ΜΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΙΣΗ

Τα μη γραμμικά μοντέλα έχουν την πιο κάτω μορφή:

$$Y_i = f\left(X_i, \gamma\right) + \varepsilon_i$$

η μορφή αυτή μοιάζει με τη μορφή που έχουμε για τα γραμμικά μοντέλα (δηλαδή η παρατήρηση Y_i είναι το άθροισμα της αναμενόμενης συνάρτησης $f(X_i,\gamma)$ με τυχαία σφάλματα ε_i.

η διαφορά είναι ότι η αναμενόμενη συνάρτηση εδώ είναι μη γραμμική

ΜΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΙΣΗ

- Τα σφάλματα είναι τυχαίες μεταβλητές με τις πιο κάτω υποθέσεις:
- $E(\varepsilon_i)=0$
- Σταθερή διασπορά
- Ανά δύο τα σφάλματα είναι ασυσχέτιστα $E(\varepsilon_i,\varepsilon_j)=0,\ \forall i \neq j$
- Επίσης πολλές φορές υποθέτουμε ότι είναι κανονικές μεταβλητές

ΑΝΕΞΑΡΤΗΤΕΣ ΜΕΤΑΒΛΗΤΕΣ

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΙΣΗ

Το λογιστικό μοντέλο είναι ένα

1. μη γραμμικό μοντέλο
2. τα σφάλματα δεν ακολουθούν κανονική κατανομή και
3. η μεταβλητή απόκρισης είναι διακριτή.

- Η λογιστική παλινδρόμηση χρησιμοποιείται σε περιπτώσεις στις οποίες επιθυμούμε να προβλέψουμε την απουσία ή την παρουσία ενός χαρακτηριστικού, ή ενός συμβάντος. Είναι μια γενίκευση της απλής γραμμικής παλινδρόμησης για την περίπτωση όπου η εξαρτημένη μεταβλητή (Y) είναι δίτιμη (δηλαδή παίρνει την τιμή 0 όταν απουσιάζει το χαρακτηριστικό ή την τιμή 1 όταν υπάρχει το χαρακτηριστικό).

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΙΣΗ

Ερμηνεία

- Απλό γραμμικό μοντέλο:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Δίτιμη μεταβλητή (0, 1)

- Επειδή ισχύει $E(\varepsilon_i)=0$

$$E(Y_i) = E(\beta_0 + \beta_1 X_i + \varepsilon_i)$$
$$= E(\beta_0 + \beta_1 X_i) + E(\varepsilon_i)$$
$$= \beta_0 + \beta_1 X_i$$

ΛΟΓΙΣΤΙΚΗ ΠΑΛΙΝΔΡΟΜΙΣΗ

Επίσης αφού είναι δίτιμη μεταβλητή η Y_i, θα είναι μια μεταβλητή Bernoulli με

- Όταν το $Y_i = 1$ έχουμε $\mathrm{P}(Y_i = 1) = \pi_i$
- Όταν το $Y_i = 0$ έχουμε $\mathrm{P}(Y_i = 0) = 1 - \pi_i$

Με βάση τον ορισμό της αναμενόμενης τιμής έχουμε

$$E(Y_i) = 1\pi_i + 0(1 - \pi_i) = \pi_i$$

Εξισώνοντας τους τύπους των δύο αναμενόμενων τιμών έχουμε

Όταν το $Y_i = 1$ έχουμε $\mathrm{P}(Y_i = 1) = \pi_i$

$$E(Y_i) = \beta_0 + \beta_1 X_i = \pi_i$$

Όταν ανεξάρτητη μεταβλητή η X_i

ΓΙΑΤΙ ΌΧΙ ΓΡΑΜΜΙΚΟ ΜΟΝΤΕΛΟ

1. Τα σφάλματα δεν είναι κανονικά

Έχουμε

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \Leftrightarrow \varepsilon_i = Y_i - \left(\beta_0 + \beta_1 X_i\right)$$

Όταν

$$Y_i = 0: \ \varepsilon_i = -\beta_0 - \beta_1 X_i$$

$$Y_i = 1: \ \varepsilon_i = 1 - \beta_0 - \beta_1 X_i$$

Όχι κανονική κατανομή σφαλμάτων

ΓΙΑΤΙ ΌΧΙ ΓΡΑΜΜΙΚΟ ΜΟΝΤΕΛΟ

2. Τα σφάλματα έχουν άνισες διασπορές

Όταν η αποκρινόμενη μεταβλητή παίρνει τις τιμές 0 ή 1 τα σφάλματα δεν έχουν ίσες διασπορές.

