1. Poisson Regression - Αρχείο **asfalies.txt**

Με βάση ένα μοντέλο της **παλινδρόμησης Poisson** να εξεταστεί η εξάρτηση του αριθμού Y αποζημιώσεων λόγω τροχαίων ατυχημάτων ανά n συμβόλαια, από την ηλικία του ασφαλισμένου (agecat X_1=0 νέος, 1 μεγάλος), την κατηγορία ασφαλίστρων (cartype X_2= 1,2,3,4) και την περιοχή διαμονής του ασφαλισμένου (district X_3=1, αν Αθήνα, X_3=0, αν σε άλλη πόλη).

1..1 Στο πρόγραμμα να δηλωθεί η μεταβλητή X_2 ως κατηγορική π.χ. μέσω της εντολής factor(cartype). Να γίνουν οι στατιστικοί έλεγχοι (Wald και Deviance), χρήση του κριτηρίου AIC.

```
library(data.table)
#read data
asfalies <- fread('asfalies.txt')
#factor cartype
asfalies$cartype <- factor(asfalies$cartype)
#fit a poisson generalized linear model
fit <- glm(y ~ cartype + agecat + district + offset(log(n)),
data = asfalies, family = 'poisson')
summary(fit)
```

```
##
## Call:
## glm(formula = y ~ cartype + agecat + district + offset(log(n)),
##     family = "poisson", data = asfalies)
##
## 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 ***
## cartype2     0.16223    0.05048   3.214 0.001309 **
## cartype3     0.39535    0.05491   7.200 6.03e-13 ***
## cartype4     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
```

►Από το παραπάνω μοντέλο που υπολογίστηκε με χρήση της R, έχουμε ότι η εξίσωση του προσαρμοσμένου μοντέλου είναι:

$$\hat{y} = \hat{\mu} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_5 x_5\right) \tag{1}$$

όπου τα x_i είναι οι αντίστοιχες μεταβλητές που έχουμε, και τα β_i είναι οι αντίστοιχοι συντελεστές κάτω από τη στήλη "estimate". Η τιμή του **ελέγχου Wald**, δηλαδή το z-value υπολογίζεται από τον λόγο $z_i = \hat{\beta}_i / se(\hat{\beta}_i)$, και με τον έλεγχο σημαντικότητας υπολογίζεται το p-value. Ουσιαστικά αυτός ο έλεγχος για κάθε μεταβλητή, συγκρίνει την αναγκαιότητα της παροουσίας της στο μοντέλο με το μοντέλο χωρίς τη μεταβλητή αυτή. Παρατηρούμε ότι **όλες οι μεταβλητές είναι στατιστικά σημαντικές**, εκτός από την μεταβλητή cartype2 με p-value ελάχιστα μεγαλύτερο του 0.001 που είναι το πιο αυστηρό διάστημα, όμως και αυτό για πολύ λίγο, επομένως δεν μπορούμε να πούμε ότι είναι ασήμαντη. Επομένως ο έλεγχος Wald μας δείχνει πως οι μεταβλητές αυτές όντως σχετίζονται με τον αριθμό Y των αποζημιώσεων.

►Στη συνέχεια, πραγματοποιούμε το εξής:

```
anova(fit,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
## cartype   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
```

Προκειμένου να ελέγξουμε το τι συμβαίνει με το μοντέλο μας, πρέπει να κατανοήσουμε τι είναι το Null και τι το residual deviance:
•Το Null Deviance, είναι ανάλογο της διαφοράς του loglikelihood του saturated μοντέλου, με το loglikelihood του null μοντέλου. Saturated μοντέλο σημαίνει ένα μοντέλο το οποίο έχει τέλειο fit καθώς έχει τόσες παραμέτρους όσες και παρατηρήσεις, ενώ το null μοντέλο έχει μόνο μια παράμετρο για όλες τις παρατηρήσεις.
•Το Residual deviance, είναι ανάλογο της διαφοράς του loglikelihood του saturated μοντέλου, με το loglikelihood ενός άλλου proposed μοντέλου, που είναι εμφωλευμένο στο saturated, δηλαδή έχει λιγότερες παραμέτρους από αυτό.
Πίσω στο μοντέλο μας, βλέπουμε ότι το Null Deviance έχει 31 βαθμούς ελευθερίας (Dfs), αφού έχουμε 32 παρατηρήσεις - 1 βαθμό ελευθερίας του Null μοντέλου. Στη συνέχεια οι βαθμοί ελευθερίας μειώνονται ανάλογα με το πόσες παραμέτρους έχει κάθε μεταβλητή. Για παράδειγμα η factored μεταβλητη cartype έχει 3 επίπεδα επομένως 3 παραμέτρους άρα οι συνολικοί βαθμοί ελευθερίας για αυτό το μοντέλο είναι 28. Έτσι βρίσκονται και τα υπόλοιπα.
Αυτός ο έλεγχος δηλαδή γίνεται προκειμένου να ελέγξουμε αν κάποιο μοντέλο εμφωλευμένο στο πλήρες δικό μας, περιγράφει καλύτερα τα δεδομένα. Από τις τιμές των ελέγχων, βλέπουμε ότι για όλες τις μεταβλητές ισχύει πως p-value < 0.05, επομένως **όλες οι μεταβλητές είναι στατιστικά σημαντικές**, άρα η μηδενική υπόθεση σε όλες τις περιπτώσεις πρέπει να απορριφθεί. Αυτό σημαίνει ότι όλες οι μεταβλητές συμβάλλουν στο να περιγράψουμε καλύτερα τα δεδομένα μας.

►Όσον αφορά την τιμή του AIC, από το summary του μοντέλου μας βλέπουμε ότι AIC = 222.15. Προκειμένου να ελέγξουμε αν αυτό το μοντέλο είναι το βέλτιστο, δηλαδή έχει το ελάχιστο AIC, πραγματοποιούμε ένα backwards test:

```
step(fit, direction = 'backward')


## Start:  AIC=222.15
## y ~ cartype + agecat + district + offset(log(n))
##
##             Df Deviance    AIC
## <none>           41.789 222.15
## - district  1   54.727 233.09
## - agecat    1  107.964 286.32
## - cartype   3  131.713 306.07
##
## Call:  glm(formula = y ~ cartype + agecat + district + offset(log(n)),
##     family = "poisson", data = asfalies)
##
## Coefficients:
## (Intercept)     cartype2     cartype3     cartype4       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
```

Παρατηρούμε ότι η αφαίρεση οποιασδήποτε μεταβλητής προκαλεί αύξηση του AIC, επομένως το μοντέλο μας είναι το βέλτιστο. Το ίδιο αποτέλεσμα έχουμε και αν δοκιμάσουμε Bidirectional elimination.

►Τέλος,μέσω της ελεγχοσυνάρτησης Deviance, μπορούμε να υπολογίσουμε πόσο απέχει το δικό μας μοντέλο, από το saturated μοντέλο, υπολογίζοντας το p-value αυτής της διαφοράς, δεδομένου ότι αυτή κατανέμεται ασυμπτωτικά με την κατανομή χ^2 με βαθμούς ελευθερίας $n - p$.

