National University of Singapore

ST3131 Regression Analysis

Group 31's report

Authors:	*Matric number:*
Tay Sze Min Rachael	A0156726X
Arijit Pramanik	A0179365N
Lai Hoang Dung	A0131125Y
Danilo Peng	A0178907M

1 Introduction

1.1 Problem description

According to the World Health Organization (WHO), obesity is a major risk factor to develop a number of chronic diseases such as diabetes and cancer. The definition of obesity is the following: the disease in which excess body fat has accumulated to such extent that health may be adversely affected. Even though this is considered as a problem in only high income countries, obesity is rising in low and middle income countries as well. Therefore, it is an important issue to be able to identify people with the fat excess in terms of medical purposes. The main purpose of this project is to develop and validate a regression model for prediction of the body fat mass (BFM) in persons body.

Body mass index (BMI) is a well known measurement which is widely used to indirectly calculate the fatness of a person. However, it is shown that the model is very poor in predicting the actual fatness. Instead, there are other highly accurate methods for measuring the BFM, such as X-ray densitometry (DXA) or hydrodensitometry. However, these methods often have little practical applicability due to their high costs and methodological efforts. Thus, one could instead develop a much cheaper and more portable method with the help of regression analysis. By common sense, it is known that different body measurements such as waist circumference, neck circumference, abdominal circumference and weight are related to the BFM. Hence, one can use these variables as predictors in multiple linear regression and develop a model that is able to describe the BFM without high methodological efforts or high financial costs.

1.2 Dataset

Human body consists of two components, lean body tissue and fat tissue. In the obtained CMU Men's Body fat Dataset [dat], body density of every individual was measured. Thereafter, the following equation system was solved in order to obtain the body fat,

$$D = \frac{1}{\frac{A}{a} + \frac{B}{b}} \tag{1}$$

$$B = \frac{1}{D}\frac{ab}{a-b} - \frac{b}{a-b} \tag{2}$$

where, D is the Body Density (g/cm^3), A is the proportion of lean body tissue and B is the proportion of fat tissue . There are two other parameters, a which is the density of lean body tissue (g/cm^3) and b is the density of fat tissue (g/cm^3). These are estimated to 1.1 and 0.9. After solving the equation system, the following equation is obtained,

$$BF = \frac{495}{D} - 450, \tag{3}$$

and is also known as Siri's equation.

However estimating body density isn't something that can be done by everyone and has a high cost. Therefore, for our project, we are only looking at the other variables that can be measured using low-cost methods.

The data set contains 13 predictor variables, as detailed below.

Predictor	**Short name**	**Unit**
Age	Age	years
Weight	Weight	lbs
Height	Height	inches
Abdomen 2 circumference	Abdomen	cm
Hip circumference	Hip	cm
Neck circumference	Neck	cm
Ankle circumference	Ankle	cm
Forearm circumference	Forearm	cm
Wrist circumference	Wrist	cm
Biceps (extended) circumference	Biceps	cm
Chest circumference	Chest	cm
Knee circumference	Knee	cm
Thigh circumference	Thigh	cm

2 Discussion of statistical analysis

2.1 Stepwise Regression

Firstly, a stepwise analysis was done in order to find the best model among the given predictors. This could easily be done in R by the built in function **step** based on the Akaike Information Criterion (AIC). The model obtained model (model A) using the stepwise method was reduced from 13 predictors to 8 including, Weight , Abdomen, Wrist, Forearm, Neck, Age, Thigh and Hip with AIC value, 738.98. The default step function used has significance levels of 15.73% for entering as well as staying in the model.

2.2 Significance and partial F-test

Even though the best model among the predictors were found, further analysis was required, such as significance of the model and partial F-tests of specific predictors. Therefore, a significance test was done on model A and a p-value of less than 2.2^{-16} was obtained (Figure 3). We therefore conclude that model A is significant. Furthermore, a partial F-test was done in order to check if specific predictors could be removed in order to obtain a simpler model. As a result, it was shown that Hip, Thigh, Age and Neck had p-values greater than 0.05, thus the null hypotheses was not rejected. Hence, we conclude that these predictors could be removed .

We thus obtained model B which only included Abdomen, Weight, Wrist and Forearm as the predictors. Model B has a R-square of 0.7308 and p-value of less than 2.2^{-16} (Figure 5).

2.3 Power and Interaction Terms

Considering second order model, we included second order terms for Abdomen, Weight, Wrist and Forearm and take a look at their significance in the model.

Observing the results in Figure 6 we could tell that only weight square has some significance amongst all the second order terms since it was the only predictor whose p value was below 5%. Hence, we take a look at a reduced model with Abdomen, Weight, Wrist, Forearm along with the Weight-squared.

Using ANOVA, we compare the reduced model with weight square to the reduced model without Weight-squared, we are able to observe that p-value is much smaller than 0.05 (Figure C). We reject the null hypothesis and conclude that the model with Weight-squared is significant and that we should include it. However, since the R square only improve slightly from the reduced model, we choose to not include any power of 2 terms.

Next, we test out interaction terms. As can be seen from Figure 8, the interaction terms have p-values ¿ 0.1 and the adjusted R square value did not increase significantly. Hence, we did not include any interaction terms in the model.

Consequently, we still stick to model B.

2.4 Data transformation

Firstly we investigated if an appropriate transformation of the data had to be made. We started by making scatter plots involving body fat as a function of each predictor to check for a convex/concave and/or decreasing/increasing behavior (Appendix C).

Since there is no significant observable pattern in the plots of the response variable against each of the predictor values, we conclude that there is no need of any transformation using higher/lower powers of the predictor and response variables. In fact, the scatter plots exhibit a near linear relationship between the response variable and each of the predictor variables, though this is more useful in the case of a single predictor variable

However, we notice that the response variable is always positive, and the minimum observed body fat is 0.7 units, and the maximum is 47.5 units. Hence, $\frac{y_{max}}{y_{min}}$ is observed to be greater than 10. So, we try to apply the Box Cox transformation and observed that the resultant value of Λ to be 1.01, which is very close to 1. There, we applied the box-cox transformation to our model, with value of lambda as 1, which only brings in a reduction of 1 unit from the response variable, body Fat. Thus, the fitted model is exactly the same as before, with the exception of the intercept term(β_0) increasing by 1. Thus our model B is still appropriate.

2.5 Lack of fit Test

We do not have any repeated measurements in the above dataset and hence we can't perform the lack of fit test.

2.6 Checking for Multicollinearity

We computed the Variance Inflation Factor (VIF) and the Tolerance (TOL) to check for multicollinearity among the 4 predictors in the model.

	Abdomen	**Weight**	**Wrist**	**Forearm**
VIF	4.789592	6.924539	2.242982	1.770719
TOL	0.20537409	0.1444139	0.4458351	0.5647424

None of the TOLs are small, therefore there is no evidence for multicollinearity among the predictors.

3 Appropriateness of Statistical Analysis

3.1 Check for Influential Observations

After we obtained model B post stepwise-regression, partial F-test and testing for higher-power and interaction terms, we check for potential influential observations and outliers.

We notice that the observations indexed 36, 39, 41, 86, 152, 159, 171, 175, 178, 204, 205, 220, 223, 224, 225 and 251 are considered influential (Appendix D).

Here, we have 4 predictors, hence an observation is deemed influential if its leverage h_{ii} for the i^{th} data point is more than $\frac{(\#ofpredictors+1)}{\#ofobservations}$, i.e. much larger than $\frac{5}{251} = 0.0199$, since the leverage describes how far the i^{th} data point is from the centre of all the data points. Similarly, the studentized residuals is a measure of the distance between the i^{th} data point and the model estimated on the rest of the data points. We tag points with $|e_i^*| > 2$ similarly for further investigation. Similarly, since we have 251 observations, we tag points having $|DFFITS| > 0.2851 = \frac{2(4+1)}{(251-4-1)}$ and $|DFBETAS| > 0.1262 = \frac{2}{251}$ for each of the predictor variables.

As seen from the residuals plot for the reduced model, we see that, observations 223, 224 have a residual value lesser than -10, and is clearly an outlier for the given data set. Also it is evident that observations 36, 39, 41 are influential observations and possess an undue influence on the fitted model. This is also supported by the fact that for observations 36 and 41, h_{ii} values are 0.07384 and 0.06033 respectively, much larger than the ideal value of 0.0199, and hence deemed as an influential observation. And for observation 39, we observe that $h_{ii} = 0.26902 >> 0.0199$, and $|DFBETA_{weight}| = 1.04 > 0.1262$ and $|DFFIT| = 1.58 > 0.2851$.

We use **influence.measures** in R to check for influential observations. These observations are tagged for further study. The potential outliers identified above are removed in order to obtain a well fitted regression model. Comparing to the reduced model before removing outliers, both the models have approximately the same lower-bound for p-value of 2.2e-16. The value of R^2 as well as adj R^2 increases from 0.7308 and 0.7264 to 0.7331 and 0.7285 respectively, The sum of residuals decreases from 4632.8 to 4001.9 for the new model after removing outliers. As seen below, the plot of the residuals seem a bit more random after removing the outliers. We also see that there are no residuals with a value lesser than -10 after removing outliers compared to the reduced model with outliers. So after removing outliers, all residual values are bounded between -10 and 10.

3.2 Test for Normality and Independence of Residuals

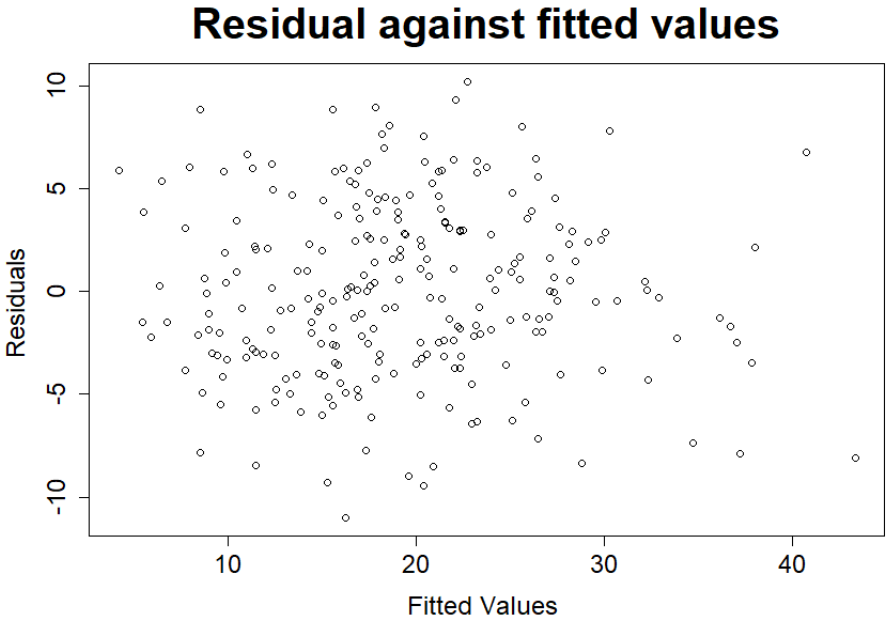


Figure 1: Scatter plot of residuals against fitted values for model B

We drew scatter plots for the residuals of the final reduced model against that of the fitted values. Since the points in the scatter plot show no specific pattern, and are random enough, we conclude that there is no need of any transformation or inclusion of other predictors into the model. Also in Appendix C, we have attached plots of the residuals against each of the predictors. Since, these plots too do not show any specific pattern and seem more or less constant throughout the range of the predictor values, the regression model is adequate. The variance seems constant throughout, so we don't need any variance stabilizing transformation.

