University Admission regression Analysis Report

Instruction:

As undergraduate students near their college graduation, one looming question persists: Should I consider graduate school, and if so, what will improve my chances of admission? The transition to postgraduate studies, though promising, is riddled with uncertainties, especially regarding the admission process. This research aims to dissect this complex process by delving into a pivotal question: How do specific academic and personal metrics influence graduate school admission probabilities? The enigma of admissions often leaves students in a quandary. By pinpointing decisive metrics, we aim to offer students a roadmap to optimize their applications. This model, while assisting students, can also refine the admissions process for institutions.

This study will scrutinize the following criteria, GRE Scores (out of 340), TOEFL Scores (out of 120), University Rating (out of 5), Statement of Purpose (SOP), and Letter of Recommendation (LOR) Strength (out of 5), Undergraduate GPA (out of 10) and Research Experience (either 0 or 1). By mapping these metrics with admission results, the model aims to offer insights into each criterion's significance in the admissions process. This clarity can help students streamline efforts. Many students believe that CGPA and GRE scores are pivotal in influencing graduate admission outcomes. However, a growing contingent argues that research experience and the Statement of Purpose (SOP) play a more significant role. This research aims to explore not just these commonly cited factors, but also other potential variables that might impact admission decisions.

Exploratory analysis:

A thorough analysis of the predictor variables has been undertaken, omitting superfluous details such as the Serial Number to ensure clarity. The refined data has been meticulously organized in a structured table, simplifying its interpretation (Table 1). The primary predictors under review include the GRE score, TOEFL score, and CGPA. For the GRE Score: the minimum stands at 290, the median is 317, the average is 316.807500, and the standard deviation is 11.4736461. When examining the TOEFL Score, the minimum is 92, the median peaks at 107, the average is 107.41, and the standard deviation is 6.0695138. In the case of CGPA, the minimum is 6.8, the median is 8.61, the average settles at 8.598925, and the standard deviation is 0.5963171. A scatter plot created to visualize the relationship between the chance of admit and the primary predictors underscored the strongest linear relationship between CGPA and the chance of admit (Figure 1)

Among the predictors, only "Research" is categorized as a dummy variable, with the rest being quantitative. In this context, "Research" being a dummy variable implies that 0 denotes students without research experience, while 1 signifies students with such experience. An insightful examination using a side-by-side boxplot illuminates a pronounced trend: students with research experience consistently showcase a higher median probability of admission (Figure 2).

The distribution of chances for admission was evaluated using a histogram, which unveiled a skew away from a standard normal distribution. This observation suggests potential transformations might be requisite to align with specific statistical model prerequisites (Figure 3)

Model development:

Since the chance of admit partially follow a normal distribution, variable transformation is made on the chance of admit, which is used in three methods, namely Log-transformation, Square-Root-transformation, and Squared transformation. Comparing three different histograms respectively, it is found that the histogram of Squared transformation is more likely to follow the normal distribution. (Figure 4)

To check whether the predictors are highly correlated with each other, it is necessary to check the variance inflation factor. Delete the predictor with VIF greater than 5 from the model. Since the CGPA VIF value is 5.207403, such that need to delete the CGPA, and the get None Multicollinearity full model.

After checking the Multicollinearity, using stepAIC function to do model selection which is called an auto-reduced model, then calculate both adjusted R squared, AIC, and BIC values. (Table 2) Auto reduced model adjust R square is 0.7473175 which is bigger than None Multicollinearity full model 0.7472653, also auto reduced model AIC -996.2463, which is less than None Multicollinearity full model, such that choose an auto reduced model. Summary auto reduced model and then using ANOVA table to compare auto reduced model and none Multicollinearity full model, got Pr(>F) is equal to 0.3384, which is bigger than 0.05, the null hypothesis is that auto reduced model has similar performance as f None Multicollinearity full model and alternative hypothesis is that None Multicollinearity full model fit better than auto reduced model, such that fail to reject null hypothesis, which means auto reduced model fit better. After model selection, need to check 4 assumptions under linear regression.

Using residual plot (Figure 5) to check auto-reduced model linearity, constant variance, and independence. The linearity in the assumption means that the population conditional mean responses are correctly specified by (Xbeta) or equivalently the population errors have mean 0, in the residual plot, since observe equally spread residuals around a horizontal line without distinct patterns, such that satisfy the linearity assumption. The points in the residual plot randomly spread satisfy independent assumption, which means the population responses (equivalently errors) are uncorrelated with each other. Also follows constant variance assumption means population errors (equivalently response) have constant spread/variance around the conditional mean.

