

ΓΕΝΙΚΕΥΜΕΝΑ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

(ΠΕΡΙΛΗΠΤΙΚΗ ΠΕΡΙΓΡΑΦΗ ΤΟΥ ΓΕΝΙΚΟΥ ΜΟΝΤΕΛΟΥ)

Στην κλασσική ανάλυση παλινδρόμησης θεωρούμε $Y_1, Y_2, \dots, Y_n$ μεταβλητές απόκρισης (dependent variables) και $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ ερμηνευτικές μεταβλητές ($\mathbf{X}_i = (1, X_{i1}, X_{i2}, \dots, X_{i,p-1})^T$) με την υπόθεση ότι υπάρχουν παράμετροι $\beta_0, \beta_1, \dots, \beta_{p-1}$ ώστε

$$Y_i \sim N(\mu_i, \sigma^2), \text{ όπου } \mu_i = \beta_0 + \beta_1 X_{i1} + \dots + \beta_{p-1} X_{i,p-1} = \mathbf{X}_i^T \boldsymbol{\beta}.$$

Σκοπός μας ήταν η μελέτη του παραπάνω μοντέλου (εκτίμηση των β_i , έλεγχος καλής προσαρμογής κτλ).

Γενικότερα τώρα, θεωρούμε ότι οι μεταβλητές απόκρισης Y_i είναι δυνατό να προέρχονται από μία γενικότερη οικογένεια κατανομών. Και σε αυτή την περίπτωση, θεωρούμε ότι η μέση τιμή μ_i του Y_i εξαρτάται από ένα σύνολο παραμέτρων $\beta_0, \beta_1, \dots, \beta_{p-1}$. Συγκεκριμένα θεωρούμε ότι

$$g(\mu_i) = \eta_i = \mathbf{X}_i^T \cdot \boldsymbol{\beta}$$

για κάποια συνάρτηση g (στο κλασσικό μοντέλο είναι $g(\mu_i) = \mu_i$) την οποία ονομάζουμε συνάρτηση σύνδεσης (link function). Μία αρκετά ευρεία κλάση κατανομών η οποία περιλαμβάνει την κανονική, την διωνυμική, την Poisson κ.α. είναι η εκθετική οικογένεια κατανομών. Θεωρούμε λοιπόν ότι οι μεταβλητές απόκρισης $Y_1, Y_2, \dots, Y_n$ προέρχονται από την εκθετική οικογένεια κατανομών, δηλαδή η Y_i έχει σ.π.π. (ή σ.π.) της μορφής

$$f(y_i; \theta_i, \phi) = \exp \left\{ \frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\}, \quad \mu_i = E(Y_i) = b'(\theta_i)$$

Σύμφωνα με τα παραπάνω, τα $\theta_i, i=1,2,\dots,n$ εξαρτώνται από ένα μικρότερο σύνολο παραμέτρων $\beta_0, \beta_1, \dots, \beta_{p-1}$, δηλαδή $\theta_i = \theta_i(\boldsymbol{\beta})$ διότι $g(b'(\theta_i)) = \mathbf{X}_i^T \cdot \boldsymbol{\beta}$ (στην κανονική κατανομή $\theta_i = \mu_i$). Σκοπός μας και πάλι είναι η μελέτη του συγκεκριμένου γενικότερου μοντέλου (εκτίμηση των β_i , έλεγχος καλής προσαρμογής κτλ).

1. Εκτίμηση των β_i

Η εκτίμηση των β_i γίνεται χρησιμοποιώντας τις εκτιμήτριες μέγιστης πιθανοφάνειας (MLE's) οι οποίες ως γνωστό προκύπτουν από τη λύση του συστήματος

$$\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}) = \mathbf{0}$$

ως προς β_i όπου $l(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta})$ και $L(\boldsymbol{\beta}) = \ln \prod_{i=1}^n f(y_i; \theta_i)$ είναι η συνάρτηση πιθανοφάνειας. Συγκεκριμένα, έχουμε το σύστημα

