Measure Your Body Fat

Zhengyuan Wen, Jiaying Jia, Hengrui Qu

Module2 of STAT 628
Data Science Practicum FALL21
20 OCTOBER 2021

Contents

Outline

Introduction

- The goal is to come up a simple, robust, and accurate "rule-of-thumb" to estimate percentage of body fat using clinically available measurements
- **252** observations
- **BODYFAT** is response variable
- **14** predictors: Age, Weight, Height, Adiposity, Neck circumference, Chest circumference, Abdomen circumference, Hip circumference, Thigh circumference, Knee circumference, Ankle circumference, Biceps circumference, Forearm circumference, Wrist circumference

Data Cleaning by summary table

- Part of Summary of the dataset

BODYFAT	DENSITY	WEIGHT	HEIGHT
Min. : 0.00	Min. :0.995	Min. :118.5	Min. :29.50
1st Qu.:12.80	1st Qu.:1.041	1st Qu.:159.0	1st Qu.:68.25
Median :19.00	Median :1.055	Median :176.5	Median :70.00
Mean :18.94	Mean :1.056	Mean :178.9	Mean :72.25
3rd Qu.:24.60	3rd Qu.:1.070	3rd Qu.:197.0	3rd Qu.:197.0
Max. :45.10	Max. :1.109	Max. :363.1	Max. :363.1

- Some individuals which have the abnormal values

Individual	variables	outliers value
172	BODYFAT	1.9
182	BODYFAT	0.0
216	BODYFAT	45.1
39	WEIGHT	363.1
42	HEIGHT	29.50

Data Cleaning by Siri's equation and BMI equation

- **The Siri's Equation**

$$\text{BODYFAT}\,\% = \frac{495}{D} - 450,$$

$$\text{D is the Body Density } \left(\text{gm}/\text{cm}^3\right)$$

- **The BMI equation**

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

- Some individuals which disobey the Siri's Equation and BMI equation

Individual	disobeyed equation
33	Siri's
48	Siri's
76	Siri's
96	Siri's
163	BMI
221	BMI

Data Cleaning Summary

- In the equation checking, we are not sure which variable has input error in original dataset,so we remove them.
- The potential outliers only consist of less than 5%, we will remove them all.
- now there are **241** individuals

