Modeling The Average College Faculty Salary In The United States Using Multiple Linear Regression

Author: Joanikij Chulev

INTRODUCTION

I will use a multiple linear regression model to analyze the average college faculty salary of instructors in the United States considering that there are many variables that are to be used as independent variables to predict the average instructor salary in a particular college.

In the multiple logistic regression, we have one dependent variable that we are predicting with the other independent variables. In this model we will see the relationship between the variable of interest (dependent) which is average faculty salary and the independent variables, also trying to see if our regression model explains a good percentage of variations using ANOVA. Further, this will enable us to see what is left to chance that cannot be explained by the regression line. So we will choose a good combination of independent variables that will be able to predict the average faculty salary which is the only dependent variable. Finally, we will know the independent variables that are most significant in predicting the average salary instructors are likely to earn in a particular faculty. More conclusions will be drawn from the model summary to gauge the reliability of the model to be used in salary estimation.

The ways of estimation of instructors' salaries in colleges have been done and the variable being used to do this estimation is not reliable and/or rational. For this estimation to be effective, we have to help in identifying the correct steps and the right way to conduct the salary estimation. Steps should be taken to minimize the difference that is there among instructor teaching in colleges. This research will be based on variables around their location, their state and the students. All these variables will culminate into the average salary estimation for a particular faculty member in the United States, considering the state and or city in which the college is located. Some of these conditions can attract a higher average faculty pay or attract a lower average college faculty pay to its instructors. To conclude we will also insert the list of factors that increase the likelihood of an instructor getting higher pay.

LITERATURE REVIEW

The Polyglot paper in 2010 looked at how a company dealt with employee pay decreases and the resulting consequences on high embezzlement rates and perceived equality. Panoptic salary decreases were imposed on two organizational units. A third unit had no wage reduction and was used as an effect cluster. The pay-cut explanations were addressed in a few different ways to the two pay-cut groups. Administration provided a great deal of material to explain its reasoning for the pay cut in the "satisfying clarification" pay-cut cluster, and they also expressed urgent penitence. The "insufficient clarification" cluster received a lot less information and showed no signs of remorse. There was no pay decrease for the management cluster (and so no explanation) (MANGALE, 2017). The people in charge and the two of the salary reduced groups started with a similar stealing rate and call for an equality push. Some time after the salary reduction, the stealing rate increased by 54% in the acceptable explanation cohort than in the control group. Further, in the "inadequate clarification" condition cohort, the stealing rate was 141% compared to the in-charge group. In this case, communication and debt had a huge, self-controlling result on workers' activities and also conducts (MANGALE, 2017).

A manager could vouch for a salary change of staff or employees to ensure there is pay equity. The Office of Human Resources and Strategic Talent Management (HRSTM) may conduct an equity change by examining the employee's past salary data, any relevant experience, education profile, job performance at work, length of service in this work, and relevant certifications/licenses to this job earned by an employee compared to other employees in the same occupational class (Personnel, 2019).

In twelve months in a college faculty, the salary payment for someone on an annual contract which is set as an annual figure is paid on a twice a week basis, in a similar way for those on a full-time employment contract. For the period of the twelve months the faculty is allowed to start a new contract or terminate an existing contract, the salary offered will be pegged on the 261 days worked, starting July 1st to June 30th . For only ten months in the faculty, the yearly payments will

be divided into twenty- two or twenty-six equal payments at the selection as a faculty member (College, 2020 - 2021).

It is required that the annual full-time rate of payment reflect the annual amount of pay one would receive if is employed on a full-time basis for the whole academic year. Some individuals are employed for short periods, like for two to four months with a small upfront payment, where the yearly FTR and the real pay are different. In the case of a Law School that yearly appoints instructors, an upfront pay fraction is 0.2 and they teach through to the end of the semester. To have salary equality it is calculated like this, the Appt. Annual FTR is calculated by the formula ((monthly compensation/Appt. Fraction) * appt. basis months). Like in this case, someone who has been appointed for three months at $4,000 per month with an appt. fraction of 0.05, if it is on eight months basis he/she would have it calculated like this: Appt. Annual FTR of $640,000 (($4000/.05) * 8). That person would earn $12,000 during the time he/she is selected for a task (MICHIGAN, 2020).

The reason to looking into all these salary studies is to determine which factors or roles may play in our study. This will ultimately lead us to the right selection of variables that are not obvious to be selected for the data analysis, in regards of these factors.

