Habib University

CS457

Data Science Techniques

Analysis of the college scorecard data

Shaikh, Mudasir
ms03831@st.habib.edu.pk

Submitted to:
Dr. Zeehasham Rasheed
May 8, 2019

Abstract

One of the most difficult questions that students from across the world ask themselves during their academic lives is the question of the best college, or to put it differently, which college suits them the best. The U.S. government in their effort to increase transparency compiled a large set of data about colleges around the U.S. Through this report, we analyze some elements of that data to answer questions such as **"How does cost of a college relate with the mean earning of its graduates?"**, **"How does admission rate of a college relate with the average SAT score of its students?"**, etc.

Contents

1 Introduction

"Find the college that's the best fit for you!" reads the tagline of the College Scorecard Data website. The College Scorecard is an online tool, created by the United States government, for consumers to compare the cost and value of higher education institutions in the United States(4). The project is designed to increase transparency, and is aimed at enabling the students and their families to have the necessary information before making a judgment on what may be termed as one of the most important journeys in a student's life. These data are provided through federal reporting from institutions, data on federal financial aid, and tax information(5). Through this project, we aim to analyze some aspects of this huge data, the analysis are due to our own understanding of the data and may not necessarily be hundred percent accurate.

2 Requirements and Resources needed

- seaborn.
- Pandas.
- NumPy.
- SciPy.
- matplotlib.
- scikit-learn.
- LaTeX for this report.

3 Dataset Description

According to the official website of College Scorecard under the US department of Education, the data available from this website covers a wide range of aspects such as Academics, Cost, Loan Repayment, Completion rates, Earnings etc(3). The data set spans nearly 20 years of data. A quick sneak peak into the dataset reveals the following information:

3.1 Rows - Observations

There are over 7000 observations in the entire dataset, each observation corresponds to a unique row in the data. Each row has measures for all the different features of the dataset.

3.2 Columns - Features

There are over 1800 features in the entire dataset, each feature corresponds to a unique column in the dataset. All but a few columns have missing values in them. Only 21 out of the 1899 data features have non-missing values, the rest have varying number of missing values. Almost all features are stored as non-null objects, which calls for some pre-processing. However, we will only process the columns needed for the analysis in this project.

3.3 Selected columns

After some basic exploratory analysis, and refining a few research questions, we reduce the scope of our analysis to the following variables:

Variable	Description	Data type
UNITID	id of Institution	integer
INSTNM	Name of institute	float
PREDDEG	Predominant undergraduate degree awarded	float
NPT4_PUB	Average price of public institution	float
NPT4_PRIV	Average price of attending a private institution	float
CONTROL	Type of Institute: Public/Private nonprofit/Private for-profit	integer
ADM_RATE	Admission Rate	float
RPY_3YR_RT_SUPP	3 year repayment rate[1]	float
COSTT4_A	Average cost of attendance (academic year institutions)	float
COSTT4_P	Average cost of attendance (program-year institutions)	float
MN_EARN_WNE_P6	Mean Earning of Graduates 6 years after entry	float
MD_EARN_WNE_P6	Median Earning of Graduates 6 years after entry	float
SAT_AVG	Average SAT equivalent score of students admitted	float

Table 1: Selected Columns

For our analysis, we restricted ourselves to institutes that have 4 years bachelors as their predominant degree. We further broke down NPT4_PRIV into Private for-profit and Private non-profit. As for the two variables of cost, we merged the two columns (as an institution would be either academic-year or program-year) and saved it into a new column **"cost"**.

4 Analysis Results

4.1 Scatter plot

4.1.1 Admission Rate vs Average Sat Score

The first question that we came up with was to analyze the relationship between the Admission rate of a college and the average SAT score of the students enrolled. Our instinct suggested that there wouldn't be any relationship between the two, however we went ahead and plotted a scatter plot of Admission rate against average Sat scores. But before plotting, we did some data cleaning, for the average sat score, there were quite a few null values in the column, and our intuition suggested that we do mean imputation, so we replaced the null values with the mean of the non-nulls. For the admission rate, we did the same.

