Project Summary (Group 15)

Introduction

Accurate measurement of body fat is inconvenient, and it is desirable to have easy methods of estimating body fat that is not inconvenient.This summary aims to show how we obtained a simple and convenient body fat prediction model from a natural body fat data set.

Data Cleaning

Initially, we analyzed the dependent variable **bodyfat**. Some sources show that the minimum body fat to stay healthy is 2% (for men). Therefore, we identified data for body fat below 2% for detailed analysis. We first considered predicting body fat in these two individuals by other relevant factors, such as density. However, we obtained body fat percentages of -3.61% and 0.70%, which remained impossible. We finally chose to delete these two rows of data. We then processed the independent variables similarly, filtered out the problematic variables, and deleted them according to the formula calculated for the variables. We ended up with a data set containing 244 data and 15 variables.

Model Selection

R-squared is a goodness-of-fit measure for linear regression models, so we chose it as the criteria of model selection.It has been indicated in literature that BMI and abdomen are significantly associated with body fat, so we first considered simple linear models associated with these two variables. The linear model based on BMI had an R-squared of 0.5143, while the linear model based on **abdomen** had a better R-squared of 0.6555.

Then, we performed a stepwise regression on this model to add or remove potential explanatory variables and obtained the following four models.

Model	R-squared
Model 1 : BODYFAT ~ ABDOMEN	0.6555
Model 2 : BODYFAT ~ ABDOMEN+WEIGHT	0.7186
Model 3 : BODYFAT ~ ABDOMEN+WEIGHT+WRIST	0.7253
Model 4 : BODYFAT ~ ABDOMEN+WEIGHT+WRIST+FOREARM	0.7339

We can find that Model 2 significantly improves compared to Model 1 (about 0.05), while Model 3 and Model 4 do not significantly improve compared to Model 2 (about 0.01). Considering the model complexity and the goodness of fit, **Model 2** is the best.

Therefore, our **final model** is

$$Bodyfat(\%) = -40.27 + 0.9143 \times Abdomen(cm) - 0.1415 \times Weight(lb).$$

This means a man (average American) whose abdomen is 40.5 inches and weighs 199.8 lb is expected to have a body fat of 25.5% based on our model. His 95% prediction interval is between 17.5% and 33.5%.

In the above equation, the units of abdomen and weight are centimeters and pounds. For every 1 cm increase in the abdomen, the model predicts a relative increase in body fat of 0.9143%. Similarly, for every 1 lb increase in weight, the model predicts a 0.1415% decrease in body fat.

The rule of thumb is “multiply your abdominal circumference by 0.9, minus 40, then minus your weight times 0.15”. A caveat of the rule of thumb is that for it tends to underestimate your body fat % compared to the non-rounded estimated model. However, this rule of thumb is within the 95% confidence intervals of the coefficients. Under this rule, the predict value of an average American man is 22.5%, which is still in the 95% prediction interval.

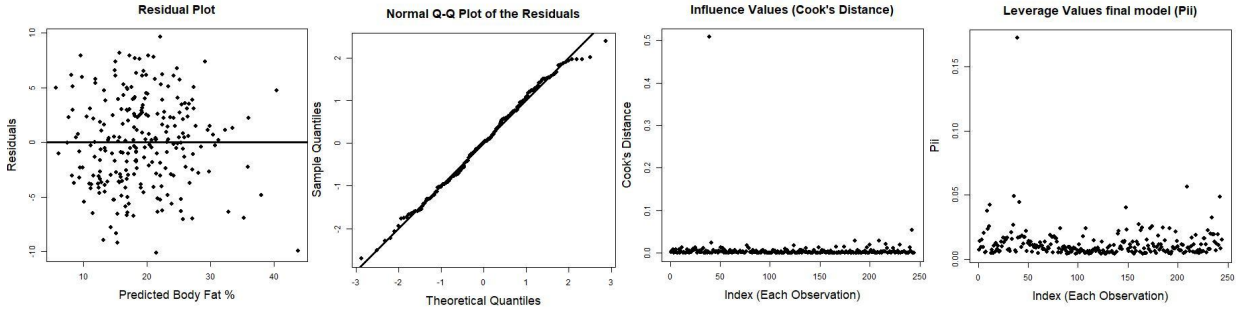
Statistical Analysis

In the model comparison above, we chose to use the **R-squared** comparison because the R-squared reflects the degree of fit of the model. Simply put, the R-square represents the proportion of variance in the dependent variable that the independent variable can explain. The larger the R-squared, the better the regression model fits the observations. Thus, we can say that Model 2 has a significant improvement compared to Model 1 because the R-squared of Model 2 is larger than that of Model 1.

Model Diagnostics

First, the residual plot is used for checking the assumption of **linearity** and **homoscedasticity**. It seems that points are randomly scattered around zero for the entire range of predicted values, and there is no clear linear trend in the residual plot. Thus, linearity and homoscedasticity are plausible.

Then we check the **normality of residuals** by **Q-Q Plot**. Residual points follow the straight dashed line, as seen in the figure. Therefore, we believe that the residuals follow a normal distribution, although a few points at both ends do not follow the straight dashed line.



The next two plots (Leverage and Influence values) show the leverage values at each point and the degree of influence on the model. The graphs show that there are still some high leverage and strong influence points. However, the data corresponding to these points are extreme, but within reasonable limits, so we keep them.

Model Strengths and Weakness

Some strengths of our model include being simple and convenient. In particular, our model satisfies the linear regression assumptions of linearity, homoscedasticity and normality of residuals, bringing credence to our model interpretation and application.

However, our model has some drawbacks. The first one is that the prediction of the model is not accurate enough. In order to keep the model simple, we sacrificed some goodness of fit. Thus, there are only 20% of predictions within +/- 5% of true value and 50% of predictions within +/- 15% of true value. Also, the model can only accept data in specific units (cm and lb), and if we input data in different units, such as inch and kg, we will get wrong predictions.

Discussions

Through continuous experimentation, we finally obtained a simple and convenient model for predicting body fat, but the model still has some shortcomings. In the future, we can consider improving the model to address these issues, such as increasing the data size and trying to find the input data with the wrong unit and convert the units.

Contributions

1. Chufan Zhou edited the data cleaning and model selection parts of the summary, worked on slides page 1 to 6. He also created code related to data cleaning and data construction.
2. Zhanpeng Xu wrote the model interpretation, tables, and usage of model parts of the summary, worked on slides 7 to 11. He also created code related to model selection and graphs.
3. Xuelan Qian wrote the model diagnostics and strength & weakness parts of the summary, worked on slides page 11 to 14. She also created code related to model diagnostics and graphs.
4. All group members contribute to the shiny code.

Reference

1. https://www.performancelab.com/blogs/fat-loss/what-is-the-lowest-body-fat-percentage#:~:text=Going%20much%20lower%20than%20this,ever%20go%20below%20these%20percentages.
2. https://www.cdc.gov/nchs/fastats/body-measurements.htm
3. https://fitnessofmens.blogspot.com/2016/05/usain-bolt-profile-weight-height-body-stats.html?m=0.
4. https://us.humankinetics.com/blogs/excerpt/normal-ranges-of-body-weight-and-body-fat.
5. Jensen NSO, Camargo TFB, Bergamaschi DP. Corrigendum to 'Comparison of methods to measure body fat in 7-to-10-year-old children: a systematic review' [Public Health 133 (April 2016) 3-13]. Public Health. 2018 Aug;161:49. doi: 10.1016/j.puhe.2018.05.013. Epub 2018 Jun 9. Erratum for: Public Health. 2016 Apr;133:3-13. PMID: 29894868.