# Linear Regression

**Yashwanth**

## Introduction

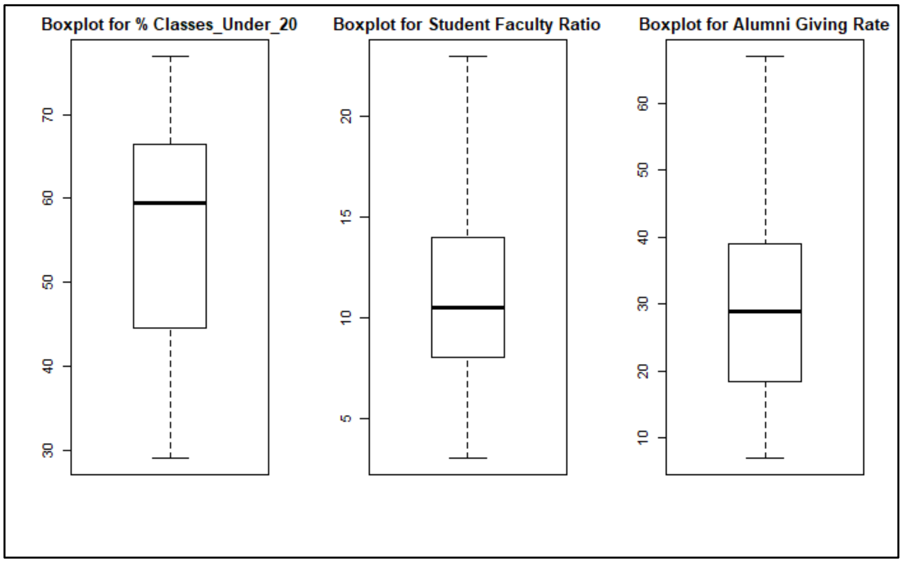
Alumni donations are an important source of revenue for colleges and universities. If administrators could determine the factors that influence increases in the percentage of alumni donation, they might be able to implement policies that could lead to increased revenues. Research shows that students 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 have done exploratory data analysis and fit linear regression models to understand implemented various linear regression model to find best model which can answer this question.

## EDA

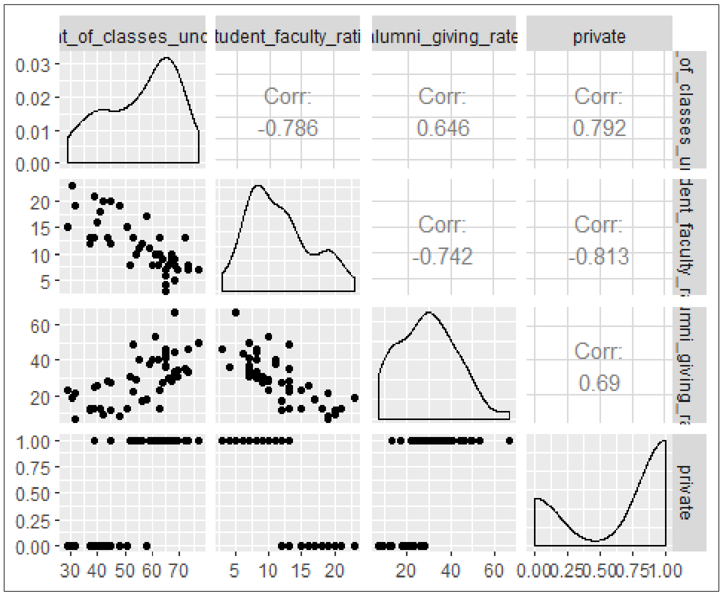
Let us look at the individual summaries for each of the 5 variables.

| Variable | Min | 1st Qu. | Median | Mean | 3rd Qu. | Max |
|---|---|---|---|---|---|---|
| Percent of classes under 20 | 29.00 | 44.75 | 59.50 | 55.73 | 66.25 | 77.00 |
| Student Faculty Ratio | 3.00 | 8.00 | 10.50 | 11.54 | 13.50 | 23.00 |
| Alumni Giving Rate | 7.00 | 18.75 | 29.00 | 29.27 | 38.50 | 67.00 |

Also, we have information for 33 private schools and 15 non – private schools. Let us understand if there are any outliers, before we proceed to look at the distributions.



