Math 430 Project

Trang Le, Yonah Barany, Claire Wiley, Naomi Oster

March 13, 2017

Abstract

This study aims to explore the relationship between the average net tuition of various higher education institutions against various predictor variables such as: percent of students receiving a specific degree type, percent of students identifying as a specific race, and percent of students receiving a Pell Grant. Using the United States Department of Education College Scorecard of January 2014, we transformed different predictor variables in order to explore the linear relationships present in the data. Through exploring data, checking validity, and transforming variables within these linear relationships, we found that specific race variables either had negative or no impact on average net tuition, the Pell Grant received variable made a negative impact on average net tuition, and a majority of specific degree types had a positive impact on average net tuition. These relationships bring to bear the hierarchical values and biases inherent in higher education. Addressing these hierarchies and biases may help students better understand their educational environment, as well as their place and perceived value within.

Introduction

The purpose of this study is to explore the biased monetary structure found within higher education by examining how the distribution of degree type, race, and grants affects the average net tuition price of universities in the U.S. This information is hoped to be used by potential students, analyzing the importance of degree type within an institution and the varying value of different degrees. Also, the interpretation of data is hoped to show these students how their own background, race, or class, would potentially affect the tuition price of schools where students like them are the majority or minority student. The data used was collected by the Department of Education in the United States. It originally illustrates a large amount of predictor variables, but has been condensed for this project to predict tuition price from the aforementioned predictors.

Additionally, direct comparisons made with degree type, race, and pell grants distributed highlights tendencies of different schools to value who studies what field. This information is important when considering social and financial motives for different groups of people pursuing different fields of study; and, because of their monetary values in specific fields, some schools will be seen as more attractive to specific groups of people.

Furthermore, thinking and making predictions about why different groups of people gravitate towards different schools and fields of study can help address and deconstruct biases in higher education institutions. This matters because of the inherent pedestalization and power of universities to create hierarchies within the different fields of study, and the power of universities to decide what is worth studying and who is worth accepting into their school.

Data

The data explored was collected in the College Scorecard by the Department of Education in the US. Its' full content spans nearly 20 years of research and covers nearly 6000 institutions. For this project, the data's predictors have been reduced into several variables in order to better interpret the data. It was necessary to combine variables that separated data between public and private colleges as well as other categories that were simply inverses of one another. To better fit the question, the category of race was broken into the variables White, Black, Hispanic, and Asian. These predictors were measured as percentages of students within the ethnic group who attend the institution. To create the category of degree type, the percentages of degrees given out in a specific major from a university were grouped into larger categories. These predictors are measured as the ratio of degrees awarded in the specific type compared to the number of degrees awarded in general at the institution. This process allows for the categories in the original data of computer science, engineering, mathematics, and physics to all be grouped into one variable: mathematics. This same process was followed for the other degree type variables as follows:

- Technology: architecture, military technology, security, construction, mechanics, precision production, transportation
- Life Science: agriculture, natural resources, medicine, biology, science technology, psychology
- Language: journalism, communication, linguistics, english
- Human Studies: culinary, liberal arts, interdisciplinary, fitness, health, arts
- Social Services: education, public administration, social science, law, business

