1 Παλινδρομηση Poisson

Τα δεδομένα της παρούσας άσκησης αφορούν το πλήθος αποζημιώσεων (Y) ανά n συμβόλαια λόγω τροχαίων ατυχημάτων. Τα υπό μελέτη χαρακτηριστικά είναι η ηλικία των ασφαλισμένων (`agecat`), η οποία δίνεται ως κατηγορική μεταβλητή βάσει της αντιστοιχίας 0 → νέος και 1 → μεγάλος, η περιοχή διαμονής του ασφαλισμένου (`district`), η οποία δίνεται επίσης ως κατηγορική μεταβλητή βάσει της αντιστοιχίας 1 → Αθήνα και 0 → άλλη περιοχή, καθώς και η κατηγορία ασφαλίστρων (`cartype`), με το πλήθος των κατηγοριών να είναι 4.

1.1 Αρχικά, τα δεδομένα φορτώνονται στο R Studio και κατόπιν η μεταβλητή `cartype` μετατρέπεται σε κατηγορική, χρησιμοποιώντας τις ακόλουθες εντολές στην R.

```
> library(data.table)
> data <- fread('asfalies.txt')
> data$cartype <- factor(data$cartype)
```

Στη συνέχεια, στα δεδομένα προσαρμόζεται ένα μοντέλο παλινδρόμησης Poisson, το οποίο συνοψίζεται ως εξής:

```
> mod <- glm(y ~ cartype + agecat + district + offset(log(n)),
                 data = data, family = 'poisson')
> summary(mod)

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

Deviance Residuals:
Min       1Q   Median       3Q      Max
-1.8590  -0.7506  -0.1297   0.6511   3.2310

Coefficients:
            Estimate  Std. Error z value Pr(>|z|)
(Intercept) -1.93522    0.05525 -35.030  < 2e-16 ***
cartype2     0.16223    0.05048   3.214 0.001309 **
cartype3     0.39535    0.05491   7.200 6.03e-13 ***
cartype4     0.56543    0.07215   7.836 4.64e-15 ***
agecat      -0.37628    0.04451  -8.453  < 2e-16 ***
district     0.21661    0.05853   3.701 0.000215 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘’  1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 207.833  on 31  degrees of freedom
Residual deviance:  41.789  on 26  degrees of freedom
AIC: 222.15

Number of Fisher Scoring iterations: 4
```

Από τα παραπάνω γίνεται εμφανές πως το προσαρμοσμένο μοντέλο είναι το

$$\hat{Y} = \exp\Big(0.162\,\text{cartype2} + 0.395\,\text{cartype3} + 0.565\,\text{cartype4} - 0.376\,\text{agecat} + 0.217\,\text{district} - 1.935\Big), \tag{1.1}$$

όπου τα νέα χαρακτηριστικά, `cartype2`, `cartype3` και `cartype4` προκύπτουν λόγω της μετατροπής της μεταβλητής `cartype` σε κατηγορική. Πριν προχωρήσει κανείς στην ερμηνεία του μοντέλου, πρέπει πρώτα να επαληθεύσει πως τα χαρακτηριστικά που υπεισέρχονται σε αυτό είναι όλα απαραίτητα. Ο πρώτος έλεγχος που πραγματοποιείται για το σκοπό αυτό είναι ο έλεγχος Wald, δηλαδή ο έλεγχος του λεγόμενου z-value, το οποίο ορίζεται ως

$$z_i = \frac{\hat{\beta}_i}{\text{se}\left(\hat{\beta}_i\right)}, \tag{1.2}$$

όπου $\hat{\beta}$ οι τιμές των συντελεστών των χαρακτηριστικών του προσαρμοσμένου μοντέλου. Η z-value κάθε χαρακτηριστικού υπολογίζεται αυτόματα από την R και φαίνεται στην προηγούμενη σύνοψη. Γίνεται εμφανές πως για όλα σχεδόν τα χαρακτηριστικά η z-value είναι μικρότερη του αυστηρότερου ορίου (0.001), καθιστώντας τα έτσι απαραίτητα για το μοντέλο. Μοναδική εξαίρεση αποτελεί το `cartype2`, για το οποίο η z-value είναι μόλις 0.0003 πάνω από το αυστηρότερο όριο. Το γεγονός αυτό από μόνο του φυσικά δεν αρκεί για την απόρριψη του χαρακτηριστικού από το μοντέλο, επιβεβαιώνει όμως την ανάγκη περαιτέρω ελέγχων.

Ο επόμενος έλεγχος σχετικά με τη σημαντικότητα των χαρακτηριστικών βασίζεται στο κριτήριο AIC. Συγκεκριμένα, η τιμή του AIC για το μοντέλο της Σχέσης (1.1) δίνεται στην προηγούμενη σύνοψη ίση με 222.15. Εάν όλα τα χαρακτηριστικά είναι απαραίτητα για το μοντέλο, αυτή θα είναι και η ελάχιστη δυνατή τιμή του AIC. Σε διαφορετική περίπτωση, εάν δηλαδή η αφαίρεση κάποιου χαρακτηριστικού οδηγεί σε χαμηλότερη τιμή για το AIC, τότε το συγκεκριμένο χαρακτηριστικό μπορεί να θεωρηθεί ως περιττό. Στα πλαίσια αυτά εφαρμόζεται η διαδικασία διαδοχικής αφαίρεσης (backward elimination), με τις σχετικές εντολές στην R και τα αποτελέσματα να φαίνονται παρακάτω. Σημειώνεται πως η παράμετρος `test="Chisq"` εισάγεται προκειμένου, αφότου αναλυθεί το κριτήριο AIC, να γίνει αναφορά στο στατιστικό έλεγχο Deviance.

```
> 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 = data)

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
```

Προκύπτει πως η αφαίρεση οποιουδήποτε χαρακτηριστικού από το μοντέλο οδηγεί σε αύξηση του AIC, γεγονός που επαληθεύει τη σημαντικότητα των χαρακτηριστικών για το μοντέλο. Αξίζει εδώ να αναφερθεί πως το ίδιο ακριβώς συμπέρασμα προκύπτει και μέσω της διαδικασίας διαδοχικής πρόσθεσης (forward selection), η οποία αντιστοιχεί στην εντολή

```
> step(glm(data$y ~ 1 + offset(log(data$n)), family="poisson"),
       direction = 'forward', scope =(~ data$cartype + data$agecat
                                        + data$district))
```

καθώς και μέσω της κατά βήματα εμπρός-πίσω επιλογής (stepwise elimination), η οποία αντιστοιχεί στην εντολή

```
> step(mod, direction = 'both')
```

Οι προκαταρκτικοί έλεγχοι ως προς τη σημαντικότητα των χαρακτηριστικών στο μοντέλο της Σχέσης (1.1) ολοκληρώνονται με το στατιστικό έλεγχο Deviance, που είναι και ο λόγος για τον οποίο η παράμετρος `test="Chisq"` προστέθηκε στην προηγούμενη διαδικασία διαδοχικής αφαίρεσης. Όπως αναφέρθηκε και για το κριτήριο AIC, ο έλεγχος Deviance πραγματοποιείται προκειμένου να εξεταστεί το ενδεχόμενο κάποιο μοντέλο εμφωλευμένο στο προκύπτον μοντέλο να είναι καλύτερο. Κρίνοντας από τις τιμές της στήλης `Pr(>Chi)` των παραπάνω αποτελεσμάτων, κάτι τέτοιο δε συμβαίνει, γεγονός που - για ακόμη μια φορά - υποδεικνύει πως όλα τα χαρακτηριστικά που έχουν ληφθεί υπ' όψιν για την κατασκευή του μοντέλου είναι στατιστικά σημαντικά. Βάσει των συμπερασμάτων αυτών, μπορεί κανείς να προχωρήσει στην ερμηνεία του μοντέλου που περιγράφει η Σχέση (1.1).

