module2

Hao Qin

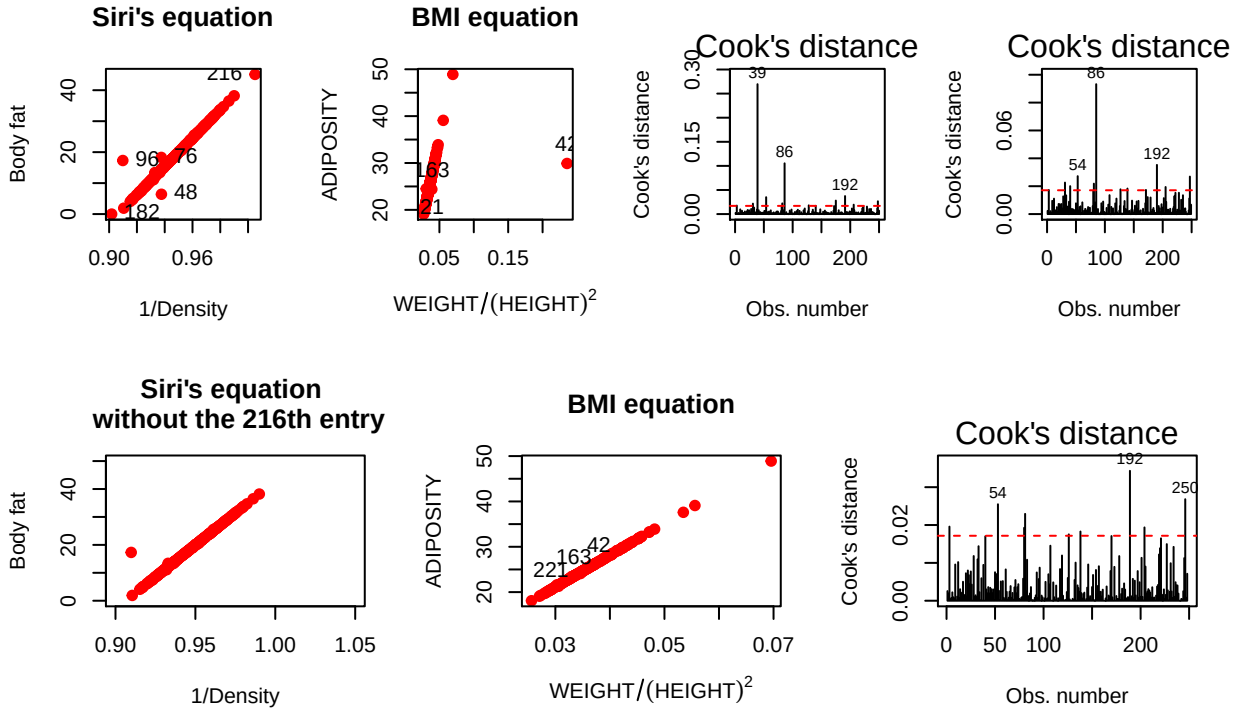
1.Cleaning data

2.Detection of the outliers and their treatment

We use three different types of method to detect not only the outliers, but also the influence points in the original dataset, which is Siri's equation, BMI formula and cook's distance. Those points will deviate from the line or the region on which most points are, thus, it is easy for us to detect them through graphics.

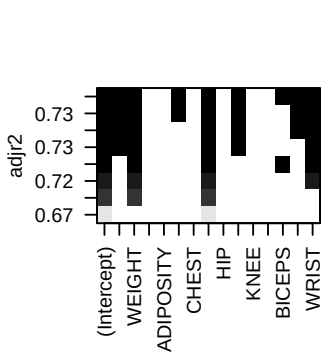
Table 1: Outliers and treatments

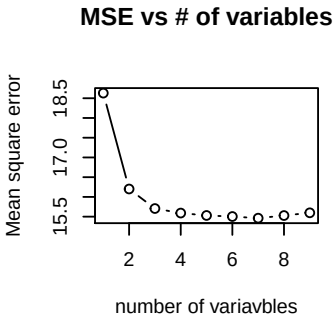
Outliers	Detection_method	Abnormal_Component	Treatment
39	Cook's distance	large Cook's distance	deleting
42	BMI formula	too short in height	imputation
48	Siri's equation	abnormal body fat	imputation
76	Siri's equation	abnormal body fat	imputation
86	influence test	deviating from the majority	deleting
96	Siri's equation	abnormal body fat	imputation
163	BMI formula	abnormal BMI index	imputation
182	Siri's equation	zero body fat	deleting
216	Siri's equation	density below than 1	deleting
221	BMI formula	lighter than normal	imputation