To test the normality of the residuals, we plot the Normal Q-Q plot of the residuals against the normal scores. Since, it approximately follows a straight line given the noise in the data, we conclude that the residuals follow a normal distribution.

Also, the same is evident from the Kolmogorov-Smirnov test, since the D value is closer to 0, and hence we conclude that the samples(residuals) were drawn from the same distribution, Also the p-value of $0.8015 > 0.05$ implies we do not reject the null hypothesis that the two samples(residuals) were drawn from the same distribution, given the mean and standard deviation of the residuals.

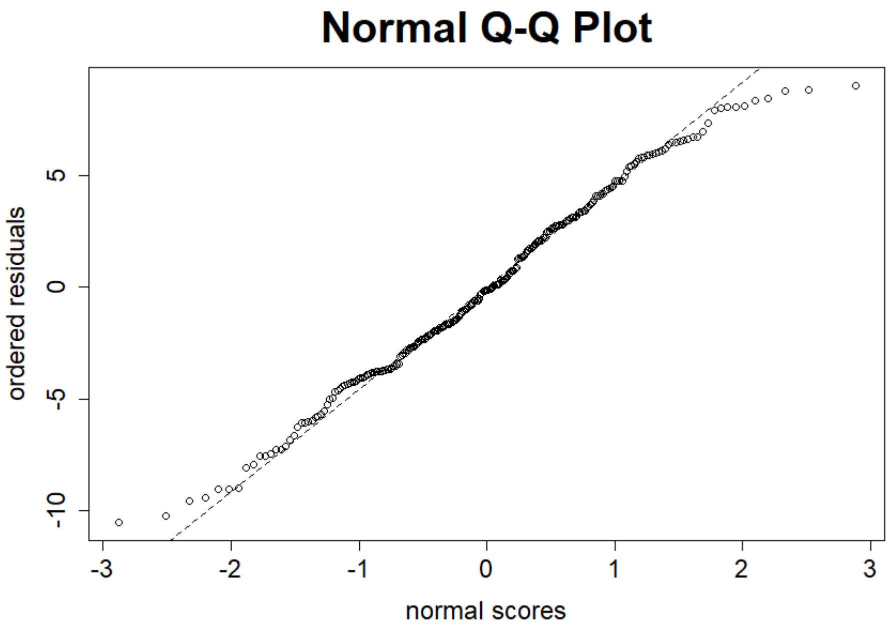


Figure 2: Normal Q-Q plot of the residuals against the normal scores

```
One-sample Kolmogorov-Smirnov test
D = 0.040639, p-value = 0.8015
alternative hypothesis:  two-sided
```

For checking the independence of the $\epsilon_i's$, we perform the Runs test as well as the Durbin-Watson test though the measurements don't have any relation with time. The D-W statistic is observed to be 1.803, and between 1.5 and 2.5, and so we can conclude that the residuals are not correlated. Moreover, we observe that the p-value for this test is $0.132 > 0.05$, and hence we can't reject the null hypothesis at the 5% significance level and we conclude that there is no autocorrelation between the residuals. Similarly, the runs test gives a p-value of $0.8333 > 0.05$, so we can't reject the null hypothesis at the 5% significance level, and hence conclude that there is no run pattern and hence, no autocorrelation.

```
Runs Test
Standard Normal = 0.21042, p-value = 0.8333
alternative hypothesis:  two.sided

lag Autocorrelation D-W Statistic p-value
1 0.09462391 1.803414 0.132
Alternative hypothesis:  rho != 0
```

4 Interpretations of Findings

4.1 Interpretation of R-squared and p-value

The final model we obtained (i.e. Model B) includes Abdomen, Weight, Wrist and Forearm as the predictors. Model B has a R-squared of 0.7308 and p-value of less than 2.2^{-16} (Figure 5). Since the R-squared is relatively high, and the p-value is low, we find that the model is useful in predicting body fat mass.

4.2 Cross Validation

To test the effectiveness of the model on unknown data, we split the original dataset into a training and testing set with ratio of 80:20.

After fitting the regression model on the training data, and validating it against the test set, we obtained the following results:

- Number of predictions within 95% prediction interval (out of 45): 43
- Number of predictions within ±5% from actually body fat percentage (out of 45): 33

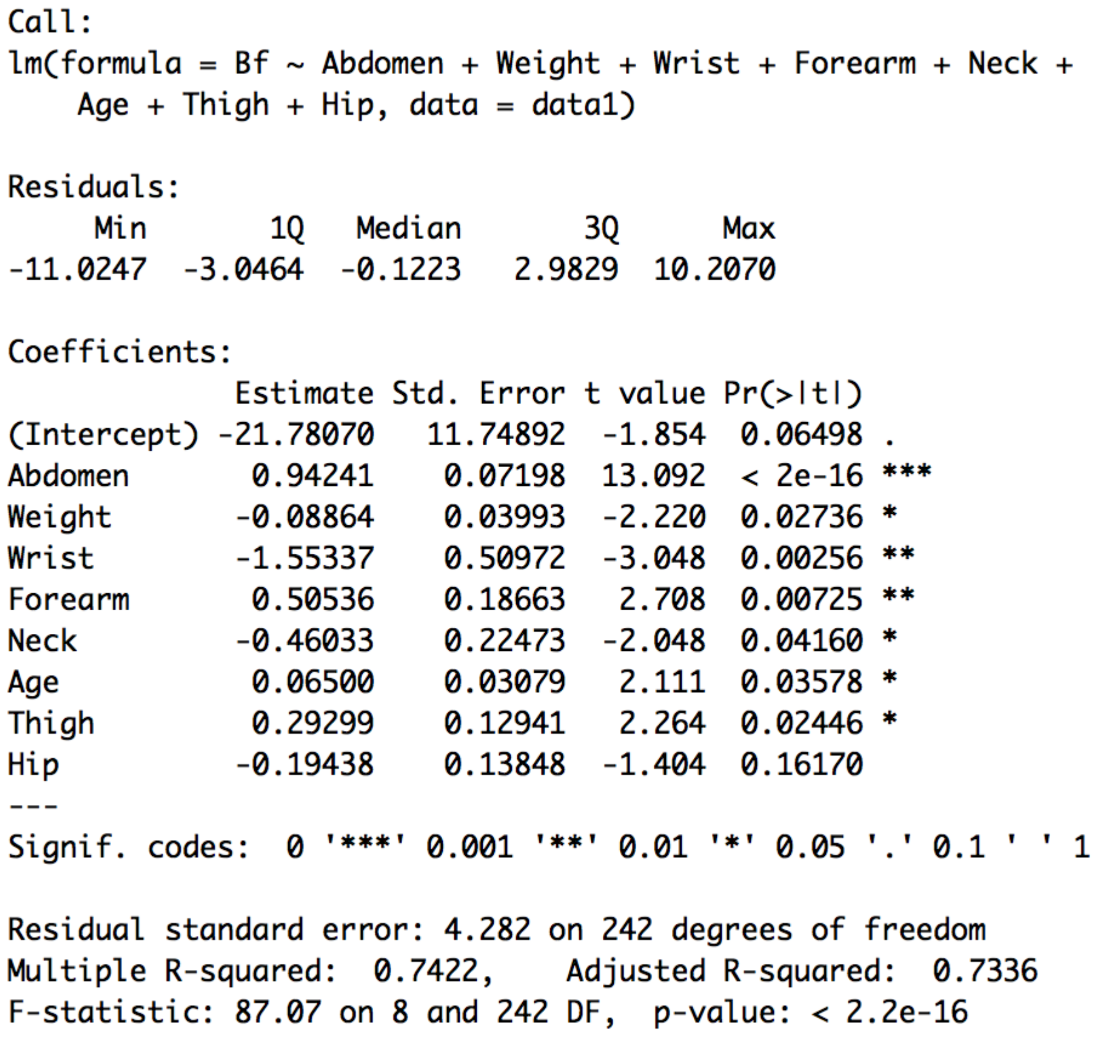
This shows that the model is quite accurate in predicting the body fat percentages.

5 Conclusion and recommendation for further work

In conclusion, the best fitted model was obtained by performing stepwise regression, partial F test and testing for higher-power and interaction terms. To validate the assumptions made for a linear regression model, residuals, outliers and influential observations were examined, and multicollinearity tests were also performed. The result is a statistically significant regression model that allows predicting body fat mass using readily-available measurements.

