ASM_Homework_Ex3

```r
#Loading necessary packages
library(car)
library(pastecs)
#library(kableExtra)
#library(tinytex)
library(carData)
```

```
##
## Attaching package: 'carData'
```

```
## The following objects are masked from 'package:car':
##
##     Adler, Blackmore, Guyer, UN, Vocab
```

```r
#Loading the data
data <- read.csv("FIFADataset.csv", header = TRUE)

#Filtering the 3 clubs of interest and the necessary variables for the analysis
fc <- data[ data$Club %in% c('Manchester City','Real Madrid CF', 'FC Barcelona'),
            c('Name', 'Agility', 'Acceleration', 'Club', 'Sprint.speed', 'Reactions', 'Balance') ]

#Showing first rows of the data
head(fc)
```

```
##                  Name Agility Acceleration            Club Sprint.speed
## 1  Cristiano Ronaldo      89           89  Real Madrid CF           91
## 2           L. Messi      90           92    FC Barcelona           87
## 4          L. Suárez      86           88    FC Barcelona           77
## 9           T. Kroos      71           60  Real Madrid CF           52
## 11      Sergio Ramos      79           75  Real Madrid CF           77
## 12      K. De Bruyne      80           76 Manchester City           75
##    Reactions Balance
## 1         96      63
## 2         95      95
## 4         93      60
## 9         86      69
## 11        85      60
## 12        88      75
```

Description of the Dataset

As it can be seen in the table the data consists of different football players from the FIFA Dataset obtained by Kaggle. This analysis covers only players from the three clubs “Manchester City”, “FC Barcelona” and “Real Madrid CF”. The following variables are relevant:

- Name: factor, contains the name of the player
- Club: factor, contains name of the club to which a player belongs
- Acceleration: numeric, theoretical range: 1-100
- Agility: numeric, theoretical range: 1-100
- Sprint.speed: numeric, theoretical range: 1-100
- Reactions: numeric, theoretical range: 1-100

- Balance: numeric, theoretical range: 1-100

A more in detail discussion of the variables can be find in the Univariate Descriptive Analysis section.

