Factors that influence the admission rate of UCLA master program

Yinzhi Chen

2021/12/17

Introduction:

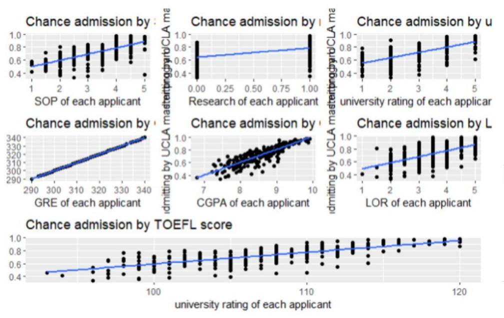
The purpose of this project is to use UCLA Master programs' applicants' background information to research what factors will influence the chance of admission to the master program. Nowadays, about 13 percent of people aged 25 and older have a master degree, which is about the same proportion that had a bachelor degree in 1960, the median wage was about 30 percent higher for employees with a master's degree than for those with a bachelor's degree so there are more and more people choosing to earn a Master's. As an undergraduate student, the importance of researching this helps us to determine the factors that could increase the chance of admission then prepare for applying for the master's program.

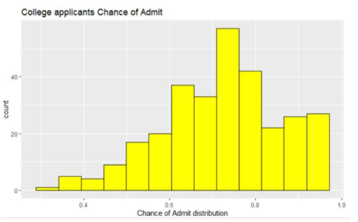
Method selection:

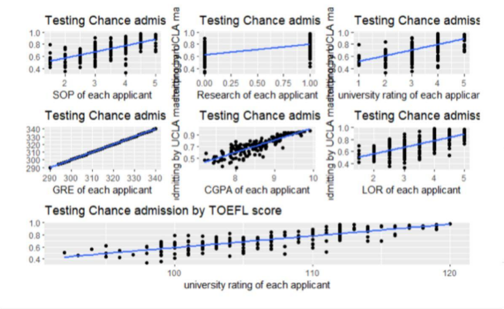
At the beginning of building a model for the analysis, to ensure the effectiveness and accuracy of the model, dataset should be separated into 2 parts which are training and testing dataset for model validation. The training dataset contains 70% of the original dataset which is used to create the model and the testing dataset contains 30% of the original dataset which is used to qualify performance. Then to start the analysis, choosing variables should be the primary step, picking out the response variable, and the numerical and categorical predictors. Since some variables don't have actual meaning such as row number or meaningless variables then we could remove them. Then prepare summary tables and draw histograms, and scatter plots to provide basic information. Whereas starting to build the initial linear regression model. To ensure the predictors are not highly correlated and the model is not influenced by multicollinearity, checking VIF for the model will identify these predictors, predictors with VIF less than 5 are acceptable, whereas the opposite. While all predictors are satisfied, we could move on to check the 2 conditions under the linear regression model. Drawing a scatter plot between Yi and Yi hat to check condition 1 which is the mean response is a single function of a linear combination of the predictors. Drawing scatter plots between predictors to observe whether each predictor is a linear function with another predictor. If both of the conditions are satisfied, move on to check whether the 4 assumptions Linearity, independence, constant variance, and normality of the linear-regression model are satisfied. For checking linearity, independence, and constant variance assumptions, drawing a residual versus fitted plot could help us to justify, if the points are randomly distributed in the graph then the linearity is satisfied; If the points aren't distributed like a cluster in the graph then independence satisfied; If the graph doesn't show a fanning pattern then constant variance satisfied. Checking the normality of the model should apply a normal QQ plot; if the points are approximately distributed around the QQ line then the normality assumption is satisfied, whereas the opposite. However, if any assumptions are violated, a model transformation is required, apply the powertransform to the variables, power these variables by their rounded power and replace the original data then fit the new model. After the model satisfies all assumptions, check whether outliers, leverage, and influential points exist, if any of them exist then determine whether these points are reasonable or find a contextual reason to remove them. Therefore we could select our model applying both automated and manual

