36-402 Data Analysis Exam 1

April 7, 2020

1 Introduction

The decision of going to college after high school is not always an easy choice to make, as many risk the financial burden placed by college until their 40s and 50s. In an ever changing economy with no guarantees, how can one make the most informed decision about the next step in their future? With steadily increasing tuition costs, is going to college worth it? How much can one expect to make upon graduation, and which universities provide the greatest value and financial security? The United States Department of Education looks to ease these burdens placed on students by providing financial-aid through Pell grants and loans, as well as valuable insights into thousands of institutions through the Collegeboard Scorecard website. **(1)** As a consequence, we have been tasked with analyzing any relationship between the cost of tuition (post financial-aid) and the median salaries earned post-graduation (10 years after entry). We would also like to explore potential differences between attending a public and private institution, and whether attending private universities is justified. Finally, we would also like to gain some insight as to how much we, as Carnegie Mellon students and students from similar institutions, can expect to make post graduation. **(2)** After an extensive study which evaluated various models and metrics, we concluded that price, upon adjusting for prior education and economic status, does have a significant impact on the average median earnings 10 years post-graduation, but the relationship is not necessarily linear. We also found that different institution types have a significant difference in the relationship that they have when it comes determining the amount of earnings post-graduation, and that students working and not-enrolled 10 years after entry from institutions similar to Carnegie Mellon can expect to make between $17,656.97 and $99,158.50.

2 Exploratory Data Analysis

2.1 Data

The data being used is from the College Scorecard website (2017), which is maintained by the United States Department of Education to better inform students as they make their enrollment decisions. The data consists of 1,294 institutions - private for-profit, private non-profit, and public universities. It is important to note that the data only consists of students who either received a loan or a grant from the federal government. Data collected from each institution included the type of institution (CONTROL), total undergraduate enrollment in the institution (UGDS), the net average price (PRICE, in dollars), the average SAT scores for admitted students (SAT, out of 1600), the percent of students that received a Pell grant (PCTPELL), the percent of students that received a federal loan (PCTFLOAN), the median earnings of students 10 years post-graduation (EARN, in dollars), and the median debt of those who completed their degrees (DEBT, in dollars).

2.2 Exploration

We began by examining the univariate distributions of each predictor variable, as well as the response variable (EARN). Immediately, we notice that the amount of data that we have on private for-profit institutions only consists of 0.3% of the entire dataset (5 of 1,294 institutions). Upon

further inspection of these five institutions, the average price is nearly twice as high as the average price of private non-profit institutions and public institutions ($32049 compared to $18575). Not only are the prices astronmically higher, but the debt is also higher, with average earnings only slightly higher (the SAT scores are almost 100 points lower, on average). This would be important to remember as we develop our models and make predictions/assumptions for this category, as we do not have sufficient data. Otherwise, the data is reasonably balanced between private non-profit institutions and public institutions. From now on, we will refer to 10 year post-graduation earnings simply as earnings. Next, we explore univariate and multivariate EDA.

