Nov 15th, 2017

# **BANA7041 HW**:(5)-MODULE II

SECTION 001

TOTAL QUESTIONS: 1 – MINI PROJECT

ASSIGNED GROUP NUMBER: (9)

GROUP MEMBER NAMES:

(Ajmal, Mohammed) M12399792

(Appalla, Sai Uday Kumar) M12383301

(Bobde, Sonal Sudhakar) M12388380

(Lalgudi Venkatesan, Renganathan) M12366827

# Determination of Factors Influencing Alumni Giving Rate

## Introduction

**Problem Statement:**

The project aims at identifying the factors influencing the alumni giving rate at universities and developing a linear regression model for predicting the alumni giving rate based on the factors considered. The data used for the project is the alumni.xls file which has data for 48 national universities (America's Best Colleges, Year 2000 Edition). The analysis was performed using the variables in the dataset along with their interactions. As an outcome of the analysis, we find that we can use only the Student/Faculty ratio variable as the predictor to make relatively precise and accurate predictions for the alumni giving rate. As a measure of the accuracy of the model we computed the R-Squared value and we find the value to be 0.5512.

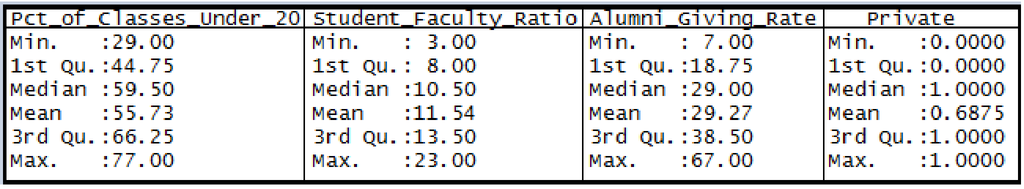
## Data Description:

There are four columns in the dataset namely College Name, % of Classes Under 20 students, Student/Faculty Ratio, Alumni giving rate and the Type of institution (Private or Public) out of which the first three are continuous variables and the later is a categorical variable. Going forward, we will refer to the alumni giving rate variable as Y (Predicted Variable) and all other variables used to make prediction about Y as predictor or X variables.

**Exploratory Data Analysis:**

(i) We would like to understand the distribution of various variables present in the dataset. The **[Table-1]** gives the summary of all the variables. We find that the alumni giving rate can vary from 7% to 67%. Also, the private variable has just two values it is categorical in nature.

**[Table-1: Summary of the variables in the dataset]**

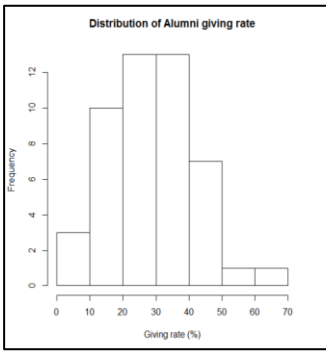


| Pct_of_Classes_Under_20 | Student_Faculty_Ratio | Alumni_Giving_Rate | Private |
|---|---|---|---|
| Min.   :29.00 | Min.   : 3.00 | Min.   : 7.00 | Min.   :0.0000 |
| 1st Qu.:44.75 | 1st Qu.: 8.00 | 1st Qu.:18.75 | 1st Qu.:0.0000 |
| Median :59.50 | Median :10.50 | Median :29.00 | Median :1.0000 |
| Mean   :55.73 | Mean   :11.54 | Mean   :29.27 | Mean   :0.6875 |
| 3rd Qu.:66.25 | 3rd Qu.:13.50 | 3rd Qu.:38.50 | 3rd Qu.:1.0000 |
| Max.   :77.00 | Max.   :23.00 | Max.   :67.00 | Max.   :1.0000 |

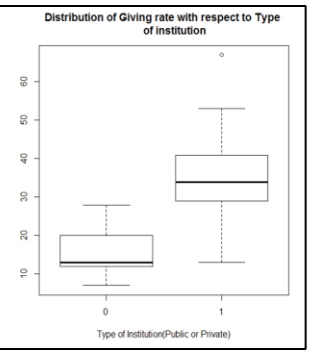
(ii) In building the linear regression model, one of our major assumptions is that Y is approximately normally distributed. We can check that by plotting a histogram of the Y values as shown in [**Figure-1].** We find that the normality assumption holds true.

