NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

BC3406 BUSINESS ANALYTIC CONSULTING

Alumni Giving Case Analysis

Name: Kong Ming Yeow

Matriculation: U2010866H

Submitted on 17 February 2023

Number of Pages: 3

Problem Statement

The problem identified is to uncover the key drivers that contributes to the alumni giving rate, and to recommend solutions for State University to undertake to improve its existing giving rate.

Establishing Significant Level

With consideration on the complexity of model and sample size of 123, we believe that α = 5% level of significance will be sufficient in providing Ms Madison the confidence to make strategic choice for improving the giving rate for State University which leads to more cash flow for the university.

Data Exploration (Correlation Matrix)

row	column	cor	p-value
SFR	LT20	-0.6919	0.00E+00
SFR	GT50	0.414978	1.82E-06
LT20	GT50	-0.58152	1.75E-12
SFR	GRAD	-0.60457	1.31E-13
LT20	GRAD	0.487248	1.10E-08
GT50	GRAD	0.017023	8.52E-01
SFR	FRR	-0.52198	5.96E-10
LT20	FRR	0.376735	1.75E-05
GT50	FRR	0.055766	5.40E-01
GRAD	FRR	0.934396	0.00E+00
SFR	GIVE	-0.54924	4.76E-11
LT20	GIVE	0.540427	1.11E-10
GT50	GIVE	-0.1751	5.27E-02
GRAD	GIVE	0.68166	0.00E+00
FRR	GIVE	0.647625	4.44E-16

From this correlation matrix, we see that GRAD and FRR has a strong correlation with a high correlation coefficient of 0.9344. This could pose as a significant issue on collinearity when these 2 variables are used as the independent variables. Hence, this matter should be investigated later on during the construction of regression model.

Furthermore, we will try to prove that there is a significant correlation between the independent variables, X and the dependent variable, Y (GIVE) through hypothesis testing.

$$H_0: p = 0 \ (no\ linear\ correlation)$$
$$H_A: p \neq 0 \ (linear\ correlation)$$

With a p-value of 0.0527 between GT50 and GIVE, we accept the null hypothesis that there is no linear correlation between these 2 variables.

Hence, with this correlation analysis, we should consider the possible of collinearity or multicollinearity between the independent variables and that there is no linear relationship between GT50 and GIVE before the construction of our regression model.

Question 1:

From this question, we are assuming that the other factors which are significant to the giving rate are not present in this analysis, and only the graduation rate was into consideration. Hence, a regression model can be derived with the equation below:

$$\sqrt{GIVE} = \beta_0 + \beta_1 \, GRAD$$

$$\sqrt{GIVE} = 0.08362 + 0.43120 \, GRAD$$

A square root transformation model will be used instead of a linear regression due to the violation of assumptions from the linear regression model as seen in Appendix A and B. With a square root transformation, the residuals of the regression model will yield normally distributed and equal residuals, this means that the model obtained is more statistically accurate, but it has become less interpretable.

With ten points (10%) increase in the graduation rate,

$$\boldsymbol{Increase\ in\ GIVE = 0.03719\ (GRAD) + 0.009071}$$

We will see an 0.03719 (GRAD) + 0.009071 increase in giving rate. This means that for instance, school B's graduation rate is 70%, we will expect the increase in giving rate to be 0.03719 (0.7) + 0.009071 = 3.5104%.

However, this only shows that there is an association between giving rate and graduation rate and does not imply causality. With an explanation power of 0.49 on this current model, this means that 49% of the giving rate can be explained by the variable, graduation rate and more analysis should be conducted.

Question 2:

From this question, we are only taking into consideration the graduation rate and student-to-faculty ratios (SFR), as such, a multiple regression model can be derived with the equation below:

$$\sqrt{GIVE} = \beta_0 + \beta_1\ GRAD + \beta_2\ SFR$$

$$\sqrt{GIVE} = 0.194951 + 0.367583\ GRAD - 0.003955\ SFR$$

A square root transformation model will be used again instead of a linear regression due to the violation of assumptions from the linear regression model as seen in Appendix C and D.
While assuming the School A and B have the identical SFR at the median of 18%:

$$\boldsymbol{Increase\ in\ GIVE = 0.02702\ (GRAD) + 0.01563}$$

This means that for instance school B's graduation rate is 70%, we will expect the increase in giving rate to be 0.02702 (0.7) + 0.01563 = 3.4544%.

When SFR is added in the consideration, we will see that the increase School A's giving rate will be slightly lower than that in Question 1. This is because SFR has an inverse effect on the giving rate, hence, when SFR is being factored into the equation, the overall giving rate will decrease even though the graduation rate remains constant at 10 points.

