Τρίτη σειρά ασκήσεων 2022-2023

Μάθημα: 858 - Στατιστική Μοντελοποίηση

ΔΠΜΣ Επιστήμη Δεδομένων και Μηχανική Μάθηση

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Ονοματεπώνυμο: Απόστολος Μουστάκης

Αριθμός Μητρώου: 03400182

Καθηγήτρια: Χρυσηίς Καρώνη – Ρίτσαρντσον

Περιεχόμενα

Άσκηση 1 - Παλινδρόμηση Poisson 2
Ερώτημα 1 2
Ερώτημα 2 5
Ερώτημα 3 7
Ερώτημα 4 12
Άσκηση 2 – Λογιστική Παλινδρόμηση 15
Ερώτημα 1 15
Ερώτημα 2 20
Ερώτημα 3 24
Ερώτημα 4 26

Άσκηση 1 - Παλινδρόμηση Poisson

Ερώτημα 1

Σε αυτή την άσκηση θα εξετάσουμε κατά πόσο εξαρτάται ο αριθμός y αποζημιώσεων λόγω τροχαίων ατυχημάτων ανά n συμβόλαια από την ηλικία του ασφαλισμένου (agecat: x_1= 0-νέος ή 1-μεγάλος), την κατηγορία ασφαλίστρων (cartype: x_2 = 1,2,3,4) και την περιοχή διαμονής του ασφαλισμένου (district: x_3 = 1 αν μένει στην Αθήνα ή x_3 = 0 αν μένει σε άλλη πόλη). Τα δεδομένα της άσκησης παρατίθενται στον παρακάτω πίνακα:

	cartype	agecat	district	y	n						
1	1	0	0	65	317	17	1	0	1	2	20
2	1	0	0	65	476	18	1	0	1	5	33
3	1	1	0	52	486	19	1	1	1	4	40
4	1	1	0	310	3259	20	1	1	1	36	316
5	2	0	0	98	486	21	2	0	1	7	31
6	2	0	0	159	1004	22	2	0	1	10	81
7	2	1	0	175	1355	23	2	1	1	22	122
8	2	1	0	877	7660	24	2	1	1	102	724
9	3	0	0	41	223	25	3	0	1	5	18
10	3	0	0	117	539	26	3	0	1	7	39
11	3	1	0	137	697	27	3	1	1	16	68
12	3	1	0	477	3442	28	3	1	1	63	344
13	4	0	0	11	40	29	4	0	1	0	3
14	4	0	0	35	148	30	4	0	1	6	16
15	4	1	0	39	214	31	4	1	1	8	25
16	4	1	0	167	1019	32	4	1	1	33	114

Πίνακας 1: Τα δεδομένα της άσκησης

Αφού πρώτα μετατρέψουμε την μεταβλητή x_2 (cartype) σε κατηγορική θα εφαρμόσουμε ένα μοντέλο παλινδρόμησης Poisson στα δεδομένα μας, καθώς αυτό κρίνεται πιο κατάλληλο στην συγκεκριμένη περίπτωση. Το summary του προσαρμοσμένου μοντέλου φαίνεται στον παρακάτω πίνακα:

```
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
```

Πίνακας 2: Summary μοντέλου glm(y ~ cartype + agecat + district + offset(log(n)), data=asfalies, family = 'poisson')

Παρατηρούμε πως εφόσον μετατρέψαμε την μεταβλητή cartype σε κατηγορική στο μοντέλο μας έχουμε τις μεταβλητές cartype2, cartype3 και cartype4 και όχι την μεταβλητή cartype1, καθώς ο συντελεστής της είναι 0 εξ ορισμού. Οι μεταβλητές cartype2, cartype3 και cartype4 δηλώνουν το πόσο θα μεταβληθεί η αποζημίωση y αν κάποιος αλλάξει κατηγορία ασφαλίστρων και από την 1 μεταβεί αντίστοιχα σε κάποια από τις υπόλοιπες τρεις κατηγορίες (2 έως 4).

Για να ελέγξουμε κατά πόσο όλες οι μεταβλητές είναι απαραίτητες στο μοντέλο μας θα χρησιμοποιήσουμε πρώτα το στατιστικό έλεγχο Wald test όπου το z-score για κάθε μεταβλητή υπολογίζεται ως z = $\hat{\beta}_i$ / $se(\hat{\beta}_i)$. Στον πίνακα 2 βλέπουμε πως τα z-score των μεταβλητών cartype3, cartype4, agecat και district αντιστοιχούν σε πιθανότητες μικρότερες του αυστηρότερου ορίου 0.001 και συνεπώς οι μεταβλητές αυτές είναι στατιστικά σημαντικές. Το z-score της μεταβλητής cartype2 αντιστοιχεί σε πιθανότητα 0.001309, η οποία είναι ελάχιστα πάνω από το όριο 0.001, όμως παραμένει πολύ χαμηλή και συνεπώς και αυτή η μεταβλητή θεωρείται στατιστικά σημαντική.

Στην συνέχεια ένας ακόμη έλεγχος που θα χρησιμοποιήσουμε για να αποφανθούμε ποιες μεταβλητές είναι απαραίτητες για το τελικό μοντέλο είναι το κριτήριο AIC. Σύμφωνα με αυτό το κριτήριο, το καλύτερο μοντέλο είναι αυτό που διαθέτει την μικρότερη τιμή AIC. Τα αποτελέσματα με χρήση Backward AIC φαίνονται στο παρακάτω πίνακα:

```
> step(mod, method="backward", test="Chisq")
Start:  AIC=222.15
y ~ 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 ***
- 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 ~ 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
```

Πίνακας 3: Αποτελέσματα με χρήση Backward AIC

Παρατηρούμε πως η αφαίρεση οποιασδήποτε μεταβλητής οδηγεί σε αύξηση του AIC και συνεπώς το αρχικό μοντέλο που περιέχει όλες τις μεταβλητές είναι το βέλτιστο με AIC = 222.1. Αξίζει να σημειωθεί πως και με την χρήση Forward ή Both AIC έχουμε τα ίδια αποτελέσματα.

Τέλος, όσον αφορά τον στατιστικό έλεγχο Deviance εξετάζουμε το πόσο αποτελεσματικά έχει προσαρμοστεί το μοντέλο στα δεδομένα. Αυτός ο έλεγχος φάνηκε προηγουμένως καθώς χρησιμοποίησα το test="Chisq" στο κριτήριο AIC αλλά φαίνεται και παρακάτω με χρήση της εντολής anova(mod, test="Chisq"). Παρατηρούμε πως όλες οι μεταβλητές είναι στατιστικά σημαντικές (αντίστοιχες p-value < 0.001) και συνεπώς το αρχικό μοντέλο περιγράφει καλύτερα τα δεδομένα.

```
> anova(mod, 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
```

Πίνακας 4: Αποτελέσματα στατιστικού ελέγχου Deviance

Το τελικό μοντέλο λοιπόν περιέχει όλες τις μεταβλητές και δίνεται από την σχέση $\hat{y} = \hat{\mu}$ = exp(-1.935 + 0.162cartype2 + 0.395cartype3 + 0.565cartype4 -0.376agecat + 0.217district)

Ερώτημα 2

Η ερμηνεία των συντελεστών $\hat{\beta}$ του τελικού μοντέλου δίνεται ως εξής:

- Αν μεταβούμε από την κατηγορία cartype1 στην κατηγορία cartype2, θα πολλαπλασιαστεί ο αριθμός των αποζημιώσεων κατά exp(0.162) =1.1759, δηλαδή ο αριθμός των αποζημιώσεων θα αυξηθεί κατά περίπου 17.59%.
- Αν μεταβούμε από την κατηγορία cartype1 στην κατηγορία cartype3, θα πολλαπλασιαστεί ο αριθμός των αποζημιώσεων κατά exp(0.395) = 1.4844, δηλαδή ο αριθμός των αποζημιώσεων θα αυξηθεί κατά περίπου 48.44%.
- Αν μεταβούμε από την κατηγορία cartype1 στην κατηγορία cartype4, θα πολλαπλασιαστεί ο αριθμός των αποζημιώσεων κατά exp(0.565) =1.7594, δηλαδή ο αριθμός των αποζημιώσεων θα αυξηθεί κατά περίπου 75.94%.
- Αν η συμμεταβλητή agecat αυξηθεί κατά μία μονάδα, δηλαδή όταν πρόκειται για ένα άτομο μεγάλης ηλικίας, θα πολλαπλασιαστεί ο αριθμός των αποζημιώσεων κατά exp(-0.376) = 0.6866, δηλαδή ο αριθμός των αποζημιώσεων θα μειωθεί κατά περίπου 31.34%.
- Αν η συμμεταβλητή district αυξηθεί κατά μία μονάδα, δηλαδή όταν πρόκειται για ένα άτομο που μένει στην Αθήνα, θα πολλαπλασιαστεί ο αριθμός των αποζημιώσεων κατά exp(0.217) = 1.2423, δηλαδή ο αριθμός των αποζημιώσεων θα αυξηθεί κατά περίπου 24.23%.

