Summary

Introduction

Body fat is always a hot topic for the public since it indicates people's health in some sense. The state-of-the-art research results reveal that some circumference and skinfold measurements may influence body fat. In this project, we exploit the factors that affect body fat and a formula to predict body fat, from a well-known dataset.

Data Cleaning

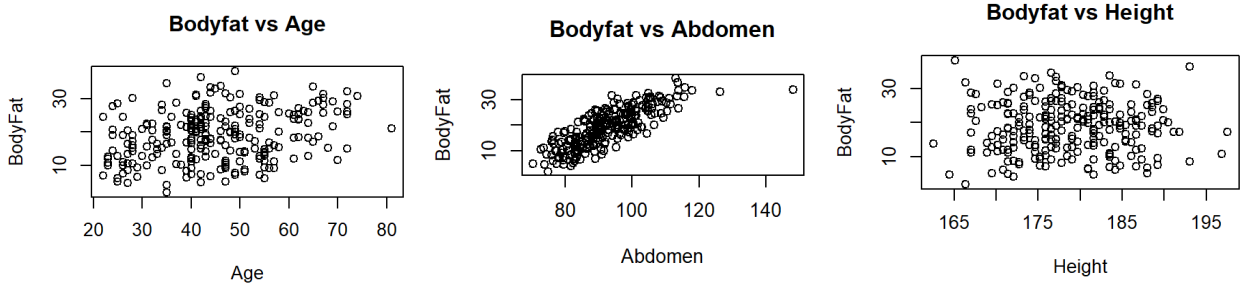
The original dataset contains 252 samples and 17 features. Some features are not useful to predict body fat, such as ID number and density, since there is a formula to convert density into body fat. We first made the measuring unit consistent, where length is in cm and weight is in kg.

After checking the data, we found some impossible values and outliers. ID 186 has a 0 body fat, which is impossible, and we cannot recover it from density since the resulting body fat is negative, ID 42 has a height of 74cm and ID 216 has very high body fat. Therefore, we remove those three samples. Finally, our cleaned dataset has 249 observations and 17 features.

Model Selection

The first and foremost task was to find the best features used in the model. It was important for us to look at factors that give us the best accuracy as well as satisfy the linear assumptions. The first step we took was to make a basic linear model with all the features and find out which features turned out to be important.

After the first step, Age, Neck, Abdomen, Forearm and Wrist came out to be important factors but these variables when you check for multicollinearity using a correlation heat map have a high correlation between them moreover there are 5 variables to be considered we were looking at a simpler model with a maximum of 3 variables. In the correlation plot, we can there are two variables that have a strong correlation with body fat i.e. Weight and Abdomen but those two themselves have a strong correlation of about 89% which is quite a huge figure. So the idea was to take either of them and go forward since in the initial model Abdomen came as important we decided to fix the abdomen and go forward with the abdomen and some other variables which are independent of the abdomen. The two other variables that we then chose were Age and Height based on the correlation plot. We first made individual graphs of the 3 variables with bodyfat to see if there is any weird pattern or any auto-correlation between them before we begin to make our model.



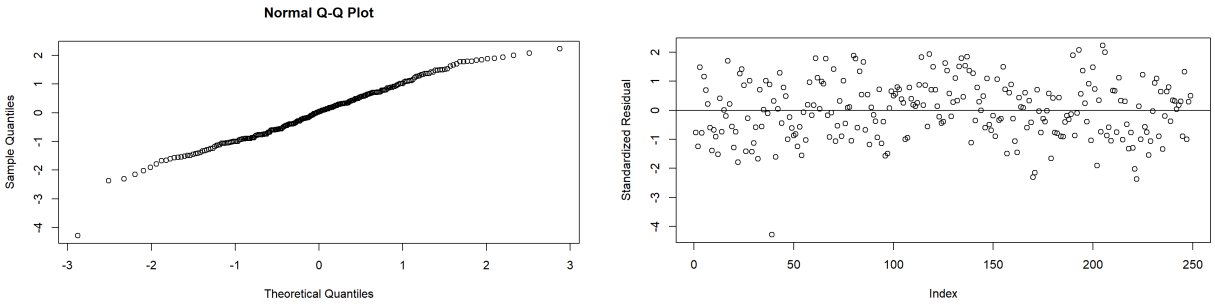
Then we proceeded with making linear models with Age, Height and Abdomen. We had 3 models in mind one just with Abdomen, another with age, height, and abdomen, and the third with height and abdomen. The idea then was to choose a model based on the highest **R square**. In the first model, Age comes as significant therefore we were left with Height and Abdomen. The r square in the abdomen comes as 0.64 and for the abdomen & height comes as 0.68 so we in the end choose the abdomen and height. Our final model is:

BodyFat% = 0.78 + 0.60 * Abdomen(cm) - 0.21*Height(cm)

The final model indicates that if a man's height increases by 10cm then he is expected to decrease about 2.1% in body fat and when a man's abdomen increases by 2cm he is expected to gain about 1.2% of body fat. The range of body fat% that we consider for different categories is based on the values given on the internet. The different categories are : Athletes (6-14%), Fit (14-18%), Average (18-25%), Obese (25%+).

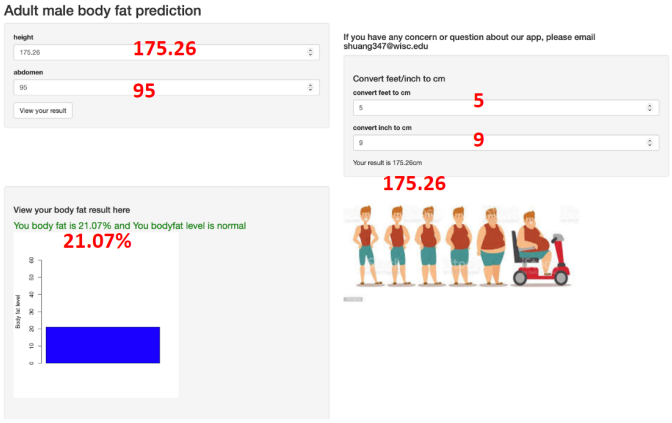
Model Diagnosis

It is crucial to diagnose the model. We check the independence, homoscedasticity, linearity, and normality of the residuals. From the line plot, there is no significant pattern for residuals, like autocorrelation or clustering, so they are independent. From the scatter plot, the points are above and below zero randomly, so they are homoscedasticity and linearity. From the qq-plot, we see they are almost in a straight line, so they follow the normal distribution.



Shiny

Our model takes two features, height and abdomen circumference, both in cm. Users could use the right upper panel to convert feet/inch to cm and then use it for the value for prediction. The result shows both digitally and graphically in the lower left corner, where the digital result is rounded to 3 decimal places, and the bar's color will vary based on the digital result. The lower right corner is a picture for users to match their body shape. If the user has any questions or suggestions, he/she can find the contact information on the top right.



Conclusion

By the trade-off between accuracy and simplicity, we build a relatively simple model with only two features, where height has a negative effect on body fat and the abdomen has a positive effect. Since all assumptions are not violated, the linear model is valid. In addition, the bivariate linear regression model is easy to interpret. However, the model is not as accurate as other complicated models.

References

- **Age in groups**

https://www.medicalnewstoday.com/articles/body-fat-percentage-chart#women

- **Body fat calculator**

https://www.calculator.net/body-fat-calculator.html

- **Predicting Body Fat Using Data on the BMI**

https://www.tandfonline.com/doi/full/10.1080/10691898.2005.11910560

Contributions

All three of the group members were equally involved in the project. We had group meetings every week for about 4 hours. The project took about 12 hours of group meetings plus individually contributed time. There was good coordination among the team, the ideas were first discussed and then implemented. We also helped each other in the individual tasks performed by suggesting ideas, solving errors, and complementing each other.

The following were the task performed by each person:

1. **Aashna Ahuja** - Feature Selection, Model fitting
2. **Shunyi Huang** - Shiny App
3. **Jingyun Jia** - Data Cleaning, Model Diagnosis