Research on Average Post-school Earning of a College or University

Principal Investigator: Shuanger Chen (schen586@wisc.edu)

Introduction

When we are choosing which college or university to attend, how much students can earn after graduation from a specific institution is a really important information under consideration. In this project, I study what factors affects average post-school earning of a college or university in the US, and how they affect the after-school earning.

In the first part of my study, by plotting different variables against post-school earning, admission rate and percentage of low-income family have negative correlation with post-school earning, and composite ACT scores, percentage of high-income family of a school have positive correlation. Averaging the post-school earning by states, we can see that during academic year of 2014-2015, Washington DC, Massachusetts and New York have the highest post-school earning, while Idaho, Arkansas, Mississippi have lowest post-school earning (income for abbreviation). And private nonprofit school have the highest mean income, follows by public school.

I also conduct OLS regression on selected variables against post-school earning and log post-school earning. The sign of these two regressions' coefficients have slightly difference. The admission rate has oddly positive coefficient on the log regression, and the original one does not. Another unexpected thing is that proportion of second highest income family have negative coefficient in both regressions. Other than that, coefficients are normal. Composite ACT score, cost of attendance, proportion of male students, proportions of Asian and Hispanic students have positive coefficients. Percentages of white and black students, completion rate in 100% amount of time, percentage of low-income family and middle-income family have negative correlation with post-school earning.

Analysis

Data:

The data is college scorecard data from academic year of 2014 to 2015. The source is US department of education. So, all the statement I make in this paper will be describing the situation during 2014-2015. And I keep the data that is useful, shown in Table 1.

Table 1: Data Description

NAMES IN DATA	ACTUALL MEANING
UNITID	Institution ID
INSTNM	Institution name
STABBR	State abbreviation
ICLEVEL	High level of reward school offered
CONROL	School type
ADM_RATE	Admission rate
SATVRMID, SATMTMID, SATWRMID	Mid-points of SAT reading, math, writing scores
ACTCM25, ACTCM75	25th, 75th percentiles of composite ACT scores
COSTT4_A	Average annual cost of attendance

UGDS	Number of student enrolled in fall semester
UGDS_WOMEN, UGDS_MEN	Gender ratios
UGDS_*(WHITE, BLACK, HISP, ASIAN, AIAN, NRA)	Race ratios

INC_PCT_*(LO, M1, M2, H1, H2)	Percentages of family income
	(INC_PCT_LO= $0-$30,000;
	INC_PCT_M1 = $30,001-$48,000;
	INC_PCT_M2 = $48,001-$75,000;
	INC_PCT_H1 = $75,001-$110,000;
	INC_PCT_H2 = $110,001+)

C100_4, C100_L4	Completion rates of 100% expected school time
MN_EARN_WNE_P10, MN_EARN_WNE_P6	Mean earning after first enrolled in school for 10 years and 6 years
MD_EARN_WNE_P10, MD_EARN_WNE_P6	Median earning after first enrolled in school for 10 years and 6 years
NOTE: DATA IS FROM 2014-2015	

For simplicity, I generally use mean earning after first enrolled in school for 10 years (MN_EARN_P10) as the post-school income. And since ratio of genders and races would add up to 1, which would cause collinearity problem, I manually pick ratio of men (UGDS_MEN), white, black, Asian and AIAN to place in the regression later. Because ACT score has composite one, while SAT does not, I use the 25th percentile of ACT scores as a measurement of student's ability.

