Στατιστική Μοντελοποίηση

- **ΔΠΜΣ Επιστήμη Δεδομένων & Μηχανική Μάθηση**
- **Ορφανουδάκης Φίλιππος Σκόβελεφ ΑΜ:03400107**
- **ΣΕΙΡΑ 3**

A) asfalies

i)

Προσαρμόζουμε το μοντέλο Poisson παλινδρόμησης θετοντας την κατηγορική μεταβλητή cartype ως factor. Θα προσπαθήσουμε να βρούμε την εξάρτηση της μεταβλητής y με τις μεταβλητές cartype, agecat και district εφαρμόζοντας την backwise selection. Η υλοποίηση πραγματοποιείται στην R η οποία πραγματοποιεί κάθε βήμα με κριτήριο την μείωση της τιμής του AIC. Προσθέτοντας την εντολή test="Chisq" μπορούμε να δούμε και την p-τιμή για την τιμή του ελέγχου Wald . Υπενθυμίζουμε ότι μια μεγάλη τιμή για τον έλεγχο Wald ή αντίστοιχα μια μικρή p-τιμή είναι ικανή να απορρίψει την μηδενική υπόθεση για κάθε μια μεταβλητή ξεχωριστά. Τα αποτελέσματα που παίρνουμε είναι τα εξής :

```
> step(model,method="backward", test="Chisq")
Start:  AIC=222.15
y ~ factor(cartype) + 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 ***
- factor(cartype)  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 ~ factor(cartype) + agecat + district + offset(log(
    family = poisson, data = asfalies)

Coefficients:
     (Intercept)  factor(cartype)2  factor(cartype)3  factor(cartype)4
         -1.9352            0.1622            0.3953            0.5654
          agecat          district
         -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
```

Η τεχνική backwards μας οδήγησε στο συμπέρασμα ότι το μοντέλο με όλες τις μεταβλητές είναι το πιο κατάλληλο. Αυτο φαίνεται από τα εξής 2 σημεία:

1. Το χαμηλότερο AIC επιτυγχάνεται με το πλήρες μοντέλο
2. Τα p-values είναι όλα <0.001 δηλαδή στατιστικά σημαντικά.

Φυσικά η τιμή της συνάρτησης deviance για το πλήρες μοντέλο είναι 41.79 και για το κενό μοντέλο ειναι 207.8, δηλαδή η διαφορά τους είναι 166.01 μια μεγάλη τιμή που οδηγεί στο να προτιμήσουμε το πλήρες από το κενό μοντέλο. Για λόγους πληρότητας θα δοκιμάσουμε διαφορετικά μοντέλα για να δείξουμε ότι το πλήρες είναι το βέλτιστο.

```
mod1<-glm(y ~ factor(cartype)+district+offset(log(n)), data = asfalies,family=poisson)

mod2<-glm(y ~ factor(cartype)+agecat+offset(log(n)), data = asfalies,family=poisson)

mod3<-glm(y ~ agecat+district+offset(log(n)), data = asfalies,family=poisson)

mod4<-glm(y ~ district+offset(log(n)), data = asfalies,family=poisson)


summary(mod1)
anova(mod1,model,test="Chisq")
AIC(mod1)

summary(mod2)
anova(mod2,model,test="Chisq")
AIC(mod2)

summary(mod3)
anova(mod3,model,test="Chisq")
AIC(mod3)
```

Τα αποτελέσματα που παίρνουμε είναι τα εξής:

	AIC	Deviance (model-modi)	p-value for Wald test
mod1	286.32	66.171	<4.125e-16 ***
mod2	233.0869	12.938	0.000322 ***
mod3	306.07	89.925	<2.2e-16 ***

Επιλέγουμε το πλήρες μοντέλο για τους εξής λόγους :

1. Έχει το χαμηλότερο AIC
2. Η τιμη της ελεγχοσυνάρτησης deviance D(modi)-D(model) είναι μεγάλη τιμή γεγονός που δείχνει ότι με την προσθήκη των μεταβλητών του πλήρες μοντέλου έχουμε βελτίωση της τιμής deviance.

3. Η p-value που αντιστοιχεί στο P(|z|>Wald-value) είναι μικρή , δηλαδή απορρίπτουμε την μηδενική υπόθεση για κάθε μια από τις μεταβλητές.

Τέλος έχοντας κάνει την επιλογή του μοντέλου μας θα υπολογίσουμε την p-value με τον εξής τρόπο :

```
1-pchisq(model$deviance,model$df.residual)
```

και σαν αποτέλεσμα έχουμε **0.02580847.**

Είναι αρκετά μικρή τιμή γεγονός που μας προβληματίζει καθώς μας οδηγεί στο συμπέρασμα ότι δεν περιγράφεται πλήρως η μεταβλητότητα της εξαρτημένης μεταβλητής Y.
Παρακάτω με τη βοήθεια των υπολοίπων θα προσπαθήσουμε να εντοπισουμε αν υπάρχει κάποιο πρόβλημα

ii)

Οι αντίστοιχες τιμές για τα διαστήματα εμπιστοσύνης 95% είναι τα εξής :

```
> confint.default(model)
                    2.5 %     97.5 %
(Intercept)      -2.04350208 -1.8269440
factor(cartype)2  0.06329746  0.2611664
factor(cartype)3  0.28772397  0.5029705
factor(cartype)4  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
factor(cartype)2 1.0653437 1.2984438
factor(cartype)3 1.3333892 1.6536260
factor(cartype)4 1.5280757 2.0275915
agecat           0.6290616 0.7489890
district         1.1072684 1.3928124
```

και η ερμηνεία του μοντέλου είναι η εξής (exp(Estimate)) :