$$Var(\varepsilon_i) = Var(Y_i - \pi_i) = Var(Y_i) + Var(-\pi_i) = Var(Y_i) + 0 = Var(Y_i)$$

$$\begin{aligned}
Var(Y_i) &= E\left\{(Y_i - E(Y_i))^2\right\} \\
&= (1-\pi_i)^2 \pi_i + (0-\pi_i)^2 (1-\pi_i) \\
&= \pi_i (1-\pi_i)[(1-\pi_i) + \pi_i] \\
&= \pi_i (1-\pi_i) \\
&= (E(Y_i))(1-E(Y_i)) \\
&= (\beta_0 + \beta_1 X_i)(1-\beta_0 - \beta_1 X_i) \\
&= Var(\varepsilon_i) \qquad (4)
\end{aligned}$$

Από την σχέση βλέπουμε πως η διασπορά των σφαλμάτων εξαρτάται από τα X_i, άρα η τιμή της διασποράς θα είναι διαφορετική για κάθε διαφορετικό X_i

ΓΙΑΤΙ ΌΧΙ ΓΡΑΜΜΙΚΟ ΜΟΝΤΕΛΟ

- **Περιορισμός στη συνάρτηση απόκρισης**

Η συνάρτηση απόκρισης επειδή παριστάνει πιθανότητες θα πρέπει να ισχύει ο περιορισμός

$$0 \le E(Y) = \pi \le 1.$$

ΑΠΛΟ ΛΟΓΙΣΤΙΚΟ ΜΟΝΤΕΛΟ

Το μοντέλο που χρησιμοποιούμε όταν η Y_i είναι δίτιμη είναι το λογιστικό, το οποίο ορίζεται ως εξής:

$$Y_i = E\left(Y_i\right) + \varepsilon_i$$

όπου Y_i ανεξάρτητη τ.μ. Bernoulli

$$E\left(Y_i\right) = \pi_i = \frac{e^{(\beta_0 + \beta_1 X_i)}}{1 + e^{(\beta_0 + \beta_1 X_i)}} = \left[1 + e^{(-\beta_0 - \beta_1 X_i)}\right]^{-1}$$

ΑΠΛΟ ΛΟΓΙΣΤΙΚΟ ΜΟΝΤΕΛΟ

Είδαμε πως η αναμενόμενη συνάρτηση πρέπει να παίρνει τιμές στο διάστημα [0,1]

$$E(Y_i) = \beta_0 + \beta_1 X_i = \pi_i \, .$$

Οι τιμές όμως της $E(Y_i)$ κυμαίνονται σε όλο το σύνολο των πραγματικών αριθμών.

Για να αντιμετωπίσουμε αυτό το πρόβλημα, μια σκέψη θα ήταν να αντικαταστήσουμε την πιθανότητα π_i της επιτυχίας του γεγονότος με τη σχετική πιθανότητα επιτυχίας, δηλαδή με το λόγο της πιθανότητας επιτυχίας του γεγονότος προς την πιθανότητα αποτυχίας του γεγονότος

$$\frac{\pi_i}{1-\pi_i} \, .$$

ΑΠΛΟ ΛΟΓΙΣΤΙΚΟ ΜΟΝΤΕΛΟ

- Το μοντέλο

$$\frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 X_i$$

πάλι δεν είναι απόλυτα σωστό γιατί παίρνει τιμές από (0, +∞). Επομένως προτείνεται ο μετασχηματισμός

$$\pi_i' = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$$

$$\pi_i' = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i$$

ΑΠΛΟ ΛΟΓΙΣΤΙΚΟ ΜΟΝΤΕΛΟ

Έχουμε

$$\pi_i' = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i$$

$$\Leftrightarrow \frac{\pi_i}{1-\pi_i} = e^{\beta_0+\beta_1 X_i} \Leftrightarrow \pi_i = e^{\beta_0+\beta_1 X_i}\left(1-\pi_i\right)$$

$$\Leftrightarrow \pi_i + \pi_i e^{\beta_0+\beta_1 X_i} = e^{\beta_0+\beta_1 X_i} \Leftrightarrow \pi_i\left(1+e^{\beta_0+\beta_1 X_i}\right) = e^{\beta_0+\beta_1 X_i}$$

$$\Leftrightarrow \pi_i = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}$$

$$\Leftrightarrow \boxed{E\left(Y_i\right) = \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}}$$

