Predictive Analysis on Master's Program Admissions

Martina Lipinska, Elias Tamagni, Bethany Walsh

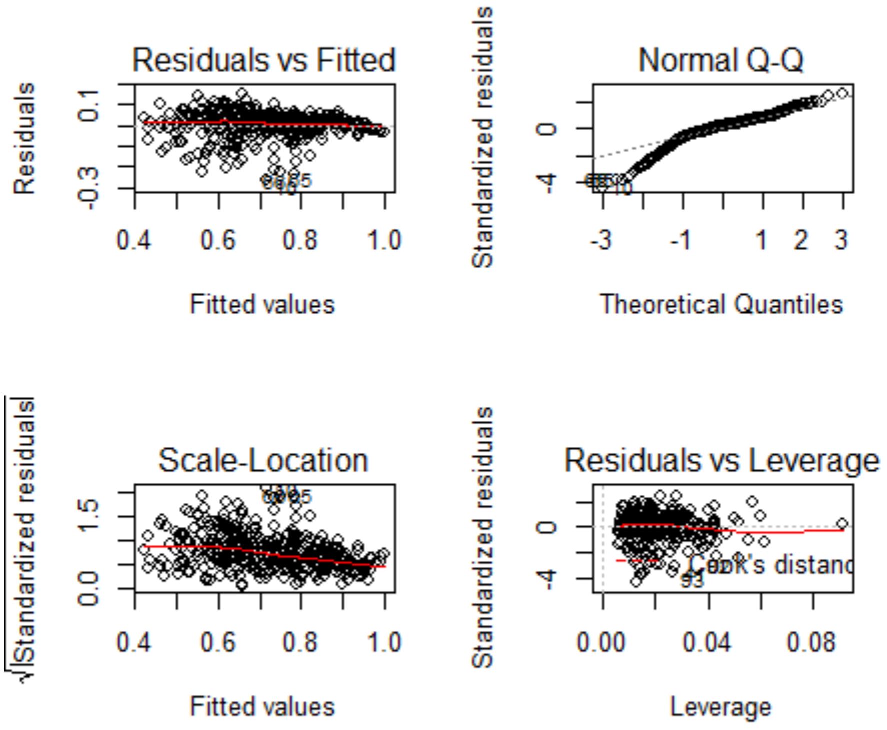
23 April 2019

This analysis examines the likelihood of admission to graduate programs. To perform this analysis, we begin with a dataset that contains the following information on 500 individual observations:

Name	**Description**
Serial.No	Serial number
GRE.Score	GRE score out of 340
TOEFL.Score	Test of English as a Foreign Language out of 120
University.Rating	Rating out of 5
SOP	Statement of Purpose strength out of 5
LOR	Letter of Recommendation strength out of 5
CGPA	Undergraduate GPA out of 10
Research	Research experience, 0 meaning no experience and 1 representing experience
Chance.of.Admit	Between 1 and 0

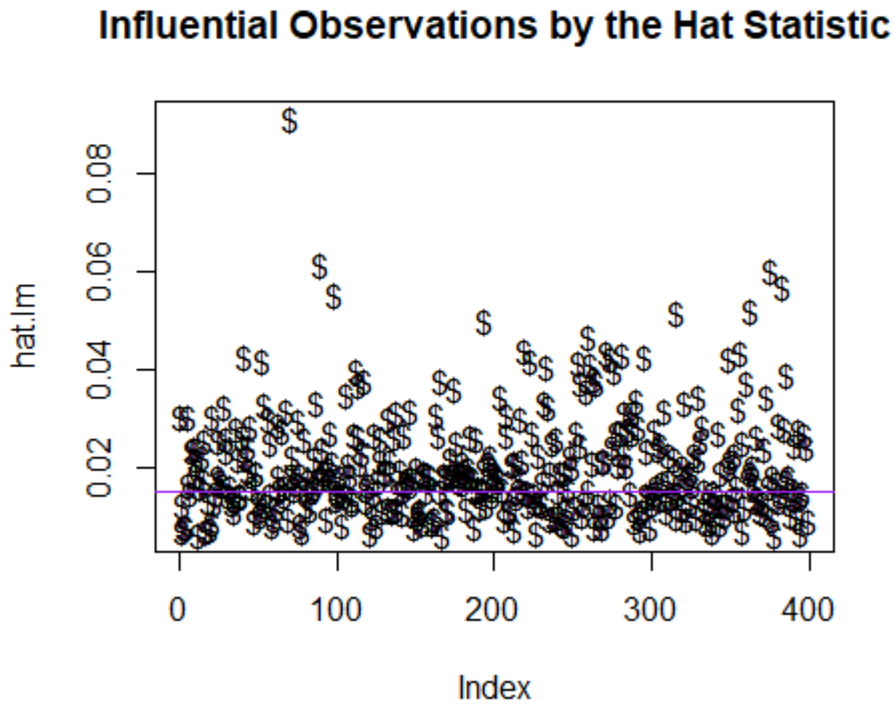
To begin, we first load, clean, and summarize the data in R. This includes removing unnecessary variables, in this case Serial.No, and omitting observations containing 'NA' values. Later in our analysis, we will be building predictive models that require the response variable to be qualitative. For this reason, we create and append an additional dummy variable, admit01, which denotes '1' for an observation with a Chance.of.Admit greater than the median value, and '0' for an observation with a Chance.of.Admit less than the median value. We interpret 1 as being accepted to the Master's program and 0 as being rejected from the program. The final step of preprocessing the data is splitting it into 400 random observations for training, and the remaining 100 observations for testing. With this, we are ready to begin examining various predictive models to assess one’s likelihood of being admitted to a master’s program.

