Case Study - Alumni Donation Data

Akanksha Bajpai, Alankrit Mahajan, Ashish Solanki, Neha Meena, Nandita Ganesh

Introduction

Alumni donations significantly contribute to college revenues. By identifying key factors influencing higher alumni donation percentages, administrators can implement strategic policies to boost overall funding. Research suggests that greater faculty accessibility correlates with increased student satisfaction. Therefore, reducing class sizes and student-faculty ratios may enhance satisfaction, potentially leading to higher alumni contributions. This project aims to develop a linear regression model to analyze the variables impacting alumni donation rates, providing actionable insights for institutions to foster stronger alumni engagement and financial support for educational advancement.

Data Description

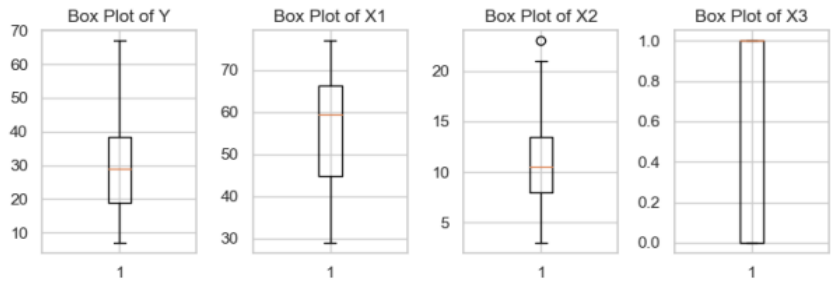
Variable	Description
school	The name of the universities
percent_of_classes_under_20	The percentage of classes offered with fewer than 20 students
student_faculty_ratio	The number of students enrolled divided by the total number of faculty
alumni_giving_rate	The percentage of alumni that made a donation to the university
private	A binary variable with '1' when the university is a private else it is '0'

