Modelli a variabili dipendenti qualitative

Giulio Palomba

Agosto 2008

Modelli con variabili dipendenti discrete (qualitative):

Binomiali (Scelte binarie): $y = 1$ oppure $y = 0$ (Logit, Probit)

Multinomiali (Risultati ordinati):

a. Count data: $y = 0, 1, 2, \ldots$ (modello di Poisson, modello binomiale)

b. Modelli gerarchici: $y = 0, 1, 2, \ldots$ A ciascun valore è assegnata una posizione gerarchica, ma non è detto che la differenza tra $y = 0$ e $y = 1$ sia ad esempio la stessa che intercorre tra $y = 2$ e $y = 3$. La numerazione serve solamente a stilare una gerarchia.

Multinomiali (Risultati non ordinati):

a. Modelli multinomiali: $y = 0, 1, 2, \ldots$, ma non c'è gerarchia.

b. Modelli annidati

1 Modelli di scelta binaria

I modelli di scelta binaria hanno la caratteristica peculiare di avere la variable dipendente discreta che può assumere solamente i valori 1 oppure 0; tale variabile dipendente è perciò dicotòmica e si configura come una variabile dummy. Il problema principale che si cerca di risolvere attraverso l'utilizzo di questi modelli è quindi quello di spiegare i valori assunti dalla variabile dipendente dicotòmica attraverso un insieme di variabili esplicative.

I modelli a scelta binaria sono sono stati concepiti per l'analisi delle scelte individuali in quanto spesso, nella realtà economica, l'individuo deve effettuare la scelta tra due distinte alternative; in quest'approccio la sfera individuale diventa perciò preponderante rispetto alla dimensione temporale nella quale le scelte vengono compiute.

L'innovazione apportata da questa classe di modelli investe soprattutto l'obietttivo della stima: se nei modelli con variabili dipendenti continue si cercava di spiegare il "quanto", attraverso i modelli con variabili dipendenti discrete si tenta di spiegare il "se", quindi il modello di regressione deve giocoforza analizzare le probabilità delle scelte individuali.

Introducendo la notazione matematica, lo scopo è quello di costruire un modello del tipo

$$p_i = Pr(y_i = j) = F(x_i'\beta), \tag{1}$$

dove y_i è la variabile dipendente dicotòmica, x_i è un vettore colonna contenente le k variabili esplicative, β è il vettore k-dimensionale contenente i parametri incogniti, p_i rappresenta la probabilità che l'i-esimo individuo effettui la j-esima scelta[1]. Tutta l'equazione (1) è riferita al generico individuo i-esimo, con $i = 1, 2, \ldots, N$. La funzione $F(x_i'\beta)$ svolge il ruolo decisivo nel modello

$$y_i = F(x_i'\beta) + \varepsilon_i \tag{2}$$

[1] Nei modelli a scelta binaria le alternative per j sono soltanto $j = 0$ oppure $j = 1$.

poiché è quella relativa alla media condizionale $E(y_i \mid x_i)$. Si tenga presente che l'argomento $x_i'\beta$ della funzione ritorna uno scalare perché è il prodotto di due vettori conformabili di dimensone rispettivamente $(1 \times k)$ e $(k \times 1)$: in questo modo il dominio di $F(x_i'\beta)$ è l'insieme dei numeri reali $\mathbb{R}$. Dato che y_i può assumere solamente i valori 1 e 0, da questo approccio emerge che:

- le variabili esplicative x_i colgono le caratteristiche individuali (gli individui scelgono),
- i parametri contenuti nel vettore β determinano l'impatto delle variabili esplicative nelle scelte individuali,
- il metodo OLS è difficilmente applicabile; generalmente si ricorre perciò alla stima MLE.

Analiticamente si ha

$$\begin{aligned} E(y_i \mid x_i) &= 1 \cdot p_i + 0 \cdot (1 - p_i) \\ &= p_i \\ &= F(x_i'\beta). \end{aligned}$$

1.1 Modello di probabilità lineare

Nel modello di probabilità lineare la funzione di regressione è lineare quindi risulta $E(y_i \mid x_i) = F(x_i'\beta) = x_i'\beta$; l'equazione (2) diventa perciò

$$y_i = x_i'\beta + \varepsilon_i \tag{3}$$

nel quale tutte le ipotesi classiche sono rispettate. Dato che la variabile dipendente è una dummy, l'unica ipotesi che non è rispettata è quella di omoschedasticità, in quanto:

$$\begin{array}{llll} \text{- se } y_i = 1 & \Rightarrow & \varepsilon_i = 1 - x_i'\beta & \Rightarrow \quad Pr(\varepsilon_i = 1 - x_i'\beta) = p_i, \\ \text{- se } y_i = 0 & \Rightarrow & \varepsilon_i = -x_i'\beta & \Rightarrow \quad Pr(\varepsilon_i = -x_i'\beta) = 1 - p_i. \end{array}$$

