Applied Regression
Case Study using R

Submitted By:
Sagar Sahoo
MS Business Analytics
University of Cincinnati

Chapter 1: Introduction

The goal of this project is to build a linear regression model that can predict alumni giving rate if certain attributes related to university are given. The regression model is built from alumni dataset that contains 48 observations and has four variables of interest - percentage of classes under 20, student faculty ratio, private and alumni giving rate.

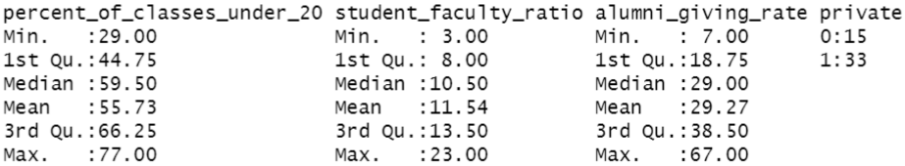
This model would be helpful in finding factors that influence alumni donation rate which would help management of an university to take necessary steps in that direction. This would help university to get more funds helping them to further improve these factors. This becomes a continuous cycle.

Chapter 2: Exploratory Data Analysis

The dataset alumni contains 48 observations and has four variables of interest - percentage of classes under 20, student faculty ratio, private and alumni giving rate. There are no missing values in these four variables.

We would be using alumni giving rate as response variable and rest of the 3 variables would be considered as predictor variables

I. Exploring Summary Statistics of all variables



```
percent_of_classes_under_20 student_faculty_ratio alumni_giving_rate private
Min.   :29.00               Min.   : 3.00         Min.   : 7.00      0:15
1st Qu.:44.75               1st Qu.: 8.00         1st Qu.:18.75      1:33
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
```

Fig 1. Summary statistics of the variables of interest

Observations:

- Private is a categorical variable
- The response variable has the highest range of 60
- The Mean and median of percent_of_class_under_20 is not similar suggesting a non-symmetric distribution. The distribution looks left skewed.

II. Exploring variables individually and interaction between variables

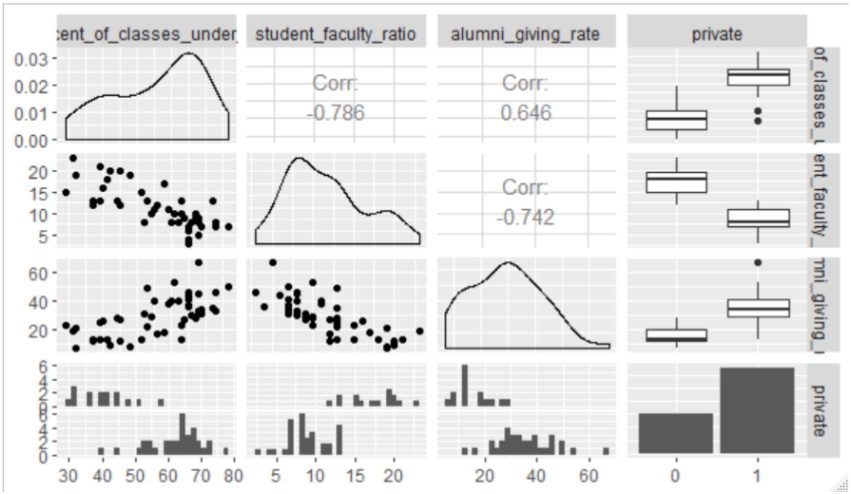


Fig 2. Variable distribution and interaction between them

Observations:

- The distribution of variables percent of classes under 200 and student faculty ratio looks bimodal. One peak for could be for private and another peak for non-private universities. The distribution of variable alumni giving rate looks approximately normal.
- There is positive linear relationship between alumni giving rate and percent ofclassesunder20
- There is negative linear relationship between alumni giving rate and student faculty ratio
- There seems to be a negative relationship between percent of classes under 20 and student faculty ratio. There can be a problem of multicollinearity. When we check the the VIF value we the value of it as 2.61. This is far less than 10. Hence there is no multicollinearity between these two variables

Chapter 3: Model Selection

For model selection, we try all models from no effect to 3- way interactions based on AIC and BIC using 3 ways of model selection – forward selection, backward elimination and step wise selection and below is the table of model metrics for all the models obtained

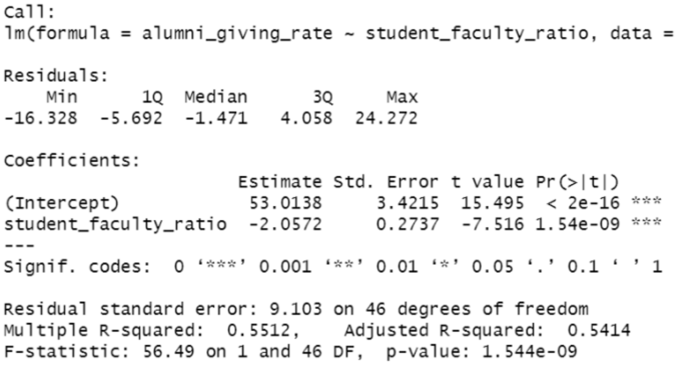
	BIC based best models			**AIC based best models**		
	be_1	**fs_1**	**ss_1**	**be_2**	**fs_2**	**ss_2**
AIC	352.196	352.196	352.196	351.365	351.137	351.365
BIC	357.81	357.81	357.81	362.592	360.493	362.592
adjR2	0.541	0.541	0.541	0.574	0.569	0.574
RMSE	9.103	9.103	9.103	8.768	8.829	8.768
PRESS	4138.88	4138.88	4138.88	4020.749	4124.727	4020.749
nterms	2	2	2	5	4	5

