final project

Yuchen Zhang

2022-12-12

Data exploratory

Distribution of Data

Import data

```
Body_df = readxl::read_excel("data/body_density_data.xlsx")
tbl_summary(Body_df)
```

```
## Table printed with `knitr::kable()`, not {gt}. Learn why at
## https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
## To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**N = 252**
id	126 (64, 189)
bodyfat_brozek	19 (13, 25)
bodyfat_siri	19 (12, 25)
body_density	1.055 (1.041, 1.070)
age	43 (36, 54)
weight	176 (159, 197)
height	70.00 (68.25, 72.25)
neck	38.00 (36.40, 39.42)
chest	100 (94, 105)
abdomen	91 (85, 99)
hip	99 (96, 104)
thigh	59.0 (56.0, 62.3)
knee	38.50 (36.98, 39.92)
ankle	22.80 (22.00, 24.00)
bicep	32.05 (30.20, 34.32)
forearm	28.70 (27.30, 30.00)
wrist	18.30 (17.60, 18.80)

```
summary(Body_df)
```

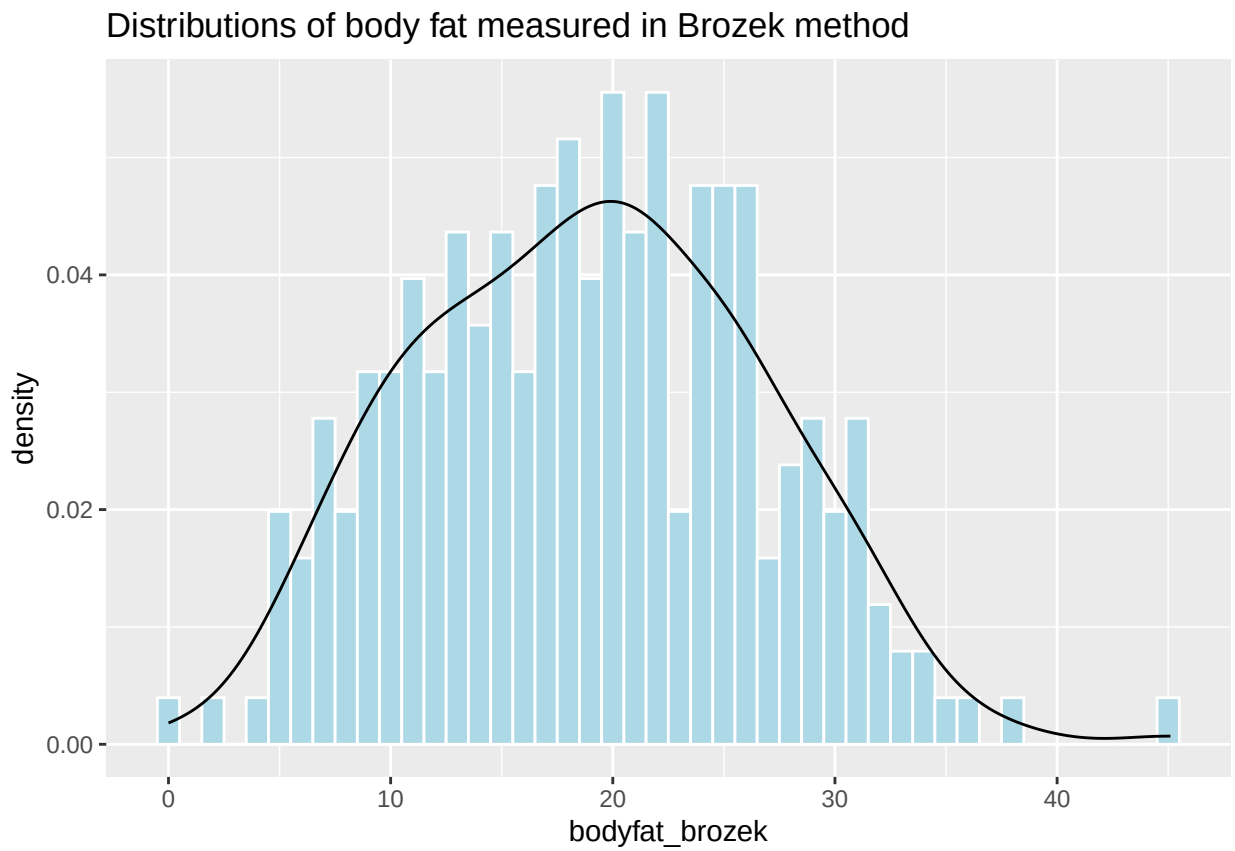
```
##        id         bodyfat_brozek   bodyfat_siri    body_density
##  Min.   :  1.00   Min.   : 0.00   Min.   : 0.00   Min.   :0.995
##  1st Qu.: 63.75   1st Qu.:12.80   1st Qu.:12.47   1st Qu.:1.041
##  Median :126.50   Median :19.00   Median :19.20   Median :1.055
```

```
##  Mean   :126.50   Mean   :18.94   Mean   :19.15   Mean   :1.056
##  3rd Qu.:189.25   3rd Qu.:24.60   3rd Qu.:25.30   3rd Qu.:1.070
##  Max.   :252.00   Max.   :45.10   Max.   :47.50   Max.   :1.109
##       age             weight          height           neck
##  Min.   :22.00   Min.   :118.5   Min.   :64.00   Min.   :31.10
##  1st Qu.:35.75   1st Qu.:159.0   1st Qu.:68.25   1st Qu.:36.40
##  Median :43.00   Median :176.5   Median :70.00   Median :38.00
##  Mean   :44.88   Mean   :178.9   Mean   :70.31   Mean   :37.99
##  3rd Qu.:54.00   3rd Qu.:197.0   3rd Qu.:72.25   3rd Qu.:39.42
##  Max.   :81.00   Max.   :363.1   Max.   :77.75   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           bicep          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
```

We chose fat density of Brozek's function as outcome and here is the distribution of bodyfat_brozek

```
ggplot(Body_df, aes(x = bodyfat_brozek)) +
 geom_histogram(aes(y = ..density..), color = "white", fill = "light blue",binwidth = 1) +
 geom_density(alpha = .2) +
 labs(title = "Distributions of body fat measured in Brozek method")
```

```
## Warning: The dot-dot notation (‘..density..‘) was deprecated in ggplot2 3.4.0.
## i Please use ‘after_stat(density)‘ instead.
```



Here are the distribution of all the variables shown in histogram

```
colnames = colnames(Body_df)
pdf("histogram.pdf")
for (i in 5:17) {
  plot =
ggplot(Body_df, aes_string(x = colnames[i])) +
 geom_histogram(aes(y = ..density..), color = "white", fill = "light blue",binwidth = 1) +
 geom_density(alpha = .2) +
 labs(title = sprintf("Distributions of %s", colnames[i]) )

  print(plot)
}
```

```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation ideoms with `aes()`
```

Here are the distribution of all the variables shown in boxplot

```
colnames = colnames(Body_df)
pdf("boxplot.pdf")
boxplot=
  for (i in 5:17) {
  plot =
ggplot(Body_df, aes_string(y = colnames[i])) +
```

```
 geom_boxplot() +
 labs(title = sprintf("Distributions of %s", colnames[i]) )

  print(plot)
}
```

From the distribution plots of the variables, we can find that, all the variables are in normal distribution with very few outliers. Thus, no transformation is required. There is a sample in which body fat calculated with Brozek's equation is 0 percent, which is abnormal, so we eliminate this point. Since our sample size is small, we won't do any further research to remove other abnormal data points that are not so obvious.

Here, we clean the data

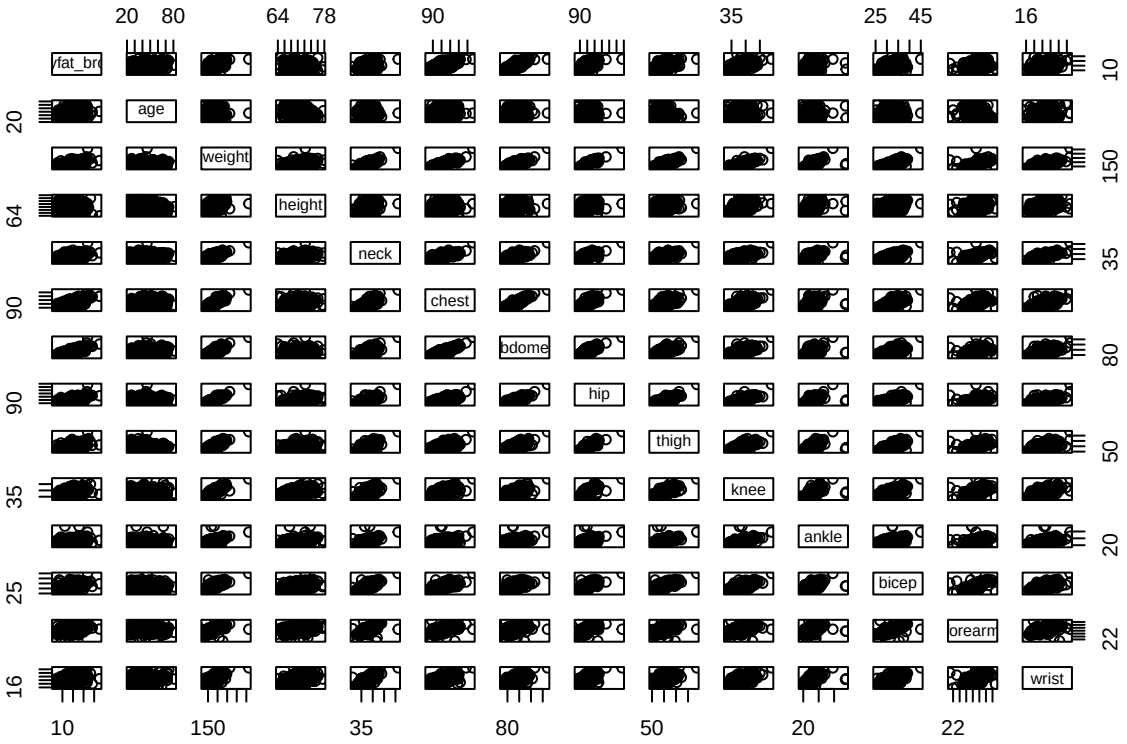
```
bodyfat_selected =
  Body_df %>%
  dplyr::select(-id,-bodyfat_siri,-body_density) %>%
  filter(bodyfat_brozek > 0)
```

Here are the correlation between the bodyfat_brozek with the predictors

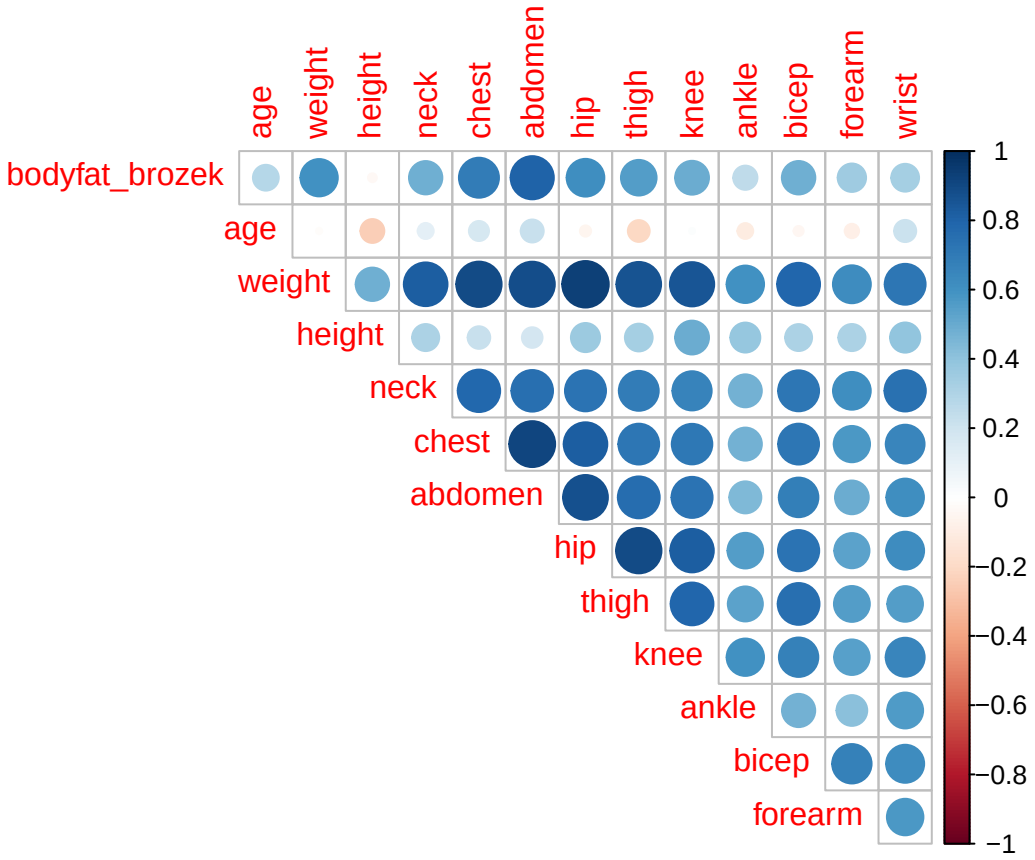
```
# Building scatter plots
pdf("correlation.pdf")
correlation =
for (i in 5:17) {
  plot =
  Body_df %>%
    ggplot(aes_string(x = colnames[i], y = "bodyfat_brozek")) + geom_point() + geom_smooth(method = 'lm
    labs(title = sprintf("Scatter plot for body fat against %s", colnames[i]) ) +
    ylab("Body Fat (Brozek)")

  print(plot)
}
```

```
# Pair the variables
pair = pairs(bodyfat_selected)
```




```
# Corrplot the variables
corplot = corrplot(cor(bodyfat_selected), type = "upper", diag = FALSE)
```



We can find that, most of the variables have linear correlation with others. It should be paid attention in later data analysis.

Building models

All the variables are normally distributed and required no transformation. As the outcome and predictors are all quantitative data, it would be improper to fit the model in Logistic Regression or exponential regression. Thus linear regression may be still the best choice for use to build the model.

Let's firstly build fit all the variables in a Multiple linear regression model.

