FINAL PROJECT

BANA 7052 – LINEAR REGRESSION

FALL 2019



Submitted by:

Abhijith Antony

Hridhay Mehta

Puneet Bhatia

Sahit Koganti

Vishnu Vijayakumar

## 1. Introduction

Alumni donations are an important source of revenue for colleges and universities. If administrators could determine the factors that influence increases in the percentage of alumni who make a donation, they might be able to implement policies that could lead to increased revenues. Research shows that students who are more satisfied with their contact with teachers are more likely to graduate. We have taken a dataset containing information of 48 national universities and studied how the different factors affect the alumni giving rate. We have implemented a multiple linear regression model to answer this question.

## 2. Data Description

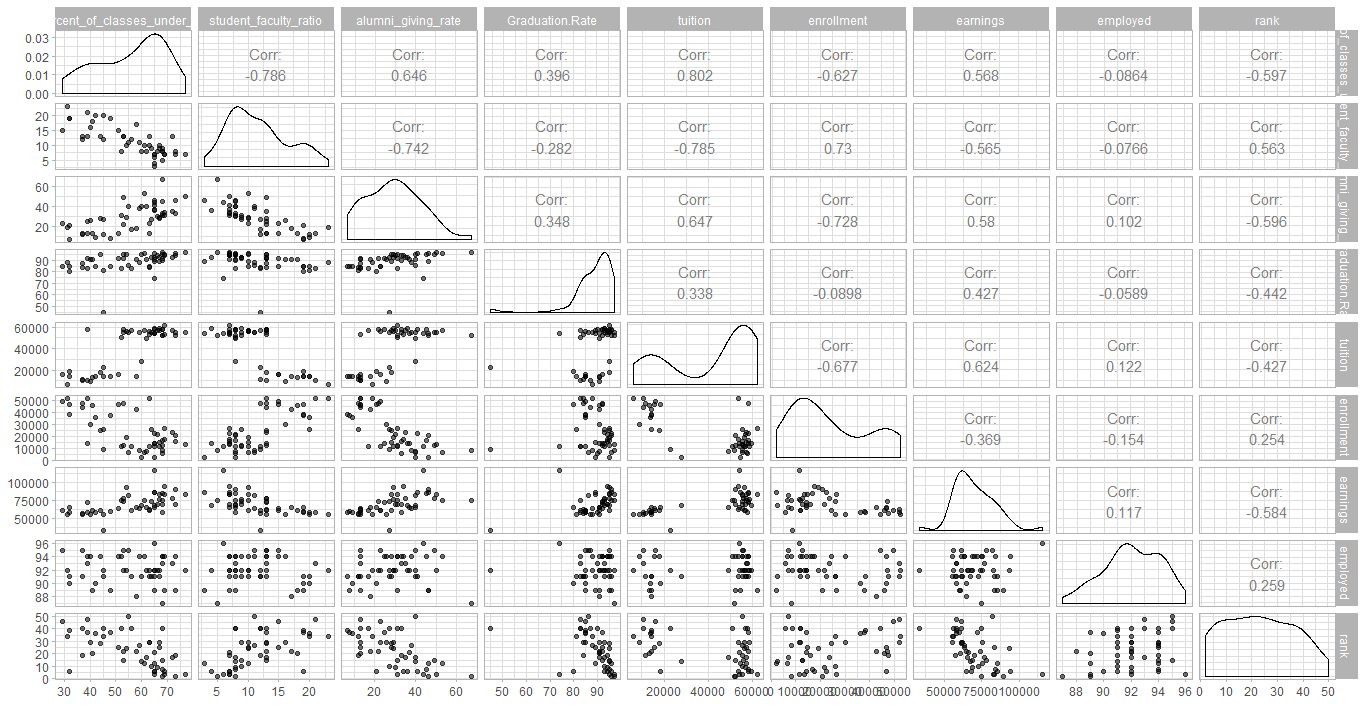
Alumni data had the following 5 variables:

* School: Name of the school
* percent\_of\_classes\_under\_20: The percentage of classes offered with fewer than 20 students
* student\_faculty\_ratio: The ratio of the students enrolled to the number of faculty in school
* alumni\_giving\_rate: The percentage of alumni that donated to the university
* private: This is an indicator variable indicating if the school is a private (1) or public institute (0)

To improve the prediction performance, we added the following variables:

* Graduation Rate: Percent of students graduating
* Tuition: Fees for the college
* Enrollment: Total number of enrolled students
* Earnings: Median earnings 6 years after graduation
* Employed: Percentage of students employed 2 years after graduation
* Rank: National University rankings

Figure 1: Correlation and data plots for each data are shown below.



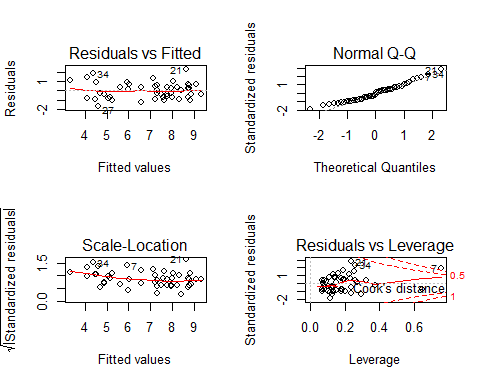
## 3. Modeling

1. **Multi-collinearity**: Starting with a basic multiple linear regression model we find that the tuition variable was causing multi-collinearity and so was dropped. This model had all Variation Inflation Factors < 5.
2. **Transformations**: The error terms seem to have some unequal variances so we try square root, log and box-cox transformations on the response variable. The Box-Cox transformation shows the best and RMSE.

