Κεφάλαιο 17

Poisson Παλινδρόμηση και Λογαριθμικά Γραμμικά Μοντέλα

Πολλές φορές σε εφαρμογές παρατηρούνται δεδομένα συχνοτήτων, π.χ. ο αριθμός των περιπτώσεων στα κελιά ενός πίνακα συνάφειας, ο αριθμός τροχαίων αυτοκινητιστικών δυστυχημάτων, ο αριθμός πελατών στην τράπεζα κ.ο.κ. Η κατανομή Poisson χρησιμεύει πιο πολύ στην ανάλυση αυτών των δεδομένων και είναι γνωστό ότι δίνεται από τον τύπο,

$$P(Y = y) = \frac{\mu^y e^{-\mu}}{y!}.$$

Η παράμετρος μ, η μέση τιμή της Poisson, είναι σημαντική και συνήθως δίνεται σαν "ρυθμός", όπως π.χ. ο αριθμός των πελατών που αγοράζουν το προϊόν Α ανά 100 που πελάτες του ιδίου καταστήματος, ο αριθμός τροχαίων ανά 1000 άτομα, κ.ο.κ.

17.1 Poisson Παλινδρόμηση

Έστω $Y_1, Y_2, \ldots, Y_n$ ανεξάρτητες τ.μ. από την $Poisson(\mu_i)$. Υποθέτουμε ότι $E(Y_i) = \mu_i = n_i\theta_i$. Αν για παράδειγμα Y_i είναι ο αριθμός απαιτήσεων σε ασφαλιστική εταιρεία για ένα μοντέλο αυτοκινήτου Α, τότε, n_i είναι ο αριθμός των μοντέλων Α που έχουν ασφαλιστεί, και θ_i μπορεί να είναι η ηλικία, η χρήση, η περιοχή κ.ο.κ. Για την ανάλυση τέτοιου είδους δεδομένων συνήθως χρησιμοποιείται

το μοντέλο

$$\theta_i = e^{x_i^T \beta},$$

οπότε το αντίστοιχο γενικευμένο γραμμικό μοντέλο δίνεται από

$$E(Y_i) = \mu_i = n_i e^{x_i^T \beta}.$$

Για παράδειγμα αν $\mathbf{x_i = 0, 1}$, τότε

$$E(Y_i|X_i = 0) = n_i, E(Y_i|X_i = 0) = n_i e^{\beta}$$

και συνεπώς το ποσοστιαίο πηλίκο δίνεται από

$$RR = \frac{E(Y_i|X_i = 1)}{E(Y_i|X_i = 0)} = e^{\beta}$$

και δείχνει την αλλαγή στην αναμενόμενη τιμή. Η εκτίμηση της παραμέτρου $\boldsymbol{\beta}$ γίνεται μέσω της θεωρίας πιθανοφάνειας για γενικευμένα γραμμικά μοντέλα.

Αν $\hat{\beta}$ είναι η Ε.Μ.Π. τότε μπορούμε να ελέγξουμε τις υποθέσεις $H_0 : \beta = \beta_0$ με score test, Wald test και έλεγχο πηλίκου πιθανοφάνειας.

Επίσης,

$$\hat{Y}_i = \hat{\mu}_i = n_i e^{\mathbf{x_i}\hat{\boldsymbol{\beta}}}.$$

Τα υπόλοιπα Pearson δίνονται από

$$r_i = \frac{O_i - E_i}{\sqrt{E_i}},$$

με $O_i = Y_i$ και $E_i = \hat{Y}_i$. Τότε,

$$X^2 = \sum r_i^2 = \sum (\frac{O_i - E_i}{\sqrt{E_i}})^2.$$

Η συνάρτηση deviance δίνεται από

$$D = 2 \sum \left\{ O_i \log \left(\frac{O_i}{E_i} \right) - (O_i - E_i) \right\},$$

και τα υπόλοιπα deviance

$$d_i = sign(O_i - E_i) \sqrt{2 \left[O_i \log \left(\frac{O_i}{E_i} \right) - (O_i - E_i) \right]}$$

Οπότε, $D = \sum_{i=1}^{n} d_i^2$, και απορρίπτω το μοντέλο αν σε επίπεδο σημαντικότητας α, είτε το D είτε το X^2 είναι μεγαλύτερο του X^2_{N-p}.

