36-402 DA Exam 1

Chae Yeon Ryu (chaeyeor)

4/3/2020

Introduction

The Department of Education is interested in knowing whether there's a relationship between students' earnings after graduation and their prior education, economic status, and the tuition. This study utilizes the College Scorecard data from 1,300 American colleges and universities. **(1)** The goal of this analysis is to quantify the statistical relationship with these variables and see if students who attend more expensive schools earn more money after graduation on average.

(2) It was found that students who attend more expensive schools tend to earn more money after graduation to some extent, and such a relationship was not the same at public, private, and for-profit institutions. This relationship was the most apparent for private schools and the least apparent for for-profit institutions. Using the model we built in our analysis, we were able to deduce that the expected mean earning for students at institutions like Carnegie Mellon is about $64,768.12.

Exploratory Data Analysis

The College Scorecard data consist of 1,300 American colleges and universities with measured scores and values such as thier average net price, mean SAT score, fraction of federal Pell grant, fraction of federal student loan, and median earnings. The dependent variable is Median earnings (dollars) of students working and not enrolled, 10 years after entry, and independent variables are Average net price, Mean equivalent SAT score for admitted students Fraction of all undergraduates who received a federal Pell grant for tuition, and Fraction of all undergraduates receiving a federal student loan. **(1)** Figure 1 shows histograms describing the distributions of each of these variables. We see that all distributions are right-skewed except with a left-skewed distribution of loan fraction.

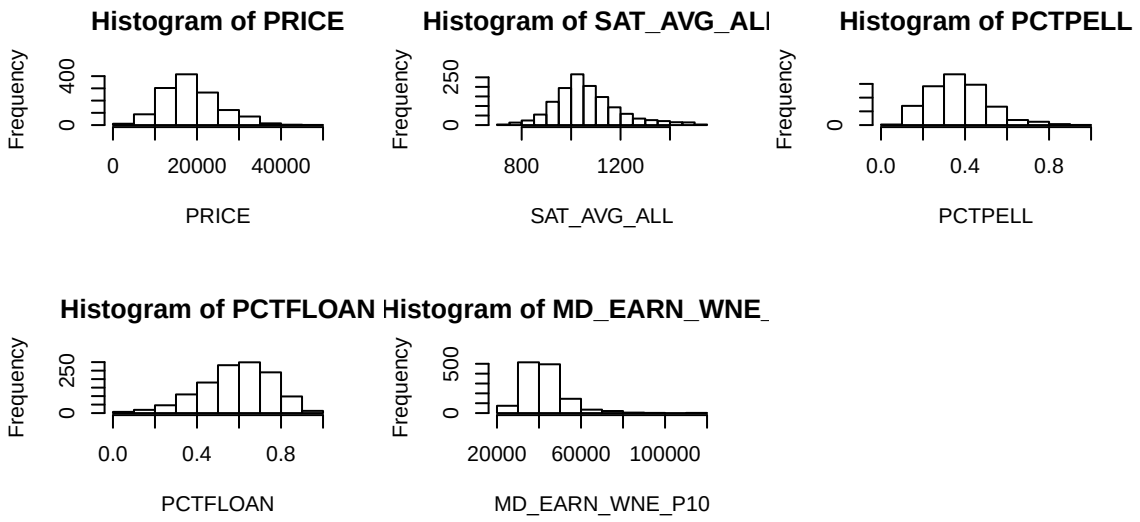

Figure 1: Histograms of Independent & Dependent variables.

