Executive Summary

October 7, 2019

1 Introduction

Body fat percentage is total mass of fat divided by total body mass. It is an important evaluation of a person's health level, too high or too low might cause diseases. There already some body fat calculators online. However, they usually require too many variables while some of them cannot be obtained easily. In this project, we try to construct a simple, moderately accurate and robust model to predict body fat percentage which obeys the "Rule of Thumb".

The dataset we use contains 252 male observations and 15 useful variables, including "BODYFAT" as response and some others as predictors. We have to construct a linear regression model with this dataset. First we have to clean the raw dataset. Then we fit a model to predict "BODYFAT", in order to obey "Rule of Thumb" we decide to choose two predictors and leave others out. After that, we construct, interpret and diagnose our final model. Finally we conclude the strengths and weaknesses of our final model.

2 Modeling

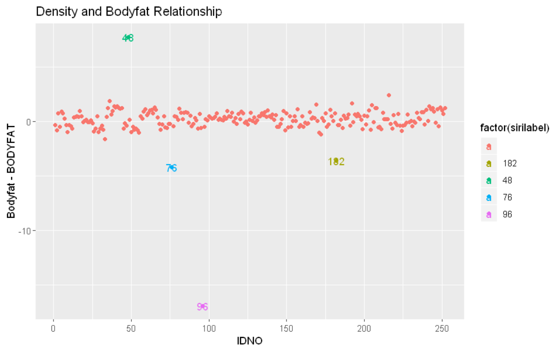
2.1 Data cleaning

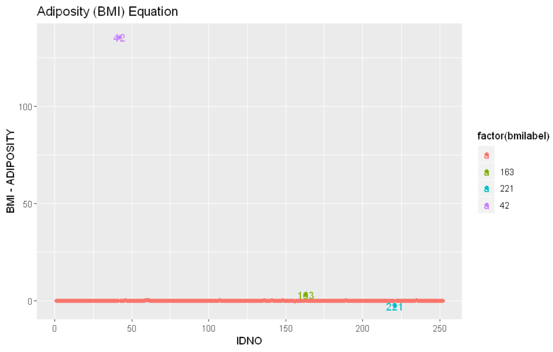
The first thing to do is to read the raw data and clean it (detect and delete abnormal observations). There are two criteria to extract abnormal observations. The first one is to delete those with abnormal feature values, such as a very short observation; The second one is to delete those with wrong relationship among some related values. For the second criterion there are two formulas, we have to check whether the points obey both of them:

$$BODYFAT \propto \frac{1}{DENSITY} \quad ; \quad ADIPOSITY(BMI) \propto \frac{WEIGHT}{HEIGHT^2}$$

By analysis, No.39 has too large weight ; No.42 is too short; No.48, 76, 96, 182 does not match the relationship between bodyfat and density; No.182 also has bodyfat 0 (abnormal value); No.42, 163, 221 don't obey the BMI equation. So we decide to delete eight points: No.39, 42, 48, 76, 96, 163, 182, 221. We remove them and make a new dataset. OutlierTest on the new set does not show any abnormality. Cook's distances also shrink to acceptable values after data cleaning.

```
No Studentized residuals with Bonferroni p < 0.05
Largest |rstudent|:
     rstudent unadjusted p-value Bonferroni p
224 -2.555635            0.01125           NA
```





2.2 Model Selection

2.2.1 Aggregation of different models

We did variable selection using these methods: Stepwise selection(AIC, BIC by backward/forward/both sides), Mallow's Cp, Adjusted R square and Lasso. And since our aim is to find a two-predictors model, we have to find out which of the predictors are frequently used. In all these nine methods, "ABDOMEN" and "WRIST" exists in all nine models, "WEIGHT" exists in eight models, other predictors exist in no more than four models. Therefore the most important three predictors are selected. We will continue our model by dropping at least one of these three predictors.

