STAT 628 Group 5

Body Fat Dataset

NAIQING CAI
YUHANG LAN
ZIHAO LI
XINKAI CHEN

Overview

Our project focuses on the measure of percentage of body fat.

In this module, we will come up with a simple, robust, accurate and precise “rule-of-thumb” method to estimate percentage of body fat.

CONTENTS

Data Preprocessing

Feature Selection

Model Selection & Evaluation

Application

Data Preprocessing

- Data Structure
- Data Correlation
- Data Cleaning

Data Preprocessing

A real data set of 252 men with measurements of their percentage of body fat and various body circumference measurements.

Data Structure

	IDNO	BODYFAT	DENSITY	AGE	WEIGHT	HEIGHT	ADIPOSITY	NECK	CHEST	ABDOMEN	HIP	THIGH	KNEE	ANKLE	BICEPS	FOREARM	WRIST
1	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	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	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	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	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	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

Data Preprocessing

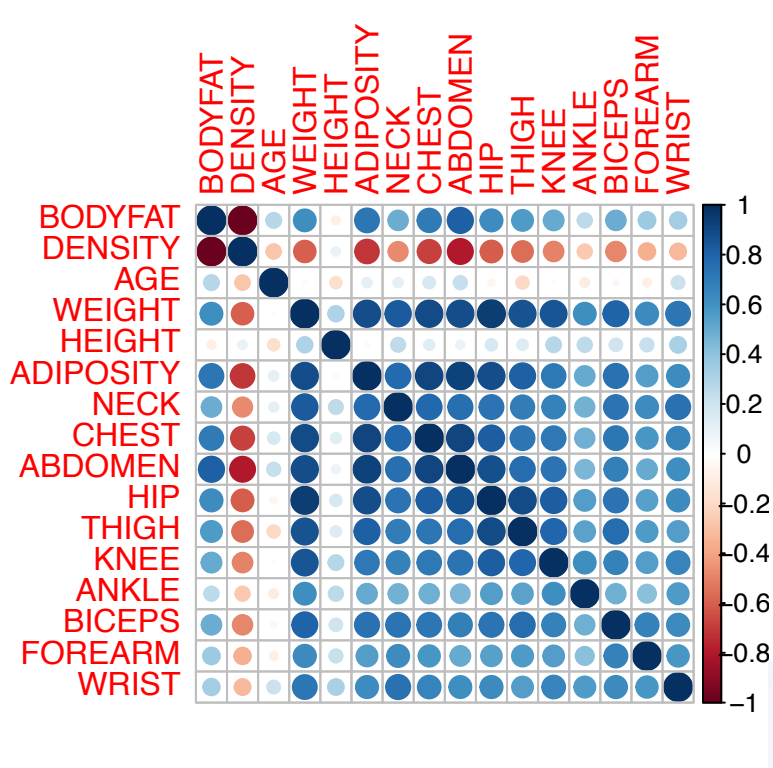
BODYFAT	DENSITY	AGE	WEIGHT	HEIGHT	ADIPOSITY
Min. : 0.00	Min. :0.995	Min. :22.00	Min. :118.5	Min. :29.50	Min. :18.10
1st Qu.:12.80	1st Qu.:1.041	1st Qu.:35.75	1st Qu.:159.0	1st Qu.:68.25	1st Qu.:23.10
Median :19.00	Median :1.055	Median :43.00	Median :176.5	Median :70.00	Median :25.05
Mean :18.94	Mean :1.056	Mean :44.88	Mean :178.9	Mean :70.15	Mean :25.44
3rd Qu.:24.60	3rd Qu.:1.070	3rd Qu.:54.00	3rd Qu.:197.0	3rd Qu.:72.25	3rd Qu.:27.32
Max. :45.10	Max. :1.109	Max. :81.00	Max. :363.1	Max. :77.75	Max. :48.90

NECK	CHEST	ABDOMEN	HIP	THIGH	KNEE
Min. :31.10	Min. : 79.30	Min. : 69.40	Min. : 85.0	Min. :47.20	Min. :33.00
1st Qu.:36.40	1st Qu.: 94.35	1st Qu.: 84.58	1st Qu.: 95.5	1st Qu.:56.00	1st Qu.:36.98
Median :38.00	Median : 99.65	Median : 90.95	Median : 99.3	Median :59.00	Median :38.50
Mean :37.99	Mean :100.82	Mean : 92.56	Mean : 99.9	Mean :59.41	Mean :38.59
3rd Qu.:39.42	3rd Qu.:105.38	3rd Qu.: 99.33	3rd Qu.:103.5	3rd Qu.:62.35	3rd Qu.:39.92
Max. :51.20	Max. :136.20	Max. :148.10	Max. :147.7	Max. :87.30	Max. :49.10

