ΣΤΑΤΙΣΤΙΚΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ

ΕΡΓΑΣΙΑ 3

ΦΟΙΤΗΤΗΣ: Προμπονάς Αντώνης

Α.Μ. : 03400232

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ: Ε.ΔΕ.Μ.Μ.

ΗΜΕΡΟΜΗΝΙΑ: Φεβρουάριος 2024

ΑΣΚΗΣΗ 1

Στην συγκεκριμένη άσκηση, κάνοντας χρήση ενός μοντέλου παλινδρόμησης Poisson, θα εξετάσουμε:

Την εξάρτηση του αριθμού Y αποζημιώσεων λόγω τροχαίων ατυχημάτων ανά n συμβόλαια, από την ηλικία του ασφαλισμένου (agecat - X1=0-νέος, 1-μεγάλος), την κατηγορία ασφαλίστρων (cartype - X2= 1,2,3,4) και την περιοχή διαμονής του ασφαλισμένου (district X3=1, αν Αθήνα, X3=0, αν σε άλλη πόλη).

Αρχικά και ύστερα από υπόδειξη της άσκησης, αρχικοποιώ τη κατηγορική μεταβλητή X2 , X2<-factor(cartype).

Έπειτα, δημιουργώ το μοντέλο παλινδρόμησης Poisson και βρίσκω τη σύνοψη του με τη χρήση των εντολών:

i. model<-glm(formula = y ~ X2 + agecat + district + offset(log(n)), family = poisson)
ii. summary(model)

```
> X2<-factor(cartype)
> model<-glm(formula = y ~ X2 + agecat + district + offset(log(n)), family = poisson)
> model

Call:  glm(formula = y ~ X2 + agecat + district + offset(log(n)), family = poisson)

Coefficients:
(Intercept)          X22          X23          X24       agecat     district
    -1.9352       0.1622       0.3953       0.5654      -0.3763       0.2166

Degrees of Freedom: 31 Total (i.e. Null);  26 Residual
Null Deviance:      207.8
Residual Deviance: 41.79        AIC: 222.1
> summary(model)

Call:
glm(formula = y ~ X2 + agecat + district + offset(log(n)), family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8590  -0.7506  -0.1297   0.6511   3.2310

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.93522    0.05525 -35.030  < 2e-16 ***
X22          0.16223    0.05048   3.214 0.001309 **
X23          0.39535    0.05491   7.200 6.03e-13 ***
X24          0.56543    0.07215   7.836 4.64e-15 ***
agecat      -0.37628    0.04451  -8.453  < 2e-16 ***
district     0.21661    0.05853   3.701 0.000215 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 207.833  on 31  degrees of freedom
Residual deviance:  41.789  on 26  degrees of freedom
AIC: 222.15

Number of Fisher Scoring iterations: 4
```

Το μοντέλο είναι exp(-1.93522+0.16223*X2_2+0.39535* X2_3 +0.56543* X2_4 -0.37628* agecat + 0.21661* district)=

exp(-1.93522+0.16223*β1+0.39535*β2 +0.56543*β3 -0.37628*β4 + 0.21661*β5)

Αν αυξηθεί η συμμεταβλητή agecat κατά μια μονάδα (δηλαδή για έναν ασφαλισμένο) θα πολλαπλασιαστεί ο αριθμός Y αποζημιώσεων λόγω τροχαίων ατυχημάτων κατά exp(-0.37628)= 0.6864101, δηλαδή θα μειωθεί η ηλικία αυτών που ενεπλάκησαν σε τροχαίο κατά 32%.

Αν αυξηθεί η συμμεταβλητή district κατά μια μονάδα (δηλαδή αν η πόλη είναι η Αθήνα) θα πολλαπλασιαστεί ο αριθμός Y των αποζημιώσεων λόγω τροχαίων κατά exp(0.21661)= 1.24, δηλαδή θα αυξηθεί ο αριθμός κατά 24%.

Χρησιμοποιώντας την εντολή factor(cartype) έχω μετατρέψει την μεταβλητή cartype από ποσοτική σε κατηγορική . Ως κατηγορική η cartype έχει 4 κατηγορίες. Οι κατηγορίες αυτές αντιπροσωπεύουν η κάθεμία το τύπο των ασφαλίστρων. Η κατηγορία 1 θεωρείται η σταθερά.

Αν αυξηθεί η κατηγορία σε 2 θα πολλαπλασιαστεί ο αριθμός Y των αποζημιώσεων λόγω τροχαίων κατά exp(0.16223)= 1.18.

Αν αυξηθεί η κατηγορία σε 3 θα πολλαπλασιαστεί ο αριθμός Y των αποζημιώσεων λόγω τροχαίων κατά exp(0.39535)= 1.48.

Αν αυξηθεί η κατηγορία σε 4 θα πολλαπλασιαστεί ο αριθμός Y των αποζημιώσεων λόγω τροχαίων κατά exp(0.56543)= 1.76.

Αφού δημιούργησα το κατάλληλο μοντέλο, πραγματοποιώ τους στατιστικούς ελέγχους (Wald και Deviance), κάνοντας χρήση και του κριτηρίου AIC.

Ο κώδικας που χρησιμοποίησα και τα αποτελέσματα του καταγράφονται παρακάτω:

i. wald_test <- summary(model)$coefficients[, "Pr(>|z|)"]
ii. deviance_test1 <- anova(model, test = "Chisq")
iii. deviance_test2 <- 1 - pchisq(model$deviance, df = model$df.residual)
iv. aic <- AIC(model)

```
> wald_test <- summary(model)$coefficients[, "Pr(>|z|)"]
> wald_test
  (Intercept)           X22           X23           X24        agecat 
7.982640e-269  1.309304e-03  6.029896e-13  4.636754e-15  2.842310e-17 
     district 
 2.148209e-04 
> deviance_test1 <- anova(model, test = "Chisq")
> deviance_test1
Analysis of Deviance Table

Model: poisson, link: log

Response: y

Terms added sequentially (first to last)


         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                        31    207.833              
X2        3   88.348        28    119.485 < 2.2e-16 ***
agecat    1   64.759        27     54.727 8.466e-16 ***
district  1   12.938        26     41.789  0.000322 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> deviance_test2 <- 1 - pchisq(model$deviance, df = model$df.residual)
> deviance_test2
[1] 0.02580847
> result <- data.frame(Wald_Test = wald_test, Deviance_Test = deviance_test)
> print(result)
                Wald_Test Deviance_Test
(Intercept) 7.982640e-269    0.04742357
X22          1.309304e-03    0.04742357
X23          6.029896e-13    0.04742357
X24          4.636754e-15    0.04742357
agecat       2.842310e-17    0.04742357
district     2.148209e-04    0.04742357
> aic <- AIC(model)
> cat("AIC:", aic, "\n")
AIC: 222.1488
```

‘Ολα τα παραπάνω συμπυκνώνονται μέσω της εντολής step(model,method="backward", test="Chisq") .

```
> step(model,method="backward", test="Chisq")
Start:  AIC=222.15
y ~ X2 + agecat + district + offset(log(n))

           Df Deviance    AIC    LRT  Pr(>Chi)    
<none>          41.789 222.15                     
- district  1   54.727 233.09 12.938  0.000322 ***
- agecat    1  107.964 286.32 66.176 4.125e-16 ***
- X2        3  131.713 306.07 89.925 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Call:  glm(formula = y ~ X2 + agecat + district + offset(log(n)), family = poisson)

Coefficients:
(Intercept)          X22          X23          X24       agecat     district  
    -1.9352       0.1622       0.3953       0.5654      -0.3763       0.2166  

Degrees of Freedom: 31 Total (i.e. Null);  26 Residual
Null Deviance:      207.8 
Residual Deviance: 41.79        AIC: 222.1
> logLik(model)
'log Lik.' -105.0744 (df=6)
```

Για κάθε έναν συντελεστή του μοντέλου, δημιουργώ διάστημα εμπιστοσύνης.

Οι εντολές που χρησιμοποιώ είναι οι: confint.default(model) και exp(confint.default(model)) και έχουν τα εξής αποτελέσματα:

```
> confint.default(model)
                  2.5 %     97.5 %
(Intercept) -2.04350208 -1.8269440
X22          0.06329746  0.2611664
X23          0.28772397  0.5029705
X24          0.42400923  0.7068487
agecat      -0.46352606 -0.2890309
district     0.10189607  0.3313250
> exp(confint.default(model))
                2.5 %    97.5 %
(Intercept) 0.1295741 0.1609045
X22         1.0653437 1.2984438
X23         1.3333892 1.6536260
X24         1.5280757 2.0275915
agecat      0.6290616 0.7489890
district    1.1072684 1.3928124
```

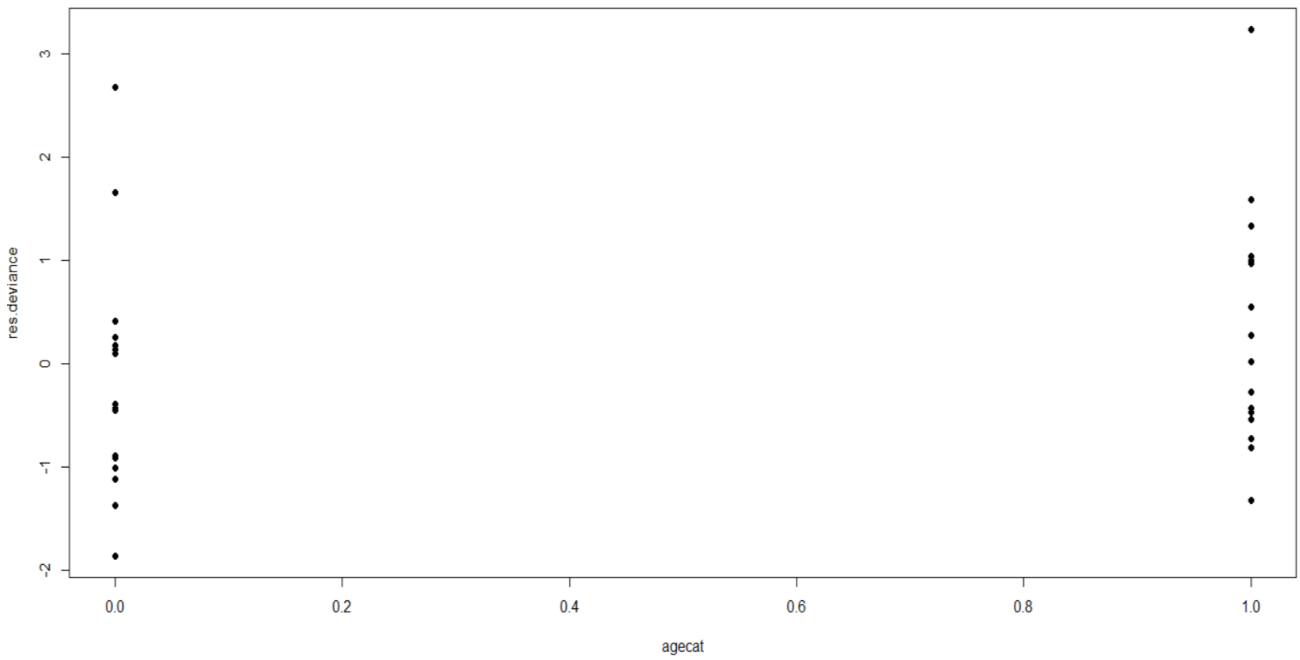
Μέσα από τους παραπάνω ελέγχους και δη τον έλεγχο Deviance, βλέπουμε ότι όλες οι μεταβλητές είναι στατιστικά σημαντικές. Το μοντέλο έχει AIC=222. Το AIC αντιπροσωπεύει το ποσοστό απώλειας πληροφορίας και η συγκεκριμένη τιμή είναι σχετικά καλή και δείχνει μία καλή προσαρμοστικότητα του μοντέλου με την πραγματικότητα. Το μοντέλο δηλαδή είναι αρκετά έγκυρο και ακριβές.

