La regressione logistica

Introduzione

Si vuole descrivere la relazione di dipendenza del possesso di un attributo dicotomico da una o piú variabili indipendenti di natura qualsiasi. Alcuni esempi di attributi dicotomici:

- per un soggetto che ha ottenuto un prestito, la restituzione/non-restituzione del prestito;
- per una banca, il fallimento/sopravvivenza dopo un dato periodo di tempo;
- per un cliente, il riscontro positivo/negativo ad un'offerta promozionale;
- per un paziente sotto osservazione, la presenza/assenza di una data malattia.

Gli obiettivi possono essere molteplici:

- individuare tra le variabili indipendenti quelle a maggiore potere esplicativo, che vanno quindi interpretate come determinanti del possesso o meno dell'attributo: a seconda che siano correlate positivamente o negativamente con il fenomeno studiato possono essere considerate rispettivamente come fattori di rischio o come fattori di protezione;
- ricercare la combinazione lineare delle variabili indipendenti che meglio discrimina fra il gruppo delle unitá che possiedono l'attributo e quello delle unitá che non lo possiedono;
- stimare la probabilitá del possesso dell'attributo per una nuova unitá statistica su cui é stato osservato il vettore di variabili X e, fissato per tale probabilitá un valore soglia, classificare l'unitá alla categoria delle unitá che possiedono l'attributo o a quello delle unitá che non lo possiedono.

Assunzioni e specificazioni del modello

Costruzione di un modello di regressione per Y, variabile risposta dicotomica con valori 0 e 1, corrispondenti rispettivamente all'assenza e alla presenza dell'attributo.

In un modello di regressione la quantitá che si ipotizza funzione di X é il valore medio aritmetico della variabile dipendente Y condizionato ad un dato x, $E(Y|x)$. Nel caso del modello di regressione logistica, questo valor medio condizionato corrisponde a $P(Y = 1|x)$, cioé alla probabilitá di possedere l'attributo in esame condizionata al fatto che il vettore delle variabili indipendenti assume valore x.

Assunzioni e specificazioni del modello

Costruzione di un modello di regressione per Y, variabile risposta dicotomica con valori 0 e 1, corrispondenti rispettivamente all'assenza e alla presenza dell'attributo.

In un modello di regressione la quantitá che si ipotizza funzione di X é il valore medio aritmetico della variabile dipendente Y condizionato ad un dato x, $E(Y|x)$. Nel caso del modello di regressione logistica, questo valor medio condizionato corrisponde a $P(Y = 1|x)$, cioé alla probabilitá di possedere l'attributo in esame condizionata al fatto che il vettore delle variabili indipendenti assume valore x.

Si vuole descrivere la funzione che lega tale probabilitá, indicata con $\pi(x)$, alla combinazione delle variabili indipendenti. Il modello di regressione per Y é dunque:

$$Y = \pi(x) + \varepsilon$$

Un modello di regressione lineare sarebbe del tutto inappropriato a questo scopo. Una funzione lineare di X, essendo non limitata (né inferiormente, né superiormente), potrebbe dare luogo a valori stimati di $\pi(x)$ esterni all'intervallo $[0, 1]$, e quindi privi di senso.

Si vuole descrivere la funzione che lega tale probabilitá, indicata con $\pi(x)$, alla combinazione delle variabili indipendenti. Il modello di regressione per Y é dunque:

$$Y = \pi(x) + \varepsilon$$

Un modello di regressione lineare sarebbe del tutto inappropriato a questo scopo. Una funzione lineare di X, essendo non limitata (né inferiormente, né superiormente), potrebbe dare luogo a valori stimati di $\pi(x)$ esterni all'intervallo $[0, 1]$, e quindi privi di senso.