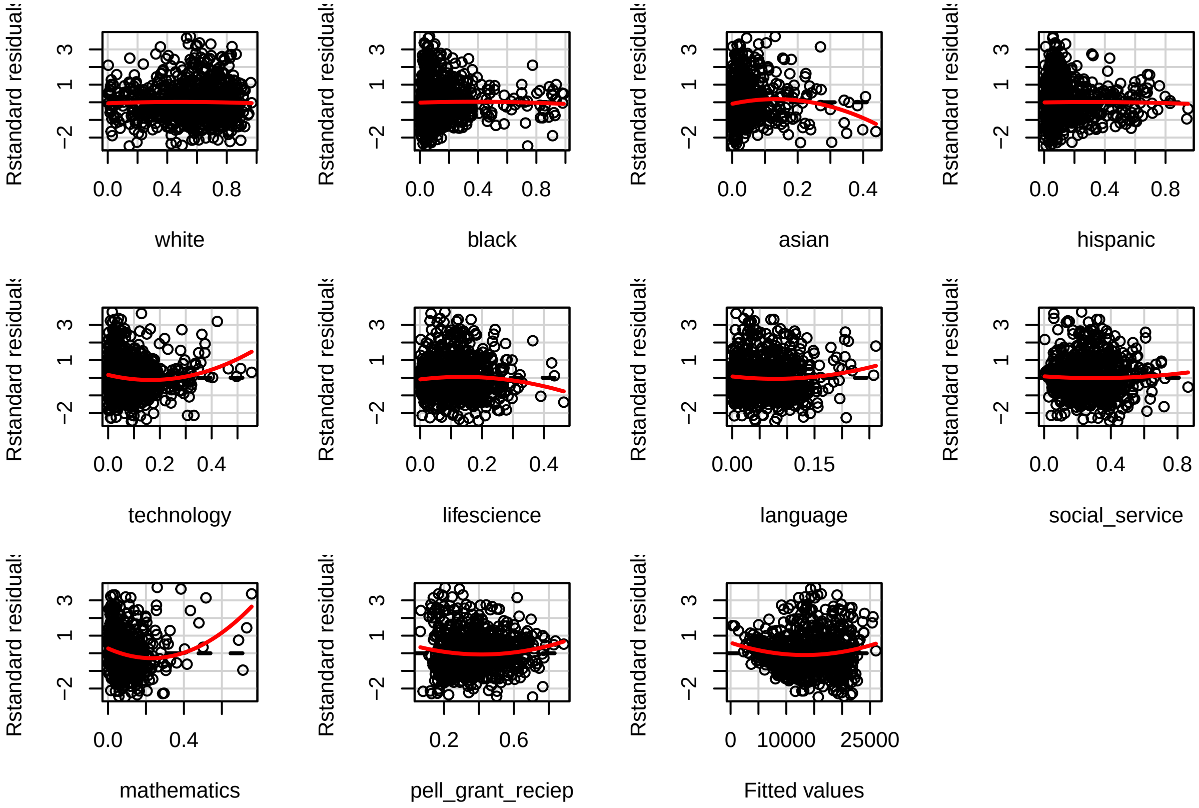
Some institutions that were removed from the data before modeling were schools in Puerto Rico and schools with a negative average net tuition. Many universities were also removed from the dataset because they did not have any data reported. The columns Pacific Islander and Native American were removed from the race category since they had a lot of missing values that did not allow a regression to be successfully run. Additionally, some columns from the degree category were also removed because of the missing values. At least one variable for every overarching category had to be removed from the model so that there were not sums of 1 for any category.

Beginning Model

The model we will look at is as follows:

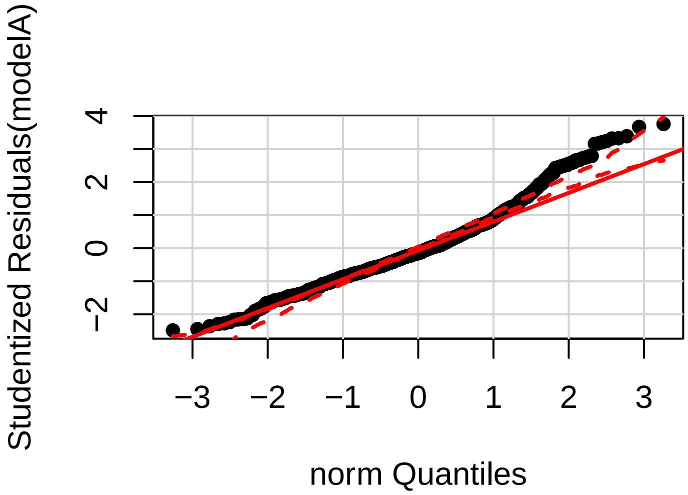
$$\begin{gathered} Average\ Net\ Tuition = \beta_0 + \beta_1 white + \beta_2 black + \beta_3 asian \\ + \beta_4 hispanic + \beta_5 technology \beta_6 lifescience + \beta_7 language \\ + \beta_8 socialservice + \beta_9 mathematics + \beta_{10} pellgrantrecipients \end{gathered}$$

In order to complete cross validation we split the data set into two smaller independent sets. The training set has 900 observations, which accounts for about 70% of the full data set. The test set has 379 observations, which account for about 30% of the full data set.