```
multifit = lm(bodyfat_brozek ~ ., data = bodyfat_selected)
```

Linear Regression

Automatic Selection Backward Elmination

```
step(multifit, direction = "backward")
```

```
## Start:  AIC=707.65
## bodyfat_brozek ~ age + weight + height + neck + chest + abdomen +
##     hip + thigh + knee + ankle + bicep + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
```

```
## - knee     1      0.05 3764.0 705.66
## - height   1      1.79 3765.8 705.77
## - chest    1      2.04 3766.0 705.79
## - ankle    1      9.00 3773.0 706.25
## - bicep    1     17.07 3781.1 706.79
## - hip      1     30.04 3794.0 707.65
## <none>                3764.0 707.65
## - weight   1     30.99 3795.0 707.71
## - thigh    1     42.91 3806.9 708.50
## - age      1     60.20 3824.2 709.64
## - neck     1     64.55 3828.5 709.92
## - forearm  1     82.23 3846.2 711.08
## - wrist    1    151.88 3915.9 715.58
## - abdomen  1   1794.96 5558.9 803.53
##
## Step:  AIC=705.66
## bodyfat_brozek ~ age + weight + height + neck + chest + abdomen +
##     hip + thigh + ankle + bicep + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - chest    1      2.02 3766.1 703.79
## - height   1      2.04 3766.1 703.79
## - ankle    1      9.04 3773.1 704.26
## - bicep    1     17.21 3781.2 704.80
## <none>                3764.0 705.66
## - hip      1     30.37 3794.4 705.67
## - weight   1     31.61 3795.6 705.76
## - thigh    1     46.33 3810.4 706.73
## - age      1     63.36 3827.4 707.85
## - neck     1     64.94 3829.0 707.95
## - forearm  1     82.45 3846.5 709.09
## - wrist    1    153.16 3917.2 713.67
## - abdomen  1   1795.12 5559.2 801.53
##
## Step:  AIC=703.79
## bodyfat_brozek ~ age + weight + height + neck + abdomen + hip +
##     thigh + ankle + bicep + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - height   1      0.95 3767.0 701.85
## - ankle    1      9.59 3775.7 702.43
## - bicep    1     16.66 3782.7 702.90
## - hip      1     28.44 3794.5 703.68
## <none>                3766.1 703.79
## - weight   1     52.14 3818.2 705.24
## - thigh    1     53.48 3819.5 705.33
## - age      1     62.83 3828.9 705.94
## - neck     1     64.61 3830.7 706.06
## - forearm  1     80.59 3846.6 707.11
## - wrist    1    152.23 3918.3 711.74
## - abdomen  1   1995.48 5761.5 808.51
##
## Step:  AIC=701.85
## bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh +
```

```
##     ankle + bicep + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - ankle    1      9.98 3777.0 700.52
## - bicep    1     18.54 3785.6 701.09
## - hip      1     27.53 3794.5 701.68
## <none>                 3767.0 701.85
## - thigh    1     56.43 3823.4 703.59
## - age      1     63.49 3830.5 704.05
## - neck     1     64.03 3831.0 704.09
## - forearm  1     80.84 3847.9 705.18
## - weight   1     99.97 3867.0 706.43
## - wrist    1    154.49 3921.5 709.94
## - abdomen  1   2714.56 6481.6 836.07
##
## Step:  AIC=700.52
## bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh +
##     bicep + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - bicep    1     17.28 3794.3 699.66
## - hip      1     28.60 3805.6 700.41
## <none>                 3777.0 700.52
## - thigh    1     59.00 3836.0 702.41
## - age      1     60.63 3837.6 702.52
## - neck     1     72.15 3849.1 703.27
## - forearm  1     80.19 3857.2 703.79
## - weight   1     90.44 3867.4 704.46
## - wrist    1    144.51 3921.5 707.94
## - abdomen  1   2721.03 6498.0 834.70
##
## Step:  AIC=699.66
## bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh +
##     forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## <none>                 3794.3 699.66
## - hip      1     32.77 3827.1 699.82
## - neck     1     65.39 3859.7 701.95
## - age      1     65.69 3860.0 701.97
## - weight   1     78.60 3872.9 702.81
## - thigh    1     82.85 3877.1 703.09
## - forearm  1    116.31 3910.6 705.24
## - wrist    1    143.34 3937.6 706.97
## - abdomen  1   2705.26 6499.5 832.76


##
## Call:
## lm(formula = bodyfat_brozek ~ age + weight + neck + abdomen +
##     hip + thigh + forearm + wrist, data = bodyfat_selected)
##
## Coefficients:
## (Intercept)          age       weight         neck      abdomen          hip
##   -19.00338      0.05827     -0.08266     -0.42436      0.87429     -0.18514
```

```
##       thigh      forearm        wrist
##     0.27507      0.47002     -1.42508
```

The Final model obtained from Backward Elimination is **lm(formula = bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + wrist, data = bodyfat_selected)**

Forward selection

```
intercept_only = lm(bodyfat_brozek ~ 1, data = bodyfat_selected)
step(intercept_only, direction = "forward", scope = formula(multifit))
```

```
## Start:  AIC=1023.93
## bodyfat_brozek ~ 1
##
##           Df Sum of Sq     RSS     AIC
## + abdomen  1    9653.8  5065.2  758.18
## + chest    1    7115.7  7603.2  860.13
## + hip      1    5623.3  9095.6  905.11
## + weight   1    5394.8  9324.1  911.34
## + thigh    1    4470.3 10248.6  935.07
## + knee     1    3643.5 11075.4  954.54
## + bicep    1    3480.6 11238.3  958.21
## + neck     1    3437.9 11281.1  959.16
## + forearm  1    1809.7 12909.2  993.00
## + wrist    1    1659.6 13059.4  995.90
## + age      1    1228.8 13490.1 1004.05
## + ankle    1     954.0 13764.9 1009.11
## <none>                14718.9 1023.93
## + height   1      16.6 14702.3 1025.65
##
## Step:  AIC=758.18
## bodyfat_brozek ~ abdomen
##
##           Df Sum of Sq    RSS    AIC
## + weight   1    860.58 4204.6 713.44
## + wrist    1    614.47 4450.7 727.72
## + neck     1    523.82 4541.3 732.78
## + height   1    508.94 4556.2 733.60
## + hip      1    474.10 4591.1 735.51
## + knee     1    289.81 4775.4 745.39
## + ankle    1    206.25 4858.9 749.74
## + chest    1    182.56 4882.6 750.96
## + age      1    165.66 4899.5 751.83
## + thigh    1    152.60 4912.6 752.50
## + bicep    1    118.06 4947.1 754.26
## + forearm  1     48.54 5016.6 757.76
## <none>                5065.2 758.18
##
## Step:  AIC=713.44
## bodyfat_brozek ~ abdomen + weight
##
##           Df Sum of Sq    RSS    AIC
## + wrist    1   138.572 4066.0 707.03
## + neck     1    73.286 4131.3 711.02
```

```
## + thigh    1    67.257 4137.3 711.39
## + forearm  1    54.703 4149.9 712.15
## + bicep    1    52.962 4151.6 712.26
## <none>                 4204.6 713.44
## + height   1     7.011 4197.6 715.02
## + knee     1     4.995 4199.6 715.14
## + age      1     2.482 4202.1 715.29
## + ankle    1     0.761 4203.8 715.39
## + chest    1     0.525 4204.1 715.41
## + hip      1     0.048 4204.5 715.43
##
## Step:  AIC=707.03
## bodyfat_brozek ~ abdomen + weight + wrist
##
##           Df Sum of Sq    RSS    AIC
## + forearm  1   106.099 3959.9 702.39
## + bicep    1    74.409 3991.6 704.39
## <none>                 4066.0 707.03
## + thigh    1    32.057 4033.9 707.04
## + neck     1    20.253 4045.8 707.77
## + age      1    16.178 4049.8 708.03
## + knee     1    12.534 4053.5 708.25
## + ankle    1    11.057 4054.9 708.34
## + hip      1     9.105 4056.9 708.46
## + height   1     5.795 4060.2 708.67
## + chest    1     0.148 4065.9 709.02
##
## Step:  AIC=702.39
## bodyfat_brozek ~ abdomen + weight + wrist + forearm
##
##          Df Sum of Sq    RSS    AIC
## + neck    1    41.640 3918.3 701.74
## <none>                 3959.9 702.39
## + age     1    29.885 3930.0 702.49
## + bicep   1    28.482 3931.4 702.58
## + thigh   1    21.731 3938.2 703.01
## + ankle   1    13.765 3946.1 703.52
## + knee    1    12.212 3947.7 703.61
## + hip     1     3.783 3956.1 704.15
## + height  1     1.864 3958.0 704.27
## + chest   1     1.523 3958.4 704.29
##
## Step:  AIC=701.74
## bodyfat_brozek ~ abdomen + weight + wrist + forearm + neck
##
##          Df Sum of Sq    RSS    AIC
## + bicep   1    38.370 3879.9 701.27
## + age     1    37.504 3880.8 701.32
## <none>                 3918.3 701.74
## + thigh   1    20.135 3898.1 702.44
## + hip     1    10.509 3907.8 703.06
## + ankle   1     7.930 3910.3 703.23
## + height  1     6.364 3911.9 703.33
## + knee    1     5.675 3912.6 703.37
```

```
## + chest   1     0.252 3918.0 703.72
##
## Step:  AIC=701.27
## bodyfat_brozek ~ abdomen + weight + wrist + forearm + neck +
##     bicep
##
##          Df Sum of Sq    RSS    AIC
## + age     1    39.303 3840.6 700.71
## <none>                3879.9 701.27
## + hip     1    12.267 3867.6 702.47
## + thigh   1     9.438 3870.5 702.65
## + ankle   1     9.208 3870.7 702.67
## + knee    1     6.142 3873.8 702.87
## + height  1     1.578 3878.3 703.16
## + chest   1     0.708 3879.2 703.22
##
## Step:  AIC=700.71
## bodyfat_brozek ~ abdomen + weight + wrist + forearm + neck +
##     bicep + age
##
##          Df Sum of Sq    RSS    AIC
## + thigh   1    35.002 3805.6 700.41
## <none>                3840.6 700.71
## + ankle   1    12.700 3827.9 701.88
## + hip     1     4.600 3836.0 702.41
## + knee    1     3.516 3837.1 702.48
## + chest   1     3.193 3837.4 702.50
## + height  1     2.751 3837.8 702.53
##
## Step:  AIC=700.41
## bodyfat_brozek ~ abdomen + weight + wrist + forearm + neck +
##     bicep + age + thigh
##
##          Df Sum of Sq    RSS    AIC
## <none>                3805.6 700.41
## + hip     1   28.5999 3777.0 700.52
## + ankle   1   11.0547 3794.5 701.68
## + height  1    0.1318 3805.5 702.40
## + chest   1    0.1129 3805.5 702.40
## + knee    1    0.0066 3805.6 702.41

##
## Call:
## lm(formula = bodyfat_brozek ~ abdomen + weight + wrist + forearm +
##     neck + bicep + age + thigh, data = bodyfat_selected)
##
## Coefficients:
## (Intercept)      abdomen       weight        wrist      forearm         neck
##   -28.94511      0.85489     -0.11737     -1.42807      0.43701     -0.39574
##       bicep          age        thigh
##     0.18172      0.05799      0.16588
```

The model obtained from Forward Selection is **lm(formula = bodyfat_brozek ~ abdomen + weight + wrist + forearm + neck + bicep + age + thigh, data = bodyfat_selected)**

Stepwise Selection

```
step(multifit, direction = "both")
```

```
## Start:  AIC=707.65
## bodyfat_brozek ~ age + weight + height + neck + chest + abdomen +
##     hip + thigh + knee + ankle + bicep + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - knee     1      0.05 3764.0 705.66
## - height   1      1.79 3765.8 705.77
## - chest    1      2.04 3766.0 705.79
## - ankle    1      9.00 3773.0 706.25
## - bicep    1     17.07 3781.1 706.79
## - hip      1     30.04 3794.0 707.65
## <none>                 3764.0 707.65
## - weight   1     30.99 3795.0 707.71
## - thigh    1     42.91 3806.9 708.50
## - age      1     60.20 3824.2 709.64
## - neck     1     64.55 3828.5 709.92
## - forearm  1     82.23 3846.2 711.08
## - wrist    1    151.88 3915.9 715.58
## - abdomen  1   1794.96 5558.9 803.53
##
## Step:  AIC=705.66
## bodyfat_brozek ~ age + weight + height + neck + chest + abdomen +
##     hip + thigh + ankle + bicep + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - chest    1      2.02 3766.1 703.79
## - height   1      2.04 3766.1 703.79
## - ankle    1      9.04 3773.1 704.26
## - bicep    1     17.21 3781.2 704.80
## <none>                 3764.0 705.66
## - hip      1     30.37 3794.4 705.67
## - weight   1     31.61 3795.6 705.76
## - thigh    1     46.33 3810.4 706.73
## + knee     1      0.05 3764.0 707.65
## - age      1     63.36 3827.4 707.85
## - neck     1     64.94 3829.0 707.95
## - forearm  1     82.45 3846.5 709.09
## - wrist    1    153.16 3917.2 713.67
## - abdomen  1   1795.12 5559.2 801.53
##
## Step:  AIC=703.79
## bodyfat_brozek ~ age + weight + height + neck + abdomen + hip +
##     thigh + ankle + bicep + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - height   1      0.95 3767.0 701.85
## - ankle    1      9.59 3775.7 702.43
## - bicep    1     16.66 3782.7 702.90
## - hip      1     28.44 3794.5 703.68
## <none>                 3766.1 703.79
```