## private percent\_of\_classes\_under\_20   
## 4.885532 4.321744   
## student\_faculty\_ratio Graduation.Rate   
## 4.570834 1.455548   
## enrollment earnings   
## 2.838109 2.084623   
## rank employed   
## 2.461304 1.320899

1. **Diagnostics**: We do diagnostic checking on the residuals of the Box-Cox transformed model.

Figure 2: Diagnostics of residuals



1. **Model selection**: We check for BIC values of the 10 best subsets of size 6 and find that with 3 predictors we achieve good BIC and values. We also performed backward, forward and stepwise selection techniques and summarized the results for each model.

Figure 3: BIC values

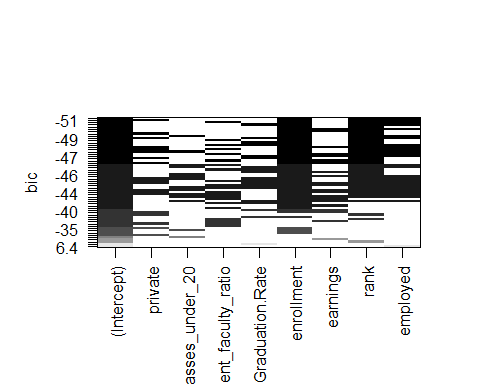
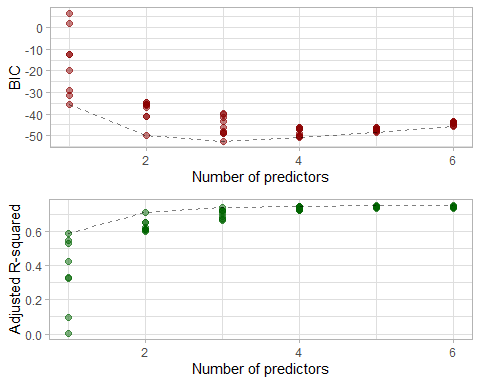


Figure 4: BIC & Adj. R-Squared vs. Number of Predictors



## be\_1 be\_2 fs\_1 fs\_2 ss\_1 ss\_2  
## AIC 137.478 68.894 137.809 102.652 137.478 68.894  
## BIC 146.834 117.546 149.036 119.492 146.834 117.546  
## adjR2 0.741 0.950 0.744 0.883 0.741 0.950  
## RMSE 0.954 0.417 0.948 0.640 0.954 0.417  
## PRESS 51.581 21.332 52.467 24.334 51.581 21.332  
## nterms 4.000 25.000 5.000 8.000 4.000 25.000

##   
## Call:  
## lm(formula = AGR2 ~ student\_faculty\_ratio + enrollment + rank +   
## employed + private + rank:employed + student\_faculty\_ratio:enrollment,   
## data = alumni\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.99872 -0.50157 0.02201 0.44597 1.23820   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 2.599e+01 6.926e+00 3.752 0.000556  
## student\_faculty\_ratio -5.590e-02 6.146e-02 -0.909 0.368535  
## enrollment -1.244e-04 2.163e-05 -5.752 1.05e-06  
## rank -2.022e+00 3.077e-01 -6.570 7.48e-08  
## employed -1.769e-01 7.671e-02 -2.306 0.026363  
## private 1.081e+00 3.892e-01 2.776 0.008320  
## rank:employed 2.112e-02 3.303e-03 6.394 1.32e-07  
## student\_faculty\_ratio:enrollment 4.999e-06 1.611e-06 3.103 0.003505  
##   
## (Intercept) \*\*\*  
## student\_faculty\_ratio   
## enrollment \*\*\*  
## rank \*\*\*  
## employed \*   
## private \*\*   
## rank:employed \*\*\*  
## student\_faculty\_ratio:enrollment \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6402 on 40 degrees of freedom  
## Multiple R-squared: 0.9005, Adjusted R-squared: 0.8831   
## F-statistic: 51.73 on 7 and 40 DF, p-value: < 2.2e-16

## 4. Results

The above is the box-cox transformed output variable. In order to select a parsimonious model we select fs\_2. The model’s equation now becomes:

1. 88.12% variance in the output variable is explained by this model
2. Residual standard error of the model is found to be 0.6402
3. Some inferences from the equation:  
   i. All held constant, with 1 unit increase in student\_faculty\_ratio, the average AGR2 decreases by 0.056 units  
   ii. All held constant, with 1 unit increase in rank, the average AGR2 decreases by 2.02 units

**Hypothesis t-tests:**  
: = 0  
: != 0  
For all p-values < 0.05, we reject   
We can see that for all variables except student\_faculty\_ratio the p-value < 0.05. Thus those estimates are significant. For student\_faculty\_ratio, even though the is not significant but we keep it as it gives a higher .

**Hypothesis F-test:**  
: All = 0  
: At least one != 0  
As p-value < 0.05, we reject . Thus our model as a whole is significant.

## 5. References

* <https://www.niche.com/colleges/search/best-colleges/>
* <https://www.usnews.com/best-colleges/rankings/national-universities>
* <https://github.com/bgreenwell/uc-bana7052>
* RESEARCH PAPER – ‘MILLENNIAL ALUMNI GIVING: FACTORS FOR DONATING TO COLLEGES AND UNIVERSITIES’ by Yolanda Barbier Gibson