Ερμηνεία των συντελεστών β

Στο logistic παράδειγμα είχαμε:

$$\log\frac{p}{1-p} = b_0 + b_1X_1 + b_2X_2$$

↳ αύξηση κατά 1 του X_2 → αύξηση της $\log\frac{p}{1-p}$ κατά β_2 (αθροιστικά)

↳ $\frac{p}{1-p} = \exp\{b_0+b_1X_1+b_2X_2\}$ → αύξηση της $\frac{p}{1-p}$ κατά e^{b_2} (πολ/κά).

↳ $p = \dfrac{e^{b_0+b_1X_1+b_2X_2}}{1+e^{b_0+b_1X_1+b_2X_2}} \longrightarrow \dfrac{\partial p}{\partial X_2} = p\cdot(1-p)\,b_2$

Συνεπώς, η μεταβολή στη X_2 επιφέρει μεγαλύτερη μεταβολή στη πιθανότητα αν p κοντά στο 0.5 παρά όταν είναι κοντά στο 0 ή στο 1.

↳ Με άλλα λόγια, $P=P(x)$ είναι συνάρτηση του X.

Οπότε: $\text{odds}(x) = \dfrac{P(x)}{1-P(x)} \longrightarrow \log(\text{odds}(x)) = b_0 + b_1X$

Αν: $X \longrightarrow X+1$ τότε $\longrightarrow \text{odds}(x) = e^{b_0+b_1X}$

↳ $\text{odds}(x+1) = e^{b_0+b_1(x+1)} = \text{odds}(x)\cdot e^{b_1}$

Οπότε: odds (ratio) $= e^{b_1}$ και

$$\log(\text{odds}(\text{ratio})) = \log\frac{\text{odds}(x+1)}{\text{odds}(x)} = b$$

Οπότε: $\frac{odds(x+1) - odds(x)}{odds(x)} = e^{b} - 1 \cong b$

αφού θεωρήσαμε τη προσέγγιση:

$e^{b} \cong 1 + b$

Διωνυμικά Δεδομένα (Binary data).

y: εξαρτημένη μεταβλητή

$y_1, y_2, \dots, y_n$ ανεξ. παρατηρήσεις από $Y_1, Y_2, \dots, Y_n$

και:

y_i: αριθμός επιτυχιών σε N_i δοκιμές με πιθανότητα p_i

Οπότε: $y_i = \sum_{j=1}^{N_i} Z_{ij}$ όπου: $Z_{ij} = \begin{cases} 1, & \text{με πιθ. } p_i \\ 0, & \text{με πιθ. } 1-p_i \end{cases}$

Συνεπώς: $f_Y(y; N, p) = \binom{N}{y} p^{y} \cdot (1-p)^{N-y}$

με: $E(Y) = Np$ και $V(Y) = Np(1-p)$

↳ Για κάθε i παρατηρούμε ένα διάνυσμα ανεξάρτητων μεταβλητών $\underline{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ik} \end{pmatrix}$.

Έτσι έχουμε:

(a) Δεδομένα κατ' άτομο

	Y_i	x_1	$\dots$	x_k
άτομο 1 :	1	---	---	---
άτομο 2 :	0	---	---	---
$\vdots$	$\vdots$	$\vdots$		$\vdots$

Σε αυτή τη περίπτωση (κατ' άτομο) έχουμε $N_i = 1$. Συνεπώς Bernoulli data (β) κατά συνδυασμό των ανεξάρτητων μεταβ.

Έχουμε:

y_i	N_i	x_1	-----	x_k
1	2	-----	-----	-----
3	5	-----	-----	-----
⋮	⋮	⋮		⋮

Όλοι οι πιθανοί συνδυασμοί των ανεξ. μεταβλητών.

Συνήθως: Διωνυμικά δεδομένα

↳ Η εναλλακτική παρουσίαση γίνεται λόγω της μείωσης του αριθ. των γραμμών (πχ. παρουσιάζουμε σε μια γραμμή όλα τα άτομα με την ίδια τιμή για την μεταβ. x_1 ... κλπ)

↳ Δε μπορεί να χρησιμοποιηθεί για συνεχή δεδομένα. (ανεξ. μεταβλητές).

