Bodyfat Calculator

Thursday group 9

October 9, 2019

Outline

1. Remove Outliers
2. Model Fitting
3. Model Evaluation
4. Calculator application
5. Strength and weakness

Remove outliers

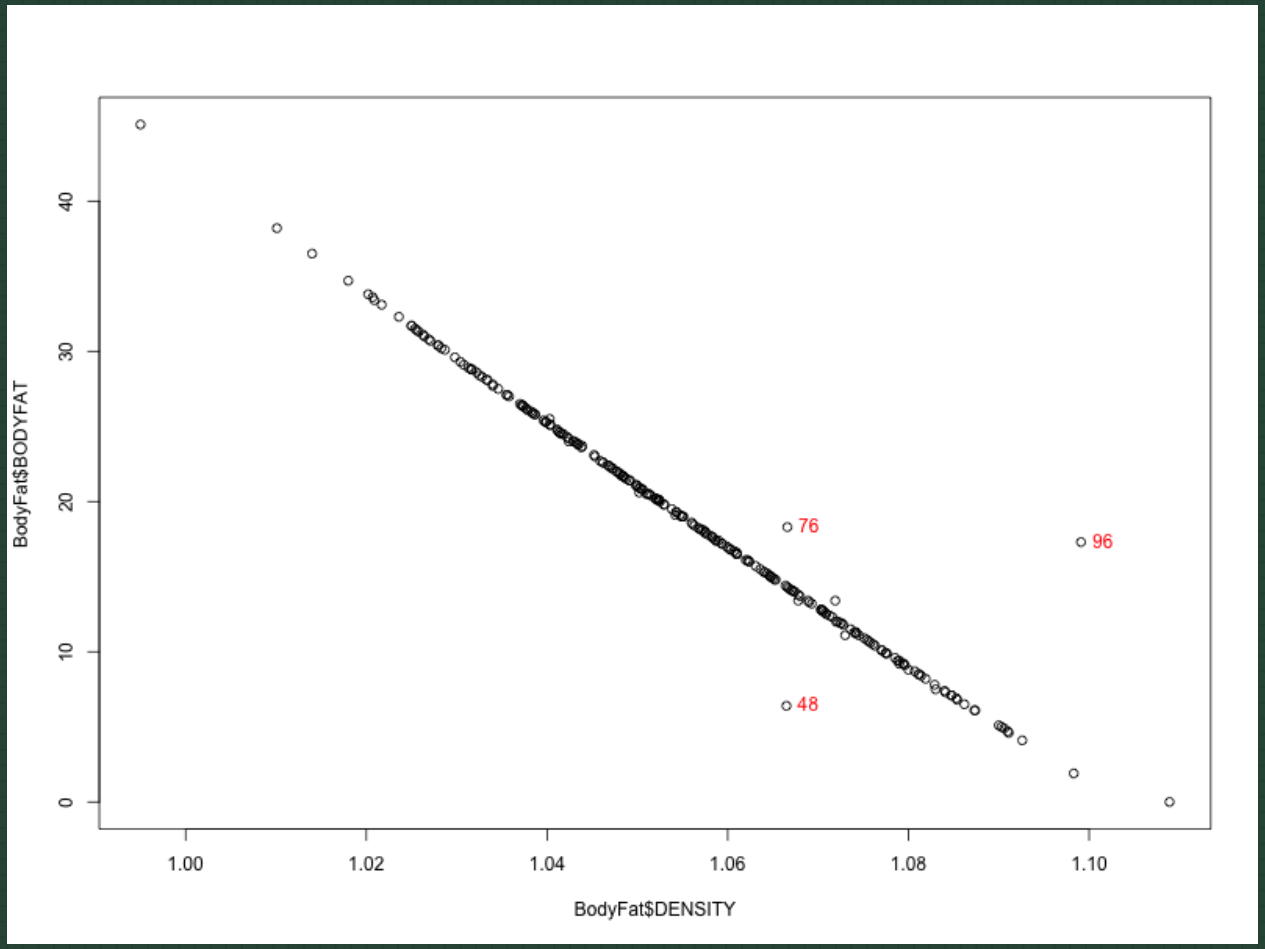

Figure: Density bodyfat graph

Remove outliers

	BODYFAT	DENSITY	cal_fat
96	17.30	1.10	0.37
48	6.40	1.07	14.14
76	18.30	1.07	14.09
182	0.00	1.11	-3.61
216	45.10	0.99	47.49

- The 96th, 48th and 76th examples should be deleted. 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.

