Body Fat % Prediction

Group 09
Kanishk Saxena, Pushpit Kaushik, Yu Luan

Executive Summary

In our effort to find an easy, accurate, and affordable way to estimate body fat percentage, our team analyzed data from 252 men who had their body fat and body measurements recorded. The main goal of our project was to create a simple and practical method, using everyday measurements, to estimate body fat accurately. This would eliminate the need for complicated and expensive tests.

This summary outlines our discoveries and the methods we used in a straightforward manner. It's designed to help other data scientists understand what we did and maybe even do the same kind of analysis.

Before the analysis, we did the data cleaning part to neglect some missing values and impossible body fat values through the density and its original body fat values. After the cleaning process, our analysis revealed that a straightforward linear regression model, using specific body circumference measurements as predictor variables, can offer a remarkably accurate estimate of body fat percentage. Through a rigorous examination of the dataset, we discovered that the ABDOMEN, CHEST, HIP, WEIGHT, THIGH are the most significant predictors through the Pearson correlation matrix's value. (This matrix is used to capture the linear relationship between the predictor and features.) These measurements are readily available and can be easily obtained, making our model practical for clinical use and self-assessment.

The resulting "rule-of-thumb" model has a high degree of accuracy, with a coefficient of determination (R-squared) of 0.717 , indicating that it explains 71.7 of the variance in body fat percentage. Our model is not only robust but also easily interpretable, as it relies on straightforward measurements that do not require specialized equipment or expertise.

To arrive at our findings, we conducted a thorough statistical analysis. First, we performed data exploration, which included examining the distribution of body fat percentage and predictor variables. We identified potential outliers and leverage points that could affect the model's performance. Through data visualization, it became apparent that the selected predictors exhibited a linear relationship with body fat percentage.

Next, we used a straightforward approach called linear regression. In this, we treated body fat percentage as the main thing we wanted to figure out (the "dependent variable"), while the sizes of the abdomen, thigh, hip size, chest and a person's weight were the things we thought could help us figure it out .From the correlation matrix we found these five variables seemed highly related to each other. So we performed recursive feature elimination first to help us maintain high performance and also can help us remove the variable. During this process, we removed the redundant variable chest. Then for our goal of robustness, we divided the dataset into training and testing parts. 80% belongs to training and 20% belongs to testing. Then based on the metric of RMSE, we use the thigh, weight, and abdomen as our final features to predict body fat value.

We also looked at how well our method worked by checking something called "residuals." This is just a fancy word for the differences between what our model predicted and what we observed. These differences were pretty even and not all over the place, which is good. We also made sure that the size measurements we used didn't confuse each other, and it turned out they didn't. They were all independent and played their own roles in helping us estimate body fat percentage.

Our model, though effective, is not without limitations. Our model only focuses on the linear trend, it may lose the information to capture some nonlinear relationship . Also we treated the part of data’s features independently, we ignored the effect of interaction terms. The data's limited size, consisting of only 252 observations, may also restrict the model's generalizability.

In conclusion, our analysis has produced a simple, robust, and accurate "rule-of-thumb" model for estimating body fat percentage in men. The combination of ABDOMEN, WEIGHT, THIGH provides a cost-effective and convenient alternative to traditional body fat measurement methods. This work lays the foundation for further research and validation on larger and more diverse datasets to refine the model and make it applicable to a broader population.

References & Contributions:

Shiny app reference: https://medium.com/mlearning-ai/machine-learning-app-with-shiny-8c088f2f4646
Chatgpt code:
Line 90-148 is modified and improved by chatgpt.
Training and testing data reference:
https://rpubs.com/cliex159/881990
Slides:
Taken reference and background knowledge from assignment document.
http://jse.amstat.org/v4n1/datasets.johnson.html (Katch and McArdle (1977), p. 113)

Contributions	**Kanishk**	**Pushpit**	**Yu Lan**
Github Repo	1. Create the github repo from scratch. 2. Create all the folders for the repo as directed in the assignment. 3. Create a proper ReadMe for the repo so that it is easy to understand for the users. 4. Sharing the github repo with everyone in the team.	1. Collaborated with the team to make sure there are no conflicts while pushing the code. 2. Pushed and published the Shiny app.	1. Collaborated with the team, pushed and revised the code of the model. 2. Set up and make the whole model and analysis code.
Executive Summary	1. Create executive summary document from scratch. 2. Apart from the code and app perspective, I added everything in document talking about introduction and cleaning of data. 3. Discussion with the team on the executive summary	4. Communicated and collaborated with the team to understand and mention important findings, advantages and disadvantages of our model.	1. Communic-ated with team members and modify the analysis part. 2. Add some sentences to include all our effort in the summary
R Shiny app	1. Gave valuable inputs to Pushpit on how can we create a good user interface so user can easily use the app to get body fat %. 2. Discussions with Pushpit on improving the app	1. Developed the Shiny app after evaluating the linear model and understanding the attributes contributing to it.	1. Provide part of plot code and the model code to Pushpit 2. Discussion with Pushpit to improve his model
Presentation	1. Made the presentation document from scratch. 2. Added the first slide, introduction, background and last 2 slides. 3. Discussion with the team on how to split the speaking part.	1. Decided and rehearsed on the contributions to the presentation and demo the Shiny app.	1. Finish the part of model fitting data cleaning and selection from slide 4-12 2. Make and include all plots and table