Body Fat Calculator

Kangyi Zhao, Xinyu Zhang, Kehui Yao

1 Introduction

1.1 Motivation

In the past, The most straightforward way to judge a person's fatness is to use the ratio of height to weight (BMI, weight divided by the square of the height). However, this method is easy to ignore the amount of body fat. Therefore, to judge a person's true degree of obesity, in addition to using BMI as a reference indicator, the body fat rate must be tested, so that the results will be more objective.

There are a lot of ways designed to test body fat rate, like body-fat scale, online body-fat calculator. We try to build a model only predicts the male, a simple, robust, accurate and precise"rule-of-thumb" method to estimate the percentage of body fat based on the 252-men data set. To make sure the simplicity of the model, we try to keep one or two variable in our model.

1.2 Description of dataset

1.2.1 Formula

We can get the formula for estimating the **body fat B(%)** from previous study

$$Percentage of BodyFat(i.e.100 \times B) = \frac{495}{D} - 450, D = BodyDensity(gm/cm^3)$$

We can get also get the formula for estimating **ADIPOSITY (bmi)**

$$ADIPOSITY(bmi) = \frac{weight(lbs) \times 703}{height(in^2)} - 450$$

1.2.2 Glimpse at the dataset

In [2]:
```
BodyFat = read.csv("BodyFat.csv"); head(BodyFat,2)
```

IDNO	BODYFAT	DENSITY	AGE	WEIGHT	HEIGHT	ADIPOSITY	NECK	CHEST	ABDOMEN	HIP	THIGH	KNEE	ANKLE	BICEPS	FOREARM	WRIST
1	12.6	1.0708	23	154.25	67.75	23.7	36.2	93.1	85.2	94.5	59.0	37.3	21.9	32.0	27.4	17.1
2	6.9	1.0853	22	173.25	72.25	23.4	38.5	93.6	83.0	98.7	58.7	37.3	23.4	30.5	28.9	18.2

We have **252** observations and **17** variables, with their precisions and units included in the parenthesis.

- Response variable: Body fat (0.1)
- Predictive variables and their precison unit: Age(1 years), Weight(0.25 lbs), Height(0.25 inches), Adiposity(0.1 bmi), Neck circumference(0.1 cm), Chest circumference(0.1 cm), Abdomen circumference(0.1 cm), Hip circumference(0.1 cm), Thigh circumference(0.1 cm), Knee circumference(0.1 cm), Ankle circumference(0.1 cm), Biceps (extended) circumference(0.1 cm), Forearm circumference(0.1 cm), Wrist circumference(0.1 cm)

2 Data Cleaning

To begin with, we used boxplot() and summary() to have an overview of the data. The following five points stand out: 216's high bodyfat, 182's zero bodyfat, 39's large weight, and 42's short height.

In [3]:
```
summary(BodyFat[,c(2,5,6)])[c(1,6),] #Boxplot in slides.
```

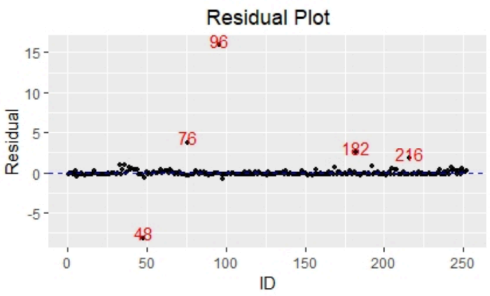
```
    BODYFAT          WEIGHT          HEIGHT     
 Min.   : 0.00   Min.   :118.5   Min.   :29.50  
 Max.   :45.10   Max.   :363.1   Max.   :77.75  
```

For data cleaning, we have summarized the following three criteria:

1. NO imputation for any records of response variable, BODYFAT.
2. Exclude records which are verified as wrong and unable to impute correctly.
3. Exclude extreme and high influential points, for the privilege of fitting a robust and common applied model.

