Κεφάλαιο 10

Λογιστική Παλινδρόμηση

Στο κεφάλαιο αυτό θα δούμε την μέθοδο της λογιστικής παλινδρόμησης η οποία χρησιμεύει στο να αναπτύξουμε σχέση μίας δίτιμης ανεξάρτητης τυχαίας μεταβλητής και συνεχών η διακριτών ανεξάρτητων μεταβλητών. Ουσιαστικά η μέθοδος αυτή γενικεύει τα γραμμικά μοντέλα, έτσι ώστε η εξαρτημένη μεταβλητή να ακολουθεί την εκθετική οικογένεια κατανομών.

10.1 Περιγραφή των Δεδομένων

Έρευνα με εργάτες της αμερικάνικης βιομηχανίας βαμβακιού θέλει να εξετάσει αν κάποιος εργάτης πάσχει από κάποια συγκεκριμένη ασθένεια του πνεύμονα. Επίσης, συγκεντρώθηκαν οι τιμές για τις ακόλουθες πέντε μεταβλητές :

- φυλή (race) (1=λευκός, 2=άλλο)
- φύλο (sex) (1=άρρεν, 2=θήλυ)
- κάπνισμα (1=καπνιστής, 2=μη καπνιστής)
- διάρκεια εργασίας (1= λιγότερο από 10 χρόνια, 2=10–22 χρόνια, 3= περισσότερο από 20 χρόνια)
- σκόνη: ποσοστό σκόνης στον εργασιακό χώρο (1=ψηλό, 2=μέτριο 3=χαμηλό)

Τα δεδομένα βρίσκονται στο παράρτημα αυτού του κεφαλαίου.

Το πρόβλημα για αυτά τα δεδομένα είναι το να εξακριβωθεί κατά πόσο οι επεξηγηματικές μεταβλητές είναι σημαντικές στην εμφάνιση αυτής της ασθένειας.

Με άλλα λόγια, ποιες από αυτές τις μεταβλητές μπορούν να χρησιμοποιηθούν για να προβλέψουν κατά πόσο ένας εργάτης πάσχει από ασθένεια του πνεύμονα. Επειδή η ανεξάρτητη μεταβλητή είναι δυαδική, θα χρησιμοποιηθεί η λογιστική παλινδρόμηση για την ανάλυση.

10.2 Λογιστική Παλινδρόμηση

Αντί να χρησιμοποιηθεί ένα γραμμικό μοντέλο για να εξεταστεί η εξάρτηση της πιθανότητας εμφάνισης της ασθένειας του πνεύμονα από τις επεξηγηματικές μεταβλητές, χρησιμοποιείται ο λογιστικός μετασχηματισμός, ο οποίος ορίζεται ως

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 x_2 + \ldots + \beta_k X_k. \tag{10.1}$$

Στο παράδειγμα, p είναι η πιθανότητα ένας εργάτης να πάσχει από ασθένεια του πνεύμονα. Στο μοντέλο υπάρχουν k (στο παράδειγμα 5) επεξηγηματικές μεταβλητές. Οι συντελεστές παλινδρόμησης εκτιμούνται με τη μέθοδο της μέγιστης πιθανοφάνειας με την υπόθεση ότι η εξαρτημένη μεταβλητή ακολουθεί τη διωνυμική κατανομή. Από την εξίσωση (10.1), το p μπορεί να υπολογιστεί από

$$p = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k)} \tag{10.2}$$

10.3 Ανάλυση στην R

Οι πρώτες δύο στήλες των δεδομένων καταγράφουν τη συχνότητα των εργατών με ή χωρίς την ασθένεια για τις αντίστοιχες τιμές (κατηγορίες) των επεξηγηματικών μεταβλητών. Η ανάλυση της λογιστικής παλινδρόμησης γίνεται με την εντολή `glm` με ανάλογο τρόπο με τη εντολή `lm` για τη γραμμική παλινδρόμησης δίνοντας και την συνάρτηση σύνδεσης (link function) με το όρισμα `family`.

```
> logreg<-read.table("logistic1.txt",header=T)
> attach(logreg)
> out1<-glm( cbind(Yes, No)~dust+race+sex+smoking+Empleng,
+ family=binomial)
> out1

Call:  glm(formula = cbind(Yes, No) ~ dust + race + sex + smoking
       +Empleng, family = binomial)
```

```
Coefficients:
(Intercept)       dust       race        sex    smoking    Empleng
    -0.4852    -1.3751     0.2463    -0.2590    -0.6292     0.3856

Degrees of Freedom: 64 Total (i.e. Null);  59 Residual
Null Deviance:      322.5
Residual Deviance: 69.51        AIC: 188.2
```

