La regressione logistica multinomiale

Introduzione

Il modello di regressione logistica puó essere generalizzato per considerare il caso nel quale la variabile dipendente Y é policotomica e assume m+1 modalitá.

- $0, 1, \ldots, m$: modalitá di Y;
- 0 categoria di riferimento;
- $x = (x_1, x_2, \ldots, x_p)$: vettore delle variabili esplicative;

Per ogni modalitá di Y si definisce la funzione logit:

$$g_1(x) = logit_1(x) = ln\left[\frac{P(Y=1|x)}{P(Y=0|x)}\right] = \beta_{10} + \beta_{11}x_1 + \ldots + \beta_{1p}x_p$$

$$g_2(x) = logit_2(x) = ln\left[\frac{P(Y=2|x)}{P(Y=0|x)}\right] = \beta_{20} + \beta_{21}x_1 + \ldots + \beta_{2p}x_p$$

$$g_3(x) = logit_3(x) = ln\left[\frac{P(Y=3|x)}{P(Y=0|x)}\right] = \beta_{30} + \beta_{31}x_1 + \ldots + \beta_{3p}x_p$$

$$g_m(x) = logit_m(x) = ln\left[\frac{P(Y=m|x)}{P(Y=0|x)}\right] = \beta_{m0} + \beta_{m1}x_1 + \ldots + \beta_{mp}x_p$$

Tutte le modalitá sono confrontate con quella di riferimento. L'i-esimo logit esprime la propensione della variabile dipendente ad assumere l'i-esima modalitá rispetto alla modalitá di riferimento 0 in corrispondenza del valore x del vettore delle variabili esplicative.

$$g_3(x) = logit_3(x) = ln\left[\frac{P(Y=3|x)}{P(Y=0|x)}\right] = \beta_{30} + \beta_{31}x_1 + \ldots + \beta_{3p}x_p$$

$$g_m(x) = logit_m(x) = ln\left[\frac{P(Y=m|x)}{P(Y=0|x)}\right] = \beta_{m0} + \beta_{m1}x_1 + \ldots + \beta_{mp}x_p$$

Tutte le modalitá sono confrontate con quella di riferimento. L'i-esimo logit esprime la propensione della variabile dipendente ad assumere l'i-esima modalitá rispetto alla modalitá di riferimento 0 in corrispondenza del valore x del vettore delle variabili esplicative.

Le probabilitá con le quali la variabile dipendente assume le diverse modalitá, condizionate ad x, risultano:

$$\pi_0(x) = P(Y = 0|x) = \frac{1}{1 + exp[g_1(x)] + \ldots + exp[g_m(x)]}$$

$$\pi_1(x) = P(Y = 1|x) = \frac{exp[g_1(x)]}{1 + exp[g_1(x)] + \ldots + exp[g_m(x)]}$$

$$\pi_m(x) = P(Y = m|x) = \frac{exp[g_m(x)]}{1 + exp[g_1(x)] + \ldots + exp[g_m(x)]}$$

Per stimare il modello la variabile dipendente é codificata mediante $m+1$ variabili dicotomiche y_j tali che $y_j = 1$ quando Y assume la j-esima modalitá e 0 altrimenti.
La funzione di verosimiglianza é data da

$$L(\beta) = \prod_{i=1}^{n} [\pi_0(x_i)^{y_{0i}} \pi_1(x_i)^{y_{1i}} \dots \pi_m(x_i)^{y_{mi}}]$$

e viene massimizzata numericamente.

Regressione logistica ordinale

Quando la variabile dipendente assume modalitá ordinate é opportuno utilizzare il modello della regressione logistica ordinale.
Si possono distinguere tre possibili modelli:

- Proportional-odds model
- Continuation-ratio model
- Adiacent-category model

Proportional-odds model

Nel proportional-odds model per ogni modalitá j della variabile dipendente Y si considera il logit fra le categorie successive a j rispetto alle precedenti:

$$c_j(x) = ln\left[\frac{P(Y > j|x)}{P(Y \leq j|x)}\right]$$

Esiste inoltre una variabile latente Y^* che dipende dai regressori. L'insieme dei valori di Y^* é suddiviso in $m+1$ sottoinsiemi sulla base di quelli che sono dei valori soglia $((-\infty, \tau_1), (\tau_1, \tau_2), \ldots, (\tau_m, +\infty))$

Il logit risulterá essere allora:

$$c_j(x) = ln\left[\frac{P(Y > j|x)}{P(Y \leq j|x)}\right]$$

Proportional-odds model

Nel proportional-odds model per ogni modalitá j della variabile dipendente Y si considera il logit fra le categorie successive a j rispetto alle precedenti:

$$c_j(x) = ln\left[\frac{P(Y > j|x)}{P(Y \leq j|x)}\right]$$

Esiste inoltre una variabile latente Y^* che dipende dai regressori. L'insieme dei valori di Y^* é suddiviso in $m+1$ sottoinsiemi sulla base di quelli che sono dei valori soglia $((-\infty, \tau_1), (\tau_1, \tau_2), \ldots, (\tau_m, +\infty))$
Il logit risulterá essere allora:

$$c_j(x) = ln\left[\frac{P(Y > j|x)}{P(Y \leq j|x)}\right]$$

Continuation-ratio model

Nel continuation-ratio model si considera il logit fra ciascuna categoria della variabile dipendente e le precedenti. Dato il valore delle variabili esplicative si ha:

$$r_j(x) = ln\left[\frac{P(Y=j|x)}{P(Y<j|x)}\right] = \beta_{j0} + \beta_{j1}x_1 + \ldots + \beta_{jp}x_p$$

Al variare della modalitá della variabile dipendente variano i parametri del logit, quindi il numero di parametri da stimare sará $p \times m$ e coincide con quello del modello multinomiale.
Questo modello é stimato mediante m regressioni logistiche binarie nelle quali si confronta la j-esima categoria della variabile dipendente rispetto alle precedenti. Il valore della log-verosimiglianza é dato dalla somma dei valori ottenuti nelle singole regressioni.

Adjacent-category model

Questo modello é utile nel confronto fra successive modalitá della variabile dipendente. Il logit fra categorie adiacenti é dato da

$$a_j(x) = ln\left[\frac{P(Y=j|x)}{P(Y=j-1|x)}\right] = \beta_0 + x_1\beta_1 + \ldots + x_p\beta_p$$

Il logit rispetto alla categoria di riferimento é dato da

$$ln\left[\frac{P(Y=j|x)}{P(Y=0|x)}\right] =$$

$$ln\left[\frac{P(Y=j|x)}{P(Y=0|x)}\right] + ln\left[\frac{P(Y=2|x)}{P(Y=1|x)}\right] + \ldots + ln\left[\frac{P(Y=j|x)}{P(Y=j-1|x)}\right]$$

Adjacent-category model

Questo modello é utile nel confronto fra successive modalitá della variabile dipendente. Il logit fra categorie adiacenti é dato da

$$a_j(x) = ln\left[\frac{P(Y=j|x)}{P(Y=j-1|x)}\right] = \beta_0 + x_1\beta_1 + \ldots + x_p\beta_p$$

Il logit rispetto alla categoria di riferimento é dato da

$$ln\left[\frac{P(Y=j|x)}{P(Y=0|x)}\right] =$$

$$ln\left[\frac{P(Y=j|x)}{P(Y=0|x)}\right] + ln\left[\frac{P(Y=2|x)}{P(Y=1|x)}\right] + \ldots + ln\left[\frac{P(Y=j|x)}{P(Y=j-1|x)}\right]$$