2.1 Check BODYFAT Using DENSITY

Since BODYFAT is the response variable, the most crucial one during the whole analysis, as criterion 1 mentions, we do not impute BODYFAT. And as the following regression plot shows, BODYFAT is inversely proportional with DENSITY. And aftering detecting the following outliers one at each time: 96, 48, 76, 182, and 216, we trained the model without aforementioned outliers to predicted their BODYFAT based on DENSITY and analyzed each outliers in an order of decreasing residual, analyzing them orderly.



Prediction	96	48	76	182	216
|Bias|	15.7	7.9	4.0	2.1	0.0

96: Keep. Given 96 has the largest prediction bias, we deducted that either DENSITY or BODYFAT must be recorded wrong. Compared with observations sharing similar DENSITY (172), similar BODYFAT (126), similar BODYFAT and HEIGHT(109), 96 has more significant difference with 172, thus the DENSITY must be wrong. Without further verification, the distribution of 96 seems reasonable, thus we keep 96.

In [6]:
```
BodyFat[c(96, 172, 126, 109),c(2,3,4,5,6,7,10,17)]
```

	BODYFAT	DENSITY	AGE	WEIGHT	HEIGHT	ADIPOSITY	ABDOMEN	WRIST
96	17.3	1.0991	53	224.50	77.75	26.1	99.2	20.4
172	1.9	1.0983	35	125.75	65.50	20.6	75.0	16.9
126	17.4	1.0587	46	167.00	67.00	26.2	89.9	17.6
109	17.2	1.0593	43	194.00	75.50	24.0	88.7	19.2

48 & 76 : Delete. 48 and 76 have similar DENSITY, but significantly different BODYFAT. Comparing them with observations sharing similar DENSITY (146 & 90), all indexes except for BODYFAT are reasonable, for which we deducted the BODYFAT records are wrong. Though we could have estimated their BODYFAT as 14.3, as the first criterion mentioned before, we do not impute response variable. Thus, we delete 48 and 76 based on the 1st and 2nd criteria.

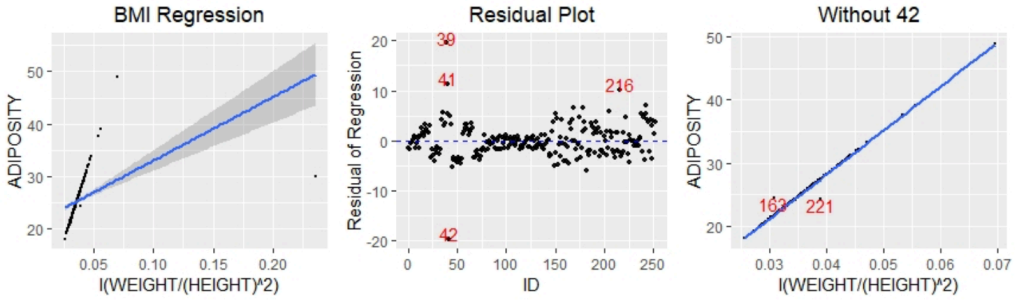
In [7]:
```
BodyFat[order(BodyFat$DENSITY)[173:176],c(2,3,4,5,6,7,10,17)]
```

	BODYFAT	DENSITY	AGE	WEIGHT	HEIGHT	ADIPOSITY	ABDOMEN	WRIST
146	14.4	1.0664	24	156.00	70.75	21.9	81.9	17.3
48	6.4	1.0665	39	148.50	71.25	20.6	79.5	17.9
76	18.3	1.0666	61	148.25	67.50	22.9	81.8	18.3
90	14.3	1.0666	48	176.00	73.00	23.3	86.5	18.7

182 & 216: Delete. As the following figure shows, red points refer to 182, purple points refer to 216. Though 182 and 216 have not too large predition bias, their extreme BODYFAT, WEIGHT, and other indxes indicate their distributions are rather extreme and uncommon. It is unreasonable for a human to have zero or negative BODYFAT like 182, and it is also rare for people to have such serious overweight phenomenon. Thus, based on the 3rd criterion, we sacrifice the information provided by 182 and 216, for the privilege of fitting a more robust and accurate model for most people.