Under diagnostic plots (Figure 6), in the Residuals vs Fitted observe equally spread residuals around a horizontal line without distinct patterns, there is no indication that the linearity assumption is violated. In the Scale-Location, there is a straight line with randomly spread points, such that no indication of heteroscedasticity. In the Normal Q-Q, the residuals partially follow a straight, there is an indication that the normality assumption is violated. In the Residuals vs Leverage, most of the points inside of the Cook's distance, just fewer points outside of the Cook's distance, therefore the influential point exits.Since the histogram does not perfectly follow the normal distribution, and there is a violation of the normality in the model, need to do variable transformation. Choose the square of the response variable and then check the VIF value. In the transformed model, also need to delete the predictor whose VIF value is>5, so need to delete CGPA since VIF (CGPA) is 6.465994.

After checking the Multicollinearity, using stepAIC function to do model selection which is called model auto reduced1, Summary model auto reduced1, and then using ANOVA table to compare model auto reduced1 and full model1, got Pr(>F) is equal to 0.3773, which is bigger than 0.05, the null hypothesis is that model auto reduced1 has similar performance as full model1, and alternative hypothesis is that full model1 fit better than model auto reduced1, such that fail to reject null hypothesis, which means model auto reduced1 fit better. After model selection, need to check 4 assumptions under linear regression. Using residual plot (Figure 7) to check model auto reduced1 linearity, constant variance, and independence, since its transformation on a response variable (chance of admit), such that changed linear regression assumptions, so the transformed model did not follow linearity, independence, and constant variance. This means the population error does not have mean of 0, and population response or errors are correlated with each other.

Using diagnostic plots (Figure 8) to double check if it fails to follow part of the assumptions under linear regression. In the Residuals vs Fitted observe not equally spread residuals around a horizontal line with distinct patterns, there is an indication that the linearity assumption is violated. In the Scale-Location, there is a straight line with clustered points compared with none transformed model, such an indication of heteroscedasticity. In the Normal Q-Q, the residuals follow a straight, there is no indication that the normality assumption is violated, compared with none transformed model which is improved normality. In the Residuals vs Leverage, most of the points inside of the Cook’s distance, fewer points outside of the Cook’s distance, therefore the

influential point exits, compared with none transformed model, the influential points increased.

Conclusions:

Before the response variable (chance of admit) transformation, the multiple R-squared of the auto-reduced model is 0.7499, and the adjusted R-squared is 0.7473. The transformed model multiple R-squared is 0.9877, and the adjusted R-squared is 0.9876. Compared those two models multiple R-squared and adjusted R-squared, the transformed model is higher than none transformed model, since the R-squared represents the coefficient of multiple determination measures the proportion of the total variation in the dependent variable that is explained by the set of independent variables, such that will prefer using transformed model. This means the transformed model does an even better job of explaining what factors onto someone's chance of getting admitted. Since choosing the transformed model, Y-hat (Chance of Admit) = -0.7071173 + 0.0006491(GRE.Score) + 0.0008687(TOEFL.Score) +).0.0079714(University. Rating). which means one unit increase in GRE Score slightly increases the chance of admission. But one unit increasing in TOEFL Score or University Rating is more likely for someone to get admitted.

Even though the R-squared increase in the transformed model, there are also assumption violations that exist in the transformed model, such as failure to satisfy linearity, constant variance, and more influential points exists. Due to the limitations of the model, while the transformed model provides great insights, caution should be exercised in its general application. It must be acknowledged that in the real world, many factors affect college admissions, many of which may not be included in the model.

Appendix:

Table 1

Description: df [9 × 7]

	Min <dbl>	X1st.Qu. <dbl>	Median <dbl>	Mean <dbl>	X3rd.Qu. <dbl>	Max <dbl>	Std.Dev <dbl>
GRE.Score	290.0000	308.0000	317.0000	316.8075000	325.0000	340.0000	11.4736461
TOEFL.Score	92.0000	103.0000	107.0000	107.4100000	112.0000	120.0000	6.0695138
University.Rating	1.0000	2.0000	3.0000	3.0875000	4.0000	5.0000	1.1437281
SOP	1.0000	2.5000	3.5000	3.4000000	4.0000	5.0000	1.0068686
LOR	1.0000	3.0000	3.5000	3.4525000	4.0000	5.0000	0.8984775
CGPA	6.8000	8.1700	8.6100	8.5989250	9.0625	9.9200	0.5963171
Research	0.0000	0.0000	1.0000	0.5475000	1.0000	1.0000	0.4983620
Chance.of.Admit	0.3400	0.6400	0.7300	0.7243500	0.8300	0.9700	0.1426093
Transformed_Chance_of_Admit	0.1156	0.4096	0.5329	0.5449695	0.6889	0.9409	0.2009675

