557_Project_2BS

Ben Straub, Hillary Koch, Jiawei Huang, Arif Masrur

3/15/2017

EDA in no particular order of sanity

Names of Variables

```
 [1] "seismic"        "seismoacoustic" "shift"          "genergy"       
 [5] "gpuls"          "gdenergy"       "gdpuls"         "ghazard"       
 [9] "nbumps"         "nbumps2"        "nbumps3"        "nbumps4"       
[13] "nbumps5"        "nbumps6"        "nbumps7"        "nbumps89"      
[17] "energy"         "maxenergy"      "class"         
```

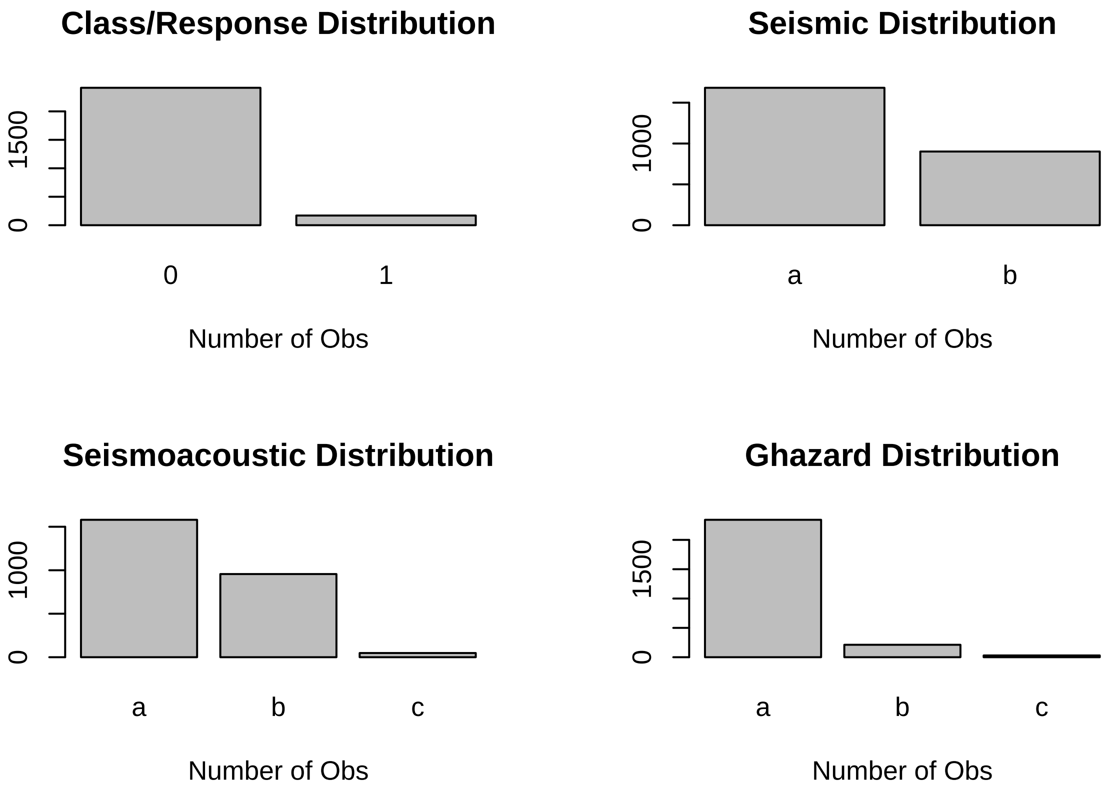
Summary Statistics

```
 seismic  seismoacoustic shift       genergy            gpuls       
 a:1682   a:1580         N: 921   Min.   :    100   Min.   :   2.0  
 b: 902   b: 956         W:1663   1st Qu.:  11660   1st Qu.: 190.0  
          c:  48                  Median :  25485   Median : 379.0  
                                  Mean   :  90242   Mean   : 538.6  
                                  3rd Qu.:  52832   3rd Qu.: 669.0  
                                  Max.   :2595650   Max.   :4518.0  
    gdenergy          gdpuls        ghazard      nbumps      
 Min.   : -96.00   Min.   :-96.000   a:2342   Min.   :0.0000  
 1st Qu.: -37.00   1st Qu.:-36.000   b: 212   1st Qu.:0.0000  
 Median :  -6.00   Median : -6.000   c:  30   Median :0.0000  
 Mean   :  12.38   Mean   :  4.509            Mean   :0.8595  
 3rd Qu.:  38.00   3rd Qu.: 30.250            3rd Qu.:1.0000  
 Max.   :1245.00   Max.   :838.000            Max.   :9.0000  
    nbumps2          nbumps3          nbumps4           nbumps5        
 Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.000000  
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.000000  
 Median :0.0000   Median :0.0000   Median :0.00000   Median :0.000000  
 Mean   :0.3936   Mean   :0.3928   Mean   :0.06772   Mean   :0.004644  
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.000000  
 Max.   :8.0000   Max.   :7.0000   Max.   :3.00000   Max.   :1.000000  
    nbumps6     nbumps7     nbumps89     energy         maxenergy     
 Min.   :0   Min.   :0   Min.   :0   Min.   :     0   Min.   :     0  
 1st Qu.:0   1st Qu.:0   1st Qu.:0   1st Qu.:     0   1st Qu.:     0  
 Median :0   Median :0   Median :0   Median :     0   Median :     0  
 Mean   :0   Mean   :0   Mean   :0   Mean   :  4975   Mean   :  4279  
 3rd Qu.:0   3rd Qu.:0   3rd Qu.:0   3rd Qu.:  2600   3rd Qu.:  2000  
 Max.   :0   Max.   :0   Max.   :0   Max.   :402000   Max.   :400000  
     class        
 Min.   :0.00000  
 1st Qu.:0.00000  
 Median :0.00000  
 Mean   :0.06579  
 3rd Qu.:0.00000
 Max.   :1.00000
```