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}} \ln \prod_{i=1}^n f(y_i; \theta_i) = \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^n \left(\frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right) = \mathbf{0} \\ \Leftrightarrow \left(\frac{\partial}{\partial \beta_0} \sum_{i=1}^n \left(\frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right), \dots, \frac{\partial}{\partial \beta_{p-1}} \sum_{i=1}^n \left(\frac{y_i \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right) \right) &= (0, 0, \dots, 0) \end{aligned}$$

Λόγω του ότι το τελευταίο σύστημα (p εξισώσεων με p αγνώστους $\beta_0, \dots, \beta_{p-1}$) βρίσκεται γενικά σε πεπλεγμένη μορφή (είναι μη γραμμικό), ο υπολογισμός των λύσεων γίνεται προσεγγιστικά μέσα από μία επαναληπτική διαδικασία (αρχικά υπολογίζουμε ένα $\boldsymbol{\beta}^{(0)}$, με βάση αυτό υπολογίζουμε ένα $\boldsymbol{\beta}^{(1)}$, με βάση το $\boldsymbol{\beta}^{(1)}$ υπολογίζουμε ένα $\boldsymbol{\beta}^{(2)}$, κ.ο.κ. ώστε $\boldsymbol{\beta}^{(k)} \rightarrow \boldsymbol{\beta}$ για $k \rightarrow \infty$).

Συγκεκριμένα χρησιμοποιούμε μία παραλλαγή της προσεγγιστικής (p -διάστατης) μεθόδου Newton-Raphson¹ η οποία καλείται Fisher Scoring Method (κατά την οποία ο πίνακας $\mathbf{J}$ που εμφανίζεται στην επαναληπτική διαδικασία αντικαθίσταται από τον $E(\mathbf{J}) = -\mathbf{Inf}(\boldsymbol{\theta})$). Μετά από απλοποιήσεις φτάνουμε σε μία επαναληπτική μέθοδο της μορφής

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

όπου τα $\mathbf{W}^{(m-1)}$, $\mathbf{z}^{(m-1)}$ υπολογίζονται με βάση τα $\boldsymbol{\beta}^{(m-1)}$. Συγκεκριμένα,

$$\mathbf{W}^{(m-1)} = \text{diag}(1/(V(\mu_i^{(m-1)})g'(\mu_i^{(m-1)})^2))_i, \quad \mathbf{z}_i^{(m-1)} = \eta_i^{(m-1)} + (y_i - \mu_i^{(m-1)})g'(\mu_i^{(m-1)})$$

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

Στο αρχικό βήμα συνήθως επιλέγουμε: $\mu_i^{(0)} = Y_i$, $\eta_i^{(0)} = g(Y_i)$, $\boldsymbol{\beta}^{(0)} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T g(\mathbf{Y})$. 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 (συνήθως όλη η παραπάνω διαδικασία για την εκτίμηση του $\boldsymbol{\beta}$ γίνεται αυτόματα από το στατιστικό πακέτο).

2. Έλεγχος καλής προσαρμογής του μοντέλου

Ο έλεγχος για την καλή προσαρμογή του μοντέλου γίνεται με την χρήση του γενικευμένου λόγου πιθανοφαιών. Συγκεκριμένα θέλουμε να ελέγξουμε την μηδενική υπόθεση $H_0: g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$. Σύμφωνα με το κριτήριο του γενικευμένου λόγου πιθανοφαιών απορρίπτουμε την H_0 όταν

$$\frac{\sup_{\boldsymbol{\mu} \in \Theta_0} L(\boldsymbol{\mu}, \mathbf{y})}{\sup_{\boldsymbol{\mu} \in R^n} L(\boldsymbol{\mu}, \mathbf{y})} < c \Leftrightarrow -2 \ln \frac{L_c(\hat{\boldsymbol{\mu}}, \mathbf{y})}{L_f(\mathbf{y}, \mathbf{y})} = 2(l_f(\mathbf{y}, \mathbf{y}) - l_c(\hat{\boldsymbol{\mu}}, \mathbf{y})) := S_{c,f}(\mathbf{y}, \hat{\boldsymbol{\mu}}) > c'$$