(iii) We also want to understand how the Y variable is distributed with respect to the type of institution by using a Box Plot as shown in the [**Figure – 2]**. We find that the spread of giving rate is a lot different for private and public institutions. We also find that there is an outlier in our giving rate for the private institution.

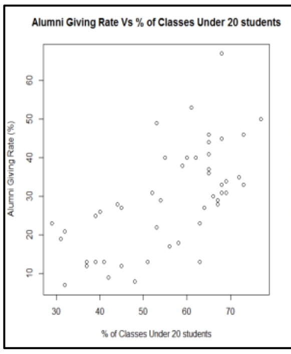
(iv) We use scatter plots to understand the relationship between the other variables. From the [**Figure-3]** we find that as the % Class under 20 students increases, the alumni giving rate increase. Also from [**Figure-4]** we find that as the Student Faculty ratio increases, the alumni giving rate decreases. We still need to perform a statistical analysis to verify these findings.



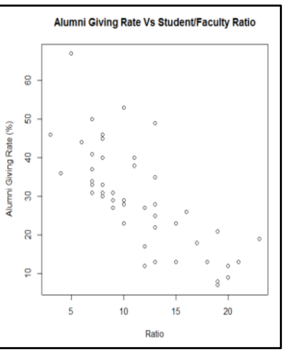
**[Figure – 1]**



**[Figure – 2]**



**[Figure – 3]**



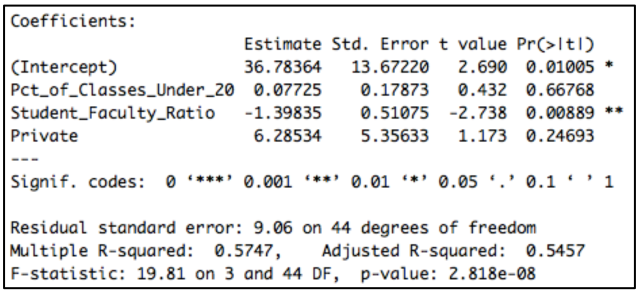
**[Figure – 4]**

## Computing Correlation Coefficient:

We find that the correlation between alumni giving rate and % Class under 20 students is 0.6456504 which supports our analysis that they are positively correlated. Also, we find that the correlation between giving rate and student faculty ratio is -0.7423975 which supports our analysis that they are negatively correlated.

## Preliminary Linear Regression Model:

**Table 2: Coefficients and significance levels for Model 1**



```
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            36.78364   13.67220   2.690  0.01005 *
Pct_of_Classes_Under_20  0.07725    0.17873   0.432  0.66768
Student_Faculty_Ratio   -1.39835    0.51075  -2.738  0.00889 **
Private                  6.28534    5.35633   1.173  0.24693
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.06 on 44 degrees of freedom
Multiple R-squared:  0.5747,    Adjusted R-squared:  0.5457
F-statistic: 19.81 on 3 and 44 DF,  p-value: 2.818e-08
```

