628 Body Fat Data Project

Group 3 Members:

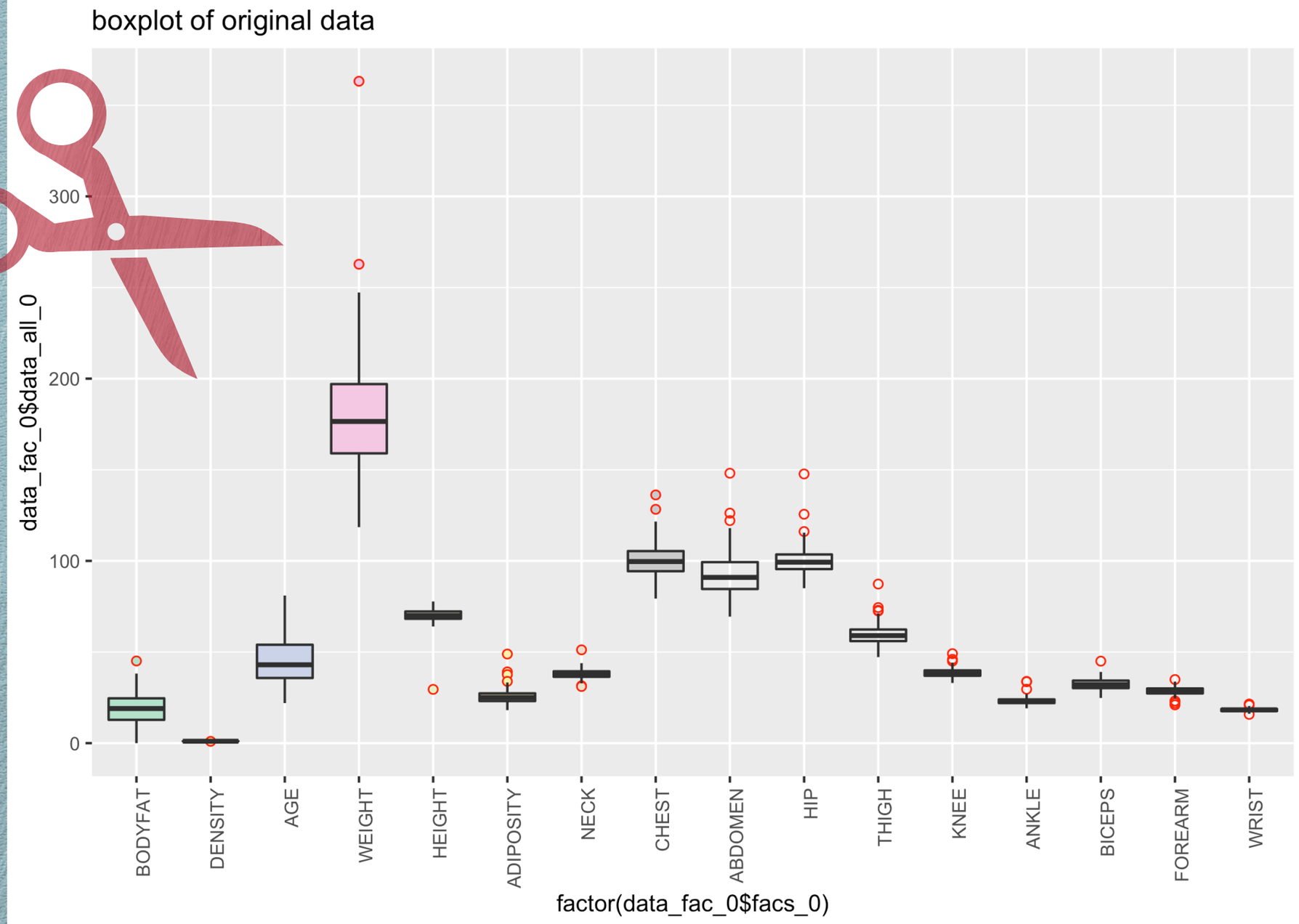
Lingfeng Zhu lzhu88@wisc.edu

Ruochen Yin ryin26@wisc.edu

Jiahan Li jli936@wisc.edu

Chong Wei cwei48@wisc.edu

OUTLINE



Data Cleaning

boxplot of original data

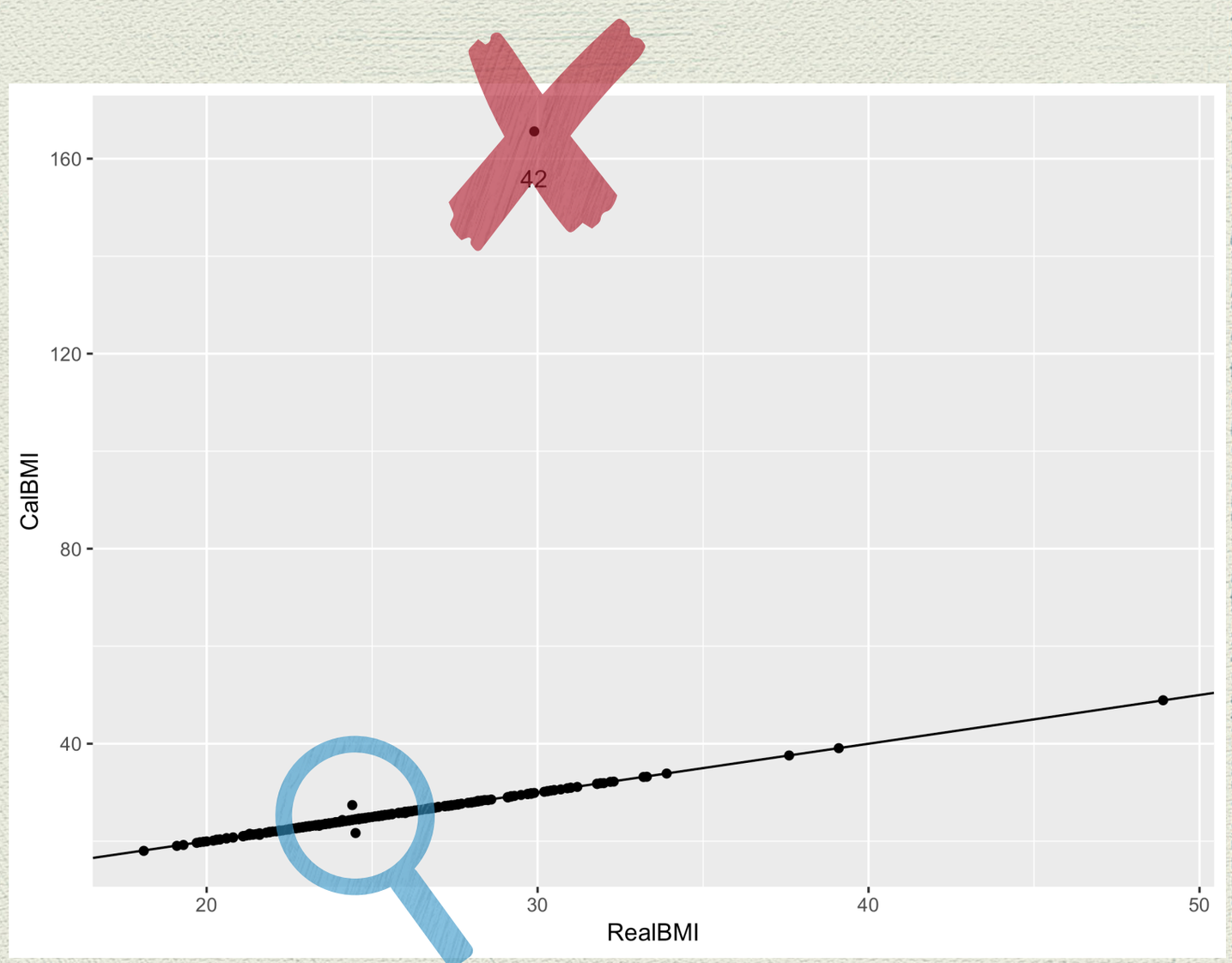
Check Quantile

- Calculating the rage of weight, height, density, bodyfat between quantile 0.01, 0.99 :
- Weight_q = (125.505, 245.720)
- Height_q = (64.3825, 76.0000)
- Density_q = (1.016040, 1.095393)
- Bodyfat_q = (4.355, 35.582)

