**CASE STUDY ON ALUMNI DONATION DATA**

**Srujana Guduru**

# INTRODUCTION

Our objective is to build a linear regression model that best explains the predictor variable i.e. alumni giving rate using other variables present in alumni data. The data set consists of 48 observations and 4 variables: percent of classes under 20, student faculty ratio, alumni giving rate and private.

The model will be useful for administrators to identify the factors influencing the increase in percentage of alumni who donate which is a key source of revenue for colleges and universities.

# DATA DESCRIPTION

Response variable (Y)= Alumni giving rate

Predictor variables:

X1 = percentage of classes with fewer than 20 students

X2 = student/faculty ratio

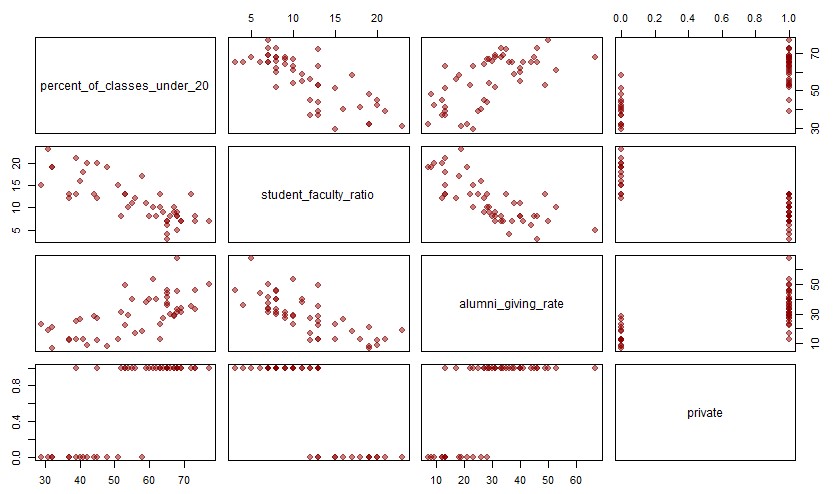
X3 = indicator variable private (1 indicates a private school)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Statistic** | **Y-Alumni giving rate** | **X1- percent of classes under 20** | **X2-student/faculty ratio** | **X3-Private school** |
| **Minimum** | 7.00 | 29.00 | 3.00 | 0.00 |
| **1st quantile** | 18.75 | 44.75 | 8.00 | 0.00 |
| **Median** | 29.00 | 59.50 | 10.50 | 1.00 |
| **Mean** | 29.27 | 55.73 | 11.54 | 0.6875 |
| **3rd quantile** | 38.50 | 66.25 | 13.50 | 1.00 |
| **Maximum** | 67.00 | 77.00 | 23.00 | 1.00 |

**Table 1: Summary Statistics**

|  |  |
| --- | --- |
| Predictor | Response Y-Alumni giving rate |
| X1- percent of classes under 20 | 0.64 |
| X2- student/faculty ratio | -0.74 |
| X3- Private | 0.69 |

**Table 2: Correlation values of response variable with all predictors.**



**Graph 1: Scatter plot matrix to understand the relationship between all the predictors and the response.**

# METHODS FOR MODEL SELECTION

Using forward selection, backward selection and step wise selection, we compare metrics of the models.

Let us consider BIC as our metric to compare the models since BIC penalizes large number of predictors and tends to favor more parsimonious models.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **be\_1** | **be\_2** | **fs\_1** | **fs\_2** | **ss\_1** | **ss\_2** |
| **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** | 0.541 | 0.541 | 0.541 | 0.541 | 0.541 | 0.541 |
| **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** | 2.000 | 2.000 | 2.000 | 2.000 | 2.000 | 2.000 |

**Table 3: Comparison of Model metrics**

We can see that all the models have the same values and gave only student faculty ratio as the predictor. Let us check Adjusted R2 value for different combinations and choose the best one since there are only 3 predictors.

|  |  |
| --- | --- |
| Predictors | Adjusted R-squared |
| X1 | 0.40 |
| X2 | 0.5414 |
| X1+X2 | 0.5418 |
| X1+X2+X3 | 0.5457 |

**Table 4: Adjusted R2 for predictor variables**

We can see that R2 value is approximately same for X2, X1+X2 and X1+X2+X3.

So, we would choose the model with only one predictor since interpretation becomes easy and variance decreases with the decrease in the number of predictors.

