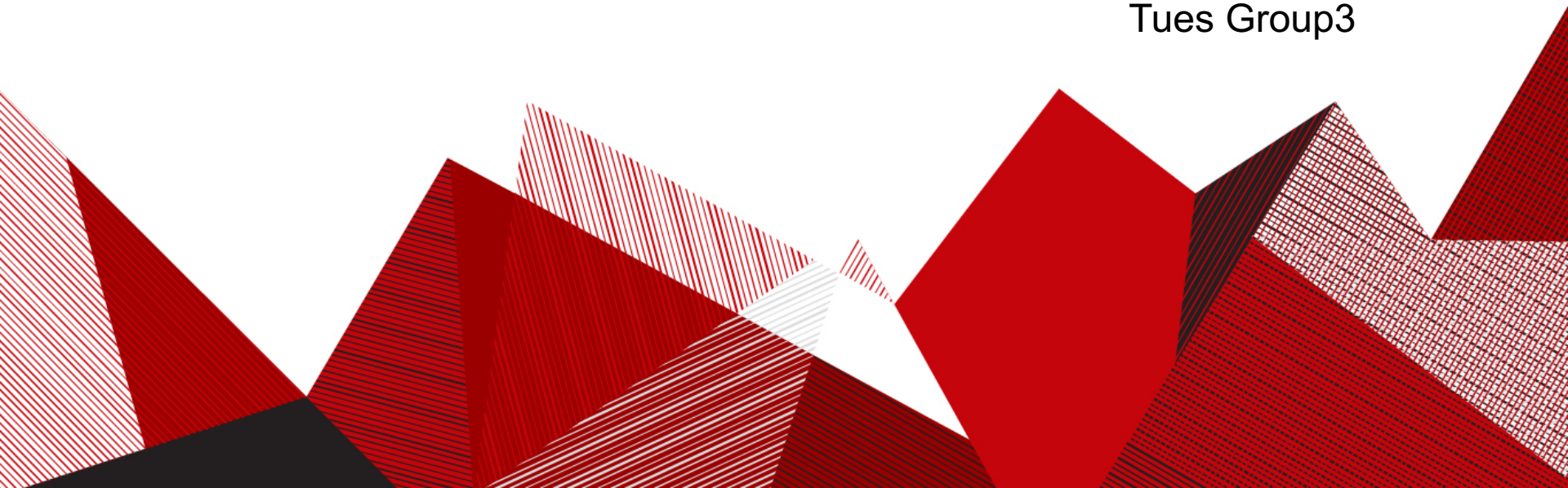
Module1: BODYFAT

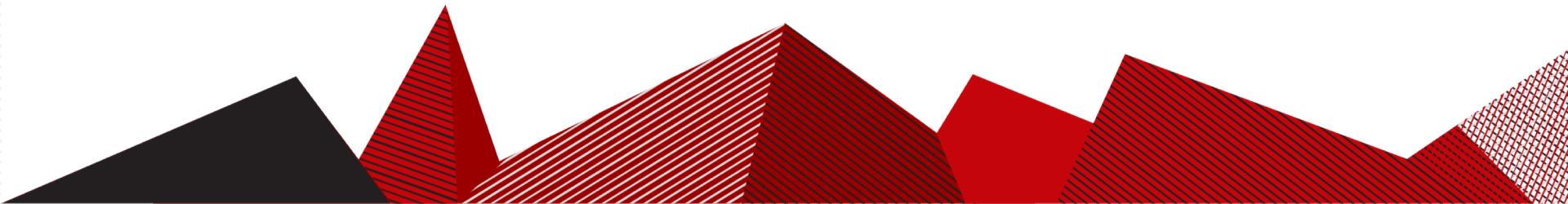
By Tianqi Li, Tiannan Huang and Siyu Wang

