Percentage of Body Fat and Body Measurements

Abstract—In this paper, we evaluate the behavior and performance of several linear regression algorithms for predicting the body fat based on other body measurements. We implement the codes in R, and we use cross validation as evaluation method, and the mean squared error as evaluation metric.

1 INTRODUCTION

The idea is to implement a regression learner to predict the body fat based on other more accessible body measurements. This learner can be part of a mobile application on personal device or a health monitoring application on a hospital server. The system accepts a patient's profile. Then, feature extraction is performed to identify the key information and generate the vector representation of patient's body measurements. These vectors are fed to a regression learner to estimate the percentage of the body fat. The variety comes from different ethnicity, overall size of the body, age and gender characteristics and error while recording the measurements.

In this article, we focus only on linear regression but investigate the impact of feature selection, LASSO, RIDGE, stepwise, Principal component, and Partial least squares regression.

Problem Statement: The objective of our analysis is to estimate the percentage of the body fat based on age, weight, height, and 10 body circumference measurements.

Data Set: The dataset provided includes 252 observations and 17 variables. For each observation the percentage of body fat is calculated using Brozek's equation and used as response value. The dataset provided includes age, weight, height, percent body fat using Siri's equation, Density, Adiposity index, Fat Free Weight and circumference of chest, neck, abdomen, hip, thigh, knee, ankle, biceps, forearm, and wrist(Johnson R. 1996). We split the data to 9:1 for training and testing.

2 EXPLORATORY DATA ANALYSIS

We start our analysis by perfuming initial investigations on data to understand the features better and to discover any patterns or anomalies in the data set. We use summary statistics and visual representations to investigate the variation of features.

Size of dataset: Training data includes 227 observations and test data includes 25 observations.

Range: The features are in different range. It is necessary to standardize the variables before using Lasso and Ridge Regression.

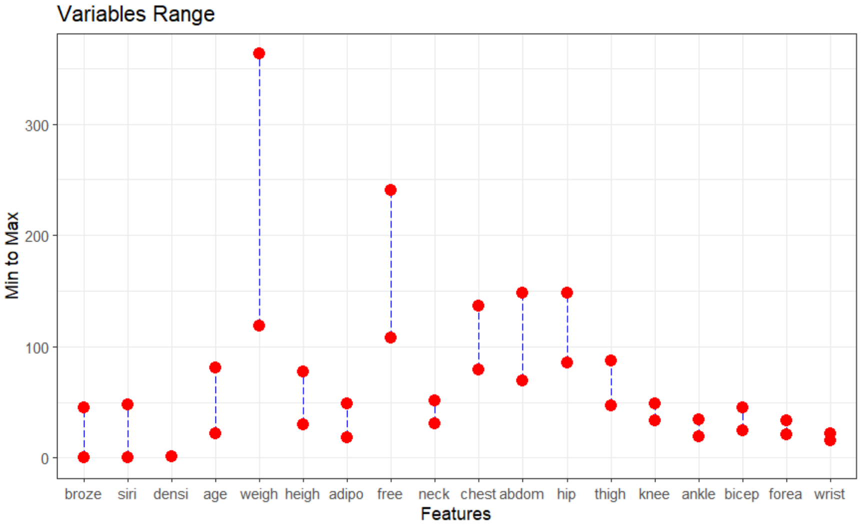

Figure 1: Range of Variables; We could variables are in different ranges.

Distribution of Response: The histogram is skewed to the right, and even removing one or two observations, doesn't cause the distribution to be normal.

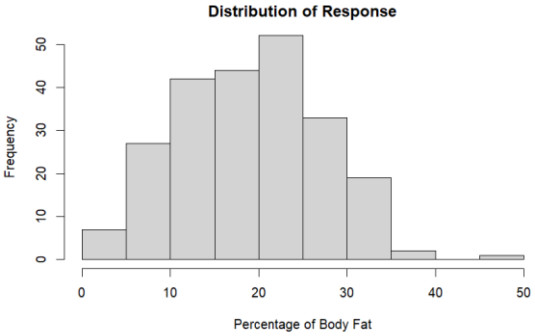

Figure 2: Distribution of response is not normal.

Correlation: Next, we investigate the correlation between each pair of features. Though, we do not remove the highly correlated features, because we want to feed the full features to our regression models and study the behaviors. It is not surprising to notice that Brozek is highly correlated to Siri because these are just two different methods to calculate the body fat. Moreover, We can see that weight and circumference of abdomen, hip and tight are highly correlated.