Το αποτέλεσμα δίνει τις εκτιμήσεις των συντελεστών των παραμέτρων, την απόκλιση (deviance) του μηδενικού μοντέλου και των υπολοίπων μαζί με τους βαθμούς ελευθερίας τους αλλά και την τιμή του κριτηρίου AIC. Πιο λεπτομερή ανάλυση των συντελεστών των παραμέτρων δίνεται με την εντολή summary, ενώ η εντολή anova παρουσιάζει τον πίνακα ανάλυσης της απόκλισης.

```
> summary(out1)

Call:
glm(formula = cbind(Yes, No) ~ dust + race + sex + smoking +
    Empleng, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.4126  -0.7573  -0.2421   0.3688   1.9804

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.4852     0.6060  -0.801 0.423312
dust         -1.3751     0.1155 -11.901  < 2e-16 ***
race          0.2463     0.2061   1.195 0.232026
sex          -0.2590     0.2116  -1.224 0.220949
smoking      -0.6292     0.1931  -3.259 0.001119 **
Empleng       0.3856     0.1069   3.607 0.000310 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

```
    Null deviance: 322.527  on 64  degrees of freedom
Residual deviance:  69.509  on 59  degrees of freedom
AIC: 188.19

Number of Fisher Scoring iterations: 5
> anova(out1)
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(Yes, No)

Terms added sequentially (first to last)


        Df Deviance Resid. Df Resid. Dev
NULL                    64     322.53
dust     1   221.96        63     100.56
race     1     1.05        62      99.51
sex      1     5.97        61      93.54
smoking  1    10.73        60      82.82
Empleng  1    13.31        59      69.51
```

Από τα πιο πάνω συμπεραίνεται ότι οι μεταβλητές dust, smoking και Empleng είναι οι πιο σημαντικές για την πρόβλεψη ασθένειας του πνεύμονα, ενώ φαίνεται ότι οι άλλες δύο μεταβλητές δεν είναι τόσο σημαντικές. Στο συμπέρασμα αυτό καταλήγουμε από το p-value τους για τον t έλεγχο, αλλά και από την συνεισφορά της κάθε μεταβλητής στην απόκλιση όταν αυτή προστεθεί στο μοντέλο, η οποία παρουσιάζεται στο πίνακα ανάλυσης της απόκλισης. Συνεπώς, εφαρμόζεται ένα νέο μοντέλο λογιστικής παλινδρόμησης με τις τρεις σημαντικές μεταβλητές και το συγκρίνεται με το προηγούμενο.

```
> out2<-glm( cbind(Yes, No)~dust+smoking+Empleng, family=binomial)
> anova(out2,out1)
Analysis of Deviance Table

Model 1: cbind(Yes, No) ~ dust + smoking + Empleng
Model 2: cbind(Yes, No) ~ dust + race + sex + smoking + Empleng
```

```
  Resid. Df Resid. Dev Df Deviance
1        61     72.562
2        59     69.509  2    3.053
> 1-pchisq(3.053,2)
[1] 0.2172949
```

Ο έλεγχος σύγκρισης μοντέλου έχει για μηδενική υπόθεση H_0 ότι το νέο μοντέλο εφαρμόζει καλύτερα τα δεδομένα. Ο έλεγχος είναι X^2 και αφού το p-value (`1-pchisq(3.053,2)`) είναι μεγαλύτερο από 0.05, δεν απορρίπτεται η μηδενική υπόθεση. Ποιο κάτω παρουσιάζεται η ανάλυση για του συντελεστές του μικρότερου μοντέλου.

