Regression Analysis on Early Career Pay

Branden Lopez, Brandon Palomino
MATH 261A, Section 02, Dr. Bremer
05/23/2022

TABLE OF CONTENTS

I INTRODUCTION/MOTIVATION

College tuition is somewhat difficult to analyze. There are many components that contribute towards the cost value of a school's tuition and how it affects the school students. Over time, questions arise about how tuition correlates to career pay for young individuals entering the workforce after graduating. Does attending private schools lead to higher salaries compared to attending public colleges? Is there a difference between a four-year college and a two-year college when analyzing a student's career pay? These questions lead to solving problems in regard to understanding college tuition and earning salaries for students.

The data set comes from the US Department of Education and has been compiled by TuitionTracker.org. The data set consists of many observations and has numerous distinct variables across multiple CSV files. Tuition and fees by college/university along with school type, degree length, state, and in-state vs out-state are organized under "tuition cost" and "tuition income." Diversity of the student demographic by college/university are organized under "diversity school." Career salary earnings for new grads are listed under "salary potential." Combining all parts includes college information such as early career pay, mid-career pay, state, type (i.e. public vs. private), degree length, in-state tuition, out-of-state tuition, and more.

For the purpose of this data analysis on the college tuition data set, Python will be used to join individual tables into a cohesive data set with around 10 predictor variables. Furthermore, Python is used for data pre-processing to mitigate missing values. Finally, R will be used to conduct a careful regression analysis: describing data, selecting predictor variables, transforming variables if necessary, studying outliers, influential points, and conducting variable selection. Following the required regression analysis will output a final fitted model that best describes the data set.

We attempt to specify predictors of early career salaries for students across numerous colleges using regression analysis. By solving this issue, we hope to gain a greater understanding of how cost and other college characteristics affect students' earnings.

II DATA PREPOCESSING

The “College tuition, diversity, and pay” spans 5 individual tables collected from various sources thus, the tables need to be joined. A natural choice for joining the tables is with an inner join; an inner join combines records from two tables whenever there are matching values in a field common to both tables. The primary key for joining is “name”, resulting in universities that are string literals of one another to remain after the join. Of the 935 instances originally in salary, 628 remain after joining. It’s worth mentioning that Fuzzy-Matching, a method that checks if two words are similar within an integer-defined number of operations (insert, deleted, replace), could increase the remaining instances by allowing for words such as San Jose and San José to be matched.

Our response, early career salary is measured when first entering the workforce, and so a predictor that describes the school’s status the year before entering the workforce can reasonably affect income shortly after graduation. Due to this, we narrow school metrics on the year 2017 a year before early career salary is recorded.

One of the 5 tables, “historical tuition” doesn’t contain any “name” variable and cannot be joined. Despite this, the national average tuition cost is present from the 1980’s to 2017, and other tables contain how much a school costs to attend after a scholarship. Due to this, we are able to create a new categorical variable $tuition\_higher\_than\_national\_average$ after scholarships. The remaining information is excluded from consideration.

Lastly, we explore an understated fact about linear models (at least in class), Null values. A linear regression model will not work on missing or Null values so these values must be mitigated. Rather than drop values, we use a simple impute method on the 6 null values, replacing the nulls with the average value in their respective columns.

Therefore we end with 3 categorical variables, $Type$, $DegreeLength$ and $tuition\_higher\_than\_national\_average$, as well as 13 continuous variables. All variables are listed below.

$y = early\_career\_pay$

$$\begin{aligned}
x_1 &= make\_world\_better\_percent \\
x_2 &= stem\_percent \\
x_3 &= total\_price \\
x_4 &= net\_cost \\
x_5 &= tuition\_higher\_than\_national\_average \\
x_6 &= type \\
x_7 &= degree\_length \\
x_8 &= room\_and\_board \\
x_9 &= in\_state\_tuition \\
x_{10} &= in\_state\_total \\
x_{11} &= out\_of\_state\_tuition \\
x_{12} &= out\_of\_state\_total \\
x_{13} &= Total.Minority \\
x_{14} &= total\_enrollment
\end{aligned}$$

III ASSUMPTIONS AND VARIABLE TRANSFORMATIONS

To determine the linearity between predictors and response, we analyze each of the predictors against early career pay. Scatter can be used to plots explore these predictors against the response; a linear correlation verifies linearity. Figure 1 illustrates the scatter plots for several predictors against early career pay.