After looking into the behavior and attributes of students in the class attributes on learning in principles of economics classes, we reached to the following conclusions. The following two classes principles of microeconomics and macroeconomics classes have the majority of the students as male. They also found that in the past male students performed better in classes dealing with economics, and finally in their research how they designed their results made them have an opposite conclusion (NGHAMBI, 2014). A deep look into gender difference found a change in the trend of the gender of a student influence in the economics classes, where female students were better off than the male students. The majority of the students in this category secured a job outside their majors, which shows the need for flexibility (Orhan Kara, 2009). Some of the variables that were based on closely, included gender, course (macro vs. micro), # of hours worked, SAT score, number of missed classes, recommending the course to a friend, instructors, being a junior, number

of economics courses taken, and interest in the course, were found to be the significant factors contributing to learning and success in terms of passing the exams as measured by grades for introductory economics classes. Other variables that were also considered in this research were as follows; GPA, age, staying in university housing, the number of mathematics classes taken, instructor's use of graphs to explain a topic, being a fourth-year student, enrolling in a class because of the reputation of an instructor, all had a positive effect on students grades though they are not statistically significant. Finally, the effect of the number of hours per week spent on studying for the class was obsolete infomation (Orhan Kara, 2009).

In a research study on student behavior they checked on time versus performance by region (eg, slow vs fast), considering government by different regions and the regulatory bodies in the education sector, which evidently found that their relationship which was complex (Adam M. Persky & b Hannah Mierzwa, 2018). Tutors normally help learners to develop the character to help overcome natural or born tendencies in individual student personality to enhance better performance. As a matter of fact, time limit examinations normally negatively affect some students personality traits (Adam M. Persky & b Hannah Mierzwa, 2018). In 2013 in a similar linear regression was run using variables that are summarized from multiple variables and are belived to have strong relation and impact on the exam score results from the students (Rozon, 2013).

The reason to looking into all of these studies and articles is choice selection for varibles used in this study. Consequently tutor salary may be concerned with student performance and student related data variables, which must be taken into account.

OBJECTIVES

Main Objective:

To fit a multiple linear regression between one dependent and many independent variables, the average faculty salary as the dependent variable and a list of relevant variables that would predict the average faculty salary. We finally want to look at how the independent variables are affecting the dependent variable. We will also check the independent variables that are significant in predicting the average pay instructors in a particular college faculty are likely to earn.

Other Objectives:

- Investigate the significance of controlling variation by the independent variables which is regression model and control of variation by chance through ANOVA.
- Check the direction in which the independent variables are affecting the dependent variable.

RESEARCH METHODOLOGY

In this case, we want to model the salary an instructor is likely to earn while in a particular college faculty depending on certain factors. We are going to use multiple linear regression to estimate earnings. I will set up the multiple linear regression model with the two categories of variables; independent variables and a dependent variable that is being predicted.

I will use multiple linear regression which has the following mathematical equation:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p,$$

- *Y* is the value being forecasted and is also known as the dependent variable.
- *X1* is the first independent variable predicting *Y*.
- *b*0 is the intercept, it can also be the predicted value of *Y* when the *X1* for i=1…10 is 0.
- *b*1 is the coefficient of the first independent variable from it we can conclude how it is affecting the dependent variable(negatively/positively)– which is how a change in the independent variable would affect the dependent variable.

While you can perform multiple linear regression by hand, this is a tedious process, since it involves data transformation a cleaning to achieve a good model, so most people use statistical programs to help them quickly analyze the data.

We will also split the data into two tests and train to see how well the model is performing in prediction.

*In this case, we will use R software.

DATA ANALYSIS & RESULTS

The first step is to have a glimpse into how our data looks like its dimension and what we are likely to have as our dependent variable in the analysis before we clean it and are subject to the analysis.

*Below is a brief look at the head of the data:

```
dim(data)
] 7112   32
head(data)
A tibble: 6 x 32
ïunitid instnm   city  stabbr zip   control adm_rate sat_avg menonly womenonl
  <dbl> <chr>    <chr> <chr>  <chr>   <dbl> <chr>    <chr>   <chr>   <chr>
 100654 Alabama~ Norm~ AL     35762       1 0.9027   929     0       0
 100663 Univers~ Birm~ AL     3529~       1 0.9181   1195    0       0
 100690 Amridge~ Mont~ AL     3611~       2 NULL     NULL    0       0
 100706 Univers~ Hunt~ AL     35899       1 0.8123   1322    0       0
 100724 Alabama~ Mont~ AL     3610~       1 0.9787   935     0       0
 100751 The Uni~ Tusc~ AL     3548~       1 0.533    1278    0       0
```

The data comes in an Excel, downloaded from the bellow link: (due to the Corona economic fluctuations the data within the variables, was taken from the MERGED2017_18_PP (year 2018)). As concluded to be the most stable economic year for doing my examination.

1. https://collegescorecard.ed.gov/data

2. https://collegescorecard.ed.gov/assets/CollegeScorecardDataDictionary.xlsx

1-DATA ZIP FILES OF ALL TRACKED YEARS.