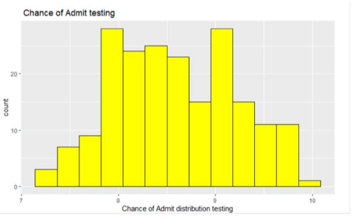
methods, applying automated model selection using R and manually selecting predictors based on common sense or logic to create a manually selected model, then comparing the 2 models using ANOVA table, AIC, and BIC value. Next, summarize both models and check out the 2 conditions and 4 assumptions of the linear regression model. Therefore, combine with the ANOVA table, AIC, BIC results to determine the better model.
Last, do the whole process above again on testing data then compare diagnostic plots and compare leverage points, outliers, and influential points on training and testing data for both models. Ultimately last, making conclusions and choosing the final model based on all of the analysis above and discuss the limitation of the analysis.

Results:
Separate the dataset into training and testing for model validation. Put 70% of the original data into the training for creating models and 30% into testing to qualify performance. Therefore, using the training dataset and starting the analysis, choosing chance.of.admit which is the chance of UCLA master program applicants admission rate as the response variable and all of the other variables CGPA, TOEFL, GRE.Score, SOP, LOR, University.Rating, Serial.No, and Research as predictors since all of them might influence the admission rate. Then summarize and draw both testing and training dataset histograms of response variables and scatter plots for all of the variables. The histogram of response variable chance of admit Reference testing data has an approximately bi-model shape, training data has a unimodal shape with a little bit of left-skewed. Scatter plots of predictors are approximately the same in both datasets.

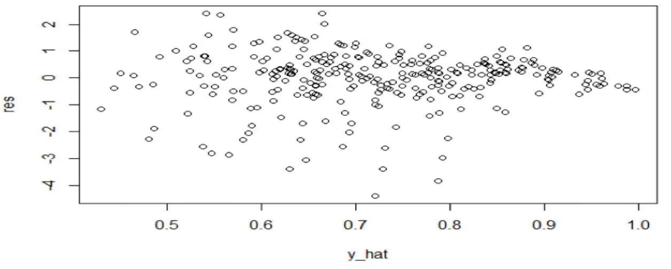


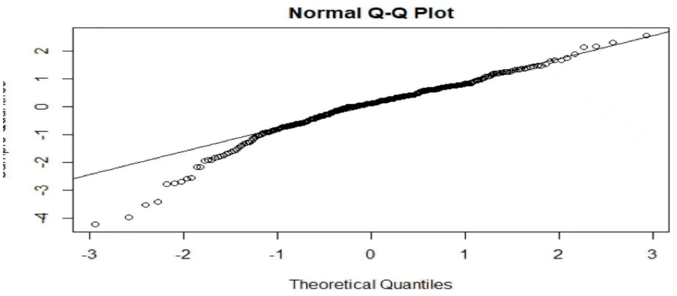






Next, draw the initial linear regression model by removing useless predictors Serial.No. and rowid since they are the variables for organizing the dataset. Then checking the VIF value for all predictors to determine whether multicollinearity exists. Then we observed that the VIF value for GRE.Score, TOEFL.Score, University.Rating, LOR, SOP, CGPA and Research are all less than 5 which proves that multicollinearity could be ignored. Therefore, move on to check out whether the model satisfies the 2 conditions under linear regression. The scatter plot between Yi and Yi_hat appears to be a linear pattern with no divergence or shrinks, so condition 1 satisfied. Next, scatter plots between each predictor appear to be a linear relationship which proves Condition 2. With satisfying both conditions, move on to check the 4 assumptions Linearity, independence, constant variance and normality of linear regression. Using residual and y hat plot shows that it satisfies linearity since points are distributed randomly, also satisfies independence since points are not gathering like a cluster but it slightly violates the constant variance because it has a fanning pattern on the left side of the graph. For checking normality assumption using the normal QQ plot, the points are divergent on the left and right tail which means it also violates normality. Whereas a model transformation is required, apply the powertransform and we observe that variables Chance.of.Admit, GRE.Score , Research and CGPA need to be transformed, so power Chance.of.Admit by 2.5, GRE.Score by 5, apply log on Research and power CGPA by 3 and replace them in the dataset then refit a new model.

