DATA 501 Final Project - Aidan Murphy

Due Date = 2024-12-20

Research Questions

What are the key factors influencing the likelihood of graduate admission?

How well can a regression model predict the chance of admission based on applicant data?

Are interaction terms between predictors significant in improving the model's performance?

Explore and Prepare Data

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

This dataset contains 9 columns and 400 entries pertaining to university admissions data and is sourced from Kaggle. The columns contain information like GRE score out of 340, TOEFL score out of 120, university rating out of 5, SOP (Statement of Purpose) and LOR (Letter of Recommendation) strength out of 5, CGPA, research experience and chance of admit.

```r
library(readr)
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
library(caTools)
library(Metrics)
library(car)
```

```
## Loading required package: carData
```

```r
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v stringr   1.5.1
## v forcats   1.0.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::recode() masks car::recode()
## x purrr::some()   masks car::some()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
data <- read_csv("adm_data.csv")
```

```
## Rows: 400 Columns: 9
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (9): Serial No., GRE Score, TOEFL Score, University Rating, SOP, LOR, CG...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
head(data)
```

```
## # A tibble: 6 x 9
##   `Serial No.` `GRE Score` `TOEFL Score` `University Rating`   SOP   LOR  CGPA
##          <dbl>       <dbl>         <dbl>               <dbl> <dbl> <dbl> <dbl>
## 1            1         337           118                   4   4.5   4.5  9.65
## 2            2         324           107                   4   4     4.5  8.87
## 3            3         316           104                   3   3     3.5  8
## 4            4         322           110                   3   3.5   2.5  8.67
## 5            5         314           103                   2   2     3    8.21
## 6            6         330           115                   5   4.5   3    9.34
## # i 2 more variables: Research <dbl>, `Chance of Admit` <dbl>
```

```r
summary(data)
```