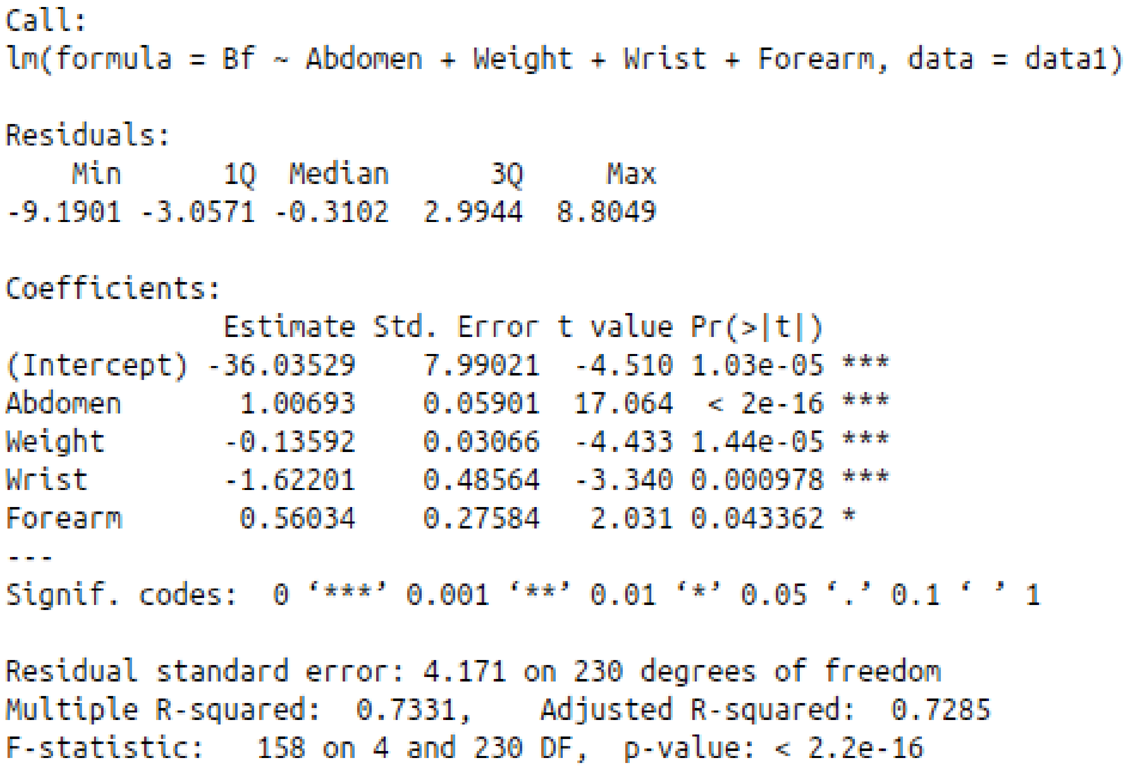
A possible limitation of this project would be that the data used was only taken from a single source. It is possible to gain more insight by examining other datasets for body fat mass such as the CDC Body Composition Dataset [cdc]. Thus the result of this project may only be applicable to a limited male demographic group. A natural extension to this project is to also to examine the effect of predictor variables on the body fat mass of female subjects. Doing so would make the model more effective in predicting the body fat percentage across various kinds of age groups and genders.

A Models' R-squared and significance



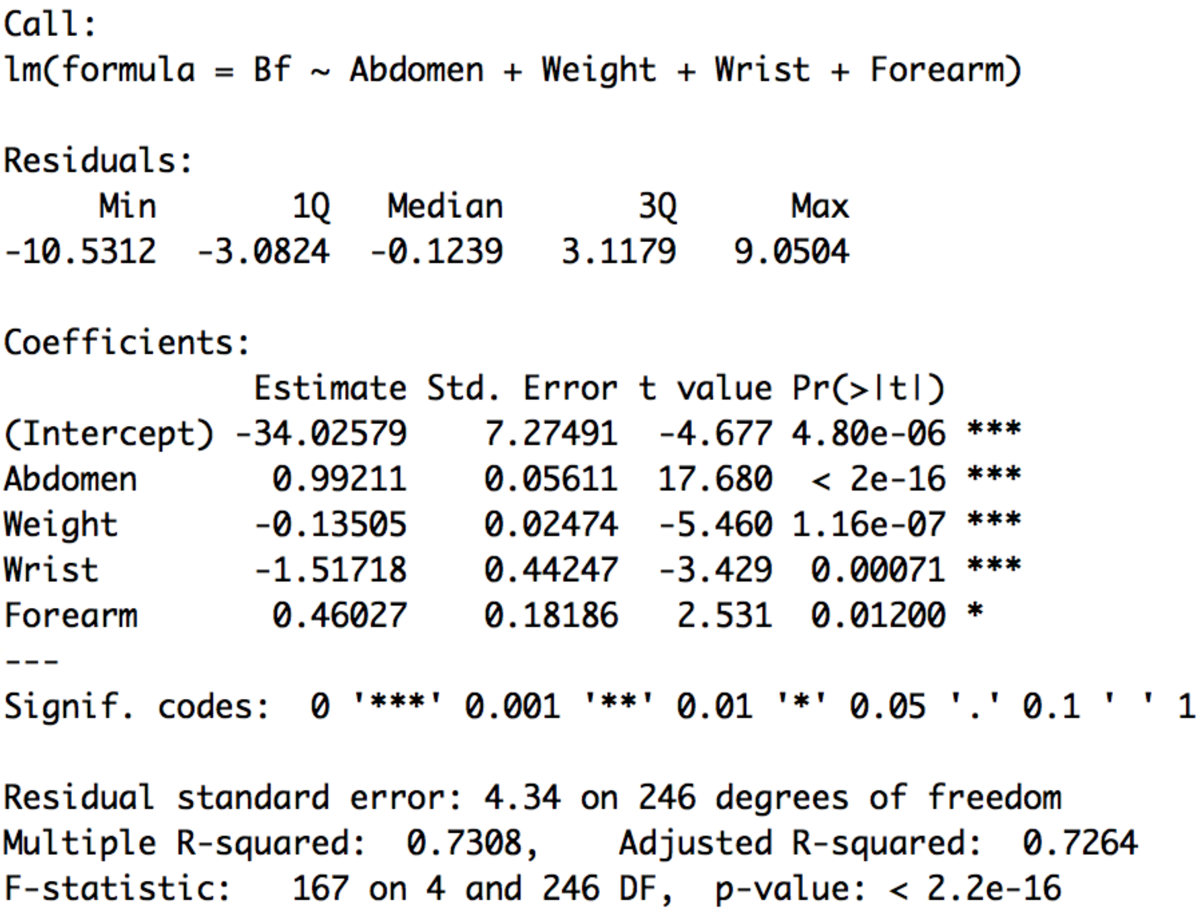
```
Call:
lm(formula = Bf ~ Abdomen + Weight + Wrist + Forearm + Neck +
    Age + Thigh + Hip, data = data1)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.0247  -3.0464  -0.1223   2.9829  10.2070 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -21.78070   11.74892  -1.854  0.06498 .  
Abdomen       0.94241    0.07198  13.092  < 2e-16 ***
Weight       -0.08864    0.03993  -2.220  0.02736 *  
Wrist        -1.55337    0.50972  -3.048  0.00256 ** 
Forearm       0.50536    0.18663   2.708  0.00725 ** 
Neck         -0.46033    0.22473  -2.048  0.04160 *  
Age           0.06500    0.03079   2.111  0.03578 *  
Thigh         0.29299    0.12941   2.264  0.02446 *  
Hip          -0.19438    0.13848  -1.404  0.16170    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.282 on 242 degrees of freedom
Multiple R-squared:  0.7422,	Adjusted R-squared:  0.7336 
F-statistic: 87.07 on 8 and 242 DF,  p-value: < 2.2e-16
```

Figure 3: Model A (post-stepwise-regression using AIC)



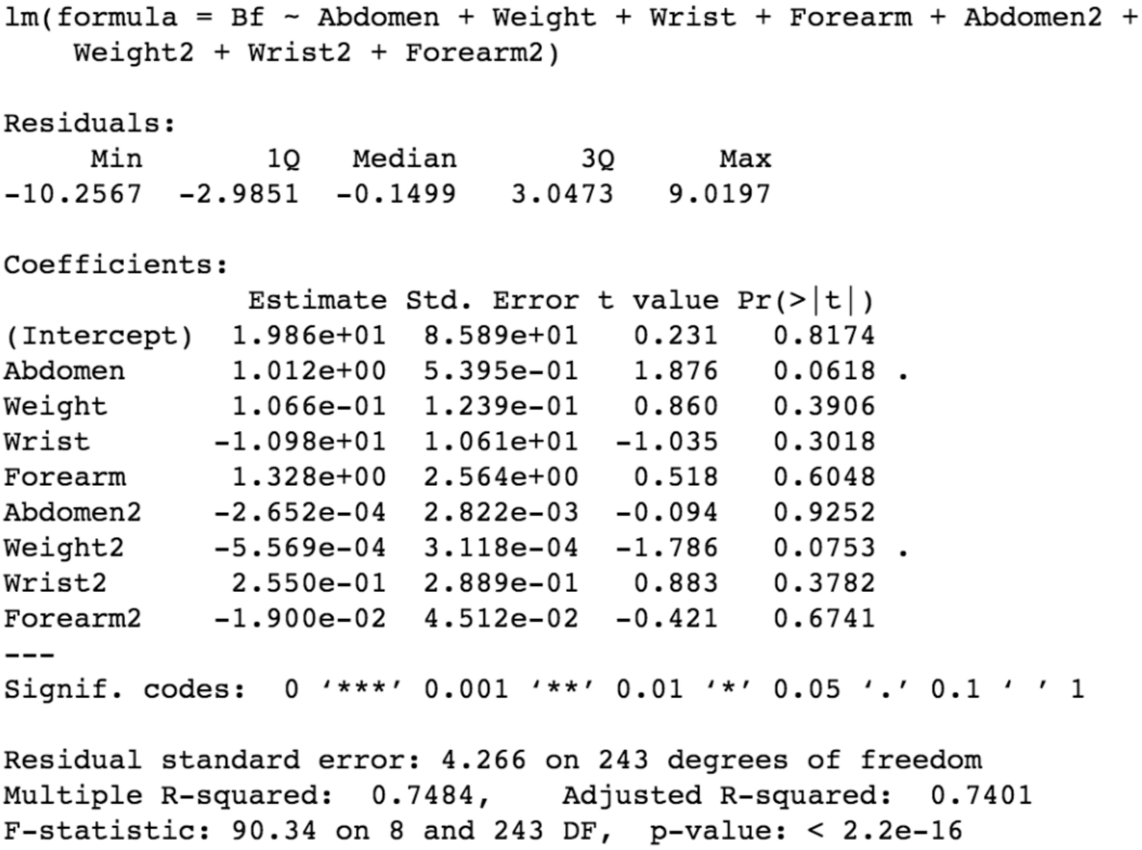
```
Call:
lm(formula = Bf ~ Abdomen + Weight + Wrist + Forearm, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.1901 -3.0571 -0.3102  2.9944  8.8049 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -36.03529    7.99021  -4.510 1.03e-05 ***
Abdomen       1.00693    0.05901  17.064  < 2e-16 ***
Weight       -0.13592    0.03066  -4.433 1.44e-05 ***
Wrist        -1.62201    0.48564  -3.340 0.000978 ***
Forearm       0.56034    0.27584   2.031 0.043362 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.171 on 230 degrees of freedom
Multiple R-squared:  0.7331,    Adjusted R-squared:  0.7285 
F-statistic:   158 on 4 and 230 DF,  p-value: < 2.2e-16
```