Tues Group3



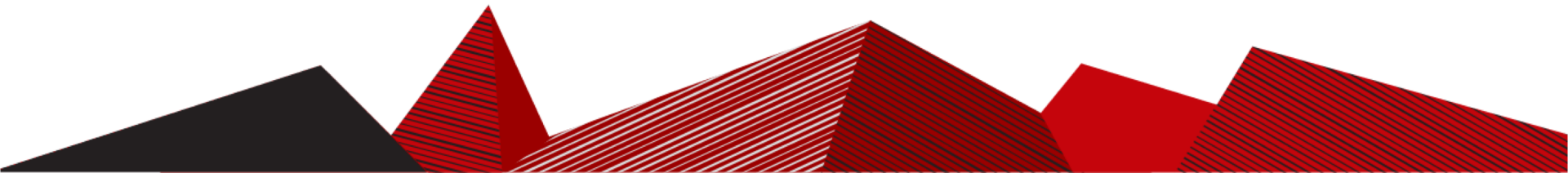
Outline of the Analysis

- Data Cleaning
- Model Selection
- Model Diagnostic

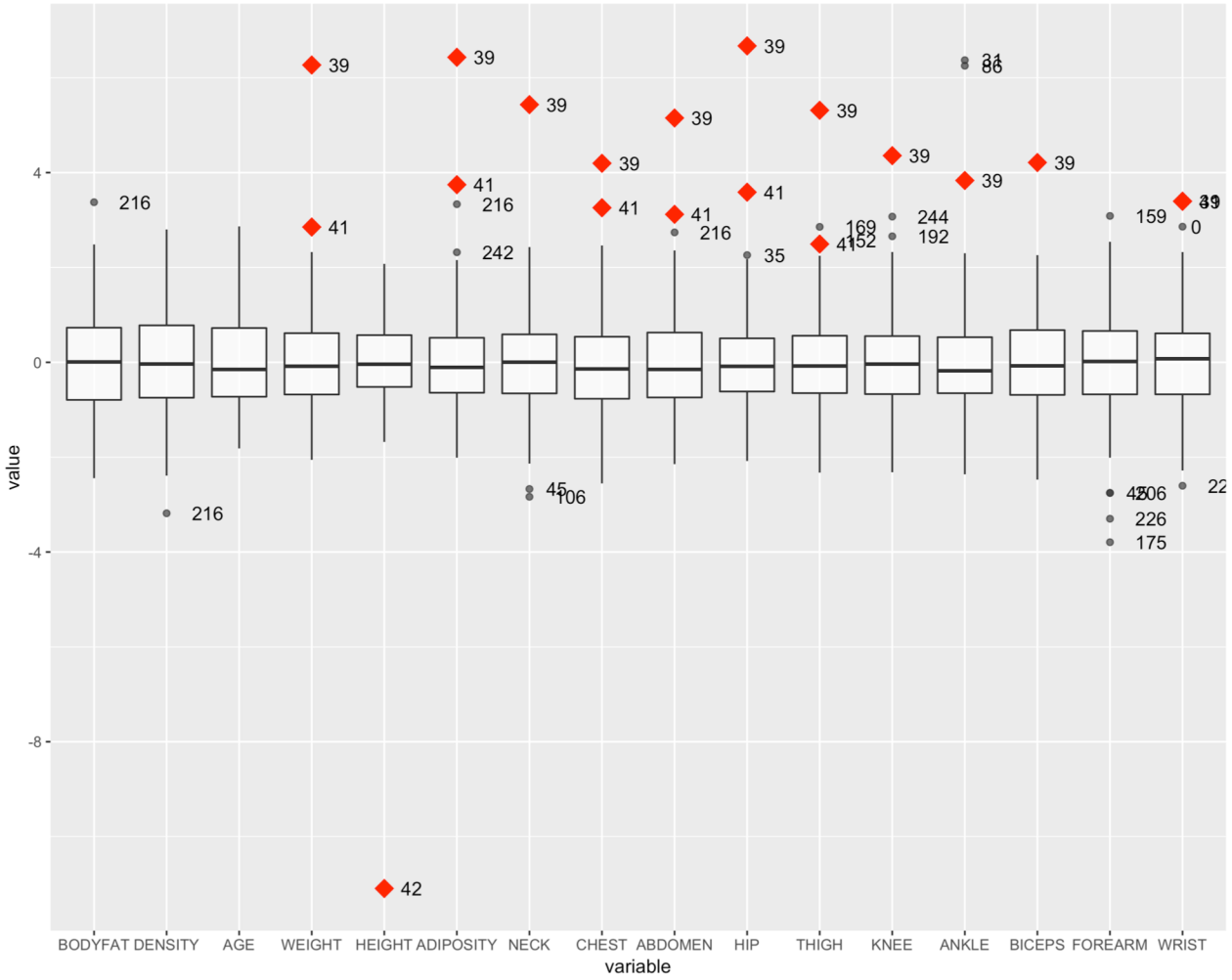


Data Cleaning

- Boxplot
- BODYFAT Calculation
- BMI Calculation

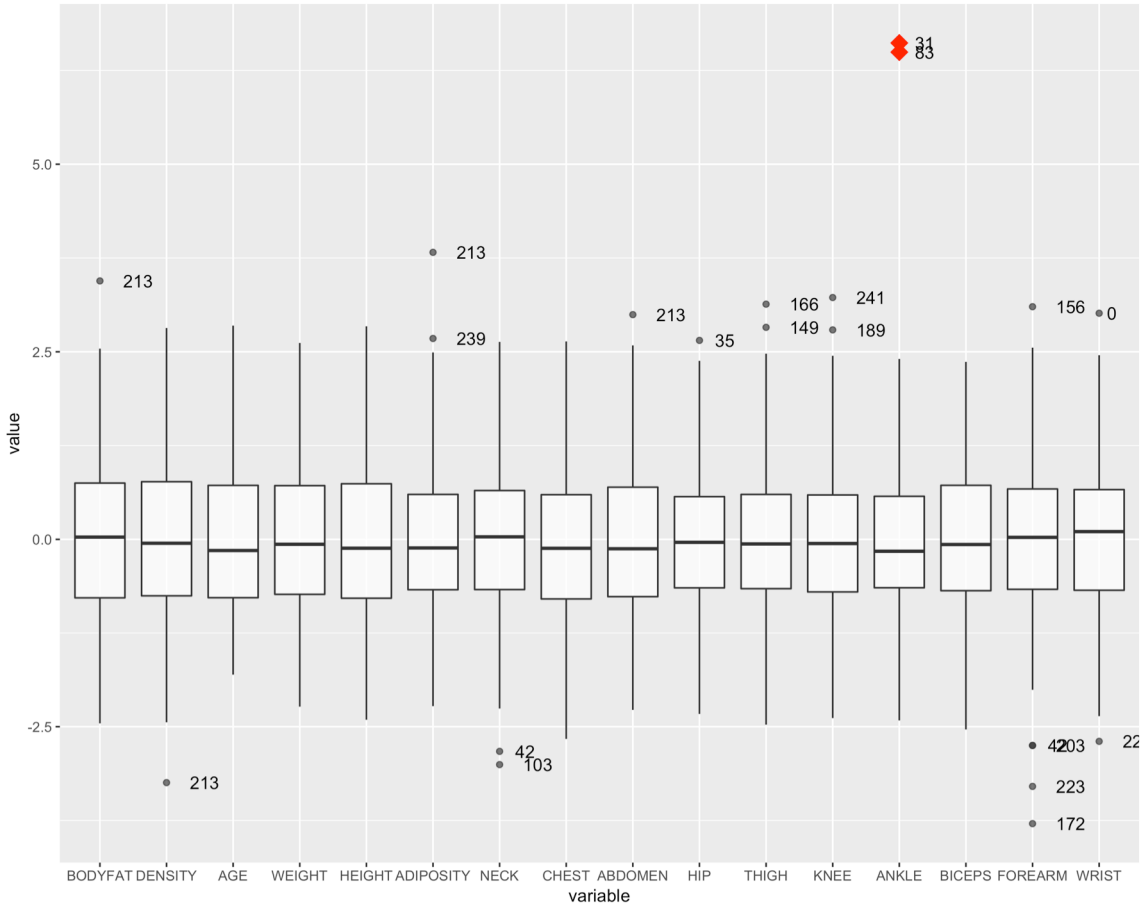


Boxplot 1


value
variable
BODYFAT
DENSITY
AGE
WEIGHT
HEIGHT
ADIPOSITY
NECK
CHEST
ABDOMEN
HIP
THIGH
KNEE
ANKLE
BICEPS
FOREARM
WRIST
4
0
-4
-8
39
41
216
242
42
45
106
35
169
244
192
159
226
175
0

