STAT 123 - Homework 4

Parker DeBruyne - V00837207

31/03/2022

Due Friday April 8th by 9:00pm

1. Download and save the AdmissionsPredict.csv dataset and read it into R. This data set consists of data regarding international students who are applying for graduate programs in English speaking countries. The data set contains 7 variables: • GRE (Graduate Record Examination score) • TOEFL (Test of English as a Foreign Language score) • University Rating (score out of 5) • SOP (Statement of Purpose, score out of 5) • LOR (Letter of Recommendation, score out of 5) • UGPA (Undegraduate Grade Point Average, score out of 10) • Chance of Admit (value between 0 and 1)

```
df = read.csv("AdmissionPredict.csv")
col_names = colnames(df)
```

   (a) The response variable is y = Chance of Admit. All other variables are possible explana- tory variables. Create a vector called xnames which contains the names of each of the explanatory variables.

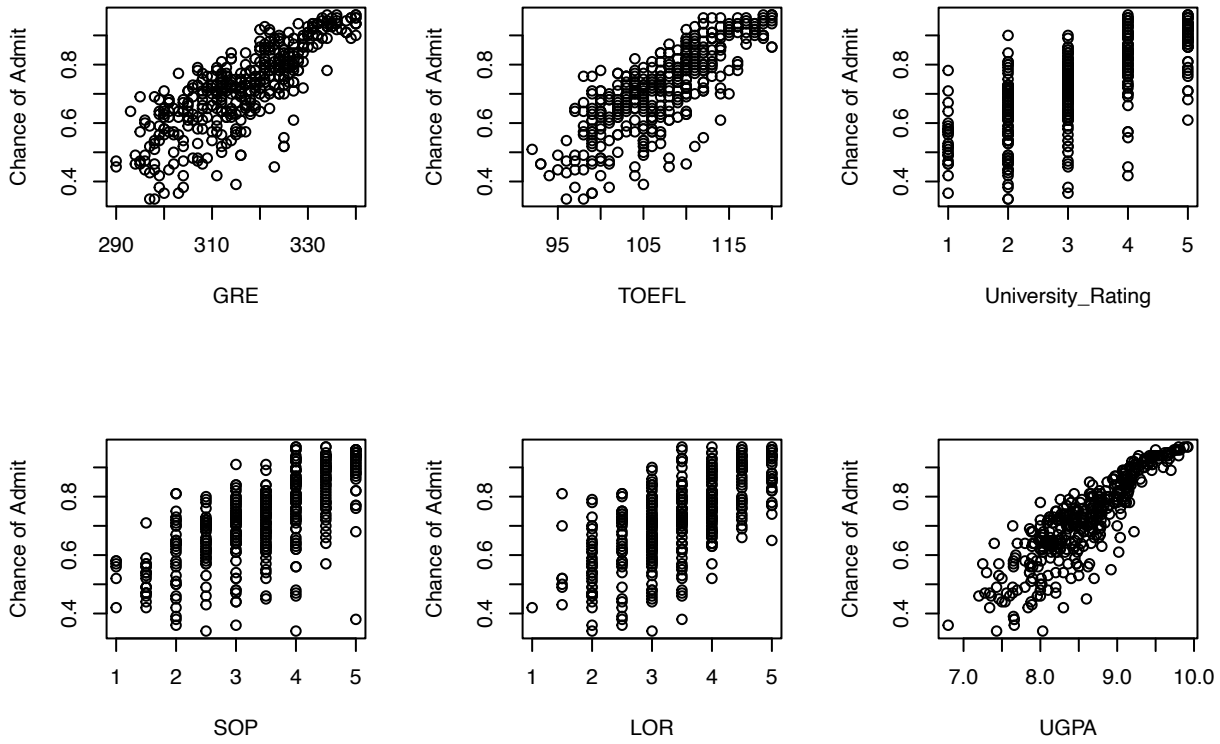
```
df2 = df %>% rename(GRE=col_names[2],
                    TOEFL=col_names[3],
                    University_Rating=col_names[4],
                    SOP=col_names[5],
                    LOR=col_names[6])
xnames = colnames(df2)

y = df2$Chance.of.Admit
x1 = df2$GRE
x2 = df2$TOEFL
x3 = df2$University_Rating
x4 = df2$SOP
x5 = df2$LOR
x6 = df2$UGPA
```

