# 36-402 Data Analysis Exam 1

## Introduction

The Department of Education is interested in knowing whether colleges and universities with higher tuition are worth the cost compared to less expensive institutions. Given in-depth information and data from about 1,300 American colleges and universities, we will to answer the following three questions. The first being whether students who attend more expensive schools earn more money upon their graduation (1). We realize that there are many variables impacting the opportunities that students have, so we will attempt to control for the student's prior education and economic status. We also plan to answer whether the relationship between the cost of the institution and how much money individuals earn after graduation is the same for different types of schools, including public, nonprofit, and for-profit institutions. Lastly, we plan to take a deeper dive at certain types universities in particular. Using the model we construct, we will determine what the expected earnings for students at institutions like Carnegie Mellon.

Our analysis led us to the conclusion that more expensive schools do not necessarily lead to earning more money after graduation, but that there is a relationship between them. Additionally, we found that the relationship between the cost of an institution and earnings after graduation is different for public, nonprofit, and for-profit schools. Lastly, we predicted that the expected earnings for students at institutions like Carnegie Mellon is approximately between $66,000 and $73,000 (2).

## Exploratory Data Analysis

The United States Department of Education maintains a College Scorecard website to track American colleges and universities. We will use a portion of this dataset from 1,294 institutions to answer the questions set forth before us. We define 6 key variables. The first is our response variable, which we will call Earnings. It is the median earnings, in US Dollars (USD), of students working 10 years after their entry to the school. Our key explanatory variable we will call Price, which is then average net cost, in USD, of

attending the college, including tuition, fees, living expenses, etc. The other three explanatory variables we will use are SAT, PercentPell, and PercentFedLoan. SAT is the mean equivalent SAT score for admitted students, while PercentPell and PercentFedLoan are the percent of undergraduates who received a federal Pell grant and the fraction of undergraduates receiving a federal student loan, respectively. We will use SAT as a proxy for student's prior education, and PercentFedLoan and PercentPell as a way to account for economic status. Lastly, we define the categorical variable, Control, which is the type of school: public, nonprofit, or for-profit. The data set also included other variables such as student debt and total enrollment, but we did not identify them as necessary explanatory or confounding variables for our analysis.

We will start by looking at each variable individually. Price and Earnings are our two most essential variables in answering the Department of Education's question. In Figure 1 below, we see the histogram of each. Earnings, our response variable, appears to be heavily right skewed with the potential for outliers (2). A boxplot was made for Earnings (figure not shown), which did indicate outliers. This raises the idea that perhaps a transformation should be used. Price takes on a much more normal shape with perhaps a slight skew to the right. Taking a look at normality plots for each (figures not shown), we see that Earnings is obviously not normal, but that Price has some normality to it, though not perfect (1). We also analyzed the distribution of the three other explanatory variables (histograms and plots not shown). SAT appeared to be very normal in its histogram, perhaps with the right tail being a tad too long, but the normality plot indicated it was not normal. PercentPell's histogram had a decent normal shape with slight right skew, but the normality plot again confirmed a lack of normality. Lastly, PercentFedLoan had a slight left skew, the only of the continuous variables to have a left skew and was not normal. We also did a quick summary of the Control variable and found that 535 of the institutions were public, 754 were nonprofit, and only 5 were for-profit. The biggest takeaway regarding Control is that it will be difficult to make any conclusions on for-profit schools with data from only 5 schools.

Next, we explored how the variables were related to each other in order to get a better idea of how to fit models to the data. Below in Figure 2, we see a scatterplot matrix with

each of our five continuous variables. First comparing the predictors to our response variable, Earnings and Price seemed to have a slightly positive relationship, but nothing too significant, while Earnings and SAT had a more distinct positive, linear relationship. PercentPell and Earnings had a weak negative relationship, which was somewhat linear, and PercentFedLoan and Earnings did not seem have any relationship, maybe barely negative (3). The scatterplot matrix also allows us to analyze relationships between explanatory variables to see if there will be any multicollinearity issues that might undermine the validity of our models. Looking at Price, this variable does not seem to have a distinct relationship with any of the other variables, except perhaps a slightly negative relationship with PercentPell and a weak positive relationship with PercentFedLoan. SAT has no real relationship with Price, but noticeable negative relationships with PercentPell and PercentFedLoan. Wrapping things up, we see a positive relationship between PercentPell and PercentFedLoan, which makes sense intuitively as they attempt to measure the same thing. We found it interesting that prior education (SAT) and economic status (PercentPell and PercentFedLoan) seemed to have a negative relationship according to the scatterplots (4).

