STAT 628 Module 2 BodyFat Group 3

Introduction:

Our motivation is to obtain a precise, concise and reasonable regression model to predict body fat percentage of a male based on his body informations. Our dataset `BodyFat` is a real data set of 252 men with measurements of their percentage of body fat and various body circumference measurements. We assume that percentage of body fat for an individual can be estimated once body density has been determined:

$$BodyFat(\%) = \frac{495}{BodyDensity(gm/cm^3)} - 450 \quad (1)$$

And the BMI can be calculated by:

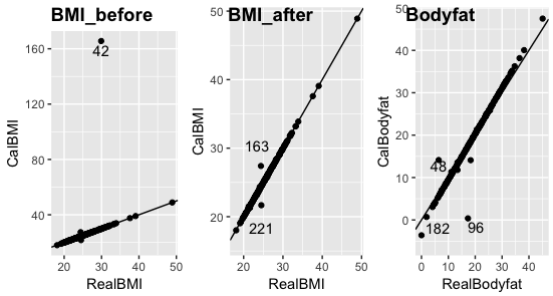
$$BMI = \frac{703 \times Weight(lbs)}{Height(inches)} \quad (2)$$

1. Exploratory data analysis

We can first intuitively analyze the data by looking at its summary tables and histograms. There are some obvious "abnormal" obsevations: For example, the minimal `BODYFAT` is 0 %, which is unlikely to be true; besides, the maximal `BODYFAT` is 36 %, which is too large for a living person. We can calculate the predicted values of these "abnormal" data using some formula or just delete these observations.

2. Data cleaning

Calculate the rages of `WEIGHT` , `HEIGHT` , `DENSITY` , `BODYFAT` between there 1% quantiles and 99% quantiles respectively, any value out of such ranges can be considered as "abnormal" observations. We can then compare the body fat percentages calculated from formula (1) (`CalBodyFat`) and given by the dataset (`RealBodyFat`) of each observation to see if there is any strong difference. Do the same thing for BMI based on formula (2).



BMI: The 42nd observation seems to be an outlier. Check his `WEIGHT` and `HEIGHT` to find out where is the problem. We can see that the *42nd* observation's height (**29.5** inches) is out of quantile range, so we use the formula (2) to calculated the height based on his weight and BMI, it turns out that his height should be **69.43** inches. After changing the height of the *42nd* observation, we can find that the 163rd and 221st observations still don't fit the line perfectly, but after similar analysis, we find out that their heights and weights is within the quantile range above, so we won't change them.

BodyFat: From the graph above, we can see that the *182nd* observation have a **ZERO** calculated body fat percentage and a **NEGATIVE** given body fat percentage, which is unlikely to be true for a normal male, so, we will just remove this observation. The *96th* obsrvation has a density far away from the density quantile range,

since people can not easily obtain body density in real life, we will just remove this observation (We also drop the column `DENSITY`). Finally, we use the 0.01 and 0.99 quantiles to replace the values out of 0.01 ~ 0.99 ranges and save the cleaned data into a `.csv` file.

3. Feature Selection

3.1 Cross Validation

We use a **10-fold cross validation** to compute the average MSE (mean-square error) of each model:

$$MSE = E((Y - \hat{Y})^2) \tag{3}$$

We shuffle the data before every CV and repeated 1000 times.

3.2 Stepwise model selection

Perform stepwise selections: We choose **AIC** and **BIC** methods: Let k denote the number of parameters, let RSS denote the residual sum of squares, let n denote the number of observations, we have:

AIC (Akaike Information Criterion):

$$AIC = 2k + n \ln \frac{RSS}{n} \tag{4}$$

Choose the model with minimal AIC:

$$BODYFAT \sim ABDOMEN + WEIGHT + WRIST + AGE + THIGH + NECK + FOREARM$$

BIC (Bayesian information criterion):

$$BIC = k \ln n + n \ln \frac{RSS}{n} \tag{5}$$

Choose the model with minimal BIC: $BODYFAT\ AGE + ABDOMEN + WRIST + HEIGHT$

3.3 LASSO

Let β denote the parameters, we have:

LASSO (least absolute shrinkage and selection operator) regression uses an objective function as:

$$f(\beta) = RSS_\beta + \lambda ||\beta||_1 \tag{6}$$

Find the β that minimizes $f(\beta)$. Since LASSO can force some parameters to become zero, we can use LASSO to perform feature selection. It is not difficult to find that **LASSO** and **BIC** methods give the same feature selection result. Since the CV_MSE of AIC and BIC model are close to each other, we will choose BIC model because it has less predictors.