```
## - weight   1     52.14 3818.2 705.24
## - thigh    1     53.48 3819.5 705.33
## + chest    1      2.02 3764.0 705.66
## + knee     1      0.04 3766.0 705.79
## - age      1     62.83 3828.9 705.94
## - neck     1     64.61 3830.7 706.06
## - forearm  1     80.59 3846.6 707.11
## - wrist    1    152.23 3918.3 711.74
## - abdomen  1   1995.48 5761.5 808.51
##
## Step:  AIC=701.85
## bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh +
##     ankle + bicep + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - ankle    1      9.98 3777.0 700.52
## - bicep    1     18.54 3785.6 701.09
## - hip      1     27.53 3794.5 701.68
## <none>                 3767.0 701.85
## - thigh    1     56.43 3823.4 703.59
## + height   1      0.95 3766.1 703.79
## + chest    1      0.94 3766.1 703.79
## + knee     1      0.19 3766.8 703.84
## - age      1     63.49 3830.5 704.05
## - neck     1     64.03 3831.0 704.09
## - forearm  1     80.84 3847.9 705.18
## - weight   1     99.97 3867.0 706.43
## - wrist    1    154.49 3921.5 709.94
## - abdomen  1   2714.56 6481.6 836.07
##
## Step:  AIC=700.52
## bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh +
##     bicep + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - bicep    1     17.28 3794.3 699.66
## - hip      1     28.60 3805.6 700.41
## <none>                 3777.0 700.52
## + ankle    1      9.98 3767.0 701.85
## - thigh    1     59.00 3836.0 702.41
## + height   1      1.34 3775.7 702.43
## + chest    1      1.16 3775.8 702.44
## - age      1     60.63 3837.6 702.52
## + knee     1      0.01 3777.0 702.52
## - neck     1     72.15 3849.1 703.27
## - forearm  1     80.19 3857.2 703.79
## - weight   1     90.44 3867.4 704.46
## - wrist    1    144.51 3921.5 707.94
## - abdomen  1   2721.03 6498.0 834.70
##
## Step:  AIC=699.66
## bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh +
##     forearm + wrist
##
```

```
##           Df Sum of Sq    RSS    AIC
## <none>                  3794.3 699.66
## - hip      1     32.77 3827.1 699.82
## + bicep    1     17.28 3777.0 700.52
## + ankle    1      8.72 3785.6 701.09
## + height   1      3.34 3790.9 701.44
## + chest    1      0.37 3793.9 701.64
## + knee     1      0.10 3794.2 701.66
## - neck     1     65.39 3859.7 701.95
## - age      1     65.69 3860.0 701.97
## - weight   1     78.60 3872.9 702.81
## - thigh    1     82.85 3877.1 703.09
## - forearm  1    116.31 3910.6 705.24
## - wrist    1    143.34 3937.6 706.97
## - abdomen  1   2705.26 6499.5 832.76
```

```
##
## Call:
## lm(formula = bodyfat_brozek ~ age + weight + neck + abdomen +
##     hip + thigh + forearm + wrist, data = bodyfat_selected)
##
## Coefficients:
## (Intercept)          age       weight         neck      abdomen          hip
##   -19.00338      0.05827     -0.08266     -0.42436      0.87429     -0.18514
##       thigh      forearm        wrist
##     0.27507      0.47002     -1.42508
```

The stepwise selection from both side get us the model to be **lm(formula = bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + wrist, data = bodyfat_selected)**

The model obtained by Forward selections is included bicep than from that obtained in Backward elimination and Stepwise selection.

Model validation

```
set.seed(1)
train = trainControl(method = "cv", number = 5)
model_for = train(bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + bicep + forearm + wris
                  trControl = train,
                  method = "lm",
                  na.action = na.pass)

model_for$finalModel
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Coefficients:
## (Intercept)          age       weight         neck      abdomen          hip
##   -19.54727      0.05613     -0.09044     -0.44847      0.87823     -0.17358
##       thigh        bicep      forearm        wrist
##     0.24067      0.16368      0.41066     -1.43100
```

```
model_back = train(bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + wrist,
                   data = bodyfat_selected,
                   trControl = train,
                   method = "lm",
                   na.action = na.pass)

model_back$finalModel
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Coefficients:
## (Intercept)          age       weight         neck      abdomen          hip
##   -19.00338      0.05827     -0.08266     -0.42436      0.87429     -0.18514
##       thigh      forearm        wrist
##     0.27507      0.47002     -1.42508
```

```
print(model_for)
```

```
## Linear Regression
##
## 251 samples
##   9 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 201, 202, 202, 199, 200
## Resampling results:
##
##   RMSE     Rsquared   MAE
##   4.02546  0.7292992  3.286613
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
print(model_back)
```

```
## Linear Regression
##
## 251 samples
##   8 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 199, 201, 201, 200, 203
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   4.074935  0.7302663  3.330233
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

The RMSE of model_for(4.025) is slightly smaller than that of model_back(4.075). Model_for is more preferred. However, considering the principle of parsimony, we will run ANOVA to further compare the two models.

```
model_for = lm(bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + bicep + forearm + wrist, d

model_back = lm(bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + wrist, data = b
anova(model_back, model_for)
```

```
## Analysis of Variance Table
##
## Model 1: bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh +
##     forearm + wrist
## Model 2: bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh +
##     bicep + forearm + wrist
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    242 3794.3
## 2    241 3777.0  1    17.284 1.1028 0.2947
```

F = 1.103, p_value = 0.2947, we fail to reject H0 and conclude that the model_for is not superior. Now, according to the principle of parsimony, we will choose the model_for.

From the procedures we done above, the final model was agreed to be **lm(formula = bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + wrist, data = bodyfat_selected)**

Test Based Procedures Then, let's try Tested Based Procedures "Cp test"

```
mat = as.matrix(bodyfat_selected)
leaps(x = mat[,2:14], y = mat[,1], nbest = 2, method = "Cp")
```

```
## $which
##        1     2     3     4     5     6     7     8     9     A     B     C
## 1  FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 1  FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 2  FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 2  FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 3  FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 3  FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 4  FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE
## 4  FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE
## 5  FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE
## 5   TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE
## 6   TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE
## 6  FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE
## 7   TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE
## 7   TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE
## 8   TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE
## 8   TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE
## 9   TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE
## 9   TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
## 10  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
## 10  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE
```

```
## 11  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
## 11  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
## 12  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
## 12  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## 13  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##        D
## 1  FALSE
## 1  FALSE
## 2  FALSE
## 2   TRUE
## 3   TRUE
## 3  FALSE
## 4   TRUE
## 4   TRUE
## 5   TRUE
## 5   TRUE
## 6   TRUE
## 6   TRUE
## 7   TRUE
## 7   TRUE
## 8   TRUE
## 8   TRUE
## 9   TRUE
## 9   TRUE
## 10  TRUE
## 10  TRUE
## 11  TRUE
## 11  TRUE
## 12  TRUE
## 12  TRUE
## 13  TRUE
##
## $label
##  [1] "(Intercept)" "1"           "2"           "3"           "4"
##  [6] "5"           "6"           "7"           "8"           "9"
## [11] "A"           "B"           "C"           "D"
##
## $size
##  [1]  2  2  3  3  4  4  5  5  6  6  7  7  8  8  9  9 10 10 11 11 12 12 13 13 14
##
## $Cp
##  [1]  71.929289 231.737101  19.742464  35.239002  13.017232  17.128022
##  [7]   8.336695  10.332052   7.714835   8.455006   7.139163   7.298868
## [13]   5.971406   6.824138   5.907725   6.620238   6.819436   7.358543
## [19]   8.190932   8.735034  10.130923  10.131938  12.003444  12.112859
## [25]  14.000000
```

The smallest Cp value we got indicate that best model: **lm(formula = bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + wrist, data = bodyfat_selected)**

```
leaps(x = mat[,2:14], y = mat[,1], nbest = 2, method = "adjr2")
```

```
## $which
```

```
##        1     2     3     4     5     6     7     8     9     A     B     C
## 1  FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 1  FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 2  FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 2  FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 3  FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 3  FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
## 4  FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE
## 4  FALSE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE
## 5  FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE
## 5   TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE  TRUE
## 6   TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE
## 6  FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE
## 7   TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE
## 7   TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE  TRUE
## 8   TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE  TRUE
## 8   TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE
## 9   TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE
## 9   TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE
## 10  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
## 10  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE FALSE  TRUE  TRUE
## 11  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
## 11  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
## 12  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE
## 12  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## 13  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
##        D
## 1  FALSE
## 1  FALSE
## 2  FALSE
## 2   TRUE
## 3   TRUE
## 3  FALSE
## 4   TRUE
## 4   TRUE
## 5   TRUE
## 5   TRUE
## 6   TRUE
## 6   TRUE
## 7   TRUE
## 7   TRUE
## 8   TRUE
## 8   TRUE
## 9   TRUE
## 9   TRUE
## 10  TRUE
## 10  TRUE
## 11  TRUE
## 11  TRUE
## 12  TRUE
## 12  TRUE
## 13  TRUE
##
## $label
```

```
##  [1] "(Intercept)" "1"           "2"           "3"           "4"
##  [6] "5"           "6"           "7"           "8"           "9"
## [11] "A"           "B"           "C"           "D"
##
## $size
##  [1]  2  2  3  3  4  4  5  5  6  6  7  7  8  8  9  9 10 10 11 11 12 12 13 13 14
##
## $adjr2
##  [1] 0.6544921 0.4813663 0.7120382 0.6951825 0.7204012 0.7159118 0.7265902
##  [8] 0.7244022 0.7283610 0.7275460 0.7300952 0.7299186 0.7325010 0.7315544
## [15] 0.7336959 0.7329017 0.7338091 0.7332056 0.7334063 0.7327948 0.7323586
## [22] 0.7323575 0.7313786 0.7312545 0.7302491
```

The largest adjusted R2 indicated the best subset to be: **lm(formula = bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + bicep + forearm + wrist, data= bodyfat_selected)**

The model obtained by the via Mallow's Cp and Ajusted R squares are slightly different. As the models obtained are the identical as what we obtained in Automatic selection, same result can be obtained via model validation.

The model we chosen would be **lm(formula = bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + wrist, data = bodyfat_selected)**

LASSO Let's use corss validation to choose lambda

```
lambda_seq = 10^seq(-3, 0, by = 0.1)
set.seed(1)
cv_bodyfat = cv.glmnet(as.matrix(bodyfat_selected[2:14]), bodyfat_selected$bodyfat_brozek, lambda = lam
cv_bodyfat
```

```
##
## Call:  cv.glmnet(x = as.matrix(bodyfat_selected[2:14]), y = bodyfat_selected$bodyfat_brozek,      la
##
## Measure: Mean-Squared Error
##
##      Lambda Index Measure     SE Nonzero
## min 0.01585    19   16.83 0.8678      12
## 1se 0.19953     8   17.64 0.9817       6
```

The Lambda minimum is 0.0794. Then, let's reun a LASSO regression using this value.

```
lasso_fit = glmnet::glmnet(as.matrix(bodyfat_selected[2:14]), bodyfat_selected$bodyfat_brozek, lambda =
coef(lasso_fit)
```

```
## 14 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept) -11.79379143
## age           0.05677414
## weight       -0.06785393
## height       -0.08715773
## neck         -0.42814041
## chest        -0.02161955
## abdomen       0.85241509
```

```
## hip          -0.16240545
## thigh         0.18984863
## knee          .
## ankle         0.11205159
## bicep         0.14002270
## forearm       0.39704078
## wrist        -1.50944944
```

The final model obtained from LASSO is **lm(formula =bodyfat_brozek ~ age + weight + height + neck + chest + abdomen + hip + thigh + ankle + bicep + forearm + wrist, data = bodyfat_selected)**

Model choose

Stepwise selection and criterion test both indicate the same model. The LASSO model includes more predictors. Further process would be done to choose from the LASSO model: **lm(formula =bodyfat_brozek ~ age + weight + height + neck + chest + abdomen + hip + thigh + ankle + bicep + forearm + wrist, data = bodyfat_selected)** and the small model **lm(formula = bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + wrist, data= bodyfat_selected)**.

Model Validation Since the lasso method incurs a different model, so we will adopt model validation to choose the better model. First, use 5-fold validation and create the training sets.

```
train = trainControl(method = "cv", number = 5)
```

Then, fit the lasso model.

```
set.seed(1)
model_lasso = train(bodyfat_brozek ~ age + weight + height + neck + chest + abdomen + hip + thigh + ankl
                    data = bodyfat_selected,
                    trControl = train,
                    method = "lm",
                    na.action = na.pass)

model_lasso$finalModel
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Coefficients:
## (Intercept)          age       weight       height         neck        chest
##   -14.80034      0.05754     -0.08070     -0.05792     -0.43777     -0.03440
##     abdomen          hip        thigh        ankle        bicep      forearm
##     0.88644     -0.18474      0.22114      0.15289      0.16610      0.41913
##       wrist
##    -1.52734
```

```
model_test = train(bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + wrist,
                    data = bodyfat_selected,
                    trControl = train,
```

```
                    method = "lm",
                    na.action = na.pass)

model_test$finalModel
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Coefficients:
## (Intercept)          age       weight         neck      abdomen          hip
##   -19.00338      0.05827     -0.08266     -0.42436      0.87429     -0.18514
##       thigh      forearm        wrist
##     0.27507      0.47002     -1.42508
```

```
print(model_lasso)
```

```
## Linear Regression
##
## 251 samples
##  12 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 201, 202, 202, 199, 200
## Resampling results:
##
##   RMSE     Rsquared   MAE
##   4.18458  0.7113264  3.39777
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```
print(model_test)
```

```
## Linear Regression
##
## 251 samples
##   8 predictor
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 199, 201, 201, 200, 203
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   4.074935  0.7302663  3.330233
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

The RMSE obtained from model_lasso is 4.1846 and that of model_test is 4.0749. The MAE obtained from model_lasso is 3.397 and that of model_test is 3.330.

As the RMSE and MAE are slightly smaller in tested model, it is more preferred. However, considering the principle of parsimony, we will run ANOVA to further compare the two models.

