Homework 5

Akhilesh Sureddi

2022-11-27

## **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 who donate, they might be able to implement policies that could lead to increased revenues.

A study shows that students who have more access to the faculty are more likely to be satisfied. 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 an increase in the percentage of alumni who donate.

In this project, we will develop a linear regression model to study the factors affecting alumni donation in schools.

## **Data**

The data set comes from the 2006 ASA Data Expo and contains data for 48 national universities (America’s Best Colleges, Year 2000 Edition).

#### Snapshot of the data:

| school | percent\_of\_classes\_under\_20 | student\_faculty\_ratio | alumni\_giving\_rate | private |
| --- | --- | --- | --- | --- |
| Boston College | 39 | 13 | 25 | 1 |
| Brandeis University | 68 | 8 | 33 | 1 |
| Brown University | 60 | 8 | 40 | 1 |
| California Institute of Technology | 65 | 3 | 46 | 1 |
| Carnegie Mellon University | 67 | 10 | 28 | 1 |

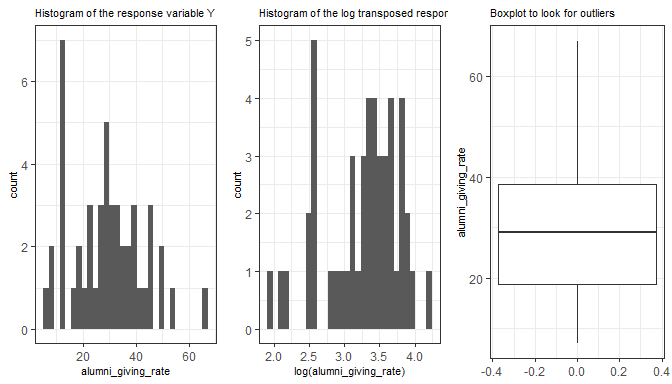
## **Exploratory Data Analysis**

**Following are the observations based on Univariate and Bivariate Analysis:**

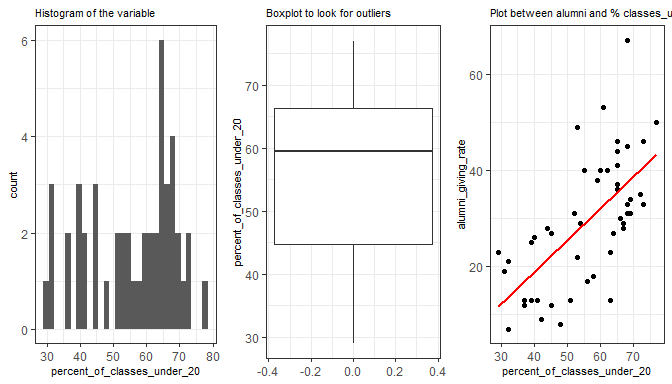
* There are no missing values and outliers in the data
* log transformation on alumni\_giving\_rate doesn’t have much impact
* There exist a positive relationship between the Percent of Classes Under 20 and alumni\_giving\_rate
* There exist a negative relationship between the Student Faculty Ratio and alumni\_giving\_rate
* There exist a relationship between the private/public and alumni\_giving\_rate
* The correlation between student faculty ratio and % classes under 20 is -0.786
* The correlation between student faculty ratio and Alumni giving rate is -0.742

### Univariate Analysis

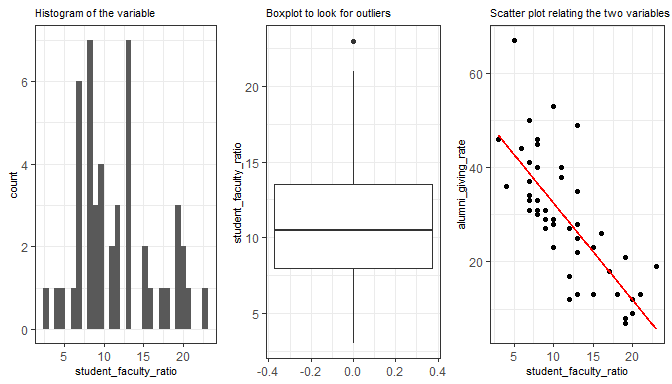
#### Response Variable: alumni\_giving\_rate



