Bodyfat Dataset Analysis

By Yan Chu, Boliang Ma, Jiancheng Xu (Group 5)

Department of Statistics

UW-Madison

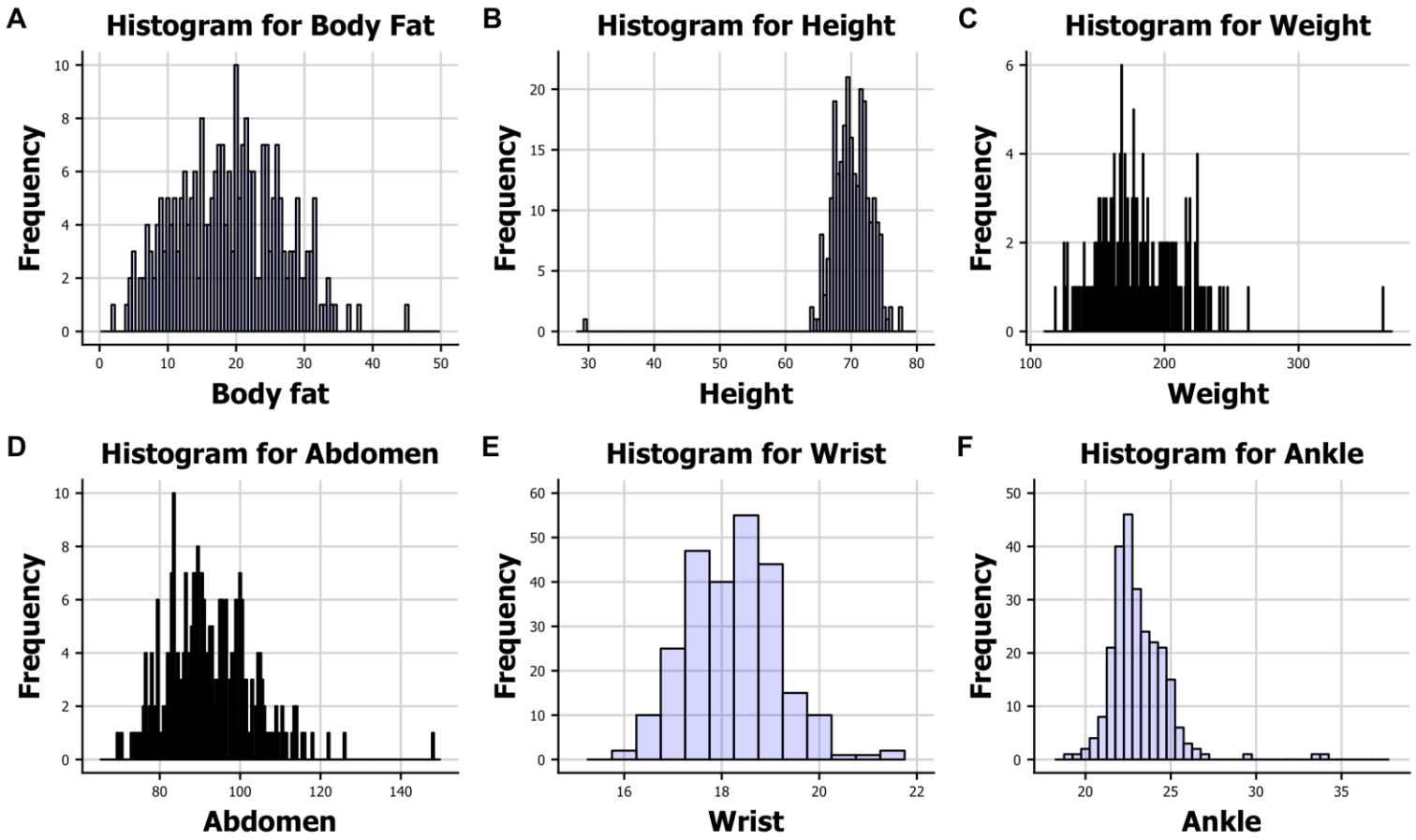
Outline

- Raw Data and Data Visualization
- Modeling and Analysis
- Diagnostics and Conclusion