```
model_small = lm(bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + wrist,
                   data = bodyfat_selected)
model_large = lm(bodyfat_brozek ~ age + weight + height + neck + chest + abdomen + hip + thigh + ankle +

anova(model_small,model_large)
```

ANOVA for MLR

```
## Analysis of Variance Table
##
## Model 1: bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh +
##     forearm + wrist
## Model 2: bodyfat_brozek ~ age + weight + height + neck + chest + abdomen +
##     hip + thigh + ankle + bicep + forearm + wrist
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    242 3794.3
## 2    238 3764.0  4    30.243 0.4781 0.7518
```

F = 0.4781, p_value = 0.7518, we fail to reject H0 and conclude that the larger model is not superior. Now, according to the principle of parsimony, we will choose the small model.

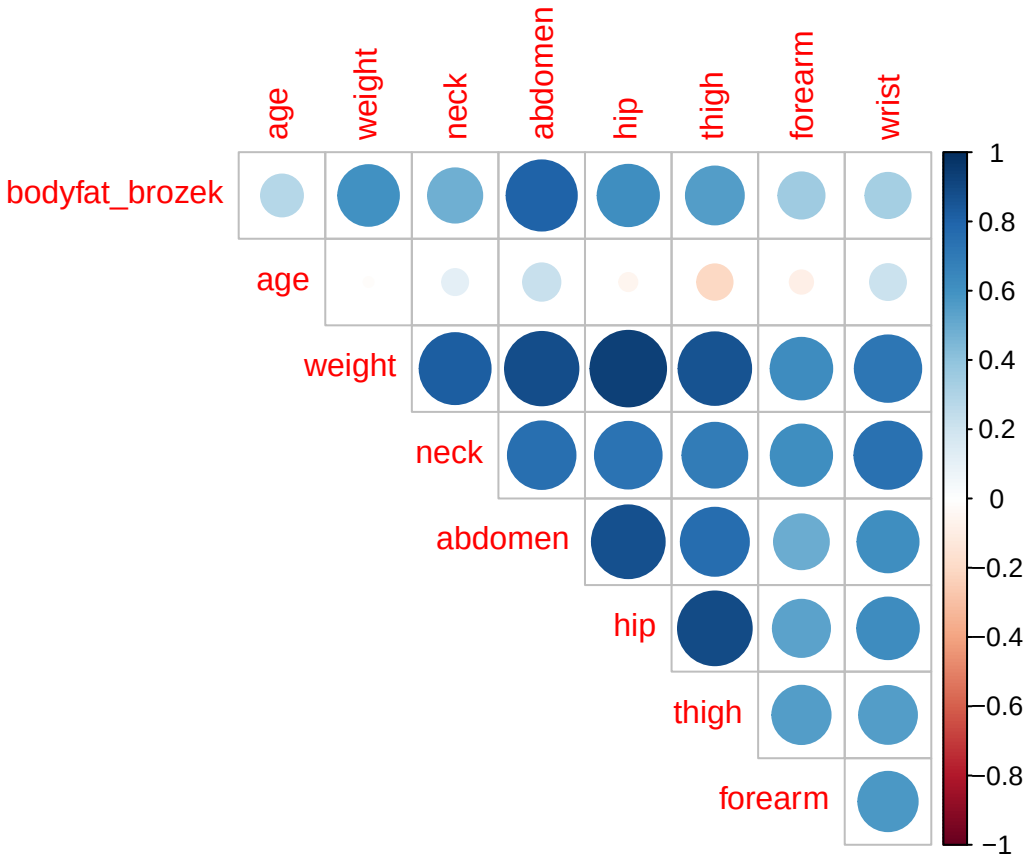
```
model_1 = model_small
summary(model_1)
```

```
##
## Call:
## lm(formula = bodyfat_brozek ~ age + weight + neck + abdomen +
##     hip + thigh + forearm + wrist, data = bodyfat_selected)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.117  -2.800  -0.223   2.709   9.477
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.00338   10.86380  -1.749  0.08152 .
## age           0.05827    0.02847   2.047  0.04175 *
## weight       -0.08266    0.03692  -2.239  0.02607 *
## neck         -0.42436    0.20780  -2.042  0.04222 *
## abdomen       0.87429    0.06656  13.136  < 2e-16 ***
## hip          -0.18514    0.12805  -1.446  0.14952
## thigh         0.27507    0.11966   2.299  0.02237 *
## forearm       0.47002    0.17257   2.724  0.00693 **
## wrist        -1.42508    0.47132  -3.024  0.00277 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.96 on 242 degrees of freedom
## Multiple R-squared:  0.7422, Adjusted R-squared:  0.7337
## F-statistic:  87.1 on 8 and 242 DF,  p-value: < 2.2e-16
```

Interaction

In the data exploration part, we found that the correlation between the variables are very significant. The predictors can interact with each other. Thus, let's check it again with our further selected data.

```
body_mod_select = bodyfat_selected %>%
  dplyr::select(bodyfat_brozek, age, weight, neck, abdomen, hip, thigh, forearm, wrist)
corrplot(cor(body_mod_select), type = "upper", diag = FALSE)
```



We can find that the correlation of the variables left are strong between each other, except age. Let's check their interactions.

```
lm.fit2 = lm(bodyfat_brozek ~ (age + weight + neck + abdomen + hip + thigh + forearm + wrist)^2,
             data = bodyfat_selected)
table = anova(lm.fit2)

inter_group = table %>%
  filter(`Pr(>F)` < 0.05)

inter_group
```

```
## Analysis of Variance Table
##
## Response: bodyfat_brozek
##               Df Sum Sq Mean Sq  F value    Pr(>F)
## age            1 1228.8  1228.8  81.1709 < 2.2e-16 ***
```

```
## weight        1 5479.4  5479.4 361.9375 < 2.2e-16 ***
## neck          1  152.5   152.5  10.0751  0.001724 **
## abdomen       1 3727.0  3727.0 246.1871 < 2.2e-16 ***
## thigh         1  112.3   112.3   7.4149  0.007002 **
## forearm       1   74.3    74.3   4.9073  0.027798 *
## wrist         1  143.3   143.3   9.4682  0.002364 **
## weight:neck   1  110.8   110.8   7.3156  0.007386 **
## neck:abdomen  1   62.2    62.2   4.1114  0.043834 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the table, based on the pvalue smaller than 0.05. We found that only the weight:neck and neck: abdomen's interaction are statistically significant. Thus, these two interactive groups would be included in the model.

The model with interaction between significant associated predictors is: **model_2 = lm(bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + wrist + weight:neck + neck:abdomen, data = bodyfat_selected)**

```
model_2 = lm(bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + wrist + weight*ne
summary(model_2)
```

```
##
## Call:
## lm(formula = bodyfat_brozek ~ age + weight + neck + abdomen +
##     hip + thigh + forearm + wrist + weight * neck + neck * abdomen,
##     data = bodyfat_selected)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2337 -2.5363 -0.2964  2.7402  9.3143
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.845207  39.141311   0.354 0.723857
## age           0.065526   0.028145   2.328 0.020735 *
## weight        0.641405   0.306566   2.092 0.037469 *
## neck         -1.155900   1.004728  -1.150 0.251099
## abdomen      -0.860353   0.957353  -0.899 0.369724
## hip          -0.150543   0.127789  -1.178 0.239941
## thigh         0.271825   0.120088   2.264 0.024496 *
## forearm       0.292488   0.181441   1.612 0.108270
## wrist        -1.607618   0.467833  -3.436 0.000695 ***
## weight:neck  -0.018230   0.007826  -2.329 0.020678 *
## neck:abdomen  0.044050   0.024734   1.781 0.076179 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.898 on 240 degrees of freedom
## Multiple R-squared:  0.7522, Adjusted R-squared:  0.7419
## F-statistic: 72.86 on 10 and 240 DF,  p-value: < 2.2e-16
```

```
broom::tidy(summary(model_2))
```

```
## # A tibble: 11 x 5
##    term          estimate std.error statistic  p.value
##    <chr>            <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)   13.8     39.1         0.354 0.724
##  2 age            0.0655   0.0281      2.33  0.0207
##  3 weight         0.641    0.307       2.09  0.0375
##  4 neck          -1.16     1.00       -1.15  0.251
##  5 abdomen       -0.860    0.957      -0.899 0.370
##  6 hip           -0.151    0.128      -1.18  0.240
##  7 thigh          0.272    0.120       2.26  0.0245
##  8 forearm        0.292    0.181       1.61  0.108
##  9 wrist         -1.61     0.468      -3.44  0.000695
## 10 weight:neck   -0.0182   0.00783    -2.33  0.0207
## 11 neck:abdomen   0.0441   0.0247      1.78  0.0762
```

Here, we got two candidate model. One without interactions: **model_1 = lm(bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + wrist, data = bodyfat_selected)**

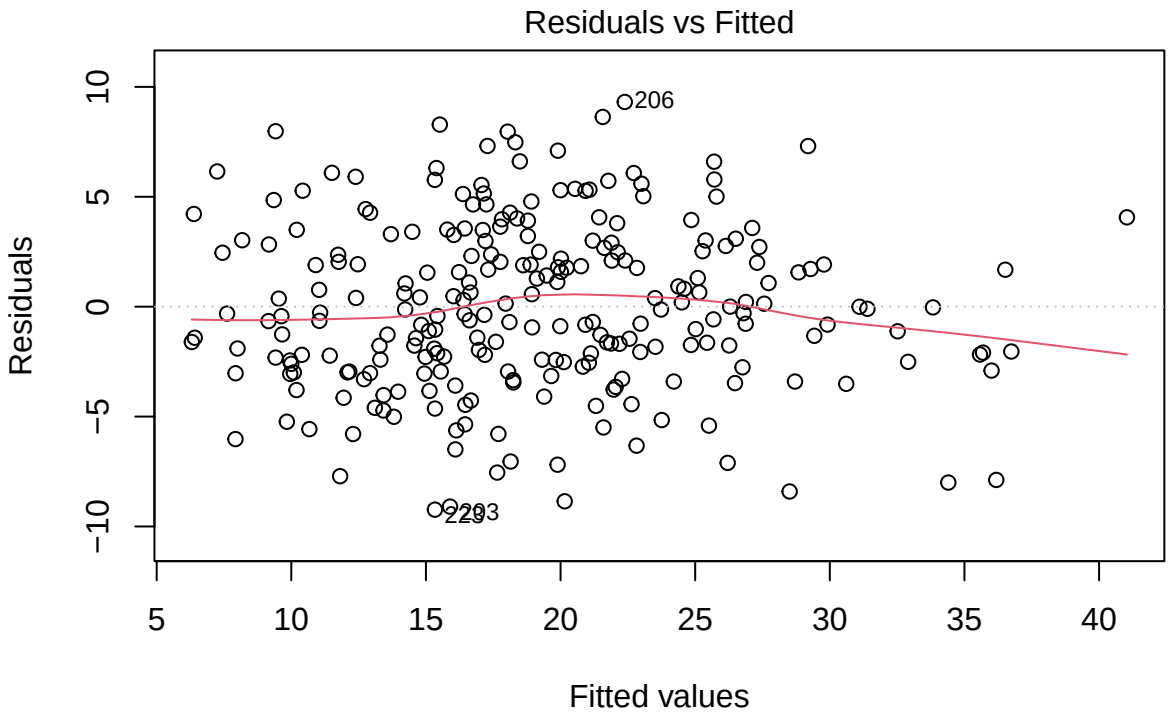
Another one with interactions: **model_2 = lm(bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + wrist + weight:neck + neck:abdomen, data = bodyfat_selected)**