2-DATA GLOSSARY AND INFORMATION.

Data Variables definition is as follows:

ïunitid : Unit ID for institution.
instnm : Institution name.
city : City.
stabbr : State postcode.
zip : Zipcode.
control : Control of institution.
adm_rate : Admission rate.
sat_avg : Average SAT equivalent score of students admitted.
menonly: Flag for men only collage.
womenonly: Flag for women only collage.
ugds: Enrollment of all undergraduate students.
ugds_white : Percentage of undergraduates students who are White.
ugds_black : Percentage of undergraduates students who are Black.
ugds_asian : Percentage of undergraduates students who are Asians.
ugds_hisp : Percentage of undergraduates students who are Hispanian.
ugds_aian : Percentage of undergraduates students who are Alaskan Native.
ugds_nhpi : Percentage of undergraduates students who are Hawaiian/Pacific Islanders.
ugds_2mor : Percentage of undergraduates students who are 2 or more races.
ugds_nra : Percentage of undergraduates students who are non-residents.
ugds_unkn : Percentage of undergraduates students whose race is unknown.
ugds_men : Percentage of undergraduates students who are men.
ugds_women : Percentage of undergraduates students who are women.
costt4_a : Average cost of attendance.
tuitfte : Net tuition revenue per full-time equivalent student.
inexpfte : Instructional expenditures per full-time equivalent student.
avgfacsal : Average faculty salary.
pftfac : Proportion of faculty that is full-time.
pctpell : Percentage of undergraduates who receive a Pell Grant.
debt_mdn : The median original amount of the loan principal upon entering repayment.
pctfloan : Percent of all undergraduate students receiving a federal student loan.
c150_4 : Completion rate for first-time, full-time students at four-year institutions (150% of expected time to completion)
ret_ft4 : First-time, full-time student retention rate at four-year institutions.

A look in the image reveals the dimension of the data that informs us of the size of the data although it is subject to reduction as we clean and remove cases with missing values.

The next step is data cleaning and transformation in a way that will be easy to analyze:

```
data$adm_rate <- as.numeric(data$adm_rate)
data$avgfacsal <- as.numeric(data$avgfacsal)
data$tuitfte <- as.numeric(data$tuitfte)
data$pctpell <- as.numeric(data$pctpell)
data$debt_mdn <- as.numeric(data$debt_mdn)
data$sat_avg <- as.numeric(data$sat_avg)
data$ugds <- as.numeric(data$ugds)
data$ugds_white <- as.numeric(data$ugds_white)
data$ugds_black <- as.numeric(data$ugds_black)
data$ugds_asian <- as.numeric(data$ugds_asian)
data$ugds_hisp <- as.numeric(data$ugds_hisp)
data$costt4_a <- as.numeric(data$costt4_a)
```

One of the variables into consideration is the race proportion in the college faculties.

After data transformation, we go directly into the model fitted.

Call:

lm(formula = avgfacsal_t1 ~ tuitfte_t1 + debt_mdn_t1 + pctpell + ugds_t1 + sat_avg_t1 + ugds_white + ugds_black + ugds_asian + ugds_hisp + costt4_a_t1, data = data)

The lm() function is used to fit linear models to data frames in the R Language.

Coefficients: (with values below the variables)

(Intercept)	tuitfte_t1	debt_mdn_t1	pctpell	ugds_t1	sat_avg_t1
-0.02779	0.07749	0.03660	-0.13644	0.14550	0.50505
ugds_white	ugds_black	ugds_asian	ugds_hisp	costt4_a_t1	
-0.04065	0.06806	0.51927	0.09469	-0.03779	

*From the formula, intercepts, and coefficients we can derive the following equation. And the intercept is to be in the first position followed by the coefficient being replaced with the values:

$$Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_5X_5 + B_6X_6 + B_7X_7 + B_8X_8 + B_9X_9 + B_{10}X_{10}$$

$$Y = -0.02779 + 0.07749X_1 + 0.03660X_2 - 0.13644X_3 + 0.14550X_4 + 0.50505X_5 - 0.04065X_6 + 0.06806X_7 + 0.51927X_8 + 0.09469X_9 - 0.03779X_{10}$$

Avgfacsal = -0.02779 + 0.07749 tuitfte + 0.03660 debt_mdn - 0.13644 pctpell + 0.14550 ugds + 0.50505 sat_avg - 0.04065 ugds_white + 0.06806 ugds_black + 0.51927 ugds_asian + 0.09469 ugds_hisp - 0.03779 costt4_a

*Where the intercept which is B_0 is -0.02779 and the coefficients B_1, B_2, B_3, B_4, B_5, B_6, B_7, B_8, B_9 0.07749, 0.03660, 0.13644, 0.145500, 0.50505, 0.04065, 0.06806, 0.51927, 0.09469 and 0.03779 respectively.