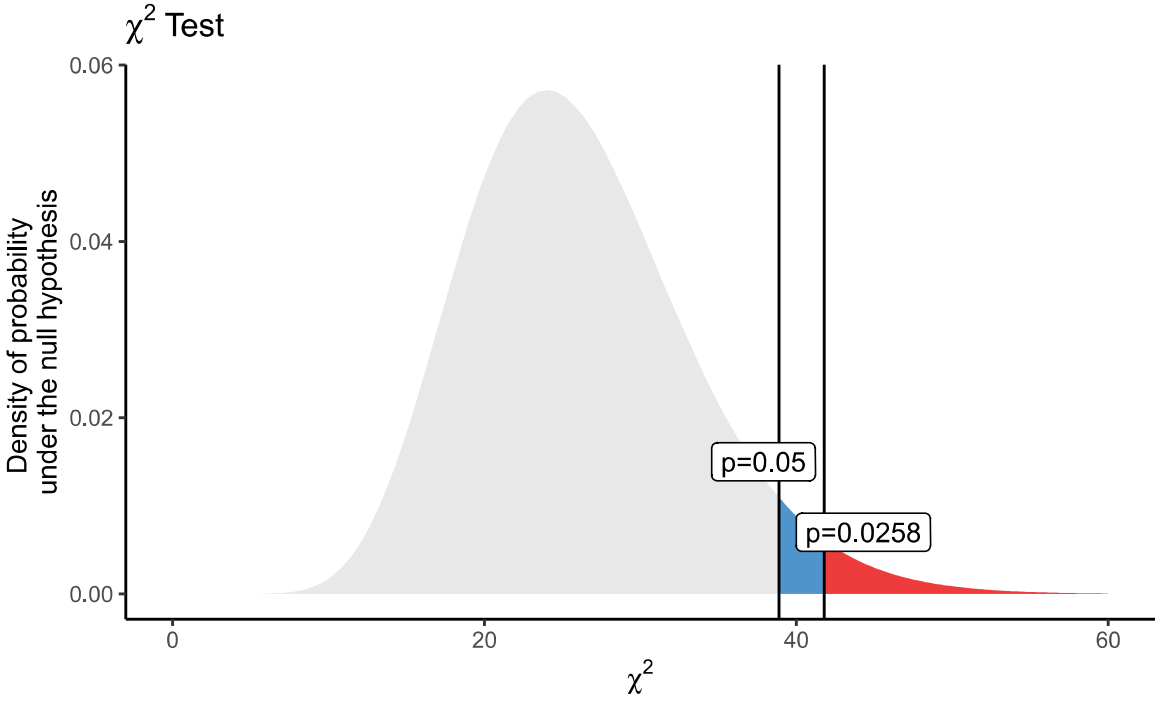
```
pchisq(fit$deviance,fit$df.residual,lower.tail=FALSE)


## [1] 0.02580847
```

Σε αυτή την περίπτωση, βλέπουμε ότι το p-value είναι αρκετά μικρό, εάν θεωρήσουμε επίπεδο σημαντικότητας $\alpha = 0.05$, επομένως βλέπουμε ότι **παρόλο που όλες οι μεταβλητές χρειάζονται μέσα στο μοντέλο μας, αυτό στο τέλος δεν είναι αρκετό να περιγράψει με αρκετή σαφήνεια τα δεδομένα μας.**Επειδή ο γραφικός έλεγχος του χ^2-test μας βοηθά να καταλάβουμε καλύτερα γιατί απορρίπτουμε αυτήν την τιμή, στο σχήμα 1 βλέπουμε ότι η p-value βρίσκεται δεξιότερα του ορίου που μας θέτει το όριο σημαντικότητας α.

```
library(ggplot2)
ourpvalue<- 1-pchisq(fit$deviance,fit$df.residual)
CI_pvalue <- 0.05
ourch_value <- qchisq(ourpvalue,fit$df.residual,lower.tail=FALSE)
CI_chvalue <- qchisq(CI_pvalue,fit$df.residual,lower.tail=FALSE)
ch <- data.frame(x = c(0, 60))
chi_plot <- ggplot() +stat_function(data = ch,aes(x=x),fun = dchisq,
                args = list(df = 26),geom='area',fill='gray91')+

  stat_function(fun = dchisq,args = list(df = 26),
                xlim = c(38.885,41.789),
                geom = "area",fill='steelblue3')+
  stat_function(fun = dchisq,args = list(df = 26),
                xlim = c(41.789,60),
                geom = "area",fill='brown2')+
  geom_vline(xintercept = 41.789)+geom_vline(xintercept=38.885)+
theme_classic()+
  geom_label(aes(CI_chvalue-1,y=0.015,label=paste0("p=",CI_pvalue)))+
  geom_label(aes(ourch_value+2.5,y=0.007,label=paste0("p=",
  round(ourpvalue,4))))+
labs(title=parse(text = expression(chi^2 ~ "Test")))+
  ylab('Density of probability \nunder the null hypothesis')+
  xlab(parse(text = expression(chi^2)))
chi_plot
```



Σχήμα 1: $\chi^2 - Test$

1..2 Να κατασκευαστούν διαστήματα εμπιστοσύνης για τους εκτιμημένους συντελεστές $\hat{\beta}$ του τελικού μοντέλου και να γίνουν ερμηνείες.

Στην R, υπολογίζουμε τα διαστήματα εμπιστοσύνης για τους συντελεστές ως εξής:

```
suppressMessages(confint(fit))
```

```
##                   2.5 %     97.5 %
## (Intercept) -2.04472348 -1.8281432
## cartype2     0.06402271  0.2619349
## cartype3     0.28814409  0.5034436
## cartype4     0.42299446  0.7059551
## agecat      -0.46278476 -0.2882587
## district     0.10011971  0.3296247
```

Παρατηρούμε πως σε καμία περίπτωση δεν περιέχεται το μηδέν στο διάστημα εμπιστοσύνης. Αυτό είναι καλό για τους συντελεστές μας καθώς αν κάποιο διάστημα περιείχε το μηδέν θα σήμαινε ότι ίσως έχουμε κάνει κάτι λάθος και έπρεπε να είχαμε απορρίψει κάποια μεταβλητή νωρίτερα.
Όμως, επειδή το μοντέλο μας έχει τη μορφή που φαίνεται στην εξίσωση (1), τα πραγματικά διαστήματα εμπιστοσύνης για τους συντελεστές είναι τα εξής:

```
suppressMessages(exp(confint(fit)))
```

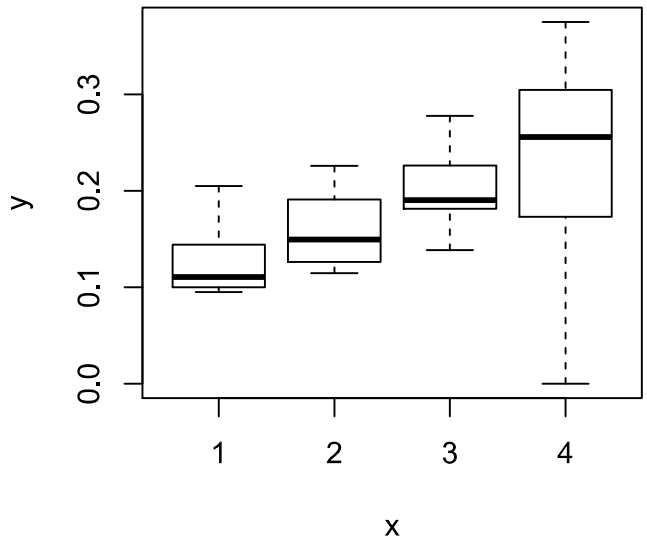
```
##                 2.5 %    97.5 %
## (Intercept) 0.1294160 0.1607117
## cartype2    1.0661166 1.2994419
## cartype3    1.3339495 1.6544086
## cartype4    1.5265258 2.0257805
## agecat      0.6295281 0.7495676
## district    1.1053032 1.3904462
```

Φέρνουμε και τους συντελεστές να τους έχουμε μπροστά μας:

```
fit$coefficients
```

```
## (Intercept)    cartype2    cartype3    cartype4      agecat    district
##  -1.9352231   0.1622320   0.3953472   0.5654289  -0.3762785   0.2166106
```