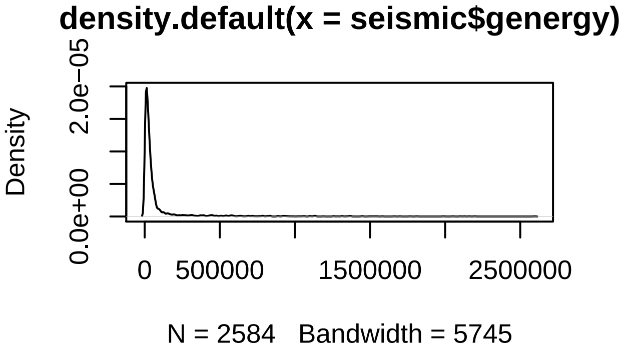
Dimensions of Data Matrix

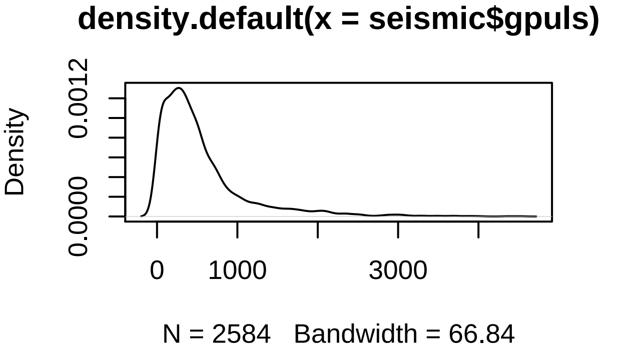
```
[1] 2584   19
```

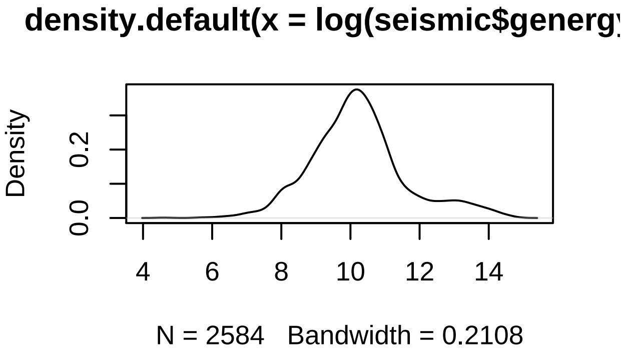
What the Factor Variables look like

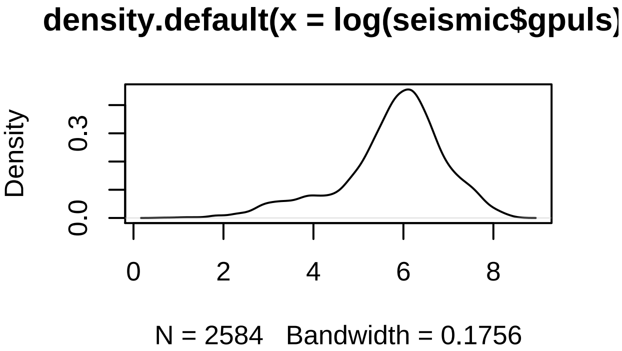


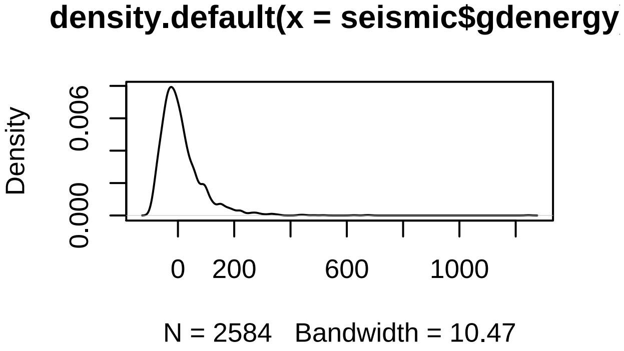
What the Continuous Variables look like

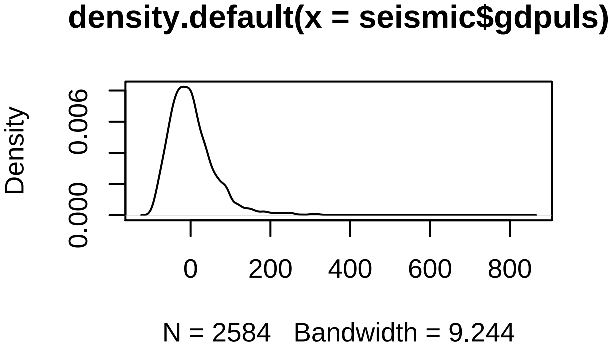


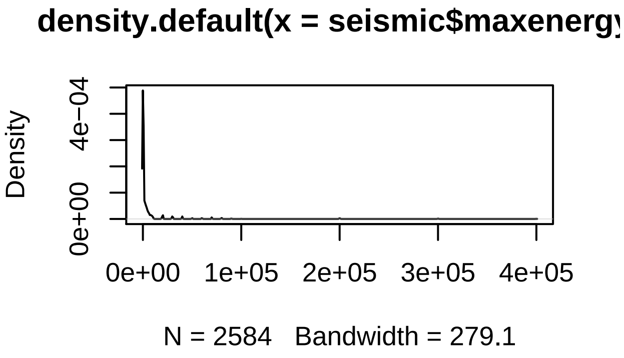


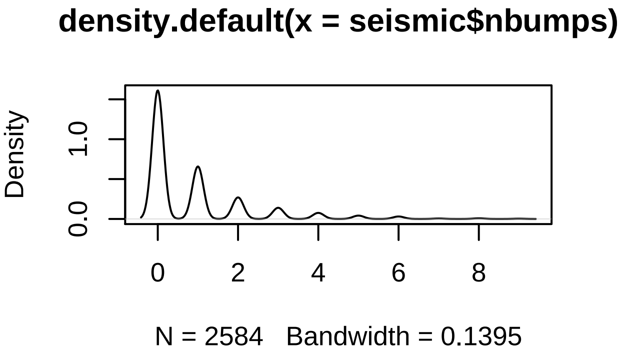