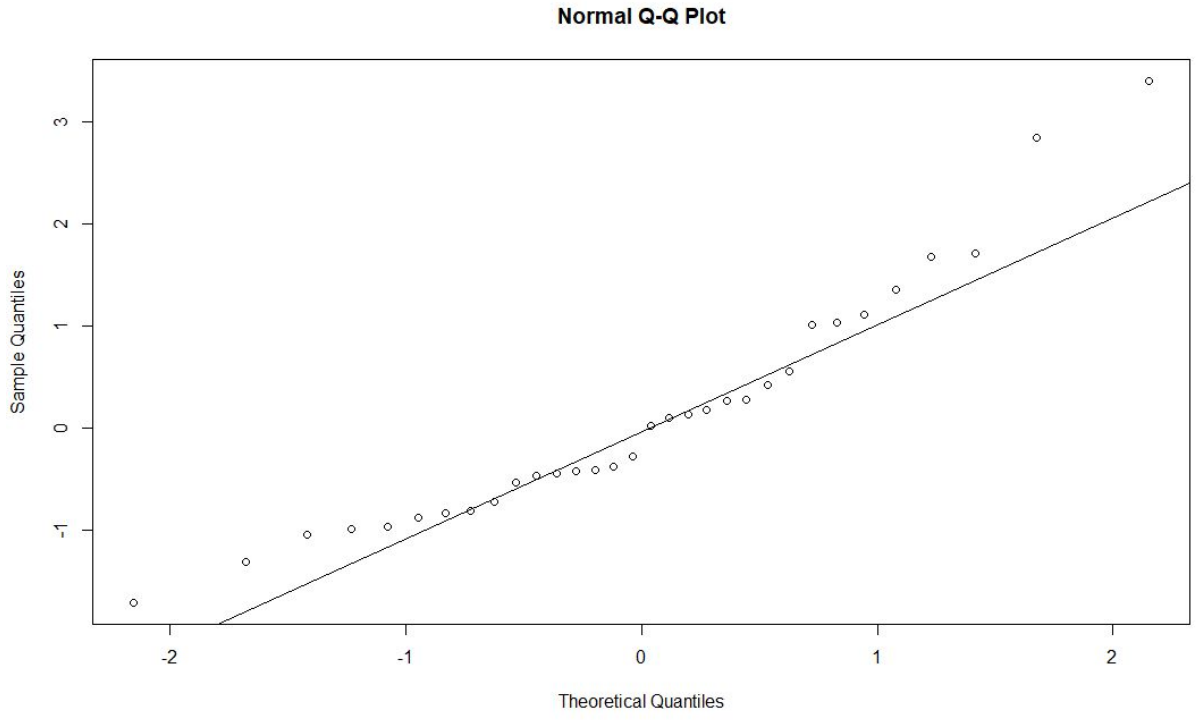
```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.93522    0.05525 -35.030  < 2e-16 ***
factor(cartype)2  0.16223    0.05048   3.214 0.001309 **
factor(cartype)3  0.39535    0.05491   7.200 6.03e-13 ***
factor(cartype)4  0.56543    0.07215   7.836 4.64e-15 ***
agecat           -0.37628    0.04451  -8.453  < 2e-16 ***
district          0.21661    0.05853   3.701 0.000215 ***
```

- Για κάθε 1 χρόνο που αυξάνεται η ηλικία έχουμε 31% μείωσης του ποσού αποζημίωσης

- Αν η περιοχή κατοικίας είναι η Αθήνα τότε έχουμε αύξηση κατα 23% του ποσού αποζημίωσης συγκριτικά με τις άλλες περιοχές κατοικίας
- Αν ο τύπος του αμαξιού είναι 2 τότε συγκριτικά με τον τύπο 1 έχω αύξηση του ποσού αποζημίωσης κατά 17% , αν ο τύπος είναι 3 τότε συγκριτικά με τον τύπο 1 έχω αύξηση 27% και τέλος αν ο τύπος είναι 4 έχω αύξηση κατα 68% συγκριτικά με τον τύπο 1

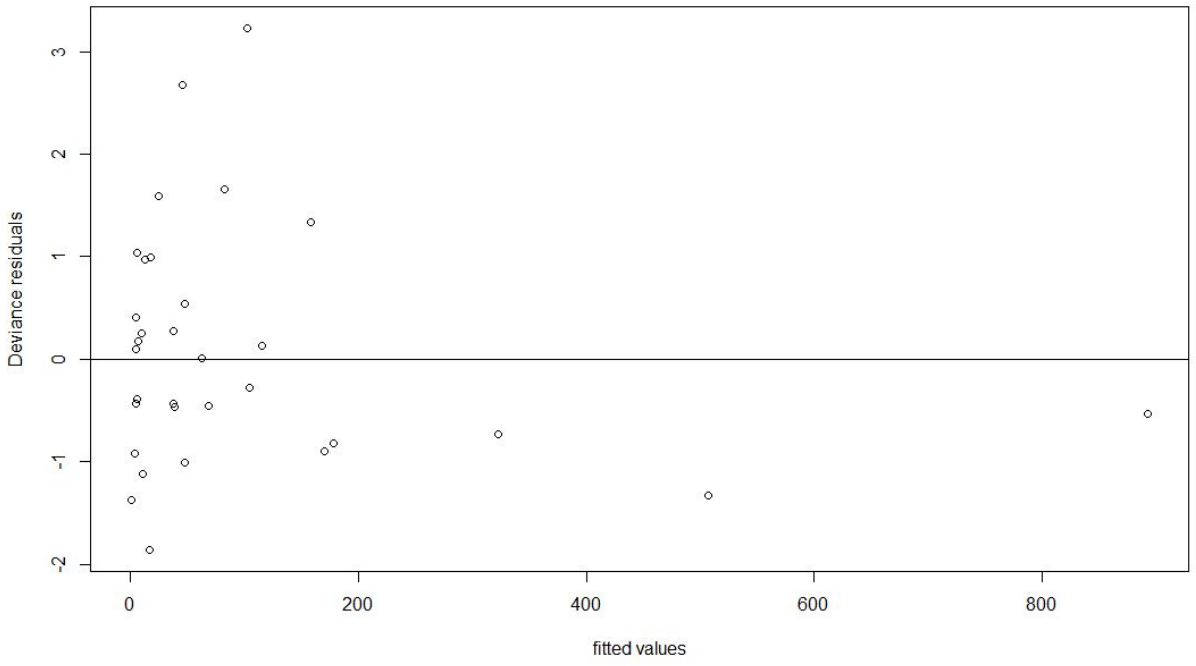
iii)

