Alumni Donation Case Study

Ankush Morey

4/7/2021

Contents

Introduction

Overview 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. As a result, one might suspect that smaller class sizes and lower student-faculty ratios might lead to a higher percentage of satisfied graduates, which in turn might lead to increases in the percentage of alumni who make a donation. The dataset contains data for 48 national universities (America's Best Colleges, Year 2000 Edition).

Problem Statement: Determine key factors that influence Alumni donation rate.

Data Description Data source - https://bgreenwell.github.io/uc-bana7052/data/alumni.csv

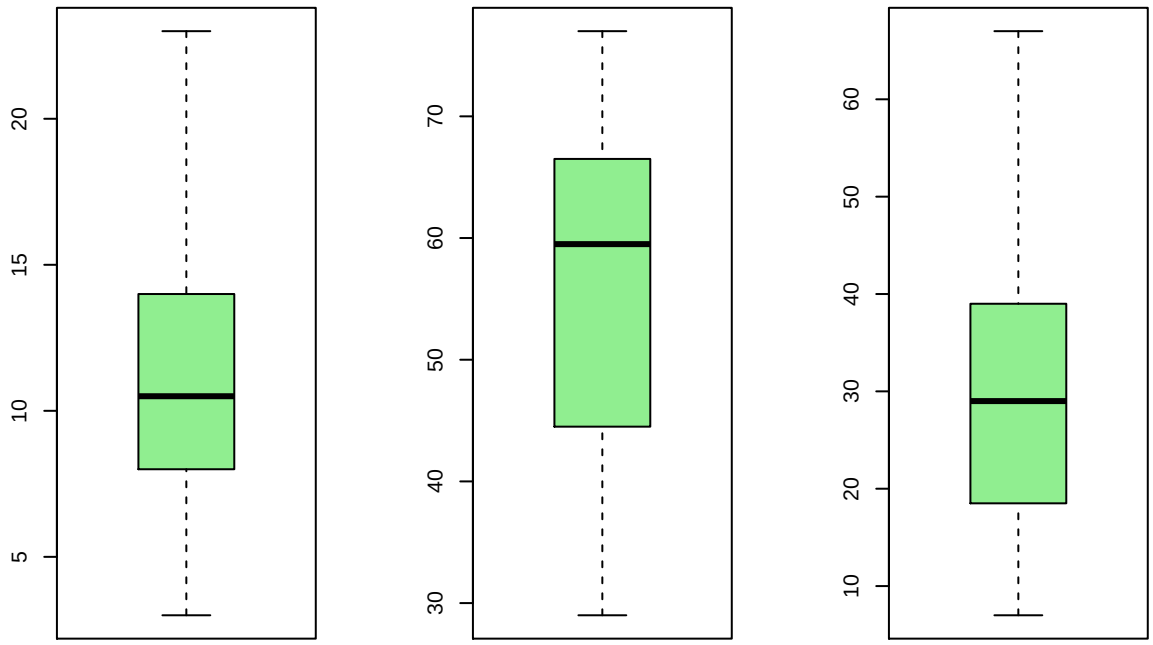
- % of Classes Under 20 - the percentage of classes offered with fewer than 20 students.(Continuous Numeric)
- Student/Faculty Ratio - the number of students enrolled divided by the total number of faculty. (Continuous Numeric)
- Alumni Giving Rate - the percentage of alumni that made a donation to the university.(Continuous Numeric)
- Private - 1 implies Private university and 0 implies Public university (Categorical Binary)

Method/Approach: This is a linear regression problem, with target variable being continuous numeric. Our objective is to determine the factors that impact our target the most. Thus we are interested in finding statistically significant Beta coefficients.

- We checked missing values and outliers
- Performed univariate and bivariate analysis
- Correlation plots for relationships between independent and dependent variables.
- Tried multiple input combinations for linear regression to arrive at best possible model.

EDA

percent_of_classes_under_20 student_faculty_ratio alumni_giving_rate Min. :29.00 Min. : 3.00 Min. : 7.00
1st Qu.:44.75 1st Qu.: 8.00 1st Qu.:18.75
Median :59.50 Median :10.50 Median :29.00
Mean :55.73 Mean :11.54 Mean :29.27
3rd Qu.:66.25 3rd Qu.:13.50 3rd Qu.:38.50
Max. :77.00 Max. :23.00 Max. :67.00
private
Min. :0.0000
1st Qu.:0.0000
Median :1.0000
Mean :0.6875
3rd Qu.:1.0000
Max. :1.0000

