Body Fat Prediction with Multiple Regression

Zihang Wang, Yinqiu Xu, Sixu Li

October 7, 2020

1 Introduction

Body fat percentage (BFR) is believed to be one of the effective indicators of obesity [2]. However, accurately estimating BFP is quite expensive and time-consuming (dual energy X-ray absorptiometry [3]), 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 splited it into training and test sets. Our model was developed on the training set based on crossvalidation 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. Basicly, we preposcess 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). **Detect and remove outliers:** After drawing some scatter plots between body fat and each anthropometric measurements, we noticed that there are two problematic data points. One of them has 0 BFP, which is obviously impossible, so we directly dropped it. And the other one has height 29.5, which is also unrealistic, but other measurements in this data point seem reasonable, so we recalculated the value of height based on the weight and BMI of this person. Besides these two points, there are also two other outliers, 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.

(iii). **Generate 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 develope our model on the training data and then evaluted it performance on the test data.

3 Model Selection

Note that we 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 invovle all the variables in the regression model and then using the AIC 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 with smallest average mean absolute percentage error (MAPE)[1]. Specifically, we randomly split data into 10 equal sized subgroups. A single subgroup is retained as the validation set to calculate the MAPE and the remaining 9 subgroups are used to run the regressions. After repeating this cross-validation process for 10 times, we will obtain an average MAPE for each candidate models and we choose the model with the smallest prediction error.

References

[1] Flegal, K., Kit, B., Orpana, H., and Graubard, B. Association of all-cause mortality with overweight and obesity using standard body mass index categoriesa systematic review and meta-analysisall-cause mortality using bmi categories. *JAMA : the journal of the American Medical Association 309* (01 2013), 71–82.

[2] 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.

[3] 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.

[4] Ortega, F., Lavie, C., and Blair, S. Obesity and cardiovascular disease. *Circulation Research 118* (05 2016), 1752–1770.

[1] 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.