Using linear regression model to predicting body fat

Taiyuan Zhang

2023-05-03

Aim

In this report, we are aiming to build linear regression model which could be used for predicting body fat using 10 body measurement variables. Although body fat is difficult to measure, it is important to help medical professionals determine risk of certain conditions. The data we have are body fat and relative body measurement variables of 202 men.

Summary

In the report, we first do some exploratory analysis to the data mainly by data visualization. Secondly, we do model selection to find the best subset of variables for regression based on AIC/BIC, Mallow’s Cp and Adjusted R-squared criterion. Thirdly, we identify and analyze outliers and high-leverage points. Fourthly, we check the model assumption by plotting QQ-plot, residual, component residual plot, etc and do the manipulation to model based on that. A comparison between our best model and the full model using test data will also be provided.

Exploratory analysis

We first delete zero of NAN value in Train as these could be considered as wrong data.

```
##   brozek neck chest abdom   hip thigh knee ankle biceps forearm wrist
## 1   12.6 36.2  93.1  85.2  94.5  59.0 37.3  21.9   32.0    27.4  17.1
## 2    6.9 38.5  93.6  83.0  98.7  58.7 37.3  23.4   30.5    28.9  18.2
## 3   10.9 37.4 101.8  86.4 101.2  60.1 37.3  22.8   32.4    29.4  18.2
## 4   27.8 34.4  97.3 100.0 101.9  63.2 42.2  24.0   32.2    27.7  17.7
## 5   19.0 36.4 105.1  90.7 100.3  58.4 38.3  22.9   31.9    27.8  17.7
```

Data summary

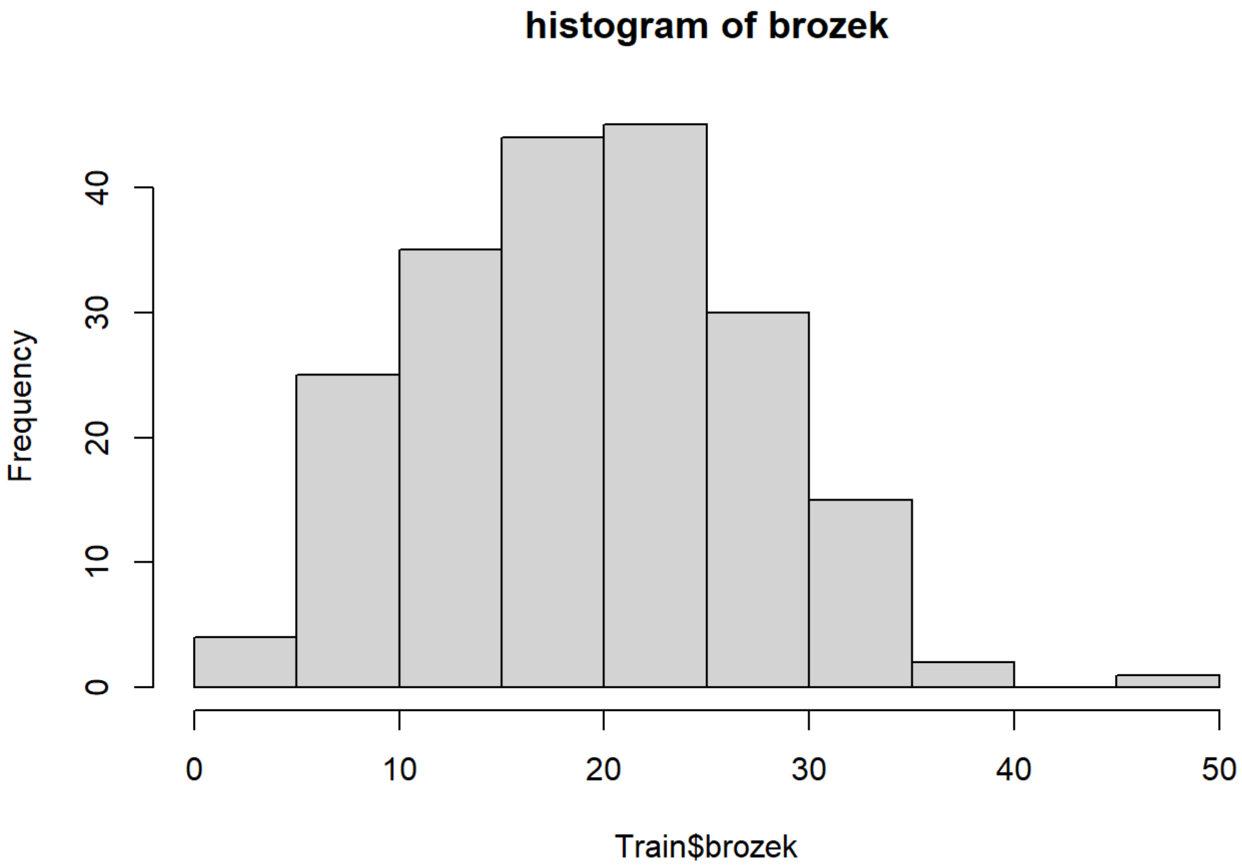
Name	Train
Number of rows	201
Number of columns	11
_______________________	
Column type frequency:	
numeric	11
________________________	
Group variables	None

Variable type: numeric

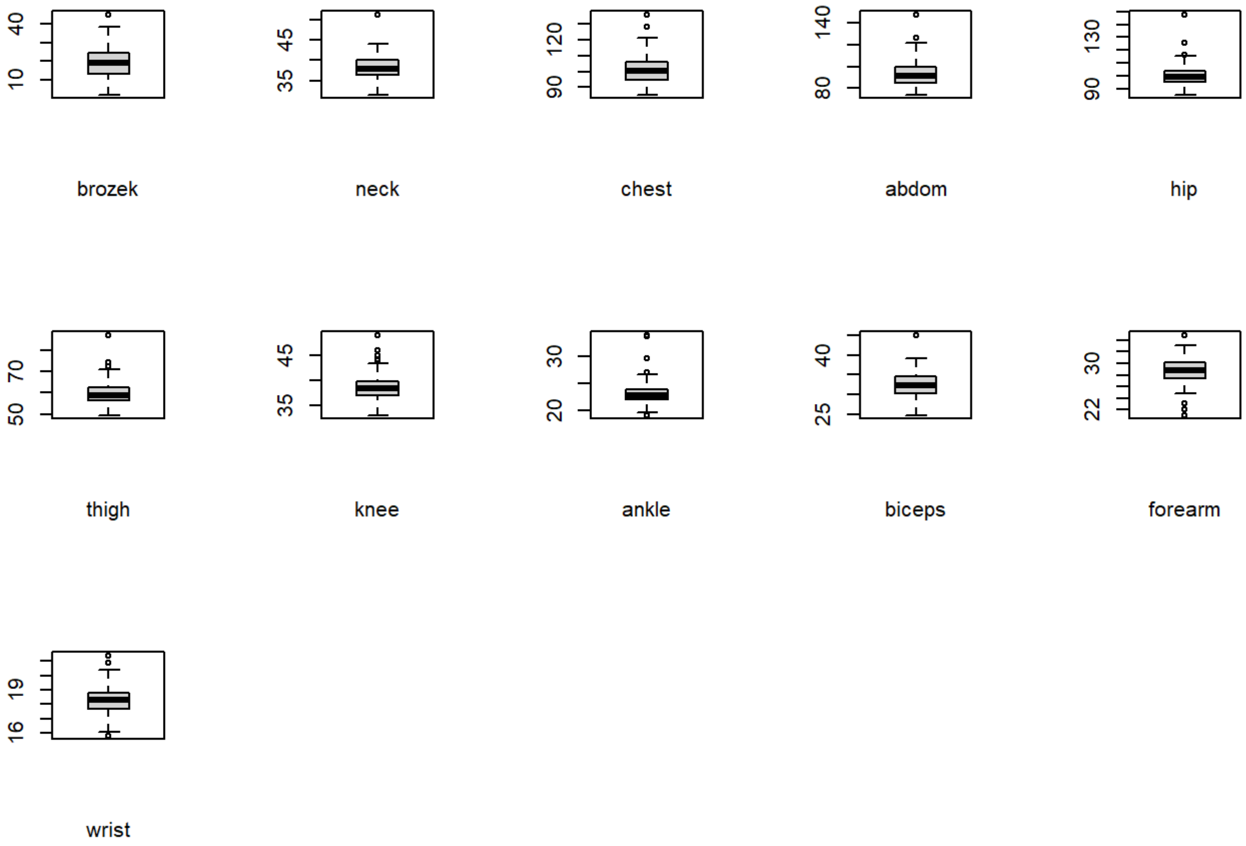
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
brozek	0	1	19.18	7.71	1.9	13.4	19.0	24.6	45.1	
neck	0	1	38.21	2.44	31.5	36.5	38.0	40.0	51.2	
chest	0	1	101.29	8.61	85.1	94.6	100.4	106.2	136.2	
abdom	0	1	93.09	10.96	73.9	84.6	91.1	99.8	148.1	
hip	0	1	100.09	7.27	85.3	95.5	99.4	103.7	147.7	
thigh	0	1	59.46	5.26	49.3	56.3	59.0	62.5	87.3	
knee	0	1	38.58	2.44	33.0	37.0	38.4	39.8	49.1	
ankle	0	1	23.08	1.77	19.1	22.0	22.8	23.9	33.9	
biceps	0	1	32.43	3.07	24.8	30.3	32.4	34.5	45.0	
forearm	0	1	28.71	2.03	21.0	27.4	28.8	30.1	34.9	
wrist	0	1	18.27	0.96	15.8	17.7	18.3	18.8	21.4	

A zero from brozek is deleted. A further check confirm that the data set after delete has all positive value. The distributions of the value of these variable are not identical, as the histograms show.

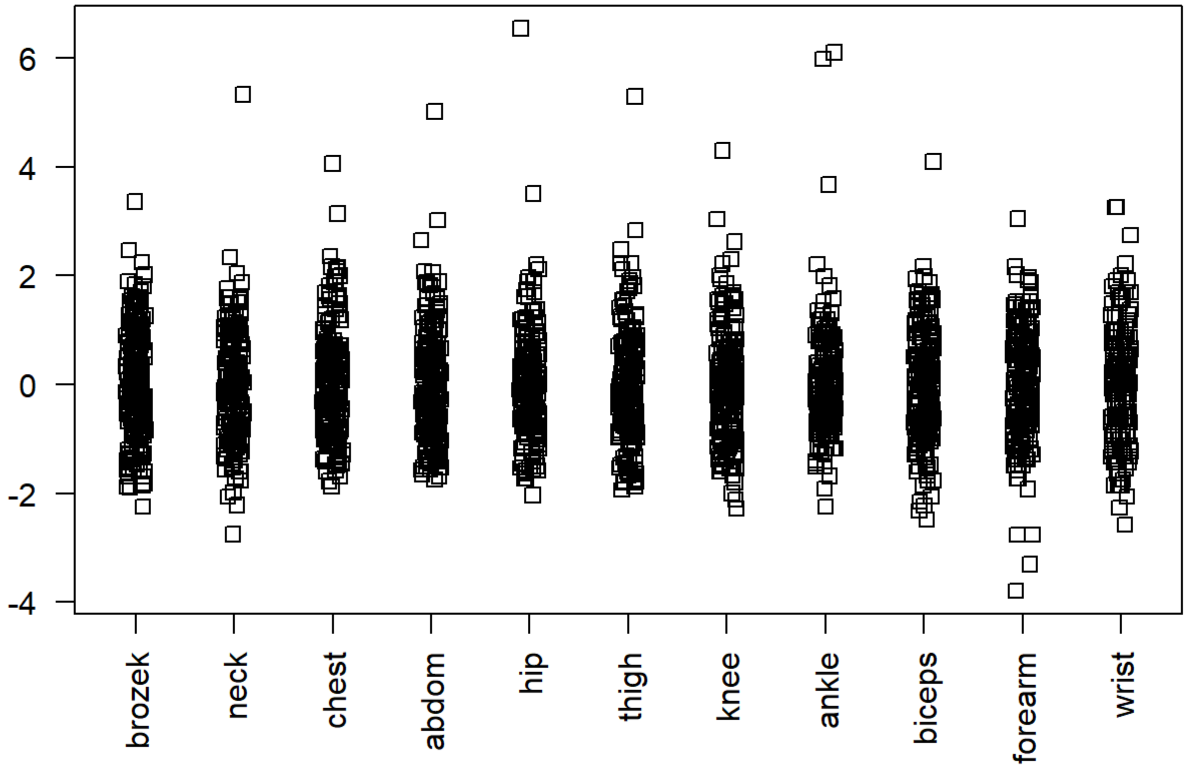
We then check the distribution of dependent variable. We see that there the dependent variable might not perfectly follow normal distribution as there is some positive skewness. However, the deviations from normality may be acceptable and only perform linear regression no not require the dependent and independent variables to be normally distributed. It is also worth noting that some variables such as hip and thigh hive strong positive skewness as the above rough histograms show, so transformation to these variables will being considered.



Thirdly, the Boxplot shows the outliers are less than 5 for all variables. Furthermore, most variables have almost all outliers bigger than normal value, except forearm.



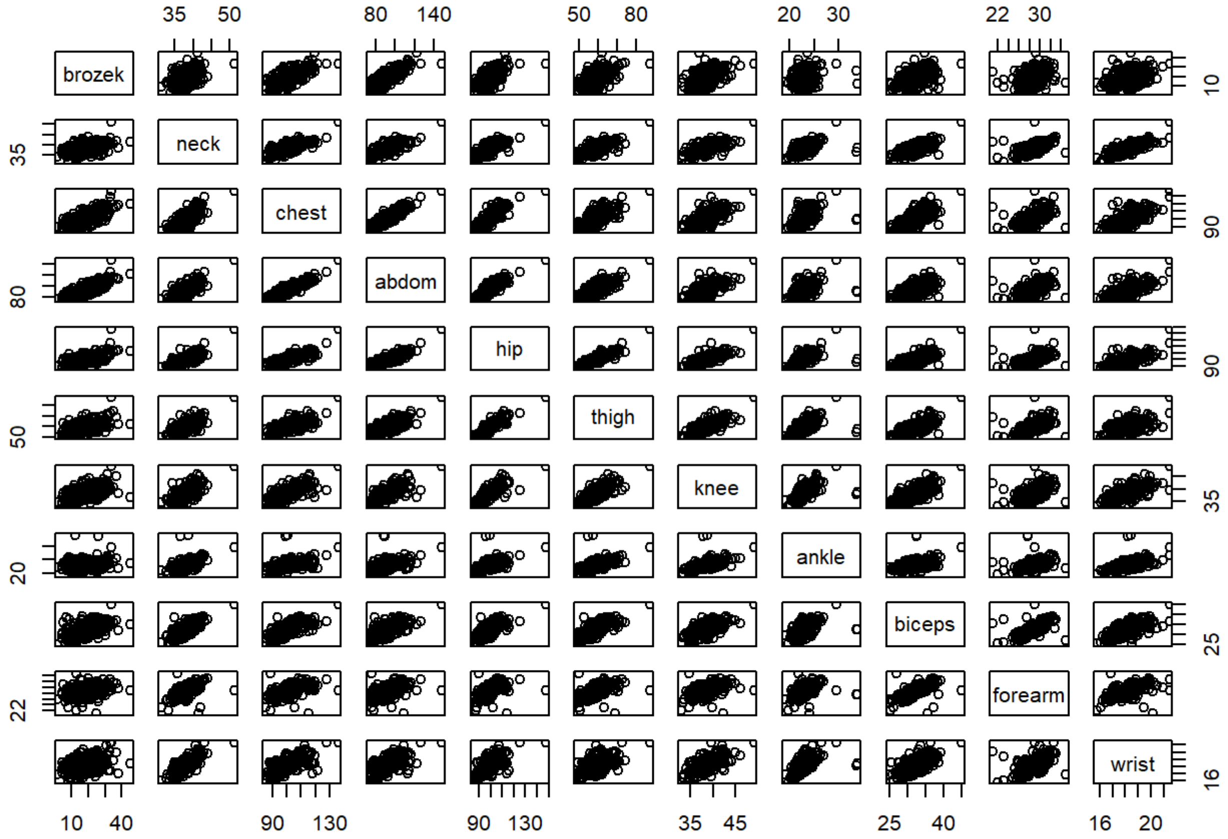
Fourthly, We check the skewness of variable using strip chart.



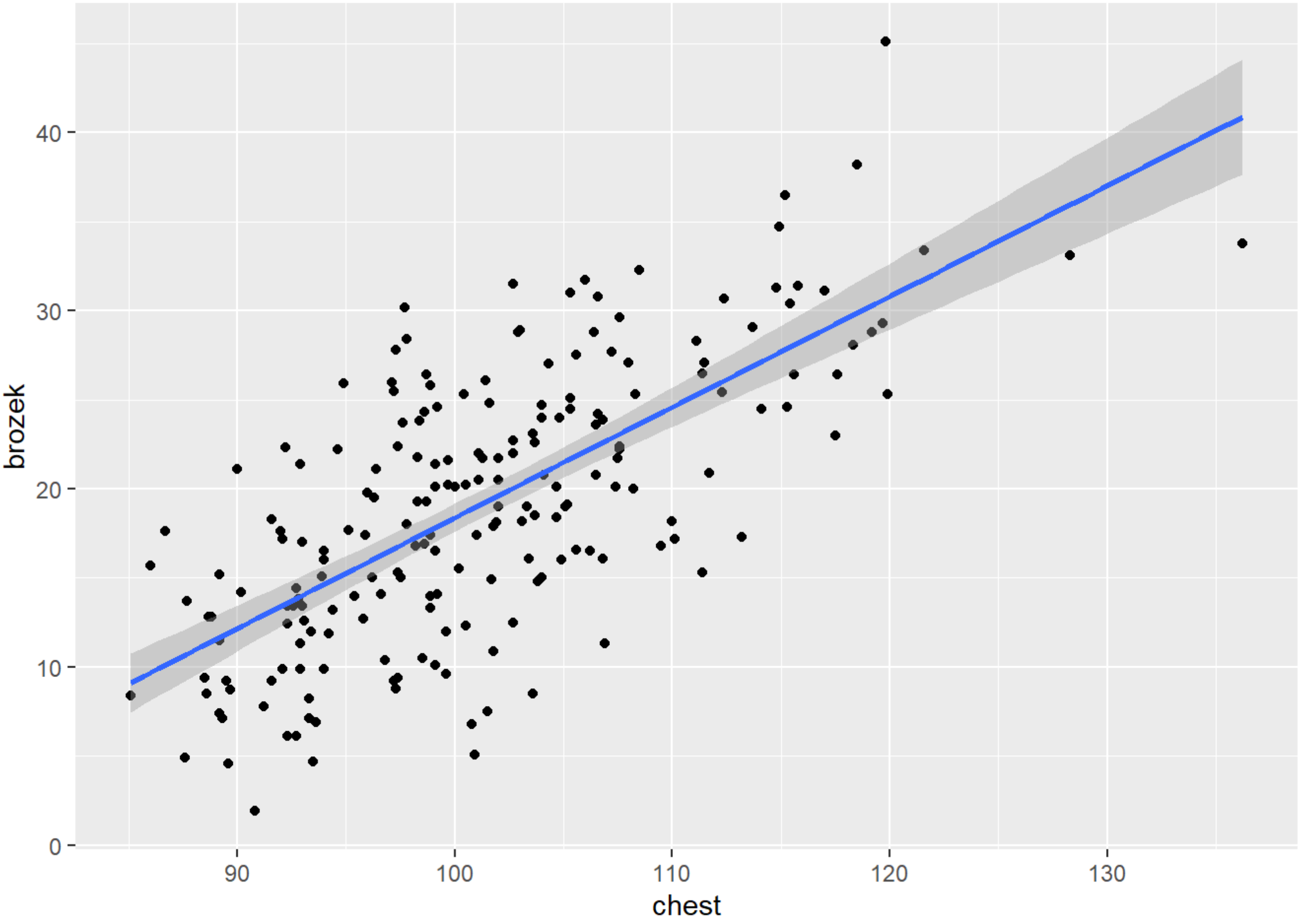
Small amount of noise is used to moving apart points to avoid severe overprint. We have also standardized the value of variable to make their distribution comparable. Generally, neck, abdom, hip, thigh and ankle are skewed.

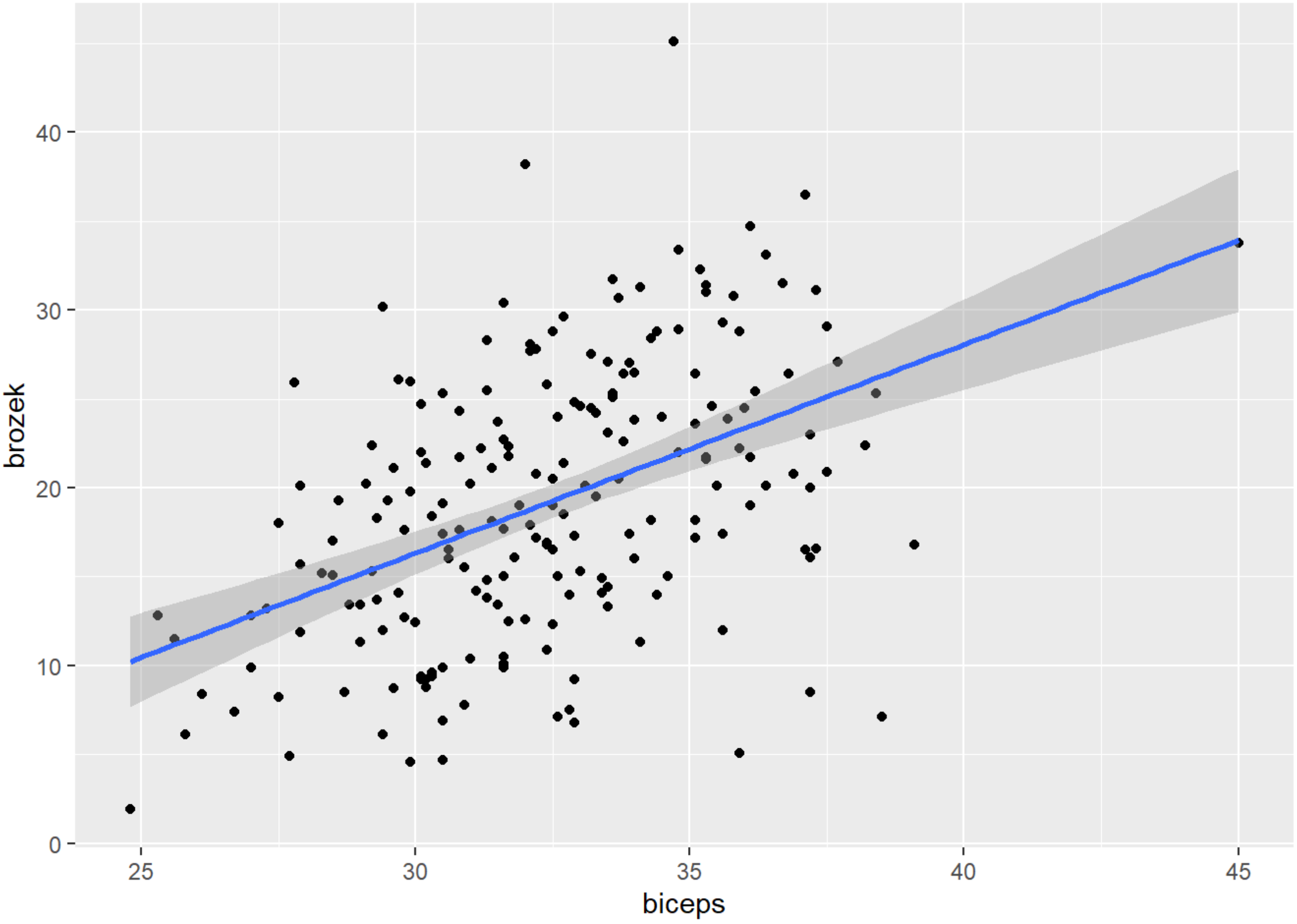
Finally, The scatter plot shows that approximately all the variable are in positive relationship with brozek

