Development of a Body Fat Percentage Estimation Model

Introduction

This report will break down the details of the development of a model for estimating body fat percentage, including process of the dataset, guidelines on the development, diagnosis, and performance of the model.

Motivation for Model

Linear Model

Body fat percentage is linearly related to variables, like weight and volume. Thus, linear models will be sufficient for the problem.

Progressive Models

For users knowing different levels of information, different models should be applied. Detailed information gives a precise estimation while less information yields a less precise estimation. The three different levels are:

1. **Essential information** (lv. 1)
   Age, Height and Weight.
2. **More detailed information** (lv. 2)
   Wrist and Hip Circumference.
3. **Advanced information** (lv. 3)
   Variables need to be measured.

Exploratory Data Analysis and Data Cleaning

First, we look at the key aspect about the data. The minimum of bodyfat is 0.00, which is obviously abnormal. We also consider the value of minimum density and minimum height abnormal, with respect to 0.995 and 118.5.

According to Siri's equation, the percentage of body fat equals $495/D - 450$, where D denotes the Body Density (gm/cm^3). So, the percentage of body fat has a linear relationship with $1/D$.

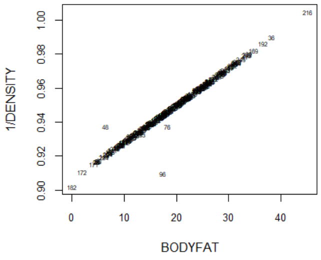

Figure 1 Linear Relationship and Abnormal Points

The image above clear indicates this linear relationship.

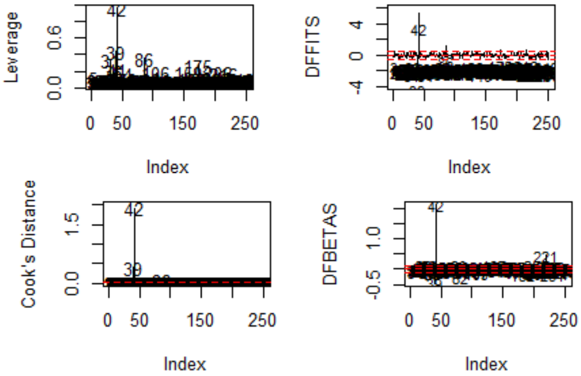

Figure 2 Different Test Methods

Besides, we decide to detect the abnormal cases by *leverage*, *DFFITS*, *Cook's distance* and *DFBETAS*. High-leverage points are the observations made at extreme or outlying values of the independent variables and lack of neighboring observations. DFFITS and Cook's distance shows how influential a point is in the regression influence of a data point, DFBETA measures the difference in each parameter estimate with and without the influential point.

Final Model

Level 1

For the users who only know his/her age, height, and weight, we have only used these 3 predictors to train our model. Model-1 is

$$Bodyfat = -26.9576 - 1.0858 \times age + 2.1092 \times weight + 0.0068 \times age \times height - 0.009 \times weight \times height$$

Level 2

For the users who know more information which includes chest, wrist, and hip circumference, we have used these 6 predictors to train our model. Model-2 is

$$Bodyfat = -265.6789 - 2.1233 \times age + 1.9143 \times height + 6.0166 \times hip - 11.8612 \times wrist - 0.0043 \times age \times weight + 0.0045 \times age \times height + 0.1011 \times age \times wrist + 0.0097 \times height \times weight - 0.0089 \times weight \times chest - 0.0331 \times height \times hip + 0.0404 \times wrist \times chest$$

Level 3

Then we have used all the predictors to train our model. For the users who can provide abdomen, neck and forearm circumference, our final model Model-3 is

$$Bodyfat = -48.0073 + 1.2725 \times hip - 0.003 \times height \times hip + 0.0216 \times abdomen \times neck - 0.024 \times hip \times neck + 0.0065 \times height \times forearm - 0.0492 \times forearm \times wrist$$

Model Selection

Besides the main effect of the predictors, we have also considered the interaction effect between two variables. To select the best combination of predictors, we have used two methods which are *exhaustive search* and *backward stepwise regression*. In level-1 and level-2 model, since the number of predictors is not large, we could use exhaustive search with R^2_{adj} as the criteria. Exhaustive search will try all the combinations of the variables and return the combination with the highest R^2_{adj}. In level-3 model, we cannot use exhaustive search because the number of predictors is too large. So, we have used backward stepwise regression with cross validation to do model selection. We set the maximum number of predictors as a hyperparameter and RMSE as the cost function. Fig 3 shows the process of our training process.

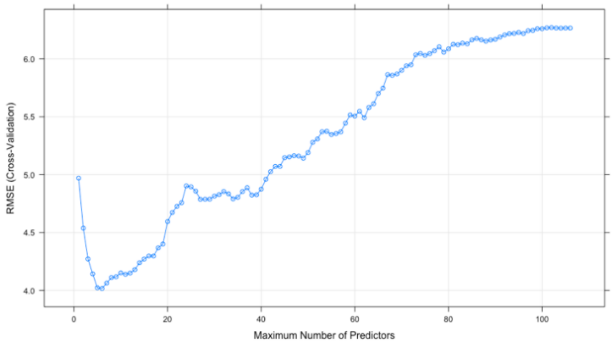

Figure 3 RMSE during a Training Process

Analysis

We used linear regression to be our model. To evaluate the performance of our models, we have randomly split the data set into training and testing set and 75% of the data is used as training data. The table below shows the RMSE and R^2_{adj} on different data sets and different models.

In Table 1 below, “RMSE train” and “R^2_{adj} train” are the RMSE and R^2_{adj} of the model trained on training data. “RMSE test” is the RMSE is the RMSE on testing data with the model trained on training data. “RMSE full” and “R^2_{adj} full” are the RMSE and R^2_{adj} of the model trained on all of the data.

We can see the model performance becomes better as the level increases.

Table 1 Performance of Each Model

	Level1	Level2	Level3
RMSE train	4.8764	4.2987	3.7550
RMSE test	4.8251	4.6265	4.7662
RMSE full	4.8294	4.3216	3.8587
R^2_{adj} train	0.5482	0.6367	0.7260
R^2_{adj} full	0.5678	0.6434	0.7220

Model Diagnostics

We have made QQ plot and scatter plot of the residuals of the 3 models and they show that the assumption of normality, homoscedasticity and linearity are not violated.

Model Strength and Weakness

For the exhaustive search, the weakness of this method is that it is not efficient. It must go through every single scenario to find the optimal solution, the selected factors, in our cases. It is very time-consuming.

In the last model, compared to the first two, the backward stepwise regression has been used. The strength of this method is that it is more efficient than exhaustive search. It has less time complexity. However, it also has criticism. First this method is prone to overfitting. And the second question is that the model may be over-simplifications of the real models of the data.

Conclusion

With cleaned data, we have trained three linear models to estimate a person’s body fat percentage, which is accurate, flexible in different situations, and easy-to-use.

Contributions

- Hao Jiang: Model Ideas, App Development, GitHub Management, Report, Presentation.
- Kou Wang: Data Cleaning, Model Ideas, App Design Ideas, Report, Presentation.
- Yicen Liu: Model Ideas, Model Building, App Design Ideas, Report, Presentation.