Group 4 Module 2 Summary

Introduction

Percentage of body fat (PBF) is an important indicator of a person's body health. It is desirable to have easy methods to estimate body fat. In this summary, our group is to find a model to estimate PBF using clinically available measurements. We came up with three models and finally decided to use a log linear regression model which estimates PBF accurately using two measurements.

Background Information of Data

The BodyFat.csv dataset comprises 252 records with measurements like age, weight, height, adiposity (calculated by BMI), and various body circumference data. Two key variables of interest are density, determined through underwater weighing, and PBF.

Firstly, we do some descriptive analysis in variables. As shown in the table:

	MIN	MEDIAN	MAX	MEAN	SD
PBF (%)	0.00	19.00	45.10	19.01	7.60
Weight (lbs)	118.50	176.50	363.10	178.33	26.76
Height (inches)	29.50	77.00	77.75	70.28	2.59
Abdomen Circumference(cm)	69.40	90.95	148.10	92.56	10.16
Wrist Circumference(cm)	15.80	18.30	21.40	18.23	0.90

From data description we know the relationships between PBF and density, adiposity and weight along with height, which are:$PBF = 495/DENSITY - 450$,

$ADIPOSITY = 0.4536 \times WEIGHT/(HEIGHT \times 0.0254)^2$

We detected potential outliers by comparing PBF and adiposity to the predicted PBF and adiposity deduced from input density, weight and height. In the first comparison group(left table), the 96th and 48th observations have significant differences between input and predicted PBF, and the 182nd observation has impossible PBF, so we deleted them. For points we could not judge whether they are outliers, we retained them.In the second comparison group(right table), the 42nd, 163rd and 221st observations have significant differences between input and predicted adiposity so we deleted them as outliers.We detected other potential outliers through boxplots.For these points we could not judge whether they are outliers, so we retained them.

Index	BODYFAT	bf_predict
96	17.3	0.3684833
48	6.4	14.1350211
76	18.3	14.0915057
182	0	-3.6116873

Index	ADIPOSITY	bmi_predict	WEIGHT	HEIGHT
42	29.9	165.62101	205	29.5
163	24.4	27.40739	184.25	68.75
221	24.5	21.67843	153.25	70.5
156	21.6	21.29397	171.5	75.25

Final Model Statement

Through fitting and comparing models, our final proposed models to predict PBF is:

$$PBF = 60.17 \times log(ABDOMEN) - 41.37 \times log(HEIGHT) - 77.08$$

That means, for example, a man with an abdomen circumference of 90 cm and a height of 77 inches is expected to have a PBF of 60.17×log(90)-41.37×log(77)-77.08=13.97(unit:%). Its 95% prediction interval is [5.82% , 22.13%]

Rationale for Final Model

We came up with three methods:

Model 1: we borrowed some idea from US Army, they have a formula for men to calculate body fat, it is like: $PBF \sim log\ (Waist - Neck) + log(Height)$

Because we do not have data of waist, we use abdomen instead of waist. The estimated model is

$$PBF = -9.237 \times log(HEIGHT - NECK) + 54.345 \times log(ABDOMEN) - 194.573$$

Model 2: Model 1 requires 3 variables. We tried to dismiss NECK and found that the result was even better. This is a log linear regression model. The log linear regression model is a transformed linear regression model and needs to comply the basic assumptions of linear regression rules. The estimated model is

$$PBF = 60.17 \times log(ABDOMEN) - 41.37 \times log(HEIGHT) - 77.08$$

Model 3: linear regression model. We limited the number of predictors within five. By exhaustive method we list all the combinations of measurements and find the best-performing model with the highest adjusted R-square. Then we got a linear regression model with five predictors:

$$PBF = -0.239 \times WEIGHT + 0.6994 \times HEIGHT + 1.2496 \times ADIPOSITY + 0.8199 \times ABDOMEN - 1.1357 \times WRIST - 74.3198$$

Through cross validation, the performance of three models are as follows.

	Avg adj r square	Avg mse	Sd mse
Model 1	0.6778	4.0393	0.7092
Model 2	0.7151	4.2525	0.2698
Model 3	0.7236	3.8491	0.2364

From the results, we could find model 2 and 3 have a better performance with higher adjusted R-square than model 1.Moreover, model 2 only has two variables, which is simpler than model 3, so we adopted model 2.

Model Diagnostics

We used Leverage plot and looked for extreme samples; See Fig 1, we found the 39th observation is an extreme sample and we removed this outlier. After rerunning the diagnostics checks, we got Fig 2.

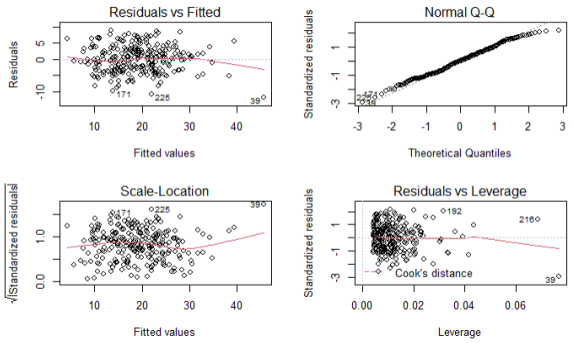

Fig 1: Before Removal

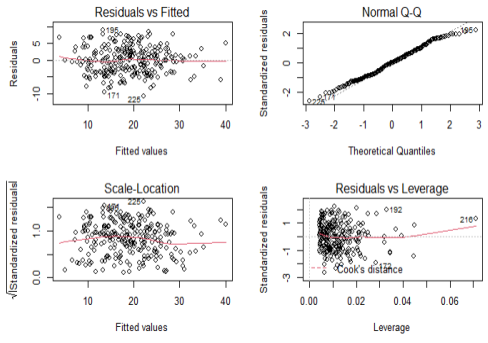

Fig 2: After Removal

In Fig 2, we could see our model fits well. The residual follows a normal distribution and the homoscedasticity assumption holds.

Model Strengths and Weaknesses

Strength: Our model is simple, while giving a fairly good R square.

Weakness： Inputs have a limited range. For example, when Abdomen is around 67, the PBF is close to 0, which is unreasonable.

Thoughts: Our model has a limited input range, how to improve it?

Reference

Contributions	Shijie Chen	Shuangyu Wang	Yuman Wu
Presentation	Responsible for feedback	Responsible for slides 1-4	Responsible for slides 5-12
Summary	Responsible for feedback and instruction	Responsible for Introduction, background information,rationale for final model	Responsible for cross validation,model diagnostics,model strength/weakness, discussion, contribution
Code	Responsible for code, uploading code to github	Responsible for organizing data and format	Responsible for organizing data and pictures
Shiny App	Responsible for code	Responsible for feedback	Responsible for feedback