Data Analysis of College Scorecard

Adarsh Srinivas
College of Information Studies
University of Maryland,
College Park
adarshs1@umd.edu

Isha Kaur Ratti
College of Information Studies
University of Maryland,
College Park
ikratti@umd.edu

Jay Sheth
College of Information Studies
University of Maryland,
College Park
jaysheth@umd.edu

1. Introduction:

There are approximately 7804 universities in the U.S. [1]. With such a huge pool of options available, the students and their families would surely have a tough time shortlisting the universities they would consider applying to and further deciding on the best available option. So how do students go about this? What factors do they consider? What factors do their families consider? This decision might be one of the most important one in their lives. Hence, we plan to make it simpler for the students and their families to make this decision through our analysis and findings. We are using the recently released College Scorecard data by the US Government. The URL file path for it is https://collegescorecard.ed.gov/data/

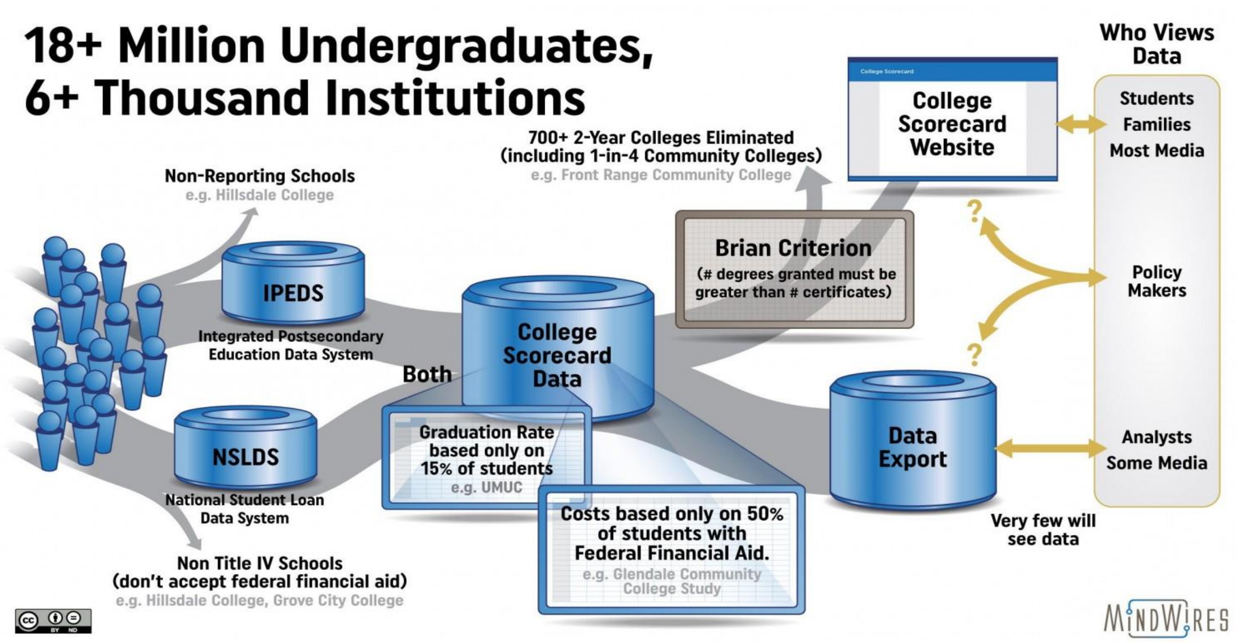
This College Scorecard initiative has been designed to increase the power in the hands of the public (students and families) to differentiate colleges and get an idea about how well these colleges are preparing their students in the path of success helping those making better informed decisions about selecting a college. As this College Scorecard is the most recent data until now (September 2015), it includes data on completion rates, loan payment, debt and post-school earnings which provides a comprehensive picture of student outcomes [5]. This recent data covers an entire range of topics including categories of data such as root, school, academics, admissions, student, cost, aid, repayment, completion and earnings. This dataset covers nearly 18 years of data consisting of several sources like Integrated Postsecondary Education System (IPEDS), National Student Loan Data System (NSLDS) and Department of Treasury. These data also provide insights into colleges that are eligible of receiving federal financial aid thereby looking at the outcomes of students at such colleges.

Having a look at statistics, in the fall of 2013, roughly 20 million students were enrolled in 4,724 degree-granting institutions, up from 10.2 million students in 3,004 institutions in 1974 [2]. At the same time, there were another 472,000 students attending over 2,500 non-degree granting institutions that often involve shorter term certificate programs [2]. These statistics do not take into account the millions of individuals who considered college but decided not to attend and the millions of students who chose to leave college.

