Alumni Donation Case Study

I. 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 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 donations. Similarly, to find various other factors that can affect the alumni donation rate, we have taken the dataset of 48 national universities (America's Best Colleges, Year 2000 Edition) and implemented various linear regression models to find best model which can answer this question.

II. Data Description

Response Variable: alumni_giving_rate
Potential Predictors:

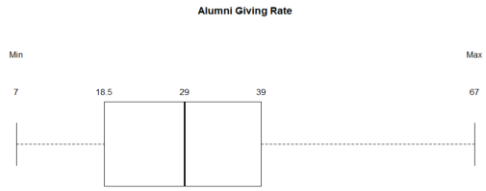
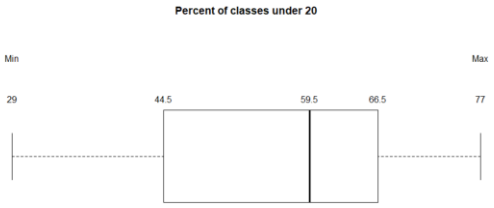
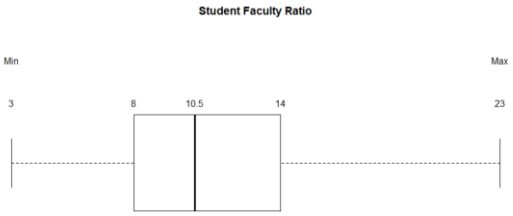
- percent_of_classes_under_20
- student_faculty_ratio
- private

Quick summary of data:

```
'data.frame'     :                 48 obs. of  5 variables:
$ ï..school      :                 Factor w/ 48 levels "Boston College",..: 1 2 3 4 5 6 7 8 9 10 ...
$ percent_of_classes_under_20:     int  39 68 60 65 67 52 45 69 72 61 ...
$ student_faculty_ratio    :       int  13 8 8 3 10 8 12 7 13 10 ...
$ alumni_giving_rate       :       int  25 33 40 46 28 31 27 31 35 53 ...
$ private                  :       int  1 1 1 1 1 1 1 1 1 1 ...
```

Since variabel "ï..school" is a row identifier, uniquely identifying each school, we remove this variable from our analysis.

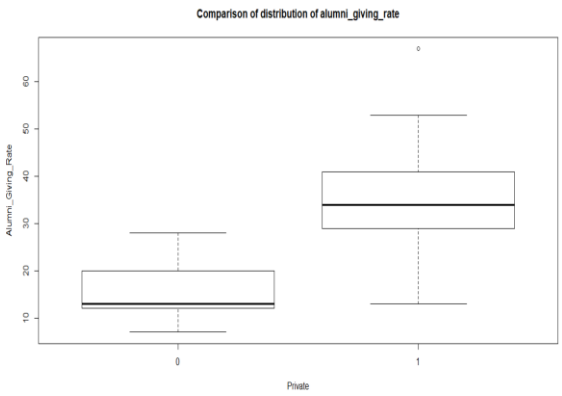
Univariate Analysis of the Response Variable and the Predictor Variables