2.3 Transformations

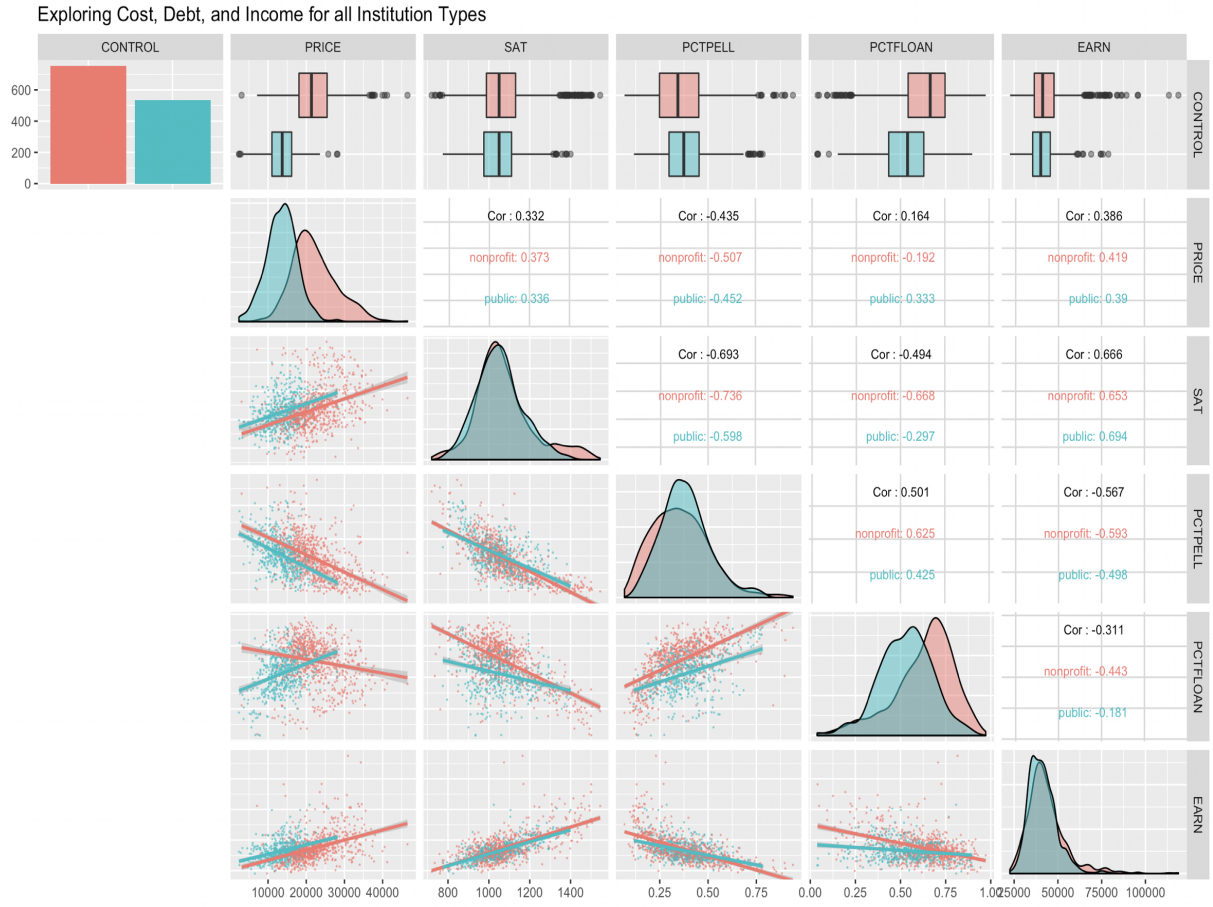

Figure 1: pairs plot describing a few variables of interest (univariate and multivariate checks), colored by CONTROL

(1) Based on Figure 1 and various other explorations considered (not shown), there was significant skew on many of the variables. As a result, a log-transformation was performed across all the variables in an attempt to rectify the skew. **(2)** Our response is EARN, which is heavily skewed to the right (pre-log transform). Note that although the distribution of a few certain covariates became worse due to the log-transformation, the integrity of the model remains valid. **(3)** From Figure 1, there seems to be a fairly strong and positive linear relationship between the price of tuition and earnings, as well as SAT score and earnings. There also seems to be a fairly strong

but negative linear relationship between the percentage of students receiving Pell grants and their earnings. It is interesting to note that there is a noticeable discrepancy in the relationship for federal loans and earnings between private non-profit institutions and public institutions. This may indicate a desire to include an interaction term later on (as a potential omitted variable). A natural belief would be to control for prior education and prior economic status in performing the following analysis as there seems to be a strong negative linear relationship between SAT scores and the percentage of students receiving Pell grants at each institution. It is further interesting to note that private institutions have an average of 63% of students requiring a federal loan compared to an average of 52% of students for public institutions. **(4)** A quick look at the highest earning institutions, clearly show that private non-profit institutions such as Ivy League schools (expensive and prestigious - driven by socioeconomic elitism), STEM focused schools (CMU, CalTech, MIT, Rose-Hulman, Stevens, RPI, etc.), or business heavy schools (Bentley, Babson, Duke, etc.) pay the most post-graduation (figure shown below).