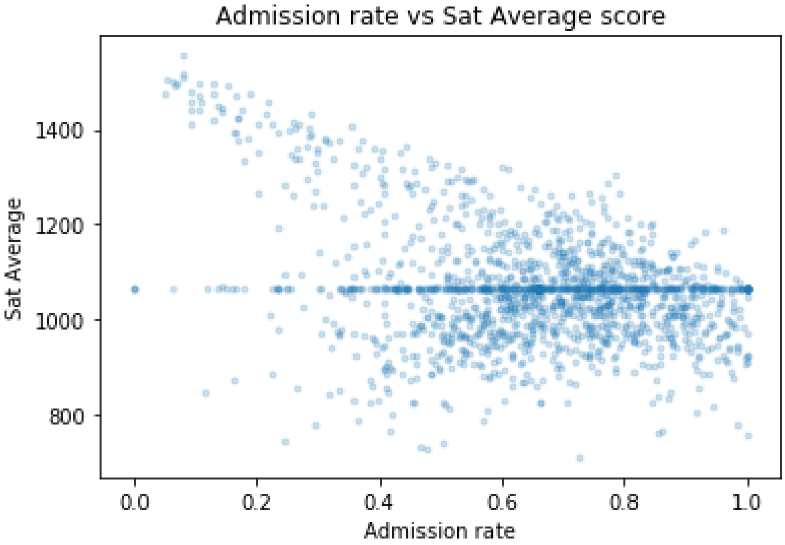


Figure 1: Admission Rate vs Average Sat Score

Firstly, observe that there are quite a few outliers. For the sake of simplicity, let's ignore them for now. The result, that is the plot suggests that there is a negative relationship between the two, albeit weak. Although this relationship is not very significant, it still gives us some information such as colleges that have low admission rate, may have students who had high SAT scores. This is what we see in real world as well, the big colleges that have low admission rate, have a very high cut-off range for SAT scores.

4.1.2 Earning of graduates vs Cost of attendance

Through this question, we intend to explore how the cost of attendance of a college relates with the earning of its graduates. We chose the 6 years earning variable, that is, earning of students 6 years after entry. Our intuition suggested that there could be a moderate relationship between the two, however we went ahead and plotted a scatter plot of earning against cost. But before plotting, we did some data cleaning, for the earning variable, we chose two variables, the mean earning and the median earning of students 6 years after entry, there were quite a few null values in the columns, and in addition to these null values, some observations had

"Privacy Suppressed" values, which may be categorized as non-random nulls. We got rid of those, and for the random nulls, our intuition suggested that we do mean imputation, so we replaced the null values with the mean of the non-nulls. For the cost of attendance, we did the same.