►Αυτό σημαίνει, πως αν αυξηθεί κατά μια μονάδα η συμμεταβλητή agecat, δηλαδή γίνει 1 = μεγάλος σε ηλικία, τότε η αποζημίωση Y θα πολλαπλασιαστεί με $e^{-0.3762785} = 0.6864$, ή αλλιώς με διάστημα 95% εμπιστοσύνης: $(0.6295281, 0.7495676)$, που σημαίνει ότι θα πάρει 25 έως 37% μικρότερη αποζημίωση.
►Αντίστοιχα, αν η συμμεταβλητή district αυξηθεί κατά μια μονάδα, (δηλαδή Αθήνα) τότε η αποζημίωση αυξάνεται κατά $(1.1053032, 1.3904462)$, δηλαδή στην Αθήνα οι αποζημιώσεις είναι μεγαλύτερες κατά 10 ως 40%.
►Όσον αφορά την μεταβλητή Cartype, καθώς αυτή έχει διαφορετικά επίπεδα, παρατηρούμε ότι εμφανίζονται στους συντελεστές μόνο τα cartypes 2,3 και 4. Αυτό συμβαίνει διότι ως default θεωρείται το cartype = 1, και εφόσον δεν γίνεται να ισχύει παραπάνω από ένα είδος ταυτόχρονα, οι συντελεστές των υπόλοιπων τριών ειδών μας δείχνουν τη μεταβολή των αποζημιώσεων τους σε σχέση με το default. Βλέπουμε ότι η cartype4 έχει μεγαλύτερο συντελεστή, δηλαδή επιφέρει 52.65% με 100% μεγαλύτερες αποζημιώσεις. Αυτό άλλωστε είναι εμφανές και από το διάγραμμα 2, όπου φαίνεται ο λόγος των αποζημιώσεων προς τον αριθμό των συμβολαίων σε σχέση με το cartype.



Σχήμα 2: y/n vs Cartype Boxplot

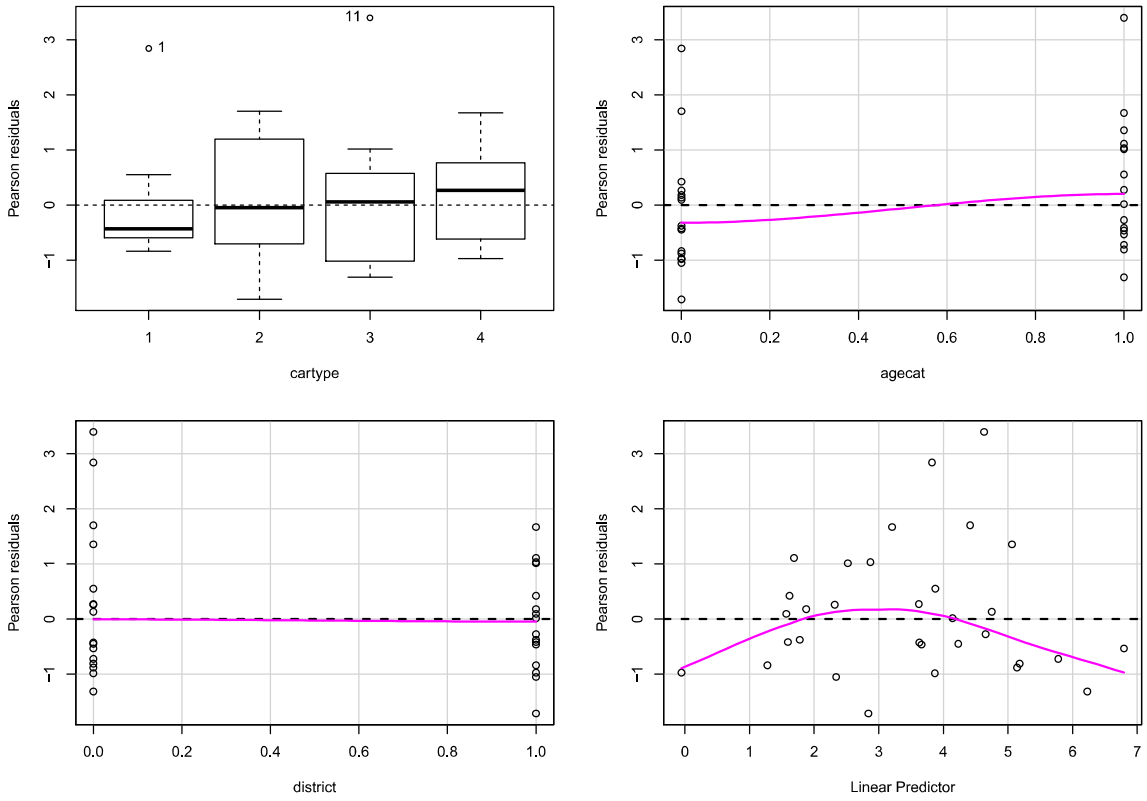
```
suppressWarnings(plot(asfalies$cartype,asfalies$y/asfalies$n))
```

1..3 Να γίνουν γραφικές παραστάσεις για τα υπόλοιπα Pearson και Deviance, index plots για τα h_{ii}, τις αποστάσεις Cook, καθώς και για τα υπόλοιπα πιθανοφάνειας.

Εδώ, παρουσιάζουμε τα διαγράμματα τα οποία θα χρησιμοποιήσουμε, και θα τα σχολιάσουμε στο τέλος.

```
#plot3
suppressMessages(library(car))
residualPlots(fit)
```

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



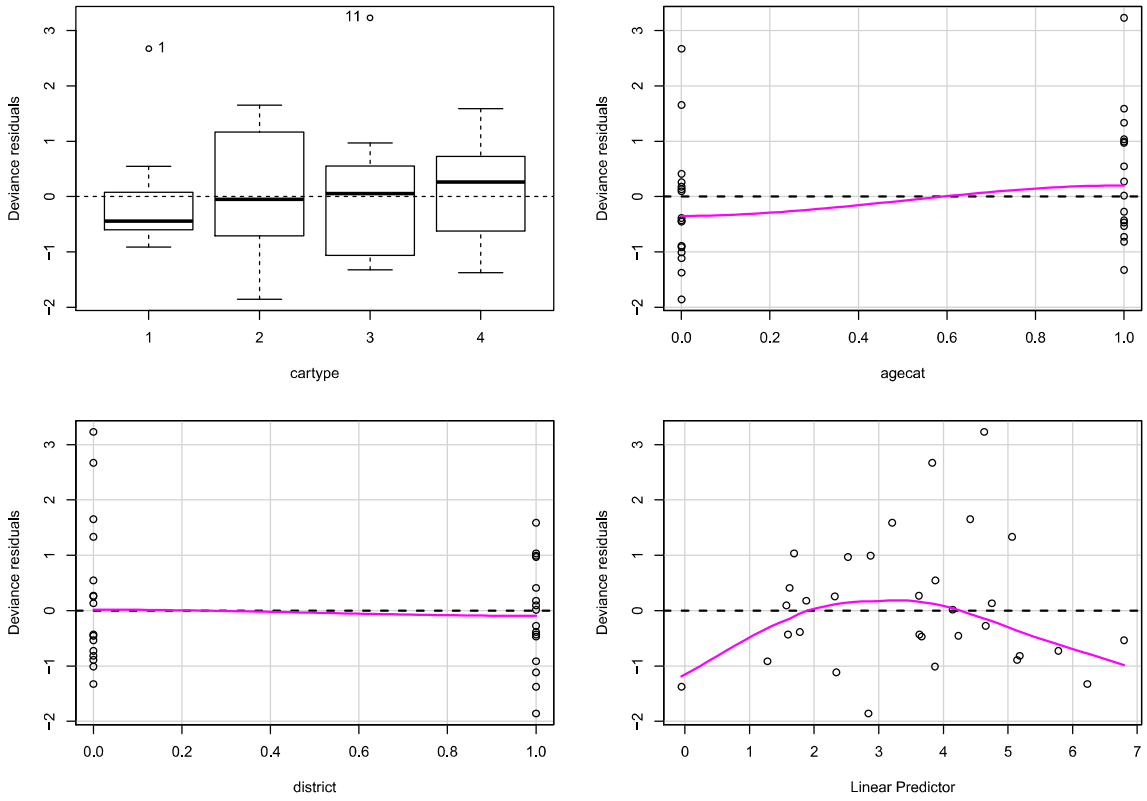
Σχήμα 3: Pearson Residuals vs Variables

```
#plot4
suppressMessages(library(car))
residualPlots(fit,type = 'deviance')
```

