



ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ
ΥΠΟΛΟΓΙΣΤΩΝ
Ε.ΔΕ.Μ.Μ
ΜΑΘΗΜΑ: ΣΤΑΤΙΣΤΙΚΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ
ΧΕΙΜΕΡΙΝΟ ΕΞΑΜΗΝΟ 2021-2022

3η ΣΕΙΡΑ ΑΣΚΗΣΕΩΝ



Άσκηση 1

Ερώτημα 1

Στην άσκηση αυτή θα μελετηθεί η εξάρτηση του αριθμού Y των αποζημιώσεων λόγω τροχαίων ατυχημάτων. Η τυχαία μεταβλητή Y ακολουθεί κατανομή Poisson και εκφράζεται ανά n αριθμό συμβολαίων (offset). Οι μεταβλητές από τις οποίες εξαρτάται ο αριθμός των αποζημιώσεων παρουσιάζεται στον Πίνακα 1.

Όνομα Μεταβλητής	Περιγραφή
agecat	ηλικία ασφαλιζόμενου ($X_1 = 0$ – νέος και $X_1 = 1$ – μεγάλος)
cartype	κατηγορία ασφαλιστρών ($X_2 = 1, 2, 3, 4$)
district	περιοχή διαμονής ασφαλισμένου ($X_3 = 1$ – Αθήνα και $X_3 = 0$ – σε άλλη πόλη)

Πίνακας 1: Επεξηγηματικές μεταβλητές από τις οποίες εξαρτάται ο αριθμός αποζημιώσεων Y

Καθώς η τ.μ. Y εκφράζεται ανά n αριθμό συμβολαίων, τότε το μοντέλο της Poisson παλινδρόμησης που θα προσαρμοστεί στα δεδομένα, θα έχει την μορφή:

$$\mu_i^* = n_i \mu_i = n_i \exp(X_i' \beta) \Rightarrow \ln\left(\frac{\mu_i^*}{n_i}\right) = X_i' \beta \quad (1)$$

όπου μ_i^* αντιστοιχεί στην αναμενόμενη τιμή της παρατήρησης i , η οποία εκφράζεται ανά n_i συμβόλαια. Αρχικά το μοντέλο της Poisson παλινδρόμησης θα προσαρμοστεί και στις 3 προαναφερόμενες μεταβλητές ώστε να γίνουν οι κατάλληλοι έλεγχοι σημαντικότητας για την κάθε μια από αυτές. Για τον σκοπό αυτό καλείται η συνάρτηση `glm()` της R και ορίζεται ως κατηγορική η μεταβλητή `cartype` μέσω της συνάρτησης `factor()`. Επιπλέον ορίζεται και το `offset` της παλινδρόμησης για τον αριθμό συμβολαίων. Ο κώδικας παρουσιάζεται στο τέλος της εργασίας. Για το κατά πόσο στατιστικά σημαντική είναι η κάθε μεταβλητή, πραγματοποιείται αυτοματοποιημένα από την τη συνάρτηση `glm()`, ο έλεγχος Wald, του οποίου η ελεγχουσυνάρτηση δίνεται από τη σχέση:

$$Z = \frac{\hat{\beta}_j - \beta_j}{\text{se}(\hat{\beta}_j)} \sim N(0, 1), j = 0, 1, 2, 3 \quad (2)$$

όπου για μηδενική υπόθεση H_0 θεωρείται ότι $\beta_j = 0$ και για την εναλλακτική H_1 ισχύει ότι $\beta_j \neq 0$. Οι p values που προκύπτουν από τον έλεγχο Wald είναι μικρότερες του 0.05 και πιο συγκεκριμένα είναι 2e-16, 0.001309, 6.03e-13, 4.64e-15 και 0.000215, για την `agecat`, την `cartype` σε κάθε κατηγορία και την `district` αντιστοίχως. Επομένως, σύμφωνα με αυτόν τον έλεγχο η ποσότητα των αποζημιώσεων εμφανίζει ισχυρή εξάρτηση από την ηλικία, την κατηγορία ασφαλιστρών αλλά και από την τοποθεσία κατοικίας του ασφαλιζόμενου.

Ο επόμενος έλεγχος που θα πραγματοποιηθεί στηρίζεται στην συνάρτηση Deviance η οποία ακολουθεί ασυμπτωτικά την κατανομή χ^2 και δίνεται από τη σχέση:

$$D(\hat{\beta}) = -2\{\ell(\hat{\beta}) - \tilde{\ell}\} \sim \chi_{n-p}^2 \quad (3)$$

όπου $\ell(\hat{\beta})$ είναι η πιθανοφάνεια του μοντέλου που εξετάζεται και $\tilde{\ell}$ η πιθανοφάνεια του κορεσμένου μοντέλου όπου για κάθε παρατήρηση αντιστοιχεί μια παράμετρος και έτσι δεν τίθεται κάποιος περιορισμός. Οι βαθμοί ελευθερίας της κατανομής χ^2 είναι $n-p=32-(5+1)=26$, καθώς το κορεσμένο μοντέλο έχει 32 παραμέτρους, όσα και τα δείγματα ενώ το προσαρμοσμένο έχει 6. Θεωρώντας ως H_0 το προσαρμοσμένο

μοντέλο και ως εναλλακτική υπόθεση H_1 το κορεσμένο, η p-value του ελέγχου ισούται με 0.02580847. Συνεπώς, η H_1 είναι στατιστικά σημαντική και έτσι το μοντέλο που προσαρμόζεται δεν είναι αρκετά ικανοποιητικό. Ωστόσο, επειδή η deviance ακολουθεί ασυμπτωτικά την χ^2 και τα δείγματα δεν είναι αρκετά, το συμπέρασμα αυτό είναι σχετικά επισφαλές. Συνεπώς, η μελέτη θα συνεχιστεί με το προσαρμοσμένο μοντέλο.

Επιπλέον, με χρήση της διαφοράς των deviance θα εξεταστεί πόσο αποτελεσματικά έχει προσαρμοστεί το μοντέλο στις τιμές. Θεωρώντας λοιπόν ως μηδενική υπόθεση H_0 ένα μοντέλο M_0 και ως εναλλακτική υπόθεση H_1 το πλήρες μοντέλο με τις 5 μεταβλητές M_1 (3 εκ των οποίων η μία είναι κατηγορική με 4 τιμές), στο οποίο το M_0 είναι εμφωλευμένο, τότε η διαφορά των συναρτήσεων Deviance ακολουθεί την κατανομή χ_q^2 , όπου q η διαφορά στον αριθμό των παραμέτρων μεταξύ των δύο μοντέλων. Η αντίστοιχη ελεγχουσυνάρτηση δίνεται από τη σχέση:

$$D(\hat{\beta}_0) - D(\hat{\beta}_1) = -2(\hat{\ell}_0 - \hat{\ell}_1) \sim \chi_q^2 \quad (4)$$

Στην συγκεκριμένη περίπτωση, ως M_0 θα θεωρηθεί το σταθερό μοντέλο, δηλαδή αυτό που δεν συμμετέχει καμία επεξηγηματική μεταβλητή και επομένως $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$. Συνεπώς, η διαφορά μεταξύ των παραμέτρων θα είναι q=5. Η διαφορά της deviance μεταξύ των δύο μοντέλων είναι 166.0446 και η p-value για την κατανομή χ_5^2 είναι σχεδόν μηδενική. Συνεπώς, είναι στατιστικά σημαντική η απόρριψη του M_0 και έτσι το M_1 παράγει αξιόπιστες προβλέψεις για τις αναμενόμενες τιμές της ποσότητας των αποζημιώσεων.

Το τελευταίο συμπέρασμα επιβεβαιώνεται χρησιμοποιώντας και βηματικές τεχνικές επιλογής μοντέλου με κριτήριο την ελαχιστοποίηση του AIC. Ειδικότερα, αξιοποιείται η συνάρτηση step() η οποία με backward elimination καταλήγει στο βέλτιστο μοντέλο με βάση την ελεγχουσυνάρτηση deviance της εξίσωσης 4 και την παράλληλη ελαχιστοποίηση του AIC. Το αποτέλεσμα του συγκεκριμένου ελέγχου καταλήγει πως πράγματι η πιθανοφάνεια μεγιστοποιείται για το πλήρες μοντέλο με τιμή AIC = 222.1.