Data Cleansing

```
#checking levels
sapply(fc, class)
```

```
##         Name      Agility Acceleration         Club Sprint.speed 
##     "factor"     "factor"     "factor"     "factor"     "factor" 
##    Reactions      Balance 
##     "factor"     "factor"
```

```
#looking at the different variables which should be numerical
levels(fc$Acceleration)[fc$Acceleration]
```

```
##  [1] "89"    "92"    "88"    "60"    "75"    "76"    "75"    "93"   
##  [9] "90"    "77"    "67"    "72"    "55"    "72"    "74"    "48"   
## [17] "77"    "77"    "63"    "54"    "38"    "93"    "73"    "61"   
## [25] "79"    "85"    "76"    "93"    "64"    "72"    "62"    "86"   
## [33] "67"    "61"    "93"    "71"    "83"    "85"    "94"    "93"   
## [41] "58"    "75"    "73"    "89"    "91"    "74"    "72"    "54"   
## [49] "76"    "76"    "80"    "32"    "71"    "79"    "72"    "72"   
## [57] "79"    "75"    "69"    "75"    "57"    "76"    "69"    "89"   
## [65] "82"    "72"    "75"    "76"    "49"    "64"    "76"    "78"   
## [73] "78"    "64"    "67"    "33+10" "52"    "44"    "70"    "68"   
## [81] "68"    "63"    "72"    "57"
```

```
levels(fc$Agility)[fc$Agility]
```

```
##  [1] "89"   "90"   "86"   "71"   "79"   "80"   "93"   "77"   "86"   "77"  
## [11] "72"   "92"   "58"   "79"   "83"   "58"   "77"   "58"   "66"   "60"  
## [21] "37"   "90"   "83"   "60"   "78"   "88"   "82"   "90"   "60"   "68"  
## [31] "69"   "70"   "72"   "58"   "89"   "67"   "82"   "77"   "91"   "93"  
## [41] "64"   "82"   "77"   "93"   "88"   "80"   "60"   "48"   "69"   "70"  
## [51] "83"   "43"   "64"   "71"   "76"   "72"   "68"   "73"   "60"   "73"  
## [61] "55"   "89"   "69"   "78"   "75"   "74"   "80"   "75"   "54"   "65"  
## [71] "79"   "79"   "83"   "71"   "71"   "36+2" "58"   "38"   "71"   "63"  
## [81] "65"   "54"   "60"   "44"
```

```
levels(fc$Sprint.speed)[fc$Sprint.speed]
```

```
##  [1] "91"   "87"   "77"   "52"   "77"   "75"   "71"   "95"   "84"   "79"  
## [11] "64"   "65"   "73"   "71"   "69"   "34"   "78"   "81"   "61"   "53"  
## [21] "50"   "93"   "70"   "64"   "77"   "78"   "86"   "89"   "63"   "75"  
## [31] "64"   "93"   "64"   "64"   "94"   "66"   "79"   "77"   "92"   "93"  
## [41] "54"   "68"   "73"   "85"   "92"   "74"   "80"   "50"   "78"   "76"  
## [51] "79"   "49"   "71"   "81"   "73"   "65"   "84"   "80"   "67"   "80"  
## [61] "54"   "72"   "77"   "93"   "85"   "75"   "70"   "78"   "52"   "61"  
## [71] "76"   "75"   "70"   "66"   "72"   "32-1" "58"   "46"   "70"   "67"  
## [81] "66"   "59"   "68"   "54"
```

```
levels(fc$Reactions)[fc$Reactions]
```

```
##  [1] "96" "95" "93" "86" "85" "88" "88" "87" "89" "87" "79" "84" "84" "88"
## [15] "79" "82" "83" "81" "83" "82" "82" "82" "87" "84" "78" "84" "81" "82"
## [29] "86" "77" "78" "81" "83" "82" "79" "79" "85" "68" "79" "78" "77" "80"
## [43] "82" "82" "82" "76" "72" "78" "78" "81" "82" "74" "75" "76" "86" "78"
## [57] "76" "79" "70" "79" "74" "72" "75" "78" "64" "75" "70" "63" "62" "62"
## [71] "59" "54" "62" "60" "64" "55" "54" "59" "60" "56" "53" "51" "44" "45"
```

```
levels(fc$Balance)[fc$Balance]
```

```
##  [1] "63"   "95"   "60"   "69"   "60"   "75"   "94"   "65"   "91"   "82"  
## [11] "66"   "89"   "42"   "84"   "85"   "55"   "60"   "47"   "57"   "61"  
## [21] "43"   "86"   "81"   "42"   "76"   "86"   "79"   "85"   "40"   "66"  
## [31] "62"   "72"   "78"   "56"   "82"   "64"   "78"   "70"   "92"   "79"  
## [41] "64"   "76"   "79"   "92"   "83"   "81"   "68"   "50"   "74"   "64"  
## [51] "76"   "58"   "57"   "80"   "78"   "73"   "71"   "65"   "57"   "65"  
## [61] "66"   "86"   "69"   "77"   "61"   "66"   "80"   "74"   "62"   "75"  
## [71] "83"   "82"   "79"   "63"   "69"   "49+2" "52"   "40"   "62"   "63"  
## [81] "70"   "69"   "64"   "64"
```

```
#truncate at the first two digits and convert to numerical
fc[, 'Acceleration'] <-as.numeric(substr(levels(fc$Acceleration)[fc$Acceleration], 0, 2))
fc[,'Agility'] <- as.numeric(substr(levels(fc$Agility)[fc$Agility], 0, 2))
fc[, 'Sprint.speed'] <-as.numeric(substr(levels(fc$Sprint.speed)[fc$Sprint.speed], 0, 2))
fc[, 'Reactions'] <-as.numeric(substr(levels(fc$Reactions)[fc$Reactions], 0, 2))
fc[, 'Balance'] <-as.numeric(substr(levels(fc$Balance)[fc$Balance], 0, 2))

#checking levels again
sapply(fc, class)
```

```
##         Name      Agility Acceleration         Club Sprint.speed 
##     "factor"    "numeric"    "numeric"     "factor"    "numeric" 
##    Reactions      Balance 
##    "numeric"    "numeric"
```

After taking a first look at the data, we see that some values do not consist only out of one number but of some additional numbers which should be truncated, so that all rows of the specific columns can be numeric.

