The relationship between cost of tuition and earnings after graduating

Zhiyuan Yang

University of California, San Diego

ECON 5/POLI 5D: Data Analytics

Spring 2021

Abstract

This report analyzes the relationship between the student's college tuition and the income after graduating. Dataset on information of colleges comes from the US department of education. The interpretation of two variables is performed on linear regression model and visualized by histogram and scatterplot. After analyzing on different subset and evaluating possible confounding variable, a result that supports the alternative hypothesis is reached, and it shows that there is a positive relationship between the student's college tuition and the income after graduating.

1. Introduction

College education is an important part for both society and individuals, most of teenagers today choose to attend college to improve themselves to get a better job or higher salary. The value of the college education is a good topic to study on. From family aspect, parents need to know if it is worthy to spend more money on children's education, from society aspect, government needs to know if it is worthy to invest more on education system. This research is going to study on "If there is a relationship between the amount of the tuition fee of college and student's level of income after graduating." In other word, for the student who attends a more expensive college, does this student tend to earn more than other students in the future?

Null hypothesis: There is no relationship between the amount of tuition fee of the college and the level of income of students graduated from this college.

Alternative hypothesis: There is a positive relationship between the amount of tuition fee of the college and the level of income of students graduated from this college.

2. Data

The data used for this research is named as "opportunity insights college data", and it comes from U.S. department of education's IPEDS database and the college scorecard from 2000 to 2013. The dataset includes 49 different measures of 2463 colleges in America, such as the average tuition of students, median income of alumni, ranking and selectivity of the college. This report is going to study on two major data: The average annual cost of attendances of colleges in 2000 and the median earnings of students graduated from corresponding college in 2011.

The average annual cost of attendances of colleges in 2000 is a numerical data, measured in dollars. It refers to the average tuition and other costs such as housing and commuting fee per year for each college. For example, a college with the value of 5000 on this data means that it costs a student 5000 dollars to attend this college in year 2000 on average.

the median earnings of students graduated from corresponding college in 2011 is also a numerical data measured in dollars. It refers to the median income of students in 2011 who are

graduated and working. For example, a college with the value of 20000 on this data means that the students graduated from this college earn a median of 20000 dollars in 2013.

Table 1: Summary of variables

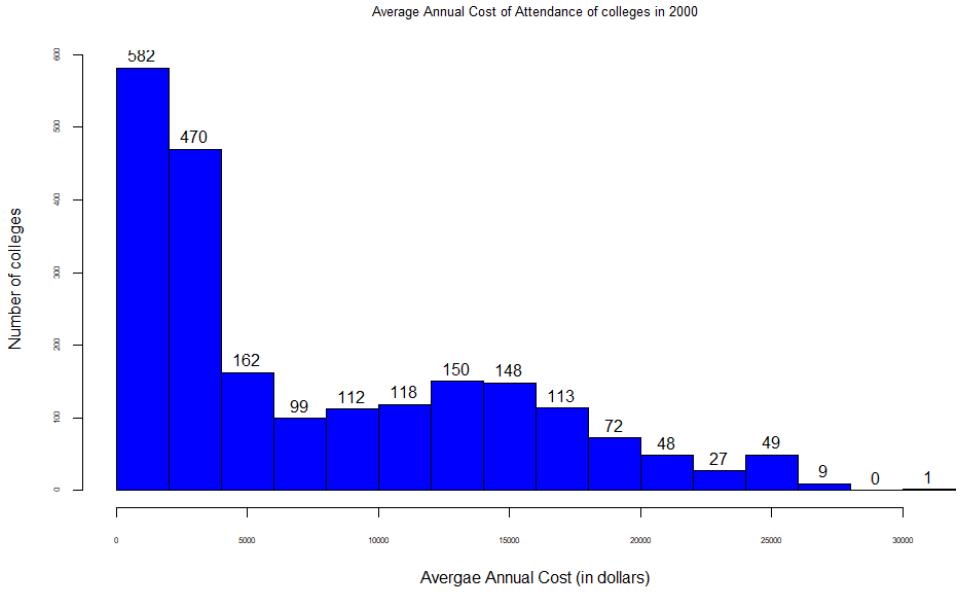
Numerical variables					
Variable name	Count	Min	Mean	Median	Max
Average annual tuition 2000		0	7596	4218	30500
Median annual earning 2011		13400	34800	36560	116400
Asian student share		0	0.015	0.014	0.79
Black student share		0	0.13	0.06	1
Categorical values					
Barron's selectivity index	1: elite	2: highly selective	3-5: selective	9: special schools	999: non-selective
College's level of degree	1: Four-year	2: Two-year	3: Less than Two-year		

3. Methods and Results

3.1 Basic analysis:

In order to understand the variables better, some descriptive visualizations of variables are performed. The first variable is the annual tuition cost. Histogram below shows the distribution of the Average Annual Cost of Attendance of colleges in 2000. It is a roughly right skewed distribution, with the mean of 7596 and the median of 4218, and most of the colleges have an annual tuition below 5000 in year 2000.

