Alumni Donation Study

First Name: Vaidiyanathan

Last Name: Lalgudi Venkatesan

I. Introduction

The intention of this case study is to determine the best model to fit the data from the Alumni Donation data. School administrators would like to understand how to best influence policies that lead to increased revenues. This problem was undertaken with the *alumni* dataset provided for the course from Brandon Greenwell's github, which in turn is supplied from *America's Best Colleges, Year 2000 Edition.* See *Data Description* for a more in-depth explanation of the data.

After exploring the dataset, modeling was done using linear regression techniques. A few models were explored, starting with simple linear regression, models with no interactions, and finally, models using all possible interactions. Residual diagnostics and transformations were combined with R's leaps() package to chose an optimal model that was easy to interpret and had a fairly high adjusted R^2 value. Our final model used *student_to_faculty_ratio* and *private* variables without any interactions:

$$alumni\_giving\_rate = 8.824 - 0.216 * student\_faculty\_ratio + 1.334 * private$$

II. Data Description

The dataset we are using has one response, or Y variable, titled *alumni_giving_rate*. This column represents a percentage (0-100) of what percent of alumni gave money for each individual university. Then the dataset also has three predictor, or independent X variables:

- *percent_of_classes_under_20:* a percentage (again, 0-100) of how many classes the university has with less than 20 students enrolled
- *student_to_faculty_ratio:* the ratio of number of students to faculty members
  private: a binary TRUE/FALSE variable that is set to 0 if the university is public (not private) and 1 if the university is private

Each of the 48 rows in the dataset represents a separate university. One additional column is present as a label, displaying the name of each university, although this is not used in analysis.

Basic Summary Statistics:

The following table (Fig 2.1) shows the most common summary statistics (measures of central tendency, dispersion, and skewness) for each of the four variables mentioned above, rounded to three decimal points:

Variable	Mean	Median	Min	Max	Range	Std.	Variance	Skewness
alumni_giving_rate	29.271	29.0	7	67	60	13.441	180.670	0.358
percent_of_classes_under_20	55.729	59.5	29	77	48	13.194	174.074	-0.485
student_to_faculty_ratio	11.542	10.5	3	23	20	4.851	13.441	0.563
Private	0.688	1.0	0	1	1	0.468	0.219	-0.809

Fig 2.1: Summary Statistics for All Variables

From Fig. 2.1 above, a few noteworthy early observations can be seen. Comparing the location of the mean relative to the median, this confirms the shape of the distributions that was calculated with skewness. The response variable and the predictor student_to_faculty_ratio both have a right skew, while percent_of_classes_under_20 has a left skew. More schools also tend to be private in the dataset than public (noticed by the skewness, as well as the mean of 0.688 being larger than 0.5)

Finally, using the method of $1.5 * IQR$, each predictor was quickly checked for any major outliers (confirmed with box plot on right, Fig. 2.2), and none were found at this stage. This topic is later revisited with residual plots.

The p value for F statistic was <0.001, indicating that at least one predictor was affecting the alumni donation rate.

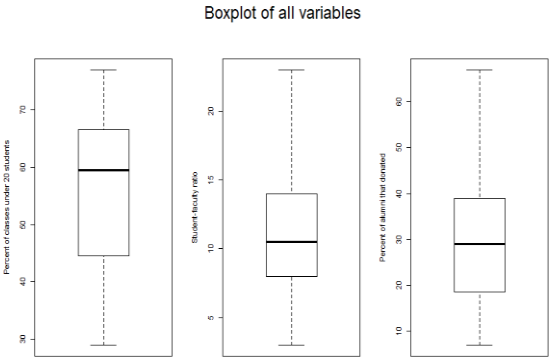


Figure 2.2: Box plots

Plots: XY Scatterplots

Figs. 2.3 and 2.4 show the change in alumni donation rate with respect to percentage of class under 20 and student/faculty ratio for both public and private schools.

Fig. 2.3

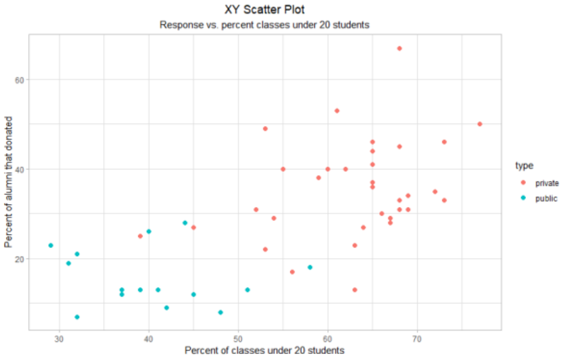
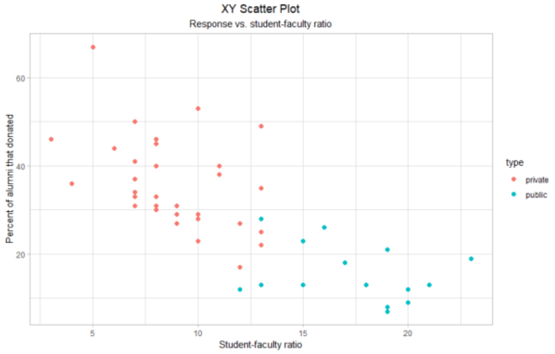