The first model that we create is a linear regression model. In this, we specify Chance.of.Admit as the response variable, and use all of the predefined as predictors. With this, GRE.Score, TOEFL.score, LOR, CGPA, and Research are significant influencers. The next step is to produce the four diagnostic plots for linear regression. The results are as follows:

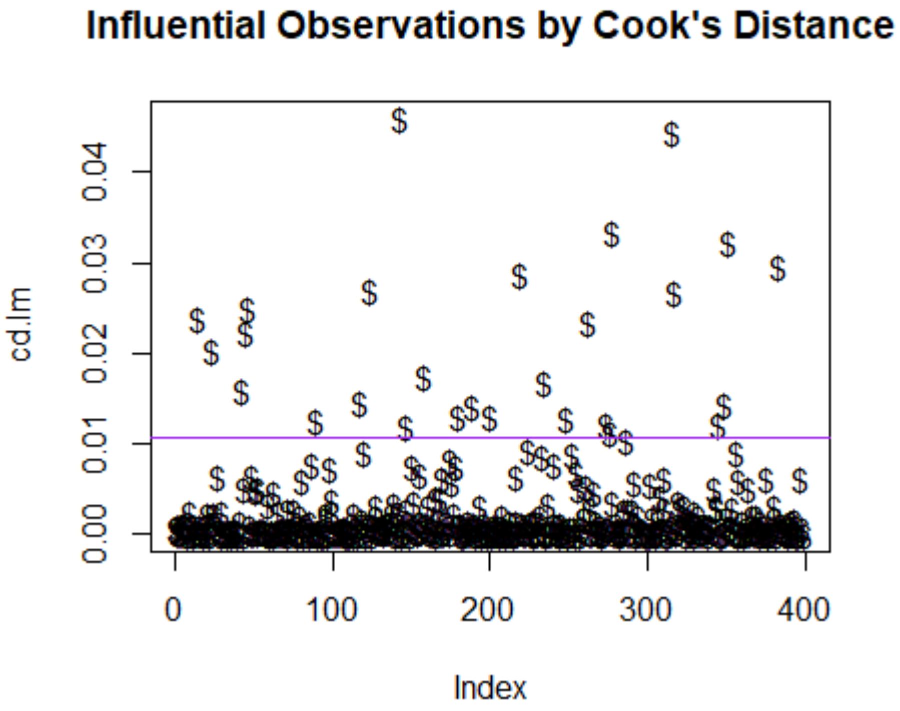


The Residuals vs. Fitted Plot resembles a linear relationship, meaning the data is simulated in a way that meets the regression assumptions very well. The Normal Q-Q Plot demonstrates that the data generally follows a straight line well, with the exception of the leftmost points. With this, we have concluded that the residuals are normally distributed. The Scale Location Plot shows that the residuals are in fact equally spread along the ranges of predictors. Therefore, we can make the assumption of equal variance. The Residuals vs. Leverage Plot, we can barely see the Cook's Distance lines because all cases are well within them. Therefore, we have no extreme cases of outliers to disregard. Based on these diagnostic plots, the linear regression model is a good fit.

