**1.Data Manipulation**

I used the following code to perform the Data manipulation:



This yields the following results:

|  |  |
| --- | --- |
| Number of missing columns removed : | 3 |
| Number of missing rows removed : | 4 |
| Number of columns with constant predictors: | 0 |
| Number of collinear columns removed: | 16 |

**2. Multiple Imputation**

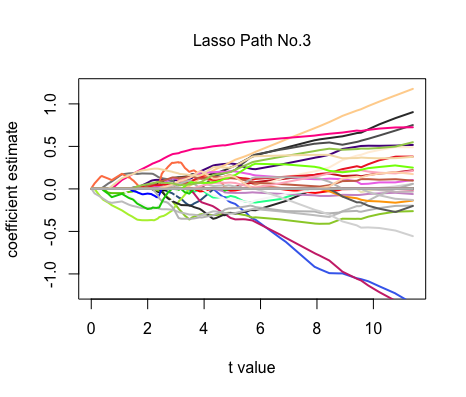
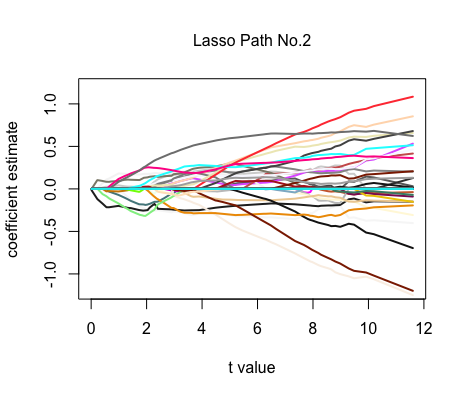
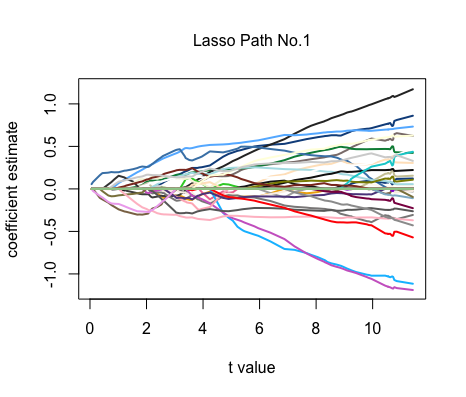
Multiple imputation is a statistical technique for analyzing incomplete data sets, that is, data sets for which some entries are missing. As a result, multiple imputation allows us to perform complete dataset analyses (which many/most statistical and machine learning procedures require), and obtain corresponding estimates of interest on each complete dataset. In this project, three imputed datasets are created through the step() function. This yields a new dataset of 150 new predictors.

**3. Lasso via Modification to Least Angle Regression**

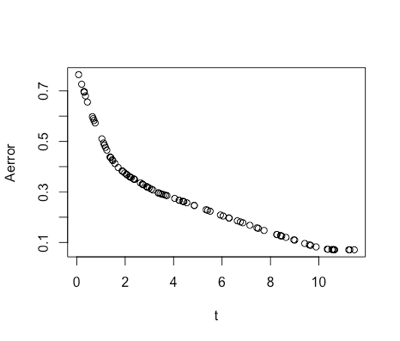
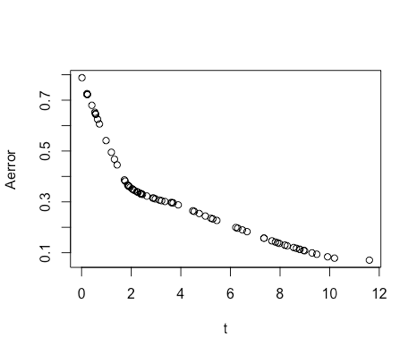
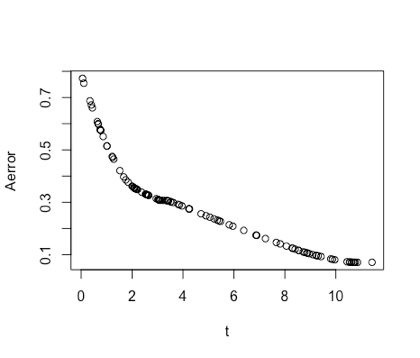
The lasso algorithm proposed by Efron is realized through R.