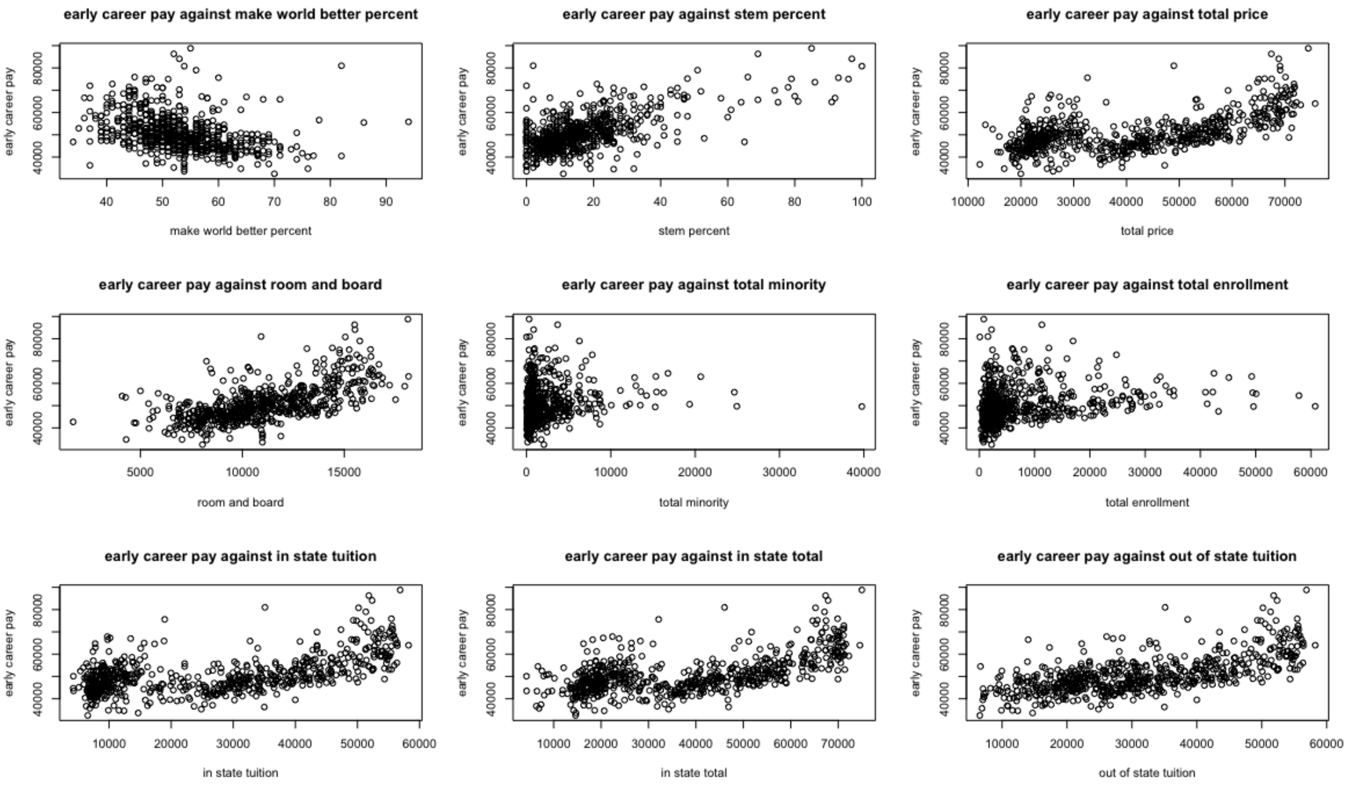

Figure 1: Predictors against early career salary

A majority of the scatter plots describe the relationship as linear. Despite this, scatter plots such as ”total minority” and ”total enrollment” have values clustered in one location and then a small linear trend. Despite this do not transform these variables as no obvious transformation is known to us.

For the full model, a linear model will be fitted against early career pay in order to create a qq-plot of the residuals as well as a scatter plot of the residuals against the predicted response. Figure 2 illustrates the residual plot and qq-plot of the predictors against early career salary.