ΑΠΛΟ ΛΟΓΙΣΤΙΚΟ ΜΟΝΤΕΛΟ

Επίσης ισχύει

$$E(Y_i) = \left(1 + e^{-\beta_0 - \beta_1 X_i}\right)^{-1}:$$

Απόδειξη

$$E(Y_i) = \frac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}} = \left(\frac{1 + e^{\beta_0 + \beta_1 X_i}}{e^{\beta_0 + \beta_1 X_i}}\right)^{-1} = \left(\frac{1}{e^{\beta_0 + \beta_1 X_i}} + \frac{e^{\beta_0 + \beta_1 X_i}}{e^{\beta_0 + \beta_1 X_i}}\right)^{-1}$$

$$= \left(e^{-\beta_0 - \beta_1 X_i} + 1\right)^{-1}$$

ΑΠΛΟ ΛΟΓΙΣΤΙΚΟ ΜΟΝΤΕΛΟ

Ορισμός

Ο λόγος $\dfrac{\pi_i}{1-\pi_i}$

ονομάζεται odds ενώ ο μετασχηματισμός $\pi_i' = \ln\left(\dfrac{\pi_i}{1-\pi_i}\right)$

ονομάζεται logit μετασχηματισμός της πιθανότητας

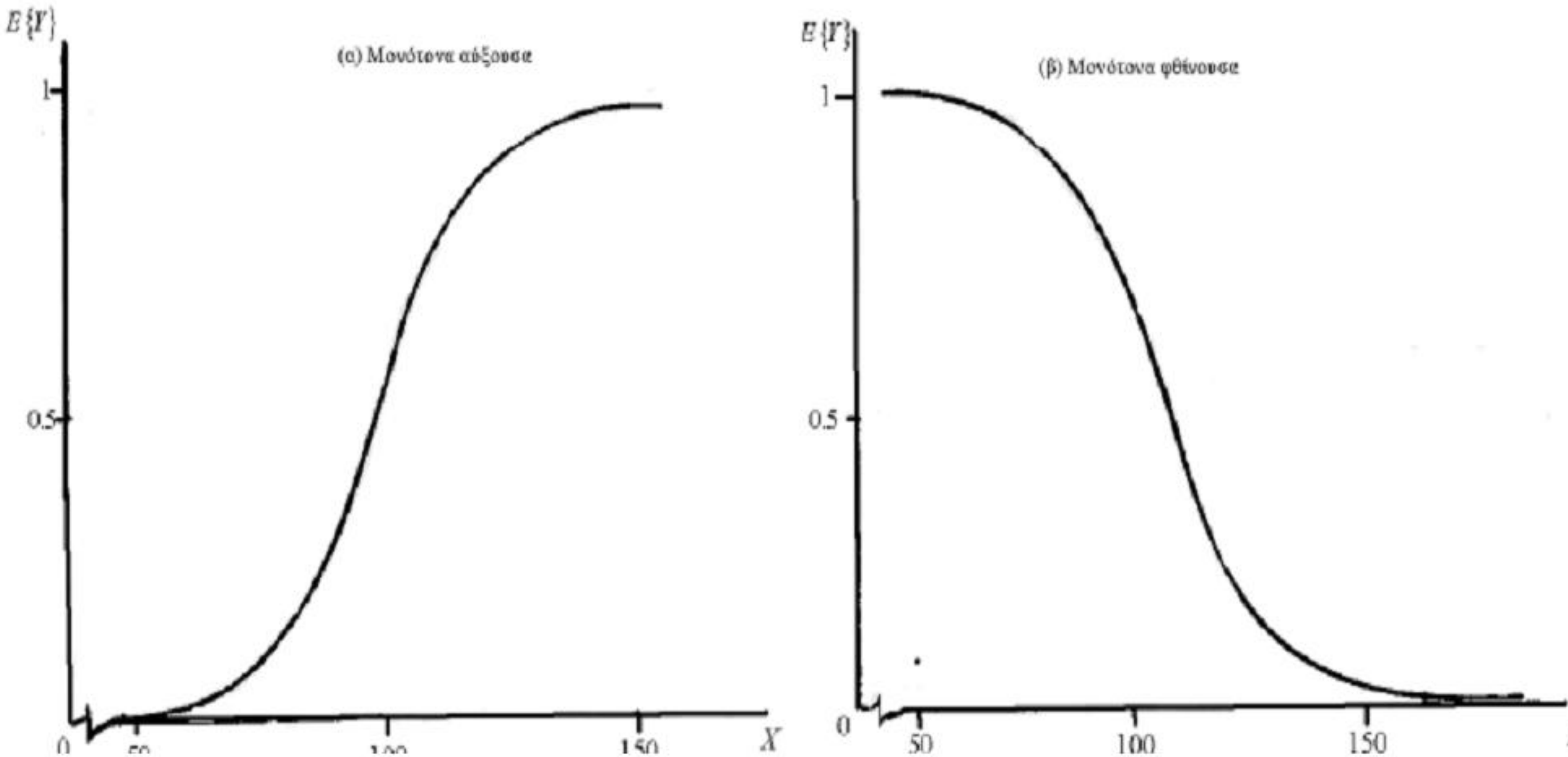
Η αναμενόμενη λογιστική συνάρτηση είναι:

a) Είτε μονότονα αύξουσα συνάρτηση είτε μονότονα φθίνουσα,

b) Είναι σχεδόν γραμμική στην περιοχή [0.2, 0.8],

c) Πλησιάζει το 0 και 1 στις ακραίες τιμές της εμβέλειας του X

ΑΠΛΟ ΛΟΓΙΣΤΙΚΟ ΜΟΝΤΕΛΟ



ΣΥΝΑΡΤΗΣΗ ΠΙΘΑΝΟΦΑΝΕΙΑΣ

Αφού τα Y_i είναι τυχαίες μεταβλητές Bernoulli όπου

$$\mathrm{P}(Y_i = 1) = \pi_i$$

$$\mathrm{P}(Y_i = 0) = 1 - \pi_i$$

η συνάρτηση πυκνότητας πιθανότητας είναι:

$$f_i(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \qquad Y_i = 0,1 \quad \text{και} \quad i = 1,\ldots,n$$

ΣΥΝΑΡΤΗΣΗ ΠΙΘΑΝΟΦΑΝΕΙΑΣ

Οι παρατηρήσεις Y_i είναι ανεξάρτητες οπότε η από κοινού συνάρτησης πιθανότητας θα είναι:

$$g(Y_1,\ldots\ldots,Y_n)=\prod_{i=1}^{n} f_i(Y_i)=\prod_{i=1}^{n}\pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$$

$$\ln g(Y_1,\ldots\ldots,Y_n)=\ln\prod_{i=1}^{n}\pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$$

$$=\sum_{i=1}^{n}\left[Y_i\ln\left(\frac{\pi_i}{1-\pi_i}\right)\right]+\sum_{i=1}^{n}\ln(1-\pi_i)$$

$$\pi_i'=\ln\left(\frac{\pi_i}{1-\pi_i}\right)=\beta_0+\beta_1X_i \qquad E(Y_i)=\pi_i=\frac{e^{(\beta_0+\beta_1X_i)}}{1+e^{(\beta_0+\beta_1X_i)}}=\left[1+e^{(-\beta_0-\beta_1X_i)}\right]^{-1}$$

ΣΥΝΑΡΤΗΣΗ ΠΙΘΑΝΟΦΑΝΕΙΑΣ

$$\ln L(\beta_0, \beta_1) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) + \sum_{i=1}^{n} \ln\left(1 - \frac{e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}\right)$$

$$= \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) + \sum_{i=1}^{n} \ln\left(\frac{1+e^{\beta_0+\beta_1 X_i} - e^{\beta_0+\beta_1 X_i}}{1+e^{\beta_0+\beta_1 X_i}}\right)$$

$$= \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) + \sum_{i=1}^{n} \ln\left(1+e^{\beta_0+\beta_1 X_i}\right)^{-1}$$

$$= \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} \ln\left(1+e^{\beta_0+\beta_1 X_i}\right)$$

Εκτιμήσεις β_0 και β_1 $\longrightarrow$ $\hat{\pi} = \dfrac{e^{b_0+b_1 X}}{1+e^{b_0+b_1 X}}$ $\longrightarrow$ $\hat{\pi}' = \ln\left(\dfrac{\hat{\pi}}{1-\hat{\pi}}\right)$

$$\hat{\pi}' = b_0 + b_1 X$$

ΕΡΜΗΝΕΙΑ

συντελεστής παλινδρόμησης b_1

Η ερμηνεία προέρχεται από την ιδιότητα που έχει ο εκτιμώμενος λόγος πιθανοτήτων (odds) $\frac{\pi_i}{1-\pi_i}$ ο οποίος πολλαπλασιάζεται με το e^{b_1} για κάθε μοναοα που αυξάνεται το Χ

$$OR = \frac{odds_2}{odds_1} = e^{b_1}$$

Αν το b_1 είναι θετικό, ο παράγοντας e^{b_1} είναι μεγαλύτερος από τη μονάδα, δηλαδή ο εκτιμώμενος λόγος πιθανοτήτων αυξάνεται. Αν το b_1 είναι αρνητικό, ο παράγοντας e^{b_1} είναι μικρότερος της μονάδας, και άρα ο εκτιμώμενος λόγος πιθανοτήτων μειώνεται.