Re-examining the Link Between College Tuition and Average Income of Recent Graduates

Group #12

Group members: Jack MacIsaac, Momoka Oguri, Seve Yuen

Abstract

Looking at whether or not to attend college can be a difficult decision for some. In this paper we attempt to recreate the research done in the paper "A Regression Analysis of College Tuition and Mean Income," and investigate whether cost of tuition influences the average income seven years after enrollment. We will also look at different variables that could help to explain our dependent variable better than the model in the original paper, such as university expenditures per student, whether the school is private or public, average family income of students, and racial demographics of the student population, among others. To examine this link, we undertook several single and multiple regression models looking at the statistical significance of the regressors as well as the R^2 values of the specified models to see how well the regressors and models explain our dependent variable, average income seven years after enrollment. We looked at data from the U.S. Department of Education's College Scorecard program from 2018. Looking at the results from our replicative research, we can see that we found that tuition cost slightly influences mean income and should be considered. We also found that the things prospective students should consider when deciding whether to go to university are tuition cost, whether a university is public or private, how much the university spends per student, and their own family's income.

Introduction

As we know, college tuition has been increasing by quite a lot over the last couple decades. This fact, as well as the idea that most people view a post-secondary institution as a necessity in order to get a good paying job in the future. This paper aims to re-investigate the link between the rising costs to attend post-secondary institutions and the average mean income that graduates make over a seven-year period following graduation. We are conducting this replication research in order to determine whether or not investing in a post-secondary education is still a good investment for people. We will examine this link using the same methodology as Johnson, Lax, and Mitnick (2016) in their paper "A Regression Analysis of College Tuition and Mean Income." We hypothesize that due to an increasing cost of tuition, that there will not be a very strong link between cost to attend a post-secondary institution and the average income that recent graduates make. We will also look at other possible variables that could explain our dependent variable, average income of recent graduates over a seven-year period and see whether or not there are more significant underlying factors that influence it than the cost of tuition. While in our replicative research we did find that cost of tuition does have a small effect on mean income as we hypothesized it would. In our replicative research we also find that the original papers final model that included all significant variables is not very effective at explaining mean income in our dataset. So, with this information, we undertook several regression models in order to find the significant regressors in our dataset. We find that with our new model, the dependent variable can be explained more effectively when including the significant regressors in our model. In this paper, we will go into detail about the literature our paper is based on, and then we will discuss the methodology with which we undertook in order to find our key results. We will then get into where we got our data from and some of the key points about it, and then look at our key findings from our replicative research.

Literature Review

In our assigned paper "A Regression Analysis of College Tuition and Mean Income," Johnson, Lax, and Mitnick (2016) explore the relationship between the cost of post-secondary education and the effect it has on graduates' earnings over the next seven years. They hypothesize that old assumptions that there would be a positive correlation between the cost of an education and the eventual earnings of a graduate are no longer valid, and that there is no longer a strong relationship between these two variables. The authors conducted regression on several hypothesized models ranging from a simple regression model that only included one explanatory variable of average cost of attending a university, to models with many more explanatory variables, such as average SAT score or average family income, to find out whether the cost of attending a post-secondary institution influences graduates' earnings. After conducting their analysis Johnson, Lax, and Mitnick concluded that, "the cost of an institution appears to have little impact on the eventual earnings that students receive after graduating from the institution." (p. 17), which was the outcome that they had hypothesized to be correct.

In looking at whether cost to attend a post-secondary institution has any influence on earnings, another question arises, is the return on your investment at elite, private institutions good enough to justify the increasing costs to attend. Fox (1993) attempts to examine how the rate of return on a person's investment into a college education is affected by the rising costs of attending university. Fox discusses how the falling rate of return on education from elite institutions may induce people to decide to attend cheaper post-secondary institutions in the future. He examines this by first estimating a standard human capital model, then estimating earning for different groups in order to compare them later. From this, the author concludes that the increasing cost difference between public and private post-secondary institutions is leading to the rate of return on a student's investment is decreasing. This paper was written in 1993, so Fox also mentions that while at the time the rate of return is comparable to the quality of the education, that in the future, attending these elite private universities would no longer be a good investment.

