Two-Page Executive Summary

Introduction and Motivation:

As we have learned, accurate measurement of body fat percentage is inconvenient and costly. Thus, we would like to find out a way to easily estimate body fat percentage based on daily measurements. Our group built a linear regression model based on the dataset. Our model is intended to be simple, accurate and robust.

Background Information and Data Cleaning:

1. From the given data, the mean of body fat percentage is 18.9 with standard error 7.8. The mean of weight is 178.92 with standard error 29.39. We found that the range of weight is quite large, from 118.5 pounds to more than 360 pounds.
2. We converted the units of all circumferences from cm to inch.
3. We found suspicious data points when inspecting body fat percentage and height.
   a. We imputed a height of 69.5 inches for IDNO 42 whose original height is 29.5 inches based on the formula of calculating adiposity.
   b. We removed IDNO 172 and IDNO 182 with body fat 1.9 and 0, respectively. According to reports, the minimum possible body fat percentage for human beings is 2%. After trying to impute their body fat and getting a worse outcome, we think that these two data points should be removed.

Choosing Model and Final Model:

1. The final model is:

$$\widehat{body\,fat}(\%) = -62.51 + 2.86 \times \text{abdomen(inch)} - 0.014 \times \text{weight (pound)} - 0.0030 \times \text{abdomen} * \text{weight}$$

2. For example, a man with a 33.54-inch abdomen circumference, who weighs 154.25 pounds is expected to have 15.48 percent body fat.
3. **a.** The estimated intercept is -62.51. Our estimated coefficients are 2.86 for abdomen circumference,
   -0.014 for weight and -0.0030 for their interaction term.
   b. For a man who weighs 150 pounds, for every 1 inch increase in abdomen circumference, the model predicts that body fat percentage will increase, on average, by 2.40. For a man with abdomen circumference 34 inches, for every 1-pound increase in weight, body fat percentage will decrease by 0.12. This means for a man with given abdomen circumference, the heavier he is, the more muscular figure he has. Thus, he has lower body fat. Due to the interaction term, the exact slopes are not constant between individuals using this model, but the concept remains the same.
4. We chose this model because of the following reasons: (1) We use a linear regression model because it can give us accurate results and is a fairly simple model for users. (2) We chose body fat as the response variable over density because we had stronger correlations between explanatory variables and body fat than we did with density. (3) We found the linear correlation between abdomen circumference and body fat percentage to be the strongest. We then added other variables respectively to our model and compared their R^2. Model with weight has the biggest R^2. (See Table 1)

Table 1: Model comparison

model	Adjusted R^2
BODYFAT~Abdomen.Inch.+WEIGHT	0.7201
BODYFAT~Abdomen.Inch.+ADIPOSITY	0.6532

Statistical Analysis

a. The test for regression coefficients is t-test. The null hypothesis is: H0: $\hat{\beta}$=0. The test statistics is t=$\frac{\hat{\beta}}{\widehat{sd(\hat{\beta})}} \sim t_{n-2}$, where n is the number of samples. The coefficient of abdomen and interaction is significant at significance level α=0.05. Therefore, we do not reject the null hypothesis. This means that the abdomen and interaction have a significant explanation of variation on body fat.

b. We found our R^2 to be 0.7201. This means that our model is moderately explanatory and should be accurate at prediction.

c. The 95% CI for coefficient is: $\hat{\beta} \pm t_{n-2}(0.025) * \widehat{sd(\hat{\beta})}$. The 95% CI for the coefficient of abdomen is: (2.420562, 3.300974). The 95% CI for the coefficient of weight is: (-0.1041277, 0.07583168). The 95% CI for the coefficient of the interaction term is: (-0.00501392, -0.00108608)

Model Diagnostics

We checked the following three assumptions for MLR. First, we checked linearity using scatter plots and correlation coefficients (see Figure 1). Second, we checked independent variance (see Figure 2). The plot shows no pattern. We believe that there is no correlation between residuals and fitted values. Third, we checked normal distribution of variance. Because residuals generally follow normal distributions except some points., we believed assumption of normality is plausible, even though the distribution is actually a little left leaned than a normal distribution.

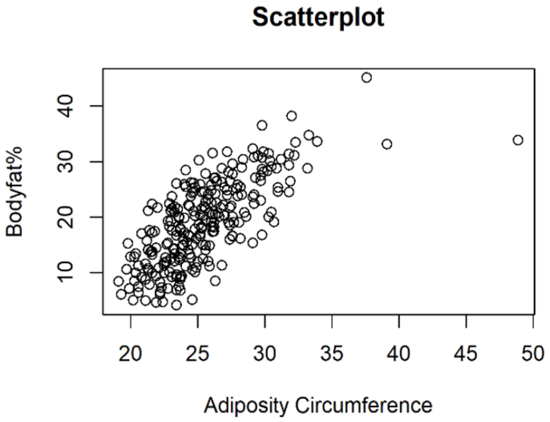

Figure 1 linearity

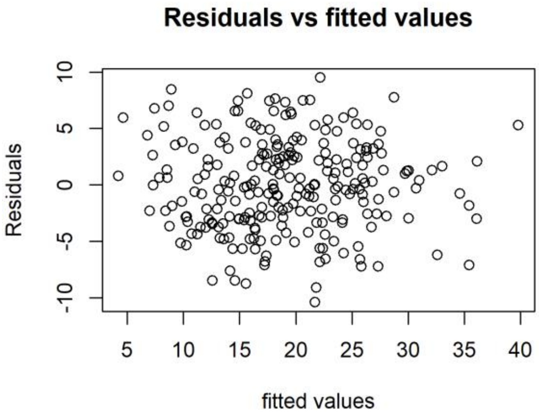

Figure 2 independence

Model Strengths and Weaknesses

a. Some strengths of our model include: **(1)** Our model is quite simple, with only two easy-to-measure variables. **(2)** Our model is quite explanatory. R^2 is over 0.7. In particular, our model satisfies the linear regression assumptions.

b. Some weaknesses of our model include our model does not treat special points specifically. That is to say, for extremely obese people or extremely skinny people, we may not give them an accurate prediction on their body fat.

Conclusion and Discussion:

In conclusion, we built a multiple linear regression model based on the body fat dataset. For each individual, we only need his abdomen circumference and weight to get his estimated body fat percentage, which is quite simple and feasible to use. Since we have a R^2 which is over 0.7, it should be accurate and robust.

Contributions:

We worked together as a whole group on data cleaning, variable screening, building a linear model and preparing for the presentation.

Xinyue Zhu wrote the majority of report. Shuguang Chen edited final model part of the report. Luke VandenHeuvel edited background information and data cleaning part of the report.

Shuguang Chen drafted the majority of the presentation slides. Luke VandenHeuvel drew the plot on the page of the slides.

Luke VandenHeuvel created the majority of the R Shiny App. Shuguang Chen and Xinyue Zhu gave suggestions to the R Shiny App.

Luke VandenHeuvel wrote the R code for correlation and scatter plots. Shuguang Chen wrote the R code for adding different explanatory variables and comparing the models.

Xinyue Zhu wrote the code for statistical analysis and model diagnosis.

Overall, we met five times for discussion and spent ten hours finishing the project.

Reference:

1. Fred Kiger. A night to remember. https://www.espn.com/sportscentury/features/00242495.html
2. William E. SIRI (1956). The gross composition of the body. Advances in Biological and Medical Physics, Volume 4, Pages 239-280