36-402 DA Exam 1

Richard Kang (rjkang)

4/3/2020

Introduction

(1)The Department of Education is interested in knowing whether more expensive institutions are worth the high tuitions.This project will attempt to answer three questions. On average, do students who attend more expensive schools earn more money after graduation? Does the type of institution affect the relationship between post-graduation salary and tuition? How much are students from institutions like Carnegie Mellon University expected to earn after graduation? (2)The result of this analysis showed that students who attended more expensive schools earned more money after graduation. Also, the type of institution affected the relationship in question. Lastly, students from institutions like CMU were expected to earn [65405, 71331] dollars after graduation, with 95% confidence.

Exploratory Data Analysis

(1)Key explanatory variables for this analysis are PRICE, GRAD_DEPT_MDN_SUPP, PCTFLOAN, PCTPELL, SAT_AVG_ALL, UGDS, CONTROL. GRAD_DEBT_MDN_SUPP and UGDS are heavily skewed, but other response variables show a fairly normal distribution. (2)The response variable of interest is MD_EARN_WNE_P10. MD_EARN_WNE_P10 is skewed right, as expected of any histogram of salary. (3)Conducting multivariate EDA shows that the response variable has a positive relationship with PRICE and SAT_AVG_ALL, a negative relationship with PCTFLOAN, PCTPELL, and a weak relationship with GRAD_DEBT_MDN_SUPP and UGDS. (4)This EDA hints at the possibility that the tuition of an institution and salary after graduation is positively related, and that schools with a higher overall SAT score has a stronger positive relationship with salarya after graduation. It is interesting to note that UGDS shows minimal interaction with any of the other variables. Also, SAT_AVG_ALL shows strong relationships with other variables, indicating that overall sat average has strong negative relationships with debt after graduation, percentage of students that receive a federal loan, and percentage

