Illinois Institute of Technology — **School of Applied Technology**
ITMD-514-01: Programming for Data Analytics — **PROJECT REPORT**

NAME: Padmashri Adgonda Malgonnavar	NAME: Chinmay Thakare
HAWKID: padgondamalgonnavar@hawk.iit.edu	HAWKID: cthakare@hawk.iit.edu
CWID: A20457612	CWID: A20454398

PROJECT: GRADUATE SCHOOL ADMIT PREDICTION

INTRODUCTION

Objective:

The number of people opting for a masters grows year on year and there is an increase in competition to get a seat at the best ranked university increase simultaneously. The most preferred country for a masters is The United States of America. Generally, as the students don't have much idea about the procedures, requirements and details of the universities in the USA they seek help from the education consultancy firms to help them successfully secure admission in the universities which are best suitable for their profile. Graduate admissions can be a mapping problem between students and universities where students always face a dilemma deciding universities of their choice while applying to a master's programs. Universities qualify students based on certain criteria, one of the most important being the test scores of a student. Students applying to any university in the United State of America, must have appeared for Graduate Record Examination (GRE)/Graduate Management Admission Test (GMAT) and International English Language Testing System (IELTS)/Test of English as a Foreign Language (TOEFL) and have a grading system from a recognized undergraduate school. Based on these scores, a student might qualify to be considered for a seat in the university.

In this project, we will be using the graduate admission dataset which is in a csv format to predict the chances of a student getting admitted to a university based on various academic and non-academic scores. Given a set of standardized scores like the GRE, TOEFL, Letter of recommendation, Statement of purpose the university may also consider other factors such as background of the student, research papers published, work experience. For this project we focus on making use of the statistical data to create a model and use it for predicting the chance of admission for a student with a given set of scores.

Dataset:

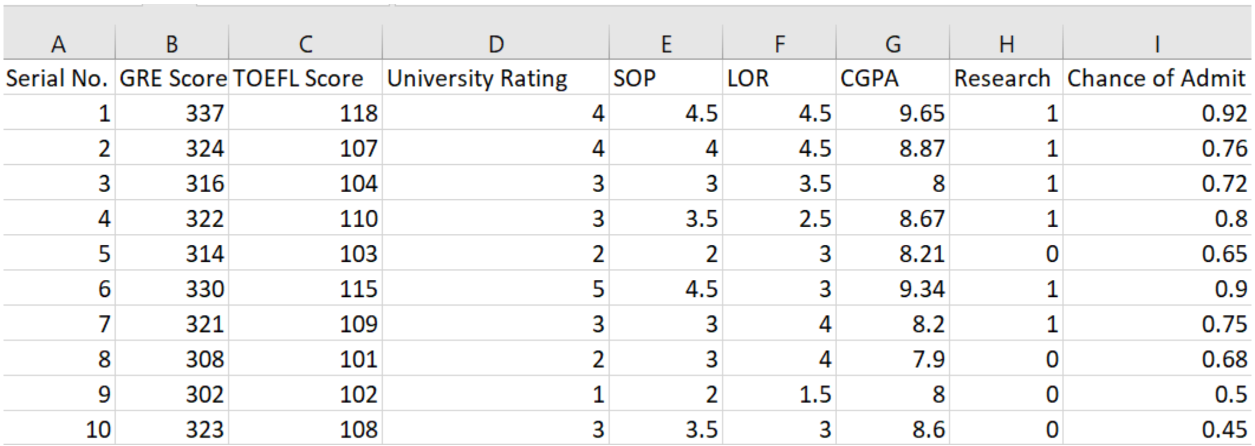
The dataset consists of 500 records with 8 variables:

1) GRE scores (out of 340)
2) TOEFL scores (out of 120)
3) University rating (out of 5)
4) SOP (out of 5)
5) LOR (out of 5)
6) CGPA
7) Research paper published (1 or 0)
8) Chance of admit (ranging from 0 to 1)

The parameters that help in determining the chances of admit are students GRE scores, TOEFL scores, University rating, SOP, LOR, CGPA, Research. Students are expected to declare these scores when applying to any university in the USA.

