Body Fat Data Analysis

Jiawen Chen, Chunyuan Jin, and Han Liao

10/8/2019

Overview

1. Introduction
2. Data Cleaning
   - Check Extreme Values with Summary Table
   - Check with Siri's Equation
   - Check with BMI Formula
   - Summary of Data Cleaning
3. Variable Selection and Statistical Modeling
   - Stepwise Linear Regression
   - Variable Selection with XGBoost
   - Stable Statistical Model
   - Model Diagnostics
4. Conclusion
5. Web-Based App

Introduction

- Body fat percentage, a measure of fitness level
- Data set

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
3	24.6	1.0414	22	154.00	66.25	24.7	34.0	95.8	87.9	99.2	59.6	38.9	24.0	28.8	25.2	16.6
4	10.9	1.0751	26	184.75	72.25	24.9	37.4	101.8	86.4	101.2	60.1	37.3	22.8	32.4	29.4	18.2
5	27.8	1.0340	24	184.25	71.25	25.6	34.4	97.3	100.0	101.9	63.2	42.2	24.0	32.2	27.7	17.7
6	20.6	1.0502	24	210.25	74.75	26.5	39.0	104.5	94.4	107.8	66.0	42.0	25.6	35.7	30.6	18.8

- **252** observations
- Response variable (BODYFAT) with a redundant variable (DENSITY)
- **14** predictive variables such as AGE, WEIGHT, etc.

Check Extreme Values with Summary Table

- Summary part of the dataset

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

- Some records which have abnormal values

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

Check with Siri's Equation

- The Siri's equation:

$$Body\ Fat\ \%\ (i.e. 100 \times B) = \frac{495}{D} - 450,$$

$$D\ is\ the\ Body\ Density\ (gm/cm^3)$$

- Records that are not aligned with the logic of Siri's Equation

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

```
> 495/BodyFatData$DENSITY[48] - 450
[1] 14.13502
> 495/BodyFatData$DENSITY[76] - 450
[1] 14.09151
> 495/BodyFatData$DENSITY[96] - 450
[1] 0.3684833
```

Check with BMI Formula

- BMI's formula:

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

- Records that are not aligned with the logic of BMI Formula

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

```
> (703*BodyFatData$WEIGHT[163])/(BodyFatData$HEIGHT[163])^2
[1] 27.40422
> (703*BodyFatData$WEIGHT[220])/(BodyFatData$HEIGHT[220])^2
[1] 22.64875
> (703*BodyFatData$WEIGHT[234])/(BodyFatData$HEIGHT[234])^2
[1] 25.14288
```

Summary of Data Cleaning

- 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 adjusted according to the WEIGHT and ADIPOSITY.
- The BODYFAT of record **48** and **76** are adjusted according to the DENSITY.
- The ADIPOSITY of record **163**, **220**, and **234** are adjusted according to the WEIGHT and HEIGHT.

