Εθνικό Μετσόβιο Πολυτεχνείο
Επιστήμη Δεδομένων και Μηχανική Μάθηση

ΣΤΑΤΙΣΤΙΚΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ

Σειρά 3

Νάνος Γεώργιος
03400144
nanosgiwrgos1997@gmail.com

"The only relevant test of the validity of a hypothesis is comparison of its predictions with experience."

– Milton Friedman - Statistician

Άσκηση 1 ~ Poisson

Μέσω της παλινδρόμησης Poisson θα προσαρμοστεί το μοντέλο και θα πραγματοποιηθεί η ανάλυση. Για την ανάλυση, μετατρέπεται η μεταβλητη n σε offset(log(n)). Το n αφορά το πλήθος των συμβολαίων. Η χρήση του offset γίνεται για να είναι όλες οι μεταβλητές στην ίδια μονάδα μέτρησης και να δύναται να μετρηθεί το rate (αριθμός αποζημιώσεων ανά συμβόλαιο) ενώ το log, λόγω του ότι έτσι μοντελοποιείται η συνάρτηση poisson εφόσον έχει link function τη λογαριθμική.

```
library(MASS)
library(car)
```

```
## Loading required package: carData
```

```
model1 <- read.table("./asfalies.txt", header=TRUE)
attach(model1)
car_cat <- factor(cartype)
mod_full <- glm(y ~ agecat + car_cat + district + offset(log(n)), family=poisson, data=model1)
summary(mod_full)
```

```
##
## Call:
## glm(formula = y ~ agecat + car_cat + district + offset(log(n)),
##     family = poisson, data = model1)
##
## 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 ***
## agecat      -0.37628    0.04451  -8.453  < 2e-16 ***
## car_cat2     0.16223    0.05048   3.214 0.001309 **
## car_cat3     0.39535    0.05491   7.200 6.03e-13 ***
## car_cat4     0.56543    0.07215   7.836 4.64e-15 ***
## 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 παραπάνω summary φαίνεται πως όλες οι μεταβλητές σχετίζονται με τον αριθμό Υ των αποζημιώσεων αφού όλες είναι στατιστικά σημαντικές σύμφωνα με το wald test. Επιπρόσθετα, φαίνεται πως κάθε μεταβλητή συνεισφέρει, στη βελτίωση της Deviance σε σχέση με το null model. Στο Deviance Resid. φαίνεται το πόσο η συγκεκριμένη μεταβλητή μειώνει την ελεγχοσυνάρτηση Deviance όταν εισάγεται, ενώ στην στήλη Dev. φαίνεται το υπόλοιπο του συνολικού Deviance σε σχέση με το saturated μοντέλο.

Γενικά, η ελεγχοσυνάρτηση Deviance μετράει την απώλεια προσαρμογής σε σχέση με το saturated (κορεσμένο - πλήρες) μοντέλο το οποίο έχει ένα συντελεστή για κάθε δείγμα.

```
anova(mod_full, test = "Chisq")
```

```
## 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
## agecat    1   62.182        30    145.652 3.132e-15 ***
## car_cat   3   90.925        27     54.727 < 2.2e-16 ***
## district  1   12.938        26     41.789  0.000322 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Από τα παραπάνω βλέπουμε πως η μεταβλητή car_cat επιφέρει τη μεγαλύτερη βελτίωση στο μοντέλο καθώς μειώνει κατα 90.925 το Deviance της συνάρτησης όταν αυτή περιέχει το agecat μόνο σαν μεταβλητή.

BACKWARD STEPWISE SELECTION MODEL

Εξετάζεται αν η αφαίρεση μεταβλητών θα επιφέρει θετικά αποτελέσματα στο μοντέλο με τη χρήση backward stepwise selection και wald tests.

```
m1 = step(mod_full, direction="backward", test="Chisq")
```

```
## Start:  AIC=222.15
## y ~ agecat + car_cat + 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 ***
## - car_cat   3  131.713 306.07 89.925 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

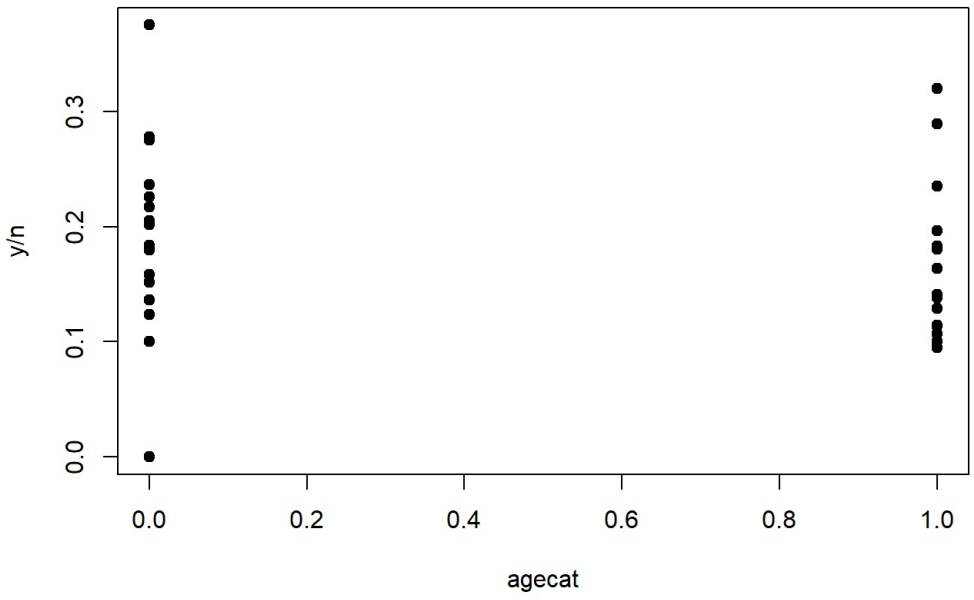
```
summary(m1)
```

```
##
## Call:
## glm(formula = y ~ agecat + car_cat + district + offset(log(n)),
##     family = poisson, data = model1)
##
## 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 ***
## agecat      -0.37628    0.04451  -8.453  < 2e-16 ***
## car_cat2     0.16223    0.05048   3.214 0.001309 **
## car_cat3     0.39535    0.05491   7.200 6.03e-13 ***
## car_cat4     0.56543    0.07215   7.836 4.64e-15 ***
## 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
```

Παρατηρούμε πως το μοντέλο με όλες τις μεταβλητές είναι το καταλληλότερο σύμφωνα με τον backward αλγόριθμο, με κριτήριο AIC = 222.15.

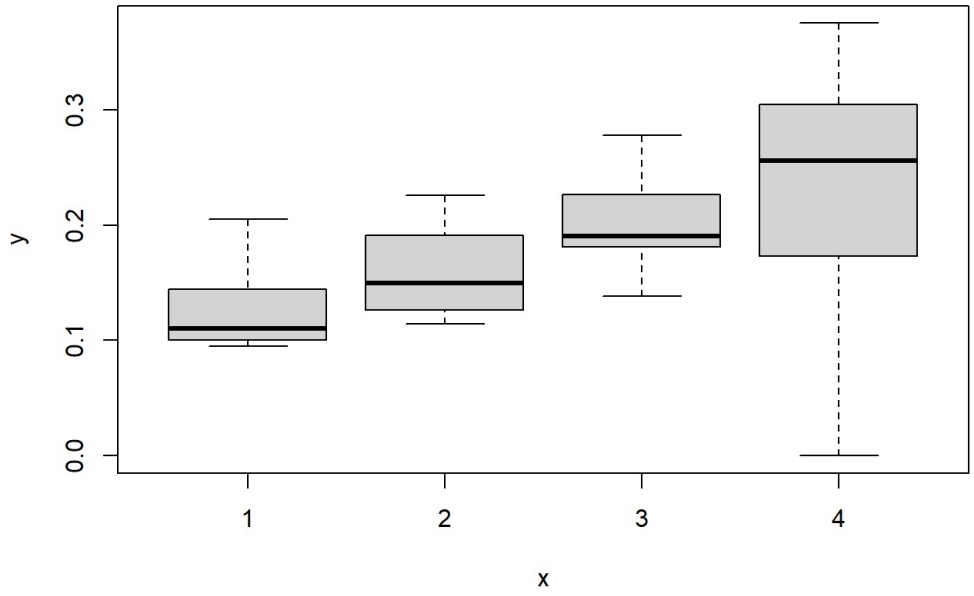
ΔΙΑΓΡΑΜΜΑ ΡΥΘΜΟΥ ΑΠΟΖΗΜΙΩΣΕΩΝ Y ~ ΚΑΤΗΓΟΡΙΑΣ ΗΛΙΚΙΩΝ

```
plot(agecat,y/n,pch=19)
```



ΔΙΑΓΡΑΜΜΑ ΡΥΘΜΟΥ ΑΠΟΖΗΜΙΩΣΕΩΝ Y ~ ΚΑΤΗΓΟΡΙΑΣ ΑΥΤΟΚΙΝΗΤΩΝ

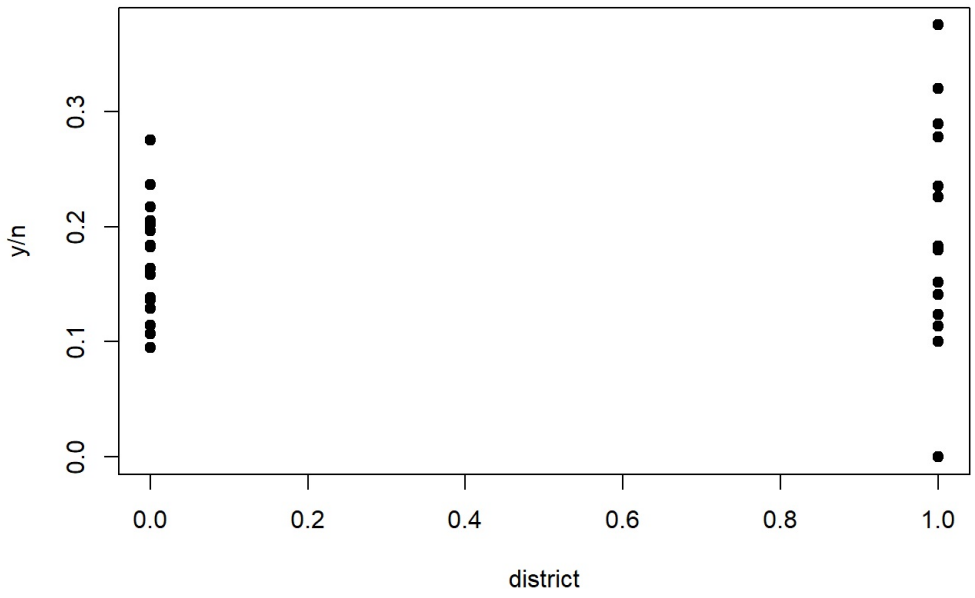
```
plot(car_cat,y/n,pch=19)
```



Στο παραπάνω διάγραμμα φαίνεται πως όσο αυξάνει το επίπεδο στην κατηγορία αμαξιών, ανεβαίνει παράλληλα και η median τιμή των αποζημιώσεων.

ΔΙΑΓΡΑΜΜΑ ΡΥΘΜΟΥ ΑΠΟΖΗΜΙΩΣΕΩΝ Y ~ ΕΠΑΡΧΙΑ/ΠΟΛΗ

