Report I - SF2930 Regression Analysis

Chen Hanmo, 990904-T072 Shen Linying, 981208-T066

9 March, 2019

Contents

Project 1

Scenario I: Body fat assessment

1 Introduction

1.1 Background

The World Health organization (WHO) reported that obesity is a major risk factor for a number of chronic diseases, including diabetes, cardiovascular diseases and cancer. Obesity is defined as ”the disease in which excess of body fat has accumulated to such extend that health may be adversely affected”. Once being considered as a problem only for high income countries, obesity is now rise in low- and middle-income countries. An important issue for medical purposes is thus is to reliably identify people with the fat excess.

1.2 Goals

As a major risk for many chronic diseases, obesity can be influenced by many factors including one’s age, height and other body indexes.

In the dataset from http://lib.stat.cmu.edu/datasets/bodyfat, it uses *body fat mass* (BFM) instead of the *body mass index*(BMI) as the measure of body fatness and other 13 indexs are included as factors that may influence the fatness.

In this project, we are trying to develop a regression model based on the body dataset and use statistical techniques to modify and validate it.

2 Model development

We establish and modify our model following the flow chart[1] presented in Montgomery’s book. Although our model-building process covers many other aspects, we will mainly focus on the **residual analysis** and **multicollinearity diagnostics** in the following illustration.

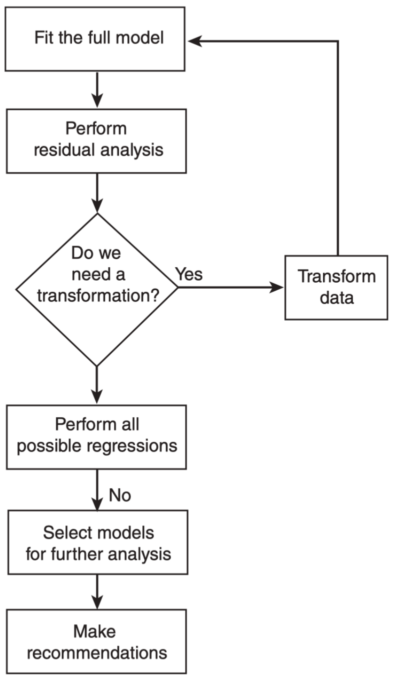
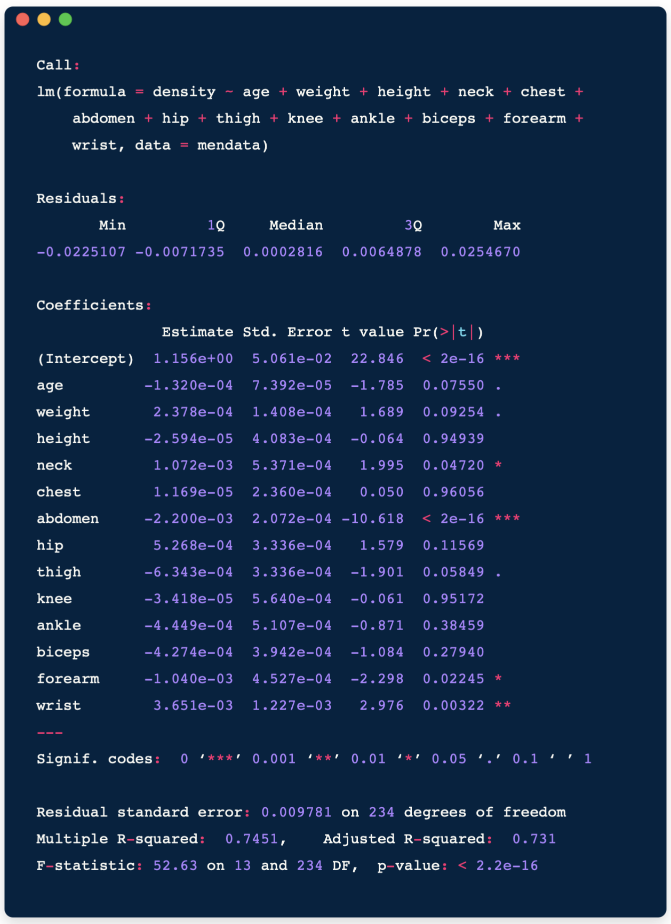

Fit the full model
Perform residual analysis
Do we need a transformation?
Yes
Transform data
Perform all possible regressions
No
Select models for further analysis
Make recommendations

Figure 1: the model-building process

2.1 The full model

Firstly, we try to establish the full model with all the columns in the data frame. And the summary of the model is presented as [2].



```
Call:
lm(formula = density ~ age + weight + height + neck + chest +
    abdomen + hip + thigh + knee + ankle + biceps + forearm +
    wrist, data = mendata)

Residuals:
      Min         1Q     Median         3Q        Max
-0.0225107 -0.0071735  0.0002816  0.0064878  0.0254670

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.156e+00  5.061e-02  22.846  < 2e-16 ***
age         -1.320e-04  7.392e-05  -1.785  0.07550 .
weight       2.378e-04  1.408e-04   1.689  0.09254 .
height      -2.594e-05  4.083e-04  -0.064  0.94939
neck         1.072e-03  5.371e-04   1.995  0.04720 *
chest        1.169e-05  2.360e-04   0.050  0.96056
abdomen     -2.200e-03  2.072e-04 -10.618  < 2e-16 ***
hip          5.268e-04  3.336e-04   1.579  0.11569
thigh       -6.343e-04  3.336e-04  -1.901  0.05849 .
knee        -3.418e-05  5.640e-04  -0.061  0.95172
ankle       -4.449e-04  5.107e-04  -0.871  0.38459
biceps      -4.274e-04  3.942e-04  -1.084  0.27940
forearm     -1.040e-03  4.527e-04  -2.298  0.02245 *
wrist        3.651e-03  1.227e-03   2.976  0.00322 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.009781 on 234 degrees of freedom
Multiple R-squared:  0.7451,    Adjusted R-squared:  0.731
F-statistic: 52.63 on 13 and 234 DF,  p-value: < 2.2e-16
```

Figure 2: Summary of the full model

The original full model is

$$\begin{aligned}
\text{density} =& 1.156121 - 1.319699 \times 10^{-4} * \text{age} \ + 2.377541 \times 10^{-4} * \text{weight} \\
& - 2.594465 \times 10^{-5} * \text{height} + 1.071540 \times 10^{-3} * \text{neck} + 1.168530 \times 10^{-5} * \text{chest} \\
& - 2.199601 \times 10^{-3} * \text{abdomen} + 5.267858 \times 10^{-4} * \text{hip} - 6.342697 \times 10^{-4} * \text{thigh} \\
& - 3.418355 \times 10^{-5} * \text{knee} - 4.448680 \times 10^{-4} * \text{ankle} - 4.273484 \times 10^{-4} * \text{biceps} \\
& - 1.040264 \times 10^{-3} * \text{forearm} + 3.651081 \times 10^{-3} * \text{wrist} \qquad (1)
\end{aligned}$$

Based on that result, it can be seen that the full model is not so satisfying. For example, the **R-square** and the **p-value** of many coefficients is not approriate for a validated model and further modification is needed.

2.2 Residual Analysis

Residual analysis is essential in our model-building to check model adequecy.