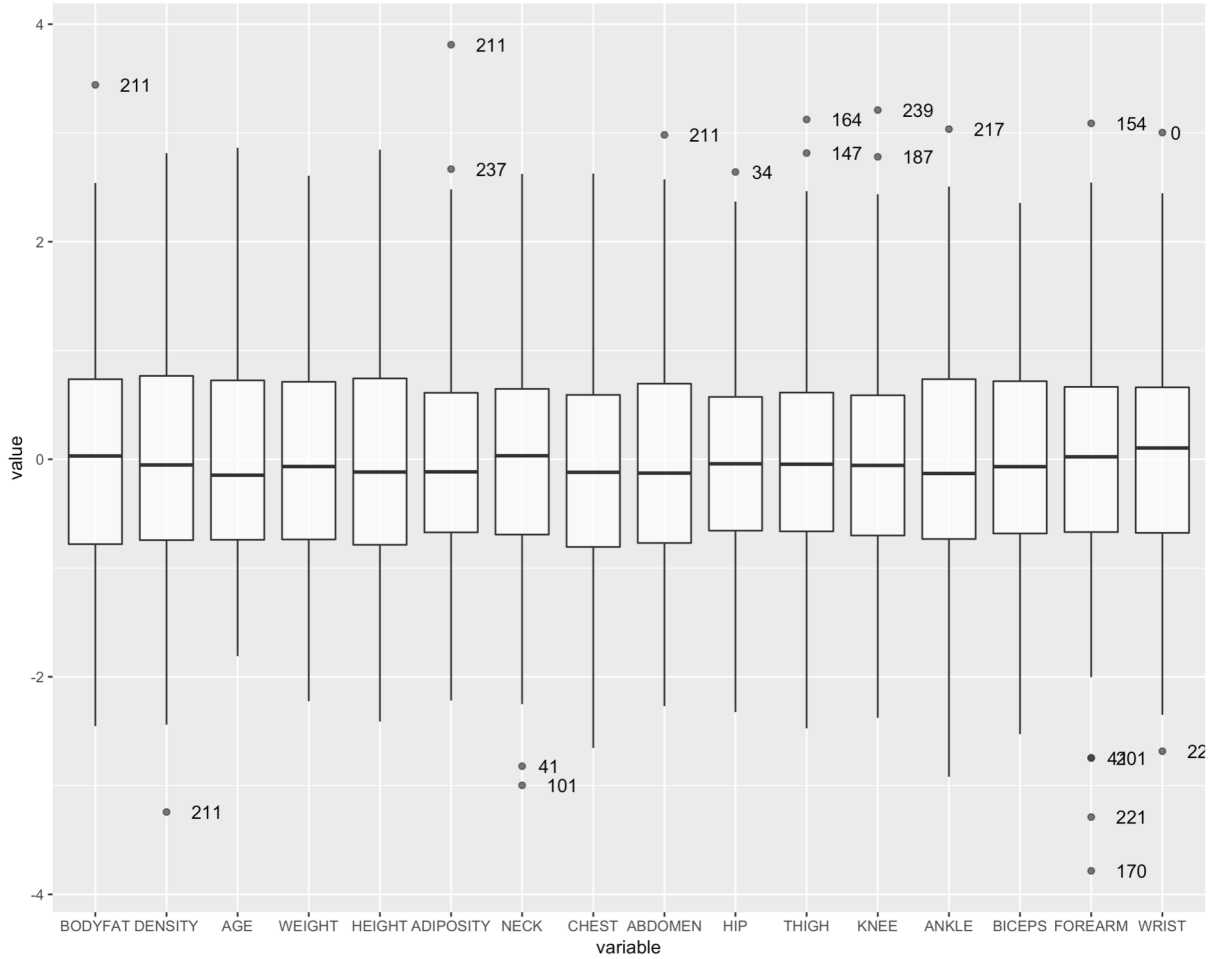
Boxplot 2

- After removing point 39, 41 and 42



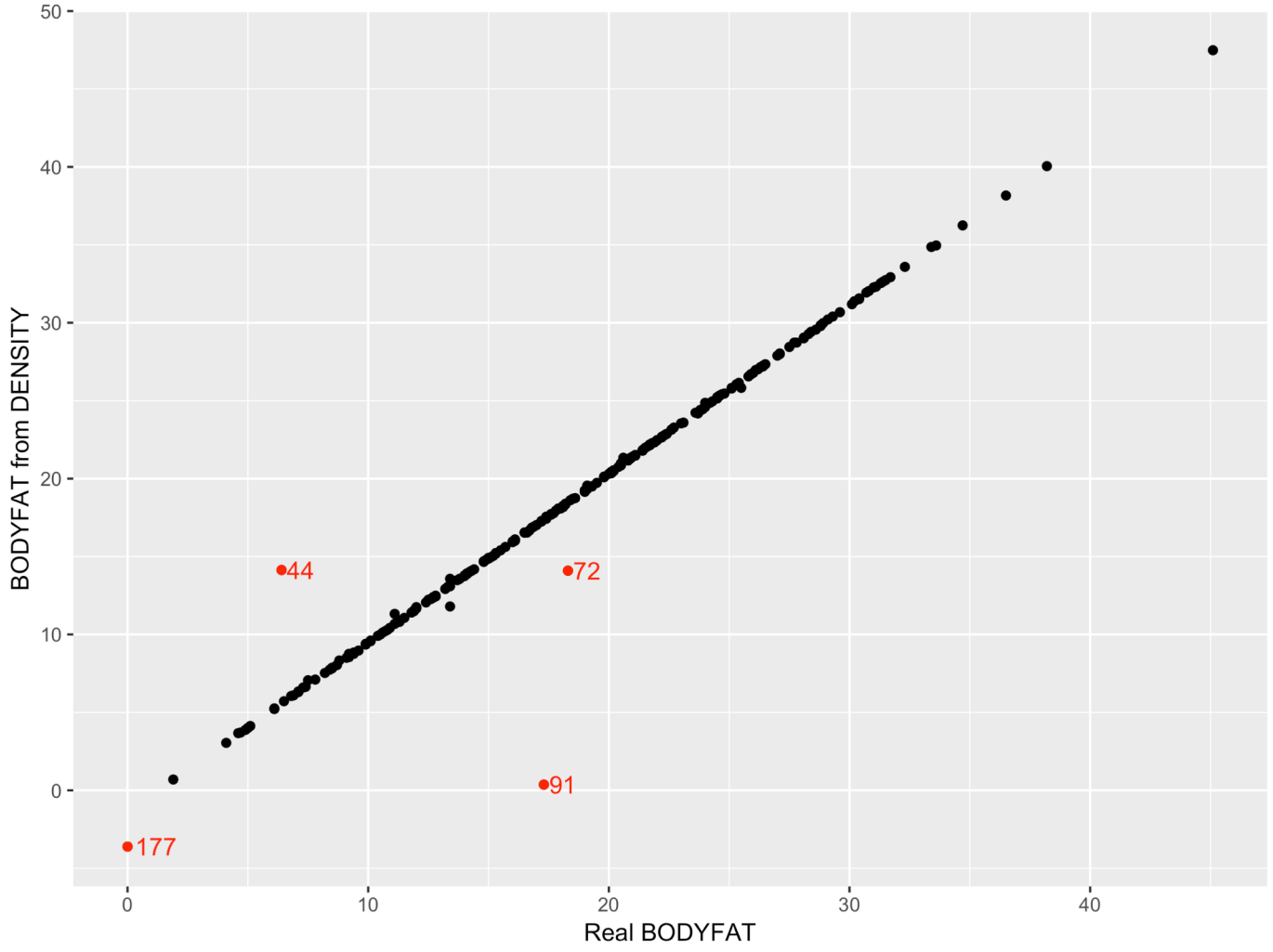
Boxplot 3

- After removing point 31 and 83



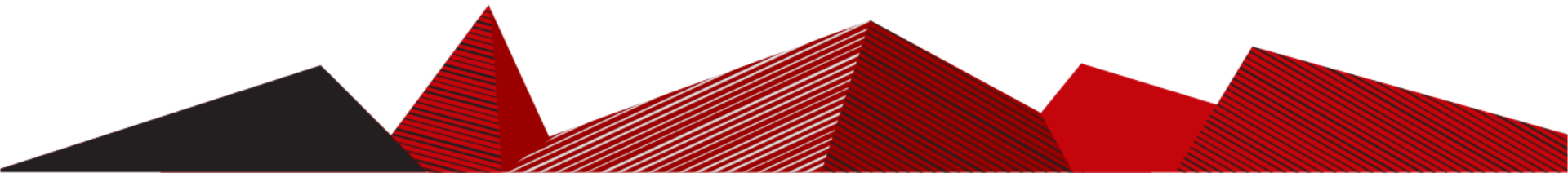
BODYFAT Calculation

- BODYFAT = 495 / DENSITY - 450



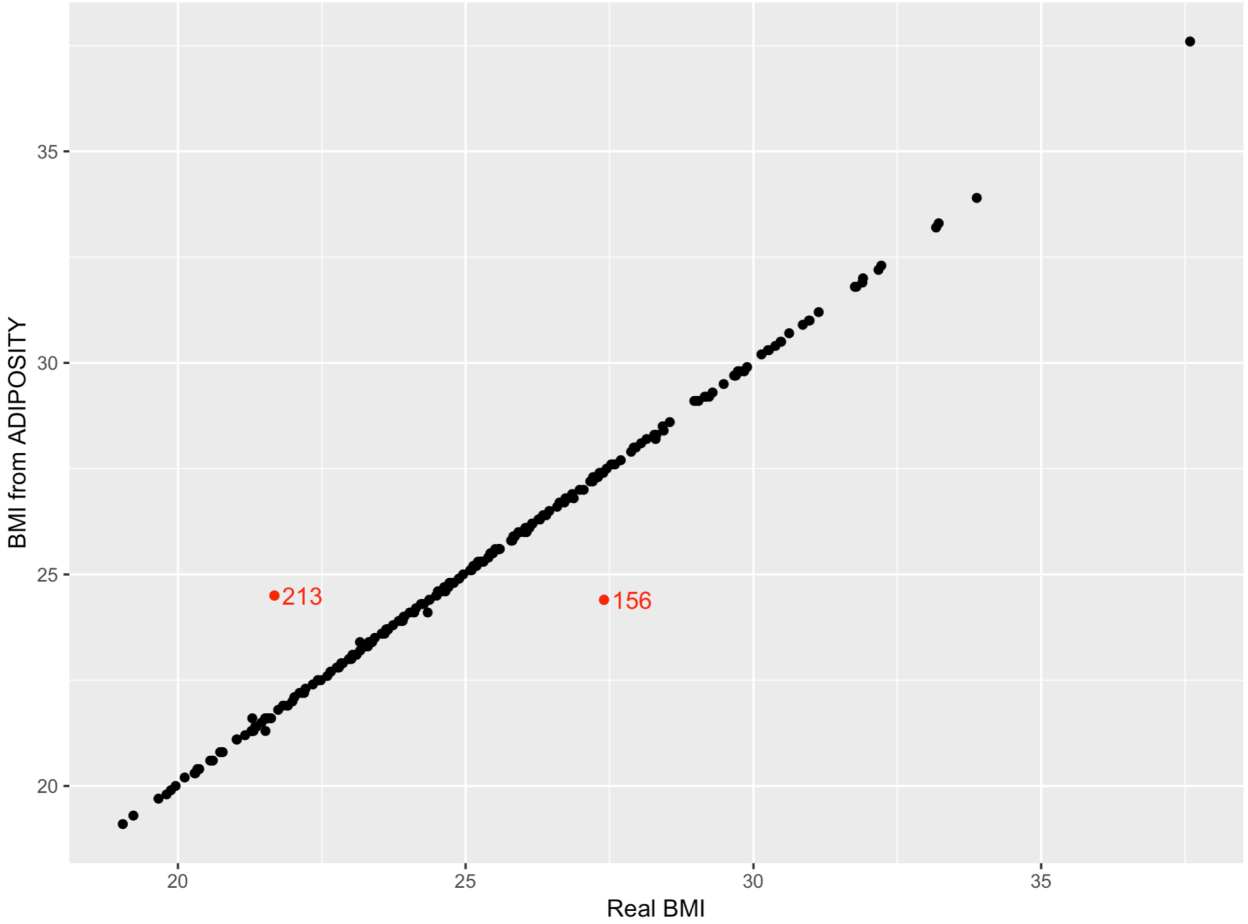
Linear Regression Prediction

	44	72	91	177
Real BODYFAT	6.40000	18.30000	**17.3000000**	0.000000
lm Prediction	10.00727	12.54937	**16.3742795**	5.736091
BODYFAT from DENSITY	14.13502	14.09151	**0.3684833**	-3.611687