```
plot(district,y/n,pch=19)
```



Έ ΕΛΕΓΧΟΣ DEVIANCE ΓΙΑ ΤΟ ΕΠΙΛΕΓΜΕΝΟ ΜΟΝΤΕΛΟ

Στη συνέχεια γίνεται χρήση της ελεγχοσυνάρτησης Deviance η οποία ακολουθεί την κατανομή Chi-square για τη σύγκριση της απώλειας που έχει το επιλεγμένο μοντέλο από το κορεσμένο (saturated).

```
pvalue <-  1 - pchisq(mod_full$deviance,mod_full$df.residual)
pvalue
```

```
## [1] 0.02580847
```

Γενικά, φαίνεται ότι το μοντέλο εξηγεί τη μεταβλητή y αρκετά αποδοτικά. Το μοντέλο παραπάνω έχει βαθμούς ελευθερίας 32-26 = 6 και η τιμή p δεν είναι ικανοποιητική, αλλά είναι πιο αξιόπιστο όταν γίνεται με βάση 2 άλλα μοντέλα και όχι οταν πραγματοποιείται σύγκριση με το saturated μοντέλο.

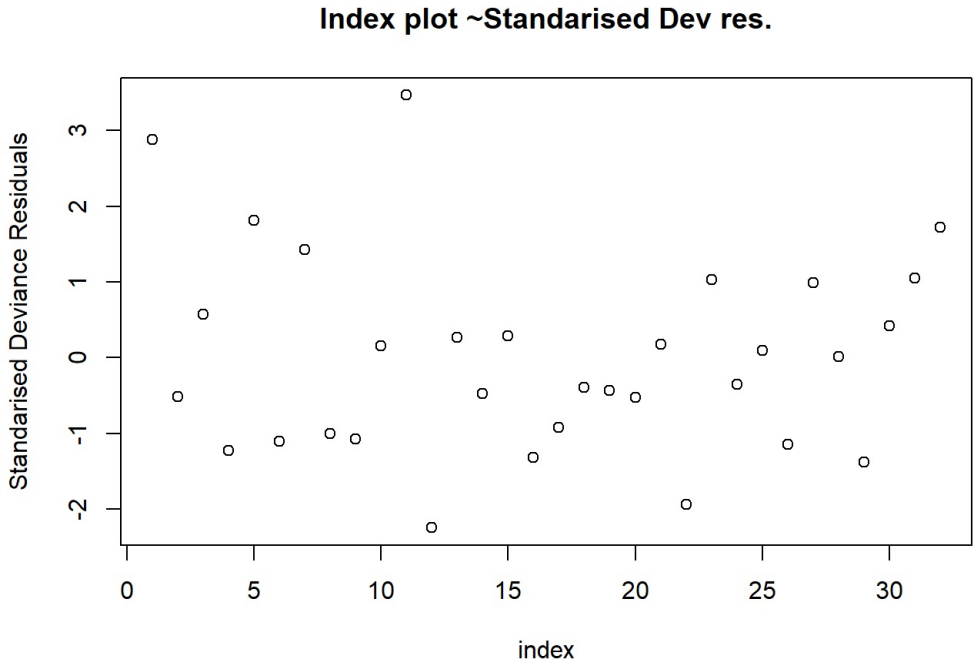
ΔΙΑΓΝΩΣΤΙΚΟΙ ΕΛΕΓΧΟΙ ~ ΠΡΟΣΑΡΜΟΓΗΣ ~ ΑΤΥΠΩΝ ΣΗΜΕΙΩΝ Αρχικά, έχουμε τα index plots των διαφόρων τύπων υπολοίπων ως προς τη σειρά παρατηρήσεων. Σε περίπτωση που εμφανίζονται ασυνήθιστα υπόλοιπα το μοντέλο δεν είναι ικανοποιητικό. Επιπλέον, υπάρχουν γραφικές παραστάσεις υπολοίπων έναντι κάθε συμμεταβλητής η του linear predictor (Χ'β) που μπορούν να αποβούν πολύ χρήσιμες στην εξέταση, ώστε είτε να συμπεριλάβουμε νέες μεταβλητές, είτε να μετασχηματιστεί μια υπάρχουσα. Οι γραφικές παραστάσεις που αναπτύσσονται συμβάλουν στον εντοπισμό των outliers. Τέλος, πραγματοποιείται και ο έλεγχος των υπολοίπων αν ακολουθούν την κανονική κατανομή για να φανεί πόσο καλά έχει προσαρμοστεί το μοντέλο. Υπολογίζονται τα τυποποιημένα σφάλματα pearson, deviance, likelihood.

```
res_pearson <- residuals(mod_full, type = "pearson")
standard_pearson_residuals <- res_pearson/(sqrt(1-hatvalues(mod_full)))

res_deviance <- residuals(mod_full, type = "deviance")
standard_res_deviance <- rstandard(mod_full)

res_lik <- sign(y - fitted.values(mod_full))*sqrt(hatvalues(mod_full)*standard_pearson_residuals^2 + (1-
hatvalues(mod_full))*standard_res_deviance^2)
```

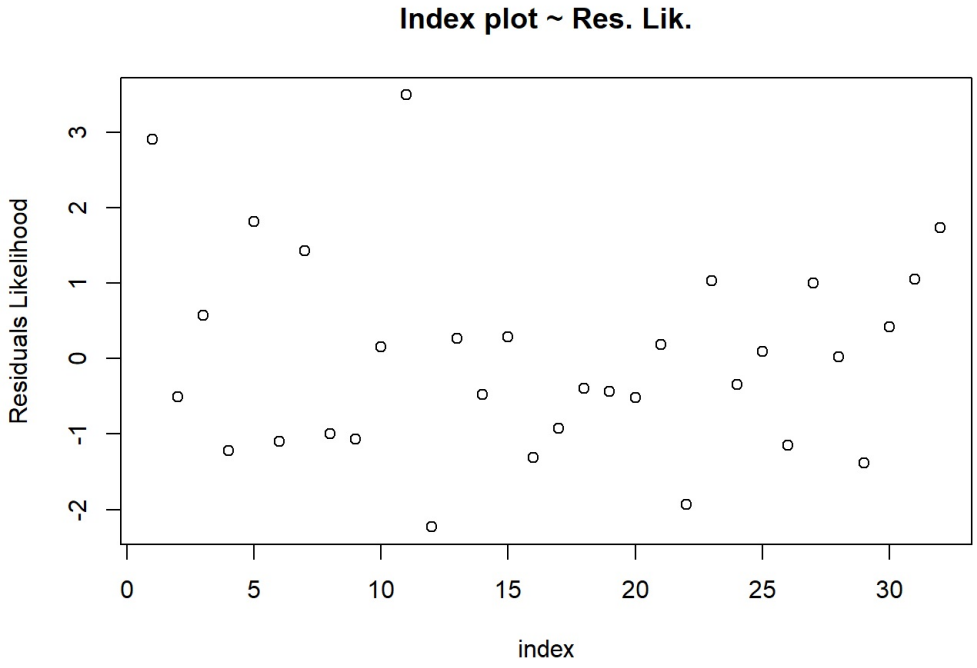
INDEX PLOTS

```
plot(standard_pearson_residuals, main='Index plot ~Standarised Pearson res.', xlab = 'index', ylab = 'Standarised
Pearson Residuals')
```


```
plot(standard_res_deviance, main='Index plot ~Standarised Dev res.', xlab = 'index', ylab = 'Standarised Deviance Residuals')
```



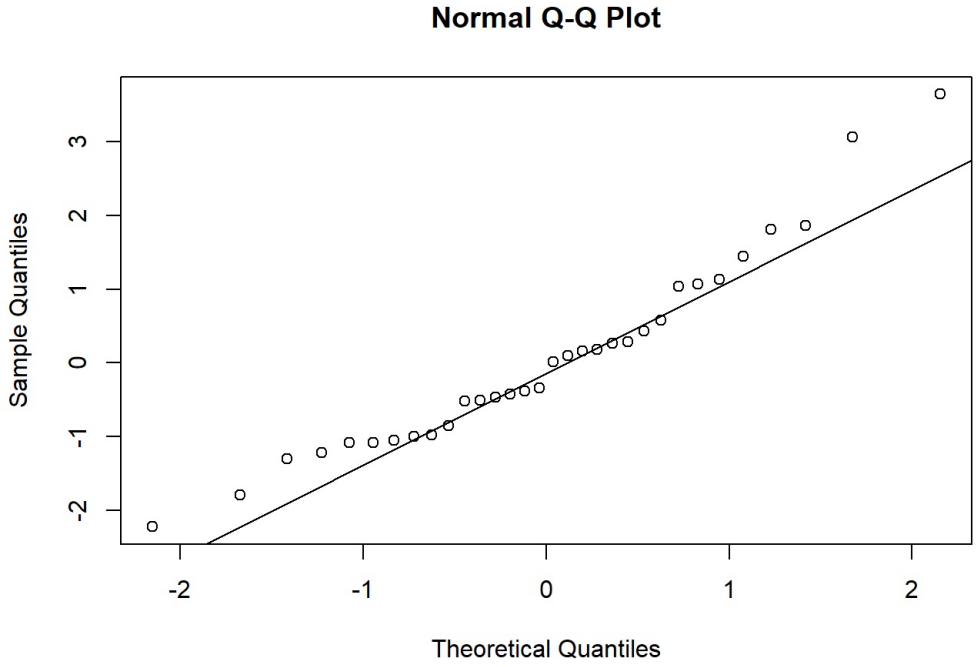