2.2.1 Standardized Residuals

First, we re-scale the residuals to the standardized values

$$d_i = \frac{e_i}{\sqrt{MS_{res}}}$$

and following [3] is the histogram of the standardized residuals.

According to the histogram, the standardized residuals all fall into the interval of $[-3, 3]$, but do not perfectly correspond with the normal distribution, which needs further inspection(as we did in the next section).

2.2.2 Normal Probability Plot

To check the normality assumption we constucted the normal probability plot of the residuals as follow [4]

As is seen in the graph, there is small departures from the normality. And the heavy-tailed error distributions indicates that there may be outliers that “pull” the least-squares fit too much in their direction. We may consider robust regression methods in the future processing.

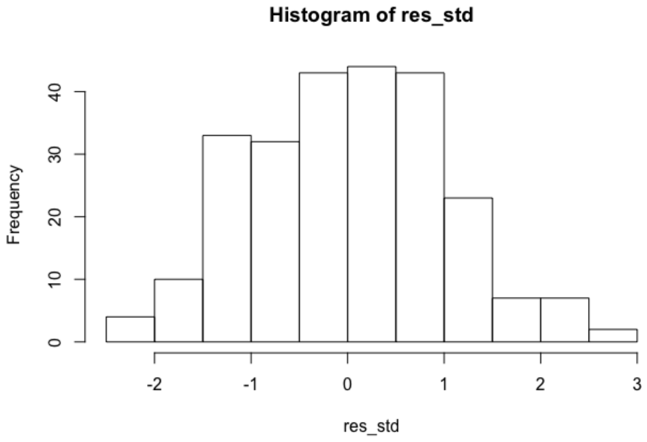


Figure 3: Histogram of the Standardized Residuals

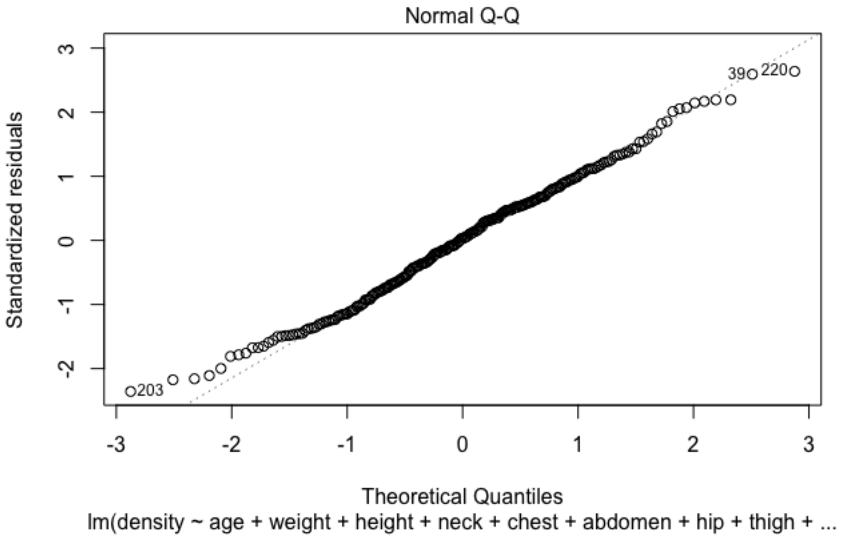


Figure 4: Normal Q-Q Plot

2.2.3 Residuals against fitted values and regressors

Another way of checking model adequency is to plot the residuals (here we use *externally studentized residuals*) against the fitted values of the model and results are as follows[5].

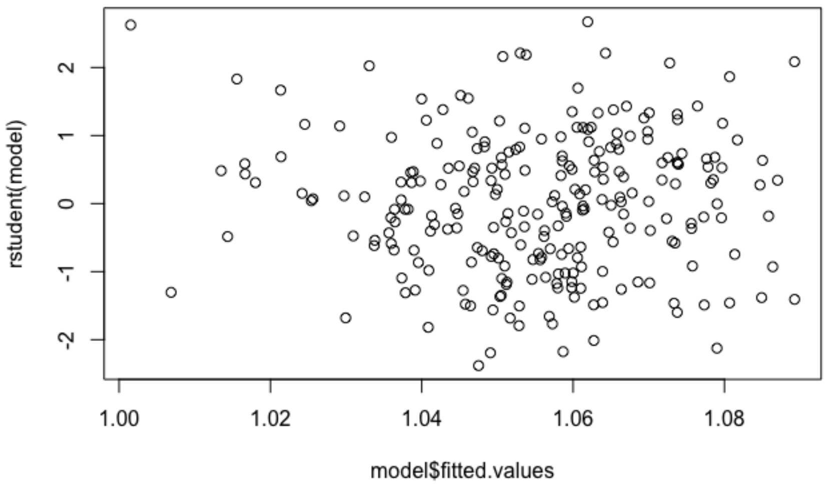


Figure 5: Residuals against fitted values

The residuals can be contained in a horizontal band, which means there are no obvious model defects.

Similarly, we can construct the 13 graphs of residuals against 13 regressors seperately and the result [6][7]is also similar with the previous one.

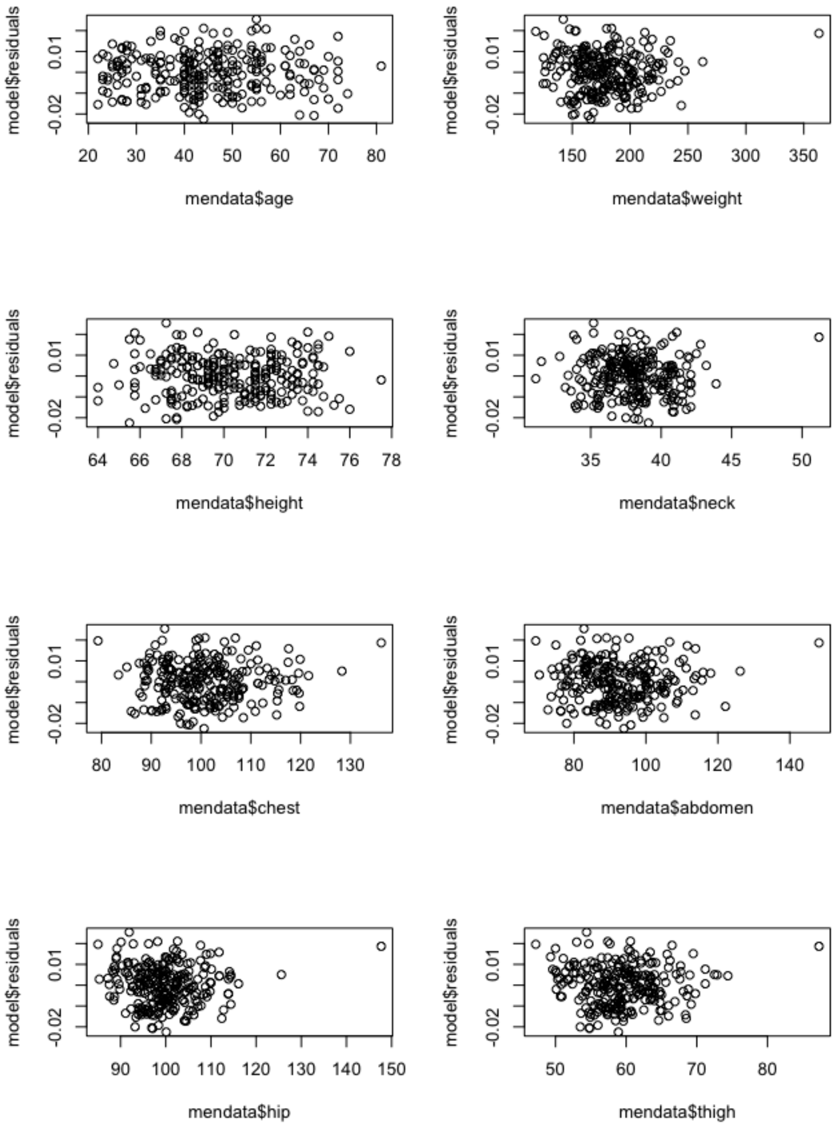


Figure 6: Residuals against regressors

2.2.4 Partial Residual Plots

Addtionally, we can construct Added-Variable Plots to see the marginal effect on the target of each regressor.

