Body Fat Analysis

Introduction

The body fat percentage (BFP) of a human or other living being is the total mass of fat divided by total body mass, multiplied by 100. It is a measure of fitness level, since it is the only body measurement which directly calculates a person's relative body composition without regard to height or weight. While the body fat percentage plays an important role in various health outcomes such as life expectancy, prognosis for disease, healthcare costs, and the general well-being of individuals, it is quite difficult and costly to calculate: one of the more accurate methods requires underwater weighing. Thus, it is very practical to come up with a simple, accurate and precise method to estimate body fat percentage.

We can use multiple regression to determine the prediction equations for the determination of body fat percentage. Since the data set contains various body measurements and an accurate estimate of the percentage of body fat from Siri's equation, body fat percentage can be fit to the other measurements using multiple regression, giving a useful predictive equation for doctor to estimate people's body fat percentage. The various measurements other than body fat are ones that are easy to obtain and serve as proxies for body fat, which is not so easily obtained.

The data set we used to analyze contains age, weight, height and 10 circumference measurements on 252 men. Each man's percentage of body fat was accurately estimated by an underwater weighing technique.

Data preprocessing

According to the Siri's equation $100\mathrm{B} = \frac{495}{D}$-450, there should be a linear relationship between the two variables body fat and density. Since we are interested in body fat, we select it as the response variable.

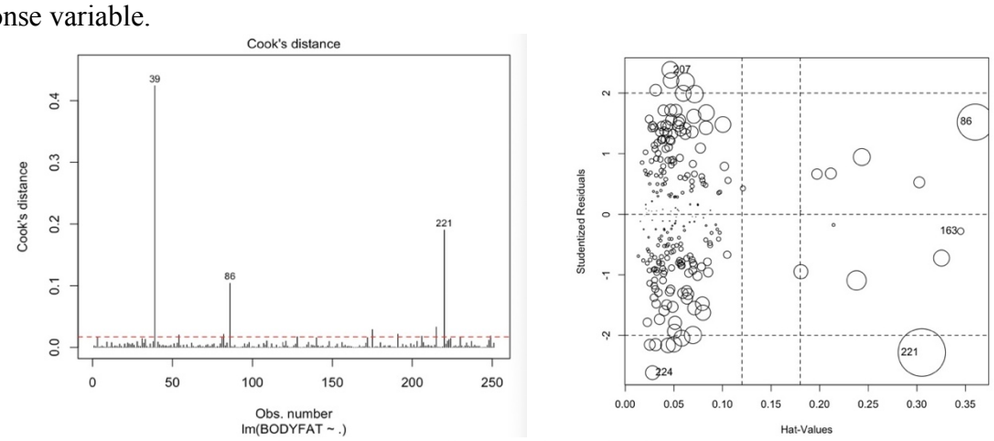

Figure 1: Cook's distance plot and Influence plot

Firstly, we look through the summary statistics of all variables. We found the there are some abnormal observations. The body fat of observation 182 is 0 which must be an error and this could be handled by computing from the Siri's formula, which should be -3.6. Since there may exist an record error, we decide to delete it. The density of observation 216 is 0.995 smaller than the density of water. Since we do not use density and the other features are normal, we just keep this observation. The height of observation 42 is 29.50 which is so small, but the other features seem quite normal. We consider to impute it by BMI calculation, which should be 69.50.

Secondly, we fit the linear regression with all observations and then use Cook's distance to find the influential points of the model. From the Cook's distance plot[1], we identify observation 39 is obvious influential point. The subject weights 363 pounds which is very abnormal and all the other features are also quite big, for instance the hip, chest, abdomen circumstances. Then, we consider to delete it.

After the data cleaning, we draw an influence plot. According to the influence plot, the studentized residuals are in the range [-3,3] and the hat values are small. There are no obvious outliers and influential points. Hence we could do next step to construct the linear regression model.

Multicollinearity

To determine which explanatory variables I need to construct the regression model, first I calculate the

variance inflation factors of each X_k, k=1,2,3,...14. A rule of thumb is that if the mean variance inflation factor is greater than 1, then there might exist multicollinearity. For each inflation factor VIF , if it's

greater than 10, the multicollinearity might have a significant influence. In this dataset, by using the VIF function in R's "car" package. We may gain VIF shown below.
Because mean VIF is 22.18 greater than 1. Hence, there might be the multicollinearity problem. From the table of VIF s shown above, we could identify the individual factor VIF of WEIGHT,HEIGHT and ADIPOSITY are much greater than 10, which indicates that multicollinearity may have a large impact on the inference.

