Estimating Body Fat Percentage Using Linear Regression

Introduction: The primary goal of this project is to create a reliable model to estimate body fat percentage using easily obtainable measurements. Balancing simplicity, robustness, and accuracy, we aim for a tool that's user-friendly and versatile for broad applications.

Data Cleaning: Our initial dataset comprised 252 data measurements, outliers and suspicious data points were identified and addressed. For instance, individuals with extreme values or inconsistencies in their measurements were removed or imputed to enhance the data's quality.

Data Cleaning Summary: We used Z-score method to identify and remove outliners.

Initial rows: 252.

Rows with missing values removed: 10.

Outlier rows removed: 10.

Final rows after cleaning: 242.

Final Model Statement:

The final model developed in this project is a linear regression model that estimates the percentage of body fat using three clinically available measurements: 'Height', 'Abdomen', and 'Wrist'. The model was selected based on its simplicity, robustness, and low cross-validation error. For instance, a man with a height of 180 cm, abdomen circumference of 85 cm, and wrist circumference of 18 cm is expected to have a body fat percentage of 13.68% based on our model. His 95% confidence interval is between 13.03% and 14.33%.

The estimated coefficients for 'Height', 'Abdomen', and 'Wrist' are -0.1528, 0.7228, and -1.4684 respectively. These coefficients represent the change in body fat percentage for each unit increase in the corresponding measurement. For instance, for every one unit increase in height (in cm), the model predicts that body fat percentage will decrease, on average, by 0.15%.

In this project, five distinct machine learning models were evaluated: Linear Regression, Lasso Regression, Principal Components Regression, Decision Trees, and Random Forests. Each model was assessed based on its predictive performance and computational efficiency.

- For the three linear models, Linear Regression has the relatively best performance. Linear Regression, despite its simplicity, demonstrated superior performance with an R-squared value of 0.7160 and a Mean Squared Error (MSE) of 15.6092. However, it assumes a linear relationship between predictors and the target variable, which may not always hold true. It's also sensitive to outliers and multicollinearity among predictors.
- Decision Trees are easily interpretable and can handle both numerical and categorical data. However, they are prone to overfitting, especially with complex data structures. In this case, the Decision Tree model achieved an R-squared value of 0.5843 and an MSE of 22.7793.
- Random Forests are effective for large datasets and can handle numerous input variables without variable deletion. They also mitigate the overfitting problem seen in Decision Trees. However, they are less interpretable than individual decision trees and can be slow in generating predictions once trained. The Random Forest model in this project achieved an R-squared value of 0.6072 and an MSE of 23.27.

Given these evaluations, the Linear Regression model was selected for its balance of predictive accuracy, computational efficiency, and interpretability.

The Linear Regression model emerged as the best model with an R^2 value of 0.7160 and an MSE of 15.6092. This suggests that the Linear Regression model explains approximately 71.60% of the variance in the body fat percentage and has the lowest average squared difference between the predicted and actual values among all models considered. The estimated test error is 16.1673, providing an estimate of how well this model is expected to perform on new, unseen data.

These results highlight the importance of model selection in achieving accurate predictions and underline the effectiveness of Linear Regression.

The final linear regression model exhibits significant statistical properties. The coefficients for 'Abdomen', 'Height', and 'Wrist' are all statistically significant at the 0.05 level based on two-sided t-tests, with p-values less than 2e-16, 0.000429, and 0.000157 respectively. This indicates that these predictors have a significant relationship with body fat percentage. The overall model is also statistically significant at the 0.05 level based on an F-test, with a p-value less than 2.2e-16. This suggests that at least one of the predictors is significantly related to the body fat percentage.

In terms of the direction of the relationships, 'Abdomen' has a positive correlation with body fat, indicating that as abdomen size increases, body fat percentage also increases. Conversely, 'Height' and 'Wrist' have negative correlations with body fat, suggesting that as these measurements increase, body fat percentage decreases.

For model diagnostics in multiple linear regression, we primarily examined three assumptions using various plots. Firstly, we assessed linearity and homoscedasticity with the residual plot. Given that the residuals scatter randomly around the horizontal axis of fitted values, the assumptions of linearity and homoscedasticity seem to be satisfied, even though there's a slight deviation where the smoothed line isn't perfectly horizontal at zero. Secondly, we checked the normality of residuals using a Q-Q plot. As the data points in the Q-Q plot approximately align with a straight line, this suggests that our residuals are approximately normally distributed.

In the Residuals vs. Leverage plot, we identified three outliers. To assess their influence, we individually removed them to evaluate any significant changes to the final model. The results indicated that their impact on the model was negligible, so we decided to retain them.

The chosen linear regression model is advantageous due to its simplicity and interpretability, making it an excellent choice for applications where understanding the model's decisions is crucial. It's also effective, explaining approximately 71.6% of the variation in body fat, indicating a substantial linear relationship between the predictors and the target variable.

However, the model does exhibit multicollinearity, a situation where two or more predictors in the model are highly correlated. This correlation can lead to unstable estimates of the model's parameters and can decrease the precision of the estimated coefficients, reducing the statistical power of the model.

To mitigate this issue, subsequent steps could involve using combined variables or implementing principal component analysis (PCA). Combined variables can help reduce multicollinearity by creating new predictors that are combinations of the original one.

Reference:

James, G., Witten, D., Hastie, T., & Tibshirani, R. (n.d.). An introduction to statistical learning: With applications in R. https://www.statlearning.com/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (n.d.). An introduction to statistical learning: With applications in Python. https://www.statlearning.com/

Contributions:

We separated our task to several parts, for each person, we are going to:

Preethi: cleaning the data, fit and evaluate decision tree model, preparing the summary document.
Yidan: cleaning the data, fit and evaluate the linear regression model, preparing the presentation materials
Bingyan: fit and evaluate random forest model, make the web app, maintaining the GitHub project

Contributions	Yidan	Preethi	Bingyan
Presentation	Responsible for making all slides except the 7 and 8.	Responsible for slide 7. Reviewed all the slides	Responsible for slide 8. Reviewed all slides.
Summary	Proofread the summary. Responsible for model diagnostics part.	Responsible for the entire summary document.	Proofread the summary. Responsible for model diagnostics part.
Code	Data cleaning code and linear regression model code.	Data cleaning code and decision tree model code.	Data cleaning code and random forest model code.
Web App (Flask)	Reviewed the web app.	Reviewed the web app.	Responsible for web app.