As is shown in the figure [8], some regressors such as *weight,neck,hip,forearm,abdomen,wrist* have obvious effect in this model while others don't such as*age and knee.*

2.3 Diagnostics of Leverage and Influence

2.3.1 Leverage Plots

Using `leveragePlots` in the `car` package, the leverage plots of each regressor are as follows [9].

2.3.2 Residuals V.S. Leverage

We can also construct the plot of Residuals V.S. Leverage [10].

2.3.3 Cook's distance

What's more, Cook's distance is an important indicator of leverage and influence.

As is shown in graph [11,12], because the maximum Cook's distance is just around 0.4(a suggested cut-off is 1 because $F_{0.5,p,n-p}$ is approximately equal to 1 when n is large) so there is no need to rule out any point.

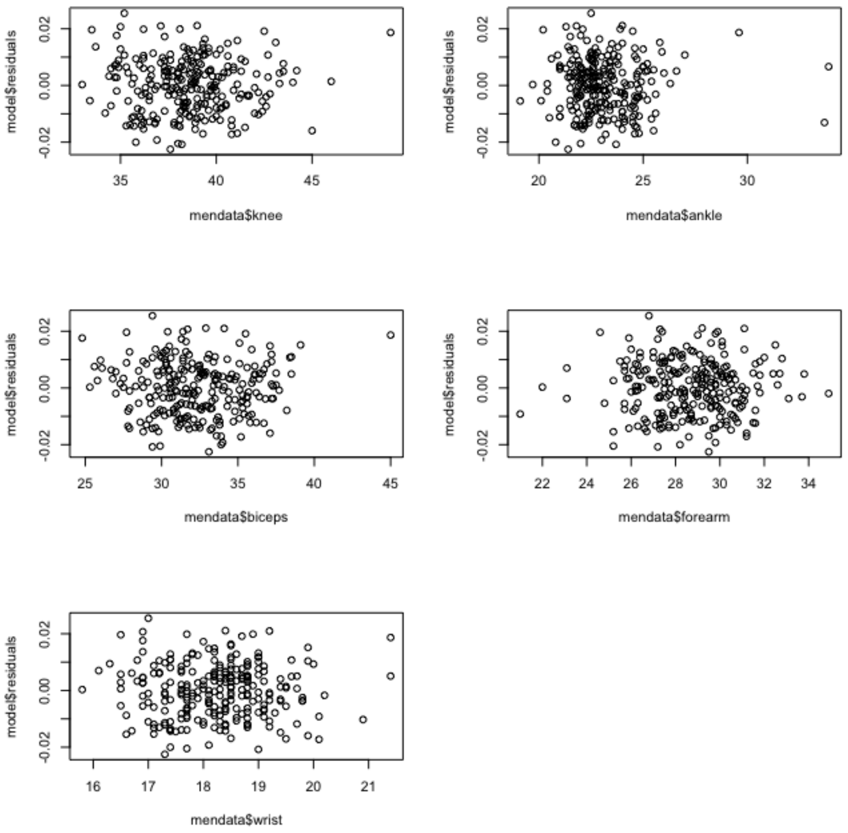


