Body Fat Assignment

Stat 628 Module 2

RUI HUANG

CHENYANG JIANG

HanGyu KANG

ENZE WANG

Body Fat Assignment

- Introduction
- Data Analysis and Data Clean
- Variables Selection
- Model Selection
- Model Diagnosis and Summary
- Shiny Application
- Acknowledge

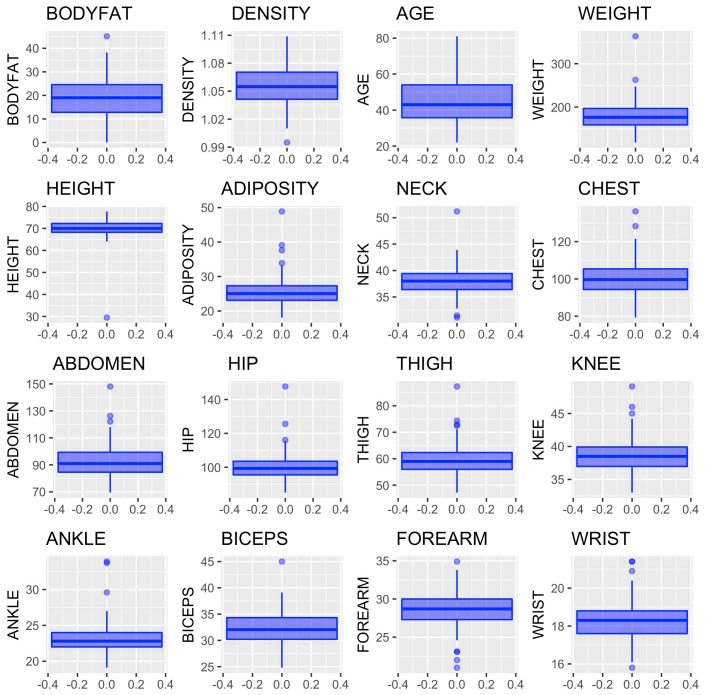
Introduction

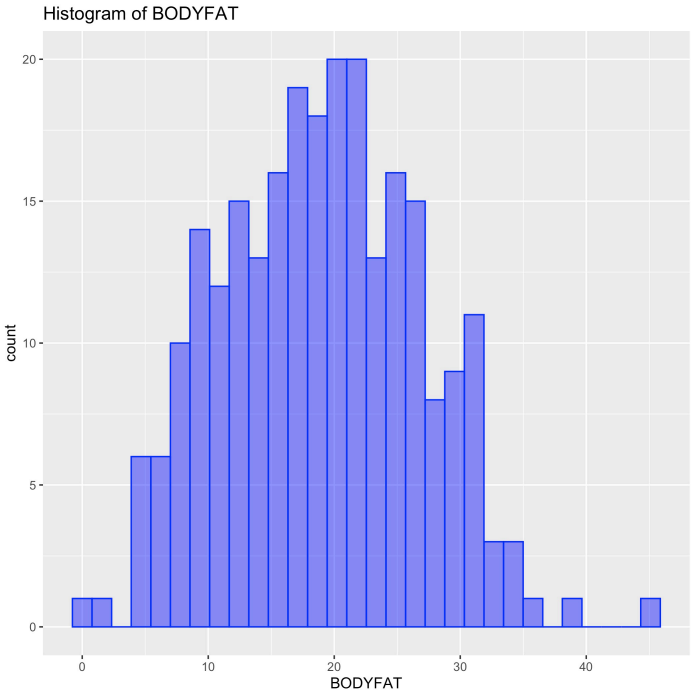
- Body Fat Percentage is the total mass of fat divided by total body mass.
- Essential body fat is necessary for healthy daily life, but large body fat is harmful to human health.
- A precise way to measure body fat is to combine submersion method with Siri equation (Body fat=495/Density-450). But it is difficult.
- Our groups want to use several human measurements including age, weight to predict body fat.

