(mfrow=c(1,1))
model2.lm =lm(DEBT~EARN+COL)
summary(model2.lm) #Adj R^2 0.1883
abline(model2.lm)
```

```
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept)  2.583e+04  4.177e+03   6.184  3.7e-08 ***
#   EARN          2.308e-01  7.412e-02   3.114 0.002678 **
#   COL          -1.641e+02  4.095e+01  -4.007 0.000151 ***
#   ---
# Signif. codes:  0 '***' .001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 7170 on 70 degrees of freedom
# Multiple R-squared:  0.2108,  Adjusted R-squared:  0.1883
# F-statistic:  9.35 on 2 and 70 DF,  p-value: 0.0002519

par(mfrow=c(2,2))
plot(model2.lm)
par(mfrow=c(1,1))


#Seeing if this models were any good
model4.lm =lm(DEBT~EARN*COL)
summary(model4.lm)#Adj R^2  0.2085
par(mfrow=c(2,2))
plot(model4.lm)
par(mfrow=c(1,1))
#I had a great amount of code here that tried different models.
#None of them gave me anything interesting, so I have left them
#out of the index in interest of readability and space.
```