2.2 Check HEIGHT & WEIGHT Using ADIPOSITY ¶

As the formula for ADIPOSITY with HEIGHT and WEIGHT introduced before, we identified and corrected outliers for WEIGHT and HEIGHT based on ADIPOSITY. Using similar approach as the last section, we obtained a similar list of outliers with decreasing residuals: 39, 42, 163, and 221. Besides, 42 is an extreme high influential point as the following plot shows.



Prediction	39	42	163	221
|Bias|	0.0	135.6	3.0	2.8

39: Delete. Though the records are correct based on the prediction result for 39, he has the highest records for the following ten variables, which indicate this person as extreme overweight. Recalling our 3rd criterion, we exclude this point from constructing a robust model.

WEIGHT	ADIPOSITY	NECK	CHEST	ABDOMEN	HIP	THIGH	KNEE	BICEPS	WRIST

42: Impute HEIGHT. 42 is a significant high influential point, and his records on HEIGHT is unreasonable as 29.50 inches. The imputed HEIGHT is 69.48, which is reasonable, however, considering the minimum precision for HEIGHT is 0.25 inch, we corrected the 42's HEIGHT into 69.50 inches.

42	Record	Prediction	Imputation
HEIGHT	29.50	69.48	69.50

163: Delete. As we can see, 163 and 221 are significant outliers based on the redisual plot. By analyzing subsets with similar ADIPOSITY, 174's 24.3 and 84's 24.5, as 163's 24.4, we found the outlier could be interpreted as each following way: WEIGHT recorded higher, or HEIGHT recorded lower, or ADIPOSITY recorded lower. Since we are unable to identify the exact mistake and impute correctly, based on our 3rd criteria, we delete 163 from consideration.

In [3]:
```
BodyFat[ order(BodyFat$ADIPOSITY)[c(105,108,109)], c(5:7)]
```

	WEIGHT	HEIGHT	ADIPOSITY
174	176.25	71.50	24.3
163	184.25	68.75	24.4
84	170.75	70.00	24.5

221: Delete. Based on the similar approach, assuming a correct ADIPOSITY, 221 might have wrong WEIGHT records, lower than the real WEIGHT, compared with 163rd and 9th observations sharing similar ADIPOSITY. However, when we look at the subset with similar WEIGHT and assume a correct ADIPOSITY record, 221 s HEIGHT might be recorded wrong, higher than the real HEIGHT. Thus, without identifying a significant abnormal behaviour for all three indexes, ADIPOSITY, WEIGHT, and HEIGHT, but confirming one of the pretictive variable would be used in analysis is wrong, we delete 221 based on our 2nd criteria.

In [6]:
```
cbind( BodyFat[ order(BodyFat$ADIPOSITY)[c(108,111,112)], c(5:7)], BodyFat[ order(BodyFat$WEIGHT)[44:46],c(1,5:7)])
```

	WEIGHT	HEIGHT	ADIPOSITY	IDNO	WEIGHT	HEIGHT	ADIPOSITY
163	184.25	68.75	24.4	191	153.00	69.25	22.5
221	153.25	70.50	24.5	221	153.25	70.50	24.5
9	191.00	74.00	24.6	3	154.00	66.25	24.7

2.3 Data cleaning summary

In conclusion, for data cleaning. We exclude seven points, 39, 48, 76, 163, 182, 216, and 221 based on our three criteria. Besides we impute the HEIGHT of 42.