**Therefore, our current model is Y = β0 + β1\*X2**

**RESULTS**

**Graph**

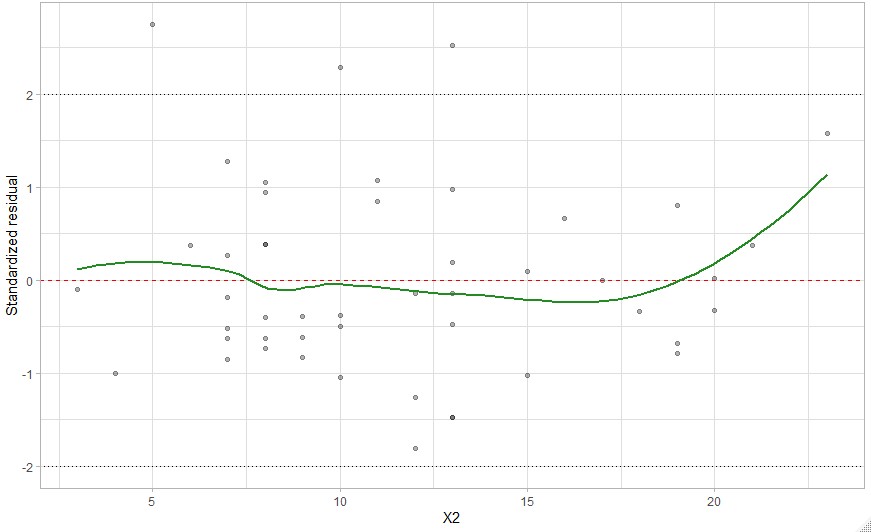
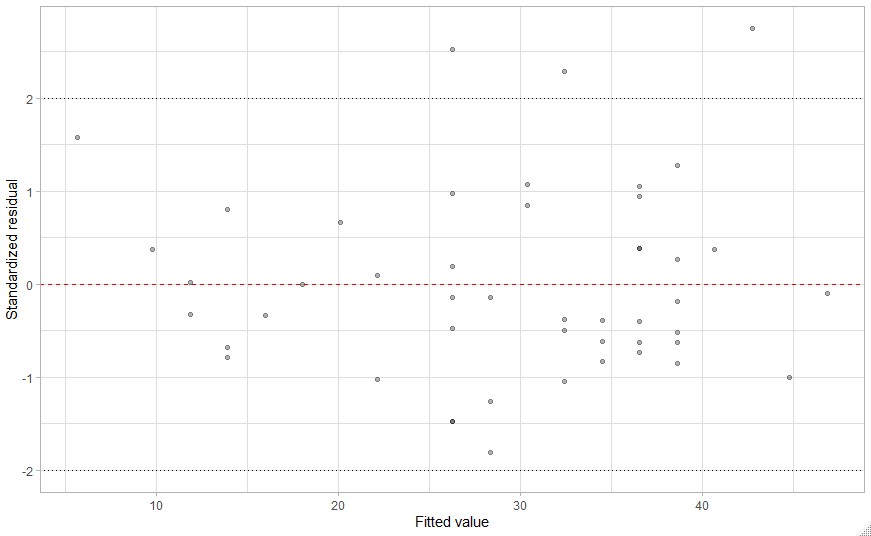
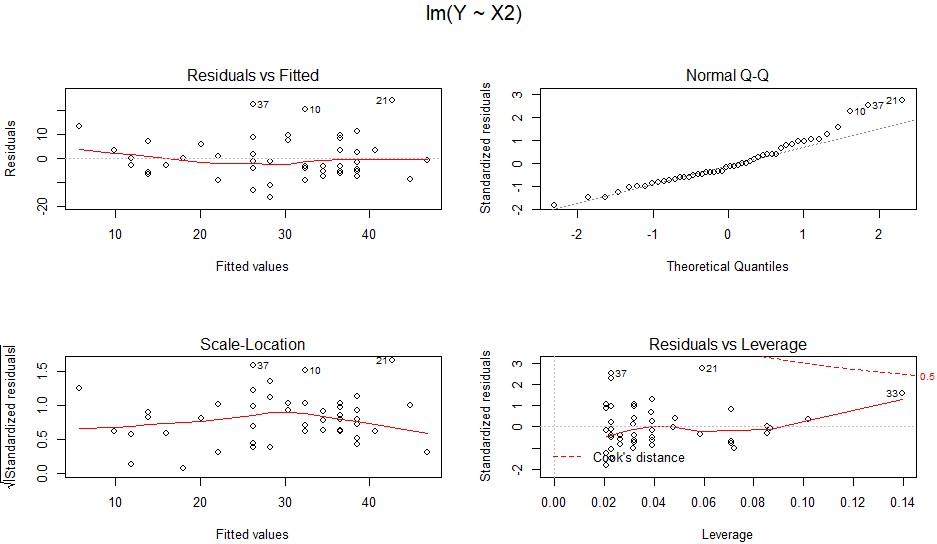
**2**

**:**

**Residual Diagnostics**

**of Model Y~X**

**2**



Checking the constant variance assumption from standardized residual vs Fitted value graph: Considering ±2 SD of standardized residual as acceptance range, there are 3 outliers. Since, there are only few outliers it can be concluded that the constant variance condition is satisfied.