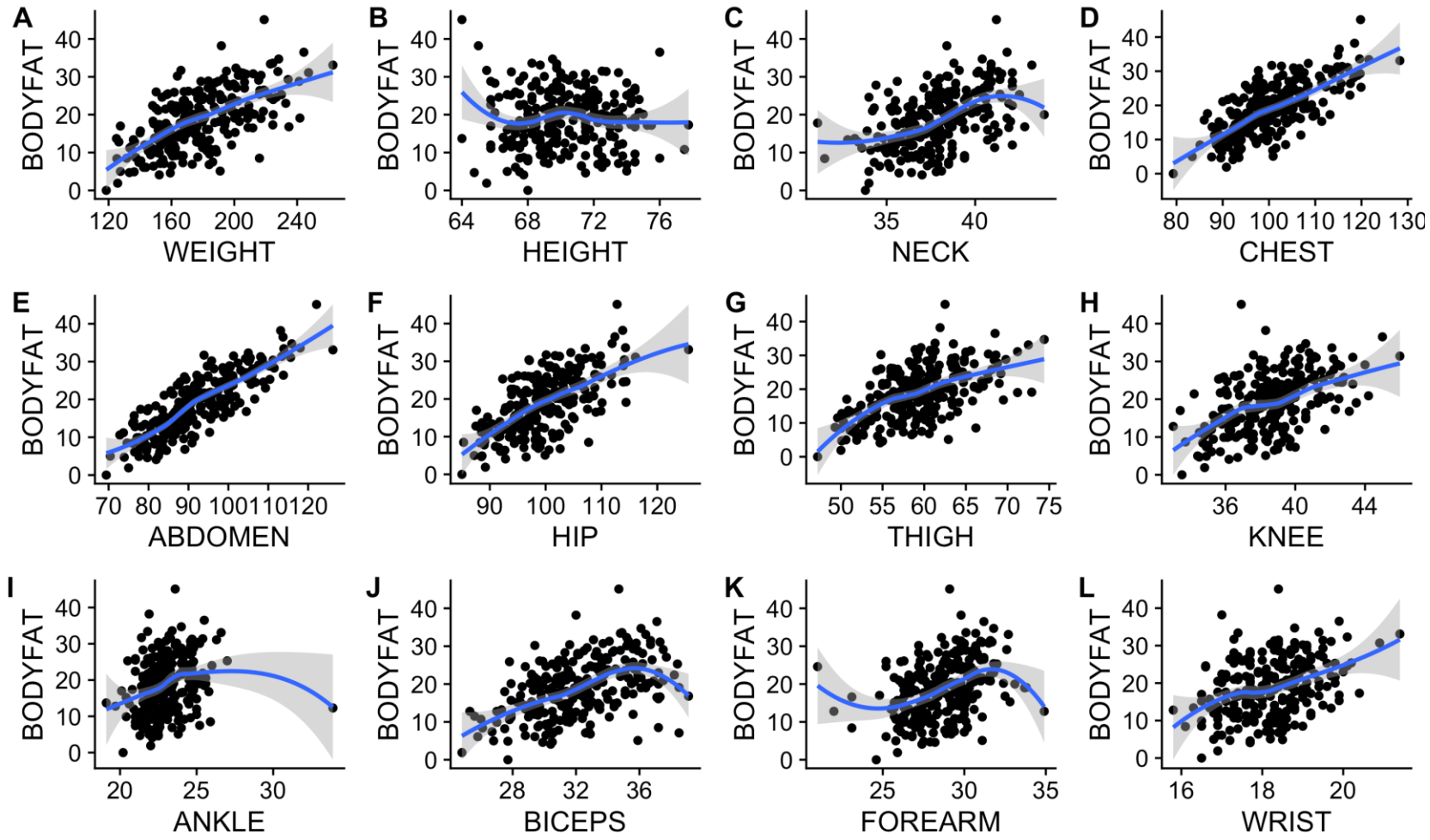
Raw Data and Data Visualization

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

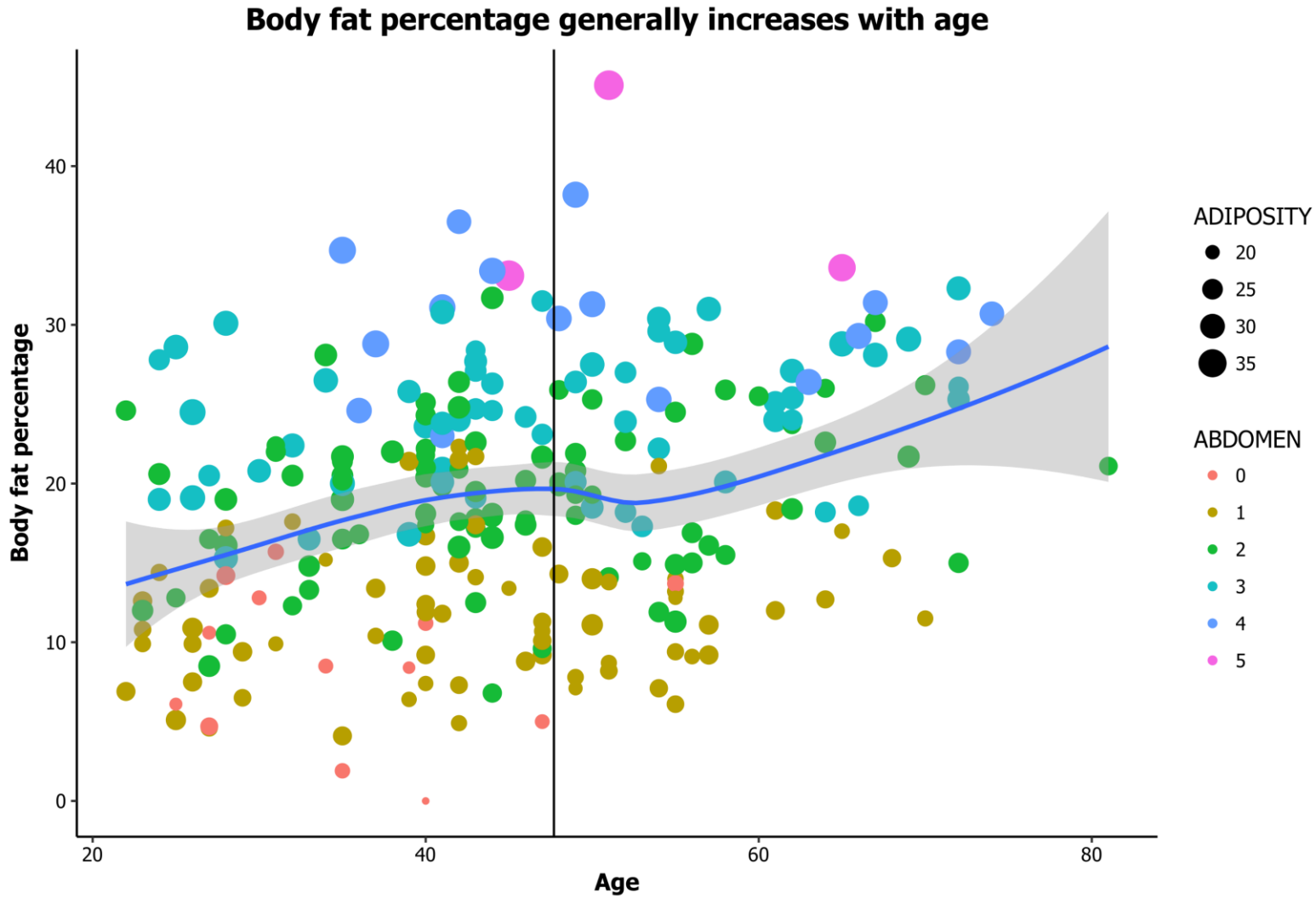
Raw Data and Data Visualization



Raw Data and Data Visualization



Raw Data and Data Visualization



Modeling and Analysis

- Full data model - use all data

- Two age level separate model - divided by age

Modeling and Analysis

- Full model is obtained use forward selection.

Method	Variable	MSE
Forward	Abdomen, Weight, Wrist, Forearm, Thigh, Age	14.74
Backward	Age, Weight, Height, Adiposity, Abdomen, Thigh, Forearm, Wrist	14.55

- Forward model is simpler and only losing a little accuracy
- Further deleting three variables, keep WEIGHT, ABDOMEN and WRIST.

Modeling and Analysis

- Separated model

- Under age 45, multicollinearity is a heavy issue.