The estimation of the rate of return on investment in human capital has been a popular topic for researchers. During the 1960s, the researchers mostly used the internal rate of return (IRR) approach and shifted to an earning function approach. In "Comparing Two Approaches to the Rate of Return to Investment in Education", Kara (2008) compares the pattern of results derived from these two methods with data from Turkey. The IRR approach equates the interest rates to the present values of benefits and costs of education, which takes into account the opportunity cost. The earning function approach takes the log of earnings as the dependent variable and a function of schooling and years of labor market experience. While the IRR method indicates a declining marginal rate of return as schooling increases, the earning function approach had the opposite pattern. Kara concluded that the IRR approach better estimates the return relative to the actual school-related cost, and the earning function approach better measures "earnings differences associated with additional schooling" (p.163). With the raising awareness of females in the work environment, the observation that being married for women decreased the labor participation may not be adaptable to the current situation. The paper using data from Turkey also limits its application to other countries.

Models and Estimation Methods

We use multiple regression models to examine the return on investment of college with the mean income as the dependent variable. In addition to the tuition, the regressors include demographic and socioeconomic factors, such as race, categorized by white, black, Hispanic and Asian, and family income, as mentioned in reviews above to improve the model. The regressors also include loans and length of college attendance as they describe key features relating to the costs of attending college.

The method is to generate three multiple regression models and compare the values of adjusted R^2 to determine which model best describes the variations in the mean income. The first model includes all the variables we listed that have potential influences on the mean income. Then, the second model is built without independently insignificant variables based on the p-values of each coefficients from the first model. Finally, the third model adds back jointly significant variables that were removed in the second model.

To test the joint significance of the race variables, we perform a F-test. The null hypothesis is that the coefficients of those variables are all zero. The conclusion of F-test is made based on the sample value of F statistic and the associated p-value at 5% significance level.

Simple Regression Model:

$$\text{l\_earn} = \beta_0 + \beta_1 lcost + u$$

The simple regression model most directly relates the cost of education and the income.

Multiple Regression Model 1:

$$\text{l\_earn} = \beta_0 + \delta_1 private + \beta_1 lcost + \beta_2 lexp + \beta_3 \text{c150\_4} + \beta_4 \text{rpy\_7yr\_rt} + \beta_5 \text{sat\_avg\_all} + \beta_6 \text{lfaminc} + \beta_7 \text{ugds\_white} + \beta_8 \text{ugds\_black} + \beta_9 \text{ugds\_hisp} + \beta_{10} \text{ugds\_asian} + u$$

The model includes all variables in the list of data that have potential effect on the mean income.

Multiple Regression Model 2:

$$\text{l\_earn} = \beta_0 + \beta_1 lexp + \beta_2 \text{rpy\_7yr\_rt} + \beta_3 \text{sat\_avg\_all} + \beta_4 \text{lfaminc} + u$$

Based on the statistical significance of each coefficient estimates in the result from the multiple regression model 1, the model excludes independently insignificant variables.

Multiple Regression Model 3:

$$\text{l\_earn} = \beta_0 + \beta_1 lexp + \beta_2 \text{rpy\_7yr\_rt} + \beta_3 \text{sat\_avg\_all} + \beta_4 \text{lfaminc} + \beta_5 \text{ugds\_white} + \beta_6 \text{ugds\_black} + \beta_7 \text{ugds\_hisp} + \beta_8 \text{ugds\_asian} + u$$

After the test of significance by F-test, the model adds back in the jointly significant variables that are excluded in the multiple regression model 2.

New Multiple Regression Model 4:

$$\text{l\_earn} = \beta_0 + \delta_1 private + \beta_1 lcost + \beta_2 lexp + \beta_6 \text{lfaminc} + u$$

Following the same method for our dataset, we found different statistical significances for the variables as the original paper, and thus our new final model is to include the four statistically significant variables. As we are going to explain later, we did not find the variables used to describe races to be jointly significant through F-test.

Model Assumptions

The Gustav Markov Assumptions ensures that the ordinary least squares (OLS) estimate is the best linear unbiased estimate (BLUE). The assumptions are linearity of parameters, randomly sampled data from the population, non-collinearity between regressors, exogeneity of regressors with the error term, and homoskedasticity.