Model Selection

Next, in order to reduce multicollinearity, we need to reduce the model parameters. There are four famous criteron to select the predictor variables. The R^2 criterion, adjusted-R^2criterion, Mallow's C_p criterion, BIC/AIC criterion. The smaller Cp, BIC/AIC and the bigger adjusted-R^2criterion and R^2, the better the performance of the model fitting.
I first use the "regsubset" function in R's "leaps" package to gain the best selection model with each best selection size 1 and selection among maximum size 14 of subsets to be examined.

Model: BODYFAT ~ WEIGHT + ABDOMEN

Coefficients	Estimate	Std.Error	t value	Pr(>|t|)
Intercept	-42.50	2.46	-17.29	< 2e-16
WEIGHT	-0.12	0.020	-6.30	1.39e-09
ABDOMEN	0.90	0.052	17.42	< 2e-16

Residual standard error: 4.06 on 247 degrees of freedom
Multiple R-squared: 0.7192, Adjusted R-squared: 0.717
F-statistic : 316.4 on 2 and 247 DF, p-value : < 2e-16

Figure 2: Criterion table and Model output

From the table shown above (Figure 2 Criterion table), we could identify that when 2 variables are selected based on the exhaustive search, the change of R^2 criterion, adjusted-R^2, C_p , BIC criterion is the largest, indicating that select two variables would be enough to construct a suitable model. Based on the "rugsubset" function, the model is BODYFAT~WEIGHT+ABDOMEN.
Next we would use t test to check the significance of the coefficients (Figure 2 Model ouput). Because the p value of both the 2 variables a much smaller than 0.05 and therefore the coefficients are significant. The model is effective.
Then I use the "step" function in R and BIC criterion.The reason that I choose BIC criterion is that AIC will tend to overfitting. The direction of this process is "both" instead of exhaustive search we implemented before, which means we either add or delete the explanatory variable and compare BIC among them and choose the smallest BIC to gain the best model fitting explanatory variables. The result after a number of steps of choosing is BODYFAT ~ AGE + ADIPOSITY + CHEST + ABDOMEN + WRIST.

Diagnostics

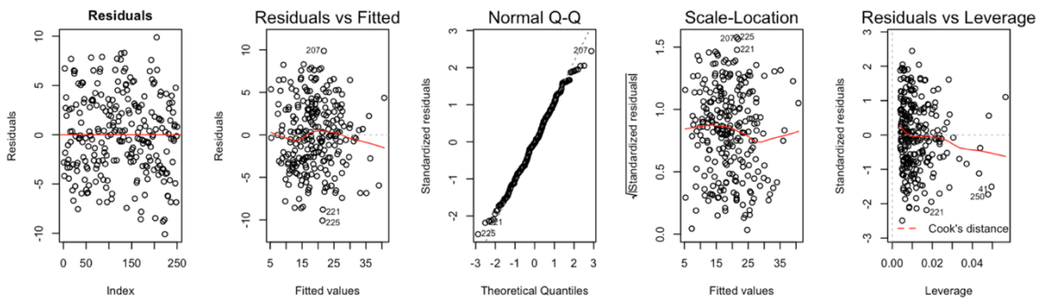

Figure 3: Diagnostic plots

From the plot shown above, for Residudals vs Index plot both of them are patternless and therefore satisfy the independence. For Residuals vs Fitted plot, the vertical distance from the points to x axis are both nearly constant and patternless and therefore satisfying the homoscedasticity. For the Normal Q-Q plot the residual points of both models are almost in a line and thus the normality assumption should not be violated.

Relation Between Bodyfat and BMI

Many papers talk about the relation between bodyfat and BMI and then they use BMI to assess bodyfat. Compared with the method mentioned above, the method on BMI is easier to measure and it is more convenient in daily life. Thus, next we compare the result of two models.
The article by Terence C. Mills provides a semi-logarithmic relation between bodyfat and BMI. Thus, we explore the relation between BODYFAT and ln(BMI).

