# Abstract

We conducted initial visualization of the body fat dataset with plots and tables, followed by model selection, model interaction, and cross-validation. Model diagnostics was also applied to check whether the model needs any transformation before omitting the outliers. We then select an optimal model that can best predict the body fat.

# Introduction

Body fat is an important factor in many medical situations. Measuring body fat can help to assess whether a person is at a healthy weight. Additionally, obese or overweight can be identified by calculating the body fat. A normal body fat is between 25 and 30 percent in women and 18 and 23 percent in men. More than 68% of US adults are considered overweight, and 35% are obese. While the number of people who are overweight or obese are gradually increasing, more people are becoming more concerned about this issue. The goal of this project is to predict the body fat in different patients based on their physical measurements.

# Methods

## \_Data Exploratory Analysis\_

We imported the original raw dataset-body density data. Among the three target variable: bodyfat\_brozek, bodyfat\_siri, and density, we have chosen bodyfat\_siri as our final target variable. Other two were omitted from the dataset.

The goal of this part is to have a sense of the distribution of each variables.Correlation was also considered to see whether any variables are highly correlated with eachother so that we can try to eliminate the multicollinearity effect. In order to visualize the above information, a descriptive statistics table was made, a boxplot for the target variable, histogram for the predictors, and a correlation plot. Multicollinearity will also be checked by calculating the VIF.

## \_Model Selection\_

In order to select the most appropriate model, we used automatic procedures (Stepwise Regression) as followed for comparison between bigger and smaller models.

## \_Model Interaction\_

To see if there are interaction effects between any two independent variables, separate regression models were graphed. If the graph showed parallel slopes, then the interaction was not significant. This interaction term should be removed. Otherwise, it should be kept. The main effects of these two variables should be taken into consideration together in the model.

## \_Model Diagnostics\_

After obtaining the model that we have chosen, several plots will be made to check whether the model meets the regression assumptions. The QQ plot will be used to determine the normality of the data. Outliers in the data will also be identified by plotting the residuals vs leverage plot and also calculating the Cook's distance, which can show the influence of a specific case on all fitted values. Finally, a model without outliers will be shown and will be compared to the original model to examine the influence of the outliers. The influential outliers will be omitted in the final model.

## \_Model Validation\_

A 5-fold cross-validation will be constructed to test the validity of the model. RMSE(root mean squared deviation), R square, and adjusted R square will be used to evaluate the model.

# Results

## \_Data Exploratory Analysis\_

The conclusion was confirmed by the results of histograms, showing that the reciprocal transformation of the five variables "neck", "abdomen", "knee", "ankle", "bicep" mitigated their original skewed, thus exhibiting a better normal distribution. Therefore, we decided to use the transformed value of "i\_neck", "i\_abdomen", "i\_knee", "i\_ankle", "i\_bicep" to participate in the subsequent calculation.

## \_Model Selection\_

The model we obtained by Stepwise Regression is bodyfat\_siri = 1.257e+02 + 5.425e-02 \* age - 4.106e-01 \* height + 6.568e+02 \* i\_neck - 6.876e+03 \* i\_abdomen + 3.315e-01 \* forearm - 1.741e+00 \* wrist.

## \_Model Interaction\_

The graphs showed that there were intersection points between forearm and height, neck and height, and abdomen and knee. The graph of all the remaining paired variables showed two parallel lines. Therefore, the revised model should be bodyfat\_siri ~ age + height + neck + abdomen + forearm + wrist + forearm\*height + height\*neck + abdomen\*knee.

## \_Model Diagnostics\_

Next, we operated diagnostics by the following diagnostic plots for checking the adequacy of the regression model.

As shown in the Residuals vs Fitted plot, residual values bounce around 0 and residuals form a generally horizontal 'band' around zero, showing no unequal error variance. There are three observations #192, #224, #225 standing out from the random pattern, making them potential outliers.

According to the QQ plot, which is almost straight, exhibiting a nice normality without heteroscedasticity. #192, #224, #225 still stand out again from the random pattern in the QQ plot.

The Scale-location plot is almost a horizontal line with equally spread points, confirming the assumption of equal variance.

In the Residuals vs Leverage plot, observations #153, #175, #216 appears to be outlying values with high statistics in Cook's distance D. Next, we compared the model estimates with and without the three observations. It came out that though excluding three values slightly increased R^square by 0.002, but all variables in the model turned out to be statistically significant. Therefore, we decide to exclude outliers #153, #175, #216 from our analysis.

Overall, these plots look like this model is fitting the data well.

## \_Model Validation\_

bodyfat\_siri = 84.7635 + 1.0080 \* age -0.4150 \* height + 2212.3159 \* i\_neck -6835.4225 \* i\_abdomen + 0.3344 \* forearm -1.7566 \* wrist -36.2415 \* (age\*i\_neck)

# Discussion