The next order of business is to check for potential outliers, specifically using the hat statistic and cook's distance. The plots are as follows:

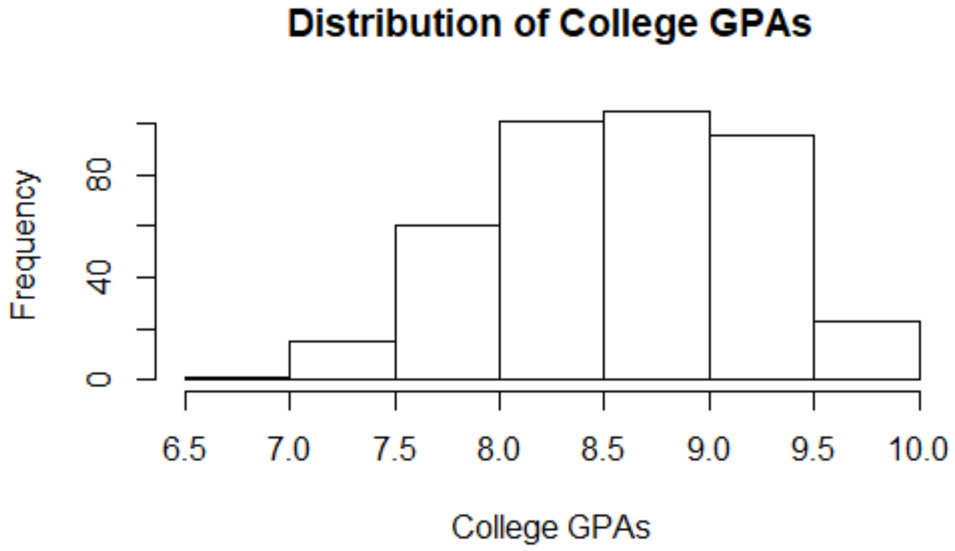


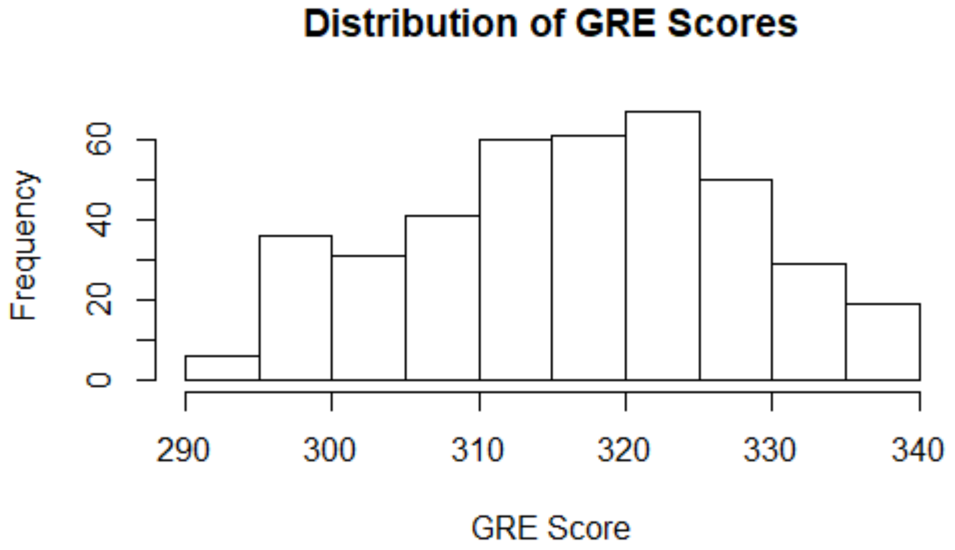
253 training observations are potential outliers/influential observations by the hat statistic.