Data Analysis

- Data Visualization

1. Any outliers
2. Data distribution
3. Null and abnormal value

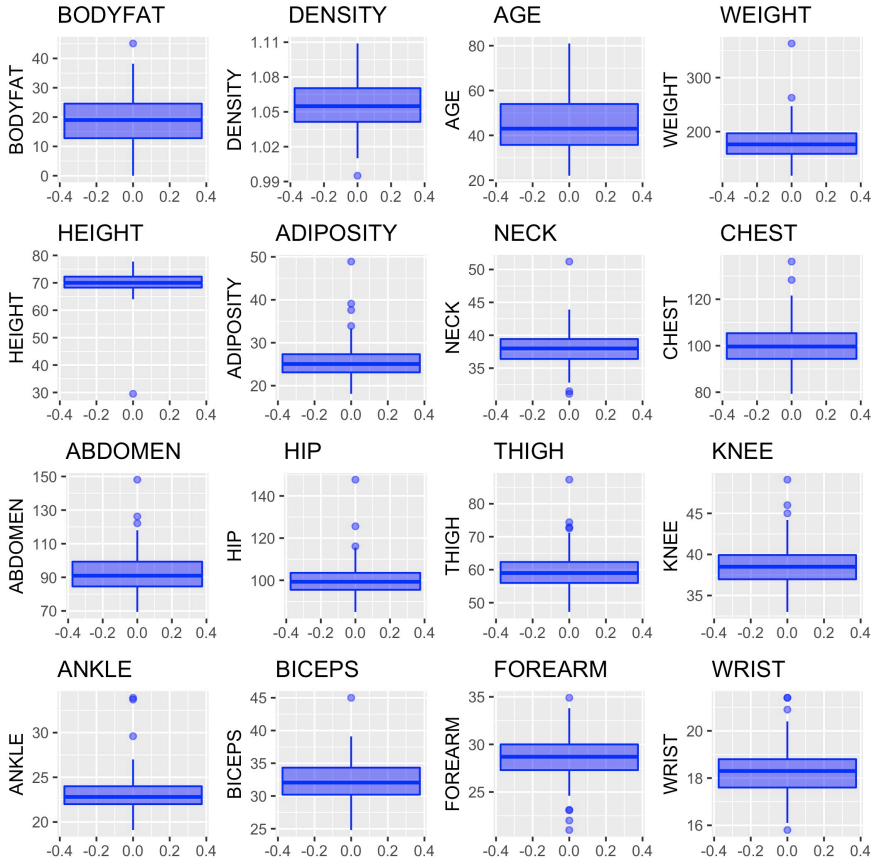




Data Clean - Boxplots

1. Method to outliers from boxplots
2. Suspicious outliers:

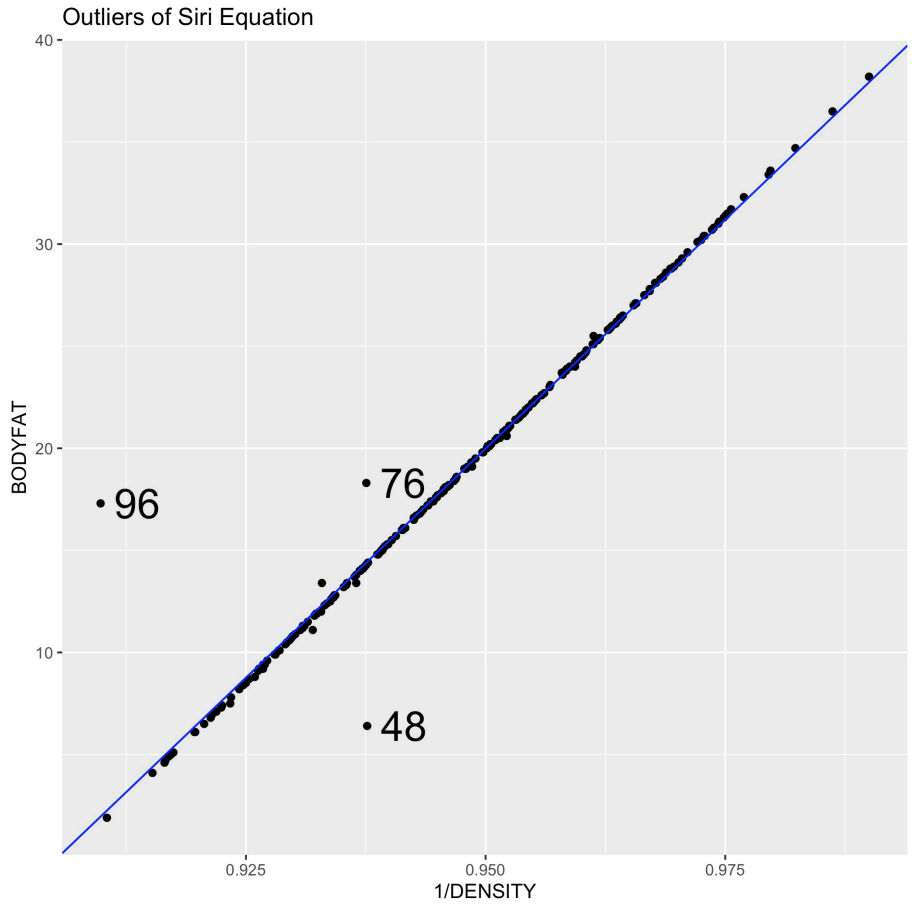
IDNO	Reason	Method
182	BODYFAT is 0	Remove
216	Extreme value in BODYFAT	Remove
39	Outliers from boxplots	Remove
42	Low HEIGHT	Fix by BMI equation
41	Outliers from boxplots	Remove



Data Clean – Siri Equation

1. Connection between BODYFAT and DENSITY in our data is called Siri equation:
2. Siri equation: BODYFAT=495/DENSITY-450
3. Outliers in regression between BODYFAT and 1/DENSITY

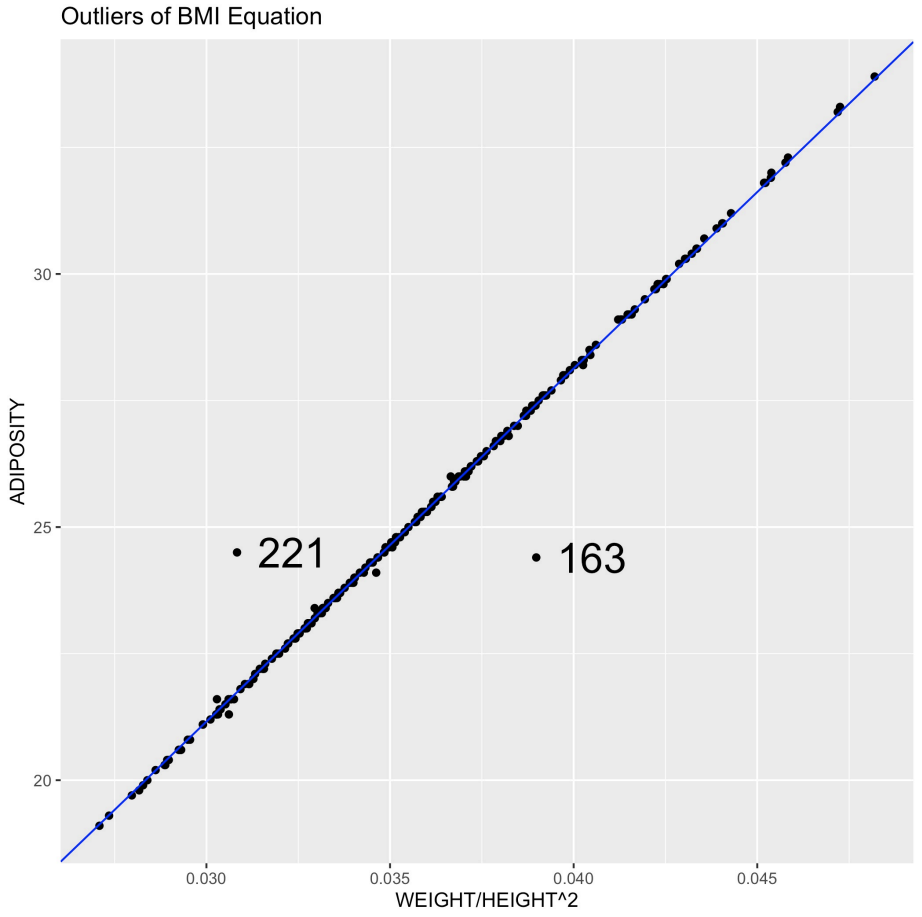
IDNO	Initial BODYFAT	New BODYFAT	Method
96	17.3	0.36	Update DENSITY: 1.06
76	18.3	14.09	Update BODYFAT: 14.09
48	6.4	14.13	Update BODYFAT: 14.13



