# Report

**Yilun Chen, Huiyu Jiang, Jiyun Chen**

# [Introduction]

## 1. Motivation

In modern society, people begin to pursue healthier lifestyles and focus on personal health. As a measure of obesity, body fat percentage plays an important role in various health outcomes.Our project aims at building a model a simple, robust, accurate and precise“rule-of-thumb” method to estimate percentage of body fat based on the 252-men data set, meaning that our model only predicts the male.

Thesis statement: Using a linear model to infer and predict the male bodyfat based on three factors.

## 2. Background of body fat and the data set

### [About body fat]

From the achievement of previous studies, we can get the formula for estimating the body fat B(%)

Percentage of Body Fat $(i.e. 100 \times B) = \frac{495}{D} - 450$, D = Body Density $(gm/cm^3)$ (1)

### [About the data set]

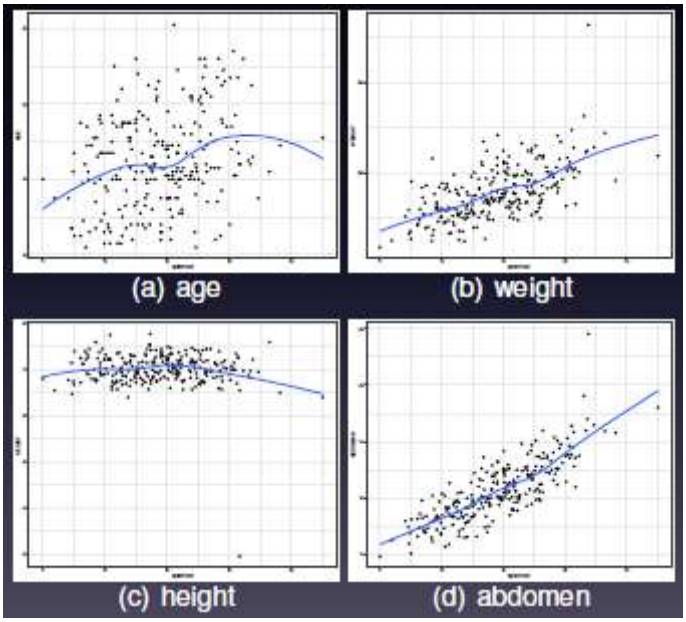
Our data is based on 252 men and has 14 independent variables:

Age(years), Weight(lbs), Height(inches), Adiposity(bmi), Neck circumference(cm), Chest circumference(cm), Abdomen circumference(cm), Hip circumference(cm), Thigh circumference(cm), Knee circumference(cm), Ankle circumference(cm), Biceps (extended) circumference(cm), Forearm circumference(cm), Wrist circumference(cm); one dependent variable: Body fat B(%); and Density, here, we think it is meaningless to our model.

| IDNO | BODYFAT | DENSITY | AGE | WEIGHT | HEIGHT | ADIPOSITY | NECK | CHEST | ABDOMEN | HIP | THIGH | KNEE | ANKLE | BICEPS | FOREARM | WRIST |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 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 | 17.1 |
| 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 | 18.2 |
| 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 | 16.6 |
| 4 | 10.9 | 1.0751 | 26 | 184.75 | 72.25 | 24.9 | 37.4 | 101.8 | 86.4 | 101.2 | 60.1 | 37.3 | 22.8 | 32.4 | 29.4 | 18.2 |
| 5 | 27.8 | 1.0340 | 24 | 184.25 | 71.25 | 25.6 | 34.4 | 97.3 | 100.0 | 101.9 | 63.2 | 42.2 | 24.0 | 32.2 | 27.7 | 17.7 |
| 6 | 20.6 | 1.0502 | 24 | 210.25 | 74.75 | 26.5 | 39.0 | 104.5 | 94.4 | 107.8 | 66.0 | 42.0 | 25.6 | 35.7 | 30.6 | 18.8 |

# [Data Analysis]

## 1. Model selection




From the graphs, we can find a linear relationship between BODYFAT and the rest variables(except HEIGHT); AGE has a weak linear relationship with BODYFAT and the other variables, like ABDOMEN and WEIGHT, have strong linear relationships with BODYFAT. Thus, we choose the linear regression model to estimate the body fat.