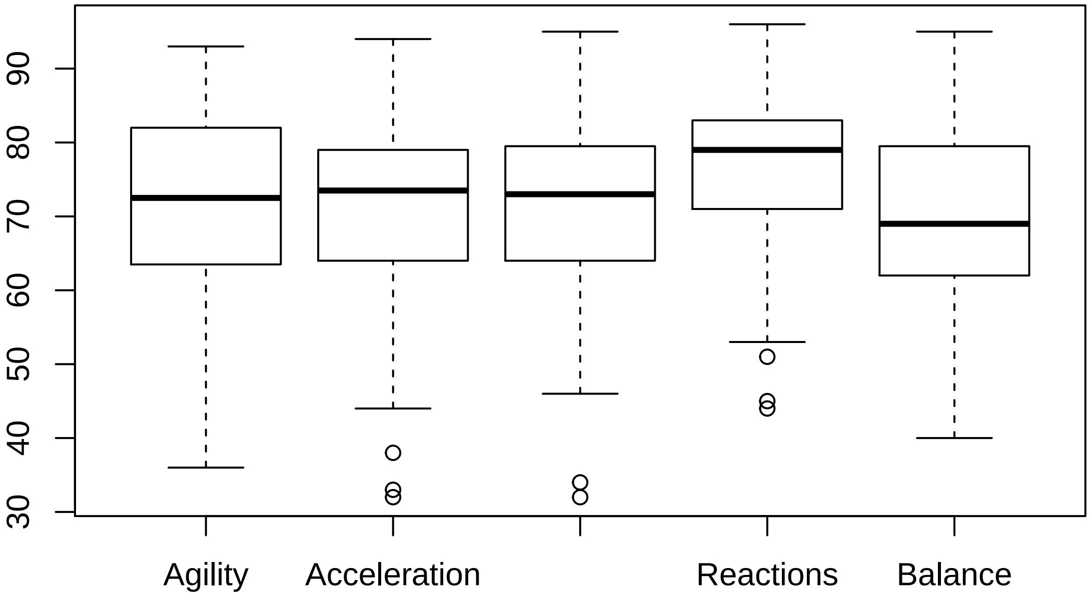
Univariate Descriptive Analysis and Outlier Detection

```
round(stat.desc(fc[,c('Acceleration', 'Agility', 'Sprint.speed', 'Balance', 'Reactions')], basic = T, de
```

```
##              Acceleration Agility Sprint.speed Balance Reactions
## nbr.val             84.00   84.00        84.00   84.00     84.00
## nbr.null             0.00    0.00         0.00    0.00      0.00
## nbr.na               0.00    0.00         0.00    0.00      0.00
## min                 32.00   36.00        32.00   40.00     44.00
## max                 94.00   93.00        95.00   95.00     96.00
## range               62.00   57.00        63.00   55.00     52.00
## sum               6044.00 6035.00      6023.00 5853.00   6368.00
## median              73.50   72.50        73.00   69.00     79.00
## mean                71.95   71.85        71.70   69.68     75.81
## SE.mean              1.48    1.48         1.46    1.43      1.24
## CI.mean.0.95         2.95    2.95         2.90    2.85      2.46
## var                184.41  184.73       178.14  172.87    129.00
## std.dev             13.58   13.59        13.35   13.15     11.36
## coef.var              0.19    0.19         0.19    0.19      0.15
```

By looking at the statistics of the numeric variables we see that there are no missing values (neither NA nor NULL) in the data. All variables are more or less in the same range. Only ‘Balance’ and ‘Reactions’ have a slightly higher Minimum. Also the means are all around 71 where ‘Balance’ has the lowest mean (of 69.68) and ‘Reactions’ the highest mean (of 75.81). The variance values among the variables are similar with the exception of ‘Reactions’ which has a much smaller variance of 129.

```
boxplot(fc[, c(2, 3, 5, 6, 7)])
```



It can be seen that there are some outliers in the Acceleration, Sprint.speed and Reactions variables. For futher analysis we will have a closer look on the instances that are smaller than Q1 - 1.5*IQR.

```
fc[fc$Acceleration < 40, ]
```

```
##                  Name Agility Acceleration           Club Sprint.speed
## 79    M. ter Stegen      37           38    FC Barcelona           50
## 468    Kiko Casilla      43           32 Real Madrid CF           49
## 11633     L. Zidane      36           33 Real Madrid CF           32
##       Reactions Balance
## 79           82      43
## 468          74      58
## 11633        55      49
```

```
fc[fc$Sprint.speed < 40, ]
```

```
##                    Name Agility Acceleration           Club Sprint.speed
## 54    Sergio Busquets      58           48    FC Barcelona           34
## 11633       L. Zidane      36           33 Real Madrid CF           32
##       Reactions Balance
## 54           82      55
## 11633        55      49
```

```
fc[fc$Reactions < 52, ]
```

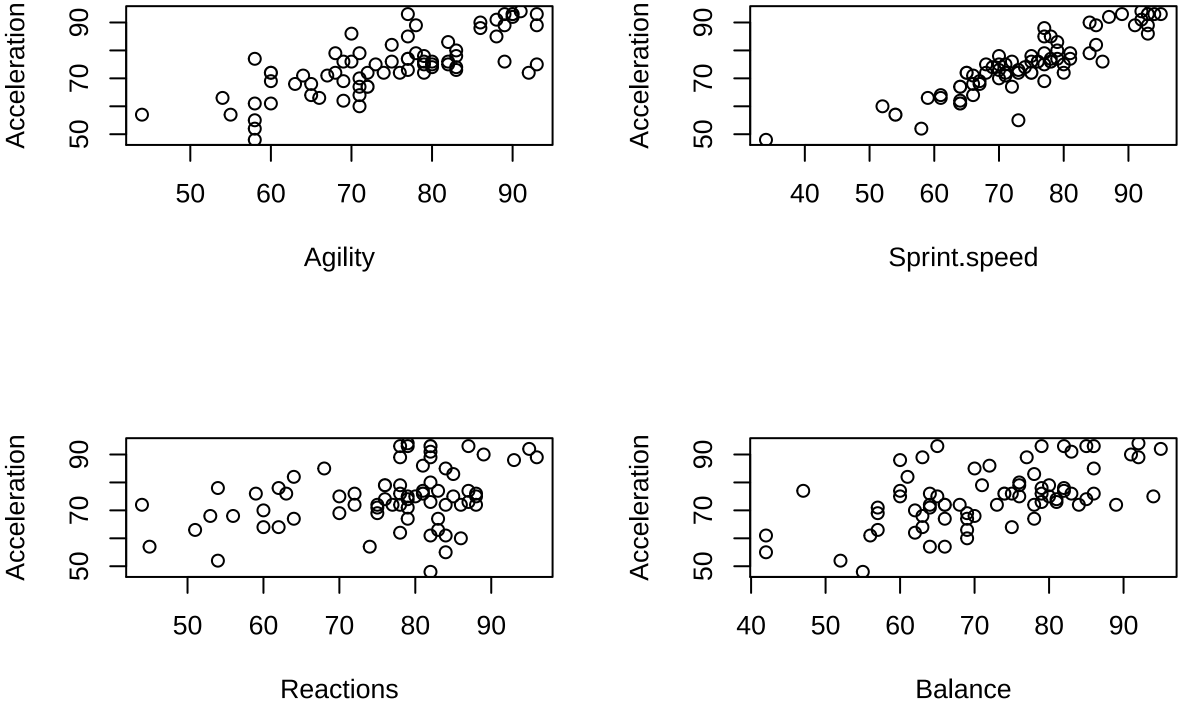
```
##                     Name Agility Acceleration            Club Sprint.speed
## 16599          M. Wood      54           63 Manchester City           59
## 17347   Erik Sarmiento      60           72 Manchester City           68
## 17962 J. Latibeaudiere      44           57 Manchester City           54
##       Reactions Balance
## 16599        51      69
## 17347        44      64
## 17962        45      64
```

