Analysis of Relative Importance of Factors Affecting the Chance of Admission of Applying to Graduate Students

Wenhan Ju [1, †, *], Heran Wang [2, †] and Zi Ye [3, †]

[1] Dulwich International High School in Suzhou, Suzhou, Jiangsu, 215127, China

[2] International Department of the Experimental High School Attached to Beijing Normal University, Beijing, 100032, China

[3] Department of Mathematics and Applied Mathematics, Guangdong University of Technology, Guangzhou, Guangdong, 510520, China

* Corresponding author: Josephine.Ju23@stu.dulwich.org

†These authors contributed equally

Abstract. The conceptualization of statistical model: the initial characteristic of linear regression revealed by Karl Pearson lay the foundation for the exploration of Sir Francis Galton. The article discusses the concept of the linear regression model and R-square and focuses on the relationship between the chance of admission and GRE/TOFEL score by the simple linear regression between independent variables and dependent variables. The collected statistical data from Kaggle about 386 students ' scores in an American university is fitting the linear regression. This concentrates on the evaluation of the simple linear regression and the discovery of coverage for drawbacks or solutions for the promotion of the accuracy of the relationship. In the demonstrating experiments, collected TOFEL/ GRE scores of 375 graduate students in college indicated new observation that the existence of positively related between two items by R-square while testifying model possess drawbacks and limitations in a degree. The exploration of solutions to offset the impact of these errors and omissions will be concentrated in the following.

Keywords: simple linear regression, regression analysis, multiple linear regression, stepwise regression.

1. Introduction

1.1. Background

Every year many college students apply for graduate admission to different universities. Nowadays, due to the reform of the education system and the more comprehensive tests for judging students' level over the years, the chance of admission for college students to apply for graduate students is influenced by various examination results. In this case, many college students are confused about the various scores of examinations. They do not know how these grades specifically influence their chance of admission, such as which grades have the greatest impact on their chance of admission. Against this background, this paper wants to specifically elaborate on the impact of various examination scores on college students' chances of admission.

1.2. Related research

In Jianfei Leng, Xu Gao, and Jiaping Zhu's paper "Application of multiple linear regression statistical prediction model" [1], authors of the paper explain the significance of establishing a multiple linear regression equation, that is, to evaluate the relative importance of the influence of each independent variable on the dependent variable. The multiple linear regression's roles are very consistent with the research theme. This paper wants to study the relative importance of the impact of different test scores on the chance of admission of College Students. With the paper's support, multiple linear regression is a crucial research method in the paper.

The paper "Graduate Admission Chance Prediction Using Deep Neural Network" proposes a deep neural network (DNN) [2] to predict the chance of getting admitted to a university according to the student's portfolio and all the selection criteria in the paper are considered to predict the chance of admission. In addition, the paper "A Machine Learning Approach for Graduate Admission Prediction" [3] develops the approach to automatically predict the possibility of postgraduate admission to help graduates recognize and target the universities which are best suitable for their profile.

The paper "A Recommender System for Predicting Students' Admission to a Graduate Program using Machine Learning Algorithms" [4] provides a recommender system for early predicting university admission based on four Machine Learning algorithms namely Linear Regression, Decision Tree, Support Vector Regression, and Random Forest Regression. Similarly, in the paper "Deep Learning in diverse Computing and Network Applications, Student Admission Predictor using Deep Learning" [5], Student Admission Predictor (SAP) is developed in the proposed work using machine learning algorithms to help the students to find the chances of getting admission to a university. The paper "Prediction Probability of Getting an Admission into a University using Machine Learning" [6] develops a model that predicts the percentage of chances into the university accurately.

The most relevant paper is "development of a predicated model for college admission chances" [7], which builds a predictive model based on regressors like CGPA, TOEFL, GRE, and SOP into account. This has some similarities with the influencing factors studied in this article.

To sum up, predecessors have studied and published a series of models to help college students predict their chance of admission, so as to help college students to be admitted better. However, in this paper, the small influencing factor in the previous large prediction model, which is the various examination scores, is specially selected. This paper will analyze in more detail how the factor of examination scores influences the chance of admission of college students. This can also further help support the previous prediction models from predecessors.