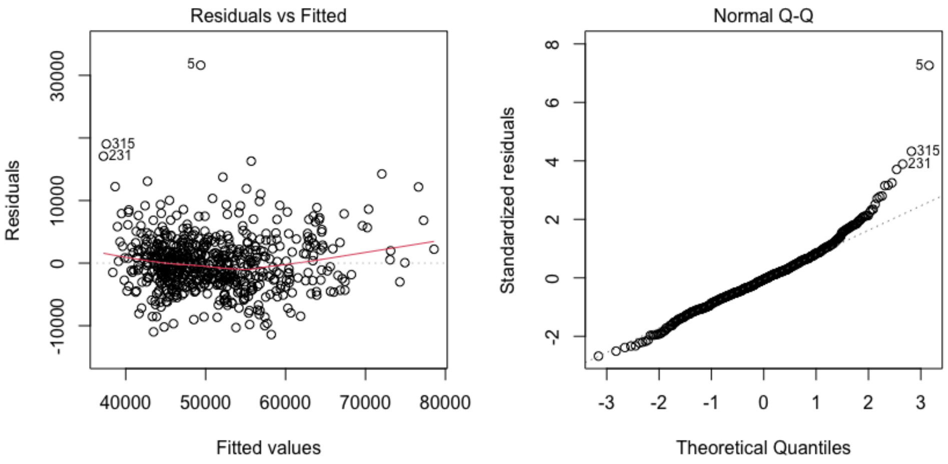

Figure 2: residual plot & qq-plot for whole model

Although arguments could be made on whether the residuals are approximately normally distributed (i.e: qq-plot points lie approximately on line), there are an abundance of outliers on the qq-plot that stem away from the line. Also, the residual plots show that the residual variance in these fitted models is nearly constant. Therefore we consider a variable transformation on our response.

The box-cox method is applied to the data set to find the best transformation on the response. Figure 3 shows the box-cox of the college tuition data.

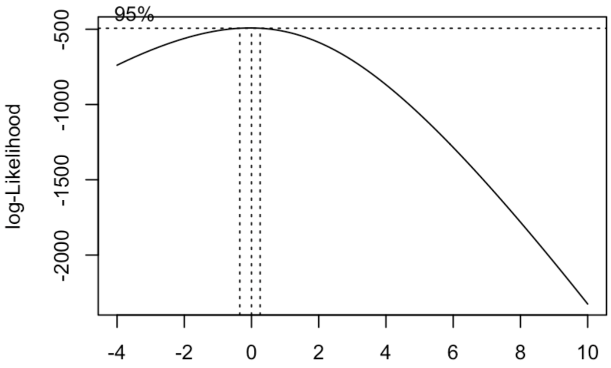

Figure 3: Predictors against early career salary

The value λ that maximizes the model log-likelihood is $\hat{\lambda} = 0$. An easily interpretative value in the 95% confidence interval for λ is 0. This corresponds to a log-transformation on the response: y → log(y).

Thus, our only transformation on the fitted model will only be a log transformation on the response. Figure 4 shows that the transformed model produced a qq-plot being normally distributed with its residuals.

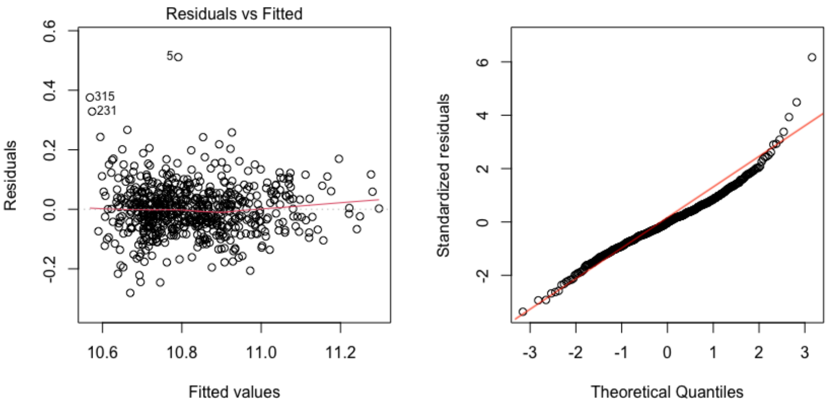

Figure 4: Residual plot & qq-plot of transformed response

IV OUTLIERS

While the logarithmic transformation reduces residual error, the residual vs fitted and qq plots showcase many outliers. We use Cook's Distance Bar Plots to verify the claim outliers exist. Many thresholds are used to determine if a cook's distance value is an outlier; while we use the threshold of 1 in class, many pragmatic articles use 4/n as the threshold. This lower threshold vindicates the exploration of believed 'outliers' even if they are not removed.