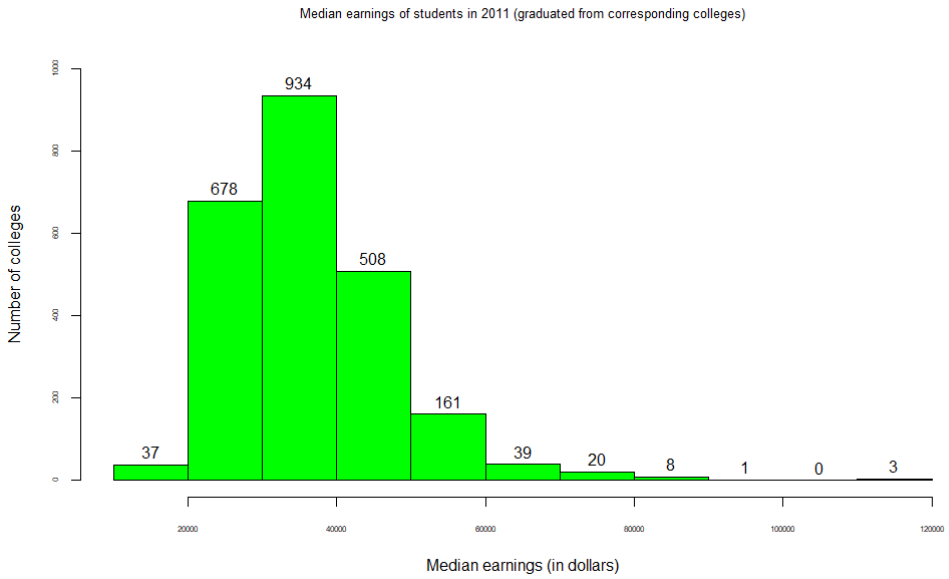
Graph 1: Average Annual Cost of Attendance of colleges in 2000




```
summary(df$sticker_price_2000)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0    1890    4218    7596   12950   30500     303
```

The second variable is the median income after graduating. Histogram below shows the distribution of the median earnings of students in 2011 who graduated from corresponding colleges. It is also a roughly right skewed distribution, with the mean of 36560 and the median of 34800, which means most of the colleges have a median annual income between 20000 and 60000 of their graduated students in year 2013.

Graph 2: Median earnings of students in 2011 (graduated from corresponding colleges)




```
summary(df$scorecard_median_earnings_2011)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
13400   29100   34800   36560   42200  116400      74
```

After study on the two variables separately, the analysis on the relationship between two variables is performed. Below is the result of the linear regression and a scatterplot of two variables.

Graph 3: Scatterplot of the annual average cost of colleges in 2000 and median earning of students in 2011

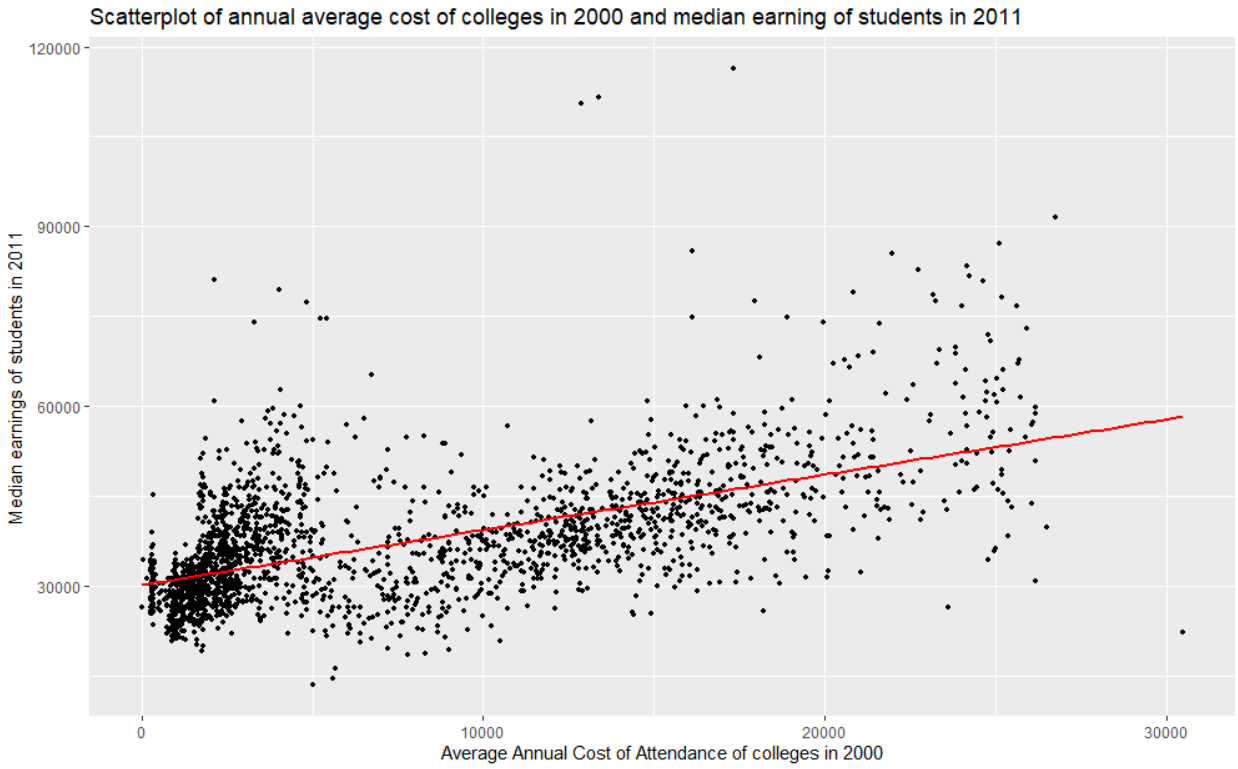


Table 2: Linear regression between the annual average cost of colleges in 2000 and median earning of students in 2011

```
Residuals:
   Min     1Q Median     3Q    Max
-36059  -5092  -1271   3689  70263

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.016e+04  2.804e+02  107.56   <2e-16 ***
sticker_price_2000 9.213e-01  2.717e-02   33.91   <2e-16 ***
```

The result shows that there is a positive correlation between the annual tuition fee and the median income after graduating. The coefficient between two variables is around 0.92, which

means that a 1-dollar increase in the annual tuition fee is expected to increase the median income after graduating by 0.92 dollars. The p value is 2.2e-16 which is very close to 0 and less than 0.05. It means that there are nearly 0% chance to see this result under the null hypothesis, hence the result is very significant.