```
> summary(out2)

Call:
glm(formula = cbind(Yes, No) ~ dust + smoking + Empleng, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.3421  -0.7700  -0.2518   0.4001   2.0523

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.14177    0.34120  -0.415 0.677783
dust        -1.46572    0.10578 -13.856  < 2e-16 ***
smoking     -0.67781    0.18871  -3.592 0.000328 ***
Empleng      0.33313    0.08861   3.760 0.000170 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 322.527  on 64  degrees of freedom
Residual deviance:  72.562  on 61  degrees of freedom
AIC: 187.24

Number of Fisher Scoring iterations: 5
```

Θεωρώντας τις τιμές των συντελεστών από πιο πάνω, το μοντέλο λογιστικής παλινδρόμησης που εφαρμόζει καλύτερα τα δεδομένα δίνεται από

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = -0.1418 - 1.4657 \times \text{dust} - 0.6778 \times \text{smoking} + 0.3331 \times \text{Empleng}$$

και είναι δυνατόν να υπολογιστεί η εκτιμώμενη τιμή της πιθανότητας κάποιος εργάτης να πάσχει από ασθένεια του πνεύμονα για κάθε συνδυασμό τιμών από τις τρεις επεξηγηματικές μεταβλητές. Για παράδειγμα, αν ένας εργάτης δουλεύει σε εργασιακό χώρο με ψηλό ποσοστό σκόνης (`dust=1`), καπνίζει (`smoking=1`) και δουλεύει για περισσότερο από 20 χρόνια (`Empleng=3`), η εξίσωση δίνει το αποτέλεσμα $\log\left(\hat{p}/(1-\hat{p})\right) = -1.286$, και συνεπώς $\hat{p} = 0.2165$.

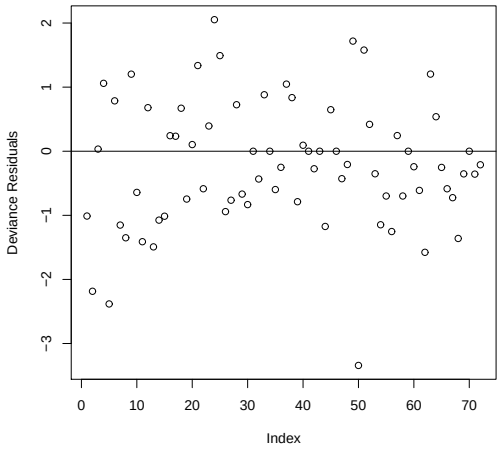
Στη συνέχεια υπολογίζονται δύο είδη υπολοίπων της λογιστικής παλινδρόμησης, τα υπόλοιπα απόκλισης και τα υπόλοιπα Pearson, και κατασκευάζεται το γράφημά τους (Σχήματα 10.1 και 10.2). Η μεθοδολογία της ανάλυσης υπολοίπων είναι παρόμοια με εκείνης της πολλαπλής γραμμικής παλινδρόμησης. Και τα δύο γραφήματα δείχνουν ότι η 50η παρατήρηση είναι λίγο προβληματική. Η αρνητική τιμή του υπολοίπου υποδεικνύει ότι η εκτιμώμενη τιμή είναι μεγαλύτερη από την παρατηρούμενη τιμή. Εξετάζοντας τα δεδομένα, παρατηρείται ότι η 50η παρατήρηση αναφέρεται στους εργάτες με μέτριο ποσοστό σκόνης στον εργασιακό τους χώρο (`dust=2`), καπνίζουν (`smoking=1`) και εργάζονται για περισσότερα από 20 χρόνια (`Empleng=3`) και άρα $\hat{p} = 0.059$. Η εκτιμώμενη πιθανότητα είναι $1/142 = 0.007$.

```
> residuals(out2, type="d")
> residuals(out2, type="pear")
> plot(residuals(out2, type="d"), xlab="Index",
+ ylab="Deviance Residuals")
> abline(h=0)
> plot(residuals(out2, type="pear"), xlab="Index",
+ ylab="Pearson Residuals")
> abline(h=0)
```

10.4 Μοντέλο Probit

Όμοια ανάλυση μπορεί να γίνει χρησιμοποιώντας το μοντέλο probit

$$p = \Phi(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_K),$$



Σχήμα 10.1: Υπόλοιπα απόκλισης.

όπου Φ η συνάρτηση πυκνότητας πιθανότητας της τυπικής κανονικής κατανομής.

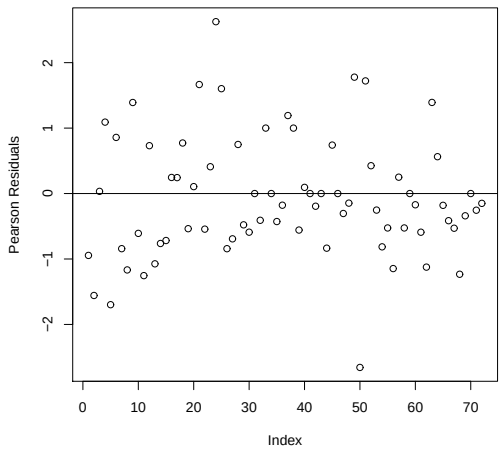
```
> out3<-glm( cbind(Yes, No)~dust+smoking+Empleng,
+ family=binomial(link=probit))
> out3

