ISYE 7406

Data Mining & Statistical Learning

Hassan Moughanieh

HW2

1. Introduction

1.1 Problem Description

The Brozek body fat formula is a method for estimating an individual's body fat percentage based on their body density. Developed by Brozek, this formula utilizes the principle that body fat is less dense than lean body mass. It typically requires precise measurements of body density, often obtained through underwater weighing or similar techniques. The formula itself involves specific constants to convert body density into an estimated body fat percentage.

In this project, we are looking to find the best statistical model that could predict the Brozek values from 17 potential parameters.

1.2 Dataset

The dataset comprises of 252 data points with 17 predictors. “Brozek” is the response we are after in this study. The following is a description of all the variables in the dataset (source: Johnson R. Journal of Statistics Education v.4, n.1 (1996)):

- **brozek** Percent body fat using Brozek’s equation, 457/Density - 414.2
- **siri** Percent body fat using Siri’s equation, 495/Density - 450
- **density** Density (gm/cm^3)
- **age** Age (yrs)
- **weight** Weight (lbs)
- **height** Height (inches)
- **adipos** Adiposity index = Weight/Height^2 (kg/m^2)
- **free** Fat Free Weight = (1 - fraction of body fat) * Weight, using Brozek’s formula (lbs)
- **neck** Neck circumference (cm)
- **chest** Chest circumference (cm)
- **abdom** Abdomen circumference (cm) at the umbilicus and level with the iliac crest
- **hip** Hip circumference (cm)
- **thigh** Thigh circumference (cm)
- **knee** Knee circumference (cm)
- **ankle** Ankle circumference (cm)
- **biceps** Extended biceps circumference (cm)
- **forearm** Forearm circumference (cm)
- **wrist** Wrist circumference (cm) distal to the styloid processes

2. Exploratory Data Analysis

The dataset was split into 90/10 training and testing dataset, respectively, resulting in 227/25 split. Looking at the summary statistics, we find that data varies widely across the board. For instance, Brozek ranges from 0% to 45.1%. The age parameter as well, covers ages from 22 to 81. Moreover, height varies from 29.5 to 77.75 inches reflecting a range.

Moreover, the correlation matrix shows a perfect or near perfect correlation between “brozek” and “siri” and “density” with values of 1 & -0.99, respectively. Other parameters like “height” and “free” had less significant values at -0.10 for ”height” & 0.01 for “Free”. The rest of the parameters ranged between 0.25 & 0.73.

By plotting the relationship between “brozek” & the other parameters, we notice a few outliers that could potentially affect the results of our models. (fig 2.1 & 2.2)

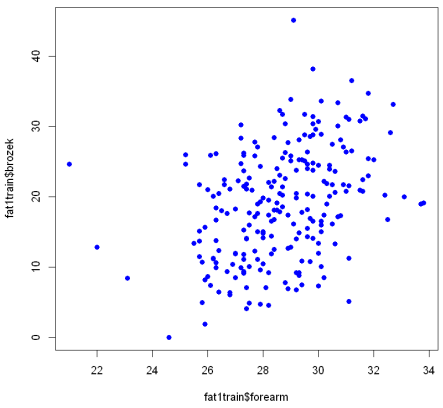

Fig 2.1

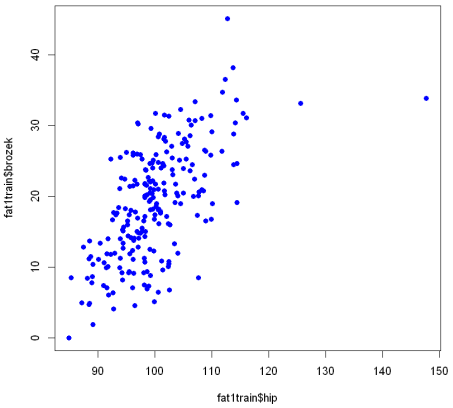

Fig 2.2

3. Methods/Methodology

The following 7 models were tested to find the best performing model with the least Mean Squared Error (MSE):

1. Linear regression with all predictors.
2. Linear regression with the best subset of k = 5 predictors variables.
3. Linear regression with variables (stepwise) selected using AIC.
4. Ridge regression.
5. LASSO.
6. Principal component regression.
7. Partial least squares.

We first built the full linear regression model with 17 parameters and “brozek” as a response. As mentoioned before, we could notice the existence of several outliers. However, we are more concerned with them being influential points. For that, we performed Cooks distance test for and looked for values larger than 1. As evident from the chart below, points 42 & 96 are influential points and will impact the performance of the models (fig 2.3).