3.2 Advanced analysis

After performing individual analysis and linear regression on two variables, it seems that there is a significant relationship between the annual tuition fee and the median income after graduating. However, the basic analysis put all universities into one population and measured them as an entity, but there are huge differences between colleges, for example, the difference between 4-year college and 2-year college, highly selective and low selective college, Asian student majority and black student majority college. To improve the preciseness of the study, the following part is going to analyze the relationship between two variables across different groups of colleges.

Group 1: Ethnicity

Asian students

We are interested in if the ethnicity of students in the college will influence the result. The dataset is going to be divided into two subsets: Colleges with relatively large share of Asian students and colleges with relatively large share of black students.

Table 3: Summary table of Asian and pacific student share

asian_or_pacific_share_fall_2000
Min. :0.00000
1st Qu.:0.00671
Median :0.01497
Mean :0.03390
3rd Qu.:0.03518
Max. :0.79159

According to the analysis of the data, the share of Asian or pacific students among colleges in US has a minimum of 0, a maximum of 0.79 and a mean of 0.034. The 3rd quantile value is

0.035, which means that if a college has a share of Asian or pacific students of 0.035, it is greater than 75% of the colleges in US. Therefore, if a college has Asian or pacific students share greater than 0.035, it will be classified in a subset that have a relatively large Asian student share.

Graph 4: Scatterplot of the annual average cost of colleges in 2000 and median earning of students in 2011 among colleges that have relatively larger share of Asian students

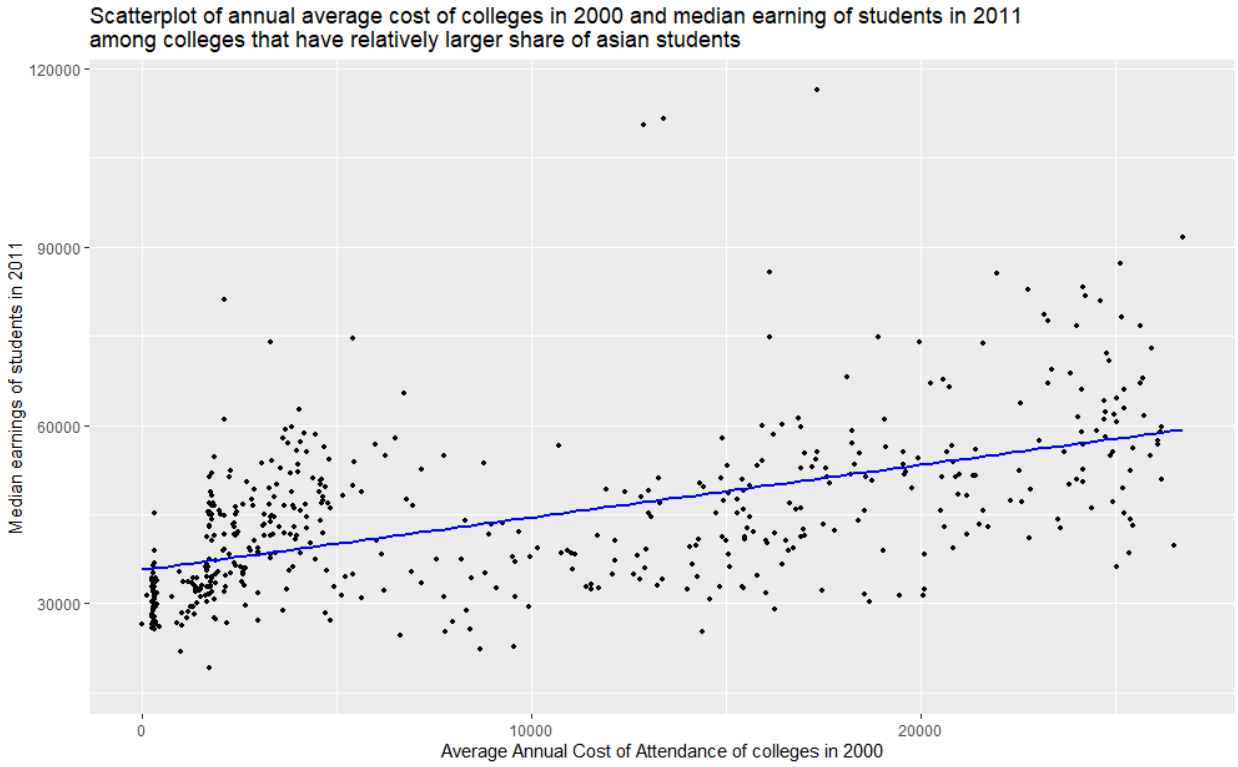


Table 4: Linear regression between annual average cost of colleges in 2000 and median earning of students in 2011 among colleges that have relatively larger share of Asian students

```
Residuals:
   Min     1Q Median     3Q    Max
-23211  -7061  -2053   5908  65389

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.569e+04  7.491e+02   47.64   <2e-16 ***
sticker_price_2000 8.835e-01  5.787e-02   15.27   <2e-16 ***
```

Above are the results of the linear regression and a scatterplot of two variables on the Asian subset. It shows that there is a positive correlation between the annual tuition and the median income after graduating among colleges that have relatively larger share of Asian students. The coefficient between two variables is around 0.88, which means that a 1-dollar increase in the annual tuition fee is expected to increase the median income after graduating by 0.88 dollars. The p value is 2.2e-16 which is very close to 0 and less than 0.05. It implies that the result is very significant.