```
plot(res_lik , main='Index plot ~ Res. Lik.', xlab = 'index', ylab = 'Residuals Likelihood')
```



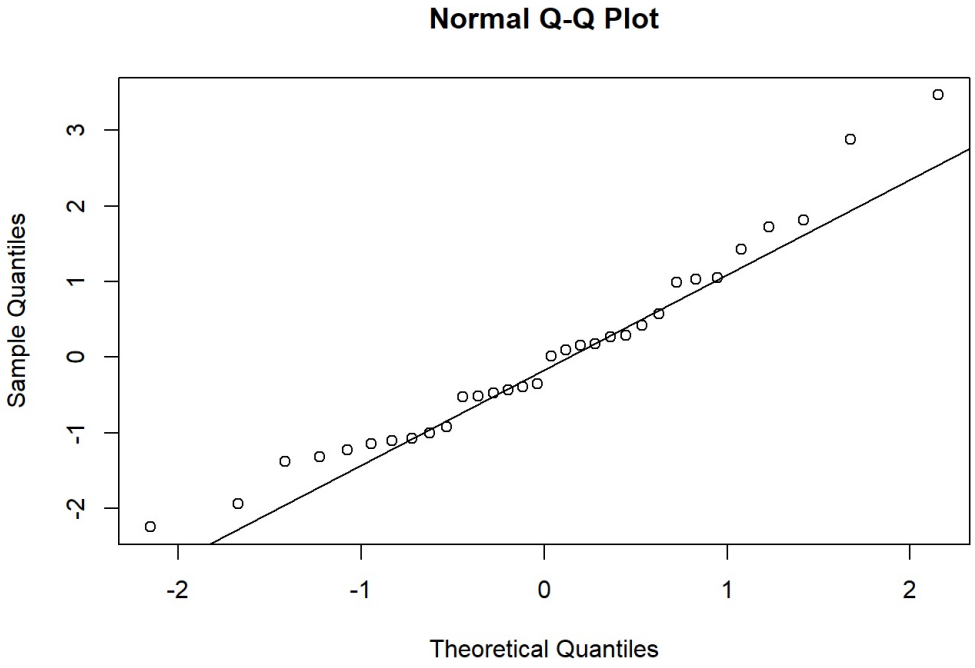
Φαίνεται ότι υπάρχουν άτυπα σημεία από το index plots.

QQ NORMAL PLOTS Σοο διάγραμμα NORMAL Q-Q PLOT παρατηρείται ότι αρκετά σημεία παρεκκλίνουν από την ευθεία, συνεπώς ενισχύεται ο ισχυρισμός της ύπαρξης άτυπων σημείων.

```
qqnorm(standard_pearson_residuals)
qqline(standard_pearson_residuals)
```

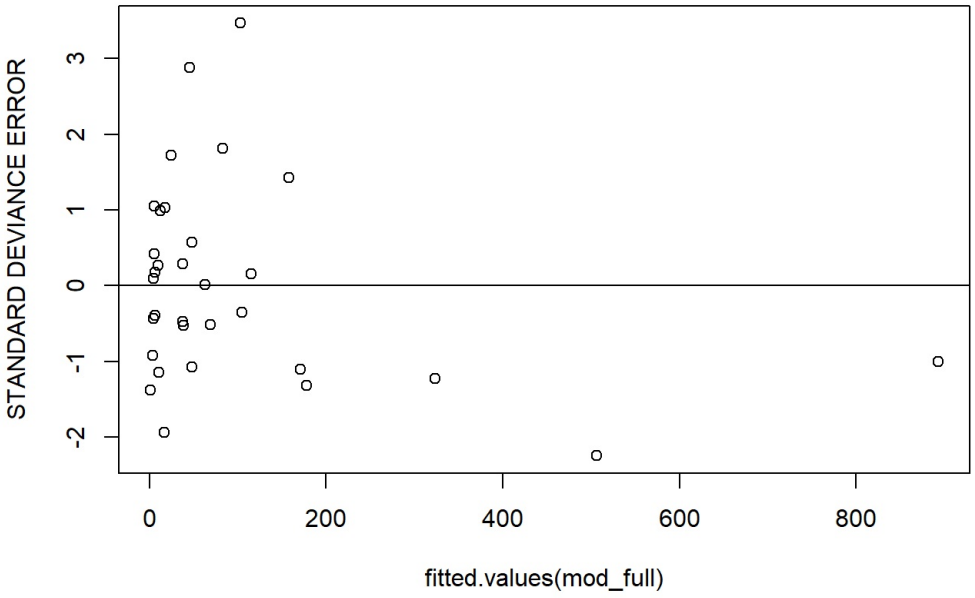



```
qqnorm(res_deviance/sqrt((1-hatvalues(mod_full))))
qqline(res_deviance/sqrt((1-hatvalues(mod_full))))
```

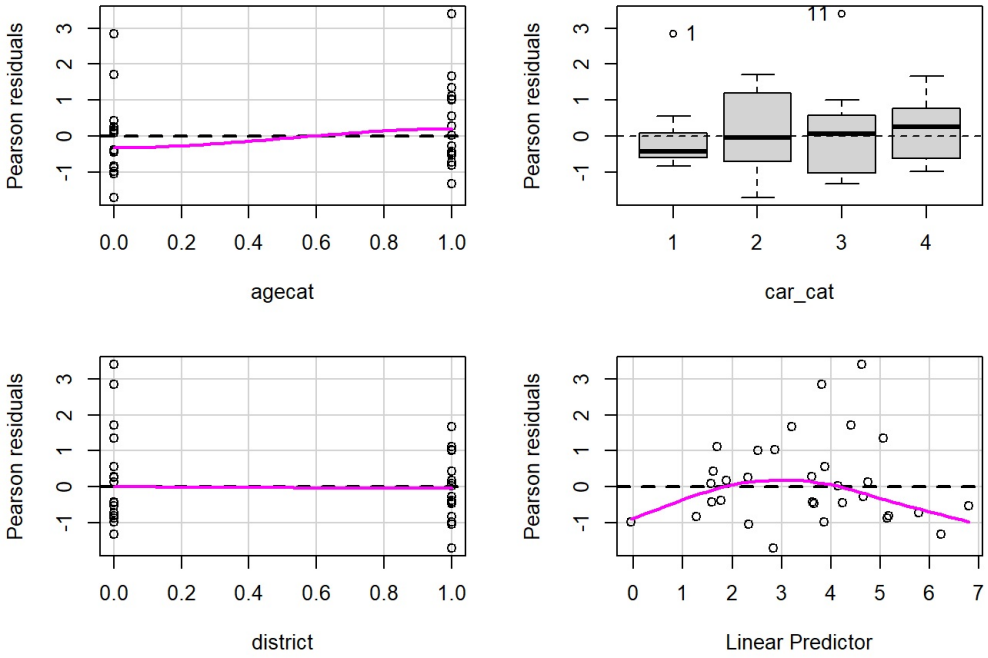


PLOTS RESIDUALS ~ COVARIATES AND LINEAR PREDICTOR(FITTED VALUES)

```
plot(fitted.values(mod_full), res_deviance/sqrt((1-hatvalues(mod_full))), ylab = "STANDARD DEVIANCE ERROR")
abline(h=0)
```