Question 3:

With reference to the data analysis in Appendix E, we can obtain the optimal regression model which variables and residuals that are statistically significant to measure the giving rate.

$$\sqrt{GIVE} = -0.30618 + 0.25391\ LT20 + 0.67233\ FRR$$

Before answering the question, we need to establish an understanding on the term "impressive giving rate." In this analysis, an "impressive giving rate" will be one who is doing better that the predicted giving rate derived from the regression model. As such, the spread between the actual giving rate and the predicted giving rate, which is called the residuals will be used to determine which schools has the most and least impressive giving rate.

Schools	LT20	FRR	GIVE	Predicted GIVE	Residuals
University of California Berkeley	62%	97%	12%	25.34%	-13.34%
Auburn University	24%	87%	31%	11.54%	19.46%

From our analysis, we see that University of California Berkeley has the least impressive giving rate, with a residual of -13.34%, this means that the university is underperforming than the forecasted giving rate. Meanwhile, Auburn University has the most impressive giving rate, with a residual of 19.46%, this implies that the university is performing better than the forecasted giving rate, exceeding the model expectation.

Question 4:

With reference to the data analysis in Appendix E, we will use the regression model to measure the giving rate for this question:

$$\sqrt{GIVE} = -0.30618 + 0.25391\,LT20 + 0.67233\,FRR$$

While only taking into variables that are statistically significant, State University has 34% of classes with fewer than 20 students and a freshman retention rate of 77%:

$$GIVE = [-0.30618 + 0.25391\,(0.34) + 0.67233\,(0.77)]^2$$

$$\boldsymbol{GIVE = 8.8711\%}$$

The estimated giving rate based on the constructed regression equation for State University is 8.8711% which is 0.8711% lesser than the current alumni giving rate of 8%. Hence, State University can consider the following recommendations in the next section to improve their current giving rate.

Recommendations:

With an alumni giving rate of 8%, State University is considered on the low end of the spectrum. I would suggest the following to the school to increase its LT20 and FRR which proved to be statistically significant in contributing to the giving rate from this analysis:

1. Even though SFR is not concluded statistically significant in this analysis, but in practical sense, the school should look into lowering the student-to-faculty ratio if possible. This will help to increase the percentage of classes with fewer than 20 students, thereby increasing student satisfaction with the school focused education system, and statistically increasing the amount of after-graduation donations.
2. The school should aim to increase its freshman retention rate by create a sense of belongings through a great freshman orientation. According to the Connected Student Report, students with a great onboarding experience are 35 times more likely to have a great overall university experience. Additionally, there is a 73% correlation between a positive onboarding experience and a positive overall university experience, showing that belonging is foundational for student retention (Salesforce, n.d.).
3. The school can focus on building relation with the graduated student by organizing alumni meeting. This can help to build strong bonds between students and university, and in hopes leads to alumni students donating more funds than the usual member.

Limitations:

I understand that State University is a large public university that is well reputed in sports, these athletic programs help to create a bond between the athletes and the school. But with most of the students being commuters, this creates a problem in creating the bond between the students and the university.

Appendix A: Q1 Regression Model and Testing of Assumptions

Linear Regression Model:

$$GIVE = \beta_0 + \beta_1\, GRAD$$

	Estimate	**Std. Error**	**T-Value**	**P-Value**
Intercept	-0.06718	0.02108	-3.187	0.00183 **
GRAD	0.32387	0.03160	10.248	< 2e-16 ***

Meeting Assumption of Regression Model:

1. Linear Relationship between X and Y

This assumption has been fulfilled in the correlation matrix above.

2. Independence of X and Y

This assumption has been fulfilled in the correlation matrix above.

3. Normality of Residuals

The Shapiro-Wilk test will be used to test the normality of the residuals. Hence, we test the following null hypothesis:

$$H_0 = Residuals\ are\ normally\ distributed$$

Against the alternative hypothesis:

$$H_A = Residuals\ are\ not\ normally\ distributed$$

From the test, with a P-value of 0.0009355, which is lesser than the threshold of 0.05, the null hypothesis can be rejected, thus the **residuals are not normally distributed**, violating the assumptions of this regression model.

4. Homoscedasticity

The Breusch-Pagan test will be used to test for heteroscedasticity of errors in regression. Hence, we test the following null hypothesis:

$$H_0 = Error\ variance\ is\ Constant\ (Homoscedasticity)$$

Against the alternative hypothesis:

$$H_A = Error\ variance\ is\ Not\ Constant\ (Heteroscedasticity\ )$$

From the test, with a P-value of 0.001799, which is lesser than the threshold of 0.05, the null hypothesis can be rejected, thus **heteroscedasticity is assumed** and violating the assumptions of this regression model.

Since, the linear regression model violates the assumptions, the result from this model may be incorrect or misleading. Thus, a data transformation can be used to address this violation.