Model Diagnostics

Rediduals vs Fitted plot

```
plot(model_1, which = 1)
```

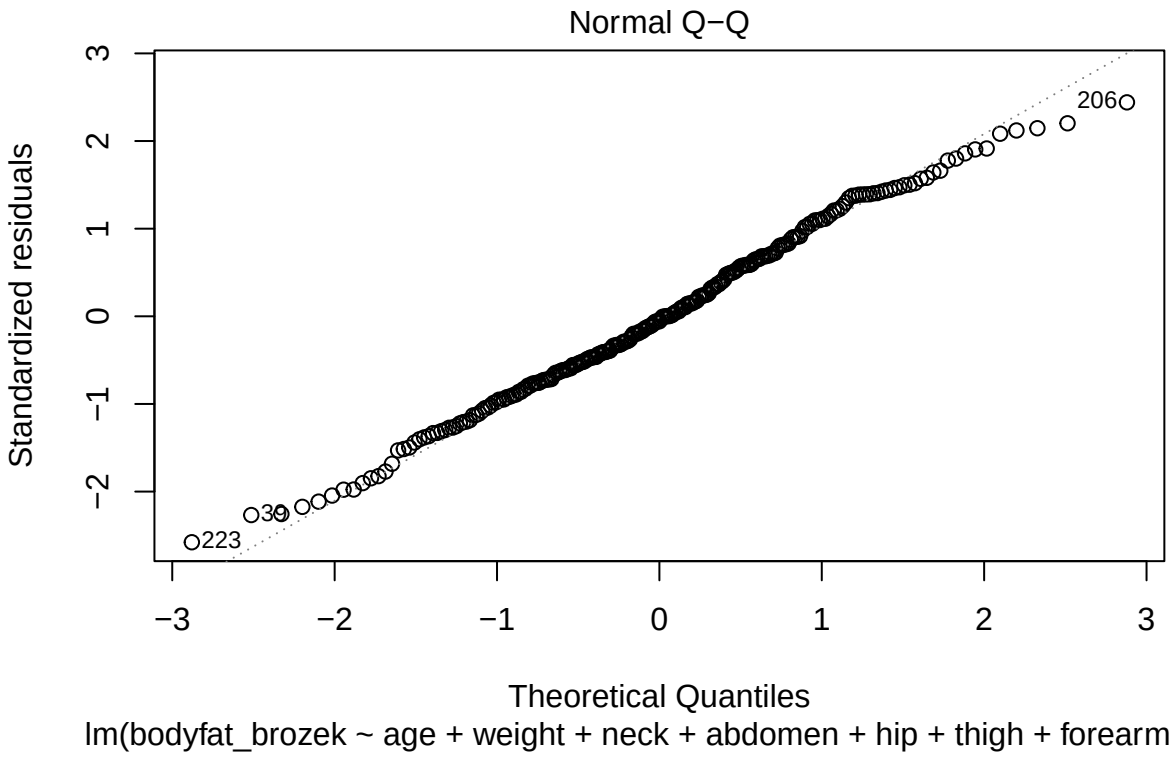

```
plot(model_2, which = 1)
```



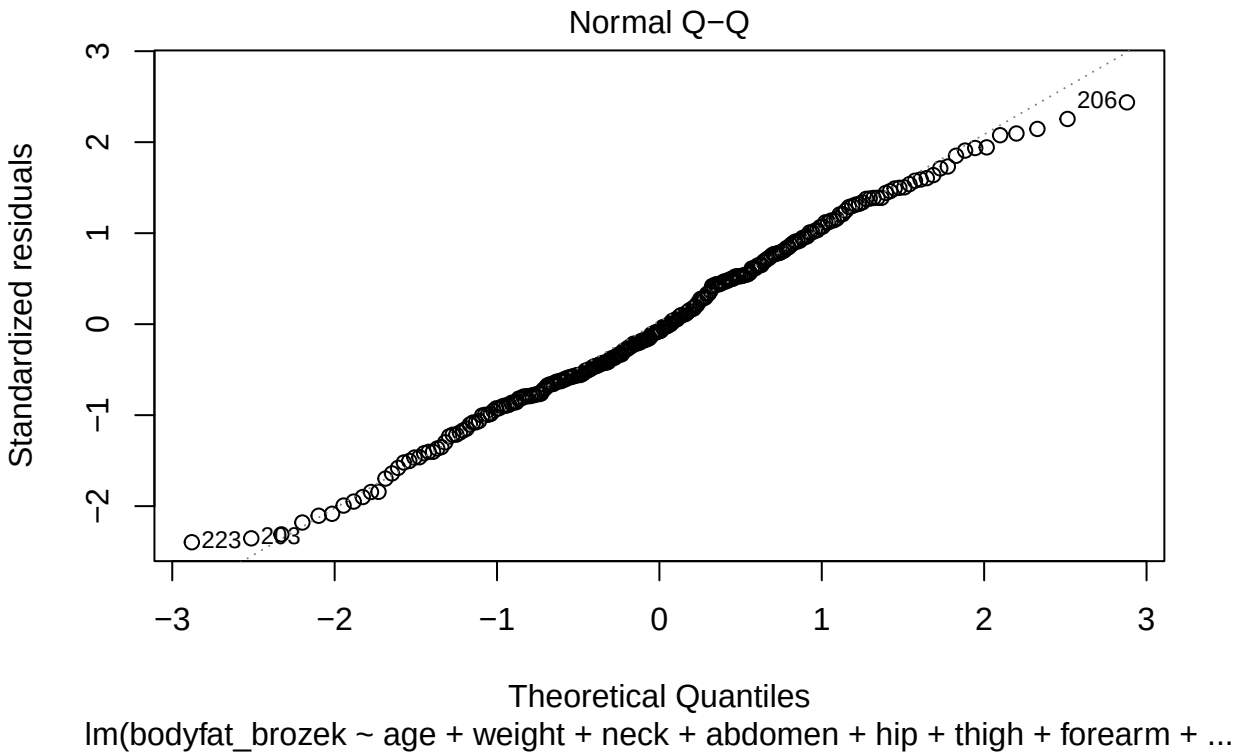
Horizontal bands around 0 are formed, indicating no heteroscedasticity and outliers were detected.

Normal QQ plot

```
plot(model_1, which = 2)
```



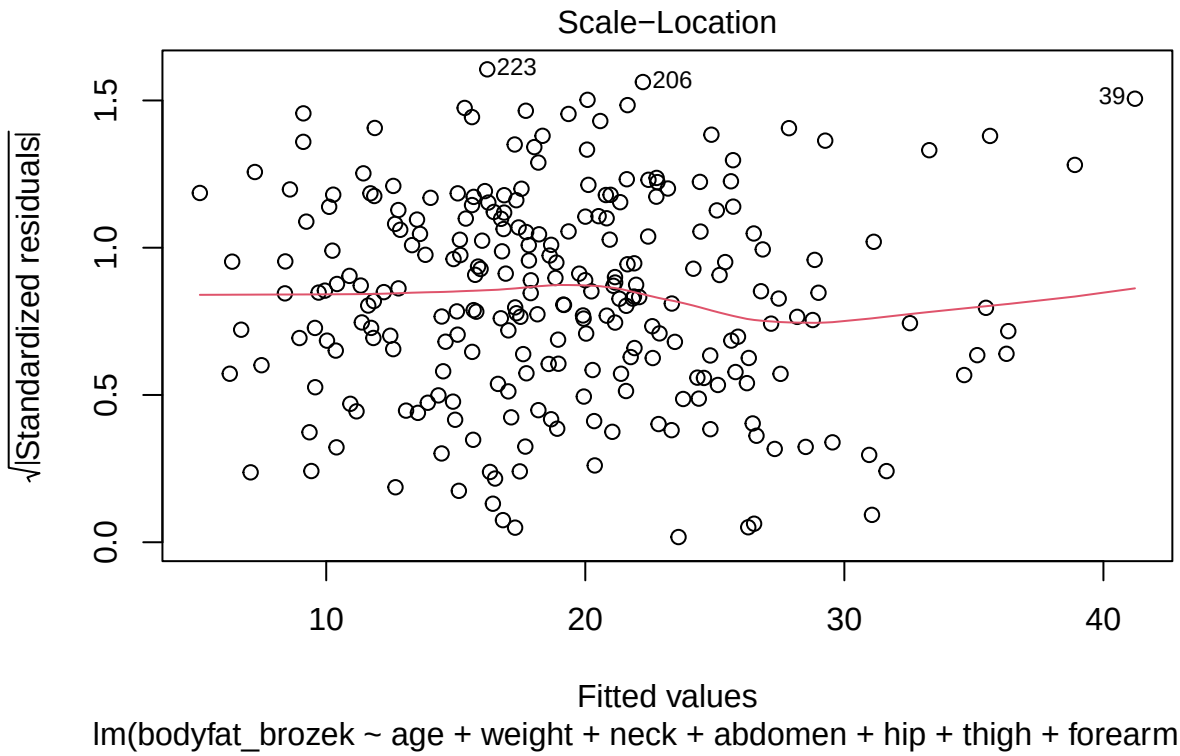

```
plot(model_2, which = 2)
```



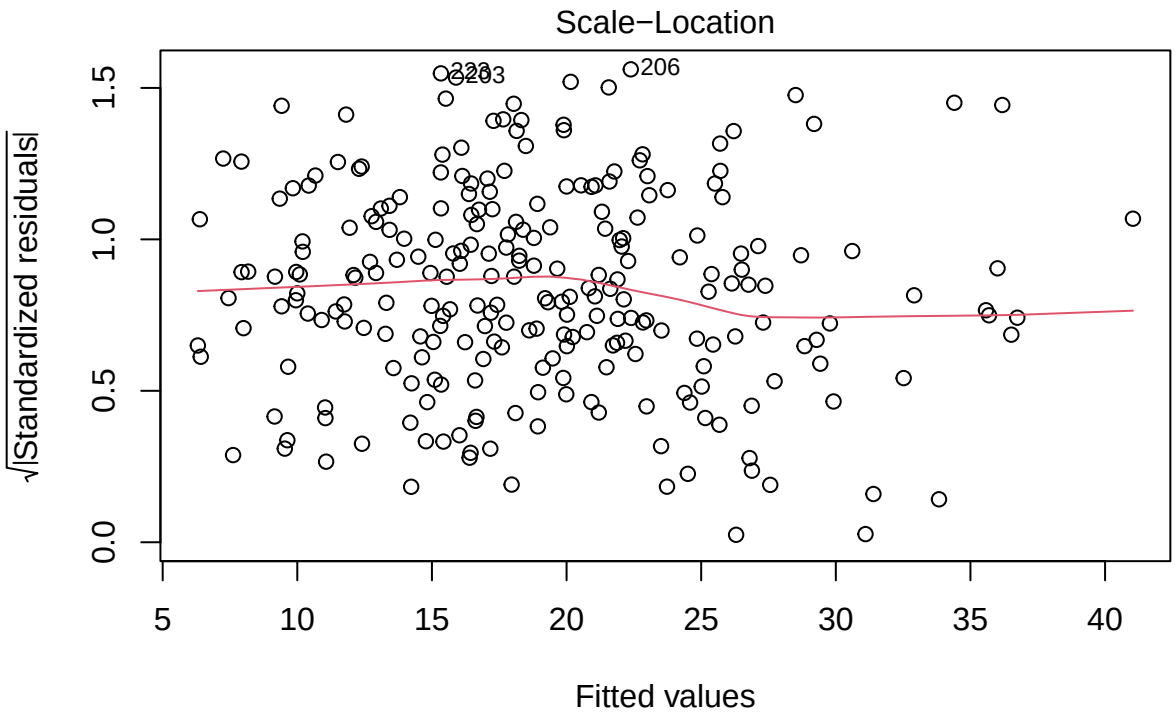
Straight lines were observed, indicating the residuals are normal.

Scale-Location plot

```
plot(model_1, which = 3)
```




```
plot(model_2, which = 3)
```

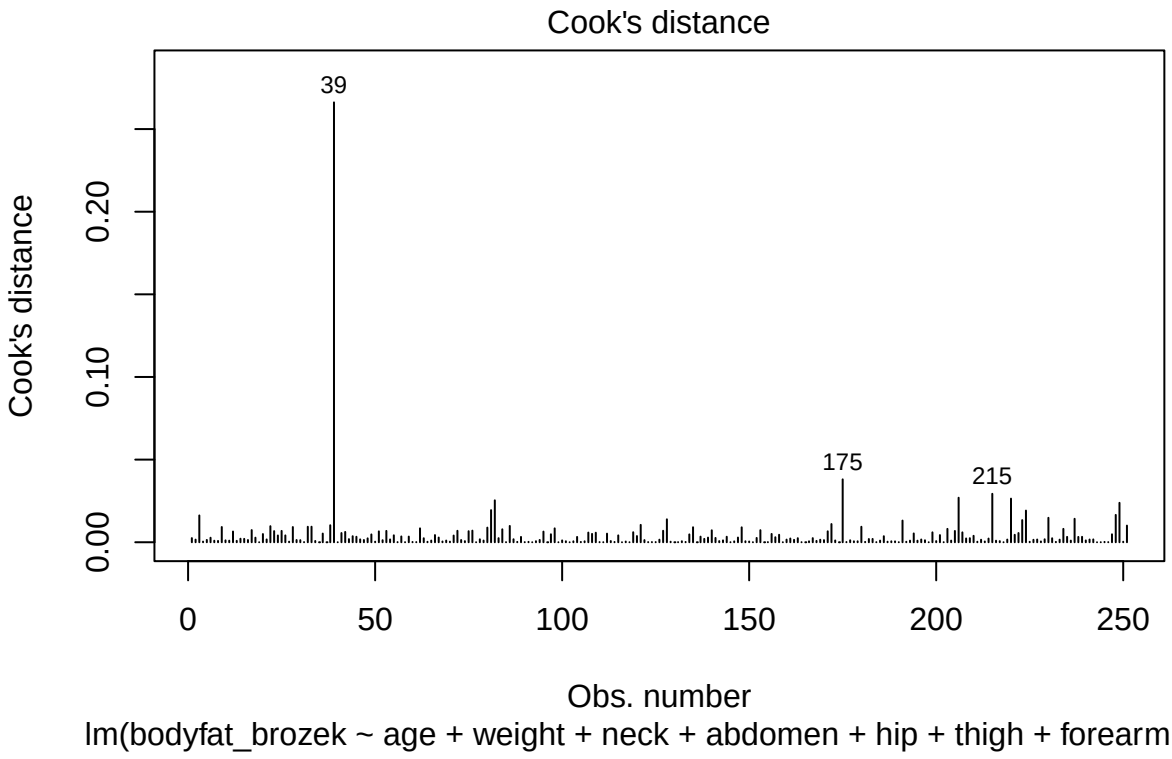


In each plot, the points were equally distributed with a horizontal line.

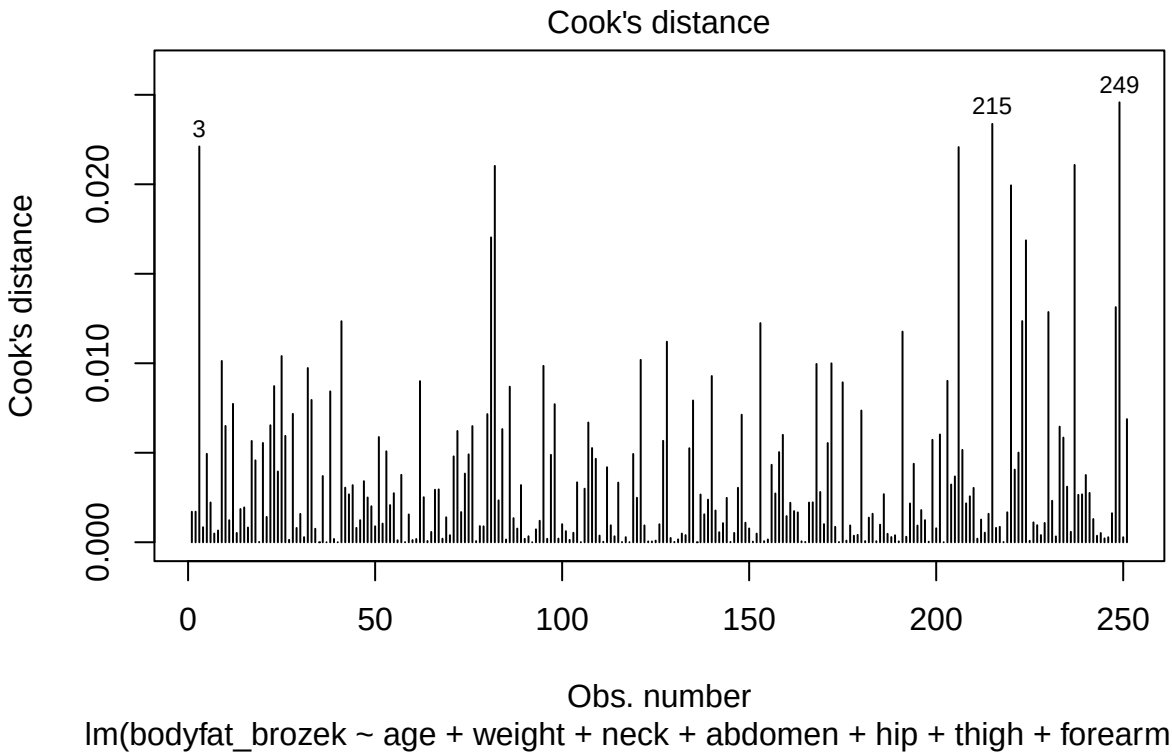
Outliers and Leverage

Cook's distance plot.

```
plot(model_1, which = 4)  #39, #175, #216
```




```
plot(model_2, which = 4)  #3, #216, #250
```



There are in total 251 samples used to build our models, so the rule of thumb for Cook's distance is D > 4/n = 0.016. There were some influential outliers were detected. Further process of outliers are required.