Ερώτημα 2

Καθώς το βέλτιστο μοντέλο είναι το πλήρες και η κατανομή που ακολουθούν ασυμπτωτικά οι παράμετροι του μοντέλου είναι η κανονική, όπως καταδεικνύει η σχέση 2, τότε για $\alpha=0.05$, κατασκευάζεται ένα 95% διάστημα εμπιστοσύνης για τον κάθε συντελεστή, όπως παρουσιάζεται στον Πίνακα 2. Ωστόσο, επειδή οι παράμετροι του μοντέλου επεξηγούν εκθετικά την εξαρτημένη μεταβλητή σύμφωνα με τη σχέση 1, οι τιμές που παρουσιάζονται στον πίνακα αντιστοιχούν στις εκθετικές τιμές των διαστημάτων εμπιστοσύνης. Επιπλέον στον Πίνακα 2 αναγράφεται και η εκθετική τιμή του κάθε συντελεστή $e^{\hat{\beta}_i}$

Παράμετρος	Τιμή $e^{\hat{\beta}_i}$	2.5%	97.5%
Πόλωση (Intercept)	0.1443921	0.1295741	0.1609045
factor(cartype)2	1.1761330	1.0653437	1.2984438
factor(cartype)3	1.4848997	1.3333892	1.6536260
factor(cartype)4	1.7602026	1.5280757	2.0275915
agecat	0.6864111	0.6290616	0.7489890
district	1.2418604	1.1072684	1.3928124

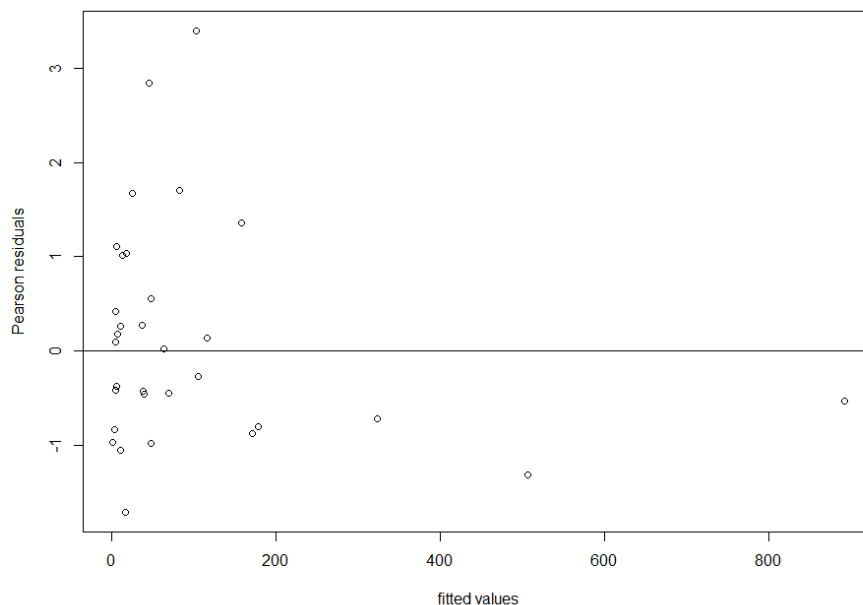
Πίνακας 2: εκτίμηση και 95% διαστήματα εμπιστοσύνης για τους εκθετικούς συντελεστές του μοντέλου

Για τους συντελεστές προκύπτουν οι παρακάτω ερμηνείες:

1. Εάν η κατηγορία ασφαλιστρου μεταβληθεί από την κατηγορία αναφοράς 1 στην κατηγορία 2, τότε η αναμενόμενη ποσότητα των αποζημιώσεων Y , θα αυξηθεί κατά $100|1-1.176| = 17.6\%$.
2. Εάν η κατηγορία ασφαλιστρου μεταβληθεί από την κατηγορία αναφοράς 1 στην κατηγορία 3, τότε η αναμενόμενη ποσότητα των αποζημιώσεων Y , θα αυξηθεί κατά $100|1-1.484| = 48.4\%$.
3. Εάν η κατηγορία ασφαλιστρου μεταβληθεί από την κατηγορία αναφοράς 1 στην κατηγορία 4, τότε η αναμενόμενη ποσότητα των αποζημιώσεων Y , θα αυξηθεί κατά $100|1-1.760| = 76\%$.
4. Εάν η ηλικία του ασφαλιζόμενου μεταβληθεί από νέος σε μεγάλος, τότε τότε η αναμενόμενη ποσότητα των αποζημιώσεων Y , θα μειωθεί κατά $100|1-0.686| = 31.4\%$
5. Εάν η περιοχή διαμονής του ασφαλιζόμενου μεταβληθεί από μια επαρχιακή πόλη στην Αθήνα, τότε τότε η αναμενόμενη ποσότητα των αποζημιώσεων Y , θα αυξηθεί κατά $100|1-1.242| = 24.2\%$.

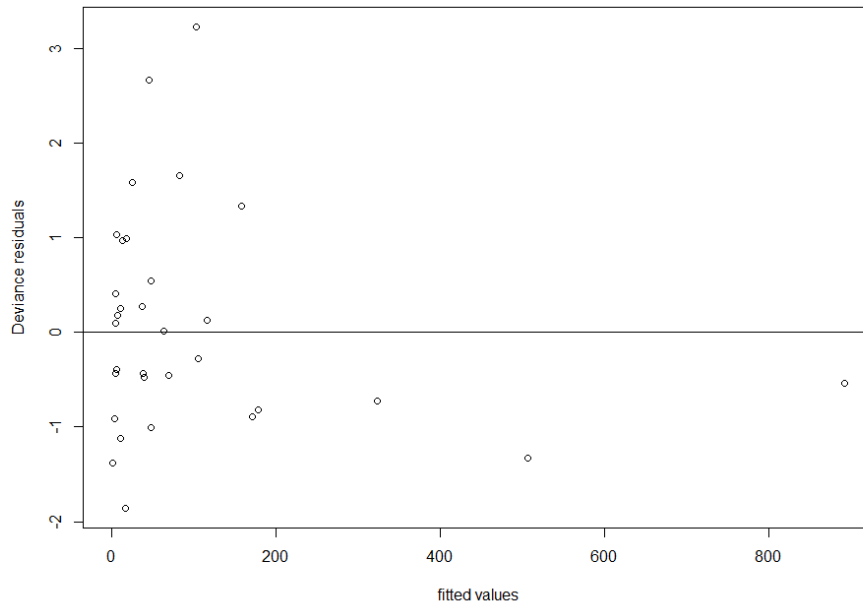
Ερώτημα 3

Για την καταλληλότητα του μοντέλου θα εξεταστούν τα υπόλοιπα Pearson καθώς επίσης και τα υπόλοιπα Deviance. Ειδικότερα, όπως παρουσιάζεται στα Γραφήματα 1 και 2 δεν εμφανίζεται κάποια εξάρτηση σχετικά με τις προσαρμοσμένες τιμές. Τόσο στα deviance υπόλοιπα όσο και στα pearson, οι περισσότερες τιμές εμφανίζονται στο εύρος $(-2, 2)$. Ορισμένες ωστόσο παρατηρήσεις προκύπτουν εκτός αυτού του εύρους και αυτό αποτελεί πιθανώς ένδειξη για outliers. Οι παρατηρήσεις αυτές αντιστοιχούν στο 1ο και στο 11ο δείγμα.



