BodyFat Analysis

Shiwei Cao, Shurong Gu, Yuwei Sun

February 2, 2018

Procedure

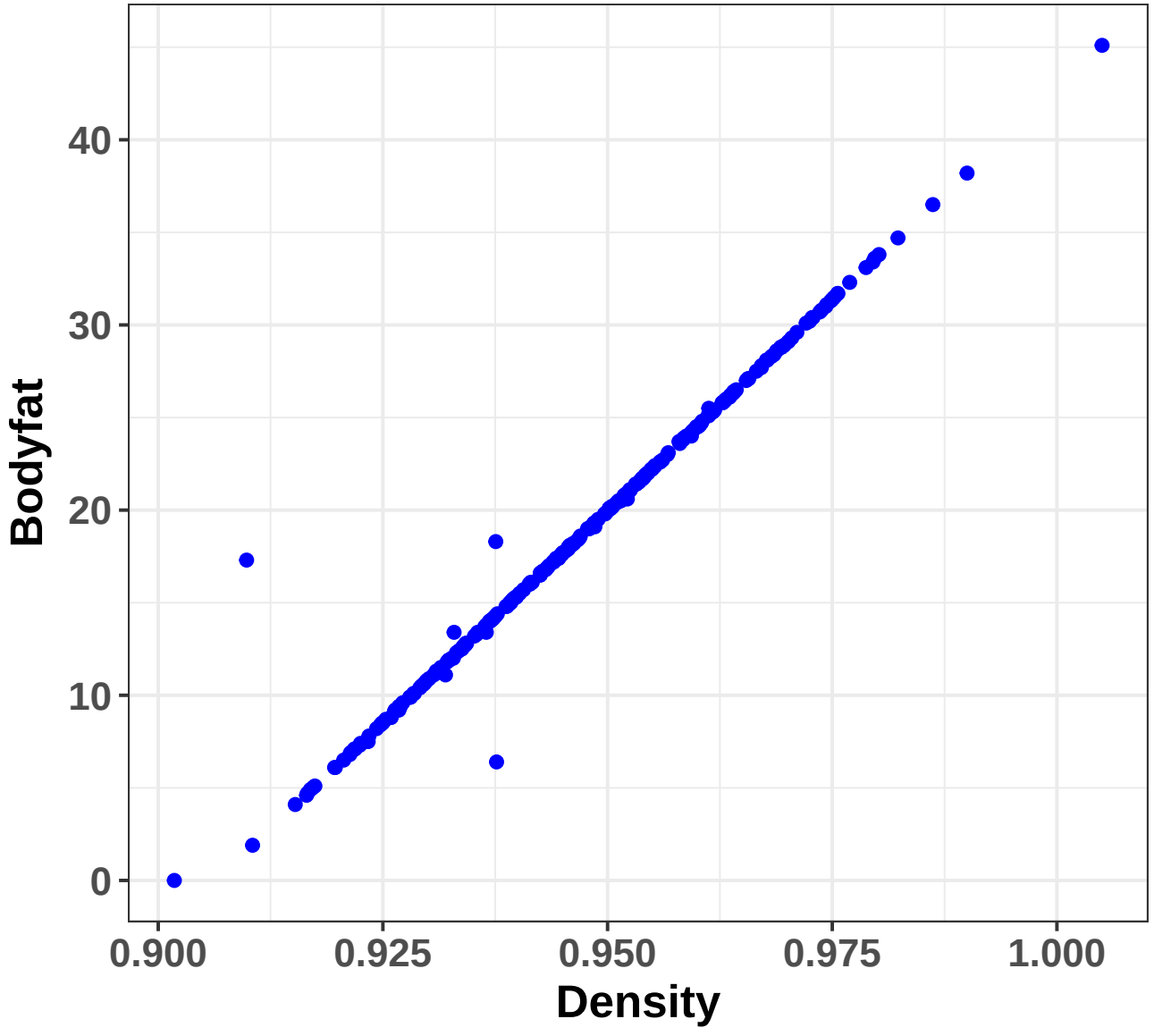
- **Data preprocessing**: EDA, data cleaning
- **Statistical Modeling**:
  - Multiple Linear Regression
    - Variable Selection: stepwise method
    - Model Diagnosis
  - Lasso
- **Model selection**:
  - Make predictions on validation set and choose the best model (the one with smallest mse)
- **Model Interpretation**, **"rule of thumb"**
- **Strengths and weaknesses**