Check outliers —- BMI

- Relationship among adiposity, weight and height:

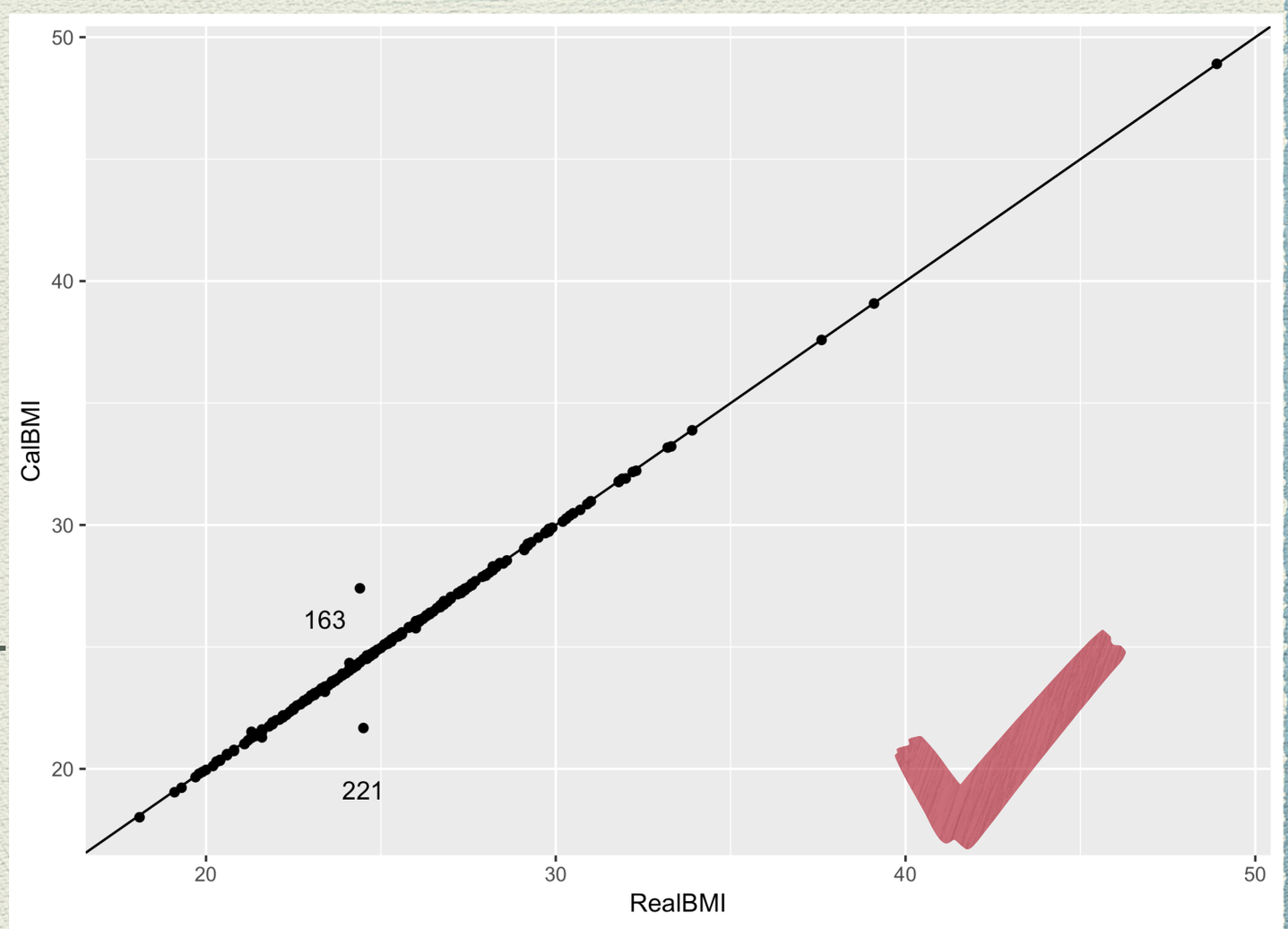
- $$BMI = \frac{703 * WEIGHT(lbs)}{HEIGHT^2(inch)}$$



Check outliers —- BMI

- Relationship among adiposity, weight and height:

- $$BMI = \frac{703 * WEIGHT(lbs)}{HEIGHT^2(inch)}$$

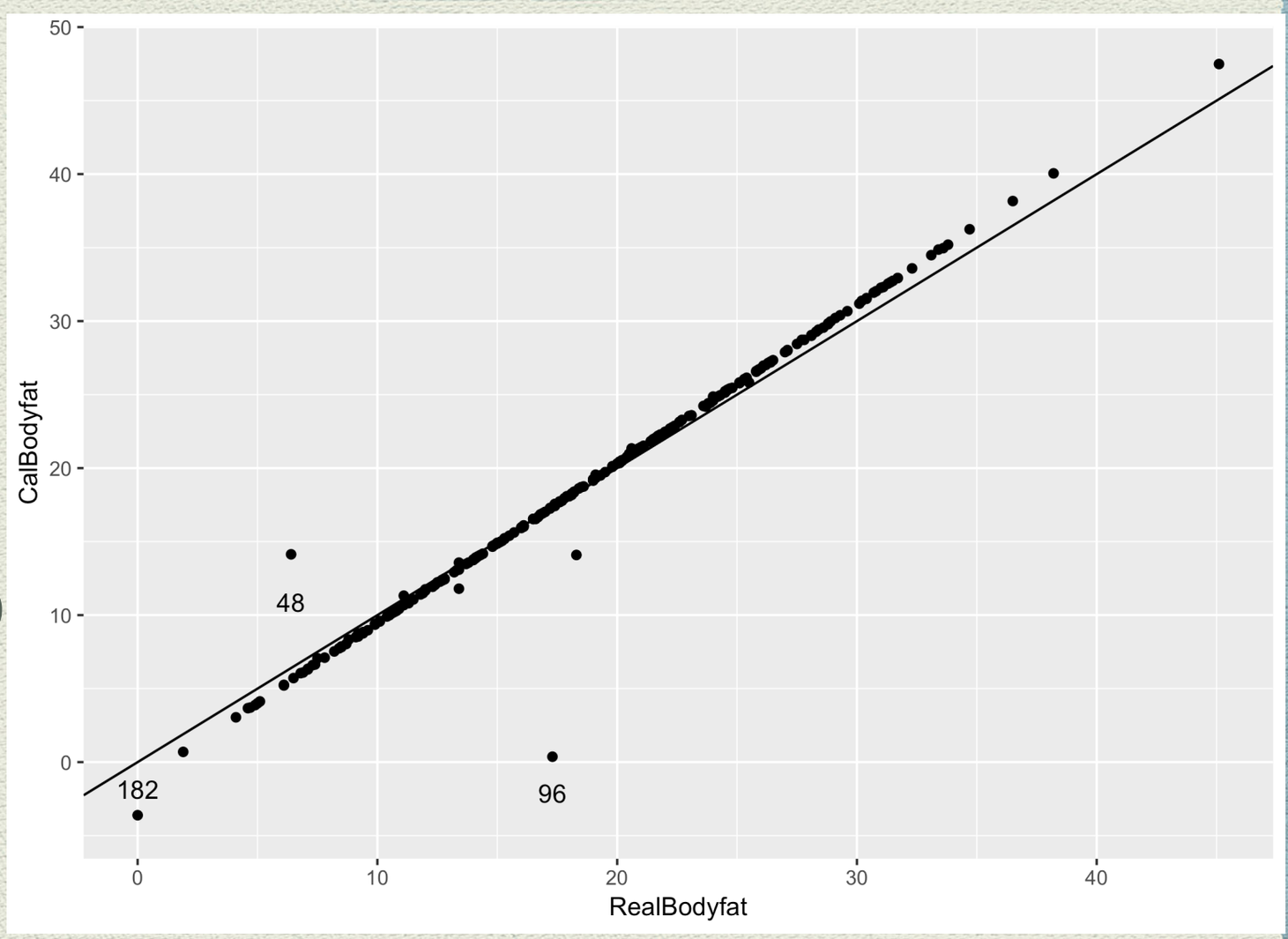


Check outliers

- From formula of BMI, we have point 42's height 29.5 is abnormal, so we change it according to the formula and let it become 69.43.
- Although point 163, 221 are not perfectly matched the calculated BMI, but their height and weight is in the range, so we won't change them.

Check outliers —- Bodyfat

- Relationship between bodyfat and density:
- $$BODYFAT = \frac{495}{DENSITY} - 450$$