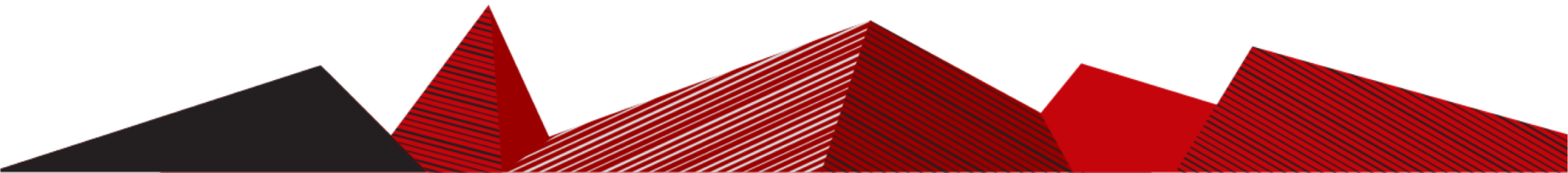
BMI Calculation

- ADIPOSITY(BMI) = 703 * WEIGHT / HEIGHT ^ 2



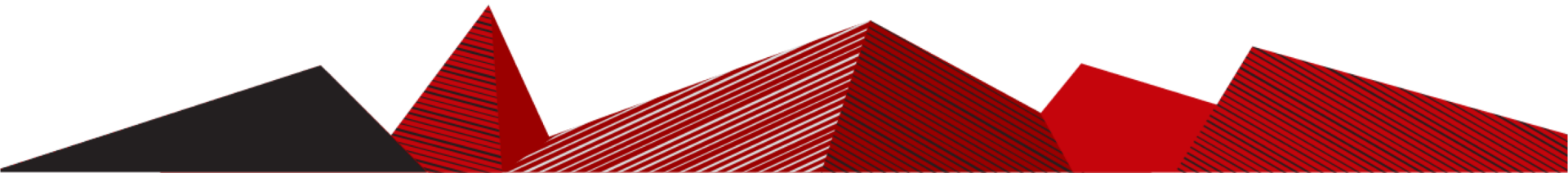
Linear Regression Prediction

	156	213
Real BMI	27.40422	21.67592
lm Prediction	16.00850	18.72592
BMI from ADIPOSITY	24.40000	24.50000



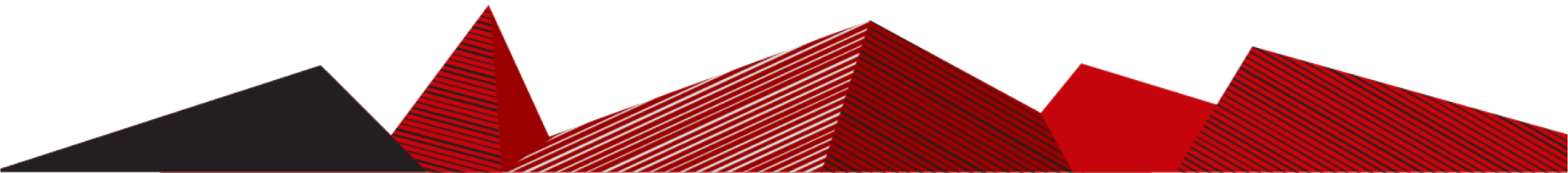
Final Data

- We also delete column DENSITY
- It has 242 rows and 15 columns



Variable Selection

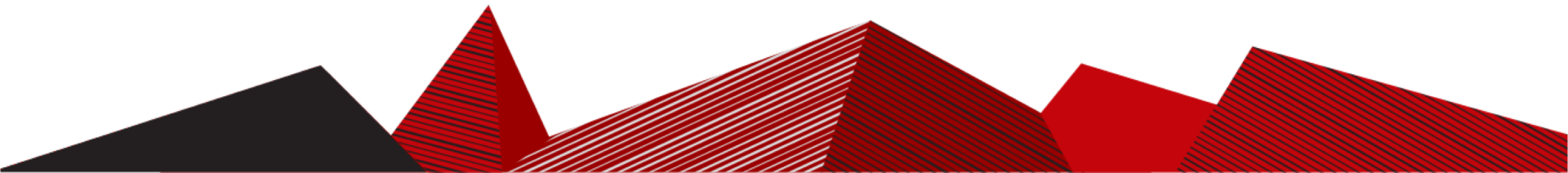
- Stepwise variable selection
- Direction: both side
- AIC as the criterion



