Body fat and New Zealand

1. Introduction

At 3.15 am of October 12, 2017, Helen Clark, a New Zealand Prime Minister, released a piece of information on Twitter which said “Shocking 3rd placing of NZ on obesity league table. Whoever’s in government needs to implement strong policy to stop this epidemic”.

The appeal came from the latest data on the OECD “obesity data update” in 2017, with nearly 1/3 of New Zealanders in the category of obesity. The number which is still growing pushes New Zealand to the third place of the world.

In fact, obesity may not only influence our daily life, but also closely related to health. Till now, more than 4 million people die of overweight or obesity every year. In OECD, Japan, known as “the most longevity country”, has the lowest obesity rate. In stark contrast, the New Zealand medical system needs to pay more than 600 million dollars for obesity every year.

In order to “stop this epidemic”, some knowledge about obesity should be learned about. Therefore, as the most important indicator of obesity, body fat percentage play a significant role when we intend to figure out our own health situation. Body fat refers to the proportion of body fat in the body's total body weight, which reflects the amount of fat in our body.

This report describes the fitting of a body fat model, assess the model. The aim is to provide a simple and accurate method for predicting body fat, so that people could have a certain understanding of their body fat and adjust their lifestyle according to their physical condition. In this study, a total of 252 people of different ages and figures participated in the survey. The samples recorded their age, height, weight, and the circumference of the main parts of their body.

The model fitting process included data description, variable selection, unusual observations check, regression diagnostics and multicollinearity detection. After establishing a multiple linear model, the variables were combined according to the actual meaning, and interaction items were added to regress to get a model with interaction. Finally, the cross-validation method and the comparison of AIC value were conducted to compare and select the best model.

In the assessment of models, the level of the model evaluation in R language was used to express if the variable was significant, shown as “.” and “*” (the variable with “.” has less influence on the model being assessed than the variable with “*”). “***” stands for the variable with the most significance for the model.

2. Method

2.1 Data Description

Firstly, a descriptive statistical analysis of data was done for a comprehensive understanding of data and general relationship between variables.

To understand the five eigenvalues and the mean of each variable, summary() function was conducted to provide the basic statistics.

The fitting process was based on the multivariate linear model, which required the corresponding variable was approximately normal distributed. If the data did not meet this requirement, it needed to be transformed. Therefore, I used a histogram (hist() function in R) and added a density curve on it to show the distribution of the responding variable, which was Bodyfat. At the same time, the boxplot (boxplot() function in R) was also used to display the outliers clearly, since the outliers should be deleted before fitting the model to prevent their influence on the precise of the model.

After the preliminary data cleaning, the scatterplot matrix (using ggpairs() function in R) was used to present the relationship between the variables.

2.2 Variable Selection

In the selection of variables, the method of stepwise regression was first performed. Since the number of variables was large, the direction of step() was set as “backward”. That means all the variables are brought into the model and the one which is least significant (based on value of AIC) will be dropped. AIC takes the statistical fitting degree of the model and the number of parameters used to fit the model into account, so the model with small AIC should be preferred.

In order to verify the result of variable selection, the regsubsets () function was conducted to carry out the all subset regression, which would consider all the possible models and show the effect of variables on model fitting.

2.3 Unusual Observation Check and Model Fitting

After variables were selected, the model needed to be continuously optimized and the variables with great p value (represents they were not significant) should be removed. Next, it was unusual observations checking.

Unusual Observations generally concludes outliers, High-leverage observations and influential observations. Outliers are the points that are not well fitted, which are usually with a large positive or negative residual. High-leverage observations are outliers related to other predictive points. In other words, they are combined by a number of abnormal predictive points and have no relation to the value of the response variable. The influential points have imbalance effect on the estimated value of the model parameters. The existence of these points would seriously affect the precise of the model. The InfluencePlot() function was used to detect these unusual observations.

This process was repeated several times. After each exception point was processed, the model was refitted to ensure that the model was universal without being affected by unusual observations.

2.4 Regression Diagnostics

There are four basic assumptions for linear regression, namely, the independent variable has a linear relationship with the dependent variable, the residual is basically a normal distribution, the residuals variance is basically the same, and the residuals are independent. The functions plot() and durbinWatsonTest() were conducted to verify the assumptions.

2.5 Multicollinearity

Ideally, each predictor of the linear model should be linear independent. If there was a collinearity between the predictors, the accuracy of the regression model would be reduced. The variance expansion factor VIF (Variance Inflation Factor) was used to measure the collinearity. In general, it is considered that VIF should be less than 10.

2.6 Interaction

Usually, we need to use the interaction between two independent variables in regression analysis to test the interaction effect between the two variables. That is to say, the relationship between one variable and the responding variable is affected by the change of another variable.

In real life, chest circumference, waist circumference and hip circumference were used to measure the body's body shape. However, a single item was not enough to determine bodyfat, which required two or three of them to codetermine. Therefore, their interaction items were added to the original full model to conduct the regression.

In order to prevent the multicollinearity brought by the interaction item, the variables contained in the interaction item were centralized before adding the interaction in the model (using scale() function in R). Then, the whole model which took interaction into account was put into stepwise regression, regression diagnosis and unusual observation checking (using qqplot() function in R), and the optimal model was obtained.

2.7 Model Comparison

Since fitting model was not only for descriptive analysis, it was also needed for making prediction, so that people could learn about their body fat percentage well. Therefore, it was necessary to evaluate the prediction performance of the models and select the best model according to the results of the evaluation.

Since the number of samples was not big enough, it was unreasonable to divide the dataset into training sets and testing sets. Here, cross-validation was a very suitable method for evaluating the model. The data set was divided into ten subsets, made each of them to be testing set in turn and the remaining nine to be the training set. The average number of ten times validation was the final result of cross validation. In addition, in order to make sure the accuracy of model selection, the AIC values were also put into the comparison (a criterion for evaluating the regression model, the smaller the AIC value is, the better the model is).

3. Result

3.1 Data Description

The variables, units and summary statistics of the dataset are displayed in Table 1.

Table 1 Summary Statistics of Bodyfat Dataset

Bodyfat	Age(yrs)	Weight(inches)	Height(inches)	Neck(inches)
Min. : 0.00	Min. :22.00	Min. :118.5	Min. :64.00	Min. :31.10
1st Qu.:12.80	1st Qu.:35.75	1st Qu.:159.0	1st Qu.:68.25	1st Qu.:36.40
Median :19.00	Median :43.00	Median :176.5	Median :70.00	Median :38.00
Mean :18.94	Mean :44.88	Mean :178.9	Mean :70.31	Mean :37.99
3rd Qu.:24.60	3rd Qu.:54.00	3rd Qu.:197.0	3rd Qu.:72.25	3rd Qu.:39.42
Max. :45.10	Max. :81.00	Max. :363.1	Max. :77.75	Max. :51.20
Chest(inches)	Abdomen(inches)	Hip(inches)	Thigh(inches)	Knee(inches)
Min. : 79.30	Min. : 69.40	Min. : 85.0	Min. :47.20	Min. :33.00
1st Qu.: 94.35	1st Qu.: 84.58	1st Qu.: 95.5	1st Qu.:56.00	1st Qu.:36.98
Median : 99.65	Median : 90.95	Median : 99.3	Median :59.00	Median :38.50
Mean :100.82	Mean : 92.56	Mean : 99.9	Mean :59.41	Mean :38.59
3rd Qu.:105.38	3rd Qu.: 99.33	3rd Qu.:103.5	3rd Qu.:62.35	3rd Qu.:39.92
Max. :136.20	Max. :148.10	Max. :147.7	Max. :87.30	Max. :49.10
Ankle(inches)	Biceps(inches)	Forearm(inches)	Wrist(inches)	
Min. :19.1	Min. :24.80	Min. :21.00	Min. :15.80	
1st Qu.:22.0	1st Qu.:30.20	1st Qu.:27.30	1st Qu.:17.60	
Median :22.8	Median :32.05	Median :28.70	Median :18.30	
Mean :23.1	Mean :32.27	Mean :28.66	Mean :18.23	
3rd Qu.:24.0	3rd Qu.:34.33	3rd Qu.:30.00	3rd Qu.:18.80	
Max. :33.9	Max. :45.00	Max. :34.90	Max. :21.40	

Figure 1 shows the histogram and density curve of Bodyfat. The following figure is the boxplot of the responding variable, which presents both distribution and outliers clearly.