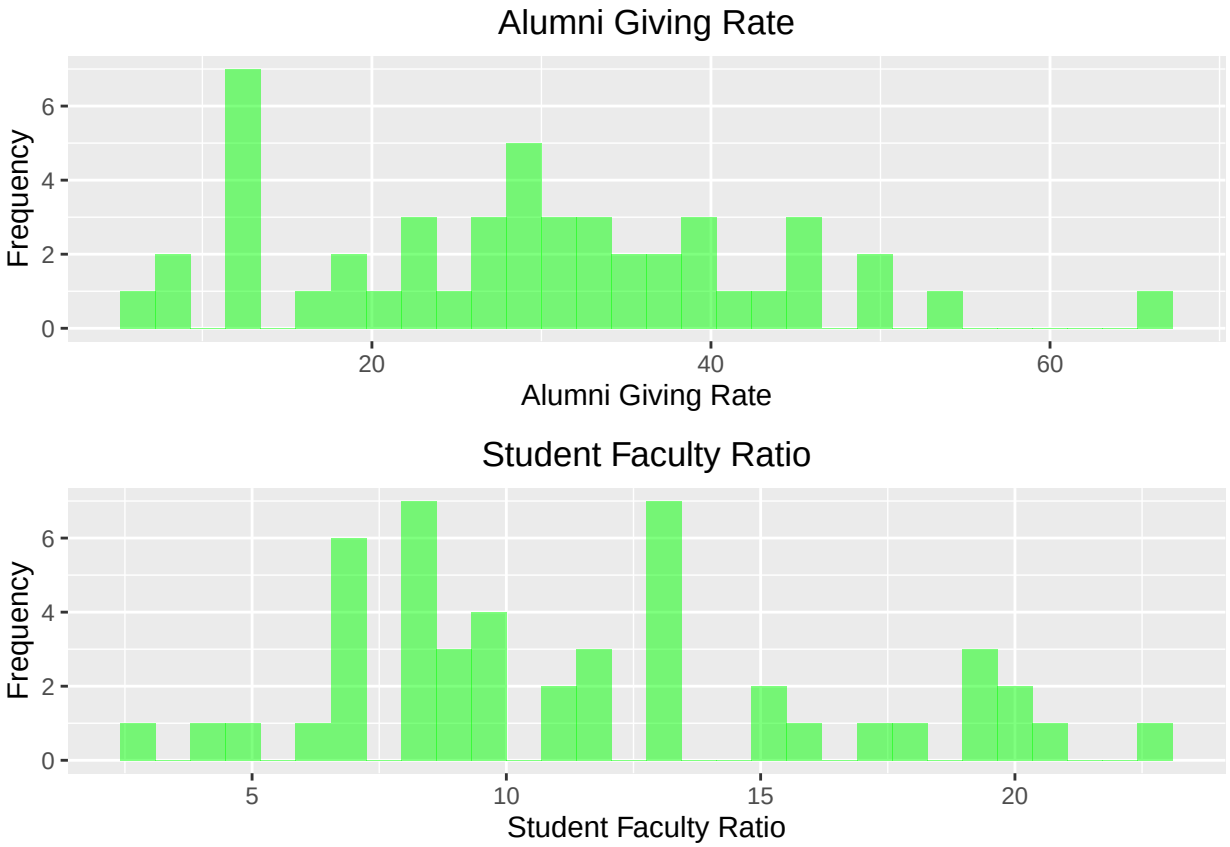


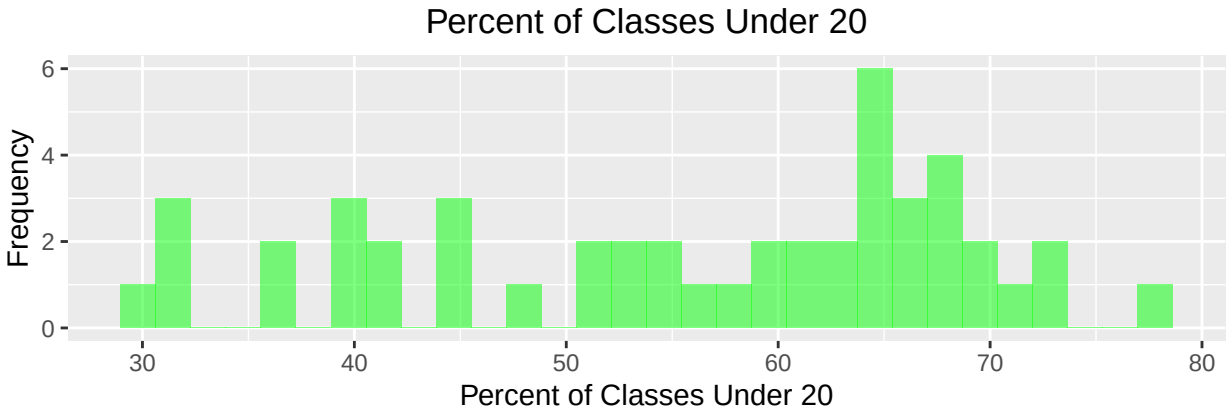
We found that the data for percent_of_classes_under_20 is distributed with mean of 55.73% and standard deviation of 13.19% with a max value of 77% and min of 29%.The median of data stands at 59.5%. There are no missing values in the dataset. We do not see any outliers. We found that the data for student_faculty_ratio is distributed with mean of 11.54 and standard deviation of 4.85 with a max value of 23 and min of 3.The median of data stands at 10.5. There are no missing values in the dataset. . We do not see any outliers.

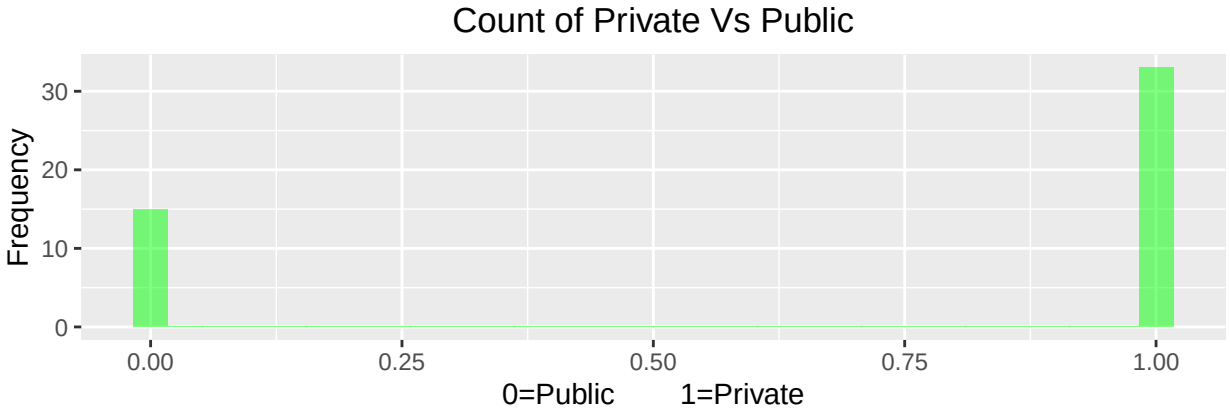
We found that the data for alumni_giving_rate is distributed with mean of 29.27% and standard deviation of 4.85% with a max value of 67% and min of 7%.The median of data stands at 29%. There are no missing values in the dataset. . We don't see any outliers.