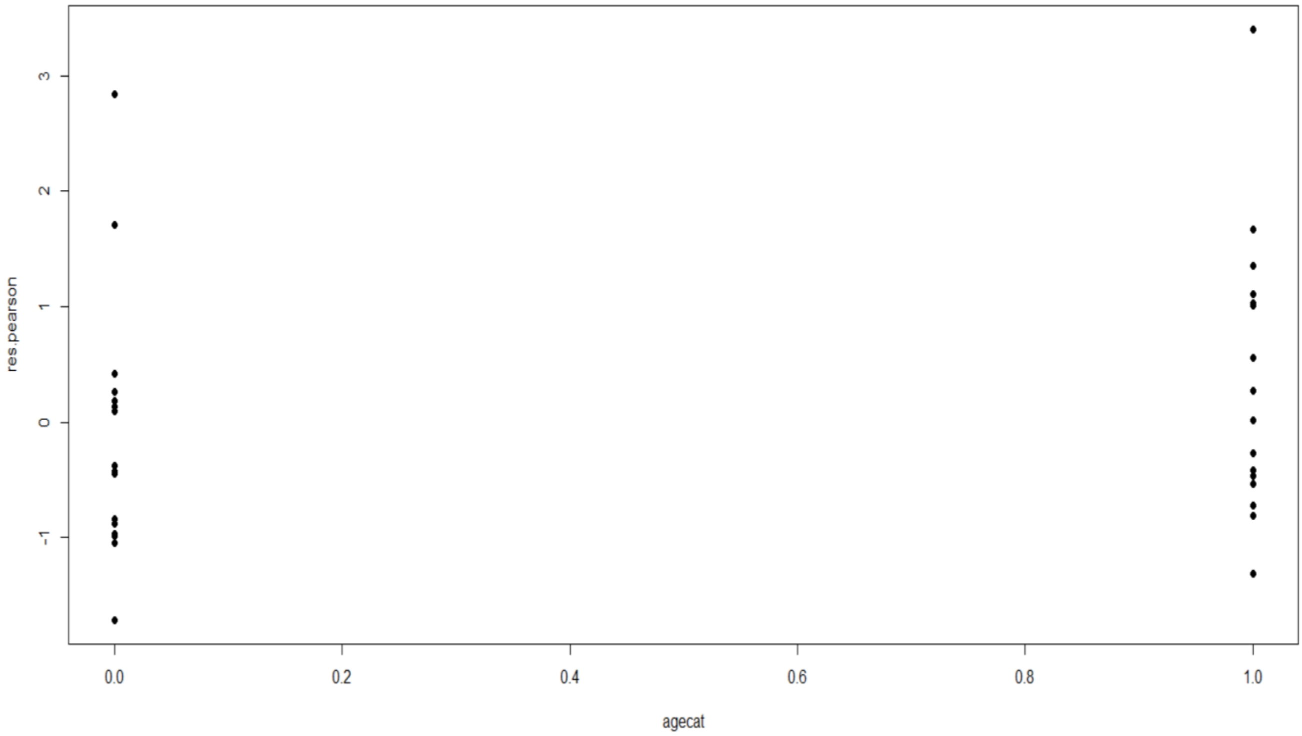
Επόμενο βήμα, για την υλοποίηση αυτής της άσκησης είναι σχεδιάσουμε κάποια διαγράμματα που θα μας βοηθήσουν να ερμηνεύσουμε καλύτερα το μοντέλο.

Τα πρώτα 2 διαγράμματα που αναπαριστώ είναι τα διαγράμματα για τα υπόλοιπα Pearson και Deviance.

Κώδικας υλοποίησης:

```
res.deviance<-residuals(model)

res.pearson<-residuals(model,type="pearson")

plot(agecat, res.deviance,pch=19)

plot(agecat, res.pearson,pch=19)
```



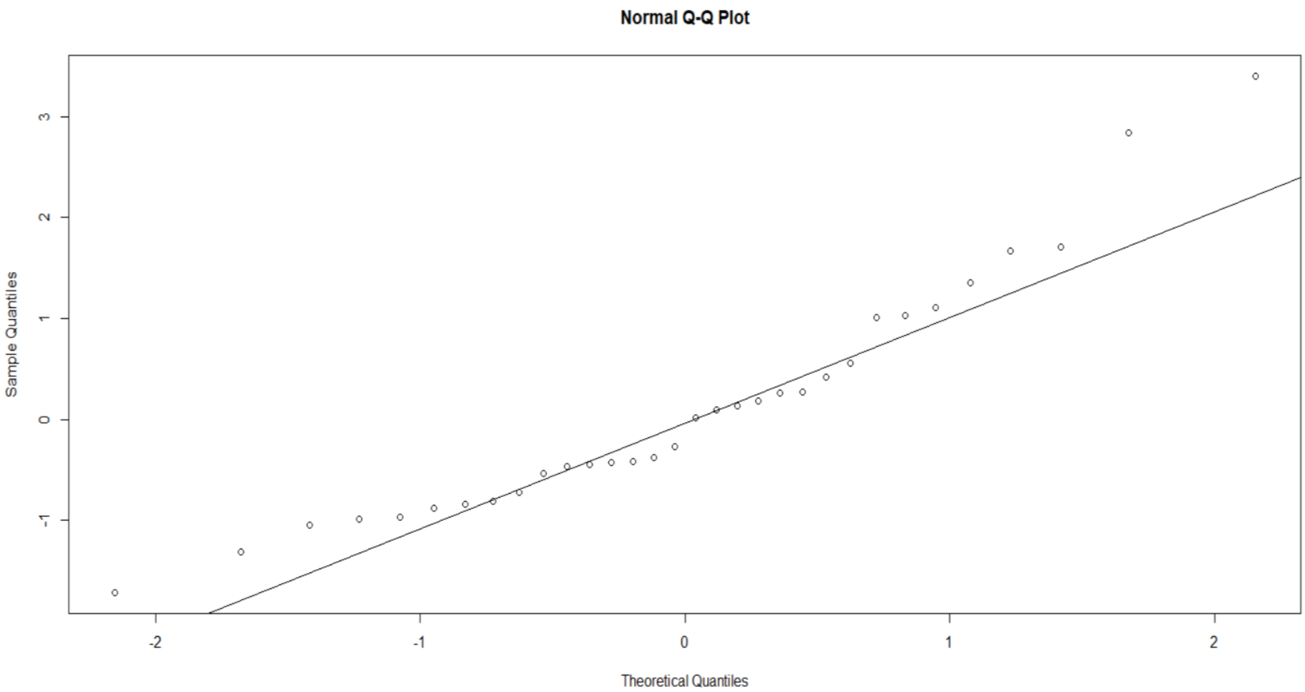


Στη συνέχεια εξετάζουμε την κανονική κατανομή των residuals Perason και Deviance.

Κώδικας υλοποίησης:

qqnorm(res.pearson)

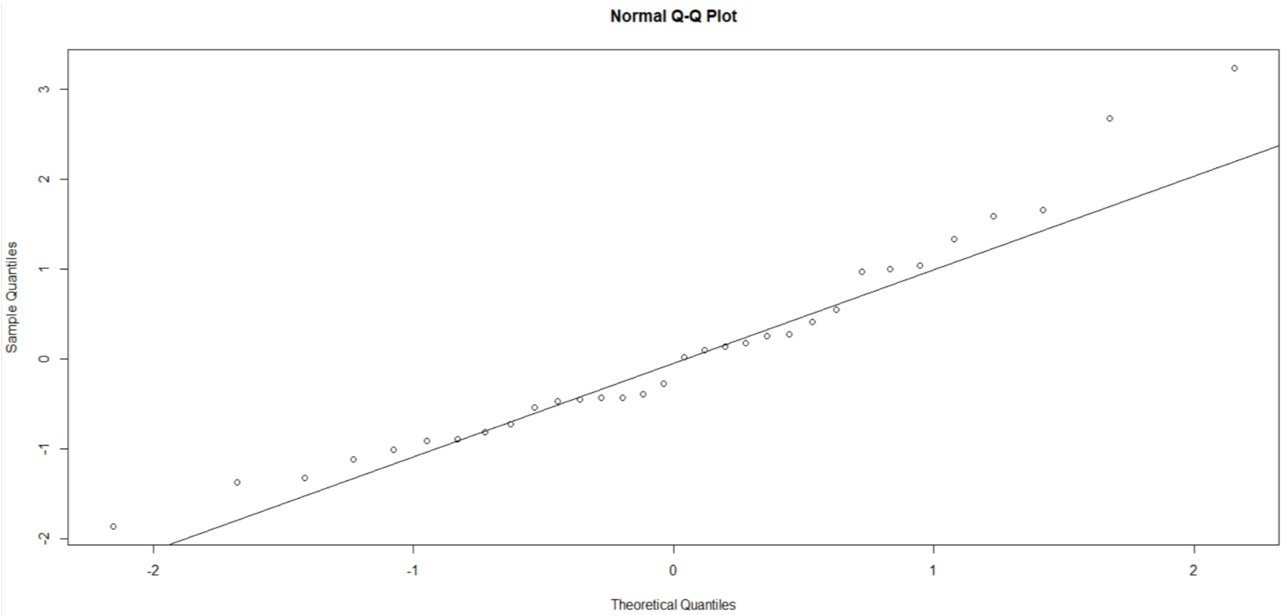
qqline(res.pearson)



Κώδικας υλοποίησης:

qqnorm(res.deviance)

qqline(res.deviance)