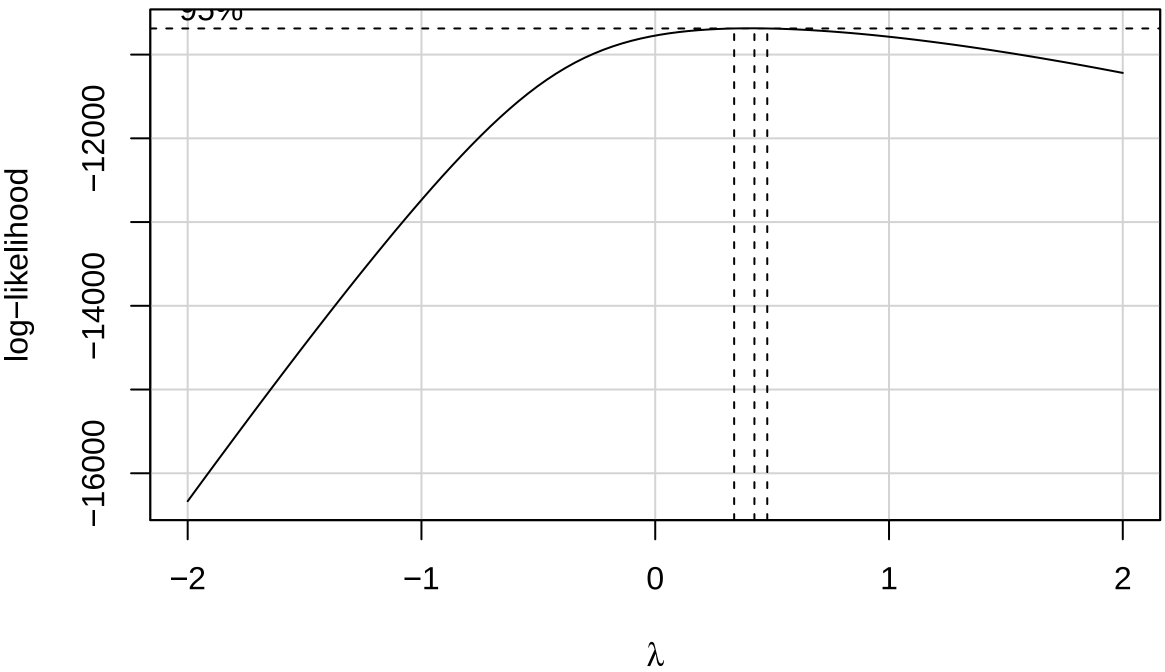
```
##                     Test stat Pr(>|t|)
## white                  -0.541    0.589
## black                  -0.520    0.603
## asian                  -3.296    0.001
## hispanic               -0.270    0.787
## technology              3.728    0.000
## lifescience            -2.106    0.036
## language                2.049    0.041
## social_service          0.920    0.358
## mathematics             6.312    0.000
## pell_grant_reciep       2.989    0.003
## Tukey test              3.388    0.001
```

As seen from the residual plots, the raw data shows a “megaphone” pattern specifically in the variables of Black, Asian, Hispanic, technology, and mathematics. The other plots do not seem to require transformation so they will be left the way they are. The added variable plots and the marginal model plots also confirm the claim stated above, so transformations should be made.