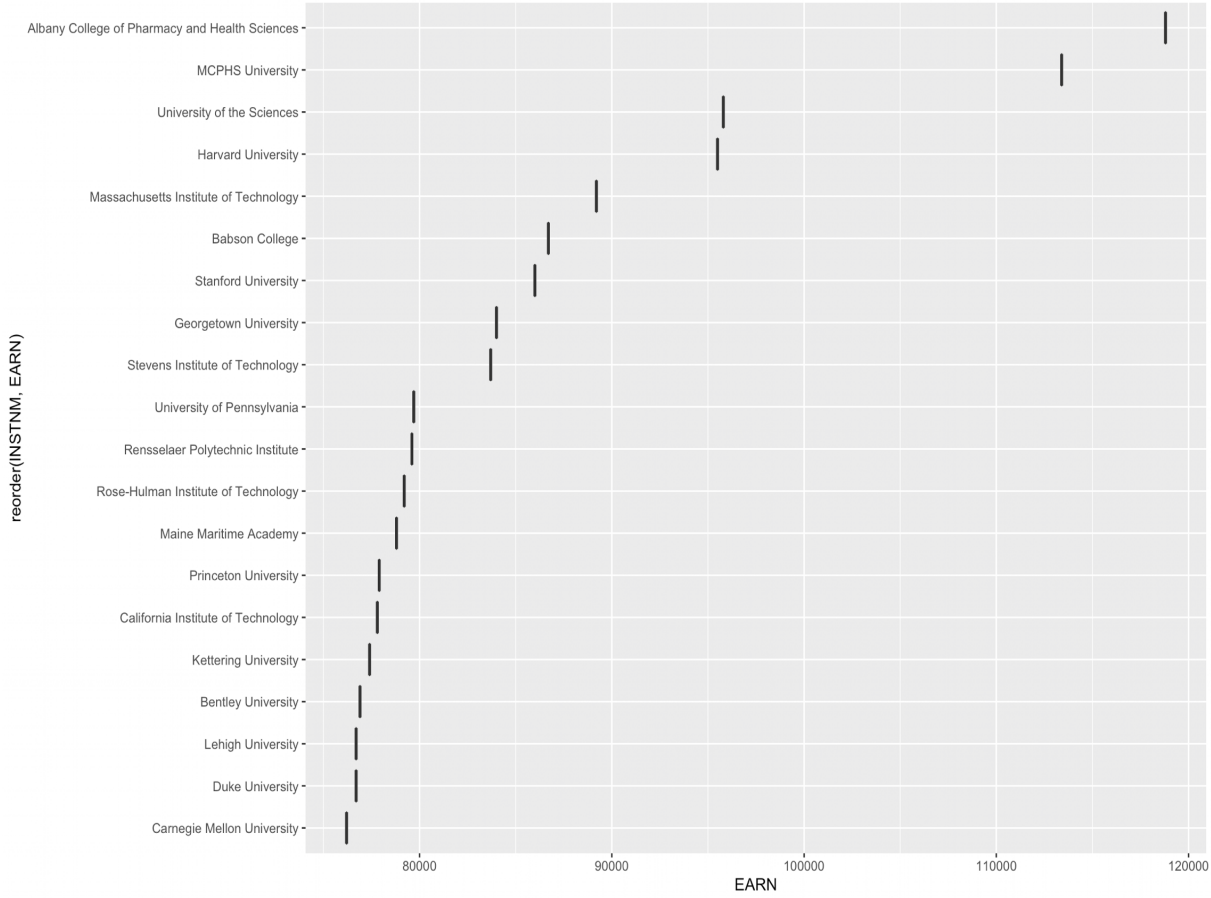

Figure 2: top 20 median earnings

3 Modeling and Diagnostics

As there are many regressions to choose from, we aim to select the most accurate and interpretable model that explains the underlying patterns in our data. Following Occam's Razor, we prefer simpler models over complex ones. **(1)** We begin by first examining linear regression models, and start by including all the covariates. We then did an exhaustive search using subsets regression to