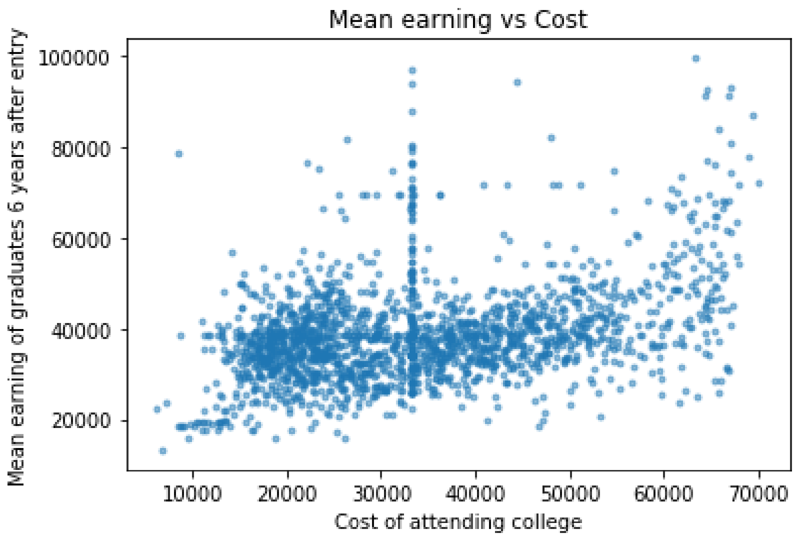


Figure 2: Mean earning of graduates 6 years after entry vs Cost of attendance

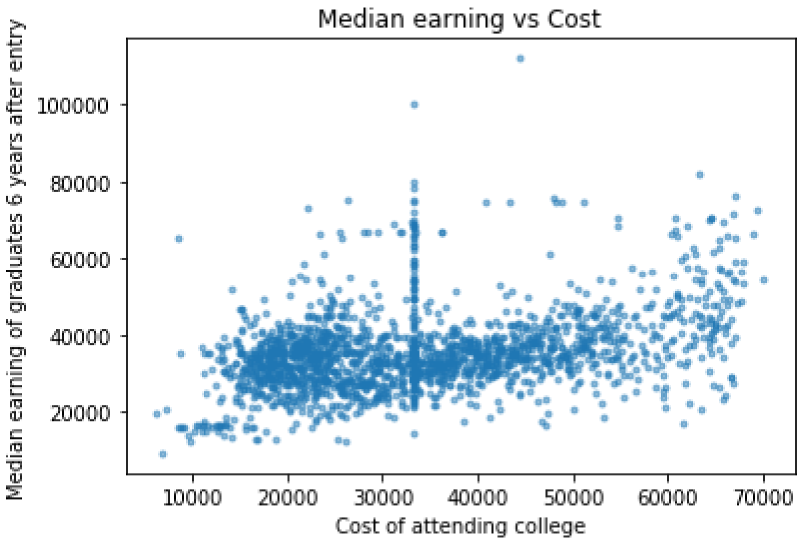


Figure 3: Median earning of graduates 6 years after entry vs Cost of attendance

Firstly, observe that there are a few outliers. For the sake of simplicity, let's ignore them for now. The result, that is the plot suggests that there is a moderate to high positive relationship between the two. The plot suggests that colleges with high cost of attendance tend to produce graduates having high earnings.

4.2 Boxplot

4.2.1 Cost of institution

We can see that the median cost is somewhere between 30000 to 40000 USD. There appears to be slightly more variability in cost between the median and lower quartile as compared to median and upper quartile, that is, colleges that have a cost above the median cost tend to have more agreement within them in terms of cost. This plot doesn't tell much, Let's plot it with categorical variables of type of institution.

