Module 2

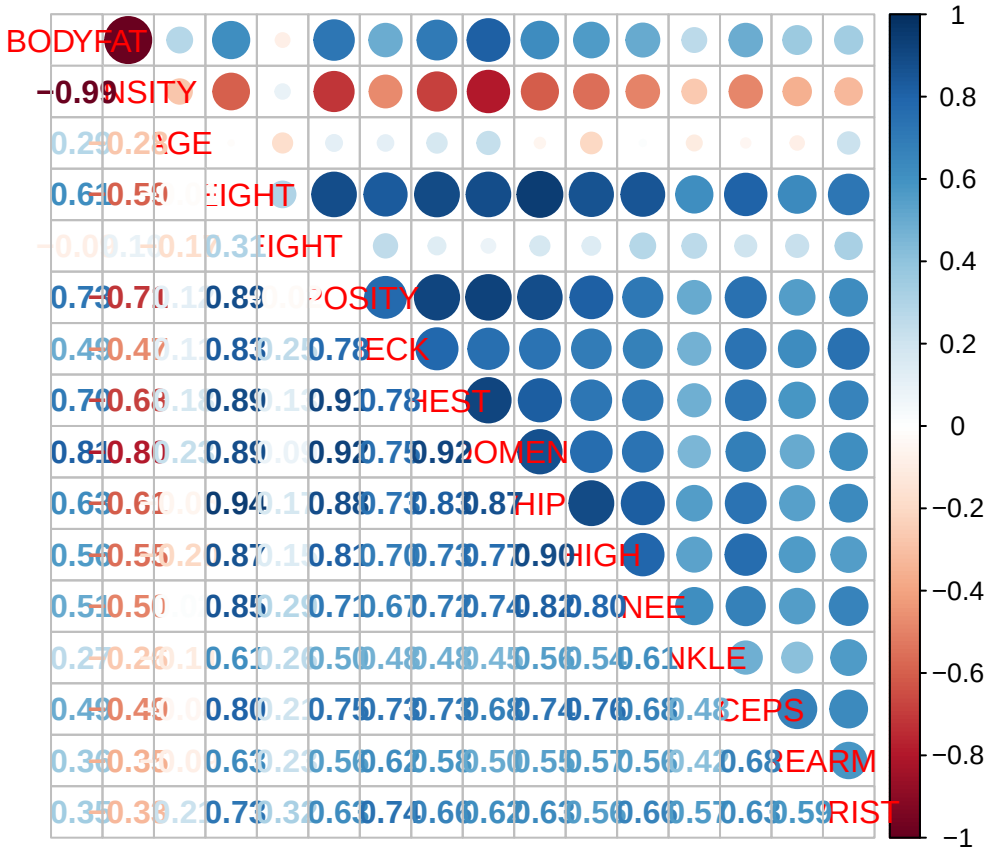
2022-10-18

```
# import needed packages
library(tidyverse)
library(corrplot)
library(MASS)
library(DAAG)
library(glmnet)
```

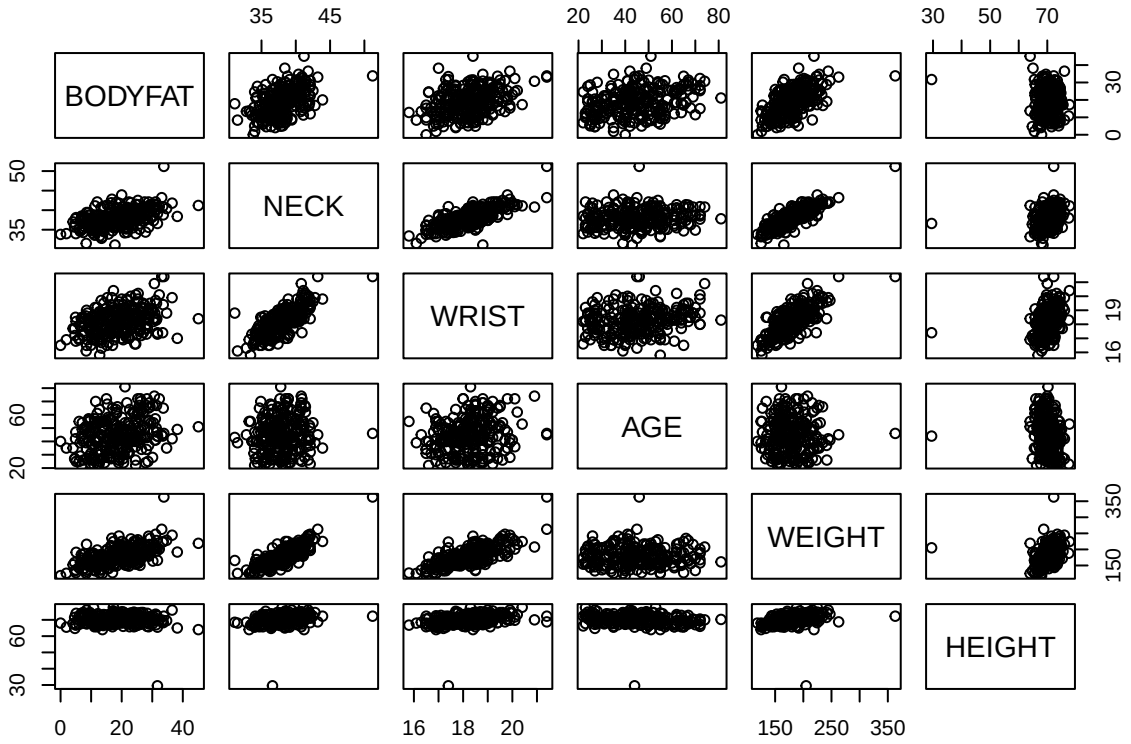
```
# read the raw dataset
df <- read.csv("../data/BodyFat.csv")
summary(df)
```

```
##       IDNO           BODYFAT         DENSITY           AGE
##  Min.   :  1.00   Min.   : 0.00   Min.   :0.995   Min.   :22.00
##  1st Qu.: 63.75   1st Qu.:12.80   1st Qu.:1.041   1st Qu.:35.75
##  Median :126.50   Median :19.00   Median :1.055   Median :43.00
##  Mean   :126.50   Mean   :18.94   Mean   :1.056   Mean   :44.88
##  3rd Qu.:189.25   3rd Qu.:24.60   3rd Qu.:1.070   3rd Qu.:54.00
##  Max.   :252.00   Max.   :45.10   Max.   :1.109   Max.   :81.00
##      WEIGHT          HEIGHT        ADIPOSITY          NECK
##  Min.   :118.5   Min.   :29.50   Min.   :18.10   Min.   :31.10
##  1st Qu.:159.0   1st Qu.:68.25   1st Qu.:23.10   1st Qu.:36.40
##  Median :176.5   Median :70.00   Median :25.05   Median :38.00
##  Mean   :178.9   Mean   :70.15   Mean   :25.44   Mean   :37.99
##  3rd Qu.:197.0   3rd Qu.:72.25   3rd Qu.:27.32   3rd Qu.:39.42
##  Max.   :363.1   Max.   :77.75   Max.   :48.90   Max.   :51.20
##      CHEST           ABDOMEN            HIP            THIGH
##  Min.   : 79.30   Min.   : 69.40   Min.   : 85.0   Min.   :47.20
##  1st Qu.: 94.35   1st Qu.: 84.58   1st Qu.: 95.5   1st Qu.:56.00
##  Median : 99.65   Median : 90.95   Median : 99.3   Median :59.00
##  Mean   :100.82   Mean   : 92.56   Mean   : 99.9   Mean   :59.41
##  3rd Qu.:105.38   3rd Qu.: 99.33   3rd Qu.:103.5   3rd Qu.:62.35
##  Max.   :136.20   Max.   :148.10   Max.   :147.7   Max.   :87.30
##       KNEE            ANKLE           BICEPS          FOREARM          WRIST
##  Min.   :33.00   Min.   :19.1   Min.   :24.80   Min.   :21.00   Min.   :15.80
##  1st Qu.:36.98   1st Qu.:22.0   1st Qu.:30.20   1st Qu.:27.30   1st Qu.:17.60
##  Median :38.50   Median :22.8   Median :32.05   Median :28.70   Median :18.30
##  Mean   :38.59   Mean   :23.1   Mean   :32.27   Mean   :28.66   Mean   :18.23
##  3rd Qu.:39.92   3rd Qu.:24.0   3rd Qu.:34.33   3rd Qu.:30.00   3rd Qu.:18.80
##  Max.   :49.10   Max.   :33.9   Max.   :45.00   Max.   :34.90   Max.   :21.40
```

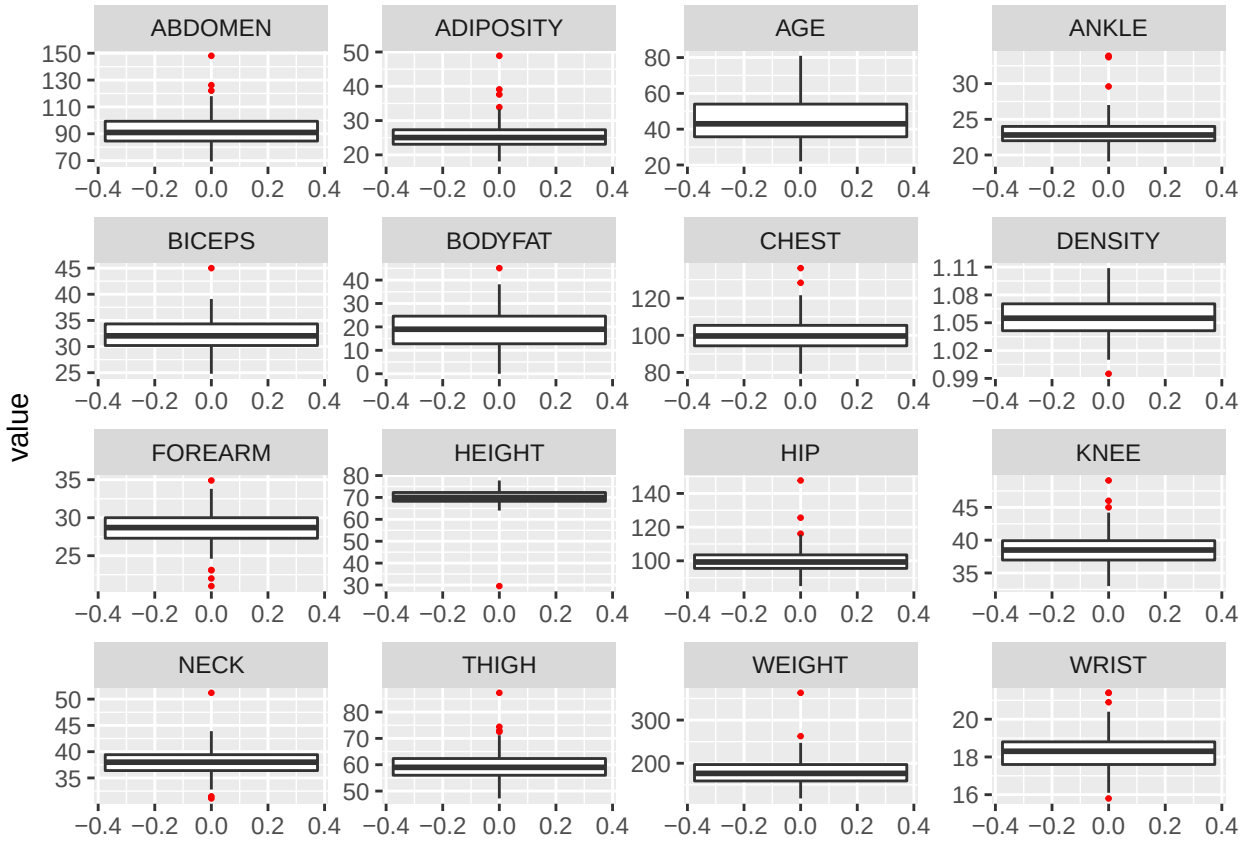
```
# overview the raw data, for pairs(), you can add other variables
corrplot.mixed(cor(df[,2:17]),lower="number",upper = "circle")
```




```
pairs(df[,c("BODYFAT","NECK",'WRIST',"AGE",'WEIGHT',"HEIGHT")])
```




```
# boxplot for each variable to see if there are outliers
df %>% gather(key=variable,value=value,-IDNO) %>%
  ggplot(aes(y=value)) +
  geom_boxplot(outlier.colour = "red",outlier.size = 0.5) +
  facet_wrap(~variable,scales="free")
```




```
# build the full model with all of the variables first and then use stepwise algorithm to select variab
full_model <- lm(BODYFAT~.-IDNO-DENSITY,data=df_cleaned)
step.model <- stepAIC(full_model,direction = "both")
```

```
## Start:  AIC=701.6
## BODYFAT ~ (IDNO + DENSITY + AGE + WEIGHT + HEIGHT + ADIPOSITY +
##     NECK + CHEST + ABDOMEN + HIP + THIGH + KNEE + ANKLE + BICEPS +
##     FOREARM + WRIST) - IDNO - DENSITY
##
##              Df Sum of Sq    RSS    AIC
## - KNEE        1      0.00 3694.7 699.60
## - CHEST       1      3.01 3697.7 699.81
## - HEIGHT      1      4.92 3699.6 699.93
## - ANKLE       1      7.07 3701.8 700.08
## - ADIPOSITY   1      9.26 3704.0 700.23
## - BICEPS      1     11.01 3705.7 700.34
## <none>                    3694.7 701.60
## - WEIGHT      1     30.65 3725.4 701.66
## - THIGH       1     35.05 3729.8 701.95
## - HIP         1     36.83 3731.5 702.07
## - AGE         1     48.75 3743.4 702.87
## - NECK        1     60.17 3754.9 703.63
## - FOREARM     1     72.81 3767.5 704.46
## - WRIST       1    141.23 3835.9 708.94
## - ABDOMEN     1   1630.23 5324.9 790.61
##
```

