Project Summary

Group9: Yun Mo, Fangyang Chen, Anne Huen Wai Wong, Richard Yang

1. Introduction

Nowadays, the problem of obesity is increasingly serious, and lots of people work out to develop good figures. It's important to have a specific metric for them to keep their plan and see their progress,and the bodyfat is one of such measurement.

Although there's a very accurate way to calculate the Bodyfat using density, it's not very convenient to measure and would be cumbersome to use for clinical purpose. So we think it's necessary to build a model to predict the Bodyfat using some available metrics, and select variables as few as possible to make the calculation tool is convenient for everyone.

2. Background information of data

Our data contains covariates of commonly available measurements and Bodyfat or Density as response:

Percent body fat from Siri's (1956) equation, Density determined from underwater weighing, Age (years), Weight (lbs), Height (inches), Adioposity (bmi), Neck circumference (cm), Chest circumference (cm), Abdomen 2 circumference (cm), Hip circumference (cm), Thigh circumference (cm), Knee circumference (cm), Ankle circumference (cm), Biceps, (extended) circumference (cm), Forearm circumference (cm), Wrist circumference (cm)

3. statement of the model

3.1 Remove outliers

3.1.1 According to bodyfat ~ density equation

As we all know, percent body fat `BODYFAT` and density determined from underwater weighing `DENSITY` are correlated. In fact, there's a linear relation between these two variables.

After seeing the relationship between `BODYFAT` and `DENSITY` , We noticed the 48th, 76th and 96th guys do not follow the linear relationship between `1/DENSITY` and `BODYFAT` . Since we don't know which measurement is inaccurate, we'll just remove them from the dataset.

Since the relationship between `DENSITY` and `BODYFAT` is given as follows:

$$bodyfat = \frac{495}{density} - 450$$

unit of measurements: $bodyfat(\%), density(g/cm^3)$

We can use `DENSITY` to calculate bodyfat, and compare it with `BODYFAT` given by the dataset.

	BODYFAT	DENSITY	cal_fat
182	0.00	1.11	-3.61
216	45.10	0.99	47.49

The 96th, 48th and 76th guys have been deleted in the former analysis. The calculated bodyfat for the 182th guy is negative and is impossible, so we remove him. The calculated bodyfat for the 216th guy is close to its `BODYFAT` in the dataset and is too much for a human, we remove him too.

3.1.2 According to the extreme value

Weight

We noticed there is a guy weights larger than 350 lbs. It is obvious that the highest weight guy has an unormal weight. We tried to recover the weight from his BMI and height, that is, `ADIPOSITY` and `HEIGHT` from the dataset.

The BMI calculation formula is:

$$BMI = \frac{weight}{height^2}$$

unit of measurements: $weight(kg), height(m)$

It's almost the same as his `WEIGHT` from the dataset, which means the data is the true value. And because it's too extreme and not representative, we can just remove data.

Height

For the 42th example, we know the guy is a Hobbit but his weight is normal, so we wanna double check it.

	IDNO	BODYFAT	DENSITY	AGE	WEIGHT	HEIGHT	ADIPOSITY	NECK	CHEST	ABDOMEN	HIP	THIGH	KNEE	ANKLE	BICEPS	FOREARM
42	42	31.70	1.02	44	205.00	29.50	29.90	36.60	106.00	104.30	115.50	70.60	42.50	23.70	33.60	28.70

We tried to recover the weight from his BMI and height, that is, `ADIPOSITY` and `HEIGHT` from the dataset.

So we can recover his height from his BMI and weight record. And the height for the 42th guy should be around 69.4 inches.

3.2 Select variables

BIC: Compare the 3 models built by stepwise variable selection(both/backward/forward) with BIC criterion, the model generated by forward selection $model.bic.forward$ has the least number of variables.
The model $model.bic.forward$ has 4 variables, in which the variable `BICEPS` is less significant. In order to get a more simplified model, we refit the model excluding `BICEPS` , i.e our model for BIC is $BODYFAT \sim ABDOMEN + WEIGHT + WRIST$.

LASSO: Variable selection method LASSO selects 4 variables: `AGE` , `HEIGHT` , `ABDOMEN` , `WRIST` . Fitting a Multiple Linear Regression model with these variables, we find Variable `AGE` is less significant in $model\_LASSO$. In order to get a more simplified model, we refit the model excluding `AGE` .

Compare the multiple R-squared of the 2 models we got:

- $model\_BIC$: 0.72
- $model\_LASSO\_final$: 0.7182

$model\_BIC$ has larger multiple R-squared, and they have the same number of variable selected, so we choose it as our final model.

4. Model elaboration

Our final model is:

$$Bodyfat = -22.90319 + 0.87320 \times Abdomen - 0.08140 \times Weight - 1.33584 \times Wrist$$

	Estimate	Std. Error	t value	Pr(>|t|)	2.5 %	97.5 %
(Intercept)	-22.90	6.22	-3.68	0.00	-35.16	-10.64
ABDOMEN	0.87	0.05	16.67	0.00	0.77	0.98
WEIGHT	-0.08	0.02	-3.63	0.00	-0.13	-0.04
WRIST	-1.34	0.40	-3.32	0.00	-2.13	-0.54

The final model contains 3 variables: **Abdomen**, **Weight** and **Wrist**. The Multiple R-squared for this model is 0.72, which means these 3 variables can explain about 72% variability of the bodyfat.

