Alumni Donation Case Study (Group 3)

# INTRODUCTION

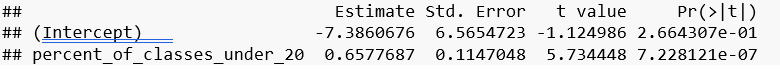
Alumni donations are a vital revenue source for universities. This study examines factors like student-teacher satisfaction, university type, national ranking, Ivy League affiliation, and location to quantify their impact on alumni donation rates. Through exploratory analyses and a streamlined linear regression model, we aim to provide administrators with actionable insights for policies to boost university revenues.

## EDA

Univariate Analysis

### percent\_of\_classes\_under\_20 (predictor variable) :

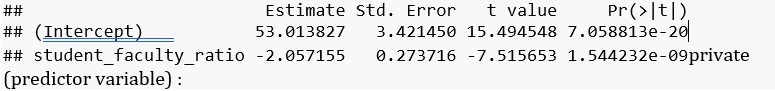
A graph of different sizes and shapes

Description automatically generated with medium confidence

The univariate regression model for this variable gives the following coefficients. With this low a p-value, we could expect this variable to influence the alumni giving rate.

### student\_faculty\_ratio (predictor variable) :

A graph of different sizes and shapes

Description automatically generated with medium confidence

The results of a univariate model is available below. Again, we can expect this variable to highly influence the predictor

A graph and a diagram

Description automatically generated A close up of numbers

Description automatically generated

In summary, the data indicates that public schools receive an average of 15.67 donation points, while private schools receive an additional 19.79 points on average.

A screenshot of a computer screen

Description automatically generated A screenshot of a computer

Description automatically generated

* On running this algortihm, we find the AIC to be 213.60 with the regression equation being
* Y=41.429−1.486∗studentfacultyratio+7.267∗private
* The summary shows that
  + R^2 = 57.47%
  + Adjusted R squared = 54.57%
  + p value is also less than 0.05, which means that null hypothesis can be rejected and alternate hypothesis can be accepted which implies there is a significant linear relation between the response variable with student\_faculty\_ratio and private

\*Observations +We see that there is no specific pattern in Residuals vs. Student Faculty Ratio graph +The Residual plot for Private is also expected as we see only two categorical values for this variable i.e. 0 and 1 +Alumni giving Rate Residual plot shows a linear trend which is somewhat concerning and violates the residuals assumption +The scatter plot for the fitted values is also on the same lines as student faculty ratio and hence it should follow a normal distribution along with constant variance just by visual inspection

A black text on a white background

Description automatically generated

The p-value from Shapiro test is < 0.05 which signifies that null hypothesis can be rejected which means that the residuals don’t follow Normal Distribution.

A black text on a white background

Description automatically generated

*To test the constant variance assumption, we observe the residuals versus fitted values plot* from the plot, we can observe that the residuals are increasing with the increase in fitted values. Hence, the assumption of constant error variance is violated, and we can say that the model is having heteroskedasticity. \*However, to confirm further we can use Breusch-Pagan test, where the p-value is greater than 0.05. Hence, we cannot reject null hypothesis. +H0 - the variance of residuals is constant (Null Hypothesis) +H1 - the variance of residuals is not constant (Alternate Hypothesis)

A close up of numbers

Description automatically generated

The **vif** values are less than the threshold value of 10, and hence we can conclude that there is no multicollinearity in the suggested model

## Remedial Measures and Transformation

Since the assumption of constant variance is violated (from the fitted vs Residual plot), we use the Box cox transformation to determine lambda value (0.3434343).

A graph with a curve

Description automatically generated with medium confidence

## Transformed Model Diagnostics

With the transformed response variable, we re-run the stepwise algorithm to determine the new predictors that forms the best fit model.

A screenshot of a computer

Description automatically generated

The AIC is significantly reduced to 0.18, with the regression equation being

The summary conclusions are:

* R-Squared = 62.98%
* Adjusted R squared = 61.34%
* p value is also less than 0.05, which means that null hypothesis can be rejected and alternate hypothesis can be accepted which implies there is a significant linear relation between the response variable with student\_faculty\_ratio and private.