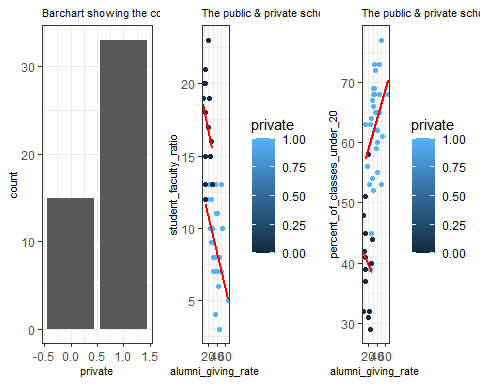
#### Dependent Variable: Percent of Classes Under 20



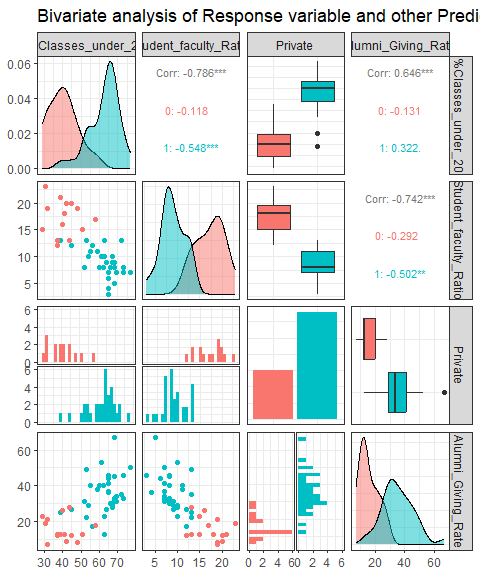
#### Dependent Variable: Student Faculty Ratio



#### Dependent Variable: Private



### Bivariate Analysis

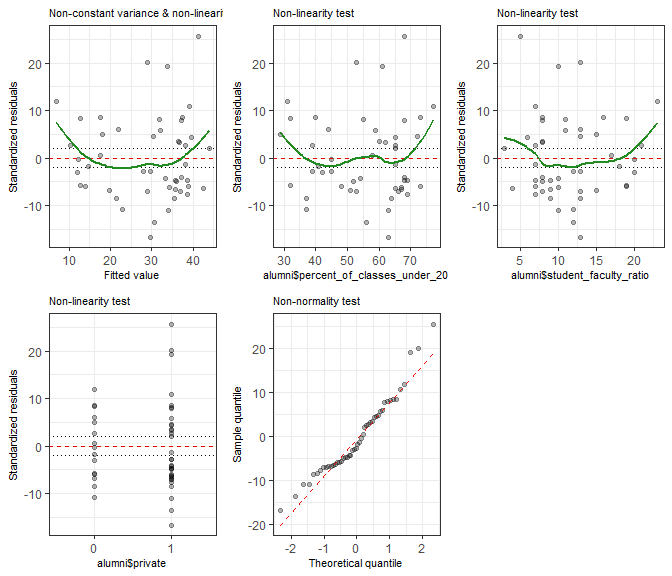


## **Model**

### Basic Model Results

##   
## Call:  
## lm(formula = alumni\_giving\_rate ~ student\_faculty\_ratio + percent\_of\_classes\_under\_20 +   
## private, data = alumni)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.757 -6.320 -2.273 5.152 25.669   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 36.78364 13.67220 2.690 0.01005 \*   
## student\_faculty\_ratio -1.39835 0.51075 -2.738 0.00889 \*\*  
## percent\_of\_classes\_under\_20 0.07725 0.17873 0.432 0.66768   
## private1 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

### Basic Model Diagnostics



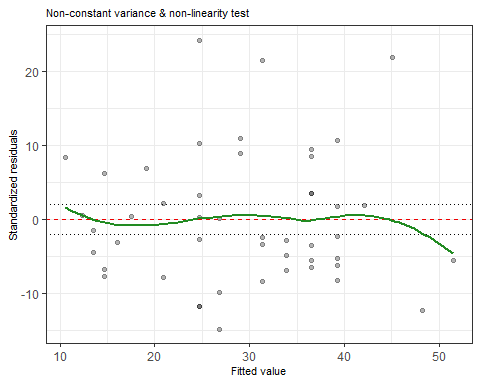
As seen from the results above, There exists a slight non linearity in the response variable and the residuals are not constantly spread. This indicates heteroskedasticity.

