An Analysis Of Graduate Admissions

Brendan Hawk

2022-08-18

Analysis of Graduate Admissions

Data

The data for this analysis is the Graduate Admission dataset[1], and was accessed on Kaggle[2]. This is version 1.1 of this dataset, which was expanded to have more observations than the first version. This is a synthetic dataset that includes 500 records of graduate school candidates' scores and ratings for the many parts of their application packages. The dataset is inspired by another synthetic dataset[3] originally created by the UCLA Office of Advanced Research Computing (OARC), formerly known as the Office of Information Technology (IDRE)[4] with similar but fewer input variables. The data for this analysis will will be included alongside this report in a separate file named `Admission_Predict_Ver1.1.csv`. The data is used as-is, with no cleaning or pre-processing necessary.

Each observation in this dataset includes values for the following variables:

- **GRE**: the Graduate Record Examinations score
- **TOEFL**: the Test of English as a Foreign Language score
- **URating**: the University Rating (out of 5)
- **SOP**: A strength rating (out of 5) for the candidates Statement of Purpose
- **LOR**: A strength rating (out of 5) for the candidates Letter of Recommendation
- **UGPA**: The candidates Undergraduate GPA (out of 10)
- **Research**: A binary flag indicating whether or not the candidate had research experience (1) or not (0), and
- **CtA**: The response variable: a percentage indicating the Chance to Admit for this candidate, presented as a decimal between 0 and 1.

We will first explore the data, then present two simple analyses attempting to statistically test the relationship between the 7 input variables and the Chance to Admit. We are interested in whether or not research experience is significantly associated with a higher Chance to Admit, and whether a regression model formed from the other values for test scores and ratings can be used to predict the expected Chance to Admit.

[1]Mohan S Acharya, Asfia Armaan, Aneeta S Antony : A Comparison of Regression Models for Prediction of Graduate Admissions, IEEE International Conference on Computational Intelligence in Data Science 2019

[2]https://www.kaggle.com/datasets/mohansacharya/graduate-admissions

[3]https://stats.oarc.ucla.edu/stat/data/binary.csv

[4]https://stats.oarc.ucla.edu/

Exploration

One of the most fundamental assumptions of the statistical methods applied in the analyses below is that the values of our input variables are approximately normal. Fig. 1 succinctly shows that this holds for the continuous variables. Note that the university rating is treated here as a continuous variable, but is in a grey area between continuous and categorical: it is a numeric value but one that is only reported in integer increments between 1 and 5 inclusive. We could, if needed for another type of analysis, treat this as categorical.

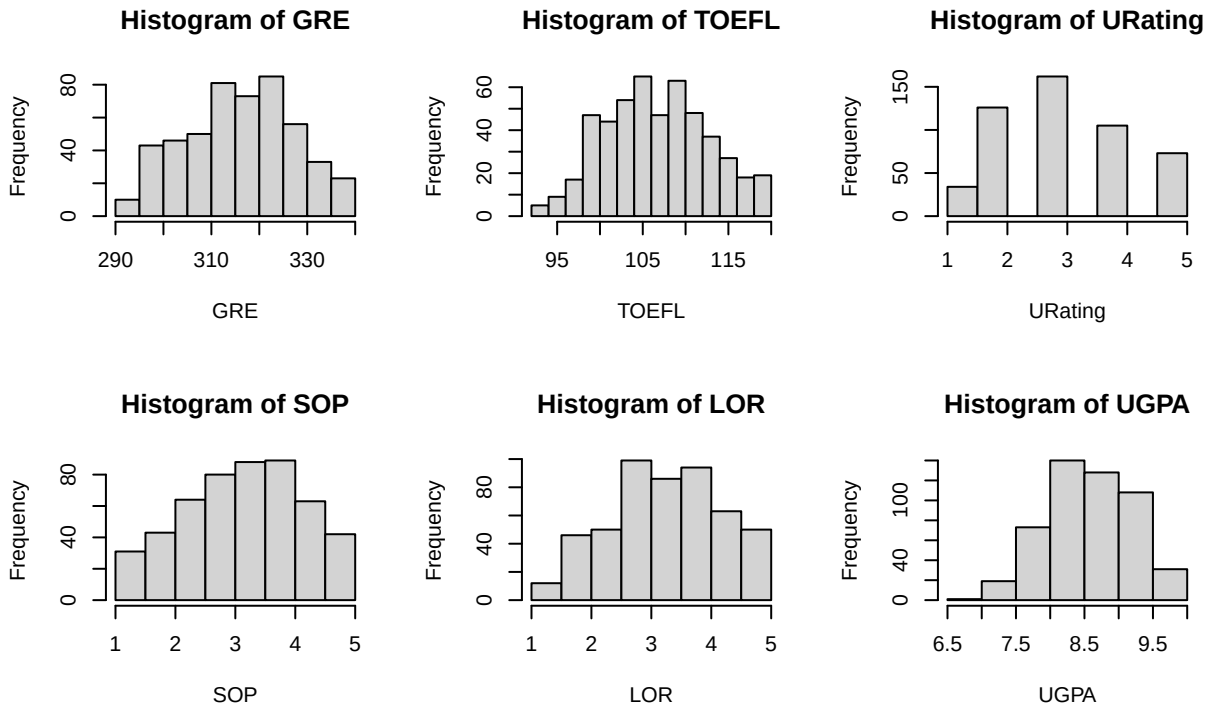

Figure 1: Normalcy of Explanatory Variables

The one categorical variable, Research, should also be inspected for incomparable amounts of variation or disproportionate differences in the response values for the two sub-populations. Fig. 2. shows that while there are differences in the means for each of these sub-populations, they are comparable in their variability.

