

557_Project_2BS

Ben Straub, Hillary Koch, Jiawei Huang, Arif Masrur

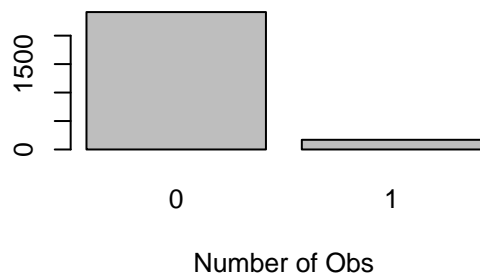
3/15/2017

No Command Lines Ever. Whoa

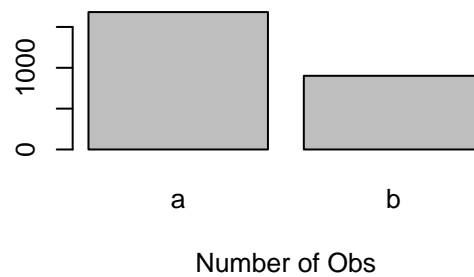
```
## function (... , recursive = FALSE) .Primitive("c")
```

What the Factor Variables look like

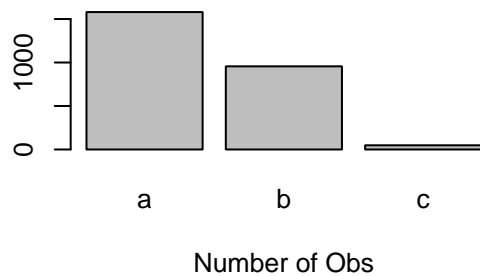
Class/Response Distribution



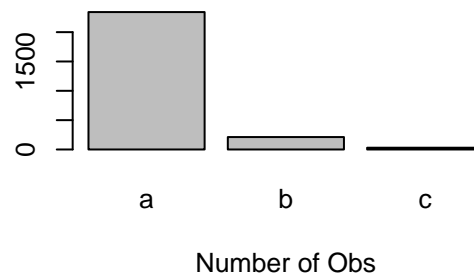
Seismic Distribution



Seismoacoustic Distribution

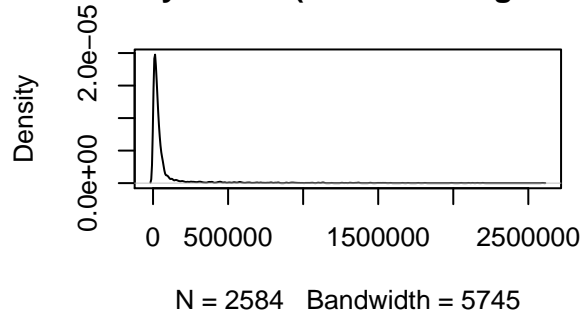


Ghazard Distribution

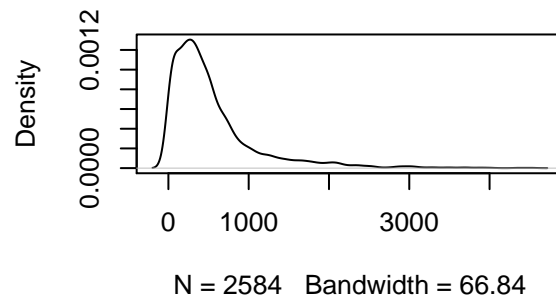


What the Continuous Variables look like

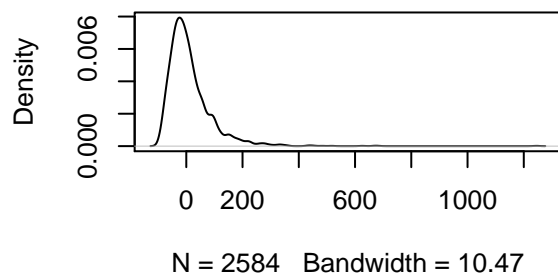
density.default(x = seismic\$genergy)



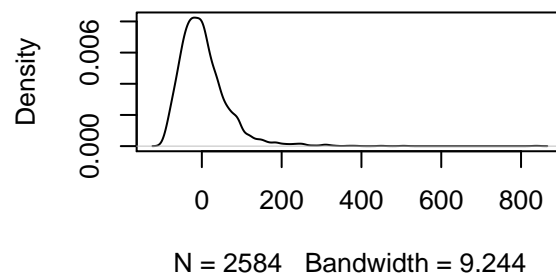
density.default(x = seismic\$gpuls)



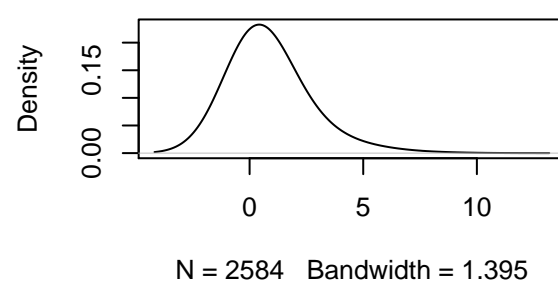
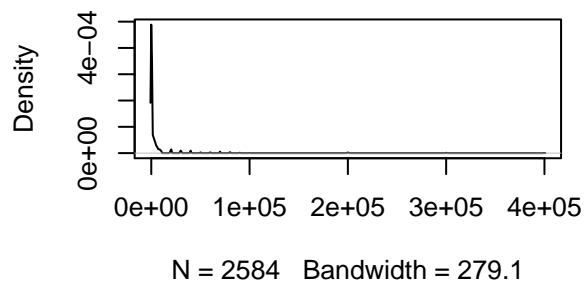
density.default(x = seismic\$gdenergy)



