Research on the Admission of Graduate Students Based on Multiple Regression Model

Bo Liu[1a]

[1]*Department of applied mathematics, The Hong Kong Polytechnic University, Hong Kong SAT, China*
bo720627.liu@connect.polyu.hk

Abstract
In light of the unprecedented high unemployment rate due to the latest spread and development of coronavirus with only a few available internship and full-time job opportunities for college graduates, this research based on large sample post-graduate admission statistics dataset and prospective post-graduates applicants' perspective mainly adopted sequential variable selection methods to generate finite number of ordinary least square multiple linear regression models, utilized best subset model selection framework in terms of comparing corresponding types of evaluation metric and criterion, incorporated various model diagnostic plots to validate important regression model assumptions and constructed an cube-transformation technique to correct the heteroskedasticity presented in the model. This research concluded that selected optimal ordinary least squared multiple linear regression models under cubed transformation could achieve satisfactory performance of forecasting the post-graduate admission chance without sacrificing the goodness of fit. This research helps post-graduate applicant to understand importance of each admission covariate contributed to overall admission chance.

Keywords: *Graduate admission; Multiple linear regression; Model selection; Model diagnostic*

1.INTRODUCTION

1.1. Research Background and Motivation

According to the statistical table from official news and report, the expected number of college senior graduates aimed to apply for post-graduate school to continue pursuing a higher degree from all over the world in each year grows exponentially. Besides, in light of the unprecedented higher unemployment rate caused by latest spread and outbreak of the coronavirus, the number of potential job applicants demanding and competing for limited number of similar type of available employment opportunities with relatively attractive working compensations within the past two years has increased drastically and due to the worker of the human resource recruitment department in the company tends to want to hire more employees with holding a higher academic degree or having professional technical skills in a specialized field, earning an post-graduate degree certificates become more and more important or even necessary before the job application or interview for college graduates holding an bachelor degree. Based on these above-mentioned facts, joining in an scientific research designated to theoretically explore for significance of several kind of post-graduate school admission determinants throughout using the feasible ordinary least squared multiple linear regression models to reasonably analyze and objectively predict the post-graduate school admission chance becomes more and more valuable and meaningful as these selected optimal theoretical constructed multiple linear regression model can practically enable potential post-graduate applicant to simulate the estimated admission chance in order to help them to have a better idea to understand the importance post-graduate admission factors contributed to the overall acceptance probability.

1.2. Literature Review

Chakrabarty N., Chowdhury S., & Rana S., (2020) have adopted the gradient boosting regression model to analyze a student's academic accomplishments along with university rankings to calculate the chance of getting admission in that university as output and attained a coefficient of determination of 0.84 as a higher statistical result than the performance of any other graduate admission chance prediction model [1]. Acharya M. S., Armaan A., & Antony A. S., (2019,