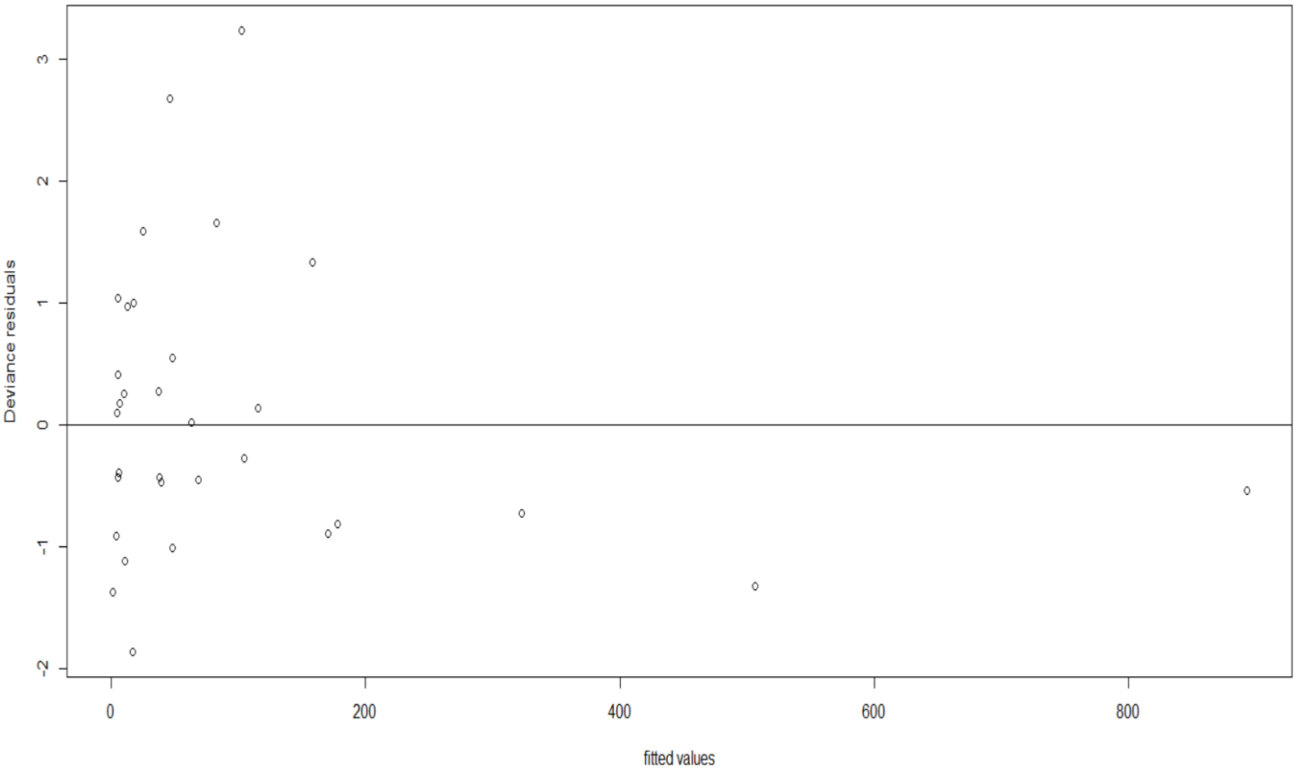


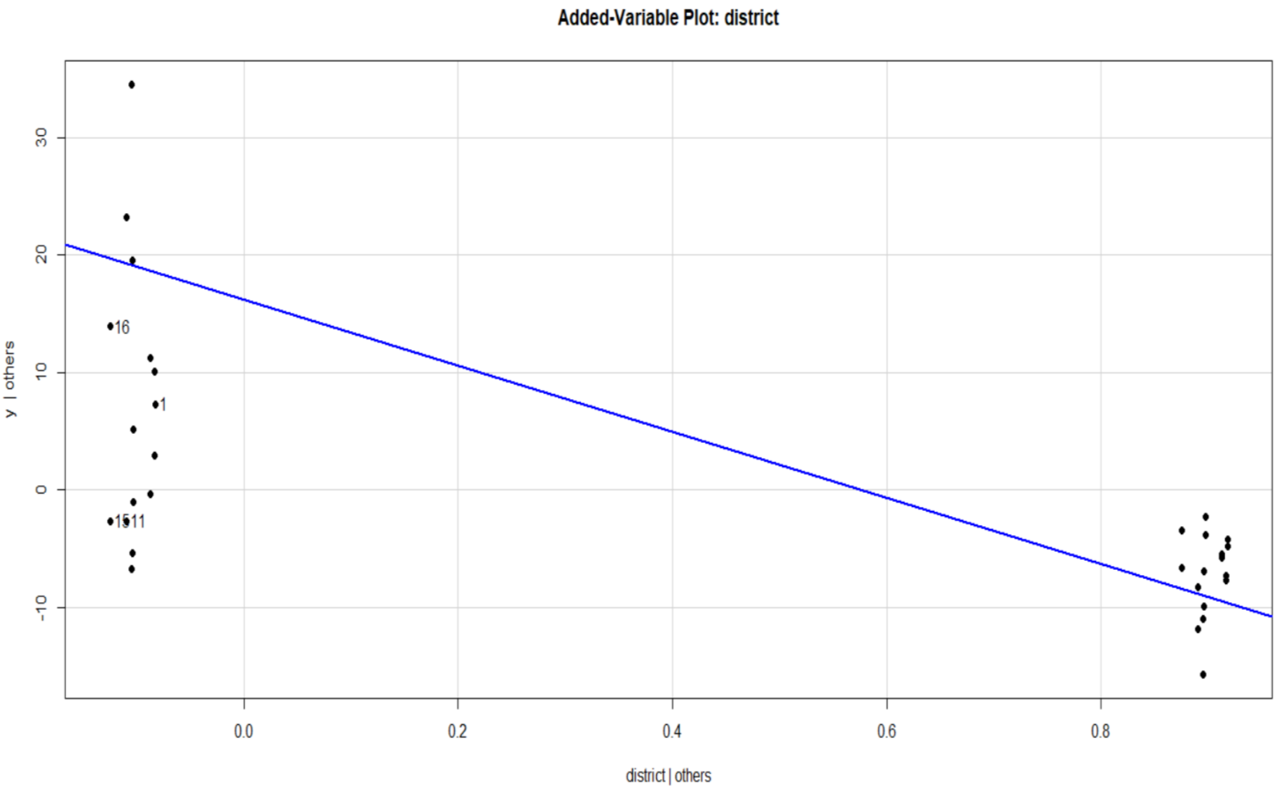
Κώδικας υλοποίησης:

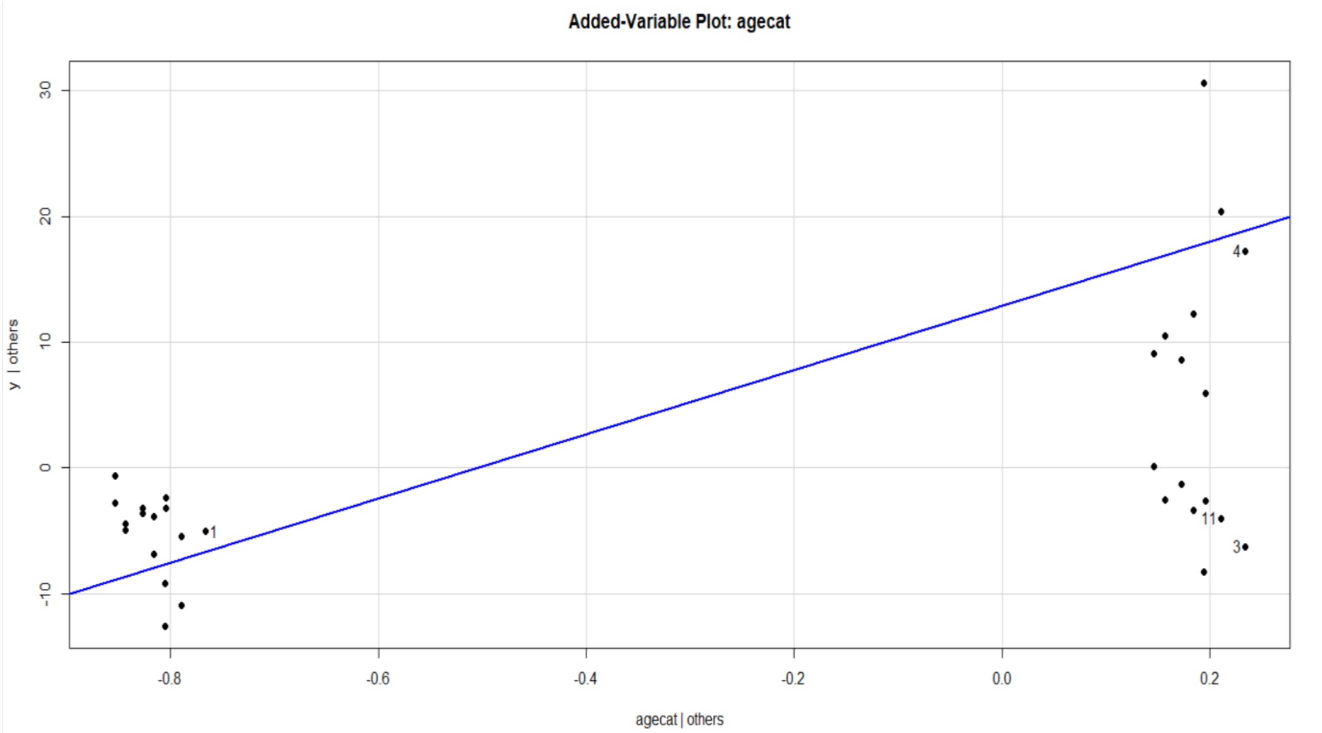
i. plot(fitted.values(model),res.deviance,xlab='fitted values', ylab='Deviance residuals')

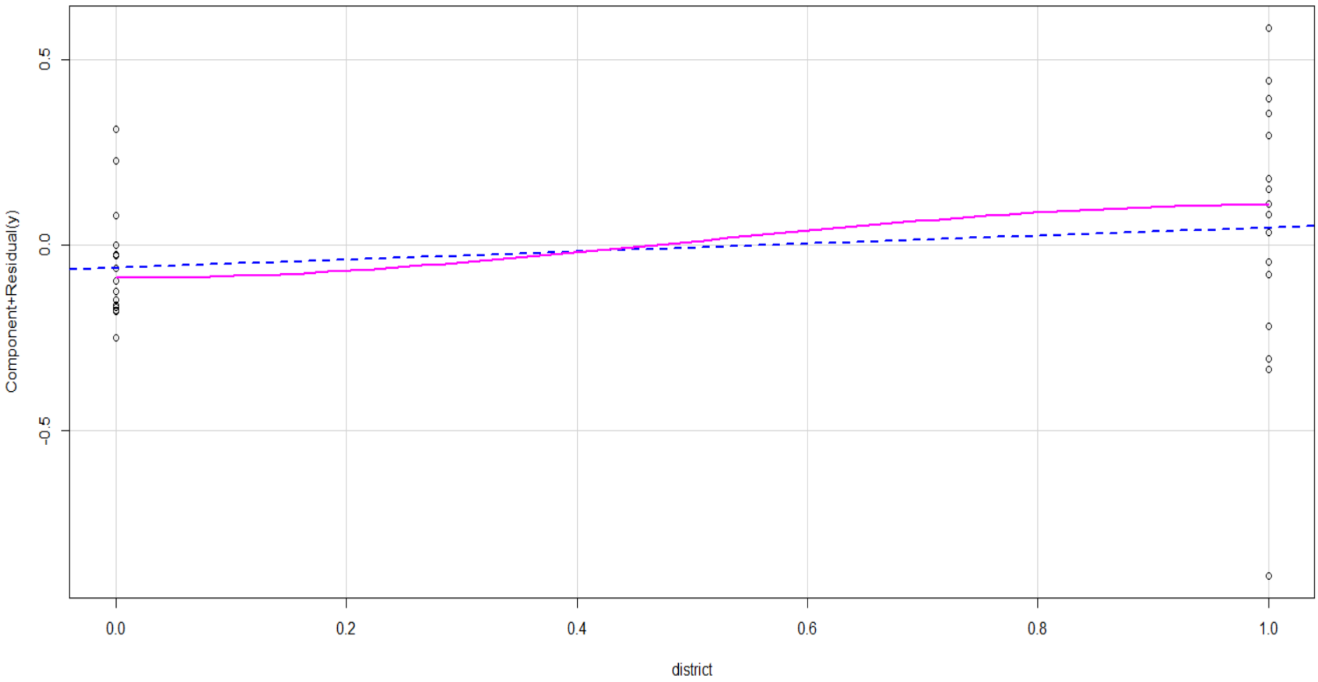
   abline(h=0)