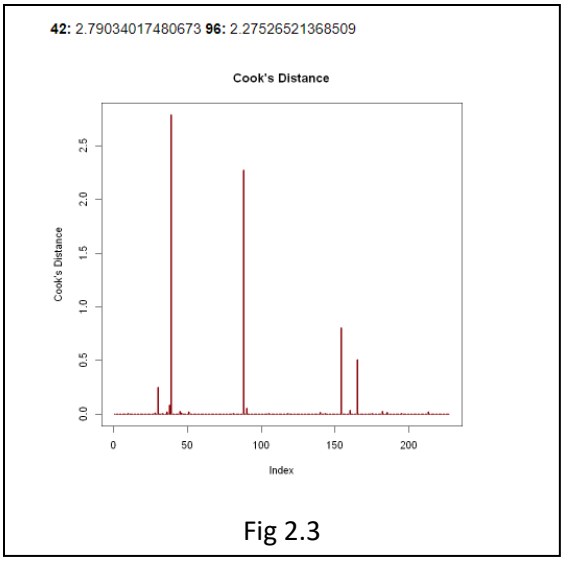

Fig 2.3

Next, we performed several tests to make sure the assumptions of linear regression hold. We started with linearity assumption. Most parameters show a linear relationship with “brozek” (fig. 2.1 & 2.2). We also checked for multicollinearity using VIF. All the values were below the threshold indicating lack of multicollinearity. Next, we checked for constant variance and independence. Both assumptions hold as evident by (fig 2.4).

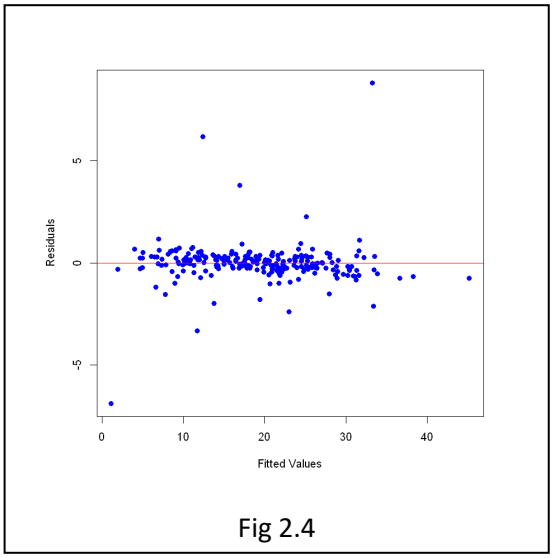

Fig 2.4

Lastly, we checked for normality using QQ-plot and histogram. Both figures (fig 2.5 & 2.6) show heavy tails on both sides indicating that the model probably needs transformation.

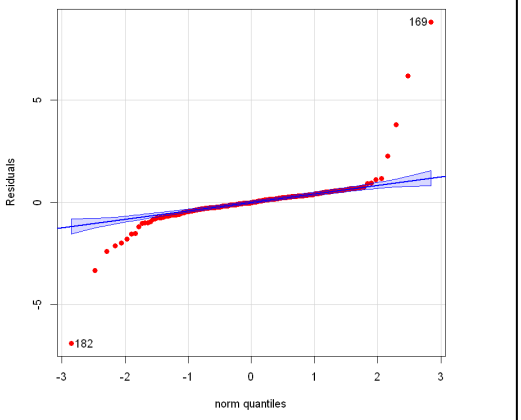

Fig 2.5

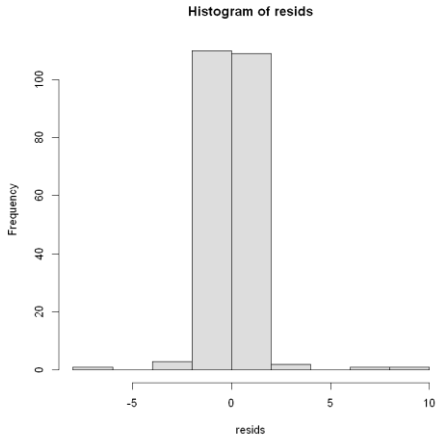

Fig 2.6

For the purposes of this project, we did not perform any transformation or remove the influential points to keep the results constant with the other students.

After that, we looked for the best performing subset model with 5 parameters. (fig 2.7) Shows as significant drop in RSS at k= 1 from 14,193.24 to 7.65. However, that improvement almost flattens out after that with RSS for k=8 at 6.83. The 5 parameters of the best performing models are “siri”, “density”, “thigh”, “knee”, and “wrist”.