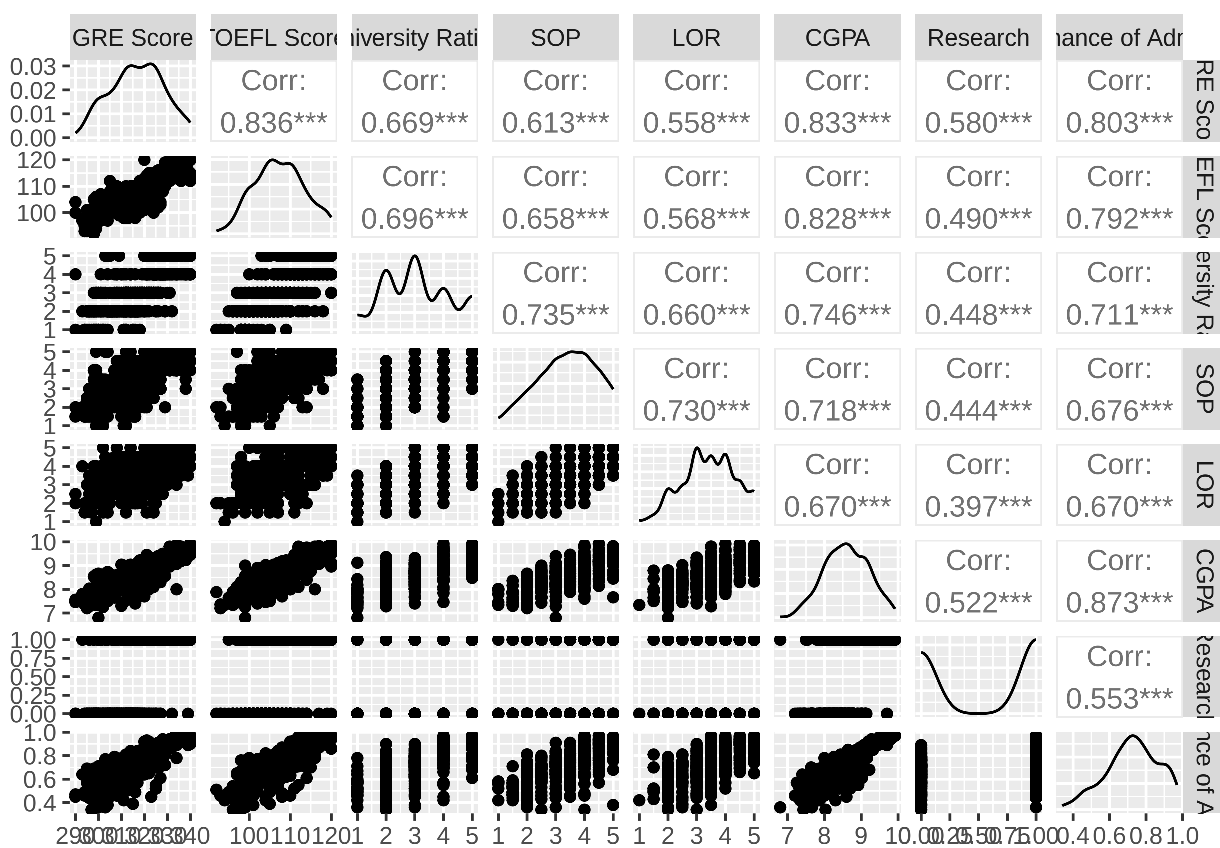
```
##    Serial No.       GRE Score      TOEFL Score    University Rating
##  Min.   :  1.0   Min.   :290.0   Min.   : 92.0   Min.   :1.000
##  1st Qu.:100.8   1st Qu.:308.0   1st Qu.:103.0   1st Qu.:2.000
##  Median :200.5   Median :317.0   Median :107.0   Median :3.000
##  Mean   :200.5   Mean   :316.8   Mean   :107.4   Mean   :3.087
##  3rd Qu.:300.2   3rd Qu.:325.0   3rd Qu.:112.0   3rd Qu.:4.000
##  Max.   :400.0   Max.   :340.0   Max.   :120.0   Max.   :5.000
##       SOP            LOR             CGPA           Research
##  Min.   :1.0   Min.   :1.000   Min.   :6.800   Min.   :0.0000
##  1st Qu.:2.5   1st Qu.:3.000   1st Qu.:8.170   1st Qu.:0.0000
```

```
##  Median :3.5   Median :3.500   Median :8.610   Median :1.0000
##  Mean   :3.4   Mean   :3.453   Mean   :8.599   Mean   :0.5475
##  3rd Qu.:4.0   3rd Qu.:4.000   3rd Qu.:9.062   3rd Qu.:1.0000
##  Max.   :5.0   Max.   :5.000   Max.   :9.920   Max.   :1.0000
##  Chance of Admit
##  Min.   :0.3400
##  1st Qu.:0.6400
##  Median :0.7300
##  Mean   :0.7244
##  3rd Qu.:0.8300
##  Max.   :0.9700
```

```r
colSums(is.na(data))
```

```
##        Serial No.         GRE Score       TOEFL Score University Rating
##                 0                 0                 0                 0
##               SOP               LOR              CGPA          Research
##                 0                 0                 0                 0
##   Chance of Admit
##                 0
```

```r
ggpairs(data, columns = c("GRE Score", "TOEFL Score", "University Rating", "SOP", "LOR", "CGPA", "Resea
```



Model Development

```r
set.seed(123)

split <- sample.split(data$`Chance of Admit`, SplitRatio = 0.7)
train_data <- subset(data, split == TRUE)
test_data <- subset(data, split == FALSE)

full_model <- lm(`Chance of Admit` ~ `GRE Score` + `TOEFL Score` + `University Rating` + SOP + `LOR` + `C
summary(full_model)
```

```
##
## Call:
## lm(formula = `Chance of Admit` ~ `GRE Score` + `TOEFL Score` +
##     `University Rating` + SOP + LOR + CGPA + Research, data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22977 -0.02183  0.01013  0.03538  0.15754
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -1.1721456  0.1477486  -7.933 5.47e-14 ***
## `GRE Score`          0.0018394  0.0007026   2.618 0.009342 **
## `TOEFL Score`        0.0021408  0.0012888   1.661 0.097859 .
## `University Rating`  0.0100230  0.0057824   1.733 0.084155 .
## SOP                 -0.0055521  0.0067518  -0.822 0.411608
## LOR                  0.0222482  0.0066237   3.359 0.000894 ***
## CGPA                 0.1138817  0.0140379   8.112 1.68e-14 ***
## Research             0.0302177  0.0095173   3.175 0.001669 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06365 on 274 degrees of freedom
## Multiple R-squared:  0.8045, Adjusted R-squared:  0.7995
## F-statistic: 161.1 on 7 and 274 DF,  p-value: < 2.2e-16
```

```r
vif_values <- vif(full_model)
print(vif_values)
```

```
##         `GRE Score`       `TOEFL Score` `University Rating`                 SOP
##            4.552965            4.256049            3.043241            3.220805
##                 LOR                CGPA            Research
##            2.530978            5.088465            1.526787
```