Η φράση «κατά περίπου» χρησιμοποιείται καθώς αναφερόμαστε στην μέση τιμή. Υπολογίζοντας διαστήματα εμπιστοσύνης για τους συντελεστές $\hat{\beta}$ μπορούμε να είμαστε πιο ακριβείς στο πως θα μεταβληθεί ο αριθμός των αποζημιώσεων σε κάθε περίπτωση. Τα 95% διαστήματα εμπιστοσύνης για τους συντελεστές $\hat{\beta}$ παρουσιάζονται παρακάτω:

```
> confint.default(mod)
                  2.5 %     97.5 %
(Intercept) -2.04350208 -1.8269440
cartype2     0.06329746  0.2611664
cartype3     0.28772397  0.5029705
cartype4     0.42400923  0.7068487
agecat      -0.46352606 -0.2890309
district     0.10189607  0.3313250
> (exp(confint.default(mod))-1)*100
                  2.5 %    97.5 %
(Intercept) -87.042586 -83.90955
cartype2      6.534369  29.84438
cartype3     33.338920  65.36260
cartype4     52.807570 102.75915
agecat      -37.093838 -25.10110
district     10.726839  39.28124
```

Πίνακας 5: 95% διαστήματα εμπιστοσύνης για τους συντελεστές $\hat{\beta}$

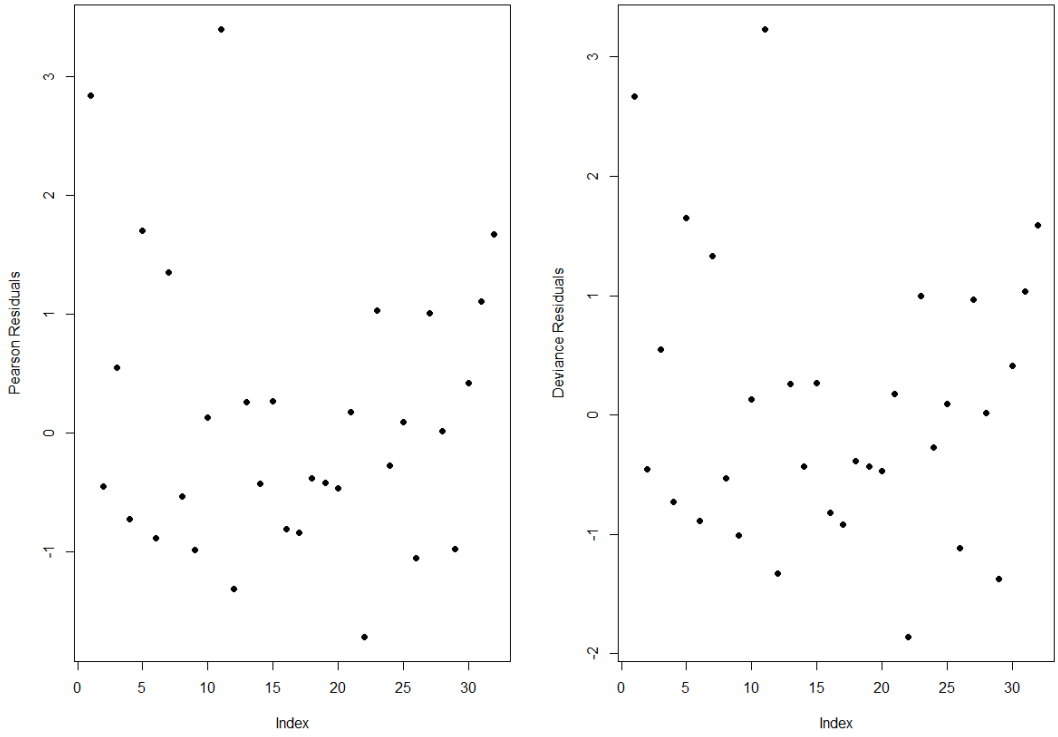
Η εντολή (exp(confint.default(mod))-1)*100 μετατρέπει κατάλληλα τα διαστήματα εμπιστοσύνης ώστε να εκφράζουν τις μεταβολές στις αποζημιώσεις όπως περιέγραψα παραπάνω. Με αυτό τον τρόπο έχουμε τις ακόλουθες ερμηνείες:

- Ένας ασφαλισμένος που επιλέγει την κατηγορία ασφαλίστρων cartype2 έναντι της cartype1 λαμβάνει υψηλότερες αποζημιώσεις κατά (6.53%,29.84%) με μέση τιμή 17.59%.

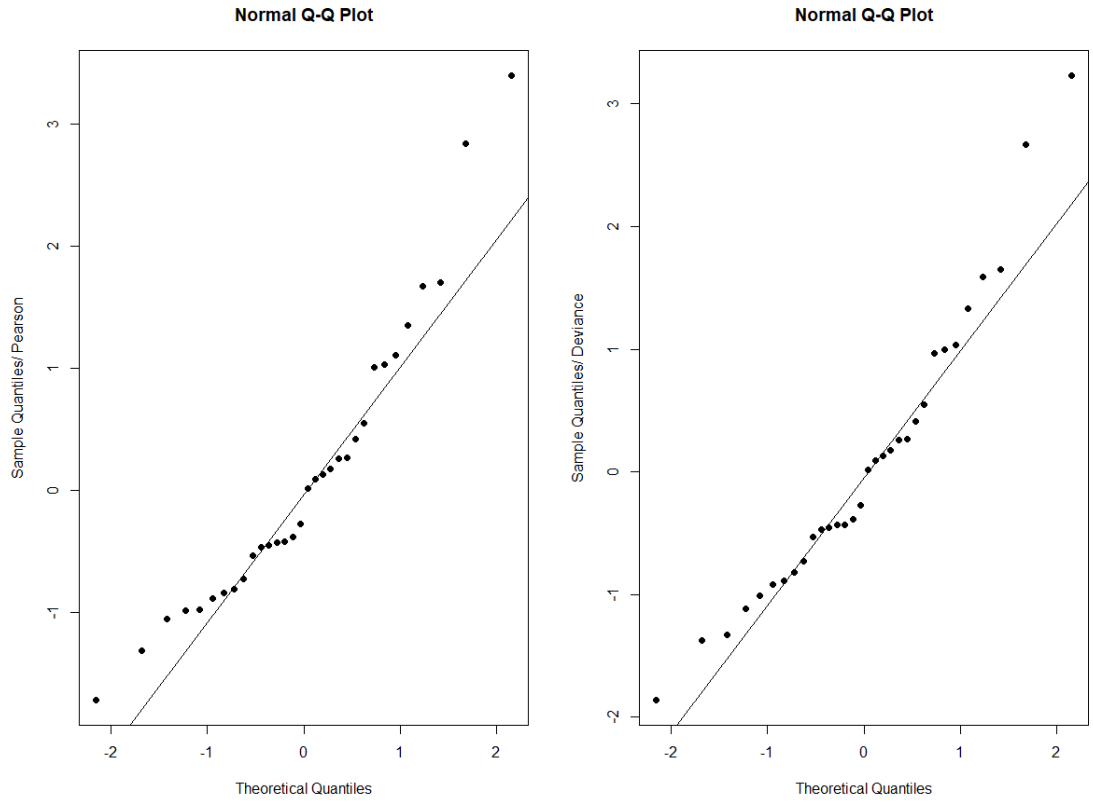
- Ένας ασφαλισμένος που επιλέγει την κατηγορία ασφαλίστρων cartype3 έναντι της cartype1 λαμβάνει υψηλότερες αποζημιώσεις κατά (33.39%, 65.36%) με μέση τιμή 48.44%.
- Ένας ασφαλισμένος που επιλέγει την κατηγορία ασφαλίστρων cartype4 έναντι της cartype1 λαμβάνει υψηλότερες αποζημιώσεις κατά (52.81%, 102.75%) με μέση τιμή 75.94%.
- Ένας ασφαλισμένος μεγάλης ηλικίας λαμβάνει χαμηλότερες αποζημιώσεις από έναν ασφαλισμένο κατά (-37.09%, -25.1%) με μέση τιμή -31.34%.
- Ένας ασφαλισμένος που μένει στην Αθήνα λαμβάνει υψηλότερες αποζημιώσεις από ένα ασφαλισμένο που μένει σε κάποια άλλη περιοχή κατά (10.73%,39.28%) με μέση τιμή 24.23%.

Ερώτημα 3

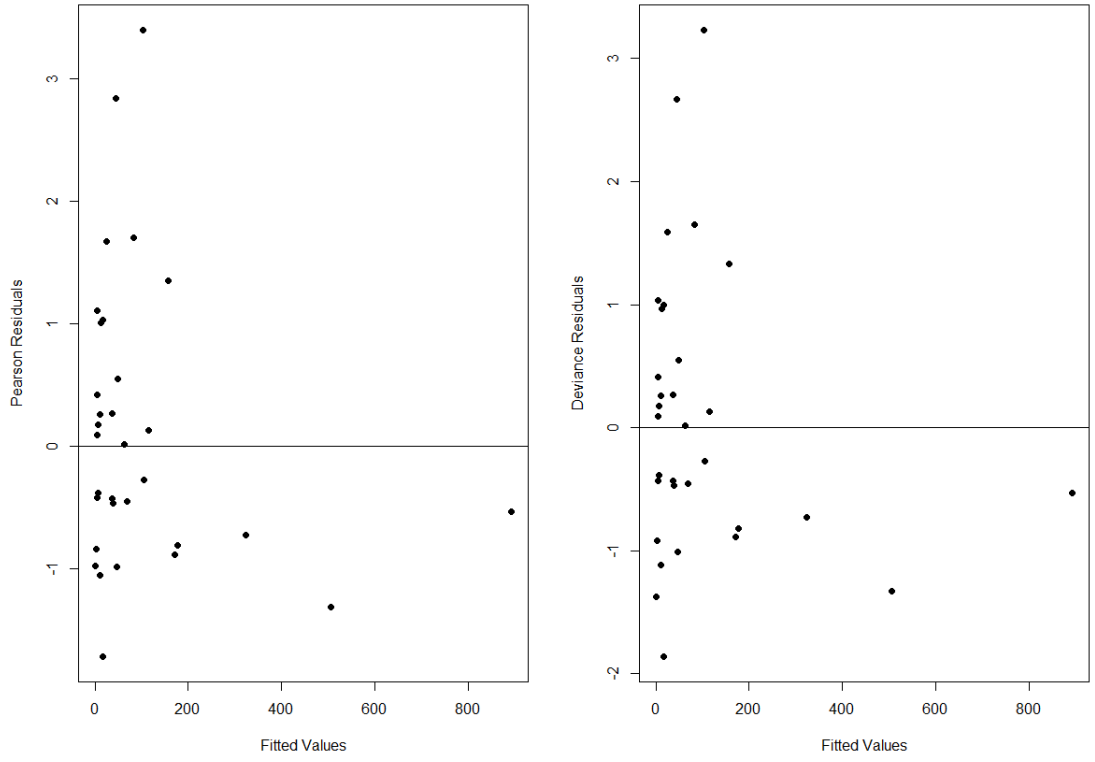
Θα ξεκινήσουμε με την δημιουργία των γραφικών παραστάσεων που αφορούν τα υπόλοιπα Pearson και Deviance, οι οποίες είναι οι εξής: Index plots για υπόλοιπα Pearson και Deviance, Q-Q plots για υπόλοιπα Pearson και Deviance, fitted values για τα υπόλοιπα Pearson και Deviance, Υπόλοιπα Pearson ανά μεταβλητή και Υπόλοιπα Deviance ανά μεταβλητή. Αυτές οι γραφικές παραστάσεις φαίνονται παρακάτω και ακολουθεί ο σχολιασμός.