ANKLE	BICEPS	FOREARM	WRIST
Min. :19.1	Min. :24.80	Min. :21.00	Min. :15.80
1st Qu.:22.0	1st Qu.:30.20	1st Qu.:27.30	1st Qu.:17.60
Median :22.8	Median :32.05	Median :28.70	Median :18.30
Mean :23.1	Mean :32.27	Mean :28.66	Mean :18.23
3rd Qu.:24.0	3rd Qu.:34.33	3rd Qu.:30.00	3rd Qu.:18.80
Max. :33.9	Max. :45.00	Max. :34.90	Max. :21.40

Data Preprocessing

Data Correlation



Data Preprocessing

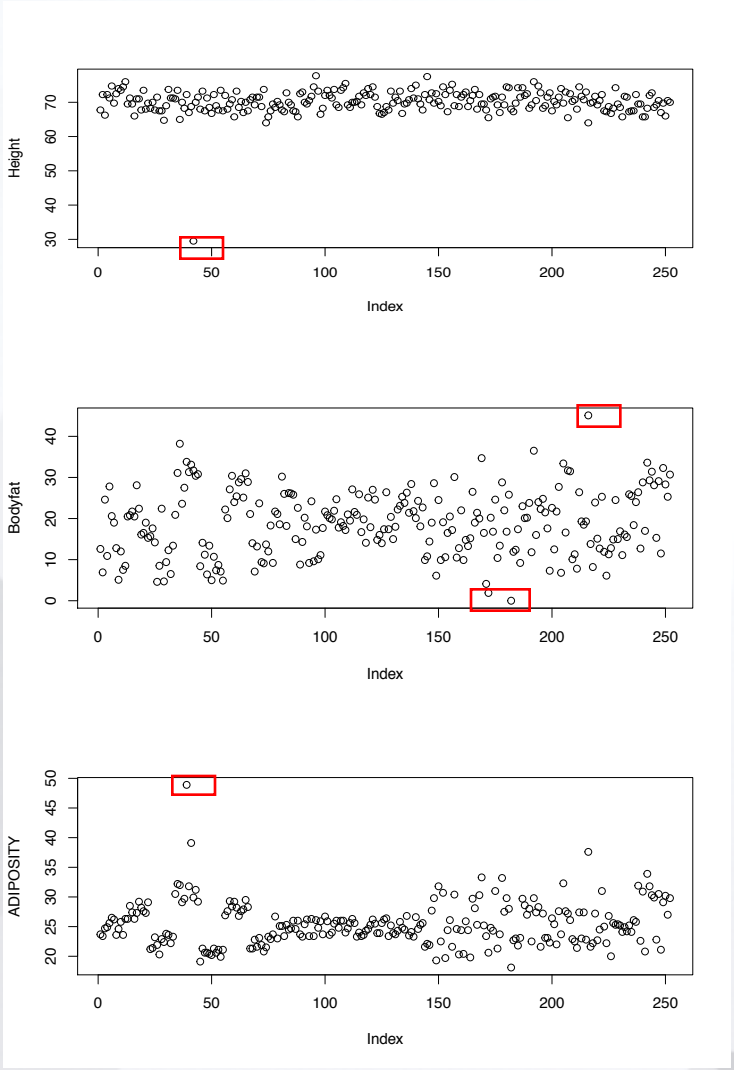
01 **Outlier Diagnosis**
Scatter Plot
Leverage
Cooks Distance

02 **Outlier Test**
DFFITS

Data Preprocessing

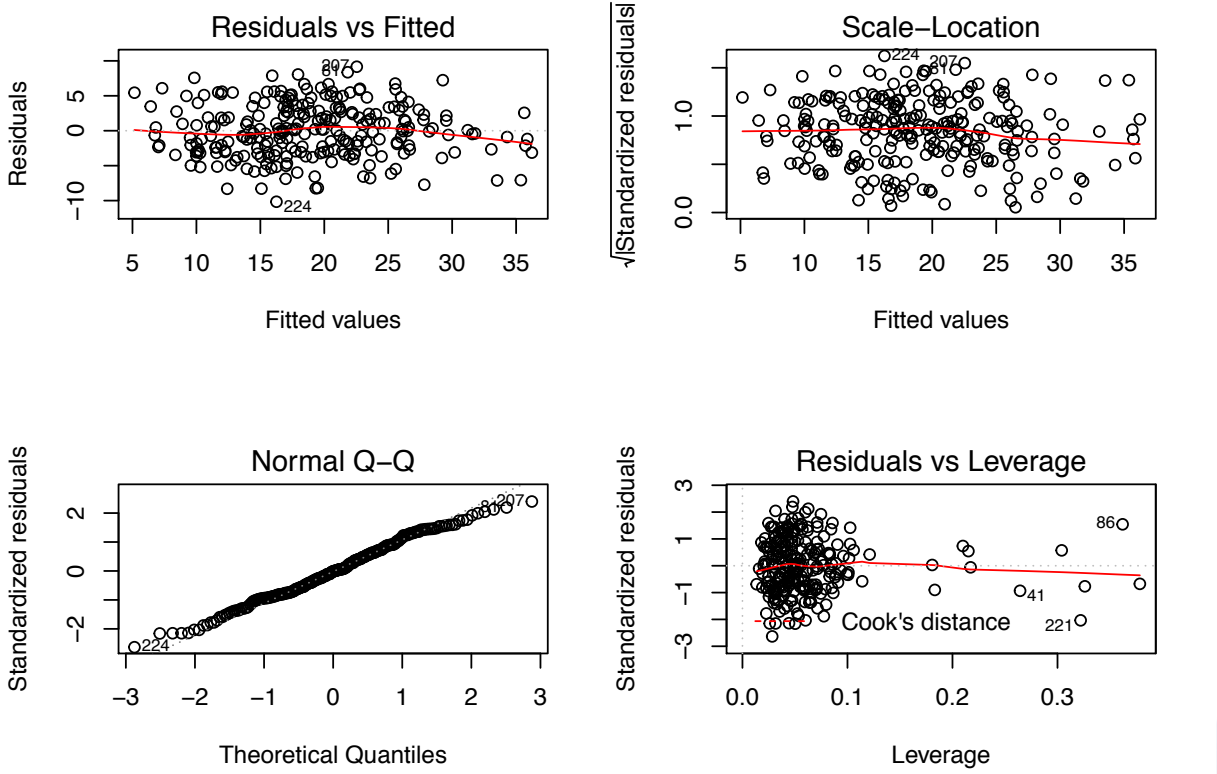
Scatter Plot
--Height
--Body Fat
--Adiposity

