Body Fat Data Analysis

Presented by Kangyi Zhao, Xinyu Zhang, Kehui Yao

University of Wisconsin Madison

2019-02-03

WISCONSIN

UNIVERSITY OF WISCONSIN–MADISON

Background

- Purpose: Estimate bodyfat percentage of male.
- Principle: Simple, robust, and interpretable.
- Dataset:

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

  - 252 observations.
  - 2 response variable: BODYFAT (DENSITY).
  - 14 predictive variables such as AGE, WEIGHT, and etc.
  - Precision and units: WEIGHT (0.25 lbs), HEIGHT (0.25 inches), Abdomen circumference(0.1 cm) and etc.

Outlier Overview

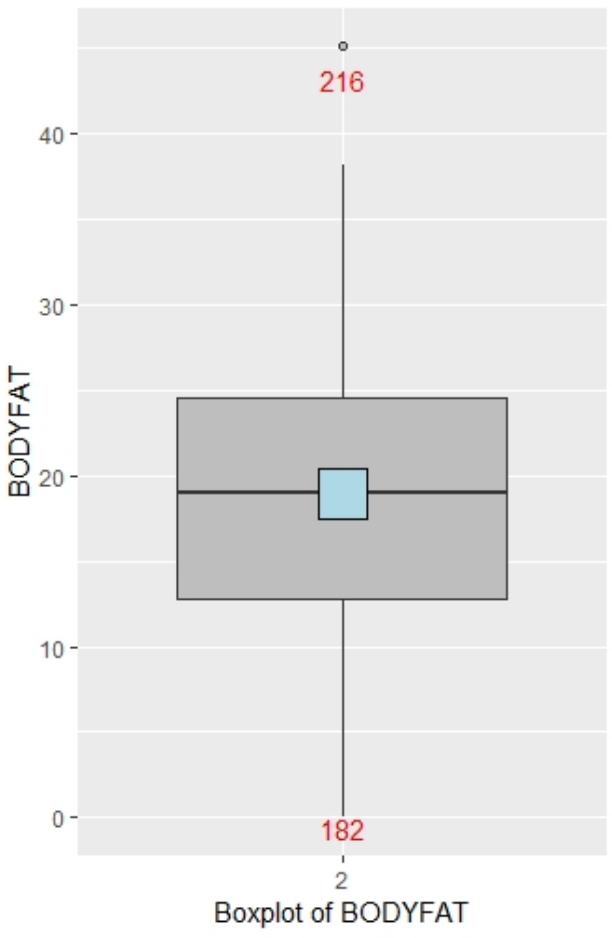
- Prior Knowledge:
  - Relation 1: bodyfat is highly correlated with density
  - Relation 2: adiposity is highly correlated with weight and height

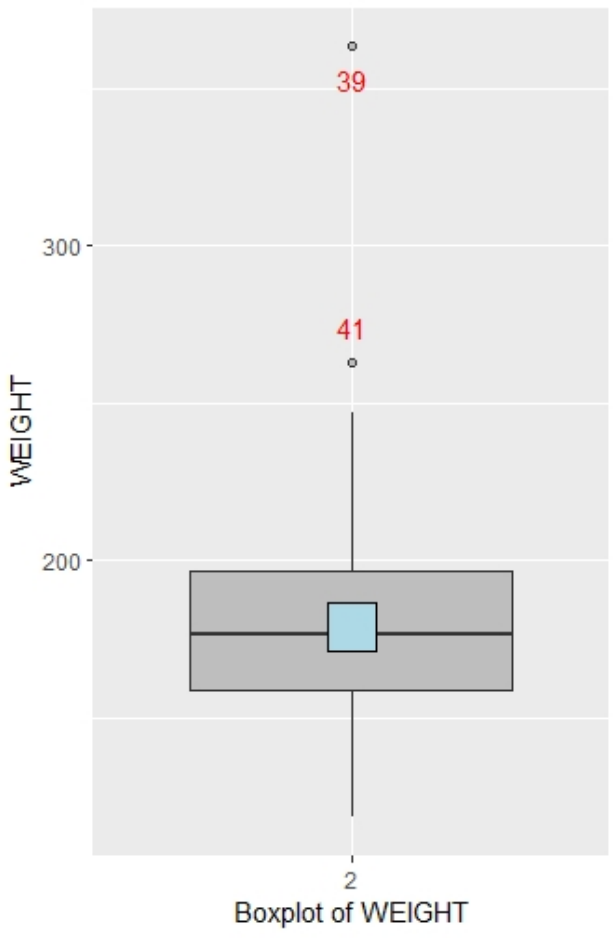
$$adiposity(bmi) = \frac{weight(lbs) \times 703}{height(in^2)} - 450$$