This project was basically designed to provide better reliable and unbiased information for better college choice and college performance [6]. It was developed to address the lack of information about university costs, quality, earnings etc. There are many colleges offering high quality education and prepare students for professional experiences. However, there are some colleges which do not serve their students very well e.g. high fees, lack of support to students for completing education and getting decent jobs [1]. There is an increasing need of more information and the need of an extensive dataset which accounts for such variations in educational opportunities and hence, College Scorecard comes as a welcome initiative.

It is important for USA as a country to have a high quality educational system as there are tens of millions of students each year deciding to pursue higher education. Also, this is important in elevating America's position in the global economy by providing education of the highest quality [1]. This initiative provides unprecedented transparency about the college costs, quality and other parameters which further helps a great deal in achieving its objective.

An analysis of the College Scorecard by Phil Hill and Russ Poulin is shown in the figure below [3]:


18+ Million Undergraduates,
6+ Thousand Institutions
Who Views Data
College Scorecard
College Scorecard Website
Students
Families
Most Media
700+ 2-Year Colleges Eliminated
(including 1-in-4 Community Colleges)
e.g. Front Range Community College
Non-Reporting Schools
e.g. Hillsdale College
?
Brian Criterion
(# degrees granted must be greater than # certificates)
Policy Makers
IPEDS
Integrated Postsecondary Education Data System
College Scorecard Data
Both
?
Data Export
Analysts
Some Media
NSLDS
National Student Loan Data System
Graduation Rate based only on 15% of students
e.g. UMUC
Costs based only on 50% of students with Federal Financial Aid.
e.g. Glendale Community College Study
Very few will see data
Non Title IV Schools
(don't accept federal financial aid)
e.g. Hillsdale College, Grove City College
MINDWIRES

2. Research Question:

In the given data set multiple factors are taken into consideration, but from the student point of view not all factors would be of equal importance. The criteria for selecting a university may differ among students. So how does one decide which factors to look at or which are the factors that need to be given more importance over the others? The answer to this question is subjective as a student tends to have heterogeneous goals related to higher education. A particular student may be interested in a university that has an excellent record of the students having high average salary post-graduation or salary may not matter to that student at all or the primary criteria could be academic interests or the location. In this way, the list may seem endless.

To overcome this dilemma, we intend to analyze the factors that according to us can best predict a university's success. However, these factors too can be influenced by multiple other variables and are thus considered as proxies for success. For this purpose, a set of independent variables are selected that are tested against each of the dependent variables.

The goal of this analysis is to test whether the success factors are under the influence of independent variables and if yes, how much each of the independent variables contribute to the variance in the dependent variables. For example, a university with low average output earnings may not necessarily correctly indicate the success of a university because there are various other factors that control the variance in the output earnings. To study these variances, this analysis is being conducted.

3. Method:

For our analysis, we are using the recently released College Scorecard dataset by the U.S. Department of Education. This dataset contains huge information about Universities statistics across the United States. The scope of the data under our study spans nearly 18 years from 1997 to 2015 [6]. This selected time span has over 7800 entries across 2000 variables, which are good enough to obtain satisfying outcomes. The sample of our analysis are the 7804 universities. For our study, the descriptive list of useful variables are as follows:

3.1 Variables:

3.1.1 ADM_RATE_ALL:

This variable describes the number of undergraduates accepted divided by the number of undergraduate applied. This variable helps the students get an idea of the acceptance rate of a particular university and their chances of getting it to one. The admission rate describes the university's stringiness in student profile evaluation and selection. But at the same time it is important to know the yield rate i.e. the number of accepted students that actually take the admission. It is thus important to measure the influence of various other factors on admission rate in order to measure a university's success. Thus, admission rate has been included as one of the dependent variables.

3.1.2 PCTFLOAN:

Another factor considered significant by most students while selecting a university is the cost of attending the university. Thus, percentage of federal grants provided by a university can help students figure the amount of loans they have to take and for how long the student would have to continue paying debt post-graduation. Besides, availability of federal grants gives an opportunity to students with low family incomes have access to and afford high education. This variable describes the percentage of undergraduate students receiving federal grants which would help the prospective students have an overview of their chance in receiving the aids and in turn gain an estimate of their cost of attendance. Since financial aids are an important aspect of the overall costs involved in pursuing higher education, it has been considered as a factor to measure a university's success.

3.1.3 md_earn_wne_p10:

The rationale behind selecting the median income of graduated students is that it allows the prospective students to have a look at the median salaries of former students over a span of 10 years. One of the most common reasons mentioned by students while picking a university is the employment opportunities they will receive post-graduation. Earnings are often considered as a measure to determine if the cost of education was worth it i.e. how long would the students take to cover their cost of education. However, it is not only about recovering the cost and paying the debt. It helps the students gain an insight about expansion of their career in future.

