- Μερικές φορές δεν είναι εφικτό να προσδιοριστεί η συνάρτηση πιθανοφάνειας. Σε αυτή την περίπτωση, αν γνωρίζουμε τις δύο πρώτες ροπές μιας τ.μ. y, τότε η μέθοδος της quasi–πιθανοφάνειας μπορεί να χρησιμοποιηθεί για την εκτίμηση των παραμέτρων του μοντέλου.

8.9 Λογιστική παλινδρόμηση

Το **μοντέλο της λογιστικής παλινδρόμησης** (logistic regression), που εξετάζεται εδώ, αποτελεί ειδική περίπτωση των γενικευμένων γραμμικών μοντέλων, αλλά λόγω της σπουδαιότητάς του παρουσιάζεται χωριστά. Έχει αναπτυχθεί εκτεταμένη βιβλιογραφία που ασχολείται ειδικά με τη μεθοδολογία αυτού του μοντέλου.

Σε πολλές εφαρμογές η εξαρτημένη μεταβλητή παίρνει δύο μόνο τιμές, οι οποίες αντιστοιχούν σε δύο ενδεχόμενα. Για παράδειγμα, το αν ο ασθενής ζει ή απεβίωσε, το αν ο άνεργος βρίσκει εργασία ή δε βρίσκει, το αν ραγίζει ή αντέχει το δοκάρι. Οι τιμές της μεταβλητής αποτελούν μια αυθαίρετη κωδικοποίηση των δύο ενδεχομένων, συνήθως 0 και 1. Επικεντρώνουμε την προσοχή μας σε ένα από τα δύο ενδεχόμενα, την «επιτυχία» $y = 1$ με πιθανότητα $p = P(\text{επιτυχία})$. Η y είναι τ.μ. της κατανομής Bernoulli, δηλαδή $y \sim B(p)$, με $E(y) = p$ και $V(y) = p(1-p)$.

Επεκτείνοντας σε μια σειρά από n δοκιμές (δηλαδή, πραγματοποιήσεων των ενδεχομένων), ορίζουμε την τ.μ.

$$y = \text{αριθμός επιτυχιών σε } n \text{ δοκιμές.}$$

Υπό την υπόθεση ότι η πιθανότητα επιτυχίας p είναι ίδια σε κάθε δοκιμή και οι δοκιμές είναι ανεξάρτητες μεταξύ τους, τότε ισχύει η Διωνυμική (binomial) κατανομή

$$y \sim b(n, p)$$

με σ.π.
$$f(y) = \binom{n}{y} p^y (1-p)^{n-y},\ y = 0, 1, 2, \ldots, n,$$

όπου η πιθανότητα επιτυχίας p είναι παράμετρος της κατανομής. Η Διωνυμική κατανομή αποτελεί τη βασική κατανομή για την περιγραφή και ανάλυση μιας μεταβλητής αυτής της φύσης. Η μέση τιμή της y είναι ίση με $E(y) = np$ και η διασπορά με $V(y) = np(1-p)$. Στην ειδική περίπτωση που $n = 1$, μιλάμε για **δυαδικά δεδομένα**, αλλιώς για **διωνυμικά δεδομένα**.

Σε πολλές περιπτώσεις η τ.μ. y ενδέχεται να εξαρτάται από κάποιες επεξηγηματικές μεταβλητές. Η εξάρτηση της y από τις επεξηγηματικές μεταβλητές $\mathbf{x}$ (ανεξάρτητες μεταβλητές ή συμμεταβλητές) εισάγεται μέσω της εξάρτησης της πιθανότητας επιτυχίας p από τις $\mathbf{x}$ (π.χ. η πιθανότητα να μείνει κάποιος άνεργος εξαρτάται από το φύλο, την ηλικία, το μορφωτικό επίπεδο κ.ά.). Πιο συγκεκριμένα, κατασκευάζεται το αποκαλούμενο **μοντέλο της λογιστικής παλινδρόμησης**, το οποίο είναι ένα γενικευμένο γραμμικό μοντέλο και εκφράζεται μέσω της σχέσης

$$\eta_x = g\left(E(y_{\mathbf{x}})\right) = g\left(\mu_{\mathbf{x}}\right) = \mathbf{x}'\boldsymbol{\beta}$$

με την ακόλουθη δομή:

1. $y_{\mathbf{x}} \sim b(n_{\mathbf{x}}, \mu_{\mathbf{x}})$ ($n_{\mathbf{x}} > 1$, διωνυμικά δεδομένα)
   ή $y_{\mathbf{x}} \sim B(\mu_{\mathbf{x}})$ ($n_{\mathbf{x}} = 1$, δυαδικά δεδομένα)

2. $\eta_{\mathbf{x}} = g\left(\mu_{\mathbf{x}}\right) = \ln \frac{\mu_{\mathbf{x}}}{n_{\mathbf{x}} - \mu_{\mathbf{x}}} = \ln \frac{p_{\mathbf{x}}}{1-p_{\mathbf{x}}} = \text{logit}(p_{\mathbf{x}}) = \mathbf{x}'\boldsymbol{\beta}$ (συνάρτηση logit)

3. ανεξαρτησία μεταξύ των παρατηρήσεων $y_{\mathbf{x}}$,

όπου $n_{\mathbf{x}}$ είναι ο αριθμός των επαναλήψεων της τιμής του διανύσματος $\mathbf{x}$ των επεξηγηματικών μεταβλητών. Αντιστρέφοντας τη συνάρτηση σύνδεσης προκύπτει

$$p_{\mathbf{x}} = e^{\eta_{\mathbf{x}}} / \left(1 + e^{\eta_{\mathbf{x}}}\right)$$

από την οποία είναι φανερό ότι ισχύει ο απαραίτητος περιορισμός $0 < p_{\mathbf{x}} < 1$.

Για κάθε παρατήρηση i το μοντέλο γράφεται ως

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_\kappa x_{ik},\ i = 1, \ldots, n,$$

όπου

$$p_i = p_{\mathbf{x}_i} = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})} \tag{8.17}$$

η πιθανότητα «επιτυχίας» και συνεπώς

$$E(y_i) = n_i p_i = n_i \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}}.$$

Η παραπάνω συνάρτηση logit είναι η κανονική συνάρτηση σύνδεσης (link function) της Διωνυμικής κατανομής και αποτελεί τη συνηθισμένη επιλογή. Εναλλακτικές συναρτήσεις σύνδεσης που χρησιμοποιούνται για ειδικές περιπτώσεις περιλαμβάνουν τις

α) $g(\mu_{\mathbf{x}}) = \ln\{-\ln(1 - p_{\mathbf{x}})\} = \mathbf{x}'\boldsymbol{\beta}$ (συνάρτηση complementary log–log)

β) $g(\mu_{\mathbf{x}}) = \Phi^{-1}(p_{\mathbf{x}}) = \mathbf{x}'\boldsymbol{\beta}$ (συνάρτηση probit).

8.9.1 Εκτίμηση παραμέτρων και ερμηνεία

Όπως με όλα τα γενικευμένα γραμμικά μοντέλα, η προσαρμογή του μοντέλου στα δεδομένα γίνεται με τη μέθοδο μέγιστης πιθανοφάνειας. Η συνάρτηση πιθανοφάνειας L ενός δείγματος τιμών $y_1, y_2, \ldots, y_n$ με μέσες τιμές $E(y_i) = \mu_i = n_i p_i$ και συμμεταβλητές $\mathbf{x}_i' = (x_{i0}, x_{i1}, \ldots, x_{ik})$, όπου n_i ο αριθμός δοκιμών της στατιστικής μονάδας i, p_i η αντίστοιχη πιθανότητα επιτυχίας και $x_{i0} \equiv 1$, γράφεται ως

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}.$$

Η πιθανοφάνεια εξαρτάται από τις άγνωστες πιθανότητες επιτυχίας p_i, οι οποίες με τη σειρά τους εξαρτώνται από τα $\boldsymbol{\beta}$ μέσω της σχέσης (8.17). Έτσι η συνάρτηση

πιθανοφάνειας μπορεί να θεωρηθεί ως συνάρτηση των $\boldsymbol{\beta}$ με

$$\begin{aligned}\ell = \ln L(\boldsymbol{\beta}) &= \sum_{i=1}^{n}\left\{\ln\binom{n_i}{y_i} + y_i \ln p_i + (n_i - y_i)\ln(1-p_i)\right\} \\ &= \sum_{i=1}^{n}\left\{\ln\binom{n_i}{y_i} + y_i \ln\left(\frac{p_i}{1-p_i}\right) + n_i \ln(1-p_i)\right\} \\ &= \sum_{i=1}^{n}\left\{\ln\binom{n_i}{y_i} + y_i \mathbf{x}_i'\boldsymbol{\beta} - n_i \ln(1+e^{\mathbf{x}_i'\boldsymbol{\beta}})\right\}. \qquad (8.18)\end{aligned}$$

Παραγωγίζοντας έχουμε

$$\begin{aligned}\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^{n} y_i x_{ij} - \sum_{i=1}^{n} n_i x_{ij} e^{\mathbf{x}_i'\boldsymbol{\beta}}(1+e^{\mathbf{x}_i'\boldsymbol{\beta}})^{-1},\ j = 0, 1, \ldots, k \\ &= \sum_{i=1}^{n}\left[y_i - n_i e^{\mathbf{x}_i'\boldsymbol{\beta}}(1+e^{\mathbf{x}_i'\boldsymbol{\beta}})^{-1}\right] x_{ij} \\ &= \sum_{i=1}^{n}(y_i - n_i p_i) x_{ij}\,.\end{aligned}$$

Αυτό το αποτέλεσμα προκύπτει και απευθείας μέσω της γενικής σχέσης (8.9)

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} \frac{y_i - \mu_i}{V(y_i)} \frac{1}{g'(\mu_i)} x_{ij}\,,$$

αφού

$$E(y_i) = \mu_i = b'(\theta_i) = n_i p_i,\ \ a(\phi) = 1,\ \ g(\mu_i) = \theta_i$$

και άρα

$$V(y_i) = a(\phi) b''(\theta_i) = b''(\theta_i) = \mathrm{v}(\mu_i) = \frac{d\mu_i}{d\theta_i} = \frac{1}{g'(\mu_i)}\,.$$

Οι εκτιμήτριες μέγιστης πιθανοφάνειας των β_j προκύπτουν με την ικανοποίηση των εξισώσεων

$$\sum_{i=1}^{n}(y_i - n_i \hat{p}_i) x_{ij} = \sum_{i=1}^{n}(y_i - \hat{\mu}_i) x_{ij} = 0,\ j = 0, 1, \ldots, k \ \ \Rightarrow \mathbf{X}'(\mathbf{y} - \hat{\boldsymbol{\mu}}) = \mathbf{0},$$

δηλαδή εξισώνοντας τις μερικές παραγώγους με μηδέν παίρνουμε ένα σύστημα από $k+1$ μη γραμμικές εξισώσεις, το οποίο λύνεται μόνο με επαναληπτικές μεθόδους (βλ. Παράγραφο 8.4). Η λύση του συστήματος μας δίνει τις τιμές των εκτιμητριών $\hat{\boldsymbol{\beta}}$. Η αντίστοιχη προσαρμοσμένη τιμή του αριθμού των επιτυχιών για την i-οστή παρατήρηση είναι $\hat{\mu}_i = n_i \hat{p}_i$.

