Stat 628 Module 2: BodyFat

YUKUN FANG MENGKUN CHEN JIAYI SHEN

1 Introduction(YF)

A variety of popular health books suggest that the readers assess their health, by estimating their percentage of body fat. The goal of the project is to come up with a simple and accurate way of determining body fat percentage of males based on readily available clinical measurements.
Our rule of thumb:

$$\text{BodyFat} = -29.7950 + 0.14W + 0.04C + 0.05A + 0.03H + 0.02T + 0.17\text{AGE}$$

W stands for weight, **C** stands for chest, **A** stands for abdomen, **H** stands for hip, **T** for thigh.

2 Background(YF)

The data set contains measurements from 252 men who had their body fat percentage accurately measured via underwater weighing. Notice that the outcome variable is body fat percentage and the set of predictors are every variable except ID number, body fat percentage, and density.

3 Motivation and Statement(YF)

3.1 Motivation

In the beginning we have used linear model step selection by using AIC or BIC to find the best model, which may lead to multicollinearity during fitting the model. So we use the principal component analysis to find the PC(principal components), and then use PCs to find linear model.

3.2 Statement of the Model

First, we calculate the covariance matrix of the data set for 15 variables without the three. Then we calculate the eigenvalues and the eigenvectors of the matrix. Then calculate the cumsum of the eigenvalues and divided by sum of eigenvectors and equals to cumvariance. This value is the main result which is used to find the principle components.
Second, we plot the cumvariance and select the top two eigenvalues as the PCs: PC1 and PC2. Then we need to remove the variables whose coefficients are too small with big p-value for t-statistics.
At last, we use PCs as variables to fit the linear model to diagnose and remedy it.

4 Estimation and Inference of Parameters(YF)

After the calculation, we get the 2 PCs:

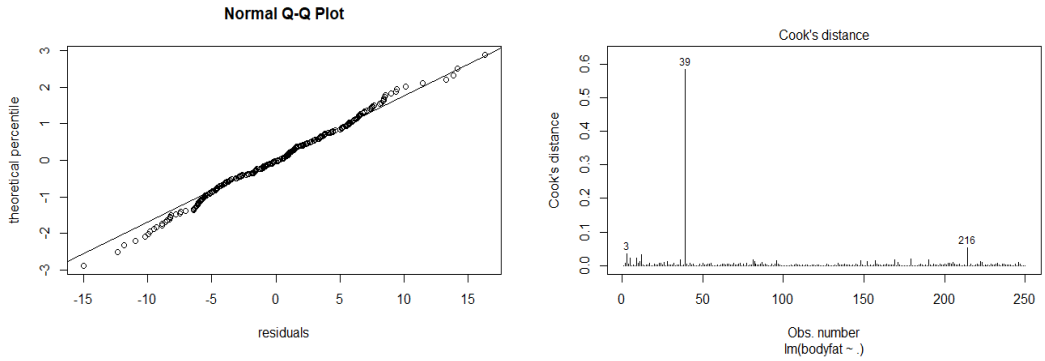
$$PC_1 = 0.88W + 0.24C + 0.30A + 0.20H + 0.14T$$

$$PC_2 = 0.96\text{AGE}$$

Then fit the model and get a not good model. After we diagnose and remedy it, we change PCs by the primary variables and get the final model.

5 Model Diagnostics(YF)

We draw these two plots and find that there exists one outlier NO.39.



After removing that, we refit the model and get the result:

$$\text{BODYFAT} = -29.7950 + 0.1644PC_1 + 0.1759PC_2$$

And the summary shows that is fits well. We can get r-square=50% and residual table

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-27.5858	3.0069	-9.17	$< 2e-16$
pc1	0.1644	0.0115	14.25	$< 2e-16$
pc2	0.1759	0.0282	6.25	$1.83e-09$

So the final model:

$$\text{BODYFAT} = -29.7950 + 0.14W + 0.04C + 0.05A + 0.03H + 0.02T + 0.17\text{AGE}$$

6 Strengths and Weakness(YF)

- **Strengths**: This method could be wildly used for any other problem and be robust to multicollinearity. Also it is easy to understand and get the result.
- **Weakness**: If someone has some body statistics out of range, it may cause bad result. Also the R-square is not close to 1 which means we need to improve the model.

7 Conclusion

Summary: Yukun Fang Code: Mengkun Chen Slides: Jiayi Shen