This one simply means that if all the other coefficients are 0 then the predicted value of our dependent variable will be the intercept which is -0.02779. If that is not the case then the coefficient which is not 0 will influence the value of the dependent variable depending on the value and the sign in the coefficient, which will be simply concluded as a negative or positive effect.

Residuals:

Min	1Q	Median	3Q	Max
-0.26669	-0.03371	-0.00327	0.03015	0.40237

Coefficients: (In depth)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.02779	0.03614	-0.769	0.442130
tuitfte_t1	0.07749	0.01927	4.021	6.12e-05 ***
debt_mdn_t1	0.03660	0.01421	2.575	0.010146 *
pctpell	-0.13644	0.02126	-6.418	1.94e-10 ***
ugds_t1	0.14550	0.01613	9.020	< 2e-16 ***
sat_avg_t1	0.50505	0.03769	13.401	< 2e-16 ***
ugds_white	-0.04065	0.02547	-1.596	0.110732
ugds_black	0.06806	0.02743	2.482	0.013206 *
ugds_asian	0.51927	0.04956	10.479	< 2e-16 ***
ugds_hisp	0.09469	0.02857	3.314	0.000946 ***
costt4_a_t1	-0.03779	0.01740	-2.172	0.030010 *

>Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06049 on 1285 degrees of freedom

Multiple R-squared: 0.6774, Adjusted R-squared: 0.6749

F-statistic: 269.8 on 10 and 1285 DF, p-value: < 2.2e-16

From the model summary, we have all the variables as significant in predicting average faculty pay. Among them, some are more significant than others we arrange them by looking at the p-value and the star *** codes. The decision criteria for significance in predicting average faculty pay is when the p-val is less than 0.05. All the listed variables have a p-val less than 0.05 and that explains our decision of saying that all the variables are significant in predicting average faculty pay except ugds_white.

Another way to order the significance from the most significant to least is by looking at the one with the minimal value in p-val followed by the others. The most significant independent variables are ugds, sat_avg, ugds_asian they have the least p-val and full star code for significance. Still, on the significance, they are followed by pctpell then tuitfte then udgs_hisp which are the only ones with three *** star codes which shows very strong significance. The remaining shows significance in this order; debt_mdn, ugds_black, costt4_a which are significant but not as strong as the ones listed above. We have ugds_white which is not significant at all in this model. We have a total of ten independent variables in which nine of the variables are significant in predicting the average faculty salary given to the instructors except for udgs_white. The race of white did not influence too much on the salary of the instructors.

The list of our transformed variable that was used in fitting the model are;tuitfte_t1. debt_mdn_t1, pctpell, ugds_t1, sat_avg_t1, ugds_white , ugds_black, ugds_asian , ugds_hisp, costt4_a_t1.

We want to understand how each of the variables affects the dependent variable and the direction in which it affects the dependent variable.

Only three variables among the ten independent variables affect the dependent variable (average faculty salary) negatively. We look at the coefficient if it is negative an increase on that variable would result in a significant decrease in the dependent variable. They are the following;

pctpell, ugds_white, costt4_a_t1

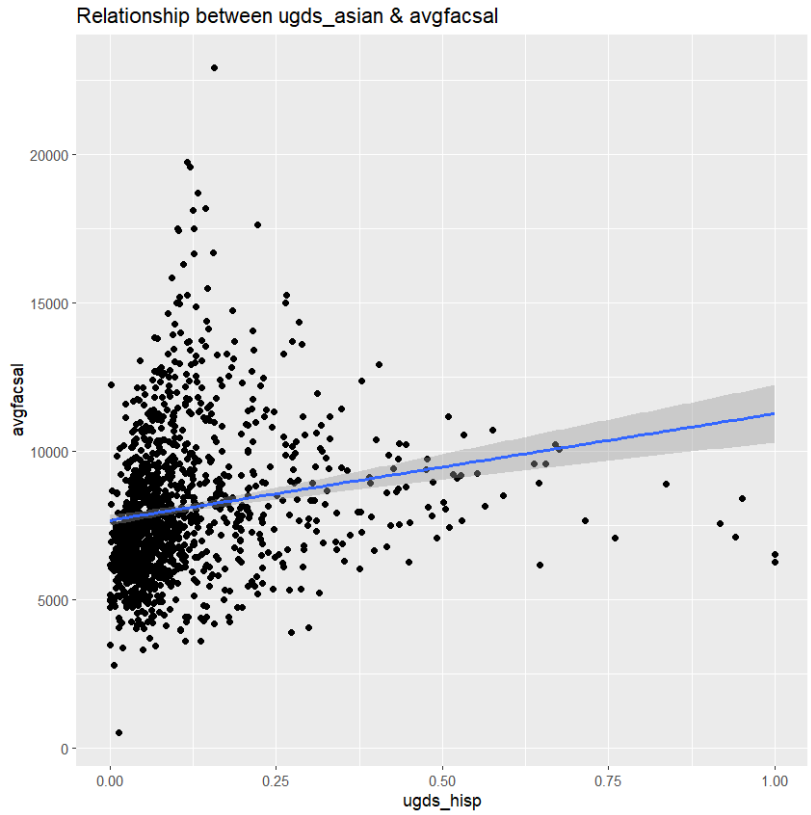
Among the three those that were significant in predicting the average faculty salary we