|  |  |  |
| --- | --- | --- |
| Dataset1 | Dataset2 | Dataset3 |

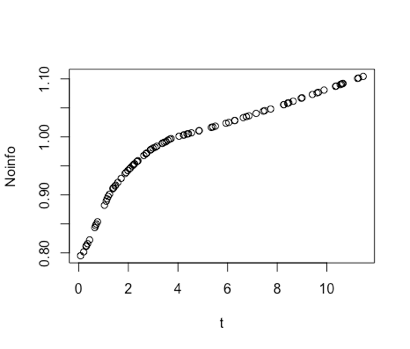
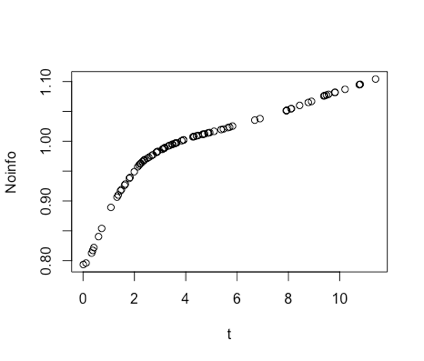
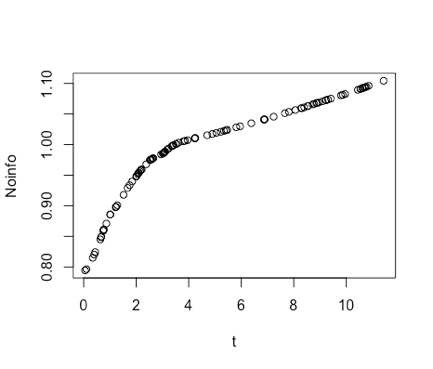
**Lasso Path**



**apparent mean absolute predictive error rate**

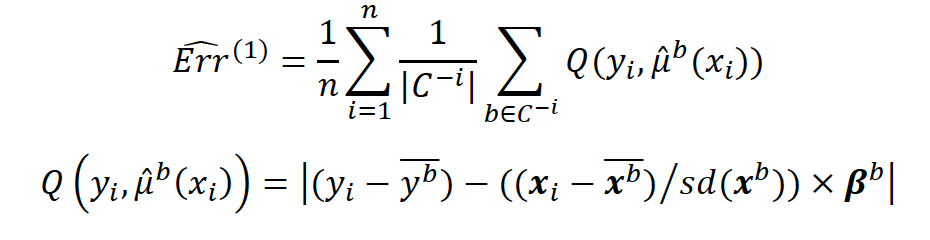


**the no information error rate**



**4. ACCURACY ASSESSMENT AND MODEL SELECTION**

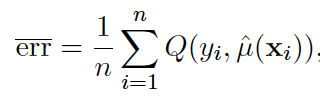
For each multiply imputed dataset, we generate B=25 bootstrap samples. Then we use them to calculate the so-called leave-one-out bootstrap estimate predictive error for every t:



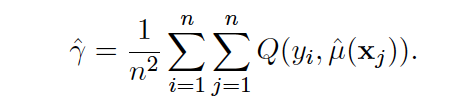
Since we have different 𝑡=Σ|𝛽𝑖| for different bootstrap sample. We need to use linear interpolation in order to get the same t values.

We compute the 0.632+ bootstrap estimate for every data set based on the relative overfitting rate, no information error rate, apparent error rate and the leave-one-out bootstrap estimate of error.