G. Vilas Bhau et al. (Eds.): MSEA 2022, ACSR 101, pp. 681–689, 2023.
https://doi.org/10.2991/978-94-6463-042-8_98

February) have adopted a machine learning-based approach in terms of incorporating linear regression, support vector regression, decision trees and random forest as four different possible models into this research and also utilized model selection techniques to determine the optimal model among these based on model performance to comparatively analyze the prediction of graduate admission to conclude whether universities of these post-graduate school applicants' choices are reasonable ones [2]. Chari D., & Potvin G., (2019) adopted the survey question to collect and compare the responses for ranking the admission criteria according to its own importance and utilized the mean and standard error to perform the data analysis for collected sample responses based on post-graduate school applicants' perspective to conclude that a seemingly "over-emphasized" admission determinants might be less advantageous to applicants' admissions-related decision making and reduce their chance of success [3]. Based on exploratory data analysis techniques, Sujay S., (2020) utilized machine-learning algorithm to perform cross-validation in terms of conducting model selection and fitting, and constructed a well-performed linear regression prediction model based on python output to analyze the admission chance of a post-graduate school applicant and conclude that the prediction of the admission chance could be successfully implemented with the help of supervised machine-learning and exploratory data analysis[4]. Iman, A., and Tian, X. (2021) have adopted Naïve Bayes, Multilayer Perceptron, Logistic Regression, Random Forest, REP Tree, Random Tree, and J48 as seven different types of machine learning classification models and utilized evaluation metrics to compare these model performance to select the optimal model to predict the graduate admission outcome of candidates with analyzing known parameters contained in a dataset of 400 applicant records and conclude Naïve Bayes to be the most accurate model available for not only helping prospective applicants to narrow down the search for selecting the right post-graduate school list for this type of dataset but also providing graduate admissions committees with the support of filtering large pools of applications to enable them to view and understand the their past admission decision patterns [5]. El Guabassi, I., and Bousalem, Z. et al. (2021) have adopted four different types of machine learning algorithms in terms of linear regression, support vector regression, decision tree regression and random forest regression and also utilized various evaluation metrics to compare the performance for each of these algorithms to construct an optimal predictive model in order to analyze the most significant parameters affecting the admission chance of candidates. This research not only demonstrated random forest regression as the most suitable machine learning algorithm available for forecasting post-graduate school admission but also showed that cumulative grade point average tends to be the most significant parameter influencing the admission chance of candidates [6]. Bag, A. (2020) has adopted three different types of multiple regression models in terms of linear regression, decision tree, and random forest and utilized evaluation metrics in terms of coefficient of determination and mean squared error used to measure the performance of each of these models to analyze and forecast admission chances of post-graduate school applicants. This research concluded linear regression to be the ideal model among these three with low MSE and high R-squared score to predict the admission chance of candidates [7]. T Goni, M. O. F., and Matin, A. et al. (2020, December) have adopted both deep neural network and other existing methods and utilized various performance metrics in terms of mean square error, root mean square error, mean absolute error, and R-squared score used to determine the optimal method to analyze and predict the chance of admission by taking into account all of these selection criteria. This research concludes deep neural network performed better than any other existing method with attaining R-squared Score of 0.8538 and MSE of 0.0031 [8]. Khan M. A., Dixit M., & Dixit A., (2020, April) have adopted machine learning approach in terms of linear regression, support vector regression, ridge regression, Bayesian ridge regression, artificial neural network, random forest regression, Ada Boost regression, K-nearest neighbors regression and decision tree regression under the first category of regression algorithm combined with artificial neural network, logistic regression, Naïve Bayes classifier, support vector machines, perceptron, K-nearest neighbors classifier, random forest classifier and decision tree classifier under another category of classification algorithm to analyze the most significant post-graduate admission determinant and construct a new approach to forecast the admission chance of candidates. This research shows artificial neural network not only becomes the optimal model to fit into this dataset with attaining the R-Squared Score of 0.890120 under the regression algorithm but also achieves satisfactory performance with attaining accuracy and F score of 0.953846 and 0.911765 respectively [9]. Aljasmi S. et al. (2020) have adopted four different machine learning models in terms of multiple linear regression, K-nearest neighbor, random forest and multi-layer perceptron to not only forecast the admission chance of a master program for candidates but also enable applicants to understand in advance whether they are likely to be admitted into a master's program. This research concludes multi-layer perceptron to be the best model used to predict the admission chance of a master program for candidates [10].

1.3. Paper Organization

The organizational framework of this paper is as follows, the first part is the introduction, which mainly includes the research background and motivation and

literature review; the second part is the method, including data description and construction of multiple linear regression models, the third part is the analysis of empirical results; the last part is the conclusion.