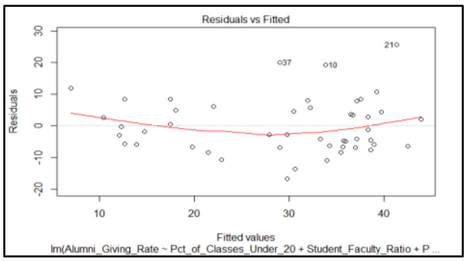
For the first model we fit has Alumni Giving Rate as the response variable and Percentage of Classes under 20 as predictor variable X1, Student Faculty Ratio as predictor variable X2, Private as predictor variable X3. The fitted model is as such:

$$\textbf{Model 1: } Y = 36.78364 + 0.07725X_1 - 1.39835X_2 + 6.28534X_3$$

**F-Statistics:** The value of the F-statistic is 19.81 with a p-value of 2.818e-08. This p-value indicates that we can reject the null hypothesis for the F-test which is $\beta_1 = \beta_2 = \beta_3 = 0$. This shows that at least one of the parameters is not zero and hence, we can assume that the response variable has a linear relationship with at least one of the predictor variables

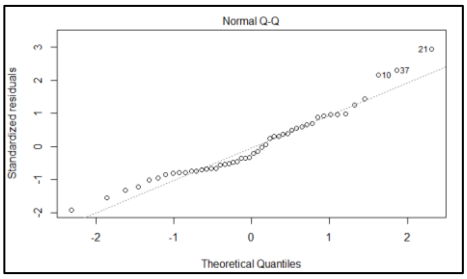
**Regression parameter estimates:** From the results shown in the table, it can be observed that the p-values for the parameters β$_1$ and β$_3$ are greater than 0.05. This indicates that the corresponding null hypotheses cannot be rejected (β$_1$=0 and β$_3$=0 respectively). Hence, from the obtained p-values, we can infer that the variables "Pct_of_Classes_Under_20" and "Private" do not have a significant effect on the regression model.

**Residual Diagnostics:**



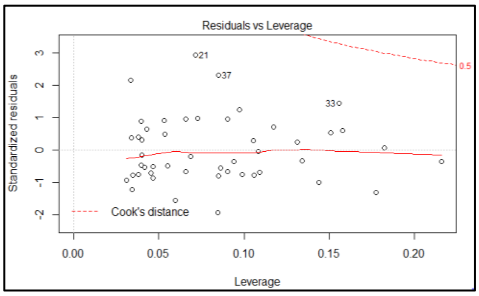
**[Figure – 5: Residual plot to test for linearity and error variance]**

**Tests for linearity and constant error variance:** Residuals look randomly distributed and don't follow any kind of non-linear pattern. Hence, we can say that relationship is linear. Also, there are no signs of any non-constant error variance.



**[Figure – 6: Quantile-Quantile plot to test normality of error]**

**Tests for error normality:** The normal q-q plot shows that the standardized residuals do not follow the 45-degree line and the distribution seems lightly-tailed at both ends. This indicates that the error is not normally distributed.

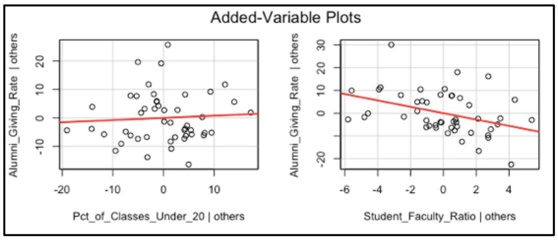


**[Figure – 7: Residuals vs Leverage plot to check for high leverage points]**

**Residuals vs Leverage:** This plot shows us that there are no high leverage points in the data as there are no residuals beyond the Cook's threshold lines.

# Statistical Methods and Analysis:

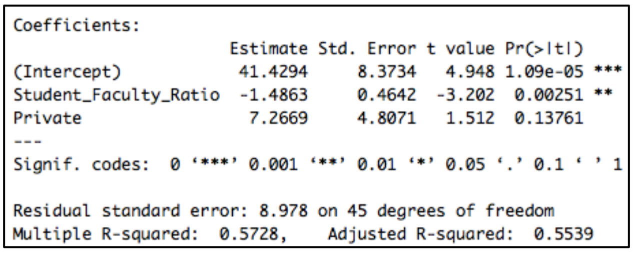
**Effect of X$_1$ (Pct_of_Classes_Under_20):** From the preliminary model summary, we see that the p-value for X$_1$ is greater than 0.05. So, it might be the case that X$_1$ has no significant effect on the response variable or is perhaps highly correlated with X$_2$ (Student_Faculty_Ratio). First, we see the correlation of X$_1$ and X$_2$. This is -0.7855. This indicates that there may be multi-collinearity if we include both. To verify this further, first we can construct added variable plots.