3.1.4 C150_4:

Concentrating on four year schools, completion rate is selected to predict a university's success as it goes beyond explaining just the time taken to graduate or finish the program. It is simultaneously associated with various other outcomes expected by students. Firstly, with completion you get the degree certificate indicating that you have graduated from college. Next amongst them is the experiences students have during their course of completion which can be described as an evaluator of how their expectations from the university have been met. Employment opportunities and the ability to repay student loan are other factors associated with completion rate [2]**.** With so many positive outcomes having associated with it, completion rate made it to our list of factors predicting success.

3.1.5 COSTT4_A:

This variable is the average cost of attendance for academic year institutions and includes tuition & fees, books and supplies, and living expenses for full time, degree seeking undergraduate students. [4]

Note: We want to consider the total average cost of attendance at a particular university for an academic year and not just tuition fees or the program year which is why we choose COSTT4_A in favor of other variables.

3.1.6 CONTROL:

This variable specifies whether the institution is public, private nonprofit, or private for profit.

3.1.7 UGDS_WHITE:

This variable tells us the percentage of undergraduate students in a given university who are White.

3.1.8 UGDS_BLACK:

This variable tells us the percentage of undergraduate students in a given university who are Black.

3.1.9 UGDS_HISP:

This variable tells us the percentage of undergraduate students in a given university who are Hispanic.

3.1.10 UGDS_ASIAN:

This variable tells us the percentage of undergraduate students in a given university who are Asian.

3.1.11 UGDS_AIAN:

This variable tells us the percentage of undergraduate students in a given university who are American Indian/ Alaska Native.

3.1.12 UGDS_NHPI:

This variable tells us the percentage of undergraduate students in a given university who are Native Hawaiian/ Pacific Islander.

Note: For our analysis, we are interested in the population of United States and we feel the above five races constitute the majority of demographics of USA which is why we do not include other variables and races like UGDS_2MOR, UGDS_NRA, UGDS_UNKN.

3.1.13 SAT_AVG_ALL:

This variable includes the average SAT score of enrolled students in a particular university which can help students to find a school that is a good match.

Note: SAT is one of the most widely taken tests across the United States and is a requirement before enrolling into any University. For this reason, we have chosen SAT_AVG_ALL among other variables as this variable gives us the average SAT score of students enrolled in any university.

3.2 Data Cleaning/Processing:

Firstly, we are interested in universities that provide a 4-year Bachelor’s degree or higher. So, we filter out cases (universities) that offer a 2-year degree or less. Secondly, all the variables given in the dataset are of type ‘String’. Since analyses and tests cannot be performed on ‘String’ variables, we have converted the String variables into Numeric variables in SPSS. Further, we perform a pairwise deletion where we filter cases in which both SAT_AVG_ALL and ADM_RATE_ALL variables have missing values. This helps us to conduct our analyses on the variables of our interest resulting in a cleaner dataset than the initial one. However, this cleaner dataset still contains lot of missing values. We filter out these values by replacing the missing values with Series Mean on SPSS. The variables ADM_RATE_ALL, COSTT4_A, CONTROL, UGDS_WHITE, UGDS_AIAN, and

UGDS_NHPI consists of less than 5% missing values which we think is marginal. Hence, the missing values of these variables are not replaced with series mean. However, other variables like SAT_AVG_ALL, UGDS_BLACK, UGDS_HISP, UGDS_ASIAN, C150_4 and md_earn_wne_p10 has more than 5% missing values and hence the missing values of these variables are replaced by series mean. Lastly, we check for normality of all the variables under test. After running descriptive statistics for each of these variables, we find that UGDS_WHITE, C150_4, md_earn_wne_p10, PCTFLOAN and ADM_RATE_ALL are not normally distributed with skewness and kurtosis not falling in the acceptable range of -1.5 to +1.5. Therefore, we transform these variables by performing a LOG transformation on SPSS (Lg10). After running descriptive statistics of the transformed variables and observing the histograms, we find that all the variables now approximate to a normal distribution. Hence, we now proceed with our analyses after cleaning the data and ensuring normal distribution of all variables.

3.3 Test performed:

Through our research question, we are interested in finding the relationship between one continuous DV (Admission Rate or Completion rate or Federal Aid or Earnings) and multiple (four) categorical/continuous IV's (SAT score, Private/Public, cost of attendance and Race). Hence, we perform a multiple regression test. This test allows us to observe the effect and measure the relationship between an IV and a continuous DV while holding other IV's constant.