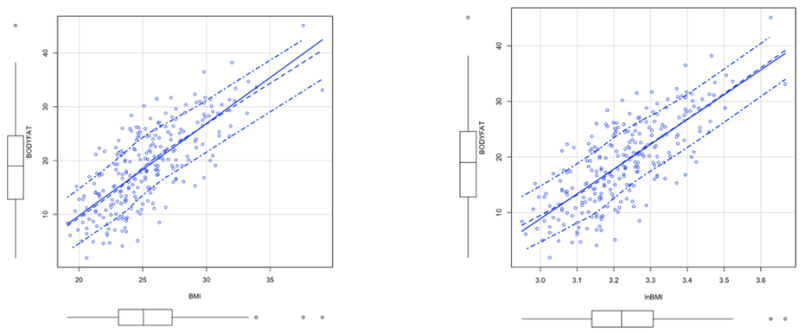

Figure 4: BMI plot and ln(BMI) plot

In addition, N. K. Mungreiphy, June Stevens and Mott JW have talked about the relation among bodyfat, BMI and age. And, the U.S. National Institutes of Health and the Centers for Disease Control and Prevention provide the criterion about BMI by 10 on the age. Thus, we rescale the age to ordinal level by 10 and add the covariate to our model, and conduct a linear model with ln(BMI) and Age_level.

Coefficients	Estimate	Std.Error	t value	Pr(>|t|)
Intercept	-117.22	8.01	-14.61	< 2e-16
ln(BMI)	42.54	2.47	17.21	< 2e-16
30-40	7.20	2.90	2.48	0.014
40-50	0.34	2.72	0.13	0.90
50-60	1.60	2.23	0.72	0.47
60-70	-0.94	1.72	-0.54	0.59
70-80	0.037	1.21	0.03	0.98
80+	1.07	0.85	1.27	0.21

Figure 5: ln(BMI)+age model output

Diagnostics

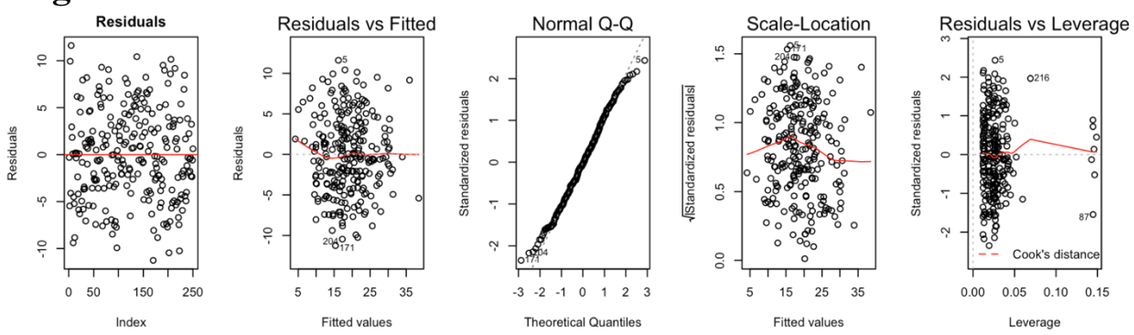

Figure 6: Diagostics plots

From the plot shown above, for Resudals vs Index plot both of them are patternless and therefore satisfy the independence. For Residuals vs Fitted plot, the vertical distance from the points to x axis are both nearly constant and patternless and therefore satisfying the homoscedasticity. For the Normal Q-Q plot the residual points of both models are almost in a line and thus the normality assumption should not be violated.

Compare two models

Here, we get two models, the one is A：BODYFAT ~WEIGHT + ABDOMEN and the other is B：BODYFAT ~ ln(BMI) + Age_level. We split the dataset to train dataset and test dataset to test the robustness of the two model.

Onmi Model	4.42
BODYFAT~WEIGHT+ABDOMEN	3.31
BODYFAT~lnBMI + Age_level	3.90

Figure 7: Model Comparison

We use Mean absolute error to test three models. From the result (Figure 7), we can find these two models perform better than Onmi Model in the dataset. For our two model, Model A is more accurate and Model B is much easier to measure. So, we provide these two model for doctor.

Model A: BODYFAT= -42.5-0.12WEIGHT+0.90ABDOMEN

Model B: BODYFAT= -117.22+42.54ln(BMI)+7.20 Age_level(30-40)+0.34 Age_level(40-50)+1.60 Age_level(50-60)-0.94 Age_level(60-70) +0.037 Age_level(70-80)+1.07 Age_level(80+)

Contribution

Shirley Zhang: Construct the linear regression model including the multicollinearity test, the model variables selection based on the common 4 criterions using the exhaustive search and the stepwise search, do model diagnostic and finally help summarize the results from the whole group all together.

Reference

[1]. Shalabh, IIT Kanpur.Regression Analysis. 2002