17.2 Παράδειγμα

Τα παρακάτω δεδομένα αναφέρονται σε μία μελέτη όπου όλοι οι Βρετανοί γιατροί απάντησαν σε ένα ερωτηματολόγιο σχετικά με το αν καπνίζουν ή όχι. Ο παρακάτω πίνακας δείχνει τον αριθμό θανάτων από στεφανιαία νόσο μετά από 10 χρόνια. Παρουσιάζει επίσης και τον ολικό πληθυσμό.

	Smokers		Non-Smokers	
Age group	Deaths	Population	Deaths	Population
35 - 44	32	52407	2	18790
45 - 54	104	43248	12	10673
55 - 64	206	28612	28	5710
65 - 74	186	12663	28	2585
75 - 84	102	5317	31	1462

Στην R το πλαίσιο των δεδομένων κατασκευάζεται ως εξης :

```
> deaths <- c(32,2,104,12,206,28,186,28,102,31)
> population <- c(52407,18790,43248,10673,28612,5710,12663,2585,5317,1462)
> smoke <-gl(2,1,10,labels=c("Yes","No"))
> age <- gl(5,2,10,labels=c("35--44","45--54","55--64", "65--74", "75--84"))
> chddata=data.frame(deaths,population,smoke,age)
> chddata
   deaths population smoke    age
1      32      52407   Yes 35--44
2       2      18790    No 35--44
3     104      43248   Yes 45--54
4      12      10673    No 45--54
5     206      28612   Yes 55--64
6      28       5710    No 55--64
7     186      12663   Yes 65--74
8      28       2585    No 65--74
9     102       5317   Yes 75--84
10     31       1462    No 75--84
```

Θα εξεταστούν τρία ερωτήματα:

1. Είναι τα ποσοστά θανάτου πιο ψηλά στους καπνιστές;

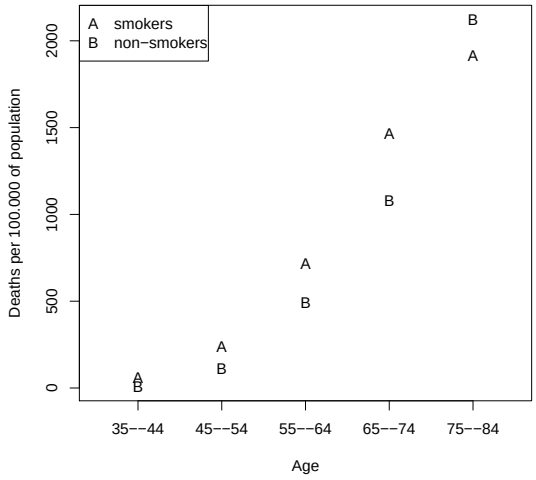
2. Αν ναι, κατα πόσο;

3. Υπάρχει διαφοροποίηση λόγω ηλικίας;

Μια πρώτη περιγραφή του προβλήματος δίνεται μέσω του γραφήματος στο Σχήμα 17.1, το οποίο παρουσιάζει τους θανάτους ανά 100 χιλιάδες άτομα του πληθυσμού συναρτήσει της ηλικιακής ομάδας για τους καπνιστές (A) και μη καπνιστές (B), αντίστοιχα. Υπάρχει ένδειξη ότι,με εξαίρεση την ηλικιακή ομάδα 75 εως 84 χρονών, τα ποσοστά θανάτου στους καπνιστές είναι μεγαλύτερα από τα αντίστοιχα στους μη καπνιστές, αλλά και η διαφορά των ποσοστών αυξάνεται όσο μεγαλώνει η ηλικία των ατόμων. Το γράφημα αυτό κατασκευάζεται χρησιμοποιώντας τις εντολές :

```
> rate= deaths*100000/population
> plot(age[smoke=="No"], rate[smoke=="No"], xlab="Age",
+ ylab="Deaths per 100.000 of population",lty=0,ylab=c(0,2500))

> points(age[smoke=="Yes"], rate[smoke=="Yes"], pch="A")
> points(age[smoke=="No"], rate[smoke=="No"], pch="B")
> legend("topleft",c("smokers","non-smokers"),pch=c("A","B"))
```



Σχήμα 17.1: Θανάτοι ανά 100 χιλιάδες άτομα συναρτήσει της ηλικιακής ομάδας.