The data is placed in google drive and the link of the same is provided below:

https://drive.google.com/open?id=18Y3j9IDSuNhjuFpGpZ5XXTrFPs-eg8Cd



A	B	C	D	E	F	G	H	I
Serial No.	GRE Score	TOEFL Score	University Rating	SOP	LOR	CGPA	Research	Chance of Admit
1	337	118	4	4.5	4.5	9.65	1	0.92
2	324	107	4	4	4.5	8.87	1	0.76
3	316	104	3	3	3.5	8	1	0.72
4	322	110	3	3.5	2.5	8.67	1	0.8
5	314	103	2	2	3	8.21	0	0.65
6	330	115	5	4.5	3	9.34	1	0.9
7	321	109	3	3	4	8.2	1	0.75
8	308	101	2	3	4	7.9	0	0.68
9	302	102	1	2	1.5	8	0	0.5
10	323	108	3	3.5	3	8.6	0	0.45

Figure: Admission predict dataset.

Dataset Information:

The dataset contains a list of scores of a student applying for graduate school such as GRE (Graduate Record Examinations), TOEFL (Test of English as a Foreign Language), CGPA (Cumulative Grade Point Average) and rating of the student’s SOP (Statement of Purpose) and LOR (Letter of Recommendation) on a scale of 5 which a student obtains from their undergraduate professors or their superior at work. The dataset also includes the applying students undergraduate school rating on a scale of 5 such that a rating of 5 is the best and a field for Research which describes if a student has a research paper published (represented by 1) or not (represented by 0). The chance of admit field describes the percentage chance, a student has in securing a seat at graduate school for a given set of scores. When applying to graduate schools in the United States a student needs to report their GRE, TOEFL and CGPA scores along with their SOP and LOR which is graded by the graduate school they are applying to. The University Rating information is collected from a credible rating agency/organization. Having a research paper published may or may not affect a student’s chance of admit and we will test the hypothesis in this project. Here the chance of Admit variable is the dependent variable and all the other variables are independent.

Summary of dataset in R:

We start with reading the dataset from the google drive folder. For this the ‘googledrive’ library must be installed. After reading the dataset we produce the summary and structure of the dataset. The summary shows information about the fields in the dataset such as the minimum, maximum, mean, median values. The First quartile (25%) which is one fourth way along from the first observation to the last observation divides the sample data in such a way that 25% of the values are less than the first quartile and 75% are more than first quartile. Whereas the third Quartile (75%) is three fourth way along the way from the first observation to the last observation which divides the sample data in such a way that 75% of the values are less than third quantile and 25% of the values are more than the third quantile. Next, we check the structure of the dataset which shows us the number of observations, number and type of variables along with first 10 values in each variable in the dataset. Next, we check for any missing/NA values in the dataset.

METHODOLOGY

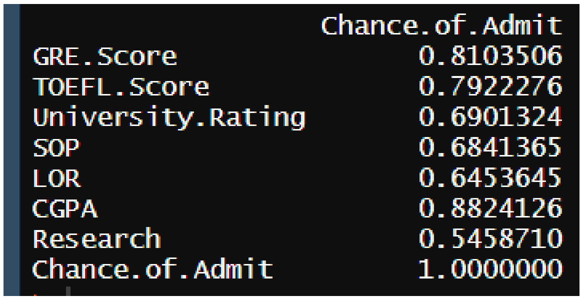
In this paper we perform hypothesis test (T-test) for second hypothesis problem where to determine if the research paper improves the chances of admission. Initially, we split our dataset into two groups based on the value of the Research column. Meaning, calculating the mean between students with research experience and students without research. Once the mean is calculated between two groups, we find that the mean between two groups were different and hence, we perform t.test to check whether these groups of student's data have different mean value of admission chance. To solve third hypothesis problem, we will build logistic regression model and the function to be called is glm(). R makes it very easy to fit a logistic regression model. We split the data into two chunks, that is, training and testing set. The training set will be used to fit our model which we will be testing over the testing set. In glm() function we specify the parameter family to binomial. Later, we then check the accuracy of our model.