The VIF value for CGPA is > 5 indicating multicollinearity. It is statistically significant however, so instead we remove University Rating, SOP, and TOEFL score from the model, which are not. We will see if this affects the fit of the model and if it solves the multicollinearity issue.

```r
reduced_model <- lm(`Chance of Admit` ~ `CGPA` + `GRE Score` + `LOR` + `Research`, data = train_data)
summary(reduced_model)
```

```
##
## Call:
## lm(formula = `Chance of Admit` ~ CGPA + `GRE Score` + LOR + Research,
##     data = train_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.237344 -0.023409  0.007814  0.037770  0.165922
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.2701090  0.1371658  -9.260  < 2e-16 ***
## CGPA         0.1256196  0.0126912   9.898  < 2e-16 ***
## `GRE Score`  0.0025709  0.0006247   4.115 5.11e-05 ***
## LOR          0.0241980  0.0056890   4.253 2.88e-05 ***
## Research     0.0296923  0.0094798   3.132  0.00192 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06405 on 277 degrees of freedom
## Multiple R-squared:  0.7999, Adjusted R-squared:  0.797
## F-statistic: 276.7 on 4 and 277 DF,  p-value: < 2.2e-16
```

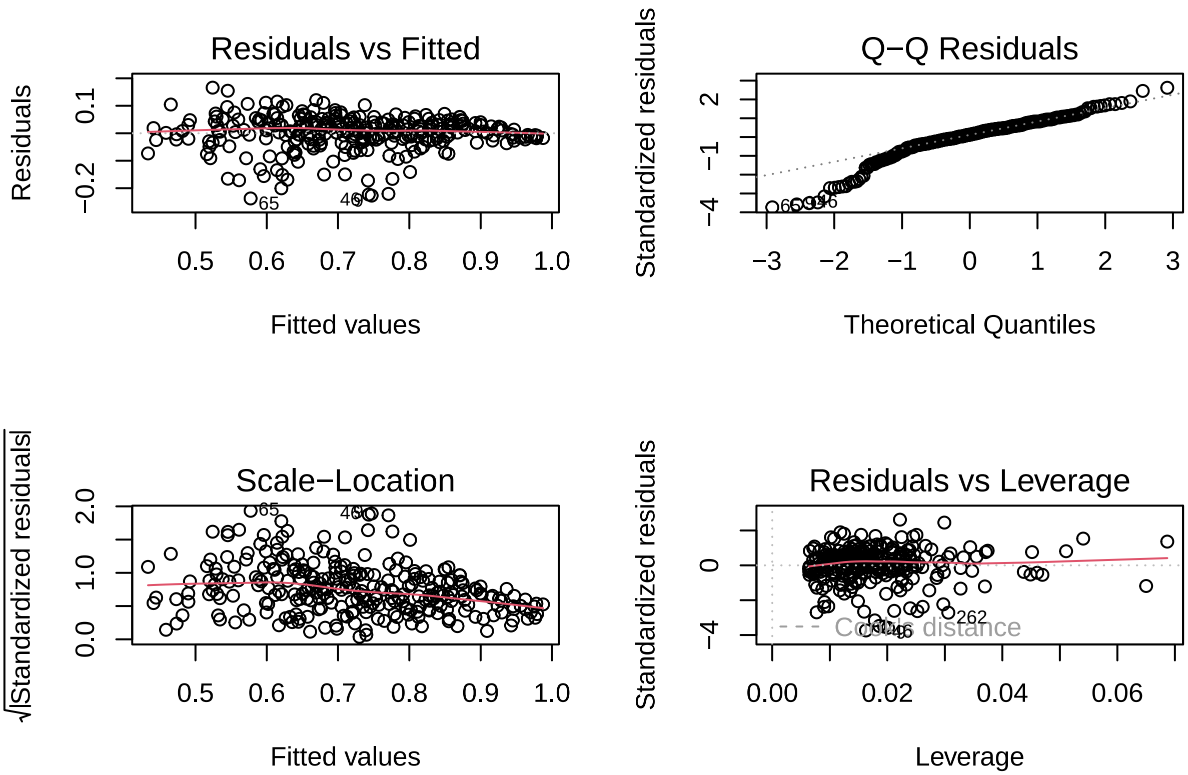
```r
vif(reduced_model)
```

```
##        CGPA `GRE Score`         LOR    Research
##    4.107126    3.554272    1.843817    1.495890
```

```r
anova_result <- anova(full_model, reduced_model)
print(anova_result)
```

```
## Analysis of Variance Table
##
## Model 1: `Chance of Admit` ~ `GRE Score` + `TOEFL Score` + `University Rating` +
##     SOP + LOR + CGPA + Research
## Model 2: `Chance of Admit` ~ CGPA + `GRE Score` + LOR + Research
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    274 1.1100
## 2    277 1.1363 -3  -0.02632 2.1657 0.09234 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
par(mfrow = c(2, 2))
plot(reduced_model)
```




```r
bptest(reduced_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reduced_model
## BP = 12.576, df = 4, p-value = 0.01354
```

