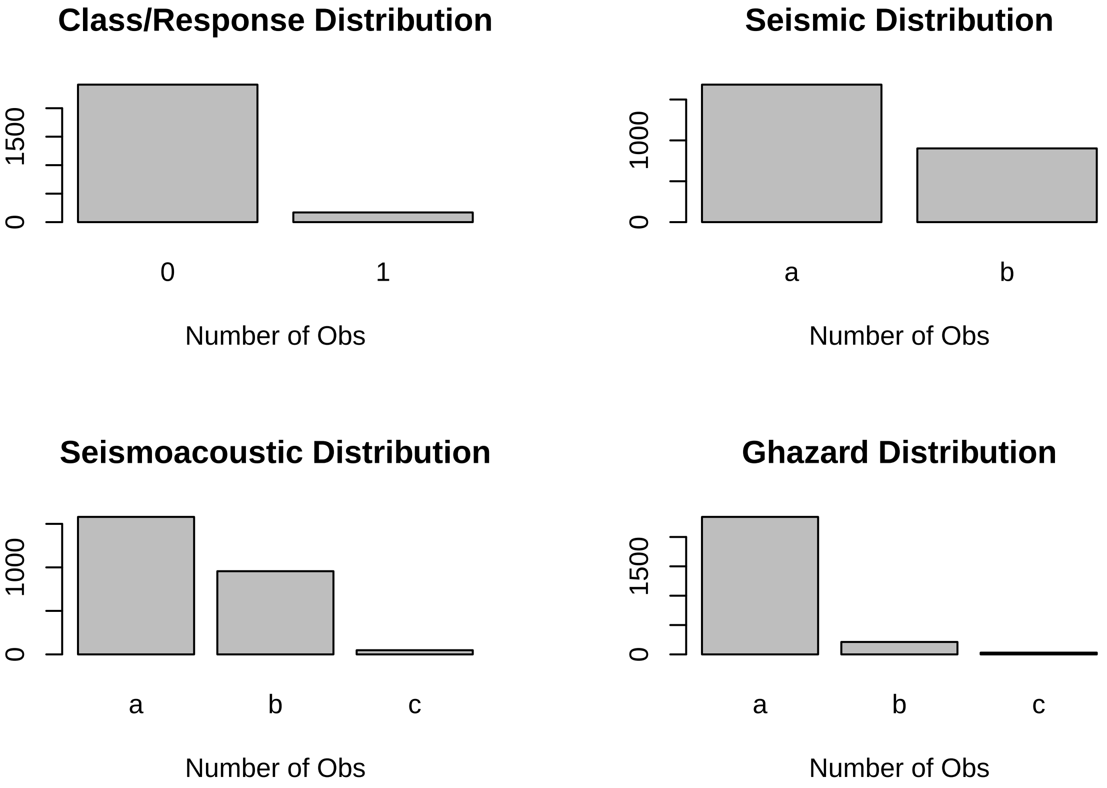
557_Project_2BS

Ben Straub, Hillary Koch, Jiawei Huang, Arif Masrur

3/15/2017

No Command Lines Ever. Whoa

What the Factor Variables look like



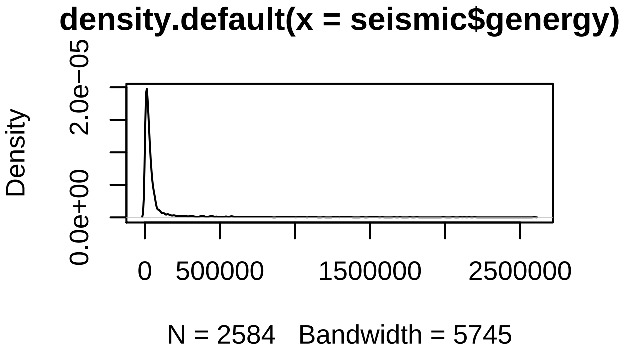
Class/Response Distribution

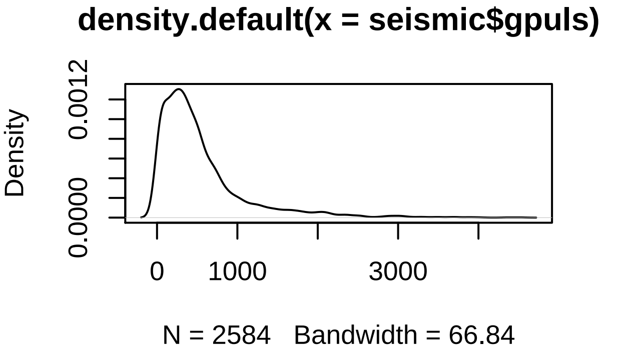
Seismic Distribution

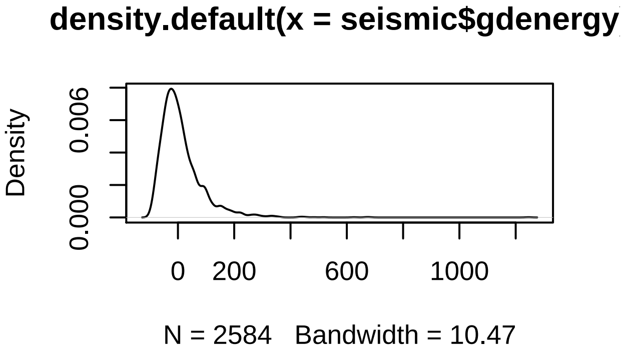
Seismoacoustic Distribution

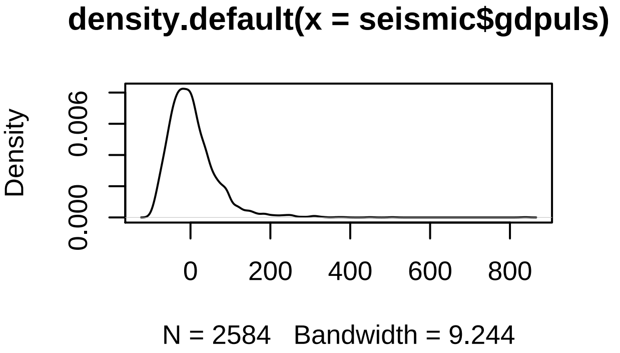
Ghazard Distribution

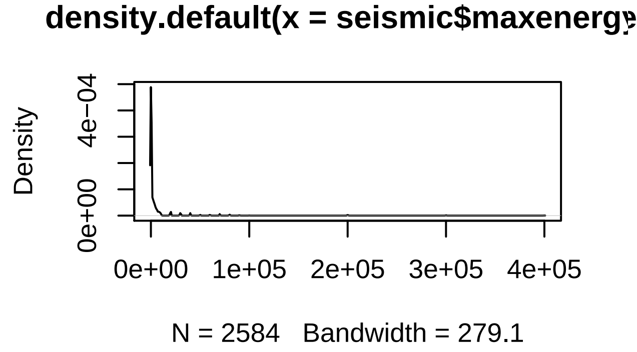
What the Continuous Variables look like

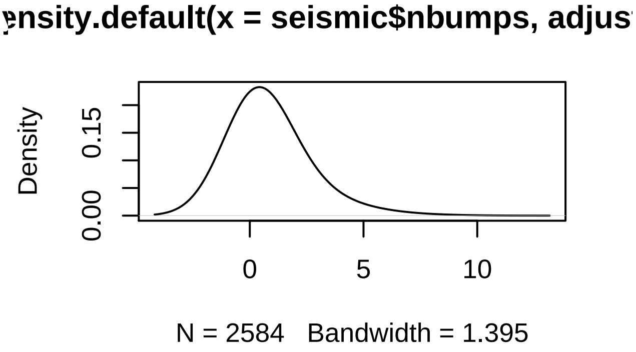


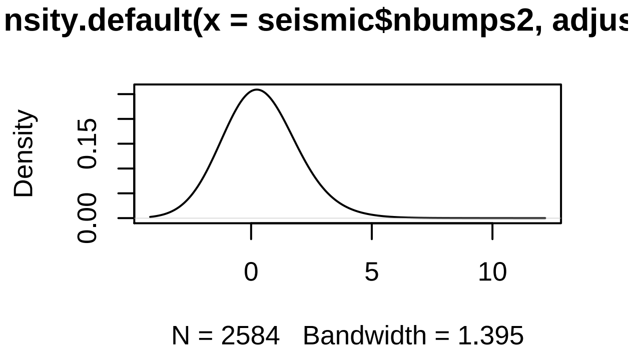


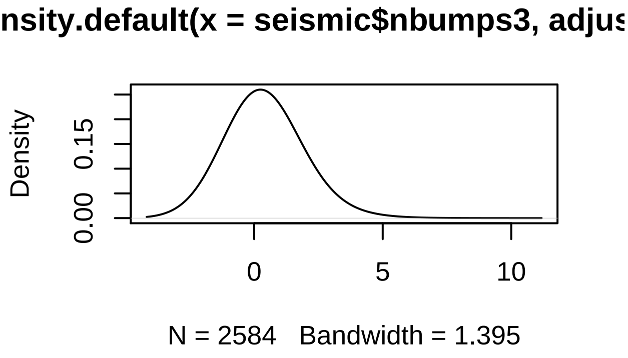








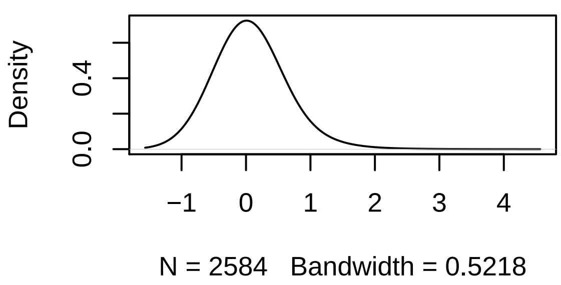




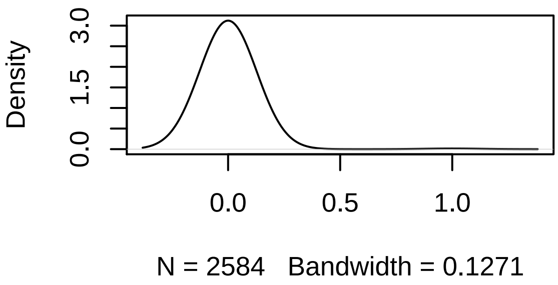
```
Call:
```

```
lm(formula = class ~ ., data = seismic)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.57549 -0.07778 -0.03812 -0.00950  1.03232 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -2.393e-02  2.565e-02  -0.933  0.35090    
seismic         1.869e-02  1.076e-02   1.737  0.08254 .  
seismoacoustic  2.610e-03  1.002e-02   0.260  0.79457    
shift           6.190e-04  1.157e-02   0.054  0.95732    
genergy        -8.698e-08  3.459e-08  -2.514  0.01199 *  
gpuls           1.019e-04  1.670e-05   6.102  1.2e-09 ***
gdenergy       -6.943e-05  1.006e-04  -0.690  0.49009    
gdpuls         -1.942e-04  1.368e-04  -1.420  0.15583    
ghazard        -1.394e-02  1.608e-02  -0.867  0.38618    
nbumps          4.674e-01  1.680e-01   2.783  0.00543 ** 
nbumps2        -4.282e-01  1.682e-01  -2.546  0.01096 *  
nbumps3        -4.260e-01  1.681e-01  -2.535  0.01131 *  
nbumps4        -4.622e-01  1.708e-01  -2.706  0.00685 ** 
nbumps5        -2.963e-01  2.332e-01  -1.270  0.20408    
energy          2.536e-07  2.395e-06   0.106  0.91568    
maxenergy      -1.054e-06  2.333e-06  -0.452  0.65164    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2371 on 2568 degrees of freedom
Multiple R-squared:  0.09128,   Adjusted R-squared:  0.08597 
F-statistic:  17.2 on 15 and 2568 DF,  p-value: < 2.2e-16
```

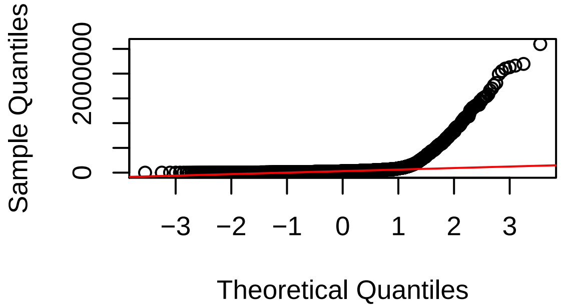
nsity.default(x = seismic$nbumps4, adjus



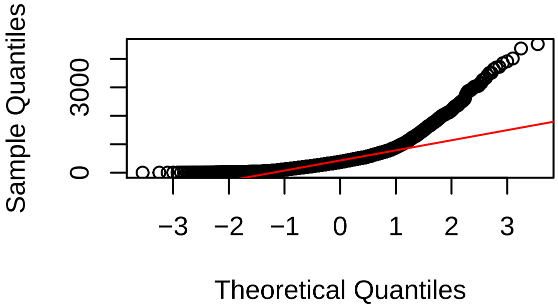
nsity.default(x = seismic$nbumps5, adjus



