HW_Logistic Regression

경제학과 2020110210 공소연

2022-11-25

1

```
rm(list = ls())

## 1
library(MASS)
```

```
## Warning: 패키지 'MASS'는 R 버전 4.2.2에서 작성되었습니다
```

```
data <- biopsy
```

2

```
str(data)
```

```
## 'data.frame':    699 obs. of  11 variables:
##  $ ID   : chr  "1000025" "1002945" "1015425" "1016277" ...
##  $ V1   : int  5 5 3 6 4 8 1 2 2 4 ...
##  $ V2   : int  1 4 1 8 1 10 1 1 1 2 ...
##  $ V3   : int  1 4 1 8 1 10 1 2 1 1 ...
##  $ V4   : int  1 5 1 1 3 8 1 1 1 1 ...
##  $ V5   : int  2 7 2 3 2 7 2 2 2 2 ...
##  $ V6   : int  1 10 2 4 1 10 10 1 1 1 ...
##  $ V7   : int  3 3 3 3 3 9 3 3 1 2 ...
##  $ V8   : int  1 2 1 7 1 7 1 1 1 1 ...
##  $ V9   : int  1 1 1 1 1 1 1 1 5 1 ...
##  $ class: Factor w/ 2 levels "benign","malignant": 1 1 1 1 1 2 1 1 1 1 ...
```

```
summary(data)
```

```
##       ID                 V1               V2               V3
##  Length:699         Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  Class :character   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000
##  Mode  :character   Median : 4.000   Median : 1.000   Median : 1.000
##                     Mean   : 4.418   Mean   : 3.134   Mean   : 3.207
##                     3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000
##                     Max.   :10.000   Max.   :10.000   Max.   :10.000
##
##        V4               V5               V6               V7
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 2.000
##  Median : 1.000   Median : 2.000   Median : 1.000   Median : 3.000
##  Mean   : 2.807   Mean   : 3.216   Mean   : 3.545   Mean   : 3.438
##  3rd Qu.: 4.000   3rd Qu.: 4.000   3rd Qu.: 6.000   3rd Qu.: 5.000
##  Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000
##                                    NA's   :16
##        V8               V9               class
##  Min.   : 1.000   Min.   : 1.000   benign   :458
##  1st Qu.: 1.000   1st Qu.: 1.000   malignant:241
##  Median : 1.000   Median : 1.000
##  Mean   : 2.867   Mean   : 1.589
##  3rd Qu.: 4.000   3rd Qu.: 1.000
##  Max.   :10.000   Max.   :10.000
##
```

```
?biopsy
```

```
## httpd 도움말 서버를 시작합니다 ... 완료
```

3

```
data <- na.omit(data)
summary(data)
```

```
##       ID                 V1               V2               V3
##  Length:683         Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  Class :character   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000
##  Mode  :character   Median : 4.000   Median : 1.000   Median : 1.000
##                     Mean   : 4.442   Mean   : 3.151   Mean   : 3.215
##                     3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000
##                     Max.   :10.000   Max.   :10.000   Max.   :10.000
##        V4               V5               V6               V7
##  Min.   : 1.00   Min.   : 1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 1.00   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 2.000
##  Median : 1.00   Median : 2.000   Median : 1.000   Median : 3.000
##  Mean   : 2.83   Mean   : 3.234   Mean   : 3.545   Mean   : 3.445
##  3rd Qu.: 4.00   3rd Qu.: 4.000   3rd Qu.: 6.000   3rd Qu.: 5.000
##  Max.   :10.00   Max.   :10.000   Max.   :10.000   Max.   :10.000
##        V8               V9               class
##  Min.   : 1.00   Min.   : 1.000   benign   :444
##  1st Qu.: 1.00   1st Qu.: 1.000   malignant:239
##  Median : 1.00   Median : 1.000
##  Mean   : 2.87   Mean   : 1.603
##  3rd Qu.: 4.00   3rd Qu.: 1.000
##  Max.   :10.00   Max.   :10.000
```