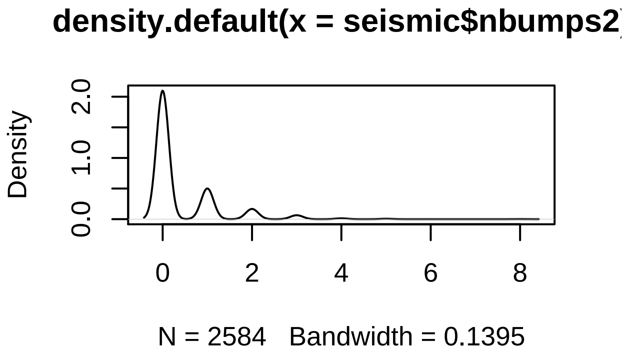


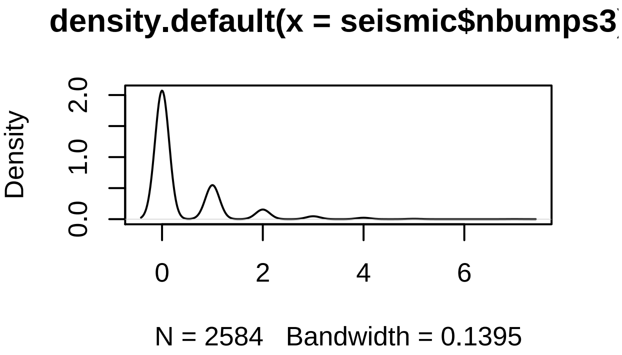


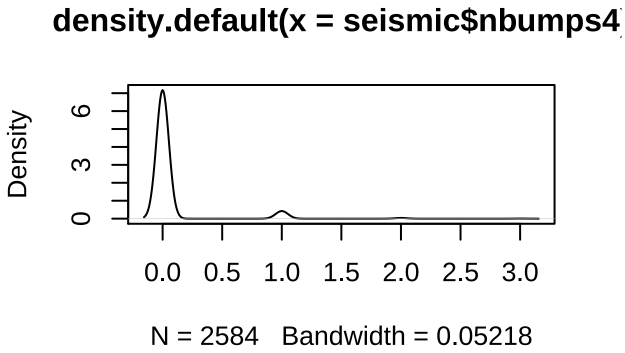


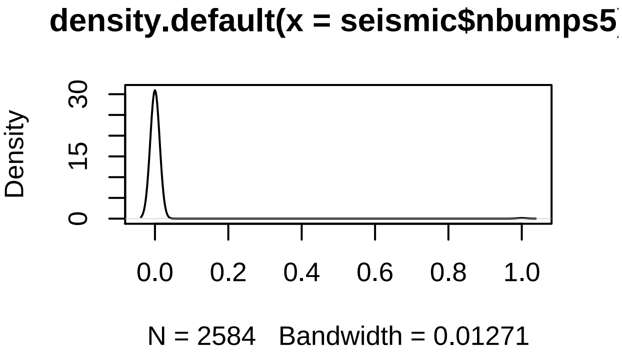






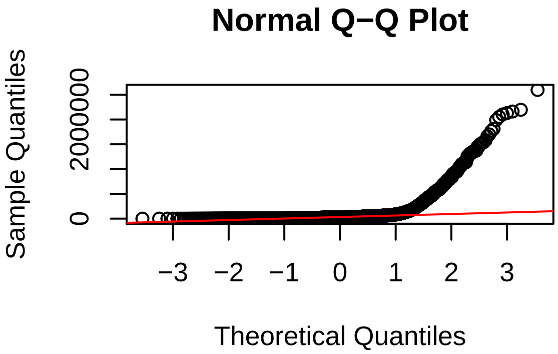


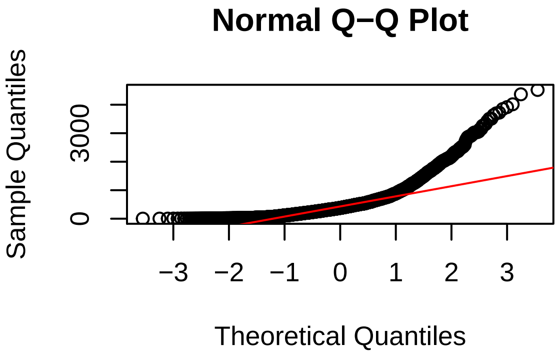


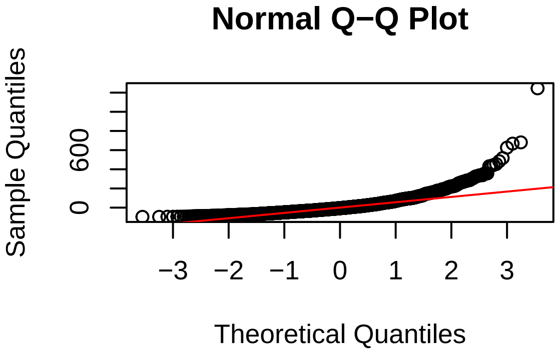


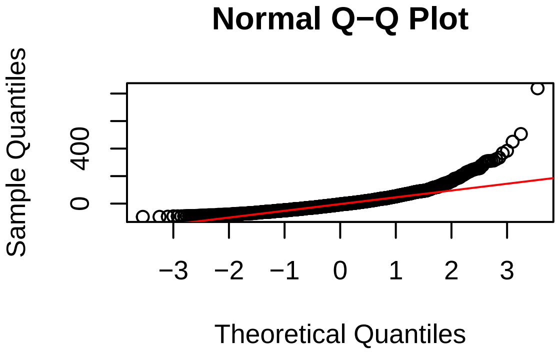
```
Call:
lm(formula = class ~ ., data = seismic)

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
```