Figure 7: Residuals against regressors

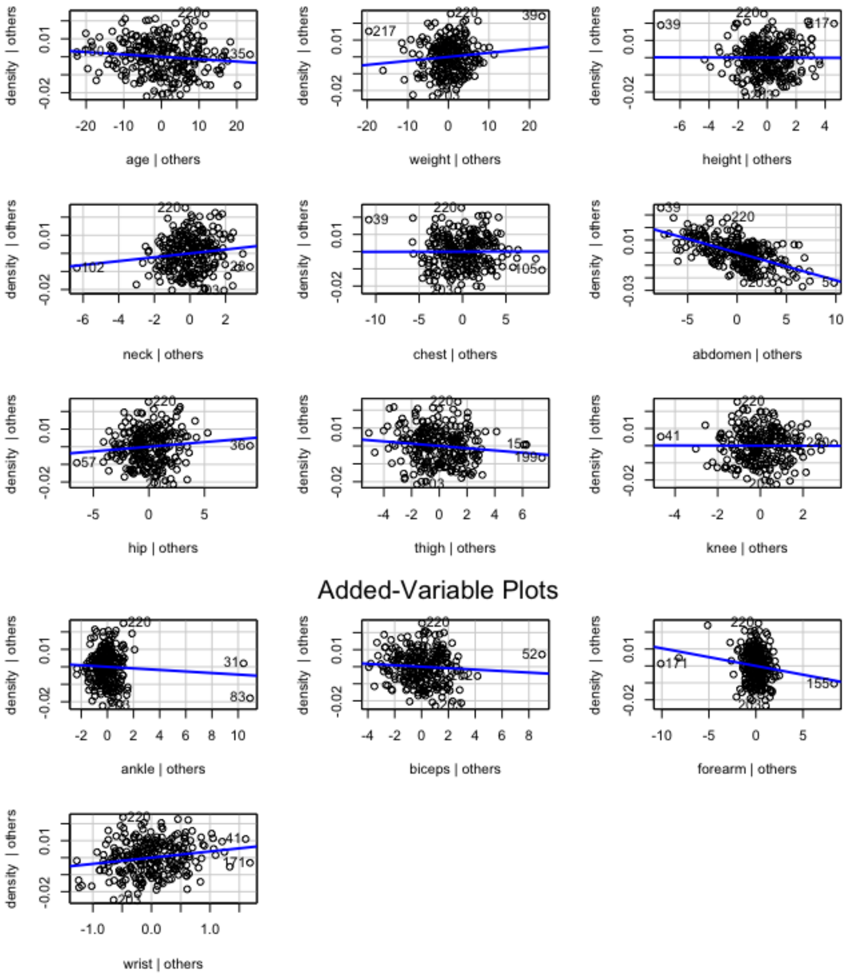


Figure 8: Residuals against regressors

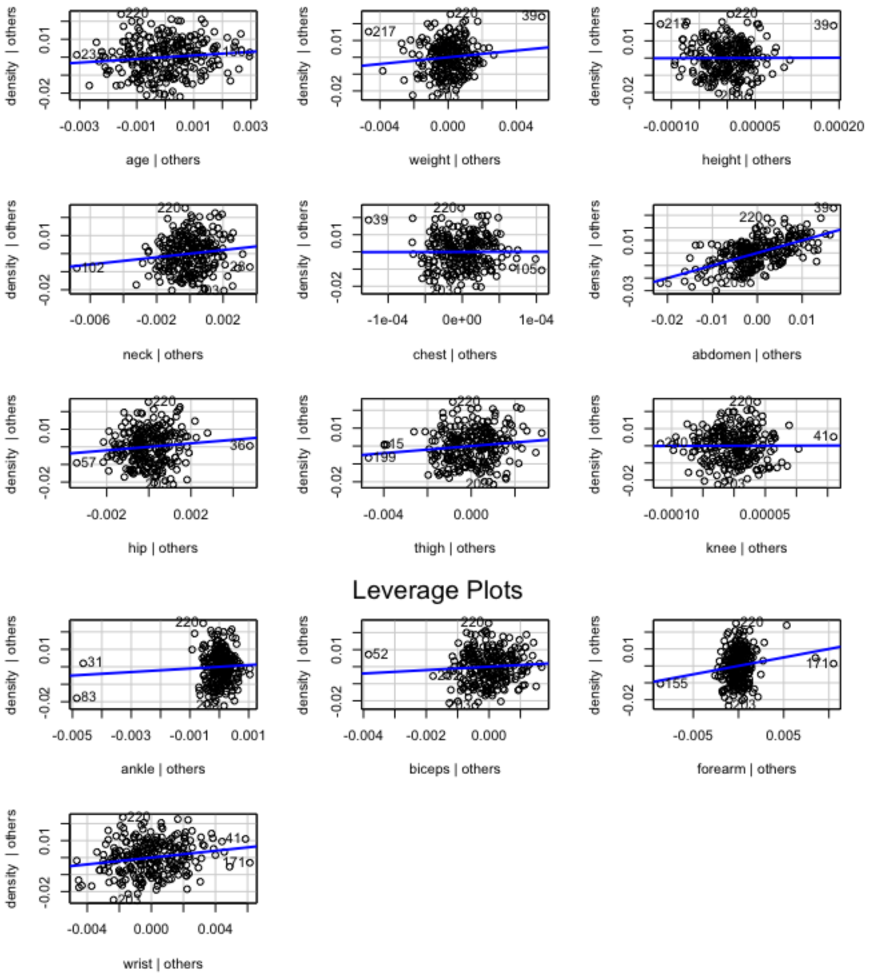


Figure 9: Leverage of regressors

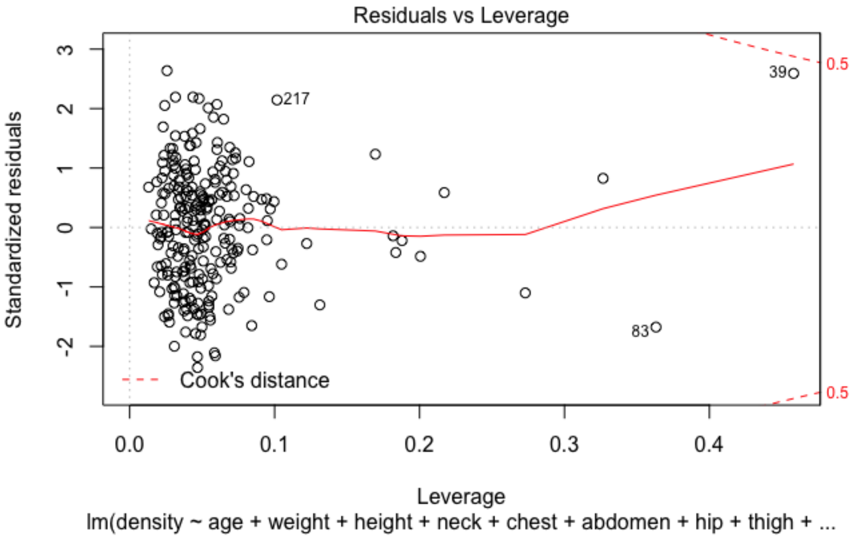


Figure 10: Residuals V.S. Leverage

2.4 Variable Transformation

Then we will consider possible variable transformation to improve the model.

According to the externally studentized residuals against fitted values plot [5], there is no obvious heteroscedasticity between fitted values. Also we can perform `ncvtest()` to test heteroscedasticity and the result is as follows [13].

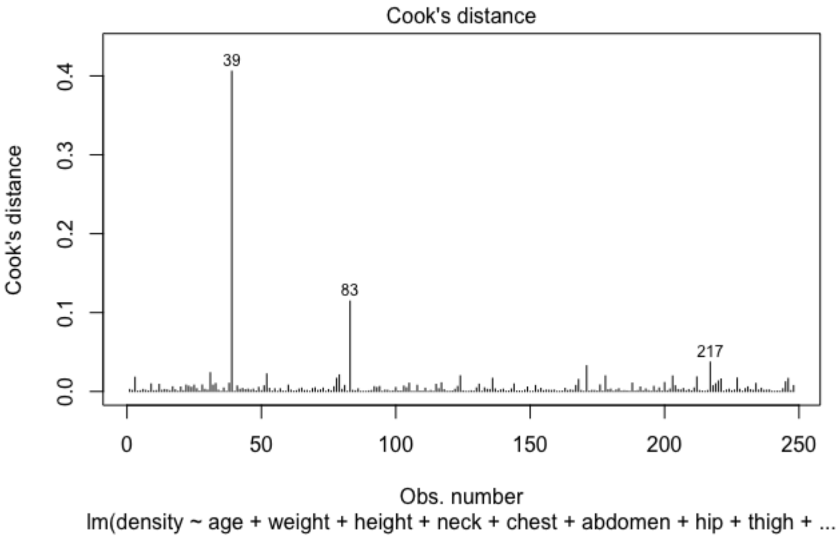


Figure 11: Cook's Distance

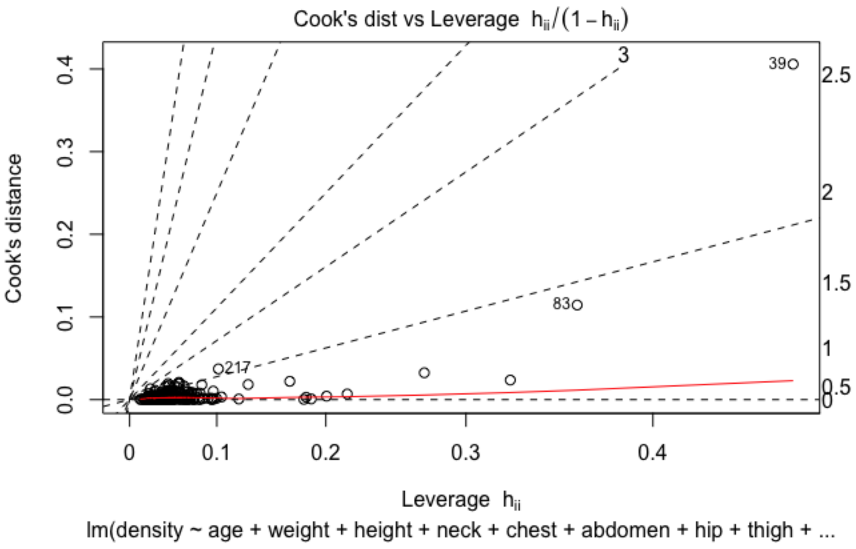
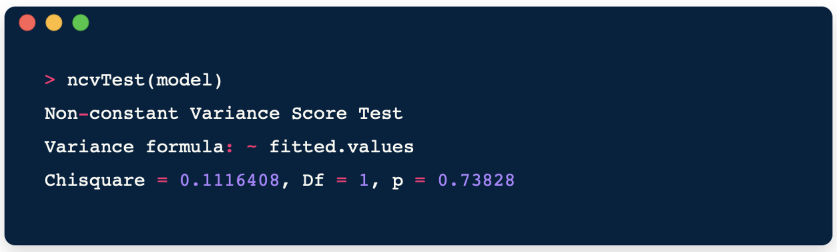


