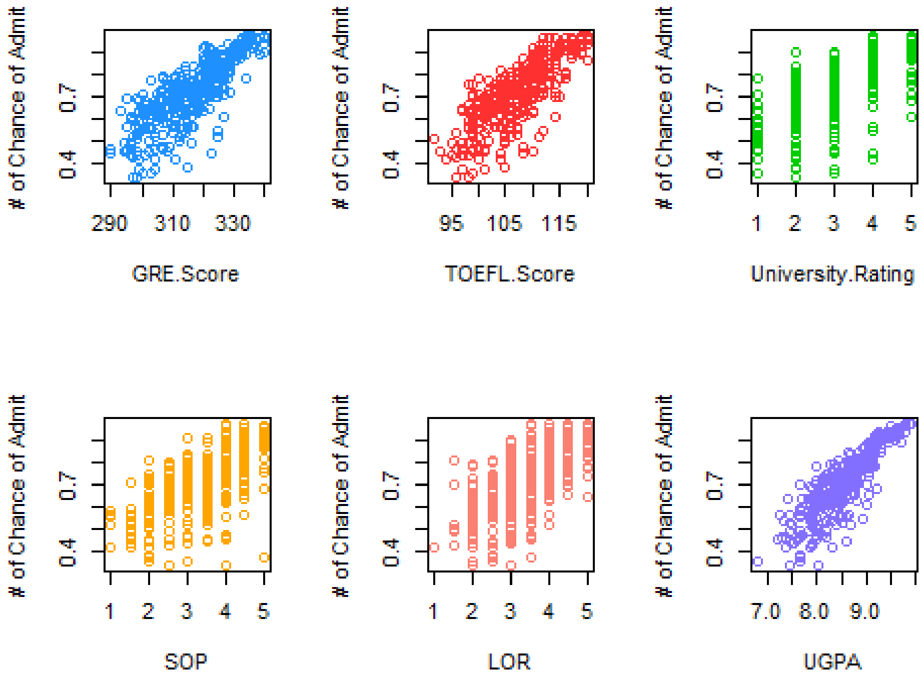
# STAT 123 Assignment 4

Eric Huber

03/04/2022

## Question 1:

```
AdmissionsPredict = read.csv("AdmissionPredict.csv")

#We can exclude "serial No." because it is just a way of indexing which is
already built in
AdmissionsPredict = AdmissionsPredict[,2:8]

y = AdmissionsPredict$Chance.of.Admit

#a
xnames = colnames(AdmissionsPredict[,1:6])


#b
par(mfrow = c(2,3))
colours = c("dodgerblue", "firebrick1", "green3", "orange", "salmon",
"slateblue1", "darkblue")
m = dim(AdmissionsPredict)[2] -1
for(i in 1:m){
  plot(AdmissionsPredict[,i], AdmissionsPredict[,7], ylab = "# of Chance of
Admit", col = colours[i], xlab = xnames[i])

  i = i+1
}
```



```
#c
cornum = numeric()
for(i in 1:m){
 cornum[i] = cor(AdmissionsPredict[,i], AdmissionsPredict[,7])
}
cornum

## [1] 0.8026105 0.7915940 0.7112503 0.6757319 0.6698888 0.8732891
```

(1.c) The explanatory variables which we are most easily able to identify the form is the GRE.Score, TOEFL.Score and UGPA. All three of which look to be positive with a linear form. The clearest and strongest one is the UPGA. This is also reflected in our correlation calculation.

