

Group 4 Body Fat Prediction

Chao Chang, Yezhou Li, Shuyang Chen, Ping Yu

October 9, 2019

Abstract

The current way of measuring body fat is complicated and inaccurate. Our group try to come up with a precise, robust and simple model with **two measurements** to calculate body fat.

1 Data Cleaning

In this section, we check some duplicate records, delete and impute some possible bad points. In the end, we delete 163rd, 182nd, 216th, 221st examples.

From the histogram of BODYFAT, the 216th example has a body fat of 45.1% and a density of 0.995, which is impossible for human. the 182nd example has a body fat of 0 and we get a negative body fat value using the body fat-density formula. Thus, these two point are deleted from the raw data set.

The histogram of HEIGHT shows that the 42rd example is only 29 inches, which is obviously a mistaken record. We impute his height using the BMI (adiposity) formula and the result is 69.4 inches.

The 39th example has the largest value in weight, abdomen, chest, hip, thigh, knee, biceps, wrist and adiposity. We believe this man is just too fat and it may contains some useful information. We keep this point.

We compute the ADIPOSITY using weight and height to see whether there are some wrong records in this variable. The 163rd and 221st ex-

ample has much larger difference between the records and computed values than other points. Thus, we believe the two points are mistaken records and delete them from raw data.

2 Model Selection

We compare the performance of two kinds of models: linear model and fraction model.

2.1 Linear model

We decide to use forward, backward selection with AIC and BIC as the criterion to choose the best linear model. The models given by AIC have more than eight variables and they are too complicated. Meanwhile, stepwise selection using BIC, no matter forward, backfard or in both direction give us the same model with four variables:

Bodyfat \sim Abdomen + Weight + Wrist + Forearm

Next, we look into the multicollinearity issues among the four variables by checking their VIF values. In this step, we divide the data set into training set (70%) and test set (30%). The random seed here is 123. Table 1 shows the result and it shows that from the full model, weight has the highest VIF value and it may have some multicollinearity problem.

In order to find two variables that has the best prediction power while suffer less from multicolliearity problem, we are going to drop the variables gradually to see whether the model performance changes a lot. Table 2 shows the result. **ABDOMEN** is an important variable

Var	Weight	Abdomen	Forearm	Wrist
VIF	7.70	5.51	1.70	2.12

Table 1: VIF of four predictors

Models	RMSE	R^2
Four variables	4.17	0.71
Abdomen + Forearm + Wrist	4.33	0.69
Forearm + Wrist	7.08	0.17
Abdomen + Wrist	4.33	0.69
Abdomen + Forearm	4.34	0.68
Weight + Abdomen	4.27	0.70

Table 2: Model performance

since by dropping it the R^2 drops to 0.17 and the RMSE increases. By making trade-off between RMSE and R^2 , we conclude that the model with **ABDOMEN** and **WRIST** is the best linear model.

$$\text{Bodyfat} = 0.682 \times \text{Abeomen} - 2.022 \times \text{Wrist} - 7.256$$

2.2 Fraction Model

Given Siri equation, we notice that there is a linear relationship between **BodyFat** and $1/\text{Density}$. By physical theorem, we know that **Density** can be calculated by **Volume** and **Weight**. Since we already possess accurate measure of **Weight**, the idea is to estimate volume using the remaining variables. Taking "Rule of Thumb" into account, we develop a new model named **Fraction Model**:

$$\text{Bodyfat} = \beta_1 \frac{X_1}{\text{Weight}} + \beta_2 \frac{1}{\text{Weight}} + \beta_3$$

We will use **10-fold Cross Validation** to find the best X_i among the remaining 13 variables. The result of Cross Validation is shown in Table 3, increasingly sorted by **RMSE**.

As shown in Table 3, **ABDOMEN** ranks top in the table no matter sorted by RMSE or R^2 . Therefore, we choose **ABDOMEN** as our

Variable	RMSE	R^2
ABDOMEN	4.06	0.72
ADIPOSIITY	5.10	0.55
CHEST	5.41	0.51
AGE	5.51	0.47
WRIST	5.73	0.43
HEIGHT	5.75	0.51
HIP	5.76	0.43
THIGH	5.84	0.40
NECK	5.87	0.41
ANKLE	5.88	0.41
KNEE	5.90	0.40
FOREARM	5.92	0.39
BICEPS	5.92	0.39

Table 3: Fraction Model Selection

optimal X_1 .