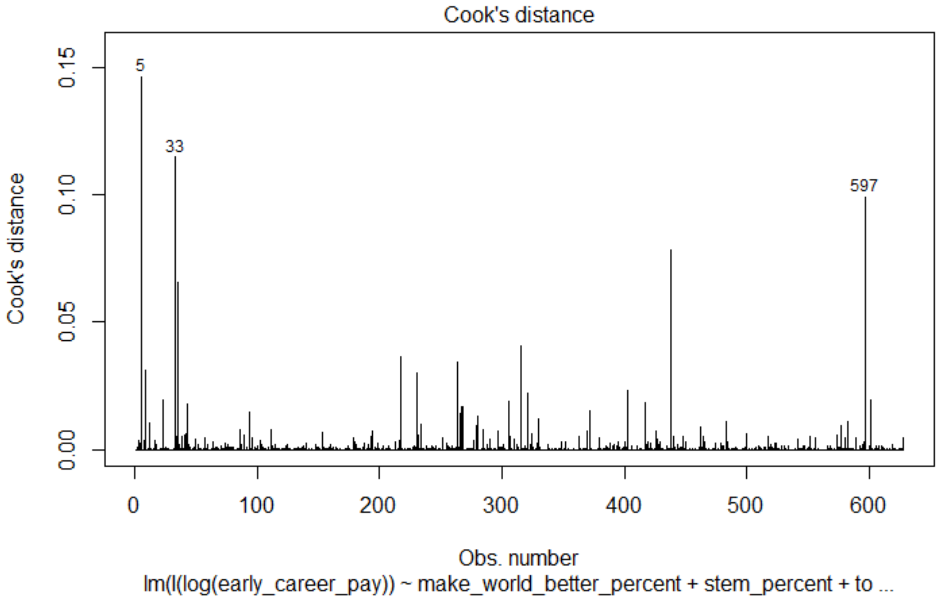

Figure 5: Cook's Distance Box Plot

With a threshold of .64, many observations are outliers. Rather than delete outliers, we investigate what makes an outlier. The first six outliers (5, 33, 35, 438, 315, 231) show nothing unique in the continuous or categorical variables; however, the 'name' label shows that two of the six are health colleges providing specialized education in the high-income fields of nursing, pharmacology, and various other health sciences. Further "by-hand" google analysis uncovers that four of the six outliers are health-science institutions. After modeling, we analyze the predicted salary vs.

the real salary and find that these health institutions have salaries Significantly higher than what is predicted.

Instance No.	Predicted Early Career Pay	Real Early Career pay
5	46781	81000
35	42000	55800
315	38333	56600
231	39200	54300

Table 1: Final model prediction on outliers.

Health science intuitions form a noticeable characteristic of our outliers; making an indicator variable to capture this behavior could produce better modeling results. However, creating this variable would require manual annotation and careful validation of 628 institutions, a time-consuming outside of the scope of this project. Instead, we remove only the most significant outlier as doing so leads to a 2% R^2 increase.

V VARIABLE SELECTION

With linearity, normality, constant variance of the residuals, and outlier investigation done, we are ready to select a subset of variables that accurately describe early career salary.

To start off variable selection, we check the Variable Inflation Factor (VIF) for multicollinearity. This shows that many variables contain high multicollinearity (VIF $>$ 10) and for good reasons. The net cost of attendance is naturally the cost of tuition plus room and board. Early linear models found these variables to be a perfect linear combination; therefore, removing variables is important to eliminate multicollinearity. Since we want to predict early career income for both in-state and out-of-state students, we remove variables that distinguish the two, solving all multicollinearity issues.

