**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. The mean of body fat percentage is 18.93849. Its standard error is 7.750856. The mean of weight is 178.9244. Its standard error is 29.38916. Our group found that the range of weight is quite large.
2. Our group found some data points of body fat percentage and height suspicious. a. We imputed IDNO 42 with height 29.5 by 68.75 based on his adiposity.

b. We removed IDNO 172 and IDNO 182 with body fat 1.9 and 0, respectively. According to the Internet data, Jordan's body fat was 4%. We think that these two body fat data points are impractically low and should be removed.

**Choosing Model and Final Model:**

1. The final model is:

= -62.511270+ 2.860768\*abdomen circumference (inch) -0.014148\*weight (pound) -0.003050\* abdomen circumference \*weight+ ε, ε

1. For example, a man with abdomen circumference 33.54 inches and with weight 154.25 pounds is expected to have body fat percentage 15.479679.
2. a. Our estimated intercept is -62.511270. Our estimated coefficients are respectively 2.860768 for abdomen circumference, -0.014148 for weight and -0.003050 for their interaction.

b. For a man with weight 150 pounds, every 1-inch increase in abdomen circumference, the model predicts that body fat percentage will increase, on average, by 2.403268. For a man with abdomen circumference 34 inches, every 1-pound increase in weight, body fat percentage will decrease by 0.117848. This means for a man with given abdomen circumference, the heavier he is, the stronger figure he has. Thus, he has lower body fat.

4. We chose this model because of the following reasons. (1) We use a linear regression model because it can give us accurate and robust result. (2) We excluded density from potential explanatory variables because according to Siri (1956), the percentage of body fat can be only determined by density. Also, we cancelled neck and knee circumference because they are mostly due to naturally born. We deleted thigh and bicep circumference because they cannot be always correctly measured. (3) We found the linear correlation between abdomen circumference and body fat percentage to be the biggest. We then add other variables respectively to our model and compare their R^2. Model with weight has the biggest R^2.