1.3. Objection

Due to today's pressure on academic application caused by population expansion and strict system, more and more college students become more anxious and confused at the stage of applying for graduate students. The motivation of this paper is to start from the perspective of exam results, so that college students can better understand the relative importance of different exams, so as to effectively help them determine the priority of exam preparation and help them make more efficient application preparation.

2. Methods

2.1. Source of data

The dataset from Kaggle is used to predict Graduate Admissions from an Indian perspective. The data consists of several parameters e.g., GRE Score and TOEFL Score which are considered important during the application for Masters's Programs. It contains 400 rows with 8 columns, and each row provides each student's score or relevant level when applying for the programs. The current study is mainly to build an admission prediction model.

2.1.1. Dependent variable

In this dataset, Chance of Admission (denoted as CA) is taken as the response variable. It ranges from 0 to 1.

2.1.2. Independent variables

There are several factors that are likely to affect the admission, such as GRE Score (out of 340), TOEFL Scores (out of 120), University Rating (out of 5), Statement of Purpose, Letter of Recommendation Strength (out of 5), Undergraduate GPA (out of 10), Research Experience (either 0 or 1). These parameters are regarded as explanatory variables.

2.2. Models

2.2.1. Simple linear regression

The paper “The conceptualization of statistical model: the initial characteristic of linear regression” [8] discusses the concept of linear regression model and R-squared value, which is a value that reflects the accuracy of the analysis.

The study wants to use the simple linear regression model to do the regression analysis, where explanatory variables are V1 / V2 / V6 and the response variables are V8.

Then the simple linear regression model can be built as below formula (1)-(3):

$$Model_1: V8 = a_1 + b_1V_1 + \varepsilon_1. \tag{1}$$

$$Model_2: V8 = a_2 + b_2V_2 + \varepsilon_2. \tag{2}$$

$$Model_6: V8 = a_6 + b_6V_6 + \varepsilon_6. \tag{3}$$

The function lm() in R is mainly used to do simple linear regression and the function summary( ) to implement regression analysis by checking the coefficients estimated, residual standard error, $R^{2,}$ and the p-value.

2.2.2. Multiple linear regression

As a crucial analyzing method, many papers have specifically explained the process of using multiple linear regression. The paper “Selection of optimal multiple linear regression model” [9] explains how to select the most appropriate multiple linear regression model to achieve the best effect through the analysis of data characteristics. In addition, the paper “Predictive modeling method of multiple linear regression” [10] specifically explains the whole process of creating the multiple linear regression’s predicting model. It also includes the later model accuracy evaluation.

According to common sense and experience (also according to the experiment results from the simple linear regression model), the Chance of Admission is influenced by all the important parameters together, which therefore caused 2 problems to be solved. The first one is, how to build a model so that it can contain multiple variables instead of a single variable. And the other is, which parameters are of significance so that they can become variables contained in the model?

To deal with the first problem, multiple linear regression is chosen. As for the second problem, it is proposed that all the parameters from the dataset should be initially put together into a multiple linear regression model, and then apply the backward and stepwise regression method to exclude those less important parameters. By stepwise regression method, the significance of the multiple linear regression model is expected to get improved.

There is another problem since some variables like Research Experience (whether students have had research experience or not corresponds to the value 1 or 0) in the dataset are qualitative rather than quantitative. These qualitative variables should not be handled the same way as those quantitative variables. With the function level () in R, a value can be assigned to them so that these qualitative variables then can be seen as quantitative variables.

The multiple linear regression model that contains all the parameters can be built as below formula (4):

$$Model: V8 = a + b_1V_1 + b_2V_2 + b_3V_3 + b_4V_4 + b_5V_5 + b_6V_6 + b_7V_7 + \varepsilon. \tag{4}$$

2.2.3. Stepwise regression

After the simple linear regression and multiple linear regression, because the p-value of significance is still not satisfactory, by using the function step (), the stepwise regression is built.

According to the Akaike information criterion, GRE, TOEFL, LOR, CGPA, and research are selected as explanatory variables.