The outliers in the Acceleration variable are all goalkeepers. Being a goalkeeper requests completely different abilities than other playing positions as goalkeepers do not have to run through the entire field. As they are a very special case in the Fifa Dataset it is reasonable to exclude all goalkeepers from the analysis. They play under completely different conditions than the other players. However, other players which have very low values e.g. in Sprint.speed or Reactions are not goalkeepers and although they have values which are much lower than the average ones they should remain in the dataset as they play under equal conditions as the other players.

```r
pos <- data[data$Club %in% c('Manchester City','Real Madrid CF', 'FC Barcelona'),
            c('Name', 'Preferred.Positions') ]
#merge the positions to the dataframe so that it is possible to exclude goalkeepers
fc_pos <- merge(pos, fc, by="Name")

#noticed that one player (Danilo from Man City) appeared twice in the original dataset -> remove duplic
fc_pos <- fc_pos[!duplicated(fc_pos$Name),]

#exclude goalkeepers and remove position variable
fc <- fc_pos[!(fc_pos$Preferred.Positions %in% 'GK '),]
fc <- fc[,c(1, 3, 4, 5, 6, 7, 8)]

#Reindex
rownames(fc) <- 1:nrow(fc)
```

Bivariate Descriptive Anaylsis

```r
# Scatter plots
par(mfrow = c(2,2))
plot(fc$Acceleration ~ fc$Agility, ylab = 'Acceleration', xlab='Agility')
plot(fc$Acceleration ~ fc$Sprint.speed, ylab = 'Acceleration', xlab='Sprint.speed')
plot(fc$Acceleration ~ fc$Reactions, ylab = 'Acceleration', xlab='Reactions')
plot(fc$Acceleration ~ fc$Balance, ylab = 'Acceleration', xlab='Balance')
```



```
cor(fc[, 'Acceleration'], fc[, 'Agility'])
```

```
## [1] 0.7659992
```

```
cor(fc[, 'Acceleration'], fc[, 'Sprint.speed'])
```

```
## [1] 0.8978477
```

```
cor(fc[, 'Acceleration'], fc[, 'Reactions'])
```

```
## [1] 0.356397
```

```
cor(fc[, 'Acceleration'], fc[, 'Balance'])
```

```
## [1] 0.5736713
```

In the scatter plots you can see that some variables are more correlated with Acceleration than others. Agility (cor=0.77) and Sprint.speed (cor=0.90) are highly correlated with Acceleration. On the other side the relation in the scatter plots for Reactions (cor=0.36) and Balance (cor=0.57) with Acceleration are not really good to see and the correlation values are low. On a bivariate level therefore Agility and Sprint.Speed should be better predictors for Acceleration than Reactions and Balance.

Hypotheses

After the uni and bivariate analysis of the variables the following hypotheses can be stated:

- The higher the Agility of a player, the higher the Acceleration.
- The higher the Sprint.Speed of a player, the higher the Acceleration.
- The higher the Reactions of a player, the higher the Acceleration.
- The higher the Balance of a player, the higher the Acceleration.
- Agility and Sprint.Speed have a higher influence than Reactions and Balance.