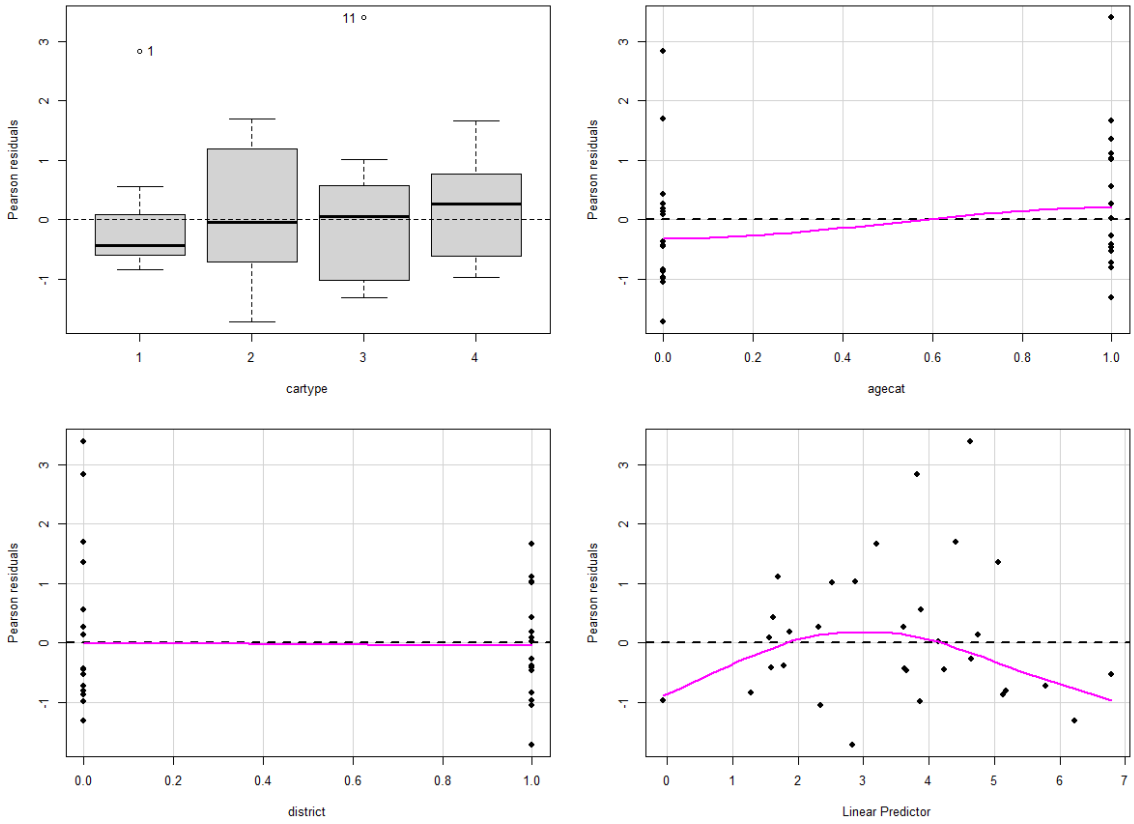
Πίνακας 6: Index plots για τα Pearson και Deviance Residuals



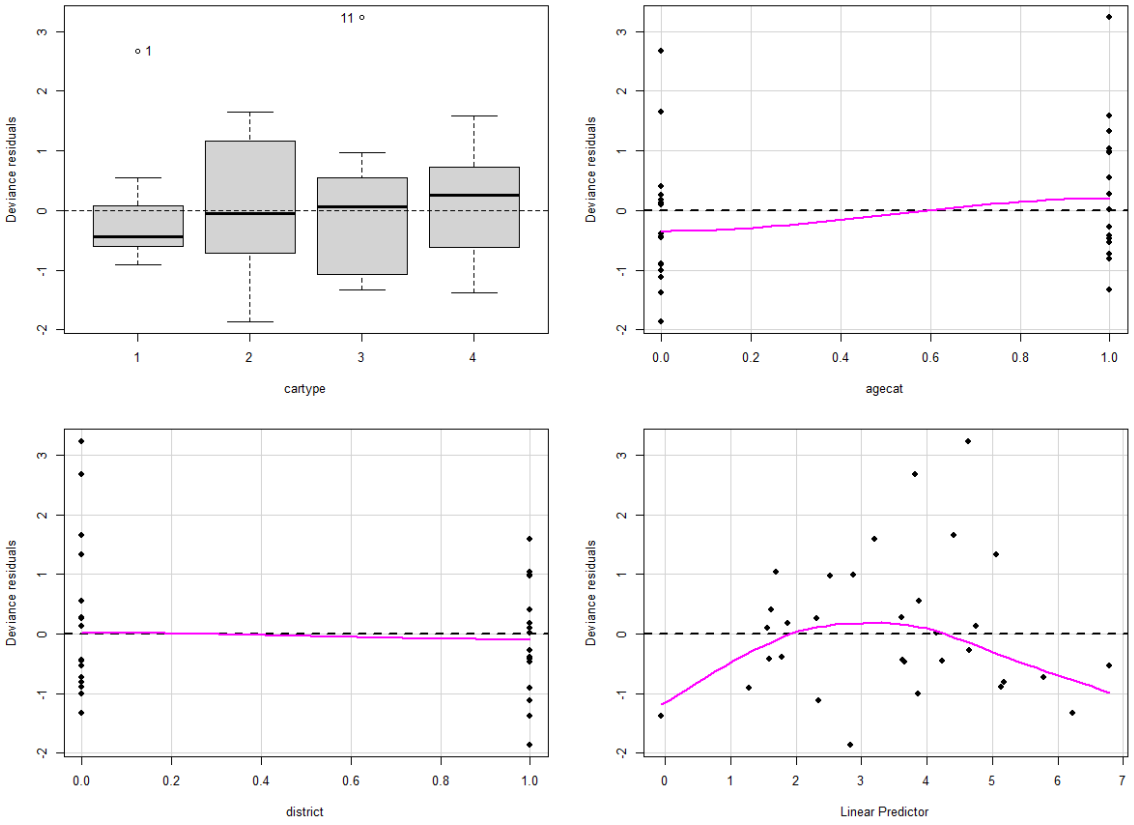
Πίνακας 7: Normal Q-Q plots για τα Pearson και Deviance Residuals



Πίνακας 8: Fitted Values VS Pearson και Deviance Residuals



Πίνακας 9: Διαγράμματα Pearson Residuals ανά μεταβλητή



Πίνακας 10: Διαγράμματα Deviance Residuals ανά μεταβλητή

Ξεκινώντας την ανάλυση παρατηρούμε πως όλα τα διαγράμματα που αφορούν τα υπόλοιπα Pearson είναι σχεδόν ίδια με τα διαγράμματα που αφορούν τα υπόλοιπα Deviance, με ελάχιστες και όχι αισθητές διαφορές. Τα σχόλια λοιπόν που θα γίνουν αφορούν και τις δύο περιπτώσεις.

Όσον αφορά τα index plots αλλά και τα διαγράμματα Pearson και Deviance residuals ανά μεταβλητή οι παρατηρήσεις 1 και 11 εμφανίζονται ως ενδεχόμενα άτυπα σημεία ή σημεία επιρροής. Στα index plots οι δύο αυτές παρατηρήσεις ξεφεύγουν με τιμές κοντά στο 3, ενώ αυτό φαίνεται ξεκάθαρα και στα δύο διαγράμματα που αφορούν την μεταβλητή cartype (Pearson Residuals vs Cartype και Deviance Residuals vs Cartype). Επιπλέον, για τα index plots κρίνεται σκόπιμο να αναφερθεί πως τα υπόλοιπα φαίνεται να κατανέμονται τυχαία γύρω από το 0, χωρίς να κρύβουν κάποια συστηματικότητα.

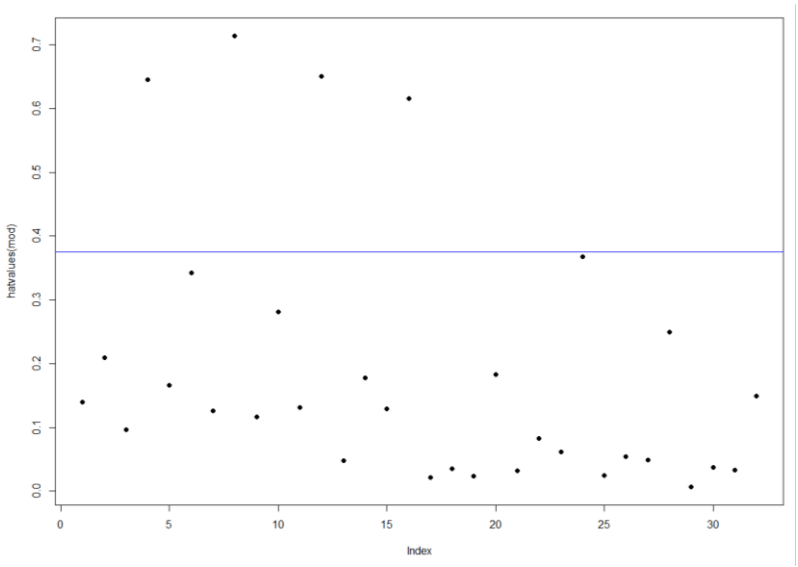
Όσον αφορά τα normal Q-Q plots ορισμένες παρατηρήσεις ξεφεύγουν αλλά γενικά φαίνεται να υπάρχει μια γραμμική τάση, η οποία φανερώνει καλή προσαρμογή του μοντέλου στα δεδομένα.

Τέλος, όσον αφορά τα διαγράμματα Fitted Values vs Pearson και Deviance Residuals παρόμοια εντοπίζουμε τις παρατηρήσεις 1 και 11 ως πιθανά άτυπα σημεία ή σημεία επιρροής ως προς τα υπόλοιπα, καθώς έχουν τιμές κοντά στο 3, ενώ όλες οι υπόλοιπες παρατηρήσεις κινούνται στο εύρος (-2,2). Ως προς τα fitted values, οι παρατηρήσεις 4, 12 και 8 ξεφεύγουν αρκετά καθώς έχουν πολύ υψηλό αριθμό αποζημιώσεων.

Θα συνεχίσουμε σχεδιάζοντας το διάγραμμα για τα h_{ii}, τα οποία είναι τα διαγώνια στοιχεία του hat matrix. Παρακάτω παρατίθεται ο πίνακας με τις ακριβείς μετρήσεις των h_{ii} και το διάγραμμα των h_{ii}.

