ΔΠΜΣ Επιστήμη Δεδομένων και Μηχανική Μάθηση

Στατιστική Μοντελοποίηση

3η Σειρά Ασκήσεων

Ονοματεπώνυμο: Ευάγγελος Τσόγκας

Αριθμός Μητρώου: 03400120

1. Παλινδρόμηση Poisson

(i) Προσαρμογή μοντέλου

Χρησιμοποιώντας τα δεδομένα στο asfalies.txt προσαρμόζουμε ένα μοντέλο παλινδρόμησης Poisson με εξαρτημένη μεταβλητή τον αριθμό αποζημιώσεων y λόγω τροχαίων ατυχημάτων ανά n συμβόλαια. Προκειμένου να λάβουμε υπόψη το μέγεθος του πληθυσμού περιλαμβάνουμε την μεταβλητή n στο μοντέλο ως **offset(ln n)** με γνωστό συντελεστή. Επίσης, δηλώνουμε την κατηγορία ασφαλίστρων cartype ως κατηγορική με την εντολή **factor(cartype)**. Στην παρακάτω εικόνα φαίνονται οι πληροφορίες του προσαρμοσμένου μοντέλου.

```
Call:
glm(formula = y ~ agecat + factor(cartype) + 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 ***
agecat           -0.37628    0.04451  -8.453  < 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 ***
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
```

Έλεγχος Wald

Για τη σημαντικότητα των επεξηγηματικών μεταβλητών στο μοντέλο παρατηρούμε από τον στατιστικό έλεγχο Wald (z-value) πως όλες σχετίζονται με τη μεταβλητή y, αφού οι p-τιμές τους είναι αρκετά μικρές, επομένως χρειάζονται στο μοντέλο.

Έλεγχος Deviance

Στη συνέχεια εξετάζουμε την ελεγχοσυνάρτηση deviance. Παρακάτω κάνουμε σύγκριση του μοντέλου μας με το κορεσμένο μοντέλο και βλέπουμε ότι η τιμή του ελέγχου είναι πολύ μεγάλη και η p-τιμή του πολύ μικρή, αρά συμπεραίνουμε ότι μπορούμε να απορρίψουμε τη μηδενική υπόθεση.

```
Analysis of Deviance Table

Model 1: y ~ 1
Model 2: y ~ agecat + factor(cartype) + district + offset(log(n))
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        31     5660.6
2        26       41.8  5   5618.8 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Μετά εξετάζουμε πόσο σημαντική είναι κάθε μεταβλητή σε σχέση με το null μοντέλο και παρακάτω βλέπουμε πως για όλες τις μεταβλητές η p-τιμή είναι αρκετά μικρή, άρα είναι σημαντικές.

```
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 ***
factor(cartype) 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
```

Επιπλέον, εξετάζουμε και την p-τιμή του ελέγχου deviance για τη σημαντικότητα του μοντέλου μας (την τιμή του ελέγχου την έχουμε ήδη: Residual deviance = 41.789). Στη παρακάτω εικόνα φαίνεται πως η p-τιμή είναι 0.0258, η οποία δεν είναι μικρή, άρα το μοντέλο δεν μπορεί να θεωρηθεί ικανοποιητικό. Παρ’ όλα αυτά το δείγμα μας είναι μικρό οπότε ο έλεγχος μάλλον είναι αναξιόπιστος, αφού ο ισχυρισμός της κατανομή χ^2 των τιμών της deviance ισχύει μόνο ασυμπτωματικά.

```
> 1 - pchisq(model$deviance, model$df.residual)
[1] 0.02580847
```

Κριτήριο AIC

Επίσης, μπορούμε να εφαρμόσουμε backward elimination συγκρίνοντας μοντέλα με βάση το κριτήριο AIC. Το αποτέλεσμα είναι ότι από το μοντέλο δεν αφαιρέθηκε καμία μεταβλητή και όπως φαίνεται και στην παρακάτω εικόνα το αρχικό μοντέλο μας έχει το μικρότερο AIC, άρα το προτιμάμε.