2.2.2 Additional analysis

Summarizing the result of these models, we find that "ABDOMEN" is the most important predictor , followed by "WRIST" and "WEIGHT". Because of our "Rule of thumb" criterion, we decide to choose one of "WRIST" and "WEIGHT", then take it along with "ABDOMEN" as our two predictors. Suppose we set "Model 1" using ABDOMEN and WEIGHT to predict BODYFAT, and "Model 2" using ABDOMEN and WRIST to predict BODYFAT. "Model 3" only use ABDOMEN as predictor. We check multicolinearity by correlation and VIF test first.

```
Correlation Coefficient (Model 1):  0.873 ; Correlation Coefficient (Model 2):  0.603
VIF of Model 1:  4.194 ; VIF of Model 2:  1.572
```

By VIF test, we suppose the predictors in "Model 1" are more correlated than "Model 2", so "WRIST" might be a better choice. Also, we can see the value of R-square are similar between two models. If we only use ABDOMEN to predict, the R-square indicates that's not reliable. So in order to reduce multicolinearity, we choose "Model 2" (ABDOMEN and WRIST) as our final model.

```
The R-square value of Model 1,2,3 are , 0.715, 0.724, 0.678
```

3 Final Model Interpretation

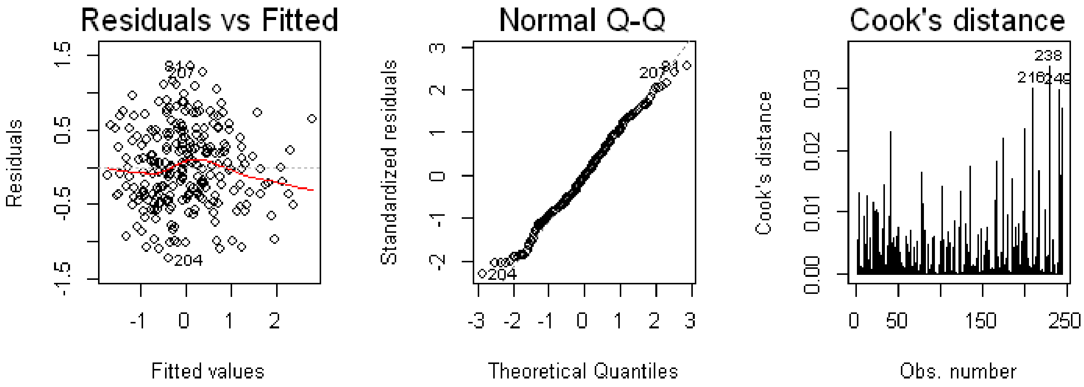
To interpret our final model, Male Bodyfat percentage can be estimated by 0.73 abdomen(cm) minus 2.03 wrist(cm) and then substract 11.2. That is:

$$Bodyfat(\%) = 0.73 * Abdomen(cm) - 2.03 * Wrist(cm) - 11.2$$

For example, a male who has abdomen 80cm and wrist 18cm, his bodyfat percentage is around 10.66%.

3.1 Diagnostics

Now we check modeling assumptions with plots. We can see the points approximately follows a normal distribution. They do not violate linear regression assumptions. So our final model is acceptable.



3.2 Strengths and weakness

Strengths of our Model:
1. Our final model is simple and straightforward. The variables included for prediction is easy to get. It is also easy to calculate.
2. The model is robust, it use the significant predictors that exist in most of models.
3. The model is accurate as the adjusted R-squared is greater than 0.7.

Weaknesses of our Model:
1. We only consider two variables and ignored the interaction between these variables as a trade off between simplicity and accuracy. 2. Our data sample only contains 252 observations, if the weight is too hig or too low, the prediction may not be so accurate.

3.3 Conclusion

Our final model only contains two variables, abdomen and wrist. They are easy to measure so that the body calculator is easy and as accurate as possible. Our Shiny App also consider the limitation of our model, we only predict the bodyfat percentage of male and the abdomen and wrist number cannot be too large or too small in order to make the bodyfat percentage reasonable.

4 Contribution

GUANQI LU: Data cleaning and model building and jupyter notebook summary.
QIAOCHU YU: Slides and plots using ggplot and jupyter notebook summary.
HAOXIANG WEI: Data analyzing and jupyter notebook summary.
YI-HSUAN TSAIShiny App building and jupyter notebook summary.