Black students

Table 5: Summary table of black student share

black_share_fall_2000
Min. :0.00000
1st Qu.:0.02796
Median :0.06364
Mean :0.13037
3rd Qu.:0.15214
Max. :1.00000

According to the analysis of the data, the share of black students among colleges in US has a minimum of 0, a maximum of 1 and a mean of 0.13. The 3rd quantile value is 0.15, which means if a college has a share of black students of 0.15, it is greater than 75% of the colleges in US. Therefore, if a college has Asian or pacific students share greater than 0.15, it will be classified in a subset that have a relatively large Asian student share.

Graph 5: Scatterplot of the annual average cost of colleges in 2000 and median earning of students in 2011 among colleges that have relatively larger share of black students

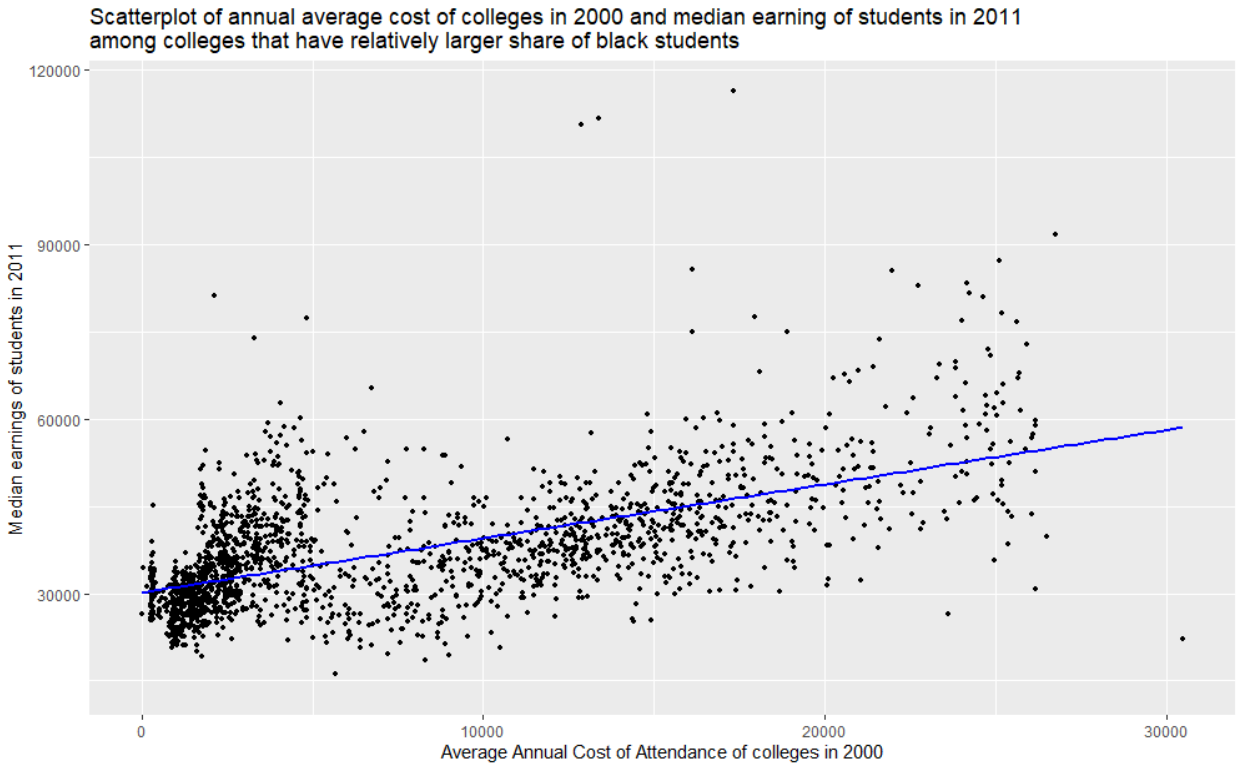


Table 6: Linear regression between annual average cost of colleges in 2000 and median earning of students in 2011 among colleges that have relatively larger share of black students

```
Residuals:
   Min     1Q Median     3Q    Max
-36387  -5112  -1282   3848  70067

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.018e+04  3.042e+02   99.21   <2e-16 ***
sticker_price_2000 9.314e-01  2.913e-02   31.97   <2e-16 ***
```

Above are the results of the linear regression and a scatterplot of two variables on the black subset. It shows that there is a positive correlation between the annual tuition and the median

income after graduating among colleges that have relatively larger share of black students. The coefficient between two variables is around 0.93, which means that a 1-dollar increase in the annual tuition fee is expected to increase the median income after graduating by 0.93 dollars. The p value is 2.2e-16 which is very close to 0 and less than 0.05. It implies that the result is very significant.