```
Start:  AIC=222.15
y ~ agecat + factor(cartype) + 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
```

(ii) Διαστήματα εμπιστοσύνης

Στην παρακάτω εικόνα φαίνονται τα διαστήματα εμπιστοσύνης για τους συντελεστές των μεταβλητών. Παρατηρούμε πως δεν περιέχεται το 0 στα διαστήματα εμπιστοσύνης, το οποίο συμφωνεί με τους στατιστικούς ελέγχους Wald, των οποίων τα p-values είναι πολύ μικρά και άρα απορρίπτουμε τις υποθέσεις H_0 για τις επεξηγηματικές μεταβλητές του μοντέλου.

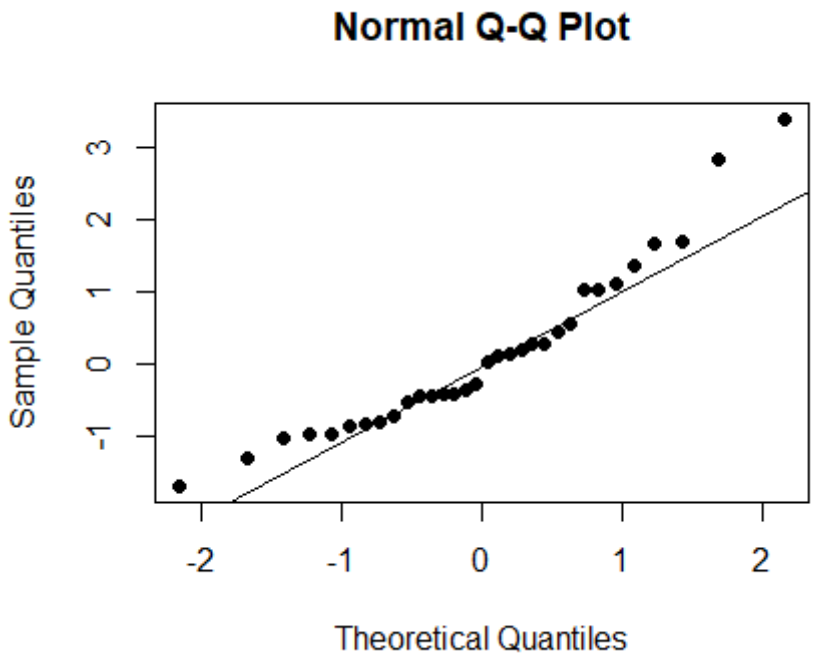
```
> confint.default(model)
                        2.5 %     97.5 %
(Intercept)       -2.04350208 -1.8269440
agecat            -0.46352606 -0.2890309
factor(cartype)2   0.06329746  0.2611664
factor(cartype)3   0.28772397  0.5029705
factor(cartype)4   0.42400923  0.7068487
district           0.10189607  0.3313250
> exp(confint.default(model))
                      2.5 %    97.5 %
(Intercept)       0.1295741 0.1609045
agecat            0.6290616 0.7489890
factor(cartype)2  1.0653437 1.2984438
factor(cartype)3  1.3333892 1.6536260
factor(cartype)4  1.5280757 2.0275915
district          1.1072684 1.3928124
```

Ερμηνεία συντελεστών

- Ο συντελεστής του **agecat** είναι -0.3762. Αν αυξηθεί η ηλικία του ασφαλισμένου κατά μια μονάδα, ο αριθμός αποζημιώσεων y θα πολλαπλασιαστεί με exp(-0.3762) = 0.69, δηλαδή αν ο ασφαλισμένος είναι μεγάλος σε ηλικία ο αριθμός αποζημιώσεων λόγω τροχαίων ατυχημάτων μειώνεται κατά 31%.
- Ο συντελεστής του **district** είναι 0.2166 και exp(0.2166) = 1.24. Άρα, αν η περιοχή διαμονής του ασφαλισμένου είναι η Αθήνα τότε ο αριθμός αποζημιώσεων λόγω τροχαίων αυξάνεται κατά 24%.
- Από τους συντελεστές των επιπέδων του **cartype** φαίνεται πως όσο μεγαλύτερο είναι το επίπεδο (μεγαλύτερος αριθμός κατηγορίας) τόσο πιο πολύ αυξάνεται ο αριθμός αποζημιώσεων. Για την κατηγορία ασφαλίστρων νούμερο 4 για παράδειγμα exp(0.5654) = 1.76, δηλαδή ο αριθμός αποζημιώσεων αυξάνεται κατά 76% σε σχέση με την κατηγορία 1.

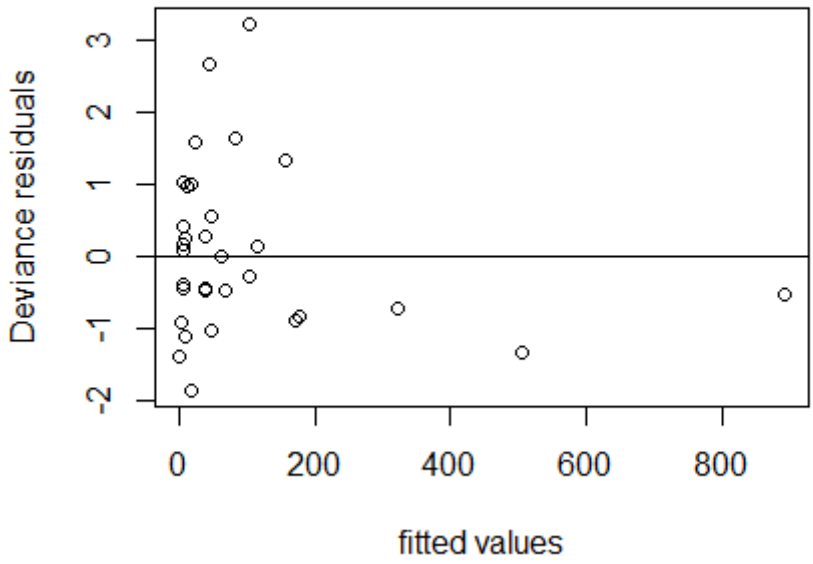
(iii) Υπόλοιπα Pearson