$$H_0 : \beta_{\text{extra predictors}} = 0$$

$$H_1 : \beta_{\text{extra predictors}} \neq 0$$

Since the p-value of the f-score is 0.092 which is $> a = 0.05$, we reject the alternative hypothesis that the full model explains more variability in the response variable than the reduced model in favor of the null hypothesis that the reduced model is sufficient. The new VIF values indicate that there is no significant multicollinearity among the four predictor variables. However, the Breusch-Pagan test indicates there is heteroscedasticity present in the model because the $p - value = 0.014 < a = 0.05$.

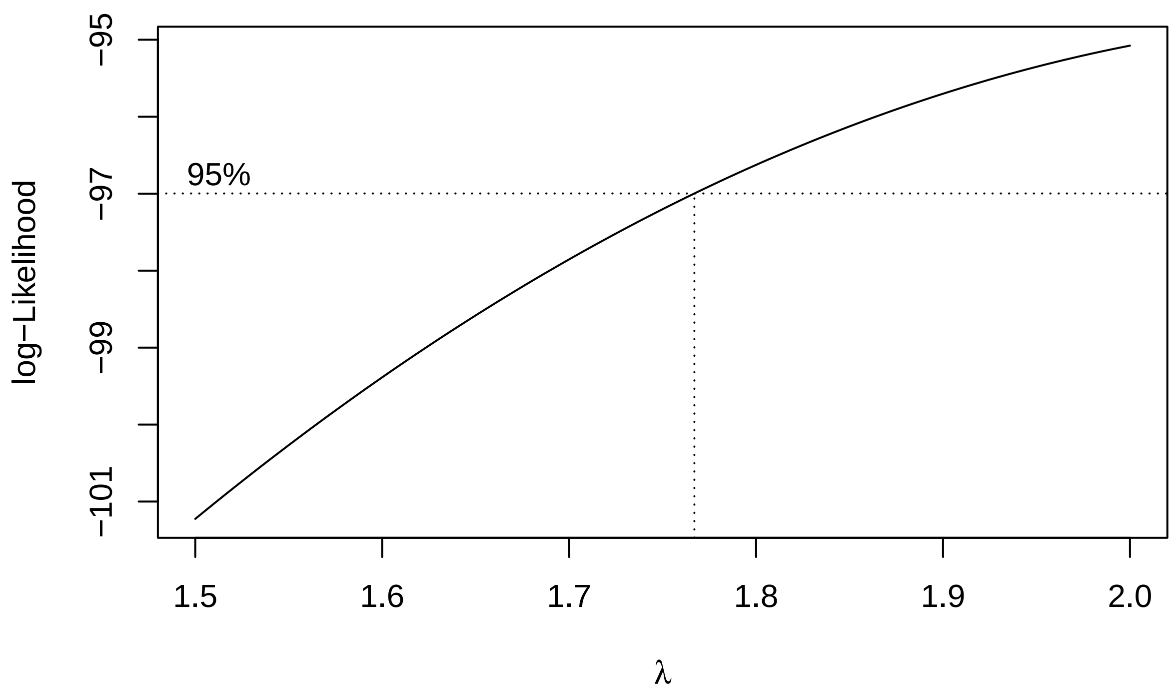
Transformation

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
boxcox_result <- boxcox(reduced_model, lambda = seq(1.5, 2, by = 0.1))
```




