George Bush (codice 1), mentre 354 hanno detto di avere preferito il candidato Democratico Michael Dukakis (codice 0). La proporzione di intervistati schierati a favore di Bush è esattamente uguale a 0,60. Se si applica a questi dati un normale modello di regressione lineare in cui la variabile dipendente è la scelta di voto (Bush contro Dukakis) e le variabili indipendenti sono quattro, si ottengono le seguenti stime dei parametri:

$$\hat{Y}_i = \underset{(0,08)}{0,07} + \underset{(0,01)}{0,13}X_{1i} + \underset{(0,01)}{0,04}X_{2i} + \underset{(0,01)}{0,14}X_{3i} + \underset{(0,04)}{0,01}X_{4i} \qquad \underset{(0,004)}{} \qquad R^2_{\text{rtf}} = 0,456$$

dove $\hat{Y}$ denota il voto atteso a favore di Bush; X_1 denota l'identificazione di partito dell'intervistato (da 0 = democratico convinto a 6 = repubblicano convinto); X_2 indica il suo orientamento politico (da 1 = estrema sinistra a 7 = estrema destra); X_3 è una variabile indicatore che rappresenta la razza dell'intervistato (1 = bianco, 0 = altrimenti); e X_4 denota gli anni di istruzione formale (da 0 a 20). Poiché la variabile dipendente può assumere solo due valori, l'equazione può essere interpretata come un **modello lineare di probabilità** del voto a favore di Bush. Ad esempio, poiché il coefficiente di regressione associato alla variabile X_1 assume un valore pari a 0,13, possiamo affermare che ogni spostamento dell'identificazione di partito verso il polo repubblicano aumenta la probabilità di votare per Bush di 0,13; analogamente, i bianchi hanno il 14% di probabilità in più di votare per George Bush, e così via.

I modelli lineari di probabilità, tuttavia, possiedono una caratteristica che li rende indesiderabili: applicando il metodo della regressione lineare a variabili dipendenti di tipo dicotomico si violano due assunti fondamentali dell'analisi di regressione. Innanzitutto, gli errori non sono più distribuiti normalmente. Come sappiamo, nella regressione l'errore equivale alla differenza fra un valore osservato e uno atteso in base al modello prescelto:

$$e_i = Y_i - \hat{Y}_i = Y_i - (\alpha + \sum \beta_{ji} X_{ji})$$

Tuttavia, poiché i valori osservati della variabile dipendente possono essere solo due, cioè 1 e 0, anche gli errori corrispondenti possono assumere solo due valori: quando $Y_i = 1$, allora $e_i = 1 - \alpha - \Sigma\beta_{ji}X_{ji}$; quando, invece, $Y_i = 0$, allora $e_i = -\alpha - \Sigma\beta_{ji}X_{ji}$. La conseguenza di ciò è che sebbene le stime OLS dei parametri β_j rimangano corrette, esse cessano di essere le stime più efficienti (cioè quelle con la varianza campionaria più piccola). Dunque, i test di significatività basati su queste stime e sui loro errori standard possono indurre il ricercatore a trarre conclusioni non valide, anche quando si analizzano campioni numerosi.

Il secondo problema legato ai modelli lineari di probabilità consiste nel fatto che alcuni valori predetti dall'equazione possono essere privi di senso. Poiché i parametri esprimono relazioni lineari multivariate che legano le variabili indipendenti a quella dipendente, i valori attesi in corrispondenza di qualche particolare combinazione delle variabili indipendenti possono ricadere al di fuori dell'intervallo 0-1. Per illustrare questa possibilità, si consideri il caso di un elettore che manifesta valori estremi in tutte e quattro le variabili indipendenti:

$$\hat{Y}_i = 0,07 + 0,13(6) + 0,04(7) + 0,14(1) - 0,01(0) = 1,27$$

È evidente che una probabilità di votare Bush pari a 1,27 non ha alcun senso. Analogamente, si consideri il caso di un altro elettore che, come il precedente, manifesta valori estremi in tutte e quattro le variabili indipendenti, ma questa volta nella direzione opposta:

$$\hat{Y}_i = 0,07 + 0,13(0) + 0,04(1) + 0,14(0) - 0,01(20) = -0,09$$

Anche in questo caso ci troviamo di fronte a un valore, specificamente una probabilità negativa, che è assolutamente privo di senso.