Στο παρακάτω διάγραμμα φαίνεται ο γραφικός έλεγχος των υπολοίπων Pearson. Δεν εξετάζουμε αν τα υπόλοιπα είναι της κανονικής κατανομής, αλλά για πιθανά άτυπα σημεία. Τα περισσότερα σημεία σχηματίζουν μια σχετικά καλά ορισμένη ευθεία, αλλά υπάρχουν 3 σημεία που πιθανώς είναι άτυπα, ένα κάτω αριστερά και δύο πάνω δεξιά, τα οποία αντιστοιχούν στις παρατηρήσεις 22, 1 και 11 αντίστοιχα.



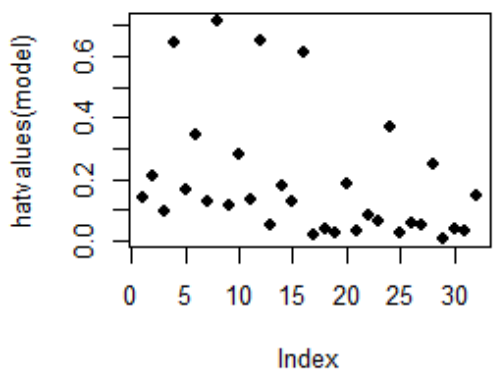
Υπόλοιπα Deviance

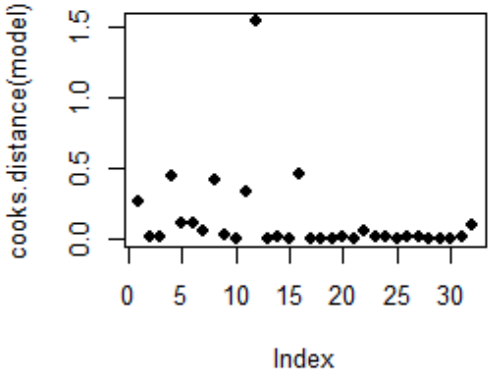
Παρακάτω φαίνεται το διάγραμμα των υπολοίπων deviance σε σχέση με τις προσαρμοσμένες τιμές θέλοντας να εξετάσουμε την υπόθεση ότι οι παρατηρήσεις είναι ανεξάρτητες μεταξύ τους. Από τη μορφή του είναι λίγο δύσκολο να βγάλουμε συμπέρασμα για την ανεξαρτησία λόγω του ότι υπάρχουν μερικά fitted values με πολύ μεγάλη τιμή, αλλά δεν φαίνεται να παραβιάζεται αυτή η υπόθεση. Πιθανώς άτυπα σημεία αποτελούν αυτά με τις μεγαλύτερες τιμές, δηλαδή οι παρατηρήσεις 1 και 11 με τιμές 2.67 και 3.23 αντίστοιχα.

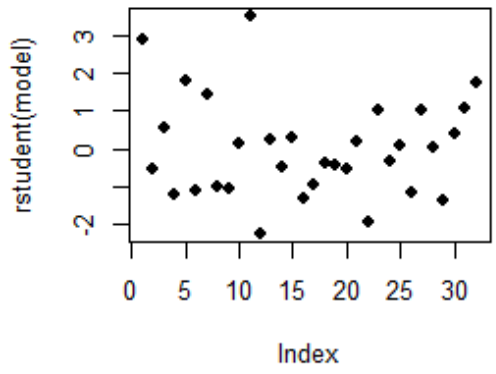


Σημεία επιρροής

Στα παρακάτω διαγράμματα φαίνονται τα index plots για τα hii, τις αποστάσεις Cook και τα υπόλοιπα πιθανοφάνειας, αντίστοιχα, με σκοπό τον εντοπισμό πιθανών σημείων επιρροής. Με βάση τα hat values πιθανά σημεία επιρροής με μεγάλες τιμές είναι τα 4, 8, 12 και 16. Με βάση τις αποστάσεις Cook το σημείο 12 έχει πολύ μεγάλη τιμή σε σχέση με τα άλλα. Τέλος, με βάση τα υπόλοιπα πιθανοφάνειας μεγάλες τιμές έχουν τα σημεία 1, 11 και 12. Είναι ενδιαφέρον ότι και οι τρεις γραφικές παραστάσεις συμφωνούν για την παρατήρηση 12, οπότε κατά πάσα πιθανότητα αποτελεί σημείο επιρροής.







2. Λογιστική Παλινδρόμηση

(i) Προσαρμογή Μοντέλου

Προσαρμόζουμε ένα μοντέλο λογιστικής παλινδρόμησης με εξαρτημένη μεταβλητή τη response, δηλαδή αν υπάρχει επιτυχής ανταπόκριση στην θεραπεία ή όχι. Το πρόβλημά μας αφορά δυαδικά δεδομένα. Στη παρακάτω εικόνα φαίνονται πληροφορίες για το προσαρμοσμένο μοντέλο.

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

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
```

