557_Project_2BS

Ben Straub, Hillary Koch, Jiawei Huang, Arif Masrur

3/15/2017

Boring Stuff on the dataset

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"
```

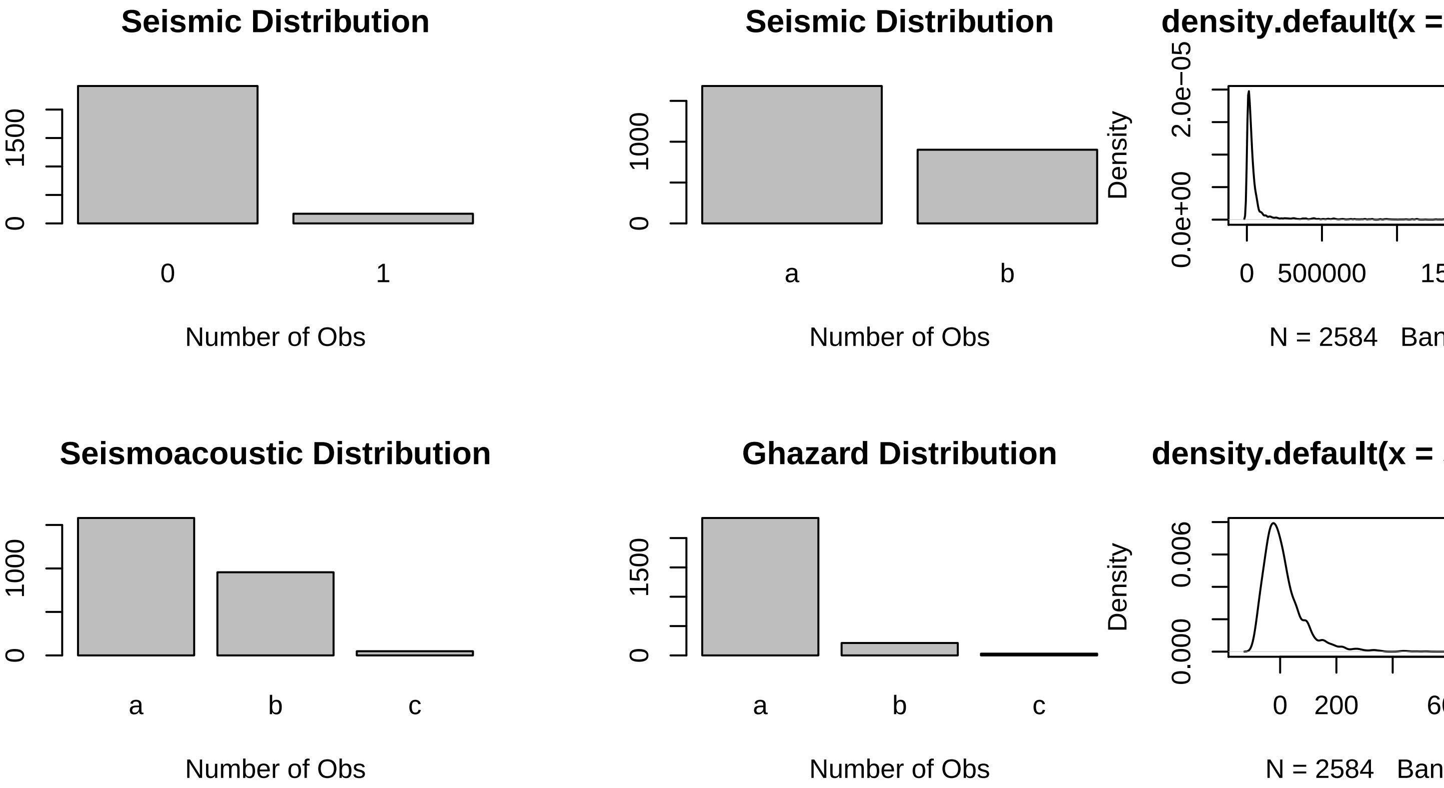
Summary Statistics

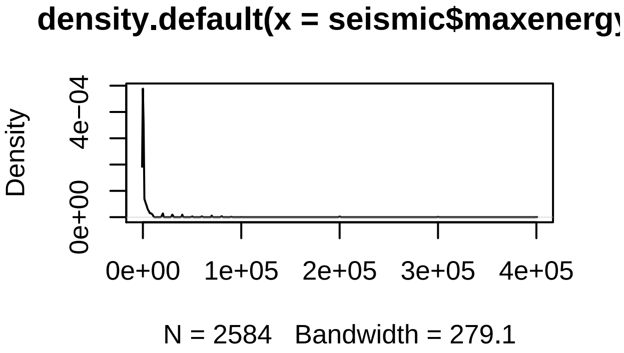
```
 seismic  seismoacoustic shift        genergy            gpuls
 a:1682   a:1580         N: 921   Min.   :    100   Min.   :   2.0
 b: 902   b: 956         W:1663   1st Qu.:  11660   1st Qu.: 190.0
          c:  48                  Median :  25485   Median : 379.0
                                  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  
```

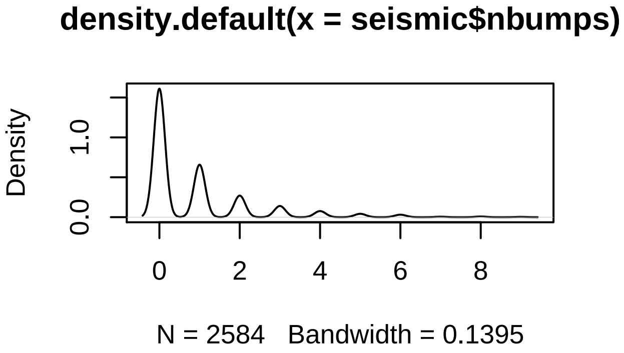
Dimensions of Data Matrix

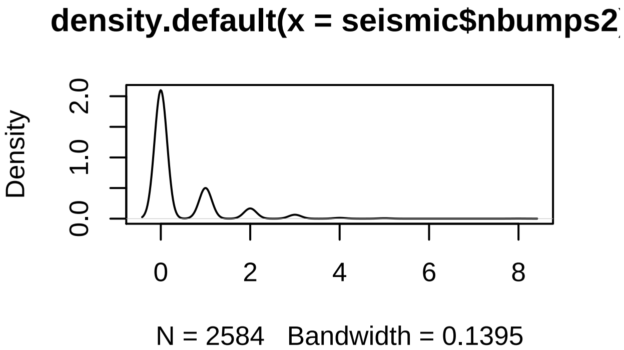
```
[1] 2584   19
```

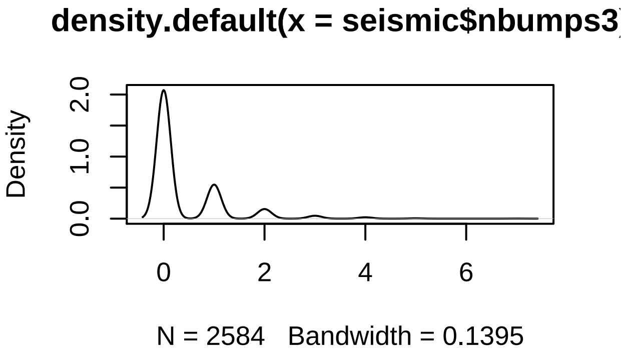
Check for Normality of Data

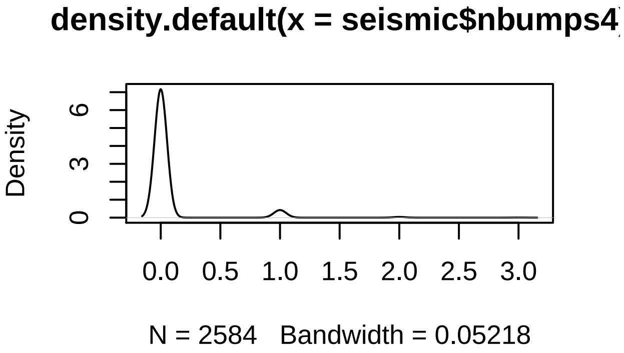


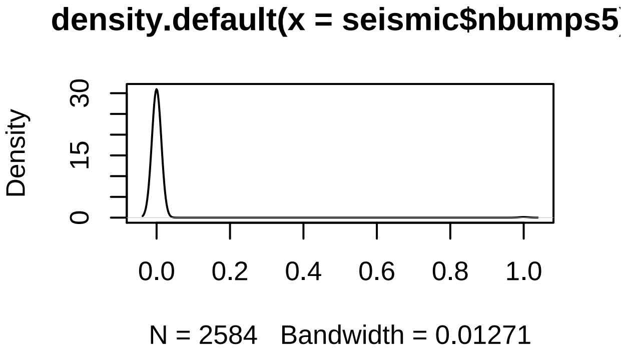

density.default(x = seismic$maxenergy
Density
0e+00
4e−04
0e+00
1e+05
2e+05
3e+05
4e+05
N = 2584   Bandwidth = 279.1

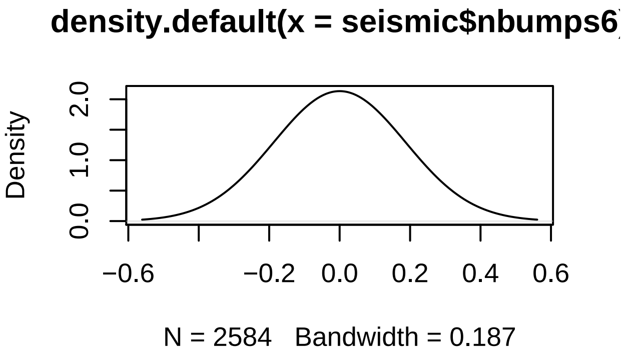

density.default(x = seismic$nbumps)
Density
0.0
1.0
0
2
4
6
8
N = 2584   Bandwidth = 0.1395

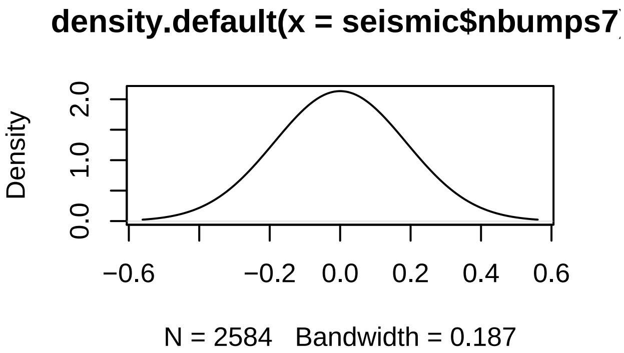

density.default(x = seismic$nbumps2