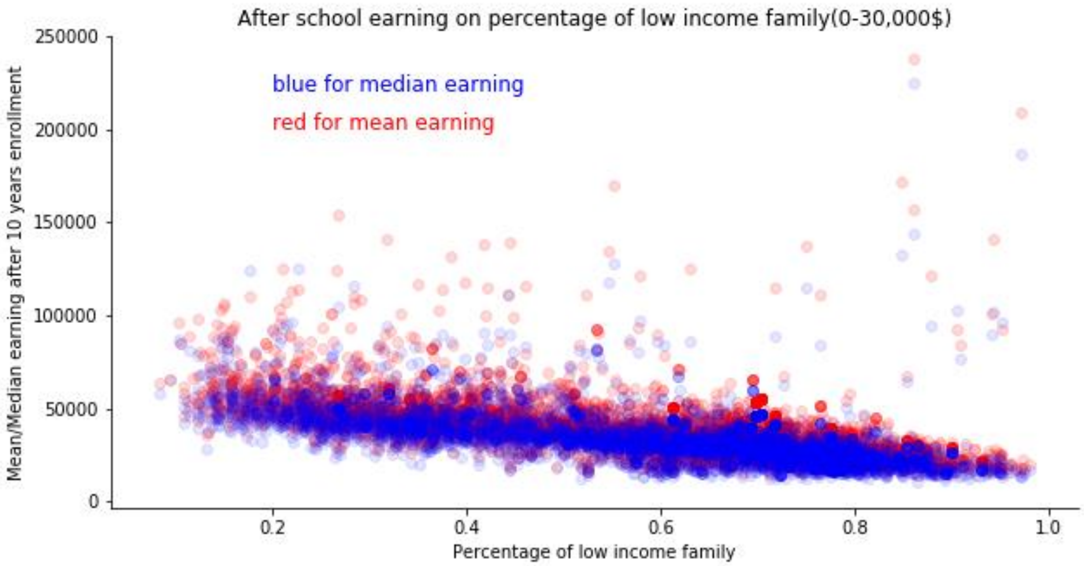
From the data, I find out that during the lowest post-school states are Idaho, Arkansas and Mississippi. The highest income states are Washington DC, Massachusetts and New York, where have either very good school or high-income population. Wisconsin ranks 11 over the whole country, which is happy to find out.

Methods and Results

Plotting

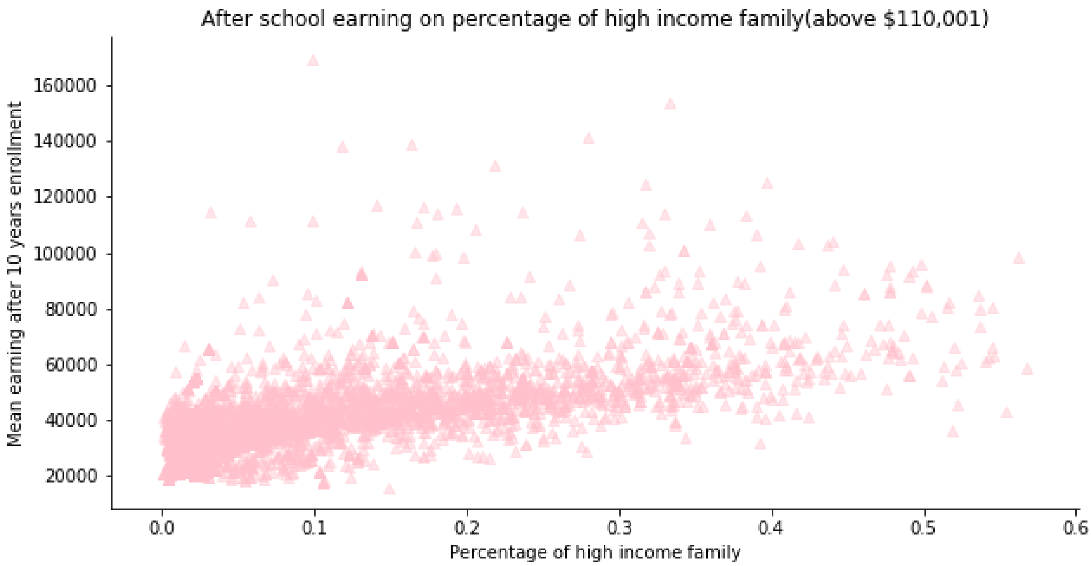
To better understand how the data distributes and getting a clear picture. I use matplotlib and seaborn to depict the data. I first plot the percentage of low-income family against mean post-school earning and median post-school earning. As seen in Figure 1, the mean is slightly higher than the median income. I interpret it as schools may have some super rich alumni that push up the mean. The negative correlation between percentage of low-income family with post-school earning is obvious.

