Body Fat Percentage Prediction

Introduction

BodyFat is an important health indicator. In this project, we use body circumferences to predict body fat percentage.

Data Cleaning

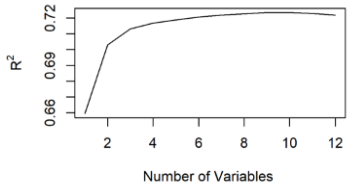
The dataset contains 252 rows and 17 columns. The BodyFat column is used as the response(Y) variable. First, the Idno, adiposity and density are excluded, and there are 13 variables left. Then we analyzed all the predictors. From the analysis summary and plots, some outliers showed up.

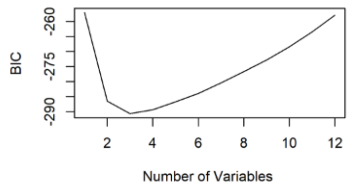
First we used the density function to recalculate the 2 Bodyfat outliers, but they are still outliers after the calculation. So we chose to delete those outlier individuals. We also deleted 2 weight and height outliers. The new dataset has 248 rows. We noticed that the data contains both metric units and US units. For simplification, we unified all our data to metric units.

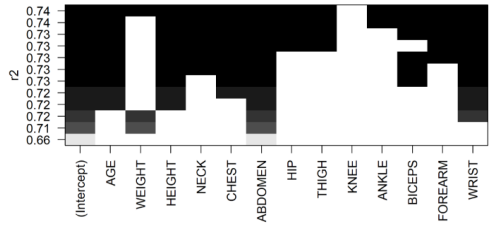
Model Selection

For the reason of robustness and simplicity, we chose Multiple Linear Regression (MLR) with at most 4 predictors as our model. The rule of thumb of our model is: Your abdomen(cm) *0.9-weight(kg)*0.2 - wrist(cm)*1.3 - 23.

The Best Subset Regression was used to select the best subset of predictors out of 13 predictors. We analyzed the R^2 and Bayesian Information Criterion (bic) of all possible subsets, and chose the subset with 3 elements {Weight, Abdomen, and Wrist).







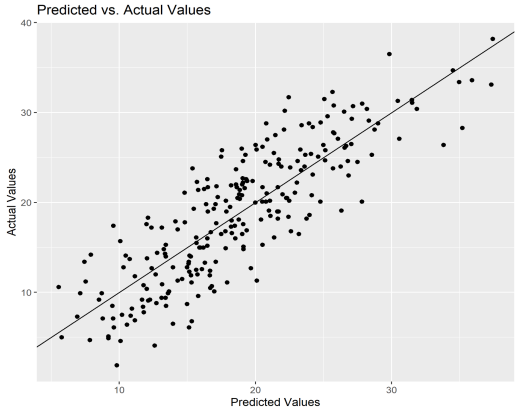
We performed the regression on our selected predictors and Y variable. The results are as follows:

```
Call:
lm(formula = BODYFAT ~ WEIGHT + ABDOMEN + WRIST, data = BodyFat_clean)

Residuals:
   Min     1Q Median     3Q    Max
-9.017 -2.979 -0.370  2.972  9.269

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.85622    6.23905  -3.824 0.000167 ***
WEIGHT       -0.18731    0.04943  -3.789 0.000190 ***
ABDOMEN       0.87468    0.05219  16.760  < 2e-16 ***
WRIST        -1.25673    0.40368  -3.113 0.002072 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.982 on 244 degrees of freedom
Multiple R-squared:  0.7166,    Adjusted R-squared:  0.7131
F-statistic: 205.6 on 3 and 244 DF,  p-value: < 2.2e-16
```



Body fat prediction equation:

$$Body\ Fat\ Pct\ =\ -23.859\ -\ 0.187Weight(kg)\ +\ 0.874Abdomen(cm)\ -\ 1.257Wrist(cm)$$

For example, for a male with weight 85.16kg, Abdomen 96.4cm, and wrist 18.2cm, his predicted body fat percentage would be 21.6%. There is a 95% probability that his body fat is between 13.8% and 29.5%.

The above scatter plot depicts the actual values of body fat percentage versus the predicted values based on our model. The plot shows that points are aligned close to the regression line.

Statistical Analysis

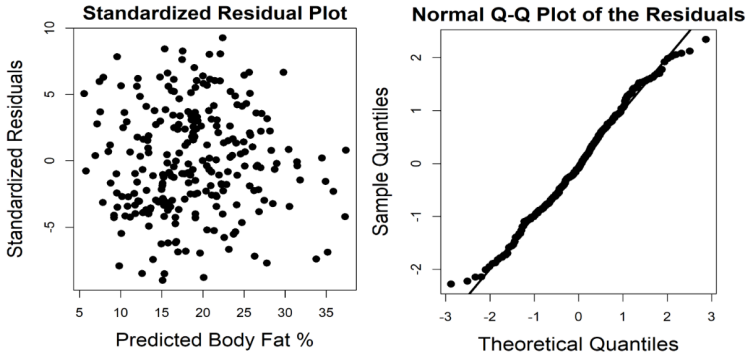
The interpretation of the coefficients can be: For every 1 cm increase in abdomen, the model predicts that the body fat percentage will increase, on average, by 0.874% (holding all other factors unchanged). We found our R^2 to be 0.71, thus the predictors {Weight, Abdomen, and Wrist} explain about 71% of the variation in body fat percentage.

To test the relationship between predictors {Weight, Abdomen, and Wrist} and Bodyfat. We suppose the hypotheses $H_0: \beta_j = 0 \; vs \; H_1: \beta_j \neq 0$, with β_j denotes the coefficients of predictor x_j.

We use the t test statistic. From the model result, we can see all coefficients are significant. Because of the p-value is $2.2 * 10^{-16}$, we can reject the null hypothesis at the $\alpha = 0.05$ level. That is, we can see the relationship between our predictors and body fat percentage is significant.

Model Diagnostics

We diagnosed the MLR assumptions by a standardized residual plot and a normal QQ plot.



1. Linearity and homoscedasticity: The linearity and homoscedasticity is reasonable based on the standardized residual plot.
2. Normally distributed errors: From the normal Q-Q plot of the residuals, the points are very close to the line. Thus the errors seem normally distributed.

Model Strengths/Weaknesses

We believe the MLR between our 3 predictors and Bodyfat is a simple but useful model. The strength of this model is its readiness. The three predictors are all easy to measure. Any person with a tape and a weighing scale can measure their own weight, abdomen and wrist easily.

The weakness of the model is that the R^2 of the model is not very high. The 0.71 R^2 means only 71% of the variation in body fat percentage can be explained. This might be due to the fact that we only selected 3 predictors out of 13 variables for simplicity. Another reason might be that our data size is not large. We might be able to get a better model if we have a larger data size.

Conclusion

On the whole, our model achieves the goal of body fat percentage prediction. A male user only needs to enter his weight, abdomen, and wrist. Then the model will predict the body fat accordingly. Our predictions might not be super accurate, but people can easily get an approximate body fat with nearly no cost.

Contributions:
CL wrote the statistics analysis, model diagnostics, and model weakness/strength part of the summary, worked on slides p9 to p12. CL also revised code to model diagnostics, and was responsible for the Shiny app code.

YZ wrote the model selection and conclusion part of the summary, worked on slides p3 to p8. YZ also revised code related to model selection and created code related to model diagnostics.

ZT wrote the introduction and data cleaning part of the summary, worked on slides p1 to p2. ZT also created the data cleaning and model selection, and plotting portion of the code.