select the predictors in an aim to balance the number of predictors and the various other goodness-of-fit metrics such as adjusted r-squared, Mallow's Cp, and Akaike Information Criterion (AIC). Based on the lowest Mallow Cp statistic and a competitive adjusted r-squared, we decided to use the following regression:

$$\text{EARN} = \beta_0 + \beta_1 \text{UGDS} + \beta_2 \text{PRICE} + \beta_3 \text{SAT} + \beta_4 \text{PCTPELL} + \beta_5 \text{PCTFLOAN}$$

There was no severe multicollinearity present after examining the variance inflation factor (square root below 2). Interaction terms were also explored, such as that between SAT and PCTPELL (assuming that higher SAT scores also drove PCTPELL down due to socioeconomic advantages), but was not impactful to the adjusted r-squared of the model. The residuals of this model gave evidence to homoscedacticity, or that the variance is approximately the same for all predictor values, and the errors seem to have mean 0. **(2)** It is also interesting to note that not only did the presence of PCTFLOAN not increase the adj. r-squared value (went down from 52.16% without PCTFLOAN to 52.12%), it actually increased the complexity and Mallow's Cp statistic; however, we decided to retain it as we believe that it affects the opportunities that students have (controlling for the covariates), and naturally their earnings post-graduation. This was also supported by the fact that private and public institutions had a drastic difference in this covariate with respect to earnings, as seen in Figure 1. We controlle for covariates by simply including them in our model.

After establishing a good linear regression model, we'd like to examine the benefits of greater flexibility and power that come from fitting with GAMs. We used a smoothing spline with four effective degrees of freedom over each of the variables of interest: UGDS, PRICE, SAT, PCTPELL, and PCTFLOAN. After analyzing the plot (shown below), it was believed that SAT and PCTPELL could be linear.

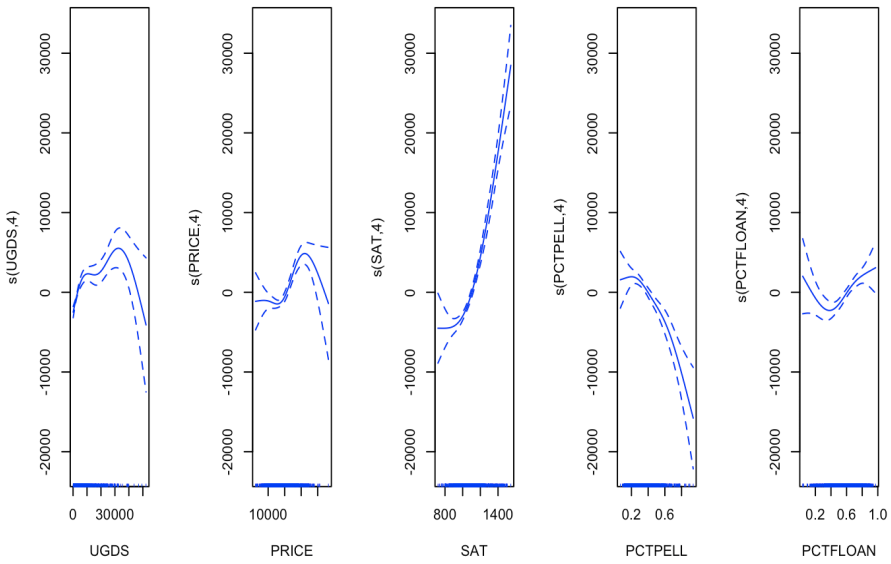

Figure 3: plot of all the smoothing splines used in the baseline GAM

After performing an iterative process of fitting with and without those variables, we utilized an ANOVA F-Test to determine the optimal GAM. The best GAM model that we found was one in which SAT and PCTPELL were both linear, while retaining the rest of the smoothed variables (all

variables were significant).