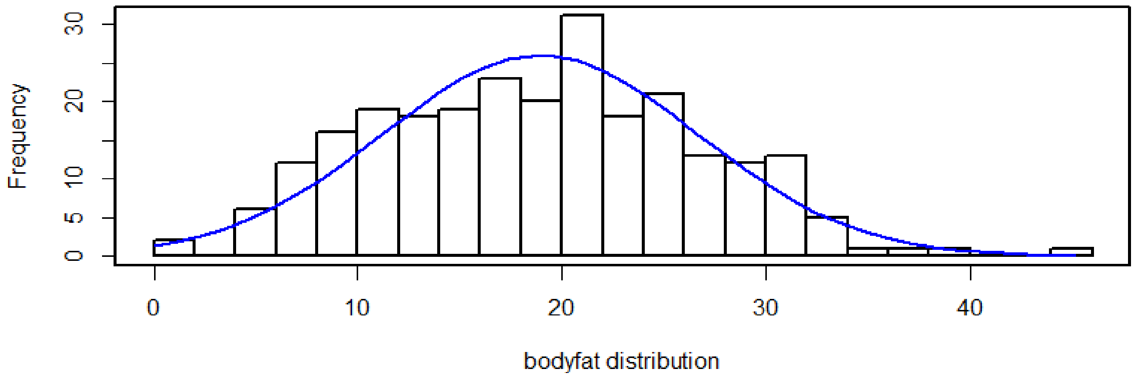


Figure 1 Distribution of 'Bodyfat'

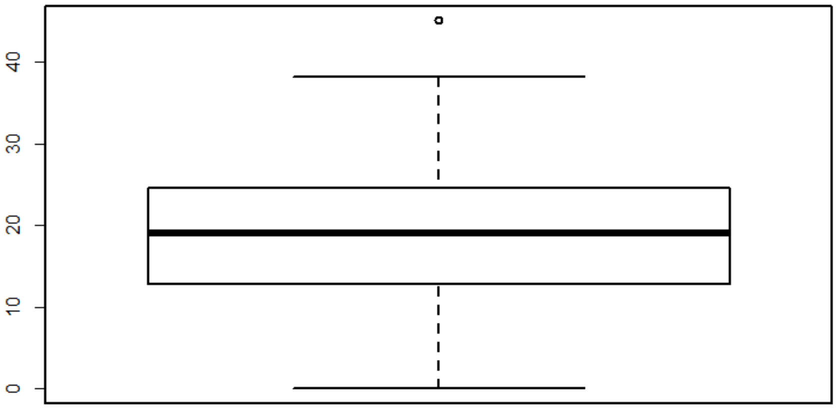


Figure 2 Boxplot of Bodyfat

The relationship between variables are explained by scatterplot as well as correlation which are shown in Figure 3. The graph shows that all the variables perform approximately normal distribution, and there is a good linear relationship between bodyfat and most of the independent variables.

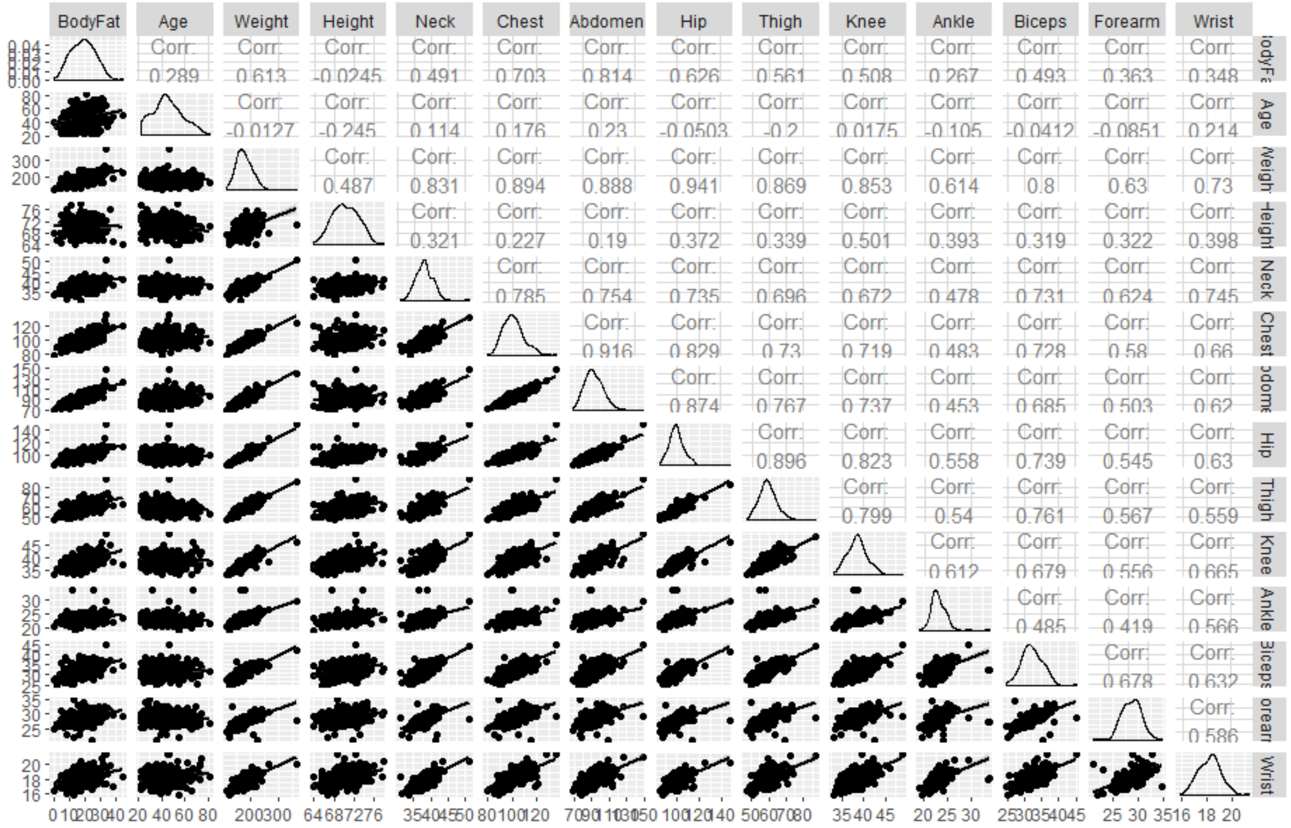


Figure 3 Scatter Plot Matrix for All Variables

3.2 Variable Selection

The summary statistics of the model got after the execution of the backward selection is as follows,

Table2 Summary Statistics of Original Model

	Estimate	Standard Error	t value	Pr(>|t|)	assessment
Intercept	7.266	7.588	0.958	0.3392	
Age	0.054	0.023	2.378	0.0182	*
Height	-0.302	0.111	-2.709	0.0072	* *
Neck	-0.293	0.202	-1.449	0.1485	
Chest	-0.136	0.081	-1.673	0.0957	.
Abdomen	0.812	0.060	13.553	<2e-16	* * *
Forearm	0.331	0.177	-1.864	0.0635	.
Wrist	-1.604	0.447	-3.585	0.0004	* * *
Multiple R-squared: 0.7444	Adjusted R-squared:0.737	p-value:<2.2e-16	F-value:100.7		

The model with age, height, chest, abdomen and wrist was finally obtained on the basis of the stepwise regression model (removed the predictor with large p value). The summary statistics is shown in Table 3.