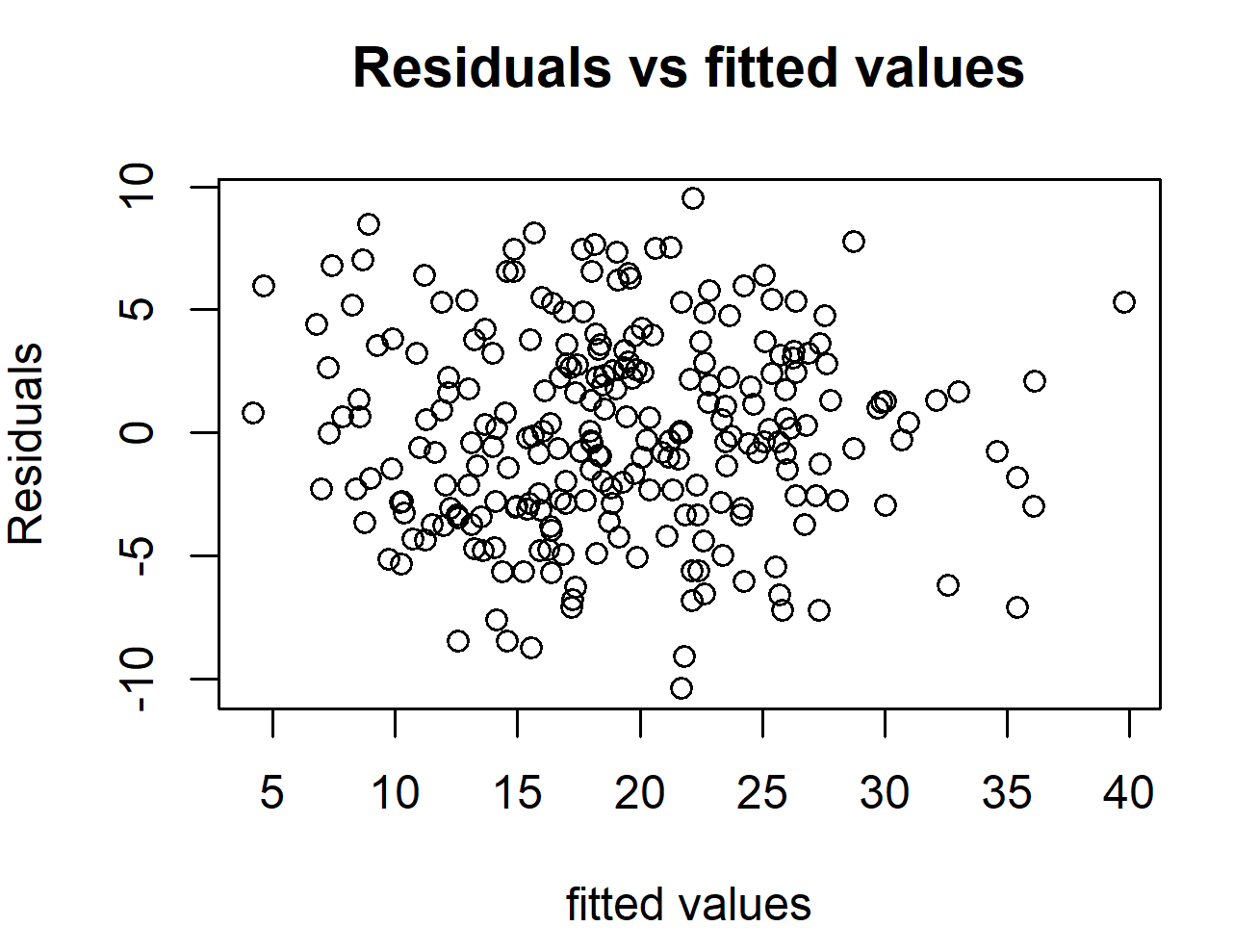
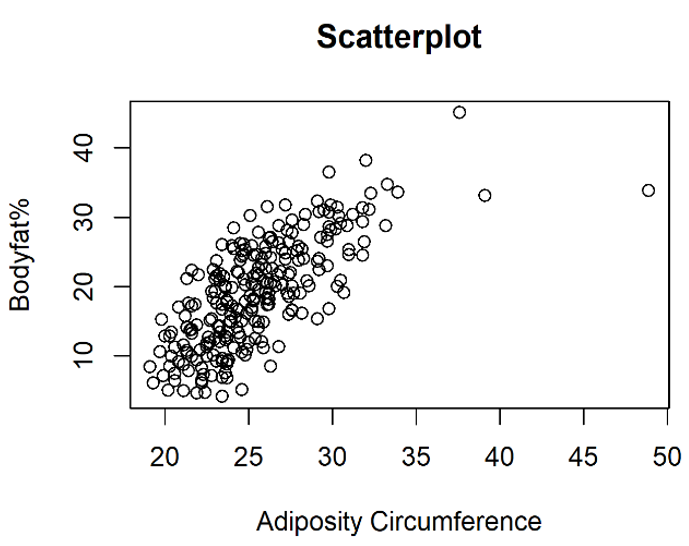
|  |  |
| --- | --- |
| model | Adjusted R^2 |
| BODYFAT~Abdomen.Inch.+WEIGHT | 0.7201 |
| BODYFAT~Abdomen.Inch.+ADIPOSITY | 0.6532 |

**Statistical Analysis**

1. The test for regression coefficients is t-test. The null hypothesis is: H0: =0. The test statistics is t=, where n is the number of samples. The coefficient of abdomen and interaction is significant under 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.
2. We found our R^2 to be 0.7201. This means that our model is moderately explanatory and should be accurate at prediction.
3. The 95% CI for coefficient is: 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 BLANK). Second, we checked independent variance. 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.



**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 obsessive 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**:

Xinyue Zhu wrote the report. Shuguang Chen made the PPT for presentation. Luke VandenHeuvel made the shiny app. Overall, we met five times, spent ten hour.

**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