36-402 DA Exam 1

Madhuri Raman (madhurir)

4/3/2020

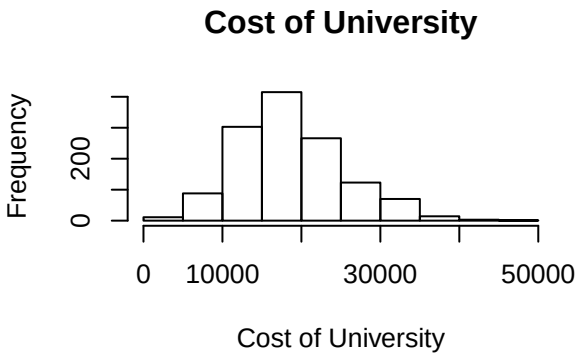
Introduction

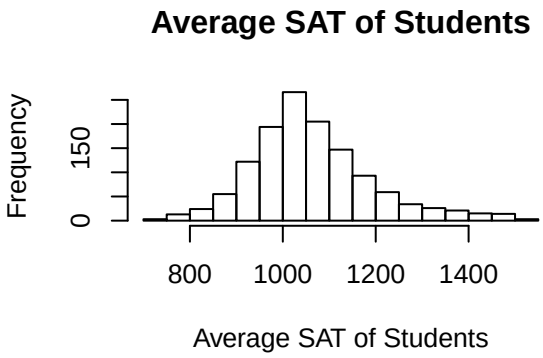
(1) The purpose of this study is to investigate the value (or lack thereof) of attending a more expensive university. We are specifically interested in the relationship between the cost of attending a university and the median earnings of that university's students after graduation, controlling for students' economic statuses and prior education before attending the university. We also want to investigate how this relationship between cost of university and post-graduation earnings may differ across different types of institutions, such as public, private, and for-profit institutions. Finally, since we are students at Carnegie Mellon, we want to use our model to predict the median earnings of CMU students post-graduation.

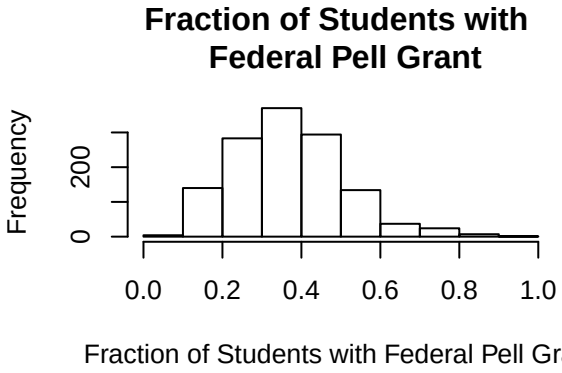
(2) In the end, we found that when we control for students' prior education, economic status, and they type of their institution, more expensive universities produce graduates with higher median earnings than their less expensive counterpart. Additionally, the type of institution one attends, whether it be private forprofit, public, or private nonprofit, does not change this relationship between price and future earnings. Finally, the expected median earnings for students at institutions like Carnegie Mellon University are between 63044.61 and 63071.49 dollars.

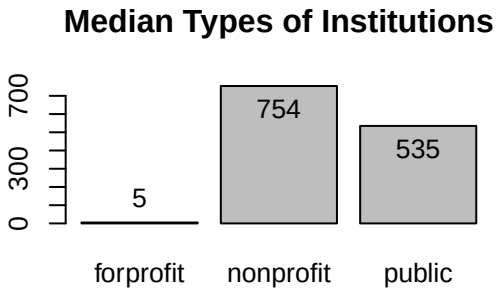
Exploratory Data Analysis

(1) The key variables we will be using to model median earnings of students 10 years after college are the average net price of the college, the mean SAT score of admitted students, the fraction of students at the college with a federal Pell grant, and the type of the institution (private non-profit, private for-profit, or public). We will examine the distributions of each of our continuous predictor variables, price of university, average SAT, and percent Pell grant students, with the following histograms and the distribution of our categorical variable, institution type, with a box plot.