```
## Step:  AIC=699.6
## BODYFAT ~ AGE + WEIGHT + HEIGHT + ADIPOSITY + NECK + CHEST +
##     ABDOMEN + HIP + THIGH + ANKLE + BICEPS + FOREARM + WRIST
##
##             Df Sum of Sq    RSS    AIC
## - CHEST      1      3.01 3697.7 697.81
## - HEIGHT     1      4.93 3699.6 697.94
## - ANKLE      1      7.31 3702.0 698.10
## - ADIPOSITY  1      9.27 3704.0 698.23
## - BICEPS     1     11.03 3705.7 698.35
## <none>                   3694.7 699.60
## - WEIGHT     1     30.88 3725.6 699.68
## - HIP        1     36.99 3731.7 700.08
## - THIGH      1     38.57 3733.3 700.19
## - AGE        1     51.88 3746.6 701.08
## + KNEE       1      0.00 3694.7 701.60
## - NECK       1     60.88 3755.6 701.67
## - FOREARM    1     73.33 3768.0 702.50
## - WRIST      1    142.09 3836.8 707.00
## - ABDOMEN    1   1630.28 5325.0 788.62
##
## Step:  AIC=697.81
## BODYFAT ~ AGE + WEIGHT + HEIGHT + ADIPOSITY + NECK + ABDOMEN +
##     HIP + THIGH + ANKLE + BICEPS + FOREARM + WRIST
##
##             Df Sum of Sq    RSS    AIC
## - HEIGHT     1      4.13 3701.8 696.08
## - ADIPOSITY  1      7.34 3705.0 696.30
## - ANKLE      1      8.29 3706.0 696.36
## - BICEPS     1     10.45 3708.2 696.51
## <none>                   3697.7 697.81
## - WEIGHT     1     32.41 3730.1 697.98
## - HIP        1     34.06 3731.8 698.09
## - THIGH      1     45.06 3742.8 698.82
## - AGE        1     50.87 3748.6 699.21
## + CHEST      1      3.01 3694.7 699.60
## + KNEE       1      0.00 3697.7 699.81
## - NECK       1     60.58 3758.3 699.85
## - FOREARM    1     71.48 3769.2 