Stepwise Linear Regression

- use BIC as criterion and directions of forward and backward

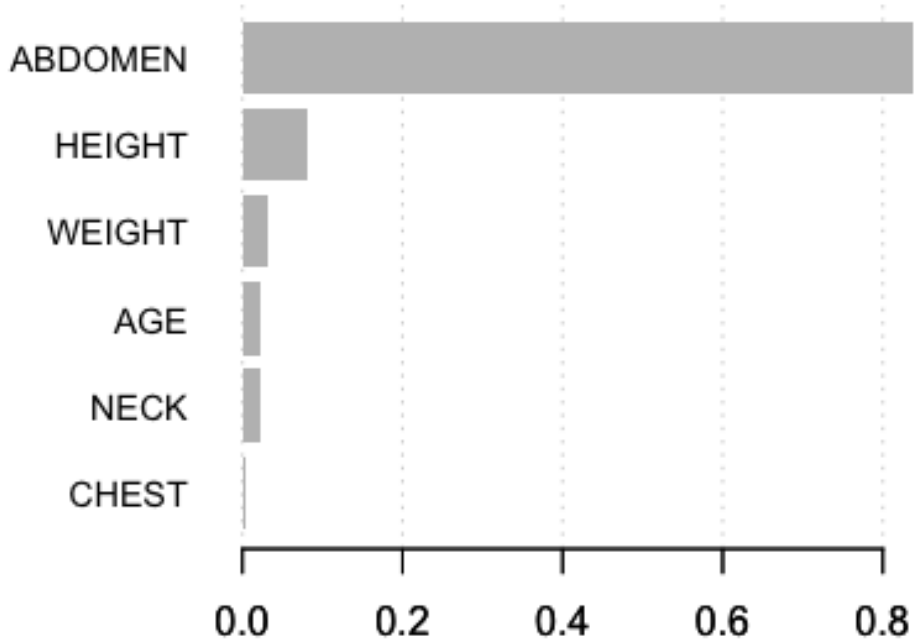
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -32.94538    6.78196  -4.858 2.12e-06 ***
ABDOMEN       0.92637    0.05166  17.933  < 2e-16 ***
WEIGHT       -0.13393    0.02302  -5.818 1.87e-08 ***
WRIST        -1.26352    0.41003  -3.082  0.00230 **
FOREARM       0.46142    0.16690   2.765  0.00613 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.966 on 244 degrees of freedom
Multiple R-squared:  0.736,     Adjusted R-squared:  0.7317
F-statistic: 170.1 on 4 and 244 DF,  p-value: < 2.2e-16
```

- VIF values to check collinearity

ABDOMEN	WEIGHT	WRIST	FOREARM
4.860632	7.159185	2.278900	1.776857

Variable Selection with XGBoost

- XGBoost is an ensemble learning method with tree models. It will output the importance of variables after establishing the model. This nonlinear method can be used to do variable selection.



Stable Statistical Model

- To make our model both stable and accurate, we only choose two predictors to build normal linear model.
- ABDOMEN is the most important variable among 14 predictors. Then, we try to add another variable among WRIST, WEIGHT, ADIPOSITY.
- the model with highest R-square 0.72

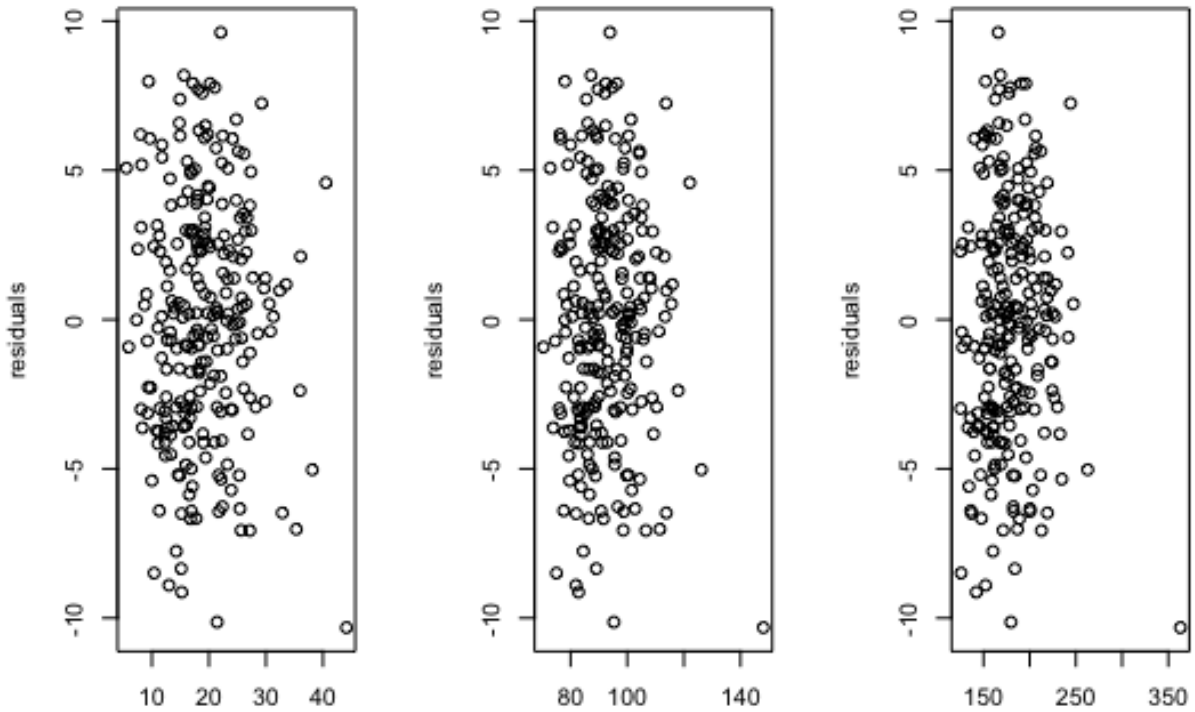
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -40.77528    2.39583 -17.019  < 2e-16 ***
ABDOMEN       0.91797    0.05203  17.642  < 2e-16 ***
WEIGHT       -0.14060    0.01911  -7.358 2.78e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.063 on 246 degrees of freedom
Multiple R-squared:  0.7206,    Adjusted R-squared:  0.7183
F-statistic: 317.2 on 2 and 246 DF,  p-value: < 2.2e-16
```

