Body Fat Prediction with Multiple Regression

Zihang Wang, Yinqiu Xu, Sixu Li

October 19, 2020

1 Introduction

Body fat percentage (BFR) is believed to be one of the effective indicators of obesity [1]. However, accurately estimating BFP is quite expensive and time-consuming (dual energy X-ray absorptiometry [2]), which is difficult to apply in clinical practice or daily life measurements. So, the objective of this study is to develop a simple but effective model to predict BFP. We used a real data set contained 252 male with measurements of their BFP and various anthropometric measurements and we randomly split it into training and test sets. Our model was developed on the training set based on cross-validation validation and stepwise selection, and it predicts BFP on the test set with average error about (16.3%) of body fat, which is relatively accurate with limited data. In the subsequent sections, we will show how we obtain this model based on the data set we have.

2 Data Preprocessing

In this section, we will talk about how we began our data analysis by first preprocessing the raw data set. Basically, we preprocess our data in the following steps:

(i). **Drop irrelevant variables:** Note that variable IDNO is just the index of each data point and variable DENSITY is not allowed to use in the prediction, so, we just dropped these two variables.

(ii). **Data cleaning:** Firstly, we directly checked the data and we noticed that there are two people whose body fat are lower than 2, but normally, even the body fat of a male athlete should be greater than 2%. So, we removed these two abnormal points. Secondly, we drew some scatter plots between body fat and each anthropometric measurements, and we noticed that there are some outliers. Data point 42 only has height 29.5, which is obviously unrealistic, but other measurements of it seem reasonable, so we fixed it by recalculating the value of height based on the weight and adiposity of this person. Besides, there are also two other outliers, data points 216 and 39, which have 45.1 BFP and 48.9 BMI correspondingly. However, these two data are reasonable in the real life, so we currently kept them and we will discuss how to deal with them later. Thirdly, we know that $\text{BMI} = \frac{\text{Weight}_{\text{lb}}}{\text{Height}^2_{\text{inch}}} \times 703$[1], so, we used the weight and height in the original data to calculate BMI and compared it to the adiposity we have and we found out that, for date points 163 and 221, the absolute value of calculated BMI minus adiposity are greater than 1. But different from data point 42, in 163 and 221, we can't tell which variable (weight, height and adiposity) is problematic, thus, we decided to drop these two data points.

(iii). **Generate new variables:** Note that some of the measurements don't have obvious positive or negative relationship to the BFP, for example, people with large hip circumference are not necessary to be obesity. But some ratios of different circumferences are meaningful. For instance, people with small waist hip ratio are often in the good shape. Therefore, we consider to generate several ratios such as AHR (abdomen hip ratio), WBR(wrist biceps ratio) that could be potentially used as variables in the regression analysis.

(iv). **Split data:** We randomly split the original data set into training (80%) and test (20%) sets, and we would develop our model on the training data and then evaluated it performance on the test data.

[1]The formula comes from Wikipedia https://en.wikipedia.org/wiki/Body_mass_index.

3 Model Selection

Note that we have many variables that could be used in the regression, which means there are tons of models that we could choose, so in this section, we will talk about how to use cross-validation to select the best model. First, we proposed several candidate models based on the stepwise regression analysis or some basic knowledge about how people usually judge they are obesity or not. In particular, for stepwise selection, we first began with different combinations of variables (one of them is learned from paper [3]) and then using the BIC criterion to do the backward selections (this process is automatically done by R). And based on some previous researches and daily life experience, we proposed another candidate that only involves variables BMI and AHR in the regression model. After obtaining the candidate models, we did 10-fold cross-validation on the training data and selected out the best model based on average mean absolute percentage error (MAPE)[2] and model complexity. Specifically, we randomly split data into 10 equal sized subgroups. A single subgroup was retained as the validation set to calculate the MAPE and the remaining 9 subgroups were used to run the regressions. After repeating this cross-validation process for 10 times, we obtained an average MAPE for each candidate model and we chose the model with relatively small prediction error and less variables involved in regression.

4 Model Diagnosis

5 Conclusion

References

[1] Ho-Pham, L., Lai, T., Nguyen, M., and Nguyen, T. Relationship between body mass index and percent body fat in vietnamese: Implications for the diagnosis of obesity. *PloS one 10* (05 2015), e0127198.

[2] Lizak-Popiołek, D., Budzowski, A., Seń, M., and Czarny, W. Anthropometric measures of body composition used in obesity diagnosis – an overview. *Hygeia Public Health 51* (01 2016), 124–133.

[3] Merrill, Z., Chambers, A., and Cham, R. Development and validation of body fat prediction models in american adults. *Obesity Science & Practice 6*, 2 (2020), 189–195.

[2] MAPE $= \frac{1}{n}\sum_{t=1}^{n}|\frac{A_t-P_t}{A_t}|$, where A_t and P_t are the actual and predict value respectively.