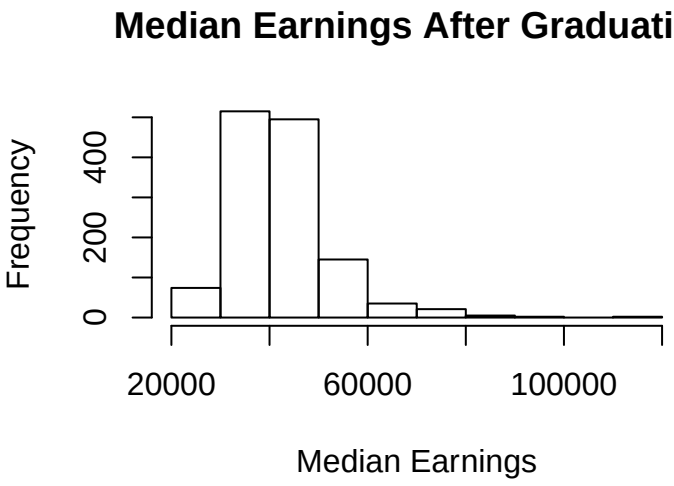






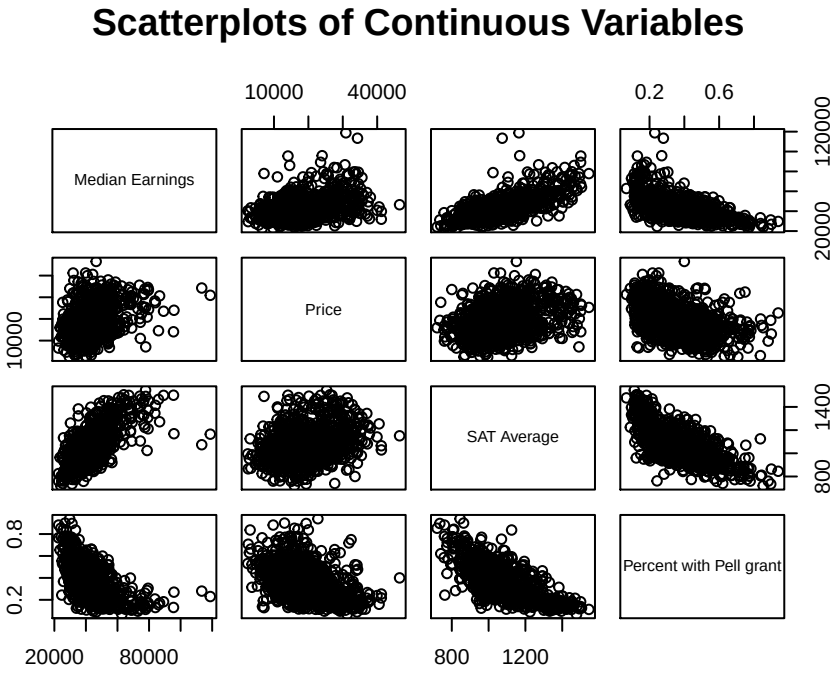
The variables all look fairly normally distributed with slight skews to the right. However, this degree of skewness should not seriously impact our model since our the assumptions that go into our model do not pertain to the distributions of the variables and only pertain to the distribution of the residuals.

(2) The response variable we will use in our model is the median earnings of students 10 years after graduation from the university. We examine its distribution in the histogram below.



The distribution of the response variable is approximately centered around 40000 but does look rather skewed to the right.