όπου $\Theta_0 = \{\boldsymbol{\mu}: g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\beta} \in R^p\}$. Η στατιστική συνάρτηση $S_{c,f}(\mathbf{y}, \hat{\boldsymbol{\mu}}) = D_{c,f}(\mathbf{y}, \hat{\boldsymbol{\mu}})/\phi$ καλείται **Scaled Deviance** και ασυμπτωτικά ακολουθεί χ^2 κατανομή με $\dim(R^n) - \dim(\Theta_0) = n - p$ β.ε (βλ. παρατήρηση στο τέλος). Άρα απορρίπτουμε την H_0 (ότι δηλαδή το μοντέλο είναι σωστό) όταν

$$S_{c,f}(\mathbf{y}, \hat{\boldsymbol{\mu}}) > \chi_{n-p,a}^2.$$

Ένα άλλο κριτήριο καλής προσαρμογής είναι το X^2 του **Pearson** (ασυμπτωτικά ισοδύναμο του $S_{c,f}(\mathbf{y}, \hat{\boldsymbol{\mu}})$):

$$X^2 = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)} \sim \chi_{n-p}^2$$

Σύμφωνα με αυτό το κριτήριο απορρίπτουμε την H_0 όταν $X^2 > \chi_{n-p,a}^2$ (τα στατιστικά πακέτα συνήθως δίνουν την τιμή της $S_{c,f}$ ή του X^2 του Pearson).

¹ Γενικά η n -διάστατη N-R μέθοδος χρησιμοποιείται για την εύρεση των ριζών ενός συστήματος εξισώσεων

$$\mathbf{f}(\mathbf{x}) = \mathbf{0} \Leftrightarrow (f_1(x_1, x_2, \dots, x_p), \dots, f_p(x_1, x_2, \dots, x_p)) = (0, 0, \dots, 0).$$

Ο γενικός αναγωγικός τύπος που χρησιμοποιείται είναι ο

$$\mathbf{x}^{(m)} = \mathbf{x}^{(m-1)} - \mathbf{J}_{\mathbf{x}^{(m-1)}}^{-1} \cdot \mathbf{f}(\mathbf{x}^{(m-1)}), \quad \mathbf{J} = \left[\frac{\partial f_i}{\partial x_j} \right]_{i,j} \quad (\text{μονοδιάστατη περίπτωση: } x^{(m)} = x^{(m-1)} - \frac{f(x^{(m-1)})}{f'(x^{(m-1)})})$$

Στη συγκεκριμένη περίπτωση είναι $f_i = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_i}$, $x_i = \beta_i$, και άρα $\mathbf{J} = \left[\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j} \right]_{i,j}$.

3. Σύγκριση μοντέλων

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

$$\frac{\sup_{\boldsymbol{\mu} \in \Theta_0} L(\mathbf{y}, \boldsymbol{\mu})}{\sup_{\boldsymbol{\mu} \in \Theta_1} L(\mathbf{y}, \boldsymbol{\mu})} < c \Leftrightarrow -2 \ln \frac{L(\mathbf{y}, \hat{\boldsymbol{\mu}})}{L(\mathbf{y}, \hat{\boldsymbol{\mu}}')} = 2(l(\mathbf{y}, \hat{\boldsymbol{\mu}}') - l(\mathbf{y}, \hat{\boldsymbol{\mu}})) = S_{c,f}(\mathbf{y}, \hat{\boldsymbol{\mu}}) - S_{c,f}(\mathbf{y}, \hat{\boldsymbol{\mu}}') > c'$$

και όμοια με παραπάνω ασυμπτωτικά ακολουθεί χ^2 κατανομή με $\dim(\Theta_1) - \dim(\Theta_0) = p' - p$ β.ε. Άρα απορρίπτουμε την H_0 (ότι δηλαδή το μικρότερο μοντέλο είναι σωστό) όταν