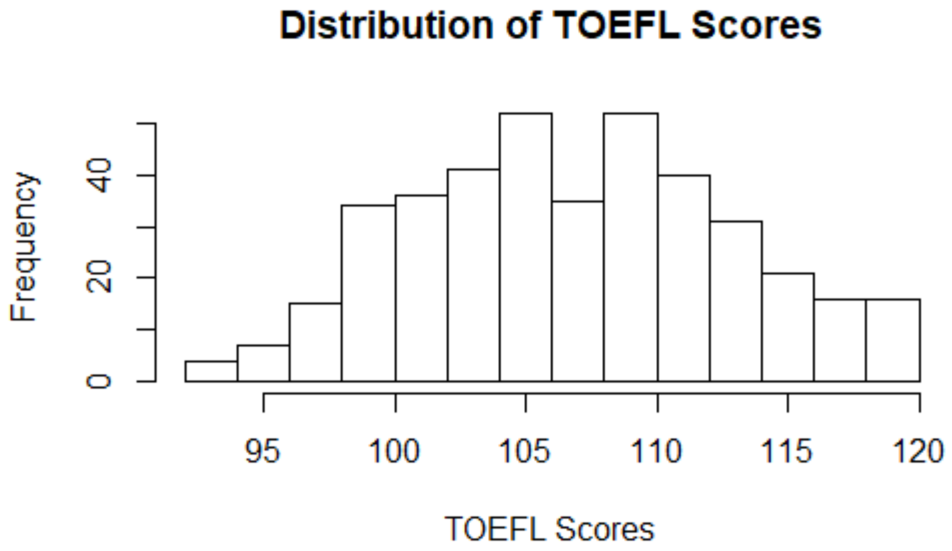


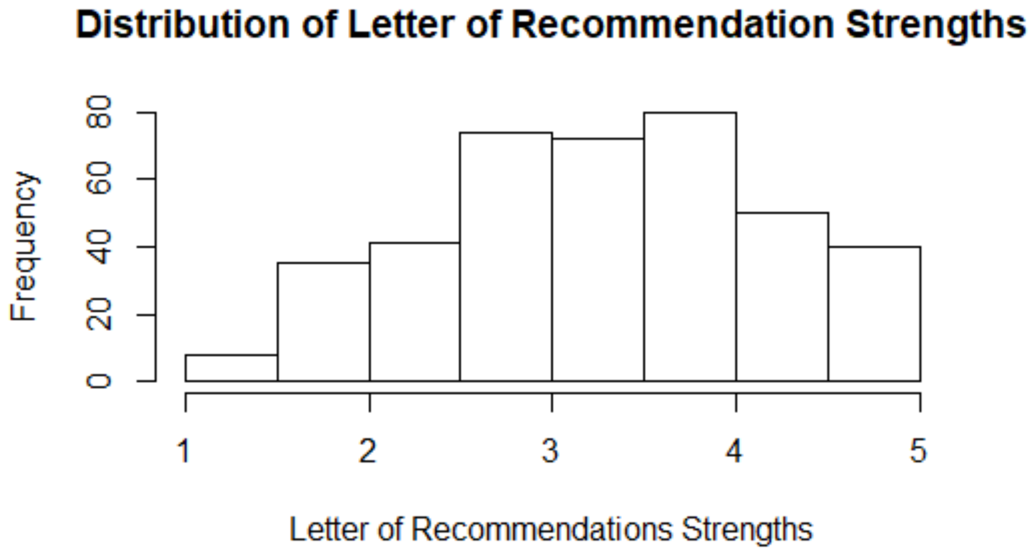
Based on Rule 1 for cook's distance, 27 training observations are potential outliers/influential observations. Based on Rule 2, none of the observations are potential outliers. Therefore, we will keep all observations and not remove any outliers.

The last step for data pre-processing is to check the variables' histograms and measures of skewness to check for even distributions or high skewness. The resulting histograms for the significant variables are as follows:


Distribution of College GPAs
Frequency
0
40
80
6.5
7.0
7.5
8.0
8.5
9.0
9.5
10.0
College GPAs


Distribution of GRE Scores
Frequency
0
20
40
60
290
300
310
320
330
340
GRE Score





Since all the distributions appear to be even and the skewness values are also very low, we assess the linear model once again as a good fit.

We then produce a confusion matrix to demonstrate the accuracy of the training set on the testing set:

	No	**Yes**
No	55	9
Yes	2	34

The error rate for the testing set based on the training model is 11.0%.

We then continue with our analysis by performing three more predictive models: logistic regression, linear discriminant analysis (LDA), and quadratic discriminant analysis (QDA). The resulting error rates are as follows:

	Logistic	**LDA**	**QDA**
Training	12.0%	12.25%	12.5%
Test	11.0%	11.0%	12.0%

To assess the best model on which to make predictions, we use the Aikake and Bayesian information criterion. Based on these, the linear regression model is the best. We then continue to make predictions based on this model. For a student with a GPA of 7 out of 10, the following represent the first 10 rows of predictions made:

	fit	**lwr**	**upr**
1	0.9558608	0.9444315	0.9672901

2	0.8021863	0.7915104	0.8128621
3	0.6517081	0.6390506	0.6643656
4	0.7442235	0.7327554	0.7556917
5	0.6359131	0.6261096	0.6457166
6	0.8664803	0.8535369	0.8794238
7	0.7104969	0.6968701	0.7241236
8	0.5997399	0.5865726	0.6129071
9	0.5554195	0.5414700	0.5693690
10	0.7159647	0.7038279	0.7281015

In conclusion, a student’s chance of admission to a master’ program can be predicted based on their GRE Score, TOEFL Score, College GPA, Letter of Recommendation strength, and research experience.