Extreme Values:

1. WEIGHT: Delete 39. 2. BODYFAT: Delete 182 and 216.

Incorrect Records:

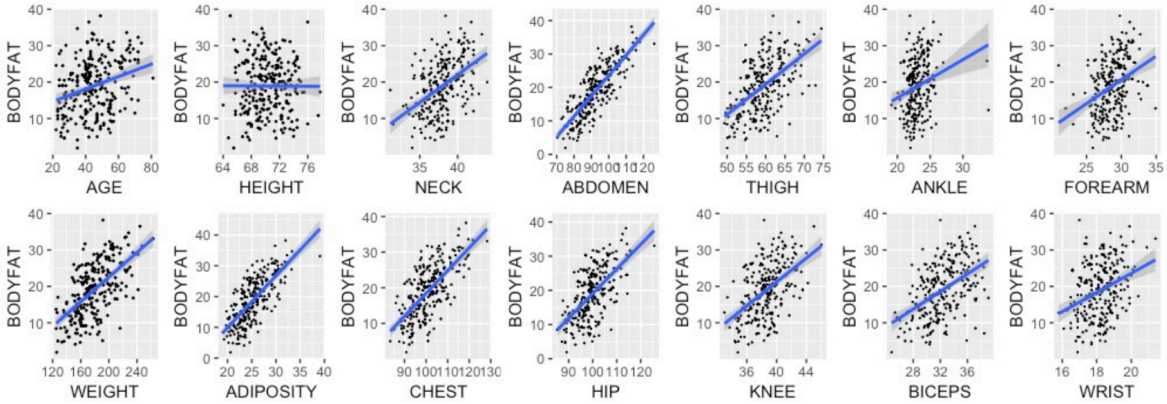
1. BODYFAT: Delete 48 and 76. 2. WEIGHT & HEIGHT: Delete 163 and 221.

Imputed Value:

1. HEIGHT: 42.

3 Variable Selection and Statistical Modeling

3.1 Tendency of variables



As we can see in this Figure, all the variables show the linear tendency, and some variable might have multicollinearity, so we try Lasso regression at first, we also use the Mallow's Cp, BIC forward and backward to select the variable.

3.2 Vaiable and Model Selection

Method	ABDOMEN	WRIST	HEIGHT	WEIGHT	AGE	R-squard
Lasso-1	0.50	0	0	0	0	0.641
Lasso-2	0.55	0	-0.18	0	0	0.671
Lasso-4	0.65	-1.14	-0.25	0	0.03	0.717
Lasso-all	-	-	-	-	-	0.739
Mallow's Cp-2	0.72	-2.05	0	0	0	0.704
Mallow's Cp-7	-	-	-	-	-	0.730
BIC forward-3	0.87	-1.34	0	-0.08	0	**0.72**
BIC forward-2	0.89	0	0	-0.12	0	**0.707**
BIC backward-3	0.71	-2.18	0	0	0.07	0.717

From the table, all the selected variables are reasonable
ABDOMEN: a direct reflection of body fat
WRIST and HEIGHT: an indicator of body frame
WEIGHT: refect body fat and muscle proportion

3.2.1 Lasso regression

As we can see in this form, we can not see the significant increase after we select **ABDOMEN** and the increase becomes much smaller after keeping the first four variables. All the model is reasonable. In Lasso-2, ABDOMEN is an indicator of body fat and HEIGHT is an indicator of the body frame. Usually, the man with larger body frame has the bigger abdomen, it also takes body frame into account. Apart from that, HEIGHT is a variable which can be easily obtained, because usually everyone knows their HEIGHT. In Lasso-4, WRIST and AGE also be considered as a factor as well.

3.2.2 Mallow's Cp

Mallow's Cp select 7 variables, the R-square is 0.73. for simplicity, we only select the most significant 2 variables, **ABDOMEN** and **WRIST**, which has the same explanation of Lasso-2.

3.2.3 BIC

We use BIC forward and backward to select the variable when compared with AIC, BIC usually tends to keep the most significant variable. Then, we try the combination of the 3 variable, because we try to keep less variable.

