Introduction and Background

Body fat is an important part of the human body because it stores energy, secretes hormones etc. According to US Navy's body fat measurement research, body fat percentage is related to circumference of parts of one's body, such as abdomen and wrist. In our project, we are trying to build a simple model for predicting body fat percentage on a data set containing 252 male samples with 16 measurements on their bodies.

Data Cleaning:

We used box plots and histograms to help us go through the data roughly. We noticed that most of the variables followed normal distribution, but there were some bad data points. Before fitting a model, we removed certain observations (ID:39,42,172,182) since they were considered corrupted. For example, ID 182 has unrealistically low body fat of 0%. Similarly, ID 39 was removed because he has a weight of 363 pounds much larger than the other samples. In this case, our fitted model cannot well predict the body fat percentage among the ultra-obese population, because there is only one ultra-obese point in the training set. And we know that adiposity can be calculate by height and weight. In order to avoid strong correlation between features, ADIPOSITY was removed. Moreover, people usually don't know their body density and it was used to calculated BODYFAT, DENSITY was removed.

Model Fitting

Our final model is

$$Bodyfat\% = 11.13 - 0.16 \times Height + 0.71 \times Abdomen - 1.57 \times Wrist$$

We have 3 variables, height, abdomen, and wrist in units of centimeters, and those three variables can be easily measured. For example, a man with height of 176 cm, abdomen circumference of 86.8 cm and wrist circumference of 17 cm, our model indicates that 95% prediction interval of his body fat percentage is between 8.89% and 24.55%.

Our estimated coefficients for height, abdomen and wrist are – 0.16, 0.71, and –1.57 in unit of centimeters. That is, for every centimeter increment in the one's abdomen, the model predicts that his body fat will increase, on average, by 0.71% and similarly for other features.

We chose this model because of the following reasons. First, due to US Navy's body fat measurement research, we have abdomen and wrist are related to body fat. Then, the table below indicates that the p-values of the 3 variables less than 0.001, which means that they are significant. Also, we have the adjusted R-squared of this model is 0.7246 and the p-value of model less than 2.2e-16. Moreover, other models with 4 significant predictors had 0.7287 R-squared, which is only 0.41% greater than our final model but carried an additional predictor. In other words, an additional predictor only brings us 0.41% increase on R-squared, which hardly dominates the 3-predictor model.

Statistical Analysis, Hypothesis Testing and Inference

We tested the significance of all coefficients. Suppose our null hypothesis is that the slope is equal to 0. The p-value is less than 2.2e-16. The p-value indicated that the probability of getting test results at least as extreme as the results observed (under the null hypothesis) is 2.2e-16, then we rejected the null hypothesis. That is, our model is significant. Our R-square is 0.7246, which

implies that 72.46% of the data fitted well in this model. The estimated slope and estimated intercept are listed in the table to left with 95% CIs.

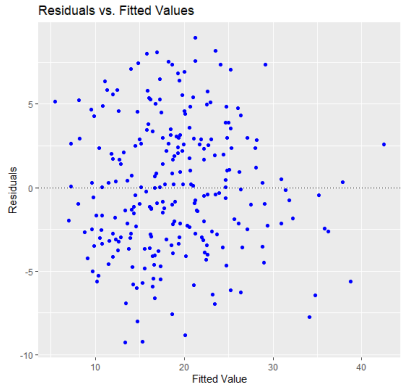
	P-value	95% CI
Intercept	0.129	[-0.25, -0.08]
Abdomen	<2e-16	[-2.30, -0.85]
Height	9.12e-05	[-2.30, -0.85]
Wrist	2.87e-05	[-3.27, 25.52]

Model Diagnostics

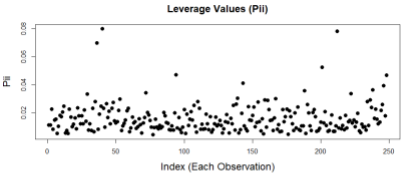
We checked the variance inflation factors of the three variables in the model, and all of them were less than 2, which indicated that there was no strong multicollinearity among the variables.

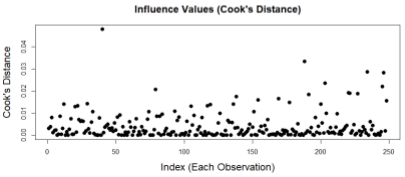
	Abdomen	Wrist	Height
VIF	1.548	1.768	1.184

We also used the plot of standardized residuals and QQ plot to check the assumption for the randomness and unpredictability of the regression model. The points in the residuals vs. fitted values plot are randomly dispersed around the horizontal axis, so we concluded that this model is appropriate for the data.



The plots of leverage value and Cook's distance were used to distinguish highly influential data points and high leverage observations. According to these two figures, the maximum leverage value is less than 0.08 and the Cook's distances of all the points are less than 0.05. Therefore, there was no outlier to be removed from the cleaned data set.





Model Strengths and Weaknesses

Strength

The strengths of our model are absence of multicollinearity among the variables and normality of residuals. In addition, we only need three body indexes (i.e., abdomen, wrist and height) to predict the body fat. Therefore, our model is simple and practical.

Weakness

Since the amount of the data of which the body fat is higher than 35 percent is too small, the linear model is not accurate enough when predicting data with high body fat.

Conclusion

The project aimed to find an easy and efficient way to accurately predict body fat percentage with minimal necessary information provided by the user. A multiple linear regression model was developed to achieve such simplicity and robustness. The essential part of building the model was selections of predictors. Best subset selection was applied and specified a model with lowest BIC and reasonably high R-squared. Meanwhile the model had only three predictors meaning that the model was simple enough. Height, abdomen and wrist are three required input to get a prediction for body fat percentage and these three predictors can be easily measured. All model assumptions were met, and the prediction precision and accuracy were good. Moreover, the model was very easy to interpret with simple linear relations. An improvement that can be made in the future is to use BMI as a replacement of height and weight to fit the model, because all the references we had implied that body fat has a statistically significant relation with BMI, but in general the project objective has been achieved.

Contributions:

We have met over four times within two weeks and every member was highly involved. Generally, every member contributed to every part of the project, but to be specific:

1. Every member in the group contributed to the raw codes of data cleaning, model fitting and model diagnostic.
2. HS wrote Introduction and Background and data cleaning part of the summary. ZZ wrote model fitting and statistical analysis part of the summary. YG wrote model diagnostics and model strengths and weaknesses and summary. SL wrote conclusion part of the summary and reviewed and revised the entire summary.
3. Every member in the group contributed to the creation of the PowerPoint file.
4. HS and ZZ wrote the codes of shiny app.
5. SL built the GitHub repository.

References:

Bohnker, B. K., Sack, D. M., Wedierhold, L., & Malakooti, M. (2005). Navy physical readiness test scores and body mass index (spring 2002 cycle). *Military medicine*, *170*(10), 851 854.