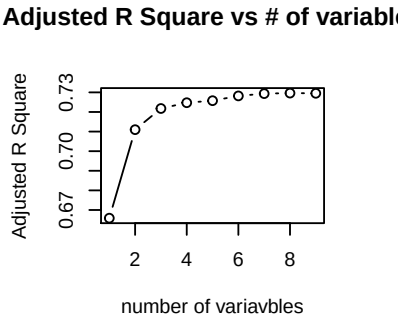


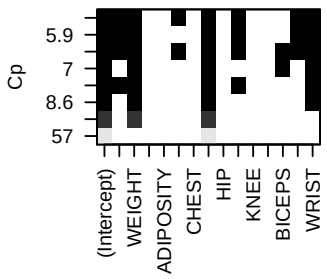
3.Selecting variables

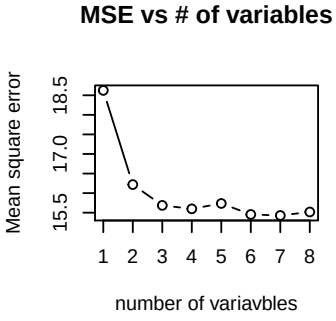
In this part, several methods have been applied, including stepwise aic and bic selection, lasso and group lasso, mallow's cp and Bess, which is a new proposed way to selecting variables. After applying these methods, since the total sample size is not large, we can use cross validation to measure the performance of each method, to decide which component should be treated as the independent variable in the final model. To be specific, the final model should be as a rule of thumb, which requires that the quantity of the independent variables should not exceed four.

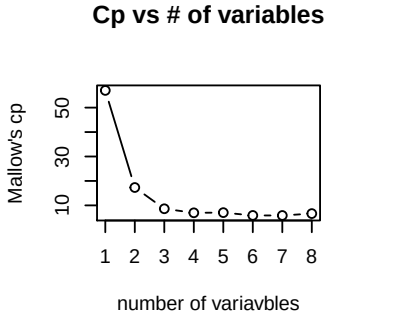




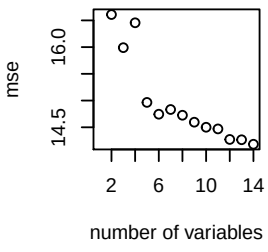


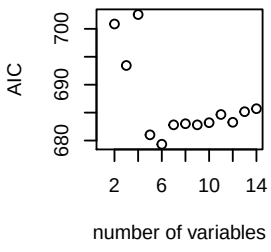


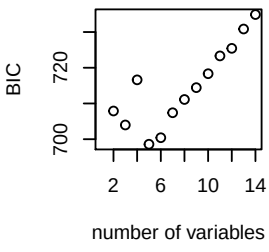





```
#lasso and group lasso
```







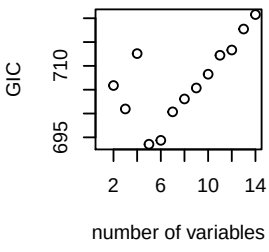


Table 2: Different methods with multiple choices of variables

	model	Mean_Sq	variables
12	Adjusted R square	15.59120	WEIGHT + ABDOMEN + BICEPS + WRIST
8	Mallow's Cp	15.59827	WEIGHT + ABDOMEN + BICEPS + WRIST
7	Mallow's Cp	15.68636	WEIGHT + ABDOMEN + WRIST
1	BIC	15.69973	WEIGHT + ABDOMEN + WRIST
11	Adjusted R square	15.70616	WEIGHT + ABDOMEN + WRIST
15	BeSS	15.93861	WEIGHT + ADIPOSITY + CHEST + ABDOMEN + WRIST
2	BIC	16.18662	WEIGHT + ABDOMEN
10	Adjusted R square	16.20085	WEIGHT + ABDOMEN
6	Mallow's Cp	16.21848	WEIGHT + ABDOMEN
14	BeSS	16.45513	ADIPOSITY + CHEST + ABDOMEN + HIP
13	BeSS	16.72340	HEIGHT + CHEST + ABDOMEN
5	Mallow's Cp	18.62434	ABDOMEN
9	Adjusted R square	18.63013	ABDOMEN
4	BIC	18.63312	ABDOMEN
3	BIC	35.20704	WEIGHT

4.Model building

From Table 2, there is no significant difference in Mean squared error among those different methods except the situation of one variable. Considering the rule of thumb, we decide to choose a model within two variables, which is also a not heavy sacrifice in accuracy. The variables we select is weight and abdomen, applied in the linear model.