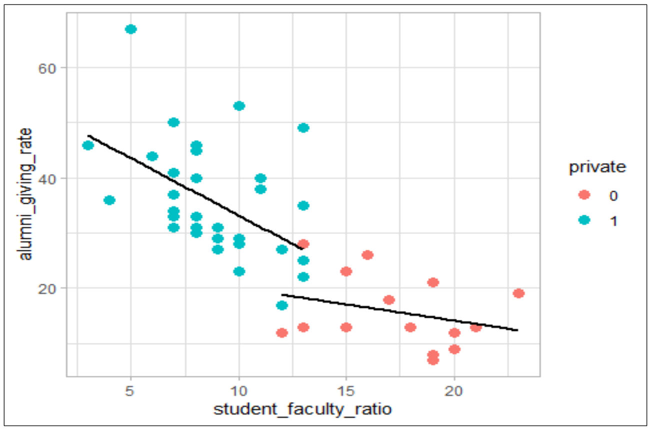
We see that there are no outliers. Let us move ahead to check distributions

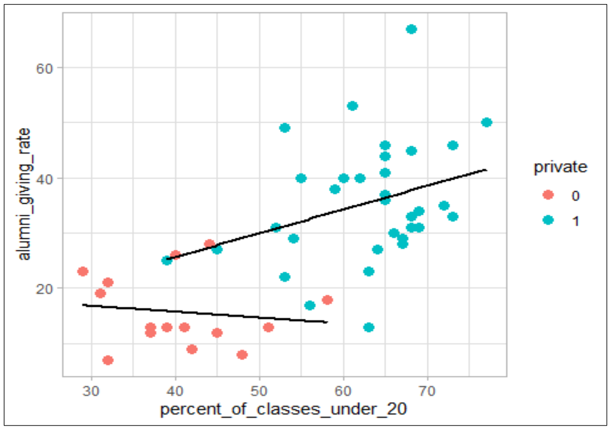


The correlations among predictors don't seem to be an issue here. We see some non - linear relationship between alumni giving rate and percent of classes under 20. This might be an issue for us and might have to apply some transformations to correct this.

We also see that the variable student faculty ratio is skewed towards the right with the bulk of the observations are having student faculty ratio between 5 to 10. The percent of classes under 20 is skewed towards the left with most schools having between 60% to 70% of their classes under 20. Also, most of the schools have ~ 20 to 40% of their alumni making donations.

Now, let us look at how the categorical variable behaves.





As the slopes do not look parallel for both the predictors percent of classes under 20 and student faculty ratio. So, we might have to include an interaction variable between percent of classes under 20 and private as well as for student faculty ratio and private.

## Modeling

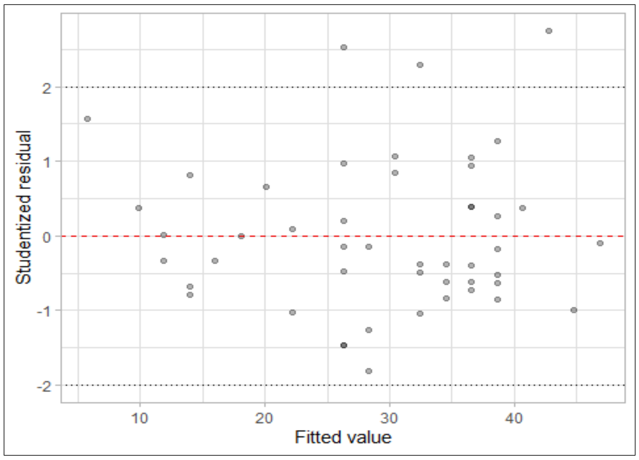
Let us use forward selection, backward elimination and stepwise selection to decide on the best performing model. The reason we are following this approach is to understand the base model we would be getting from this automated procedure and if it will be in line with the EDA results we have seen above. The scope we are giving for this exercise is – **Min**: Only intercept model, **Max:** Model which includes every variable – up to 3-way interactions. Note that these selection techniques use BIC as the accuracy metric.