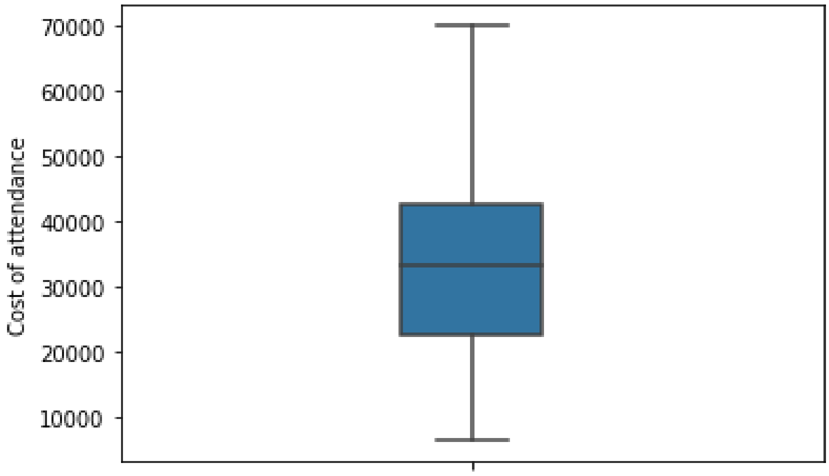


Figure 4: Boxplot of Cost of institution

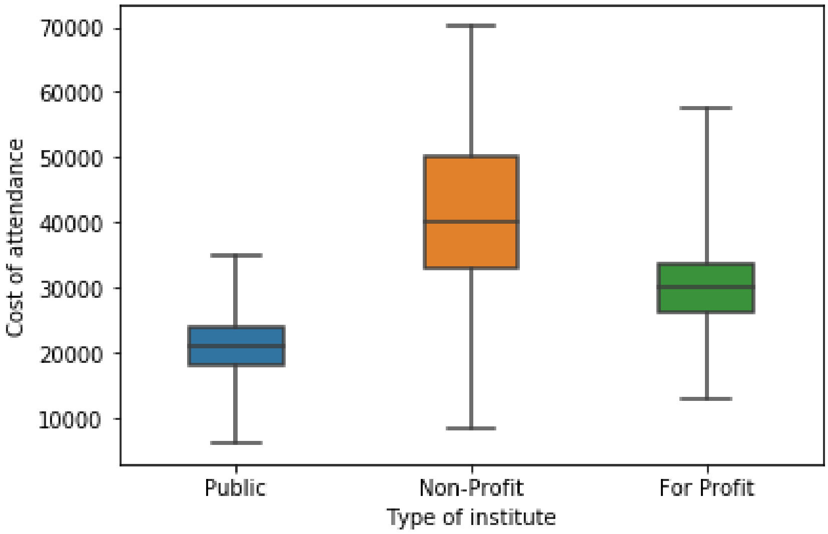


Figure 5: Boxplot of Cost with Categorical Variables (Type of institution)

The plot above groups the colleges in terms of their funding type and then compares their cost. From the plot it is clear that, colleges of different types vary in terms of cost, which is what we would expect. Private, Non-profit colleges tend to have higher cost, the median, upper quartile, lower quartile and whiskers all are the highest, and have more variability for non profits than for the other two. The high cost is not what we would expect in general from non-profit colleges, but this is what the data suggests. Also notice that the Public colleges fall at the bottom in terms of cost, there is also the least amount of variability in cost of Public colleges, which is what we would expect.