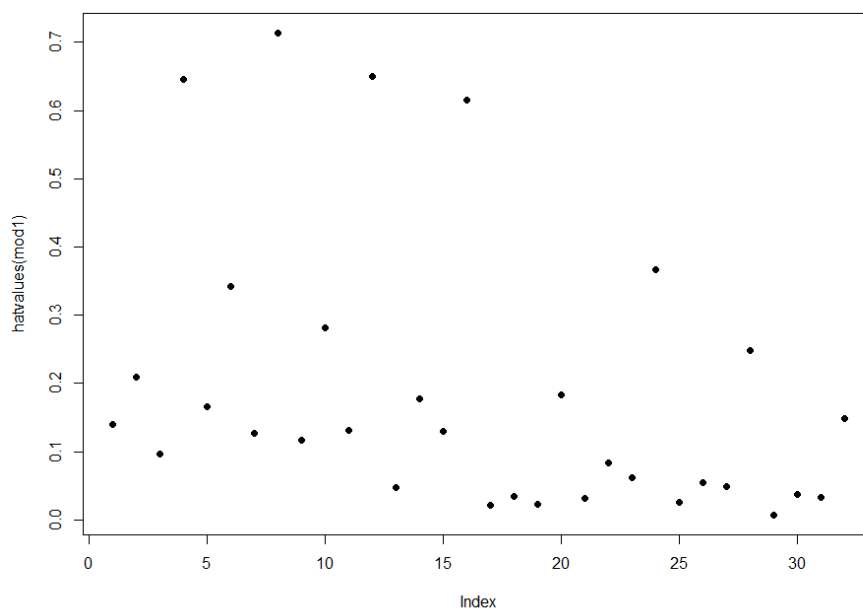
Εικόνα 1: Pearson Residuals diagnostics plot

Οι έλεγχοι για outliers συνεχίζονται με τον υπολογισμό των στοιχείων h_{ii} . Ειδικότερα, οι παρατηρήσεις που μπορούν να αποτελέσουν πιθανά σημεία επιρροής θα πρέπει να εμφανίζουν τιμές $h_{ii} > 2p/n = 0.375$, όπου $p = 6$ ο αριθμός των συντελεστών του μοντέλου poisson και $n = 32$ οι παρατηρήσεις του συνόλου των δεδομένων. Τα δείγματα που ικανοποιούν αυτή τη συνθήκη αντιστοιχούν στις

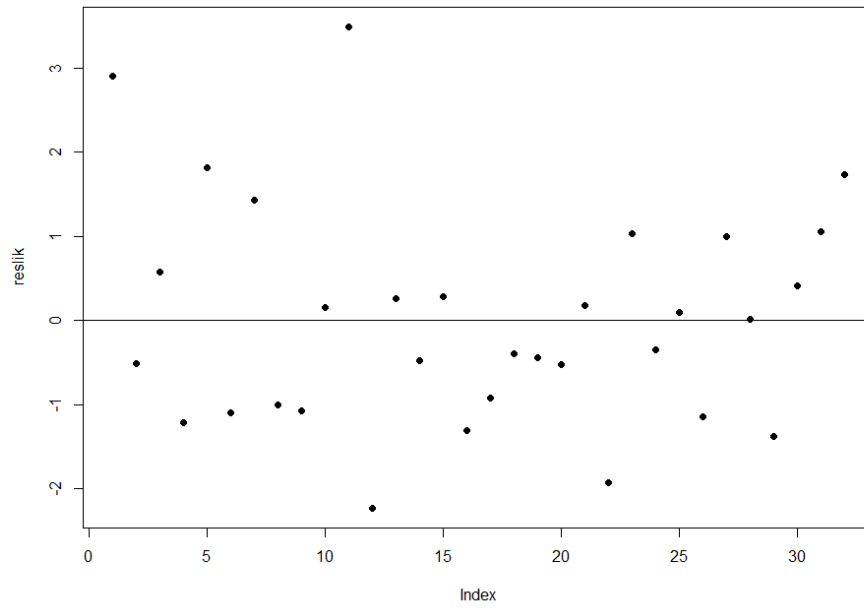


Εικόνα 2: Deviance Residuals diagnostics plot

παρατηρήσεις 4,8,12 και 16 όπως παρατηρείται και στο Γράφημα 3. Επιπροσθέτως, οι τιμές των h_{ii} επηρεάζουν και τα υπόλοιπα πιθανοφάνειας. Για τις συγκεκριμένες παρατηρήσεις ελέγχονται οι τιμές αυτών των υπολοίπων και παρατηρείται πως η υψηλότερη τιμή που εμφανίζεται με διαφορά είναι το 2.230321 για την παρατήρηση 12. Οι παρατηρήσεις που ξεπερνούν αυτή την τιμή για το υπόλοιπο είναι η 1η και η 11η, όπως παρουσιάζεται και στο Γράφημα 4. Το γεγονός αυτό συμφωνεί και με τις παρατηρήσεις στα υπόλοιπα Pearson και Deviance. Επομένως η 1η και η 11η παρατήρηση είναι σημεία επιρροής, ως προς τα υπόλοιπα πιθανοφάνειας, τα οποία εκφράζουν την μεταβολή της Deviance αν παραληφθεί από το μοντέλο η κάθε μια από αυτές τις παρατηρήσεις.

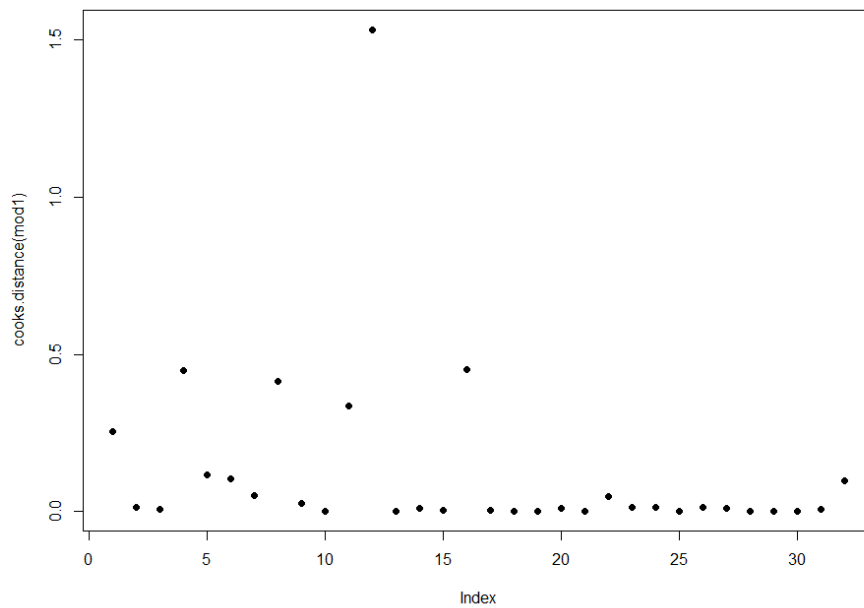


Εικόνα 3: Hat values



Εικόνα 4: Likelihood residuals diagnostics plot

Το ισχυρότερο όμως από τα σημεία επιρροής αποτελεί η 12η παρατήρηση καθώς προκύπτει και με βάση τον έλεγχο της απόστασης Cook, όπως παρουσιάζεται στο Γράφημα 5. Ειδικότερα, η απόσταση Cook για την συγκεκριμένη παρατήρηση είναι μεγαλύτερη του 1, το οποίο αποτελεί ισχυρή ένδειξη σημείου επιρροής. Επιπρόσθετα σε αυτό τον έλεγχο, συμφωνούν και οι έλεγχοι που πραγματοποιήθηκαν στα h_{ii} καθώς επίσης και στα υπόλοιπα πιθανοφάνειας.