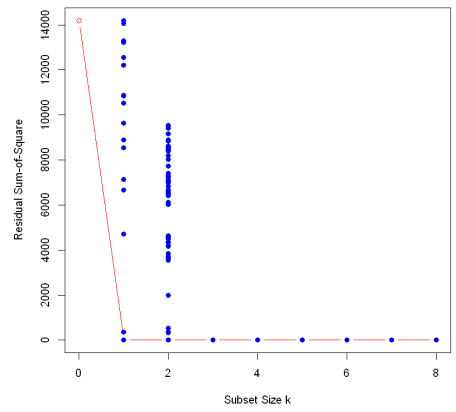

Fig 2.7

Next, we perform stepwise regression on the full model using AID as a selection criteria. The resultant model had 10 parameters of the 17 total with and Adjusted R-squared value of 0.9995.

The next model was ridge. We built the model with all possible lambda values from 0 to 100 with increments of 0.001. (Fig 2.8) shows how the beta coefficients change with lambda values. We then computed the optimal lambda and built the corresponding model. Note that ridge is not considered a feature selection method.

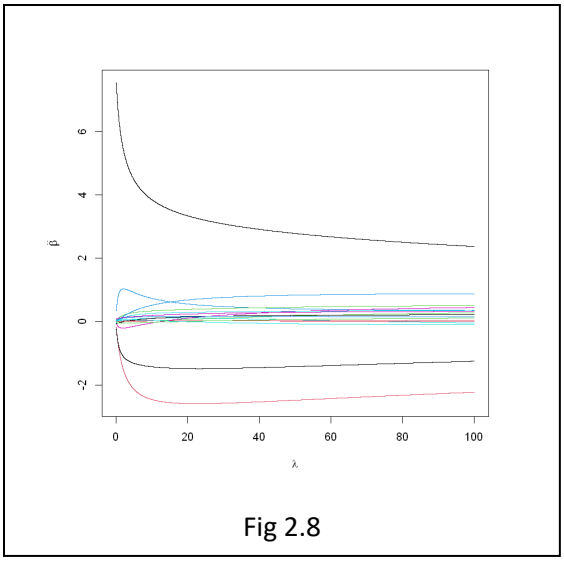

Fig 2.8

We then used LASSO to perform feature selection by choosing the lambda that minimizes Mellow's Cp criteria. (Fig 2.9) shows the relationship of penalty parameter and the standardized coefficients.

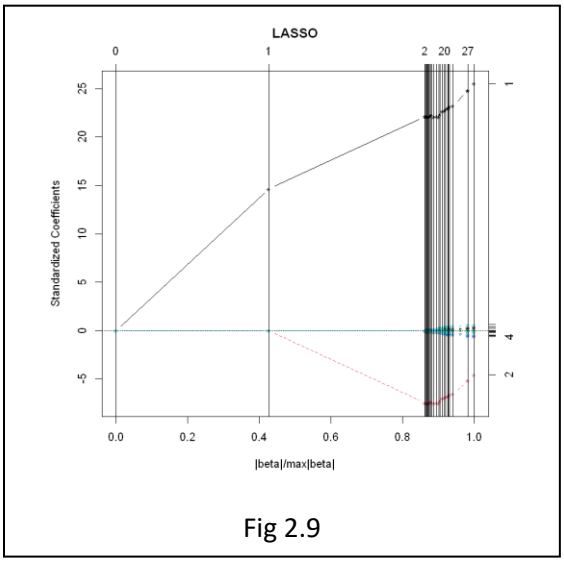

Fig 2.9

The sixth model is Principal Component Regression (PCR). We ran the linear regression on all possible PCs. (fig 2.10) shows the impact of number of parameters on the Root Mean Squared Error of Prediction (RMSEP). The full model with 17 parameters performed the best.

The last model was Partial Least Squares (PLS) Regression. The optimal number of parameters was 17 (full model)

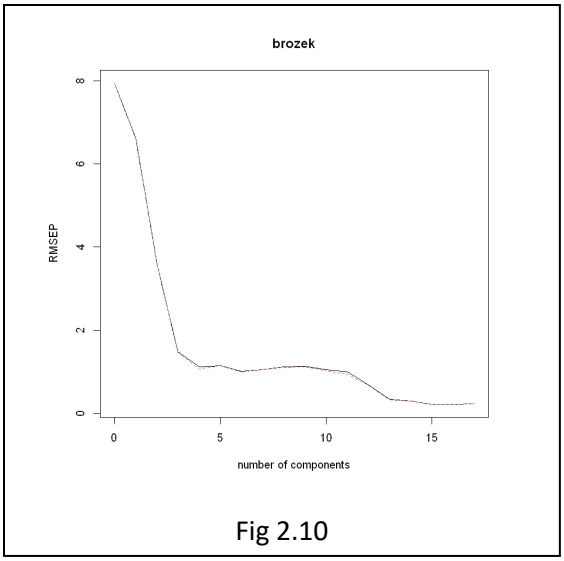

Fig 2.10

In all the methods, training and testing errors were calculated to evaluate the performance. Due to the small size of the sample (252 datapoints), we performed cross validation using the Monte Carlo method with 100 iterations to ensure the robustness of the models. The training and testing error were randomly split in the same proportions as before 90/10, respectively, at each round.