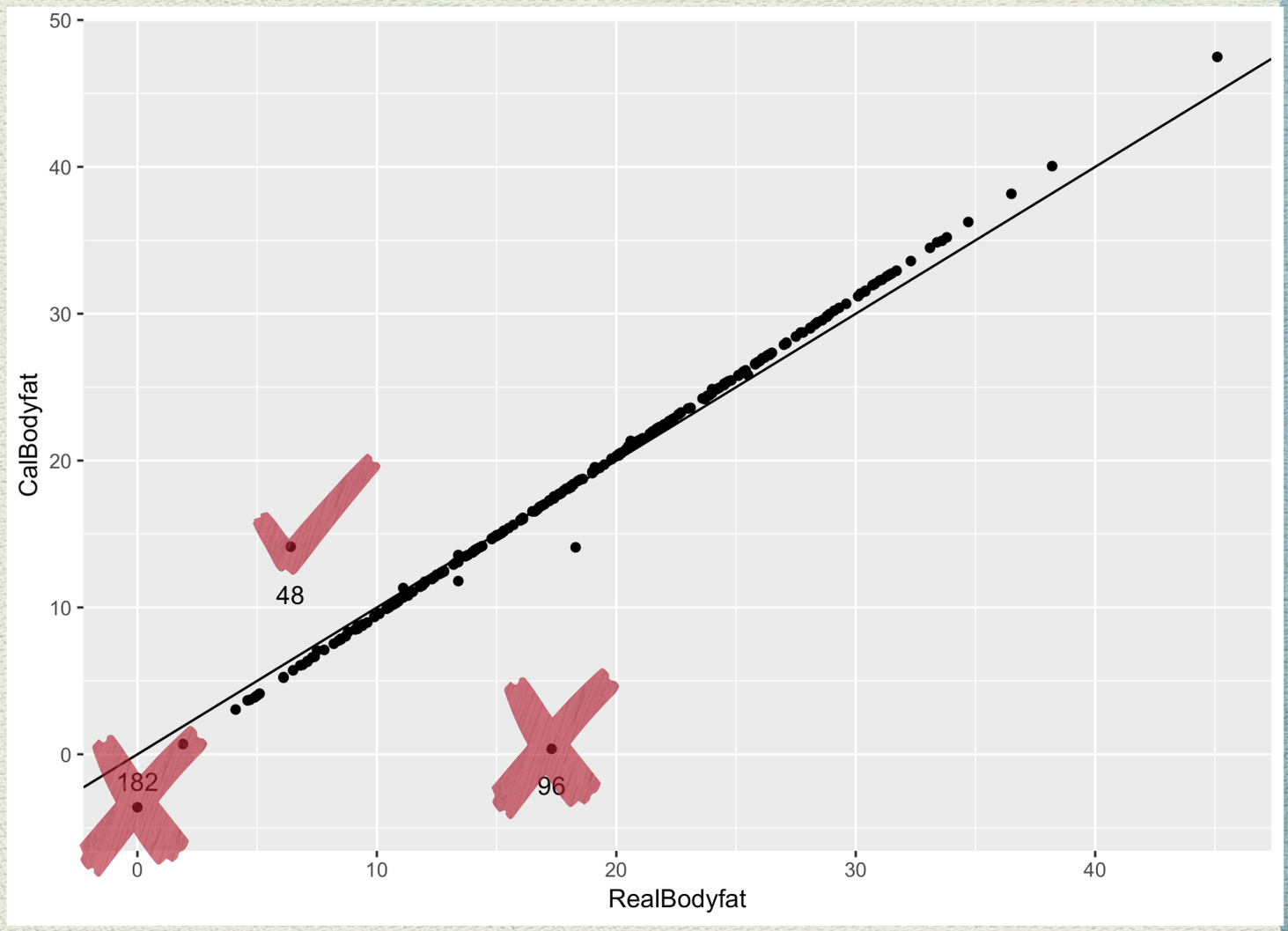


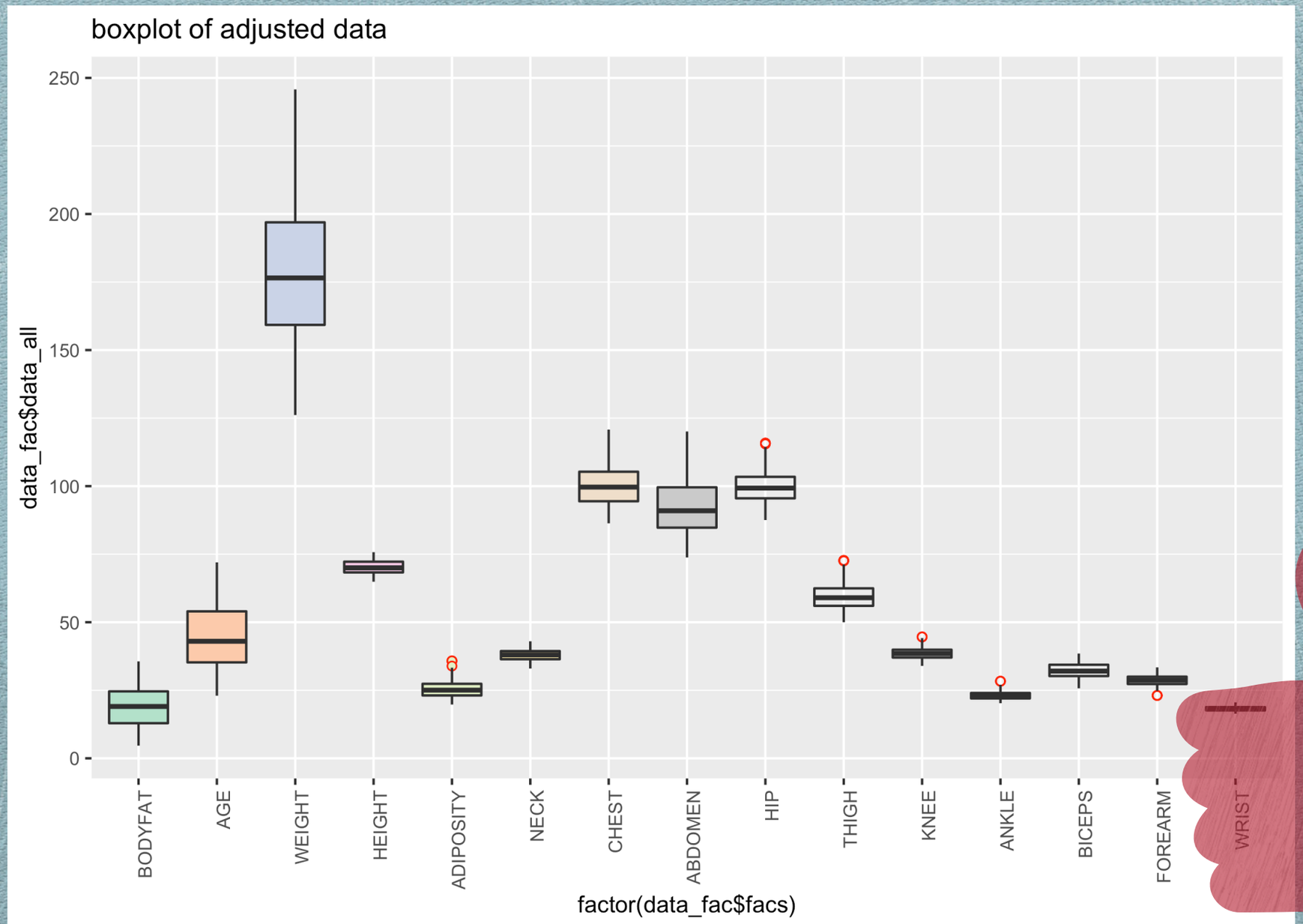
Check outliers

- From formula of Bodyfat, point 182 has the real bodyfat lower than zero, which is impossible. Besides, the density is also close to 0 which means we can’t adjust it, so we just delete 182.
- Point 48, 96 is away from the calculated bodyfat, too. Point 48 has normal density and bodyfat, so we will keep it. However, point 96 has density away from the range, so we will delete point 96.
- In conclusion, we delete 2 points:96,182

Check outliers —- Bodyfat

- Relationship between bodyfat and density:
- $$BODYFAT = \frac{495}{DENSITY} - 450$$





