# Tsolakidis_6

2024-04-17

## Εργασία 6 - Τσολακιδης Βασιλειος

Να δημιουργήσετε ένα R markdown αρχείο (.rmd) με ένα case study Λογιστικής Παλινδρόμησης. Θα χρησιμοποιήσετε τα δεδομένα του Framingham Heart Study για να δημιουργήσετε ένα Μοντέλο Λογιστικής Παλινδρόμησης το οποίο θα δίνει προβλέψεις για την εξαρτημένη μεταβλητή TenYearCHD, δηλαδή την πιθανότητα εμφάνισης στεφανιαίας νόσου στους ασθενείς της βάσης για την επόμενη δεκαετία.

```
# Εισαγωγή των απαιτούμενων βιβλιοθηκών
library(caTools)
library(ROCR)
```

## Φορτωμα δεδομένων του Framingham Heart Study και χωρισμος της βάση σε training και testing sets με τυχαίο τρόπο. Το training set θα είναι το 65% της βάσης και θα ορίσουμε το seed σε 971.

```
# Φόρτωση των δεδομένων
data <- read.csv("framingham.csv")

# Ορισμός του seed
set.seed(971)

# Διαχωρισμός της βάσης σε training και testing sets
split <- sample.split(data$TenYearCHD, SplitRatio = 0.65)
train <- subset(data, split == TRUE)
test <- subset(data, split == FALSE)

# Για να εκκαθαρίσουμε τις εγγραφές που περιέχουν κελιά με τιμές NA.
# cleaned_train <- train[complete.cases(train), ]
# cleaned_test <- test[complete.cases(test), ]
# Εμφάνιση του αριθμού των εγγραφών μετά την εκκαθάριση
# cat("Αριθμός καταχωρήσεων στο cleaned training set:", nrow(cleaned_train),
"\n")
#cat("Αριθμός καταχωρήσεων στο cleaned test set:", nrow(cleaned_test), "\n")

# Αριθμός καταχωρήσεων σε κάθε set
cat("Αριθμός καταχωρήσεων στο training set:", nrow(train), "\n")

## Αριθμός καταχωρήσεων στο training set: 2756

cat("Αριθμός καταχωρήσεων στο test set:", nrow(test), "\n")

## Αριθμός καταχωρήσεων στο test set: 1484
```

## Δημιουργια τΜοντέλου Λογιστικής Παλινδρόμησης στο training set και αναλυση των συσχετίσεων των ανεξάρτητων μεταβλητών με την εξαρτημένη μεταβλητή TenYearCHD.

```
# Μοντέλο Λογιστικής Παλινδρόμησης
framinghamLog <- glm(TenYearCHD ~ ., data = train, family = binomial)

# Εμφάνιση των συντελεστών (coefficients) του μοντέλου
summary(framinghamLog)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ ., family = binomial, data = train)
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -7.710026   0.870156  -8.861  < 2e-16 ***
## male             0.543274   0.133725   4.063 4.85e-05 ***
## age              0.059171   0.008137   7.272 3.55e-13 ***
## education       -0.050854   0.060348  -0.843 0.399414
## currentSmoker    0.224119   0.192240   1.166 0.243683
## cigsPerDay       0.012071   0.007757   1.556 0.119704
## BPMeds           0.189399   0.278797   0.679 0.496920
## prevalentStroke  0.173518   0.731660   0.237 0.812536
## prevalentHyp     0.222979   0.172137   1.295 0.195196
## diabetes         0.068638   0.384132   0.179 0.858187
## totChol          0.001411   0.001384   1.019 0.307998
## sysBP            0.015979   0.004658   3.431 0.000602 ***
## diaBP           -0.008711   0.007719  -1.129 0.259108
## BMI              0.013347   0.015911   0.839 0.401528
## heartRate       -0.003626   0.005149  -0.704 0.481317
## glucose          0.007232   0.002792   2.590 0.009584 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2018.4  on 2377  degrees of freedom
## Residual deviance: 1805.6  on 2362  degrees of freedom
##   (378 observations deleted due to missingness)
## AIC: 1837.6
##
## Number of Fisher Scoring iterations: 5
```

### Σχολιασμος

Από τα αποτελέσματα του μοντέλου μπορούμε να δούμε τους συντελεστές για κάθε ανεξάρτητη μεταβλητή. Οι μεταβλητές με ισχυρή συσχέτιση με την εξαρτημένη μεταβλητή TenYearCHD και με σημαντικότητα (significance) θα έχουν μεγάλο απόλυτο τιμή του συντελεστή (coefficient) και μικρή p-value.

## Βελτιστο Μοντέλο Λογιστικής Παλινδρόμησης

```
# ------------ Βελτιστο Μοντέλο Λογιστικής Παλινδρόμησης ------------
# Επιλογή μόνο των βέλτιστων μεταβλητών
selected_variables <- c("age", "male", "BMI", "glucose", "diabetes",
"TenYearCHD")

# Μοντέλο Λογιστικής Παλινδρόμησης με τις επιλεγμένες μεταβλητές
framinghamLogOptimal <- glm(TenYearCHD ~ ., data =
train[,selected_variables], family = binomial)

# Εμφάνιση των συντελεστών (coefficients) του νέου μοντέλου
summary(framinghamLogOptimal)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ ., family = binomial, data = train[,
##     selected_variables])
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.114378   0.552495 -12.877  < 2e-16 ***
## age          0.070966   0.007011  10.122  < 2e-16 ***
## male         0.577230   0.116728   4.945 7.61e-07 ***
## BMI          0.030018   0.014154   2.121  0.03394 *
## glucose      0.008127   0.002649   3.068  0.00216 **
## diabetes     0.147935   0.371130   0.399  0.69018
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2124.4  on 2490  degrees of freedom
## Residual deviance: 1951.6  on 2485  degrees of freedom
##   (265 observations deleted due to missingness)
## AIC: 1963.6
##
## Number of Fisher Scoring iterations: 5
```

### Σχολιασμος

Το βέλτιστο μοντέλο Λογιστικής Παλινδρόμησης που δημιουργήσαμε βασίζεται σε μια πιο απλοποιημένη εκδοχή του αρχικού μοντέλου, περιλαμβάνοντας μόνο τις πιο σημαντικές μεταβλητές που επηρεάζουν την εξαρτημένη μεταβλητή TenYearCHD, όπως ο ηλικία, το φύλο, το Δείκτης Μάζας Σώματος (BMI), η γλυκόζη και η διαβήτης.Η αφαίρεση μη σημαντικών μεταβλητών από το μοντέλο συμβάλλει στη μείωση της πολυπλοκότητας του μοντέλου και τη βελτίωση της απόδοσής του, ενώ ταυτόχρονα μειώνει τον κίνδυνο υπερ-προσαρμογής (overfitting).

**Προβλέψεις στο test set χρησιμοποιώντας το μοντέλο που δημιουργήσαμε και να εξετάσουμε τι ακριβώς μας δείχνει η εντολή predict.**

```
# Προβλέψεις στο test set
predictTest <- predict(framinghamLog, newdata = test, type = "response")

# Εμφάνιση των πρώτων 10 προβλέψεων
head(predictTest)
```

```
##          8         11         12         14         17         19
## 0.06324532 0.08338104 0.04551406 0.07234750 0.17142047 0.03962362
```