Questa relazione genera eteroschedsticità, infatti

$$\begin{aligned} Var(\varepsilon_i) &= [(\varepsilon_i \mid y_i = 1) - E(\varepsilon_i)]^2 Pr(\varepsilon_i \mid y_i = 1) + [(\varepsilon_i \mid y_i = 0) - E(\varepsilon_i)]^2 Pr(\varepsilon_i \mid y_i = 0) \\ &= [1 - x_i'\beta]^2 x_i'\beta + [-x_i'\beta]^2 (1 - x_i'\beta) \\ &= x_i'\beta(1 - x_i'\beta) \\ &= p_i(1 - p_i) \end{aligned} \tag{4}$$

L'equazione (4) è quella della varianza di una variabile casuale bernoulliana[2] nella quale la probabilità che $y_i = 1$ sia p_i. L'errore non è omoschedastico in quanto dipende dai valori di x_i e del vettore dei parametri β, quindi lo stimatore OLS non rispetta il Teorema di Gauss-Markov[3]. Per ottenere una stima efficiente occorre perciò stimare un modello GLS con matrice dei pesi data da

$$\Omega = \text{diag}(\omega_i),$$

con $\omega_i = [x_i'\beta(1 - x_i'\beta)]^{1/2}$ e $i = 1, 2, \ldots, N$.

Il modello di probabilità lineare soffre di due problemi:

1. può accadere che $x_i'\beta(1 - x_i'\beta) < 0$, cosa non ammissibile per una varianza,
2. può accadere che $x_i'\beta \notin [0, 1]$, cosa non ammissibile per una probabilità.

[2] Si ricorda che la varianza della variabile casuale bernoulliana con funzione di probabilità $p(x) = p^x(1 - p)^{1-p}$ è data da $Var(X) = p(1 - p)$.

[3] In particolare lo stimatore OLS non è BLUE.

1.2 Modelli Logit e Probit

Per ovviare ai due problemi che affliggono il modello di probabilità lineare occorre trovare una forma funzionale per $F(x_i'\beta)$ tale per cui

$$\begin{cases} \lim_{x_i'\beta \to -\infty} F(x_i'\beta) = 0 \\ \\ \lim_{x_i'\beta \to +\infty} F(x_i'\beta) = 1. \end{cases} \tag{5}$$

La naturale candidata a svolgere questo ruolo è la funzione di ripartizione poiché $F(x_i'\beta) : \mathbb{R} \longrightarrow [0,1]$: ciò garantisce che $p_i = F(x_i'\beta) \in [0,1]$, quindi anche che $Var(\varepsilon_i) \geq 0$ per ogni i.

La scelta di $F(x_i'\beta)$ è decisiva in questo contesto e le due specificazioni principali introdotte in letteratura utilizzano la funzione di ripartizione di due variabili casuali continue con funzione di densità simmetrica e funzioni di ripartizioni non molto "differenti" (si vedano le Figura 1 e 2). In particolare si hanno:

- Modello **Probit**: distribuzione normale con $F(x_i'\beta) = \Phi(x_i'\beta)$

$$p_i = Pr(y_i = 1) = \Phi(x_i'\beta) = \int_{-\infty}^{x_i'\beta/\sigma} \frac{1}{2\pi} e^{-\frac{z^2}{2}} dz \tag{6}$$

- Modello **Logit**: distribuzione logistica con $F(x_i'\beta) = \Lambda(x_i'\beta)$

$$p_i = Pr(y_i = 1) = \Lambda(x_i'\beta) = \int_{-\infty}^{x_i'\beta/\sigma} \frac{e^z}{[1+e^z]^2} dz = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}} \tag{7}$$

  dove il parametro (scalare) σ è inserito per ovviare al problema dell'identificazione[4] durante la fase della stima. In entrambe le formulazioni $x_i'\beta \in (-\infty, +\infty)$ e le funzioni di media condizionale non sono lineari nelle variabli esplicative x_i.

Le variabili casuali normale standardizzata e logistica hanno rispettivamente funzione di densità $\phi(x_i'\beta)$ e $\lambda(x_i'\beta)$ simmetriche (si veda la Figura 1); da questa caratteristica deriva che

$$\begin{cases} \Phi(x_i'\beta) = 1 - \Phi(-x_i'\beta) \\ \\ \Lambda(x_i'\beta) = 1 - \Lambda(-x_i'\beta) = \frac{1}{1-e^{-x_i'\beta}}. \end{cases} \tag{8}$$

Inoltre, per la variabile casuale logistica, valgono le seguenti proprietà:

1. distribuzione molto simile alla distribuzione normale standardizzata, soprattutto per $x_i'\beta \in [-1.2, 1.2]$ (cfr. Greene),
2. la distribuzione è leptocurtica, cioè comprende una massa di probabilità lungo le code maggiore di quella che si ha per la variabile casuale normale (si veda la Figura 1): la probabilità associata agli eventi "estremi" ($y_i = 0$ oppure $y_i = 1$) è perciò maggiore nel Logit rispetto al Probit,
3. la varianza della variabile casuale logistica è pari a $\pi^2/3$,
4. è sempre possibile determinare analiticamente la primitiva nell'integrale della curva di densità, cosa non possibile per a variabile casuale normale[5],
5. la funzione di densità è data da
$$\lambda(x_i'\beta) = \Lambda(x_i'\beta)[1 - \Lambda(x_i'\beta)], \tag{9}$$

[4] Di fatto σ rappresenta un coefficiente di proporzionalità che può essere posto uguale ad 1 senza determinare perdità di generalità del modello. Una spiegazione intuitiva sarà fornita a pag. 11.

