Group 2 Body Fat Prediction Summary

1. Introduction

Estimating body fat percentage is one of the crucial components in assessing an individual’s health and risk for various diseases. One of the traditional methods of measuring a person’s body fat percentage is to weigh them on dry land and then weigh them while in the water. These methods are inconvenient and costly in clinical settings. This project aims to develop a simple, robust, and accurate model to estimate body fat percentage using available clinical data such as age, weight, height, BMI, and body circumference measurements. By using the real data set of 252 men with measurements of their body fat percentage and various body circumference measurements, we will explore the trade-offs between simplicity, robustness, and accuracy, making the methods convenient for everyday use in clinical settings.

2. Data Cleaning

We aim to explore a method of data cleaning to assess the performance of most of the population in a healthy range. In this model, we selected the IQR (Interquartile Range Method) Method to clean the data set. It is a common method for detecting outliers in the dataset. IQR is required to calculate the first and third quartile of the data. By using these values, we can calculate the IQR which is the third quartile value subtracted by the first quartile value (IQR = Q3 – Q1). Then, we can determine the outlier by setting the lower bound, typically defined as the first quartile subtracted by 1.5 times IQR (Lower Bound = Q1 – 1.5 * IQR), and the upper bound, defined as the third quartile plus 1.5 times IQR (Upper Bound = Q3 + 1.5 * IQR). In this case, we will remove an outlier where IDNO is 216 due to the high body fat that is outside the upper bound. Also, we imputed the body fat percentage where IDNO is 182 due to 0% body fat, obviously impossible for humans, into 4.42% which is estimated by the popular body fat estimation model from the internet.

2. Feature Selection

Feature selection is essential to improve model performance, prevent overfitting, and reduce computational complexity. To identify the most influential features and streamline the model, we utilized Recursive Feature Elimination (RFE) to perform feature selection. It works by fitting the model multiple times, each time removing the least significant feature based on the model's coefficient or importance score, until the desired number of features is selected In detail, we performed RFE iteratively from 1 to the total number of features, then plotted the results (Figure 1). Based on the trade-off between model complexity and performance, 6 features were finally selected.

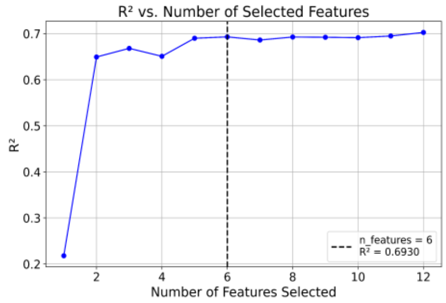

Figure 1 R² Scores Across Different Numbers of Selected Features Using RFE

4. Model Training and Evaluation

For this study, we employed multiple linear regression (MLR) as the predictive model to estimate body fat percentage based on the selected features. MLR was chosen for its simplicity, interpretability, and effectiveness in modeling linear relationships between dependent and independent variables. In this task, the dataset was split into training and validation sets in a 4:1

ratio. After training, the coefficient of determination (R^2) and root mean square error (RMSE) metrics were used for evaluating the model performance on validation set. R^2 reflects how well the model explains the variability in the target variable, while RMSE provides insight into the average prediction error. The agreement between the observed and the predicted bodyfat percentage are shown in Figure 2. Our model achieved an R^2 of 0.6930 and RMSE of 4.4288.

