STAT 628 Module 2 - Body Fat Percentage: Executive Summary

Body fat percentage can be used as a measure of an individual's overall health and fitness. However, common methods of estimating body fat percentage, such as skin fold caliper measurements or using body density as determined by water displacement, require specialized tools and/or are inconvenient to measure[1,2]. A 'rule-of-thumb' body fat model allows people to easily approximate their body fat percentage. This summary describes the development of a male body fat percentage model, with the goal that the model is simple, accurate, and robust.

For data cleaning, we first converted to metric units for HEIGHT and WEIGHT. Then we verified the derived BMI column. 10 rows were calculated incorrectly, and it was observed that there were abnormal heights in the data, such as the 42nd row whose height is only 0.7 meters. Hence chose to believe the remaining two columns of data and fixed the incorrect height. Second, it was found that there were impossible body fat percentages in the data, such as row 182 with 0% body fat, whereas a normal male has more than 5% body fat. So we imputed the data using the body fat percentage formula with DENSITY[3]. Since the results were still abnormal, we instead used the BMI method[4] to impute body fat percentage for those rows.

Models were evaluated for their simplicity, accuracy, and robustness. To keep the model simple, we limited our analysis to only linear models with no more than 2 parameters. Accuracy was measured using the correlation coefficient (R^2) , a measure of the linear relationship between an outcome variable and its predictors on a scale of 0 to 1. Robustness was measured by checking the sum of square differences against a prior model: the US Navy method[4].

First, each possible 1-parameter linear model was evaluated for accuracy, and it was found that abdomen circumference was the best predictor with respect to R^2. Then each 2-parameter combination of ABDOMEN and another predictor was evaluated to try to improve the accuracy metric. From this analysis, 3 models stood out: ABDOMEN, ABDOMEN + HEIGHT, and ABDOMEN + WEIGHT.

Our final model is: Body Fat ~ -36.92+ 0.88*ABDOMEN-0.31*WEIGHT. We selected this model as it achieved the highest R^2 of 0.69 and the lowest Sum of Squared Differences (SSD) of 7.98. The R^2 indicates that approximately 69% of the variance in body fat percentage can be explained by these two predictors and the p-values for all predictor variables in all models were significant. This indicates our model can explain changes in body fat percentage well, and all variables are important.

We evaluated the robustness by calculating the Sum of Squared Differences (SSD) for 4 data points with the US Navy method. The SSD of the model is 7.98. A lower SSD indicates that our model's predictions closely align with the prior model, enhancing the model's credibility. To assess collinearity, we also calculated the Variance Inflation Factor (VIF) values (Table 1). The VIF for our final model was 4.73, suggesting low collinearity among the variables. Typically, VIF values below 5 are considered acceptable[5], indicating that coefficient estimates are stable.

Model	**ABDOMEN+WEIGHT**	**ABDOMEN+HEIGHT**
VIF value	4.73	1.04

Table 1: VIF values for the 2 factor models

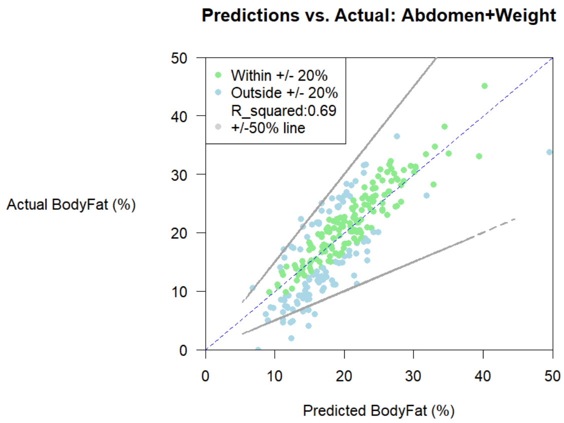

Figure 1: Fitted vs. Actual plot

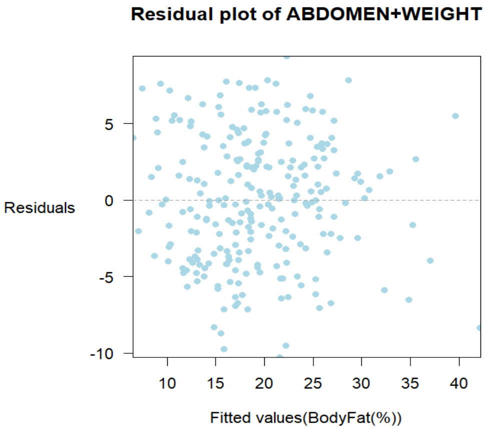

Figure 2: Residual plot

In the Fitted vs. Actual value plot (Fig.1), the blue dotted line is the fitted line, which shows that the intercept approached 0 and the slopes approached 1. This validates the strong agreement between predicted and actual values. In this plot, we highlighted predictions within ±20% of the true value in green, accounting for 147 out of 252 data points (whereas the other two models had 145 out of 252). The two gray lines indicate the ±50% range of the true value, where 230 out of 252 data points fall within this range.

Additionally, the residual plot (Fig.2) revealed no clear patterns, supporting the linearity and homoscedasticity assumptions of this model. The lack of patterns suggests that the residuals are randomly distributed and consistent across levels of the independent variables, reinforcing the model's robustness and enhancing the reliability of its predictions.

Some strengths of our model include strong fit, simplicity, and robustness. Due to the simplicity of the model, it offers strong interpretability, making it easy to convey the logic behind predictions to non-experts. The model is robust to small noise in the input data, leading to stable predictions.

When a model reports that only 58% of predictions fall within ±20% of the true value, this highlights important limitations in its predictive accuracy. The relatively low proportion of accurate predictions could lead to incorrect body fat assessments, potentially impacting health-related decision-making or interventions.

In conclusion, the chosen model using abdomen circumference and weight provides a simple and robust metric to estimate male body fat percentage, with accuracy equal to or better than other simple models. Further improvements could be made by using an expanded dataset to develop the model, especially if there is enough data for a test/train split so that test error can be properly evaluated. Adding complexity to the model via additional variables or considering non-linear models could also improve the accuracy in future iterations of this work.

References

1. Bailey, Covert (1994). Smart Exercise: Burning Fat, Getting Fit, Houghton-Mifflin Co., Boston, pp. 179-186.
2. Durnin, J. V. G. A., and Womersley, J., (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 pp. 77–97. Web.
3. Siri, W.E. (1956), "Gross composition of the body", in Advances in Biological and Medical Physics, vol. IV, edited by J.H. Lawrence and C.A. Tobias, Academic Press, Inc., New York.
4. “Body Fat Calculator.” *Calculator.Net*, www.calculator.net/body-fat-calculator.html. Accessed 16 Oct. 2024.
5. Akinwande, M. , Dikko, H. and Samson, A., (2015). “Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis.” Open Journal of Statistics, 5, pp. 754-767. Web.

Contributions

Contributions	Amy Merkelz	Chenyu Jiang	Yifan Zhang
Presentation	Made slides 2,4,5,9-10. Reviewed all slides.	Made slide 3. Reviewed all slides.	Made slides 6,7,8. Reviewed all slides.
Summary	Wrote introduction , model selection, and conclusion sections.	Wrote data cleaning and model strengths sections.	Wrote statistical analysis and model diagnostics sections.
Code	Responsible for robustness test code and provided feedback on result visualizations.	Responsible for data cleaning code and model selection code.	Responsible for result visualizations code for Figures 1 and 2.
Shiny App	Responsible for the Shiny app.	Responsible for contour plot code and provided feedback on the Shiny app.	Provided feedback on the Shiny app.