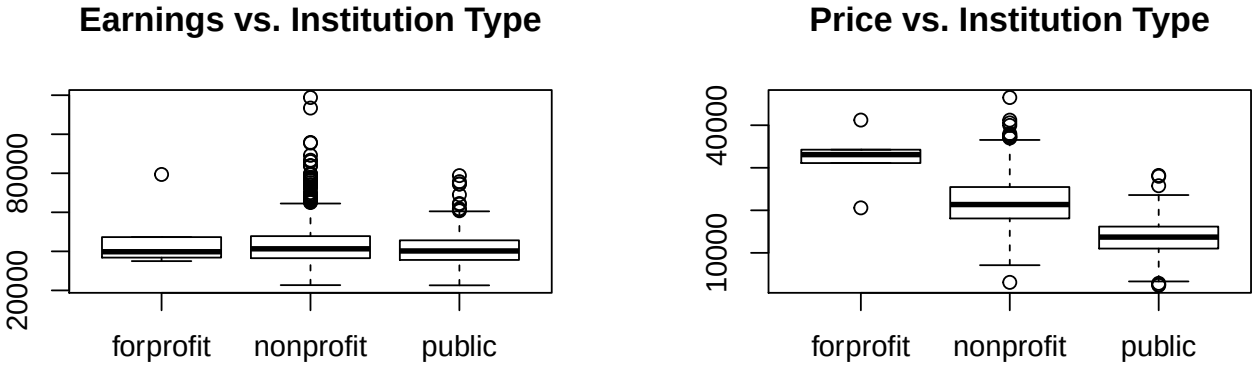
(3) We will now examine the pairwise relationships of each predictor with our response variable. For the three continuous predictors, we will construct a pairs plot to display the scatterplots of the predictors against each other and against the response variable.



(4) Based on the scatterplots, we can see a somewhat linear relationship between median earnings and all three continuous predictors. Specifically, there is a positive association between the response and price and the response and SAT average, while there is a negative association between the response and percent of students with a federal Pell grant. As price of university increases and as student SAT average increases, median earnings of students appear to increase as well. As the percent of student with a Pell grant increases, median earnings generally decrease. Specifically, based on the positive direction of the scatterplot of median earnings vs. price, we suspect that students who attend a more expensive school may earn more money after graduation than those who attend a less expensive school.

However, the associations we see are not entirely linear. For example, in the scatterplots of percent with Pell grant, the data trend appear to have a bit of an "elbow" or curve in the data. Specifically, the lower end of the distribution of Pell grant percentage have a steeper slope. Similarly, we see this behavior of a steeper slope at the right end of the distribution of SAT average in the plots of Earnings vs. SAT average and Price vs. SAT average. Additionally, in the scatterplot for median earnings vs. price, there appears to be a slight decrease in the slope, almost like a downward trend. These areas indicate potential nonlinearities in the feature space. Thus, we may consider creating an additive model with nonlinear transformations for the features price, SAT average, and percent with Pell grant.

Finally, to examine the relationship between the categorical predictor, institution type, and the continuous response, median earnings, we can look at the following box plot of median earnings for each type of institution to compare the distributions of each individual type of institution.



We see that the distributions of median earnings is not the same for the three institution types. The range of the median earnings for private nonprofit schools is larger than for public or private forprofit schools. Private nonprofit schools have a more right skewed distribution and slightly higher 50% and 75% quantiles of median earnings. However, private forprofit schools have a much higher minimum median earnings than public and private nonprofit schools.

In the boxplot of university price for each type of institution, private forprofit schools have the highest median cost of attendance, private nonprofit schools the next highest median price, and public schools the lowest median price. These distributions are a bit less skewed than those of median earnings vs. institution type. Again, the ranges of the distributions vary; private nonprofit schools have the largest range while private forprofit have the smallest range. Additionally, it is interesting to note that the maximum cost of attendance of a public institution is less than the 25% quantile of the distribution of private forprofit school costs.

Based on these two boxplots, we suspect that the relationship between price and earnings will not be the same at public, private nonprofit, and private forprofit institutions. Specifically, we saw in our EDA that all three types of schools had similar median and quantile values of earnings, while in the distributions of price, private forprofit schools have a clearly higher price than private nonprofit schools, which have a higher price than public schools. This suggests that the relationship between price and earnings may be different at different types of schools.

Modeling & Diagnostics

(1) We will construct a linear model and an additive model to answer the research questions. As we mentioned previously, we identified price, SAT average, and percent with Pell grant as the variables for which we may need a nonlinear function, based on subtle trends we could observe from their scatterplots. We will fit smoothing splines with 4 degrees of freedom to these predictors in our additive model. Note that since institution type is a categorical variable with three levels, we include two "dummy variables" or indicator variables in our model to capture the institution type variable.

Linear Model: Earnings = β_0 + β_1(Price) + β_2(SAT Average) + β_3(Percent Pell) + β_4(Institution Type Nonprofit Indicator) + β_5(Institution Type Public Indicator) where $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are linear terms.