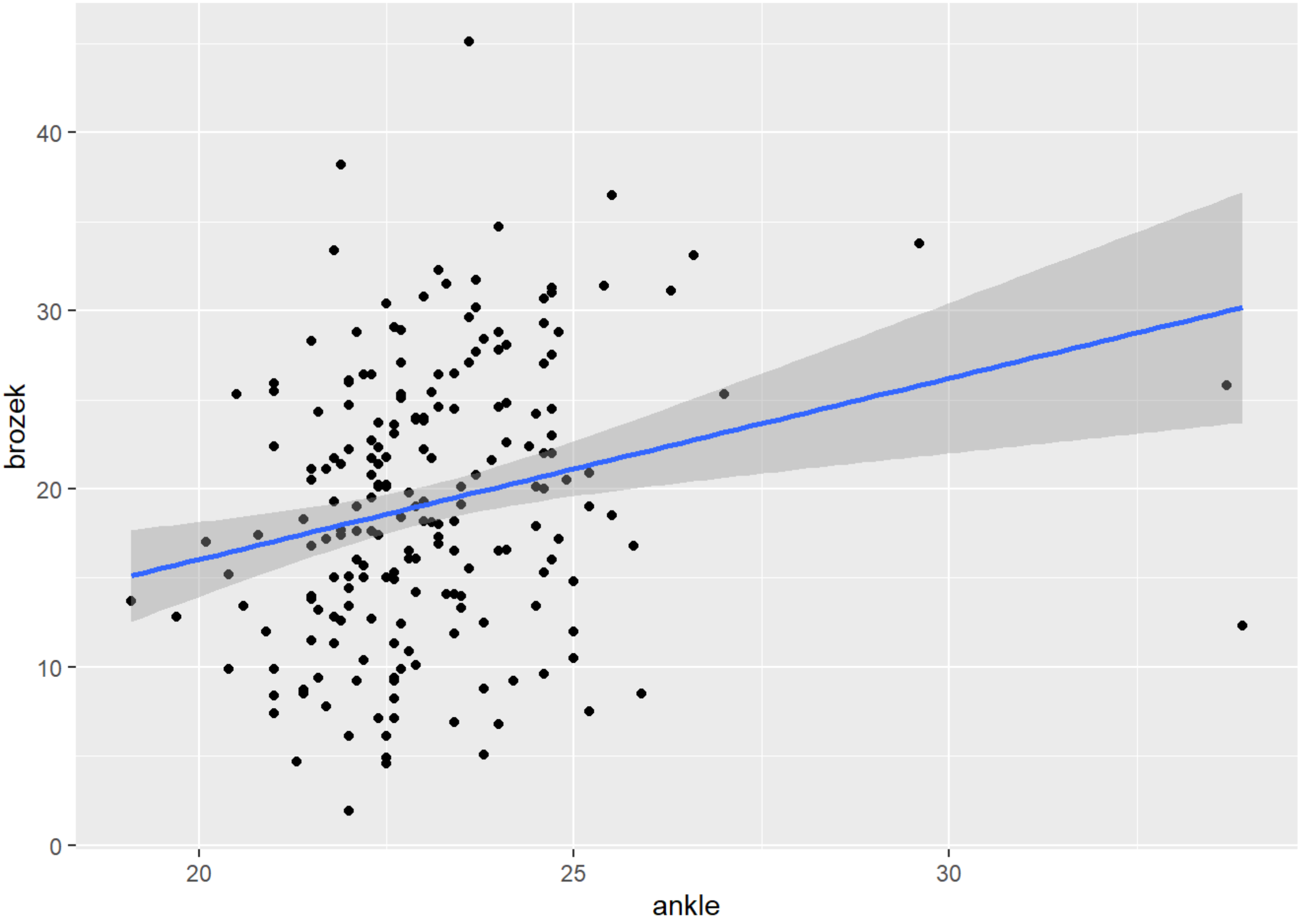
```
plot(Train)
```

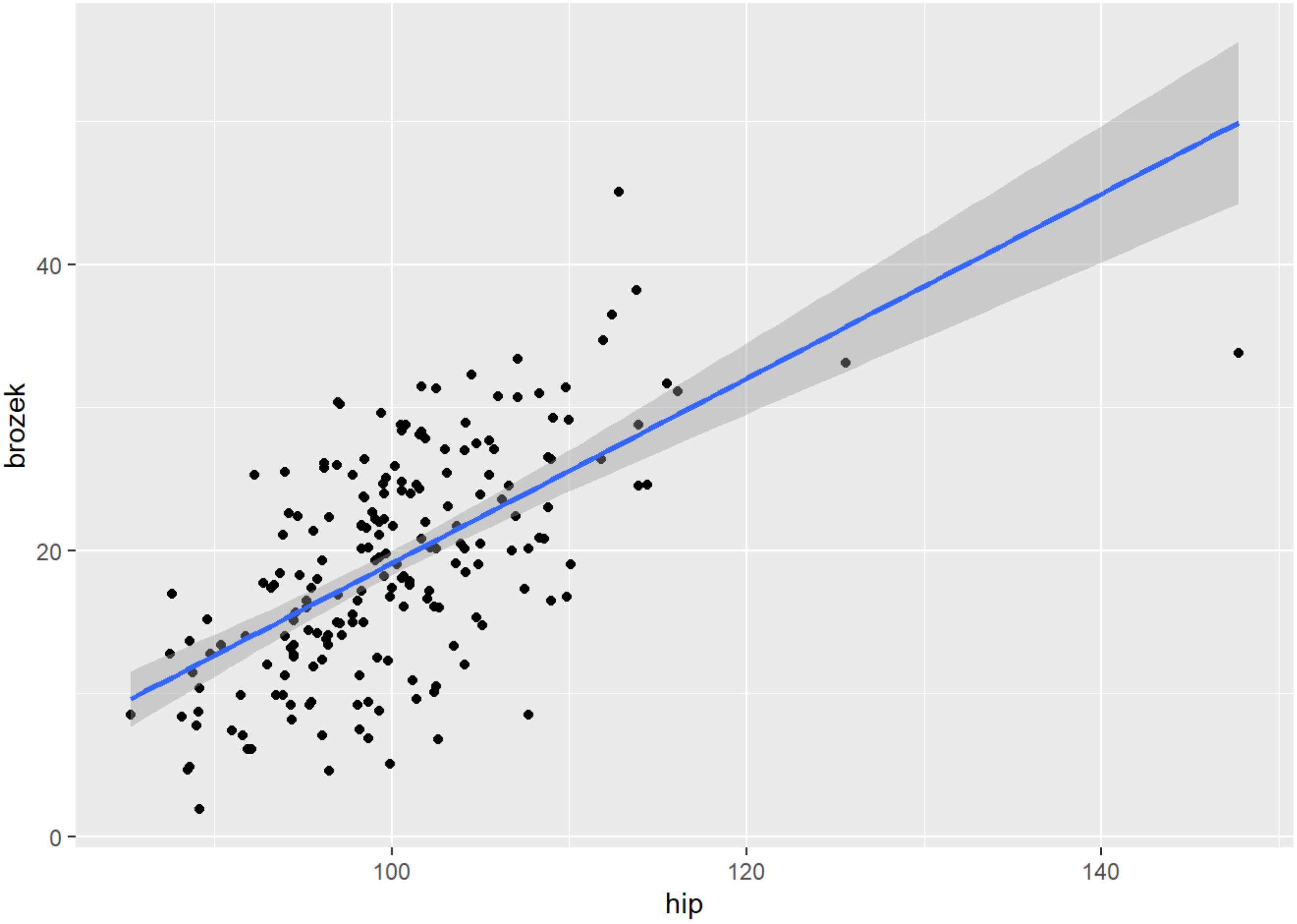


scatter plot plus the linear fit of some variables with 95% confidence interval shown in grey again confirm that there is strong positive relationship between these variables and brozek.


brozek
0
10
20
30
40
90
100
110
120
130
chest


brozek
0
10
20
30
40
25
30
35
40
45
biceps





A further look to one of these linear fit shows that estimating the value of brozek only by one variable is not accurate. The next part will justify which variables should included in this model.

```
##
## Call:
## lm(formula = brozek ~ hip, data = Train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.1173  -3.5959  -0.2852   4.2881  17.7143
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -45.43826    5.97844   -7.60 1.14e-12 ***
## hip           0.64560    0.05957   10.84  < 2e-16 ***
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## Residual standard error: 6.126 on 199 degrees of freedom
## Multiple R-squared:  0.3711, Adjusted R-squared:  0.368
## F-statistic: 117.4 on 1 and 199 DF,  p-value: < 2.2e-16
```

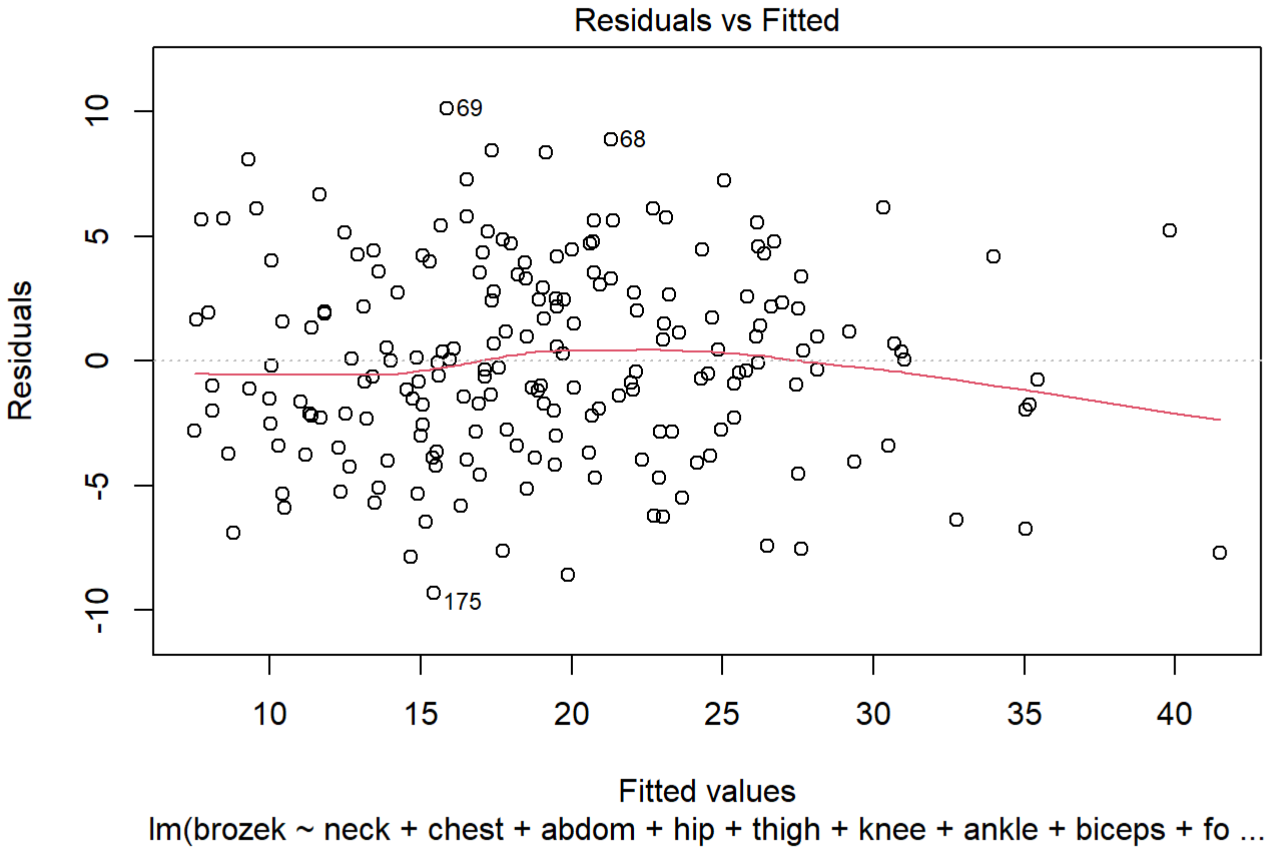
Explore on the full model

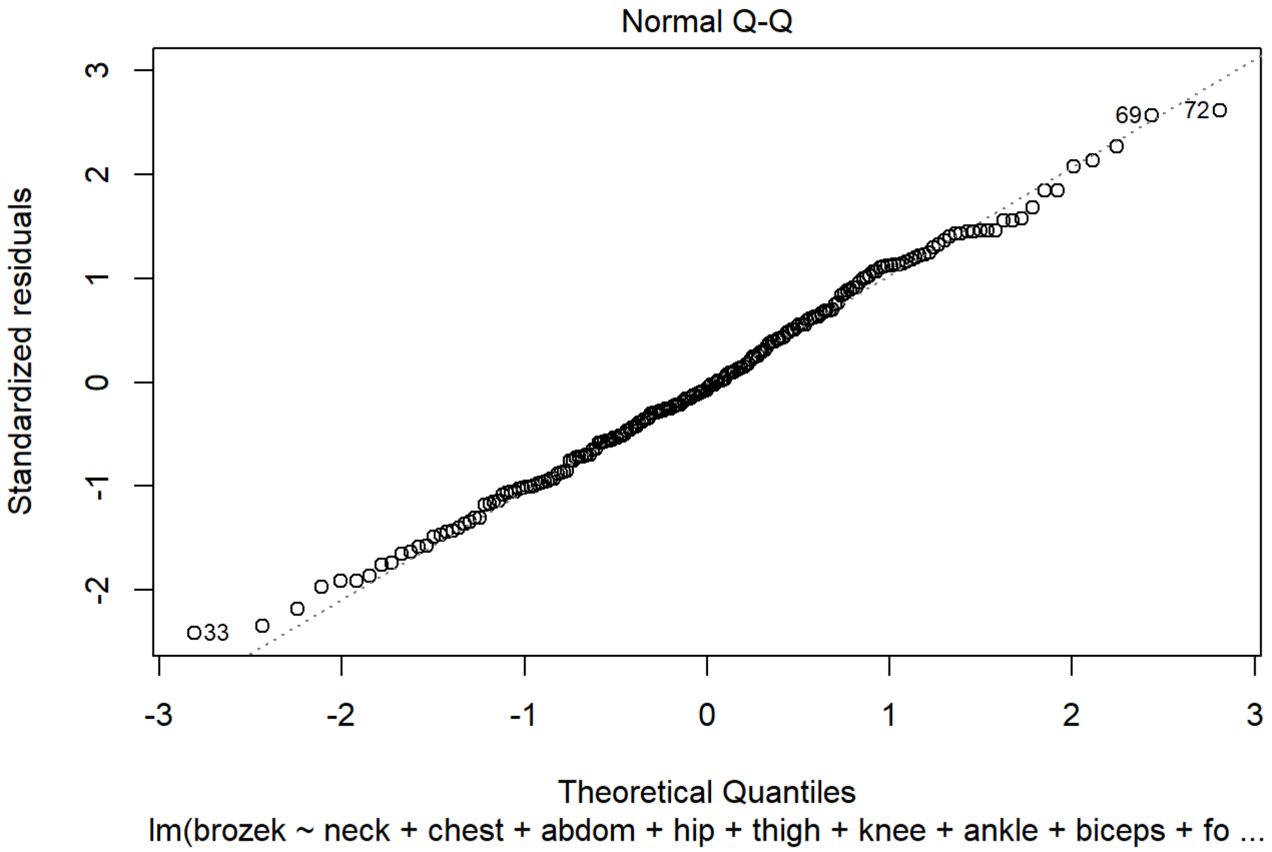
We start with the full model:

```
##
## Call:
## lm(formula = brozek ~ neck + chest + abdom + hip + thigh + knee +
##     ankle + biceps + forearm + wrist, data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3175 -2.7935 -0.2838  2.7578 10.1297
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.866246   6.844036   1.295  0.19673
## neck        -0.691259   0.241338  -2.864  0.00465 **
## chest       -0.108806   0.091820  -1.185  0.23750
## abdom        0.983531   0.077163  12.746  < 2e-16 ***
## hip         -0.339114   0.127061  -2.669  0.00827 **
## thigh       -0.004294   0.141129  -0.030  0.97576
## knee        -0.107833   0.223287  -0.483  0.62970
## ankle       -0.101910   0.210521  -0.484  0.62888
## biceps       0.173812   0.170848   1.017  0.31028
## forearm      0.357765   0.195377   1.831  0.06864 .
## wrist       -1.039871   0.508978  -2.043  0.04243 *
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## Residual standard error: 4.023 on 190 degrees of freedom
## Multiple R-squared:  0.741,  Adjusted R-squared:  0.7274
## F-statistic: 54.36 on 10 and 190 DF,  p-value: < 2.2e-16
```

This is could be regard as a pre-analysis before model selection. From the p-value above, we find that neck, abdom and hip have strong correlation to brozek while forearm and wrist have weaker correlation. The other variable do not show correlation with brozek. Generally, if we do Single parameter hypothesis test, only the 5 low p-value variables' test will reject the hypothesis that the parameter is zero, which means we could then combine these variable in the model.

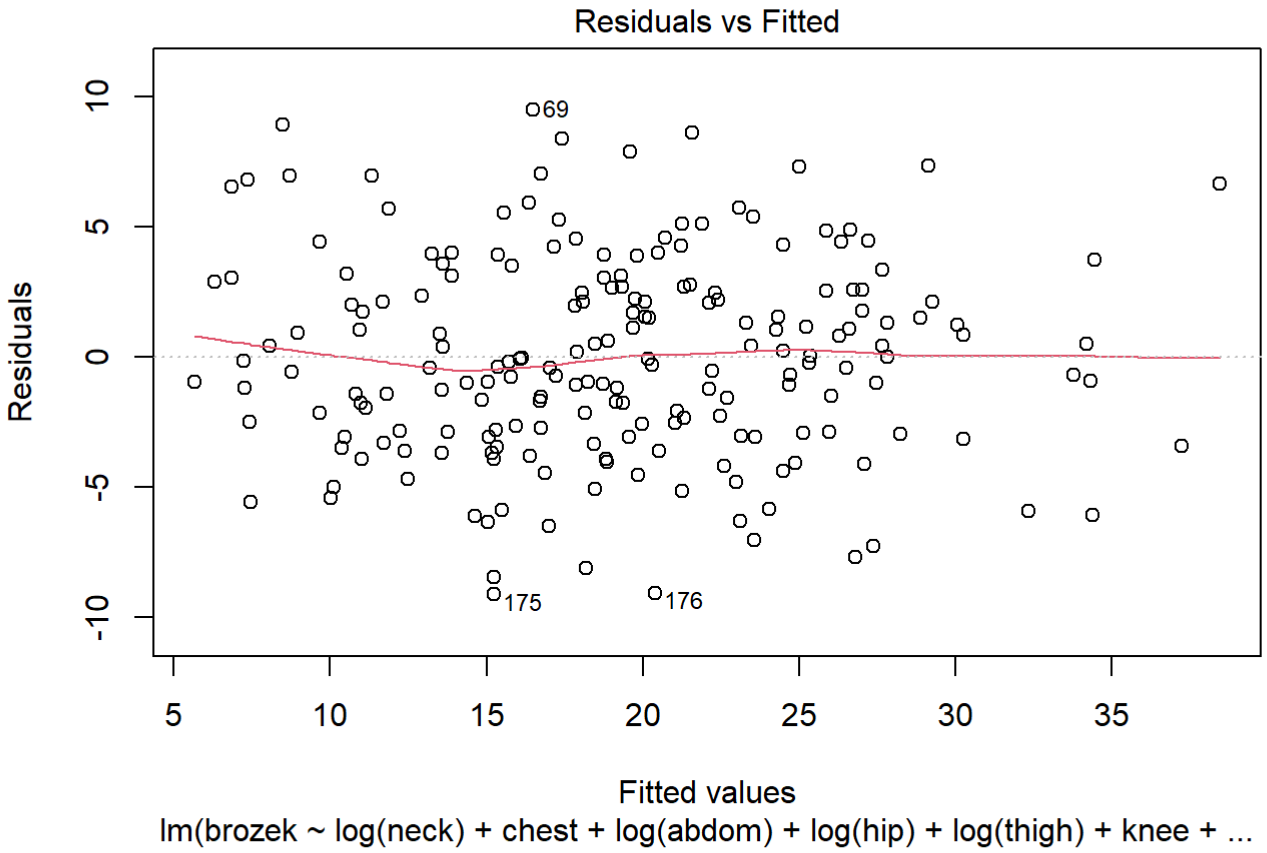
As a exploration, we plot plot the residual plot of full model here and find the offside of residual-fitted plot is below zero, which might caused by the positive skewness of some variables that we have identified above. The QQ-Plot looks fine which is assuring.

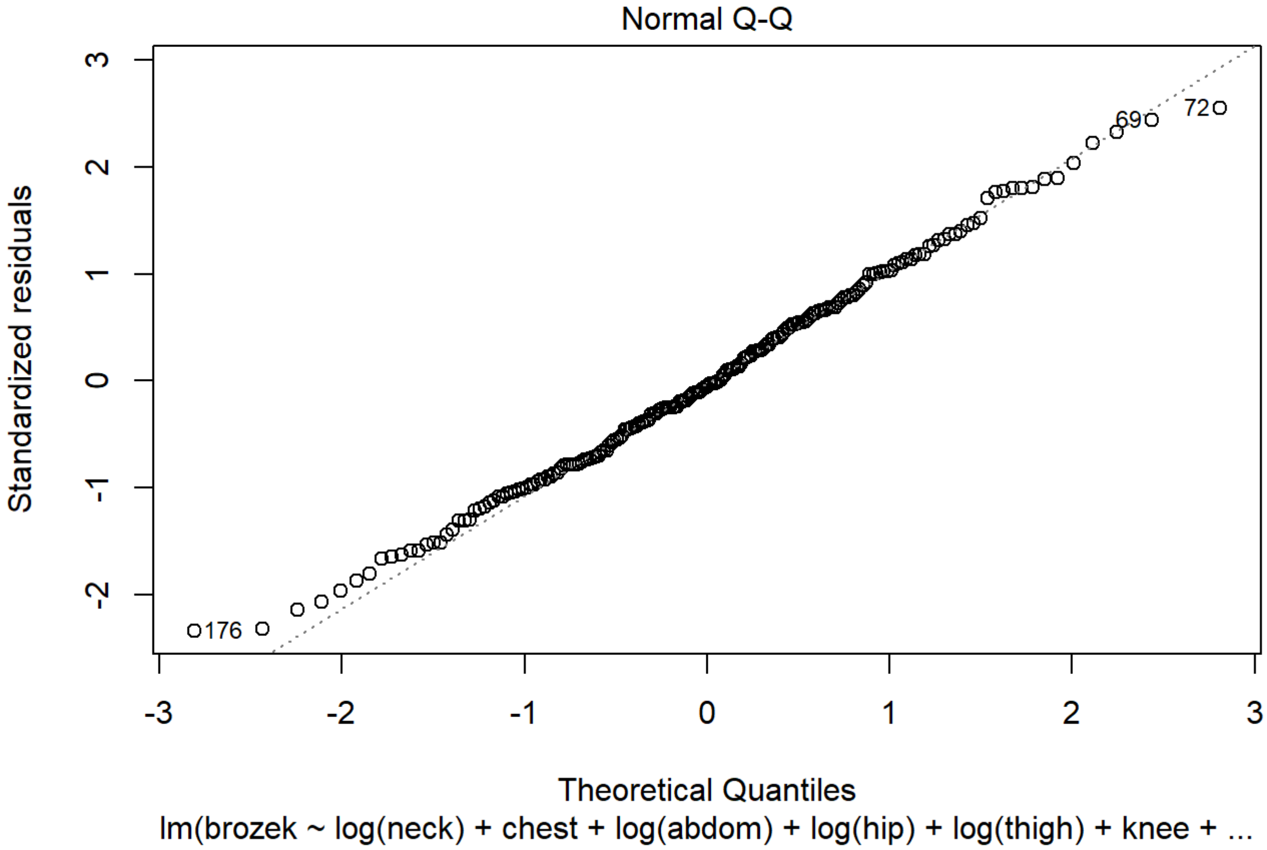




Although normally should not be perform here, We try to make transform to some variable based on the their skewness to see if we can get better result as we know there are skewness of some variables.

```
##
## Call:
## lm(formula = brozek ~ log(neck) + chest + log(abdom) + log(hip) +
##     log(thigh) + knee + log(ankle) + biceps + forearm + wrist,
##     data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1270 -2.8823 -0.2061  2.6739  9.4980
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -160.30875   43.47665  -3.687 0.000296 ***
## log(neck)    -23.42902    9.15251  -2.560 0.011249 *
## chest         -0.09793    0.08940  -1.095 0.274758
## log(abdom)    89.84217    6.96437  12.900  < 2e-16 ***
## log(hip)     -21.47214   13.16034  -1.632 0.104424
## log(thigh)    -1.54920    8.44816  -0.183 0.854697
## knee          -0.23996    0.22527  -1.065 0.288116
## log(ankle)    -3.09265    5.39617  -0.573 0.567242
## biceps         0.14427    0.16808   0.858 0.391808
## forearm        0.28411    0.19280   1.474 0.142252
## wrist         -1.13269    0.50595  -2.239 0.026333 *
## ---
## Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
##
## Residual standard error: 3.979 on 190 degrees of freedom
## Multiple R-squared:  0.7467, Adjusted R-squared:  0.7333
## F-statistic:    56 on 10 and 190 DF,  p-value: < 2.2e-16
```





The residual plot indeed give a better result now, particularly for the residual for large fitted values. However, we con not sure whether to perform it right now as the transformation make the explanation of model more problematic. However, we will remember this finding and try this transformation in the formal model selection.

Model selection

We now turn our focus to the formal model selection, we generally follow the so-called “Criterion-Based Procedure”.

Above all, we use the regsubsets() function in R which uses an algorithm called “exhaustive search” to perform best subset selection. The algorithm works by considering all possible models that can be formed from the available predictors, starting with the null model (which includes only the intercept) and progressing to the full model (which includes all predictors).

```
## Subset selection object
## Call: regsubsets.formula(brozek ~ ., data = Train)
## 10 Variables  (and intercept)
##         Forced in Forced out
## neck        FALSE      FALSE
## chest       FALSE      FALSE
## abdom       FALSE      FALSE
## hip         FALSE      FALSE
## thigh       FALSE      FALSE
## knee        FALSE      FALSE
## ankle       FALSE      FALSE
## biceps      FALSE      FALSE
## forearm     FALSE      FALSE
## wrist       FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          neck chest abdom hip thigh knee ankle biceps forearm wrist
## 1  ( 1 ) " "  " "   "*"   " " " "   " "  " "   " "    " "     " "  
## 2  ( 1 ) " "  " "   "*"   "*" " "   " "  " "   " "    " "     " "  
## 3  ( 1 ) "*"  " "   "*"   "*" " "   " "  " "   " "    " "     " "  
## 4  ( 1 ) "*"  " "   "*"   "*" " "   " "  " "   " "    " "     "*"  
## 5  ( 1 ) "*"  " "   "*"   "*" " "   " "  " "   " "    "*"     "*"  
## 6  ( 1 ) "*"  "*"   "*"   "*" " "   " "  " "   " "    "*"     "*"  
## 7  ( 1 ) "*"  "*"   "*"   "*" " "   " "  " "   "*"    "*"     "*"  
## 8  ( 1 ) "*"  "*"   "*"   "*" " "   "*"  " "   "*"    "*"     "*"  
```

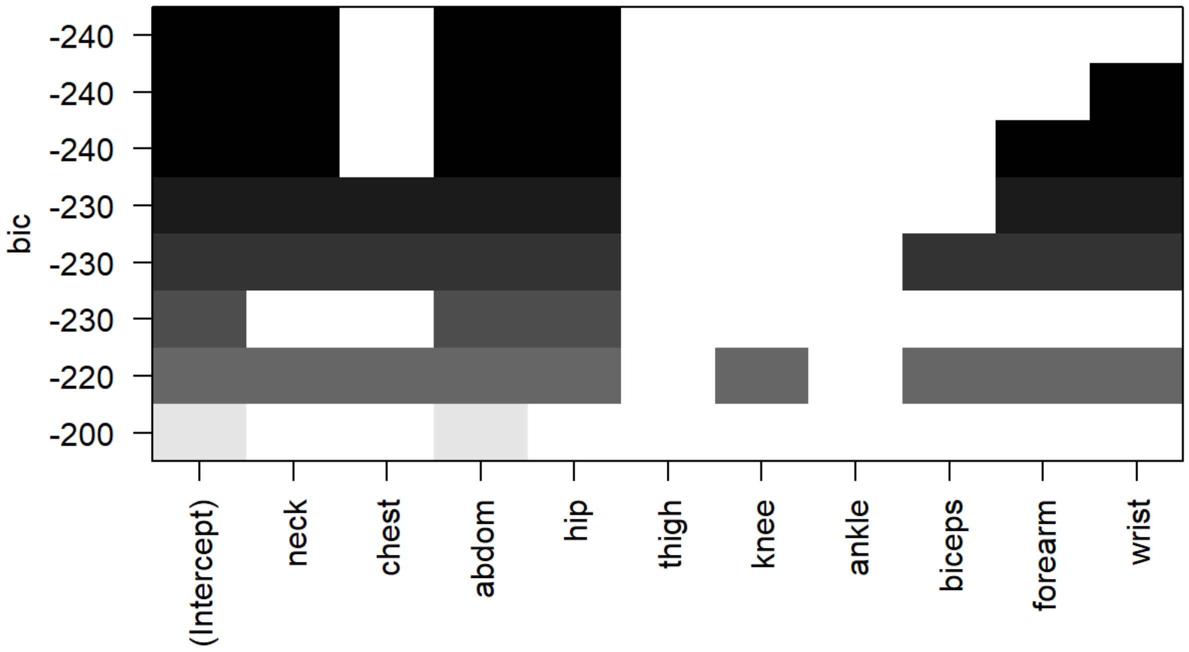
The result shows the order of including variables in our model. For example, we should build model using neck, hip and thigh if the criterion below believe the model with 3 variables is the best one if we use Mallow’s Cp as the criterion to choose candidate models.