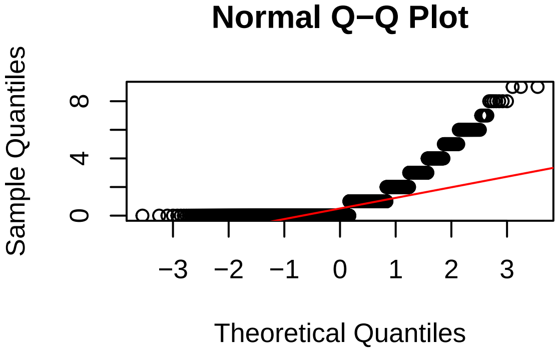
```
Residual standard error: 0.2371 on 2568 degrees of freedom
Multiple R-squared:  0.09128,   Adjusted R-squared:  0.08597
F-statistic:  17.2 on 15 and 2568 DF,  p-value: < 2.2e-16
```

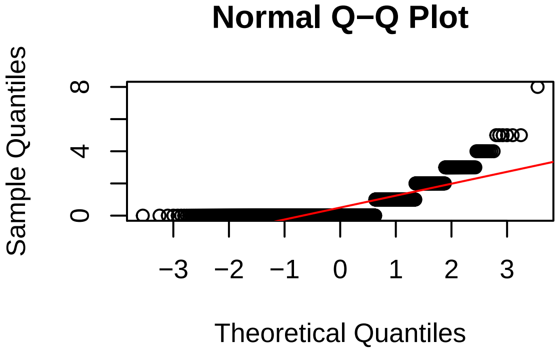


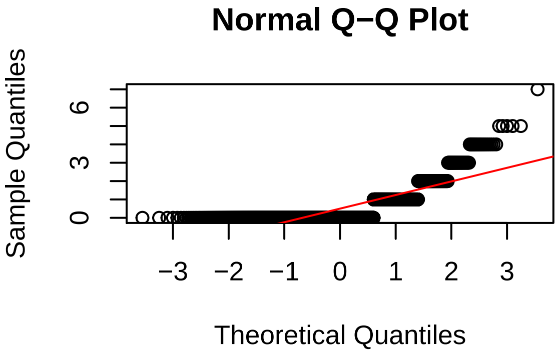


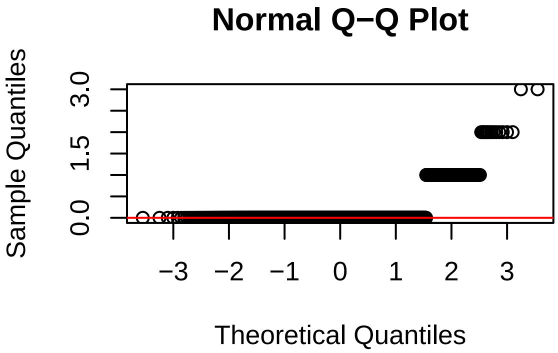


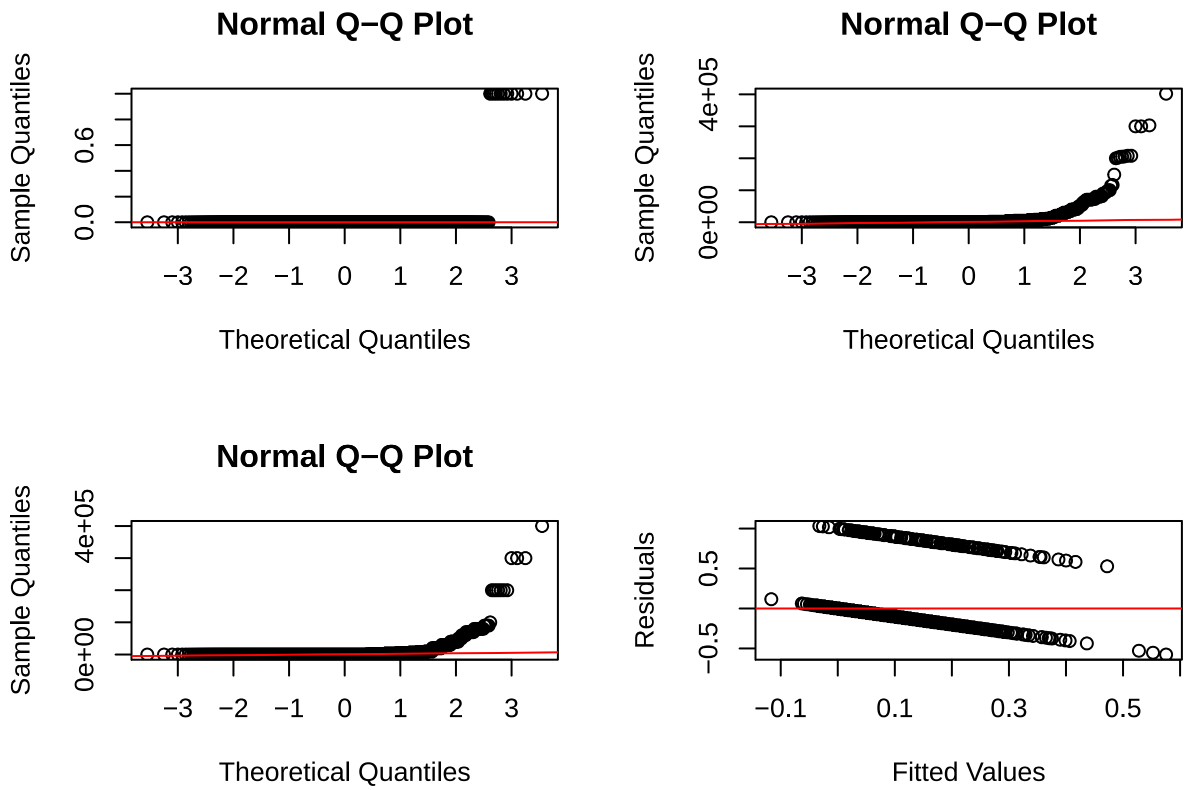


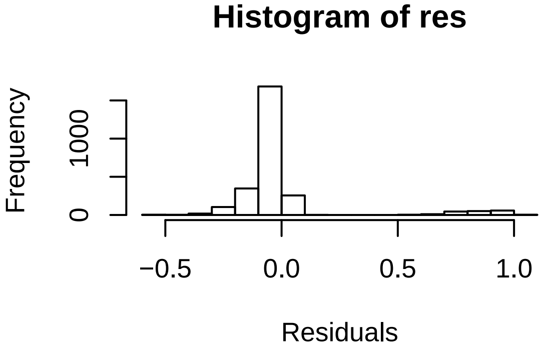










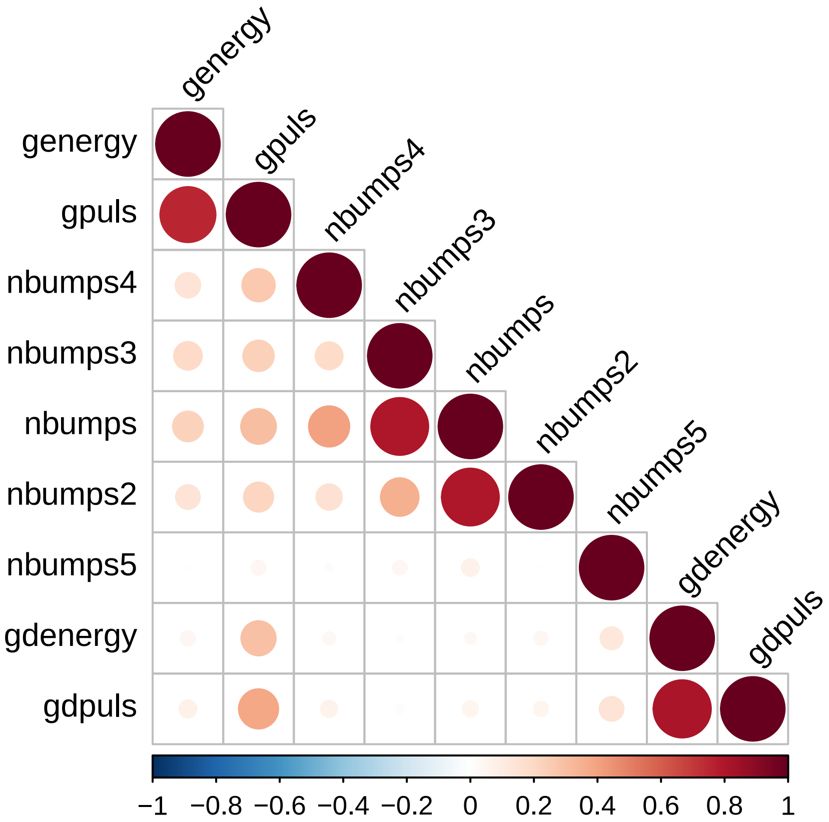


Lots of multicollinearity to worry about during variable selection

```
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



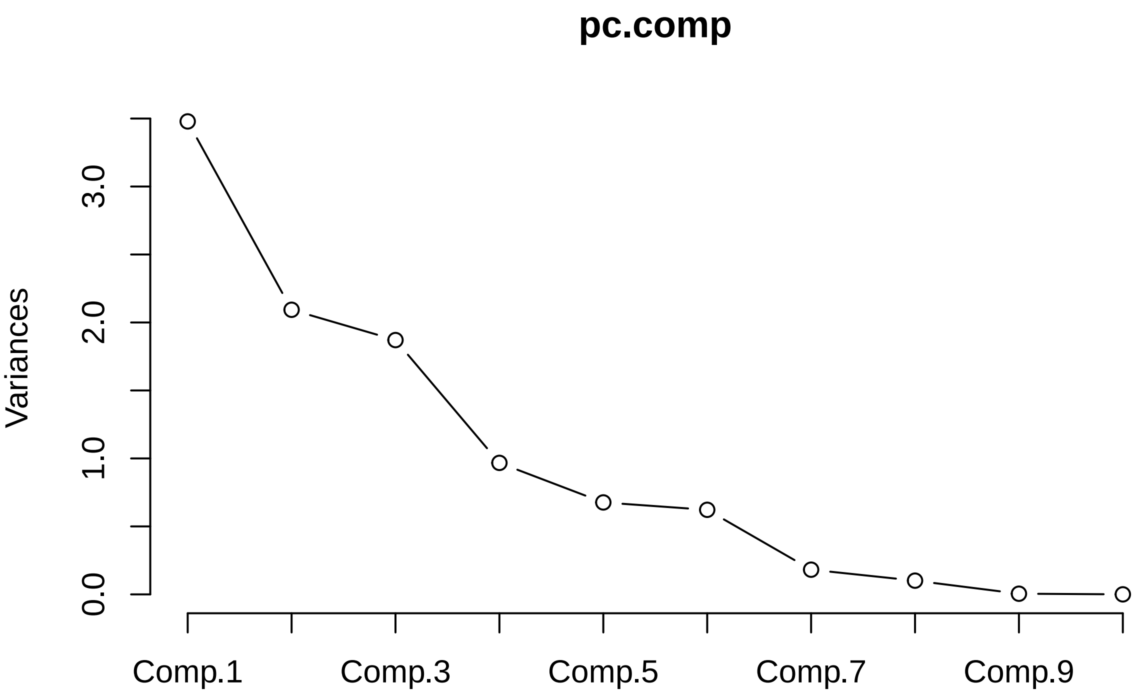
```
$r
        row   column     cor       p