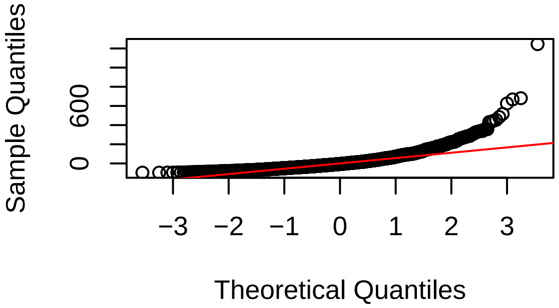
Normal Q−Q Plot



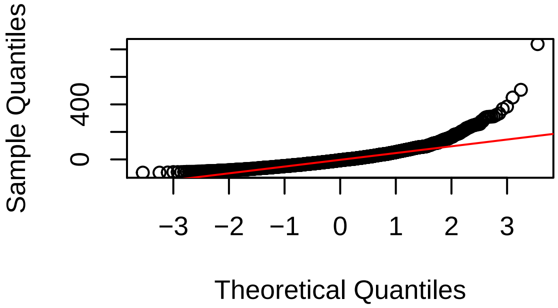
Normal Q−Q Plot



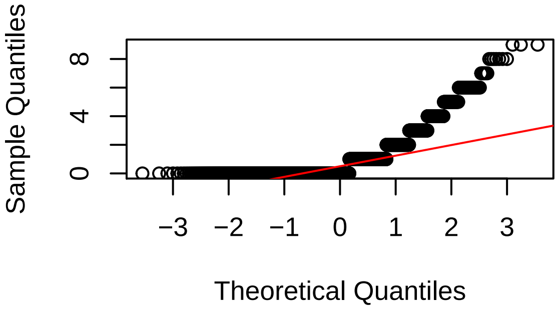
Normal Q−Q Plot



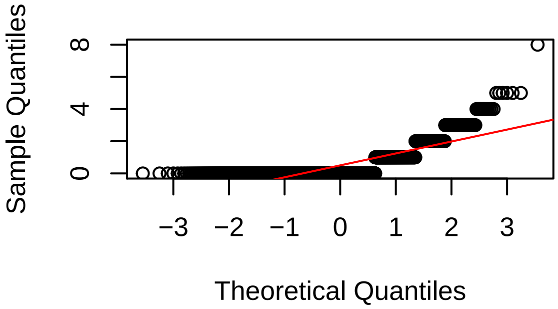
Normal Q−Q Plot

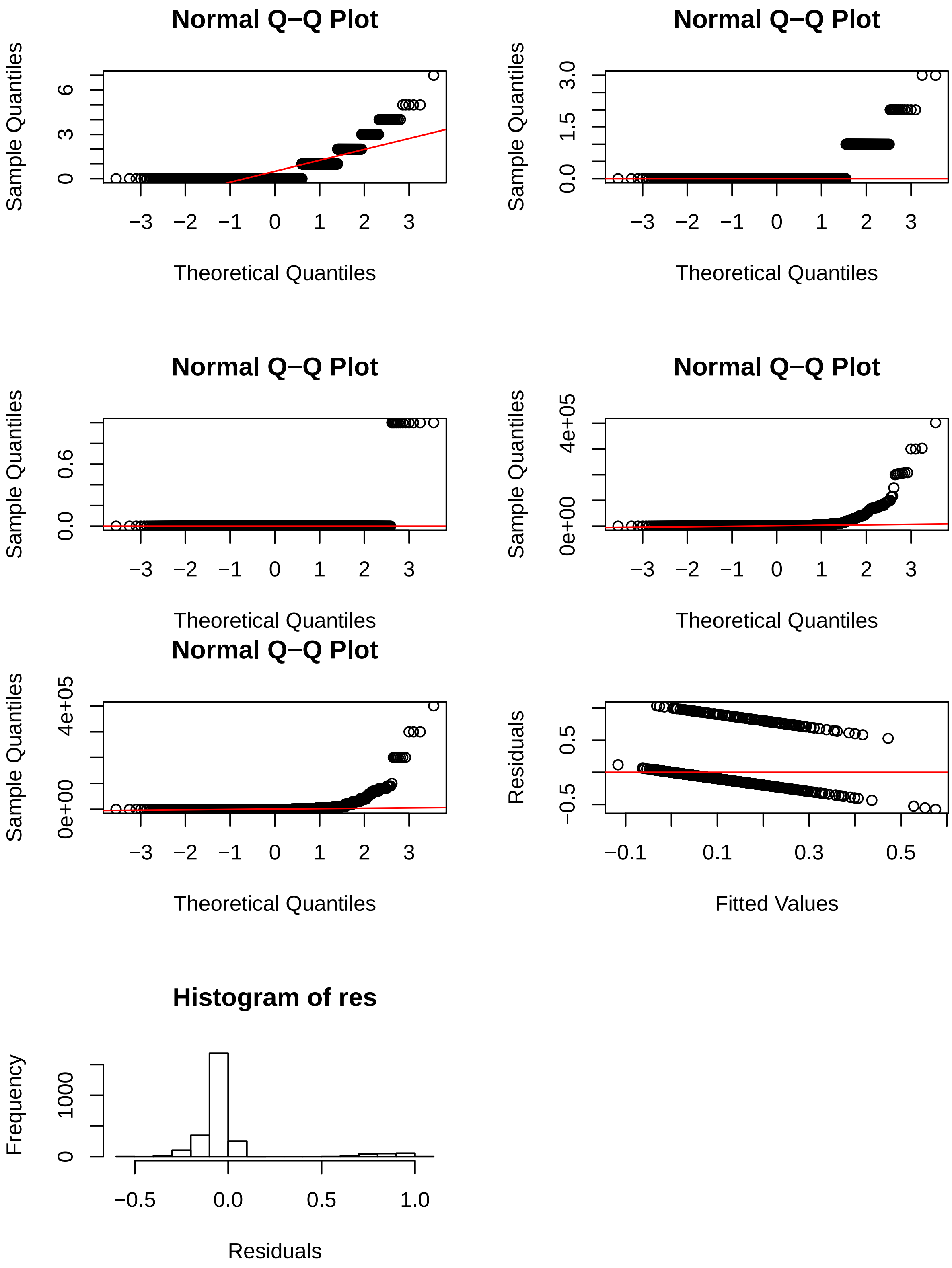


Normal Q−Q Plot



Normal Q−Q Plot



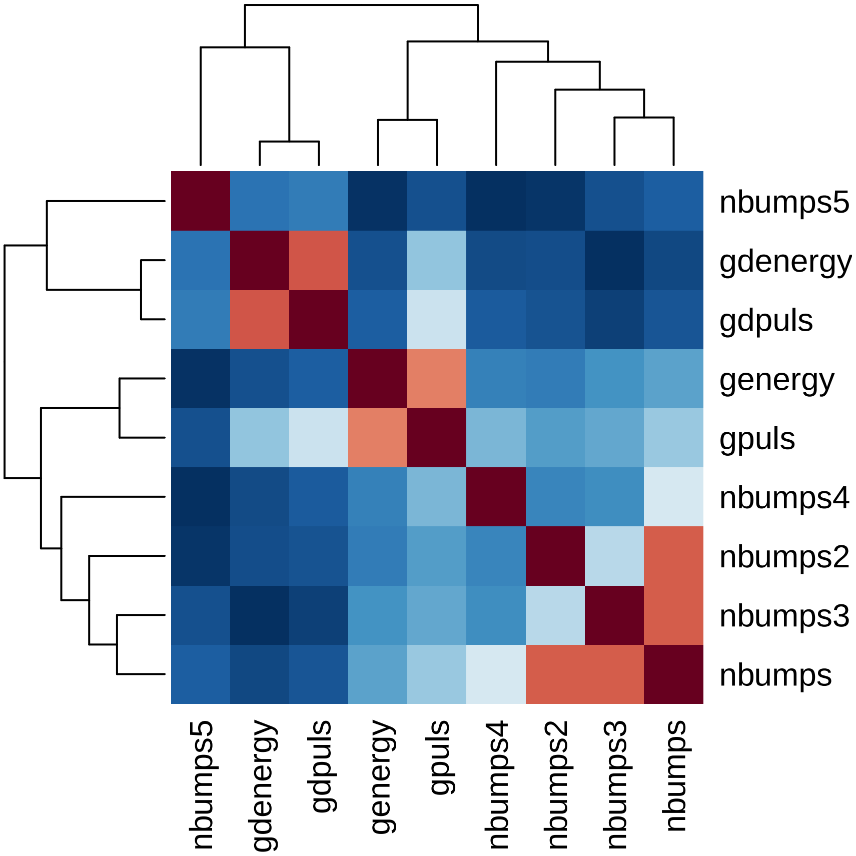

Normal Q−Q Plot
Sample Quantiles
0
3
6
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Normal Q−Q Plot
Sample Quantiles
0.0
1.5
3.0
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Normal Q−Q Plot
Sample Quantiles
0.0
0.6
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Normal Q−Q Plot
Sample Quantiles
0e+00
4e+05
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Normal Q−Q Plot
Sample Quantiles
0e+00
4e+05
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Residuals
−0.5
0.5
−0.1
0.1
0.3
0.5
Fitted Values
Histogram of res
Frequency
0
1000
−0.5
0.0
0.5
1.0
Residuals

Lots of multicollinearity to worry about during variable selection

```r
vif(fit)
```

```
##        seismic seismoacoustic          shift        genergy          gpuls
##       1.209814       1.286183       1.411216       2.889651       4.057018
##       gdenergy         gdpuls        ghazard         nbumps        nbumps2
##       3.000282       3.430524       1.395598    2414.689538     798.964152
##        nbumps3        nbumps4        nbumps5         energy      maxenergy
##     769.131960     104.402690      11.562237     110.283444      93.762895
```

Correlation of the Variables



```
$r
         genergy gpuls nbumps4 nbumps3 nbumps nbumps2 nbumps5 gdenergy
genergy        1
gpuls       0.75     1
nbumps4     0.15  0.26       1
nbumps3     0.19  0.23    0.18       1
nbumps      0.22   0.3     0.4     0.8      1
nbumps2     0.14  0.21    0.16    0.35    0.8       1
nbumps5  -0.0099 0.049  -0.017   0.046   0.07 -0.0053       1
gdenergy   0.049  0.29   0.037  -0.012   0.03   0.041    0.12        1
```

```
gdpuls     0.072  0.38   0.066   0.015  0.058   0.051    0.14     0.81
         gdpuls
genergy
gpuls
nbumps4
nbumps3
nbumps
nbumps2
nbumps5
gdenergy
gdpuls        1

$p
         genergy gpuls nbumps4 nbumps3 nbumps nbumps2 nbumps5 gdenergy
genergy        0
gpuls          0     0
nbumps4  1.4e-14     0       0
nbumps3        0     0       0       0
nbumps         0     0       0       0      0
nbumps2  2.2e-13     0       0       0      0       0
nbumps5     0.62 0.012     0.4   0.018  4e-04    0.79       0
gdenergy   0.014     0   0.061    0.54    0.13   0.036 3.3e-10        0
gdpuls   0.00027     0 0.00076    0.45 0.0032  0.0094 5.9e-13        0
         gdpuls
genergy
gpuls
nbumps4
nbumps3
nbumps
nbumps2
nbumps5
gdenergy
gdpuls        0

$sym
         genergy gpuls nbumps4 nbumps3 nbumps nbumps2 nbumps5 gdenergy
genergy  1
gpuls    ,       1
nbumps4                1
nbumps3                        1
nbumps                 .       ,       1
nbumps2                        .       ,      1
nbumps5                                               1
gdenergy                                                      1
gdpuls           .                                            +
         gdpuls
genergy
gpuls
nbumps4
nbumps3
nbumps
nbumps2
nbumps5
gdenergy
```

```
gdpuls    1
attr(,"legend")
[1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

Separating into Test and Training Sets

```
##------------------------------------
## Setting up Test and Training Sets
##------------------------------------

n <- dim(seismic)[1]
p <- dim(seismic)[2]

set.seed(2016)
test <- sample(n, round(n/4))
train <- (1:n)[-test]
seismic.train <- seismic[train,]
seismic.test <- seismic[test,]

dim(seismic)
```

```
[1] 2584   16
```

```
dim(seismic.train)
```

```
[1] 1938   16
```

```
dim(seismic.test)
```

```
[1] 646  16
```

```
#View(seismic.train)
#View(seismic.test)
```

Linear regression of an indicator matrix

```
##----------------------------------------
## Linear regression of indicator matrix
##----------------------------------------

responseY <- seismic$class
predictorX <- seismic[,-16]

# Following Le Bao's code
class1 <- which(responseY==1)
class0 <- which(responseY==0)
```

```
Y <- matrix(data = rep(0,length(responseY)*2),nrow = length(responseY))
Y[class0,1] <- 1
Y[class1,2] <- 1

betaHat <- solve(t(as.matrix(predictorX))%*%as.matrix(predictorX))%*%t(as.matrix(predictorX))%*%Y
Y1 <- as.matrix(predictorX)%*%betaHat[,1]
Y2 <- as.matrix(predictorX)%*%betaHat[,2]

pred.mx <- cbind(Y1,Y2)
pred <- rep(NA,length(Y1))
for(i in 1:length(Y1)){
  pred[i] <- which.max(pred.mx[i,]) - 1
}

# Confusion matrix
mx <- cbind(pred,responseY,pred-responseY)

confusion <- matrix(rep(NA,4), nrow = 2)
correct <- which(mx[,3] == 0)
confusion[1,1] <- length(which(mx[correct,1] == 0))
confusion[2,2] <- length(which(mx[correct,1] == 1))
confusion[1,2] <- length(which(mx[,3] == -1))
confusion[2,1] <- length(which(mx[,3] == 1))
confusion
```

```
##      [,1] [,2]
## [1,] 2411  169
## [2,]    3    1
```

```
sensitivity <- confusion[2,2]/sum(confusion[,2])
specificity <- confusion[1,1]/sum(confusion[,1])
error.rate <- (confusion[1,2] + confusion[2,1])/sum(confusion)
c(sensitivity, specificity, error.rate)
```

```
## [1] 0.005882353 0.998757249 0.066563467
```

Linear Discriminant Analysis on full model

Quadratic Discriminant Analysis -INCOMPLETE

```
##-------------------------------------
## Fit QDA model
##-------------------------------------

## Currently, can't perform QDA.  This is probably due to multicollinearity in the model
## (can't invert covariance matrix) but should be possible after variable selection

#qda.fit <- qda(class~., data = seismic, subset = train)
```

Regularized Discriminant Analysis -INCOMPLETE

```
##-------------------------------------
## Fit RDA model
##-------------------------------------

## Currently, can't perform RDA.  This is probably due to multicollinearity in the model
## (can't invert covariance matrix) but should be possible after variable selection

rda.fit <- rda(class~., data=seismic.train)

# Using Training model on train Data
rda.pred=predict(rda.fit, seismic.train, type="response")

rda.class.train <- rda.pred$class

posterior.train <- rda.pred$posterior
truth.train <- as.integer(seismic.train$class)