[5] Per questo motivo la variabile casuale normale standardizzata è tabulata.

Figura 1: Funzioni di densità

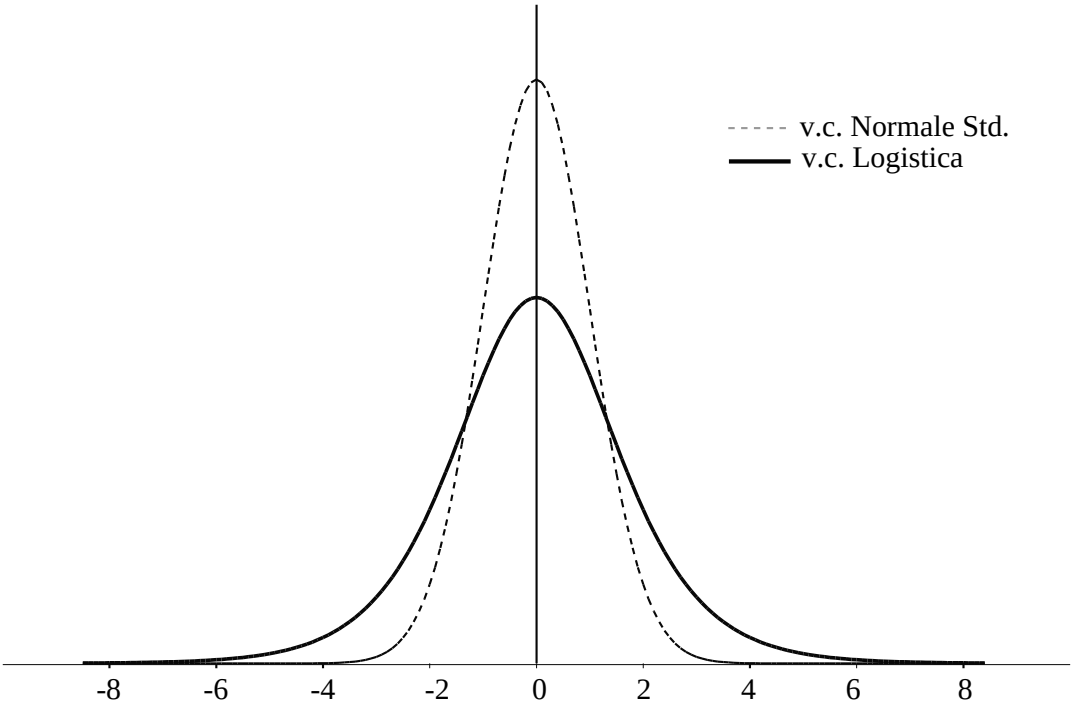
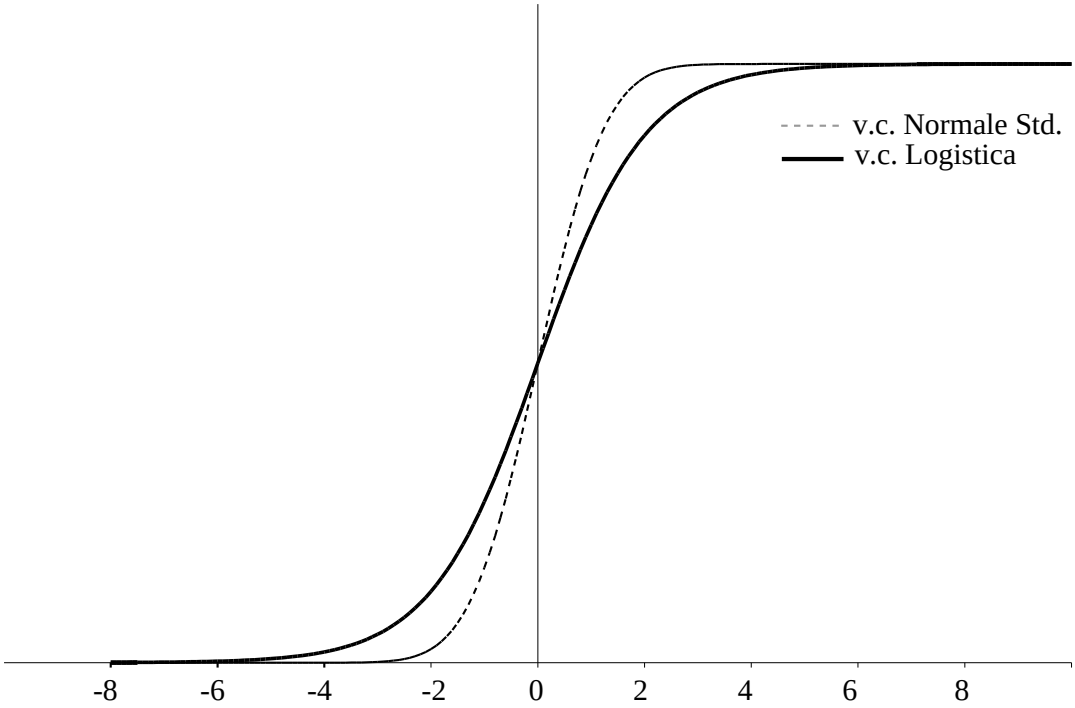