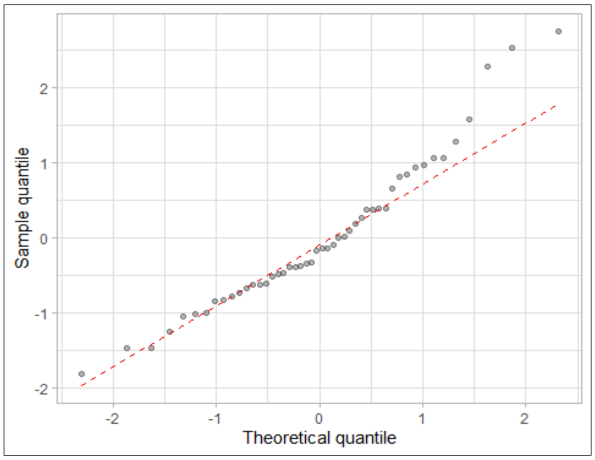
As we have built the 3 models using forward selection, backward elimination and step wise selection, let us compare the model metrics and understand how we should be proceeding. For this, let us define a function which will provide the metrics to compare the models.

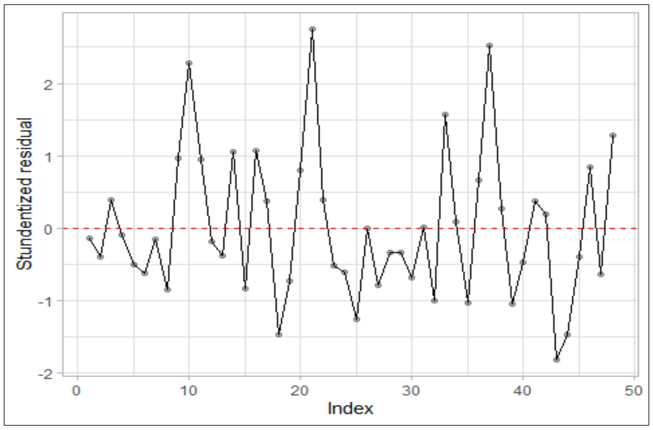
```
##           fit_be   fit_fs fit_step
## AIC      352.196  352.196  352.196
## BIC      357.810  357.810  357.810
## adjR2      0.541    0.541    0.541
## RMSE       9.103    9.103    9.103
## PRESS  4138.880 4138.880 4138.880
## nterms     2.000    2.000    2.000

## 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 ***
## 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
```