Ερμηνεία των συντελεστών β

Ένα μεγάλο πλεονέκτημα της λογιστικής παλινδρόμησης έναντι άλλων μοντέλων για διωνυμικά ή δυαδικά δεδομένα είναι η δυνατότητα ερμηνείας των τιμών των συντελεστών $\hat{\boldsymbol{\beta}}$.

Εφόσον εκτιμηθούν οι παράμετροι $\hat{\boldsymbol{\beta}}$, η σχέση μεταξύ της προσαρμοσμένης πιθανότητας απόκρισης και των τιμών των $x_0, x_1, x_2, \ldots, x_k$ επεξηγηματικών μεταβλητών μπορεί να εκφραστεί ως $\hat{p} = \frac{e^{\mathbf{x}'\hat{\boldsymbol{\beta}}}}{1+e^{\mathbf{x}'\hat{\boldsymbol{\beta}}}}$ ή ισοδύναμα μέσω του **λόγου των συμπληρωματικών πιθανοτήτων** (odds)

$$\frac{\hat{p}}{1-\hat{p}} = e^{\mathbf{x}'\hat{\boldsymbol{\beta}}} = e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2+\ldots+\hat{\beta}_k x_k},$$

όπου $x_0 \equiv 1$. Από τη σχέση αυτή προκύπτει ότι η ποσότητα $e^{\hat{\beta}_j}$ είναι ο παράγοντας επί τον οποίο πολλαπλασιάζεται ο λόγος των συμπληρωματικών πιθανοτήτων πραγματοποίησης του γεγονότος «επιτυχία» (odds), όταν η αντίστοιχη ανεξάρτητη μεταβλητή x_j αυξηθεί κατά μία μονάδα, με δεδομένο πάντα ότι οι υπόλοιπες συμμεταβλητές παραμένουν σταθερές. Αν ο εκτιμημένος συντελεστής $\hat{\beta}_j$ είναι θετικός, ο παράγοντας $e^{\hat{\beta}_j}$ είναι μεγαλύτερος από τη μονάδα, γεγονός που σημαίνει πως το odds $\hat{p}/(1-\hat{p})$ αυξάνεται με την αύξηση της x_j, αντίθετα αν το $\hat{\beta}_j$ είναι αρνητικό, ο παράγοντας $e^{\hat{\beta}_j}$ είναι μικρότερος της μονάδας και το odds μειώνεται με την αύξηση της x_j.

Οι παράμετροι της παλινδρόμησης μπορούν να εκφραστούν και μέσα από το **σχετικό λόγο των συμπληρωματικών πιθανοτήτων**, δηλαδή μέσα από το **λόγο των odds** (odds ratio). Γενικώς ο λόγος των odds ενός ατόμου με τιμές συμμεταβλητών $\mathbf{x}_1$ σε σχέση με ένα άτομο με τιμές $\mathbf{x}_2$ των ίδιων συμμεταβλητών

προκύπτει ως

$$\frac{\hat{p}_1}{1-\hat{p}_1} / \frac{\hat{p}_2}{1-\hat{p}_2} = \frac{\text{odds}\,(y=1|\,\mathbf{x}_1)}{\text{odds}\,(y=1|\,\mathbf{x}_2)} = \frac{\exp(\mathbf{x}_1'\hat{\boldsymbol{\beta}})}{\exp(\mathbf{x}_2'\hat{\boldsymbol{\beta}})} = \exp\{(\mathbf{x}_1-\mathbf{x}_2)'\,\hat{\boldsymbol{\beta}}\}\,.$$

Ειδικότερα, αν θεωρήσουμε ένα μοντέλο με δύο συμμεταβλητές x_1 και x_2, όπου η x_1 είναι μια δείκτρια μεταβλητή με $x_1 = 0,\ 1$, τότε

$$\text{odds}\,\{y=1|x_1=1, x_2\} = \exp\left(\hat{\beta}_0+\hat{\beta}_1+\hat{\beta}_2 x_2\right)$$

$$\text{odds}\,\{y=1|x_1=0, x_2\} = \exp\left(\hat{\beta}_0+\hat{\beta}_2 x_2\right).$$

Συνεπώς ο λόγος των odds (odds ratio) ισούται με $e^{\hat{\beta}_1}$ και είναι ανεξάρτητος της x_2. Αυτό το μοντέλο δείχνει ότι για κάθε τιμή της x_1 οι ευθείες ln (odds) ή logit($\hat{p}$) έχουν την ίδια κλίση, δηλαδή θα είναι παράλληλες (βλ. Παράγραφο 5.4 για τη σύγκριση μεταξύ δύο μοντέλων της πολλαπλής γραμμικής παλινδρόμησης).

Διαστήματα εμπιστοσύνης

Όπως αναφέραμε στις Παραγράφους 8.2.2 και 8.4.2, με βάση τη στατιστική συνάρτηση του Wald μπορούμε να κατασκευάσουμε ένα $100(1-\alpha)\%$–διάστημα εμπιστοσύνης για την παράμετρο β_j, ως $\hat{\beta}_j \pm z_{\alpha/2} se(\hat{\beta}_j)$, $j = 0, 1, 2, \ldots, k$. Από το δ.ε. της β_j μπορούμε να προσδιορίσουμε και ένα $100(1-\alpha)\%$–διάστημα εμπιστοσύνης $\exp\left\{\hat{\beta}_j \pm z_{\alpha/2} se(\hat{\beta}_j)\right\}$ για το λόγο των odds (odds ratio) e^{β_j}.

8.9.2 Ελεγχοσυνάρτηση deviance

Μέσω της σχέσης $\ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i'\boldsymbol{\beta}$ το μοντέλο επιβάλλει μια δομή στα δεδομένα. Χωρίς καμία δομή έχουμε απλώς ανεξάρτητες τιμές από διαφορετικές Διωνυμικές

κατανομές

$$y_i \sim b(n_i, p_i)\,,$$

όπου $\mu_i = n_i p_i$ είναι χωρίς άλλο περιορισμό εκτός του ότι $\mu_i > 0$, και συνεπώς θα εκτιμάται από το $\tilde{\mu}_i = y_i$. Τότε η συμβολή της i–οστής παρατήρησης στη μεγιστοποιημένη τιμή του λογαρίθμου της πιθανοφάνειας, υπό την υπόθεση H_S του χορεσμένου μοντέλου M_S, είναι

$$\tilde{\ell}_{Si} = \ln \binom{n_i}{y_i} + y_i \ln \tilde{p}_i + (n_i - y_i) \ln(1 - \tilde{p}_i)\,.$$

Η αντίστοιχη συμβολή της i–οστής παρατήρησης στη μεγιστοποιημένη τιμή του λογαρίθμου της πιθανοφάνειας, υπό την H_0 του υπό εξέταση μοντέλου M_0, είναι

$$\hat{\ell}_{0i} = \ln \binom{n_i}{y_i} + y_i \ln \hat{p}_i + (n_i - y_i) \ln(1 - \hat{p}_i)\,,$$

(βλ. και Παράγραφο 8.4.1). Η διαφορά των δύο τιμών είναι

$$\begin{aligned}
\hat{\ell}_{0i} - \tilde{\ell}_{Si} &= y_i (\ln \hat{p}_i - \ln \tilde{p}_i) + (n_i - y_i) [\ln(1 - \hat{p}_i) - \ln(1 - \tilde{p}_i)] \\
&= y_i \ln \left(\frac{\hat{p}_i}{\tilde{p}_i} \right) + (n_i - y_i) \ln \left(\frac{1 - \hat{p}_i}{1 - \tilde{p}_i} \right) \\
&= y_i \ln \left(\frac{\hat{\mu}_i}{y_i} \right) + (n_i - y_i) \ln \left(\frac{n_i - \hat{\mu}_i}{n_i - y_i} \right)\,,
\end{aligned}$$

όπου $\tilde{p}_i = \frac{y_i}{n_i}$ και $\hat{p}_i = \frac{\hat{\mu}_i}{n_i}$.

Όπως με το μοντέλο της παλινδρόμησης Poisson (βλ. και Παράδειγμα 8.4.1), ορίζουμε την τυποποιημένη ελεγχοσυνάρτηση deviance ως

$$\begin{aligned}
D(\hat{\boldsymbol{\beta}}) = D(\mathbf{y}; \hat{\boldsymbol{\mu}}) &= -2 \left\{ \hat{\ell}_0 - \tilde{\ell}_S \right\} \\
&= \sum_{i=1}^{n} 2 \left\{ y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) + (n_i - y_i) \ln \left(\frac{n_i - y_i}{n_i - \hat{\mu}_i} \right) \right\} \\
&= \sum_{i=1}^{n} d_i \left(y_i, \hat{\mu}_i \right) = \sum_{i=1}^{n} d_i(\hat{\boldsymbol{\beta}})\,,
\end{aligned}$$

η οποία αποτελεί ένα μέτρο σύγκρισης μεταξύ των παρατηρήσεων y_i και των εκτιμηθέντων $\hat{\mu}_i$ και η οποία ταυτίζεται με τη συνάρτηση deviance, αφού για τη Διωνυμική κατανομή ισχύει $a(\phi) = 1$. Υπενθυμίζουμε ότι η συνάρτηση deviance χρησιμοποιείται κυρίως για τη σύγκριση και ανάπτυξη μοντέλων (βλ. Παραγράφους 8.2.3 και 8.4.1).

Εδώ σημειώνουμε ότι **στην ειδική περίπτωση των δυαδικών δεδομένων**, δηλαδή όταν $n_i = 1$, $\forall i$, η **ελεγχοσυνάρτηση deviance δε μας παρέχει πληροφορίες για την καταλληλότητα ενός μοντέλου** και αυτό διότι εξαρτάται μόνο από τις προσαρμοσμένες ή εκτιμημένες τιμές $\hat{\mu}_i$ ως ακολούθως. Η συνάρτηση πιθανοφάνειας για n δυαδικές παρατηρήσεις είναι

$$L = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i}$$

με $E(y_i) = p_i = \mu_i$ και επομένως η λογαριθμοποιημένη πιθανοφάνεια είναι

$$\ell = \ln L = \sum_{i=1}^{n} \{y_i \ln \mu_i + (1 - y_i) \ln(1 - \mu_i)\} . \tag{8.19}$$

Εκτιμώντας το πλήρες ή κορεσμένο μοντέλο ισχύει $\tilde{\mu}_i = y_i$ και αφού οι όροι $y_i \ln y_i$ και $(1 - y_i) \ln(1 - y_i)$ είναι ίσοι με το μηδέν για τις δύο δυνατές τιμές του y_i, 0 και 1, θα ισχύει $\tilde{\ell}_S = 0$. Επομένως η ελεγχοσυνάρτηση deviance για δυαδικές παρατηρήσεις ($n_i = 1$) δίνεται από τη σχέση

$$\begin{aligned} D(\hat{\boldsymbol{\beta}}) &= -2 \sum_{i=1}^{n} \{y_i \ln \hat{\mu}_i + (1 - y_i) \ln(1 - \hat{\mu}_i)\} \\ &= -2 \sum_{i=1}^{n} \{y_i \ln[\hat{\mu}_i/(1 - \hat{\mu}_i)] + \ln(1 - \hat{\mu}_i)\} . \end{aligned} \tag{8.20}$$

Για το υπό εξέταση μοντέλο, όπου $n_i = 1$, η λογαριθμοποιημένη συνάρτηση

πιθανοφάνειας των σχέσεων (8.18) και (8.19) γράφεται ως

$$\ell = \ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i \mathbf{x}_i' \boldsymbol{\beta} - \ln(1 + e^{\mathbf{x}_i' \boldsymbol{\beta}}) \right\}$$

