Lab 8

Attila Lazar

02.12.2020

1)

data

We load the dataset. We select observations with response *1* or *3*

```
#install.packages("rrcov")
data(olitos, package="rrcov")
olitos.a <- olitos[which(olitos$grp %in% c(1,3)), -26]
grp <- olitos[which(olitos$grp %in% c(1,3)), "grp"]
y <- ifelse(grp==1, 1, 0)
olitos.a <- cbind(olitos.a, y)

set.seed(1234)
n <- nrow(olitos.a)
train.a <- sample(1:n, round(n*2/3))
test.a <- (1:n) [-train.a]
```

a)

We train our model using the training dataset and use only variables *X1* to *X2*

```
modelglm <- glm(y~X1+X2+X3+X4+X5+X6, data=olitos.a, family="binomial", subset=train.a)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
modelglm
```

```
##
## Call:  glm(formula = y ~ X1 + X2 + X3 + X4 + X5 + X6, family = "binomial",
##     data = olitos.a, subset = train.a)
##
## Coefficients:
## (Intercept)           X1           X2           X3           X4
##   -422.3264       3.2339       7.6684    -141.7143     -68.3168
##          X5           X6
##     -0.4534     -28.0174
##
## Degrees of Freedom: 55 Total (i.e. Null);  49 Residual
## Null Deviance:       73
## Residual Deviance: 15.81     AIC: 29.81
```

```
summary(modelglm)
```

```
##
## Call:
## glm(formula = y ~ X1 + X2 + X3 + X4 + X5 + X6, family = "binomial",
##     data = olitos.a, subset = train.a)
```

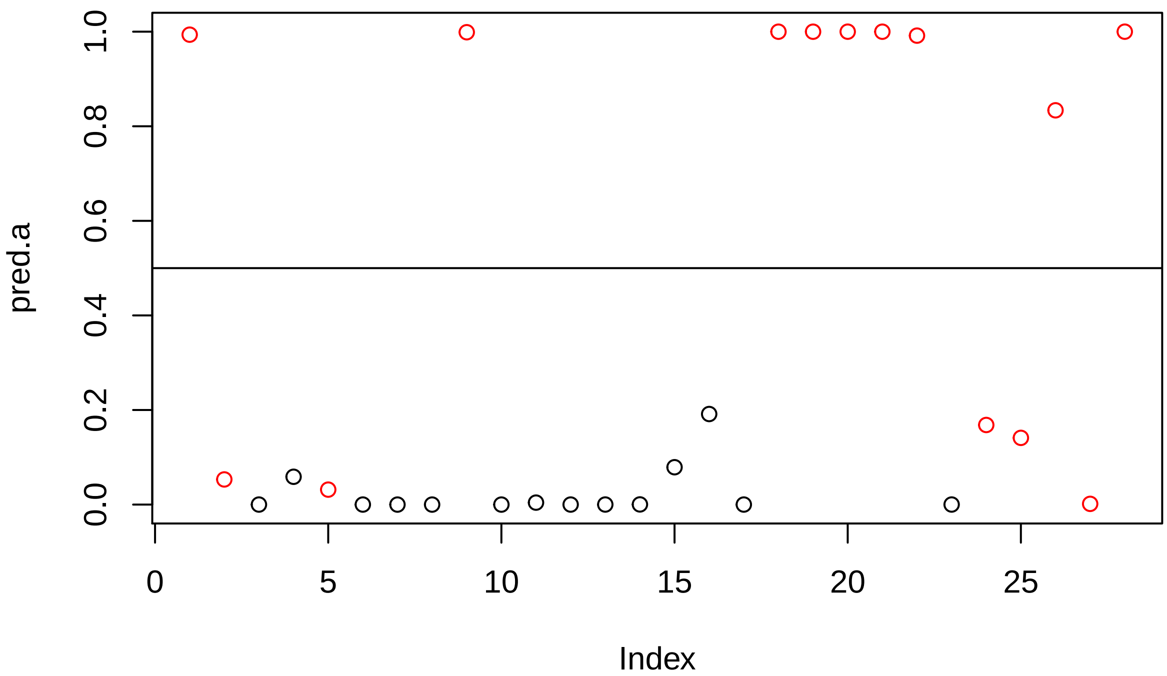
```
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.74125  -0.00531   0.00441   0.16661   2.42507
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -422.3264   442.3161  -0.955   0.3397
## X1             3.2339     1.8515   1.747   0.0807 .
## X2             7.6684     7.2318   1.060   0.2890
## X3          -141.7143    61.4912  -2.305   0.0212 *
## X4           -68.3168   355.5670  -0.192   0.8476
## X5            -0.4534     0.7624  -0.595   0.5520
## X6           -28.0174    11.5332  -2.429   0.0151 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 72.997  on 55  degrees of freedom
## Residual deviance: 15.808  on 49  degrees of freedom
## AIC: 29.808
##
## Number of Fisher Scoring iterations: 8
```