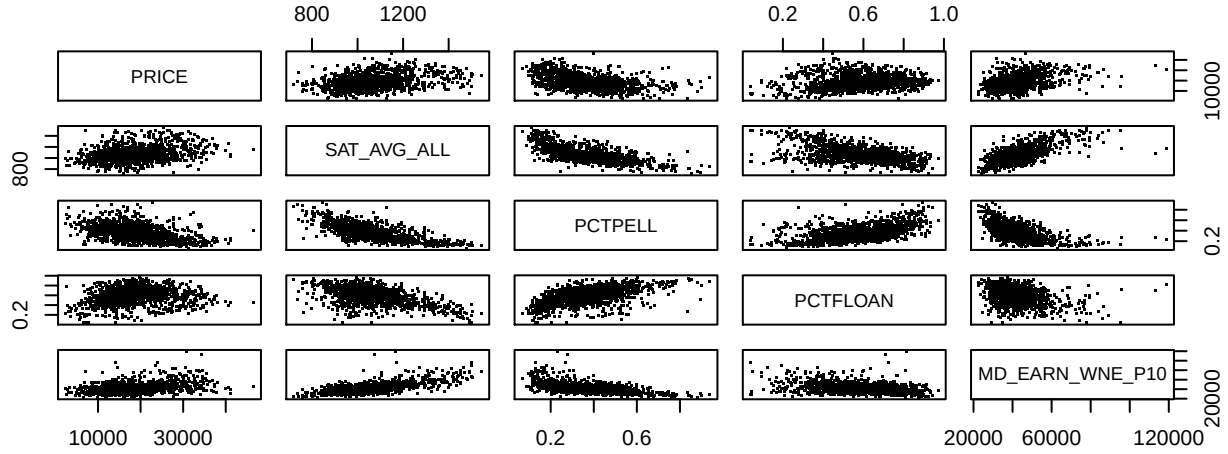

Figure 2: Bivariate relationships between variables.

They all have a unimodal shape, and independent variables display roughly symmetric shape while slightly skewed. **(2)** Median earnings, a response variable, also have a unimodal shape with right-skewedness, range from 22600 to 118800, and have a median 40800 and a mean 42547. 50% of the values are within 36000 and 46800, showing that values are highly concentrated around that interval. This suggests that it might be better to use a log transformation on Median earnings later in the modeling section. **(3)** Figure 2 shows bivariate relationships between variables. It is observed that price and laon display non-linear relationships while other variables display linear relationships. SAT with Pell grant (cor = -0.6924474), SAT with loan (cor = -0.4932500), and Pell grant with earnings (cor = -0.5652098) seem to have a negative relationship, and SAT with earnings (cor = 0.6645637) and Pell grant with loan (cor = 0.5017520) seem to have a positive relationship. **(4)** We see that both Pell grant and loan have a negative relationship with SAT and these two seem to be related to each other as well. Hence, it might be helpful to take into account thier relationship in the models.

Modeling & Diagnostics

(1) Based on the EDA, it is reasonable to consider SAT score, Pell grant, loan, and average net price for analyzing an association with the median earnings. The first linear model has these five factors as independent variables, and median earnings as a response variable. The second model is additive, with price and loan being non-linear according to the previous EDA. The response variable is used with log transformation, for it was a highly skewed variable that's better to be transformed into a more normalized dataset. **(2)** It also includes a multiplicative term to control for the covariates, Pell grant and loan, as we discovered a possible association between those two variables in EDA. Hence, the additive model takes into account an interaction between Pell grant and loan. The reason we only want to control for Pell grant and loan is that they both had a negative relationship between SAT so it might over-adjust for covariates if we also add multiplicative terms for SAT with those variables. It also turns out that R-squared value is the highest when only one multiplicative term between Pell grant and loan is included. **(3)** Figure 3 and Figure 4 depict Diagnostic residual plots with normal Q-Q plots for both linear and additive models. For the linear model, we see that the residual plot doesn't display the traits of non-constant variation and the values are randomly dispersed around the horizontal axis 0. There are only few outliers, so a linear regression model seems to be appropriate for the data. On the other hand, if we look at the normal Q-Q plot, although a lot of points are clustered on the 45 degree line, points start to diverge in the end quite a lot. This implies that the data may not be normally distributed, violating linearity assumption about general linear regression model. Yet, the addtive model shows an improvement. Its residual plot shows more dispersed points

around 0 and its normal Q-Q plot presents points clustered on the line quite well and only few points diverge from it. Such linearity of the points suggests that the data are normally distributed; hence, our inference about the model is more likely to be accurate. Furthermore, R-squared value is bigger for an addtive model (0.521 vs. 0.483), suggesting that the proportion of the variance for a dependent variable that's explained by an independent variable is bigger when the additive model is used.

