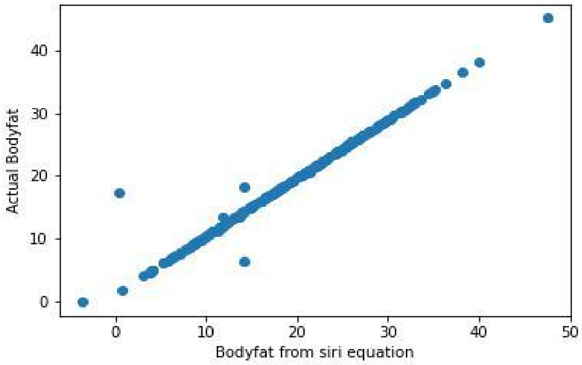
Group 17 - Module 2 : BodyFat Prediction Modeling

Introduction

This is the summary document for body fat prediction modeling. Our group tried to build a model as simple with a few number of features, which gave us a linear regression model with two features left, Abdomen and Weight.

EDA and Feature Selection

BODYFAT is highly correlated with the feature DENSITY (-0.988) as either one can be reciprocally calculated by Siri Equation. (Siri, 1956). <BODYFAT = 495 / DENSITY - 450>. Given the equation and IQR, five outliers can be filtered, 48th, 76th, 96th, 182th, and 216th rows of the data file. In that sense, once a model is finalized, the target values for these five data points might be adjusted.



Our group selected important features by filtering out highly-correlated features, features having zero coefficients from Lasso, and features with lower feature importance levels from the attribute 'feature importances_' of random forest model. As a result, nine features were dropped off of the features set on top of removing the column DENSITY, which gave us the five candidates features, AGE, WEIGHT, HEIGHT, ABDOMEN, and WRIST.

Modeling

Since there are only 5 features left, the only factors we will use in the final model will be age, weight, height, abdomen and wrist.

We use the best subset regression to select the best model. Since it's not easy to compare models with different numbers of independent variables, we firstly choose the best models among the models with the same number of independent variables, then compare them to decide which is the best.

To compare the models with the same number of independent variables, we believe that the best model is the model with the highest r-square, and when models' r-squares are close (differences less than 2), we prefer the one with the lowest Mallow's cp.

Model Selection

Here are the best models.

1. Bodyfat ~ Abdomen	r-square = 0.63
2. Bodyfat ~ Weight + Abdomen	r-square = 0.69
3. Bodyfat ~ Weight + Abdomen + Wrist	r-square = 0.70
4. Bodyfat ~ Weight + Height + Abdomen + Wrist	r-square = 0.70

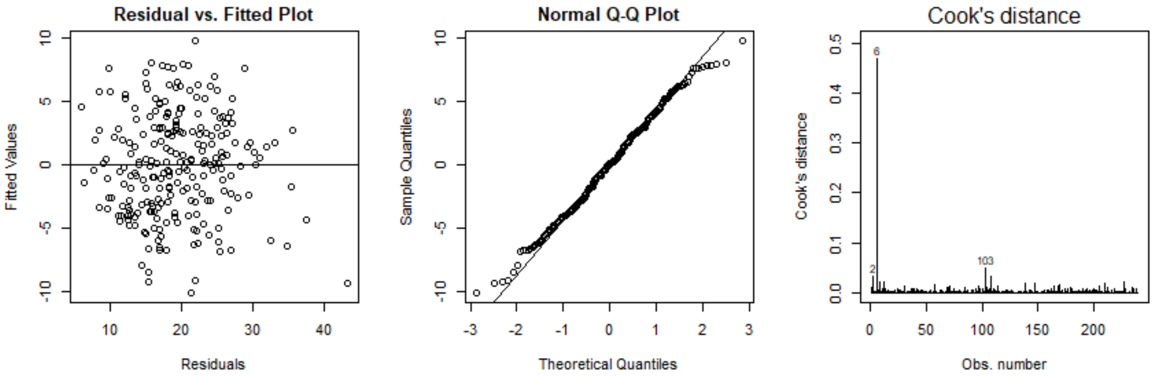
We can find that in these models, the latter model is always the former model with an extra variable, which implies the five factors have a decrease effect on the model. According to the r-square, we can conclude that there’s no significant difference among model 2, 3 and 4 in validity, when model 1 is obviously unconvincing, so in order to keep our model as simple as possible, we prefer the model with 2 factors.

In summary, we believe that model Bodyfat ~ Weight + Abdomen is the best model. The predicted bodyfat can be calculated by the function below.

Bodyfat(%) = -38.57 + 0.88*Abdomen circumference (cm) – 0.14*Weight (lbs)

Model Diagnostics

First the residual plot is used to check linearity and homoscedasticity assumptions. From the leftmost plot below, the points are relatively evenly distributed and there is no significant trend. So, the assumption of homoscedasticity and linearity for a linear regression are met. Next, a QQ-Plot is used to check the normality assumption. According to the plot in the middle, the normality assumption is not violated. Finally, the check for influential points is performed by examining the cook’s distance plot (right) and there are no violations



The model presented in this project has several strengths. Compared to the famous US Navy model (Hodgdon, 1984), which has 5 factors, the model presented is much more simple with only two predictors. This means the user can get a result with only two easily accessible measurements. Another strength is that it’s relatively accurate. With a r-squared value of 0.69, which means 69% of the variance in the dependent variable that can be explained by the independent variable. Possible weakness of the model is that abdomen circumference is not the easiest to measure and weight fluctuates during the day.

Conclusion

The project created a model using over 200 data samples that predicts body fat percentage using height and abdominal circumference. Feature selection is performed with Correlation, Lasso, and Random Forest. A linear regression is fitted to finish the model. The model does not violate the assumptions, is easy to use and yields accurate results.

References

1. Hodgdon, James A., and Marcie B. Beckett. “Prediction of Percent Body Fat for U.S. Navy Women from Body Circumferences and Height.” 1984, https://doi.org/10.21236/ada146456.
2. Siri, W.E. (1956) *Body Composition from Fluid Spaces and Density: Analysis of Methods*, UCRL-3349 University of California Radiation Laboratory, Berkeley (California).

Contributions

Sungrim Lee: worked on EDA with Python, Summary, Presentation, Git Repo, accordingly.
Xiao Li: worked on Modeling with R, Summary, Presentation, Git Repo, accordingly.
Max Zou: worked on Shiny App and Model Diagnostics for Summary, Presentation, and Git Repo, accordingly.