Figura 2: Funzioni di ripartizione



6. il Log-Odds Ratio è lineare nelle variabili e nei parametri, infatti

$$\begin{aligned}\log \Lambda(x_i'\beta) &= \log\left[\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right] \\ &= x_i'\beta - \log(1+e^{x_i'\beta})\end{aligned}$$

e

$$\begin{aligned}\log[1-\Lambda(x_i'\beta)] &= \log\left[\frac{1}{1+e^{-x_i'\beta}}\right] \\ &= -\log(1+e^{-x_i'\beta}).\end{aligned}$$

Poiché il log-odds ratio è dato dal logaritmo del rapporto tra la probabilità che $y_i = 1$ e la probabilità che $y_i = 0$, si ha

$$\begin{aligned}\log\frac{p_i}{1-p_i} &= \log\frac{\Lambda(x_i'\beta)}{1-\Lambda(x_i'\beta)} \\ &= \log\Lambda(x_i'\beta) - \log[1-\Lambda(x_i'\beta)] \\ &= x_i'\beta \end{aligned} \tag{10}$$

La scelta tra un Logit o un Probit dipende esclusivamente da ragioni pratiche in quanto di fatto non esistono motivi teorici per a scelta. Amemiya (1981) discute queste ragioni, ma la questione ancora oggi resta aperta. In diverse applicazioni, inoltre, non sembrano esserci sostanziali differenze tra un approccio o l'altro.

1.2.1 Interpretazione dei coefficienti

Nel modello lineare di probabilità i parametri contenuti in β rappresentano gli effetti marginali che le variabili esplicative x_i esercitano sulla variabile dipendente, infatti

$$\frac{\partial E(y_i \mid x_i)}{\partial x_i} = \beta.$$

Tale effetto marginale è indipendente da $i = 1, 2, \ldots, N$. A prescindere dal modello utilizzato nella stima, nel Logit e nel Probit tale condizione non è vera perché il vettore dei parametri è inserito all'interno di una funzione non lineare, quindi risulta

$$\frac{\partial E(y_i \mid x_i)}{\partial x_i} = \frac{\partial F(x_i'\beta)}{\partial x_i} = f(x_i'\beta)\beta$$

che invece dipende necessariamente dalle variabili esplicative x_i. In particolare risulta

$$f(x_i'\beta)\beta = \begin{cases} \phi(x_i'\beta)\beta & \text{Probit} \\ \\ \lambda(x_i'\beta)\beta = \Lambda(x_i'\beta)[1-\Lambda(x_i'\beta)]\beta & \text{Logit} \end{cases} \tag{11}$$

Una quantità spesso utilizzata nel confronto tra le stime Logit e Probit è la pendenza (slope at mean) $m(\bar{x})$ che ricorre al vettore delle medie aritmetiche di x_i per valutare gli effetti marginali[6]. La pendenza è perciò definita come segue:

$$m(\bar{x}) = \frac{\partial F(\bar{x}'\beta)}{\partial x} = f(\bar{x}'\beta)\beta \tag{12}$$

dove $\bar{x}$ vettore k-dimensionale contenente le medie aritmetiche delle variabili esplicative x_i.

In grandi campioni la pendenza del Probit e del Logit tendono a coincidere; se ciò accade anche in campioni finiti, in pratica non c'è sostanziale differenza tra le due stime.

[6] Ad esempio, il pacchetto econometrico `Gretl` per default utilizza le pendenze al posto degli standard error dei parametri nelle stime Logit e Probit.

E' inoltre possibile effettuare un confronto tra i diversi modelli basandosi sulla stima dei cefficienti β. Tenendo presente che la varianza della variabile casuale normale standardizzata è 1, mentre quella della variabile casuale logistica è $\pi^2/3$, standardizzando la stima Logit si ottiene

$$\hat{\beta}_P \approx \frac{\hat{\beta}_L}{\sqrt{\pi^2/3}} = \frac{\sqrt{3}}{\pi}\hat{\beta}_L = 0.551\hat{\beta}_L$$

Amemiya (1981) dimostra che si ottengono risultati più accurati se si utilizza un coefficiente pari a 0.625, in quanto approssima meglio la conversione quando $x_i'\beta = 0$, cioè quando $F(x_i'\beta) = 0.5$. Data la leptocurtosi della variabile casuale logistica (si veda la Figura ???), per campioni sbilanciati[7] tale coefficiente tende a diminuire lungo le code della distribuzione.

Considerando anche il modello lineare di probabilità risulta

$$\hat{\beta}_{PL} \approx 0.25\hat{\beta}_L$$

mentre per la costante si ha

$$\hat{\alpha}_{PL} \approx 0.5 + 0.25\hat{\alpha}_L.$$

In questo modo sono possibili tutte le coversioni tra $\hat{\beta}_P$, $\hat{\beta}_L$ e $\hat{\beta}_{PL}$.