Figure 4: Model B with 4 predictors before removing outliers



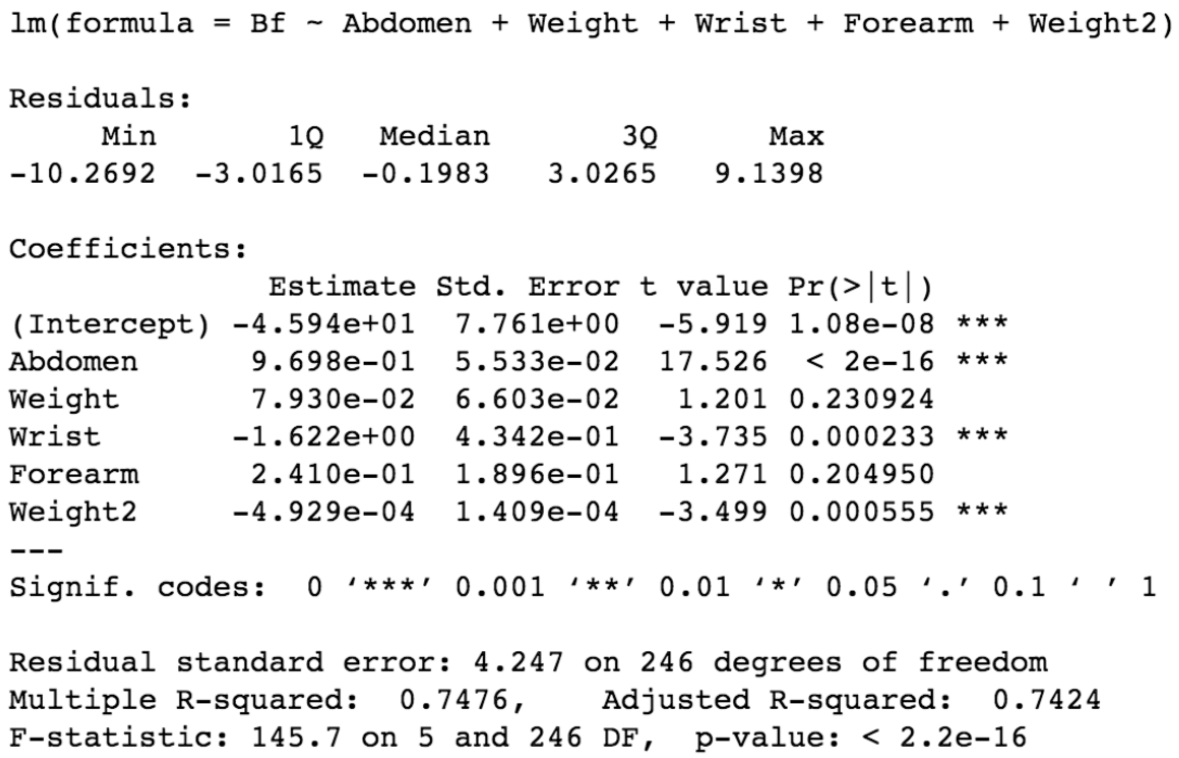
```
Call:
lm(formula = Bf ~ Abdomen + Weight + Wrist + Forearm)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.5312  -3.0824  -0.1239   3.1179   9.0504 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -34.02579    7.27491  -4.677 4.80e-06 ***
Abdomen       0.99211    0.05611  17.680  < 2e-16 ***
Weight       -0.13505    0.02474  -5.460 1.16e-07 ***
Wrist        -1.51718    0.44247  -3.429  0.00071 ***
Forearm       0.46027    0.18186   2.531  0.01200 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.34 on 246 degrees of freedom
Multiple R-squared:  0.7308,	Adjusted R-squared:  0.7264 
F-statistic:   167 on 4 and 246 DF,  p-value: < 2.2e-16
```

Figure 5: Model B with 4 predictors after removing outliers



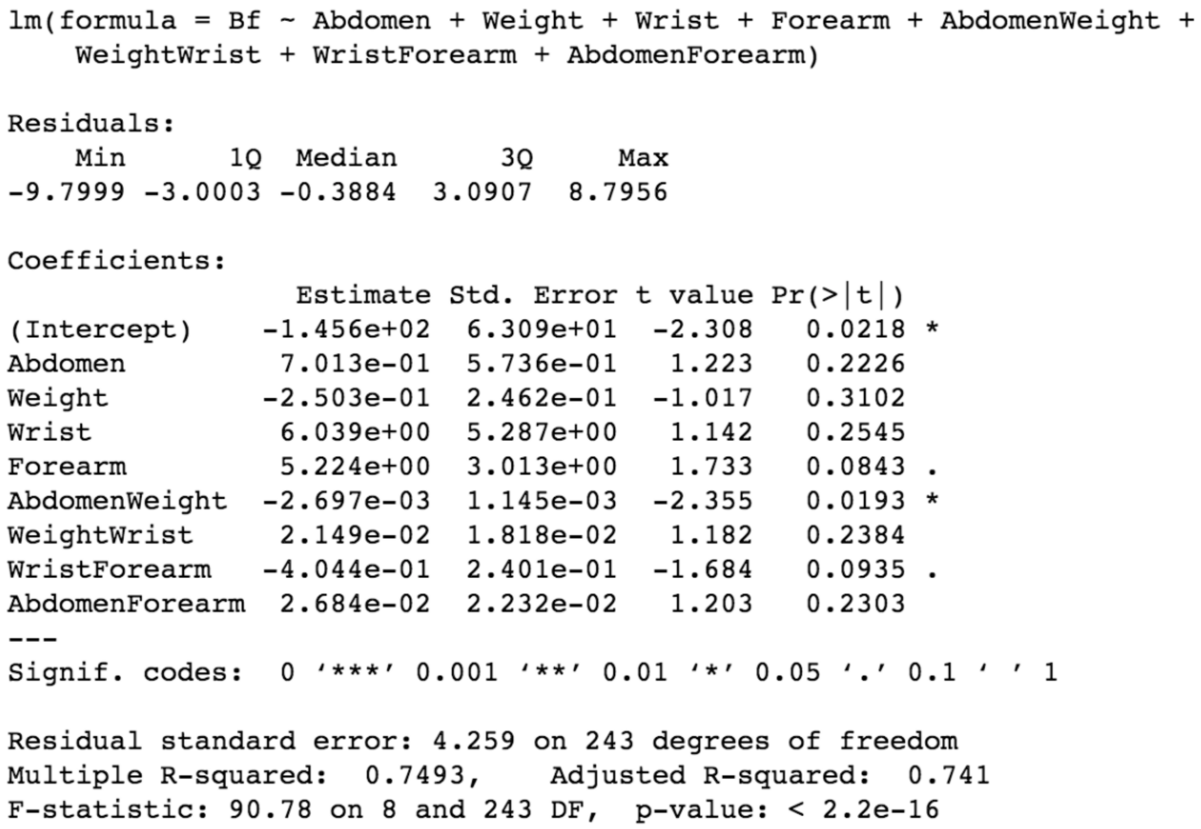
```
lm(formula = Bf ~ Abdomen + Weight + Wrist + Forearm + Abdomen2 +
    Weight2 + Wrist2 + Forearm2)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.2567  -2.9851  -0.1499   3.0473   9.0197 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)  1.986e+01  8.589e+01   0.231   0.8174  
Abdomen      1.012e+00  5.395e-01   1.876   0.0618 .
Weight       1.066e-01  1.239e-01   0.860   0.3906  
Wrist       -1.098e+01  1.061e+01  -1.035   0.3018  
Forearm      1.328e+00  2.564e+00   0.518   0.6048  
Abdomen2    -2.652e-04  2.822e-03  -0.094   0.9252  
Weight2     -5.569e-04  3.118e-04  -1.786   0.0753 .
Wrist2       2.550e-01  2.889e-01   0.883   0.3782  
Forearm2    -1.900e-02  4.512e-02  -0.421   0.6741  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.266 on 243 degrees of freedom
Multiple R-squared:  0.7484,    Adjusted R-squared:  0.7401 
F-statistic: 90.34 on 8 and 243 DF,  p-value: < 2.2e-16
```

Figure 6: Model B with Power-2 Terms



```
lm(formula = Bf ~ Abdomen + Weight + Wrist + Forearm + Weight2)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.2692  -3.0165  -0.1983   3.0265   9.1398 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.594e+01  7.761e+00  -5.919 1.08e-08 ***
Abdomen      9.698e-01  5.533e-02  17.526  < 2e-16 ***
Weight       7.930e-02  6.603e-02   1.201 0.230924    
Wrist       -1.622e+00  4.342e-01  -3.735 0.000233 ***
Forearm      2.410e-01  1.896e-01   1.271 0.204950    
Weight2     -4.929e-04  1.409e-04  -3.499 0.000555 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.247 on 246 degrees of freedom
Multiple R-squared:  0.7476,	Adjusted R-squared:  0.7424 
F-statistic: 145.7 on 5 and 246 DF,  p-value: < 2.2e-16
```