↳ Ζητάμε λοιπόν μια σχέση ανάμεσα στο P_i και X_i's

↳ Παίρνουμε τον μετασχηματισμό των "odds"

άρα: $\underbrace{\text{logit}(P_i)}_{g(P_i)} = \log \frac{P_i}{1-P_i} = \underset{\sim}{x}_i' \underset{\sim}{b} \quad \in (-\infty, +\infty).$

Συμπερασματολογία

Υπόθ: $f(y_i;P_i) \propto P_i^{y_i}(1-P_i)^{N_i-y_i}$

Log-likelihood

$$\log f(\underline{y};P_i) = \sum_{i=1}^{n} y_i \log(P_i) + (N_i-y_i)\log(1-P_i)$$

Η ποσότητα αυτή (για συγκεκριμένο i) ανήκει στην ΕΟΚ με κανονική παράμετρο

$$\theta_i = \log\frac{P_i}{1-P_i} \quad \text{και} \quad b(\theta_i) = \log(1+e^{\theta_i})\cdot N_i$$

Άρα $\mu_i = E(Y_i) = b'(\theta_i) = N_i\dfrac{e^{\theta_i}}{1+e^{\theta_i}} = N_iP_i$

και κανονικό link: $g(\mu) = b'^{-1}(P_i) = \log\dfrac{P_i}{1-P_i}$

↳ Επίλυση των $\hat{\beta}$ με Fisher's Scoring

↳ Επίλυση των $\hat{V}(\hat{\beta})$ για Ε.Υ./Δ.Ε.

Deviance

$$Los = 2 \cdot \sum_{i=1}^{n} \frac{y_i\left(\hat{\theta}_i^{(s)} - \hat{\theta}_i^{(0)}\right) - \left(b(\hat{\theta}_i^{(s)}) - b(\hat{\theta}_i^{(0)})\right)}{a(\varphi_i)}$$

όπου: $\theta_i = \log \frac{p_i}{1-p_i}$, $a(\varphi_i) = 1$

και: $b(\theta_i) = N_i \log(1 + e^{\theta_i}) = -N_i \log(1-p_i)$

Άρα: $$Los = 2 \cdot \sum_{i=1}^{n} \left[y_i \left(\log \frac{\hat{p}_i^{(s)}}{1-\hat{p}_i^{(s)}} - \log \frac{\hat{p}_i^{(0)}}{1-\hat{p}_i^{(0)}} \right) + N_i \left(\log(1-\hat{p}_i^{(s)}) - \log(1-\hat{p}_i^{(0)}) \right) \right] =$$

$$= 2 \cdot \sum_{i=1}^{n} \left[y_i \log \frac{\hat{p}_i^{(s)}}{\hat{p}_i^{(0)}} + (N_i - y_i) \log \frac{1-\hat{p}_i^{(s)}}{1-\hat{p}_i^{(0)}} \right] =$$

$$= 2 \sum_{i=1}^{n} \left[y_i \log \frac{y_i/N_i}{\hat{p}_i^{(0)}} + (N_i - y_i) \log \frac{1 - y_i/N_i}{1-\hat{p}_i^{(0)}} \right] =$$

$$= 2 \sum_{i=1}^{n} \left[y_i \log \frac{y_i}{E_0(Y_i)} + (N_i - y_i) \log \frac{N_i - y_i}{N_i - E_0(Y_i)} \right]$$

αφού για το saturated μοντέλο ισχύει: $\hat{p}_i^{(s)} = \frac{y_i}{N_i}$

και: $N_i \hat{p}_i^{(0)} = E_0(Y_i)$ ο αναμενόμενος # επιτυχιών κάτω από το μηδενικό μοντέλο.

↳ Η Deviance συγκρίνει τους παρατηρούμενους αρ. επιτυχιών με τους αναμενόμενους για το συγκεκριμένο (λογιστικό) μοντέλο.

Έλεγχος καλής προσαρμογής.

↳ Κάτω από την H_0 έχουμε: $L_{OS} \overset{\text{ασυμ.}}{\sim} \chi^2_{n-p}$

↳(!) Θα πρέπει: $\sum_{i=1}^{n} N_i \to \infty$ και $N_i P_i \to \infty$

↳ Αν για κάποια i έχουμε: $N_i P_i \not\to \infty$ τότε χρησιμοποιούμε το χ^2 του Pearson

↳ Ομοίως, για ελέγχους συνολικής σημαντικότητας ισχύει: $L_{OS} - L_{1S} \sim \chi^2_1$ ή χ^2_{q-p}

Σημ. η προσέγγιση χ^2 είναι αρκετά ακριβής για διαφορές αποκλίσεων παρόλο που δεν είναι για τις ίδιες τις αποκλίσεις.

Άλλα Link Functions

↳ Η logit(p) δεν είναι η μοναδική link συν.

↳ Γενικώς οι link functions θα πρέπει να έχουν την ιδιότητα

$$g : (0,1) \longrightarrow \mathbb{R}$$

πχ: $g(p) = F^{-1}(p)$ όπου F αθρ. συνάρτηση κατανομής πιθανότητας τ.μ.

(παράδειγμα με κριτής συνεχής τ.μ.)

Link functions

α) $\text{logit}(p) = \log \frac{p}{1-p}$ -- αντίστροφη αθροιστικής της logistic κατανομής

β) $\text{probit}(p) = \Phi^{-1}(p)$ --- της κανονικής

γ) comp log-log : $\log(-\log(1-p))$ .... log(εκθετικής)