Smallest Height: 42
(we can recover it)
Two Smallest Body Fat: 172,182
Largest Body Fat: 216
(We need to delate them)
Largest Adiposity: 39
(We need to delate them)



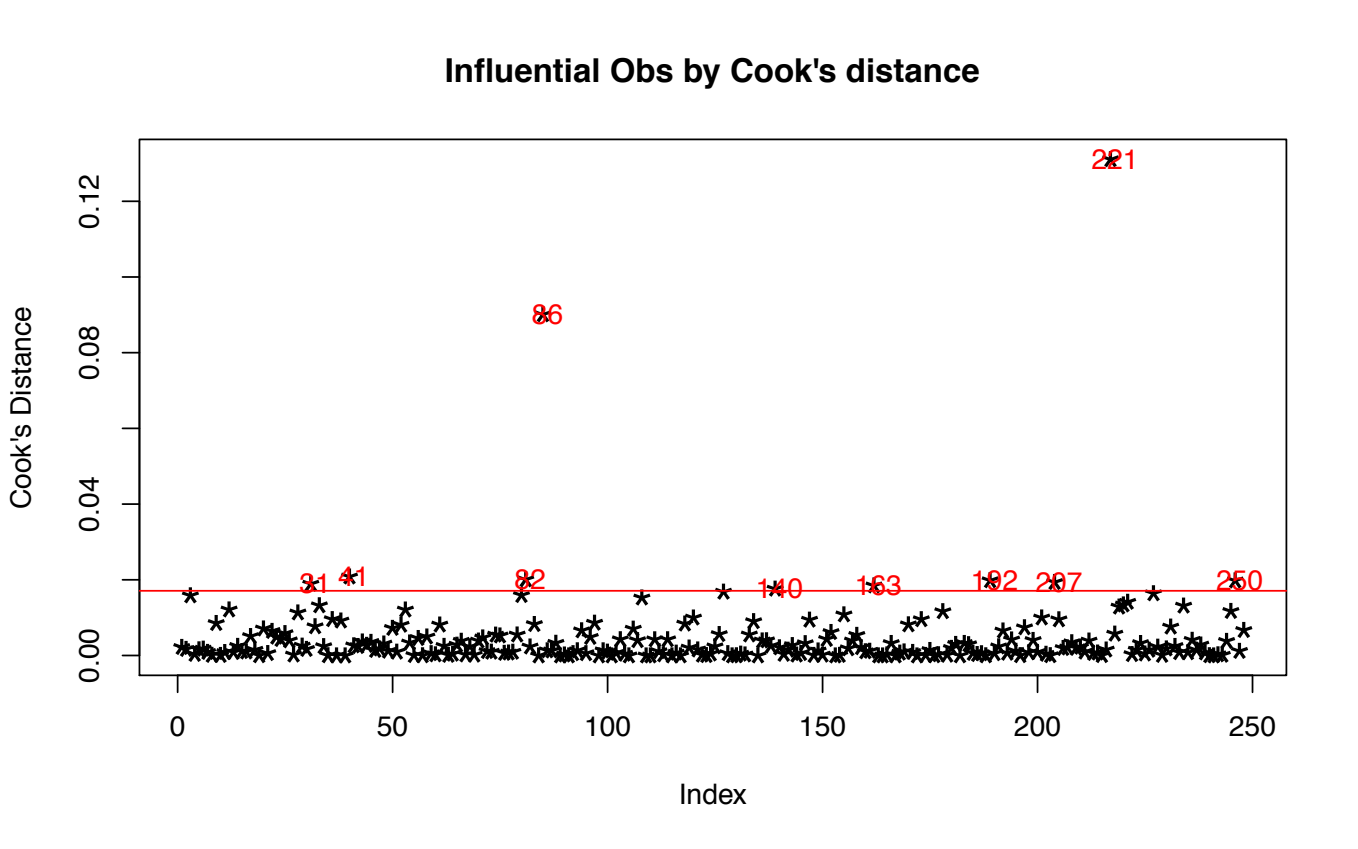
Data Preprocessing

Leverage —— 86, 221



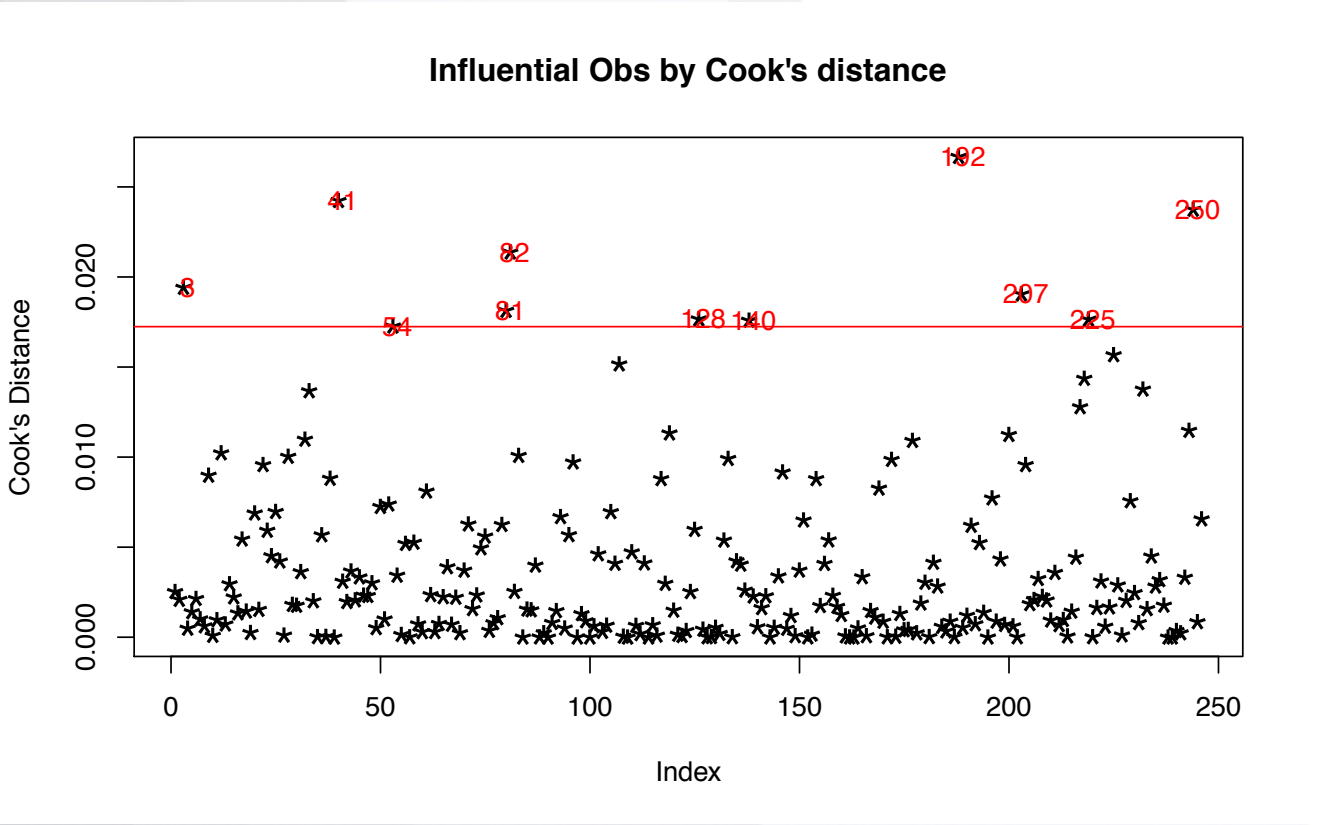
Data Preprocessing

Cooks Distance —— 86, 221



Data Preprocessing

Cooks Distance



Data Preprocessing