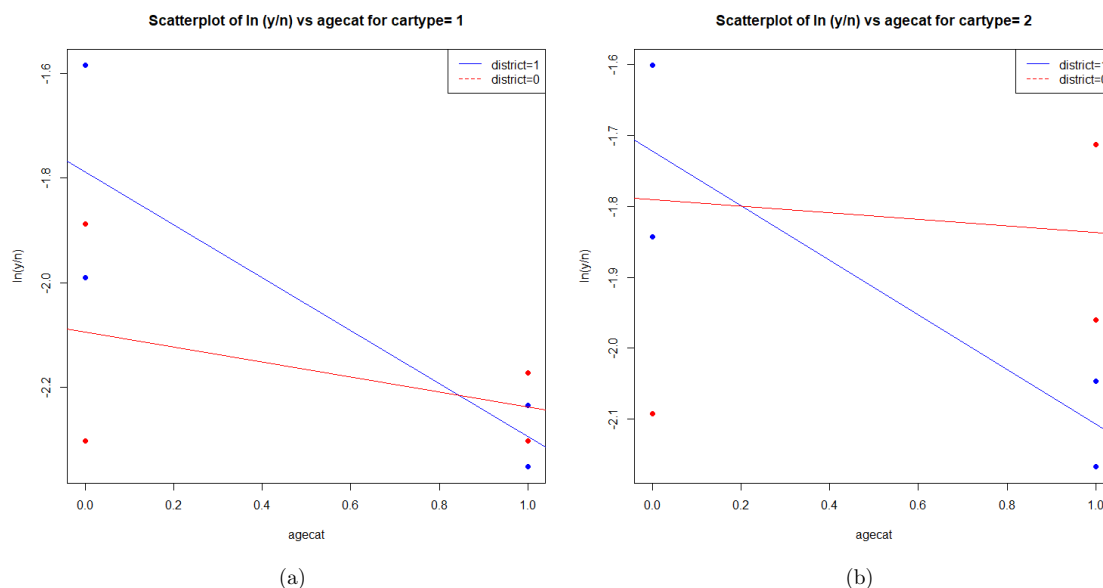


Εικόνα 5: Απόσταση Cook

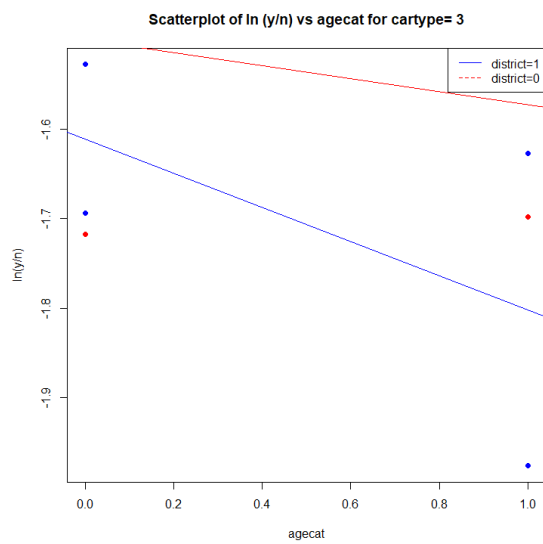
Ερώτημα 4

Το τελευταίο βήμα της ανάλυσης αφορά την εύρεση κατάλληλης αλληλεπίδρασης μεταξύ των μεταβλητών, η οποία θα περιγράψει αποτελεσματικότερα την σχέση του αριθμού των αποζημιώσεων Y ανά αριθμό συμβολαίων n . Ειδικότερα εξετάστηκαν οι αλληλεπιδράσεις $\text{factor}(\text{cartype}) * \text{agecat}$, $\text{factor}(\text{cartype}) * \text{district}$ και $\text{agecat} * \text{district}$, από τις οποίες μόνο η τελευταία προέκυψε οριακά στατιστικά σημαντική σύμφωνα με τον έλεγχο Wald. Ειδικότερα, η p -value του ελέγχου Wald για την νέα μεταβλητή $\text{agecat} * \text{district}$ προκύπτει 0.05633, η οποία θα γίνει οριακά αποδεκτή. Ωστόσο, κατά την προσαρμογή του μοντέλου, ενώ το σύνολο των μεταβλητών συμπεριλαμβανομένης και της αλληλεπίδρασης προκύπτει στατιστικά σημαντικό, η μεταβλητή district δεν είναι στατιστικά σημαντική με $p\text{-value} = 0.70125$. Πάραυτα, το μοντέλο θα διατηρηθεί καθώς η σημαντικότητα της district προκύπτει μέσω της αλληλεπίδρασης. Ειδικότερα, το εκθετικό του συντελεστή για την αλληλεπίδραση είναι 1.387 και επομένως ο τόπος κατοικίας και συγκεκριμένα η Αθήνα ($\text{district}=1$), θα αυξήσει τις αποζημιώσεις κατά 38.7% αν ο οδηγός είναι μεγάλος σε ηλικία σε σύγκριση με τον νεαρό.

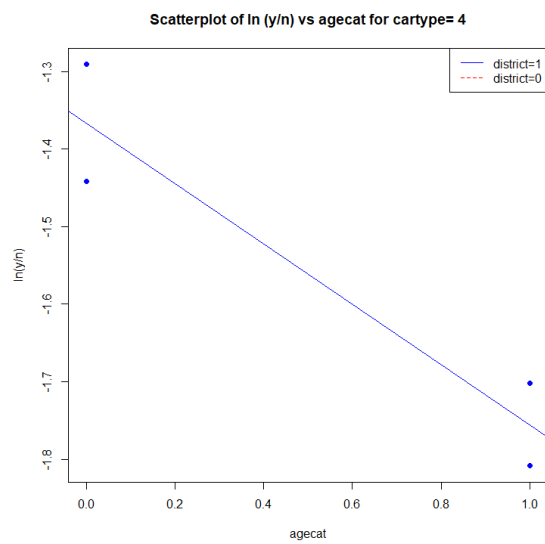
Στα διαγράμματα 6 και 7 παρουσιάζεται γραφικά η διαφορετική πτωτική τάση του λογαρίθμου των αποζημιώσεων ανά συμβόλαιο σε σχέση με την ηλικία (agecat) του ασφαλιζόμενου ανάλογα με το αν διαμένει στην Αθήνα ($\text{district}=1$) ή όχι ($\text{district}=0$). Οι τέσσερις γραφικές παραστάσεις αντιστοιχούν στα 4 διαφορετικά είδη συμβολαίου. Καθώς οι δύο γραφικές παραστάσεις για την διαμονή και μη στην Αθήνα παρουσιάζουν σημείο τομής, πράγματι η αλληλεπίδραση $\text{agecat} * \text{district}$ είναι σημαντική. Στην γραφική παράσταση για το την κατηγορική μεταβλητή $\text{cartype}=4$, δεν υπήρχαν αποζημιώσεις για την επαρχία ($\text{district}=0$).



Εικόνα 6: Scatterplot του $\ln(y/n)$ σε σχέση agecat για $\text{cartype}=1,2$



(a)



(b)

Εικόνα 7: Scatterplot του $\ln(y/n)$ σε σχέση agecat για cartype=3,4

Άσκηση 2

Ερώτημα 1

Στην άσκηση αυτή θα μελετηθεί η εξάρτηση της πιθανότητας ανταπόκρισης μιας θεραπείας για την λευχαιμία από τις συµμεταβλητές age, smear, infiltrate, index, blasts και temperature. Για τον σκοπό αυτό θα χρησιμοποιηθεί το µοντέλο της λογιστικής παλινδρόµησης, το οποίο θα προσαρµοστεί σύµφωνα µε τη σχέση:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik}, i = 1, \dots, n \quad (5)$$

όπου η πιθανότητα ανταπόκρισης p_i στη θεραπεία προκύπτει από τη σχέση:

$$p_i = p_{x_i} = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})} \quad (6)$$

Ο στόχος είναι η εύρεση του καλύτερου δυνατού µοντέλου, δηλαδή του αριθµού των παραµέτρων που περιγράφουν την εξαρτηµένη δυαδική µεταβλητή response. Η µεταβλητή αυτή ακολουθεί την κατανοµή Bernoulli µε παράµετρο p , την πιθανότητα της ανταπόκρισης στην θεραπεία. Οι συµμεταβλητές του προβλήµατος και η σηµασία τους παρουσιάζονται στον Πίνακα 3.