Additive Model: Earnings = β_0 + r_1(Price) + r_2(SAT Average) + r_3(Percent Pell) + r_4(Institution Type Nonprofit Indicator) + r_5(Institution Type Public Indicator) where r_1, r_2, r_3 are smoothing splines with 4 degrees of freedom and r_4, r_5 are linear terms.

(2) Our main interest with respect to the research questions is the relationship between earnings and price. However, we included 3 other variables in the model, SAT average, percent with Pell grants, and institution type, because we want to measure the relationship between earnings and price when these variables are controlled for, or kept fixed. This means that we can be sure that a change in price by an amount x will lead to a change in predicted earnings by an amount y with no influence by the other three variables in our model, SAT average, percent with Pell grant, and institution type. This is what we want to be able to report to the Department of Education when they ask if the cost of education influences median earnings for the purpose of funding federal loans. We don't want to report an answer that may actually be influenced by confounding variables such as SAT scores and prior economic status, so we need to include such variables in our model to control for them.

(3) We will now examine model diagnostics to evaluate the assumptions of our models and possible improvements and modifications we can make to them.

Based on the plots of the partial response functions of the additive model, the fit of smoothing splines to the continuous predictors in our model seems appropriate because each partial response function looks quite nonlinear.

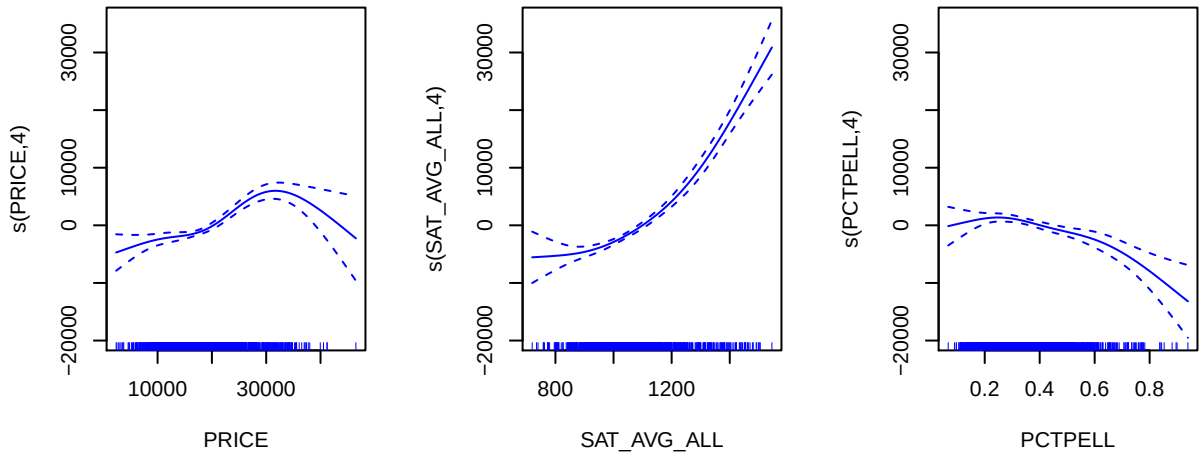
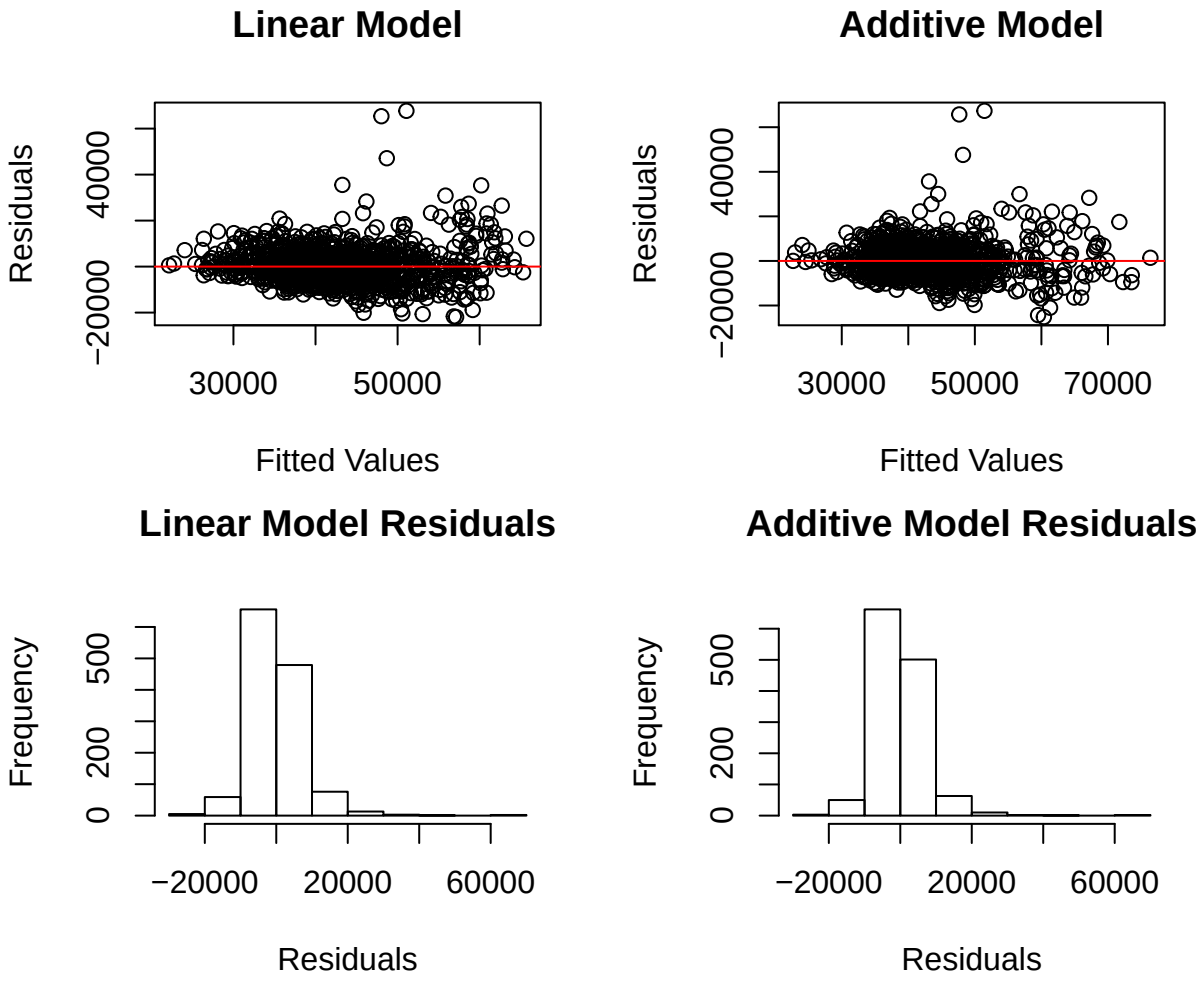
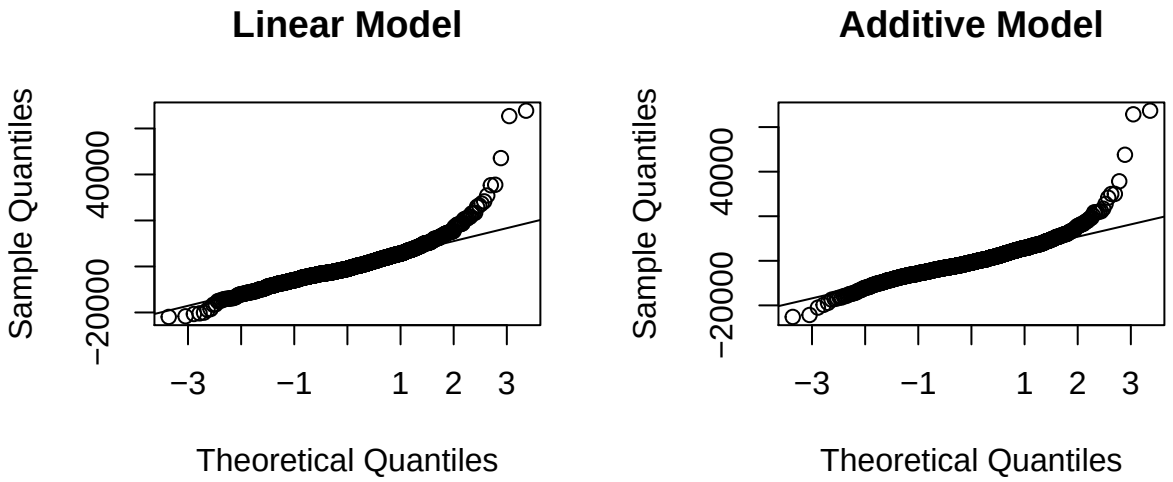


