Statistics 628 Module 2 Group 11 Body Fat Study Executive Summary

Brian Tsai, Tinghui Xu, Shengwen Yang

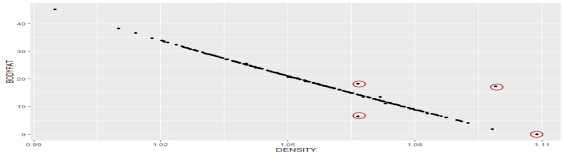
October 21, 2021

1 Introduction

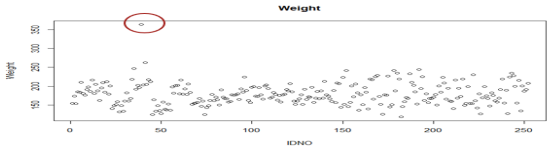
In this project, we built a model to estimate a person's body fat percentage by certain measurements based on previous data. This report will go over our data cleaning, model selection, model diagnostics, and final conclusions.

2 Data Cleaning

For data cleaning, we first tried to recover data by the linear relationship between body fat and density and between height, weight and adiposity. We plotted body fat against density (Figure a) and found that there were a few data points that were away from the line, including IDNO48, 76, 96, and 182. We then used Siri's equation to impute the body fat values based on their density for IDNO48 and 76. For IDNO96 and 182, however, we obtained negative body fat percentage according to the equation, so we chose to remove them. Next, we found that the height for IDNO42 was unreasonably low, hence we used adiposity and weight to impute the height. In the end, we searched for outliers by the summary statistics of all variables. We found IDNO39 was an outlier based on plots of weight (Figure b). IDNO79 was the only person who was over 80. IDNO172 had an abnormal body fat that was below 3%. Therefore, we removed all of them. For the consistency of the unit of variables, we also changed the units of weight and height to kilograms and centimeters.



(a) Body Fat V.S. Density



(b) Weight V.S. IDNO

3 Model Selection

3.1 Models Comparison

We tried different models to fit the data, including linear regression, random forest and XGBoost models. In the end, we chose the linear regression model because of the following reasons. First, compared with the other two models, linear regression is straightforward to understand and easy to explain. Second, different from the random forest and XGBoost model with R^2 of 0.62 and 0.63, the linear regression model has a better R^2 value of 0.735. Third, MSE for random forest and XGBoost model are 21.34 and 20.67, which are higher compared to 15.75 of linear regression model.

3.2 Variables Selection

Based on the coefficient of Lasso regression from Table 1, we chose the most significant variables, including age, abdomen, height, and wrist.

Table 1: Coefficients of Lasso Regression

	Intercept	Age	Height	Abdomen	Wrist	Others
Coefficient	−7.931	0.004	−0.119	0.563	0.209	∗

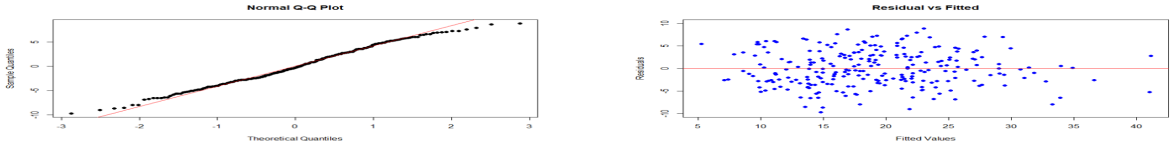
3.3 Final Model

To sum up, we decided to predict the body fat by age, abdomen, aeight, and wrist using a multiple linear regression model. As the result of the study[1] showed, body fat changed with age differently between ages 20 to 39 and 40 to 84.We decided to treat age as a categorical variable and cut off at 40. Since the dataset was small with only 247 rows, we chose to use the Leaving One Out Cross Validation method to train our final model. The equation for our final multiple linear regression model is:

$$BodyFat = 9.399 + 0.707Abdomen - 0.138Height - 1.742Wrist + 1.149(Age40 - 80) \quad (1)$$

Those coefficients mean that for a 1cm increase in abdomen,the model predicts that body fat will increase 0.707%, and for a 1cm increase in height and wrist, the model predicts that body fat will decrease 0.138% and 1.742%. For people in age group of 40 to 80, the model predicts that body fat will increase 1.149%. An example to use the model is that: Usain Bolt at the age of 35 with 83.82cm abdomen, 195cm height, and 18.4cm wrist is expected to have a body fat of 9.58% based on our model. His 95% prediction interval is between 8.06 %and 11.09%.

4 Model Diagnostics



(a) Normal QQ Plot (b) Residual V.S. Fitted

Figure 2: Assumptions Check

We checked the assumptions for our final multiple linear regression model. First, we checked normality by the shapiro test and qq-plot. The p-value of the shapiro normality test is 0.076 for our model, so we failed to reject the null hypothesis.In normal qq-plot (Figure2 a), most of the residuals are approximately in a line. Then we checked the homoscedasticity by the residual plot(Figure2 b). We also checked VIF values for each variable is lower than two, so there is no multicollinearity between variables.

Strength of our model is that our model is simple, easy to interpret. We found our R^2 to be 0.735, which implies 73.5% of variance in the dependent variable that can be explained by the independent variable and the data fits the model well. Some weakness of our model include that for male older than 80 and younger than 20, our model might not be accurate.

Reference

[1] A. J. Silver, C. P. Guillen, M. J. Kahl, and J. E. Morley. Effect of aging on body fat. *Journal of the American Geriatrics Society*, 41(3):211–213, 1993.

5 Contributions

Brian Tsai: Data cleaning, LASSO, and slides; Tinghui Xu: Linear regression model and Shiny app development; Shengwen Yang: Decision tree models and summary