Introduction

Our research is to develop a method to help people estimate body fat percentage to help them assess health. We conducted statistical analysis over the data of a group of people's body information and worked out a numeric model to estimate the body fat percentage.

Background Information/Data Cleaning

1. Our research is based on a dataset of 252 records of the men's body fat percentage, basic information and some other circumference measurements, elements including:
   Age, Weight, Height, Adiposity (also known as BMI), and Circumferences of Neck, Chest, Abdomen, Hip, Thigh, Knee, Ankle, Biceps, Forearm, Wrist
2. Data pre-processing:
   - We explore the data by looking at the dimensions, some rows and variable quantiles.
   - We removed individuals with body fat of 0%, which are practically impossible.
   - We split the data set randomly into two groups--the training group (200 data points) and testing group (the rest), which will be used to build and assess the model.
   - We believe the rest of the data are convincible and neat.

Final Model Adopted

The model we adopted to measure body fat percentage is:

bodyfat = -30.333 + 0.917*abdomen – 0.125*weight – 1.405*wrist + 0.432*forearm

This means that for every 1 cm increase in abdomen circumference, the model predicts that body fat will increase, on average, by 0.917%**.** For example: a man who weighs 154.25lbs with 85.2cm abdomen circumference, 17.1cm wrist circumference of and 27.4cm forearm circumference is expected to have a body fat percentage of 27.27%.

Rule of thumb: multiply your abdomen circumference with 0.58, then minus 34.8, and the result is roughly your body fat percentage.

	Estimate	Std. Error	T-value	Pr(>T)
Intercept	-30.333	6.726	-4.510	<0.001
Abdomen	0.917	0.052	17.679	<0.001
Weight	-0.125	0.023	-5.461	<0.001
Wrist	-1.405	0.409	-3.435	<0.001
Forearm	0.432	0.168	2.567	0.011
Adjusted R^2: 0.727				

We chose this model after the comparison with other possible models. We firstly build our model with the variables mentioned in background info (density, neck, chest, etc.). Then, we find all the influential point, by cook's distance, and remove them. This process is to remove the (abnormal) data points which greatly affect the result. We check the p-value of variables and remove the variables which are not significant (p-value > 0.05). Then we compare several models (like log transformation) and find the one with the least RMSE (about 3.7-3.8) on the testing group.

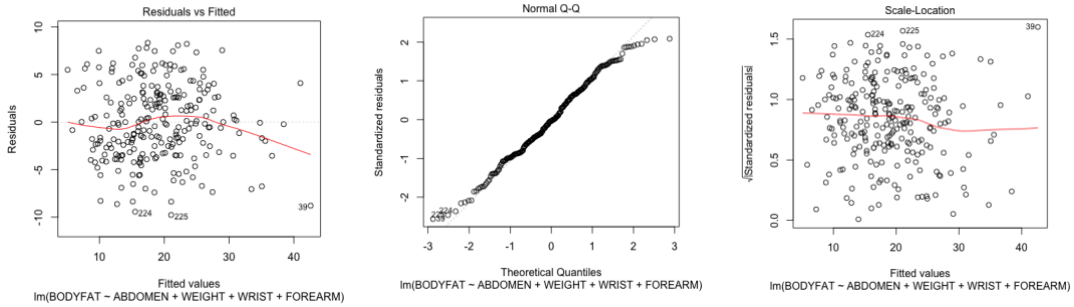
Statistical Analysis

It is shown that all the variables in our model have p-value less than 0.05. This is to say these variables (abdomen, weight, wrist and forearm) are all statistically significant and meaningful to

our model. We also found the R-square to be 0.73, which implies the model explains 73% of variability of the response data around the mean, so it is a simple but strong model.

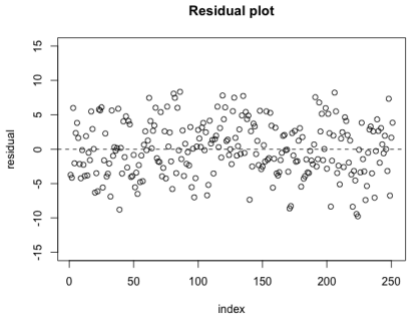
Model Diagnostics

Because we are using linear regressions, there are several assumptions that should be met.



Linearity: From the first plot, we see that the red line is not strictly flat but still acceptable
Normality: From the second plot, we see that the data forms an almost straight line in the diagonal, so we conclude that the assumption of normality is met.
Constant Variance: From the third plot, we see the red line is nearly flat. We can conclude the assumption of constant variance is met.
Multicollinearity: By calculating VIF and correlation between each factor, we find weight has high correlation with other factors. However, since weight explains great portion of the variance of the data, we keep it in our model.

Model Performance



After we fit the model with our data and calculate the residual, we get the graph on the left. We see that the residuals lie around the line of y=0 with no obvious pattern. We can see that the range of residual is not small. However, we are still satisfied with the result as we have small dataset. We are confident that we will have a higher precision with more data.

Conclusion

We found that there is a linear relationship between bodyfat percentage and abdomen, weight, wrist and forearm. This model was achieved after we cleaned data, removed the outliers, used multiple methods to and pick the factors that were most statistically significant. It shows that our model, which explains 73% of variability from the data, and meets most of assumptions of linear regression model. In the future, we can collect more data to train the model for higher precision.

Contribution: HT modified the modeling and conclusion part of summary, worked on data modeling part of slides; **BY** is main drafter of summary and slides, and is responsible for parts of: Intro, Data cleaning, final model, Statistical analysis, conclusion; **JX** is responsible for diagnostics and model performance of both slide and report, and revised and updated both files.