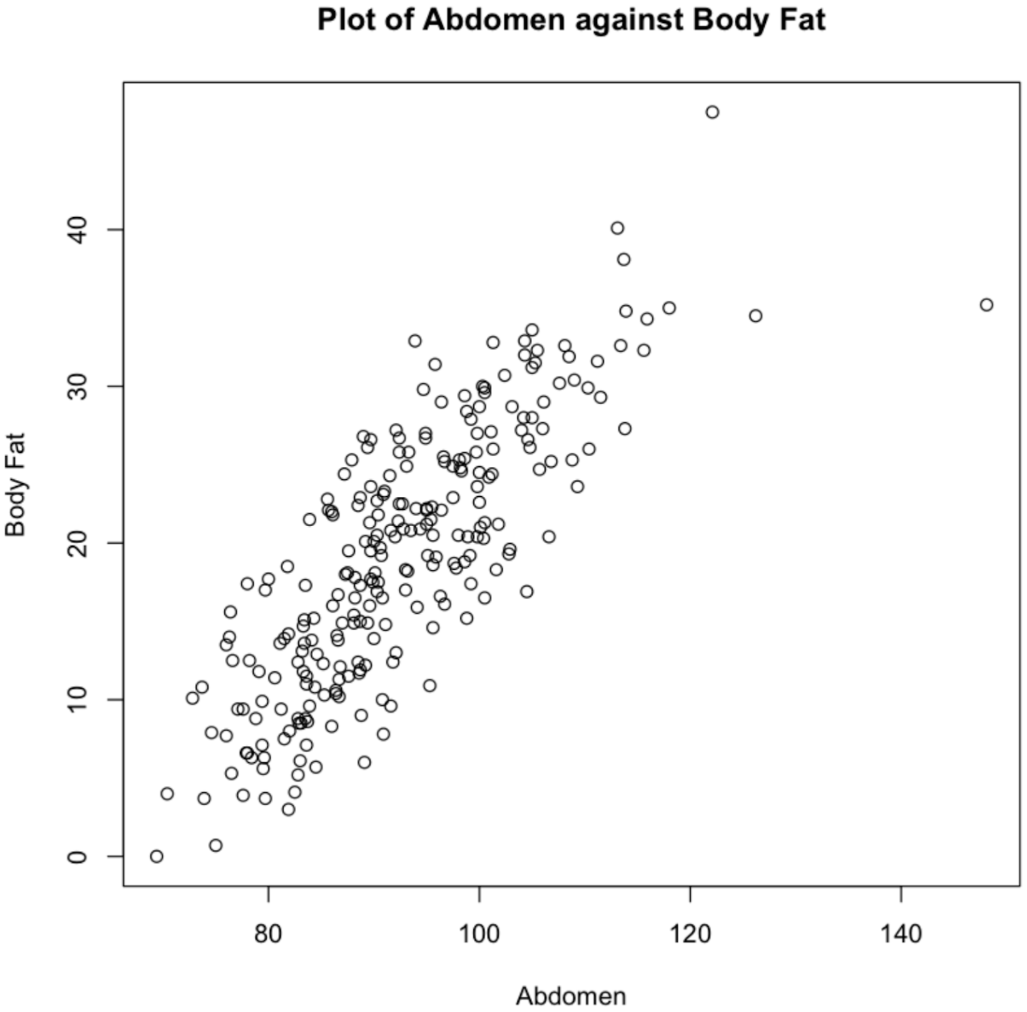
Figure 7: Model B with Weight-Squared

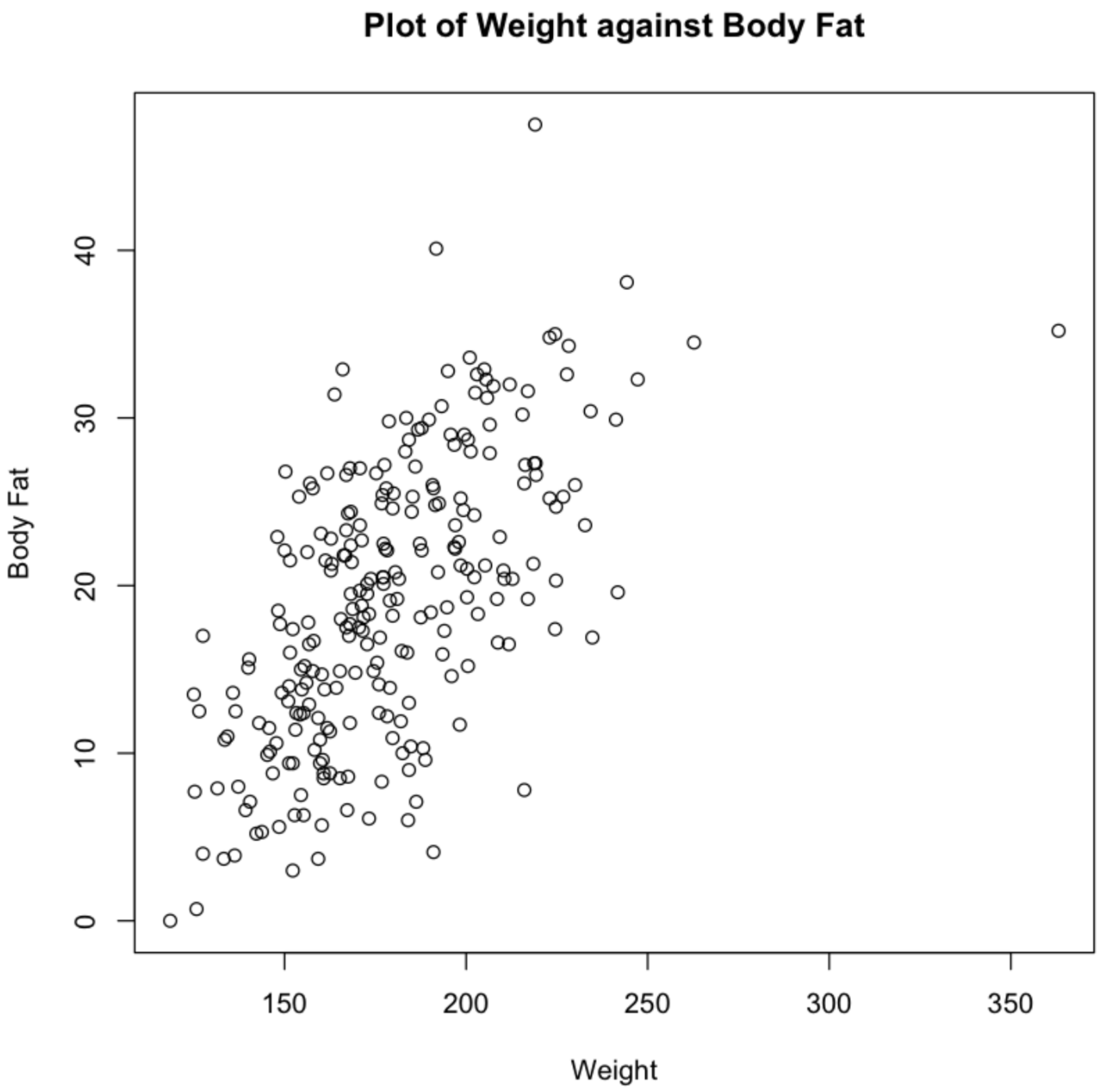


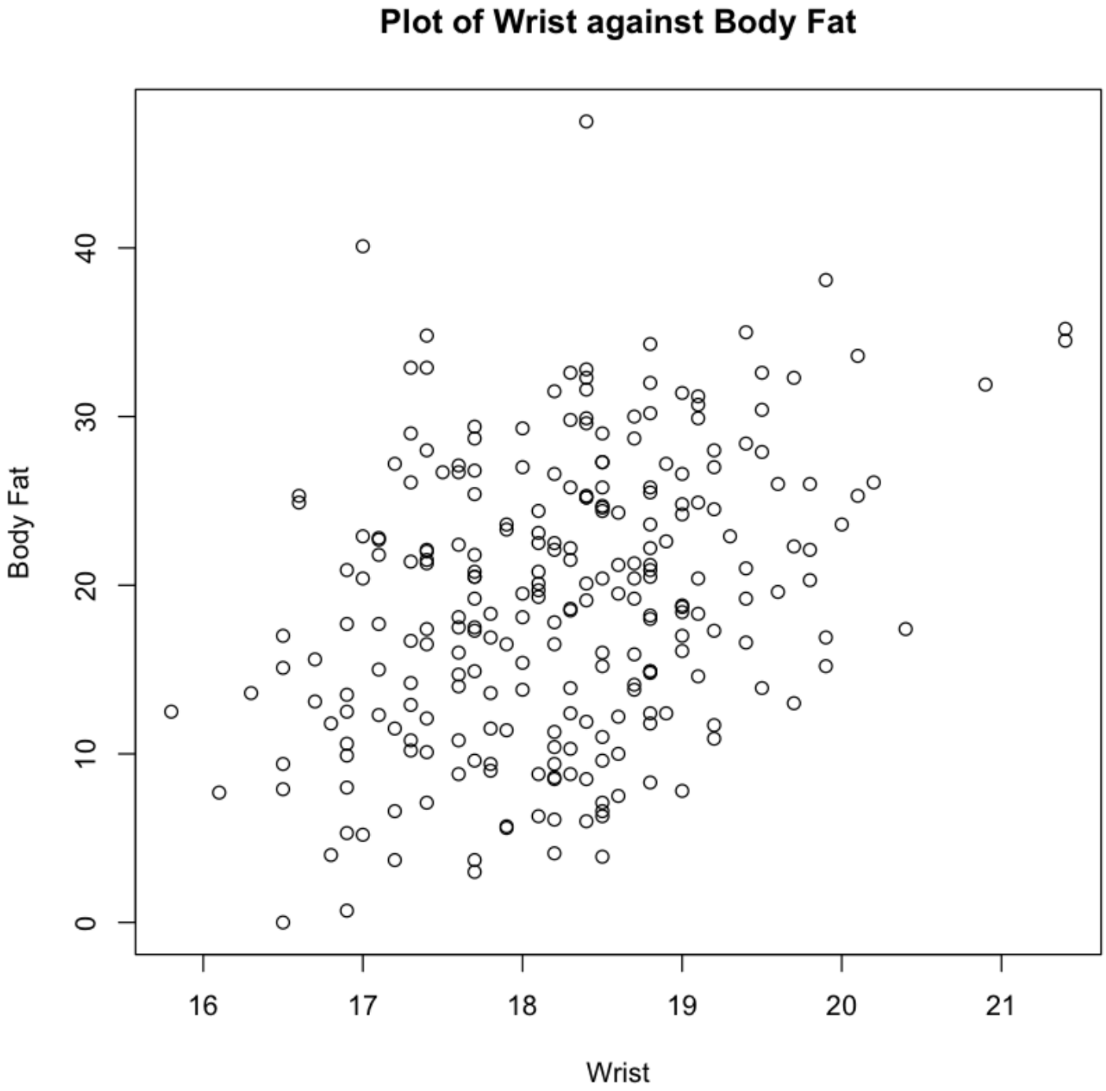
```
lm(formula = Bf ~ Abdomen + Weight + Wrist + Forearm + AbdomenWeight +
    WeightWrist + WristForearm + AbdomenForearm)

Residuals:
    Min      1Q  Median      3Q     Max
-9.7999 -3.0003 -0.3884  3.0907  8.7956

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -1.456e+02  6.309e+01  -2.308   0.0218 *
Abdomen         7.013e-01  5.736e-01   1.223   0.2226
Weight         -2.503e-01  2.462e-01  -1.017   0.3102
Wrist           6.039e+00  5.287e+00   1.142   0.2545
Forearm         5.224e+00  3.013e+00   1.733   0.0843 .
AbdomenWeight  -2.697e-03  1.145e-03  -2.355   0.0193 *
WeightWrist     2.149e-02  1.818e-02   1.182   0.2384
WristForearm   -4.044e-01  2.401e-01  -1.684   0.0935 .
AbdomenForearm  2.684e-02  2.232e-02   1.203   0.2303
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.259 on 243 degrees of freedom
Multiple R-squared:  0.7493,    Adjusted R-squared:  0.741
F-statistic: 90.78 on 8 and 243 DF,  p-value: < 2.2e-16
```