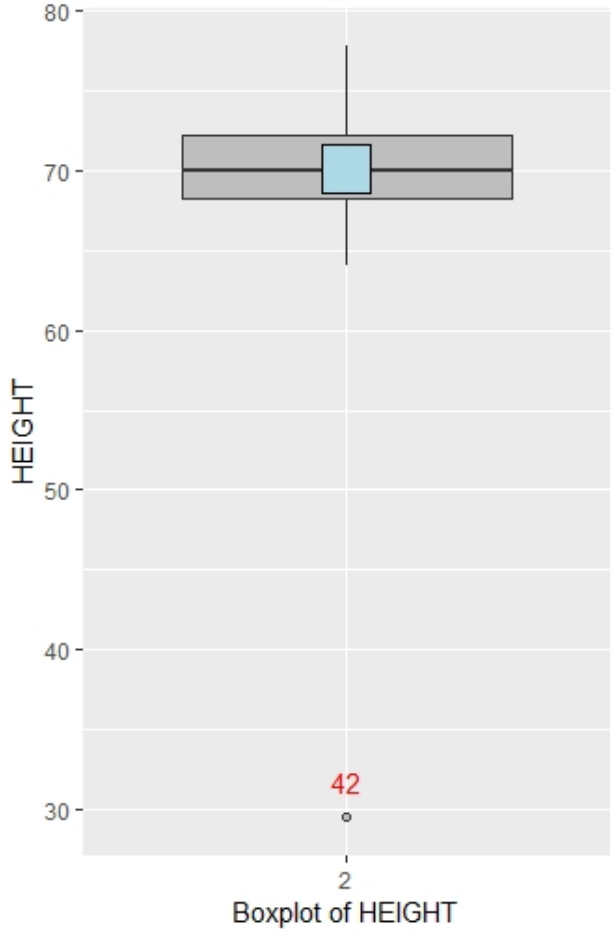
- Procedure:
  - Outlier Detection: boxplot() & summary()
  - Check Relations.
  - Analyze each outlier.
- Method:
  - Overall distribution.
  - Prediction bias based on prior knowledge.
  - Analysis based on specific subsets.

Outlier Dectection

Five outliers based on three variables stand out.