We found that the data for private is distributed with mean of 0.6875 and standard deviation of 0.47 with a max value of 1 and min of 0.The median of data stands at 1. There are no missing values in the dataset. This is binary categorical variable.

Histograms and Bar plot

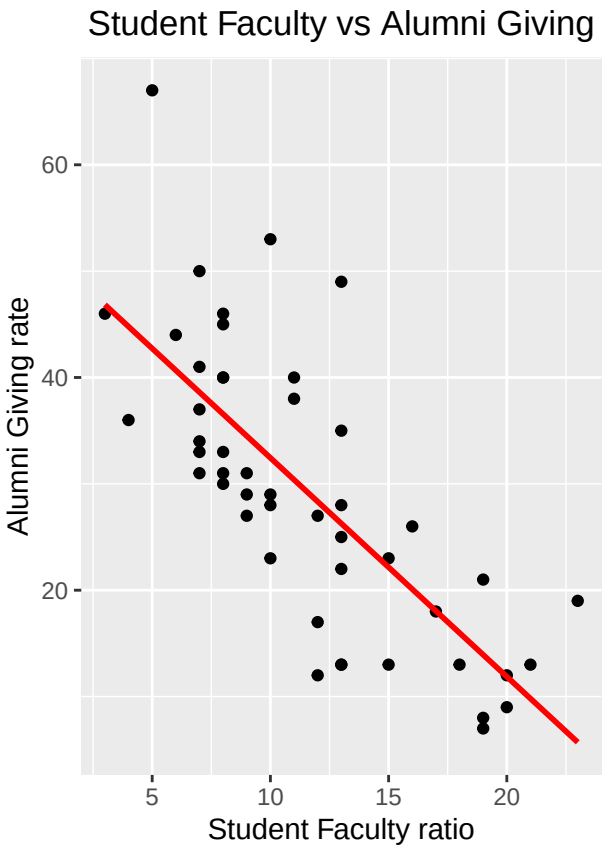


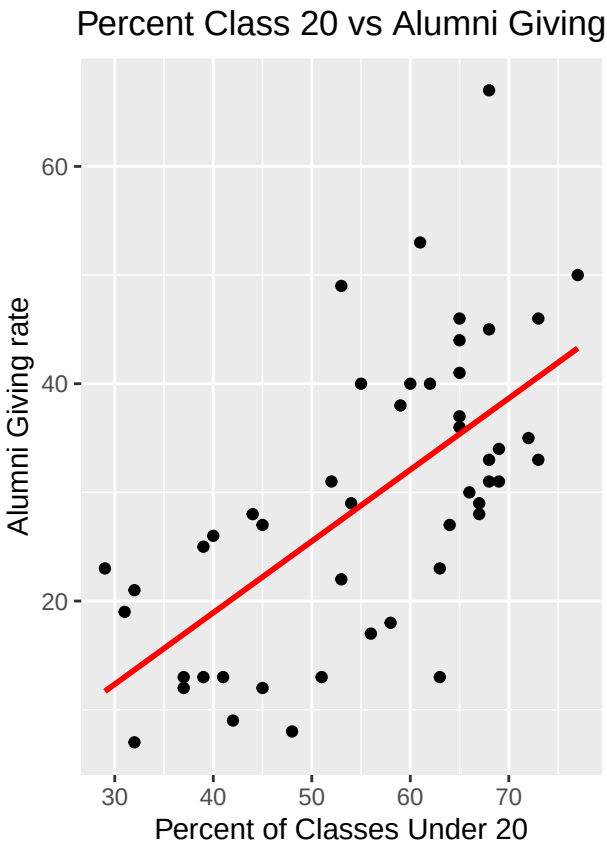




None of the variables i.e. Alumni giving rate, Student faculty ratio and Percentage class below 20 is normally distributed. Also, count of public universities = 15 and private universities = 33.