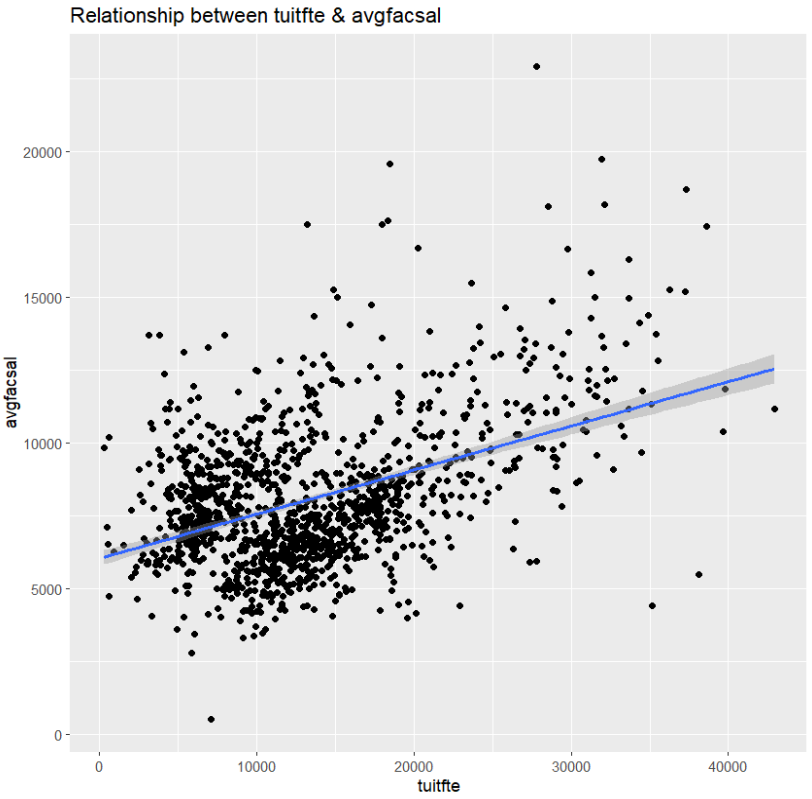
Pctpell and costt4_a_t1.

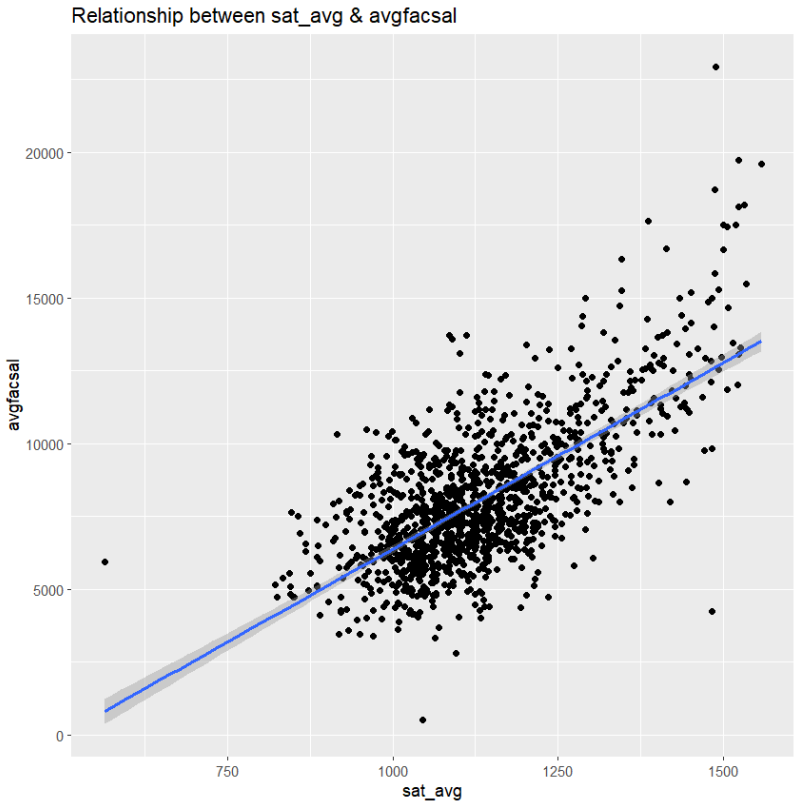
This leaves us with seven variables among the significant variables that are affecting the dependent variable positively.