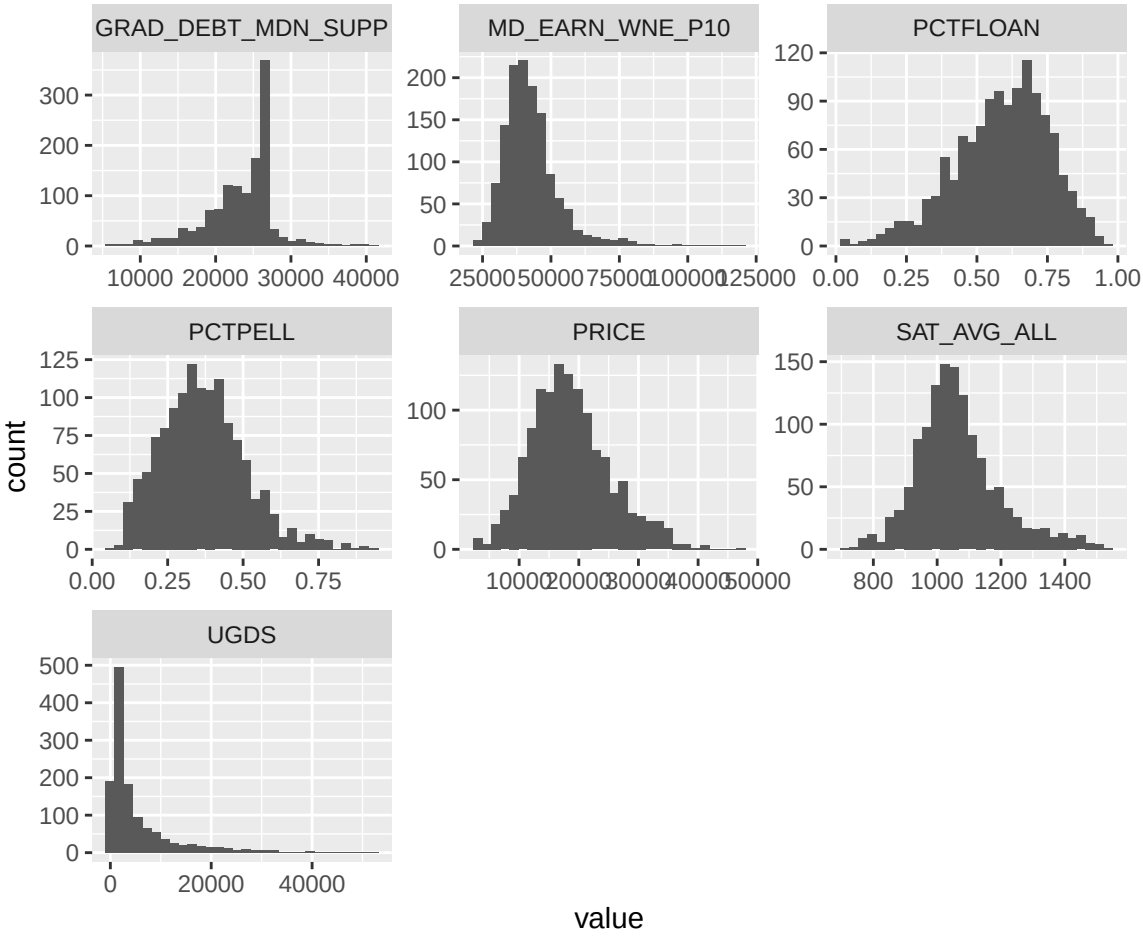


Figure 1: Univariate EDA

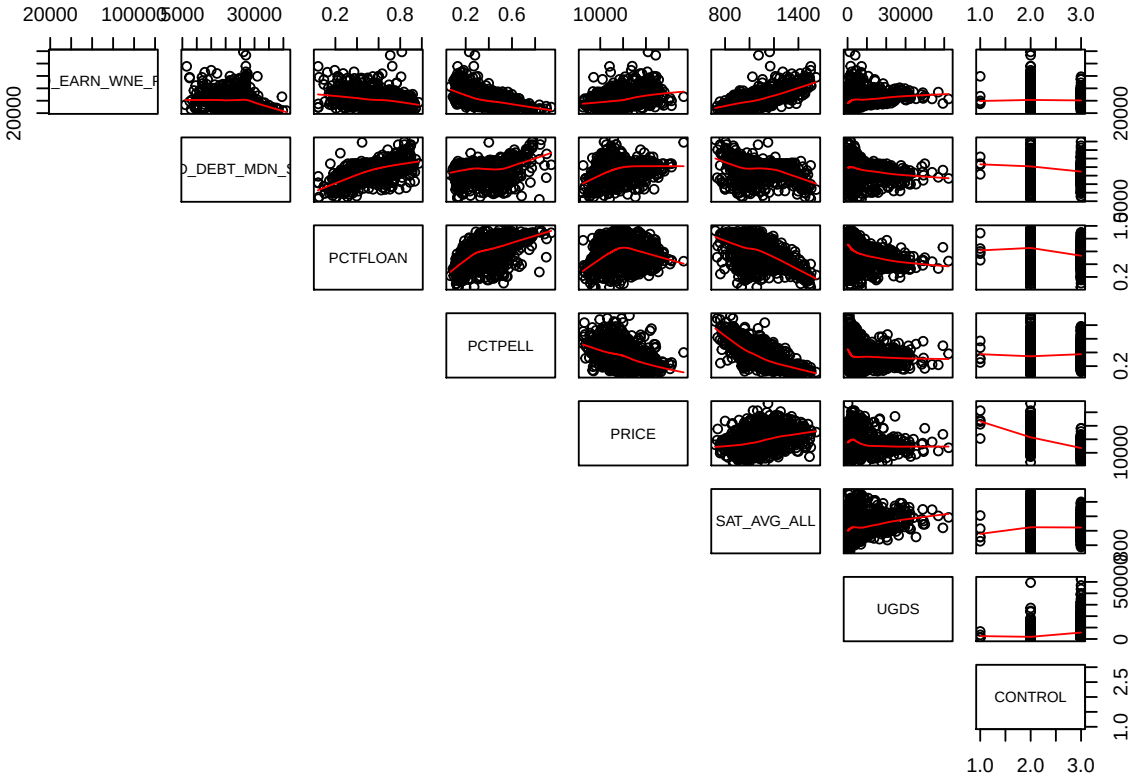


Figure 2: Multivariate EDA

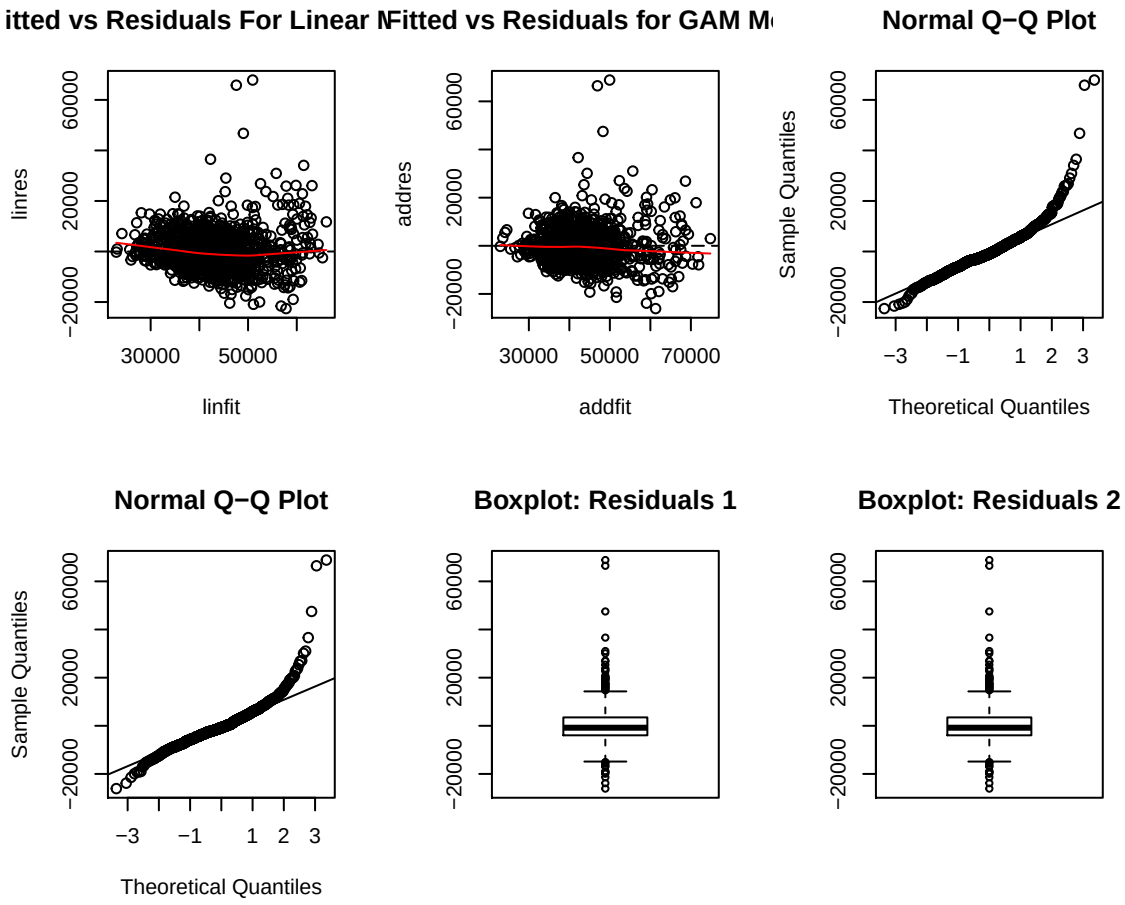


Figure 3: Residual Analysis

of students receiving a federal pell grant. Lastly, a positive relationship between price and sat_avg_all raises the possibility that the schools with higher SAT overalls tend to have higher tuition.