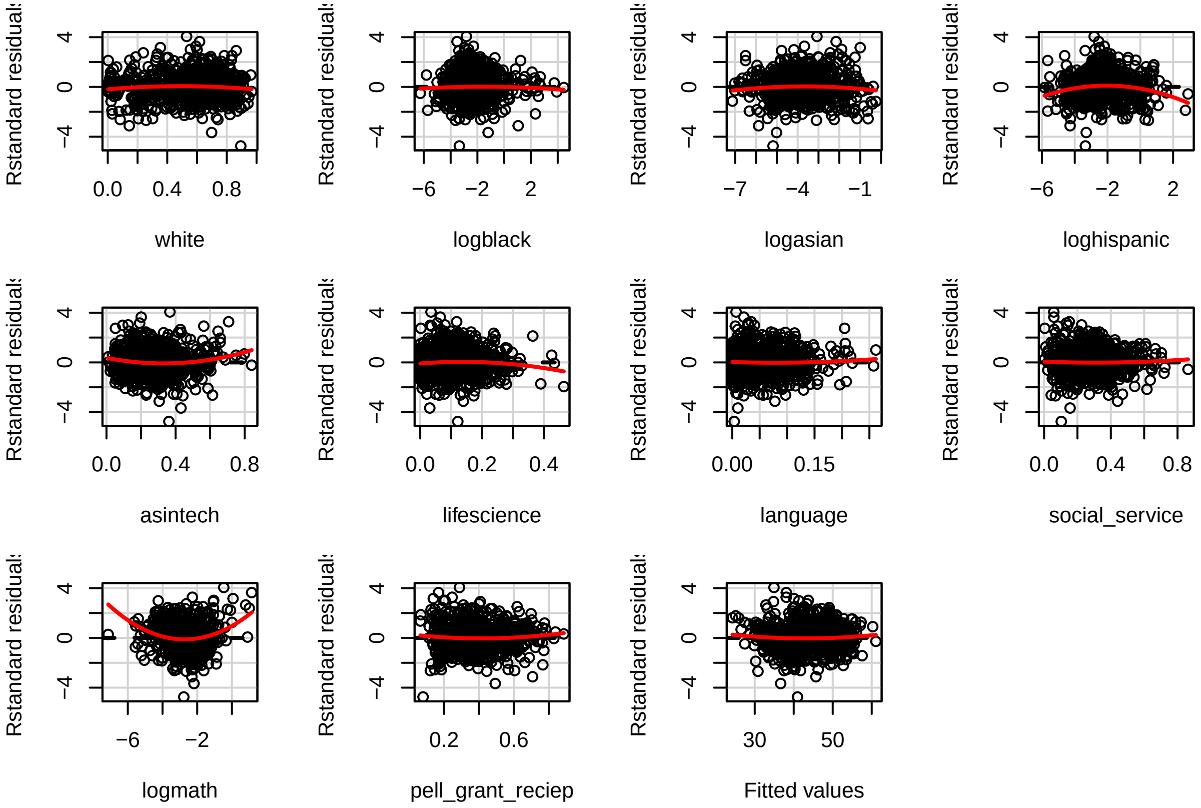
Based on the qqplot, the normality assumption has been violated since the raw data model is heavy tailed.

Transforming the Beginning Model



```
## bcPower Transformation to Normality
##    Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
## Y1    0.4072   0.0361           0.3365           0.4779
##
## Likelihood ratio tests about transformation parameters
##                              LRT df pval
## LR test, lambda = (0) 173.8076  1    0
## LR test, lambda = (1) 201.4231  1    0
```

The Box-Cox transformation suggests a $\lambda = 0.4072$. However, this model will round the suggested λ to 0.4 to make the model easier to read. This is acceptable becuase 0.4 is still in the interval of $(0.3365, 0.4779)$.



```
##                    Test stat Pr(>|t|)
## white                 -2.126    0.034
## logblack              -0.946    0.344
## logasian              -1.901    0.058
## loghispanic           -5.465    0.000
## asintech               3.269    0.001
## lifescience           -1.967    0.049
## language               0.808    0.419
## social_service         0.752    0.452
## logmath                7.244    0.000
## pell_grant_reciep      1.730    0.084
## Tukey test             1.301    0.193
```

This model requires the use of the $sin^{-1}(\sqrt{x})$ and the $log(\frac{x}{1-x})$ transformations since the data uses percentages. Each predictor variable in question was transformed using both of the transformations and the better of the two was picked for the new model. Again, the untouched variables are left the way they were since they their plots show no concern and do not require any transformation. The constant variance assumption is no longer being violated. The qqplot also shows that the assumption of normality is no longer being violated.

