Executive Summary

Angela Wei, Yanrun Lu, Ruizhen Jing

Introduction

Body fat percentage is a crucial health and fitness indicator, but it is difficult to measure precisely. In the project, we aim to find a simple, robust, and accurate model to estimate body fat from the BodyFat dataset, which contains data from 252 men on 14 body composition variables. We cleaned the data by removing anomalies and outliers to improve its reliability and relevance. We tested several common fitting models and found that the linear model with ABDOMEN and WRIST as the variables was the best choice based on various factors. We will explain our model, assess its performance, and conclude in the next sections.

Background Information & Data Cleaning

Observing the data itself, we find that while most variables seem relatively unimodal, there are certain values that are outrageously valued outliers. For example, there are data points for BODYFAT where there were recorded values at 0% and 1.9%. It is generally acknowledged that essential body fat levels required for men to live is approximately 3% of body weight (University of Pennsylvania). Observing the other end of the scale for BODYFAT, we see that someone was recorded as having 45% body fat. When we compare this against the CDC's 1999-2000 Examination Data - Continuous NHANES (our closest information against 1970), provides the range of 3.20% to 64.20% in body fat for Americans (CDC). Also, even given that it is possible for someone with dwarfism to be 29.50 inches tall, they would not be representative of the average American male, so they also must be removed from the dataset. Other records wherein the values were reasonable, but were identified as outliers, also had to be removed from the data set to support our usage of linear regression given that it relies on the assumption of normality in its independent variables.

Final Model

For our final model, we decided to use a linear regression that depends on `ABDOMEN` and `WRIST` as shown in the equation below:

$$[BodyFat] = 0.73[Abdomen] - 2.08[Wrist] - 9.83 \quad (1)$$

BodyFat: This is the response or outcome variable we're trying to predict or explain. **Abdomen**: This is a predictor variable. For every unit increase in the "Abdomen" measurement, the "BodyFat" is expected to increase by 0.73 units, assuming "Wrist" remains constant. **Wrist**: Another predictor variable. For every unit increase in the "Wrist" measurement, "BodyFat" is expected to decrease by 2.08 units, keeping "Abdomen" constant. **-9.83**: This is the y-intercept. It represents the expected value of "BodyFat" when both "Abdomen" and "Wrist" are 0.

To show this in an example, an adult male whose abdomen circumference is 80 cm and wrist is 17 cm would be estimated to have a body fat percentage of 12.57%.

Final Model Rationale

In order to find the most suitable model for this case we chose three indicators to decide whether a model performs well or not in our research. First, we used R^2 to evaluate the accuracy of the model. Second, the team set a threshold of 3 variables to ensure simplicity in the end model. Third, we divided data into training and test sets with a 75% and 25% split, and compared the R^2 of the models in the test set with the one in the training set to judge their robustness. From there, we choose 3 different kinds of models to observe their performance. Through our exploration, we found that the robustness of Decision Tree was insufficient in comparison and the components of PCA were conflicting and thus hard to interpret. So, we decided to choose a linear regression model. However, using a model with 8 whole variables does not

fulfill our requirements for model simplicity. We applied the leaps library to assess the change in R^2 compared to the number of factors. We found that the R^2 ranged from 0.68 to 0.75, with the biggest jump going from 1(Abdomen) to 2 (Abdomen & Wrist) factors (R^2 of 0.68 and 0.73, respectively). We decided on two factors because there is not a large change in accuracy when going from 2 to 3 factors and it would encourage model simplicity.

After establishing the model, we also needed to do relevant model diagnostics to show whether it is reasonable in statistics. We checked the following four assumptions for the linear regression model.First, we checked the significance of the overall model and the coefficients which are used to prove that the model does work and the predictors actually have influence on body fat. We use a F-test based on the 95% CI to verify the first part and 2 t-tests based on the 95% CI to show the second part. From the results below we can get that the model does work.

	F value	t value	p value
Whole Model	304.5		<0.01
Abdomen		22.8	<0.01
Wrist		-5.97	<0.01

Secondly, we checked the normality of residuals because the fundamental assumption of classical linear regression is that the residuals are normally distributed. Through the Shapiro-Wilk test based on the 95% CI, we got the p-value of 0.13 which didn't reject the normality assumption. Thirdly, we checked that residuals have constant variance across levels of the independent variables for the violated assumption will lead to unreliable t-statistics and p-values. Through White test based on the 95% CI, we got the p-value of 0.88 which didn't reject the assumption. Last but not least, we checked multicollinearity of the two predictors because when predictors are highly correlated, coefficient estimates will be unstable and sensitive to minor changes in the model. For the VIF values were 1.54, we think that there is no significant multicollinearity between ABDOMEN and WRIST.

Model Strengths/ Weaknesses

Linear regression offers notable strengths and weaknesses in its application. On the positive side, it is lauded for its simplicity, ease of interpreting output coefficients, and facilitating clear insights into variable relationships. It can achieve consistent fits even with limited data, enhancing its reliability in small datasets. Its results can also be effectively visualized through scatter plots and regression lines. However, linear regression exhibits weaknesses, notably inaccuracy with too few variables which is balanced by increased complexity with the addition of variables. Additionally, linear models require variable uncorrelation, which is seldom true for real data, and this influences variable selection. Furthermore, the model is sensitive to outliers, which can significantly impact the regression line's slope and intercept, potentially leading to misleading conclusions.

Conclusion & Discussion

In this research, we embarked on a journey to predict body fat percentage using easy-to-measure body composition variables. After comprehensive data cleaning and analysis, the linear regression model incorporating the ABDOMEN and WRIST variables emerged as the most fitting, striking a balance between simplicity and accuracy. While our model showcased its efficacy in predicting body fat, it's imperative to acknowledge the inherent limitations of linear regression, especially regarding its sensitivity to outliers and potential multicollinearity issues. Nevertheless, our rigorous data cleaning, coupled with statistical tests, has mitigated many such concerns. For practitioners and health enthusiasts, our model provides a practical means to estimate body fat without resorting to intricate methods. This study underscores the potential of simple models in making headway into complex health metrics, but it also emphasizes the continuous need for further refinement and validation.

Citations

CDC. “1999-2000 Examination Data - Continuous NHANES.” *Centers for Disease Control and Provention*, 9 March 2019, https://wwwn.cdc.gov/Nchs/Nhanes/Search/DataPage.aspx?Component=Examination&Cycle=1999-2000. Accessed 15 October 2023.

University of Pennsylvania. “Body Composition Information and FAQ’s Sheet.” *PennRec*, 2022, http://pennshape.upenn.edu/files/pennshape/Body-Composition-Fact-Sheet.pdf. Accessed 15 October 2023.

Contributions

	Angela Wei	Ruizhen Jing	Yanrun Lu
Presentation	Created the presentation Responsible for presenting slides 1 - 4	Reviewed/edited and provided feedback on the whole slide. Responsible for presenting slides 5-7.	Reviewed/edited and provided feedback on the whole slide. Responsible for presenting slides 8-11.
Summary	Background Info & Data Cleaning, Final Model, Final Model Rationale Reviewed/edited whole document	Introduction & Model Strengths/ Weaknesses. Reviewed/edited and provided feedback on the whole document.	Final Model, Final Model Rationale Reviewed, Conclusion & Discussion Reviewed/edited whole document
Code	Data cleaning, PCA, Decision tree,	Build model, Visualization	Model diagnosis, Visualization
Shiny App	Started histogram and sliders	Create interactive 3D images	Created the calculator