2.MANUSCRIPT PREPARATION

2.1. Benchmark Model

This research paper might first consider constructing an ordinary least squared benchmark model containing full parameters.

$$Y= \beta_0 +\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\beta_4 X_4+ \beta_5 X_5 + \beta_6 X_6+ \beta_7 X_7+ \varepsilon \quad (1)$$

The dataset contains several parameters:

Table 1: Parameters and Description

Number	Variable symbols	Explanation
1	X_1	GRE Scores (out of 340)
2	X_2	TOEFL Scores (out of 120)
3	X_3	University Rating (out of 5)
4	X_4	Statement of Purpose
5	X_5	Letter of Recommendation Strength (out of 5)
6	X_6	Undergraduate GPA (out of 10)
7	X_7	Research Experience (either 0 or 1)
8	Y	Chance of Admit (ranging from 0 to 1)
9	β_0	the intercept of this full parameter multiple linear regression model
10	β_i, i=1,2,3,4,5,6,7	corresponding coefficients for this full parameter multiple linear regression model.

Every of each single column in the middle of dataset between the first and last one is an independent variable to characterize the range of GRE score and TOEFL score for the graduate school applicants, indicate the university rating, the quality of their statements of purpose and the strength of letters of recommendation, and describe the their overall undergraduate academic performance in terms of GPA indicators as well as whether applicants have research work experience of relevant academic field in their application.

The last column is the response variable, which is a continuous value between 0 and 1, indicating the chances of getting admission.

Table 2: Descriptive Statistics of These Seven Covariates

Variable symbols	Min	Mean	Median	Max	Skewness	Kurtosis
X1	290	316.472	317	340	-0.03972223	2.28405
X2	92	107.192	107	120	0.09531393	2.341293
X3	1	3.114	3	5	0.09002387	2.186022
X4	1	3.374	3.5	5	-0.2282849	2.289344
X5	1	3.484	3.5	5	-0.1448541	2.249712
X6	6.8	8.57644	8.56	9.92	-0.02653261	2.432343
X7	0	0.56	1	1	-0.2417469	1.058442
Y	Min	Mean	Median	Max	Skewness	Kurtosis

2.2. Model Selection

This research is intended to generate 2^7=128 possible subset ordinary least squared multiple linear regression models by adopting sequential variable selection method with sequentially adding or dropping one independent variable at a time to select the best model among all of these 128 possible subset candidate models based on various model selection statistics and criterion listed down below.

Table 3: Model Evaluation Metrics with its Corresponding Selection Criterion

Number	Model Selection Statistics	Corresponding Model Selection Criterion
1	coefficient of determination	higher the better
2	adjusted coefficient of determination	higher the better
3	AIC (Akaike Information Criterion)	lower the better
4	BIC (Bayesian Information Criterion)	lower the better
5	MSE (mean square error)	lower the better
6	Mallow Cp	should be close to the number of predictors in model
7	MAPE (Mean Absolute Percentage Error)	lower the better
8	Min_Max Accuracy	higher the better

Best Subset Selection Method: Coefficient of Determination R^2. R^2, coefficient of determination, describes how much proportion of total variation of probability of getting post-graduate schools' admissions for college graduates can be well-explained by variation of range of GRE scores, range of TOEFL scores, levels of university rating, quality of statements of purpose, strength of letters of Recommendation Strength, overall undergraduate GPA and availability of research experience as all of these seven covariates incorporated in the regression model.

Since this research is aimed to select a model with the highest coefficient of determination under the Criterion of R-squared statistics and the property of coefficient of determination indicates that the value of coefficient of determination is proportional to the number of independent variables or covariates, we could form a conclusion that the full regression model

containing all parameters will always have the highest value of coefficient of determination.

This research has also validated theoretically hypothesized full parameter ordinary least squared benchmark model.

$$Y = -1.275725 + 0.001859X_1 + 0.002778X_2 + 0.005941X_3 + 0.001586X_4 + 0.016859X_5 + 0.118385X_6 + 0.024307X_7 \quad (2)$$

By calculating the R-squared statistic for each model using the statistical computing software R^2 and comparing the statistical results with each other, the highest coefficient of determination of 0.8219007 was indeed obtained among all these 128 possible subset multiple linear regression models. Since coefficient of determination R^2 provides a better goodness of fit for multiple linear regression models with sacrificing the model prediction accuracy in terms of characterizing proportion of the total variation explained by regression model, this research would also like to implement the adjusted squared statistics to perform model selection to offset some of the increase in value of R^2 due to the increased number of independent covariates in the model.