1. Linearity

The first assumption of linearity can be expressed algebraically as below:

$$y_t = \beta_0 + \beta_1 X_{t2} + \cdots + \beta_k X_{tk} + u_t \text{ , or}$$

$$y_t = x_t' \beta + u_t$$

Since all the multiple regressions follows the same form as this, the models in this paper satisfy this assumption.

2. Randomly sampled

The original dataset from the source includes almost all universities related to our hypothesis. We only included universities with all the variables required for the regression models, and then selected the sample size to be 1316 in EViews to randomly select samples from the dataset. The sample size of 1316 is the size used in the original paper. The exclusion of institutions without sufficient variables should not lead to any bias as the institutions' decisions on providing specific data is arbitrary.

3. Non-collinearity

The table below lists the correlations between the variables used in our regression models. The cells highlighted in yellow if the correlation is greater than |0.5| and highlighted in red if the correlation is greater than |0.7|. Although there are two strong correlations, the table shows that there is no perfect collinearity between them. Thus, the assumption is satisfied.

	l_earn	Priv	lcost	lexp	C150	Rpy	Sat	lfam	UgdsW	UgdsB	UgdsH	UgdsA
l_earn	1.00											
Priv	-0.34	1.00										
lcost	0.62	-0.82	1.00									
lexp	0.54	-0.18	0.50	1.00								
C150	-0.05	0.03	-0.03	-0.05	1.00							
Rpy	-0.01	0.04	-0.02	0.04	-0.09	1.00						
Sat	0.03	-0.04	0.03	-0.03	-0.01	-0.02	1.00					
lfam	0.66	-0.50	0.76	0.56	-0.04	0.03	0.03	1.00				
UgdsW	0.04	-0.03	0.03	0.01	-0.09	0.02	-0.03	0.03	1.00			
UgdsB	0.003	-0.002	-0.04	-0.01	0.03	0.03	0.01	0.002	-0.02	1.00		
UgdsH	-0.06	0.03	-0.04	-0.05	0.03	-0.05	0.01	-0.08	0.01	0.00	1.00	
UgdsA	0.01	0.03	-0.02	-0.01	-0.01	-0.002	0.03	0.01	0.00	-0.02	0.00	1.00

	Ugds_white	Ugds_black	Ugds_hisp	Ugds_asian
Ugds_white	1.0000			
Ugds_black	-0.0244	1.0000		

Ugds_hisp	-0.0001	-0.0003	1.0000	
Ugds_asian	0.0177	0.0191	0.0098	1.0000

Figure 1: Correlation coefficient Tables

4. Exogeneity / Zero conditional mean

According to the table of correlation above (Figure1), there are overall more positive correlations between the variables included in the multiple regression model. Since the simple regression model do not include those variables, it is likely that the simple regression has a positive bias. Thus, the assumption would not be met for the simple regression model. It is hard to know if we have omitted any relevant variables in the multiple regression models. However, we included a variety of variables that provide different aspects. They should help our models to be unbiased.

5. Homoskedasticity

The graphs below plots the residuals for the simple regression, the three multiple regressions from the paper, and the new model 4 (Appendix A: Figure 2-4). Although there are a few sharp spikes, the overall variance of the residuals are relatively constant, indicating homoskedasticity. Thus, the simple and multiple regression models meet the assumption.

F-test

To test the joint significance the variables describing race, we perform a F-test on Model (1):

$$\text{l\_earn} = \beta_0 + \delta_1 private + \beta_1 lcost + \beta_2 lexp + \beta_3 \text{c150\_4} + \beta_4 \text{rpy\_7yr\_rt} + \beta_5 \text{sat\_avg\_all} + \beta_6 \text{lfaminc} + \beta_7 \text{ugds\_white} + \beta_8 \text{ugds\_black} + \beta_9 \text{ugds\_hisp} + \beta_{10} \text{ugds\_asian} + u$$

The null hypothesis is that the coefficients of the variables describing race, $\beta_7, \beta_8, \beta_9, and\ \beta_{10}$ are all zero. The alternative hypothesis is that at least one of them is not zero.

$$H_0: \beta_7 = \beta_8 = \beta_9 = \beta_{10} = 0$$

$$vs.$$

$$H_1: not\ H_0 (at\ least\ one\ \beta_i \neq 0, i = 7, \ldots, 10)$$

We then find the sample value of F-statistical and the associated p-value. The conclusion is drawn based on the p-value at the 5% significance level. We reject the null hypothesis in favor of the alternative hypothesis, if the p-value is less than 0.05.

