Body Fat Prediction with Multiple Regression

Zihang Wang, Yinqiu Xu, Sixu Li

October 25, 2020

1 Introduction

Body fat percentage (BFP) is believed to be one of the effective indicators of obesity [1]. However, accurately estimating BFP is quite expensive and time-consuming. For example, using dual energy X-ray absorptiometry [2] 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 individuals with measurements of their BFP and various anthropometric measurements and selected our model based on stepwise selection and cross-validation. In the subsequent sections, we will concretely show how we obtain our final model based on the data set we have.

2 Data Preprocessing

In this section, we preprocessed our data in the following steps:

(i). **Data cleaning:** Firstly, we directly checked the data, and noticed that there are two people whose BFP are lower than 2%. However, males' BFP should be greater than 2%, even for the athletes. So, we removed these two abnormal individuals.
Secondly, we drew some scatter plots between body fat and each anthropometric measurements, and noticed that there are some outliers. Individual 42 only has height 29.5, which is obviously unrealistic, but other measurements of him seem reasonable. So, we fixed it by recalculating the value of height based on the weight and adiposity of this person. Besides, there is also another outlier, individual 39, who has 48.9 adiposity. However, this seems reasonable in the real life, so we chose to keep it and will discuss it in the Section 4.
Thirdly, we know that $\text{Adiposity}_{\text{BMI}} = \frac{\text{Weight}_{\text{lb}}}{\text{Height}^2_{\text{inch}}} \times 703$[1]. So, we used the weight and height to calculate adiposity and compared it to the provided value. And we found out that, they are not consistent for individuals 163 and 221. Because adiposity is obtained from weight and height, it is quite possible that the value of adiposity in individuals 163 and 221 was wrongly calculated. Therefore, we replaced these two original adiposity with the calculated ones.

(ii). **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 considered generating several ratios such as AHR (abdomen hip ratio), WBR(wrist biceps ratio) that could be potentially used as variables in the regression analysis.

After data preprocessing, there are 250 individuals with 23 variables in our data set and the average of body fat percentage is 19.08% with standard error 7.21.

3 Model Selection and Experiment Results

Note that we had 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 stepwise selection and cross-validation to select out the best model. First, we proposed several candidate models based on the stepwise regression analysis or some basic knowledge of BFP prediction. In particular, for stepwise selection, we began with different combinations of variables, which were chosen based on some background research, i.e., what variables people usually used in the previous literature about BFP prediction [1–3], and then used the BIC criterion to do the stepwise selections. Besides, from daily life experience, we proposed two other simple candidates. In particular, one only involves variables adiposity and AHR, and the other uses variables abdomen and weight in the regression model. After obtaining the candidate models, we did 10-fold cross-validation and selected the best model based on average Mean

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

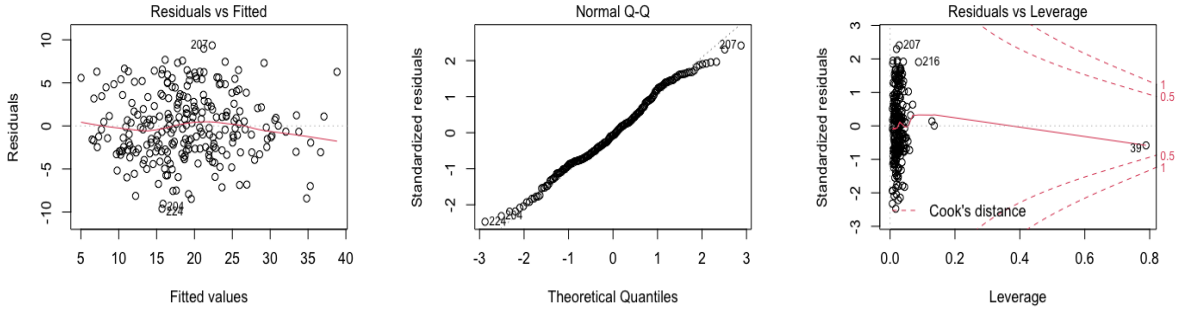
Absolute Percentage Error (MAPE)[2] and model complexity. Specifically, we randomly split data into 10 equal sized subgroups. One single subgroup was retained as the validation set to calculate the MAPE and the remaining 9 subgroups were used to run the regressions. After repeating this cross-validation process for 10 times, we obtained an average MAPE for each candidate model and chose the model with relatively small average prediction error and low model complexity.
After running the whole process above in R, we obtained our final model (the initial combinations of variables in this model mainly comes from [3]):