Figure 1: Additive Model Partial Response Functions





The distributions of the residuals in the two models look extremely similar. Both distributions are right skewed, but still centered around 0 as the residuals all hover the red horizontal line. The right skewness is also apparent on the normal Q-Q plots of the residuals for both model by the points at the right tail of the distribution lying far from the normal Q-Q line. However, in general, the residuals for both models lie well on the normal Q-Q line. The variances of the residuals in both models look fairly constant, with a very slight increase in the residuals of both models for higher fitted values approximately above 60000. Both models have several cases in the middle of their fitted-value ranges, around 50000, that have unusually high residuals. These points correspond to universities with higher median earnings of students than one would expect. However, in general, we can conclude that the assumptions of our multiple linear regression model and our additive model are satisfied in order to proceed with analysis and inference.

(4) Now we will perform 5-fold cross-validation to determine whether the linear or additive model fits best to the data in terms of prediction error.

The resulting estimate of prediction error (mean squared error) for the linear model was 54816181 and for the additive model was 52499629. Based on the metric of minimum prediction error, the additive model has the best fit to the data, so this is the model we will proceed with for inference.

(5) From the 5-fold cross-validation, the estimated (naively-calculated) standard error for the linear model was 7107712 and for the additive model was 7543164. The difference in cross-validation prediction errors between the two models is 2316552. This difference is much smaller than one standard error estimate and we also note that both models performed with similar prediction errors and residual distributions. This compels us to proceed with a more simple model like the linear model over the more complex additive model.

(6) Overall, based on the residual diagnostic plots previously created, such as the histograms of the

Table 1: Linear Regression Model Output

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1169.5671	4578.0648	0.2555	0.7984
PRICE	0.3176	0.0454	6.9949	0.0000
SAT_AVG_ALL	40.3733	2.1549	18.7360	0.0000
PCTPELL	-8982.3056	2170.6585	-4.1381	0.0000
CONTROLnonprofit	-4877.3672	3336.0487	-1.4620	0.1440
CONTROLpublic	-3299.4419	3411.4659	-0.9672	0.3336

residuals and the plots of residuals vs. fitted values, we believe that the variance of the residuals in the linear model is close to constant. However, from our previous detailed analysis of the linear model residual plots in section (3), we note that there are data points with very high positive residuals as well as some very high negative. Thus, we do not 100% trust the linear model for its mean response and distributions of noise, so we will proceed with a resample residuals bootstrap for this data.

Results

(1) Using the linear model, we will address the research question and determine whether students who attend more expensive school earn more money after graduation. To do this, we examine the regression coefficient for price in our model from Table 1. The coefficient for price in the regression is 0.3176 with a standard error estimate of 0.0454.

Thus, controlling for institution type, SAT average, and percent with Pell grants, we see that with every 1 unit increase in price of university there is a 0.3176 increase in median earnings 10 year after graduation. After conducting a t-test for this coefficient, we could reject the null hypothesis that the coefficient is equal to zero with a p-value of 4.25e-12 , and conclude that there is a significant positive association between university price and median earnings after graduation controlling for the other variables we included in the model.

(2) Now we will use our linear model to determine whether the relationship between price and median earnings is the same at public, private, and for-profit schools. In other words, we want to know if there is a statistically significant interaction between price and institution type.