Then we investigate further into the fraction model associated **ABDOMEN**. The summary of this model indicates that **intercept** is **not** significant given its p value is 0.48, while the other 2 variables are both significant. Hence, we drop intercept, which produces a model of 0.96 R^2 :

$$\text{Bodyfat}_i = 153 \times \frac{\text{ABDOMEN}}{\text{Weight}} - \frac{1.06 \times 10^4}{\text{Weight}}$$

2.3 Model Comparison

We will compare the linear model with the fraction model to determine our final model. Table 4 shows the comparison of these two models. Given that fraction model's full data R^2 is noticeably larger than its 10-fold CV R^2 , we conclude that the Fraction model **overfits** on full data. In spite of that, we still choose it as our final model, because for both criteria, RMSE and R^2 , **Fraction** model behaves **better** than **Linear model** no matter judged by Cross Validation or full data. This suggests that on average Fraction model has better generalization ability than Linear model.

The above paragraph shows that Fraction model is better than Linear model. Before

	RMSE		R^2	
model	CV	Full Data	CV	Full data
Linear	4.20	4.21	0.70	0.68
Fraction	4.03	4.04	0.72	0.96

Table 4: Model Comparison

we jump to any conclusion, we should check whether Fraction model satisfies model assumptions of Linear Regression Model. Figure 1 shows that model satisfies Linearity, Homoscedasticity, Normality of error term in general. However, in the upper-left figure, we notice a "bowed" pattern of the red line, indicating that Fraction model makes systematic errors especially when it's predicting low Bodyfat.

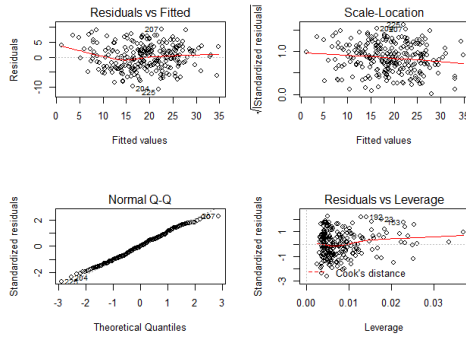


Figure 1: Fraction Model Assumptions Check

3 Shiny Application

Our body fat prediction shiny app address is https://clarefrost.shinyapps.io/shiny_628_group4/.

In this app, we set different unit for weight (kg & lbs) and abdomen (cm & inches). We also put the histogram of body fat in the app for users' reference.

4 Conclusion

In conclusion, we select **Fraction model as the final model.**

$$\text{Bodyfat}_i = 153 \times \frac{\text{ABDOMEN}}{\text{Weight}} - \frac{1.06 \times 10^4}{\text{Weight}}$$

The rule of thumb here is: 153 times the ratio between abdomen and weight, and then minus 1.06×10^4 times the inverse of weight.

One example can be: if a person has a weight of 184.5lbs and his abdomen circumference is 86.4inches. Then his body fat is $153 \times \frac{86.4}{184.5} - \frac{1.06 \times 10^4}{184.5} = 13.75\%$.

This model is good in the sense that:

- Simple model: Use only two accessible measurements, weight and abdomen.
- Precise model: Relatively low RMSE and high R^2 .
- Logical model: Utilize Siri equation and estimation of density to develop model, so that there is physical logics in the model.

However, this model also the following flaws:

- Overfits on full data.
- Systematically less precise predicting low and high Bodyfat.
- Restriction on the amount of variables leads to a crude estimation of volume.

5 Contribution

- Chao Chang: Linear model variable selection with BIC/AIC in both direction; Build the shiny App.
- Yezhou Li: Fraction model variable selection with cross validation; Report summary.
- Shuyang Chen: Linear model variable selection with backward/ forward BIC/AIC; Organize code, figure and data file.
- Ping Yu: Linear model variable selection with Mallor's C_p and R^2 ; Report summary; Presentation slides.