Το μοντέλο που θα χρησιμοποιηθεί για την ανάλυση είναι το ακόλουθο:

$$\log(deaths_i) = \log(population_i) + \beta_1 + \beta_2 smoke_i + \beta_3 agecat_i + \beta_4 agesq_i + \beta_5 smkage_i$$

όπου $i = 1, 2, \ldots, 5$ για τους καπνιστές και $i = 6, 7, \ldots, 10$ για τους μη καπνιστές. Επίσης,

$$smoke_i = \begin{cases} 1 & \text{για} \quad \text{ΝΑΙ} \\ 0 & \text{για} \quad \text{ΟΧΙ} \end{cases}, \quad agecat_i = \begin{cases} 1 & \text{για} \quad \text{35-44} \\ 2 & \text{για} \quad \text{45-54} \\ 3 & \text{για} \quad \text{55-64} \\ 4 & \text{για} \quad \text{65-74} \\ 5 & \text{για} \quad \text{75-84}, \end{cases},$$

και

$$agesq_i = (agecat_i)^2, \quad smkage_i = \begin{cases} agecat_i & \text{για} \quad \text{καπνιστές} \\ 0 & \text{για} \quad \text{μη καπνιστές} \end{cases},$$

Για την εφαρμογή της ανάλυσης στην R χρησιμοποιήθηκαν οι ακόλουθες εντολές :

```
> age <- as.numeric(age)
> age
 [1] 1 1 2 2 3 3 4 4 5 5
> smoke <- ifelse(smoke=="Yes",1,0)
> smoke
 [1] 1 0 1 0 1 0 1 0 1 0
> agesq <- age^{2}
> agesq
 [1]  1  1  4  4  9  9 16 16 25 25
> agesm <-ifelse(smoke==0, age, 0)
> agesm
 [1] 0 1 0 2 0 3 0 4 0 5
> populationl <- log(population)

> fit1 <- glm(deaths~offset(populationl)+smoke+age+agesq+agesm, family=poisson)
> summary(fit1)

Call:
glm(formula = deaths ~ offset(populationl) + smoke + age + agesq +
    agesm, family = poisson)

Deviance Residuals:
       1         2         3         4         5         6         7         8
```

```
 0.43820  -0.83049  -0.27329   0.13404  -0.15265   0.64107   0.23393  -0.41058
       9        10
-0.05700  -0.01275

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.79176    0.45008 -23.978  < 2e-16 ***
smoke         1.44097    0.37220   3.872 0.000108 ***
age           2.06893    0.18170  11.386  < 2e-16 ***
agesq        -0.19768    0.02737  -7.223 5.08e-13 ***
agesm         0.30755    0.09704   3.169 0.001528 **
---
Signif. codes:  0 ‘‘***’’ 0.001 ‘‘**’’ 0.01 ‘‘*’’ 0.05 ‘‘.’’ 0.1 ‘‘ ’’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 935.0673  on 9  degrees of freedom
Residual deviance:   1.6354  on 5  degrees of freedom
AIC: 66.703

Number of Fisher Scoring iterations: 4
> rate.ratio <- exp(fit1$coef[-1])
> rate.ratio
    smoke       age     agesq     agesm
4.2247998 7.9163500 0.8206353 1.3600862
```

Από τον πίνακα συντελεστών συμπεραίνεται ότι και οι 4 επεξηγηματικές μεταβλητές είναι σημαντικές για το μοντέλο. Συνεπώς, η πιθανότητα θανάτου επηρεάζεται από το αν κάποιος είναι καπνιστής αλλά και από την ηλικία του. Το `rate.ratio` που υπολογίστηκε στο τέλος, δείχνει ότι για αυτούς που καπνίζουν, το ρίσκο θανάτου είναι 4 φορές μεγαλύτερο.