## Confusion matrix
rda.train.confusion <- table(rda.class.train,seismic.train$class)
rda.train.sensitivity <- rda.train.confusion[2,2]/sum(rda.train.confusion[,2])
rda.train.specificity <- rda.train.confusion[1,1]/sum(rda.train.confusion[,1])

# Sensitivity is slightly worse here
rda.train.confusion
```

```
##
## rda.class.train    0    1
##               0 1785  126
##               1   22    5
```

```
rda.train.sensitivity
```

```
## [1] 0.03816794
```

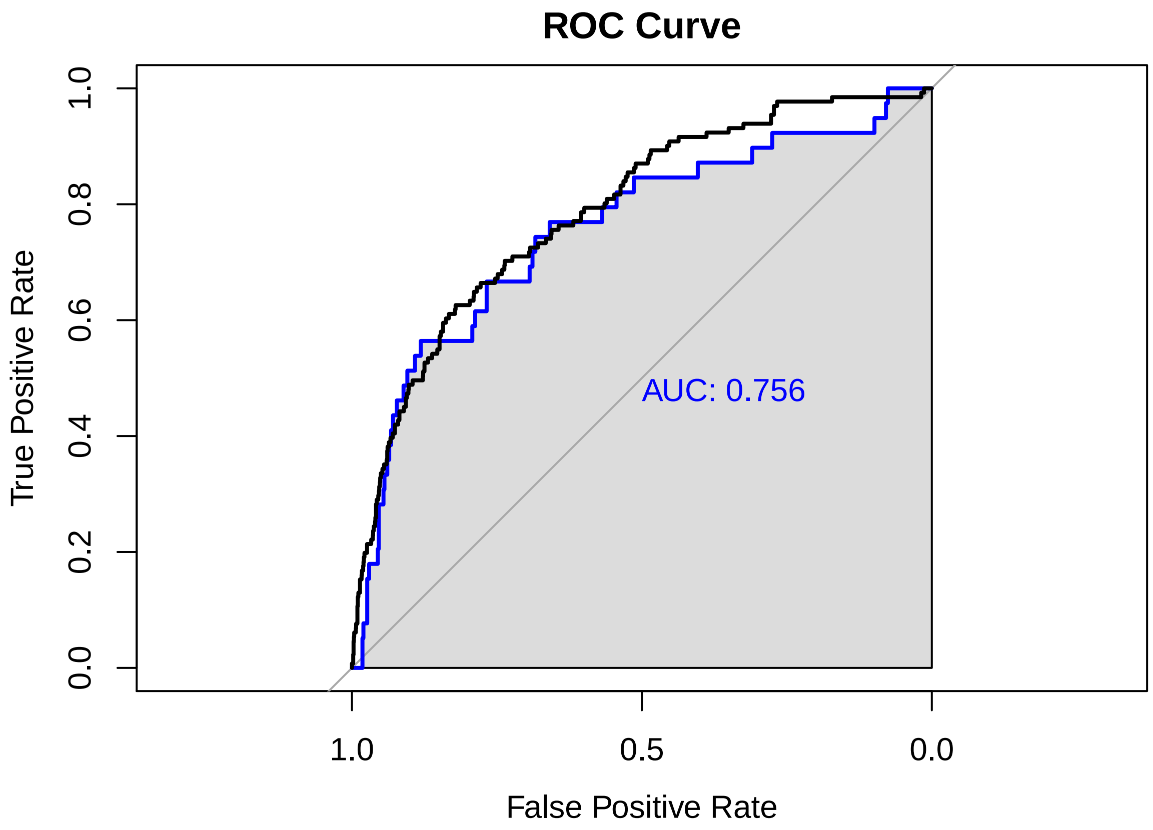
```
rda.train.specificity
```

```
## [1] 0.9878251
```

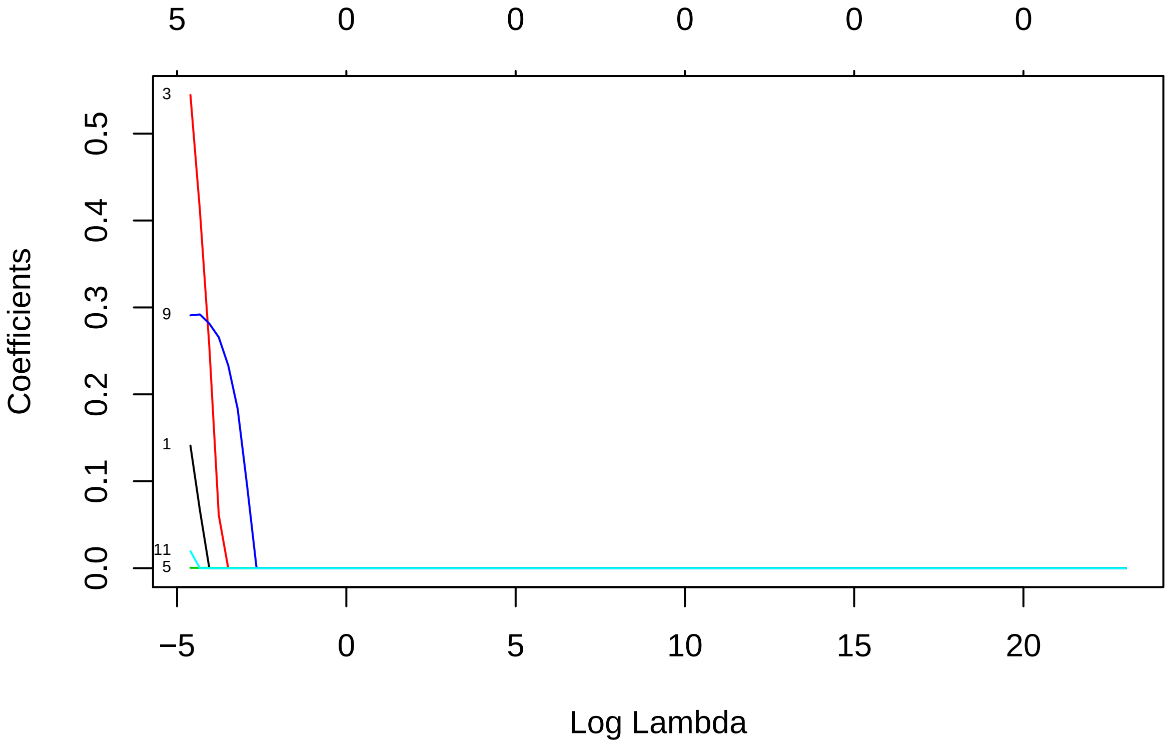
Logistic Regression.

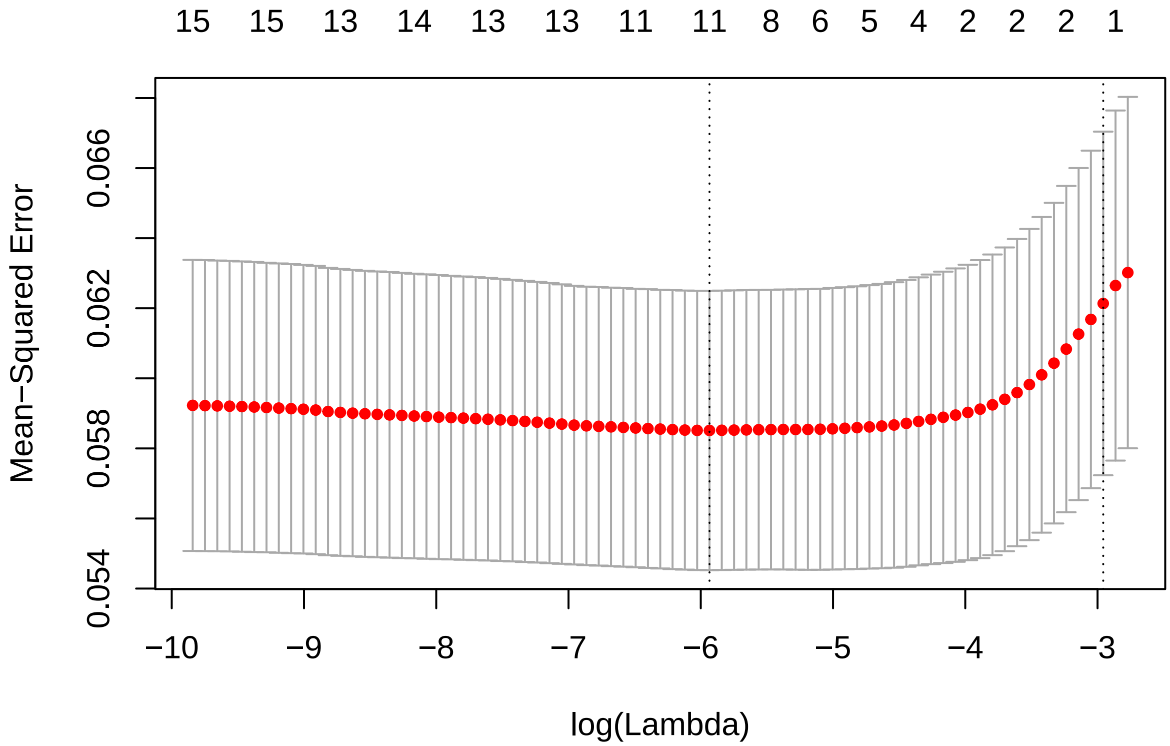
```
Call:
glm(formula = class ~ ., family = binomial, data = seismic.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8471  -0.3860  -0.2851  -0.1566   3.0825
```

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)    -6.343e+00  7.721e-01  -8.215  < 2e-16 ***
seismic         4.808e-01  2.111e-01   2.278 0.022727 *  
seismoacoustic  2.159e-01  1.993e-01   1.084 0.278524    
shift           1.179e+00  3.573e-01   3.301 0.000965 ***
genergy        -2.471e-07  5.044e-07  -0.490 0.624239    
gpuls           7.095e-04  2.474e-04   2.868 0.004136 ** 
gdenergy       -1.904e-04  2.177e-03  -0.087 0.930292    
gdpuls         -2.997e-03  3.093e-03  -0.969 0.332500    
ghazard        -2.335e-01  3.509e-01  -0.666 0.505671    
nbumps          1.807e+01  5.354e+02   0.034 0.973080    
nbumps2        -1.773e+01  5.354e+02  -0.033 0.973590    
nbumps3        -1.771e+01  5.354e+02  -0.033 0.973611    
nbumps4        -1.806e+01  5.354e+02  -0.034 0.973097    
nbumps5        -1.604e+01  5.354e+02  -0.030 0.976095    
energy          1.622e-06  4.033e-05   0.040 0.967929    
maxenergy      -7.101e-06  3.969e-05  -0.179 0.858012    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 958.82  on 1937  degrees of freedom
Residual deviance: 813.40  on 1922  degrees of freedom
AIC: 845.4

Number of Fisher Scoring iterations: 12

[1] 0.9329205



glm.pred    0    1
       0 1802  125
       1    5    6

[1] 0.04580153

[1] 0.997233

[1] 0.9349845



glm.pred   0   1
       0 604  39
       1   3   0

[1] 0

[1] 0.9950577
```



