Bodyfat Analysis

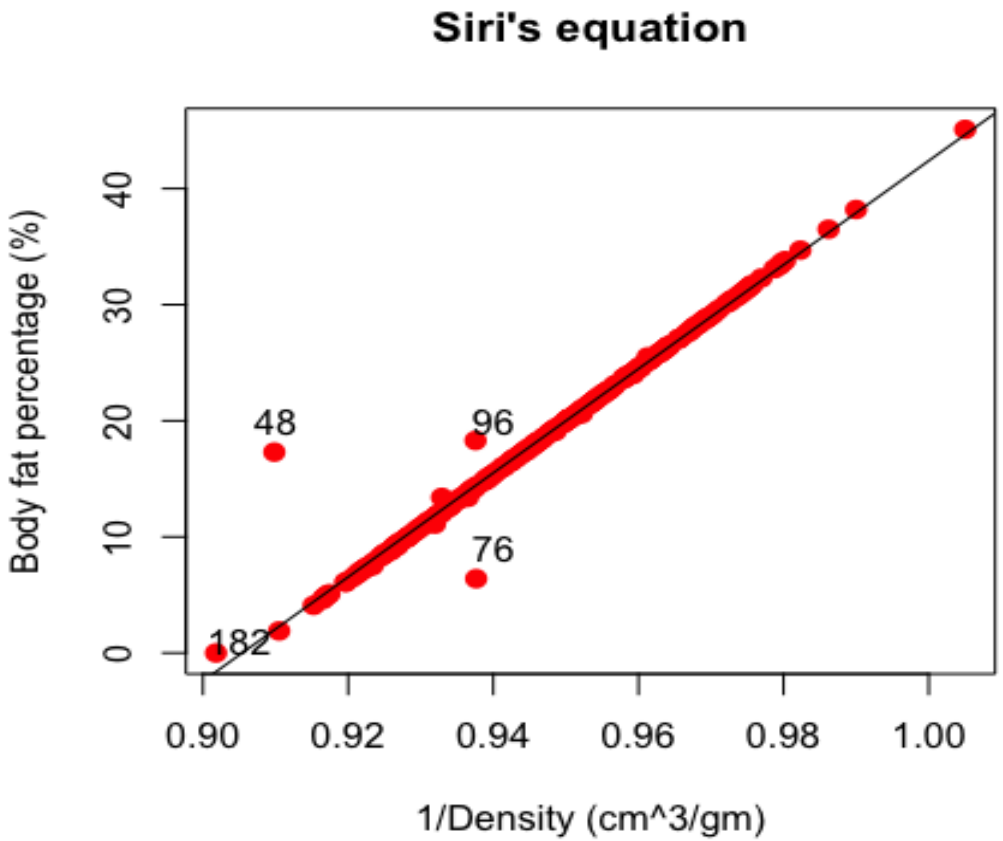
Group1:Yunwen Jiang, Luwei Liang, Jingyu Ji

Feb.7 2019

Data Clean

1.Check body fat records

$$Percentagebodyfat(i.e.100 * BODYFAT) = 495/DENSITY - 450 \quad (1)$$



Data Clean

Suspect that this is due to wrong calculation.So we compute body fat through siri's equation. The result is :

	BODYFAT	DENSITY	AGE	WEIGHT	HEIGHT	ADIPOSITY	NECK	C
48	6.4	1.0665	39	148.50	71.25	20.6	34.6	
76	18.3	1.0666	61	148.25	67.50	22.9	36.0	
96	17.3	1.0991	53	224.50	77.75	26.1	41.1	
182	0.0	1.1089	40	118.50	68.00	18.1	33.8	

$$\text{48th: } BODYFAT = 495/1.0665 - 450 = 14.14$$
$$\text{76th: } BODYFAT = 495/1.0666 - 450 = 14.09$$
$$\text{96th: } BODYFAT = 495/1.0991 - 450 = 0.37$$
$$\text{182nd: } BODYFAT = 495/1.1089 - 450 = -3.61$$

The result of 182nd observation is negative!