Data Cleaning

boxplot of adjusted data

AIC

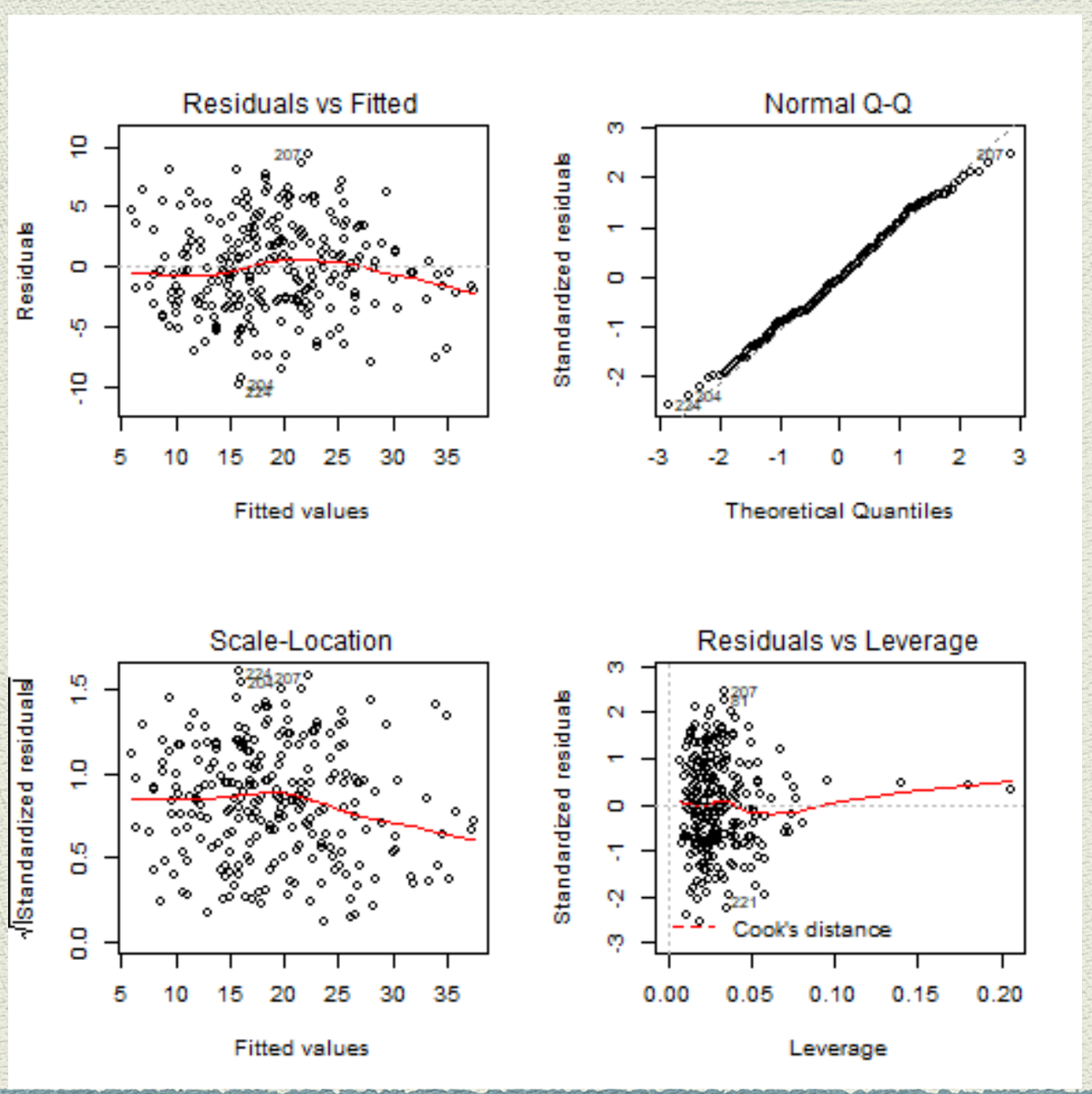
BIC

LASSO

Feature Selection

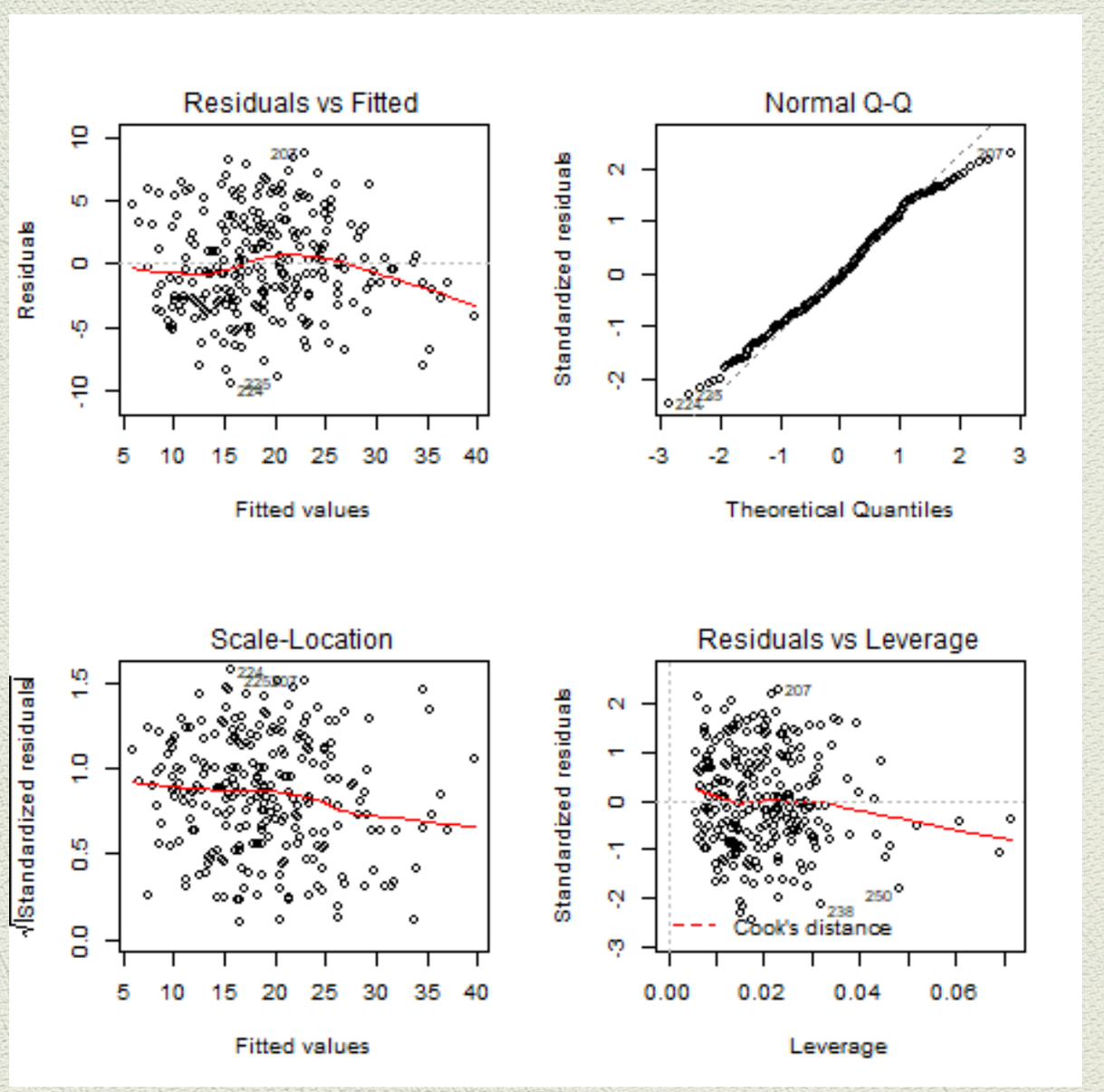
AIC

BODYFAT ~ ABDOMEN + WEIGHT + WRIST + AGE + THIGH + NECK + FOREARM



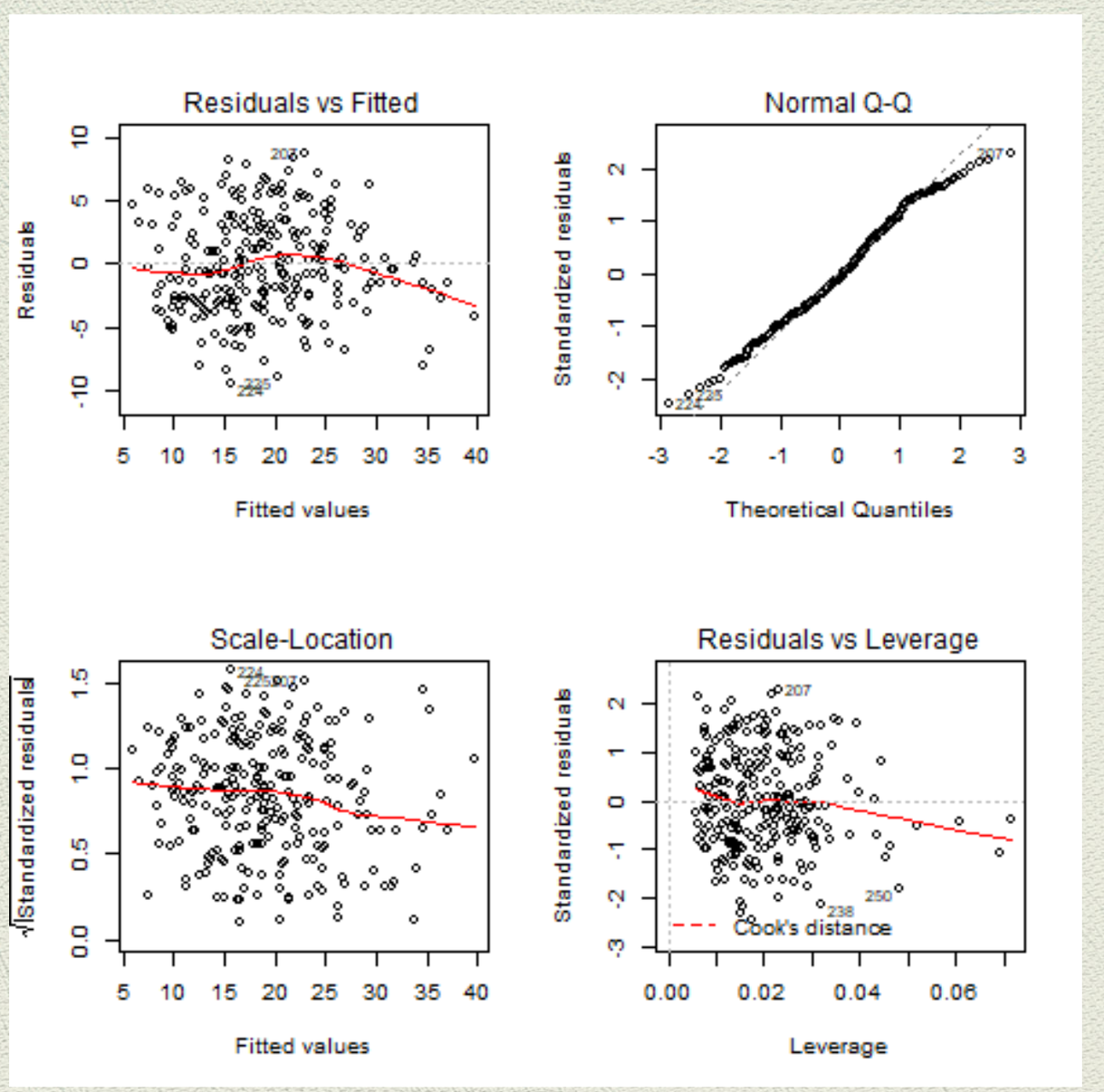
BIC

BODYFAT ~ AGE + ABDOMEN + WRIST + HEIGHT



LASSO

BODYFAT ~ AGE + HEIGHT + ABDOMEN + WRIST



Comparison

AIC

BODYFAT ~ ABDOMEN + WEIGHT + WRIST + AGE + THIGH + NECK + FOREARM

R-square : 0.7418 Cv-mean: 15.4747

BIC

BODYFAT ~ AGE + HEIGHT + ABDOMEN + WRIST

R-square : 0.7336 Cv-mean: 15.6591

LASSO

BODYFAT ~ AGE + HEIGHT + ABDOMEN + WRIST

R-square : 0.7336 Cv-mean: 15.6591

Model 1: Bodyfat ~ Weight + Abdomen

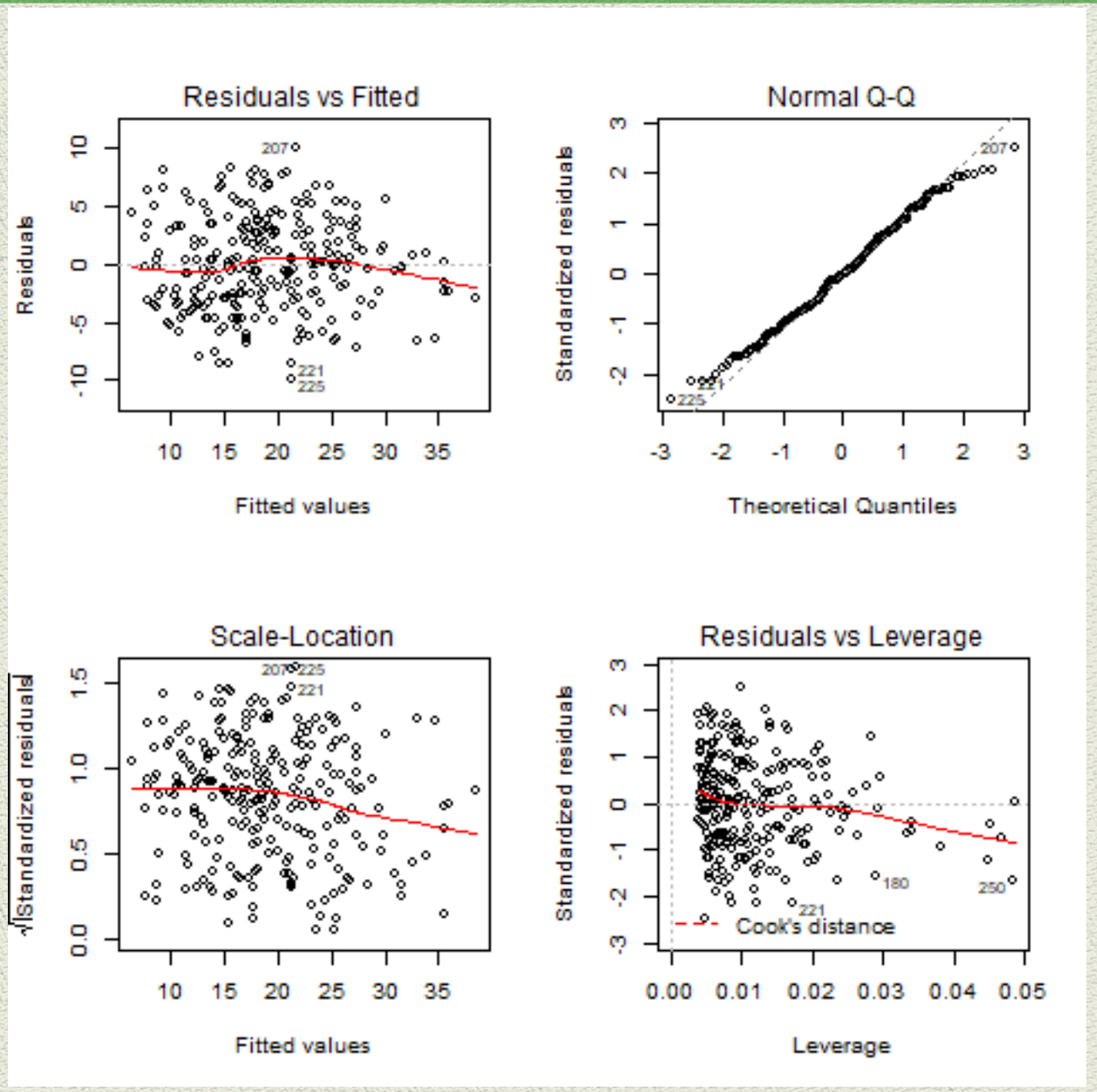
Model 2: Bodyfat ~ Weight * Abdomen

Model 3: Bodyfat ~ Abdomen

Trying other models

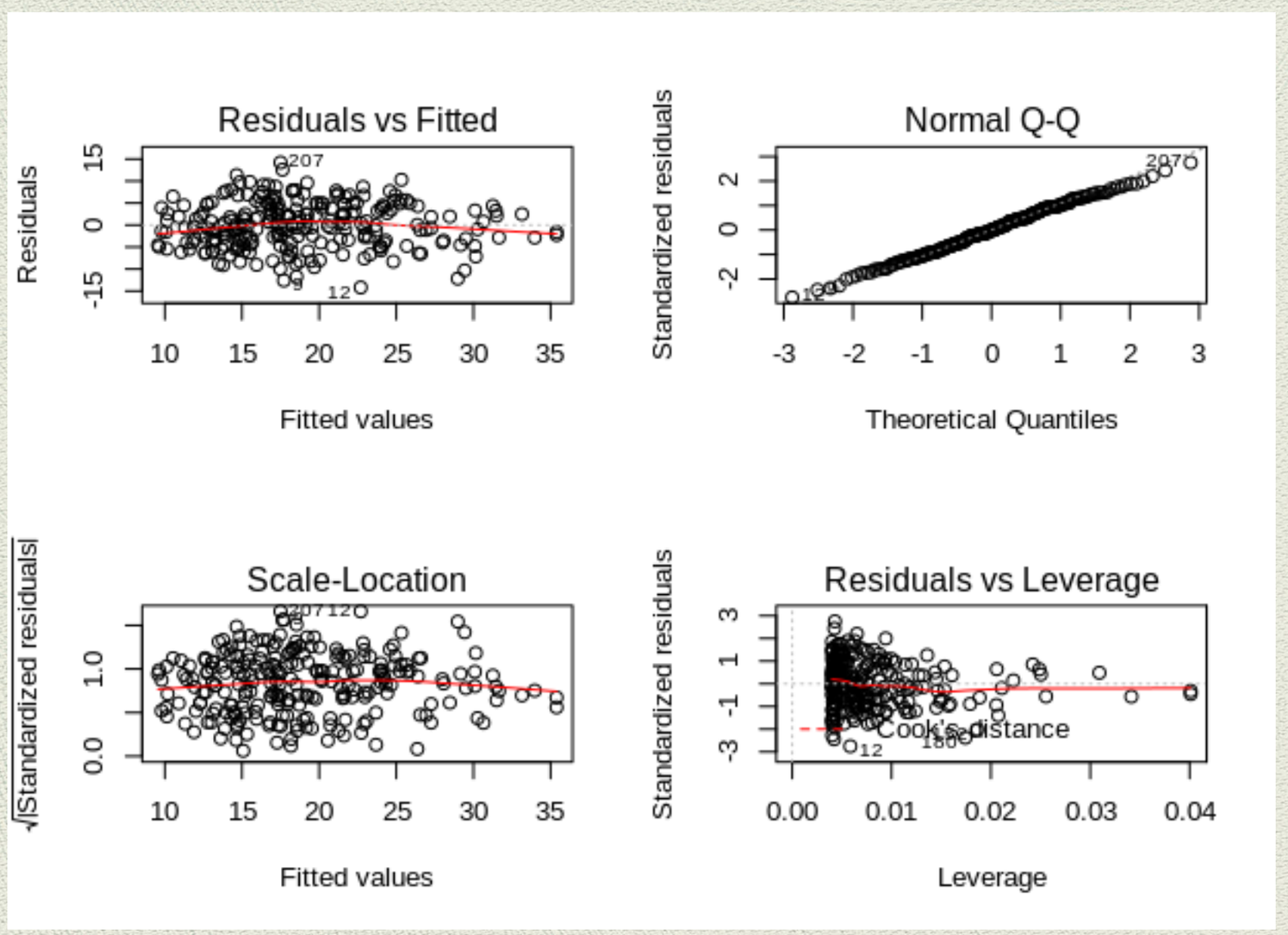
Model 1: Bodyfat ~ Weight + Abdomen