Before selecting which model to use, we examine potential violations that may have occurred when building the model (see below).

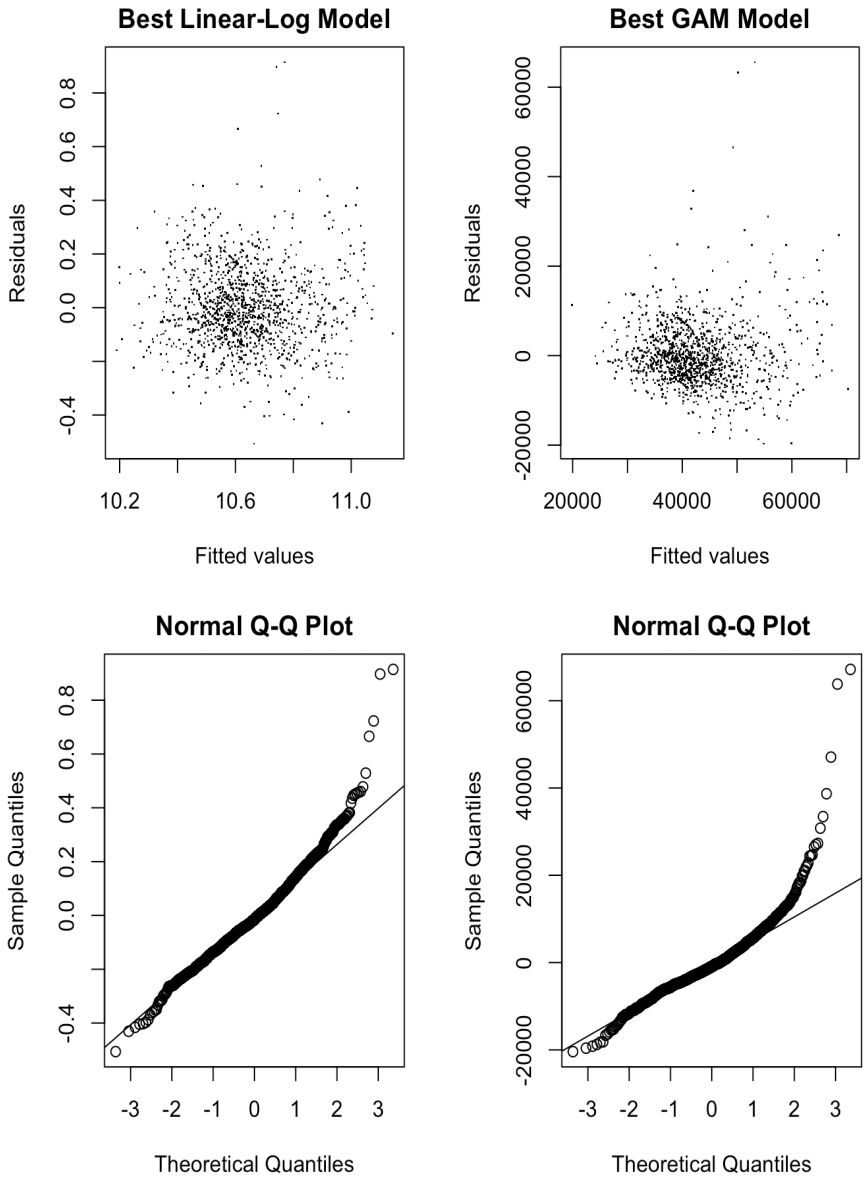

Figure 4: residual and normal q-q plots linear-log and GAM, respectively

(3) Clearly, the residuals obtained by the linear-log regression model seem to be the better of the two, as it shows no clear patterns with constant scattering, a mean of 0, and approximately normal errors. On ther other hand, the optimal GAM fit has a noticeable clustering and fanning shape (left to right, heteroscedasticity). Thus, we would not be inclined to believe that the residuals were evenly distributed. Moreover, the q-q plot of the GAM model has a noticeable skew, indicating an abscence of normality. This seems to be less of an issue for the linear-log regression model, where the points on the q-q plot lie closer to the normal line. Possible improvements to both of these models would be to increase their respective complexities; however, an increase in the dimensional-

ity of our models (through introducing complex terms, perhaps the product of multiple, etc) would come at the cost of reduced interpretability of our results. As a consequence, we will stick with these models that do an acceptable job at explaining the data.