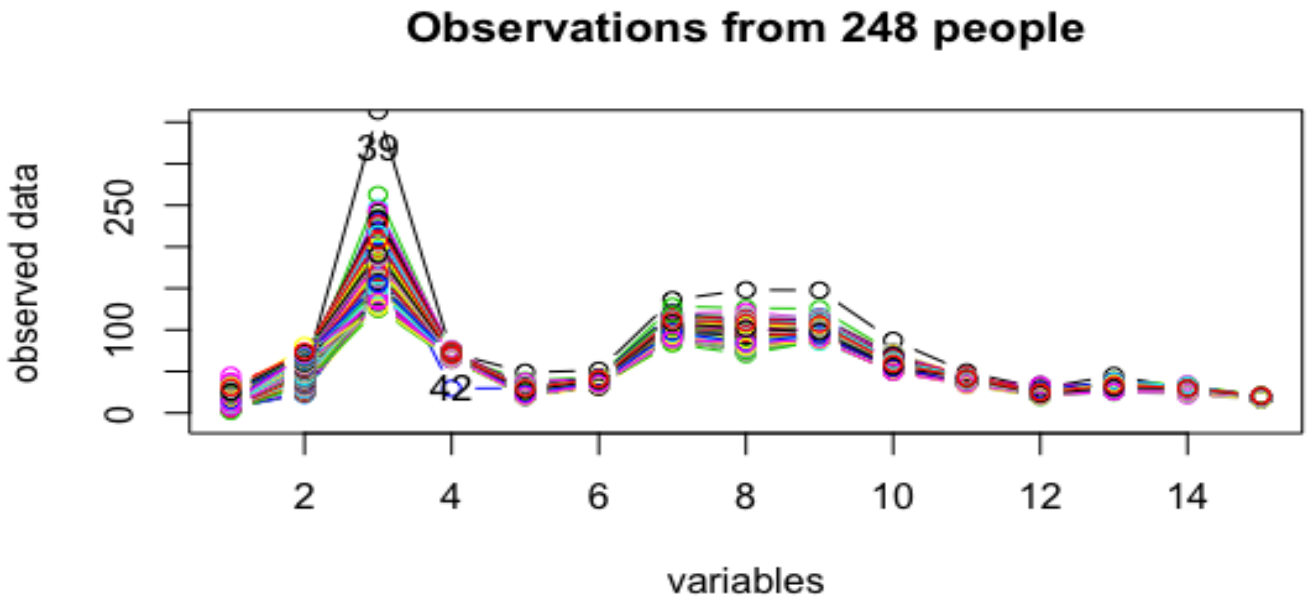
Possibility of wrong density record!

No evidence to tell which one is wrong: density or bodyfat? So delete them!

Data Clean

2.'Stick out' points

For 39th point:



Value of most variables of 39th point is extremely large.But BODYFAT of it is just 33.8.

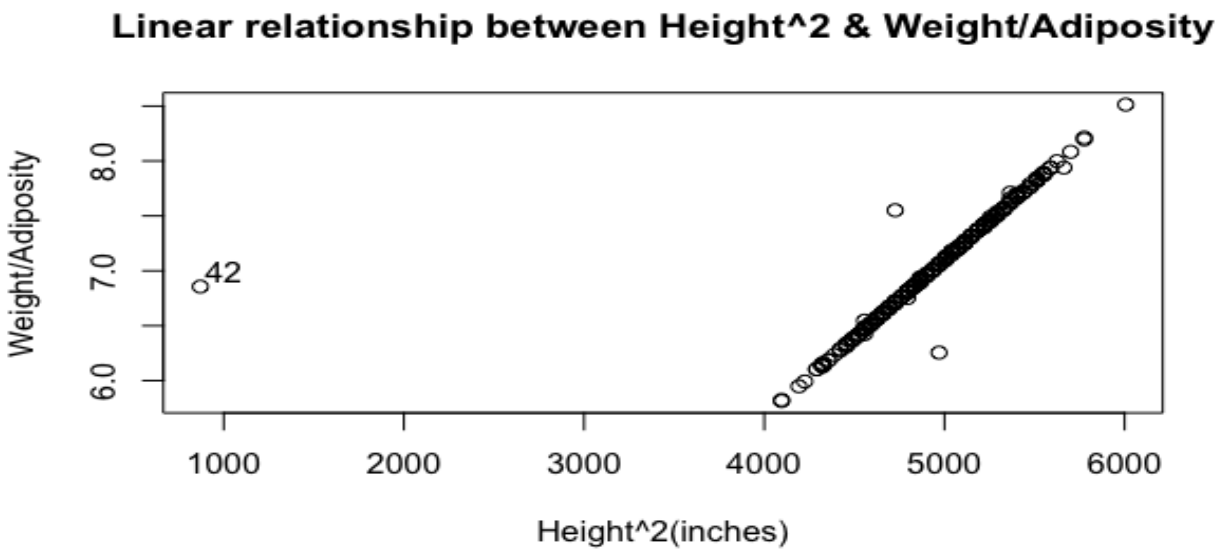
Unnormal,possible to cause high influence on the model.Delete it!