Dovrebbe essere evidente, dunque, che il modello lineare di probabilità è insoddisfacente e deve essere sostituito da una valida alternativa. Fortunatamente questa alternativa esiste ed è applicabile all'analisi delle variabili dipendenti sia di tipo dicotomico che di tipo discreto a più categorie.

3. La trasformazione logistica e le sue proprietà

Le percentuali (%) e le proporzioni (p) non rappresentano gli unici modi per misurare una variabile dipendente. La **trasformazione logistica di *p*** costituisce una valida alternativa caratterizzata da alcune proprietà interessanti. L'unità di probabilità logistica relativa a una data osservazione i, detta **logit**, si ottiene formando l'odds di p_i rispetto al suo reciproco $(1 - p_i)$ e calcolando il logaritmo naturale di questo rapporto. In altri termini, il logit equivale al logaritmo naturale di un odds:

$$L_i = \ln\left(\frac{p_i}{1 - p_i}\right)$$

Il logit si distribuisce simmetricamente intorno a un valore centrale. Quando $p_i = 0,50$ il suo reciproco assume lo stesso valore: $1 - 0,50 = 0,50$. Il logaritmo naturale del rapporto fra questi due valori è $L_i = \ln(0,50/0,50) = \ln 1 = 0$. Mano a mano che la dicotomia si sposta

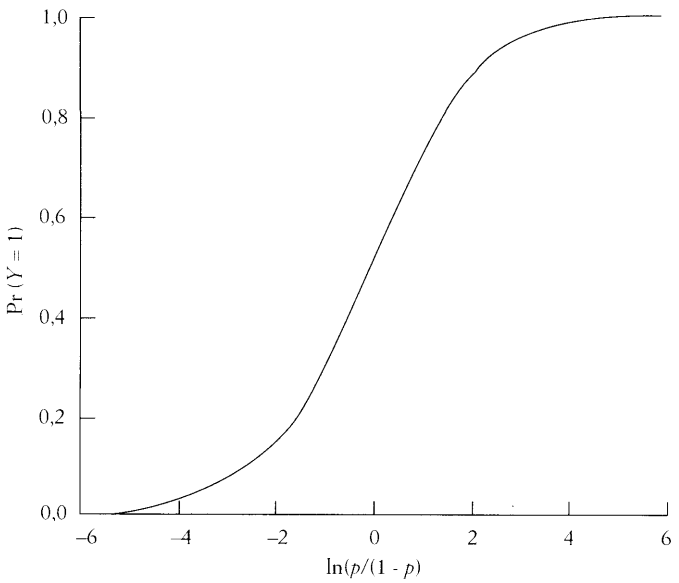

FIG. 9.4. Trasformazione logistica delle probabilità.

verso una delle due direzioni, approssimando 0 o 1, i valori del logit si allontanano da 0, come mostra la seguente tabella:

p_i	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90
$1 - p_i$	0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10
logit	–2,20	–1,39	–0,85	–0,41	0,00	0,41	0,85	1,39	2,20

Come si può vedere, sebbene tutte le probabilità siano equidistanti (con un intervallo pari a 0,10), la distanza fra i logit corrispondenti aumenta sempre di più mano a mano che ci si allontana da $p_i = 0{,}50$. Inoltre, sebbene non esista alcun limite inferiore o superiore per il logit, quando p_i è esattamente uguale a 1 o a 0 il logit risulta indefinito. La figura 9.4 illustra la trasformazione continua delle probabilità nei loro logit. Questa curva a S mostra una forte somiglianza con la distribuzione cumulativa delle probabilità della variabile normale standardizzata. Nell'intervallo compreso fra $p_i = 0{,}25$ e $p_i = 0{,}75$ la trasformazione logistica è pressoché lineare; di conseguenza, all'interno di questa gamma di valori il modello lineare di probabilità produce risultati molto simili a quelli del modello logistico. Tuttavia, quando la probabilità si avvicina ai due estremi della distribuzione la non linearità della trasformazione logistica si accentua. È interessante osservare che mano a mano che il valore assoluto del logit aumenta, la probabi-

lità corrispondente si avvicina sempre di più ai suoi valori estremi (0 o 1), senza però raggiungerli mai. Questo vincolo imposto ai valori attesi della trasformazione logistica offre a quest'ultima un vantaggio rilevante rispetto al modello lineare di probabilità.