1.2.2 Stima

Poiché la forma funzioale della media condizionale di y_i date le variabili esplicative x_i non è lineare cade un'importante ippotesi del modello lineare classico, quindi non è possibile stimare un Probit o un Logit col metodo OLS. Generalmente questo tipo di modelli viene stimato ricorrendo al metodo MLE.

Se si considerano N eventi i.i.d. dicotòmici con $Pr(y_i = 1) = p_i$ e $Pr(y_i = 0) = 1 - p_i$, in pratica si ha un campionamento da una popolazione bernoulliana nel quale la funzione di probabilità per ciascuna osservazione è data da $p(y_i) = p_i^{y_i}(1-p_i)^{1-y_i}$. La funzione di probabilità congiunta del campione o funzione di verosimiglianza è perciò[8]

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{N} p_i^{y_i}(1-p_i)^{1-y_i} \\ &= \prod_{i=1}^{N} F(x_i'\beta)^{y_i}[1 - F(x_i'\beta)]^{1-y_i}. \end{aligned} \tag{13}$$

La log-verosimiglianza è

$$\ell(\beta) = \sum_{i=1}^{N} y_i \log F(x_i'\beta) + (1 - y_i)\log[1 - F(x_i'\beta)]. \tag{14}$$

[7] Un campione si definisce sbilanciato se la numerosità delle $y_i = 1$ e quella delle $y_i = 0$ differiscono sostanzialmente.

[8] Si tenga presente che la distribuzione bernoulliana è un caso particolare della distribuzione binomiale con numero delle prove pari a 1. In questo caso il coefficiente binomiale associato a ciascuna osservazione vale

$$\begin{pmatrix} 1 \\ y_i \end{pmatrix} = 1$$

per ogni y_i.

Lo score è

$$\begin{aligned} s(\beta) &= \sum_{i=1}^{N} y_i \frac{\partial \log F(x_i'\beta)}{\partial \beta} + (1-y_i)\frac{\partial \log[1-F(x_i'\beta)]}{\partial \beta} \\ &= \sum_{i=1}^{N} y_i \frac{\partial F(x_i'\beta)/\partial \beta}{F(x_i'\beta)} + (1-y_i)\frac{\partial [-F(x_i'\beta)]/\partial \beta}{1-F(x_i'\beta)} \\ &= \sum_{i=1}^{N} \frac{\partial F(x_i'\beta)/\partial \beta [y_i - F(x_i'\beta)]}{F(x_i'\beta)[1-F(x_i'\beta)]} \\ &= \sum_{i=1}^{N} \frac{y_i - F(x_i'\beta)}{F(x_i'\beta)[1-F(x_i'\beta)]} f(x_i'\beta)x_i. \end{aligned} \tag{15}$$

In particolare si ha

$$s(\beta) = \begin{cases} \displaystyle\sum_{i=1}^{N} \frac{y_i - \Phi(x_i'\beta)}{\Phi(x_i'\beta)[1-\Phi(x_i'\beta)]}\phi(x_i'\beta)x_i & \text{Probit} \\ \displaystyle\sum_{i=1}^{N} \frac{y_i - \Lambda(x_i'\beta)}{\Lambda(x_i'\beta)[1-\Lambda(x_i'\beta)]}\lambda(x_i'\beta)x_i = \sum_{i=1}^{N}[y_i - \Lambda(x_i'\beta)]x_i & \text{Logit} \end{cases} \tag{16}$$

In pratica, per entrambi i modelli, la funzione score può essere scritta nella forma

$$s(\beta) = \sum_{i=1}^{N} g_i x_i, \tag{17}$$

dove $g_i = y_i - \Lambda(x_i'\beta)$ per il Logit, mentre g_i assume una forma più complessa per il Probit[9].

Per ottenere lo stimatore MLE occorre che la condizione del primo ordine sullo score $s(\beta) = 0$ sia soddisfatta; poiché β è inserito all'interno della funzione non lineare $F(x_i'\beta)$, la soluzione del problema non può essere ottenuta in forma chiusa, quindi occorrono algoritmi di calcolo numerico per ottimizzare la funzione log-verosimiglianza. Il metodo di Newton rappresenta l'algoritmo più conveniente per raggiungere la convergenza. Una volta ottenuta la stima per il vettore β è possibile calcolare le quantità $e^{x_i\hat{\beta}}$, $\phi(x_i\hat{\beta})$, $\Phi(x_i\hat{\beta})$, $\lambda(x_i\hat{\beta})$ e $\Lambda(x_i\hat{\beta})$.