The model has selected only student\_faculty\_ratio as the significant predictor variable and has Adj.R squared value of 0.55.

#### Student Faculty Ratio only Model

First step is to include student\_faculty\_ratio and a polynomial term of student\_faculty\_ratio to check:

##   
## Call:  
## lm(formula = alumni\_giving\_rate ~ student\_faculty\_ratio + I(student\_faculty\_ratio^2),   
## data = alumni)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.799 -6.305 -2.318 4.197 24.308   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 61.99693 8.17756 7.581 1.41e-09 \*\*\*  
## student\_faculty\_ratio -3.69561 1.38320 -2.672 0.0105 \*   
## I(student\_faculty\_ratio^2) 0.06354 0.05259 1.208 0.2333   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.057 on 45 degrees of freedom  
## Multiple R-squared: 0.5653, Adjusted R-squared: 0.5459   
## F-statistic: 29.25 on 2 and 45 DF, p-value: 7.249e-09



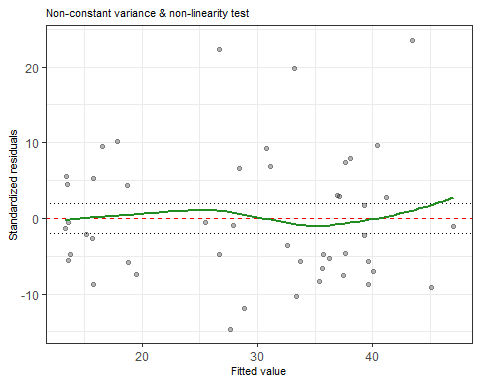
Though the non linearity has reduced the polynomial term doesn’t hold any significance in prediction. So, our approach is to add interaction terms to the models.

#### Adding interaction Terms

As we observed in the EDA above, we noticed that there seems to be an interaction between: + Private & student\_faculty\_ratio + Private & percent\_of\_classes\_under\_20

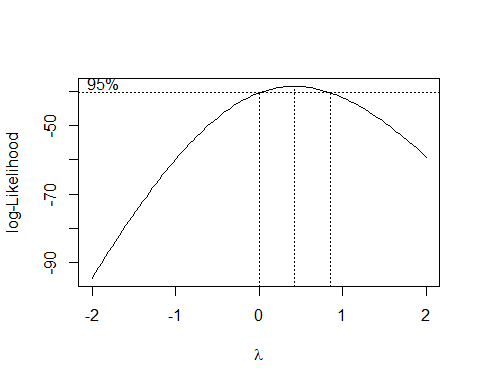
Next step is to include the interaction terms and check the model results:

##   
## Call:  
## lm(formula = alumni\_giving\_rate ~ student\_faculty\_ratio + percent\_of\_classes\_under\_20 +   
## private + private:student\_faculty\_ratio + private:percent\_of\_classes\_under\_20,   
## data = alumni)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.624 -5.768 -1.769 5.323 23.564   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 32.0206 18.9423 1.690 0.0984 .  
## student\_faculty\_ratio -0.6258 0.7466 -0.838 0.4067   
## percent\_of\_classes\_under\_20 -0.1363 0.3018 -0.452 0.6538   
## private1 14.9514 26.7319 0.559 0.5789   
## student\_faculty\_ratio:private1 -1.2949 1.0335 -1.253 0.2171   
## percent\_of\_classes\_under\_20:private1 0.2255 0.3806 0.593 0.5566   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.999 on 42 degrees of freedom  
## Multiple R-squared: 0.5994, Adjusted R-squared: 0.5517   
## F-statistic: 12.57 on 5 and 42 DF, p-value: 1.739e-07