Table 3 Summary Statistics of Optimized Model

	Estimate	Standard Error	t value	Pr(>|t|)	assessment
Intercept	7.315	7.573	0.966	0.335	
Age	0.045	0.022	2.058	0.0407	*
Height	-0.307	0.112	-2.741	0.0066	**
Chest	-0.134	0.078	-1.722	0.0864	.
Abdomen	0.801	0.060	0.056	0.0003	* * *
Wrist	-1.609	0.391	-4.117	5.25E-05	* * *
Multiple R-squared: 0.7396	Adjusted R-squared:0.7343	p-value:<2.2e-16	F-value:138.6		

The running result of all subsets regression could be checked in Figure 4.

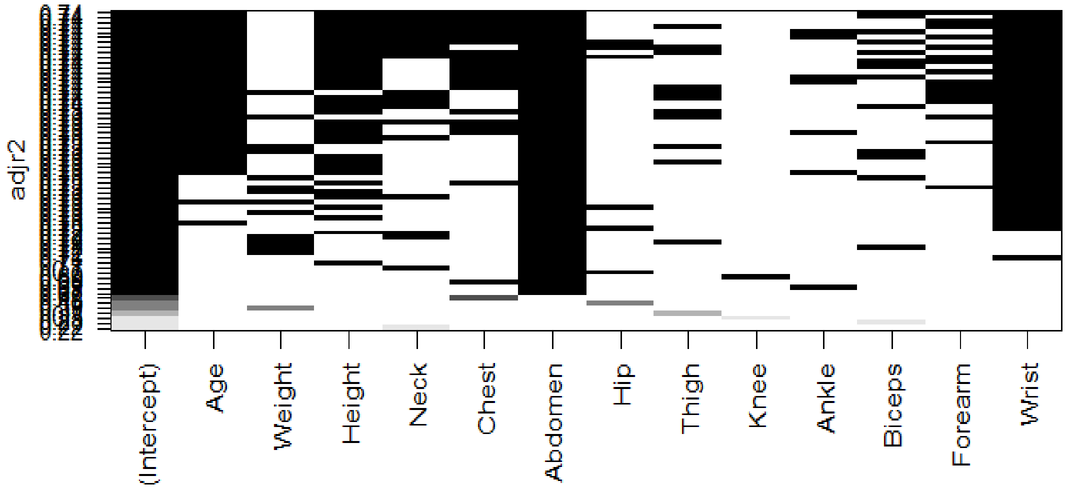


Figure 4 Result of All Subsets Regression

3.3 Unusual Observation Check

The detection of outliers, high-leverage observations and influential observations is shown in Figure 5. States above +2 or below -2 on the vertical axis could be considered outliers (222), and those above 0.2 or 0.3 on the horizontal axis had high leverage (40). The size of the circle was proportional to the influence of the point, and the large circle might have a strong disproportionately influence on the parameters of the model (248).

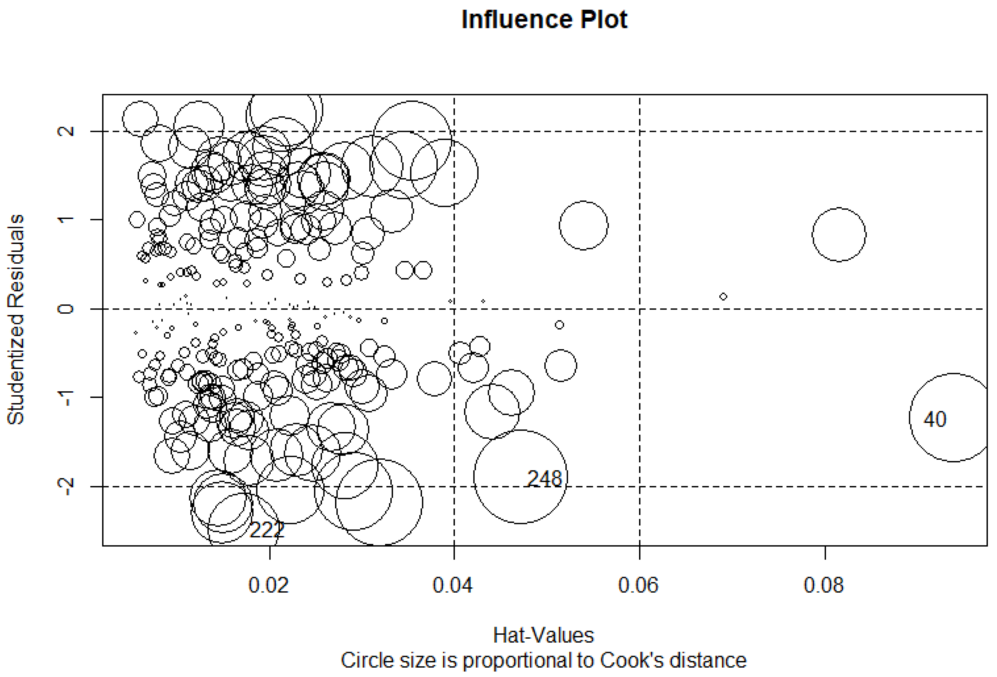


Figure 5 Unusual Observations

The final model after dropping unusual observations (Influence Plot of final model could be checked in Figure 1 in Appendix 2) and refitting is shown as follows (3dp),

Bodyfat = 6.75 + 0.06 * Age - 0.351 * Height + 0.728 * Abdomen - 1.807 * Wrist (MODEL 1)

Table 4 Summary Statistics of Final Linear Model

	Estimate	Standard Error	t value	Pr(>|t|)	assessment
Intercept	6.750	7.344	0.919	0.359	
Age	0.060	0.022	2.737	0.0067	* *
Height	-0.351	0.112	-3.121	0.002	* *
Abdomen	0.728	0.032	23.097	<2e-16	* * *
Wrist	-1.807	0.374	-4.833	2.41E-06	* * *
Multiple R-squared: 0.7377	Adjusted R-squared:0.7333		p-value:<2.2e-16		F-value:168

3.4 Regression Diagnostics

The result of conducting plot() towards the fitted model which was used for verifying the assumptions of linear regression is shown in Figure 6.

