ISYE 7406 Homework 2

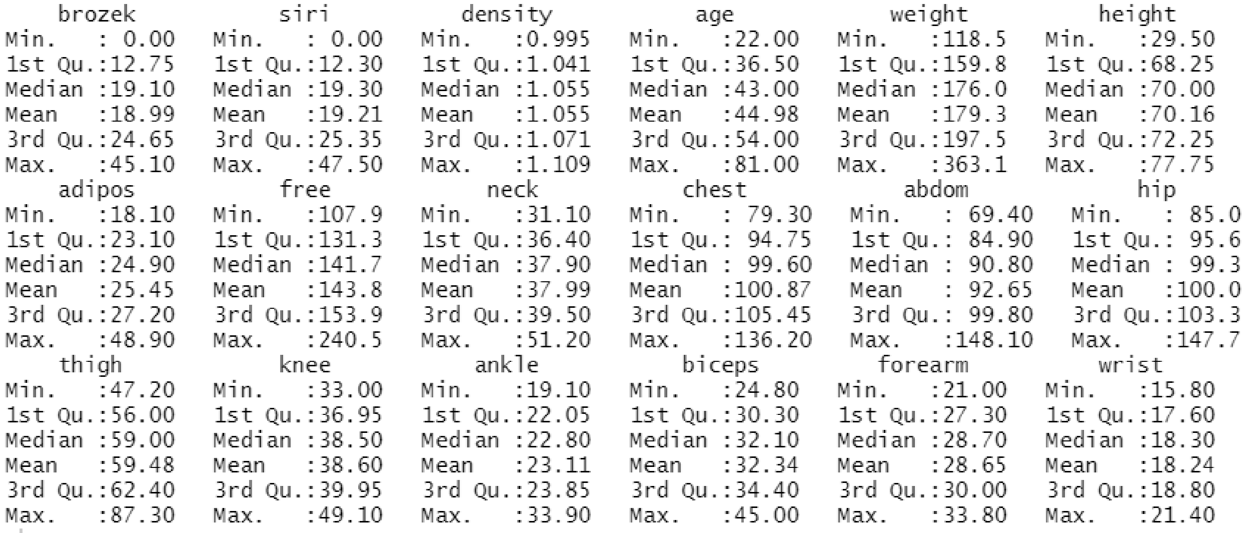
August 12, 2021

Introduction

This report will review Percentage of Body Fat and Body Measurements dataset which includes the Age, weight, height, and 10 body circumference measurements of 252 men. The purpose of analyzing the dataset is to generate models (Linear Regression w/ all predictors, Linear regression with the best subset of k = 5 predictors variables, Linear regression with variables (stepwise) selected using AIC, Ridge regression, LASSO, Principal component regression, and Partial least squares) to identify the Training & Test errors. Then similar to Homework 1, the models are evaluated using a 100-run Cross Validation to identify which model has the best Test error. The purpose of this report is to also identify other findings and visually explore the dataset for important/unusual patterns.

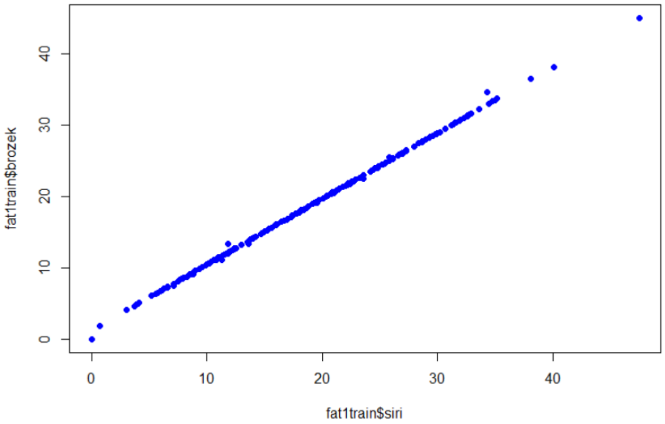
Exploratory Data Analysis

The exploratory data analysis will check for important/unusual patterns in the data through plots and summary statistics. According to the summary statistics shown below, it was interesting to see how siri (Percent body fat using Siri's equation) & brozek (Percent body fat using Brozek's equation) which is also the response variable had a minimum of 0.