**[Figure – 8: Added variable plots for all X$_1$ and X$_2$]**

From the added variable plot, we can see that the slope for Pct_of_Classes_under_20 is almost at zero. So, it might be a good idea to leave out X$_1$. We also performed a partial F-test.

**[Table 3: Coefficients and significance levels for Model 2]**

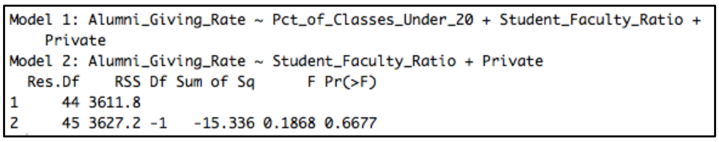


```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           41.4294     8.3734   4.948 1.09e-05 ***
Student_Faculty_Ratio -1.4863     0.4642  -3.202  0.00251 ** 
Private                7.2669     4.8071   1.512  0.13761    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.978 on 45 degrees of freedom
Multiple R-squared:  0.5728,	Adjusted R-squared:  0.5539
```

The reduced model used $X_2$ (Student_Faculty _Ratio) and $X_3$ (Private) as regressors. **Model 2: $Y = 41.4294 - 1.4863X_2 + 7.2669X_3$.** We see a slight increase in adjusted R-squared, again this indicates that dropping $X_1$ may be alright.

**[Table 4: Results of Partial F-test]**

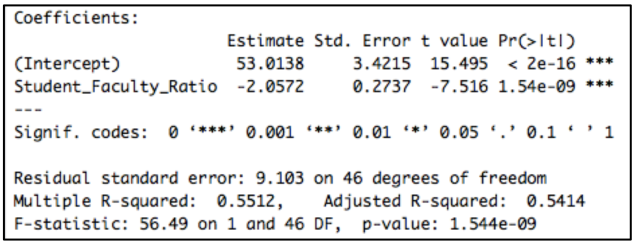


```
Model 1: Alumni_Giving_Rate ~ Pct_of_Classes_Under_20 + Student_Faculty_Ratio + 
    Private
Model 2: Alumni_Giving_Rate ~ Student_Faculty_Ratio + Private
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     44 3611.8                           
2     45 3627.2 -1   -15.336 0.1868 0.6677
```

Here, we see that p-value is greater than 0.05. So, we cannot reject the null hypothesis that $\beta_1 = 0$. So, we can drop ‘percentage of classes under 20’ from the model. Now, we will be proceeding with the reduced model and assess effects of $X_3$ because we can see that $X_3$ has a p-value > 0.05 even in this model.

**Effect of $X_3$ (Private):** From the preliminary model summary as well as from reduced model summary, we see that the p-value for $X_3$ is greater than 0.05. So, maybe that $X_3$ also has no significant effect on the response variable.

**[Table 5: Coefficients and significance levels for Model 3]**

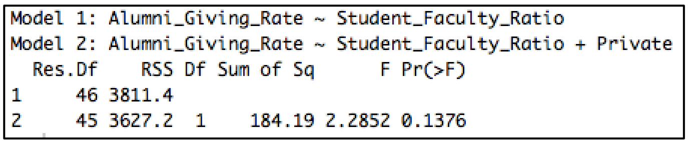


```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)            53.0138     3.4215  15.495  < 2e-16 ***
Student_Faculty_Ratio  -2.0572     0.2737  -7.516 1.54e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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
```

Consider, model 3 with only $X_2$ as regressor. **Model 3: $Y = 53.0138 – 2.0572X_3$.** As this is a categorical variable, we will proceed with F-test and not added variable plots (Reference included).