IMPLEMENTATION

Hypothesis 1 -

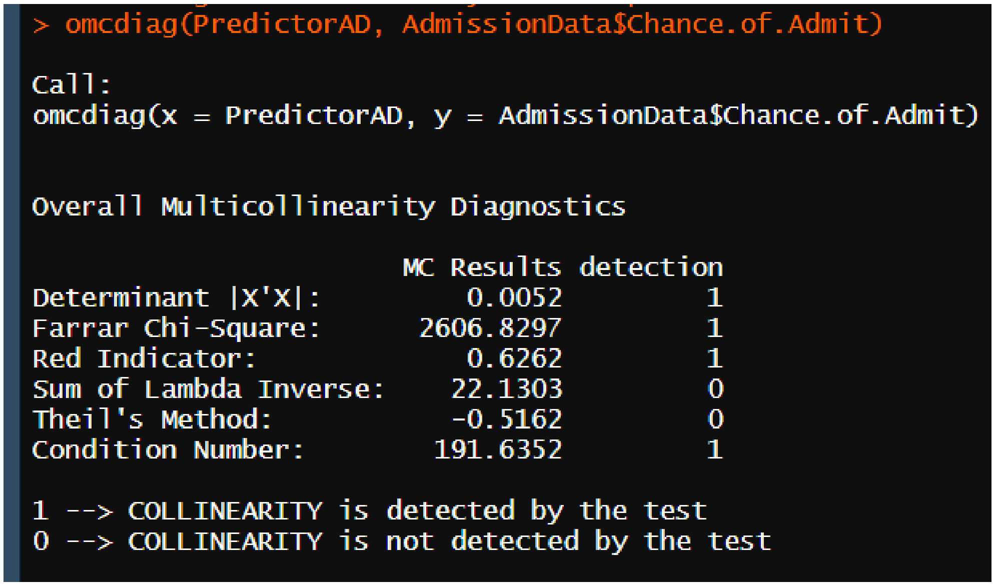
Determine if there is a relation among the various factors affecting the admission decision chances.

For this we check how the variables are correlated with each other.



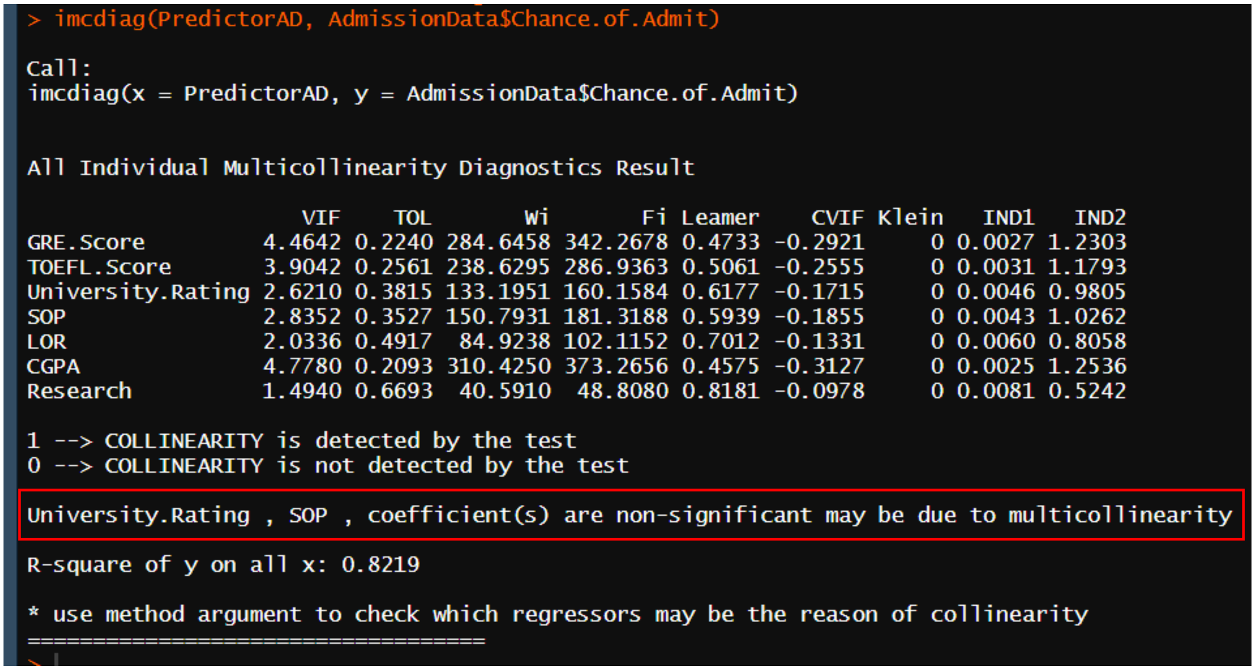
```
                  Chance.of.Admit
GRE.Score               0.8103506
TOEFL.Score             0.7922276
University.Rating       0.6901324
SOP                     0.6841365
LOR                     0.6453645
CGPA                    0.8824126
Research                0.5458710
Chance.of.Admit         1.0000000
```