Nel modello di regressione lineare l'errore si distribuisce normalmente, con media nulla e varianza costante. Questa assunzione non é valida quando Y é una variabile dicotomica, perché in tal caso l'errore puó assumere solo 2 valori:

$$\varepsilon = Y - \pi(x) = \begin{cases} 1 - \pi(x) & \text{con prob} \quad \pi(x) \\ -\pi(x) & \text{con prob} \quad 1 - \pi(x) \end{cases}$$

con media

$$E(\varepsilon) = [1 - \pi(x)]\pi(x) - \pi(x)[1 - \pi(x)] = 0$$

e varianza

$$Var(\varepsilon) = [1 - \pi(x)]^2\pi(x) + \pi(x)^2[1 - \pi(x)] = \pi(x)[1 - \pi(x)]$$

che dipende dal valore di X e quindi non é costante

Nel modello di regressione lineare l'errore si distribuisce normalmente, con media nulla e varianza costante. Questa assunzione non é valida quando Y é una variabile dicotomica, perché in tal caso l'errore puó assumere solo 2 valori:

$$\varepsilon = Y - \pi(x) = \begin{cases} 1 - \pi(x) & \text{con prob} \quad \pi(x) \\ -\pi(x) & \text{con prob} \quad 1 - \pi(x) \end{cases}$$

con media

$$E(\varepsilon) = [1 - \pi(x)]\pi(x) - \pi(x)[1 - \pi(x)] = 0$$

e varianza

$$Var(\varepsilon) = [1 - \pi(x)]^2\pi(x) + \pi(x)^2[1 - \pi(x)] = \pi(x)[1 - \pi(x)]$$

che dipende dal valore di X e quindi non é costante

Il modello

Per descrivere la relazione di dipendenza della probabilitá

$$\pi(x) = P(Y = 1|x)$$

dai valori di $X = (X_1, X_2, \ldots, X_p)$ si puó usare la distribuzione logistica.

$$\pi(x) = \frac{e^{\eta}}{1 + e^{\eta}}$$

η puó essere espresso come $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$

$$\pi(x) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_J}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_J}}$$

Il grafico della precedente funzione descrive una curva monotona a forma di S allungata (sigmoide), limitata superiormente dalla retta y=1 e inferiormente dalla retta y=0, alle quali tende asintoticamente.

Il modello

Per descrivere la relazione di dipendenza della probabilitá

$$\pi(x) = P(Y = 1|x)$$

dai valori di $X = (X_1, X_2, \ldots, X_p)$ si puó usare la distribuzione logistica.

$$\pi(x) = \frac{e^{\eta}}{1 + e^{\eta}}$$

η puó essere espresso come $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$

$$\pi(x) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_J}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_J}}$$

Il grafico della precedente funzione descrive una curva monotona a forma di S allungata (sigmoide), limitata superiormente dalla retta y=1 e inferiormente dalla retta y=0, alle quali tende asintoticamente.

Il modello

Per descrivere la relazione di dipendenza della probabilitá

$$\pi(x) = P(Y = 1|x)$$

dai valori di $X = (X_1, X_2, \ldots, X_p)$ si puó usare la distribuzione logistica.

$$\pi(x) = \frac{e^{\eta}}{1 + e^{\eta}}$$

η puó essere espresso come $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$

$$\pi(x) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_J}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_J}}$$

Il grafico della precedente funzione descrive una curva monotona a forma di S allungata (sigmoide), limitata superiormente dalla retta y=1 e inferiormente dalla retta y=0, alle quali tende asintoticamente.

Il modello

Per descrivere la relazione di dipendenza della probabilitá

$$\pi(x) = P(Y = 1|x)$$

dai valori di $X = (X_1, X_2, \ldots, X_p)$ si puó usare la distribuzione logistica.

$$\pi(x) = \frac{e^{\eta}}{1 + e^{\eta}}$$

η puó essere espresso come $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$

$$\pi(x) = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_J}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j x_J}}$$

Il grafico della precedente funzione descrive una curva monotona a forma di S allungata (sigmoide), limitata superiormente dalla retta y=1 e inferiormente dalla retta y=0, alle quali tende asintoticamente.