Σε ό,τι αφορά τα χαρακτηριστικά `cartype2`, `cartype3` και `cartype4`, υπενθυμίζεται πως αυτά προκύπτουν από τη μετατροπή της μεταβλητής `cartype` σε κατηγορική. Το χαρακτηριστικό `cartype1` δεν υπεισέρχεται ρητά στο μοντέλο, καθώς ο συντελεστής του είναι ταυτοτικά ίσος με το μηδέν, αφού αντιστοιχεί στη default τιμή της `cartype`. Έτσι, οι συντελεστές των `cartypeX`, με `X` = 2,3,4, αντιστοιχούν σε ένα μέτρο της μεταβολής της αποζημίωσης εάν κανείς μεταβεί από την κατηγορία 1 στην κατηγορία `X`. Συγκεκριμένα, και οι τρεις συντελεστές, $\hat{\beta}_2$, $\hat{\beta}_3$, $\hat{\beta}_4$ είναι θετικοί και μάλιστα ισχύει πως $\hat{\beta}_2 < \hat{\beta}_3 < \hat{\beta}_4$, γεγονός που υποδεικνύει πως η αποζημίωση αυξάνεται καθώς αυξάνεται το `X` που αντιστοιχεί στην κατηγορία ασφαλίστρων. Η μετάβαση από την κατηγορία 1 στην κατηγορία 2 οδηγεί σε μέση αύξηση 17.58% για την αποζημίωση, η μετάβαση από την κατηγορία 1 στην κατηγορία 3 οδηγεί σε μέση αύξηση 48.48% για την αποζημίωση, ενώ η μετάβαση από την κατηγορία 1 στην κατηγορία 4 οδηγεί σε μέση αύξηση 75.94% για την αποζημίωση (κατ' αντιστοιχία μπορεί να υπολογιστεί και η σχετική αύξηση για άλλες μεταβάσεις). Σχετικά με το συντελεστή του χαρακτηριστικού `agecat`, προκύπτει πως η αύξηση της συμμεταβλητής αυτής κατά μία μονάδα πολλαπλασιάζει την αποζημίωση κατά έναν παράγοντα $\exp(-0.376) \simeq 0.6864$. Με άλλα λόγια, ένας μεγάλος σε ηλικία άνθρωπος αναμένεται να λάβει μειωμένη αποζημίωση σε σχέση με έναν νέο, με μέση μείωση 31.36%. Τέλος, ο συντελεστής του χαρακτηριστικού `district` υποδεικνύει πως η αύξηση της συμμεταβλητής αυτής κατά μία μονάδα οδηγεί σε έναν πολλαπλασιαστικό παράγοντα $\exp(0.217) \simeq 1.2423$, πράγμα το οποίο σημαίνει πως στην Αθήνα οι αποζημιώσεις είναι υψηλότερες, με τη μέση αύξηση να ισούται με 24.23%.

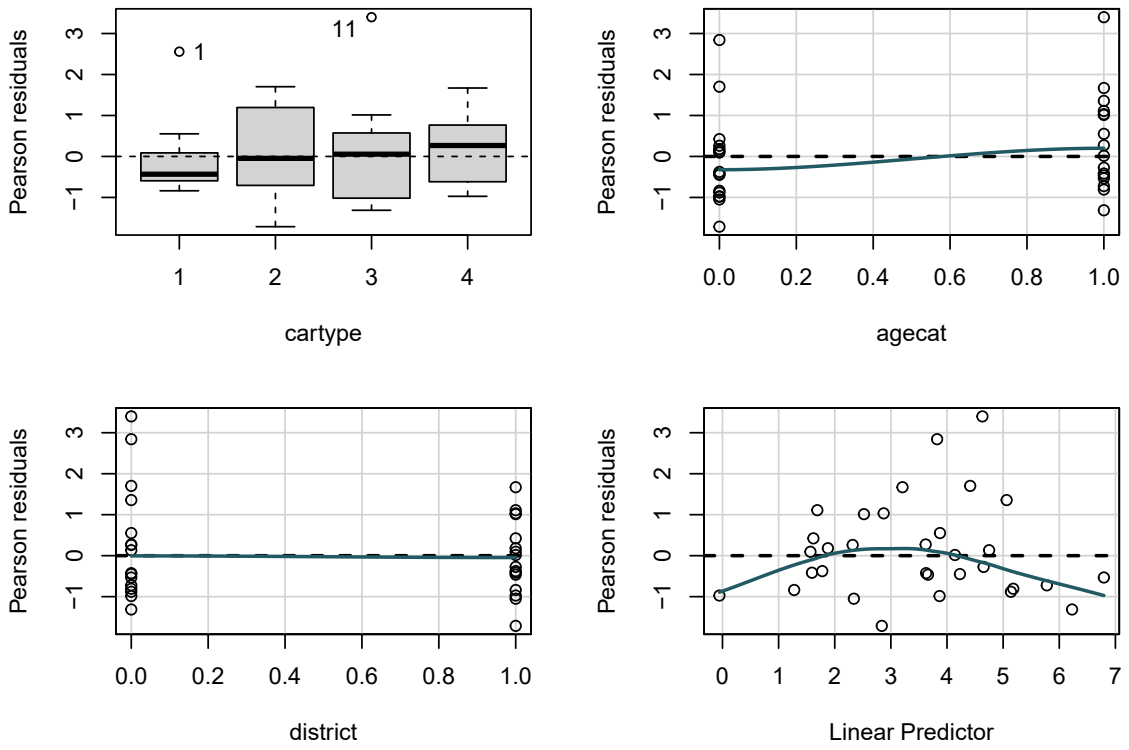
1.2 Παρότι τα αποτελέσματα αυτά επαρκούν για την εξαγωγή κάποιων βασικών συμπερασμάτων, με το κυριότερο να είναι πως η κατηγορία ασφαλίστρων 4 για έναν νέο οδηγό στην Αθήνα αποφέρει τη μέγιστη αποζημίωση, δεν αποτελούν παρά μέσες τιμές για τις αντίστοιχες μεταβολές, συνεπώς μια ακριβέστερη ποσοτικοποίησή τους κρίνεται απαραίτητη. Για το σκοπό αυτό, υπολογίζονται τα διαστήματα εμπιστοσύνης 95% για τους συντελεστές της Σχέσης (1.1), από τα οποία μπορεί να εξαχθεί το πλήρες εύρος των προαναφερθέντων μεταβολών και όχι μόνο η μέση τιμή τους. Οι σχετικές εντολές στην R μαζί με τα αποτελέσματά τους φαίνονται 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
```

Σε ό,τι αφορά τα διαστήματα εμπιστοσύνης των συντελεστών, η βασική παρατήρηση είναι πως η τιμή μηδέν δεν περιέχεται στο διάστημα κανενός εξ αυτών, γεγονός το οποίο συνάδει με το προηγουμένως εξαχθέν συμπέρασμα σχετικά με τη σημαντικότητα όλων των χαρακτηριστικών για το μοντέλο. Από την άλλη, για τις μεταβολές των παραμέτρων του μοντέλου οι ποιοτικές προβλέψεις που έγιναν παραπάνω επαληθεύονται, ενώ ποσοτικά μπορεί κανείς να είναι πιο ακριβής, λέγοντας πως:

1. Η κατηγορία ασφαλίστρων 2 οδηγεί σε **υψηλότερη** αποζημίωση σε σχέση με την κατηγορία 1 κατά 6.53% έως 29.84%.
2. Η κατηγορία ασφαλίστρων 3 οδηγεί σε **υψηλότερη** αποζημίωση σε σχέση με την κατηγορία 1 κατά 33.34% έως 65.36%.
3. Η κατηγορία ασφαλίστρων 4 οδηγεί σε **υψηλότερη** αποζημίωση σε σχέση με την κατηγορία 1 κατά 52.81% έως 102.76%.
4. Ένας μεγάλος οδηγός αναμένεται να λάβει **χαμηλότερη** αποζημίωση σε σχέση με έναν νέο οδηγό κατά 25.10% έως 37.09%.
5. Η ασφάλιση στην Αθήνα οδηγεί σε **υψηλότερη** αποζημίωση σε σχέση με ασφάλιση εκτός Αθήνας κατά 10.73% έως 39.28%.

1.3 Το επόμενο βήμα της ανάλυσης αποτελεί η μελέτη ορισμένων γραφημάτων, ξεκινώντας από τα διαγράμματα για τα υπόλοιπα Pearson κάθε χαρακτηριστικού, τα οποία προκύπτουν μέσω της εντολής `residualPlots(mod,type='pearson')` στην R, δεδομένης της βιβλιοθήκης `car`, και απεικονίζονται στην Εικόνα 1.1.

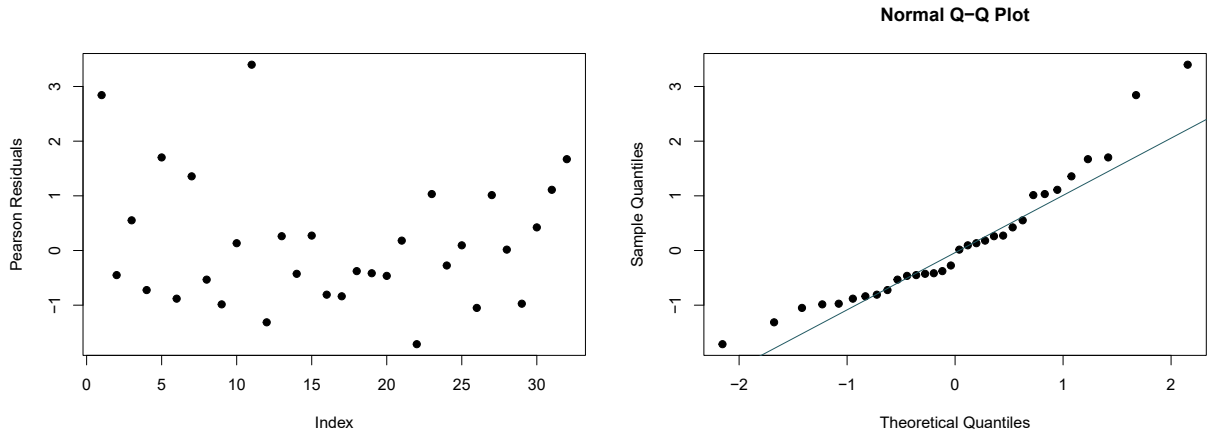


Εικόνα 1.1: Διαγράμματα των υπολοίπων Pearson ανά χαρακτηριστικό.

Με εξαίρεση τα σημεία με δείκτη 1 και 11, τα οποία φαίνεται να αντιστοιχούν σε outliers, δε φαίνεται να υπάρχει κάποιος συστηματικός παράγοντας. Για πληρότητα, μέσω των εντολών