We then check for presence of multi collinearity in the dataset which requires the ‘mctest’ library to be installed. We test all the independent variables against the dependent variable (Chance of Admit) to detect multi collinearity with the function ‘omcdiag. This function uses various methods such as Chi-Square, Theil’s Method to detect multi collinearity.



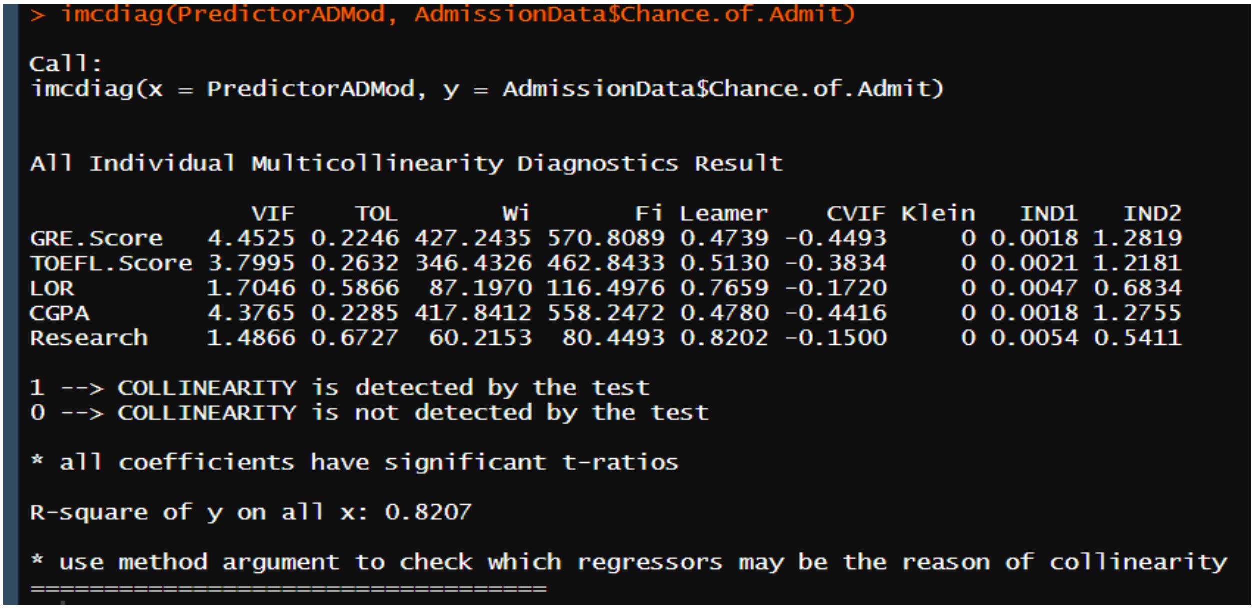
```
> omcdiag(PredictorAD, AdmissionData$Chance.of.Admit)

Call:
omcdiag(x = PredictorAD, y = AdmissionData$Chance.of.Admit)


Overall Multicollinearity Diagnostics

                       MC Results detection
Determinant |X'X|:         0.0052         1
Farrar Chi-Square:      2606.8297         1
Red Indicator:             0.6262         1
Sum of Lambda Inverse:    22.1303         0
Theil's Method:           -0.5162         0
Condition Number:        191.6352         1

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test
```