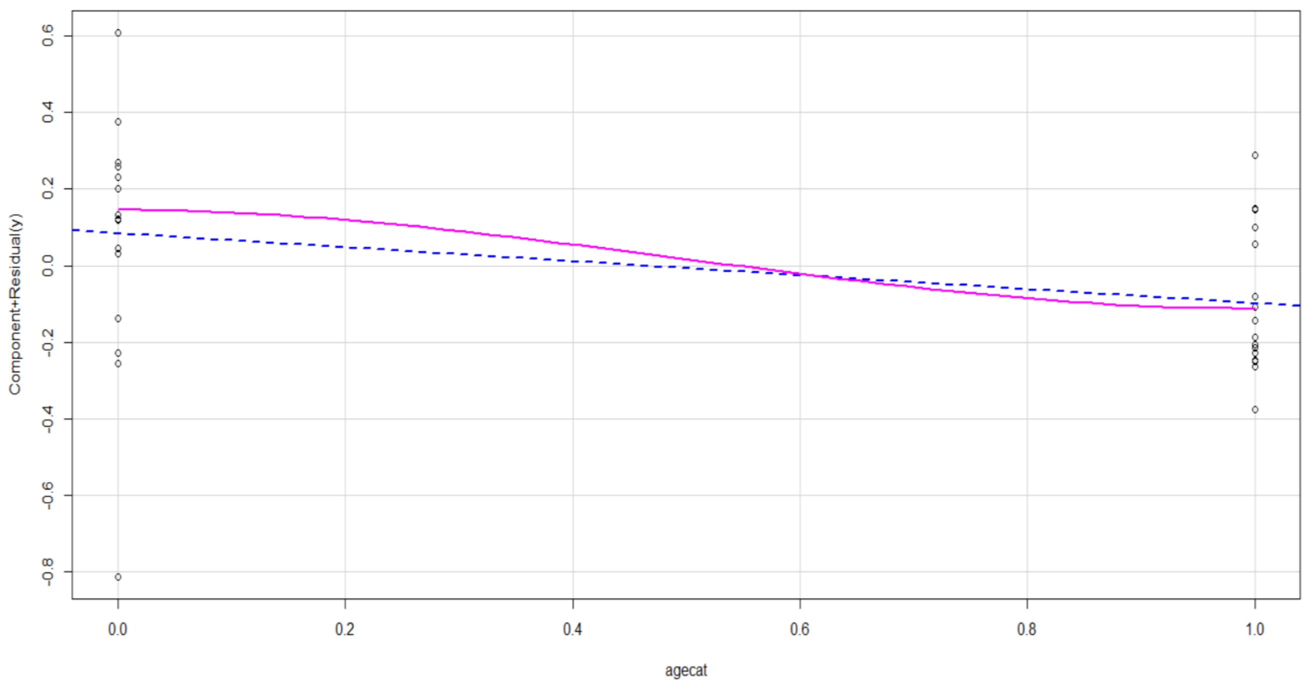
ii. avPlot(model, variable=district, pch=19)
iii. avPlot(model, variable=agecat, pch=19)
iv. crPlot(model, variable=district)
v. crPlot(model, variable=agecat)
vi. crPlot(model, variable=X2)


3
2
1
0
-1
-2
Deviance residuals
0
200
400
600
800
fitted values

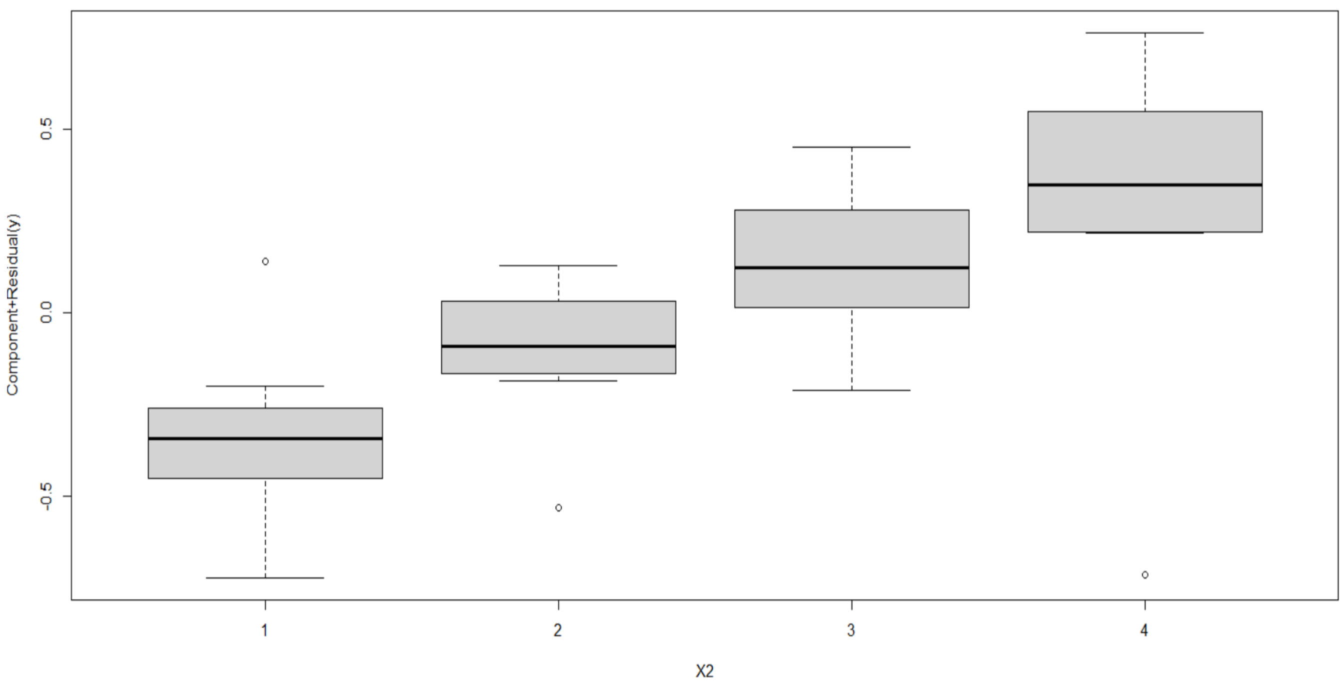

Added-Variable Plot: district
30
20
10
0
-10
y | others
16
1
1511
0.0
0.2
0.4
0.6
0.8
district | others


Added-Variable Plot: agecat
y | others
agecat | others
-0.8
-0.6
-0.4
-0.2
0.0
0.2
-10
0
10
20
30
1
4
11
3


Component+Residual(y)
district
0.0
0.2
0.4
0.6
0.8
1.0
-0.5
0.0
0.5



Component+Residual(y)
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
0.0
0.2
0.4
0.6
0.8
1.0
agecat

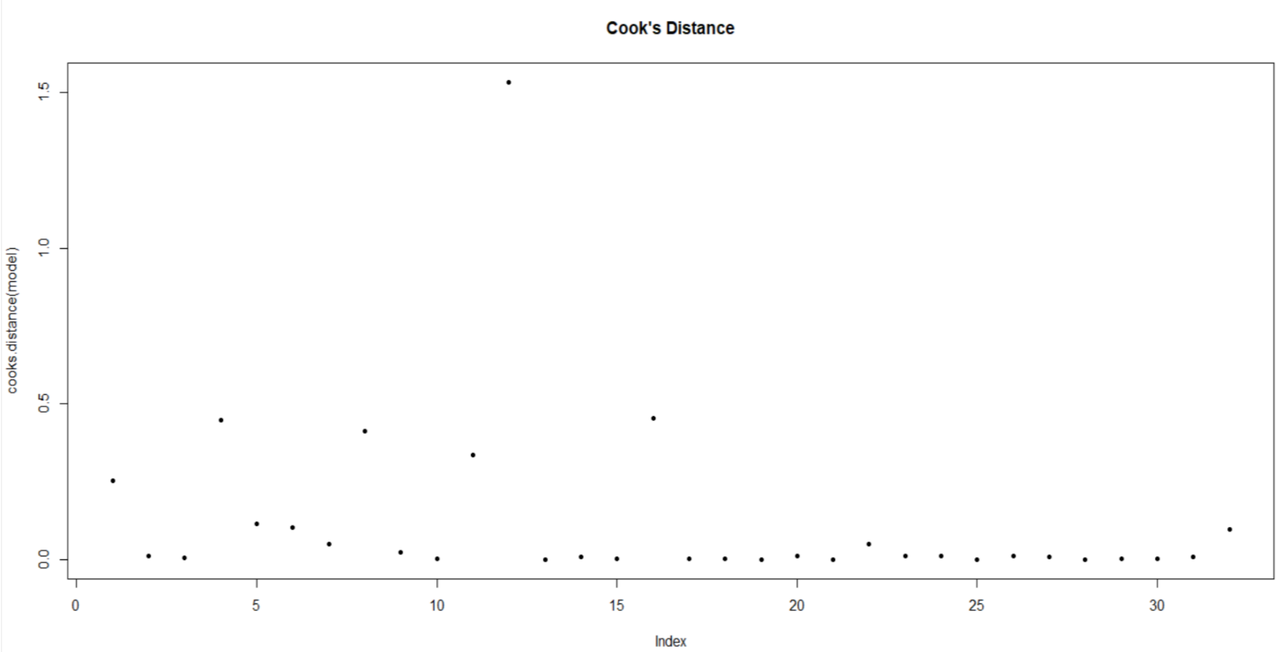


Component+Residual(y)
0.5
0.0
-0.5
1
2
3
4
X2

Cook Distance’s diagram:

Κώδικας υλοποίησης:

```
plot(cooks.distance(model), pch = 20, main = "Cook's Distance")
```



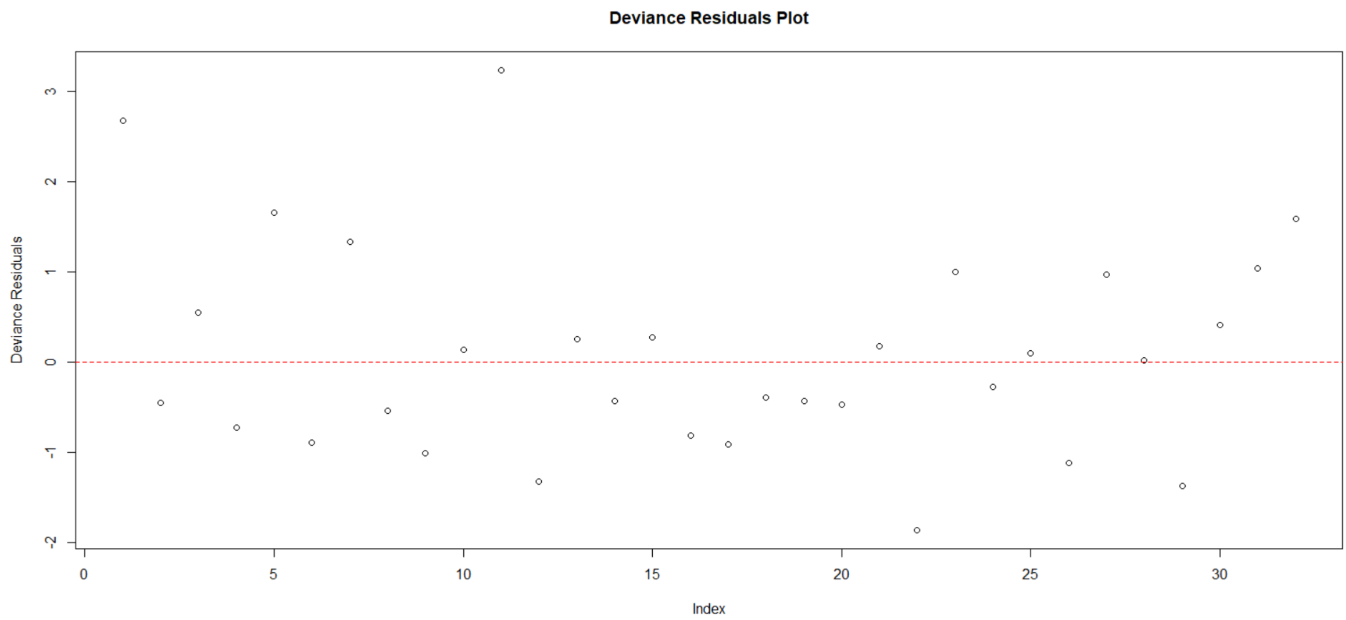
Το τελευταίο διάγραμμα που δημιουργούμε απεικονίζει τα υπόλοιπα πιθανοφάνειας.

