Prediction for Body Fat Percentage

Jiayi Li

This project develops a predictive model for estimating body fat percentage using observable physical characteristics, bypassing the need for specialized equipment. Significant variables were selected, and a Multiple Linear Regression Model was built and validated. After identifying and removing influential outliers with Cook's Distance, the refined model showed improved accuracy, demonstrating that body fat percentage can be effectively estimated using common measurements.

Body Fat Prediction | Multiple Linear Regression | Outlier Detection | Cook's Distance | Health Monitoring

Introduction

In this project, we aim to develop a model to predict body fat percentage based on observable characteristics. This model can help individuals estimate their body fat percentage without specialized equipment, promoting healthier lifestyle choices.

Exploration of the Data

The dataset initially consists of 16 variables, capturing various physical and demographic measurements, along with body fat percentage. Before model training, we removed the **Density** variable, as body fat percentage was derived directly from it, which would otherwise lead to circular logic and bias in model predictions. On top of that, we also found that the correlationship coefficient between Abdomen and Waist is equal to 1, which indicates a perfect positive linear relationship. And it would cause errors in prediction model, so we also delete Waist as well.

After cleaning, the remaining variables include:

- **Pct.BF**: Body Fat Percentage, the target variable.
- **Age**: Age of the individual.
- **Weight** and **Height**: Physical dimensions, with weight in pounds and height in inches.
- **Neck, Chest, Abdomen, Hip, Thigh, Knee, Ankle, Bicep, Forearm, Wrist**: Circumferential measurements in inches across various body parts, providing detailed indicators of body composition and distribution.

After these steps, the final dataset used for modeling contains 14 variables, with **Pct.BF** as the dependent variable and 13 predictors.

Building the Model

Finding Candidates. Bidirectional Selection (Stepwise Selection) was used to identify significant predictors. The selection results are shown in Fig. 1.

Using a 5% significance threshold, we selected Age, Height, Neck, Abdomen, and Wrist as the independent variables for the model.

Fitting the Model. The Multiple Linear Regression model constructed with these variables is shown in Fig. 2.

Assumption Checking

To ensure the validity of our model, we checked key assumptions.

Linearity. As shown in Fig. 3, the residuals are randomly scattered around zero, indicating the linearity assumption is met.

Homoscedasticity. As shown in Fig. 3, the residuals plot does not show any discernible pattern, confirming that homoscedasticity is maintained.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.04038    8.35881   0.603 0.547074
Age          0.07258    0.03030   2.396 0.017361 *
Height      -0.26807    0.12612  -2.125 0.034567 *
Neck        -0.45133    0.21774  -2.073 0.039252 *
Abdomen      0.82271    0.06880  11.958  < 2e-16 ***
Hip         -0.19488    0.12984  -1.501 0.134689
Thigh        0.22387    0.12900   1.735 0.083943 .
Forearm      0.29550    0.19166   1.542 0.124440
Wrist       -1.73072    0.49360  -3.506 0.000542 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.241 on 241 degrees of freedom
Multiple R-squared:  0.7469,    Adjusted R-squared:  0.7385
F-statistic:  88.9 on 8 and 241 DF,  p-value: < 2.2e-16
```

Fig. 1. Stepwise Selection Results showing significant predictors.

```
Residuals:
    Min      1Q  Median      3Q     Max
-10.434  -2.997  -0.277   3.326   9.985

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.62108    8.16427   0.566 0.571905
Age          0.05163    0.02399   2.152 0.032385 *
Wrist       -1.58685    0.47129  -3.367 0.000883 ***
Neck        -0.28652    0.20710  -1.384 0.167767
Abdomen      0.80118    0.04004  20.011  < 2e-16 ***
Height      -0.31378    0.12151  -2.582 0.010395 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.269 on 244 degrees of freedom
Multiple R-squared:  0.7404,    Adjusted R-squared:  0.7351
F-statistic: 139.2 on 5 and 244 DF,  p-value: < 2.2e-16
```

Fig. 2. Multiple Linear Regression Model for predicting body fat percentage.

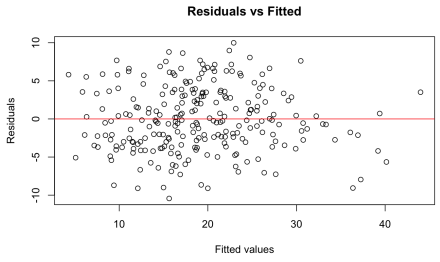

Fig. 3. Residual Plot showing residuals scattered around zero, indicating linearity.

Normality. The points in the Q-Q plot (Fig. 4) fall approximately along the line, indicating normality.

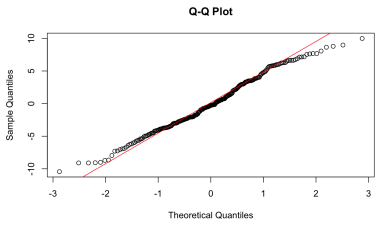

Fig. 4. Q-Q Plot confirming normality assumption.

Multicollinearity. As shown in Fig. 5, all VIF values are less than 5, confirming minimal multicollinearity.