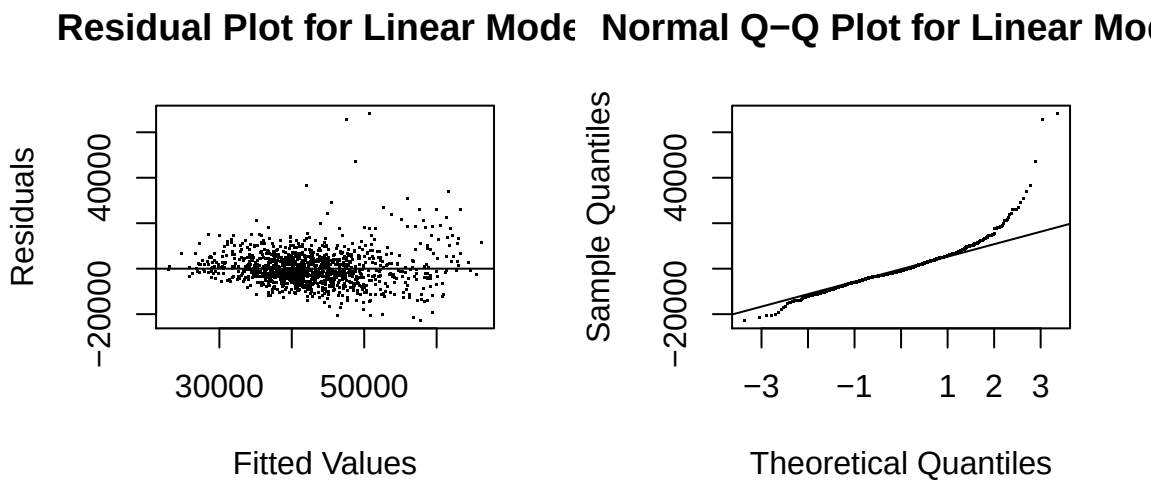

Figure 3: Diagnostic residual plot for linear model.

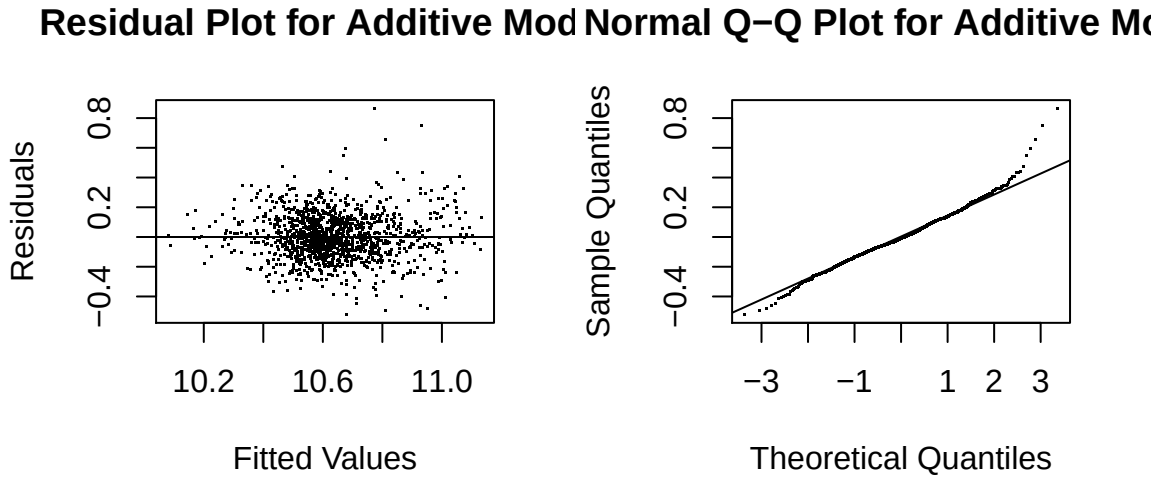

Figure 4: Diagnostic residual plot for additive model.