La garanzia che lo stimatore MLE $\hat{\beta}$ risolva un problema di ricerca di un massimo deriva dal fatto che la

[9] In particolare, sfruttando la proprietà $\Phi(x_i'\beta) = 1 - \Phi(-x_i'\beta)$, per il Probit la funzione score può essere scritta nella forma

$$\begin{aligned} s(\beta) &= \left[\sum_{y_i=1} \frac{\phi(x_i'\beta)}{\Phi(x_i'\beta)} + \sum_{y_i=0} \frac{-\phi(x_i'\beta)}{\Phi(-x_i'\beta)}\right] x_i \\ &= \sum_{i=1}^{N} \frac{q_i \phi(q_i x_i'\beta)}{\Phi(q_i x_i'\beta)} x_i \\ &= \sum_{i=1}^{N} g_i x_i, \end{aligned}$$

dove $q_i = 2y_i - 1$ in modo da trasformare i valori 0 e 1 nei valori ± 1.

matrice di informazione è almeno semidefinita positiva, infatti risulta

$$I(\beta) = \begin{cases} \displaystyle\sum_{i=1}^{N} \frac{\phi(x_i'\beta)^2}{\Phi(x_i'\beta)[1-\Phi(x_i'\beta)]} x_i x_i' & \text{Probit} \\ \displaystyle\sum_{i=1}^{N} \lambda(x_i'\beta) x_i x_i' & \text{Logit} \end{cases} \tag{18}$$

Dalla (18) risulta che in entrambi i modelli la matrice di informazione è una forma quadratica di dimensione $k \times k$, quindi risulta almeno definita positiva (Hessiana $H(\beta) = -I(\beta)$ almeno definita negativa).

La matrice asintotica delle covarianze di β è data da

$$V_\beta = [I(\beta)]^{-1} \tag{19}$$

che rappresenta anche l'estremo di Cramer-Rao. Lungo la diagonale principale di tale matrice sono disposti gli standard error $se(\beta)$ utili per la verifica di ipotesi sui parametri del vettore β stimato.

Poiché nei modelli Probit e Logit la probabilità prevista è pari a $\hat{p}_i = F(x_i'\hat{\beta})$, mentre gli effetti marginali previsti ammontano a $f(x_i'\hat{\beta})\hat{\beta}$, lo standard error degli stessi effetti marginali viene calcolato via Delta Method attraverso gli elementi diagonali della matrice

$$\Sigma_{\hat{\beta}} = \left\{ \left[\frac{\partial F(x_i'\hat{\beta})}{\partial \beta} \right]' V_\beta \left[\frac{\partial F(x_i'\hat{\beta})}{\partial \beta} \right] \right\}^{1/2}.$$

1.2.3 Verifica di ipotesi

Nei modelli Probit e Logit tutti i test statistici standard per la verifica di ipotesi sui parametri stimati $\hat{\beta}$ si possono applicare (test t, W, LR e LM).

Un test corrispondente al test F nel modello lineare classico per l'ipotesi nulla $H_0 : \beta_j = 0$, con $j = 2, 3, \ldots, k$, è dato dal rapporto di verosimiglianza

$$\text{LR} = 2[\ell(\hat{\beta}) - \ell(\tilde{\beta})], \tag{20}$$

dove $\ell(\hat{\beta})$ è la funzione log-verosimiglianza valutata in corrispondenza della stima MLE, $\ell(\tilde{\beta})$ è la funzione log-verosimiglianza valutata sotto H_0 nella quale tutti i parametri, esclusa la costante del modello[10], siano posti uguali a zero. Poiché l'ipotesi nulla prevede $(k-1)$ vincoli, la distribuzione del test è LR$\sim \chi^2_{k-1}$.

Sotto H_0 la funzione $\ell(\tilde{\beta})$ non deve essere stimata in quanto può essere agevolmente calcolata attraverso i seguenti passaggi; partendo dalla (14), poiché il vettore dei parametri è vincolato ad essere nullo in tutte le sue componenti ad eccezione di β_1, si ottiene

$$\ell(\tilde{\beta}) = \sum_{i=1}^{N} y_i \log F(\beta_1) + (1-y_i) \log[1 - F(\beta_1)].$$

Dato che $p_i = F(x_i'\beta)$, se tutti i valori di β (ad eccezione della costante) sono nulli risulta che la probabilità stimata di $y_i = 1$ non dipende da x_i quindi $p = F(\beta_1)$. Sostituendo in $\ell(\tilde{\beta})$ si ha

$$\ell(\tilde{\beta}) = \sum_{i=1}^{N} y_i \log p + (1-y_i) \log(1-p).$$

[10] In questo caso è implicita l'ipotesi che la prima colonna della matrice X delle le variabili esplicative sia data dal vettore ι_N contenente N elementi pari ad 1.

Tenendo conto che $\sum_{i=1}^{N} y_i = N_1$, $\sum_{i=1}^{N}(1-y_i) = N_0$ e $N = N_1 + N_0$ risulta

$$\begin{aligned}\ell(\tilde{\beta}) &= N_1 \log p + N_0 \log(1-p) \\ &= N\left[\frac{N_1}{N}\log p + \frac{N_0}{N}\log(1-p)\right] \\ &= N[p \cdot \log p + (1-p)\cdot \log(1-p)]. \end{aligned} \tag{21}$$

Un'altra procedura di test molto utile all'interno dei modelli Probit e Logit è il test LM ottenuto partendo dalla condizione (17) sullo score che può essere riscritta nella forma matriciale

$$s(\beta) = X'G\iota_N$$

dove la matrice G è diagonale di dimensione $N \times N$ con i valori g_i disposti lungo la diagonale principale.