```
     Age    Wrist     Neck  Abdomen   Height
1.258769 2.534915 3.073979 2.281892 1.381069
```

Fig. 5. Variance Inflation Factor (VIF) results indicating multicollinearity levels.

Model Refinement

We refined the model by identifying and removing influential outliers using Cook's Distance.

Outlier Identification. Cook's Distance plot (Fig. 6) identifies points above the threshold as influential, which were removed for model refinement.

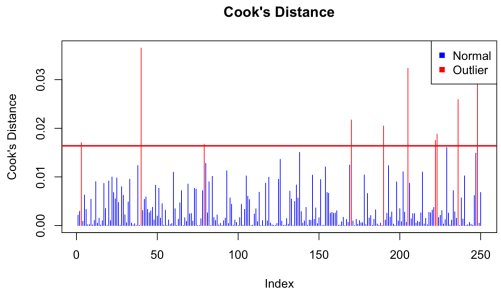

Fig. 6. Cook's Distance Plot showing influential data points.

Refined Model. The refined model (Fig. 7) achieved higher accuracy with improved R-squared and lower residual error.

```
Residuals:
    Min      1Q  Median      3Q     Max
-9.1556 -2.9241 -0.3798  3.0997  8.7763

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.27773    7.97713   0.912 0.362537
Age          0.06008    0.02315   2.596 0.010039 *
Wrist       -1.56251    0.45524  -3.432 0.000708 ***
Neck        -0.29858    0.19725  -1.514 0.131447
Abdomen      0.80802    0.03890  20.771  < 2e-16 ***
Height      -0.36454    0.11817  -3.085 0.002281 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.986 on 234 degrees of freedom
Multiple R-squared:  0.7581,    Adjusted R-squared:  0.7529
F-statistic: 146.6 on 5 and 234 DF,  p-value: < 2.2e-16
```

Fig. 7. Final refined model with improved R-squared and lower Residual Standard Error.

Results

The final predictive model identified **Age, Height, Neck, Abdomen, and Wrist** as the most significant predictors of body fat percentage, with **Abdomen** showing the highest correlation. This suggests that abdominal measurements have a strong influence on body fat prediction, which is consistent with known associations between abdominal fat and overall body fat.

After refining the model by removing influential outliers identified through *Cook's Distance*, the accuracy of predictions improved notably. The refined model demonstrated an **increased R-squared value** and a **decrease in Residual Standard Error**, indicating a better fit and reduced variance in prediction errors.

Model Performance. The refined model's performance was evaluated based on metrics like **R-squared** and **Residual Standard Error**:

- **R-squared**: The model achieved a high R-squared value which is 0.76, signifying that a large proportion of the variance in body fat percentage could be explained by the selected predictors.
- **Residual Standard Error**: The lower Residual Standard Error in the refined model signifies improved precision in the predictions.

Predictor Significance and Interpretation.

- **Abdomen**: As the most influential variable, it highlights a direct relationship with body fat percentage, aligning with findings in health research.
- **Other Predictors** (Age, Height, Neck, and Wrist): These variables, while not as impactful as Abdomen, contribute to refining the body fat estimation, providing a balanced prediction model that does not rely on a single measurement.

Discussion and Conclusion

In the current study, the sample size of the dataset we used was extremely limited and only male, so the current model is likely to be biased from widespread use. Therefore, in the future, we will collect more data to build and train models to improve their universality.