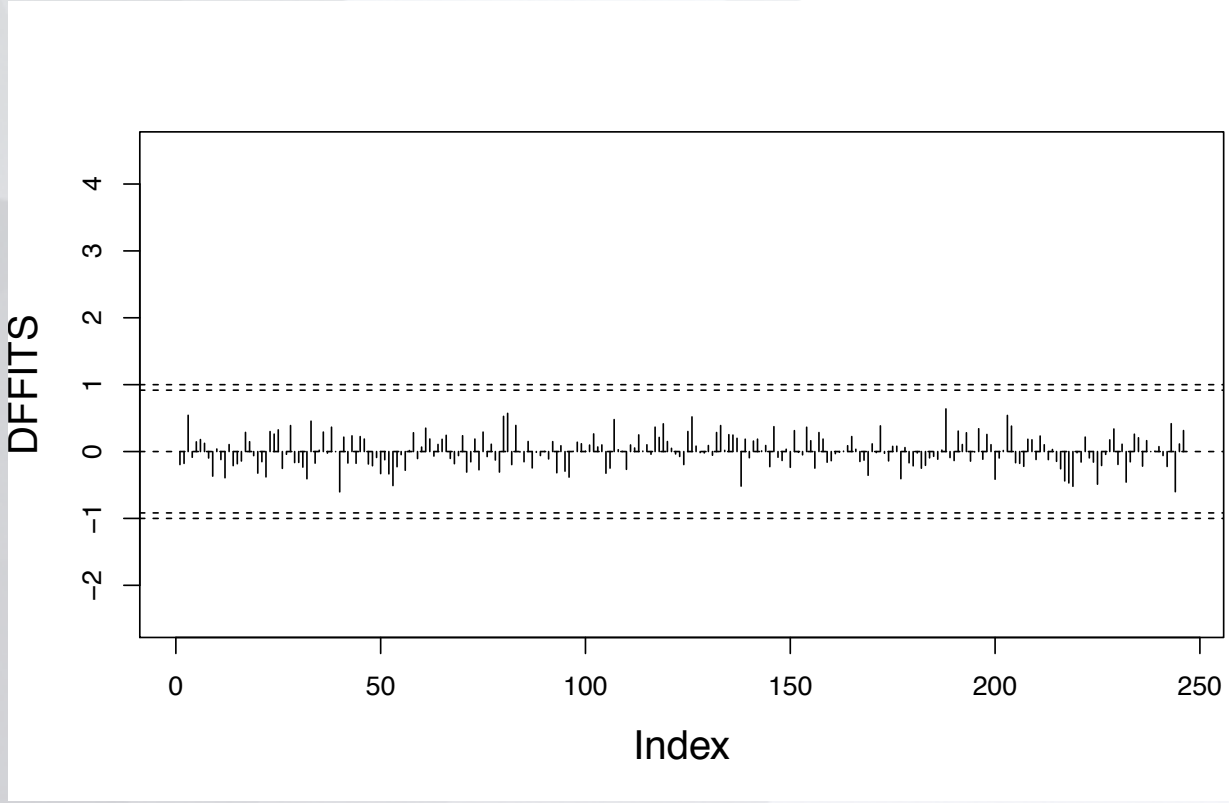
Outlier Test

> **library**(car)
> outlierTest(model)

```
No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
     rstudent unadjusted p-value Bonferonni p
224 -2.558296           0.011161           NA
```

Data Preprocessing

DFFITS



Feature Selection

- Eyeballing
- Mallow CP
- Adjusted R^2
- AIC
  - Forward
  - Backward
  - Stepwise
- BIC
  - Forward
  - Backward
  - Stepwise
- VIF