A black text on a white background

Description automatically generated

The p-value from Shapiro test is > 0.05 which signifies that null hypothesis cannot be rejected which means that the residuals follow Normal Distribution.

**Breusch-Pagan test for constant variance**

##   
## studentized Breusch-Pagan test  
##   
## data: alumni\_fitAll2  
## BP = 2.8191, df = 3, p-value = 0.4204

* To test the constant variance assumption, we observe the residuals versus fitted values plot
* From the plot for constant variance, we see an improvement from model 1.
* However, to confirm further we can used Breusch-Pagan test, where the p-value is greater than 0.05. Hence, we cannot reject null hypothesis and conclude that the residuals show constant variance as compared to model 1.

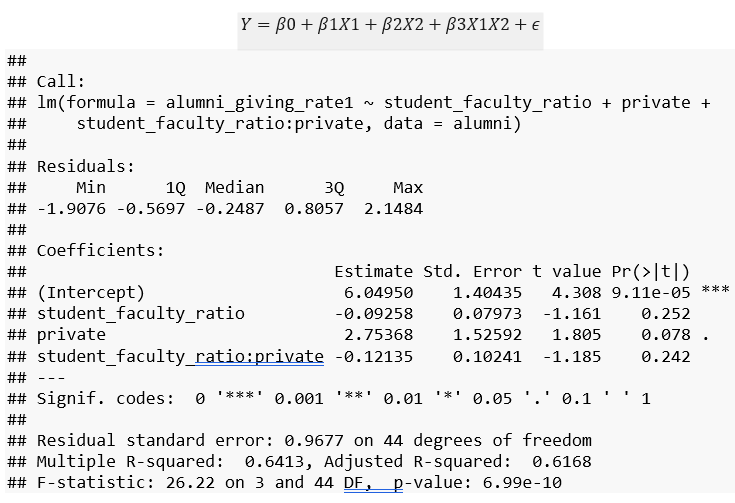
A close-up of a number

Description automatically generated

The vif values are less than the threshold value of 10, and hence we can conclude that there is no multicollinearity in the suggested model.

# Interactive Model

We will now explore another model that incorporates an interaction between the two predictor variables. This additional model aims to delve into the categorical nature of the ‘Private’ variable within the dataset.



From the summary stats of the interaction model, we observe that none of the predictor variables are statistically significant due to the p-values being larger than the threshold of 0.05. We hence conclude that the model’s performance in this case will not be as good as model 2 and we avoid further analysis.

### CONCLUSION

| Parameters | Model\_1 | Model\_2 | Model\_3 |
| --- | --- | --- | --- |
| Regression Equation | 𝑌=41.429−1.486 \* student\_faculty\_ratio+7.267 ∗ private | 𝑌=7.3245−0.1661 ∗ student\_faculty\_ratio +1.0529 ∗ private | 𝑌=6.04950−0.09258 ∗student\_faculty\_ratio + 2.75368 ∗ private + 0.12135 \* student\_faculty\_ratio ∗ private |
| R-squared value | 57.47% | 62.98% | 64.13% |
| Adjusted R-squared | 54.57% | 61.34% | 61.68% |
| Assumption of constant variance | Non - Constant | Constant | NA |
| Assumption of normality | Not Normal | Normal | NA |
| Multicollinearity | Not Observed | Not Observed | NA |
| AIC | 213.6 | 0.18 | NA |

We used the **stepwise algorithm** to determine the best model for the given *predictor variables*.

As per the conclusion table, we can see the best performance is shown by **Model 2**. The model was achieved through **Box-Cox variance stabilizing transformation**. Post the transformation, the model’s residuals showed a constant variance (Breusch-Pagan test) and also normality (Shapiro Wilk Test). The model also exhibited extremely low value of AIC reflecting the best fit in the given predictor variables.