(1.d)

```
full_model = lm(Chance.of.Admit ~ GRE.Score + TOEFL.Score + University.Rating
+ SOP + LOR + UGPA  , data = AdmissionsPredict)
summary(full_model)

##
## Call:
## lm(formula = Chance.of.Admit ~ GRE.Score + TOEFL.Score + University.Rating
+
##     SOP + LOR + UGPA, data = AdmissionsPredict)
##
```

```
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.279178 -0.023112  0.009864  0.035841  0.159383 
## 
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       -1.4138594  0.1154455 -12.247  < 2e-16 ***
## GRE.Score          0.0022761  0.0005779   3.938 9.70e-05 ***
## TOEFL.Score        0.0027534  0.0010999   2.503   0.0127 *  
## University.Rating  0.0060620  0.0048204   1.258   0.2093    
## SOP               -0.0019614  0.0056041  -0.350   0.7265    
## LOR                0.0227486  0.0055995   4.063 5.86e-05 ***
## UGPA               0.1198749  0.0123470   9.709  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.06447 on 393 degrees of freedom
## Multiple R-squared:  0.7987, Adjusted R-squared:  0.7956 
## F-statistic: 259.9 on 6 and 393 DF,  p-value: < 2.2e-16
```

$y = -1.4138594 + 0.0022761(x1) + 0.0027534(x2) + 0.0060620(x3) - 0.0019614(x4) + 0.0227486(x5) + 0.1198749(x6)$

(1.e)

```
summary(full_model)
```

```
## 
## Call:
## lm(formula = Chance.of.Admit ~ GRE.Score + TOEFL.Score + University.Rating 
+ 
##     SOP + LOR + UGPA, data = AdmissionsPredict)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.279178 -0.023112  0.009864  0.035841  0.159383 
## 
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       -1.4138594  0.1154455 -12.247  < 2e-16 ***
## GRE.Score          0.0022761  0.0005779   3.938 9.70e-05 ***
## TOEFL.Score        0.0027534  0.0010999   2.503   0.0127 *  
## University.Rating  0.0060620  0.0048204   1.258   0.2093    
## SOP               -0.0019614  0.0056041  -0.350   0.7265    
## LOR                0.0227486  0.0055995   4.063 5.86e-05 ***
## UGPA               0.1198749  0.0123470   9.709  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
```

```
## Residual standard error: 0.06447 on 393 degrees of freedom
## Multiple R-squared:  0.7987, Adjusted R-squared:  0.7956
## F-statistic: 259.9 on 6 and 393 DF,  p-value: < 2.2e-16
```

(1.e) No, not all terms are significant. We can see from looking at our summary() that “SOP” and “University.Rating” could be removed from our model.

(1.f)

```
new_model = lm(Chance.of.Admit ~ GRE.Score + TOEFL.Score + LOR + UGPA  , data
= AdmissionsPredict)
summary(new_model)

##
## Call:
## lm(formula = Chance.of.Admit ~ GRE.Score + TOEFL.Score + LOR +
##     UGPA, data = AdmissionsPredict)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.279714 -0.022678  0.009575  0.036309  0.160523
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.4630686  0.1057458 -13.836  < 2e-16 ***
## GRE.Score    0.0023179  0.0005761   4.023 6.88e-05 ***
## TOEFL.Score  0.0029252  0.0010761   2.718  0.00685 **
## LOR          0.0239713  0.0048405   4.952 1.09e-06 ***
## UGPA         0.1228233  0.0118475  10.367  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06443 on 395 degrees of freedom
## Multiple R-squared:  0.7979, Adjusted R-squared:  0.7959
## F-statistic: 389.9 on 4 and 395 DF,  p-value: < 2.2e-16
```

y = -1.4630686 + 0.0023179(x1) + 0.0029252(x2) + 0.0239713(x3) + 0.1228233(x4)

(1.g)

```
print("GRE.Score Range")

## [1] "GRE.Score Range"

range(AdmissionsPredict$GRE.Score)

## [1] 290 340

print("TOEFL.Score Range")

## [1] "TOEFL.Score Range"
```

```
range(AdmissionsPredict$TOEFL.Score)

## [1]  92 120

print("University.Rating Range")

## [1] "University.Rating Range"

range(AdmissionsPredict$University.Rating)

## [1] 1 5

print("SOP Range")

## [1] "SOP Range"

range(AdmissionsPredict$SOP)

## [1] 1 5

print("LOR Range")

## [1] "LOR Range"

range(AdmissionsPredict$LOR)

## [1] 1 5

print("UGPA Range")

## [1] "UGPA Range"

range(AdmissionsPredict$UGPA)

## [1] 6.80 9.92
```

(1.h)

```
y = (-1.4630686  + 0.0023179*(320) + 0.0029252*(101) + 0.0239713*(4) +
0.1228233*(8.4))*100
y

## [1] 70.17055
```

There is roughly a 70.17% chance that the student gets accepted into the graduate program.