Density
0.0
1.0
2.0
0
2
4
6
8
N = 2584   Bandwidth = 0.1395

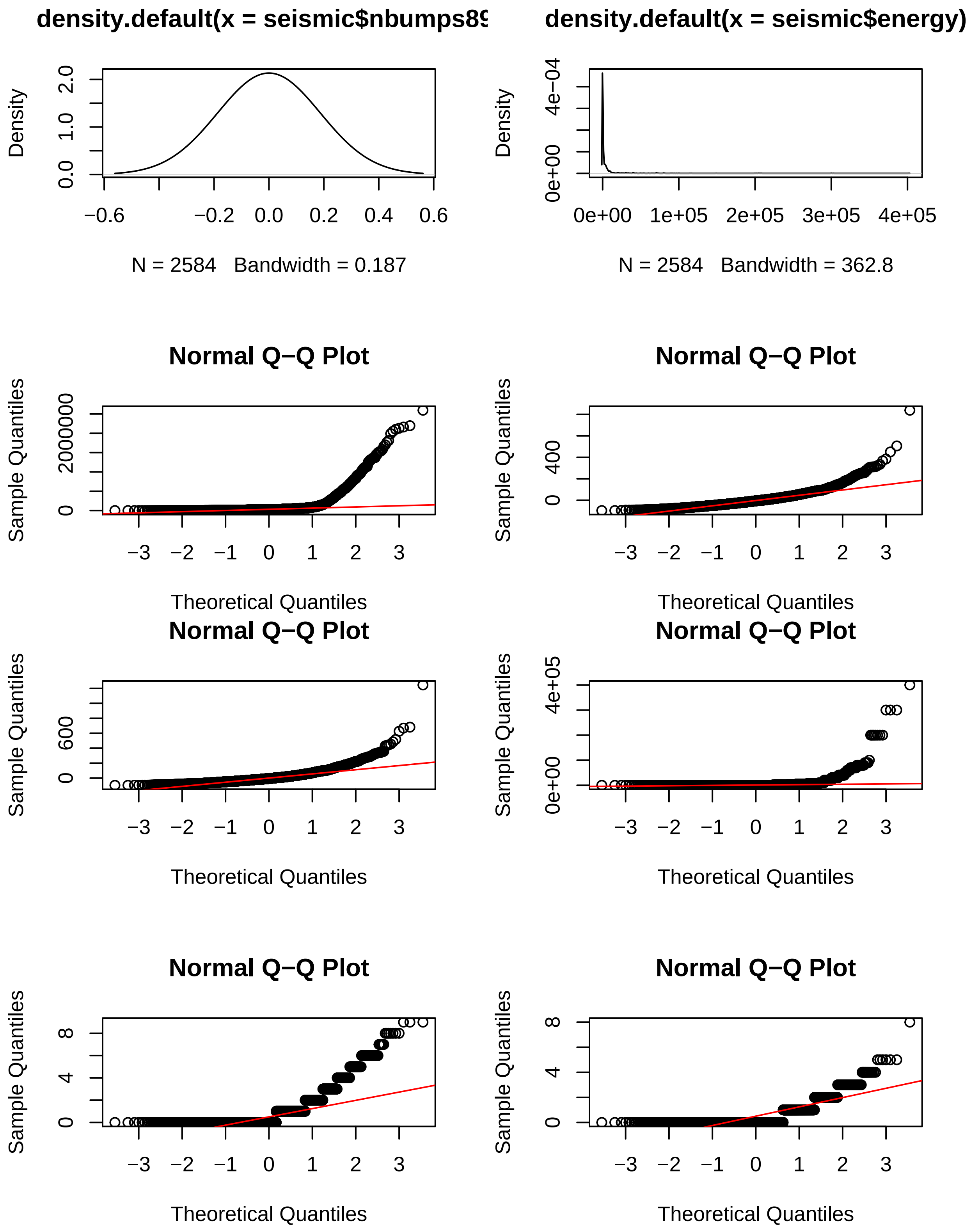

density.default(x = seismic$nbumps3
Density
0.0
1.0
2.0
0
2
4
6
N = 2584   Bandwidth = 0.1395

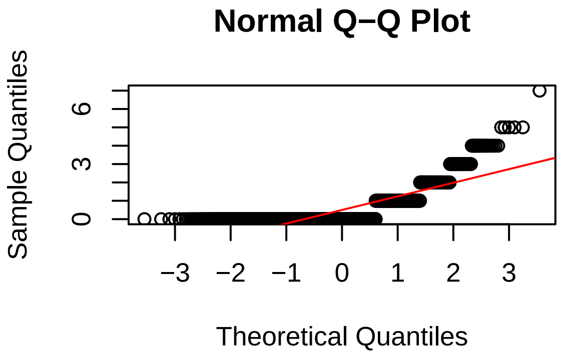

density.default(x = seismic$nbumps4
Density
0
3
6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
N = 2584   Bandwidth = 0.05218

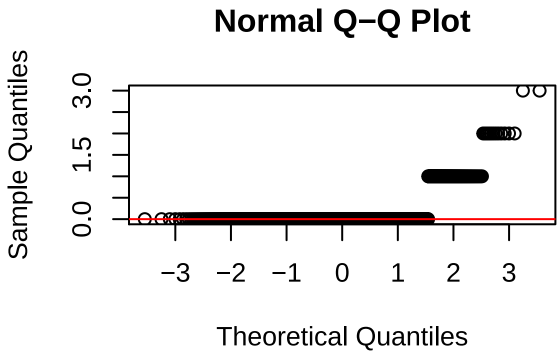

density.default(x = seismic$nbumps5
Density
0
15
30
0.0
0.2
0.4
0.6
0.8
1.0
N = 2584   Bandwidth = 0.01271

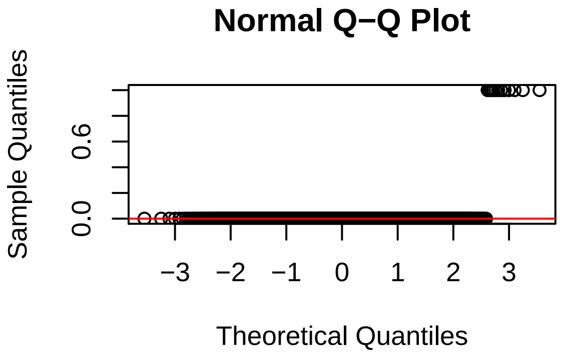

density.default(x = seismic$nbumps6
Density
0.0
1.0
2.0
−0.6
−0.2
0.0
0.2
0.4
0.6
N = 2584   Bandwidth = 0.187

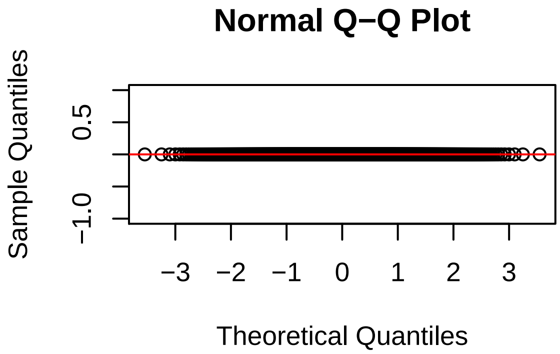

density.default(x = seismic$nbumps7
Density
0.0
1.0
2.0
−0.6
−0.2
0.0
0.2
0.4
0.6
N = 2584   Bandwidth = 0.187

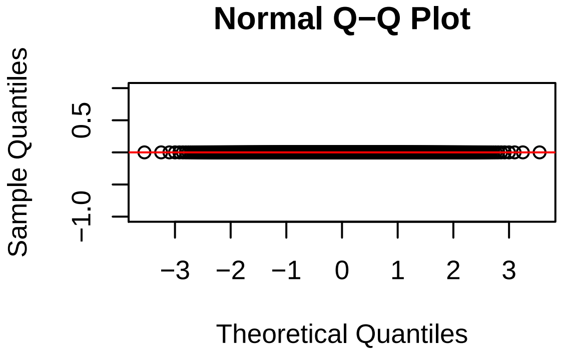

density.default(x = seismic$nbumps89
Density
0.0
1.0
2.0
−0.6
−0.2
0.0
0.2
0.4
0.6
N = 2584   Bandwidth = 0.187
density.default(x = seismic$energy)
Density
0e+00
4e−04
0e+00
1e+05
2e+05
3e+05
4e+05
N = 2584   Bandwidth = 362.8
Normal Q−Q Plot
Sample Quantiles
0
2000000
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Normal Q−Q Plot
Sample Quantiles
0
400
