Introduction/Background Information

Our ultimate goal is to find a simple and accurate model that well predicts the body fat of people. To gain a better understanding of the data set, we draw a histogram of variable **body fat%**, which turns out to be a bell shape with potential outliers. A correlation plot of all explanatory variables is also drawn, through which we get the insight that a huge fraction of the data set is highly correlated (for example **weight** and **hip** have an correlation of 0.94). As a consequence, we decide to build a linear model with uncorrelated variables after eliminating potential outliers.

Data Cleaning

Through observing the data set and checking the background information about body fat, we found some observations may not be suitable for the model training. 3 different types of suspicious outliers are removed in data cleaning process:

a) Remove individual whose body fat equals to 0, weight equals to 363.15 and height equals to 29.5. (Use head() and tail() to find some extreme points in our data.)

b) We know that BMI follows $weight(\text{kg})/height^2(m^2)$, then we remove the points which are out of the linear relationship between ADIPOSITY and BMI.

c)We get the *Siri equation*: bodyfat fit= 495/density - 450 online. Then remove individual out of the linear relationship between *Siri equation* and the bodyfat in the dataset.

Motivation for Model/Choosing Model/Final Model/Rule of Thumb

Aiming at catering various user demands, we want to find a model that satisfies both conciseness and accuracy. The simplest one should be the one that only need 3 most frequently used measurements (age, height and weight). Using anova, we realize there is no need to take interaction or square into consideration. On the other hand, the most accurate model should be the one selected through BIC criterion or best subsets regression(abdomen, weight, wrist and biceps). Fortunately, the improvement of accuracy using different mode is not significant and the sacrifice of providing other measurements is not that unacceptable, therefore, the best model should be a trade off between these two options mentioned.

We managed to get a model by selecting the significant variables in BIC stepwise result, a model with 3 variables(**abdomen, weight and wrist**). And compare the results with the "simplest" and "most accurate" one to see the advantages. As it's shown in the table, our model yields a high adjusted R^2, low *MSE* and requires relative less inputs.

	adj R-square	Number of variables	CV	True MSE
Simplest model	0.5713	3	25.4	24.8
BIC selected model	0.7329	4	15.8	15.4
Final model	0.7297	3	15.9	15.6

The final model selected is:

$$\begin{aligned} BODYFAT = 0.8914 * ABDOMEN - 0.0922 * WEIGHT \\ - 1.1628 * WRIST - 25.7672 \end{aligned}$$

Rule of thumb:

$$BODYFAT = ABDOMEN - 0.1 * WEIGHT - 1.2 * WRIST - 26$$

In the final model, the estimated coefficients are 0.8914, -0.0922 and -1.1628, which are in the units of *abdomen*, *weight* and *wrist* respectively. This means that for every 1 increase in *abdomen*, the model predicts that body fat % will increase, on average, by 0.8914. Same results can be drawn with other units.

The following example is employed to explain the calculation process: a man with $abdomen = 85.2cm$, $weight = 154.25lb$ and $wrist = 17.1cm$ is expected to have a body fat percentage % of 16.1 based on our model. The corresponding 95% prediction interval is between 15.3% and 16.9%.

Statistical Analysis

We conducted following tests to see whether predictors chosen are significant for prediction. The p-value of F-test is $< 2e^{-16}$, so there is clearly a linear relationship between our 3 variables and body fat % and this linearity is significant at α=0.05 confidence level. The p-values of three predictors are significant respectively.

The adjusted R^2 is 0.7297, which implies that variables in our model can explain 72.97% of the *body fat%*.

Model Diagnostics

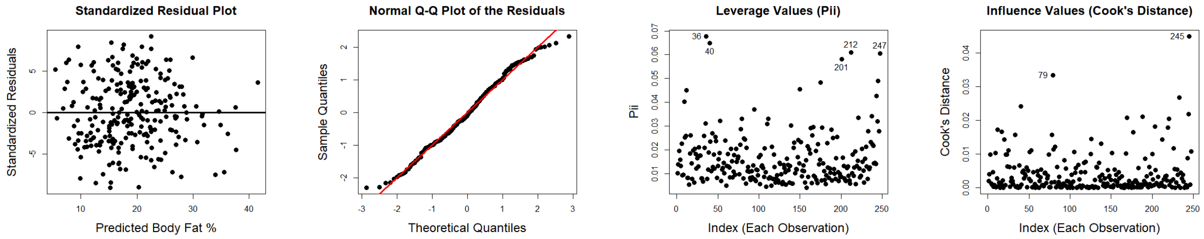

Figure 1: Model Diagnosis

After model fitting, we can first check the multicollinearity, after checking the VIF, we find that there are no severe multicollinearity. Then we perform model diagnose by residual plot (good for diagnosing linearity vioations, homoscedasticity violations, and detecting outliers in Y) and a QQ plot(good for diagnosing violations to Normality). Linearity seems reasonable because there are no obvious non-linear trends in the residual plot; the points look randomly scattered around the X axis. Homoscedasticity is plausible, and normality also seems reasonable because the points in the QQ plot hug the 45 degree line very closely as illustrated in Figure 1-(1)(2).

Check leverage and influential points. For leverage points, we use p_{ii} measures from the lectures. For influential points, we use both the Cook's distance and the p_{ii} measures. From Figure 1-(3)(4), there may be five leverage values but no influential points this time.

Strengths and Weaknesses

Strengths: 1. Our model achieves both conciseness and accuracy, which is a simple linear model and requires only 3 inputs and yields a high accuracy. 2. As it's based on a small data set, our model excluded potential outliers and influence points, yielding a widely application. 3. The assumptions of SLR hold for our model: Homoscedasticity, linearity and no multicollinearity.

Weaknesses: 1. The data set is small and , biased with only males' data which may reduce the robustness, limit the application and impact the prediction accuracy. 2. Some outliers are deleted, which may contain important information.

Conclusion In summary, we build a simple and accurate body fat prediction model that takes *abdomen*, *weight* and *wrist* as input, with and adjusted R^2 of 0.7229 and MSE 0.156.

Contributions

Yiran Wang: Descriptive analysis and data cleaning part of report, slides 1-6.

Xiaofeng Wang: Model testing, selection and statistical analysis of report, slides 7-12.

Jiawei Huang: Model diagnosis and conclusion of report, slides 13-18.