Figure 12: Cook's Distance vs Leverage



```
> ncvTest(model)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.1116408, Df = 1, p = 0.73828
```

Figure 13: Non-Constant Variance Test

However, we can also perform Box-Cox Transform to see if there is any possible power transforms.

$$\lambda = -3.724051$$

So we get the tranformed model [2].

$$\begin{aligned}\text{density}^{-4} =& 0.5393006 + 3.706681 \times 10^{-4} * \text{age} \; - 6.524177 \times 10^{-4} * \text{weight}\\ &- 3.844684 \times 10^{-4} * \text{height} - 3.170274 \times 10^{-3} * \text{neck} - 8.605261 \times 10^{-5} * \text{chest}\\ &+ 6.732445 \times 10^{-3} * \text{abdomen} - 1.366551 \times 10^{-3} * \text{hip} + 1.544550 \times 10^{-3} * \text{thigh}\\ &- 1.500961 \times 10^{-4} * \text{knee} + 1.437862 \times 10^{-3} * \text{ankle} + 1.170640 \times 10^{-3} * \text{biceps}\\ &+ 3.199862 \times 10^{-3} * \text{forearm} - 1.152693 \times 10^{-2} * \text{wrist} \end{aligned} \tag{2}$$

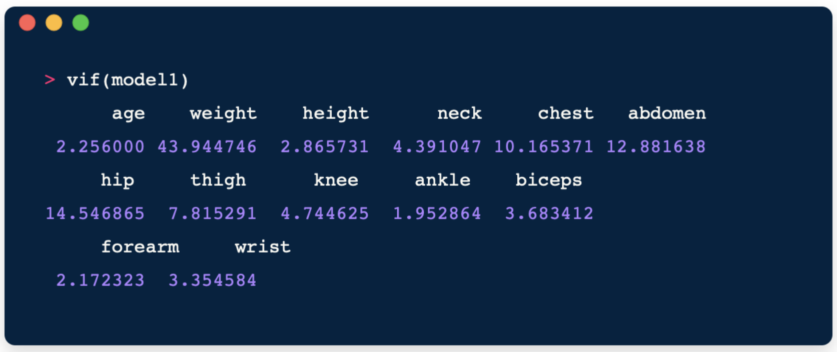
2.5 Multicollinearity diagnostics and treatments

2.5.1 Variance Inflation factors

One way to check multicollinearity is to calculate VIFs.

$$VIF_j = C_{jj} = (1 - R_j^2)^{-1}$$

The results are as follows[14].



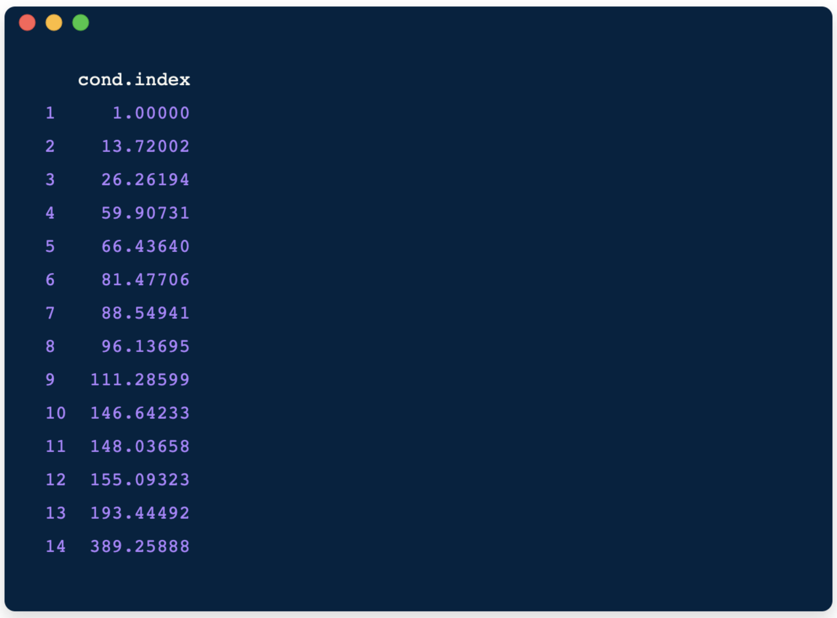
```
> vif(model1)
      age    weight    height      neck     chest   abdomen
 2.256000 43.944746  2.865731  4.391047 10.165371 12.881638
      hip     thigh      knee     ankle    biceps
14.546865  7.815291  4.744625  1.952864  3.683412
  forearm     wrist
 2.172323  3.354584
```

Figure 14: VIF

Because the maximum value is 43.94 so there exists multicollinearity.

2.5.2 Eigensystem Analysis

Another way of diagnostics is Eigensystem Analysis to calculate condition indices



```
   cond.index
1     1.00000
2    13.72002
3    26.26194
4    59.90731
5    66.43640
6    81.47706
7    88.54941
8    96.13695
9   111.28599
10  146.64233
11  148.03658
12  155.09323
13  193.44492
14  389.25888
```