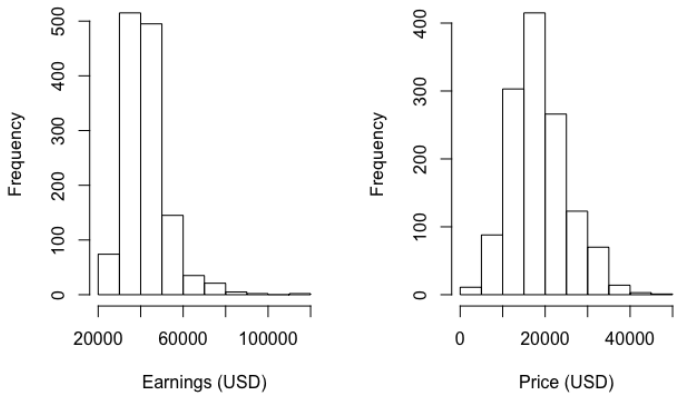


**Figure 1:** Histograms showing the distributions of variables Earnings and Price.

The last exploratory analysis that we conducted was take a look at the summaries for Earnings and Price based on each specific type of school. A few interesting takeaways were that all three types seemed to have very similar median earnings, while for-profit had the highest mean and nonprofit had largest third quartile and maximum. Public institutions actually had the smallest value of the three for each major measurement,

though the differences were minimal. It is important to remember there are only 5 for-profit schools to analyze so its measurements are not as significant with such a small sample size. Looking at the Price of each institution, we see that for-profit had the highest median and mean followed by nonprofit and then public. An interesting finding is that the differences in Price were much more than that of Earnings. Public schools had the lowest prices overall, followed by nonprofit and then for-profit with the highest cost. Overall, it is difficult to decipher if the relationship between Price and Earnings differs based on the type of institution from this exploratory analysis.

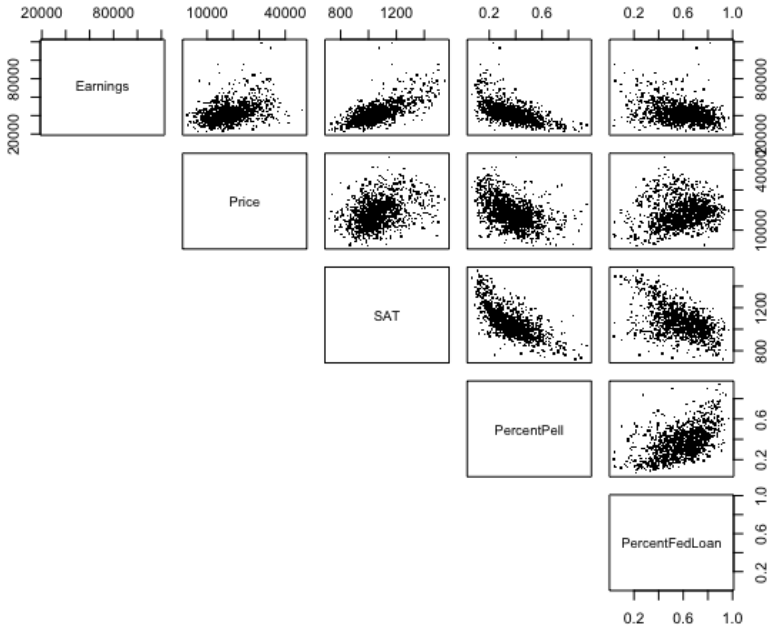


**Figure 2:** Scatterplot matrix showing relationships between relevant continuous variables.

## Modeling & Diagnostics

We constructed a linear model and an additive model in order to answer the questions the Department of Education asked of us, which can be seen below (1). It is important to note that s() indicates a nonlinear smoothing function with four degrees of freedom.