La matrice di informazione può essere stimata in modo consistente attraverso il metodo BHHH

$$I(\beta) = \frac{\partial \ell(\beta)}{\partial \beta}' \frac{\partial \ell(\beta)}{\partial \beta} = X'G'GX,$$

dove ovviamente $G' = G$. Il test LM può essere così ottenuto attraverso la statistica

$$\begin{aligned}\text{LM} &= s(\tilde{\beta})'[I(\tilde{\beta})]^{-1}s(\tilde{\beta}) \\ \text{LM} &= \iota_N'G'X(X'G'GX)^{-1}X'G\iota_N \\ \text{LM} &= N\left[\frac{1}{N}\iota_N'G'X(X'G'GX)^{-1}X'G\iota_N\right].\end{aligned}$$

Dato che $N = \iota_N'\iota_N$, il test LM è pari a

$$\text{LM} = NR_\iota^2, \tag{22}$$

cioè N volte l'indice di determinazione non centrato della regressione *Outer Product Gradients* (OPG regression) di ι_N su GX. Questa procedura permette di testare diverse ipotesi sui parametri, non solo quella di azzeramento di tutti i parametri, esclusa la costante; poiché richiede la stima del solo modello vincolato, questo test è molto utile per condurre test di specificazione (soprattutto di variabili omesse) o test di eteroschedasticità. Naturalmente, se H_0 impone q vincoli, il test ha distribuzione LM$\sim \chi_q^2$.

1.2.4 Bontà di adattamento

Diverse misure di bontà di adattamento sono state proposte per i modelli di scelta binaria; un'ipotesi molto sfruttata in questo contesto è quella introdotta a pag. 8 in base alla quale tutti i coefficienti contenuti nel vettore β sono azzerati ad eccezione della costante. Sotto H_0 la funzione log-verosimiglianza è data dall'equazione (21).

Tutte le misure di bontà di adattamento nei modelli Probit e Logit prevedono l'utilizzo della funzione log-verosimiglianza libera $\ell(\hat{\beta})$ e della log-verosimiglianza vincolata data da $\ell(\tilde{\beta})$.

Pseudo-R^2

$$R^2 = 1 - \frac{1}{1 + 2N[\ell(\hat{\beta}) - \ell(\tilde{\beta})]} \tag{23}$$

I casi estremi di quest'indice indicano che, quando nel modello libero i coefficienti stimati sono nulli, log-verosimiglianza libera e vincolata coincidono, quindi lo pseudo-R^2 è nullo. Viceversa, quando c'è perfetto adattamento, $\hat{p}_i = y_i$ quindi $F(x_i'\hat{\beta}) = y_i$ per ogni $i \in [1, N]$. Sostituendo questo risultato nella (14) si ottiene immediatamente che $\ell(\hat{\beta}) = 0$. Dato che per definizione $\ell(\hat{\beta}) \geq \ell(\tilde{\beta})$, ciò significa che la log-verosimiglianza è negativa, quindi lo pseudo-R^2 diventa

$$R^2 = 1 - \frac{1}{1 - 2N\ell(\tilde{\beta})}$$

ed è sempre minore di 1. In definitiva $R^2 \in [0, 1)$.

R^2 di McFadden

$$R^2 = 1 - \frac{\ell(\hat{\beta})}{\ell(\tilde{\beta})} \tag{24}$$

In questo caso $R^2 \in [0,1]$ (l'estremo superiore è incluso) poiché se le due log-verosimiglianze coincidono l'indice è nullo, mentre nel caso di perfetto adattamento è nulla solo la log-verosimiglianza libera, quindi l'indice è pari a 1.

R^2 di previsione

Questo indice mette a confronto le previsione corrette con quelle non corrette. Una previsione è definita corretta se risulta

$$\begin{array}{lllll} \text{-se } y_i = 1 & \Rightarrow & F(x_i'\hat{\beta}) > 0.5 & \Rightarrow & x_i'\hat{\beta} > 0 \\ \text{-se } y_i = 0 & \Rightarrow & F(x_i'\hat{\beta}) \leq 0.5 & \Rightarrow & x_i'\hat{\beta} \leq 0. \end{array}$$

Nel caso del modello vincolato, la migliore previsione è fornita da

$$\text{-se } y_i = 1 \quad \Rightarrow \quad \frac{N_1}{N} > 0.5$$

$$\text{-se } y_i = 0 \quad \Rightarrow \quad \frac{N_0}{N} \leq 0.5.$$

L'errore di previsione è perciò definito nel modello libero come

$$\widehat{EP} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2,$$

mentre in quello vincolato è

$$\widetilde{EP} = \begin{cases} 1 - \hat{p} & \text{se } \hat{p} > 0.5 \\ \hat{p} & \text{se } \hat{p} \leq 0.5. \end{cases}$$

In questo caso $\widetilde{EP} \leq 0.5$ per definizione. L'indice di bontà di adattamento è analogo all'R^2 di McFadden, ma utilizza l'errore di previsione, quindi

$$R^2 = 1 - \frac{\widehat{EP}}{\widetilde{EP}} \tag{25}$$

Nel caso di perfetto adattamento $\widehat{EP}$ è ovviamente nullo quindi l'indice assume valore massimo pari ad 1. Il problema sorge quando si cerca un valore minimo per R^2, in quanto può accadere che $\widehat{EP} > \widetilde{EP}$, interpretabile come una migliore previsione da parte del modello vincolato rispetto a quella ottenibile dal modello libero: questa affermazione è naturalmente insostenibile dato che nel modello vincolato molte variabili esplicative sono escluse. In questo caso l'indice assume valori negativi quindi, nella pratica, è perciò auspicabile ottenere almeno un valore positivo per tale indice.