In order to ascertain the validity of the models, we'd like to estimate the predictive power and standard error of each in order to select the best one. As we have already presented the model diagnostics (both seem reasonable), we now present the results of a 10-fold cross-validation test as can be seen Table 1.

Linear-Log Reg.	GAM
53267653	49835722
6498466	5403114

Table 1: MSE (first row) and SE (second row) from 10-fold CV

(4) Table 1 includes both the 10-fold cross-validation test MSE (the first row) and a rough estimate of the standard error (second row). Our GAM model not only performs better in terms of prediction error, but also has a smaller standard error. We also know that there are more degrees of freedom in our GAM model (15 compared to 6 in the linear regression model), which allows for a greater flexibility when fitting. Although the residuals are not uniformly distributed and lacking in normality, we select the GAM as our chosen model as it gives a pretty significant increase in predictive power with a much smaller standard error. It is important to note that the difference between the two models seems to be relatively significant (upon first glance). **(5)** There is an approximately 6% decrease in prediction error, and nearly a 20% decrease in standard erro when we choose the GAM model. **(6)** Looking at the diagnostic plots, it seems that bootstrapping by resampling cases is the most appropriate choice to proceed with, as the plots lack homoscedasticity and does not give strong signs of normality.

4 Results

(1) As requested by the United States Department of Education, we found that the price of attending school and the amount of money earned after graduation does not necessarily have a linear relationship. That is, an increase in the price of a school is not linearly related to the amount of money earned after graduation. This analysis was drawn by the smoothing spline fitted by our GAM, as seen in Figure 3. Price seems to dip up and down, with an almost sinusoidal pattern. However, this analysis should be taken with caution - there may be many confounding variables and variables that were not included in this study that would probably have affected this relationship (such as location, major, race/ethnicity etc).

Additionally, there exists a statistically significant difference among public, private, and for-profit universities when it comes to the relationship between price and earnings. The null hypothesis is that the relationship (as defined by the model parameters in our GAM, above) among public, private, and for-profit universities are the same (price vs earnings). The alternative hypthesis is that the relationship is different (at least one of the coefficients among our parameters in our GAM model is different than the rest). **(2)** After performing an ANOVA test on our GAM model, which includes interaction terms, it was determined that all smoothing and parametric terms were

significant (p-value < 0.05, for all terms), indicating that at least one term was different and that the relationship between price and earnings was not the same among private non-profit, private for-profit, and public universities. It is important to note that any conclusion drawn here would not be ideal, since there is a deficiency in the quantity of the data for private for-profit institutions. A separate analysis was conducted by removing these data points and disregarding this category of institutions (which make up of less than 0.38% of the total data), and private non-profit and public institutions still remained statsitically significant for its equivalent hypothesis test. This belief can be reinforced below by examining a noticeable difference in the residual plot for private non-profit and public institutions, as seen below.