και παραγωγίζοντας ως προς τις παραμέτρους β_j έχουμε ότι

$$\frac{\partial \ln L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - p_i) x_{ij} = \sum_{i=1}^{n} (y_i - \mu_i) x_{ij} \,,$$

από την οποία συνεπάγεται ότι

$$\begin{aligned} \sum_{j=0}^{k} \beta_j \frac{\partial \ln L(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^{n} (y_i - \mu_i) \sum_{j=0}^{k} \beta_j x_{ij} \\ &= \sum_{i=1}^{n} (y_i - \mu_i) \ln\{\mu_i/(1-\mu_i)\} \,, \end{aligned}$$

όπου $\mu_i = p_i = \frac{e^{\mathbf{x}_i' \boldsymbol{\beta}}}{1+e^{\mathbf{x}_i' \boldsymbol{\beta}}}$. Επειδή η $\hat{\boldsymbol{\beta}}$ είναι η εκτιμήτρια μέγιστης πιθανοφάνειας της $\boldsymbol{\beta}$, η παράγωγος στην αριστερή πλευρά της εξίσωσης μηδενίζεται στο $\hat{\boldsymbol{\beta}}$. Επομένως οι προσαρμοσμένες πιθανότητες $\hat{\mu}_i = \hat{p}_i$ πρέπει να ικανοποιούν την εξίσωση

$$\sum_{i=1}^{n} (y_i - \hat{\mu}_i) \text{logit}(\hat{\mu}_i) = 0$$

και άρα

$$\sum_{i=1}^{n} y_i \text{logit}(\hat{\mu}_i) = \sum_{i=1}^{n} \hat{\mu}_i \text{logit}(\hat{\mu}_i) \,.$$

Τέλος, αντικαθιστώντας την $\sum_{i=1}^{n} y_i \text{logit}(\hat{\mu}_i)$ στη σχέση (8.20) για την $D(\hat{\boldsymbol{\beta}})$ λαμβάνουμε την τελική έκφραση για την ελεγχοσυνάρτηση deviance

$$D(\hat{\boldsymbol{\beta}}) = -2 \sum_{i=1}^{n} \{ \hat{\mu}_i \text{logit}(\hat{\mu}_i) + \ln(1 - \hat{\mu}_i) \} \,,$$

η οποία εξαρτάται μόνο από τις προσαρμοσμένες τιμές $\hat{\mu}_i$ και όχι άμεσα από τις

παρατηρήσεις y_i. Συνεπώς δεν μπορεί να κριθεί η καλή προσαρμογή του μοντέλου. Σε αυτήν την περίπτωση των δυαδικών παρατηρήσεων, όπου όλα τα $n_i = 1$, μπορεί να δειχθεί ότι η deviance δεν ακολουθεί την κατανομή χ^2 ούτε προσεγγιστικά. Ωστόσο **μεταβολές** της deviance εξακολουθούν να είναι της κατανομής χ^2.

8.9.3 χ^2–έλεγχοι καλής προσαρμογής

Για τη λογιστική παλινδρόμηση έχουν αναπτυχθεί και οι ακόλουθοι έλεγχοι καλής προσαρμογής, οι οποίοι παρουσιάζονται στη συνέχεια.

Pearson

Για τη μελέτη της καταλληλότητας του μοντέλου εκτός από την ελεγχοσυνάρτηση deviance χρησιμοποιείται και η ελεγχοσυνάρτηση Pearson, που δίνεται από τον τύπο

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i\,(1-\hat{p}_i)}\,.$$

Η ελεγχοσυνάρτηση deviance καθώς και αυτή του Pearson είναι ασυμπτωτικά ισοδύναμες και ακολουθούν την ίδια κατανομή χ^2, όταν το προσαρμοσμένο μοντέλο είναι το ορθό (βλ. Άσκηση 7). Οι τιμές τους γενικά διαφέρουν, ωστόσο σπάνια στο βαθμό που να οδηγούν σε διαφορετικά συμπεράσματα. Μεγάλες διαφορές μεταξύ των δύο στατιστικών συναρτήσεων μπορεί να θεωρηθεί ως ένδειξη ότι για τη μία από τις δύο αυτές συναρτήσεις η προσέγγιση της κατανομής χ^2 δεν είναι ικανοποιητική. Βέβαια η deviance έχει το πλεονέκτημα ότι μπορεί να χρησιμοποιηθεί για τη σύγκριση μεταξύ δύο εμφωλευμένων μοντέλων, αφού από τη διαφορά στη συνάρτηση deviance μεταξύ των δύο μοντέλων αξιολογείται η σημαντικότητα των επιπρόσθετων όρων. Αντίθετα η ελεγχοσυνάρτηση του Pearson δεν μπορεί να χρησιμοποιηθεί με αυτόν τον τρόπο.

Αν κάθε παρατήρηση είναι έτσι ώστε $y = 0$ ή $y = 1$ με $n_i = 1$, τότε οι ελεγχοσυναρτήσεις deviance (βλ. προηγούμενη Παράγραφο 8.9.2) και του Pearson δεν είναι χρήσιμες, διότι δε θα ισχύει η ασυμπτωτική θεωρία. Σε αυτήν την περίπτω-

ση, μεταξύ εναλλακτικών ελέγχων, προτείνεται η χρήση του ακόλουθου ελέγχου των Hosmer–Lemeshow.

Hosmer–Lemeshow

Σε αντίθεση με την ελεγχοσυνάρτηση deviance ο έλεγχος των Hosmer and Lemeshow (1980) είναι ένα μέτρο καταλληλότητας του μοντέλου που μπορεί να χρησιμοποιηθεί αρχικά σε μη ομαδοποιημένα δυαδικά δεδομένα ($n_i = 1$). Στη συνέχεια όμως οι παρατηρήσεις ομαδοποιούνται σύμφωνα με τις εκτιμημένες πιθανότητες. Πιο συγκεκριμένα, για να υπολογιστεί ο έλεγχος αυτός, οι παρατηρήσεις διατάσσονται κατά αύξουσα σειρά σύμφωνα με την τιμή της εκτιμημένης πιθανότητας $\hat{p}_i$ και χωρίζονται σε ομάδες του ίδιου περίπου αριθμού παρατηρήσεων. (Για παράδειγμα, τα γνωστά στατιστικά πακέτα SPSS και MINITAB δημιουργούν 10 ομάδες για αυτόν τον έλεγχο). Ας υποθέσουμε ότι στην i–οστή από τις g συνολικά ομάδες υπάρχουν m_i παρατηρήσεις, όπου ο συνολικός αριθμός επιτυχιών είναι o_i, και ο αντίστοιχος αναμενόμενος αριθμός επιτυχιών είναι e_i. (Οι συχνότητες o_i και e_i προκύπτουν από το άθροισμα των y_j και $\hat{\mu}_j$ των $j = 1, \ldots, m_i$ παρατηρήσεων της ομάδας i αντίστοιχα).

Τότε μπορούμε να εφαρμόσουμε ένα έλεγχο χ^2 του Pearson στον $g \times 2$ πίνακα συνάφειας και έτσι διαμορφώνεται ο έλεγχος χ^2 καλής προσαρμογής των Hosmer–Lemeshow, ο οποίος δίνεται από τον τύπο

$$X^2_{HL} = \sum_{i=1}^{g} \frac{(o_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)},$$

όπου $\hat{\pi}_i = \frac{e_i}{m_i}$ η μέση πιθανότητα επιτυχίας της i–οστής ομάδας. Από προσομοιώσεις έχει βρεθεί ότι ο έλεγχος αυτός ακολουθεί προσεγγιστικά την κατανομή χ^2 με $(g-2)$ βαθμούς ελευθερίας. Ωστόσο, αφού η τιμή της ελεγχοσυνάρτησης εξαρτάται από το χωρισμό των παρατηρήσεων σε ομάδες και από τον αριθμό τους σε καθεμία από αυτές, θεωρείται ως ένα ανεπίσημο μέτρο αξιολόγησης προσαρμογής του μοντέλου.

8.9.4 Υπόλοιπα

Υπόλοιπα Pearson

Επειδή το μοντέλο της λογιστικής παλινδρόμησης είναι ένα γενικευμένο γραμμικό μοντέλο, τα υπόλοιπα ορίζονται όπως και στη γενική περίπτωση. Επειδή για το μοντέλο αυτό $\mathrm{v}(\hat{\mu}_i) = \widehat{V}(y_i) = n_i\hat{p}_i(1-\hat{p}_i)$, το υπόλοιπο Pearson δίνεται ως

$$r_i^P = \frac{y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}}\,,\ i = 1, \ldots, n\,,$$

(βλ. Παραγράφους 8.2.4, 8.3, 8.5.1 και 8.9.1).

Τα τυποποιημένα υπόλοιπα Pearson ορίζονται μέσω της σχέσης

$$r_i^{PS} = \frac{y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1-\hat{p}_i)(1-\hat{h}_{ii})}} = \frac{r_i^P}{\sqrt{1-\hat{h}_{ii}}}\,,$$

όπου $\hat{h}_{ii}$ είναι το i–οστό διαγώνιο στοιχείο του $n \times n$ πίνακα

$$\widehat{\mathbf{H}} = \widehat{\mathbf{W}}^{1/2}\mathbf{X}(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\widehat{\mathbf{W}}^{1/2},$$

$\mathbf{X}$ ο $n \times p$ πίνακας σχεδιασμού και $\widehat{\mathbf{W}}$ ο $n \times n$ διαγώνιος πίνακας, με i–οστό στοιχείο το $n_i\hat{p}_i(1-\hat{p}_i)$, που αποτελεί την εκτιμημένη διασπορά $\widehat{V}(y_i)$ της απόκρισης y_i (βλ. Παράγραφο 8.4, σελίδα 368). Επίσης $E(r_i^{PS}) \simeq 0$ και $V(r_i^{PS}) \simeq 1$, ωστόσο η κατανομή αυτών των υπολοίπων δεν προσεγγίζεται καλά από την Κανονική.

Υπόλοιπο deviance

Όπως είδαμε στις Παραγράφους 8.2.4, 8.4.1 και 8.5.1 το υπόλοιπο deviance προκύπτει από την τυποποιημένη συνάρτηση deviance

$$D(\hat{\boldsymbol{\beta}}) = D\left(\mathbf{y}; \hat{\boldsymbol{\mu}}\right) = \sum_{i=1}^{n} d_i(y_i, \hat{\mu}_i) = \sum_{i=1}^{n} d_i(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \left(r_i^D\right)^2$$

και αποτελεί την προσημασμένη τετραγωνική ρίζα της συμβολής της i–οστής παρατήρησης σε αυτήν, η οποία δίνεται από τη σχέση

$$r_i^D = \text{sgn}(y_i - \hat{\mu}_i) \left\{ 2y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + 2\left(n_i - y_i\right) \ln\left(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\right) \right\}^{1/2}.$$

Το τυποποιημένο υπόλοιπο deviance ορίζεται ως $r_i^{DS} = r_i^D / \sqrt{1 - \hat{h}_{ii}}$.

Αν τα δεδομένα είναι σε δυαδική μορφή, το υπόλοιπο deviance ανάγεται στο

$$r_i^D = \text{sgn}(y_i - \hat{\mu}_i) \left\{ -2 \left[y_i \ln \hat{\mu}_i + (1 - y_i) \ln\left(1 - \hat{\mu}_i\right) \right] \right\}^{1/2},$$

όπου τώρα $\hat{\mu}_i = \hat{p}_i$, αφού $n_i = 1$ (βλ. και Παράγραφο 8.9.2).