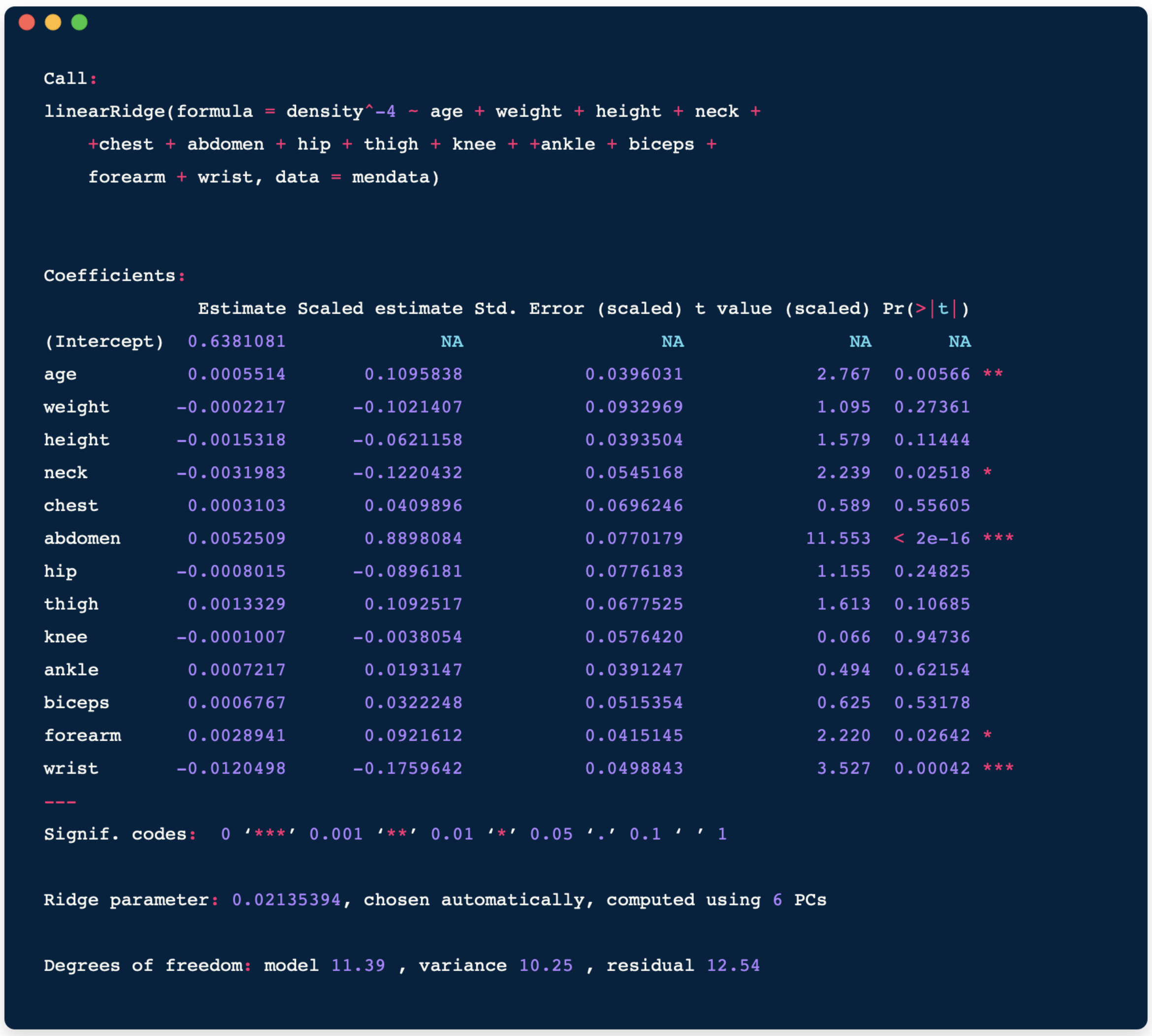
Figure 15: Condition Indices

Since many values exceed 100, there exists multicollinearity.

2.5.3 Ridge Regression

To address the multicollinearity, we can use ridge regression and the result is as follows [16].
And we get the scaled ridge regression model [3]

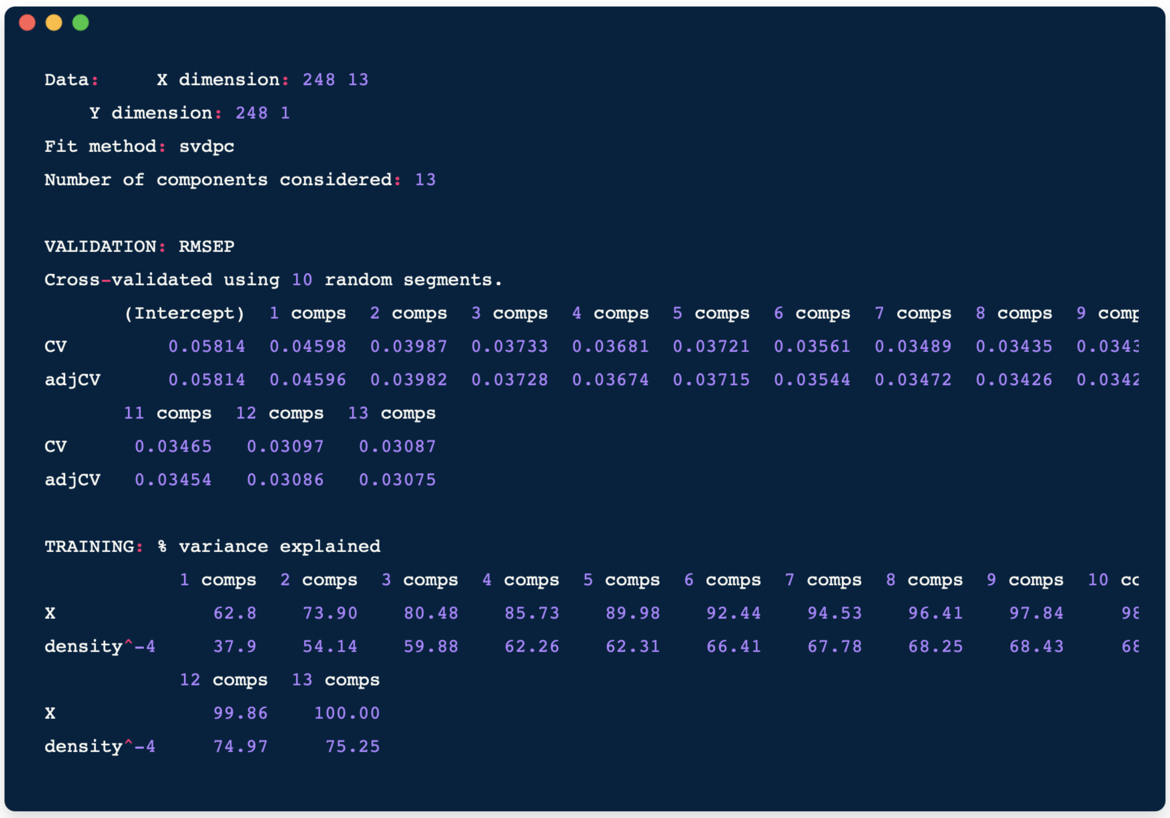
$$\begin{aligned}\text{density}^{-4} =& 0.6381081 + 0.0005514 * \text{age} - 0.0002217 * \text{weight}\\ & - 0.0015318 * \text{height} - 0.0031983 * \text{neck}0.0003103 * \text{chest}\\ & 0.0052509 * \text{abdomen} - 0.0008015 * \text{hip} + 0.0013329 * \text{thigh}\\ & - 0.0001007 * \text{knee} + 0.0007217 * \text{ankle} + 0.0006767 * \text{biceps}\\ & + 0.0028941 * \text{forearm} - 0.0120498 * \text{wrist} \end{aligned} \tag{3}$$



```
Call:
linearRidge(formula = density^-4 ~ age + weight + height + neck +
    +chest + abdomen + hip + thigh + knee + +ankle + biceps +
    forearm + wrist, data = mendata)


Coefficients:
              Estimate Scaled estimate Std. Error (scaled) t value (scaled) Pr(>|t|)
(Intercept)  0.6381081              NA                  NA               NA       NA
age          0.0005514       0.1095838           0.0396031            2.767  0.00566 **
weight      -0.0002217      -0.1021407           0.0932969            1.095  0.27361
height      -0.0015318      -0.0621158           0.0393504            1.579  0.11444
neck        -0.0031983      -0.1220432           0.0545168            2.239  0.02518 *
chest        0.0003103       0.0409896           0.0696246            0.589  0.55605
abdomen      0.0052509       0.8898084           0.0770179           11.553  < 2e-16 ***
hip         -0.0008015      -0.0896181           0.0776183            1.155  0.24825
thigh        0.0013329       0.1092517           0.0677525            1.613  0.10685
knee        -0.0001007      -0.0038054           0.0576420            0.066  0.94736
ankle        0.0007217       0.0193147           0.0391247            0.494  0.62154
biceps       0.0006767       0.0322248           0.0515354            0.625  0.53178
forearm      0.0028941       0.0921612           0.0415145            2.220  0.02642 *
wrist       -0.0120498      -0.1759642           0.0498843            3.527  0.00042 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Ridge parameter: 0.02135394, chosen automatically, computed using 6 PCs

Degrees of freedom: model 11.39 , variance 10.25 , residual 12.54
```