Figure 1



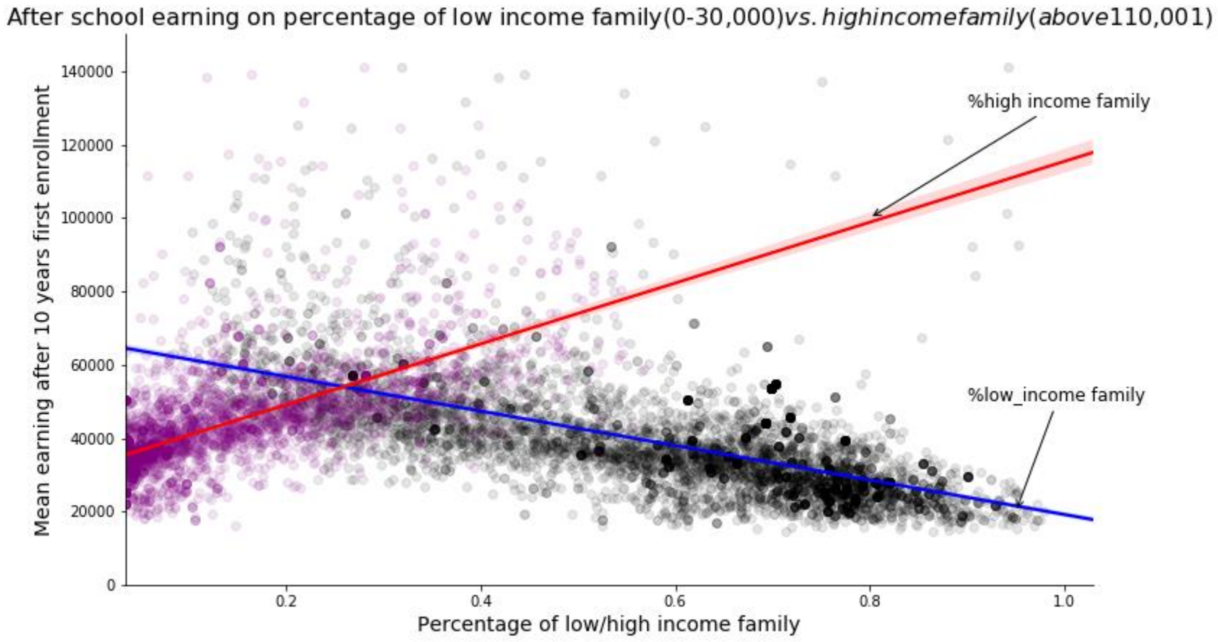
In contrast, as shown in Figure 2, percentage of high-income family has positive correlation with after-school income.

Figure 2



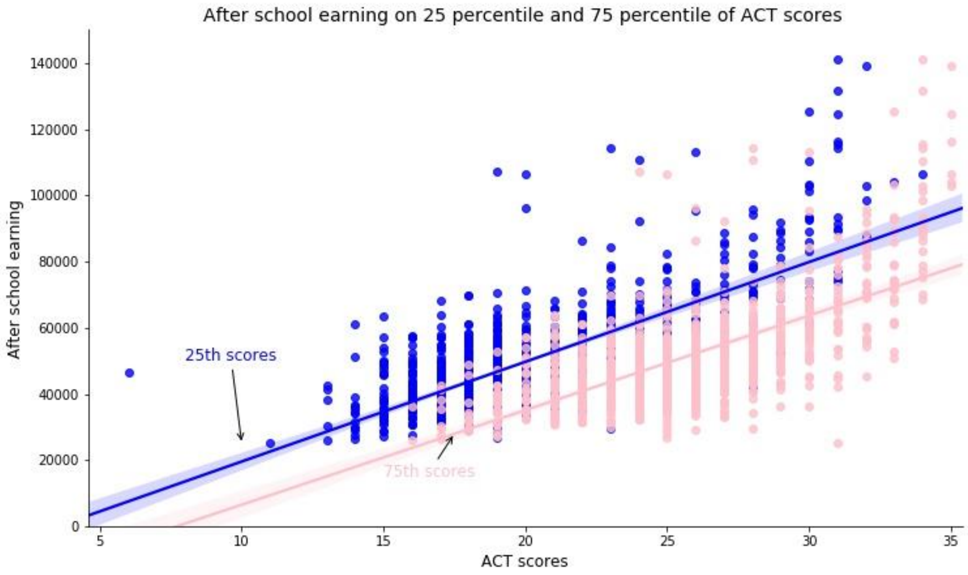
I also use seaborn to construct a regression on the Figure 3, for percentage of low-income family and high-income family against post-school earning. The 95% confidence interval are narrow in the figure.

