Homework 5

BANA 7052: Applied Linear Regression

Section 1: Wednesday 6:00 – 9:50

Group 5

Tess Newkold

Aabhaas Sethi

Anirudh Chekuri

James Brand

Abstract

Alumni donations are a crucial part of a universities' revenue. Understanding what factors are contributing to a higher donation rate and implementing a few changes could lead to an increase in donations by future alumni. We used data from America's Best Colleges which included 48 national universities and four variables to analyze these factors. Several linear regression models were made to determine best fit. Our final model included student to faculty ratio and private/public variables in relation to alumni giving. We concluded that a lower student to faculty ratio results in higher percentage of alumni giving, and private schools also had higher alumni giving. Since a university cannot change whether it is a private or public university, we suggest decreasing the student to faculty ratio. This may lead to an increase in alumni donations and thus an increase in revenue.

Data Description

The data has 5 variables and 48 observations. Three variables (*Percent of classes under 20, Student Faculty Ratio, Alumni Giving Rate*) are continuous and two variables *(I School, Private*) are categorical. The box plots (Figure 1) reveals no outliers, but we can observe that the variable *Percent of Classes Under 20* is left skewed. To get a better view, we can look at histograms. The histograms reveal *Percent of Class Under 20* is indeed left skewed, and *Student Faculty Ratio* is right skewed (Figure 2). Summaries of variables are shown in Table 1 and Table 2. The first model we tried is Y = X1 + X2 + X3 (Table 3) and summary statistics shown below (Table 4). Looking at residual diagnostics the Residuals vs Fitted plot, the predictor and outcome variables seem to have a nonlinear relationship, but this small pattern can be ignored. From the Normal Q-Q plot, the residuals have a have distribution that is almost similar to normal distribution. From the Scale-Location plot, as the fitted values increased, the model had more values distributed to the right. The assumption of homoscedasticity is not perfectly applied in the model. From the Residuals vs Leverage plot, there are very few influential points in the model (Figure 3).

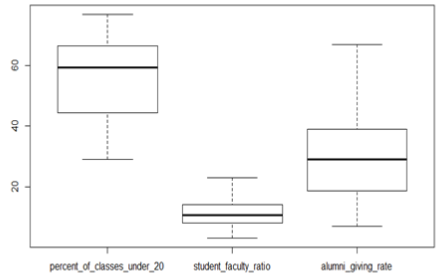

Figure 1. Boxplots of three variables, Percent of classes under 20, Student Faculty Ratio, Alumni Giving Rate

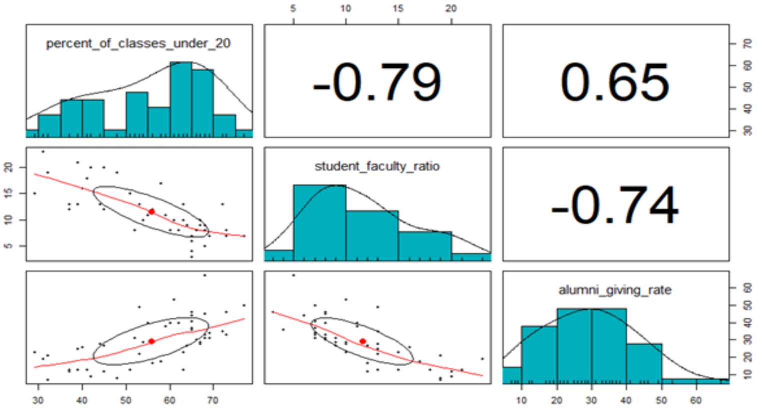

Figure 2. Correlation matrix of the three continuous variables.

Variable	Min	Median	Mean	Max	Std. Dev.
% of Classes Under 20	29	59.5	55.7	77	13.19
Student Faculty Ratio	3	10.5	11.5	23	4.85
Alumni Giving Rate	7	29	29.3	67	13.44

Table 1. Continuous variable summary

Variable	Number of 0's	Number of 1's
Private	15	33

Table 2. Categorical variable summary

Model	University Type
Alumni Giving Rate = 43.071+0.077*percent of Classes under 20 - 1.398*Student Faculty Ratio Y = 43.071 + 0.077*X1 – 1.398*X2	Private
Alumni Giving Rate = 36.654+0.077*percent of Classes under 20 - 1.398*Student Faculty Ratio Y = 36.654 + 0.077*X1 – 1.398*X2	Non-private

Table 3. First model shown in private universities and non-private universities

Variable	Coefficient	Std. Error	P Value	Adjusted R^2
Student Faculty Ratio	-1.398	0.511	0.009	0.546
Percent of classes under 20	0.077	0.179	0.668	
Private	6.280	5.356	0.246	

