Body Fat Prediction, STAT 628 (Module 2)

Authors (Thurs_Group3): Zheng Ni, Jingpeng Weizhou, Zifeng Wang, Ke Chen

1. Introduction

As a measurement of obesity, Body Fat Percentage is a crucial index for describing health condition. In fact, many reserachers have proposed various accurate way to calculate body fat percentage, while these methods always require costly measurement. In hence, we would like to construct a simple but also precise "rule of thumb" method to predict body fat percentage of males using available clinical measurements.

In this project, we try with multiple linear regression models with different subset of features and finally choose the best model with highest accuracy and robustness.

2. Data Description

The dataset contains 252 men with measurements of their percentage of body fat and various body circumference measurements. Overall, the response variable is **BODYFAT** and there are 16 explained variables including **AGE,WEIGHT,HEIGHT**, etc. The following table shows the structure of our dataset.

In [23]:
```
data = read.csv('BodyFat.csv',header = TRUE);data[1:3,]
```

IDNO	BODYFAT	DENSITY	AGE	WEIGHT	HEIGHT	ADIPOSITY	NECK	CHEST	ABDOMEN	HIP	THIGH	KNEE	ANKLE	BICEPS	FOREARM	WR
1	12.6	1.0708	23	154.25	67.75	23.7	36.2	93.1	85.2	94.5	59.0	37.3	21.9	32.0	27.4	1
2	6.9	1.0853	22	173.25	72.25	23.4	38.5	93.6	83.0	98.7	58.7	37.3	23.4	30.5	28.9	1
3	24.6	1.0414	22	154.00	66.25	24.7	34.0	95.8	87.9	99.2	59.6	38.9	24.0	28.8	25.2	1

Based on previous research, we know that **DENSITY** is almost perfectly correlated with **BODYFAT**, so we cannot view **DENSITY** as a reliable variable. The following table are the explained variables we would use in the our model.

AGE	WEIGHT	HEIGHT	ADIPOSITY	NECK	CHEST	ABDOMEN	HIP	THIHG	KNEE	ANKLE	BICEPS	FOREARM	WRIST
years	lbs	inches	bmi	cm	cm	cm	cm	cm	cm	cm	cm	cm	cm

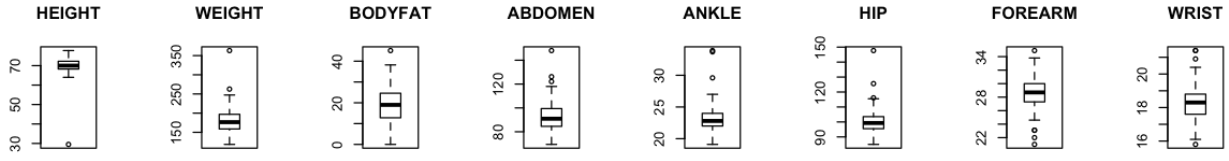
3. Data Preprocessing

Before we go into modeling part, we have to check whether there are missing values and outliers in our model. Based on the code below, we find there is no missing values in our dataset, so we don't need to use imputation method. As for the outliers, we would check the boxplot first to see if there are some data beyond 75% quantile.

From the boxplots below, we find some people have extremely low height, some have extremely high weight and bodyfat and even one man has 0 bodyfat which are beyond the possible range. So we think these records are outliers.

In [17]:
```
options(repr.plot.width=10, repr.plot.height=2)
cat('Number of missing values:',sum(is.na(data)),'\n');par(mfrow=c(1,8))
boxplot(data$HEIGHT,main='HEIGHT');boxplot(data$WEIGHT,main='WEIGHT') ;boxplot(data$BODYFAT,main='BODYFAT')
boxplot(data$ABDOMEN,main='ABDOMEN');boxplot(data$ANKLE,main='ANKLE');boxplot(data$HIP,main='HIP')
boxplot(data$FOREARM,main='FOREARM');boxplot(data$WRIST,main='WRIST')
cat('Outliers:',unique(c(which(data$HEIGHT<40),which(data$WEIGHT>300),which(data$BODYFAT>40),which(data$ABDOMEN>120),
                         which(data$ANKLE>30),which(data$HIP>120),which(data$FOREARM>34),which(data$FOREARM<22),which(data$WRIST<16))))
```

```
Number of missing values: 0
Outliers: 42 39 216 41 31 86 159 175 226
```



After detecting the outliers, we try with the following methods to deal with the outliers.