```
> res.pearson <- residuals(mod,type='pearson')

> par(mfrow=c(1,2))
> plot(res.pearson,xlab='Index', ylab='Pearson Residuals',pch=19)
> qqnorm(res.pearson, pch=19)
> qqline(res.pearson)
```

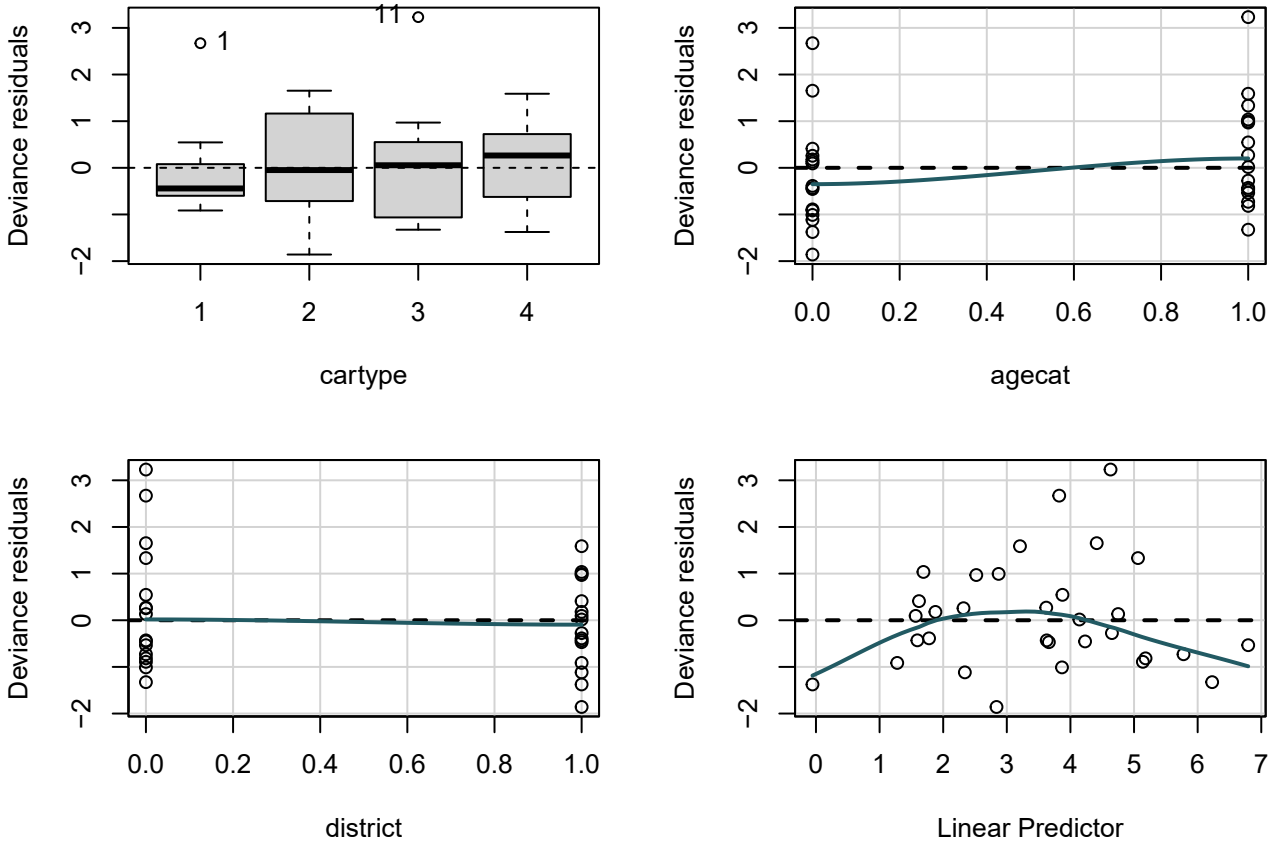
στην R, κατασκευάζονται και τα index plot και QQ-plot που αφορούν τα υπόλοιπα Pearson και απεικονίζονται στην Εικόνα 1.2.



Εικόνα 1.2: Index plot και QQ-plot για τα υπόλοιπα Pearson.

Και στην περίπτωση αυτή, η ιδιαίτερη συμπεριφορά των σημείων 1 και 11 γίνεται εμφανής, αφού τα υπόλοιπα που αντιστοιχούν σε αυτά φαίνεται να λαμβάνουν τιμές ~ 3 και να αποκλίνουν αρκετά από το προσαρμοσμένο μοντέλο. Σχετικά με το index plot των υπολοίπων, αυτά φαίνεται να έχουν αρκετά τυχαία κατανομή, γεγονός που ενισχύει το συμπέρασμα πως το μοντέλο δεν παρουσιάζει κάποια εμφανή συστηματικότητα. Τέλος, σε ό,τι αφορά την προσαρμογή του μοντέλου που αποτυπώνεται στο qq-plot, αυτή φαίνεται ικανοποιητική για τα περισσότερα σημεία, με ορισμένα όμως να εμφανίζουν σημαντικές αποκλίσεις.

Η ίδια διαδικασία επαναλαμβάνεται και στην περίπτωση των υπολοίπων Deviance, ξεκινώντας με την εντολή `residualPlots(mod,type='deviance')` στην R για τα διαγράμματά τους ανά χαρακτηριστικό, τα οποία απεικονίζονται στην Εικόνα 1.3.

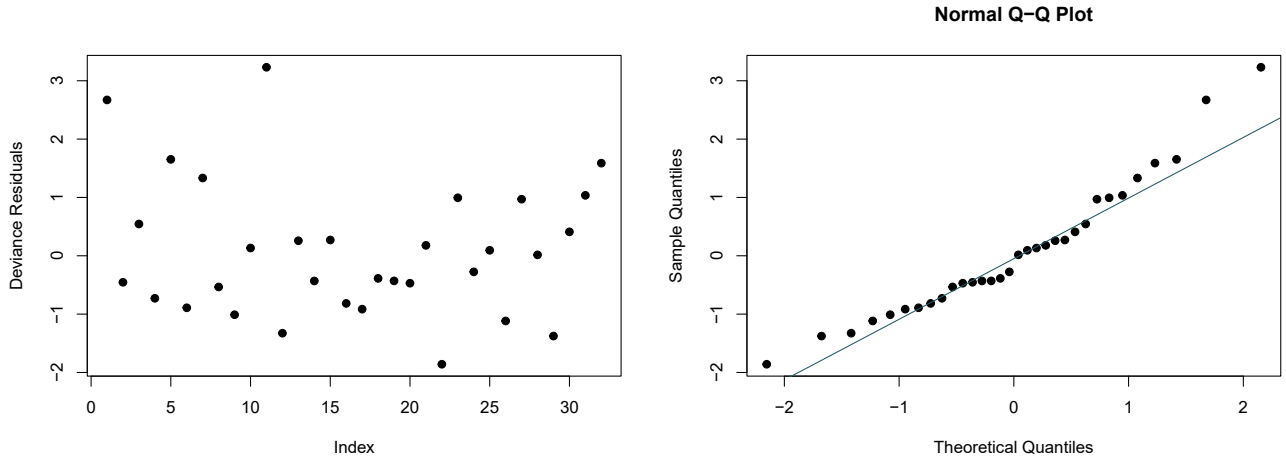


Εικόνα 1.3: Διαγράμματα των υπολοίπων Deviance ανά χαρακτηριστικό.

Η ομοιότητα με τα διαγράμματα της Εικόνας 1.1 είναι ολοφάνερη, γεγονός που σημαίνει πως τα προηγούμενα σχόλια αντιστοιχούν και στην περίπτωση των υπολοίπων Deviance. Σε ό,τι αφορά το αντίστοιχο index plot και QQ-plot, οι εντολές στην R είναι οι