Fig. 2.4



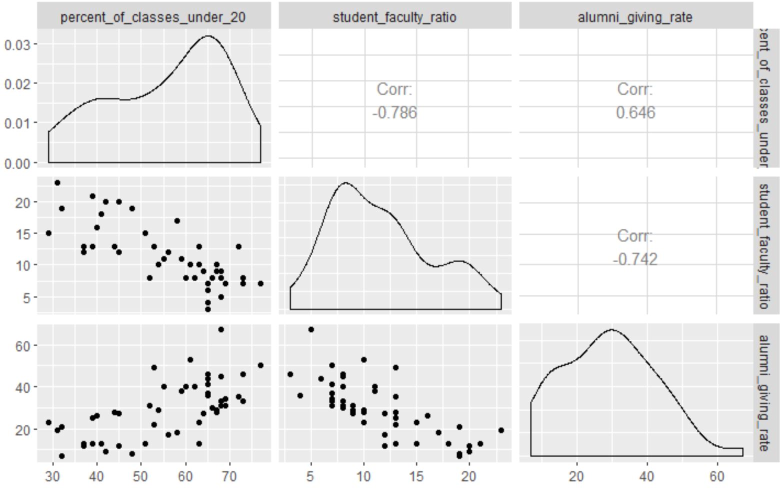


Fig 2.5 shows the correlation between variables

III. Modeling

1. Initial model selection without interaction – Using BIC as the selection criteria (Adj R^2 = 0.5414)

We initially selected a model without interactions between variables as the best fit model, using the least BIC as the election criteria. Because there are only three predictor variables, all possible combinations were ran with R's leaps() package. The best fit model consisted of regressing alumni donation rate on the student faculty ratio. This achieved an adjusted R^2 value of 0.5414.

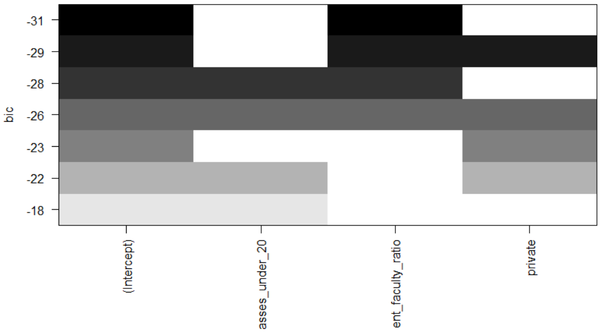


Fig. 3.1: Predictor Selection Chart

$$alumni\_giving\_rate \ = \ 53.014 - 2.057 * student\_faculty\_ratio$$

2. Transformation (Adj R^2: 0.5844)

On checking the model diagnostics, it was observed that the errors were not normally distributed (Fig. 3.3). This violates the assumptions of linear regression. Hence, we transformed alumni donation rate using Box Cox transformation to compensate for the errors and observed an increase in adjusted R^2 value to 0.5844.