(4) Using the 5-fold cross-validation, we are able to obtain two models' prediction error and and the corresponding standard errors. For the linear model, the average of the five cross-validation error was 54683159 and its standard error was 7579869. For the additive model, the average of the five cross-validation error was 52126860 and its standard error was 7943152. Based on the prediction error, additive model fits best to the data. **(5)** However, given the huge standard error, which is above 7,000,000 for both models, our estimates of the prediction error are quite uncertain. Hence,

the difference between two models doesn't appear significant. Along with the residual diagnostics, we've seen that the additive model performs better in both justifing the use of linear regression models and the prediction error. **(6)** On the other hand, we don't know the right distribution for the noise and the shape for the regression function is only assumed, so parametric bootstrap doesn't seem to be ideal for this data. Since resampling cases assumes nothing about the shape of the regression function or the distribution of the noise, resampling cases seems to be appropriate for this data to be safe.

Results

(1) Using the additive model, we are able to determine that students who attend more expensive school earn more money after graduation. There is a strong evidence that the true slope is not zero as p-value for average net price is 3.03 X 10^-09 after accounting for other factors as well. Figure 5 displays a positive relationship between average net price and the fitted values of log median earnings, showing that more tuition is linked to higher median earnings.

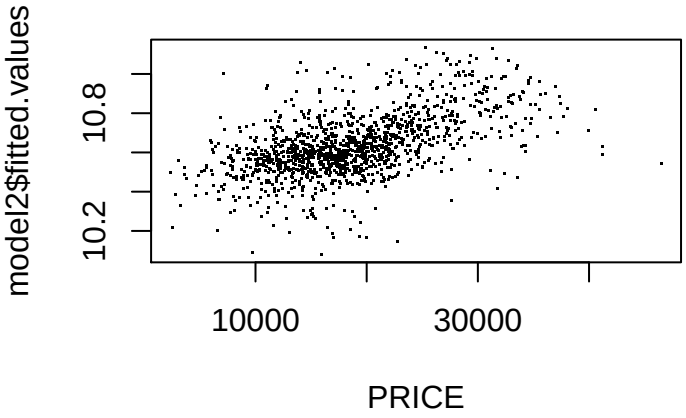

Figure 5: Scatter plot between average net price and fitted log median earnings

In order to see whether the relationship between price and earnings is the same at public, private, and for-profit, the additive model is further adjusted by having CONTROL that distinguishes the type of institutions, and the interaction term between price and control is added to account for thier relationship since private universities are expected to have a higher tuition. F-test is used to compare the orginal additive model with the adjusted one. **(2)** The null hypothesis states that the original additive model with no control and interaction variables fits the data as well as the adjusted model. The alternative hypothesis states that the new model taking into account the type of universities fits the data better than the original model. In order to do F-test, we assume that the data are normally distributed and that samples are independent from one another. **(2)** The F-statistic is about

8, indicating that there is a significant difference between the means of the samples being compared (variation among group means is also expected to be significant). This implies that the relationship between price and earnings is different at public, private, and for-profit institutions. Figure 6 also shows that the positive correlation between earnings and average net price is the strongest for private, followed by public and for-profit.