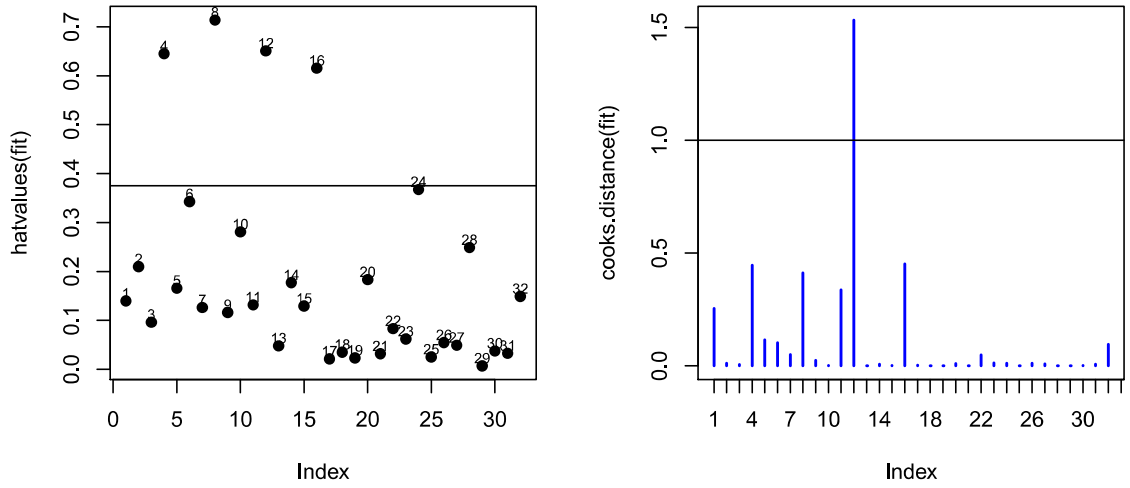
```
##          Test stat Pr(>|Test stat|)
## cartype
## agecat           0                1
## district         0                1
```

```
#plot5
par(mfrow = c(1, 2))
plot(hatvalues(fit),pch=19)
text(hatvalues(fit),offset = 0.1,pos=3,cex=0.7)
abline(h=2*6/32)

plot(cooks.distance(fit), type = "h", lwd = 2,col = "blue",xaxt='n')
axis(side=1, 1:50)
abline(h=1)
```

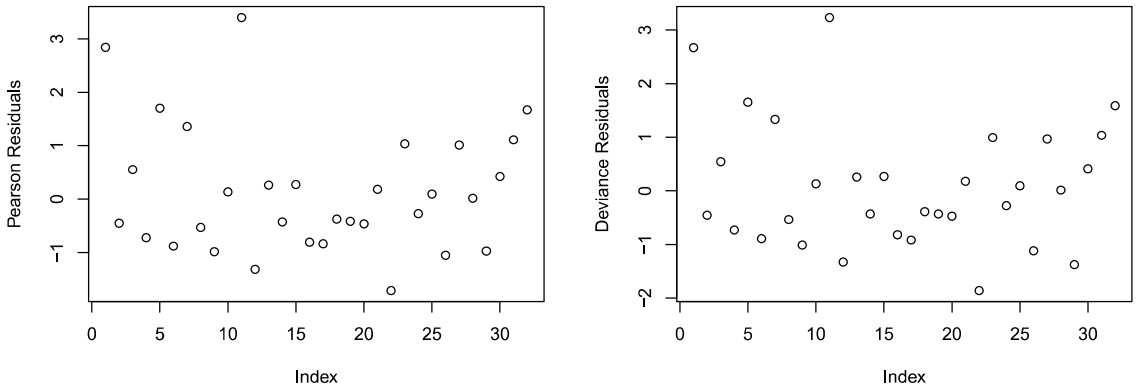


Σχήμα 4: Deviance Residuals vs Variables



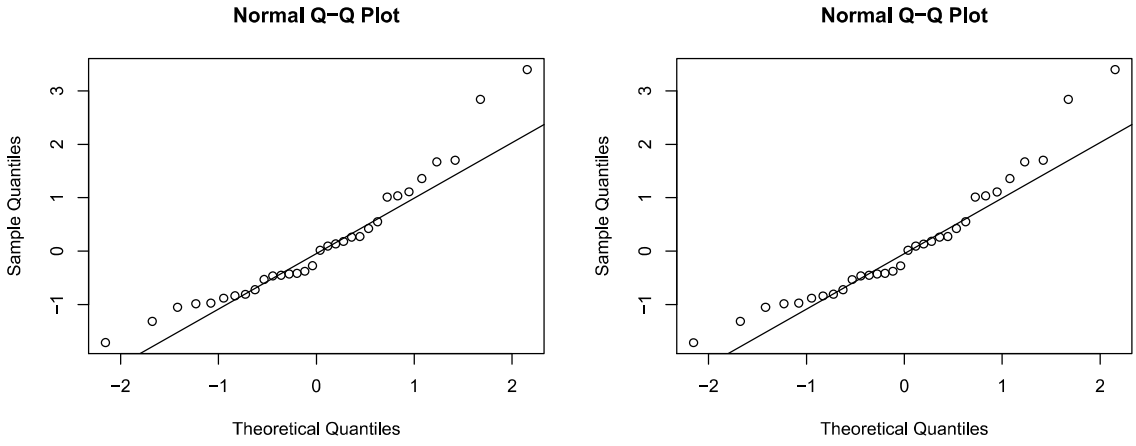
Σχήμα 5: Hat Values and Cook's Distance vs Index

```
asfalies$pearson <-residuals(fit,type="pearson")
asfalies$res.deviance <-residuals(fit)
par(mfrow = c(1, 2))#Plot6
plot(asfalies$pearson,xlab='Index',ylab='Pearson Residuals')
plot(asfalies$res.deviance,xlab='Index',ylab='Deviance Residuals')
```



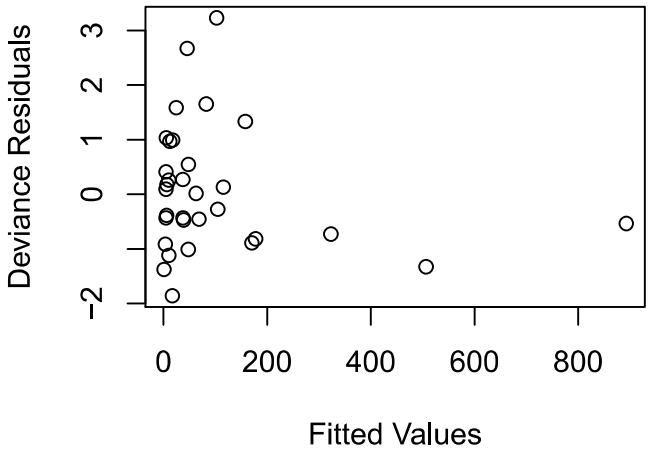
Σχήμα 6: Pearson and Deviance Residuals Index Plots

```
par(mfrow=c(1,2))#Plot7
qqnorm(asfalies$pearson)
qqline(asfalies$res.deviance)
qqnorm(asfalies$pearson)
qqline(asfalies$res.deviance)
```