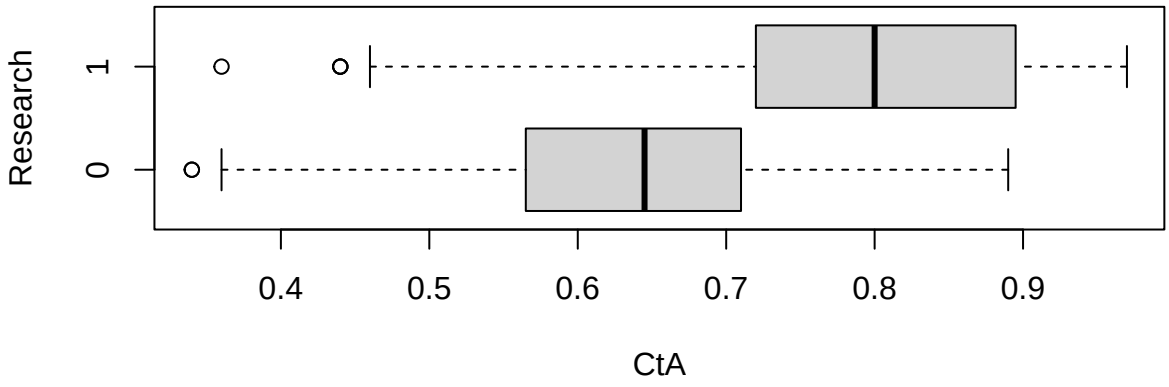

Figure 2: Comparison of Mean Chance to Admit by Research

Now that we have seen that the explanatory variables are not themselves skewed or otherwise problematic, we can begin to explore their relationship with the response variable. Figs. 3 and 4 show pairplots to begin visually inspecting the relationships and correlations in this dataset. We can see that all the continuous explanatory variables have a moderate to very strong positive linear correlation with the response variable. This will be particularly important in the second analysis performed.

