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/Depression/depressi
df = read.table(file_path, header=TRUE, sep="", dec=".")
```

Summary of dataset - Note: Y = effectiveness score

```
str(df)
```

```
## 'data.frame':    36 obs. of  5 variables:
##  $ y  : int  56 41 40 28 55 25 46 71 48 63 ...
##  $ age: int  21 23 30 19 28 23 33 67 42 33 ...
##  $ x2 : int  1 0 0 0 1 0 0 0 0 1 ...
##  $ x3 : int  0 1 1 0 0 0 1 0 1 0 ...
##  $ TRT: chr  "A" "B" "B" "B" ...
```

Transform categorical columns to as.factor data types

```
categorical_cols <- c('x2', 'x3', 'TRT')
df[categorical_cols] <- lapply(df[categorical_cols], as.factor)
```

Ensuring column data types are correct now.

```
str(df)
```

```
## 'data.frame':    36 obs. of  5 variables:
##  $ y  : int  56 41 40 28 55 25 46 71 48 63 ...
##  $ age: int  21 23 30 19 28 23 33 67 42 33 ...
##  $ x2 : Factor w/ 2 levels "0","1": 2 1 1 1 2 1 1 1 1 2 ...
##  $ x3 : Factor w/ 2 levels "0","1": 1 2 2 1 1 1 2 1 2 1 ...
##  $ TRT: Factor w/ 2 levels "A","B": 1 2 2 2 1 2 2 2 2 1 ...
```

Model 1: Y =y (the effectiveness score) and X1=age

```r
model <- lm(y ~ age, data=df)
```

Model 1 summary

```r
summary(model)
```

```
##
## Call:
## lm(formula = y ~ age, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.8916  -5.7463  -0.4105   4.7013  16.4607
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.33935    4.08258   6.207 4.65e-07 ***
## age          0.67619    0.08797   7.687 6.15e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.613 on 34 degrees of freedom
## Multiple R-squared:  0.6347, Adjusted R-squared:  0.624
## F-statistic: 59.08 on 1 and 34 DF,  p-value: 6.155e-09
```

Model 2: Y=effectiveness score, X1=age, X2=TRT (treatment), X3=age*TRT

```
model_2 <- lm(y ~ age + TRT + age*TRT, data=df)
```

Model 2 summary

```
summary(model_2)
```

```
##
## Call:
## lm(formula = y ~ age + TRT + age * TRT, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5262  -3.4552   0.3882   3.7915   7.4342
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  47.5156     4.8471   9.803 3.68e-11 ***
## age           0.3305     0.1033   3.201  0.00309 **
## TRTB        -31.5774     5.8051  -5.440 5.53e-06 ***
## age:TRTB      0.4842     0.1243   3.895  0.00047 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.973 on 32 degrees of freedom
## Multiple R-squared:  0.8533, Adjusted R-squared:  0.8395
## F-statistic: 62.04 on 3 and 32 DF,  p-value: 1.975e-13
```