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Normal Q−Q Plot
Sample Quantiles
0
600
−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
0
4
8
−3
−2
−1
0
1
2
3
Theoretical Quantiles
Normal Q−Q Plot
Sample Quantiles
0
4
8
−3
−2
−1
0
1
2
3
Theoretical Quantiles

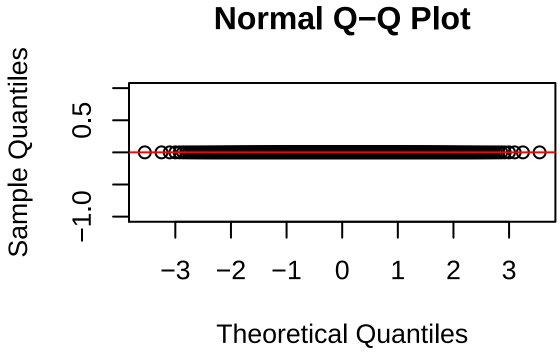

Normal Q−Q Plot
Sample Quantiles
0
3
6
−3
−2
−1
0
1
2
3
Theoretical Quantiles

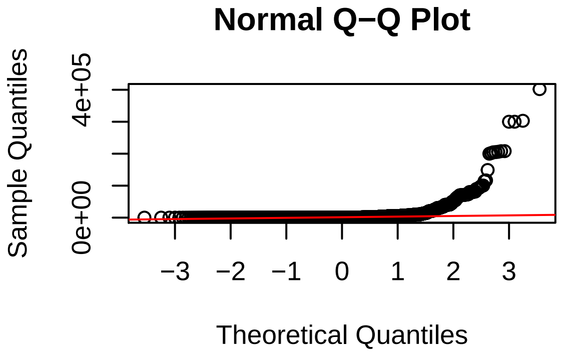

Normal Q−Q Plot
Sample Quantiles
0.0
1.5
3.0
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Normal Q−Q Plot
Sample Quantiles
0.0
0.6
−3
−2
−1
0
1
2
3
Theoretical Quantiles

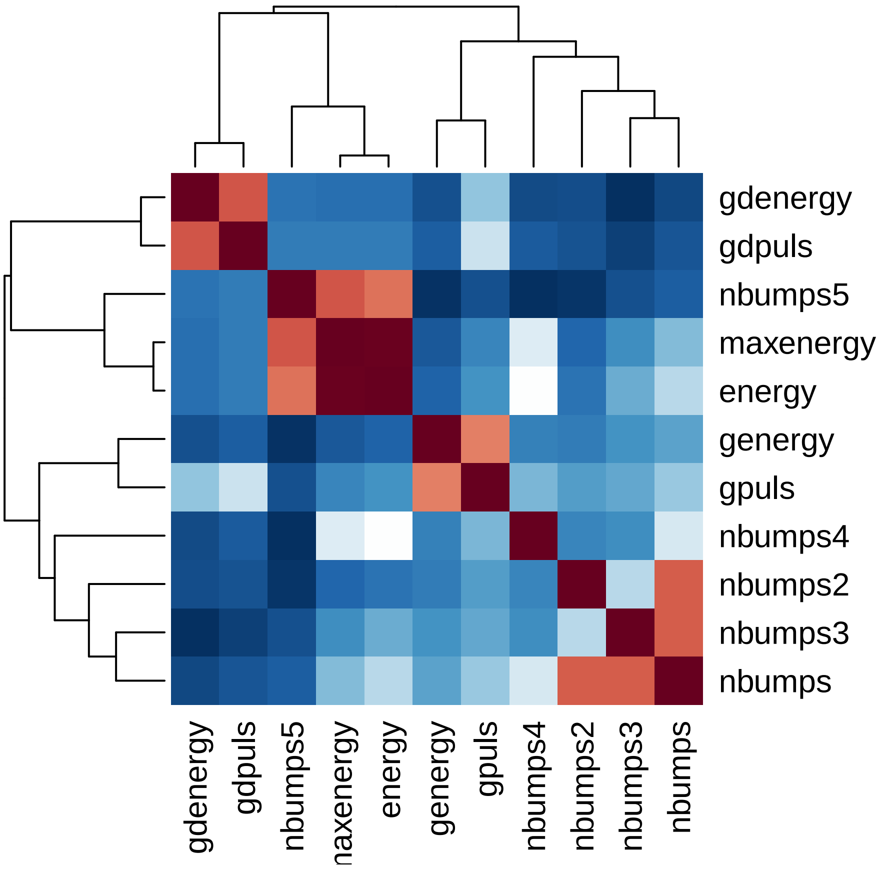

Normal Q−Q Plot
Sample Quantiles
−1.0
0.5
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Normal Q−Q Plot
Sample Quantiles
−1.0
0.5
−3
−2
−1
0
1
2
3
Theoretical Quantiles


Normal Q−Q Plot
Sample Quantiles
−1.0
0.5
−3
−2
−1
0
1
2
3
Theoretical Quantiles

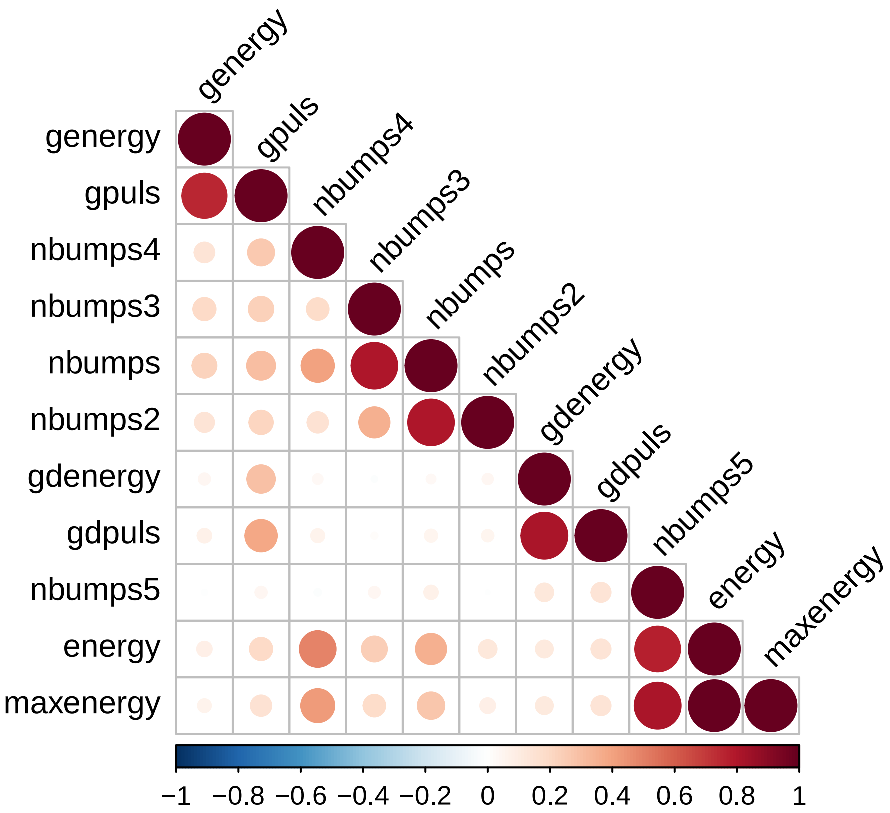

Normal Q−Q Plot
Sample Quantiles
0e+00
4e+05
−3
−2
−1
0
1
2
3
Theoretical Quantiles

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  gdenergy  0.0490 1.4e-02
17   gpuls  gdenergy  0.2900 0.0e+00
18 nbumps4  gdenergy  0.0370 6.1e-02
19 nbumps3  gdenergy -0.0120 5.4e-01
20  nbumps  gdenergy  0.0300 1.3e-01