Firstly,The AIC/BIC criteria naturally balance between fit and simplicity of model selection. The difference between two model is that BIC penalizes larger model more heavily so will prefer smaller models. We will see if that will make a difference.

we first use AIC criterion without the transformation on variable

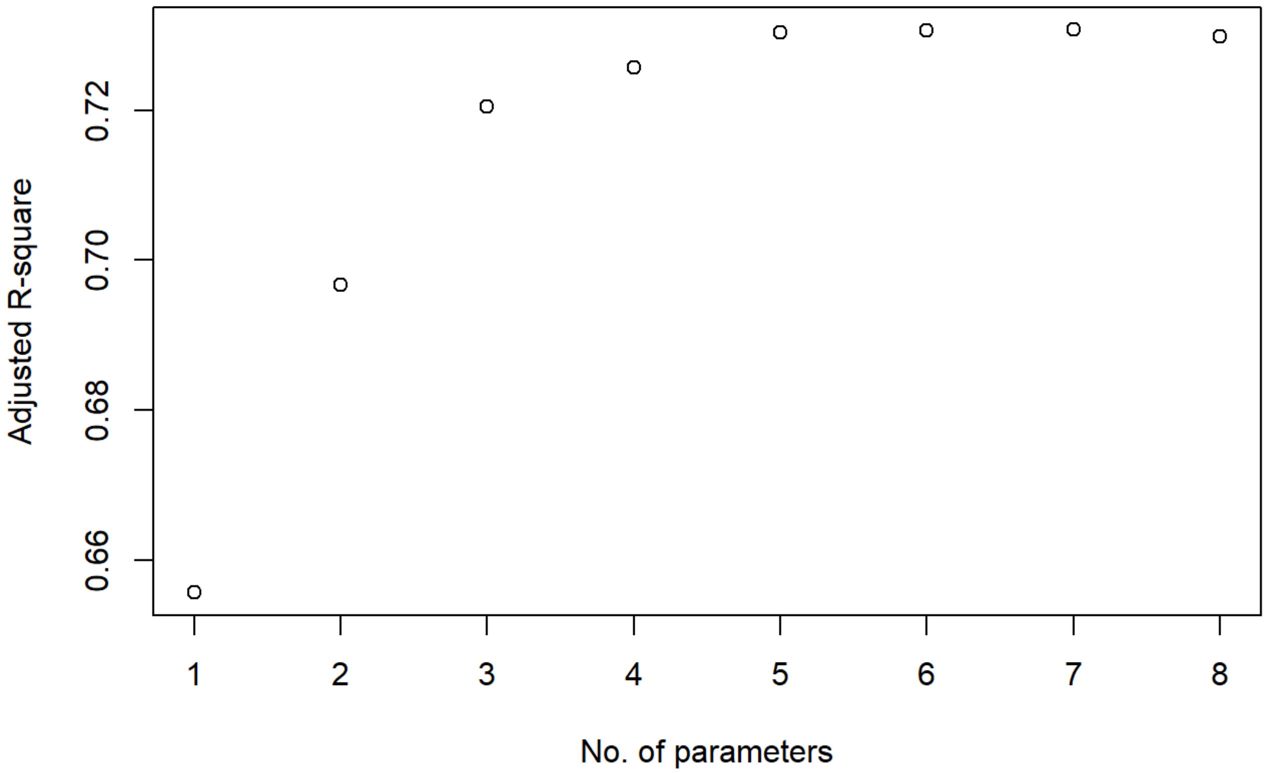
```
##
## Call:
## lm(formula = brozek ~ neck + abdom + hip + forearm + wrist, data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3029 -2.9009 -0.1711  2.9554 10.0308
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.30925    6.28436   0.845  0.39924
## neck        -0.67743    0.23371  -2.899  0.00418 **
## abdom        0.92489    0.05721  16.167  < 2e-16 ***
## hip         -0.35699    0.08416  -4.242 3.42e-05 ***
## forearm      0.37489    0.17799   2.106  0.03646 *
## wrist       -1.16945    0.46715  -2.503  0.01312 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.001 on 195 degrees of freedom
## Multiple R-squared:  0.7371, Adjusted R-squared:  0.7304
## F-statistic: 109.4 on 5 and 195 DF,  p-value: < 2.2e-16
```

The result of AIC shows that the best model should be (brozek ~ neck + abdom + hip + forearm + wrist). we then use BIC criterion without the transformation on variable



Both AIC and BIC method show that the model should include 3, 4 or 5 variables choose from neck, abdom, hip, forearm and wrist. This is predictable as BIC generally prefer simplier model. We will try all the three model below.

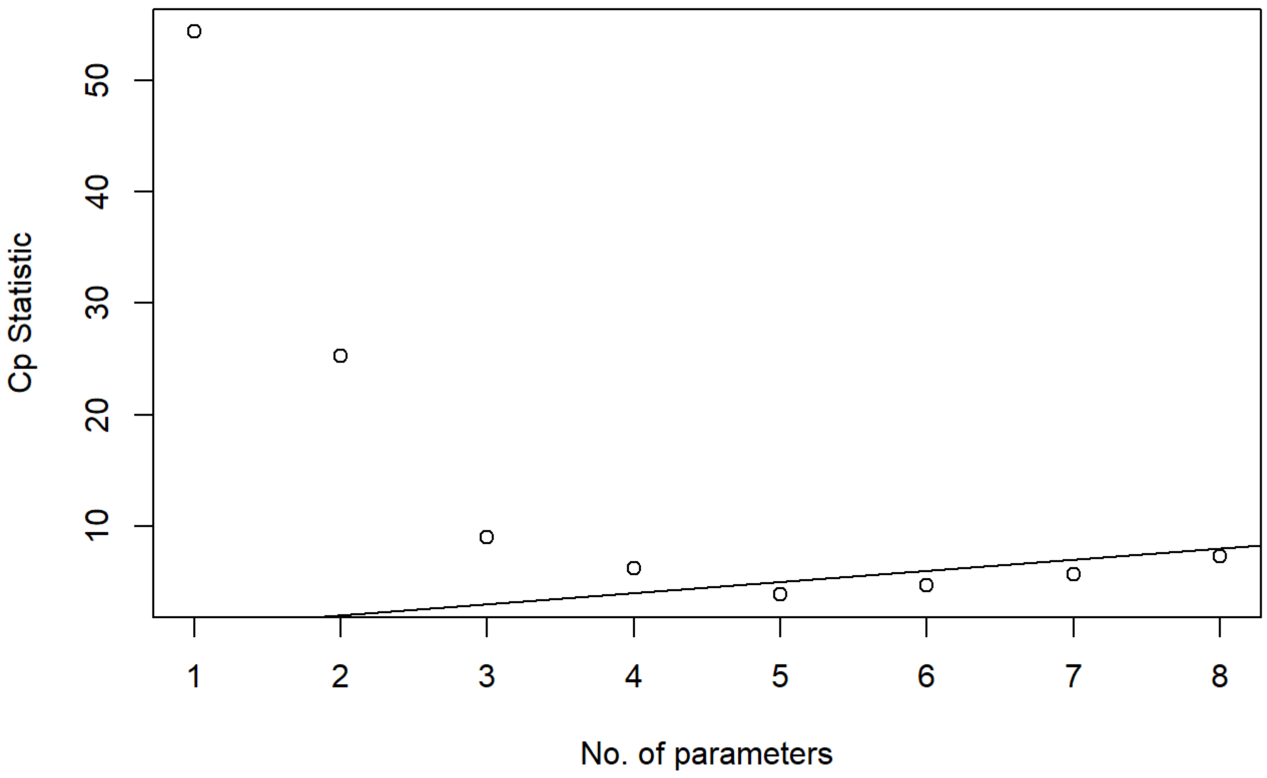
Secondly, we turn to another commonly used criterion called adjusted Adjusted R-square, which improved from the R-square. This criterion is more easy to understand, which also taken account into both fit and simplicity of model. We output the number of variables in the best model.



```
## [1] 7
```

Adjusted R-square shows the maximum occurs at using 7 parameters, in which case the model will exclude thigh, knee and ankle. Nevertheless, we should note that there is no obvious increase in value if increase from 3 to 7 parameters, thus Adjusted R-square do not give us particular information.

Thirdly, We consider Mallow’s Cp, which takes account of the average mean square error. It showed be point that all criterion we use trade-off fit in terms of RSS against complexity(p)



As the plot shows, the best model should use 5 parameters including neck, abdom, hip, wrist and forearm, which is the first point below Cp=p line, indicating the simplest model with good fits.

Although not particularly useful as there is not a high amount of computation, Stepwise Regression is finally used as a check.

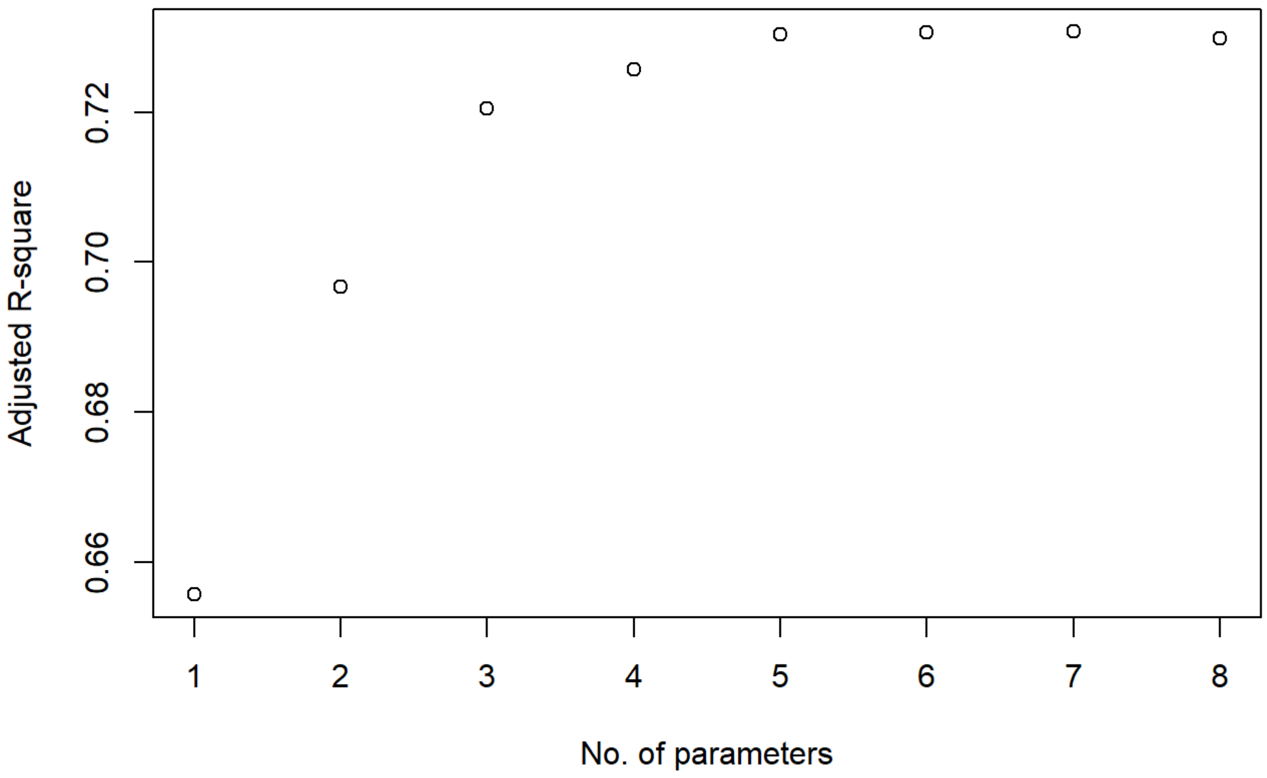
```
## Start:  AIC=570.32
## brozek ~ neck + chest + abdom + hip + thigh + knee + ankle + 
##     biceps + forearm + wrist
## 
##           Df Sum of Sq    RSS    AIC
## - thigh    1      0.01 3075.7 568.33
## - knee     1      3.78 3079.5 568.57
## - ankle    1      3.79 3079.5 568.57
## - biceps   1     16.75 3092.5 569.42
## - chest    1     22.73 3098.4 569.80
## <none>                 3075.7 570.32
## - forearm  1     54.28 3130.0 571.84
## - wrist    1     67.57 3143.3 572.69
## - hip      1    115.31 3191.0 575.72
## - neck     1    132.81 3208.5 576.82
## - abdom    1   2629.95 5705.6 692.53
## 
## Step:  AIC=568.33
## brozek ~ neck + chest + abdom + hip + knee + ankle + biceps + 
##     forearm + wrist
## 
##           Df Sum of Sq    RSS    AIC
## - ankle    1      3.81 3079.5 566.57
## - knee     1      4.01 3079.7 566.59
## - biceps   1     18.12 3093.8 567.51
## - chest    1     23.29 3099.0 567.84
## <none>                 3075.7 568.33
## - forearm  1     54.55 3130.3 569.86
## + thigh    1      0.01 3075.7 570.32
## - wrist    1     69.38 3145.1 570.81
## - neck     1    133.20 3208.9 574.85
## - hip      1    191.68 3267.4 578.48
## - abdom    1   2632.28 5708.0 690.61
## 
## Step:  AIC=566.57
## brozek ~ neck + chest + abdom + hip + knee + biceps + forearm + 
##     wrist
## 
##           Df Sum of Sq    RSS    AIC
## - knee     1      6.35 3085.9 564.99
## - biceps   1     17.92 3097.5 565.74
## - chest    1     24.31 3103.8 566.16
## <none>                 3079.5 566.57
## - forearm  1     54.62 3134.2 568.11
## + ankle    1      3.81 3075.7 568.33
## + thigh    1      0.03 3079.5 568.57
## - wrist    1     79.13 3158.7 569.67
## - neck     1    133.80 3213.3 573.12
## - hip      1    205.98 3285.5 577.59
## - abdom    1   2730.06 5809.6 692.16
## 
## Step:  AIC=564.99
## brozek ~ neck + chest + abdom + hip + biceps + forearm + wrist
## 
##           Df Sum of Sq    RSS    AIC
## - biceps   1     16.43 3102.3 564.06
## - chest    1     23.23 3109.1 564.50
## <none>                 3085.9 564.99
## - forearm  1     52.39 3138.3 566.37
## + knee     1      6.35 3079.5 566.57
## + ankle    1      6.15 3079.7 566.59
## + thigh    1      0.48 3085.4 566.96
## - wrist    1     97.33 3183.2 569.23
## - neck     1    131.52 3217.4 571.38
## - hip      1    290.26 3376.1 581.06
## - abdom    1   2725.32 5811.2 690.21
##
## Step:  AIC=564.06
## brozek ~ neck + chest + abdom + hip + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## - chest    1     19.40 3121.7 563.31
## <none>                 3102.3 564.06
## + biceps   1     16.43 3085.9 564.99
## + ankle    1      5.58 3096.7 565.69
## + knee     1      4.86 3097.5 565.74
## + thigh    1      0.37 3101.9 566.03
## - forearm  1     83.61 3185.9 567.40
## - wrist    1     90.36 3192.7 567.83
## - neck     1    119.31 3221.6 569.64
## - hip      1    274.62 3376.9 579.11
## - abdom    1   2709.60 5811.9 688.24
##
## Step:  AIC=563.31
## brozek ~ neck + abdom + hip + forearm + wrist
##
##           Df Sum of Sq    RSS    AIC
## <none>                 3121.7 563.31
## + chest    1      19.4 3102.3 564.06
## + biceps   1      12.6 3109.1 564.50
## + ankle    1       6.5 3115.2 564.89
## + knee     1       4.1 3117.6 565.04
## + thigh    1       1.5 3120.2 565.21
## - forearm  1      71.0 3192.7 565.83
## - wrist    1     100.3 3222.0 567.67
## - neck     1     134.5 3256.2 569.79
## - hip      1     288.0 3409.8 579.05
## - abdom    1    4184.1 7305.9 732.22
```

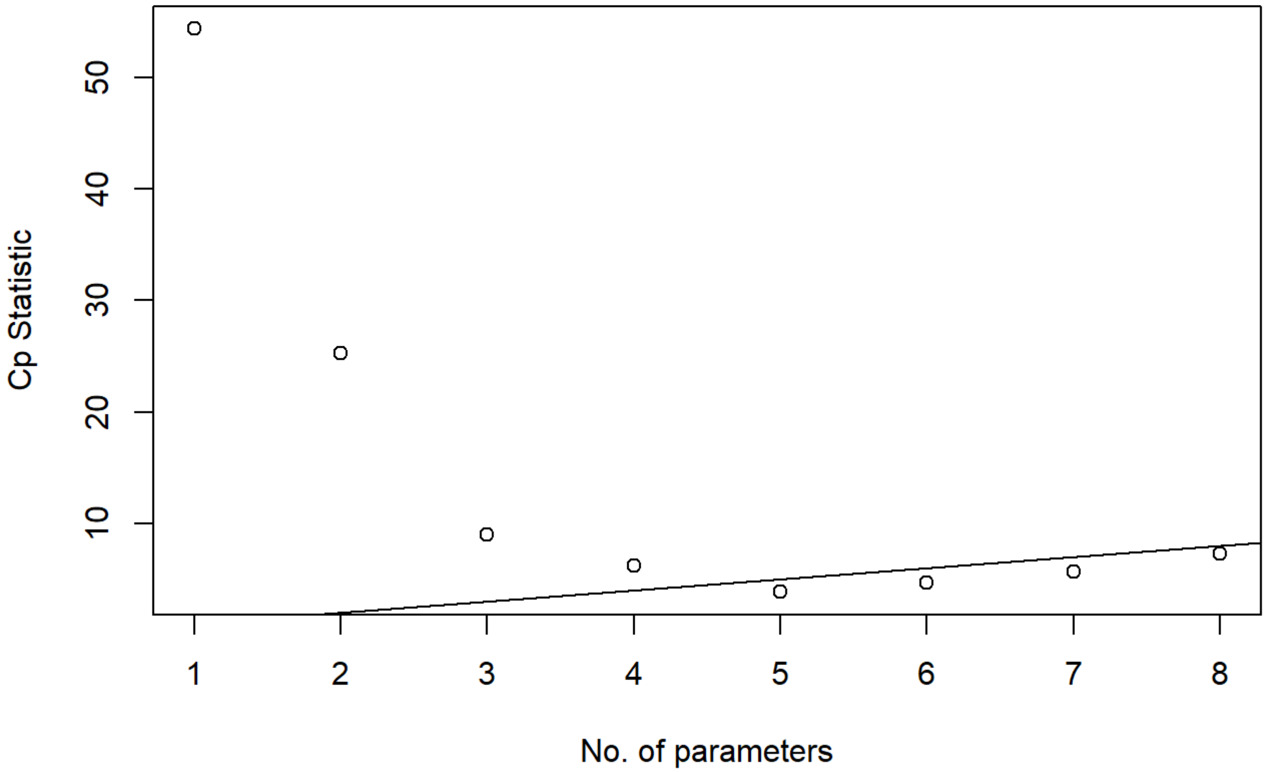
```
##
## Call:
## lm(formula = brozek ~ neck + abdom + hip + forearm + wrist, data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3029 -2.9009 -0.1711  2.9554 10.0308
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.30925    6.28436   0.845  0.39924
## neck        -0.67743    0.23371  -2.899  0.00418 **
## abdom        0.92489    0.05721  16.167  < 2e-16 ***
## hip         -0.35699    0.08416  -4.242 3.42e-05 ***
## forearm      0.37489    0.17799   2.106  0.03646 *
## wrist       -1.16945    0.46715  -2.503  0.01312 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.001 on 195 degrees of freedom
## Multiple R-squared:  0.7371, Adjusted R-squared:  0.7304
## F-statistic: 109.4 on 5 and 195 DF,  p-value: < 2.2e-16
```

The Stepwise Regression gives a result which is consistent to what we find before that the model with 5 parameters (brozek ~ neck + abdom + hip + forearm + wrist) is the best one.

As we have shown that the transformation might make fitting better, We then add the transformation to the model to check if the best model is different. This is useful as transformation may be changing the nature of the relationship between the variables and the response variable, which could affect the goodness of fit of the model.

```
##
## Call:
## lm(formula = brozek ~ log(neck) + log(abdom) + log(hip) + forearm +
##     wrist, data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1077 -2.8505 -0.1034  2.6718  9.4370
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -130.8915    27.2368  -4.806 3.07e-06 ***
## log(neck)    -23.3747     8.8498  -2.641 0.008929 **
## log(abdom)    85.1838     5.3056  16.055  < 2e-16 ***
## log(hip)     -29.1715     8.5425  -3.415 0.000776 ***
## forearm        0.2894     0.1789   1.618 0.107321
## wrist         -1.3382     0.4638  -2.885 0.004350 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.964 on 195 degrees of freedom
## Multiple R-squared:  0.742,  Adjusted R-squared:  0.7353
## F-statistic: 112.1 on 5 and 195 DF,  p-value: < 2.2e-16
```





We find the value of adjusted R-squared is very close from 4 parameter to 8 parameters. Also, Cp shows the best model should contain 5 parameters(brozek ~ log(neck) + log(abdom) + log(hip) + forearm + wrist). This is consistent with the finding above.

We also use Stepwise regression method to make further analysis.

```
## Start:  AIC=565.89
## brozek ~ log(neck) + chest + log(abdom) + log(hip) + log(thigh) + 
##     knee + log(ankle) + biceps + forearm + wrist
## 
##              Df Sum of Sq    RSS    AIC
## - log(thigh)  1      0.53 3009.1 563.92
## - log(ankle)  1      5.20 3013.8 564.24
## - biceps      1     11.66 3020.2 564.67
## - knee        1     17.97 3026.5 565.09
## - chest       1     19.00 3027.6 565.15
## <none>                    3008.6 565.89
## - forearm     1     34.38 3042.9 566.17
## - log(hip)    1     42.15 3050.7 566.68
## - wrist       1     79.36 3087.9 569.12
## - log(neck)   1    103.76 3112.3 570.70
## - log(abdom)  1   2635.13 5643.7 690.33
## 
## Step:  AIC=563.92
## brozek ~ log(neck) + chest + log(abdom) + log(hip) + knee + log(ankle) + 
##     biceps + forearm + wrist
## 
##              Df Sum of Sq    RSS    AIC
## - log(ankle)  1      5.40 3014.5 562.28
## - biceps      1     11.22 3020.3 562.67
## - chest       1     18.49 3027.6 563.16
## - knee        1     19.50 3028.6 563.22
## <none>                    3009.1 563.92
## - forearm     1     33.89 3043.0 564.18
## + log(thigh)  1      0.53 3008.6 565.89
## - log(hip)    1     78.78 3087.9 567.12
## - wrist       1     79.67 3088.8 567.18
## - log(neck)   1    104.86 3114.0 568.81
## - log(abdom)  1   2637.94 5647.0 688.45
## 
## Step:  AIC=562.28
## brozek ~ log(neck) + chest + log(abdom) + log(hip) + knee + biceps + 
##     forearm + wrist
## 
##              Df Sum of Sq    RSS    AIC
## - biceps      1     11.15 3025.6 561.03
## - chest       1     19.54 3034.0 561.58
## - knee        1     27.00 3041.5 562.08
## <none>                    3014.5 562.28
## - forearm     1     33.34 3047.8 562.49
## + log(ankle)  1      5.40 3009.1 563.92
## + log(thigh)  1      0.73 3013.8 564.24
## - log(hip)    1     87.79 3102.3 566.05
## - wrist       1     93.82 3108.3 566.44
## - log(neck)   1    105.37 3119.9 567.19
## - log(abdom)  1   2756.30 5770.8 690.81
## 
## Step:  AIC=561.03
## brozek ~ log(neck) + chest + log(abdom) + log(hip) + knee + forearm + 
##     wrist
## 
```

```
##              Df Sum of Sq    RSS    AIC
## - chest       1     16.64 3042.3 560.13
## - knee        1     24.69 3050.3 560.66
## <none>                    3025.6 561.03
## + biceps      1     11.15 3014.5 562.28
## - forearm     1     52.09 3077.7 562.46
## + log(ankle)  1      5.33 3020.3 562.67
## + log(thigh)  1      0.03 3025.6 563.02
## - log(hip)    1     78.17 3103.8 564.15
## - wrist       1     89.19 3114.8 564.87
## - log(neck)   1     96.72 3122.4 565.35
## - log(abdom)  1   2747.81 5773.5 688.90
## 
## Step:  AIC=560.13
## brozek ~ log(neck) + log(abdom) + log(hip) + knee + forearm + 
##     wrist
## 
##              Df Sum of Sq    RSS    AIC
## - knee        1      22.1 3064.4 559.58
## <none>                    3042.3 560.13
## + chest       1      16.6 3025.6 561.03
## - forearm     1      44.6 3086.9 561.05
## + biceps      1       8.3 3034.0 561.58
## + log(ankle)  1       6.3 3036.0 561.71
## + log(thigh)  1       0.7 3041.6 562.08
## - log(hip)    1      92.7 3135.0 564.16
## - wrist       1      97.8 3140.1 564.49
## - log(neck)   1     112.6 3154.9 565.43
## - log(abdom)  1    4072.1 7114.4 728.88
## 
## Step:  AIC=559.58
## brozek ~ log(neck) + log(abdom) + log(hip) + forearm + wrist
## 
##              Df Sum of Sq    RSS    AIC
## <none>                    3064.4 559.58
## + knee        1      22.1 3042.3 560.13
## - forearm     1      41.1 3105.5 560.26
## + chest       1      14.1 3050.3 560.66
## + log(ankle)  1      13.2 3051.2 560.72
## + biceps      1       6.6 3057.8 561.15
## + log(thigh)  1       0.0 3064.4 561.58
## - log(neck)   1     109.6 3174.0 564.65
## - wrist       1     130.8 3195.2 565.99
## - log(hip)    1     183.3 3247.7 569.26
## - log(abdom)  1    4050.9 7115.3 726.91
```

```
##
## Call:
## lm(formula = brozek ~ log(neck) + log(abdom) + log(hip) + forearm +
##     wrist, data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1077 -2.8505 -0.1034  2.6718  9.4370
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -130.8915    27.2368  -4.806 3.07e-06 ***
## log(neck)    -23.3747     8.8498  -2.641 0.008929 **
## log(abdom)    85.1838     5.3056  16.055  < 2e-16 ***
## log(hip)     -29.1715     8.5425  -3.415 0.000776 ***
## forearm        0.2894     0.1789   1.618 0.107321
## wrist         -1.3382     0.4638  -2.885 0.004350 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.964 on 195 degrees of freedom
## Multiple R-squared:  0.742,  Adjusted R-squared:  0.7353
## F-statistic: 112.1 on 5 and 195 DF,  p-value: < 2.2e-16
```

It shows the best model with 5 parameters, which is the consistent with the result of Cp and result before the Transformation.

