Εργασία 6

2024-04-13

Διαχωρισμός dataset

```
library(caTools)
library(readr)
framingham <- read_csv("C:/Users/UserA/Downloads//framingham.csv",show_col_types = FALSE)
set.seed(925)
split <- sample.split(framingham$TenYearCHD,SplitRatio=0.65)
train <- subset(framingham,split==TRUE)
test <- subset(framingham,split==FALSE)
```

Καταχωρίσεις training set:

```
nrow(train)
```

```
## [1] 2756
```

Καταχωρίσεις test set:

```
nrow(test)
```

```
## [1] 1484
```

Δημιουργία μοντέλου λογιστικής παλινδρόμησης

```
framinghamLog <- glm(TenYearCHD ~ ., data=framingham, family=binomial)
```

Συσχετίσεις μεταβλητών:

```
summary(framinghamLog)
```

```
##
## Call:
## glm(formula = TenYearCHD ~ ., family = binomial, data = framingham)
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -8.328186   0.715449 -11.641  < 2e-16 ***
## male             0.555279   0.109033   5.093 3.53e-07 ***
```

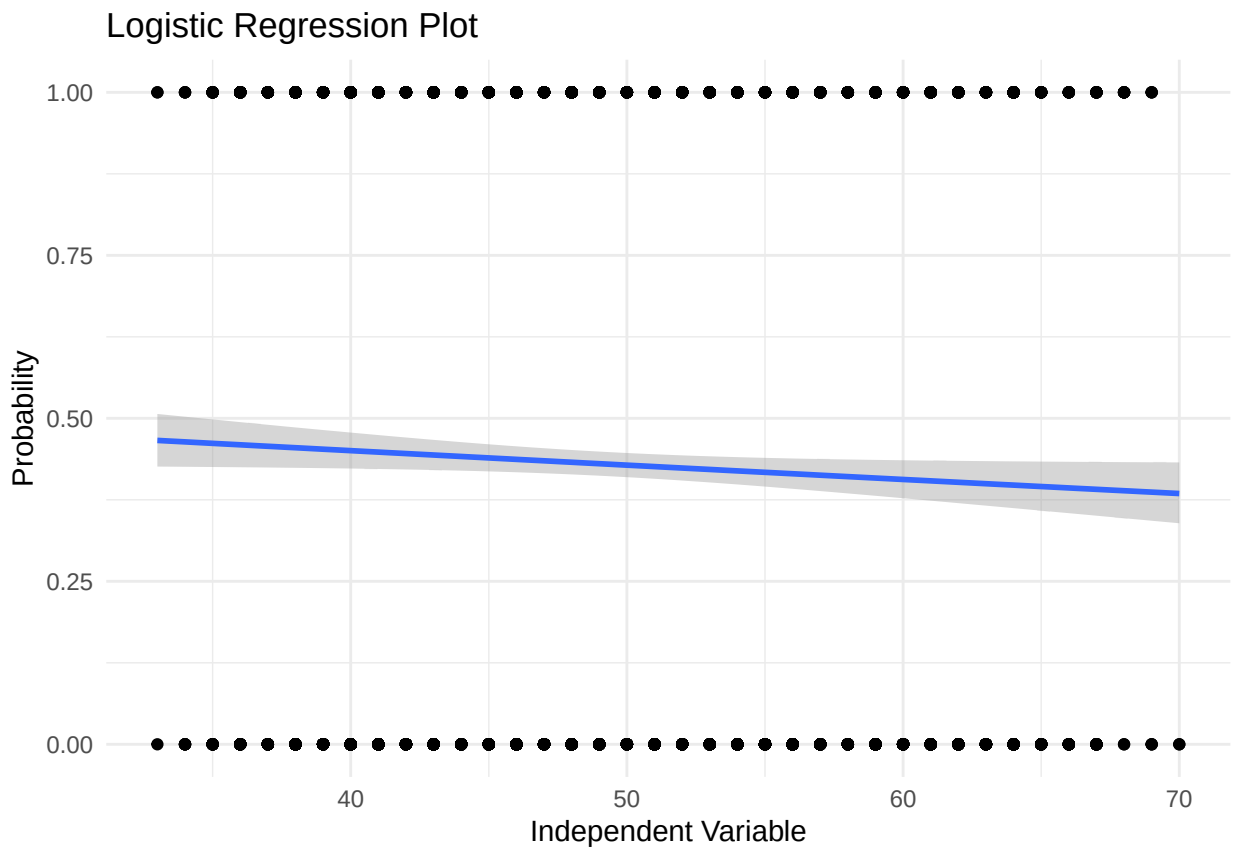
```
## age              0.063515   0.006679   9.509  < 2e-16 ***
## education       -0.047767   0.049395  -0.967  0.33353
## currentSmoker    0.071601   0.156752   0.457  0.64783
## cigsPerDay       0.017914   0.006238   2.872  0.00408 **
## BPMeds           0.162496   0.234326   0.693  0.48802
## prevalentStroke  0.693660   0.489569   1.417  0.15652
## prevalentHyp     0.234208   0.138026   1.697  0.08973 .
## diabetes         0.039167   0.315506   0.124  0.90120
## totChol          0.002332   0.001127   2.070  0.03850 *
## sysBP            0.015403   0.003808   4.044 5.24e-05 ***
## diaBP           -0.004159   0.006438  -0.646  0.51831
## BMI              0.006672   0.012758   0.523  0.60097
## heartRate       -0.003246   0.004211  -0.771  0.44082
## glucose          0.007127   0.002234   3.190  0.00142 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3121.2  on 3657  degrees of freedom
## Residual deviance: 2754.5  on 3642  degrees of freedom
##   (582 observations deleted due to missingness)
## AIC: 2786.5
##
## Number of Fisher Scoring iterations: 5
```

Παρατηρούμε ότι υπάρχουν έξι ανεξάρτητες μεταβλητές οι οποίες έχουν από τουλάχιστον ένα αστέρι, το οποίο σημαίνει ότι είναι σημαντικές για το μοντέλο μας. Αυτές είναι με ένα αστέρι: totChol, με δύο αστέρια: glucose και cigsPerDay, με τρία αστέρια: sysBP, male και age.

Διάγραμμα

```
library(ggplot2)
ggplot(train, aes(age, male)) +
         geom_point() +
         geom_smooth(method = "glm", method.args = list(family = "binomial")) +
         labs(title = "Logistic Regression Plot",
                        x = "Independent Variable",
                        y = "Probability") +
         theme_minimal()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Προβλέψεις στο test set

```
predictTest <- predict(framinghamLog, type='response', newdata=test)
```

Το predict μας δείχνει την πιθανότητα εμφάνισης στεφανιαίας νόσου στους ασθενείς την επόμενη δεκαετία.