Stat 628-Module 2 Summary

Fengxia Dong, Haohao Su, Huitong Kou

Introduction

Accurate measurement of body fat can be inconvenient or costly and thus we hope to develop some easy methods of estimating body fat. In this project, based on the given data, our group came up with a simple, robust, and accurate "rule-of-thumb" to estimate percentage of body fat using clinically available measurements. In the following parts of this summary, we will discuss about our model in detail.

Background Info and Data Cleaning

The data set contains measurements, including age, weight, height, bmi, and various body circumference measurements, from 252 men who had their body fat percentage accurately measured via underwater weighing. In this project, BODYFAT (body fat percentage) should be our target variable, while all other variables except IDNO (ID number) and DENSITY (density determined from underwater weighing) make up the set of predictors.

To clean the data, we extracted and removed extreme outliers with the help of boxplots for each predictor by setting the outlier outer fence as the inter-quartile range multiplying by 3. Four extreme outliers are removed which are observations of 31, 39, 42, and 86.

Motivation for Model

Multicollinearity is detected by several measures including Farrar Chi-Square, Red Indicator, Sum of Lambda Inverse, and Theil's Method. Therefore, we choose from ridge regression, lasso regression, and elastic net regression, which can handle multicollinearity issues. Each of the three models is fitted with all variables standardized. By comparing mean squared errors (MSE), the elastic net regression is selected as our best model with the smallest MSE value. The elastic net regression has the benefits of both ridge and lasso regressions by including both L-1 regularization (which penalizes the sum of absolute values of the coefficients) and L-2 regularization (which penalizes sum of squared coefficients). It gives us a simple model with AGE, HEIGHT, ABDOMEN and WRIST as predictors.

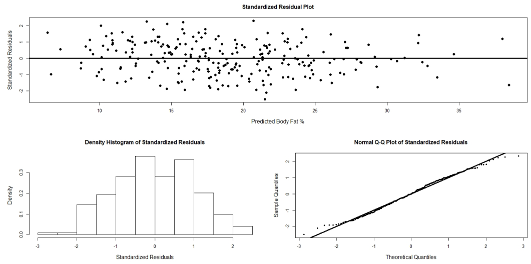
$$\begin{aligned}\textit{Body Fat Percentage} = &-10.093 + 0.019 Age - 0.256 Height \\ &+ 0.604 \textit{Abdomen 2 circumference} - 0.529 \textit{Wrist circumference}\end{aligned}$$

According to this equation, age, height, abdomen circumference and wrist circumference explain about 72.62% variation in body fat percentage (based on R^2) and its corresponding RMSE (10-fold-cross validation) is about 4.067.

Our estimated coefficients are -10.093, 0.019, -0.256, 0.604 and -0.529, which are separately in the units of intercept, age, height, abdomen and wrist circumference. This means that, controlling other variables, for every 1 inch increase in height, the model predicts that body fat % will decrease, on average, by 0.256; for every 1 cm increase in abdomen circumference, the model predicts that body fat % will increase, on average, by 0.604; for every 1 cm increase in wrist circumference, the model predicts that body fat % will decrease, on average, by 0.529. Also, every year, a man can gain, if he keeps his body shape invariant,on average 0.019 of body fat %.

Model Diagnostics

After getting the best model, we hope to diagnose some assumptions, including linearity, homoskedasticity and normality, with a residual plot, a residual density histogram and a QQ plot.



According to the figure above, the points look basically scattered around the X axis at random in the residual plot and there exist no obvious non-linear trends. So, linearity and homoskedasticity seem reasonable.

However, by the QQ plot of standardized residuals, the points do not hug the 45 degree line closely and the residual histogram shows that distribution of standardized residuals looks bimodal but not unimodal. Thus, it is possible that normality assumption is violated.

Rule of Thumb

Now, we can get a rule of thumb to estimate body fat.

”Multiply your age by 0.02 and minus a quarter of your height in inch. Multiply your abdomen circumference by 0.6 and minus a half of your wrist circumference. Sum these up and then minus 10.1.”

For example, a 5ft 8in tall graduate student (e.g. age 23), with 93 cm abdomen circumference and 17 cm wrist circumference, can be predicted to have 20.037% body fat.

Strengths and Weakness of Model

1. Linearity and homoskedasticity seem reasonable by the residual plot;

2. The distribution of residuals is likely to be bimodal and the normality assumption is possibly violated;

3. The predictors selected by elastic net regression explain about 72.62% of the variation in body fat %, which is not very high but enough for daily usage.

4. Most predictors and their corresponding coefficients seem reasonable, especially the ”Abdomen” variable. When other features are controlled, a man with larger abdomen circumference has higher possibility of obesity, which implies higher body fat percentage. However, the ”Wrist” variable with its negative coefficient may be not easy to interpret.

5. The elastic regression with 20-fold-cross validation and 20 repeats combines L_1 and L_2 penalties and offer an accurate and reasonable model with a simple form. But still, it is not an easy method to use, especially for a beginner to statistics or quantitative analysis.

6. While age and height is easy to attain, it is not very likely for a man to remember his own abdomen and wrist circumference.

Conclusion

Overall, our model provides a simple way of estimating the body fat percentage based on age, height, abdomen and wrist circumference. This model can explain around 73% variation in body fat percentage and some assumptions like linearity, and homoskedasticity seem plausible. However, the normality assumption is possibly violated and considering accessibility and modeling process, further simplicity can be expected.

Contribution

FD: created the main structure R code script and fixed some errors in summary file.

HS: added the plotting part and comments in R script, worked on the main part of summary file and PowerPoint slides.

HK: fixed some errors in the script, built the shiny app and managed the GitHub repo.