Linear Model

To check the hypotheses a multivariate linear regression will be performed.

```
mod <- lm(Acceleration ~ Agility+Sprint.speed+Reactions+Balance, fc)
summary(mod)
```

```
##
## Call:
## lm(formula = Acceleration ~ Agility + Sprint.speed + Reactions +
##     Balance, data = fc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.8156  -2.1325   0.1261   1.9745   9.5391
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.06239    3.74481   0.551 0.583595
## Agility       0.27917    0.07503   3.721 0.000401 ***
## Sprint.speed  0.64800    0.04634  13.983  < 2e-16 ***
## Reactions    -0.02584    0.04392  -0.588 0.558278
## Balance       0.08161    0.05617   1.453 0.150749
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.572 on 69 degrees of freedom
## Multiple R-squared:  0.8923, Adjusted R-squared:  0.8861
## F-statistic: 142.9 on 4 and 69 DF,  p-value: < 2.2e-16
```

From the multivariate model we can observe that 50% of the **residuals** are in the range between -2.1325 and 1.9745, being this a low error range, BUT, we got a residual with an error of nearly -12, for that player the model over-predicted its agility atribute. In the other hand, we have a residual that under-predicted the agility value by aproximately 9.5391. With the values of the **coefficients** and their **p-values** we can say that Agility and Sprint.speed are the variables with high influence in the response variable. Because we are in a multivariate case, we see (small) differences in the values of **multiple R-squared** and **adjusted R-squared** being the second the one that is valuable to us because penalizes the complexity of the model (number of coefficients). If we re-do the model taking off the non significant variables, maybe we will see an increase in this statistic. The **residual standard error** is 3.572, different than 0, meaning that there's uncertainty in the model. Also we observe that the model have 69 degrees of freedom (n-p-1 = 74-4-1). Lastly, we have the **F-Statistic** that indicates the relationship between the response and the predictor variables and indicates which variables are good predictors. In the following ANOVA table we can see te relationship in detail.

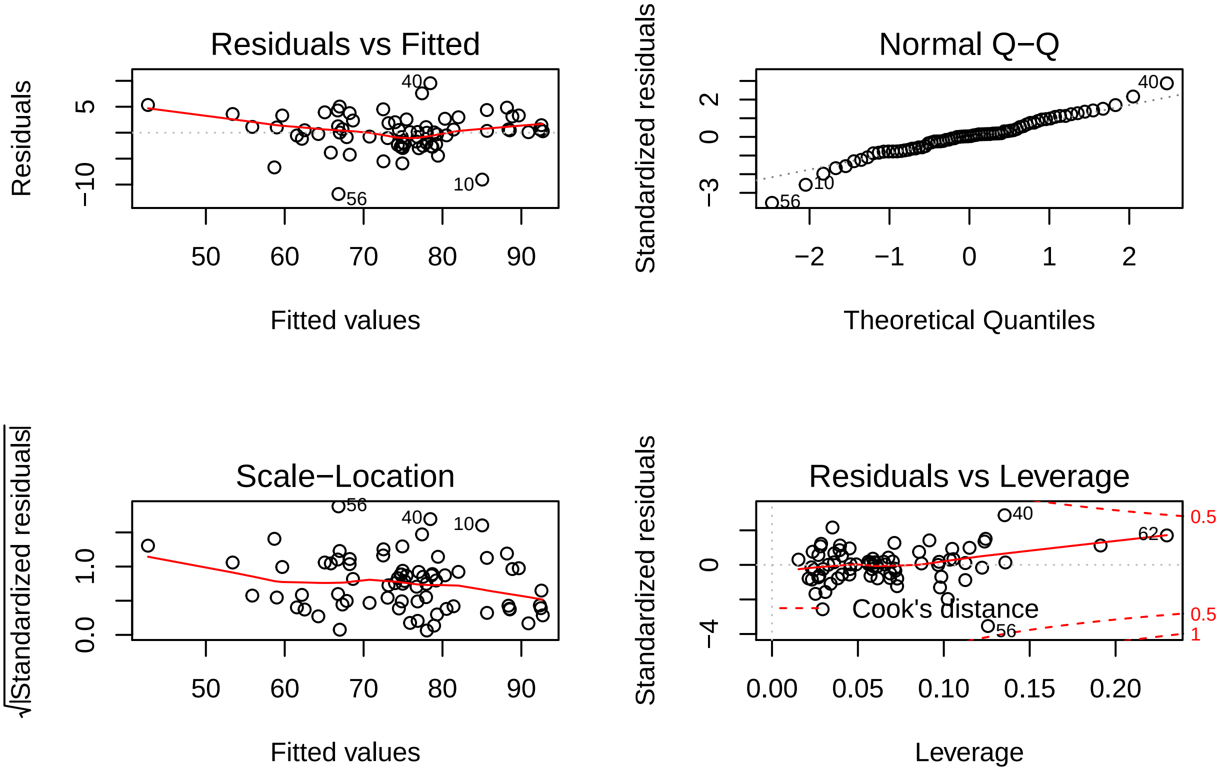
```
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: Acceleration
##              Df Sum Sq Mean Sq  F value Pr(>F)
## Agility       1 4796.3  4796.3 375.9571 <2e-16 ***
## Sprint.speed  1 2455.7  2455.7 192.4871 <2e-16 ***
## Reactions     1   15.1    15.1   1.1842 0.2803
## Balance       1   26.9    26.9   2.1113 0.1507
## Residuals    69  880.3    12.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again we see that Agility and Sprint.speed are the ones that explain better the variance of Acceleration.

Linear Model hypothesis

