Project Summary

Introduction

Body fat is an important measurement of the human body, which is closely related to body shape and health. Measuring body fat percentage could be inconvenient, so we are supposed to find proper ways to estimate body fat percentage. Based on a real data set of 252 men with their body fat percentage and 16 body measurements, we are trying to build a simple model to estimate male body fat as accurately as possible.

Data Cleaning

We first drew boxplots for each variable to know how the data are distributed. The plots show that there are a few extreme values. By tracking these values, we found the data with ID 42, 172, and 182 might have wrong data. Data 172 and 182 have extremely low body fat, 0% and 1.9%, which are impossible for human beings[1]. We failed to calculate them using density, so we deleted them. We also deleted data 42 which has a very low height, 29.5 inches, but his other measurements are in normal ranges, so we consider it as the wrong data.

Fitting Model

A matrix of scatterplots shows that Bodyfat and other variables are linearly correlated. In this case, multiple linear regression seems to be a reasonable model. To reach a balance between simplicity, interpretability and prediction accuracy, we conducted stepwise regression from both sides to select predictors. That is, beginning with the full model, adding or removing a variable in succession and testing for statistical significance after each iteration. To adjust the R-Squared value for the model size, we chose to consider the AIC criterion[2] which is defined by $AIC = -2logL + 2d$ where L is the maximized value of the likelihood function for the estimated model and d is the total number of predictors used. Generally, we select the model that has the lowest AIC value.

With multicollinearity, predictors tend to be linearly dependent, and thus the design matrix tends to be singular. We used VIF (variance inflation factor) to assess multicollinearity. The VIF of weight is 18.37, and the VIF of hip is 13.27, which were removed since their VIF are greater than 10, which indicates there exists severe multicollinearity[3].

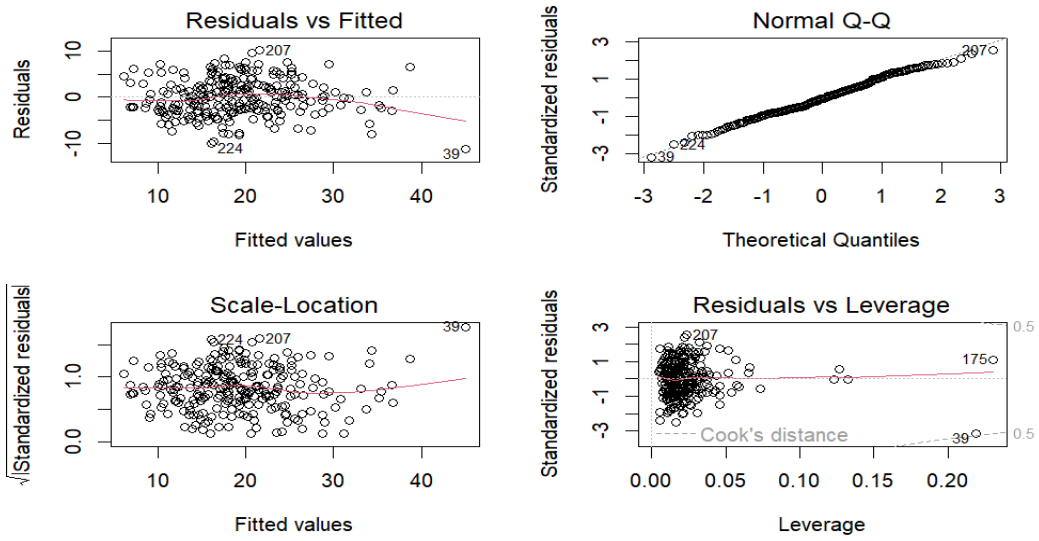
Finally, we dropped the variable THIGH for similarity because it is statistically insignificant with p-value=0.794. After dropping the outlier with ID 39, we got the final model:

$$BodyFat = -8.02 + 0.08 \times AGE - 0.39 \times NECK + 0.73 \times ABDOMEN + 0.27 \times FOREARM - 2.04 \times WRIST$$

Diagnostics

In this part we evaluated the four model assumptions with diagnostic plots. The residuals plot shows that the assumptions of linearity and homoscedasticity hold. The assumption of normality was evaluated using the QQ-plot. The fourth plot is Cook's distance which is a measure of the influence of each observation on the model. We found the observation with ID 39 is not only an

outlier but also an influential point. So we removed it and got the final model shown above.



Statistical Analysis

We conducted the hypothesis test to see whether the predictors we have chosen are significant in predicting the outcome. Suppose our null hypothesis is that the slope is equal to 0. We found the p-value is less than 2.2e-16, which indicated that the probability of getting test results at least as extreme as the results observed (under the null hypothesis) is 2.2e-16. So we rejected the null hypothesis based on 95% CI, which means our model is significant. Our R-square is 0.7291, which implies that 72.91% of the data fitted well in this model. Besides, all of the VIF(variance inflation factors) are less than 3.5, which indicated that there was no strong multicollinearity among the variables.

Model Strengths and Weaknesses

The strengths of our model are the absence of multicollinearity among the variables and the normality of residuals. And compared with the initial model, our final model reached a balance between prediction accuracy and simplicity.

For weaknesses, firstly our model requires the measurement of five values which may not be very user-friendly. Besides, our model has two negative coefficients which are hard for interpretation.

Conclusion

In this project, we developed a multiple regression model for body fat using five predictors in the BodyFat dataset and tried to find an efficient way to accurately predict body fat percentage. The essential part of building the model is the selection of predictors, we used stepwise to specify a model with the lowest AIC and reasonably high R-squared. The model satisfies typical assumptions for linear regression, works well with respect to all metrics, and can be used to predict body fat. Moreover, the model was very easy to interpret with simple linear relations. With further research and a well-formulated understanding of additional factors(such as BMI) in the future, we are confident that a statistical regression model like this can be used to provide even more accurate and substantial results.

Contributions

1. Every member of our group contributed to the establishment of the model.
2. RZ wrote the main part of the code which is used to build the model and analyzed it.
3. For the summary, ZZ wrote Introduction and Data cleaning. RZ edited Model fitting and Diagnostics. HX edited Statistical analysis, Model strengths and weaknesses and Conclusion. All of us contributed to the final review and revision.
4. ZZ wrote the code for the shiny App.
5. HX and ZZ created the slides.
6. RZ built the GitHub repository.

Reference

1. Friedl, KARL E., et al. "Lower limit of body fat in healthy active men." Journal of applied physiology 77.2 (1994): 933-940.
2. Bozdogan, Hamparsum. "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions." Psychometrika 52.3 (1987): 345-370.
3. Marquardt, D. W. (1970). "Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation". Technometrics. 12 (3): 591–612 [pp. 605–7].