X3 and *X6* and also *X1* seem to be the significatly contributing variables.

b)

we plot predictions for the test set

```
pred.a <- predict(modelglm, olitos.a[test.a,], type="response")
plot(pred.a, col=as.numeric(olitos.a[test.a,"y"]+1))
abline(h=0.5)
```



and calculate the confusion matrix, and the classification error

```
T <- table(olitos.a[test.a,"y"], pred.a>0.5)
T
```

```
##
##     FALSE TRUE
##   0    14    0
##   1     5    9
```

```
e1 <- 1-sum(diag(T))/sum(T)
e1
```

```
## [1] 0.1785714
```

c)

Now we train the model with all variables

```
modelglm.c <- glm(y~.,data=olitos.a, family="binomial", subset=train.a)
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
modelglm.c
```

```
##
## Call:  glm(formula = y ~ ., family = "binomial", data = olitos.a, subset = train.a)
##
## Coefficients:
## (Intercept)           X1           X2           X3           X4
##  -5792.0872      29.8579      49.0796     -39.2290     155.3744
##          X5           X6           X7           X8           X9
##    -13.0026     -86.9438     -58.9490     -21.7592      52.3894
##         X10          X11          X12          X13          X14
```

```
##    -39.6360     -41.1873      -0.3633      40.9740      -4.5058
##         X15          X16          X17          X18          X19
##      5.1552      31.8056     -14.7121      -5.7769      -2.6569
##         X20          X21          X22          X23          X24
##     -3.0258      -1.4192      -0.2279      12.8424      -0.5300
##         X25
##      1.0865
##
## Degrees of Freedom: 55 Total (i.e. Null);  30 Residual
## Null Deviance:       73
## Residual Deviance: 9.636e-10     AIC: 52
```

