Body Fat Calculator

Hao Tong, Yuan Cao, Shushu Zhang

1. Introduction

Body fat is a highly specialized organ, critically important for health and longevity [1]. It can be one of the most useful index for determining health and achieving fitness goal of an individual, even more than BMI. However, measurement techniques such as underwater weighing and air displacement plethysmography are always cumbersome to apply. Thus, it is meaningful to come up with a easier method to estimate the body fat with clinically available measurements, like body weight. Therefore, we are given a data set containing 252 observations whose body fat is accurately measuared and calculated by underwater weighing, together with 14 accessible predictors.

2. Overall Findings & Rule of Thumb

Regressed from the 14 predictors in the data set, "Bodyfat" can be estimated as

$$\begin{aligned}\mathrm{Bodyfat} = -28.81 - 0.13\mathrm{WEIGHT} - 0.38\mathrm{NECK} + 0.94\mathrm{ABDOMEN} + 0.23\mathrm{BICEPS} \\ +0.41\mathrm{FOREARM} - 1.14\mathrm{WRIST}.\end{aligned} \tag{1}$$

In this rule of thumb, the predictors are easily obtained in daily life. These six measurements are correlated to some extent. For example, one's abdomen circumference will typically increase as he gains weight. However, to obtain more accurate estimation of bodyfat, we typically need to know how much abdomen circumference has been increased per pound he gains. In the extreme case, if a men gains 1lb in weight but does not have any change in other variables sugguesting that he is building muscles instead of gaining fat, the model indicates that he will lose 0.13 in bodyfat.

3. Statistical Analysis

To obtain the aforementioned rule of thumb, we need to first clean the data in Section 3.1, and then fit a proper model in Section 3.2.

3.1. Data Cleaning

In current context, it is natural to detect outliers by each variable seperately using techniques such as box plot. However, in data cleaning, we intend to impute the outliers that rise due to human errors instead of natural deviations in population distributions. In our case, the predictor variables have great redundancy (i.e. highly related) in this case. For example, "ADIPOSITY", "WEIGHT", and "HEIGHT" are strictly related with the following equation

$$\mathrm{ADIPOSITY}(\mathrm{bmi}) = \mathrm{BODYWEIGHT}(\mathrm{kg}) : \mathrm{BODYHEIGHT}(\mathrm{m}^2).$$

Anomalous observations that are supposed to be imputed or removed in advance, must have some inconsistency in different variables, while univariate outliers does not necessarily indicate anomalous observations. For example, observation No. 39 is way outlined in every variable, which will definitely be considered as an outlier if we consider variables seperately. However, it deviated in almost every variable at a similar degree, making it a reasonable observation of a highly overweight and obese person, thus the observation shouldn't been imputed.

In order to detect the anomalous observations to be imputed based on the aforementioned criteria, we introduce multivariate detections, namely, fitting linear regressions for every variable with respect to other related variables using Residuals or Cook's Distance as a anomalous measure. Then, we concentrate on the extreme (maximum/minimum) values of each variable to make sure they make sense with practical meaning of the certain variable. To be more specific, in step 1, we fit "DENSITY" with all other variables except "IDNO" and "BODYFAT" to make sure there is no obvious observation error in "DENSITY". Then, we fit "BODYFAT" with "DENSITY", resulting in two outliers *No. 48* and *No. 96*. In addition, we detect the relationship among "ADIPOSITY", "WEIGHT", and "HEIGHT" by the aforementioned equation, leading to an abnormal observation *No. 42*. Then, the rest of the variables are regressed similarly with all other variables except "IDNO", resulting in anomalous observations of *No. 31* and *No. 86* for "ANKLE", *No. 175* for "FOREARM". We impute all of the aforementioned anomalous observations with fitted values in the linear model or the defined equation. We can even infer the error mechanisms for some imputations. No. 42, which has anomalously low value only in "HEIGHT", can be imputed by the induced (or say redundent) variable "ADIPOSITY" defined as $\mathrm{ADIPOSITY(bmi)} = \mathrm{BODYWEIGHT(kg)} : \mathrm{BODYHEIGHT(m^2)}$. As a result, "HEIGHT" OF No. 42, 29.50, is replaced by 69.45, indicating 6 is probably mistyped by 2. Note that No. 163 and No. 221 also have slight mismatches in the relation of "ADIPOSITY" with "WEIGHT" and "HEIGHT". But all three variables are in reasonable ranges, making it hard to determine the "bad" observations. Therefore, we determine to leave as they are.

In step 2, we use boxplot to detect the outliers in the target variable "BODYFAT", resulting in finding the higher-end outlier *No.216* and lower-end unreasonble point *No. 182* with zero body fat. Owing to the uncertainty of the anomalous mechanism, we simply remove these two observations.