Figure 16: Ridge Regression Model

2.5.4 Principal Component Regression



```
Data:     X dimension: 248 13
    Y dimension: 248 1
Fit method: svdpc
Number of components considered: 13

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comp
CV         0.05814  0.04598  0.03987  0.03733  0.03681  0.03721  0.03561  0.03489  0.03435  0.0343
adjCV      0.05814  0.04596  0.03982  0.03728  0.03674  0.03715  0.03544  0.03472  0.03426  0.0342
       11 comps  12 comps  13 comps
CV      0.03465   0.03097   0.03087
adjCV   0.03454   0.03086   0.03075

TRAINING: % variance explained
            1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 co
X              62.8    73.90    80.48    85.73    89.98    92.44    94.53    96.41    97.84      98
density^-4     37.9    54.14    59.88    62.26    62.31    66.41    67.78    68.25    68.43      68
            12 comps  13 comps
X              99.86    100.00
density^-4     74.97     75.25
```

Figure 17: Principal Component Regression Model

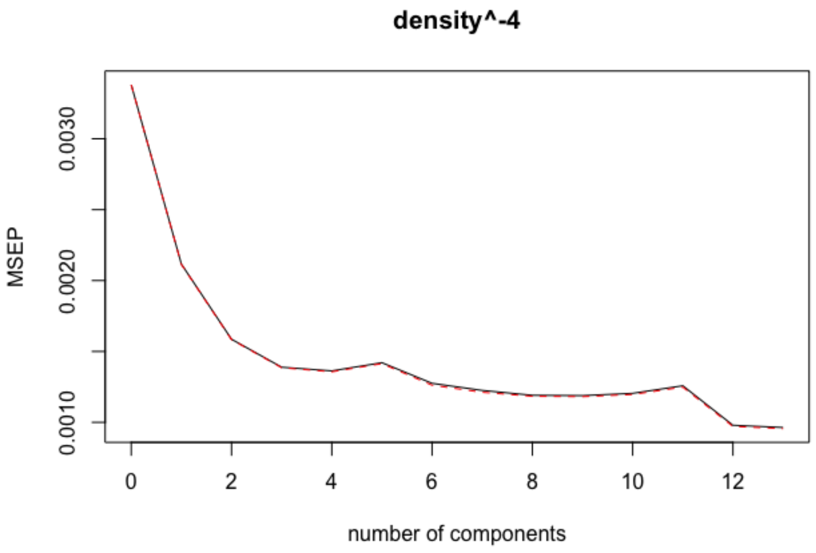


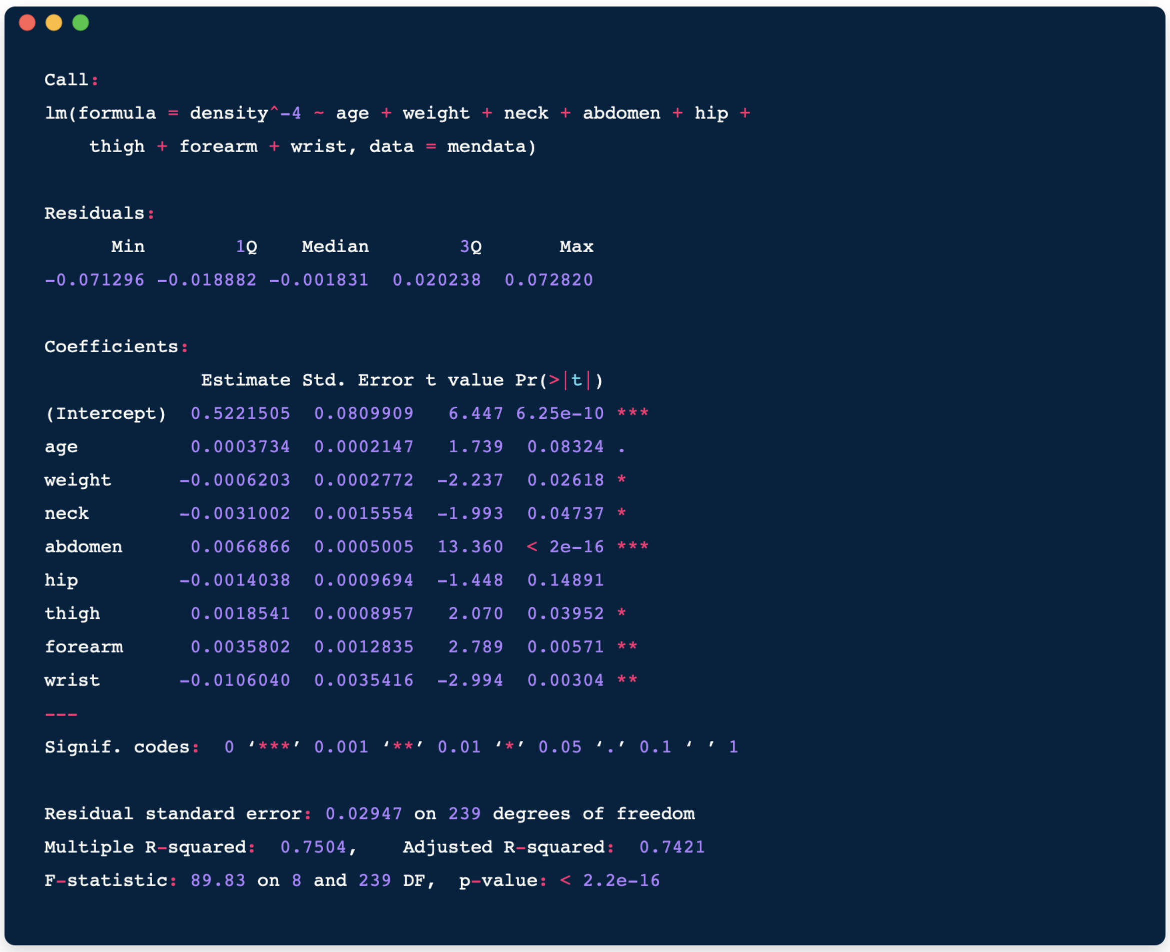
Figure 18: MSEP

And we get the scaled Principal Component Regression Model [4] using the first 6 Principal Component ,

$$\begin{aligned}\text{density}^{-4} =& 0.7161708 + 0.014240774 * \text{age} - 0.005425814 * \text{weight} \\ & - 0.012197553 * \text{height} - 0.007726027 \text{neck} 0.011736343 * \text{chest} \\ & + 0.016157224 * \text{abdomen} + 0.010209759 * \text{hip} + 0.008134732 * \text{thigh} \\ & + 0.010946211 * \text{knee} - 0.003781667 \text{ankle} + 0.001847463 * \text{biceps} \\ & + 0.003295693 * \text{forearm} - 0.012448422 * \text{wrist} \end{aligned} \tag{4}$$

2.6 Variable Selection

We can use stepwise regression for variable selection.The modified model is as follows [19].



