Bodyfat Calculator

Hao Qin, Jiacheng Miao, Qiaoyu Wang, Yuhan Meng

1. Introduction

Nowadays, the physical health has become a serious concern in the modern society since excess storage of fat will lead to a variety range of diseases and highly leave up the risk of coronary heart disease. To measure the fat of person, we also have developed several methods, including BMI index, which is the quotient of weight divided by height squared. In this article, we develop a sample method to measure the bodyfat percentage of one person, which only requires few easily accessible variables.

2. Data Processing

2.1 Outliers Detection

The data we have is BodyFat.csv file which includes **252** persons physical data in total **17** variables. In order to use these data to build the model, we use three different types of methods 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.

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	37.3	21.9	32	27.4	17.1
2	6.9	1.0853	22	173.25	72.25	23.4	38.5	93.6	83	98.7	58.7	37.3	23.4	30.5	28.9	18.2
3	24.6	1.0414	22	154	66.25	24.7	34	95.8	87.9	99.2	59.6	38.9	24	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	1.034	24	184.25	71.25	25.6	34.4	97.3	100	101.9	63.2	42.2	24	32.2	27.7	17.7

2.1.1 Siri's equation formula: $Percentage\ of\ Body\ Fat = \frac{495}{Body\ Density} - 450$

2.1.2 BMI equation formula: $BMI = \frac{Weight}{Height^2}$

2.1.3 Data Diagnosis

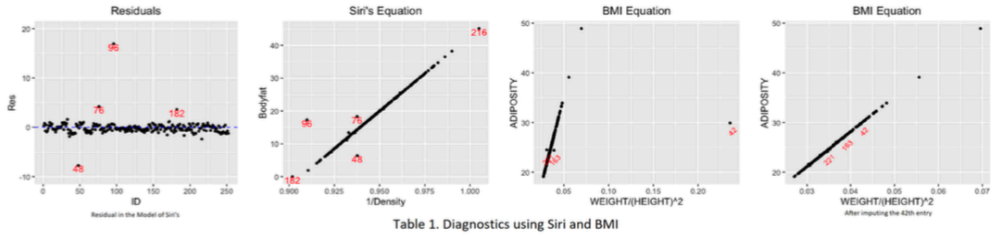


Table 1. Diagnostics using Siri and BMI

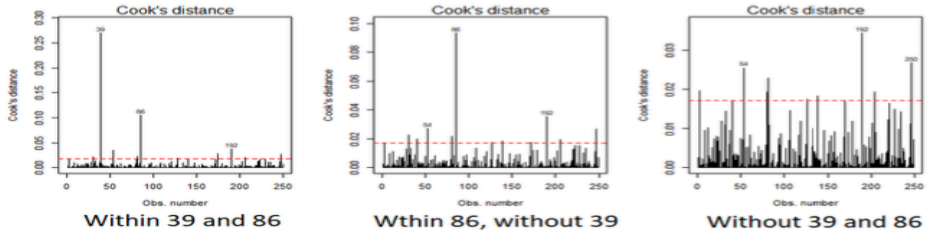
The criteria we follow in the treatment are three clauses:

Retain, **Impute** or **Exclude**. If the data have no abnormal variables comparing to the majority points, we keep it into the final model fitting process. On the contrary, if the data has been detected from one of above three methods, we assume that some components are gathering incorrectly and need to impute the wrong component to retain the imputation.

However, the if the imputed data does not reasonable neither, we have to delete it from the dataset. For the high influence point, we prefer to delete those points. One obstacle in front of us is to detect which component should be responsible for the outliers. We select several normal points to compare the abnormal one, which they have many similar components with the outlier, thus we can assume the component does not match with each other should be corrected.

2.2 Influence points checking

In this part, we use Cook's distance to measure the influence of each point and eliminate those effects from a few points. Thus, we have detected the **39th** and **86th** entries have large effect on the coefficients. After deleting those points, the Cook's distance diagnostic is quite good so we stop at this time.



2.3 Data cleaning summary

Deleting: 39, 86, 182, 216

Imputation: 42, 48, 76, 96, 163, 221

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

Table 2. Outliers and their treatments

3. Variable Selection and Model fitting

In this part, several methods have been applied, including stepwise ***AIC*** and ***BIC*** selection, ***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.

To be specific, any stepwise variable selection methods has been combined with the linear model such as ***AIC*** and ***BIC***. When we calculate the ***MSE*** of each model we select, cross validation has been applied to avoid outfitting. Thus, we should focus on whether the ***MSE*** value of these models are the same or deviating from each other since there will be a fluctuation in each time we calculate.