Appendix B: Q1 Regression Model – Square Root Transformation

Squared Root Transformation of Regression Model:

$$\sqrt{GIVE} = \beta_0 + \beta_1\, GRAD$$

	Estimate	**Std. Error**	**T-Value**	**P-Value**
Intercept	0.08362	0.02672	3.13	0.00219 **
GRAD	0.43120	0.04006	10.77	< 2e-16 ***

Meeting Assumption of Regression Model:

1. Linear Relationship between X and Y

This assumption has been fulfilled in the correlation matrix above.

2. Independence of X and Y

This assumption has been fulfilled in the correlation matrix above.

3. Normality of Residuals

The Shapiro-Wilk test will be used to test the normality of the residuals. Hence, we test the following null hypothesis:

$$H_0 = Residuals\ are\ normally\ distributed$$

Against the alternative hypothesis:

$$H_A = Residuals\ are\ not\ normally\ distributed$$

From the test, with a P-value of 0.5743, which is higher than the threshold of 0.05, the null hypothesis cannot be rejected, thus the **residuals are normally distributed**.

4. Homoscedasticity

The Breusch-Pagan test will be used to test for heteroscedasticity of errors in regression. Hence, we test the following null hypothesis:

$$H_0 = Error\ variance\ is\ Constant\ (Homoscedasticity)$$

Against the alternative hypothesis:

$$H_A = Error\ variance\ is\ Not\ Constant\ (Heteroscedasticity\ )$$

From the test, with a P-value of 0.4734, which is higher than the threshold of 0.05, the null hypothesis cannot be rejected, thus **homoscedasticity is assumed.**

Hence, the square root transformation of model is able to satisfy all 4 assumptions and will be statistically significant in this analysis.

Appendix C: Q2 Regression Model and Testing of Assumptions

Linear Regression Model:

$$GIVE = \beta_0 + \beta_1 \, GRAD + \beta_2 \, SFR$$

	Estimate	Std. Error	T-Value	P-Value
Intercept	0.041475	0.045846	0.905	0.36746
GRAD	0.261793	0.038722	6.761	5.27e-10 ***
SFR	-0.003860	0.001455	-2.652	0.00908 **

Meeting Assumption of Regression Model:

1. Normality of Residuals

The Shapiro-Wilk test will be used to test the normality of the residuals.

From the test, with a P-value of 0.001021, which is lesser than the threshold of 0.05, the null hypothesis can be rejected, thus the **residuals are not normally distributed**, violating the assumptions of this regression model.

2. Homoscedasticity

The Breusch-Pagan test will be used to test for heteroscedasticity of errors in regression.

From the test, with a P-value of 0.03688, which is lesser than the threshold of 0.05, the null hypothesis can be rejected, thus **heteroscedasticity is assumed** and violating the assumptions of this regression model.

Since, the linear regression model violates the assumptions, the result from this model may be incorrect or misleading. Thus, a data transformation can be used to address this violation.

Appendix D: Q2 Regression Model – Square Root Transformation

	Estimate	Std. Error	T-Value	P-Value
Intercept	0.194951	0.058694	3.321	0.00119 **
GRAD	0.367583	0.049574	7.415	1.89e-11 ***
SFR	-0.003955	0.001863	-2.122	0.03585 *

Meeting Assumption of Regression Model:

1. Normality of Residuals

The Shapiro-Wilk test will be used to test the normality of the residuals.

From the test, with a P-value of 0.5629, which is higher than the threshold of 0.05, the null hypothesis cannot be rejected, thus the **residuals are normally distributed**.

2. Homoscedasticity

The Breusch-Pagan test will be used to test for heteroscedasticity of errors in regression.

From the test, with a P-value of 0.7017, which is higher than the threshold of 0.05, the null hypothesis cannot be rejected, thus **homoscedasticity is assumed**.

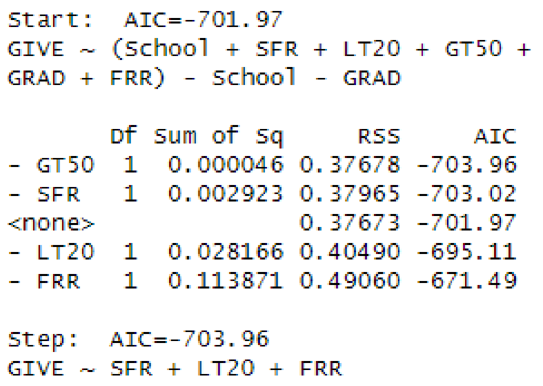
Appendix E: Q3 - Construction of Regression Model

Checking for Multicollinearity with Variance Inflation Factor (VIF)

Variables	**SFR**	**LT20**	**GT50**	**GRAD**	**FRR**
VIF	2.490582	2.859001	1.971972	10.667675	8.481457
VIF (W/O Grad)	2.350077	2.520429	1.808985	NA	1.645934

As mentioned in the correlation matrix, we suspect that there may be multicollinearity present between the independent variables. As multicollinear variables can skew the regression models, VIF is used to identify and remove these variables. Thus, with a VIF of 10.6677 which is above the threshold of 5, **GRAD is being removed from the model**. After GRAD has been removed, the VIF for the remaining variables remain within the threshold and there is no longer issue of multicollinearity.