Explore the data

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

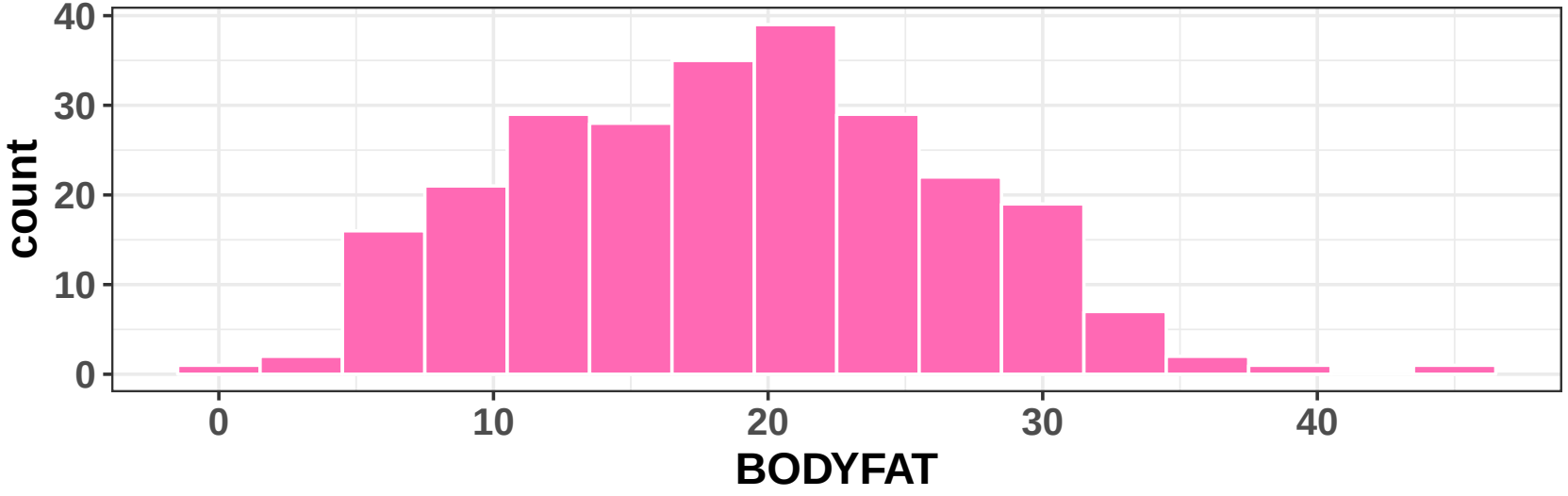
	IDNO	BODYFAT	DENSITY	AGE	WEIGHT	HEIGHT	ADIPOSITY	NECK	CHEST	ABDOMEN	HIP	THIGH	KNEE	ANKLE	BICEPS	FOREARM	WRIST
247	247	29.1	1.0308	69	215.50	70.50	30.5	40.8	113.7	107.6	110.0	63.3	44.0	22.6	37.5	32.6	18.8
248	248	11.5	1.0736	70	134.25	67.00	21.1	34.9	89.2	83.6	88.8	49.6	34.8	21.5	25.6	25.7	18.5
249	249	32.3	1.0236	72	201.00	69.75	29.1	40.9	108.5	105.0	104.5	59.6	40.8	23.2	35.2	28.6	20.1
250	250	28.3	1.0328	72	186.75	66.00	30.2	38.9	111.1	111.5	101.7	60.3	37.3	21.5	31.3	27.2	18.0
251	251	25.3	1.0399	72	190.75	70.50	27.0	38.9	108.3	101.3	97.8	56.0	41.6	22.7	30.5	29.4	19.8
252	252	30.7	1.0271	74	207.50	70.00	29.8	40.8	112.4	108.5	107.1	59.3	42.2	24.6	33.7	30.0	20.9