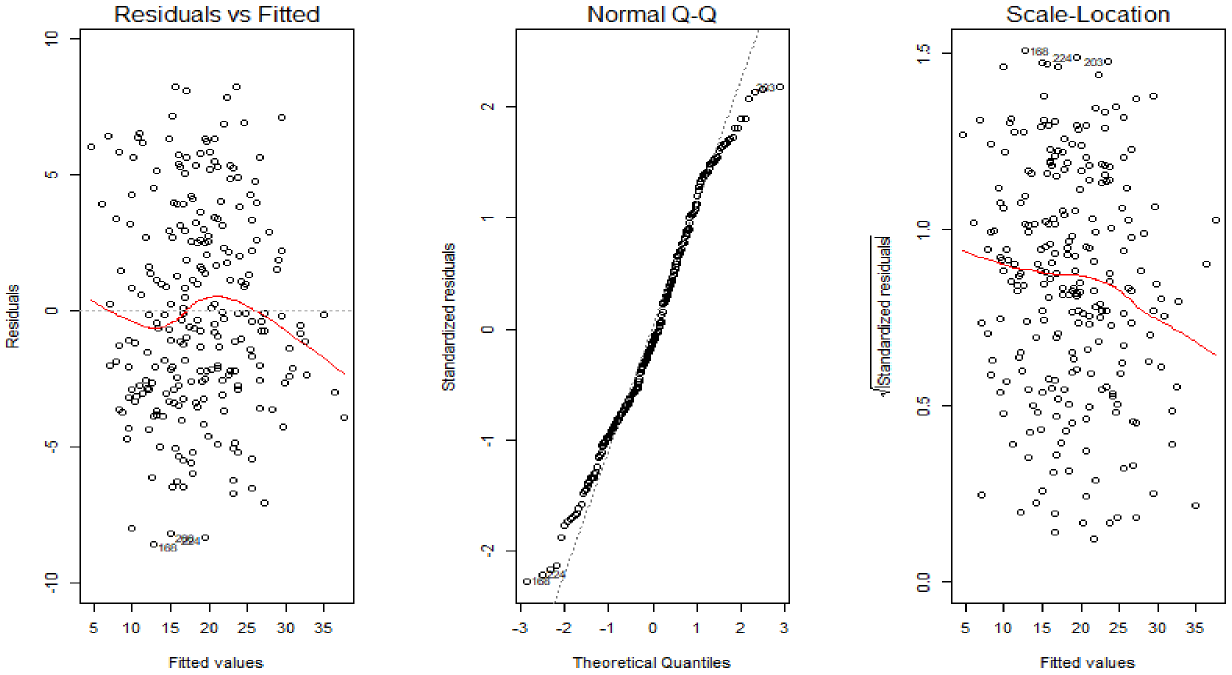


Figure 6 Regression diagnostics of percent bodyfat

The mutual independence between the residuals was done by durbinWatsonTest () function (using R), and resulting that p value was greater than 0.05 (p=0.192), which indicated that the residuals were independent.

3.5 Multicollinearity

The VIF values of variables presented Table 5 were all below 10, therefore, there was no problem of multicollinearity in this model.

Table 5 VIF Value of Variables in Final Linear Model

Age	Height	Abdomen	Wrist
1.240	1.404	1.605	1.884

3.6 Interaction

After centralization of the three variables, the model with interaction and its summary statistics are shown as follows, (the regression diagnosis and unusual points checking of the model are shown in Figure 2 and Figure 3 in Appendix 2).

Bodyfat = 76.432 - 0.432 * Height -0.125 * Chest + 0.825 * Abdomen

- 1.465 * Wrist – 0.005 * Chest: Abdomen (MODEL 2)

Table 6 Summary Statistics of Final Linear Model

	Estimate	Standard Error	t value	Pr(>|t|)	assessment
Intercept	76.432	8.418	9.079	<2e-16	* * *
Height	-0.432	0.106	-4.082	6.06E-05	* * *
Chest	-0.125	0.079	-1.593	0.112	
Abdomen	0.825	0.059	14.019	<-2e-16	* * *
Wrist	-1.465	0.379	-3.869	0.0001	* * *
Chest: Abdomen	-0.005	0.002	-1.121	0.035	*
Multiple R-squared: 0.7399	Adjusted R-squared:0.7346		p-value:<2.2e-16		F-value:138.8

3.7 Model comparison

The result of cross-validation and AIC value of two fitted more are presented in Table 7.

Table 7 Summary Statistics of Final Linear Model

	Original R-Square	Cross-Validated R-Square	AIC
MODEL1	0.73	0.719	1348
MODEL2	0.74	0.715	1402

4. Discussion

4.1 Discussion on the Process of Regression

· Figure 1 shows that the distribution of Bodyfat is approximately normal distributed, which means there is no need to transform the responding variable.

According to the boxplot of Bodyfat (Figure 2), both the maximum (outlier in boxplot) and minimum (which is equal to 0) is unreasonable. Therefore, they were dropped at the beginning.

· After choosing prediction variables, I used the all subsets regression to check the results of stepwise regression, which also made the influence of each variable on the model more intuitive. It could be seen from the Figure 4 that the models with age, height, chest, abdomen, wrist perform better, which support the result we get in step regression.

· In Figure 6, the scatter points in "Residuals Figure" perform a random distribution around the line, which indicates that the linear relationship is good. The same condition could be found in "Standard Residuals Figure", which shows that the variance of the residual is roughly the same. The scattered points in "Q-Q Graph" are mostly concentrated on the straight lines, therefore, normality of the residuals is good enough.

Since the model was multielement, the car () package was also used to verify the linear regression hypothesis of respective variables (Result is shown in Figure 4 in Appendix 2).

· What we could figure out from Table 6 is that the interaction is significant, indicating that the relationship between circumference of chest and bodyfat depends on the circumference of abdomen to a certain extent. As abdomen circumference varies, the impact of chest circumference on bodyfat is different. This is of practical significance. The reason why being with large chest circumference for a lot of people may be congenital reason or physical fitness instead of high bodyfat.

· Table 7 tells that the change of R-Square of MODEL1 is smaller, which also offers a lower AIC value. Therefore, although the R-Square of MODEL2 is slightly greater, we choose MODEL1 as the final model.

· In addition, I did not detect outliers, influential and high-leverage observations separately in the detection of unusual values. Instead, to use the influencePlot() function to test all the unusual observations at a time. This is not only more convenient, but also a better display of the result by figure (more clearly than Figure 3 in Appendix).

When adding an interaction, there were a lot of problems. First, it was the selection of interaction, which should be based on the actual significance of the data and the previously selected, and then continue to carry out a regression attempt. The second problem was multicollinearity, which was solved by centralization.

4.2 Model Interpretation

From the selected model 1, we could aware that the factors that have relatively great influence on bodyfat are circumference of Age, Height, Abdomen and Wrist.

Among them, age and waist have a positive correlation to bodyfat.

That is to say, with the increase in age, the bodyfat increases slightly (the coefficient is 0.06). About this positive relationship, my understanding is that for most people, if the condition of life remains unchanged, the metabolism of human body will slow down with increasing age and make fat easy to accumulate inside the body.

The increase in the size of the waist also contributes to the increase of the body fat. The circumference of chest and hip may vary a lot because of gender, fitness, or innate reasons. However, the change in abdominal circumference caused by abdominal fat increases, to a certain extent, could reflect the bodyfat and whether there is a state of overweight and obesity.

The bodyfat is negatively correlated with height and wrist. Based on the analysis of the database, the growth of height will reduce the bodyfat slightly. At the same time, the growth of the wrist circumference has a similar effect. Usually, the circumference of wrist is measured to determine the size of the body's skeleton. Therefore, in this model, height and wrist can describe the frame of the human body. According to the model, when the age and abdomen circumference are in the model, the bodyfat of people with larger human frame will be lower. This result is partly influenced by gender since the bodyfat of women is generally higher than that of men, at the same time, the frame is smaller than that of men.

Conclusion

The final prediction model for this regression covers Age, Height, Abdomen circumference and Wrist circumference. The concrete form of the model is,

Bodyfat = 6.75 + 0.06 * Age - 0.351 * Height + 0.728 * Abdomen - 1.807 * Wrist (3 dp)

The model is a simple multiple linear model with 73.33% R-SQURE, high fitting degree and good performance in cross-validation.

Therefore, the bodyfat can be measured by this model after measuring the waist circumference, height, and wrist circumference (unit: inches). The normal bodyfat of women should be less than 28%, and the male should be less than 18%. If it is beyond the normal range, a healthy body should be guaranteed by proper control of the diet and exercise.

Number of Friends on Facebook

1. Introduction

According to <The Atlantic>, in 1950, less than 10% of the Americans lived alone, but by 2010, 27% of the Americans chose to live by themselves. In 1985, the average number of "close friends" was 2.94, and only 10% reported that they had no good friends to chat with. However, by 2004, the average number of "close friends" was 2.08, and people who had no friends to chat with had already accounted 25%. The largest change in the past 20 years was the emergence of the Internet. Although the Internet era provides more and more social methods, the relationship between people seems to be more and more distant. We don't want to be intimate with people, and even invite people to visit our home.

In order to get a better understanding of the relationship between the real social level and virtual sociality, this report interpret the number of friends on Facebook, which is the most popular social application around the world, based on the social level. In this study, 12 students (6 males and 6 females) were investigated for their social level and the number of friends on Facebook.

The dataset provides students' social skills in school, the number of Facebook friends (in hundreds), and the gender of the students.