Poiché «appiattiscono» le probabilità molto elevate e quelle molto basse, i logit risultano utili quando si tratta di effettuare confronti fra proporzioni a diversi livelli. La tabella 9.1 mostra i percorsi scolastici seguiti da quattro coorti di cittadini americani che hanno cominciato la quinta elementare rispettivamente negli anni 1960, 1955, 1951 e 1945. Il riquadro superiore riporta le proporzioni di ciascuna coorte rispettivamente per la terza media, il diploma di maturità e l'iscrizione all'università. Come si può vedere, nel periodo preso in considerazione i tassi di passaggio da un livello scolastico a quello successivo sono aumentati progressivamente. I confronti fra coorti, tuttavia, sono complicati dal fatto che i tassi i partenza di ciascun livello scolastico sono diversi. Se, per esempio, confrontiamo la coorte del 1945 con quella del 1960 possiamo vedere che il tasso di frequenza della terza media è aumentato di 0,109 punti, mentre la proporzione di diplomati e quella di iscritti all'università sono aumentate rispettivamente di 0,265 e 0,218 punti. Se adottiamo un'ottica diversa, però, possiamo affermare che dal 1945 al 1960 il tasso di ingresso all'università è quasi raddoppiato, il tasso di diplomati è aumentato del 50% e la frequenza della terza media è aumentata di circa un ottavo. Entrambe queste interpretazioni delle proporzioni suggeriscono che nel periodo 1945-1960 la frequenza scolastica è aumentata in modo diverso secondo il livello scolastico. Tuttavia, poiché le proporzioni e le percentuali non possono assumere valori al di fuori dell'intervallo 0-1 (0-100 nel caso delle percentuali), questi confronti non sono appropriati: un cambiamento dell'1% in un tasso pari al 50% non è uguale a un cambiamento dell'1% in un tasso del 98%.

In virtù della sua natura simmetrica, la trasformazione logistica di un odds $(p_i/(1-p_i))$ e del suo reciproco $((1-p_i)/p_i)$ dà luogo a logit di uguale valore ma di segno opposto. Per esempio, alle proporzioni 0,75 e 0,25 corrisponde un odds pari a $0{,}75/(1-0{,}75) = 3$; il reciproco di questo odds è uguale a $0{,}25/(1-0{,}25) = 0{,}333$. Se calcoliamo i logaritmi naturali di questi due valori otteniamo i seguenti logit: 1,099 e −1,099. Il secondo e il terzo riquadro della tabella 9.1 riportano i vari tassi di partecipazione scolastica rispettivamente in forma di odds e di logit. La figura 9.5 rappresenta graficamente questi logit, distinti per coorte e livello scolastico. Come si può vedere, per ogni livello scolastico l'andamento nel tempo assume una forma pressoché lineare. Possiamo dunque concludere che, nel periodo considerato, la partecipazione scolastica è cresciuta a tassi approssimativamente costanti.

I logit costituiscono la base per costruire un'alternativa al modello lineare di probabilità. Secondo quest'ultimo, la probabilità che l'osservazione i assuma valore 1 in corrispondenza della variabile dipendente

TAB. 9.1. *Tassi di accesso a tre diversi livelli scolastici, secondo la coorte*

Anno di inizio della quinta elementare	Livelli scolastici successivi		
	Accesso alla terza media	Conseguimento diploma di maturità	Iscrizione all'università
		Proporzioni	
1960	0,967	0,787	0,452
1955	0,948	0,642	0,343
1951	0,921	0,582	0,308
1945	0,858	0,522	0,234
		Odds	
1960	29,303	3,695	0,825
1955	18,231	1,793	0,522
1951	11,658	1,392	0,445
1945	6,042	1,092	0,305
		Logit	
1960	3,378	1,307	–0,193
1955	2,903	0,584	–0,650
1951	2,456	0,331	–0,809
1945	1,799	0,088	–1,186