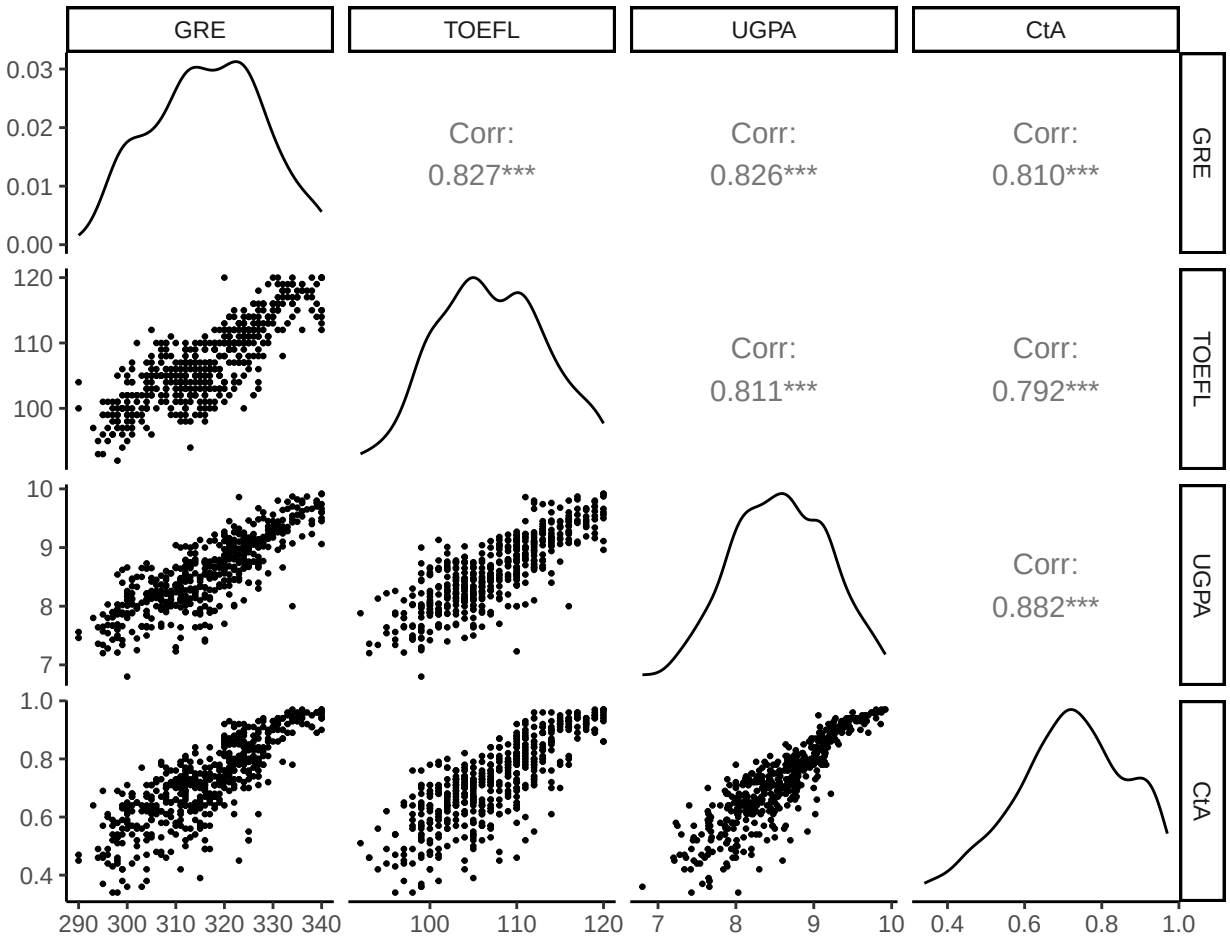

Figure 3: Correlation of Student Academic Scores and Chance to Admit

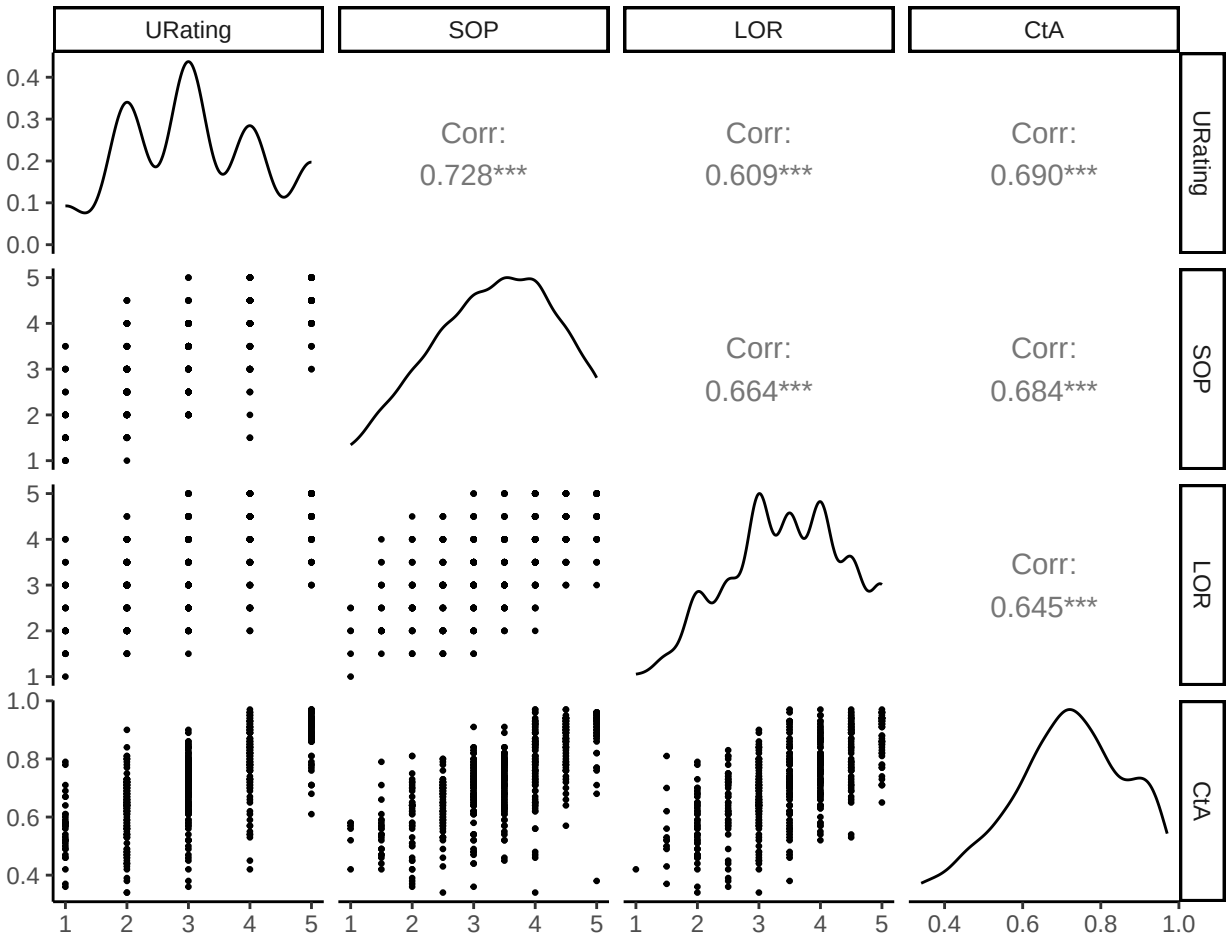

Figure 4: Correlation of Student Candidate Ratings and Chance to Admit

Research

Analysing the Effect of Research on Chance to Admit

The first analysis will be a comparison of mean values for Chance to Admit given the presence or absence of research experience on the candidates' application package. Particularly in the sciences, this could very well be the make-or-break factor when determining an applicant's candidacy for a graduate program, or something that makes a candidate's application package stand out from their peers'.

