Stat.628 Report

Zhengyuan Wen, Jiaying Jia, Hengrui Qu

1 Introduction and Data Cleaning

1.1 Introduction

This project aims to come up with a simple, robust, accurate and precise "rule-of-thumb" method to estimate percentage of body fat using clinically available measurements. 252 observations and 17 variables are included in the dataset.

1.2 Data Cleaning

- Check by summary table of dataset

  Some extreme values were found in some variables after using the `summary()` function in `R`.

  As response variable, the value of BODYFAT ranges from 2 to 40. Data of Individual 172,216,182 were unreasonable. As for WEIGHT and HEIGHT, Individual 39 had extremely large values, while Individual 42 had extremely small values. The extreme values are not suitable for us to train the model, so we decided to delete these individuals entirely.

- Check by Siri Equation and BMI Equation

  We built a linear model between the bodyfat percentage calculated by Siri's equation and the bodyfat percentage in the data set. And we also built a linear model between BMI(calculated by data of Height and Weight) and the ADIPOSITY in the data set.

  According to the residual plot and the QQ plot, we could see that Individuals 33, 48, 76, 96 were inconsistent with Siri's results, Individual 163, 221 were inconsistent with those computed by the BMI equation. Because we are not sure which variable has input error in original dataset, we deleted all of these individuals.

- Finally, We removed Individuals 33,39,42,48,76,96,163,172,182,216,221,now there are 241 observations.

2 Modeling and Analysis

2.1 Candidate Models and Motivations

- **Simple Linear Regression with BIC Criteria**. The first method we tried was building a linear model and using two-way stepwise method with BIC criteria. The main idea is that if one new variable is to be added in the model but has high multilinearity with the existing variables, then it's useless to add this new variable since it would barely add any additional information.
- **Lasso**. Then we tried to select variables using lasso method. The main idea is to add a L_1 penalty for both fitting and coefficients based on OLS. The lasso estimates share features of both ridge and best subset selection regression and can set some of the coefficients to zero, or in other words, shrink the magnitude.
- **Generalized Additive Model (GAM)**. To further extract the information from the data, we tried to improve the linear model and added some nonlinear trend, which came out to be GAM. Since all variables are continuous, we choose Gaussian family as our assumption and link function as the identity. To select an appropriate subset of predictors, we firstly used Double Penalty Approach[1] to delete the variables which were shrunk to zero. Then we removed the insignificant predictors step by step, at last we obtained a model with all significant predictors. Here we didn't use some information criteria since AIC would tend to select a more complicated model and BIC is not suitable for the models which have "infinite" parameters[3].

2.2 Final Model Selection

We ended up choosing the best model by 10-fold Cross Validation (CV). We compared the three models above, used training data to train the model, and used test data to calculate RMSE. We ran CV with $n = 300$. That CV gave us following results.

Table 1: Average RMSE for Linear Model, Lasso and GAM in CV (Keep 4 decimal places)

	lm_test	lm_train	lasso_test	lasso_train	GAM_test	GAM_train
Average RMSE($n = 300$)	3.9607	3.8892	4.0381	3.8202	3.8663	3.6073

From the above criteria, we chose GAM as our final model since it has the minimal RMSE in test set[1]. Final model should look like (1). $_{(\cdot)}$ is estimated P-value for coefficients or smooth terms.

$$\begin{aligned}\widehat{BODYFAT}_i &= 26.035_{(0.000)} + 0.060_{(0.033)}AGE_i - 0.108_{(0.001)}WEIGHT_i + 0.332_{(0.055)}FOREARM_i \\ &\quad + f_1(ABDOMEN_i)_{(0.000)} + f_2(THIGH_i)_{(0.028)} + f_3(WRIST_i)_{(0.001)} \end{aligned} \tag{1}$$

All numbers keep 3 decimal places. Estimation method for smooth parameters is REML. Type of smoothing splines is Thin Plate Regression Splines. These two methods have been proven to be more stable in most of the cases[2]. We obtained our adjusted R^2 as 0.747.