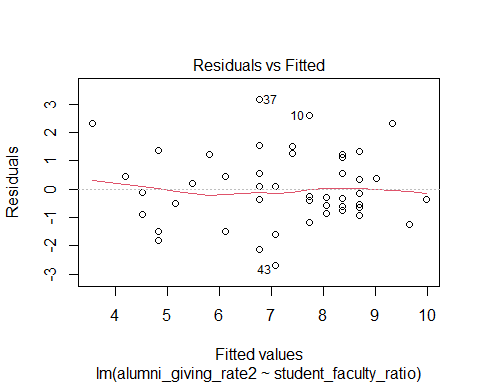
Using the interaction terms also we didn’t find any significant improvement in the model as well as predictors. So, our approach is to add the Box cox transformation to the non linearity. And also we used only student faculty ratio because it only has significance.

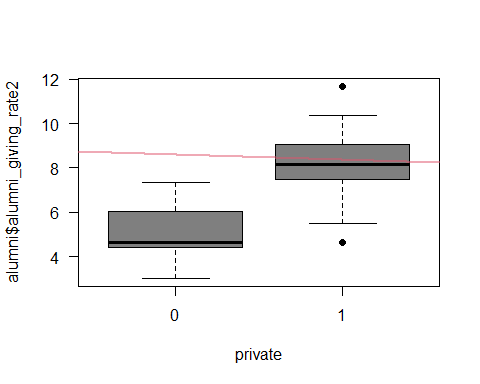
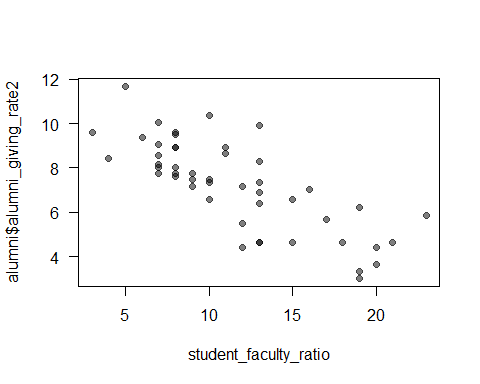
#### Box Cox Transformation:



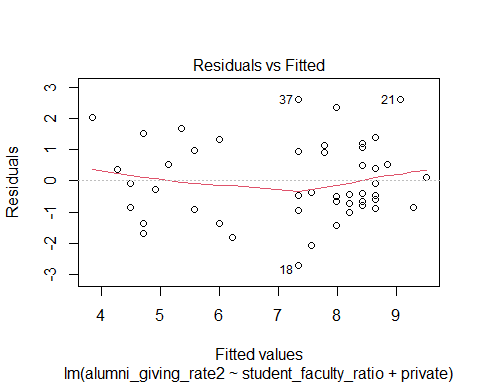
## [1] 0.4242424

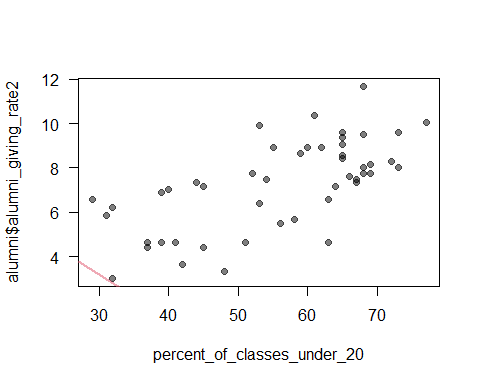
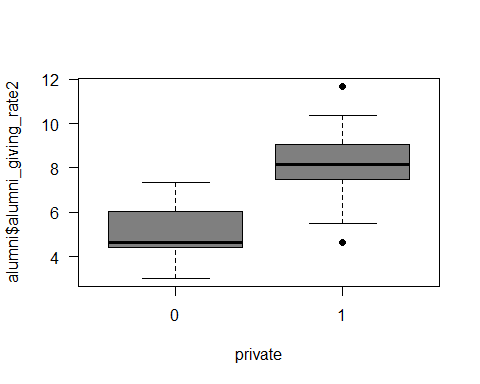
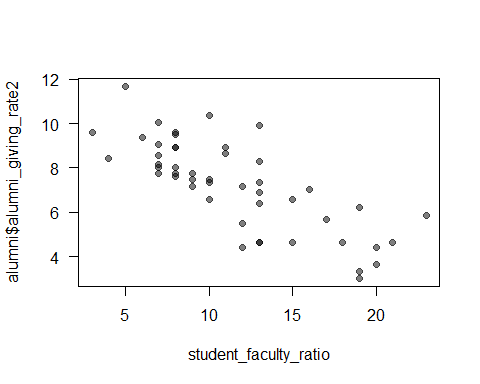
##   
## Call:  
## lm(formula = alumni\_giving\_rate2 ~ student\_faculty\_ratio, data = alumni)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.6870 -0.7874 -0.2050 0.6987 3.1568   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.95027 0.49036 22.331 < 2e-16 \*\*\*  
## student\_faculty\_ratio -0.32134 0.03923 -8.192 1.55e-10 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.305 on 46 degrees of freedom  
## Multiple R-squared: 0.5933, Adjusted R-squared: 0.5844   
## F-statistic: 67.1 on 1 and 46 DF, p-value: 1.546e-10