All the method above suggest a good model should contain variables from: 1.{neck,abdom,hip} 2.{neck,abdom,hip,wrist} 3.{neck,abdom,hip,forearm}

We thus build model for each of them and try to choose a best one:

```
fit0<- lm(brozek ~  neck + chest + abdom + hip + thigh + knee + ankle+ biceps + forearm + wris
t,data=Train)
fit1<- lm(brozek~neck+abdom+hip,data=Train)
fit2<- lm(brozek~neck + abdom + hip +wrist, data=Train)
fit3<- lm(brozek~neck + abdom + hip + forearm + wrist,data=Train)
```

```
##
## Call:
## lm(formula = brozek ~ neck + chest + abdom + hip + thigh + knee +
##     ankle + biceps + forearm + wrist, data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3175 -2.7935 -0.2838  2.7578 10.1297
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.866246   6.844036   1.295  0.19673
## neck        -0.691259   0.241338  -2.864  0.00465 **
## chest       -0.108806   0.091820  -1.185  0.23750
## abdom        0.983531   0.077163  12.746  < 2e-16 ***
## hip         -0.339114   0.127061  -2.669  0.00827 **
## thigh       -0.004294   0.141129  -0.030  0.97576
## knee        -0.107833   0.223287  -0.483  0.62970
## ankle       -0.101910   0.210521  -0.484  0.62888
## biceps       0.173812   0.170848   1.017  0.31028
## forearm      0.357765   0.195377   1.831  0.06864 .
## wrist       -1.039871   0.508978  -2.043  0.04243 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.023 on 190 degrees of freedom
## Multiple R-squared:  0.741,  Adjusted R-squared:  0.7274
## F-statistic: 54.36 on 10 and 190 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = brozek ~ neck + abdom + hip, data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0242 -2.9805 -0.2037  2.8074  9.3424
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.74506    5.57608   0.134    0.894
## neck        -0.81970    0.19421  -4.221 3.72e-05 ***
## abdom        0.91934    0.05808  15.828  < 2e-16 ***
## hip         -0.35787    0.08457  -4.232 3.55e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.074 on 197 degrees of freedom
## Multiple R-squared:  0.7247, Adjusted R-squared:  0.7205
## F-statistic: 172.8 on 3 and 197 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = brozek ~ neck + abdom + hip + wrist, data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4052 -2.8903 -0.1877  2.8081  8.8491
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.20700    6.27372   1.149   0.2521
## neck        -0.55332    0.22814  -2.425   0.0162 *
## abdom        0.91627    0.05756  15.918  < 2e-16 ***
## hip         -0.33681    0.08434  -3.993 9.21e-05 ***
## wrist       -1.01051    0.46504  -2.173   0.0310 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.036 on 196 degrees of freedom
## Multiple R-squared:  0.7311, Adjusted R-squared:  0.7257
## F-statistic: 133.3 on 4 and 196 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = brozek ~ neck + abdom + hip + forearm + wrist, data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3029 -2.9009 -0.1711  2.9554 10.0308
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.30925    6.28436   0.845  0.39924
## neck        -0.67743    0.23371  -2.899  0.00418 **
## abdom        0.92489    0.05721  16.167  < 2e-16 ***
## hip         -0.35699    0.08416  -4.242 3.42e-05 ***
## forearm      0.37489    0.17799   2.106  0.03646 *
## wrist       -1.16945    0.46715  -2.503  0.01312 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.001 on 195 degrees of freedom
## Multiple R-squared:  0.7371, Adjusted R-squared:  0.7304
## F-statistic: 109.4 on 5 and 195 DF,  p-value: < 2.2e-16
```

we finally choose the model with five variables namely fit3 for now because: 1.it give us the best Adjusted R-squared and Residual standard error among three model. 2.Other two model are nested in it. We can use F-test to determine if the more complex model provides a significantly better fit to the data than the simpler model.

Combining with the result above from Mallow’s Cp and Stepwise regression, the fit3 (brozek ~ neck + abdom + hip + forearm + wrist) is the best one for now is highly possible.

We now perform F-test to determine if fit3 provides a significantly better fit to the data than the fit1 and fit2

```
## Analysis of Variance Table
##
## Model 1: brozek ~ neck + abdom + hip
## Model 2: brozek ~ neck + abdom + hip + wrist
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    197 3269.6
## 2    196 3192.7  1    76.915 4.7218 0.03098 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Variance Table
##
## Model 1: brozek ~ neck + abdom + hip + wrist
## Model 2: brozek ~ neck + abdom + hip + forearm + wrist
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    196 3192.7
## 2    195 3121.7  1    71.021 4.4364 0.03646 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

With the p-value for both group below 4%, we can conclude that the more complex model provides a significantly better fit to the data than the simpler model at 5% significant level.

The models with transformation are also tried below to check if it indeed make better fit.

```
fit01<- lm(brozek ~ log(neck) + chest + log(abdom) + log(hip) + log(thigh) + knee + log(ankle)+
biceps + forearm + wrist,data=Train)
fit11<- lm(brozek~log(neck) + log(abdom) + hip,data=Train)
fit21<- lm(brozek~ log(neck) + log(abdom) + hip + wrist, data=Train)
fit31<- lm(brozek ~ log(neck) + log(abdom) + log(hip) + forearm + wrist,data = Train)
```

```
##
## Call:
## lm(formula = brozek ~ log(neck) + chest + log(abdom) + log(hip) +
##     log(thigh) + knee + log(ankle) + biceps + forearm + wrist,
##     data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1270 -2.8823 -0.2061  2.6739  9.4980
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -160.30875   43.47665  -3.687 0.000296 ***
## log(neck)    -23.42902    9.15251  -2.560 0.011249 *
## chest         -0.09793    0.08940  -1.095 0.274758
## log(abdom)    89.84217    6.96437  12.900  < 2e-16 ***
## log(hip)     -21.47214   13.16034  -1.632 0.104424
## log(thigh)    -1.54920    8.44816  -0.183 0.854697
## knee          -0.23996    0.22527  -1.065 0.288116
## log(ankle)    -3.09265    5.39617  -0.573 0.567242
## biceps         0.14427    0.16808   0.858 0.391808
## forearm        0.28411    0.19280   1.474 0.142252
## wrist         -1.13269    0.50595  -2.239 0.026333 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.979 on 190 degrees of freedom
## Multiple R-squared:  0.7467, Adjusted R-squared:  0.7333
## F-statistic:    56 on 10 and 190 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = brozek ~ log(neck) + log(abdom) + hip, data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6847 -2.9714 -0.1226  2.5452 10.0933
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -220.68602   22.39751  -9.853  < 2e-16 ***
## log(neck)    -31.43595    7.32997  -4.289 2.81e-05 ***
## log(abdom)    84.78487    5.26966  16.089  < 2e-16 ***
## hip           -0.29459    0.07934  -3.713 0.000267 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.032 on 197 degrees of freedom
## Multiple R-squared:  0.7304, Adjusted R-squared:  0.7263
## F-statistic: 177.9 on 3 and 197 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = brozek ~ log(neck) + log(abdom) + hip + wrist, data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0367 -2.9161 -0.2283  2.8163  9.3701
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -245.46746   23.82204 -10.304  < 2e-16 ***
## log(neck)    -18.91569    8.53868  -2.215 0.027891 *
## log(abdom)    84.68942    5.18491  16.334  < 2e-16 ***
## hip           -0.26882    0.07863  -3.419 0.000765 ***
## wrist         -1.25627    0.45868  -2.739 0.006734 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.967 on 196 degrees of freedom
## Multiple R-squared:  0.7403, Adjusted R-squared:  0.735
## F-statistic: 139.7 on 4 and 196 DF,  p-value: < 2.2e-16
```

```
##
## Call:
## lm(formula = brozek ~ log(neck) + log(abdom) + log(hip) + forearm +
##     wrist, data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1077 -2.8505 -0.1034  2.6718  9.4370
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -130.8915    27.2368  -4.806 3.07e-06 ***
## log(neck)    -23.3747     8.8498  -2.641 0.008929 **
## log(abdom)    85.1838     5.3056  16.055  < 2e-16 ***
## log(hip)     -29.1715     8.5425  -3.415 0.000776 ***
## forearm        0.2894     0.1789   1.618 0.107321
## wrist         -1.3382     0.4638  -2.885 0.004350 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.964 on 195 degrees of freedom
## Multiple R-squared:  0.742,  Adjusted R-squared:  0.7353
## F-statistic: 112.1 on 5 and 195 DF,  p-value: < 2.2e-16
```

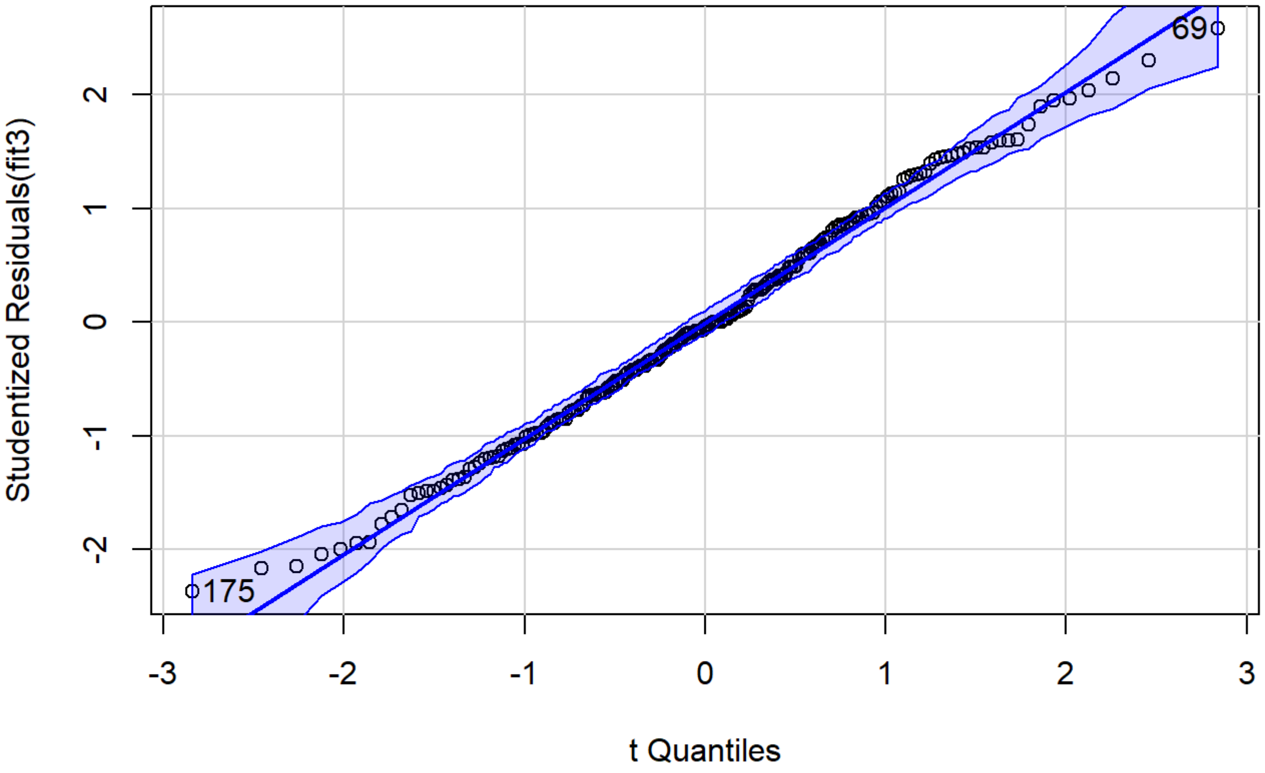
The further check confirm that the transformation indeed make the model estimate better with better Residual standard error and Adjusted R-squared. We will remember this and try this model in the next part, Model checking and validation

Model checking and validation

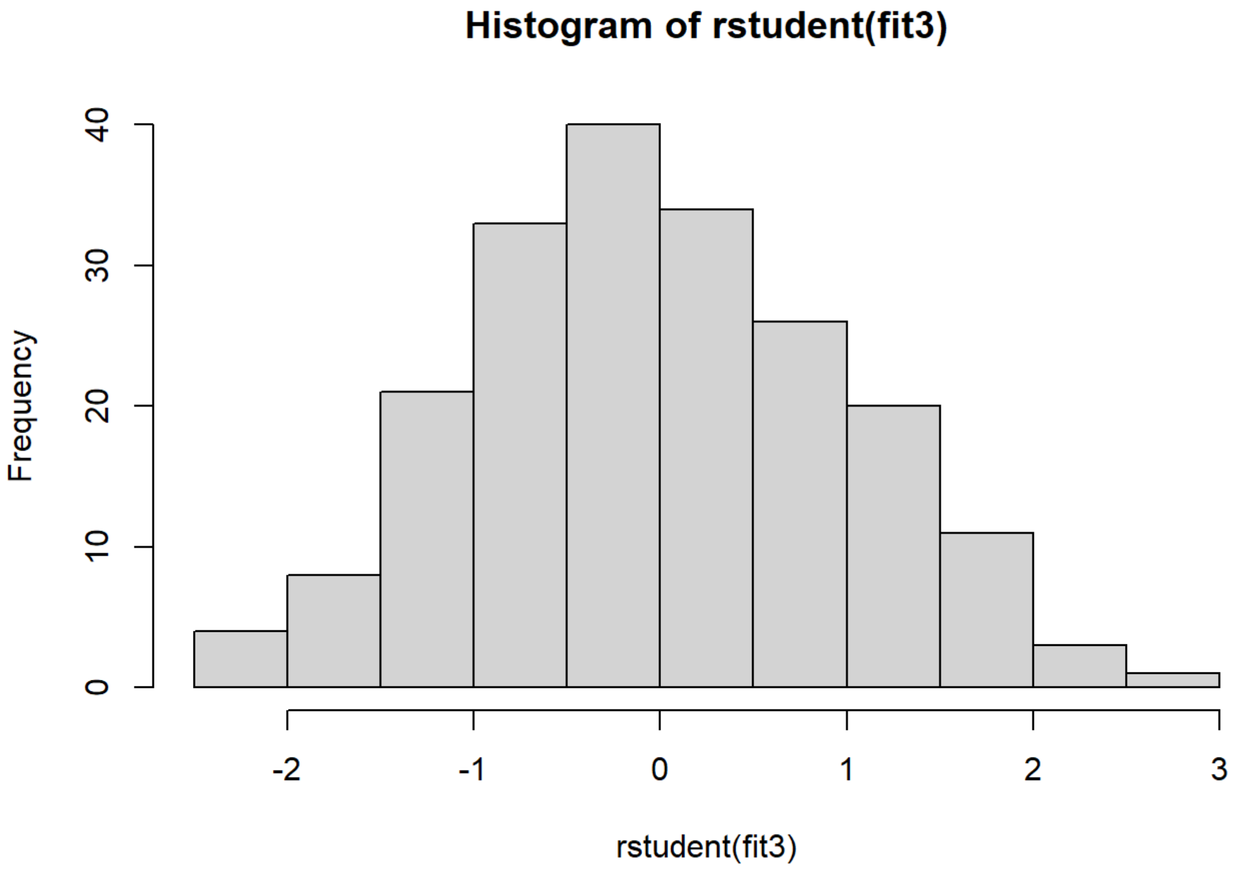
We have finish the model selection by now, next to do is to check if meet all the assumption of linear regression. There are four assumption in linear regression, namely:

1. Linear relationship
2. Independence of residual
3. Homoscedasticity(The residuals have constant variance at every level of x)
4. Normality(he residuals of the model are normally distributed)

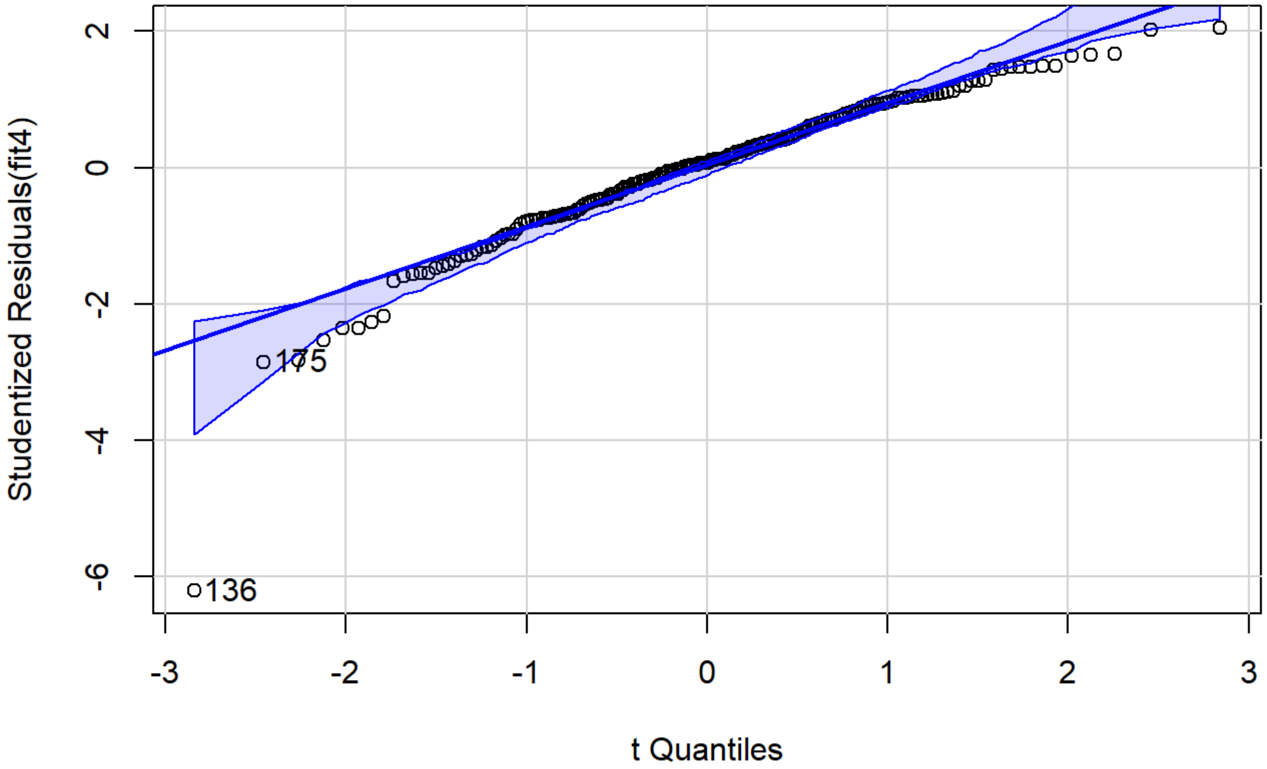
We will check them from now. Firstly, we check the QQ-plot and the histogram of the residuals.



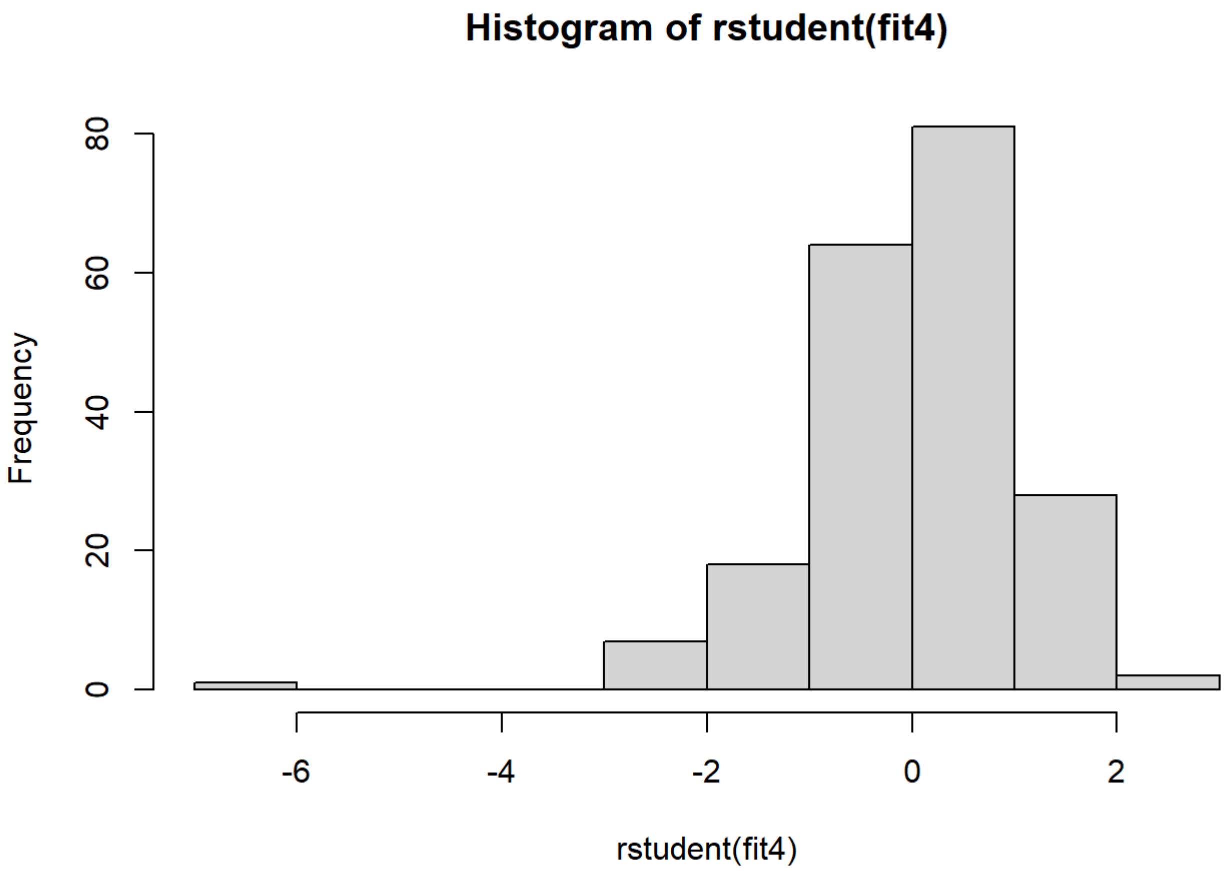
```
## [1]  69 175
```



The QQ-plot and histogram of suggest a small degree of positive skew. We give a try to fix this by transform the dependent variable using log.

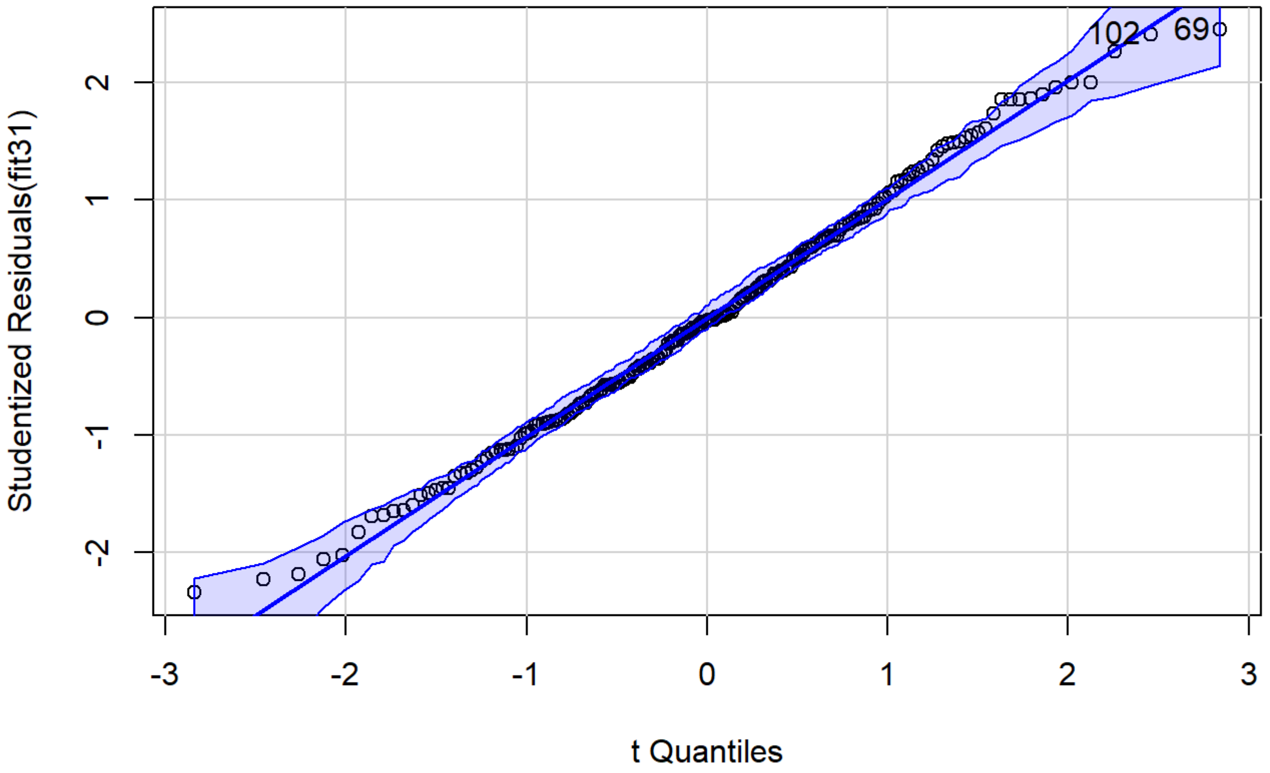