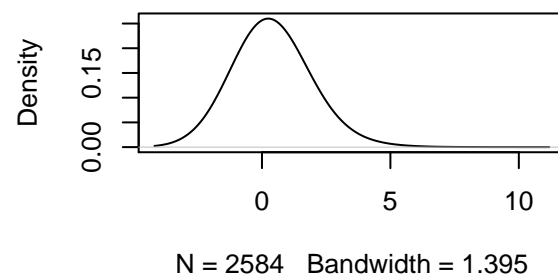
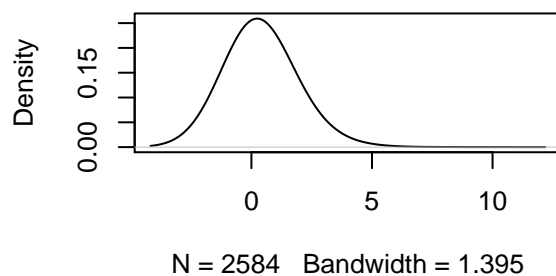
density.default(x = seismic\$gdpuls)



density.default(x = seismic\$maxenergy, adjus



nsity.default(x = seismic\$nbumps2, adjus



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

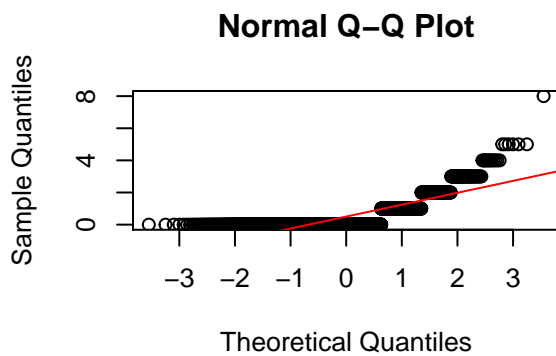
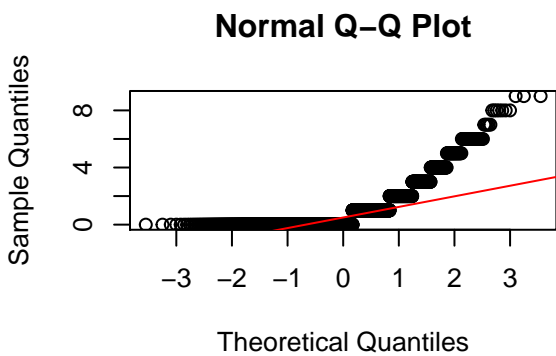
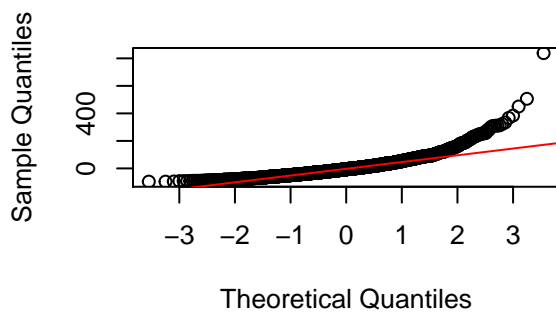
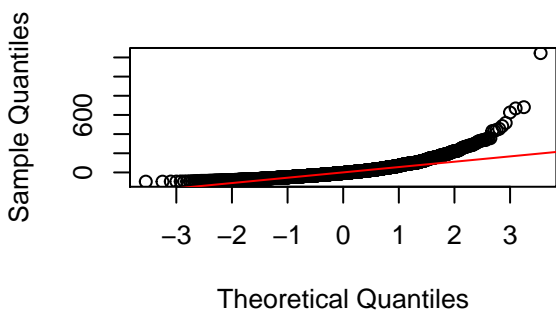
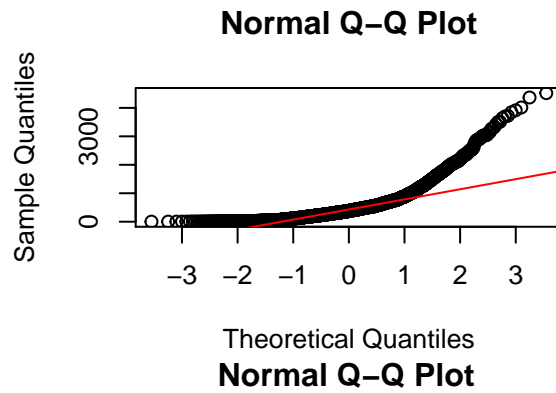
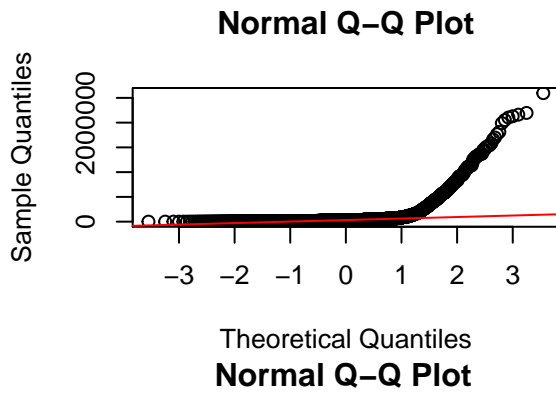
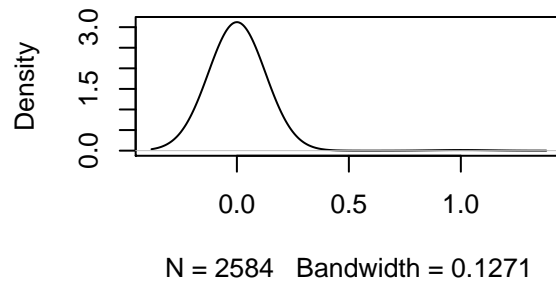
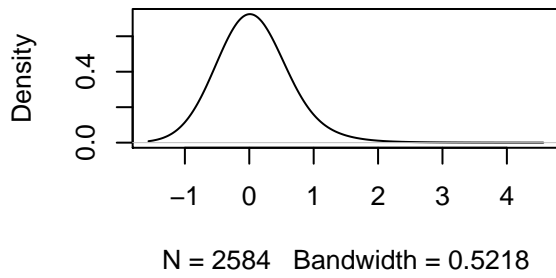
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