Exploratory Data Analysis

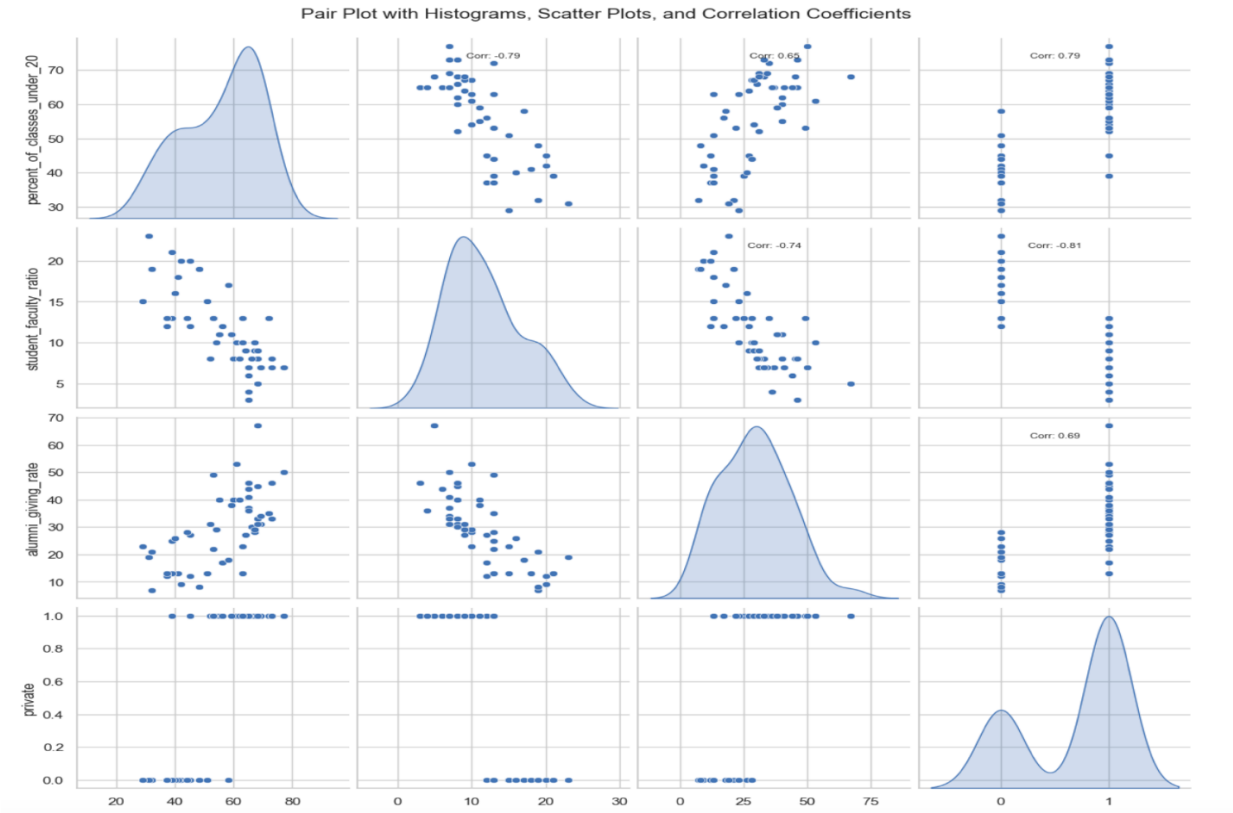
Summary of variables

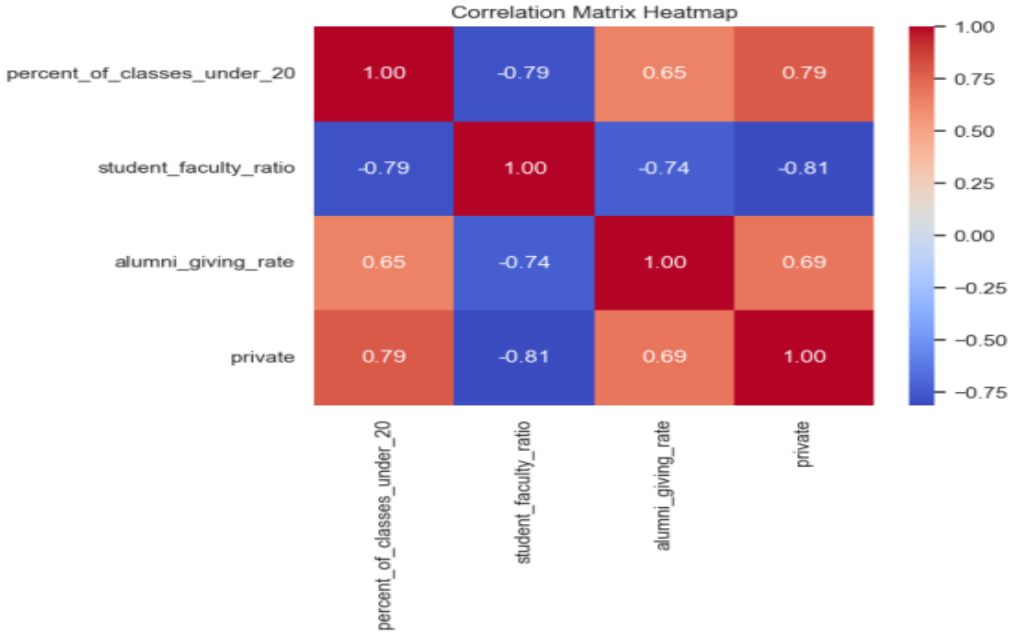
	Y_Summary	X1_Summary	X2_Summary	X3_Summary
count	48.00	48.00	48.00	48.00
mean	29.27	55.73	11.54	0.69
std	13.44	13.19	4.85	0.47
min	7.00	29.00	3.00	0.00
25%	18.75	44.75	8.00	0.00
50%	29.00	59.50	10.50	1.00
75%	38.50	66.25	13.50	1.00
max	67.00	77.00	23.00	1.00

Plot and observations



We can see in the Box plot of response variable Y and predictor variable (X1, X2), there are no outliers in the data. X3 is categorical data with binary values of 1 or 0. Therefore, there are no outliers in the predictor X3 variable data.





1. Y is linearly correlated to Predictor variable X1 in positive direction with correlation coefficient 0.645.
2. Y is linearly correlated to Predictor variable X2 in negative direction with correlation coefficient -0.742.
3. Response variable Y is correlated to Predictor variable X3 in with correlation coefficient of 0.689.
4. The strong correlations between predictor variables suggest that Multicollinearity could be present, which can affect the stability of the regression coefficients.

Model Analysis

We are using stepwise regression between baseline model (lm(Y ~ 1)) and (lm(Y ~ X1 + X2 + X3 + X2:X3 +X3:X1 + X1:X2) considering both forward and backward steps to identify a parsimonious model that balances goodness of fit and complexity. The approach involves sequentially adding or removing potential explanatory variables and conducting statistical significance tests after each iteration. The selection process is guided by evaluating AIC (Akaike Information Criterion) values produced by the model, aiming to refine the model by choosing the most relevant variables.

