STAT628_Module_2_Group_9

1 Introduction

1.1 Background

Body fat percentage is a measure of fitness level, which is not easy to obtain during clinical application. A popular way to estimate the body fat is by using the Siri's equation. However, body density is also hard to obtain.

This project aims to come up with a simple, robust, accurate and precise "rule-of-thumb" method to estimate percentage of body fat using clinically available measurements.

1.2 Description of Dataset

1.2.1 Formula

$$Body\ Fat\ \%\ (i.e. 100 \times B) = \frac{495}{D} - 450,\ D = Body\ Density\ (gm/cm^3)$$

$$ADIPOSITY\ (BMI) = \frac{Weight\ (lbs) \times 703}{[Height\ (inch)]^2}$$

1.2.2 Overview of the dataset

We have **252** observations and **17** variables, their units are given in the parenthesis.

Response variable: Body fat in percentage Predictors: Age (years), Weight (lbs), Height (inches), Adiposity (bmi), Neck circumference (cm), Chest circumference (cm), Abdomen circumference (cm), Hip circumference (cm), Thigh circumference (cm), Knee circumference (cm), Ankle circumference (cm), Biceps (extended) circumference (cm), Forearm circumference (cm), Wrist circumference (cm)

2 Data Cleaning

2.1 Use Summary Table to Check Extreme Values

We use the summary table to get an overview of the data. What surprises us is the extreme values in some variables

From the summary table, we can see there exist extreme values in the some variables.

BODYFAT	DENSITY	WEIGHT	HEIGHT	ADIPOSITY
Min. : 0.00	Min. :0.995	Min. :118.5	Min. :29.50	Min. :18.10
1st Qu.:12.80	1st Qu.:1.041	1st Qu.:159.0	1st Qu.:68.25	1st Qu.:23.10
Median :19.00	Median :1.055	Median :176.5	Median :70.00	Median :25.05
Mean :18.94	Mean :1.056	Mean :178.9	Mean :70.15	Mean :25.44
3rd Qu.:24.60	3rd Qu.:1.070	3rd Qu.:197.0	3rd Qu.:72.25	3rd Qu.:27.32
Max. :45.10	Max. :1.109	Max. :363.1	Max. :77.75	Max. :48.90

The observations that have extreme values are

IDNO	BODYFAT	DENSITY	AGE	WEIGHT	HEIGHT	ADIPOSITY	NECK	CHEST	ABDOMEN	HIP	THIGH	KNEE	ANKLE	BICEPS	FOREARM	WRIST
172	1.9	1.0983	35	125.75	65.50	20.6	34.0	90.8	75.0	89.2	50.0	34.8	22.0	24.8	25.9	16.9
182	0.0	1.1089	40	118.50	68.00	18.1	33.8	79.3	69.4	85.0	47.2	33.5	20.2	27.7	24.6	16.5
216	45.1	0.9950	51	219.00	64.00	37.6	41.2	119.8	122.1	112.8	62.5	36.9	23.6	34.7	29.1	18.4
39	33.8	1.0202	46	363.15*	72.25	48.9*	51.2*	136.2*	148.1*	147.7*	87.3*	49.1*	29.6	45.0*	29.0	21.4*
42	31.7	1.0250	44	205.00	29.50	29.9	36.6	106.0	104.3	115.5	70.6	42.5	23.7	33.6	28.7	17.4

BODYFAT is the response variable whose reasonable value ranges from 2% to 39%. Individual 172 has lowest possible body fat, which can be considered as essential fat; Individual 216 is sever obesity, which is possible; Individual 182, it's impossible to have 0% of bodyfat, and after checking the siri's equation to his density, the corresponding bodyfat becomes negative, thus we filter this records out of our analysis.

There also exists extreme value in **WEIGHT**, which occurs in individual 39. This man also has the largest value in ADIPOSITY, NECK, CHEST, ABDOMEN, HIP, THIGH, KNEE, BICEPS, and WRIST. Which indicates that this record does exist.

As for **HEIGHT**, individual 42's height is only 29.5 which is quite abnormal. After checking the corresponding weight by BMI formula, we can assume that this is a wrong record. Thus we fix his height by applying the BMI formula.