```
residualPlots(mod_full)
```




```
##          Test stat Pr(>|Test stat|)
## agecat           0                1
## car_cat
## district         0                1
```

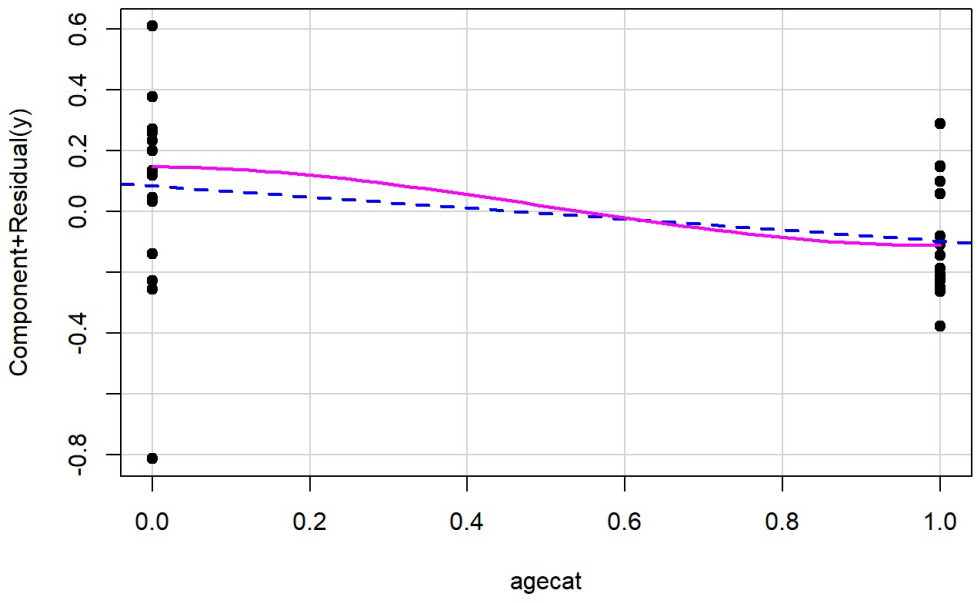
Από τα γραφήματα residuals~agecat και residuals~district παρουσιάζονται ορισμένα χρήσιμα αποτελέσματα καθώς δεν παρατηρείται κάποια συστηματικότητα ή συμπεριφορά. Το γράφημα του linear predictor είναι εμφανές πως δεν σκεδάζεται τυχαία αλλά έχουν μια συγκέντρωση προς το κέντρο ενώ εν συνεχεία στο τελευταίο γράφημα βλέπουμε πως πάνε κυρίως κάτω από το 0 οι τιμές.

CR PLOTS

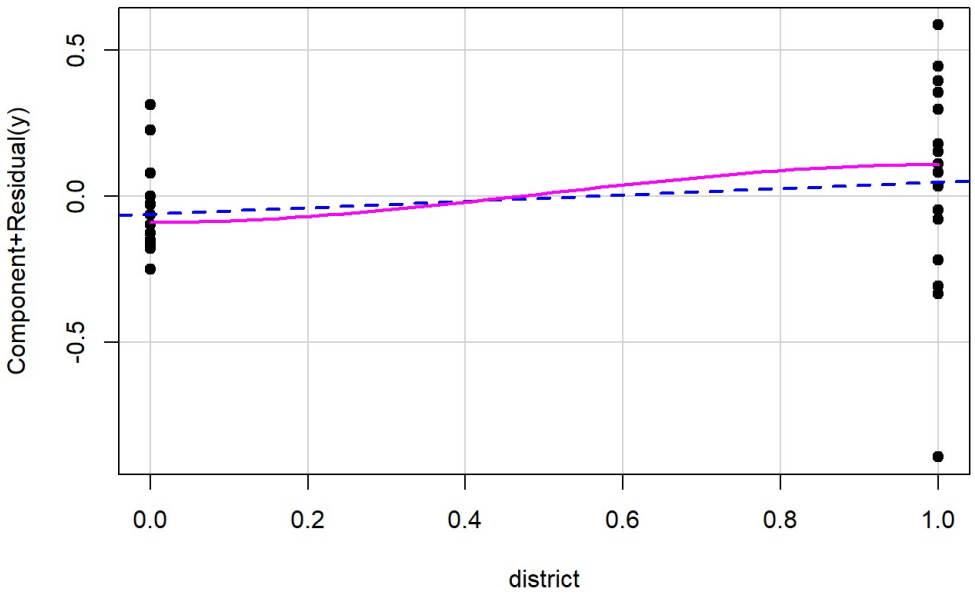
CR PLOTS

Στα παρακάτω διαγράμματα οι μεταβλητές είναι δύσκολο να μετασχηματιστούνε, για να βελτιώθεί η γραμμικότητά τους εφόσον είναι binary τιμές, ωστόσο φαίνεται να έχουν μια σημαντική γραμμικότητα,

```
crPlot(mod_full, variable=agecat, pch=19)
```




```
crPlot(mod_full, variable=district, pch=19)
```

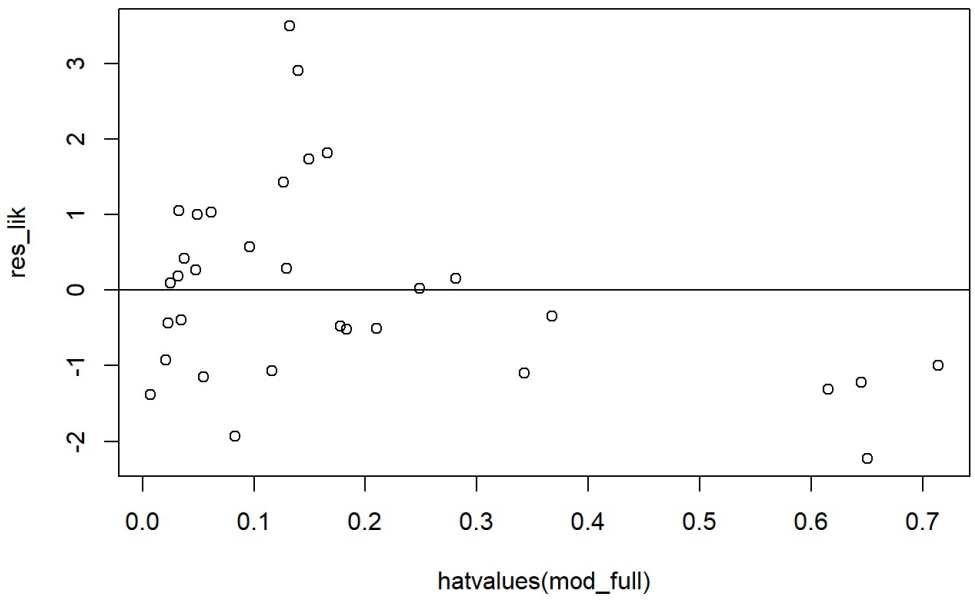


ΕΞΕΤΑΣΗ ΣΗΜΕΙΩΝ ΕΠΙΡΡΟΗΣ

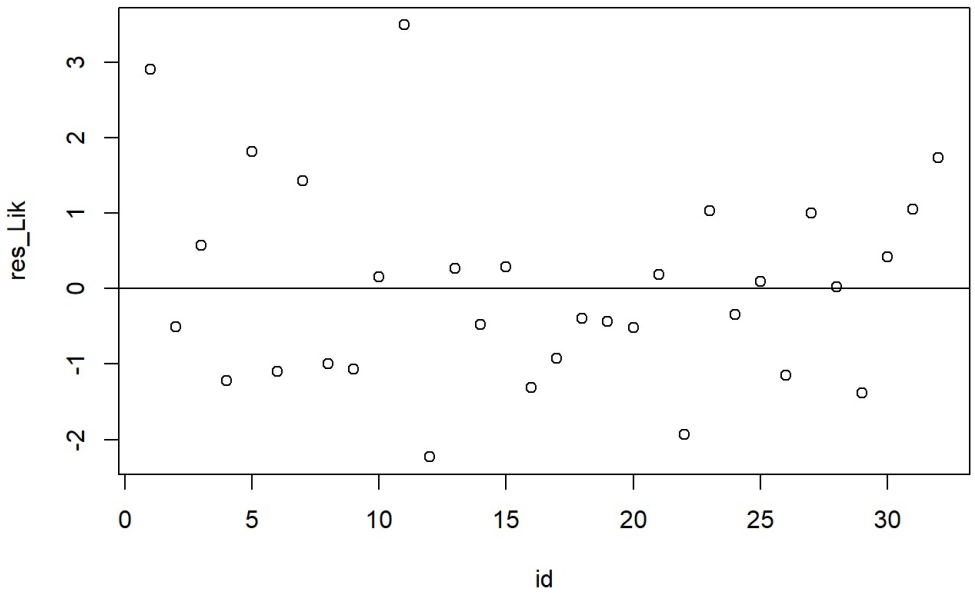
Στα πρώτα 2 διαγράμματα φαίνεται η ύπαρξη ορισμένων άτυπων σημείων. Πιο αναλυτικά στο διάγραμμα με τις αποστάσεις Cook η παρατήρηση 12 έχει cooks distance μεγαλύτερο του 1.

Όσον αφορά την μόχλευση ύποπτα θεωρούνται σημεία με τιμή μεγαλύτερη της 2p/n = 0.1875 και φαίνεται ότι υπάρχουν αρκετά τέτοια σημεία, όπως οι παρατηρήσεις 4, 8, 12, 16 και 24.

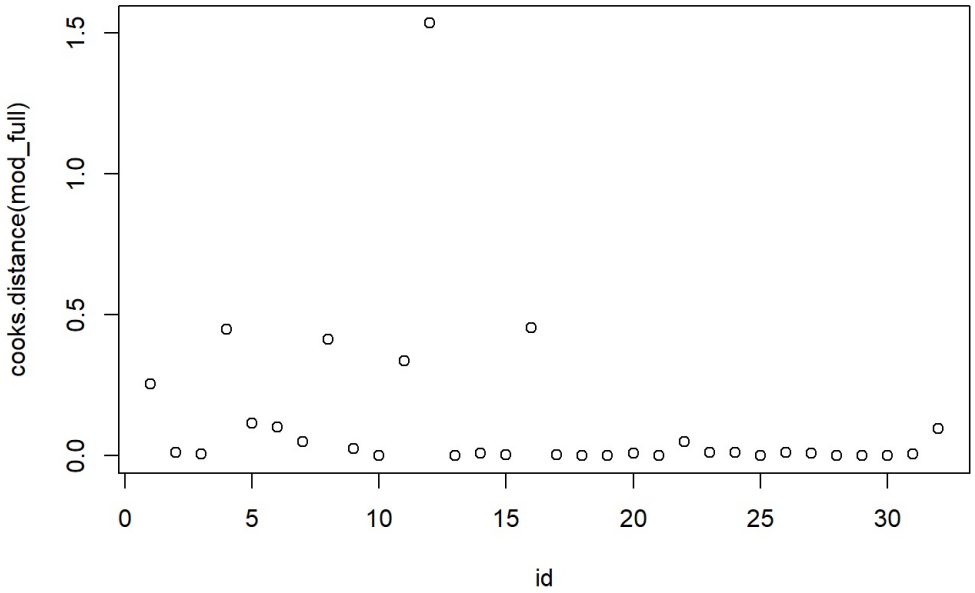
```
id = c(1:32) # observation's number

#hii probability balance plot
plot(hatvalues(mod_full),res_lik)
abline(h=0)
```



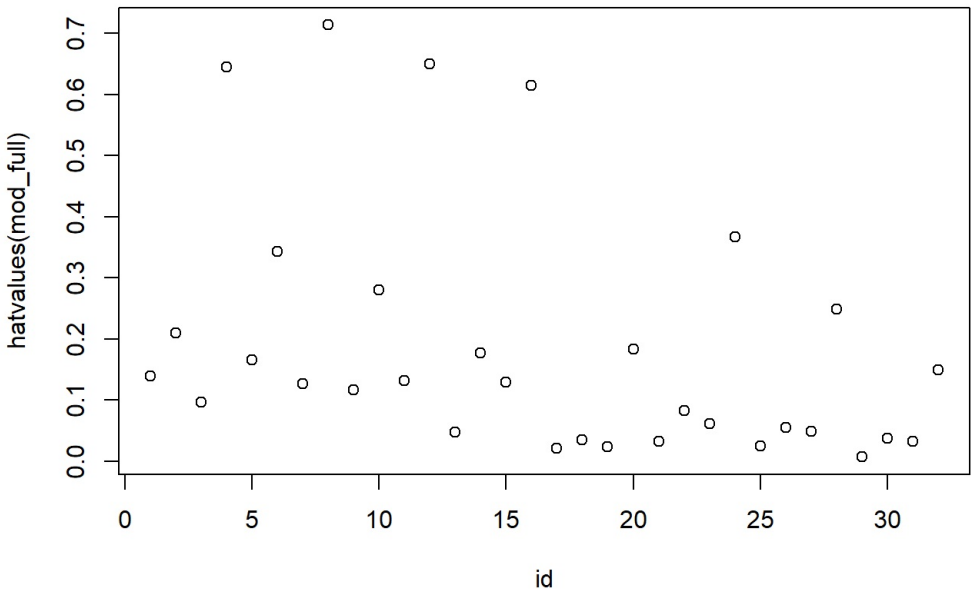