21 nbumps2  gdenergy  0.0410 3.6e-02
22 genergy    gdpuls  0.0720 2.7e-04
```

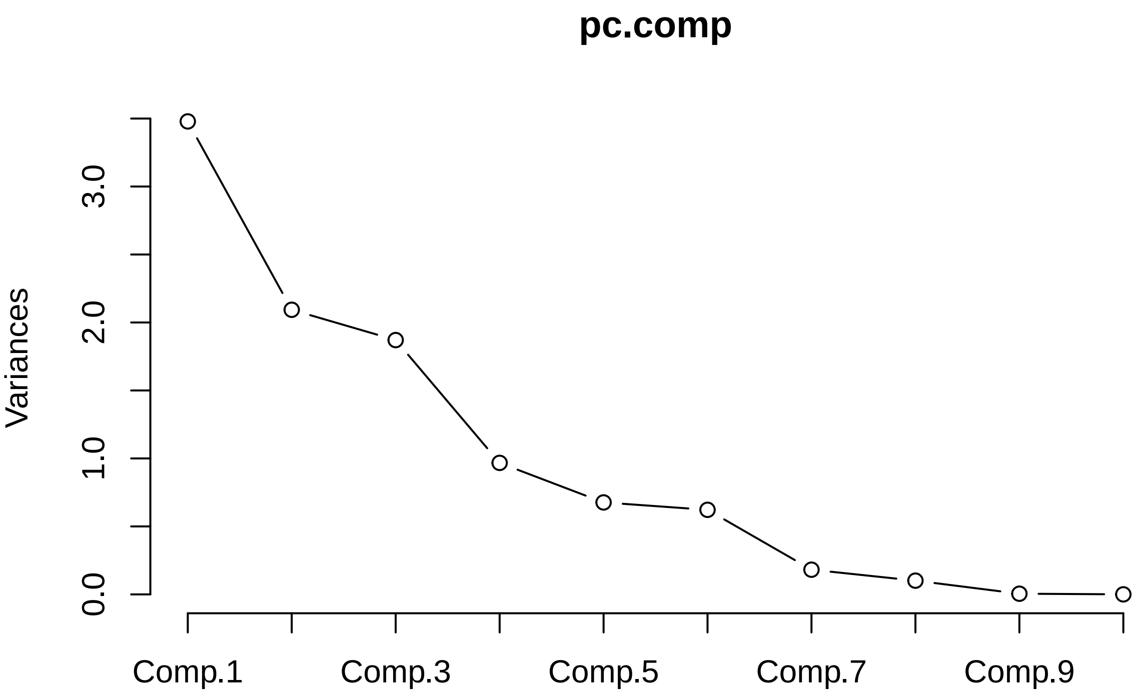
```
23    gpuls    gdpuls  0.3800 0.0e+00
24  nbumps4    gdpuls  0.0660 7.6e-04
25  nbumps3    gdpuls  0.0150 4.5e-01
26   nbumps    gdpuls  0.0580 3.2e-03
27  nbumps2    gdpuls  0.0510 9.4e-03
28 gdenergy    gdpuls  0.8100 0.0e+00
29  genergy   nbumps5 -0.0099 6.2e-01
30    gpuls   nbumps5  0.0490 1.2e-02
31  nbumps4   nbumps5 -0.0170 4.0e-01
32  nbumps3   nbumps5  0.0460 1.8e-02
33   nbumps   nbumps5  0.0700 4.0e-04
34  nbumps2   nbumps5 -0.0053 7.9e-01
35 gdenergy   nbumps5  0.1200 3.3e-10
36   gdpuls   nbumps5  0.1400 5.9e-13
37  genergy    energy  0.0810 3.9e-05
38    gpuls    energy  0.1900 0.0e+00
39  nbumps4    energy  0.4900 0.0e+00
40  nbumps3    energy  0.2400 0.0e+00
41   nbumps    energy  0.3500 0.0e+00
42  nbumps2    energy  0.1200 2.0e-10
43 gdenergy    energy  0.1100 6.7e-08
44   gdpuls    energy  0.1400 2.5e-13
45  nbumps5    energy  0.7700 0.0e+00
46  genergy maxenergy  0.0640 1.1e-03
47    gpuls maxenergy  0.1600 0.0e+00
48  nbumps4 maxenergy  0.4200 0.0e+00
49  nbumps3 maxenergy  0.1800 0.0e+00
50   nbumps maxenergy  0.2700 0.0e+00
51  nbumps2 maxenergy  0.0850 1.5e-05