Figure 3



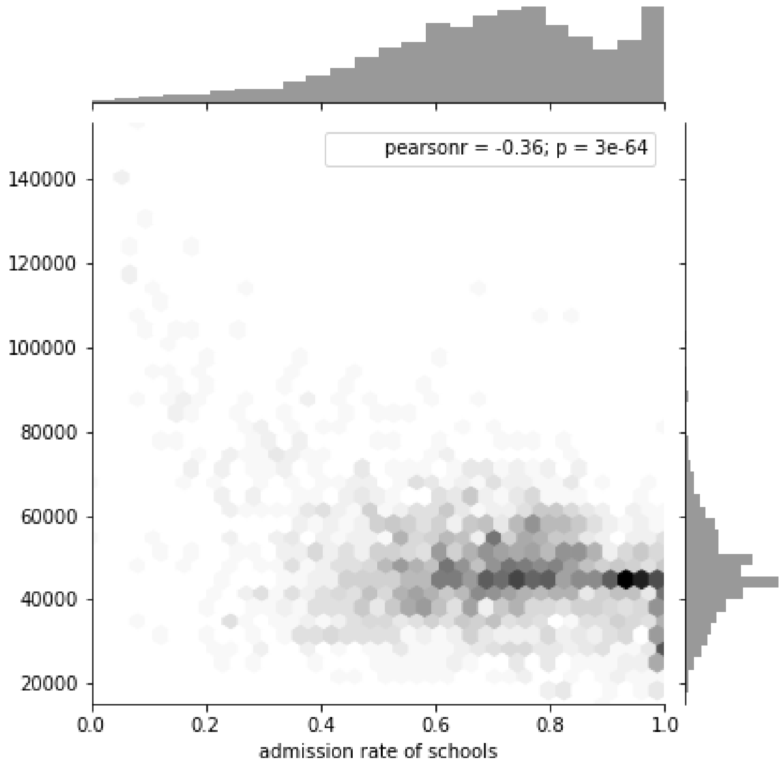
Out of curiosity of the impact of ACT scores, I plot it against student income in Figure 4. The 75th percentile regression is below the 25th one. It matches the fact that for the same score, if that is the score for 25th percentile of school A, and for 75th percentile for school B, then school A's students have higher score than B. They are likely to make more money in the future. Also, the slope of these two line are almost the same, so I decide to use just one of them(25th percentile) in my regression.

Figure 4



Jointplot is a very useful tool in sketching data. Therefore, I use it to depict the density of admission rate and post-school earning. The correlation is relatively fuzzy compared to those shown above. Result shown in Figure 5.