```
par(mfrow=c(2,2))
plot(mod)
```



We observe that the residuals are not normally distributed, thus indicates that our model is not capable of handle all the uncertaintly in the data with the variables used. The observations 56, 40 and 10 cannot be predicted well with the model and that means that we have heavy tails in the distribution of the residuals.

Using Agility and Sprint.speed as predictors

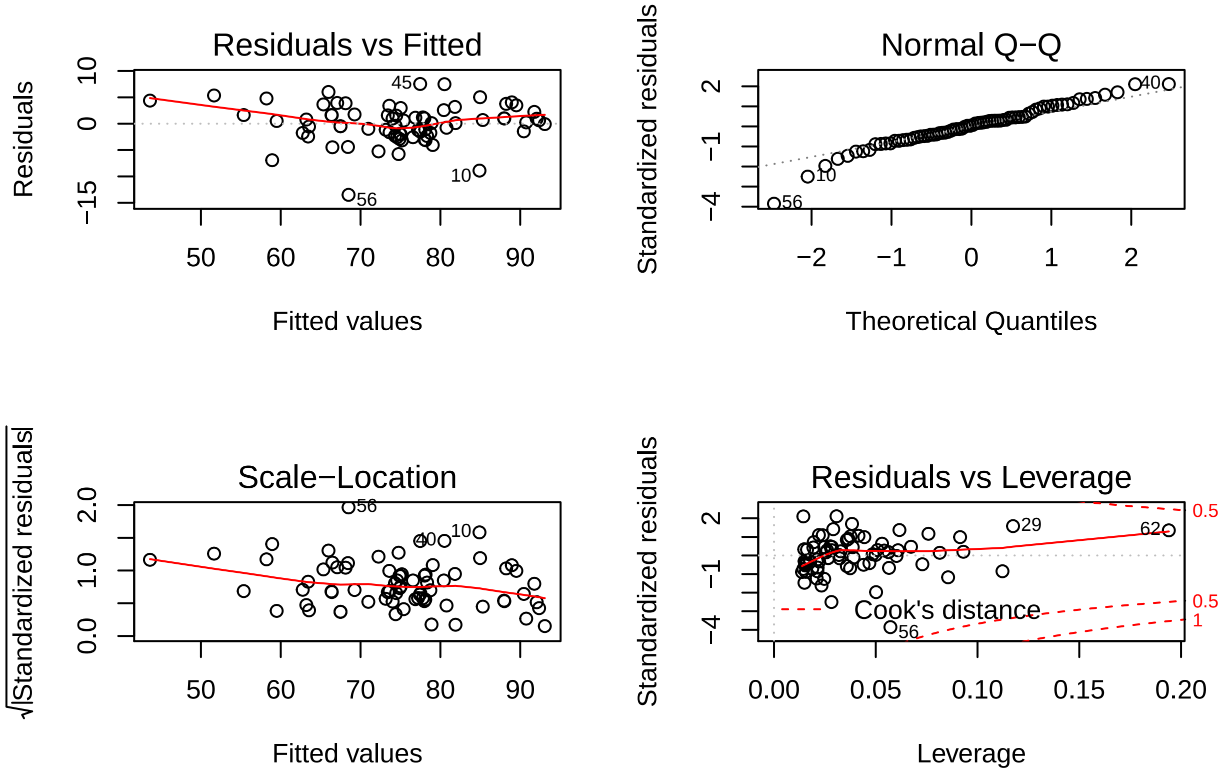
```
#checking without insignificant variables from first model
mod2 <- lm(Acceleration ~ Agility+Sprint.speed, fc)
summary(mod2)
```

```
##
## Call:
## lm(formula = Acceleration ~ Agility + Sprint.speed, data = fc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.4962  -1.8757   0.1775   1.7247   7.5324
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.33172    3.15096   0.740    0.462
## Agility         0.33788    0.04732   7.141 6.47e-10 ***
## Sprint.speed    0.63791    0.04640  13.749  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.604 on 71 degrees of freedom
## Multiple R-squared:  0.8872, Adjusted R-squared:  0.884
## F-statistic: 279.1 on 2 and 71 DF,  p-value: < 2.2e-16
```

Unexpectedly we got a similar **adjusted R-squared** as the first model (a little smaller value!). Meaning that somehow the not so important variables (Reactions and Balance) has influence in the response variable.

```
par(mfrow=c(2,2))
plot(mod2)
```



Also this model have similar problems with observations 56, 40 and 10 where their residual values are far from the expected value.