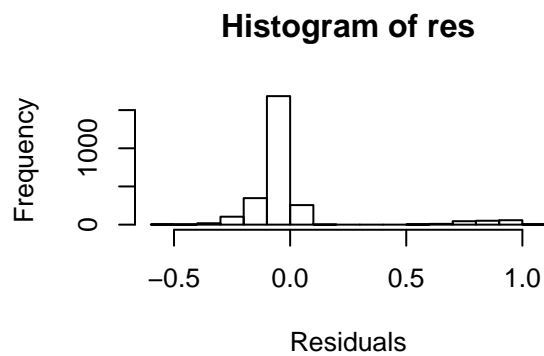
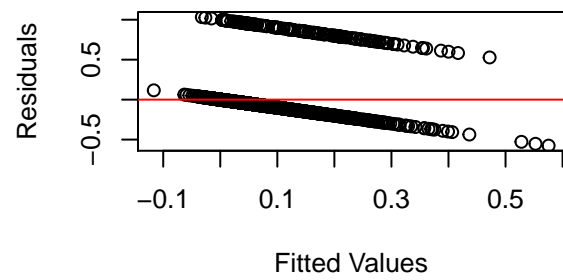
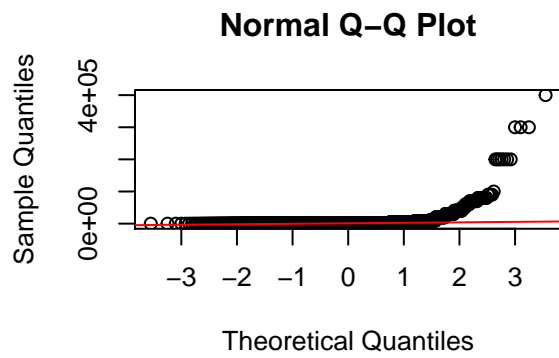
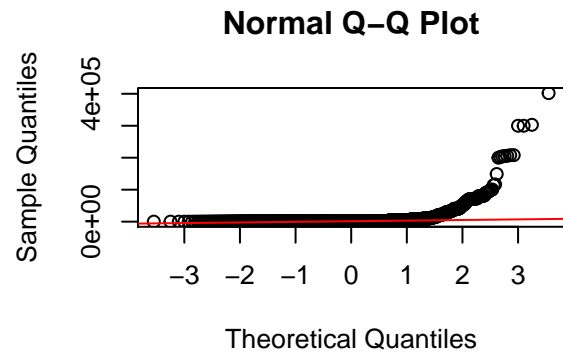
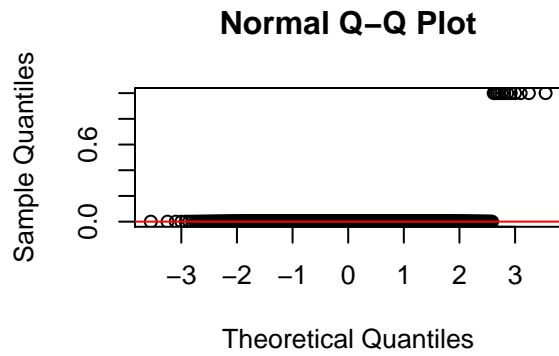
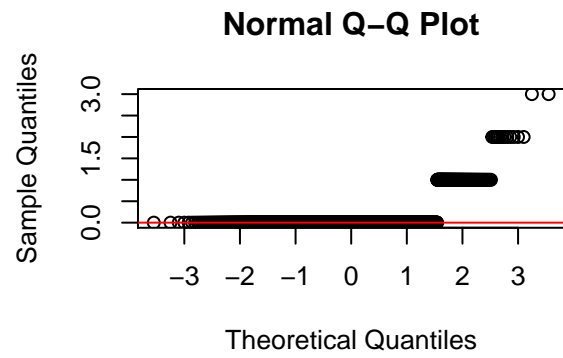
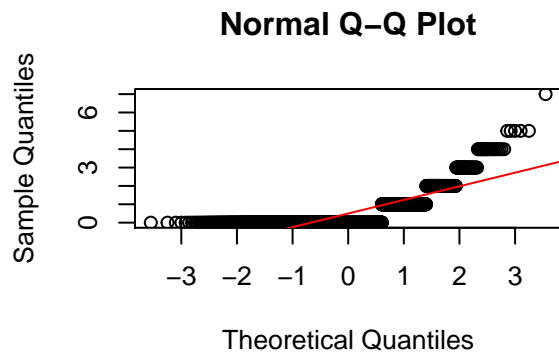
Residual standard error: 0.2371 on 2568 degrees of freedom

