Body Fat Analysis

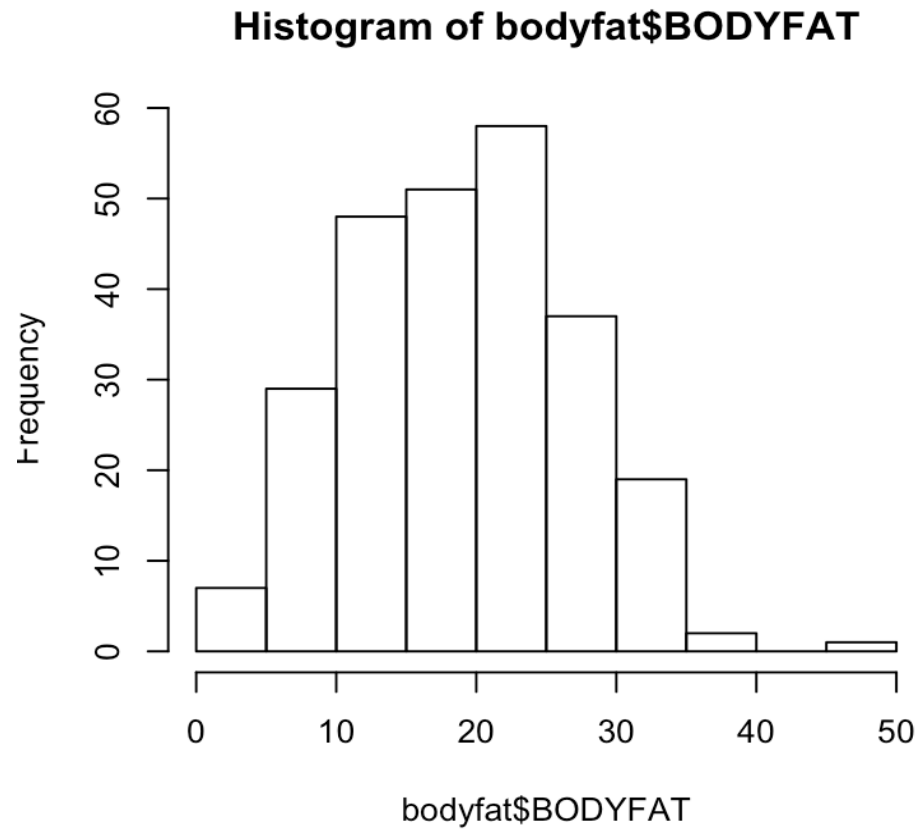
STAT 628 Module 1

Yanqi Huang, Wenjin Li, Xiawei Wang

February 7, 2018

Data Cleaning

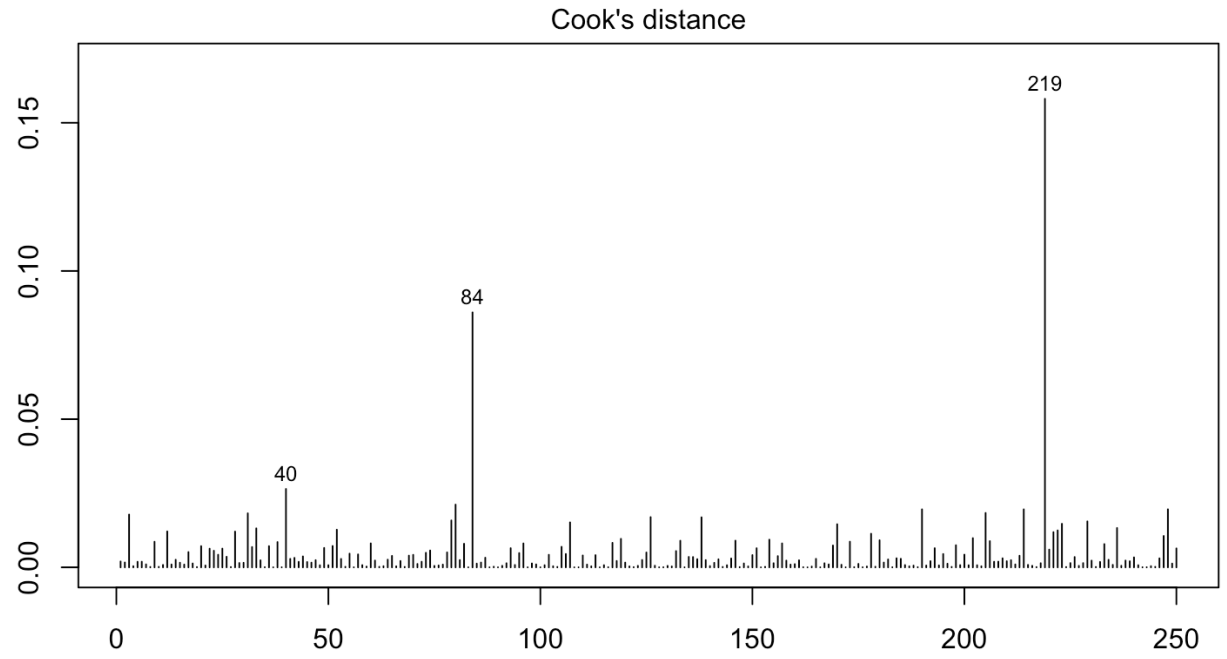
- Summarized the data, the minimum of BODYFAT and HEIGHT seem abnormal, which are 0% and 29 inches correspondingly.




```
> bodyfat[c(42,182),]
    IDNO BODYFAT AGE WEIGHT HEIGHT ADIPOSITY NECK CHEST ABDOMEN   HIP THIGH KNEE ANKLE BICEPS FOREARM WRIST
42    42    31.7  44  205.0   29.5      29.9 36.6 106.0   104.3 115.5  70.6 42.5  23.7   33.6    28.7  17.4
182  182     0.0  40  118.5   68.0      18.1 33.8  79.3    69.4  85.0  47.2 33.5  20.2   27.7    24.6  16.5
```