2.3. T-test and regression diagnosis

2.3.1. T-test

The T-test is used to test the data in order to have a more accurate regression analysis, where x_1 and x_2 are the two sets of data to be compared, s^2 is the combined standard error of data, and n_1 and n_2 are the different groups. The T-test formula is shown below.

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\left(s^2\left(\frac{1}{n} + \frac{2}{n}\right)\right)}}. \tag{5}$$

2.3.2. Regression diagnosis

In any meaningful linear model, not only the p-value obtained by the t-test of each parameter should be considered, but also the p-value obtained by the F-test of the overall model. Besides, the linear model is based on the following five assumptions (including Normal i.i.d. errors, Constant error variance, Absence of influential cases, Linear relationship between predictors and outcome variable, and Collinearity). If these five assumptions are not met, the prediction effect of the model is not acceptable enough even if it passes the t-test and F-test. Additionally, the model can also get further optimization through the study of these five indicators (like using some other nonparametric statistical methods). It can be easily visualized using R.

One of the five assumptions is, the residual should be a normal distribution, which is independent of the estimated value. If the residual error is related to the estimated value, it indicates the model has room to improve. It can plot the figure about Residual vs Fitted.

A normal QQ plot is also used to find whether the residual is normally distributed, which nearly does similar work to the Residual vs Fitted plot.

Another assumption is, it is assumed that the variance in the prediction model should be a fixed value. If the variance is not a fixed value, the reliability of this model will be greatly reduced. It is shown in the Scale-Location plot where the Y axis is squared standardized residuals and the X axis is fitted values.

The plot of Residuals vs Leverage is mainly to check whether there are extreme points.

2.4. Research framework

The whole research process first finds the data source of a graduate's chance of admission on Kaggle. After that, a simple linear regression is performed on the variables to do the analysis. After the first regression analysis, to solve some problems and improve the accuracy of the analysis, the paper respectively conducts the multiple linear regression and the stepwise regression. Eventually, according to the regression analysis above, the T-test and regression diagnosis are done, which reflects some characteristics of the conducted mathematical model.

3. Results

3.1. Simple linear regression

From the Table 1 below, $\widehat{b_1}$= 0.0099759, and $\widehat{a_1}$= -2.430842. It can also be found that the R-squared of Model_1 is 0.6442, which means the regression line explains only 64% of the variation. Fig. 1 shows the regression line of the estimate Model_1.

Table 1. The estimated coefficients, R-squared, F-statistic can be seen from Model$_1$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.4360842	0.1178141	-20.68	$<2e^{-16}$ ***
V1(GRE)	0.0099759	0.0003716	26.84	$<2e^{-16}$ ***
Residual standard error: 0.08517 on 398 degrees of freedom. Multiple R-squared: 0.6442, Adjusted R-squared: 0.6433. F-statistic: 720.6 on 1 and 398 DF. p-value: < $2.2e^{-16}$.				

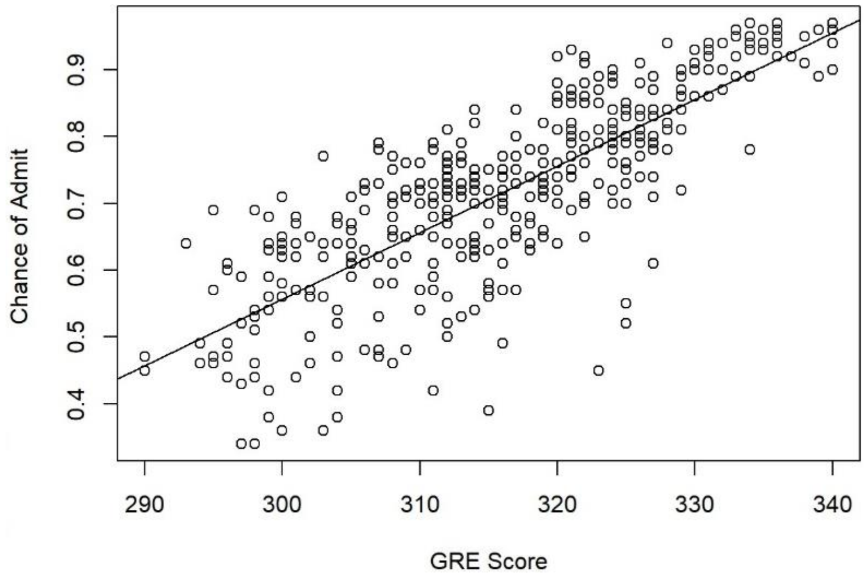

Figure 1. The scatter plot of V1 and V8 and regression line (Photo credit: Original)

