Predicting Body Fat Applying Multivariable Linear Regression

Jingchen Chai, Ruihan Zhang, Wenjia Zhu

Abstract

Body fat is an important factor in many medical situations. Measuring body fat can help to assess whether a person is at a health weight. Additionally, obese or overweight can be identified by calculating the body fat. We conducted initial visualization of the body fat dataset with plots and tables, followed by model selection, model interaction, and cross-validation. Model diagnostics was also applied to check whether the model needs any transformation before omitting the outliers. We then select an optimal model that can best predict the body fat.

Introduction

A normal body fat is between 25 and 30 percent in women and 18 and 23 percent in men [1]. More than 68% of US adults are considered overweight, and 35% are obese [2]. While the number of people who are overweight or obese are gradually increasing, more people are becoming more concerned about this issue. The goal of this project is to predict the body fat in different patients based on their physical measurements.

Methods

Data Exploratory Analysis

We imported the original raw dataset-body density data. Among the three target variable: bodyfat_brozek, bodyfat_siri, and density, we have chosen bodyfat_siri as our final target variable. Other two were omitted from the dataset. The goal of this part is to have a sense of the distribution of each variables.Correlation was also considered to see whether any variables are highly correlated with eachother so that we can try to eliminate the multicollinearity effect. In order to visualize the above information, a descriptive statistics table was made, a boxplot for the target variable, histogram for the predictors, and a correlation plot. Multicollinearity will also be checked by calculating the VIF.

Model Selection

In order to select the most appropriate model, we applied stepwise selection for comparasion.

_Model interaction-

Model Diagnostics (need compare?)

After obtaining the model that we have chosen, several plots will be made to check whether the model meets the regression assumptions. The QQ plot will be used to determine the normality of the data. Outliers in the

data will also be identified by plotting the residuals vs leverage plot and also calculating the Cook's distance , which can show the influence of a specific case on all fitted values. Finally, a model without outliers will be shown and will be compared to the original model to examine the influence of the outliers. The influential outliers will be omitted in the final model. VIF again?

Cross-Validation (need compare?)

A 5-fold cross-validation will be constructed to test the validity of the model.RMSE(root mean squared deviation), R square, and adjusted R square will be used to evaluate the model.

```
body_df=read_excel("data/body_density_data.xlsx") %>%
  janitor::clean_names() %>%
  select(-bodyfat_brozek,-body_density,-id)
```

```
# check multicollinearity
body_corr=
body_df %>%
cor()
corr=corrplot(cor(body_corr),
         method = "color",
         type = "upper",
         addCoef.col = "black",
         number.cex = 0.6,
         diag = FALSE)
```