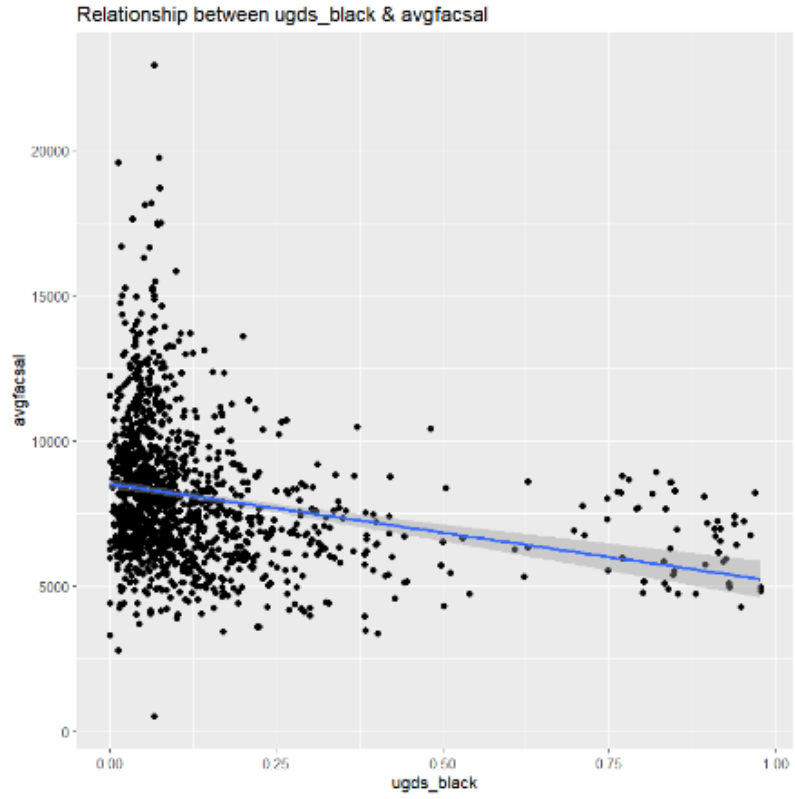
The Multiple R squared and Adjusted R squared (67%) are showing the extent to which our listed independent variables can explain the dependent variable which is average faculty salary.

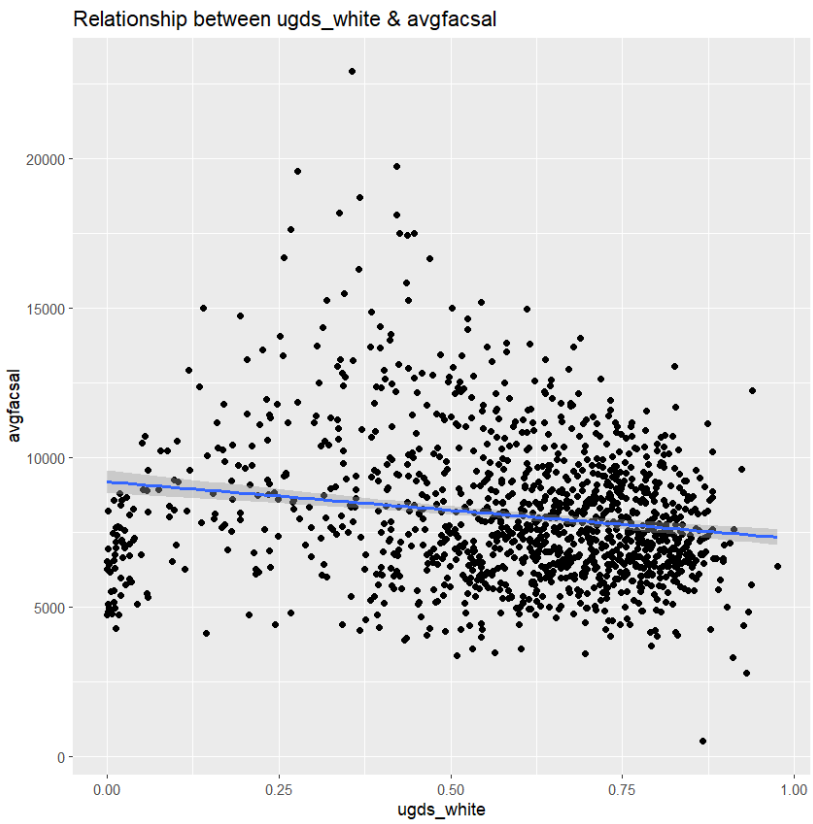
Some of the visualized relationships between the avg faculty salary and the variables (including positive and negative)