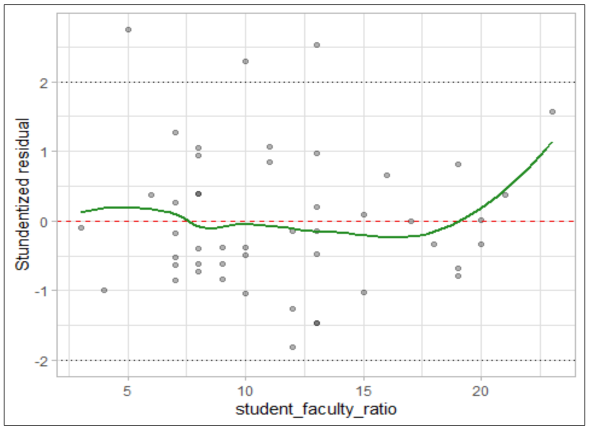
### Residual Diagnostics

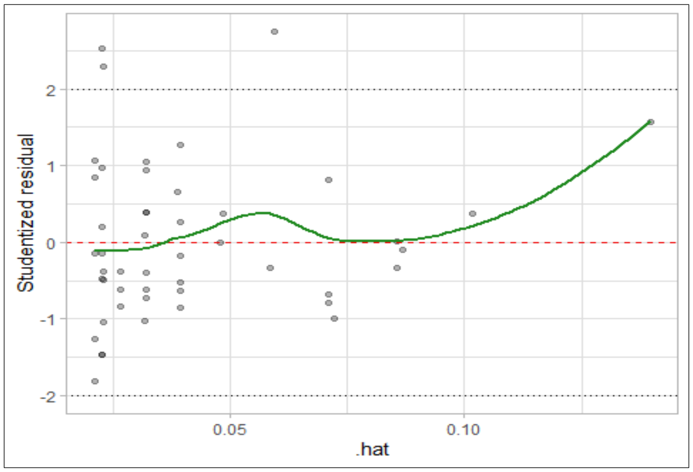
All the 3 approaches end up giving the same model that shows student_faculty_ratio as the only significant variable which contributes towards the variability of the response variable. Now let us check the diagnostics.







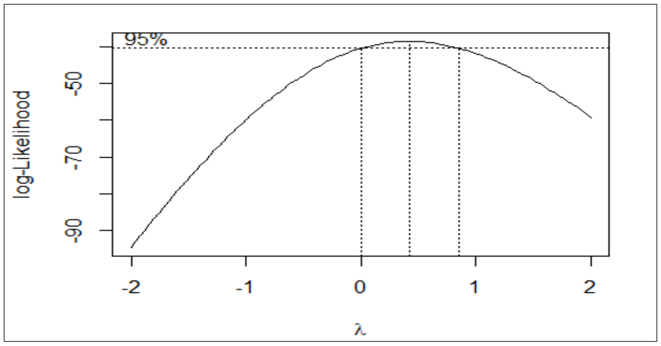




We do not see any mean structure present, but we see there might be a problem with non-constant variance. It is a skew right data – note that we have only 48 observations. We see that serial correlation is not a problem here, as the gap between the points is quite random and does not show a pattern. This seems to be a random scatter as we are do not see any shape here except for the line trying to follow the scatter. We do see an outlier which is changing the direction of the fitted line - this point is an influential point.

## Transformation (Box - Cox Procedure)

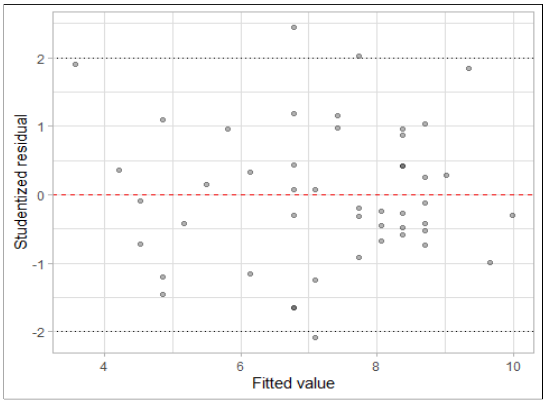
As we see a problem of non - constant variance as well as a problem with non - normality, we will apply box - cox trasformation to check if we can fix this.