4.2.2 Family Income

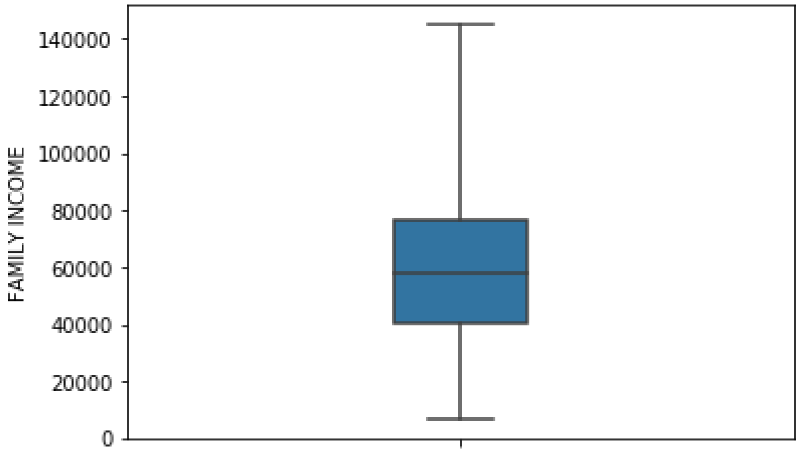


Figure 6: Boxplot of Family Income

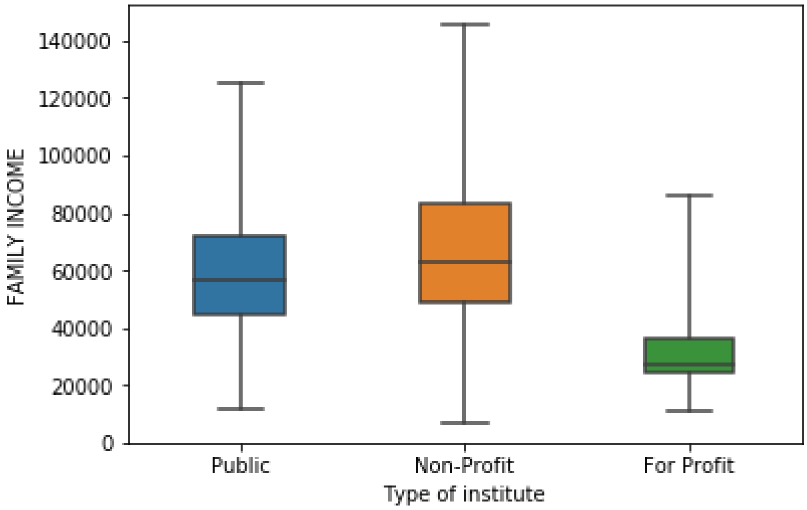


Figure 7: Boxplot of Family Income with Categorical Variables (Type of institution)

The analysis for Family income is similar to what we have outlined in the boxplot analysis of cost, which is why the common analysis is left out for this plot. What might seem interesting, and perhaps surprising is that the Family income of students from For-profit colleges is the least, with the least median, and inter-quartile ranges.

The family income of students from Non-profit colleges is the highest, with more variability than the rest, which is what we would expect following our analysis for cost.

4.2.3 Repayment Rate

Similar to the variables that we have dealt with in previous sections, we first processed the repayment rate column [1] by cleaning it and getting rid of the null values. We did mean imputation in this case.

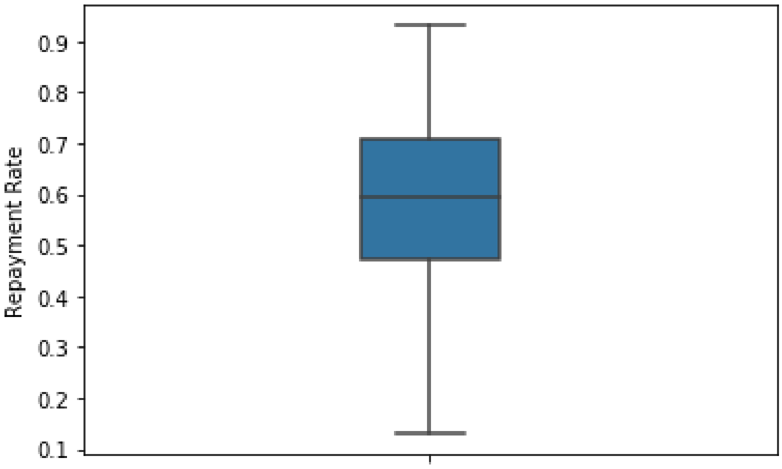


Figure 8: Boxplot of Repayment Rate

[1] Depicts the fraction of borrowers at an institution who are not in default on their federal loans. A borrower is considered in repayment if his or her loan payments, at the time of measurement, covers all accrued interest (post-separation) and at least $1 more(5)

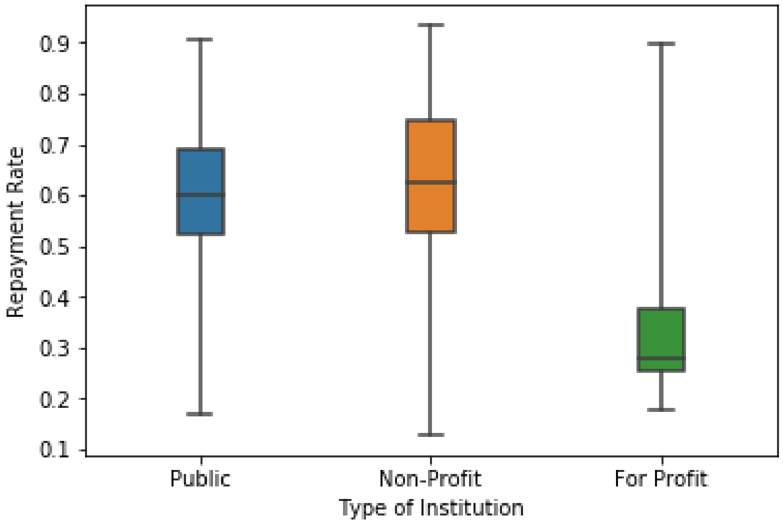


Figure 9: Boxplot of Repayment Rate with Categorical Variables (Type of institution)