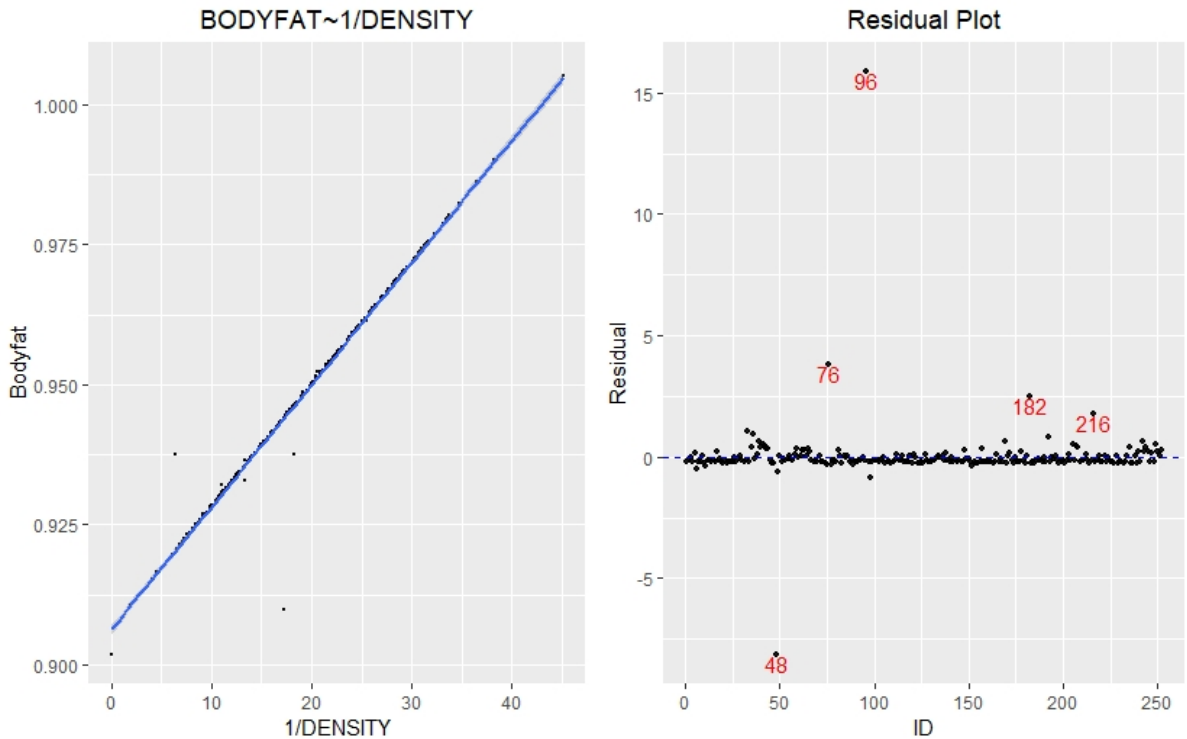






Check Bodyfat on Density

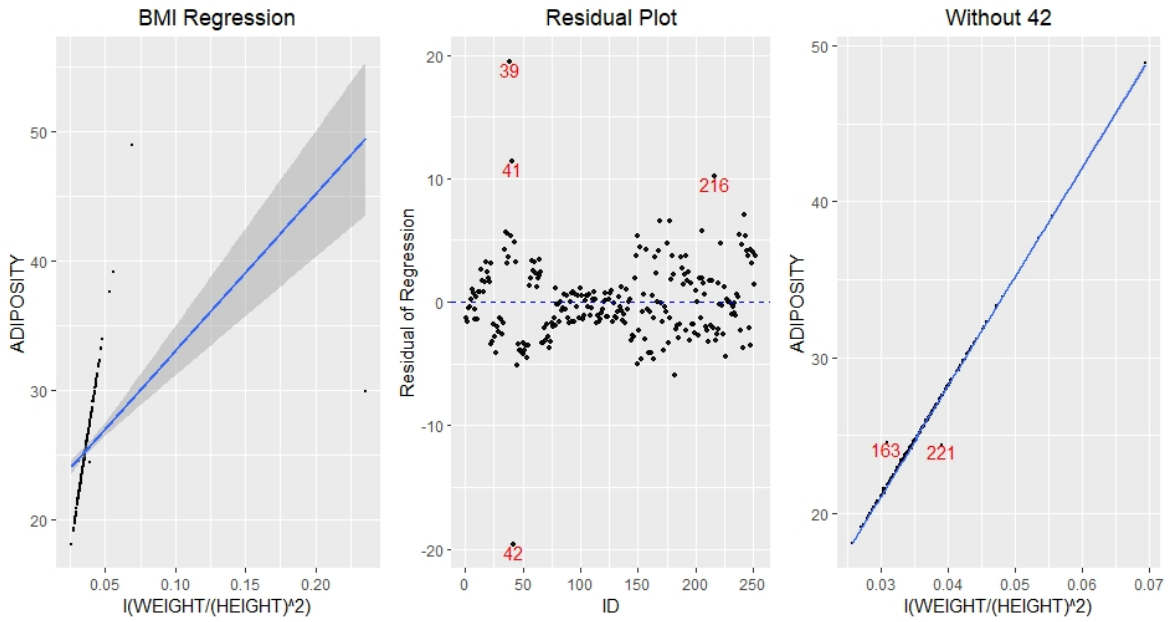
Based on the regression model, we obtain a list of possible outliers marked as red in the graph.