1	2	3	4	5	6	7	8	9
0.139699568	0.209769698	0.096198142	0.645081779	0.165873592	0.342668903	0.126274179	0.713845176	0.116202166
10	11	12	13	14	15	16	17	18
0.280865325	0.131760643	0.650674510	0.047908509	0.177261482	0.129243200	0.615415050	0.021214389	0.035003741
19	20	21	22	23	24	25	26	27
0.023191226	0.183210687	0.031816274	0.083132844	0.061922896	0.367476860	0.025193908	0.054586801	0.049225274
28	29	30	31	32				
0.249021976	0.007038849	0.037540529	0.032676587	0.149005237				

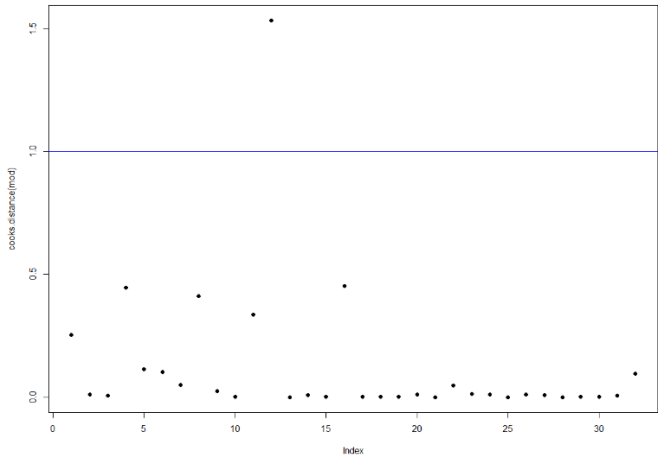
Πίνακας 11: h_{ii} για τις 32 παρατηρήσεις



Διάγραμμα 1: Index plot για τα h_{ii} για τις 32 παρατηρήσεις

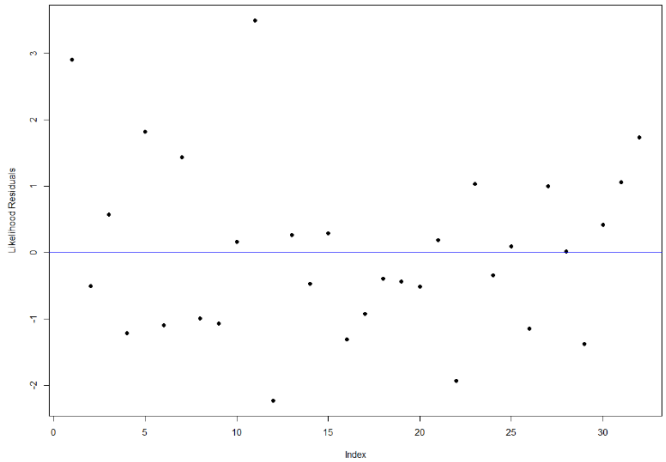
Ο υπολογισμός των h_{ii} μπορεί να χρησιμοποιηθεί ως διαγνωστικός έλεγχος για την πιθανή παρουσία άτυπων σημείων ή σημείων επιρροής. Σύμφωνα με το κριτήριο h_{ii} μία παρατήρηση i θεωρείται σημείο επιρροής αν ισχύει h_{ii} > 2p/n, όπου p = k +1 (k ανεξάρτητες μεταβλητές) και n το πλήθος των παρατηρήσεων. Στη περίπτωση μας έχουμε p = 6 και n = 32 και άρα πρέπει h_{ii} > 0.375 (Η μπλε γραμμή που σχεδίασα στο διάγραμμα). Κοιτάζοντας τον πίνακα 11 βλέπουμε πως οι παρατηρήσεις 4,8,12 και 16 ξεπερνούν κατά πολύ το όριο και συνεπώς μπορούν να θεωρηθούν σημεία επιρροής, ενώ οι παρατηρήσεις 6 και 24 βρίσκονται πολύ κοντά στο όριο αλλά παραμένουν κάτω από αυτό.

Ένα ακόμη κριτήριο για τον εντοπισμό άτυπων σημείων ή σημείων επιρροής αποτελεί η απόσταση Cook. Σύμφωνα με αυτό το κριτήριο αν D_i >> 1 η παρατήρηση i θεωρείται σημείο επιρροής. Παρακάτω παρατίθεται το αντίστοιχο διάγραμμα στο οποίο φαίνεται πως μόνο η παρατήρηση 12 θεωρείται σημείο επιρροής καθώς ξεπερνάει αρκετά το 1.



Διάγραμμα 2: Υπολογισμός απόστασης Cook για κάθε παρατήρηση i

Τέλος, παρακάτω παρατίθεται το διάγραμμα των υπολοίπων πιθανοφάνειας, το οποίο εκφράζει την μεταβολή της deviance αν κάθε φορά παραλείπεται η i-οστή παρατήρηση. Παρατηρούμε πως το διάγραμμα αυτό δεν έχει κάποια διαφορά με τα index plots για τα Pearson και Deviance Residuals, τα οποία σχεδιάστηκαν παραπάνω, και συνεπώς εξάγουμε τα ίδια συμπεράσματα.



Διάγραμμα 3: Index plot για υπόλοιπα πιθανοφάνειας

Ερώτημα 4

Για να δούμε το πόσο καλή προσαρμογή έχει το μοντέλο μας στα δεδομένα σε σχέση με το κορεσμένο, θα εφαρμόσουμε τον παρακάτω έλεγχο σύγκρισης deviance:

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

Πίνακας 12: Σύγκριση μοντέλου με το κορεσμένο

Ο έλεγχος αυτός μας δείχνει πως δεν έχουμε καθόλου καλή προσαρμογή του μοντέλου στα δεδομένα. Ένας τρόπος για να βελτιώσουμε αυτή την προσαρμογή είναι να εισάγουμε αλληλεπιδράσεις μεταξύ των μεταβλητών. Αρχικά εξετάζουμε την αλληλεπίδραση agecat*cartype. Το summary το νέου μοντέλου φαίνεται παρακάτω:

```
Call:
glm(formula = y ~ cartype + agecat + district + agecat * cartype +
    offset(log(n)), family = "poisson", data = asfalies)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8233  -0.7028  -0.1240   0.8065   3.0988

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.32786    0.17802  -7.459 8.73e-14 ***
cartype2        -0.09045    0.21711  -0.417  0.67695
cartype3         0.05361    0.23803   0.225  0.82180
cartype4         0.25981    0.33566   0.774  0.43892
agecat          -0.50774    0.09894  -5.132 2.87e-07 ***
district         0.21692    0.05853   3.706  0.00021 ***
cartype2:agecat  0.14338    0.11953   1.200  0.23032
cartype3:agecat  0.19298    0.13081   1.475  0.14014
cartype4:agecat  0.17233    0.18184   0.948  0.34330
---
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:  39.434  on 23  degrees of freedom
AIC: 225.79

Number of Fisher Scoring iterations: 4
```

Πίνακας 13: Summary μοντέλου glm(y ~ cartype + agecat + district + agecat*cartype + offset(log(n)), data=asfalies, family = 'poisson')

Παρατηρούμε πως τα z-scores των μεταβλητών cartype2, cartype3, cartype4 και των αλληλεπιδράσεων αντιστοιχούν σε p-values > 0.05 και συνεπώς είναι στατιστικά μη σημαντικά. Επιπλέον, το AIC αυξήθηκε από 222.15 σε 225.79. Συνεπώς, η αλληλεπίδραση agecat*cartype δεν βελτιώνει την προσαρμογή του μοντέλου.

Στην συνέχεια θα εξετάζουμε την αλληλεπίδραση district*cartype. Το summary το νέου μοντέλου φαίνεται παρακάτω:

```
Call:
glm(formula = y ~ cartype + agecat + district + district * cartype +
    offset(log(n)), family = "poisson", data = asfalies)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7226  -0.6658   0.0260   0.4098   3.2367

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.54718    0.09056 -17.084  < 2e-16 ***
cartype2          0.15317    0.05291   2.895   0.0038 **
cartype3          0.38172    0.05770   6.616 3.69e-11 ***
cartype4          0.51016    0.07750   6.583 4.62e-11 ***
agecat           -0.37562    0.04452  -8.438  < 2e-16 ***
district          0.07745    0.15269   0.507   0.6120
cartype2:district 0.09978    0.17654   0.565   0.5719
cartype3:district 0.14557    0.18866   0.772   0.4404
cartype4:district 0.44498    0.22036   2.019   0.0434 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 207.83  on 31  degrees of freedom
Residual deviance:  37.27  on 23  degrees of freedom
AIC: 223.63

Number of Fisher Scoring iterations: 4
```

Πίνακας 14: Summary μοντέλου glm(y ~ cartype + agecat + district + district*cartype + offset(log(n)), data=asfalies, family = 'poisson')

Παρατηρούμε πως τα z-scores των μεταβλητών district και των αλληλεπιδράσεων cartype2:district και cartype3:district αντιστοιχούν σε p-values > 0.05 και συνεπώς είναι στατιστικά μη σημαντικά. Το z-score της αλληλεπίδρασης cartype4:district βγαίνει οριακά στατιστικά σημαντικό (p-value = 0.0434), όμως αυτό δεν είναι αρκετό. Επιπλέον, το AIC αυξήθηκε από 222.15 σε 223.63. Συνεπώς, η αλληλεπίδραση district*cartype δεν βελτιώνει την προσαρμογή του μοντέλου.

Τέλος, θα εξετάζουμε την αλληλεπίδραση agecat*district. Το summary το νέου μοντέλου φαίνεται παρακάτω:

```
Call:
glm(formula = y ~ cartype + agecat + district + agecat * district +
    offset(log(n)), family = "poisson", data = asfalies)

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

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -1.51178    0.09245 -16.353  < 2e-16 ***
cartype2         0.16279    0.05048   3.225  0.00126 **
cartype3         0.39565    0.05491   7.205 5.79e-13 ***
cartype4         0.56639    0.07216   7.849 4.18e-15 ***
agecat          -0.40282    0.04629  -8.702  < 2e-16 ***
district        -0.38890    0.32555  -1.195  0.23225
agecat:district  0.32763    0.17167   1.908  0.05633 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

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

Number of Fisher Scoring iterations: 4
```

Πίνακας 15: Summary μοντέλου glm(y ~ cartype + agecat + district + agecat*district + offset(log(n)), data=asfalies, family = 'poisson')

Παρατηρούμε πως τα z-scores των μεταβλητών district και της αλληλεπίδρασης agecat:district αντιστοιχούν σε p-values > 0.05 και συνεπώς είναι στατιστικά μη σημαντικά. Το AIC σε αυτή την περίπτωση μειώνεται λίγο, από 222.15 σε 220.25, αλλά η μείωση είναι πολύ μικρή. Συνεπώς, η αλληλεπίδραση agecat*district δεν βελτιώνει την προσαρμογή του μοντέλου.