We begin by summarising the two populations disaggregated by whether or not they had research experience. Fig. 2 shows a noticeable difference in the mean Chance to Admit for these two groups. Table 1 below validates this, as well as validating that the variability in each group is approximately comparable.

Table 1: Summary of Mean Chance to Admit by Research

Research	N	Mean	SD
0	220	0.6349	0.1119
1	280	0.7900	0.1232

We can begin a formal test of the theory that having Research experience increases the expected mean Chance to Admit by setting up the following hypotheses:

$$H_0 : \mu_0 = \mu_1$$
$$H_1 : \mu_0 < \mu_1$$

That is, the null hypothesis is that the means of both groups are equal, and our alternative hypothesis will be that the mean of group 0 is less than the mean of group 1. We will conduct this one-tailed test at the $\alpha = 0.05$ level.

Using software, our derived critical value from the t-distribution will be -1.648, which is the left-hand value of this distribution at our given α with 487.605 degrees of freedom. Therefore, since this is a one-sided comparison and we are testing for $\mu_1 < \mu_2$ our decision rule will be to reject H_0 if $t < -1.648$.

This test will derive it's results from the test statistic for a two-sample comparison of means.

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Using software, this value is calculated as -14.7073. Therefore, we reject the null hypothesis H_0 that the means of group 0 and group 1 are equal. There is significant evidence at the $\alpha = 0.05$ level that the mean of group 0 is less than the mean of group 1 ($p = 4.3351 \times 10^{-41}$). The 95% confidence interval for the difference in means between group 0 and group 1 is (-0.1725, -0.1377). That is, we are 95% confident that the true difference in the population means between group 0 and group 1 is between -0.1725 and -0.1377. In agreement with our original research question, this indicates that the presence of research experience on a candidate's application is expected to result in a Chance to Admit that is higher by 13.77% to 17.25%.

Multiple Linear Regression

The second analysis will determine if the continuous explanatory variables in this dataset can be formally proven to have a predictive association with the Chance to Admit. To do this, we will perform a Multiple Linear Regression with 6 input variables: GRE, TOEFL, URating, SOP, LOR, and UGPA. It should be noted again that I have chosen to treat the University Rating as a continuous variable. It would be prudent in future work to treat this as a categorical variable and note the difference in the resulting regression.

Before we perform this regression, it is important to consider the underlying assumptions that make this kind of analysis appropriate for our given data:

1. **The true relationship is linear**: this is not one we can prove *a priori*, but Figs. 3 and 4 do seem to indicate linear relationships where we need them.
2. **The observations are independent**: this is another unprovable assumption that has more to do with the derivation of the sample data than the data itself. However given the source of this dataset this assumption is one we can make.
3. **The variation of the response variable around the regression line is constant**: This will have to be revisited after the regression is performed.
4. **The residuals are normally distributed**: This will have to be revisited after the regression is performed.

We can also take another look at the correlation strengths from Figs. 3 and 4, succinctly reproduced in Table 2 below. This is useful for double-checking that we aren't running the risk of including a set of colinear inputs for our regression. In particular, it stands out that UGPA is strongly correlated CtA ($r = 0.8824$) as well as with GRE ($r = 0.8259$) and TOEFL ($r = 0.8106$). TOEFL and GRE are also strongly correlated with each other ($r = 0.8272$). Including all of these together could cause unexpected interactions to occur when the regression is run. For example: since GRE and UGPA are strongly correlated with each other *and* with the response variable, future work might consider running the same regression twice including only one of these variables and comparing the residuals or other metrics. That said, experimenting with which input variables to include and exclude could be an entire paper by itself so for brevity I am going to keep all variables and see what a first attempt at this regression looks like.

