In recent years, accurately predicting body fat has become crucial for health assessments, as it provides an important indicator for health issues such as obesity. Having a simple yet accurate method to assess body fat is both convenient and significant for understanding personal health. Therefore, in this report, we aim to develop a statistical approach to effectively measure body fat percentage.

Our dataset contains measurements from 252 individuals, including a range of body dimensions and other key variables used for body fat prediction. The variables can be categorized as follows:

- **General Information**: Includes the individual's ID, age, weight, height, and adiposity (BMI).
- **Body Fat and Density**: Percent body fat (calculated using Siri's 1956 equation) and body density measured through underwater weighing.
- **Body Circumference Measurements**: Neck, chest, abdomen, hip, thigh, knee, ankle, biceps (extended), forearm, and wrist circumferences (all in centimeters).

These variables will be utilized, either fully or selectively, to build a statistical model for predicting body fat percentage.

Before using these data to fit a model, we first process and clean our data. We use IQR(Interquartile Range) to find outliers and remove them. According to the outcome, we remove individuals No.39, No.41 and No.216 since they have multiple outliers. Also, noticing that height of No.42 is significantly below average, we use adiposity together with weight to adjust it. Moreover, to mitigate the influence of different units, we scale the data.

We choose linear regression for its simplicity and high interpretability. With proper selection of predictors, we may also achieve relatively high accuracy. In the task we assume that factors are independent from each other. In model training, we use 10-fold cross validation to guarantee robustness of model and full utilization of data. We will choose different predictor combination from various aspects.

Firstly, We want to find the combination of predictors with highest AIC. This is achieved by running code for stepwise regression in R. Then, by enumerating all models containing only one variable and two variables, i.e.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

and

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i,$$

Here and all following $\epsilon_i \overset{\text{iid}}{\sim} \mathbf{N}(0, 1)$.

Therefore, we have candidates for best predictors. Meanwhile, considering that different variables may interact with each other, we also enumerate models like

$$Y_i = \beta_0 + \beta_1 X_{1i} X_{2i} + \epsilon_i.$$

From results above we notice that variable "abdomen" is involved in all of the best models. Thus we explore more about it by fitting models like

$$Y_i = \beta_0 + \beta_1 Abdomen_i^2 + \beta_2 X_i + \epsilon_i$$

where X_i is one of other factors. Combining all of above results, our final model is

$$Y_i = \beta_0 + \beta_1 Abdomen_i + \beta_2 weight_i + \epsilon_i.$$

since it has relatively large R^2 value and small MSE. At the same time this model is very easy to use and interpret.

Coefficients given by R are given above. P-value of coefficients suggest that factors "weight"

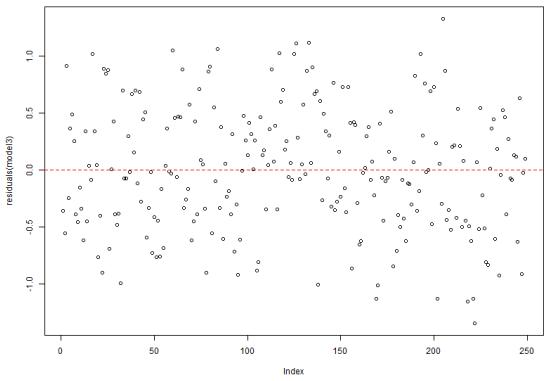
```
Residuals:
     Min       1Q   Median       3Q      Max
-1.34245 -0.38753 -0.01659  0.39215  1.32695

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.914e-16  3.425e-02   0.000        1
WEIGHT      -4.189e-01  6.964e-02  -6.015 6.47e-09 ***
ABDOMEN      1.182e+00  6.964e-02  16.967  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5404 on 246 degrees of freedom
Multiple R-squared:  0.7103,    Adjusted R-squared:  0.7079
F-statistic: 301.5 on 2 and 246 DF,  p-value: < 2.2e-16
```

and "abdomen" are statistically significant. R^2 indicates that it is a fair model.

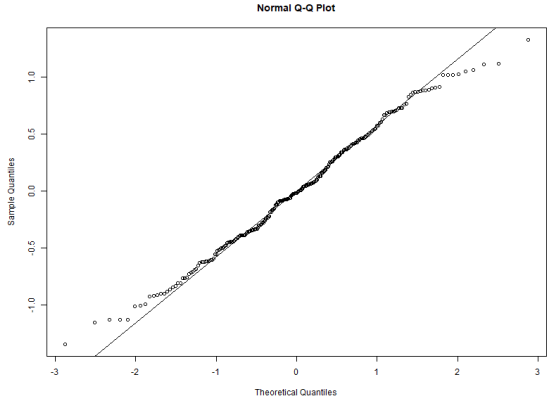
From original data, we have $\mu_{weight} = 177.69(lbs), \sigma_{weight} = 26.48(lbs), \mu_{abdomen} = 92.08(cm), \sigma_{abdomen} = 9.85(cm)$ and $\mu_{bodyfat} = 18.94\%, \sigma_{bodyfat} = 7.75\%$. where μ stands for sample mean and σ for sample standard deviation. Given an individual with weight 200 lbs and abdomen circumference 100 cm, we can calculate his body fat by

$$(1.18 \times \frac{100 - 92.08}{9.85} - 0.42 \times \frac{200 - 177.69}{26.48}) \times 7.75\% + 18.94\% = 23.55\%.$$

Finally, here is the scatter plot and qqplot of residuals of our model. From the scatter plot of



(a) scatter plot of residuals



(b) qqplot

residuals, we find that residuals appear to be randomly scattered around the zero line, with no clear pattern or trend, which indicates independence and homoscedasticity of residuals. Together with qqplot, we may conclude that residuals follows $\mathbf{N}(0, 1)$ distribution, which matches assumption on our model.

In conclusion, this is an easy-to-use and interpretable model which maintaining a certain level of accuracy. However, since we explore only a few combination of predictors and use the simplest model, this result is far from optimal. With more powerful models and more training data, estimation of body fat can be much more precise.

References

[1] Bailey, Covert (1994). Smart Exercise: Burning Fat, Getting Fit, Houghton-Mifflin Co., Boston, pp. 179-186.

[2] Behnke, A.R. and Wilmore, J.H. (1974). Evaluation and Regulation of Body Build and Composition, Prentice-Hall, Englewood Cliffs, N.J.

[3] Siri, W.E. (1956), "Gross composition of the body", in Advances in Biological and Medical Physics, vol. IV, edited by J.H. Lawrence and C.A. Tobias, Academic Press, Inc., New York.

[4] Katch, Frank and McArdle, William (1977). Nutrition, Weight Control, and Exercise, Houghton Mifflin Co., Boston.

[5] Wilmore, Jack (1976). Athletic Training and Physical Fitness: Physiological Principles of the Conditioning Process, Allyn and Bacon, Inc., Boston.

Contribution

Xupeng Tang	: most of code
Xiangsen Dong	: report
Zhengyong Chen	: Github repository and Shiny
Zhaoqing Wu	: part of code