Feature Selection

Eyeballing

Abdomen, Wrist are most important

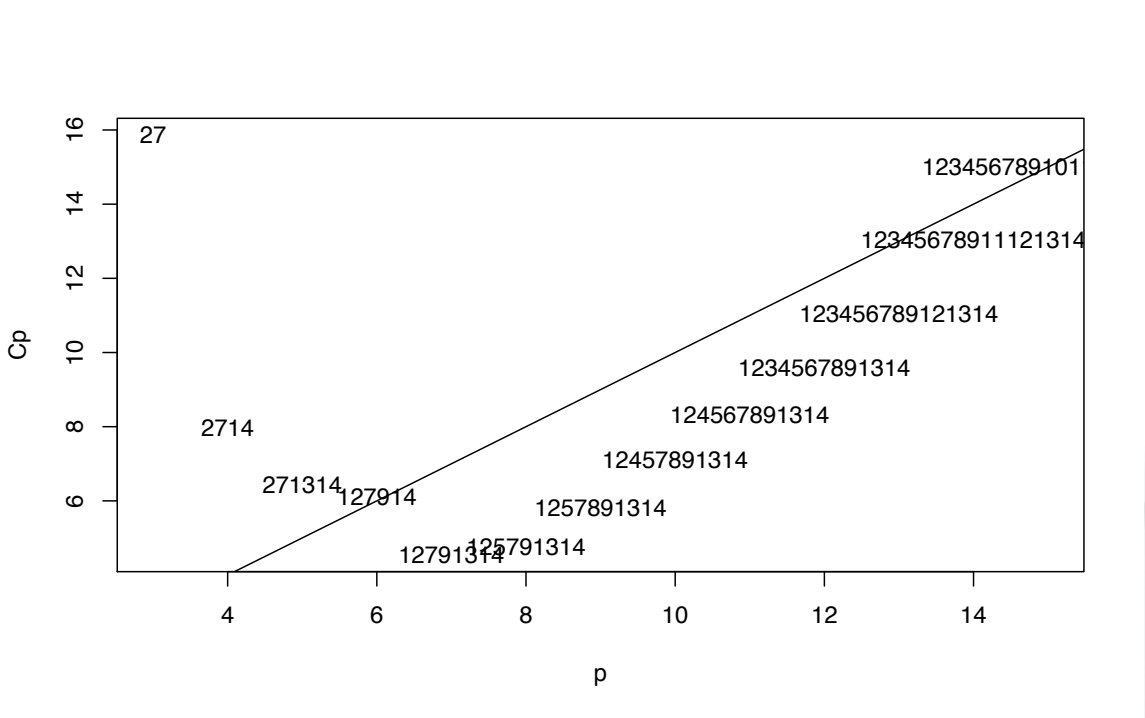
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -53.30060   45.44215  -1.173  0.24203
AGE           0.05515    0.02990   1.844  0.06642 .
WEIGHT       -0.17720    0.12721  -1.393  0.16495
HEIGHT        0.54173    0.61775   0.877  0.38143
ADIPOSITY     1.01268    0.86902   1.165  0.24509
NECK         -0.32802    0.21486  -1.527  0.12820
CHEST        -0.09944    0.10357  -0.960  0.33798
ABDOMEN       0.83355    0.08523   9.780  < 2e-16 ***
HIP          -0.16396    0.13373  -1.226  0.22143
THIGH         0.21225    0.13745   1.544  0.12391
KNEE          0.03994    0.22871   0.175  0.86153
ANKLE        -0.04475    0.24829  -0.180  0.85712
BICEPS        0.11376    0.15721   0.724  0.47003
FOREARM       0.28971    0.19109   1.516  0.13088
WRIST        -1.40790    0.49538  -2.842  0.00488 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.878 on 231 degrees of freedom