Variable Selection-LASSO





```
[1] 8.670049
```

```
  (Intercept)        seismic seismoacoustic          shift        genergy
-8.144581e-03   8.800484e-03   0.000000e+00   7.977504e-03   0.000000e+00
        gpuls       gdenergy         gdpuls        ghazard         nbumps
 4.677101e-05   0.000000e+00   0.000000e+00   0.000000e+00   3.117955e-02
      nbumps2        nbumps3        nbumps4        nbumps5         energy
 0.000000e+00   0.000000e+00   0.000000e+00   0.000000e+00   0.000000e+00
    maxenergy
 0.000000e+00
```

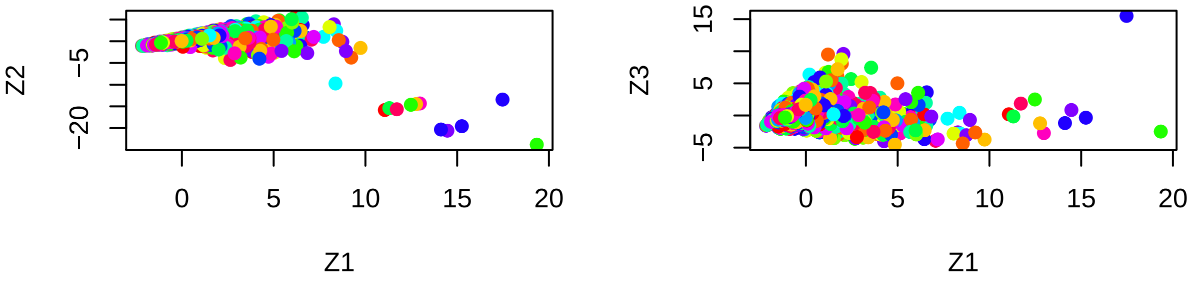
Variables selected through LASSO

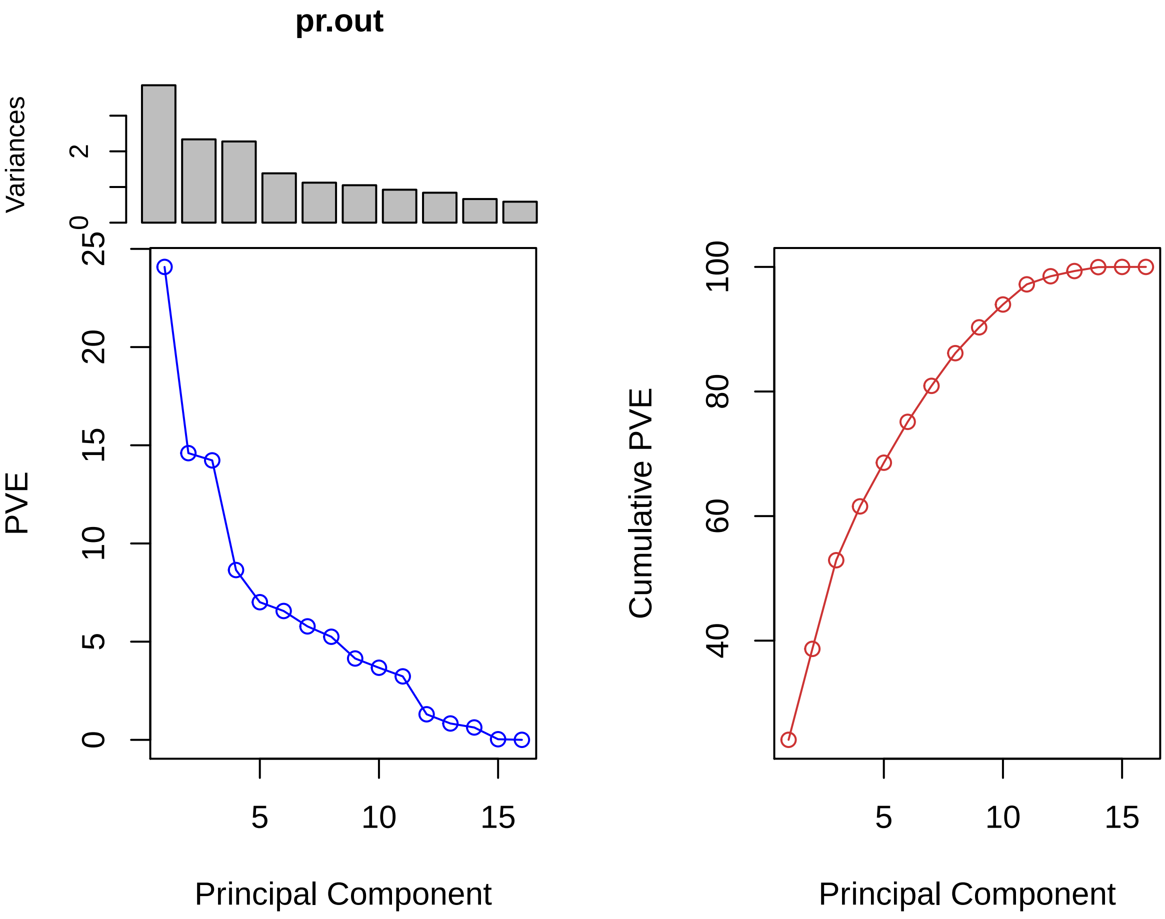
```
  (Intercept)       seismic         shift         gpuls        nbumps
-8.144581e-03  8.800484e-03  7.977504e-03  4.677101e-05  3.117955e-02
```

Principal Component Analysis from the Book - INCOMPLETE

```
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6
Standard deviation     1.9629 1.5284 1.5089 1.17618 1.05902 1.02457
Proportion of Variance 0.2408 0.1460 0.1423 0.08646 0.07009 0.06561
Cumulative Proportion  0.2408 0.3868 0.5291 0.61559 0.68568 0.75129
                           PC7    PC8     PC9    PC10    PC11    PC12
Standard deviation     0.96145 0.9165 0.81413 0.76650 0.71908 0.45631
Proportion of Variance 0.05777 0.0525 0.04143 0.03672 0.03232 0.01301
```

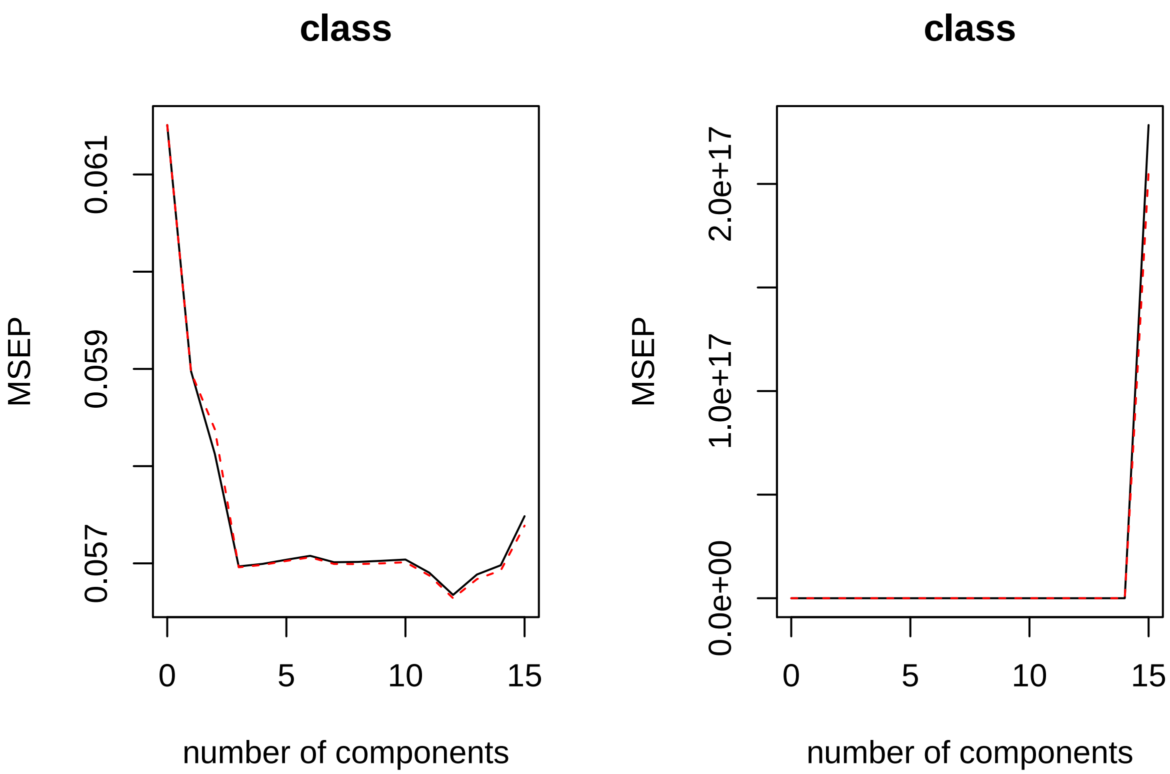
```
Cumulative Proportion  0.80907 0.8616 0.90299 0.93971 0.97203 0.98504
                          PC13    PC14     PC15     PC16
Standard deviation     0.36522 0.3174 0.07039 0.01562
Proportion of Variance 0.00834 0.0063 0.00031 0.00002
Cumulative Proportion  0.99338 0.9997 0.99998 1.00000
```





```
Data:   X dimension: 2584 15
    Y dimension: 2584 1
Fit method: svdpc
Number of components considered: 15

VALIDATION: RMSEP
Cross-validated using 10 random segments.
```