From Table 2 below, $\widehat{b_2}$= 0.0185993, and $\widehat{a_2}$= -1.2734005 and the R-squared of Model$_2$ is 0.6266, which is even less than the R-squared of Model$_1$.

Table 2. The estimated coefficients, R-squared, and F-statistic can be seen from Model$_2$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.2734005	0.0774217	-16.45	$<2e^{-16}$ ***
V2(TOEFL)	0.0185993	0.0007197	25.84	$<2e^{-16}$ ***
Residual standard error: 0.08525 on 398 degrees of freedom. Multiple R-squared: 0.6266, Adjusted R-squared: 0.6257. F-statistic: 667.9 on 1 and 398 DF. p-value: < $2.2e^{-16}$.				

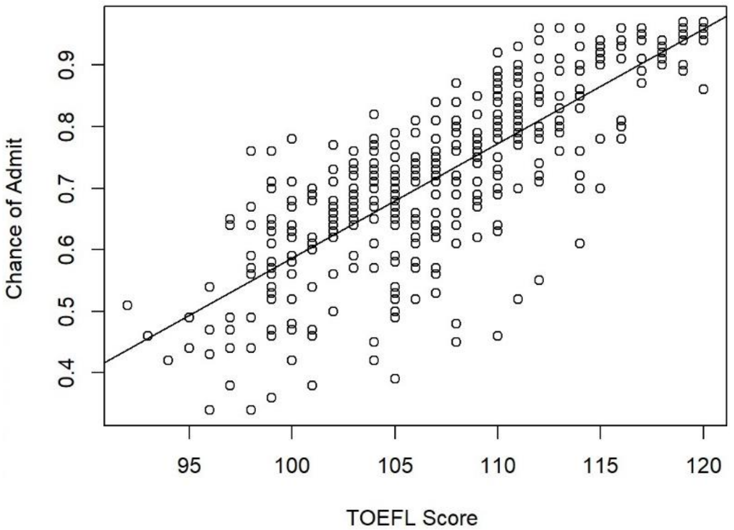

Figure 2. The scatter plot of V2 and V8 and regression line (Photo credit: Original)

From Table 3 below, $\widehat{b_6}$= 0.20885, and $\widehat{a_6}$= -1.07151. It can also be found that the R-squared of Model$_6$ is 0.7626, which is a relatively acceptable result. Fig. 3 shows the regression line of the estimate model_6. It can be easily found that the regression line of Model$_6$ fits the scatter plot best, compared with the other two regression lines above.

Table 3. The estimated coefficients, R-squared, and F-statistic can be seen from Model$_6$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.07151	0.05034	-21.29	$<2e^{-16}$ ***
V2(TOEFL)	0.20885	0.00584	35.76	$<2e^{-16}$ ***
Residual standard error: 0.06957 on 398 degrees of freedom. Multiple R-squared: 0.7626, Adjusted R-squared: 0.7620. F-statistic: 1279 on 1 and 398 DF. p-value: < $2.2e^{-16}$.				