Υπόλοιπα Pearson - QQ plot



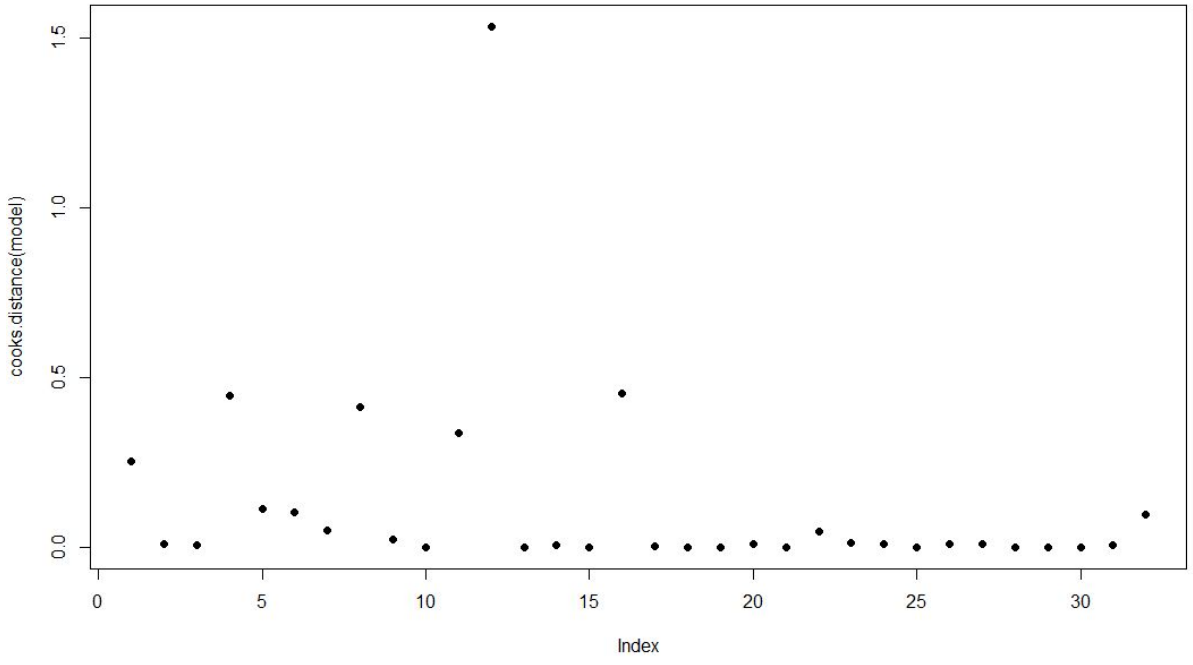
- Παρατηρούμε ότι 2 σημεία είναι υποψήφια άτυπα σημεία όπως φαίνεται και στη γραφική παράσταση. Τα σημεία αυτά είναι τα πάνω δεξιά που αντιστοιχούν στην 1η και στην 11η παρατήρηση. Τα υπόλοιπα σημεία σχηματίζουν μια αρκετά καλά ορισμένη ευθεία.

Υπόλοιπα Deviance - fitted values



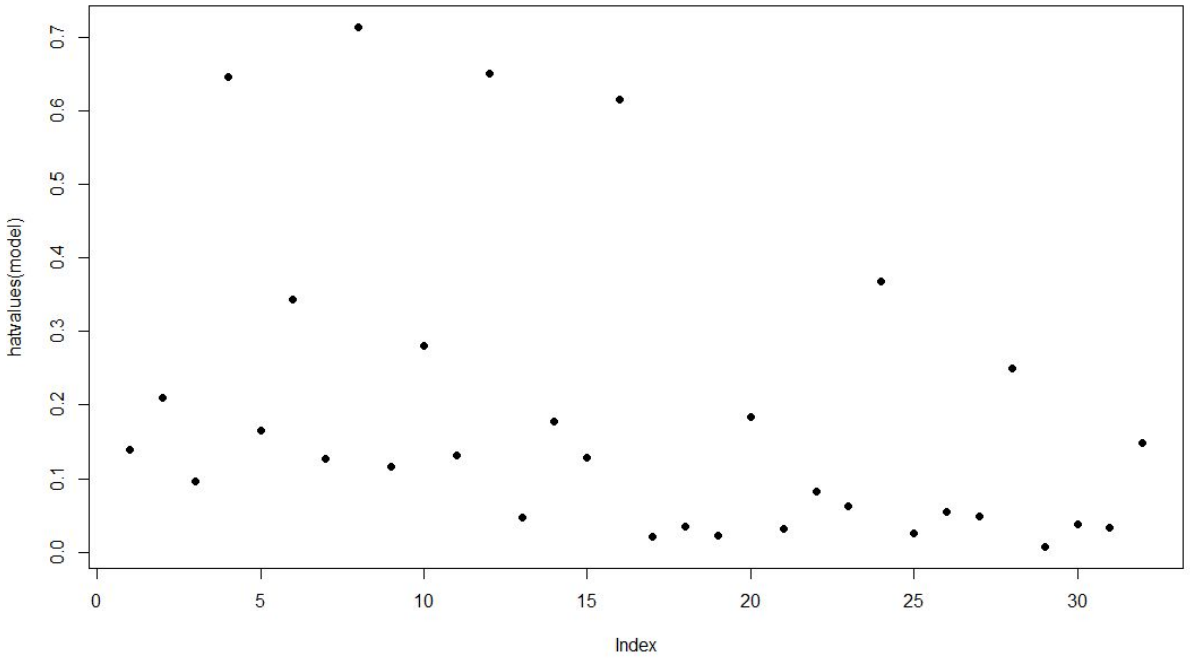
- Δεν παρατηρούμε κάποια εξάρτηση μεταξύ των παρατηρήσεων. Επίσης φαινεται να μην παραβιάζεται η ομοσκεδαστικότητα.