Παρακάτω παρουσιάζεται ο έλεγχος σύγκρισης deviance των μοντέλων που εξετάστηκαν με τις αλληλεπιδράσεις σε σχέση με το κορεσμένο μοντέλο. Και με αυτόν τον τρόπο συμπεραίνουμε πως η προσαρμογή του μοντέλου συνεχίζει να μην είναι καλή σε καμία από τις περιπτώσεις.

```
> 1-pchisq(mod1$deviance, mod$df.residual)
[1] 0.04429191
> 1-pchisq(mod2$deviance, mod$df.residual)
[1] 0.07066585
> 1-pchisq(mod3$deviance, mod$df.residual)
[1] 0.06200743
```

Πίνακας 16: Σύγκριση των μοντέλων που περιέχουν αλληλεπιδράσεις με το κορεσμένο

Αξίζει να αναφερθεί πως υπάρχουν διάφοροι τρόποι για να γίνει καλύτερη η προσαρμογή του μοντέλου στα δεδομένα. Ενδεικτικοί τρόποι αποτελούν η εισαγωγή παραπάνω δεδομένων, η εισαγωγή νέων μεταβλητών στο μοντέλο (ανεξάρτητων των μεταβλητών που υπάρχουν ήδη) και η αφαίρεση δεδομένων - παρατηρήσεων που αποτελούν σημεία επιρροής ή άτυπα σημεία. Ενώ γενικότερα δεν αφαιρούμε δεδομένα χωρίς να έχουμε παραπάνω πληροφορίες θα το δοκιμάσω στην συγκεκριμένη άσκηση σαν παράδειγμα. Έτσι λοιπόν θα αφαιρέσω τις παρατηρήσεις 1, 4, 8, 11 και 12 που βρέθηκαν ότι αποτελούν πιθανά άτυπα σημεία ή σημεία επιρροής από ελέγχους που έγιναν στο τρίτο ερώτημα. Παρακάτω παραθέτω τον έλεγχο deviance με το κορεσμένο μοντέλο και το summary του νέου μοντέλου, όπου με μία πρώτη εικόνα η προσαρμογή του στα νέα δεδομένα είναι πάρα πολύ καλή.

```
> 1-pchisq(mod4$deviance, mod$df.residual)
[1] 0.8412714

Call:
glm(formula = y ~ cartype + agecat + district + offset(log(n)),
    family = "poisson", data = new_model)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6525  -0.6926   0.1359   0.4748   1.7240

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.00931    0.08522 -23.578  < 2e-16 ***
cartype2     0.22872    0.08862   2.581  0.00986 **
cartype3     0.43877    0.10232   4.288 1.80e-05 ***
cartype4     0.56755    0.10104   5.617 1.94e-08 ***
agecat      -0.28249    0.06682  -4.228 2.36e-05 ***
district     0.16948    0.07281   2.328  0.01992 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 74.945  on 26  degrees of freedom
Residual deviance: 18.884  on 21  degrees of freedom
AIC: 162.28

Number of Fisher Scoring iterations: 4
```

Πίνακας 17: Έλεγχος με το κορεσμένο μοντέλο και summary νέου ενδεικτικού μοντέλο

Άσκηση 2 – Λογιστική Παλινδρόμηση

Ερώτημα 1

Σε αυτή την άσκηση θα εξετάσουμε κατά πόσο εξαρτάται η ανταπόκριση (response) στην θεραπεία για λευχαιμία (ναι=1, όχι=0) από την ηλικία του ασθενή (age), το ποσοστό επίστρωσης βλαστοκυττάρων (smear), το ποσοστό κυττάρων στο μυελό των οστών (infiltrate), τον δείκτη κυττάρων λευχαιμίας (index), τα βλαστοκύτταρα (blasts) και την υψηλότερη θερμοκρασία πριν τη θεραπεία (temperature). Τα δεδομένα της άσκησης φαίνονται στον παρακάτω πίνακα:

	age	smear	infiltrate	index	blasts	temperature	response
1	20	78	39	7	0.6	990	1
2	25	64	61	16	35.0	1030	1
3	26	61	55	12	7.5	982	1
4	26	64	64	16	21.0	1000	1
5	27	95	95	6	7.5	980	1
6	27	80	64	8	0.6	1010	0
7	28	88	88	20	4.8	986	1
8	28	70	70	14	10.0	1010	1
9	31	72	72	5	2.3	988	1
10	33	58	58	7	5.7	986	0
11	33	92	92	5	2.6	980	1
12	33	42	38	12	2.5	984	1
13	34	26	26	7	7.0	982	0
14	36	55	55	14	4.5	986	1
15	37	71	71	15	4.4	1020	0
16	40	91	91	9	35.0	986	1
17	40	52	49	12	2.1	988	1
18	43	74	63	4	0.1	986	0
19	45	78	47	14	4.2	980	1
20	45	60	36	10	0.6	992	1
21	45	82	32	10	28.1	1016	0
22	45	79	79	4	1.1	1030	0
23	47	56	28	2	0.9	990	0
24	48	60	54	10	2.2	1002	0
25	50	83	66	19	11.6	996	1
26	50	36	32	14	4.5	992	1
27	51	88	70	8	0.5	982	0
28	52	87	87	7	10.3	986	0
29	53	75	68	13	2.3	980	1
30	53	65	65	6	2.3	982	0
31	56	97	92	10	16.0	992	1
32	57	87	83	19	21.6	1020	0
33	59	45	45	8	1.1	999	0
34	59	36	34	5	0.0	1038	0
35	60	39	33	7	0.9	988	0
36	60	76	53	12	0.4	982	0
37	61	46	37	4	1.4	1006	0
38	61	39	8	8	0.3	990	0
39	61	90	90	1	9.9	990	0
40	62	84	84	19	11.5	1020	1
41	63	42	27	5	0.3	1014	0
42	65	75	75	10	20.0	1004	0
43	71	44	22	6	0.3	990	0
44	71	63	63	11	10.0	986	1
45	73	33	33	4	0.5	1010	0
46	73	93	84	6	38.0	1020	0
47	74	58	58	10	2.4	1002	1
48	74	32	30	16	6.7	988	0
49	75	60	60	17	8.2	990	1
50	77	69	69	9	1.5	986	1
51	80	73	73	7	1.5	986	0

Πίνακας 18: Τα δεδομένα της άσκησης

θα εφαρμόσουμε ένα μοντέλο Λογιστικής Παλινδρόμησης στα δεδομένα μας, καθώς αυτό κρίνεται πιο κατάλληλο στην συγκεκριμένη περίπτωση. Το summary του προσαρμοσμένου μοντέλου φαίνεται στον παρακάτω πίνακα:

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

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
```

Πίνακας 19: Summary μοντέλου glm(response ~ age + smear + infiltrate + index + blasts + temperature, family='binomial')

Για να ελέγξουμε κατά πόσο όλες οι μεταβλητές είναι απαραίτητες στο μοντέλο μας θα χρησιμοποιήσουμε πρώτα το στατιστικό έλεγχο Wald test όπου το z-score για κάθε μεταβλητή υπολογίζεται ως z = $\hat{\beta}_i$ / $se(\hat{\beta}_i)$, όπως περιγράψαμε και στην πρώτη άσκηση. Παρατηρούμε πως καμία από τις μεταβλητές δεν διαθέτει z-score με πιθανότητα < 0.001, το οποίο είναι το αυστηρότερο όριο. Οι μεταβλητές που έχουν z-score με πιθανότητα < 0.05 και συνεπώς μπορούν να θεωρηθούν στατιστικά σημαντικές είναι οι age, index και temperature. Οι υπόλοιπες μεταβλητές έχουν z-scores που αντιστοιχούν σε πολύ υψηλές πιθανότητες. Παρόμοια είναι τα αποτελέσματα και με τον έλεγχο Deviance, ο οποίος πραγματοποιείται με την εντολή anova(mod1,test="Chisq"), του οποίου τα αποτελέσματα φαίνονται στον παρακάτω πίνακα.

```
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
```

Πίνακας 20: Έλεγχος Deviance για το μοντέλο glm(response ~ age + smear + infiltrate + index + blasts + temperature, family='binomial')

Και σε αυτή την περίπτωση μόνο οι μεταβλητές age, index και temperature θεωρούνται στατιστικά σημαντικές. Εφόσον όλες οι μεταβλητές δεν είναι στατιστικά σημαντικές θα χρησιμοποιήσουμε το κριτήριο AIC ώστε να αποφανθούμε ποιες μεταβλητές είναι απαραίτητες για το τελικό μοντέλο. Σύμφωνα με αυτό το κριτήριο, το καλύτερο μοντέλο είναι αυτό που διαθέτει την μικρότερη τιμή AIC. Τα αποτελέσματα με χρήση Backward AIC φαίνονται στο παρακάτω πίνακα:

```
> step(mod1, method="backward", test="Chisq")
Start:  AIC=54.06
response ~ age + smear + infiltrate + index + blasts + temperature

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

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

             Df Deviance    AIC     LRT  Pr(>Chi)
- blasts      1   40.136 50.136  0.0626  0.802420
<none>            40.074 52.074
- infiltrate  1   42.615 52.615  2.5412  0.110913
- age         1   46.216 56.216  6.1421  0.013200 *
- temperature 1   48.346 58.346  8.2727  0.004025 **
- index       1   56.308 66.308 16.2346 5.596e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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

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

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

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
```

Πίνακας 21: Αποτελέσματα με χρήση Backward AIC

Παρατηρούμε πως με βάση το κριτήριο AIC το τελικό μοντέλο περιέχει τις μεταβλητές age, temperature, index και infiltrate με AIC = 50.14. Η μεταβλητή infiltrate προστίθεται στο τελευταίο βήμα, παρόλο που θεωρείται στατιστικά μη σημαντική από τον έλεγχο deviance, καθώς οδηγεί σε μικρότερο AIC. Αξίζει να σημειωθεί πως και με την χρήση Forward ή Both AIC έχουμε τα ίδια αποτελέσματα.

Επιπλέον βλέποντας το summary του μοντέλου που περιέχει την μεταβλητή infiltrate (response ~ age + infiltrate + index + temperature) παρατηρούμε πως με βάση τον

έλεγχο Wald το z-score της μεταβλητής infiltrate αντιστοιχεί σε p-value = 0.10077 > 0.05 και συνεπώς η μεταβλητή αυτή είναι στατιστικά μη σημαντική.

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

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
```

