Summery

October 7, 2019

Part I Background and Motivation

The body fat percentage (BFP) of a human is defined as the total mass of fat divided by total body mass, multiplied by 100. BFP is of great importance in measuring the well-being of individuals and predicting the risk of illness. However, the measurement of BFP has faced a trade-off problem between measurement accuracy and approach flexibility for a long time. Therefore finding a method that achieves both simplicity and exactitude in calculating BFP is our ultimate goal in conducting this project.

The dataset obtained has a total of 252 observations with 17 features including their percentage of body fat and various body circumference measurements.

Part II Data Preprocessing

In the data preprocessing part, we first examine whether the data has abnormal observations from the following 4 aspects:

- According to the histogram plot of `BODYFAT`, we find that observation `IDNO=182` has feature `BODYFAT=0`, which is apparently can not happen in the real world, therefore we remove this observation.
- As Siri's equation shown, `BODYFAT` goes linear with `1/DENSITY`. After fitting the linear regression model between `BODYFAT` and `1/DENSITY` we find observations with `IDNO=48,76,96` are outliers according to remedies plots. Since we can not decide neither `BODYFAT` nor `1/DENSITY` is trustworthy, we choose to delete these three points.
- According to the calculation formula of BMI (BMI=703Weight/Height^2), we find observation `IDNO=42` has the wrong height, we remove this observation and re-fit the linear model between `ADIPOSITY` and `WEIGHT/HEIGHT^2`. According to the remedies plots of the re-fitted model, observations with `IDNO=163,221` are outliers. After looking through the whole data, we cannot determine which feature's value is convincing and finally remove these two outliers. As for observation `IDNO=42`, we use the re-fitted model to correct its height.
- Except for the above approaches, we also take a look at other features. The correlation plot shows that there is strong multicollinearity between certain variables. The observation with `IDNO=39` has extremely large values in certain features. While we think this kind of person may truly exist, we choose to keep this observation in our dataset.

After removing several suspectable observations, we fit the full model with remain data and do further detection for the linear regression model. It can be shown from the remedies plots of the full model that there are no obvious outliers anymore. Except for outliers, we make a normal Q-Q plot for residuals and also perform Box-Cox transformation of the full model. Both of these two methods show that remained data are good enough so we do not need to do the transformation.

Part III Variable Selection

3.1 Mallow's Cp

Below are the results, model that chooses different variables will get different Adjusted R-squared values:

model	Adjusted R^2
BODYFAT ~ ABDOMEN	0.659
BODYFAT ~ ABDOMEN+WEIGHT	0.718
BODYFAT ~ ABDOMEN+WEIGHT+WRIST	0.724
BODYFAT ~ ABDOMEN+WEIGHT+WRIST+FOREARM	0.731

3.2 AIC & BIC

Criteria	Model	Adjusted R^2
	Both direction	
AIC	BODYFAT ~ AGE + WEIGHT + HEIGHT + ADIPOSITY + NECK +ABDOMEN + HIP + THIGH +FOREARM + WRIST	0.7413
BIC	BODYFAT ~ WEIGHT + ABDOMEN + FOREARM + WRIST	0.7318
	Backward direction	
AIC	BODYFAT ~ AGE + WEIGHT + HEIGHT + ADIPOSITY + NECK +ABDOMEN + HIP + THIGH + FOREARM + WRIST	0.7413
BIC	BODYFAT ~ WEIGHT + ABDOMEN + FOREARM + WRIST	0.7318
	Forward direction	
AIC	BODYFAT ~ ABDOMEN + WEIGHT + WRIST + FOREARM + NECK + BICEPS	0.7349
BIC	BODYFAT ~ WEIGHT + ABDOMEN + FOREARM + WRIST	0.7318

3.3 F-test

Factors in final model using this method:
BODYFAT ~ WEIGHT+HEIGHT+ADIPOSITY+ABDOMEN+THIGH+FOREARM+WRIST

3.4 LASSO

In order to deal with the problem of multicollinearity and do variable feature selection to choose a simpler model, we used lasso with lambda computed according to 1se rule. Then, we get 6 variables for the model: BODYFAT ~ AGE+HEIGHT+NECK+ABDOMEN+FOREARM+WRIST Even though the multicollinearity has been reduced, the model is not simple enough. Therefore, we do not choose LASSO as our final model.

