Modelli a risposta binaria

Se la variabile da spiegare Y è di tipo dicotomico, allora la distribuzione di probabilità sottostante è la Bernoulli, che assegna probabilità π all'evento successo y=1 e probabilità 1-π all'evento insuccesso y=0.

La particolarità di tale distribuzione è che la probabilità π coincide con il valore atteso di Y. Infatti:

$$E(Y)=0\times P(Y=0)+1\times P(Y=1)=P(Y=1)=\pi$$

Modelli a Risposta Binaria - Introduzione 2

Specificare quindi un modello di regressione che spieghi, in funzione di una o più variabili esplicative, il valore atteso di un fenomeno rappresentato da una variabile binaria equivale a specificare un modello per la probabilità di successo π.

In altri termini, l'interesse principale risiede nella relazione

$$\mathrm{P}(Y=1|X_1, \ldots, X_k)= \mathrm{P}(Y=1|\boldsymbol{X})= h(\boldsymbol{X}, \boldsymbol{\beta}).$$

Modelli a Risposta Binaria - Introduzione 3

Esempio: Supponiamo di voler spiegare la partecipazione alle forze lavoro e di estrarre un campione casuale di individui. Consideriamo quindi la variabile *Y* che assume valore 1, se nel periodo di riferimento dell'analisi i rispondenti hanno lavorato o cercato lavoro, e valore 0 in caso contrario.

Nell'ipotesi che la decisione presa possa dipendere da un insieme di caratteristiche individuali, quali ad esempio l'età, lo stato coniugale, il livello di istruzione, la storia lavorativa ed eventuali altri fattori che possano influire sulla decisione, si dovrà individuare la specifica funzione $h(\boldsymbol{X}, \boldsymbol{\beta})$ che leghi a tali caratteristiche la probabilità di partecipazione alle F.L. e che ci consenta di valutare l'effetto marginale di un determinato fattore.

Modelli a Risposta Binaria - Introduzione 4

Un modello che miri a spiegare la probabilità con cui si verifichi ciascuna modalità di una qualsiasi variabile di interesse *Y*, categorica o discreta, rientra nell'ambito generale dei

modelli di probabilità.

In tale ambito, se le diverse modalità della dipendente *Y* corrispondono alle possibili alternative di scelta che un individuo ha di fronte in un determinato contesto, i modelli di probabilità consentono di spiegare la probabilità di effettuare una determinata scelta in funzione (solitamente non lineare) di una o più variabili esplicative (spesso corrispondenti a determinate caratteristiche individuali).

Modelli a Risposta Binaria - Introduzione 5

Nell'ambito generale dei modelli di probabilità, si avranno ovviamente specificazioni alternative a seconda del tipo di variabile Y considerata (dicotomica, categorica, quantitativa discreta) e del tipo di funzione scelta per esplicitare la $h(\boldsymbol{X}, \boldsymbol{\beta})$.

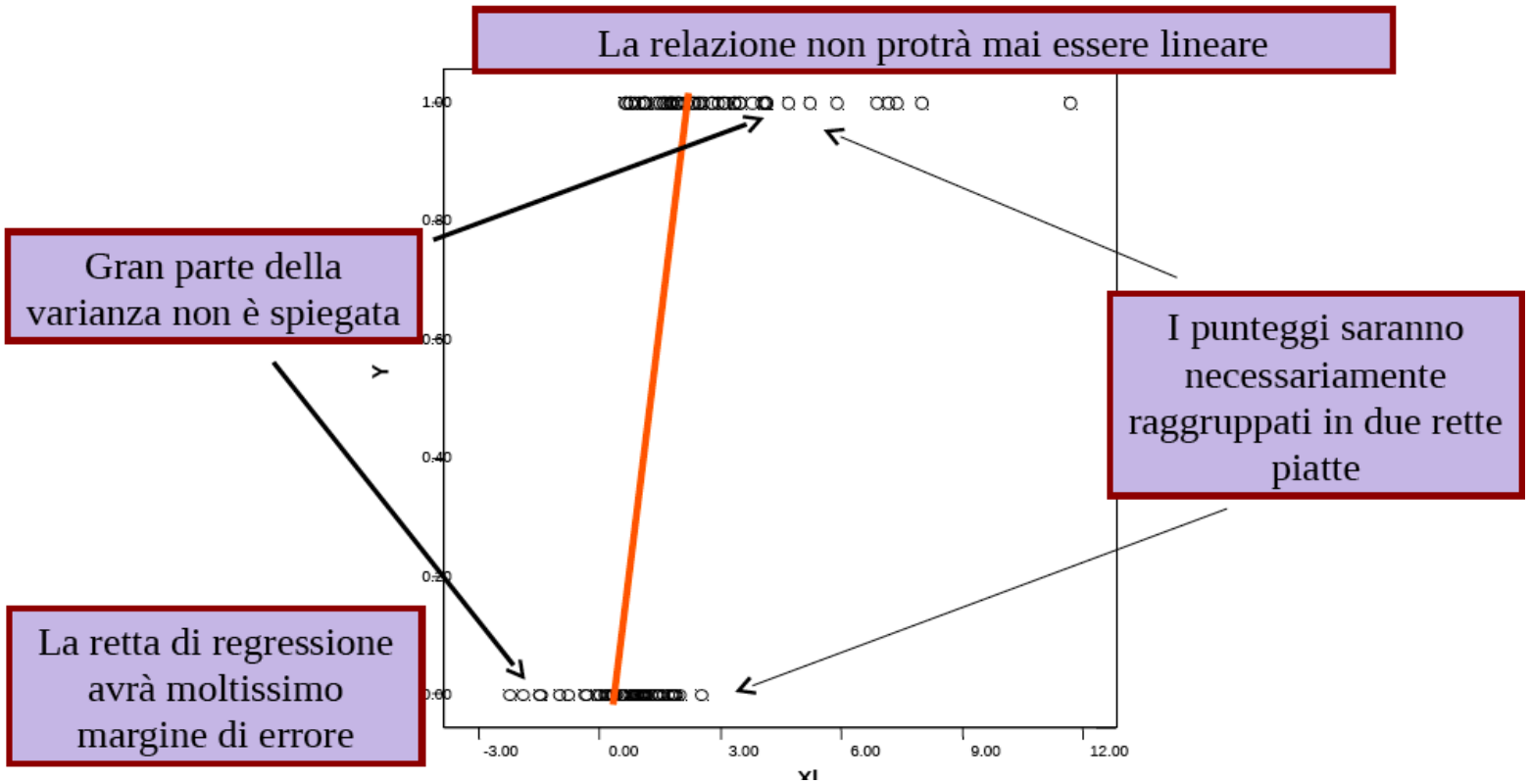
Per il momento ci limitiamo alla spiegazione di fenomeni dicotomici.

La specificazione più semplice di tutte è ovviamente la funzione lineare nei parametri.

Tuttavia, come risulterà ancora più evidente nel seguito, questa specificazione non è affatto idonea a spiegare le relazioni tra la dipendente e le esplicative e a stimare le probabilità π_i.

Modelli a Risposta Binaria - Introduzione 6

Innanzitutto, si può osservare che lo scatterplot sarà sempre costituito da due linee parallele di punti e qualsiasi retta interpolante si adatterà molto poco alla nuvola dei punti:



In altri termini, gran parte della variabilità dei dati non potrà essere spiegata.

Per questo motivo, il modello lineare di probabilità rappresenta solo un riferimento teorico, non utilizzato nelle applicazioni.

Linear Probability Model

Il modello lineare è specificato mediante la seguente equazione:

$$E(Y|X_1, \ldots, X_k) = P(Y=1|X_1, \ldots, X_k) = \boldsymbol{X\beta},$$

dove il generico elemento β_j del vettore $\boldsymbol{\beta}$ rappresenta come di consueto l'effetto marginale della corrispondente variabile esplicativa, misura cioè la variazione nella probabilità di successo dovuta ad una variazione in X_j tenendo fermi tutti gli altri fattori esplicativi.

Esso può essere stimato mediante gli OLS, tuttavia le condizioni che rendono ottimi gli stimatori OLS in questo caso non sono più soddisfatte. In particolare, la componente ε di disturbo risulta eteroschedastica.

Inoltre, l'assunzione di normalità degli errori, necessaria per le procedure inferenziali, non è più valida visto che la Y è dicotomica.

Linear Probability Model 3

Riguardo al problema della non normalità, la conseguenza più grave è l'impossibilità di fare inferenza, non potendo ricorrere alle distribuzioni *t* e *F* impiegate nelle procedure inferenziali

Riguardo all'eteroschedasticità, per meglio comprendere da dove essa scaturisca, è sufficiente ricordare che per una variabile bernoulliana si ha $\mathrm{Var}(Y)=\pi(1-\pi)$.
Ciò significa che qualsiasi fattore che influenzi il valore atteso $E(Y|\boldsymbol{X})$, cioè la probabilità π, produrrà effetti anche sulla varianza delle osservazioni.