Data Cleaning

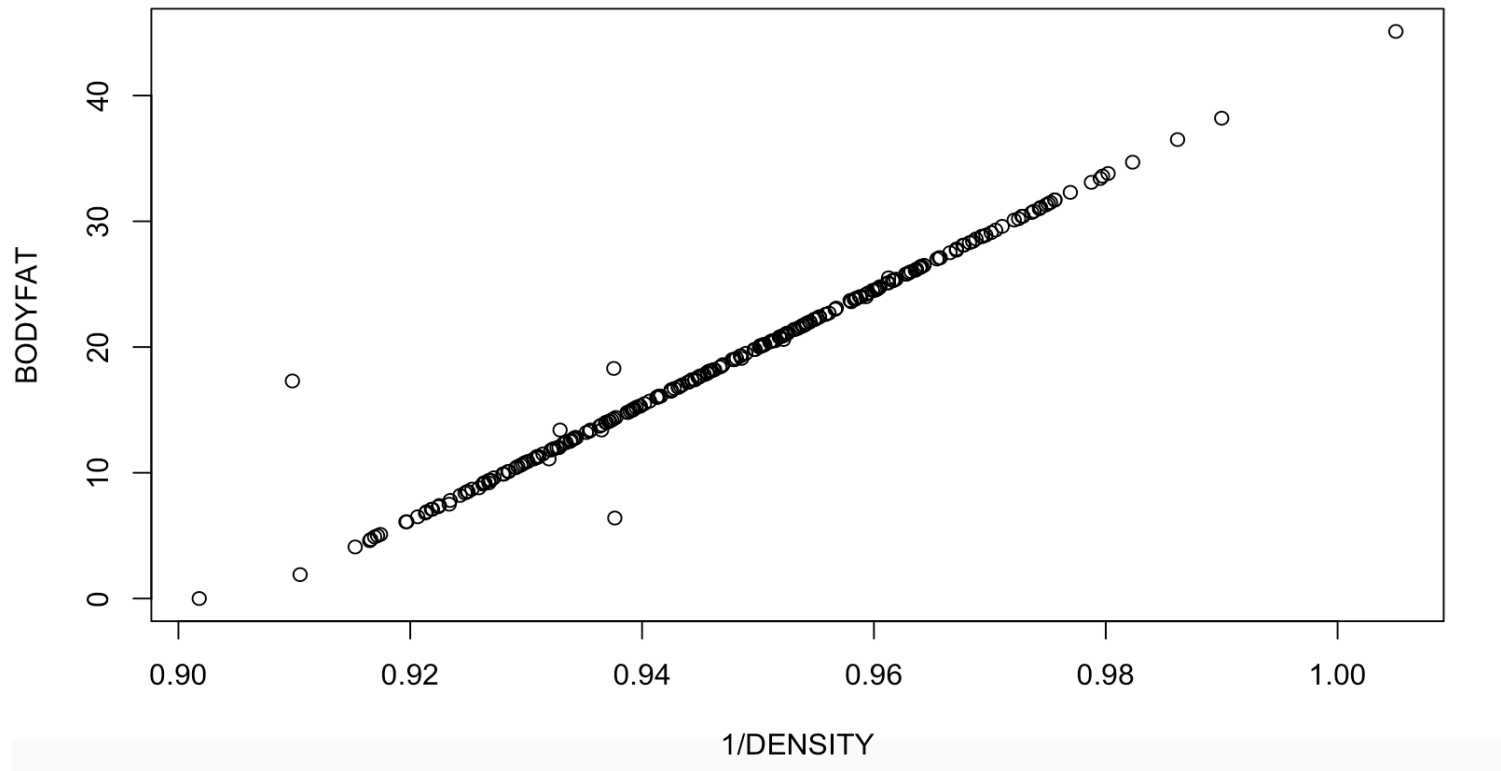
- Regression on BODYFAT after removing the 42nd and 182nd observations.
- The 39th point is shown as the high leverage point.




```
> bodyfat[39,]
   IDNO BODYFAT AGE WEIGHT HEIGHT ADIPOSITY NECK CHEST ABDOMEN   HIP THIGH KNEE ANKLE BICEPS FOREARM WRIST
39   39    33.8  46 363.15  72.25      48.9 51.2 136.2   148.1 147.7  87.3 49.1  29.6     45      29  21.4
```