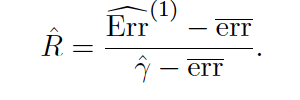
Apparent error rate:



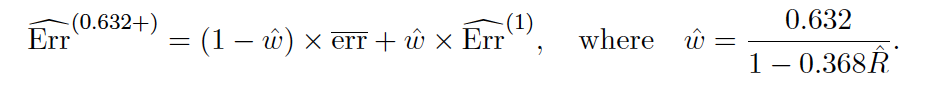
No information error rate:



Relative rate of overfitting:

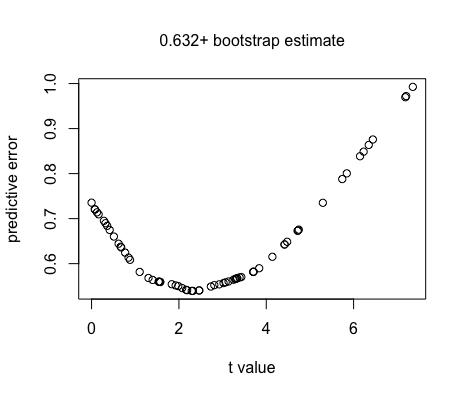
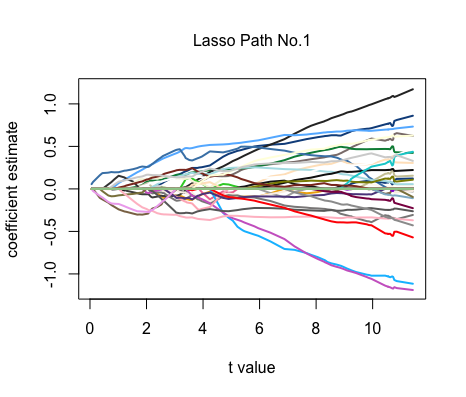


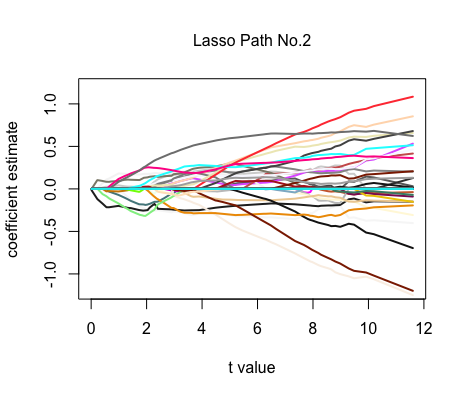
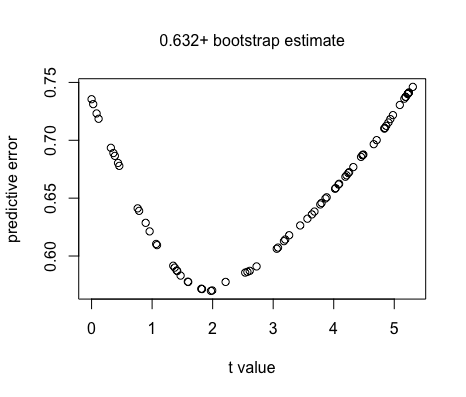
The 0.632+ bootstrap estimate of predictive error:

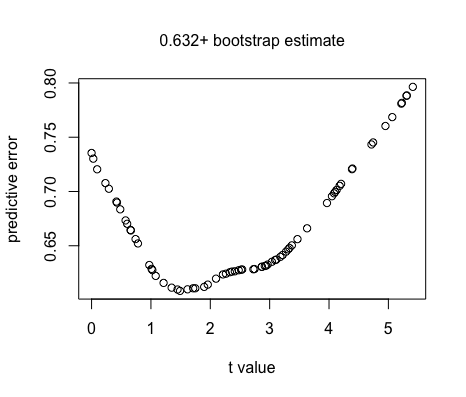
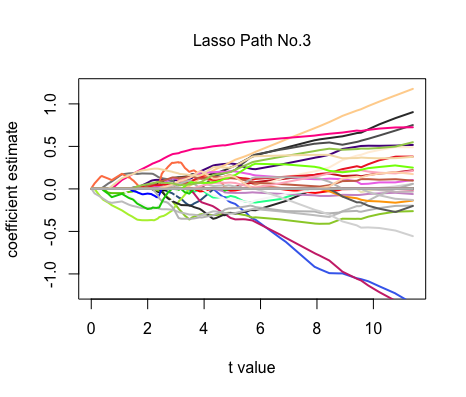