After running the tests, from the correlation matrix, we see that none of the IV's are significantly correlated with the DV since the correlation coefficient of each IV is less than 0.7. Also, from the collinearity statistics in the coefficients table, we see that Variance Inflation Factor (VIF) is < 3 and Tolerance is >.3 for all variables. Hence, we say that the assumption of multicollinearity has not been violated. Scatterplot provides the best view of scedasticity. From the scatterplot, we see that variance around the regression line is same for all values and hence we can conclude that the Homoscedasticity assumption has not been violated.

4 Analyses:

In our study as we are looking at predicting the values of success factors based on the values of multiple other variables. In multiple regression, the variables we are interested in predicting are known dependent variables and the one's used for predicting the dependent variable are known as the independent variables. It allows you to measure the relationship between a dependent variable and an independent variable while keeping other independent variables constant. Since there are multiple dependent variables that need to be tested, each of the dependent variable is taken at a time and then multiple regression is performed using the set of independent variables. Hence, multiple regression is the best option in order to conduct this analyses.

In multiple regression, it is important to test the variables for the following assumptions of multiple regression:

- Homoscedasticity: This assumption states that the variances along the line of best fit remain similar as you move along the line. Violation to homoscedasticity leads to heteroscedasticity in which the distribution is concentrated at one end and is loose at another end. The assumption is checked using plots in which actual residuals are plotted by predicted residuals in DV.
- Multicollinearity: Data should not show multicollinearity, which occurs when you have two or more independent variables that are highly correlated with each other. If two or more IV's are correlated at 0.7 or higher, multicollinearity occurs. Collinearity Diagnostics" can be used to test variables for multicollinearity. This table provides two additional details in your coefficient table, Variance Inflation

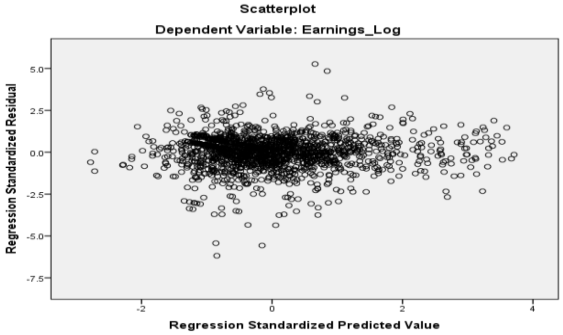
Factor (VIF) which should be low (<3) and Tolerance should be high (> .3).

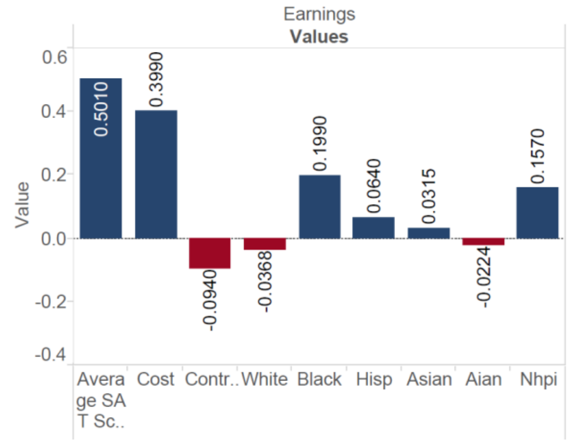
- Linearity: There should be a linear relationship between the dependent variable and each of the independent variables and the dependent variable and the independent variables together.
- Independence of observations: This assumption states that only one value has been recorded for each case in a particular variable.
- Normality: This assumption states that each of the variables being considered in the test should have normal distribution.

The tests performed for each of the dependent variables and their corresponding outputs are discussed below:

4.1 md_earn_wne_p10 (Earnings)

For the given analysis, the relationship between each predictor and earnings is computed holding constant the effect of the other predictors. In the default mode R square= 0.362 i.e. 36.2% variance in the median output earnings is accounted by all of the mentioned independent variables together. The ANOVA test tells us whether the model is good fit for data and from the given results,
F (9, 1717) =108.382, p<0.001 it can be said that the results are statistically significant and the model is a good fit. Also with the VIF values less than 3 and tolerance greater than 0.3, the assumption of multicollinearity is not violated. From the scatterplot it can be seen that the distribution is concentrated in the center and thus it can stated that assumption of homoscedasticity is also not violated.