700.57
## - WRIST      1    140.94 3838.7 705.12
## - ABDOMEN    1   1729.77 5427.5 791.36
##
## Step:  AIC=696.08
## BODYFAT ~ AGE + WEIGHT + ADIPOSITY + NECK + ABDOMEN + HIP + THIGH +
##     ANKLE + BICEPS + FOREARM + WRIST
##
##             Df Sum of Sq    RSS    AIC
## - ADIPOSITY  1      6.71 3708.6 694.54
## - ANKLE      1      9.67 3711.5 694.73
## - BICEPS     1     10.11 3712.0 694.76
## <none>                   3701.8 696.08
## - HIP        1     32.33 3734.2 696.25
## - THIGH      1     42.31 3744.2 696.91
## - AGE        1     48.81 3750.6 697.35
```

```
## + HEIGHT     1      4.13 3697.7 697.81
## + CHEST      1      2.21 3699.6 697.94
## + KNEE       1      0.01 3701.8 698.08
## - NECK       1     62.89 3764.7 698.28
## - FOREARM    1     77.67 3779.5 699.25
## - WEIGHT     1     95.54 3797.4 700.43
## - WRIST      1    139.82 3841.7 703.32
## - ABDOMEN    1   1792.47 5494.3 792.41
##
## Step:  AIC=694.54
## BODYFAT ~ AGE + WEIGHT + NECK + ABDOMEN + HIP + THIGH + ANKLE +
##     BICEPS + FOREARM + WRIST
##
##             Df Sum of Sq    RSS    AIC
## - ANKLE      1     11.25 3719.8 693.29
## - BICEPS     1     13.72 3722.3 693.45
## - HIP        1     28.26 3736.8 694.43
## <none>                   3708.6 694.54
## - THIGH      1     46.59 3755.1 695.64
## - AGE        1     49.63 3758.2 695.85
## + ADIPOSITY  1      6.71 3701.8 696.08
## + HEIGHT     1      3.51 3705.0 696.30
## - NECK       1     57.53 3766.1 696.37
## + KNEE       1      0.29 3708.3 696.52
## + CHEST      1      0.09 3708.5 696.53
## - FOREARM    1     80.46 3789.0 697.88
## - WEIGHT     1    101.48 3810.0 699.26
## - WRIST      1    144.34 3852.9 702.04
## - ABDOMEN    1   2739.29 6447.8 830.26
##
## Step:  AIC=693.29
## BODYFAT ~ AGE + WEIGHT + NECK + ABDOMEN + HIP + THIGH + BICEPS +
##     FOREARM + WRIST
##
##             Df Sum of Sq    RSS    AIC
## - BICEPS     1     12.64 3732.4 692.13
## - HIP        1     29.51 3749.3 693.26
## <none>                   3719.8 693.29
## - AGE        1     46.98 3766.8 694.41
## + ANKLE      1     11.25 3708.6 694.54
## - THIGH      1     49.07 3768.9 694.55
## + ADIPOSITY  1      8.30 3711.5 694.73
## + HEIGHT     1      4.21 3715.6 695.01
## + CHEST      1      0.18 3719.6 695.28
## + KNEE       1      0.00 3719.8 695.29
## - NECK       1     65.56 3785.4 695.64
## - FOREARM    1     79.71 3799.5 696.57
## - WEIGHT     1     90.93 3810.7 697.30
## - WRIST      1    133.19 3853.0 700.05
## - ABDOMEN    1   2743.02 6462.8 828.84
##
## Step:  AIC=692.13
## BODYFAT ~ AGE + WEIGHT + NECK + ABDOMEN + HIP + THIGH + FOREARM +
##     WRIST
```

```
##
##             Df Sum of Sq    RSS    AIC
## <none>                   3732.4 692.13
## - HIP        1     32.68 3765.1 692.31
## + BICEPS     1     12.64 3719.8 693.29
## + ADIPOSITY  1     11.96 3720.5 693.34
## + ANKLE      1     10.18 3722.3 693.45
## - AGE        1     50.53 3783.0 693.48
## + HEIGHT     1      7.01 3725.4 693.67
## - NECK       1     60.16 3792.6 694.12
## + KNEE       1      0.10 3732.3 694.13
## + CHEST      1      0.00 3732.4 694.13
## - THIGH      1     67.34 3799.8 694.59
## - WEIGHT     1     81.43 3813.9 695.51
## - FOREARM    1    111.62 3844.1 697.47
## - WRIST      1    132.27 3864.7 698.81
## - ABDOMEN    1   2730.78 6463.2 826.85
```

```
summary(step.model)
```

```
##
## Call:
## lm(formula = BODYFAT ~ AGE + WEIGHT + NECK + ABDOMEN + HIP +
##     THIGH + FOREARM + WRIST, data = df_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.1278  -2.7502  -0.1838   2.6860   9.4442
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -18.47088   10.84518  -1.703  0.08984 .