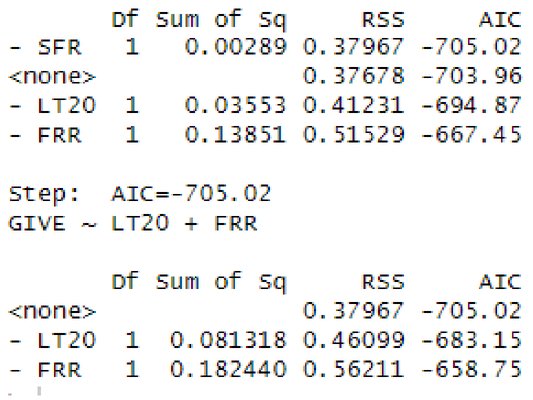
Removing Insignificant Variables using Backward Elimination with Akaike Information Criteria (AIC)



```
Start:  AIC=-701.97
GIVE ~ (School + SFR + LT20 + GT50 +
GRAD + FRR) - School - GRAD

       Df Sum of Sq     RSS     AIC
- GT50  1  0.000046 0.37678 -703.96
- SFR   1  0.002923 0.37965 -703.02
<none>              0.37673 -701.97
- LT20  1  0.028166 0.40490 -695.11
- FRR   1  0.113871 0.49060 -671.49

Step:  AIC=-703.96
GIVE ~ SFR + LT20 + FRR
```



```
       Df Sum of Sq     RSS     AIC
- SFR   1   0.00289 0.37967 -705.02
<none>              0.37678 -703.96
- LT20  1   0.03553 0.41231 -694.87
- FRR   1   0.13851 0.51529 -667.45

Step:  AIC=-705.02
GIVE ~ LT20 + FRR

       Df Sum of Sq     RSS     AIC
<none>              0.37967 -705.02
- LT20  1  0.081318 0.46099 -683.15
- FRR   1  0.182440 0.56211 -658.75
```

Linear Regression Model:

After removing multicollinear variables and insignificant variables, we are left with the following:

$$GIVE = \beta_0 + \beta_1 \, LT20 + \beta_2 \, FRR$$

	Estimate	**Std. Error**	**T-Value**	**P-Value**
Intercept	-0.36055	0.05145	-7.008	1.52e-10 ***
LT20	0.20815	0.04106	5.070	1.47e-06 ***
FRR	0.49732	0.06549	7.594	7.48e-12 ***

Meeting Assumption of Regression Model:

1. Normality of Residuals

The Shapiro-Wilk test will be used to test the normality of the residuals. From the test, with a P-value of 0.0006026, which is lesser than the threshold of 0.05, the null hypothesis can be rejected, thus the **residuals are not normally distributed**, violating the assumptions of this regression model.

2. Homoscedasticity

The Breusch-Pagan test will be used to test for heteroscedasticity of errors in regression. From the test, with a P-value of 0.03592, which is lesser than the threshold of 0.05, the null hypothesis can be rejected, thus **heteroscedasticity is assumed** and violating the assumptions of this regression model.

Since, the linear regression model violates the assumptions, the result from this model may be incorrect or misleading. Thus, a data transformation can be used to address this violation.

$$\sqrt{GIVE} = \beta_0 + \beta_1 \, LT20 + \beta_2 \, FRR$$

	Estimate	Std. Error	T-Value	P-Value
Intercept	-0.30618	0.06632	-4.617	9.86e-06 ***
LT20	0.25391	0.05292	4.798	4.66e-06 ***
FRR	0.67233	0.08442	7.964	1.07e-12 ***

Meeting Assumption of Regression Model:

1. Normality of Residuals

The Shapiro-Wilk test will be used to test the normality of the residuals.

From the test, with a P-value of 0.6298, which is higher than the threshold of 0.05, the null hypothesis cannot be rejected, thus the **residuals are normally distributed**.

2. Homoscedasticity

The Breusch-Pagan test will be used to test for heteroscedasticity of errors in regression.

From the test, with a P-value of 0.7125, which is higher than the threshold of 0.05, the null hypothesis cannot be rejected, thus **homoscedasticity is assumed**.

Hence, the square root transformation of model can satisfy all four assumptions and will be statistically significant in this analysis.

References

SalesForce. (n.d.). *Top Student Retention Strategies for Higher Education*. Retrieved from SalesForce: https://www.salesforce.org/resources/article/student-retention-strategies/