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. While the body fat percentage plays an important role in various health outcomes such as life expectancy, healthcare costs and the so on, it is quite difficult and costly to calculate. So it is practical to come up with a simple and accurate method to estimate body fat percentage.

We can use multiple regression to determine the prediction equations for the determination of body fat percentage. 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 Siri's equation.

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.

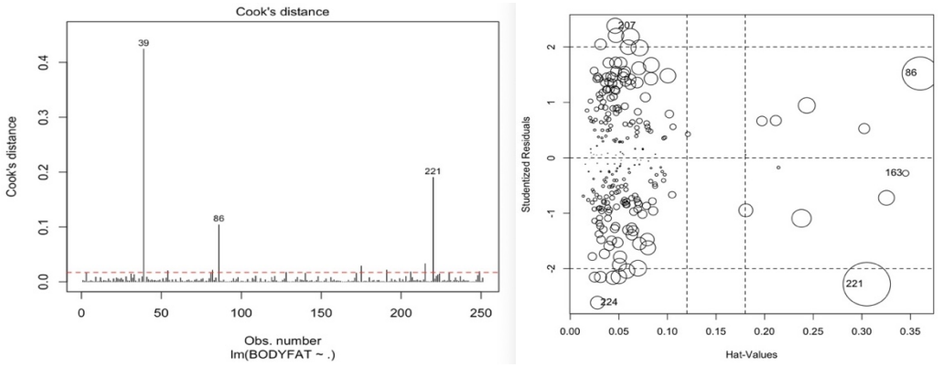

Figure 1: Cook's distance plot and Influence plot

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. The studentized residuals are within [-3,3] and the hat values are small. There are no obvious outliers. Then 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 VIFs 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 criterions 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.

We 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.

rsq	adjr2	cp	bic
0.6742	0.6729	60.1617	-269.3042
0.7192	0.717	19.8244	-300.9901
0.7314	0.7282	10.3594	-306.576
0.7364	0.732	7.7344	-305.6788
0.7395	0.7342	6.7556	-303.1817
0.7422	0.7358	6.2277	-300.2556
0.7444	0.737	6.2044	-296.8312
0.7462	0.7377	6.5122	-293.0772
0.7473	0.7379	7.4015	-288.7226
0.7482	0.7377	8.5727	-284.0755
0.7492	0.7376	9.666	-279.514
0.7497	0.737	11.2284	-274.4572
0.7498	0.7361	13.0652	-269.1093
0.7499	0.735	15	-263.6572

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 selecting two variables would be enough to construct a suitable model. Based on the "regsubset" function, the model would be BODYFAT ~ WEIGHT +ABDOMEN.

Next we would use t test to check the significance of the coefficients (Figure 2 Model output). 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 we use the "step" function in R^2 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

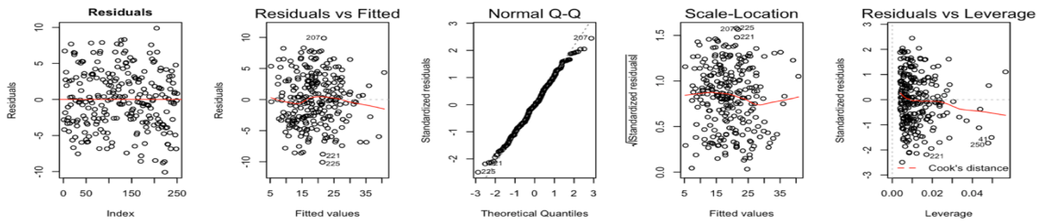

Figure 3: Diagnostic plots

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.

Relationship Between Bodyfat and BMI

Many materials have talked about relationships between BODYFAT and BMI, like Terence C. Mills provides a semi-logarithmic relation between bodyfat and BMI, and BMI is much easier to measure than ABDOMEN. Thus, we consider lnBMI to simple our model by lnBMI.

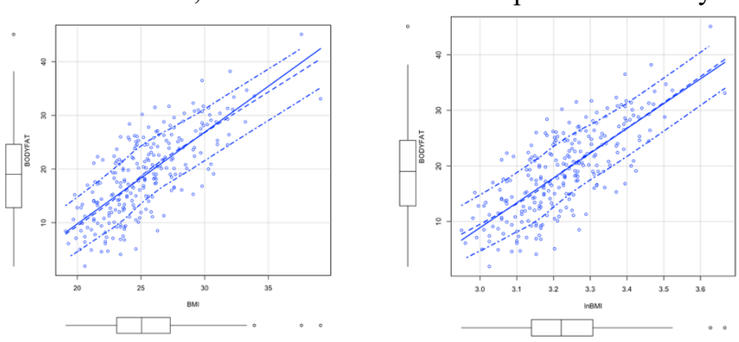

Figure 4: BMI plot and ln(BMI) plot

From the scatter plot of BMI and lnBMI, we can find linear relation between BODYFAT and lnBMI is much better. Then, we need to test whether lnBMI can explain ABDOMEN well.

