Summary

Abigail Terzis, Xingze Wang, Jianwei Ren

1. Introduction

Body fat is essential for humans to live as it acts as protection for our bones and organs, much like skin does. However, similar to everything, too much or too little body fat can be dangerous to a person's health. When calculating this percentage is not as easy in practice so we must come up with a simple, concise way to make this prediction to stay healthy. We have cleaned data, built a model, and developed an easy way for men to calculate their body fat percentage. The data was collected from a group of 252 men with 17 different variables that were measured.

2. Data Cleaning

We first plotted a histogram of each variable to view potential outliers, in particular, outliers of values that can be recalculated using other variables. For instance, the gentleman corresponding to ID number 172, 182 and 216 showed abnormal values for BODYFAT, so we used the density variable to recalculate it using the Siri's equation:

$$(495/DENSITY) \; - \; 450$$

However, even after doing so, the values for their body fat percentages were still abnormal, because of this we decided to remove those observations. Additionally, we observed that the gentleman with ID number 39 had measurements outside of the normal range. We removed this observation since there is no way we can try to recalculate the circumferences.

3. Model

BodyFat= -22.74 - 0.086*Weight + 0.876*Abdomen - 1.313*Wrist

A man with weight of 187.75 pounds, Abdomen of 96.4cm and wrist of 18.2 is expected to have a body fat percentage of 21.73 based on our model. His 95% prediction interval is between 13.872 and 29.495.

Our estimated coefficients are -0.084, 0.876 and -1.313, which are in the units of pound, cm and cm, respectively. As a man's weight increases by 1 pound with other variables held constant, he is expected to gain -0.086% in body fat. As men's abdomen increases by 1 cm with other variables held constant, he is expected to gain 0.876% in body fat. As men's increases by 1 cm with other variables held constant, he is expected to gain -1.313% in body fat.

From *Figure 1,* we have four plots with four metrics: Rsq, AdjR2, BIC and Cp. As we can see from the BIC plot, we have lowest values with the 3-variable model which is the best. But will still want our model as simple as possible but evaluate the good performance. Both 2-variable model and 3-variable model give us good correlations but with a relatively big difference on Mallow's Cp. Therefore, based on the discussion above, 3-variable model with Weight, Abdomen and Wrist would be our best model.

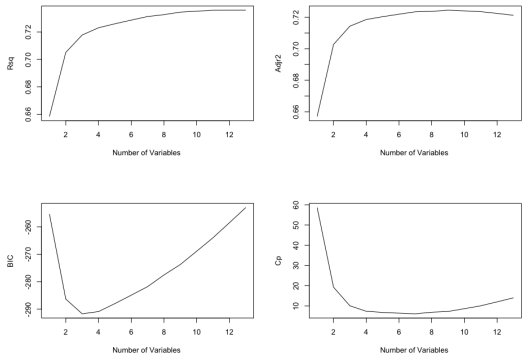
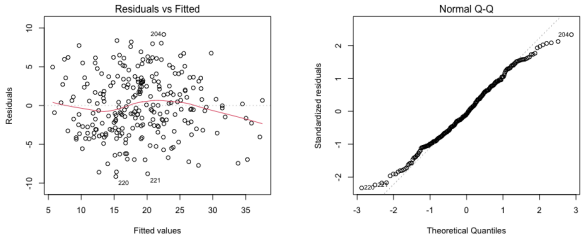


Figure 1. Metrics by Number of variables

4. Model Diagnostics

After model fitting, we diagnose the MLR assumptions with a residual plot and a QQ plot. The residual plot can diagnose linearity violations and homoscedasticity violations. We can see that there are no obvious non-linear trends in the residual plot and the points are evenly distributed, so linearity and homoscedasticity are satisfied. The QQ plot diagnoses the violation of normality. From the plot below we can see that the points hug the 45 degree line closely except a few points, so the normality assumption is also satisfied.



5. Model Strengths/Weaknesses

Some strengths of our model include that we can select the smallest number of predictors with guaranteed prediction accuracy, as well as we can compare using various methods such as Adjusted R^2, Cp and BIC. The weakness of our model is that the choice of the number of variables varies from person to person. For instance, if one prefer the Adjusted R^2 method, he/she might choose a larger number of predictors because the R^2 increases with the number of variables.

6. Conclusion

Through our processes of selecting a model we were able to find an equation to predict a man's body fat percentage and produce a clean web-based application for any male to easily use. While body fat percentage is not the only measure of health, it is a very useful tool to evaluate lifestyle choices.

Contributions:

AT: Cleaned the data and recalculated the points that appeared to be outliers. Created Shiny web-based app, wrote the introduction, data cleaning, and conclusion of executive summary. Readme section of GitHub

JR: Created the Github repo and uploaded all files. Found the best subset for our model. Wrote the Motivation/Choosing Model section of executive summary

XW: Created Powerpoint slides, wrote the model diagnostics and model strengths and weaknesses of executive summary. Also ran a model in R verify results

Together: fit model and reviewed the significance of variables, edited the executive summary, reviewed powerpoint slides. All group members communicated well together and showed up to every scheduled meeting prepared.

References:

https://www.topendsports.com/testing/siri-equation.htm#:~:text=Many%20body%20composition%20equations%20derive,%2F%20Body%20Density)%20%2D%20450.