STAT 628 Module 2 Summary

LEC 001 Group 19
Jingshan Huang,Xiaotian Wang,Yuanyou Yao

October 10, 2020

1 Introduction

According to a variety of popular health books, people are trying to use their percentage of body fat to assess their health. Hence, a simple, robust, and accurate way for estimating body fat is necessary.

Motivated by this, our team collect more than 20,000 men's body data that include age, weight, height and other 11 commonly available measurements, and estimated body fat based on Siri's (1956) equation.

When we draw some plots and look at the correlation between these measurements and body fat, obvious linear relationships are found. It encourages us to try linear model. Then, considering the correlation between those variables, an AIC based step-wise linear model is the core of this project.

2 Data Cleaning

We implement four rounds of data cleaning based on four different cleaning methods. Firstly, we use boxplots to detect outliers for each variables. Noticing the relatioinship among height, weight and adiposity (*adiposity* is proportional to $weight/height^2$), we regard any points not following this relationship as outliers. Thirdly, we detect the outliers based on our experience. To be not an outlier, the weight must within 300 lbs, the height must be higher than 30 inches and the bodyfat must be 2% 45%. Finally, we first fit the baseline linear model(variable density not used)

$$lm(bodyfat \sim .)$$

and then find the influential points based on Cook's distance and DFFITS to detect the outliers.

Totally, we detect 36 outliers and delete them from our dataset.

3 Modeling

3.1 Model Selection

Considering the interactions between any two variables, we use backwards selection choosing AIC as the criteria to find the best model. Starting with the full model

$$lm(bodyfat \sim . * .),$$

we finally find the best model

$$lm(bodyfat \sim abdomen + weight + wrist + forearm + neck + age + thigh + knee + hip + chest + weight : thigh + abdomen : neck + forearm : chest),$$

with the locally minimum AIC.

3.2 Final Model and Interpretation

After fitting the best model found by backward selection with criteria AIC, we have the best model

$$Bodyfat = 7.60–1.36abdomen + 0.37weight - 1.66wrist + 4.90forearm - 6.03neck + 0.07age + 1.62thigh–0.37knee - 0.20hip + 1.07chest–0.01weight \times thigh + 0.06abdomen \times neck - 0.04forearm \times chest$$

We interpret the model as the following examples:

- As a person get one year older, he is expected to gain 0.07% in body fat.
- As a person's weight increases by one pound, he is expected to gain (0.37 – 0.01thigh)% in body fat.

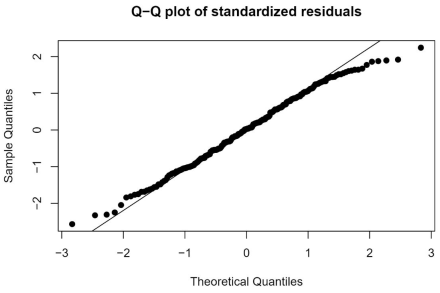
3.3 Statistical Inference

After implementing F test, we know 7 out of 13 variables are significant at 0.05 level, and other 3 variables are significant at 0.1 level. For the other 3 variables, their p values are lower than 0.2 and it is acceptable.

4 Model Diagnostic

4.1 Normality

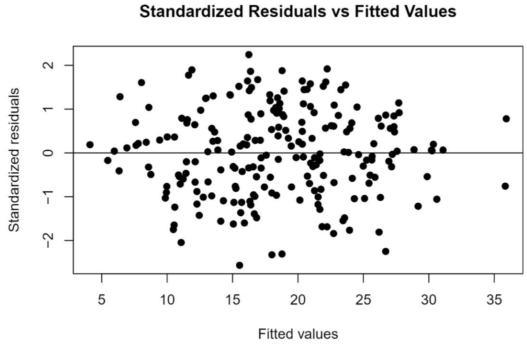
We draw the Q-Q plot of the standardized residuals to check the normality of the model.



From the Q-Q plot, we can see that the normality for the model is acceptable.

4.2 Heteroscedasticity

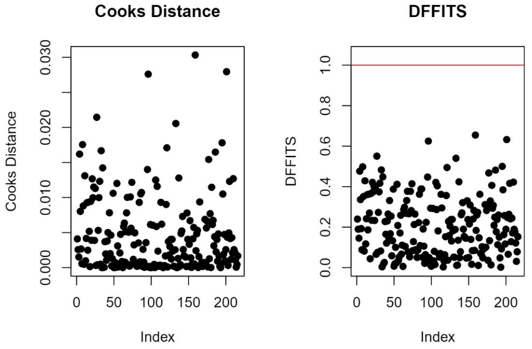
We can see the standardized residuals v.s. fitted values plot to check the equal variance assumption.



According to the plot we conclude that there is no severe heteroscedasticity problem.

4.3 Influential Points

We use the Cook's distance and DFFITS to detect whether there is any influential point.



From the two plots and the rule of thumb, there is no obvious influential points.

5 Discussion

Interactions are considered in our model, so collinearity problems can be avoided to some extent. Because this is a linear model, it is easy to make some statistical inference.

However, so many predictor variables are included in this model, and thus it can be hard to predict bodyfat for individuals with missing data.