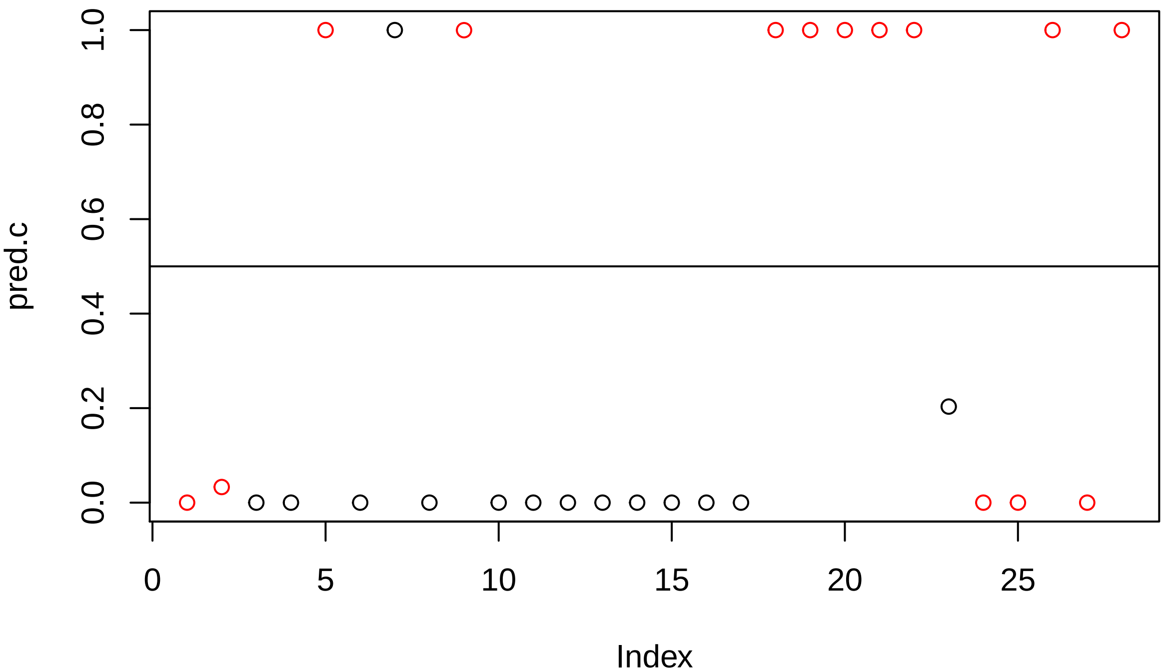
```r
summary(modelglm.c)
```

```
##
## Call:
## glm(formula = y ~ ., family = "binomial", data = olitos.a, subset = train.a)
##
## Deviance Residuals:
##        Min          1Q      Median          3Q         Max
## -7.784e-06  -2.110e-08   2.110e-08   4.538e-07   1.004e-05
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.792e+03  4.780e+07       0        1
## X1           2.986e+01  2.448e+05       0        1
## X2           4.908e+01  6.930e+05       0        1
## X3          -3.923e+01  4.995e+06       0        1
## X4           1.554e+02  5.972e+07       0        1
## X5          -1.300e+01  1.677e+05       0        1
## X6          -8.694e+01  1.403e+06       0        1
## X7          -5.895e+01  4.552e+05       0        1
## X8          -2.176e+01  3.627e+05       0        1
## X9           5.239e+01  7.830e+05       0        1
## X10         -3.964e+01  1.049e+06       0        1
## X11         -4.119e+01  1.360e+06       0        1
## X12         -3.633e-01  3.276e+03       0        1
## X13          4.097e+01  4.374e+05       0        1
## X14         -4.506e+00  5.367e+05       0        1
## X15          5.155e+00  2.429e+05       0        1
## X16          3.181e+01  6.340e+05       0        1
## X17         -1.471e+01  9.346e+04       0        1
## X18         -5.777e+00  1.102e+05       0        1
## X19         -2.657e+00  9.134e+04       0        1
## X20         -3.026e+00  8.382e+04       0        1
## X21         -1.419e+00  9.300e+04       0        1
## X22         -2.279e-01  2.639e+04       0        1
## X23          1.284e+01  1.810e+05       0        1
## X24         -5.300e-01  5.158e+04       0        1
## X25          1.087e+00  2.573e+04       0        1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 7.2997e+01  on 55  degrees of freedom
## Residual deviance: 9.6358e-10  on 30  degrees of freedom
```

```
## AIC: 52
##
## Number of Fisher Scoring iterations: 25
```

Here the inference does not work. we also get a warning, probably becouse we have to few samples. We plot the predictions

```
pred.c <- predict(modelglm.c, olitos.a[test.a,], type="response")
plot(pred.c, col=as.numeric(olitos.a[test.a,"y"]+1))
abline(h=0.5)
```



The confusion Matrix shows that we get worse results using all variables.

```
T <- table(olitos.a[test.a,"y"], pred.c>0.5)
T
```

```
##
##     FALSE TRUE
##   0    13    1
##   1     5    9
```

```
e2 <- 1-sum(diag(T))/sum(T)
e2
```

```
## [1] 0.2142857
```

2)

a)