From Table 3, there is no significant difference in Mean squared error among those different methods except the situation

model	Mean_Sq	variables
Adjusted R square	15.61004	WEIGHT + ABDOMEN + BICEPS + WRIST
Mallow's Cp	15.61351	WEIGHT + ABDOMEN + BICEPS + WRIST
BIC	15.66767	WEIGHT + ABDOMEN + WRIST
Mallow's Cp	15.67571	WEIGHT + ABDOMEN + WRIST
Adjusted R square	15.69975	WEIGHT + ABDOMEN + WRIST
BeSS	15.91708	WEIGHT + ADIPOSITY + CHEST + ABDOMEN + WRIST
Mallow's Cp	16.19411	WEIGHT + ABDOMEN
Adjusted R square	16.22663	WEIGHT + ABDOMEN
BIC	16.22798	WEIGHT + ABDOMEN
BeSS	16.43950	ADIPOSITY + CHEST + ABDOMEN + HIP
BeSS	16.70302	HEIGHT + CHEST + ABDOMEN
LASSO	17.17719	HEIGHT + AGE + ABDOMEN + WRIST
BIC	18.61008	ABDOMEN
Adjusted R square	18.64164	ABDOMEN
Mallow's Cp	18.65939	ABDOMEN
BIC	35.17510	WEIGHT

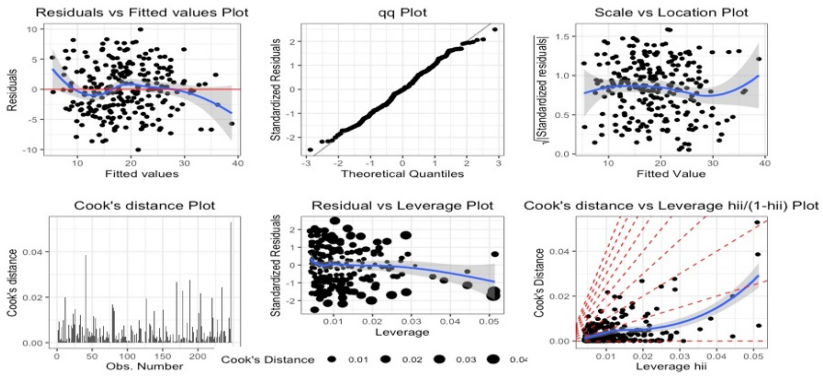
Table 3. Model comparison based on MSE

of one variable. Considering the rule of thumb, we decide to choose a model within two variables, which is not a heavy sacrifice in accuracy. The variables we select are weight and abdomen, applied in the sample linear model. Our final model is shown below:

BODYFAT(percentage) = - 41.9007 - 0.1230WEIGHTS(lbs) + 0.8956ABDOMEN(cm)

4. Model Diagnostics

After fitting the model, we need to test the normality of the residual and if there exist any pattern.



Apparently, there is no significant pattern among those plots, and the normality of the residual can be guaranteed. As for the Homoscedasticity and Influential points, there are no significant points showing in the plot. Thus we can accept the fitted model.

5. Conclusion

Advantage:

1. Rule of Thumb: The model we use only includes two variables, which is easy for test person to acquire its bodyfat percentage quickly.

2. Simplicity: The model only uses around 250 data to build a linear model, which is easy to handle for a modern computer.

Disadvantage:

1. Not a well-founded model This model is founded by about 250 people, which can hardly represent all of human since there might be some differences among different district.

2. Measurement error: Since the abdomen has been included in the model, it is easily to be measured incorrectly for normal people.

6. Shinny APP

Link: https://jiacheng-miao.shinyapps.io/Module2/

7. Contribution

Hao Qin: BeSS model building, slide editing, jupytor file editing and pdf summary editing

Qiaoyu Wang: Analyzing data with Mallow's Cp and Adjusted R square, integrating the ipynb and check the assumption by model diagnosis.

Jiacheng Miao: Set up Lasso model and Shinny App

Yuhan Meng: Data Cleaning, slide making, using stepwise variables selection model buidling method and drawing diagnostic plots

8. Reference

[1]. Centers for Disease Control and Prevention. (2011). Body mass index: considerations for practitioners. Cdc [Internet], 1-4.

[2]. Fuller, N. J., & Elia, M. (1990). Calculation of body fat in the obese by Siri's formula. European journal of clinical nutrition, 44(2), 165.

[3]. Wen, C., Zhang, A., Quan, S., & Wang, X. (2017). BeSS: An R Package for Best Subset Selection in Linear, Logistic and CoxPH Models. arXiv preprint arXiv:1709.06254.