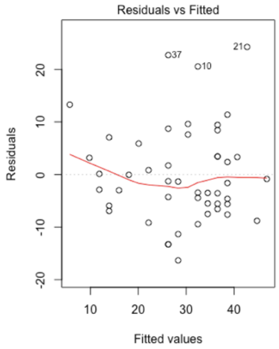
**[Table 6: Results of Partial F-test]**

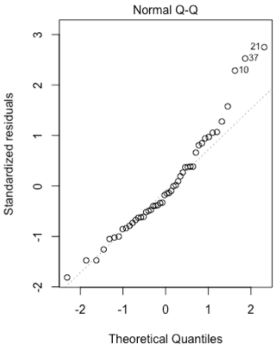


```
Model 1: Alumni_Giving_Rate ~ Student_Faculty_Ratio
Model 2: Alumni_Giving_Rate ~ Student_Faculty_Ratio + Private
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     46 3811.4                           
2     45 3627.2  1    184.19 2.2852 0.1376
```

Here, p-value > 0.05. So, we cannot reject the null hypothesis : $\beta_1 = 0$. So, we can drop ‘Private’ from the model. Now, we will be proceeding with model 3 and looking at ways to improve the model.

## Residual Diagnostics for final model:





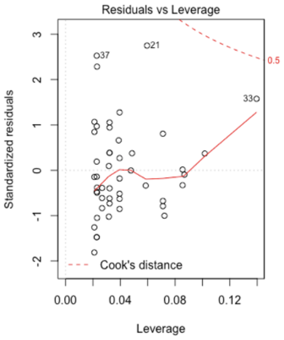


**Figure 9: Residuals vs Fitted** **Figure 10: Normal QQ Plot** **Figure 11: Residuals vs Leverage**

- **Residuals vs Fitted:** Tests for linearity and constant error variance

Residuals look randomly distributed and don't follow any kind of non-linear pattern. Hence, we can say that relationship is linear. Also, there are no signs of any non-constant error variance.

- **Normal QQ Plot:** Tests for error normality

Standardized residuals approximately follow normal distribution and is slightly right-skewed.

- **Residuals vs Leverage:**

This plot shows us that there are three points 37, 21, 33 can be considered as influential points. However, none of them lie outside the Cook's distance (dotted red line). So, we can flag them, but should not ideally remove them.

## Results

**Final model*: $Y = 53.0138 – 2.0572X_3$** *Model summary: Refer Table 5 above

**Interpretation:** According to this model, we see that every unit increase in $X_3$ (Student_Faculty_Ratio) decreases the Alumni Giving Rate by 2.0572. Also, at Student Faculty Ratio of zero (given zero is in the scope of the model), the Alumni Giving Rate will be 53.0138.

## Discussion

1. **Exploring interaction terms for between variables:** We tried to include interaction between the predictor variables to check if that leads to a better model. We transformed the predictor variables $X_1$ and $X_2$ by centering them around their respective means and then included the product as an interaction term. We still found that $X_1$ or the interaction term was not significant. Furthermore, we included interaction between $X_3$ and $X_2$, $X_1$ i.e. interaction between categorical and both continuous variables. However, the results of the regression were not encouraging. Hence these were not included in the final model.
2. **Limitations:** The current dataset is small and hence certain variables like Private are not coming out to be significant in the regression analysis.
3. **Remedy for Outliers (not applied):** We built a model by removing points 37, 21 and 33 which seem to be outliers. We also got a higher R-squared value of 0.625. But we did not remove them in the final model because they do not lie outside the Cook's distance of 0.5 (refer figure 11).

## References:

1. Whether to use added variable plots for categorical variables: http://www-hsc.usc.edu/~eckel/biostat2/notes/notes11.pdf
2. For interaction between continuous variables: http://www.psychwiki.com/wiki/Interaction_between_two_continuous_variables