3.3 Model summary

Model

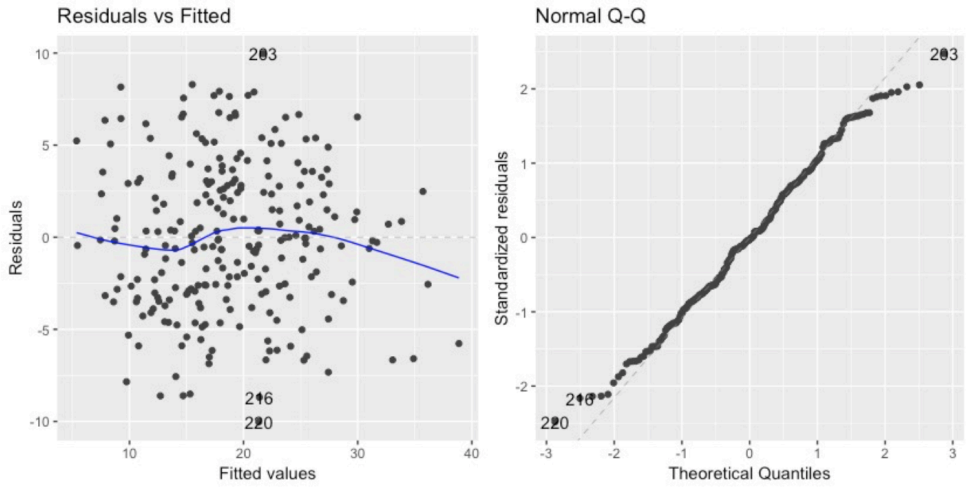
We use R-squard to select model, BIC forward-3 and BIC forward-2 give a good proformence. We choose **BIC forward-2** because weight is common but not everyone know their WRIST, it needs take time and cost to measure it. Additionally, the incease of R-squard is only 0.013, which is small, so we delete the variable WRIST.

Explanation

In BIC forward-2, the ABDOMEN and WEIGHT have a negative coefficient. In this Figure, we can see the person in the left side has a big ABDOMEN, admittedly, his WEIGHT is also big, but the density of muscle is larger than the density of fat, with the large body fat instead of muscle, the increase of weight is small compared with the increase of ABDOMEN. We can get a big body fat of this person. For the person on the right side, the ABDOMEN is not small, because ABDOMEN almost contains muscle. As a result, the increase in weight is bigger than the increase of ABDOMEN, we can get a small body fat of this person.

3.4 Model diagnostics

After model fitting, We diagnose the Model assumptions with a residual plot and a QQ plot.



There is no absolute pattern in the residual plot which means that the model can achieve the assumption of independent.
In the QQ-plot, there is almost a line, so the assumption of normality also can be satisfied.

4 Conclusion

Possible rule of thumb: "multiply your abdomen (cm) by 0.9 , subtract your weight (kg) by 0.3, and subtract 41

$$Percentage of Body Fat = -41(Constant) + 0.9ABDOMEN(cm) - 0.3WEIGHT(kg)$$

This rule of thumb is close to our estimated model by constructing 95% confidence intervals of the slope and the intercept.

4.1 Advantage & Disadvantage

Our model has the following advantages:

1. **Simplicity**: Aiming at building a simple model, our final model is convenient to remember and apply into daily life. Abdomen circumference and weight
2. **Robustness**: Based on information from various kinds of models, the final model is made more accurate and robust.
3. **Interpretability**: Based on common sense, the larger the abdomen circumferen, the larger the size of a person, and based on same body size, the larger the weight is, the more muscle a person owns, and the less body fat, for which our final model could be interpreted in consistency with common sense.

However, in a trade off between simplicity and accuracy, we prefer the former one. Thus, the accuracy might be sacrificed according to the parsimonious selection of variables, decisive deletion of outliers, and the simplification of the final coefficients.

5 Contribution

Kangyi Zhao: Build the Lasso and Mallow's Cp prediction models; contribute to the report introduction, and summary the report.

Xinyu Zhang: Clear up the code and images; contribute to data cleaning and BIC model selection of the report.

Kehui Yao: Make the tidy ggplot images; contribute to the presentation slides and web based app.