Data Clean

For 42nd point:

It has extremely short height: 29.5 while other variables are in normal range.

Suspect that it is wrongly recorded.Use linear relationship between $height^2$ and $weight/adiposity$ to check.



Choose to recalculate!

Multicollinearity Check

There are too many variables, also severe multicollinearity problem exists, which means that variable selection is needed:

AGE	WEIGHT	HEIGHT	ADIPOSITY	NECK	CHEST	ABDOMEN	HIP
2.259946	123.087875	27.628498	92.478830	3.859629	11.015050	12.252360	12.161626
THIGH	KNEE	ANKLE	BICEPS	FOREARM	WRIST		
7.145256	4.369946	1.852809	3.399834	2.393246	3.178639		

Figure: VIF of each variable

Variable Selection

We use four basic criteria: Mallow's Cp, adjusted R^2, stepwise AIC and stepwise BIC to narrow our variable selection.

Method	Number	Selected variables
Mallow's Cp	6	Age, Height, Chest, Abdomen, Biceps, Wrist
Adjusted R^2	9	Age, Adiposity, Neck, Chest, Abdomen, Hip, Thigh, Forearm, Wrist
Stepwise AIC	4	Abdomen, Weight, Wrist, Biceps
Stepwise BIC	3	Abdomen, Weight, Wrist

Variable Selection

Then we use cross validation to help us judge the performance of these four 'best' models based on RMSE.
We try 1000 times for each model. Each time we select 200 observations out of 247 as the train dataset, and the remaining 47 ones as the test dataset. Then we rebuild the linear regression models using the variables selected by each criterion.

Method	RMSE	Selected variables
Mallow's Cp	4.0312	Age, Height, Chest, Abdomen, Biceps, Wrist
Adjusted R^2	4.0319	Age, Adiposity, Neck, Chest, Abdomen, Hip, Thigh, Forearm, Wrist
Stepwise AIC	4.0241	Abdomen, Weight, Wrist, Biceps
Stepwise BIC	4.0431	Abdomen, Weight, Wrist

Model Building

The BIC method choose these three variables: WEIGHT, ABDOMEN, and WRIST.

(Bodyfat%)=-23.04+ 0.88 Abdomen(cm)-0.08Weight(lbs)-1.36WRIST(cm)

After some intense discussion, we decide to delete the variable WRIST based on the following reasons.

Model Building

1) **ABDOMEN** is a really important variable for this dataset.

- For single simple regression $BODYFAT \sim ABDOMEN$. The R^2 could be around 0.7, which is extremely high for a single variable.
- ABDOMEN appears in all four selected models with large coefficients and broad range at the same time, which results in huge influence on the final result.

Model Building

$$\hat{\beta}_j \cdot [Q_3(x_j) - Q_1(x_j)]$$

Variable	Result
Abdomen	12.54
Weight	-3.11
Wrist	-1.63

ABDOMEN is definitely of decisive importance of the whole dataset, all the other variables are just playing the role of cooperation, adjustment and optimization.
However, it is hard to explain the cooperation relationship between wrist and abdomen. It is also hard to explain the relationship between wrist and bodyfat. **Wrist** is less informative and intutively understandable to users.

Model Building

2) The linear model with variables ABDOMEN+WEIGHT gets relatively larger R^2 than the one with ABDOMEN+WRIST.

3) The linear model with variables ABDOMEN+WEIGHT still performed well in CV based on RMSE:4.1038.

Variable	Result
ABDOMEN+WEIGHT	0.71744
ABDOMEN+WRIST	0.7148

Model Interpretation

The final model includes two variables: Abdomen and Weight.

$$(Bodyfat\%) = -42.4587 + 0.8982 Abdomen(cm) - 0.1210 Weight(lbs)$$

- In general cases, men's body fat percent is determined by the abdomen circumference. The larger your abdomen circumference is, you'll get higher percentage of body fat.
- For people with the same abdomen circumference, the heavier people will get a lower percent of bodyfat.