Αποστάσεις Cook



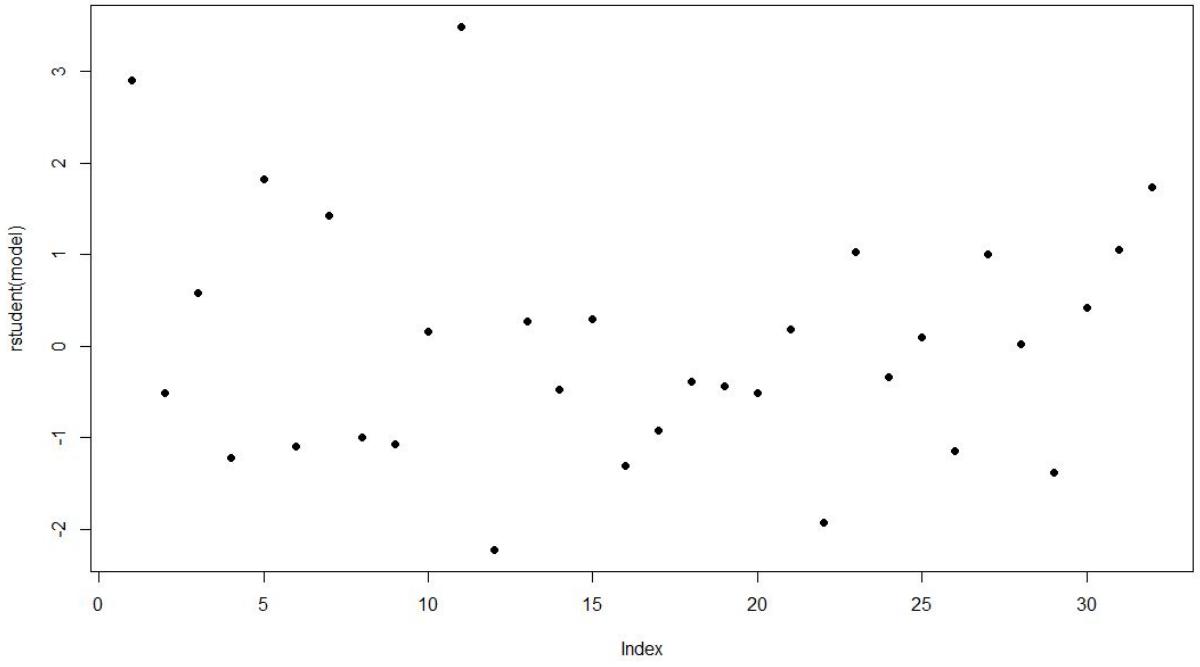
- Η 12η παρατήρηση βλέπουμε ότι έχει τιμή μεγαλύτερη από 1 επομένως θεωρούμε ότι είναι σημείο επιρροής. Φυσικά θα ελέγξουμε και τα Hi(leverages) για να έχουμε μια ολοκληρωμένη εικόνα.

Hii (leverages)



- Από το παραπάνω διάγραμμα προκύπτουν αρκετά σημεία επιρροής τα οποία δεν έχουν φανεί από τις αποστάσεις Cook. Παρόλα αυτά η 12η παρατήρηση προκύπτει και εδώ σαν σαν σημείο επιρροής, επομένως μπορούμε με σιγουριά να πούμε ότι η 12η παρατήρηση είναι σημείο επιρροής.

Likelihood residuals



- Η παρατήρηση 11 αλλά και οριακά η παρατήρηση 1 προκύπτουν σαν outliers. Παρατηρώντας την αντίστοιχη student t κατανομή για 27 βαθμούς ελευθερίας βλέπουμε ότι η πλειοψηφία των σημείων θα έπρεπε να είναι μέσα στις τιμές (-2.051831 ,2.051831).

B) leukaemia

i) Ορίζουμε το πλήρες μοντέλο των 6 συμμεταβλητών και παίρνουμε τα εξής αποτελέσματα:

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

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
```

Παρατηρούμε απο τις p-τιμές των ελέγχων Wald ότι στο πλήρες μοντέλο δεν είναι στατιστικά σημαντικές οι συμμεταβλητές, καθώς όλες έχουν τιμή μεγαλύτερη από 0.001. Για τον λόγο αυτό θα εξετάσουμε διαφορετικά μοντέλα με σκοπό να βελτιώσουμε τόσο τις τιμές ελέγχων Wald , αλλά να επιτύχουμε χαμηλότερο deviance και AIC.

Αρχικά δοκιμάζουμε ορισμένους μετασχηματισμούς. Πιο συγκεκριμένα κατασκευάσαμε τα εξής μοντέλα :