Όνοµα Μεταβλητής	Περιγραφή
age	ηλικία του ασθενούς
smear	ποσοστό επίστρωσης βλαστοκυττάρων
infiltrate	ποσοστό κυττάρων στο μυελό των οστών
index	δείκτης κυττάρων λευχαιµίας
blasts	βλαστοκύτταρα
temperature	υψηλότερη θερμοκρασία πριν τη θεραπεία ($\times 10^\circ \text{ F}$)

Πίνακας 3: Μεταβλητές προς προσαρµογή

Αρχικά, θα προσαρµοστεί το µοντέλο της λογιστικής παλινδρόµησης στο σύνολο των συµμεταβλητών που αναφέρθηκαν προκειµένου να ελεγχθεί εάν είναι στατιστικώς σηµαντικές για το µοντέλο. Ο κώδικας παρουσιάζεται αναλυτικά στο τέλος της εργασίας. Σύµφωνα µε τον έλεγχο Wald για την κάθε παράµετρο $\hat{\beta}_j$ του µοντέλου προκύπτει πως µόνο για τις µεταβλητές age, index, temperature και τον σταθερό όρο απορρίπτεται η µηδενική υπόθεση για τον µηδενισµό των συντελεστών τους µε p -values 0.02714, 0.00269, 0.01448 και 0.01588 αντίστοιχα. Εποµένως, το µοντέλο µε τις 6 επεξηγηµατικές µεταβλητές δεν είναι το βέλτιστο. Αυτό, επίσης επιβεβαιώνεται και µέσω της σύγκρισης της ελεγχοσυνάρτησης Deviance για την πρόσθεση των µεταβλητών στο µοντέλο. Ειδικότερα, µέσω της εντολής `anova()` για το µοντέλο µε τις 6 επεξηγηµατικές µεταβλητές και µε βάση των έλεγχο χ^2 ελέγχεται η µεταβολή της Deviance µε την πρόσθεση της κάθε µεταβλητής. Επιπλέον και αυτός ο έλεγχος καταδεικνύει πως οι µοναδικές στατιστικά σηµαντικές µεταβλητές είναι αυτές που προέκυψαν από τον έλεγχο Wald. Εποµένως, κρίνεται απαραίτητο να χρησιµοποιηθούν βηµατικές τεχνικές για την επιλογή του βέλτιστου µοντέλου.

Η επιλογή του βέλτιστου µοντέλου θα εξεταστεί µε χρήση της εντολής `step()` τόσο µε βάση το forward όσο και µε βάση το backward κριτήριο. Για τον έλεγχο της σηµαντικότητας των µεταβλητών αξιοποιείται η ελεγχοσυνάρτηση Deviance και ο έλεγχος χ^2 ενώ για την επιλογή του βέλτιστου µοντέλου χρησιµοποιείται το κριτήριο AIC. Και οι δύο µέθοδοι καταλήγουν στο ίδιο µοντέλο σύµφωνα µε το οποίο η εξαρτηµένη µεταβλητή response και κατ' επέκταση η σχετική πιθανότητα ανταπόκρισης

στη θεραπεία εξαρτώνται από της συµµεταβλητές age, infiltrate, index, temperature µε αντίστοιχο AIC=50.14. Ωστόσο, η συµµεταβλητή infiltrate δεν προκύπτει στατιστικά σηµαντική, καθώς η p-value του ελέγχου Wald είναι 0.10077 εποµένως δεν µπορεί να απορριφθεί η υπόθεση πως η µεταβλητή δεν ανήκει στο µοντέλο. Συνεπώς, θα πρέπει να µελετηθεί το µοντέλο το οποίο περιέχει µόνο τις µεταβλητές age, index και temperature.

Αρχικά, το AIC του µοντέλου $M_1 = (\text{age, index, temperature})$ προκύπτει στα 51.26538, το οποίο ενώ είναι µεγαλύτερο από το AIC=50.14 του $M_2 = (\text{age, index, temperature, infiltrate})$, η διαφορά δεν είναι αξιοσηµείωτη. Επιπλέον, θα εξεταστεί η πρόσθεση της µεταβλητής infiltrate αυστηρά µε βάση τον έλεγχο χ^2 . Ειδικότερα, για τα δύο µοντέλα $M_1 \subset M_2$, αντιστοιχίζεται η µηδενική υπόθεση H_0 στο M_1 και η εναλλακτική υπόθεση H_1 στο M_2 . Η µεταβολή της deviance είναι:

$$D(M_1) - D(M_2) = 43.265 - 40.136 = 3.129 \sim \chi_1^2 \quad (7)$$

και η p-value του ελέγχου είναι $0.0769 > 0.05$. Εποµένως, η µηδενική υπόθεση δεν απορρίπτεται και συνεπώς µε βάση το έλεγχο deviance διατηρείται το µοντέλο $M_1 = (\text{age, index, temperature})$. Η συγκεκριµένη µελέτη πραγµατοποιήθηκε µέσω της εντολής anova() για τα µοντέλα M_1 και M_2 µε βάση τον έλεγχο χ^2 . Επιπλέον για το µοντέλο M_1 πραγµατοποιείται και έλεγχος για την µεταβολή της deviance σε σχέση µε το σταθερό µοντέλο M_0 , όπου $M_0 \subset M_1$, µε στόχο να εξεταστεί η καλή προσαρµογή του µοντέλου στα δεδοµένα. Η p-value του ελέγχου είναι 5.195202×10^{-6} και εποµένως το µοντέλο M_1 θα διατηρηθεί για την λογιστική παλινδρόµηση των συγκεκριµένων δεδοµένων.

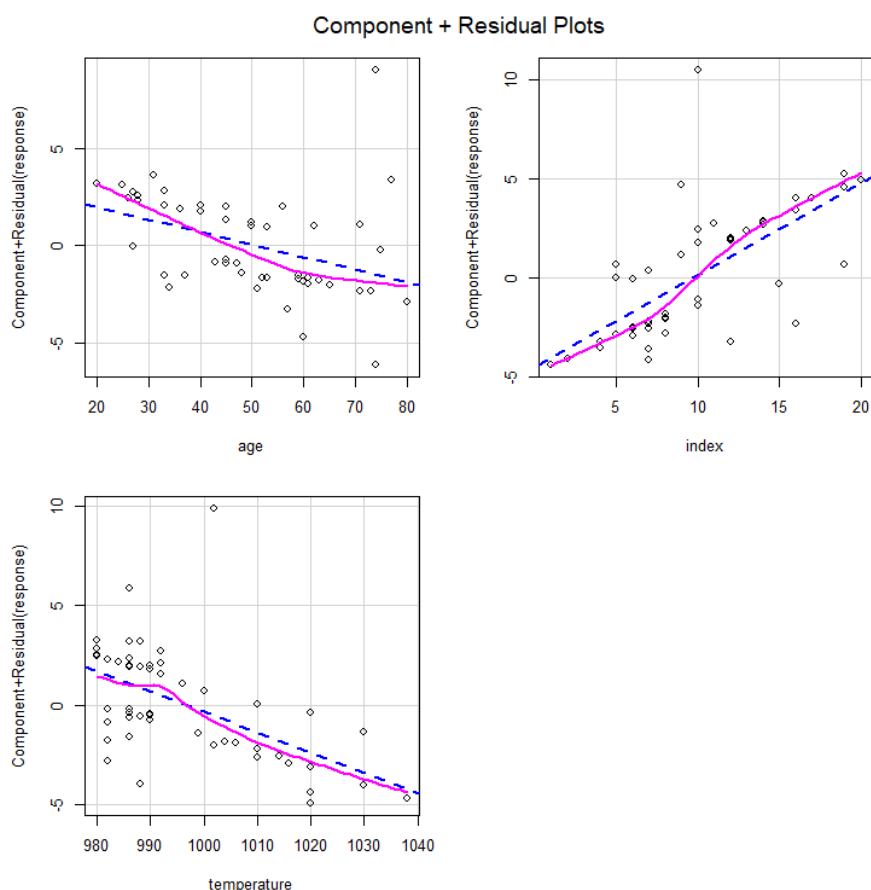
Ερώτηµα 2

Για το µοντέλο $M_1 = (\text{age, index, temperature})$ θα δηµιουργηθούν οι γραφικές παραστάσεις των µερικών υπολοίπων, των υπολοίπων Deviance µε την ηµι-κανονική κατανοµή, τα index plots των hii και των αποστάσεων Cook για την ανίχνευση outliers, καθώς επίσης και των υπολοίπων πιθανοφάνειας.