```
## [1] 136 175
```

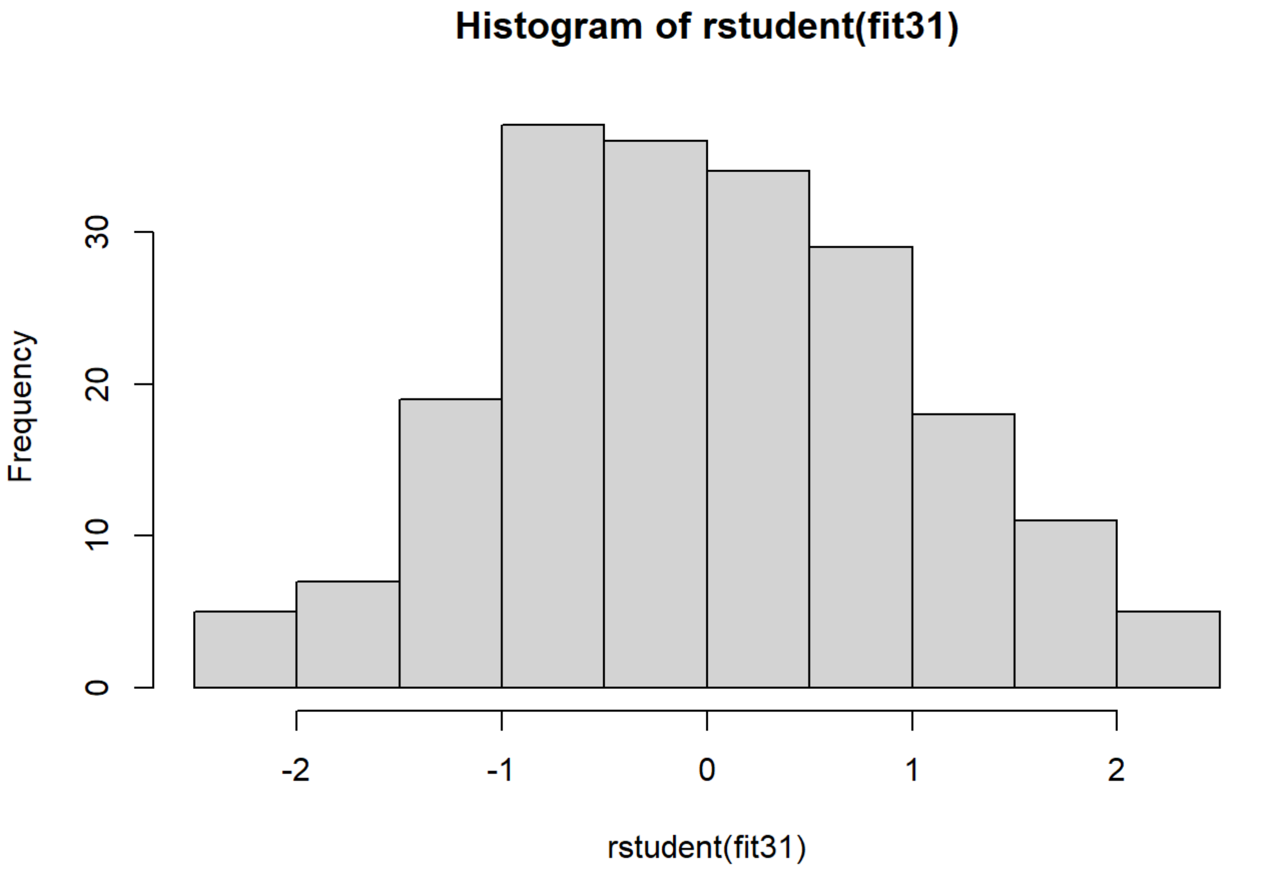


The results of QQ-plot and histogram are not satisfying, we disuse this transformation.

The transformation to some independent variable could now be checked again

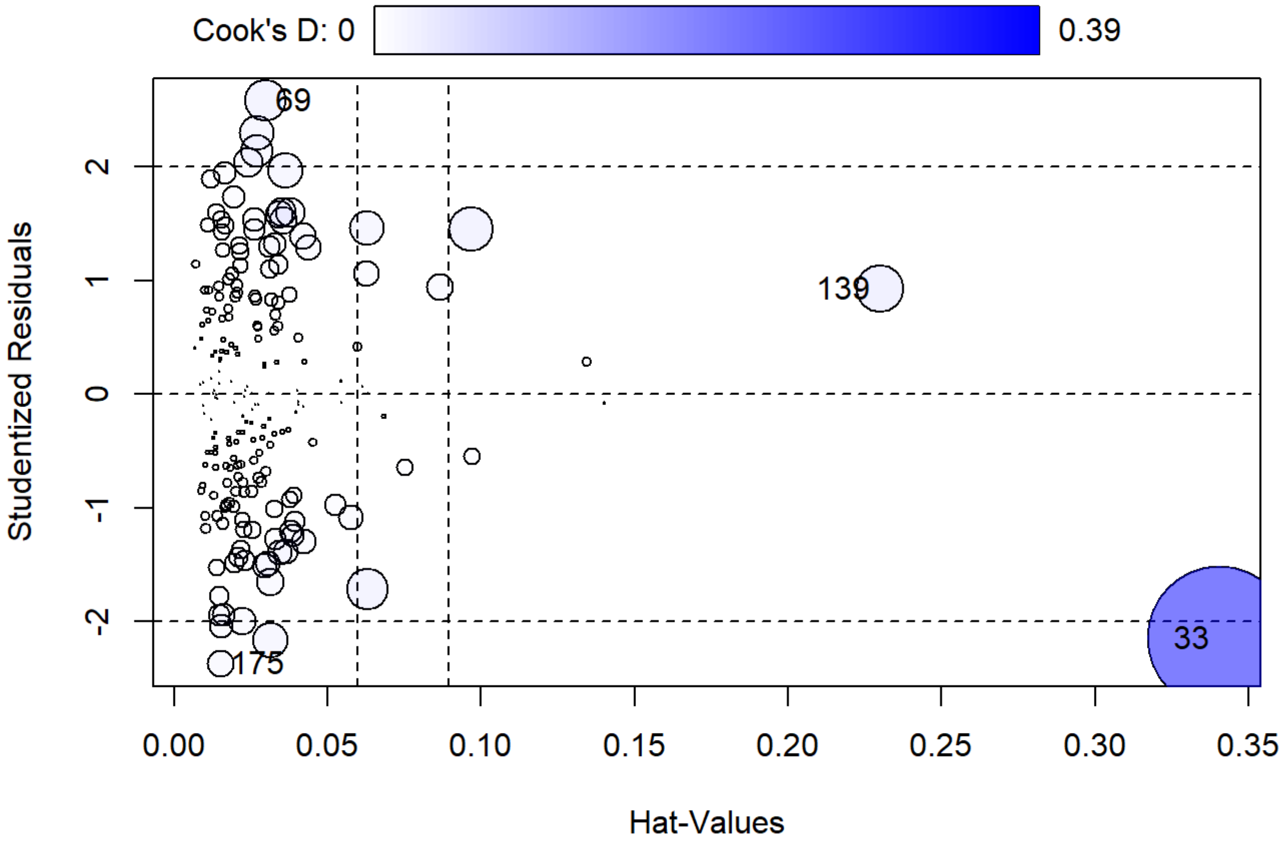

Studentized Residuals(fit31)
t Quantiles
-3
-2
-1
0
1
2
3
-2
-1
0
1
2
102
69

```
## [1]  69 102
```

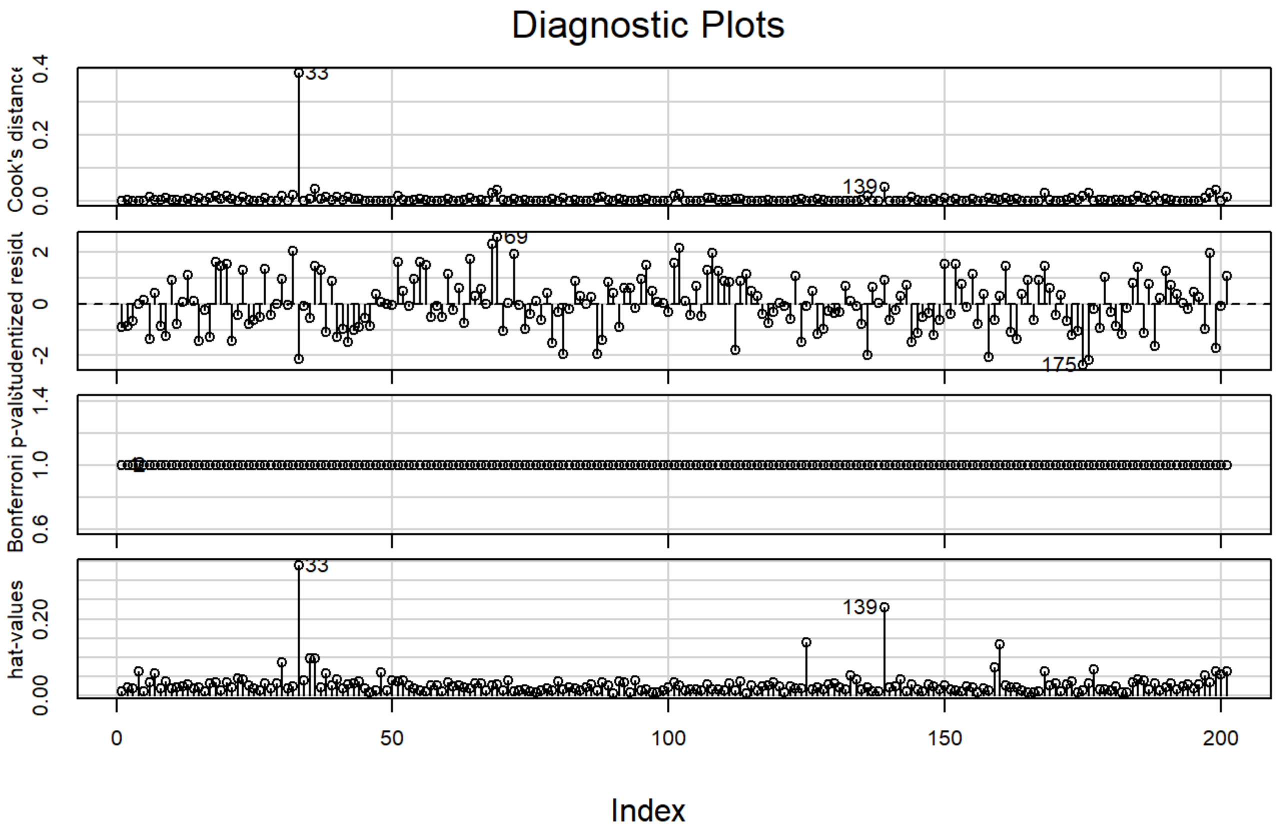

Histogram of rstudent(fit31)
Frequency
0
10
20
30
-2
-1
0
1
2
rstudent(fit31)

There is no particular progress after transformation.

Secondly, we turn our attention to the problem of data itself by finding high leverage points and outliers



```
##         StudRes        Hat     CookD
## 33  -2.1486757 0.34046983 0.38998966
## 69   2.5819692 0.02982927 0.03319747
## 139  0.9253069 0.22998693 0.04265263
## 175 -2.3705037 0.01515979 0.01408282
```



The ‘influence plot’ shows outlyingness, leverage and influence (represented as CookD distance) sepaartely. In this case: (Hat):high leverage point in this case is point with has pii>2p/n=12/200 = 0.06,we can see #33 and ##139 could be regarded as high leverage (studRes):outliers could be checked using StudRes. Generally,|ti|>2 is considered to be large, we find #33, #69 and #175 are outliers. By the analysis above and as the graph shows, point 33 should be deleted and we make further try on the basis of that.

we first delete the line 33 and check the result

```
##
## Call:
## lm(formula = brozek ~ neck + abdom + hip + wrist + forearm, data = Train1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9962 -2.8339 -0.1931  2.9506  9.5926
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.07359    6.53146   0.164  0.86961
## neck        -0.54950    0.23911  -2.298  0.02262 *
## abdom        0.90246    0.05764  15.657  < 2e-16 ***
## hip         -0.28481    0.08991  -3.168  0.00178 **
## wrist       -1.26716    0.46510  -2.724  0.00703 **
## forearm      0.23718    0.18765   1.264  0.20777
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.965 on 194 degrees of freedom
## Multiple R-squared:  0.7385, Adjusted R-squared:  0.7318
## F-statistic: 109.6 on 5 and 194 DF,  p-value: < 2.2e-16
```

Thus we believe we need to remove the line 33. Now we check if there are more data that need to be removed. We then delete the line 33 and 69.

```
##
## Call:
## lm(formula = brozek ~ neck + abdom + hip + wrist + forearm, data = Train2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9237 -2.6821 -0.1293  2.8856  9.1644
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.50795    6.44942   0.079  0.93731
## neck        -0.60999    0.23720  -2.572  0.01088 *
## abdom        0.90687    0.05691  15.936  < 2e-16 ***
## hip         -0.28879    0.08874  -3.255  0.00134 **
## wrist       -1.22328    0.45931  -2.663  0.00839 **
## forearm      0.30726    0.18730   1.640  0.10253
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.912 on 193 degrees of freedom
## Multiple R-squared:  0.7456, Adjusted R-squared:  0.739
## F-statistic: 113.1 on 5 and 193 DF,  p-value: < 2.2e-16
```

Residual standard error and R-adjusted become better. We choose to try to delete line 69 before fit.

We thirdly try to delete the line 33 and 139.

```
##
## Call:
## lm(formula = brozek ~ neck + abdom + hip + wrist + forearm, data = Train3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0215 -2.8248 -0.1137  2.9589  9.7127
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.46678    6.63577   0.221  0.82529
## neck        -0.56247    0.24231  -2.321  0.02132 *
## abdom        0.90520    0.05826  15.537  < 2e-16 ***
## hip         -0.29329    0.09311  -3.150  0.00189 **
## wrist       -1.29282    0.47151  -2.742  0.00668 **
## forearm      0.27738    0.21846   1.270  0.20571
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.973 on 193 degrees of freedom
## Multiple R-squared:  0.738,  Adjusted R-squared:  0.7312
## F-statistic: 108.7 on 5 and 193 DF,  p-value: < 2.2e-16
```

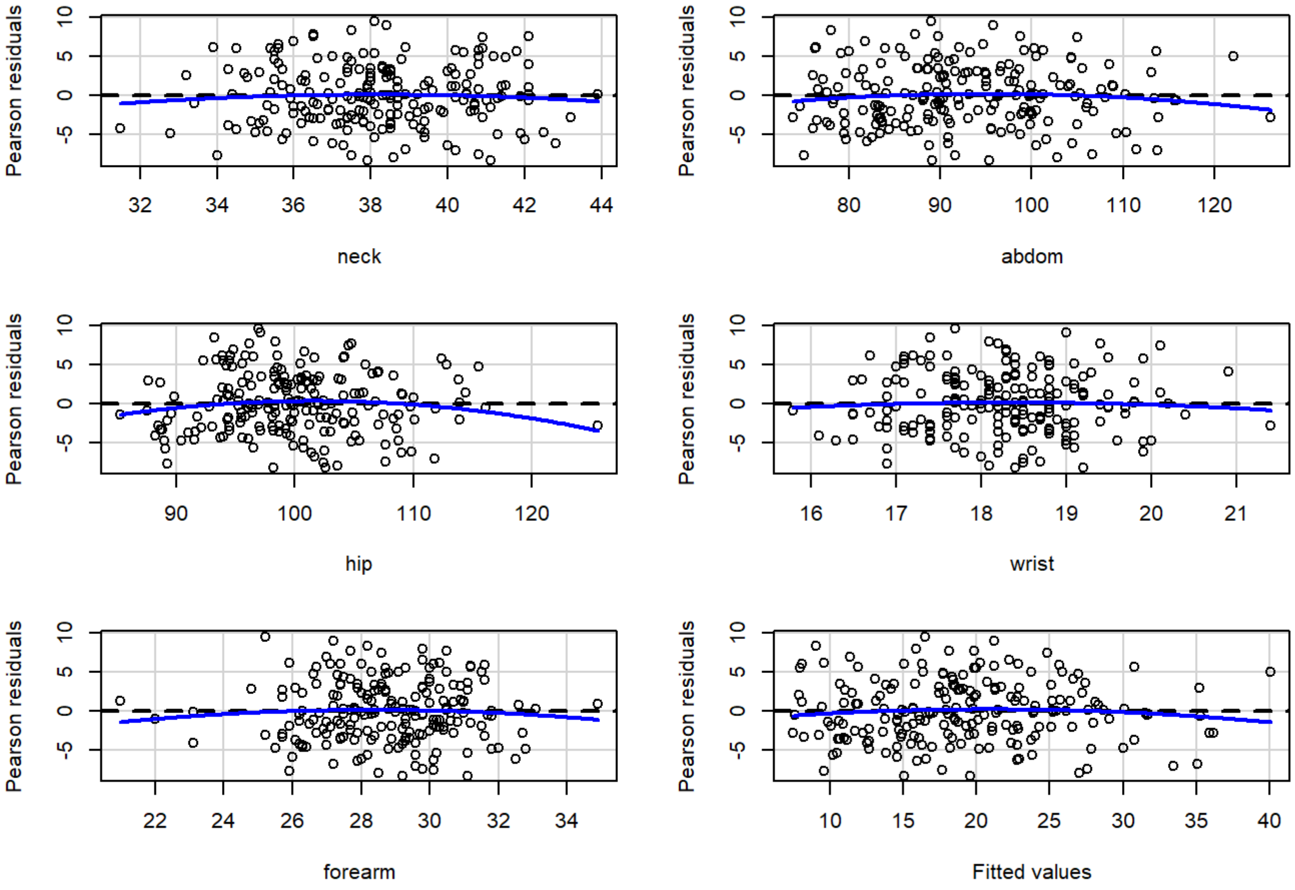
The result dose not getting better.

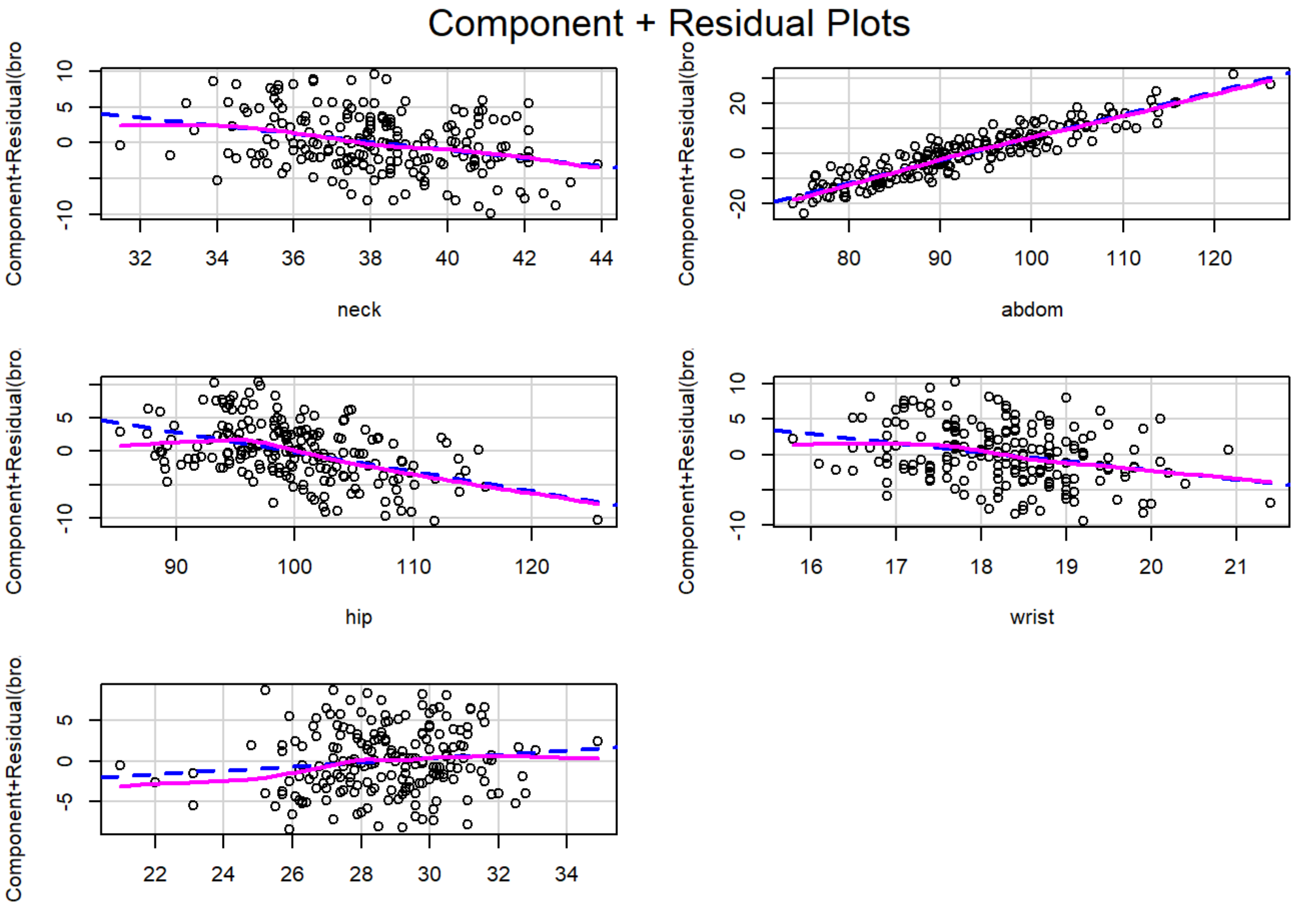
We finally try to delete the line 33 and 175.