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, adjusnsity.default(x = seismic\$nbumps5, adjus



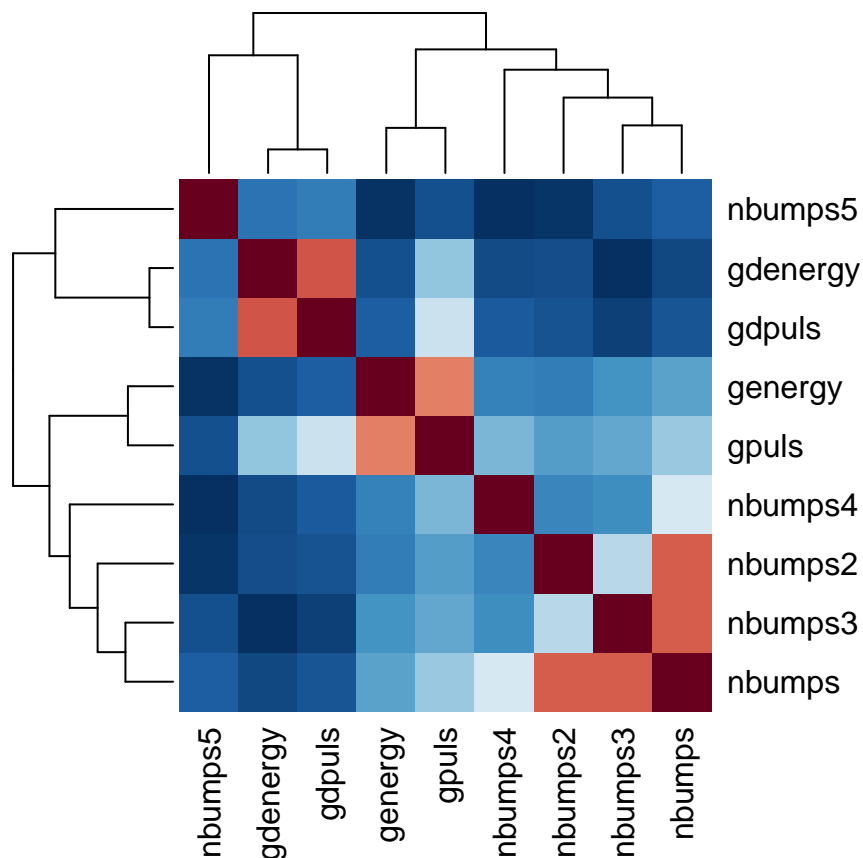


Lots of multicollinearity to worry about during variable selection

```
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
      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
gdenenergy
gdpuls

1

\$p

	genergy	gpuls	nbumps4	nbumps3	nbumps	nbumps2	nbumps5	gdenenergy
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	
gdenenergy	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
gdenenergy
gdpuls

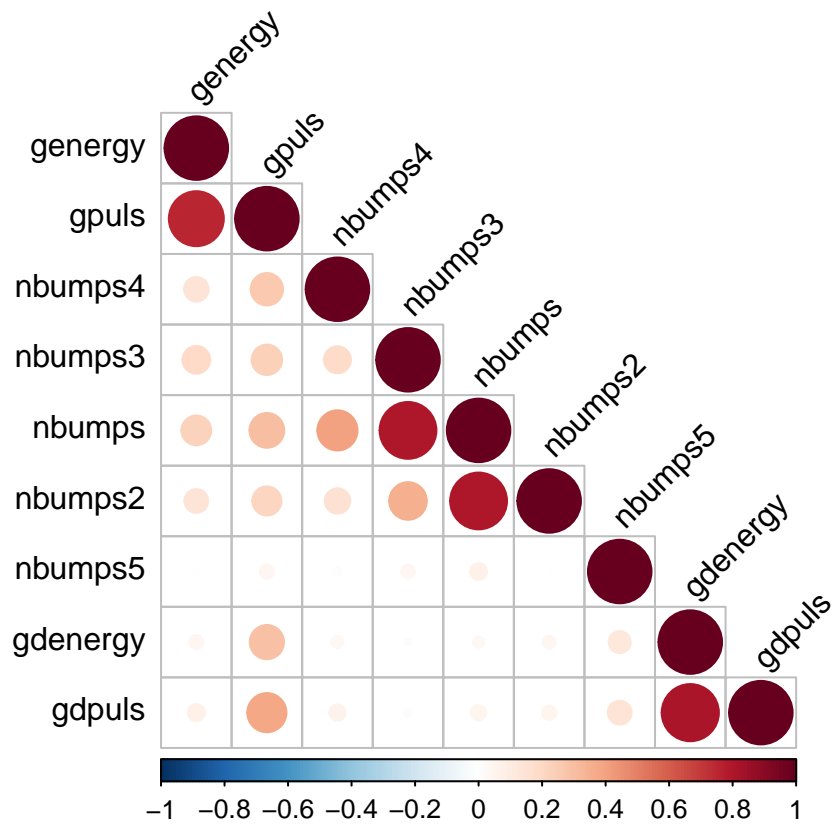
0

\$sym

	genergy	gpuls	nbumps4	nbumps3	nbumps	nbumps2	nbumps5	gdenenergy
genergy	1							
gpuls	,	1						
nbumps4			1					
nbumps3				1				
nbumps		.	,	1				
nbumps2			.	,	1			
nbumps5						1		
gdenenergy							1	
gdpuls		.					+	
gdpuls								

genergy
gpuls
nbumps4
nbumps3
nbumps
nbumps2
nbumps5
gdenenergy

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