Con riferimento alla componente di disturbo ε, poiché

$$Y = \mathrm{E}(Y|\boldsymbol{X}) + \varepsilon = \mathrm{X}\boldsymbol{\beta} + \varepsilon,$$

si ha che anche $\varepsilon = Y - \mathrm{E}(Y|\boldsymbol{X})$ può assumere soltanto due valori:

- $1 - \boldsymbol{X\beta}$, con probabilità $\boldsymbol{X\beta}$ (quando y=1),
- $-\boldsymbol{X\beta}$, con probabilità $1 - \mathrm{X}\boldsymbol{\beta}$ (quando y=0).

Risulta pertanto:

$$\mathrm{E}(\varepsilon|\boldsymbol{X}) = (1-\boldsymbol{X\beta})\boldsymbol{X\beta} - \boldsymbol{X\beta}(1-\boldsymbol{X\beta}) = 0,$$

$$\mathrm{Var}(\varepsilon|\boldsymbol{X}) = (1-\boldsymbol{X\beta})^2\,\boldsymbol{X\beta} + (-\boldsymbol{X\beta})^2(1-\boldsymbol{X\beta}) = \boldsymbol{X\beta}(1-\boldsymbol{X\beta}) = \mathrm{Var}(Y|\boldsymbol{X})$$

dove la quantità $\boldsymbol{X\beta}$ assume un valore diverso per ogni unità.

Tuttavia, la principale limitazione di questa specificazione è costituita dal fatto che, a meno di non effettuare degli aggiustamenti ad hoc, stime delle y_i esterne all'intervallo 0-1 condurrebbero a valori impossibili per le probabilità (cioè negativi o maggiori di 1), nonché a stime negative delle varianze, date da: $\hat{\sigma}_i^2 = \hat{y}_i(1 - \hat{y}_i)$!

Per questo motivo, grazie anche alla disponibilità di software per la stima di modelli più complicati, il modello lineare viene sempre più utilizzato soltanto come base di confronto (benchmark minimale) con altri modelli più appropriati.

Modello Generale di Probabilità

Per ovviare al problema di stime incoerenti per le probabilità di successo, è sufficiente scegliere la funzione *h*, che lega la probabilità di successo a X$\boldsymbol{\beta}$, tra le funzioni (in genere non lineari) che assumono valori unicamente nell'intervallo (0, 1):

$$P(Y=1|X_1, \ldots, X_k)=h(\beta_1 X_1 + \cdots + \beta_k X_k), \qquad h(.)\in(0,1)$$

A seconda di come venga specificata la funzione *h*, si hanno le diverse tipologie di modelli a risposta binaria.

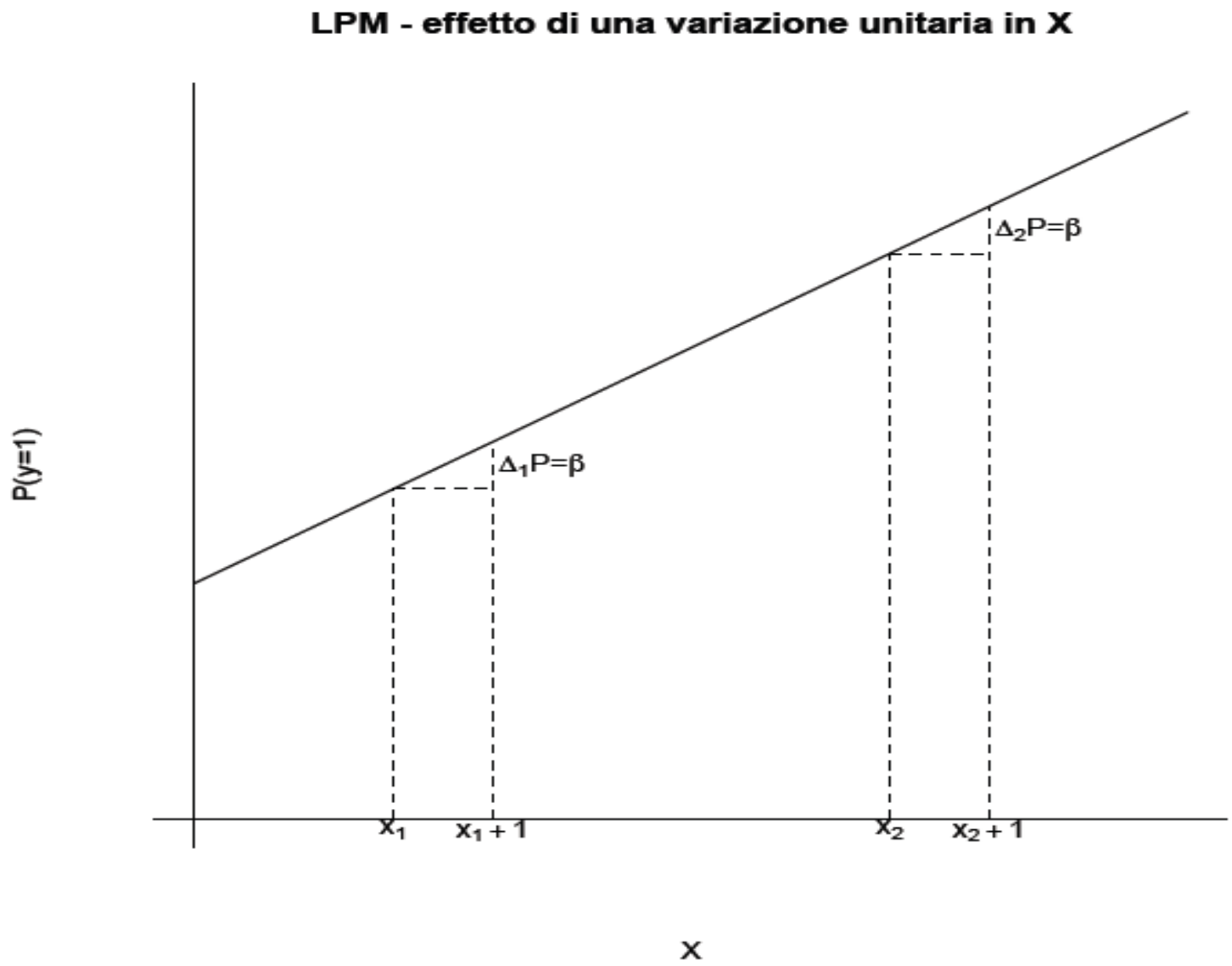
Un'idea molto semplice è specificare la *h* come una funzione di ripartizione *F*(.), per definizione compresa in (0, 1) e quindi tale da garantire

$$\mathrm{P}(Y=1) \xrightarrow[\boldsymbol{X\beta}\to\infty]{} 1, \qquad \mathrm{P}(Y=1) \xrightarrow[\boldsymbol{X\beta}\to-\infty]{} 0$$