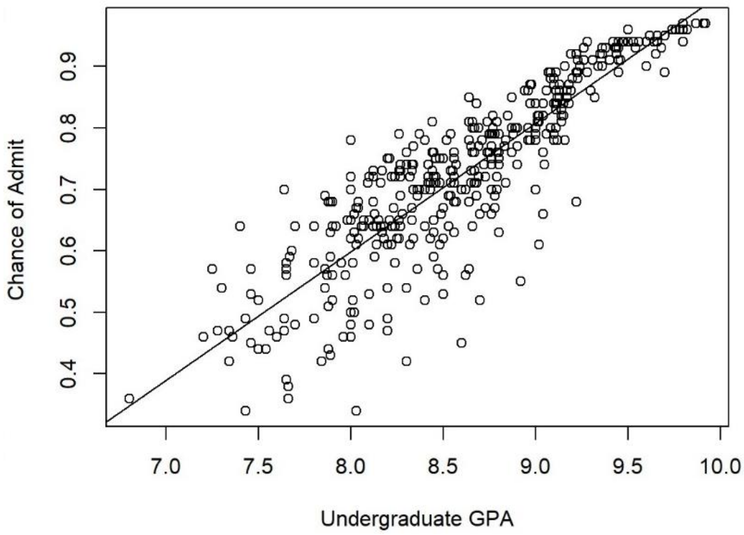

Figure 3. The scatter plot of V6 and V8 and regression line (Photo credit: Original)

3.2. Multiple linear regression

3.2.1 Multiple linear regression that contains all the parameters

From Table 4 below, $\widehat{b_1}$= 0.0017374, $\widehat{b_2}$= 0.0029196, $\widehat{b_3}$= 0.0057167, $\widehat{b_4}$= -0.0033052, $\widehat{b_5}$= 0.0223531, $\widehat{b_6}$= 0.1189395, $\widehat{b_7}$= 0.0245251 and $\hat{a}$= -1.2594325. The R-squared of the Model is 0.8035, which is a nearly satisfactory result.

Table 4. The estimated coefficients, R-squared, and F-statistic can be seen in Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.2594325	0.1247307	-10.097	$<2e^{-16}$ ***
V1(GRE)	0.0017374	0.0005979	2.906	0.00387 **
V2(TOEFL)	0.0029196	0.0010895	2.680	0.00768 **
V3(US)	0.0057167	0.0047704	1.198	0.23150 **
V4(SOP)	-0.0033052	0.0055616	-0.594	0.55267 **
V5(LOR)	0.0223531	0.0055415	4.034	$6.6e^{-5}$ ***
V6(CGPA)	0.1189395	0.0122194	9.734	$<2e^{-16}$ ***
V7(Research)	0.0245251	0.0079598	3.081	0.00221 **
Residual standard error: 0.06378 on 392 degrees of freedom. Multiple R-squared: 0.8035, Adjusted R-squared: 0.8000. F-statistic: 228.9 on 7 and 392 DF. p-value: < $2.2e^{-16}$.				

3.2.2. After stepwise regression

By using function step () in R to do the stepwise regression, V1, V2, V5, V6, and V7 are chosen to be those significant variables as Table 5. Based on these selected variables, the final model is built as below:

$$Model_{final}: V8 = a + b_1V_1 + b_2V_2 + b_5V_5 + b_6V_6 + b_7V_7 + \varepsilon \tag{6}$$

Table 5. The final result of stepwise regression

Step: AIC = -2196.38				
	DF	Sum of Sq	RSS	AIC
<none>			1.6008	-2196.4
V2(TOEFL)	1	0.3292	1.6338	-2190.2
V1(GRE)	1	0.3638	1.6372	-2189.4
V7(Research)	1	0.3912	1.6400	-2188.7
V5(LOR)	1	0.09133	1.6922	-2176.2
V6(CGPA)	1	0.43201	2.0328	-2102.8

From Table 6 below, $\widehat{b_1}$= 0.0017820, $\widehat{b_2}$= 0.0030320, $\widehat{b_5}$= 0.0227762, $\widehat{b_6}$= 0.1210042, $\widehat{b_7}$= 0.0245769 and $\hat{a}$= -1.2984636. In Table 6, it can also be found that the R-squared of $Model_{final}$ is 0.8027, which is relatively satisfactory, while the significance level gets improved by excluding V3 and V4 (It is easy to be found by comparing the two columns of Pr(>|t|) in Table 4 and Table 6).