Group 2: Type of colleges

After evaluating the composition of the students share, we also want to study on if the result has any difference between different type of colleges. According to the dataset, there are two major types of colleges in U.S. Four-year college and two-year college. There are 1465 four-year colleges and 913 two-year colleges. Since they have different length of time, different types of instruction, and different educational objectives, it is better to analyze them separately and to see if they both reach the same result or there is any difference. The subsets are chosen based on a categorical value called ic level, where a value of 1 refers to a four-year college, and a value of 2 refers to a two-year college.

Graph 6: Scatterplot of the annual average cost of colleges in 2000 and median earning of students in 2011 among four-year colleges

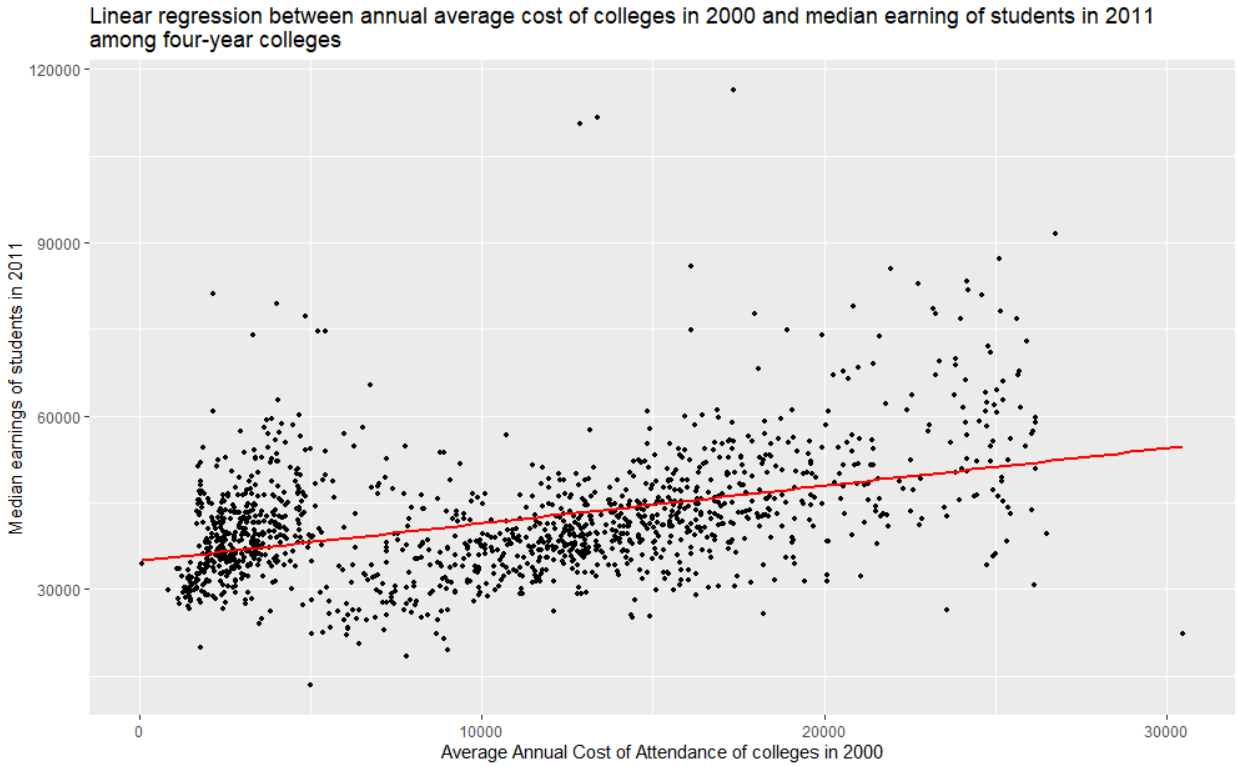


Table 7: Linear regression between annual average cost of colleges in 2000 and median earning of students in 2011 among four-year colleges

```
Residuals:
   Min     1Q Median     3Q    Max
-32582  -5780  -1302   4740  70179

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.494e+04  4.795e+02   72.86   <2e-16 ***
sticker_price_2000 6.507e-01  3.787e-02   17.18   <2e-16 ***
```

Four-year colleges:

Above are the results of the linear regression and a scatterplot of two variables on the four-year colleges subset. It shows that there is a positive correlation between the annual tuition and the median income after graduating among four -year colleges. The coefficient between two

variables is around 0.65, which means that a 1-dollar increase in the annual tuition fee is expected to increase the median income after graduating by 0.65 dollars. The p value is 2.2e-16 which is very close to 0 and less than 0.05. It implies that the result is very significant.