The analysis for Repayment Rate is again very similar to what we have outlined in our previous two boxplot analysis. Which is why the common analysis is left out for this plot. What seemed interesting to me was that the Repayment Rate of students from For-profit colleges is the lowest, with the least median. Which is not what we would expect initially following our analysis of cost, because according to our intuition, the higher is the cost, the lower should be its repayment rate, however, this is not to say that the cost is the only or even strong indicator of repayment rate, but cost according to our understanding should affect the repayment rate, which is what the plot doesn't suggest, since non-profits, that have the highest cost, also have the highest repayment rate. One possible explanation of this could be that, students who attend colleges with high-costs more often belong to richer families (this is also backed up by our previous boxplot) and hence may receive support by their families to repay the loan. In addition to this, our scatter plot of cost vs earning suggested that non-profits (that is, colleges with higher cost) tend to produce graduates with higher mean earning, and higher earning in turn may result in high repayment rate.

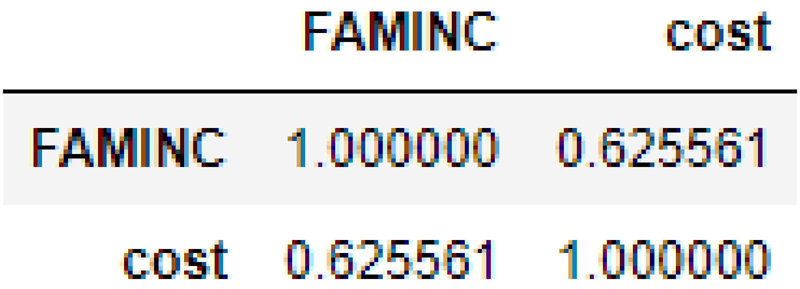
4.3 Correlation analysis

4.3.1 Family Income and Cost of College

For correlation analysis, we came up with a seemingly absurd question:

Is the family income of the student and the cost of the college they go to related?

The answer, to our surprise, was yes! In fact, they have a positive correlation coefficient of around 0.6, which can be regarded as highly moderate. This result is very interesting, and to the best of our understanding means that students from families having higher income tend to go to colleges with higher cost, and vice versa. This interpretation doesn't sound as absurd, as our initial question did.



	FAMINC	cost
FAMINC	1.000000	0.625561
cost	0.625561	1.000000

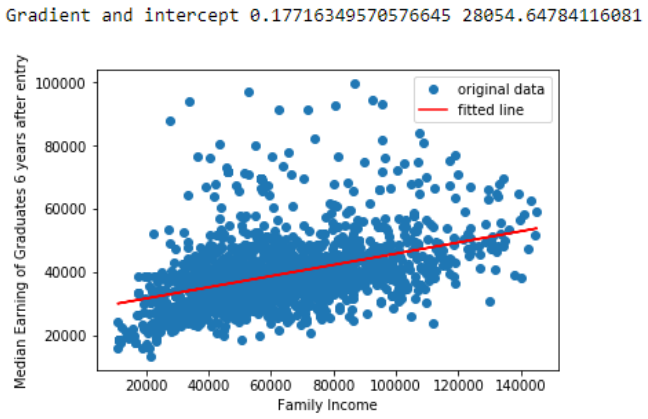
Figure 10: Correlation matrix of family income(FAMINC) vs Cost of college(cost)

4.4 Simple Regression: Linear Model

For the regression analysis, we fitted Median Income (response variable) against three different explanatory variables:

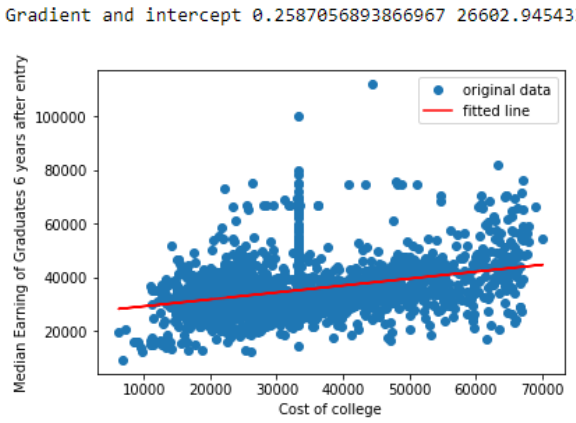
- Family Income
- College Cost
- Admission Rate