2.2 Check Siri's Equation

Known that the body fat percentage can be estimated by the density with the Siri's equation. We build a linear model between the bodyfat percentage estimated by Siri's equation and the bodyfat percentage in the data set. The residual plot and the QQ plot of this model is shown below. We can see that record 48, 76, 96 are possible outliers.

IDNO	BODYFAT	DENSITY	AGE	WEIGHT	HEIGHT	ADIPOSITY	NECK	CHEST	ABDOMEN	HIP	THIGH	KNEE	ANKLE	BICEPS	FOREARM	WRIST
48	6.4	1.0665	39	148.50	71.25	20.6	34.6	89.8	79.5	92.7	52.7	37.5	21.9	28.8	26.8	17.9
76	18.3	1.0666	61	148.25	67.50	22.9	36.0	91.6	81.8	94.8	54.5	37.0	21.4	29.3	27.0	18.3
96	17.3	1.0991	53	224.50	77.75	26.1	41.1	113.2	99.2	107.5	61.7	42.3	23.2	32.9	30.8	20.4

Individual 96's other variables all have normal value, which indicates his desity might be wrongly recorded.

Individual 48 and 76 have similar values in other variables, thus their body fat percentage should also be similar. Thus we use the Siri's equation to fix their body fat percentage.

2.3 Check the BMI Formula

Known that the ADIPOSITY can be estimated using WEIGHT and HEIGHT. We build a linear model between the BMI estimated by equation and the ADIPOSITY in the data set. The residual plot and the QQ plot of this model is shown below. We can see that record 163, 220, 234 are possible outliers.

IDNO	BODYFAT	DENSITY	AGE	WEIGHT	HEIGHT	ADIPOSITY	NECK	CHEST	ABDOMEN	HIP	THIGH	KNEE	ANKLE	BICEPS	FOREARM	WRIST
163	13.3	1.0690	33	184.25	68.75	24.4	40.7	98.9	92.1	103.5	64.0	37.3	23.5	33.5	30.6	19.7
220	15.1	1.0646	53	154.50	69.25	22.7	37.6	93.9	88.7	94.5	53.7	36.2	22.0	28.5	25.7	17.1
234	25.9	1.0384	58	161.75	67.25	25.2	35.1	94.9	94.9	100.2	56.8	35.9	21.0	27.8	26.1	17.6

For individual 163, 220, and 234, their weights vary a lot, which means their BMI should not be similar to each other. Thus we adopt the calculated BMI as their ADIPOSITY.

2.4 Data cleaning summary

Record **182** is filtered out because it has 0 body fat and there's no way to fix that.

The HEIGHT of record **42** is fixed according to the weight and adiposity.

The body fat of record **48** and **76** are fixed according to the density.

The adiposity of record **163**, **220**, and **234** are fixed according to the weight and height.

3 Variable Selection and Statistical Modeling

3.1 Variable Selection

3.1.1 Stepwise Backward and Forward LR

Considering there is a tradeoff between model variance and accuracy, we'd better not use all 14 predictors to establish the model. Although the mean square of error (MSE) will be small, the model itself will be unstable. Firstly, we try normal linear regression using the stepwise method to select features. We chose BIC as criterion and directions of forward and backward. The result was shown respectively.

-

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-29.9234561	6.69406028	-4.470150	1.193678e-05
ABDOMEN	0.9133283	0.05163326	17.688759	9.869649e-46
WEIGHT	-0.1238172	0.02276101	-5.439883	1.285958e-07
WRIST	-1.4082304	0.40713801	-3.458853	6.391933e-04
FOREARM	0.4253252	0.16733889	2.541700	1.164621e-02

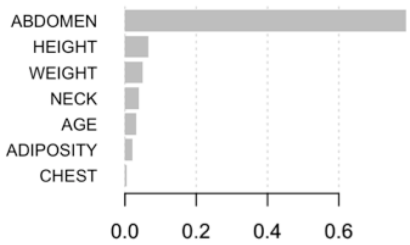
The results of backward and forward method are the same, WEIGHT, ABDOMEN, FOREARM, WRIST were chosen as predictors and they are all significant. R-square of the model is 0.736, which indicates the accuracy of the model is quite good. However, in the linear model, collinearity of predictors will increase the variance of model. VIF is a value to identify collinearity. There are collineary between variables if VIF is larger than 5.

```
 ABDOMEN   WEIGHT    WRIST  FOREARM
4.789592 6.924539 2.242982 1.770719
```

The model is still unstable, so we try other methods to do variable selection.