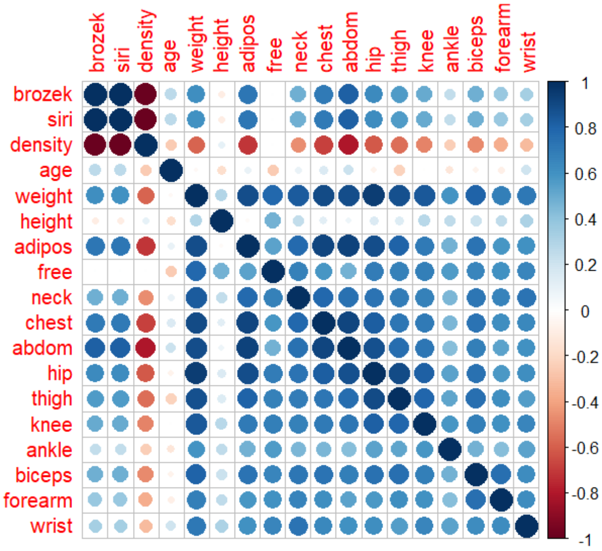

Figure 3: Correlation Heat map

Statistical Summary and Outliers: Next, we investigate the outliers using boxplot, but decided not to remove the outliers and just work with the data set as it is. Because of the variety in health conditions and, we may have rare values-but it is still expected and indicates of a specific health situation.

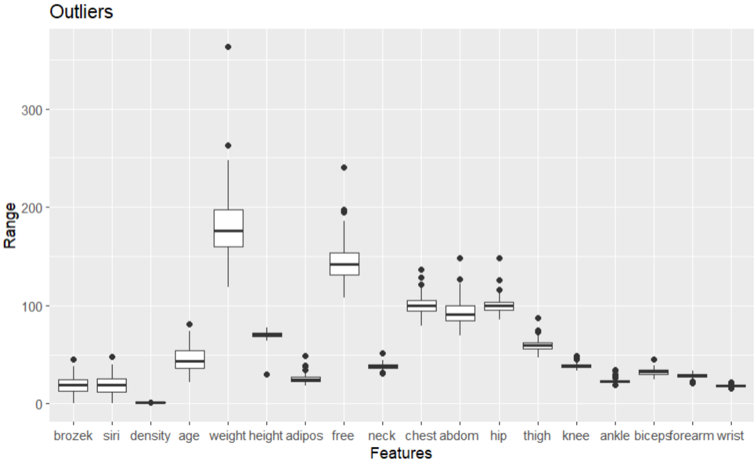

Figure 4: Boxplot; visualization of outliers, median and quartiles

Variation: Next, we explore the variation of each predictor for different response. The goal is to develop an **intuitive** understanding about the data. From figure 4, we can **anticipate** that Siri, density, abdomen, chest and weight would be more useful than height and ankle for estimating the Brozek body fat.