Data Clean – BMI Equation

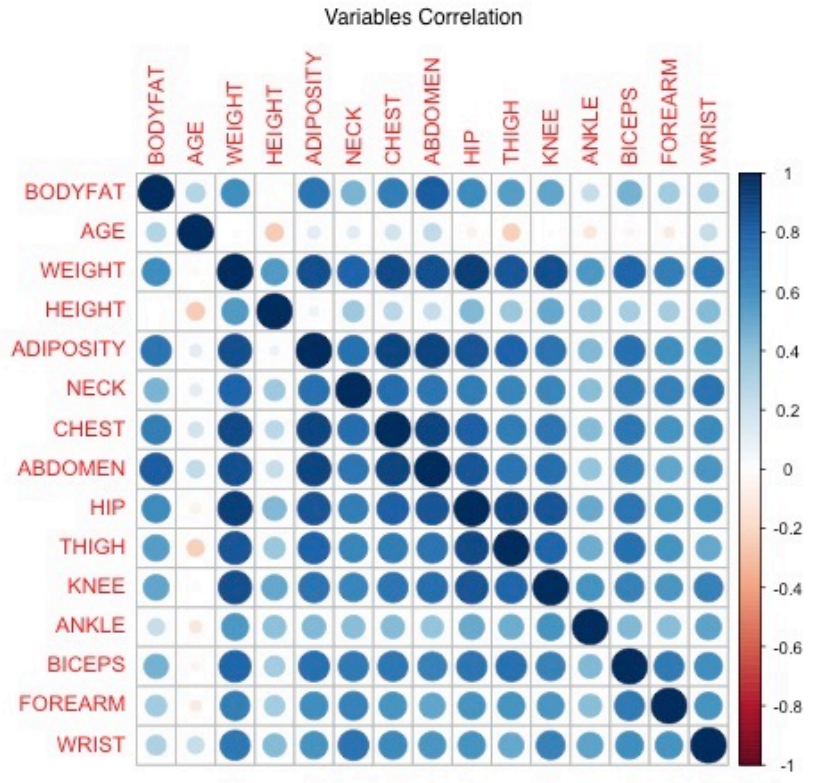
1. Connection among ADIPOSITY, WEIGHT and HEIGHT in our data is called BMI equation:
2. BMI equation: ADIPOSITY=703 WEIGHT/HEIGHT^2
3. Outliers in regression between ADIPOSITY and WEIGHT/HEIGHT^2

IDNO	Initial ADIPOSITY	New ADIPOSITY	Method
163	24.4	27.4	Update ADIPOSITY: 1.06
221	24.5	21.7	Update ADIPOSITY: 14.09



Clean Data - Summary

1. It is a trade off to keep or remove outliers.
2. There is high correlation among variables after clean.

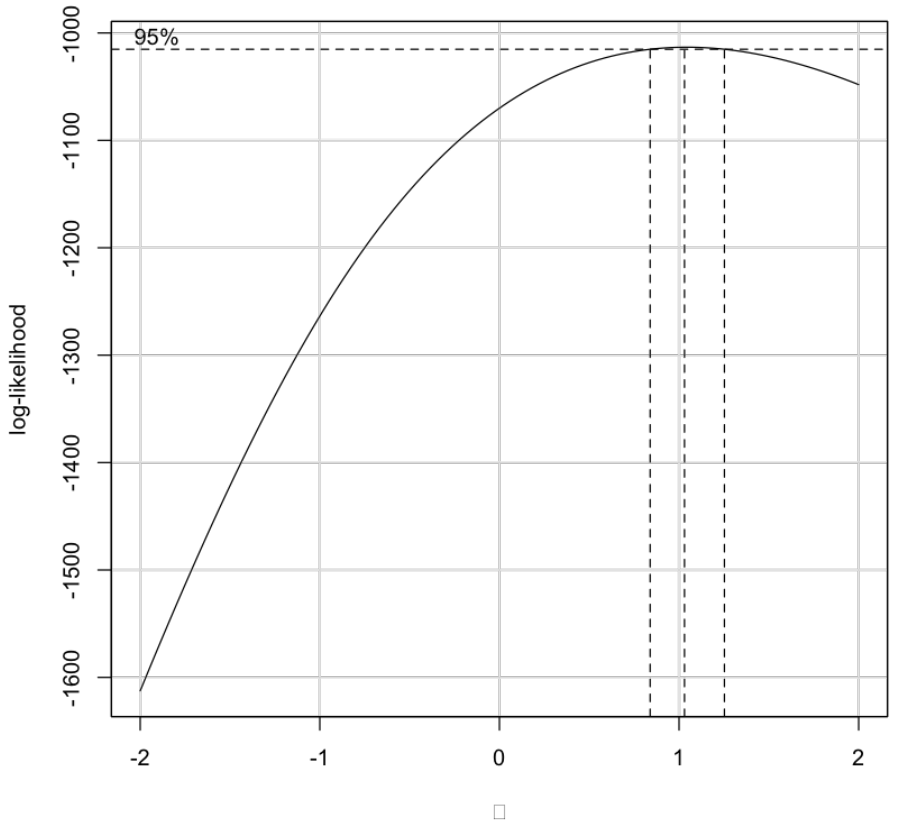


Fixed_IDNO	Method
182	Delete for extreme value
216	Delete for extreme value
39	Delete for extreme value
41	Delete for extreme value
42	Fix HEIGHT by BMI equation
96	Fix DENSITY by Siri equation
76	Fix BODYFAT by Siri equation
48	Fix BODYFAT by Siri equation
221	Fix ADIPOSITY by BMI equation
163	Fix ADIPOSITY by BMI equation

Variables Selection – Model Build

1. It is unnecessary to transformation on y according to BOXCOX.
2. It is useless to use PCA for model build
3. From simple linear regression of all variables, many variables is not significant and there is serious multicollinearity, the mean of VIF is 42.

Therefore, it is necessary to select variables.



Variables Selection – Basic Idea

- Method:

1. Lasso Regression
2. Subsets Method with Cp and BIC
3. Forward Directions Search with AIC and BIC

We prefer both simple and precise model.

- Full Model:

1. Full Model with all variables: BODYFAT~AGE+WEIGHT+HEIGHT+ADIPOSITY+NECK +CHEST+ABDOMEN+HIP+THIGH+KNEE+ANKLE+BICE PS+FOREARM+WRIST
2. Full Model with all variables after log (log()) transformation
3. Full Model with all variables after square(^2) transformation
4. Full Model with all variables from 1, 2, 3.

We still want to figure out any improvement after transformation on x

Variables Selection – Lasso Regression

1. Results: BODYFAT ~ AGE + HEIGHT + NECK + ABDOMEN + BICEPS + FOREARM + WRIST + THIGH
2. The multicollinearity is not so serious, the mean of VIF is 2.88.

	s0
AGE	0.04168313
WEIGHT	.
HEIGHT	-0.28032210
ADIPOSITY	.
NECK	-0.22089134
CHEST	.
ABDOMEN	0.69012676
HIP	.
THIGH	0.02151504
KNEE	.
ANKLE	.
BICEPS	0.02220848
FOREARM	0.06154111
WRIST	-1.24613508

Variables Selection – Subsets Method

Here we only give results from all variables including square and log transformation.

- BODYFAT~ABDOMEN R squared = 0.66
- BODYFAT~ABDOMEN + sqWEIGHT R squared = 0.71
- BODYFAT~LogWRIST + sqHEIGHT + LogABDOMEN R squared = 0.72

From four full model, the ABDOMEN is the most important variable. WEIGHT, WRIST are also very important.