Let's remove the influential points.

```
bodyfat_out_1 = bodyfat_selected[-c(39,175,216),]
bodyfat_out_2 = bodyfat_selected[-c(3,216,250),]
```

Then, let's fit the model with and without influential points. **Check model_1**

```
with1 = lm(bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + wrist, data = bodyfa

without1 = lm(bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + wrist, data = bo

summary(with1)
```

```
##
## Call:
## lm(formula = bodyfat_brozek ~ age + weight + neck + abdomen +
##     hip + thigh + forearm + wrist, data = bodyfat_selected)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.117  -2.800  -0.223   2.709   9.477
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.00338   10.86380  -1.749  0.08152 .
## age           0.05827    0.02847   2.047  0.04175 *
## weight       -0.08266    0.03692  -2.239  0.02607 *
## neck         -0.42436    0.20780  -2.042  0.04222 *
## abdomen       0.87429    0.06656  13.136  < 2e-16 ***
## hip          -0.18514    0.12805  -1.446  0.14952
## thigh         0.27507    0.11966   2.299  0.02237 *
## forearm       0.47002    0.17257   2.724  0.00693 **
## wrist        -1.42508    0.47132  -3.024  0.00277 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.96 on 242 degrees of freedom
## Multiple R-squared:  0.7422, Adjusted R-squared:  0.7337
## F-statistic:  87.1 on 8 and 242 DF,  p-value: < 2.2e-16
```

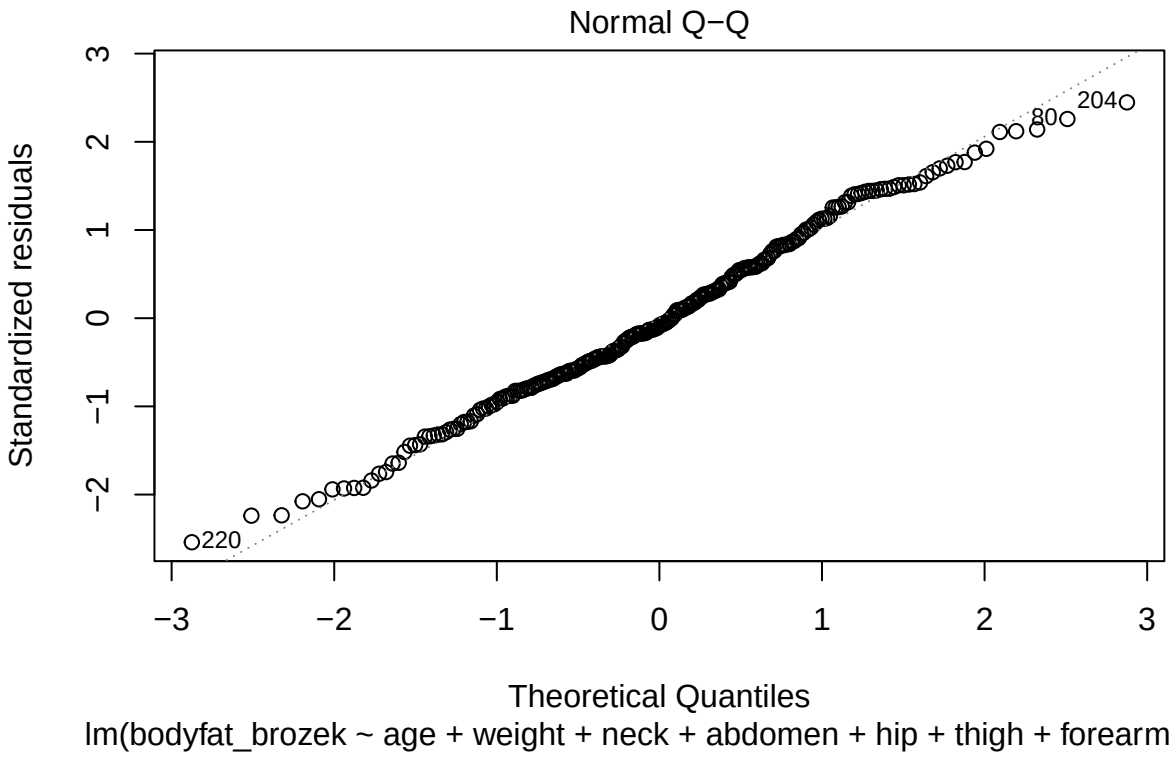
```
summary(without1)
```

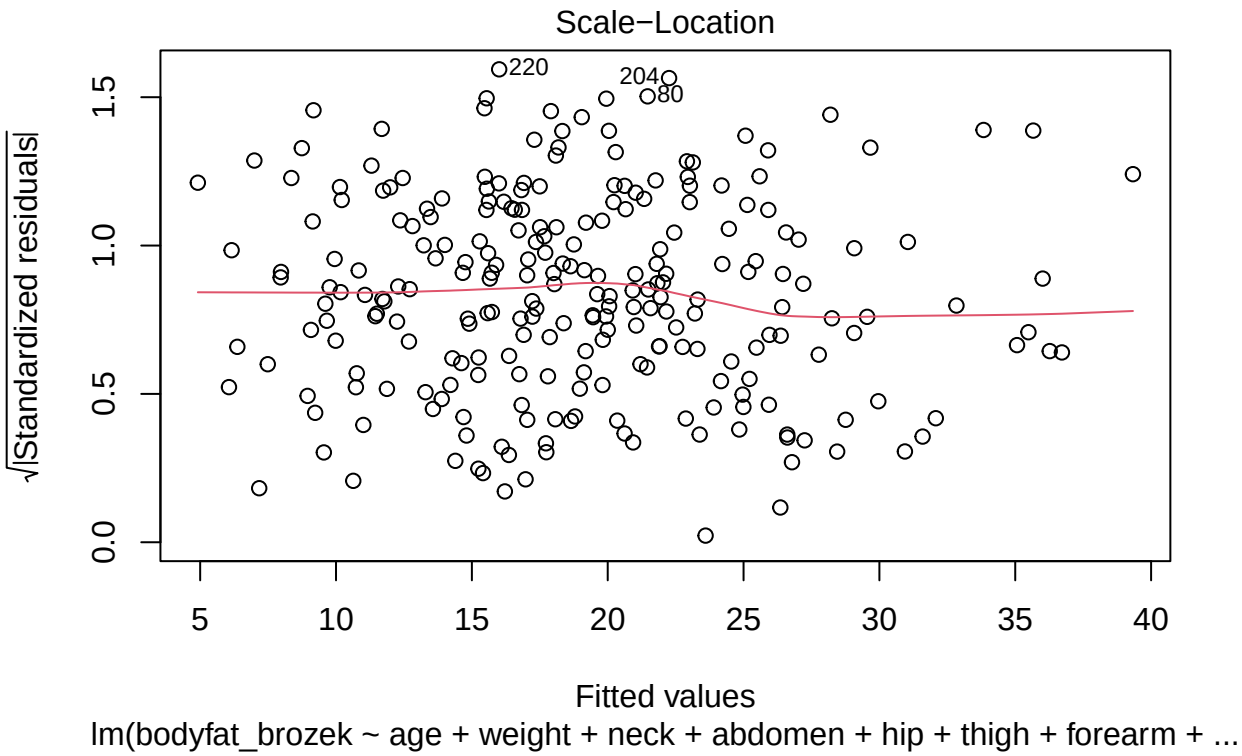
```
##
## Call:
## lm(formula = bodyfat_brozek ~ age + weight + neck + abdomen +
##     hip + thigh + forearm + wrist, data = bodyfat_out_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9050 -2.6794 -0.3214  2.6927  9.4414
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.72718   10.84731  -1.819 0.070220 .
## age           0.06484    0.02866   2.262 0.024580 *
## weight       -0.07015    0.03709  -1.891 0.059803 .
## neck         -0.34600    0.21220  -1.631 0.104301
## abdomen       0.84556    0.06779  12.473  < 2e-16 ***
## hip          -0.13902    0.13099  -1.061 0.289622
## thigh         0.25536    0.11955   2.136 0.033702 *
## forearm       0.39693    0.20854   1.903 0.058188 .
## wrist        -1.61427    0.48206  -3.349 0.000943 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.937 on 239 degrees of freedom
## Multiple R-squared:  0.7435, Adjusted R-squared:  0.7349
## F-statistic: 86.58 on 8 and 239 DF,  p-value: < 2.2e-16
```

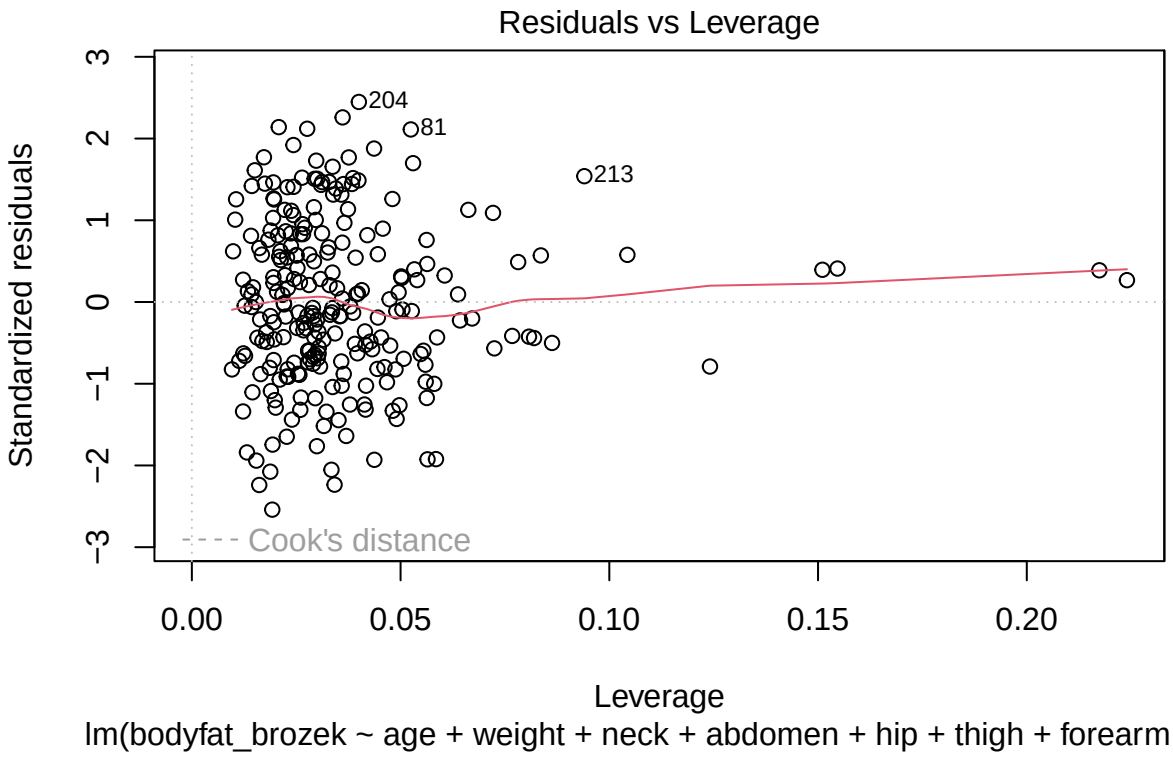
check without1 diagnostics

```
plot(without1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + ...
204
201
220
10
5
0
−5
−10
5
10
15
20
25
30
35
40


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + ...
220
80
204
−3
−2
−1
0
1
2
3


Scale−Location
220
204
80
√|Standardized residuals|
0.0
0.5
1.0
1.5
5
10
15
20
25
30
35
40
Fitted values
lm(bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + ...



check model 2

```
with2 = lm(bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + wrist + weight*neck

without2 = lm(bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + wrist + weight*ne