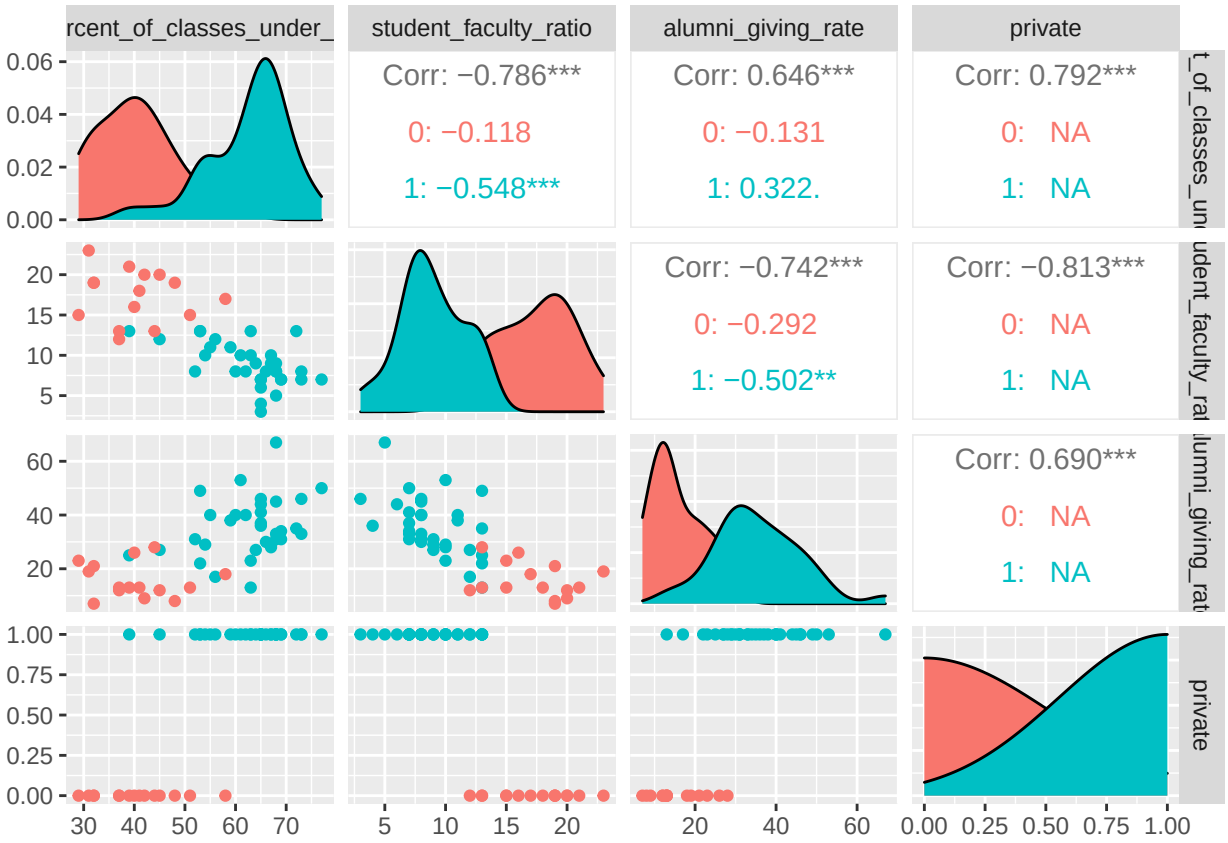
Scatter Plots





1) Alumni giving rate is highly correlated with Student-faculty ratio and class under 20.
2) Student faculty ratio has negative correlation with alumni giving rate.

Correlation Plots and Heat map Lets calculate the correlation coefficients:



1) Alumni giving rate is highly correlated with Student-faculty ratio and class under 20. Student faculty ratio has a negative correlation coefficient of -0.742 with alumni giving rate. Percentage of classes under 20 has a positive correlation of 0.646 with alumni giving rate.

2) All predictor variables have strong correlation with Response variable i.e Alumni giving Rate.

Model Building

Since we have very few data points (only 48), dividing the dataset further into train and test would affect the model training/learning. Hence we have considered whole dataset for training. And later if we have access to new data set we can test our models. But for this case study we would rely on metrics like Adjusted R squared, residual diagnostics for selecting a suitable model.

Model 1 Building 1st model with all input parameters. As all predictor variables have strong correlation matrix with Response variable, we have chosen the following predictor variables : Y = response variable= Alumni giving rate X1= predictor variable= class under twenty X2= predictor variable= student faculty ratio X3= predictor variable= Private

Call: lm(formula = alumni_giving_rate ~ ., data = df_alumni)

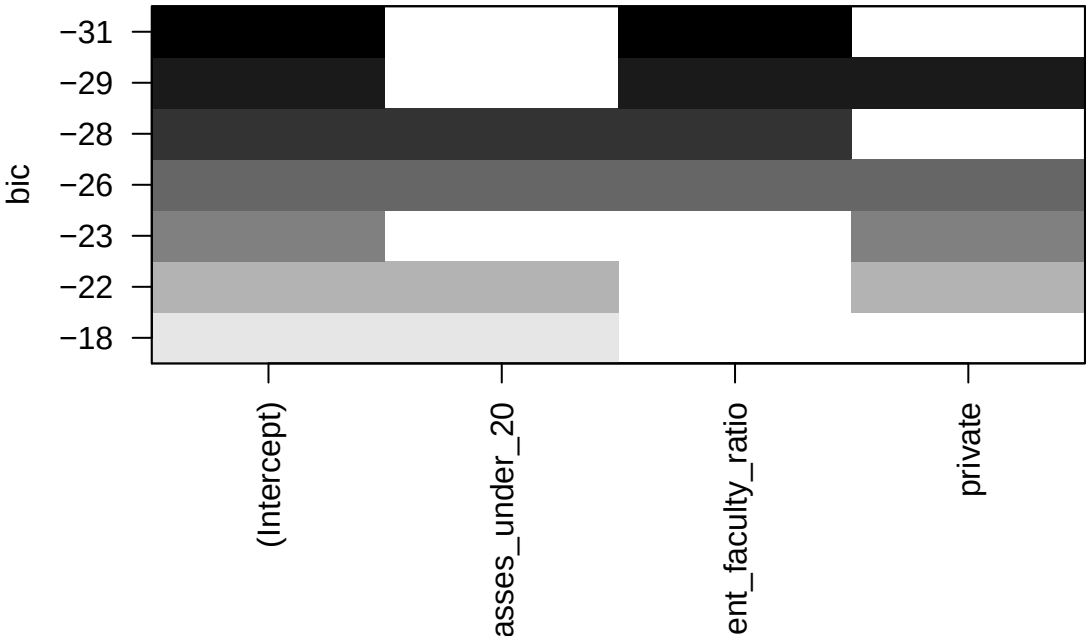
Residuals: Min 1Q Median 3Q Max -16.757 -6.320 -2.273 5.152 25.669