4. Results

As previously mentioned, we calculated the training and testing error of all the models, in addition to, the testing errors with CV. The results are as follows:

Model	**Training Error**	**Testing error**
Linear regression with all predictors.	**0.02930823**	0.00875598
Linear regression with the best subset of k = 5 predictors variables.	0.03146801	0.00278622
Linear regression with variables (stepwise) selected using AIC.	0.02945827	0.00895597
Ridge regression.	0.02930890	0.00885923
LASSO regression.	0.12214236	**0.00059628**
Principal component regression.	**0.02930823**	0.00875598
Partial least squares.	0.03022288	0.00837090

Model	**Testing error with CV**	**Mean variance with CV**
Linear regression with all predictors.	0.05699344	0.00641112
Linear regression with the best subset of k = 5 predictors variables.	0.04660547	0.00642014
Linear regression with variables (stepwise) selected using AIC.	0.02942632	0.00109592
Ridge regression.	0.01754126	0.00165336
LASSO regression.	**0.00635245**	**0.00009396**
Principal component regression.	0.05643630	0.00646890
Partial least squares.	0.05673612	0.00627806

5. Findings/Conclusion

From the tables above we see the both the full and the PCR models yielded the same MSE. Note that the PCR model used the full model (17 parameters). LASSO on the other hand had the largest MSE with a value of **0.12214236**. Surprisingly, LASSO (mod5) had the smallest testing error at **0.00059628.** When performing cross validation, LASSO model outperformed the rest by a large margin as evident in the tables above and fig 4.1 & 4.2.

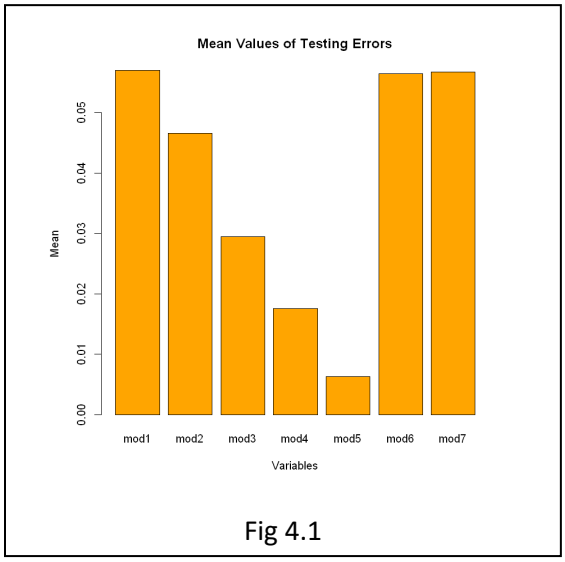

Fig 4.1

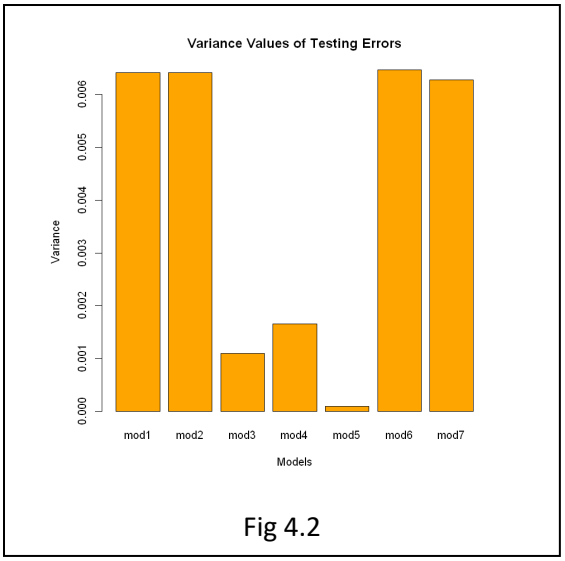

Fig 4.2

We can also notice that both mod6 (PCR) & mod7 (PLS), which used all 17 parameters, performed very similar to the full model.

6. Appendix

The analysis using R is attached for reference.