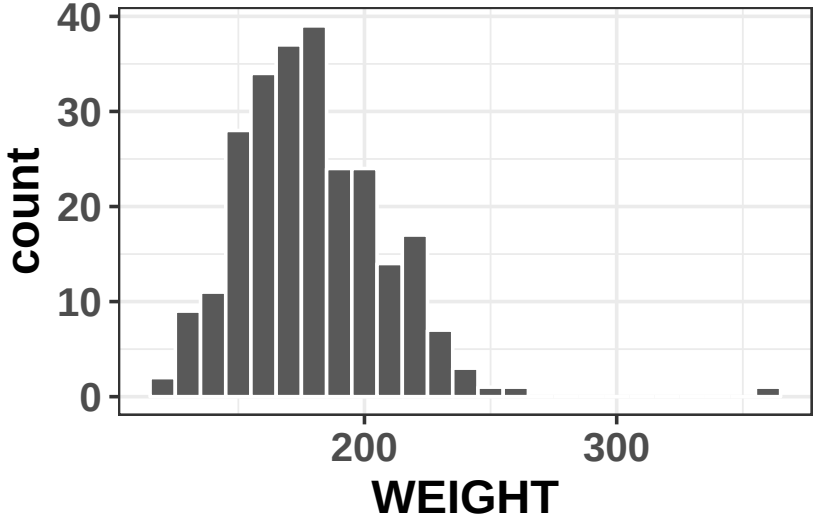
Data Cleaning

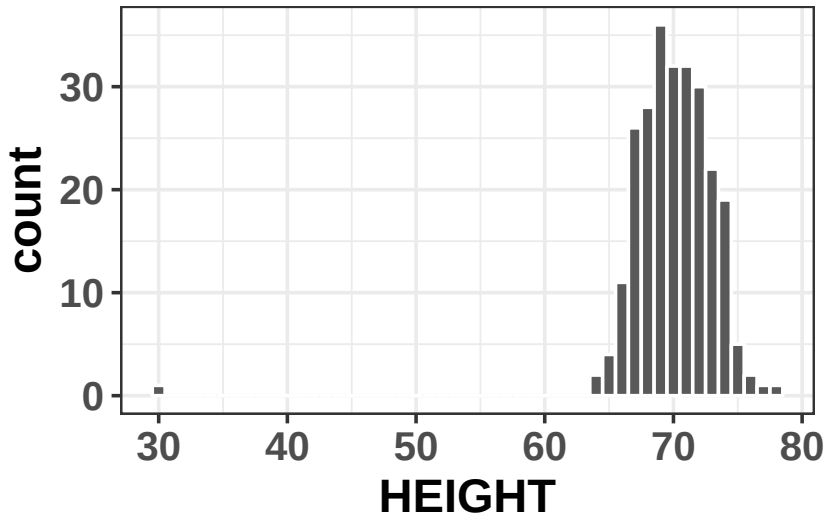


We decide to remove 3 points for further analysis:

The one with BODYFAT=0; the one with HEIGHT=29.5 inches (only 75cm tall); the one with WEIGHT=363.15 pounds







Variable Selection

- Divide the whole data into train (80%) and validation (20%) set
- Train set: Use several different methods to choose the best subset of variables

1. Mallow's Cp: leaps() in R (do an exhaustive search)
2. Stepwise regression based on AIC
3. Stepwise regression based on BIC