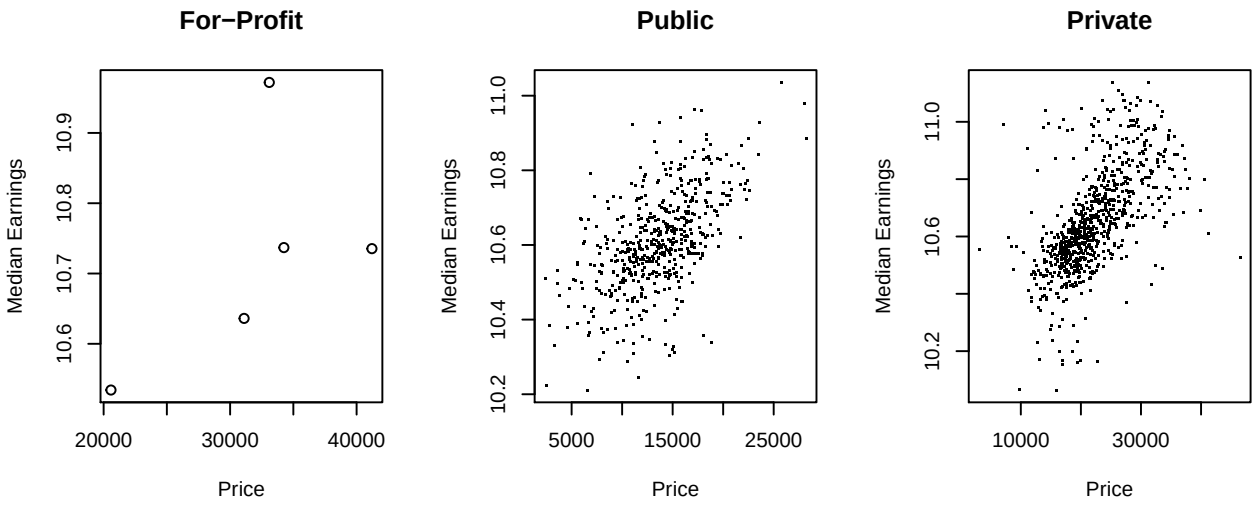

Figure 6: Scatter plot between average net price and predicted log median earnings for three types of college

(3) Using an additive model, we can build a confidence interval for the mean earnings of students after graduation for a school just like Carnegie Mellon. A confidence interval with 95% confidence is [62554.26, 67060.33]. Hence, we are 95% confident that the interval [62554.26, 67060.33] captures the true mean earnings of students after graduation for a school like Carnegie Mellon. Here, we are assuming that the samples from each population is independent of one another and that the populations from which the samples are taken are normally distributed. **(4)** The pivotal confidence interval from bootstrapping is also obtained, and it's [62252.71, 67393.44], which is quite close to the one we found by predicting. We believe that 95% pivotal confidence interval from bootstrapping is more reliable and safer. The confidence interval from predicting assumes normal distribution and variable independence, while bootstrapping by resampling cases assumes nothing. However, as a bias-variance trade-off, a bootstrapping confidence interval gives a wider and weaker bounds, losing precision.

Conclusions

(1) The results of the analysis provide strong evidence for a relationship between average net price and median earnings; On average, students who attend more expensive schools earn more money after graduation, and this pattern was the most apparent when the univeristy is private and the least apparent when the university is for-profit. Based on the model, the expected mean earning for students at institutions like Carnegie Mellon is about $64,768.12 using bootstrapping by resampling cases.

(2) The higher tuition is, the more resources and faculties for education. Hence, it could be possible that higher average net price leads to a better education for students, making them more competitive for jobs. The reason this pattern is more obvious for private schools could be that they are not regulated by the state and federal funding, so they can choose to invest more in education. For for-profit schools, the goal is to operate as a business and generate a positive return from the students, so it's most likely that they will focus mostly on profits for thier shareholders instead of the quality of education. Hence, the relationship between the average net price and earnings by students could be the weakest for for-profit schools. However, it's important to note that there may be other confounding factors influencing our analysis. For instance, such a relationship between the average net price and median earnings could vary by race, minorities, location, etc. Thus, for the further analysis, it is suggested to account for other possible factors to decide whether more expensive institutions are worth their high tuitions. In addition, since our boostrapping analysis lacks accuracy, it might be useful to further examine whether our model is correct (i.e. we have a right distribution for the noise and a right shape for the regression function) and if it is, we could employ the parametric bootstrap for precision.