summary(with2)
```

```
##
## Call:
## lm(formula = bodyfat_brozek ~ age + weight + neck + abdomen +
##     hip + thigh + forearm + wrist + weight * neck + neck * abdomen,
##     data = bodyfat_selected)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.2337 -2.5363 -0.2964  2.7402  9.3143
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.845207  39.141311   0.354 0.723857
## age           0.065526   0.028145   2.328 0.020735 *
## weight        0.641405   0.306566   2.092 0.037469 *
## neck         -1.155900   1.004728  -1.150 0.251099
## abdomen      -0.860353   0.957353  -0.899 0.369724
## hip          -0.150543   0.127789  -1.178 0.239941
```

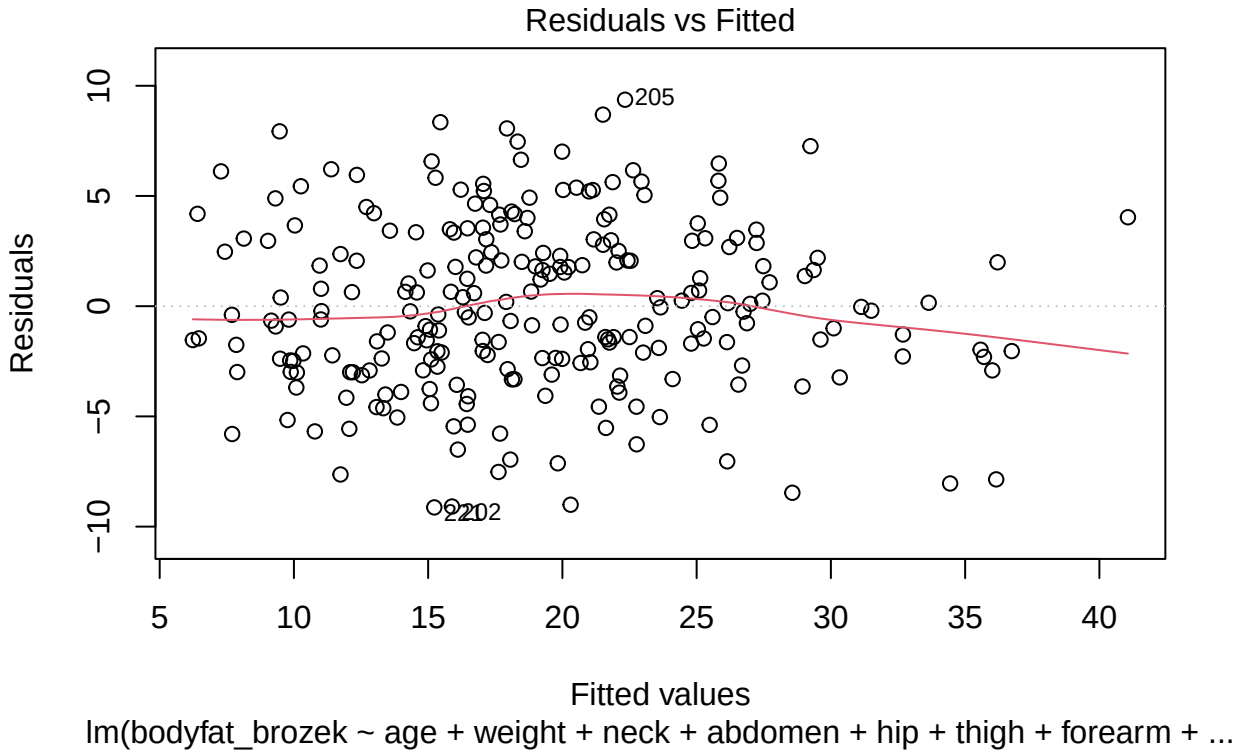
```
## thigh         0.271825   0.120088   2.264 0.024496 *
## forearm       0.292488   0.181441   1.612 0.108270
## wrist        -1.607618   0.467833  -3.436 0.000695 ***
## weight:neck  -0.018230   0.007826  -2.329 0.020678 *
## neck:abdomen  0.044050   0.024734   1.781 0.076179 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.898 on 240 degrees of freedom
## Multiple R-squared:  0.7522, Adjusted R-squared:  0.7419
## F-statistic: 72.86 on 10 and 240 DF,  p-value: < 2.2e-16
```

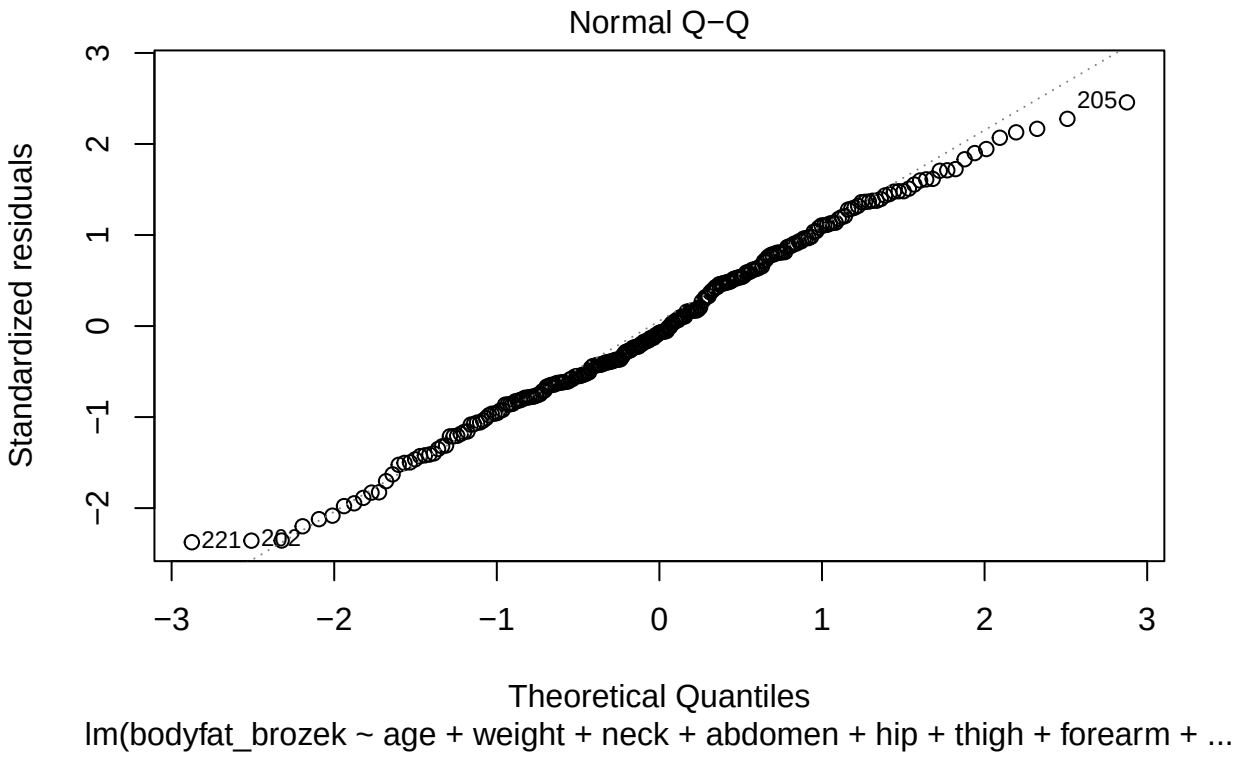
```
summary(without2)
```

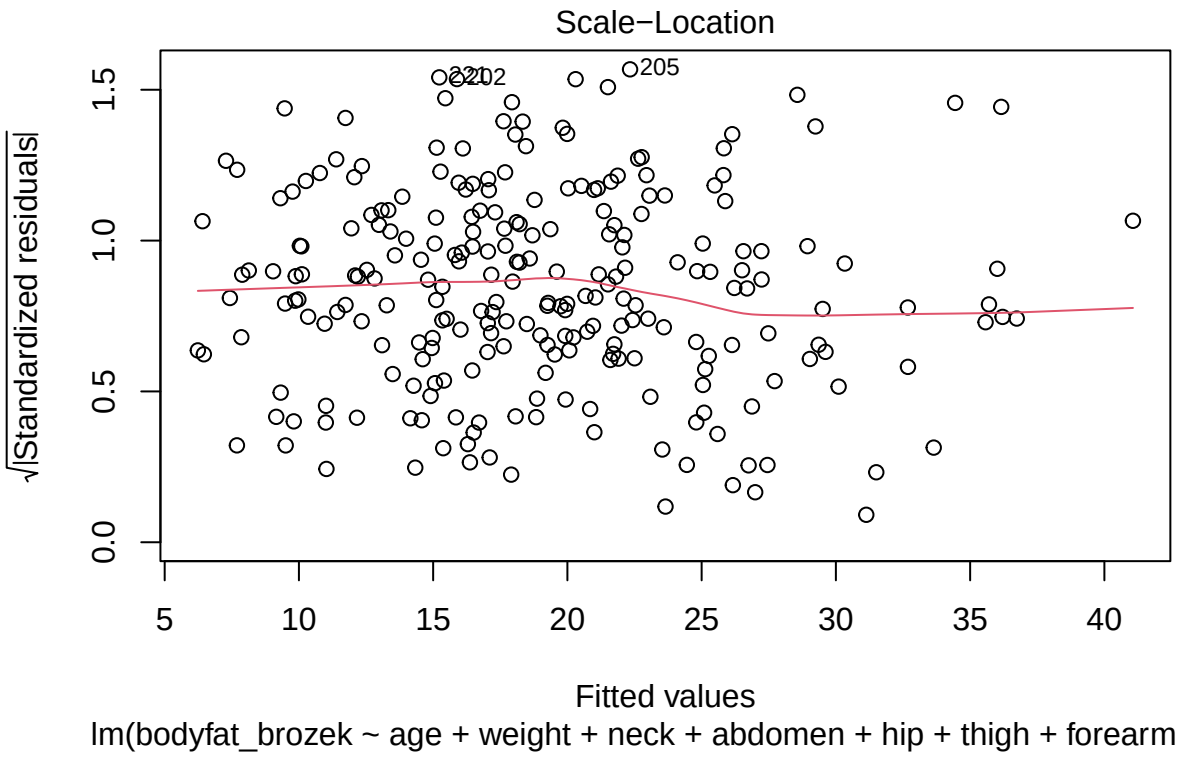
```
##
## Call:
## lm(formula = bodyfat_brozek ~ age + weight + neck + abdomen +
##     hip + thigh + forearm + wrist + weight * neck + neck * abdomen,
##     data = bodyfat_out_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1282 -2.4949 -0.2874  2.8912  9.3663
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.853295  39.208651   0.506 0.613082
## age           0.071342   0.028341   2.517 0.012489 *
## weight        0.705852   0.307935   2.292 0.022771 *
## neck         -1.286374   1.006519  -1.278 0.202485
## abdomen      -1.049332   0.961528  -1.091 0.276242
## hip          -0.167534   0.128919  -1.300 0.195027
## thigh         0.275772   0.119969   2.299 0.022394 *
## forearm       0.314021   0.181339   1.732 0.084633 .
## wrist        -1.639190   0.469204  -3.494 0.000569 ***
## weight:neck  -0.019748   0.007855  -2.514 0.012599 *
## neck:abdomen  0.048699   0.024823   1.962 0.050948 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.889 on 237 degrees of freedom
## Multiple R-squared:  0.7548, Adjusted R-squared:  0.7445
## F-statistic: 72.96 on 10 and 237 DF,  p-value: < 2.2e-16
```

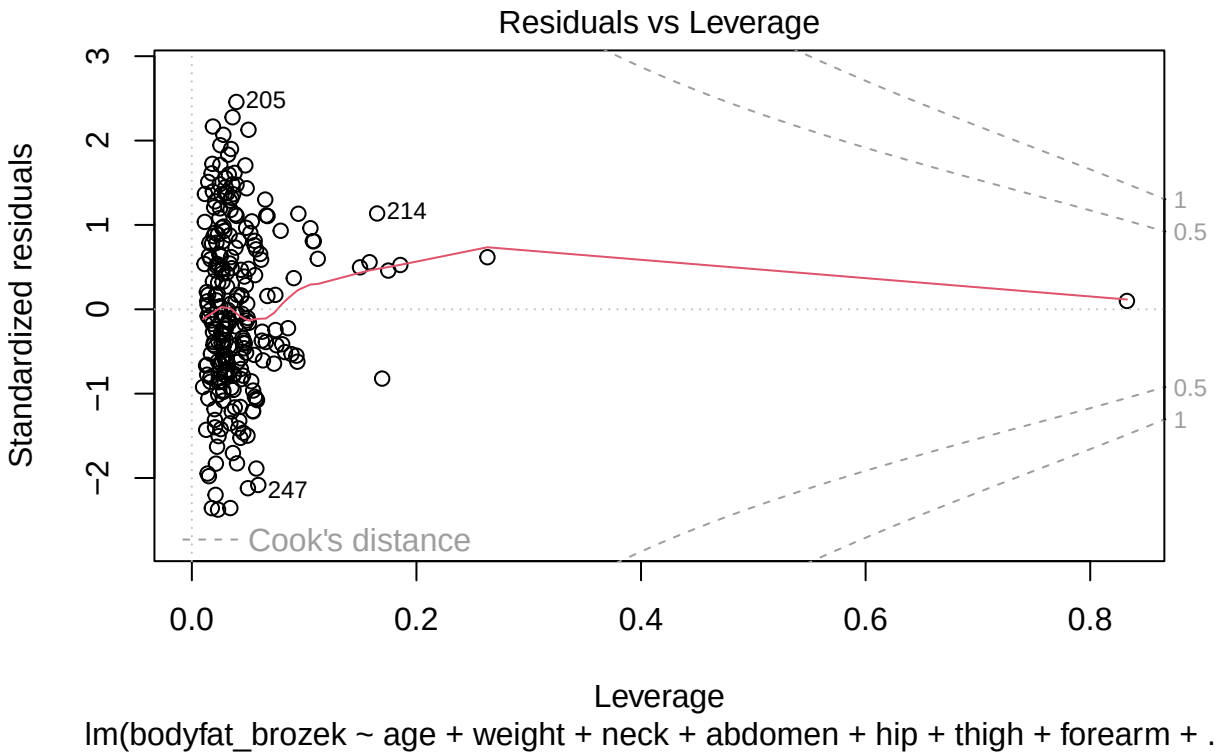
check without2 diagnostics

```
plot(without2)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + ...
221
202
205
−3
−2
−1
0
1
2
3

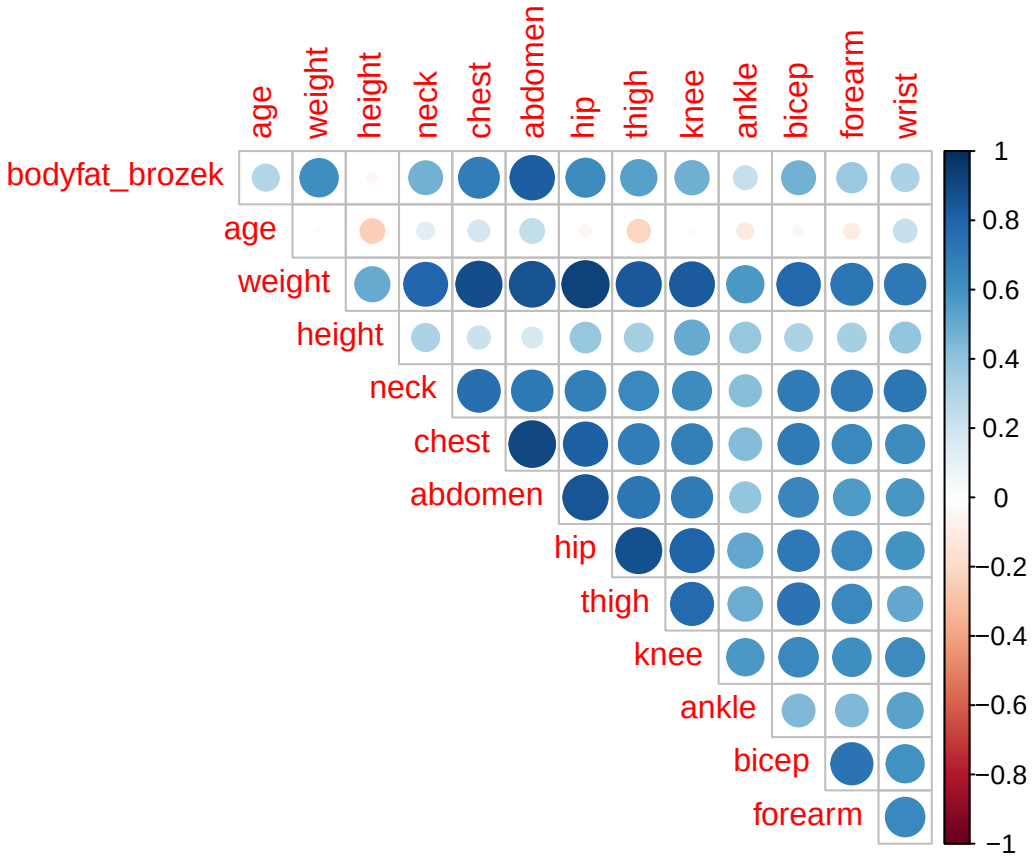

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
5
10
15
20
25
30
35
40
Fitted values
lm(bodyfat_brozek ~ age + weight + neck + abdomen + hip + thigh + forearm + ...