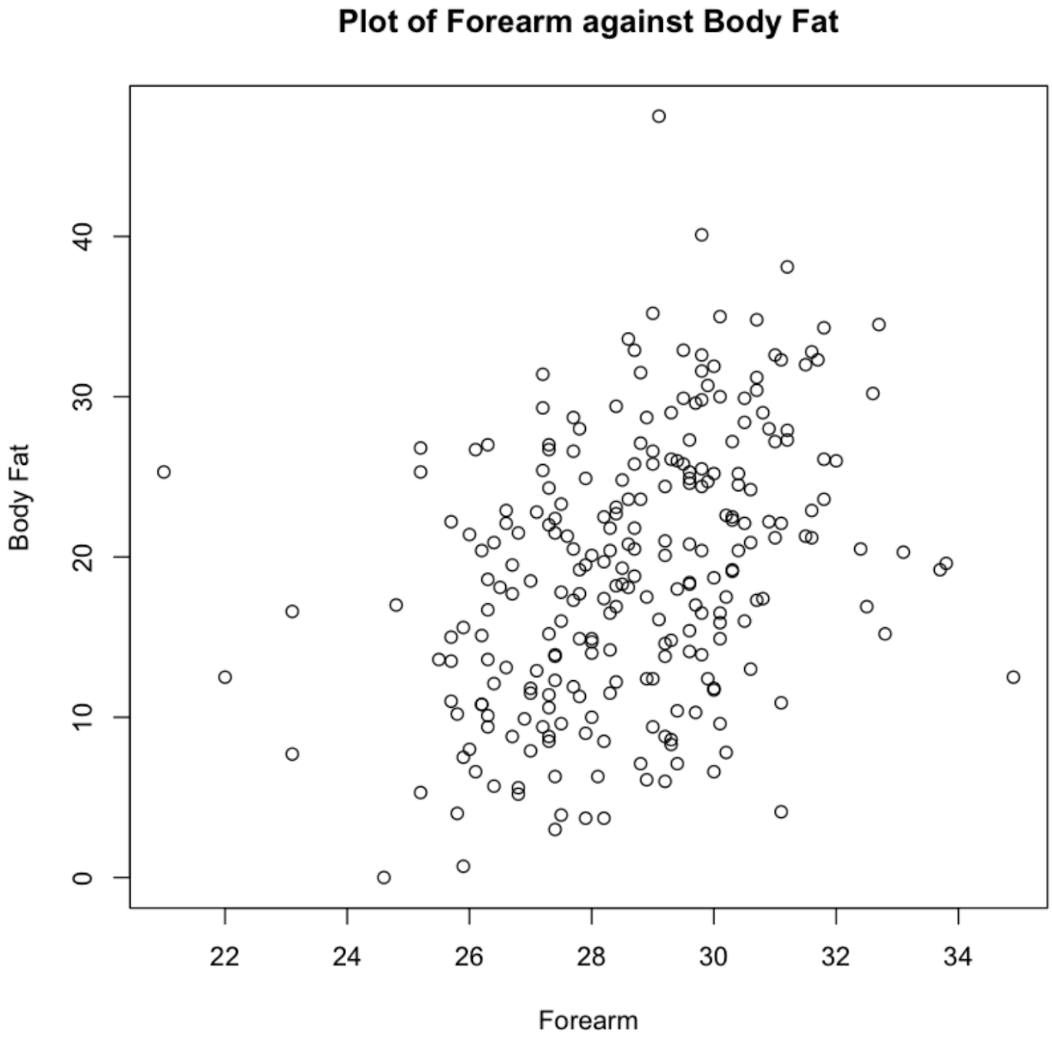
Figure 8: Model B with Interaction Terms

B Scatter plots of Body Fat against each predictor in Model B

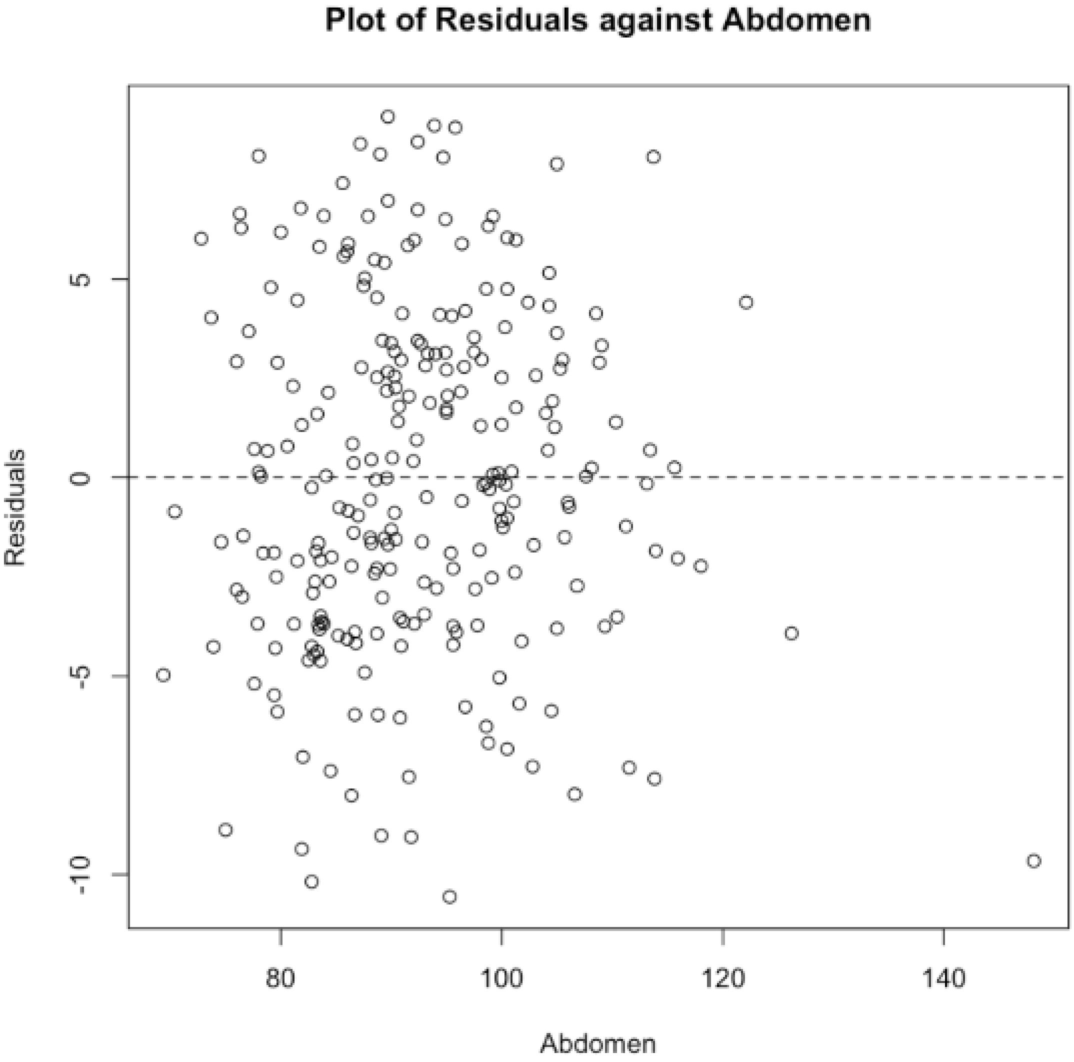


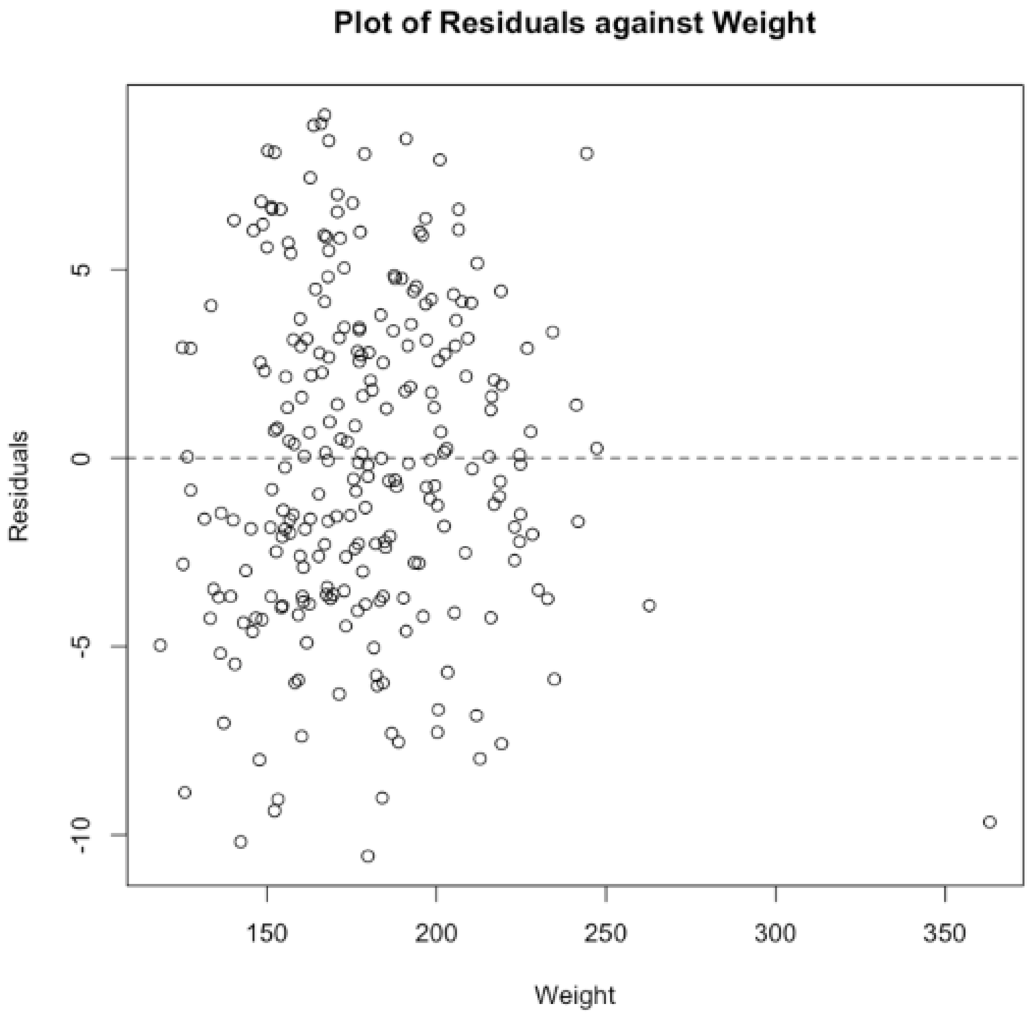

Plot of Weight against Body Fat
Body Fat
0
10
20
30
40
Weight
150
200
250
300
350