Since R^2 always increases with the number of independent variables or covariates, the adjusted R^2 is defined to fix this issue: Adjusted R-squared, Adjusted coefficient of determination, describes how much proportion of total variation of probability of getting post-graduates' admissions for college graduates can be explained by variation of only a relevant subset of these seven independent variables. Since this research is aimed to select a model with the highest adjusted coefficient of determination under the Criterion of adjusted R^2 statistics and the property of adjusted coefficient of determination indicates that the value of adjusted coefficient of determination is inversely proportional to the number of independent variables or covariates, we could form a conclusion that adjusted coefficient of determination can offset some of the increase in value of coefficient of determination and strike important balance goodness of fit for all subset models and model prediction accuracy.

This research has validated ordinary least squared multiple linear regression model.

$$Y = -1.280014 + 0.001853X_1 + 0.002807X_2 + 0.006428X_3 + 0.017287X_5 + 0.118999X_6 + 0.024354X_7 \quad (3)$$

By calculating the adjusted R-squared statistic for each model using the statistical computing software R and comparing the statistical results with each other, the highest value of the adjusted coefficient of determination of 0.8196889 was indeed obtained among all the 128 possible subset multiple linear regression models.

Since the summary of coefficient of determination R^2 (0.8219007) and adjusted coefficient of determination R^2 (0.8196889) indicate the ordinary least squared benchmark full-parameters multiple linear regression model or alternative ordinary least squared multiple linear regression model with an independent variable of quality of the statement of purpose eliminated would be the best subset model among all of these 128 possible multiple linear regression models in the middle procedure of the research, this research is interested in digging deeper to analyze, search for and validate the true optimal subset model based on various other relevant model selection statistics and criterion.

The Akaike information criterion (AIC) is an estimator of out-of-sample prediction error and thereby relative quality of statistical models for a given set of data.

$$AIC = 2p - 2\log L \quad (4)$$

L is the maximum value of the likelihood function for the model. The model with the lowest AIC is preferred. The Akaike information criterion (AIC) is a criterion for model selection among infinite and relatively high-dimensional set of models. AIC measures the quality of each model, relative to each of the other models among all of these 128 possible models.

Since this research is aimed to select a model with the lowest values of AIC under the Criterion of the Akaike information criterion, we have found that the ordinary least squared multiple linear regression model.

$$Y = -1.280014 + 0.001853X_1 + 0.002807X_2 + 0.006428X_3 + 0.017287X_5 + 0.118999X_6 + 0.024354X_7 \quad (5)$$

Of all the 128 possible models based on the R output, it has the lowest AIC value of -1386.35. Bayesian information criterion (BIC) or Schwarz information criterion (also SIC, SBC, SBIC) is a criterion for model selection among a finite set of models.

$$BIC = p \times \log n - 2\log L \quad (6)$$

The model with the lowest BIC is preferred. Since this research aimed to select a model with the lowest values of BIC under the Criterion of the Bayesian information criterion, we have found that ordinary least squared multiple linear regression models.

$$Y = -1.335702 + 0.001889X_1 + 0.003017X_2 + 0.019320X_5 + 0.122980X_6 + 0.025165X_7 \quad (7)$$

Out of all these 128 possible models based on the R output, the lowest value of the BIC has been reached, which is -1355.80.

Mean Square Error (MSE) is used to describe and characterize the measurement of average squared

difference between actual and estimated values of probability of getting post-graduate school admissions.