- No imputation for response variable.
- Delete any records verified as incorrect.

Check Height and Weight on Adiposity

Based on the following formula
We did similar linear regression for adiposity. We can then correct the unreasonable records according to the residual plot.



- Correct unreasonable records if able to identify specific error.
- Otherwise, delete any records which are verified as incorrect and are unable to impute correctly.

Glimpse at Outliers

- Delete 163: unable to impute correctly.
  Uncertained of incorrect records: weight, height, or adiposity

	WEIGHT	HEIGHT	ADIPOSITY
174	176.25	71.50	24.3
163	184.25	68.75	24.4
84	170.75	70.00	24.5

- Impute 42's Height: attention to precision units.

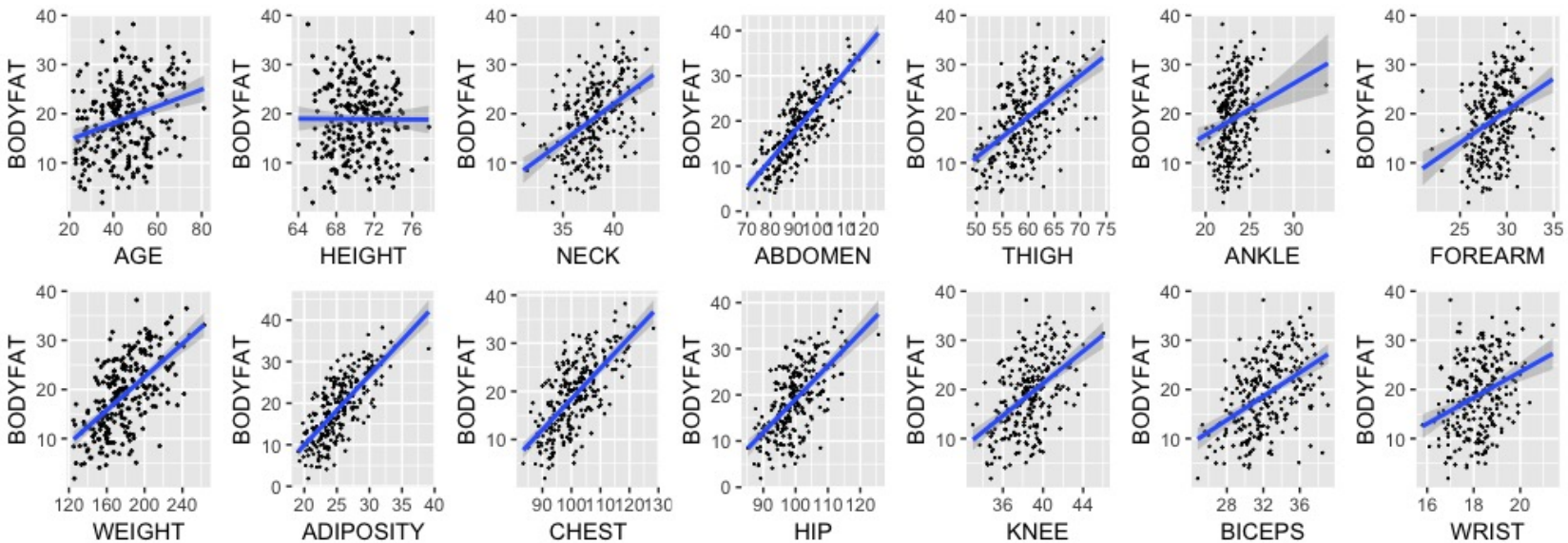
	Record	Prediction	Imputation
42	29.5	69.48	69.5

Remove outliers

- Delete Extreme values:
  - Weight: delete ID 39.
  - Bodyfat: delete ID 182 and 216.
- Delete Incorrect records:
  - Bodyfat: delete ID 48 and 76.
  - Weight and height: delete ID 163 and 221.
- Impute based on adiposity:
  - Height: ID 42.

