Summary of BodyFat Analysis

Introduction and Motivation

Body fat percentage is the percentage of total body weight that is fat multiplied by 100. it includes both essential and stored fat, which is vital for life and reproduction, and is an indicator of the fitness of an individual's body composition. Measuring body fat accurately can be expensive, so people want to find simpler ways to estimate body fat. Our team's goal was to develop an inexpensive, efficient, reliable and accurate method using readily available clinical data.
We find that a multiple linear model with the variables WEIGHT, ABDOMEN, FOREARM and WRIST predicts BODYFAT very accurately and remains simple and robust.

Background Information and Data Cleaning

We're not factoring in DENSITY as a variable since it's intrinsically tied to body fat through a specific formula $100\times B\ =\ \frac{495}{D}-450$ and can't be directly measured while BodyFat can be measured.
First, we take a rough look at the relationship between BODYFAT and 1/DENSITY. Most of the data points lie on a diagonal line, so we can assume that this diagonal line may indicate a linear relationship between BODYFAT and 1/DENSITY. Our goal is to find data points that do not lie on this line, i.e. outliers. We initially apply the lm() function for a linear model fit. For linear correlations, residuals should be near zero. Data points deviating significantly, especially beyond 1.5 times the standard deviation, are considered outliers. They are marked in red on the plot.

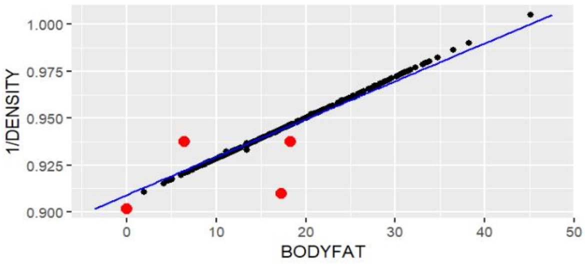

Figure 1

Next, we'll check for additional outliers or influential points. The 42nd data point in the Height histogram indicates a height of 29.5 inches (75cm), unusually short. Yet, its Weight and Adiposity values are normal, suggesting the Height might be a typo.

Final Model Statement

The final multiple linear model is:

$BodyFat\,\% = -34.91 - 0.35\times WEIGHT + 1.00\times ABDOMEN + 0.50\times FOREARM - 1.37\times WRIST$.

According to our model, a man weighing approximately 85 kg, with a waist circumference of 92 cm, a forearm circumference of 31 cm and a wrist circumference of 20 cm would have a predicted body fat percentage of 16.4%. The 95% prediction interval for his body fat percentage lies between 14.9% and 17.9%.
This prediction equation indicates that on average, body fat increases by nearly 10% for every 10 cm increase in abdominal circumference, holding weight, wrist, and forearm measurements constant. In addition, the negative coefficients for weight and wrist measurements follow a certain logic that increases in weight or wrist size may come from sources other than fat. These increases may be due to increased muscle mass or increased bone density, resulting in a decrease in body fat relative to body weight.

Rationale for Final Model & Statistical Analysis

In order to achieve a simple, robust, and accurate model, considering that there may be too many 14 variables, we attempted to perform stepwise model selection based on AIC and BIC standards, and fitted multiple linear models based on the selected variables. And to prevent overfitting, we also attempted Lasso regression.
Comparing the following four indicators, it was found that
1. Accuracy: R^2 of the three models are very close, but the MSE of the BIC selection model is significantly smaller than the other two models, indicating that the BIC model has smaller testing errors and better predictive performance.

2. Simplicity: only four variables in the BIC and Lasso models are smaller than those in the AIC model, indicating that the BIC and Lasso models have lower complexity and are simpler and more convenient.
3. Multicollinearity: the VIF of AIC is greater than 10, indicating that the correlation between model independent variables may be higher. Therefore, we have decided to choose the BIC model.

METHOD	R.Squared	MSE.test	VIF	Sim
AIC	0.7441	17.3468	>10	9
BIC	0.7320	15.5411	<10	4
LASSO	0.7416	17.1149		4

First, for the selected BIC models we test whether the p-value is less than the significance level of 0.05, which means that the four model variables have a significant effect on the dependent variable. Second, we find that the R^2 is 0.73, which indicates that the model fits the data relatively well and that about 73% of the variability in the dependent variable (Bodyfat) can be explained by the independent variables in the model. Then, the VIF is less than 10, which usually means that there is less problem of multicollinearity between the independent variables and the model is more stable.

Model Validation

The Q-Q plot (Figure 2) shows that although slightly skewed at both ends, the residual is a normal distribution approximately.

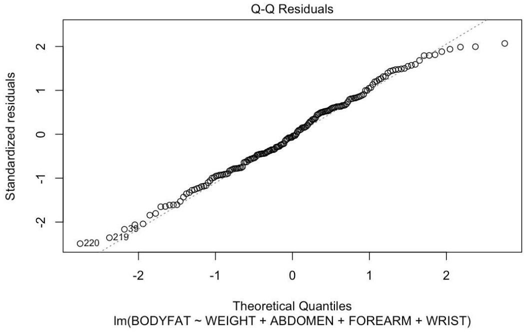

Figure 2: QQ Plot

From figure 3, the residuals "randomly bounce" around the 0 line, which indicates that the assumed relationship is linear. The residuals form a "horizontal band" roughly around the 0 line, indicating that the variance of the error terms is equal.

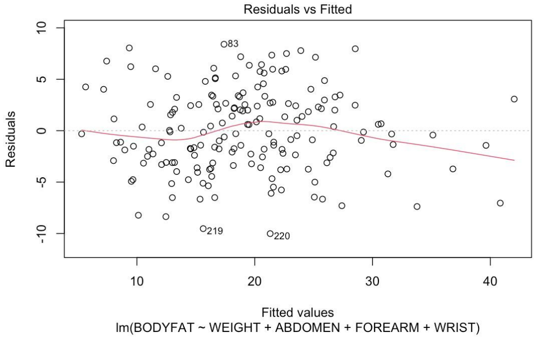

Figure 3: Residuals vs. Fitted Value Plot

Model Strengths & Weaknesses

Strength:
1. Our model has only four independent variables, which is relatively simple and easy to interpret. It is easier for the user to measure.
2. Our model has a high R^2 and the model has a high degree of explanation.
3. Our model meets the assumptions of normality, linearity, and homoscedasticity in linear regression, which adds confidence to our results and interpretations when conducting statistical analysis and making predictions.
Weakness:
1. We divided the raw data into a training and test set, which may lose some information when building the model.
2. Our model is only for normal males, especially for certain groups of males with specific body types (e.g., dwarfism, etc.), which may not yield accurate predictive values.

Conclusion & Discussion

We compared three models and selected the optimal BIC stepwise regression linear model, which has lower complexity and higher accuracy, and satisfies the assumptions of the linear model. Although it may have weaknesses, such as the inability to accurately predict the body fat rate of special populations. We can collect more comprehensive data for dynamic updates of the model in the future. In summary, we believe that we have developed a simple and robust method for calculating body fat.

Contributions:

Jinchen Gong:

Code: Data cleaning code; Model and evaluation indicator code
Report: Background, Data Cleanup, Model, and Statistical Analysis
PPT

Lanxi Zhang:

Code : Model and model hypothesis validation code; Shiny App code
Report: Statistical Analysis, Model Validation, Model Strengths & Weaknesses and Conclusion & Discussion
PPT

Yiyuan Li

Code: Data cleaning code and analysis