Γενικευμενα Γραμμικα Μοντελα για διτιμα μοντελα

Σ. Ζημερας
Επικ. Καθ.
Τμημα Στατιστικης και Αναλογιστικων – Χρηματοοικονομικων Μαθηματικων
Καρλοβασι
Σαμος

Μοντελο

Εστω τ,μ, Υ με

$$Y = \begin{cases} 1, \text{ επιτυχια με } p\left(Y=1\right)=\pi \\ 0, \text{ αποτυχια με } p\left(Y=0\right)=1-\pi \end{cases}$$

Η σ.π.π δινεται από την σχεση

$$f\left(y,\pi\right)=\pi^{y}\left(1-\pi\right)^{1-y}$$

Η από κοινου σ.π. δινεται από την σχεση

$$f\left(y,\pi\right)=\prod_{i=1}^{n}\pi_i^{y_i}\left(1-\pi_i\right)^{1-y_i}=\exp\left\{\sum_i y_i\log\left(\frac{\pi_i}{1-\pi_i}\right)+\sum_i\log\left(1-\pi_i\right)\right\}$$

Εκθετικη οικογενεια κατανομων

Μοντελο

Αν τα π_j είναι όλα ίσα μεταξύ τους, μπορούμε να ορίσουμε τη συνάρτηση $Y = \sum_{i=1}^{n} Y_i$, όπου το Y παριστάνει τον αριθμό των επιτυχιών σε n ανεξάρτητες προσπάθειες και ακολουθεί τη διωνυμική κατανομή, με παραμέτρους n, π και συνάρτηση μάζας πιθανότητας:

$$P(Y = y) = \begin{pmatrix} n \\ y \end{pmatrix} \pi^y (1-\pi)^{n-y} \quad , \quad \text{όπου } y = 0, 1, ..., n.$$

Αν $Y_i \sim b(n_i, \pi_i)$ τότε η λογαριθμική συνάρτηση πιθανοφάνειας είναι:

$$l(\pi_1, ..., \pi_N; y_1, ..., y_N) = \sum_{i=1}^{N} [y_i log(\frac{\pi_i}{1-\pi_i}) + n_i log(1-\pi_i) + log \begin{pmatrix} n_i \\ y_i \end{pmatrix}].$$

Μοντέλο

Έστω N ανεξάρτητες μεταβλητές Y_i για τις οποίες ισχύει: $Y_i \sim b(n_i, \pi_i)$.

Μοντελοποιούμε τις πιθανότητες π_i, ως:

$$g(\pi_i) = \eta_i = \mathbf{x}_i^T \mathbf{b},$$

όπου $\mathbf{x}_i$ το διάνυσμα επεξηγηματικών μεταβλητών, το $\mathbf{b}$ το διάνυσμα των παραμέτρων και g η συνάρτηση σύνδεσης. Οι τιμές του π_i ανήκουν στο διάστημα $[0, 1]$ επειδή είναι πιθανότητες. Αυτό μας περιορίζει σχετικά με τα $\mathbf{x}_i$ και $\mathbf{b}$.

Μοντέλο

Συναρτήσεις σύνδεσης

• logit: $g(\pi_i) = log(\frac{\pi_i}{1-\pi_i})$,

οπότε θα είναι: $\pi_i = \frac{e^{n_i}}{1+e^{n_i}}$.

• probit: $g(\pi_i) = \Phi^{-1}(\pi_i)$, όπου με Φ συμβολίζεται η συνάρτηση κατανομής της Κανονικής κατανομής και Φ^{-1} είναι η αντίστροφή της,

οπότε $\pi_i = \Phi(n_i)$.

• complementary log-log ή cloglog: $g(\pi_i) = log(-log(1-\pi_i))$,

οπότε: $\pi_i = 1 - exp(-e^{n_i})$.

Λογιστική παλινδρομηση

Στη Λογιστική Παλινδρόμηση η μεταβλητή απόκρισης ακολουθεί την κατανομή Bernoulli ή την διωνυμική κατανομή.

Σε περίπτωση όπου η μεταβλητή απόκρισης είναι μία τυχαία μεταβλητή από την κατανομή Bernoulli έχουμε:

$$E(y_i) = \pi_i = P(x_i),$$

$$Var(y_i) = \pi_i(1 - \pi_i).$$

Σύμφωνα με το μοντέλο $Y_i = b_0 + b_1 x_i + \varepsilon_i$, έχουμε:

$$E(Y_i) = b_0 + b_1 X_i,$$

άρα:

$$b_0 + b_j X_i = \pi_i.$$

Λογιστική παλινδρομηση

Επίσης,

$$P(Y_i = 1) = \pi_i \quad , \quad P(Y_i = 0) = 1 - \pi_i.$$

Άρα η μέση τιμή $E(Y_i) = b_0 + b_1 X_i$ είναι η πιθανότητα ότι η $Y_i = 1$.

Αν υποθέσουμε ότι $y = 1$ επιτυχία και $y = 0$ αποτυχία με πιθανότητα π και $1 - \pi$ αντίστοιχα, τότε έχουμε:

$$E(Y_i) = \frac{exp(b_0 + b_1 X_i)}{1 + exp(b_0 + b_1 X_i)}.$$

Λογιστική παλινδρομηση

Για να εκτιμήσουμε τις παραμέτρους, θα χρησιμοποιήσουμε τη μέθοδο μέγιστης πιθανοφάνειας. Η συνάρτηση πιθανοφάνειας είναι:

$$L(b) = \prod_{i=1}^{n} \begin{pmatrix} n_i \\ y_i \end{pmatrix} \pi_i^{y_i} (1 - \pi_i)^{n_i - y_i},$$

όπου το π_i είναι:

$$\pi_i = \frac{e^{b_0 + b_1 X_{1i} + \ldots + b_k X_{ki}}}{1 + e^{b_0 + b_1 X_{1i} + \ldots + b_k X_{ki}}}.$$

Αυτό που ψάχνουμε, είναι να βρούμε τα $\hat{b}_0, \hat{b}_1, \ldots, \hat{b}_k$ που μεγιστοποιούν το $L(b)$.