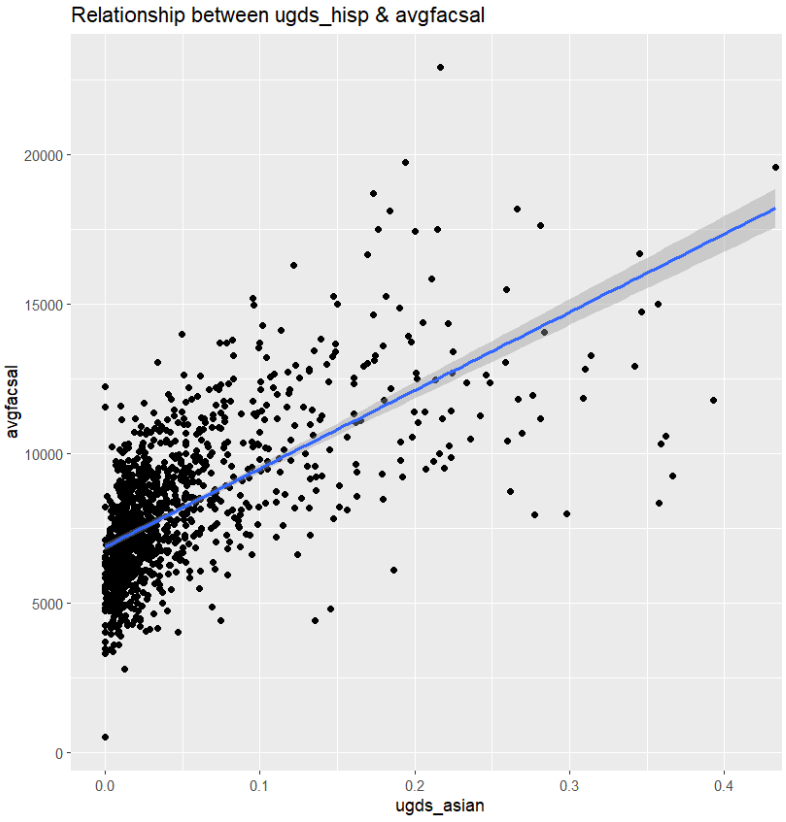


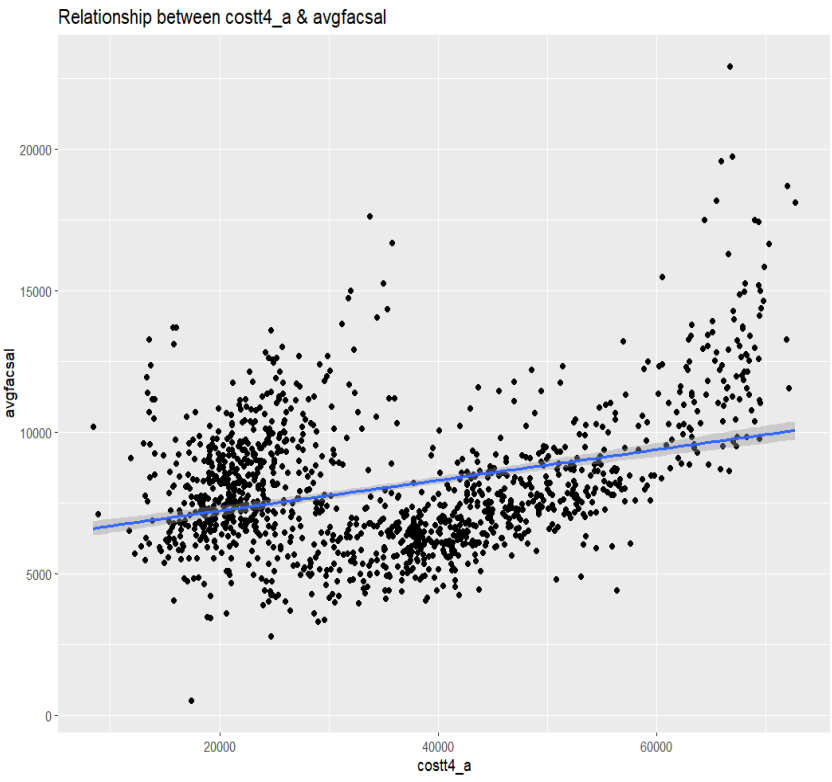





Relationship between ugds_black & avgfacsal
avgfacsal
20000
15000
10000
5000
0
0.00
0.25
0.50
0.75
1.00
ugds_black


