Background

Measurement of body fat can be inconvenient and expensive, thus we aim to construct a simple and accurate model which can be used to predict the body fat. In the body fat dataset, we have 252 male with measurements of percentage of their body fat and various body circumferences.

Step 1: Preprocessing the data

By looking at the summary table of our dataset, we can have a general idea. There exists some abnormal points which be potential outliers of the bodyfat dataset.

In order to detect all the outliers, we first consider Siri's Equation,

$$BodyFat = \frac{495}{Density} - 450$$

We plot the actual bodyfat percentage in the dataset against the bodyfat percentage calculated by Siri's Equation. From the plot, we can tell that point 48, 76, 96 and 182 are far off the line.

IDNO	Bodyfat	Treatment
48	14.1%	change 6.4% to 14.1%
76	14.1%	change 18.3% to 14.1%
96	0.37%	keep original value, impute density with 1.0593
182	-3.6%	check BMI, use BMI to impute Bodyfat with 14.7%

In addition, by looking at the boxplot, we easily identify there are some people whose measurements seem to be very large compared to other datapoints.

IDNO	Problem	Treatment
39	extremely overweight	delete data
41	include many extreme values	delete data
42	too short	use BMI to impute Height = 69.43
216	density<1	delete data

Step 2: Feature Selection

We adopted several approaches here: BIC selection, Lasso feature selection, Boruta Algorithm Selection, and relative importance of different vairables contribution to R-square. Here is the overview of feature Selection Result:

Approach	Selected Features
P Value	Abdomen, Wrist
R-Squared Relative Importance	Abdomen, Chest
BIC	Abdomen, Height, Wrist
Subset Selection	Abdomen, Weight
Lasso	Abdomen, Height
Boruta Algorithm	Abdomen, Chest

Step 3: Model Selection

After the feature selection, we listed several candidate models and used 10-fold cross validation.

Models	R-Squared	F Statistics	MSE(Test)	MSE(Train)
BODYFAT~ABDOMEN+WRIST+HEIGHT	0.7229	218.4	16.40444	16.37404
BODYFAT~ABDOMEN+CHEST	0.6827	269.9	18.66096	18.63678
BODYFAT~ABDOMEN+HEIGHT	0.7028	296.7	17.52315	17.49791
BODYFAT~ABDOMEN+WEIGHT	0.7115	309.3	16.97938	16.95667
BODYFAT~ABDOMEN+WRIST	0.7079	304	17.18307	17.16051
BODYFAT~ABDOMEN	0.6664	500	19.55060	19.53293

Step 4: Final Model Interpretation and Diagnostic

Our selected model is

$$BODYFAT \sim -42 + 0.9ABDOMEN - 0.1WEIGHT$$

The bodyfat can be calculated as the above formula. The rule of thumb:

- If you want to calculate your bodyfat, measure your abdomen and weight circumference, and calculate 90% abdomen minus 10% weight and minus 42, then you will have your estimated bodyfat.

Model Interpretation

As we noticed that the coefficient of abdomen is 0.89470 meaning that as abdomen increased by one unit, the bodyfat will be increased by 0.89470, while as weight increased by one unit, the bodyfat will be decreased by 0.1241. The negative coefficient of weight can be explained that there is tradeoff between abdomen and weight. In real life, skinny people may have high bodyfat.

Model diagnostic

1.Linear relationship: from the Residuals vs Fitted value plot, an approxmately horizontal line, but no distinct pattern is an indication for a linear relationship, which satisfy linearity assumption.

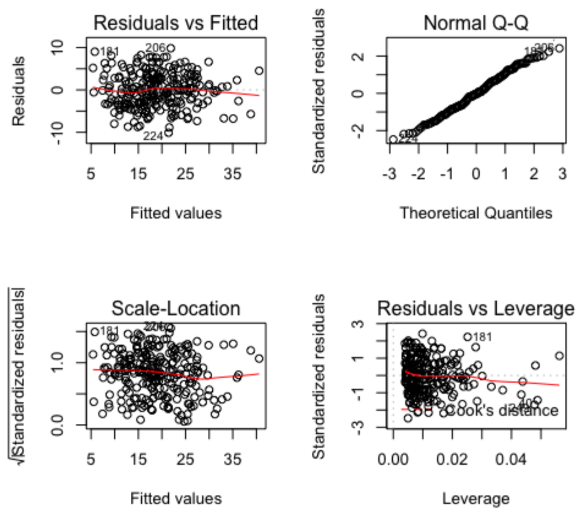
2.Multivariate normality: residuals points mostly follow the dashed line except few head and tail points which is acceptable to retain the normality assumption.

3.No multicollinearity: the VIF of Abdomen and Weight is 4.24 which is below 5, so there is no significant effect of multicollinearity.

4.Homoscedasticity: from the scattter plot of body fat, the null hypothesis that variance is a constance can be retained since there is no certain trend or pattern of variance with index increase.

5.Influential points: in the diagnostics plots, there is several points which stand out, 181 and 206. We remove the two points one by one, however, the increase of R-squared is very limited, so we decide to keep the two points.

6.Robustness: as for model robustness we buid linear model including 39th observation which is a outlier and compare the model with our original model to test robustness. We realize that the coefficients change is smaller than 0.01 and the p-values are still significant.



Strength and Weakness

Strength

- Our model is simple, explicit and interpretable.
- We only have two predictors which makes the interpretation easier.

Weakness

- This dataset contains biased sample, thus our model may only be applicable to limited range of people. Our model can only be used be predict male bodyfat, since our dataset only contains male observations; our model only contains male age from 20 to 80, and the distribution over various age difference are uneven, thus our model may not be effectively generalized.
- The coefficient of weight is negative. If one has light weight but with large abdomen circumference measurements, our model prediction may not be accurate.

Contribution

- Fangfei Lin: Jupytor Notebook Editing, Buruto Model Building, Slides Editing
- Lu Chen: Data Cleaning, Slides Editing, Reference Research, Outlier detection
- Qintao Ying: Lasso Regression, R-shiny, Cross-Validation
- Yansong Mao: Model Diagnostic, Stepwise Methods, Model Selection