Fonte: U.S. Census Bureau, *Historical Statistics of the United States: Colonial Times to 1970. Part 1*, Washington, D.C., U.S. Government Printing Office, Series H587-597, 1975, p. 379.

può essere espressa come funzione lineare e additiva di K variabili indipendenti. Formalmente:

$$\Pr(Y_i = 1) = p_i = \alpha + \sum_{j=1}^{K} \beta_j X_{ji}$$

Se, per semplificare la notazione, poniamo:

$$Z = \alpha + \sum_{j=1}^{K} \beta_j X_{ji}$$

possiamo scrivere la forma funzionale generale di un'equazione logistica come segue:

$$F(Z) = \frac{e^Z}{1 + e^Z} = \frac{1}{1 + e^{-Z}}$$

Facendo le opportune sostituzioni, la probabilità che l'osservazione i assuma valore 1 in corrispondenza della variabile dipendente può essere calcolata come segue:

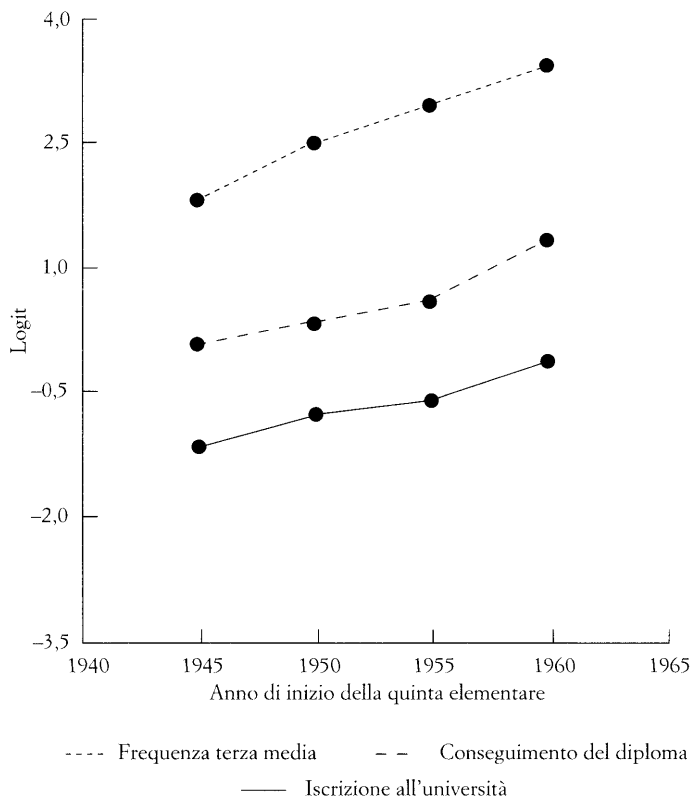

FIG. 9.5. Rappresentazione grafica dei logit di accesso a tre diversi livelli scolastici, secondo la coorte anagrafica.

$$\hat{p}_i = \frac{1}{1 + e^{-\alpha - \Sigma \beta_j X_{ji}}}$$

Poiché il logit associato all'osservazione i corrisponde al logaritmo naturale dell'odds, si ha:

$$L_i = \ln\left(\frac{p_i}{1 - p_i}\right) =$$

$$= \ln(e^{\alpha + \Sigma \beta_j X_{ji}}) =$$

$$= \alpha + \Sigma \beta_j X_{ji}$$

Il quadro 9.1 illustra i dettagli di questa derivazione.

QUADRO 9.1. Derivazione del logit

In primo luogo, per semplificare la notazione si ponga:

$$Z = \alpha + \sum_{j=1}^{K} \beta_j X_{ji}$$

Poiché la probabilità che l'osservazione i assuma valore 1 in corrispondenza della variabile dipendente è uguale a:

$$p_i = \frac{1}{1 + e^{-Z}}$$

il suo reciproco deve essere:

$$1 - p_i = 1 - \frac{1}{1 + e^{-Z}} = \frac{1 + e^{-Z} - 1}{1 + e^{-Z}} = \frac{e^{-Z}}{1 + e^{-Z}}$$

