Summary of Module 2

Group 1: Qizheng Xia, Jinghao Liu, Marit McQuaig

19th October 2021

1 Introduction

In this report, we will create and analyze a model to calculate body fat percentage given measurements including age, weight, height, etc.[1]

2 Data Cleaning

The dataset contained two incorrect body fat values (they were far under 3%, roughly the amount needed for survival). There were five instances where plugging the density values into Siris's equation resulted in a different body fat value than the dataset had as well as three cases where BMI calculated from weight and height values didn't agree with the adiposity values. These observations totaled only 3.6% of the total data and there was no discernible pattern between the affected results, so we removed them from the dataset rather than attempting imputation. After the cleaning process, our dataset had 243 values and the body fat measurements were roughly normally distributed with a mean of 19.03 and a standard deviation of 7.43.

3 Basis for Model Selection

We identified three criteria to guide our model selection. First, we looked for the model with the highest R-squared/adjusted R-squared, since this indicates it is the model that explains the most of the dependent variable (body fat percent). We also aimed to find a model with five or fewer variables, for simplicity's sake. Our model must also include variables highly correlated with the response variable.

4 Final Model

After comparing six models, we selected a simple multiple linear regression model to predict body fat with three explanatory variables: age (grouped), abdomen circumference, and wrist circumference.

$$Y = -155.7 + 3.8 * AGE.GROUP2 + 1.4 * AGE.GROUP3 + 62.9 * log(ABDOMEN) - 38.3 * log(WRIST)$$

where we divided the age variable into three groups: group one contains people between 22 and 30 years old, group two is people between 30 and 60, and group three for participants older than 60.

A layperson can easily use the above model to evaluate their body fat as they only need to know their age and measure their abdomen and wrist circumference (cm). For example, our group member Qizheng: he is 22 years old and his abdomen and wrist circumference are 80cm and 20cm. Our model indicates his body fat is 16.1%, and the prediction interval is $[7.827901, 24.34394]$.

From the model we can conclude that with the same abdomen and wrist circumference, the body fat of people under 30 years old is 3.8% and 1.4% less than people in age groups two and three respectively. For people in the same age group, increasing abdomen circumference by 1% leads to their body fat increasing 0.629%, and increasing abdomen circumference by 1% leads to their body fat decreasing 0.383%.

Of all the models we explored, this model had the highest R-squared and adjusted R-squared and all coefficients were significant at a 0.1 significance level. Plus it only has three variables, so it meets all of our criteria for model selection.

5 Statistical Analysis

All coefficients are significant, which means that the coefficients in our model not equal to 0. Taking a look at the overall model, the F-statistic is 151, and its p-value is $2.2 * 10^{-16}$ (nearly equal to 0) so we can conclude that the included variables have an effect on the predicted variable (body fat).

The R-squared of final model is 0.7126, which means that more than 70% of the data can be explained by the model. Generally, an R-squared value greater than 0.7 is considered a strong effect. This implies that age, abdomen, and wrist circumference are good predictors for body fat.

6 Model Diagnostics

Since we just the linear model to analysis the relationship between body fat and body feature of people, the model need to follow some assumptions, such as normality, equal variance, linearity and etc.

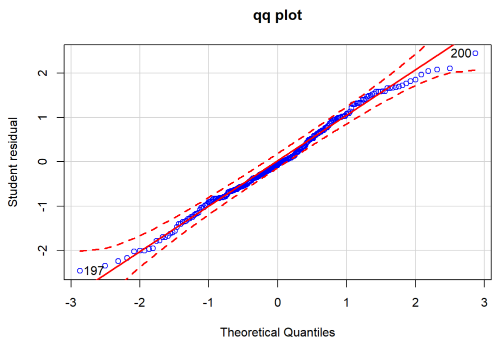

figure1. Q-Q plot

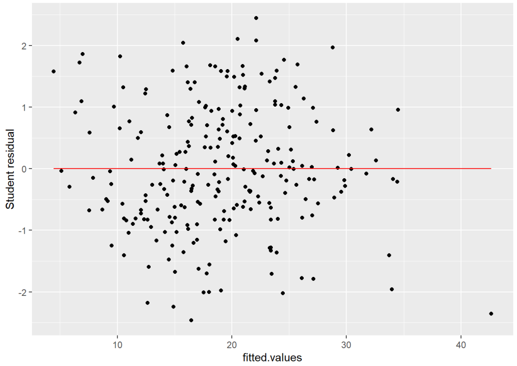

figure2.student residual vs. fitted-values

First of all, we use the Q-Q plot to check the normality of the model. Most of the points are on the line of $y = x$, which means that the distribution of Student residual of model is almost normal distribution. In the second figure points are relatively evenly distributed on either side of 0, and there is no obvious trend with these points.

To check the robustness of our model, we used two thirds of the data (chosen at random) as the training set, and the other third as the test set. The estimation, R square, and other statistics of the model based on training set and test set are all close to our initial model, so we can conclude that the model is robust.

7 Model Strengths and Weaknesses

7.1 Strengths

As discussed above, our model performs well based on R square and p-values. It only contains three variables making it simple to use. Since age is grouped, users do not need to tell how exactly how year old they are, just which group they fit in. Plus, the model meets all assumptions of linear models.

7.2 Weaknesses

This model uses a log transformation on two variables which makes interpretation more difficult and if there are values equal to 1, the model may not be useful. A second weakness is that the size of age group one is not very large compared with other groups, which may make estimation for this age group less accurate. The third is that the model can only calculate body fat of people over 22 years old.

8 Conclusion

In conclusion, our model calculates body fat percentage based on age, abdomen circumference, and wrist circumference. After the statistical analysis and model diagnostics, we can confidently say that the model is significant and useful for people to get easily and accurately predict their body fat percentage.

9 Reference

[1] 6 Ways to Measure Body Fat Percentage, Healthline, accessed 5 March 2018, <https://www.healthline.com/health/how-to-measure-body-fat>