2. Method

2.1 Data Description

Since the data are classified according to gender, we need to understand the difference in the distribution of number of their friends on Facebook. Therefore, a density curve was plotted firstly, which could also tell whether Friends is the normal distributed.

Since gender is qualitative data, it needs to be converted into factor (using factor()), and then added into the prediction model. In order to build the model more accurately, using the scatterplot to show the relationship between the independent variable (social level) and the dependent variable (number of Facebook friends). However, Figure does not show whether males and females have interaction terms, because the value of male and female social does not coincide. Therefore, the prediction and response variables were exchanged for redrawing the scatter plots. At this point, the X axis (Friends) of the two genders coincides with each other, so the two fitting lines can be observed to determine whether there is an interaction.

2.2 Variable Selection and Model Fitting

The full model was built using lm (), and the model's statistics were calculated with summary (). Then I used update () to gradually drop the variable that had less significance (larger p value) and get the final model. In order to check the result, all subsets regression was conducted using regsubsets ()

2.3 Regression Diagnosis

Finally, the stepwise regression method was used for linear regression, and the plot () and durbinWatsonTest () function was executed for regression diagnosis.

3. Result

Figure 1 shows the distribution of the number of friends on Facebook of males and females, which are both approximately normal distributed.

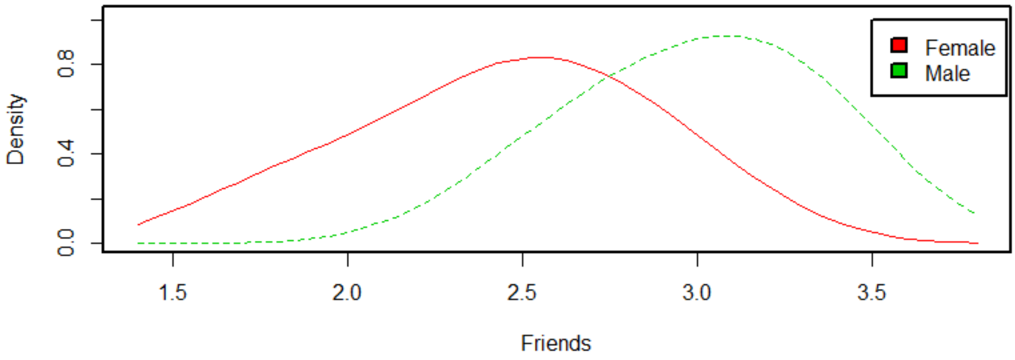


Figure 1 Distribution of Friends on Facebook of Females and Males

The relationship between the independent variable (social level) and the dependent variable (number of Facebook friends) is displayed as follows.

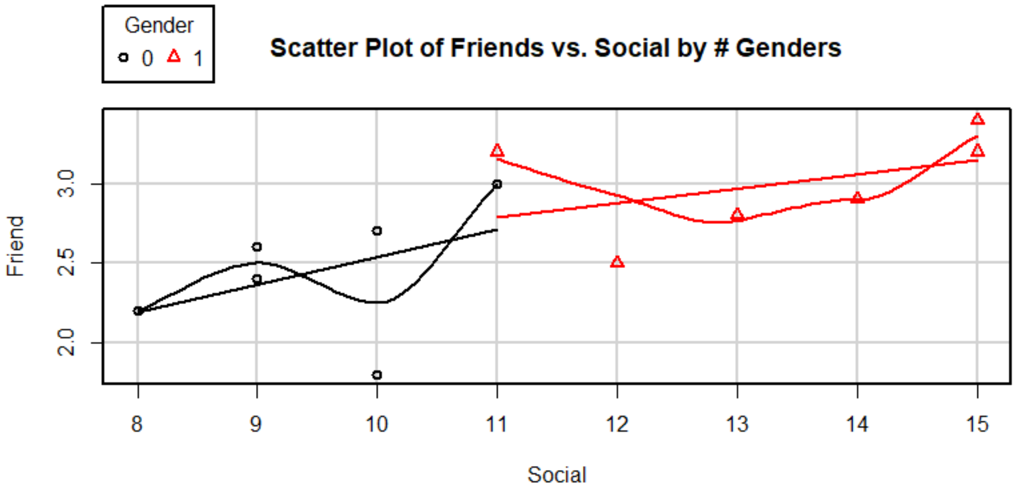


Figure 2 Relationship between Social Level and Number of Friends on Facebook

After interchanging the predictive variable with the response variable, we got the relation plot figure as Figure 3.

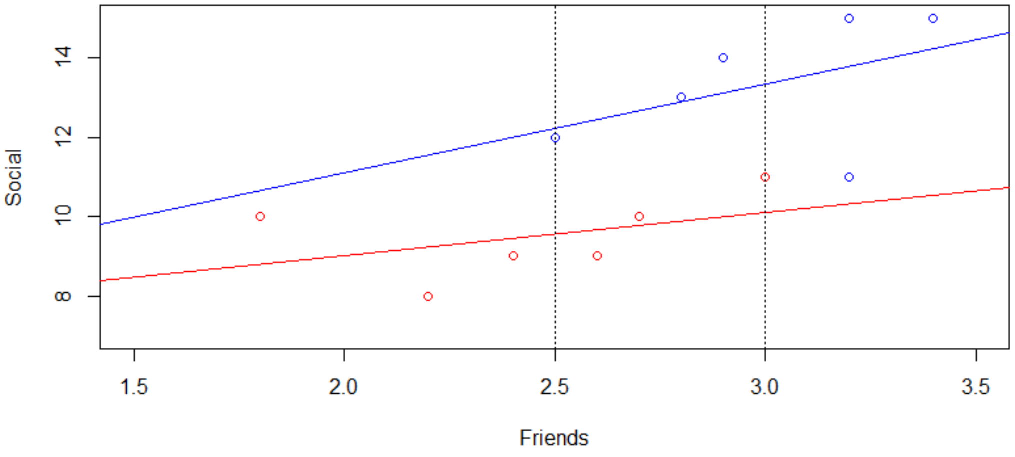


Figure 3 The scatter plot of Friends and Social

The summary statistics of the full model is shown in Table 1. In the full model, there was no predictor had a small p value that could pass the assessment.

Table 1 Summary Statistics of the Full Model

	Estimate	Standard Error	t value	Pr(>|t|)	assessment
Intercept	0.809	1.539	0.526	0.613	The significance of all items did not pass the evaluation.
Social	0.173	0.161	1.072	0.315	
GenderMale	0.991	2.073	0.478	0.645	
Social: GenderMale	-0.083	0.192	0.432	0.677	
Multiple R-squared: 0.5079	Adjusted R-squared:0.3234		p-value:0.1121		F-value: 2.275

The item with greatest p value (Social: Gender) was dropped, and got a model with statistics shown as follows. The assessment was still not fixed, however, R-Square improved significantly.