Data Cleaning

- In fact, point 39 has the maximum value of 10 variables, we delete it due to its high influence on the dataset
- According to Siri’ s Equation, BODYFAT is proportional to the inverse of DENSITY. Remove points that are not aligned on the line, that is, point 48,96,76.



Multicollinearity Check

- Based on common sense, we know there exists correlation between explanatory variables, shown as below:

ANKLE	AGE	FOREARM	WRIST	BICEPS	NECK	KNEE
1.851834	2.290695	2.393534	3.201316	3.398873	3.868223	4.323974
THIGH	CHEST	HIP	ABDOMEN	HEIGHT	ADIPOSITY	WEIGHT
7.025465	10.996107	12.196766	12.258507	27.617778	91.789505	122.668687

- To avoid multicollinearity between variables, we performed Mallow’ s Cp procedure on the cleaned dataset

Multicollinearity Check

- We noticed that when degree of freedom is greater than 5, Mallow' s Cp and RSS do not decrease significantly. so we choose those five variables to perform the regression process.

```
> summary(laa)
LARS/LASSO
Call: lars(x = x, y = y)
   Df     Rss      Cp
0   1 14176.2 672.071
1   2  4930.1  76.588
2   3  4444.0  47.172
3   4  4392.1  45.819
4   5  3849.8  12.778
5   6  3800.7  11.606
6   7  3787.5  12.751
7   8  3734.7  11.339
8   9  3726.6  12.815
9  10  3701.0  13.157
10 11  3683.6  14.038
11 12  3603.4  10.853
12 13  3597.1  12.447
13 14  3593.1  14.190
14 15  3579.0  15.276
15 14  3578.2  13.225
16 15  3574.7  15.000
```

Regression

- The first five variables given by Mallow’ s Cp are ABDOMEN HEIGHT AGE WRIST NECK

```
> laa

Call:
lars(x = x, y = y)
R-squared: 0.748
Sequence of LASSO moves:
     ABDOMEN HEIGHT AGE WRIST NECK BICEPS FOREARM THIGH CHEST ANKLE HIP KNEE ADIPOSITY WEIGHT HEIGHT HEIGHT
Var        7      3   1    14    5     12      13     9     6    11   8   10         4      2     -3      3
Step       1      2   3     4    5      6       7     8     9    10  11   12        13     14     15     16
```

Model Comparsion

We use different criteria to evaluate models generated by different methods. These methods give different selection of variables.