Model Summary

$$(Bodyfat\%) = -42.4587 + 0.8982 Abdomen(cm) - 0.1210 Weight(lbs)$$

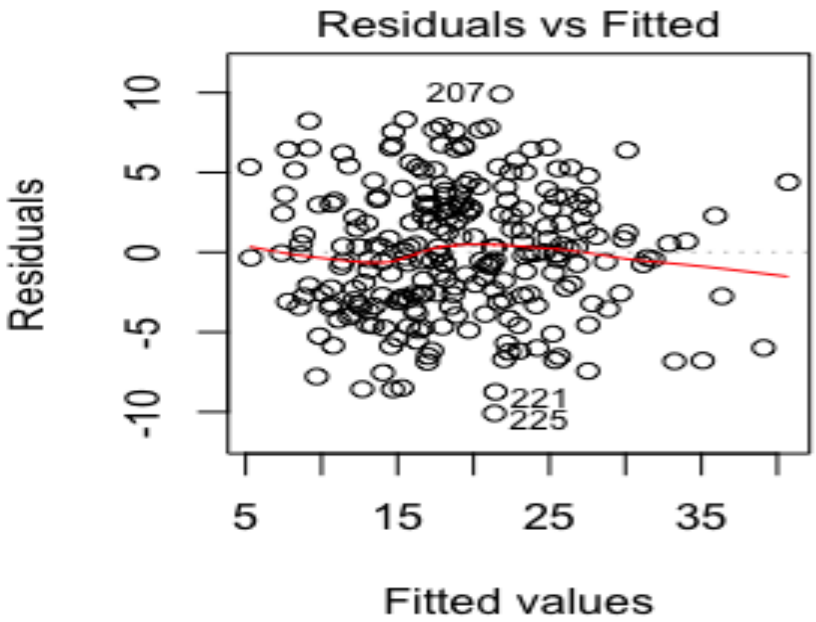
	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	-42.45868	2.47043	-17.187	$< 2e-16$
WEIGHT	-0.12103	0.01979	-6.116	$3.79e-09$
ABDOMEN	0.89817	0.05215	17.223	$< 2e-16$

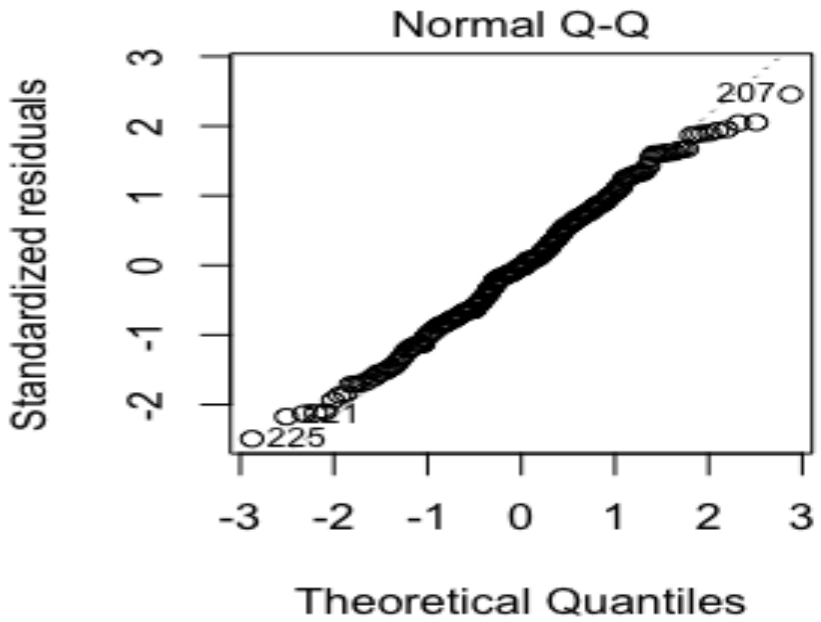
Residual standard error: 4.058

Adjusted R-squared: 0.7174

p-value: $< 2.2e-16$

Model Diagnostic






```
        Shapiro-Wilk normality test

data:  fat3_lm_wa$residuals
W = 0.99211, p-value = 0.2092
```

