Module 2-Group 17

Introduction It's very challenging to measure the body fat accurately. Although we can get a very precise result, it requests lots of body component measurements of the person. In this module, we aim to predict the simple method to estimate the body fat using a few body component measurements which can be obtained easily.

Background Information/Data Cleaning First, notice that ID number has no effect on our prediction, and the density is an one to one function of body fat, so we remove those columns.
Then We used 1) Eyeballing summary of the data to identify any extreme points, errors in the data and missing values 2) By plotting histogram we checked if the potential outliers exists and removed those data points. For example,
1.a person has 0 body fat which is impossible.
2.a 44 years old man has 29.5-inch HEIGHT.
3.WEIGHT of a man is 363 pounds.
4. a man with 45.1 bodyfat

Motivation for Model/Choosing Model/Final Model/Rule of Thumb We want to build the most accurate model but meanwhile the variables used in the model should be easy to get and the number of variables cannot be very large. The simplest way is using all the variables and doing 'lm' to fit the model, which is used as the benchmark in our analysis. But this model will cause a multiple collinearity which may affect the prediction.
Basically, We consider 3 strategies to get over collinearity: 1.choose a model by AIC in stepwise algorithm;2.use significant variables based on ANOVA to do linear regression;3.Ridge regression. For the second strategy, noticing that there are 4 extremely significant variables(with p-value$< 2.2 \times e^{-16}$) , we not only consider a model that uses all significant variables, but also a model that uses only these four extremely important variables.
10-fold Cross-validation is used to test the performance of all possible models, and we will consider both number of variables used in the model(simplicity) and MSE(accuracy). The result is showed in **Table 1**.

Table 1: Results of cross-validation

	Number of variables	mean-MSE(based on CV)
Simple lm()	14	14.2
Stepwise model	7	15.6
ANOVA	8	14.1
ANOVA(4 extremely important variables)	4	15.3
Ridge regression	8	15.2
Ridge(4 extremely important variables)	4	16.0

As we can observe, the second strategy,use significant variables based on ANOVA, which is better than other methods. Because ANOVA has smaller MSE when has similar number of variables, and uses fewer number of variables when the MSE is similar. Comparing simple lm and ANOVA, we can see that use of all of the variables is unnecessary and we can omit few variables. Based on this analysis, we get two final models:
Model 1: $BodyFat = -27.8 - 0.092 \times Weight + 0.837 \times Abdomen - 0.001 \times Age - 0.201 \times Height$
Model 2:

$$BodyFat = -11.87 - 0.049 \times Weight + 0.734 \times Abdomen + 0.047 \times Age - 0.033 \times Height + 0.343 \times Adiposity - 0.261 \times Neck - 1.470 \times Wrist$$

The first model is based on 4 really important variables and the second is based on all 8 significant variables. The second one is more accurate but we also consider the situation when the user doesn't know some measurements like the thigh circumference. In this case, we will estimate the body fat based on model 1 and all of the variables in model 1 are very common such that most users know these measurements. For example, if one user gives us that $Weight = 154.25,\ Abdomen = 85.2,\ Age = 23,\ Height = 67.75$, then we will estimate the body fat based on model 2 and the estimation is15.3 with 95% confidence interval (7.5, 23.1).
The coefficients, take model 1 for example, are -0.095, 0.844, -0.004, -0.117. This means, for example, for every weight increases in 1 while keeping other variables the same, the model predicts that body fat percentage will decrease, on average, by 0.116. However, here we should point out that this doesn't mean people with higher weight tend to have lower body fat percentage since values of other variables will change when the weight changes.

Statistical Analysis In the final 2 models, we conducted F-test,the H_0 is the variable can be dropped from the model.From Table2, the p-values of these predictors are very small, all the p-values are smaller than 10^{-14} in model

1,this means that if the variable have no influence on our model,we only have less than 10^{-14} possibility that we get this coefficient of this variable,which shows strong evidence that these variables are significant for prediction,also, in model 2, all the p-values are smaller than 10^-3, so these models are good enough that we can't drop any variables in both models.

Table 2: F-test of model 1

	estimate	F-value	p-value
AGE	-9.792×10^{-4}	70.933	3.195×10^{-15}
WEIGHT	-9.248×10^{-2}	317.070	2.2×10^{-16}
HEIGHT	-2.008×10^{-1}	84.853	$< 2.2 \times 10^{-16}$
ABDOMEN	8.370×10^{-1}	121.452	$< 2.2 \times 10^{-16}$

Further more, we found the R^2 in our models are 0.7089 and 0.7267, which means more than 70% variation of the body fat are interpreted by our predictors.

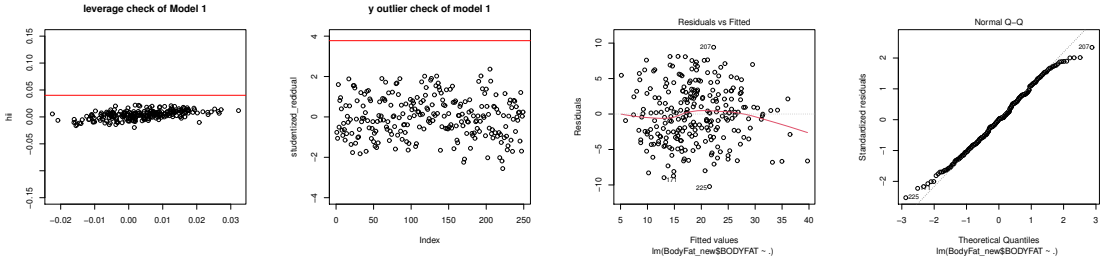

Figure 1: Diagnostic plots of model 1

Model Diagnostics We checked assumptions in the MLR.First we use residual plot to check the independence, equal variance and linearity, because the points are evenly scattered in the both sides of 0, so we can believe that the independence, equal variance hold, also, there is no obvious trend of the residual, so linearity should be reasonable.Then we using QQ-plot to check the normality assumption, the points fit the QQ-line well, so the normality is plausible. For outlier and influential, we use leverage h_{ii} to check x outliers and Bonferroni correction of t-test to check y outliers, we can see that there is no outliers in the data.And we use Cook's distance to check the influential, from the plot, we can see there is no influential.

Model Strengths/Weaknesses Our models have the following strengths: (1)Accurate and stable.Our model passed lots of test and cross validation, and was compared with lots software, its accuracy and stationary can be guaranteed. (2)Robust and flexible.For some situations, the person who wants to test his body fat may not be able to give all the variables in the data set, our model can cover this problem well, If one can give at least 4 pieces of information which can be measured easily, we can estimate the body fat percentage for him. If he can give more body information, we can give him a more precise prediction. (3)Simple but efficient. Our 2 models are both MLR model and the explanatory are carefully selected, which lead a simple calculation but a precise result.
But there does exist some weakness: To get over multiple collinearity issue, when we selected the explanatory for our models, we dropped some variables. These variables may contain some information, but these variables are not used in our model.

Conclusion In a conclusion, to get over multiple collinearity the model has to do some sacrifice that it dropped some variables, but our models still have good performance on various tests, cross validation and most important, prediction.It can not only use very few information to get a relatively accurate result, but also give a satisfactory prediction if information is enough.

Contribution

Ashvini Fulpagar:Introduction, data clean, conclusion part of summary and slides.
Hongyi Liu:Statistical analysis, model diagnostic part of summary and slides.
Luyang Fang:Model motivation and selection part of summary and slides