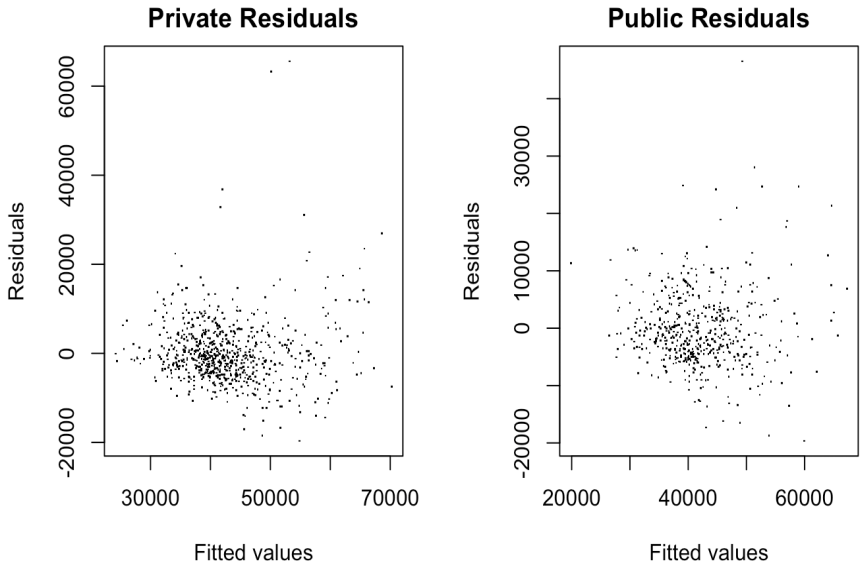

Figure 5: residuals plots for private non-profit institutions and public institutions

Finally, using two different methods, we found two different 95% confidence intervals that described the mean earnings of students after graduation from schools just like Carnegie Mellon University. Here, we define "schools just like Carnegie Mellon University" as institutions that have an undergraduate enrollment size between 4,000 and 10,000, with SAT scores above a 1400 (these were some of the most significant covariates from the above analysis). These institutions included M.I.T, Stanford, Harvard, etc. Using the standard error from the `predict` function, we obtained a 95% confidence interval of [61423.66, 65726.08]. **(3)** That is, one can expect that the true mean earnings of students 10 years after entry from schools like Carnegie Mellon University is between \$61,423.66 dollars and \$65,726.08. It is important to note that the assumptions made by the `predict` involves uncorrelated, normally distributed errors with mean 0 and equal variance (homoscedasticity) - which is not necessarily the case (as seen above in the diagnostic plots). To remedy this, we performed a parametric bootstrapping by resampling cases (with the number of bootstrap samples being 10,000). The resulting 95% pivotal confidence interval was [17656.97, 99158.5]. That is, one can expect that the mean earnings of students after graduation from schools like Carnegie Mellon University is between \$17,656.97 and \$99,158.50. The more reliable confidence interval is the one obtained by bootstrapping, as it assumes nothing of the original distribution (although the confidence interval may be too wide to be useful). More precisely, since the residual plots did not inspire confidence for the normality and homoscedacticity. **(4)** Thus, we believe that the confidence interval obtained by bootstrapping is more reliable - additionally, this interval contains the previous interval that made the assumptions above (which is unsurprising, as stronger assumptions made

by `predict` would imply a tighter bound on the interval).

5 Conclusion

Going to college is a financial risk that many are not willing to make. As a result, making informed decisions about where to go is of paramount importance. After this analysis, we hope that the US Department of Education can use our analysis to their advantage when implementing educational policies and financial aid. **(1)** Although we found price to be a significant factor when determining eventual earnings, the relationship was not necessarily linear. **(2)** This makes sense, as earnings is probably more driven by intelligence and work ethic (most aptly presented by SAT, which turned out to be strongly linear in our EDA and linear in our model) than the price of a university. It is also important to note that this data does not include those who did not need financial aid. Those who come from an advantageous background due to the socioeconomic standing of their parents are not only more likely to succeed, but also more likely to go to private schools. The relationship between financial advantages, attending private school, and earning more, are closely related. Although there was a significant difference in the types of institutions, we were severely limited in our analysis for private for-profit institutions as we simply did not have enough data. Additional limitations include variables that would play a drastic role, but were omitted (and thus imposing bias), such as major, ethnicity/race, and school ranking. The best indicator of monetary success seems to be SAT. It is interesting to note that debt did not play a significant role in determining the earnings - that is, debt was insignificant in the final model that we chose. This would lead one to believe that price would also perhaps be insignificant, or perhaps a less significant role, but it does play an important role (although not necessarily a linear one). Institutions such as Carnegie Mellon University, where a conservative measure (through bootstrapping) for the expected pay is between $17,656.97 and $99,158.50, continue to be common destinations for high achieving students that are willing to pay a high tuition for a quality education that furthers their career.