$$\text{BFP} = 0.077*\text{age} + 2.383*\text{adiposity} - 0.036*\text{adiposity}^2 - 0.474*\text{neck} + 0.631*\text{abdomen} - 1.985*\text{wrist} - 25.414 \quad (1)$$

Although this model is more complicated than those models only involved two variables, it significantly outperforms them on the prediction accuracy. So, we decided to sacrifice some model simplicities for better performances, and our final model's average MAPE in the 10-fold cross-validation is about 22%, which is relatively accurate with limited data. Besides, all the p-values of the parameters in our model are less than 0.015, which means that all the parameters are significant. And the adjusted R^2 is 0.7349, which implies that our model explains about 73.5% variation of BFP. Note that there are 7 coefficients in our model, for example, our estimated coefficient for variable age is 0.077 in the unit of year. This means that when the individual's age increases 1 year, the predicted BFP will increase about 0.077%. Additionally, our model could not only provide us with the exact prediction value, but also give us the prediction interval (PI) with given confidence level. For example, a 45 years old man with 25.5 adiposity, 38 cm neck circumference, 92.7 cm abdomen circumference and 18.2 cm wrist circumference is expected to have about 19.65% BFP based on the prediction of our model and the PI is between 19.12% and 20.18%.

4 Model Diagnostics

In this section, we will check the assumptions of our quadratic regression model, i.e., homoscedasticity and normality, and then deal with the influential points, by using the diagnostic plots. Firstly, we plotted residuals against fitted values to check constant variance.As displayed in Fig.1(a), the residuals distributed fairly random around zero. Therefore, the assumption on homoscedasticity is reasonable. Secondly, we used QQ-plot to check the normality of residuals. And as we can see in Fig.1(b), most of points lie on the diagonal, so the normality assumption holds. Finally, the standardized residuals against leverage was plotted to find high-leverage points. Notice in Fig.1(c), all of data points are within the dashed Cook's distance line, although 39 is close. So, we removed it in the data set and ran the cross-validation again on our model. And since the average MAPE had little change, we tended to not consider individual 39 as an influential point.



(a) Residuals plot (b) Q-Q plot (c) Leverage plot

Figure 1: Diagnostics plots

5 Conclusion

Because our model was selected out by cross-validation and used average MAPE as the criterion, it should be one of the most fitted regression model for the given data. Also, it would generalize well in the prediction and be robust to some extreme data. However, there are also some weaknesses of our model. Besides relatively high model complexity, the interpretability of this regression model might be problematic or in other words, the interpretation of the parameter in our model might contradict the common sense. For example, the parameter of variable neck is -0.474, which means that people with larger neck circumference would have lower BFP, which is obviously different from our daily experience. There might be two reasons for the strange behaviors of the parameters in our model. One is that our current data is too limited to find the best rule-of-thumb for BFP prediction and the other is that maybe our common sense on the BFP prediction is not fully correct. Therefore, in the future work, we could try to collect more data and fit our regression model again to figure out the true reason and improve the interpretability of our model.

[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.

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.

Contributions

- Zihang Wang wrote the introduction and data preprocessing parts in the summary and slides.
- Yinqiu Xu wrote model diagnostics parts in the summary and slides.
- Sixu Li wrote the model selection and conclusion parts in the summary and slides.