```
fc[c(56,40,10),]
```

```
##          Name Agility Acceleration           Club Sprint.speed Reactions
## 56      Piqué      58           55   FC Barcelona           73        84
## 40 L. Suárez      86           88   FC Barcelona           77        93
## 10   Carvajal      82           76 Real Madrid CF           86        81
##    Balance
## 56      42
## 40      60
## 10      79
```

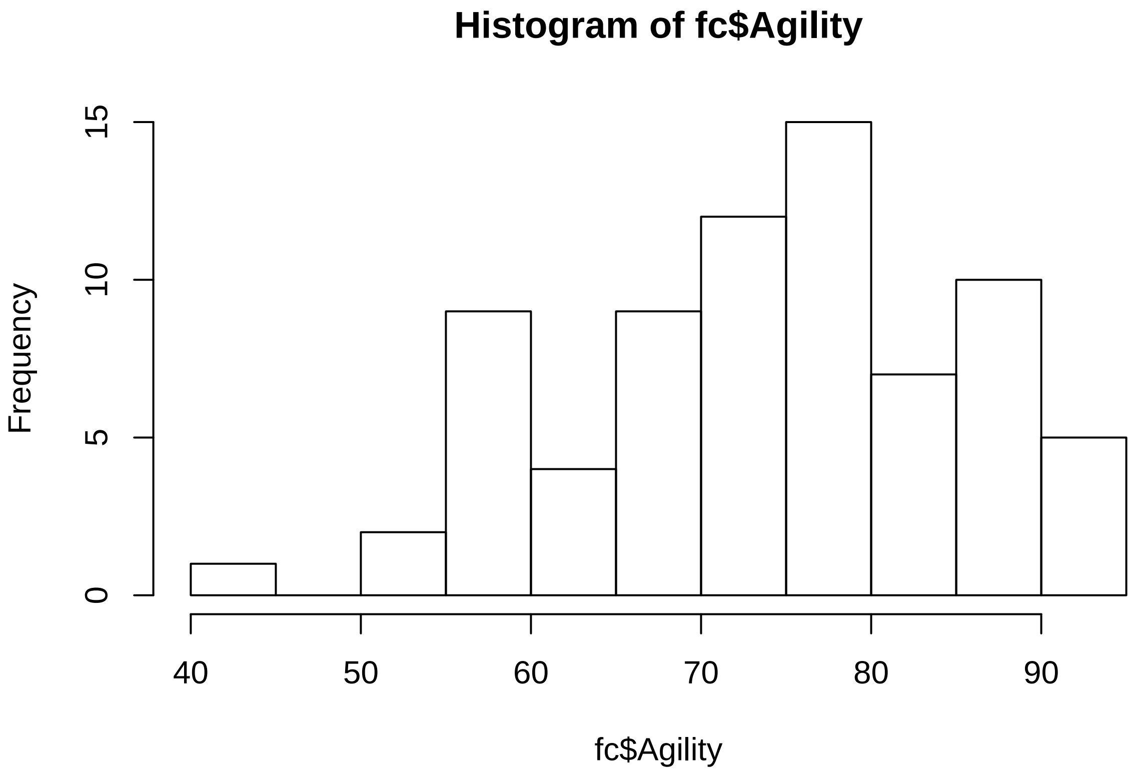
```
mod$fitted.values[c(56,40,10)]
```

```
##       56       40       10
## 66.81557 78.46091 85.03691
```

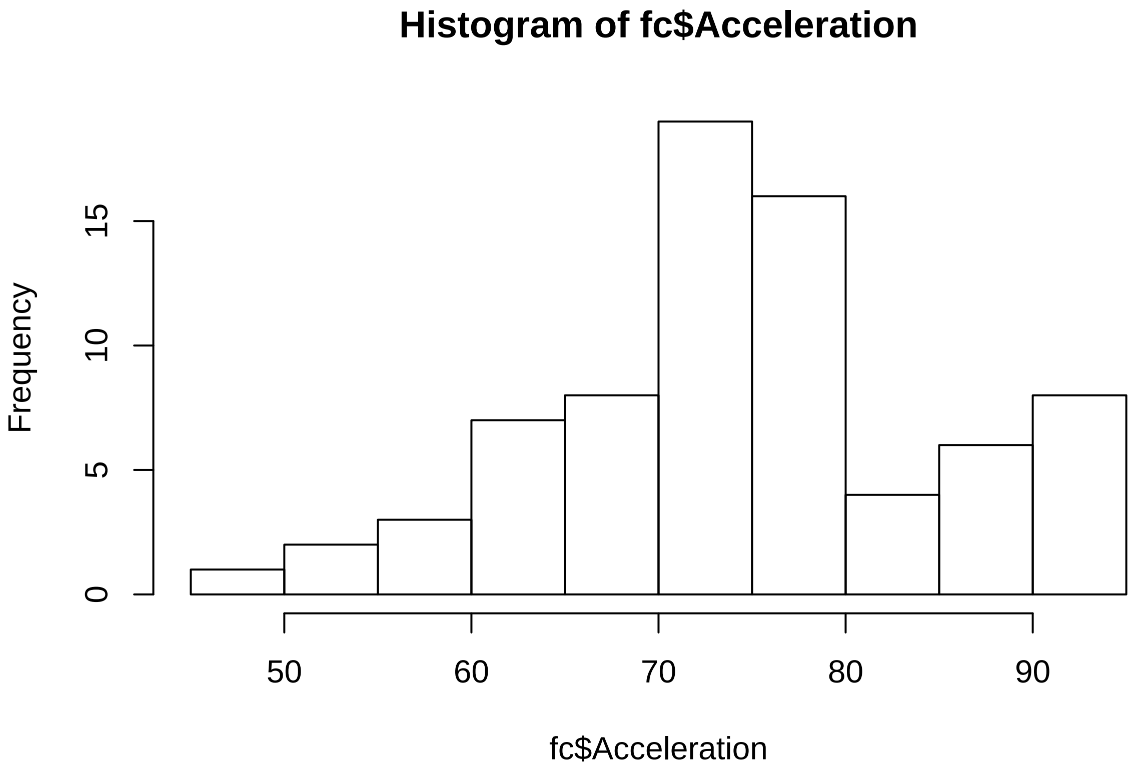
```
mod$residuals[c(56,40,10)]
```