Remove outliers

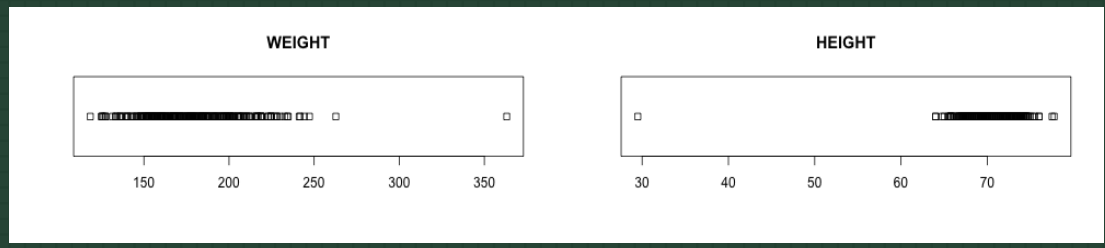

Figure: The extreme value

According to the extreme weight:

(1) We noticed there is a guy's weight larger than 350 lbs.

(2) Using $BMI = weight/height^2$, the calculated weight is the same as dataset, which means the data is the true value. And because it's too extreme and not representative, we can just remove data.

According to the extreme height:

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

(2) We tried to recover the height from his BMI and height. And the height for the 42nd guy should be around 69.4 inches.

Remove outliers

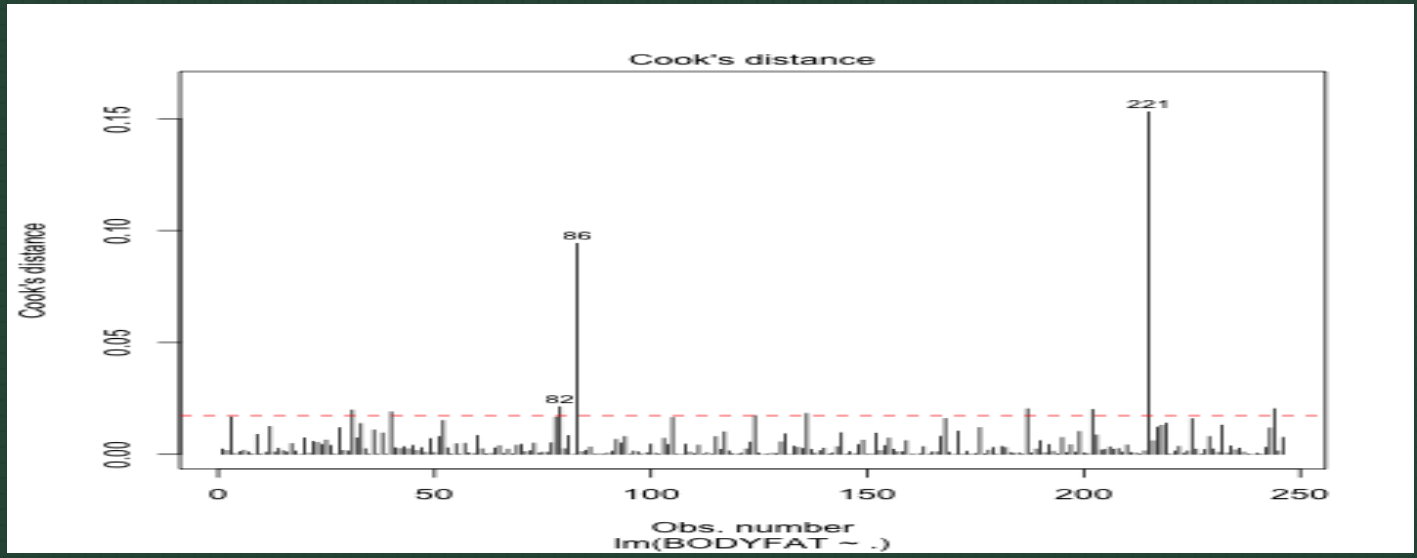

Figure: The cook distance

Although some cases are extreme in terms of cook distance, we just don't have enough evidence or reason to remove such data, so we just use the final data with 48th, 76th, 96th, 182nd, 216th data removed and 42nd data modified.