52 gdenergy maxenergy  0.1100 3.2e-08
53   gdpuls maxenergy  0.1400 2.2e-13
54  nbumps5 maxenergy  0.8100 0.0e+00
55   energy maxenergy  0.9900 0.0e+00

$p
NULL

$sym
NULL
```

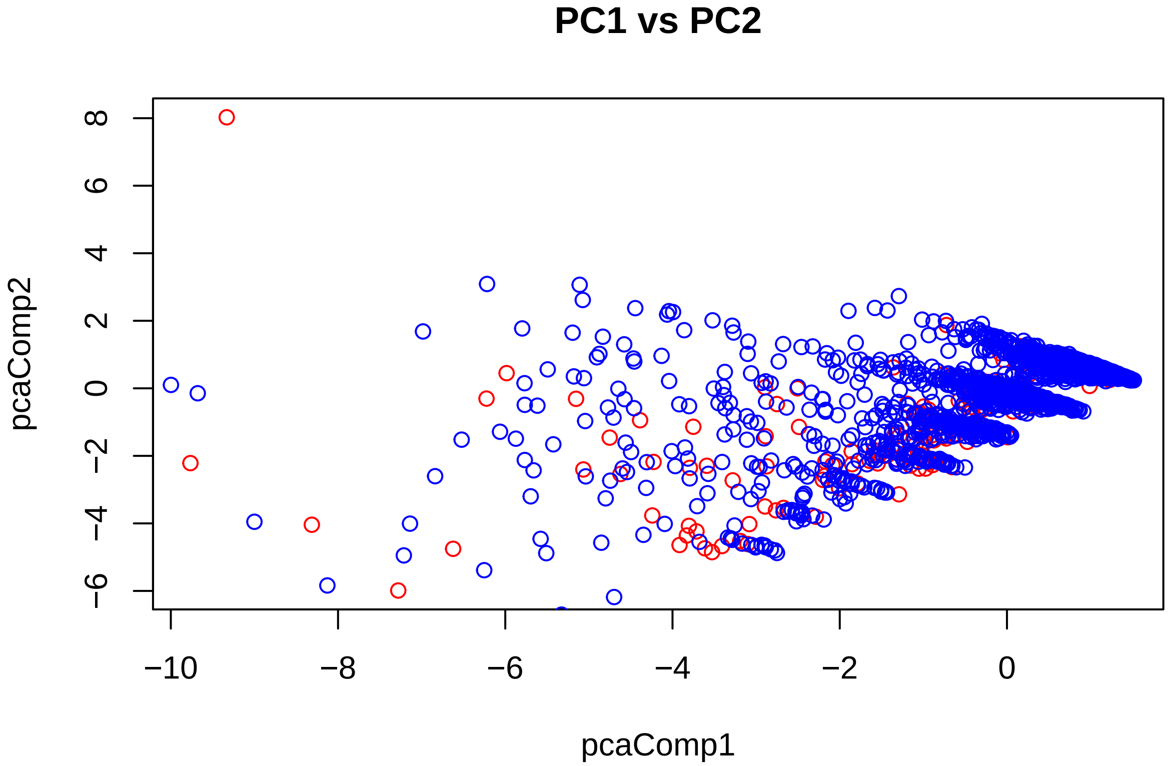


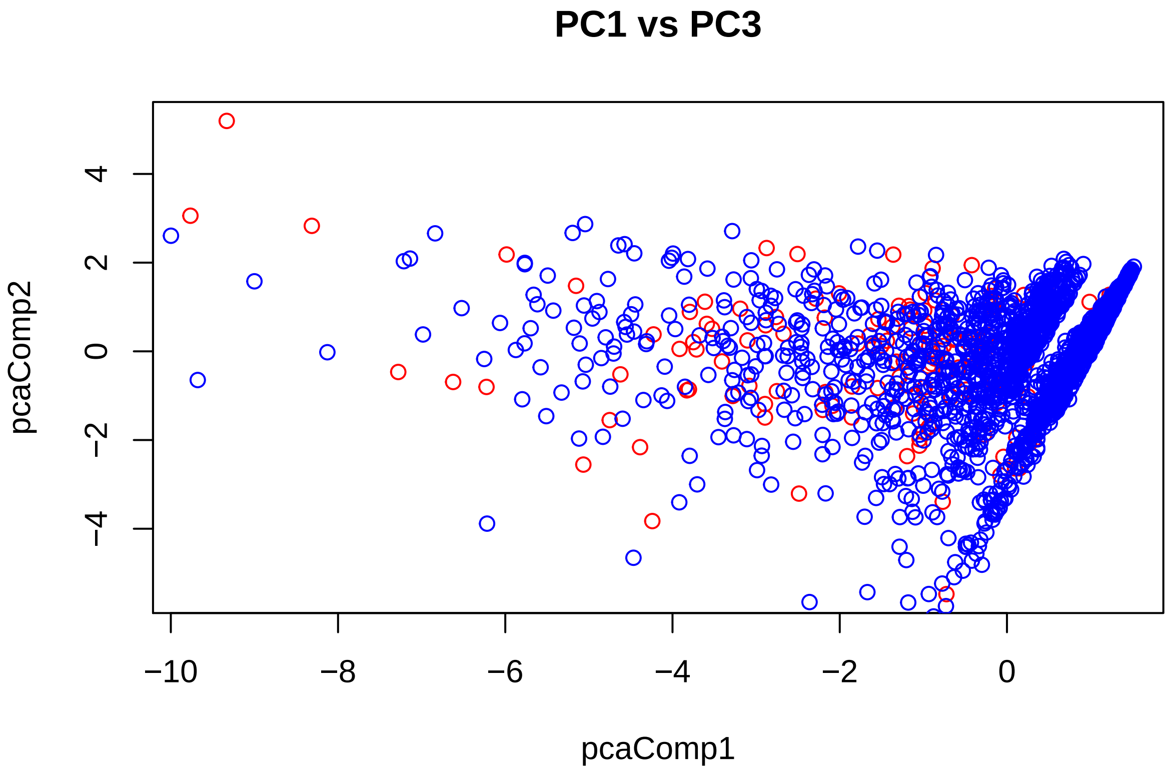
```
$r
        row    column     cor       p
1   genergy     gpuls  0.7500 0.0e+00
2   genergy   nbumps4  0.1500 1.4e-14
3     gpuls   nbumps4  0.2600 0.0e+00
4   genergy   nbumps3  0.1900 0.0e+00
5     gpuls   nbumps3  0.2300 0.0e+00
6   nbumps4   nbumps3  0.1800 0.0e+00
7   genergy    nbumps  0.2200 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  gdenergy  0.0490 1.4e-02
17    gpuls  gdenergy  0.2900 0.0e+00
18  nbumps4  gdenergy  0.0370 6.1e-02
19  nbumps3  gdenergy -0.0120 5.4e-01
20   nbumps  gdenergy  0.0300 1.3e-01
21  nbumps2  gdenergy  0.0410 3.6e-02
22  genergy    gdpuls  0.0720 2.7e-04
23    gpuls    gdpuls  0.3800 0.0e+00
24  nbumps4    gdpuls  0.0660 7.6e-04
25  nbumps3    gdpuls  0.0150 4.5e-01
26   nbumps    gdpuls  0.0580 3.2e-03
```

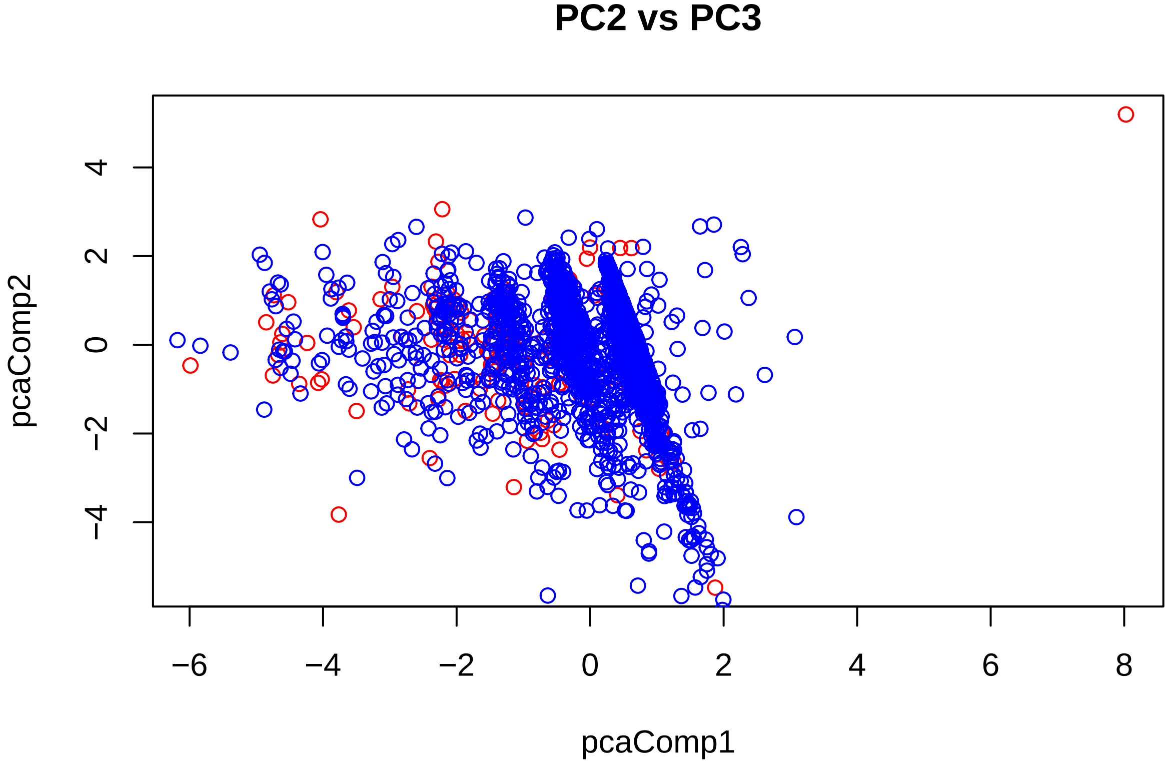
```
27  nbumps2    gdpuls  0.0510 9.4e-03
28 gdenergy    gdpuls  0.8100 0.0e+00
29  genergy   nbumps5 -0.0099 6.2e-01
30    gpuls   nbumps5  0.0490 1.2e-02
31  nbumps4   nbumps5 -0.0170 4.0e-01