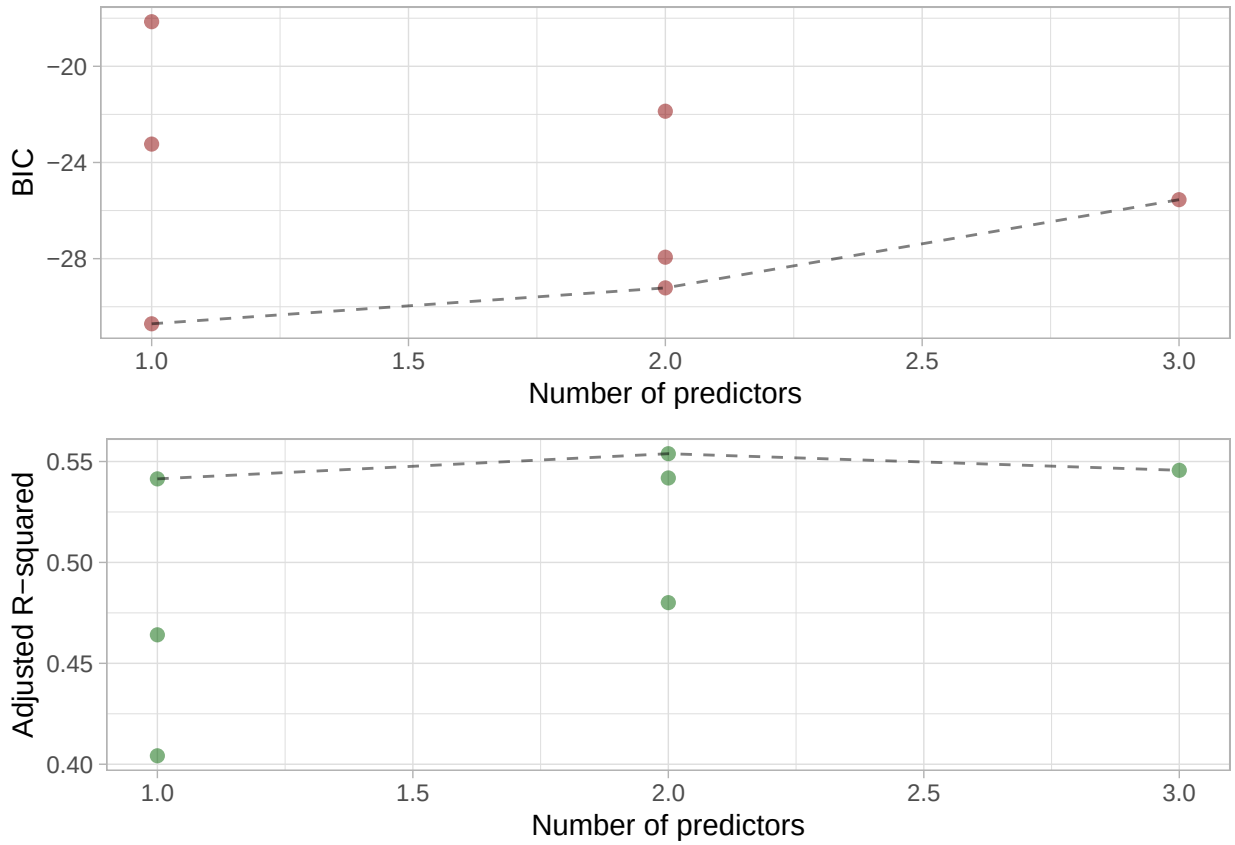
Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.78364 13.67220 2.690 0.01005 * percent_of_classes_under_20 0.07725 0.17873 0.432 0.66768 student_faculty_ratio -1.39835 0.51075 -2.738 0.00889 ** private 6.28534 5.35633 1.173 0.24693
— Signif. codes: 0 ‘’ ***0.001*** ’’ *0.01* ’’ 0.05 ‘.’ 0.1 ’ ’ 1

Residual standard error: 9.06 on 44 degrees of freedom Multiple R-squared: 0.5747, Adjusted R-squared: 0.5457 F-statistic: 19.81 on 3 and 44 DF, p-value: 2.818e-08

Based on the linear regression fitted we find that Adjusted R-squared: 0.5457 and Residual standard error is 9.06. The summary statistics of the model 1 shows that mean estimates of X1 and X3 have a p-value greater than 0.05 and therefore we cannot reject the null hypothesis of coefficients of X1 and X3 being equal to zero. P-value for estimate of X2 is less than 0.05 and therefore we can reject null hypothesis of coefficient of X2 =0. The fitted model is Y= 36.78 + 0.77 * X1 - 1.40 * X2 + 6.29 * X3

Best parameter combination using regsubsets Post this result we tried to verify the variables with help of regsubsets to get the best possible number of variables to make the model.





Based on the BIC values we conclude that the best model can be made with the help of intercept and student faculty ratio.

Model 2 Based on the inputs provided by Model 1 and regsubset analysis , we try to fit a modes as mentioned below : Y = response variable= Alumni giving rate X2 = predictor variable= student faculty ratio

Call: lm(formula = alumni_giving_rate ~ student_faculty_ratio, data = df_alumni)