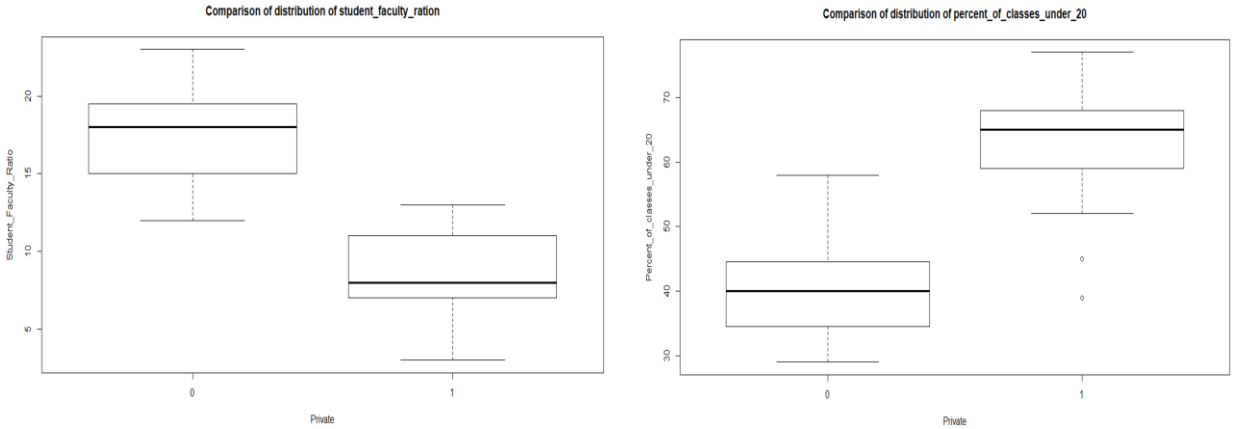
alumni_giving_rate Descriptive Statistics: Min: 7.00 ; Q_1: 18.75 Median: 29.00 ; Mean: 29.27 Q_3: 38.50 ; Max: 67.00	
percent_of_classes_under_20 Descriptive Statistics: Min: 29.00 ; Q_1: 44.75 Median: 59.50 ; Mean: 55.73 Q_3: 66.25 ; Max: 77.00	
student_faculty_ratio Descriptive Statistics: Min: 3.00 ; Q_1: 8.00 Median: 10.50 ; Mean: 11.54 Q_3: 13.50 ; Max: 23.00	

Descriptive Analysis of the categorical predictor variable: Private

Number of Private Universities	33
Number of Non-Private Universities	15

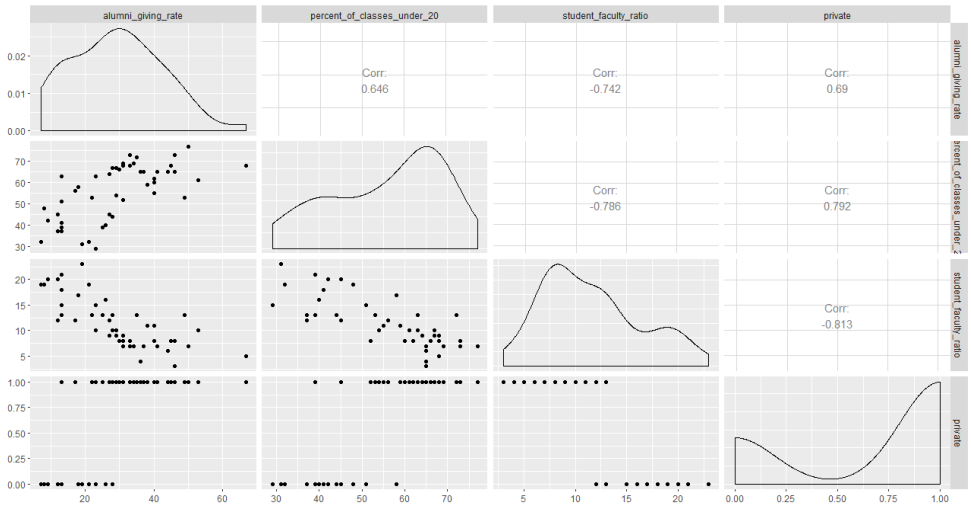
Since it is a categorical variable, which can take only 2 values, 0 or 1, we see distribution of each of the other variables with this variable





The Alumni_Giving_Rate and Percent_of_Classes_under_20 tend to be higher in the case of private universities as compared to non-private. On the other hand, student/faculty ratio is greater for non-private.

Analyzing the relation between each pair of variables: ***Correlation Analysis***



From the above correlation graphs, student_faculty_ratio and private seem to be correlated with each other. The student_faculty_ratio have some non-linear relationship with our response variable alumni_giving_rate and we might have to apply transformation to correct this. Overall, our response variable doesn't seem to have any strong correlation with any single predictor variable by looking at the correlation values.

Coming to the distributions, the student_faculty_ratio is skewed towards right with bulk of the observations coming between 5 and 10. The percent_of_classes_under_20 is looking left skewed with majority of the observations falling between 60-70%.

III. **Methods:** Here we tried 2 approaches (a) Heuristic (b)Automated using step function

Heuristic Approach:

Iteration 1: *We use all the 3 predictor variables to fit the regression line*

$$\hat{Y} = 37.78 - 1.39*(\text{student\_faculty\_ratio}) \; ; \; R^2 = 0.5747$$

In this iteration, we observe that of all the 3 predictor variables, only "student_faculty_ratio" comes out to be the significant variable.