	(Intercept)	AGE	WEIGHT	HEIGHT	ADIPOSITY	NECK	CHEST	ABDOMEN	HIP	THIGH	⋯	sqKNEE	sqANKLE	sqBICEPS	sqFOREARM	sqWRIST	
1	1	0	0	0	0	0	0	1	0	0	⋯	0	0	0	0	0	0.661
2	1	0	0	0	0	0	0	1	0	0	⋯	0	0	0	0	0	0.707
3	1	0	0	0	0	0	0	0	0	0	⋯	0	0	0	0	0	0.720
4	1	0	0	0	0	0	0	0	0	0	⋯	0	0	0	0	0	0.725
5	1	1	0	0	0	0	0	0	0	0	⋯	0	0	0	0	0	0.728

Variables Selection – Forward Directions Search

Here we only give results from BIC:

1. BODYFAT ~ ABDOMEN + WEIGHT + WRIST
   BIC=700.41
2. BODYFAT ~ LogABDOMEN + LogWRIST + LogHEIGHT
   BIC=697.82
3. BODYFAT ~ sqABDOMEN + sqWEIGHT + sqWRIST
   BIC=705.87
4. BODYFAT ~ ABDOMEN + sqWEIGHT + LogWRIST
   BIC=697.68

BIC will give simpler models than AIC.

ABDOMEN is also the most important variable.

```
Step:  AIC=700.41
BODYFAT ~ ABDOMEN + WEIGHT + WRIST

            Df Sum of Sq    RSS    AIC
<none>                    3822.9 700.41
+ BICEPS     1    47.398 3775.5 702.83
+ ADIPOSITY  1    46.899 3776.0 702.87
+ FOREARM    1    37.849 3785.0 703.46
+ HEIGHT     1    34.458 3788.4 703.68
+ THIGH      1    26.693 3796.2 704.19
+ AGE        1    24.213 3798.6 704.35
+ NECK       1    19.330 3803.5 704.67
+ ANKLE      1    12.236 3810.6 705.13
+ CHEST      1     6.471 3816.4 705.51
+ KNEE       1     3.818 3819.0 705.68
+ HIP        1     0.224 3822.6 705.91
```

Variables Selection – Summary

11 alternative models:

1. BODYFAT ~ ABDOMEN + sqWEIGHT + LogWRIST
2. BODYFAT ~ ABDOMEN + sqWEIGHT
3. BODYFAT ~ sqABDOMEN + sqWEIGHT + sqWRIST
4. BODYFAT ~ sqABDOMEN + sqWEIGHT
5. BODYFAT ~ sqABDOMEN
6. BODYFAT ~ LogABDOMEN + LogWRIST + LogHEIGHT
7. BODYFAT ~ LogABDOMEN + LogWRIST
8. BODYFAT ~ LogABDOMEN
9. BODYFAT ~ ABDOMEN + WEIGHT + WRIST
10. BODYFAT ~ ABDOMEN + WEIGHT
11. BODYFAT ~ ABDOMEN

Model Selection

– 30 - Repeated with 10 - fold Cross Validation

ID	Model	Root Mean Squared Error	R-squared
1	BODYFAT ~ ABDOMEN + sqWEIGHT + LogWRIST	3.937862	0.7247008
2	BODYFAT ~ ABDOMEN + sqWEIGHT	4.003475	0.7136670
3	BODYFAT ~ sqABDOMEN + sqWEIGHT + sqWRIST	4.003981	0.7160629
4	BODYFAT ~ sqABDOMEN + sqWEIGHT	4.042249	0.7085140
5	BODYFAT ~ sqABDOMEN	4.320769	0.6634867
6	BODYFAT ~ LogABDOMEN + LogWRIST + LogHEIGHT	3.936800	0.7203500
7	BODYFAT ~ LogABDOMEN + LogWRIST	4.035844	0.7090924
8	BODYFAT ~ LogABDOMEN	4.294042	0.6701187
9	BODYFAT ~ ABDOMEN + WEIGHT + WRIST	3.955417	0.7190513
10	BODYFAT ~ ABDOMEN + WEIGHT	4.025416	0.7134265
11	BODYFAT ~ ABDOMEN	4.292844	0.6716008

Model Diagnosis and Summary

We choose ABDOMEN, WEIGHT and WRIST as our predictors.

The final model is:

BODYFAT=

0.88 ABDOMEN-0.08 WEIGHT-1.26 WRIST -24.23.

The adjusted R squared is 0.7133.

Residual standard error is 3.958.

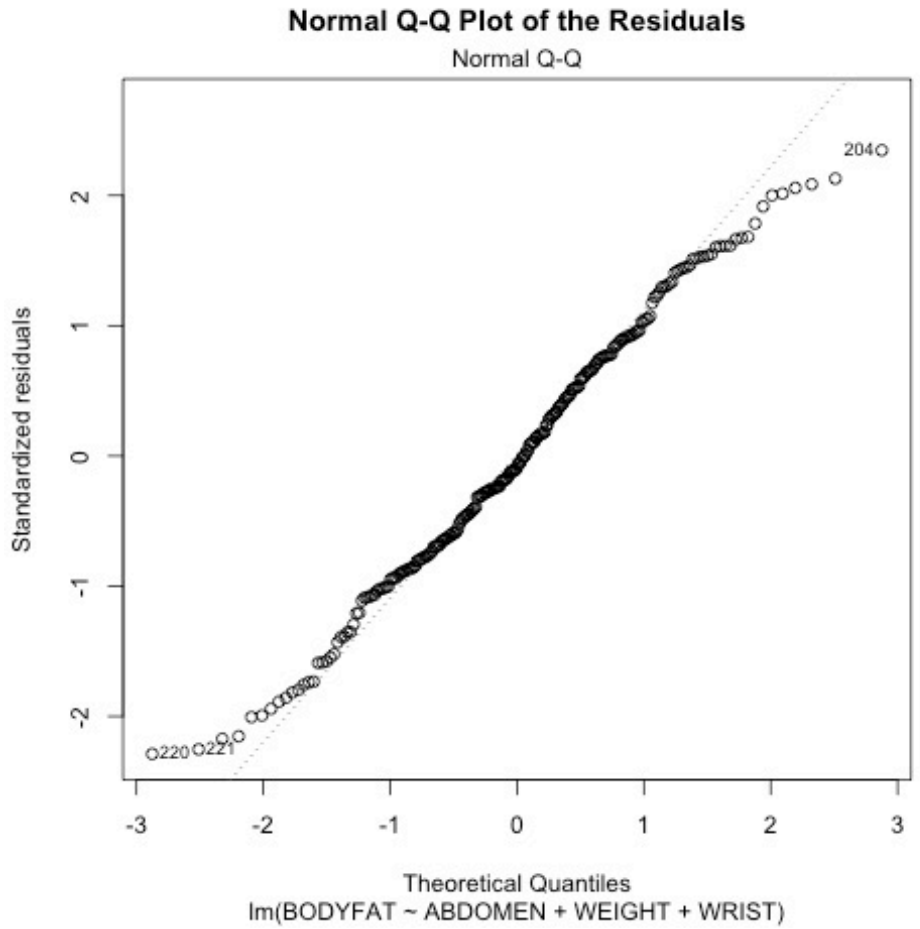
All variables are significant.