Formando il rapporto fra questi due reciproci e semplificando si ottiene:

$$\frac{p_i}{1 - p_i} = \frac{1/(1 + e^{-Z})}{e^{-Z}/(1 + e^{-Z})} = \frac{1}{e^{-Z}} = e^{Z}$$

Se, a questo punto, si calcola il logaritmo naturale del rapporto fra le due probabilità si ottiene:

$$\ln\left(\frac{p_i}{1 - p_i}\right) = \ln\left(\frac{1}{e^{-Z}}\right) = \ln(e^{Z}) = Z$$

Infine, sostituendo Z e definendo il risultato come logit L dell'osservazione i si ha:

$$\ln\left(\frac{p_i}{1 - p_i}\right) = L_i = \alpha + \sum \beta_j X_{ji}$$

La figura 9.6 illustra la differenza fra due ipotetiche curve di regressione calcolate utilizzando gli stessi dati: la prima rappresenta il modello lineare di probabilità, mentre la seconda rappresenta il modello logistico. Come si può osservare, la probabilità attesa in base al modello lineare risulta minore di zero e maggiore di 1 in corrispondenza dei valori estremi di Z; al contrario, la probabilità attesa in base al modello logistico non supera mai questi limiti, qualunque sia il valore di Z.

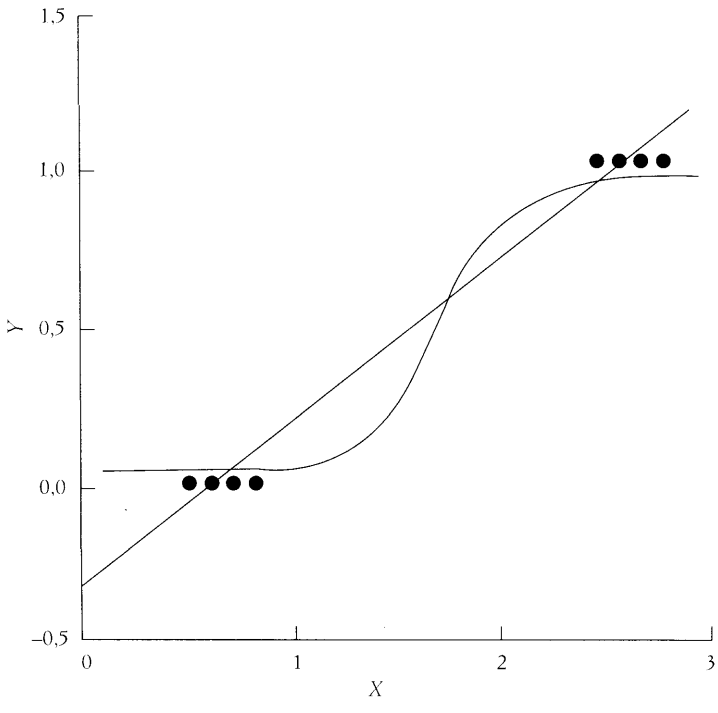

FIG. 9.6. Confronto fra il modello lineare di probabilità e il modello logistico.

Sebbene la probabilità sottostante non sia una funzione lineare delle variabili indipendenti, la trasformazione logistica fa sì che i logit siano una funzione lineare delle variabili indipendenti. Ogni logit è direttamente interpretabile come il logaritmo naturale del rapporto fra la probabilità che $Y = 1$ e la probabilità che $Y = 0$. Data la simmetria della curva logistica mostrata nella figura 9.4, quando la probabilità che $Y = 1$ è esattamente uguale a 0,50 allora il logit risulta pari a zero. Quando, invece, la probabilità che $Y = 1$ è maggiore della probabilità che $Y = 0$, allora il logit assume valori maggiori di zero. Infine, quando la probabilità che $Y = 1$ è minore della probabilità che $Y = 0$, allora il logit assume valori minori di zero. Il logit risulta indefinito quando la probabilità che $Y = 0$ è esattamente uguale a zero: qualsiasi divisione per zero, infatti, è matematicamente impossibile. Tuttavia, quando la probabilità che $Y = 1$ approssima la certezza (cioè $p_{Y=1} \rightarrow 1$) – e, quindi, la probabilità che $Y = 0$ si avvicina a zero (cioè $p_{Y=0} \rightarrow 0$) – allora il loro odds approssima l'infinito ($p_{Y=1}/\ p_{Y=0} \rightarrow \infty$), così come il loro logit. Nel caso opposto in cui la probabilità che $Y = 0$ approssima la certezza – e, quindi, la probabilità che $Y = 1$ si avvicina a zero –

allora l'odds corrispondente si avvicina a zero mentre il logit approssima l'infinito negativo. Queste proprietà del logit ne fanno una forma funzionale molto utile per l'analisi multivariata.