```
> res.deviance <- residuals(mod,type='deviance')

> par(mfrow=c(1,2))
> plot(res.deviance,xlab='Index', ylab='Deviance␣Residuals',pch=19)
> qqnorm(res.deviance, pch=19)
> qqline(res.deviance)
```

και τα αντίστοιχα γραφήματα παρατίθενται στην Εικόνα 1.4. Κατ' αντιστοιχία με τα διαγράμματα ανά χαρακτηριστικό, και αυτά τα γραφήματα είναι πρακτικά πανομοιότυπα με αυτά της Εικόνας 1.2.

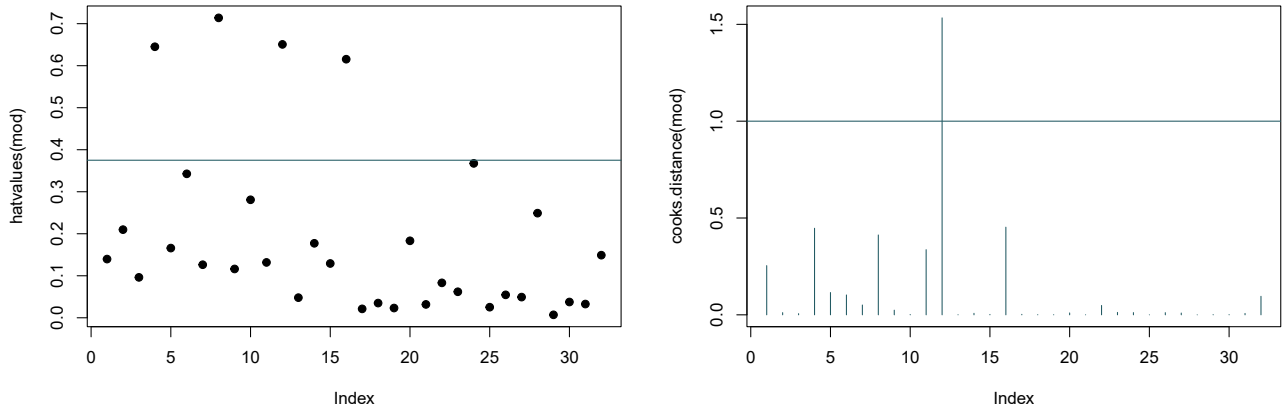


Εικόνα 1.4: Index plot και QQ-plot για τα υπόλοιπα Deviance.

Προχωρώντας, μέσω της εντολής

```
> par(mfrow=c(1,2))

> plot(hatvalues(mod), pch=19)
> plot(cooks.distance(mod),pch=19)
```

στην R, δημιουργούνται τα γραφήματα για τα λεγόμενα hat values, δηλαδή τα διαγώνια στοιχεία του πίνακα H, καθώς και για τις αποστάσεις Cook και στη συνέχεια, απεικονίζονται στην Εικόνα 1.5.



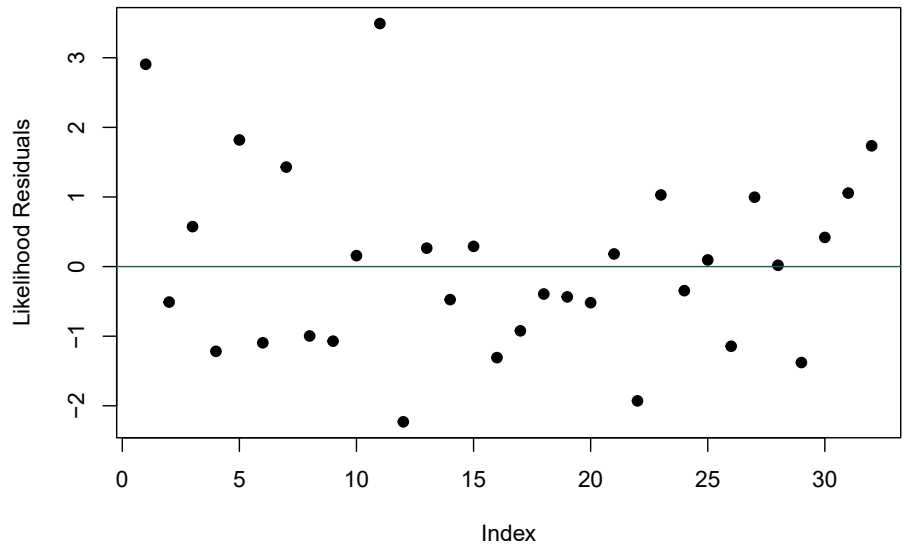
Εικόνα 1.5: Index plots για τα hat values (h_{ii}) και τις αποστάσεις Cook.

Όπως είναι γνωστό, τα γραφήματα αυτά υποδεικνύουν τα σημεία επιρροής του μοντέλου. Σε ό,τι αφορά τα hat values, το σχετικό κατώφλι ορίζεται ως $2p/n$, με p το πλήθος των χαρακτηριστικών του μοντέλου και n το σύνολο των δειγμάτων, ενώ για την απόσταση Cook το σχετικό κατώφλι θεωρείται ίσο με τη μονάδα. Σε καθεμιά από τις δύο περιπτώσεις, το σημείο με δείκτη 12 φαίνεται να αποτελεί σημείο επιρροής, καθώς απέχει (σημαντικά μάλιστα) από το αντίστοιχο κατώφλι. Σε ό,τι αφορά τα hat values, υπάρχουν επίσης ενδείξεις για τα σημεία με δείκτες 4, 8 και 16 ως πιθανά σημεία επιρροής, ενώ τα σημεία με δείκτες 6 και 24 είναι μόνο οριακά κάτω από το κατώφλι.

Κλείνοντας την ανάλυση των γραφικών παραστάσεων, παρατίθεται το index plot και για τα υπόλοιπα πιθανοφάνειας, το οποίο δημιουργείται στην R μέσω των εντολών

```
> reslik <- rstudent(mod)

> plot(reslik, xlab='Index', ylab='Likelihood Residuals', pch=19)
> abline(h=0)
```

και απεικονίζεται στην Εικόνα 1.6.



Εικόνα 1.6: Index plot για τα υπόλοιπα πιθανοφάνειας.

Το διάγραμμα αυτό δεν παρουσιάζει οποιαδήποτε αισθητή διαφορά σε σχέση με τα index plots των Εικόνων 1.2 (για τα υπόλοιπα Pearson) και 1.4 (για τα υπόλοιπα Deviance).

1.4 Με βάση την έως τώρα ανάλυση, μπορεί κανείς με ασφάλεια να ισχυριστεί πως τα χαρακτηριστικά που έχουν επιλεχθεί για την κατασκευή του μοντέλου της Σχέσης (1.1) είναι όλα σημαντικά, χωρίς όμως αυτό να σημαίνει πως το μοντέλο είναι και το βέλτιστο. Οι προηγούμενοι γραφικοί έλεγχοι υπέδειξαν πως η προσαρμογή των δεδομένων σε αυτό δεν είναι ιδανική, κάτι το οποίο ενδέχεται να οφείλεται σε συγκεκριμένα outliers (σημεία με δείκτες 1 και 11) ή γενικά σημεία υψηλής επιρροής (π.χ. τα σημεία με δείκτες 4, 8 ή 12). Η σημαντικότητα του μοντέλου συνολικά μπορεί να καθοριστεί μέσω της p-value που προκύπτει από τη σύγκριση του μοντέλου με το αντίστοιχο κορεσμένο μοντέλο, δηλαδή το μοντέλο το οποίο προσαρμόζεται βέλτιστα στα δεδομένα, έχοντας πλήθος χαρακτηριστικών ίσο με το πλήθος των δεδομένων. Η σχετική εντολή στην R μαζί με το αποτέλεσμά της φαίνονται παρακάτω.

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

[1] 0.02580847
```

Η προκύπτουσα p-value υποδεικνύει πως η προσαρμογή του μοντέλου είναι έχει σημαντική στατιστική διαφορά από το κορεσμένο μοντέλο και ως εκ τούτου αυτό δε μπορεί να θεωρηθεί ικανοποιητικό [1]. Το γεγονός αυτό μπορεί σίγουρα να αποδοθεί μέχρι ένα βαθμό στο μικρό πλήθος των διαθέσιμων δεδομένων, καθώς τα outliers και τα σημεία επιρροής επηρεάζουν ευκολότερα ένα σύνολο με λίγα δεδομένα.