```
$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

```

##-----
## 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)
```

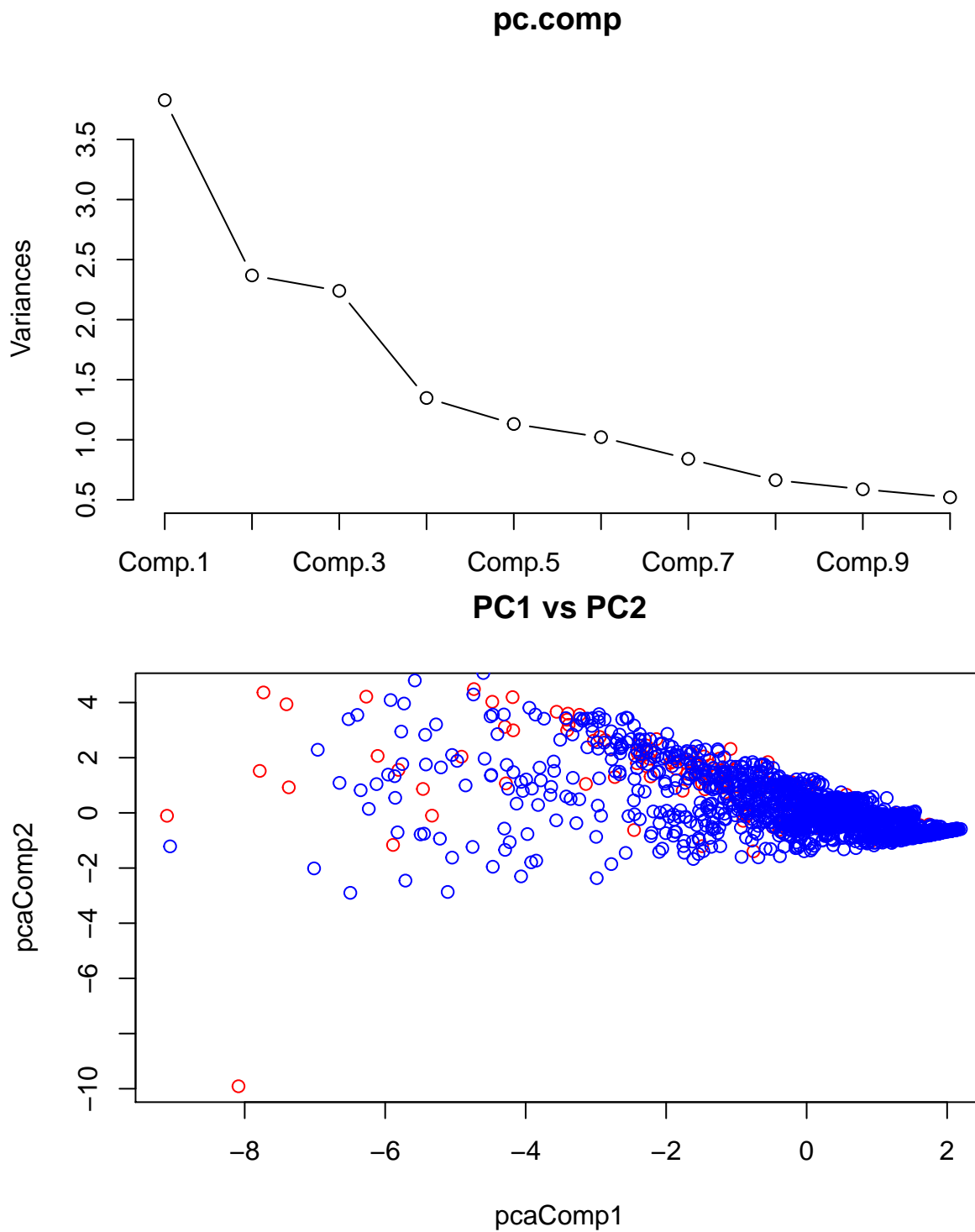
```
[1] 1938  16
```

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