```
model1<- glm(response ~ age+smear+infiltrate+log(index)+blasts+temperature, data =
leukaimia,family=binomial)

model2<- glm(response ~ log(age)+smear+infiltrate+index+blasts+temperature, data =
leukaimia,family=binomial)

model3<- glm(response ~ age+log(smear)+infiltrate+index+blasts+temperature, data =
leukaimia,family=binomial)

model4<- glm(response ~ age+smear+log(infiltrate)+index+blasts+temperature, data =
leukaimia,family=binomial)

model5<- glm(response ~ age+smear+infiltrate+index+blasts+log(temperature), data =
leukaimia,family=binomial)

model6<- glm(response ~ age^2+smear+infiltrate+index+blasts+temperature, data =
leukaimia,family=binomial)
```

΄Υστερα από αυτές τις δοκιμές το AIC και το deviance μειώθηκαν περίπου κατα μια μονάδα. Μια τόσο μικρή αλλαγή σε ένα μικρό dataset φυσικά δεν μας ενδιαφέρει, επομένως παραμένουμε στο αρχικό πλήρες μοντέλο. Στη συνέχεια θα πραγματοποιήσουμε stepwise

backwards μέθοδο και θα κάνουμε τους απαραίτητους ελέγχους για να δούμε αν θα επιλέξουμε ένα μοντέλο με λιγότερες συμμεταβλητές.

Το προτεινόμενο μοντέλο είναι το εξής :

```
glm(formula = response ~ age + infiltrate + index + temperature,
    family = binomial, data = leukaimia)
```

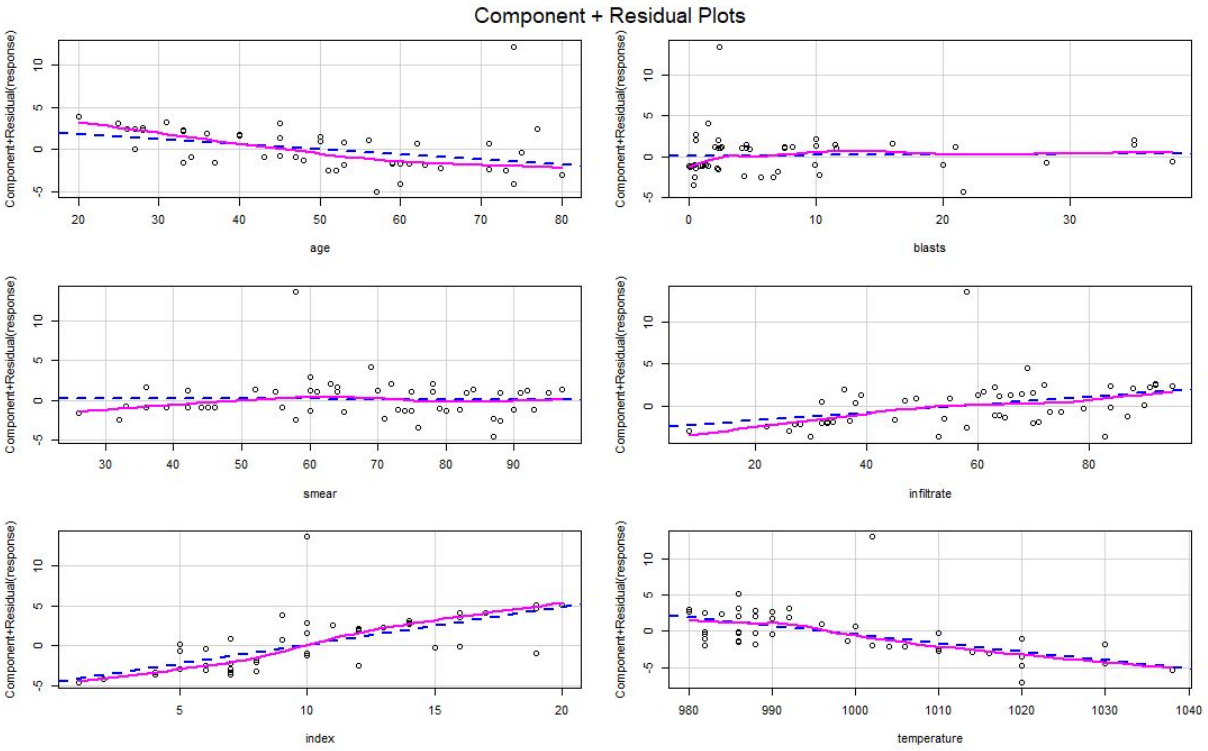
Με το εξής χαρακτηριστικά :