Ένας τρόπος να ελεγχθεί το κατά πόσο το μοντέλο επιδέχεται βελτιώσεων με τα υπάρχοντα δεδομένα και χωρίς την αφαίρεση των outliers ή των σημείων επιρροής είναι η εισαγωγή αλληλεπιδράσεων μεταξύ των διαφόρων χαρακτηριστικών και ο έλεγχος του κατά πόσο αυτές είναι

σημαντικές ή όχι, από την αντίστοιχη z-value. Αρχικά, στο μοντέλο εισάγεται η αλληλεπίδραση μεταξύ των χαρακτηριστικών `agecat` και `cartype`, μέσω ενός απλού πολλαπλασιασμού. Η σχετική εντολή στην R με τα αποτελέσματά της φαίνονται παρακάτω.

```
> mod2a <- glm(y ~ cartype + agecat + district + agecat*cartype
               + offset(log(n)), data=data, family='poisson')
> summary(mod2a)

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)     -1.83560    0.08555 -21.456  < 2e-16 ***
cartype2         0.05293    0.10464   0.506  0.61300
cartype3         0.24660    0.11481   2.148  0.03173 *
cartype4         0.43213    0.16289   2.653  0.00798 **
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
```

Τόσο τα z-values όσο και η τιμή του AIC (το οποίο αυξήθηκε ελαφρώς) υποδεικνύουν πως η συγκεκριμένη αλληλεπίδραση όχι μόνο δε βελτιώνει το μοντέλο, αλλά μειώνει τη σημαντικότητα άλλων χαρακτηριστικών. Η επόμενη αλληλεπίδραση που δοκιμάζεται είναι η αλληλεπίδραση μεταξύ των χαρακτηριστικών `district` και `cartype`:

```
> mod2b <- glm(y ~ cartype + agecat + district + district*cartype
               + offset(log(n)), data=data, family='poisson')
> summary(mod2b)

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)       -1.92280    0.05648 -34.042  < 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
```

Τα αποτελέσματα είναι καλύτερα από πριν, όμως η τιμή του AIC παραμένει αυξημένη, ενώ ακόμα και η καλύτερη αλληλεπίδραση, δηλαδή αυτή μεταξύ `district` και `cartype4`, δε δίνει αρκετά χαμηλό z-value προκειμένου να χαρακτηριστεί ως σημαντική. Η τελευταία αλληλεπίδραση που δοκιμάζεται είναι αυτή μεταξύ των χαρακτηριστικών `district` και `agecat`:

```
> mod2c <- glm(y ~ cartype + agecat + district + district*agecat
               + offset(log(n)), data=data, family='poisson')
> summary(mod2c)

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)     -1.91460    0.05599 -34.196  < 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        -0.06127    0.15970  -0.384  0.70125
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
```

Ξανά, η αλληλεπίδραση δεν είναι αρκετά σημαντική βάσει του z-value της, παρ' όλα αυτά η τιμή του AIC είναι ελαφρώς μειωμένη σε σχέση με αυτήν που προέκυπτε από το μοντέλο της Σχέσης (1.1). Δυστυχώς, η προσαρμογή θα μπορέσει να γίνει με πιο ικανοποιητικό τρόπο είτε μέσω αύξησης του πλήθους των δεδομένων (και ενδεχομένως την απομάκρυνση ορισμένων), είτε μέσω της προσθήκης επιπλέον χαρακτηριστικών χωρίς κοινή πληροφορία με τα ήδη υπάρχοντα, με την ελπίδα το μοντέλο να πλησιάζει το αντίστοιχο κορεσμένο λίγο περισσότερο.

2 Λογιστικη Παλινδρομηση

Τα δεδομένα της άσκησης αυτής αφορούν το κατά πόσο ασθενείς που πάσχουν από λευχαιμία ανταποκρίνονται σε θεραπεία (`response` = 1) ή όχι (`response` = 0). Τα υπό μελέτη χαρακτηριστικά (συμμεταβλητές) είναι η ηλικία του ασθενή (`age`), το ποσοστό επίστρωσης βλαστοκυττάρων (`smear`), το ποσοστό κυττάρων στο μυελό των οστών (`infiltrate`), ο δείκτης των κυττάρων λευχαιμίας (`index`), τα βλαστοκύτταρα (`blasts`), καθώς και η υψηλότερη θερμοκρασία πριν τη θερμοκρασίας (`temperature` - $\times 10^{o}$F).

2.1 Αφότου τα δεδομένα φορτωθούν στο R Studio, προσαρμόζεται σε αυτά ένα μοντέλο λογιστικής παλινδρόμησης, το οποίο θα ελεγχθεί με βάση τους στατιστικούς ελέγχους που αξιοποιήθηκαν και στην προηγούμενη άσκηση. Οι σχετικές εντολές στην R μαζί με τα αποτελέσματά τους φαίνονται παρακάτω:

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

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

Deviance Residuals:
Min        1Q    Median        3Q       Max
-1.73878  -0.58099  -0.05505   0.62618   2.28425

Coefficients:
            Estimate  Std. Error z value Pr(>|z|)
(Intercept) 98.52361   40.85385   2.412  0.01588 *
age         -0.06029    0.02729  -2.210  0.02714 *
smear       -0.00480    0.04108  -0.117  0.90698
infiltrate   0.03621    0.03934   0.921  0.35728
index        0.39845    0.13278   3.001  0.00269 **
blasts       0.01343    0.05782   0.232  0.81627
temperature -0.10223    0.04181  -2.445  0.01448 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '  1

(Dispersion parameter for binomial family taken to be 1)
```

```
Null deviance: 70.524  on 50  degrees of freedom
Residual deviance: 40.060  on 44  degrees of freedom
AIC: 54.06

Number of Fisher Scoring iterations: 6
```

Σε ό,τι αφορά το κριτήριο Wald, δηλαδή τον έλεγχο του z-value, φαίνεται πως το μοναδικό χαρακτηριστικό με z-value που αντιστοιχεί σε πιθανότητα μικρότερη του 0.01 είναι το `index`. Τα χαρακτηριστικά `age` και `temperature` φαίνεται να έχουν κάποια σημαντικότητα, αν και μικρότερη σε σχέση με το `index`, ενώ για τα `smear`, `infiltrate` και `blasts` τα αντίστοιχα z-value αντιστοιχούν σε πιθανότητες μεγαλύτερες (κατά πολύ) του 0.05. Έτσι, ο στατιστικός έλεγχος Wald υποδεικνύει πως από τα 6 χαρακτηριστικά, μόνο τα `age`, `temperature` και `index` έχουν σχετικά υψηλή στατιστική σημασία στο συγκεκριμένο μοντέλο, υποδεικνύοντας πως πρέπει να πραγματοποιηθεί τροποποίησή του.

Πριν κανείς προχωρήσει σε μια τέτοια τροποποίηση, οφείλει να πραγματοποιήσει περαιτέρω ελέγχους. Για το σκοπό αυτό, μέσω της εντολής `anova(mod,test="Chisq")` πραγματοποιείται ο έλεγχος Deviance, τα αποτελέσματα του οποίου είναι τα ακόλουθα:

```
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
```

Ο έλεγχος Deviance επαληθεύει τα πορίσματα που προέκυψαν από τον έλεγχο Wald: οι μεταβλητές `smear`, `infiltrate` και `blasts` φαίνεται να μην είναι στατιστικά σημαντικές για το συγκεκριμένο μοντέλο. Έτσι, κρίνεται απαραίτητη η τροποποίησή του, η οποία πραγματοποιείται μέσω της διαδικασίας διαδοχικής αφαίρεσης, βάσει του κριτηρίου AIC. Η σχετική εντολή σε R μαζί με τα αποτελέσματά της φαίνεται παρακάτω.