Our new model including the interaction between the variables (for each of the two indicators for institution type) is:

Earnings = β_0 + β_1(Price) + β_2(SAT Average) + β_3(Percent Pell) + β_4(Institution Type Nonprofit

Table 2: ANOVA F Test Output

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1288	69576333459	NA	NA	NA	NA
1286	69526270511	2	50062947	0.463	0.6295

Indicator) + β_5(Institution Type Public Indicator) + β_6(Price)(Institution Type Nonprofit Indicator) + β_7(Price)(Institution Type Public Indicator) where $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$ are linear terms.

We will conduct a hypothesis test for the significance of the indicator terms in order to determine if there is a statistically significant interaction between price and institution type. The following are our null and alternative hypotheses for the test: $H_0 : \beta_6 = \beta_7 = 0$; H_A : at least one of β_6, β_7 are nonzero. The assumptions we made in conducting this test are that a linear model is appropriate for the data, the noise is independent and identically distributed from a Gaussian with constant variance, and that the noise is independent of X.

The test statistic for our ANOVA F test is $\frac{SS_{reg}/df_{reg}}{SS_{res}/df_{res}} = 0.463$ with a p-value of 0.6295. Thus, assuming our model assumptions are correct, we conclude that there is not a significant decrease in the MSE coming from adding the interaction between price and institution type to the model that did not arise from noise. The p-value is much larger than $\alpha = 0.05$, so there is not enough evidence to reject the null hypothesis of a linear model with no interaction term. This implies that relationship between price and earnings is not significantly different at public, private, and for-profit universities.

(3) We will use the original linear model (with no interaction terms) in order to build a 95% confidence interval for the mean earnings of students after graduation for a school just like Carnegie Mellon. Again, the assumptions that this confidence interval method makes are the usual multiple linear regression assumptions which are that the true model is linear, the errors are independent and identically distributed from a normal distribution with mean 0 and constant variance, and that the errors are independent of the predictors.

The 95% confidence interval for the mean earnings of students 10 years after graduation from a school like Carnegie Mellon is (61751.11, 64391.58). This means that we are 95% confident that the true mean earnings for students from a school like Carnegie Mellon is between 61751.11 and 64391.58 dollars.

(4) Lastly, we will compare this confidence interval to one calculated by bootstrapping by resampling residuals. With this method, we will assume that the distribution of the residuals is similar for each x and for each model with resampled noise in our bootstrap.

The bootstrapping method resulted in a normal confidence interval of (63044.61, 63071.49), which

means that we are 95% confident that the true mean earnings for students from a school like Carnegie Mellon is between 63044.61 and 63071.49 dollars. This is a much narrower interval than our previously made confidence interval from the original linear model. The bootstrap method assumes that our residuals are similar for each x but resamples them, while the original confidence interval method has more assumptions regarding their mean and distribution and results in a wider interval. Thus, I would consider the bootstrap confidence interval to be more reliable based on the assumptions involved and the resulting more precise 95% confidence interval for the mean response.

Conclusions

(1) Based on our analysis, we found that students who attend more expensive schools earn more money after graduation when controlling for the average SAT of students at their school (i.e. prior education), the percent of students at their school with a Pell grant (i.e. economic status) and the type of institution they attend (public, private for-profit, or private non-profit). We also hypothesized that this relationship between price of university and earnings would be different for different types of institutions, but in fact we did find enough evidence to prove this true. Institution type does not directly interact with or affect the relationship between price and median earnings. Finally, we found that the expected median earnings for students at institutions like Carnegie Mellon University is between 63044.61 and 63071.49 dollars. The Department of Education, which funds federal student loans, wants to use this information to understand if attending an expensive institutions is worth it in the long run; that is, will someone who attended a more expensive institution ultimately make more money? is it worth funding their loans if they decide to attend an expensive institution like Carnegie Mellon? Controlling for all other prominent factors, indeed the answer is yes.

(2) Our results are sensible. It makes sense that attending a more expensive institution can lead to higher earnings later in life, because more expensive universities generally have more and better quality resources that will prepare you for your career path. This fact does not change if someone attends a public school vs. a private for-profit school vs. a private non-profit school, because in general, all universities convert money (tuition) into resources for their students. Finally, the median earnings of students at schools like Carnegie Mellon University is around 63058.05, which is reasonable considering our main discovery. Carnegie Mellon has above average cost of attendance, so we would expect their students' median earnings to be above average as well, which they are.