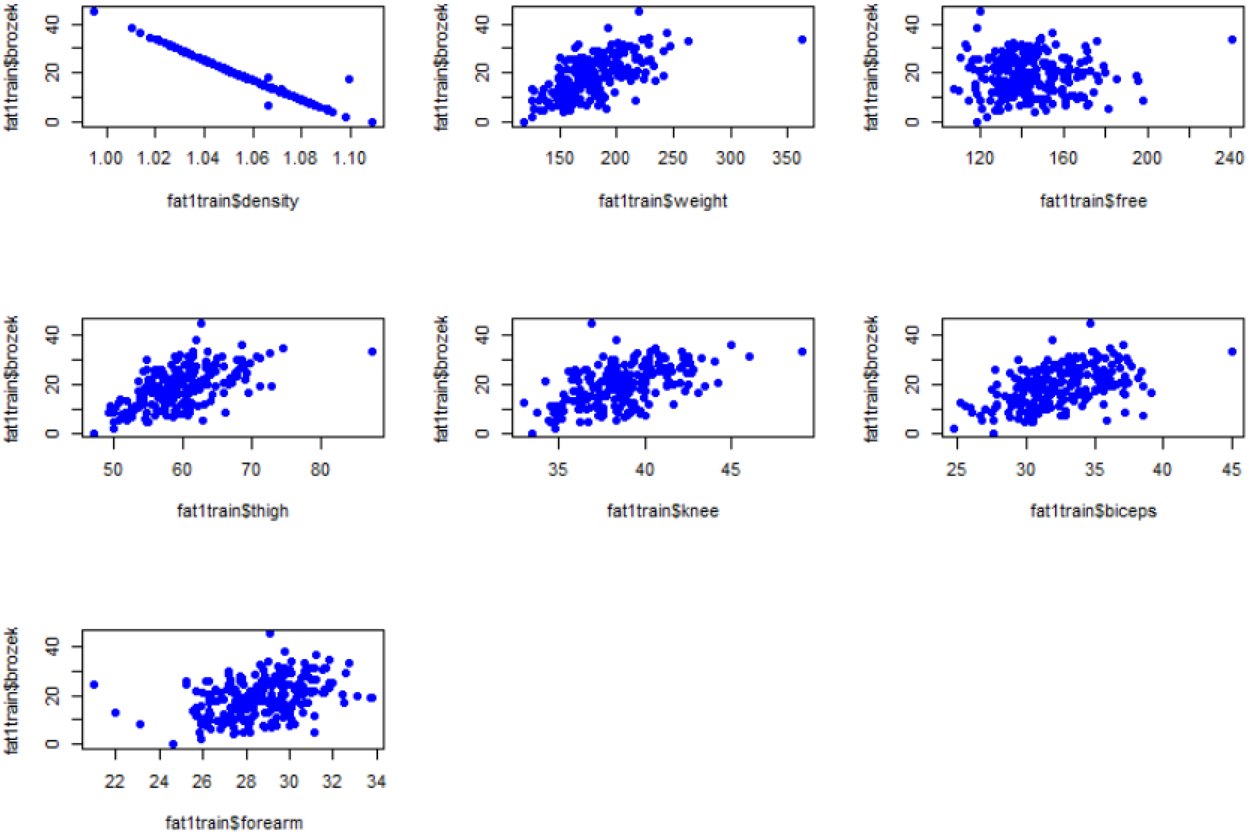


```
     brozek           siri          density           age            weight          height     
 Min.   : 0.00   Min.   : 0.00   Min.   :0.995   Min.   :22.00   Min.   :118.5   Min.   :29.50  
 1st Qu.:12.75   1st Qu.:12.30   1st Qu.:1.041   1st Qu.:36.50   1st Qu.:159.8   1st Qu.:68.25  
 Median :19.10   Median :19.30   Median :1.055   Median :43.00   Median :176.0   Median :70.00  
 Mean   :18.99   Mean   :19.21   Mean   :1.055   Mean   :44.98   Mean   :179.3   Mean   :70.16  
 3rd Qu.:24.65   3rd Qu.:25.35   3rd Qu.:1.071   3rd Qu.:54.00   3rd Qu.:197.5   3rd Qu.:72.25  
 Max.   :45.10   Max.   :47.50   Max.   :1.109   Max.   :81.00   Max.   :363.1   Max.   :77.75  
     adipos           free            neck           chest            abdom             hip       
 Min.   :18.10   Min.   :107.9   Min.   :31.10   Min.   : 79.30   Min.   : 69.40   Min.   : 85.0  
 1st Qu.:23.10   1st Qu.:131.3   1st Qu.:36.40   1st Qu.: 94.75   1st Qu.: 84.90   1st Qu.: 95.6  
 Median :24.90   Median :141.7   Median :37.90   Median : 99.60   Median : 90.80   Median : 99.3  
 Mean   :25.45   Mean   :143.8   Mean   :37.99   Mean   :100.87   Mean   : 92.65   Mean   :100.0  
 3rd Qu.:27.20   3rd Qu.:153.9   3rd Qu.:39.50   3rd Qu.:105.45   3rd Qu.: 99.80   3rd Qu.:103.3  
 Max.   :48.90   Max.   :240.5   Max.   :51.20   Max.   :136.20   Max.   :148.10   Max.   :147.7  
     thigh           knee           ankle          biceps         forearm          wrist      
 Min.   :47.20   Min.   :33.00   Min.   :19.10   Min.   :24.80   Min.   :21.00   Min.   :15.80  
 1st Qu.:56.00   1st Qu.:36.95   1st Qu.:22.05   1st Qu.:30.30   1st Qu.:27.30   1st Qu.:17.60  
 Median :59.00   Median :38.50   Median :22.80   Median :32.10   Median :28.70   Median :18.30  
 Mean   :59.48   Mean   :38.60   Mean   :23.11   Mean   :32.34   Mean   :28.65   Mean   :18.24  
 3rd Qu.:62.40   3rd Qu.:39.95   3rd Qu.:23.85   3rd Qu.:34.40   3rd Qu.:30.00   3rd Qu.:18.80  
 Max.   :87.30   Max.   :49.10   Max.   :33.90   Max.   :45.00   Max.   :33.80   Max.   :21.40  
```