Αρχικά, εξετάζοντας τα διαγράµµατα µερικών υπολοίπων στην Εικόνα 8 παρατηρείται πως καµία από τις µεταβλητές που χρησιµοποιήθηκαν δεν χρειάζεται κάποιον µετασχηµατισµό και εποµένως, ο λογάριθµος την σχετικής πιθανότητας ανταπόκρισης στην θεραπεία εξαρτάται γραµµικά από αυτές. Ειδικότερα, για κάθε µια από της µεταβλητές, πραγµατοποιείται λογιστική παλινδρόµηση ως προς τις υπόλοιπες µεταβλητές του µοντέλου και τα υπόλοιπα που προκύπτουν απεικονίζονται συναρτήσει της µεταβλητής που απέχει από το µοντέλο. Εάν η εξάρτηση που εµφανίζεται είναι γραµµική, το οποίο συµβαίνει σε αυτή την περίπτωση και για τις τρεις µεταβλητές, τότε οι µεταβλητές δεν χρειάζεται να υποστούν κάποιο µετασχηµατισµό.

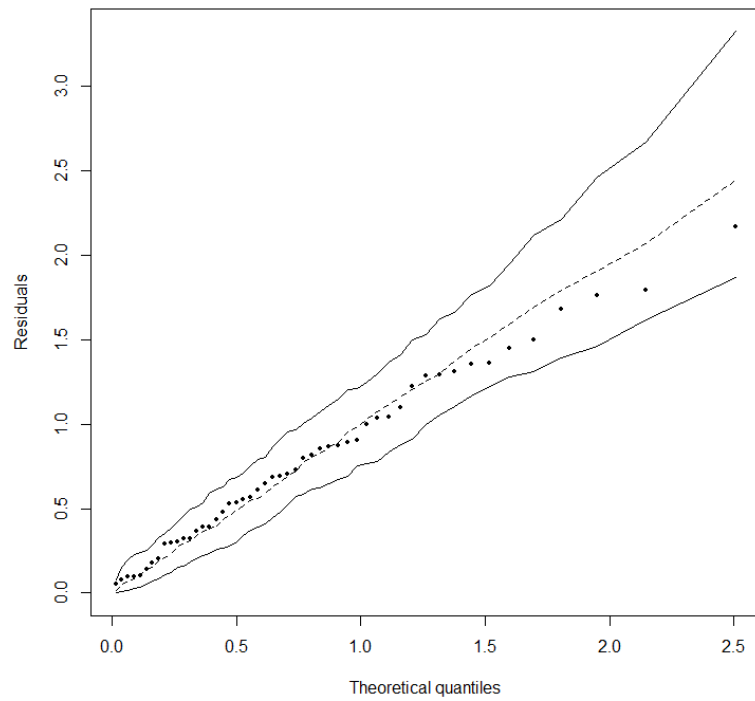
Στη συνέχεια εξετάζεται η προσαρµογή του µοντέλου µε βάση το διάγραµµα της Ηµι-κανονικής κατανοµής, το οποίο παρουσιάζεται στην Εικόνα 9. Καθώς, το σύνολο των σηµείων εµφανίζεται εντός του simulation envelope, τότε το µοντέλο έχει προσαρµοστεί αποτελεσµατικά στα δεδοµένα. Επιπλέον, η απουσία σηµείων µε µεγάλη διασπορά δηλώνει επίσης και την πιθανή απουσία outliers στα συγκεκριµένα δεδοµένα για το µοντέλο αυτό. Όµως για την ανίχνευση outliers θα εξεταστούν επιπρόσθετα τα hat values και οι αποστάσεις Cook της κάθε παρατήρησης.

Η συστηµατικότερη ανίχνευση πιθανών outliers ξεκινά µε τη µελέτη των hat values. Ειδικότερα, το κατώφλι ώστε µια παρατήρηση να θεωρηθεί πιθανό outlier είναι να ισχύει $h_{ii} > 2p/n = 2 * 4/51 =$

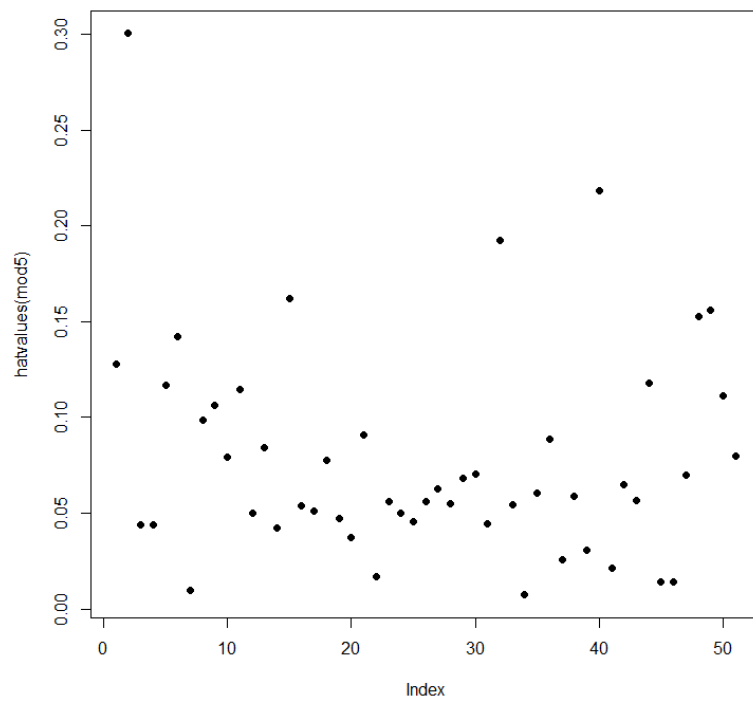


Εικόνα 8: Partial residual plots για το μοντέλο με τις μεταβλητές age, index, temperature

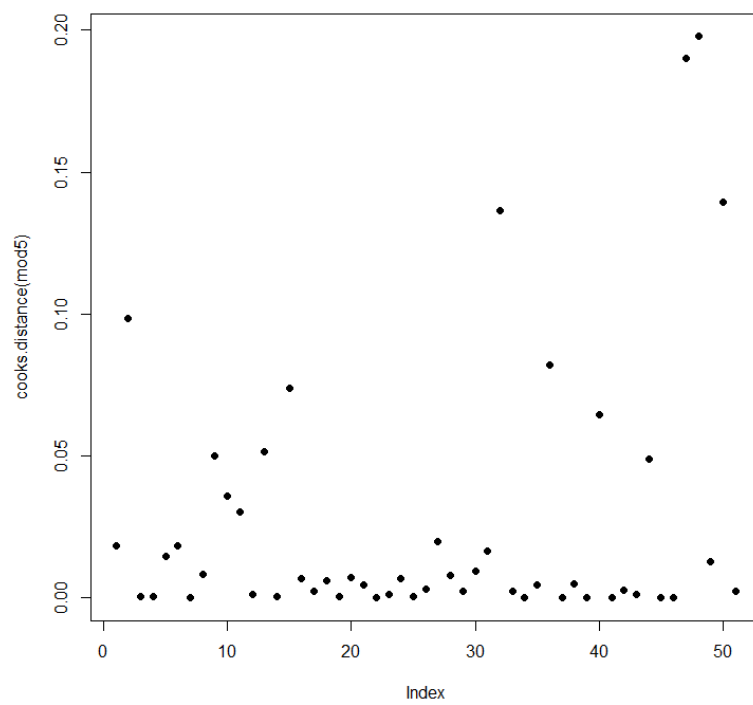
0.1568627. Για το συγκεκριμένο σέτ δεδομένων, σύμφωνα και με το Γράφημα 10, η συνθήκη αυτή ισχύει για τις παρατηρήσεις 2, 15, 32 και 40, τις οποίες θα θεωρήσουμε πιθανά outliers. Ωστόσο, αυτό δεν επιβεβαιώνεται από το διάγραμμα των αποστάσεων Cook καθώς όπως παρατηρείται στην Εικόνα 11, καμία από τις παρατηρήσεις δεν υπερβαίνει την τιμή 1. Συνεπώς, σύμφωνα με το μέτρο της απόστασης Cook, η αφαίρεση καμίας από τις παρατηρήσεις από το μοντέλο δεν θα επηρεάσει σημαντικά την εκτίμηση των παραμέτρων του. Τέλος, οι παρατηρήσεις επιβεβαιώνονται και από τα υπόλοιπα πιθανοφάνειας τα οποία παρουσιάζονται στην Εικόνα 12 και εκφράζουν την μεταβολή της deviance αν αφαιρεθεί από το μοντέλο η i -οστή παρατήρηση. Ειδικότερα οι περισσότερες παρατηρήσεις παρουσιάζουν υπόλοιπα στο εύρος $[-1.5, 1.5]$ με ομοιόμορφη κατανομή εντός αυτού του εύρους. Οι παρατηρήσεις που το ξεπερνούν ελάχιστα είναι οι 13, 32, 36, 47, 48, 50 αλλά η απόκλιση δεν είναι σημαντική όπως παρουσιάζεται και στο γράφημα. Συνεπώς, η αφαίρεση καμίας από τις παρατηρήσεις δεν θα μεταβάλει σημαντικά την Deviance και άρα δεν προκύπτει κάποιο ισχυρά σημαντικό σημείο επιρροής.