Έλεγχος Wald

Παρατηρώντας τα αποτελέσματα του ελέγχου Wald βλέπουμε ότι καμία από τις επεξηγηματικές μεταβλητές δεν έχει p-τιμή < 0.001, επομένως δεν μπορούμε να απορρίψουμε τις μηδενικές υποθέσεις και το μοντέλο θα πρέπει να βελτιωθεί.

Κριτήριο AIC

Θα χρησιμοποιήσουμε backward elimination με βάση το κριτήριο AIC, ώστε να βρούμε κάποιο καλύτερο υποσύνολο μεταβλητών για το μοντέλο μας. Το μοντέλο που επιλέχθηκε όπως φαίνεται παρακάτω είναι αυτό από το οποίο έχουν αφαιρεθεί οι μεταβλητές smear και blasts. Έχει λίγο μικρότερο AIC από το αρχικό μοντέλο, αλλά και πάλι τα αποτελέσματα του ελέγχου Wald δεν είναι ικανοποιητικά. Παρ' όλα αυτά οι μεταβλητές που αφαιρέθηκαν πέρα από το γεγονός ότι βελτίωσαν το AIC ήταν και αυτές με τις μεγαλύτερες (με διαφορά) p-τιμές πράγμα που μας ωθεί στο να προτιμήσουμε το νέο μοντέλο.