See results of this test outlined in Appendix B.

Data

The variables in this paper are used to investigate the mean income students earn 7 years after enrollment throughout the 1316 universities and colleges in America. These institutions consist of colleges, private and public universities; however, only four-year programs are offered. After analyzing the average cost of a student attending an institution in the simple regression it was evident that lcost was significant; however, it was not enough to clearly describe the true model as its R^2 came out quite low. This is likely due to lack of explanatory variables therefore, to analyze further the next models will include more variables used to model mean income.

Figure 5 below will define the variables used in the four models depicting the mean income. The most important variables included to depict the mean income of students after graduating would be the mean income of students', the percentage of students that obtain a degree within 6 years, the percentage of students with debt, their SAT scores, and whether the institution is private or public. These specific variables influence the mean income of students' because they target the student's capabilities which also shows value of a specific institution. The remaining variables based on family and race will be used to determine if they provide similar influences on a student's wealthy success.

Variable Name	Description	Unit
l_earn	Mean income of students 7 years after enrollment	Log of US Dollars
private	Variable representing whether a university is private or public	Boolean variable (= 0 or 1)
lcost	Average total cost of attendance	Log of US Dollars
lexp	Expenditures per student by the university	Log of US Dollars
c_150_4	Percentage of students that earned a degree within 4-6 yrs	Percentage

rpy_7yr_rt	Percentage of students that have not defaulted on, have paid off, or are paying off student loans 7 years after enrolling	Percentage
sat_avg_all	Average SAT score of accepted students	SAT Score
lfaminc	Average family income of students	Log of US Dollars
udgs_white	Percentage of students at each university who are White	Percentage
udgs_black	Percentage of students at each university who are Black	Percentage
udgs_hisp	Percentage of students at each university who are Hispanic	Percentage
udgs_asian	Percentage of students at each university who are Asian	Percentage

Figure 5: Descriptions of Variables included in this paper

This paper's data was obtained from the U.S. Department of Education's College Scorecard program (https://collegescorecard.ed.gov/data/). The data used will represent the school year of 2018.

Summary statistics for each variable can be found in Appendix C.

Empirical Results

Simple Regression

The simple regression is used to see if the log of the student's attendance has clear influence on the log of the mean income students have 7 years after graduating. From EViews, we obtained the following:

$$\underset{}{l\_earn} = \underset{(0.105059)}{7.970142} + \underset{(0.010471)}{0.245304} * lcost$$

$R^2 = 0.380370$

The coefficient of lcost is positive signifying that it has a positive impact on the earnings. That is, for a 1% increase in costs will equate to a 0.245% increase in earnings. In Figure 1 the correlation between earnings and cost is +0.62 which is quite strong. Although the lcost is a significant regressor the R^2 of 0.38 is seen to be quite small. This means that the cost of attending alone is not enough to explain the earnings.

Multiple Regression 1 (All variables included):

$$
\begin{aligned}
l\_earn = {} & \underset{(0.183554)}{6.228881} + \underset{(0.022497)}{0.129592} * private + \underset{(0.205649)}{0.205649} * lcost + \underset{(0.013631)}{0.064919} * lexp - \underset{(0.025189)}{0.019718} * c150\_4 \\
& + \underset{(0.030691)}{0.048896} * rpy\_7yr\_rt - \underset{(0.0000436)}{0.0000856} * sat\_avg\_all + \underset{(0.017953)}{0.148766} * logfaminc \\
& - \underset{(0.020061)}{0.024794} * ugds\_white + \underset{(0.025341)}{0.0370} * ugds\_black - \underset{(0.025673)}{0.033992} * ugds\_hisp - \underset{(0.074937)}{0.014253} * ugds\_asian
\end{aligned}
$$