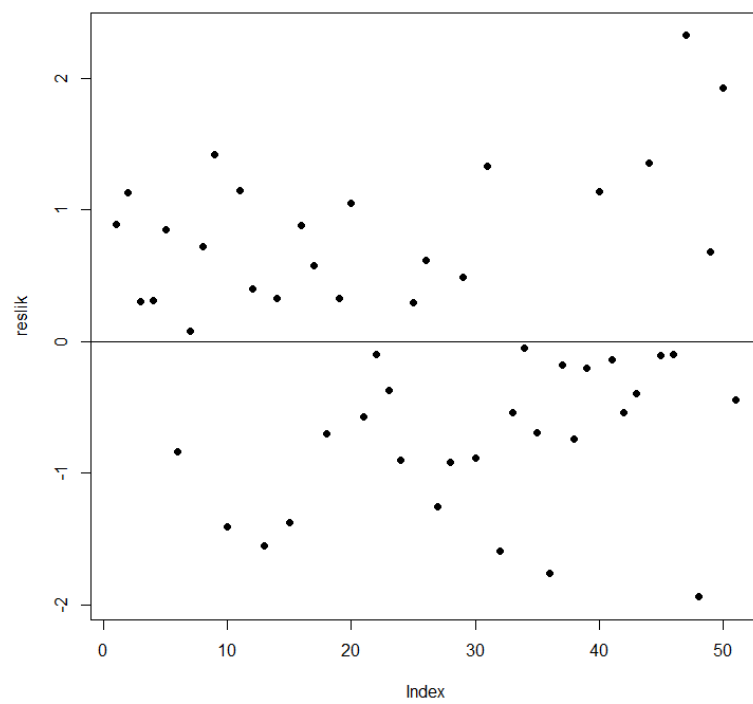
Εικόνα 9: Deviance Residuals με την ημι-κανονική κατανομή



Εικόνα 10: Hat values για το μοντέλο με τις μεταβλητές age, index, temperature



Εικόνα 11: Απόσταση Cook για το μοντέλο με τις μεταβλητές age, index, temperature



Εικόνα 12: Υπόλοιπα πιθανοφάνειας για το μοντέλο με τις μεταβλητές age, index, temperature

Ερώτημα 3

Για την κατασκευή ενός 95%, δηλαδή για $\alpha=0.05$, διαστήματος εμπιστοσύνης για τους συντελεστές του μοντέλου αξιοποιείται η σχέση:

$$\hat{\beta}_j \pm z_{\alpha/2} \text{se}(\hat{\beta}_j) \quad (8)$$

και καθώς αυτό που χρρίζει ερμηνείας σύμφωνα με τη σχέση 5, δεν είναι ο λογάριθμος αλλά η σχετική πιθανότητα ανταπόκρισης στη θεραπεία, τότε θα πρέπει το διάστημα που θα βρεθεί καθώς επίσης και οι εκτιμημένες τιμές των παραμέτρων να τοποθετηθούν εντός εκθετικού. Ο κώδικας παρουσιάζεται στο τέλος της εργασίας και τα αποτελέσματά του εμφανίζονται στον Πίνακα 4.

Παράμετρος	Τιμή	2.5%	97.5%
Πόλωση (Intercept)	8.956604e+37	5.889831e+07	1.362021e+68
age	0.9431767	0.8970617	0.9916623
index	1.469506	1.158070	1.864695
temperature	0.9148701	0.8524284	0.9818857

Πίνακας 4: εκτίμηση και 95% διαστήματα εμπιστοσύνης για τους εκθετικούς συντελεστές του μοντέλου λογιστικής παλινδρόμησης

Για τις ερμηνείες των συντελεστών προκύπτει ότι:

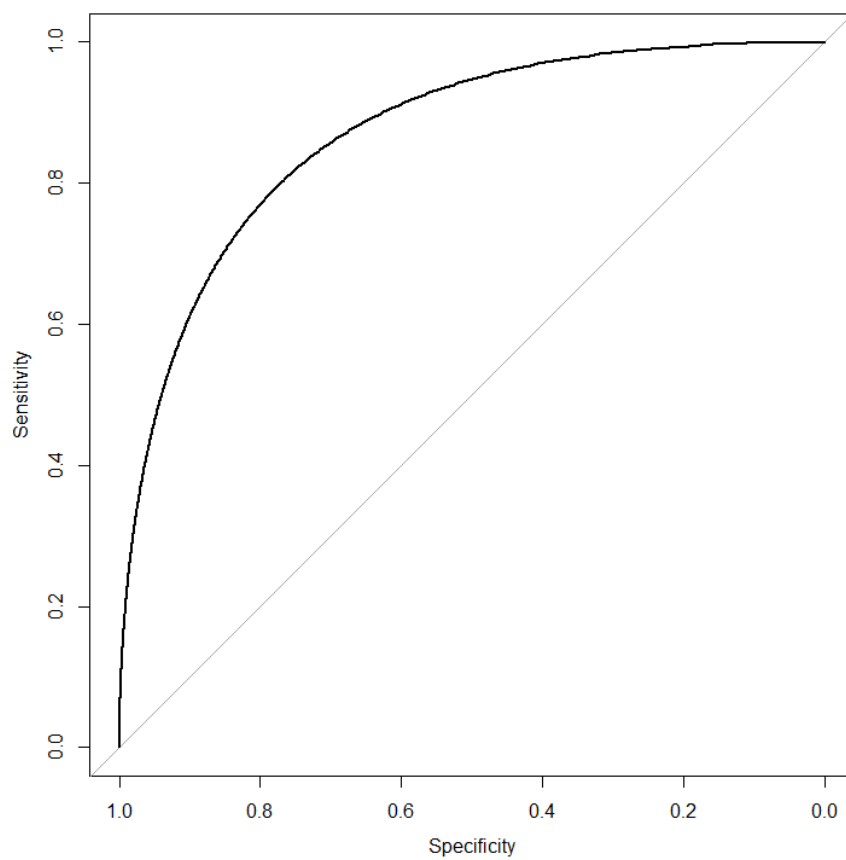
1. Εάν η ηλικία (age) του ασθενή αυξηθεί κατά μία μονάδα τότε η σχετική πιθανότητα ανταπόκρισης στη θεραπεία θα μειωθεί κατά $100|1-0.9431767| = 5.7\%$.
2. Εάν ο δείκτης κυττάρων λευχαιμίας (index) αυξηθεί κατά μία μονάδα, τότε η σχετική πιθανότητα ανταπόκρισης στη θεραπεία του ασθενούς θα αυξηθεί κατά $100|1-1.469506| = 46.9\%$.
3. Εάν η υψηλότερη θερμοκρασία του ασθενούς πριν τη θεραπεία (temperature) αυξηθεί κατά μία μονάδα, τότε η σχετική πιθανότητα ανταπόκρισης στη θεραπεία του ασθενούς θα μειωθεί κατά $100|1-0.9148701| = 8.5\%$.