R-square : 0.7181 Cv-mean: 16.2636



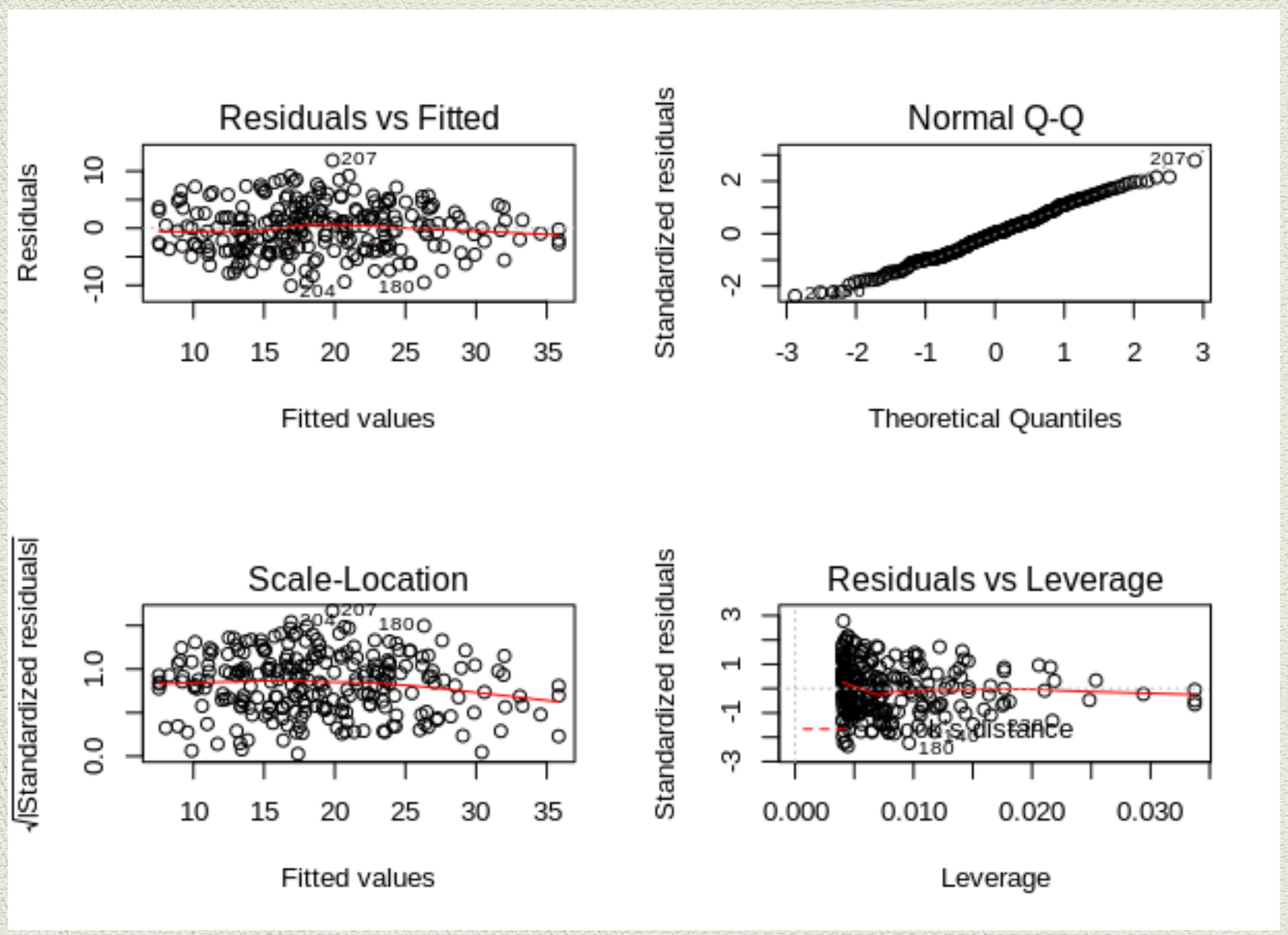
Model 2: Bodyfat ~ Weight * Abdomen

R-square : 0.5293 Cv-mean: 26.9237



Model 3: Bodyfat ~ Abdomen

R-square : 0.6785 Cv-mean: 18.3760



Comparison

BIC

BODYFAT ~ AGE + HEIGHT + ABDOMEN + WRIST

R-square : 0.7336 Cv-mean: 15.6591

Model 1

BODYFAT ~ WEIGHT + ABDOMEN

R-square : 0.7181 Cv-mean: 16.2636

Model 2

BODYFAT ~ WEIGHT * ABDOMEN

R-square : 0.5293 Cv-mean: 26.9237

Model 3

BODYFAT ~ ABDOMEN

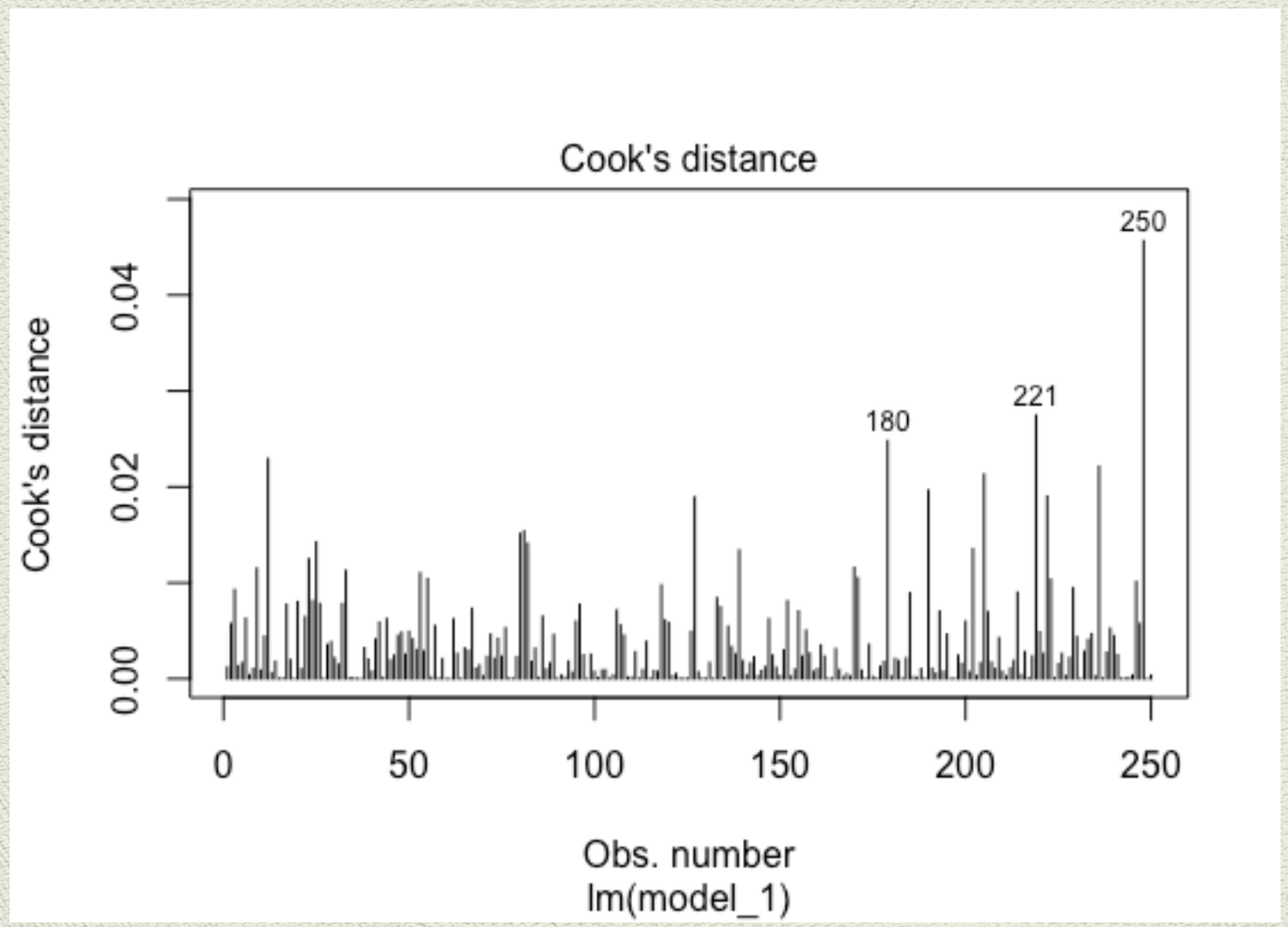
R-square : 0.6785 Cv-mean: 18.3760

Model One

BODY FAT = 2.94338 + 0.05205 * AGE + 0.69798 * ABDOMEN - 1.71840 * WRIST - 0.27799 * HEIGHT

R-square : 0.7336 Cv-mean: 15.6591 P-value < 2.2e-16

Possible Rule of Thumb:
Your body fat equal to
Your abdomen (cm)
multiply by 0.7
minus wrist (cm) multiply
by 1.7 minus height(inch)
multiply by 0.28 add 3.0