Graph 7: Scatterplot of the annual average cost of colleges in 2000 and median earning of students in 2011 among two-year colleges

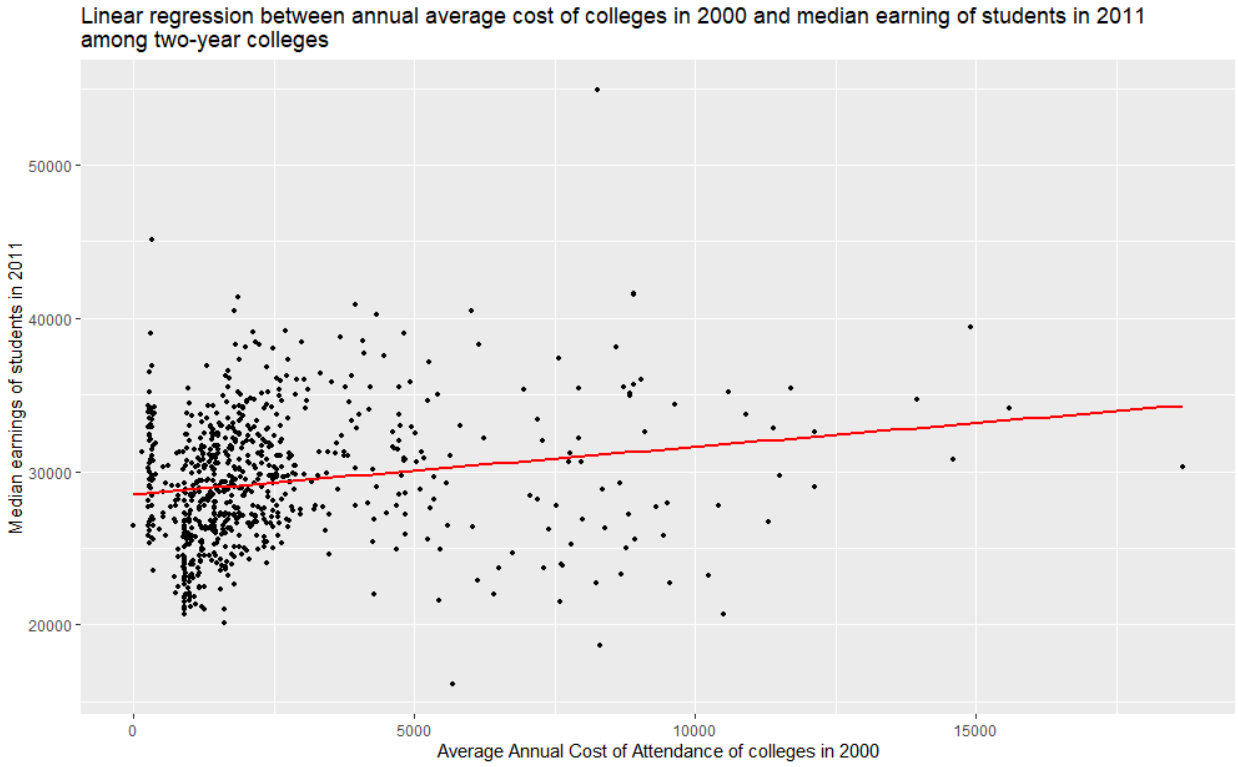


Table 8: Linear regression between annual average cost of colleges in 2000 and median earning of students in 2011 among four-year colleges

```
Residuals:
     Min       1Q   Median       3Q      Max
-14175.7  -2784.0    -48.8   2682.7  23823.0

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.851e+04  2.150e+02  132.59  < 2e-16 ***
sticker_price_2000 3.107e-01  6.189e-02    5.02 6.44e-07 ***
```

Two-year colleges:

Above are the results of the linear regression and a scatterplot of two variables on the two-year colleges subset. It shows that there is a positive correlation between the annual tuition and the median income after graduating among two-year colleges. The coefficient between two variables is around 0.31, which means that a 1-dollar increase in the annual tuition fee is expected to increase the median income after graduating by 0.31 dollars. The p value is 6.4e-7 which is very close to 0 and less than 0.05. It implies that the result is very significant.

Group 3: Selectivity

The third factor is the selectivity of the college. The dataset is going to be divided into two subsets: Highly selective college and non-selective college. The subsets are chosen based on a categorical value called Barron's selectivity index, where a value of 1 means an elite, highly selective school, and a value of 999 means a normal, non-selective school.

Graph 8: Scatterplot of the annual average cost of colleges in 2000 and median earning of students in 2011 among highly selective colleges