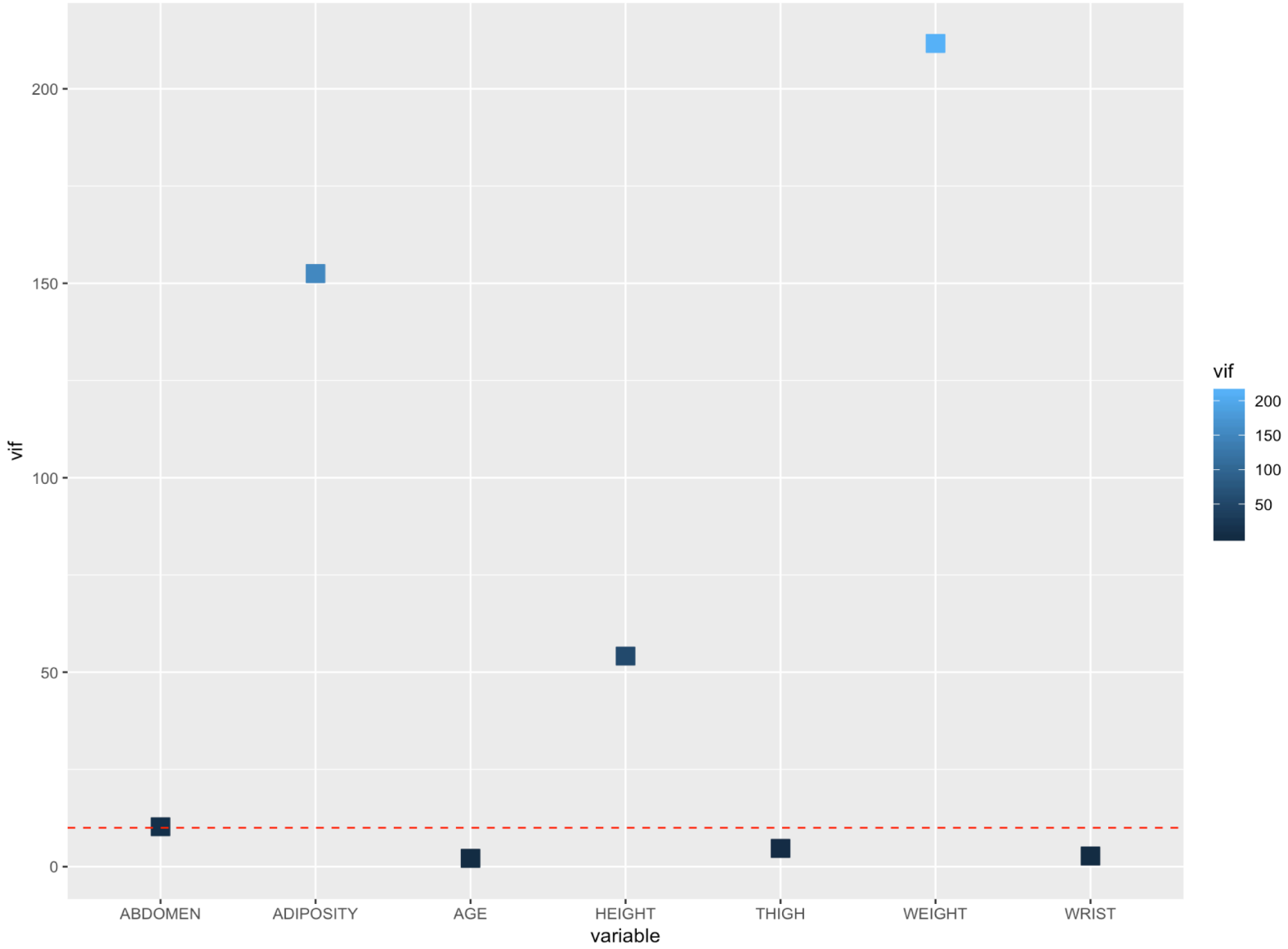
BODYFAT ~ AGE + HEIGHT + ABDOMEN + WRIST + ADIPOSITY + THIGH + WEIGHT

Not simple!

Collinearity problem!



VIF Plot


vif
200
150
100
50
0
ABDOMEN
ADIPOSITY
AGE
HEIGHT
THIGH
WEIGHT
WRIST
variable
vif
200
150
100
50

Models with Single Predictor

We consider 3 kinds of Model

1. Predictor: only one variable

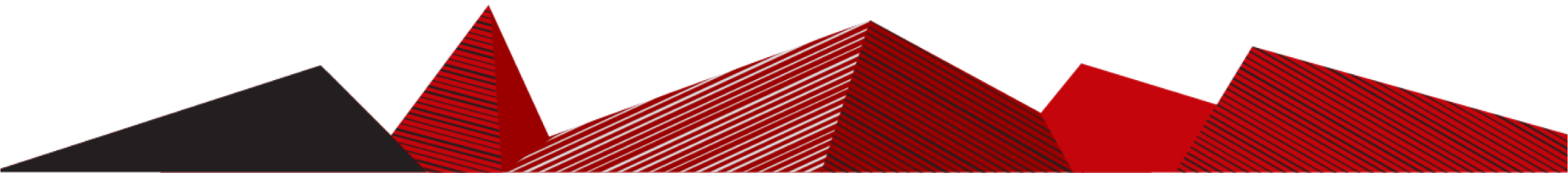
$$Y \sim X$$

2. Predictor: product of two variables

$$Y \sim X1 * X2$$

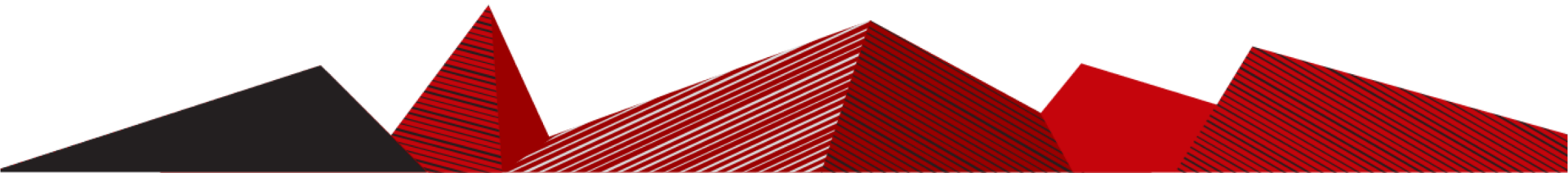
3. Predictor: ratio of two variables

$$Y \sim X1 / X2$$



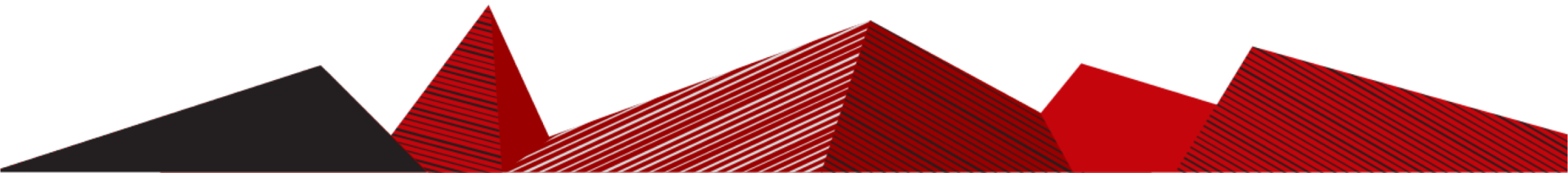
Models with Lowest MSE(under 3 forms)