```r
lambda_best <- boxcox_result$x[which.max(boxcox_result$y)]


lambda <- 1.75
train_data$transformed_chance <- (train_data$`Chance of Admit`)^lambda

transformed_model <- lm(transformed_chance ~ CGPA + `GRE Score` + LOR + Research, data = train_data)

summary(transformed_model)
```

```
##
## Call:
## lm(formula = transformed_chance ~ CGPA + `GRE Score` + LOR +
##     Research, data = train_data)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29541 -0.03364  0.01398  0.05243  0.21140
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.150866   0.174118 -12.353  < 2e-16 ***
## CGPA         0.171608   0.016110  10.652  < 2e-16 ***
## `GRE Score`  0.003561   0.000793   4.490 1.05e-05 ***
## LOR          0.031316   0.007222   4.336 2.03e-05 ***
## Research     0.040661   0.012034   3.379 0.000832 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0813 on 277 degrees of freedom
## Multiple R-squared:  0.8212, Adjusted R-squared:  0.8186
## F-statistic: 318.1 on 4 and 277 DF,  p-value: < 2.2e-16
```

```r
bptest(transformed_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  transformed_model
## BP = 5.7591, df = 4, p-value = 0.2179
```

```r
par(mfrow = c(2, 2))
```

To achieve homoscedasticity, we can use a Box-Cox transformation. The plot of log-Likelihood suggested the appropriate lambda value was near 1.5-2 so I adjusted the axis and chose 1.75 as λ around where the curve intersects the 95% line. We can then raise the response variable to the power of lambda. After the transformation, I performed another Breusch_Pagan test to see if the heteroscedasticity problem has been resolved. The p-value this time was 0.218 indicating no statistically significant signs of heteroscedasticitiy.

Model with Interaction Term

```r
model_without_interaction <- lm(`Chance of Admit` ~ CGPA + `GRE Score` + LOR + Research, data = train_da
model_with_interaction <- lm(`Chance of Admit` ~ CGPA * `GRE Score` + LOR + Research, data = train_data)

anova(model_without_interaction, model_with_interaction)
```

```
## Analysis of Variance Table
##
## Model 1: `Chance of Admit` ~ CGPA + `GRE Score` + LOR + Research
## Model 2: `Chance of Admit` ~ CGPA * `GRE Score` + LOR + Research
##   Res.Df    RSS Df  Sum of Sq      F Pr(>F)
## 1    277 1.1363
## 2    276 1.1360  1 0.00034907 0.0848 0.7711
```

From the ANOVA table, it is clear that adding an interaction term to this model does absolutely nothing to improve the fit so we will proceed without one.

Predictions

```r
new_data <- tibble(
  CGPA = c(9.0, 8.5, 7.3),
  `GRE Score` = c(330, 320, 340),
  LOR = c(4.5, 4.0, 3.9),
  Research = c(1, 0, 1)
)

predictions <- predict(transformed_model, newdata = new_data, interval = "prediction")

back_transformed <- data.frame(
  fit = predictions[, "fit"]^(1 / 1.75),
  lwr = predictions[, "lwr"]^(1 / 1.75),
  upr = predictions[, "upr"]^(1 / 1.75)
)

print(back_transformed)
```

```
##         fit       lwr       upr
## 1 0.8485300 0.7390966 0.9482576
## 2 0.7270534 0.6015719 0.8380164
## 3 0.6537203 0.5019388 0.7826259
```

With a new tibble containing some random predictor values, we can make predict the chance of admit for new data. For each new set of values, the function estimates a chance of admit value, and then gives an upper or lower prediction interval bound. The interval provides a range of plausible values for which the response variable may reside in. The result is then back transformed to account for the Box-Cox transformation that was applied so that the result is on the right scale.

Results

This analysis focused on building a model to predict graduate admission chances using applicant data. Key variables like GRE Score, CGPA, LOR, and Research were identified as the most important predictors, while multicollinearity issues were resolved through evaluation and removal of unnecessary predictors. Testing revealed that adding interaction terms didn't significantly improve the model, allowing for a streamlined approach. CGPA appeared to be the most influential factor, highlighting its key role in admissions decisions. By addressing heteroscedasticity with a Box-Cox transformation $\lambda = 1.75$, the model achieved an adjusted R-squared value of 0.8186, indicating that approximately 82% of the variability in admission likelihood could be explained by the predictors. Diagnostic plots confirmed that the model satisfied regression assumptions, and predictions for new applicants were back-transformed to provide results in the original scale, along with prediction intervals to show the uncertainty. With a low residual standard error and highly significant F-statistic, the model proves to be reliable. Hypothesis testing indicated that interaction terms did not significantly improve the model's performance so they were left out. In the future, adding new predictors and using a much larger dataset than the sample one used could improve this model significantly by capturing more variability.