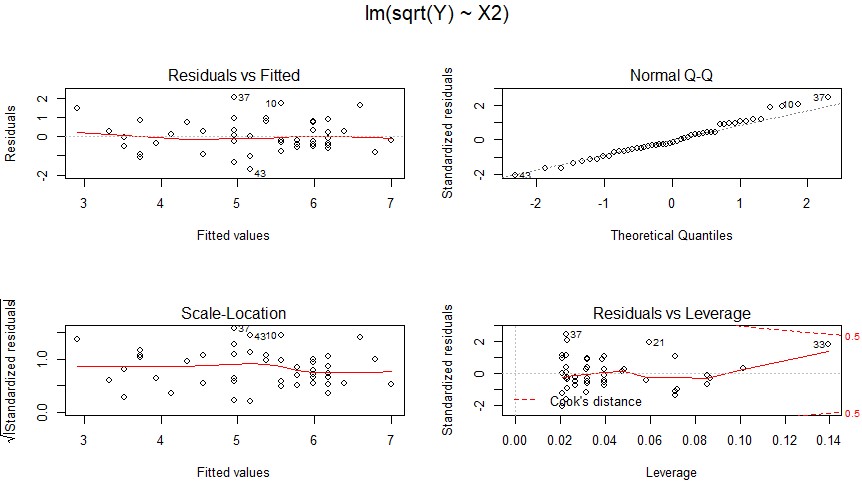
Checking non-linearity from Standardized residual vs predictor: The graph looks non-linear at the right extreme because of the outliers. Since there is only one outlier it can be ignored and can conclude that linearity is satisfied.

Normality from Normal Q-Q plot: At the right extreme there are few outliers which can be ignored as large portion of data is in line with normality and can conclude that the normality condition is satisfied

**TRANSFORMATIONS:**

Let us try some transformations and check if we could further increase the accuracy. We try transformation of Y to square root(Y):

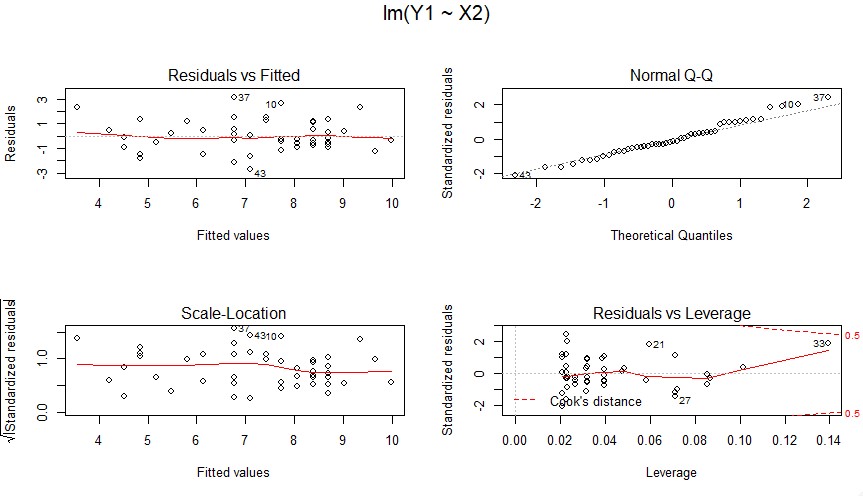
**Graph 3: Residual Diagnostics of Model sqrt(Y)~X2**



By doing the square root transformation, our R2 value (accuracy) further increased to 58.09 from 54.14.

Since, there is some non-linearity and non-normality at the right extreme in our original model we can try using Box-cox transformation to correct it.

By transforming Y to Yλ,we obtained lambda () value as 0.424.



**Graph 4: Residual Diagnostics of Model Yλ ~ X2**

By Box-cox transformation, our R2 value increased to 58.44 from 54.14 in our original model.

**CONCLUSION:**

Firstly, getting to know the data is very important in any model building problem. We have analyzed the dataset using various summary statistics and found the correlation values. We used forward selection, backward elimination and stepwise selection to find the best model and arrived at the model with one predictor. So, from the available 3 predictors we have built the model with only one predictor i.e., Student faculty ratio since the model with all the predictors is equivalent to it. We select a parsimonious model, the simplest model with the least assumptions and variables as it has greatest explanatory power and is better than the model with increased number of predictors.

After selecting the model, we have analyzed the residual diagnostics and found that the linear regression assumptions are satisfied with few exceptions. In our quest to further increase the accuracy, we performed transformations on the data. With the box cox transformation our R2 (accuracy) increased to 58.44 from 54.14 in our original model.

We can see that the increase in accuracy is only 4% after the transformation, we think it cannot be traded against our interpretation. Since our goal was to build a parsimonious model, we finalized the simple model with one predictor and without any transformation. The final model is: **(alumni giving rate) = 53.0138- 2.0572 \* (student faculty ratio)**