Στην περίπτωση των δυαδικών δεδομένων η κατανομή των υπολοίπων Pearson και deviance (και των αντίστοιχων τυποποιημένων), δεν προσεγγίζεται από την Κανονική, ακόμα κι αν το προσαρμοσμένο μοντέλο είναι το σωστό. Ωστόσο για τα υπόλοιπα αυτά, γραφήματα δείκτη (index plots) μπορούν να φανούν χρήσιμα για την εξέταση της καταλληλότητας ενός μοντέλου, καθώς επίσης και γραφήματα της Κανονικής και ημι-Κανονικής κατανομής (half-normal plots), ενδεχομένως σε συνδυασμό με προσομοιωμένα διαστήματα εμπιστοσύνης για καλύτερη ερμηνεία (βλ. Παράγραφο 8.5.1 και Παράρτημα Δ).

Υπόλοιπο πιθανοφάνειας

Στο γενικό κεφάλαιο των γενικευμένων γραμμικών μοντέλων παρουσιάστηκαν τα υπόλοιπα πιθανοφάνειας που προσδιορίζονται από τις μεταβολές στη deviance αφαιρώντας την κάθε παρατήρηση με τη σειρά της. Η μεταβολή αυτή μπορεί να προσεγγιστεί από το σταθμισμένο συνδυασμό των παραπάνω υπολοίπων $\hat{h}_{ii}\left(r_i^{PS}\right)^2 + \left(1 - \hat{h}_{ii}\right)\left(r_i^{DS}\right)^2 = \left(r_i^L\right)^2$. Ομοίως με το υπόλοιπο deviance r_i^D, το υπόλοιπο πιθανοφάνειας r_i^L παίρνει το πρόσημο $\text{sgn}(y_i - \hat{\mu}_i)$ (βλ. και Παράγραφο 8.5.1). Στο MINITAB 17, το υπόλοιπο r_i^L καλείται deleted deviance residual.

8.9.5 Κριτήρια επιλογής μοντέλου της λογιστικής παλινδρόμησης

Όπως και στη γενική περίπτωση των γενικευμένων γραμμικών μοντέλων, έτσι και στη λογιστική περίπτωση μπορούν να χρησιμοποιηθούν τα κριτήρια AIC και BIC καθώς και οι διάφοροι συντελεστές προσδιορισμού για την επιλογή του βέλ-

τιστου μοντέλου. Οι δείκτες αυτοί για την επιλογή ενός μοντέλου στη λογιστική παλινδρόμηση παρουσιάζονται στη συνέχεια.

Κριτήρια AIC, BIC

Στην περίπτωση της λογιστικής παλινδρόμησης τα κριτήρια AIC και BIC παίρνουν τη μορφή

$$\begin{aligned} AIC &= -2\sum_{i=1}^{n}\left[\ln\binom{n_i}{y_i} + y_i \ln \hat{p}_i + (n_i - y_i)\ln(1-\hat{p}_i)\right] + 2p \\ &= 2\sum_{i=1}^{n}\left[n_i \ln\left(1+e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}}\right) - y_i \mathbf{x}_i'\hat{\boldsymbol{\beta}} - \ln\binom{n_i}{y_i}\right] + 2p \end{aligned}$$

και

$$BIC = 2\sum_{i=1}^{n}\left[n_i \ln\left(1+e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}}\right) - y_i \mathbf{x}_i'\hat{\boldsymbol{\beta}} - \ln\binom{n_i}{y_i}\right] + p\ln n\,.$$

Οι μικρότερες τιμές υποδεικνύουν το καλύτερο μοντέλο. Παρατηρήσαμε νωρίτερα ότι συχνά στα στατιστικά πακέτα παραλείπονται οι σταθεροί όροι. Στη λογιστική παλινδρόμηση ο όρος αυτός είναι ο $\ln\binom{n_i}{y_i}$.

Κριτήρια R^2

Όπως αναφέραμε στην Παράγραφο 8.6.2 τα περισσότερα υπολογιστικά προγράμματα για γενικευμένα γραμμικά μοντέλα δίνουν έμφαση στα κριτήρια AIC και BIC, και γενικώς δεν υπολογίζουν δείκτες τύπου R^2. Βέβαια μπορούμε χωρίς δυσκολία να τους βρούμε από τις τιμές της μεγιστοποιημένης λογαριθμοποιημένης πιθανοφάνειας. Τα πιο γνωστά μέτρα είναι ο ψευδο–R^2_M της Παραγράφου 8.6.2

$$R^2_M = 1 - \left(\frac{\hat{L}_0}{\hat{L}_1}\right)^{2/m},$$

όπου $m = \sum_{i=1}^{n} n_i$ και ο διορθωμένος συντελεστής του R^2_M,

$$R^2_N = \frac{R^2_M}{\max R^2_M}$$

του Nagelkerke. Βλ. περισσότερα σχόλια στο τέλος του Παραδείγματος 8.9.1.

Τα δύο αυτά ψευδο–R^2 μέτρα καθώς και άλλα, που είδαμε στην Παράγραφο 8.6.2, συχνά παρουσιάζουν χαμηλές τιμές στη λογιστική παλινδρόμηση – ιδιαίτερα στην περίπτωση των δυαδικών δεδομένων – σε σύγκριση με αυτά που συνηθίζονται στα γραμμικά μοντέλα, παρότι άλλοι δείκτες προτείνουν την καλή προσαρμογή του μοντέλου. Αυτό συμβαίνει, διότι το μοντέλο εξηγεί ή προβλέπει μόνο την πιθανότητα επιτυχίας $p = E(Y)$ και όχι τις ατομικές τιμές y (0 ή 1). Δεδομένης της p, η επιτυχία ή αποτυχία είναι τυχαίο γεγονός που το μοντέλο δεν μπορεί να προβλέψει. Επομένως μεγάλο μέρος της συνολικής μεταβλητότητας των δεδομένων δεν μπορεί να εξηγηθεί, άρα ένας δείκτης τύπου R^2 αναγκαστικά παίρνει χαμηλή τιμή. Παρόμοιο πρόβλημα αναφέρθηκε και στη γραμμική παλινδρόμηση, όταν οι ίδιες τιμές των μεταβλητών $\mathbf{x}$ επαναλαμβάνονται (βλ. Παράγραφο 3.5).

8.9.6 Χρήση R – Παράδειγμα

Για την προσαρμογή ενός μοντέλου λογιστικής παλινδρόμησης αλλά και την εκτέλεση όλων των προαναφερθέντων ελέγχων θα βασιστούμε στην R.

Παράδειγμα 8.9.1. *Στον Πίνακα 8.3 παρουσιάζονται τα αποτελέσματα μιας δοκιμής ενός φαρμάκου σε ποντίκια. Χρησιμοποιήθηκαν δύο διαφορετικές μέθοδοι παρασκευής του φαρμάκου (PR=1 ή 2) και διάφορες δόσεις (DOSE). Κάθε συνδυασμός δόσης και μεθόδου παρασκευής φαρμάκου δοκιμάστηκε σε μια ομάδα ποντικιών. Το μέγεθος της ομάδας είναι το στοιχείο* Trials *του πίνακα. Για κάθε ομάδα καταγράφηκε ο αριθμός (Events) των ποντικιών με αρνητική αντίδραση. Για παράδειγμα ($3^η$ γραμμή δεδομένων), το φάρμακο που παρασκευάστηκε με την πρώτη μέθοδο με δόση των 7 μονάδων δοκιμάστηκε σε 38 ποντίκια, τα 11 εκ των οποίων αντέδρασαν αρνητικά. Με το συμβολισμό των προηγούμενων παραγράφων, $y_3 = 11$, ο αριθμός επιτυχιών σε $n_3 = 38$ δοκιμές. Ζητείται να ερευνήσουμε την ύπαρξη της εξάρτησης της πιθανότητας ένα ποντίκι να παρουσιάσει αρνητική αντίδραση από τη δόση και τη μέθοδο παρασκευής του φαρμάκου.*

Πίνακας 8.3: Πίνακας δεδομένων για το Παράδειγμα 8.9.1

id	1	2	3	4	5	6	7	8	9	10	11	12	13	14
PR	1	1	1	1	1	1	1	1	1	2	2	2	2	2
DOSE	3.4	5.2	7	8.5	10.5	13	18	21	28	6.5	10	14	21.5	29
Events	0	5	11	14	18	21	23	30	27	2	10	18	21	27
Trials	33	32	38	37	40	37	31	37	30	40	30	40	35	37

Θεωρούμε ότι η τ.μ. Events *που εκφράζει τον αριθμό των ποντικιών με αρνητική αντίδραση ακολουθεί τη Διωνυμική κατανομή. Η ενδεχόμενη εξάρτησή της από τις επεξηγηματικές μεταβλητές* PR *και* DOSE *εισάγεται μέσω της εξάρτησης της* $p = P(\text{αρνητική αντίδραση})$ *από τις επεξηγηματικές μεταβλητές με τη βοήθεια ενός μοντέλου λογιστικής παλινδρόμησης.*

Ορίζουμε τη δείκτρια μεταβλητή $drug$

$$drug = \begin{cases} 1\,, & \alpha\nu\ PR = 1 \\ 0\,, & \alpha\nu\ PR = 2\,. \end{cases}$$

Εισάγοντας τις ακόλουθες εντολές

```
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
sf <- cbind(Events, Trials-Events)
fit <- glm(sf ~ drug + DOSE, family=binomial)
summary(fit)
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

στην R *λαμβάνουμε τα Αποτελέσματα 8.2 για την προσαρμογή του μοντέλου λογιστικής παλινδρόμησης με τη συνάρτηση σύνδεσης* logit. *Υπενθυμίζεται ότι στην* R, Residual deviance *είναι η* deviance *του μοντέλου που προσαρμόζεται,*

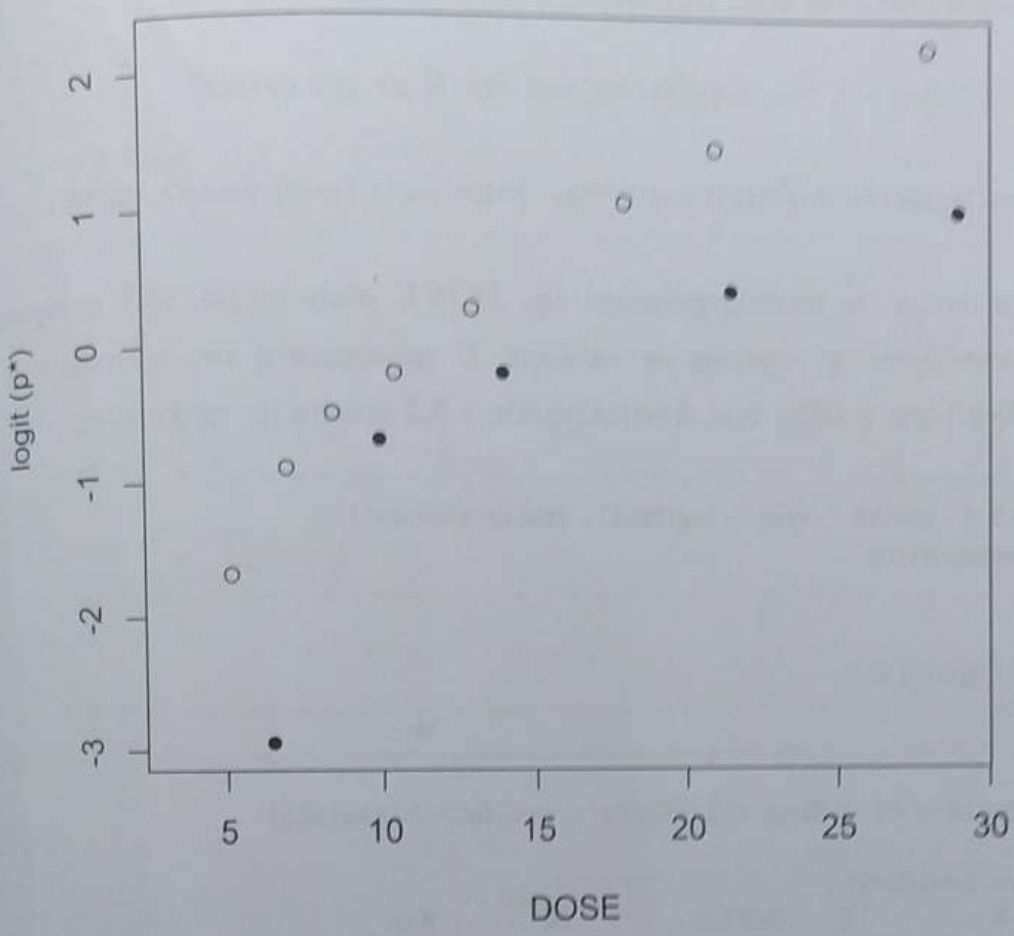
ενώ Null deviance είναι η deviance του μοντέλου που περιέχει μόνο το σταθερό όρο.

Αποτελέσματα 8.2

