Motivation:

The percentage of body fat is an important indicator of personal health. However, it is hard to implement an accurate way of measurement of body fat. Thus, here we are trying to discover a model to answer this question with other measurements that are much easier to get.

Data Cleaning:

The data set is about 252 men with measurements of their body fat percentage and various body circumference measurements. The Y, body fat percentage, range from 0 to 45.10, has a mean 18.94 and median 19.

During checking the data set, there are several suspicious data points arise:

1. As we know from Katch and McArdle (1977), p. 111 or Wilmore (1976), p. 123, the bodyfat that is calculated from the density shall be accurate. However, there are several outliers whose body fat does not meet the calculated body fat from density. We cannot tell which one is wrong, and the body fat is our Y. Thus, we drop these data points. (48, 76, 182, 96)
2. Another suspicious point is point 42. It only has a height of 30, which is below half of all other data points. There is some error in the data, or there is some exceptional disease (that person is over 40 years old, thus cannot be explained by age). Either way, we think it does not suit our data set.
3. There are also some points (e.g., 39 and 41) that have measure values away from the majority. However, after look deep into those data, we think the data is reasonable and can be explained with extremely obsess.

Purposed Model:

Our final model is:

$$BODYFAT = -47.93 + 1.164 * ABDOMEN - 1.433 * WRIST + 0.06032 * WEIGHT - 0.001518 * WEIGHT * ABDOMEN$$

And, in order to make it easy-to-use, we think the rule of thumb shall be:

$$BODYFAT = -48 + 1.2 * ABDOMEN - 1.4 * WRIST + 0.06 * WEIGHT - 0.0015 * WEIGHT * ABDOMEN$$

So, for example, a man with 104.3 as the abdomen, 18.8 as wrist, and 212 as weight is expected to have a body fat percentage of 30.3926, and the real body fat percentage is 30.8.

Our estimated coefficients are -47.93, 1.164, -1.433, 0.06032 and -0.001518, which are in the centimeters, centimeters, lbs and lbs*centimeter. This means that for every abdomen increase in 1 centimeter, the model predicts that body fat % will increase, on average, by 1.2. Every wrist increases in 1 centimeter, the predicted body fat % will decrease, on average, by 1.4. Every weight increase in 1 lb, the model predicts that body fat % will increase, on average, by 0.06. Every result of weight times abdomen increases in 1 centimeter*lbs, the predicted body fat % will decrease, on average, by 1.4.

We chose this model because of the following reasons.

First, it is common sense that body fat is related to weight and abdomen, so we organized our model based on these traits. Second, we searched over different model based on BIC, and used cross validation to choose models. This model has the best mean R-squared after 12-folds cross-validation. Third, the prediction ability is stable in different groups, which satisfies the robustness requirement.

Statistical Inference

As mentioned in the previous section, our body fat prediction model is based on wrists, weights, and abdomen. We conducted the following t-test to see the significance of the parameters we chose. The null hypothesis for each parameter is zero, which means this

parameter is unrelated to bodyfat. We will do a t-test to check the correctness of the hypothesis. After calculating the t-test statistics, we obtained the following results.

	Estimate	Std.Error	t value	Pr(>|t|)
Intercept	-4.770e+01	9.007e+00	-5.296	2.80e-07
WRIST	-1.621e+00	4.250e-01	-3.814	1.76e-04
ABDOMEN	1.181e+00	8.822e-02	13.392	<2e-16
WEIGHT	8.205e-02	4.919e-02	1.668	0.096
ABDOMEN:WEIGHT	-1.647e-03	3.999e-04	-4.118	5.35e-05

We will analyze these statistical results by focusing on the parameter "weight," and the rest are similar. As we can see, the estimated coefficient of weight is 8.205e-02, which means that since the weight increase 1 unit, the bodyfat will increase 8.205e-02 unit, roughly speaking. Then we notice that the t value for weight is 1.668, and the p-value is 0.096. We can reject the null hypothesis if the Type I error we are willing to tolerate is 10%. In other words, we can reject the null hypothesis 90% sure, which means that parameter "weight" is significant to the bodyfat. We can analyze other parameters in the same way. It shows that all these parameters are significant to bodyfat.

We also do the F-test, and our null hypothesis is that all the coefficients of parameters are equal to zero. We obtain the p-value of F-statistics is less than 2.2e-16, which gives us the confidence to reject the null hypothesis.

Finally, we also check the R square, in other words, the coefficient of determination, whose value is 0.748. It means that 74.8% of body fat variation can be explained by the parameters we used. It is also a strong proof to show our model's performance.

Model Diagnostics

We checked the following four assumptions for MLR.

First, we checked linearity using residuals v.s. fitted value plot. (see Figure 2). Since no clear patterns were discovered, we believed linearity is plausible, even though there are slight linearity violations. Second, we checked the normality assumption with a normal Q-Q plot (see Figure 3). We believe normality is satisfied. Thirdly, we checked for the equal variance assumption. The plot (see Figure 4) shows there might be validation on the assumption, so we performed a non-constant variance score test. The p-value is 0.42, so we believe the assumption is satisfied.

We also checked for outliers. No outliers were detected after the data cleaning process.

Model Strengths/Weaknesses

Pros: Since we search all models with level-1 interaction, we can think that our model's accuracy is comparatively high. Also, we examed our assumptions, so our result is more reliable.

Cons: We did not set a hard limit for the number of parameters, so it might still be a little bit hard to calculate in mind.

Conclusion/Discussion:

In Conclusion, we came up with a model that men can use to determine their body fat with simple measurement and a calculator with good accuracy.

Contributions:

For the writing part of this summary:

HT: Summary: Data Cleaning, Model Strengths/Weaknesses, Reviewing. Slides: Background to Model Selection

ZJ: Summary: Statistical Inference. Slides: Inference

YZ: Purposed Model, Model Diagnostics. Slides: Model Diagnostics and Further Improvement

References:

Katch, Frank and McArdle, William (1977). _Nutrition, Weight Control, and Exercise_, Houghton Mifflin Co., Boston.

Wilmore, Jack (1976). _Athletic Training and Physical Fitness: Physiological Principles of the Conditioning Process_, Allyn and Bacon, Inc., Boston.