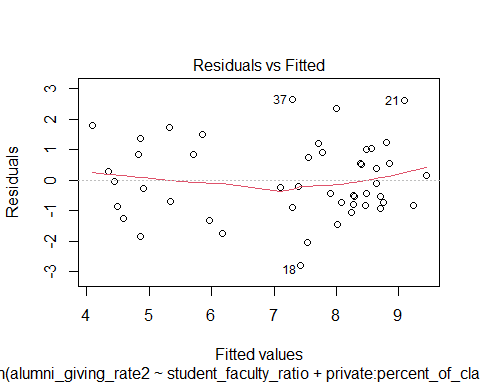


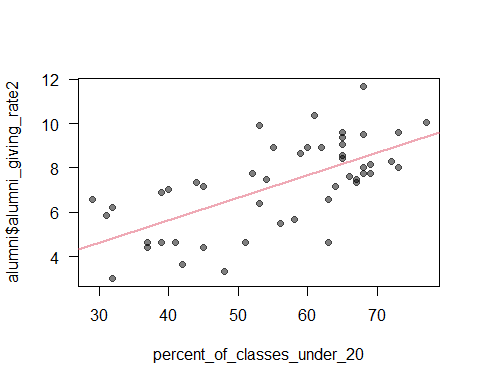
##   
## Call:  
## lm(formula = alumni\_giving\_rate2 ~ student\_faculty\_ratio + private,   
## data = alumni)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.7016 -0.8641 -0.3297 0.9600 2.5987   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.82417 1.18042 7.475 2.02e-09 \*\*\*  
## student\_faculty\_ratio -0.21658 0.06544 -3.310 0.00185 \*\*   
## private1 1.33369 0.67767 1.968 0.05524 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.266 on 45 degrees of freedom  
## Multiple R-squared: 0.6255, Adjusted R-squared: 0.6089   
## F-statistic: 37.58 on 2 and 45 DF, p-value: 2.525e-10



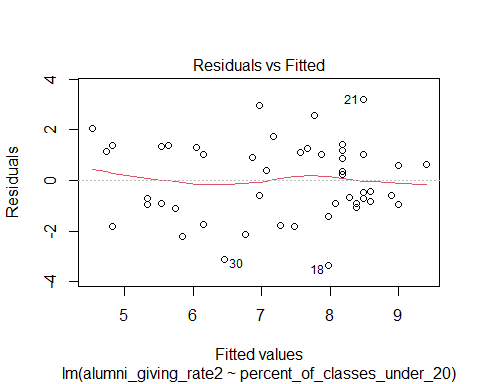


##   
## Call:  
## lm(formula = alumni\_giving\_rate2 ~ student\_faculty\_ratio + private:percent\_of\_classes\_under\_20,   
## data = alumni)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.7882 -0.8242 -0.2158 0.8649 2.6345   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.16615 1.88250 4.869 1.48e-05 \*\*\*  
## student\_faculty\_ratio -0.19848 0.07271 -2.730 0.00908 \*\*   
## private0:percent\_of\_classes\_under\_20 -0.01669 0.03059 -0.546 0.58801   
## private1:percent\_of\_classes\_under\_20 0.01338 0.02318 0.577 0.56663   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.268 on 44 degrees of freedom  
## Multiple R-squared: 0.6326, Adjusted R-squared: 0.6075   
## F-statistic: 25.25 on 3 and 44 DF, p-value: 1.178e-09