```
Call:
glm(formula = sf ~ drug + DOSE, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5967  -0.8187   0.2498   0.6074   1.6034

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.95352    0.31953  -9.243  < 2e-16 ***
drug         0.87525    0.23393   3.742 0.000183 ***
DOSE         0.16126    0.01601  10.069  < 2e-16 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 166.834  on 13  degrees of freedom
Residual deviance:  27.098  on 11  degrees of freedom
AIC: 80.951

Number of Fisher Scoring iterations: 4
```

Από τα Αποτελέσματα 8.2 παρατηρούμε από τις p–τιμές των ελέγχων Wald ότι και οι δύο επεξηγηματικές μεταβλητές είναι στατιστικά σημαντικές με $p < 0.001$. Από την άλλη όμως το ίδιο το μοντέλο δεν μπορεί να θεωρηθεί ως ένα καλό εναλλακτικό μοντέλο έναντι του κορεσμένου μοντέλου, αφού η p–τιμή του ελέγχου για τη σημαντικότητά του ισούται με $P(X^2 > 27.098) = 0.0044$ της κατανομής χ^2 με 11 β.ε. (Νωρίτερα εκφράσαμε επιφυλάξεις ως προς την ακρίβεια προσέγγισης μιας τιμής της deviance από την κατανομή χ^2. Ωστόσο εδώ δεν μπορούμε να αγνοήσουμε την p–τιμή, η οποία είναι πολύ μικρή). Η τιμή αυτή υπολογίζεται από την R με την εντολή

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



Σχήμα 8.4: Logit(p^*) σε σχέση με τη δόση φαρμάκου για τα δεδομένα του Παραδείγματος 8.9.1 (μαύροι κύκλοι, $drug = 0$: ανοιχτοί κύκλοι, $drug = 1$)

Βασικός λόγος για την αδυναμία του μοντέλου είναι η ανάγκη μετασχηματισμού της μεταβλητής DOSE. Αυτό γίνεται πιο κατανοητό από το Σχήμα 8.4, όπου παρουσιάζονται οι τιμές του $\text{logit}(p^*) = \ln \frac{p^*}{1-p^*}$ *έναντι των τιμών της DOSE, όπου* $p^* = y/n = Events/Trials$*, το ποσοστό των ποντικιών με αρνητική αντίδραση ανά ομάδα. Το Σχήμα 8.4 λαμβάνεται από την R με τις εντολές*

```
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
pstar <- Events/Trials

plot(log(pstar/(1-pstar))~DOSE,  ylab="logit (p*)", pch=c(1,16)[PR])
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```

Είναι φανερό ότι το logit δε σχετίζεται γραμμικά με τη δόση (βλ. ενότητα «Ερμηνεία των συντελεστών β» της Παραγράφου 8.9.1). Ο μετασχηματισμός που προτείνεται εδώ είναι ο $\ln(DOSE)$.

Από το Σχήμα 8.5, που λαμβάνεται από την R με την εντολή

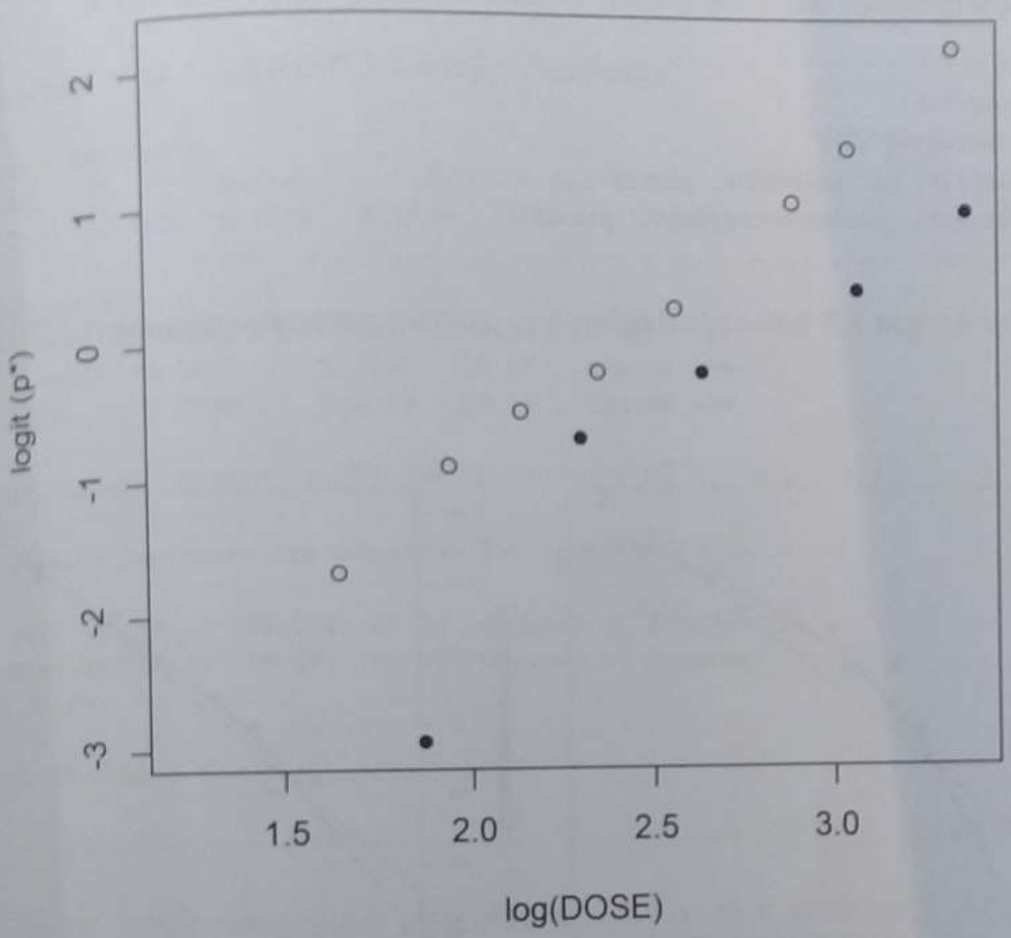
```
plot(log(pstar/(1-pstar))~log(DOSE), ylab="logit (p*)", pch=c(1,16)[PR])
```

φαίνεται ότι με το μετασχηματισμό της $DOSE$ *στην* $\ln(DOSE)$ *πετύχαμε τη γραμμικοποίηση της σχέσης με το logit. Η προσαρμογή του μοντέλου με την* $\ln(DOSE)$ *και η λήψη των Αποτελεσμάτων 8.3 γίνεται με τις εντολές*

```
fit2 <- glm(sf ~ drug + log(DOSE), family=binomial)
summary(fit2)
```

Αποτελέσματα 8.3

```
Call:
glm(formula = sf ~ drug + log(DOSE), family = binomial)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.03026  -0.39653  -0.01947   0.34974   1.24987

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.4822     0.6276  -10.33  < 2e-16 ***
drug          0.9290     0.2334    3.98 6.89e-05 ***
log(DOSE)     2.2972     0.2196   10.46  < 2e-16 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 166.8335  on 13  degrees of freedom
Residual deviance:   8.7912  on 11  degrees of freedom
AIC: 62.644

Number of Fisher Scoring iterations: 4
```



Σχήμα 8.5: Logit(p^*) σε σχέση με το λογάριθμο της δόσης για τα δεδομένα του Παραδείγματος 8.9.1 (μαύροι κύκλοι, $drug = 0$: ανοιχτοί κύκλοι, $drug = 1$)

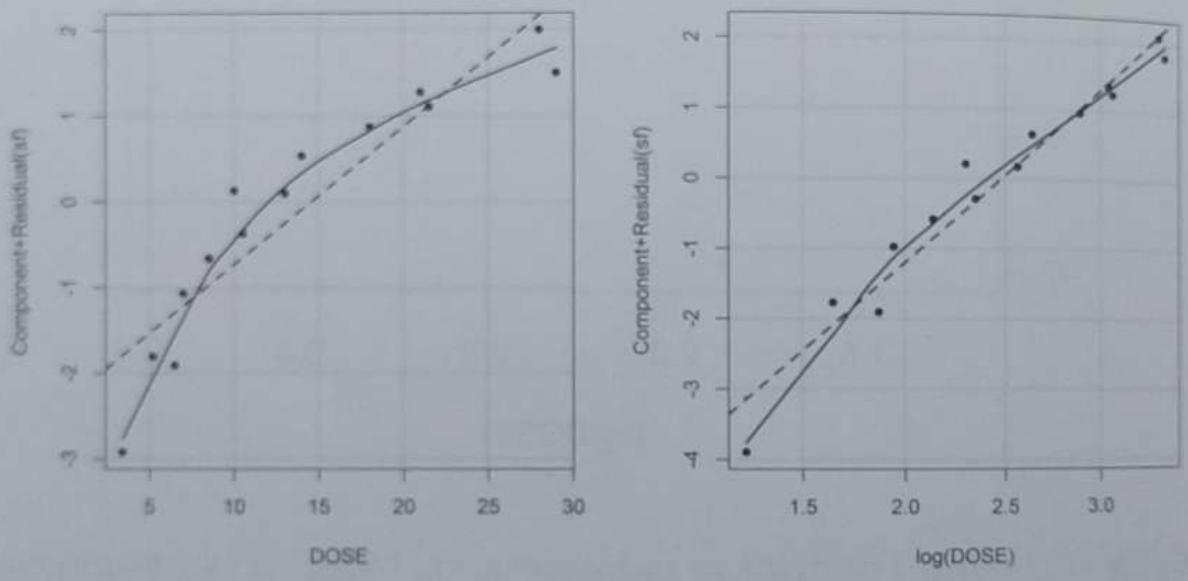
Η σημαντικότητα των συμμεταβλητών δεν επηρεάστηκε από τον προτεινόμενο μετασχηματισμό. Επιπρόσθετα, η τιμή της deviance είναι αισθητά μικρότερη σε σχέση με την τιμή της deviance στα Αποτελέσματα 8.2. Η τιμή αυτή μάλιστα είναι τέτοια ώστε το μοντέλο αυτό να είναι ένα καλό εναλλακτικό του κορεσμένου μοντέλου, αφού η p–τιμή του ελέγχου για τη deviance ισούται με $P(X^2 > 8.79) = 0.641$, όπως μπορούμε να διαπιστώσουμε με την εντολή

```
1-pchisq(fit2$deviance,fit2$df.residual)
```

Πάλι για την εξέταση τη συναρτησιακής σχέσης της μεταβλητής DOSE *στο μοντέλο, μπορούμε να κάνουμε χρήση των διαγραμμάτων των μερικών υπολοίπων (βλ. Παραγράφους 5.3.2 και 8.6.3). Οι ακόλουθες εντολές από την R*

```
library(car)
par (mfrow=c(1,2))
crPlot(fit, variable=DOSE, pch=19)
crPlot(fit2, variable=log(DOSE), pch=19)
```

δίνουν το Σχήμα 8.6 που επιβεβαιώνει την ανάγκη μετασχηματισμού της DOSE.