Considerando invece il seguente rapporto si ha la funzione definita logit:

$$logit(\pi(x)) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right]$$

Logaritmo del rapporto della probabilitá di possedere l'attributo con la probabilitá di non possederlo. Il rapporto é invece definito odds. Si dimostra che

$$logit(\pi(x)) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

che é quindi funzione lineare delle variabili esplicative.

Considerando invece il seguente rapporto si ha la funzione definita logit:

$$logit(\pi(x)) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right]$$

Logaritmo del rapporto della probabilitá di possedere l'attributo con la probabilitá di non possederlo. Il rapporto é invece definito odds. Si dimostra che

$$logit(\pi(x)) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

che é quindi funzione lineare delle variabili esplicative.

Considerando invece il seguente rapporto si ha la funzione definita logit:

$$logit(\pi(x)) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right]$$

Logaritmo del rapporto della probabilitá di possedere l'attributo con la probabilitá di non possederlo. Il rapporto é invece definito odds. Si dimostra che

$$logit(\pi(x)) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

che é quindi funzione lineare delle variabili esplicative.

I modelli lineari generalizzati

Il modello logistico appartiene alla famiglia dei modelli lineari generalizzati (GLM). Un modello di questo tipo mette in relazione una funzione del valore atteso della variabile dipendente Y con le variabili esplicative attraverso un'equazione lineare.

Esso é specificato da tre componenti:

- La componente aleatoria $(Y_1, Y_2, \ldots, Y_n)$ costituita da un insieme di variabili aleatorie assunte reciprocamente indipendenti e con distribuzione di probabilitá appartenente alla famiglia esponenziale;
- la componente sistematica $\sum_{j=1}^{p} \beta_j x_{ij}$ che specifica una combinazione lineare delle variabili esplicative nel modello;
- la funzione legame $g(E(Y_i)) = \sum_{j=1}^{p} \beta_j x_{ij}$ che mette in relazione la componente aleatoria e la componente sistematica del modello, specificando quale funzione g del valore atteso di Y_i dipende linearmente dalle variabili esplicative.

Specificando diverse funzioni come funzione legame si ottengono i seguenti casi particolari di modello lineare generalizzato:

- prendendo come funzione legame la funzione identitá $g(E(Y_i)) = E(Y_i)$ si ottiene $E(Y_i) = \sum_{j=1}^{p} \beta_j x_{ij}$ che é il tradizionale modello di regressione lineare.
- prendendo come funzione legame la funzione logit, $g(E(Y_i)) = ln\left[\frac{E(Y_i)}{1-E(Y_i)}\right]$ si ha $ln\left[\frac{E(Y_i)}{1-E(Y_i)}\right] = \sum_{j=1}^{p} \beta_j x_{ij}$ che, considerando Y_i dicotomica, é il modello di regressione logistica;
- prendendo come funzione legame la funzione logaritmo $g(E(Y_i)) = ln\left[E(Y_i)\right]$ si ha $ln\left[E(Y_i)\right] = \sum_{j=1}^{p} \beta_j x_{ij}$ che é definito modello log-lineare

Stima dei parametri

Poiché non vale l'omoschedasticitá dei residui non é possibile adottare il metodo di stima dei minimi quadrati. Si puó usare il metodo della massima verosimiglianza.

Per semplicitá si considera il modello con una sola variabile indipendente:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

che in termini di logit é

$$g(x) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x$$

Ricordando l'ipotesi di indipendenza reciproca delle variabili campionarie, la funzione di verosimiglianza del campione osservato sará:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} f(y_i|x_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\left[1-\pi(x_i)\right]^{(1-y_i)}$$

Stima dei parametri

Poiché non vale l'omoschedasticitá dei residui non é possibile adottare il metodo di stima dei minimi quadrati. Si puó usare il metodo della massima verosimiglianza.
Per semplicitá si considera il modello con una sola variabile indipendente:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

che in termini di logit é

$$g(x) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x$$

Ricordando l'ipotesi di indipendenza reciproca delle variabili campionarie, la funzione di verosimiglianza del campione osservato sará:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} f(y_i|x_i) = \prod_{i=1}^{n} \pi(x_i)^{y_i}\left[1-\pi(x_i)\right]^{(1-y_i)}$$