```
#plot for residual likelihood vs id
plot(id,res_lik,ylab = "res_Lik")
abline(h=0)
```




```
#cooks distance
plot(id,cooks.distance(mod_full))
```




```
#hat values
plot(id,hatvalues(mod_full))
```



Συγκεκριμένα τα άτυπα σημεία φαίνονται παρακάτω:

```
cooks.distance(mod_full)[which(cooks.distance(mod_full)>1)]
```

```
##       12
## 1.533511
```

```
hatvalues(mod_full)[which(hatvalues(mod_full)> 0.1875)]
```

```
##         2         4         6         8        10        12        16        24
## 0.2097697 0.6450818 0.3426689 0.7138452 0.2808653 0.6506745 0.6154150 0.3674769
##        28
## 0.2490220
```

Συμπερασματικά, το μοντέλο ίσως φαίνεται αδύναμο σε σχέση με το κορεσμένο λόγω αρκετών άτυπων σημείων. Αν αφαιρεθούν τα άτυπα σημεία τότε σίγουρα θα βελτιωθεί αρκετά η προγνωστική του ικανότητα.

DEVIANCE R SQUARED

Ο συντελεστής προσδιορισμού λαμβάνει τιμές μεταξύ 0 και 1 και αφορά το ποσοστό της μεταβλητότητας που εξηγείται από το συστηματικό μέρος του μοντέλου. Ισούται με 0,79.

```
1 - (mod_full$deviance / mod_full$null.deviance)
```

```
## [1] 0.7989323
```

Αν αφαιρεθούν τα άτυπα σημεία ελέγχουμε αν θα υπάρξει βελτίωση του συντελεστή προσδιορισμού.

```
model2 <- read.table("./asfalies_outliers.txt", header=TRUE)
attach(model2)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     id
```

```
## The following objects are masked from model1:
##
##     agecat, cartype, district, n, y
```

```
car_cat <- factor(cartype)
mod_no_outliers <- glm(y ~ agecat + car_cat + district + offset(log(n)), family=poisson, data=model2)


1 - (mod_no_outliers$deviance / mod_no_outliers$null.deviance)
```

```
## [1] 0.8148468
```

Όπως αναμενόταν η αφαίρεση των άτυπως σημείων αθξάνει την προγνωστική ικανότητα του μοντέλου.

ΕΡΜΗΝΕΙΑ ΣΥΝΤΕΛΕΣΤΩΝ ΤΟΥ ΕΠΙΛΕΓΜΕΝΟΥ ΜΑΣ ΜΟΝΤΕΛΟΥ

agecat coefficient: −0.37628

Ο συντελεστής της κατηγορίας ηλικίας είναι αρνητικός, γεγονός που δείχνει ότι όσο αυξάνεται η ηλικία μειώνεται ο αριθμός των αποζημιώσεων λόγω τροχαίων ατυχημάτων. Ουσιαστικά, αν έχουμε σταθερές όλες τις άλλες μεταβλητές, ο αριθμός των αποζημιώσεων για την ηλικιακή κατηγορία 1 είναι 0.686 φορές ο αντίστοιχος της κατηγορίας 0. Αυτό σημαίνει ότι όταν από την κατηγορία 0 που είναι οι μικρότερες ηλικίες μεταβούμε στην κατηγορία 1 που είναι οι μεγαλύτερες, η μεταβλητή των αποζημιώσεων πολλαπλασιάζεται με το e^-0.37, που σημαίνει πως η μεταβλητή y ελαττώνεται.

district coefficient: 0.21661

Ο συντελεστής της κατηγορίας για τις περιοχές είναι θετικός που σημαίνει ότι λειτουργεί αυξητικά στην εξαρτημένη μεταβλητή μας όταν μεταβαίνουμε από άλλες πόλεις (0) στην Αθήνα (1). Ο αριθμός αποζημιώσεων για την ομάδα στην περιοχή της Αθήνας είναι e^0.21661=1.242 φορές ο αντίστοιχος αριθμός εξαρτημένης μεταβλητής στην άλλη περιοχή.

cartype coefficient ~ κατηγορηματική μεταβλητή - 4 levels - 3 dummy variables:

Η συγκεκριμένη μεταβλητή ουσιαστικά Θέτει ως σημείο αναφοράς το 1ο επίπεδο ενώ τα άλλα επίπεδα δηλώνουν τη μεταβολή από το επίπεδο αναφοράς στο αντίστοιχα δικό τους επίπεδο. Παρακάτω φαίνονται οι συντελεστές για το κάθε επίπεδο έχοντας ως αναφορά το 1ο. 1->2: 0.16223 1->3: 1.485 1->4: 1.760

Το συμπέρασμα είναι πως το επίπεδο και το rate αυξάνονται ανάλογα.

Άσκηση 2 ~ Logistic

Η εξαρτημένη μεταβλητή είναι η response ενώ οι συμμεταβλητές είναι οι smear, infiltrate, index, blasts, temperature. Στην εκφώνηση αναφέρεται ότι ισχύει n=1, συνεπώς τα δεδομένα αντί για διωνυμικά είναι δυαδικά. Αρχικά δημιουργούμε ένα μοντέλο λογιστικής παλινδρόμησης.

```
#data reading
library(MASS)
library(car)
```

```
## Loading required package: carData
```

```
mydata = read.table("./leukaemia.txt", header=TRUE)
attach(mydata)
# logistic regression model
logmod1 <- glm(response ~ age + smear + infiltrate + index + blasts + temperature, family=binomial,data=mydata)

summary(logmod1)
```

```
##
## Call:
## glm(formula = response ~ age + smear + infiltrate + index + blasts +
##     temperature, family = binomial, data = mydata)
##
## 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
##
## Number of Fisher Scoring iterations: 6
```

Φαίνεται από το wald test ότι οι μεταβλητές smear, infiltrate και blasts έχουν υψηλό p-value, συνεπώς δεν είναι στατιστικά σημαντικές και δεν έχουν ισχύ στην μεταβλητή θεραπείας.

```
cor(mydata)
```

```
##                    age       smear   infiltrate       index      blasts
## age          1.00000000 -0.20378215 -0.136998888 -0.12425459 -0.07054114
## smear       -0.20378215  1.00000000  0.847132591  0.10269246  0.40390824
## infiltrate  -0.13699889  0.84713259  1.000000000  0.14437713  0.38057840
## index       -0.12425459  0.10269246  0.144377132  1.00000000  0.28791696
## blasts      -0.07054114  0.40390824  0.380578397  0.28791696  1.00000000
## temperature  0.08458907 -0.02824923 -0.006709947  0.07052914  0.33131167
## response    -0.35012742  0.21526022  0.264831006  0.48820317  0.13211841
##              temperature   response
## age          0.084589073 -0.3501274
## smear       -0.028249230  0.2152602
## infiltrate  -0.006709947  0.2648310
## index        0.070529145  0.4882032
## blasts       0.331311665  0.1321184
## temperature  1.000000000 -0.2647615
## response    -0.264761500  1.0000000
```

Από τον πίνακα συσχέτισης παρατηρείται ότι υπάρχει συσχέτιση μεταξύ ορισμένων μεταβλητών. Η μεταβλητή smear με τη μεταβλητή infiltrate συσχετίζονται σε αρκετά μεγάλο βαθμό με συσχέτιση της τάξεως 0.8. Αυτός ο βαθμός συσχέτισης ερμηνεύεται, αφού το ποσοστό κυττάρων στο μυελό των οστών, επηρεάζει το ποσοστό επίστρωσης βλαστοκυττάρων.

Γενικότερα στην περίπτωση όπου υπάρχουν binary data (ni=1) η Deviance δεν παρέχει πληροφορίες σχετικά με την προσαρμογή του μοντέλου αφού υπάρχει εξάρτηση μόνο από τις εκτιμώμενες τιμές μi. Όμως, η Deviance μπορεί να χρησιμοποιηθεί για τη σύγκριση των δύο μοντέλων ώστε να ερευνηθεί τη μεταβολήτά της, η οποία ακολουθεί την κατανομή chi square.

BACKWARD STEPWISE SELECTION WITH AIC CRITERION

```
backw = step(logmod1, direction="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
```

```
summary(backw)
```

```
##
## Call:
## glm(formula = response ~ age + infiltrate + index + temperature,
##     family = binomial, data = mydata)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.73886  -0.56473  -0.05442   0.62185   2.26516
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 95.56766   38.59482   2.476  0.01328 *
## age         -0.06026    0.02678  -2.250  0.02445 *
## infiltrate   0.03413    0.02079   1.641  0.10077
## index        0.40673    0.13034   3.121  0.00181 **
## temperature -0.09944    0.03954  -2.515  0.01191 *
## ---
## 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: 40.136  on 46  degrees of freedom
## AIC: 50.136
##
## Number of Fisher Scoring iterations: 6
```

Η παραπάνω διαδικασία οδήγησε σε μοντέλο που αφορά τις μεταβλητές index, temperature, age, infiltrate. Οι μεταβλητές smear και blasts εξαιρέθηκαν καθώς από ότι φαίνεται είχαν συσχέτιση με την infiltrate, επομένως δεν προσέφεραν κάποια επιπλέον πληροφορία. Επίσης, από το summary βλέπουμε πως όλες οι μεταβλητές είναι στατιστικά σημαντικές εκτός της infiltrate. Όμως η αφαίρεση της θα επέφερε αύξηση στο συντελεστή του AIC.

