Summary

1. Background

Nowadays, people are paying more and more attention to their own body health. The percentage of body fat becomes an increasing popular way for people to know whether they are in a good health. However, it is not convenient to measure the body fat very accurately and it can also be very costly. So here comes a question: how can we estimate the body fat by some factors that can be easily measured and how can we do it as accurate as possible?

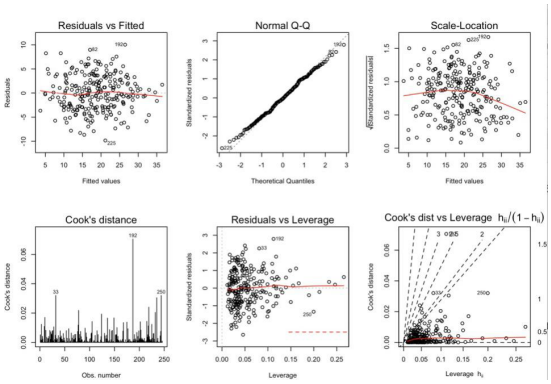
There are some factors that may be associated with body fat and can be easily measured in our daily life. They are age(years), weight(lbs), height(inches), adiposity(bmi), and the circumference(cm) of neck, chest, abdomen, hip, thigh, knee, ankle, biceps(extended), forearm and wrist. We want to build a model of body fat and some of the factors above. Then if we can get the data of these factors, we can estimate the body fat by this model. The easiest and the most widely used model is regression model. Thus, we choose to build a multiple regression model at first. Regard body fat as response variable and other factors as explanatory variables. Our aim is to select a model that fits our existing data best.

2. Data Cleaning

The first step of building a regression model is to find out the outliers and remove them. We decide to find the outliers from 3 aspects. First, we scale the data and draw the boxplot to find outliers from quantiles. Those far away from other data are obviously outliers(175^{th}, 31^{st},42^{nd}). We use bar plot to check who show outliers in quite a few(>3) variables(39^{th},41^{st},216^{th}). Second, some data's Bodyfat and Density don't meet "Siri's equation"(even allow a little error). We use boxplot to detect them(96^{th},48^{th}). Third, some data's Weight, Height and Adiposity don't meet BMI formula. We use boxplot to find them(42^{nd}). (Daiyi Shi wrote the part 1,2.)

3. Model Selection

After removing all the outliers, we get the clean data. Now we come to the second stage, that is model selection. From what we learned before, we should choose the model with smallest AIC, BIC, C_p and largest $R^2_{a,p}$. First, we take two-way interactions into consideration, and use stepwise selection to select models by AIC and BIC respectively. Second, we use regsubset to select one-way models by $R^2_{a,p}$,$C_p$and BIC. regsubset() is a R function which can do model selection by exhaustive search, forward or backward stepwise, or sequential replacement, so this way can do the selection thoroughly. In addition, after we get the 4 models, we need to do model diagnosis to check whether these models meet our assumptions well. We use R to draw some plot as follows.



For example, here is the plot of model 1. From the plot, we can see that model 1 meets the assumptions well, so we can choose model 1 to be in our list. Similarly, we can draw the plot of other 3 models, and the result are the same, they all meet our assumptions well. (Haoran Teng wrote the part 3.)

4. Model Validation

Then we come to the final part, that is choose a model from the 4 models above. We need to do model validation to check a selected model against independent data. There are several possible approaches. Here we choose K-fold cross validation. We split data randomly into 5 roughly equal parts and do training-test. By R, we get the mean square error of each model.

	Model 1	Model 2	Model 3	Model 4
RMSE	3.9297	4.0076	4.0275	3.9959

Model 1 has the lowest error. However, noting that 1st(AIC)'s error is not much less than 2nd(BIC)'s, but much more complicated. We use anova to figure the significance of difference. If the difference is not significant, we can choose a much more concise model(2nd). Unfortunately, it is significant by anova table($p=0.00037<0.05$).

so we choose model 1 to be our best model, that is:

$$BF = 174.9061 - 1.3484Ab + 1.1144We - 5.5715Wr - 0.4743H - 0.9374Ag - 0.1550C - 5.1855N + 0.3360F - 0.0048Ab * We + 0.0809Wr * Ag + 0.0834Ab * N - 0.0052Ab * Ag - 0.0159We * N$$

where BF refers to body fat, Ab refers to abdomen, We refers to weight, Wr refers to wrist, H refers to height, Ag refers to age, C refers to chest, N refers to neck, F refers to forearm.

5. Strengh and Weakness

The strength of this model is that the model is very simple to understand and all the variables we use to estimate the body fat are easy to measure. It can save out time and money. However, we haven't considered if there are three-way, four-way or even more interactions. Also, we haven't considered the existence of square, cube or log of some variables. Thus, the model may not be the most accurate.

6. Conclusion

In conclusion, through the real data set of 252 men with measurements of their percentage of body fat and various body circumference measurements, we come up with a simple way to estimate body fat, that is multiple regression. The formula of the model seems complex, but the principle of it is quite easy and robust. (Zijin Wang wrote the part 4,5,6.)