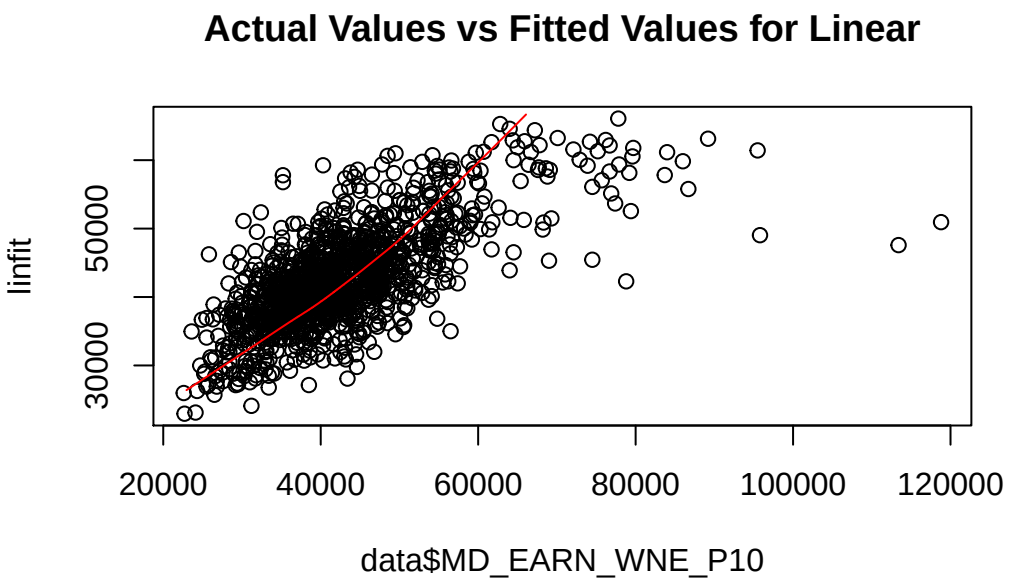
Modeling & Diagnostics

(1)The two models that were fit were:

$$LinearModel : Earnings = -1003 + .2404PRICE + 40.22SATAVGALL - 10160PCTPELL$$

$$AdditiveModel : Earnings = 38010 + 0.244PRICE + s(PCTPELL) + s(SATAVGALL)$$

We want to know the relationship between PRICE and Earnings, but also are concerned about the potential confounding variables, which is why PCTPELL and SATAVGALL were