Table 2 Summary Statistics of Model 2

	Estimate	Standard Error	t value	Pr(>|t|)	assessment
Intercept	1.365	0.803	1.701	0.123	The significance of all items did not pass the evaluation.
Social	0.114	0.083	1.374	0.203	
GenderMale	0.112	0.380	0.295	0.771	
Multiple R-squared: 0.4964	Adjusted R-squared:0.3845		p-value:0.04563		F-value: 4.436

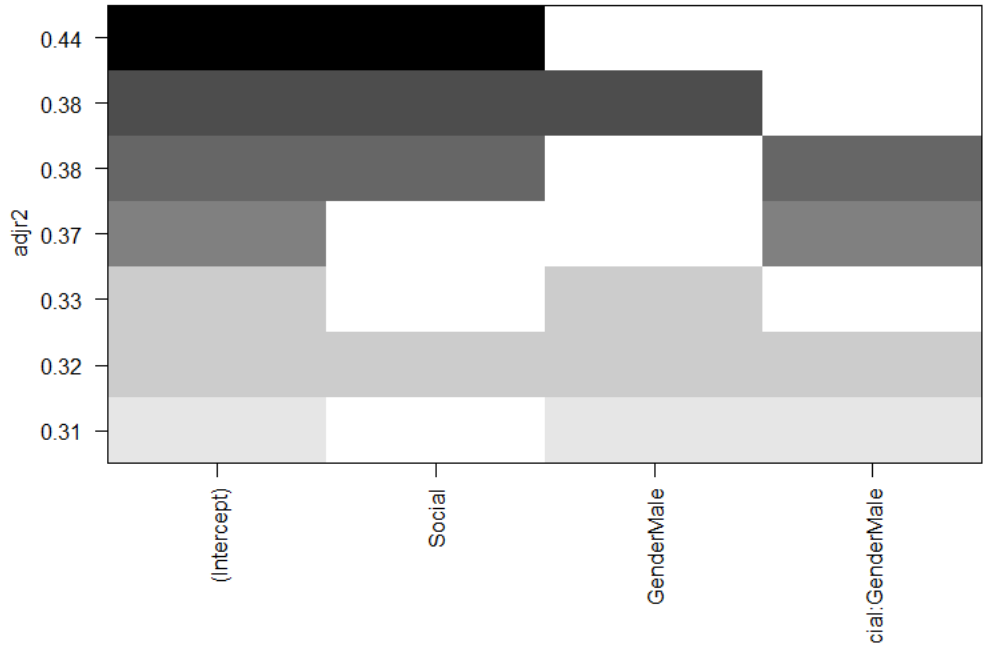
After Gender was dropped, a relatively good model was got. The model and its statistics are shown as follows.

Table 3 Summary Statistics of Model 2

	Estimate	Standard Erroe	t value	Pr(>|t|)	assessment
Intercept	1.187	0.504	2.353	0.0404	*
Social	0.135	0.043	3.109	0.0111	*
Multiple R-squared: 0.4915		Adjusted R-squared:0.4407	p-value:0.01108		F-value: 9.667

Friends = 1.187 + 0.135 * Social (Model)

The result of all subsets regression is shown in Figure 4, which shows that Social is significant for the best model as what the model tells.



After modelling, regression diagnostics was executed and the result was shown in Figure 4.

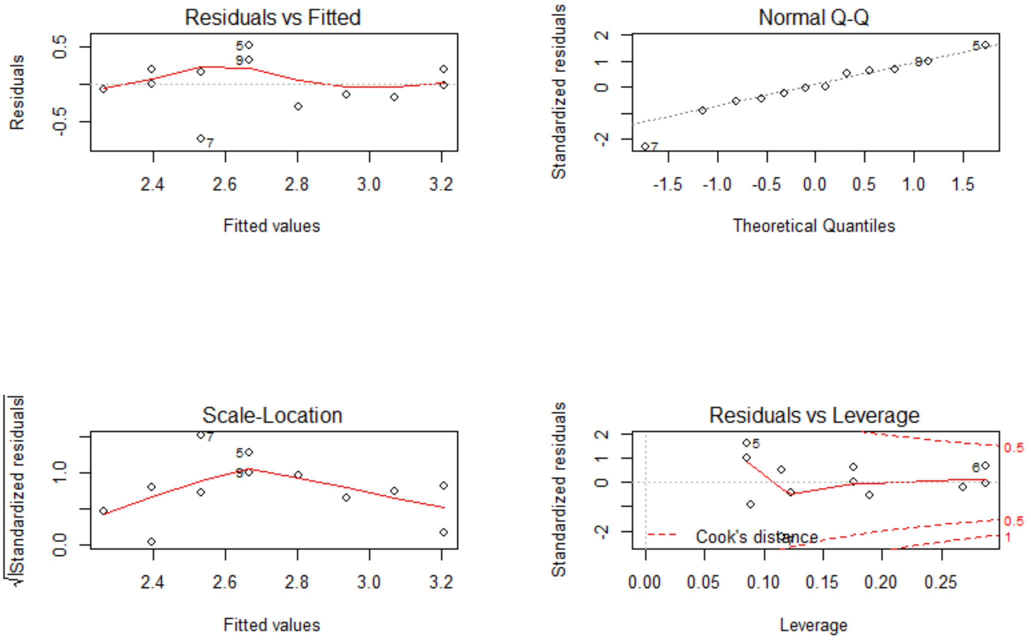


Figure 4 Regression diagnostics of Number of Friends on Facebook

The mutual independence between the residuals was done by durbinWatsonTest () function (using R), and resulting that p value was greater than 0.05 (p=0.876), which indicated that the hypothesis was reasonable.

4. Discussion

· What can be seen from Figure 2 is that the curve could fit the points well, but as the amount of data is too small, the addition of the square term may cause overfitting. Therefore, the square term was not taken into account.

In Figure 3, the space between the dotted line is the overlap. According to the Figure 3, the two straight lines have little slope gap in the range of the data given, thus there is no obvious interaction. This inference is confirmed in the process of variable selection.

However, in Figure 1 and figure 3, the number of friends on Facebook of female is generally more than that of male, which could not support that Gender was dropped in variable selection. This situation may be due to a significant gap between the Social level of male and female in the sample. As shown in Figurev2, there is no comparison when using Social level as predictor.

· The final model just includes the social level. This means that based on this sample, the relationship between the number of friends on Facebook and the social level has no difference with males and females. The number of friends on Facebook increases with the improvement of social level. According to scientific research, this phenomenon could be explained as that there will be higher social need when the human brain has larger capacity. The social need of the students in this sample are manifested both in the real life and the virtual network. It does not affect their performance on the other social platform because of an excessive dependence on one kind of social method.

· Since the amount of data used for modelling is too small, it is impossible to deal with outliers and evaluate the predictability of the model, which would have a certain impact on the model fitting. At the same time, the R-Square of the model could just indicate a moderate relationship between social level and the number of Facebook friends.

Conclusion

The final model is,

$$\textbf{Friends} = \mathbf{1.187 + 0.135 * Social}$$

The model is a linear model with 44.07% R-Square. Therefore, in the selected sample, students' performance in real and on Facebook is consistent. That is to say, to a certain extent, we can estimate the social level on internet through the actual social level. However, because of the limited number of observations in the sample, investigations and analysis are still needed for precise results.