Iteration 2: *We use all the 3 predictor variables and their interactions to fit the regression line*

With introduction of interactions, we observe that no variable (including intercept) comes out to be significant. And therefore, we discard this model.

Iteration 3: *We use "student_faculty_ratio" and "private" as predictor variables*

From the results obtained in Iteration 1, we see that "percent_of_classes_under_20" is least significant of all. And therefore, we eliminate that variable from our model. After fitting the model, we applied Box-Cox transformation on the response variable and obtained the model given below –

$$\hat{Y} = 7.32 - 0.166*(\text{student\_faculty\_ratio}) + 1.05*(\text{private}) \; ; \; R^2 = 62.98\%$$

Where $\hat{Y}$ is corresponding to the transformed response variable ($\lambda = 0.34$)

Automated Procedure:

We calculated that if we assume there are no interaction there will be close to 8 linear models' candidates for 3 predictors. Whereas considering 2-way interactions we will get 64 combinations of predictor terms.

We chose Matrices of AIC, BIC, Adjusted R2, RMSE, PRESS for model selection. As we do not have automated procedure to calculate press statistic to work along with a step function, we had to write a function.

We also used Forward selection, Backward selection and both selection with step function to find the Matrix of selection.

	Be1	Be2	Fs1	Fs2	Ss1	Ss2
AIC	352.196	352.196	352.196	352.196	352.196	352.196
BIC	357.810	357.810	357.810	357.810	357.810	357.810
AdjR2	.541	.541	.541	.541	.541	.541
RMSE	9.103	9.103	9.103	9.103	9.103	9.103
PRESS	4138.880	4138.880	4138.880	4138.880	4138.880	4138.880
nterms	2.00	2.00	2.00	2.00	2.00	2.00

Upon observing these values in above table, we see that all selection matrices have same values, which signifies that Interaction between predictors does not play any role & and all approaches suggest the same linear model.

Summary for Fs1:

	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***

$\hat{Y}$ = 53.0138 - 2.0572*(student_faculty_ratio) ; R^2 = 55.12%

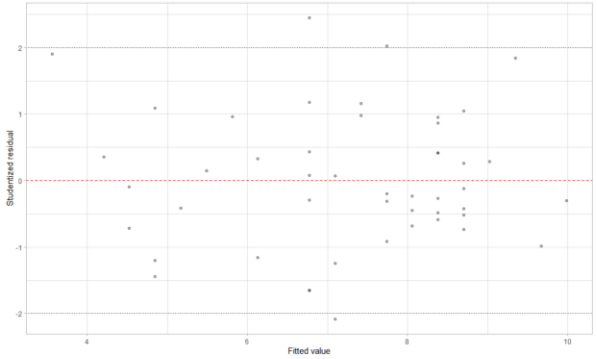
After applying Box-Cox transformation on the response variable:

$\hat{Y}$ = 10.9502 – 0.32134*(student_faculty_ratio) ; R^2 = 59.33%;

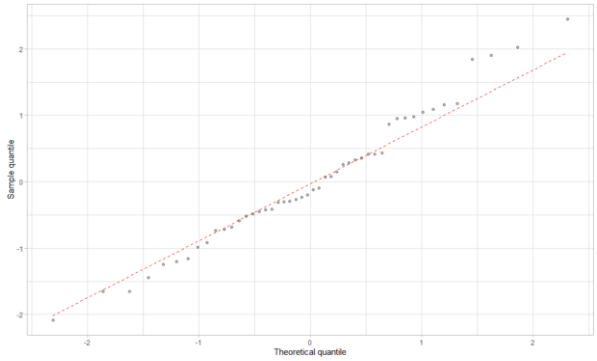
Where **$\hat{Y}$** is corresponding to the transformed response variable (λ = 0.4242)

Residual Analysis for the above model with transformed Y:

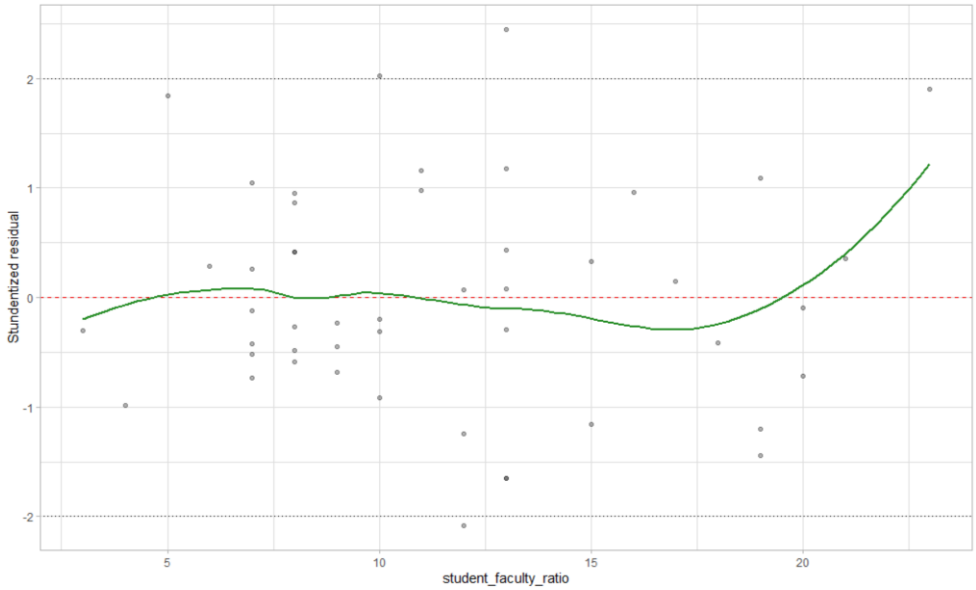
Fitted vs Residuals



Normality check for residuals



Residuals vs Predictor



- From plot of **"Residuals vs Fitted"** we observe that the residuals are randomly scattered around 0. **This suggests the fit of the transformed Y as a linear regression function of predictor variable X.** This can also be observed from the "Residual vs Predictor" graph. **Also, the variance of error terms is more or less constant**

- From the Q-Q plot it appears that the data is little skewed towards the right
  The assumption of normality seems to be violated, but that is not a major point of concern

IV. Results

From our analysis, we learned something important about the relationship between the response & predictor variables:

- Both the heuristic and automated approaches outlined above indicated the response variable (alumni_giving_rate) is negatively correlated with the student_faculty_ratio. This means there is evidence to support the following statement: Smaller student-to-faculty ratios lead to a higher alumni giving rates.
- This evidence is consistent with what we know intuitively (as mentioned in the introduction), which is that students are generally more satisfied with their education upon having more direct interaction with their professors. A small extrapolation from this point could be that students who are more satisfied with their education would be more likely to give to their alma mater.

V. Discussion

As in most prediction model selection scenarios, we were presented with a decision between accuracy and complexity in our model as we determined the "best" model to use. Based on our analysis, we determined the following models were the top 2 we evaluated:

1. **Heuristic Model – Iteration 3 (prioritizing model accuracy):** This model contains the categorical variable private and the continuous variable student_faculty_ratio as its predictor variables and a Box-Cox-transforme response variable. Based on R^2 value, we determined that the predictors of this model explain 62.98% of the variability in the response variable, which is the highest of all the models we evaluated.
2. **Automated Model (prioritizing model simplicity):** This model was developed using forward selection, and only contains the student_faculty_ratio as its predictor variable. Though it does have a Box-Cox transformation, this model seems to be the most easily interpretable model since it only contains one predictor variable.

Our conclusion is that the second model would be most appropriate for this situation, as schools are looking for actionable information about how to increase their alumni donation rate. Increasing their student-to-faculty ratio is a tangible step they could take, and the model would be easily explainable to those making hiring decisions in the future. Universities are much less likely to convert from a public to a private institution, which is a variable contained in the first model.

VI. Reference:

- http://rpubs.com/Aabhaas/Predicting_Alumni_Donation
- https://rpubs.com/Varsha_agarwalla/474946
- https://bgreenwell.github.io/uc-bana7052/slides/lecture-05#1