```
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
```

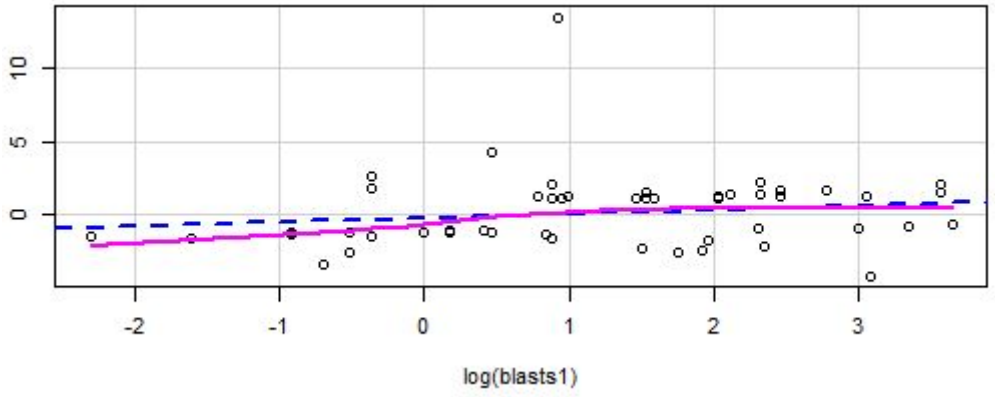
Παρατηρούμε ότι οι τιμές του deviance και AIC δεν είναι αισθητά βελτιωμένες για να επιλέξουμε με σιγουριά αυτό το μοντέλο. Τα p-values απο τον έλεγχο Wald είναι ελαφρώς βελτιωμένα , απολύτως αναμενόμενο αφού έχουμε λιγότερες συμμεταβλητές αλλά ακόμα και σε αυτό το μοντέλο δεν δείχνουν να είναι στατιστικά σημαντικές.

ii)

<u>Partial Residuals</u>



- Από τα παραπάνω διαγράμματα παρατηρούμε τυχον μετασχηματισμούς που μπορούμε να πραγματοποιήσουμε. Ο μετασχηματισμός που αποφασίζουμε να κάνουμε είναι ο εξής log(blasts1) <- log(blasts+0.1) και έχουμε πλέον το νέο διάγραμμα για την συμμεταβλητή log(blasts1).



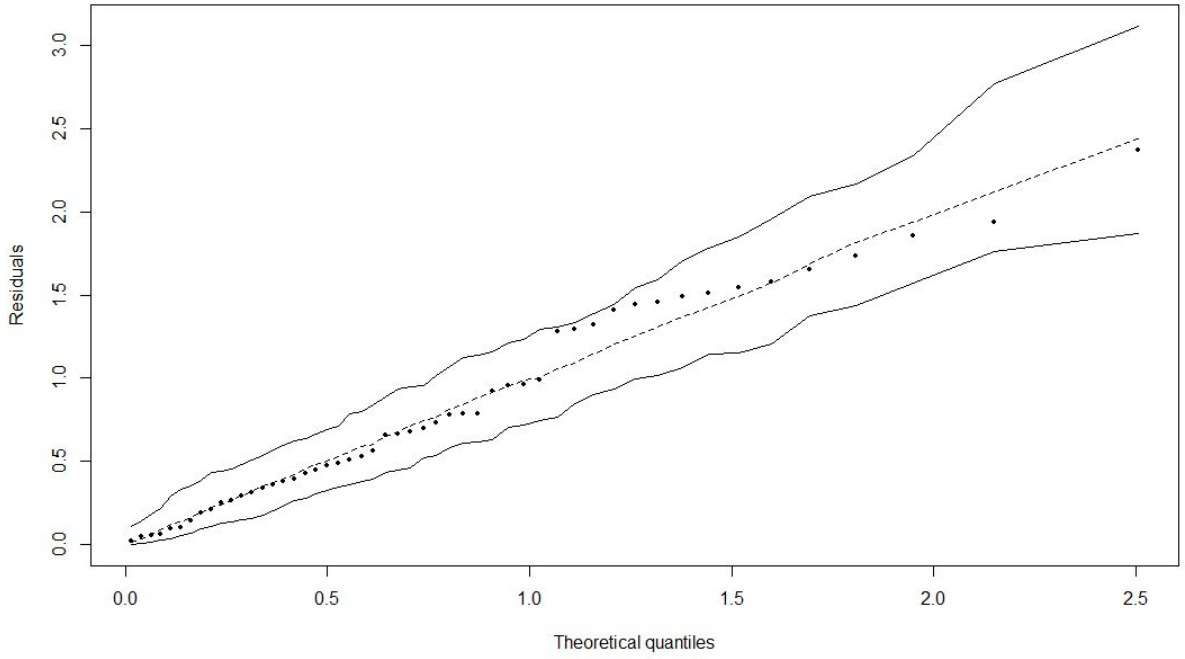
Το νέο μοντέλο έχει :

- AIC : 54.007
- Deviance : 40.007
- και τα αντιστοιχα p-values για τον Wald έλεγχο είναι

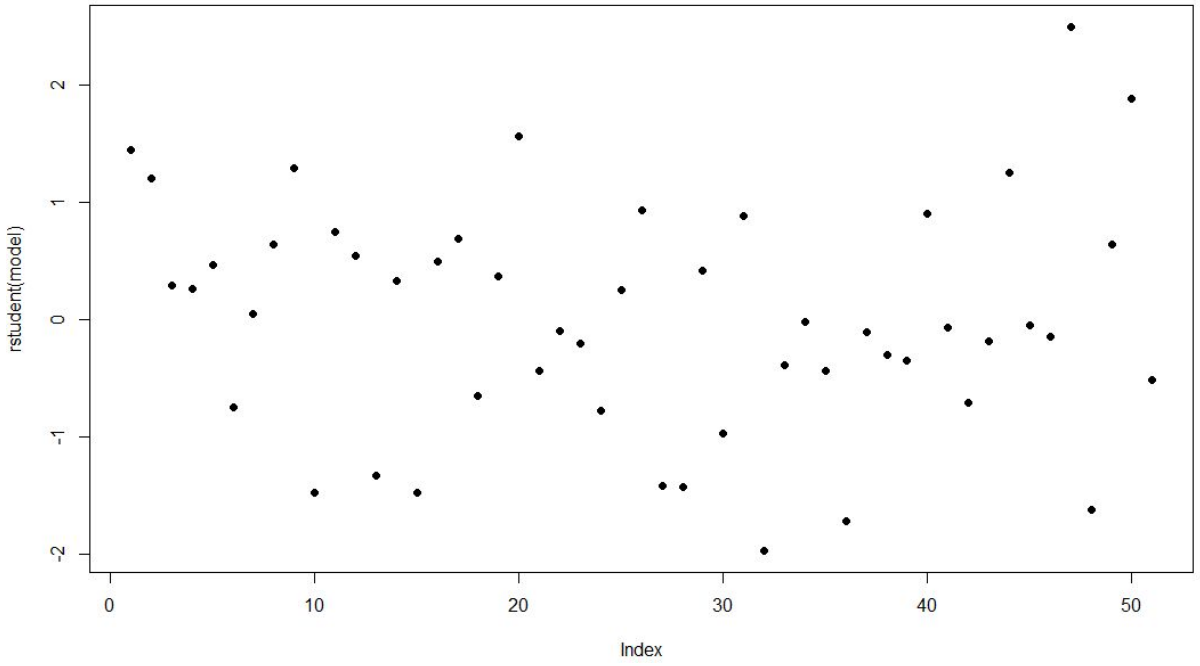
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 96.939337  38.696224   2.505  0.01224 *
age         -0.059638   0.027337  -2.182  0.02914 *
log(blasts1) 0.140176   0.423084   0.331  0.74040
smear        0.001078   0.045047   0.024  0.98091
infiltrate   0.030134   0.044638   0.675  0.49964
index        0.388173   0.137913   2.815  0.00488 **
temperature -0.100672   0.039714  -2.535  0.01125 *
```

Συνεπώς η αλλαγή που έγινε δεν φαίνεται να χειροτερεύει το μοντέλο .