	Lasso	**Backward**	**Lasso.step**	**Full**
Adjusted R^2	0.7294	0.7233	0.7252	0.7326
AIC	1381.7638	1386.2765	1384.6215	1388.4901
BIC	1402.7958	1403.8032	1402.1481	1444.5754
MSE	15.337	15.7485	15.6429	14.5314
Variables	HEIGHT, ABDOMEN, WRIST, AGE	HEIGHT, ABDOMEN, WRIST	ABDOMEN, WRIST, AGE	All

Thus, the model given by Lasso gives the best results, that is , it has the smallest ad-R^2, AIC and MSE, and BIC value doesn’ t differ from the smallest one significantly.

Model Summary

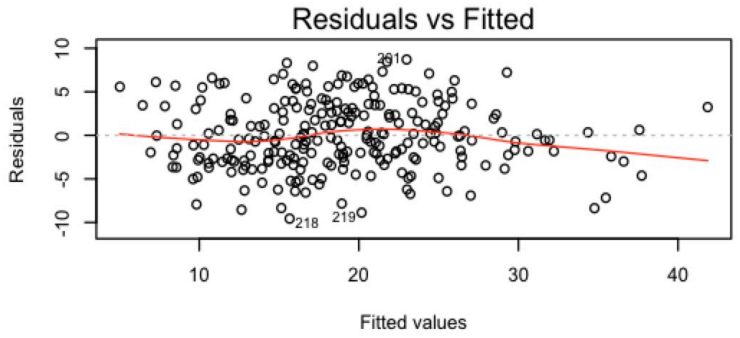
The final model includes four variables: ABDOMEN, AGE, WRIST, HEIGHT:

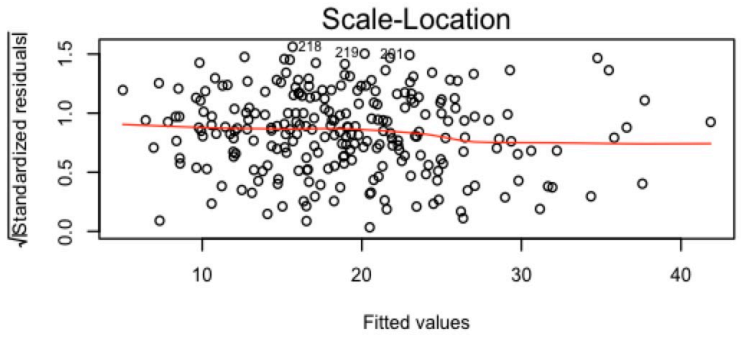
%Body Fat=4.05−0.29*Height+0.71*Abdomen−1.79*Wrist+0.05*Age

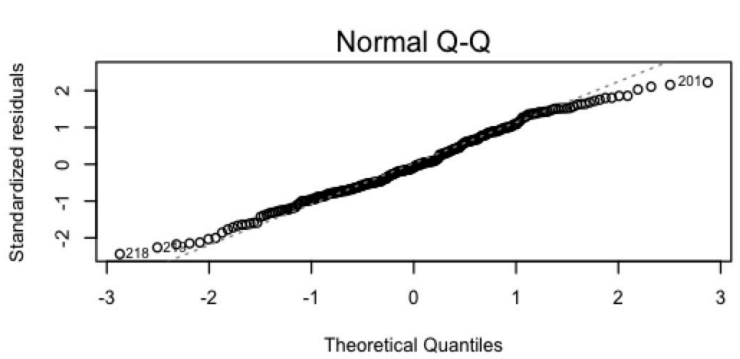
	Estimate	Std. Error	t value	Pr(>|t|)	Confidence Intervals
(Intercept)	4.0542	7.6801	0.5279	0.5981	(-10.9985, 19.1068)
HEIGHT	-0.2895	0.1139	-2.5427	0.0116	(-0.5126, -0.0663)
ABDOMEN	0.7115	0.0315	22.5524	0	(0.6497, 0.7733)
WRIST	-1.7945	0.3813	-4.7062	0	(-2.5419, -1.0472)
AGE	0.0486	0.0221	2.1923	0.0293	(0.0051, 0.092)