9 rows

Figure 1

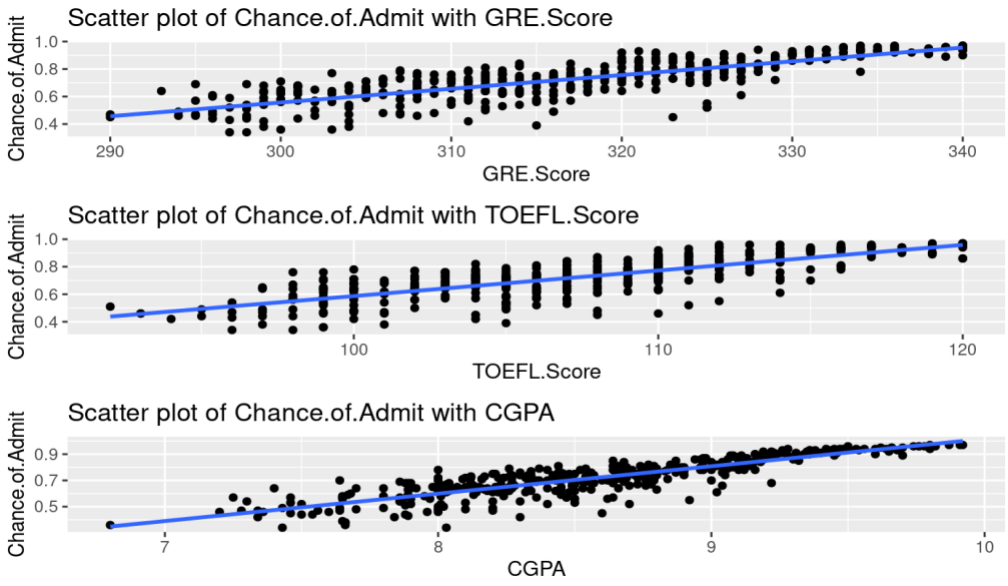


Figure 2

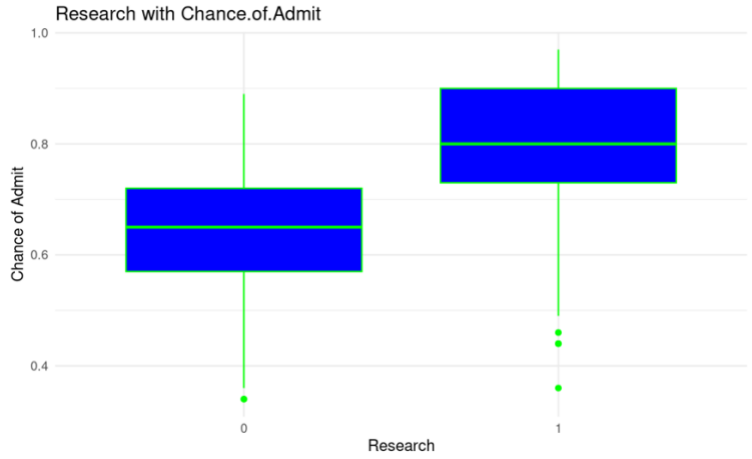


Figure 3

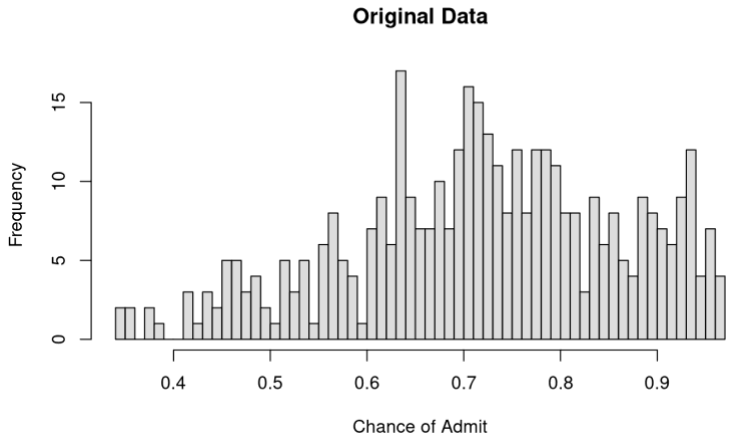


Figure 4

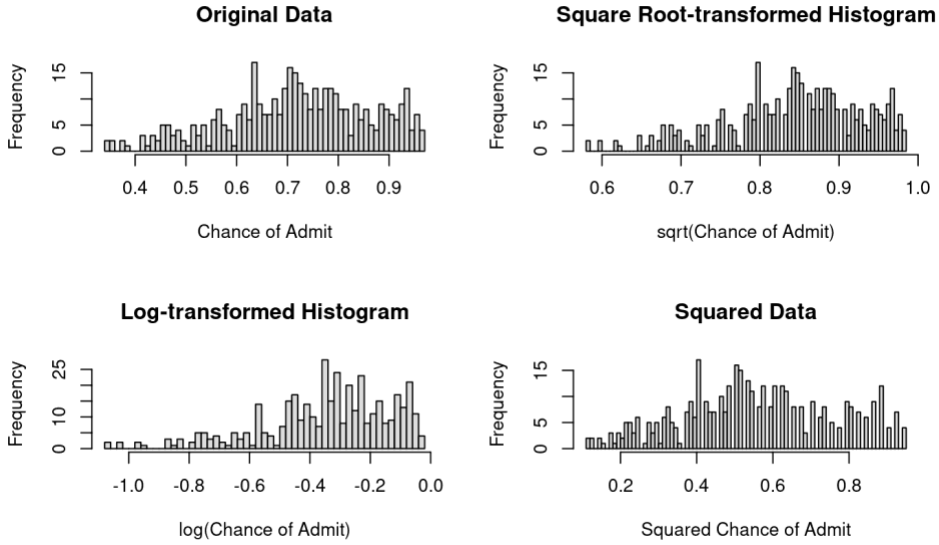


Table 2

Description: df [2 × 4]

Model <chr>	Adjusted_R_Squared <dbl>	AIC <dbl>	BIC <dbl>
Non_Multicollineaity_full_model	0.7472653	-965.1777	-937.2375
model_auto_reduced	0.7473175	-966.2463	-942.2975