When consider F-test for null hypothesis: $\beta_1 = \beta_2 = \beta_3 = 0$, the corresponding p-value is less than $2.2e^{-16}$. So we reject the null hypothesis.

For the coefficients for the 3 variables, when testing for null hypothesis: $\beta_i = 0$, the p-values for each test is very significant (less than 0.01). This means the possibility for falsely rejecting each null hypothesis is less than 0.01.

We can also further check the 95% confidence intervals which list in above table. None of the confidence intervals for the 3 variables including 0, which also support that the 3 coefficients are not 0.

5. Layman interpretation

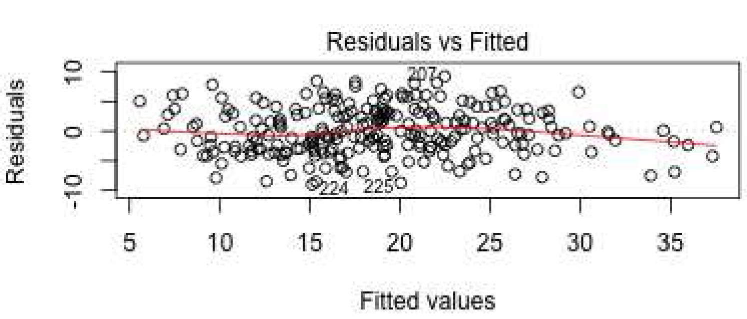
The approximate of final model is:

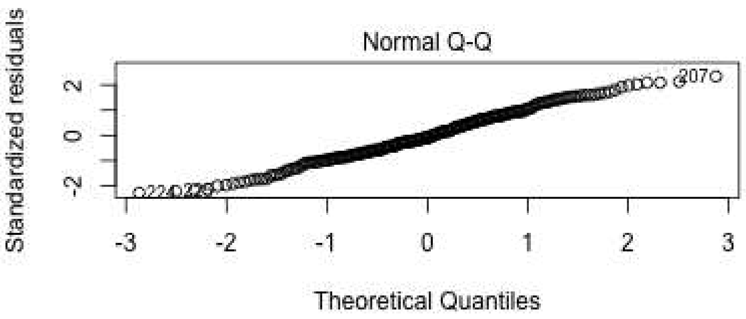
$$Bodyfat = -23 + 0.9 * Abdomen - 0.08 * Weight - 1.3 * Wrist$$

That is, when fixing other two variables, 1cm increasment in **Abdomen** leads to 0.9 increasment in Bodyfat; 10 lbs increasment in **Weight** leads to 0.8 decreasment in Bodyfat; 1cm increasment in **Wrist** leads to 1.3 decreasment in Bodyfat.

6.Model diagnostics

In this part, we check 3 assumptions of a Simple Linear Regression model by using diagnostic plots.





- **Normality.** From the Normal Q-Q plot, the residuals may have a heavy right tail, but in general, the points line on the diagonal can support the normality.
  We also used the Shapiro-Wilk normality test to check the normality, which gets a p-value of 0.09072. So we can not reject the normality assumption.
- **Equal variance.**
  From the Residuals vs Fitted plot, the variance of residuals are approximately equal.
- **Independence.**
  We calculated the autocorrelation of the residuals. It shows that for all lags, the autocorrelation are inside the critical bound. So the independece assumption is satisfied.

7. Strength of our analysis

- Our model has considered all possible outliers in the procedure of preprocessing data.
- The model has included multiple strategies for variable selection, which helps it generalize better in real data.
- We implemented the whole procedure of diagnostics and can guarantee the model meets most assumption
- The model is concise enough for just considering the linear relationship between the variables and responses, which would be easy to understand to doctors with limited statistics knowledge。

8. Weakness of our analysis

- Our data has lots of outliers when evaluated in different aspects, which makes the analysis not that reliable.
- The size of the dataset is too small, so we cannot really tell what's the random noise, and some analysis maybe too subjective.
- The model is too simple because of the doctor's knowledge level, which is not complicated enough to capture their relationship.
- The variables selected doesn't coincide with the common sense, which maybe due to the noise in our data.

9. Contribution

- Yun Mo was responsible for variables selection and model diagnosis.
- Fangyang Chen was responsible for coding and testing the Shiny app.
- Anne Huen Wai Wong played an major role in building model for prediction.
- Richard Yang wrote the summary and slides for presentation.

10. Conclusion

- The best choice of linear model features are **Abdomen**, **Weight** and **Wrist**, which can roughly characterize the `BODYFAT` and at the meantime is concise enough to explain.
- Our model satisfies most assumptions, which means our analysis or inference is valid in general.
- We can just measure above three easily acquired features to get a not precise estimate of male's Bodyfat.