1   genergy    gpuls  0.7500 0.0e+00
2   genergy  nbumps4  0.1500 1.4e-14
3     gpuls  nbumps4  0.2600 0.0e+00
4   genergy  nbumps3  0.1900 0.0e+00
5     gpuls  nbumps3  0.2300 0.0e+00
6   nbumps4  nbumps3  0.1800 0.0e+00
7   genergy   nbumps  0.2200 0.0e+00
8     gpuls   nbumps  0.3000 0.0e+00
9   nbumps4   nbumps  0.4000 0.0e+00
10  nbumps3   nbumps  0.8000 0.0e+00
11  genergy  nbumps2  0.1400 2.2e-13
12    gpuls  nbumps2  0.2100 0.0e+00
13  nbumps4  nbumps2  0.1600 0.0e+00
14  nbumps3  nbumps2  0.3500 0.0e+00
15   nbumps  nbumps2  0.8000 0.0e+00
16  genergy  nbumps5 -0.0099 6.2e-01
17    gpuls  nbumps5  0.0490 1.2e-02
18  nbumps4  nbumps5 -0.0170 4.0e-01
19  nbumps3  nbumps5  0.0460 1.8e-02
20   nbumps  nbumps5  0.0700 4.0e-04
21  nbumps2  nbumps5 -0.0053 7.9e-01
22  genergy gdenergy  0.0490 1.4e-02
23    gpuls gdenergy  0.2900 0.0e+00
```

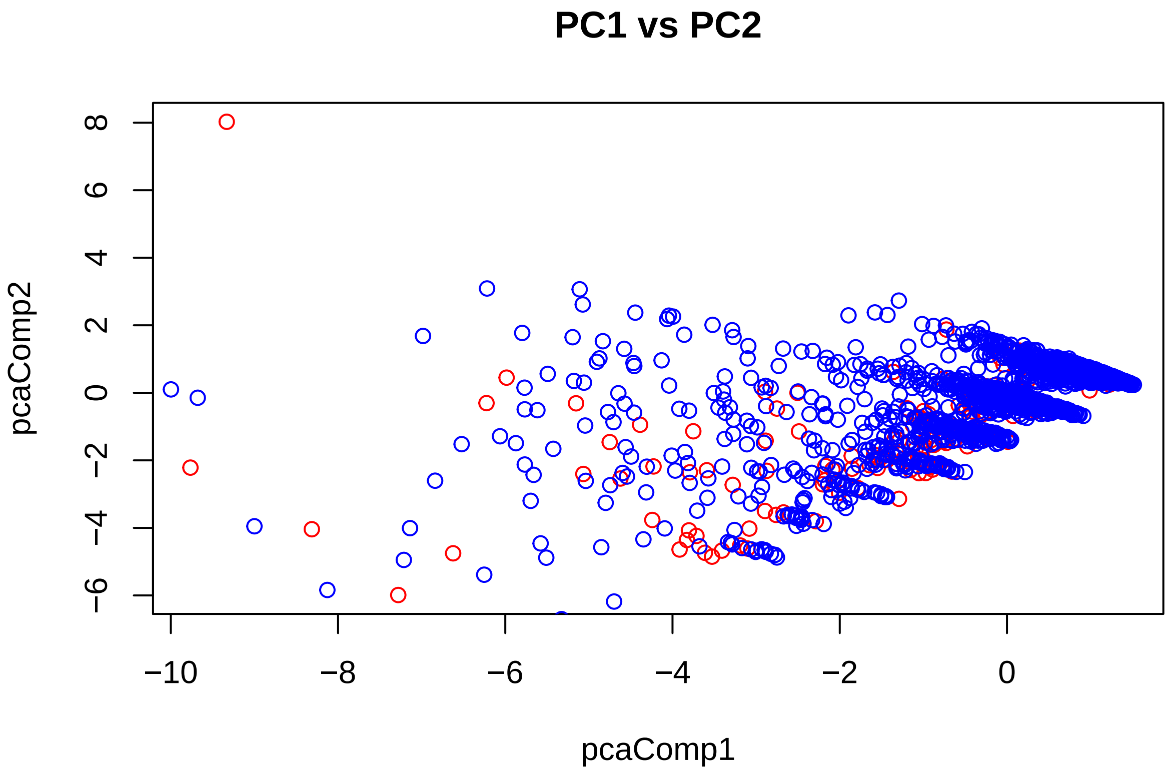
```
24  nbumps4 gdenergy  0.0370 6.1e-02
25  nbumps3 gdenergy -0.0120 5.4e-01
26   nbumps gdenergy  0.0300 1.3e-01
27  nbumps2 gdenergy  0.0410 3.6e-02
28  nbumps5 gdenergy  0.1200 3.3e-10
29  genergy   gdpuls  0.0720 2.7e-04
30    gpuls   gdpuls  0.3800 0.0e+00
31  nbumps4   gdpuls  0.0660 7.6e-04
32  nbumps3   gdpuls  0.0150 4.5e-01
33   nbumps   gdpuls  0.0580 3.2e-03
34  nbumps2   gdpuls  0.0510 9.4e-03
35  nbumps5   gdpuls  0.1400 5.9e-13
36 gdenergy   gdpuls  0.8100 0.0e+00