3.4 Other Models

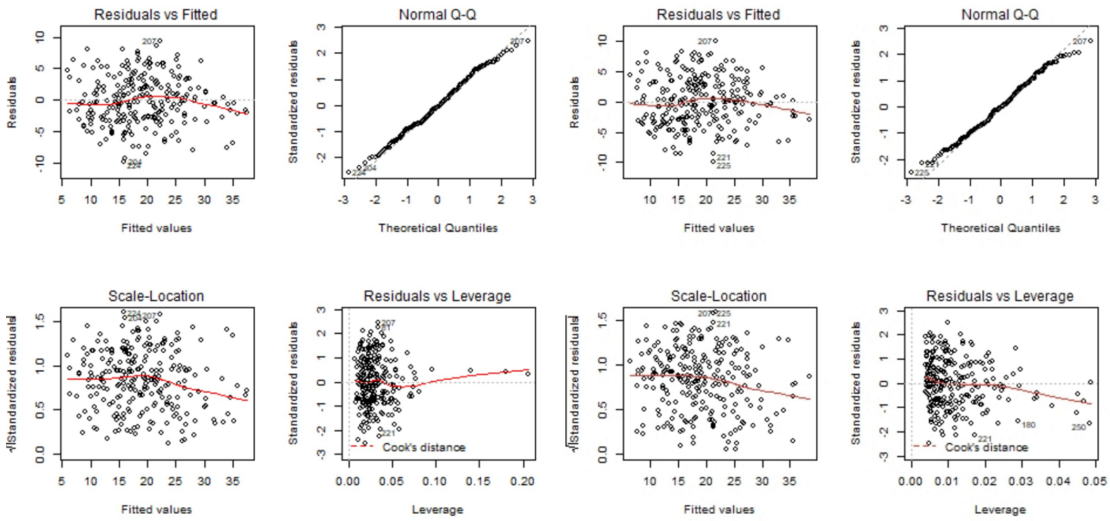
According to the correlation of bodyfat and other variables, abdomen and weight have relatively higher correlation, so we will try to use this two variables and build other models to see what happen:

$$1st : BODYFAT \sim WEIGHT + ABDOMEN$$
$$2nd : BODYFAT \sim ABDOMEN : WEIGHT$$
$$3rd : BODYFAT \sim ABDOMEN$$

Since the 1st model of the three other models has the lowest CV_MSE, we will chooswe this model.

Model	AIC	BIC	Model1	Model2	Model3
R-squared	0.7418	0.7336	0.7181	0.5293	0.6785
MSE	15.4747	15.6591	16.2636	26.9229	18.3756
p-value	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$	$< 2.2e-16$

3.5 Model Diagnostics Plots



1. From the residual plots we can see that linearity and equal variance assumptions are satisfied for both **BIC** model and **1st** model, because the residual points are evenly distributed on both sides of the line and they are both very close to the x-axis.
2. From the QQ plots, we can see that mornality assumptions are satisfied for both **BIC** model and **1st** model.
3. From the cook's distances in the residuals vs leverage plots, we can assume there is no outlier.

4. Strength & weakness

Strength: 1. To deal with outliers, for each feature, we use 99% quantile to replace the values greater than the 99% quantile and 1% quantile to replace the values less than it, by doing so, we can prevent our model from the influences of extreme values. 2. We repeat 10-fold Cross Validation 1000 times, which can make our results more convincible. 3. Our final model is simple linear regression model, which contains only two significant predictors, and they are easy to measure during daily life in real world. That makes our model concise and easy to interpret.

Weakness: 1. The methods we use in data cleaning may cause some internal changes of the data structure, which can affect our model results. 2. We only choose two predictors for simplicity, but there is possibility that model with more predictors may give us better results in some specific cases.

5. Conclusion

Choose the model with less predictors, our final model is:

$$BODYFAT = -41.60 - 0.12 \times WEIGHT + 0.88 \times ABDOMEN$$

Possible rule of thumb: "To calculate your body fat percentage, mutiply your ABDOMEN by 0.88, minus your WEIGHT multiplied by 0.12, then minus 41.60."

6. Contribution

- Lingfeng ZHU: Implemented exploratory data analysis and copleted the Shiny App.
- Ruochen YIN: Completed presentation slides and presentation-related works.
- Jiahan LI: Implemented data cleaning and part of model diagnostic plots.
- Chong WEI: Implemented part of model selection and cross validation.