An Exhaustive search is lackluster in distinguishing subsets of predictors as the top 3 models for each level of predictors 4 or more clusters around: $R^2 = .7$, $MS_{Res} \approx .006$ and fails to produce Mallow's CP value less than the number of predictors used. Instead, a "by-hand" Forward and backward variable selection is performed. Both of these methods arrive at the same predictors for our final model.

$$\begin{aligned} log(y) &= 10.36 + 3.945 * 10^{-3}x_2 + 5.76 * 10^{-6}x_3 \\ &+ .009x_6 + 8.6 * 10^{-6}x_8 - 4.805 * 10^{-6}x_{13} \\ &+ 5.09 * 10^{-6}x_{14} \end{aligned}$$

With all model assumptions checked and model selection done we adopt this as our final model.

VI FINAL MODEL

To save time flipping between the variables page we convert the variables to their respective names.

$$\begin{aligned} log(y) = 10.36 &+ 3.945 * 10^{-3}StemPercent + 5.76 * 10^{-6}TotalPrice \\ &+ .009Type + 8.6 * 10^{-6}RoomAndBoard - 4.805 * 10^{-6}TotalMinority \\ &+ 5.09 * 10^{-6}TotalEnrollment \end{aligned}$$

Our model produces a low $SE_{RES} = .08$, with $R^2 \approx .72$. Thus our final model explains 72% of the variability in early career income, and while this is satisfactory, the log-transformed response changes the interpretation of coefficients. Consider a Simple linear model where the marginal change in a variable is additive, but in a log-transformed response linear model it is now multiplicative. Let's illustrate this fact with a simple linear model (Note: In R log is natural log)

$$\begin{aligned} log(y) &= \beta_0 x_1 + \beta_1 x_1 \\ y &= e^{\beta_0 x_1 + \beta_1 x_1} \\ y &= e^{\beta_0 x_1} e^{\beta_1 x_1} \end{aligned}$$

Thus a one unit increase in Stem Percent (holding all else fixed) is a $exp(.0003945) = 1.003953$ or 0.4% increase in early career salary. Likewise, each additional minority student results in a 0.00048% decrease in early career salary.

Unfortunately, the model tells us that increasing the number of minority students at an institution decreases early career salary. There are many reasonable assumptions to this conclusion: perhaps the model is wrong, or the model's attribution of minorities reflects de-facto discrimination in the United States. The class focused on the theory of regression and not ethics, so we cannot confidently come to a conclusion by omitting the Total Minority Variable, as we rejected the null hypothesis that this variable's coefficient should be zero.

Overall our model tells us that Large-expensive public intuitions with a high percentage of STEM students produce higher average early career salaries and that minority students decrease average early career salary.

VII Conclusions

In summary, applying transformation techniques, removing outliers, and conducting variable selection validated Regression Modeling on the college tuition data. 72% of the variation in "early career salary" can be explained through our final subset of predictors. This final model indicates that large expensive institutions with a large stem percentage tend to produce graduates with higher early career salaries. Another interesting note about the final model was that STEM students produce higher average salaries whereas a large minority student population results to lower average salaries. However, this data applies to students in the US at already good universities and does not reflect students' individual performances.

Given more resources, several improvements could be made to better reflect "early career salary". After joining tables 628 out of 935 instances remained, and Fuzzy matching during the pre-processing phase could have resulted in more instances for modeling. Utilizing more of the tuition data could also lead to better coefficient descriptions and even change residual variance. Institutions that specialize in healthcare-related professions were outliers; thus, an extra categorical variable to account for this could increase performance. Finally, implementing a model that is not discriminatory would have given the model other reasons for why students from specific schools were receiving different salaries compared to others.

While linear regression might not have been an optimal model for the college tuition data, many trends have been confirmed and even uncovered about how costs, STEM, and minorities can influence a student's earnings when entering the workforce.

VIII REFERENCES

[1] Mock, Thomas. (2018) College tuition, diversity, and pay, Version 1. Retrieved November 17 2022, from https://www.kaggle.com/datasets/jessemostipak/college-tuition-diversity-and-pay?select=tuition_income.csv.

[2] Ford, Clay. ”Interpreting log transformations in a linear model,” University of Virginia Library StatLab articles. August 17, 2018. Accessed December 5, 2022.

[3] Montgomery, Peck & Vining, Introduction to Linear Regression Analysis. 5th edition, 2012. ISBN 9780470542811