Summary of Body Fat Project - Group 2

1 Introduction

The body fat percentage is a measure of fitness level, which can be calculated as the total mass of fat divided by total body mass. In this project, a simple, robust, and accurate method to estimate the percentage of body fat using available measurements is proposed, related to man's weight, circumferences of abdomen, and thigh.

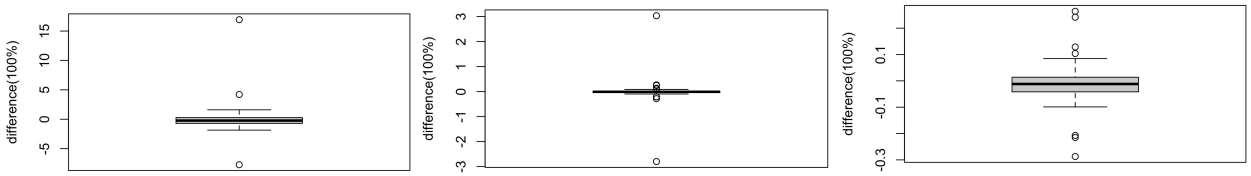
2 Data Preprocessing

In the data set, there are 252 observations and 17 variables. There are two relationships of some variables:

$$BODYFAT = 495/DENSITY - 450 \tag{1}$$

$$ADIPOSITY = 0.454 \times WEIGHT/(0.0254 \times HEIGHT)^2 \tag{2}$$

The predicted variable BODYFAT is supposed to be in a valid range, so we delete three observations whose BODYFAT value is smaller than 3 or larger than 40, which are the 172nd, 182nd, and the 216th with values: 1.9, 0, 45.1 respectively. We noticed the smallest Height is only 29.5 inches from the 42nd observation which is thought to be too short. We impute the invalid height by (2) and the value is 69.5 which is more reliable.
Based on (1) we recalculate the Body Fat by Density. Inspired by Meyer and Cole (2019), these three observations,who are 48th, 76th, 96th, do not match according to the box plot of the differences between original data and recalculated ones (as the left figure shows below). Since it is not possible to tell which variable is more reliable, we deleted these three observations. Similarly, after checking Adiposity by Weight and Height using (2) and drawing the boxplot of differences again (the middle figure shows below), we find there are obvious outliers: the 163rd, 221st and decide to delete them. Then we make a new boxplot (as the right figure shows below) without the 163rd and 221st observations and find although some points seem abnormal, the bias of the most extreme point is only about 0.3. Therefore, we don't view them as outliers.



By inspecting the pairwise scatter plots, some candidates are far from the majority. Via the data set and star plots, we noticed 39th and 41st observations are two big guys with large measurements in almost every aspect and are statistically different from others, so we delete them.
In conclusion, we delete 39th, 41st, 48th, 76th, 96th, 163rd, 172nd, 182nd, 216th, 221st observations and impute the Height of the 42nd observation.

3 Model Fitting and Selection

4 models are fitted in total. Model 1 is fitted based on backward AIC selection Akaike (1974) and then removes the variables that have a VIF value greater than 10. The result is:

$$\begin{aligned} BODYFAT = &-10.78494 + 0.08673 \times AGE - 0.44667 \times NECK + 0.69946 \times ABDOMEN \\ &+ 0.08762 \times THIGH + 0.27643 \times FOREARM - 1.91675 \times WRIST \end{aligned} \tag{3}$$

Model 2 is fitted by selecting three variables with the highest correlation coefficient with Body Fat, and it is shown as:

$$BODYFAT = -27.04418 + 0.77101 \times ABDOMEN + 0.21530 \times ADIPOSITY - 0.30407 \times CHEST \tag{4}$$

Model 3 is fitted by selecting correlation coefficient higher than 0.5 and then do the backward AIC selection, which is:

$$BODYFAT = -49.10679 + 0.90497 \times ABDOMEN - 0.15878 \times WEIGHT + 0.21646 \times THIGH \tag{5}$$

Model 4 is fitted by background searching. Body fat has a linear relationship with the circumferences of the waist, neck, and body weight. We replace the waist with the abdomen, and get the following result:

$$BODYFAT = 6.33705 - 0.47185 \times HEIGHT + 0.73534 \times ABDOMEN - 0.58141 \times NECK \tag{6}$$

To compare the performance of the four models above, we collect the adjusted R-squares (R^2) and Root Mean Squared Error (RMSE) of each model Usuelli (2014)and come out with the following table:

Table 1: Performance of Models

Model	Number of variables used	Adjusted R^2	Rank of R^2	RMSE	Rank of RMSE
1	6	0.7125	1	3.894257	1
2	3	0.6736	4	4.175837	4
3	3	0.7119	2	3.92342	2
4	3	0.7106	3	3.932186	3

According to the above table, we can see that Model 1 has the highest R-square and the lowest RMSE but it requires 6 variables. Compared with Model 1, Model 3 has similar R-square and RMSE and requires fewer variables so we decided to use Model 3 to be our final model. So, the final model is,

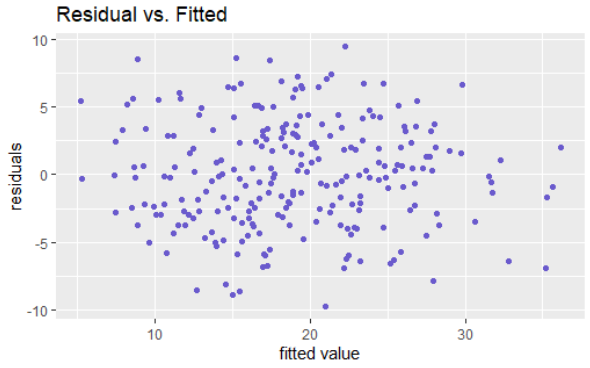
$$BODYFAT = -49.10679 + 0.90497 \times ABDOMEN - 0.15878 \times WEIGHT + 0.21646 \times THIGH \quad (7)$$

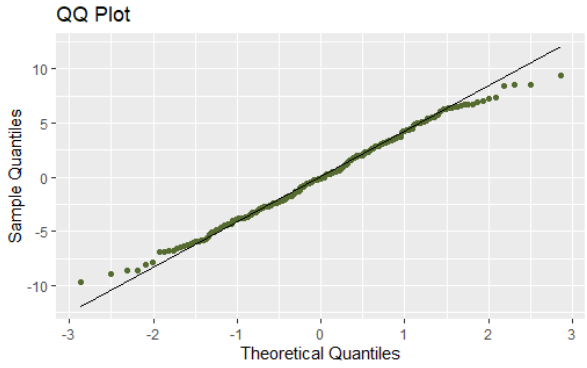
The final model shows men's body fat is positively correlated with the circumference of the abdomen and thigh while negatively correlated with body weight. The estimated coefficient of ABDOMEN is 0.905 which means holding other variables stable, a 1 cm increase of abdomen circumference will result in a 0.905 percent increase of body fat on average. Similarly, we can interpret other coefficients. For example, the body fat percentage of a male who is 180 lbs weight, whose circumference of the thigh is 60 cm with 100 cm abdomen circumference will be located between 24.90054 and 26.6938 with 95 percent confidence.

4 Model Diagnostics

From the residuals versus fitted values plot, we do not find any pattern of the residual points. Random scatter indicates no serious departure from linearity. Therefore, the assumption of linearity is not violated. In addition, the residual points are distributed evenly and randomly. Therefore, the assumption of homoscedasticity is not violated. Furthermore, there is no outlier in the plot.

From the normal QQ plot, we find that the points at the tails are not close to the line, so the normality assumption may be violated. However, for estimating and predicting of values of the response variable, the results will not be affected by the normality assumption. Therefore, we will not make a remedy to our model.





5 Model Strengths and Weaknesses

Strengths: All predictors in the final model are significant under alpha = 0.05. In addition, the model is quite simple but gives a fairly R-square and RMSE. Moreover, the data of variables are easy to get.

Weakness: The predicted range of the model is limited. For example, the estimated body fat for a male with 150 cm of abdomen circumference, 60 cm thigh circumference, and 200 lbs body weight is 67 percent which is too high for a person. Thus, the model is accurate only when data is within a certain range. Another thing we noticed is that the correlation between BODYFAT and WEIGHT is 0.59. However, the coefficient for WEIGHT is negative which does not match with the correlation coefficient.

References

Akaike, H., "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, 1974, *19* (6), 716–723.

Meyer, Stephanie and Renee Cole, "Army Body Composition Program Study Results Concerning: Enrollees Are More Over Fat Than Expected," *Military Medicine*, 03 2019, *184* (Supplement$_1$), 400 – –408.

Usuelli, Michele, *R Machine Learning Essentials*, Packt Publishing, 2014.

Website https://www.healthline.com/health/how-to-measure-body-fat

Website https://www.bizcalcs.com/body-fat-navy/

Contributions

Bowen Tian: Contribute to data preprocessing, model diagnostics, report writing.

Ouyang Xu: Contribute to R Shiny app, Github construction, report writing.

Tianhang Li: Contribute to model selection, model evaluation, report writing, slides writing.