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

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
```

Έλεγχος Deviance

Από τη στιγμή που έχουμε δυαδικά δεδομένα δεν μπορούμε να χρησιμοποιήσουμε την ελεγχοσυνάρτηση deviance για να διαπιστώσουμε την καταλληλότητα του μοντέλου, αλλά θα συγκρίνουμε το αρχικό μοντέλο με αυτό που επιλέχθηκε από το backward elimination. Στην παρακάτω εικόνα φαίνεται ο πίνακας της ανάλυσης deviance για τα δύο μοντέλα. Το αρχικό μοντέλο έναντι αυτού που επιλέχθηκε από το backward elimination δεν έχει σημαντικά ελαττωμένο υπόλοιπο deviance και η p-τιμή είναι αρκετά μεγάλη, επομένως ως καλύτερο μοντέλο θα επιλέξουμε αυτό που είναι απλούστερο και με μικρότερο AIC, δηλαδή το:

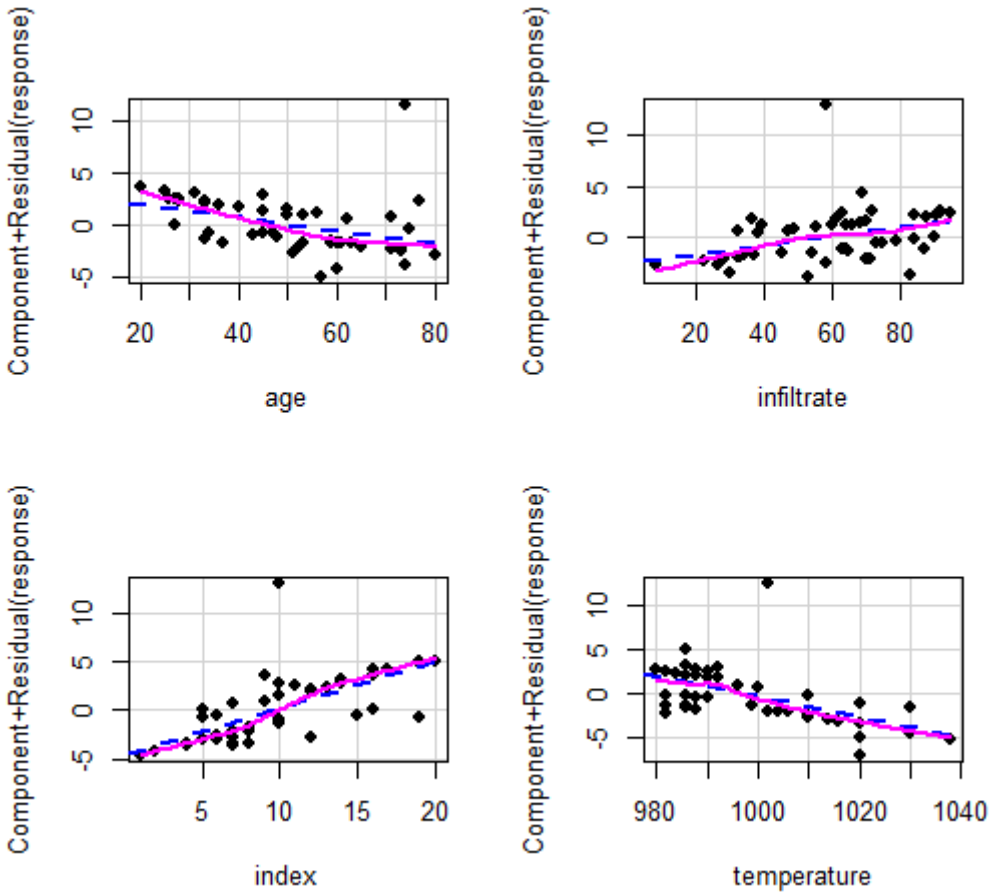
response ~ age + infiltrate + index + temperature