Since multi collinearity is detected by this function we check for the variable that are non-significant using the ‘imcdiag’ function.



```
> imcdiag(PredictorAD, AdmissionData$Chance.of.Admit)

Call:
imcdiag(x = PredictorAD, y = AdmissionData$Chance.of.Admit)


All Individual Multicollinearity Diagnostics Result

                     VIF    TOL       Wi       Fi Leamer    CVIF Klein   IND1   IND2
GRE.Score         4.4642 0.2240 284.6458 342.2678 0.4733 -0.2921     0 0.0027 1.2303
TOEFL.Score       3.9042 0.2561 238.6295 286.9363 0.5061 -0.2555     0 0.0031 1.1793
University.Rating 2.6210 0.3815 133.1951 160.1584 0.6177 -0.1715     0 0.0046 0.9805
SOP               2.8352 0.3527 150.7931 181.3188 0.5939 -0.1855     0 0.0043 1.0262
LOR               2.0336 0.4917  84.9238 102.1152 0.7012 -0.1331     0 0.0060 0.8058
CGPA              4.7780 0.2093 310.4250 373.2656 0.4575 -0.3127     0 0.0025 1.2536
Research          1.4940 0.6693  40.5910  48.8080 0.8181 -0.0978     0 0.0081 0.5242

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

University.Rating , SOP , coefficient(s) are non-significant may be due to multicollinearity

R-square of y on all x: 0.8219

* use method argument to check which regressors may be the reason of collinearity
===================================
```

After removing the variables identified by the ‘imcdiag’ function that are non-significant we perform the same test again to check for multi collinearity. After the second iteration no multi collinearity is detected.



```
> imcdiag(PredictorADMod, AdmissionData$Chance.of.Admit)

Call:
imcdiag(x = PredictorADMod, y = AdmissionData$Chance.of.Admit)


All Individual Multicollinearity Diagnostics Result

               VIF    TOL       Wi       Fi Leamer    CVIF Klein   IND1   IND2
GRE.Score   4.4525 0.2246 427.2435 570.8089 0.4739 -0.4493     0 0.0018 1.2819
TOEFL.Score 3.7995 0.2632 346.4326 462.8433 0.5130 -0.3834     0 0.0021 1.2181
LOR         1.7046 0.5866  87.1970 116.4976 0.7659 -0.1720     0 0.0047 0.6834
CGPA        4.3765 0.2285 417.8412 558.2472 0.4780 -0.4416     0 0.0018 1.2755
Research    1.4866 0.6727  60.2153  80.4493 0.8202 -0.1500     0 0.0054 0.5411

1 --> COLLINEARITY is detected by the test
0 --> COLLINEARITY is not detected by the test

* all coefficients have significant t-ratios

R-square of y on all x: 0.8207

* use method argument to check which regressors may be the reason of collinearity
===================================
```

Based on the above tests we can conclude that variables like GRE score, TOEFL score and CGPA are highly correlated (~>0.8) with our independent variable (Chance of Admit). The SOP and University Rating variables are determined to be non-significant may be due to multi collinearity by the ‘imcdiag’ function.

Hypothesis 2 -

Determine if publishing research paper improves the chances of admission.