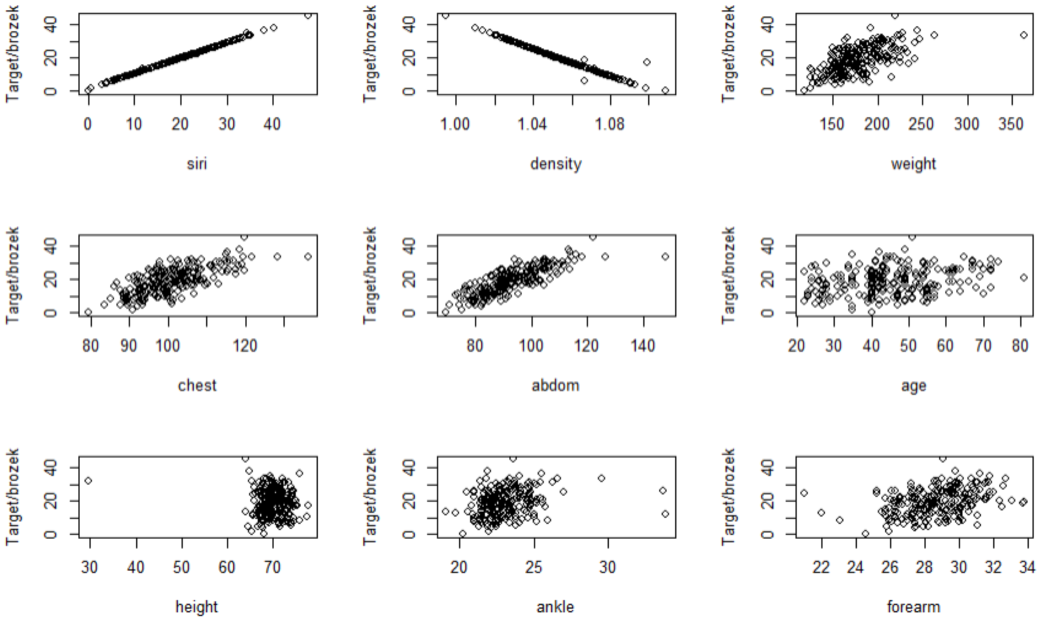

Figure 5: Sample plot of variation of response vs. each feature.

3 METHODOLOGY

In this section, we discuss regression models based on different subset of features. We utilize stepwise, ridge, lasso, principal component analysis, and partial least squares to build different models, and then compare their mean squared error (MSE) with simple linear regression using all features.

Standardization: Before training the models, first we standardize our training set, but not the target value. Then, we perform standardization on test set using the mean and variance from training data set. Note that we did not perform standardization on whole data set, because we wanted to make sure no information is leaked from testing data set to training data set (impact on mean and sigma).

Cross-Validation: To properly compare these methods, we utilize Monte Carlo Cross-Validation algorithm with B=100. In each step, we randomly split the dataset into 9:1 training: test set and repeat the above computations. Finally, we compute and compare the "average" performances of each model. Unlike last section, here we identify the best number of PCAs to be used for regression

4 RESULTS AND FINDINGS

Frist, we present the results from models without cross validations. Based on Table 2 and without performing any cross-validation, we can say that regression based on 5 variables of [Siri, density, thigh, knee, wrist] performs better on test set. It is interesting that for almost all models Testing MSE is better than Training MSE. This is a probably because test set is not chosen randomly.

Method	**Training MSE**	**Testing MSE**
Full model	0.0293	0.0087
The best subset model (siri+density+thigh+knee+wrist)	0.0315	**0.0028**
Stepwise variable selection with AIC	0.0294	0.0089
Ridge model	0.0296	0.0083
LASSO model	0.0308	**0.0031**
Principal Component model (nPC=10)	0.5389	0.3687
Principal Component model (nPC=17)	0.0293	0.0087
Partial Least Squares (PLS) model	0.0295	0.0103

Table 1: mean squared error for each method without using cross-validation.

Cross-Validation Results are listed in table 2. We can see that the LASSO model, and the best subset model with k=5 led to smallest average MSE over cross validation iterations. It is interesting to notice that standard variation for all models is almost same. You can see that the average MSE for PCA model in Table 2 is much smaller than the MSE reported in Table 1. It is because for the case without cross validation we did not list the MSE for optimum number of PCs. It is very important to notice PCA doesn't necessary lead the better model.

Method	**Average Testing MSE**	**SD Testing MSE**
Full model	0.0550	0.0860
The best subset model	**0.0462**	0.0896
Stepwise variable selection with AIC	0.0521	0.0892
Ridge model	0.0559	0.0866
LASSO model	**0.0443**	0.0823
Principal Component model (nPC=opt)	0.0550	0.0860
Partial Least Squares (PLS) model	0.0550	0.0860

Table 2: mean squared error for each method with using cross-validation.

5 SUMMARY AND CONCLUSION

We analyzed the behavior and performance of 7 linear regression model. We learned that PCA would not be always useful in dimension reduction. We also learned that comparison of different algorithm with or without cross-validation may lead to different decision. It is always better to utilize cross validation when comparing different algorithm.

Even though, we did not manually remove the highly correlated features, we can see that the features selection techniques remove these features- though it may not be the best choice, and that is why LASSO performed better than Ridge. (Melkumova et al .2017)

And finally, we find out that LASSO is the best model to estimate the percentage of body fat . We report the estimated MSE to be 0.0443.

6 REFERENCES

1. Johnson R. Journal of Statistics Education v.4, n.1 (1996).
2. Melkumova, L.E. & Shatskikh, S.Ya. (2017). Comparing Ridge and LASSO estimators for data analysis. Procedia Engineering. 201. 746-755. 10.1016/j.proeng.2017.09.615