```
##
## Call:
## lm(formula = brozek ~ neck + abdom + hip + wrist + forearm, data = Train4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2716 -2.8225 -0.1397  2.8648  9.5257
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.65042    6.49526   0.408  0.68369
## neck        -0.56314    0.23654  -2.381  0.01825 *
## abdom        0.90616    0.05702  15.891  < 2e-16 ***
## hip         -0.29244    0.08897  -3.287  0.00120 **
## wrist       -1.30316    0.46023  -2.832  0.00512 **
## forearm      0.23945    0.18558   1.290  0.19849
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.921 on 193 degrees of freedom
## Multiple R-squared:  0.7418, Adjusted R-squared:  0.7351
## F-statistic: 110.9 on 5 and 193 DF,  p-value: < 2.2e-16
```

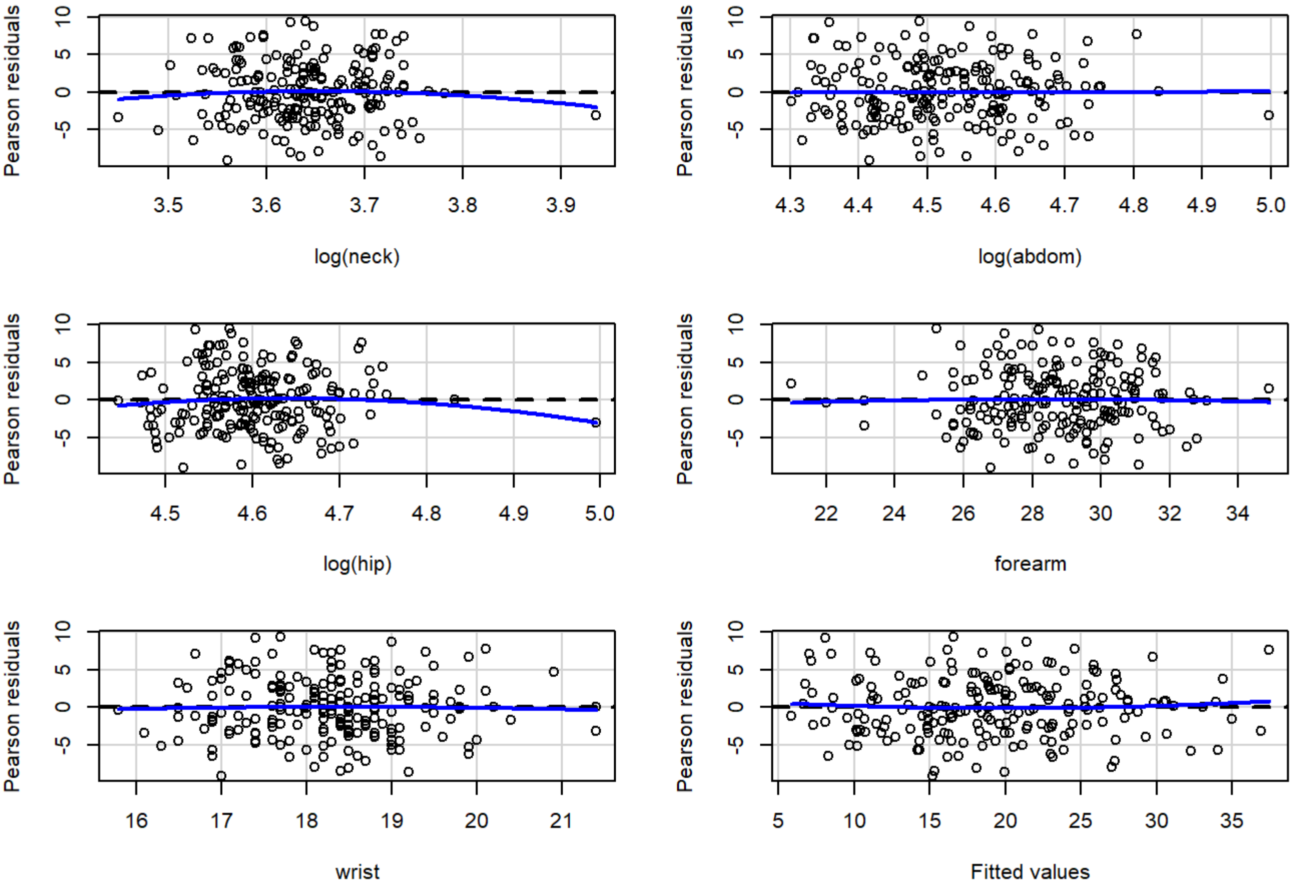
Delete data more than line 33 and line 69 will not make the Residual standard error and Adjusted R-squared better. Thus, we only delete line 33 and 69.

Thirdly, we check the residual plot check if the error of the regression really meet assumption.

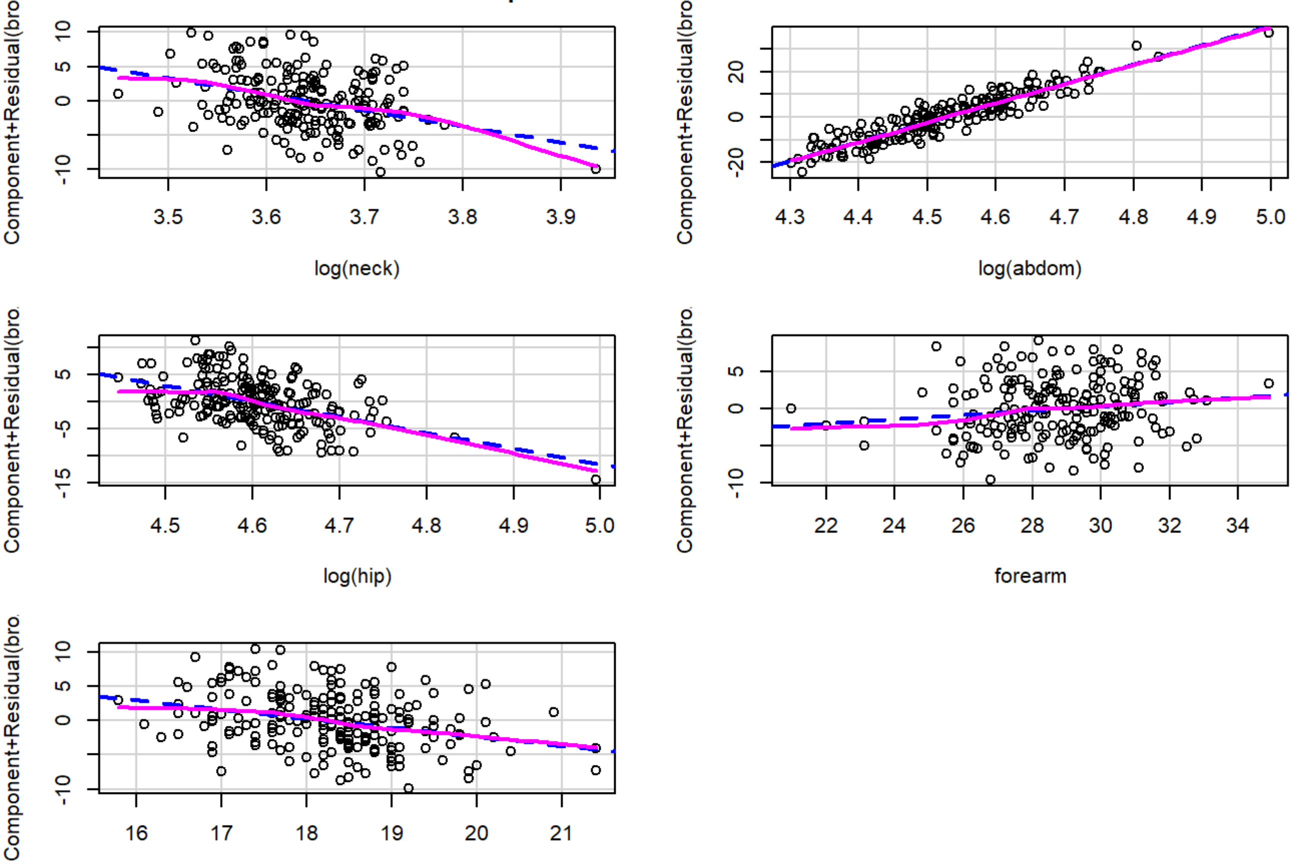




We find the plots do not meet the assumption of error. Particularly, linearity and independence of errors. There are some non-linearity and non-independence, especially how the response depend on abdom, hip and forearm. As before, we try to transforming some variables by moving down the ladder of powers.



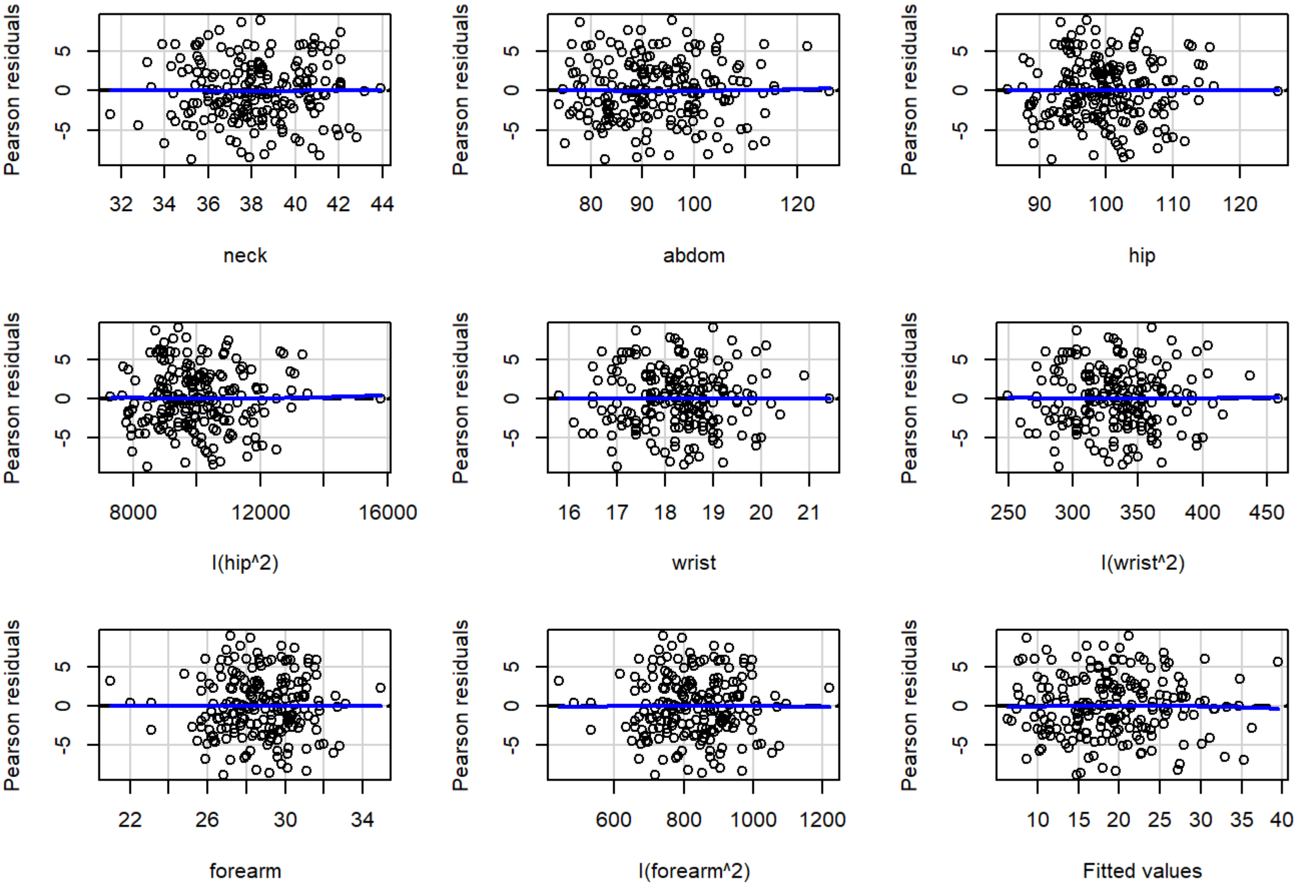
Component + Residual Plots

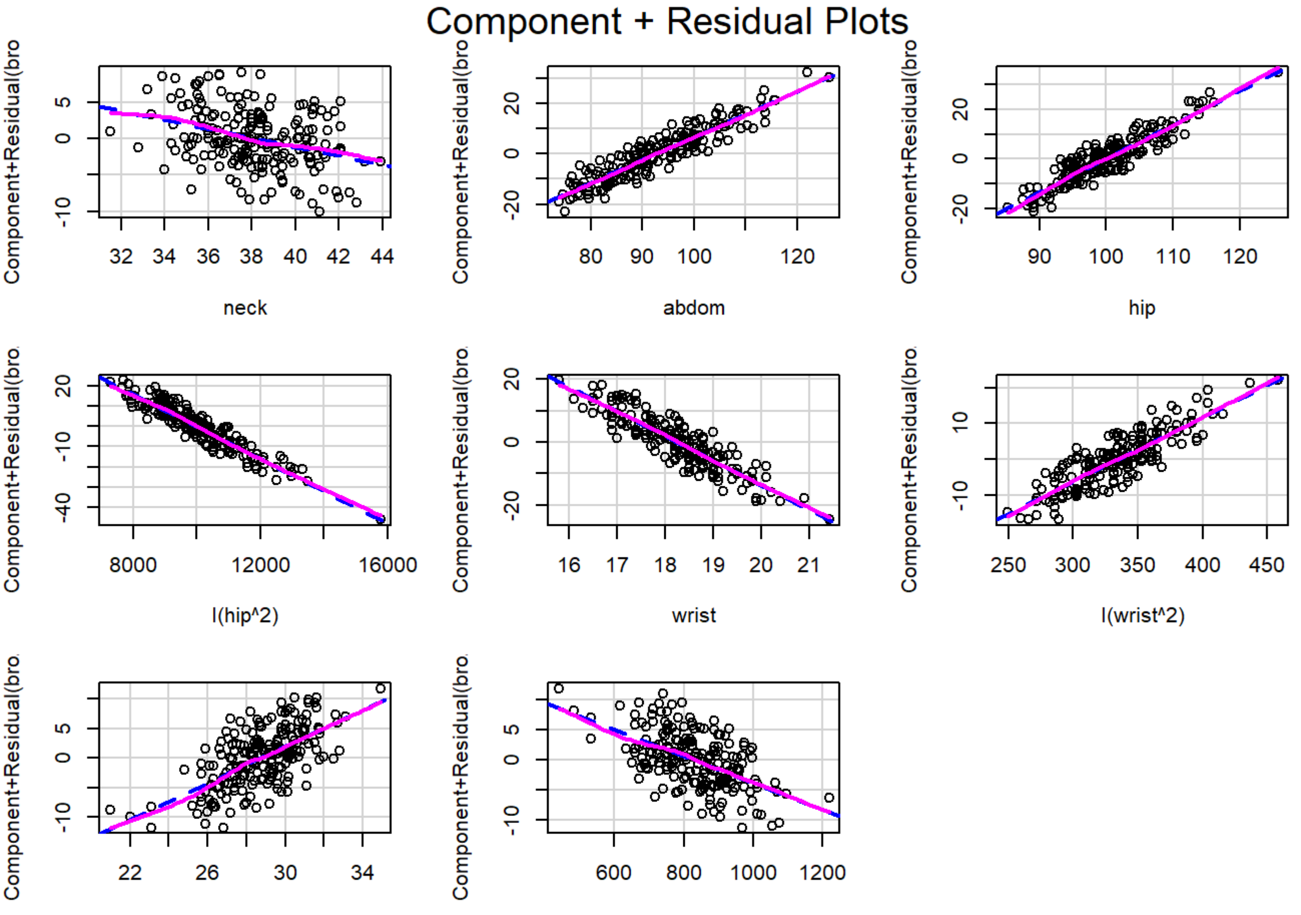


```
##
## Call:
## lm(formula = brozek ~ log(neck) + log(abdom) + log(hip) + forearm +
##     wrist, data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1077 -2.8505 -0.1034  2.6718  9.4370
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -130.8915    27.2368  -4.806 3.07e-06 ***
## log(neck)    -23.3747     8.8498  -2.641 0.008929 **
## log(abdom)    85.1838     5.3056  16.055  < 2e-16 ***
## log(hip)     -29.1715     8.5425  -3.415 0.000776 ***
## forearm        0.2894     0.1789   1.618 0.107321
## wrist         -1.3382     0.4638  -2.885 0.004350 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.964 on 195 degrees of freedom
## Multiple R-squared:  0.742,  Adjusted R-squared:  0.7353
## F-statistic: 112.1 on 5 and 195 DF,  p-value: < 2.2e-16
```

The result get better, but as the relationship looks like it could be non-monotone, we can now try to transform some variables up the ladder of power by including higher order terms to see if we can do better than log transformation. We now build three model which include three, two and one variables with higher order term. We only consider include second order term here.

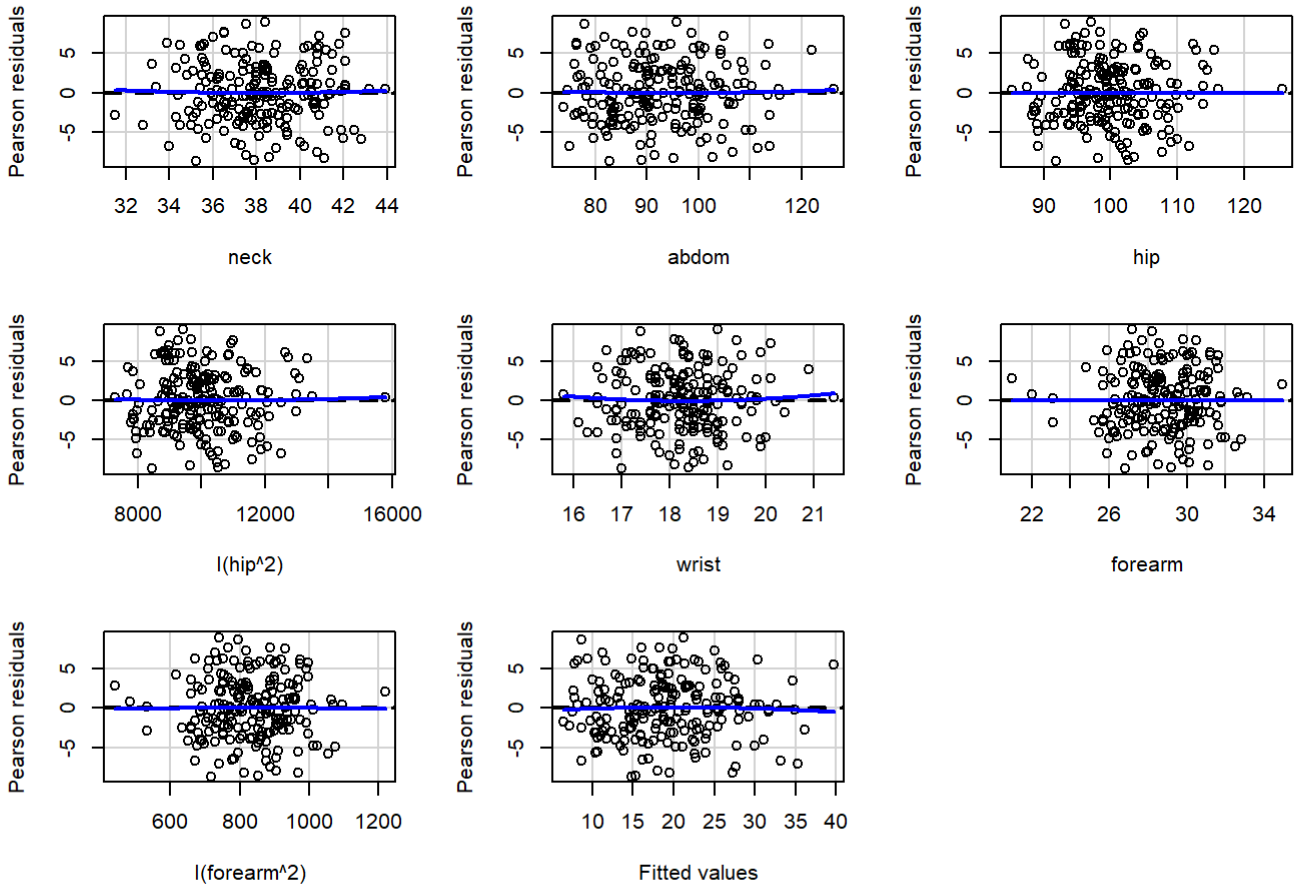
In fit5, we add higher order term to hip, wrist and forearm, as these variables have Residual and Crplot that do not meet assumption.

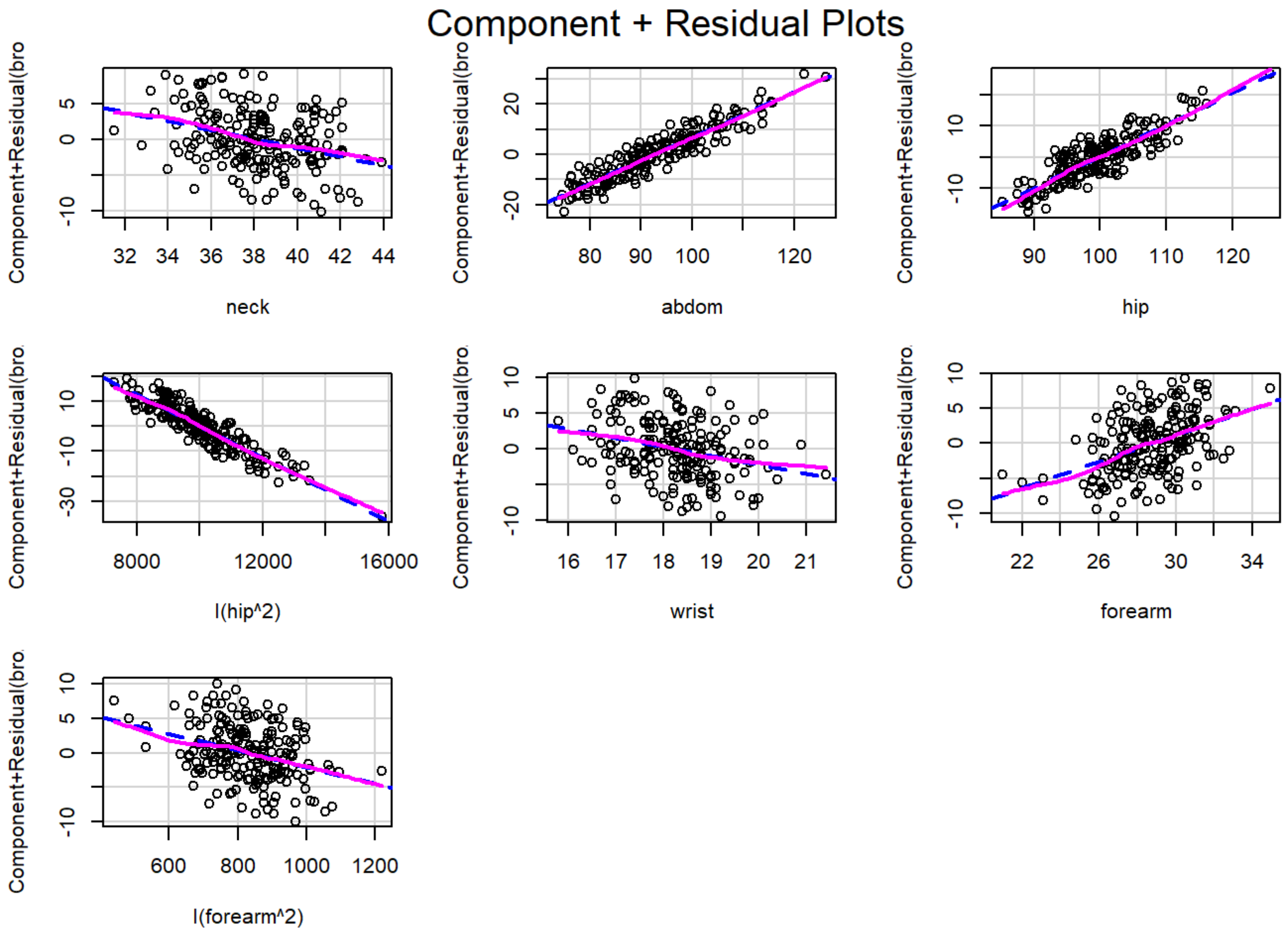

Pearson residuals
neck
abdom
hip
I(hip^2)
wrist
I(wrist^2)
forearm
I(forearm^2)
Fitted values


Component + Residual Plots
Component+Residual(bro
neck
abdom
hip
I(hip^2)
wrist
I(wrist^2)
forearm
I(forearm^2)