##   
## Call:  
## lm(formula = alumni\_giving\_rate2 ~ percent\_of\_classes\_under\_20,   
## data = alumni)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.3389 -0.9401 -0.0917 1.1491 3.1863   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.58279 0.97023 1.631 0.11   
## percent\_of\_classes\_under\_20 0.10154 0.01695 5.990 3e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.533 on 46 degrees of freedom  
## Multiple R-squared: 0.4382, Adjusted R-squared: 0.426   
## F-statistic: 35.88 on 1 and 46 DF, p-value: 2.995e-07



## **Variable selection (FS)**

##   
## Call:  
## lm(formula = alumni\_giving\_rate2 ~ . + .^2, data = alumni[, c(2,   
## 3, 5, 6)])  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.3949 -0.8651 -0.1731 0.8433 2.6731   
##   
## Coefficients:  
## Estimate Std. Error t value  
## (Intercept) -1.04665 7.43649 -0.141  
## percent\_of\_classes\_under\_20 0.21984 0.18829 1.168  
## student\_faculty\_ratio 0.41226 0.41366 0.997  
## private1 1.13591 3.79714 0.299  
## percent\_of\_classes\_under\_20:student\_faculty\_ratio -0.01421 0.01062 -1.339  
## percent\_of\_classes\_under\_20:private1 -0.05718 0.08744 -0.654  
## student\_faculty\_ratio:private1 0.22910 0.31368 0.730  
## Pr(>|t|)  
## (Intercept) 0.889  
## percent\_of\_classes\_under\_20 0.250  
## student\_faculty\_ratio 0.325  
## private1 0.766  
## percent\_of\_classes\_under\_20:student\_faculty\_ratio 0.188  
## percent\_of\_classes\_under\_20:private1 0.517  
## student\_faculty\_ratio:private1 0.469  
##   
## Residual standard error: 1.269 on 41 degrees of freedom  
## Multiple R-squared: 0.6571, Adjusted R-squared: 0.6069   
## F-statistic: 13.09 on 6 and 41 DF, p-value: 3.242e-08

## Start: AIC=70.53  
## alumni\_giving\_rate2 ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + student\_faculty\_ratio 1 114.199 78.288 31.223  
## + private 1 102.858 89.629 37.717  
## + percent\_of\_classes\_under\_20 1 84.351 108.135 46.727  
## <none> 192.486 70.535  
##   
## Step: AIC=31.22  
## alumni\_giving\_rate2 ~ student\_faculty\_ratio  
##   
## Df Sum of Sq RSS AIC  
## + private 1 6.2043 72.083 31.131  
## <none> 78.288 31.223  
## + percent\_of\_classes\_under\_20 1 1.6279 76.660 34.086  
##   
## Step: AIC=31.13  
## alumni\_giving\_rate2 ~ student\_faculty\_ratio + private  
##   
## Df Sum of Sq RSS AIC  
## <none> 72.083 31.131  
## + student\_faculty\_ratio:private 1 2.47163 69.612 33.328  
## + percent\_of\_classes\_under\_20 1 0.05895 72.024 34.963

##BIC

## wagereg1 wagereg2 wagereg3 wagereg4 fwdBIC   
## 31.22339 31.13138 34.09106 46.72712 31.13138

## wagereg1 wagereg2 wagereg3 wagereg4 fwdBIC   
## 0.3000322149 0.3141567021 0.0715253857 0.0001289953 0.3141567021

## **Discussion of Results**

**Following are the observations:**

* Based on the BIC forward selection resultys we choose the model with least BIC which is alumni\_giving\_rate2 ~ student\_faculty\_ratio + private
* Model statistics for best model is Adjusted R-squared: 61% and BIC value is: 31.13138
* Alumni rate is transformed using Lambda value of 0.4242424
* Intercept: 8.82417, Beta of student\_faculty\_ratio: -0.21658, Beta of private1: 1.33369