4

```
data <- data[,-1]
str(data)
```

```
## 'data.frame':    683 obs. of  10 variables:
##  $ V1   : int  5 5 3 6 4 8 1 2 2 4 ...
##  $ V2   : int  1 4 1 8 1 10 1 1 1 2 ...
##  $ V3   : int  1 4 1 8 1 10 1 2 1 1 ...
##  $ V4   : int  1 5 1 1 3 8 1 1 1 1 ...
##  $ V5   : int  2 7 2 3 2 7 2 2 2 2 ...
##  $ V6   : int  1 10 2 4 1 10 10 1 1 1 ...
##  $ V7   : int  3 3 3 3 3 9 3 3 1 2 ...
##  $ V8   : int  1 2 1 7 1 7 1 1 1 1 ...
##  $ V9   : int  1 1 1 1 1 1 1 1 5 1 ...
##  $ class: Factor w/ 2 levels "benign","malignant": 1 1 1 1 1 2 1 1 1 1 ...
```

```
levels(data$class) <- c(0,1)
str(data)
```

```
## 'data.frame':    683 obs. of  10 variables:
##  $ V1   : int  5 5 3 6 4 8 1 2 2 4 ...
##  $ V2   : int  1 4 1 8 1 10 1 1 1 2 ...
##  $ V3   : int  1 4 1 8 1 10 1 2 1 1 ...
##  $ V4   : int  1 5 1 1 3 8 1 1 1 1 ...
##  $ V5   : int  2 7 2 3 2 7 2 2 2 2 ...
##  $ V6   : int  1 10 2 4 1 10 10 1 1 1 ...
##  $ V7   : int  3 3 3 3 3 9 3 3 1 2 ...
##  $ V8   : int  1 2 1 7 1 7 1 1 1 1 ...
##  $ V9   : int  1 1 1 1 1 1 1 1 5 1 ...
##  $ class: Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 1 1 ...
```

```
summary(data)
```

```
##        V1              V2              V3              V4
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.00
##  1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 1.00
##  Median : 4.000   Median : 1.000   Median : 1.000   Median : 1.00
##  Mean   : 4.442   Mean   : 3.151   Mean   : 3.215   Mean   : 2.83
##  3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 5.000   3rd Qu.: 4.00
##  Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.00
##        V5              V6              V7              V8
##  Min.   : 1.000   Min.   : 1.000   Min.   : 1.000   Min.   : 1.00
##  1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.: 1.00
##  Median : 2.000   Median : 1.000   Median : 3.000   Median : 1.00
##  Mean   : 3.234   Mean   : 3.545   Mean   : 3.445   Mean   : 2.87
##  3rd Qu.: 4.000   3rd Qu.: 6.000   3rd Qu.: 5.000   3rd Qu.: 4.00
##  Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.00
##        V9         class
##  Min.   : 1.000   0:444
##  1st Qu.: 1.000   1:239
##  Median : 1.000
##  Mean   : 1.603
##  3rd Qu.: 1.000
##  Max.   :10.000
```

```
head(data[10], 10)
```

```
##    class
## 1      0
## 2      0
## 3      0
## 4      0
## 5      0
## 6      1
## 7      0
## 8      0
## 9      0
## 10     0
```

```
set.seed(1)
library(caret)
```

```
## 필요한 패키지를 로딩중입니다: ggplot2
```

```
## 필요한 패키지를 로딩중입니다: lattice
```

```
parts <- createDataPartition(data$class, p=0.7, list=F)
training <- data[parts,]
testing <- data[-parts,]
glm <- glm(class ~ .,
           data = training,
           family = binomial)
summary(glm)
```

```
##
## Call:
## glm(formula = class ~ ., family = binomial, data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2439  -0.1242  -0.0687   0.0217   2.0597
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.84556    1.32967  -7.405 1.32e-13 ***
## V1           0.46854    0.15925   2.942  0.00326 **
## V2           0.16288    0.24086   0.676  0.49888
## V3           0.33456    0.26272   1.273  0.20286
## V4           0.33811    0.13491   2.506  0.01220 *
## V5           0.09569    0.18171   0.527  0.59846
## V6           0.30553    0.10834   2.820  0.00480 **
## V7           0.45154    0.19394   2.328  0.01990 *
## V8           0.11815    0.14058   0.840  0.40065
## V9           0.52288    0.34885   1.499  0.13391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 620.686  on 478  degrees of freedom
## Residual deviance:  78.758  on 469  degrees of freedom
## AIC: 98.758
##
## Number of Fisher Scoring iterations: 8
```

```
head(fitted(glm),10)
```

```
##           1           2           3           4           6          13
## 0.015145977 0.882511709 0.008111754 0.817135456 0.999977924 0.252260684
##          14          15          16          17
## 0.004330061 0.999787519 0.732767256 0.006091007
```

