Body Fat Prediction

Shuo Li, Yunze Wang, Xinrui Zhong

1 Introduction

In this project, we aim to develop a simple and accurate model to estimate body fat percentage using accessible body measurements. The dataset includes 252 male individuals with recorded body fat percentages and various body metrics.

2 Data Preprocessing and Cleaning

2.1 Data Preprocessing

We removed the variable *IDNO*, which is unrelated to predicting body fat, as well as *DENSITY*, which can directly calculate body fat. For outliers, we identified and removed them using box plots, and applied the KNN method (K=3) to fill in the missing values. Finally, we normalized the data using the MinMaxScaler.

2.2 BMI Processing

Since BMI can be calculated from weight and height, we compared the theoretical BMI with the actual BMI values in the dataset, identified outliers using box plots, and analyzed weight and height using Z-scores to compare their deviation levels with BMI. Based on this analysis, we removed the most significant outliers and recalculated BMI using the cleaned weight and height data, thus completing the preprocessing of BMI.

2.3 BODYFAT Handling

For the response vector *BODYFAT*, we observed the presence of outliers. After removing these outliers, we needed to select an appropriate method to handle the missing values.

Here, we introduced a new formula to calculate body fat percentage (BFP):

$$\text{BFP} = \frac{1.0324 - 0.19077 \times \log_{10}(\text{waist} - \text{neck}) + 0.15456 \times \log_{10}(\text{height})}{495} - 450$$

We compared the theoretical BFP values with the actual *BODYFAT* values from the dataset and performed residual analysis by computing the absolute residual (Abs_Residual). We then replaced *BODYFAT* values with BFP for samples where the absolute residual was greater than 0.5, completing the data cleaning process.

3 Model Selection

3.1 Multiple Linear Regression (MLR)

Multiple linear regression is a statistical method used to analyze the linear relationship between two or more independent variables and a dependent variable. It is simple and easy to understand, making it suitable for predicting *BODYFAT*.

3.2 Feature Extraction

We used three feature selection methods: forward selection, backward elimination, and step-wise regression, with a significance level of $p = 0.05$. The selected features from each method are as follows:

- Forward: 'ABDOMEN', 'WRIST', 'HEIGHT'
- Backward: 'WEIGHT', 'HEIGHT', 'ADIPOSITY', 'ABDOMEN', 'WRIST'
- Step-wise: 'ABDOMEN', 'WRIST', 'HEIGHT'

Since the forward selection and step-wise methods resulted in the same variables, we combined them into MLR_Forward, while the backward method was denoted as MLR_Backward. We analyzed both MLR models, and the results are shown in Table 1.

3.3 Decision Tree

Decision trees are a type of supervised learning algorithm used for both classification and regression tasks. By recursively splitting the dataset, a tree structure is formed to predict the target variable. After ranking the importance of the features, the following were selected:

- 'ABDOMEN', 'WEIGHT', 'BICEPS', 'NECK', 'WRIST', 'AGE'

These features were then fed into the decision tree model for predicting *BODYFAT*, with the results shown in Table 1.

	MLR_Forward	MLR_Backward	Decision Tree
R^2	0.73	0.73	NA
Adjusted R^2	0.72	0.72	NA
Mean Squared Error (MSE)	13.25	13.19	8.47
Cross-validation MSE	13.84	14.00	28.45
F-test (p-value)	2.58e-70	3.56e-68	NA
Jarque-Bera (JB) test (p-value)	0.112	0.110	NA

Table 1: Comparison of different models

3.4 Model Analysis

From Table 1, we can observe that both MLR models have similar R^2, adjusted R^2, MSE, and cross-validation MSE values. Although the decision tree has a lower MSE, its cross-validation MSE is significantly higher after 5-fold cross-validation, about twice that of the MLR models, indicating overfitting. Therefore, we eliminate the decision tree model.

3.5 Model Diagnostics And Final Choice

For the remaining MLR models, we conducted F-tests and Jarque-Bera (JB) tests. As shown in Table 1, both models exhibit similar robustness. Given its simplicity, we ultimately selected MLR_Forward. Based on this model, we found that *ABDOMEN* has the most significant impact on *BODYFAT*, which is consistent with biological knowledge. Fat distribution in the human body is uneven, with the abdomen being a major fat storage area, especially for visceral fat, making the model logically sound.

4 Model Advantages and Disadvantages

4.1 Advantages

- The multiple linear regression model is simple in structure, making it easy to implement and use.

4.2 Disadvantages

- MLR relies heavily on the assumption that the residuals (errors) follow a normal distribution
- It's difficult to clearly explain the individual effect of each variable

5 Contribution

- **Shuo Li**: Data cleaning, model building, coding
- **Yunze Wang**: Model building, summary writing, GitHub maintenance
- **Xinrui Zhong**: Model building, Shiny App development