2 rows

Figure 5

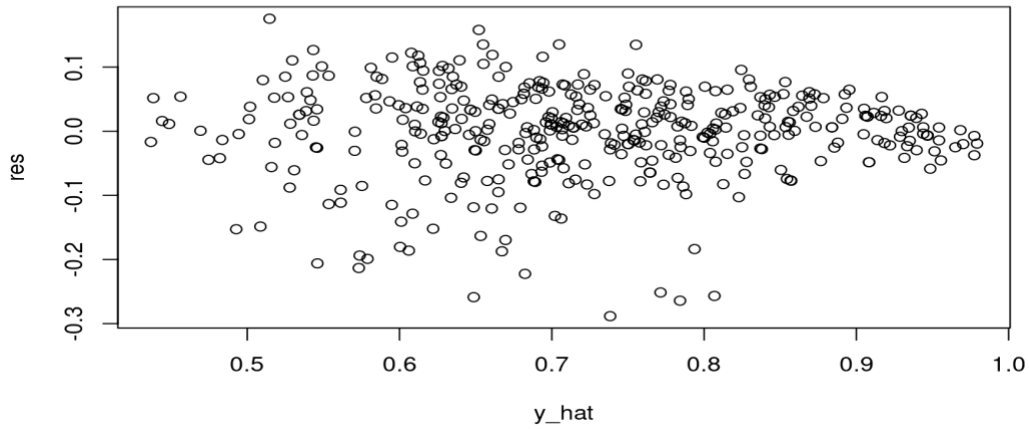


Figure 6

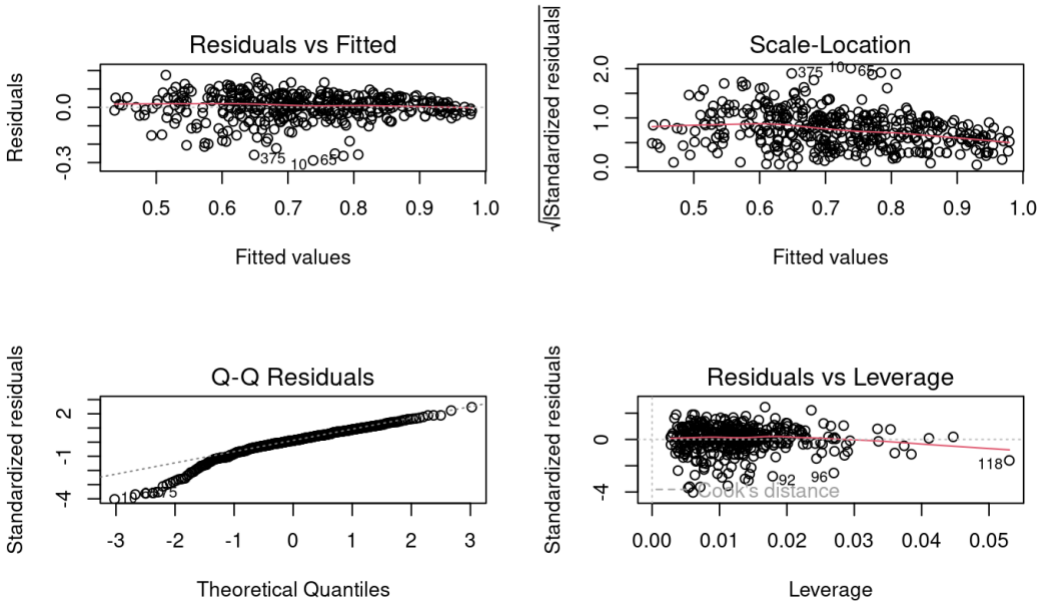


Figure 7

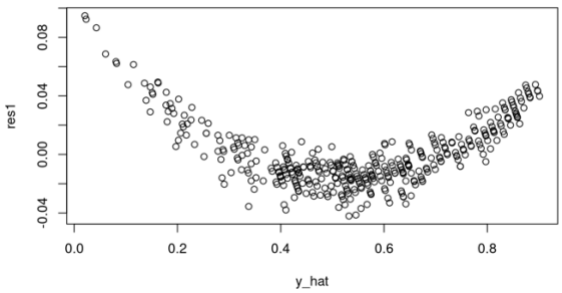
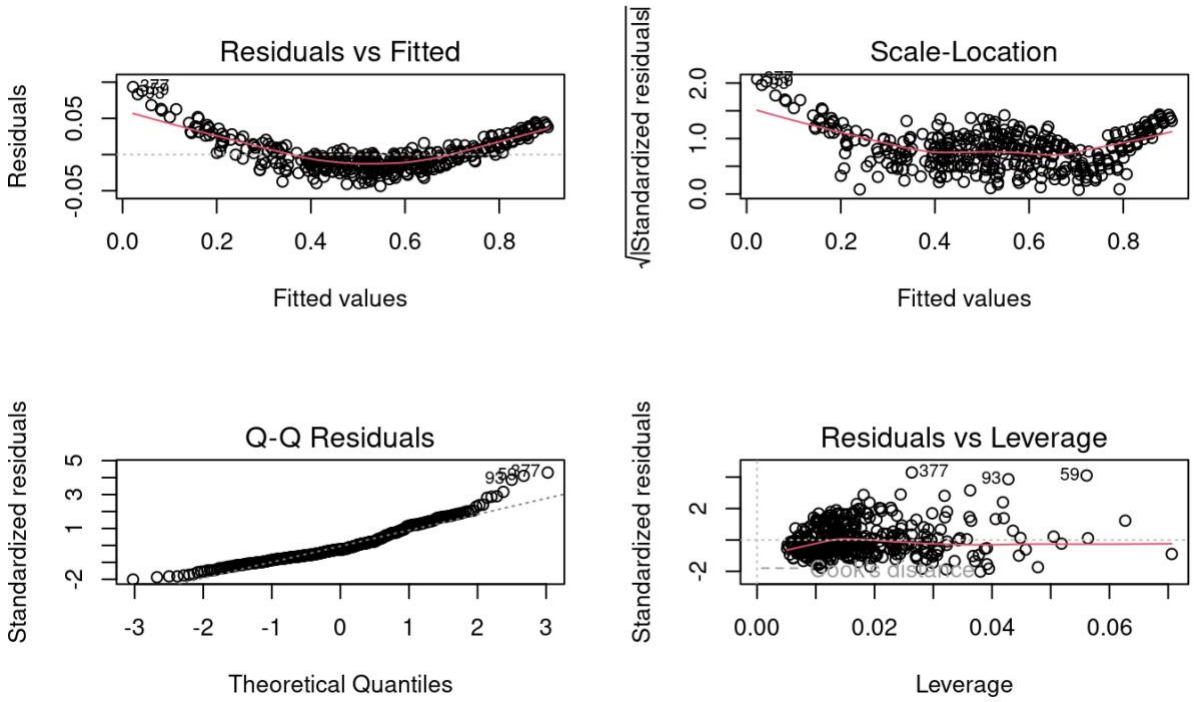


Figure 8



Citation:

kylek1. (2023, January 30). *University admission regression analysis*. Kaggle. https://www.kaggle.com/code/kylek1/university-admission-regression-analysis