32  nbumps3   nbumps5  0.0460 1.8e-02
33   nbumps   nbumps5  0.0700 4.0e-04
34  nbumps2   nbumps5 -0.0053 7.9e-01
35 gdenergy   nbumps5  0.1200 3.3e-10
36   gdpuls   nbumps5  0.1400 5.9e-13
37  genergy    energy  0.0810 3.9e-05
38    gpuls    energy  0.1900 0.0e+00
39  nbumps4    energy  0.4900 0.0e+00
40  nbumps3    energy  0.2400 0.0e+00
41   nbumps    energy  0.3500 0.0e+00
42  nbumps2    energy  0.1200 2.0e-10
43 gdenergy    energy  0.1100 6.7e-08
44   gdpuls    energy  0.1400 2.5e-13
45  nbumps5    energy  0.7700 0.0e+00
46  genergy maxenergy  0.0640 1.1e-03
47    gpuls maxenergy  0.1600 0.0e+00
48  nbumps4 maxenergy  0.4200 0.0e+00
49  nbumps3 maxenergy  0.1800 0.0e+00
50   nbumps maxenergy  0.2700 0.0e+00
51  nbumps2 maxenergy  0.0850 1.5e-05
52 gdenergy maxenergy  0.1100 3.2e-08
53   gdpuls maxenergy  0.1400 2.2e-13
54  nbumps5 maxenergy  0.8100 0.0e+00
55   energy maxenergy  0.9900 0.0e+00

$p
NULL

$sym
NULL
```

Linear regression of an indicator matrix









Logistic Regression on the Training and Test Sets

```
Call:
```

```
glm(formula = y.train ~ ., family = binomial, data = seismic.train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.8338  -0.3884  -0.2855  -0.1560   3.0819  

Coefficients: (3 not defined because of singularities)