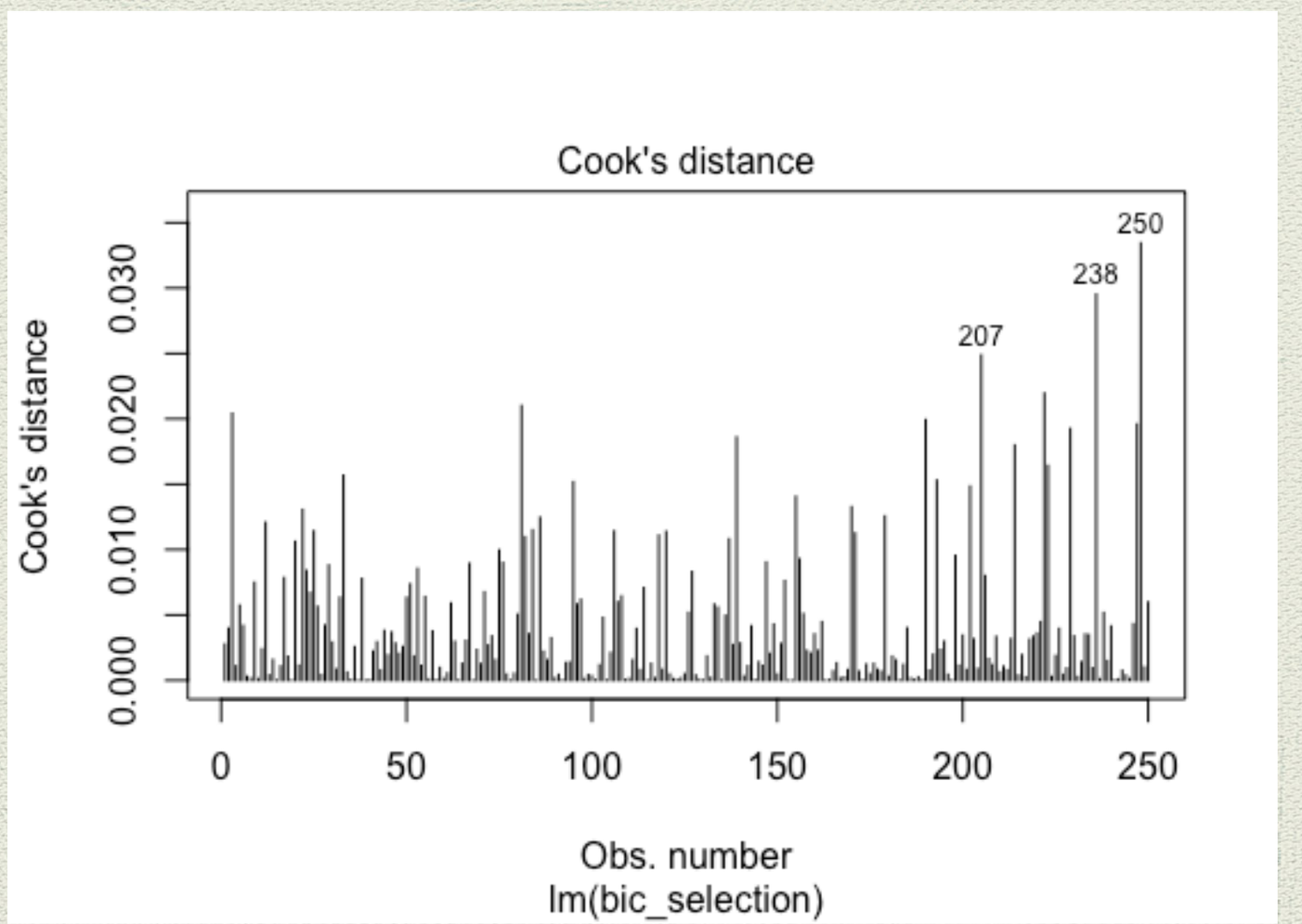
Model Two

BODY FAT = -41.60272 – 0.11553 * WEIGHT
+ 0.87794 * ABDOMEN

R-square : 0.7181 Cv-mean: 16.2636 P-value < 2.2e-16

Possible Rule of Thumb:

Your body fat equal to
Your abdomen (cm)
multiply by 0.88
minus weight (lbs)
multiply by 0.12 minus
41.6

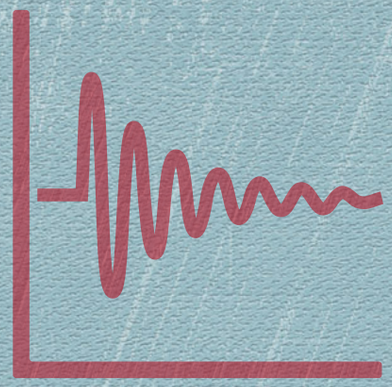


Strengths:

1. To deal with outliers, for each feature, we use 99% quantile to replace the values greater than the 99% quantile and 1% quantile to replace the values less than it, by doing so, we can prevent our model from the influences of extreme values.

2. We repeat 10-fold Cross Validation 1000 times, which can make our results more convincible.

3. Our final model is simple linear regression model, which contains only two significant predictors, and they are easy to measure during daily life in real world. That makes our model concise and easy to interpret.



Weaknesses

1. The methods we use in data cleaning (such as using the quantiles) may cause some internal changes of the data structure, which can affect our model results.

2. We only choose two predictors for simplicity, but there is possibility that model with more predictors may give us better results in some specific cases.

Summary

To draw a conclusion, we obtain two linear models, one of which contains 2 predictors while the other one has 4. Since the performances of these two models are close to each other, our shiny application contains both. Users can choose either of them to calculate body fat based on the data he has. But for simplicity, we prefer the model with less predictor (in this case, the model with only two predictors) if we have to pick one model as the best one.

Thank You !

Here is our shiny app:
https://lingfengzhu.shinyapps.io/bodyfat/