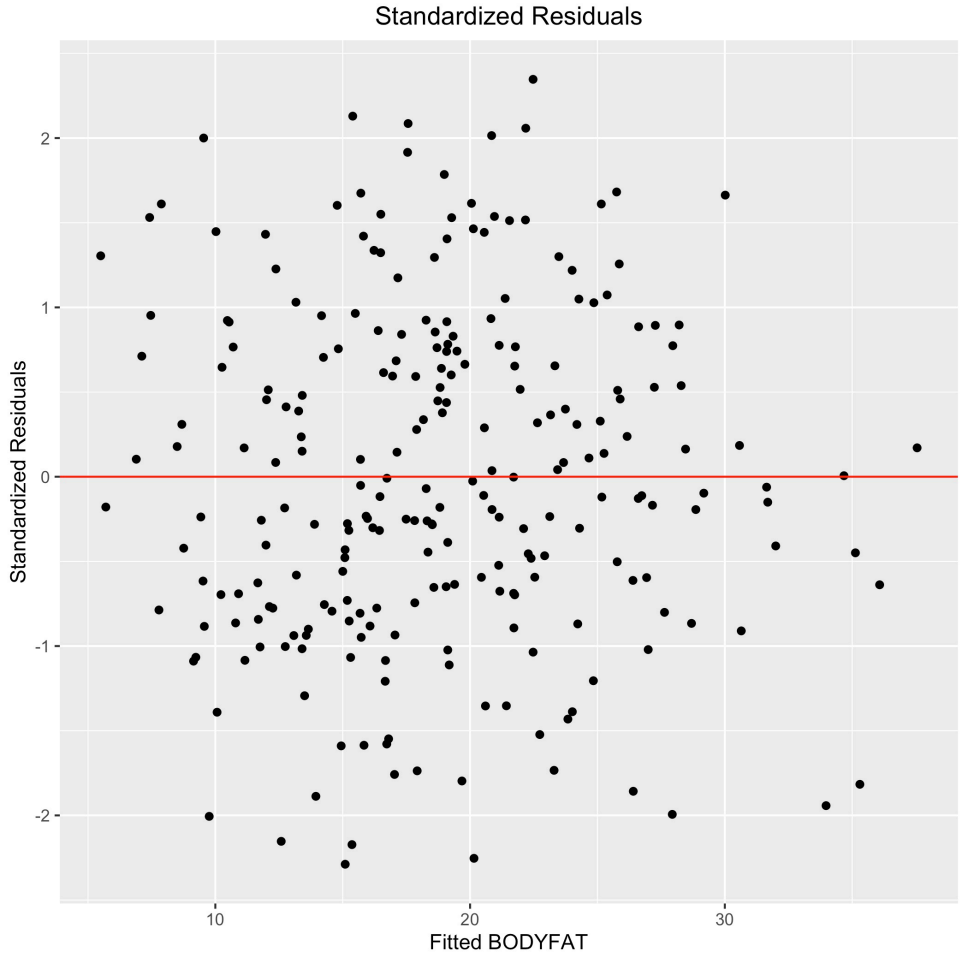
No serious multicollinearity, mean VIF=3.8.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -24.22925    6.27922  -3.859 0.000146 ***
ABDOMEN       0.87792    0.05210  16.851  < 2e-16 ***
WEIGHT       -0.08384    0.02226  -3.766 0.000208 ***
WRIST        -1.26116    0.40007  -3.152 0.001822 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.958 on 244 degrees of freedom
Multiple R-squared:  0.7168,    Adjusted R-squared:  0.7133
F-statistic: 205.9 on 3 and 244 DF,  p-value: < 2.2e-16
```

Model Diagnosis
– Standardized Residuals, QQ plots for Residuals





Model Diagnosis – Leverage Plots

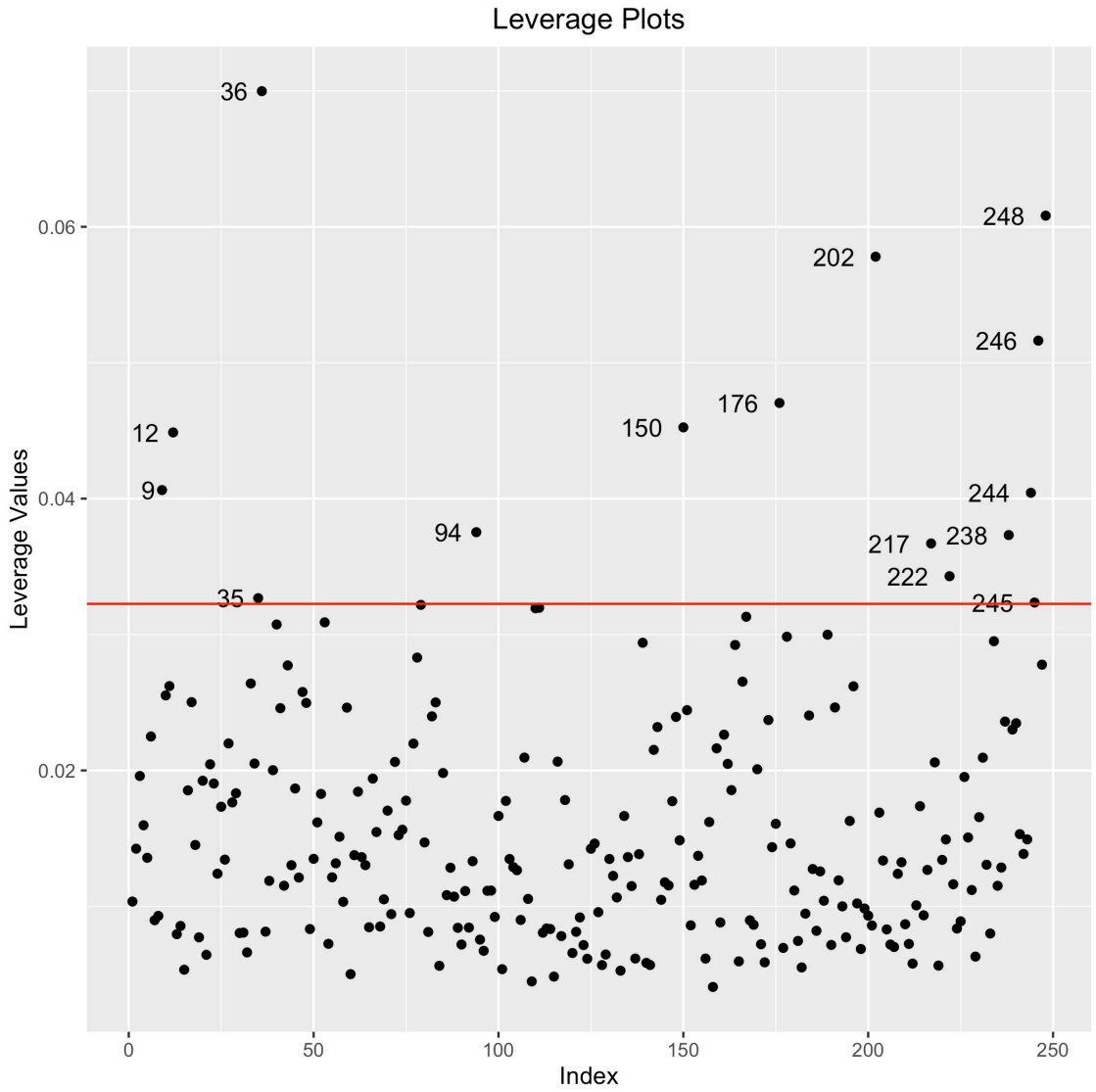
The leverage is diagonal elements from $H=X(X'X)^{-1}X'$

There is a lot of outliers in Leverage plots.

But the coefficients of model change little after removing the outliers and fitting again.

We keep the initial model .

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -22.43134    6.99358  -3.207  0.00153 **
ABDOMEN       0.87763    0.06018  14.584  < 2e-16 ***
WEIGHT       -0.07934    0.02585  -3.069  0.00240 **
WRIST        -1.39633    0.44868  -3.112  0.00209 **
```