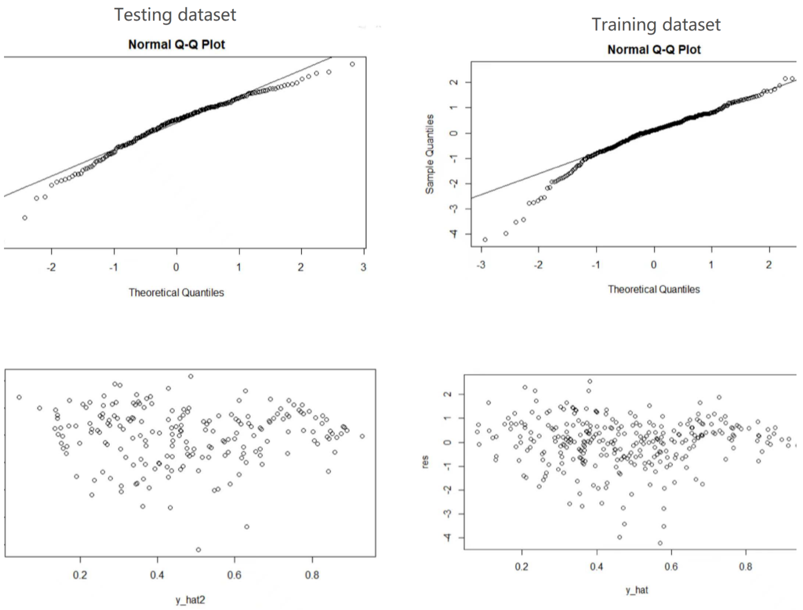




Then it's time to pick out the final model, apply automated model and manually select predictors for manual model, using anova table to compare them with full model, automated model has a p value 0.6248 which is significant so the automated model is better than full model. Manual model has an anova value 0.0006148 which is not really significant so full model is better. Then calculate adjusted r square, AIC and BIC value, the automated model performs better since it has the biggest adjusted R square value, the smallest AIC value and the smallest BIC value.

	Adjusted R square	AIC	BIC
Full model	0.839478	-633.0576	-599.7235
Automated model	0.839895	-634.8114	-605.1811
Manual model	0.831433	-621.3198	-599.0972

Then observe the outliers, leverage points, and influential points, both automated and manual selected models have outliers and influential points, but all of them are valid since the level of applicants' academic background has huge gaps so it's acceptable. Next, repeating the same process of checking conditions and assumptions under linear regression for automated and manual models, we found out that they satisfied both conditions and satisfied 4 assumptions but slightly violated normality, the normality gets way better than the model before transformed.



Therefore, we could start doing the whole process above for testing dataset again and finally compare the training data models and testing data models diagnostic plots. We observe the graphs for checking conditions and assumptions from training and testing datasets are approximately the same which proves validation.

Then compare leverage points, outliers, and influential points in training and testing datasets.

	Leverage point	Outlier	Influential point
Training auto-model	12	1	0
Training manual-model	16	2	0
Testing auto-model	12	1	0
Testing manual-model	15	1	0

But training and testing manual-model have different numbers of Leverage points and outliers.

In conclusion, we could choose the model with automated stepwise selection since it performs the best in all models. Then the interpretation of the final model shows that the admission rate of the UCLA master program is positively influenced by applicants' TOEFL score, university rating, recommendation letters, GRE score, number of researches, and CGPA.

Discussion and limitations:

First, the model still slightly violates the assumption of normality after transformation which should be included in limitations. Second, the dataset contains a lot of leverage points and outliers because the applicants might have huge gaps in their academic backgrounds which will bring influential consequences but they are still valid points because they are not recorded incorrectly.

Reference:

Hymowitz, K. (2021, August 14). More students than ever chase a graduate degree - and society is suffering. New York Post. Retrieved December 18, 2021, from https://nypost.com/2021/08/14/more-students-chasing-graduate-degrees-isnt-good-for-society/

U.S. Bureau of Labor Statistics. (n.d.). Should I get a master's degree? : Career outlook. U.S. Bureau of Labor Statistics. Retrieved December 18, 2021, from https://www.bls.gov/careeroutlook/2015/article/should-i-get-a-masters-degree.htm