In the second hypothesis, we separated the dataset into two such that one group has data for students with research paper and the other group without any research paper. Next, we calculate the mean of Chance of Admit for each group. We then made use of T-test (t.test) which shows us whether the means of two groups are equal. The assumption for the test is that both groups are sampled from normal distributions with equal variances. The null hypothesis is that the two means are equal, and the alternative is that they are not. Based on this test we observe that the mean of the group with research is more than the mean of the group without research. Thus, we can conclude that students with research paper have a higher chance of admit than students with no research paper.

The hypothesis can also be tested using the ‘ggplot’. We need to have the ‘ggplot2’ library installed and factorize the binary variable Research to create a ggplot. The interval variable University rating is factorized for hypothesis 3.

The boxplot created by the ‘ggplot’ function shows us that the students with Research paper (1) have a higher chance of Admit than students without research paper (0) except for a few outliers.

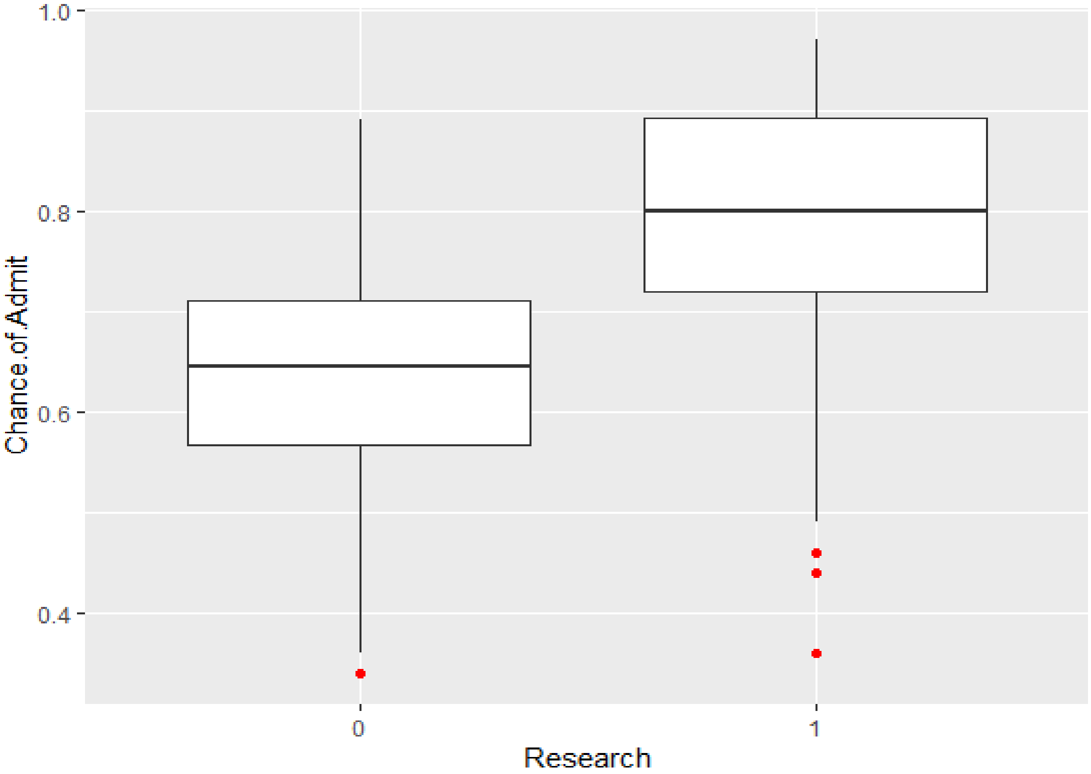


Figure: Boxplot of research variable against Chance of Admit.

