Introduction

Although there are many already-built models and apps, we found that most of them require many inconvenient measurements to use. As such, our goal was to develop a simple and robust method that can calculate a user's body fat percentage through simple measurements and without losing significant accuracy.

Data Pre-Processing

The original data set consisted of 252 observations and 17 variables. During data cleaning, two records (ID's 172 and 182) were deleted, as their body fat measurements were corrupted and unrecoverable by imputation. The extraneous columns of IDNO and Density were filtered out as well, leaving us with 250 observations with one response variable (Body Fat) and 14 predictors.

Motivation for Model

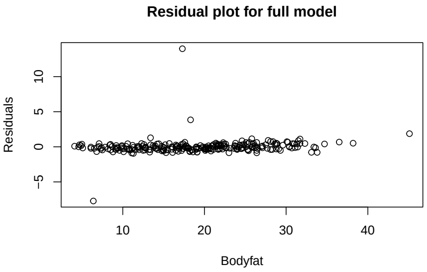

Figure 1: Residual plot for original full model

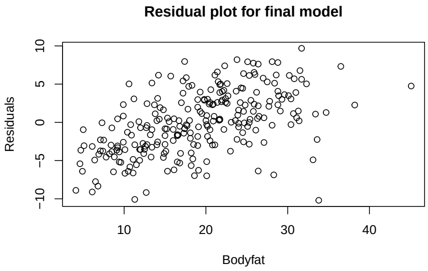

Figure 2: Residual plot for model picked

We chose to use linear regression to build our final model, mainly because the data set is not too large and there is only one dependent variable. At first, we tried to load all predictor variables to our linear model, which gave a 97.83% R^2, which was overfitted and a poor model, as shown by the residual plot.

We then used the Variance Inflation Factor (VIF) to find predictors with high multicollinearity, and removed all predictors that had VIF > 10, leaving us with age, weight, neck, abdomen, hip, thigh, forearm, & wrist. A linear model with these remaining predictor gave a 74.07% R^2, suggesting that our model is still accurate, as well as a better fitting residual plot, as shown in figure 2.

Next, we tested simplified combinations of the remaining variables, and found that using just WEIGHT and ABDOMEN to fit the linear model would give us a similar R^2 value (71.34%), similar residual plot, all with clearer interpretability. This left us with a final model of:

$$BODYFAT = -40.4706 - 0.31 \cdot WEIGHT + 0.91 \cdot ABDOMEN$$

This means that if abdomen circumference does not change, for every 3 KG increase there is in weight, body fat will decrease body fat by roughly 1%, and if weight stays constant, then every centimeter increase in abdomen circumference, will increase body fat by about 1%.

Model Diagnostics

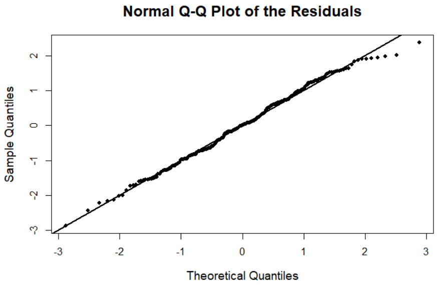

Figure 3: Q-Q plot for prediction

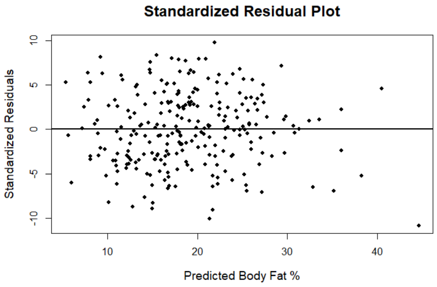

Figure 4: Residual plot for prediction

The Q-Q plot of the final simple linear regression model hugs the 45 degree line closely, although it has some slightly left skew issue and skinny tail. The residual plot of the predict body fat shows the model is reasonable since no obvious trend was observed.

Model Strengths & Weaknesses

Strengths: The model provides good accuracy with only two easy to obtain predictors.

Weaknesses: Because of only using two variables as predictors, this model loses accuracy at the extremities. Additionally, it requires specific units for accurate use.

References

1. **DG: Wrote code for data cleaning, images, wrote Github README, developed 2 page executive summary, reviewed Shiny App**
2. **SL: Wrote code for stepwise model and PCA Model, developed the PowerPoint Presentation, edited summary, reviewed Shiny App**
3. **ML: Wrote code for SLR testing, Developed/Maintains/Deployed the Shiny App, managed Github, proofread/revised summary and PowerPoint**