- Validation set: See if the models generalize well on unseen data

```
##          Model    MSE
## 1 Mallow's Cp 15.069
## 2          AIC 14.702
## 3          BIC 13.989
```

BIC results

Start:

```
> model.BIC <- step(model, k=log(249))
Start:  AIC=624.5
BODYFAT ~ AGE + WEIGHT + HEIGHT + ADIPOSITY + NECK + CHEST +
    ABDOMEN + HIP + THIGH + KNEE + ANKLE + BICEPS + FOREARM +
    WRIST

            Df Sum of Sq    RSS    AIC
- KNEE       1      0.16 3002.1 618.99
- BICEPS     1      6.13 3008.1 619.39
- NECK       1      6.38 3008.3 619.41
- CHEST      1     11.80 3013.7 619.77
- HIP        1     15.46 3017.4 620.01
- ANKLE      1     20.35 3022.3 620.33
- FOREARM    1     20.53 3022.4 620.35
- THIGH      1     25.75 3027.7 620.69
- HEIGHT     1     32.21 3034.1 621.12
- AGE        1     38.57 3040.5 621.54
- ADIPOSITY  1     42.18 3044.1 621.77
- WEIGHT     1     47.07 3049.0 622.09
<none>                   3001.9 624.50
- WRIST      1    121.38 3123.3 626.91
- ABDOMEN    1   1117.80 4119.7 682.29
```

End:

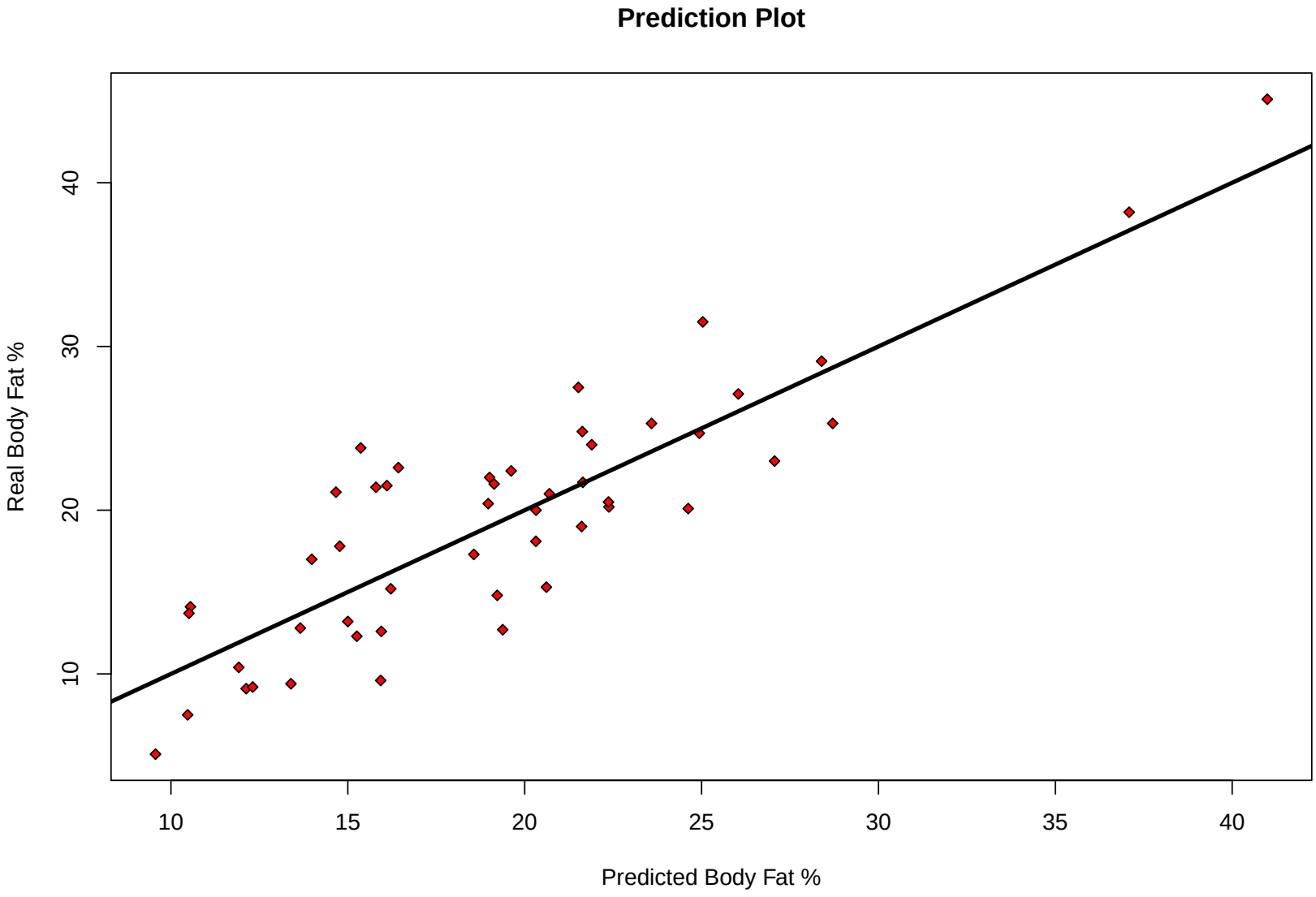
```
Step:  AIC=576.93
BODYFAT ~ WEIGHT + ABDOMEN + WRIST

            Df Sum of Sq    RSS    AIC
<none>                   3205.4 576.93
- WRIST      1     121.7 3327.1 578.86
- WEIGHT     1     138.9 3344.3 579.89
- ABDOMEN    1    3348.7 6554.1 714.46
```