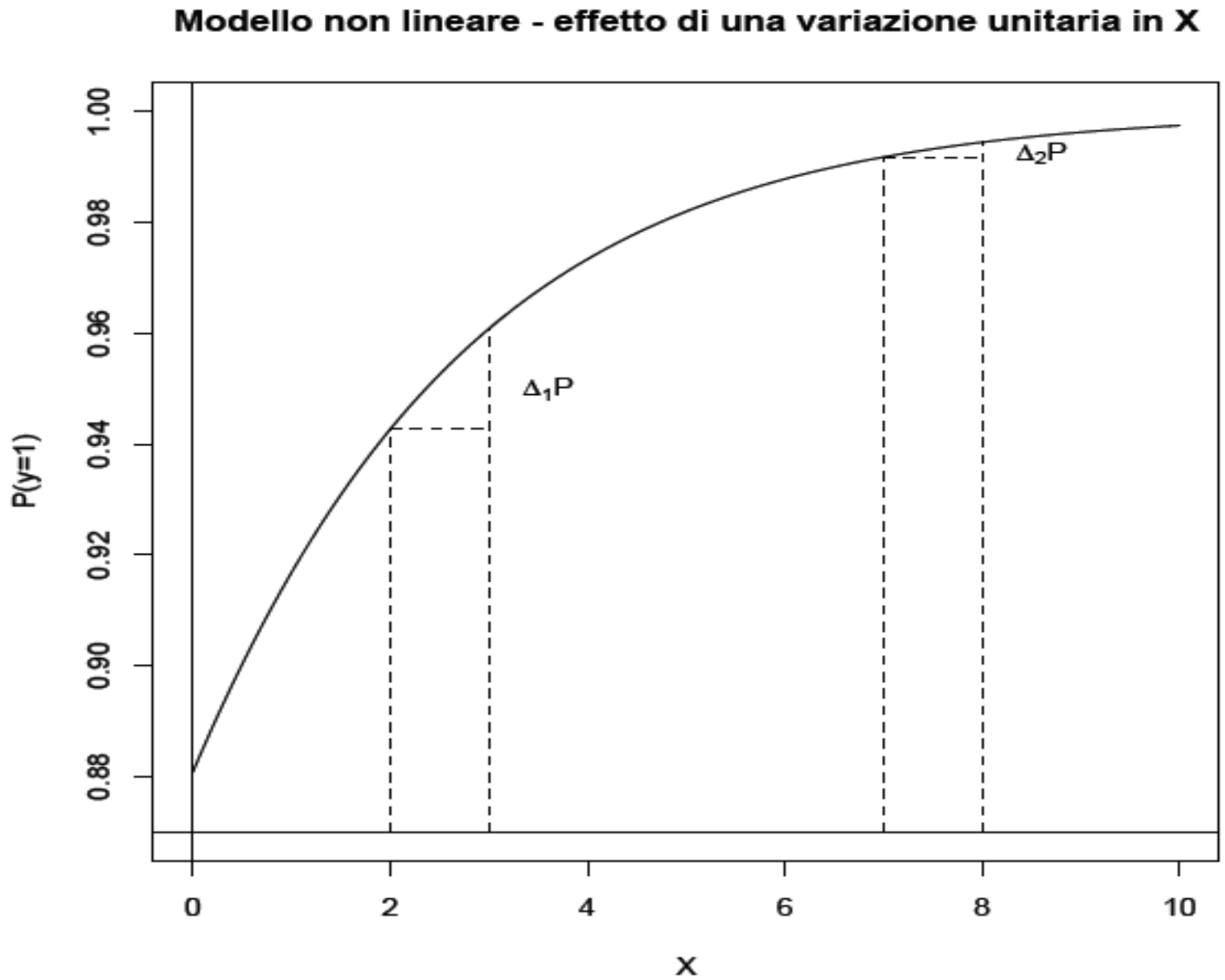
Il prezzo da pagare riguarda l'interpretabilità dei risultati, in quanto i parametri del modello in genere non coincidono con gli effetti netti delle corrispondenti variabili esplicative e, inoltre, l'effetto di una variazione unitaria in una esplicativa cambia a seconda del valore da cui parte la variazione, analogamente a quanto visto nel modello lineare classico per i regressori che entrano in forma non lineare nella funzione di regressione.

Ciò risulta evidente nei seguenti due grafici, riferiti a due modelli di probabilità (nella sola esplicativa X): il primo lineare e l'altro no.

Modello Generale di Probabilità 4



Modello Generale di Probabilità 5



Inoltre, come si è già avuto modo di osservare, nei modelli di regressione lineare se una esplicativa entra in forma non lineare, benché il suo effetto netto cambi al variare dei valori da essa assunta, esso non dipende dai valori delle altre esplicative.

Al contrario, quando il modello è non lineare anche nei parametri (come avviene se la funzione *h* è una funzione di ripartizione), allora l'effetto netto di una esplicativa dipende anche dai valori assunti dalle altre covariate.

Infatti risulta in generale:

$$\partial \mathrm{E}(Y|\boldsymbol{X})/\partial X_j = \partial F(\boldsymbol{X\beta})/\partial X_j = f(\boldsymbol{X\beta})\, \partial \boldsymbol{X\beta}/\partial X_j = f(\boldsymbol{X\beta})\beta_j$$

dove $f(.)$ è la funzione di densità corrispondente alla cumulata $F(.)$, ipotizzata continua, e l'effetto netto di X_j dipende anche dai valori di tutte le altre esplicative attraverso $f(\boldsymbol{X\beta})$, che al pari di $\boldsymbol{X\beta}$ assume un valore diverso per ogni unità.

Una determinata esplicativa, cioè, presenta in generale un diverso effetto parziale per ciascuna unità considerata.

Tuttavia, poiché la funzione di densità *f* non può essere negativa, l'effetto parziale della covariata X_j su π presenta lo stesso segno del corrispondente coefficiente di regressione β_j.

Inoltre, per distribuzioni unimodali e simmetriche intorno a zero, l'effetto parziale in corrispondenza della *i*-esima unità, per la quale si osserva il vettore di covariate $\mathbf{x}_i$, sarà tanto più elevato quanto più $|\mathbf{x}_i\boldsymbol{\beta}| \rightarrow 0$ e, viceversa, tanto più esiguo quanto più elevato è $|\mathbf{x}_i\boldsymbol{\beta}|$.

Come precedentemente osservato, a seconda del modello distributivo scelto per specificare la funzione di ripartizione $F(.)$, si avrà un diverso modello a risposta binaria.

Un modo alternativo di qualificare le diverse specificazioni consiste nel considerare una particolare funzione $g(.)$ da applicare alla probabilità π, che possa essere espressa come funzione lineare delle variabili esplicative:

$$\eta_i = g(\pi_i) = \mathbf{x}_i\boldsymbol{\beta}$$

La quantità η è detta *predittore* e la $g(.)$ è detta *funzione link*.

La grossa differenza rispetto al modello lineare classico è che adesso la struttura lineare non è più assunta per la Y (o per il suo valore atteso π), bensì per η, funzione non lineare di π.

Questo è l'approccio tipico dei GLM (acronimo di Generalized Linear Models, cioè modelli lineari generalizzati) e la specificazione di modelli diversi dipende dalla particolare scelta della funzione di link.

Lo stesso modello a probabilità lineare può essere visto come un caso particolare (banale) del GLM: basta definire la $g(.)$ come funzione identità, in modo che risulti $\eta=\pi=\boldsymbol{X\beta}$.

Per quanto precedentemente visto, ovviamente tale scelta non è appropriata poiché non garantisce il vincolo $\pi \in (0,1)$.

Le specificazioni di $g(.)$ che si sono maggiormente affermate in letteratura sono sostanzialmente tre, le prime due riferite a distribuzioni simmetriche e la terza ad una asimmetrica:

- link logistico (logit):

$$\eta = g(\pi) = \log[\pi/(1-\pi)];$$

in questo caso la probabilità di successo risulta

$$\pi = g^{-1}(\eta) = e^{\eta}/(1+e^{\eta}) = \Lambda(\eta),$$

cioè la funzione di ripartizione di una densità logistica standard, definita come

$$\lambda(\eta) = e^{\eta}/(1+e^{\eta})^2;$$

- link probit: $\eta = g(\pi) = \Phi^{-1}(\pi)$ e $\pi = \Phi(\eta)$, dove Φ è la cdf di una N(0;1);

- link log-log:

$\eta = g(\pi) = -\log[-\log(\pi)]$ e

$\pi = \exp[-\exp(-\eta)]$, espressione che coincide anch'essa con la funzione di ripartizione di una particolare densità.

Spesso viene utilizzata la forma complementare (clog-log), con:

$\eta = g(\pi) = \log[-\log(1-\pi)]$

e $\pi = 1-\exp[-\exp(\eta)]$.

La quantità $\pi/(1-\pi)$ che compare nel link logit è detta *odds* e rappresenta una scala alternativa con cui misurare il grado di fiducia nel verificarsi di un evento, molto utilizzata soprattutto in ambito biostatistico (… e nelle scommesse!).

In particolare, con riferimento all'evento $y=1$, si definisce *odds* la seguente quantità:

$$\text{Od}(1)=\text{P}(Y=1)/[1-\text{P}(Y=1)]=\pi/(1-\pi).$$

Essa indica una sorta di rischio relativo circa il verificarsi di un determinato evento, o anche una sorta di pronostico riguardo al possesso o meno di una certa caratteristica o all'effettuazione di una determinata scelta.
Ad esempio, se P(Y=1)=0.25, allora il rapporto risulta pari a 0.25/0.75=1/3, cioè è di 3 a 1 in favore dell'evento y=0.

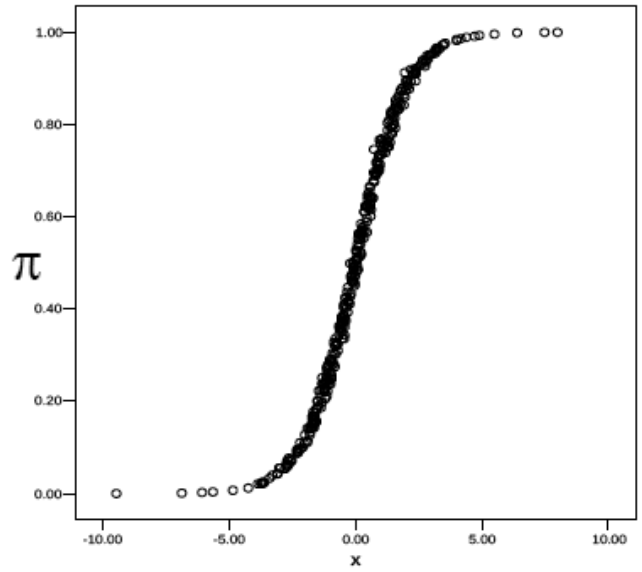
L'odds, inoltre, può assumere qualsiasi valore positivo,
cresce con la probabilità di successo e
per il modello logit vale exp($\boldsymbol{X\beta}$).