La funzione di log-verosimiglianza sará:

$$l(\beta_0, \beta_1) = \sum_{i=1}^{n} \{y_i ln\left[\pi(x_i)\right] + (1 - y_i) ln\left[1 - \pi(x_i)\right]\} =$$

$$= \left[y_i ln\left[\frac{\pi(x_i)}{1 - \pi(x_i)}\right] + ln\left[1 - \pi(x_i)\right]\right] =$$

$$= \left[y_i(\beta_0 + \beta_1 x_i) + ln\left(1 - \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}\right)\right] =$$

$$= \left[y_i(\beta_0 + \beta_1 x_i) + ln\left[\frac{1}{1 + e^{\beta_0 + \beta_1 x_i}}\right]\right]$$

$$= \sum_{i=1}^{n} \left[y_i(\beta_0 + \beta_1 x_i) - ln\left[1 + e^{\beta_0 + \beta_1 x_i}\right]\right]$$

Per la stima é necessario calcolare le derivate parziali e porle uguali a 0, ricavando il sistema delle equazioni di verosimiglianza.

Le equazioni che si ottengono non sono lineari nelle incognite e quindi la loro soluzione non é immediata, ma richiede l'impiego di metodi numerici iterativi.

Gli stimatori di massima verosimiglianza godono della proprietá di equivarianza rispetto a trasformazioni funzionali differenziabili. La stima di $\pi(x_i)$ risulta quindi:

$$\hat{\pi}(x_i) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}}$$

e rappresenta il valore di Y stimato dal modello di regressione logistica in corrispondenza di $X = x_i$.

Asintoticamente, sotto condizioni non particolarmente restrittive, gli stimatori di massima verosimiglianza sono corretti, normodistribuiti ed efficienti.

Per la stima é necessario calcolare le derivate parziali e porle uguali a 0, ricavando il sistema delle equazioni di verosimiglianza.

Le equazioni che si ottengono non sono lineari nelle incognite e quindi la loro soluzione non é immediata, ma richiede l'impiego di metodi numerici iterativi.

Gli stimatori di massima verosimiglianza godono della proprietá di equivarianza rispetto a trasformazioni funzionali differenziabili. La stima di $\pi(x_i)$ risulta quindi:

$$\hat{\pi}(x_i) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}$$

e rappresenta il valore di Y stimato dal modello di regressione logistica in corrispondenza di $X = x_i$.

Asintoticamente, sotto condizioni non particolarmente restrittive, gli stimatori di massima verosimiglianza sono corretti, normodistribuiti ed efficienti.

Per la stima é necessario calcolare le derivate parziali e porle uguali a 0, ricavando il sistema delle equazioni di verosimiglianza.

Le equazioni che si ottengono non sono lineari nelle incognite e quindi la loro soluzione non é immediata, ma richiede l'impiego di metodi numerici iterativi.

Gli stimatori di massima verosimiglianza godono della proprietá di equivarianza rispetto a trasformazioni funzionali differenziabili. La stima di $\pi(x_i)$ risulta quindi:

$$\hat{\pi}(x_i) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 x_i}}$$

e rappresenta il valore di Y stimato dal modello di regressione logistica in corrispondenza di $X = x_i$.

Asintoticamente, sotto condizioni non particolarmente restrittive, gli stimatori di massima verosimiglianza sono corretti, normodistribuiti ed efficienti.

Per la stima é necessario calcolare le derivate parziali e porle uguali a 0, ricavando il sistema delle equazioni di verosimiglianza.

Le equazioni che si ottengono non sono lineari nelle incognite e quindi la loro soluzione non é immediata, ma richiede l'impiego di metodi numerici iterativi.