6

```
pred <- predict(glm, newdata = testing, type = "response")
yi <- as.numeric(testing$class==1)
data.frame(pred, yi)
```

```
##             pred yi
## 5   0.0185772310  0
## 7   0.0356019191  0
## 8   0.0052419489  0
## 9   0.0122255195  0
## 10  0.0071607707  0
## 11  0.0021404550  0
## 12  0.0023951706  0
## 25  0.0023548993  0
## 27  0.0040856480  0
## 30  0.0018645668  0
## 33  0.9989218672  1
## 42  0.9154862414  1
## 43  0.9994587134  1
## 45  0.9994764965  1
## 49  0.0185772310  0
## 51  0.9499721864  1
## 55  0.9982861418  1
## 56  0.9656309024  1
## 64  0.3017558925  1
## 65  0.0015005257  0
## 68  0.8670816884  1
## 70  0.0026494582  0
## 75  0.8968032688  1
## 86  0.9989594136  1
## 89  0.0095341646  0
## 90  0.0036912400  0
## 95  0.0037570228  0
## 97  0.0021388888  0
## 98  0.0151459774  0
## 108 0.9938576621  1
## 109 0.0018997468  0
## 111 0.0317612650  0
## 113 0.9996790257  1
## 115 0.0210322196  0
## 120 0.0095330178  0
## 126 0.0015005257  0
## 128 0.0059889689  0
## 131 0.0187583352  0
## 133 0.9995177574  1
## 139 0.0084905556  0
## 147 0.5065511640  1
## 156 0.9130181082  1
## 158 0.0037570228  0
## 160 0.9999877034  1
## 162 0.0107170839  0
## 166 0.0144913144  0
## 167 0.9984170506  1
## 171 0.0024361289  0
## 172 0.0023548993  0
## 173 0.0015005257  0
## 174 0.9999994952  1
## 175 0.9900277913  1
## 180 0.6522445488  1
## 189 0.9992617736  1
```

```
## 191 0.9999696505  1
## 198 0.0406827459  0
## 205 0.0023548993  0
## 208 0.0021404550  0
## 217 0.0015005257  0
## 218 0.0023548993  0
## 219 0.9990317541  1
## 222 0.9996475252  1
## 223 0.0358408379  1
## 226 0.0015005257  0
## 232 0.9991255539  1
## 234 0.9921211637  1
## 237 0.9999875848  1
## 239 0.9999968588  1
## 252 0.9995185328  1
## 254 0.9994587134  1
## 256 0.8986667502  1
## 257 0.0024361289  0
## 258 0.0038211800  0
## 265 0.9965770258  1
## 272 0.0151459774  0
## 277 0.0038211800  0
## 279 0.0023548993  0
## 281 0.0059889689  0
## 284 0.9854932604  1
## 289 0.5848857107  1
## 291 0.0009558278  0
## 300 0.9966772675  1
## 302 0.0023548993  0
## 308 0.0023548993  0
## 310 0.0480339628  0
## 314 0.0009558278  0
## 317 0.8673621254  1
## 327 0.8980045983  1
## 332 0.0252863167  0
## 335 0.9884751044  1
## 336 0.0009558278  0
## 337 0.9810342033  1
## 346 0.0009558278  0
## 348 0.0010999763  0
## 351 0.0121091815  0
## 353 0.3829874262  0
## 355 0.0015005257  0
## 360 0.9865271037  1
## 366 0.0023951706  0
## 367 0.9999980438  1
## 368 0.9999419436  1
## 370 0.0026593351  0
## 371 0.0128837952  0
## 375 0.0053313279  0
## 376 0.0009558278  0
## 380 0.0657470209  0
## 381 0.0009558278  0
## 386 0.0228236705  0
## 390 0.0161919414  0
## 391 0.0021029174  0
```

```