$R^2 = 0.526238$

Regression model 1 includes all the variables and based on the R^2 of 0.53 this model starts to explain the true model a lot better than the simple regression. In model 1 the most impactful regressors were lcost, lfaminc and private, all having a positive effect on l_earn. Additionally, lcost, lfaminc, and private yielded a 0.206%, 0.149%, and a 0.13% increase to l_earn if there were a 1% increase to just one of those three regressors in their respective cases. Other interesting things in this model to see is that based on race only the black culture provided a positive response to l_earn, this could indicate that diversification in race in these institutions may decrease the future earnings. Another idea given in this model is that the SAT average scores seem to be negligible in affecting l_earn therefore, regardless of one's SAT mark they could be equally as successful as another student with a different mark.

Multiple Regression 2 (All independently significant variables included):

$$
\begin{aligned}
l\_earn = {} & \underset{(0.150915)}{6.223375} + \underset{(0.012304)}{0.152475} * lexp + \underset{(0.036193)}{0.051981} * rpy\_7yr\_rt - \underset{(0.0000534)}{0.0000258} * sat\_avg\_all + \\
& + \underset{(0.014693)}{0.266704} * lfaminc
\end{aligned}
$$

$R^2 = 0.457557$

In model 2 we are observing how the individual impacting variables affect the students' earnings. All the other variables are seen to not give an insight of how it can impact an individual's earnings. After observing model 2 it is seen that this model is worse than the first model given its R^2 of 0.46. The fact that we observe these specific variables while leaving out the other significant variables such as private and lcost influences the decrease in R^2.

Multiple Regression 3 (Final Model from Original Paper):

$$\underset{(0.151602)}{\text{l\_earn} = 6.244893} + \underset{(0.012321)}{0.151837} * \text{lexp} + \underset{(0.036231)}{0.054721} * \text{rpy\_7yr\_rt} - \underset{(0.0000536)}{0.0000231} * \text{sat\_avg\_all}$$
$$+ \underset{(0.014740)}{0.266244} * \text{logfaminc} - \underset{(0.023968)}{0.032570} * \text{ugds\_white} + \underset{(0.030492)}{0.028234} * \text{ugds\_black} - \underset{(0.030184)}{0.033631} * \text{ugds\_hisp}$$
$$- \underset{(0.093929)}{0.041410} * \text{ugds\_asian}$$

$R^2 = 0.459519$

Multiple Regression 4 (Our improved model):

$$\underset{(0.173852)}{\text{l\_earn} = 6.109085} + \underset{(0.022532)}{0.131606} * \text{private} + \underset{(0.024556)}{0.209156} * \text{lcost} + \underset{(0.013648)}{0.065245} * \text{lexp}$$
$$+ \underset{(0.017914)}{0.147503} * \text{lfaminc}$$

$R^2 = 0.519596$

In model 4, by just using the statistically significant variables we were able to generate an R^2 of 0.520 compared to 0.526 of model 1 while using 7 less variables. This indicates that the chosen variables have the largest impact on l_earn while all having a positive effect too. In this paper lcost was included based off a 5% significance level test observed from all the models. We can also see that our R^2 for the improved model is higher than the R2 for the original final model (Model 3).

F-Tests for Significance:

Dependent variable l_earn					
Independent variables	Simple Regression	Model 1	Model 2	Model 3	Model 4
private		0.129592 (0.022497) (0.0000)			0.131606 (0.022532) (0.0000)
lcost	0.245304 (0.010471) (0.0000)	0.205649 (0.024564) (0.0000)			0.209156 (0.024556) (0.0000)
lexp		0.064919 (0.013631) (0.0000)	0.152475 (0.012304) (0.0000)	0.151837 (0.012321) (0.0000)	0.065245 (0.013648) (0.0000)
c_150_4		-0.019718 (0.025189) (0.4339)			