Figure 5

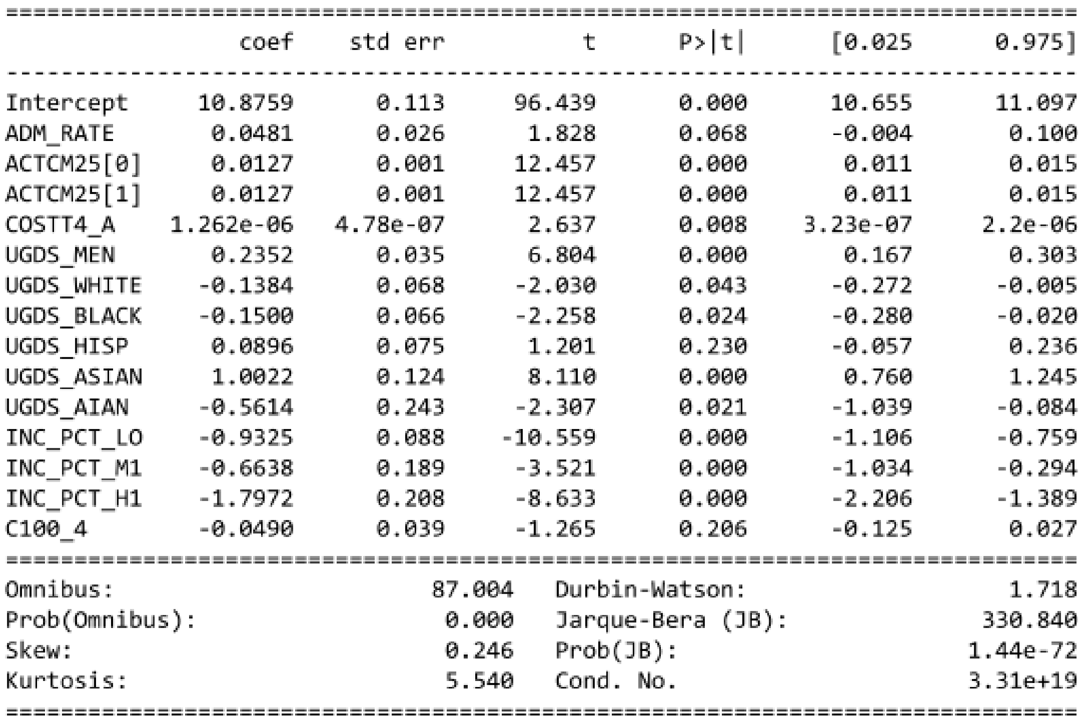


Regression

Since the income is quite a big number, log(post-school earning) might be a good try at first. Table 2 is the result of that. There are some unusual things appearing in the coefficient table. Unlike I expected, the admission rate completion rate, and percentage of high-income family have positive impact on the post-school earning. And an interesting thing is that the more white people school has, the lower the post-school income. Others are pretty I expected. The R square is 0.7. It indicates the independent variables explain the dependent variable well.