Σχήμα 8.6: Διαγράμματα μερικών υπολοίπων πριν και μετά το μετασχηματισμό της μεταβλητής *DOSE* και δεδομένου ότι η μεταβλητή *drug* είναι στο μοντέλο

Αν και είναι φανερή η αναγκαιότητα διατήρησης της μεταβλητής drug στο μοντέλο, θα παρουσιάσουμε δύο διαφορετικούς τρόπους εξέτασης της αναγκαιότητας αφαίρεσης ή διατήρησης μιας μεταβλητής στο μοντέλο. Η προσαρμογή του μοντέλου χωρίς την drug γίνεται με τις εντολές

```
fit3 <- glm(sf ~ log(DOSE), family=binomial)
summary(fit3)
```

με τις οποίες λαμβάνουμε τα Αποτελέσματα 8.4.

Αποτελέσματα 8.4

```
Call:
glm(formula = sf ~ log(DOSE), family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5717  -1.5203   0.6578   1.0829   1.1951

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.3617     0.5305  -10.11   <2e-16 ***
log(DOSE)     2.0820     0.2059   10.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 166.834  on 13  degrees of freedom
Residual deviance:  25.577  on 12  degrees of freedom
AIC: 77.43

Number of Fisher Scoring iterations: 4
```

Ο πρώτος τρόπος σύγκρισης των δύο μοντέλων είναι ο δείκτης AIC. Όπως έχουμε αναφέρει, το βέλτιστο μοντέλο με βάση αυτό το κριτήριο είναι το μοντέλο με τη μικρότερη τιμή AIC. Αυτό είναι το μοντέλο με τις μεταβλητές $\ln(DOSE)$ και drug, με $AIC = 62.644$ (Αποτελέσματα 8.3) καθώς η τιμή του δείκτη AIC για το μοντέλο μόνο με την $\ln(DOSE)$ είναι 77.43 (Αποτελέσματα 8.4).

Ο δεύτερος τρόπος είναι με τη σύγκριση των τιμών της ελεγχοσυνάρτησης deviance. Για τη σύγκριση των δύο αυτών εμφωλευμένων μοντέλων εκτελούμε την εντολή

```
anova(fit3,fit2, test = "Chisq")
```

με την οποία λαμβάνουμε τα Αποτελέσματα 8.5, στα οποία παρουσιάζεται

ο πίνακας ανάλυσης της deviance για τα δύο μοντέλα.

Αποτελέσματα 8.5

```
Analysis of Deviance Table

Model 1: sf ~ log(DOSE)
Model 2: sf ~ drug + log(DOSE)
  Resid. Df Resid. Dev Df Deviance P(> Chi)
1        12    25.5773
2        11     8.7912  1   16.786 4.184e-05 ***
---
Signif. codes:  0 "***" 0.001 "**" 0.01 "*" 0.05 "." 0.1 " " 1
```

Από την τιμή του ελέγχου παρατηρούμε ότι απορρίπτεται το πιο απλό μοντέλο (αυτό χωρίς την drug), διότι η πρόσθεση της $DOSE$ μειώνει σημαντικά τη deviance. Με παρόμοια διαδικασία επιβεβαιώνεται ότι η μεταβλητή $\ln(DOSE)$ δεν πρέπει να αφαιρεθεί από το μοντέλο.

Άρα το βέλτιστο μοντέλο είναι το μοντέλο με τις μεταβλητές drug και $\ln(DOSE)$, για το οποίο η προσαρμοσμένη εξίσωση παλινδρόμησης των Αποτελεσμάτων 8.3 είναι η

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = -6.4822 + 0.9290\, drug + 2.2972\, \ln(DOSE),$$

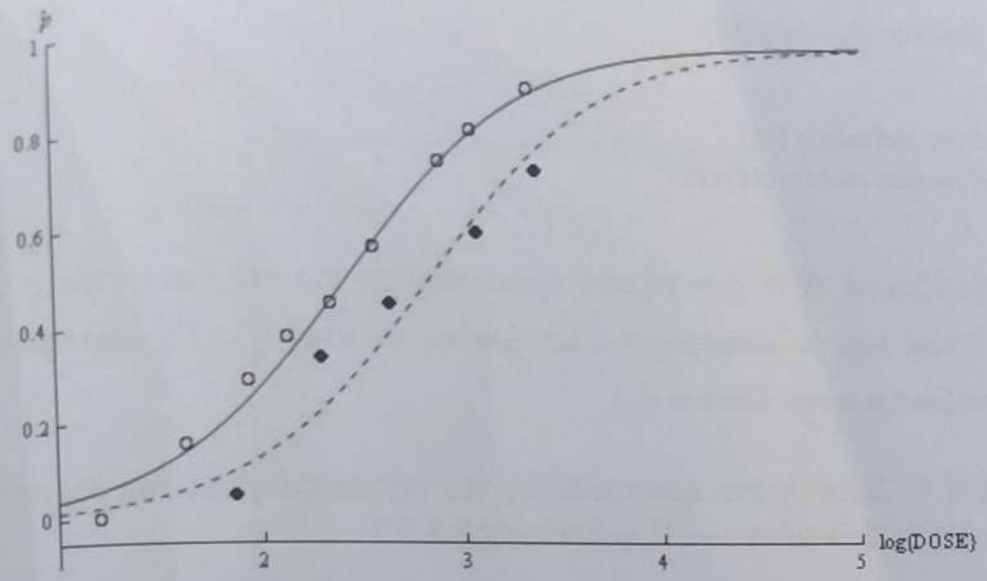
όπου $\hat{p}$ η εκτιμημένη τιμή της πιθανότητας αρνητικής αντίδρασης ενός ποντικιού.

Η προηγούμενη σχέση μπορεί να εκφραστεί ισοδύναμα ως

$$\hat{p} = \frac{\exp(-6.4822 + 0.9290\, drug + 2.2972\, \ln(DOSE))}{1 + \exp(-6.4822 + 0.9290\, drug + 2.2972\, \ln(DOSE))},$$

και ο αναμενόμενος αριθμός αρνητικών αντιδράσεων (Events) σε ομάδα n_i ποντικιών (Trials) δίνεται από την $\hat{\mu}_i = n_i \hat{p}_i$.

Οι εκτιμώμενες καμπύλες για την πιθανότητα αρνητικής αντίδρασης $\hat{p}$ για τις δύο μεθόδους παρασκευής των φαρμάκων παρουσιάζονται στο Σχήμα 8.7.



Σχήμα 8.7: Προσαρμοσμένες τιμές $\hat{p}$ από το μοντέλο με $\ln(DOSE)$ και $drug$ στα δεδομένα του Παραδείγματος 8.9.1 (μαύροι κύκλοι, $drug = 0$: ανοιχτοί κύκλοι, $drug = 1$)

Η ποσότητα $\exp(\hat{\beta}_j)$ *είναι ο παράγοντας επί τον οποίο πολλαπλασιάζεται το* odds *(ο λόγος συμπληρωματικών πιθανοτήτων) πραγματοποίησης του γεγονότος (εδώ της αρνητικής αντίδρασης), όταν η ανεξάρτητη μεταβλητή* x_j *αυξηθεί κατά μία μονάδα, με δεδομένο πάντα ότι οι υπόλοιπες μεταβλητές παραμένουν σταθερές. Αποτελεί το λόγο δύο* odds *(το ένα υπολογισμένο στο* $x_j + 1$*, το δε άλλο στο* x_j*), δηλαδή το* odds ratio *της Παραγράφου 8.9.1. Αν το* $\hat{\beta}_j$ *είναι θετικό, ο παράγοντας* $\exp(\hat{\beta}_j)$ *είναι μεγαλύτερος από τη μονάδα, γεγονός που σημαίνει πως ο λόγος των συμπληρωματικών πιθανοτήτων (*odds*) αυξάνεται ενώ αντίθετα, αν το* $\hat{\beta}_j$ *είναι αρνητικό, ο παράγοντας* $\exp(\hat{\beta}_j)$ *είναι μικρότερος της μονάδας και ο λόγος συμπληρωματικών πιθανοτήτων μειώνεται.*

Για το μοντέλο του παραδείγματος είναι φανερό ότι ανάμεσα στις δύο μεθόδους παρασκευής του φαρμάκου υπάρχει διαφορά με το λόγο των συμπληρωματικών πιθανοτήτων να μειώνεται για τη δεύτερη μέθοδο. Επειδή ο συντελεστής της μεταβλητής $\ln(DOSE)$ *είναι θετικός, ο λόγος των συμπληρωματικών πιθανοτήτων αυξάνεται καθώς αυξάνεται η τιμή της μεταβλητής* $DOSE$*, άρα και της* $\ln(DOSE)$.

Με τις ακόλουθες εντολές

```
confint.default(fit2)
exp(confint.default(fit2))
```

κατασκευάζονται 95% διαστήματα εμπιστοσύνης για τις παραμέτρους του μοντέλου αλλά και τα αντίστοιχα διαστήματα των $\exp(\beta_j)$. *Τα διαστήματα αυτά παρουσιάζονται στον Πίνακα 8.4.*

Πίνακας 8.4: Διαστήματα εμπιστοσύνης για τις παραμέτρους του βέλτιστου μοντέλου για τα δεδομένα του Παραδείγματος 8.9.1

	β_j		$\exp(\beta_j)$	
drug	0.472	1.387	1.602	4.001
$\ln(DOSE)$	1.867	2.728	6.467	15.296

Ολοκληρώνοντας το παράδειγμα, παρουσιάζουμε την εφαρμογή των διαγνωστικών ελέγχων για το βέλτιστο μοντέλο, οι οποίες βασίζονται στα υπόλοιπα και στα μέτρα επιρροής.