Πίνακας 22: Summary μοντέλου: response ~ age + infiltrate + index + temperature

Ο έλεγχος Deviance που είδαμε προηγουμένως μπορεί να πραγματοποιηθεί επίσης μεταξύ των δύο εμφωλευμένων μοντέλων M1: response ~ age + index + temperature (mod2) και M2: response ~ age + infiltrate + index + temperature (mod3) με την εντολή anova(mod3,mod2, test="Chisq"), τα αποτελέσματα του οποίου φαίνονται παρακάτω:

```
> anova(mod3,mod2, test="Chisq")
Analysis of Deviance Table

Model 1: response ~ age + index + temperature
Model 2: response ~ age + infiltrate + index + temperature
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        47     43.265
2        46     40.136  1   3.1292   0.0769 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Πίνακας 23: Αποτελέσματα του ελέγχου Deviance μεταξύ των μοντέλων M1 και M2

Τα αποτελέσματα προφανώς συμφωνούν με προηγουμένως: Η Deviance μειώνεται από 43.265 σε 40.136 αλλά η μείωση είναι στατιστικά μη σημαντική (0.0769 > 0.05). Συνεπώς με βάση τους ελέγχους Wald και Deviance επιλέγουμε να διώξουμε την μεταβλητή infiltrate, καθώς βγαίνει στατιστικά μη σημαντική. Η συμπερίληψη της μεταβλητής αυτής μπορεί να οδηγεί σε μικρότερο AIC, αλλά η διαφορά είναι πολύ μικρή (από 51.265 σε 50.14). Το τελικό μοντέλο, λοιπόν, είναι το M1: response ~ age + index + temperature, το summary του οποίου φαίνεται παρακάτω:

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

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

(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
```

Πίνακας 24: Summary τελικού μοντέλου: response ~ age + index + temperature

Το τελικό μοντέλο δίνεται από την σχέση:

$$ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = 87.388 - 0.059\ age + 0.385\ index - 0.089\ temperature$$

ή ισοδύναμα:

$$\frac{\hat{p}}{1-\hat{p}} = exp(87.388 - 0.059\ age + 0.385\ index - 0.089\ temperature)$$

όπου η πιθανότητα επιτυχίας (θετική ανταπόκριση στην θεραπεία για λευχαιμία) δίνεται από την σχέση:

$$\hat{p} = \frac{exp(87.388 - 0.059\ age + 0.385\ index - 0.089\ temperature)}{1 + exp(87.388 - 0.059\ age + 0.385\ index - 0.089\ temperature)}$$

Τέλος, για να δούμε το πόσο καλή προσαρμογή έχει το τελικό μας μοντέλο στα δεδομένα σε σχέση με το κορεσμένο, θα εφαρμόσουμε τον παρακάτω έλεγχο σύγκρισης deviance:

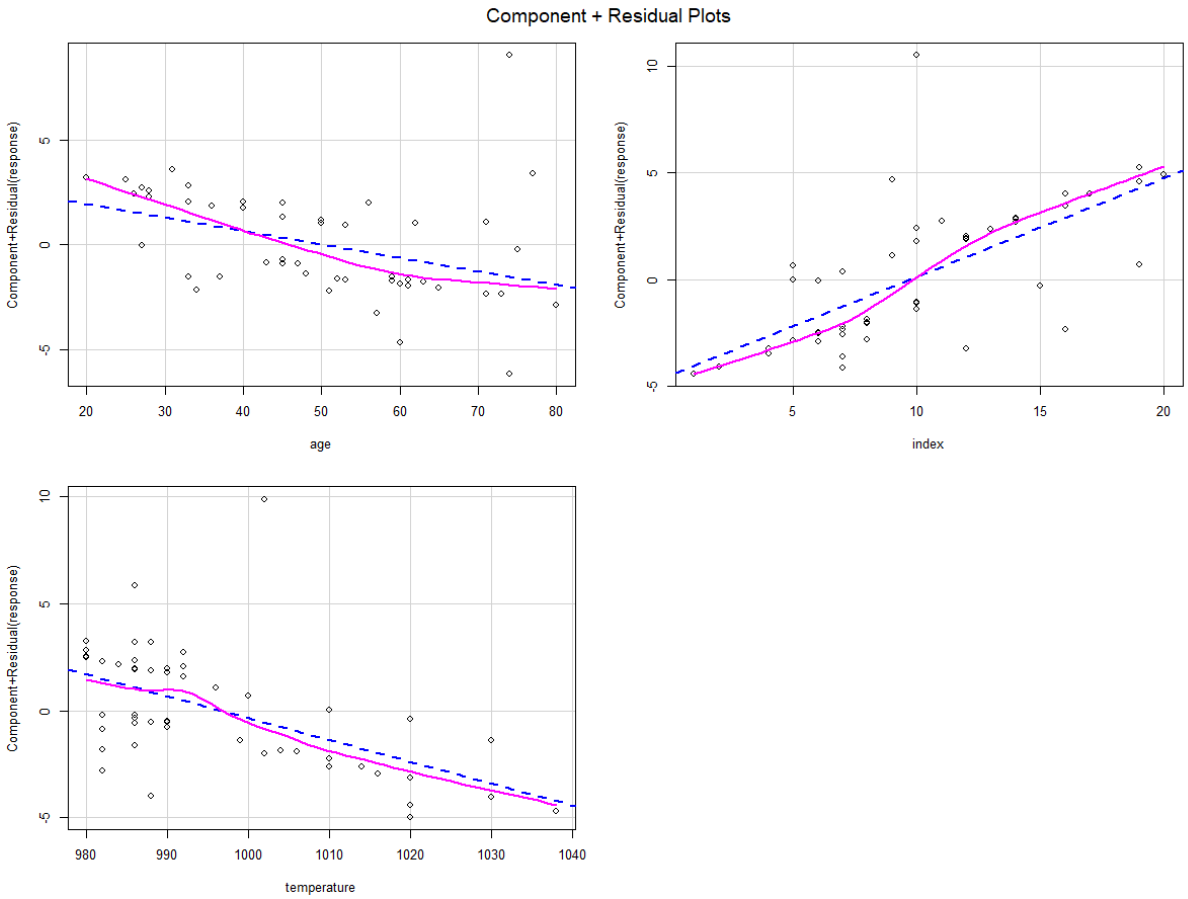
```
> 1-pchisq(mod3$deviance, mod3$df.residual)
[1] 0.6280164
```

Πίνακας 25: Σύγκριση τελικού μοντέλου με το κορεσμένο

Ο έλεγχος αυτός μας δείχνει πως έχουμε μια καλή προσαρμογή του μοντέλου στα δεδομένα.

Ερώτημα 2

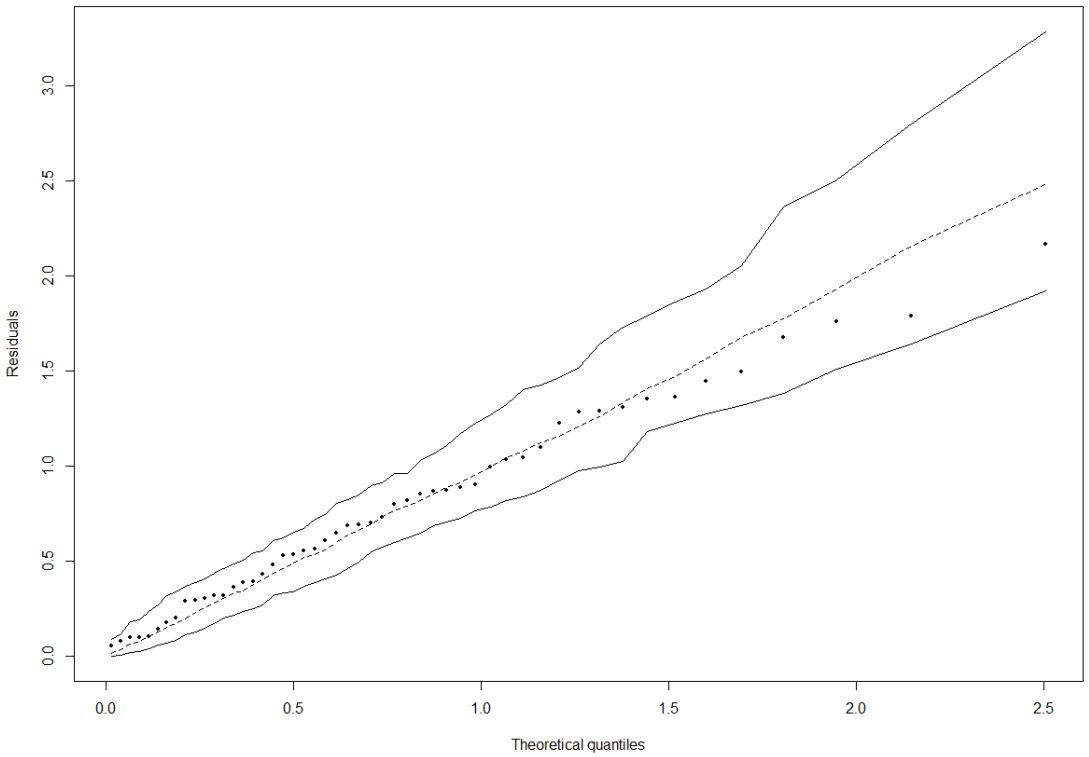
Θα ξεκινήσουμε σχεδιάζοντας τα διαγράμματα των μερικών υπολοίπων, τα οποία μας δείχνουν αν χρειάζεται κάποιος μετασχηματισμός στις μεταβλητές που εισέρχονται στο μοντέλο, δηλαδή στις μεταβλητές age, index και temperature. Τα διαγράμματα φαίνονται παρακάτω:



Πίνακας 26: Διαγράμματα μερικών υπολοίπων για τις μεταβλητές age, index και temperature