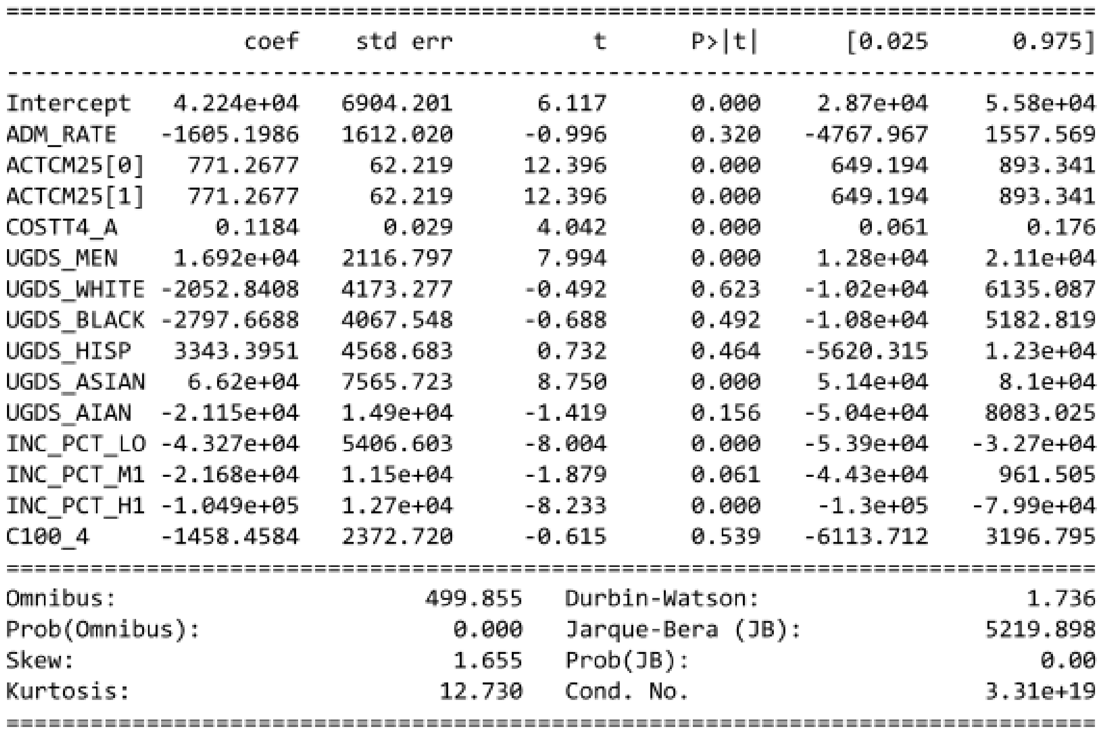
The result in the first regression on log(post-school earning) seems not working out. So, I just regress the post-school earning without log to see what will happen. The summary of the second regression is shown in Table 3. This time, the admission rate have negative impact on the post-school earning, which fits the fact that the lower the admission rate, the more power for schools to choose students. Therefore, student with high capability would enter school, and the can earn more after school. However, the percentage of high-family still has negative coefficient. Even though the r square drops a little bit to 0.685, I consider it a better model than the log one for the reason that it is closer to the reality.

Table 2: regression on log(post-school earning) against other covariates



```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     10.8759      0.113     96.439      0.000      10.655      11.097
ADM_RATE       0.0481      0.026      1.828      0.068      -0.004       0.100
ACTCM25[0]     0.0127      0.001     12.457      0.000       0.011       0.015
ACTCM25[1]     0.0127      0.001     12.457      0.000       0.011       0.015
COSTT4_A    1.262e-06   4.78e-07      2.637      0.008    3.23e-07     2.2e-06
UGDS_MEN       0.2352      0.035      6.804      0.000       0.167       0.303
UGDS_WHITE    -0.1384      0.068     -2.030      0.043      -0.272      -0.005
UGDS_BLACK    -0.1500      0.066     -2.258      0.024      -0.280      -0.020
UGDS_HISP      0.0896      0.075      1.201      0.230      -0.057       0.236
UGDS_ASIAN     1.0022      0.124      8.110      0.000       0.760       1.245
UGDS_AIAN     -0.5614      0.243     -2.307      0.021      -1.039      -0.084
INC_PCT_LO    -0.9325      0.088    -10.559      0.000      -1.106      -0.759
INC_PCT_M1    -0.6638      0.189     -3.521      0.000      -1.034      -0.294
INC_PCT_H1    -1.7972      0.208     -8.633      0.000      -2.206      -1.389
C100_4        -0.0490      0.039     -1.265      0.206      -0.125       0.027
==============================================================================
Omnibus:                       87.004   Durbin-Watson:                   1.718
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              330.840
Skew:                           0.246   Prob(JB):                     1.44e-72
Kurtosis:                       5.540   Cond. No.                     3.31e+19
==============================================================================
```