Κώδικας υλοποίησης:

deviance_residuals <- residuals(model, type = "deviance")

plot(deviance_residuals, type = "p", main = "Deviance Residuals Plot", ylab = "Deviance Residuals")

abline(h = 0, col = "red", lty = 2)



Ασκηση 1.4

Στο ερώτημα αυτό,

Κώδικας υλοποίησης:

1.

new_var<- agecat*cartype

model1<-glm(formula = y ~ X2 + agecat + district+ new_var + offset(log(n)), family = poisson)

model1

summary(model1)

```
Call:
glm(formula = y ~ X2 + agecat + district + new_var + offset(log(n)),
    family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.91182  -0.80290   0.03817   0.74627   3.12287

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.87140    0.07312 -25.594  < 2e-16 ***
X22          0.11197    0.06349   1.764 0.077797 .
X23          0.29155    0.09711   3.002 0.002678 **
X24          0.40461    0.14422   2.805 0.005025 **
agecat      -0.52696    0.12419  -4.243  2.2e-05 ***
district     0.21688    0.05853   3.705 0.000211 ***
new_var      0.06693    0.05182   1.291 0.196565
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 207.833  on 31  degrees of freedom
Residual deviance:  40.112  on 25  degrees of freedom
AIC: 222.47
```

2.

new_var<- district*cartype

model1<-glm(formula = y ~ X2 + agecat + district+ new_var + offset(log(n)), family = poisson)

model1

summary(model1)

```
Call:
glm(formula = y ~ X2 + agecat + district + new_var + offset(log(n)),
    family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6987  -0.6552  -0.0200   0.4988   3.3224

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.91973    0.05569 -34.469  < 2e-16 ***
X22          0.15096    0.05076   2.974  0.00294 **
X23          0.37079    0.05630   6.586 4.51e-11 ***
X24          0.52222    0.07585   6.885 5.78e-12 ***
agecat      -0.37581    0.04452  -8.442  < 2e-16 ***
district    -0.08316    0.16993  -0.489  0.62456
new_var      0.12657    0.06628   1.910  0.05617 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 207.833  on 31  degrees of freedom
Residual deviance:  38.157  on 25  degrees of freedom
AIC: 220.52
```

Το μοντέλο είναι exp(-1.91793+0.15096*X2_2+0.37079* X2_3 +0.52222* X2_4 -0.37581* agecat - 0.08316* district + 0.12657*new_var)=

exp(-1.91793+0.15096*β1+0.37079* β2 +0.52222* β3 -0.37581* β4 - 0.08316* β5+ 0.12657*β6)

Σε αυτό το μοντέλο η μεταβλητή που έχουμε προσθέσει είναι η new_var , η οποία όμως αποτελεί αποτέλεσμα του cartype*district. Αν η μεταβλητή αυτή αυξηθεί κατά 1 μονάδα, τότε ο αριθμός Y αποζημιώσεων λόγω τροχαίων ατυχημάτων ανά n συμβόλαια θα πολλαπλασιαστεί κατά exp(0.12657)=1.13, δηλαδή θα αυξηθεί κατά 13%.

Επίσης υπάρχει μία διαφοροποίηση στη μεταβλητή district, η οποία πλέον όταν αυξάνεται κατά 1 μονάδα, ο αριθμός Y αποζημιώσεων λόγω τροχαίων ατυχημάτων ανά n συμβόλαια θα πολλαπλασιαστεί κατά exp(-0.08316)=0.92, δηλαδή θα μειωθεί κατά 8%.

3.

new_var<- agecat*district

model1<-glm(formula = y ~ X2 + agecat + district+ new_var + offset(log(n)), family = poisson)

model1

summary(model1)

```
Call:
glm(formula = y ~ X2 + agecat + district + new_var + offset(log(n)),
    family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2106  -0.6509  -0.2148   0.8084   3.2908

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.91460    0.05599 -34.196  < 2e-16 ***
X22          0.16279    0.05048   3.225  0.00126 **
X23          0.39565    0.05491   7.205 5.79e-13 ***
X24          0.56639    0.07216   7.849 4.18e-15 ***
agecat      -0.40282    0.04629  -8.702  < 2e-16 ***
district    -0.06127    0.15970  -0.384  0.70125
new_var      0.32763    0.17167   1.908  0.05633 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 207.833  on 31  degrees of freedom
Residual deviance:  37.889  on 25  degrees of freedom
AIC: 220.25
```

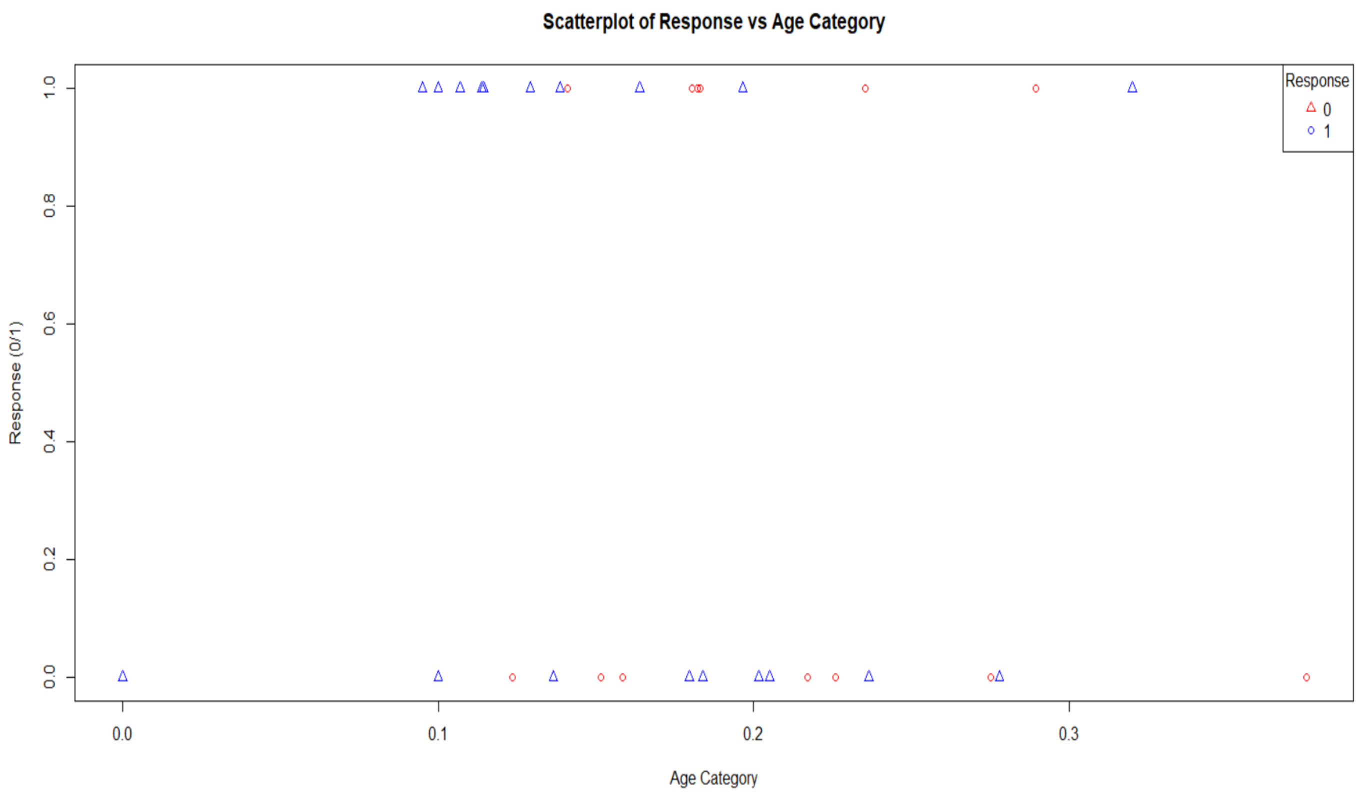
Το μοντέλο είναι exp(-1.91460+0.16729*X2_2+0.39565* X2_3 +0.56639* X2_4 -0.40282* agecat - 0.06127* district + 0.32763*new_var)=

exp(-1.91460+0.16729*β1+0.39565* β2+0.56639* β3 -0.40282* β4- 0.06127* β5+ 0.32763*β6

Σε αυτό το μοντέλο η μεταβλητή που έχουμε προσθέσει είναι η new_var , η οποία όμως αποτελεί αποτέλεσμα του agecat*district. Αν η μεταβλητή αυτή αυξηθεί κατά 1 μονάδα, τότε ο αριθμός Y αποζημιώσεων λόγω τροχαίων ατυχημάτων ανά n συμβόλαια θα πολλαπλασιαστεί κατά exp(0.32763)=1.39, δηλαδή θα αυξηθεί κατά 39%.

Επίσης υπάρχει μία διαφοροποίηση στη μεταβλητή district, η οποία πλέον όταν αυξάνεται κατά 1 μονάδα, ο αριθμός Y αποζημιώσεων λόγω τροχαίων ατυχημάτων ανά n συμβόλαια θα πολλαπλασιαστεί κατά exp(-0.06127)=0.94, δηλαδή θα μειωθεί κατά 6%.

Τα μοντέλα παλινδρόμησης Poisson που μου είναι στατιστικά σημαντικά με την είσοδο της νέας μεταβλητής είναι αυτά για new_var<- district*cartype και για new_var<- agecat*district . Και τα δύο έχουν p-value στα πέριξ του 5%.Τα δύο αυτά μοντέλα είναι σχεδόν παρόμοια. Έχουν και τα 2 σχεδόν ίδιο AIC.



ΑΣΚΗΣΗ 2

Στην συγκεκριμένη άσκηση κάνοντας χρήση ενός μοντέλου λογιστικής παλινδρόμησης θα εξετάσουμε την εξάρτηση της πιθανότητας ανταπόκρισης της θεραπείας από τις συμμεταβλητές age, smear, infiltrate, index, blasts και temperature κάνοντας χρήση των στατιστικών ελέγχων Wald και Deviance καθώς και του κριτηρίου AIC.

Κώδικας υλοποίησης:

Εφαρμογή μοντέλου λογιστικής παλινδρόμησης

```
logistic_model <- glm(response ~ age + smear + infiltrate + index + blasts + temperature, data = data, family = "binomial")

# Εκτύπωση των αποτελεσμάτων

summary(logistic_model)

#Εφαρμογή στατιστικών ελέγχων

wald_test <- summary(logistic_model)$coefficients[, "Pr(>|z|)"]

wald_test

deviance_test1 <- anova(logistic_model, test = "Chisq")

deviance_test1

cat("Deviance Test:\n")

deviance_test2 <- 1 - pchisq(logistic_model$deviance, df = logistic_model$df.residual)

deviance_test2

# Εκτύπωση των αποτελεσμάτων

result <- data.frame(Wald_Test = wald_test, Deviance_Test = deviance_test)

print(result)

 step(logistic_model,method="backward", test="Chisq")

logLik(logistic_model)

 aic <- AIC(logistic_model)

cat("AIC:", aic, "\n")
```

```
Call:
glm(formula = response ~ age + smear + infiltrate + index + blasts +
    temperature, family = "binomial", data = data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.73878  -0.58099  -0.05505   0.62618   2.28425

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 98.52361   40.85385   2.412  0.01588 *
age         -0.06029    0.02729  -2.210  0.02714 *
smear       -0.00480    0.04108  -0.117  0.90698
infiltrate   0.03621    0.03934   0.921  0.35728
index        0.39845    0.13278   3.001  0.00269 **
blasts       0.01343    0.05782   0.232  0.81627
temperature -0.10223    0.04181  -2.445  0.01448 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 70.524  on 50  degrees of freedom
Residual deviance: 40.060  on 44  degrees of freedom
AIC: 54.06
```

Το μοντέλο είναι exp(-98.52361 -0.06029*age -0.0048* smear +0.03261* infiltrate +0.39845* index+ 0.01343* blasts –0.10223*temperature)=

exp(-98.52361 -0.06029*β1-0.0048* β2+0.03261* β3 +0.39845* β4 0.01343* β5 –0.10223*β6)

Αν αυξηθεί η συμμεταβλητή age κατά μια μονάδα (δηλαδή για έναν ασθενή) θα πολλαπλασιαστεί ο αριθμός Υ ανταπόκρισης της θεραπείας κατά exp(-0.06)= 0.94, δηλαδή θα μειωθεί η ηλικία των ασθενών κατά 6%.