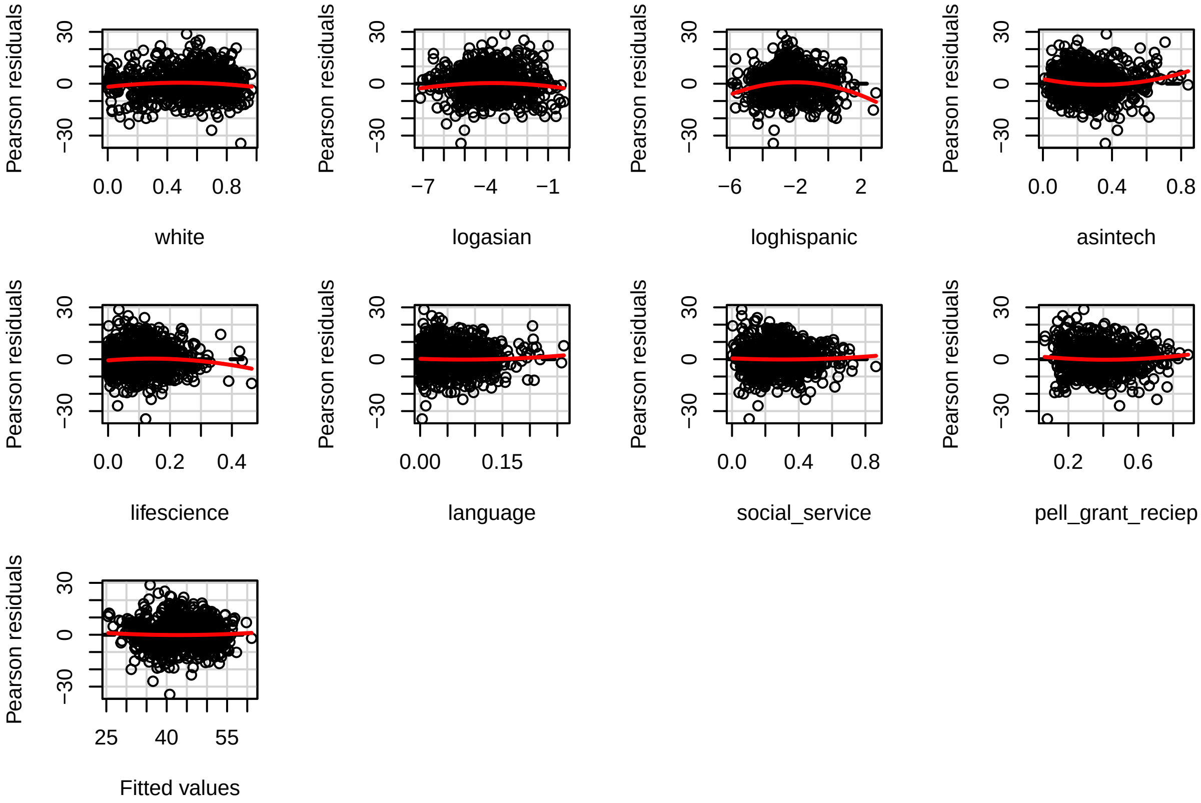
```
##             white          logblack          logasian       loghispanic
##          3.409834          2.555589          1.851716          2.738479
##          asintech       lifescience          language    social_service
##          1.514808          1.385510          1.574489          1.453627
##           logmath pell_grant_reciep
##          1.152891          1.806486
```

Additionally, checking multicollinearity from the variance inflation factor shows that none of the values exceed 5, so there are no suspicious variables.

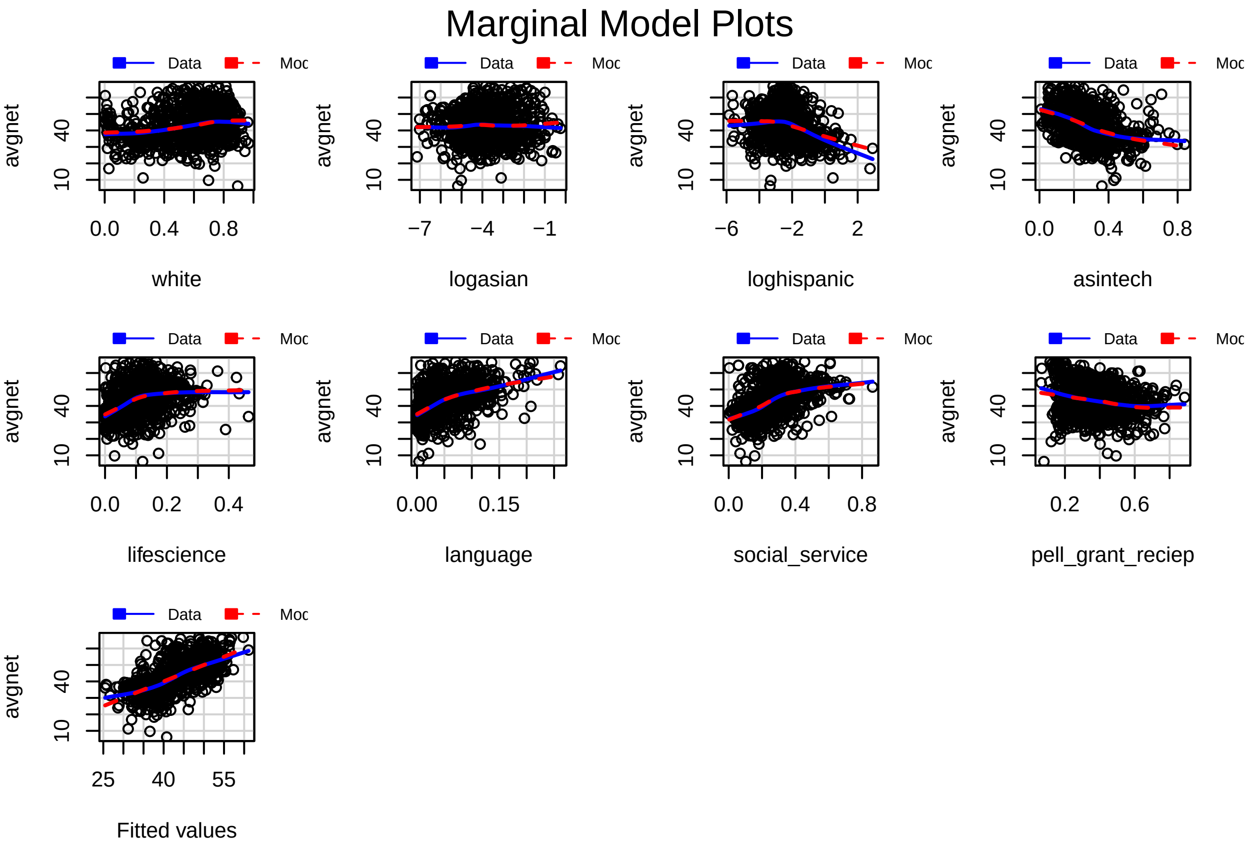
```
##
```

```
## Call:
## lm(formula = avgnet ~ white + logasian + loghispanic + asintech +
##     lifescience + language + social_service + pell_grant_reciep,
##     data = train)
##
## Coefficients:
##       (Intercept)              white           logasian
##           33.0828             4.2914             0.6037
##       loghispanic           asintech        lifescience
##           -1.5703            -6.0405            16.4444
##          language     social_service  pell_grant_reciep
##           53.5568            21.9240            -7.1132
```