## AGE           0.05164    0.02865   1.803  0.07272 .
## WEIGHT       -0.08447    0.03691  -2.288  0.02299 *
## NECK         -0.40850    0.20769  -1.967  0.05035 .
## ABDOMEN       0.88102    0.06649  13.251  < 2e-16 ***
## HIP          -0.18774    0.12952  -1.450  0.14848
## THIGH         0.24979    0.12004   2.081  0.03850 *
## FOREARM       0.46077    0.17199   2.679  0.00789 **
## WRIST        -1.37525    0.47156  -2.916  0.00388 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.944 on 240 degrees of freedom
## Multiple R-squared:  0.7383, Adjusted R-squared:  0.7296
## F-statistic: 84.66 on 8 and 240 DF,  p-value: < 2.2e-16
```

```
# assess the multicollinearity via vif()
vif(step.model)
```

```
##     AGE  WEIGHT    NECK ABDOMEN     HIP   THIGH FOREARM   WRIST
##  2.0976 18.3650  4.0147  8.0153 13.2690  6.0735  1.9031  3.0493
```

```
# remove weight and hip because their vif are greater than 10
# remove thigh because it becomes insignificant after the last step.
md2 <- lm(BODYFAT ~ AGE  + NECK + ABDOMEN +
```

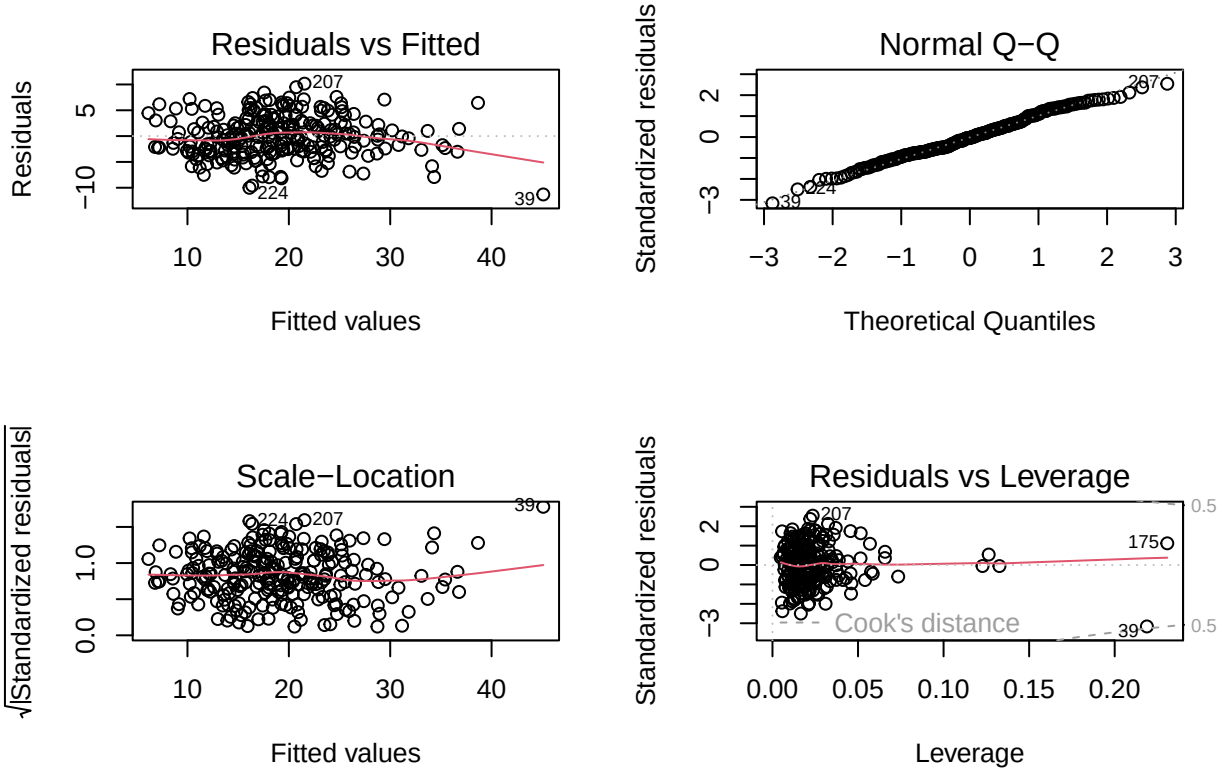
```
    FOREARM + WRIST, data = df_cleaned)
summary(md2)
```

```
##
## Call:
## lm(formula = BODYFAT ~ AGE + NECK + ABDOMEN + FOREARM + WRIST,
##     data = df_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.3001  -2.8490  -0.1616   2.6663  10.1543
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.37319    5.43044  -0.989  0.32342