Outline

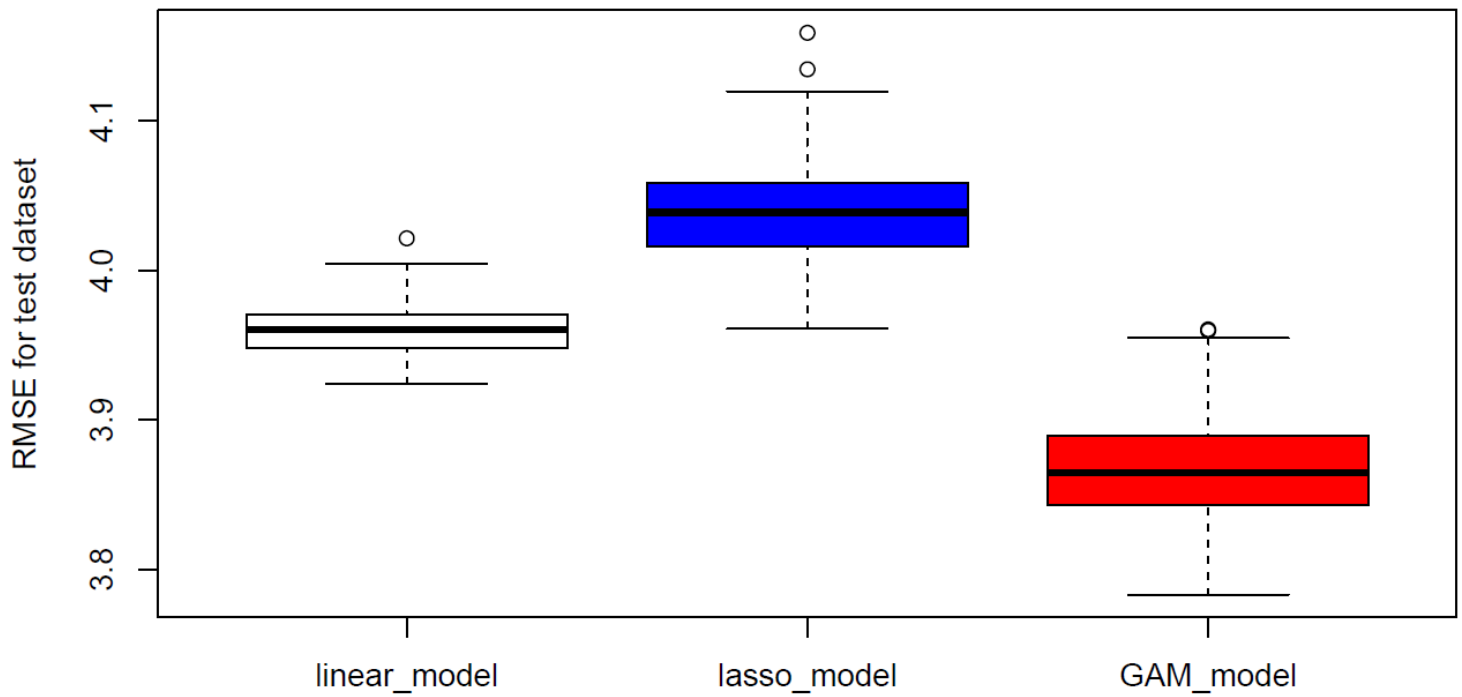
Candidate Models

- Simple Linear Regression with BIC Criteria
- Lasso
- Generalized Additive Model (GAM)

Candidate Model	Variables
SLR with BIC	WEIGHT ABONMEN WRIST
Lasso	AGE HEIGHT NECK ABDOMEN WRIST
GAMs	AGE WEIGHT FOREARM ABDOMEN THIGH WRIST

Metric for Model performance

- We ended up choosing the best model by **10-fold Cross Validation (CV)**. We compared three models above, used training data to train the model, and used test data to calculate **RMSE**



Variable Selection in GAMs

Variables	Models (0)	(1)	(2)	(3)	(4)	(5)
Intercept	0.0000	0.0000	0.0130	0.0035	0.0044	0.0003
AGE	0.0151	0.019	0.0163	0.0137	0.0185	0.0326
WEIGHT	0.1516	0.2084	0.0325	0.0044	0.0033	0.0009
HEIGHT	0.0612	0.6452				
NECK	0.0138	0.0606	0.0639	0.0758	0.1215	
ABDOMEN	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
HIP	0.0428	0.6217	0.4628			
THIGH	0.1616	0.0809	0.0433	0.0529	0.0180	0.0278
BICEPS	0.0253	0.2353	0.1900	0.1777		
FOREARM	0.2369	0.0897	0.0723	0.0661	0.0185	0.0550
WRIST	0	0.0017	0.0018	0.0018	0.0031	0.0008
SELECT	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE
AIC	1337.549	1337.911	1335.550	1334.177	1335.817	1337.629

Four decimal places number represent the P-value for smoothing terms or coefficients. The numbers with blue are treated as smoothing terms.

Outline

Model Description

- **Final Model**

$$\widehat{BODYFAT}_i = 26.035 + 0.060AGE_i - 0.108WEIGHT_i$$
$$+ 0.332FOREARM_i + f_1(ABDOMEN_i) + f_2(THIGH_i) + f_3(WRIST_i)$$

- **Linear terms**
- As men get older by one year/his weight get lower by one pound/his forearm circumferences larger by one centimeters , he is expected to gain 0.06%/0.108%/0.332% in body fat