Table 2: Correlation Coefficients for Continuous Explanatory Variables

	TOEFL	URating	SOP	LOR	UGPA	CtA
GRE	0.8272	0.6354	0.6135	0.5247	0.8259	0.8104
TOEFL		0.6498	0.6444	0.5416	0.8106	0.7922
URating			0.7280	0.6087	0.7053	0.6901
SOP				0.6637	0.7122	0.6841
LOR					0.6375	0.6454
UGPA						0.8824

Formally testing this analysis begins by setting up a global test for the complete model.

$$H_0 : \beta_{GRE} = \beta_{TOEFL} = \beta_{URating} = \beta_{SOP} = \beta_{LOR} = \beta_{UGPA}$$

$$H_1 : \beta_{GRE} \neq 0 \; OR \; \beta_{TOEFL} \neq 0 \; OR \; \beta_{URating} \neq 0 \; OR \; \beta_{SOP} \neq 0 \; OR \; \beta_{LOR} \neq 0 \; OR \; \beta_{UGPA} \neq 0$$

This test can be determined by either the F-test value or its related p value. The decision rule will be either:

- Reject H_0 if $F \geq 2.117$, the critical value from the F distribution with 6 and 493 degrees of freedom at the $\alpha = 0.05$ level, or
- Reject H_0 if $p < \alpha$, with the threshold once again set at $\alpha = 0.05$.

The F-test statistic for this global test is calculated using software to be 366.8295, which has a p value of 3.2194×10^{-178}. Therefore on both accounts we reject the null hypothesis H_0 that all coefficients are 0. There is significant evidence at the $\alpha = 0.05$ level that one or more of the coefficients is not 0. That is GRE, TOEFL, URating, SOP, LOR, and UGPA taken together have a linear association that is predictive of Chance to Admit. The adjusted R^2 value for this model is 0.817, meaning approximately 81.70% of the variance in the response variable can be explained using these explanatory variables.

Table 3 lists all the calculated model coefficients along with their their standard errors, t-test, and p values. This table also includes the lower and upper bounds of each coefficient's 95% confidence interval.

Table 3: MLR: Estimates, Errors, Test Statistics, and 95% CI of Coefficients

	Estimate	SE	t	P(>|t|)	Lower CI	Upper CI
(Intercept)	-1.426775	0.0970903	-14.6953	8.336e-41	-1.6175371	-1.236014
GRE	0.002387	0.0004874	4.8967	1.323e-06	0.0014291	0.003345
TOEFL	0.002625	0.0008824	2.9742	3.081e-03	0.0008908	0.004358
URating	0.006716	0.0038440	1.7471	8.124e-02	-0.0008368	0.014268
SOP	0.001925	0.0046194	0.4167	6.771e-01	-0.0071513	0.011001
LOR	0.017700	0.0041837	4.2307	2.779e-05	0.0094799	0.025920
UGPA	0.119249	0.0098249	12.1374	7.623e-30	0.0999447	0.138552

At this point we can begin to formally analyse each coefficient, determining whether or not the results are significant and interpreting their values.

GRE The estimated coefficient for GRE indicates that for each point increase on the GRE the predicted Chance to Admit would rise by 0.0024. There is significant evidence at the $\alpha = 0.05$ level that we would reject a null hypothesis $H_0 : \beta_{GRE} = 0$ meaning that GRE score is predictive of the Chance to Admit when controlling for all other variables ($p = 1.3234 \times 10^{-6}$). We are 95% confident that the true population value for this coefficient is between 0.0014 and 0.0033. That is, for every point increase in GRE score the true increase in the Chance to Admit would be between 0.0014 and 0.0033.

TOEFL The estimated coefficient for TOEFL indicates that for each point increase on the TOEFL the predicted Chance to Admit would rise by 0.0026. There is significant evidence at the $\alpha = 0.05$ level that we would reject a null hypothesis $H_0 : \beta_{TOEFL} = 0$ meaning that TOEFL score is predictive of the Chance to Admit when controlling for all other variables ($p = 0.0031$). We are 95% confident that the true population value for this coefficient is between 0.0009 and 0.0044. That is, for every point increase in TOEFL score the true increase in the Chance to Admit would be between 0.0009 and 0.0044.

URating The estimated coefficient for University Rating would indicate that for every 1 unit increase in this rating the expected Chance to Admit would increase 0.0067 after controlling for all other variables. However, this is not found to be statistically significant. There is not strong enough evidence at the $\alpha = 0.05$ level to reject a null hypothesis $H_0 : \beta_{URating} = 0$ ($p = 0.0812$). This is further confirmed by the 95% confidence interval (-0.0008, 0.0143), which includes 0 between its upper and lower bounds.

SOP The estimated coefficient for Statement of Purpose rating would indicate that for every 1 unit increase in this rating the expected Chance to Admit would increase 0.0019 after controlling for all other variables. However, this is not found to be statistically significant. There is not strong enough evidence at the $\alpha = 0.05$ level to reject a null hypothesis $H_0 : \beta_{SOP} = 0$ ($p = 0.6771$). This is further confirmed by the 95% confidence interval (-0.0072, 0.011), which includes 0 between its upper and lower bounds.

LOR The estimated coefficient for Letter of Recommendation rating indicates that for each 1 unit increase in this rating the predicted Chance to Admit would rise by 0.0177. There is significant evidence at the $\alpha = 0.05$ level that we would reject a null hypothesis $H_0 : \beta_{LOR} = 0$ meaning that Letter of Recommendation rating is predictive of the Chance to Admit when controlling for all other variables ($p = 2.7786 \times 10^{-5}$). We are 95% confident that the true population value for this coefficient is between 0.0095 and 0.0259. That is, for every 1 unit increase in Letter of Recommendation rating the true increase in the Chance to Admit would be between 0.0095 and 0.0259.