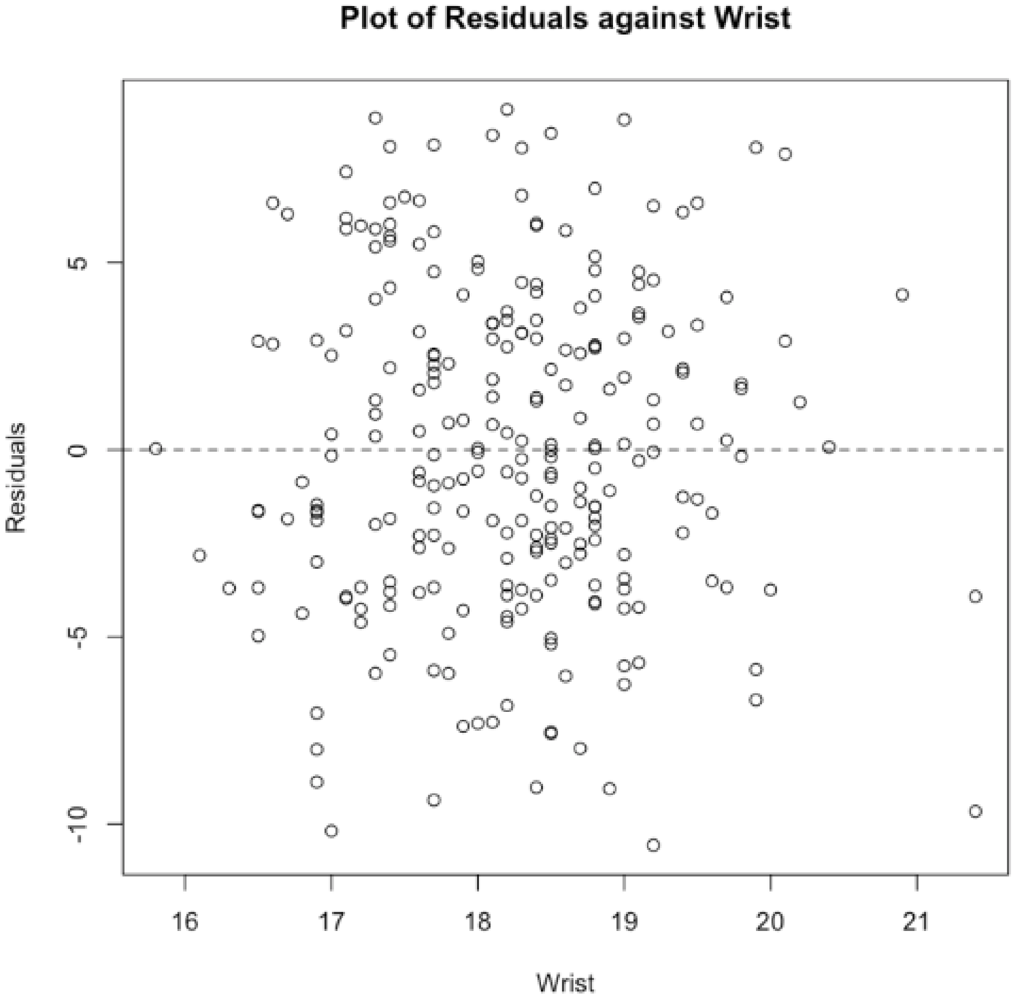

Plot of Wrist against Body Fat
Body Fat
0
10
20
30
40
16
17
18
19
20
21
Wrist

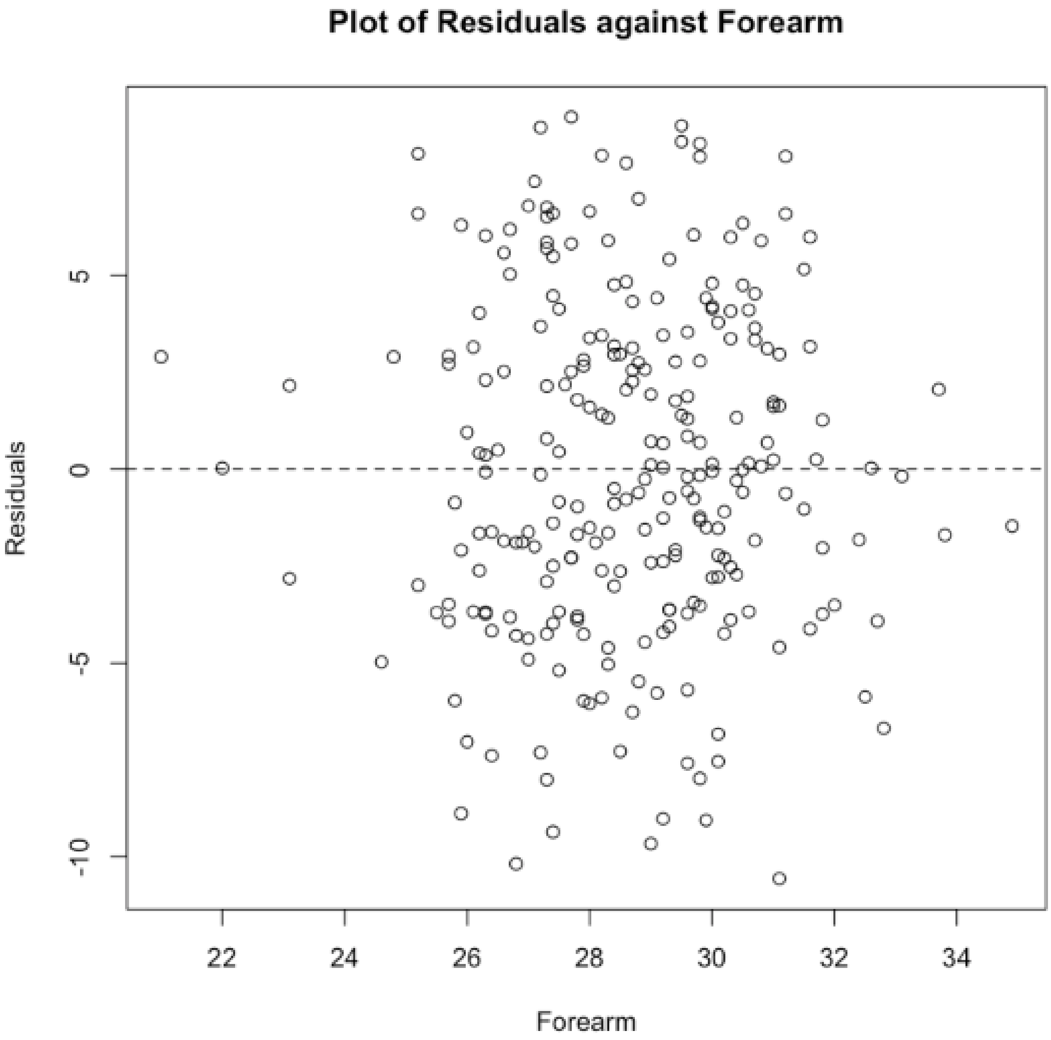

Plot of Forearm against Body Fat
Body Fat
0
10
20
30
40
22
24
26
28
30
32
34
Forearm

C Residual plots of residual values against each predictor in Model B

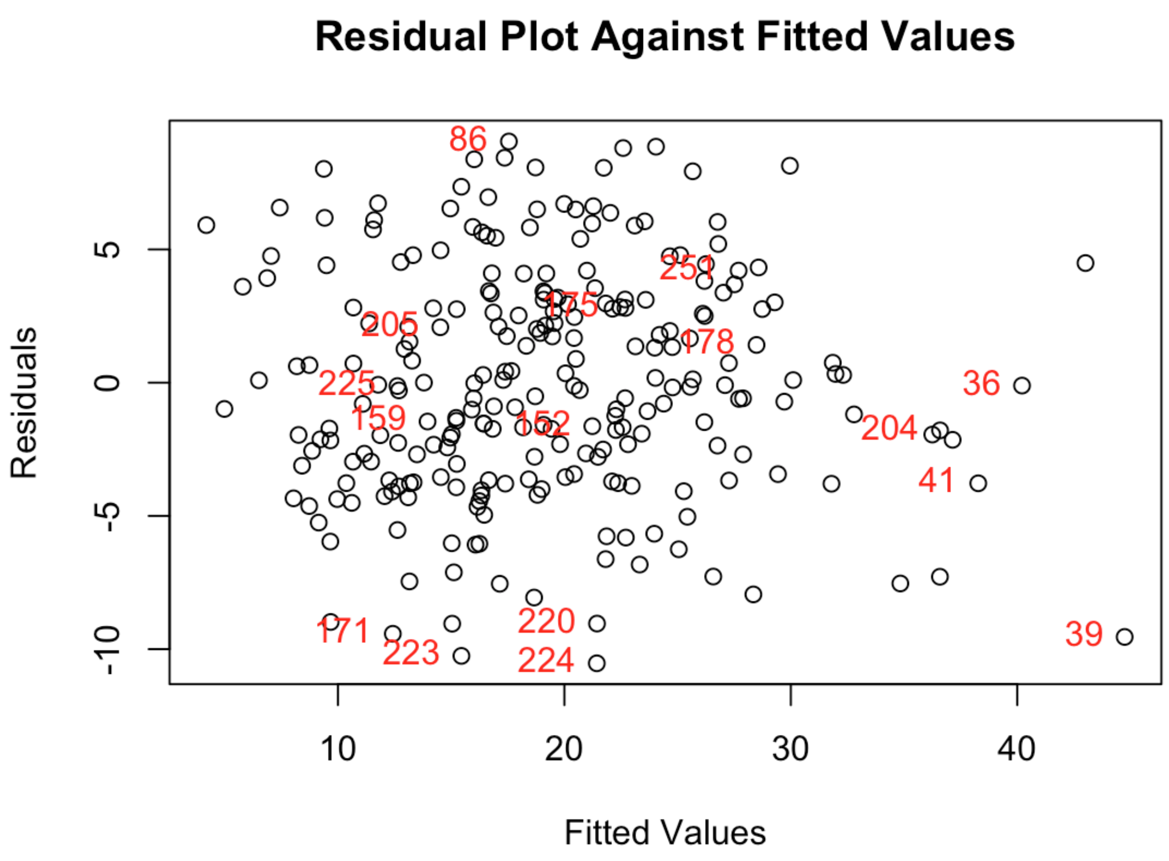



Plot of Residuals against Weight
Residuals
Weight
5
0
-5
-10
150
200
250
300
350


Plot of Residuals against Wrist
Residuals
5
0
-5
-10
16
17
18
19
20
21
Wrist


Plot of Residuals against Forearm
Residuals
5
0
-5
-10
22
24
26
28
30
32
34
Forearm