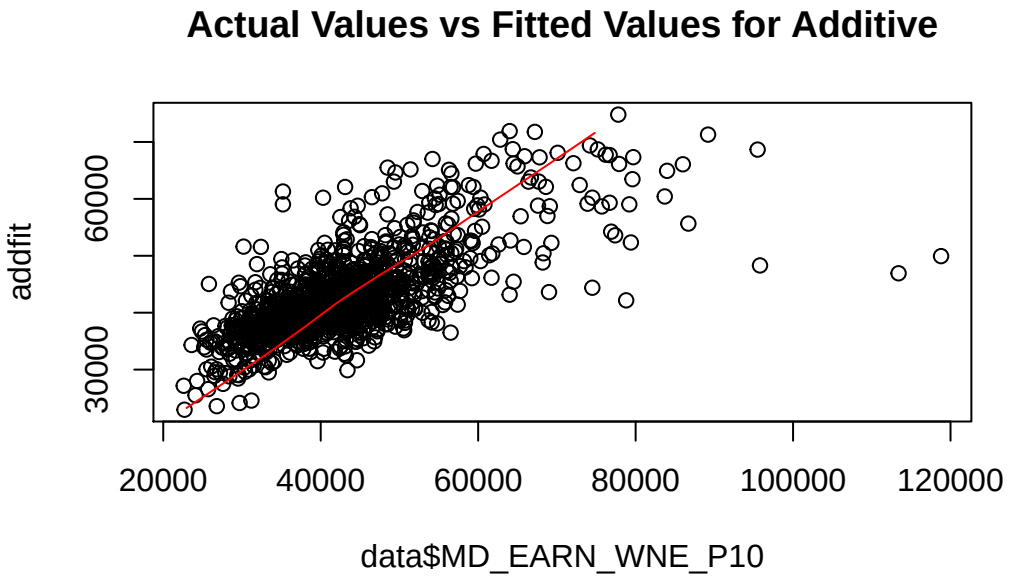


Figure 4: Residual Analysis 2

included. (2)Controlling for a variable simply means accounting for that variable in the model. Including PRICE, PCTPELL, SAT_AVG_ALL seemed sensible because these variables showed the strongest relationships in multivariate EDA. (3)Linear model's diagnostics were done through a fitted vs residuals plot, a normal qqplot, and a boxplot of the residuals. The fitted vs residuals plot and normal qqplot seem to satisfy the normality assumptions. However, the residual boxplot shows a handful of the datapoints lying outside the first and third quartiles. The gam model's diagnostics show similar results as the linear model's diagnostics. The number of points lying beyond the quantiles are not large, and they may not be too significant, and can be assumed to be outliers, for now. Lastly, a scatter plot of actual vs fitted values for the linear model seems to take on a slightly nonlinear curve, whereas the additive model's curve is more linear. (4)After running a 5-fold cross validation, the additive model returned lower prediction errors. (5)It is hard to tell whether there is a significant difference between the models, because the linear model had lower variance. The difference between prediction errors was approximately 2,300,000, and the difference in standard errors is approximately 200,000. The standard error for both models was approximately 5,300,000, much greater than the difference in prediction errors. This means that the difference in prediction error could be insignificant. These numbers are so large because the variable of interest is salary, which can go up to large numbers, and evaluating prediction errors involves squaring these large numbers. This is something to keep in mind. (6)The fitted vs residuals plot for the GAM model has a good patternless scatter around 0, which is good for the normality assumption. Also, the qqline seems to have the majority of points on the qqline. One drawback is the handful of points that lie outside the quartiles for the residual boxplot. Although we have strong support for the normality assumption for the additive model, nonparametric bootstrapping seems like a better choice, taking into account the distribution of the residual boxplot. It is better to assume nothing, than to incorrectly assume.

