Summary

Group 7 Ruijing Chen, Tianyu Yao, Kangxin Zheng

Introduction/Motivation:

We aim to develop a simple, reliable model for predicting body fat using accessible measurements, offering a practical alternative to existing methods.

Background information/Data cleaning:

1. Consistency check: We observed that body fat can be calculated from body density, and adiposity (BMI) can be calculated from height and weight. To ensure consistency, we compared the calculated values with the original data, focusing on discrepancies:

Calculation error : Density was accurate, but the body fat rate was miscalculated.We replaced the incorrect values with the recalculated ones based on density.

Measurement error : The density measurement itself was incorrect, but the body fat calculation was otherwise valid.We set the body density value to NA for these entries.

2. Find outliers : Using box plots and the IQR (Interquartile Range) method, we identified potential outliers. Some individuals frequently had extremely high values, suggesting they were overweight. Others, however, exhibited large values in only one feature (e.g., someone with a normal body but an unusually large neck measurement). We classified the latter as outliers.

Additionally, as the IQR method may not capture all outliers, we manually defined a reasonable range for body fat percentages. We set the acceptable range between 3% (the minimum essential body fat) and 40% (an abnormally high value, especially for men).We set these outliers to NA.

3. Imputation : For outliers and missing values, we used regression imputation. The outlier or missing column was treated as the dependent variable, and we predicted its value using other relevant features.

Model Selection and Motivation:

We used multiple linear regression as the model. To choose the variables, we applied all-subsets regression, which went through all possible combinations of variables, which is 2^p−1 possible models of p predictors, to select the best regression model.From this process, we selected the top-performing model within each subset size.Then we evaluated these models with three criterion:

1. Accuracy:calculated by adjusted r-square.

2. Simplicity,:which is the number of predictors in the model.

3. Robustness:applying bootstrap resampling techniques and evaluating the variance in model performance across different samples.

From Plot1, you can easily see these three points of trade off.

After evaluating several candidate models using a bubble plot,we ultimately chose the model using abdomen and weight as predictors, for it's the relatively robust and simple model,and maintains good accuracy among the candidates. It strikes the ideal balance across all three dimensions – simplicity, robustness, and accuracy. Its overall performance is more likely to yield reliable and efficient results compared to other models that either lack robustness or simplicity.

Final Model:

The model predicts body fat percentage using two variables: abdomen circumference and body weight. The linear regression equation is:

$$Body\ Fat\% \ = \ -40.12 \ + 0.92 \ * \ Abdomen(cm) \ - 0.14 \ * \ Weight(lb)$$

According to the model, for every 1 lb increase in weight, while keeping abdomen circumference constant, body fat percentage decreases by 0.14%. Conversely, for every 1 cm increase in abdomen circumference, while keeping weight constant, body fat percentage increases by 0.92%. We divided the population into five categories based on body fat, and from Plot2, it is simple to see which category one belongs to.

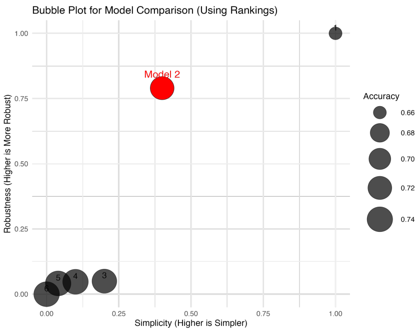
Model Diagnostics and Strength and Weakness:

The model was assessed through residual analysis and significance testing. The residuals follow a normal distribution as confirmed by the QQ plot. Both the abdomen and weight coefficients are statistically significant, with p-values much smaller than 0.05. Additionally, there is no significant multicollinearity between the variables, as indicated by a VIF value of 4.77, which is below the threshold of 5.

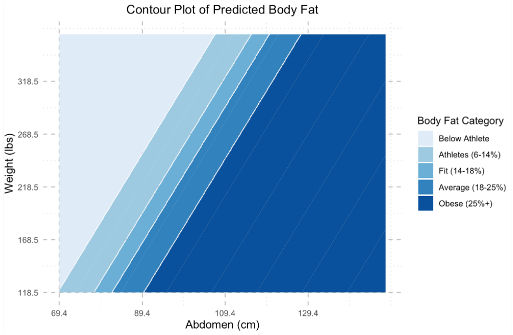
The main strength of the model is its simplicity, as it uses only two easily measurable variables while explaining 72.1% of the variation in body fat percentage. However, the model's predictions are not very accurate, with only 12.7% of predictions within ±3% of the true value, and 32.94% of predictions falling within ±10% of the true value. Thus, while useful, the model could be improved to enhance prediction accuracy.

Conclusion:

We developed a simple model using abdomen circumference and weight to predict body fat percentage, explaining 72.1% of the variability. While practical and easy to use, its accuracy is limited, with only a third of predictions within ±10% of the true value. Future improvements could boost precision by adding more features or refining the model.



Plot1



Plot2

Contribution：

Kangxin Zheng was in charge of the data description and cleaning part, and made constructive suggestions for the Shiny app, such as suggesting the addition of the ability to add images and frame the images based on the results.

Tianyu Yao was in charge of the model selection and building part and add error reporting for shiny app.

Ruijing Chen was in charge of the final modeling part as well as the making of shiny.

We wrote the R code, summary, and ppt of our respective parts, and each of us integrated one in the end. We had three meetings.

Reference:

1.Body Fat Prediction：

https://www.kaggle.com/code/yeganehbavafa/body-fat-prediction

2.Body Fat Prediction Dataset:

https://www.kaggle.com/datasets/fedesoriano/body-fat-prediction-dataset

3. Kabacoff, R. I. (2015). R in action: Data analysis and graphics with R (2nd ed.). Manning Publications.