Call:  glm(formula = cbind(Yes, No) ~ dust + smoking
+ Empleng, family = binomial(link = probit))

Coefficients:
(Intercept)         dust      smoking      Empleng
    -0.4044      -0.6268      -0.2840       0.1406

Degrees of Freedom: 64 Total (i.e. Null);  61 Residual
Null Deviance:      322.5
Residual Deviance: 84.59        AIC: 199.3
> summary(out3)

Call: glm(formula = cbind(Yes, No) ~ dust + smoking + Empleng,
family = binomial(link = probit))
```



Σχήμα 10.2: Υπόλοιπα Pearson.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5085  -0.7912  -0.2626   0.2894   2.5515

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.40438    0.15877  -2.547 0.010867 *
dust        -0.62685    0.04632 -13.532  < 2e-16 ***
smoking     -0.28397    0.08214  -3.457 0.000546 ***
Empleng      0.14065    0.04056   3.468 0.000525 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 322.527  on 64  degrees of freedom
Residual deviance:  84.587  on 61  degrees of freedom
AIC: 199.26

Number of Fisher Scoring iterations: 5
```

Δεδομένα Κεφαλαίου 10

	Yes	No	dust	race	sex	smoking	Empleng
1	3	37	1	1	1	1	1
2	0	74	2	1	1	1	1
3	2	258	3	1	1	1	1
4	25	139	1	2	1	1	1
5	0	88	2	2	1	1	1
6	3	242	3	2	1	1	1
7	0	5	1	1	2	1	1
8	1	93	2	1	2	1	1
9	3	180	3	1	2	1	1
10	2	22	1	2	2	1	1
11	2	145	2	2	2	1	1
12	3	260	3	2	2	1	1
13	0	16	1	1	1	2	1
14	0	35	2	1	1	2	1
15	0	134	3	1	1	2	1
16	6	75	1	2	1	2	1
17	1	47	2	2	1	2	1
18	1	122	3	2	1	2	1
19	0	4	1	1	2	2	1
20	1	54	2	1	2	2	1
21	2	169	3	1	2	2	1
22	1	24	1	2	2	2	1
23	3	142	2	2	2	2	1
24	4	301	3	2	2	2	1
25	8	21	1	1	1	1	2
26	1	50	2	1	1	1	2
27	1	187	3	1	1	1	2
28	8	30	1	2	1	1	2
29	0	5	2	2	1	1	2
30	0	33	3	2	1	1	2
31	0	0	1	1	2	1	2
32	1	33	2	1	2	1	2
33	2	94	3	1	2	1	2
34	0	0	1	2	2	1	2

35	0	4	2	2	2	1	2
36	0	3	3	2	2	1	2
37	2	8	1	1	1	2	2
38	1	16	2	1	1	2	2
39	0	58	3	1	1	2	2
40	1	9	1	2	1	2	2
41	0	0	2	2	1	2	2
42	0	7	3	2	1	2	2
43	0	0	1	1	2	2	2
44	0	30	2	1	2	2	2
45	1	90	3	1	2	2	2
46	0	0	1	2	2	2	2
47	0	4	2	2	2	2	2
48	0	4	3	2	2	2	2
49	31	77	1	1	1	1	3
50	1	141	2	1	1	1	3
51	12	495	3	1	1	1	3
52	10	31	1	2	1	1	3
53	0	1	2	2	1	1	3
54	0	45	3	2	1	1	3
55	0	1	1	1	2	1	3
56	3	91	2	1	2	1	3
57	3	176	3	1	2	1	3
58	0	1	1	2	2	1	3
59	0	0	2	2	2	1	3
60	0	2	3	2	2	1	3
61	5	47	1	1	1	2	3
62	0	39	2	1	1	2	3
63	3	182	3	1	1	2	3
64	3	15	1	2	1	2	3
65	0	1	2	2	1	2	3
66	0	23	3	2	1	2	3
67	0	2	1	1	2	2	3
68	3	187	2	1	2	2	3
69	2	340	3	1	2	2	3
70	0	0	1	2	2	2	3
71	0	2	2	2	2	2	3

72	0	3	3	2	2	2	3