1.2.5 Giustificazione teorica

Dal punto di vista teorico i modelli Probit e Logit possono essere giustificati in base a due principi differenti.

A) Massimizzazione di una funzione di utilità

Ipotesi:

1. U_i^0 e U_i^1 sono le funzioni di utilità per l'i-esimo individuo per l'alternativa binaria 1 o 0,
2. ciacuna funzione di utilità è composta da una parte sistematica ed una casuale, cioè $U_i^j = S_i^j + \varepsilon_i^j$,

3. l'individuo sceglie l'alternativa $j = 1$ se e solo se $U_i^1 > U_i^0$.

Date queste ipotesi si ottiene:

$$\begin{aligned} Pr(y_i = 1) &= Pr(U_i^1 > U_i^0) \\ &= Pr(S_i^1 + \varepsilon_i^1 > S_i^0 + \varepsilon_i^0) \\ &= Pr(\varepsilon_i^0 - \varepsilon_i^1 < S_i^1 - S_i^0) \\ &= F(S_i^1 - S_i^0) \end{aligned}$$

Si ottengono i modelli Probit e Logit semplicemente assumendo che la differenza $(\varepsilon_i^0 - \varepsilon_i^1)$ abbia distribuzione normale oppure logistica. Se tale differenza ha varianza pari a σ^2, l'espressione di cui sopra diventa

$$Pr(y_i = 1) = F\left(\frac{S_i^1 - S_i^0}{\sigma}\right).$$

Ovviamente, specificando la componente sistematica come $S_i^j = (x_i^j)'\beta^*$, si ottiene

$$\begin{aligned} Pr(y_i = 1) &= Pr\left(\frac{(x_i^1 - x_i^0)'\beta^*}{\sigma}\right) \\ &= Pr\left(\frac{x_i'\beta^*}{\sigma}\right) \\ &= Pr(x_i'\beta), \end{aligned}$$

formulazione nella quale risulta evidente il fatto che i parametri contenuti in β non possono essere stimati separatamente da σ. Occorre perciò un vincolo di normalizzazione da imporre al parametro della varianza dell'errore in modo da avere esatta identificazione.

B) Variabile latente

Ipotesi:

1. esiste una variabile latente y_i^* non osservabile,
2. la variablie y_i è osservabile e può assumere solo i valori 0 e 1,
3. esiste una soglia c tale per cui si osservano

$$\begin{cases} y_i = 1 & \text{se } y_i^* > c \\ y_i = 0 & \text{se } y_i^* \leq c \end{cases}$$

4. $y_i^* = (x_i^1 - x_i^0)'\beta^* + (\varepsilon_i^1 - \varepsilon_0) = x_i'\beta + \varepsilon_i$

Date queste ipotesi si ha

$$\begin{aligned} Pr(y_i = 1) &= Pr(y_i^* > c) \\ &= Pr(y_i^* - x_i'\beta^* > c - x_i'\beta^*) \\ &= Pr(\varepsilon_i > c - x_i'\beta^*) \\ &= Pr\left(\frac{\varepsilon_i}{\sigma} > \frac{c - x_i'\beta^*}{\sigma}\right) \\ &= 1 - Pr\left(\frac{\varepsilon_i}{\sigma} < \frac{c - x_i'\beta^*}{\sigma}\right) \\ &= 1 - F\left(\frac{c - x_i'\beta^*}{\sigma}\right) \\ &= F\left(\frac{x_i'\beta^* - c}{\sigma}\right) \\ &= F(x_i'\beta - C), \end{aligned}$$

dove $\beta = \beta^*/\sigma$ e $C = c/\sigma$. In questo caso i parametri contenuti in β non possono essere stimati separatamente da σ e da c, quindi si possono fissare $\sigma = 1$ e $c = 0$ per ottenere l'equazione analoga alle espressioni (6) e (7). Tale imposizione gerantisce l'identificazione del modello e non determina alcuna perdita di generalità in quanto la stima di β mantiene β^* e σ sempre nella stessa proporzione. Infine, se il modello contiene una costante ($\alpha = \beta_1$), anche la scelta di $c = 0$ non produce alcun effetto sulle stime: ponendo $c = c_0$ infatti si può definire una nuova costante pari ad $\alpha - c_0/\sigma$ e stimare un modello nelle stesse condizioni in cui C è nullo.

Questo scenario è identico a quello basato sull'approccio della massimizzazione dell'utilità: se $c = 0$ infatti basta definire la variabile latente come $y_i^* = U_i^1 - U_i^0$. Quando $U_i^1 > U_i^0 \Rightarrow y_i^* > 0$ e si osserva $y_i = 1$ mentre, quando $U_i^1 \leq U_i^0 \Rightarrow y_i^* \leq 0$, e in questo caso si osserva $y_i = 0$.