1. Fisrtly, we use some BMI function(**BMI(ADIPOSITY) = WEIGHT/(HEIGHT^2) * 703**) and the relation between BODYFAT and DENSITY proposed by researchers (**BODYFAT = 495/DENSITY - 450**). These functions can help us to compute the value of outliers.
2. Secondly, we utilize regression imputation method. We regress the outlier feature on other related features and get the estimate of the real value of the outliers.
3. Finally, for those data which are consistent with the functions above, they might not due to inaccurate measurement. They may be some extreme values which are far from other data, we mainly choose to drop these records(39,41,216) in order to get more precise analysis.

In [18]:
```
# BY BMI FUNCTION
data[,'HEIGHT'][42] = sqrt(205*703/29.9);data[,'ADIPOSITY'][221] = 153.25/70.50^2*703
# Regression IMPUTATION
data[,'WRIST'][226] = 6.5+0.26*data[,'ANKLE'][226]+0.18*data[,'KNEE'][226]+0.02*data[,'AGE'][226] # lm(WRIST~ANKLE+KNEE+AGE)
data[,'ANKLE'][31] = 3.1+0.56*data[,'WRIST'][31]+0.28*data[,'KNEE'][31]-0.02*data[,'AGE'][31] # lm(ANKLE~WRIST+KNEE+AGE))
data[,'ANKLE'][86] = 3.1+0.56*data[,'WRIST'][86]+0.28*data[,'KNEE'][86]-0.02*data[,'AGE'][86] # lm(ANKLE~WRIST+KNEE+AGE)
# DROP OUTLIERS
data = data[-c(39,41,216),]
```

4. Model Selection

After data cleaning, we find there are so many variables which may be correlation among them. In order to get a precise model, we would use stepwise method for feature selection. In order to find the best model, we try with both forward and backward selection using criterions as AIC,BIC,R2. These methods have some consistent results indicating **ABDOMEN, WEIGHT, WRIST** are the most important features. Here, we mainly shows the result of forward selection using BIC criterion.

In [10]:
```
X <- data[,4:17]; Y <- data$BODYFAT; min_model <- lm(Y~1,data = X);biggest <- formula(lm(Y~.,X))
bic_forward = step(min_model,direction = 'forward',scope = biggest,trace = 0,k=log(length(data[,1])));bic_forward
```

```
Call:
lm(formula = Y ~ ABDOMEN + WEIGHT + WRIST, data = X)

Coefficients:
(Intercept)      ABDOMEN       WEIGHT        WRIST
  -25.68435      0.88743     -0.08671     -1.20254
```

Since the **ABDOMEN** is the most significant one with smallest p-value, it must be the most important variable to be selected. For **WEIGHT** and **WRIST**, they are both statistically significant, however, based on the "Rule of Thumb" we had better drop one of them in order to make our model simple. To make better trade-off, we further compare these two models to decide which variable should be kept.

In [4]:
```
Model_1 <- lm(BODYFAT ~ ABDOMEN + WEIGHT,data);Model_2 <- lm(BODYFAT ~ ABDOMEN + WRIST,data)
cat('R2 of model with WEIGHT',summary(Model_1)$r.squared,',R2 of model with WRIST',summary(Model_2)$r.squared)
Model_1
```

```
R2 of model with WEIGHT 0.7102751 ,R2 of model with WRIST 0.7030952

Call:
lm(formula = BODYFAT ~ ABDOMEN + WEIGHT, data = data)

Coefficients:
(Intercept)      ABDOMEN       WEIGHT
   -43.1062       0.9005      -0.1187
```

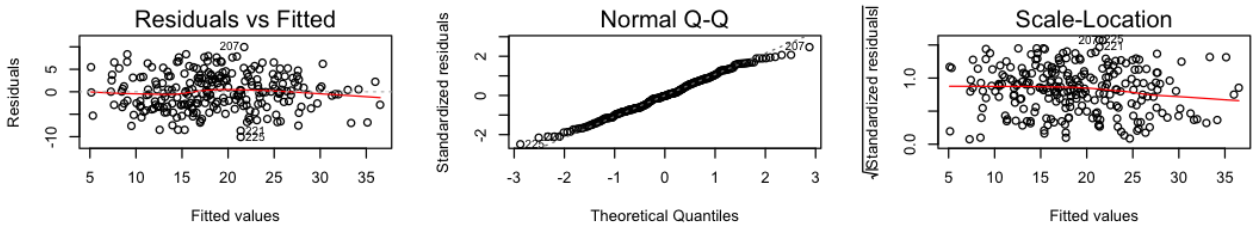
Comparing the models with **ABDOMEN,WEIGHT** and **ABDOMEN,WRIST**, we can see the R-squared for the model with **WEIGHT** is higher which suggests it has better performance. In hence, we determine our final model in the form of linear regression with variables **ABDOMEN** and **WEIGHT**.