Model Fitting

- We use two variable selection strategies to decide our final model. One is step-wise selection using BIC as the criterion and the other is using regression penalization method LASSO, which both could shrink the number of parameters efficiently.

(1) [Step-wise] Using forward/backward/both-side to find out the predictors with least BIC value and organized a simplest model with three variables {Abdomen, Weight, Wrist}. The fitted linear model has a coefficient of determination of $R^2 = 0.72$.

Model Fitting

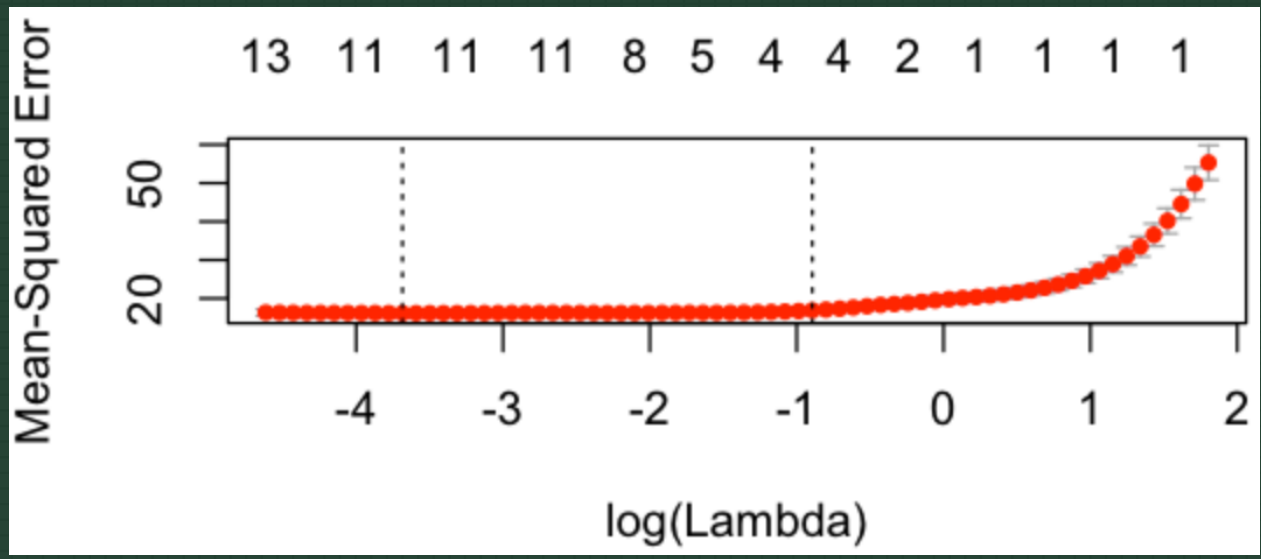

Figure: CV for choosing penalization parameter λ

(2) [LASSO] After we decide the penalization parameter λ using cross-validation, we get the predictors {Abdomen,Height,Wrist} with the $R^2 = 0.7182$