From the correlation descriptive in multiple regression, the following can be concluded about strong and moderate relationships:

- SMEAN (SAT_AVG_ALL) (Average SAT score): It is significantly correlated with earnings as p<0.001 and with r = 0.501 it is the most highly correlated variable possessing a strong relationship.
- C150_4_Log (Completion rate): It is significantly correlated with median output earnings as p<0.001 and with r = 0.353 it is moderately correlated.
- COSTT4_A (Average Cost of Attendance): It is significantly correlated with earnings as p<0.001 and the strength of the relationship is moderate with r=0.399

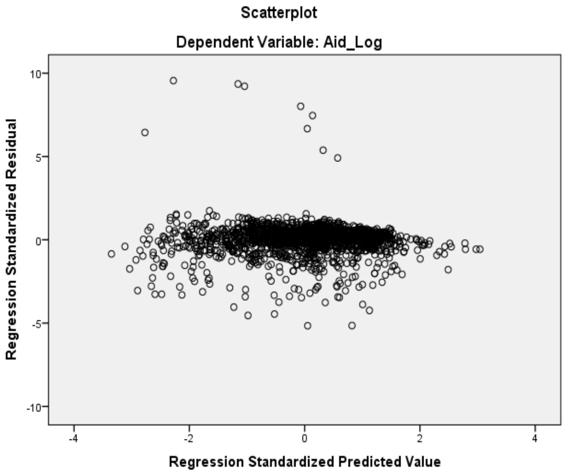
Finally based on the unstandardized beta coefficients, the equation for regression line can be given as:
y=4.233-.011(UGDS_White_Log)
+0.001(SMEAN (UGDS_BLACK))-
0.002(SMEAN (UGDS_HISP)) + 0.000(SMEAN
(UGDS_ASIAN)) +0.000(SMEAN
(SAT_AVG_ALL)) +2.422*10^{-6}(COSTT4_A)—
0.036(CONTROL)-0.001(UGDS_AIAN)
+0.003(UGDS_NHPI).

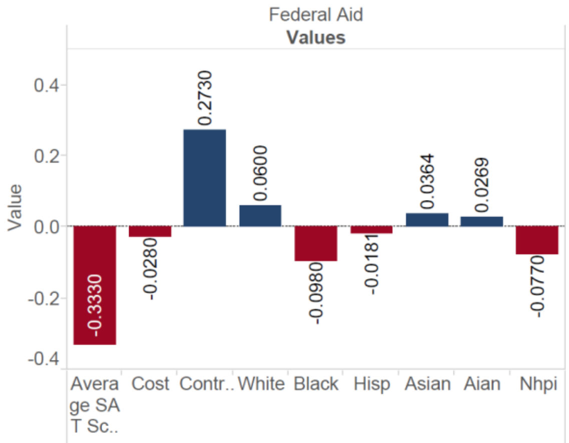
For given values of the IV, the value for median output earnings can be calculated. Thus, with the available results it can be seen that the median output earnings is most highly correlated with average SAT score and using standardized beta coefficient, one standard deviation increase in SAT score leads to a 0.393 standard deviation increase in predicted output earnings, with the other variables held constant.

With a high correlation coefficient of R =0.602 , a high F score value of 108.38 and low standard error of 0.079, it can be concluded that the regression model is statistically significant with median output earnings as the dependent variable and can be considered as a proxy for success of a university while taking into consideration multiple independent variables.

4.2 PCTFLOAN (Financial Aid)

In this analysis R square= 0.241 i.e. 24.1% variance in the median output earnings is accounted by all of the mentioned independent variables together in the default mode. The ANOVA test tells us that F (9, 1665) =58.595, p<0.001 indicating that at least one of the independent variables is linearly related with the financial aids. The results are thus statistically significant and the model is a good fit. The VIF values are less than 3 and tolerance is greater than 0.3 indicating that the assumption of multicollinearity is not violated. From the scatterplot it can be seen that the distribution is concentrated in the center and not on either sides of the distribution. Besides a few cases which are away from the best fit line, it can be stated that assumption of homoscedasticity is also not violated.





From the correlation descriptive in multiple regression, amongst all the variables control variable and SAT score have a significant relationship with the dependent variable:

- SMEAN (SAT_AVG_ALL) (Average SAT score): It is significantly correlated with financial aid as p<0.001 and with r = -0.333 it is the most highly correlated variable possessing a moderate relationship.
- CONTROL (Public/Private): It is significantly correlated with earnings as p<0.001 and with r = 0.273 it possess a near moderate relationship.