**Linear regression assumption**:1.Linear relationship 2.Multivariate normality 3.No or little multicollinearity 4.No auto-correlation 5.Homoscedasticity

## 2. Data preprocessing

### 2.1 Unreasonable record deletion

| | BODYFAT | AGE | WEIGHT | HEIGHT | ADIPOSITY | NECK | CHEST | ABDOMEN | HIP | THIGH | KNEE | ANKLE | BICEPS | FOREARM | WRIST |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **182** | 0 | 40 | 118.5 | 68 | 18.1 | 33.8 | 79.3 | 69.4 | 85 | 47.2 | 33.5 | 20.2 | 27.7 | 24.6 | 16.5 |

BODYFAT(182nd)=495/1.1089-450=-3.612

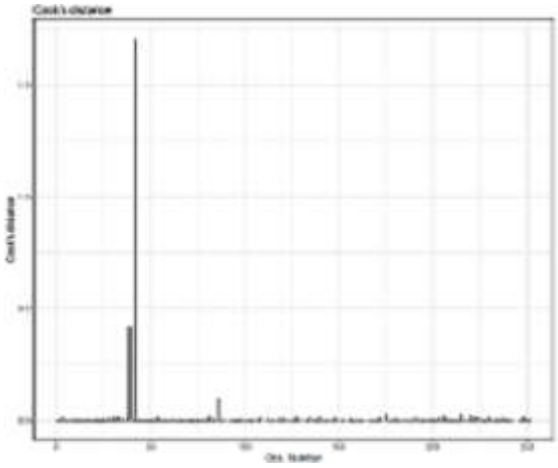
We notice that the 182nd has bodyfat=0;and we calculate the body fat with the density formula (1) and the result is negative. Thus, we regard this record as one unreasonable observarion and delete it.

### 2.2. Outlier detection

Each time we throw one sample based on the cook distance until the value of cook distance is not greater than 2p/n.

First time we drop the 42nd record,

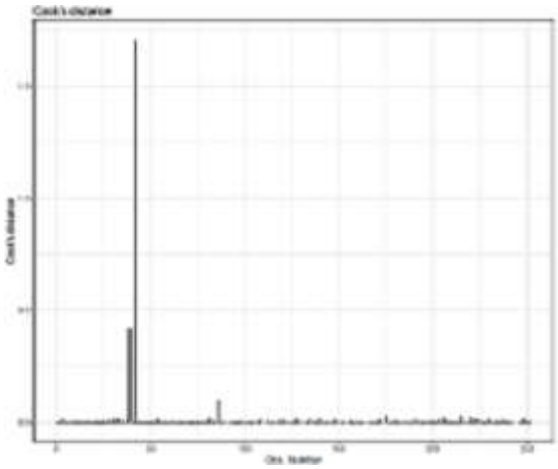
| | BODYFAT | AGE | WEIGHT | HEIGHT | ADIPOSITY | NECK | CHEST | ABDOMEN | HIP | THIGH | KNEE | ANKLE | BICEPS | FOREARM | WRIST |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **42** | 31.7 | 44 | 205 | 29.5 | 29.9 | 36.6 | 106 | 104.3 | 115.5 | 70.6 | 42.5 | 23.7 | 33.6 | 28.7 | 17.4 |



Interpretation: the 42nd observation has the biggest cook distance value. He is really short and the record seems to be fake, thus we delete this record.

Second time we drop the 39th record,

| | BODYFAT | AGE | WEIGHT | HEIGHT | ADIPOSITY | NECK | CHEST | ABDOMEN | HIP | THIGH | KNEE | ANKLE | BICEPS | FOREARM | WRIST |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **39** | 33.8 | 46 | 363.15 | 72.25 | 48.9 | 51.2 | 136.2 | 148.1 | 147.7 | 87.3 | 49.1 | 29.6 | 45 | 29 | 21.4 |



Interpretation: the 39th observation has the biggest cook distance value and he is extremely heavy, thus we delete this record. After droping the 42th and 39th sample, the data set looks good for us.

## 3. Variable selection

### 3.1 Stepwise method

#### *3.1.1 Using AIC to select variables.*

$$lnL(\mu, \sigma^2) = -\frac{n}{2}ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

AIC: $2k - 2ln(\hat{L})$

According to the AIC, we pick five variables include abdomen, weight, wrist, biceps, and age.We can get the model,

$(BodyFat\%) = -23.717 + 0.838Abdomen - 0.085Weight - 1.656Wrist + 0.039Age + 0.277Bicep$

Since the BIC is more conservative, we use the stepwise model based on BIC criterion.