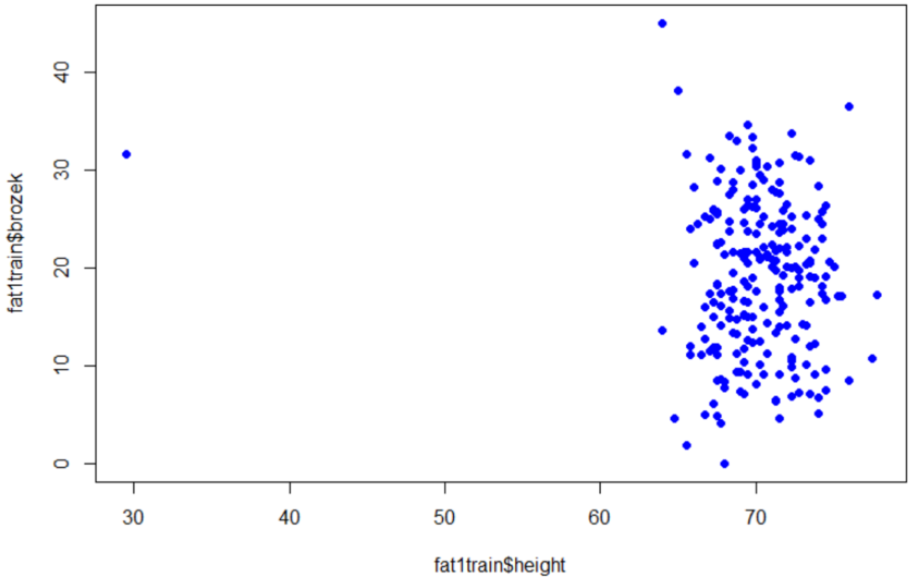
I explored this further through a scatterplot and saw that they have a definitive positive correlation. This makes sense since their computations are fairly similar.