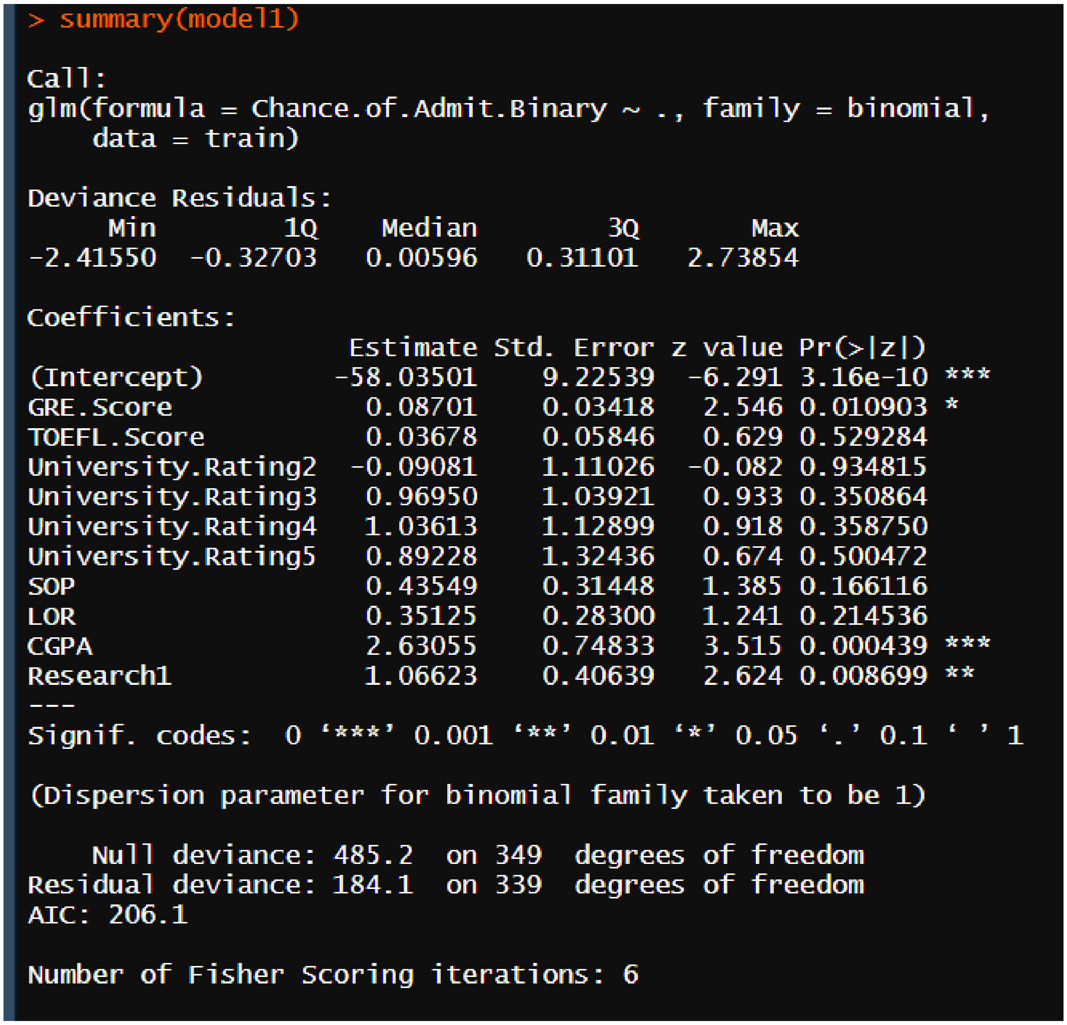
Hypothesis 3-

Create a model to predict the chance of admit

In the third hypothesis, we will use logistic regression and create a model to predict the chance of admit and test its accuracy. For performing logistic regression, the outcome/dependent variable should be binary. For this the 'Chance.Of.Admit' variable should be transformed to binary. We start with testing number records having chance of admit greater than 0.5. Here we find that the data gets divided unequally with a large difference in the number of records. We ultimately found 0.72 as the optimum value for chance of admit dividing the data equally. We then created a separate binary field 'Chance.of.Admit.Binary' which stores the value '1' if chance of admit is greater than/equal to 0.73 and '0' if chance of admit is less than 0.73

We set the Chance of Admit and Serial No. variable to null as they are of no use for performing logistic regression. We have then created 2 datasets: ‘train’ and ‘test’ which has 70% and 30% records from the main data, respectively. The ‘train’ set would be used to predict the outcome for ‘test’ set.

