Introduction

Body fat percentages, historically, are measured by calipers to determine the thickness of fat underneath the skin (Durnin et. al, 1974). This is not an easy task to complete from home. Many new methods have surfaced using simple measures people can complete by themselves including well known factors of height, weight, and age. Given a dataset with various measures of body circumference from 252 males we strive to produce a rule of thumb that can easily be used based on one's own measurements to give a quick and robust body fat percentage estimate.

Data Cleaning

By running boxplots on each variable, two specific outliers were identified. Two individuals had a reported body fat percentage below 4.0% (0% and 1.9%), which according to Gallager et al. (2000) is scientifically impossible for healthy men of any race.

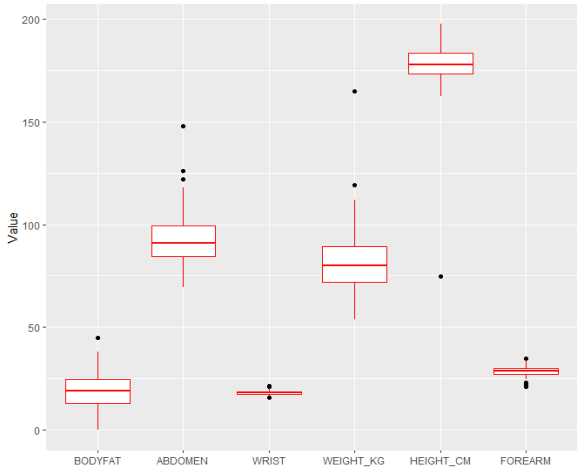

Figure 1

We initially wanted to reproduce the body fat percentage from other reported measures using outside sources. Upon running randomly chosen observations' information through said calculators, the results were too far off the dataset's reported body fat and led us to the decision to delete these two individuals. The boxplot of height (Figure 1) displayed another extreme outlier with height of 29.50 inches which is scientifically impossible for healthy men without dwarfism (Mayo Clinic Staff, 2018). To fix this, we used the formula for BMI and recalculated height from this individual's weight and adiposity measures. We also took out an additional outlier when running model diagnostics, specifically Cook's Distance (Figure 2). This specific individual had vastly large measures across many variables and a Cook's Distance of > 0.5.

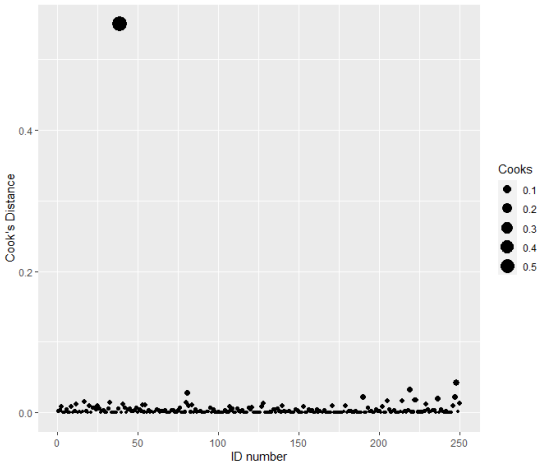

Figure 2

Model Building

For the simplicity concern, we use stepwise selection to choose variable. The stepwise selection consists of iteratively adding and removing predictors, in the predictive model, in order to find the subset of variables in the data set resulting in the best performing model, that is a model that lowers prediction error. The final variables that are contained in the model are: weight, wrist and abdomen which means these variables have the biggest effect on body fat. Then, we use these variables to build a linear regression model.

Final Model

Our final model was produced as the following:

$$\begin{aligned} Body\ Fat\ \% \ = \ & -22.894 \\ & - 0.193 Weight(kg) \\ & - 1.336 Wrist \\ & + 0.886 Abdomen \end{aligned}$$

A rule of thumb for men to measure body fat percentage is to:
Multiply 0.9 by abdomen (cm), subtract 0.19 times weight (kg), subtract 1.4 times wrist (cm), and subtract 23.

On average, for every 10cm increase in abdomen circumference, we can expect close to an 8.9 % increase in body fat, holding weight and wrist measures constant. Due to the positive relationships between almost all variables, we see a negative intercept value in our model. This means, when all predictors are set to zero, the expected value of body fat is -23.894. Additionally, the negative coefficients on weight and wrist measures fall in line with the logic that it is possible to gain weight or wrist size from other sources than fat. These increases could be due to muscle gain or higher bone density, leading to the proportion of body fat relative to weight decreasing as the smaller negative coefficients suggest.

Model Validation

The Q-Q plot (Figure 3) suggests that though slightly skewed at both tails, the residuals are approximately normally distributed.