The fitted models with their gradient(slope) and intercept are shown below:



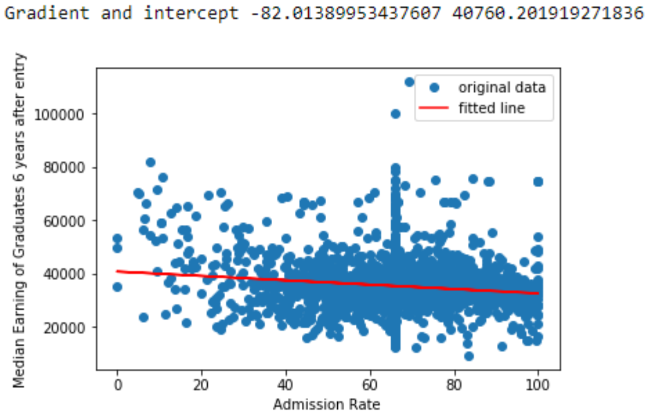
```
Gradient and intercept 0.17716349570576645 28054.64784116081
```

Figure 11: Linear Model of Family Income(Explanatory) and Median Income(Response)



```
Gradient and intercept 0.2587056893866967 26602.94543396997
```

Figure 12: Linear Model of College Cost(Explanatory) and Median Income(Response)



```
Gradient and intercept -82.01389953437607 40760.201919271836
```

Figure 13: Linear Model of Admission Rate(Explanatory) and Median Income(Response)

Let's analyse the 2nd graph, i.e., the linear model of College Cost(Explanatory) and Median Income(Response). Let M be the Median earning, and C be the cost then the equation of the fit is

$$M = 0.26 * C + 26602$$

The value of the slope-coefficient is not very high, but still has significance: it tells the amount of change in M that can be expected to result from a unit increase in C. That is, if the cost of a college increases by 1000$ the Median earning of its graduates is `expected` to increase by 0.26*1000 = 260$.
The other two graphs, or models, can be analysed similarly.

4.5 Hypothesis testing

Let's formulate our hypotheses. For hypothesis testing, we select the following two features `"ownership of institute"` i.e. whether the institute is for profit, public or nonprofit. The other variable we select is `"Average cost of institute"` i.e. how much it costs to attend an institute.

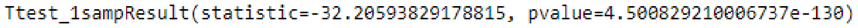
4.5.1 Null hypotheses

The average cost of public institutions is not different from average cost of population(all institutions).

4.5.2 Alternate hypotheses

The average cost of public institutions is less than average cost of population(all institutions).

4.5.3 Result of Hypothesis testing



```
Ttest_1sampResult(statistic=-32.20593829178815, pvalue=4.500829210006737e-130)
```

Figure 14: Result of hypotheses test

The average cost(mean) of Public institutions was `14010` whereas the average cost(mean) of all institutes was `19561`. We performed a one sample t-test on the cost of Public institutions against the mean of cost of population (all institutes). The result was a p value of approximately 0, and since a small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis(6), we reject the Null hypothesis. This, and the means of the two samples, indicates that the average cost of public institutions is less than the average cost of all institutes.

References

[1] Dataset “College Scorecard Data” (Office of Planning, Evaluation, and Policy Developments, U.S. Department of Education, 2017).

[2] ‘Using Federal Data To Measure And Improve The Performance Of U.S. Institutions Of Higher Education’ Executive Office of the President of the United States

[3] Data Documentation, College Scorecard Data

[4] The College Scorecard Data, Wikipedia

[5] Full Data Documentation, College Scorecard Data

[6] what-a-p-value-tells-you-about-statistical-data