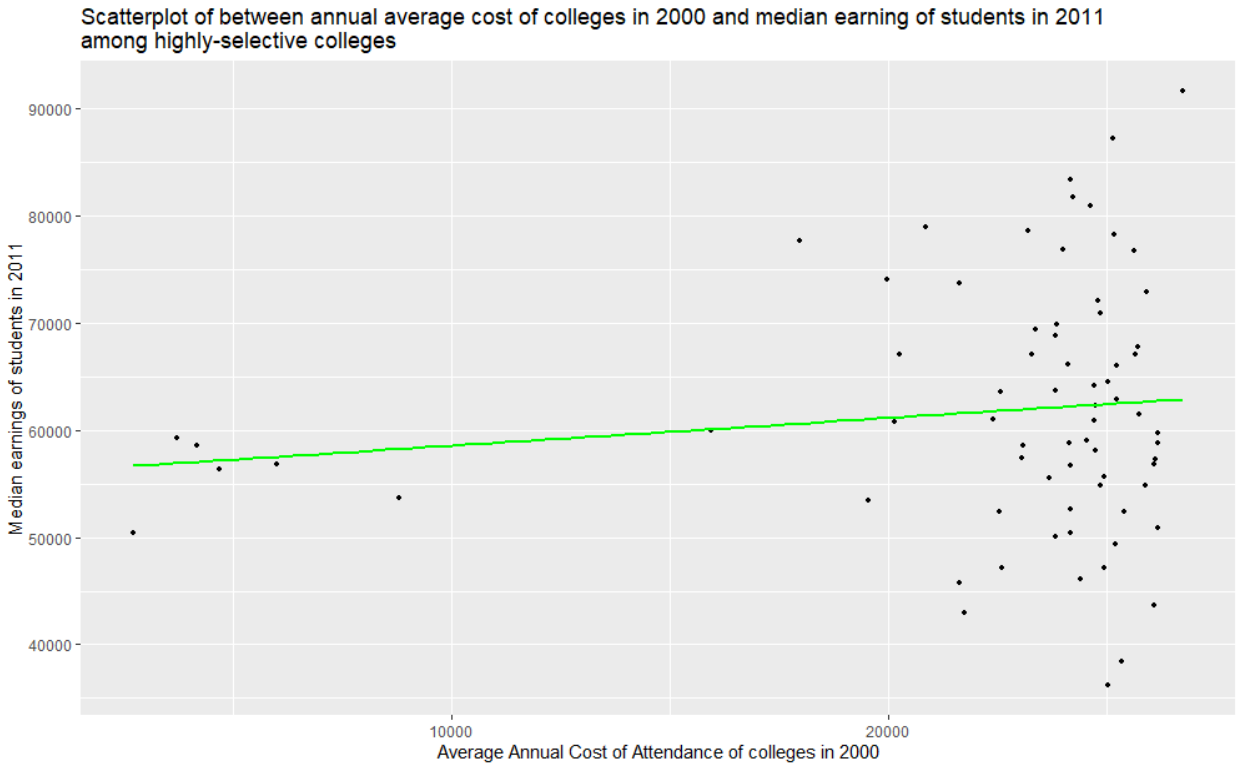


Table 9: Linear regression between annual average cost of colleges in 2000 and median earning of students in 2011 among highly selective colleges

```
Residuals:
     Min       1Q   Median       3Q      Max
-26225.2  -6999.8   -726.5   6861.4  28728.3

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        5.595e+04  5.597e+03   9.996 4.06e-15 ***
sticker_price_2000 2.587e-01  2.428e-01   1.065     0.29
```

Above are the results of the linear regression and a scatterplot of two variables on the Asian subset. It shows that there is a very small positive relationship between the annual tuition and the

median income after graduating among colleges that have relatively larger share of black students. The coefficient between two variables is around 0.26, which means that a 1-dollar increase in the annual tuition fee is expected to increase the median income after graduating by 0.26 dollars. The p value is 0.29 which is greater than 0.05. It implies that the result is not significant.

Graph 9: Scatterplot of the annual average cost of colleges in 2000 and median earning of students in 2011 among non-selective colleges

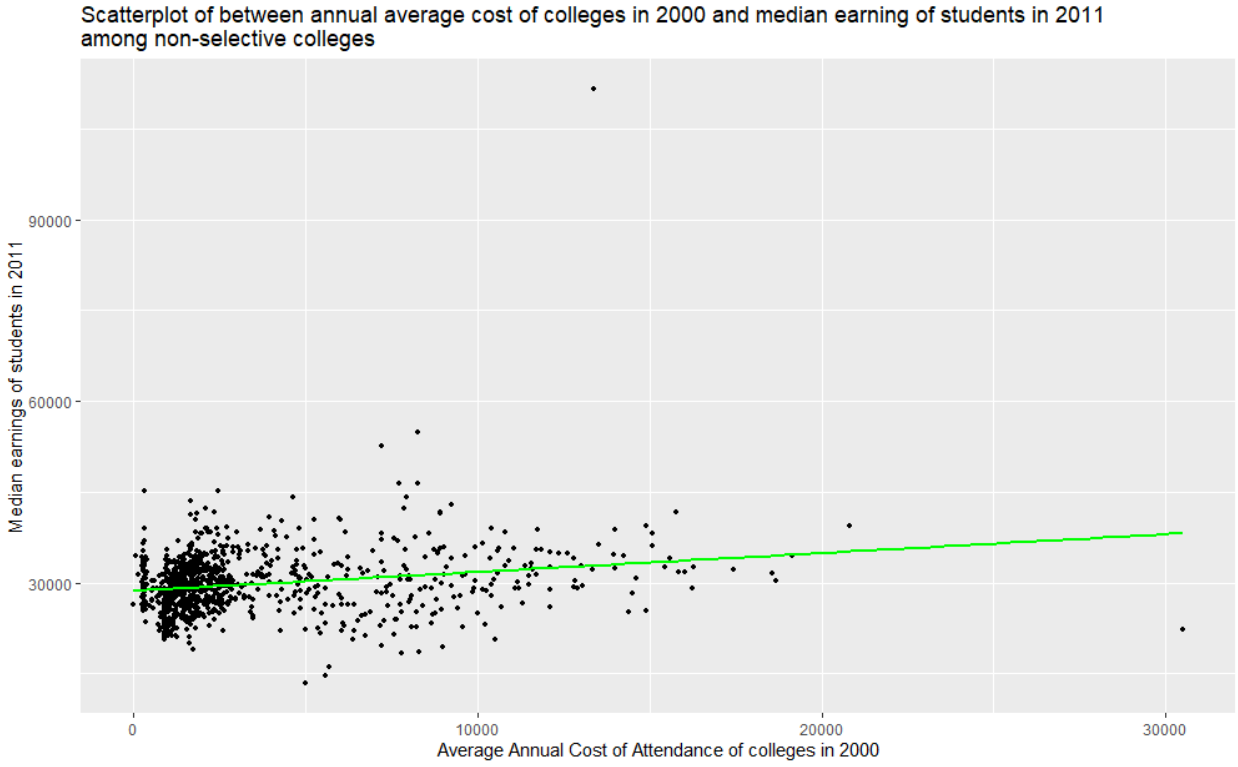


Table 10: Linear regression between annual average cost of colleges in 2000 and median earning of students in 2011 among non-selective colleges

```
Residuals:
   Min     1Q Median     3Q    Max
-16862  -2942   -168   2638  78723

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.870e+04  2.314e+02 124.024  < 2e-16 ***
sticker_price_2000 3.115e-01  4.636e-02   6.718 3.15e-11 ***
```