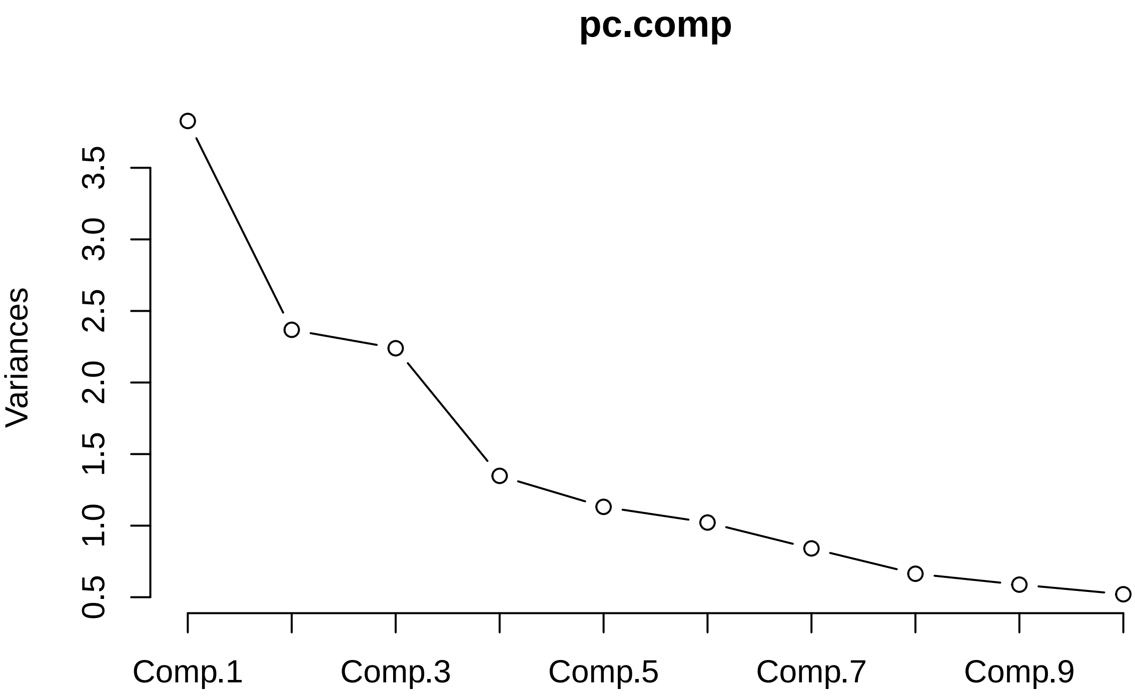
```
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV           0.248   0.2428   0.2411   0.2387   0.2387   0.2388   0.2389
adjCV        0.248   0.2428   0.2416   0.2387   0.2387   0.2388   0.2389
       7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
CV      0.2388   0.2388   0.2388    0.2388    0.2385    0.2381    0.2385
adjCV   0.2387   0.2387   0.2387    0.2388    0.2385    0.2380    0.2384
       14 comps  15 comps
CV       0.2387    0.2398
adjCV    0.2386    0.2396

TRAINING: % variance explained
       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
X       25.306   40.680   55.704   64.926   72.401   79.396   85.185
class    4.225    5.285    7.573    7.577    7.584    7.592    7.792
       8 comps  9 comps  10 comps  11 comps  12 comps  13 comps  14 comps
X       89.627   93.557    97.005    98.398    99.294     99.97    99.998
class    7.917    8.022     8.026     8.289     8.847      8.87     8.872
       15 comps
X       100.000
class     9.128
```



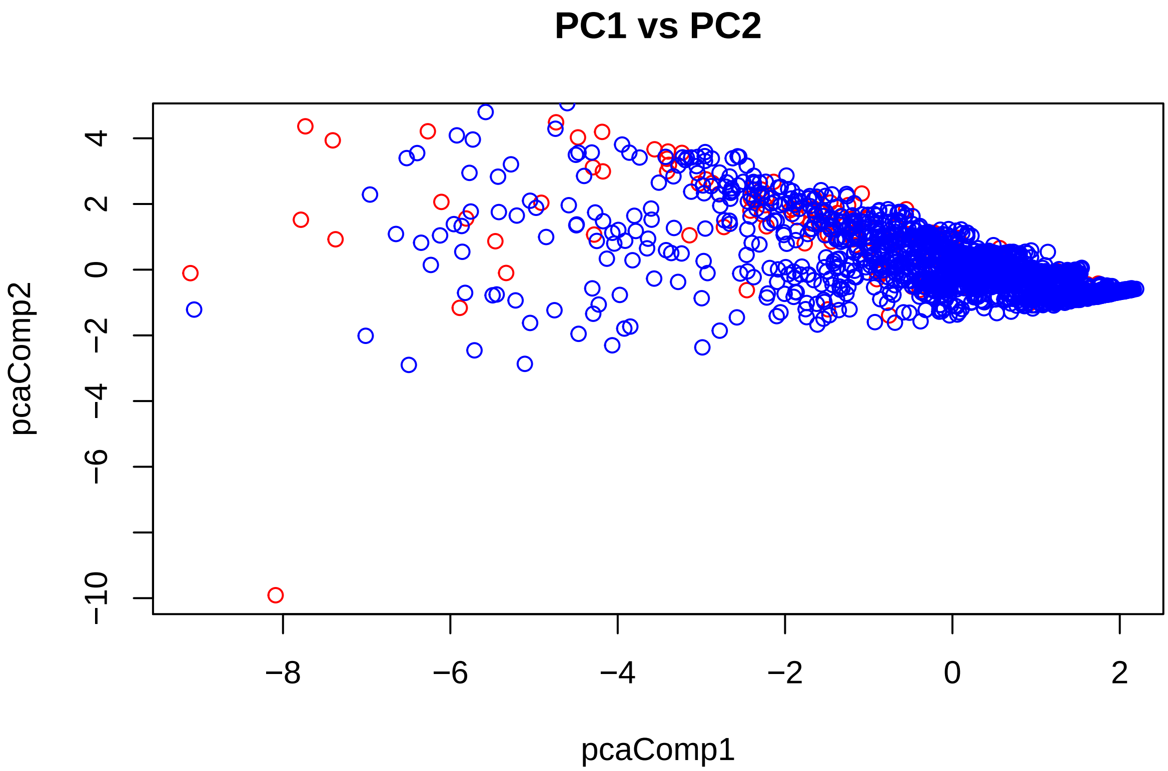
```
[1] 0.05357258

Data:    X dimension: 1938 15
    Y dimension: 1938 1
Fit method: svdpc
Number of components considered: 15

VALIDATION: RMSEP
```