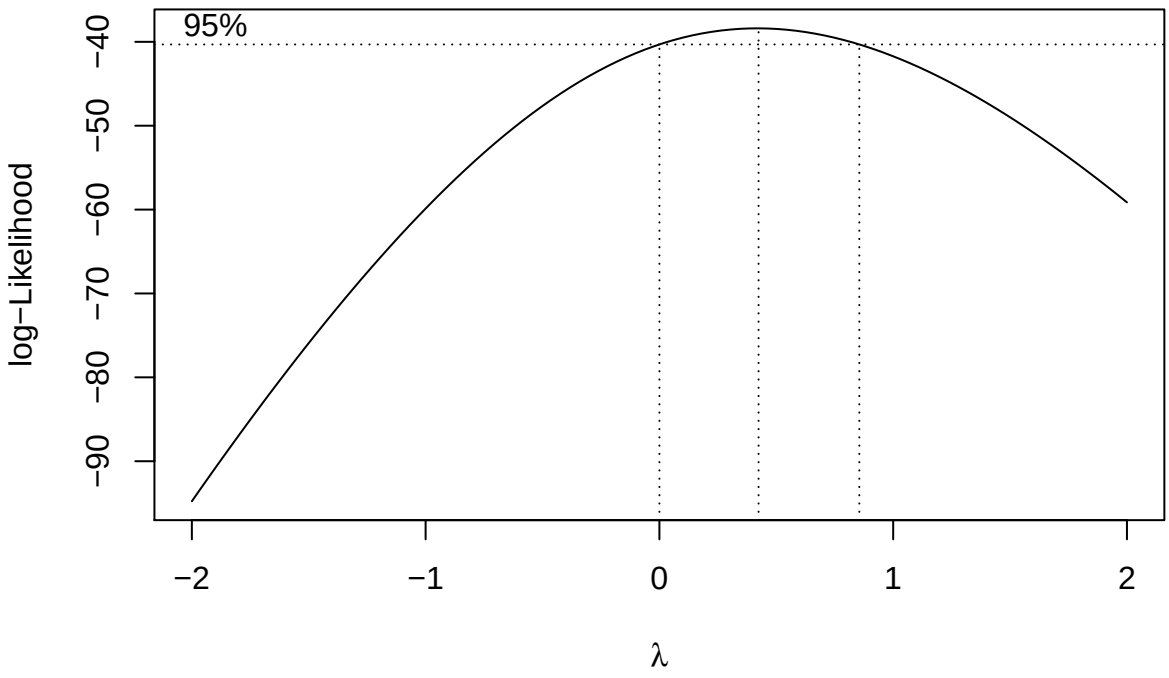
Residuals: Min 1Q Median 3Q Max -16.328 -5.692 -1.471 4.058 24.272

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 53.0138 3.4215 15.495 < 2e-16 ***student_faculty_ratio -2.0572 0.2737 -7.516 1.54e-09***
— Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.103 on 46 degrees of freedom Multiple R-squared: 0.5512, Adjusted R-squared: 0.5414 F-statistic: 56.49 on 1 and 46 DF, p-value: 1.544e-09

From above linear regression we find that Adjusted R-squared: 0.5414 and Residual standard error is 9.103. The summary statistics of the model 2 shows that mean estimate of X2 has a P-value for less than 0.05 and therefore we can reject null hypothesis of coefficient of X2 is 0.The fitted model is Y= 53.01 - 2.06 * X2

Model 3 We tried Box Cox transformation based on which we transform the response variable according to the value of lambda:



Call: lm(formula = alumni_giving_rate2 ~ student_faculty_ratio, data = df_alumni)

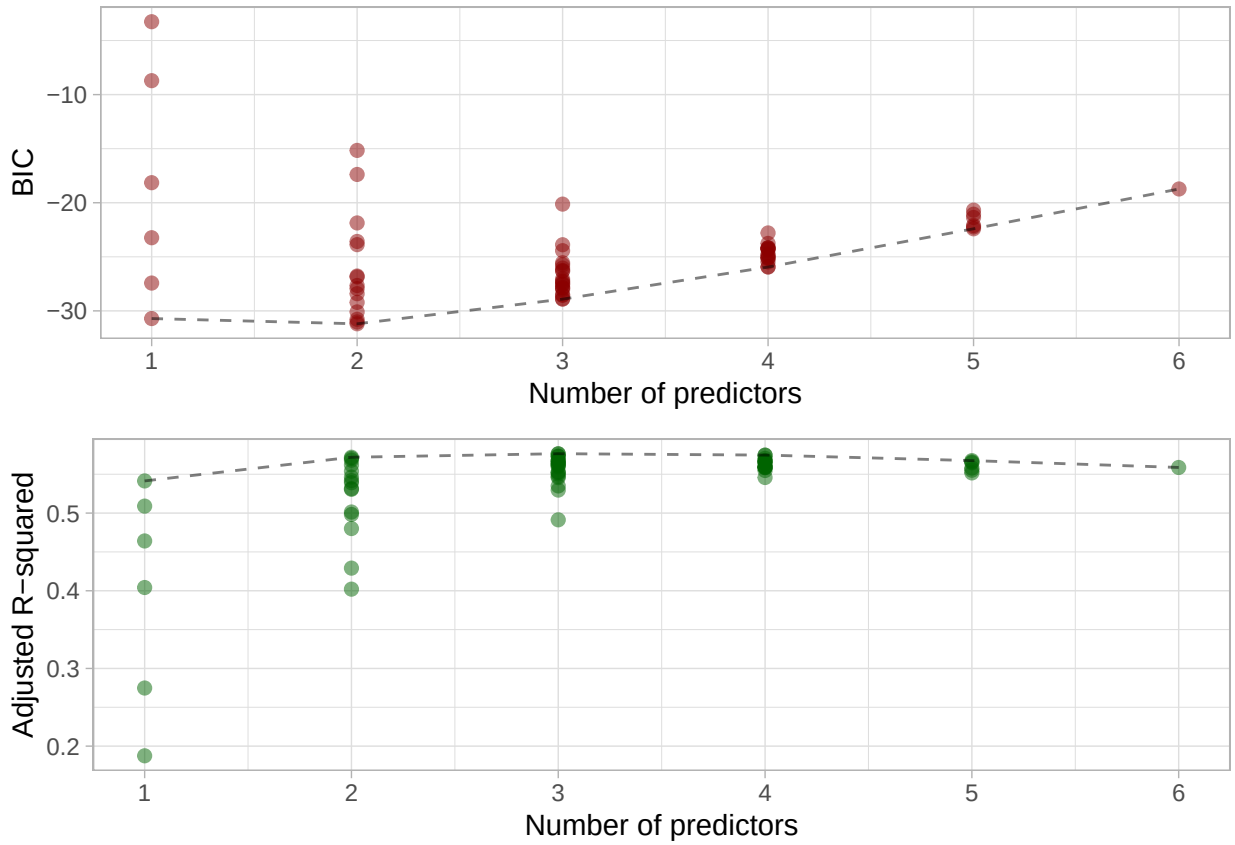
Residuals: Min 1Q Median 3Q Max -2.6870 -0.7874 -0.2050 0.6987 3.1568

