Introduction

In today's fast-paced world, people are increasingly paying attention to their health. According to the World Health Organization and numerous clinical studies, there is a direct correlation between increased body fat and the risk of heart diseases (“Obesity”). Therefore, body fat percentage can serve as a good indicator of one's physical condition. This report explores the relationship between various body measurements and their potential predictive power for body fat in U.S. adult men. Through statistical analysis, our goal was to find a simple, robust, and accurate model for individuals to measure body fat in adult men.

Background Information/Data Cleaning:

The original dataset has 252 rows, each representing a male individual, and 17 columns, each representing a body measurement. First, we converted all metrics to the metric system. Next, we used density to estimate their body fat, correcting for miscalculated body fat data. According to a report published by the CDC and common sense: we excluded men who weighed more than 150 kg (IDNO: 39), had more than 40% body fat (IDNO: 216), and were less than 150 cm tall (IDNO: 42). Since body fat = 495 / density - 450 and body fat should be at least greater than 3%, it can be concluded that density is at least less than 1.093 (IDNO: 96, 172, 182) (Fryar CD). Ultimately, the remaining dataset we can use is 246 individuals with 14 possible predictors.

Final Model

The final model chosen was BODYFAT = -10.1784 + 0.7132*ABDOMEN - 2.0181*WRIST. The estimated coefficients are 0.7132 for abdomen and -2.0181 for wrist. The coefficients are in units of body fat percentage per centimeter. Assuming wrist measurement remains constant, each centimeter increase in abdomen increases the body fat percentage by 0.7132. Assuming abdomen measurement remains constant, each centimeter increase in wrist decreases body fat percentage by 2.0181. Therefore an average man with a 102cm abdomen and 17cm wrist is expected to have a body fat percentage of 28.26 based on our model. The 95% prediction interval for his body fat is between 26.81% and 29.70%.

Rationale for your Final Model

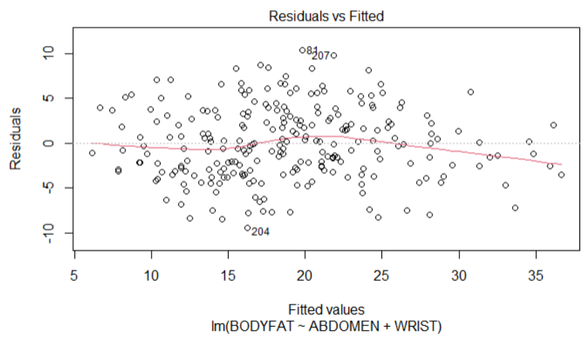
Through correlation testing and using a decision tree, we found that the variable abdomen was the best at predicting body fat. Other variables that showed promise at playing a predictive role in body fat were height, ankle, neck,and wrist.We built three main models: abdomen + height (model 1), abdomen + wrist (model 2), and abdomen + neck (model 3). We used r-squared to measure how accurate our model is. Using just the abdomen variable as a predictor gave an R^2 value of 0.6576. The model 1 has an R^2 value of 0.6939, which means that the model explains about 69.39% of the variation in body fat. The model 2 has an R^2 value of 0.6977. The model 3 has an R^2 value of 0.6863. Model 2 had the highest R^2 value out of all of the models and is still simple with the two predictors of abdomen and wrist.

Relevant Statistical Analysis

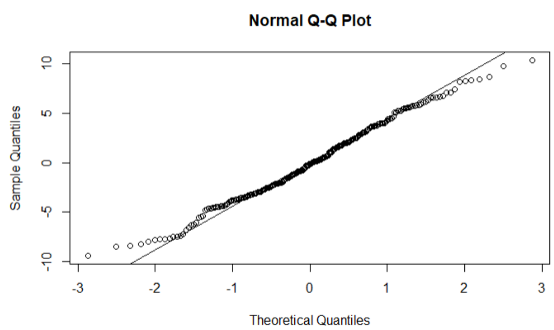
In order to see if the predictor variables we chose are significant in predicting the outcome, we performed the following tests. Our null hypothesis was the predictor variable has no relationship with the outcome, coefficient is 0. The alternative hypothesis being that the predictor variable has a relationship with the outcome, coefficient is not 0. We then used the t-statistic for this test and drew conclusions from the p-value. Both abdomen and wrist had p-values well below 0.05, indicating a significant relationship with body fat, while all are under the Type I error that we are willing to tolerate. The R^2 value is an important indicator of how good a model is, and a higher R^2 value means that the model is better able to explain changes in the data. Our final model has an R^2 value of 0.6977 which means the model explains about 69.77% of the variation in body fat.

Model Diagnostics

We checked the following assumptions for linear regression: linearity, homoscedasticity, normality, and multicollinearity. To test for linearity and homoscedasticity we can plot the fitted values vs the residuals as shown below. A concern is that the relationship between the predictors and target variable is not entirely linear due to the red line on the graph not being mostly flat at zero. The spread of residuals also don't look consistent which could indicate heteroscedasticity. Trying a few variable transformations to fix this, did not make any significant improvements.



To test for normality of the residuals we used a Q-Q plot as shown below. Most of the residuals follow the straight line in the plot so we can assume normality.



To test multicollinearity we used variance inflation factor (VIF). Both abdomen and wrist variables had VIF of about 1.57 which indicates little to no multicollinearity since the score is well below five.

Model Strengths/Weaknesses

Some strengths of our model include a high R^2 value, significant predictor variables, and its simplicity. Overall the model is able to help people be able to attempt to predict their body fat at home by themselves. Some weaknesses of our model include limited data size which allows the prediction only for adult males, not strong assumptions of linearity and heteroscedasticity, and the abdomen size can vary throughout the day ("Stomach"). There could be further improvements to the model but this would make the model more complex.

Conclusion

In conclusion, our model suggests that the percentage of body fat in adult men can be estimated relatively accurately by measuring abdominal and wrist dimensions, but its reliability may be limited by the fact that the sample used in analysis is not representative of the entire population and the number of variables selected.

References

Fryar CD, Carroll MD, Gu Q, Afful J, Ogden CL. Anthropometric reference data for children and adults: United States, 2015–2018. National Center for Health Statistics. Vital Health Stat 3(46). 2021.

"Stomach." Cleveland Clinic, Cleveland Clinic Foundation, Publication Date (if available), https://my.clevelandclinic.org/health/body/21758-stomach.

"Obesity." World Health Organization, https://www.who.int/health-topics/obesity#tab=tab_1.

"National Library of Medicine" https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3854278/

Contributions	Kai Shukla	Abigail Sikora	Yeming Wei
Presentation	Slides 2,3, and 12 Reviewed/edited all slides.	Slides 8,9,10,11. Reviewed/edited all slides.	Slides 4,5,6,7. Reviewed/edited all slides.
Summary	Responsible for the Introduction section. Reviewed all sections.	Responsible for Rationale for Final Model, Model Diagnostics, Strengths/Weaknesses ,and References sections. Reviewed/edited all sections.	Responsible for Rationale for Final Model, Relevant Statistical Analysis, and Conclusion. Reviewed/edited all sections.
Code	Created README and Shiny App Code. Reviewed all code.	Responsible for images in the slides and summary. Responsible for some data cleaning code. Reviewed all code.	Responsible for some data cleaning code and analysis code. Reviewed all code.
Shiny App	Created the Shiny App	Reviewed/edited and provided feedback on Shiny app	Reviewed/edited and provided feedback on Shiny app