Group 5 Body Fat Data Project

February 4, 2019

The goal is to come up with a simple, precise and accurate way of determining body fat percentage of males based on readily available clinical measurements.

1 Step 1: Analyzing Raw Data

In this step, we check duplicate data, fix and amend badly-formatted, incorrect and amend incomplete data from original file. If the potential outliers only consist of less than 10%, we will remove them all.

- Relationship between *bodyfat* and *density*

No.48/76/96/182 disobey Siri's equation. If we want to do backward to impute the body fat percentage of No.182, which is 0 in original file, we will get a negative value. And it's not reasonable. Thus, we delete all four data points.

- Relationship among *adiposity*, *weight* and *height*

Adiposity is also known as BMI, which is defined as:

$$BMI = \frac{703 \times WEIGHT}{HEIGHT^2}$$

According to this formula, we remove No.42/163/221 because we have no idea which vari- able has input error in original file.

- Relationship among *abdomen* and other variables

abdomen has strong positive correlations with *hip*, *thigh*, *knee* and etc. We need be careful to deal with the multicollinearity problem.

2 Step 2: Model Selection

In this step, we split the data into two parts(training/test), training set includes 80% and testing set includes 20% respectively. We set the random seed as 1234 to split the data set.

- Forward stepwise selection For computational reasons, the best subset cannot be applied for any large n due to the 2n complexity. We want to select the single best model using Mallow's Cp, AIC, BIC, adjusted R2. For Mallow's Cp, it is equivalent to AIC in this case due to the Gaussian linear regression model. We ignore this part in order to show the more important result. The table below is the variable selection results based on three different criteria.

Criteria	Variables
lowest AIC	*Abdomen*, *Weight*, *Thigh*, *Forearm*, *Wrist*
lowest BIC	*Abdomen*, *Weight*
highest Adjusted R^2	*Abdomen*, *Weight*, *Thigh*, *Forearm*, *Wrist*, *Ankle*

- Lasso regression The LASSO is a regression method that involves penalizing the absolute size of the regression coefficients. We did a grid search on the best in order to get a better performance.

Best λ	Variables
1.21622	*Weight*, *Abdomen*, *Thigh*

- Model Comparison There are a number of criteria that one may use; typically we choose the smallest model that minimizes an estimate of the expected prediction error. The testing set Mean Square Error of four different criteria are shown as following:

Method	test MSE
AIC	12.23
BIC	13.00
Adjusted R^2	12.40
Lasso	13.07

We can notice that there is no obvious difference among four values. Based on the principle of parsimony, we choose the simplest model, which is generated by applying BIC criteria.

3 Step 3: Model Diagnostic

Remember that there is a strong positive correlation between abdomen and some other variables. Thus, it's important for us to check the multicollinearity in case violates the model assumption.

- Calculate the VIF

  A rule of thumb is that if $VIF > 10$, then there exists high multicollinearity.

VIF	Variable
162.64	Abdomen
162.64	Weight

We dropped *weight* due to the small effect size compared with *abdomen*.

4 Step 4: Model Rebuild

In this section, we will regress *bodyfat* on *abdomen* and do a series of model dignostic.

If you take a glance at the scatter plot (images/3-1.png) of the regression model, you will notice that there is a point far away from the red line, which may have a significant effect on the slope and intercept.

So we checked the leverage (~ 0.14, which is three times larger than the second largest one) and absolute standardized residual (~ 4). If we dig into the data point (No.39), the abdomen of No.39 is 148.1 (cm), which is much larger than the second maximum value. We can identify it as an outlier which distort the outcome and accuracy of the model.

We remove No.39 and rebuild the same regression model (images/5-1.png).

- Standarized Residuals vs Fitted Values (images/4-1.png)

The residuals fall within an area representing a horizonal band, which indicates the model assumptions of constant variance and zero error mean are correct.

- Test for error normality (images/4-2.png)

The graph confirms the Jarque-Bera test. The points are approximately aligned.

- Detection of outliers(images/4-3.png & images/4-4.png)

The data points with high leverage are people with extreme but reasonable abdomen values. So we won't delete these data points.

Based on the plots above, the OLS model assumptions are correct.

5 Step 5: Proposed Model

Our proposed model is to use abdomen to predict body fat percentage:

$$BodyFat(\%) = -38.15 + 0.62 \times ABDOMEN(cm)$$

Possible rule of thumb: "mutiply your ABDOMEN by 0.6, then minus 38"

- The coefficient of determination is equal to $R^2 = 0.67$.
- The regression is globally significant at the 0.05 level [F-statistic = 396.3, with Prob (F- statistic) = 1.16e-48]
- All the coefficients seem also siginificant at the 0.05 level.
- Every addtional centimeter in *abdomen* you can expect body fat percentage to increase by an average of 0.62%.

Stengths and Weakness of Model: I think the OLS model is a reasonable model between body fat percentage and abdomen, despite some caveats.

- Linearity and Additivity: The expected value of dependent variable is a straight-line function of each independent variable, holding the others fixed; The slope of that line does not depend on the values of the other variables; Our proposed model only have one independent variable.

- Statitical independence of the errors: The individuals collected in the data set are independent (no duplicate data points).
- Homoscedasticity of the errors: It's reasonable based on the diagnostic plots.
- Normality of the error distribution: Though there is a little tail in our QQ plot, the assumption is not violated.
- The precision of our body fat estimation: As shown below, our proposed model follows the spotlight in real life.
- Adjusted R^2 here is only about 0.67, which means this model may not have a satisfactory accuracy.

However, body fat percentage is a measurement of obesity. In practice, we notice that people only care about the description of body fat percentage rather than the exact value of body fat percentage:

Description	Men
Essential fit	3-5%
Athletes	6-13%
Fitness	14-17%
Average	18-24%
Obese	25%+

So we decided to classfify men into three levels: fitness people $(bodyfat17)$, average people $(17 < bodyfat < 25)$ and obese people $(bodyfat25)$. And the classfication results on test data are wonderful. None of these people is wrongly classified. Furthermore, the conclusion confirms our proposed model is right.

/	Predict fitness	Predict average	Predict obese
Actual fitness	25	0	0
Actual average	0	18	0
Actual obese	0	0	6

6 Contribution

Y.D. implemented principle component regression which is not included in the report. Y.D. planned and made former slides.

Q.H. implemented the forward variable selection, did part of model diagnostic and also completed the model interpretation.

Z.Z. implemented the LASSO regression and checked the incorrect values in the original file and also did part of model diagnostic.