Above are the results of the linear regression and a scatterplot of two variables on the Asian subset. It shows that there is a positive linear relationship between the annual tuition and the median income after graduating among colleges that have relatively larger share of black students. The coefficient between two variables is around 0.31, which means that a 1-dollar increase in the annual tuition fee is expected to increase the median income after graduating by 0.31 dollars. The p value is 3.15e-16 which is very close to 0 and less than 0.05. It implies that the result is very significant.

3.3 Confounding variables

As stated above, the estimated coefficient in the linear regression model of the selective college subset is not very significant, and the p value is 0.29, which is much greater than the significant level 0.05, hence we want to explore that if the selectivity of colleges is a confounding variable of the two variables. In other word, are the highly selective colleges tend to have higher tuition and higher income? To pursue this study, the following analysis are made.

Table 11: Linear regression between median earning of students in 2011 and selectivity of colleges

```
Residuals:
     Min       1Q   Median       3Q      Max
-12704.7  -6340.0    600.4   5660.1  21503.1

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)               7755.6      413.7  18.749   <2e-16 ***
scorecard_rej_rate_2013   9174.7     1022.2   8.975   <2e-16 ***
```

Above is the result of the linear regression of median income as Y variable and selectivity of colleges as X variable. It shows that there is a positive relationship between the selectivity of colleges and the median income after graduating. The coefficient between two variables is around 9175, which means that a 1 percent increase in the annual tuition fee is expected to increase the median income after graduating by 91.75 dollars. The p value is 2.2e-16 which is very close to 0 and less than 0.05. It implies that the result is very significant.

Table 12: Linear regression between annual average cost of colleges in 2000 and selectivity of colleges

```
Residuals:
   Min     1Q Median     3Q    Max
-28471  -5936   -762   5068  78245

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              36262.9      584.7   62.02   <2e-16 ***
scorecard_rej_rate_2013  17539.5     1442.2   12.16   <2e-16 ***
```

Above is the result of the linear regression of annual tuition as Y variable and selectivity of colleges as X variable. It shows that there is a positive relationship between the selectivity of

colleges and the median income after graduating. The coefficient between two variables is around 17539, which means that a 1 percent increase in the annual tuition fee is expected to increase the median income after graduating by 175.39 dollars. The p value is 2.2e-16 which is very close to 0 and less than 0.05. It implies that the result is very significant.

Discussion

The analysis on the entire dataset shows that there is a significant positive relationship between the amount of tuition of the college and the level of income of students graduated from this college.

Results from three subsets strengthen the hypothesis, that is, this relationship also holds for different groups of colleges. The coefficient of the varies egression varies in different groups. The coefficient in Asian majority colleges is 0.88, and in black majority colleges is 0.91, the close distance implies that the relationship does not differs a lot across different ethnicities. The coefficient in two-year colleges is 0.31, and in four-year colleges is 0.65, it is almost doubled, which implies that the cost-income relationship differs in different type of colleges. In other word, a unit increase of tuition in four-year college tends to increase student's future income more than in two-year colleges. When evaluating the result in selectivity subset, we find that the result is significant in non-selective colleges but not in highly selective colleges, and this drives us to make an investigation on the possible confounding variable.

The analysis on the last part shows that the selectivity of colleges is a confounding variable that influences both the tuition and future income because it has a high positive estimated coefficient (9174 and 17539) on both two variables, and a small p value infers that it is a significant result.

The possible reason of the result might be that a high-quality college education may needs a higher cost hence a higher tuition, and good education results in a better chance of work hence a higher future income. The one surprising phenomenon is that this relationship holds for non-selective colleges but not very evident for highly selective colleges, because at the beginning we believed that the correlation should be strong since students in highly selective colleges tend to have better economic conditions and better chances in finding the job. One of the biggest limitations of this study is lacking cross-time analysis, it would be better if we can establish a

crossing analysis on this relationship in different years, which will enhance the validity of the result. Another drawback might be the incompleteness of the data, there are hundreds of null values needed to be dropped, and some desirable data such as the industry the student works in and the average employment rate among students are not included in the dataset, hence if there are any other research might reach a different conclusion, we are happy to discuss and improve our methods and dataset.

Conclusion

The result proves the alternative hypothesis that there is a significant positive relationship between the amount of tuition of the college and the level of income of students graduated from this college, a 1-dollar increase in the annual tuition fee is expected to increase the median income after graduating by 0.92 dollars, the coefficient varies in different subsets but the correlation is valid in all groups except the elite colleges group. Selectivity of the college is a confounding variable that will cause both tuition and income to change. For the future potential studies, I suggest researchers to explore on the magnitude of the income change and which factor is the major reason that influences the income.

References

IPEDS. “Overview of IPEDS Data.” *The Integrated Postsecondary Education Data System*, 2021, nces.ed.gov/ipeds/use-the-data/overview-of-ipeds-data.