```
##
## Call:
## lm(formula = brozek ~ neck + abdom + hip + I(hip^2) + wrist +
##     I(wrist^2) + forearm + I(forearm^2), data = Train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.766 -2.794 -0.106  2.786  9.023
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -38.905787  73.522167  -0.529    0.597
## neck          -0.610990   0.240907  -2.536    0.012 *
## abdom          0.905885   0.057250  15.823   <2e-16 ***
## hip            1.353654   1.109145   1.220    0.224
## I(hip^2)      -0.008005   0.005409  -1.480    0.141
## wrist         -7.852872  10.524516  -0.746    0.456
## I(wrist^2)     0.179035   0.285823   0.626    0.532
## forearm        1.541810   2.662966   0.579    0.563
## I(forearm^2)  -0.022400   0.047250  -0.474    0.636
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.913 on 190 degrees of freedom
## Multiple R-squared:  0.7494, Adjusted R-squared:  0.7389
## F-statistic: 71.04 on 8 and 190 DF,  p-value: < 2.2e-16
```

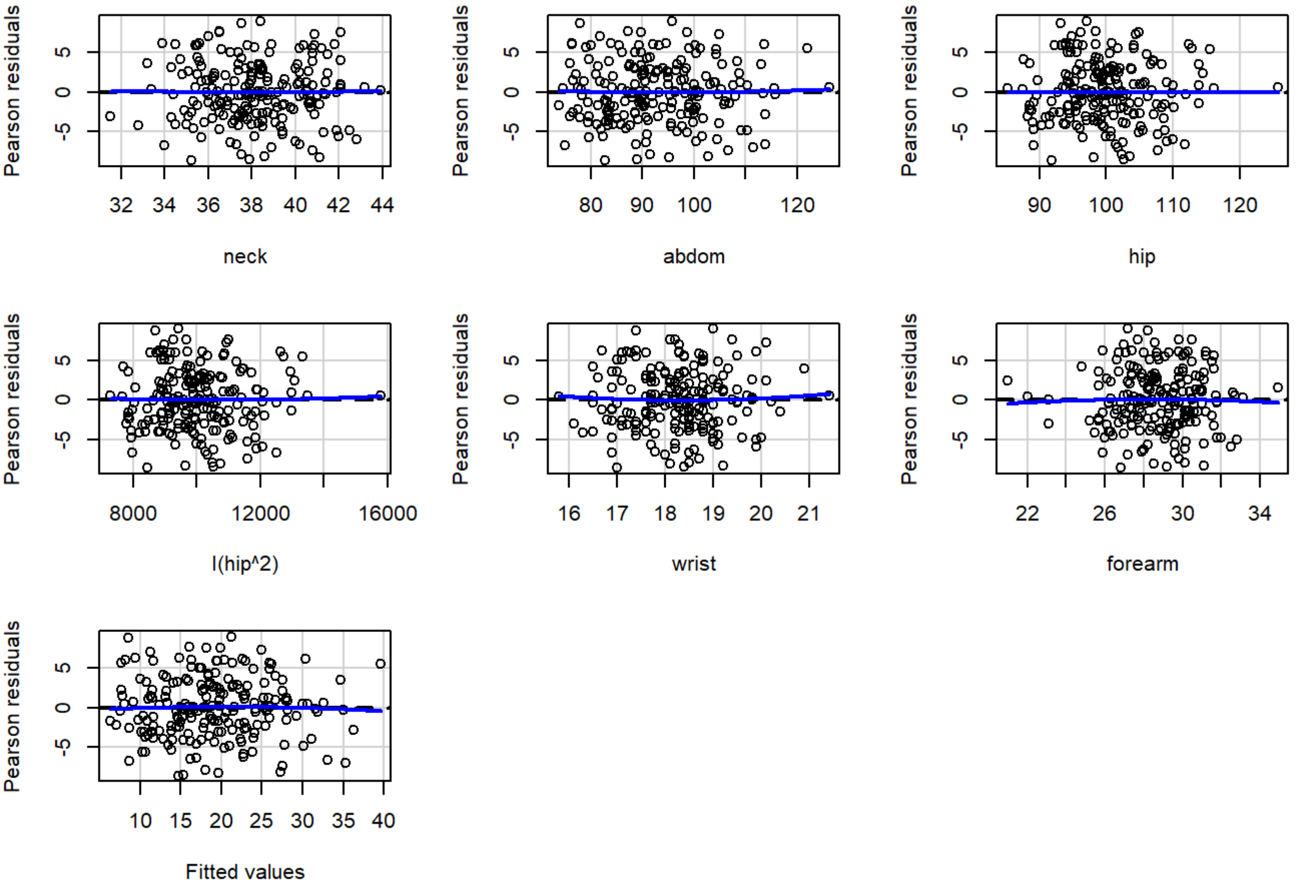
In fit6, we delete the wrist^2 term as wrist is the variable that shows lightest degree of violation to assumption.



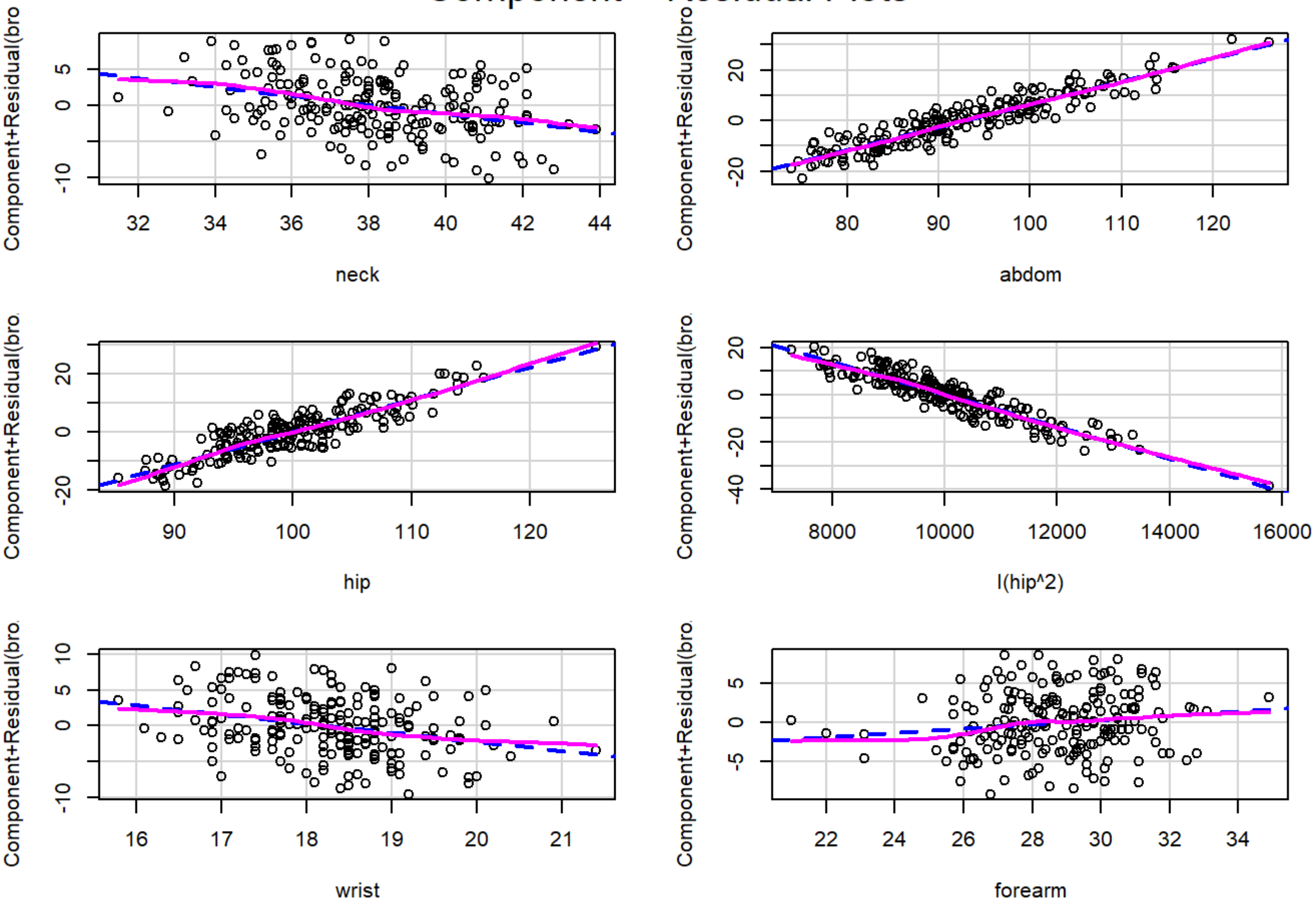


```
##
## Call:
## lm(formula = brozek ~ neck + abdom + hip + I(hip^2) + wrist +
##     forearm + I(forearm^2), data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6760 -2.6906 -0.1785  2.7759  8.9712
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -73.538860  48.383809  -1.520  0.13019
## neck          -0.618077   0.240258  -2.573  0.01085 *
## abdom          0.908862   0.056961  15.956  < 2e-16 ***
## hip            1.015513   0.967368   1.050  0.29515
## I(hip^2)      -0.006360   0.004721  -1.347  0.17953
## wrist         -1.266818   0.460254  -2.752  0.00649 **
## forearm        0.947130   2.484020   0.381  0.70341
## I(forearm^2)  -0.012033   0.044186  -0.272  0.78566
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.907 on 191 degrees of freedom
## Multiple R-squared:  0.7489, Adjusted R-squared:  0.7397
## F-statistic: 81.39 on 7 and 191 DF,  p-value: < 2.2e-16
```

We do the same in fit7, delete the forearm^2 term.



Component + Residual Plots



```
##
## Call:
## lm(formula = brozek ~ neck + abdom + hip + I(hip^2) + wrist +
##     forearm, data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6570 -2.7037 -0.1445  2.7553  9.0058
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -68.480259  44.568380  -1.537  0.12606
## neck         -0.628506   0.236614  -2.656  0.00857 **
## abdom         0.909777   0.056725  16.038  < 2e-16 ***
## hip           1.111096   0.899277   1.236  0.21814
## I(hip^2)     -0.006838   0.004372  -1.564  0.11940
## wrist        -1.272443   0.458680  -2.774  0.00608 **
## forearm       0.272593   0.187911   1.451  0.14851
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.898 on 192 degrees of freedom
## Multiple R-squared:  0.7488, Adjusted R-squared:  0.741
## F-statistic:  95.4 on 6 and 192 DF,  p-value: < 2.2e-16
```

It turns out that the best performing one is fit7, which has lowest Residual standard error and highest Adjusted R-squared. It include the second order term for hip, which shows the worst residual and component+residual Plot. fit7: brozek~neck+abdom+hip+I(hip^2)+wrist+forearm

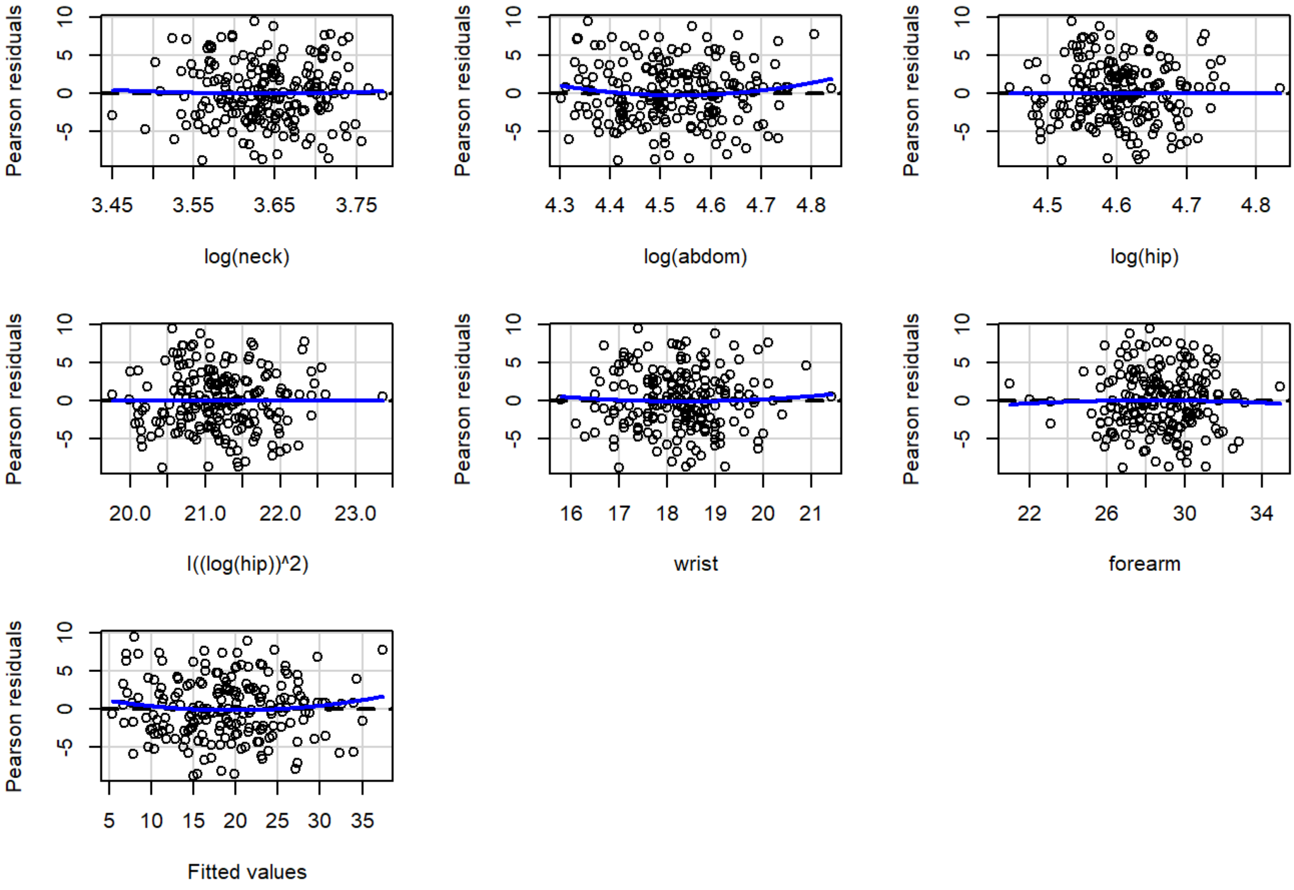
We try ANOVA method as fit6 and fit7 are nested in fi5 which respectively delete one two order term which has the less severe non-linearity.

```
## Analysis of Variance Table
##
## Model 1: brozek ~ neck + abdom + hip + I(hip^2) + wrist + I(wrist^2) +
##     forearm + I(forearm^2)
## Model 2: brozek ~ neck + abdom + hip + I(hip^2) + wrist + forearm + I(forearm^2)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    190 2909.7
## 2    191 2915.7 -1   -6.0087 0.3924 0.5318
```

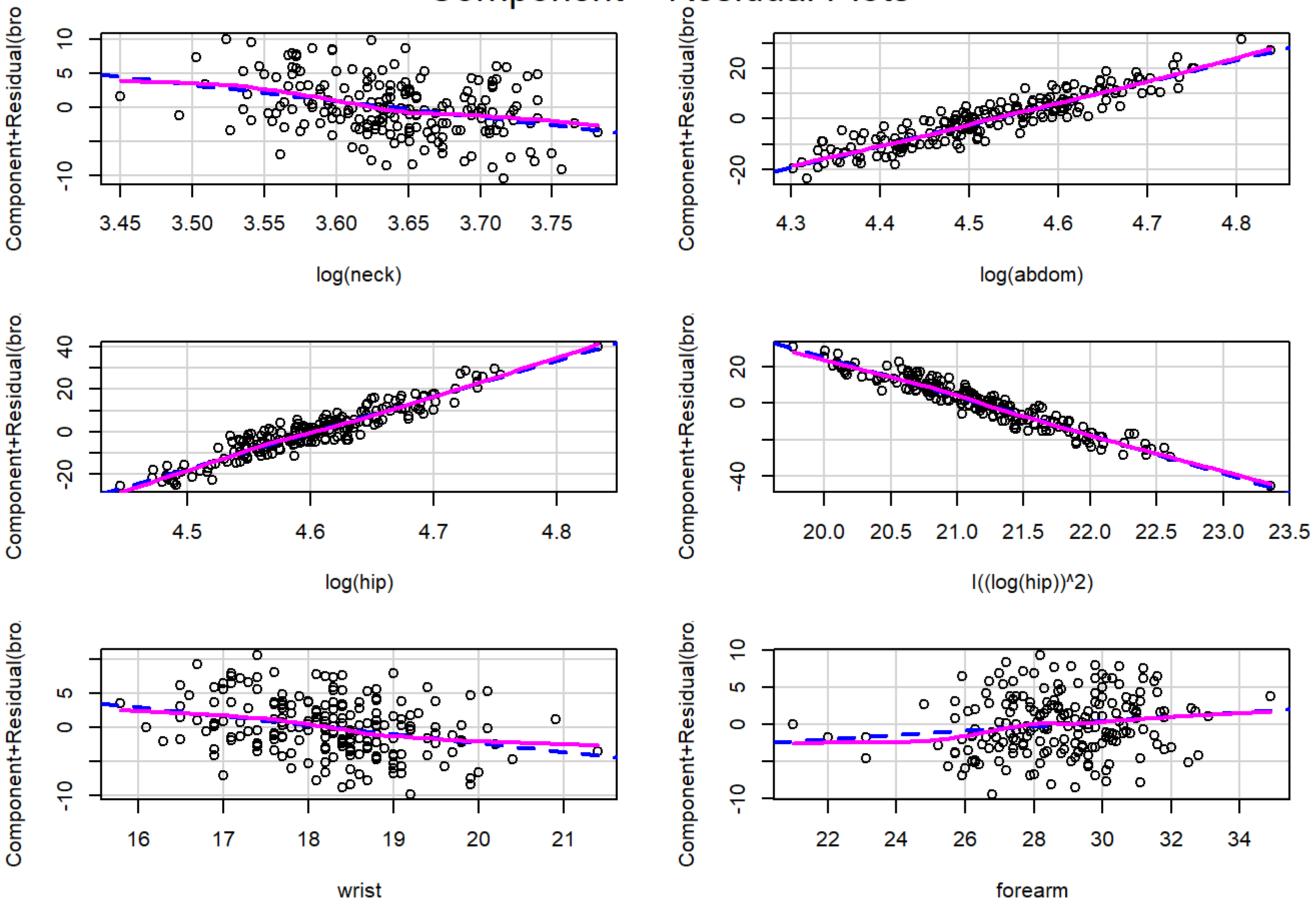
```
## Analysis of Variance Table
##
## Model 1: brozek ~ neck + abdom + hip + I(hip^2) + wrist + forearm + I(forearm^2)
## Model 2: brozek ~ neck + abdom + hip + I(hip^2) + wrist + forearm
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    191 2915.7
## 2    192 2916.9 -1   -1.1322 0.0742 0.7857
```

The method also shows what we aim for higher order term really is to meet the assumption. As the P-value is higher than 75%, including higher order actually not benefit the fitting performance.

However, what if we combine the ‘log’ transformation and higher order terms? will that even better?



Component + Residual Plots



```
##
## Call:
## lm(formula = brozek ~ log(neck) + log(abdom) + log(hip) + I((log(hip))^2) +
##     wrist + forearm, data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8425 -2.7257 -0.1025  2.5553  9.4692
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -590.7856   990.9394  -0.596  0.55175
## log(neck)        -23.9432     9.0400  -2.649  0.00876 **
## log(abdom)        84.6929     5.3467  15.840  < 2e-16 ***
## log(hip)         169.3240   429.8341   0.394  0.69407
## I((log(hip))^2)  -21.2311    46.5228  -0.456  0.64865
## wrist             -1.3264     0.4612  -2.876  0.00448 **
## forearm            0.2959     0.1898   1.560  0.12050
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.926 on 192 degrees of freedom
## Multiple R-squared:  0.7452, Adjusted R-squared:  0.7372
## F-statistic: 93.57 on 6 and 192 DF,  p-value: < 2.2e-16
```

Although the log transformation perform well above, we find it does not make progress to the normal model adding some higher order terms.

In conclusion, our best choice is fit7. fit7: brozek~neck+abdom+hip+I(hip^2)+wrist+forearm The model with coefficient is:

brozek = -68.48 - 0.63*neck + 0.91*abdom + 1.11*hip - 0.07*(hip^2) - 1.27*wrist + 0.27*forearm

Evaluate the quality of prediactors

Finally, we will use the test data to evaluate how is over best model perform and make comparison with the full model. The MSE of fit0, fit3, fit31, fit5 and fit7.