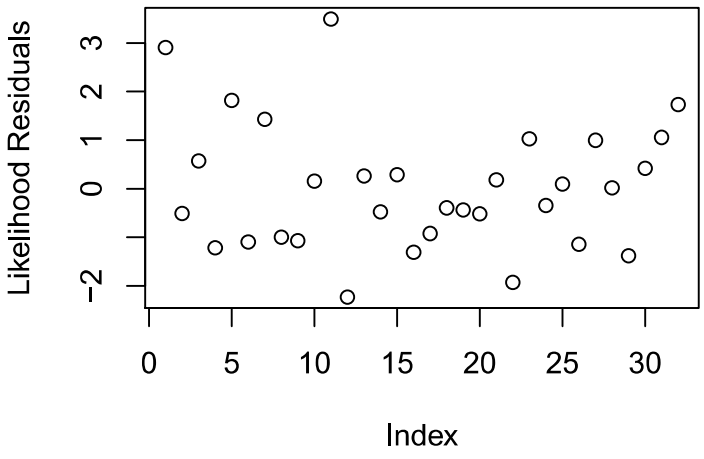
Σχήμα 7: QQ-Plots for Pearson and Deviance Residuals

```
plot(fitted.values(fit),asfalies$res.deviance,xlab='Fitted Values',
ylab='Deviance Residuals')#Plot8
```



Σχήμα 8: Deviance Residuals vs Fitted Values

```
asfalies$likelihood.res <- rstudent(fit)#plot9
plot(asfalies$likelihood.res,xlab='Index',ylab='Likelihood Residuals')
```



Σχήμα 9: Likelihood Residuals Index Plot

Σχολιασμός: ►Στα σχήματα 3 και 4, βλέπουμε διαγράμματα των τριών μεταβλητών και του linear predictor, για τα υπόλοιπα Pearson και Deviance αντίστοιχα. Δεδομένου ότι αυτά διαφέρουν ελάχιστα, ο σχολιασμός θα είναι ο ίδιος και για τα δύο ποιοτικά. Σε αυτά δεν παρατηρούμε κάποια συστηματικότητα και πως όλα βαίνουν καλώς, με την ύπαρξη όμως κάποιων outliers, για παράδειγμα οι τιμές 1 και 11.
►Στο σχήμα 7, βλέπουμε τα γραφήματα κανονικής κατανομής των υπολοίπων Pearson και Deviance. Αυτός ο έλεγχος έχει πιο πολύ νόημα για τον έλεγχο καλής προσαρμογής, και όχι σαν ένδειξη της κανονικότητας των υπολοίπων. Βλέπουμε όμως ότι τα υπόλοιπα κατανέμονται καλώς στην ευθεία, όμως υπάρχουν κάποια σημεία outliers (τα σημεία 1,11 στα δεξιά και το σημείο 12 στα αριστερά) τα οποία υποδεικνύουν την όχι τόσο καλή προσαρμογή του μοντέλου μας, με τα οποία θα πρέπει να είμαστε προσεκτικοί.
►Στα σχήματα 6 και 9 βλέπουμε Index plots των υπολοίπων Pearson, Deviance και likelihood. Και στις τρείς περιπτώσεις, βλέπουμε πως τα υπόλοιπα κατανέμονται τυχαία με ικανοποιητικό τρόπο, και δεν υπάρχει κάποιο trend.
►Στο σχήμα 5 βλέπουμε Index plots των Hatvalues και Cook's Distance. Αυτά δείχνουν ποια σημεία ασκούν επιρροή στο μοντέλο μας. Στην περίπτωση των Hatvalues, σημεία επιρροής θεωρούνται αυτά με $h_{ii} > 2p/n$ όπου p=6 ο αριθμός των παραμέτρων του μοντέλου μας, και n=32 ο αριθμός των παρατηρήσεων. Στην περίπτωση της απόστασης Cook το μάτι μας επικεντρώνεται σε τιμές μεγαλύτερες της μονάδας. Έτσι, βλέπουμε ότι το σημείο 12 είναι σημείο επιρροής καθώς βρίσκεται εκτός ορίων και στις δύο περιπτώσεις, και επιρροή φαίνεται να ασκούν και τα σημεία 4,8 και 16.
►Στο σχήμα 8, βλέπουμε διάγραμμα των fitted values από το μοντέλο μας με τα υπόλοιπα Deviance. Αρχικά, φαίνεται ότι τα περισσότερα σημεία βρίσκονται μαζεμένα στην αρχή, όμως αυτό είναι εξαιτίας της μεγάλης κλίμακας. Το σημείο 8, είναι ένα σημείο το οποίο έχει πολύ μεγάλη αποζημίωση με μεγάλο αριθμό συμβολαίων, επομένως έχει και μεγάλη τιμή η πρόβλεψή του από το μοντέλο. **Συμπέρασμα:** Είναι εμφανές, ότι μέσω των ελέγχων Wald και Deviance, όλες οι μεταβλητές μας είναι σημαντικές, όμως παρ'όλα αυτά, κάποιες έκτοπες τιμές ασκούν επιρροή στο μοντέλο μας και χαλούν την καλή του προσαρμογή, το οποίο και αποτυπώνεται στα αντίστοιχα διαγράμματα, όπως σχολιάστηκε παραπάνω. Γι'αυτό, δοκιμάζουμε να αφαιρέσουμε κάποιες έκτοπες τιμές[1], και να κάνουμε τον τελικό έλεγχο της ελεγχοσυνάρτησης Deviance ώστε να δούμε αν η κατάσταση βελτιώθηκε. Αφαιρούμε λοιπόν, τις τιμές: 1,8,11 και 12. Όπως βλέπουμε, αυτή η p-value τιμή είναι σαφώς μεγαλύτερη, την οποία μπορούμε και να αποδεχτούμε, έχοντας ως αποτέλεσμα ότι το μοντέλο μας έχει πλέον καλή προσαρμογή.

```
asfalies_new <-asfalies[!c(1,11,12,8)]
fit_new <- glm(y ~ cartype + agecat + district + offset(log(n)),
        data = asfalies_new, family = 'poisson')
1-pchisq(fit_new$deviance,fit_new$df.residual)


## [1] 0.6323179
```

[1]Στην πραγματικότητα, δεν αφαιρούμε τόσο εύκολα δεδομένα αν δεν ξέρουμε πως αυτά έχουν δημιουργηθεί, όμως το κάνουμε τώρα σαν παράδειγμα για να δούμε ότι όντως λίγες τιμές χαλάνε την καλή προσαρμογή του μοντέλου.

2. Logistic Regression - Αρχείο **leukaemia.txt**

Εξαρτημένη Μεταβλητή: Ανταπόκριση στη θεραπεία, ναι=1, όχι=0.

Συμμεταβλητές:

1. age - ηλικία του ασθενή
2. smear - ποσοστό επίστρωσης βλαστοκυττάρων
3. infiltrate - ποσοστό κυττάρων στο μυελό των οστών
4. index - δείκτης κυττάρων λευχαιμίας
5. blasts - βλαστοκύτταρα
6. temperature - υψηλότερη θερμοκρασία πριν τη θεραπεία $(\times 10^o F)$.

2..1 Να εξεταστεί η 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.