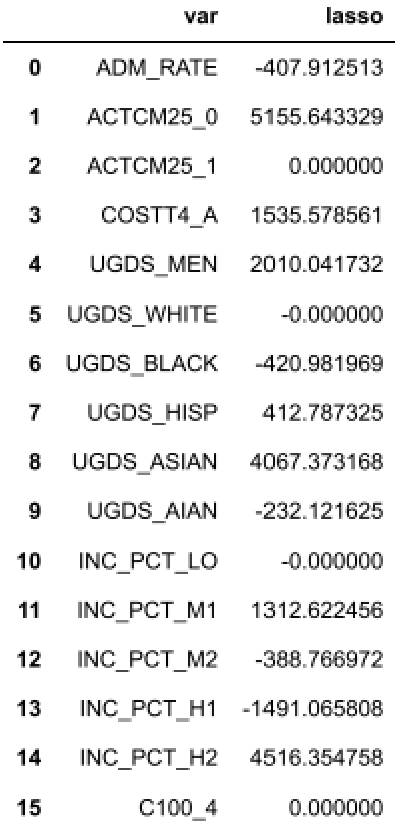
Table 3: Regression on Post-school Earning



```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   4.224e+04   6904.201      6.117      0.000    2.87e+04    5.58e+04
ADM_RATE   -1605.1986   1612.020     -0.996      0.320   -4767.967    1557.569
ACTCM25[0]   771.2677     62.219     12.396      0.000     649.194     893.341
ACTCM25[1]   771.2677     62.219     12.396      0.000     649.194     893.341
COSTT4_A       0.1184      0.029      4.042      0.000       0.061       0.176
UGDS_MEN    1.692e+04   2116.797      7.994      0.000    1.28e+04    2.11e+04
UGDS_WHITE -2052.8408   4173.277     -0.492      0.623   -1.02e+04    6135.087
UGDS_BLACK -2797.6688   4067.548     -0.688      0.492   -1.08e+04    5182.819
UGDS_HISP   3343.3951   4568.683      0.732      0.464   -5620.315    1.23e+04
UGDS_ASIAN   6.62e+04   7565.723      8.750      0.000    5.14e+04     8.1e+04
UGDS_AIAN  -2.115e+04   1.49e+04     -1.419      0.156   -5.04e+04    8083.025
INC_PCT_LO -4.327e+04   5406.603     -8.004      0.000   -5.39e+04   -3.27e+04
INC_PCT_M1 -2.168e+04   1.15e+04     -1.879      0.061   -4.43e+04     961.505
INC_PCT_H1 -1.049e+05   1.27e+04     -8.233      0.000    -1.3e+05   -7.99e+04
C100_4     -1458.4584   2372.720     -0.615      0.539   -6113.712    3196.795
==============================================================================
Omnibus:                      499.855   Durbin-Watson:                   1.736
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             5219.898
Skew:                           1.655   Prob(JB):                         0.00
Kurtosis:                      12.730   Cond. No.                     3.31e+19
==============================================================================
```

The result of the OLS regression above is not as good as I expected. I decide to perform a lasso regression through 10-fold cross validation to find out what variables are actually useful in predicting the post-school income. The best α lies on 81.06. Percentages of white people and low-income family are ruled out from the model along with completion rate. And percentage of highest income family has positive coefficient, eventually. But the percentage of second highest income family has negative coefficient. That might because the collinearity. Same thing happens in the percentage of middle income family. This is the best model I considered for now. The coefficient is shown below in Table 4.

Table 4: Coefficients of Lasso



	var	lasso
0	ADM_RATE	-407.912513
1	ACTCM25_0	5155.643329
2	ACTCM25_1	0.000000
3	COSTT4_A	1535.578561
4	UGDS_MEN	2010.041732
5	UGDS_WHITE	-0.000000
6	UGDS_BLACK	-420.981969
7	UGDS_HISP	412.787325
8	UGDS_ASIAN	4067.373168
9	UGDS_AIAN	-232.121625
10	INC_PCT_LO	-0.000000
11	INC_PCT_M1	1312.622456
12	INC_PCT_M2	-388.766972
13	INC_PCT_H1	-1491.065808
14	INC_PCT_H2	4516.354758
15	C100_4	0.000000

Conclusions and directions for future research

From the study above, the higher admission score, more Asian students, more male students, and higher cost, the more students would earn after graduation. Admission rate, proportions of white and black students have negative correlation with earning after graduation. These all somehow fit the reality.

The difficulty I encounter is that the negative coefficient on percentage of high-income family. It would be a good direction for my later study.