

Γενικευμένα Γραμμικά Μοντέλα για Δίτιμα Δεδομένα

Θεωρούμε ότι οι μεταβλητές απόκρισης $Y_1, Y_2, \dots, Y_n$ προέρχονται από την διωνυμική κατανομή η οποία αποτελεί υποπερίπτωση της εκθετικής οικογένειας κατανομών. Συγκεκριμένα η Y_i έχει σ.π.

$$f(y_i; \pi_i) = \binom{m_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i} = \exp \left\{ y_i \ln \frac{\pi_i}{1 - \pi_i} + m_i \ln(1 - \pi_i) + \ln \binom{m_i}{y_i} \right\}.$$

Σύμφωνα με τους συμβολισμούς που χρησιμοποιήσαμε για την εκθετική οικογένεια κατανομών,

$$\theta_i = \ln \frac{\pi_i}{1 - \pi_i}, \quad b(\theta_i) = -m_i \ln(1 - \pi_i), \quad a_i(\phi) = 1, \quad c(y_i, \phi) = \ln \binom{m_i}{y_i}, \quad \mu_i = m_i \pi_i.$$

Σε αυτή την περίπτωση είναι προτιμότερο να εκφράσουμε την συνάρτηση πιθανοφάνειας και την link function συναρτήσει των π_i . Τα $\pi_i, i=1, 2, \dots, n$ εξαρτώνται από ένα μικρότερο σύνολο παραμέτρων $\beta_0, \beta_1, \dots, \beta_{p-1}$, δηλαδή $\pi_i = \pi_i(\boldsymbol{\beta}) = \pi_i(\beta_0, \dots, \beta_{p-1})$. Συγκεκριμένα θεωρούμε ότι

$$g(\pi_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_{p-1} x_{i,p-1} = \mathbf{X}_i^T \cdot \boldsymbol{\beta}$$

για κάποια συνάρτηση σύνδεσης g . Στην περίπτωση της διωνυμικής κατανομής, οι συνηθέστερες συναρτήσεις σύνδεσης είναι οι:

- **Logit:** $g(\pi_i) = \ln \frac{\pi_i}{1 - \pi_i} \Leftrightarrow \pi_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}$
- **Probit:** $g(\pi_i) = \Phi^{-1}(\pi_i) \Leftrightarrow \pi_i = \Phi(\eta_i), \quad (\Phi: \text{σ.κ. της τυπικής κανονικής } N(0,1)).$
- **c-log-log:** $g(\pi_i) = \ln(-\ln(1 - \pi_i)) \Leftrightarrow \pi_i = 1 - \exp\{-e^{\eta_i}\}$

Γενικά, $\pi_i = F(\eta_i)$ για κάποια σ.κ. F (Παραπάνω έχουμε Logistic, Normal και Extreme Value Distribution αντίστοιχα).

Εκτίμηση των β_i

Σύμφωνα με τη γενική θεωρία που έχουμε αναπτύξει για τα γενικευμένα γραμμικά μοντέλα, η εκτίμηση των β_i γίνεται χρησιμοποιώντας την επαναληπτική διαδικασία

$$\boldsymbol{\beta}^{(m)} = (\mathbf{X}^T \mathbf{W}^{(m-1)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(m-1)} \mathbf{z}^{(m-1)}$$

όπου στην περίπτωση της διωνυμικής κατανομής προκύπτει ότι

$$\mathbf{z}_i^{(m-1)} = \frac{y_i - m_i \pi_i^{(m-1)}}{m_i} g'(\pi_i^{(m-1)}) + \eta_i^{(m-1)}, \quad \mathbf{W}^{(m-1)} = \text{diag} \left[\frac{m_i}{\pi_i^{(m-1)} (1 - \pi_i^{(m-1)}) g'(\pi_i^{(m-1)})^2} \right]_i$$

$$\eta_i^{(m-1)} = \mathbf{x}_i^T \boldsymbol{\beta}^{(m-1)}, \quad \pi_i^{(m-1)} = g^{-1}(\eta_i^{(m-1)}).$$