## 393 0.0038211800  0
## 394 0.0008686762  0
## 398 0.0038864286  0
## 401 0.9999612140  1
## 406 0.0015005257  0
## 407 0.0099775211  0
## 410 0.0053313279  0
## 411 0.0015005257  0
## 420 0.0028132556  0
## 424 0.0187583352  0
## 425 0.0024361289  0
## 426 0.9999976447  1
## 432 0.0395821617  0
## 435 0.9811485195  0
## 436 0.9976888018  1
## 437 0.9678037074  1
## 447 0.0009558278  0
## 453 0.0047792601  0
## 457 0.9998108317  1
## 461 0.0121091815  0
## 463 0.0192072584  0
## 466 0.9998833497  1
## 467 0.9993594594  1
## 468 0.9975135395  1
## 474 0.0038864286  0
## 478 0.0038864286  0
## 481 0.0056326325  0
## 484 0.9998354245  1
## 486 0.0031400288  0
## 491 0.0009558278  0
## 492 0.9950718841  1
## 494 0.9998534347  1
## 496 0.0034736766  0
## 504 0.0095341646  0
## 514 0.0034736766  0
## 523 0.9874996428  1
## 526 0.0034128381  0
## 529 0.0265434505  0
## 531 0.9848109694  1
## 534 0.0038211800  0
## 536 0.0064214462  0
## 537 0.0151459774  0
## 540 0.0154016129  0
## 545 0.0065308045  0
## 547 0.9999339123  1
## 550 0.9943469821  1
## 554 0.0740128504  0
## 555 0.0024361289  0
## 556 0.0456979971  0
## 563 0.0023548993  0
## 569 0.9168034850  1
## 570 0.9999555348  1
## 573 0.0038211800  0
## 576 0.0210372216  0
## 577 0.0096960158  0
## 579 0.0015005257  0
```

```
## 580 0.0023548993  0
## 584 0.0024361289  0
## 585 0.0358614172  0
## 587 0.9999944823  1
## 589 0.9951876592  1
## 590 0.0061947732  0
## 591 0.9795508978  1
## 605 0.9907129651  1
## 606 0.9998180293  1
## 609 0.9999555605  1
## 610 0.0061947732  0
## 611 0.9994705917  1
## 614 0.0033144598  0
## 616 0.0118241586  0
## 624 0.0009558278  0
## 627 0.9915350300  1
## 628 0.0051617212  0
## 630 0.0038864286  0
## 639 0.0038864286  0
## 640 0.0121091815  0
## 641 0.0076138487  0
## 642 0.0038211800  0
## 643 0.0038211800  0
## 644 0.0009558278  0
## 646 0.0038211800  0
## 652 0.0034658348  0
## 658 0.2345216016  0
## 666 0.0009558278  0
## 668 0.0059889689  0
## 678 0.0061947732  0
## 679 0.0009558278  0
## 681 0.9999998968  1
## 683 0.0170132037  0
## 687 0.0009558278  0
## 689 0.0038864286  0
## 693 0.0024361289  0
## 695 0.0036343694  0
## 699 0.9940593698  1
```

```
sqrt(sum((yi-pred)^2)/204) # RMSE
```

```
## [1] 0.1260875
```

```
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
accuracy(pred, yi)
```

```
##                  ME      RMSE        MAE  MPE MAPE
## Test set 0.00639005 0.1260875 0.03338298 -Inf  Inf
```

```
confusionMatrix(as.factor(ifelse(pred>0.5,'1','0')),
                testing$class) # cutoff=0.5, accuracy=0.985
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0 132   2
##          1   1  69
##
##                Accuracy : 0.9853
##                  95% CI : (0.9576, 0.997)
##     No Information Rate : 0.652
##     P-Value [Acc > NIR] : <2e-16
##
##                   Kappa : 0.9675
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.9925
##             Specificity : 0.9718
##          Pos Pred Value : 0.9851
##          Neg Pred Value : 0.9857
##              Prevalence : 0.6520
##          Detection Rate : 0.6471
##    Detection Prevalence : 0.6569
##       Balanced Accuracy : 0.9822
##
##        'Positive' Class : 0
##
```