Results

(1)The additive model included SAT scores, Percent of PELL Grants and PRICE of the institution as the explanatory variables. The plot of the fitted values from the additive model vs PRICE shows a positive, rather strong relationship between PRICE and the predicted MD_EARN_WNE_P10 values. Also, the summary of the model tells us that a unit increase in PRICE leads to an increase in MD_EARN_WNE_P10 of 0.243, approximately, with a standard error of 0.034. For the majority of students, attending a higher cost institution

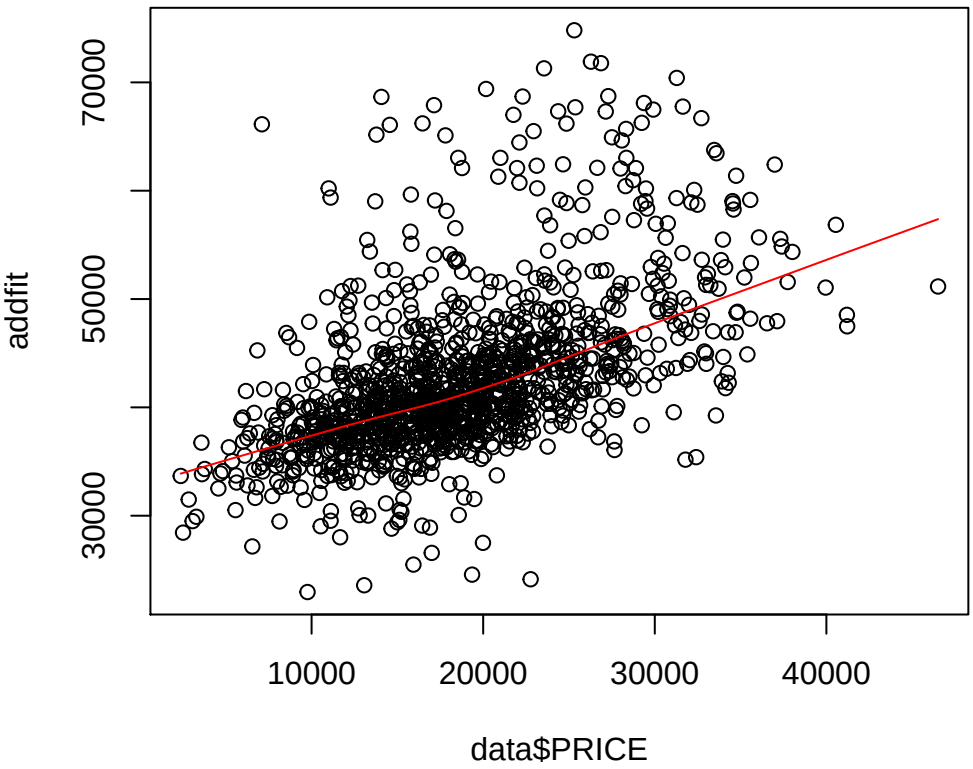


Figure 5: Fitted Values vs PRICE

will lead to a higher income salary. (2)To determine whether the type of the institution affected the relationship between price and earnings, a new model with an interaction term was fit and compared to the original additive model using ANOVA's f-test. Because it was an f-test, it assumed that the data are normal and independent of each other. Normality assumptions were shown previously, and the data are independent, so f-test's assumptions were satisfied. The null hypothesis for this test was that the two models had no significant difference. However, the pvalue of 0.0013 indicates that there is in fact a significant difference between the models, showing that the type of institution affects the relationship. (3)Students that go to institutions like CMU are expected to earn between [66200, 70179] dollars after 10 years with 95% confidence. This confidence interval used the student's t-distribution, assuming random sampling, normal distribution of the data and homogeneity of variance, all of which seemed satisfied. (4)Confidence Interval from Bootstrapping cases gave an interval of [65405, 71331]. This method of bootstrapping assumes the data are independent, and no other assumptions are made. Because no assumptions are made when bootstrapping cases, there is less risk using the bootstrapped confidence interval.

Conclusions

(1)From this data analysis, we can say that PRICE and MD_EARN_WNE_P10 have a positive relationship, when controlling for PCTPELL and SAT_AVG_ALL. After a better fitting model was chosen, an interaction term was introduced to the model to see if the type of institution had an effect on the relationship between PRICE and MD_EARN_WNE_P10. The type of institution in fact did turn out to have an impact, as shown by the anova's f-test. The conclusion of this analysis is that PRICE and MD_EARN_WNE_P10 have a positive relationship, and that the type of institution affects this relationship. (2)Some possible explanations for the affect of type of institution on PRICE and MD_EARN_WNE_P10's relationship is that private institutions tend to be more selective, and thus the majority of the students tend to move towards higher-wage jobs. On the other hand, public institutions have a larger pool of students, some of which may decide to take on a low paying job, and others that get a very high paying job. The variance in the choice of students within institutions affects the relationship of interest. Also, PRICE and MD_EARN_WNE_P10 may have a positive relationship because higher cost could mean better facilities, better student-to-faculty-ratio, and an overall better environment to learn.