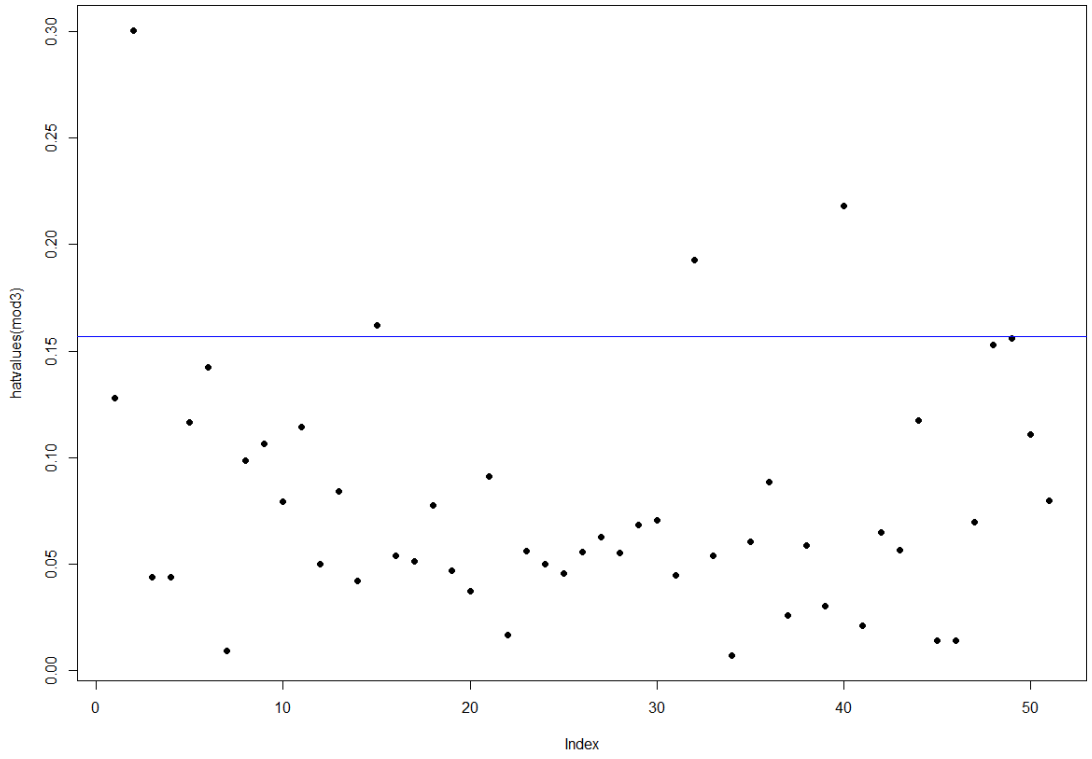
Παρατηρούμε πως καμία μεταβλητή δεν χρειάζονται κάποιον μετασχηματισμό, καθώς οι αντίστοιχες ροζ ευθείες προσομοιάζουν αρκετά καλά τις μπλε ευθείες και συνεπώς υπάρχει μια σχετικά γραμμική εξάρτηση.

Το επόμενο διάγραμμα αφορά τα υπόλοιπα deviance με χρήση της ημι-κανονικής. Στο συγκεκριμένο διάγραμμα, το οποίο φαίνεται παρακάτω, παρατηρούμε πως παρόλο που τα υπόλοιπα δεν ακολουθούν την κανονική κατανομή εμφανίζουν μία γραμμική τάση. Επιπλέον, όλα τα υπόλοιπα βρίσκονται εντός των ορίων εμπιστοσύνης (simulation envelope). Συνεπώς, έχουμε μία πολύ καλή προσαρμογή του μοντέλου στα δεδομένα.



Διάγραμμα 4: Υπόλοιπα Deviance με την ημι-κανονική κατανομή

Θα συνεχίσουμε σχεδιάζοντας το διάγραμμα για τα h_{ii}, τα οποία είναι τα διαγώνια στοιχεία του hat matrix. Παρακάτω παρατίθεται αυτό το διάγραμμα, καθώς και ο πίνακας με τις ακριβείς μετρήσεις των h_{ii}.



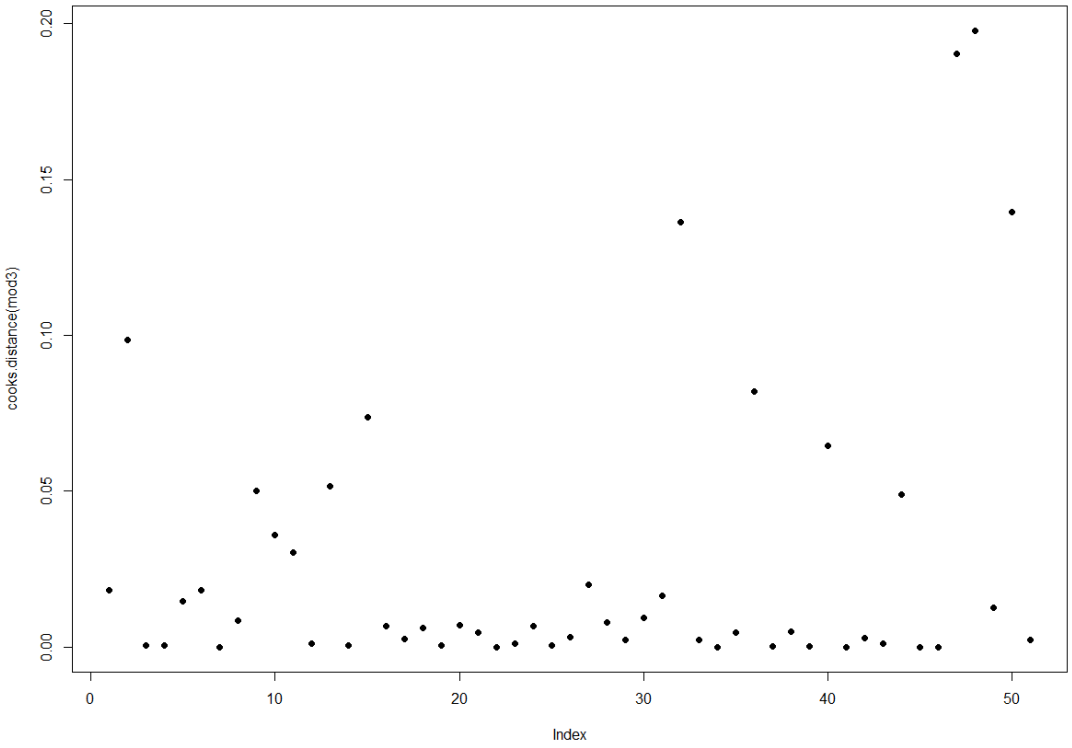
Διάγραμμα 5: Index plot για τα h_{ii} για κάθε μία από τις 51 παρατηρήσεις

```
> hatvalues(mod3)
          1           2           3           4           5           6           7           8           9          10
0.128021121 0.300277730 0.043805585 0.043996020 0.116645535 0.142310570 0.009540398 0.098448143 0.106407457 0.079425832
         11          12          13          14          15          16          17          18          19          20
0.114513474 0.050150857 0.084042179 0.042030916 0.162053954 0.053907734 0.051340310 0.077443933 0.047036451 0.037339803
         21          22          23          24          25          26          27          28          29          30
0.091051119 0.016720424 0.056212268 0.050015940 0.045811600 0.055973338 0.062784355 0.055138323 0.068372591 0.070467458
         31          32          33          34          35          36          37          38          39          40
0.044733773 0.192465647 0.054149739 0.007292811 0.060494759 0.088454319 0.025883426 0.058733891 0.030446914 0.218250406
         41          42          43          44          45          46          47          48          49          50
0.021235315 0.065077553 0.056527486 0.117633961 0.014047366 0.014043800 0.069877622 0.152818057 0.155756460 0.110975694
         51
0.079815587
```

Πίνακας 27: h_{ii} για τις 51 παρατηρήσεις

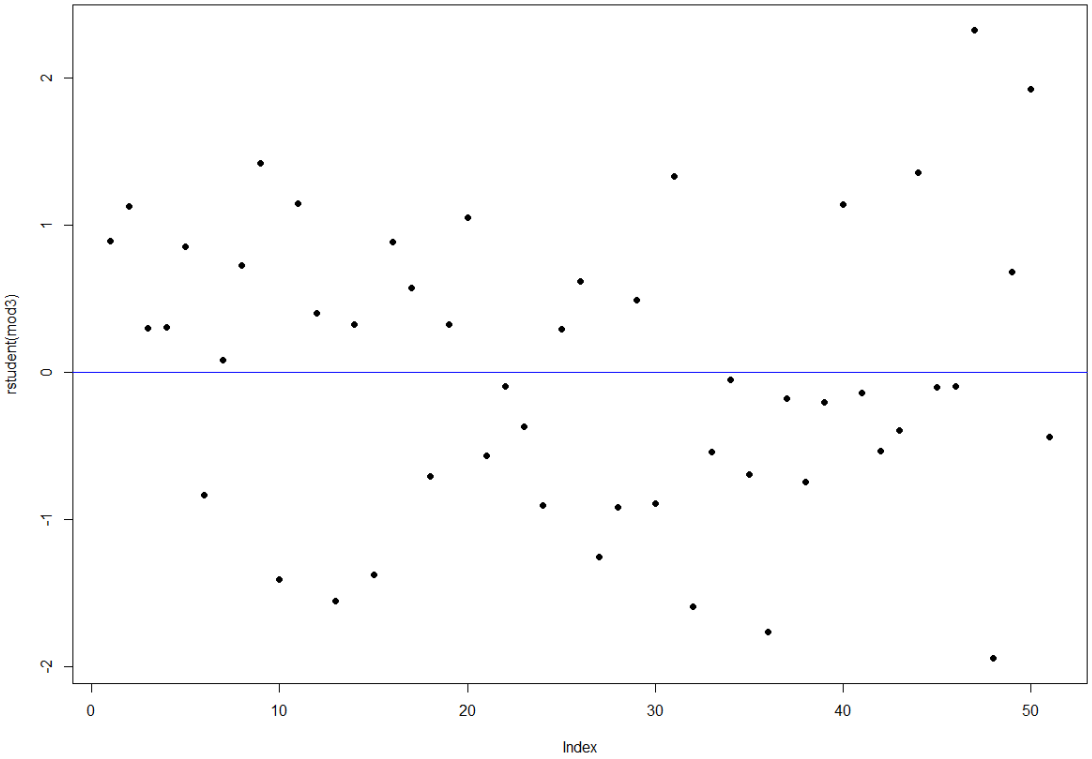
Όπως περιγράψαμε και στην πρώτη άσκηση ο υπολογισμός των h_{ii} μπορεί να χρησιμοποιηθεί ως διαγνωστικός έλεγχος για την πιθανή παρουσία άτυπων σημείων ή σημείων επιρροής. Σύμφωνα με το κριτήριο h_{ii} μία παρατήρηση i θεωρείται σημείο επιρροής αν ισχύει $h_{ii} > 2p/n$, όπου p = k +1 (k ανεξάρτητες μεταβλητές) και n το πλήθος των παρατηρήσεων. Στη περίπτωση μας έχουμε p = 4 και n = 51 και άρα πρέπει $h_{ii} > 0.1569$ (η μπλε γραμμή που σχεδίασα στο διάγραμμα). Κοιτάζοντας τον πίνακα 27 βλέπουμε πως οι παρατηρήσεις 2,15,32 και 40 ξεπερνούν το όριο και συνεπώς μπορούν να θεωρηθούν σημεία επιρροής, ενώ οι παρατηρήσεις 6, 48 και 49 βρίσκονται πολύ κοντά στο όριο αλλά παραμένουν κάτω από αυτό.