5. Model Diagnostics

In this part, we will first check the assumptions for the linear model that we used in our model. We know that there are 3 assumptions for the linear model.

Linearity of the data Homogeneity of variance Normality of residuals Independence of residuals error terms To check these assumptions, we will use diagnostic plots like the following. And we will analysis all the plots one by one.

```
In [22]: options(repr.plot.width=9, repr.plot.height=2);par(mfrow = c(1,3));plot(Model_1,1:3)
```



1. From the first plot, we can see that the residual plot shows nearly no fitted pattern. The red line is very close to the zero line. That is to say the **linearity** of the data and **independence** of the reisduals are both checked in our model.
2. From the second plot, all the points fall approximately close to this reference line, which means the **normality** of residuals are checked.
3. From the third plot, we can see that the residuals are spread in the same range all along the fitted values. We can assume that the residuals share the same variance in our model. That is to say the **homogeneity** of variance is checked.

Based on Cook's Distance and Leverage Effect, we also find there are 3 potential influential points in our model. From the data, we can see that the points are all among the old who are all about 70 years old. That is to say when people get old, their body may easily get fat than the young people. We can see that they all have the weight near the boundary, and huge body shape.

That is to say the results for our model are much accurater when applied to the young people who is younger than 60 years old.

6. Conclusion

Final model: $BODY\_FAT = 0.9005 \cdot ABDOMEN - 0.1187 \cdot WEIGHT - 25.68435$

Rule of Thumb: Summarily, we only need the abdomen and weight to calculate the approximate body fat percentage, while in our model abdomen is positively correlated with body fat and weight is negative. These two can be the two general indices for the human body shape and weight. While two persons share the same body shape (abdomen), the one with higher body weight may have stronger body with lower body fat percentage. This can be similarly applied to the body weight.

Strength: The accuracy of our model can be guaranteed while at the same time, the model is simple enough and easy to understand.

Weakness:

- Accuracy of our model is very high only based on the people with normal body size. However, the people whose body is extremely out of shape, cannot be estimated by this model accurately.
- As what we discussed in the model diagnosis part, there are still some influential points in our data. Trying to keep the origin of the data, we remained them in our model, which means there are still some noises in calculation. To get more accurate results, we recommend to using the common body shape and weight. (Reference: 25<Age<65, 150<Weight<200, 20<Abdomen<30)

Contribution: Each member in our group contributes much to this project and we all participate slides design, report compiling. The table is our duty for this project.

Member	Contribution
Ke Chen	Summary statistics, data preprocessing
Jingpeng Weizhou	Feature selection, model comparison
Zheng Ni	Model diagnostics, image files
Zifeng Wang	Shiny app design, model interpretation

Reference:

1. Bailey, Covert (1994). *Smart Exercise: Burning Fat, Getting Fit*, Houghton-Mifflin Co., Boston, pp. 179-186.
2. Behnke, A.R. and Wilmore, J.H. (1974). *Evaluation and Regulation of Body Build and Composition*, Prentice-Hall, Englewood Cliffs, N.J.
3. Siri, W.E. (1956), "Gross composition of the body", in *Advances in Biological and Medical Physics*, vol. IV, edited by J.H. Lawrence and C.A. Tobias, Academic Press, Inc., New York.
4. Katch, Frank and McArdle, William (1977). *Nutrition, Weight Control, and Exercise*, Houghton Mifflin Co., Boston. 5. Wilmore, Jack (1976). _Athletic Training and Physical Fitness: Physiological Principles of the Conditioning
5. https://en.wikipedia.org/wiki/Body_fat_percentage (https://en.wikipedia.org/wiki/Body_fat_percentage)
6. https://en.wikipedia.org/wiki/Body_mass_index (https://en.wikipedia.org/wiki/Body_mass_index)