Multiple R-squared:  0.7415,    Adjusted R-squared:  0.7258
F-statistic: 47.32 on 14 and 231 DF,  p-value: < 2.2e-16
```

Feature Selection

Mallow's Cp

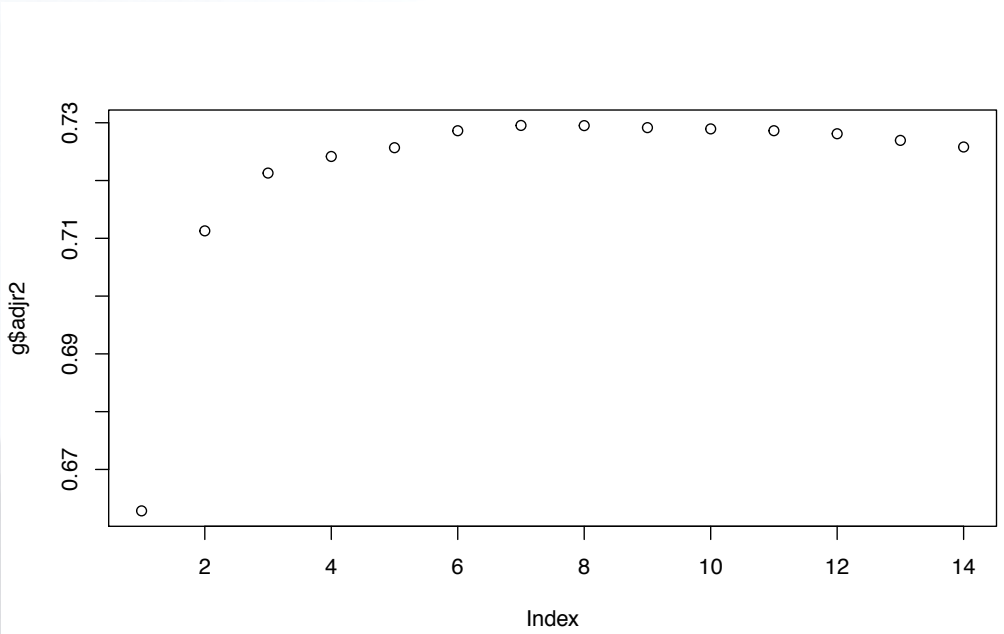
Age, Weight, Abdomen, Thigh, Wrist seem to be a good choice.



Feature Selection

Adjusted R^2

Age, Weight, Neck, Abdomen, Thigh, Forearm, Wrist seem to be a good choice.



	BODYFAT	AGE	WEIGHT	NECK	ABDOMEN	THIGH	FOREARM	WRIST
1	12.6	23	154.25	36.2	85.2	59.0	27.4	17.1
2	6.9	22	173.25	38.5	83.0	58.7	28.9	18.2
3	24.6	22	154.00	34.0	87.9	59.6	25.2	16.6
4	10.9	26	184.75	37.4	86.4	60.1	29.4	18.2
5	27.8	24	184.25	34.4	100.0	63.2	27.7	17.7
6	20.6	24	210.25	39.0	94.4	66.0	30.6	18.8

Feature Selection

AIC and BIC Backward

Weight, Abdomen, Wrist seem to be a good choice.

We also tried AIC, BIC Forward, Both, but they did not perform as well as BIC Backward.

AIC——Backward