UGPA The estimated coefficient for Undergraduate GPA indicates that for each 1 point increase the predicted Chance to Admit would rise by 0.1192. There is significant evidence at the $\alpha = 0.05$ level that we would reject a null hypothesis $H_0 : \beta_{UGPA} = 0$ meaning that Undergraduate GPA is predictive of the Chance to Admit when controlling for all other variables ($p = 7.6232 \times 10^{-30}$). We are 95% confident that the true population value for this coefficient is between 0.0999 and 0.1386. That is, for every 1 point increase in Undergraduate GPA the true increase in the Chance to Admit would be between 0.0999 and 0.1386.

A Note on Further Understanding Coefficient Values One important thing to remember when looking at the significant coefficients here is to keep in mind the *qualitative* story they tell in reference to the Chance to Admit. These regression coefficients, and the Chance to Admit value itself in the original data, are decimal representations of a percentage. It's easy to see a value such as $\beta_{UGPA} = 0.1192$ and consider it small, however the quality of this information becomes more clear as you think further about what a GPA is and what this slope is really telling us. Going from e.g. a 6.5 to a 7.5 GPA (out of 10) is a "one unit" increase, but requires a monumental improvement in one's grades. This is reflected in the coefficient for this variable, where making such a huge improvement would grant a predicted increase of 11.92% Chance to Admit. This difference between e.g. a 83.08% and 95% Chance to Admit is as grand as the required improvement to one's Undergraduate GPA.

A similar observation can be made for the standardized test scores. A single point on the GRE may "only" net a predicted increase in Chance to Admit of 0.0024, but the GRE is scored on a scale out of 340 possible points. Chances are that even small meaningful improvements in your GRE score would be on the scale of tens of points, not single points. So e.g. a 10 point increase would instead add 2.39% Chance to Admit.

Checking Assumptions We must now return to our final two assumptions for this linear regression. We need to confirm that **the variation of the response variable around the regression line is constant** and **the residuals are normally distributed**.