Ξεκινάμε, κάνοντας προσαρμογή ενός μοντέλου λογιστικής παλινδρόμησης στα δεδομένα μας, όπως φαίνεται στην συνέχεια. Παρατηρούμε, ότι για επίπεδο σημαντικότητας $\alpha = 0.05$, οι μεταβλητές smear, infiltrate και blasts δεν θεωρούνται στατιστικά σημαντικές, καθώς έχουν p-value > 0.05 .Επομένως, με βάση τον έλεγχο Wald, αυτές οι μεταβλητές θα έπρεπε να αφαιρεθούν από το μοντέλο. Για να έχουμε όμως μια καλύτερη εικόνα, θα πραγματοποιήσουμε και τον έλεγχο Deviance, να δούμε αν βγάζει κάποια διαφορετικά αποτελέσματα.

```
data <- fread('leukaemia.txt')
model <- glm(response ~ age+smear+infiltrate+index+blasts+temperature,
             data=data,family='binomial')
summary(model)
```

```
##
## 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
##
## Number of Fisher Scoring iterations: 6
```

```
anova(model,test='Chisq')
```

```
## 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
```

```r
step(model,direction ='backward')
```

```
## Start:  AIC=54.06
## response ~ age + smear + infiltrate + index + blasts + temperature
##
##               Df Deviance    AIC
## - smear        1   40.074 52.074
## - blasts       1   40.115 52.115
## - infiltrate   1   41.023 53.023
## <none>             40.060 54.060
## - age          1   46.157 58.157
## - temperature  1   48.277 60.277
## - index        1   55.823 67.823
##
## Step:  AIC=52.07
## response ~ age + infiltrate + index + blasts + temperature
##
##               Df Deviance    AIC
## - blasts       1   40.136 50.136
## <none>             40.074 52.074
## - infiltrate   1   42.615 52.615
## - age          1   46.216 56.216
## - temperature  1   48.346 58.346
## - index        1   56.308 66.308
##
## Step:  AIC=50.14
## response ~ age + infiltrate + index + temperature
##
##               Df Deviance    AIC
## <none>             40.136 50.136
## - infiltrate   1   43.265 51.265
## - age          1   46.438 54.438
## - temperature  1   48.971 56.971
## - index        1   57.602 65.602
##
## 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
```

Παρατηρούμε ότι και ο έλεγχος Deviance, βγάζει τα ίδια ποιοτικά αποτελέσματα, καθώς επιβεβαιώνει τις υποψίες μας πως οι μεταβλητές smear, infiltrate και blasts δεν προσθέτουν αρκετή πληροφορία στο μοντέλο μας. Επομένως, μέσω των μεθόδων stepwise selection, θα επιλέξουμε ένα καταλληλότερο μοντέλο. Όπως φαίνεται στην προηγούμενη σελίδα, η μέθοδος stepwise selection με προς τα πίσω κατεύθυνση, μας δείχνει ότι το βέλτιστο μοντέλο είναι αυτό που περιέχει τις μεταβλητές age, infiltrate, index και temperature, καθώς ελαττώνει το AIC σε 50.14, ενώ στην αρχή ήταν 54.06. Για να ελέγξουμε πόσο καλή προσαρμογή έχει το αρχικό μας μοντέλο σε σχέση με το τελικό, πραγματοποιούμε έλεγχο Deviance συγκρίνοντας τα μοντέλα μας με το αντίστοιχο saturated μοντέλο.

```
model2 <- glm(response ~ age+index+temperature+infiltrate,
        data=data,family='binomial')
summary(model2)
```

```
##
## Call:
## glm(formula = response ~ age + index + temperature + infiltrate,
##     family = "binomial", data = data)
##
## 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 *
## index        0.40673    0.13034   3.121  0.00181 **
## temperature -0.09944    0.03954  -2.515  0.01191 *
## infiltrate   0.03413    0.02079   1.641  0.10077
## ---
## 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
```

```
anova(model2,test='Chisq')
```

```
## 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 *
## index        1  12.6168        48     51.387 0.0003823 ***
## temperature  1   8.1216        47     43.265 0.0043741 **
## infiltrate   1   3.1291        46     40.136 0.0769039 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Παρατηρούμε ότι οι ελέγχοι Wald και Deviance, δείχνουν ότι η μεταβλητή infiltrate δεν θεωρείται στατιστικά σημαντική, γι'αυτό και αποφασίζουμε να την αφήσουμε[2], ακόμα κι αν αυτό έχει έναν μικρό αντίκτυπο στο AIC.

```
model3 <- glm(response ~ age+index+temperature,
        data=data,family='binomial')
summary(model3)
```

```
##
## Call:
## glm(formula = response ~ age + index + temperature, family = "binomial",
##     data = data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.76104  -0.68683  -0.09747   0.67388   2.16510
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 87.38804   35.45816   2.465  0.01372 *
## age         -0.05850    0.02558  -2.287  0.02218 *
## index        0.38493    0.12152   3.168  0.00154 **
## temperature -0.08897    0.03607  -2.467  0.01363 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

[2]Επιπλέον, παρακάτω στα διαστήματα εμπιστοσύνης των συντελεστών, σε αυτή τη μεταβλητή περιεχόταν το μηδέν, οπότε αυτά τα δεδομένα είναι αρκετά να μας κάνουν να την αφήσουμε.

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 70.524  on 50  degrees of freedom
## Residual deviance: 43.265  on 47  degrees of freedom
## AIC: 51.265
##
## Number of Fisher Scoring iterations: 6
```

Για να ελέγξουμε την προσαρμογή του τελικού μας μοντέλου, κάνουμε τον έλεγχο Deviance συγκρίνοντάς το με το saturated μοντέλο.

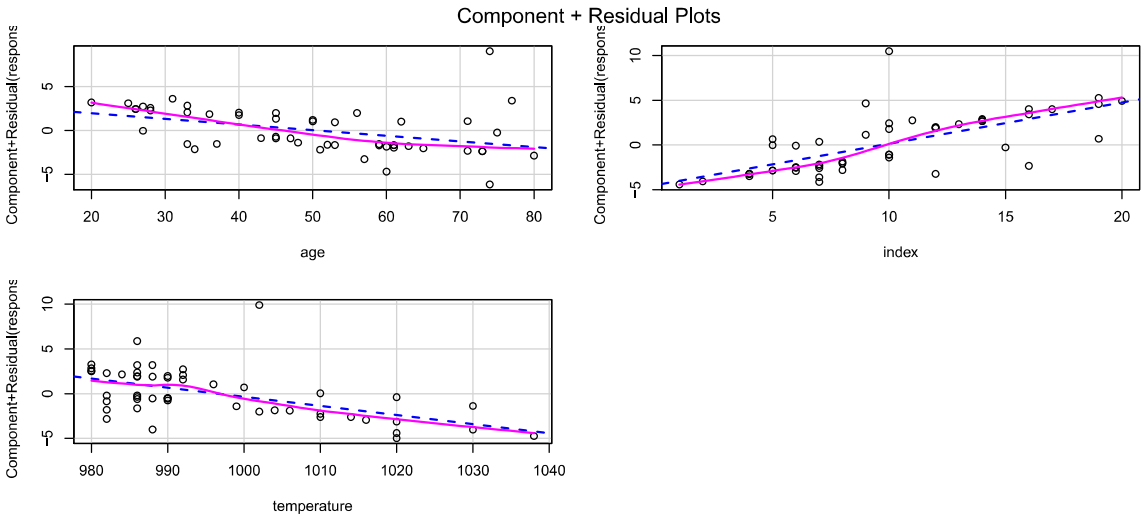
```
p_val3<- 1-pchisq(model3$deviance,model3$df.residual)
p_val3