Relationship between ugds_white & avgfacsal
avgfacsal
20000
15000
10000
5000
0
0.00
0.25
0.50
0.75
1.00
ugds_white


Relationship between ugds_hisp & avgfacsal
avgfacsal
20000
15000
10000
5000
0
0.0
0.1
0.2
0.3
0.4
ugds_asian


Relationship between costt4_a & avgfacsal
avgfacsal
20000
15000
10000
5000
0
20000
40000
60000
costt4_a

The ANOVA on the model explains to what percentage our regression model can explain residuals leaving only a few to chance.

Analysis of Variance Table

Response: avgfacsal_t1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
tuitfte_t1	1	2.9281	2.92810	800.3364	< 2.2e-16 ***
debt_mdn_t1	1	0.0462	0.04620	12.6286	0.0003937 ***
pctpell	1	1.3197	1.31973	360.7217	< 2.2e-16 ***
ugds_t1	1	2.6980	2.69797	737.4348	< 2.2e-16 ***
sat_avg_t1	1	1.2849	1.28490	351.2010	< 2.2e-16 ***
ugds_white	1	1.0281	1.02812	281.0147	< 2.2e-16 ***
ugds_black	1	0.0947	0.09467	25.8760	4.181e-07 ***
ugds_asian	1	0.4079	0.40792	111.4972	< 2.2e-16 ***
ugds_hisp	1	0.0455	0.04551	12.4384	0.0004354 ***
costt4_a_t1	1	0.0173	0.01727	4.7192	0.0300098 *
Residuals	1285	4.7013	0.00366		

>Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

From this model ANOVA, we have the results in an ANOVA table and here we can draw a conclusion from the percentage of variance controlled by the model and still narrow it down to check on how individual independent variable is controlling the variation in the dependent variable. The test statistic is a measure that allows us to assess whether the differences among the sample means (numerator) are more than would be expected by chance.

In the table, we can arrange the variable in an order of the impact of variation they are causing on the dependent variable. The variable that controlled most the variation in the dependent variable is tuitfte followed by ugds then with the following pctpell, sat_avg, ugds_white, ugds_asian in the same order. We also have those that are causing minimal variation to the dependent variable and they are the following debt_mdn, ugds_hisp, and lastly costt4_a. All the variables in the ANOVA table of the test are significant. We can firmly state that the results of the independent variables affecting the dependent variable are not by chance.

Conclusion

From the analysis, we can conclude that races like Asian and Hispanic in undergraduates were significant in determining the average faculty salaries given, because they double up in translation to those languages that are not used or are national languages in the US.

This variable udgs_white was not significant in predicting the model because mostly white students are all speaking English or the American national language, thus that is why this variable is not significant in predicting average faculty pay to the instructors.

Other significant variables in predicting average faculty pay are; udgs, debt_mdn, pctpell, and sat_avg. This means that a small deviation in them also affects the average pay an instructor would receive in them tutoring in that particular faculty.

Recommendation

The regression model is good and can be used to estimate the salary of the instructors in the United States across many colleges.

REFERENCES

Adam M. Persky, P., & b Hannah Mierzwa, P. (2018). *Factors Affecting Student Time to Examination Completion.* North Carolina: American Journal of Pharmaceutical Education.

College, V. (2020 - 2021). *Salary Schedules for the Fiscal Year 2020-2021.* Orlando, Florida: Valencia College.

MANGALE, N. M. (2017). *THE EFFECTS OF COMPENSATION ON EMPLOYEE PRODUCTIVITY.* Nairobi: The MAnagement University of Africa Repository.

MICHIGAN, U. O. (2020). *SALARY RATE OF FACULTY AND STAFF.* Ann Arbor, Michigan: UNIVERSITY OF MICHIGAN.

NGHAMBI, G. H. (2014). *FACTORS CONTRIBUTING TO POOR ACADEMIC PERFORMANCE IN CERTIFICATE OF SECONDARY EDUCATION EXAMINATION FOR COMMUNITY SECONDARY SCHOOLS IN URAMBO DISTRICT.* TABORA, TANZANIA: Digital Library of the Open University of Tanzania.

Orhan Kara, F. B. (2009). *Factors Affecting Students' Grades in Principles of Economics.* Chester University, USA: American Journal of Business Education.

Personnel. (2019). *POLICY Board of Trustees - Montgomery College.* Maryland.

Rozon, O. B. (2013). *Standardized Tests What factors affect how a student does on these exams?* Chicago: UNIVERSITY OF PUGET SOUND.