Variables selected by BIC:

At this point, we come up with a multiple linear regression:

- $Bodyfat \sim Abdomen + Wrist + Weight$

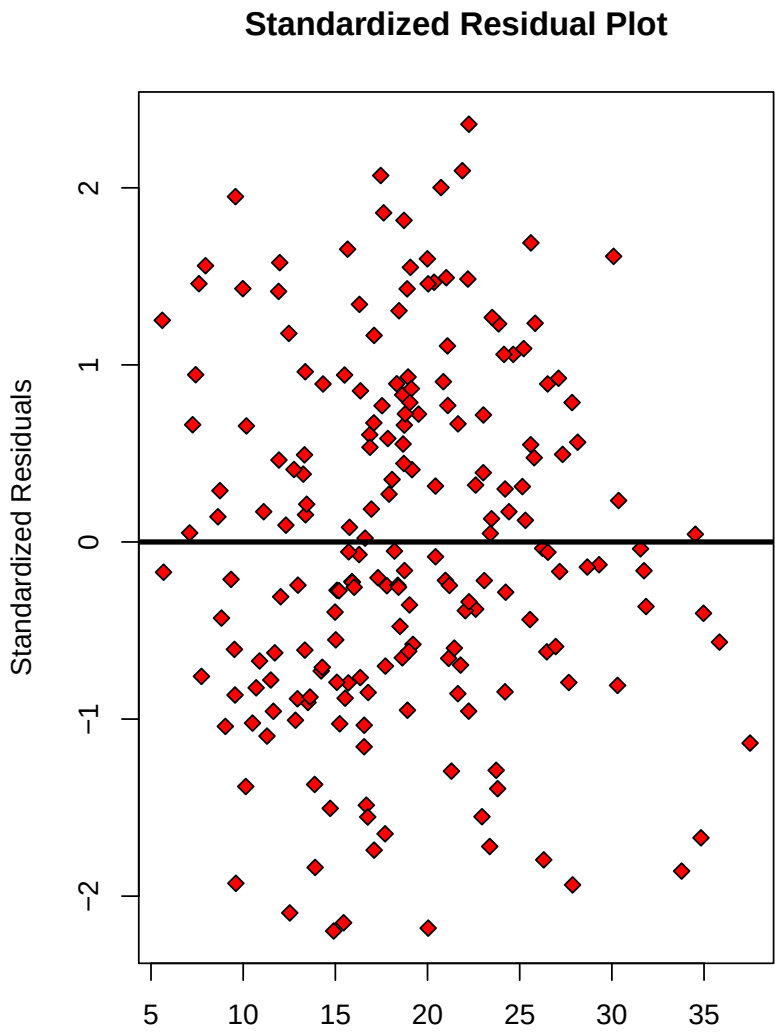
Prediction on validation set

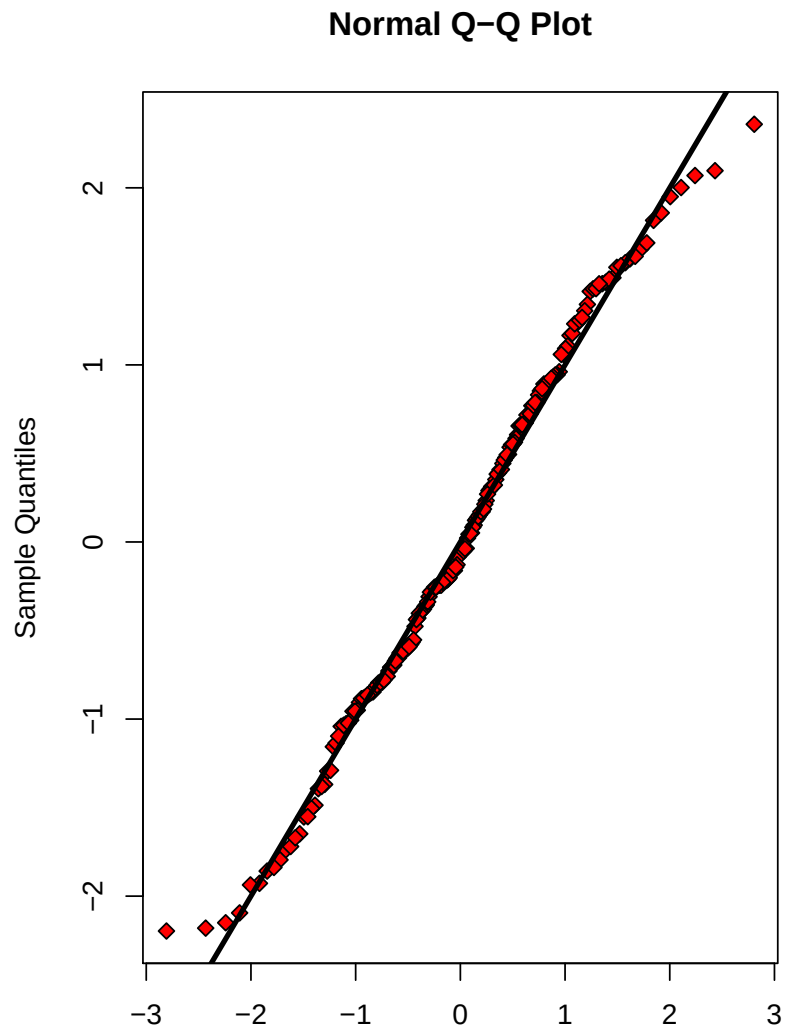


Model Diagnosis

Adjusted R-squared for the final model: 0.7175

Linear regression assumptions?