Once all the requirements are ready we created a model using logistic regression that has ‘Chance.of.Admit.Binary’ as the independent variable and all other columns as predictors from the ‘train’ data set. The summary of the model gives us the intercept and co-efficient of each variable along with the standard error.



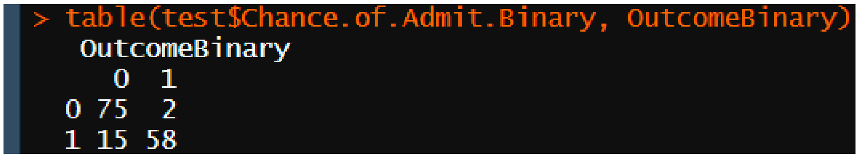
```
> summary(model1)

Call:
glm(formula = Chance.of.Admit.Binary ~ ., family = binomial,
    data = train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.41550  -0.32703   0.00596   0.31101   2.73854

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -58.03501    9.22539  -6.291 3.16e-10 ***
GRE.Score            0.08701    0.03418   2.546 0.010903 *
TOEFL.Score          0.03678    0.05846   0.629 0.529284
University.Rating2  -0.09081    1.11026  -0.082 0.934815
University.Rating3   0.96950    1.03921   0.933 0.350864
University.Rating4   1.03613    1.12899   0.918 0.358750
University.Rating5   0.89228    1.32436   0.674 0.500472
SOP                  0.43549    0.31448   1.385 0.166116
LOR                  0.35125    0.28300   1.241 0.214536
CGPA                 2.63055    0.74833   3.515 0.000439 ***
Research1            1.06623    0.40639   2.624 0.008699 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 485.2  on 349  degrees of freedom
Residual deviance: 184.1  on 339  degrees of freedom
AIC: 206.1

Number of Fisher Scoring iterations: 6
```

Figure: Logistic regression model output.

Here we observe that University Rating variable is represented into multiple variables (‘University.Rating2’, ‘University.Rating3’, ‘University.Rating4’, ‘University.Rating5’) since we have factorized the variable (University.Rating). It is important to note that for categorical predictors, R uses one of the levels as a reference. So, all coefficients for a categorical predictor measure the impact of that level with respect to the reference level. Due to this ‘Univerity.Rating1’ is not seen in the result. In calculating the predicted probability,

We then make use of predict function to calculate the chance of admit on the ‘test’ data and the values get stored in the ‘Outcome’ variable. Since the predicted values are in the range 0 to 1 we convert these values back to binary to check the accuracy with the original ‘test’ data using the same conditions (>=0.73 values transform to 1, >0.73 transform to 0).

We then check the accuracy of our model by comparing the true positives and true negatives.



```
> table(test$Chance.of.Admit.Binary, OutcomeBinary)
   OutcomeBinary
     0  1
  0 75  2
  1 15 58
```

The Accuracy is calculated by,

Accuracy = ((True Positive + True Negative) / (True Positive + True Negative + False Positive + False Negative)) * 100

Accuracy = ((75+58) / (75+58+15+6=2)) * 100

= 89%

Thus, we conclude that our model is 89% accurate.

Conclusion:

In this project we have created a model using logistic regression, in predicting the chances of admission based on given parameters. We conclude from the first hypothesis that, by checking the correlation between various factors, the SOP and University Rating variables are determined to be non-significant may be due to multi collinearity by the multi collinearity test. We also conclude that publishing research paper improves the chances of admission, based on boxplot and T-test. Using graduate admission dataset which contains a list of scores in tests, SOP, LOR, University Rating and Research paper information we created a model to predict the chance of admission and calculated the model's accuracy for the third hypothesis. We found the accuracy of our model to be 89%.