3.1.2 Variable Selection with XGBoost

XGBoost is an ensemble learning method with tree models. It will output the importance of variables after establishing the model. The criterion for ranking predictors is how many times the variable is chosen as the node in decision trees. The more frequent it is chosen the larger score it will get. Moreover, if the variable is chosen near root of tree, the score will has a large weight because its influence to the decision tree is large.



3.2 Stable Statistical Model

In both of our methods, we find that ABDOMEN is the most important variable among 14 predictors. WEIGHT ranks high in both methods comparing with other variables. WRIST is significant in linear stepwise model and ADIPOSITY scores high in nonlinear model. Considering the rule of thumb, we select ABDOMEN as our first variable in predicting body fat. Also, to increase the accuracy we choose another variable among WRIST, WEIGHT, ADIPOSITY as our second variable. The number of predictors are too small to use any tree model or ensemble method in machine learning. We still choose normal linear regression as our model. The collinearity are not significant between two variables, so we may not add regulization to our model. After comparing three models, we choose ABDOMEN and WEIGHT as our predictors, because it has highest R-square 0.72.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-40.6768857	2.41526710	-16.841568	5.953549e-43
ABDOMEN	0.9082391	0.05222159	17.392023	7.810903e-45
WEIGHT	-0.1364208	0.01914546	-7.125489	1.124027e-11

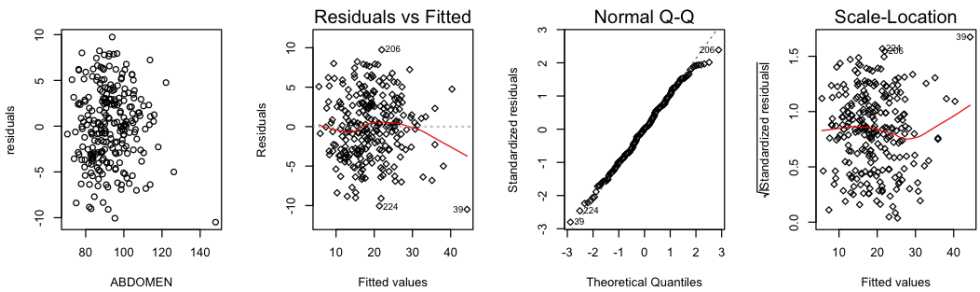
3.3 Model Diagnostics & Strength and Weekness

3.3.1 Final Model:

$$BODYFAT(\%) = 0.908 * ABDOMEN(cm) - 0.136 * WEIGHT(lb)$$

3.3.2 Model Diagnostics:

Firstly, use VIF to check collinearity. There is no obvious collinearity between ABDOMEN and WEIGHT. Then, plot the residuals vs fitted values. Residuals are seperated randomly near 0 and there is no correlation between fitted value and residuals. Also, there is no correlation between residuals and predictors. QQ-plot of residuals show that residuals generally follows normal distribution. It is normal that there are still some outliers in the model.



3.3.3 Strength and Weakness of our model:

Generally, this model is a simple, robust, accurate and efficient model. It satisfies assumptions in linear regression model and explains 72% of the variation in body fat % among men although. It also has weaknesses as it cannot capture higher order effects and interactions.

4 Shiny APP

Based on the model above, we developed the shiny app.

Extra packages Beside the "shiny" package, we also used "shinybulma", "grid", "png", "bs4Dash", "rsconnect", "ggplot2", "shinyalert".

Function

Body Fat calculator. Size converter(kg to lbs, inch to cm). Detail information about body fat.

Web-based app Link https://jiawen1014.shinyapps.io/BodyFat/

5 Contribution

Jiawen Chen: Web-based App
Chunyuan Jin: Model building and model diagnosing
Han Liao: Data cleaning