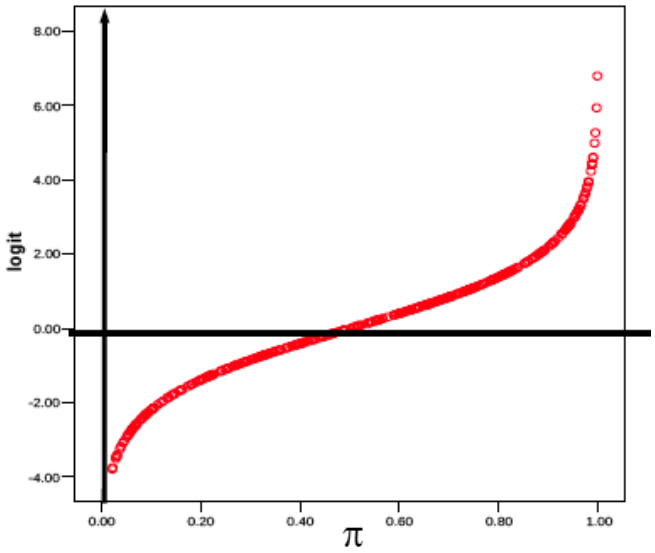
Infatti, il logit non è altro che il log(Od):
$\text{logit}(1) = \log[\text{Od}(1)] = \log[\pi/(1-\pi)] = \boldsymbol{X\beta}$.

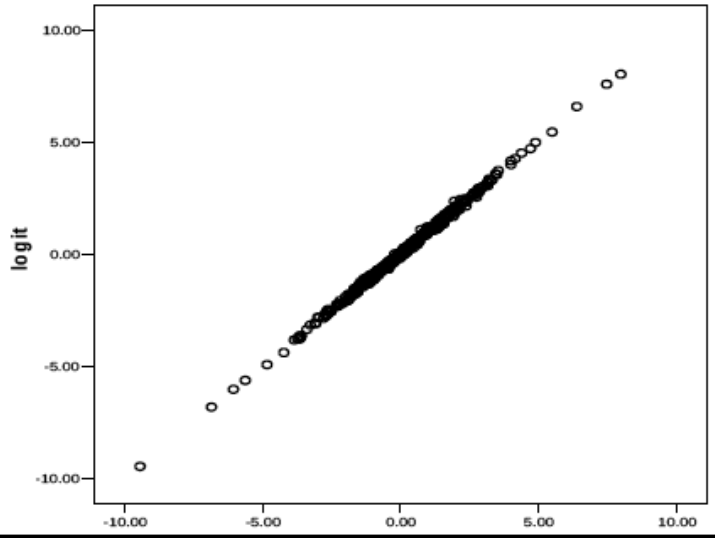
In sintesi, il modello logit non è altro che un modello di regressione lineare specificato per il log-odds dell'evento successo.

Le relazioni tra regressore X, probabilità di successo π e logit possono essere rappresentate dai seguenti grafici

Modello Generale di Probabilità 17







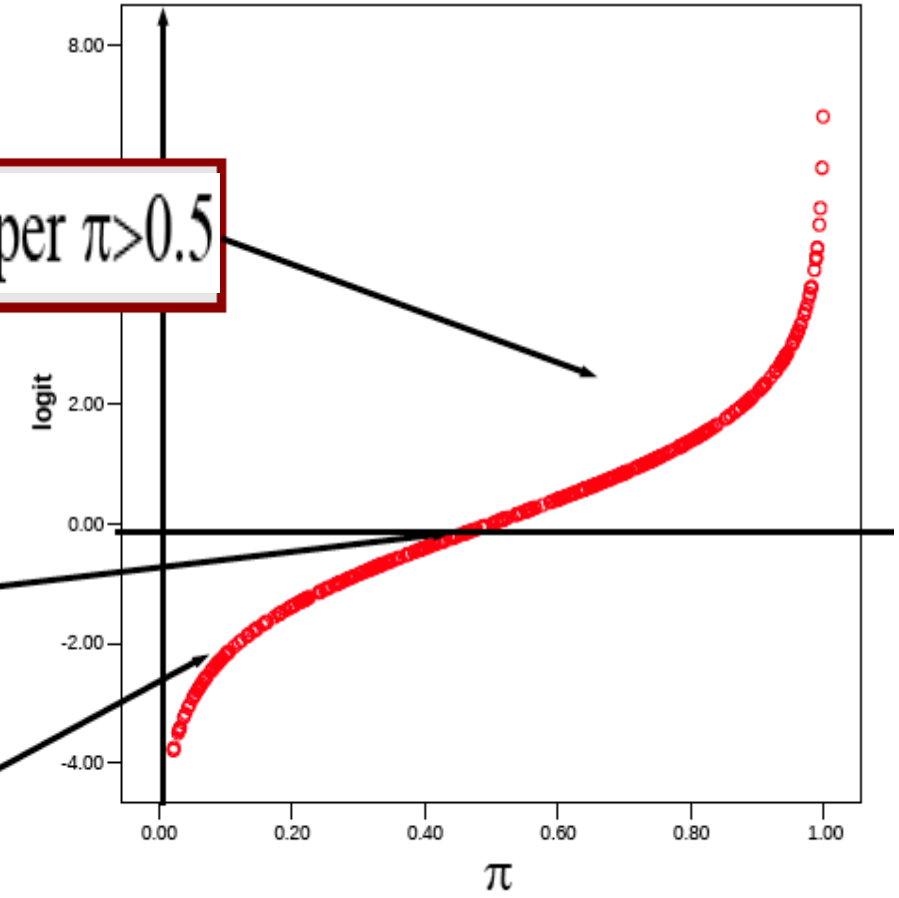
In particolare, nella relazione tra π e il logit si ha:

$$\log it = \ln\left(\frac{\pi}{1-\pi}\right)$$

Positivo per $\pi>0.5$

Centrato a 0 quando $\pi=0.5$

Negativo per $\pi<0.5$



Riepilogando:

- il logit risulta positivo se e solo se la probabilità dell'evento 'successo' è maggiore di quella dell'evento contrario;

- esso tende a $-\infty$ per $\pi \rightarrow 0$, mentre per $\pi \rightarrow 1$ il logit tende a $+\infty$;

- pertanto, logit negativi corrispondono a probabilità π inferiori a 0.5 e logit positivi a probabilità più elevate.

Le tre funzioni di link su richiamate danno nome ai modelli che su di esse si basano.

La terza specificazione, meno frequente nelle applicazioni, viene utilizzata quando una delle risposte è rara.

Data la più elevata diffusione dei modelli *logit* e *probit* e i molti punti in comune che li caratterizzano, l'attenzione verrà focalizzata su questi due.

In realtà le due specificazioni presentano poche differenze sostanziali e dal punto di vista meramente teorico non è possibile operare una scelta tra esse.

Nelle applicazioni, soprattutto in ambito biostatistico, si tende a preferire il modello logit per la più conveniente notazione matematica.

Per contro, il modello probit è spesso preferito dagli economisti soprattutto nella formalizzazione con variabile latente (richiamata più avanti), che consente alcune estensioni come quella che gestisce la presenza di eteroschedasticità (problema particolarmente sentito nelle applicazioni econometriche).

Modelli *Logit* e *Probit*

Per quanto visto in merito alle funzioni link, nel modello probit la probabilità di successo è data da:

$$\pi = \mathrm{P}(Y = 1) = \Phi(\mathbf{x}\boldsymbol{\beta}) = \int_{-\infty}^{x\beta} \phi(t)dt$$

dove $\phi(.)$ è la densità della normale standard.

