# MCAT Modeling & Analysis

Tarek El-Hajjaoui

2023-02-25

**Data loading & pre-processing**

Loading the dataset

```
file_path = '/Users/Tarek/Documents/UCI_MDS_Coding/Stats210P/R_Statistical_Modeling/MCAT/MedGPA.txt'
df = read.table(file_path, header=TRUE, sep="", dec=".")
```

Summary of dataset

```
str(df)
```

```
## 'data.frame':    55 obs. of  11 variables:
##  $ Accept    : chr  "D" "A" "A" "A" ...
##  $ Acceptance: int  0 1 1 1 1 1 1 0 1 1 ...
##  $ Sex       : chr  "F" "M" "F" "F" ...
##  $ BCPM      : num  3.59 3.75 3.24 3.74 3.53 3.59 3.85 3.26 3.74 3.86 ...
##  $ GPA       : num  3.62 3.84 3.23 3.69 3.38 3.72 3.89 3.34 3.71 3.89 ...
##  $ VR        : int  11 12 9 12 9 10 11 11 8 9 ...
##  $ PS        : int  9 13 10 11 11 9 12 11 10 9 ...
##  $ WS        : int  9 8 5 7 4 7 6 8 6 6 ...
##  $ BS        : int  9 12 9 10 11 10 11 9 11 10 ...
##  $ MCAT      : int  38 45 33 40 35 36 40 39 35 34 ...
##  $ Apps      : int  5 3 19 5 11 5 5 7 5 11 ...
```

Transforming categorical columns to factor data type.

```
categorical_cols <- c('Accept', 'Acceptance', 'Sex')
df[categorical_cols] <- lapply(df[categorical_cols], as.factor)
```

Ensuring column data types are correct now.

```
str(df)
```

```
## 'data.frame':    55 obs. of  11 variables:
##  $ Accept    : Factor w/ 2 levels "A","D": 2 1 1 1 1 1 1 2 1 1 ...
##  $ Acceptance: Factor w/ 2 levels "0","1": 1 2 2 2 2 2 2 1 2 2 ...
##  $ Sex       : Factor w/ 2 levels "F","M": 1 2 1 1 1 2 2 2 1 1 ...
##  $ BCPM      : num  3.59 3.75 3.24 3.74 3.53 3.59 3.85 3.26 3.74 3.86 ...
##  $ GPA       : num  3.62 3.84 3.23 3.69 3.38 3.72 3.89 3.34 3.71 3.89 ...
##  $ VR        : int  11 12 9 12 9 10 11 11 8 9 ...
##  $ PS        : int  9 13 10 11 11 9 12 11 10 9 ...
##  $ WS        : int  9 8 5 7 4 7 6 8 6 6 ...
##  $ BS        : int  9 12 9 10 11 10 11 9 11 10 ...
##  $ MCAT      : int  38 45 33 40 35 36 40 39 35 34 ...
##  $ Apps      : int  5 3 19 5 11 5 5 7 5 11 ...
```

### Splitting up data to supplement analysis in 2-sample t-test

Separating the dataset into Female observations & Male observations

```
df_female <- df[df$Sex == 'F',]
df_male <- df[df$Sex == 'M',]
```

Female dataset summary

```
str(df_female)
```

```
## 'data.frame':    28 obs. of  11 variables:
##  $ Accept    : Factor w/ 2 levels "A","D": 2 1 1 1 1 1 1 1 1 1 ...
##  $ Acceptance: Factor w/ 2 levels "0","1": 1 2 2 2 2 2 2 2 2 2 ...
##  $ Sex       : Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
##  $ BCPM      : num  3.59 3.24 3.74 3.53 3.74 3.86 4 3.35 3.26 3.71 ...
##  $ GPA       : num  3.62 3.23 3.69 3.38 3.71 3.89 3.97 3.49 3.54 3.71 ...
##  $ VR        : int  11 9 12 9 8 9 11 11 12 13 ...
##  $ PS        : int  9 10 11 11 10 9 9 8 8 10 ...
##  $ WS        : int  9 5 7 4 6 6 8 4 8 8 ...
##  $ BS        : int  9 9 10 11 11 10 11 8 10 10 ...
##  $ MCAT      : int  38 33 40 35 35 34 39 31 38 41 ...
##  $ Apps      : int  5 19 5 11 5 11 6 9 6 6 ...
```