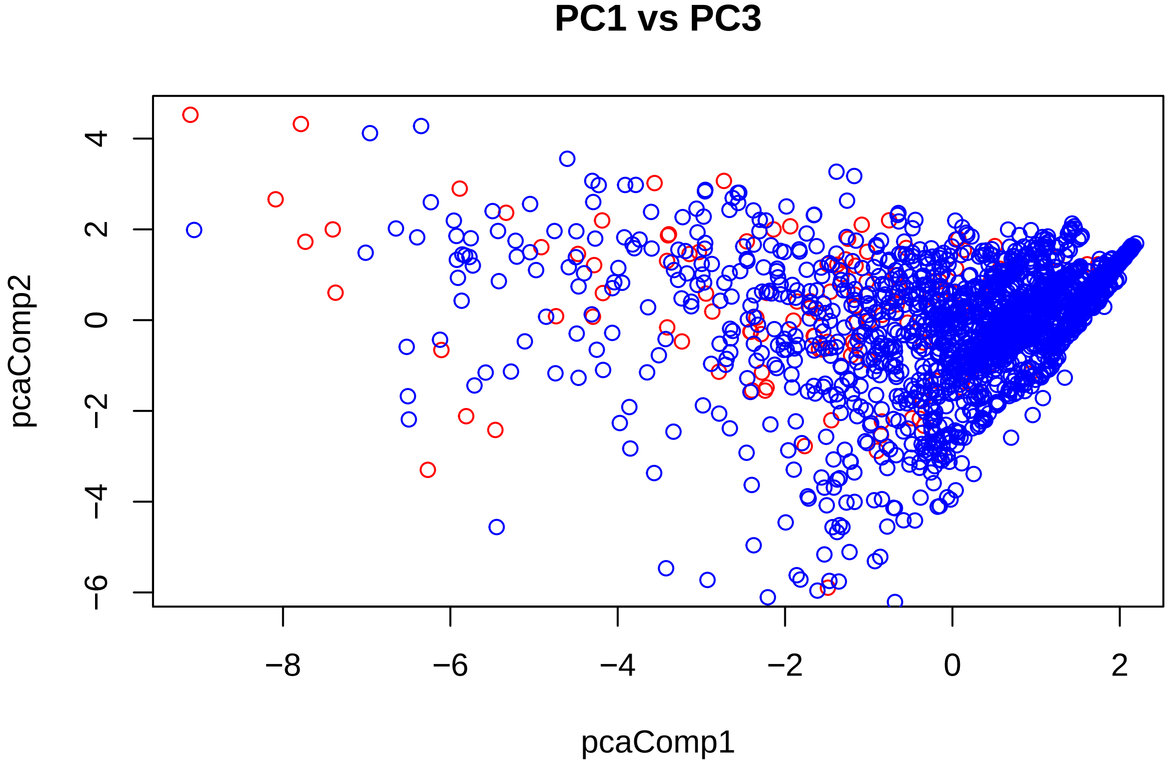
```
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV          0.2512   0.2454   0.2417   0.2413   0.2417   0.2417    0.242
adjCV       0.2512   0.2453   0.2415   0.2413   0.2416   0.2417    0.242
       7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
CV      0.2421   0.2421   0.2421    0.2422    0.2422    0.2420    0.2422
adjCV   0.2420   0.2420   0.2420    0.2421    0.2421    0.2419    0.2421
       14 comps   15 comps
CV       0.2431  477953634
adjCV    0.2429  453530602

TRAINING: % variance explained
       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
X       25.528   41.325   56.263   65.253   72.800   79.615   85.224
class    4.815    7.847    7.952    7.953    7.967    8.025    8.255
       8 comps  9 comps  10 comps  11 comps  12 comps  13 comps  14 comps
X       89.653   93.574    97.048    98.456    99.324    99.978    99.999
class    8.362    8.466     8.502     8.574     8.905     8.943     8.981
       15 comps
X       100.000
class     9.781
```

