Introduction/Motivation:

The body fat percentage is a measurement of fitness level. Comparing to another widely used index BMI, body fat percentage is more useful for determining the health of an individual.

However, measuring body fat percentage sometime is very difficult. Using the body density method to measure is a very accurate method, but it is very difficult to achieve in practice. Therefore, we want to find out a method to accurately predict body fat percentage using easy-to-measure indicators.

Background Information/Data Cleaning:

Our data include 252 data points with 16 dimensions of variable. The dependent variable is body fat rate and independent variables are some of anthropometric data, such as height, weight, abdomen, neck and ankle. Height is measured in inch and other anthropometric data is measured in centimeter. Here is a brief summary of the data:

	BODYFAT	AGE	WEIGHT	HEIGHT
Min	4.1	22	125	29.5
Mean	19.08	44.9	179	70.1
Max	45.10	81	363	77.7

We will ignore variable 'density' because it is difficult to obtain in real situation. Human body fat rate can never be less than 3%, otherwise this person will die. We have found and removed two outliers for No.172 with body fat 0% and No.182 with body fat 1.9%.

Choosing Model/Final Model/Rule of Thumb

Our rule of thumb is to find a simple but precise model. We want to get a linear model with three or less independent variables. The performance of the model is judged by R-square and RMSE.

We have four candidate model with three of them are existing methods which are YMCA, BMI approach and U.S. Navy method. The other one is a 3-variable linear model obtained by stepwise regression. We will call it AWN model based of the first character of 3 variables. We using backward stepwise regression according to AIC and marginal ANOVA.

R-square is the proportion of deviation can be explained by the model which is the larger the better and RMSE represents the average difference between predicted values and true values which is the less the better.

	YMCA	BMI	Navy	AWN
R-square	56.1%	55.3%	71.2%	71.9%
RMSE	5.031	5.078	4.073	4.029

The interpretation of R-square is 71.9% deviation of bodyfat can be explain by the model. And RMSE is usually used to construct a confidence interval which means the true value has 0.68 probability falling within Predicted Value$\pm$RMSE and 0.9 within Predicted Value$\pm$2*RMSE. We choose it as our final model because it has the largest R-square and the least RMSE.

The Navy model has similar performance as AWN, but it contains log transformation which is hard to calculate and interpret. Based on simplicity principal, AWN is better.

Statistical Inference/ Hypothesis Testing

The Final model is AWN model:

$$BODYFAT = -30.26 + 0.92 * ABDOMEN - 0.11 * WEIGHT - 0.4 * NECK$$

$$coeff\ P\ value: \quad 8 * 10^{-7} \quad 0.03 \quad 2 * 10^{-16}$$

$$F\ statistics: 209.5 \quad P\ value: 2 * 10^{-16}$$

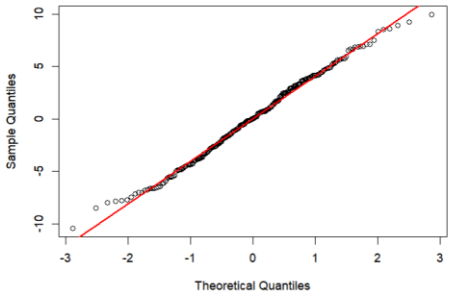
We can see all coefficients in the model are significant on 0.05 level. And F statistics also tells the model is overall significant.

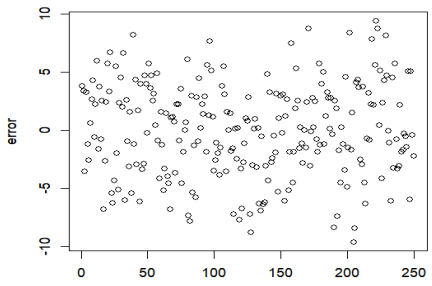
The coefficients mean that while controlling all other indicators unchanged, for every 1cm increase in abdomen, the model predicts that body fat will increase 0.92%; 1lb increase in with body fat decrease 0.11% and 1cm increase in neck circumference with body fat decrease 0.4%.

For example, an individual with 180lb weight, 91.6cm abdomen and 38.4cm neck is expected to have a body fat 17.84%. And the 90% confidence interval is between 12.72 and 22.96.

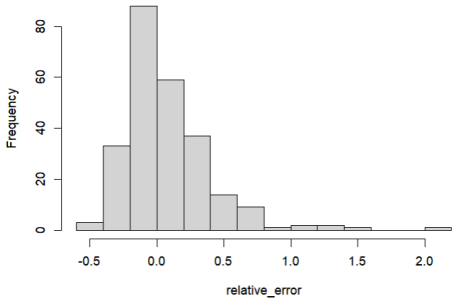
Model Diagnostics

As we diagnose the model with residual. The QQ-plot shows the residuals comply normal distribution and the scatter plot shows it is homoscedasticity. Thus, the model seems do not violate any assumption of linear model.





In order to reflect the accuracy of the model more intuitively, we calculate the relative error instead of the absolute error, which is: $Relative\ error = \frac{Absolute\ error}{True\ value}$



By plotting the Histogram of relative error, we can see most of the relative errors are within 0.5 and more than 60% are within 0.25.

Model Strengths/Weaknesses

The strength of the model is that it is very simple. With the best performance, it is just a simple linear model contains 3 variables. In addition, it is easy to explain, every coefficient has a specific meaning.

But also, the weakness is it is a little bit too simple. It might be not robust enough when facing more data. We have to consider some nonlinear relationship in big data.

Reference

[1] A simple technique for measurement of percent body fat in man. Wright HF, Dotson CO, Davis PO; U. S. Navy Medicine, 01 May 1981, 72(5):23-27

[2] YMCA Body Fat Formula. https://www.easycalculation.com/formulas/body-fat-ymca-formula.html

[3] Body fat percentage wikipedia. https://en.wikipedia.org/wiki/Body_fat_percentage

[4] Relationship between Body mass index (BMI) and body fat percentage, estimated by bioelectrical impedance. Chathuranga Ranasinghe, BMC Public Health 797 (2013)

[5] Body Mass Index. https://www.nhlbi.nih.gov/health/educational/lose_wt/bmitools.htm