Table 4. Summary statistics for first model

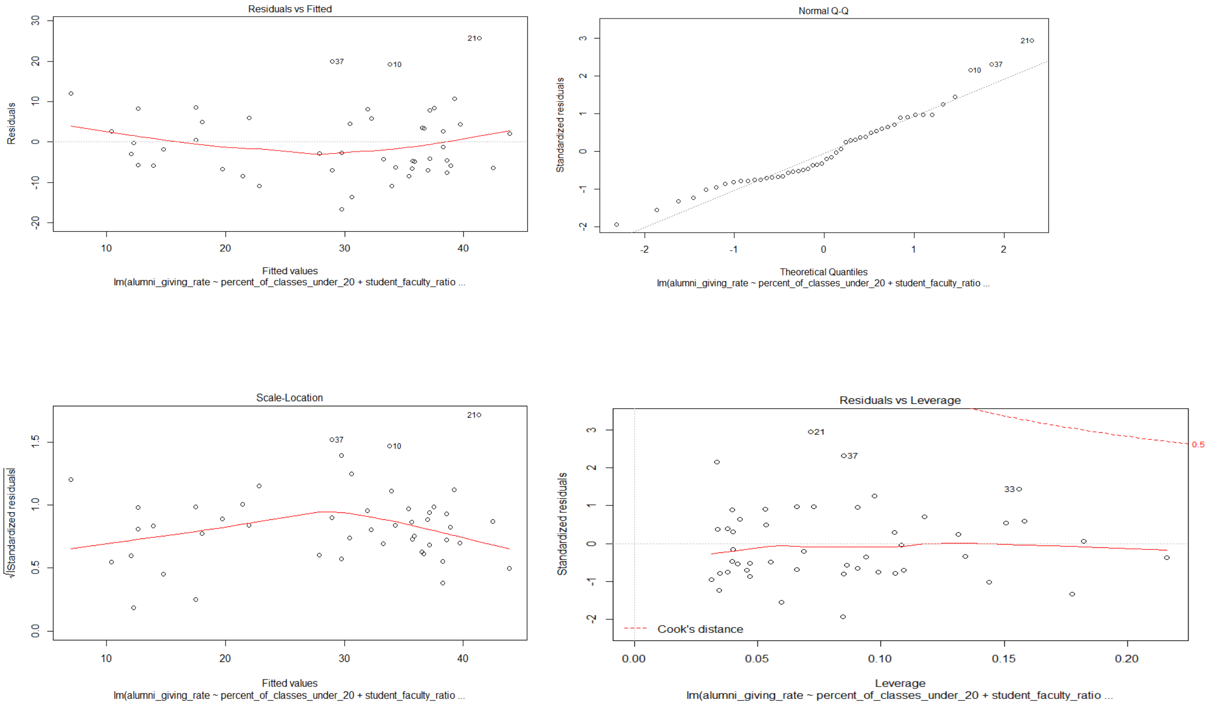

Figure 3. Residual Diagnostic plots, starting at top left and going clock-wise, Residuals vs. Fitted plot, Normal Q-Q plot, Scale-Location plot, Residuals vs. Leverage plot.

Methods

We started our analysis by building a multiple linear regression model with each of the three predictors. This beginning model used all four columns in their original form (without any transformations). The F-test for this first model indicated that the model was significant overall. However, only X2 (Student Faculty Ratio) was a significant predictor in the pairwise t-tests. X1 (% of Classes Under 20 Students) was the least significant predictor, so we eliminated it from the model. The new model with just X2 and X3 (Private School Yes/No) was much more promising. This model also improved the Adjusted R^2. We tried several other models with different combinations of predictor variables and interactions, but this model with only X2 and X3 remained our favorite. It is worth noting that we attempted quadratic transformations on X1 and X2, but these made our model needlessly complex and did not offer any substantial improvement to the Adjusted R^2. We also thoroughly investigated including the interaction between X2 and X3 in our model, to allow the slopes to differ for private and public universities. However, based on our plots and model metrics, we decided that a same slope – different intercept model fit our data better. Some of the models that used both X1 and X2 as predictors suffered from multicollinearity, since there is a relatively strong linear relationship between X1 and X2. Once we had decided to use a model with X2 and X3 as predictors, we examined the residual diagnostics. The plot of studentized residuals vs. fitted Y values showed a possible violation of the constant variance assumption. The variance seemed to increase as the fitted values of Y increased, which indicated a violation of homoscedasticity. We applied a Box-Cox transformation (lambda ≈ 0.34) to Y to resolve this issue. The residual plots of the Box-Cox transformed model satisfy our constant variance assumption.