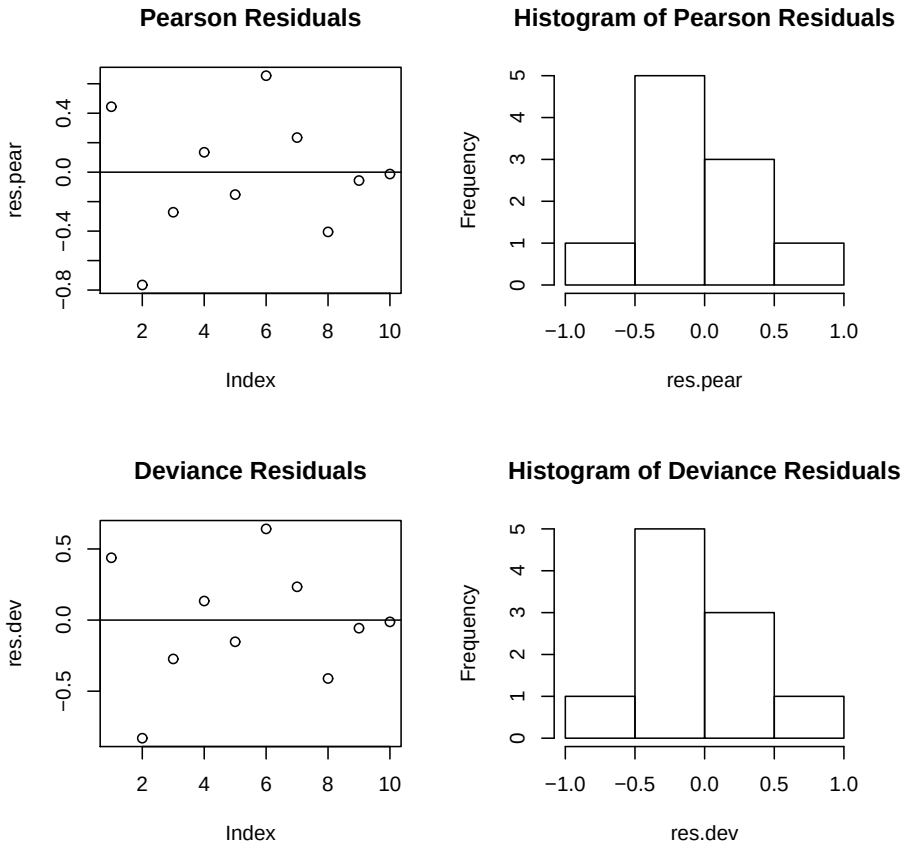
Τέλος, θα εξεταστεί η καταλληλότητα του μοντέλου με τη βοήθεια της ανάλυσης των υπολοίπων. Για το λόγο αυτό υπολογίζονται τα υπόλοιπα `pearson` και `deviance`, όπως και οι εκτιμώμενες τιμές της εξαρτημένης μεταβλητής, η οποία στο πιο πάνω παράδειγμα είναι ο αριθμός των θανάτων. Επίσης, κατασκευάζεται ένας πίνακας ο οποίος παρουσιάζει τα αποτελέσματα αυτά για κάθε συνδυασμό των παραγόντων `age` και `smoke`. Στη συνέχεια δίνεται ο έλεγχος καλής προσαρ-

μογής και για τα δύο είδη υπολοίπων, υποδεικνύοντας την καταλληλότητα του μοντέλου. Από το γράφημα και το ιστόγραμμα των υπολοίπων είναι φανερή η τυχαιότητα και η κανονικότητά τους. Φυσικά δεν μπορούν να εξαχθούν ακριβή συμπεράσματα λόγω του μικρού αριθμού των υπολοίπων.

```
> res.pear <- residuals(fit1, type="pearson")
> res.dev  <- residuals(fit1, type="deviance")
> predict.fit <- predict(fit1, type="response")
> cbind(age, smoke, deaths, predict.fit, res.pear, res.dev)
   age smoke deaths predict.fit    res.pear     res.dev
1    1     1     32   29.584734  0.44404929  0.43820403
2    1     2      2    3.414801 -0.76561908 -0.83049031
3    2     1    104  106.811960 -0.27208163 -0.27328873
4    2     2     12   11.541629  0.13492231  0.13404370
5    3     1    206  208.198646 -0.15237591 -0.15264528
6    3     2     28   24.743377  0.65469354  0.64106682
7    4     1    186  182.827893  0.23459923  0.23392570
8    4     2     28   30.229155 -0.40544060 -0.41058325
9    5     1    102  102.576767 -0.05694769 -0.05700118
10   5     2     31   31.071038 -0.01274427 -0.01274913

> devian.fit <- sum(res.dev^{2})
> 1-pchisq(devian.fit, df=10-5)
[1] 0.8969393
> pear.fit <- sum(res.pear^{2})
> 1-pchisq(pear.fit, df=10-5)
[1] 0.907199

> par(mfrow=c(2,2))
> plot(res.pear,main="Pearson Residuals")
> abline(h=0)
> hist(res.pear,main="Histogram of Pearson Residuals")
> plot(res.dev,main="Deviance Residuals")
> abline(h=0)
> hist(res.dev,main="Histogram of Deviance Residuals")
```



Σχήμα 17.2: Γραφήματα υπολοίπων.