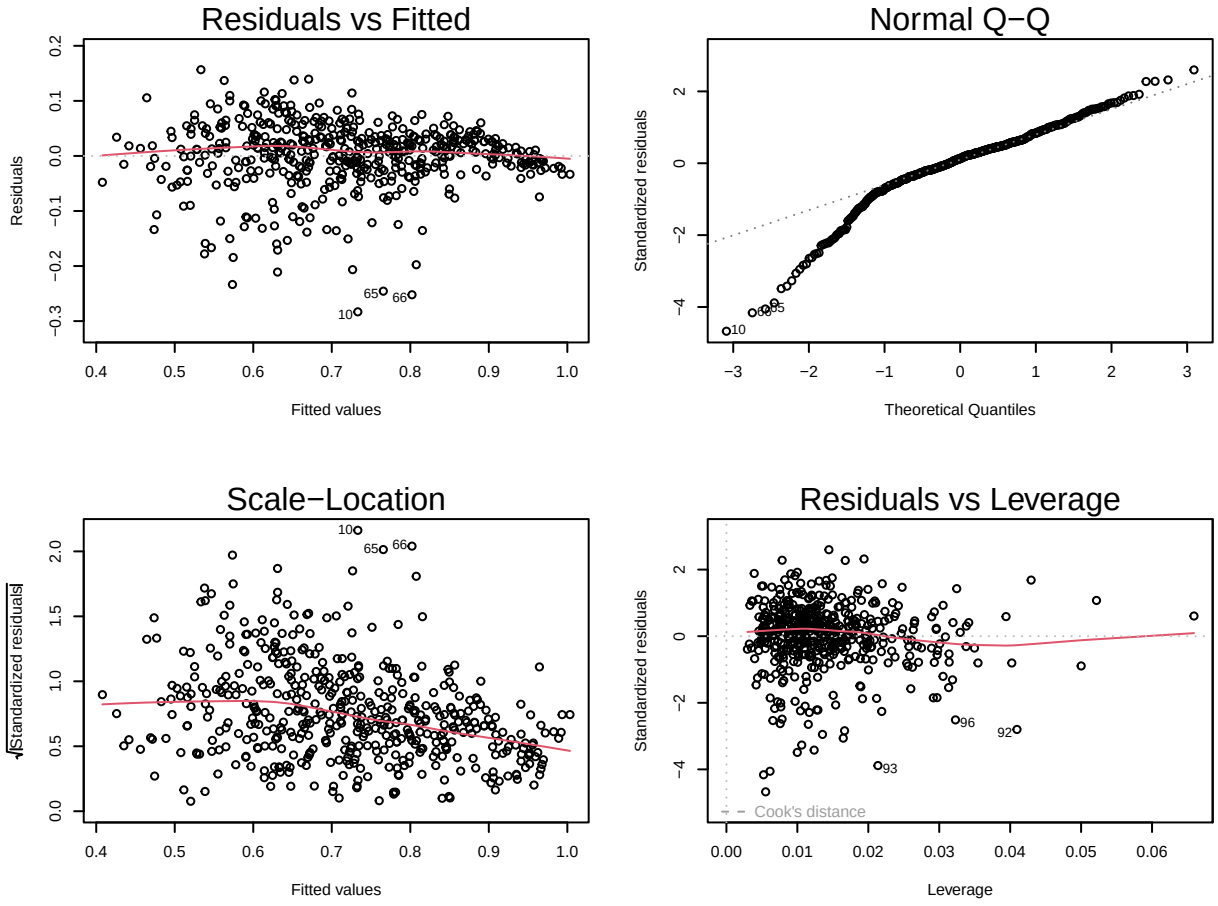

Figure 5: Diagnostic Plots of MLR

While the first, third, and fourth plots in Fig. 5 are as expected, there does seem to be some concerning skew in the Q-Q plot. The assumption about variation would seem to hold, but there is reason to believe the residuals are not as normally distributed as we would hope. In fact there appears to be a significant left skew to these. We can confirm this with further plotting of these residuals.

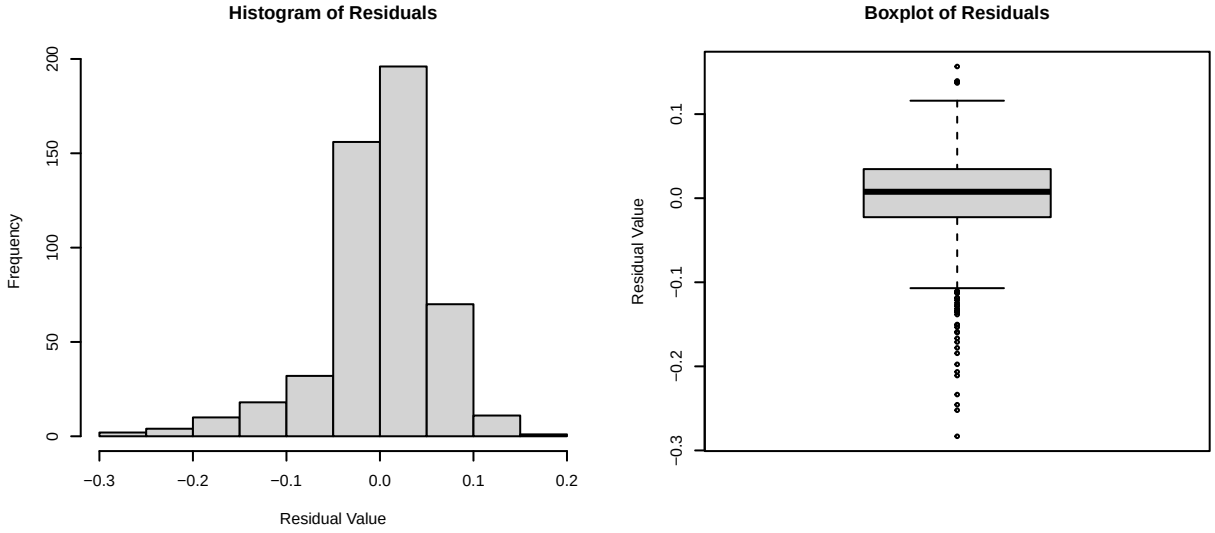

Figure 6: Non-Normalcy of Residuals

Fig. 6 confirms that there is clearly a left skew to the histogram of residuals. We can also see that this is caused by a large number of low-valued outliers in the residuals. Luckily, the inference above is not sensitive to this limitation, however it is a cause for possible further analysis. This would indicate a failure of the underlying assumption that the error terms of our explanatory variables are normal.

Conclusion

Both of the above analyses have statistically significant results. It is clear that there are several approaches to analysing this data that could yield meaningful associations and allow us to understand the relationships between graduate program candidates' application packages and their Chance to Admit. The first analysis in particular clearly shows how strong the predictive relationship of these values can be of a student's Chance to Admit.

There are, however, further questions to be asked. In particular, the regression analysis has skewed residuals that lead to questioning its fundamental assumptions. Future work would need to dig deeper into the explanatory variables and their possible interdependence. It's also possible that some of the continuous variables such as the University Rating are more appropriately handled as categorical variables and thus yield different results in a regression setting.