   (b) Use the command par(mfrow = c(2,3)) and then write a for-loop which plots each of the explanatory variables against the response variable. Make sure each plot has an x-axis title and a y-axis title. Use xnames from part (a) to create the x-axis title. Hint: You did a very similar question in lab with Steve.

```
par(mfrow = c(2,3))

for (i in 2:(length(xnames)-1)){
  plot(df2[[i]], df2[[8]],
       xlab = xnames[i],
       ylab = "Chance of Admit")
}
```



(c) For which explanatory variables are you able to identify the form of the relationship with y? What is the form that you see?

```
#Ans: GRE, TOEFL, and UGPA. It appears to be a linear relationship.
```

(d) Create a linear regression model called full model which includes all of the possible explanatory variables. Write out the model that you obtain. Example: $y = 0.3 + 0.1(x1) - 0.5(x2) + \ldots$

```
options(scipen=999)
full_model = lm(y~x1 + x2 + x3 + x4 + x5 + x6)
summary(full_model)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6)
##
## 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 < 0.0000000000000002 ***
## x1           0.0022761  0.0005779   3.938            0.0000970 ***
## x2           0.0027534  0.0010999   2.503               0.0127 *
## x3           0.0060620  0.0048204   1.258               0.2093
## x4          -0.0019614  0.0056041  -0.350               0.7265
## x5           0.0227486  0.0055995   4.063            0.0000586 ***
## x6           0.1198749  0.0123470   9.709 < 0.0000000000000002 ***
## ---
## 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: < 0.00000000000000022
```

```
# Following anser contains significant and insignificant variables.
#Ans: y = -1.4138594 + 0.0022761(x1) + 0.0027534(x2) + 0.0060620(x3) - 0.0019614(x4) + 0.0227486(x5)
```

(e) Are all terms significant? Identify any variables which should be removed from the model (and show the code that you are using to make this decision).

```
# No, not all of them are. x3 and x4 should be removed.
sig_vars_index = summary(full_model)$coefficients[,4] <= 0.05
sig_vars = summary(full_model)$coefficients[sig_vars_index,4]
sig_vars
```

```
##                             (Intercept)                                      x1
## 0.00000000000000000000000000001962317 0.00009701383931434515780090760417309
##                                      x2                                      x5
## 0.01270649220590770733796404101667576 0.00005861371781203874063237782565672
##                                      x6
## 0.00000000000000000004069244400445243
```

```
print(paste(names(sig_vars), "Was found to be significant"))
```

```
## [1] "(Intercept) Was found to be significant"
## [2] "x1 Was found to be significant"
## [3] "x2 Was found to be significant"
## [4] "x5 Was found to be significant"
## [5] "x6 Was found to be significant"
```

(f) Create a new model called new model which contains only the terms which were significant in the full model. Write out the model that you obtain.

```
new_model = -1.4138594 + 0.0022761*(x1) + 0.0027534*(x2) + 0.0227486*(x5) + 0.1198749*(x6)

# ans: y = -1.4138594 + 0.0022761*(x1) + 0.0027534*(x2) + 0.0227486*(x5) + 0.1198749*(x6)
```

(g) What is the range of values for each variable included in the new model that we can use for prediction (so that we avoid extrapolation).

```
# for (i in 2:7){
#   min = min(df2[[i]])
#   max = max(df2[[i]])
#
#   print(paste("Range of values for", xnames[i], "is", min(df2[[i]]), "to", max(df2[[i]])))
# }

for (i in c(2,3,6,7)){
  min = min(df2[[i]])
  max = max(df2[[i]])

  print(paste("Range of values for", xnames[i], "is", min(df2[[i]]), "to", max(df2[[i]])))
}
```

```
## [1] "Range of values for GRE is 290 to 340"
## [1] "Range of values for TOEFL is 92 to 120"
## [1] "Range of values for LOR is 1 to 5"
## [1] "Range of values for UGPA is 6.8 to 9.92"
```

(h) Consider a student with a GRE score of 320, a TOEFL score of 101, applying to a University with a rating of 4, with a SOP score of 3, a LOR score of 4 and an under-graduate GPA of 8.4. Use your model from part (f) to predict this students chance of being accepted to the graduate program at their University of choice.

```
s_GRE = 320
s_TOEFL = 101
s_LOR = 4
s_UGPA = 8.4

s_predict = -1.4138594 + 0.0022761*(s_GRE) + 0.0027534*(s_TOEFL) + 0.0227486*(s_LOR) + 0.1198749*(s_
print(paste(round((s_predict*100),2), "%"))
```

```
## [1] "69.05 %"
```

2. Type in the following vectors which represent people of various ages (in years) who are each timed (in seconds) running the same distance.