Based on the above parameters mainly PRESS and adjusted R-squated value, we have 2 major choices

Model 1:

alumni_giving_rate ~ student_faculty_ratio

Summary:



```
Call:
lm(formula = alumni_giving_rate ~ student_faculty_ratio, data = 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
```

Fig 3. Summary of Model1

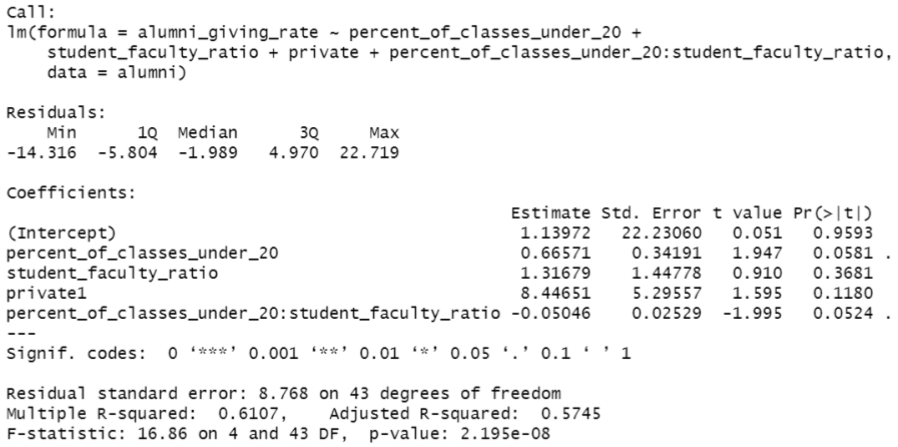
Observations:

- This model has only one predictor variable(student_faculty_ratio) making it simple model
- The sign of the estimated parameter is in sync with the negative relationship between response variable and student_faculty_ratio. The p-value<0.05 of the t-test helps us to reject the null hypothesis($\beta_1 = 0$)
- The adjusted R-squared value of this model is 0.5414
- The p-value of F-statistic is also less than 0.05 which helps to reject the null hypothesis ($\beta_1 = 0$) which is similar to t-test in this case.
- The PRESS value of this model is 4138.88

Model 2:

alumni_giving_rate ~ percent_of_classes_under_20 +
student_faculty_ratio + private + percent_of_classes_under_20:student_faculty_ratio

Summary:



```
Call:
lm(formula = alumni_giving_rate ~ percent_of_classes_under_20 + 
    student_faculty_ratio + private + percent_of_classes_under_20:student_faculty_ratio, 
    data = alumni)

Residuals:
    Min      1Q  Median      3Q     Max 
-14.316  -5.804  -1.989   4.970  22.719 

Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)  
(Intercept)                                        1.13972   22.23060   0.051   0.9593  
percent_of_classes_under_20                        0.66571    0.34191   1.947   0.0581 .
student_faculty_ratio                              1.31679    1.44778   0.910   0.3681  
private1                                           8.44651    5.29557   1.595   0.1180  
percent_of_classes_under_20:student_faculty_ratio -0.05046    0.02529  -1.995   0.0524 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.768 on 43 degrees of freedom
Multiple R-squared:  0.6107,	Adjusted R-squared:  0.5745 
F-statistic: 16.86 on 4 and 43 DF,  p-value: 2.195e-08
```

Fig 4. Summary of Model2

Observations:

- This model has only four predictor variable(student_faculty_ratio) including an interaction variable between percent_classes_under_20 and student_faculty_ratio
- The sign of the estimated parameter of student_faculty_ratio is not in sync with the negative relationship between response variable and student_faculty_ratio. The p-value>0.05 of the t-test doesnot allow us to reject the null hypothesis($\beta_1 = 0$)
- The adjusted R-squared value of this model is 0.5745
- The p-value of F-statistic is less than 0.05 which helps to reject the null hypothesis

  ($\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$)

- The PRESS value of this model is 4020.749

Considering below factors we choose our model to be model 1.

a) Model 1 is a very simple model with only 1 predictor variable

b) There is not much improvement in the value of adjusted R-squared as we include more variables and interaction between variables as seen in model 2

c) The model 2 doesnot capture the correct trend between response variable and predictor variable. (sign of estimated parameter of student_faculty_Ratio is opposite)

Therefore our model is

alumni_giving_rate = 53.0138 - 2.0572(student_faculty_ratio)

The adjusted R-squated value for this model is 0.5414.