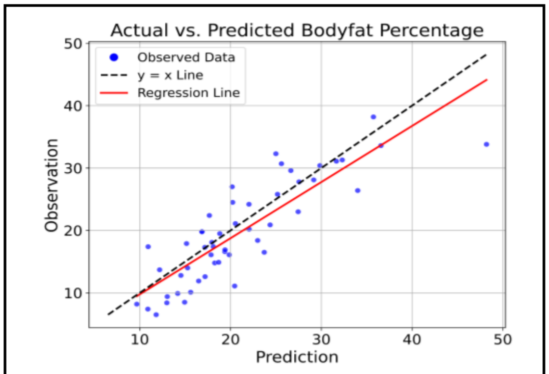

Figure 2 Scatter plot of predicted vs. actual bodyfata percentage

5. Model diagnostics

We checked the following four assumptions for MLR. First, we checked linearity using a scatter plot of residuals versus fitted values (Figure 3). Because the points were mostly randomly scattered around the horizontal axis without distinct patterns, we believed that the linearity assumption is plausible. Second, we checked homoscedasticity. The same plot (Figure 1) suggested consistent variance across most fitted values, though slightly larger residual dispersion was observed for higher fitted values (above 25), indicating a slight violation of the assumption. Third, we checked the normality of residuals using a QQ plot (Figure 2). Since the majority of the points closely align with the diagonal line, we believe that the normality assumption is largely satisfied. However, minor deviations observed at the tails of the distribution indicates the presence of mild heavy tails or potential outliers. These deviations are minor and are not expected to have a significant impact on the overall model performance. Fourth, we checked independence of residuals using the Durbin-Watson test (Equation 1). Because the test statistic was 1.6762 (close to 2), we believed that the independence assumption holds.

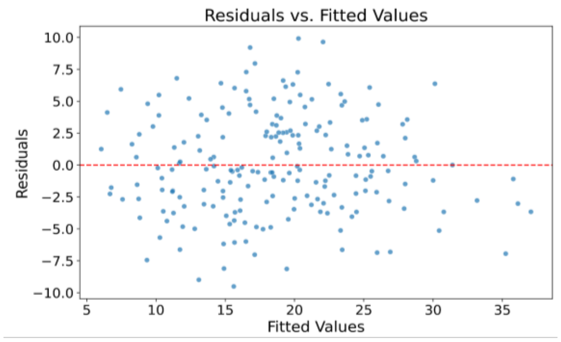

Figure 2 Scatter plot of predicted vs. actual bodyfata percentage

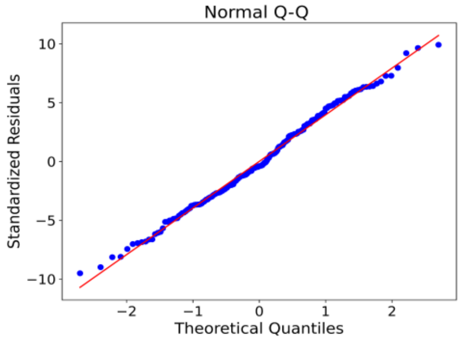

Figure 3 Q-Q Plot of Standardized Residuals

6 Conclusion

In this study, we explored MLR model for predicting body fat percentage. At first, IQR method were adopted to clean the given data set. Through Recursive Feature Elimination (RFE), we selected the most relevant features, ensuring the model maintains a balance between performance and simplicity while reducing overfitting risks. However, slight heteroscedasticity and minor deviations were observed while checking the assumptions of MLR. Despite these limitations, the model demonstrates solid predictive power with an R^2 of 0.6930 and RMSE of 4.4288. Although the model performs well for linear relationships, further improvements may be needed to address potential non-linear patterns.

Reference:

1.Katch, Frank and McArdle, William (1977). Nutrition, Weight Control, and Exercise, Houghton Mifflin Co., Boston.

2.Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, *46*, 389-422.

3.FreeDieting. (n.d.). Body fat calculator. https://www.freedieting.com/body-fat-calculator

Contribution	**Zhenke Peng**	**Zekai Xu**	**Jiren Lu**
Presentation	Responsible for sildes 1-3(Introduction and data cleaning)	Responsible for sildes 4 - 8(Feature Selection, Model Training, Evaluation, Discussion and Conclusion)	Responsible for silde 9 and providing feedback on slides
Summary	Responsible for Introduction and data cleaning Reviewed Feature selection	Responsible for Feature Selection and Model Training and Evaluation Reviewed discussion and Conclusion	Responsible for Discussion and Conclusion Reviewed Introduction and data cleaning
Code	Responsible for Data cleaning and plotting graphs	Responsible for Feature Selection and Model Training and Evaluation	Responsible for discussion and plotting graphs
Shiny App	Review and provide feedback for Shiny App	Review and provide feedback for Shiny App	Responsible for Shiny App