Since this research aimed to select a model with the lowest values of Mean Squared Error under the Criterion of the Mean Squared Error, we have found and concluded that theoretically hypothesized ordinary least square benchmark full-parameters multiple linear regression model.

$$Y= -1.275725+ 0.001859X1+ 0.002778X2 + 0.005941\ X3+ 0.001586\ X4 +0.016859\ X5 +0.118385X6+0.024307X7 \quad (8)$$

Of all the 128 models based on the R^2 output, the mean squared error was the lowest at 0.003540751.

Mallows' Cp can help us to identify and select the optimal model among of those possible different multiple regression models. Mallows' Cp Criterion is used to compare the full benchmark model to each of these 128 possible subset models to helps us strike an important balance with the number of predictors in the model because a model contained too many covariates can be relatively imprecise while a model contained only a few independent variables can produce biased estimates. Since this research aimed to select a model with a Mallows' Cp value close to the number of predictors plus the constant to indicate that the model produces relatively precise and unbiased estimates under the Criterion of the Mallows' Cp, this article conclude that model 128 produces a Mallows' Cp value of 8 exactly equal to the number of predictors plus the constant contained in model 128 itself among all of these 128 possible models based on R^2.

Since this research aimed to select a model with the lowest values of Mean Absolute Squared Error under the Criterion of the Mean Absolute Squared Error, we have found and concluded that ordinary least squared multiple linear regression model.

$$Y=-1.280014 + 0.001853\ X_1+ 0.002807\ X_2 + 0.006428\ X_3+0.017287X_5+0.118999X_6 +0.024354X_7 \quad (9)$$

Across all 128 models based on Excel output, the minimum mean absolute squared error was 0.068512498. The results show that the average absolute percentage error between the model predicted graduate admission probability and the actual graduate admission probability is 6.8514482%.

Based on above experimental observations and statistical data analysis, this research has narrowed down the optimal candidate ordinary least R^2 multiple linear regression model to be either.

$$Y_1= -1.275725+ 0.001859X_1+ 0.002778X_2 + 0.005941\ X_3+ 0.001586\ X_4 +0.016859\ X_5 + 0.118385\ X_6 + 0.024307\ X_7 \quad (10)$$

$$Y_2=-1.280014 + 0.001853\ X_1+ 0.002807\ X_2 + 0.006428\ X_3+ 0.017287\ X_5 + 0.118999\ X_6 + 0.024354X_7 \quad (11)$$

3.EMPIRCAL RESULTS

This research would like to draw a variety of different kind of diagnostic plots to verify and check the validity of above selected two optimal models' assumptions in terms of linearity, normality, homoscedasticity (equal variance), and existence of outliers or influential points.

3.1. Residual VS Fitted Plot

The Residual vs Fitted plot displays whether residuals of outcome variables exhibit non-linear patterns. An implicitly or potentially existed non-linear pattern between independent covariates and a response variable could be characterized in this plot if the model itself hasn't been able to capture the non-linear relationship. If this model has equally distributed residuals around a horizontal line without distinct patterns, then this flat trend yields none of kind of non-linearity relationships to be incorporated into the model.

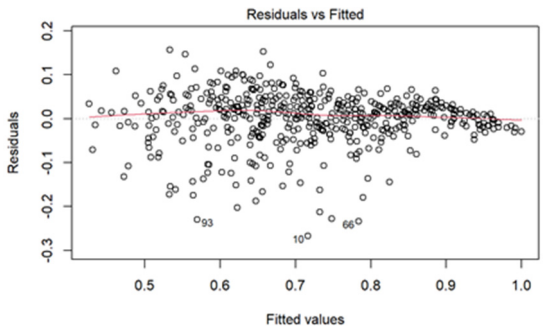

Figure 1. Residual versus Fitted Model I Performance Evaluation Plot

The residuals of model I exhibit a linear pattern since no unique pattern can be seen in the above plot.

3.2. Normal Q-Q Plot

The residuals from model I follows a straight line well to indicate that residuals from model I are almost normally distributed.