2.3 Analysis

2.3.1 Theoretical Interpretation

From the above linear coefficients, as a man gets older by one year, he is expected to gain 0.06% in body fat. And other linear coefficients could be interpreted similarly. We noticed that the weight had a negative effect on body fat. This result did not accord with our intuition. But after thinking deeply, a man who has a greater weight means that he could have greater muscle mass and since fat is less dense than muscle. This result gave us an interesting insight of body fat. Next, when it comes to the non-parametric part, the model gave us following figures.

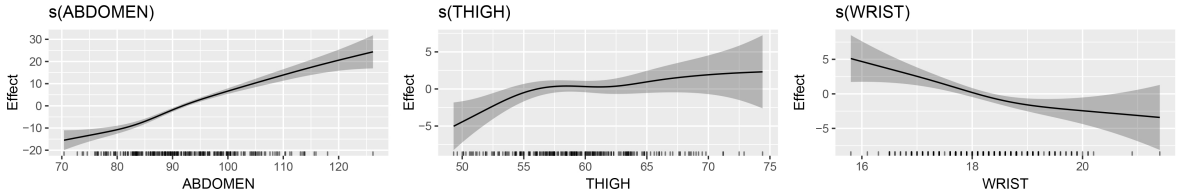

Figure 1: Estimation for Non-parametric part

We could take a rough look at the figures. For the Y axis in the figure, the abdomen measurement has the dominant effect on bodyfat. Also, holding other predictors unchanged, as abdomen measurement increases, bodyfat will increase faster and gradually has a stable increasing rate. The thigh diagram can be interpreted similarly. As the thigh measurement increases, the rate of increase in bodyfat declines to near zero, and then slightly increases. Also, the bodyfat will go down as the circumference of wrist goes up. This might seem strange, but in fact, if a man has a larger circumference of wrist, it turns out he should have a larger bone structure which results in a less body fat.

Also we can make a prediction. Given a man with 22 years old, weighing 154 pds weight, 85.2 cm abdomen, 59 cm thigh, 27.4 cm forearm and 17.1 cm wrist, the model gives a fitted value 15.74% bodyfat with a 95% confidence interval $[14.18, 17.31]$.

2.3.2 Model diagnosis

We checked the residuals plot and residuals are indeed randomly normal distributed. Also we checked the effect degree of freedom (EDF) that we chose in our final model, none of them hit the ceiling of degree of freedom constraint, so we say we have an effective estimation on smoothing splines. We also check our model concurvity (similar to multicolinearity), it turns out that our predictors have concurvity issue, which results in a underestimation of variance and tends to inflate the type 1 errors.

3 Strengths and Weakness

Strengths. Overall, the final model has a more precise prediction on test set. It is robust, a lot better than Lasso Method. It can also capture the non linear trend in variables.

Weakness. Predictors are correlated which might result in a misinterpret in the model. And heavy computations are needed to estimate the model. A non-parametric method works well within the range of predictors given, but it performs bad in generalization.

[1] We just compared the relative value, so it's okay for us to predetermine the model using the whole dataset.

A References

References

[1] Giampiero Marra and Simon N. Wood. Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, 55(7):2372–2387, 2011.

[2] Simon N. Wood. *Generalized additive models: An introduction with R. Boca Raton, FL: Chapman & Hall*. CRC, 2006.

[3] Yongli Zhang and Yuhong Yang. Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, 187(1):95–112, 2015.

B Contributions

Zhengyuan Wen built the GAM model and wrote R code for this model and 10-fold CV. He also wrote the Final Model Selection part, Analysis and Strengths and Weakness part of the summary. And he was in charge of Shiny code.

Hengrui Qu wrote the Introduction and Data Cleaning part of the summary and R code for data cleaning, and was responsible for the whole PowerPoint.

Jiaying Jia built the stepwise model and lasso model, and wrote R code for both models and R code for calculating RMSE. She also wrote Candidate Models and Motivations part of the summary and was responsible for editing and polishing the article.

We met 5 times and spent 15 hours in discussion. And we each spent around 10 hours after the discussion.