Ένα ακόμη κριτήριο για τον εντοπισμό άτυπων σημείων ή σημείων επιρροής αποτελεί η απόσταση Cook. Σύμφωνα με αυτό το κριτήριο αν $D_i >> 1$ η παρατήρηση i θεωρείται σημείο επιρροής. Παρακάτω παρατίθεται το αντίστοιχο διάγραμμα στο οποίο φαίνεται πως όλες οι παρατηρήσεις βρίσκονται αρκετά κάτω από το όριο 1 και συνεπώς καμία δεν μπορεί να θεωρηθεί άτυπο σημείο ή σημείο επιρροής. Αυτό έρχεται σε αντίθεση με το κριτήριο h_{ii}.



Διάγραμμα 6: Απόσταση Cook για κάθε μία από τις 51 παρατηρήσεις

Τέλος, παρακάτω παρατίθεται το διάγραμμα των υπολοίπων πιθανοφάνειας, το οποίο εκφράζει την μεταβολή της deviance αν κάθε φορά παραλείπεται η i-οστή παρατήρηση. Τα υπόλοιπα πιθανοφάνειας είναι τυχαία κατανεμημένα γύρω από το 0 και κανένα δεν ξεφεύγει σε μεγάλο βαθμό. Συνεπώς, δεν εντοπίζουμε σημεία επιρροής και έχουμε μία καλή προσαρμογή του μοντέλου.



Διάγραμμα 7: Index plot για υπόλοιπα πιθανοφάνειας

Ερώτημα 3

Η ερμηνεία των συντελεστών $\hat{\beta}$ του τελικού μοντέλου δίνεται ως εξής:

- Αν η ηλικία του ασθενούς αυξηθεί κατά 1, ενώ οι υπόλοιπες συμμεταβλητές παραμείνουν σταθερές, θα πολλαπλασιαστεί η πιθανότητα ανταπόκρισης στην θεραπεία για λευχαιμία κατά exp(-0.059) = 0.9427, δηλαδή η πιθανότητα ανταπόκρισης στην θεραπεία για λευχαιμία θα μειωθεί κατά περίπου 5.73%.
- Αν ο δείκτης κυττάρων λευχαιμίας αυξηθεί κατά 1, ενώ οι υπόλοιπες συμμεταβλητές παραμείνουν σταθερές, θα πολλαπλασιαστεί η πιθανότητα ανταπόκρισης στην θεραπεία για λευχαιμία κατά exp(0.385) = 1,4696, δηλαδή η πιθανότητα ανταπόκρισης στην θεραπεία για λευχαιμία θα αυξηθεί κατά περίπου 46,96%.
- Αν η υψηλότερη θερμοκρασία πριν την θεραπεία αυξηθεί κατά 1, ενώ οι υπόλοιπες συμμεταβλητές παραμείνουν σταθερές, θα πολλαπλασιαστεί η πιθανότητα ανταπόκρισης στην θεραπεία για λευχαιμία κατά exp(-0.089) = 0.9148, δηλαδή η πιθανότητα ανταπόκρισης στην θεραπεία για λευχαιμία θα μειωθεί κατά περίπου 8,52%.

Η φράση «κατά περίπου» χρησιμοποιείται καθώς αναφερόμαστε στην μέση τιμή. Υπολογίζοντας διαστήματα εμπιστοσύνης για τους συντελεστές $\hat{\beta}$ μπορούμε να είμαστε πιο ακριβείς στο πως θα μεταβληθεί η πιθανότητα ανταπόκρισης στην θεραπεία για λευχαιμία σε κάθε περίπτωση. Τα 95% διαστήματα εμπιστοσύνης για τους συντελεστές $\hat{\beta}$ παρουσιάζονται παρακάτω:

```
> confint.default(mod3)
                  2.5 %        97.5 %
(Intercept) 17.8913230 156.884756210
age         -0.1086306  -0.008372656
index        0.1467548   0.623097373
temperature -0.1596660  -0.018280336
> (exp(confint.default(mod3))-1)*100
                    2.5 %        97.5 %
(Intercept)  5.889831e+09  1.362021e+70
age         -1.029383e+01 -8.337703e-01
index        1.580699e+01  8.646948e+01
temperature -1.475716e+01 -1.811426e+00
```

Πίνακας 28: 95% διαστήματα εμπιστοσύνης για τους συντελεστές $\hat{\beta}$

Η εντολή (exp(confint.default(mod3))-1)*100 μετατρέπει κατάλληλα τα διαστήματα εμπιστοσύνης ώστε να εκφράζουν τις μεταβολές στην πιθανότητα ανταπόκρισης στην θεραπεία, όπως περιέγραψα παραπάνω. Ωστόσο, επειδή η σταθερά είναι ένας πολύ μεγάλος αριθμός σε σχέση με τις μεταβλητές age, index και temperature η R μας δείχνει τους αριθμούς σε scientific form, τους οποίους εμείς θα μετατρέψουμε σε κανονική μορφή με χρήση της εντολής options(scipen = 999). Επιπλέον, θα αγνοήσουμε την σταθερά. Τα κατάλληλα, λοιπόν, διαστήματα εμπιστοσύνης φαίνονται στον παρακάτω πίνακα:

```
                2.5 %     97.5 %
age         -10.29383 -0.8337703
index        15.80699 86.4694762
temperature -14.75716 -1.8114265
```

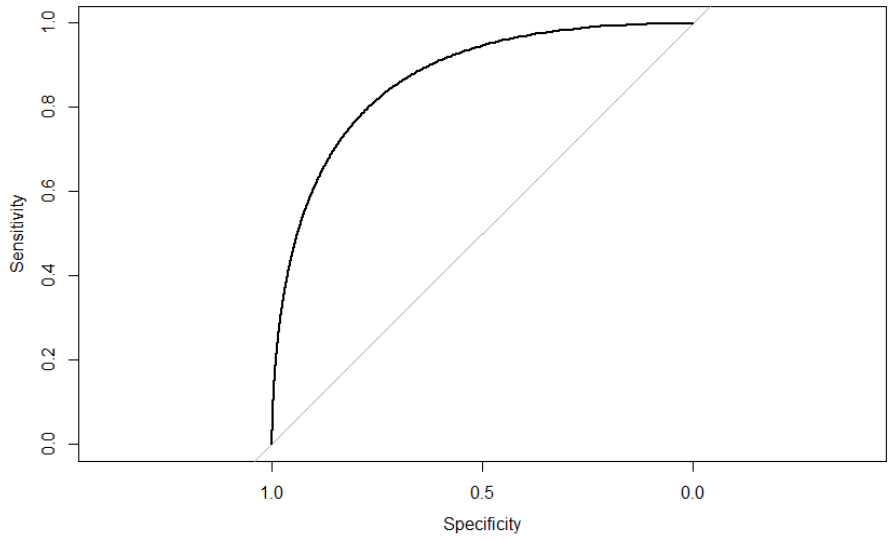
Πίνακας 29: Κατάλληλη μετατροπή στα 95% διαστήματα εμπιστοσύνης των συντελεστών $\hat{\beta}$

Με αυτό τον τρόπο έχουμε τις ακόλουθες ερμηνείες:

- Αν η ηλικία του ασθενούς αυξηθεί κατά 1 η πιθανότητα ανταπόκρισης στην θεραπεία για λευχαιμία μειώνεται κατά (-10.29, -0.83) με μέση τιμή -5.73%.
- Αν ο δείκτης κυττάρων λευχαιμίας αυξηθεί κατά 1 η πιθανότητα ανταπόκρισης στην θεραπεία για λευχαιμία αυξάνεται κατά (15.81, 86,47) με μέση τιμή 46.96%.
- Αν η υψηλότερη θερμοκρασία πριν την θεραπεία αυξηθεί κατά 1 η πιθανότητα ανταπόκρισης στην θεραπεία για λευχαιμία μειώνεται κατά (-14,76, -1,81) με μέση τιμή -8,52%.

Ερώτημα 4

Η καμπύλη ROC (receiver operating characteristic curve) απεικονίζει την προβλεπτική ικανότητα του μοντέλου (τιμές ευαισθησίας και ειδικότητας) καθώς το όριο p_0 μεταβάλλεται στο εύρος [0,1]. Μεγάλη επιτυχία πρόβλεψης συνεπάγεται πως το εμβαδόν κάτω από την καμπύλη (area under the curve - AUC) θα είναι υψηλό και κοντά στο 1. Η καμπύλη ROC για το τελικό μοντέλο response ~ age + index + temperature απεικονίζεται στο παρακάτω διάγραμμα:



Διάγραμμα 8: Καμπύλη ROC

```
Call:
roc.default(response = response, predictor = fitted.values(mod3),     smooth = TRUE, ci = TRUE, plot = TRUE)

Data: fitted.values(mod3) in 27 controls (response 0) < 24 cases (response 1).
Smoothing: binormal
Area under the curve: 0.8686
95% CI: 0.7558-0.9447 (2000 stratified bootstrap replicates)
```

Πίνακας 30: Καμπύλη ROC – Area under the curve

Παρατηρούμε πως υπάρχουν τιμές του ορίου p_0 με υψηλή ευαισθησία και ταυτόχρονα υψηλή ειδικότητα. Αυτό φαίνεται και από τον εμβαδόν κάτω από την καμπύλη, το οποίο είναι ίσο με 0.8686, μία αρκετά υψηλή τιμή και κοντά στο 1. Συνεπώς το τελικό μοντέλο μας έχει μια καλή προβλεπτική ικανότητα.