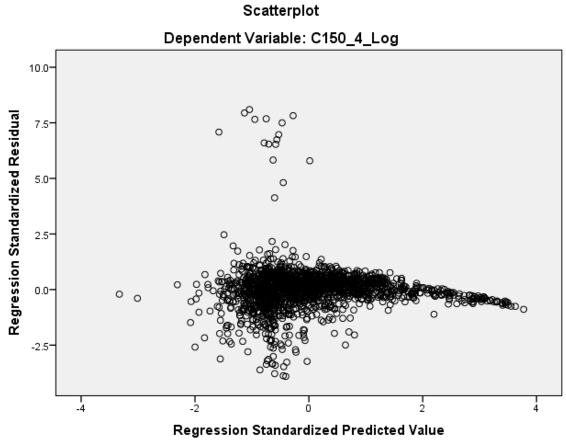
Finally based on unstandardized beta coefficients, the equation for regression line can be given as:
y=0.206+0.054(UGDSWhite_Log) +0.001(SMEAN (UGDS_BLACK)) +0.001(SMEAN (UGDS_HISP)) + 0.001(SMEAN (UGDS_ASIAN))-0.001(SMEAN (SAT_AVG_ALL))-2.572*10^{-6}(COSTT4_A)—0.086(CONTROL)-0.000(UGDS_AIAN)-4.656*10^{-6}(UGDS_NHPI).

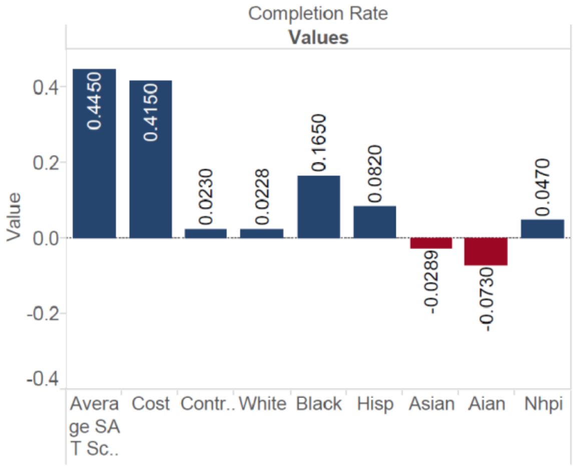
For given values of the IV, the value for aid can be calculated. Thus, with the available results it can be seen that aid is most highly but negatively correlated with average SAT score and using standardized beta coefficient, one standard deviation increase in SAT score leads to a 0.386 standard deviation decrease in predicted aid, with the other variables held constant. This is followed by control variable in which one standard deviation increase in control leads to a 0.329 standard deviation increase in predicted aid.

With independent variables explaining 24.1 variance in financial aid, a correlation coefficient of R =0.490, F score value of 58.595 and low standard error of 0.1396, it can be concluded that the regression model is statistically significant with p<0.001 and can be used as a proxy factor for a university's success.

4.3 C150_4 (Completion Rate)

Tests conducted for the dependent variable completion rate yield R square=0.296 explaining that about 29.6% variance in completion rate is together contributed by all the independent variables. The ANOVA test yields F (9, 1717) =80.001, p<0.001 indicating statistically significant results and that at least one of the independent variables have a linear relationship with completion rate. The multicollinearity assumption of multiple regression is not violated with VIF<3 and Tolerance>0.3. As for homoscedasticity, the scatter plot for completion rate shown below depicts that maximum of the distribution is concentrated at the center and the distribution at the right is also along the line of best fit. Hence, the assumption of homoscedasticity is not violated.





The correlation descriptive indicate that amongst all the variables SAT score and Cost of attendance have a significant relationship with the completion rate:

- SMEAN (SAT_AVG_ALL) (Average SAT score): It is significantly correlated as p<0.001 and with r = 0.4450 it is the most highly correlated variable possessing a moderate relationship.
- COSTT_4A (Average cost of attendance): It is significantly correlated with completion rate as p<0.001 and with r = 0.4150 it possesses a near moderate relationship.

Using standardized beta coefficients, the equation for regression line can be given as: y=-0.966+0.025(UGDSWhite_Log) +0.001(SMEAN (UGDS_BLACK))-0.001(SMEAN (UGDS_HISP)) - 0.002(SMEAN (UGDS_ASIAN)) +0.001(SMEAN (SAT_AVG_ALL)) +5.212*10^{-6}(COSTT4_A) - 0.042(CONTROL)-0.005(UGDS_AIAN) +0.001(UGDS_NHPI).