$p
NULL

$sym
NULL
```

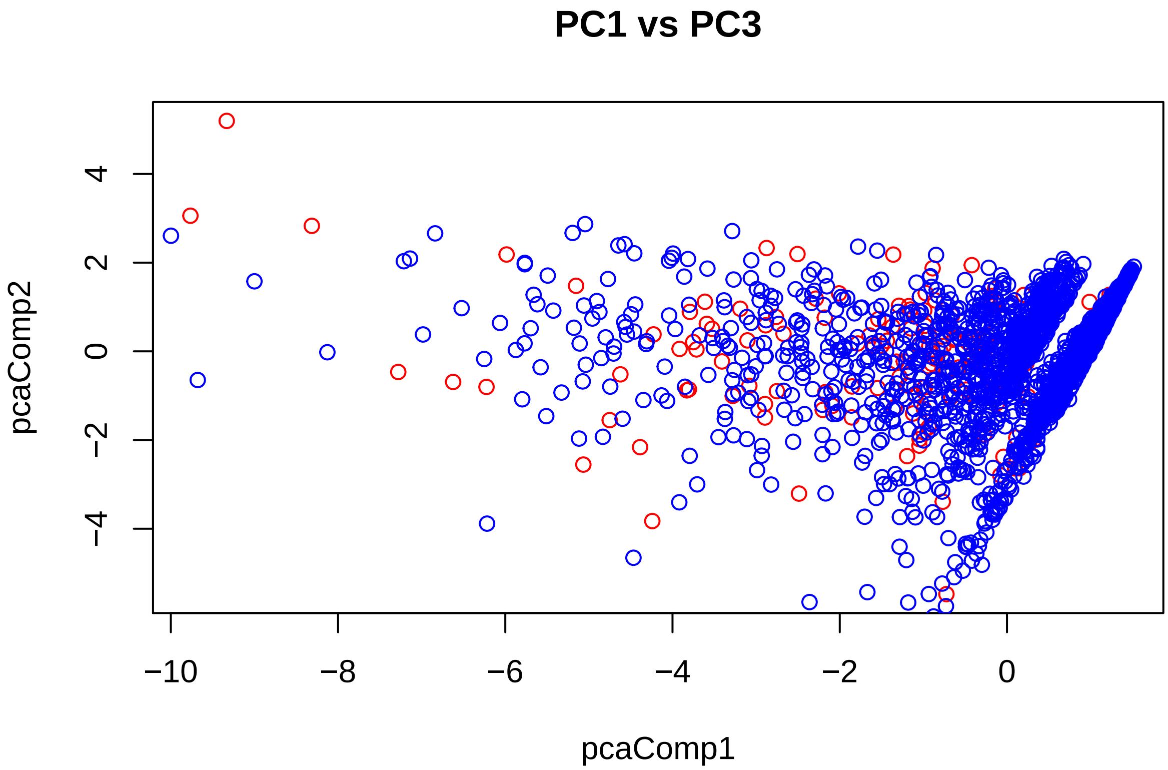
Separating into Test and Training Sets

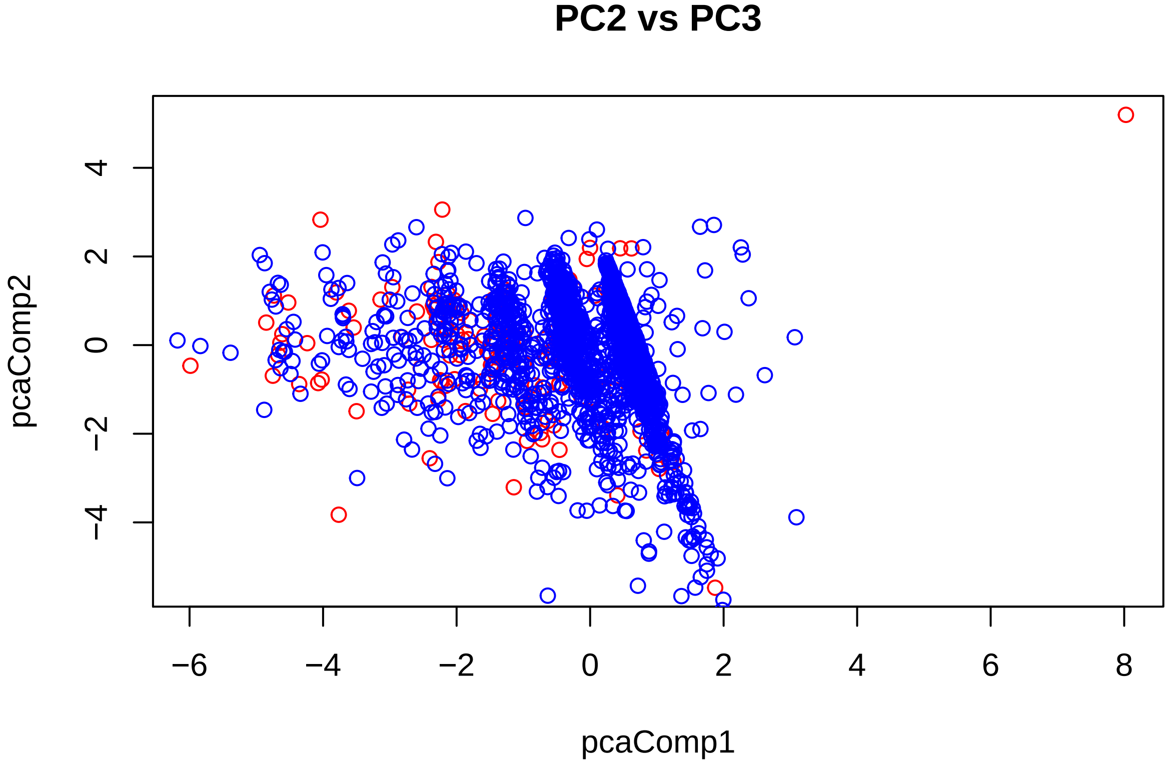
```r
##------------------------------------
## Setting up
##------------------------------------

