STAT 628 Module 2

Lu Li*, Jiawei Wu*, Nan Yan*, and Youhui Ye*

*Department of Statistics

1 Introduction

As obesity becomes a common phenomenon nowadays, a measure of fitness level is needed. Body fat is exactly the one which suits for the purpose well. However, due to the limitation of realistic conditions, it is hard to evaluate the body fat percentage directly. In such a scenario, we are trying to find a simple but also powerful linear model to predict body fat percentage accurately with a set of easily obtained body measurements, such as heights, weights and etc. Step-wise selection method and exhaustive search method are used to find out the optimal model, implemented with AIC, BIC and Mallow's Cp criterions respectively. Finally, we determined our optimal model by comparing different models' R-squares.

2 Background Information

One adequate method is to obatin the body fat by the body density. Actually, there exists a simple equation between the percentage of boday fat, body density and densities of different body tissues. However, it is also very hard to directly get our body density. Thus, we wish to use some simply available body circumference measurements to predict the percentage of body fat.

The data set we use satisfies all of our described purposes. It consists of 252 men with measurements of their body fat percentages, body densities and multifarious circumference measurements. Since men and women have different body structures, it would be inaccurate to use our formula or calculator on women's body fat percentage estimation.

3 Construct the Models

3.1 Data cleaning

First, consider extreme values in the boxplot and samples whose BODYFAT value is far away from the function $B = \frac{459}{D} - 450$ as potential outliers. In this process, we modify 48th data with the relationship between BODYFAT and DENSITY, delete 182nd,216th data.

Then, fit a original linear regression model, delete points with high Cook's distance and corresponding standardized residual whose absolute value is larger than 2.5. In this part, data points with index {39,42,86,220} are removed.

3.2 Motivation and Statement of Models

We firstly construct a regression model with all variables as predictors in the data set exclude the Bodyfat. However, with p-values, we could see that there are some insignificant variables and also by com-

puting out the VIF values, there exists significant multicollinearity within the variables. By this reason and the benefits of less variables, we would like to do the variable selection.

As stated below, AIC, BIC and Mallows'Cp are three statistics used to balance the number of parameters and the prediction errors. And we use stepwise method and exhaustive search method to search for the best subsets, implemented with AIC, BIC and Mallows'Cp. To be detailed, our first model is built under the criterion AIC and stepwise method with both, backward and forward direction. The second model is similar to first model but with BIC criterion. For the lase one, it is built under the criterion Mallows'Cp and exhaustive search method.

$$AIC = 2k - 2\ln L; BIC = k\ln n - 2\ln L \ ; C_p = \frac{SSE_p}{MSE(full)} - (n - 2p)$$

where $\hat{L}$ is the likelihood function of $\{x_1, ...x_n\}$ under normality assumption.

After splitting our data set into train set and test set with a proportion 3:1, we get three selected models. By comparing their performance on the train set and test set like R^2, adjusted R^2, MSE, the largest VIF value and number of predictors, we finally choose the model with three significant variables which are Weight, Abdomen and Wrist. Then refit this model on the whole data set.

4 Statistical Analysis

4.1 Diagnostics

Our final model has been stated below. The Multiple R-squared is 0.7180797 which means about 72% variance could be explained by the predictors. The Adjusted R-squared is 0.7145848 and MSE is 15.38403. Since the p-value for the whole model is extremely less than α=0.05, so we may reject the null hypothesis and conclude that the model is significant. And the p-values for the individual coefficients are all smaller than α=0.05, so we may conclude that all the coefficients are significantly not equal to zero.

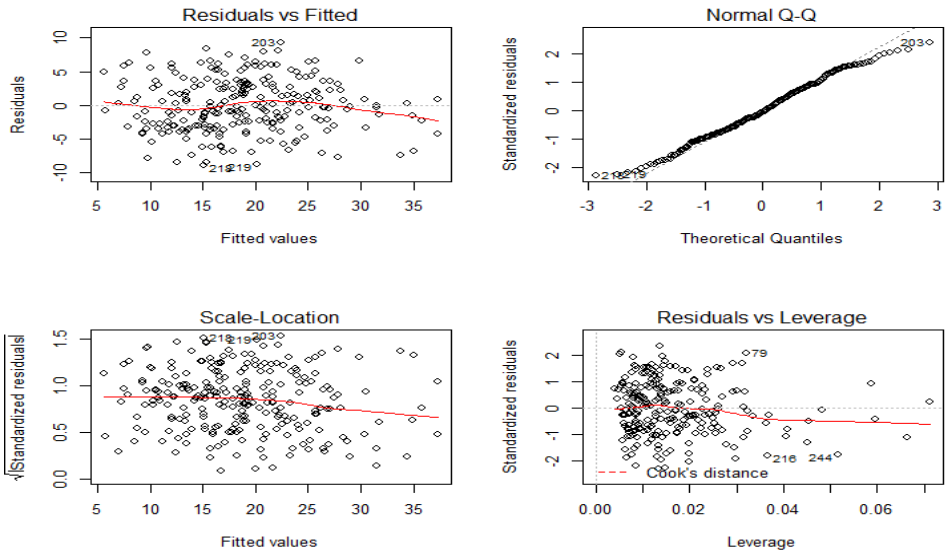
4.1.1 Model estimation

R^2_{adj} 0.715	R^2 0.718	MSE 15.638

—	Estimate	Std.Error	Pr(>|t|)	2.5 %	97.5 %	int.length
(Intercept)	-23.868	6.204	1.531e-04	-36.090	-11.646	24.443
ABDOMEN	0.870	0.052	1.996e-42	0.768	0.972	0.204
WEIGHT	-0.084	0.022	2.254e-04	-0.127	-0.040	0.088
WRIST	-1.245	0.401	2.129e-03	-2.036	-0.455	1.580

4.1.2 Diagnostic plot

From the first and third plot, we may belive that the constant variance assumption is true and the fitted values are independent from residuals. With the second plot, the residuals are roughly normaly distributed with a thinner tail than the normal distribution. From the last plot, there is no significant leverage point.



4.2 Interpretation

We can interpret each predictors X_i with coefficient β_i in this way:when the others fixed, BODYFAT changes with value β_i as X_i changes 1 unit. It can be checked that ABDOMEN is the most important predictor for the body fat, and it can be linear fitted by the other two variables WEIGHT and WRIST with $R^2 = 0.75$. But considering our aim is to do the prediction nor the explanation and with these two variables the performance of our model will be better, we will remain the WEIGHT and WRIST variables. Thus, although it seems strange BODYFAT has a negative linear relationship with WEIGHT and WRIST, we may interpret this phenomenon in such way: as coefficient of ABDOMEN is larger than the scenario when it's the only predictor, WEIGHT and WRIST sort of provide some corrections, which may work like penalty.

According to length of confident intervals of predictors, estimated coefficients are acurate excpet β_0.

4.3 Strength and Weakness

- Strength: The model makes a good trade off between accuracy and computational efficiency.
- Weakness: For people who don not know their ABDOMEN and WRIST, it's hard for them to get a precise outcome.

5 Conclusion

Since the R^2 is greater than 70%, it is resonable for us to conclude that the percentage of body fat can be predicted by weight,abdomen and wrist to some degree.

6 Contributions

Youhui and Jiawei contributed to the data cleaning and the corrsponding writing part of summary and presentation slides, as well as the creation of our shiny APP. Lu Li and Nan Yan were responsible for the modeling and inference part with corresponding writing part of summary and presenatation sildes.