Αν π.χ. χρησιμοποιήσουμε logit link, τότε $g(\pi) = \log \frac{\pi}{1 - \pi} \Rightarrow g'(\pi) = \frac{1}{\pi(1 - \pi)}$ και αντικαθιστώντας το στους παραπάνω τύπους τελικά θα έχουμε ότι

$$z_i^{(m-1)} = \left[\frac{y_i - m_i \pi_i^{(m-1)}}{m_i \pi_i^{(m-1)} (1 - \pi_i^{(m-1)})} + \eta_i^{(m-1)} \right], \quad \mathbf{W}^{(m-1)} = \text{diag}[m_i \pi_i^{(m-1)} (1 - \pi_i^{(m-1)})]$$

και

$$\eta_i^{(m-1)} = \mathbf{x}_i^T \boldsymbol{\beta}^{(m-1)}, \quad \pi_i^{(m-1)} = g^{-1}(\eta_i^{(m-1)}) = \frac{\exp(\eta_i^{(m-1)})}{1 + \exp(\eta_i^{(m-1)})}.$$

Στο αρχικό βήμα συνήθως επιλέγουμε:

$$\pi_i^{(0)} = \begin{cases} y_i / m_i, y_i \neq 0, m_i \\ 0.5 / m_i, y_i = 0 \\ 1 - 0.5 / m_i, y_i = m_i \end{cases}, \quad \eta_i^{(0)} = g(\pi_i^{(0)}).$$

Η επαναληπτική διαδικασία σταματάει όταν $\sum_{i=1}^p (\beta_i^{(m)} - \beta_i^{(m-1)})^2 < \varepsilon$ για δεδομένο ε , ή μετά από προκαθορισμένο αριθμό βημάτων.

Έλεγχος καλής προσαρμογής του μοντέλου

Απορρίπτουμε την H_0 : $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$ (ότι δηλαδή το μοντέλο είναι σωστό) όταν

$$D_{c,f}(\mathbf{y}, \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^n \left(y_i \ln \frac{y_i}{m_i \hat{\pi}_i} + (m_i - y_i) \ln \frac{m_i - y_i}{m_i - m_i \hat{\pi}_i} \right) > X_{n-p;a}^2, \quad \hat{\pi}_i = g^{-1}(\mathbf{x}_i^T \hat{\boldsymbol{\beta}}).$$

($S_{c,f}(\mathbf{y}, \hat{\boldsymbol{\mu}}) = D_{c,f}(\mathbf{y}, \hat{\boldsymbol{\mu}})$ διότι $\varphi = 1$). Σύμφωνα με το X^2 του Pearson απορρίπτουμε την H_0 : $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$ όταν

$$X_{pearson}^2 = \sum_{i=1}^n \frac{(y_i - m_i \hat{\pi}_i)^2}{m_i \hat{\pi}_i (1 - \hat{\pi}_i)} > X_{n-p;a}^2$$

(Ασυμπτωτικά, τα δύο κριτήρια είναι ισοδύναμα. Για μικρά δείγματα συνήθως προτιμάται το $X_{pearson}^2$ διότι πλησιάζει πιο γρήγορα την κατανομή χι-τετράγωνο από ότι το D)

Σύγκριση μοντέλων

Αν θέλουμε να ελέγξουμε υπόθεση της μορφής: H_0 : $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$ έναντι του μεγαλύτερου μοντέλου H_1 : $g'(\boldsymbol{\mu}) = \mathbf{X}'\boldsymbol{\beta}'$, τότε πάλι σύμφωνα με το κριτήριο του γ.λ.π. απορρίπτουμε την H_0 (ότι δηλαδή το μικρότερο μοντέλο είναι σωστό) όταν

$$D_{c,f}(\mathbf{y}, \hat{\boldsymbol{\mu}}) - D_{c,f}(\mathbf{y}, \hat{\boldsymbol{\mu}}') > X_{p'-p}^2$$