Coefficients	Estimate	Std.Error	t value	Pr(>|t|)
Intercept	-142.54	6.63	-21.52	< 2e-16
ln(BMI)	72.85	2.05	35.49	< 2e-16

Multiple R-squared: 0.8355, Adjusted R-squared: 0.8349

Figure 5: ABDOMEN ~ ln(BMI) output

We can see R^2 statistics is 0.8355, which means lnBMI can explain most information of ABDOMEN. So we can use lnBMI. And for reducing approximation error, we use WEIGHT and HEIGHT to construct a linear model.

Coefficients	Estimate	Std.Error	t value	Pr(>|t|)
Intercept	347.72	48.03	7.24	5.67e-12
ln(WEIGHT)	44.44	2.53	17.55	< 2e-16
ln(HEIGHT)sq	-50.49	5.10	-9.91	< 2e-16

Multiple R-squared: 0.5558, Adjusted R-squared: 0.5522

Figure 6: BODYFAT~ ln(WEIGHT)+ln(HEIGHT^2) output

The R^2 is 0.5558, which is not bad under only one variable. So we get the model BODYFAT ~ ln(WEIGHT) + ln(HEIGHT^2). Next, we check the model assumption.

Diagnostics

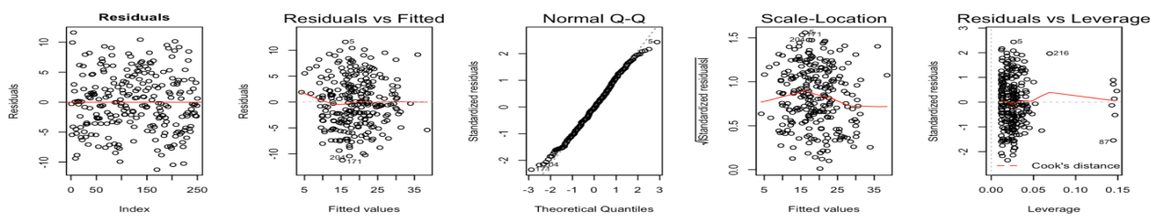

Figure 7: Diagnostics plots

From the plot shown above, for Residuals 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.

Model Evaluation

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.

ABFP Model	3.88
Model A: BODYFAT~WEIGHT+ABDOMEN	3.29
Model B: BODYFAT~ln(WEIGHT) +ln(HEIGHT^2)	4.14

Figure 8: 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.

Summary

From the analysis above, we decide these three variables: WEIGHT, HEIGHT and ABDOMEN. Under the consideration of convenience, we will use model Model B: BODYFAT=347.72+44.44 ln(WEIGHT) -50.49 ln(HEIGHT^2)+ ε. And if we want a more accruate result, we will use Model A: BODYFAT= -42.5-0.12WEIGHT+0.90ABDOMEN+ ε, where $\varepsilon \sim N(0, \sigma^2)$.

For model A, the mean change in BODYFAT will be 0.12% per pound change in WEIGHT,will be 0.9% change per cm change in ABDOMEN, holding each other predictor fixed.For model B, the mean BODYFAT will be 44.439% change per exponential pound unit change in WEIGHT, will be 50.491*2% change per exponential cm unit change in height.

Contribution

He Wang: He was responsible for the data pre-progressing. He wrote the introduction, background and detected the inaccurate records and outliers in the raw data. What's more, He finished the same part in the code, slides and summary.

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.

Chushi Shi: Analyze the relation between bodyfat and BMI including constructing weight and height model to simplify the original model, and model evaluation for all models.

Siqi Shen: He was responsible for the construction and deployment of Shiny. He also created the repository in Github and manage this daily.

Reference

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

[2]. AceFitness https://www.acefitness.org/education-and-resources/lifestyle/blog/112/what-are-the-guidelines-for-percentage-of-body-fat-loss

[3]. Terence C. Mills, Predicting Body Fat Using Data on the BMI. 2005

[4]. Army Regulation 600–9, Army Regulation 600–9. 2013