APPENDIX 1 R CODE USED IN THE REPORT

R CODE FOR QUESTION 1

```
#R CODE FOR QUESTION1
bodyfat.data<-read.table('D:/2018 first semester/linear regression/ASSIGNMENT1/bodyfat.txt', header = TRUE)
dim(bodyfat.data)
head(bodyfat.data)
summary(bodyfat.data)
h<-hist(bodyfat.data_all$BodyFat,breaks=20,xlab='bodyfat distribution')
xfit<seq(min(bodyfat.data_all$BodyFat),max(bodyfat.data_all$BodyFat),length=40)
yfit<-dnorm(xfit,mean=mean(bodyfat.data_all$BodyFat),sd=sd(bodyfat.data_all$BodyFat))
yfit<-yfit*diff(h$mid[1:2])*length(bodyfat.data_all$BodyFat)
lines(xfit,yfit,col='blue',lwd=2)
box()
boxplot(bodyfat.data$BodyFat,main='Boxplot of Bodyfat')
match(0,bodyfat.data[,2])
bodyfat.data<-bodyfat.data[-182,]
match(148.10,bodyfat.data[,8])
bodyfat.data[39,]
bodyfat.data<-bodyfat.data[-39,]
attach(bodyfat.data)
library(GGally)
ggpairs(bodyfat.data[,2:length(bodyfat.data)])
lm_all<-lm(BodyFat~Age+Weight+Height+Neck+Chest+Abdomen+Hip+Thigh+Knee+Ankle+Biceps+Forearm+Wrist)
lm.step<-step(lm_all,direction = 'backward')
summary(lm.step)
lm.step1<-update(lm.step,~.-Neck)
summary(lm.step1)
lm.step2<-update(lm.step1,~.-Forearm)
summary(lm.step2)
leaps <-regsubsets(BodyFat~Age+Weight+Height+Neck+Chest+Abdomen+Hip+Thigh+Knee+Ankle+Biceps+Forearm+Wrist,data=bodyfat.data,nbest=8)
plot(leaps,scale = "adjr2")
par(mfrow=c(2,2))
plot(lm.step2)
library(car)
```

```
crPlots(lm.step2)
qqPlot(lm.step2,labels = row.names(bodyfat.data),id.method =
"identify",simulate = TRUE,main = "Q-Q Plot")
lm.step3<-update(lm.step2,~.-Chest)
summary(lm.step3)
influencePlot(lm.step3,id.method = "identity", main="Influence
Plot",sub="Circle size is proportional to Cook's distance")
bodyfat.data<-bodyfat.data[-248,]
bodyfat.data<-bodyfat.data[-222,]
bodyfat.data<-bodyfat.data[-40,]
detach(bodyfat.data)
attach(bodyfat.data)
lm_all<-
lm(BodyFat~Age+Weight+Height+Neck+Chest+Abdomen+Hip+Thigh+Knee+Ankle+Bice
ps+Forearm+Wrist)
lm.step<-step(lm_all,direction = 'backward')
summary(lm.step)
lm.step1<-update(lm.step,~.-Neck)
summary(lm.step1)
lm.step2<-update(lm.step1,~.-Forearm)
summary(lm.step2)
lm.step3<-update(lm.step2,~.-Chest)
summary(lm.step3)
influencePlot(lm.step3,id.method = "identity", main="Influence
Plot",sub="Circle size is proportional Influence")
bodyfat.data<-bodyfat.data[-234,]
bodyfat.data<-bodyfat.data[-221,]
bodyfat.data<-bodyfat.data[-213,]
detach(bodyfat.data)
attach(bodyfat.data)
lm_all<-
lm(BodyFat~Age+Weight+Height+Neck+Chest+Abdomen+Hip+Thigh+Knee+Ankle+Bice
ps+Forearm+Wrist)
lm.step<-step(lm_all,direction = 'backward')
summary(lm.step)
lm.step1<-update(lm.step,~.-Neck)
summary(lm.step1)
lm.step3<-update(lm.step,~.-Neck,-Forearm,-Chest)
summary(lm.step3)
lm.step3<-update(lm.step,~.-Neck-Forearm-Chest)
summary(lm.step3)
par(mfrow=c(1,1))
```

```
influencePlot(lm.step3,id.method = "identity", main="Influence
Plot",sub="Circle size is proportional Influence")
qqPlot(lm.step3,labels = row.names(bodyfat.data),id.method =
"identify",simulate = TRUE,main = "Q-Q Plot")
par(mfrow=c(1,3))
plot(lm.step3,which = 1:3)
durbinWatsonTest(lm.step3)
vif(lm.step3)
detach(bodyfat.data)
bodyfat.data1<-read.table('D:/2018 first semester/linear
regression/ASSIGNMENT1/bodyfat.txt', header = TRUE)
attach(bodyfat.data1)
Chest1<-scale(Chest,center=T,scale=F)
Hip1<-scale(Hip,center=T,scale=F)
Abdomen1<-scale(Abdomen,center=T,scale=F)
head(bodyfat.data1)
bodyfat.data1<-bodyfat.data1
bodyfat.data1[,7]<-Chest1
bodyfat.data1[,8]<-Abdomen1
bodyfat.data1[,9]<-Hip1
detach(bodyfat.data1)
attach(bodyfat.data1)
lm.all.1<-
lm(BodyFat~Age+Weight+Height+Neck+Chest+Abdomen+Hip+Thigh+Knee+Ankle+Bice
ps+Forearm+Wrist+Chest:Abdomen)
lm.step.1<-step(lm.all.1,direction = 'backward')
summary(lm.step.1)
lm.step.2<-update(lm.step.1,~.-Forearm-Neck)
summary(lm.step.2)
lm.step.3<-update(lm.step.2,~.-Age)
summary(lm.step.3)
par(mfrow=c(1,1))
qqPlot(lm.step.3,labels = row.names(bodyfat.data),id.method =
"identify",simulate = TRUE,main = "Q-Q Plot")
durbinWatsonTest(lm.step.3)
vif(lm.step.3)
AIC(lm.step3,lm.step.3)
shrinkage <- function(fit, k = 10) {
  require(bootstrap)
  # define functions
  theta.fit <- function(x, y) {
    lsfit(x, y)
  }
  theta.predict <- function(fit, x) {
    cbind(1, x) %*% fit$coef
  } # matrix of predictors
  x <- fit$model[, 2:ncol(fit$model)] # vector of predicted values
  y <- fit$model[, 1]
  results <- crossval(x, y, theta.fit, theta.predict, ngroup = k)
  r2 <- cor(y, fit$fitted.values)^2
  r2cv <- cor(y, results$cv.fit)^2
  cat("Original R-square =", r2, "\n")
  cat(k, "Fold Cross-Validated R-square =", r2cv, "\n")
  cat("Change =", r2 - r2cv, "\n")
}
shrinkage(lm.step3)
shrinkage(lm.step.3)
```

R CODE FOR QUESTION 2

```
# R CODE FOR QUESTION 2
facebook<-read.table('D:/2018 first semester/linear
regression/ASSIGNMENT1/facebook.txt',header = TRUE)
GenderX<-factor(facebook$Gender,levels=c(0,1),labels=c('Female','Male'))
facebook[,2]<-GenderX
attach(facebook)
library(car)
library(sm)
par(mfrow=c(1,1))
sm.density.compare(Friends,Gender, xlab='Friends')
colfill<-c(2:(1+length(levels(Gender))))
legend(3.5,1,levels(Gender),fill=colfill)
scatterplot(Friends~Social|Gender,lwd=2,main='Scatter Plot of Friends vs.
Social by #
Genders',xlab='Social',ylab='Friend',legend.plot=TRUE,id.method='identify
')
fit1<-lm(Social[1:6]~Friends[1:6])
fit2<-lm(Social[7:12]~Friends[7:12])
plot(Social[1:6]~Friends[1:6],xlab='Friends',ylab='Social',xlim=c(1.5,3.5
),ylim=c(7,15),col='blue')
points(Social[7:12]~Friends[7:12],col='red')
abline(fit1,col='blue')
abline(fit2,col='red')
abline(v=2.5,lty=3,h=y)
abline(v=3,lty=3,h=y)
fit.all<-lm(Friends~Social+Gender+Social:Gender)
summary(fit.all)
```

```
fit.1<-update(fit.all,~.-Social:Gender)
summary(fit.1)
fit.2<-update(fit.1,~.-Gender)
summary(fit.2)
library(leaps)
leapsfacebook<-
regsubsets(Friends~Social+Gender+Social:Gender,data=facebook,nbest = 3)
plot(leapsfacebook,scale='adjr2')
par(mfrow=c(2,2))
plot(fit.2)
durbinWatsonTest(fit.2)
```