**5.Visualization**

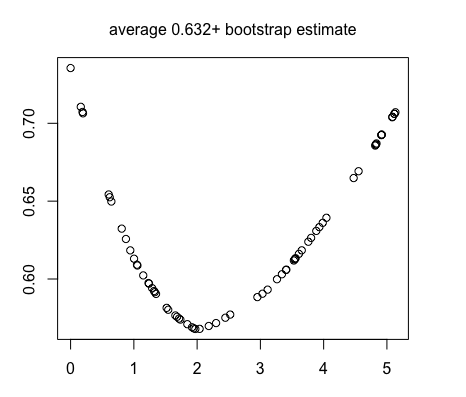
We plot a 3 row by 2 column panel plot: the first column contains the lasso paths for each of the datasets. The second column contains the 0.632+ bootstrap estimate of expected mean absolute error for the 3 data sets



****



average estimate of expected mean absolute error across three multiple imputation datasets is plotted.



**6.Interpretation**

In this example, we choose t = 2. Then I averaged the Coefficients of the regression to assess the quality of the fit. I calculated the error between the actual response and the prediction.

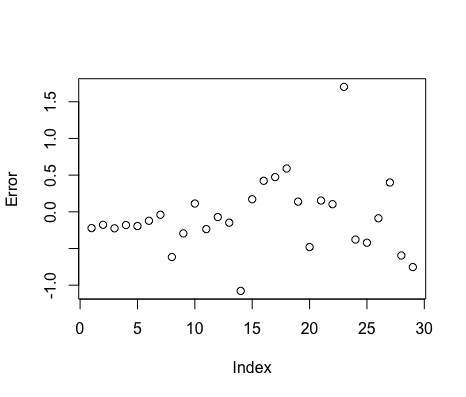
The observations 14, 24 weren’t fitted very well (largest error).

The signs of the coefficients are very important. They show which predictors predict an increase or a decrease of the PFS . We should be cautious about the interpretation of the absolute values of the coefficients. In fact, their impact depends on the range of the predictors.

The predictor " v1 High.Intensity.Large.Zone.Emphasis average " has the largest coefficient 0.281901

The predictor " v1 Texture.Strength average " has the largest absolute coefficient with value -0.287835

If the coefficient is positive, the increase of the corresponding predictor will increase the PFS, vice versa. The increase of 1 unit of the predictor(Texture Strength) will lead the PFS to decrease 0.75



[1] "v1 Correlation1 average"

[2] "v1 Maximum.Probability average"

[3] "v1 Information.Measure.of.Correlation1 average"

[4] "v1 Homogeneity1.or.Inverse.Difference average"

[5] "v1 Cluster.Prominence.1 average"

[6] "v1 Coarseness average"

[7] "v1 Complexity average"

[8] "v1 Texture.Strength average"

[9] "v1 Entropy.2 average"

[10] "v1 Mean.SD average"

[11] "v1 Large.Zone.Emphasis average"

[12] "v1 Zone.Percentage average"

[13] "v1 High.Intensity.Large.Zone.Emphasis average"

[14] "v1 Cluster.Shade difference"

[15] "v1 Sum.of.Squares.or.Variance difference"

[16] "v1 Sum.Variance difference"

[17] "v1 Difference.Variance difference"

[18] "v1 Information.Measure.of.Correlation1 difference"

[19] "v1 Inverse.Difference.Moment.Normalized difference"

[20] "v1 Contrast.1 difference"

[21] "v1 Cluster.Shade.1 difference"

[22] "v1 Sum.Average.1 difference"

[23] "v1 Sum.Variance.1 difference"

[24] "v1 Difference.Entropy.1 difference"

[25] "v1 Information.Measure.of.Correlation1.1 difference"

[26] "v1 Inverse.Difference.Moment.Normalized.1 difference"

[27] "v1 Long.Run.Emphasis difference"

[28] "v1 Coarseness difference"

[29] "v1 Texture.Strength difference"

[30] "v1 Entropy.2 difference"

[31] "v1 Maximum.Intensity difference"

[32] "v1 Mean.Intensity difference"

[33] "v1 High.Intensity.Zone.Emphasis difference"

R Code

**R Code**