Model Diagnostics

- Firstly, use VIF to check collinearity.

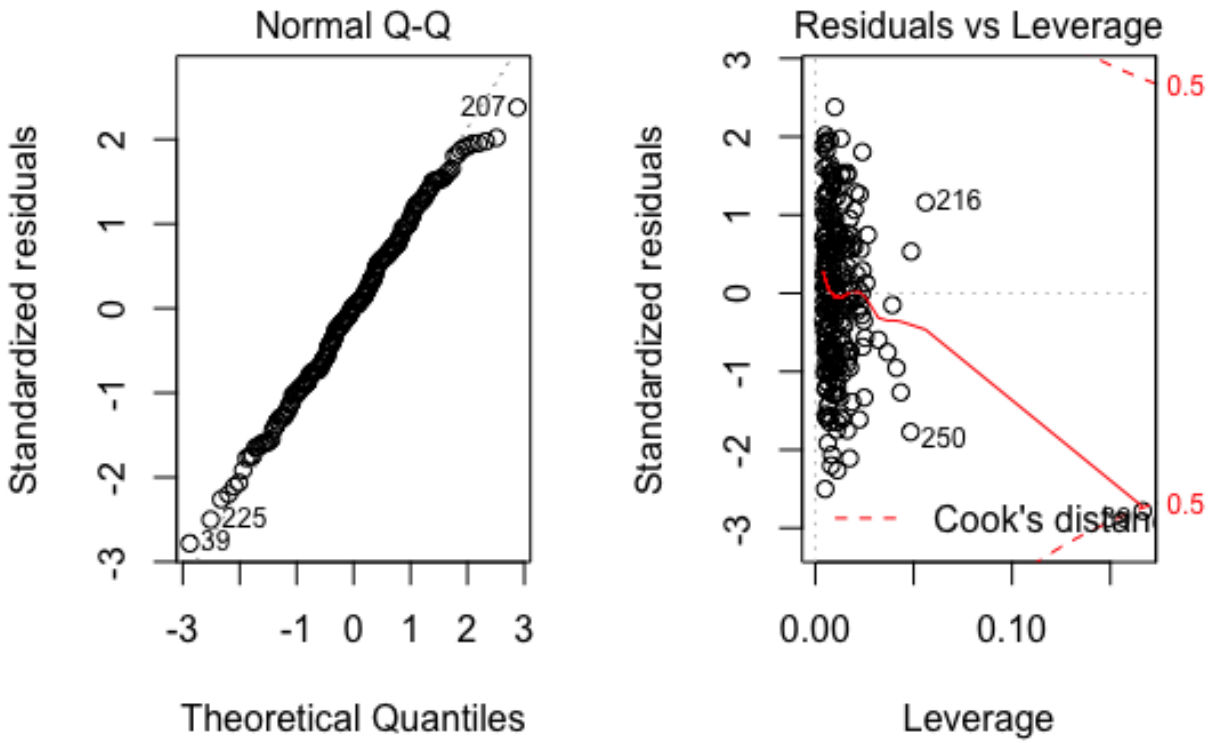
ABDOMEN	WEIGHT
4.698426	4.698426

- Then, plot the residuals vs fitted values. There is no correlation between residuals and predictors.



Model Diagnostics

- QQ-plot of residuals show that residuals generally follows normal distribution.
- It is normal that there are still some outliers in the model.



Conclusion

- Final Model:

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

- 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 percentage of the variation in body fat among men although. It also has weaknesses as it cannot capture higher order effects and interactions.

Web-Based App

https://jiawen1014.shinyapps.io/BodyFat/