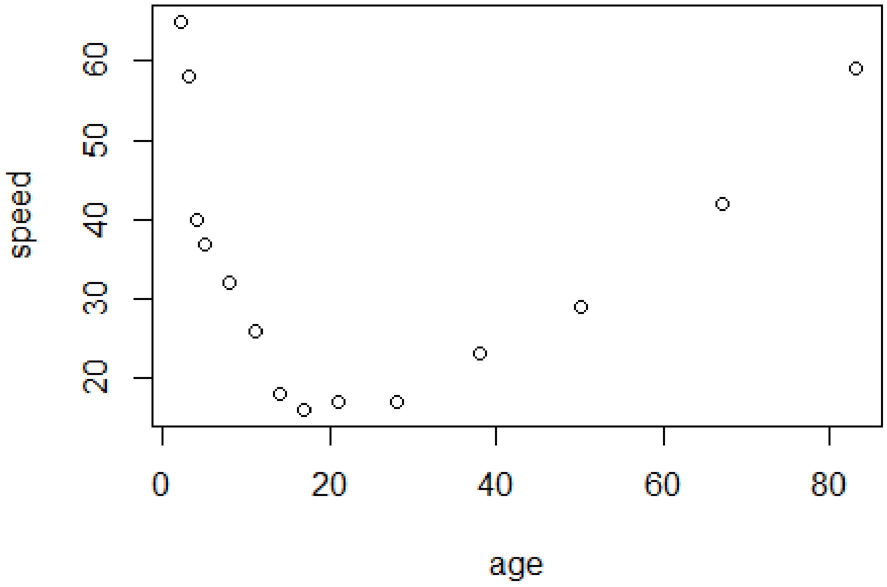
## Question 2:

```
age = c(2,3,4,5,8,11,14,17,21,28,38,50,67,83)

speed = c(65,58,40,37,32,26,18,16,17,17,23,29,42,59)
```

(2.a) Since we are trying to determine the relationship between ages (in years) and speed (in seconds). We need to use speed as the response variable and age as the explanatory variable.

(2.b)

```
plot(age,speed)
```



The form appears to be Quadratic.

(2.c)

```
age2 = age^2
age3 = age^3

cor(speed, age+age2)

## [1] 0.3225813

cor(speed, age+age2+age3)

## [1] 0.4062131
```

```
FourC_Model = lm(speed ~ age + age2 + age3
                )
summary(FourC_Model)
```

```
##
## Call:
## lm(formula = speed ~ age + age2 + age3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.070 -4.614 -1.288  4.139 11.091
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 61.4882398  4.5854592  13.409 1.02e-07 ***
## age         -3.9793958  0.6228406  -6.389 7.94e-05 ***
## age2         0.0960492  0.0191914   5.005 0.000534 ***
## age3        -0.0005889  0.0001560  -3.774 0.003636 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.718 on 10 degrees of freedom
## Multiple R-squared:  0.8761, Adjusted R-squared:  0.8389
## F-statistic: 23.57 on 3 and 10 DF,  p-value: 7.485e-05
```

y = 61.4882398 -3.9793958(x) + 0.0960492(x)^2 - 0.0005889(x)^3

(2.d)

```
y =  61.4882398 -3.9793958*(70) + 0.0960492*(70)^2 - 0.0005889*(70)^3
y
```

```
## [1] 51.57891
```

It would take a 70 year old approx ~51.58 seconds to run that distance.

(2.e) Roughly 87.6% of the response variable variation can be explained by the explanatory variable in my model.