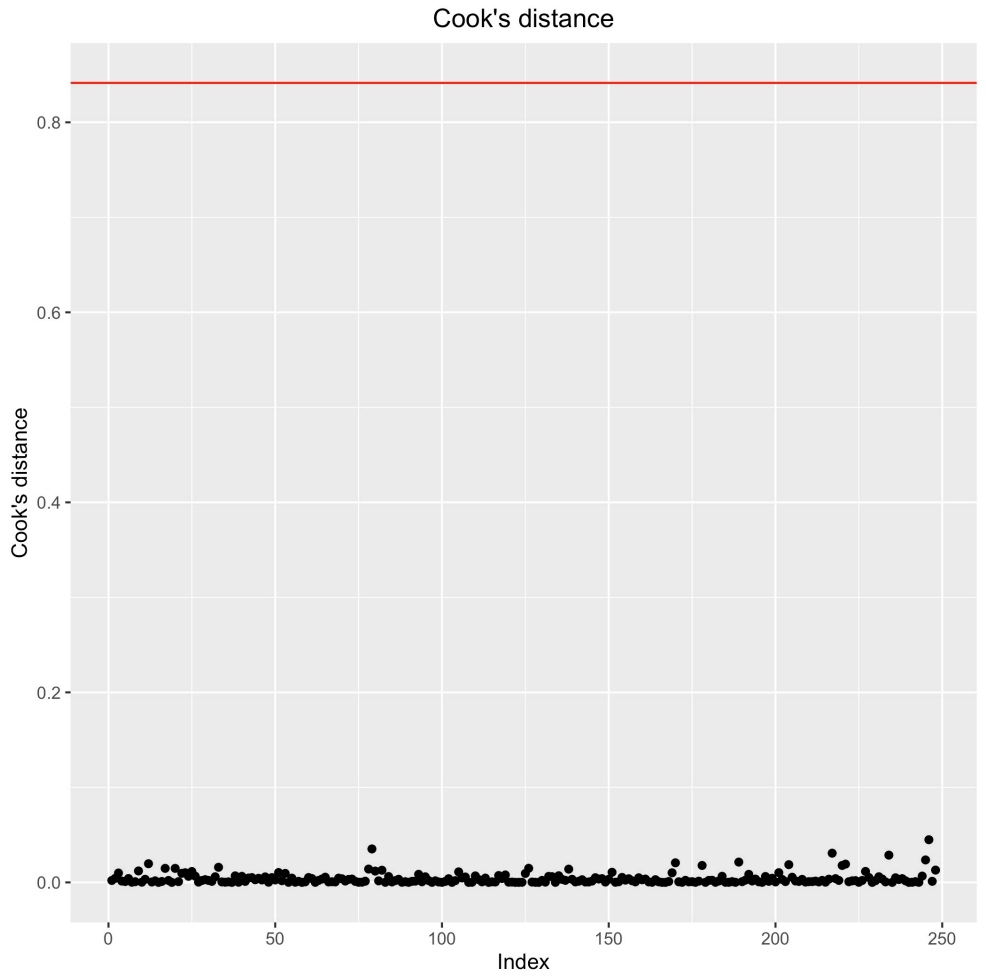


Model Diagnosis – Cook's Distance

Cook's distance $D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{p\hat{\sigma}^2}$

measure the influence of i^{th} observation on all n fitted values.

There is no outliers in Cook's Distance.