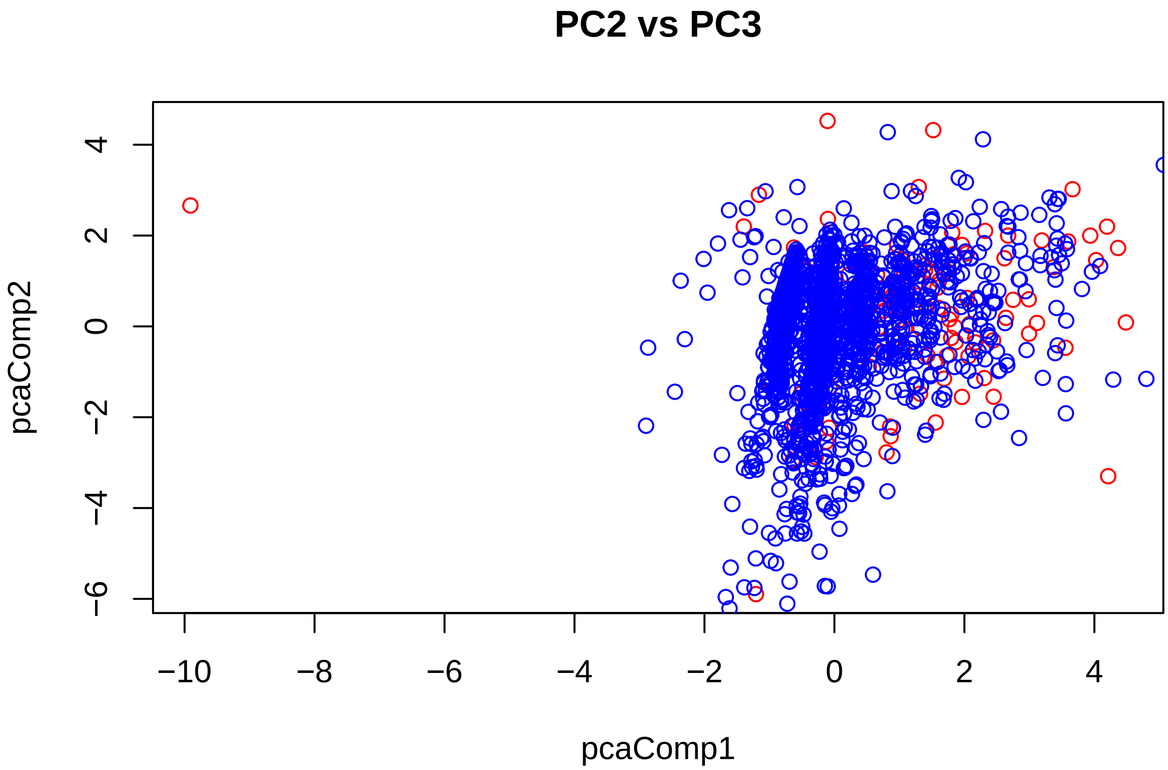
Variable Selection - PCA - INCOMPLETE



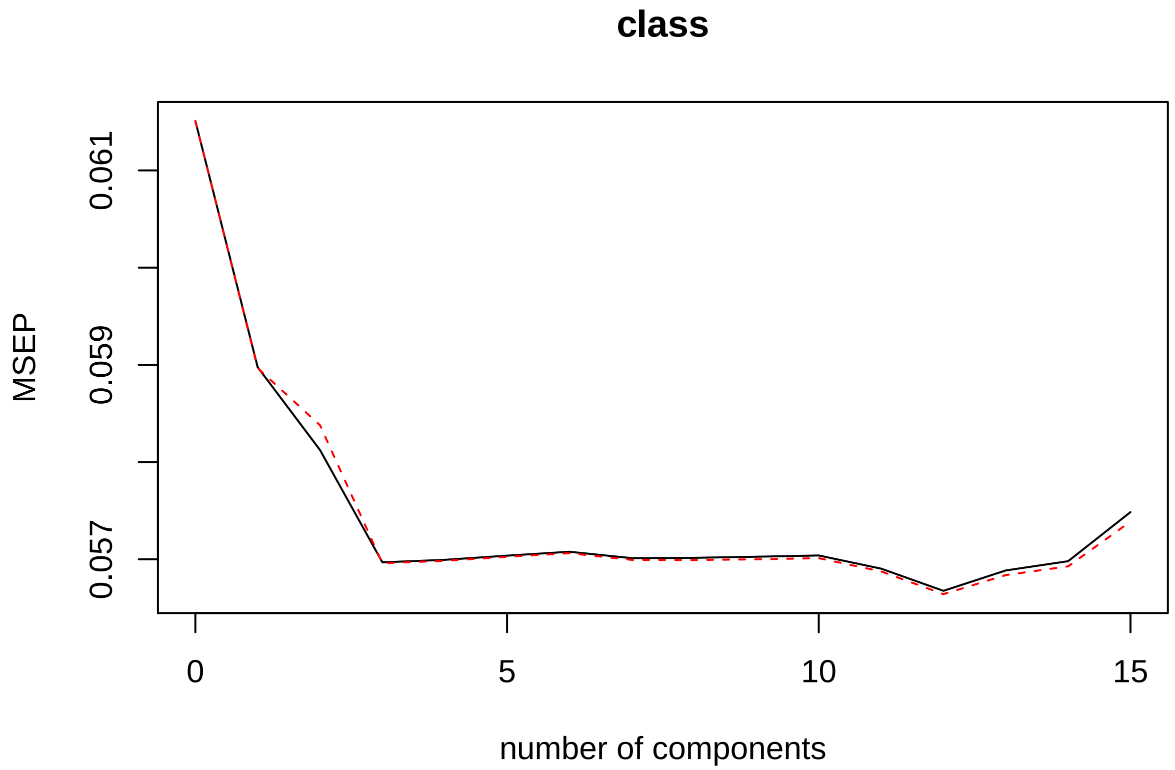
Variable Selection - PCA - INCOMPLETE

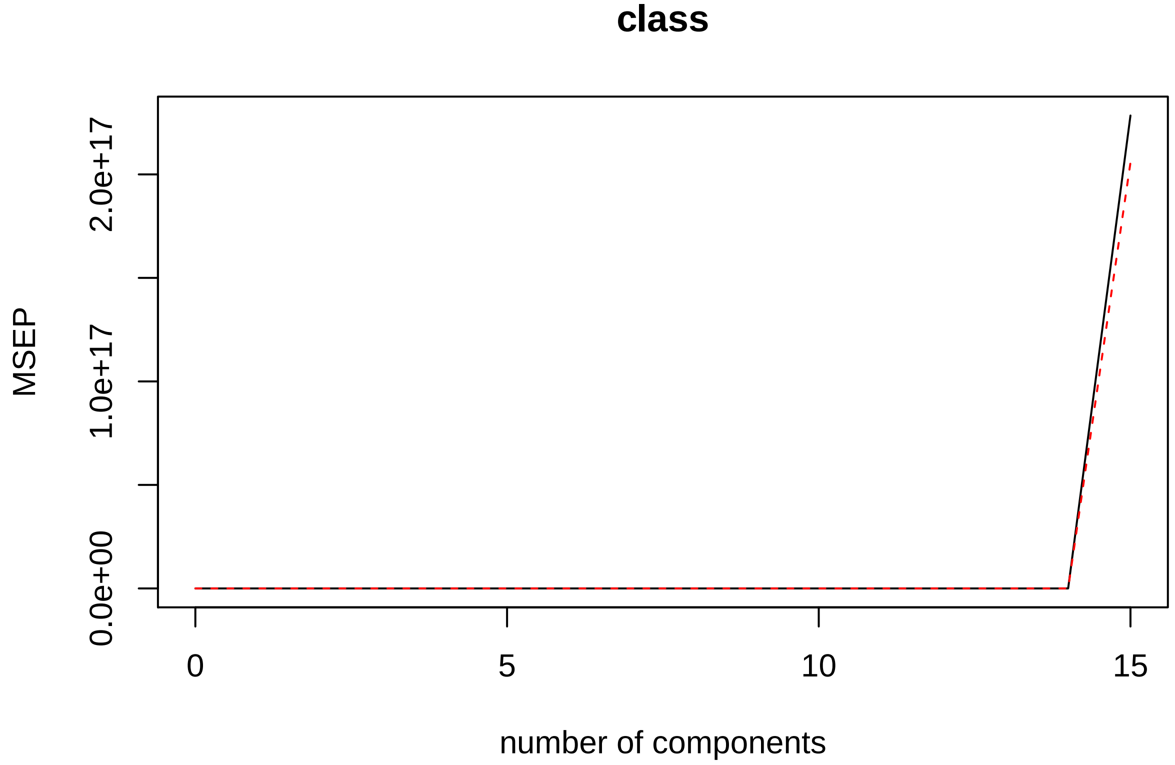

PC1 vs PC2


PC1 vs PC3



```
Data:   X dimension: 2584 15
    Y dimension: 2584 1
Fit method: svdpc
Number of components considered: 15

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV           0.248   0.2428   0.2411   0.2387   0.2387   0.2388   0.2389
adjCV        0.248   0.2428   0.2416   0.2387   0.2387   0.2388   0.2389
       7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
CV      0.2388   0.2388   0.2388    0.2388    0.2385    0.2381    0.2385
adjCV   0.2387   0.2387   0.2387    0.2388    0.2385    0.2380    0.2384
       14 comps  15 comps
CV       0.2387    0.2398
adjCV    0.2386    0.2396

TRAINING: % variance explained
       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
X       25.306   40.680   55.704   64.926   72.401   79.396   85.185
class    4.225    5.285    7.573    7.577    7.584    7.592    7.792
       8 comps  9 comps  10 comps  11 comps  12 comps  13 comps  14 comps
X       89.627   93.557    97.005    98.398    99.294     99.97    99.998
class    7.917    8.022     8.026     8.289     8.847      8.87     8.872
       15 comps
X       100.000
class     9.128
```

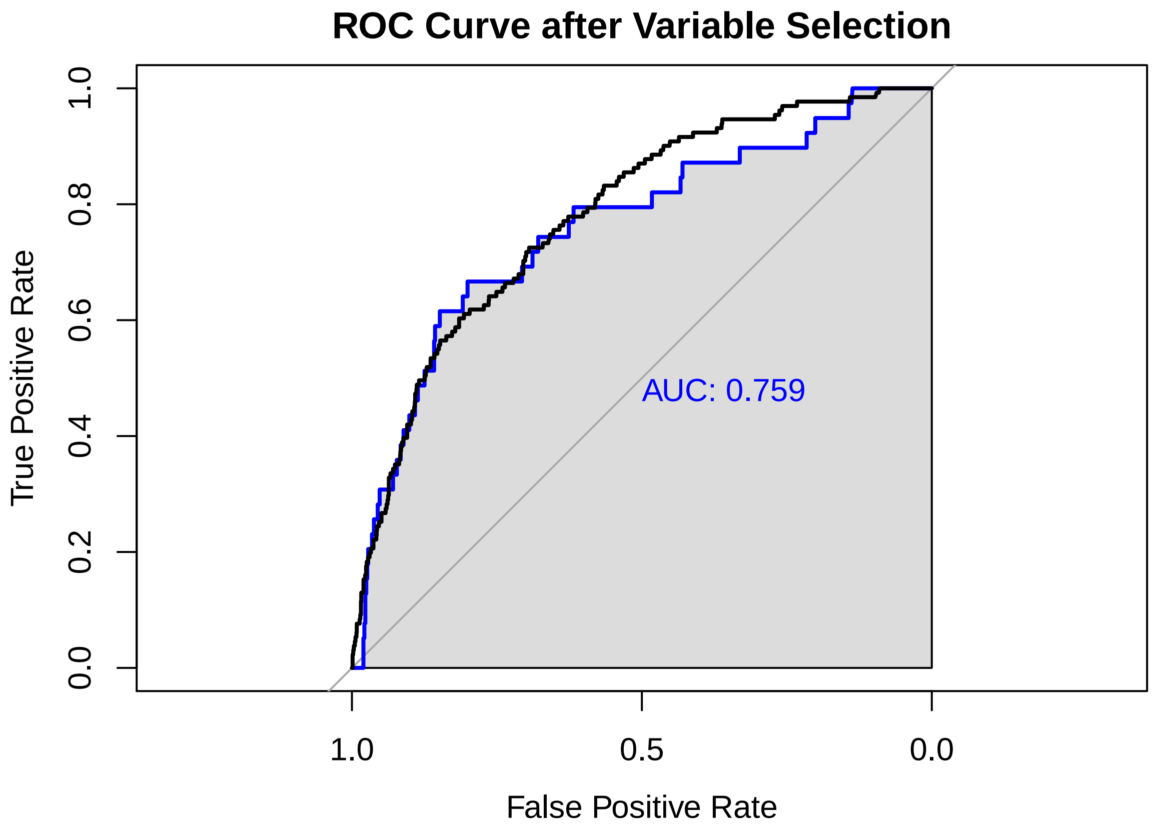

class
MSEP
0.061
0.059
0.057
0
5
10
15
number of components


class
MSEP
2.0e+17
1.0e+17
0.0e+00
0
5
10
15
number of components

```
[1] 0.05357258
```

Logistic Regression after Variable Selection

```
Call:
glm(formula = class ~ seismic + shift + gpuls + nbumps, family = binomial,
    data = seismic.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6270  -0.3846  -0.2947  -0.1627   2.9781

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.9508244  0.6490468  -9.169  < 2e-16 ***
seismic      0.3641160  0.1944250   1.873 0.061098 .
shift        1.1371057  0.3402674   3.342 0.000832 ***
gpuls        0.0004913  0.0001283   3.829 0.000129 ***
nbumps       0.3231048  0.0507286   6.369  1.9e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 958.82  on 1937  degrees of freedom