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,]

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

Linear regression of an indicator matrix





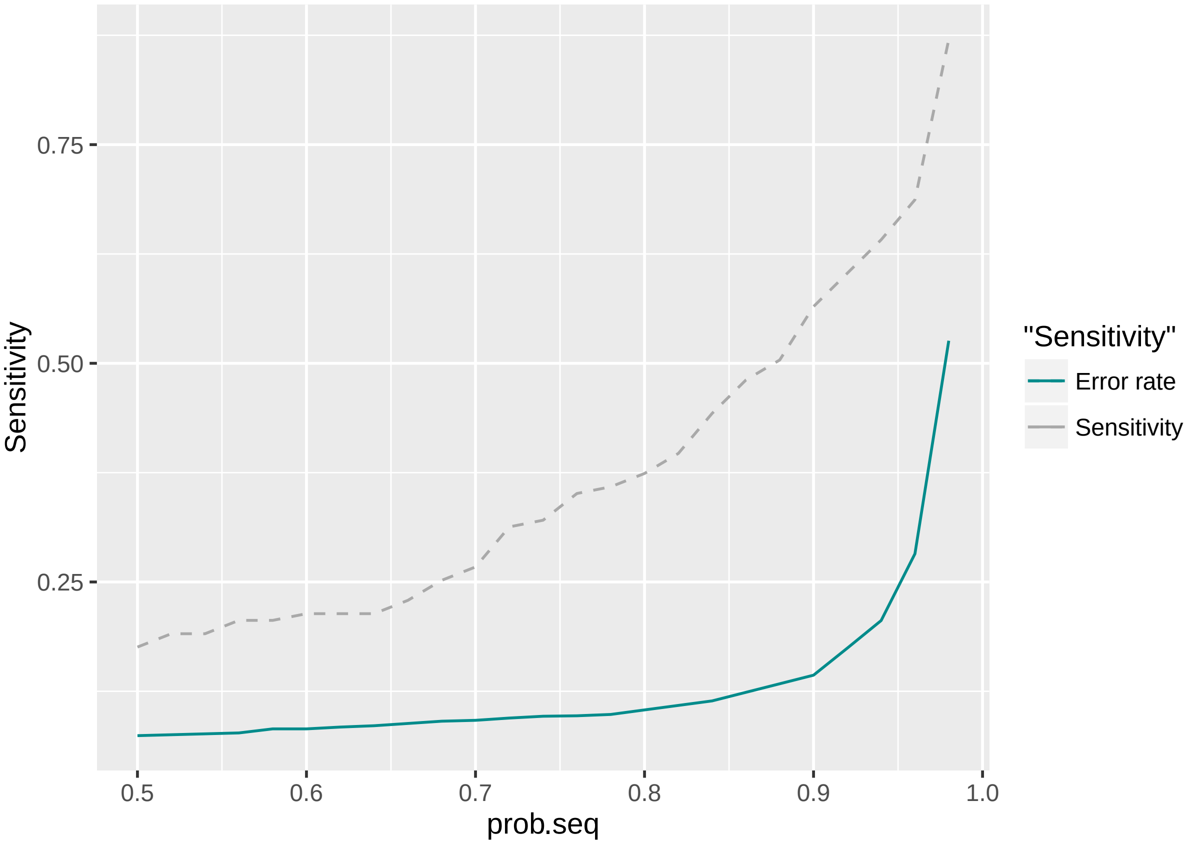

PC1 vs PC3


PC2 vs PC3

```
lda.class    0    1
        0 1771  108
        1   36   23
```

```
[1] 0.1755725
```

```
[1] 0.9800775
```



```
lda.class   0   1
        0 591  34
        1  16   5
```

```
[1] 0.1282051
```

```
[1] 0.9736409
```

Logistic Regression on the Training and Test Sets

```
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
```

The predictors that are significant in our logistic model are seismic, shift and gpuls. The predictors nbumps6, nbumps7 and nbumps89 were removed as they did not provide any data.

```
glm.pred    0    1
       0 1802  125
       1    5    6
```

```
[1] 0.9329205
```

The diagonal elements of the confusion matrix indicate correct predictions, while the off-diagonals represent incorrect predictions. Hence our model on the training data set correctly predicted that the seismic activity would be of no harzard on 1786 observations and that it would be of hazard on 0 observations, for a total of 1786 + 0 = 1786 correct predictions. The mean() function can be used to compute the fraction of hazards for which the prediction was correct. In this case, logistic regression correctly predicted the class of hazard 92 percent of the time. The bad part about this 92 percent of the time is that it did not get any of our actual real hazards observations correct!!!

```
##
## glm.pred   0   1
##        0 602  37
##        1   5   2
```

```
## [1] 0.9349845
```

The diagonal elements of the confusion matrix indicate correct predictions, while the off-diagonals represent incorrect predictions. Hence our model on the testing data set correctly predicted that the seismic activity would be of no harzard on 605 observations and that it would be hazardous on 2 observations, for a total of 602 + 2 = 604 correct predictions. The mean() function can be used to compute the fraction of seismic activity for which the prediction was correct. In this case, logistic regression correctly predicted class of hazard 93.5 % of the time. However, again worrisome, is that the model miss 5 observations that were hazardous instances and 37 that were not hazardous.

Recall that the logistic regression model had only 3 predictors that were significant from an avaiable 19. Perhaps by removing the variables that appear not to be helpful in predicting seismic hazard, we can obtain a more effective model. After all, using predictors that have no relationship with the response tends to cause a deterioration in the test error rate (since such predictors cause an increase in variance without a corresponding decrease in bias), and so removing such predictors may in turn yield an improvement [straight from the book]