**Linear:** $Earnings \sim \beta_0 + \beta_1 * Price + \beta_2 * SAT + \beta_3 * PercentPell$

**Additive:** $Earnings \sim \beta_0 + s(Price) + s(SAT) + s(PercentPell) + s(PercentFedLoan)$

There was much analysis done to decide the variables to choose for each model. For the linear model, we initially started with all four continuous explanatory variables (Price, SAT, PercentPell, PercentFedLoan). This produced a model where all variables were significant ($p<0.001$) except PercentFedLoan ($p>0.4$). Transformations were performed on both explanatory and response variables, such as logarithmic and exponential, to see if it would better improve the model, but none produced significant results. We then used an F-test to compare a model including PercentFedLoan and one without it, and the results showed that not including PercentFedLoan made no statistical difference, so are final linear model used only Price, SAT, and PercentPell. For our additive model, we tried a variety of models with different combinations of nonlinear smoothers and linear functions on the explanatory variables. Performing a F-test on all these models, we ended up with an additive model where all four explanatory variables used a nonlinear smoothing function with 4 degrees of freedom, which can be seen above. All of the functions in this model were statistically significant ($p<0.02$).

We identified earlier that there are confounding factors, prior education and economic status, when measuring the relationship between Price and Earnings. To control for these confounders, we have included them into both of our models (2). SAT, the measure for prior education, is in both models, and at least one of either PercentPell and PercentFedLoan, the measures for economic status, are in both models. It crucial to account for confounding variables in our analysis because they have statistically significant relationships to Earnings, and not including them would undermine the validity of our analysis.

Looking next at our model diagnostics with Figure 3 below, we see a plot for each model of the fitted values versus Earnings as well as a plots of the fitted values versus the model's residuals. Starting with the fitted values versus the response variable, there is not much different between the models. In the plots, we have included a $y=x$ line, which is the perfect case where each fitted value is the exact same as the response variable. We see that both models follow the $y=x$ trend, but are obviously not perfect. The plots look extremely similar and it is difficult to evaluate which model has a better fit from this, but overall, it seems that both models fit the data well (3). The lower plots

within Figure 3 show us the model's residuals plotted against the fitted values. The linear model's residuals appear to be evenly scattered above and below the 0-line with constant variance, which satisfies the linear regression assumptions. There do seem to be some outliers at the top of the plot, which was semi-expected given that our response variable is right skewed.

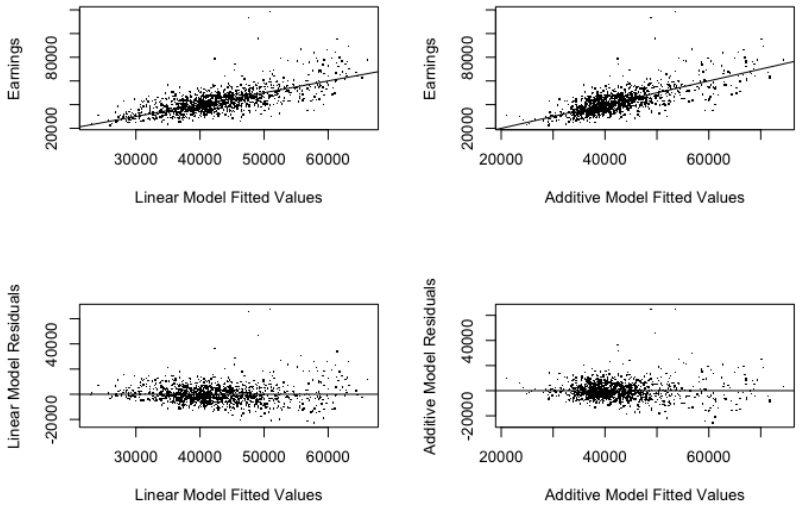


**Figure 3:** Diagnostic plots analyzing the relationships between Earnings and our model's fitted values as well as the model's fitted values versus their respective residuals.