Model Description

- **Non-parametric terms**

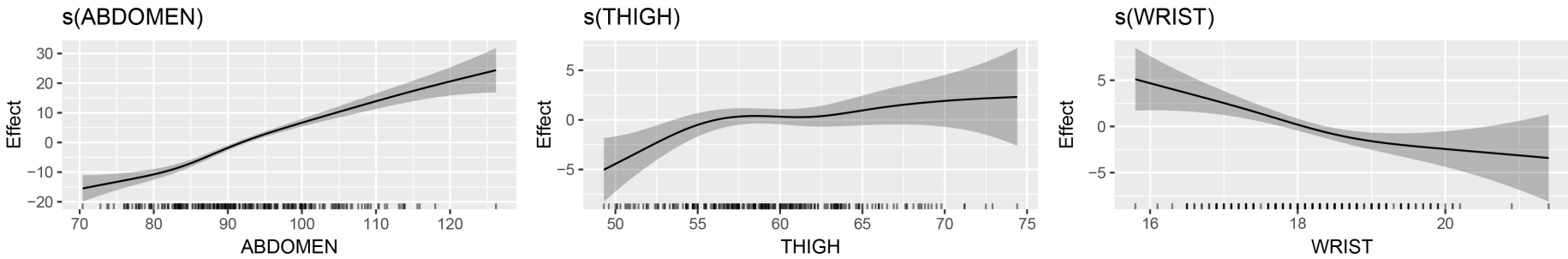


Figure: Estimation for Non-parametric part

- For Y axis in the figure, the abdomen has the dominant effect on bodyfat
- As thigh increase, increasing rate of bodyfat goes down to near the zero, and then slightly goes up as thigh reach around 63 cm.
- The bodyfat will go down as the circumference of wrist goes up.

Statistical Properties of Final Model

- Parametric Cofficients

Variables	EST	std.error	T	P-value
Intercept	26.03456	7.08668	3.674	0.000298
AGE	0.05996	0.02788	2.151	0.032563
WEIGHT	-0.10784	0.03207	-3.363	0.000906
FOREARM	0.33212	0.17217	1.929	0.054978

- Approximate significance of smooth terms

Terms	edf	F-statistics	p-value
s(ABDOMEN)	4.357	32.164	<2e-16
s(THIGH)	3.589	2.743	0.02784
s(WRIST)	2.215	6.006	0.00079

- Adjusted R-squared is 0.747

Example Usage of Model

Athletes	Fit	Average	Obese
6.5%-14.0%	14.0%-19.0%	19.0%-23.4%	23.4%-37.1%

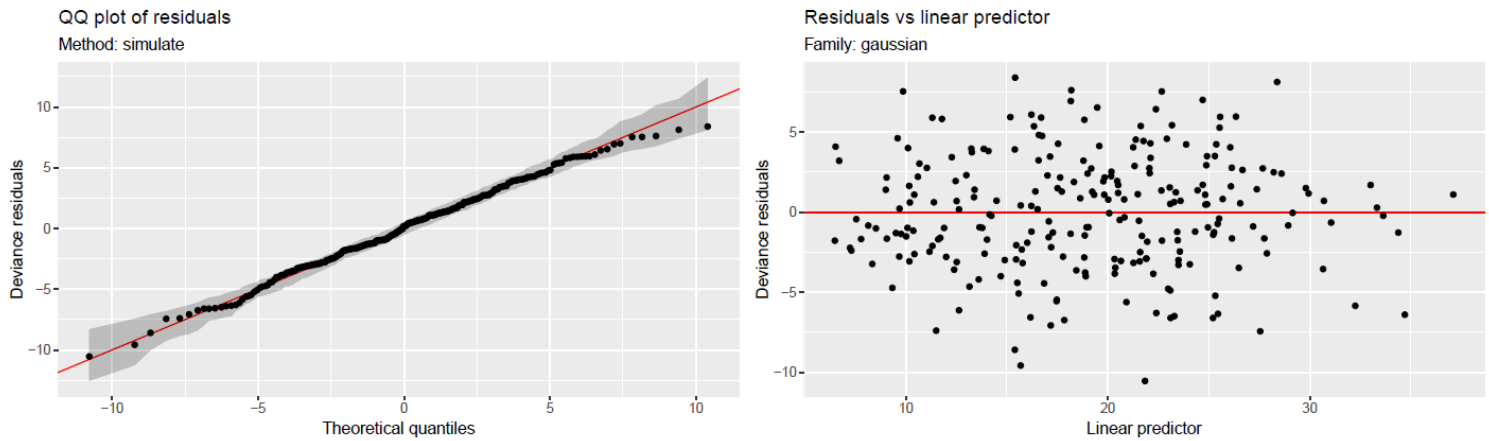
- One Example

  Given a man with 22 years old, 154 pls weight, 85.2 cm abdomen, 59 cm thigh, 27.4 cm forearm and 17.1 cm wrist, model gave a fitted value 15.74% bodyfat and a 95% confidence interval $[14.18, 17.31]$.

- Shiny

  https://wennroy.shinyapps.io/shiny/

Model Diagnosis



- Checked the residuals plot and residuals are indeed randomly normal distributed

Strength and weakness

Strength

- Final model has a great prediction on data
- Intuitive figure present.
- Variable Interpretations can be more precise.
- Capture the non linear trends.