```
Step:  AIC=671.2
BODYFAT ~ AGE + WEIGHT + ABDOMEN + THIGH + FOREARM + WRIST

          Df Sum of Sq    RSS    AIC
<none>                  3557.7 671.20
- FOREARM  1     53.60 3611.3 672.87
- AGE      1     54.36 3612.0 672.93
- THIGH    1     67.73 3625.4 673.83
- WRIST    1    178.38 3736.0 681.23
- WEIGHT   1    196.29 3754.0 682.41
- ABDOMEN  1   2369.70 5927.4 794.77
```

BIC——Backward

```
Step:  AIC=688.84
BODYFAT ~ WEIGHT + ABDOMEN + WRIST

          Df Sum of Sq    RSS    AIC
<none>                  3699.6 688.84
- WRIST    1     148.6 3848.2 693.02
- WEIGHT   1     253.9 3953.5 699.66
- ABDOMEN  1    4498.8 8198.4 879.08
```

Feature Selection

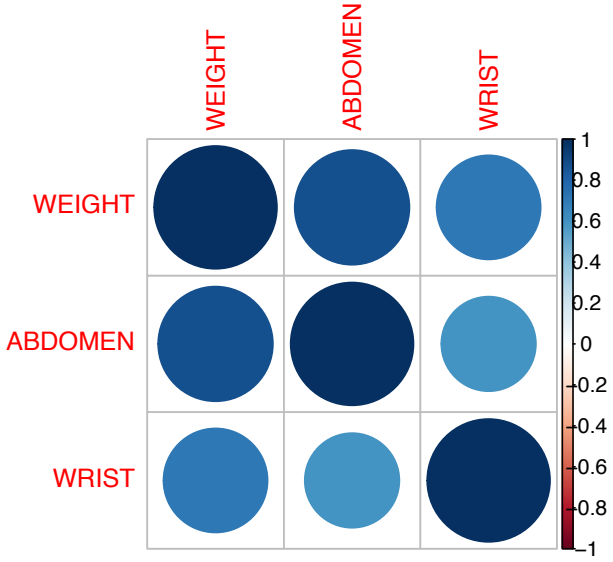
Check VIF

lm(BODYFAT ~ WEIGHT + ABDOMEN + WRIST, data)

1/(1-R^2)	3.63240101707229
WEIGHT	5.67699445648803
ABDOMEN	4.23747817633164
WRIST	2.08944656965687

lm(BODYFAT ~ ABDOMEN + WRIST, data)

1/(1-R^2)	3.39904826648538
ABDOMEN	1.5443138433835
WRIST	1.5443138433835



Model Selection & Evaluation

- R^2
- MSE
- Diagnosis

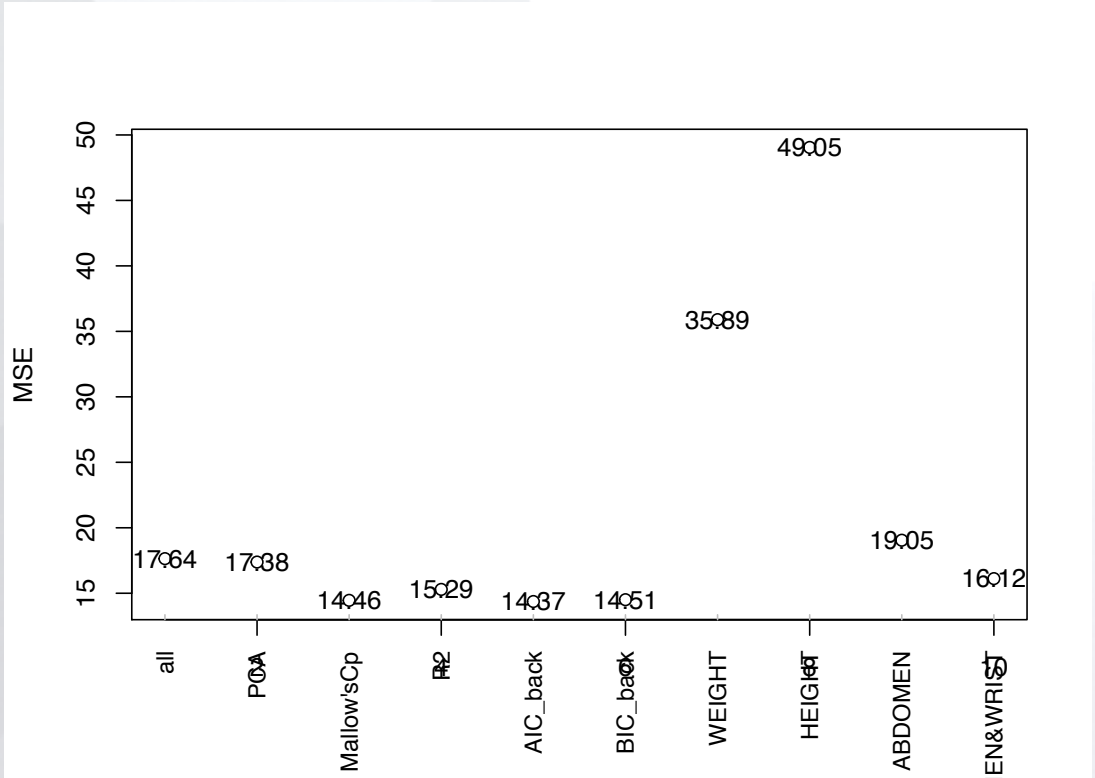
Model Selection & Evaluation

Methods	R^2	Number of Variables
All Variables	0.7415	14
Mallow's CP	0.7313	5
Adjusted R^2	0.7373	7
AIC BACKWARD	0.7353	6
BIC BACKWARD	0.7247	3
BIC BACKWARD and VIF	0.7058	2

Model Selection & Evaluation

Compare the MSEs of different models with that of PCA model

Separate dataset
--Training set: 2/3
--Validation set: 1/3
Cross Validation



Model Selection & Evaluation

BODYFAT = 0.717*ABDOMEN - 2.06008*WRIST-9.7813

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.78130    5.22738  -1.871   0.0625 .
ABDOMEN      0.71700    0.03211  22.329  < 2e-16 ***
WRIST       -2.06008    0.35140  -5.863 1.48e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.034 on 243 degrees of freedom
Multiple R-squared:  0.7058,    Adjusted R-squared:  0.7034
F-statistic: 291.5 on 2 and 243 DF,  p-value: < 2.2e-16
```