We compute a model using all response variables and the explanatory variables *X1* to *X6*

```
#install.packages("VGAM")
```

```r
n <- nrow(olitos)
train <- sample(1:n, round(n*2/3))
test <- (1:n) [-train]



library(VGAM)
```

```
## Loading required package: stats4
```

```
## Loading required package: splines
```

```r
?vglm

modelvglm <- vglm(grp~X1+X2+X3+X4+X5+X6,data=olitos, family="multinomial", subset=train)
summary(modelvglm)
```

```
##
## Call:
## vglm(formula = grp ~ X1 + X2 + X3 + X4 + X5 + X6, family = "multinomial",
##     data = olitos, subset = train)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept):1 -2865.24    2346.98  -1.221    0.222
## (Intercept):2 -2498.06    2343.67      NA       NA
## (Intercept):3 -2634.70    2335.46  -1.128    0.259
## X1:1            -28.37      23.73      NA       NA
## X1:2            -32.51      23.76  -1.368    0.171
## X1:3            -30.55      23.75  -1.286    0.198
## X2:1             45.92      37.27      NA       NA
## X2:2             39.83      37.21   1.070    0.284
## X2:3             41.52      37.08      NA       NA
## X3:1            -84.80     162.95  -0.520    0.603
## X3:2            -13.18     161.67  -0.082    0.935
## X3:3             15.00     159.96   0.094    0.925
## X4:1          -2802.13    2606.56      NA       NA
## X4:2          -3057.51    2612.48  -1.170    0.242
## X4:3          -2879.86    2614.13  -1.102    0.271
## X5:1             12.70      11.28      NA       NA
## X5:2             12.68      11.28   1.123    0.261
## X5:3             13.82      11.28      NA       NA
## X6:1            -33.57      48.86  -0.687    0.492
## X6:2            -29.29      48.93  -0.599    0.549
## X6:3            -18.48      48.87  -0.378    0.705
##
## Names of linear predictors: log(mu[,1]/mu[,4]), log(mu[,2]/mu[,4]),
## log(mu[,3]/mu[,4])
##
## Residual deviance: 71.5228 on 219 degrees of freedom
##
## Log-likelihood: -35.7614 on 219 degrees of freedom
##
## Number of Fisher scoring iterations: 20
##
```

```
## Warning: Hauck-Donner effect detected in the following estimate(s):
## '(Intercept):2', 'X1:1', 'X2:1', 'X2:3', 'X4:1', 'X5:1', 'X5:3'
##
##
## Reference group is level  4  of the response
```

According to the inference table none of the variables is significatly contributig

b)

We compute the confusion matrix and calculate the missclassification rate

```r
pred.2 <- predict(modelvglm, olitos[test,], type="link")
#plot(pred.2, col=as.numeric(olitos[test,"grp"]))

T <- table(olitos[test,"grp"], apply(pred.2, 1, which.max))
T
```

```
##
##      1  2  3
##   1 17  1  1
##   2  4  4  0
##   3  0  0  8
##   4  2  1  2
```

```r
e1 <- 1-sum(diag(T))/sum(T)
e1
```

```
## [1] 0.275
```

3)

a)

We use the function *cv.glmnet()*

```r
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loading required package: foreach
```

```
## Loaded glmnet 2.0-16
```

```r
?cv.glmnet
X <- as.matrix(olitos[train, -26])
y <- as.numeric(olitos[train, "grp"])-1
modelvglm.3 <- cv.glmnet(x=X, y=y, family="multinomial")
#summary(modelvglm.3)
```