#### *3.1.2 Using BIC to select variables.*

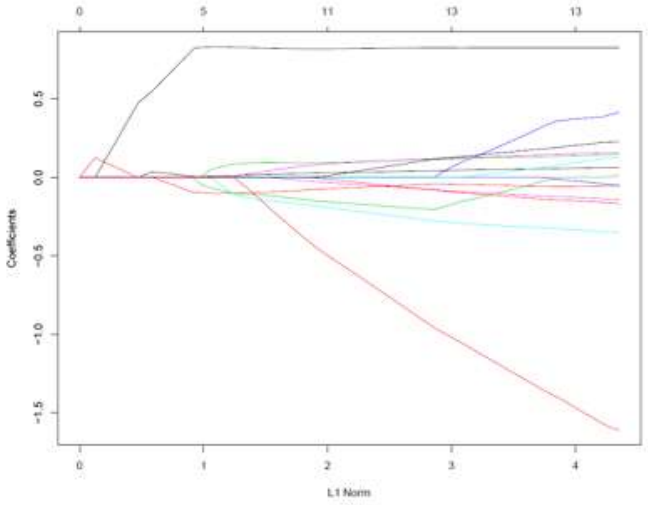
$$\text{BIC: } ln(n)k - 2ln(\hat{L})$$

The result of BIC shows abdomen, weight, and wrist are the most important variables for interpret the bodyfat.We can get the model,

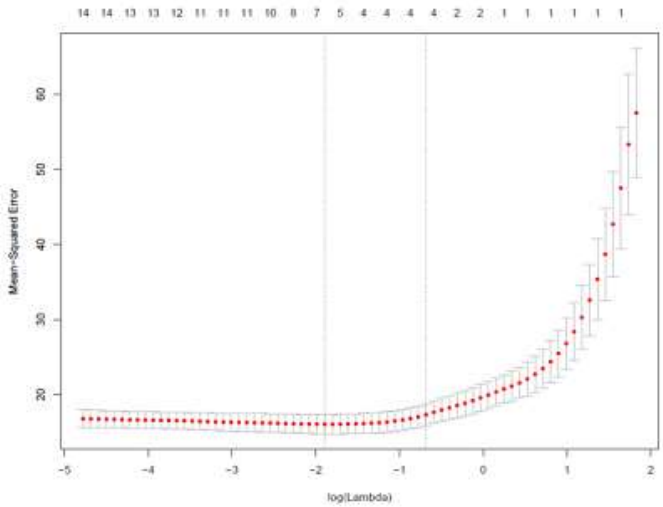
$$(BodyFat\%) = -23.994 + 0.885Abdomen - 0.087Weight - 1.282Wrist; \;\; MultipleR^2 : 0.7292$$

#### 3.2 Elastic net

$$\hat{\beta} = argmin_{\beta}(||y - X\beta||^2 + \lambda(1-\alpha)||\beta||^2 + \lambda\alpha||\beta||_1)$$




regularization path




Cross-validation

We use the grid search to determine the alpha based on the regularization path plot. Next, cross validation is adopted to choose the proper lambda. According to the graphs about the regularization path and Cross-validation, we can get the final model selects the age, height, adiposity, abdomen, and wrist. We can get the model,

$$(BodyFat\%) = 5.717 + 0.736Abdomen - 0.336Height - 1.692Wrist + 0.045Age - 0.091; \;\; Multi$$

#### 3.3 Which one to choose?

For the sake of simplicity, we choose the stepwise model based on BIC criterion as our final model. On the one hand, the performance of the regressions does not differ a lot between each model. To be specific, each model's R square is very close to the others'. Both AIC and elastic net pick too many variables. One the other, it is strange that some estimations in the regularization path of elastic net change the sign as the penalty grows. Thus we give up the elastic net model. As a result, we choose variables selected by BIC stepwise method.

## 4.Model fitting

Hypothesis $H_0 : \beta_{Weight}, \beta_{Abdomen}, \beta_{Wrist} = 0$

The final model is

$$(BodyFat\%) = -23.994 + 0.885Abdomen(cm) - 0.087Weight(lbs) - 1.282Wrist(cm)$$

For the sake of simplicity, we simplify our linear model as,

$$(BodyFat\%) = -24 + 0.9Abdomen(cm) - 0.1Weight(lbs) - 1.3Wrist(cm)$$

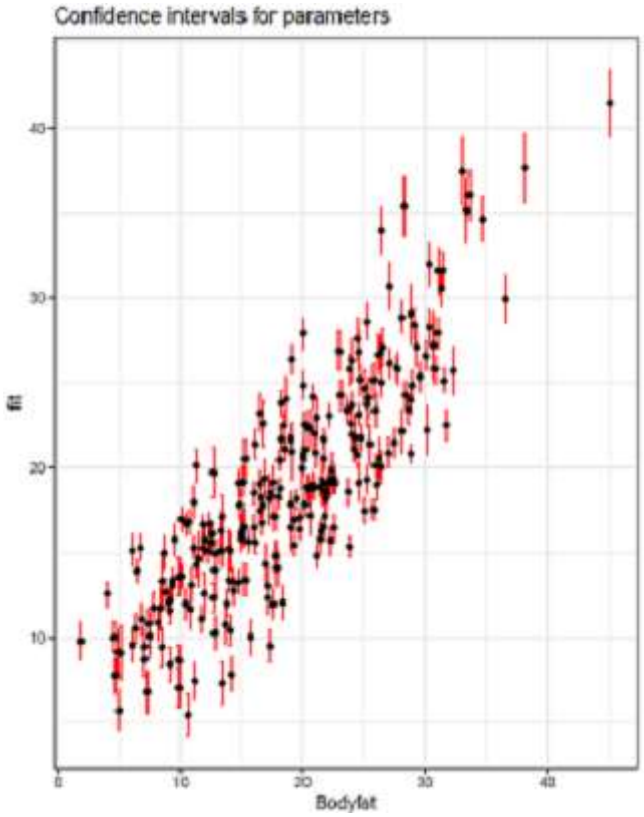
Layman's Interpretation:

e.g. If we fix the Weight and Wrist circumference, the Abdomen circumference increases 1 lbs, the body fat will increase 0.9%

$$Multiple\ R^2:\ 0.7292;\ Residual\ standard\ error:\ 3.981;\ p-value:\ < 2.2e-16$$

Layman's Interpretation:

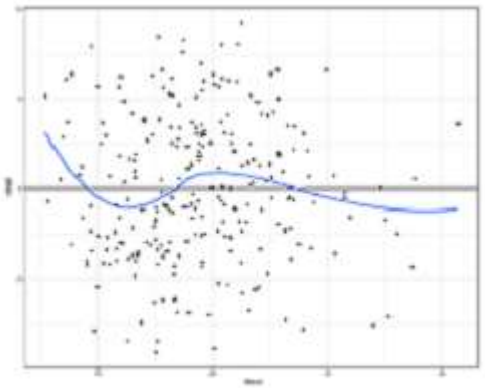
1. The residual standard error is 3.981. It's a measurament of how our predictions are different from original bodyfat records. So the result tells us that our model tends to return an estimation with 3.98 differnce from the original data.
2. The $R^2$ represents to what extend our model explains the dependent variable. To be specific, 0.7292 means the three variables we choose can explain about 73% information contained in original data.
3. The p-value means the probability of null hypothesis happening with given data. Since p-value is tiny, we draw the conclusion that null hypothesis is rejected. In other words, the coefficients of the variables we choose are significantly different from 0.
4. The 95% confidence interval is a region where we are 95% confident that the unknown bodyfat with given data will be between the lower bound and upper bound.

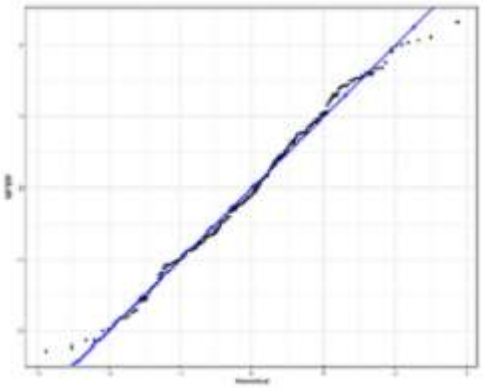



From this plot, we can see all obeservations' prediction values given by our model and their 95% confidence intervals. We can also see that our model fits well, the fitted values and the orginal values almost stay around the line y=x.

## 5.Diagnostic

| Variables | VIF |
|---|---|
| Weight | 5.6178 |
| Abdomen | 4.1857 |
| Wrist | 2.0988 |


Residual plot


QQ-plot

For our five assumptions,

1. Linear: We have shown before
2. Normality: The QQ plot indicates that the residuals generally follow the normal distribution. Although it has a thinner tail, we do not think it will effect our model performance.
3. No Multicollineanity: As shown in the variance inflation factor, all of them are below 5 which means there is not a multicollinearity in our model.
4. No auto-correlation and homoscedasticity: The residual is very similar to a white noise. Hence, we do not need to worry about the auto correlation and heteroscedasticity.

# [Conclusion]

## 1.Model

We can get our simple robust rule of thumb model,

$$(BodyFat\%) = -24 + 0.9Abdomen(cm) - 0.1Weight(lbs) - 1.3Wrist(cm)$$

## 2.Advantage and Disadvantage

**Advantage:** 1. Easy to implement; 2. Do not need a lot of information.

**Disadvantage:** Lack of accuracy (relatively large residuals)

## 3.Statements

1. Abdomen, weight and wrist can explain the badyfat of men partly.
2. When we fix the abdomen circumference and wrist circumference, the men with more weight will more likely to have lower bodyfat.
3. Linear model perform well in the bodyfat database.

## 4. Contribution

Yilun Chen: Build the prediction models; contribute to the report diagnostic and summary.

Jiyun Chen: Clear up the code and our images; contribute to the main body of the report.

Huiyu Jiang: Make the tidy ggplot images; contribute to the presentation slides and Layman's Interpretation part