rpy_7yr_rt		0.048896 (0.030691) (0.1115)	0.051981 (0.036193) (0.1512)	0.054721 (0.036231) (0.1312)	
sat_avg_all		-8.56E-05 (4.36E-05) (0.0500)	-2.58E-05 (5.34E-05) (0.6296)	-2.31E-05 (5.36E-05) (0.6660)	
lfaminc		0.148766 (0.017953) (0.0000)	0.266704 (0.014693) (0.0000)	0.266244 (0.014740) (0.0000)	0.147503 (0.017914) (0.0000)
ugds_white		-0.024794 (0.020061) (0.2168)		-0.032570 (0.023968) (0.1745)	
ugds_black		0.037000 (0.025673) (0.1446)		0.028234 (0.030492) (0.3547)	
ugds_hisp		-0.033992 (0.025673) (0.1858)		-0.033631 (0.030184) (0.2654)	
ugds_asian		-0.014253 (0.074937) (0.8492)		0.041410 (0.093929) (0.6594)	
Observations	1316	1316	1316	1316	1316
R^2	0.380370	0.526238	0.457557	0.459519	0.519596

Figure 6: Coefficient, standard error, p-value of the variables in each model

In Figure 6, the values in each cell are presented in the order of coefficient first, standard error in the first set of parentheses and p-value in the next set of parentheses. We decided to test all the regressors under a 5% significance test and given the results the final model only included private, lcost, lexp, and lfaminc yielding the best representation of the true model given only these variables. Also, the reason for the exclusion of the udgs variables was due to the Wald test (shown in Appendix B), which indicated that the race demographic variables were not jointly significant.

Conclusion

In this paper, we aimed to investigate the link between increasing tuition costs on average earnings of graduates. We also looked at other variables that could be influencing average earnings of graduates. We undertook several single and multiple regression models and investigated significance of regressors using F-Tests and R^2 values in order to find an improved model that could explain our dependent variable. From our study we found that the regressors that influence average earnings are whether a school is public or private, average total cost of attendance for a student, expenditures per student by the university, and average family income of students. This study could be expanded upon by finding regressors not included in this paper and attempting to improve the model even further.

References

Fox, M. (1993). Is it a good investment to attend an elite private college? *Economics of Education Review, 12*(2), 137-152. Retrieved March 24, 2021

Johnson, W., Lax, K., & Mitnick, C. (2016, April). *A Regression Analysis of College Tuition and Mean Income.* Retrieved from Georgia Tech Library: https://smartech.gatech.edu/handle/1853/54738

Kara, O. (2008). Comparing Two Approaches to the Rate of Return to Investment in Education. *Education Economics, 18*(2), 153-165. doi:10.1080/09645290802416486

Appendix A:

Residual Graphs used to investigate homoskedasicity assumption

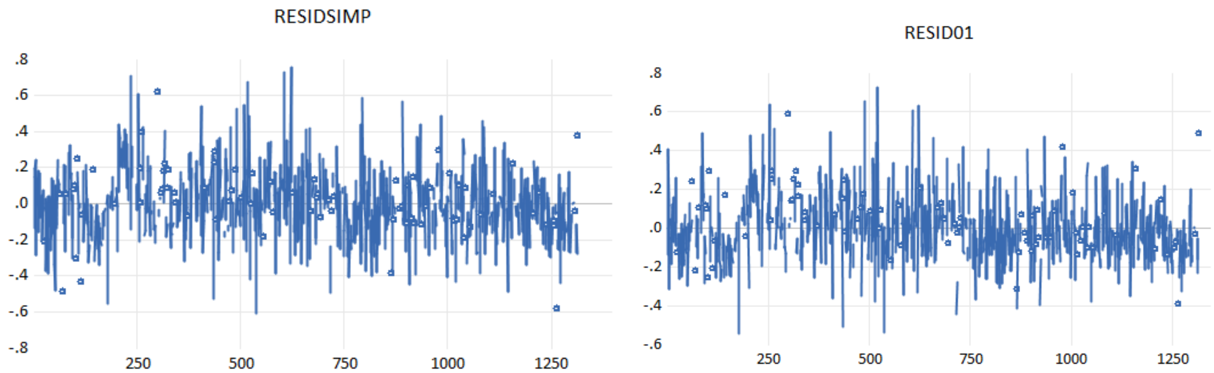

Figure 2: Residual plots for simple regression (left) and multiple regression model 1 (right)