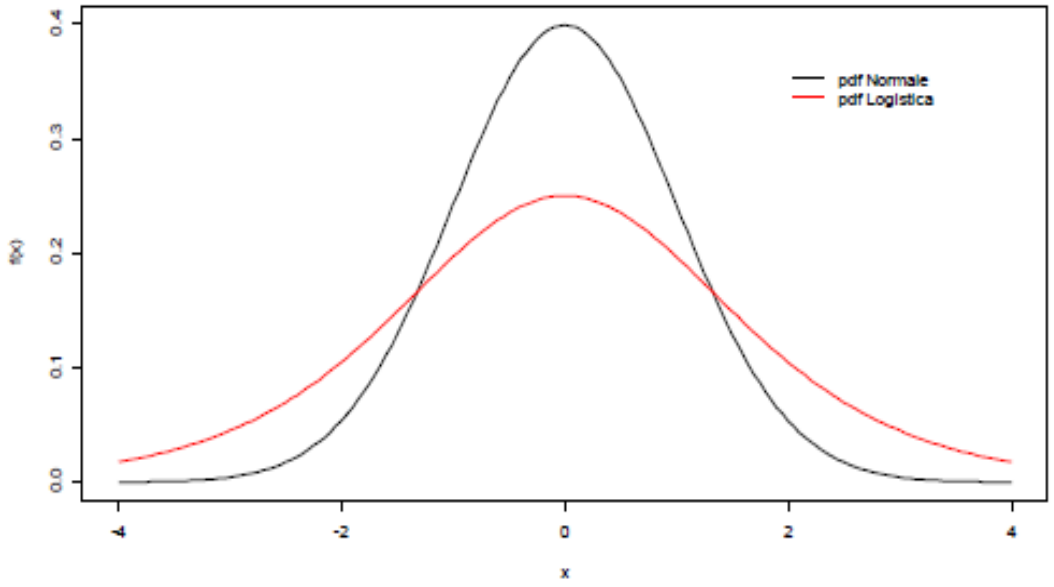
Il modello logit può essere definito mediante le due probabilità complementari:

$$\pi = \mathrm{P}(Y=1) = \frac{\exp(\mathbf{x}\boldsymbol{\beta})}{1+\exp(\mathbf{x}\boldsymbol{\beta})} = \Lambda(\mathbf{x}\boldsymbol{\beta}),$$

$$1-\pi = \mathrm{P}(Y=0) = \frac{1}{1+\exp(\mathbf{x}\boldsymbol{\beta})} = 1-\Lambda(\mathbf{x}\boldsymbol{\beta})$$

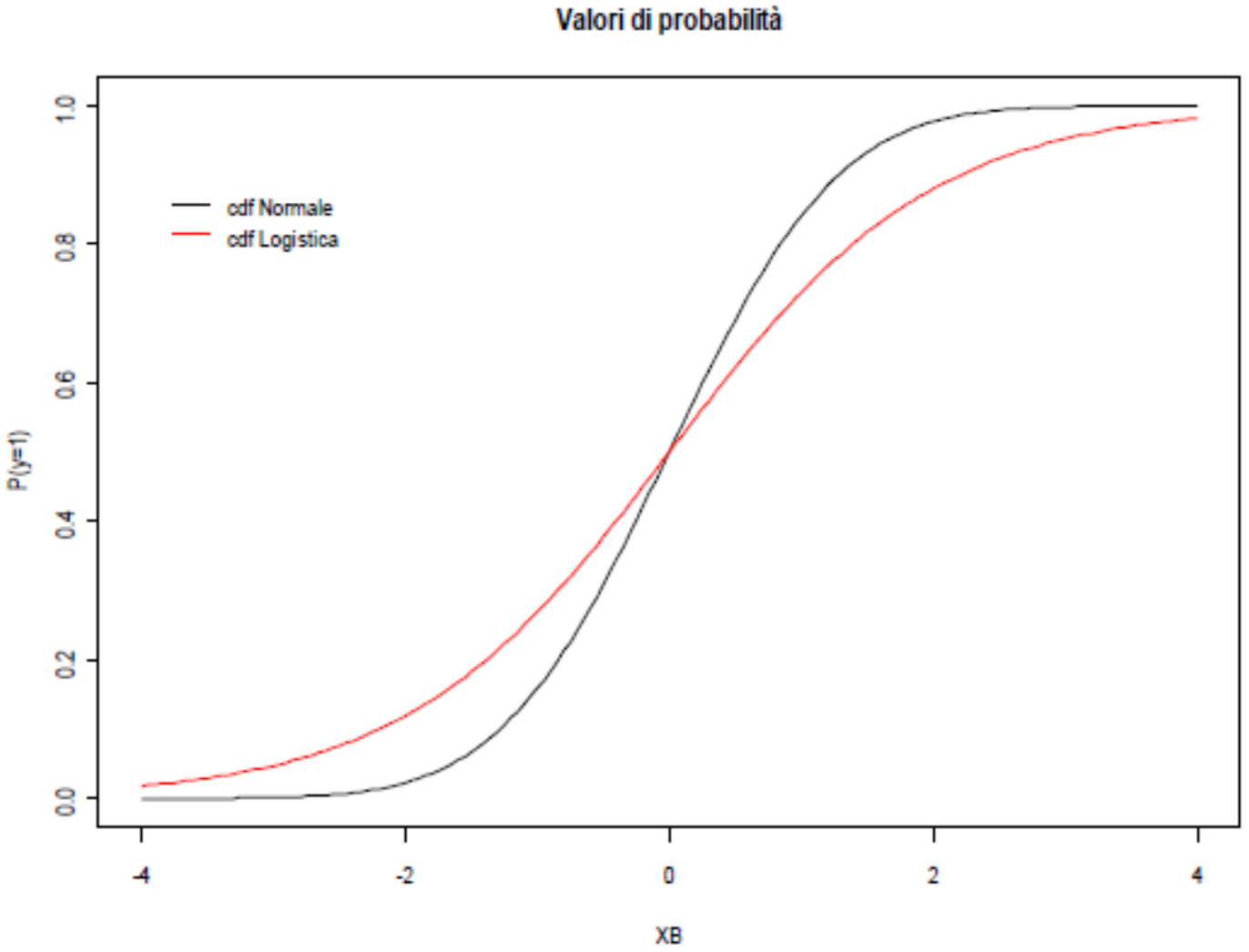
dove $\Lambda(\mathbf{x}\boldsymbol{\beta})$ è il valore che assume la funzione di ripartizione della logistica standard nel punto $\mathbf{x}\boldsymbol{\beta}$.

Le due distribuzioni logistica e normale sono in realtà abbastanza simili, fatta eccezione per le code che sono più pesanti nella logistica (potrebbe essere paragonata ad una Student-*t* con 7 gdl), a causa della maggiore variabilità.



Ciò implica che per valori di $\boldsymbol{X\beta}$ prossimi a zero, le due distribuzioni tendono a fornire probabilità molto simili, oltre che effetti più elevati dei regressori.
Si ricordi, infatti, che per entrambe le distribuzioni l'effetto del generico regressore X_j sulla probabilità di successo π, pari a $f(\mathrm{X}\boldsymbol{\beta})\beta_j$, è più elevato per valori di $\boldsymbol{X\beta}$ prossimi a zero.
Per contro, come si evince anche dal seguente grafico, la distribuzione logistica, rispetto alla normale, tende ad attribuire a y=1 una probabilità più elevata per valori estremamente piccoli di $\boldsymbol{X\beta}$ ed una probabilità più piccola per valori molto grandi di $\boldsymbol{X\beta}$.

Modelli *Logit* e *Probit* 5



Al di là della maggiore difficoltà interpretativa, la circostanza che l'effetto di una covariata sulla probabilità π vari al variare dell'intero vettore $\boldsymbol{x}$ è in realtà molto più ragionevole della costanza degli effetti implicata dal modello lineare.

Se, ad esempio, l'effetto di una covariata può plausibilmente incrementare di 0.1 la probabilità π quando questa è attestata intorno al 50%, lo stesso effetto diventerebbe irrealistico se il valore di partenza della probabilità fosse 0.95!

⇒ un effetto costante nella scala logit si traduce in effetti variabili nella scala delle probabilità, che si aggiustano automaticamente, quando la probabilità si approssima agli estremi 0 e 1.

Proprio perché gli effetti parziali non sono costanti, ma presentano in linea di principio *n* diversi valori, per interpretare i risultati è utile sintetizzare questi effetti marginali secondo uno dei seguenti tre criteri alternativi:

a) risposta media degli individui

$$n^{-1}\sum_{i=1}^{n}\partial \mathrm{E}(Y_i|\mathbf{x}_i)/\partial \mathbf{x}_i$$

b) risposta dell'individuo medio

$$\partial \mathrm{E}(Y|\mathbf{x})/\partial \mathbf{x}|_{\mathbf{x}=\bar{\mathbf{x}}}$$

c) risposta di un individuo (rappresentativo), con specifiche caratteristiche **x***,

$$\partial \mathrm{E}(Y|\mathbf{x})/\partial \mathbf{x}|_{\mathbf{x}=\mathbf{x}^*}$$

Nel modello lineare queste tre misure coincidono tutte con β, mentre nei modelli non lineari in generale differiscono tra loro.

Per grandi campioni, tuttavia, le prime due misure tendono a fornire risultati simili.

Ricordando inoltre che, nei casi
tipici in cui la densità f è unimodale e simmetrica intorno allo zero (come avviene appunto per il logit e il probit), l'effetto più grande si ha in corrispondenza di $\boldsymbol{X}\boldsymbol{\beta}=0$, possiamo considerare che:

nel modello probit risulta $f(0)=\phi(0)\cong 0.4$;

nel modello logit, invece,

$$f(0)=\lambda(0)=\exp(0)/[1+\exp(0)]^2=0.25.$$

Tali valori possono servire per confrontare le stime degli effetti, e quindi dei parametri, ottenute con i due modelli.