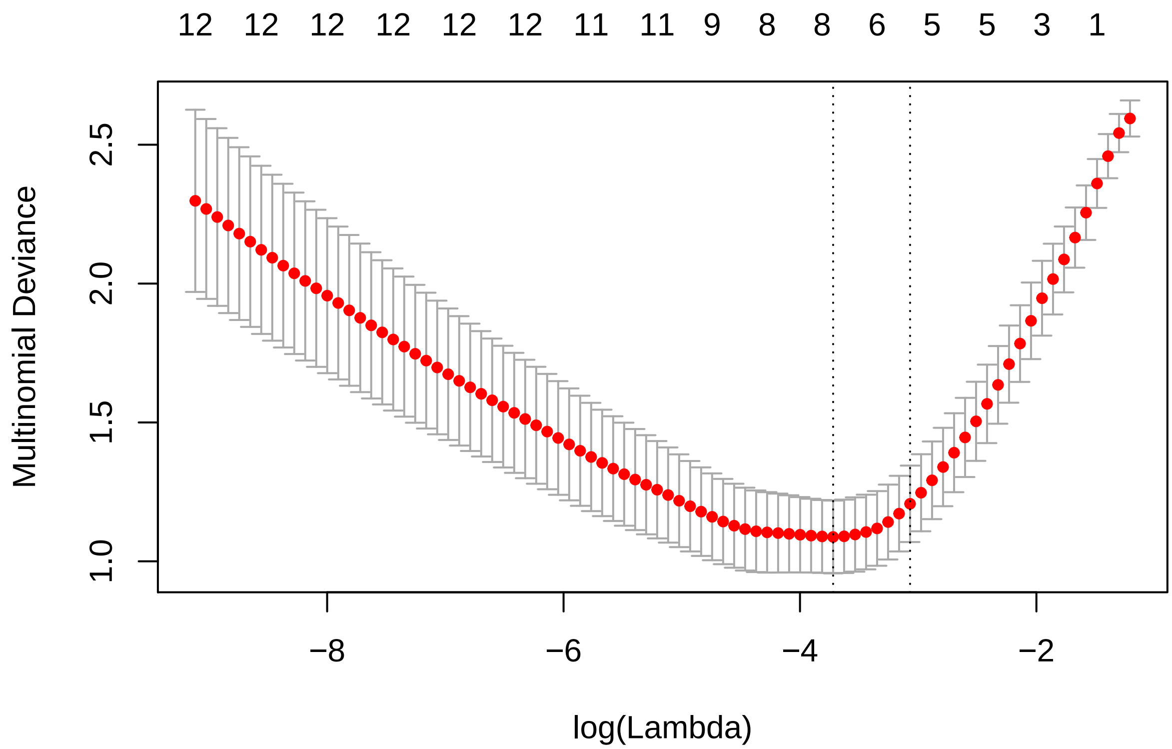
durring training we get the following warning:

Warning in lognet(x, is.sparse, ix, jx, y, weights, offset, alpha, nobs, :

one multinomial or binomial class has fewer than 8 observations; dangerous

ground

```
plot(modelvglm.3)
```



The plot shows us the optimal lambda parameter around log(-4)

b)

We use this lambda parameter to predict values of the test set anc compute our missclassification rate

```
modelvglm.3b <- glmnet(x=X, y=y, family="multinomial")
```

```
## Warning in lognet(x, is.sparse, ix, jx, y, weights, offset, alpha, nobs, :
## one multinomial or binomial class has fewer than 8 observations; dangerous
## ground
```

```
pred.3b <- predict(modelvglm.3b, as.matrix(olitos[test,-26]), type="link", s=0.0001)

T <- table(olitos[test,"grp"], apply(pred.3b, 1, which.max))
T
```

```
##
##      1  2  3  4
##   1 19  0  0  0
##   2  1  7  0  0
##   3  0  0  7  1
```

```
##    4  0  2  0  3
```

```
e1 <- 1-sum(diag(T))/sum(T)
e1
```

```
## [1] 0.1
```

4)

a)

We split or data in train and test sets

```
bank <- read.csv2("data/bank.csv")
set.seed(1234)
train <- sample(1:nrow(bank), 3000)
test <- (1:nrow(bank)) [-train]
```

train using *glm* on the train set

```
modelglm.4 <- glm(y~.,data=bank, family=binomial, subset=train)
modelglm.4
```

```
##
## Call:  glm(formula = y ~ ., family = binomial, data = bank, subset = train)
##
## Coefficients:
##        (Intercept)                  age      jobblue-collar
##         -2.161e+00           -5.405e-03          -6.862e-01
##    jobentrepreneur         jobhousemaid       jobmanagement
##         -4.969e-01           -2.754e-02          -1.277e-01
##         jobretired     jobself-employed         jobservices
##          2.013e-01           -3.237e-01          -3.851e-02
##         jobstudent        jobtechnician       jobunemployed
##          6.016e-01           -4.467e-01          -4.659e-01
##         jobunknown       maritalmarried       maritalsingle
##          7.470e-01           -2.286e-01          -1.243e-01
## educationsecondary    educationtertiary    educationunknown
##         -3.791e-02            2.514e-01          -8.069e-01
##         defaultyes              balance          housingyes
##          6.477e-01           -4.172e-06          -5.750e-01
##            loanyes     contacttelephone      contactunknown
##         -6.142e-01            5.099e-02          -1.373e+00
##                day             monthaug            monthdec
##          8.043e-03           -2.493e-01          -1.644e-01
##           monthfeb             monthjan            monthjul
##         -1.932e-02           -1.165e+00          -5.609e-01
##           monthjun             monthmar            monthmay
##          5.605e-01            1.337e+00          -5.715e-01
##           monthnov             monthoct            monthsep
##         -8.104e-01            1.536e+00           1.062e+00
##           duration             campaign               pdays
##          4.213e-03           -6.022e-02           9.916e-04
##           previous        poutcomeother     poutcomesuccess
##          1.055e-02            3.545e-01           1.847e+00
```

```
##    poutcomeunknown
##         -1.281e-01
##
## Degrees of Freedom: 2999 Total (i.e. Null);  2957 Residual
## Null Deviance:       2198
## Residual Deviance: 1469  AIC: 1555
```

```r
summary(modelglm.4)
```

