Body Fat Prediction With Common Features

Introduction

As an important standard for the growing public awareness for a healthy lifestyle, body fat percentage has become a common estimate for measuring obesity. The traditional way of measuring body fat percentage for an individual usually consists of measuring the body volume and body density, which might be inconvenient to measure.[1] In this paper, we propose a model that accurately predicts the body fat percentage of an individual using only features commonly available.

Dataset

In this paper we used a dataset consisting of 252 men with measurements of their percentage of body fat and various body circumference measurements. Some commonly available measurements include age, weight, height, BMI, etc.. Since all data points are from male , gender is not a factor in this study. During data cleaning, one data point was removed due to having 0 percentage of body fat, which seemed extremely unlikely. Another data point where the weight exceeded 363 pounds was removed as an outlier. And one data point with less than 1 body density was removed. After examining the distributions of the independent variables, we noticed that the majority of them seemed normally distributed and thus did not require transformation. The cleaned dataset can be found in the data repo.

Model Selection

Our final model and rule of thumb for estimating body fat percentage is stated the following.

$$BODYFAT = -33.4 + 0.114 * AGE + 0.985 * BMI + 0.217 * CHEST$$

where BODYFAT denotes body fat percentage, AGE denotes age, BMI denotes body mass index and CHEST denotes chest measurements.

Using this model, a man with 24 age, 26.5 BMI and 104.5 chest is expected to have a body fat % of 18.12. A 95% prediction interval for his body fat percentage is [8.21, 28.03]. The coefficients for age, BMI and CHEST are 0.114, 0.985 and 0.217. This indicates that for every 1 unit increment in age, holding other variables constant, the body fat percentage of a man is expected to increase by 0.114. Similar conclusions can be reached from other factors in the model.

In model and feature selection, we combined our background research with the statistical analysis of the variables. First, from our background research, we noticed that variables such as density are not commonly available variables. Second, from EDA results, especially the correlation plots, we avoided including highly correlated features such as BMI and weight. Third, we conducted the variable importance analysis. Specifically, we trained LASSO, Random Forest and stepwise regression using cross validation to find the best fit. In LASSO, we noticed that DENSITY, AGE, CHEST and ABDOMEN were selected. In random forest, we noticed that DENSITY, ABDOMEN, BMI and CHEST have the highest variable importance by increase in Gini's node impurity.[2] Thus, AGE, CHEST, BMI and ABDOMEN are now our candidates. ABDOMEN was removed due to high VIF score against the rest of the candidates. Last but not least, several other models were tested as comparisons to our final model.

Hypothesis Testing

We conducted a t-test to see whether the predictors we chose are significant in predicting, by testing if the coefficients of our predictors are significantly different from 0. The hypothesis are as followed:

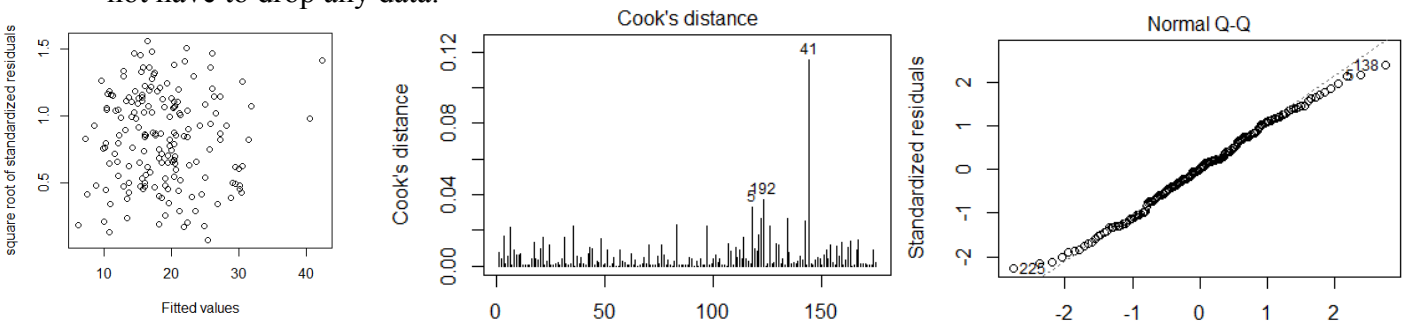
$$BODYFAT=\beta_0+\beta_{age}*Age+\beta_{BMI}*BMI+\beta_{chest}CHEST$$

1. $H_0: \beta_{age}=0$ $H_1: \beta_{age}\neq 0$ 2. $H_0: \beta_{BMI}=0$ $H_1: \beta_{BMI}\neq 0$ 3. $H_0: \beta_{chest}=0$ $H_1: \beta_{chest}\neq 0$

We calculated the t-test statistic and corresponding p-value, which is the probability of obtaining a more extreme value compared to the t-test statistic we calculated based on t-distribution. The tolerance level of falsely rejecting null hypothesis is 10%. According to the test result, all the three predictors passed the hypothesis test. Which means that all three predictors are significant in our linear model.

Model Diagnostics

We checked the following three assumptions for the linear regression model.
The first assumption is the homoscedasticity of residuals. If not, the OLS is not BLUE anymore. In order to check that, we apply the residual plot which shows that the square root of the residuals are in the same range as the estimated values get larger. So the homoscedasticity assumption of our model holds. After that, we checked whether the residuals follow normal distribution. Because the QQ plot is well fitted, we believed that the residuals following normal distribution are plausible, even though there are slight violations of when the fitted values are extremely small and large. Finally, we calculated the cook's distances which is an estimate of the influence of a data point to the estimated coefficients. All the cook's distances of the data points are in the reasonable range and we do not have to drop any data.



Model Strengths/Weakness

For the strengths of our models, we applied servel tests and used some plots to check the assumptions of the OLS model like the residuals are homoscedasticity and there does not exist autocorrelation among residuals. And it seemed that all the assumptions are meeted, which denoted that our model is reasonable. Apart from that, our model is relatively simple and its parameters can be easily interpreted.
For the weaknesses, the R squared and adjusted R squared of our model are relatively small which might imply there are still some variables which have significant influence on the body fat that are not found.

Conclusion

In conclusion, we established a model that accurately predicts the body fat percentage. The model has high interpretability and high predictive power. For future steps, we believe that increasing the complexity of the model might further increase the model's ability to generalize.

Reference

[1] Bailey, Covert (1994). _Smart Exercise: Burning Fat, Getting Fit_, Houghton-Mifflin Co., Boston, pp. 179-186.
[2] Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001).

Contribution:
Shuren He wrote the model diagnostics and model strengths/weakness parts of the report, edited the slides 8-15 and was responsible for producing all visualization code and model diagnostics code.
Xiaoyu Liu wrote the hypothesis testing and conclusion parts of the report, page 2-4 of the slides, built the shiny app and was responsible for producing the stepwise regression and VIF analysis.
Xintong Shi wrote the introduction, dataset and model selection parts of the report, edited the slides 5-7, and was responsible for producing the model selection and final model building code(LASSO, RF, final model).