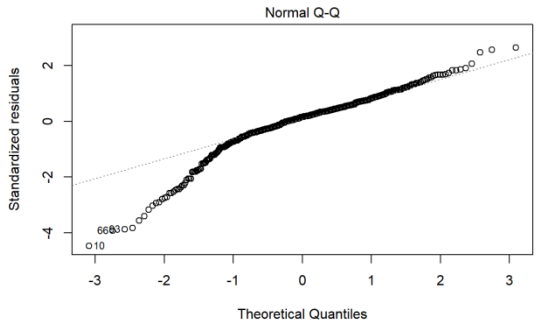

Figure 2. Normal Q-Q Model I Performance Evaluation Plot

3.3. Scale-Location Plot

This Scale-Location Plot is also referred to as Spread-Location plot. This plot is used to determine whether residuals of outcome variable are equally distributed with respect to the values within the ranges of these several covariates in order to validate whether this particular multiple linear regression model satisfies the homoskedasticity assumption. This paper relies on a horizontal line with evenly distributed points as a criterion to judge whether our chosen linear regression model violates the equal variance assumption. The residuals of our first selected optimal multiple linear regression model begin to spread denser gradually along the x-axis as it passes around 0.6. As the residuals become denser, the smooth line begins to drop, with relatively small angles of steepness in the above image. Meanwhile, the assumption of homoskedasticity is unlikely to be satisfied for this particular optimal model since the red line viewed to be roughly unparalleled with respect to fitted values axis across the plot with unequal distribution of residuals at all fitted values.

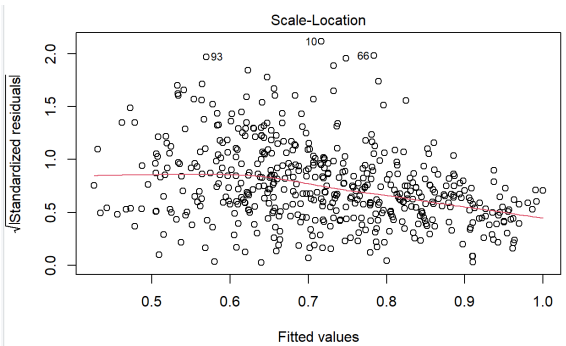

Figure 3 Scale and Location Model I Performance Evaluation Plot

3.4. Residual VS Leverage Plot

This plot was designated to identify influential points from this whole dataset studying about graduate admission statistics (i.e., subjects) if any. On the one hand, Of course, this research cannot simply make an objective claim to assume every of each outlier inside this above plot to be influential in linear regression analysis because even though the selected dataset might contain extreme values, they might not have strong influence when it comes to determine a regression line. With that being stated, the resulting statistical outcomes wouldn't deviate too much regardless of whether this research either include or exclude them from analysis as they follow the trend mostly. On the other hand, some scenarios corresponded by an subset of these dots could make a huge difference even if they seemingly appeared to be within a reasonable range of the values because these data points could potentially become extreme circumstances against a regression line and can modify the results drastically if they were chosen to be excluded from analysis as they don't get along with the trend most of the times.

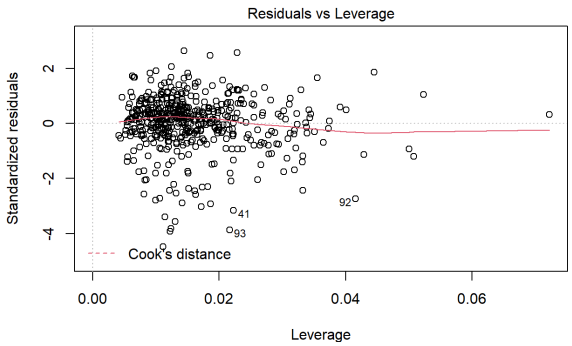

Figure 4. Residual versus Leverage Model I Performance Evaluation Plot

Unlike any other model diagnostic plots, patterns are not that crucial in the Residual-Leverage plot this time. This research primarily looked up to search for outlying values located either at the upper right corner or at the lower right corner but outside of a dashed Cook's distance indication line enclosed in the Residual-Leverage Plot to identify those spots as the places where scenarios can be influential against a regression line. When these dotted points fall outside of the Cook's distance with having higher Cook's distance scores, the corresponding scenarios are influential to the regression results. The regression results will be modified significantly if those cases were chosen to be excluded from analysis.