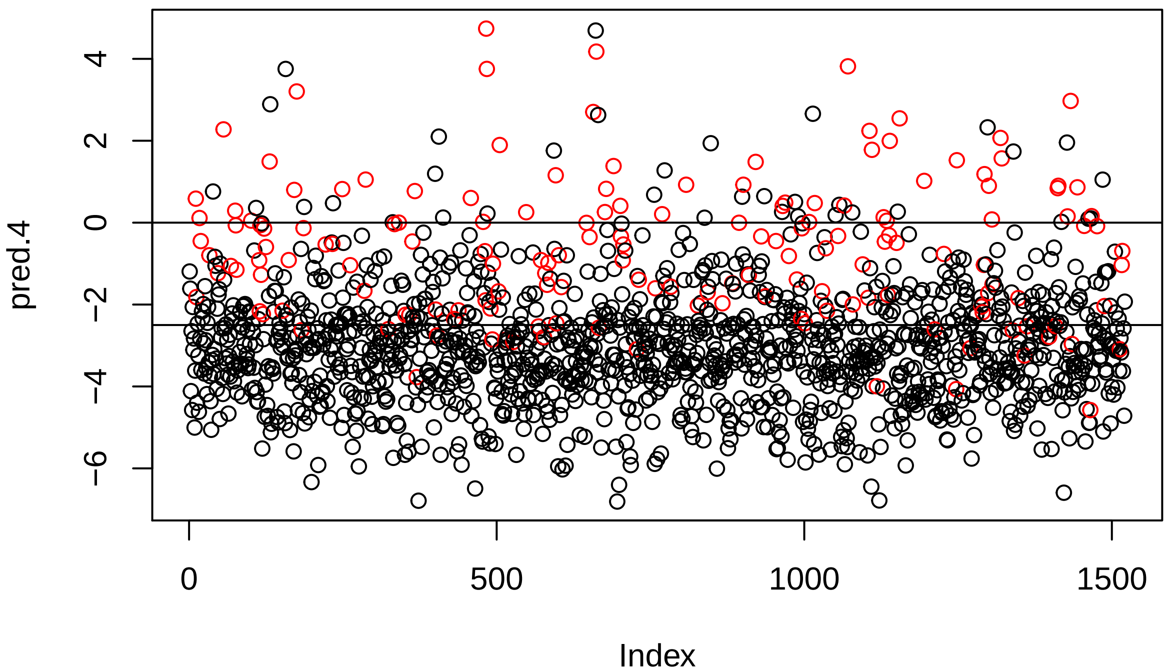
```
##
## Call:
## glm(formula = y ~ ., family = binomial, data = bank, subset = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0982  -0.3917  -0.2548  -0.1460   2.9657
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -2.161e+00  7.309e-01  -2.956 0.003112 **
## age                -5.405e-03  8.641e-03  -0.625 0.531651
## jobblue-collar     -6.862e-01  2.968e-01  -2.312 0.020763 *
## jobentrepreneur    -4.969e-01  4.580e-01  -1.085 0.278002
## jobhousemaid       -2.754e-02  4.784e-01  -0.058 0.954096
## jobmanagement      -1.277e-01  2.841e-01  -0.449 0.653221
## jobretired          2.013e-01  3.772e-01   0.534 0.593654
## jobself-employed   -3.237e-01  4.140e-01  -0.782 0.434237
## jobservices        -3.851e-02  3.184e-01  -0.121 0.903717
## jobstudent          6.016e-01  4.311e-01   1.395 0.162866
## jobtechnician      -4.467e-01  2.789e-01  -1.602 0.109203
## jobunemployed      -4.659e-01  5.064e-01  -0.920 0.357604
## jobunknown          7.470e-01  6.129e-01   1.219 0.222898
## maritalmarried     -2.286e-01  2.230e-01  -1.025 0.305278
## maritalsingle      -1.243e-01  2.568e-01  -0.484 0.628379
## educationsecondary -3.791e-02  2.467e-01  -0.154 0.877880
## educationtertiary   2.514e-01  2.809e-01   0.895 0.370787
## educationunknown   -8.069e-01  4.468e-01  -1.806 0.070946 .
## defaultyes          6.477e-01  5.007e-01   1.293 0.195849
## balance            -4.172e-06  1.965e-05  -0.212 0.831862
## housingyes         -5.750e-01  1.706e-01  -3.371 0.000749 ***
## loanyes            -6.142e-01  2.419e-01  -2.539 0.011105 *
## contacttelephone    5.099e-02  2.726e-01   0.187 0.851632
## contactunknown     -1.373e+00  2.832e-01  -4.849 1.24e-06 ***
## day                 8.043e-03  9.807e-03   0.820 0.412156
## monthaug           -2.493e-01  2.994e-01  -0.833 0.404988
## monthdec           -1.644e-01  7.561e-01  -0.217 0.827832
## monthfeb           -1.932e-02  3.685e-01  -0.052 0.958179
## monthjan           -1.165e+00  4.632e-01  -2.514 0.011933 *
## monthjul           -5.609e-01  3.012e-01  -1.863 0.062529 .