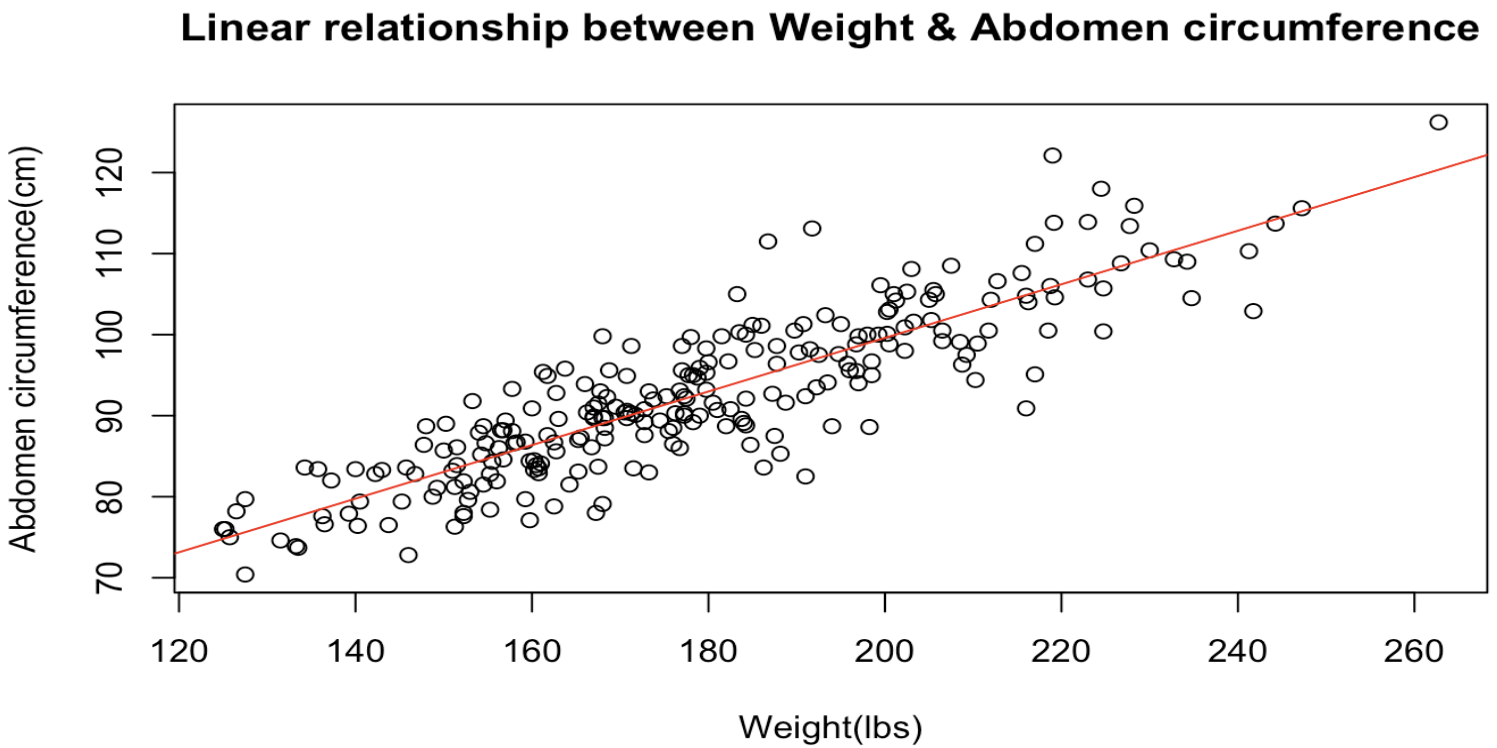
Figure: normality test

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.2549745    Df = 1     p = 0.6135939
```

Figure: homoskedasticity test

Model Diagnostic

We also find there's actually a linear relationship between weight and abdomen(but not enough to cause a colliearity problem with VIF=4.166731), which indicates that if you lose your weight by your effort, it's highly likely to reduce your abdomen circumference at the same time.



Discussion

Strengths

- Linearity: Our model follows the linear regrssion assumptions, no normality and homoskedasticity violations.
- Simplicity: Our model is simple to understand and implement. All variables we used are easy to measure.

Weakness

- Precision: The precision of the prediction is not so high. The model is not suitable for business prediction.