Model Form	Best Model's predictors	Lowest MSE of CV
Y ~ X	ABDOMEN	19
Y ~ X1 * X2	ADIPOSITY * ABDOMEN	22
Y ~ X1 / X2	ABDOMEN / HEIGHT	17



Comparison of Two Models

Model	R^2
BODYFAT ~ ABDOMEN / HEIGHT	71%
BODYFAT ~ AGE + HEIGHT + ABDOMEN + WRIST + ADIPOSITY + THIGH + WEIGHT	73%

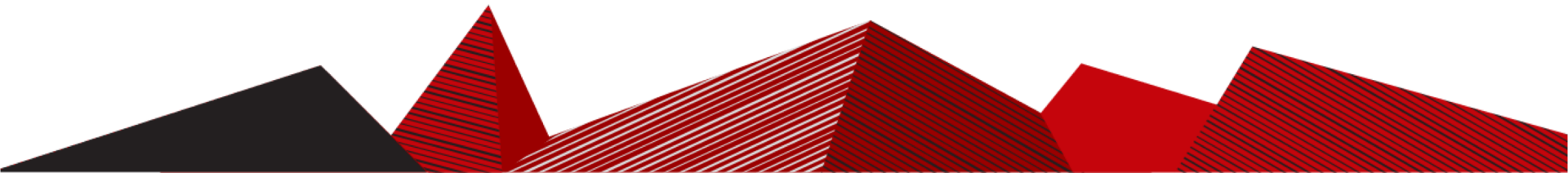


Final Model

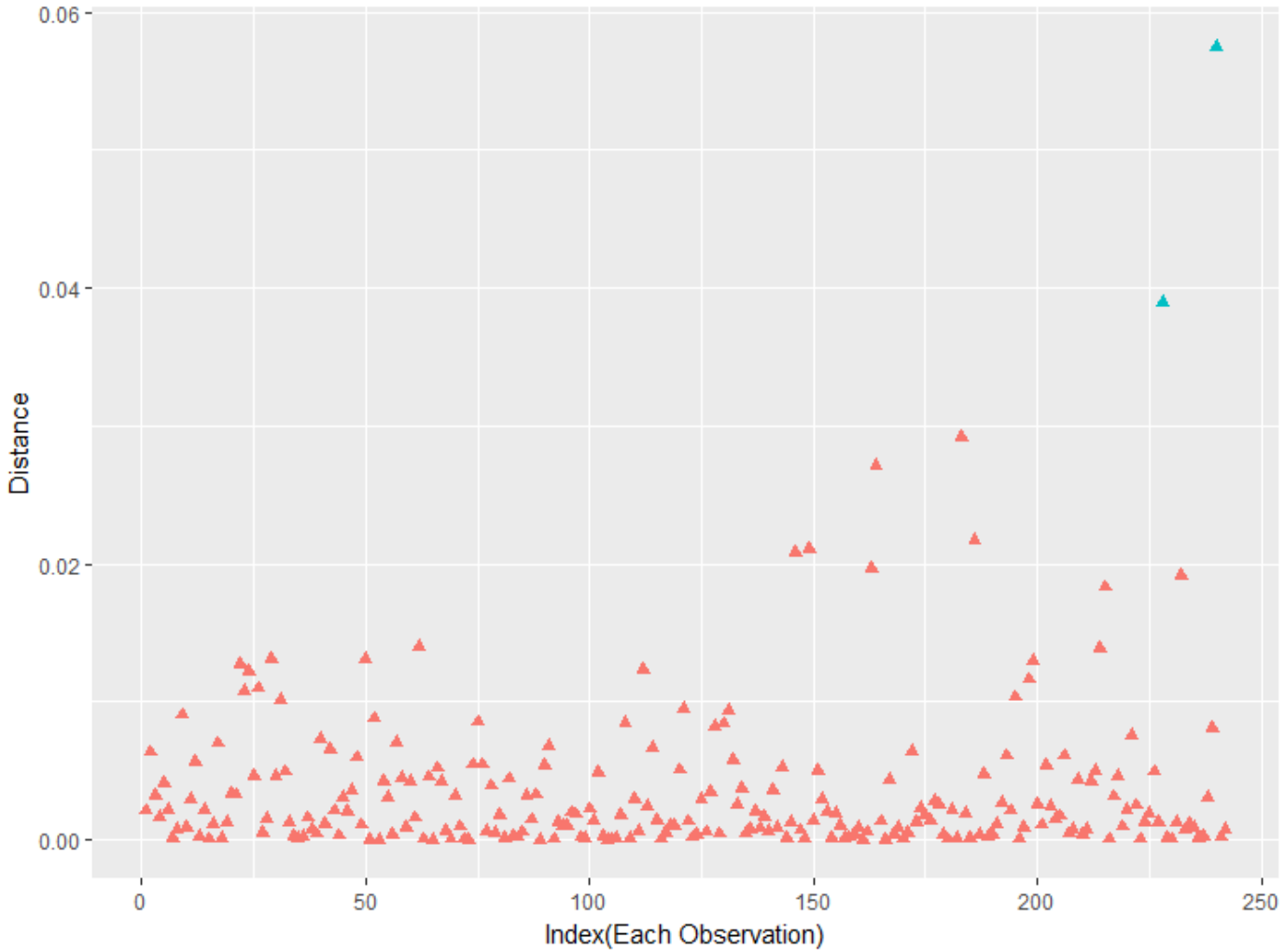
$$\widehat{BODYFAT} = 44.72 \times \frac{ABDOMEN}{HEIGHT} - 39.82$$

Rule of Thumb

Divide your abdomen circumference(cm) by your height(inches), then multiple 45 and then minus 40



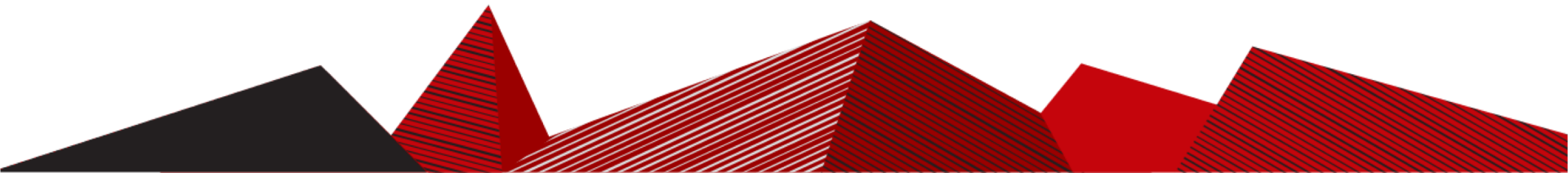
Influential values (Cook's distance)



Recheck these 2 point:

	Point 228	Point 240	1st Qu	mean	3rd Qu
Height(inch)	69.5	66.0	68.3	70.1	72.3
Abdomen(cm)	113.8	111.5	84.8	92.4	99.2

Since these two most "strange" points have reasonable data, we consider all the points could be involved in our model.