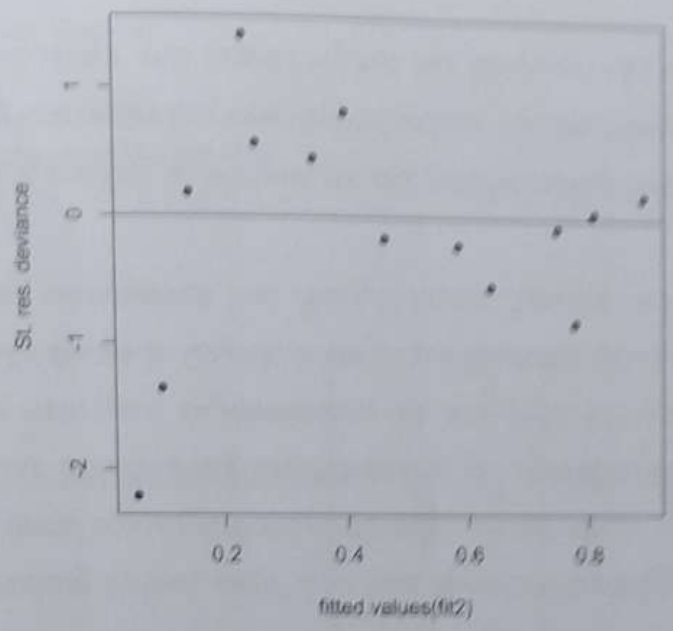
Τα τυποποιημένα υπόλοιπα deviance *και τα υπόλοιπα πιθανοφάνειας (*likelihood *ή* studentized residuals*) λαμβάνονται ευθέως από την R με τις εντολές*

```
stand.deviance <- rstandard(fit2)
res.lik <- rstudent(fit2)
```

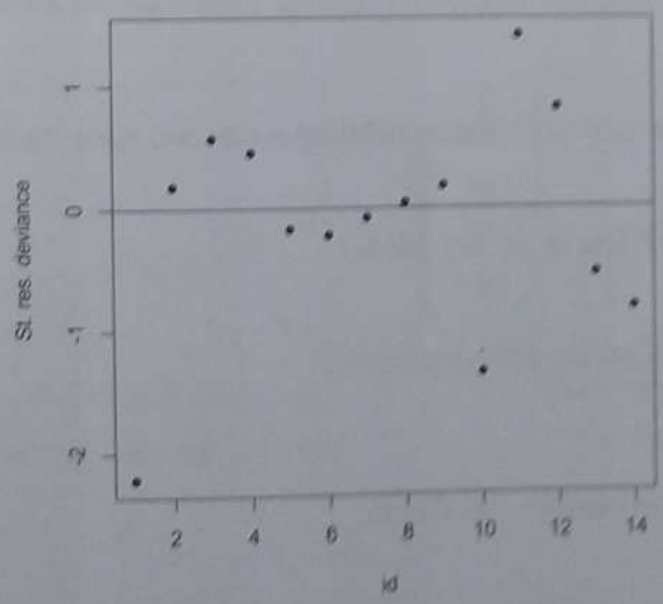
Τα Σχήματα 8.8 και 8.9 έχουν κατασκευαστεί με τις εντολές

```
plot(fitted.values(fit2), stand.deviance, pch=19, ylab="St. res. deviance")
abline(h = 0)
plot(id, stand.deviance, pch=19, ylab="St. res. deviance")
abline(h = 0)
```

και παρουσιάζουν τα γραφήματα των τυποποιημένων υπολοίπων deviance *σε σχέση με τις εκτιμημένες τιμές* ($\hat{p}_i$) *και σε σχέση με τη σειρά των δεδομένων (*id*).*



Σχήμα 8.8: Τυποποιημένα υπόλοιπα deviance ως προς τις προσαρμοσμένες τιμές $\hat{p}_i$ για το Παράδειγμα 8.9.1



Σχήμα 8.9: Τυποποιημένα υπόλοιπα deviance σε σχέση με τον αριθμό της παρατήρησης *id* για το Παράδειγμα 8.9.1

Με τις δύο αυτές γραφικές παραστάσεις μπορούμε

- *να ελέγξουμε την υπόθεση της ανεξαρτησίας των παρατηρήσεων, την οποία εδώ δεν μπορούμε να την απορρίψουμε, αφού τα υπόλοιπα δεν παρουσιάζουν κάποια ιδιαίτερη συμπεριφορά και κατανέμονται τυχαία γύρω από το μηδέν*

- *να εντοπίσουμε πιθανές παρατηρήσεις που αποκλίνουν σημαντικά από τις υπόλοιπες. Τέτοια παρατήρηση είναι η πρώτη, η οποία έχει σημαντικά μεγαλύτερη απόλυτη τιμή για τα τυποποιημένα υπόλοιπα σε σχέση με τις υπόλοιπες παρατηρήσεις. Η συγκεκριμένη παρατήρηση αντιστοιχεί σε εκείνη την ομάδα, στην οποία παρατηρήθηκε μηδενικός αριθμός ποντικιών με αρνητική αντίδραση, γεγονός που από μόνο του τη διαφοροποιεί από όλες τις άλλες.*

Να σημειώσουμε εδώ ότι με id στις προηγούμενες εντολές δηλώνουμε τον αύξοντα αριθμό της παρατήρησης.

Για τον εντοπισμό πιθανών σημείων επιρροής μπορούμε να κατασκευάσουμε

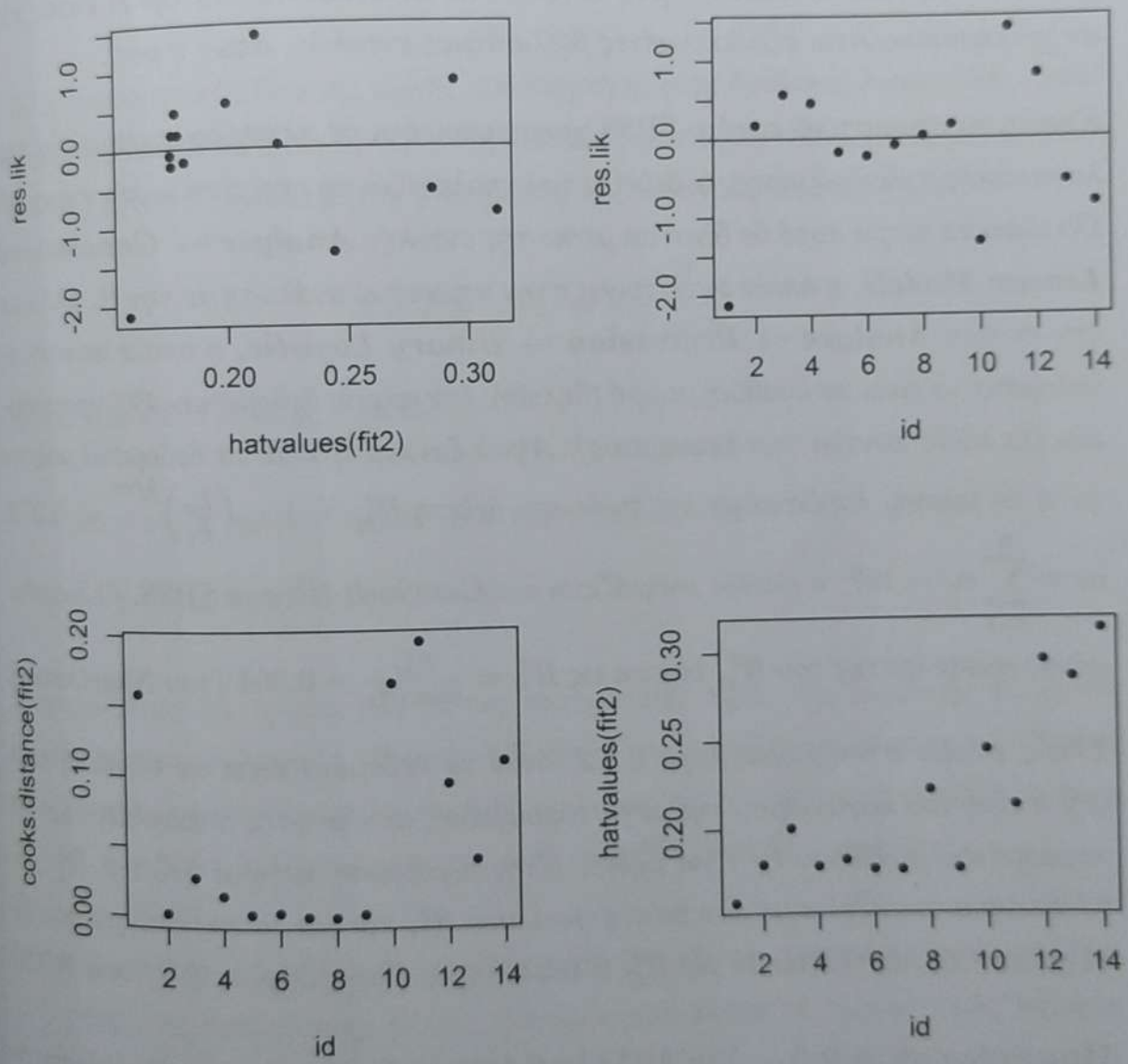
- *το διάγραμμα των υπολοίπων πιθανοφάνειας ως προς τα $\hat{h}_{ii}$*

- *τα γραφήματα δείκτη (index plots)*

  - *των υπολοίπων πιθανοφάνειας,*
  - *των h_{ii},*
  - *των αποστάσεων του Cook.*

Το διάγραμμα των υπολοίπων πιθανοφάνειας ως προς τα $\hat{h}_{ii}$ και τα τρία προαναφερθέντα γραφήματα δείκτη παρουσιάζονται στο Σχήμα 8.10, το οποίο κατασκευάζεται από την R με τις ακόλουθες εντολές:

```
par(mfrow = c(2, 2))
plot(hatvalues(fit2),res.lik, pch=19)
abline(h = 0)
plot(id,res.lik, pch=19)
abline(h = 0)
plot(id,cooks.distance(fit2), pch=19)
plot(id,hatvalues(fit2), pch=19)
```



Σχήμα 8.10: Υπόλοιπα πιθανοφάνειας σε σχέση με $\hat{h}_{ii}$, καθώς και γραφήματα των υπολοίπων πιθανοφάνειας, απόσταση Cook και $\hat{h}_{ii}$ ως προς τον αύξοντα αριθμό της παρατήρησης id, από την ανάλυση του Παραδείγματος 8.9.1

Από την εικόνα των γραφικών αυτών παραστάσεων είναι φανερό ότι δεν υπάρχει κανένα χαρακτηριστικό που να δημιουργήσει σοβαρές αμφιβολίες ως προς την ορθότητα του μοντέλου.

Παρατήρηση 8.9.1. *Κλείνουμε το παράδειγμα με μία αναφορά στις τιμές των δεικτών ψευδο-R^2, τονίζοντας για άλλη μία φορά τις επιφυλάξεις μας ως προς τη χρησιμότητά τους. Ο υπολογισμός τους δε γίνεται αυτόματα από την R εκτός και αν χρησιμοποιούνται εξειδικευμένες βιβλιοθήκες εντολών, όπως η pscl.*

Επειδή το στατιστικό πακέτο SPSS χρησιμοποιείται σε πληθώρα εφαρμογών της λογιστικής παλινδρόμησης, οι δείκτες που υπολογίζονται εκεί είναι οι πιο γνωστοί. Ωστόσο τα μέτρα αυτά δε δίνονται μέσω της εντολής **Analyze → Generalized Linear Models**, *η οποία αντιστοιχεί στην παραπάνω ανάλυση με την R, αλλά με την εντολή* **Analyze → Regression → Binary Logistic**, *η οποία απαιτεί τα δεδομένα να είναι σε δυαδική μορφή (δηλαδή, ένα αρχείο δεδομένων 497 γραμμών, μία για κάθε ποντίκι του πειράματος). Αφού ξαναγράψουμε τα δεδομένα μας σε αυτή τη μορφή, λαμβάνουμε τις τιμές του δείκτη $R^2_M = 1 - \left(\frac{L_0}{L_1}\right)^{2/m} = 0.272$, $m = \sum_{i=1}^{n} n_i = 497$, ο οποίος ονομάζεται και Cox-Snell R^2 στο SPSS. Ο διορθωμένος συντελεστής του R^2_M δίνεται ως $R^2_N = \frac{R^2_M}{\max R^2_M} = 0.364$ (του Nagelkerke).*

Όπως είδαμε στην Παράγραφο 8.9.2, όταν τα δεδομένα είναι σε δυαδική μορφή, η τιμή του λογαρίθμου της μεγιστοποιημένης συνάρτησης πιθανοφάνειας του χορεσμένου μοντέλου $\tilde{\ell}_S$ είναι μηδέν. Στην περίπτωση αυτή, ο δείκτης R^2_L του McFadden ταυτίζεται με τον δείκτη deviance R^2_D και για το παράδειγμά μας δίνεται από το MINITAB 17 ως $R^2_D = 0.2306$ (βλ. Παραγράφους 8.6.2 και 8.9.5).