Variable	Age	Weight	Height	Adiposity	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
VIF Value	1.41	339.28	60.88	252.94	4.56	10.70	14.07	18.42	9.82	5.26	4.07	5.36	2.05	4.39

- Delete variable ADIPOSITY
- Finally find ABDOMEN, WEIGHT, WRIST, BICEPS and ANKLE.

Modeling and Analysis

- Separated model
- Beyond 45
- Forward and backward give the same solution: ABDOMEN and WEIGHT. Simple and accurate.

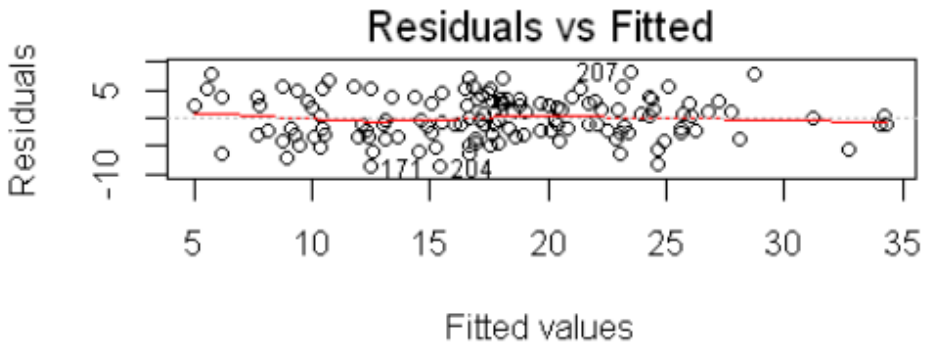
Modeling and Analysis

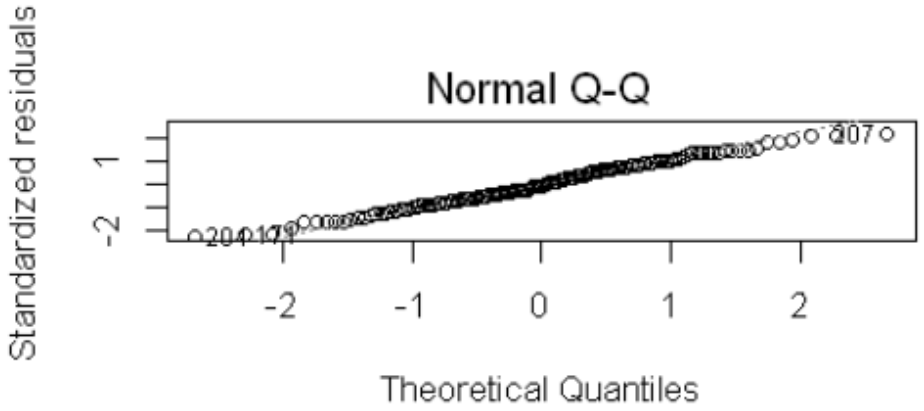
- Compared Mean square error of the full data model and separated two age level model

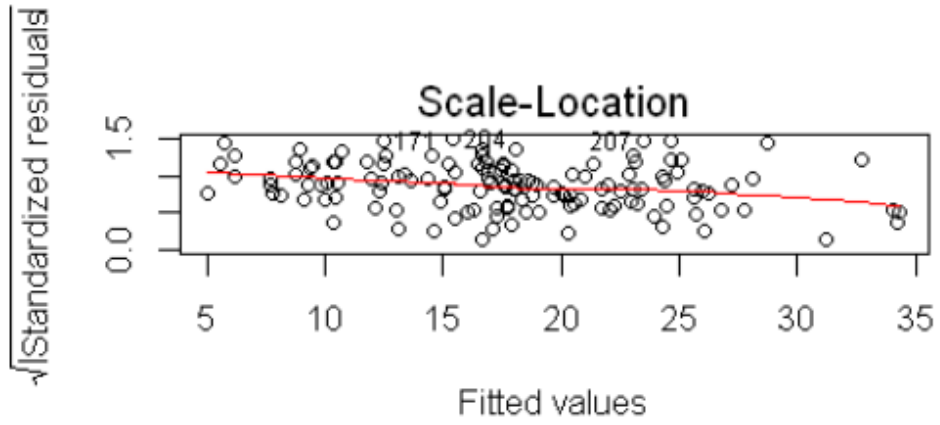
	Full data model	**Combined model**
MSE	15.32	14.44