The Residual-Leverage Diagnostic Plot above corresponds to the first selected optimal multiple linear regression model without containing any influential observation beyond the Cook's distance lines because we can hardly observe a dashed Cook's distance indication lines of which all observations fall inside in the plot above.

Similarly, this paper will use the same diagnostic method for the second model, and the residual sequence diagram of the diagnosis is shown in Figure 5.

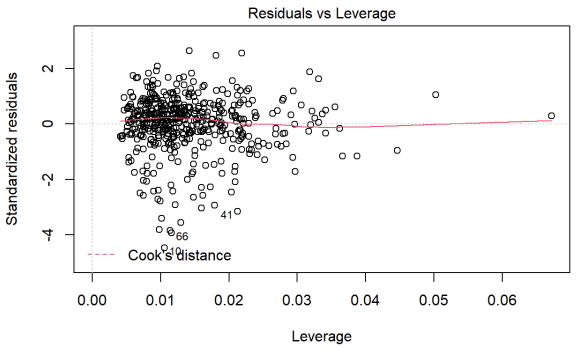

Figure 5. Residual versus Leverage Model II Performance Evaluation Plot

The Residual-Leverage Diagnostic Plot above corresponds to the second selected optimal multiple linear regression model without containing any influential observation beyond the Cook's distance lines because we can hardly observe a dashed Cook's distance indication lines of which all observations fall inside in the plot above. To formally validate whether these two selected optimal multiple linear regression models have met the equal-variance assumption, this research would like to consider adopting the Breusch-Pagan Statistical Test to help us decide whether to implement transformation techniques in terms of generalized least squared model to improve or adjust the optimal model based on the test result.

Since the resulting p-value of Breusch-Pagan Statistical Test heteroskedastic robustness standard error test with respect to these two selected models are 7.634e-05 and 0.0001118 respectively, this research does have sufficient evidence to claim these two selected optimal multiple linear regression models does not satisfy homoskedasticity assumption.

Thus, this research would consider adopting cube transformation techniques generalized least square model to correct the heteroskedasticity of these two selected optimal ordinary least squared multiple linear regressions these two selected optimal ordinary least squared multiple linear regression models by cubing dependent variable of admission chance to improve their performance. Moreover, this research has re-utilized Breusch-Pagan Statistical Test to continue moving on to validate the homoskedasticity assumption for these two selected optimal ordinary least square multiple linear regression models under cube transformation. Hence, both model I and model II under corresponding cube transformation correction have satisfied the homoskedasticity assumption by attaining the higher p-value of 0.3632 and 0.3214 respectively as statistical results achieved based on Breusch-Pagan Statistical Test.

4.DISCUSSION

From the above results, this research has adopted a transformation technique to achieve a satisfactory outcome for making progress on the two selected optimal ordinary least squared multiple linear regression model without sacrificing the goodness of fit. In contrast with other research in this relevant field, this research project has undergone somewhat similar strict model analytical procedures in terms of generating finite number of possible subset ordinary least squared multiple linear regression models under the implementation of sequential variable selection method, relying on different types of model performance evaluation metric paired up with its corresponding selection criterion to formulate the model optimization outcome, utilizing model diagnostic plots to enable researchers to not only be able to detect any violation of linearity, normality, and homoskedasticity respectively as major model assumptions contained in linear regressions but also identify influential observations that might potentially exist in the selected optimal candidate models and adopting power transformation techniques to correct heteroskedasticity of diagnosed model with striking the important balance between prediction accuracy and goodness of fit in the model. This research is also different from other relevant research with exhaustively enumerating all of possible subset of different combinations of ordinary least squared multiple linear regression under the mechanism of sequential variable selection method to further analyze the model performance on a case-by-case basis. Besides, this research topic could be further suggested to dig deeper into analysis of influence of other possible implicit relevant determinants unlisted in the datasets on admission chance estimation.