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

4. Έλεγχοι υποθέσεων για το $\boldsymbol{\beta}$

Οι έλεγχοι για το $\boldsymbol{\beta}$ βασίζονται στο γεγονός ότι αν ισχύει η $H_0: \boldsymbol{\beta} = \boldsymbol{\beta}_0$, τότε ασυμπτωτικά (βλ. παρατήρηση παρακάτω):

$$\text{i) } \mathbf{U} \sim N(0, \text{Inf}(\boldsymbol{\beta}_0)) \Rightarrow \mathbf{U}^T (\text{Inf}(\boldsymbol{\beta}_0))^{-1} \mathbf{U} \sim \chi_p^2, \quad \text{Inf}(\boldsymbol{\beta}) = \mathbf{X}^T \mathbf{W} \mathbf{X}, \quad \mathbf{W} = \text{diag}(1/V(y_i)(g'(\mu_i))^2)$$

$$\text{ii) } \hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}_0, \text{Inf}(\boldsymbol{\beta}_0)^{-1}) \Rightarrow (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \text{Inf}(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \sim \chi_p^2$$

$$\text{iii) } -2 \ln \frac{\sup_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\beta}_0)}{\sup_{\boldsymbol{\theta} \in \Theta_1} L(\hat{\boldsymbol{\beta}})} = 2(l(\hat{\boldsymbol{\beta}}) - l(\boldsymbol{\beta}_0)) \sim \chi_p^2.$$

5. Κατάλοιπα (residuals)

Κατάλοιπα Pearson:

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{V}(y_i)}}$$

Τυποποιημένα κατάλοιπα Pearson:

$$r_i^{PS} = \frac{r_i^P}{\sqrt{1 - h_{ii}}} \quad ((h_{ij}) = \mathbf{W}^{-\frac{1}{2}} \mathbf{X}(\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{-\frac{1}{2}})$$

Κατάλοιπα απόκλισης:

$$r_i^D = \text{sgn}(y_i - \hat{\mu}_i) \sqrt{2\phi(l_f(y_i, y_i) - l_c(\hat{\mu}_i, y_i))}, \quad (D_{c,f} = \sum_{i=1}^n (r_i^D)^2)$$

Παρατήρηση. (Ασυμπτωτικά αποτελέσματα για τις εκτιμήτριες μέγιστης πιθανοφάνειας (MLE's) και την Score function $\mathbf{U}$). Έστω $Y_1, Y_2, \dots, Y_n$ ανεξάρτητες και ισόνομες τ.μ. με συνάρτηση πιθανοφάνειας $L(\boldsymbol{\theta}) = L(\theta_1, \theta_2, \dots, \theta_p)$. Ως Score function του δείγματος θα καλείται η τ.μ.

$$\mathbf{U} = \frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ln L(\boldsymbol{\theta}) = \sum_{i=1}^n \frac{\partial}{\partial \boldsymbol{\theta}} \ln f(Y_i; \boldsymbol{\theta}).$$

Κάτω από κατάλληλες συνθήκες ομαλότητας της $L(\boldsymbol{\theta})$ (βλ. π.χ. GLM - McCullagh & Nelder p.469, J.D.Gibbons p.16, Rao (1973) p.364) ισχύει ότι

$$\text{i) } E(\mathbf{U}) = \mathbf{0}, \quad V(\mathbf{U}) = -E\left(\frac{\partial^2 l}{\partial \boldsymbol{\theta}^2}\right) \equiv \mathbf{Inf}(\boldsymbol{\theta}) \quad (\text{πίνακας πληροφορίας κατά Fisher})$$

$$\text{ii) } \mathbf{U} \xrightarrow{n \rightarrow \infty} N(0, \mathbf{Inf}(\boldsymbol{\theta}))$$

$$\text{iii) } \hat{\boldsymbol{\theta}} \xrightarrow{n \rightarrow \infty} N(\boldsymbol{\theta}, \mathbf{Inf}(\boldsymbol{\theta})^{-1}), \quad \text{όπου } \hat{\boldsymbol{\theta}} \text{ η ε.μ.π. (MLE) του } \boldsymbol{\theta}.$$

$$\text{iv) } -2 \ln \frac{\sup_{\boldsymbol{\theta} \in \Theta_0} L(\boldsymbol{\theta})}{\sup_{\boldsymbol{\theta} \in \Theta_1} L(\boldsymbol{\theta})} \xrightarrow{n \rightarrow \infty} \chi^2_{\dim(\Theta_1) - \dim(\Theta_0)} \quad (\dim(\Theta) = \text{διάσταση του χώρου } \Theta)$$