Since the assumptions from the transformed model are not violated, a stepwise procedure can be used. The procedure yields a model in which the mathematics and black variables have been removed. This new model will be fitted to the training data set.



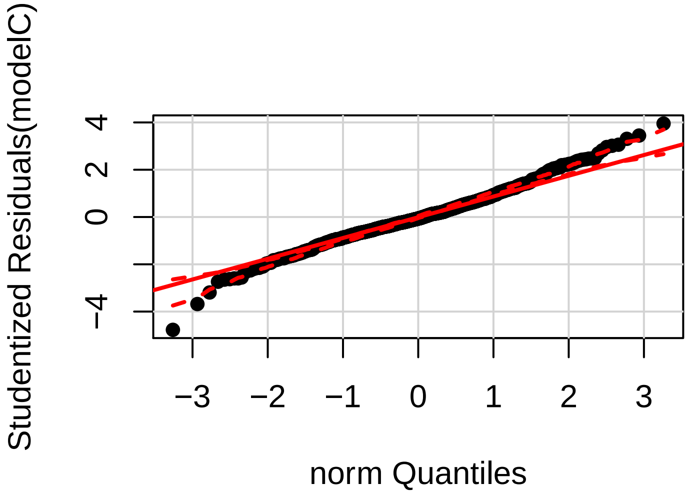
```
##                   Test stat Pr(>|t|)
## white                -2.538    0.011
## logasian             -2.279    0.023
## loghispanic          -5.474    0.000
## asintech              3.279    0.001
## lifescience          -2.029    0.043
## language              0.924    0.356
## social_service        0.817    0.414
## pell_grant_reciep     1.597    0.111
## Tukey test            0.761    0.447
```



The new residual plots for this model shows better constant variance, and no “megaphone” patterns can be arguably seen. The residual plots looks similar to the transformed variable residual plots except now two variables have been removed, which has not affected the plots significantly. Additionally, the marginal model plots confirm our model since the model and data follow the same trend.

```
##             white          logasian       loghispanic          asintech
##          1.975963          1.766702          1.834047          1.514061
##       lifescience          language    social_service pell_grant_reciep
##          1.353003          1.572246          1.365685          1.726768
```

Checking the variance inflation factors for multicollinearity, none of the variables show any suspicion since none of the factors are greater than the cutoff of 5.