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)     -4.696e+00  3.381e-01 -13.890  < 2e-16 ***
seismicb         4.625e-01  2.119e-01   2.183  0.02902 *  
seismoacousticb  2.153e-01  2.108e-01   1.021  0.30704    
seismoacousticc  4.176e-01  8.106e-01   0.515  0.60647    
shiftW           1.174e+00  3.576e-01   3.284  0.00102 ** 
genergy         -2.508e-07  5.044e-07  -0.497  0.61896    
gpuls            7.122e-04  2.474e-04   2.879  0.00400 ** 
gdenergy        -1.428e-04  2.143e-03  -0.067  0.94686    
gdpuls          -3.077e-03  3.058e-03  -1.006  0.31441    
ghazardb        -6.873e-02  3.784e-01  -0.182  0.85586    
ghazardc        -1.373e+01  5.335e+02  -0.026  0.97947    
nbumps           2.106e+01  2.400e+03   0.009  0.99300    
nbumps2         -2.072e+01  2.400e+03  -0.009  0.99311    
nbumps3         -2.071e+01  2.400e+03  -0.009  0.99312    
nbumps4         -2.106e+01  2.400e+03  -0.009  0.99300    
nbumps5         -1.915e+01  2.400e+03  -0.008  0.99363    
nbumps6                 NA         NA      NA       NA    
nbumps7                 NA         NA      NA       NA    
nbumps89                NA         NA      NA       NA    
energy           2.558e-06  4.030e-05   0.063  0.94939    
maxenergy       -7.703e-06  3.966e-05  -0.194  0.84600    
---
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: 811.70  on 1920  degrees of freedom
AIC: 847.7

Number of Fisher Scoring iterations: 15
```

The predictors that are significant in our logistic model are genergy, gpuls and ghazardb and a couple more. The predictors nbumps6, nbumps7 and nbumps89 are not defined due to singularities, which may indicated collinearity.

```
        y.train
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 1176 observations and that it would be a low hazard on 230 observations, for a total of 1176 + 230 = 1406 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 the movement of the market 73 percent of the time.

```
##         y.test
## glm.pred   0   1
##        0 603  37
##        1   4   2
```

```
## [1] 0.9365325
```

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 352 observations and that it would be a low hazard on 110 days, for a total of 352 + 110 = 462 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 the movement of the market 71.5 percent of the time.

Recall that the logistic regression model had only 7ish predictors that were significant from an avaiable 17. 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]