Ερώτημα 4

Καθώς ο στόχος της λογιστικής παλινδρόμησης είναι να παράγει θετικές και αρνητικές προβλέψεις με βάση την κατάλληλα εκτιμημένη πιθανότητα επιτυχίας \hat{p} :

$$\hat{p} = \hat{P}(Y = 1) = \frac{e^{\mathbf{x}'\hat{\beta}}}{1 + e^{\mathbf{x}'\hat{\beta}}} \quad (9)$$

τότε θα πρέπει να οριστεί ένα κατώφλι p_0 για την πιθανότητα αυτή. Ειδικότερα για τιμές της $\hat{p} > p_0$, προβλέπεται ότι $Y = 1$ και αντίθετα προβλέπεται $Y = 0$. Η προβλεπτική ικανότητα του μοντέλου αξιολογείται με βάση τα μέτρα της ευαισθησίας, η οποία είναι το ποσοστό των αληθώς θετικών προβλέψεων και της ειδικότητας, δηλαδή το ποσοστό των αληθώς αρνητικών προβλέψεων. Ωστόσο, τα μέτρα αυτά θα πρέπει να υπολογιστούν για πολλαπλές τιμές κατωφλίου πιθανότητας p_0 και να προκύπτουν υψηλά για τα περισσότερα από αυτά, προκειμένου το μοντέλο να έχει τη δυνατότητα γενίκευσης και να είναι αποτελεσματικό. Το μέτρο που πιστοποιεί ότι πράγματι υπάρχουν αρκετά p_0 για τα οποία τα μέτρα ευαισθησίας και ειδικότητας είναι υψηλά είναι το εμβαδόν κάτω από την καμπύλη ROC (Receiver-Operator Curve) ή με συντομία AUC (Area Under the Curve). Όσο πιο κοντά στη μονάδα είναι το AUC, τόσο πιο αξιόπιστο είναι το μοντέλο. Όπως παρουσιάζεται στο Γράφημα 13, η καμπύλη ROC προσεγγίζει την άνω αριστερή γωνία του διαγράμματος, η οποία αντιστοιχεί στην μεγιστοποίηση των ανωτέρω μέτρων, χωρίς ωστόσο να την φτάνει. Ειδικότερα, το AUC ισούται με 0.8686, το οποίο είναι μια αρκετά υψηλή τιμή και συνεπώς υπάρχει μια ποικιλία κατωφλίων πιθανότητας για τα οποία το μοντέλο πραγματοποιεί αληθείς προβλέψεις.



Εικόνα 13: Καμπύλη ROC για το μοντέλο με τις μεταβλητές age, index, temperature.
Προκύπτει ότι $AUC = 0.8686$

Άσκηση 1 – Κωδικας στην R

```
cdat<-read.table("C:/Users/user/Desktop/dsml/statistical modeling/hw3/asfalies.txt", header=TRUE)
cdat
attach(cdat)

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

#wald test in summary
summary(mod1)
exp(coef(mod1))

#deviance test with saturated model
pvalue_sat = 1 - pchisq(mod1$deviance,mod1$df.residual)
pvalue_sat

#deviance test with null model
DDEV = mod1$null.deviance - mod1$deviance
df = mod1$df.null - mod1$df.residual
pvalue = 1 - pchisq(DDEV,df)
data.frame(DDEV, df, pvalue)

step(mod1,method="backward", test="Chisq")

confint.default(mod1)
exp(confint.default(mod1))

res.deviance<-residuals(mod1)
res.pearson<-residuals(mod1,type="pearson")
##
qqnorm(res.pearson,pch=19)
qqline(res.pearson)

res.deviance[res.deviance>2]
res.pearson[res.pearson>2]

plot(fitted.values(mod1),res.deviance,xlab='fitted values', ylab='Deviance residuals')
abline(h=0)

plot(fitted.values(mod1),res.pearson,xlab='fitted values', ylab='Pearson residuals')
abline(h=0)

reslik<-rstudent(mod1)

plot(hatvalues(mod1), reslik, pch=19)
abline(h=0)

plot(reslik, pch=19)
abline(h=0)

plot(cooks.distance(mod1), pch=19)
plot(hatvalues(mod1), pch=19)

hatvalues(mod1)[hatvalues(mod1)>0.375]
cooks.distance(mod1)[cooks.distance(mod1)>1]
reslik[c(4,8,12,16)]
reslik[abs(reslik) > 2.230321]
##
#αλληλεπίδρασεις
mod2<-glm(y~ factor(cartype) + factor(cartype)*agecat + agecat + district + offset(log(n)), family=poisson)
mod3<-glm(y~ factor(cartype) + factor(cartype)*district + agecat + district + offset(log(n)), family=poisson)
mod4<-glm(y~ factor(cartype)+ agecat*district + agecat + district + offset(log(n)), family=poisson)
summary(mod2)
summary(mod3)
summary(mod4)

exp(coef(mod4))
```



```

minor = cdat[-c(29),]#to remove a zero value of y- non infinity values in log
x1 <- minor$cartype
y <- minor$y
n <- minor$n
x2 <- minor$agecat
x3 <- minor$district

for (i in c(1,2,3,4)) {
  y1<-y[x1==i & x3==0]
  n1<-n[x1==i & x3==0]
  y2<-y[x1==i & x3==1]
  n2<-n[x1==i & x3==1]
  yy1 =log(y1/n1)
  yy2 =log(y2/n2)

  xx1<-x2[x1==i & x3==0]
  xx2<-x2[x1==i & x3==1]

  plot(xx1,yy1, main=sprintf(paste("Scatterplot of ln (y/n) vs agecat for cartype=", i)), xlab="agecat",
       ylab="ln(y/n)", col="blue",pch=19,xlim=c(min(xx1),max(xx1)), ylim=c(min(yy1),max(yy1)))
  abline(lm(yy1~xx1), col="blue")

  points(xx2,yy2,col="red",pch=19,xlim=c(min(xx2),max(xx2)), ylim=c(min(yy2),max(yy2)))
  abline(lm(yy2~xx2), col="red")

  legend("topright", c("district=1", "district=0"), col=c("blue","red"), lty=1:4)
}

```

Άσκηση 2 – Κωδικας στην R

```

cdat<-read.table("C:/Users/user/Desktop/dsml/statistical modeling/hw3/leukaemia.txt", header=TRUE)
cdat
attach(cdat)
library(car)

##
mod1 = glm(response ~ age + smear + infiltrate + index + blasts + temperature, family=binomial)
summary(mod1)

DDEV = mod1$null.deviance - mod1$deviance
df = mod1$df.null - mod1$df.residual
pvalue = 1 - pchisq(DDEV,df)
data.frame(DDEV, df, pvalue)

anova(mod1,test="Chisq")

##

mod2 = glm(response ~ 1, family=binomial)
mod3 = step(mod2,response ~ age + smear + infiltrate + index + blasts + temperature,method="forward", test="Chisq")
mod4 = step(mod1,method="backward", test="Chisq")
summary(mod4)
mod5 = glm(response ~ age + index + temperature, family=binomial)
summary(mod5)
AIC(mod5)

##
anova(mod4,mod5,test = "Chisq")
##
DDEV = mod5$null.deviance - mod5$deviance
df = mod5$df.null - mod5$df.residual
pvalue = 1 - pchisq(DDEV,df)
data.frame(DDEV, df, pvalue)

##

```

```
crPlots(mod5)

res.deviance<-residuals(mod5)
library(hnp)
hnp(res.deviance,pch=19)

reslik<-rstudent(mod5)
plot(reslik, pch=19)
abline(h=0)

plot(cooks.distance(mod5), pch=19)
plot(hatvalues(mod5), pch=19)
hatvalues(mod5)[hatvalues(mod5)>2*4/51]

###
confint.default(mod5)
exp(confint.default(mod5))
exp(coef(mod5))
##
library(stats)
roc(response, fitted.values(mod5), smooth=TRUE, plot=TRUE)
```