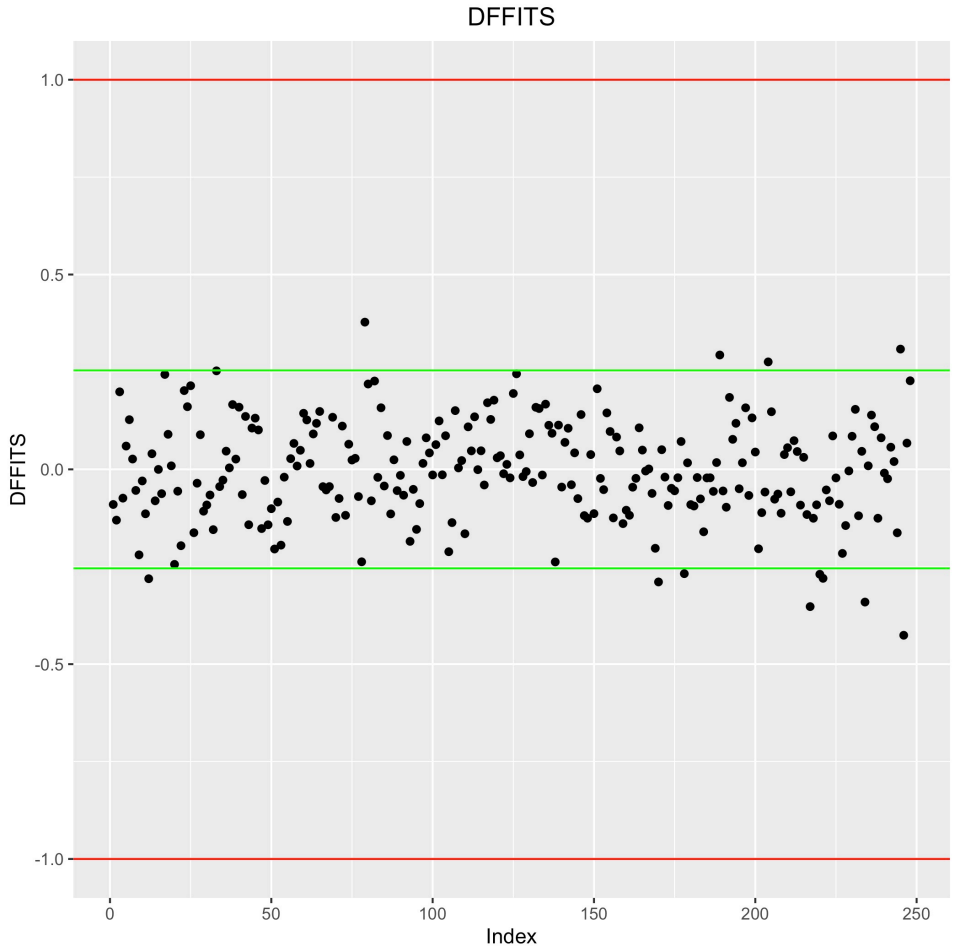
Model Diagnosis - DFFITS

DFFITS $\mathrm{DFFITS}_i = \dfrac{\hat{Y}_i - \hat{Y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}}$

measures the influence of the i^{th} observation on the fitted value yi.

The rule of thumb for small dataset is red lines, for large dataset is green lines.

There is no obvious outliers in DFFITS.



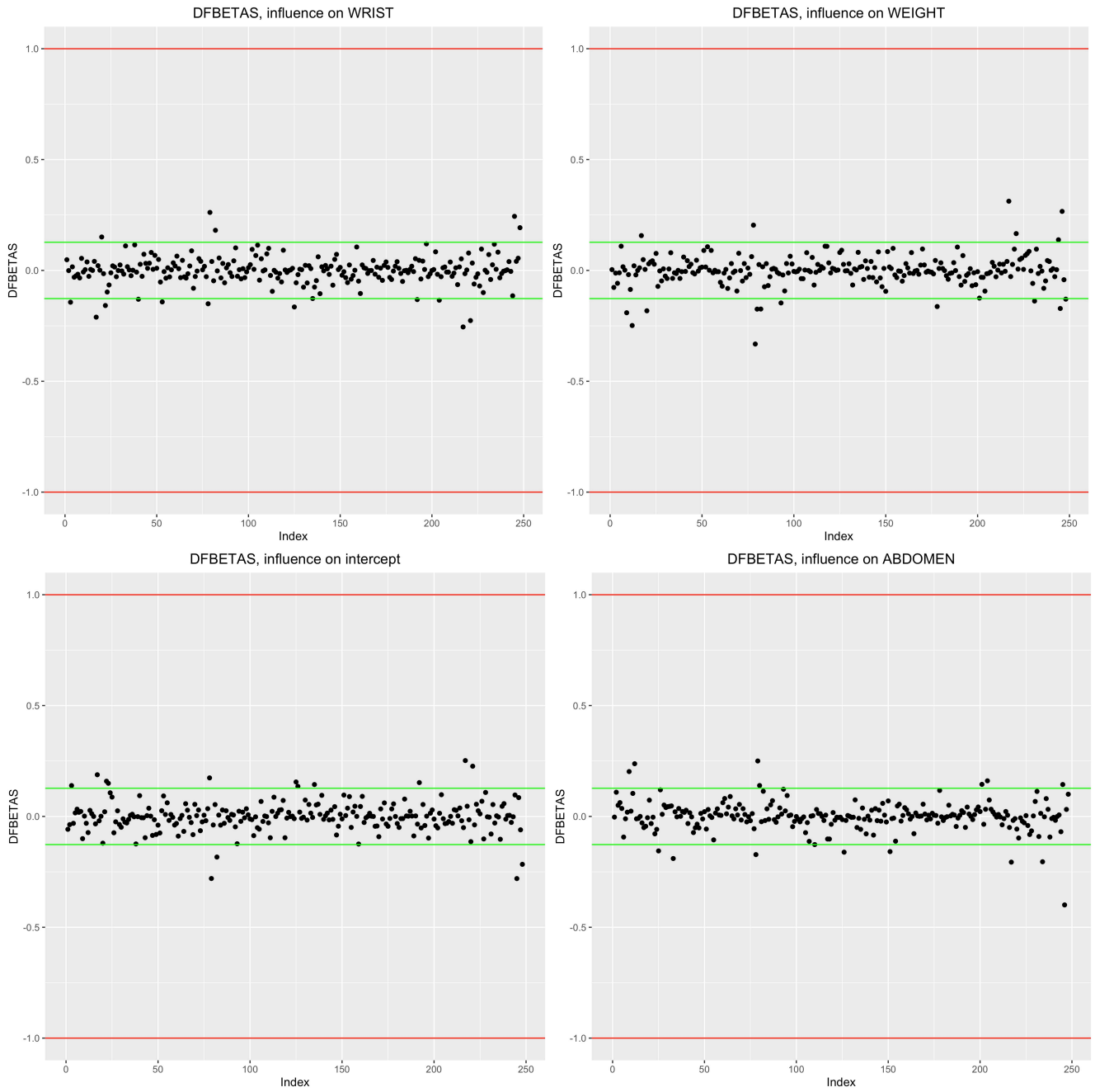
Model Diagnosis - DFBETAS

DFBETAS $\mathrm{DFBETAS}_{k(i)} = \dfrac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{\hat{\sigma}^2_{(i)}(X^T X)^{-1}_{kk}}}$

measures the influence of i[th] observation on the fit of the regression coefficient β_k .

The rule of thumb for small dataset is red lines, for large dataset is green lines.

There is no obvious outliers in DFBETAS.



Model Diagnosis and Summary

BODYFAT = 0.88 ABDOMEN-0.08 WEIGHT-1.26 WRIST -24.23.

Man with 154 lbs WEIGHT, 17 cm WRIST and 85 cm ABDOMEN, his BODYFAT will be 16(%).

Advantages: Perform well in diagnosis, no serious multicollinearity, linear regression assumptions hold.

Disadvantages: Contains three inputs and some of them are still not convenient to measure. Our model is based on data with only men.