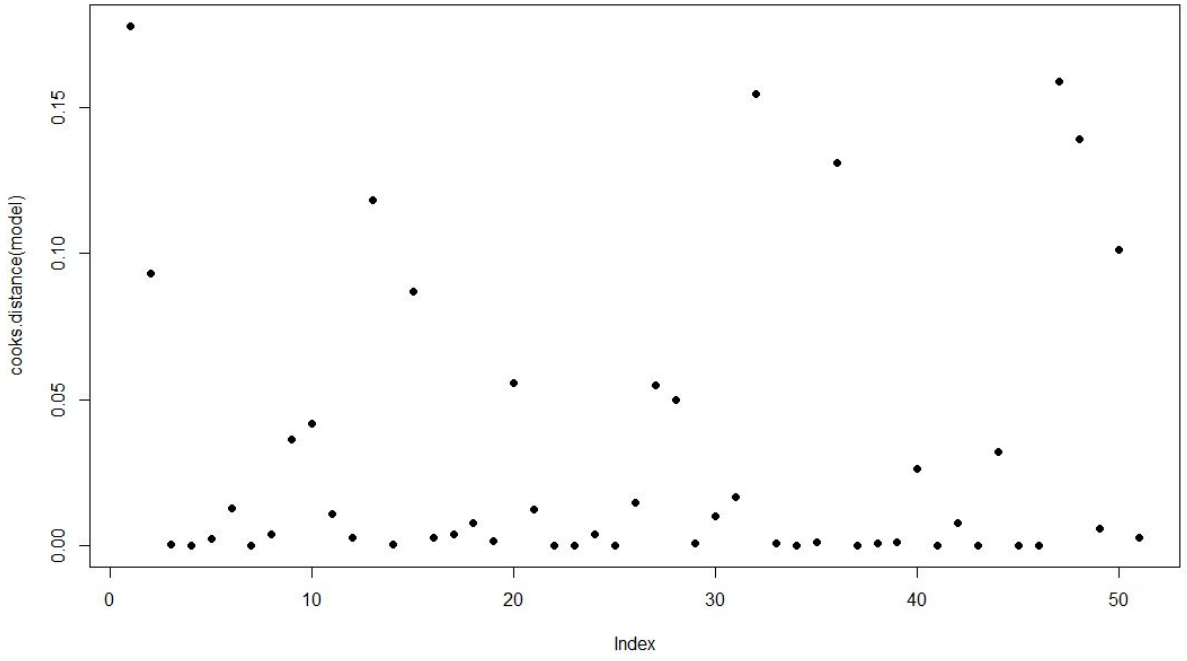
Υπόλοιπα Deviance (Ημι-κανονική κατανομή)



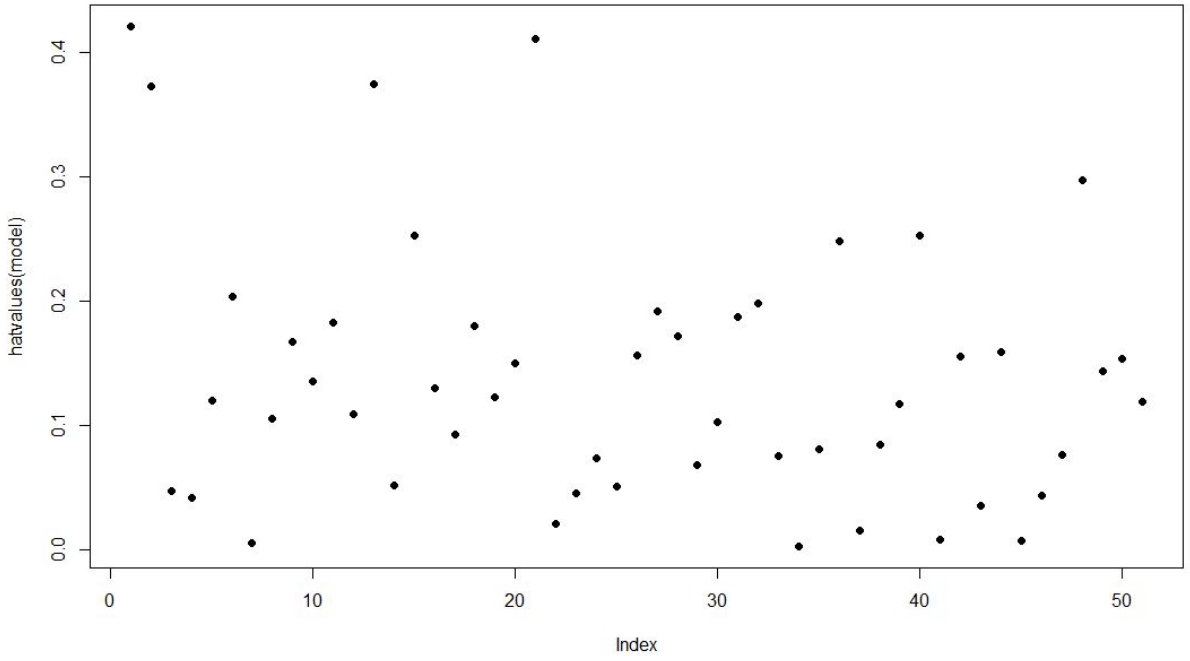
Υπόλοιπα Πιθανοφάνειας



Απόσταση Cook



Hii (leverages)