D Scatter plot of residuals against fitted values for model B, with labeled outliers



E R Code

```
library(car)
library(MASS)
library(tseries)
library(perturb)
library(caret)

data1<-read.table("Bodyfat.csv", header=T, sep = ";")
attach(data1)
#Correlation matrix ryx_i
Matrix<-cor(data1)# Highly correlated with Abdomen
```

```
#Doing the stepwise method

nullmodel <- lm(Bf~1, data=data1)
fullmodel <- lm(Bf~., data=data1)

#Stepwise function
model<-step(nullmodel,data=data1,
            scope = list(upper=fullmodel,lower=nullmodel),
            direction=”both”, k=2, test=”F”)
anova(model)
summary(model)

# Plotting residuals for the model, having Predictors
# : Abdomen + Weight + Wrist + Forearm + Neck + Age + Thigh + Hip

fitmodel<-model$fitted.values
resmodel<-model$res
plot(resmodel~fitmodel, main=”Residual Plot for Stepwise Model”,
     xlab=”Fitted Values (Bf)”, ylab=”Residuals (Bf)”)
abline(10, 0)
abline(-10, 0)

# Plot to check which powers of predictors are appropriate
plot(Abdomen, Bf, xlab=”Abdomen”, ylab=”Body Fat”)
plot(Weight, Bf, xlab=”Weight”, ylab=”Body Fat”)
plot(Wrist, Bf, xlab=”Wrist”, ylab=”Body Fat”)
plot(Forearm, Bf, xlab=”Forearm”, ylab=”Body Fat”)
plot(Neck, Bf, xlab=”Neck”, ylab=”Body Fat”)
plot(Age, Bf, xlab=”Age”, ylab=”Body Fat”)
plot(Thigh, Bf, xlab=”Thigh”, ylab=”Body Fat”)
plot(Hip, Bf, xlab=”Hip”, ylab=”Body Fat”)
# To check if the relationships are linear

#Analysis of obtained model
```

```
#Partial F-test
modelpartial1<-lm(Bf~Abdomen + Weight + Wrist + Forearm
                  + Neck + Age + Thigh ) # Testing Hip
anova(modelpartial1,model) # Remove Hip
#Testing thigh
modelpartial2<-lm(Bf~Abdomen + Weight + Wrist
                  + Forearm + Neck + Age )
anova(modelpartial1,modelpartial2) # Remove thigh
modelpartial3<-lm(Bf~Abdomen + Weight + Wrist
                  + Forearm + Neck ) #Testing age
anova(modelpartial2,modelpartial3) #remove age
modelpartial4<-lm(Bf~Abdomen + Weight + Wrist + Forearm ) #testing Neck
anova(modelpartial3,modelpartial4) #remove neck
modelpartial5<-lm(Bf~Abdomen + Weight + Wrist ) #testing forearm
anova(modelpartial4,modelpartial5) #keep forearm
modelpartial5<-lm(Bf~Abdomen + Forearm + Wrist ) #testing weight
anova(modelpartial4,modelpartial5) #keep weight
modelpartial5<-lm(Bf~Abdomen + Weight + Forearm ) #testing wrist
anova(modelpartial4,modelpartial5) #keep wrist
modelpartial5<-lm(Bf~Abdomen + Weight + Forearm ) #testing abdomen
anova(modelpartial4,modelpartial5) #keep abdomen

# lambda close to 1, so we take lambda = 1 for boxcox,
# which gives the same model

#Reduced model after partial F-tests
redmodel<-lm(Bf ~ Abdomen + Weight + Wrist + Forearm )
summary(redmodel) # Multiple R-squared:  0.7308,
# Adjusted R-squared:  0.7264
anova(redmodel)

# Remove outliers
influObs <- influence.measures(redmodel)
influObs
removeRows <- data.matrix(which(apply(influObs$is.inf, 1, any)))
# tagged for future studies
for (i in 1:NROW(removeRows)){
  data1 <- data1[-removeRows[i]+i-1,]
}

# Testing reduced model without outliers
redmodel_without_outliers <- lm(Bf ~ Abdomen + Weight + Wrist
                              + Forearm, data = datax)
summary(redmodel_without_outliers)
anova(redmodel_without_outliers)
res_without_outliers <- redmodel_without_outliers$res
fv_without_outliers <- redmodel_without_outliers$fitted.values
plot(res_without_outliers~fv_without_outliers)

#Reduced model
redmodel<-lm(Bf ~ Abdomen + Weight + Wrist + Forearm, data = data1 )
summary(redmodel)
anova(redmodel)

#Plotting graph for reduced model
redres<-redmodel$res
redfit<-redmodel$fitted.values
plot(redres~redfit)
with(datax[removeRows,], text(redres[removeRows]~redfit[removeRows], labels
plot(density(redres))
plot(Abdomen~redfit)
plot(Weight~redfit)
plot(Wrist~redfit)
plot(Forearm~redfit)

#Testing independece for reduced model
runs.test(factor(sign(redres)))
durbinWatsonTest(redmodel)

#Testing normality for reduced model
ks.test(redres,"pnorm",mean(redres),sd(redres))
qqnorm(redres,xlab="normal scores", ylab="ordered residuals")
qqline(redres,lty=2)
```

```
# whether should include higher power
Abdomen2 =  Abdomenˆ2;
Weight2 = Weightˆ2;
Wrist2 = Wristˆ2;
Forearm2 = Forearmˆ2;

power2model<−lm(Bf˜Abdomen + Weight + Wrist
          + Forearm + Abdomen2 + Weight2 + Wrist2 + Forearm2)
summary(power2model)
anova(power2model,redmodel)

# with only weight2
power2model2<−lm(Bf˜Abdomen + Weight + Wrist + Forearm + Weight2)
summary(power2model2)
anova(power2model2,redmodel)
# conclusion: add weight2

# without weight
power2model3<−lm(Bf˜Abdomen + Wrist + Forearm + Weight2)
summary(power2model3)
anova(power2model3,power2model2)
# conclusion: should not remove weight

#Plotting graph for reduced model with weight2
redres2<−power2model2$res
redfit2<−power2model2$fitted.values
par(mfrow=c(1,1))
plot(redfit2, redres2, xlab=”Fitted Values”, ylab=”Residuals”,
      main=”Plot of Residuals against Fitted Value”)
abline(h=0, lty=2)
# plot Weight and Weight2 against residual
par(mfrow=c(2,1))
plot(Weight, redres2, xlab=”Weight”, ylab=”Residuals”,
      main=”Plot of Residuals against Weight”)
abline(h=0, lty=2)
plot(Weight2, redres2, xlab=”Wrist”, ylab=”Residuals”,
      main=”Plot of Residuals against Weight2”)
```

```
abline(h=0, lty=2)
par(mfrow=c(1,1))

# whether should include interaction terms
AbdomenWeight =  Abdomen*Weight;
WeightWrist = Weight*Wrist;
WristForearm = Wrist*Forearm;
AbdomenForearm =  Abdomen*Forearm;

intmodel<-lm(Bf~Abdomen + Weight + Wrist + Forearm + Weight2
             + AbdomenWeight + WeightWrist
             + WristForearm + AbdomenForearm)
summary(intmodel)
anova(intmodel,power2model2)
# conclusion: won't include any cos all bad

# Test for homogenity of variance with Bartlett's test
# Not applicable since only 1 sample

modelweight2<-lm(Bf~Abdomen + Weight + Wrist + Forearm + Weight2)
summary(modelweight2)

# Box-cox transformation
modelbc <- boxcox(modelfinal, lambda = seq(-2,2,0.01))
val <- modelbc$x[modelbc$y==max(modelbc$y)] #Lambda = 1.01
#Laidata1$Bf <- data1$Bf - round(val, digits = 0)

# Variance Inflation Test for Reduced Model to detect multicollinearity
vif(modelfinal)
1/vif(modelfinal)
# All TOLs are not small, no evidence for multicollinearity

# Cross-validation
# Split Bf data into training set & testing set with ratio 80:20
split <- 0.80
trainIndex <- createDataPartition(data1$Bf, p=split, list=FALSE)
data_train <- data1[ trainIndex,]
```

```
data_test <- data1[-trainIndex ,]

train_model <-lm(Bf~ Abdomen + Weight + Wrist + Forearm + Weight2,
                 data=data_train)
summary(train_model)

X_test <- data.frame(Abdomen = data_test$Abdomen,
                     Weight = data_test$Weight,
                     Wrist = data_test$Wrist,
                     Forearm = data_test$Forearm)
actual <- data_test$Bf
predictions <- predict(train_model, X_test)

# Get 95% CI for mean and value
prediction_ci <-
        data.frame(predict(train_model, X_test,
                interval=”predict”, level=0.95))

# Count number of observations that is within
# the value intervals and
# within 5 from predicted values
nrow(data_test[data_test$Bf >= prediction_ci$lwr & data_test$Bf
               <= prediction_ci$upr, ]) # 44/45
nrow(data_test[abs(data_test$Bf - prediction_ci$fit)
               <= 5, ]) #30/45
nrow(data_test[data_test$Bf - prediction_ci$fit < 0, ])
nrow(data_test[data_test$Bf - prediction_ci$fit > 0, ])
```

References

[cdc] Body Composition Data for Individuals 8 Years of Age and Older: U.S. Population, 1999-2004. , Retrieved from https://www.cdc.gov/nchs/data/series/sr_11/sr11_250.pdf.

[dat] CMU Man Body fat Dataset. , Retrieved from http://lib.stat.cmu.edu/datasets/bodyfat.