Stat 628 Module 2 Summary - Group 15

Wangwen Yi, Yudi Mu, Yuechuan Chen

Introduction

The body fat percentage (BFP) of a human or other living being is the total mass of fat divided by total body mass. Its true value is difficult to measure, but we can estimate it through some easy-to-measure indicators of our body. We made a body fat calculator which enables us to easily figure out the body fat percentage.

Data Cleaning

After checking the summary of the data set, there are some potential outliers (weight.png, height.png, bodyfat.png). We referred two formulas to do the data cleaning:

- Test the BMI (bmi.png), we noticed that the value of BMI can be calculated using this formula:
  $bmi = 730 * Weight/Height^2$. For each data, we compared this BMI to the ADIPOSITY and deleted the points which have big differences (163 and 221).
- Then we used the "Siri's equation" (siri.png): $cal.bodyfat = 495/DENSITY - 450$ to pre-calculate the body fat values and compare these values to the body fat in origin data set and deleted those which have big differences (point 48, 76, 96, 172, 182).

Choosing Model

The process of getting this model

- Used AIC and BIC criteria to do the parameter choice.
- Locked three linear models as well as the model used in US navy.
- Did the cross validation to the four models and found that their results are very close. So, we chose the model with the fewest parameters, which is ABDOMEN + WEIGHT + WRIST + FOREARM.
- Removed the influential point 39 after the diagnostic of the model (cook's distance.png).
- Found that the parameter FOREARM is not significant after the removal of point 39.
- Simplified the model by remove the insignificant parameter and got the final model.

Cross Validation

- we did a cross validation to help with the model selection.
- The scale of training set and test set is 9:1 and we made 100 random assignments
- We calculated the mean square error (MSE) as the basis for reference.

Statistical Analysis

- We use F-test to test the overall model to see whether body fat % is a linear function of ABDOMEN, WEIGHT and WRIST. Suppose the null hypothesis is that ABDOMEN, WEIGHT and WRIST are (linearly) unrelated to body fat %, and the alternative is that it isn't, i.e.

  $$H_0 : \beta_i = 0, i = 1, 2, 3$$

  $$H_1 : at\ least\ one\ \beta_i \neq 0, i = 1, 2, 3$$

  We can get the associated p-values$<2e-16$. We can declare that there is a linear relationship between ABDOMEN, WEIGHT, WRIST and body fat % with significant level 0.05.
- We calculate $R^2 = 0.734$, which implies ABDOMEN, WEIGHT and WRIST explains about 73.4% of the variation in body fat %.
- We conducted the following two-sided t test to see whether the predictors we have chosen are significant in predicting the outcome. Suppose the null hypothesis is that ABDOMEN/WEIGHT/WRIST is (linearly) unrelated to body fat %, and the alternative is that

it isn't, i.e.

$$H_0 : \beta_i = 0 <-> H_1 : \beta_i \neq 0,\ i = 1, 2, 3$$

We can get the associated p-values(see Table 1):

Table 1: p-value of the coefficients

Predictors	β_1	β_2	β_3
P-value	<2e-16	4.87e-05	0.002957

The coefficients are all significant with significant level 0.05. We can declare that there is a linear relationship between ABDOMEN/WEIGHT/WRIST and body fat%.

- We also get the 95% confidence intervals for the coefficients(see Table 2): For example, the 95% confidence interval for β_1 is (2.01, 2.52), which means that we are 95% confident that the interval (2.01, 2.52) contains the true coefficient value β_1.

Table 2: 95% confidence intervals of the coefficients

Percent	2.5%	97.5%
Intercept	-37.0241512	-12.33467477
β_1(ABDOMEN)	2.0114751	2.52267511
β_2(WEIGHT)	-0.1375737	-0.04881664
β_3(WRIST)	-5.1153256	-1.06265543

Model Diagnostic

- We checked the following assumptions for MLR. First, we check linearity and homo-skedasticity using Standardized Residual plot(residual.png). Linearity seems reasonable because there are no obvious non-linear trends in the residual plot; the points look randomly scattered around the X axis. Homoskedasticity is also plausible because the residuals appear to spread randomly along the X axis. Second, we check Normality of the error terms using QQ plot(QQ.png). Normality looks reasonable because the points in the QQ plot hug the 45 degree line very closely. But, there may be possibly skinny tail (see -3 to -2 and 2 to 3 region).

- We also looked at leverage points and influential points in regression models. We checked for leverage points using p_{ii} measures and checked for influential points using both the Cook's distance and the p_{ii} measures (influential.png). There does not seem to be any leverage points and influential points.

Model Strengths and Weakness

Strengths

- The model is simple and easy to interpret and calculate, and it fits all the assumptions of linear regression.

- Even comparing with the US Navy's model for body fat prediction, our model is still pretty good in sense of cross validation and (adjusted) R-square.

Weakness

- We did not include interaction, because we thought otherwise the model might be tedious.

- We excluded several potential "outliers", therefore, our model is not suitable to predict people whose measurements are disproportionate.

Rule of Thumb

Therefore, we use linear model to predict body fat percentage for men, and the predictors are abdomen(inch), weight(lbs), wrist(inch). The coefficients are rounded to two decimals, and the model is:

$$Bodyfat(\%) = -24.68 + 2.27 Abdomen(inch)$$

$$-0.09 Weight(lbs) - 3.09 Wrist(inch).$$

Contributions

Wenyi Wang: wrote code for model diagnostic, confidence intervals; wrote summary and PPT of parts: statistical analysis and morel diagnostic;

Yudi Mu: Shiny App; wrote code for data cleaning, modeling, validation; wrote summary and PPT of parts: strengths and weakness; Github.

Yuechuan Cheng: revised code for cross validation; wrote first four parts of summary and related parts in PPT; made vedio.