Τα παραπάνω 3 διαγράμματα τα χρησιμοποιούμε έτσι ώστε να εντοπίσουμε σημεία επιρροής . Από τα υπόλοιπα πιθανοφάνειας ύποπτα σημεία είναι όσα βρίσκονται εκτός του διαστήματος(-2,2) όπως παρατηρούμε απο την t student κατανομή για τους αντίστοιχους βαθμούς ελευθερίας, επομένως έχουμε την 47η παρατήρηση πιθανό σημείο επιρροής. Από την απόσταση Cook δεν προκύπτει κάποια έντονη ένδειξη, όμως από τις τιμές των hii έχουμε σαν κριτήριο την τιμή το 0.27 , που όπως παρατηρούμε είναι 5 παρατηρήσεις ύποπτες. Παρόλα αυτά αν συνδυάσουμε και τα 3 κανένα σημείο δεν εμφανίζεται με συνέπεια σε τουλάχιστον από τα διαγράμματα μπορούμε να συνεχίσουμε με το μοντέλο μας με αρκετή ασφάλεια ως προς την ορθότητα του.

iii)

Οι αντίστοιχες τιμές για τα διαστήματα εμπιστοσύνης 95% είναι τα εξής :

```
> confint.default(model)
                    2.5 %        97.5 %
(Intercept)  21.09613163 172.782543091
age          -0.11321795  -0.006058765
log(blasts1) -0.68905367   0.969405851
smear        -0.08721238   0.089368086
infiltrate   -0.05735568   0.117622938
index         0.11786832   0.658477263
temperature  -0.17850896  -0.022834321
> exp(confint.default(model))
                     2.5 %       97.5 %
(Intercept)  1.451889e+09 1.092710e+75
age          8.929560e-01 9.939596e-01
log(blasts1) 5.020510e-01 2.636378e+00
smear        9.164824e-01 1.093483e+00
infiltrate   9.442582e-01 1.124820e+00
index        1.125096e+00 1.931848e+00
temperature  8.365166e-01 9.774244e-01
```

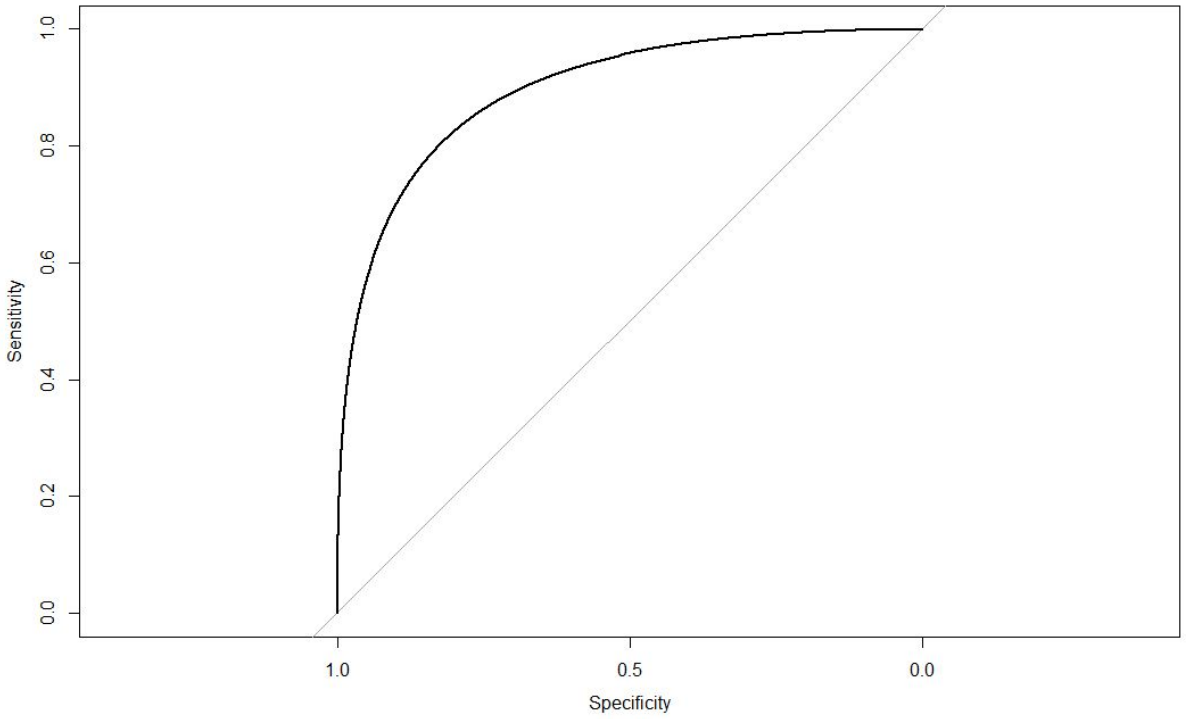
Και οι αντίστοιχη ερμηνεία είναι η εξής (exp(Estimate)):

```
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  96.939337  38.696224   2.505  0.01224 *
age          -0.059638   0.027337  -2.182  0.02914 *
log(blasts1)  0.140176   0.423084   0.331  0.74040
smear         0.001078   0.045047   0.024  0.98091
infiltrate    0.030134   0.044638   0.675  0.49964
index         0.388173   0.137913   2.815  0.00488 **
temperature  -0.100672   0.039714  -2.535  0.01125 *
```

- Με την αύξηση κατα 1 χρόνο της ηλικίας η πιθανότητα να έχουμε ανταπόκριση στη θεραπεία μειώνεται κατα 5%
- Με την αυξηση κατα 1 μονάδας του log(blast1) , του smear , του infiltrate και του index έχουμε αύξηση της πιθανότητας ανταπόκρισης κατα 15%, 0.1%, 3% και 46% αντίστοιχα.
- Με την αυξηση κατα 1 μονάδας της θερμοκρασίας έχουμε μείωση της πιθανότητας ανταπόκρισης κατά 10%

iv)

Η καμπύλη ROC του τελικού μας μοντέλου :



Area under the curve: 0.8964