Model	AIC
Y ~ 1	250.43
Y ~ X2	213.98
Y ~ X2 + X3	213.6
Y ~ X2 + X3 + X2:X3	212.92

The stepwise regression process suggests that the model *Y = X2+X3+X2:X3 * with the interaction term X2:X3 has the lowest AIC value (212.92), indicating that it is the preferred model based on the AIC criterion.

```
## lm(formula = Y ~ X2 + X3 + X2:X3)
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.8235    12.8129   2.015   0.0500 *
## X2           -0.5860     0.7274  -0.806   0.4248
## X3           28.0851    13.9220   2.017   0.0498 *
## X2:X3        -1.4854     0.9343  -1.590   0.1190
## Residual standard error: 8.829 on 44 degrees of freedom
## Multiple R-squared:  0.596,  Adjusted R-squared:  0.5685
## F-statistic: 21.64 on 3 and 44 DF,  p-value: 9.206e-09
```

Since p values are not significant in this model, we will choose the model with the 2nd lowest AIC value (lm(formula = Y ~ X2 + X3).

```
## lm(formula = Y ~ X2 + X3)
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.4294     8.3734   4.948 1.09e-05 ***
## X2           -1.4863     0.4642  -3.202  0.00251 **
## X3            7.2669     4.8071   1.512  0.13761
## Residual standard error: 8.978 on 45 degrees of freedom
## Multiple R-squared:  0.5728, Adjusted R-squared:  0.5539
## F-statistic: 30.17 on 2 and 45 DF,  p-value: 4.877e-09
```

Conclusion: Based on the significance of coefficients and model fit statistics, *Model 2 (Y~X2+X3)* appears to be a better choice. It includes fewer variables, all of which are statistically significant, and it has a higher Adjusted R-squared and F-statistic.

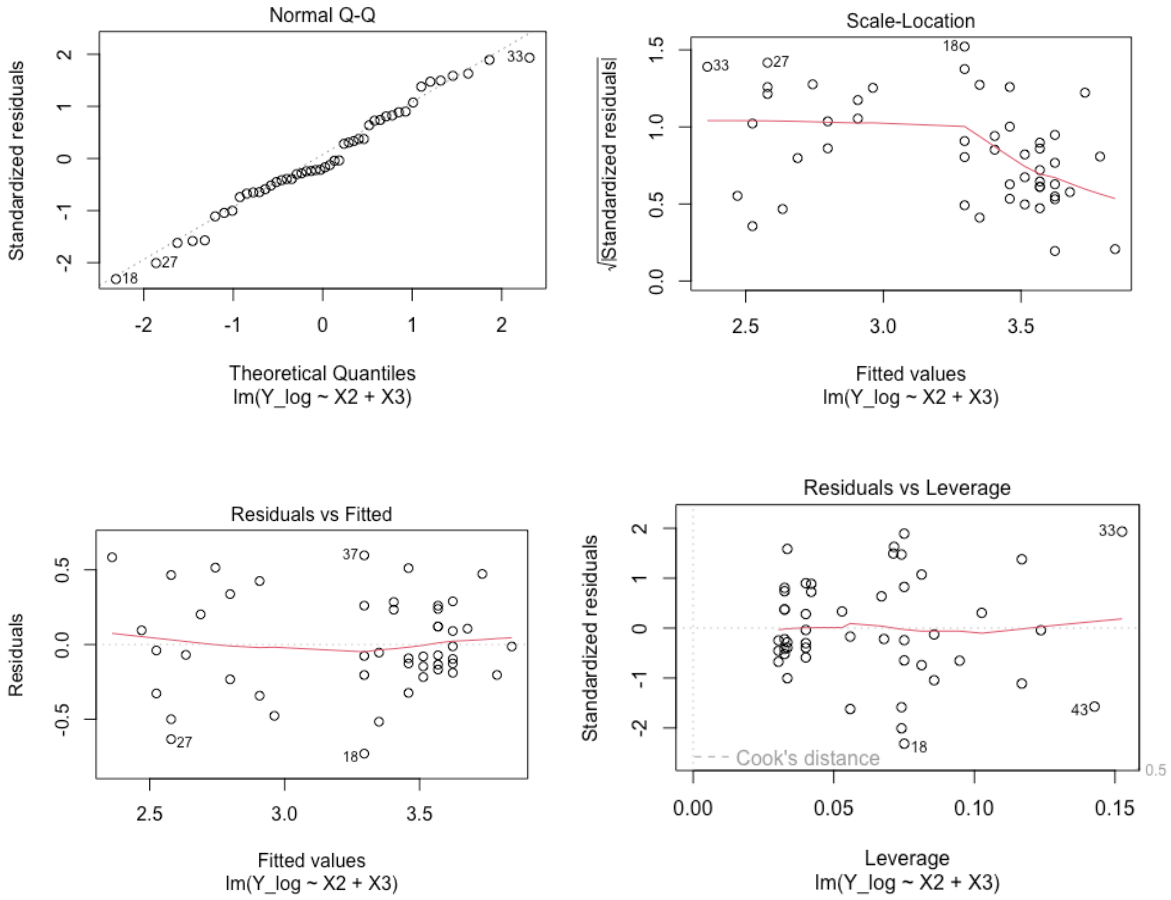
P-value for X3 is still not significant, we will apply different transformations.