Παρατήρηση 8.9.2. *Στο MINITAB 16 η ανάλυση ενός μοντέλου λογιστικής παλινδρόμησης εκτελείται με την εντολή* **Stat→Regression→Binary Logistic Regression** *από τη γραμμή εργαλείων. Στη συνέχεια, για δυαδικά δεδομένα, η εξαρτημένη μεταβλητή (y=0 ή 1) εισάγεται στο πλαίσιο* **Response in**

response/frequency format. Για διωνυμικά δεδομένα, επιλέγουμε **Response in event/trial format** *και ο αριθμός επιτυχιών (στην προκειμένη περίπτωση, Events) εισάγεται στο πλαίσιο* **Number of events** *και ο αριθμός δοκιμών (στο παράδειγμά μας, Trials) στο* **Number of trials**. *Και στις δύο περιπτώσεις, οι επεξηγηματικές μεταβλητές εισάγονται στο πλαίσιο* **Model**.

8.9.7 Καμπύλη ROC

Ένας διαφορετικός δείκτης καλής προσαρμογής μιας δυαδικής λογιστικής παλινδρόμησης βασίζεται στην προβλεπτική ικανότητα του μοντέλου, ως εξής: για κάθε μονάδα, έχουμε την εκτίμηση ή πρόβλεψη της πιθανότητας επιτυχίας

$$\hat{p} = \widehat{P}(Y = 1) = \frac{e^{x'\hat{\beta}}}{1 + e^{x'\hat{\beta}}}.$$

Έστω τώρα ένα όριο p_0 έτσι ώστε

- αν $\hat{p} > p_0$, προβλέπεται $Y = 1$ για τη μονάδα αυτή
- αν $\hat{p} \leq p_0$, προβλέπεται $Y = 0$.

Συγκρίνοντας τις προβλέψεις με τις πραγματικές τιμές της δυαδικής Y, κατασκευάζεται ο ακόλουθος Πίνακας 8.5, όπου για παράδειγμα, a είναι το πλήθος των μονάδων με πραγματική τιμή $Y = 1$ για τις οποίες και η πρόβλεψη είναι $\hat{Y} = 1$. Με βάση αυτόν τον πίνακα, ορίζονται δύο βασικές ποσότητες:

- Ευαισθησία (sensitivity)$= a/(a + c)$, δηλαδή το ποσοστό ορθής πρόβλεψης της κατάστασης $Y = 1$ ή αλλιώς «το ποσοστό των αληθώς θετικών αποτελεσμάτων» (true positive rate)
- Ειδικότητα (specificity) $= d/(b+d)$, δηλαδή το ποσοστό ορθής πρόβλεψης της κατάστασης $Y = 0$ ή αλλιώς «το ποσοστό των αληθώς αρνητικών αποτελεσμάτων» (true negative rate).

Πίνακας 8.5: *Πίνακας ταξινόμησης προβλέψεων*

		Πραγματική κατάσταση		
		$Y=1$	$Y=0$	
Πρόβλεψη	$Y=1$	a	b	$a+b$
	$Y=0$	c	d	$c+d$
		$a+c$	$b+d$	n

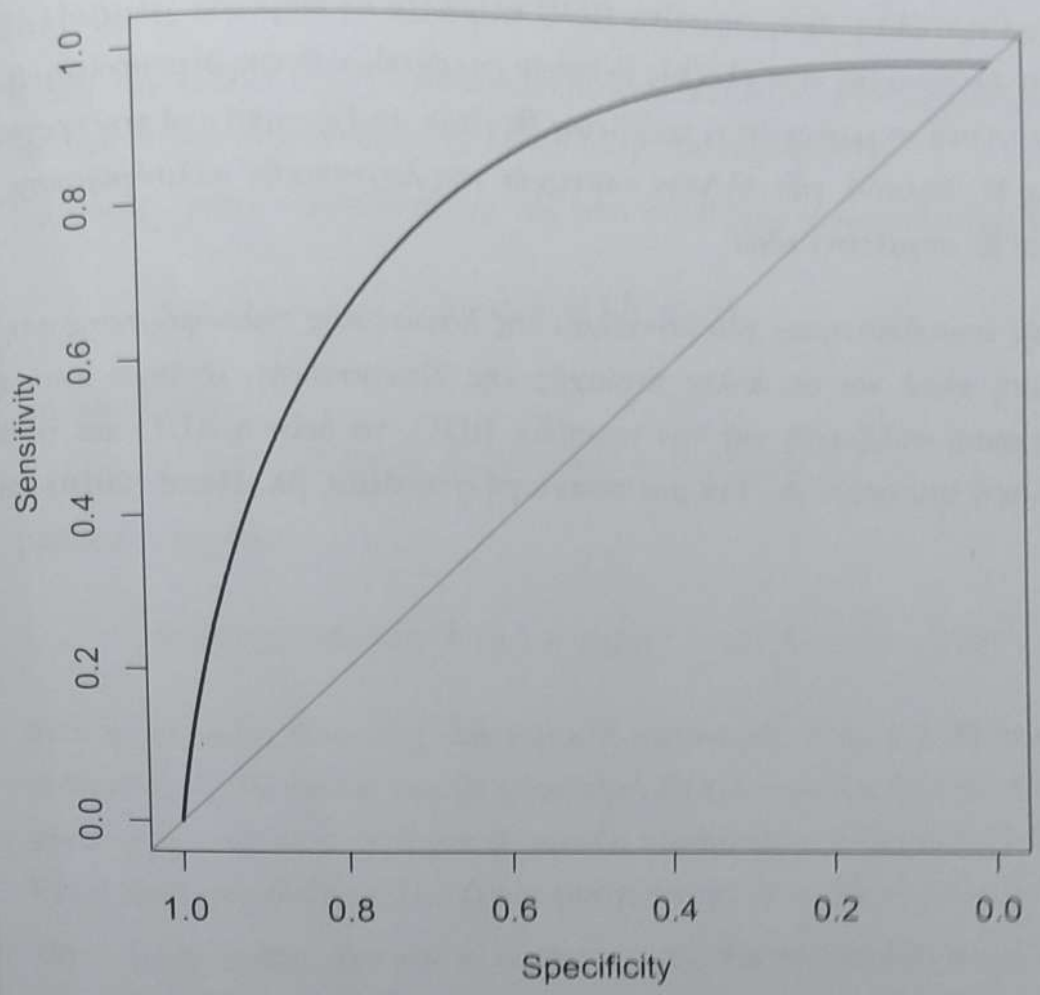
Χρήσιμη είναι και η ποσότητα 1-ειδικότητα $= b/(b+d)$, το ποσοστό των ψευδώς θετικών (false positive rate) αποτελεσμάτων (δηλαδή προβλέπεται $Y=1$, ενώ στην πραγματικότητα ισχύει $Y=0$). Αν υπολογιστούν οι τιμές της ευαισθησίας και της ειδικότητας για κάθε p_0 στο εύρος $[0,1]$, μπορεί να σχηματιστεί η χαρακτηριστική καμπύλη ROC (**receiver operating characteristic curve**), η οποία απεικονίζει την προβλεπτική ικανότητα του μοντέλου καθώς το όριο p_0 μεταβάλλεται.

Το όνομα ROC καθιερώθηκε από τις πρώτες εφαρμογές της μεθόδου σε ραντάρ.

Παράδειγμα 8.9.1 (συνέχεια)

Όπως είδαμε στην Παρατήρηση 8.9.1 πιο πάνω, έτσι και εδώ απαιτείται το αρχείο των δεδομένων (Πίνακας 8.3) να είναι σε δυαδική μορφή ώστε σε κάθε ποντίκι να αντιστοιχεί η τιμή $Y=1$ όταν υπάρχει αντίδραση στο φάρμακο και $Y=0$ αλλιώς. Στο Σχήμα 8.11 παρουσιάζεται η καμπύλη ROC για τα αποτελέσματα της ανάλυσης του Παραδείγματος 8.9.1 και λαμβάνεται από την R εκτελώντας τις εντολές

```
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
library(pROC)
mod<-glm(Y~log(DOSE)+drug, family=binomial)
roc(Y, fitted.values(mod), smooth=TRUE, plot=TRUE)
- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
```



Σχήμα 8.11: Καμπύλη ROC για το Παράδειγμα 8.9.1

Παρατηρείται ότι η κλίμακα του οριζοντίου άξονα είναι από 1 έως 0. Εναλλακτικά η ίδια καμπύλη μπορεί να παρουσιαστεί με 1-ειδικότητα στον οριζόντιο άξονα και τότε η κλίμακα θα είναι η συνηθισμένη, από 0 έως 1.

*Μεγάλη επιτυχία πρόβλεψης σημαίνει ότι υπάρχουν τιμές του ορίου p_0 με υψηλή ευαισθησία και ταυτόχρονα υψηλή ειδικότητα. Τότε, η καμπύλη ROC πλησιάζει την πάνω αριστερή γωνία του τετραγώνου του σχήματος. Ένας δείκτης, που μετρά κατά πόσο την πλησιάζει αυτήν τη γωνία, είναι το εμβαδόν κάτω από την καμπύλη (**area under the curve**, AUC), με μέγιστη τιμή το 1.*

Στο παράδειγμά μας το εμβαδόν αυτό είναι $AUC = 0.814$, μία αρκετά υψηλή τιμή. Από την άλλη, αν η καμπύλη ROC πλησίαζε τη διαγώνια γραμμή ($AUC = 0.5$), τότε τα ποσοστά των αληθώς θετικών και ψευδώς θετικών αποτελεσμάτων θα ήταν ίσα. Αυτό σημαίνει ότι η πρόβλεψη θα ήταν ανεξάρτητη από την πραγματική τιμή της Y, δηλαδή, μία πλήρης αποτυχία της λογιστικής παλινδρόμησης, κάτι που όμως δε συμβαίνει εδώ.

Λόγω της σπουδαιότητας του μοντέλου της λογιστικής παλινδρόμησης στη Βιοστατιστική αλλά και σε άλλες περιοχές της Στατιστικής, υπάρχει εκτεταμένη βιβλιογραφική συζήτηση για την καμπύλη ROC, το δείκτη AUC και τη βέλτιστη επιλογή του ορίου p_0. Για μια εισαγωγή στο θέμα, βλ. Hand (2010) και Liu (2012).