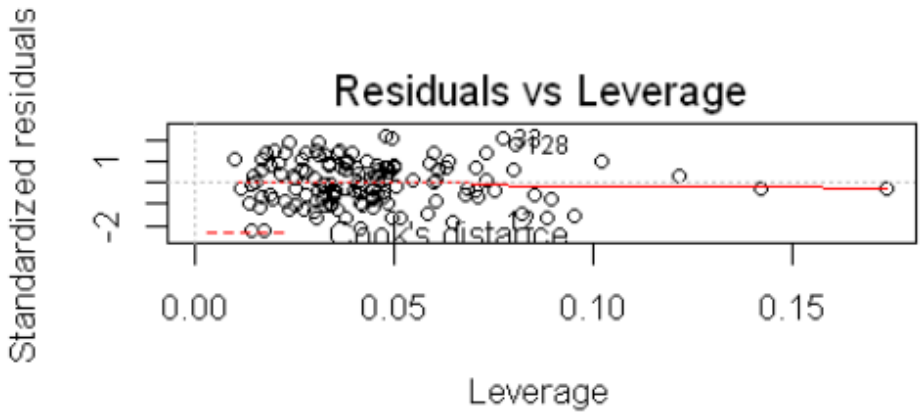
Model Diagnostics

- Under 45



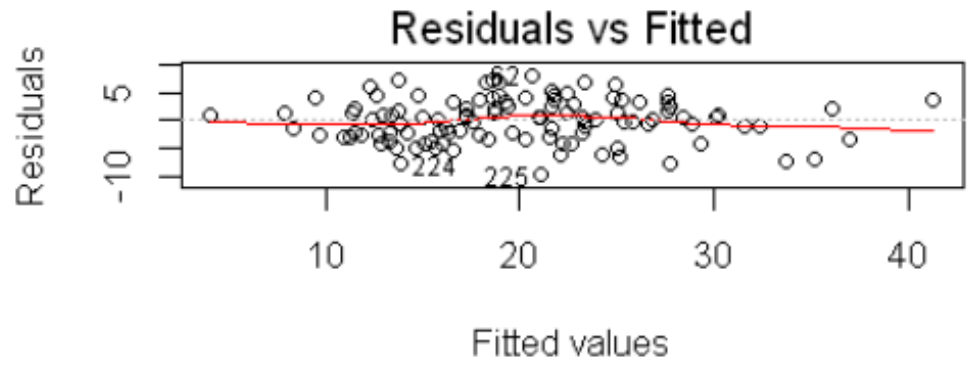


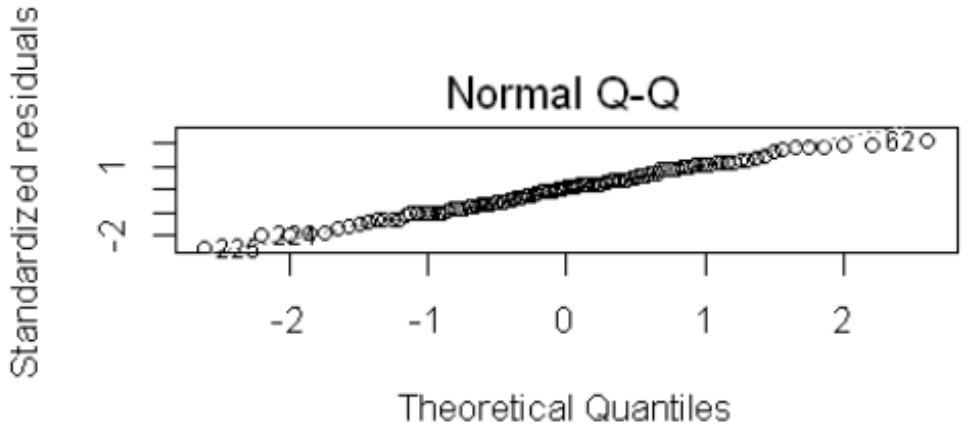


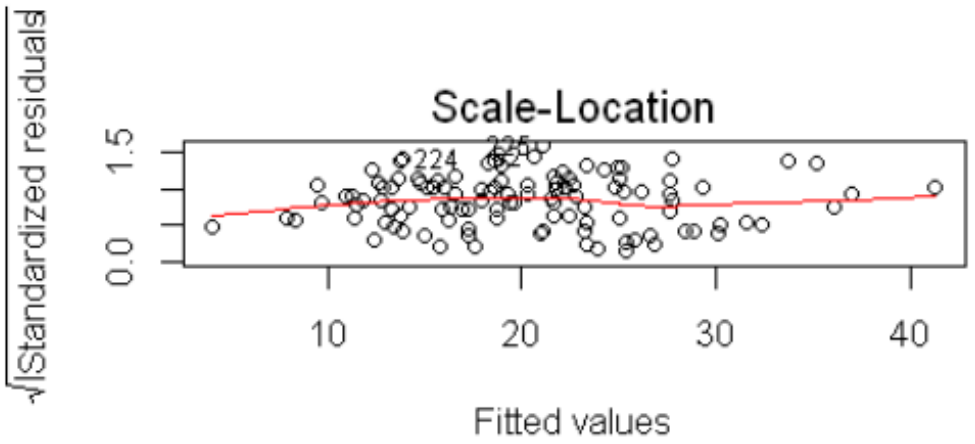


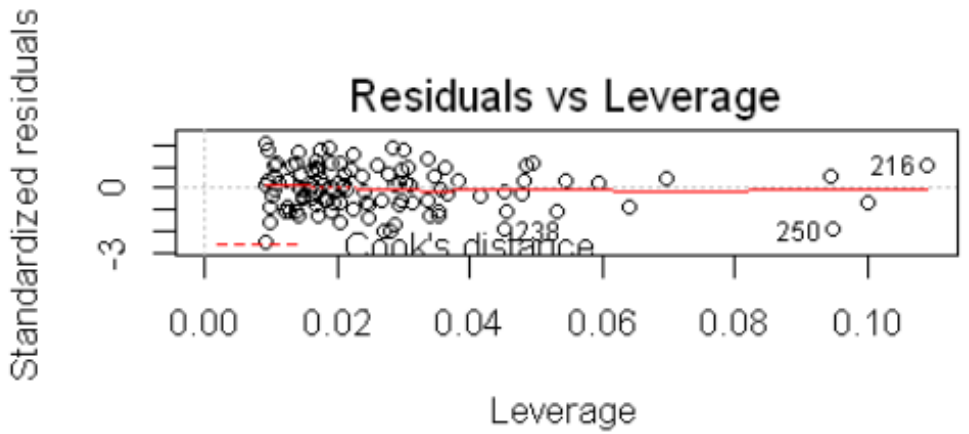
Model Diagnostics

- Beyond 45









Conclusion

$$BodyFat\% = \begin{cases} -28.927 + 0.920ABDOMEN - 0.150WEIGHT - 2.498WRIST + 0.565BICEPS + 0.723ANKLE, & if\ AGE \leq 45 \\ -46.224 + 0.919ABDOMEN - 0.112WEIGHT, & if\ AGE > 45 \end{cases}$$

Possible rule of thumb:

Under 45

$$BodyFat\% = -30 + ABDOMEN - 0.1WEIGHT - 2.5WRIST + 0.5BICEPS + ANKLE$$

Beyond 45

$$BodyFat\% = -46 + ABDOMEN - 0.1WEIGHT$$

Strength and Weakness

- All assumptions are well satisfied.
- No advanced techniques for dealing with multicollinearity is employed since we just delete the most correlated variables.
- Due to data constraints, our model is unable to predict bodyfat percentage for children or even more older people.
- We can only predict male's bodyfat.

Simple Example

- For a 47 years old man, he has a abdomen circumference 90cm and weight 200 lbs. His bodyfat can be estimated as 90 minus 200 multiplied by 0.1 and then minus 46, which is 24%.

$$BodyFat\% = -46 + ABDOMEN - 0.1WEIGHT$$

Future work

- PCA, LASSO, and PLS for multicollinearity
- GMC by Prof. Zhengjun Zhang for variable selection
- For large dataset, generate a relatively small sample to start variable selection
- …

Thank you!