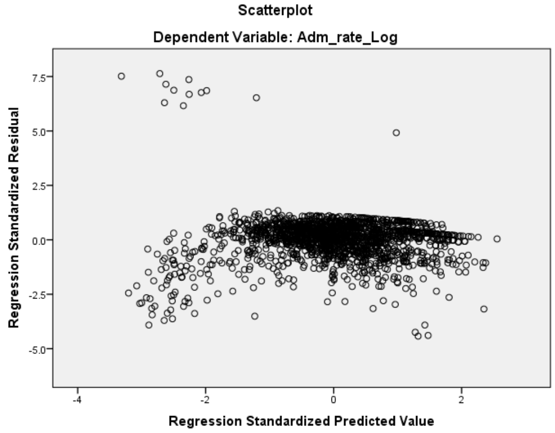
The value for completion rate can be predicted using the above equation for given values of all the independent variables using standardized beta values. With maximum variance contributed by the average cost of attendance, it can be interpreted that one standard deviation increase in COSTT_4A score leads to a 0.332 standard deviation increase in predicted completion rate, with the other variables held constant. This is followed by SAT score in which one standard deviation increase in control leads to a 0.325 standard deviation increase in predicted completion rate.

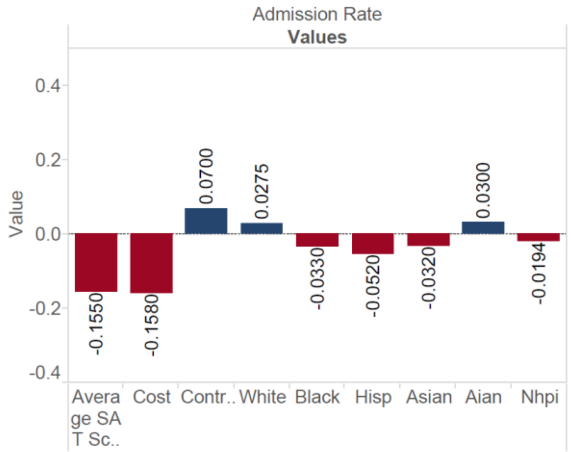
Thus, the results are statistically significant with p< 0.001 and the independent variables contributing about 29.6% variance in completion rate.

4.4 ADM_RATE_ALL (Admission Rate)

The independent variables contribute about 5.8% variance in admission rate with

R square= 0.058 which is low as compared to the contribution in other dependent variables. F (9, 1716) =11.778, p<0.001 results obtained from ANOVA indicate that the model is statistically significant and the hypothesis of this test that the regression coefficients are zero can be rejected. However again the F statistic value is not as high as the other dependent variables resulting in proportion of explained variances to be low. The multicollinearity assumption is not violated with VIF<3 and Tolerance>0.3 for all the independent variables. The assumption of homoscedasticity can be studied from the scatter plot and it can be seen that the distribution is concentrated at the center.

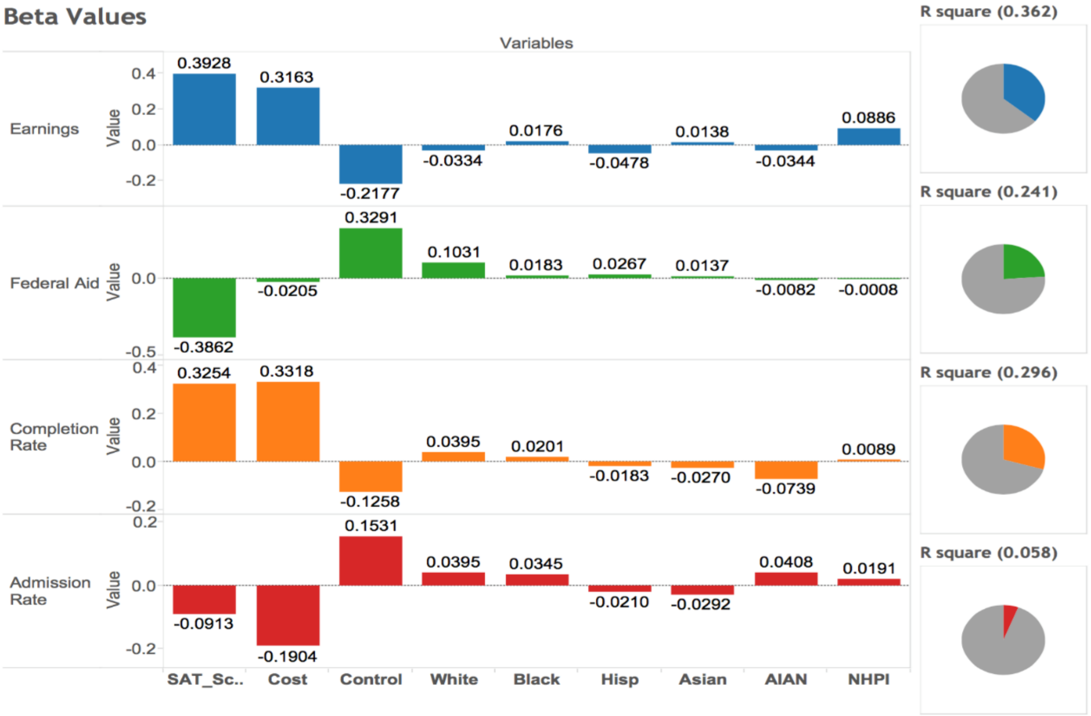




For admission rate none of the independent variables have a strong or moderate correlation with admission rate. From the standardized beta values, completion rate can be predicted using the equation for regression line as follows:

Using unstandardized beta coefficients, the equation for regression line can be given as:
y=-0.041+0.022(UGDSWhite_Log) +0.002(SMEAN (UGDS_BLACK))-0.001(SMEAN (UGDS_HISP)) - 0.002(SMEAN (UGDS_ASIAN)) +0.000(SMEAN (SAT_AVG_ALL))-2.596*10^{-6}(COSTT4_A) +0.044(CONTROL) +0.003(UGDS_AIAN) + 0.001(UGDS_NHPI).

With the independent variables accounting for low variance in the dependent variable, the maximum contribution is by average cost of attendance. With the help of standardized beta coefficient, it can be interpreted that one standard deviation increase in COSTT_4A score leads to a 0.190 standard deviation decrease in predicted admission rate, with the other variables held constant. The results are statistically significant but the variance accounted by independent variables is low with 5.8%.



5 Interpretation of Findings / Discussion:

After conducting extensive statistical tests on the college scorecard dataset we have come up with the following findings:

- Earnings is positively correlated with SAT score and cost of attendance i.e. we can infer that higher SAT score and the cost of attendance would generally lead to higher earnings.
- Federal Aid is associated with the Control variable i.e. public universities are more likely to get federal aid from the government which they will pass on to the students in the form of scholarships. However, the correlation between SAT score and Federal Aid is negative, indicating that students with a high SAT score have low chances of receiving federal aid.

- Completion Rate is positively correlated to SAT score and Cost of Attendance i.e. students with higher SAT scores and those who pay more for their education are more likely to complete their degree faster.

- Admission Rate is a variable which did not have a significant association with any of the independent variables.

These findings can be helpful for students and their families in the process of selecting a university. It can give students a clearer picture of what they should aim and consider while shortlisting a university. This knowledge would help them make better informed decisions about their careers. Stakeholders like recruiters, alumni, faculty etc. could benefit from these results giving them more substantial information about universities. Along with providing important information to the public in choosing colleges, these results could help the policy makers in the government to make decisions about federal aid and also assist Federal, State and Local Governments perspective in understanding if the investments in such colleges have been worth it and paid off.

Since, we obtained statistically significant results in our regression model for all the dependent variables, we could say that these proxies could serve as some of the many important factors which can be considered for evaluating a university's success. Our rationale behind selecting these variables as DV's is further justified and supported by the statistically significant findings obtained from the multiple regression test. Students could consider these variables while choosing a university which could help them selecting a college accurately. Future scope of this project could include taking into consideration 2-year or less degree granting institutions. Also, other factors that could predict university success could be considered in answering this research question and analyses could be conducted on the same.

References:

1. U.S. Department of Education. (September 2015). Better Information for Better College Choice & Institutional Performance. Retrieved November 29, 2015 from https://collegescorecard.ed.gov/assets/BetterInformationForBetterCollegeChoiceAndInstitutionalPerformance.pdf

2. Executive Office of the President of the United States. (September 2015). Using Federal Data to measure and improve the performance of U.S. institutions of higher education. Retrieved November 29, 2015 from https://collegescorecard.ed.gov/assets/UsingFederalDataToMeasureAndImprovePerformance.pdf

3. Svrluga, S. (October 2015). Hundreds of colleges missing from Obama's College Scorecard? The Washington Post. Retrieved November 29, 2015 from https://www.washingtonpost.com/news/gradepoint/wp/2015/10/15/hundreds-of-colleges-missing-from-obamas-college-scorecard/

4. Whitehurst, J.G. & Chingos, M.M. (October 15, 2015). Deconstructing and reconstructing the College Scorecard. Brookings. Retrieved November 29, 2015 from http://www.brookings.edu/research/papers/2015/10/15-deconstructing-reconstructing-college-scorecard-whitehurst-chingos

5. Rothwell, J. (September 28, 2015). Understanding the College Scorecard. Brookings. Retrieved November 29, 2015 from http://www.brookings.edu/research/opinions/2015/09/28- understanding-the-college-scorecard-rothwell

6. Office of the Press Secretary. (September 12, 2015). Fact Sheet: Empowering Students to Choose the College that is Right for Them. The White House. Retrieved November 29, 2015 from https://www.whitehouse.gov/the-press-office/2015/09/12/fact-sheet-empowering-students-choose-college-right-them