$$alumni\_giving\_rate \ = \ 10.950 - \ 0.321 * student\_faculty\_ratio$$

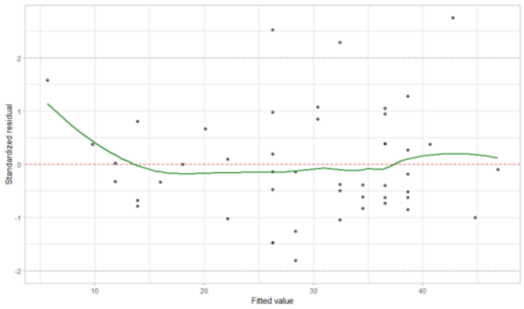


Fig. 3.2: Residuals vs Fits plot

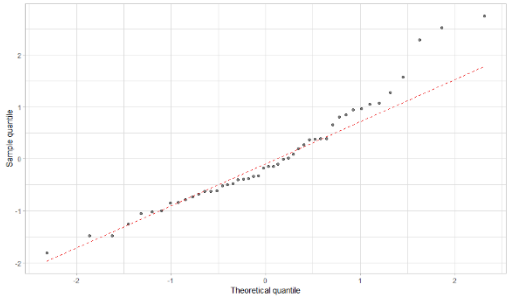


Fig. 3.3: Q-Q plot

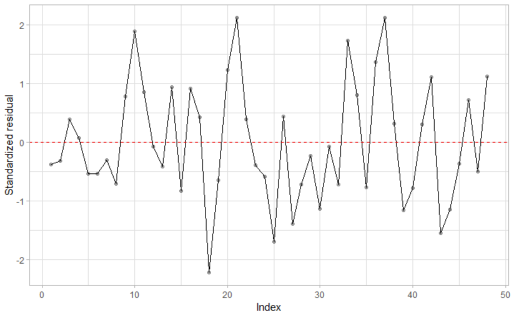


Fig. 3.4: Residuals vs order plot

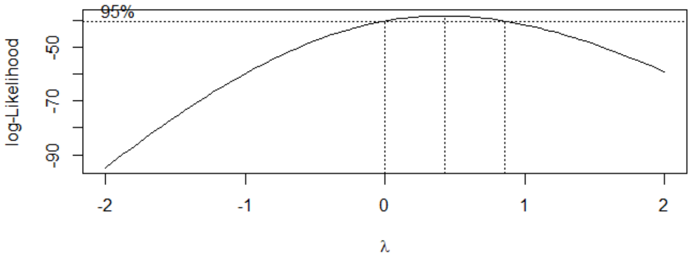
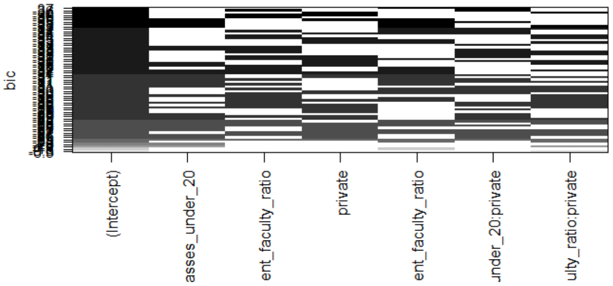
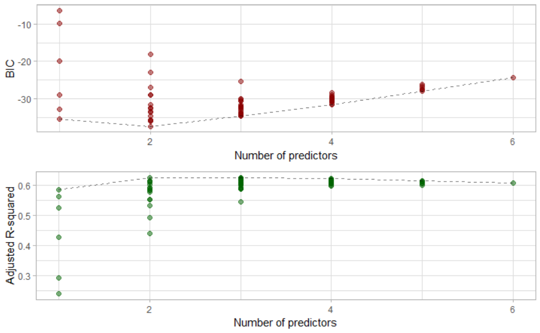


Fig 3.5: Box-cox Transformation, Ideal Lambda – 0.424

3. <u>Final model selection with 2-way interaction (Adjusted R^2: *0.6089):*</u>





Figs. 3.6-3.7: Predictor Selection Charts - with interactions

We again find the best model, this time adding all 2-way interactions with the leaps() package and using minimum BIC as the selection criteria. As seen in Figs. 3.6 and 3.7, the best fit model should contain 2 predictors in order to have high adjusted R^2 value and low BIC. However, the model with the lowest BIC had two predictors that were both 2-way interactions and no individual predictors, so we went to the next-best model that had at least one individual predictor.

IV. Results

The best fit model satisfying the above constraints and having at least one individual predictor for better interpretability was found as below:

$$alumni\_giving\_rate = 8.824 - 0.216 * student\_faculty\_ratio + 1.334 * private$$

This model also satisfied all the assumptions of linear regression: independent observations, constant variance and normally distributed errors. In the residuals vs. fitted values plot (Fig. 4.1) the residuals were non-constant variance and non-linearity. The QQ plot (Fig. 4.2) also shown an improvement in the extreme observations, especially on the higher side, compared to the previous plot in Fig. 3.3.

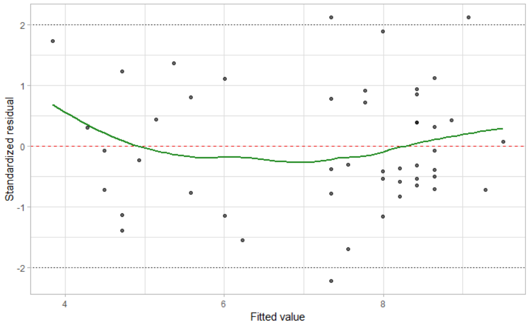


Fig. 4.1: Residuals vs Fits plot

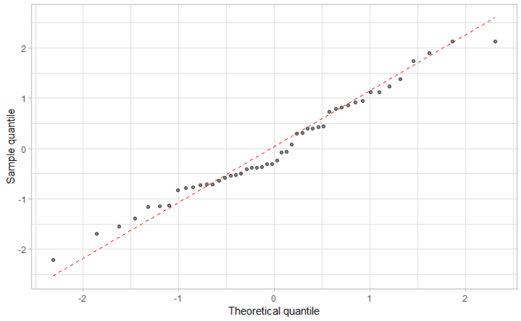


Fig. 4.2: Q-Q plot

The final plot that was observed was the residuals vs. order plot (Fig. 4.3), which helped show that the observations are independent and that there are no trends in the order the data was taken.