Variable Selection with Different Models

After data cleaning, we regress bodyfat on each predicted variables. We can see that all the variables show the linear tendency, and some variablies might have multicollinearity, so we try Lasso regression at first,and then use the Mallow's Cp, BIC forward and backward methods to double check our model.



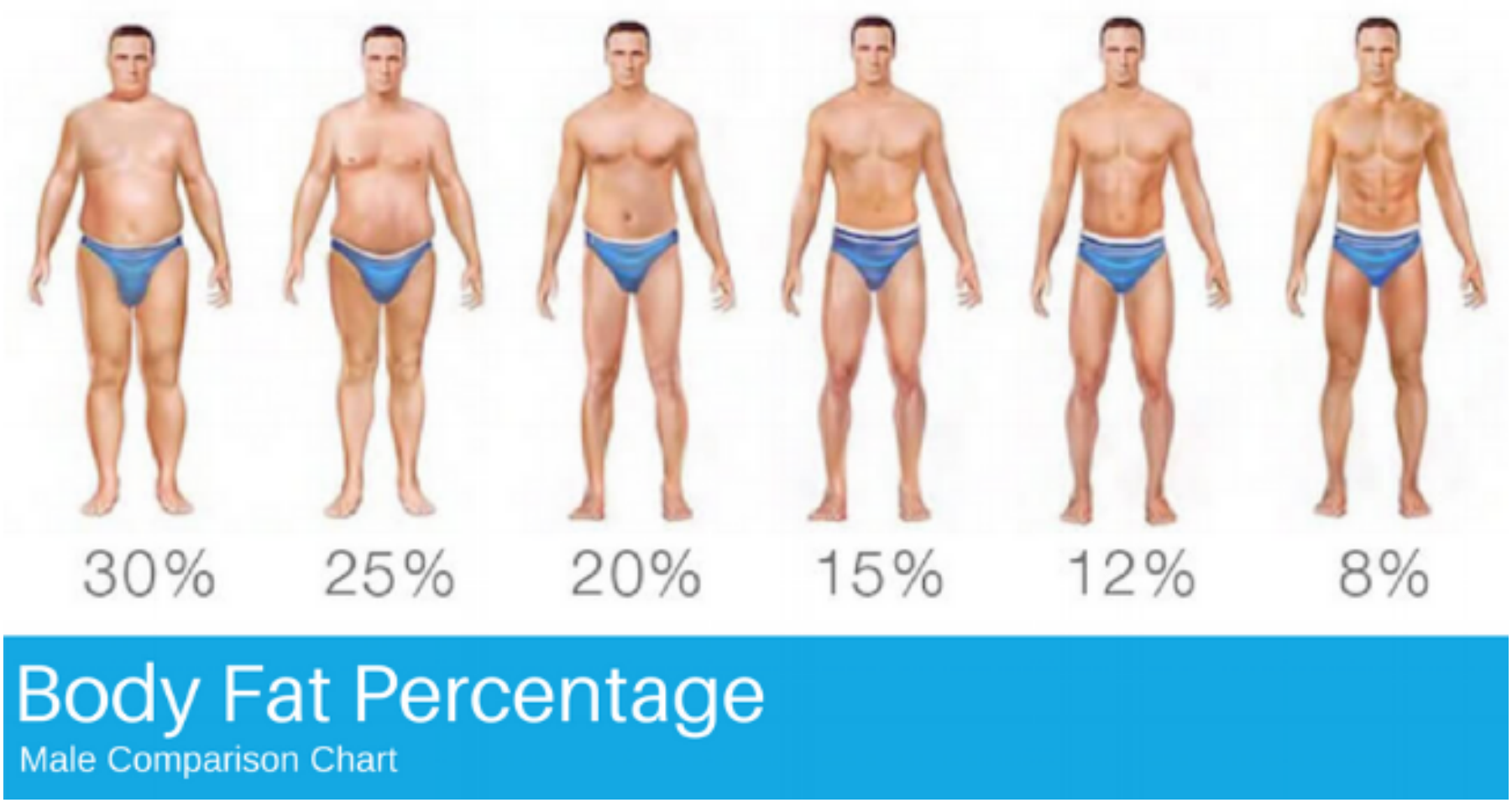
Variable Selection with Different Models