Checking once again for normality, it can be seen that this plot holds a better normal model than the two models before this one.

```
##   r.squared adj.r.squared    sigma statistic       p.value df    logLik
## 1 0.4691615     0.4631903 5379.491   78.5709 4.534357e-115 11 -9002.825
##        AIC      BIC    deviance df.residual
## 1 18029.65 18087.28 25726705369         889
```

```
##   r.squared adj.r.squared    sigma statistic       p.value df    logLik
## 1 0.4558666      0.450981 7.380901  93.30825 2.576248e-112  9 -3071.528
##        AIC     BIC deviance df.residual
## 1 6163.056 6211.08 48539.63         891
```

While the adjusted r-squared value has decreased from the untouched model's value, the difference between the two values is fairly small, and now the multiple linear regression model assumptions are not violated.

Proposed Final Model

$$Average\ Net\ Tuition^{0.4} = \beta_0 + \beta_1 white + \beta_3 log(asian)$$
$$+\beta_4 log(hispanic) + \beta_5 arcsin(\sqrt{technology}) + \beta_6 lifescience + \beta_7 language$$
$$+\beta_8 socialservice + \beta_9 log(mathematics) + \beta_{10} pellgrantrecipients$$

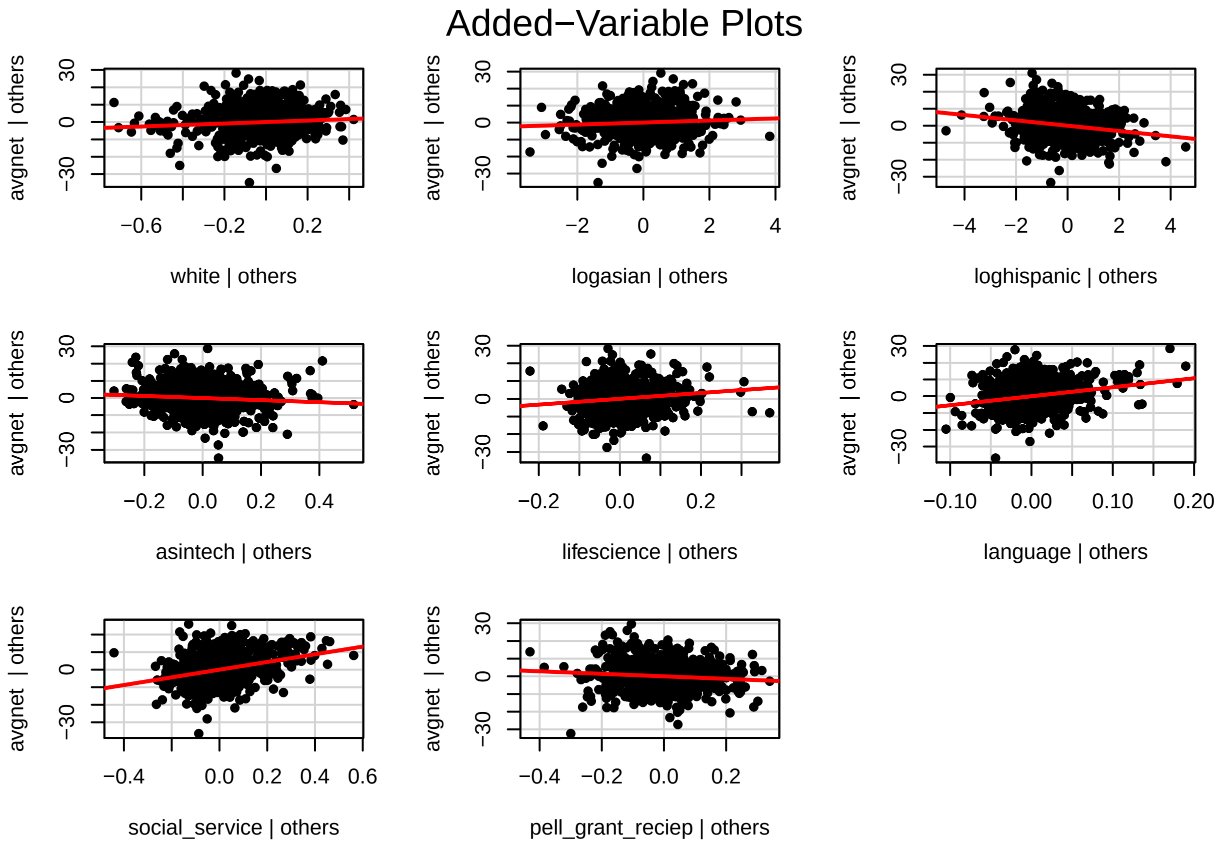
Test Set

Now that a model has been found for the training data set, it must be checked for validity on the test set. Doing so yields the same results as found with the training data set. The residual plots show spreading patterns that indicate constant variance, and the added variable plots look linear.