Gli stimatori di massima verosimiglianza godono della proprietá di equivarianza rispetto a trasformazioni funzionali differenziabili. La stima di $\pi(x_i)$ risulta quindi:

$$\hat{\pi}(x_i) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_i}}$$

e rappresenta il valore di Y stimato dal modello di regressione logistica in corrispondenza di $X = x_i$.

Asintoticamente, sotto condizioni non particolarmente restrittive, gli stimatori di massima verosimiglianza sono corretti, normodistribuiti ed efficienti.

Verifica d'ipotesi

Le precedenti proprietá permettono di costruire opportune statistiche-test per il controllo di ipotesi sui parametri e per costruire intervalli di confidenza.

$$G = -2ln\frac{\text{veros. senza la var. di interesse}}{\text{veros. con la variabile}}$$

La precedente statistica é il test rapporto di verosimiglianza (likelihood ratio test).
Sotto l'ipotesi $H_0 : \beta_1 = 0$ che l'inserimento della variabile X nel modello non apporti un contributo significativo la variabile campionaria G si distribuisce asintoticamente come una variabile aleatoria $\chi^2_{(1)}$.
Confrontando il p-value corrispondente al valore di G, calcolato sul campione osservato, con un prefissato livello di significativitá, é possibile trarre le opportune conclusioni.

Verifica d'ipotesi

Le precedenti proprietá permettono di costruire opportune statistiche-test per il controllo di ipotesi sui parametri e per costruire intervalli di confidenza.

$$G = -2ln\frac{\text{veros. senza la var. di interesse}}{\text{veros. con la variabile}}$$

La precedente statistica é il test rapporto di verosimiglianza (likelihood ratio test).
Sotto l'ipotesi $H_0 : \beta_1 = 0$ che l'inserimento della variabile X nel modello non apporti un contributo significativo la variabile campionaria G si distribuisce asintoticamente come una variabile aleatoria $\chi^2_{(1)}$.
Confrontando il p-value corrispondente al valore di G, calcolato sul campione osservato, con un prefissato livello di significativitá, é possibile trarre le opportune conclusioni.

Interpretazione dei parametri

Nel modello di regressione lineare il valore del parametro rappresenta la variazione media della Y al crescere di un'unitá della X. Nel modello di regressione logistica l'interpretazione é differente.

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

In termini di logit

$$g(x) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x$$

e dunque

$$g(x+1) - g(x) = \beta_0 + \beta_1(x+1) - \beta_0 - \beta_1 x = \beta_1$$

β_1 esprime la variazione del logit corrispondente ad un incremento unitario di X.

Interpretazione dei parametri

Nel modello di regressione lineare il valore del parametro rappresenta la variazione media della Y al crescere di un'unitá della X. Nel modello di regressione logistica l'interpretazione é differente.

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

In termini di logit

$$g(x) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x$$

e dunque

$$g(x+1) - g(x) = \beta_0 + \beta_1(x+1) - \beta_0 - \beta_1 x = \beta_1$$

β_1 esprime la variazione del logit corrispondente ad un incremento unitario di X.

Interpretazione parametri

Variabile indipendente continua

Se la X é continua si ha

$$ln\left(\frac{\text{odds per X=x+1}}{\text{odds per X=x}}\right) = ln(\text{odds per X=x+1}) - ln(\text{odds per X=x}) =$$

$$g(x+1) - g(x) = \beta_0 + \beta_1(x+1) - (\beta_0 - \beta_1 x) = \beta_1$$

quindi l'odds ratio corrispondente ad un incremento unitario di X é uguale a e^{β_1}.
Se si vuole considerare un incremento di c unitá invece che un incremento unitario si ha:

$$ln\left(\frac{\text{odds per X=x+c}}{\text{odds per X=x}}\right) = \beta_0 + \beta_1(x+c) - (\beta_0 - \beta_1 x) = c\beta_1$$

e quindi l'odds ratio per un incremento di X pari a c unitá vale $e^{c\beta_1}$

Interpretazione parametri
Variabile indipendente dicotomica

X assumerá solo due valori e lo stesso vale per l'odds:

$$\frac{P(Y=1|X=0)}{1-P(Y=1|X=0)} = \frac{\pi(0)}{1-\pi(0)}$$

$$\frac{P(Y=1|X=1)}{1-P(Y=1|X=1)} = \frac{\pi(1)}{1-\pi(1)}$$

Il rapporto dei due valori, cioé l'odds ratio sará uguale a:

$$\text{odds ratio} = \frac{\pi(1)}{1-\pi(1)} \Big/ \frac{\pi(0)}{1-\pi(0)} = \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}}{\frac{1}{1+e^{\beta_0+\beta_1}}} \Bigg/ \frac{\frac{e^{\beta_0}}{1+e^{\beta_0}}}{\frac{1}{1+e^{\beta_0}}} = e^{\beta_0+\beta_1} / e^{\beta_0} = e^{\beta_1}$$

Interpretazione parametri

Variabile indipendente dicotomica

X assumerá solo due valori e lo stesso vale per l'odds:

$$\frac{P(Y=1|X=0)}{1-P(Y=1|X=0)} = \frac{\pi(0)}{1-\pi(0)}$$

$$\frac{P(Y=1|X=1)}{1-P(Y=1|X=1)} = \frac{\pi(1)}{1-\pi(1)}$$

Il rapporto dei due valori, cioé l'odds ratio sará uguale a:

$$\text{odds ratio} = \frac{\pi(1)}{1-\pi(1)} \Big/ \frac{\pi(0)}{1-\pi(0)} = \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}}{\frac{1}{1+e^{\beta_0+\beta_1}}} \Bigg/ \frac{\frac{e^{\beta_0}}{1+e^{\beta_0}}}{\frac{1}{1+e^{\beta_0}}} = e^{\beta_0+\beta_1} / e^{\beta_0} = e^{\beta_1}$$

Interpretazione parametri

Variabile indipendente categorica

$$ln\frac{\text{odds per X= modalitá i-esima}}{\text{odds per X= modalitá di riferimento}} =$$

$$= ln(\text{odds per X=modalitá i-esima}) - ln(\text{odds per X= modalitá di riferimento}) =$$

$$= g(D_1 = 0, \ldots, D_i = 1, \ldots, D_{k-1} = 0) - g(D_1 = 0, \ldots, D_i = 0, \ldots, D_{k-1} = 0)$$

$$= \beta_0, \beta_{1,1}0 + \ldots + \beta_{1,i}1 + \ldots + \beta_{1,k-1}0 - (\beta_0, \beta_{1,1}0 + \ldots + \beta_{1,j}0 + \ldots + \beta_{1,k-1}0) = \beta_{1,i}$$

Quindi l'odds ratio di questo gruppo rispetto al gruppo di riferimento é uguale a $e^{\beta_{1,i}}$

Valutazione bontá di adattamento

Esistono modi diversi per misurare la divergenza tra il valore osservato per la variabile risposta e il corrispondente valore stimato dal modello. In particolare i residui piú utilizzati sono il residuo di Pearson e il deviance residuo.

Per definirli é necessario introdurre una serie di misure.

- J: numero di combinazioni diverse di valori delle variabili indipendenti osservate sulle n unitá statistiche (numero di logit stimati).
- n_k: numero di unitá statistiche che portano una generica combinazione di valori x_k con $k = 1, \ldots, J$
- y_k e $\hat{y}_k$ numero osservato e numero stimato di unitá statistiche per cui $Y = 1$ in corrispondenza della combinazione x_k cioé

$$\hat{y}_k = n_k \hat{\pi}(x_k) = n_k \frac{e^{\hat{g}(x_k)}}{1 + e^{\hat{g}(x_k)}}$$

Valutazione bontá di adattamento

Esistono modi diversi per misurare la divergenza tra il valore osservato per la variabile risposta e il corrispondente valore stimato dal modello. In particolare i residui piú utilizzati sono il residuo di Pearson e il deviance residuo.