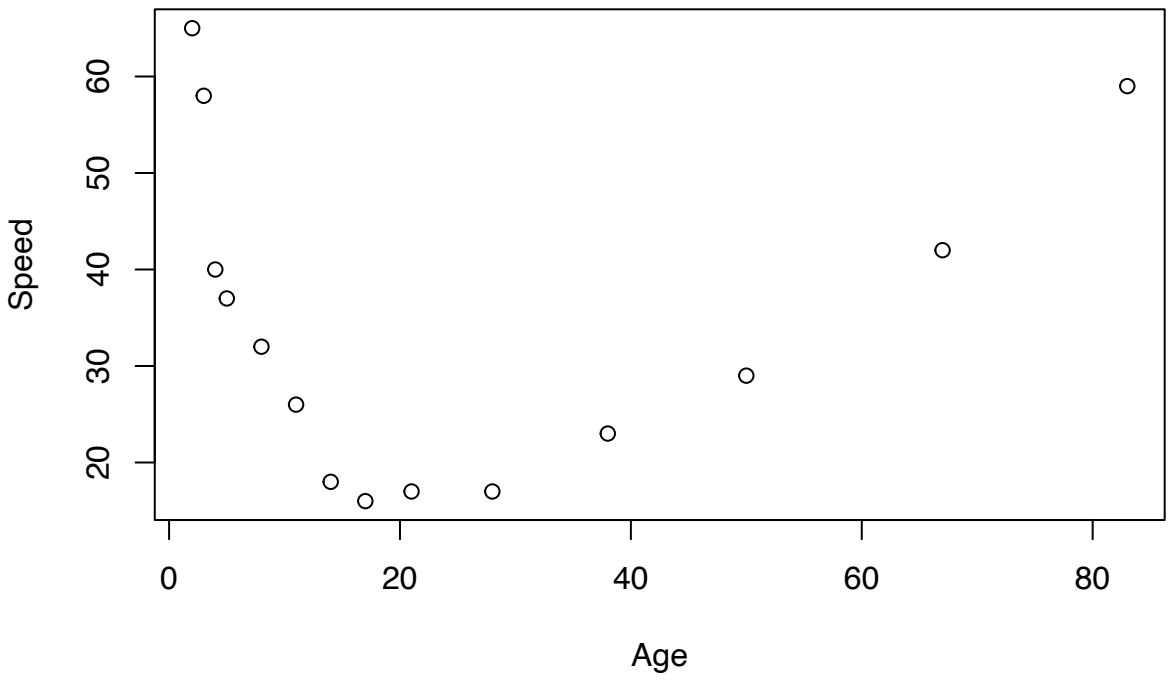
```
age = c(2,3,4,5,8,11,14,17,21,28,38,50,67,83)
speed = c(65,58,40,37,32,26,18,16,17,17,23,29,42,59)
```

(a) Which is the response variable and which is the explanatory variable?

```
#Ans: Age is the explanetory variable, Speed is the response variable.
```

(b) Plot the variables (you do not need any titles). What form does the relationship seem to have?

```
plot(age,speed, xlab="Age", ylab="Speed")
```




```
# The relationship appears to be quadratic.
```

(c) Fit a model to the form that you identified in part (b). Write out the model that you obtain.

```
xsq = age * age
quad_model = lm(speed~age + xsq)
summary = summary(quad_model)
```

```
p_vals = summary$coefficients[,4]
coefs = summary$coefficients[,1]

# AvS_model = 50.32303289 - 1.85810455*(age) + 0.02470455*(xsq)
```

(d) Use your model to predict how long it would take for a 70 year old to run that distance.

```
run_length_70 = 50.32303289 - 1.85810455*(70) + 0.02470455*(70*70)
print(paste(run_length_70, "seconds"))
```

```
## [1] "41.30800939 seconds"
```

(e) What percentage of the variation in the response variable can be explained by the variation in the explanatory variables in the model? Page 2

```
rsq = round(((summary$r.squared)*100),2)
print(paste(rsq, "%"))
```

```
## [1] "69.96 %"
```