## [1] 0.6280164
```

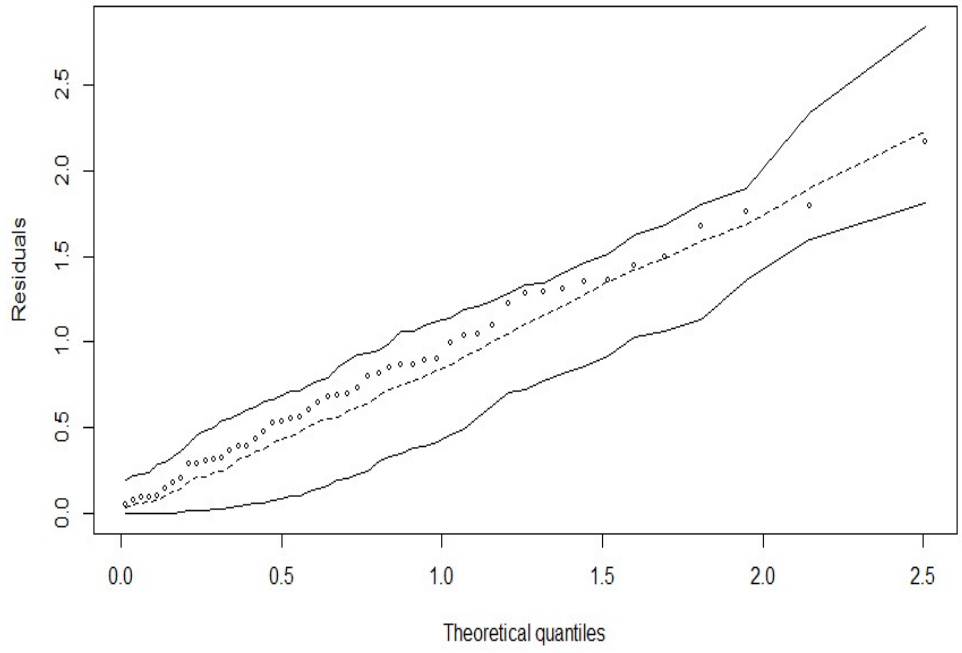
Συμπεραίνουμε ότι η προσαρμογή του τελικού μας μοντέλου είναι καλή, δεδομένου ότι το P-value είναι αρκετά μεγάλο (>0.05) και το αποδεχόμαστε.

2..2 Να γίνουν γραφικές παραστάσεις των μερικών υπολοίπων, των υπολοίπων Deviance (με την ημι κανονική κατανομή), index plots των h_{ii}, των αποστάσεων Cook, καθώς και των υπολοίπων πιθανοφάνειας.

```
crPlots(model3)
```



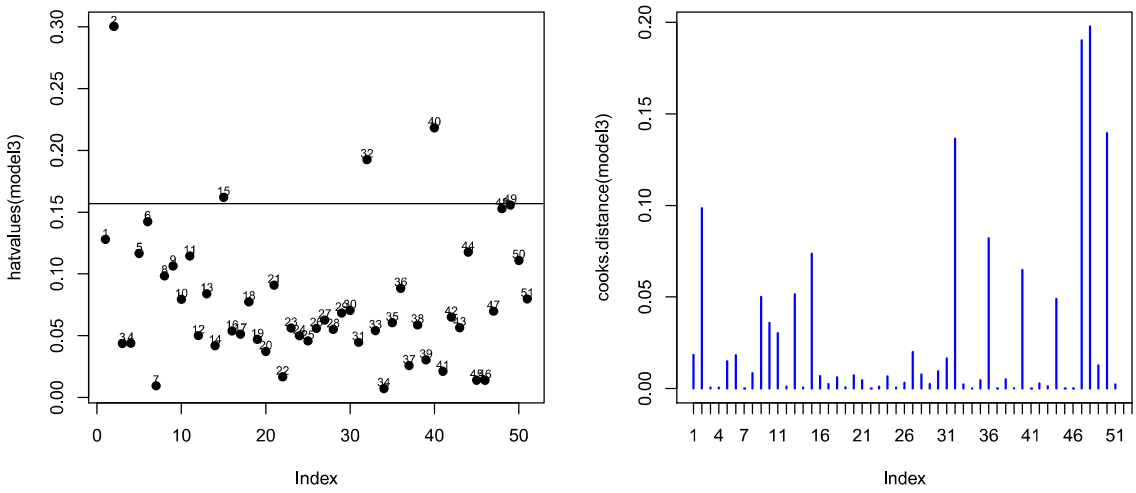
Σχήμα 10: Partial Residual Plots



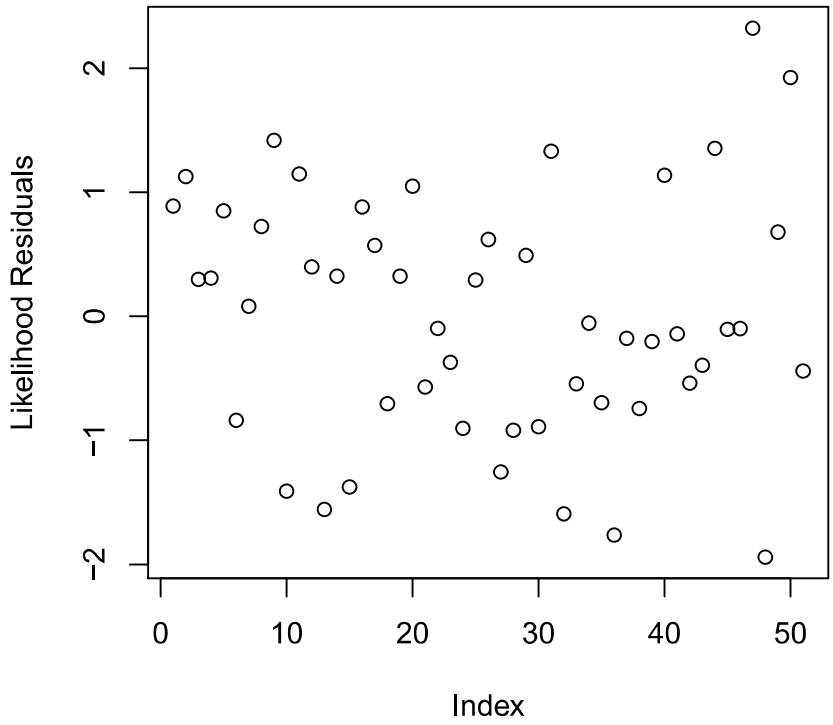
Σχήμα 11: Pearson Residual Half-Normal Plot with Simulation Envelopes

```r
#plot12
par(mfrow = c(1, 2))
plot(hatvalues(model3),pch=19)
text(hatvalues(model3),offset = 0.1,pos=3,cex=0.7)
abline(h=2*4/51)
plot(cooks.distance(model3), type = "h", lwd = 2,col = "blue",xaxt='n')
axis(side=1, 1:55)
abline(h=1)
```

```r
#plot13
plot(rstudent(model3),xlab='Index',ylab='Likelihood Residuals')
```



Σχήμα 12: Hatvalues and Cook's Distance Index Plots



Σχήμα 13: Likelihood Residuals Index Plot

Σχολιασμός:

- Στο σχήμα 10 βλέπουμε τις γραφικές παραστάσεις μερικών υπολοίπων για κάθε μία από τις μεταβλητές του μοντέλου μας. Εφόσον και για τις τρείς μεταβλητές τα γραφήματα παρουσιάζουν ικανοποιητικά γραμμική τάση, επομένως όλες αυτές οι μεταβλητές παραμένουν στο μοντέλο, και δεν χρειάζονται κάποιον μετασχηματισμό.
- Στο σχήμα 11, παρόλο που τα υπόλοιπά μας δεν προέρχονται από κανονική κατανομή, η εμφάνιση ενός trend αυτών, δείχνει καλή προσαρμογή του μοντέλου μας. Δεδομένου ότι όλα τα υπόλοιπα βρίσκονται εντός των ορίων εμπιστοσύνης (envelope), το μοντέλο μας δείχνει καλή προσαρμογή.
- Στο σχήμα 12, παρατηρούμε ότι στα hatvalues υπάρχουν κάποιες τιμές οι οποίες ξεπερνούν το εμπειρικό όριο, δηλαδή είναι $h_{ii} > 2p/n$ όπου $p = 4$ όσες και οι παραμέτροι του μοντέλου και $n = 51$ όσες και οι παρατηρήσεις μας. Στο διάγραμμα με τις αποστάσεις Cook, βλέπουμε ότι καμία τιμή δεν ξεπερνάει το όριο $D_i > 1$. Αυτό, μας δείχνει ότι ενδεχομένως οι outlier τιμές που βλέπουμε στο διάγραμμα με τα hatvalues, μπορεί να είναι outliers αλλά όχι σημεία επιρροής, καθώς όπως βλέπουμε και στο σχήμα 11, υπάρχουν κάποιες τιμές δεξιότερα στον άξονα X, χωρίς όμως να αποκλίνουν σημαντικά από την ευθεία.
- Τέλος, στο σχήμα 13, βλέπουμε το index plot των υπολοίπων πιθανοφάνειας, τα οποία φαίνονται να είναι αρκετά τυχαία διασκορπισμένα, επομένως αυτό μας αρκεί για την καλή προσαρμογή του μοντέλου μας.

2..3 Να κατασκευαστούν διαστήματα εμπιστοσύνης για τους εκτιμημένους συντελεστές $\hat{\beta}$ του τελικού μοντέλου και να γίνουν ερμηνείες.