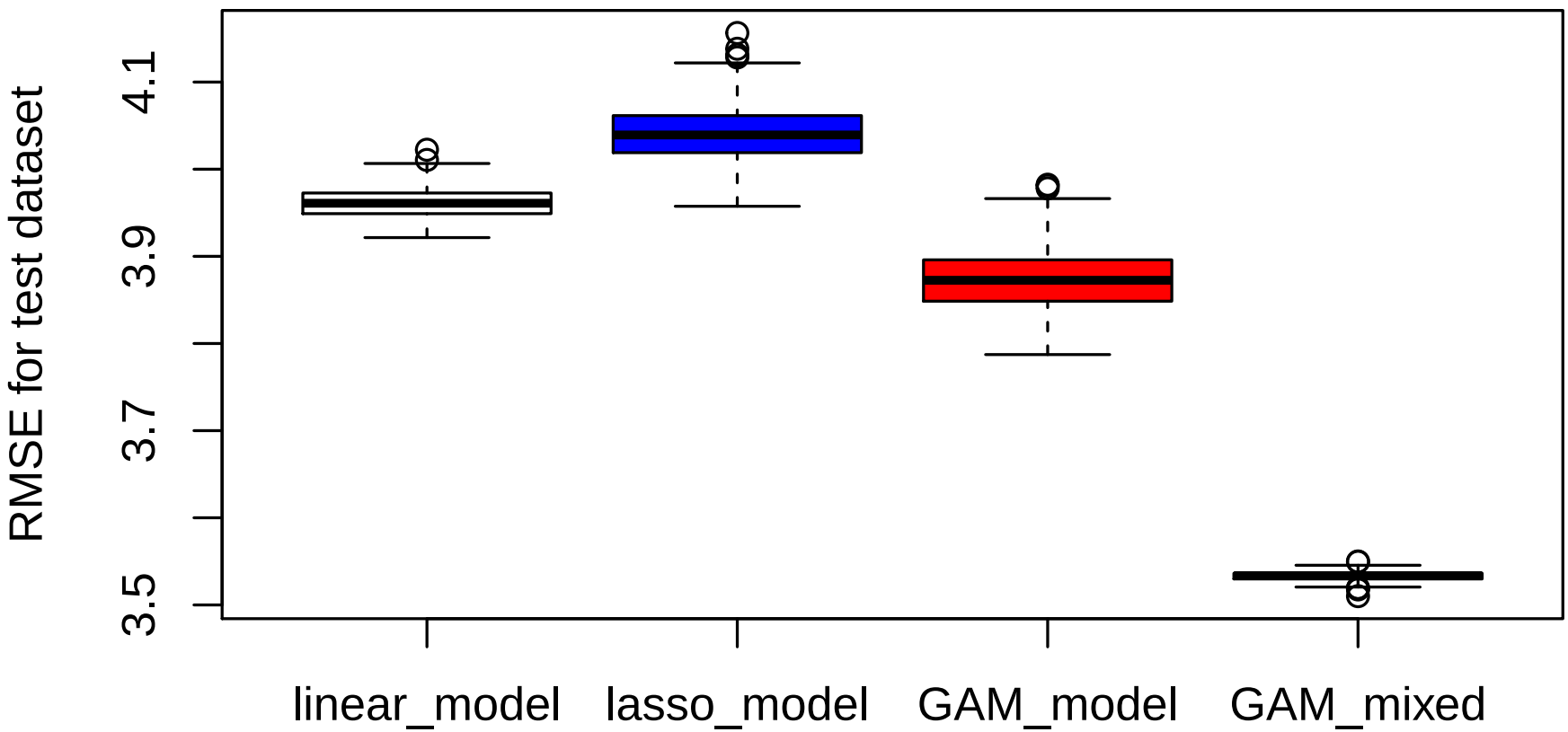
Weakness

- Some of predictors are highly correlated, then there might exist some interaction effects on Body Fat.
- Large computation should be done.
- Insufficient generalization ability.

Prospects

- Construct a GAM model considering joint distribution of predictors. (Intersection terms)
- Try our best to solve the correlation problem among predictors.

Appendix


RMSE for test dataset
3.5
3.7
3.9
4.1
linear_model
lasso_model
GAM_model
GAM_mixed

Thanks!