The additive model's residuals versus fitted values plot shows us nearly identical results as the linear model with the residuals evenly spread around the 0-line with constant variance, but the potential for a few outliers. Looking at the normality of the residuals (figures not shown), we see again very similar plots. Both plots show a majority of the residuals are normally distributed, but we do see at the upper end of the plot that the points begin to tail off. The fact that the residuals are not completely normal does slightly undermine the validity of each models ability to perform inference, but it is normal enough where these can take place. Transforming the response variable to rid it of its skewness could have improved the residuals normality, but the ones we attempted were not successful. In the case of a bootstrap, nonparametrically resampling by cases would be the best choice because it only assume the residuals are independent (6). We do not know for certain the distribution of the residuals because they are not completely normal, and the outliers do raise questions about the constant variance.

We performed a 5-fold cross validation to estimate the prediction errors of each model in order to determine which model better fit the data given that we were not able to come to a conclusion by just looking at the diagnostics and goodness-of-fit. Below in Table 1 we see the results of our cross validation. The additive model had both a smaller estimated mean squared error and estimated standard error of the MSE (4). While at first glance the prediction errors do not appear that different, it is important to realize the units. The additive model provides an estimated MSE that is three million less and a standard error that is seven hundred thousand less. These are substantial numbers, and I believe the differences between the models is significant (5), though we did not perform a formal test to verify this.

**Table 1:** Estimated prediction error for each model and standard error based on 5-fold cross validation.

| | **Linear Model** | **Additive Model** |
|---|---|---|
| **Estimated MSE** | 54,648,188 | 51,656,395 |
| **Standard Error** | 2,788,311 | 2,064,143 |

Given the results of our cross-validation along with the fact that both models had very similar model diagnostics and nearly identical residuals, we decided to choose the additive model as the model to answer the Department of Education's questions.

## Results

Using our additive model to answer the Department of Education's questions, we performed a two-step process to see whether a more expensive school leads to earning more money. First, we create a reduced additive model, as defined below, which was the same as our chosen model, but eliminated Price as an explanatory variable.

**Reduced:** $Earnings \sim \beta_0 + s(SAT) + s(PercentPell) + s(PercentFedLoan)$

An F-test was performed between the models, and we found that including a nonlinear smoothing function with Price resulted in statistically significant improvement ($p<0.001$). Now with it established that Price and Earnings have a significant relationship while accounting for the confounding variables, we needed to take a look at what this relation looked like, which we see below in Figure 4. This plot showed us the impact that Price

has on average Earnings ($42,547.14). Below we see that the relationship is not at all linear and a mix of both negative and positive slopes. It appears that Earnings increases as Price nears $30,000, but falls sharply after that. Due to this up and down shape, we concluded that, on average, more expensive schools do not necessarily lead to students making more money, though it seems schools that charge around $30,000, which is relatively expensive, have the students that earn the most (1). It is important to note that the significance we established in the F-test meant that the relationship depicted in Figure 4 was significant, but not positive. If the plot would have been overtly positive, we could have changed our conclusion.

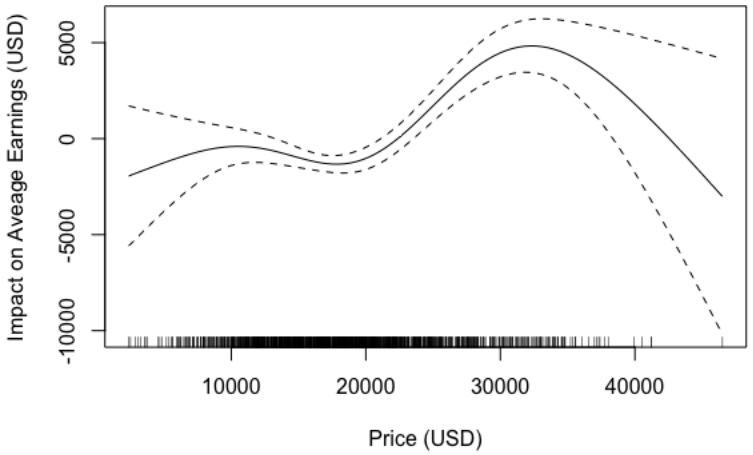


**Figure 4:** The nonlinear relationship between Price and Earnings according to the additive model while accounting for prior education and economic status.