(τα $D_{c,f}(\mathbf{y}, \hat{\boldsymbol{\mu}}), D_{c,f}(\mathbf{y}, \hat{\boldsymbol{\mu}}')$ είναι τα Deviances που αντιστοιχούν στο μικρό και στο μεγάλο μοντέλο αντίστοιχα). Ο ίδιος έλεγχος μπορεί να γίνει και με τη χρήση του $X_{pearson}^2$.

Διαστήματα εμπιστοσύνης, έλεγχοι υποθέσεων για το $\boldsymbol{\beta}$

Οι έλεγχοι για το $\boldsymbol{\beta}$ βασίζονται στο γεγονός ότι ασυμπτωτικά:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \text{Inf}(\boldsymbol{\beta})^{-1}) \Rightarrow (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \text{Inf}(\boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \chi_p^2,$$

$$(\text{εδώ } \mathbf{Inf}(\boldsymbol{\beta}) = \mathbf{X}^T \mathbf{W} \mathbf{X}, \quad \mathbf{W} = \text{diag} \left[\frac{m_i}{\pi_i (1 - \pi_i) g'(\pi_i)^2} \right]_i).$$

Επειδή ο πίνακας πληροφορίας $\mathbf{Inf}(\boldsymbol{\beta})$ δεν είναι γνωστός εκτιμάται από τον $\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X}$ όπου $\hat{\mathbf{W}}$ είναι ο τελευταίος πίνακας $\mathbf{W}^{(m)}$ της επαναληπτικής διαδικασίας που χρησιμοποιήθηκε για την εκτίμηση των β_i . Προσεγγιστικά λοιπόν θεωρούμε ότι $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1})$. Π.χ. θα ισχύει προσεγγιστικά ότι $\hat{\beta}_i \sim N(\beta_i, s^2(\hat{\beta}_i))$ όπου $s^2(\hat{\beta}_i)$ είναι το (i,i) στοιχείο του πίνακα $(\mathbf{X}^T \hat{\mathbf{W}} \mathbf{X})^{-1}$ και άρα το

$$(\hat{\beta}_i - s(\hat{\beta}_i) Z_{\alpha/2}, \hat{\beta}_i + s(\hat{\beta}_i) Z_{\alpha/2})$$

είναι ένα προσεγγιστικό δ.ε. για το β_i συντελεστού $1-\alpha$. Επίσης απορρίπτουμε την υπόθεση $H_0: \beta_i = b$ έναντι της $H_0: \beta_i \neq b$ σε ε.σ. ασυμπτωτικά ίσο με $1-\alpha$ όταν

$$\left| \frac{\hat{\beta}_i - b}{s(\hat{\beta}_i)} \right| > Z_{\alpha/2}.$$

Κατάλοιπα (residuals)

Στην περίπτωση της διωνυμικής κατανομής τα κατάλοιπα θα έχουν για $i = 1, 2, \dots, n$ την εξής μορφή :

- *Κατάλοιπα Pearson:*
$$r_i^P = \frac{y_i - m_i \hat{\pi}_i}{\sqrt{m_i \hat{\pi}_i (1 - \hat{\pi}_i)}}$$

- *Κατάλοιπα απόκλισης:*
$$r_i^D = \text{sgn}(y_i - m_i \hat{\pi}_i) \sqrt{2} \sqrt{y_i \ln \frac{y_i}{m_i \hat{\pi}_i} + (m_i - y_i) \ln \frac{m_i - y_i}{m_i - m_i \hat{\pi}_i}}$$

Υπενθυμίζεται ότι $D_{c,f} = \sum_{i=1}^n (r_i^D)^2$, $X_{Pearson}^2 = \sum_{i=1}^n (r_i^P)^2$.