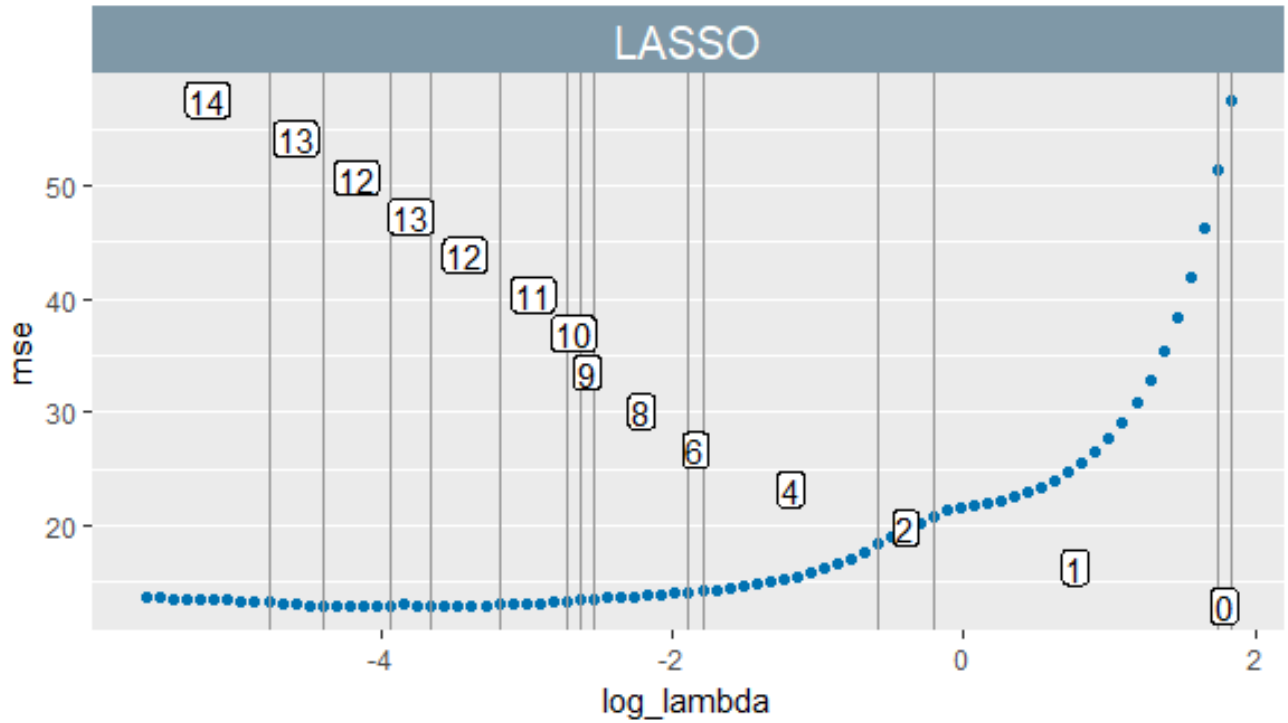
Still exist multicollinearity?

```
vif(m2)
```

```
##    WEIGHT   ABDOMEN     WRIST
## 5.669772 4.488463 1.989001
```

Another approach: Lasso

$$\min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i - \cdots - \beta_p x_p)^2 + \lambda \left| \sum_{j=1}^{p} \beta_j \right|$$



Conclusion

- Our proposed linear model to predict body fat %:

$$\begin{aligned}&(\text{BodyFat \%})= \\ &-23.7937 + 0.8519 \times \textit{Abdomen} - 1.2582 \times \textit{Wrist} - 0.0735 \times \textit{Weight}\end{aligned}$$

- Possible rule of thumb:
  - Your *% Bodyfat* =
  - Your *abdomen* circumference (cm) $\times 0.85$
  - minus *wrist* circumference (cm) $\times 1.26$
  - minus *weight* (lbs) $\times 0.07$
  - minus 24
- For a normal graduate male student, with circumferences: Abdomen=85cm, Wrist=18cm, Weight=130lbs, his predicted body fat percentage would be around 16.43%. There is a 95% probability that his body fat is between 8.26% and 24.59%.

Strengths and Weaknesses

- Strengths

1. Use a separate validation set to avoid overfitting
2. Simple, easy to interpret

- Weaknesses

1. May lose information using only 200 data points
2. Trade off between simplicity and precision