5.CONCLUSION

This research paper would like to build the best multiple regression admission chance forecast model by taking into account all the relevant admission determinants to provide a guideline for helping post-graduate applicants wished to simulate their likelihood of getting admission with their corresponding admission inputs throughout four different stages of model generation, selection, diagnostic and correction undergone by this research. The main discovery of this paper would be that the relationship between admission chance and admission covariates could be characterized by ordinary least square multiple linear regression under the cubed transformation without sacrificing the goodness of fit. This research could be improved in the future with making use of a variety of different types of statistical regressions in terms of support vector regression, K-nearest neighbors regression, Bayesian ridge regression, random forest regression and decision tree regression as alternatives to compare the performance of the best multiple regressions model under this research with each of these possible alternatives under different evaluation metrics and criterion to make this constructed on-going model

become more credible and practical. In the future, this research might consider adopting equivalent regression form of muti-factor model commonly applied in behavior finance research field to measure other important explicit and implicit admission parameters such as sentiment or mood of admission officer to help prospective applicant better and systematically understand post-graduate admission factors. Besides, this research could also incorporate some stochastic actuarial model to measure the randomness in graduate admission process. Nowadays, higher education is characterized by the diversity of subject content, the diversity of talent training goals, and the diversity of students' personalities. However, because higher education resources are still relatively limited and cannot meet the requirements of all people for higher education, this study optimizes the allocation of higher education resources, especially the resources of elite education, in order to exert its maximum effect.

REFERENCES

[1] A. Bag, (2020). A comparative study of regression algorithms for predicting graduate admission to a university.

[2] A. Iman, & X. Tian, (2021). A Comparison of Classification Models in Predicting Graduate Admission Decision. Journal of Higher Education Theory & Practice, 21(7).

[3] D. Chari, & G. Potvin, (2019). Understanding the importance of graduate admissions criteria according to prospective graduate students. Physical Review Physics Education Research, 15(2), 023101.

[4] I. El Guabassi, Z. Bousalem, R. Marah, & A. Qazdar (2021). A Recommender System for Predicting Students' Admission to a Graduate Program using Machine Learning Algorithms.

[5] M. A. Khan, M. Dixit, & A. Dixit, (2020, April). Demystifying and Anticipating Graduate School Admissions using Machine Learning Algorithms. In 2020 IEEE 9th International Conference on Communication Systems and Network Technologies (CSNT) (pp. 19-25). IEEE.

[6] M. O. F. Goni, A. Matin, T. Hasan, M. A. I. Siddique, O. Jyoti, & F. M. S. Hasnain, (2020, December). Graduate admission chance prediction using deep neural network. In 2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE) (pp. 259-262). IEEE.

[7] M. S. Acharya, A. Armaan, & A. S. Antony, (2019, February). A comparison of regression models for prediction of graduate admissions. In 2019 international conference on computational intelligence in data science (ICCIDS) (pp. 1-5). IEEE.

[8] N. Chakrabarty, S. Chowdhury, & S. Rana, (2020). A statistical approach to graduate admissions' chance prediction. In Innovations in Computer Science and Engineering (pp. 333-340). Springer, Singapore.

[9] S. Aljasmi, A. B. Nassif, I. Shahin, & A. Elnagar, (2020).Graduate Admission Prediction Using Machine Learning.

[10] S. Sujay, (2020). Supervised machine learning modelling & analysis for graduate admission prediction. Int. J. Trend Res. Dev, 7(4), 5-7.