Results

Our final regression line created after regressing the transformed response variable against two predictors, student faculty ratio and private (Figure 4) is shown below. We can see that the alumni giving rate (after transformation) decreases by 0.16613 for every unit increase in the student faculty ratio. However, the intercepts for the private and non-private universities are different. Also, the alumni giving rate is higher for private universities than the non-private universities. It decreases as the student faculty ratio increases for both types of universities. The final model is shown below (Table 5) split between private and non-private universities. The hypothesis tested is that the coefficients of the predictor variables are 0 (individually). We can see that p-value of both these predictors are less than 0.05. Hence, we reject the hypothesis and both these variables are significant. The F test also gives p value less than 0.05 indicating that we have sufficient evidence to reject the hypothesis that coef. of student faculty ratio = coef. Of Private = 0. The adjusted Rsq (0.613) is better than most of the other models that we tried (Table 6). The standard residual vs fitted values plot indicate that the assumption of linearity and constant variance is being met as the data points are scattered randomly across the plot. The Q-Q plot does not indicate any substantial deviation from the normality (Figure 5). Hence, the assumptions for linear regression are being met by our model.

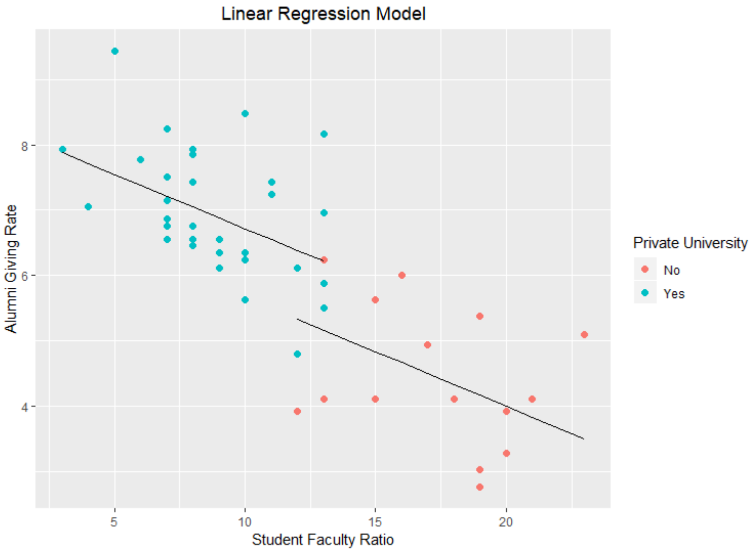

Figure 4. Final regression line

Model	University Type
Transformed Alumni Giving Rate = 8.377 - 0.166 * Student Faculty Ratio $(Y^{0.34} - 1)/0.34 = 8.377 – 0.166*X2$	Private
Transformed Alumni Giving Rate = 7.324 - 0.166 * Student Faculty Ratio $(Y^{0.34} - 1)/0.34 = 7.324 – 0.166*X2$	Non-private

Table 5. Final regression model split between private and non-private universities

Variable	Coefficient	Std. Error	P Value	Adjusted R^2
Student Faculty Ratio	-0.166	0.050	0.002	0.613
Private	-0.166	0.520	0.049	

Table 6. F test summary

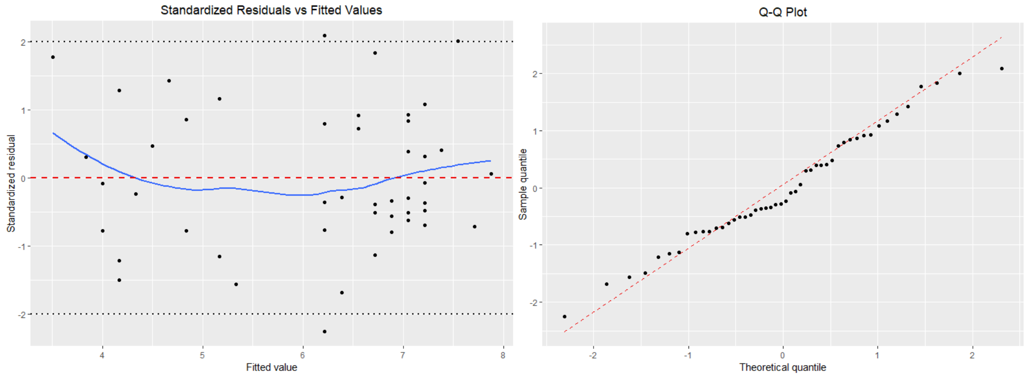

Figure 5. Residual diagnostic charts, standardized residuals vs. fitted values and Q-Q plot.

Discussion

Our results show a very interesting connection between alumni donations, student to faculty ratio, and private vs public schools. Universities with high student to faculty ratios receive fewer alumni donations, and public universities receive even fewer alumni donations. On average, public universities also have higher student to faculty ratios, so the effect is magnified. The data set used in this analysis consists of just 48 universities. It would be interesting to see if our final model could accurately predict alumni giving for all universities.

References

Greenwell, Brandon. BANA 7052: Applied Linear Regression. (2018). GitHub Repository. https://github.com/bgreenwell/uc-bana7052.

Kutner, M. H., Nachtsheim, C., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). Boston, MA: McGraw-Hill.