This model states that **for every one unit increase in student_faculty_ratio, the alumni_giving_rate decreases by 2.0572 units**

Chapter 3: Model Diagnostic Analysis

We look at different diagnostic plots to know the adequacy of the selected regression model

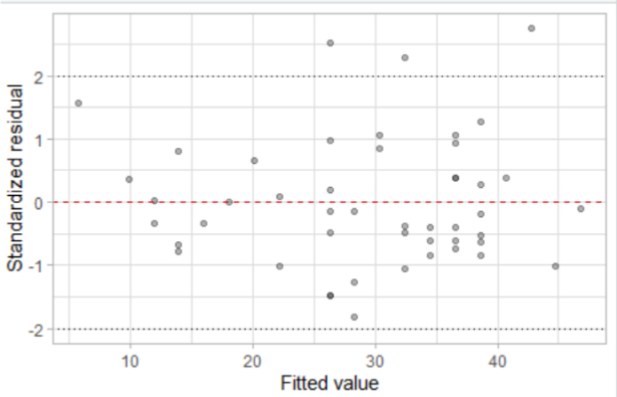


Fig 5. Residuals Vs Fitted values

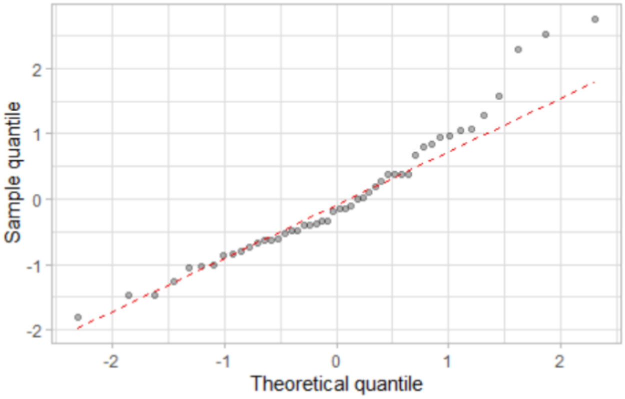


Fig 6. Normal QQ Plot

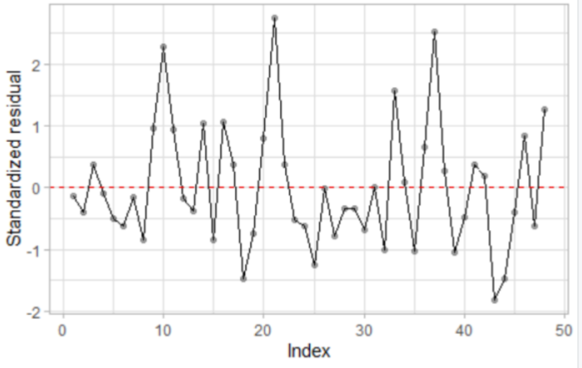


Fig 7. Residual Vs Index

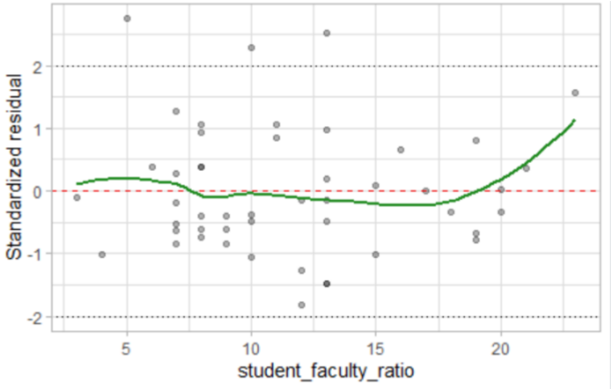


Fig 8. Residuals Vs Student_faculty_ratio

Observations:

- The varainace looks constant – The residuals are randomly scattered against both fitted values and predictor variable student_faculty_ratio(Fig5 and Fig8)
- There seems to be no misspecification of mean
- The error look approximately normally distributed from QQ plot(fig6) . It looks slightly right skewed

Therefore we can conclude that our model looks ideal.

Our Final model is

alumni_giving_rate = 53.0138 - 2.0572(student_faculty_ratio)

Chapter 5: Discussion

1) Scope for improvement with variable tranformation: We haven't considered variable transformation for both predictor and response variables which could help in building model with better model metrics.
2) Interaction between variables: Though the value of VIF between percent of classes under 20 and student faculty ratio is very low, there is clear negative linear relation between these two variables. This has to be further explored.

Chapter 6: Summary

Using Linear Regression, we developed a model to predict the alumni giving rate of universities. According to the final model, the alumni giving rate is dependent on student faculty ration of the university. For every unit increase in the student faculty ratio, the alumni giving rate decreases by 2.0572. This suggests universities to either have fewer intake of students or hire more faculty to maintain a better student faculty ratio. This would increase the university's income from alumni.