Infatti, indicando con $\beta_{k|\ell}$ e con con $\beta_{k|p}$ il parametro corrispondente alla covariata X_k nella specificazione logit e, rispettivamente, in quella probit, l'effetto di tale covariata nel punto $\boldsymbol{X\beta}=0$ sarà misurato in modo equivalente da $0.4\beta_{k|p}$ e $0.25\beta_{k|\ell}$.

Poiché l'effetto da misurare è lo stesso e le due diverse misure dovrebbero essere equivalenti,

cioè $0.4\beta_{k|p} \cong 0.25\beta_{k|\ell}$. dovrà necessariamente risultare $\beta_{k|p} < \beta_{k|\ell}$.

Non solo, valgono anche, almeno in via approssimativa, le seguenti relazioni:

$$1 \cong \frac{0.4\hat{\beta}_{k|p}}{0.25\hat{\beta}_{k|\ell}} = 1.6\frac{\hat{\beta}_{k|p}}{\hat{\beta}_{k|\ell}}$$

$$\Rightarrow \hat{\beta}_{k|\ell} \cong 1.6\hat{\beta}_{k|p}, \qquad \hat{\beta}_{k|p} \cong 0.625\hat{\beta}_{k|\ell}$$

In altri termini, per le stime logit possiamo attenderci valori più elevati di 1.6 volte quelli delle stime probit, cioé queste ultime dovrebbero essere moltiplicate per 1.6 per ottenere valori comparabili con le stime logit.

Per contro, moltiplicando le stime logit per 0.625, le si rendono confrontabili con le stime probit.

Per quanto riguarda invece il confronto tra gli effetti marginali di due variabili esplicative qualsiasi, se si considera l'effetto relativo questo è costante e non dipende da $\boldsymbol{X}$.

Infatti, per due variabili continue X_j e X_h, si ha:

$$\frac{\partial \pi / \partial x_j}{\partial \pi / \partial x_h} = \frac{\beta_j}{\beta_h}.$$

Va infine precisato che, nel caso di una variabile esplicativa binaria, l'effetto parziale di un cambiamento da 0 a 1 è più propriamente (e semplicemente) misurato mediante la differenza tra cdf; ad esempio, considerando la *k*-esima esplicativa, si ha:

$$F(\beta_1+\beta_2X_2+\ldots+\beta_{k-1}X_{k-1}+\beta_k)-F(\beta_1+\beta_2X_2+\ldots+\beta_{k-1}X_{k-1})$$

Ovviamente, anche in questo caso l'effetto dipende dai valori delle altre esplicative presentati da ciascuna unità; generalmente esso viene calcolato in corrispondenza dei loro valori medi o modali (cioè con riferimento all'individuo 'medio').

Esempio: se *Y* è un indicatore di partecipazione alle forze lavoro e la variabile X_k è una dummy per il sesso, allora la precedente espressione ci fornisce la differenza nella probabilità di far parte delle forze lavoro tra un uomo e una donna, tenendo costanti le altre caratteristiche individuali considerate nell'analisi, quali ad esempio l'età, lo stato coniugale, il livello di istruzione, la storia lavorativa.

Anche nel caso di X_k dicotomica, il segno di β_k consente soltanto di determinare se l'effetto è positivo o negativo, ma non la sua entità.

Con il criterio della differenza tra cdf si calcola anche l'effetto parziale di variabili esplicative discrete (ad esempio il numero di figli).

In particolare, l'effetto sulla probabilità di una variazione di X_k da x_k a x_k+1 risulta:

$$F[\beta_1+\beta_2 x_2+\ldots+\beta_{k-1}x_{k-1}+\beta_k(x_k+1)] - F(\beta_1+\beta_2 x_2+\ldots+\beta_{k-1}x_{k-1}+\beta_k x_k).$$

Tuttavia, in entrambi i casi di esplicativa X_k non continua, una valutazione ben approssimata del suo effetto si può ottenere semplicemente effettuando la derivata parziale, come se fosse continua.

Qualora, inoltre, tra le esplicative compaiano trasformazioni delle stesse (quadrati, logaritmi, prodotti, rapporti, ecc.), la valutazione dell'effetto parziale va sempre effettuata mediante la derivata parziale della probabilità di successo.

Esempio: Con riferimento al modello

$P(Y=1|X)=F[\beta_1+\beta_2X_2+\beta_3X_3+\beta_4X_2X_3+...+\beta_kX_k]$, l'effetto parziale di X_2 sulla probabilità di risposta è dato da

$\partial P(Y=1|X)/\partial x_2=f(X\boldsymbol{\beta})(\beta_2+\beta_4X_3)$.

Stime di massima verosimiglianza

In questo ambito, per stimare il vettore di parametri $\boldsymbol{\beta}$ si ricorre al metodo di massima verosimiglianza.

Per ricavare la funzione di verosimiglianza per ciascuna delle specificazioni adottate, è opportuno tener presente che, sulla base di un campione casuale di *n* unità, per ciascuna unità estratta la variabile risposta Y_i presenta la seguente distribuzione di Bernoulli:

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

dove y_i vale 0 oppure 1.

Ovviamente, a seconda della specificazione adottata, si modifica l'espressione da attribuire a π_i.

In generale, quindi, la densità di Y_i condizionata al vettore $\mathbf{x}_i$ è:

$$f(y_i|\mathbf{x}_i;\boldsymbol{\beta}) = [F(\mathbf{x}_i\boldsymbol{\beta})]^{y_i}[1 - F(\mathbf{x}_i\boldsymbol{\beta})]^{1-y_i},$$

y_i=0,1.

Pertanto la verosimiglianza per un campione di n osservazioni risulta:

$$L = \prod_{i=1}^{n}[F(\mathbf{x}_i\boldsymbol{\beta})]^{y_i}[1 - F(\mathbf{x}_i\boldsymbol{\beta})]^{1-y_i}$$

Fatta eccezione per il modello di probabilità lineare, in genere non esiste una soluzione esplicita (analitica) per lo stimatore ML, ma si ricorre a metodi numerici, di tipo iterativo, che in genere sono facilmente disponibili.

In ogni caso, dai risultati generali degli stimatori di massima verosimiglianza, segue che entrambi gli stimatori logit e probit di β sono consistenti e asintoticamente normali.

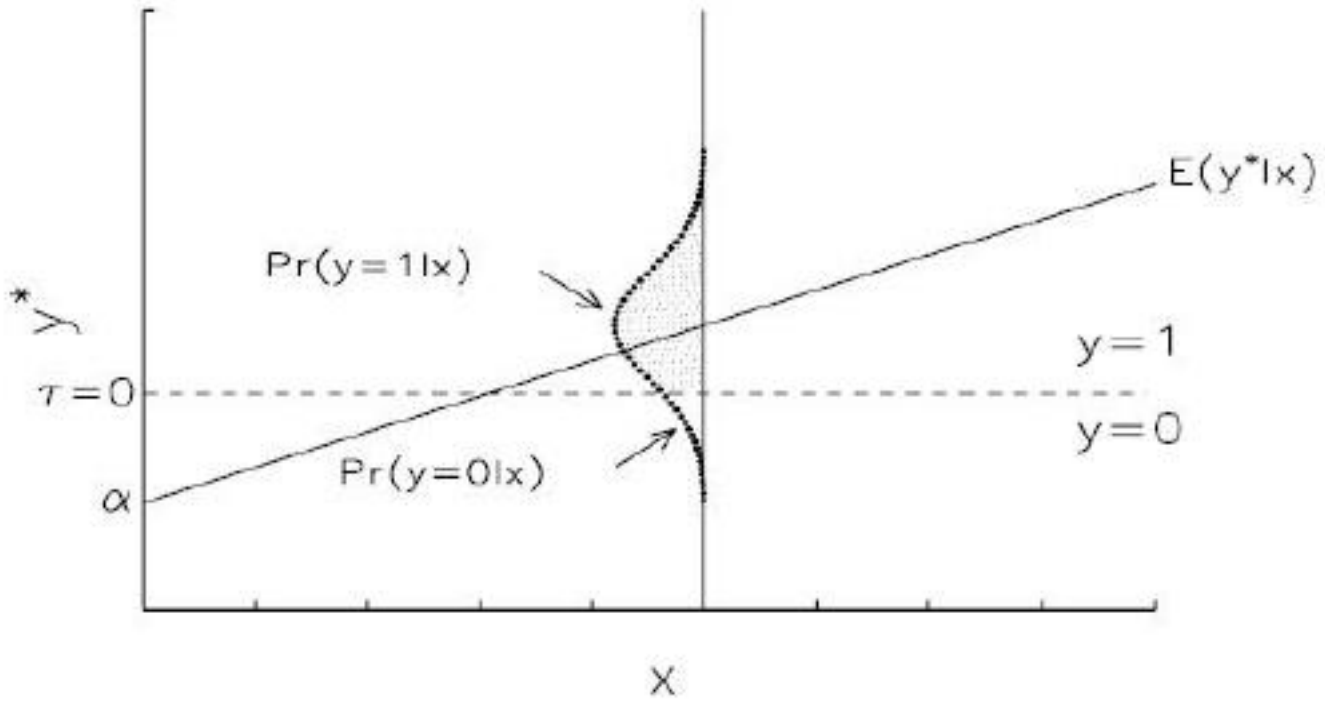
Specificazione con variabile latente

Con riferimento ad una variabile continua Y^* non osservabile, si assume che: $Y^* = X\beta + \varepsilon$
e per tale relazione si assumono le stesse ipotesi del modello lineare classico.