4. Stima e test delle equazioni di regressione logistica

4.1. Le stime dei parametri

La regressione logistica ha molti punti in comune con l'analisi della regressione multipla; la differenza sostanziale fra le due tecniche consiste nel fatto che nella prima la variabile dipendente è il logit di una classificazione dicotomica, mentre nella seconda è una misura continua. Esattamente come nella regressione lineare multipla, nella regressione logistica le variabili indipendenti possono essere di tipo continuo, dicotomico o discreto a k categorie, nonché comprendere termini interattivi. La forma fondamentale dell'**equazione di regressione logistica dicotomica** con K variabili indipendenti è la seguente:

$$\hat{L}_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki}$$

Come si può vedere, questa equazione afferma che il valore atteso del logaritmo naturale del rapporto fra le due probabilità associate alla dicotomia $(p_i/(1-p_i)$ è una funzione lineare delle K variabili indipendenti prescelte. Se si calcola l'antilogaritmo di tale equazione è possibile evidenziare il modo in cui le variabili indipendenti influiscono sull'odds:

$$e^{\hat{L}_i} = e^{\ln(p_{Y=1}/p_{Y=0})} = \frac{p_{Y=1}}{p_{Y=0}} = e^{\alpha+\beta_1 X_{1i}+\beta_2 X_{2i}+\ldots+\beta_K X_{Ki}}$$

L'equazione di regressione logistica assomiglia a quella di regressione lineare multipla in quanto ciascun coefficiente β_j indica la misura in cui il logit cambia quando la corrispondente variabile indipendente X_i aumenta o diminuisce di una unità.

I parametri della regressione logistica non possono essere stimati utilizzando il metodo dei minimi quadrati comuni. Piuttosto, bisogna ricorrere a un metodo noto come **stima di massima verosimiglianza** (MLE, dall'inglese *Maximum Likelihood Estimation*). In breve, la stima di massima verosimiglianza si basa su una serie di approssimazioni successive ai valori incogniti dei veri parametri della popolazione, α e β_j. L'obiettivo è quello di utilizzare i dati campionari per ottenere stime dei parametri a e b_j che massimizzino la probabilità di ottenere i valori campionari osservati. A differenza del metodo dei minimi quadrati comuni, che valuta la «bontà di adattamento» ai dati del modello calcolando la somma delle differenze al quadrato fra i valori predetti e

quelli osservati, il metodo della massima verosimiglianza calcola la probabilità di osservare ciascun possibile valore campionario Y_i assumendo che un dato insieme di parametri sia quello vero. L'insieme di parametri che risulta associato alla probabilità più elevata costituisce l'insieme delle stime di massima verosimiglianza ricercate.

Poiché alla stima di massima verosimiglianza non sono associate formule algebriche simili alle equazioni normali utilizzate nella stima dei minimi quadrati comuni, la sua soluzione richiede l'uso di un algoritmo – incorporato in un programma per computer – che consenta di esaminare in modo iterativo numerosi insiemi di parametri, fino a quando viene identificato l'insieme migliore. La prima fase della procedura consiste nello stabilire le stime iniziali dei parametri, generalmente fissate a zero. Una serie di *iterazioni* produce in successione nuove stime e le confronta con quelle precedenti. Le iterazioni continuano fino a quando le stime ottenute in un dato ciclo differiscono da quelle ottenute nel ciclo precedente di una quantità inferiore a una certa soglia prestabilita. Nei campioni sufficientemente numerosi le stime di massima verosimiglianza dei parametri sono corrette, efficienti e distribuite normalmente; quest'ultima proprietà permette di effettuare test di significatività utilizzando le statistiche che abbiamo già incontrato nei capitoli precedenti.