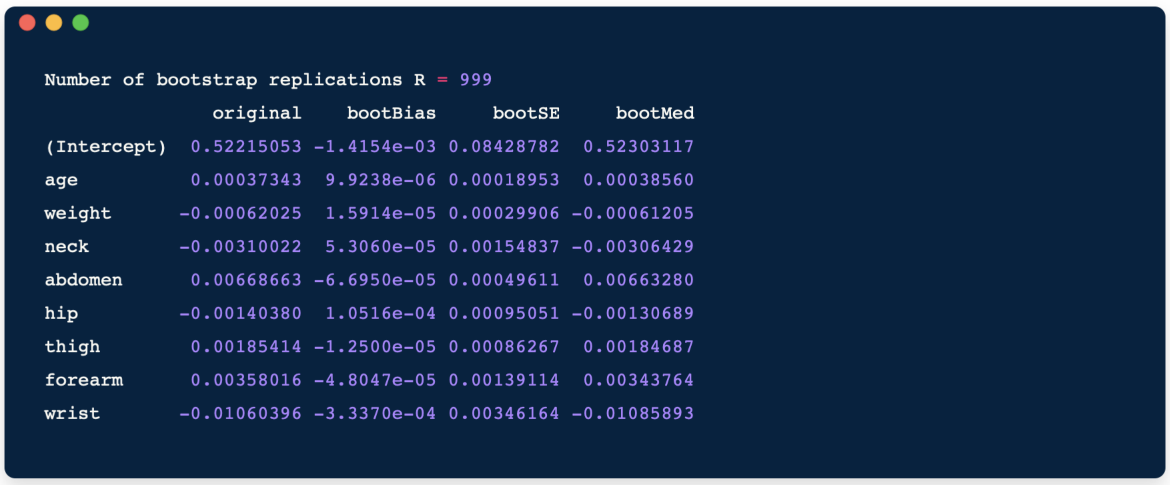
```
Call:
lm(formula = density^-4 ~ age + weight + neck + abdomen + hip + 
    thigh + forearm + wrist, data = mendata)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.071296 -0.018882 -0.001831  0.020238  0.072820 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.5221505  0.0809909   6.447 6.25e-10 ***
age          0.0003734  0.0002147   1.739  0.08324 .  
weight      -0.0006203  0.0002772  -2.237  0.02618 *  
neck        -0.0031002  0.0015554  -1.993  0.04737 *  
abdomen      0.0066866  0.0005005  13.360  < 2e-16 ***
hip         -0.0014038  0.0009694  -1.448  0.14891    
thigh        0.0018541  0.0008957   2.070  0.03952 *  
forearm      0.0035802  0.0012835   2.789  0.00571 ** 
wrist       -0.0106040  0.0035416  -2.994  0.00304 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02947 on 239 degrees of freedom
Multiple R-squared:  0.7504,    Adjusted R-squared:  0.7421 
F-statistic: 89.83 on 8 and 239 DF,  p-value: < 2.2e-16
```

Figure 19: Stepwise Regression

$$
\begin{aligned}
density^{-4} =& 0.5221505 + 0.0003734 * \text{age} - 0.0006203 * \text{weight} - 0.0031002 * \text{neck} + 0.0066866 * \text{abdomen} \\
& - 0.0014038 * \text{hip} + 0.0018541 * \text{thigh} + 0.0035802 * \text{forearm} - 0.0106040 * \text{wrist}
\end{aligned} \tag{5}
$$

2.7 Bootstraping

The last step is to use bootstraping to assess the model [5] and we get the final model [6]



```
Number of bootstrap replications R = 999
              original   bootBias     bootSE     bootMed
(Intercept)  0.52215053 -1.4154e-03 0.08428782  0.52303117
age          0.00037343  9.9238e-06 0.00018953  0.00038560
weight      -0.00062025  1.5914e-05 0.00029906 -0.00061205
neck        -0.00310022  5.3060e-05 0.00154837 -0.00306429
abdomen      0.00668663 -6.6950e-05 0.00049611  0.00663280
hip         -0.00140380  1.0516e-04 0.00095051 -0.00130689
thigh        0.00185414 -1.2500e-05 0.00086267  0.00184687
forearm      0.00358016 -4.8047e-05 0.00139114  0.00343764
wrist       -0.01060396 -3.3370e-04 0.00346164 -0.01085893
```

Figure 20: Bootstraping

$$
\begin{aligned}
density^{-4} =& 0.5230312 + 0.0003856 * \text{age} - 0.00061205 * \text{weight} - 0.0030643 * \text{neck} + 0.0066328 * \text{abdomen} \\
& - 0.00130689 * \text{hip} + 0.00184687 * \text{thigh} + 0.00343764 * \text{forearm} - 0.01085893 * \text{wrist}
\end{aligned} \tag{6}
$$

3 Results

3.1 Our Models

First we establish the original full model[2.1] by simple linear regresstion [1]. Then after residual analysis[2.2]and leverage and influence diagnostic [2.3] we usr Box-Cox Tranform [2.4] to build a new model [2] Through multicollinearity diagnostics, we try to address the multicollinearity with two different approaches– Ridge Regression [3] and Principal Component Regression(PCR) [4]. Based on previous models, we use stepwise regression to eliminate some regressors and construct the simplified model[5]. The last step is to assess last model with bootstraping, through this, we achieve the final model[6].

3.2 Assessment

The first model[1] is easiest to build , however, is not validated. Through Box-Cox Transform we get the second model[1],which improves slightly comparing to the first one. But it still have problems such as multicollinearity.To handle this, we construct the 3rd[3] and 4th model[4], but these two models still lacks practicality because of too many variables, that's why we need to simplify the model and get[5].Finally we use bootstraping to add a small modification to the final model[6].

However, it is far from perfect, one possible way to improve it is to add some non-linear components to the regression model.

4 Conclusion

In this project, we try to establish a useful model for bodyfat. It is easy to raise a model, but it is even easier to overthrow it. So the most difficult part is how to modify the model to be more sturdy.

Through all of those struggle, we finally manage to set up our model.

$$
\begin{aligned}
density^{-4} =& 0.5230312 + 0.0003856 * \text{age} - 0.00061205 * \text{weight} - 0.0030643 * \text{neck} + 0.0066328 * \text{abdomen} \\
& - 0.00130689 * \text{hip} + 0.00184687 * \text{thigh} + 0.00343764 * \text{forearm} - 0.01085893 * \text{wrist}
\end{aligned}
$$

It is not so beautiful, viewed from a mathmatical prospective, but it is built and tested on the basis of practical data. It is hard to interpret what the equation really means (maybe it doesn't have any meaning at all) but if it really means something, I can only say the plain fact such as one's fatness generally increases with her or his weight, neck, wrist and hip.

The conclusion sounds boring, but it takes much work to comes to it, maybe that is why it is more like empirical rather than theoretical work.