```
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
```

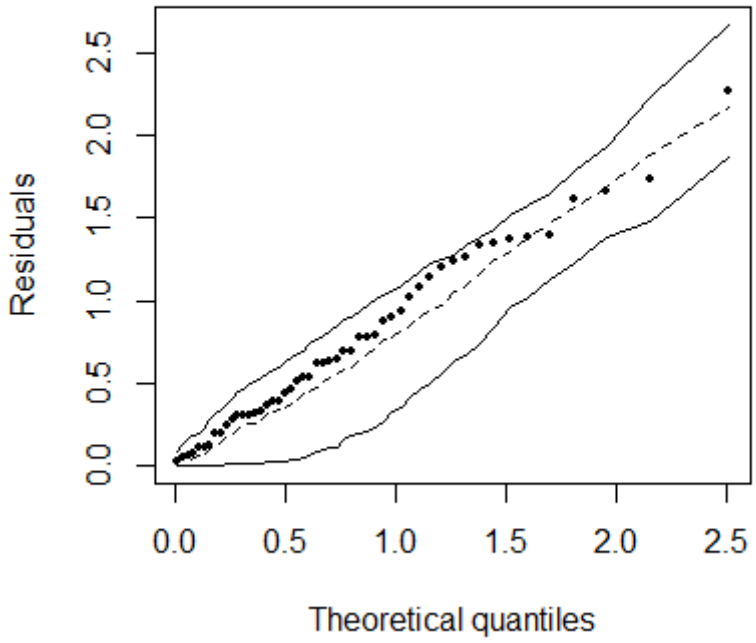
(ii) Έλεγχος μερικών υπολοίπων

Παρακάτω φαίνονται τα διαγράμματα μερικών υπολοίπων για τις επεξηγηματικές μεταβλητές του μοντέλου που επιλέχθηκε. Μόνο το temperature παρουσιάζει μια αρκετά καλά ορισμένη ευθεία, αλλά ακόμα και στις άλλες μεταβλητές δεν φαίνεται να υπάρχει κάποια ξεκάθαρη μη γραμμική σχέση, επομένως δεν θα προβούμε σε κάποιον μη γραμμικό μετασχηματισμό.



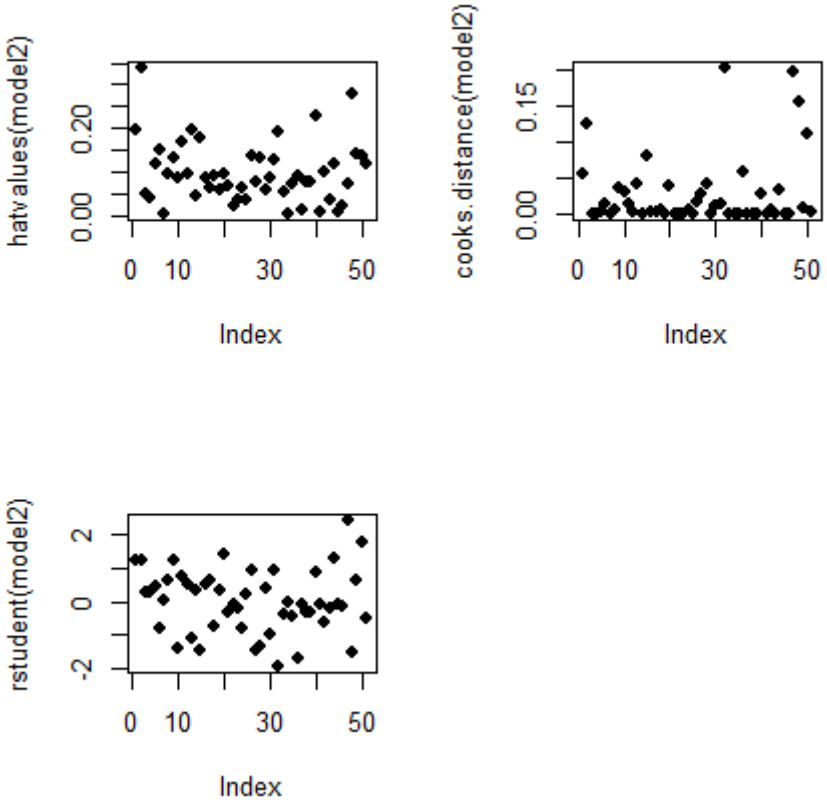
Υπόλοιπα Deviance

Στη συνέχεια βλέπουμε τη γραφική παράσταση της ημι-κανονικής κατανομής για τα υπόλοιπα Deviance με προσομοιωμένα διαστήματα εμπιστοσύνης. Σκοπός είναι το να πάρουμε πληροφορία σχετικά με την καλή προσαρμογή του μοντέλου και όχι για την κανονικότητα των υπολοίπων. Γενικά δεν βλέπουμε τιμές που να ξεφεύγουν πολύ από την ευθεία επομένως δεν προκύπτει κάποιο πρόβλημα κακής προσαρμογής του μοντέλου.



Σημεία επιρροής

Όπως και στην περίπτωση της παλινδρόμησης Poisson θα χρησιμοποιήσουμε τα index plots για τα hii, τις αποστάσεις Cook και τα υπόλοιπα πιθανοφάνειας, αντίστοιχα, για να εντοπίσουμε πιθανά σημεία επιρροής. Από τα διαγράμματα γενικά δεν παρατηρούμε σημεία με ιδιαίτερα μεγάλες τιμές, επομένως το μοντέλο μάλλον δεν επηρεάζεται αρνητικά από άτυπες παρατηρήσεις.



(iii) Διαστήματα εμπιστοσύνης

Από τα διαστήματα εμπιστοσύνης φαίνεται πως μάλλον υπάρχει πρόβλημα με τη μεταβλητή infiltrate, καθώς εμπεριέχεται το 0 στο διάστημα. Παρ' όλα αυτά δεν θα τη αφαιρέσουμε, βασιζόμενοι στους προηγούμενους ελέγχους.