APPENDIX 2 FIGURES

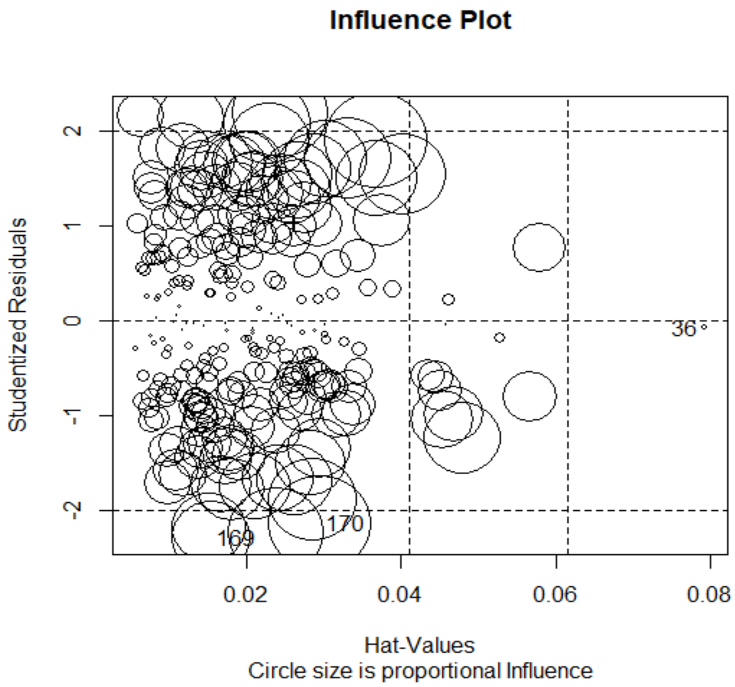


Figure 1 Influence Plot of Model 1

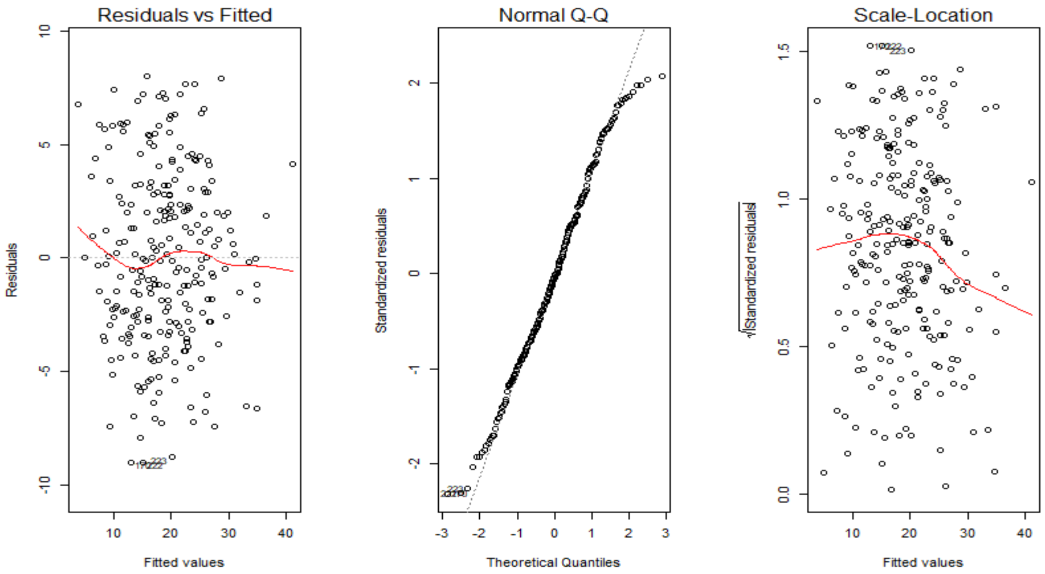


Figure 2 Regression diagnostics of percent bodyfat (Model 2)

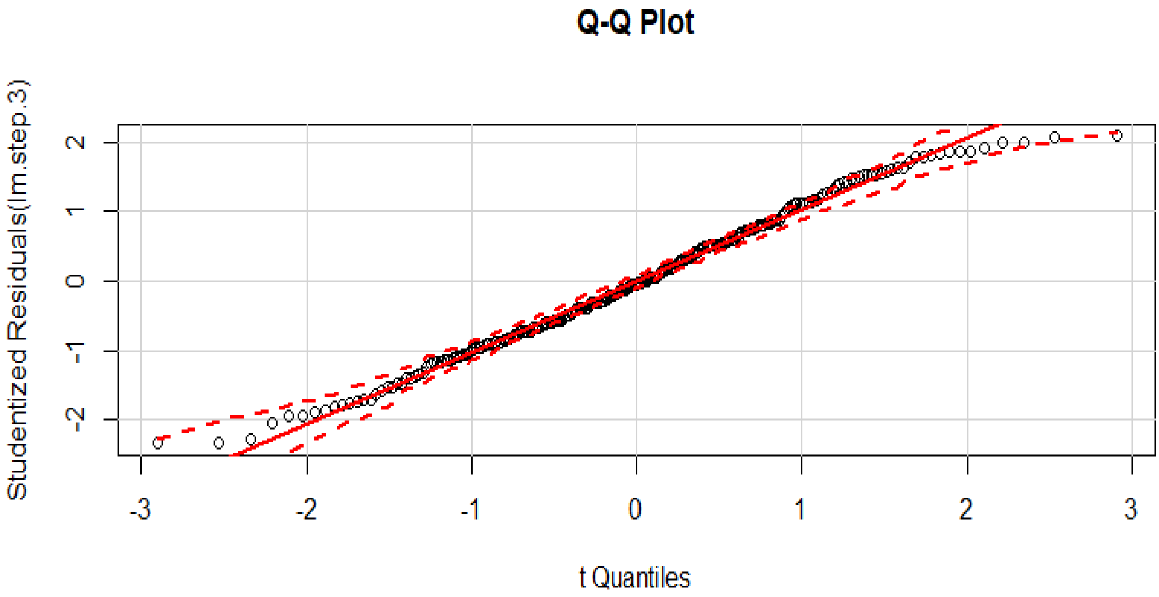


Figure 3 Q-Q Plot for Model2

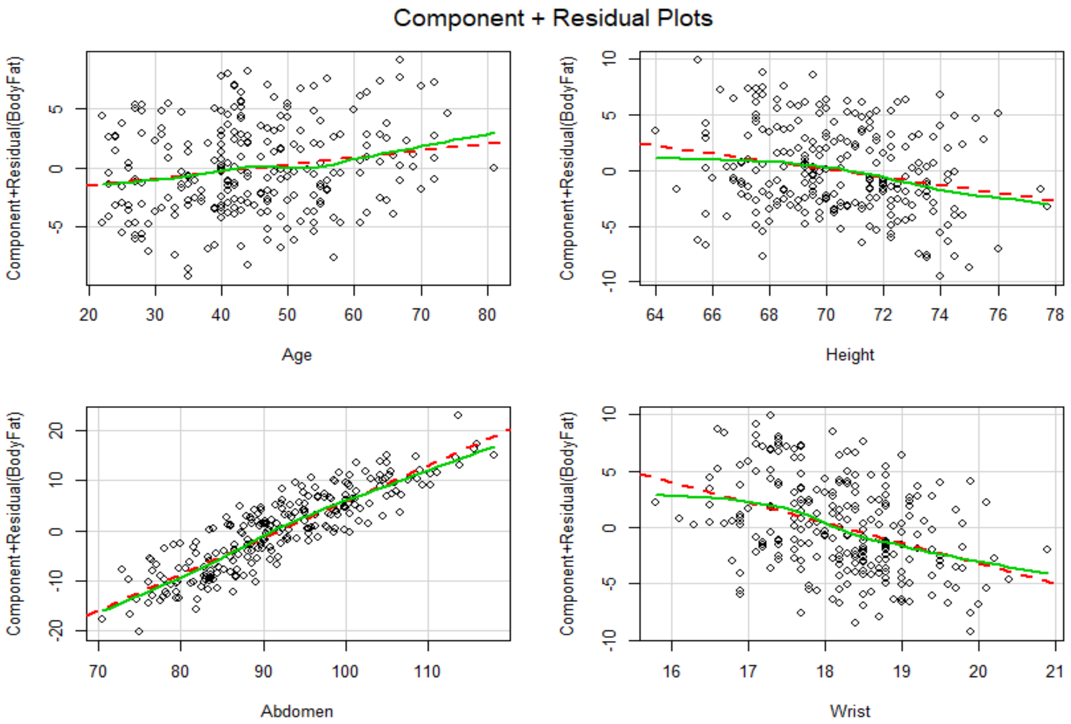


Figure 4 Residual Plot for each predictor