Model Evaluation

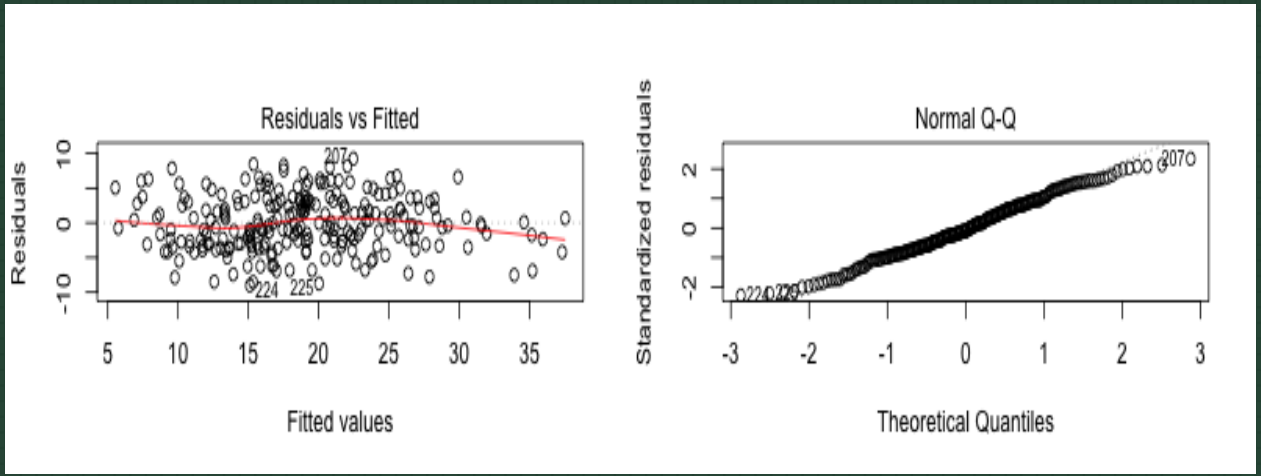

Figure: Diagnostics

- **Normality**: From the Normal Q-Q plot, 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.
- **Constant variance:** From the Residuals vs Fitted plot, the variance of residuals are approximately equal.

Model Evaluation

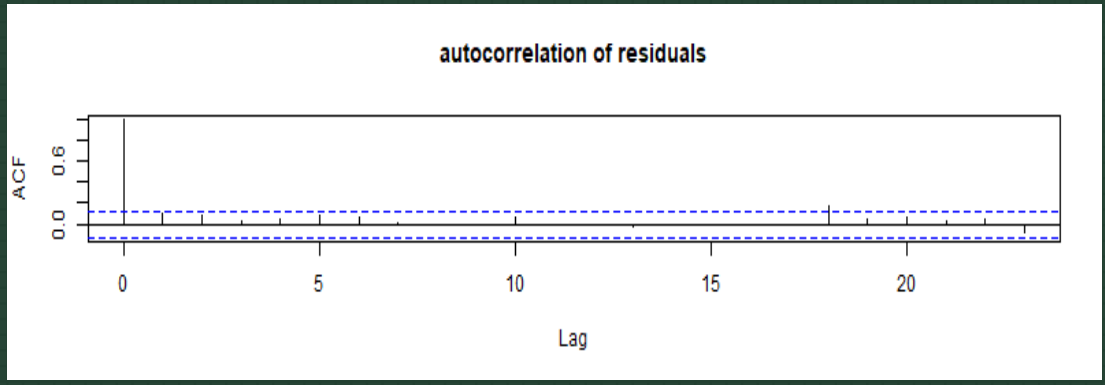

Figure: Diagnostics

- **Independence:** We calculated the autocorrelations of the residuals. It shows that for all lags, the autocorrelations are inside the critical bound. So the independence assumption is satisfied.

Model Evaluation

- **Final Model**:
  $Bodyfat =$
  $-22.903 + 0.873 \times Abdomen - 0.081 \times Weight - 1.336 \times Wrist$

	Estimate	Pr(>|t|)	2.50%	97.50%
Intercept	-22.903	2.877e-04	-35.163	-10.643
ABDOMEN	0.873	4.504e-42	0.770	0.976
WEIGHT	-0.081	3.520e-04	-0.126	-0.037
WRIST	-1.336	1.036e-03	-2.128	-0.543

Model Evaluation

- **Layman interpretation**:

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

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.

Calculator application

The url for our app is:

`https://team9bodyfat.shinyapps.io/628BodyFat/`

Strength and weakness

Strength:

(1) Our model has considered all possible outliers in the procedure of preprocessing data, and implemented the whole procedure of diagnostics and can guarantee the model meets most assumption.

(2) The model has included multiple strategies for variable selection, which helps it generalize better in real data.

(3) 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.

Weakness:

(1) The size of the dataset is too small, so we cannot really tell what's the random noise, and some analysis maybe too subjective.

(2) The model is too simple because of the doctor's knowledge level, which is not complicated enough to capture their relationship.

(3) Some variables selected doesn't coincide with the common sense, which maybe due to the noise in our data.

Thank you.