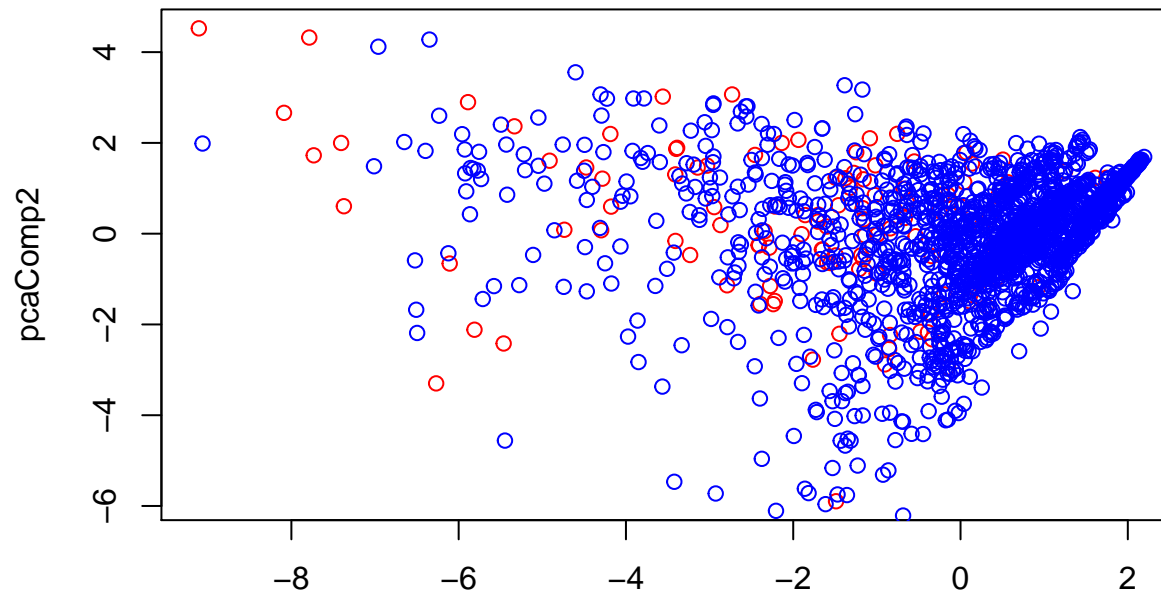
```
[1] 646  16
```

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

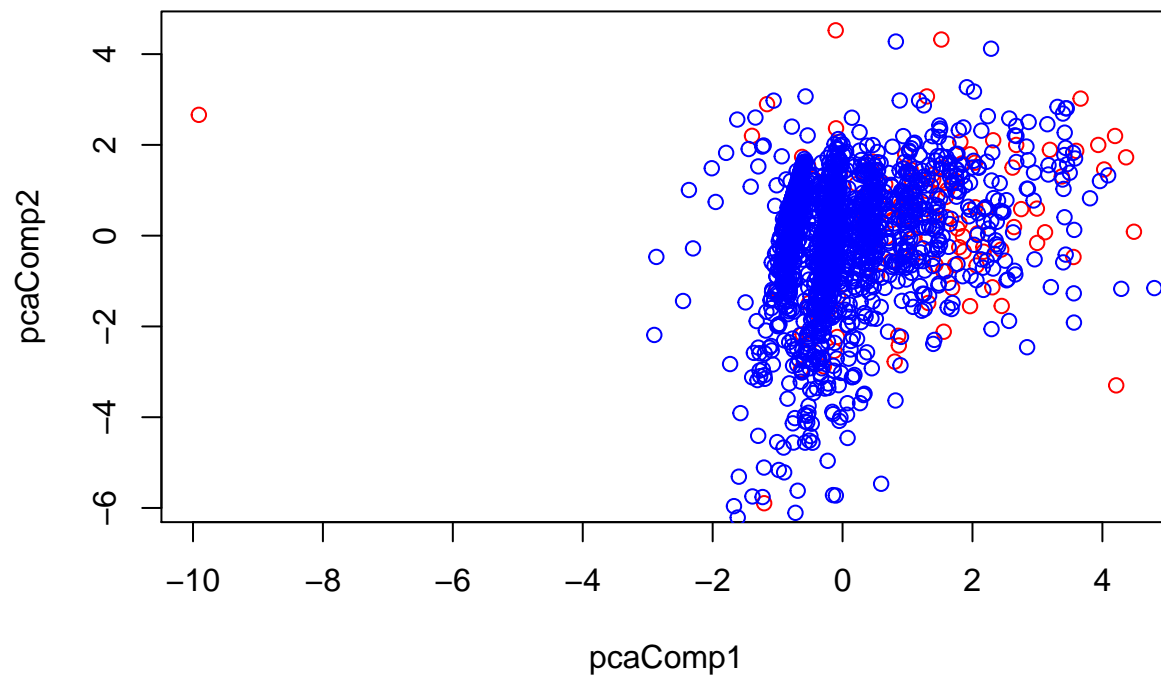
Linear regression of an indicator matrix