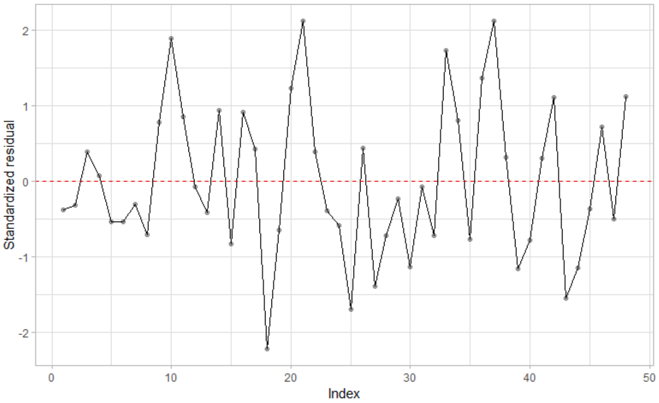


Fig. 4.3: Residuals vs order plot

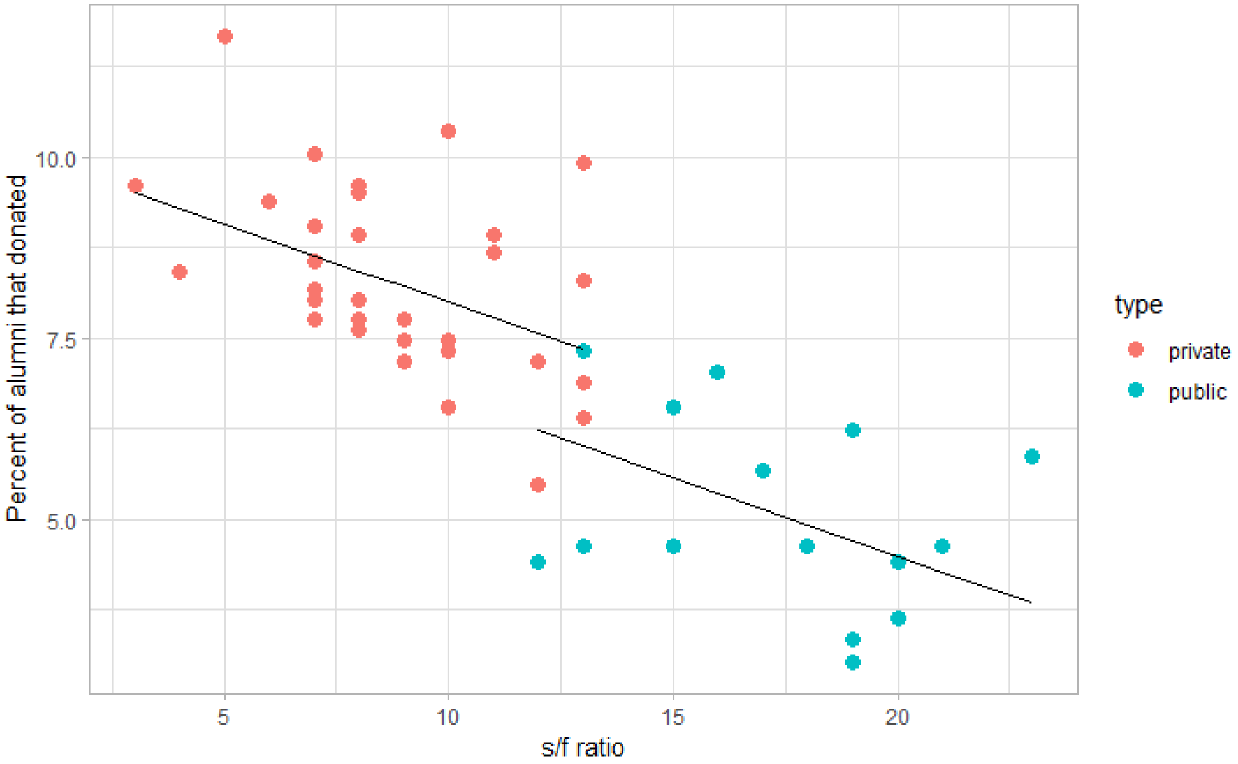


Fig. 4.4: Plot of Final Model with predictors; Student/Faculty Ratio and type (Private or Public University)

Finally, from the above graph it can be observed that the percent of alumni that donated varies linearly in a decreasing manner with student faculty ratio, clearly indicating a higher percentage of alumni having donated towards private and a lower towards public, which is reflected in the model equation.