Applying log transformation to Y

```
## lm(formula = Y_log ~ X2 + X3)
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.61808    0.30584  11.830 2.09e-15 ***
## X2          -0.05467    0.01696  -3.224  0.00235 **
```

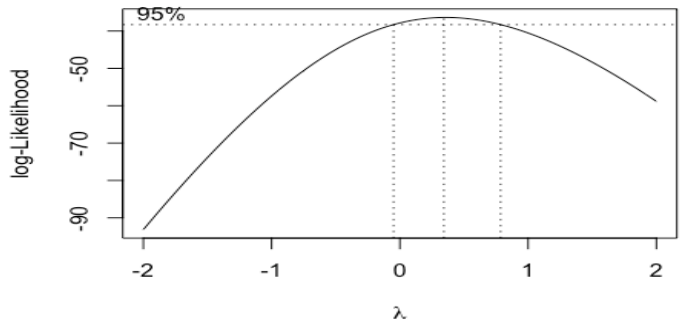
```
## X3          0.38773    0.17558   2.208  0.03236 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3279 on 45 degrees of freedom
## Multiple R-squared:  0.6383, Adjusted R-squared:  0.6222
## F-statistic:  39.7 on 2 and 45 DF,  p-value: 1.158e-10
```

The **log transformation** is often applied when the assumptions of linear regression are better met after transforming the response variable. Model 3 (Y_log~X2+X3) seems to perform better based on the lower residual standard error, higher adjusted R-squared, and higher F-statistic.



We will try to tune the model further to improve the value of adjusted R^2, AIC and BIC.

Applying **Box-Cox transformation**, which is a statistical technique used to stabilize the variance and make a dataset more closely approximate a normal distribution.



```
> model4 <- lm(Y_transformed ~ X2 + X3)
> summary(model4)

Call:
lm(formula = Y_transformed ~ X2 + X3)

Residuals:
     Min       1Q   Median       3Q      Max
-2.1031 -0.6559 -0.2420  0.7583  1.9529

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.32445    0.90662   8.079 2.65e-10 ***
X2          -0.16613    0.05026  -3.305  0.00187 **
X3           1.05289    0.52049   2.023  0.04905 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9721 on 45 degrees of freedom
Multiple R-squared:  0.6298,    Adjusted R-squared:  0.6134
F-statistic: 38.28 on 2 and 45 DF,  p-value: 1.946e-10
```

We are using stepwise regression between baseline model lm(Y_transformed ~ 1) and lm(Y_transformed ~ X1 + X2 + X3).

The stepwise regression process suggests that the model4 lm(Y_transformed ~ X2 + X3) has the lowest AIC value, indicating that it is the preferred model based on the AIC criterion among other boxcox models.

Now we have two model to compare as both the models are giving significant values for AIC, BIC, R-Square and adj R square.

Comparing model3 (lm(Y_log ~ X2 + X3) and model4 (lm**(Y_transformed ~ X2 + X3))**

Model	AIC	BIC	R-squared	Adj R-squared	F Statistic	Residual Standard Error
Y_log ~ X2 + X3	-104.1355	-106.8397	0.6383	0.6222	39.7	0.3279
Y_transformed ~ X2 + X3(using boxcox)	0.18	-2.520596	0.6298	0.6134	38.28	0.9721
Significance	Prefer lower AIC during model selection	Prefer lower BI C during model selection	Higher values of R-squared indicate a better fit	Higher values of adjusted R-squared indicate a better fit	higher F-statistic indicates a better fit.	Prefer lower residual standard error

Considering all these parameters, the model with the response variable Y_{log} and predictors X_2 and X_3 seems to be better based on the information above.

Final linear model equation:

Y=exp($\beta 0$+$\beta 2 \cdot X2$+$\beta 3 \cdot X3$) = exp (3.61-0.05 $X2$ + 0.38 $X3$)

Limitation

- Limited by dataset size and diversity, which may affect the robustness of conclusions.
- Incomplete inclusion of relevant variables may hinder model comprehensiveness.
- Simplicity of modeling techniques may overlook nuanced relationships.