```
anova(backw,logmod1,test="Chisq")
```

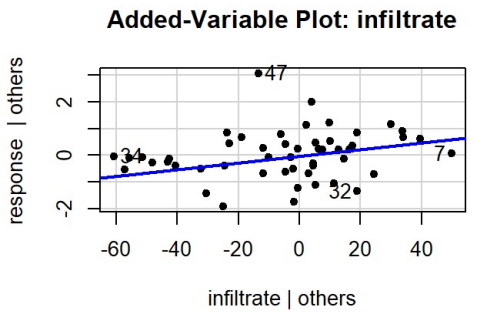
```
## Analysis of Deviance Table
##
## Model 1: response ~ age + infiltrate + index + temperature
## Model 2: response ~ age + smear + infiltrate + index + blasts + temperature
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        46     40.136
## 2        44     40.060  2 0.076321   0.9626
```

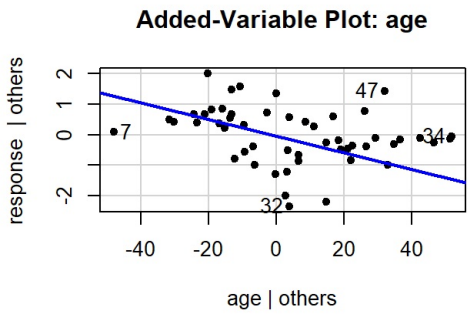
Το backward μοντέλο είναι όμοιο με το μοντέλο που αναπτύχθηκε αρχικά και δεν προκύπτουν ιδιαίτερες διαφορές. Επομένως, το μοντέλο με τις λιγότερες μεταβλητές είναι η αποδιτικότερη επιλογή, διότι έχει καλύτερο AIC.

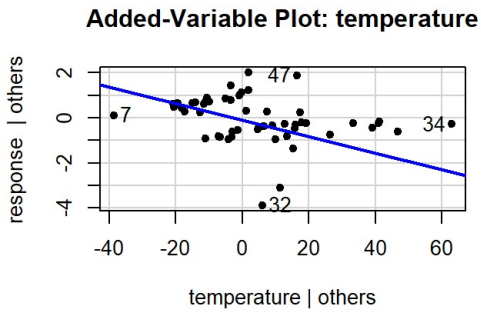
AV PLOTS

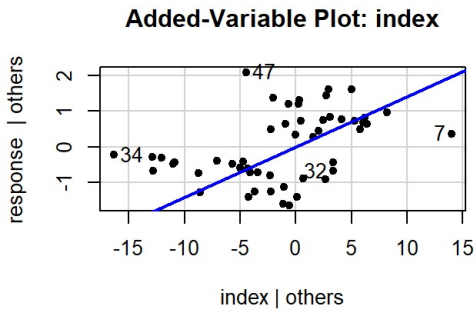
Στα av plots φαίνεται ορισμένα outliers, αλλά οι περισσότερες μεταβλητές έχουν στοιχεία κοντά στη γραμμή. Επομένως, και όλες οι μεταβλητές(4) χρειάζονται για την επεξήγηση της y.

```
library(car)
par (mfrow=c(2,2))
avPlot(backw, variable=infiltrate, pch=19)
avPlot(backw, variable=age, pch=19)
avPlot(backw, variable=temperature, pch=19)
avPlot(backw, variable=index, pch=19)
```





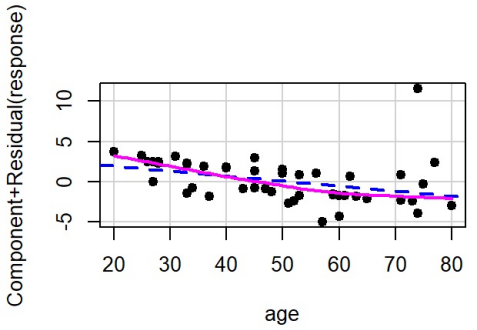


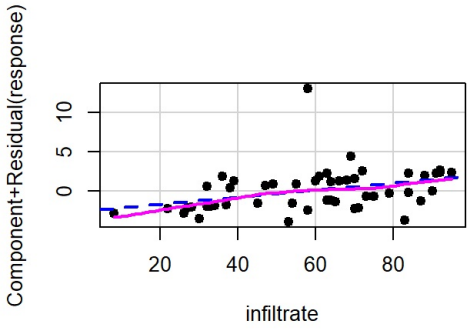


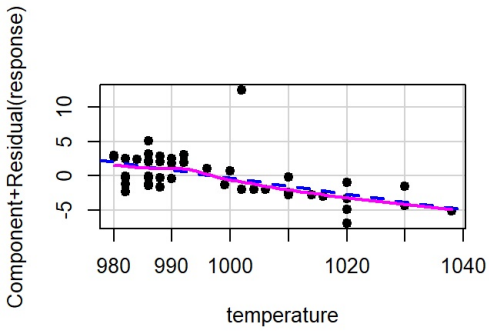
CR PLOTS

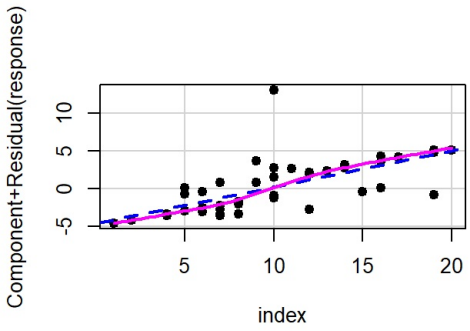
Τα cr plots δείχνουν πως όλες οι μεταβλητές παρουσιάζουν μια ευθύγραμμη τάση.

```
library(car)
par (mfrow=c(2,2))
crPlot(backw, variable=age, pch=19)
crPlot(backw, variable=infiltrate, pch=19)
crPlot(backw, variable=temperature, pch=19)
crPlot(backw, variable=index, pch=19)
```









Μετασχηματισμός της index μεταβλητής.

ΒΕΛΤΙΣΤΟ ΜΟΝΤΕΛΟ

Λογαρίθμηση παραμέτρου index.

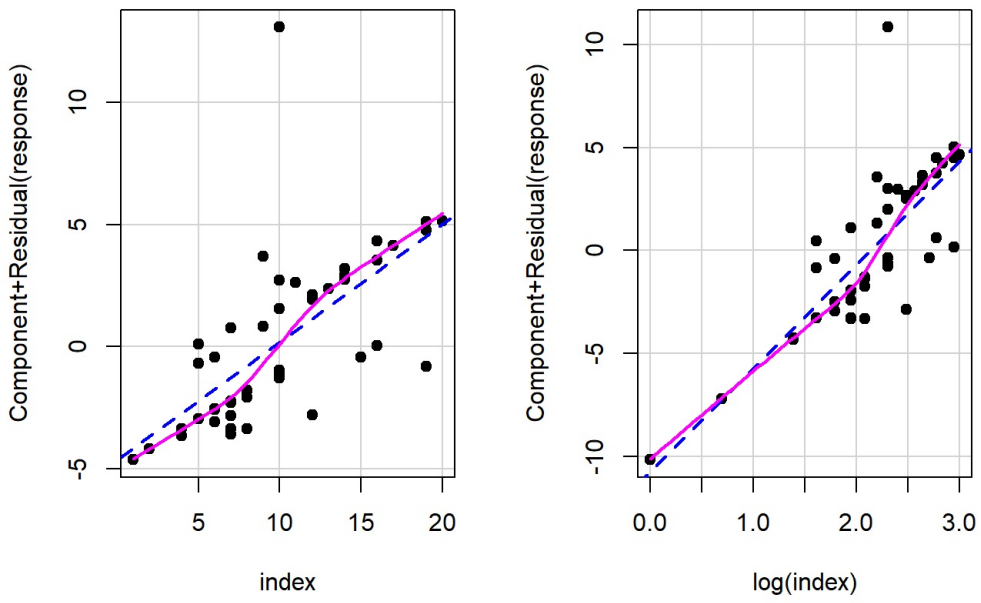
```
fit <- glm(response ~ age  + log(index)  + temperature + infiltrate , family=binomial, data=mydata)
summary(fit )
```

```
##
## Call:
## glm(formula = response ~ age + log(index) + temperature + infiltrate,
##     family = binomial, data = mydata)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.71466  -0.46372  -0.01814   0.56454   2.15254
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 88.58023   37.63528   2.354  0.01859 *
## age         -0.06267    0.02756  -2.274  0.02297 *
## log(index)   4.27521    1.44334   2.962  0.00306 **
## temperature -0.09809    0.03989  -2.459  0.01393 *
## infiltrate   0.03938    0.02165   1.819  0.06894 .
## ---
## 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: 39.118  on 46  degrees of freedom
## AIC: 49.118
##
## Number of Fisher Scoring iterations: 6
```

Από το παραπάνω summary φαίνεται πως υπάρχει μείωση στο AIC και πως όλες οι μεταβλητές είναι στατιστικά σημαντικές, εκτός της infiltrate.

Παρακάτω θα δούμε τα Component residual plots για τη μετασχηματισμένη μεταβλητή, ωστέ να δούμε αν υπήρξε βελτίωση στη γραμμική της τάση.

```
par (mfrow=c(1,2))
crPlot(backw, variable=index, pch=19)
crPlot(fit, variable=log(index), pch=19)
```



Ο συγκεκριμένος μετασχηματισμός έχει εμφανώς σημαντικότερη απόδοση.

ΑΦΑΙΡΕΣΗ INFILTRATE

Με την αφαίρεση της μεταβλητής infiltrate, όλες οι μεταβλητές του μοντέλου είναι στατιστικά σημαντικές αλλά αυξάνεται το AIC και θα μειώνεται η προσαρμοστικότητα του μοντέλου. Παρακάτω φαίνεται πως η πρόσθεση της μεταβλητής infiltrate μείωνει τη deviance κατά 3,986 σε σχέση με το μοντέλο χωρίς αυτήν. Η p-value τιμή είναι μικρότερη του 0.05 επομένως μπορούμε να απορρίψουμε το εμφωλευμένο μοντέλο.