Residual deviance: 828.98  on 1933  degrees of freedom
AIC: 838.98

Number of Fisher Scoring iterations: 6

[1] 0.9318885


glm.pred    0    1
       0 1803  128
       1    4    3

[1] 0.02290076

[1] 0.9977864

[1] 0.9380805


glm.pred   0   1
       0 606  39
       1   1   0

[1] 0

[1] 0.9983526
```



Quadratic Discriminant Analysis after variable selection

```
##------------------------------------------
## Fit QDA model after variable selection
##------------------------------------------

# Model 1
qda.fit <- qda(class~seismic+shift+gpuls+nbumps, data=seismic.train)
qda.class=predict(qda.fit,seismic.test)$class
confusion <- table(qda.class ,seismic.test$class)

sensitivity <- confusion[2,2]/sum(confusion[,2])
specificity <- confusion[1,1]/sum(confusion[,1])

confusion
```

```
##
## qda.class   0   1
##         0 565  27
##         1  42  12
```

```
sensitivity
```

```
## [1] 0.3076923
```

```
specificity
```

```
## [1] 0.9308072
```

```
# Model 2
qda.fit <- qda(class ~ genergy + gpuls + nbumps + nbumps2 + nbumps4, data=seismic.train)
qda.class=predict(qda.fit,seismic.test)$class

confusion <- table(qda.class ,seismic.test$class)

sensitivity <- confusion[2,2]/sum(confusion[,2])
specificity <- confusion[1,1]/sum(confusion[,1])

confusion
```

```
##
## qda.class   0   1
##         0 527  23
##         1  80  16
```

```
sensitivity
```

```
## [1] 0.4102564
```

```
specificity
```

```
## [1] 0.8682043
```

Regularized Discriminant Analysis after variable selection

```
rda.class   0   1
        0 595  35
        1  12   4

[1] 0.1025641

[1] 0.9802306
```

```
rda.class   0   1
        0 572  33
        1  35   6

[1] 0.1538462

[1] 0.9423394
```

Pre-Variable Selection

Model	Test Specificity	Test Sensitivity	Training Specificity	Training Sensitivity
Indicator	123	123	123	123
LDA	123	123	123	123
QDA	123	123	123	123
RDA	123	123	123	123
Log Regression	123	123	123	123

Post-Variable Selection

Model	Test Specificity	Test Sensitivity	Training Specificity	Training Sensitivity
Indicator	123	123	123	123
LDA	123	123	123	123
QDA	123	123	123	123
RDA	123	123	123	123
Log Regression	123	123	123	123