## monthjun            5.605e-01  3.660e-01   1.532 0.125627
## monthmar            1.337e+00  4.711e-01   2.838 0.004540 **
## monthmay           -5.715e-01  2.875e-01  -1.988 0.046806 *
## monthnov           -8.104e-01  3.378e-01  -2.399 0.016429 *
## monthoct            1.536e+00  3.911e-01   3.928 8.58e-05 ***
## monthsep            1.062e+00  4.796e-01   2.214 0.026830 *
```

```
## duration            4.213e-03  2.477e-04  17.009  < 2e-16 ***
## campaign           -6.022e-02  3.238e-02  -1.860 0.062930 .
## pdays               9.916e-04  1.136e-03   0.873 0.382820
## previous            1.055e-02  5.038e-02   0.209 0.834132
## poutcomeother       3.545e-01  3.223e-01   1.100 0.271252
## poutcomesuccess     1.847e+00  3.388e-01   5.453 4.95e-08 ***
## poutcomeunknown    -1.281e-01  3.944e-01  -0.325 0.745283
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2197.6  on 2999  degrees of freedom
## Residual deviance: 1469.3  on 2957  degrees of freedom
## AIC: 1555.3
##
## Number of Fisher Scoring iterations: 6
```

We see that some variables are more significant than others. *contactunknown*, *monthoct*, *duration* and *poutcomesuccess* contribute the most

We plot predictions on the test-set.

```r
pred.4 <- predict(modelglm.4, bank[test,], type="link")
plot(pred.4, col=as.numeric(bank[test, "y"]))
abline(h=0)
abline(h=-2.5)
```



We calculate the confusion table. To minimaze false negatives we shift the decision boundary to -2.5

```r
T <- table(bank[test,"y"], pred.4>0)
T
```

```
##
##        FALSE TRUE
```

```
##   no   1324   35
##   yes   108   54
```

```
T <- table(bank[test,"y"], pred.4>-2.5)
T
```

```
##
##        FALSE TRUE
##   no    957  402
##   yes    23  139
```

```
#modelglm.4b <- glm(y~.,data=bank, family=binomial, subset=train, weights = seq(1,16))
```