Module 2 – Group 9

Runshi Tang | Zixiang Xu | Zeyu Li

Introduction (*Zeyu Li is responsible for following part in both slides and summary*):

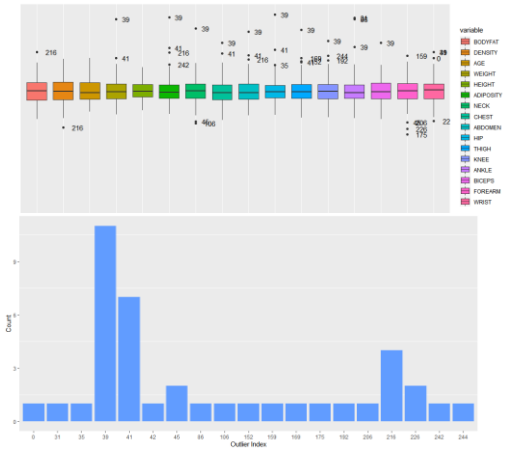
Having excessive amount of bodyfat might lead to tons of health problems. To keep our body in health, it is necessary to keep track of the percentage of our bodyfat regularly. However, sometimes it is inconvenient for some people who might be too busy to go to hospital or clinic to measure their bodyfat. This motivates us to develop this model to help these people, so that they can easily estimate their bodyfat at home.

Background about Data:

In our real dataset named "BodyFat.csv", there are ID number of individuals, percent body fat from equation, density determined from underwater weighing, age(years), weight(lbs), height(inches) and different circumference measurements of 252 men.

Analysis and Interpretation:

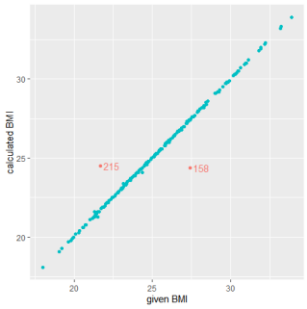
- *Identify and Eliminate Outliers*



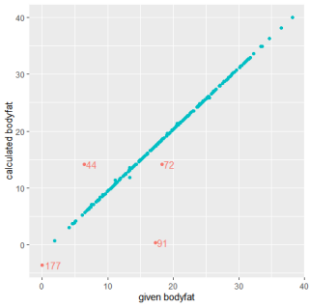
We used interquartile ranges to determine the outliers in the dataset from a multidimensional perspective. In other words, we considered participants who at least have one or more measurements beyond the normal range as outliers. According to analysis, we decided to remove the 31st, 39th, 41st, 42nd, 86th, and 216th data point. Because these participants are beyond the normal range in multiple factors. Taking the 39th participants as an example, he/she has 11 measurements of variable, which are much higher than that of other participants.

- *Check Consistency* ***(Runshi Tang is responsible for following part in both slides and summary)***

As we all know, body fat percentage and BMI can be calculated. To ensure our prediction more accurate, we further cleaned our dataset by eliminating the data points with inconsistency between given and calculated value in body fat percentage and BMI. After analyzed and visualized the data, we determine to get rid of the 44th, 72nd, 91st, 158th, 177th, and 215th data points which largely deviate from the main trend.



BMI



Bodyfat

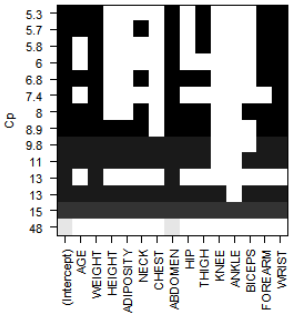
- *Construct Models*

We found our potential models by regression subset selection, which tests all possible combination of the predictor variables and select the best model based on several statistical criteria. However, to perform this test, we need to check whether interaction effects exist within variables first. Implemented the method of forward search, our algorithm checked models with the combinations of all predictors from one to all two-way interactions to see whether the addition of new variable will bring more accuracy to the model. Based on the result that there is no interaction term in final mode of forward search. After

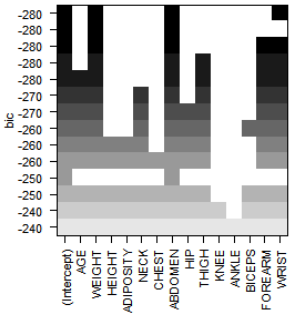
excluding interaction, we use global search to consider all possible combinations of variables based on 3 criteria which are the number of variables, Mallow Cp value, and BIC value. Mallow Cp value addresses the issue of overfitting, while BIC value tells us if the model is true. We intended to choose the model with highest accuracy and least number of predictors. So, we want both values to be small. According to analysis, we selected two candidates with relatively smaller BIC and Cp value compared to other combinations. And then, we conducted F test to see whether the predictors are significant in those two candidates. In F test for the model with 5 variables, the p-value is 0.1389 for variable "Thigh", which is bigger than significant level 0.1. So, we removed it and obtained Model 1. F test for the other candidate shows all variables are significant, which is Model 2.

Model 1: $Bodyfat \sim Weight + Abdomen + Forearm + Wrist$

Model 2: $Bodyfat \sim Weight + Abdomen + Wrist$



Model Selection (CP)



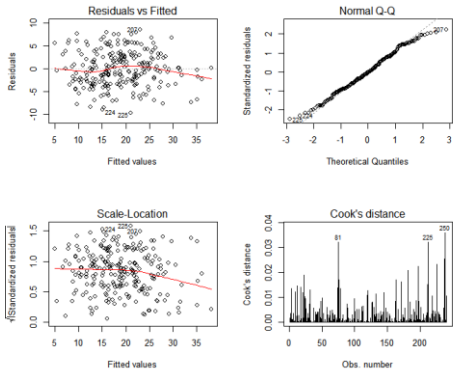
Model Selection(BCI)

Number of Variables	CP	BIC
1	48.404507	-254.5129
2	12.609903	-283.2538
3	7.430516	-284.8639
4	5.971949	-282.8753
5	5.769164	-279.6456
6	5.314538	-276.6983
7	6.755193	-272.8781

CP and BIC value

- *Model Comparison and Diagnostics* ***(Zixiang Xu is responsible for the following part in both slides and summary)***

We tried to figure out if Model 1 with 4 predictors is as complex as necessary to describe our dataset compared to Model 2. We implemented the "Anova test" for comparison and found P-value is 0.06472 which is less than significant level 0.1. The statistics shows that the complexity of Model 1 did significantly improve fit over Model 2. So, we chose Model 1 as our final model.



After we confirmed our final model, we checked model validation plots. As you can see in the right Plot, the linearity is reasonable because the points randomly scattered around the X axis. And on the Normal Q-Q plot, the points mostly align along with line, so we can know the normality is also reasonable. Besides, based on the residual plot the homoscedasticity is satisfied as well. From the bottom right plot we can see that there is no influential point, for all cook's distances are very small.

Rule of Thumb:

- A man with 154lbs weight, 90cm abdomen circumference, 20cm forearm circumference, and 20cm wrist circumference expected to 14.388% of bodyfat. The easy-to-use formular is:

$$BF\ Pct = -31.80 - 0.11weight + 0.90abdomen + 0.33Forearm - 1.24writst$$

Strengths and Weakness:

- *Strength:* Our model is reasonable and give us relatively accurate estimation of people's bodyfat percentage. And it is a simple process model requiring only 4 variables for prediction.
- *Weakness:* a. We did not further check the interaction effects among variables by the method of "Backward" and "Both-Step" due to the computational limitations. b. We gave up on the model which required user providing lesser information to predict their bodyfat. c. Our model cannot predict bodyfat if user provided extreme value.