```
> step(mod, 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", data = data)

Coefficients:
(Intercept)      age      infiltrate      index      temperature
95.56766     -0.06026      0.03413      0.40673     -0.09944

Degrees of Freedom: 50 Total (i.e. Null);  46 Residual
Null Deviance:      70.52
Residual Deviance: 40.14        AIC: 50.14
```

Η διαδικασία διαδοχικής αφαίρεσης οδηγεί σε ένα μοντέλο 2 λιγότερων συμμεταβλητών με AIC μικρότερο από αυτό του αρχικού μοντέλου (50.14 έναντι της τιμής 54.06). Παρ' όλα αυτά, ο στατιστικός έλεγχος Deviance (οι σχετικές ενδείξεις φαίνονται στα αποτελέσματα της διαδικασίας διαδοχικής αφαίρεσης ως `Pr(>Chi)`, χάρη στην παράμετρο `test="Chisq"`) εξακολουθεί να απορρίπτει τη συμμεταβλητή `infiltrate` ως μη στατιστικά σημαντική. Στο ίδιο ακριβώς συμπέρασμα καταλήγει και ο επανέλεγχος Wald, ο οποίος, μέσω των εντολών

```
> mod2 <- glm(response ~ age+index+temperature+infiltrate,
              data=data, family='binomial')
> summary(mod2)
```

στην R, αποδίδει τώρα στο χαρακτηριστικό `infiltrate` z-value τέτοιο, ώστε η αντίστοιχη πιθανότητα να ισούται με 0.1. Δεδομένης της διαφωνίας αυτής μεταξύ των στατιστικών ελέγχων Wald και Deviance και του κριτηρίου AIC, τον τελευταίο λόγο σε ό,τι αφορά τη συμμεταβλητή `infiltrate` θα έχει η σύγκριση του μοντέλου με το αντίστοιχο κορεσμένο, στην περίπτωση όπου η `infiltrate` περιλαμβάνεται σε αυτό (mod2), καθώς και στην περίπτωση όπου η `infiltrate` δεν περιλαμβάνεται σε αυτό (mod3):

```
> mod3 <- glm(response ~ age+index+temperature,
data=data, family='binomial')
```

Έτσι, το μοντέλο που βρίσκεται πιο κοντά στο κορεσμένο θα είναι αυτό το οποίο θα επιλεγεί τελικά για περαιτέρω ανάλυση. Οι σχετικές εντολές στην R μαζί με τα αποτελέσματά τους φαίνονται παρακάτω.

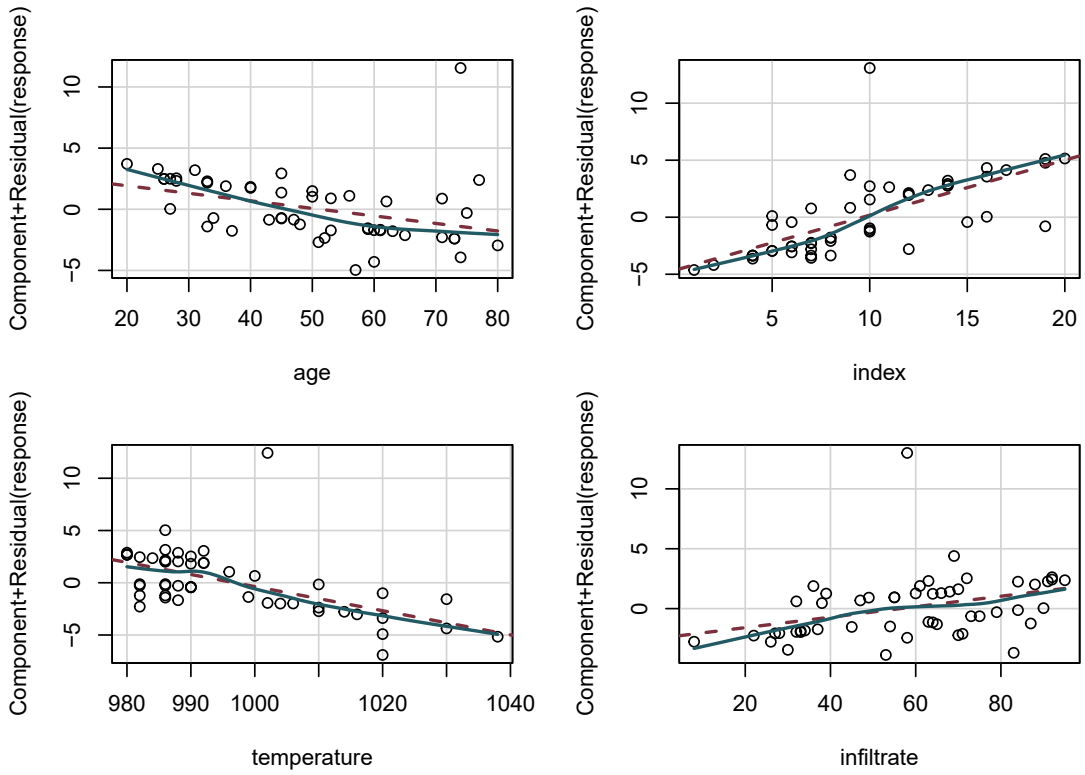
```
> 1-pchisq(mod2$deviance,mod2$df.residual)

[1] 0.7153545


> 1-pchisq(mod3$deviance,mod3$df.residual)