```
## Mean Squared Error of full model: 18.42898
```

```
## Mean Squared Error of model with 5 variable: 17.62825
```

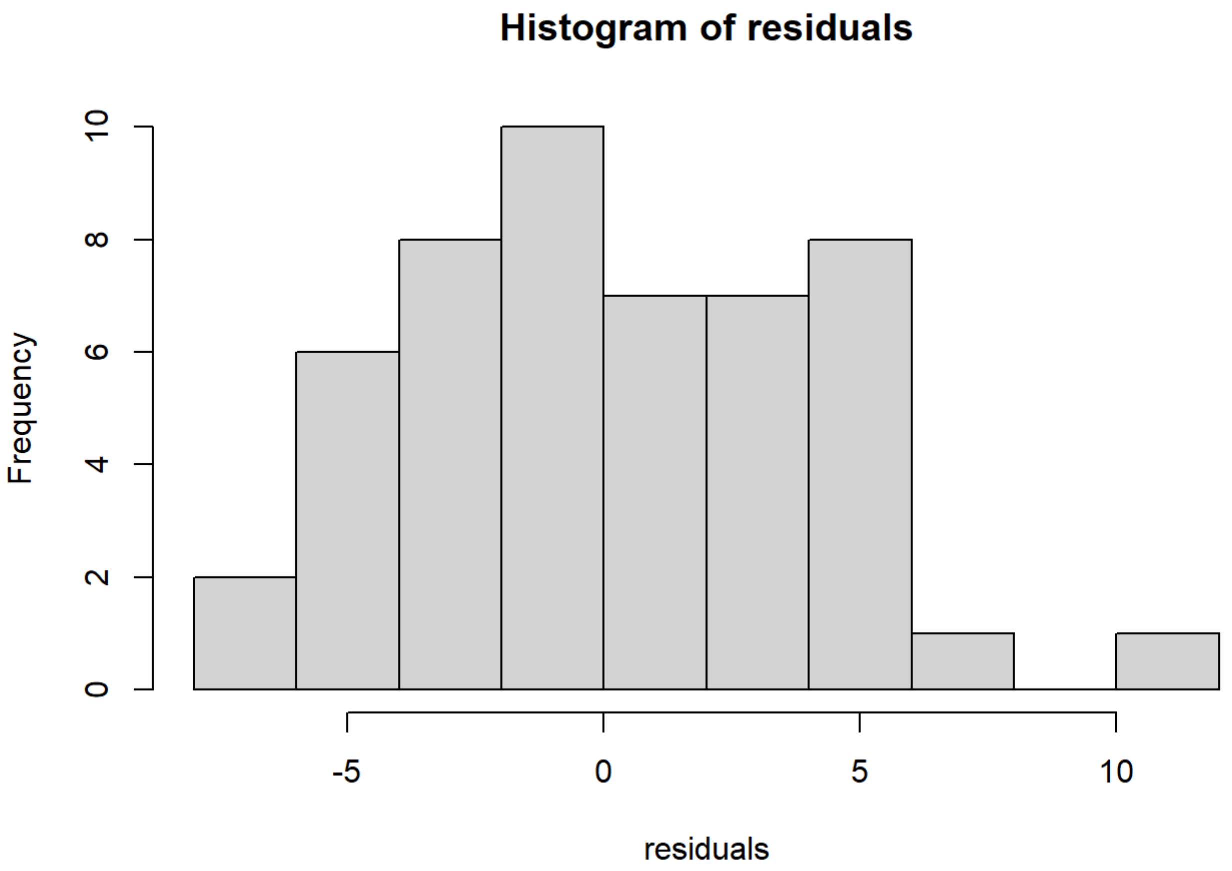
```
## Mean Squared Error of model with 5 variable with logtransformation: 17.56492
```

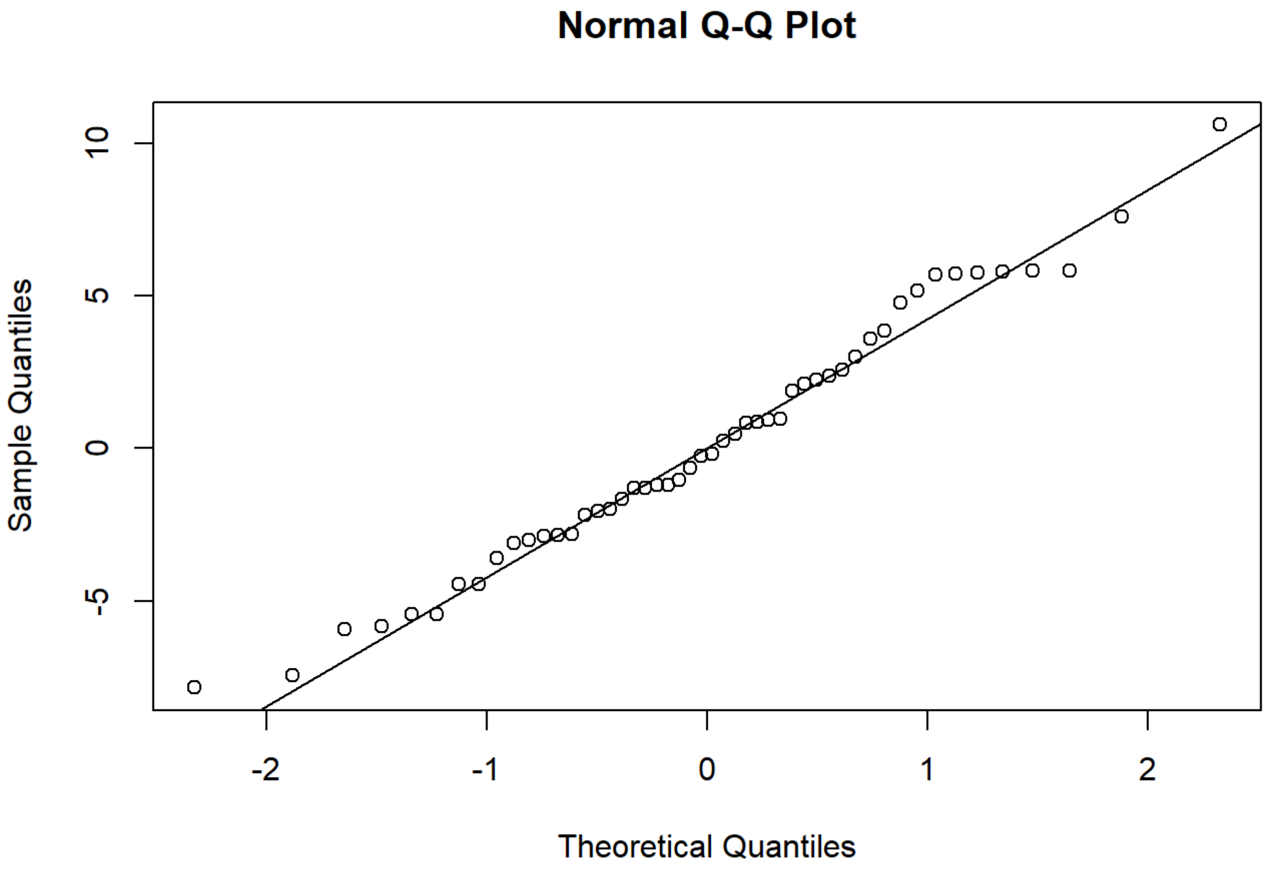
```
## Mean Squared Error of model with 5 variable plus three higher order term: 16.72531
```

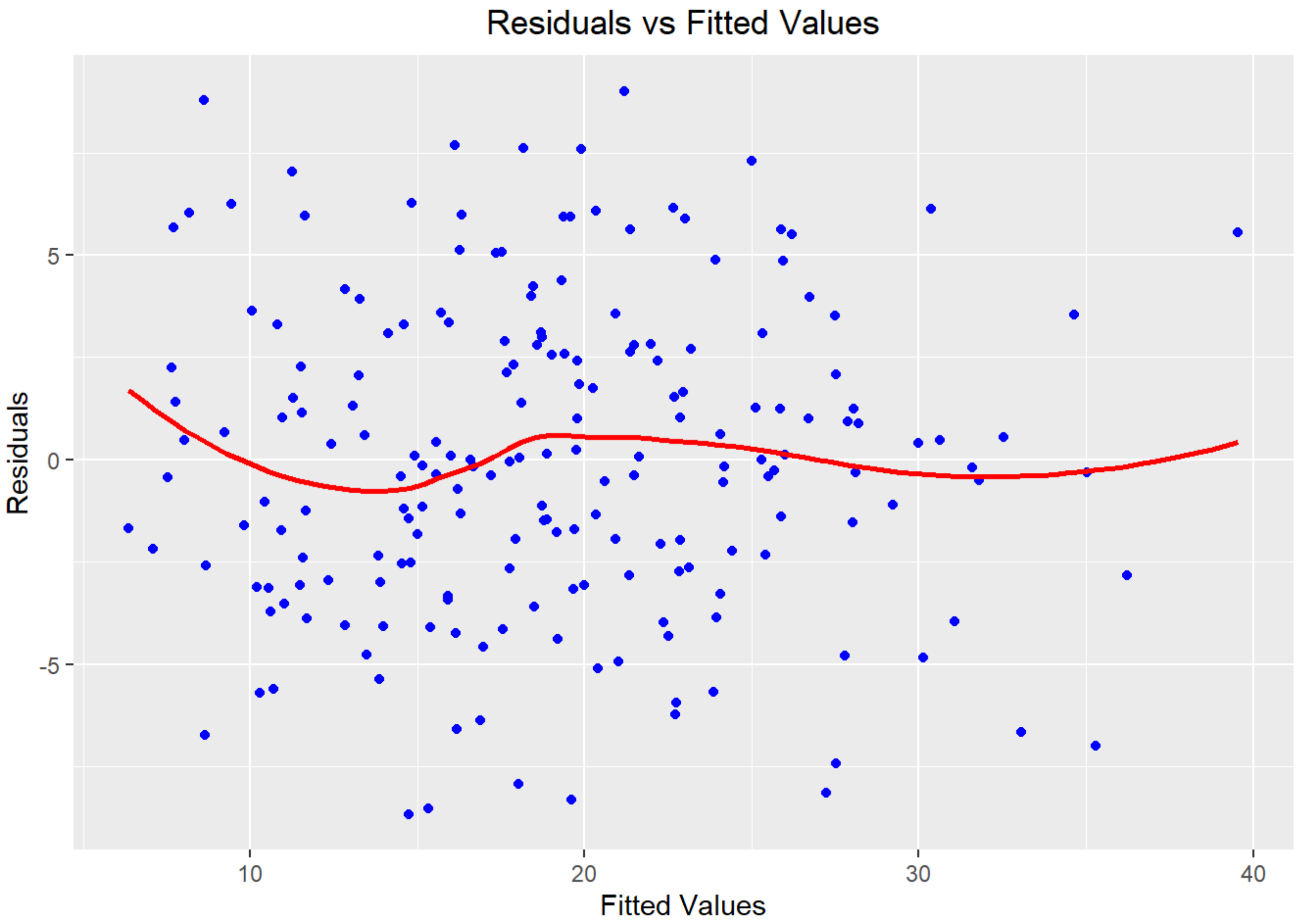
```
## Mean Squared Error with 5 variable plus one higher order term: 16.9855
```

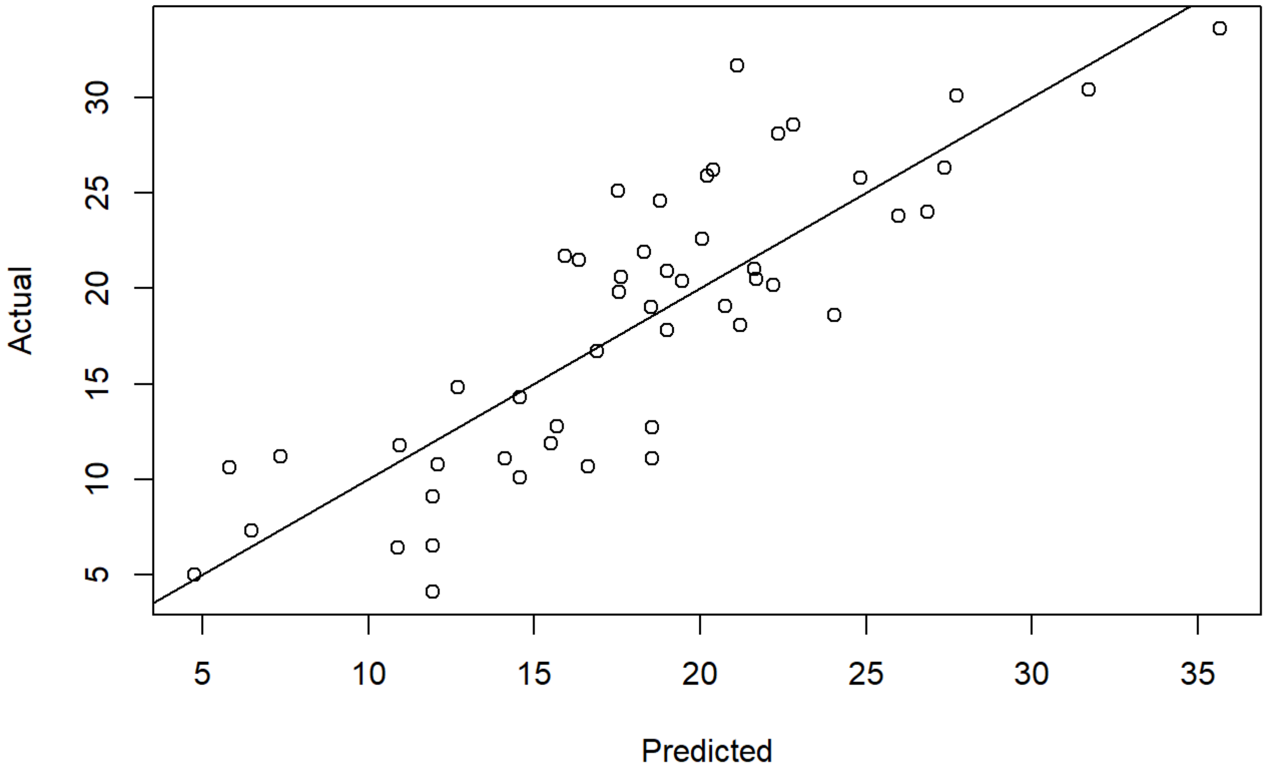
We can see that compare to the full model, the MSE is better if we use model3, the 5 variables model. It will get better if using ‘log’ transformation. Adding second order term to three variables could give the lowest MSE. However, after doing the ANOVA F-test and considering Adjusted R-squared, the the best model we give is the one with 5 variables and take a second order on hip.

In general, the MSE of the best model using test data is 16.9855, compare to the full model which is 18.42898. The Adjusted R-squared is 0.741, compare to the full model which is 0.7274









Although QQ-plot do not give much evidence to the violation of normality of the residuals, after plotting the histogram of residuals again we find the histogram has more deviation from the normal distribution. But we decide not to manipulate this considering the 50 pieces of data in Test is not large enough to provide reliable evidence.

Furthermore, the goal of linear regression is to try to balance between the assumptions of linearity and independence of errors while also trying to achieve normally distributed residuals. In practice, it is not always possible to meet all of the assumptions perfectly.

The Plot the residuals vs fitted values shows that most big deviation of fitted value from true value appears under value of 18. As the mean of brozek is greater than 18 in both Train and Test data, we can say that the error of model is generally in a smaller scale for a higher figure of 50%

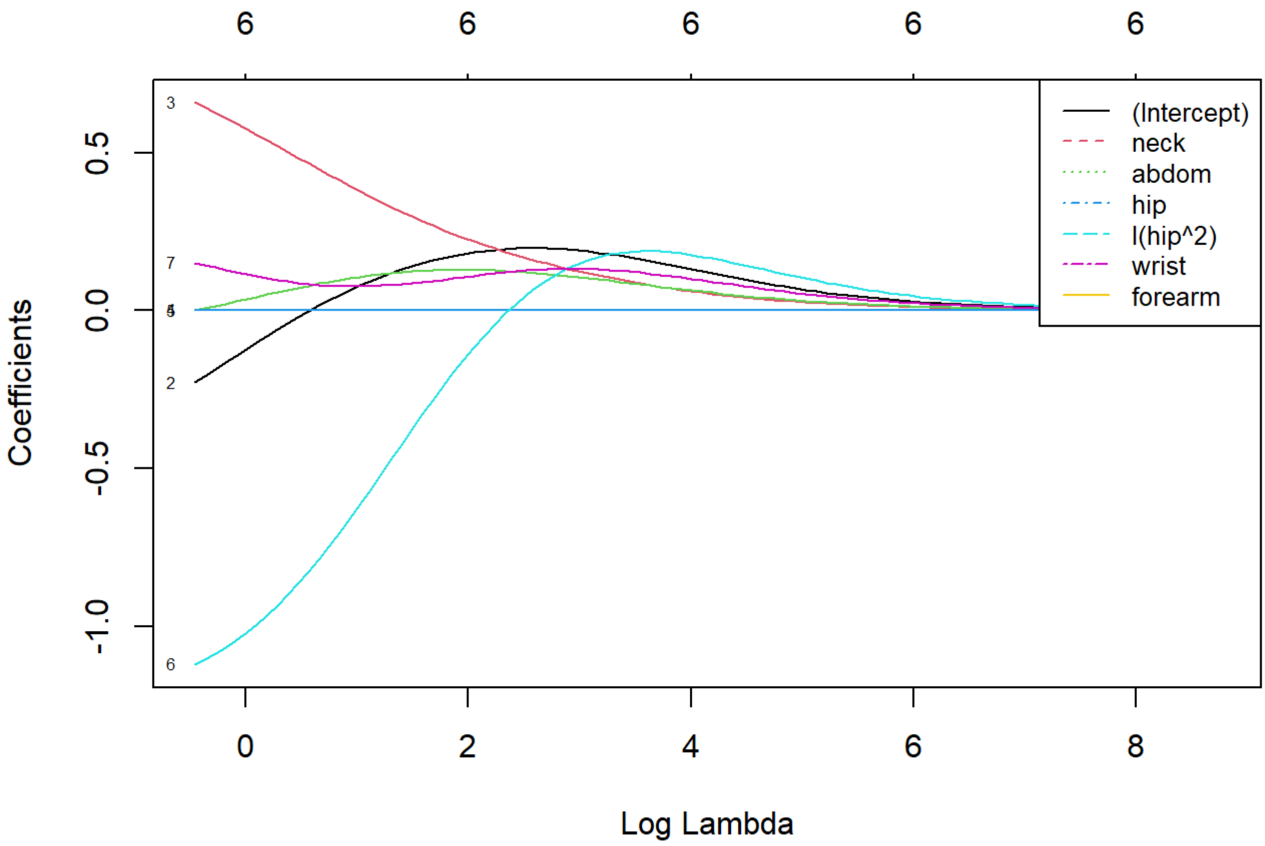
```
## Mean absolute error: 3.377553
```

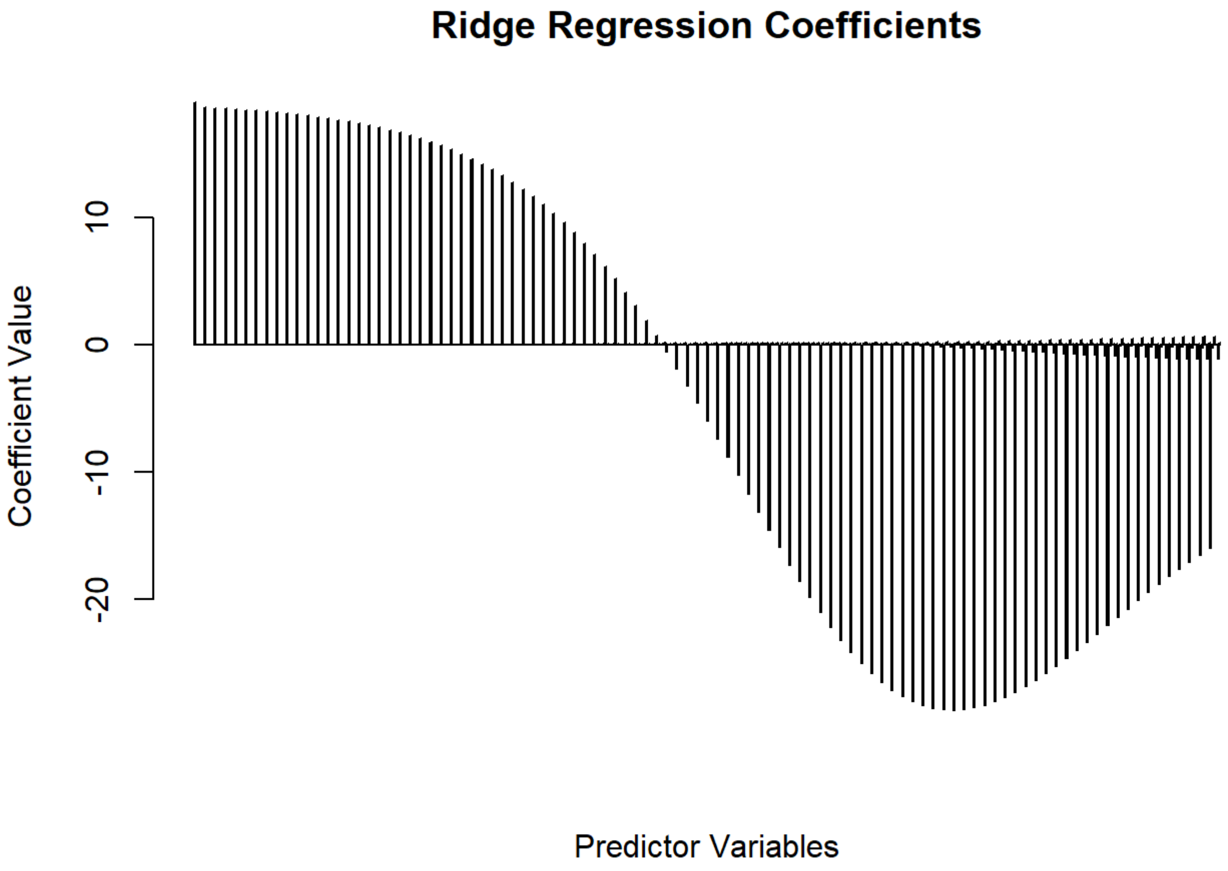
```
## Root mean squared error: 4.121347
```

```
## Mean Squared Error: 16.9855
```

Further exploration: Ridge Regreesion

Although there may not have particular progress after applying Ridge Regression, we explore it below.





```
##
## Call:
## lm(formula = ridge_formula, data = Train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6570 -2.7037 -0.1445  2.7553  9.0058
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -68.480259  44.568380  -1.537  0.12606
## neck         -0.628506   0.236614  -2.656  0.00857 **
## abdom         0.909777   0.056725  16.038  < 2e-16 ***
## hip           1.111096   0.899277   1.236  0.21814
## I(hip^2)     -0.006838   0.004372  -1.564  0.11940
## wrist        -1.272443   0.458680  -2.774  0.00608 **
## forearm       0.272593   0.187911   1.451  0.14851
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.898 on 192 degrees of freedom
## Multiple R-squared:  0.7488, Adjusted R-squared:  0.741
## F-statistic:  95.4 on 6 and 192 DF,  p-value: < 2.2e-16
```

```
## Mean Squared Error: 17.32063
```

We find there is no particular change to the model. This is possible regression as ridge regression is used to prevent overfitting and reduce the impact of multicollinearity, but if the variables are not highly correlated and the model is not overfitting, the impact of the regularization may be small.

MSE is getting worse after applying ridge regression, it might be an indication that the regularization parameter lambda is too high. In other words, the model is penalizing the coefficients too heavily, resulting in a less accurate fit to the data.

The result of ridge regression meet our intuition because our model include 6 variables, which is not complex compare to the Train data. Noisy or irrelevant features in the dataset have also been deleted when we perform model selection.

Find codes below:

```
knitr::opts_chunk$set(echo = TRUE)
library('ggplot2')
library(dplyr)
library(car)

TrainORI<-read.table('TrainORI.txt',header = TRUE)
Test<-read.table('Test.txt',header = TRUE)
TrainORI[1:5,]
Train <- TrainORI
Train<-filter(Train,brozek>0)
Train = filter(Train,!is.na(brozek),!is.na(neck),!is.na(chest),!is.na(abdom),!is.na(hip),!is.na
(thigh),!is.na(knee),!is.na(ankle),!is.na(biceps),!is.na(forearm),!is.na(wrist))
skimr::skim(Train)
hist(Train$brozek,main='histogram of brozek')
par(mfrow=c(3,5))
boxplot(Train$brozek ,xlab='brozek')
boxplot(Train$neck ,xlab='neck')
boxplot(Train$chest ,xlab='chest')
boxplot(Train$abdom ,xlab='abdom')
boxplot(Train$hip,xlab='hip')
boxplot(Train$thigh,xlab='thigh')
boxplot(Train$knee ,xlab='knee')
boxplot(Train$ankle ,xlab='ankle')
boxplot(Train$biceps ,xlab='biceps')
boxplot(Train$forearm ,xlab='forearm')
boxplot(Train$wrist ,xlab='wrist')
stripchart(data.frame(scale(Train)),method='jitter',las=2,vertical=TRUE)
plot(Train)
qplot(chest,brozek,data=Train,geom = c('point'))+stat_smooth(method='lm')
qplot(biceps,brozek,data=Train,geom = c('point'))+stat_smooth(method='lm')
qplot(ankle,brozek,data=Train,geom = c('point'))+stat_smooth(method='lm')
qplot(hip,brozek,data=Train,geom = c('point'))+stat_smooth(method='lm')
summary(lm(brozek~hip,Train))
fit0<- lm(brozek ~ neck + chest + abdom + hip + thigh + knee + ankle+ biceps + forearm + wrist,
data=Train)
summary(fit0)

plot(fit0,1:2)

fit01<- lm(brozek ~ log(neck) + chest + log(abdom) + log(hip) + log(thigh) + knee + log(ankle)+
biceps + forearm + wrist,data=Train)

summary(fit01)
plot(fit01,1:2)
library(leaps)
a <- regsubsets(brozek ~ ., data = Train)
summary.out <- summary(a)
summary.out
library(MASS)
full.model <- lm(brozek ~ ., data = Train)
step.model <- stepAIC(full.model, direction = "both", trace = FALSE)
summary(step.model)
plot(a,scale='bic')
plot(1:8,summary.out$adjr2,xlab = 'No. of parameters',ylab = 'Adjusted R-square')
which.max(summary.out$adjr2)
```

```
plot(1:8,summary.out$cp,xlab = 'No. of parameters',ylab = 'Cp Statistic')
abline(0,1)
fit0_step <- step(fit0, direction = "both")
summary(fit0_step)
full.model <- lm(brozek~log(neck) + chest + log(abdom) + log(hip) + log(thigh) + knee + log(ank
le)+ biceps + forearm + wrist,data=Train)
step.model <- stepAIC(full.model, direction = "both", trace = FALSE)
summary(step.model)
plot(1:8,summary.out$adjr2,xlab = 'No. of parameters',ylab = 'Adjusted R-square')
plot(1:8,summary.out$cp,xlab = 'No. of parameters',ylab = 'Cp Statistic')
abline(0,1)
fit01_step <- step(fit01, direction = "both")
summary(fit01_step)
fit0<- lm(brozek ~  neck + chest + abdom + hip + thigh + knee + ankle+ biceps + forearm + wris
t,data=Train)
fit1<- lm(brozek~neck+abdom+hip,data=Train)
fit2<- lm(brozek~neck + abdom + hip +wrist, data=Train)
fit3<- lm(brozek~neck + abdom + hip + forearm + wrist,data=Train)
summary(fit0)
summary(fit1)
summary(fit2)
summary(fit3)
anova(fit1,fit2)
anova(fit2,fit3)
fit01<- lm(brozek ~ log(neck) + chest + log(abdom) + log(hip) + log(thigh) + knee + log(ankle)+
biceps + forearm + wrist,data=Train)
fit11<- lm(brozek~log(neck) + log(abdom) + hip,data=Train)
fit21<- lm(brozek~ log(neck) + log(abdom) + hip + wrist, data=Train)
fit31<- lm(brozek ~ log(neck) + log(abdom) + log(hip) + forearm + wrist,data = Train)
summary(fit01)
summary(fit11)
summary(fit21)
summary(fit31)
qqPlot(fit3)
hist(rstudent(fit3))
fit4<- lm(log(brozek)~neck + abdom + hip + forearm + wrist,data=Train)
qqPlot(fit4)
hist(rstudent(fit4))
qqPlot(fit31)
hist(rstudent(fit31))
influencePlot(fit3)
influenceIndexPlot(fit3)
#we first delete the line 33 and check the result
Train1 <- Train[-c(33),]
fit3<- lm(brozek~neck+abdom+hip+wrist+forearm,data=Train1)
summary(fit3)
Train2 <- Train[-c(33,69),]
fit3<- lm(brozek~neck+abdom+hip+wrist+forearm,data=Train2)
summary(fit3)
Train3 <- Train[-c(33,139),]
fit3<- lm(brozek~neck+abdom+hip+wrist+forearm,data=Train3)
summary(fit3)
Train4 <- Train[-c(33,175),]
fit3<- lm(brozek~neck+abdom+hip+wrist+forearm,data=Train4)
summary(fit3)
Train<-Train2
```

```
TestResponse  = Test$brozek
TrainResponse  = Train$brozek
residualPlots(fit3,tests=FALSE)
crPlots(fit3)
residualPlots(fit31,tests=FALSE)
crPlots(fit31)
summary(fit31)
fit5<- lm(brozek~neck+abdom+hip+I(hip^2)+wrist+I(wrist^2)+forearm+I(forearm^2),data=Train)
residualPlots(fit5,tests=FALSE)
crPlots(fit5)
summary(fit5)
fit6<- lm(brozek~neck+abdom+hip+I(hip^2)+wrist+forearm+I(forearm^2),data=Train)
residualPlots(fit6,tests=FALSE)
crPlots(fit6)
summary(fit6)
fit7<- lm(brozek~neck+abdom+hip+I(hip^2)+wrist+forearm,data=Train)
residualPlots(fit7,tests=FALSE)
crPlots(fit7)
summary(fit7)
anova(fit5,fit6)
anova(fit6,fit7)
fit8<- lm(brozek~log(neck)+log(abdom)+log(hip)+I((log(hip))^2)+wrist+forearm,data=Train)
residualPlots(fit8,tests=FALSE)
crPlots(fit8)
summary(fit8)
predictions0<-predict(fit0,newdata = Test)
mse_bestsubset0<-mean((predictions0-TestResponse)^2)
cat("Mean Squared Error of full model:", mse_bestsubset0, "\n")

predictions3<-predict(fit3,newdata = Test)
mse_bestsubset3<-mean((predictions3-TestResponse)^2)
cat("Mean Squared Error of model with 5 variable:", mse_bestsubset3, "\n")

predictions31<-predict(fit31,newdata = Test)
mse_bestsubset31<-mean((predictions31-TestResponse)^2)
cat("Mean Squared Error of model with 5 variable with logtransformation:",mse_bestsubset31,
"\n")

predictions5<-predict(fit5,newdata = Test)
mse_bestsubset5<-mean((predictions5-TestResponse)^2)
cat("Mean Squared Error of model with 5 variable plus three higher order term:", mse_bestsubset
5, "\n")

predictions7<-predict(fit7,newdata = Test)
mse_bestsubset7<-mean((predictions7-TestResponse)^2)
cat("Mean Squared Error with 5 variable plus one higher order term:", mse_bestsubset7, "\n")
# Use the model to make predictions on the test data
predictions7 <- predict(fit7, newdata = Test)

# Calculate the residuals and the evaluation metrics
residuals <- Test$brozek - predictions7
hist(residuals)
# Check for normality of the residuals
qqnorm(residuals)
qqline(residuals)
```

```
# Plot the residuals vs fitted values
ggplot(data = data.frame(residuals = residuals(fit7), predictions = predict(fit7)), aes(x = pre
dictions, y = residuals)) +
  geom_point(color = "blue") +
  geom_smooth(method = "loess", se = FALSE, color = "red") +
  labs(title = "Residuals vs Fitted Values",
       x = "Fitted Values",
       y = "Residuals") +
  theme(plot.title = element_text(hjust = 0.5))

plot(predictions7, Test$brozek, xlab = "Predicted", ylab = "Actual")
abline(0, 1)
mae <- mean(abs(residuals))
rmse <- sqrt(mean(residuals^2))
mse <- mean(residuals^2)

cat("Mean absolute error:", mae, "\n")
cat("Root mean squared error:", rmse, "\n")
cat("Mean Squared Error:", mse, "\n")
library(glmnet)
library(ggplot2)

# Define predictor and response variables for ridge regression
x <- model.matrix(brozek ~ neck + abdom + hip + I(hip^2) + wrist + forearm, data = Train)
y <- Train$brozek

# Fit ridge regression model using glmnet
fit_ridge <- glmnet(x, y, alpha = 0)

# Plot the coefficient paths
plot(fit_ridge, xvar = "lambda", label = TRUE)

# Add a legend to the plot
legend("topright", legend = colnames(x), col = 1:ncol(x), lty = 1:ncol(x), cex = 0.8)

# Get coefficients from ridge model as numeric vector
coef_ridge <- as.numeric(coef(fit_ridge))

# Plot coefficients
barplot(coef_ridge, main = "Ridge Regression Coefficients", xlab = "Predictor Variables", ylab
= "Coefficient Value")
library(glmnet)

# Extract the predictor variables and the response variable from the Train dataset
x_train <- as.matrix(Train[, c("neck", "abdom", "hip", "wrist", "forearm")])
y_train <- Train$brozek

# Add a new variable hip^2 to the predictor matrix
x_train <- cbind(x_train, hip2 = Train$hip^2)

# Set up a sequence of lambda values
lambda_seq <- 10^seq(10, -2, length.out = 100)

# Fit the ridge regression model using the cv.glmnet() function
fit_ridge <- cv.glmnet(x_train, y_train, alpha = 0, lambda = lambda_seq, standardize = TRUE)
```

```
# Choose the lambda value that gives the lowest cross-validation error
best_lambda <- fit_ridge$lambda.min

# Refit the model using the chosen lambda value
fit_ridge_best <- glmnet(x_train, y_train, alpha = 0, lambda = best_lambda, standardize = TRUE)

# Extract the coefficients for the chosen lambda value
coef_ridge_best <- coef(fit_ridge_best)

# Convert the coefficients to a vector
coefficients <- as.vector(coef_ridge_best)[-1]

# Add the intercept to the coefficients
intercept <- coef_ridge_best[1]
coefficients <- c(intercept, coefficients)

# Create a formula for the ridge regression model
ridge_formula <- as.formula(paste(″brozek˜neck+abdom+hip+I(hipˆ2)+wrist+forearm ″, paste0(names
(coefficients[-1]), collapse = ″ + ″)))

# Fit the ridge regression model using the lm() function
ridge_model <- lm(ridge_formula, data = Train)

# Print the coefficients for the ridge regression model
summary(ridge_model)



# Fit ridge regression model on training data
x_train <- model.matrix(brozek ˜ neck + abdom + hip + I(hipˆ2) + wrist + forearm, data = Train)
y_train <- Train$brozek
fit_ridge <- glmnet(x_train, y_train, alpha = 0, lambda = 0.01)

# Make predictions on test set using model after ridge
x_test <- model.matrix(brozek ˜ neck + abdom + hip + I(hipˆ2) + wrist + forearm, data = Test)
pred_ridge <- predict(fit_ridge, newx = x_test)

# Compute MSE of predictions on test set
mse_ridge <- mean((pred_ridge - Test$brozek)ˆ2)
cat(″Mean Squared Error:″, mse_ridge, ″\n″)
```