```
                    2.5 %        97.5 %
(Intercept) 19.923208792 171.212118584
age         -0.112744214  -0.007767232
infiltrate  -0.006629639   0.074884222
index        0.151268237   0.662198781
temperature -0.176943980  -0.021940100
> exp(confint.default(model2))
                    2.5 %       97.5 %
(Intercept) 4.493033e+08 2.272366e+74
age         8.933791e-01 9.922629e-01
infiltrate  9.933923e-01 1.077759e+00
index       1.163309e+00 1.939051e+00
temperature 8.378267e-01 9.782988e-01
```

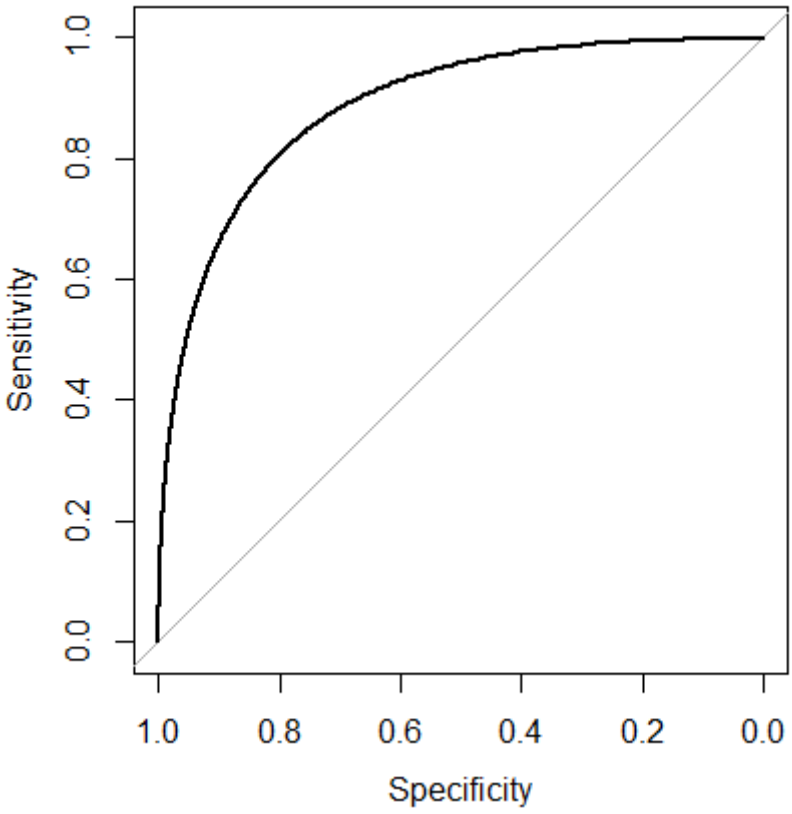
Ερμηνεία συντελεστών

Για τους συντελεστές γνωρίζουμε ότι αν είναι θετικοί τότε το αποτέλεσμα της συνάρτησης exp θα είναι μεγαλύτερο από 1, ενώ θα είναι μικρότερο αν είναι αρνητικοί. Επομένως συμπεραίνουμε τα εξής για τους συντελεστές του μοντέλου μας:

- age, temperature: οι μεταβλητές έχουν αρνητικό συντελεστή, άρα όταν αυξάνεται η ηλικία του ασθενούς ή η θερμοκρασία του πριν τη θεραπεία, μειώνεται η πιθανότητα ανταπόκρισης σε αυτήν.
- infiltrate, index: οι μεταβλητές έχουν θετικό συντελεστή, άρα όταν αυξάνεται το ποσοστό των κυττάρων στο μυελό των οστών ή ο δείκτης κυττάρων λευχαιμίας, αυξάνεται η πιθανότητα ανταπόκρισης στη θεραπεία.

(iv) Καμπύλη ROC

Τέλος, εξετάζουμε την προβλεπτική ικανότητα του μοντέλου χρησιμοποιώντας την καμπύλη ROC και την τιμή AUC (area under the curve). Παρατηρούμε πως η καμπύλη πλησιάζει την πάνω αριστερή γωνία και το εμβαδόν είναι AUC=0.8867 που είναι μια πολύ καλή τιμή, άρα το μοντέλο μας έχει καλή προβλεπτική ικανότητα.



Smoothing: binormal

Area under the curve: 0.8867