[1] 0.6280164
```

Παρότι και τα δύο μοντέλα φαίνεται να προσαρμόζονται παραπάνω από ικανοποιητικά στα δεδομένα, είναι εμφανές πως το σχετικό p-value για το μοντέλο που περιλαμβάνει τη συμμεταβλητή `infiltrate` είναι υψηλότερο και για αυτό επιλέγεται τελικά το μοντέλο που περιγράφεται από τη σχέση

$$\hat{p} = \frac{\exp\left(-0.06\ \text{age} + 0.034\ \text{infiltrate} + 0.407\ \text{index} - 0.099\ \text{temp.} + 95.568\right)}{1 + \exp\left(-0.06\ \text{age} + 0.034\ \text{infiltrate} + 0.407\ \text{index} - 0.099\ \text{temp.} + 95.568\right)}. \tag{2.1}$$

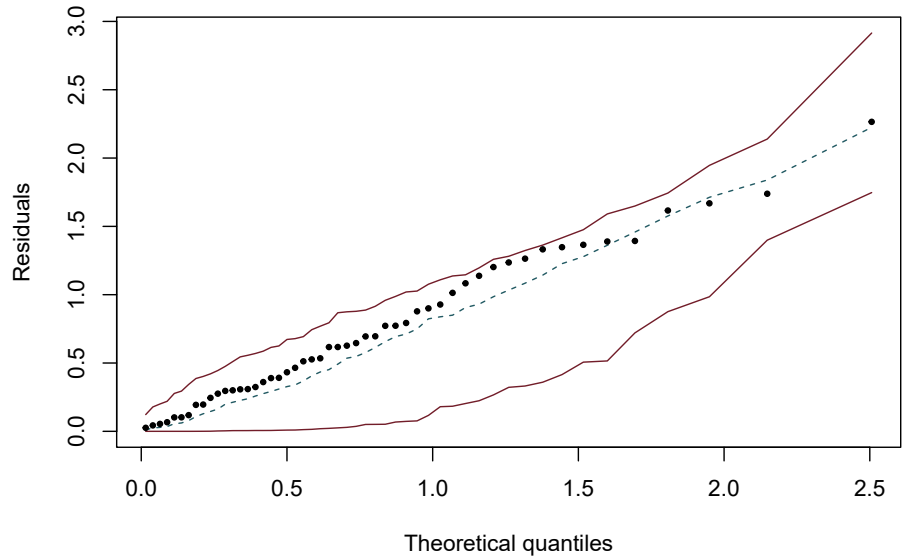
Η υψηλή τιμή για τη σταθερά (intercept) δεν αποτελεί λόγο ανησυχίας: εάν κανείς λάβει υπ' όψιν πως η μέση τιμή της θερμοκρασίας στα δεδομένα (`mean(data$temperature)`) είναι 996.137 και άρα το γινόμενο $-0.099 \cdot 996.137$ ισούται με -98.618, φαίνεται πως το πολύ υψηλό αυτό νούμερο «διορθώνεται». Η Σχέση (2.1) θα σχολιαστεί περαιτέρω αφότου υπολογιστούν τα διαστήματα εμπιστοσύνης για τους συντελεστές των συμμεταβλητών, όπου θα δοθούν και ερμηνείες για αυτούς.

2.2 Προχωρώντας στην απεικόνιση των αντίστοιχων γραφημάτων, παρατίθενται αρχικά τα διαγράμματα των μερικών υπολοίπων ανά χαρακτηριστικό, τα οποία προκύπτουν μέσω της εντολής `crPlots(mod2)` στην R και απεικονίζονται στην Εικόνα 2.1. Εκεί, φαίνεται πως τα υπόλοιπα και για τις 4 συμμεταβλητές του μοντέλου έχουν μια σχετικά γραμμική τάση, αν και η προσαρμογή των υπολοίπων της `age` και της `infilitrate` στην αντίστοιχη ευθεία θα μπορούσε να είναι και καλύτερη. Αξίζει να σημειωθεί πως σε κάθε διάγραμμα παρατηρείται ένα ακριβώς σημείο το οποίο εμφανώς αποτελεί outlier και αναμένεται να φανεί στη συνέχεια της ανάλυσης ως σημείο επιρροής (το σημείο με δείκτη 47).



Εικόνα 2.1: Διαγράμματα μερικών υπολοίπων ανά χαρακτηριστικό.

Το επόμενο γράφημα, το οποίο φαίνεται στην Εικόνα 2.1, αποτελεί τη γραφική παράσταση των υπολοίπων Deviance με την ημι-κανονική κατανομή, είναι δηλαδή ένα γράφημα το οποίο μπορεί να διαβαστεί ως «μισό QQ-plot», και προκύπτει μέσω της βιβλιοθήκης `hnp`.

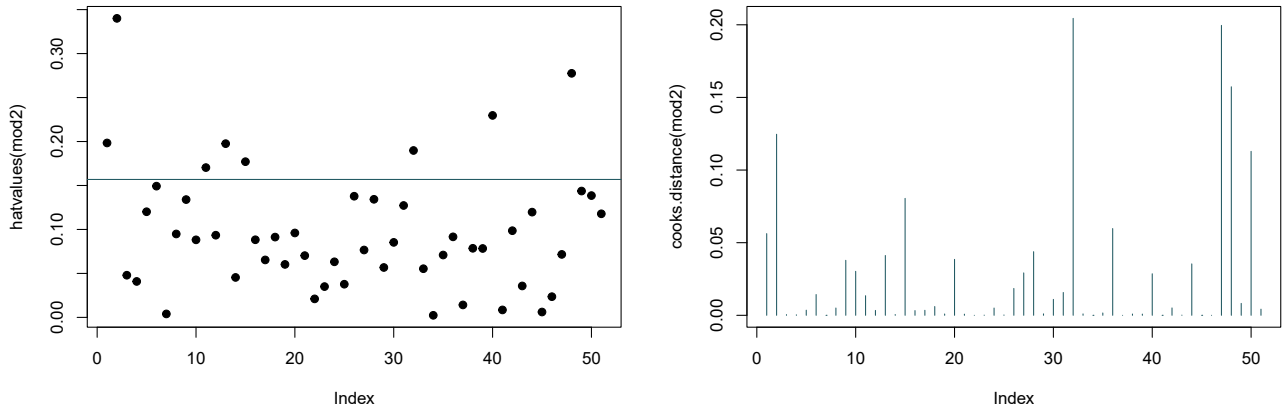


Εικόνα 2.2: Υπόλοιπα Deviance σε ημι-κανονική κατανομή.

Παρότι τα υπόλοιπα της προσαρμογής δεν αναμένεται να κατανέμονται βάσει της τυποποιημένης κανονικής κατανομής, φαίνεται να υπάρχει μια γραμμική τάση σε αυτά. Επιπλέον, δεδομένου πως τα όρια εμπιστοσύνης του γραφήματος (οι κόκκινες γραμμές, οι οποίες συνθέτουν

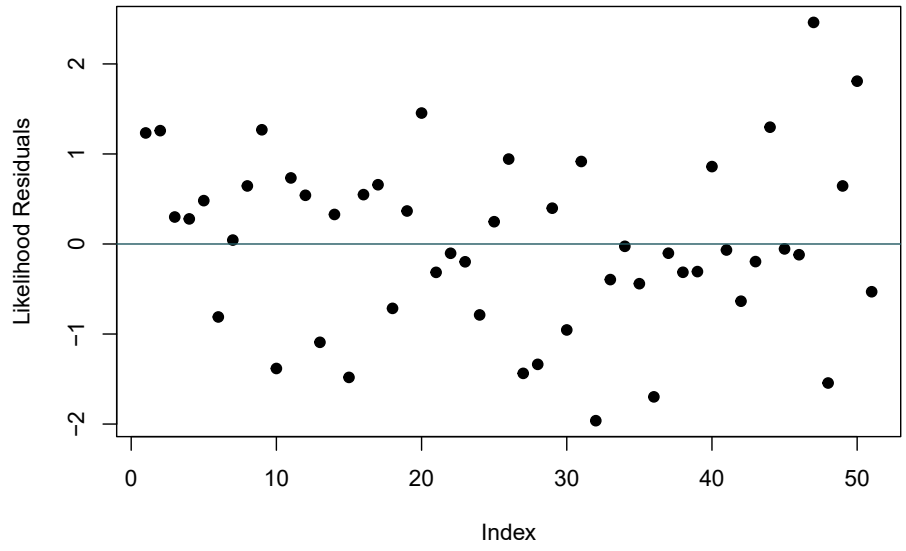
το λεγόμενο envelope) περικλύουν πλήρως τα σημεία που αντιστοιχούν στα υπόλοιπα, η καλή προσαρμογή του μοντέλου στα δεδομένα επαληθεύεται ξανά.

Για τα υπόλοιπα γραφήματα οι σχετικές εντολές στην R έχουν ήδη δοθεί στην προηγούμενη άσκηση, επομένως εδώ αυτά παρατίθενται απλώς στις Εικόνες 2.3 (Index plots για τα hat values και τις αποστάσεις Cook) και 2.4 (Index plot για τα υπόλοιπα πιθανοφάνειας).



Εικόνα 2.3: Index plots για τα hat values (h_{ii}) και τις αποστάσεις Cook.

Δεδομένου του εμπειρικού κατωφλίου $2p/n$, υπάρχουν αρκετά σημεία των οποίων τα hat-values το ξεπερνούν, με χαρακτηριστικά αυτό με δείκτη 2 και αυτό με δείκτη 47. Ειδικά το σημείο με δείκτη 47 προέκυψε και προηγουμένως στην ανάλυση ως outlier στα γραφήματα της Εικόνας 2.1. Το γεγονός πως τα σημεία αυτά αντιστοιχούν σε outliers και όχι σε σημεία επιρροής τεκμαίρεται και από το γεγονός πως αυτά δεν εμφανίζονται στο αντίστοιχο διάγραμμα με τις αποστάσεις Cook, όπου κανένα σημείο δεν ξεπερνά το μοναδιαίο κατώφλι, κατά μεγάλο μάλιστα περιθώριο.



Εικόνα 2.4: Index plot για τα υπόλοιπα πιθανοφάνειας.

Τέλος, σε ό,τι αφορά το Index plot για τα υπόλοιπα πιθανοφάνειας, τα σημεία φαίνονται αρκετά τυχαία κατανεμημένα γύρω από την ευθεία $y = 0$, το οποίο αποτελεί άλλη μια ένδειξη πως το μοντέλο που περιγράφεται από τη Σχέση (2.1) είναι ικανοποιητικό.

2.3 Σε ό,τι αφορά τα διαστήματα εμπιστοσύνης για τους εκτιμημένους συντελεστές των χαρακτηριστικών του μοντέλου, αυτά προκύπτουν μέσω των ακόλουθων:

```
> confint(mod2)