La variabile Y^* rappresenta una caratteristica non osservabile, che sta alla base della scelta effettuata o della possibilità di possedere o meno una certa caratteristica, che può essere assimilata alla *propensione al verificarsi dell'evento considerato*.

Ipotizziamo quindi che da tale variabile dipendano i valori osservati per una variabile dicotomica Y, secondo la relazione: $Y=1[Y^*>0]$,
dove la funzione 1[.] è la funzione indicatrice che assume valore 1 quando si verifica la condizione tra parentesi e 0 altrimenti.
In altri termini, la precedente relazione equivale a dire che si realizza $y=1$ se $y^*>0$ e $y=0$ se $y^*\leq 0$.

La relazione tra le possibili realizzazioni della Y^* latente, della Y osservata e delle corrispondenti due probabilità può essere così rappresentata:



Si può osservare come la specificazione con variabile latente Y^* e disturbo ε, coerentemente con quanto visto nella regressione lineare, consenta di giustificare, anche formalmente, la circostanza che due diversi individui (unità i e j), con medesime caratteristiche ($\mathbf{x}_i=\mathbf{x}_j$), possano effettuare due scelte diverse ($y_i \neq y_j$).

Per la componente ε si assume l'indipendenza da $\mathbf{x}$ e che si distribuisca secondo la generica densità f ; se questa è simmetrica intorno allo zero, risulterà $F(w)=1-F(-w)$, per qualsiasi w reale.

Pertanto la probabilità di risposta per Y risulta:

$$P(Y=1|X)=P(Y^*>0|X)=P(\varepsilon > -X\beta)|X)$$

$$=1-P(\varepsilon < -X\beta)|X)$$

$$= 1-F(-X\beta)=F(X\beta),$$

coerentemente con la specificazione inizialmente adottata per i modelli a risposta binaria.

Procedure di Test per Modelli *Logit* e *Probit*

Per la verifica di ipotesi riguardanti singoli coefficienti, si può fare ricorso al test z asintotico proposto nell'ambito del modello lineare.

Per verificare una restrizione multipla, partendo dal modello

$$P(Y=1|\mathbf{x}, \mathbf{z})=F(\mathbf{x}'\boldsymbol{\beta}+\mathbf{z}'\boldsymbol{\gamma}),$$

con $\dim(\mathbf{x})= k\times 1$ e $\dim(\mathbf{z})= q\times 1$,
si sottopone a verifica l'ipotesi nulla $H_0:\boldsymbol{\gamma}=0$, che implica q restrizioni di esclusione.

La **z** può comprendere sia variabili esplicative aggiuntive, sia forme funzionali delle **x** (ad esempio, quadrati o fattori di interazione).

In questo ambito, in alternativa al test *F* (basato sulle devianze residue dei modelli completo e ristretto), si fa riferimento al

test del rapporto di verosimiglianza

basato sul logaritmo del rapporto L_r / L_u.

La corrispondente statistica LR (da Likelihood Ratio test) è data da:

$$LR = -2[\log L_r - \log L_u],$$

dove L_r e L_u sono le funzioni di verosimiglianza valutate, rispettivamente, in corrispondenza del modello ristretto (con le sole esplicative $\mathbf{x}$) e di quello non ristretto (con tutte le esplicative $\mathbf{x}$ e $\mathbf{z}$).

La distribuzione asintotica sotto H_0 è una χ^2_q.

Tale statistica è denominata *deviance statistic* e per questo motivo è spesso indicata con D.

La denominazione *deviance statistic* è dovuta al fatto che la quantità $-2\log L$ può essere vista come una generalizzazione della devianza residua; infatti, in un modello con dati distribuiti normalmente, coincide proprio con la *SSE*.

La statistica *D* è quindi una misura della perdita di adattamento del modello ai dati e l'idea sottostante al test *LR* è del tutto simile a quella del test *F* per un modello lineare.

Infatti, se da una parte il test F si basa sull'aumento della devianza residua prodotto dall'esclusione di alcuni regressori dal modello,
il test LR misura la corrispondente riduzione nella log-verosimiglianza.

Infatti, l'esclusione di variabili in generale implica una riduzione della log-verosimiglianza, cioé ci si aspetta $\log L_u > \log L_r$,
il problema diventa allora stabilire quand'è che la riduzione nella log-likelihood sia abbastanza grande da far ritenere necessarie le variabili escluse.

Se tutte le variabili esplicative sono sottoposte alla restrizione, ad eccezione del termine costante, facendo quindi riferimento al cosiddetto *modello nullo*, la log-verosimiglianza ristretta (in questo caso indicata con $\log L_0$) risulta identica per i modelli probit e logit e pari a:

$$\log L_0 = n[P \log P + (1-P)\log(1-P)]$$

dove P è la proporzione campionaria dei successi, cioè di osservazioni con variabile dipendente pari a 1.

Bontà di adattamento dei modelli Logit e Probit

In letteratura sono state proposte molte misure, nessuna delle quali però è riuscita a riscuotere il successo dell'indice di determinazione R^2 nell'ambito della regressione lineare.

La misura più diffusa è l'*indice del rapporto delle verosimiglianze* (LRI),

proposto da McFadden, che si basa sul confronto tra il valore della log-verosimiglianza nel modello stimato e quello ottenuto nel modello nullo (con la sola intercetta).

Bontà di adattamento per Logit e Probit 2

In quest'ultimo caso si assume quindi che tutti gli altri parametri siano nulli; inoltre, come visto in precedenza, la corrispondente $\log L_0$ dipende unicamente dalle proporzioni campionarie dei successi.

L'indice LRI è considerato l'omologo in ambito non lineare, dell'indice R^2 nella regressione lineare:

$$LRI = 1 - \log L / \log L_0;$$

$$LRI_{adj} = 1 - (\log L - k)/\log L_0.$$

Bontà di adattamento per Logit e Probit 3

L'adattamento del modello è da ritenere tanto migliore, quanto maggiore è *LRI*.

La caratteristica interessante di questo indice, rispetto alle misure alternative proposte in questo ambito, è che esso varia tra 0 e 1.
Infatti, poiché le log-verosimiglianze non sono altro che somme di logaritmi di probabilità, cioè di numeri negativi, è sempre vero che

$$\log L_0 \leq \log L \leq 0$$

e quindi che

$$0 \leq \log L / \log L_0 \leq 1.$$

Bontà di adattamento per Logit e Probit 4

In particolare, il valore 0 viene assunto quando tutti i coefficienti β (esclusa l'intercetta) sono nulli, cioè quando $\log L = \log L_0$.

L'estremo superiore, invece, non è mai raggiunto e costituisce solo un limite superiore di riferimento.

Bontà di adattamento per Logit e Probit 5

Tale indice non va però interpretato come percentuale di variabilità spiegata (come l'indice R^2), quanto piuttosto come indicatore della "vicinanza" tra modello e osservazioni.

Si può poi osservare che i valori ottenuti per questo pseudo-R^2 sono generalmente molto più bassi di quelli che si ottengono in una regressione lineare per l'R^2 e l'utilizzo dell'indice è finalizzato soprattutto al confronto tra le performance di specificazioni alternative di un modello.

Si ricorda inoltre che sempre sulla verosimiglianza si basano i criteri di Akaike e Schwarz, che consentono di confrontare l'adattamento di modelli alternativi, portando a preferire quelli a cui corrispondono i valori più bassi di tali indici:

$$AIC = -2\log L + 2k$$

$$BIC = -2\log L + k \log n.$$

Un'ulteriore misura che viene usualmente riportata è la percentuale di osservazioni correttamente stimate, ottenuta confrontando i valori osservati per la Y con quelli stimati.