```
fit_new <- glm(response ~ age  + log(index)  + temperature , family=binomial, data=mydata)
summary(fit_new)
```

```
##
## Call:
## glm(formula = response ~ age + log(index) + temperature, family = binomial,
##     data = mydata)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.74904  -0.69474  -0.05176   0.70981   2.01278
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 76.44403   33.33945   2.293  0.02185 *
## age         -0.05883    0.02548  -2.308  0.02098 *
## log(index)   3.80893    1.27036   2.998  0.00271 **
## temperature -0.08258    0.03487  -2.368  0.01787 *
## ---
## 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: 43.104  on 47  degrees of freedom
## AIC: 51.104
##
## Number of Fisher Scoring iterations: 6
```

```
anova(fit_new,fit,test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: response ~ age + log(index) + temperature
## Model 2: response ~ age + log(index) + temperature + infiltrate
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        47     43.104
## 2        46     39.118  1    3.986  0.04588 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Συνεπώς επιλέγεται το μοντέλο με 4 μεταβλητές. Η εξίσωση για τη μεταβλητή response σύμφωνα με το βέλτιστο μοντέλο γράφεται ως εξής: $p = \frac{e^{88,58023 - 0,06267 * age + 4,27521 * log(index) + 0,03938 * infiltrate - 0,09809 * temperature}}{e^{88,58023 - 0,06267 * age + 4,27521 * log(index) + 0,03938 * infiltrate - 0,09809 * temperature}}$

ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ ΓΙΑ ΤΑ $\hat{\beta}$

```
# confidence interval of model
confint(fit)
```

```
## Waiting for profiling to be done...
```

```
##                     2.5 %       97.5 %
## (Intercept) 22.0148022166 173.03679004
## age         -0.1252552435  -0.01387272
## log(index)   1.8828660546   7.65151411
## temperature -0.1883546554  -0.02842773
## infiltrate   0.0006735671   0.08734412
```

```
#confidence interval for odds model p/1-p
exp(confint(fit))
```

```
## Waiting for profiling to be done...
```

```
##                    2.5 %       97.5 %
## (Intercept) 3.638372e+09 1.409039e+75
## age         8.822717e-01 9.862231e-01
## log(index)  6.572315e+00 2.103829e+03
## temperature 8.283209e-01 9.719725e-01
## infiltrate  1.000674e+00 1.091272e+00
```

Παρουσιάζονται τα διαστήματα εμπιστοσύνης της τάξεως του 95%, για τις παραμέτρους του μοντέλου αλλά και για τα αντίστοιχα διαστήματα των e^{B_j}

ΕΡΜΗΝΕΙΑ ΣΥΝΤΕΛΕΣΤΩΝ Η ποσότητα e^{B_j} είναι ο παράγοντας που πολλαπλασιάζεται με τη σχετική πιθανότητα (odds) πραγματοποίησης του γεγονότος, όταν η ανεξάρτητη μεταβλητή X_j αυξάνεται κατά μια μονάδα, δεδομένου ότι οι υπόλοιπες μεταβλητές παραμένουν σταθερές.

Αν το B_j > 0 τότε το e^{B_j} > 1 που σημαίνει πως ο λόγος odds αυξάνεται, ενώ αντίθετα εάν B_j < 0 τότε το e^{B_j} < 1 που σημαίνει πως ο λόγος odds μειώνεται.

Στην παρούσα ανάλυση για τη μεταβλητή *age* υπάρχει αρνητικός συντελεστής επομένως, αν η age αυξηθεί κατά μια μονάδα τότε η σχετική απόκριση στη θεραπεία πολλαπλασιάζεται με το 0.93925337616 < 1. Επομένως όσο αυξάνεται η ηλικία, η πιθανότητα για απόδοση της θεραπείας μειώνεται το οποίο έχει μία λογική.

Η μεταβλητή *temperature* περιέχει επίσης αρνητικό συντελεστή συνεπώς, αύξηση της θερμοκρασίας συνεπάγεται στο ότι η πιθανότητα για απόδοση της θεραπείας πολλαπλασιάζεται με 0.90656730902 <1.

Η μεταβλητή *log(index)* αφορά τον δείκτη των κυττάρων λευχαιμίας. Αν αυξηθεί κατά μια μονάδα, η πιθανότητα για απόδοση στη θεραπεία βελτιώνεται καθώς πολλαπλασιάζεται με όρο μεγαλύτερο του 1 , συγκεκριμένα τον 71.895235724.

Η μεταβλητή *infiltrate* έχει θετικό συντελεστή επομένως, η αύξηση κατά μια μονάδας της τιμής αυτής θα πολλαπλασιάσει την πιθανότητα ανταπόκρισης στη θεραπεία με το συντελεστή 1.04016567151 > 1.

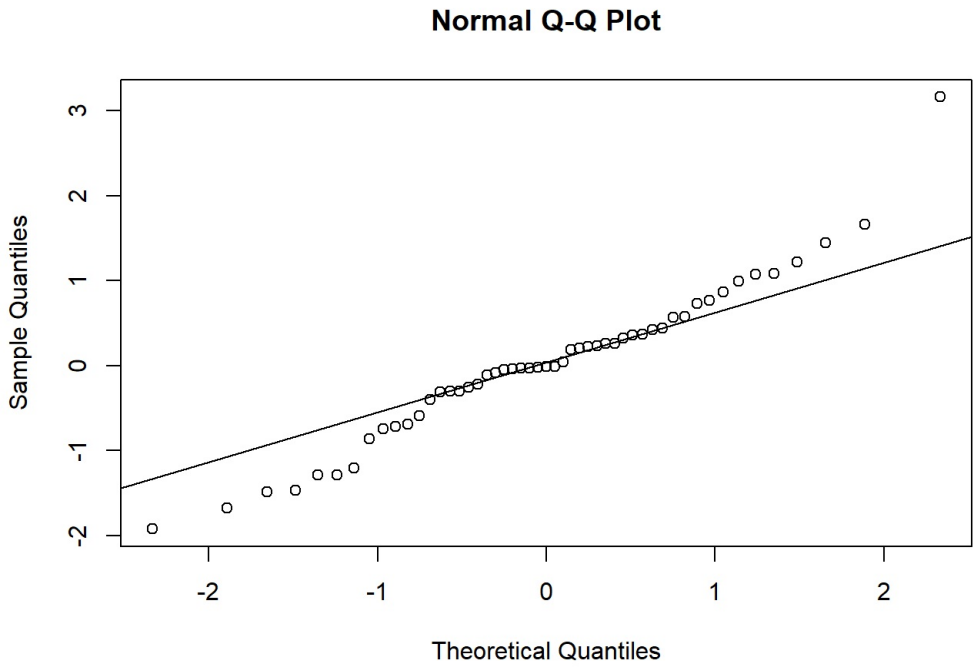
Τέλος για το *intercept coefficient* δεν δύναται να δοθεί συγκεκριμένη ερμηνεία καθώς ποτέ οι συμμεταβλητές δε θα γίνουν όλες μαζί 0.

ΔΙΑΓΝΩΣΤΙΚΟΙ ΕΛΕΓΧΟΙ -ΠΡΟΣΑΡΜΟΓΗ- ΣΗΜΕΙΑ ΕΠΙΡΡΟΗΣ

RESIDUALS QQ PLOT

Έχοντας υπολογίσει τα residuals αναπτύσσονται στη συνέχεια ορισμένα διαγράμματα που αφορούν τα σφάλματα της κανονικής κατανομή. Τα σημεία που παρεκκλίνουν της ευθείας αφορούν πιθανά outliers. Βάση του αν τα σημεία ακολουθούν ή όχι μπορούμε να καταλάβουμε πόσο καλά έχει προσαρμοστεί το μοντέλο.

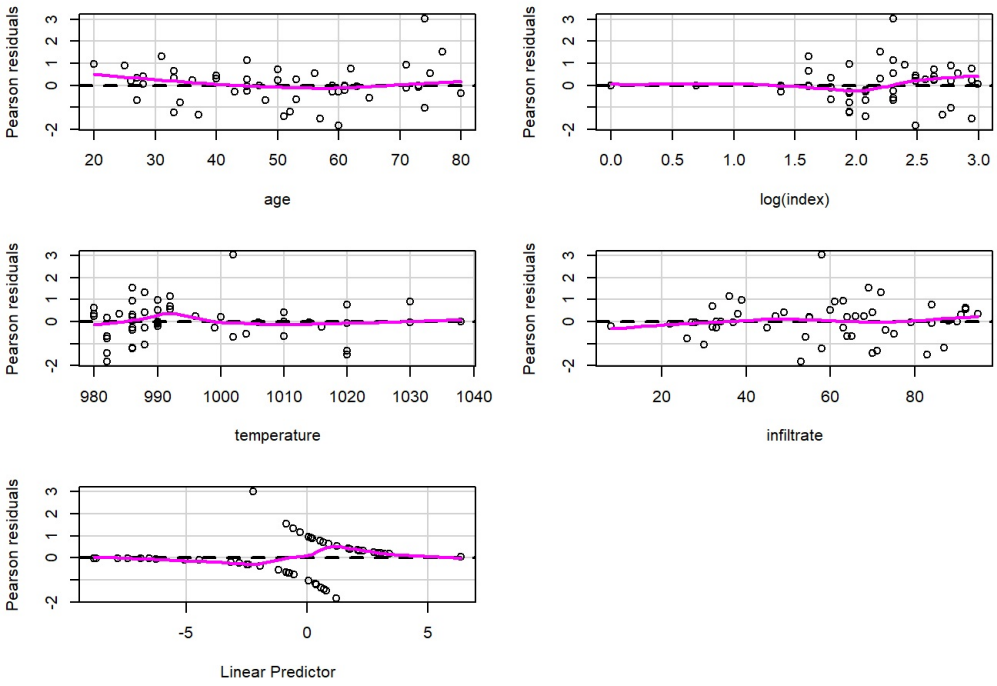
```
qqnorm(stand.pearson.residuals)
qqline(stand.pearson.residuals)
```



Φαίνεται σε ορισμένα σημεία να υπάρχει παραβίαση της κανονικότητας.