Λογιστική παλινδρομηση

Το γενικό λογιστικό μοντέλο, έχει τη μορφή:

$$logit\pi_i = log(\tfrac{\pi_i}{1-\pi_i}) = \mathbf{x}_i^T\mathbf{b},$$

όπου το $\mathbf{x}_i$ είναι ένα διάνυσμα με συνεχείς τιμές, που αντιστοιχούν σε συμμεταβλητές και εικονικές μεταβλητές. Το **b**, είναι το διάνυσμα παραμέτρων.

Οι εκτιμήσεις μέγιστης πιθανοφάνειας των παραμέτρων **b**, οπότε και των πιθανοτήτων: $\pi_i = g(\mathbf{x}_i^T\mathbf{b})$ προκύπτουν από τη μεγιστοποίηση της λογαριθμικής συνάρτησης πιθανοφάνειας:

$$l(\pi; y) = \sum_{i-1}^{N}\{y_i log(\pi_i) + (n_i - y_i)log(1-\pi_i) + log\begin{pmatrix} n_i \\ y_i \end{pmatrix}\}.$$

Λογιστική παλινδρομηση

Για να ελέγξουμε την καλή προσαρμογή του μοντέλου, χρησιμοποιούμε την στατιστική συνάρτηση deviance:

$$D = 2[l(\hat{\pi}_{max}; y) - l(\hat{\pi}; y)].$$

Αν οι μεταβλητές απόκρισης $Y_1, ...Y_N$ είναι ανεξάρτητες και ακολουθούν τη διωνυμική κατανομή,τότε, η λογαριθμική συνάρτηση πιθανοφάνειας, είναι:

$$l(\mathbf{b}, \mathbf{y}) = \sum_{i=1}^{T} \{y_i log\pi_i - y_i log(1-\pi_i) + n_i log(1-\pi_i) + log \begin{pmatrix} n_i \\ y_i \end{pmatrix}\}.$$

Τα π_i είναι όλα διαφορετικά. Έτσι, $\mathbf{b} = [\pi_i, ..., \pi_N]^T$. Οι εκτιμητές μέγιστης πιθανοφάνειας είναι: $\hat{\pi}_i = y_i/n_i$, ώστε η μέγιστη τιμή της λογαριθμικής συνάρτησης πιθανοφάνειας, είναι:

$$l(\mathbf{b}_{max}; \mathbf{y}) = \sum [y_i log \begin{pmatrix} y_i \\ n_i \end{pmatrix}] - y_i log(\tfrac{n_i - y_i}{n_i}) + n_i log(\tfrac{n_i - y_i}{n_i}) + log \begin{pmatrix} n_i \\ y_i \end{pmatrix}.$$

Λογιστική παλινδρομηση

Για οποιοδήποτε άλλο μοντέλο με $p < N$ παραμέτρους, ας υποθέσουμε ότι $\hat{\pi}_i$ δηλώνει την εκτιμήτρια μέγιστης πιθανοφάνειας για τις πιθανότητες και $\hat{y}_i = n_i\hat{\pi}_i$ δηλώνει τις προσαρμοσμένες τιμές. Τότε, η λογαριθμική συνάρτηση πιθανοφάνειας, για αυτές τις τιμές, είναι:

$$l(\mathbf{b};\mathbf{y}) = \sum[y_i log\begin{pmatrix} \hat{y}_i \\ n_i \end{pmatrix}] - y_i log(\tfrac{n_i-\hat{y}_i}{n_i}) + n_i log(\tfrac{n_i-\hat{y}_i}{n_i}) + log\begin{pmatrix} n_i \\ y_i \end{pmatrix},$$

οπότε, η deviance είναι:

$$D = 2[l(\mathbf{b}_{max};\mathbf{y}) - l(\mathbf{b};\mathbf{y})]$$

$$= 2\sum_{i=1}^{N}\{y_i log\begin{pmatrix} y_i \\ \hat{y}_i \end{pmatrix} + (n_i - y_i)log(\tfrac{n_i-y_i}{n_i-\hat{y}_i})\},$$

και θα έχει τη μορφή:

$$D = 2\sum olog\tfrac{o}{e},$$

όπου το o, παριστάνει τις συχνότητες που παρατηρούμε y_i και $(n_i - y_i)$ και το e δείχνει τις αναμενόμενες συχνότητες που παρατηρούμε, $n_i\hat{\pi}_i$ και $(n_i - n_i\hat{\pi}_i)$.

Λογιστική παλινδρομηση

Το D δεν περιέχει το σ^2, σε αντίθεση με την περίπτωση όπου η απόκριση ακολουθεί Κανονική κατανομή, οπότε, μπορεί να χρησιμοποιηθεί κατ' ευθείαν για τον έλεγχο καλής προσαρμογής, όπως και για ελέγχους υποθέσεων, με την προσέγγιση:

$$D \sim \chi^2_{N-p},$$

όπου p ο αριθμός των παραμέτρων που εκτιμήθηκαν και N ο αριθμός συμμεταβλητών.