Table 6. The estimated coefficients, R-squared, and F-statistic can be seen from $Model_{final}$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.2984636	0.1172905	-11.070	$<2e^{-16}$ ***
V1(GRE)	0.0017820	0.0005955	2.992	0.00294 **
V2(TOEFL)	0.0030320	0.0010651	2.847	0.00465 **
V5(LOR)	0.0227762	0.0048039	4.741	$2.97e^{-6}$ ***
V6(CGPA)	0.1210042	0.0117349	10.312	$<2e^{-16}$ ***
V7(Research)	0.0245769	0.0079203	3.103	0.00205 **
Residual standard error: 0.06374 on 394 degrees of freedom. Multiple R-squared: 0.8027, Adjusted R-squared: 0.8002. F-statistic: 320.6 on 5 and 394 DF. p-value: < $2.2e^{-16}$.				

Therefore, the final model is:

$$Model_{final}: V8 = -1.2984636 + 0.001782V_1 + 0.003032V_2 + 0.0227762V_5 + 0.12100426V_6 + 0.0245769V_7 \quad (7)$$

3.2.3. T-test and regression diagnosis

The T-test can also be used to check the stepwise regression result (seen in Table 7).

Table 7. The result of the t-test

t-test of coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.25943248	0.21327439	-5.9052	$7.640e^{-9}$ ***
V1(GRE)	0.00173741	0.00078846	2.2036	0.02814 *
V2(TOEFL)	0.00291958	0.00120519	2.4225	0.01587 *
V3(US)	0.00571666	0.00601035	0.9511	0.34212
V4(SOP)	-0.00330517	0.00740597	-0.4463	0.65564
V5(LOR)	0.02235313	0.00531793	4.2033	$3.260e^{-5}$ ***
V6(CGPA)	0.11893945	0.01501526	7.9212	$2.454e^{-14}$ ***
V7(Research)	0.02452511	0.01010870	2.4261	0.01571 *
Residual standard error: 0.06378 on 392 degrees of freedom. Multiple R-squared: 0.8035, Adjusted R-squared: 0.8000. F-statistic: 228.9 on 7 and 392 DF. p-value: < $2.2e^{-16}$.				

Some regression diagnoses are also made (seen in Fig. 4 and Fig. 5). Considering the plot of Residuals vs Fitted, the two figures below both show an obvious linear relationship, where the red line is nearly a straight line parallel to the X-axis, which means the residuals are not affected by the independent variables. As for Normal Q-Q, they both indicate the residual satisfies normal distribution roughly. In terms of Scale-Location, when the red line is approximately a straight line parallel to the x-axis, it indicates that the variance of the residual is homogeneous. However, in the two results, the variance of the residual decreases with the increase of the value of the fitting result variable to some degree. Residuals vs Leverage is used to identify outliers, that is, when these data are included or excluded from the analysis, the regression results will be significantly different. It can be found there is no outlier in the results. Besides, it is obvious that the leverage gets evidently improved in $Model_{final}$.