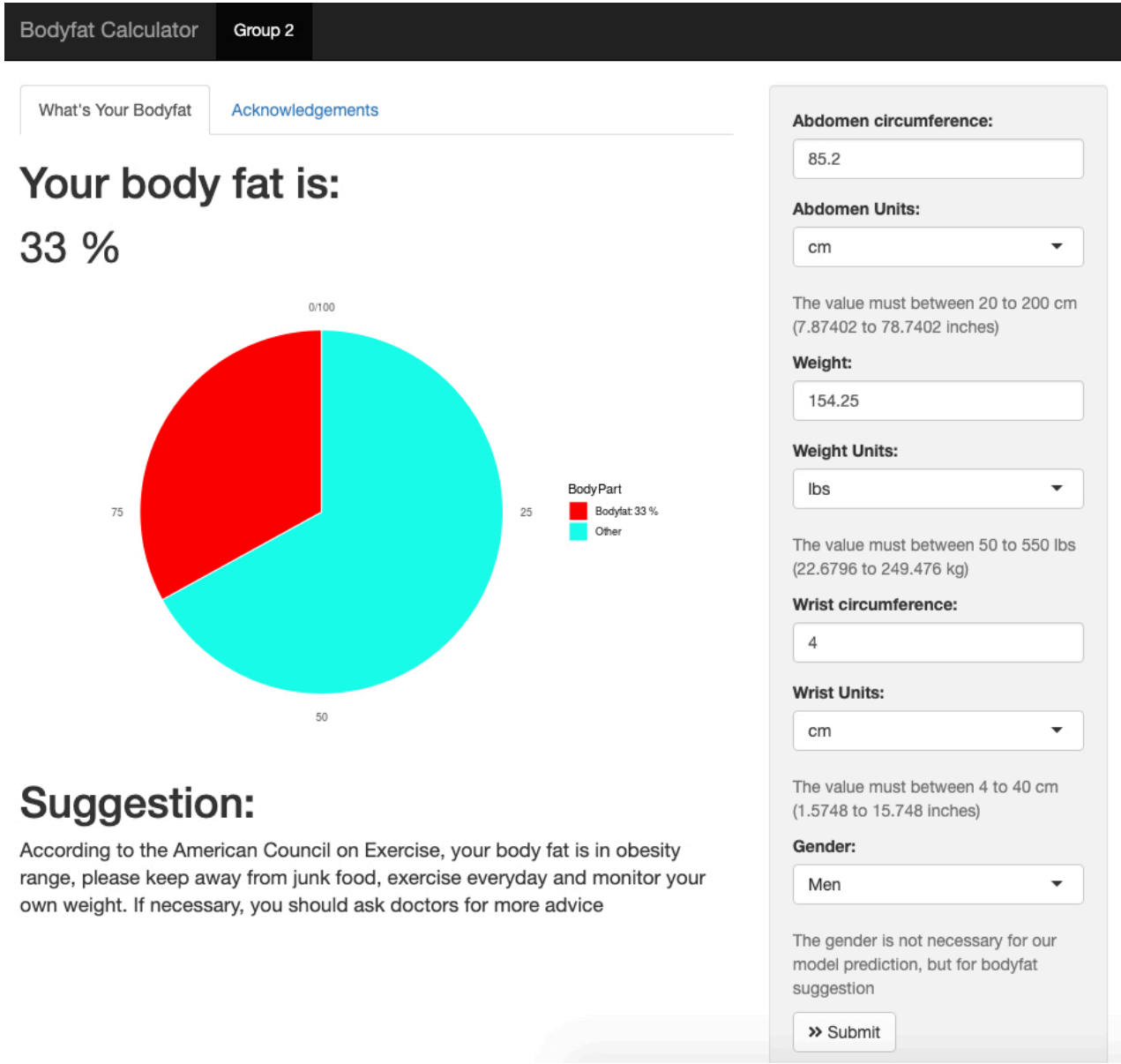
The larger the ABDOMEN is, the larger BODYFAT will be.

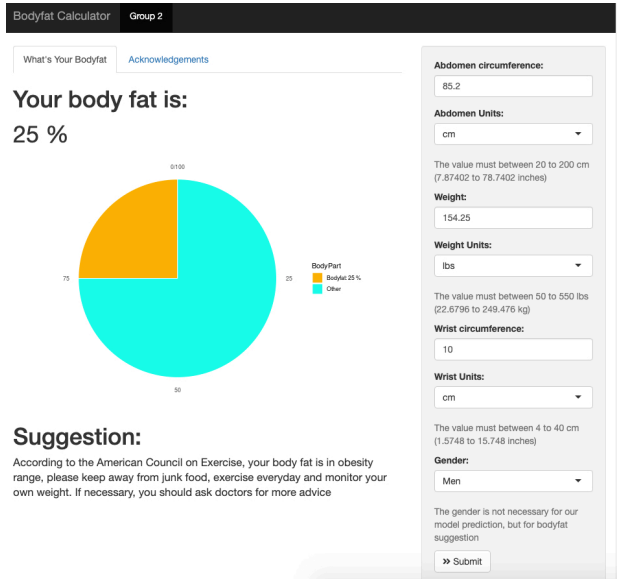
The larger the WEIGHT is, the smaller BODYFAT will be.

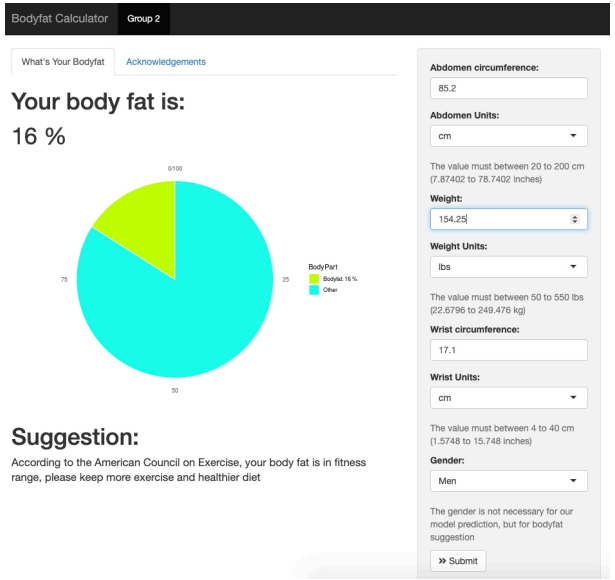
The larger the ABDOMEN is, the smaller BODYFAT will be.

Shiny Application

1. Unit Change
2. Color Warning
3. Body fat suggestion from American Council on Exercise







Acknowledge

- **GitHub, Main Code**

1. Main code edited by CHENYANG JIANG and ENZE WANG
2. PCA edited by HanGyu KANG
3. Model selection fixed by RUI HUANG
4. GitHub maintained by ENZE WANG

- **Shiny App**

1. Main edited by RUI HUANG and HanGyu KANG
2. The interface, ggplot and unit part are fixed, edited and maintained by ENZE WANG
3. Other problems are fixed by CHENYANG JIANG

- **Presentation**

1. PPT edited by ENZE WANG
2. Video edited by HanGyu KANG
3. Presented with CHENYANG JIANG and RUI HUANG

- **PDF Summary**

1. Main edited by ENZE WANG
2. Fixed by RUI HUANG, CHENYANG JIANG and HanGyu KANG