Next we wanted to see if the relationship between Earnings and Price is the same based on Control (the type of school). We wanted to test the significance of this, so we created an expanded model, as seen below, to perform an F-test with our chosen additive model to see if adding interaction terms would be significant.

**Expanded:** $Earnings \sim \beta_0 + s(Price):for\text{-}profit + s(Price):nonprofit + s(Price):public + s(SAT) + s(PercentPell) + s(PercentFedLoan)$

We defined our null hypothesis that adding the interaction terms would have no effect on the fit, while our alternative hypothesis was that adding them would have significant impact. In more detail, our null hypothesis says that s(Price) in our additive model

accounts for the relationships of all types of institutions, and our alternative says each type institution's relationship with Price and Earnings is different. If we see a significant difference between models, we know that the interaction variables relationship improves the fit of the model and s(Price) does not account for them all. We performed an F-test between these two models with an $\alpha$=0.05 and defined our test statistic as a measure for testing the significance of the fit between the expanded and chosen additive model using each model's residual sum of squares and degrees of freedom, along with the size of our sample data. The results of the F-test indicated we could reject the null hypothesis and conclude that the expanded model was significantly better at predicting Earnings than our chosen additive model ($p<0.001$). Furthermore, we could answer the Department of Education's question and conclude that the relationship between Price and Earnings is different for public, nonprofit, and for profit institutions (2). It is important to note with this test, we assume the residuals of both models are approximately normal, which we established was mostly satisfied back in our diagnostic analysis.

The last question the Department of Education asked of us was to find the expected earnings for the average student at a school like Carnegie Mellon. To find this, we constructed two 95% confidence intervals, which can be found below in Table 2. Both intervals took the explanatory variable values of Carnegie Mellon University and used the expanded model to predict Earnings. We decided to use the expanded model for this rather than out original additive model because it takes into account Carnegie Mellon's Control. Both confidence intervals appear very similar with our pivotal bootstrap having a slightly wider range, about 1,000 wider on each limit.

**Table 2:** 95% Confidence Intervals for the expected earnings of students at Carnegie Mellon University.

| Predict Function Confidence Interval | Pivotal Bootstrap Confidence Interval |
|---|---|
| [$66,568.56, $71,310.53] | [$65,786.10, $72,554.44] |

Comparing the intervals more in depth, we know that our predict confidence interval assumes that the residuals of the model we use are normally distributed with constant variance (3), while our bootstrap confidence interval, which used 1000 bootstrap sample, makes no assumption on the residuals other than the fact that they are independent of each other. As we saw back in our residual analysis, the right skewness

of response variable leads to a few residual outliers and a mostly normal distribution, but not perfect. As a result, I believe our bootstrap confidence interval is more reliable because none of its assumptions are challenged because the residuals are independent (4). Furthermore, we are confident the expected mean earnings of students who attended schools like Carnegie Mellon is approximately between $66,000 and $73,000.

## Conclusions

In this study, we used an additive model to examine the relationship between the cost of an institution and the money earned by students who graduated from that institution. We concluded that while there was a significant relationship between the price and the money earned, which peaked around a cost of $30,000, more expensive schools did not always lead to more money earned (1). We accounted for confounding variables such as the student's prior education or economic status to get a true look at the relationship. We also found that the relationship between the cost of an institution and the money earned by students was different for public, nonprofit, and for-profit schools, which likely impacted our initial findings that the cost of an institution and the money earned from students who graduated from there did not have a totally positive relationship. Some expensive institutions are not designed to lead to careers that earn tremendous salaries, but rather very specialized careers, such as in the arts (2). Lastly, we found that students who attend schools such as Carnegie Mellon can expect to make between about $66,000 and $73,000 after graduation, which is substantially higher than the mean earnings of all the institutions in our study, which was about $43,000.

Looking at how we could improve our analysis, I believe incorporating debt of student's at specific institutions, such as the earnings the first five years upon graduation minus student debt, could give us an even better idea at the relationship between earnings and cost because many students graduate with substantial loans, though this data was not readily available in this study. Additionally, a larger sample size of for-profit institutions would help us better understand its relationship with price and earnings because we only had data from five of these schools. More specific measures for economic status, such as family income, could even better account for this confounding variable.