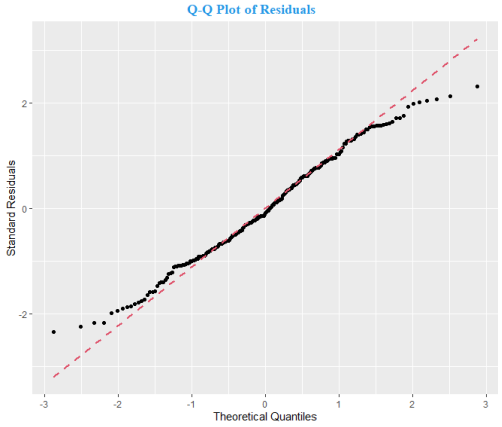

Figure 3

From Figure 4, the residuals "bounce randomly" around the 0 line, suggesting the assumption that the relationship is linear. The residuals roughly form a "horizontal band" around the 0 line, suggesting the variances of the error terms are equal. No one residual "stands out" from the basic random pattern of residuals, suggesting there are no outliers.

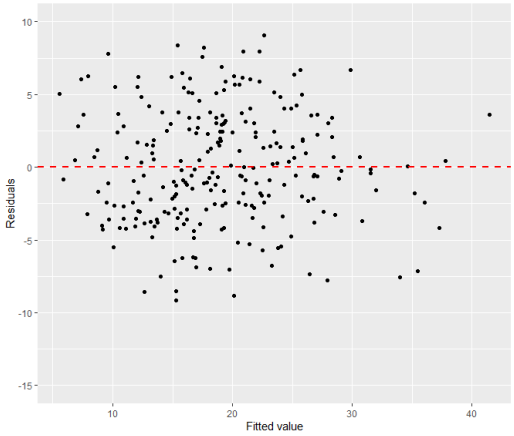

Figure 4

Model Analysis

To test the strength of our model, we use R-squared. R-squared is a statistical measure that represents the proportion of the variance for a dependent variable that is explained by an independent variable or variables in a regression model. In our model, R-squared is 0.7271 which means that approximately 72.71% of the observed variation can be explained by the model's input. We ultimately chose our final model when there was less than a 1% change in R-squared respective to one additional predictor.

Discussion

Despite the shown strengths of our model, it does have a few downfalls: our model is more precise when used for people with average figures, ignores the possible collinearity between variables, and it is a static model. For improvement, we could produce an iterative algorithm to build a dynamic model once we get new inputs that are not outliers.
All in all, we believe we have developed an easy, yet robust way of calculating body fat percentage. Completed with a pared-down version also acting as a great rule of thumb for quicker calculations.

References and Contributions

Durnin, J. V., & Womersley, J. V. G. A. (1974). Body fat assessed from total body density and its estimation from skinfold thickness: measurements on 481 men and women aged from 16 to 72 years. *British Journal of Nutrition*, 32(1), 77-97. https://doi.org/10.1079/bjn19740060

Gallagher, D., Heymsfield, S. B., Heo M., Jebb, S. A., Murgatroyd, P. R., Sakamoto Y. (2000) Healthy percentage body fat ranges: an approach for developing guidelines based on body mass index. *The American Journal of Clinical Nutrition, 72*(3), 694-701 https://doi.org/10.1093/ajcn/72.3.694

Mayo Clinic Staff. "Dwarfism." *Mayo Clinic*, Mayo Foundation for Medical Education and Research, 17 Aug. 2018, https://www.mayoclinic.org/diseases-conditions/dwarfism/symptoms-causes/syc-20371969. Accessed 15 Oct. 2022.

CD created draft for stepwise selection on linear model. Used ggplot2 to create all clean plots used in summary and slides (Residuals vs. Fitted, QQ, Cook's Distance, Boxplots, PCA selection, Final Model with Confidence Interval). Wrote weakness and further improvement. Created GitHub repository and add all collaborators. Ran stepwise selection on GLM, created PCA plus linear regression model (both are shown in the "initial thoughts" slide and GitHub). Formatted summary. Edited & Formatted "Initial Thoughts", "Model Validation" on slides.

CZ ran stepwise selection on GLM, created PCA plus linear regression model (both are shown in the "initial thoughts" slide and GitHub), the draft for stepwise selection on linear model. Revised the strengths and weakness and built the framework for the slides.

RS created and edited code for data cleaning (CD – code for clean graphs), simple linear model selection and diagnostics (with draft plots) (CD - code for clean plots), and the Shiny app. Set up the GitHub repository structure. Wrote and edited the introduction, data cleaning, and final model paragraphs of the summary. Wrote and formatted "Data Cleaning," "Model Development," and "Final Model" slides of the PowerPoint.