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.95027 0.49036 22.331 < 2e-16 ***student_faculty_ratio -0.32134 0.03923 -8.192 1.55e-10*** — Signif. codes: 0 '*** 0.001 **' *0.01* ** 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.305 on 46 degrees of freedom Multiple R-squared: 0.5933, Adjusted R-squared: 0.5844 F-statistic: 67.1 on 1 and 46 DF, p-value: 1.546e-10 Based on the transformation done we got Lambda= 0.424 and the linear regression fitted we find that Adjusted R-squared: 0.5844 and Residual standard error is 1.305. The summary statistics of the model 3 shows that mean estimate of X2 has a P-value for less than 0.05 and therefore we can reject null hypothesis of coefficient of X2 is 0.The fitted model is Yˆ(lambda) = 10.95 - 0.32 * X2

Model 4 Lets check if there are interactions within the the independent variables that would affect our target variable

[1] "16" "1,024" "16,384" "32,768"




```
Estimate Std. Error t value Pr(>|t|)
```

(Intercept) 15.667 2.271 6.899 0.000 private 38.242 5.880 6.503 0.000 private:student_faculty_ratio -2.071 0.584 -3.546 0.001

Call: lm(formula = form, data = df_alumni)

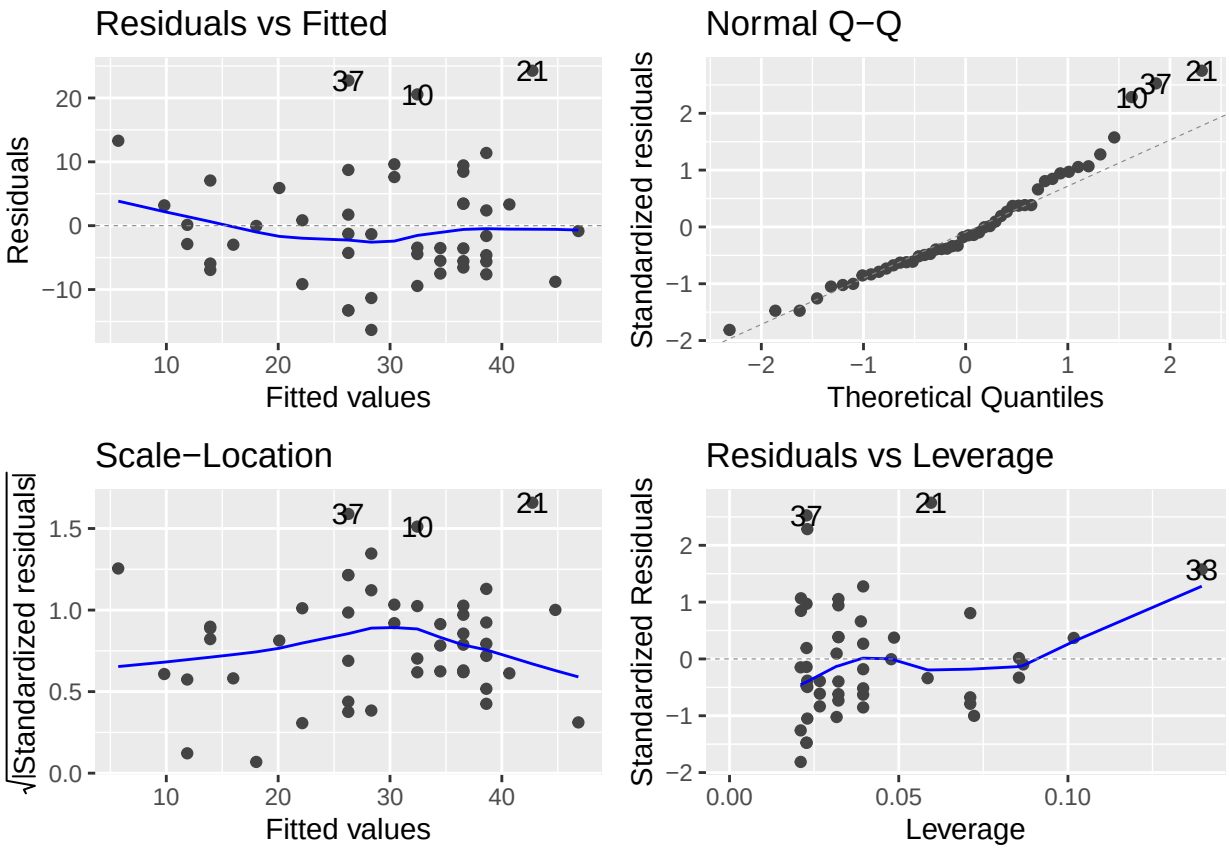
Residuals: Min 1Q Median 3Q Max -13.981 -6.284 -2.667 5.719 23.448

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 15.6667 2.2708 6.899 1.44e-08 ***private 38.2419 5.8804 6.503 5.57e-08*** private:student_faculty_ratio -2.0714 0.5841 -3.546 0.000926 *** — Signif. codes: 0 '**' ***0.001*** '*' *0.01* '' 0.05 '.' 0.1 ' ' 1