There are 16 other variables that I could scatterplot but I decided to instead run a Linear Regression with all predictors to see which predictor variables were significant based on their P-values. Based on the Regression summary results, I only plotted the significant variables listed below. It's worth noting that Density had a very definitive negative correlation but weight, thigh, knee, and biceps have a positive correlation with brozek. However, free (fat free weight) had no correlation.



I noticed in the summary statistics that height had an unusual pattern so I decided to also scatterplot this pattern as well. It was interesting to see how someone who is ~30 inches in height had the same amount of body fat% compared to men who are 70 inches in height (more than twice as much).



Methodology – Modeling for Training and Test errors

I performed the following (Linear Regression w/ all predictors, Linear regression with the best subset of k = 5 predictors variables, Linear regression with variables (stepwise) selected using AIC, Ridge regression, LASSO, Principal component regression, and Partial least squares) to retrieve training and test errors when predicting brozek. The testing data is used to evaluate the model's true performance since the training data may have excluded/included certain data that may have a significant positive/negative affect on the training error. I then performed a 100-run cross validation for each model to obtain the test errors and observe which one has the best test error. Running a Monte Carlo cross-validation helps reduce the variance compared to just running one test run and depict which model is truly the most accurate on a large scale.

Results & Findings

<u>Linear Regression w/ all predictors (i)</u> – When running the Linear Regression model, the testing error resulted at 0.008755981 which was a lot smaller compared to the training error of 0.02930823. When running the cross-validation, the testing error increased to 0.05094221 which was the 3rd largest testing error compared to the other models. The adjusted R-squared resulted at 0.9995 which is ridiculously high with a residual standard error of 0.1784

<u>Linear Regression w/ subset of k=5 (ii)</u> – The results showed that the best subset included predictor variables (siri, density, thigh, knee, and wrist) and it's worth noting that it also had an adjusted R-squared of 0.9995 but had a residual standard error of 0.1798 which is larger than the previous model with all predictors. However, you'll see that the training error resulted at 0.0314 which is larger than the Linear Regression w/ all predictors but the test error resulted at

0.002786218 which was a lot smaller compared to the previous model. The cross validation run performed a testing error of 0.04080453 which is very large compared to our previous test error but closer to the cross-validation test error from model (i). This is a great reason why we run cross-validation because running a single test error can sometimes have a large variance compared to running multiple which is why we average the test errors from each run. This model was the runner-up for the "Best Performing Model" award but it was slightly beaten by one of the other models soon to come.

Linear regression with variables (stepwise) selected using AIC (iii) – The Akaike Information Criterion (AIC) is used when trying to decide between multiple different model types but includes a penalty that increases based on the number of parameters. The AIC model chose predictor values (siri, density, weight, adipos, free, thigh, knee, biceps, forearm, and wrist) and I retrieved the training error of 0.02945827 & testing error of 0.008955971 which surprisingly was both worse than model (i). This however could have been due to variance of the model so after running the cross-validation, the testing error resulted in 0.05348192 which was still higher than model (i). This means that it makes more sense to just run model (i) instead of this model when it comes to this specific dataset.

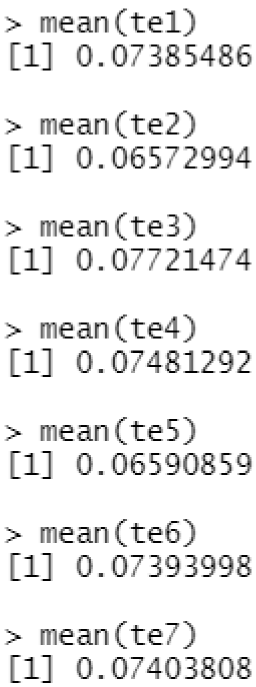
Ridge Regression (iv) – Ridge regression is useful when the predictor variables experience high collinearity. This model had a training error of 0.02930890 and a testing error of 0.008859234 which is slightly higher compared to model (i) on both accounts. When applying cross-validation, the testing error also resulted higher at 0.05158402.