Αν αυξηθεί η συμμεταβλητή smear κατά μια μονάδα (δηλαδή το ποσοστό επίστρωσης των βλαστοκυττάρων) θα πολλαπλασιαστεί ο αριθμός Υ ανταπόκρισης της θεραπείας κατά exp(-0.1)= 0.9, δηλαδή θα μειωθεί σχεδόν κατά 10%.

Αν αυξηθεί η συμμεταβλητή infiltrate κατά μια μονάδα (δηλαδή το ποσοστό κυττάρων στο μυελό των οστών) θα πολλαπλασιαστεί ο αριθμός Υ ανταπόκρισης της θεραπείας κατά exp(0.036)= 1.03, δηλαδή θα αυξηθεί ελάχιστα.

Αν αυξηθεί η συμμεταβλητή index κατά μια μονάδα (δηλαδή ο δείκτης κυττάρων λευκαιμίας) θα πολλαπλασιαστεί ο αριθμός Y ανταπόκρισης της θεραπείας κατά exp(0.39845)= 1.49, δηλαδή θα αυξηθεί κατά 50%.

Αν αυξηθεί η συμμεταβλητή blasts κατά μια μονάδα (δηλαδή τα βλαστοκύτταρα) θα πολλαπλασιαστεί ο αριθμός Y ανταπόκρισης της θεραπείας κατά exp(0.01343)= 1.01, δηλαδή θα αυξηθεί ελάχιστα.

Αν αυξηθεί η συμμεταβλητή temperature κατά μια μονάδα (δηλαδή η θερμοκρασία πριν από τη θεραπεία) θα πολλαπλασιαστεί ο αριθμός Y ανταπόκρισης της θεραπείας κατά exp(-0.0048)= 0.995, δηλαδή δεν θα μειωθεί σχεδόν καθόλου.

Το AIC είναι 54.06 που δείχνει μία πολύ καλή προσαρμοστικότητα του μοντέλου με τη πραγματικότητα. Δηλαδή έχει πολύ χαμηλή απώλεια πληροφορίας.

```
> wald_test <- summary(logistic_model)$coefficients[, "Pr(>|z|)"]
> wald_test
(Intercept)         age       smear  infiltrate       index      blasts 
0.015882192 0.027138297 0.906977044 0.357277947 0.002691784 0.816268791 
temperature 
0.014480775 
> deviance_test1 <- anova(logistic_model, test = "Chisq")
> deviance_test1
Analysis of Deviance Table

Model: binomial, link: logit

Response: response

Terms added sequentially (first to last)


            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)    
NULL                           50     70.524              
age          1   6.5207        49     64.004 0.0106626 *  
smear        1   1.2549        48     62.749 0.2626219    
infiltrate   1   1.8047        47     60.944 0.1791485    
index        1  12.1251        46     48.819 0.0004975 ***
blasts       1   0.5416        45     48.277 0.4617513    
temperature  1   8.2175        44     40.060 0.0041487 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> cat("Deviance Test:\n")
Deviance Test:
> deviance_test2 <- 1 - pchisq(logistic_model$deviance, df = logistic_model$df.residual
> deviance_test2
[1] 0.6411612
> # Εκτύπωση των αποτελεσμάτων
> result <- data.frame(Wald_Test = wald_test, Deviance_Test = deviance_test)
> print(result)
              Wald_Test Deviance_Test
(Intercept) 0.015882192    0.04742357
age         0.027138297    0.04742357
smear       0.906977044    0.04742357
infiltrate  0.357277947    0.04742357
index       0.002691784    0.04742357
blasts      0.816268791    0.04742357
temperature 0.014480775    0.04742357
```

```
> step(logistic_model,method="backward", test="Chisq")
Start:  AIC=54.06
response ~ age + smear + infiltrate + index + blasts + temperature

              Df Deviance    AIC     LRT Pr(>Chi)    
- smear        1   40.074 52.074  0.0137 0.906781    
- blasts       1   40.115 52.115  0.0547 0.815120    
- infiltrate   1   41.023 53.023  0.9628 0.326491    
<none>             40.060 54.060                     
- age          1   46.157 58.157  6.0969 0.013542 *  
- temperature  1   48.277 60.277  8.2175 0.004149 ** 
- index        1   55.823 67.823 15.7628 7.18e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Step:  AIC=52.07
response ~ age + infiltrate + index + blasts + temperature

              Df Deviance    AIC     LRT  Pr(>Chi)    
- blasts       1   40.136 50.136  0.0626  0.802420    
<none>             40.074 52.074                      
- infiltrate   1   42.615 52.615  2.5412  0.110913    
- age          1   46.216 56.216  6.1421  0.013200 *  
- temperature  1   48.346 58.346  8.2727  0.004025 ** 
- index        1   56.308 66.308 16.2346 5.596e-05 ***
```

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

Step:  AIC=50.14
response ~ age + infiltrate + index + temperature

              Df Deviance    AIC     LRT  Pr(>Chi)
<none>             40.136 50.136
- infiltrate   1   43.265 51.265  3.1291  0.076904 .
- age          1   46.438 54.438  6.3019  0.012061 *
- temperature  1   48.971 56.971  8.8344  0.002956 **
- index        1   57.602 65.602 17.4658 2.925e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Call:  glm(formula = response ~ age + infiltrate + index + temperature,
    family = "binomial", data = data)

Coefficients:
(Intercept)          age   infiltrate        index  temperature
   95.56766     -0.06026      0.03413      0.40673     -0.09944

Degrees of Freedom: 50 Total (i.e. Null);  46 Residual
Null Deviance:      70.52
Residual Deviance: 40.14        AIC: 50.14
```

Σε αυτό το σημείο θα υπολογίσω το διάσημα εμπιστοσύνης για τους εκτιμημένους συντελεστές του μοντέλου που μόλις δημιούργησα.

Κώδικας υλοποίησης:

i. confint.default(logistic_model)
ii. exp(confint.default(logistic_model))

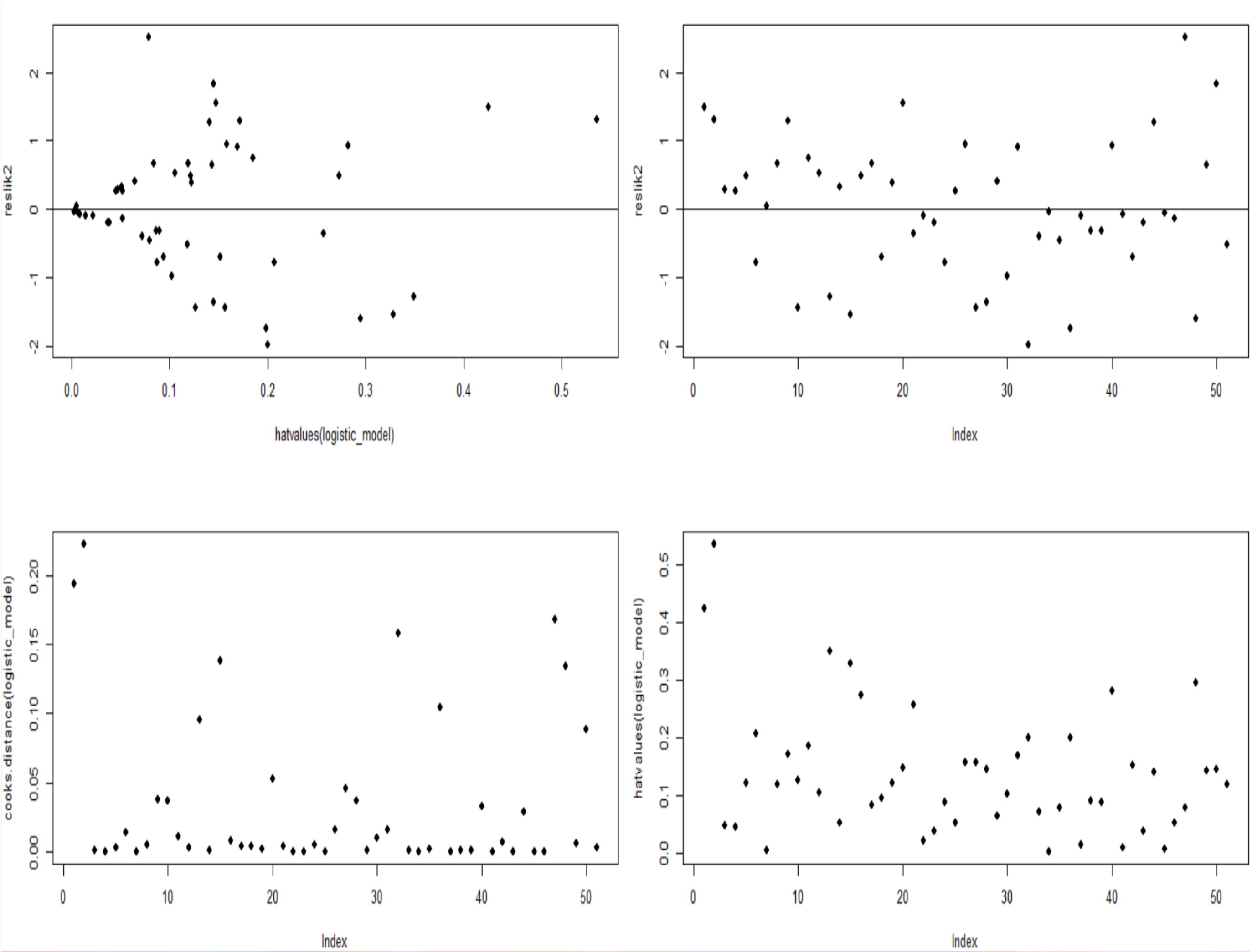
```
> confint.default(logistic_model)
                  2.5 %       97.5 %
(Intercept) 18.45154163 178.595679709
age         -0.11377522  -0.006809858
smear       -0.08530518   0.075705764
infiltrate  -0.04088792   0.113314275
index        0.13821163   0.658682569
blasts      -0.09989165   0.126760437
temperature -0.18417357  -0.020283691
> exp(confint.default(logistic_model))
                   2.5 %       97.5 %
(Intercept) 1.031342e+08 3.656943e+77
age         8.924585e-01 9.932133e-01
smear       9.182320e-01 1.078645e+00
infiltrate  9.599367e-01 1.119984e+00
index       1.148219e+00 1.932245e+00
blasts      9.049355e-01 1.135145e+00
temperature 8.317914e-01 9.799206e-01
```

Υλοποίηση διαγραμμάτων:

Διάγραμμα απεικόνισης υπολοίπων πιθανοφάνειας και cook distance

Κώδικας υλοποίησης:

```
reslik2<-rstudent(logistic_model)

par (mfrow=c(2,2))

plot(hatvalues(logistic_model), reslik2, pch=19)

abline(h=0)

plot(reslik2, pch=19)

abline(h=0)

plot(cooks.distance(logistic_model), pch=19)

plot(hatvalues(logistic_model), pch=19)
```



Διάγραμμα fitted values

Κωδικας υλοποιησης:

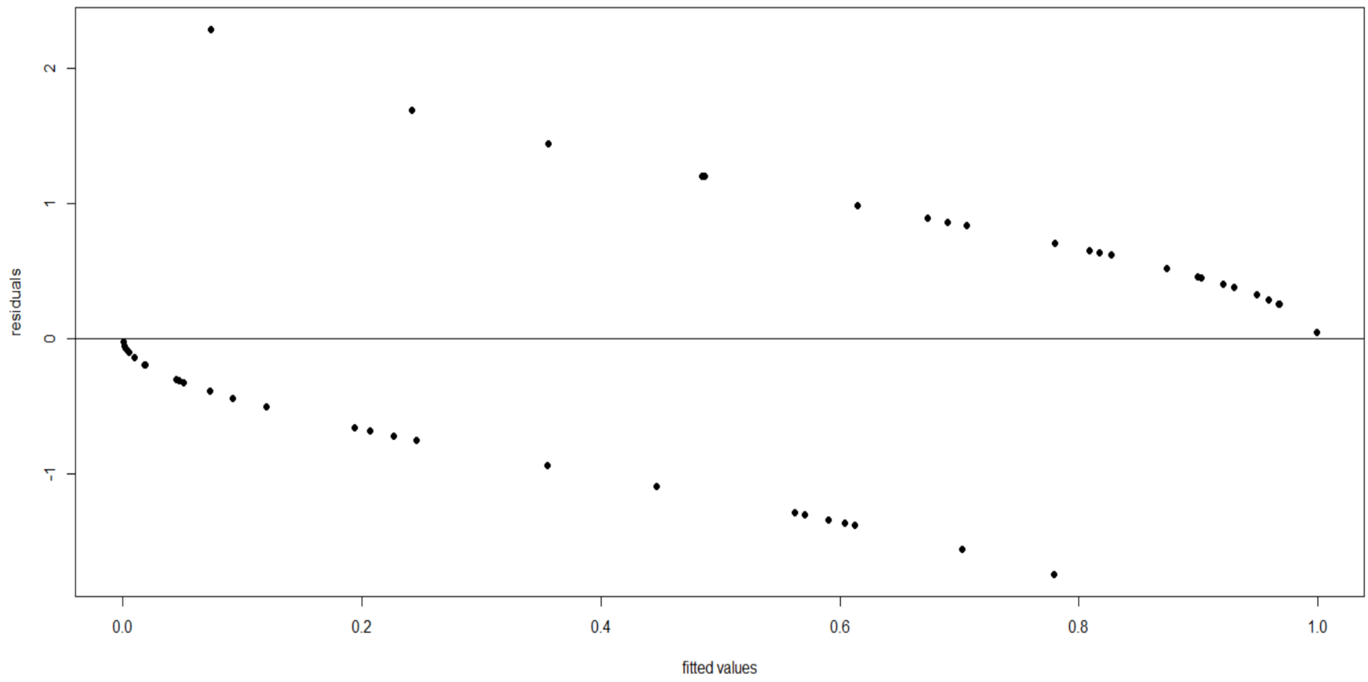
residuals(logistic_model)

rstandard(logistic_model)

plot(fitted.values(logistic_model), residuals(logistic_model), xlab='fitted values', ylab='Deviance

residuals', pch=19)

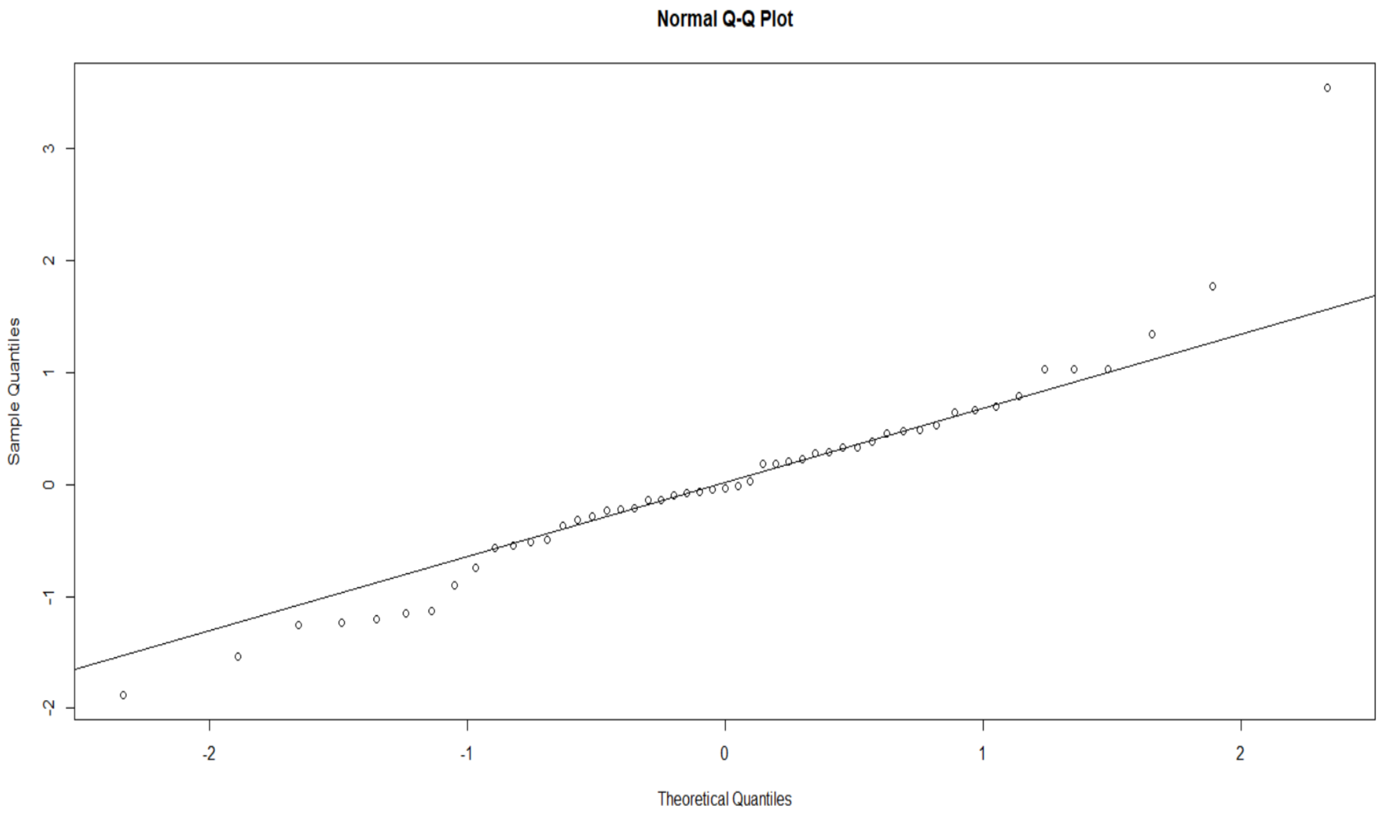
abline(h=0)



Διάγραμμα qqplot για pearson residuals

Κώδικας υλοποίησης:

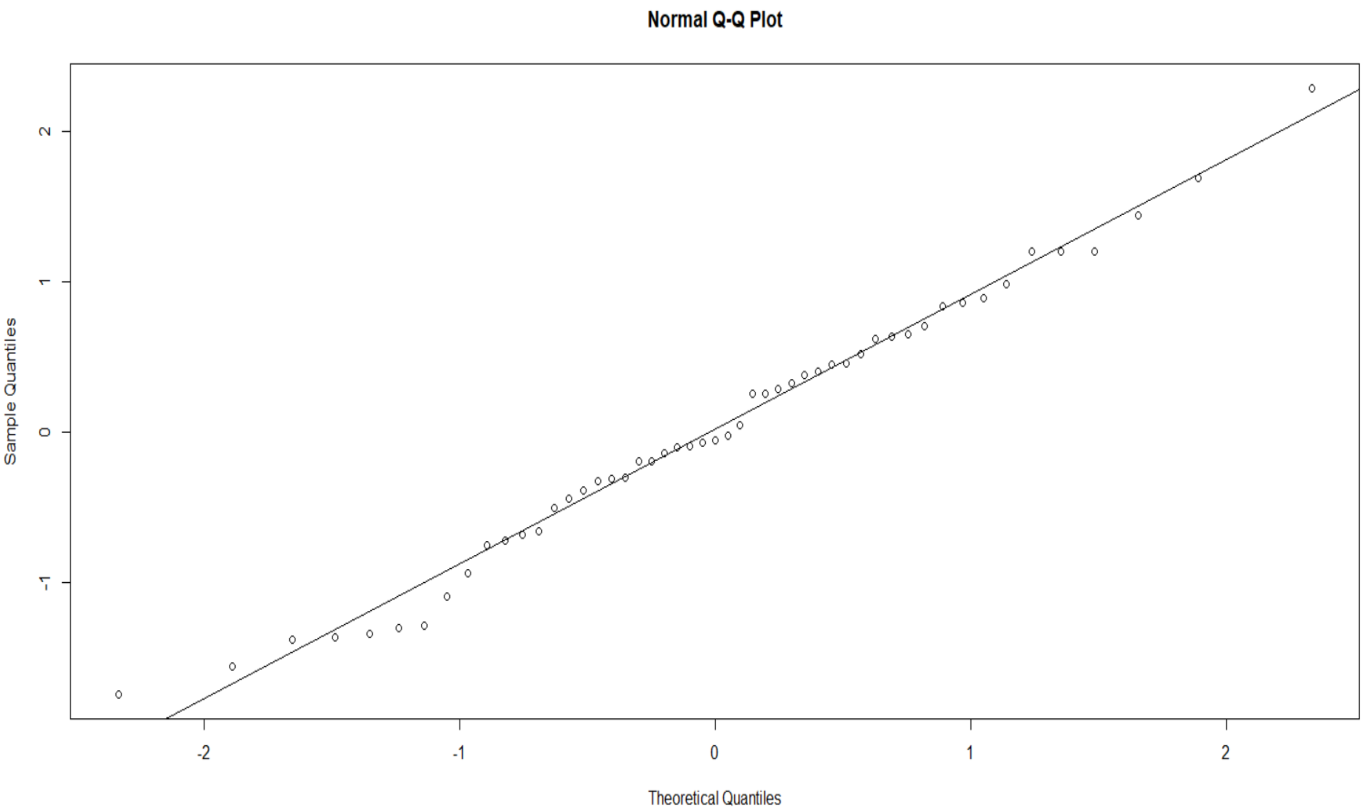
```
res.pearson2<-residuals(logistic_model,type="pearson")

qqnorm(res.pearson2)

qqline(res.pearson2)
```



Διάγραμμα qqplot για deviance residuals

Κώδικας υλοποίησης:

```
res.deviance<-residuals(logistic_model,type="deviance")
qqnorm(res.deviance)
qqline(res.deviance)
```

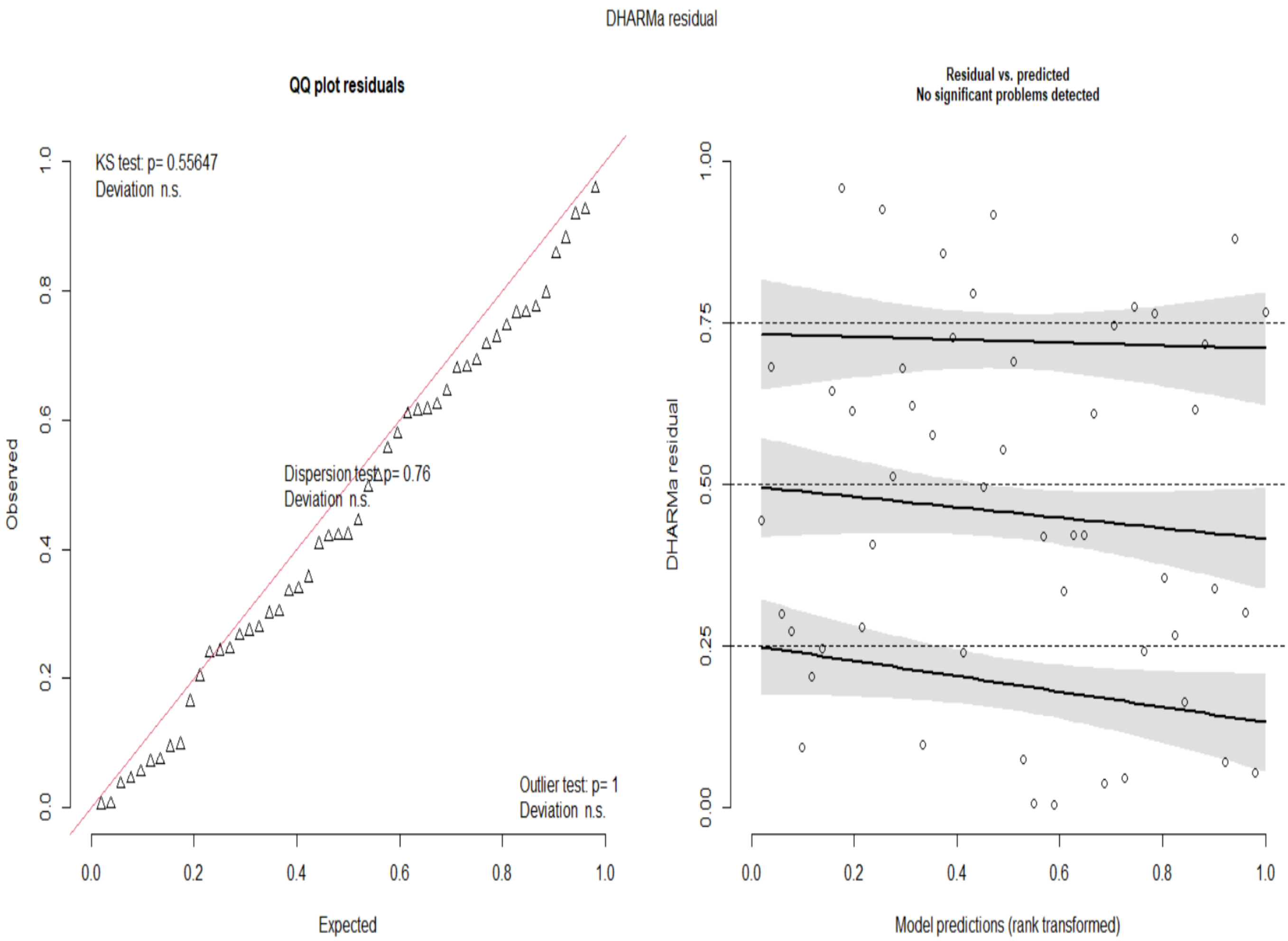


Διαγράμματα για την απεικόνιση μερικών υπολοίπων, των υπολοίπων Deviance (με την ημι-κανονική κατανομή)

Κώδικας υλοποίησης:

residuals_dharma <- simulateResiduals(fittedModel = logistic_model, n = 1000)

plot(residuals_dharma)



Τέλος ελέγχουμε τη προβλεπτική ικανότητα του μοντέλου μας, μέσω της καμπύλης ROC.

Κώδικας υλοποίησης:

1ος τρόπος:

install.packages("pROC")

library(pROC)

 # Προβλέψεις πιθανοτήτων για τα δεδομένα εκπαίδευσης

predicted_probabilities <- predict(logistic_model, type = "response")

predicted_probabilities

Δημιουργία αντικειμένου ROC

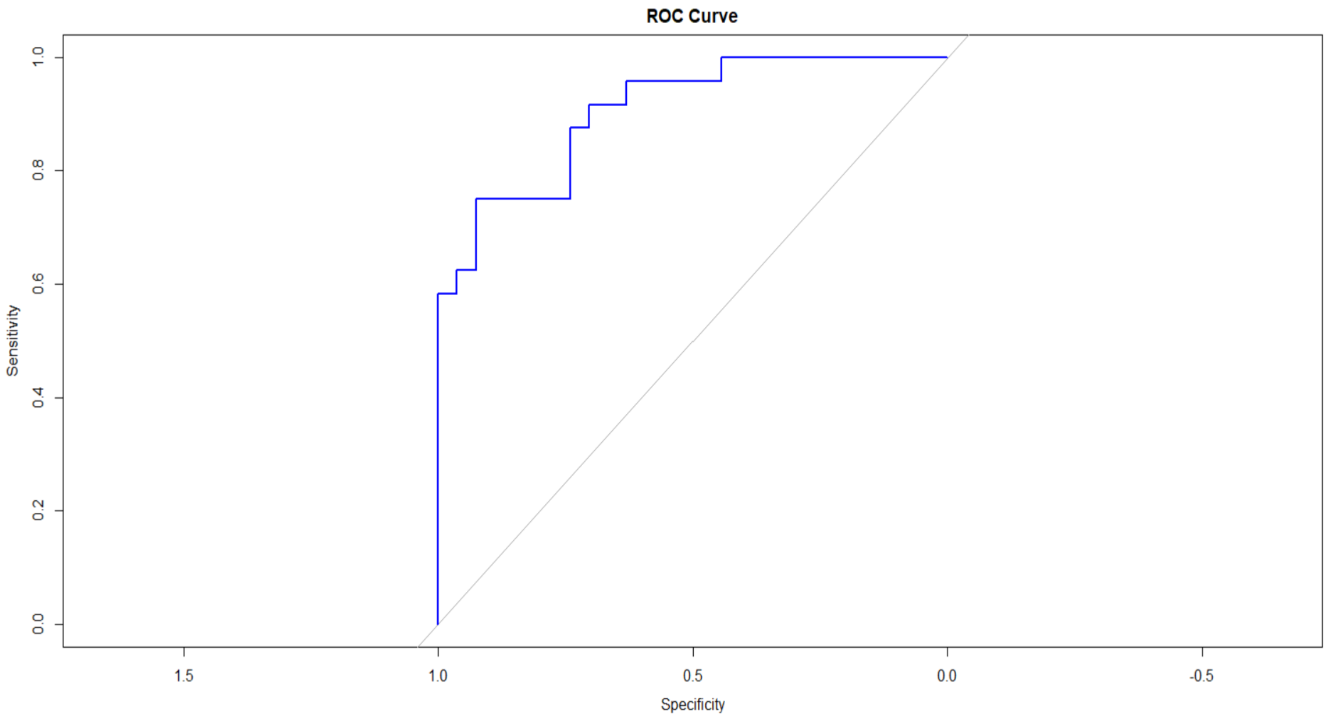
roc_curve <- roc(response, predicted_probabilities)

roc_curve

Κατασκευή και εμφάνιση της καμπύλης ROC

plot(roc_curve, main="ROC Curve", col="blue", lwd=2)

lines(c(0, 1), c(0, 1), col="gray", lty=2, lwd=1.5) # Τυπική καμπύλη

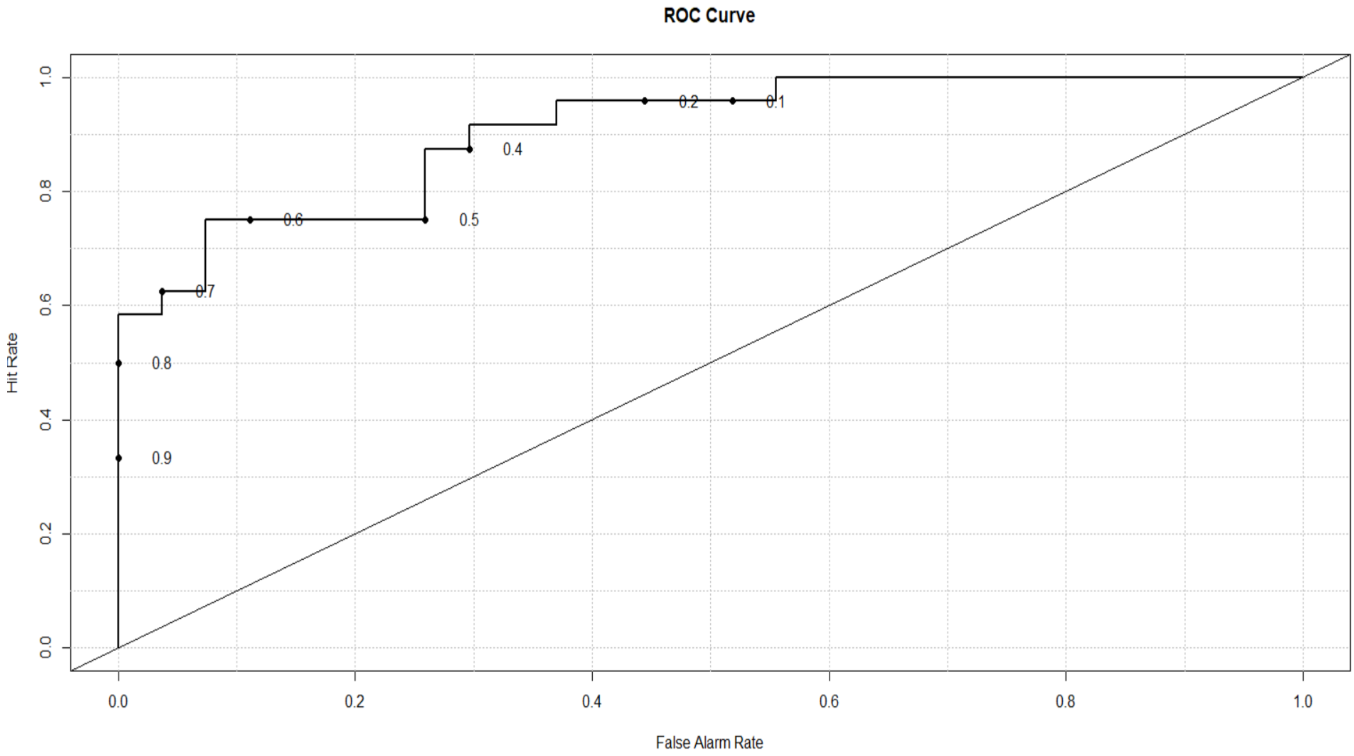


2ος τρόπος:

install.packages("verification")

library(verification)

roc.plot(response,predicted_probabilities)



Ο άξονας Χ αντιπροσωπεύει το ποσοστό των λανθασμένων θετικών προβλέψεων στο σύνολο των πραγματικά αρνητικών παρατηρήσεων.
Ο άξονας Υ αντιπροσωπεύει το ποσοστό των σωστών θετικών προβλέψεων στο σύνολο των πραγματικά θετικών παρατηρήσεων.
Όταν η καμπύλη βρίσκεται προς τα πάνω και προς τα αριστερά, υποδεικνύει καλύτερη απόδοση του μοντέλου.

Η γωνία στην επάνω αριστερή γωνία (0,1) αντιστοιχεί σε ένα τέλειο μοντέλο που δεν κάνει καμία λανθασμένη ταξινόμηση.

Η AUC (Area Under the Curve) είναι ένα μέτρο της συνολικής απόδοσης του μοντέλου. Μια τιμή AUC κοντά στο 1 υποδεικνύει καλή απόδοση, ενώ μια τιμή κοντά στο 0.5 υποδεικνύει τυχαία ταξινόμηση.

Καλή απόδοση σε ένα ROC curve εμφανίζεται όταν η καμπύλη είναι κοντά στην επάνω αριστερή γωνία και η AUC είναι υψηλή.

Άρα στο 90% των σωστών θετικών προβλέψεων και στο 30% των λανθασμένων θετικών προβλέψεων , σε αντιστοιχία κλίμακα, το μοντέλο μας αρχίζει να έχει μία καλή και σωστή ταξινόμηση, το οποίο στη συνέχεια το βοηθάει να μεγιστοποιήσει την απόδοση του.