3.2. Model Selection

In choosing the appropriate kind of model for handling this problem, linear regression outperforms several advanced machine learning models such as neural network in terms of both statistical accuracy and computational costs. For predictor selections, we use "olsrr:ols_step_best_subset" function to obtain a subset of predictors that contribute to a large R^2 value with less variables. As a matter of fact, "abdomen" alone contributes to 0.63 R^2, making it the most important predictor among all, followed by "weight" variable. For better statistical accuracy, we finally use "WEIGHT", "NECK", "ABDOMEN", "BICEPS", "FOREARM", "WRIST" these six variables as predictors in linear regression model with R^2 0.74, and other estimation and inference information listed in Table 1.

4. Diagnostics

For the linear model, we have the following assumptions: (1) observations are independent; (2) homoscedasticity; (3)linearity of the model; (4) error terms are normally distributed. The qq plot

Table 1. Estimated Coefficients

	Estimate	Std. Error	t value	p value
(Intercept)	-28.80735	7.13151	-4.039	7.17e-05 ***
WEIGHT	-0.12741	0.02557	-4.982	1.19e-06 ***
NECK	-0.37724	0.20620	-1.829	0.06854 .
ABDOMEN	0.93666	0.05212	17.971	< 2e-16 ***
BICEPS	0.22729	0.15156	1.500	0.13498
FOREARM	0.40741	0.18000	2.263	0.02448 *
WRIST	-1.14274	0.43254	-2.642	0.00877 **

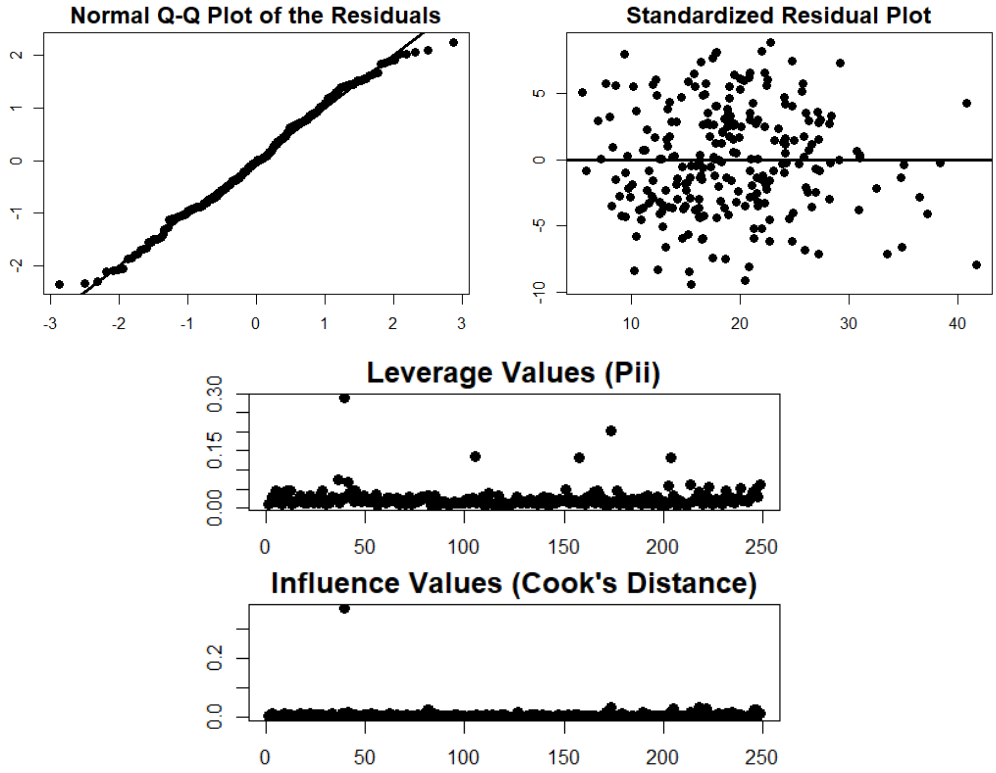

Figure 1. Diagnostics for Final Model

Section 4 indicates that the residuals which are the estimations of error terms are approximately normally distributed. The standardized residual plot in Section 4 suggests that the model is linear and homoscedastic, and there is no obvious outliers in the model, demonstrating the effectiveness of data cleaning. Based on Pii measure and Cook's distance in Section 4, we notice there are several leverage values and influence values, indicating there are points contributing to the model more than other points. However, they are not outliers. In conclusion, the assumptions of the model are satisfied.

5. Strengths and Weaknesses

The strengths of our model lies in the interpretability, computation efficiency and predictability (i.e. statistical accuracy) as mentioned above. On the other hand, although the linearity assumption

is approximately satisfied in Section 4, these variables may not be linearly related to bodyfat in practice.

6. Conclusion

As we come to a conclusion, we can use Equation (1) to calculate our body fat where abodomen circumference is the most importance predictor, followed by weight.

7. Contributions

SZ:

References

[1] David Ludwig. *Always Hungry?: Conquer Cravings, Retrain Your Fat Cells, and Lose Weight Permanently*. Hachette UK, 2016.