PC1 vs PC3



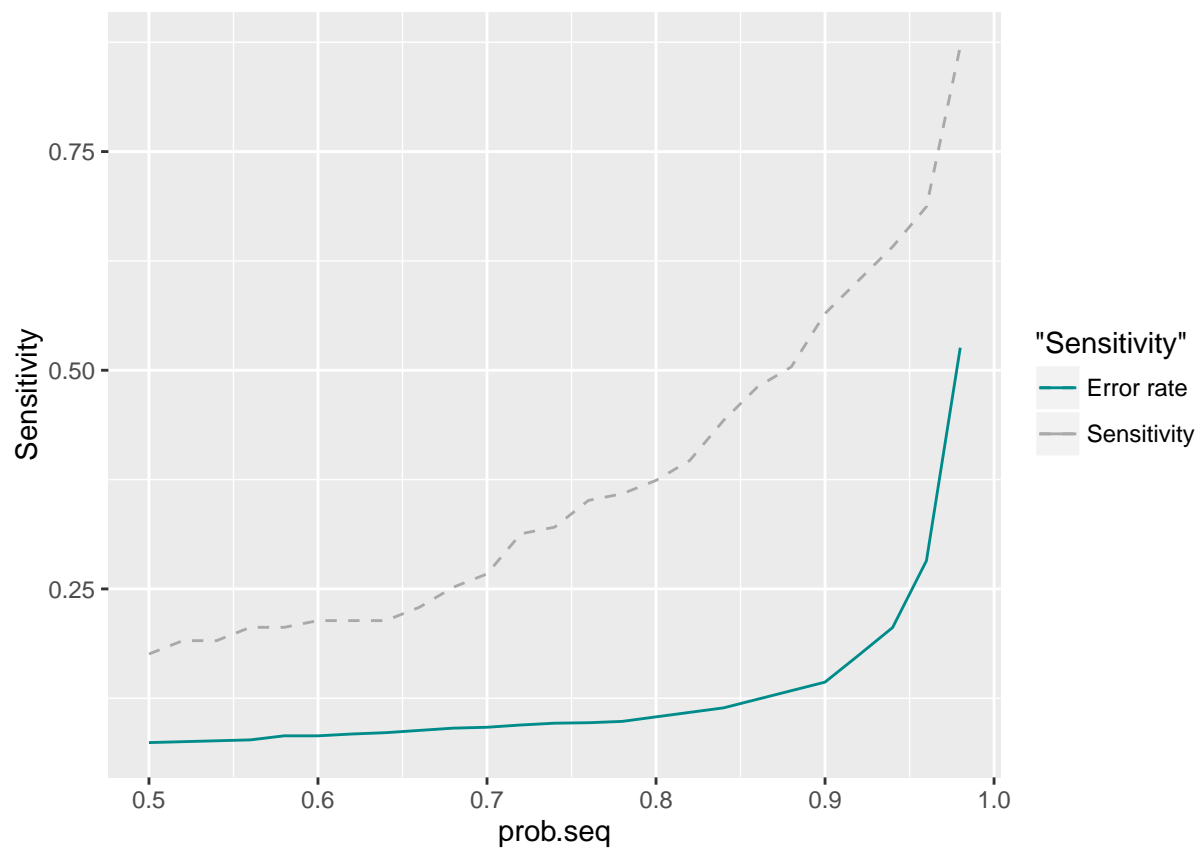
PC2 vs PC3



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

```
[1] 0.1755725
```

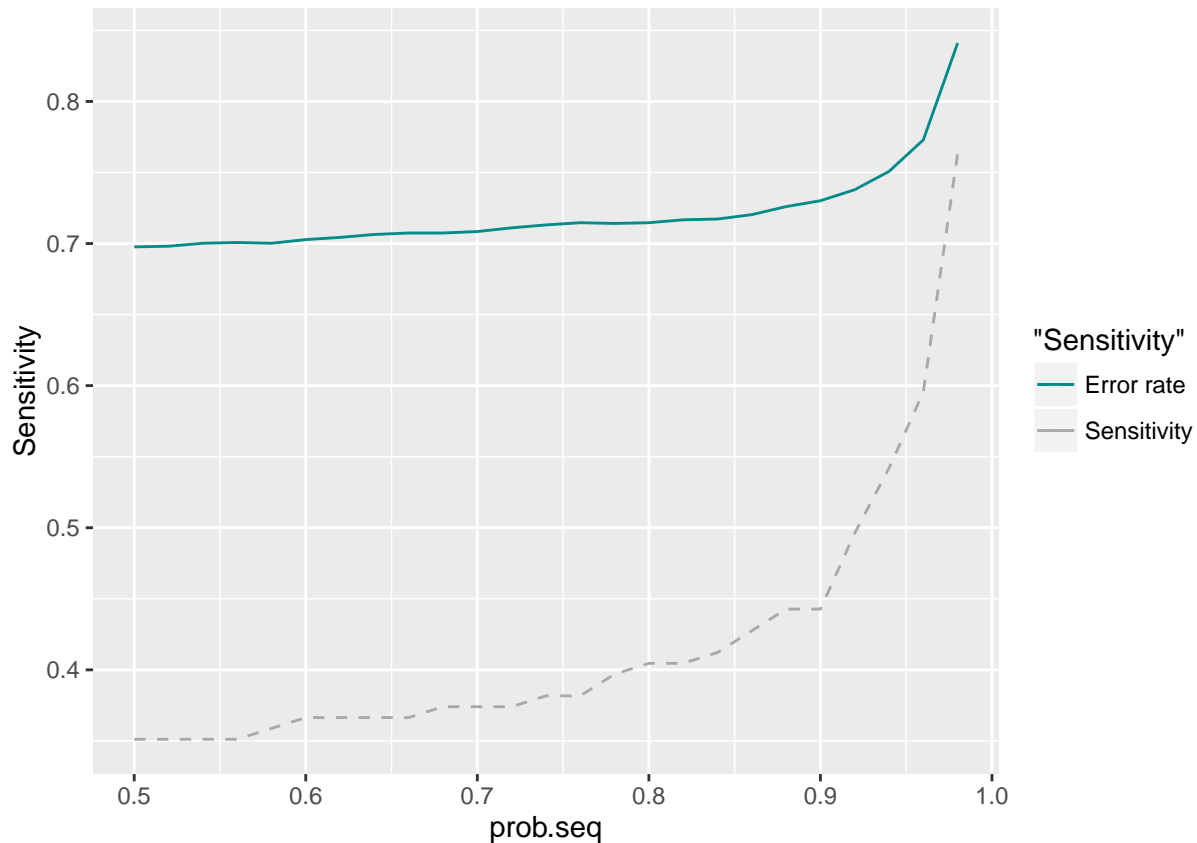
```
[1] 0.9800775
```



```
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.

```
[1] 0.9329205
```

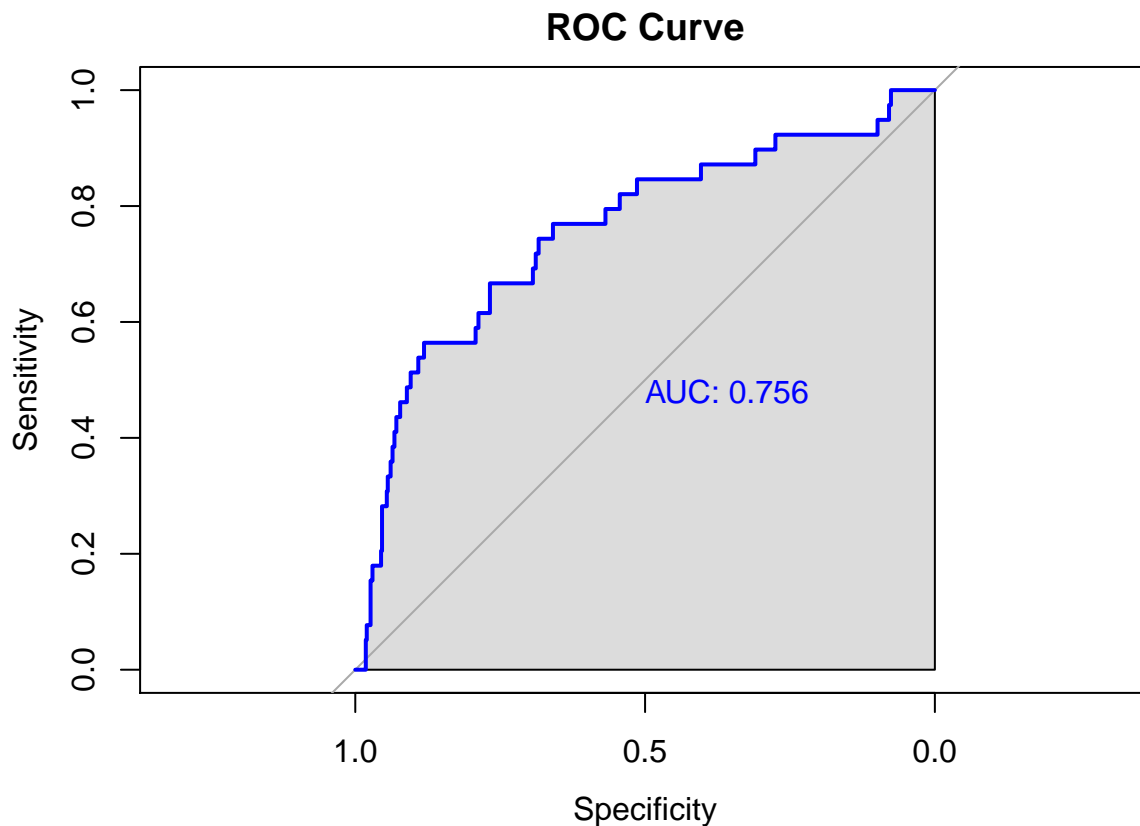
```

glm.pred    0    1
           0 1802 125
           1    5    6

```

```
[1] 0.04580153
```

```
[1] 0.997233
```



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 hazard 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!!!

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

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

```
## [1] 0.05128205
```

```
## [1] 0.9917628
```

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 hazard 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 available 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]