                  2.5 %       97.5 %
(Intercept) 28.037466906 182.78486822
age         -0.120608820  -0.01240951
index        0.188126797   0.71036499
temperature -0.189030929  -0.03052219
infiltrate  -0.003456431   0.07985518
```

Σε ό,τι αφορά τις συμμεταβλητές `age`, `index` και `temperature`, ο αριθμός μηδέν δεν περιέχεται στο διάστημα εμπιστοσύνης τους, γεγονός που επαληθεύει τη σημαντικότητά τους για το μοντέλο. Από την άλλη, το ίδιο δεν ισχύει για τη συμμεταβλητή `infiltrate`, στο 95% διάστημα εμπιστοσύνης της οποίας περιέχεται και ο αριθμός 0. Το γεγονός αυτό δεν προκαλεί έκπληξη, αφού ούτως ή άλλως η τιμή του z-value που αντιστοιχούσε στην `infiltrate` μέσω του ελέγχου Wald υποδείκνυε την απόρριψή της από το μοντέλο. Ο μόνος λόγος για τον οποίο αυτή δεν απορρίφθηκε ήταν η τιμή του AIC σε συνδυασμό με τη σύγκριση του μοντέλου με το αντίστοιχο κορεσμένο. Αξίζει επίσης να παρατηρηθεί πως, παρότι το 0 τυπικά ανήκει στο διάστημα εμπιστοσύνης 95%, πρακτικά αποτελεί το αριστερό άκρο του διαστήματος αυτού και δε βρίσκεται, για παράδειγμα, στο κέντρο του διαστήματος (για το λόγο αυτό, εάν κανείς υπολογίσει το αντίστοιχο διάστημα εμπιστοσύνης 90% θα δει πως το 0 παύει πλέον να περιέχεται σε αυτό). Με βάση τα αποτελέσματα αυτά, μπορεί κανείς να προχωρήσει στην ερμηνεία του μοντέλου.

Αξίζει αρχικά να σημειωθεί πως η Σχέση (2.1) μπορεί να γραφεί ισοδύναμα στη μορφή

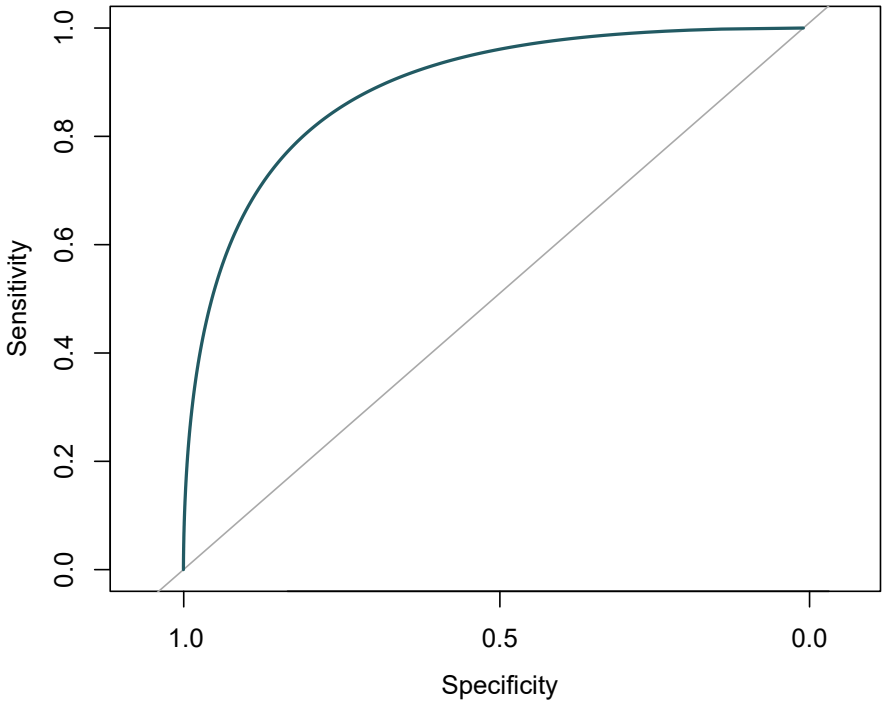
$$\frac{\hat{p}}{1-\hat{p}} = \exp\left(-0.06\,\text{age} + 0.034\,\text{infiltrate} + 0.407\,\text{index} - 0.099\,\text{temp.} + 95.568\right). \qquad (2.2)$$

Από τη μορφή αυτή φαίνεται πως για κάθε συντελεστή $\hat{\beta}_i$, η ποσότητα $\exp\left(\hat{\beta}_i\right)$ αποτελεί τον παράγοντα επί τον οποίο πολλαπλασιάζεται ο λόγος των συμπληρωματικών πιθανοτήτων πραγματοποίησης του γεγονότος `response` = 1. Αυτό, μάλιστα, αποτελεί χαρακτηριστικό κάθε μοντέλου λογιστικής παλινδρόμησης [1].

Σε ό,τι αφορά το χαρακτηριστικό `age`, με βάση τα παραπάνω, η πληροφορία που προκύπτει από το συντελεστή είναι πως για δύο ασθενείς εκ των οποίων ο ένας είναι κατά ένα χρόνο μεγαλύτερος, ο γηραιότερος ασθενής έχει από 1.23% έως και 11.36% μειωμένη πιθανότητα να ανταποκριθεί στη θεραπεία, με τη μέση μείωση να ισούται με 5.85%. Σχετικά με το δείκτη των κυττάρων λευχαιμίας, `index`, ένας ασθενής για τον οποίο ο δείκτης αυτός είναι μεγαλύτερος κατά μία μονάδα σε σχέση με τον αντίστοιχο ενός άλλου ασθενή έχει αυξημένη πιθανότητα ανταπόκρισης στη θεραπεία κατά 20.69% έως και 103.47%, με μέση τιμή 50.19%. Από την άλλη, η αύξηση του χαρακτηριστικού `temperature` κατά μία μονάδα οδηγεί σε μείωση της πιθανότητας ανταπόκρισης σε ποσοστό που κυμαίνεται από 3% έως και 17.22%, με μέσο ποσοστό μείωσης ίσο με 9.47%. Τέλος, το γεγονός πως η τιμή 0 περιέχεται στο διάστημα εμπιστοσύνης 95% της συμμεταβλητής `infiltrate` σημαίνει πως η αύξηση της τιμής της κατά μία μονάδα δεν οδηγεί με βεβαιότητα σε αύξηση ή μείωση της αντίστοιχης πιθανότητας ανταπόκρισης στη θεραπείας. Το εύρος είναι από -0.35% (με το αρνητικό πρόσημο να υποδηλώνει μειωμένη πιθανότητα) έως και 8.31% (με το θετικό πρόσημο να υποδηλώνει αυξημένη πιθανότητα), ενώ η μέση τιμή ισούται με 3.47%, γεγονός που υποδεικνύει πως κατά μέσο όρο η αύξηση του `infiltrate` κατά μία μονάδα οδηγεί σε αυξημένες πιθανότητες ανταπόκρισης.

2.4 Κλείνοντας τη μελέτη των δεδομένων για τους ασθενείς με λευχαιμία μέσω λογιστικής παλινδρόμησης, παρατίθεται μια καμπύλη ROC (Receiver Operating Characteristic), η οποία αποτελεί ένα δείκτη της ευαισθησίας (sensitivity) προς την ειδικότητα (specificity) του μοντέλου. Γενικά, όσο πιο μακριά από την ευθεία που διέρχεται από τα $(0, 1)$ και $(1, 0)$ βρίσκεται η προκύπτουσα καμπύλη, τόσο καλύτερη είναι η προβλεπτική ικανότητα του μοντέλου. Ένας από τους δείκτες που αντιστοιχούν στην απόσταση αυτή είναι το λεγόμενο «area under the curve» (AUC), το οποίο στην ιδανική περίπτωση της τέλειας πρόβλεψης ισούται με τη μονάδα. Η καμπύλη φαίνεται στο σχήμα της Εικόνας 2.5, ενώ οι σχετικές εντολές σε R [1] είναι οι ακόλουθες:

```
> library(pROC)
> roc(data$response,fitted.values(mod2),smooth=TRUE,plot=TRUE)

Data: fitted.values(mod2) in 27 controls
      (data$response 0) < 24 cases (data$response 1).
Smoothing: binormal
Area under the curve: 0.8867
```



Εικόνα 2.5: Καμπύλη ROC για το μοντέλο της Σχέσης (2.1).

Η προκύπτουσα τιμή για το AUC είναι 0.8867, γεγονός που υποδεικνύει πως το μοντέλο της Σχέσης (2.1) έχει αρκετά καλή προβλεπτική ικανότητα, επομένως δίνει μια επιπλέον αφορμή για να γίνεται λόγος για ένα ικανοποιητικό, αν όχι καλό μοντέλο.

Αναφορες

[1] Χ. Καρώνη και Π. Οικονόμου, *Στατιστικά Μοντέλα Παλινδρόμησης με χρήση MINITAB και R*. Εκδόσεις Συμεών, 2020.