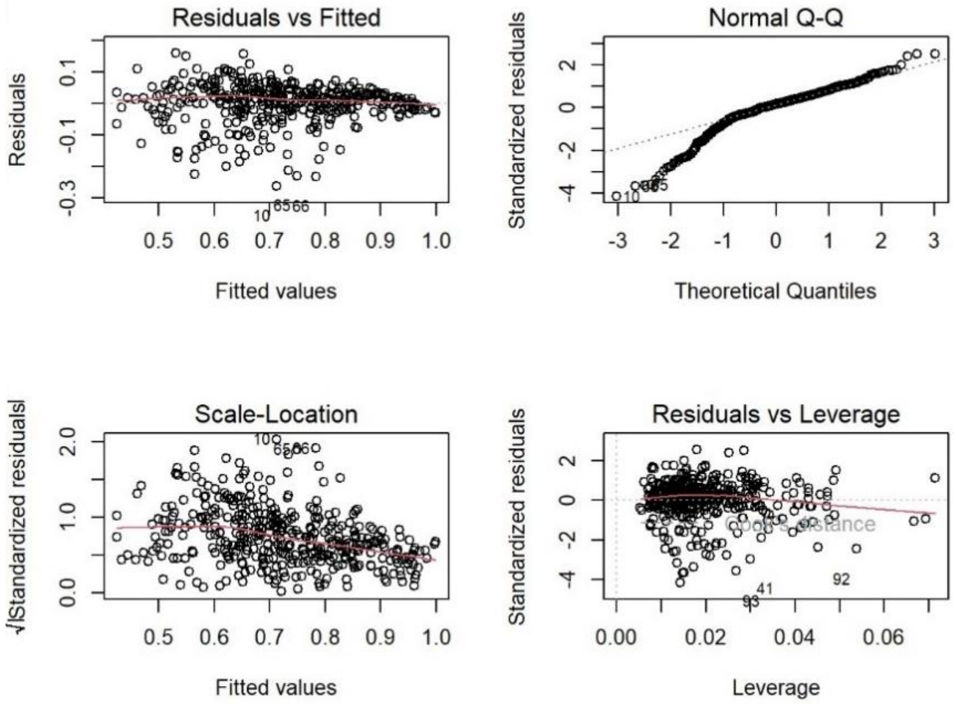

Figure 4. Regression diagnosis of Model (Photo credit: Original)

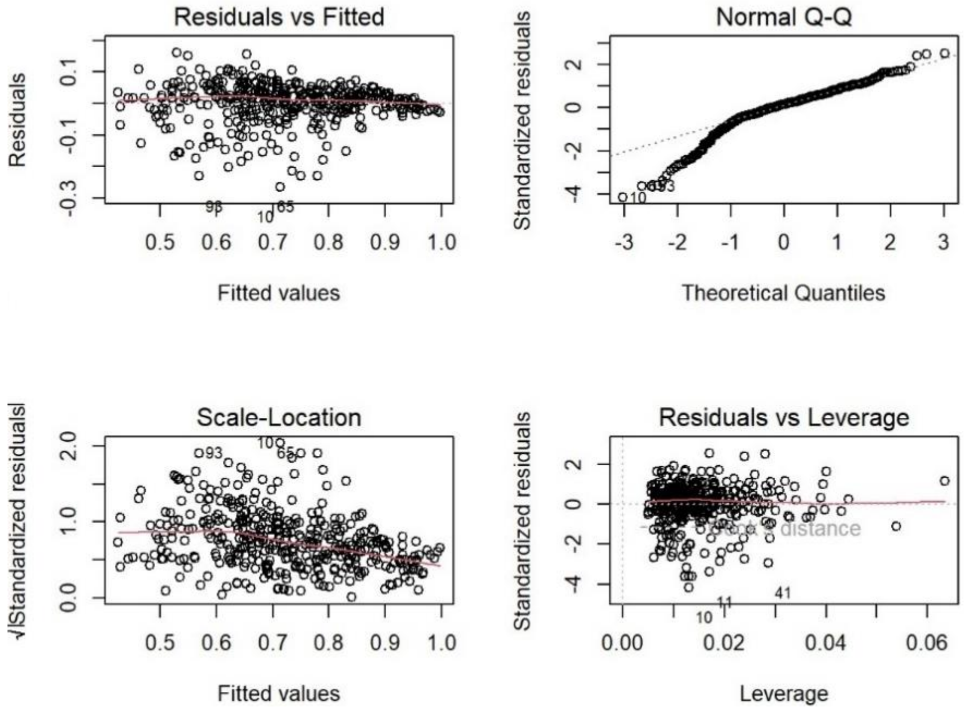

Figure 5. Regression diagnosis of Model$_{final}$ (Photo credit: Original)

4. Discussion

According to the analysis above, one of the methods used is simple linear regression. The simple linear regression model is the usage of an algebraic equation, because of the foundation of mathematics principles, which is easier than others to comprehend, such as the least-square error. Simultaneously, the calculation of the linear regression is based on the statistical parameter, and after the statistics, the linear regression is indicative of the general trend of data, which is clearer, more presentable, and more competitive than other models. At last, the most crucial reason for the wide spread of regression models is that the tools can be easily achieved by application on computers.

Nevertheless, the collinear problems are occurred with using simple linear regression, since there is the likelihood that the strong correlation for the model to easily cause a collinear equation, and the data testing contends that an increasing number of independent variables will generate instability and unreliability since more independent variables disturb results in degree in simple linear regression. Furthermore, the model possesses significant limitations in that the model only can be used for linearity, and it is not taken into account nonlinearity. After matching the simple linear graph, the data measurement is obtained by the R-square, which is beneficial to indicate the relationship between the evaluation of a dependent variable and the evolution of an independent variable can be represented. Furthermore, it also slightly assists the data fitting model due to the regression line on the graph.