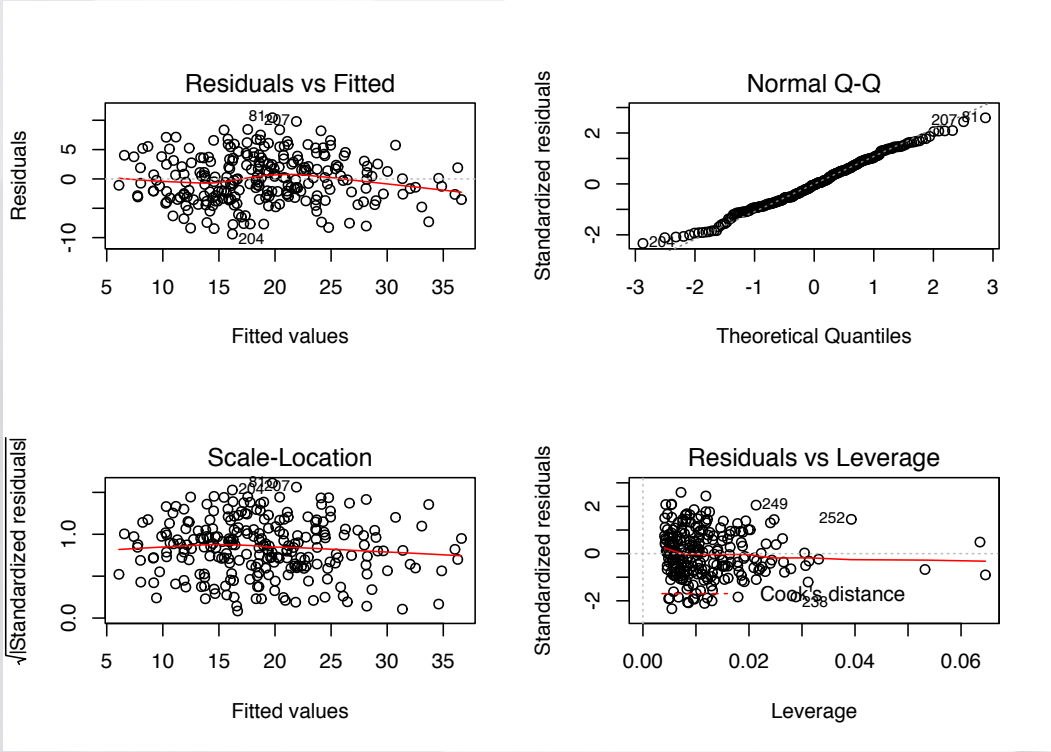
Model Selection & Evaluation

BODYFAT = 0.717*ABDOMEN - 2.06008*WRIST-9.7813

Normality

Equal Variance

Independence



Model Selection & Evaluation

- Cost Less
- No Multi-collinearity
- R^2= 0.7058
- No Overfitting

- Trade off between Cost and Performance

Rule of Thumb

- Simple Model
- Cost and Performance

Application

- Shiny

Link: https://ericchenzhang.shinyapps.io/body_fat_calculator/

Thanks for Listening