3.5 Cross-Validation Based on Selected Variables

According to the result based on above citeria, we find that WEIGHT,ABDOMEN,WRIST,FOREARM are most selected. In order to further detect necessary variables in our final model, we use both R^2 and Cross-Validation to evaluate different model constructed by these variables.

Following are the results:

No.	Variable Numbers	Model	10-fold-CV	R^2
1	2	BODYFAT ~ ABDOMEN+WEIGHT	17.8358	0.7203
2	2	BODYFAT ~ ABDOMEN+WRIST	17.8831	0.6990
3	2	BODYFAT ~ ABDOMEN+FOREARM	20.7023	0.6631
4	2	BODYFAT ~ WEIGHT+WRIST	42.3430	0.3894
5	2	BODYFAT ~ WEIGHT+FOREARM	45.9751	0.3713
6	2	BODYFAT ~ WRIST+FOREARM	64.6050	0.1588
7	3	BODYFAT ~ ABDOMEN+WEIGHT+WRIST	17.5758	0.7276
8	3	BODYFAT ~ ABDOMEN+WEIGHT+FOREARM	17.2560	0.7255
9	3	BODYFAT ~ WEIGHT+WRIST+FOREARM	42.3504	0.3894
10	4	BODYFAT ~ ABDOMEN+WEIGHT+WRIST+FOREARM	16.8968	0.7362

Part IV Final Model & Outliers

4.1 Final model description We choose weight and abdomen as predictiors for body fat as our final model. Our final model is

$$\text{BodyFat} = -42.95790 - 0.11994 \cdot \text{WEIGHT} + 0.90152 \cdot \text{ABDOMEN}$$

The p-values of the hypothesis test on coefficients are all much smaller than 0.01. Therefore, coefficients are significant.The adjusted R-squared is 0.7148. The residual standard error is 4.083.
The laymen's interpretation is that when your weight increases one additional unit(lbs), the bodyfat will decrease 0.12%. When abdomen increases one additional unit(cm), the bodyfat will increase 0.90%.

4.2 Advantages and disadvantages

4.2.1 Advantages: The model is very simple since there are only two variables and it is convenient for users to gain their weight and abdomen data. The multicollinearity has been reduced and the model is robust.

4.2.2 Disadvantages: Since there are only two variables, we sacrificed some accuracy of this model. The model is based on the male data whose age is between 22-81, which means it may not be accurate to predict neither female nor male with age out of the range. Besides, there are still some outliers. Meanwhile, the data of users' abdomen may not be as convenient as variables like height and weight to obtain.

4.3 Model Diagnosis & Outliers

4.3.1 Assumption diagnosis and outlier detection To further estimate our model, we draw plots to check the assumptions and outliers for this model. As we can see from the residuals vs fitted values plot, the assumption of equal variance and linearity have been satisfied. Besides, from the qqplot, we can see that the normality assumption are satisfied.We also draw standard/studentized residuals vs fitted values plots and find that there is no significant outlier.

4.3.2 Influential points detection To determine some influential points, we draw Leverage plot, Diffits plot, Cook's distance plot, and Debates plot. Aside from Leverage plot, the other three plots do not find influential points. From the Leverage plot, we find 8 influential points(9,12,35,36,152,216,242). However, there is no enough evidence showing that they are outliers even though they are influential.

Part V Conclusion

According to above analysis, we obatin the bodyfat calculation in following way:

$$\text{BodyFat} = -42.95790 - 0.11994 \cdot \text{WEIGHT} + 0.90152 \cdot \text{ABDOMEN}$$

The heavier one is, the smaller his/her body fat percentage is. The larger his/her abdomen value is, the larger his/her body fat percentage is.

Part VI Contribution

Haifeng Liu writes the variable selection.ipynb in code folder and corresponding parts in summary.ipynb and presentation notes, and the README.md.
Yue Wu writes dataprocessing.ipynb, the cross-validation, cp criterion, and the corresponding summary and presentation notes. She also writes the code for all ggplot visualizations.
Hongwei Pan writes final model and outliers.ipynb, LASSO model, and the related parts in summary and presentation notes. Besides, she writes the code for Shiny app.