Final Model

The final model yields the following summary:

```
## 
## Call:
## lm(formula = avgnet ~ white + logasian + loghispanic + asintech + 
##     lifescience + language + social_service + pell_grant_reciep, 
##     data = train)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -34.521  -4.417  -0.449   4.244  28.816 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)        33.0828     1.5694  21.080  < 2e-16 ***
## white               4.2914     1.5598   2.751  0.00606 ** 
## logasian            0.6037     0.2755   2.192  0.02866 *  
## loghispanic        -1.5703     0.2571  -6.107 1.51e-09 ***
## asintech           -6.0405     2.3028  -2.623  0.00886 ** 
## lifescience        16.4444     3.8461   4.276 2.11e-05 ***
## language           53.5568     6.7517   7.932 6.43e-15 ***
## social_service     21.9240     2.0895  10.492  < 2e-16 ***
## pell_grant_reciep  -7.1132     2.3158  -3.072  0.00219 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.381 on 891 degrees of freedom
## Multiple R-squared:  0.4559, Adjusted R-squared:  0.451
## F-statistic: 93.31 on 8 and 891 DF,  p-value: < 2.2e-16
```



The findings can be analyzed as follows. Percentage of mathematics degree type, White students, and Asian students, had little to no trend on the response variable. In interpreting our model, this means that as the percent of Asian students, White students, and degrees awarded in the mathematics type increases, there is no change in the average price a student pays to go to that university. This can justified reasonably for race when taking into consideration that the majority of students who attend college are White or Asian. A negative trend in data was found in the model for the percent of Pell Grants awarded, degree in technology type awarded, and hispanic students attending. This negative trend is clear in Pell Grants, that as the number of students who receive federal aid increases, the average net tuition price of an institution should decrease. Culturally, a negative trend in percent of Hispanic students attending an institution makes sense because of the historical disenfranchisement of hispanic communities in America. If an ethnic community has been historically targeted or marginalized, they currently may either need more aid from higher education institutions due to being in a lower socioeconomic class or they may not have the ability to attend expensive higher educational institutions because of their high tuition prices. Positive trends in data were found in three different degree types: language, life science, and social service. This positive trend can be explained easily when considering real world applications. When students show greater commitment to a degree area, it is in the best interest of the institutions to put resources into that field. Thus, a greater push for resources into a specific field would drive tuition prices up for students to fund that interest.

Limitations

The model shows how the distribution of degree type, race, and Pell grants distributed affect the average net tuition price of universities in the U.S., with an aim to to explore any biased monetary structure placed within higher education institutions. Limitations in the model arise when accounting for the author's personal bias when grouping degree types and races. Error may be found in the groupings of degree types and races when considering the personal biases of the authors and their groupings. For example, if another group of statisticians grouped the thirty degrees into seven degree types, their degree type categories would most likely look different than the ones used in this study. Limitations can also be found in the data removed from the original data set. As explored earlier, the data set does not include the original categories of Native American, Pacific Islander, Mixed Race, and Other. Earlier the decision to exclude these was defended by the low percentage and missing data from these categories in the College Scorecard data. After further examination, it may have been a mistake to exclude these categories because it could reinforce the bias that we were hoping to examine; a better choice may be to group these four categories together into an Other group. Yet, the decision remains to keep these groups removed from our data because many institutions had missing values in the four removed categories and the effect that they would have on our final would have been quite small. The last limitation can be found in the the interpretability of the model. For readers outside of the statistics community, the response variable may be hard to interpret with the transformation we applied.

Conclusion

The explorations and discoveries made in this study are relevant and matter because they empower students to take ownership of their education by being conscientious of their perceived value in an educational environment. Furthermore, it is important to understand how these hierarchies and biases might imply the inclusions and exclusions present in higher education. Usage of the findings of this study create an opportunity to deconstruct the uneven playing field for targeted and marginalized demographics; so that the educational environment can be rebuilt in such a way that allows for a more even and diverse discourse about who benefits most from higher education.

In order to rebuild an environment in such a way that allows for a more even and diverse discourse about who benefits most from higher education.

Reference:

United States Department of Education. (2014). Most Recent Data [Data Set]. Retrieved from https://collegescorecard.ed.gov/data/