Both plots without influential point showed to better than those with influential points.

Assessing Multicollinearity

```
corrplot(cor(bodyfat_out_1), type = "upper", diag = FALSE)
```



Still, even after excluding the influential data, all the predictors, except age, are highly correlated. Let's check the model without interactions terms.

```
check_collinearity(without1)
```

```
## # Check for Multicollinearity
##
## Low Correlation
##
##     Term   VIF     VIF 95% CI Increased SE Tolerance Tolerance 95% CI
##      age  2.11   [1.78, 2.56]         1.45      0.48     [0.39, 0.56]
##  forearm  2.65   [2.21, 3.24]         1.63      0.38     [0.31, 0.45]
##   weight 15.62 [12.43, 19.69]         3.95      0.06     [0.05, 0.08]
##      hip 11.23 [ 8.97, 14.12]         3.35      0.09     [0.07, 0.11]
##
## Moderate Correlation
##
##   Term  VIF   VIF 95% CI Increased SE Tolerance Tolerance 95% CI
##  wrist 3.03 [2.51, 3.73]         1.74      0.33     [0.27, 0.40]
##  thigh 5.41 [4.38, 6.74]         2.32      0.18     [0.15, 0.23]
##
## High Correlation
##
##     Term  VIF   VIF 95% CI Increased SE Tolerance Tolerance 95% CI
##     neck 3.68 [3.03, 4.56]         1.92      0.27     [0.22, 0.33]
##  abdomen 7.46 [6.00, 9.34]         2.73      0.13     [0.11, 0.17]
```

It's found that, there VIF value is very high for weight, hip and abdomen terms, indicating their strong correlation with other variables.

Let's remove the weight term with highest VIF(15.62).

```
without1_new = lm(bodyfat_brozek ~ age + neck + abdomen + hip + thigh + forearm + wrist, data = bodyfat_
check_collinearity(without1_new)
```

```
## # Check for Multicollinearity
##
## Low Correlation
##
##   Term  VIF    VIF 95% CI Increased SE Tolerance Tolerance 95% CI
##    age 1.96  [1.67, 2.38]         1.40      0.51     [0.42, 0.60]
##   neck 3.37  [2.77, 4.16]         1.84      0.30     [0.24, 0.36]
##    hip 8.64 [6.92, 10.84]         2.94      0.12     [0.09, 0.14]
##  thigh 5.37 [4.35,  6.70]         2.32      0.19     [0.15, 0.23]
##
## Moderate Correlation
##
##     Term  VIF    VIF 95% CI Increased SE Tolerance Tolerance 95% CI
##  forearm 2.58  [2.15, 3.16]         1.61      0.39     [0.32, 0.46]
##    wrist 2.65  [2.21, 3.25]         1.63      0.38     [0.31, 0.45]
##  abdomen 6.21 [5.01,  7.77]         2.49      0.16     [0.13, 0.20]
```

The correlation between the variables is lowered, but we still have hip(8.64), thigh(5.37) and abdomen(6.21) with high VIF. Then we remove the hip term.

```
without1_new_2 = lm(bodyfat_brozek ~ age  + neck + abdomen + thigh + forearm + wrist, data = bodyfat_ou
check_collinearity(without1_new_2)
```

```
## # Check for Multicollinearity
##
## Low Correlation
##
##     Term  VIF   VIF 95% CI Increased SE Tolerance Tolerance 95% CI
##      age 1.86 [1.59, 2.25]         1.36      0.54     [0.44, 0.63]
##     neck 3.33 [2.74, 4.12]         1.83      0.30     [0.24, 0.36]
##  abdomen 3.82 [3.12, 4.73]         1.95      0.26     [0.21, 0.32]
##    thigh 3.74 [3.07, 4.64]         1.93      0.27     [0.22, 0.33]
##  forearm 2.57 [2.14, 3.15]         1.60      0.39     [0.32, 0.47]
##    wrist 2.54 [2.12, 3.11]         1.59      0.39     [0.32, 0.47]
```

After removing these two variables, we have got all predictors with low VIF.

The final model we will have with and without interaction would be.

```
model_1f = lm(bodyfat_brozek ~ age  + neck + abdomen + thigh + forearm + wrist, data = bodyfat_out_1)
summary(model_1f)
```

```
##
## Call:
## lm(formula = bodyfat_brozek ~ age + neck + abdomen + thigh +
```

```
##     forearm + wrist, data = bodyfat_out_1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7338 -2.7938 -0.1265  2.7111 10.0812
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.77652    5.56083  -1.758 0.079999 .
## age          0.09324    0.02729   3.416 0.000745 ***
## neck        -0.41559    0.20453  -2.032 0.043258 *
## abdomen      0.70680    0.04914  14.383  < 2e-16 ***
## thigh        0.09021    0.10085   0.894 0.371962
## forearm      0.30520    0.20829   1.465 0.144155
## wrist       -2.14698    0.44709  -4.802 2.76e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.991 on 241 degrees of freedom
## Multiple R-squared:  0.7342, Adjusted R-squared:  0.7276
## F-statistic:   111 on 6 and 241 DF,  p-value: < 2.2e-16
```

```
model_2f = lm(bodyfat_brozek ~ age  + neck + abdomen + thigh + forearm + wrist + neck*abdomen, data = bo
summary(model_2f)
```

```
##
## Call:
## lm(formula = bodyfat_brozek ~ age + neck + abdomen + thigh +
##     forearm + wrist + neck * abdomen, data = bodyfat_out_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8236 -2.7998 -0.2458  2.5657  9.9140
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -72.677880  20.896807  -3.478  0.00060 ***
## age            0.090281   0.027192   3.320  0.00104 **
## neck           1.225430   0.581292   2.108  0.03606 *
## abdomen        1.393665   0.223466   6.237 2.00e-09 ***
## thigh          0.062708   0.099650   0.629  0.52976
## forearm        0.300439   0.180493   1.665  0.09731 .
## wrist         -2.053941   0.436674  -4.704 4.32e-06 ***
## neck:abdomen  -0.017789   0.005543  -3.209  0.00151 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.982 on 240 degrees of freedom
## Multiple R-squared:  0.7398, Adjusted R-squared:  0.7322
## F-statistic: 97.46 on 7 and 240 DF,  p-value: < 2.2e-16
```

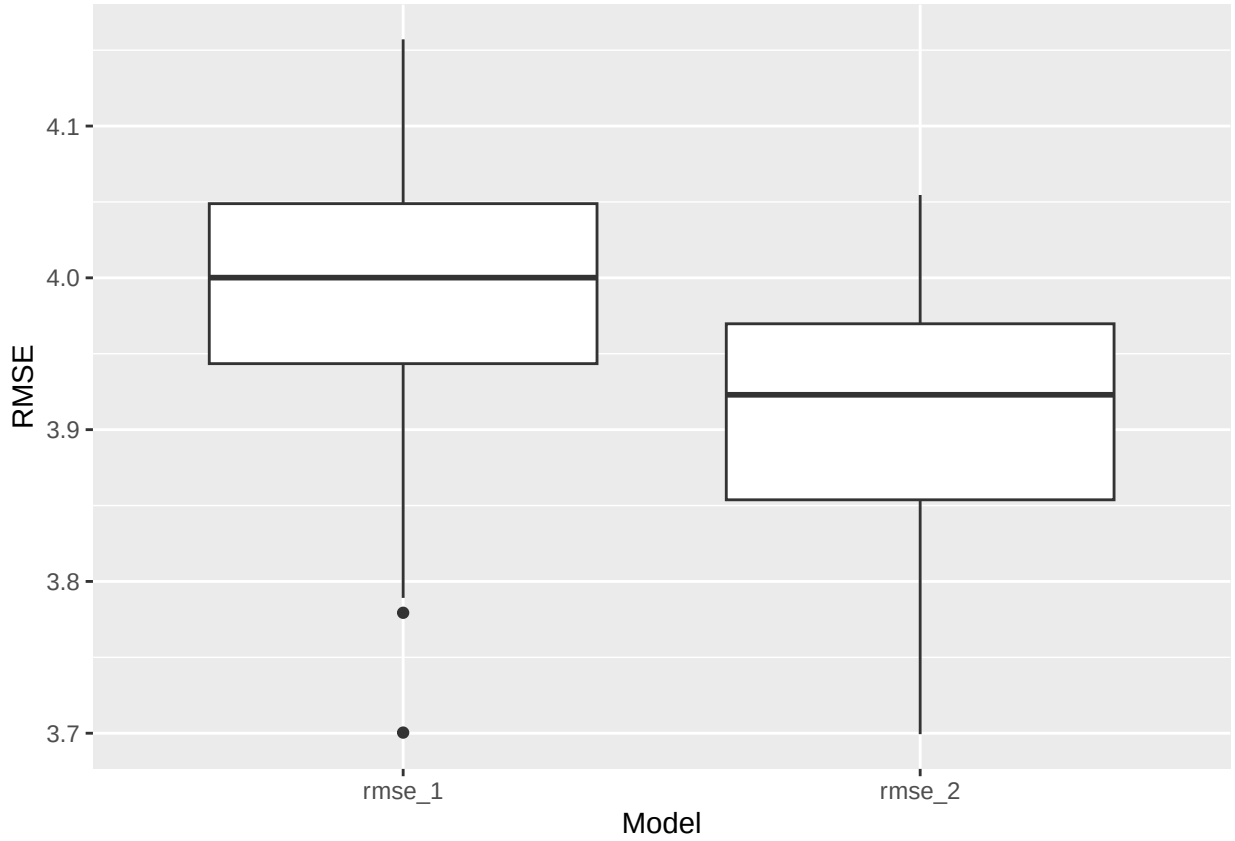
Cross validation

Let's create RMSE distribution boxplot

```
set.seed(1)
cv_df = crossv_mc(bodyfat_selected, 100) %>%
  mutate(train = map(train, as_tibble),
         test = map(test, as_tibble)) %>%
  mutate(
    mod_1 = map(.x = train, ~lm(bodyfat_brozek ~ age  + neck + abdomen + thigh + forearm + wrist, data =
    mod_2 = map(.x = train, ~lm(bodyfat_brozek ~ age  + neck + abdomen + thigh + forearm + wrist + neck
  ) %>%
  mutate(
    rmse_1 = map2_dbl(.x = mod_1, .y = test, ~rmse(model = .x)),
    rmse_2 = map2_dbl(.x = mod_2, .y = test, ~rmse(model = .x))
  )

rmse_boxplot = cv_df %>%
  dplyr::select(rmse_1, rmse_2) %>%
  pivot_longer(rmse_1:rmse_2,
               names_to = "model",
               values_to = "rmse",
               names_prefix = "remse_") %>%
  ggplot(aes(x = model, y = rmse)) + geom_boxplot() +
  labs(x = "Model", y = "RMSE")

rmse_boxplot
```



By plotting the RMSE distribution of two model, we can find that the interaction model has the lower RMSE, which is more prefered.

```
set.seed(1)
train = trainControl(method = "cv", number = 10)

mod_1_cv = train(bodyfat_brozek ~ age  + neck + abdomen + thigh + forearm + wrist, data = bodyfat_out_1

mod_2_cv = train(bodyfat_brozek ~ age  + neck + abdomen + thigh + forearm + wrist + neck*abdomen, data =

RMSE = bind_rows(
  as_tibble(mod_1_cv$results),
  as_tibble(mod_2_cv$results)
)
RMSE %>%
  knitr::kable(digit = 3)
```

intercept	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
TRUE	4.019	0.732	3.289	0.329	0.068	0.254
TRUE	3.989	0.736	3.268	0.563	0.056	0.590

Model 2's RMSE(3.989) is slightly smaller than that of model 1(4.019).

Finally, based on the data analysis, the final model we would choose to is the model 2 with interactions terms.

The final model would be bodyfat_brozek = -72.7 + 0.090*age* + *1.23*neck + 1.39*abdomen* + *0.0627*thigh + 0.300*forearm* -*2.05*wrist - 0.0178*neck:abdomen

```
final_model = lm(bodyfat_brozek ~ age  + neck + abdomen + thigh + forearm + wrist + neck*abdomen, data =
summary(final_model)
```

```
##
## Call:
## lm(formula = bodyfat_brozek ~ age + neck + abdomen + thigh +
##     forearm + wrist + neck * abdomen, data = bodyfat_out_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8236 -2.7998 -0.2458  2.5657  9.9140
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -72.677880  20.896807  -3.478  0.00060 ***
## age            0.090281   0.027192   3.320  0.00104 **
## neck           1.225430   0.581292   2.108  0.03606 *
## abdomen        1.393665   0.223466   6.237 2.00e-09 ***
## thigh          0.062708   0.099650   0.629  0.52976
## forearm        0.300439   0.180493   1.665  0.09731 .
## wrist         -2.053941   0.436674  -4.704 4.32e-06 ***
## neck:abdomen  -0.017789   0.005543  -3.209  0.00151 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.982 on 240 degrees of freedom
## Multiple R-squared:  0.7398, Adjusted R-squared:  0.7322
## F-statistic: 97.46 on 7 and 240 DF,  p-value: < 2.2e-16
```

```
table_2= data.frame(broom::tidy(summary(final_model)))
table_2 %>%
  knitr::kable(digit = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-72.678	20.897	-3.478	0.001
age	0.090	0.027	3.320	0.001
neck	1.225	0.581	2.108	0.036
abdomen	1.394	0.223	6.237	0.000
thigh	0.063	0.100	0.629	0.530
forearm	0.300	0.180	1.665	0.097
wrist	-2.054	0.437	-4.704	0.000
neck:abdomen	-0.018	0.006	-3.209	0.002