The following table shows the result of variable selection with different models.

Method	ABDOMEN	WRIST	HEIGHT	WEIGHT	AGE	R-squard
Lasso-1	0.50	0	0	0	0	0.641
Lasso-2	0.55	0	-0.18	0	0	0.671
Lasso-4	0.65	-1.14	-0.25	0	0.03	0.717
Lasso-all	-	-	-	-	-	0.739
Mallow's Cp-2	0.72	-2.05	0	0	0	0.704
Mallow's Cp-7	-	-	-	-	-	0.730
BIC forward-3	0.87	-1.34	0	-0.08	0	**0.72**
BIC forward-2	0.89	0	0	-0.12	0	**0.707**
BIC backward-3	0.71	-2.18	0	0	0.07	0.717

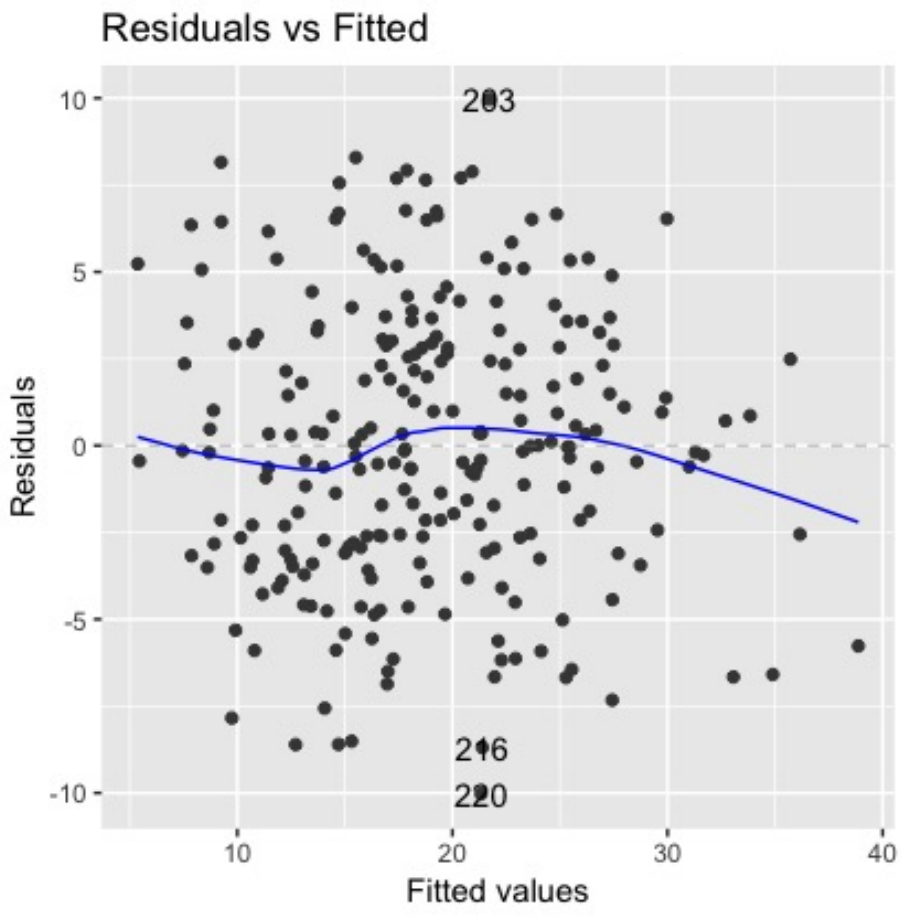
Model Explanation

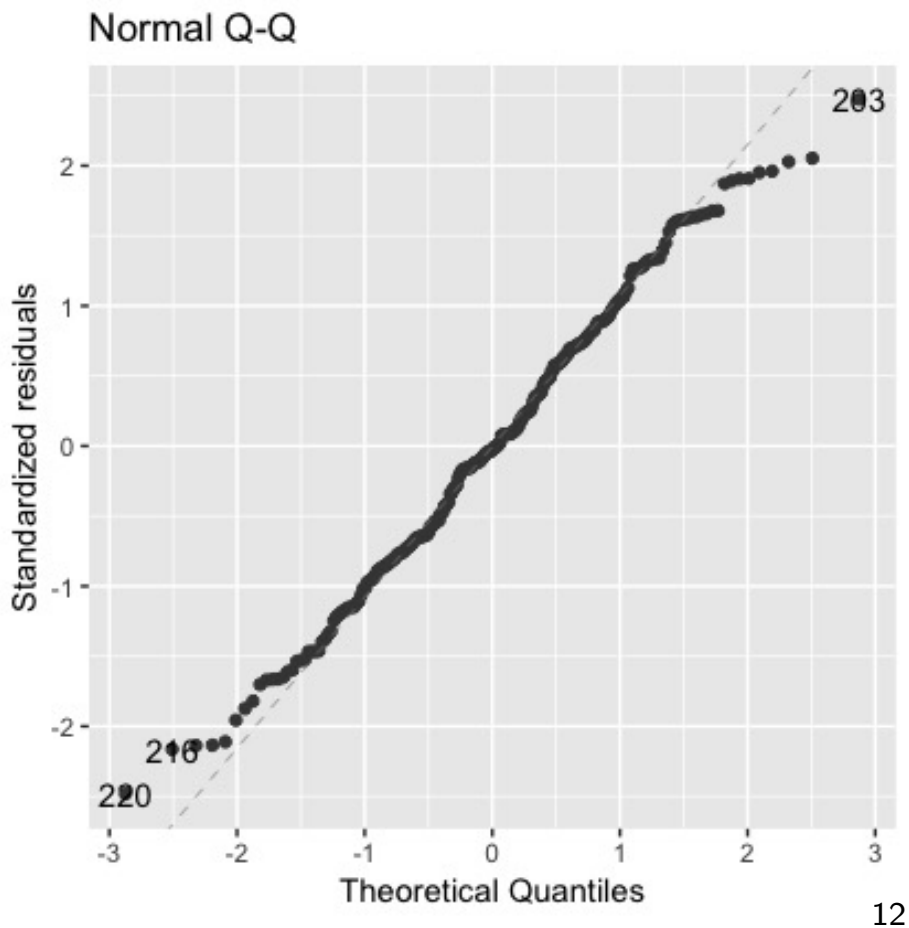
In BIC forward-2 method, we see that the abdomen and weight have a negative coefficient. How can we interpret that?



Model Diagnostics

The diagnostic plot shows no absolute pattern in the residual plot, which means the model can achieve the assumption of indenpendency. In the Normal QQ-plot, there appears to be almost a line, so the assumption of normality also can be satisfied.





Rule-of-Thumb Model

- We use R-square to select the model, and the final model we select can be written as:
  Bodyfat (percentage) = -41 (Constant) +0.9 ABDOMEN (cm) -0.1 WEIGHT (kg)
- Thus, our model shows that abdomen and weight are two most important variables when calculating the bodyfat percentage of male.

Take Home Message

How can this model predict the outlier? For example, the model will give a very small or even negative bodyfat prediction when it trys to predict a man with a large weight but small abdomen. Is there any remedy?