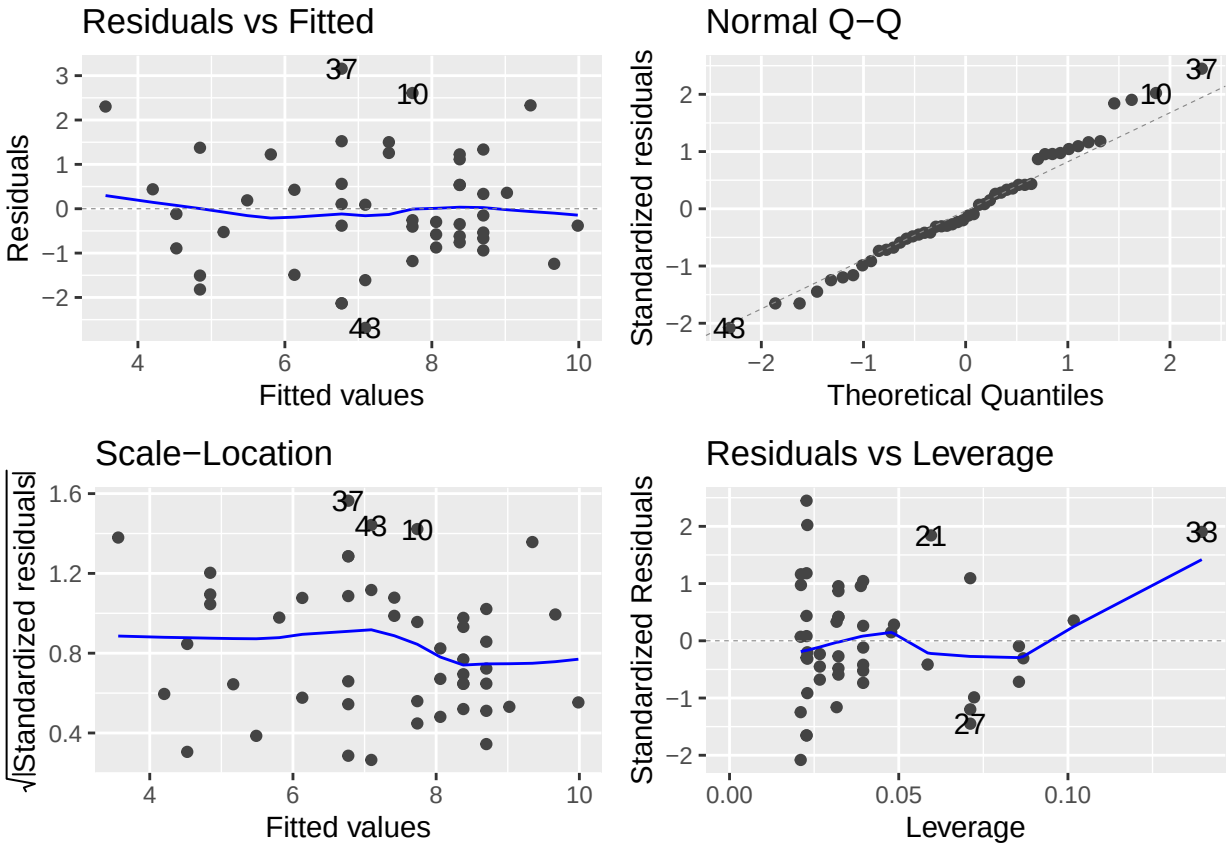
Residual standard error: 8.795 on 45 degrees of freedom Multiple R-squared: 0.5901, Adjusted R-squared: 0.5719 F-statistic: 32.39 on 2 and 45 DF, p-value: 1.929e-09 We see that there is statistically significant interaction between private and student faculty ratio. With this model we get Adjusted R-squared of 0.57 with residual error 8.8. But, the improvement in model is not that significant as compared to Box cox. moreover, interactions increase complexity of model. We would prefer a model which is simple and provides optimum prediction. Hence we will carry forward Model 2(single predictor = Student faculty ratio) and Model 3 (Box Cox transformed) for testing model assumptions.

Assumptions

Residual Diagnostics for Model 2 (Single predictor - Student faculty ratio)



Residual Diagnostics for Model 3(Box Cox transformed)



Residual Vs Fitted plot can be used to test whether the model is compliant with the assumptions that are integral to a linear regression model. The residual vs fitted plot obtained for the model 1 (with single variable Student Faculty Ratio), indicates presence of a small degree of Heteroscedasticity (non-constant variance), and the parabolic curve fitted to the residuals is indicative of a small degree of non-linearity between the independent and dependent variables. TO mitigate this effect, we can use the box cox transformation. The residuals are normally distributed, evident in the Normal QQ plot. The scale location plot is further affirmation that the residuals have non-constant variance across the range of predictors. The second set of plots are obtained for the model after box cox transformation of the data. WE can see significant improvements in the Residual vs Fitted plot, increased Homoscedasticity (Scale Location plot).

Conclusion

Model Selection As per Occam's Razor principle on problem-solving "entities should not be multiplied without necessity", or more simply, the simplest explanation is usually the right one. Here as well we try to find the simplest best possible model. Including multiple interaction terms might work well on training data but are highly likely to overfit, resulting in poor performance on new data.

Based on the Model development process followed model 2 and model 3 looks to be better . 1) When we compare these two models, we find that Adjusted R-square for Model 3 is 0.5844 compared to 0.5414 of model 2, 2) We find that Residual Standard error for model 3 is less than model 2. 3) Plot of Residual with fitted value of Model 3 is closer to constant variance compared to Model 2. 4) We also find that Model 3 Q-Q plot to be closer to normal distribution compared to Model-2 In Light of the above we would choose Model 3 as for predicting alumni giving rate.

$\hat{Y}(0.424) = 10.95 - 0.32 * X2$

Insights Hence to increase the Alumni donations universities should concentrate on improving the student to faculty ratio. This makes logical sense as well, improving student faculty ratio implies higher personalized instructions, more one-to-one sessions thus improving student engagement and experience as a whole. Such students are likely to have a deep sense of belonging to the university and same could reflect in Alumni donations.