Summary

Lanxin Xiang, Tianyue Luo, Yuyang He
October 20,2021

Introduction: The percentage of body fat a significant indicator to evaluate health condition. However, accurate estimation is conducted by measuring one's density, which is inconvenient and costly. Based on the given dataset, we propose a body fat prediction model. In this report, we will break down the detail of modeling, including data cleaning, modeling criteria, model analysis and conclusion.

Data Cleaning: We **removed 8** data and **imputed 1**. First, by checking the extreme value of variables we found two unrealistic samples with respective 0% and 1.9% body fat and one with 45.10% body fat which effects the coefficient of prediction model. The height 29.5 inches is also suspicious. We calculated the height from BMI and weight and then impute the suspicious height. Second, we checked whether the redundancy data matched each other (ie. if height, weight and adiposity matched, and bodyfat and density matched) and removed those neither match nor can be imputed. Third, applying cook's distance, we removed sample 86 which was an outlier.

Model:

Final model and interpretation: Considering people may not know or do not have measuring instrument to measure the dimentions, we put forward one **"Daily"** model and one **"Advanced"** model. Each of the model consists of three subsections dividing by age.

Equation (1) refers to the "Daily" model. Equation (2) refers to the "Advanced"model.

$$\mathrm{BF} = \begin{cases} -0.14\mathrm{W} - 0.19\mathrm{H} + 0.92\mathrm{Abd} - 28.12 & 22 \leq age < 45 \\ -0.06\mathrm{W} - 0.14\mathrm{H} + 0.82\mathrm{Abd} - 36.21 & 45 \leq age < 60 \\ -0.02\mathrm{W} - 0.17\mathrm{H} + 0.56\mathrm{Abd} - 16.93 & age \geq 60 \end{cases} \tag{1}$$

$$\mathrm{BF} = \begin{cases} -0.39\mathrm{H} - 0.55\mathrm{N} + 0.74\mathrm{Abd} + 0.30\mathrm{Bic} - 2.20\mathrm{Wst} + 29.07 & 22 \leq age < 45 \\ 0.68\mathrm{Abd} + 0.29\mathrm{Thi} - 1.32\mathrm{Ank} + 0.59\mathrm{Fora} - 1.80\mathrm{Wst} - 15.27 & 45 \leq age < 60 \\ -0.18\mathrm{W} + 0.77\mathrm{N} + 0.67\mathrm{Abd} + 1.81\mathrm{Ank} + 0.75\mathrm{Bic} - 0.80\mathrm{Fora} - 82.07 & age \geq 60 \end{cases} \tag{2}$$

The reasons we chose the model are as follows. **First**, previous studies shows percentage body fat increased slightly between ages 20 to 39 years and 40 to 84 years[1] and the pattern of body fat distribution has been indicated to be related to age[2]. However, the distribution of our data's age variable has several peaks. To cope with the problem and utilize the features mentioned above, we divide the data to three subsets: young adult($22 \leq age < 45$), middle age($45 \leq age < 60$), and old(≥ 60). **Second**, some important dimensions are not commonly used in daily life. Considering the utility of our model, we pre-

sented two versions for different people to use. **Third**, for "Advanced" model (equation 2), we applied backward stepwise regression to select the variables with AIC.

Statistical Analysis:

Our model is based on linear regression. Adjusted R^2 and RMSE are adopted to verify the performance of modeling. Table 1 lists the two value of final model. From the large Adjusted R^2 value and small RMSE value, we can tell the model performs well.

We also tried two other models: 1. treat "age" as a factorial variable instead of the criteria of divide data; 2. two-hidden layer neural network model. However, the first one's RMSE is larger than our final model and the second one's Adjusted R^2 is similar, yet the model itself is more complicated and lack of practical explanation.

Overall	Advanced	/	3.49
Whole	Advanced	0.7262	3.80
Factor	Advanced	0.7254	3.81

Table 1: Adjusted R^2 and RMSE table

Model Diagnostics:

Normality. We applied QQ-plot and Shapiro-Wilk test to check the normality of our model. Both "Daily" and "Advanced" models' QQ-plot falls along the line, and both models passed the test with p-value larger than 0.05 within each subset.

Linearity and Homoscedasticity. There were no pattern of the points on Residuals vs. Fitted value plots. They distributed randomly around 0, which suggests the assumptions of linearity and homoscedasticity are not violated.

Figure 1 has two sub figures. They are the QQ-plot and Residuals vs. Fitted-value plots of our "Daily" model of young adult group. The other plots are similar.

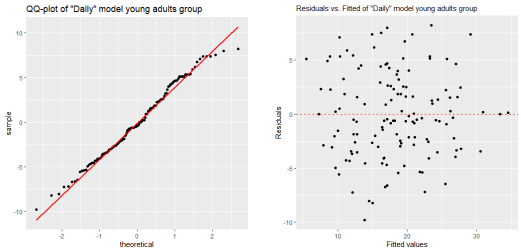

Figure 1: Main name

Model Strengths and Weaknesses:

- Strength: **Goodness of fitting**. By dividing the data to three parts, our model fits better. **Efficient**. We selected model with RMSE and Adjusted R^2, which guaranteed it not be overfitted, while stepwise regression method ensures the minimum AIC. **Useful**. Our model is easy to understand and present suitable solutions for different people.

- Weakness: The goodness of fit is constrained by linear model itself and thus, difficult to make more improvement. To make grouping have more practical significance, the sample size of each subset is uneven.

Conclusion: We cleaned the data set and fitted several candidate models. By comparing the RMSE, Adjusted R^2 and usefulness, we chose the three-subsectional model discussed above as our final model. Not only does it valid statistically, but also practically useful.

References

[1] Silver, A. J., Guillen, C. P., Kahl, M. J., & Morley, J. E. (1993). Effect of aging on body fat. Journal of the American Geriatrics Society, 41(3), 211–213.

[2] Hiroshi Shimokata, Jordan D. Tobin, Denis C. Muller, Dariush Elahi, Patricia J. Coon, Reubin Andres, Studies in the Distribution of Body Fat: I. Effects of Age, Sex, and Obesity, Journal of Gerontology, Volume 44, Issue 2, March 1989, Pages M66–M73.

Contributions:

- Lanxin Xiang: summary, factorial model, model analysis, residual plots.
- Tianyue Luo: slides, data cleaning, model selection and analysis, model plots.
- Yuyuan He: shiny app, two-hidden layer neural network model.