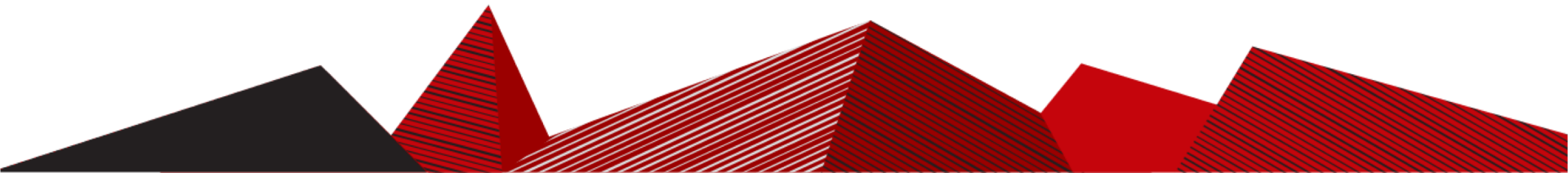
Diagnostics

1. Linearity:

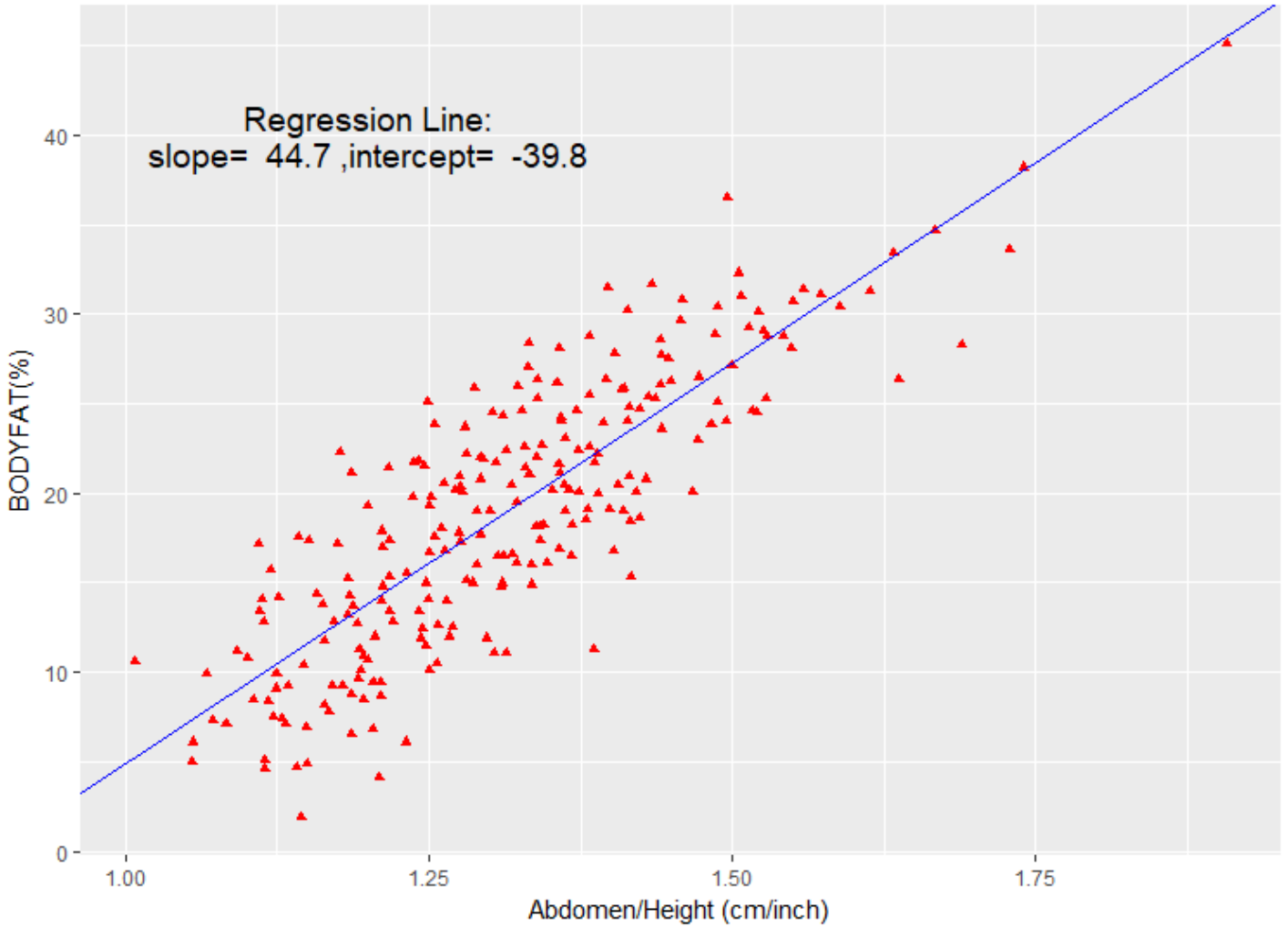
   Scatter plot with the regression line
2. Constant variance:

   Residual plot
3. Normally distributed errors:

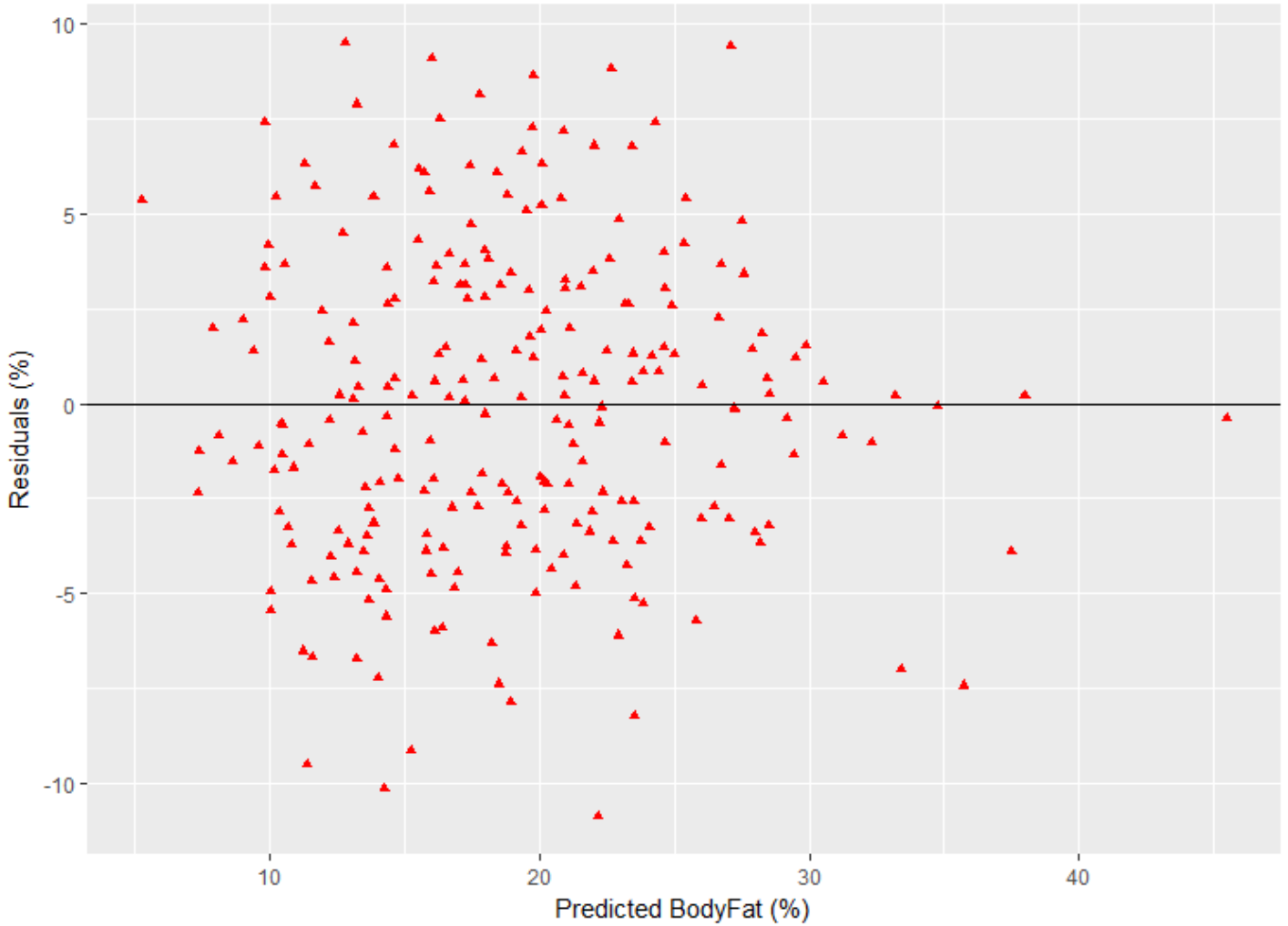
   Normal QQ plot



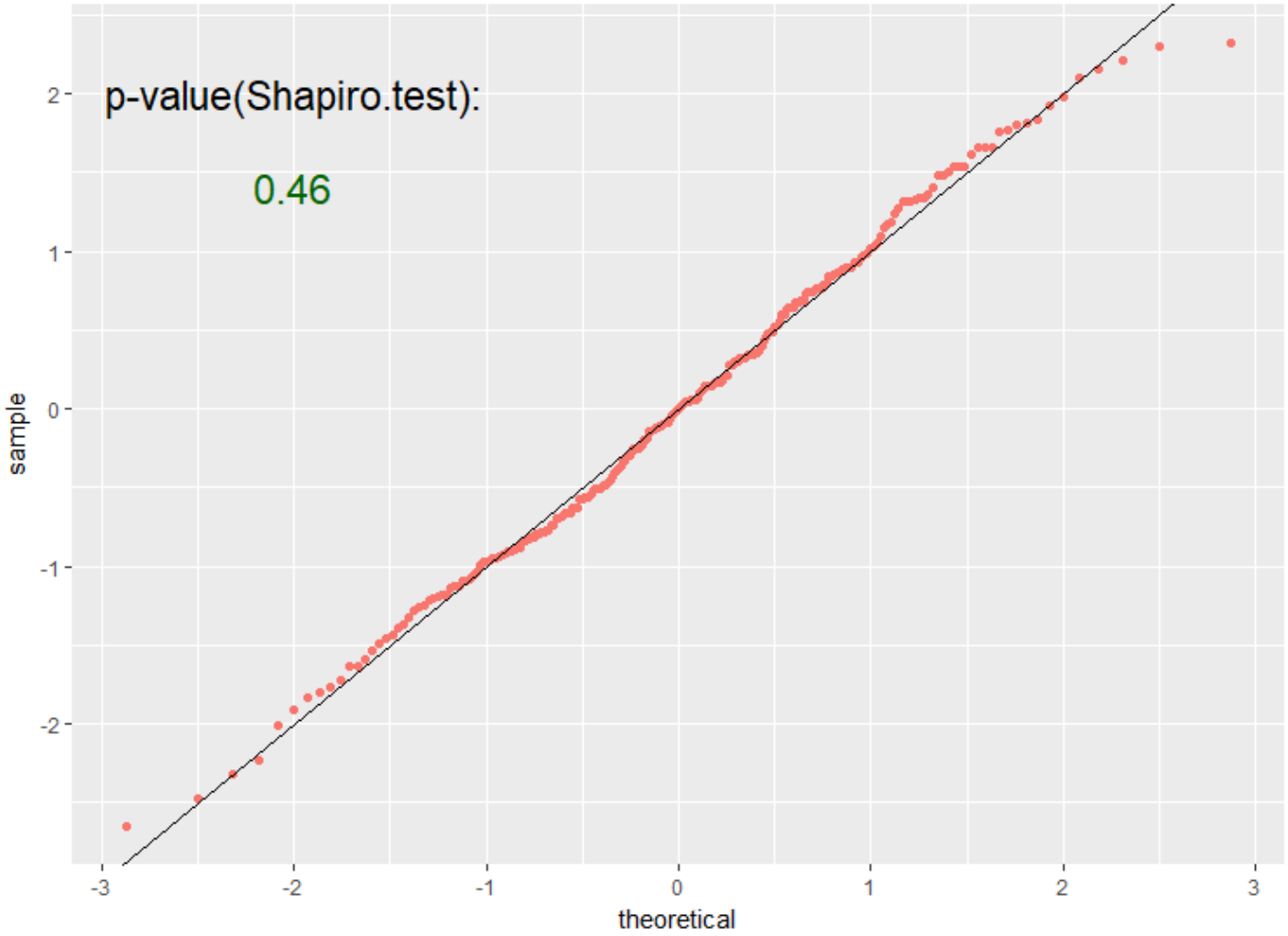
Scatter plot for BodyFat % vs Abdomen/Height



Residual plot



Normal QQ plot



Strength and Weakness of Model

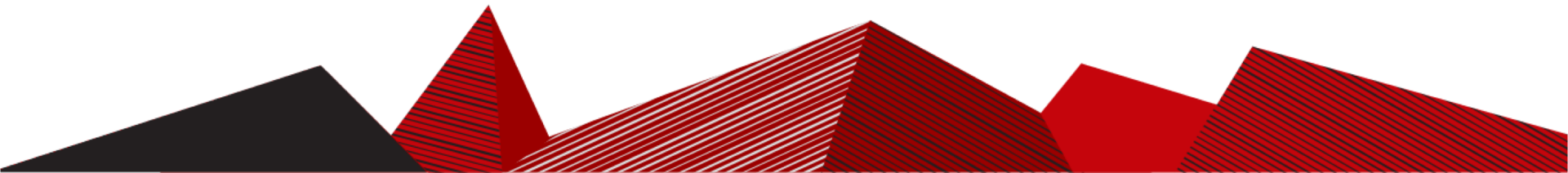
- Data cleaning:

  Seems completed
- Simplicity:

  2 variables, 1 coefficient, 1 intercept
- Robustness:

  Remain robust to different body-sized men
- Accuracy:

  Could improve if adding more variables into model



Summary

- There is almost surely a linear relationship between BodyFat % and Abdomen / Height, this is significant at the $\alpha=0.05$ level (with p-value $<2*10^{(-16)}$). And, Abdomen / Height explains about 71% of all the variation in BodyFat %, so this model is considered simple and accurate.
- According to this model, applied to people with a 70 inch height (the mean of data), a 1.57 cm increase in Abdomen may lead to a 1% increase in BodyFat %.