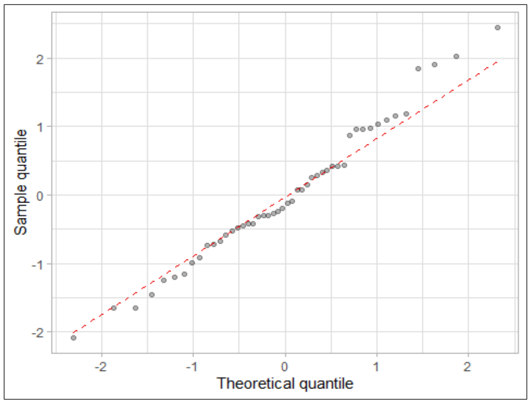


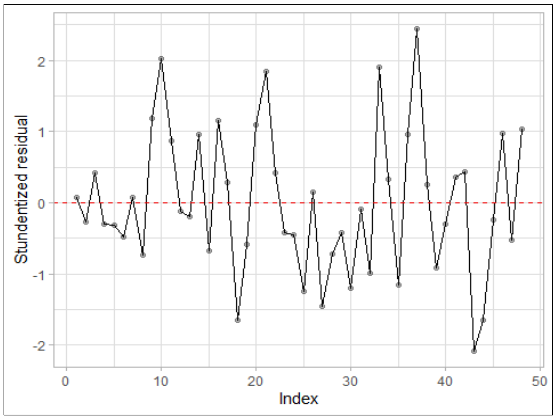
The lambda value for transformation is 0.42424. We see that the model variance being explained increased by ~4% (Adj. R^2 increased from 0.54 to 0.58). We also managed to **decrease the RMSE by ~85%** from 9.1 to 1.3

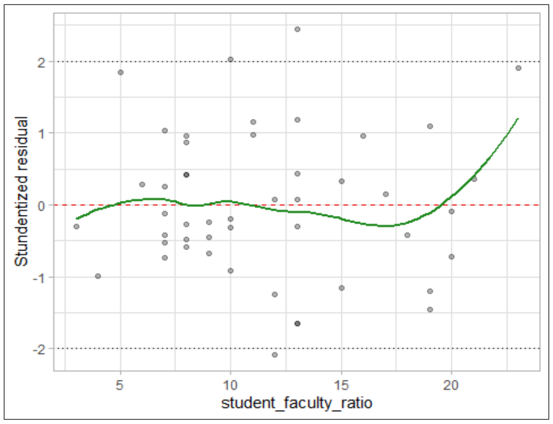
## Residual Diagnostics

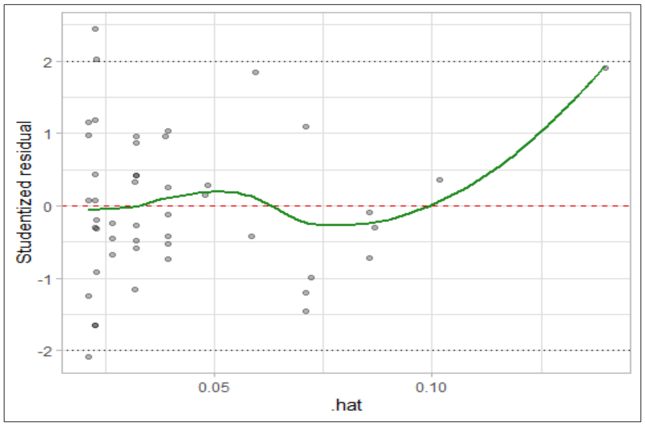
Now that we have transformed the response variable, let us look at the model diagnostics again -











We do not see any mean structure present and we also have fixed the non - constant variance to a great extent through box - cox method. However, we couldn't fix the issue of normality as it still shows exhibits skewness towards the right. However, we need not be very concerned about this violation of normality in this case. We see that serial correlation is not a problem here, as the gap between the points is quite random and does not show a pattern. This seems to be a random scatter as we are do not see any shape here except for the line trying to follow the scatter. We do see an outlier which is changing the direction of the fitted line. This point is the influential point.

All in all, we seem to have found a model which explains the response variable reasonably well and does not have any major issues with the basic assumptions of linear regression.

## Conclusion

So, our final model is -

**Estimate of Alumni_giving_rate (Ŷ) = (0.42424 * (10.9507 - 0.32134(Student_faculty_ratio)) + 1) ^ (1/0.42424).**

Model parameters - adjusted R^2 is 0.5844, RMSE is 1.3 and the value of BIC is 171.312.

A few final comments on the model – As explained above, we are giving more importance to BIC as the accuracy metric as compared to other accuracy metrics such as Adj. R^2, PRESS, AIC etc. During the analysis, we have also seen that if we consider Adj R^2 as the metric, we were able to increase the Adj. R^2 further by 4 points on including the "private" variable. However, we wanted to favor parsimonious models and thus used BIC, unless the increase is by a huge margin.

Next Steps: In further iterations, we would try to add in some external variables to see if we can enhance the performance of the model (not in the scope of this report).