PLOTS WITH LINEAR PREDICTOR AND EACH COVARIATE

```
library(car)
residualPlots(fit)
```

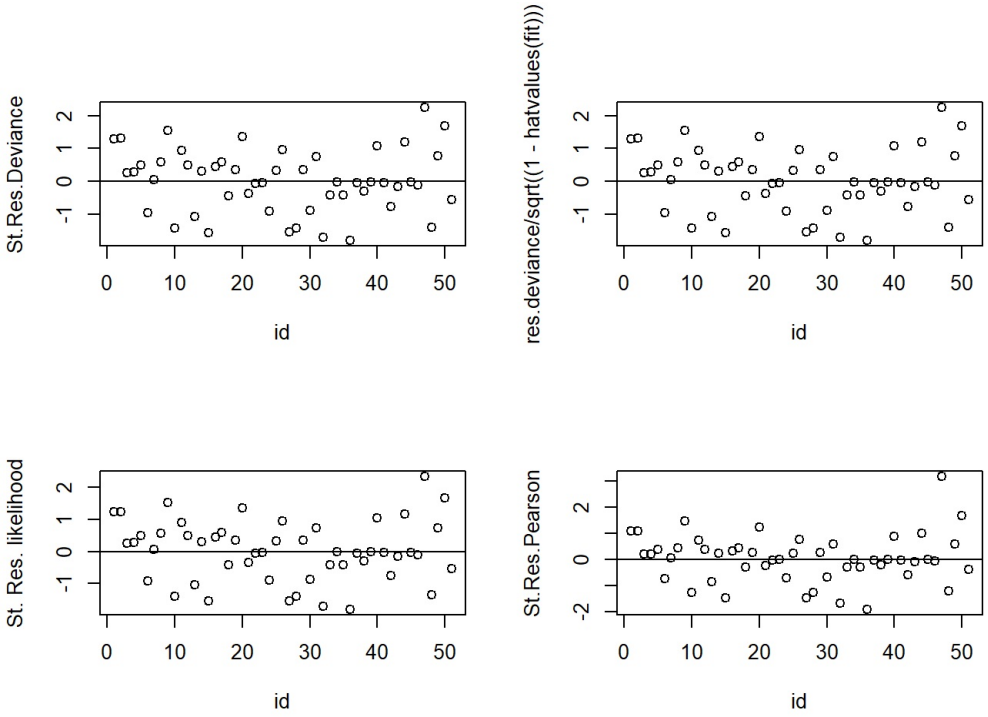



```
##             Test stat Pr(>|Test stat|)
## age            5.5315          0.01868 *
## log(index)     0.0000          0.99944
## temperature    0.6284          0.42794
## infiltrate     0.0415          0.83852
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

INDEX PLOT

Από το διάγραμμα τυποποιημένων deviance υπολοίπων με βάση το id, παρατηρούμε πως τα υπόλοιπα κατανέμονται τυχαία γύρω από το 0, επομένως οι παρατηρήσεις είναι ανεξάρτητες μεταξύ τους, αφού δεν παρουσιάζουν κάποια ιδιαίτερη συμπεριφορά.

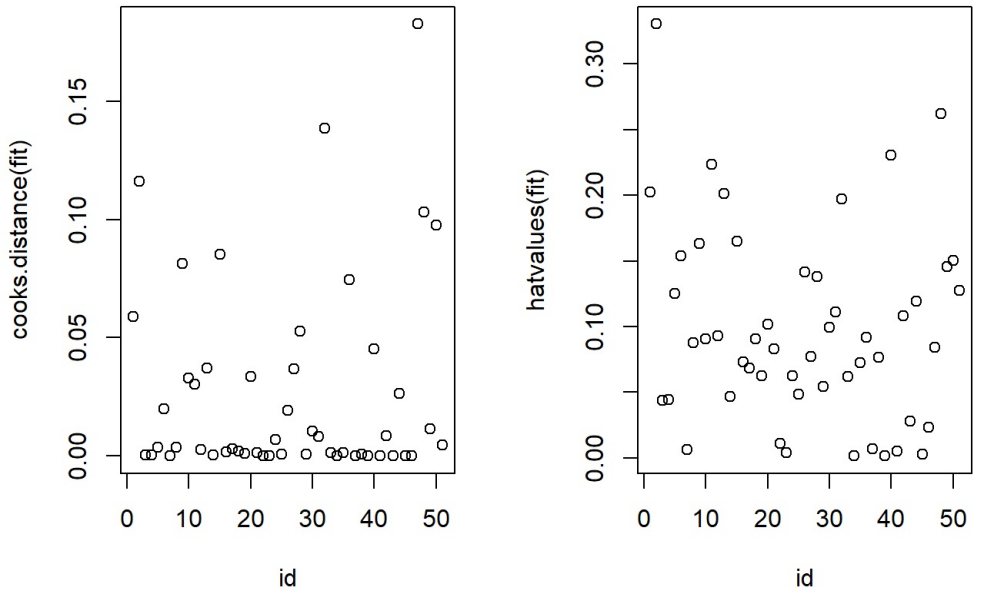
```
par(mfrow = c(2, 2))
id = c(1:51)
plot(id,res.deviance/sqrt((1-hatvalues(fit))),ylab="St.Res.Deviance")
abline(h=0)


plot(id,res.deviance/sqrt((1-hatvalues(fit))))
abline(h=0)

plot(id, res.lik, ylab = "St. Res. likelihood")
abline(h=0)

plot(id, stand.pearson.residuals, ylab = "St.Res.Pearson")
abline(h=0)
```



ΣΗΜΕΙΑ ΕΠΙΡΡΟΗΣ

Σύμφωνα με το διάγραμμα για το Cook's Distance δεν υπάρχει κάποιο σημείο επιρροής, καθώς όλες οι τιμές είναι μικρότερες του 1. Το διάγραμμα με τα hatvalues - id παρουσιάζει κάποια σημεία επιρροής καθώς υπάρχουν αρκετές παρατηρήσεις με τιμή 2p/n = 2*5/51 = 0.19607843137

```
par(mfrow = c(1, 2))

plot(id,cooks.distance(fit)) # points of influence if their value is  > 1

plot(id,hatvalues(fit)) # 2p/n = 2*3/51 =6/51= 0.11764705882
```



Παρακάτω βλέπουμε τα σημεία επιρροής από το διάγραμμα με τα hatvalues vs id.

```
hatvalues(fit)[hatvalues(fit)>0.19607843137]
```

```
##         1         2        11        13        32        40        48
## 0.2022412 0.3305063 0.2233724 0.2013889 0.1973466 0.2307939 0.2623983
```

ROC CURVE

Μεγάλη επιτυχία στο prediction σημαίνει πως υπάρχει μια τιμή p0(threshold) όπου υπάρχει υψηλό sensitivity και specificity αντίστοιχα. Σε αυτή την περίπτωση η καμπύλη ROC είναι κοντά στην πάνω αριστερή γωνία του διαγράμματος. Η περιοχή κάτω από την καμπύλη (AUC) δείχνει πόσο κοντά βρίσκονται σε αυτή τη γωνία. Η μεγιστή τιμή που μπορεί να λάβει είναι η 1 (εφόσον έχουμε 1x1 διαστάσεις).

Ορισμός ROC CURVE:

sensitivity = a/(a + c), rate of correct prediction of Y = 1 (true positive rate)

specificity = d/(b + d), rate of correct prediction of Y = 0 (true negative rate)

1-specificity είναι το ίδιο με το false positive rate.

Plot sensitivity VS 1-specificity για κάθε p0 από 0 εώς 1.

```
par(mfrow = c(1, 1))
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

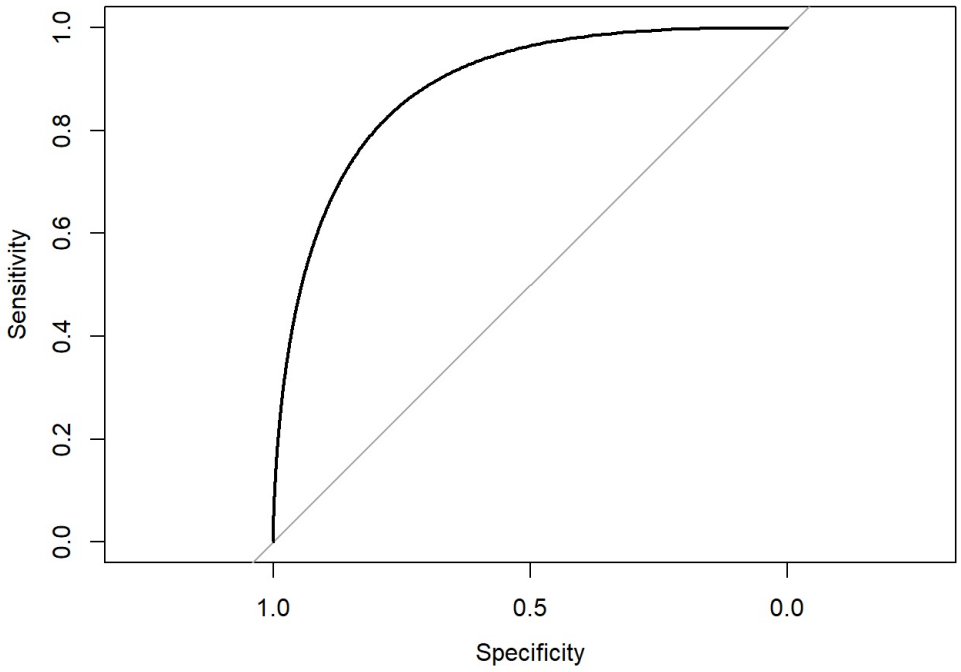
```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
roc(response, fitted.values(fit), smooth=TRUE, plot=TRUE)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```




```
##
## Call:
## roc.default(response = response, predictor = fitted.values(fit),     smooth = TRUE, plot = TRUE)
##
## Data: fitted.values(fit) in 27 controls (response 0) < 24 cases (response 1).
## Smoothing: binormal
## Area under the curve: 0.8838
```