Per ottenere questi ultimi si calcola, per ogni i, la probabilità condizionata di successo, mediante il valore assunto dalla funzione F in corrispondenza dei valori stimati per i β_j,

$$\hat{P}(Y_i|\boldsymbol{x}_i) = F(\boldsymbol{x}_i\boldsymbol{\beta})$$

Se questa risulta maggiore di una certa soglia prefissata F^*, generalmente pari a ½, allora il valore stimato per y_i è 1, altrimenti gli si attribuisce il valore 0.

La percentuale di volte in cui i valori stimati coincidono con quelli osservati costituisce la suddetta misura di valutazione (*percent correctly predicted*) …che però va presa con le pinze!

Infatti, in molte situazioni è più facile stimare correttamente una sola delle due modalità e, se questa è quella più frequente, la percentuale di predizioni esatte può continuare ad essere elevata.

In tal caso, la misura risulta fuorviante.

Esempio: Supponiamo che in un campione di 200 unità si osservino 180 determinazioni y=0 e che 150 di queste vengano stimate correttamente.
Anche se in nessun caso venisse stimato correttamente il valore y=1, la percentuale di stime corrette sarebbe comunque pari al valore non banale del 75%!

E’ preferibile quindi calcolare la suddetta percentuale separatamente per ciascuna modalità della *Y*.

Ovviamente, la percentuale globale è ottenibile come media ponderata delle due, con pesi pari alla proporzione di 0 e 1 nel campione.

Le sintesi effettuabili si possono schematizzare nella cosiddetta *confusion matrix*:

Bontà di adattamento per Logit e Probit 11

y previste y osservate	0	1	Totale
0	n_{11}	n_{12}	$n_{1.}$
1	n_{21}	n_{22}	$n_{2.}$
Totale	$n_{.1}$	$n_{.2}$	$n_{..}$

e le percentuali calcolabili sono:

$(n_{11}+n_{22})/n$; $n_{11}/n_{1.}$; $n_{22}/n_{2.}$.

In particolare, il rapporto $n_{22}/n_{2.}$, cioè la proporzione di predizioni corrette della modalità 1, è detto *sensitivity* (sensibilità).

Il rapporto $n_{11}/n_{1.}$, cioè la proporzione di predizioni corrette della modalità 0, è detto *specificity* (specificità).

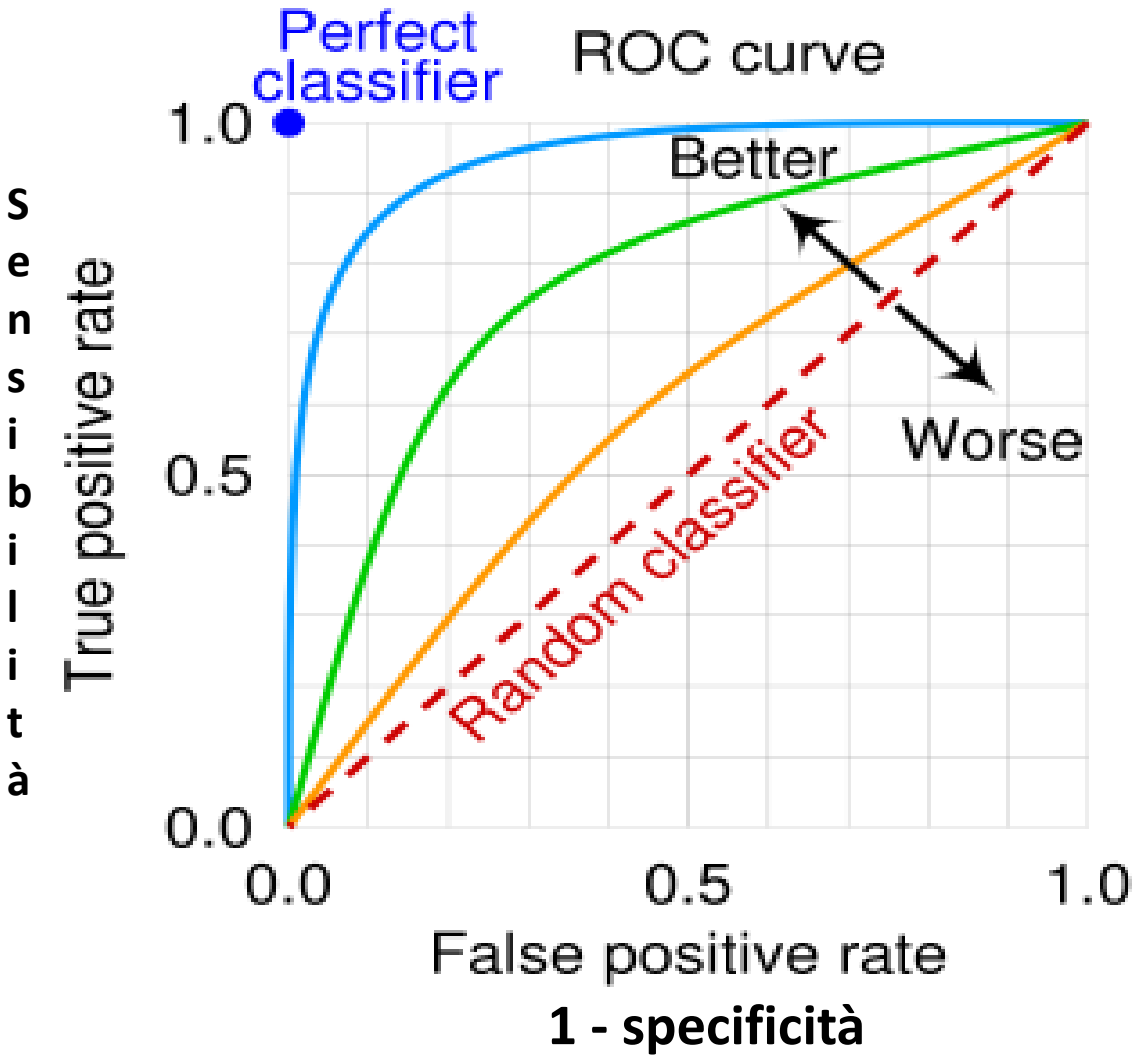
Una ulteriore misura di accuratezza, particolarmente usata anche nei test diagnostici di una malattia, si ottiene mettendo a confronto la percentuale di 'veri positivi' con quella di 'falsi positivi', cioè sensibilità vs 1-specificità.

Se il confronto è fatto graficamente, usando come coordinate le suddette proporzioni, si ottiene una curva chiamata

curva ROC

L'area sottostante alla curva ROC, denominata AUC (acronimo di Area Under the Curve), misura l'accuratezza delle stime ottenute.
Essa varia tra 0.5 - corrispondente all'area sottostante la bisettrice e a una situazione in cui non si riesce a discriminare tra veri e falsi positivi - e 1, corrispondente alla situazione in cui si discrimina perfettamente tra i due gruppi.

Bontà di adattamento per Logit e Probit 14



Un punto di debolezza di tali misure, tuttavia, risiede nella scelta della soglia F^*.

Infatti, dal precedente esempio risulta anche evidente che per campioni sbilanciati, cioè con una modalità molto più frequente dell'altra, il valore ½ risulta certamente inadeguato, in quanto non consente quasi mai di predire correttamente la modalità meno frequente.

In situazioni del genere, sarebbe ovviamente opportuno "calibrare" la soglia rispetto al campione osservato, ma ciò si pagherebbe con una diminuzione della proporzione di predizioni corrette per la modalità più frequente.

Ad esempio, se nel campione vi fosse una preponderanza di 0, come nel precedente esempio, allora per aumentare le predizioni corrette per la modalità 1 occorrerebbe abbassare la soglia rispetto a ½, ma ciò comporterebbe automaticamente una riduzione delle predizioni corrette di 0.

In altri termini, variazioni nella soglia F^*, che riducano la probabilità di un tipo di errore, necessariamente aumenterebbero quella dell'altro tipo e inoltre diventerebbero alquanto soggettive.

Ciò che va considerato a questo riguardo è che i due tipi di errore che possiamo commettere (predire 0 al posto di 1 e viceversa) non sono in genere simmetrici nei costi che comportano e quindi la scelta di F^* va valutata accuratamente.

Per meglio comprendere la differenza tra i due tipi di errore, si può considerare quanto segue.

Esempio: per un istituto di credito classificare erroneamente una richiesta di finanziamento come ad alto rischio (e quindi da evitare) può costituire un'opportunità mancata, ma classificare come valida una richiesta che non lo è può comportare ben altri costi!

Bontà di adattamento per Logit e Probit 19

La gamma delle misure proposte non si esaurisce qui, ma è bene considerare che per questo tipo di modelli è molto più importante la significatività, statistica ed economica, delle covariate, piuttosto che la bontà di adattamento del modello.

Inoltre, non bisogna dimenticare che l'ottica con cui si stimano i coefficienti del modello non è quella di massimizzare una qualche misura di adattamento, come avviene nella regressione classica, dove lo stimatore **B** massimizza l'indice R^2.