Appendix A: Code

```r
# Import required libraries, installing any that are missing
required_libraries <- c("ggplot2", "GGally", "dplyr")

for (library_name in required_libraries) {
  if (!require(library_name, character.only = TRUE)) {
    install.packages(library_name, repos = "http://cran.us.r-project.org")
    library(library_name, character.only = TRUE)
  }
}

# By default, keep to a readable amount of decimal places
options(digits = 4, scipen = 1, knitr.kable.NA = "")

# Set a base theme for all ggplot2 charts
theme_set(theme_classic())

# See: https://gist.github.com/burchill/8873d2ade156b27e92a238a774ce2758
# Fix for plots floating out of place in final rendered document
knitr::knit_hooks$set(plot = function(x, options)  {
  paste0(knitr::hook_plot_tex(x, options), "\n\\FloatBarrier\n")
})

# Read the data from disk
# Renaming columns for brevity, there are a LOT of columns to fit into tables headers!
data <- read.csv(
  "Admission_Predict_Ver1.1.csv",
  row.names = "Serial.No.",
  col.names = c(
    "Serial.No.",
    "GRE",
    "TOEFL",
    "URating",
    "SOP",
    "LOR",
    "UGPA",
    "Research",
    "CtA"
  )
)

# We don't need the row names, so remove them
data$`Serial.No.` <- NULL
row.names(data) <- NULL

# Convert this binary value into a factor with levels (0, 1)
data$Research <- factor(data$Research)

# Plot histograms of all the continuous explanatory variables
par(mfrow=c(2,3))
with(
  data,
```

```
  {
    hist(GRE)
    hist(TOEFL)
    hist(URating)
    hist(SOP)
    hist(LOR)
    hist(UGPA)
  }
)

# Compare the means of CtA by Research values
boxplot(
  CtA ~ Research,
  data,
  horizontal = TRUE
)

# Create a pairplot of GRE, TOEFL, and UGPA against CtA
ggpairs(
  data[,c(1,2,6,8)],
  lower = list(continuous = wrap("points", size=0.5))
)

# Create a pairplot of URating, SOP, and LOR against CtA
ggpairs(
  data[,c(3,4,5,8)],
  lower = list(continuous = wrap("points", size=0.5))
)

# Summarise the count, mean, and SD of CtA values by Research
data %>%
  group_by(Research) %>%
  summarise(N = n(), Mean = mean(CtA), SD = sd(CtA)) %>%
  knitr::kable(
    caption = "Summary of Mean Chance to Admit by Research"
  )

# A one-sided two-sample test of means
# H0 : mu1 = mu2
# H1 : mu1 < mu2
alpha <- 0.05
means.test <- t.test(
  CtA ~ Research,
  data,
  conf.level = 1 - alpha,
  alternative = "less"
)

# Get the critical values from the test summary
t <- means.test$statistic[["t"]]
t.df <- means.test$parameter[["df"]]
t.crit <- qt(alpha, df = t.df)
p <- means.test$p.value
```

```r
# Calculate a confidence interval by hand to include as values in the report
# For some reason, t.test() reports this interval as (-Inf, -0.1377)
means.confint <- (0.6349 - 0.7900) +
  (t.crit * c(1, -1) * sqrt((.1119**2/220) + (.1232**2/280)))

# Recap the correlation values numerically.
# In order to make this easier to read, we will only
# present the upper triangle of values in this matrix.
cor_matrix <- cor(data[,-7])
# This replaces the duplicated lower triangle of values with NA
cor_matrix[lower.tri(cor_matrix) | cor_matrix == 1] <- NA
# This removes the first column and last row, which are now completely blank
cor_matrix <- cor_matrix[-nrow(cor_matrix),-1]
knitr::kable(
  cor_matrix,
  caption = "Correlation Coefficients for Continuous Explanatory Variables"
)

# Write out the formula, done the long way here for clarity
# Equivalent to CtA ~ . -Research
mlr.formula <- CtA ~ GRE + TOEFL + URating + SOP + LOR + UGPA

# Create the model, and get the summary
mlr.model <- lm(mlr.formula, data)
mlr.summary <- summary(mlr.model)

# pull out some critical values
df <- c(mlr.summary$fstatistic[["numdf"]], mlr.summary$fstatistic[["dendf"]])
f.crit <- qf(alpha, df[1], df[2], lower.tail = FALSE)
f <- mlr.summary$fstatistic[["value"]]
p.f <- pf(f, df[1], df[2], lower.tail = FALSE)

# Create a single summary table of the coefficients, with their
# standard errors, t-test values, p-values, and 95% CI bounds
mlr.summary.with.ci <- cbind(mlr.summary$coefficients, confint(mlr.model))
knitr::kable(
  mlr.summary.with.ci,
  caption = "MLR: Estimates, Errors, Test Statistics, and 95% CI of Coefficients",
  col.names = c(
    "Estimate",
    "SE",
    "t",
    "P(>|t|)",
    "Lower CI",
    "Upper CI"
  ),
  digits = 200
)

# in a 2x2 grid, plot model diagnostics
par(mfrow = c(2, 2), cex = 0.5)
plot(mlr.model)
```

```r
# In a 1x2 grid, plot a histogram and boxplot
# of the residuals for further investigation
par(mfrow = c(1, 2), cex = 0.5)
hist(
  mlr.model$residuals,
  main = "Histogram of Residuals",
  xlab = "Residual Value"
)
boxplot(
  mlr.model$residuals,
  main = "Boxplot of Residuals",
  ylab = "Residual Value"
)
```