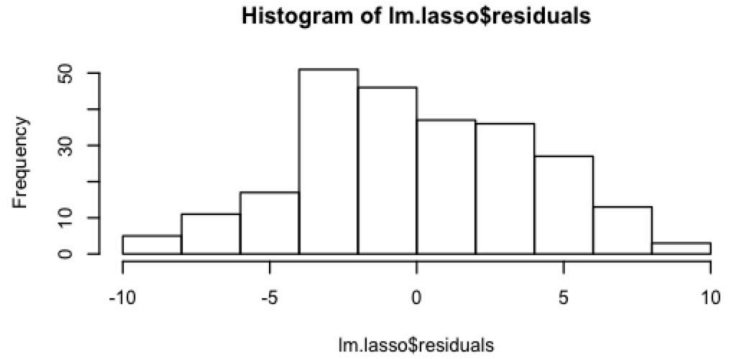
Diagnostic Plots

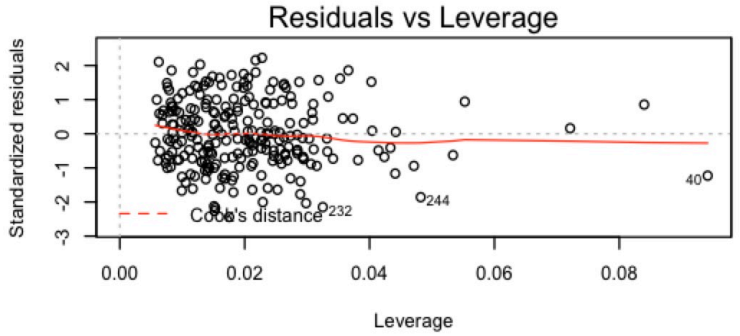
The final model includes :
ABDOMEN,
AGE,
WRIST,
HEIGHT

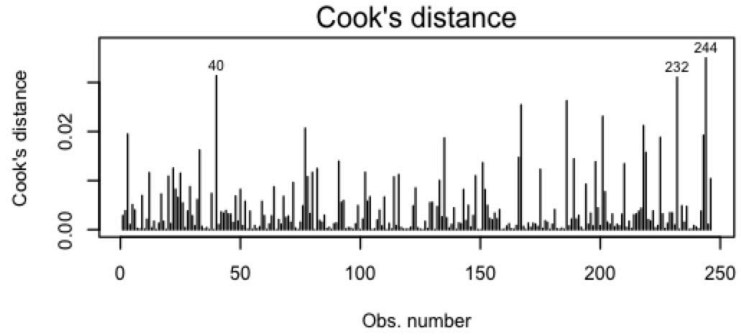












Discussion

Strengths

- Linearity: Scatterplots of Bodyfat against each of the predictors are reasonably straight. And the The residuals appear to be randomly scattered and have no pattern with respect to the predicted values, which indicates that we capture the linearity and there is no non-linear relation between %bodyfat and predictors.
- Simplicity: The final model only has 4 predictors: age, height, circumferences of abdomen and wrist. All of the measurements are easy to obtain.
- The final model conform to our experience.

Weakness

- Homoscedasticity: The scale-location plot shows that variation of %bodyfat seems less when fitted values get larger.
- Normality: According to the qq-plot of residuals, it is suspected that the normal distribution is light-tailed.
- Precision: Our practical methods have a 3% to 4% error factor in the prediction of body fat.
- Restriction: Since the dataset is observed from American men, the conclusion will not be that accurate when it is used for women. The problem can be solved by scaling.
- There may be non-linear relationships between bodyfat and circumferences.

Rule of Thumb

For easier calculation, we use values estimated by the parameters:

$$\%\text{Body Fat} = 4 - 0.3 * \text{Height} + 0.7 * \text{Abdomen} - 1.8 * \text{Wrist} + 0.05 * \text{Age}$$

- Each inch in height is associated with about a 0.3 decrease in %body fat among men.
- Each centimeter of abdomen is associated with a increase in %body fat of about 0.7 among men.
- Each centimeter of wrist is associated with a increase in %body fat of about 1.8 among men.
- One-year increase of age will in general lead to the decrease in %body fat of about 0.05.

Example

- We take points 48,76 and 96 as examples. They are excluded for the regression procedure previously. Given their data of height, abdomen, age and wrist, we have their predicted body fat.

HEIGHT	ABDOMEN	AGE	WRIST	BODYFAT
71.25	79.5	39	17.9	8.005
67.50	81.8	61	18.3	11.120
72.00	100.0	54	18.9	16.045

- Another Example: For a man aged 30 with 73 inches height, 90 cm abdomen Cir. and 17 cm wrist Cir., his predicted body fat % percentage would be around 17.87%. With the rule of thumb, we get about 13.2% as the predicted body fat 16%.