## AGE          0.08530    0.02204   3.870  0.00014 ***
## NECK        -0.54648    0.19905  -2.745  0.00649 **
## ABDOMEN      0.71306    0.03780  18.863  < 2e-16 ***
## FOREARM      0.44551    0.17305   2.575  0.01063 *
## WRIST       -2.05581    0.43712  -4.703  4.3e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.041 on 243 degrees of freedom
## Multiple R-squared:  0.7218, Adjusted R-squared:  0.7161
## F-statistic: 126.1 on 5 and 243 DF,  p-value: < 2.2e-16
```

```
vif(md2)
```

```
##     AGE    NECK ABDOMEN FOREARM   WRIST
##  1.1823  3.5121  2.4678  1.8349  2.4955
```

```
par(mfrow=c(2,2))
plot(md2)
```




```
# remove the data with ID 39 because it is not only an outlier but also an influential point.
# refit the final model and then evaluate it
df_c1=df_cleaned[-which(df_cleaned$IDNO == 39),]
md3 <- lm(BODYFAT ~ AGE  + NECK + ABDOMEN +
    FOREARM + WRIST , data = df_c1)
summary(md3)
```

```
##
## Call:
## lm(formula = BODYFAT ~ AGE + NECK + ABDOMEN + FOREARM + WRIST,
##     data = df_c1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6492 -2.7694 -0.1038  2.6696 10.0557
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.01665    5.39108  -1.487  0.13831
## AGE          0.07598    0.02182   3.482  0.00059 ***
## NECK        -0.39208    0.20109  -1.950  0.05236 .
## ABDOMEN      0.73243    0.03757  19.493  < 2e-16 ***
## FOREARM      0.27437    0.17790   1.542  0.12431
## WRIST       -2.03587    0.42895  -4.746 3.55e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.965 on 242 degrees of freedom
## Multiple R-squared:  0.7291, Adjusted R-squared:  0.7235
## F-statistic: 130.3 on 5 and 242 DF,  p-value: < 2.2e-16
```

```
vif(md3)
```

```
##     AGE    NECK ABDOMEN FOREARM   WRIST
##  1.2033  3.2753  2.2554  2.0139  2.3790
```

```
par(mfrow=c(2,2))
plot(md3)
```