Male dataset summary

```
str(df_male)
```

```
## 'data.frame':    27 obs. of  11 variables:
##  $ Accept    : Factor w/ 2 levels "A","D": 1 1 1 2 1 2 2 2 1 2 ...
##  $ Acceptance: Factor w/ 2 levels "0","1": 2 2 2 1 2 1 1 1 2 1 ...
##  $ Sex       : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ BCPM      : num  3.75 3.59 3.85 3.26 3.77 3.6 3.29 3.75 3.51 3.27 ...
##  $ GPA       : num  3.84 3.72 3.89 3.34 3.77 3.61 3.3 3.65 3.54 3.25 ...
##  $ VR        : int  12 10 11 11 8 9 11 8 9 8 ...
##  $ PS        : int  13 9 12 11 10 9 8 8 10 9 ...
##  $ WS        : int  8 7 6 8 7 4 6 8 9 5 ...
##  $ BS        : int  12 10 11 9 10 10 7 11 11 10 ...
##  $ MCAT      : int  45 36 40 39 35 32 32 35 39 32 ...
##  $ Apps      : int  3 5 5 7 5 8 15 6 1 5 ...
```

### 2-sample t-test: Y=MCAT scores, split dataset on Sex (male or female)

```
two_sample_t_test <- t.test(df$MCAT[df$Sex=="F"],df$MCAT[df$Sex=="M"], var.equal=TRUE)
two_sample_t_test
```

```
##
##  Two Sample t-test
##
## data:  df$MCAT[df$Sex == "F"] and df$MCAT[df$Sex == "M"]
## t = 0.020173, df = 53, p-value = 0.984
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.603913  2.656823
## sample estimates:
## mean of x mean of y
##  36.28571  36.25926
```

### Model 1

```
model <- lm(MCAT ~ GPA + Sex, data=df)
```

### Y=MCAT scores, X1=GPA, X2=Sex (male or female)

```
summary(model)
```

### Summary of Model 1

```
##
## Call:
## lm(formula = MCAT ~ GPA + Sex, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.5825  -2.5260  -0.0993   2.6574   8.4228
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.4820     7.0736   0.492    0.625
## GPA           9.1695     1.9652   4.666 2.19e-05 ***
## SexM          0.4261     1.1158   0.382    0.704
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.121 on 52 degrees of freedom
## Multiple R-squared:  0.2951, Adjusted R-squared:  0.268
## F-statistic: 10.89 on 2 and 52 DF,  p-value: 0.0001125
```

**Model 2 (adding interaction term, GPA*Sex)**

```
model_2 <- lm(MCAT ~ GPA + Sex + GPA*Sex, data=df)
```

**Y=MCAT scores, X1=GPA, X2=Sex, X3=GPA*Sex**

```
summary(model_2)
```

**Summary of Model 2**

```
##
## Call:
## lm(formula = MCAT ~ GPA + Sex + GPA * Sex, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3726  -2.5536  -0.2759   2.6843   8.3184
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.2996    11.8203   0.448   0.6558
## GPA           8.6614     3.2968   2.627   0.0113 *
## SexM         -2.4089    14.7357  -0.163   0.8708
## GPA:SexM      0.7964     4.1276   0.193   0.8478
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.16 on 51 degrees of freedom
## Multiple R-squared:  0.2956, Adjusted R-squared:  0.2542
## F-statistic: 7.135 on 3 and 51 DF,  p-value: 0.0004313
```

**Predictions** predicted MCAT score for a female with a 4.0 GPA:

```
predict(model_2, data.frame(GPA=4.0, Sex='F'), se.fit=TRUE)
```

```
## $fit
##        1
## 39.94515
##
## $se.fit
## [1] 1.599415
##
## $df
## [1] 51
##
## $residual.scale
## [1] 4.159835
```

predicted MCAT for a male with a 4.0 GPA:

```
predict(model_2, data.frame(GPA=4.0, Sex='M'), se.fit=TRUE)
```

```
## $fit
##        1
## 40.72194
##
## $se.fit
## [1] 1.419227
##
## $df
## [1] 51
##
## $residual.scale
## [1] 4.159835
```