```
suppressMessages(confint(model3))
```

```
##                  2.5 %       97.5 %
## (Intercept) 24.6827051 166.27140975
## age         -0.1149750  -0.01237056
## index        0.1797435   0.66649232
## temperature -0.1693614  -0.02529030
```

```
suppressMessages(exp(confint(model3)))
```

```
##                    2.5 %       97.5 %
## (Intercept) 5.242792e+10 1.624635e+72
## age         8.913885e-01 9.877056e-01
## index       1.196910e+00 1.947394e+00
## temperature 8.442038e-01 9.750268e-01
```

```
model3$coefficients


## (Intercept)         age       index temperature
## 87.38803961 -0.05850162  0.38492608 -0.08897319
```

Αρχικά παρατηρούμε ότι το μηδέν δεν περιέχεται σε κανένα διάστημα εμπιστοσύνης, οπότε μέχρι στιγμής όλα βαίνουν καλώς. Το τελικό μοντέλο είναι, όπου $\hat{p}$ η πιθανότητα ανταπόκρισης στη θεραπεία:

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.059 \cdot age + 0.385 \cdot index - 0.089 \cdot temperature \tag{2}$$

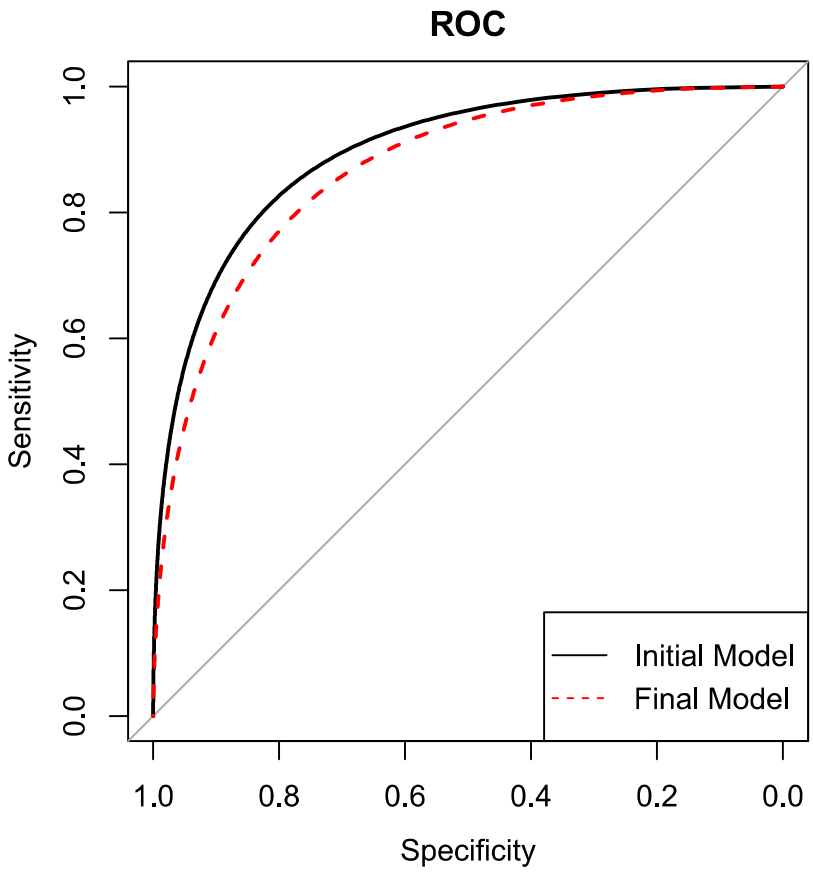
και ισοδύναμα:

$$\hat{p} = \frac{\exp(-0.059 \cdot age + 0.385 \cdot index - 0.089 \cdot temperature)}{1 + \exp(-0.059 \cdot age + 0.385 \cdot index - 0.089 \cdot temperature)} \tag{3}$$

►Όσον αφορά τη μεταβλητή age, δηλαδή την ηλικία του ασθενή, η ερμηνεία του συντελεστή είναι ότι αν αυξηθεί κατά ένα χρόνο η ηλικία, τότε οι πιθανότητες ανταπόκρισης στη θεραπεία πολλαπλασιάζονται με $\exp(-0.059) = 0.9427068$, δηλαδή μειώνονται κατά περίπου 6%. Τα όρια εμπιστοσύνης αυτής της μεταβολής δίνονται από τους συντελεστές μέσα από την εκθετική συνάρτηση, δηλαδή δίνουν μείωση από περίπου 11% έως περίπου 1%.
►Αντίστοιχα για τον δείκτη κυττάρων λευχαιμίας (index), αν αυξηθεί αυτός ο δείκτης κατά μια μονάδα, η πιθανότητα ανταπόκρισης στη θεραπεία αυξάνεται κατά $\exp(0.385) = 1.469614$, δηλαδή κατά περίπου 47%. Τα αντίστοιχα όρια εμπιστοσύνης είναι από περίπου 20% έως περίπου 95% αύξηση.
►Τέλος, για τη μεταβλητη temparature, μεταβολή κατά μια μονάδα (δηλαδή $\times 10^o F$) δείχνει μείωση της πιθανότητας ανταπόκρισης στη θεραπεία κατά $\exp(-0.089) = 0.91$ δηλαδή μείωση κατά περίπου 9% με τα όρια εμπιστοσύνης να είναι από περίπου 16% μείωση έως και 2.5%.

2..4 Να εξεταστεί η προβλεπτική ικανότητα του τελικού μοντέλου μέσω μιας καμπύλης ROC. Χ

```
suppressMessages(library(pROC))
suppressMessages(roc1<-roc(data$response,fitted.values(model),smooth=TRUE))
suppressMessages(roc2<-roc(data$response,fitted.values(model3),smooth=TRUE))
plot(roc1, col = 1, lty = 1, main = "ROC")
plot(roc2, col = 2, lty = 2, add = TRUE)
legend("bottomright", c("Initial Model", "Final Model"),
col = c(1,2), lty = c(1,2), merge = TRUE)
```



Σχήμα 14: ROC curve of initial and final model

Παραπάνω, βλέπουμε την καμπύλη ROC για το αρχικό μοντέλο, με όλες τις μεταβλητές, καθώς και για το τελικό μοντέλο. Το αρχικό μοντέλο έχει AUC (Area Under Curve) ίσο με 0.8962 ενώ **το τελικό έχει AUC ίσο με 0.868624**. Και τα δύο νούμερα είναι αρκετά υψηλά, δεδομένου ότι η μέγιστη τιμή είναι η μονάδα, και παρατηρούμε ότι το τελικό μας μοντέλο παρότι έχει καλύτερη προσαρμογή, εμφανίζει μια ελάχιστα μικρότερη προβλεπτική ικανότητα, που όμως δεν είναι σημαντική, δεδομένου του αριθμού των παρατηρήσεων που έχουμε να κάνουμε.