LASSO (v) – Lasso is useful when the true model is sparse with a large number of predictor variables. This model had a training error of 0.03085618 which is larger compared to models (i, iii, & iv) but has a smaller testing error of 0.003158102. This is also a great example of whether cross-validation can provide a more accurate depiction. The cross-validation resulted in a test terror of 0.4026165 which was the best test error compared to all other models. I thought this was odd since there wasn't a large number of predictor variables in this model so I decided to run a different seed (set.seed =1, 2, & 8) and saw that the LASSO model still outperformed Ridge Regression. I still find this odd because it seems like Ridge would be more appropriate since the predictor variables are somewhat collinear such as (height and weight) but that is only 2 of 17 predictor variables so it may not have been enough to truly be helpful for the Ridge regression model.

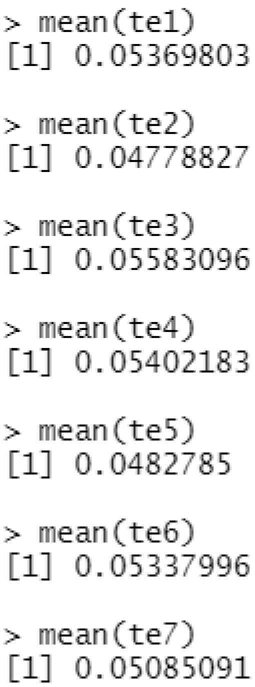
Principal Component Regression (vi) – The purpose of the Principal Component Regression is to estimate the unknown regression coefficients through linear regression. However, we concentrate on using this when the x variables are highly correlated/collinear and we eliminate any columns that don't have a significant variance. The results show that the optimal # of components for the PCR model is 17. Interestingly, the training and testing errors result in the same output as model (i) of 0.02930823 & .008755981 respectively. However, I tried different seeds and noticed that this does not consistently happen (example set.seed = 1 ends up with a higher result). After performing cross-validation, the test error resulted at 0.05073222 which ranked 4th of the 7 models.

Partial Least Squares (vii) – This is similar to the PCR method but instead it is supervised (PCR is unsupervised). It was interesting to see that the training and test errors were also the same compared to the full model (model i) so I checked a different seed (seed=1) and saw that it performed better than PCR on both accounts. After running the cross-validation, the test error 0.04980483 also resulted in a better result compared to PCR which ranked 3rd.

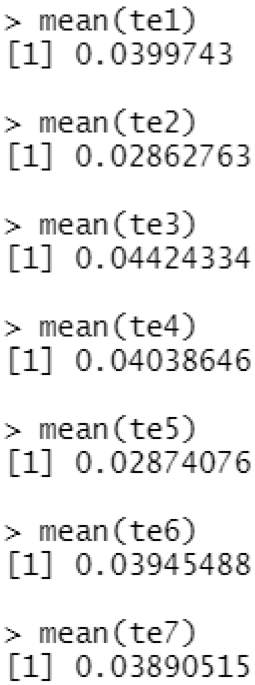
When analyzing all the cross-validated models, it is clear that the LASSO model performed the best. However, I doubled checked with 3 different seeds (1,2,3) shown below and it was interesting to see how all seeds had Linear Regression with subset k=5 perform the best. This leads me to believe that although seed (7406) had LASSO as the winner, it is not definitive unless we run a 100-run cross-validation of different seeds.



```
> mean(te1)
[1] 0.07385486

> mean(te2)
[1] 0.06572994

> mean(te3)
[1] 0.07721474

> mean(te4)
[1] 0.07481292

> mean(te5)
[1] 0.06590859

> mean(te6)
[1] 0.07393998

> mean(te7)
[1] 0.07403808
```



```
> mean(te1)
[1] 0.05369803

> mean(te2)
[1] 0.04778827

> mean(te3)
[1] 0.05583096

> mean(te4)
[1] 0.05402183

> mean(te5)
[1] 0.0482785

> mean(te6)
[1] 0.05337996

> mean(te7)
[1] 0.05085091
```



```
> mean(te1)
[1] 0.0399743

> mean(te2)
[1] 0.02862763

> mean(te3)
[1] 0.04424334

> mean(te4)
[1] 0.04038646

> mean(te5)
[1] 0.02874076

> mean(te6)
[1] 0.03945488

> mean(te7)
[1] 0.03890515
```