

MET CS 555 Term Project

Research Scenario Description

This is a general description of the use case assuming that the reader does not have any specialized knowledge about this topic. All acronyms/abbreviations are defined here.

When pursuing a postgraduate education, applicants must fulfill different requirements for the admission committee of the university to review and decide whether to admit said applicant. In majority cases, a Graduate Record Examination (GRE) exam, Test of English as a Foreign Language (TOEFL) exam, a Statement of Purpose (SOP) essay, two to three Letters of Recommendation (LOR) and an official transcript including their Cumulative Grade Point Average (CGPA) from the university of their undergraduate studies are all required. Additionally, possessing research experience may also serve as an additional benefit to applicants. The purpose of this research is to determine the probability of being admitted after considering all the above requirements.

Data set description

Below is a description of each of the columns of the data set **that is used in the analysis**. A step-by-step description of the data cleaning process with reasoning can also be found here.

Attached below is the link to download the original dataset:

<https://www.kaggle.com/datasets/akshaydattatraykhare/data-for-admission-in-the-university>

The dataset contains 400 different applicants applying to different universities with each of the requirements per column and their scores. The university names are not given, but instead a ranking scale out of 5 is given to the universities. Each column in this dataset is as described:

GRE Score: Exam score out of 340

TOEFL Score: Exam score out of 120

University Rating: Random pool of universities ranked out of 5

SOP: Essay graded out of 5 (Assume that the necessary points for a good essay is scaled out of 5)

LOR: Letter is scaled out of 5 (Assume that the necessary points for a good letter is scaled accordingly)

CGPA: Cumulative Grade Point Average out of 10

Research: 1 = Yes, 0 = No

Chances of Admit: A probability score out of 1 with 1 being a guaranteed to being admitted.

Data Cleaning:

This dataset did not require much cleaning except for removing the extra Serial Number column which represented the applicant number. The same serial number can be obtained using the row number. All other columns in the dataset should not be cleaned or removed for potential outliers because in the real-world context, it is very much possible to score full marks for both the GRE and TOEFL exams. Similarly, any university has the potential to score a full 5 out of 5 in the University rating (depending on Ivy-League status, cost of tuition and other factors). The other factors, SOP, LOR and CGPA are all completely possible to obtain a full score. Research is a simple yes or no that should not be removed as some applicants are potentially looking for research opportunities that could potentially stand the applicant out. Finally, the Chances of Admission is the variable that the study will focus more on. The probability score will be studied and observed to see if each of the previously mentioned variables play a vital role in the admission chances.

Research Question

Description the main research question

Research Question:

To determine if each factor considered when applying for postgraduate education at a university has equal or unequal association with probability of being admitted to a randomly selected university.

Statistical Analysis

All relevant assumptions of the statistical technique(s) applied are stated here along with evidence as to whether these assumptions are met.

1. Run the cor() function and display the correlation pair graphs. We want to look at each factor compared against the admission chances. From the cor() function, we can see that each factor has different correlation values that are significantly enough to be different.
2. Calculate the R² to determine how the proportion of the variation in the admission chances is affected by the MLR model containing all the independent variables.
3. Calculate the Least Squares Regression Equation that predicts Admission Chances based on all the above factors. We get the equation below:
Admission Chances = $-1.2594325 + 0.0017374\text{GRE} + 0.0029196\text{TOEFL} + 0.0057167\text{UniRank} - 0.0033052\text{SOP} + 0.0223531\text{LOR} + 0.1189395\text{CGPA} + 0.0245251\text{Research}$
4. Formally test if these factors are associated with admission chances at $\alpha = 0.01$
Using the 5-step method, we set up the hypothesis, choose the appropriate test statistic, state the decision rule, compute the test statistic and conclude.
 $H_0: \beta_{\text{GRE}} = \beta_{\text{TOEFL}} = \beta_{\text{UniRank}} = \beta_{\text{SOP}} = \beta_{\text{LOR}} = \beta_{\text{CGPA}} = \beta_{\text{Research}} = 0$
 $H_1: \text{Any of the } \beta \text{ predictors are not equal to } 0$
Using Global F-Test: $F = \text{MS Reg} / \text{MS Res}$
Decision Rule: Using $qf(0.95, df1 = 7, df2 = 392) = 2.685086$, We reject H_0 if $F \geq 2.685086$, else FAIL to reject
From the summary() function of the model, the F-statistic value is 228.9. So, we REJECT H_0 since $228.9 > 2.685086$
5. Since we rejected the H_0 , the overall model is significant. So we run an inference t-test for each of the parameters to determine the relative contribution of each factor. Here we test at $\alpha = 0.05$ and determine the decision rule to reject H_0 if $|t| \geq 1.965927$, else FAIL to reject.
For each variable, below is the calculated t statistic:
GRE = $0.0017374 / 0.0005979 = 2.905837097$ (Reject H_0)
TOEFL = $0.0029196 / 0.0010895 = 2.67976135842$ (Reject H_0)
University Rating = $0.0057167 / 0.0047704 = 1.19836910951$ (FAIL to Reject H_0)

SOP = $-0.0033052 / 0.0055616 = -0.59428941311$ (FAIL to Reject H0)
 LOR = $0.0223531 / 0.0055415 = 4.03376342146$ (Reject H0)
 CGPA = $0.1189395 / 0.0122194 = 9.73366122723$ (Reject H0) Research = $0.0245251 / 0.0079598 = 3.08112012865$ (Reject H0)

6. Calculate the Confidence Interval at the 95% level:

GRE.Score 0.0005619253 0.002912898
 TOEFL.Score 0.0007775204 0.005061633
 University.Rating -0.0036621610 0.015095477
 SOP -0.0142395495 0.007629211
 LOR 0.0114583790 0.033247876
 CGPA 0.0949156272 0.142963280
 Research 0.0088759540 0.040174259

7. Check the 4 assumptions of the Multiple Linear Regression via plotting the Model (See below at number 7 for each plot used to test assumptions):

a. Linearity

We can confirm Linearity by referring to the Residuals vs Fitted plot. The line of best fit forms a straight line. Here, it is not a perfect line, a subtle rise but still within acceptable range to be considered linear.

b. Independence

They are independent due to the residuals vs fitted plot being scattered. This means the result of an observation is not affected by the results of the previous datapoint. c.

Homoscedasticity

Using the Scale-Location plot, we can see the datapoints form a funnel shape. This indicates a violation of the homoscedasticity.

d. Normality

Use the Normal Q-Q plot. Based on the plot, an estimated +40% of the 400 data points are not on or near the dotted line, all of which reside at the lower end of the plot. Due to this deviation, the Normality assumption is also considered to be violated.

Conclusion

Based on the analysis, 2 out of the 4 assumptions for Multiple Linear Regression have failed (Homoscedasticity and Normality). These violations potentially damage the interpretations about whether each factor considered when applying for postgraduate education at a university has equal or unequal association with probability of being admitted to a randomly selected university. It is highly recommended to consider alternative approaches to overcome the violations, including but not limited to transforming the attributes or using a more robust regression method that are less sensitive to violations of assumptions.

Plots

Below are all the plots used for analysis