```
##         56          40          10
## -11.815569   9.539091  -9.036906
```

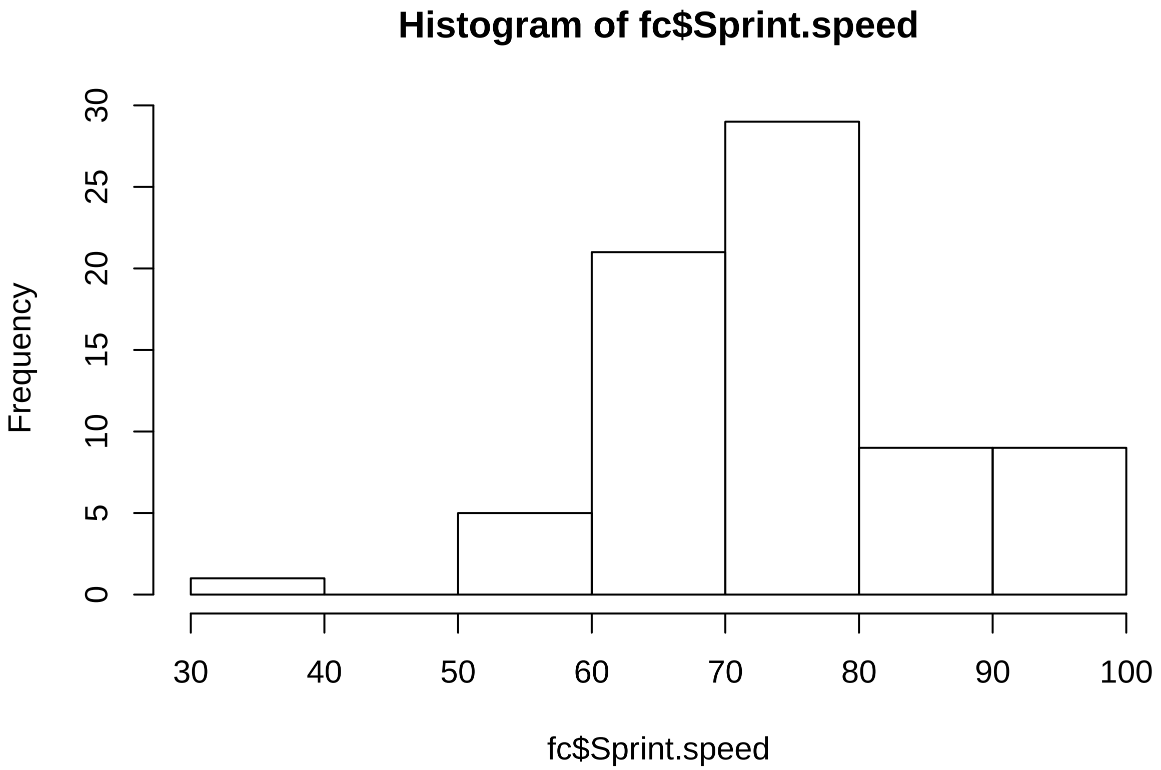
```
hist(fc$Agility)
```



```
hist(fc$Acceleration)
```



```
hist(fc$Sprint.speed)
```



The case of Carvajal is strange because it has good agility and sprint.speed but lower Acceleratin than expected. In the case of Suarez it has a good Acceleration that cannot be explained with its values on Agility and Sprint.Speed. FOr the case of Pique Acceleration would behave similar to Agility but has a good

Sprint.Speed that has an effect on over-predicting the Acceleration.

Parameter Interpretation

The first model is the one with greater **Adjusted R-squared**, it seems that is the better one. Its formula is Y = 0.27917x_1 + 0.64800x_2 - 0.02584x_3 + 0.08161x_4 + 2.06239 + e meaning that an increase in one point of Acceleration increases in an average: 0.27917 for Agility, 0.64800 for Sprint.Speed, 0.02584 for Reactions and 0.08161 for Balance.

Answer hypothesis and Final Comments

From our analsis we can conclude that Acceleration is linearly related with Agilit and Sprint.Speed in the sample of data of football players that we have. We have seen that a higher Acceleration is related to a high Agility and a high Sprint.Speed but the same does not happens with Reactions and Balance. We can conclude that Balance and Reactions have little influence on Acceleration, also seems that reactions have a negative impact on Acceleration. Finally we must say that the model is not capable of explaining the variance of Acceleration using those variables and that maybe more variables are needed in order to be able to solve this.