However, there are some limitations to the R-square. First, it is not as accurate as the public would like, as it varies between 0 and 1, but there is no way to track the exact accuracy. Moreover, it doesn't even claim accuracy about the model whether it is correct or not, even if it shows a relationship. Meanwhile, in theory, it is believed that a higher R-value means a good model, but in reality, this is not always the case, and there are times when a good model will calculate a high R-value.

Therefore, R-square is a way to evaluate and predict the model essentially with some limitations, still depending on the testing context, and python is also a suitable language for predicting models. The enhancement of accuracy that using python and R-square together will be achieved when testing for the regression analysis. Therefore, there are still many details that require to improve in the regression model, and the promoted method will concentrate on the control variable of testing context, and different methods to measure models.

5. Conclusion

The data is processed and experimented with through simple linear regression, multiple linear regression, and stepwise regression methods.

Based on the analysis results above, the paper eventually gets some models (formulas) about how to estimate the chance of admission. According to the coefficient relationship, it can be seen that undergraduate GPA has the greatest impact on the chance of admission. Meanwhile, with similar coefficients, letters of recommendation and research experience also influence the chance of admission to some degree. Coefficients sharing the same order of magnitude, GRE and TOEFL scores have the least influence on the chance of admission, among all the factors that may be relative.

At the same time, on the basis of regression analysis and regression diagnosis, the three different methods used can be compared partly. Using simple linear regression is only applicable to the case of one single variable, which is difficult to be found in practical situations and data. Compared with it, multiple linear regression can deal with problems when there are multiple variables, but it is still not suitable to cope with more complicated circumstances (the fact can be seen from the final R-squared value, which is not high enough). In addition to that, stepwise regression is suggested to be used to select variables flexibly so that the confidence level and the level of each hypothesis test get improved.

This study provides a brief basis for the prediction of master's admission opportunities and finds some influencing factors of importance. Different linear regression methods are also compared respectively. As for further refinement of those basics and factors to build a more developed and realistic model, it is expected to improve the data process and methods (like nonlinear models) in later studies.

References

[1] Leng Jianfei, Gao Xu, Zhu Jiaping. Application of multiple linear regression statistical prediction model. Statistics and Decision Making. 2016 (07): 82 - 85.

[2] M. Omaer Faruq Goni, et al. "Graduate Admission Chance Prediction Using Deep Neural Network." 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE) IEEE (2020): 259 - 262.

[3] Alghamdi, A., et al. "A Machine Learning Approach for Graduate Admission Prediction." IVSP '20: 2020 2nd International Conference on Image, Video and Signal Processing, (2020):155 - 158.

[4] El Guabassi, I., et al. A Recommender System for Predicting Students' Admission to a Graduate Program using Machine Learning Algorithms. International Association of Online Engineering, 2021.

[5] Nandal, P., "Deep Learning in diverse Computing and Network Applications Student Admission Predictor using Deep Learning." Proceedings of the International Conference on Innovative Computing & Communications (ICICC), 2020.

[6] Sivasangari, A., et al. "Prediction Probability of Getting an Admission into a University using Machine Learning." 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (2021):1706 - 1709.

[7] Jianqing, Gao. "DEVELOPMENT OF A PREDICTED MODEL FOR COLLEGE ADMISSION CHANCES." European Journal of Education and Applied Psychology (2021): 3 - 8.

[8] Klotz, J. H., "Updating simple linear regression." Statistica Sinica 5.1 (1995): 399 - 403.

[9] Jin Hao, Suying Gao. Selection of optimal multiple linear regression Model. Journal of Hebei University of Technology. 2002 (05): 10 - 14.

[10] Wang Huiwen, Jie Meng. Predictive Modeling method of multiple Linear regression. Journal of Beijing University of Aeronautics and Astronsutics. 2007 (04): 500 - 504.