Per definirli é necessario introdurre una serie di misure.

- J: numero di combinazioni diverse di valori delle variabili indipendenti osservate sulle n unitá statistiche (numero di logit stimati).
- n_k: numero di unitá statistiche che portano una generica combinazione di valori x_k con $k = 1, \dots, J$
- y_k e $\hat{y}_k$ numero osservato e numero stimato di unitá statistiche per cui $Y = 1$ in corrispondenza della combinazione x_k cioé

$$\hat{y}_k = n_k \hat{\pi}(x_k) = n_k \frac{e^{\hat{g}(x_k)}}{1 + e^{\hat{g}(x_k)}}$$

Poiché ognuno dei J valori y_k corrisponde a una numerositá differente e ad una diversa probabilitá di successo $\hat{\pi}(x_k) = \hat{\pi}_k$, i residui $(y_k - \hat{y}_k)$ sono difficili da interpretare.
Si possono confrontare dividendo ciascun residuo per il corrispondente scarto quadratico medio ottenendo il residuo di Pearson, definito come:

$$r(y_k, \hat{\pi}_k) = \frac{y_k - n_k\hat{\pi}_k}{\sqrt{n_k\hat{\pi}_k(1 - \hat{\pi}_k)}}$$

Il corrispondente residuo di devianza Ã¨ dato da:

$$d(y_k, \hat{\pi}_k) = \left[2\left[y_k ln\left(\frac{y_k}{n_k\hat{\pi}_k}\right) + (n_k - y_k) ln\left(\frac{n_k - y_k}{n_k(1 - \hat{\pi}_k)}\right)\right]\right]^{1/2}$$

Le misure di adattamento globale che si basano su questi residui sono costruite come somma dei quadrati dei residui. Per i residui di Pearson si ha:

$$\chi^2 = \sum_k r(y_k, \hat{\pi}_k)^2$$

I residui di devianza portano invece alla devianza:

$$D = \sum_k d(y_k, \hat{\pi}_k)^2$$

La distribuzione asintotica di queste due statistiche nell'universo dei campioni nell'ipotesi che il modello adattato rappresenti adeguatamente i dati é quella di un $\chi^2_{(j-(p+1))}$.
Valori piccoli indicano un buon adattamento, mentre valori grandi suggeriscono che il divario tra l'osservato e l'atteso non é da attribuire al solo errore di campionamento.

Diagnostiche sui residui

L'ispezione dei residui consente in primo luogo di controllare la validitá delle assunzioni dalle quali l'analisi ha preso le mosse. Per esempio, é possibile controllare l'ipotesi di linearitá della relazione fra il logit$(P[Y = 1|X = x])$ e un dato regressore continuo X attraverso la rappresentazione grafica dei punti di coordinate $(x_k, \hat{y}_k)$.

Se la numerositá campionaria non é troppo elevata, puó essere utile analizzare un semplice grafico dei residui (in ordinata) corrispondenti alle varie uná statistiche (elencate in ascissa). Dato che in un buon modello i residui dovrebbero essere prossimi allo 0, l'utilitá di questo grafico sta nella possibilitá di evidenziare la presenza di residui grandi (in valore assoluto; di solito esterni all'intervallo $[-2, 2]$), cioé di valori che il modello non é in grado di spiegare.

Diagnostiche sui residui

Un altro grafico utile per valutare l'adeguatezza del modello é quello contenente i valori stimati in ascissa e i residui in ordinata: in un buon modello tali punti dovrebbero essere disposti casualmente intorno all'asse delle ascisse. Se invece si evidenziano andamenti particolari potrebbe non essere corretta la scelta del logit come funzione legame. Questa eventualitá puó rappresentare una spiegazione anche per comportamenti difformi dall'atteso nel grafico che controlla la normalitá dei residui. La ricerca di valori anomali puó essere effettuata anche valutando la differenza nella stima dei parametri conseguente all'esclusione dal data set di un'unitá alla volta.