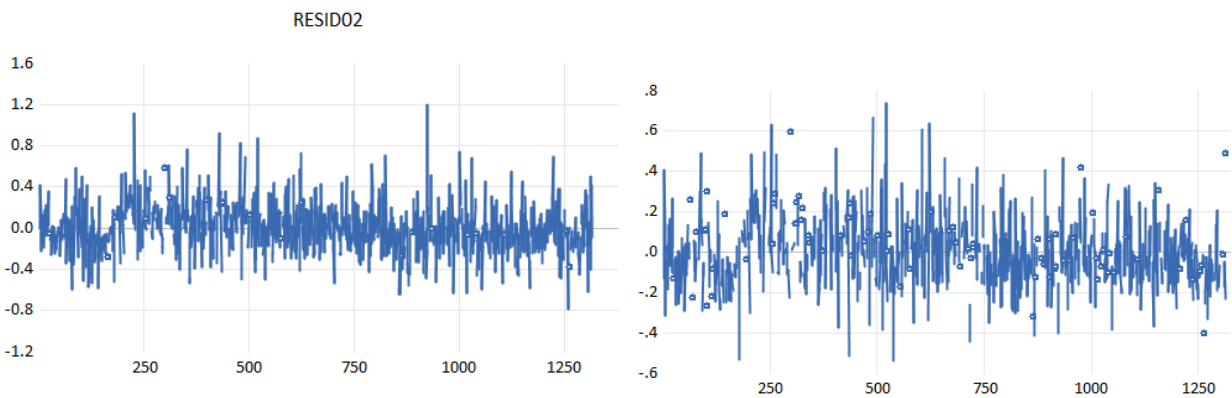

Figure 3: Residual plots for multiple regression model 2 (left) and model 3 (right)

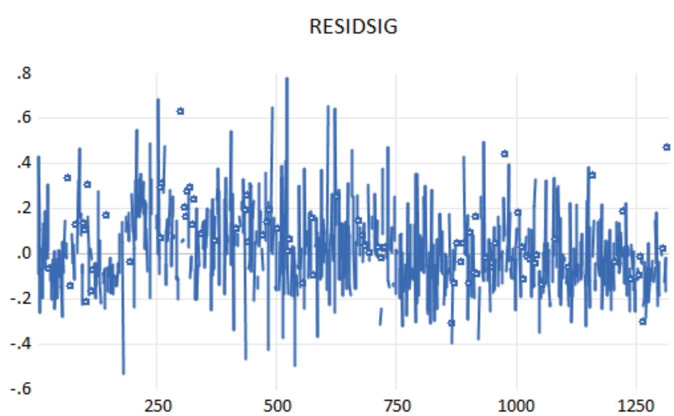

Figure 4: Residual plot for new multiple regression model

Appendix B:

Wald test for Joint Significance of racial demographic variables

H_0: $\beta_7 = \beta_8 = \beta_9 = \beta_{10} = 0 \;\; vs. \;\; H_1: not\ H_0 (at\ least\ one\ \beta_i \neq 0, i = 7, \dots, 10)$

F-Stat = 0.806330

P-Value = 0.5212

We therefore fail to reject the null hypothesis of joint significance of the racial demographic variables with our dataset.

Appendix C:

Below we have included summary statistics for each included variable in the paper

Variables	Observation	Mean	Standard Deviation	Minimum	Maximum
l_earn	1316	10.34533	0.349382	9.392662	12.05699
private(if=0)	1316	0.432371	0.495594	0.0	1.0
lcost	1316	10.01450	0.559596	8.703175	11.21861
lexp	1316	8.703125	0.781537	5.549076	12.18671
c150_4	1316	0.496304	0.223677	0.00	1.00
rpy_7yr_rt	1316	0.546958	0.184698	0.106599	0.958869
sat_avg_all	1316	1144.978	125.4514	809.00	1566.00
lfaminc	1316	10.43455	0.555863	7.055467	11.80264
ugds_white	1316	0.498881	0.275755	0.00	1.00
ugds_black	1316	0.178025	0.220595	0.00	1.00
ugds_hisp	1316	0.174496	0.222573	0.00	1.00
ugds_asian	1316	0.036821	0.071410	0.00	0.787900