Nel paragrafo 2 abbiamo illustrato un esempio in cui il voto espresso alle elezioni presidenziali americane del 1988 (Bush contro Dukakis) era visto come funzione di quattro variabili indipendenti: identificazione di partito, orientamento politico, razza e istruzione. Se utilizziamo queste stesse variabili per stimare un'equazione di regressione logistica, otteniamo i seguenti risultati (errori standard fra parentesi):

$$\hat{L}_i = \underset{(0,61)}{-2,68} + \underset{(0,05)}{0,79X_{1i}} + \underset{(0,08)}{0,33X_{2i}} + \underset{(0,33)}{0,87X_{3i}} - \underset{(0,03)}{0,08X_{4i}}$$

I valori t calcolati per ciascuna variabile indipendente sono tutti statisticamente significativi con una probabilità $p < 0,01$ di commettere un errore di I tipo.

L'interpretazione diretta dei vari effetti è problematica perché richiede di pensare in termini di logit che, certamente, non costituiscono una unità di misura «naturale». Tuttavia, il segno associato a ciascun parametro – positivo o negativo – offre una prima indicazione sulla direzione della relazione (diretta o inversa) esistente fra la variabile indipendente corrispondente e la variabile dipendente. Nel nostro esempio, i coefficienti positivi indicano che la propensione (espressa in termini di logit) a votare per Bush (anziché per Dukakis) è maggiore fra coloro che si identificano con il Partito repubblicano (X_1), hanno un orientamento politico conservatore (X_2) e sono bianchi (X_3), mentre diminuisce all'aumentare degli anni di istruzione (X_4).

La procedura di costruzione di un intervallo di confidenza intorno alla stima puntuale di un coefficiente di regressione logistica è identica a quella utilizzata nel caso della regressione lineare. Per un dato livello α, i limiti di confidenza inferiore e superiore intorno alla stima b_j equivalgono a:

$$b_j \pm t_{\alpha/2} s_{b_j}$$

dove:
s_{b_j} = errore standard della stima del coefficiente b_j.

A titolo di esempio, i limiti dell'intervallo di confidenza al 95% intorno al coefficiente b_1 (che esprime l'effetto dell'identificazione di partito) sono LIC = 0,79 – (1,96)(0,05) = 0,692 e LSC = 0,79 + (1,96)(0,05) = 0,888.

Una semplice trasformazione consente di interpretare l'effetto netto esercitato da una variabile indipendente dicotomica (variabile indicatore) sulle probabilità associate alla variabile dipendente. Tale trasformazione consiste semplicemente nel moltiplicare la stima del parametro di interesse, b_j, per la varianza della variabile dipendente, cioè $(p_i)(1 - p_i)$. Il valore così ottenuto indica l'effetto proporzionale esercitato dalla variabile indipendente sulla probabilità che la variabile dipendente assuma valore 1, al netto degli effetti esercitati dagli altri predittori. Tornando al nostro esempio, la probabilità di votare per Bush ($p_{Y=1}$) è pari a 0,60 e, quindi, la varianza della variabile dipendente risulta uguale a (0,60)(1 – 0,60) = 0,24. Pertanto, l'effetto relativo esercitato dalla razza bianca sulla probabilità di votare per Bush è pari a (0,87)(0,24) = 0,21; in altri termini, fra i bianchi la probabilità di votare per Bush supera del 21% (in termini relativi) l'analoga probabilità osservata fra i non bianchi.

Nel paragrafo 2 abbiamo visto che, utilizzando il modello lineare di probabilità, i valori estremi delle variabili indipendenti possono dare luogo a predizioni che vanno oltre i limiti «naturali» della variabile dipendente (cioè 0 e 1). Questo problema viene superato nel caso della regressione logistica. Se applichiamo i seguenti valori estremi (a favore di Bush) alla nostra equazione di regressione logistica otteniamo:

$$\hat{L}_i = -2,68 + 0,79(6) + 0,33(7) + 0,87(1) - 0,08(0) =$$

$$= -2,68 + 4,74 + 2,31 + 0,87 + 0 = 5,24$$

Per trasformare questo logit pari a 5,24 in una probabilità dobbiamo rammentare la forma della funzione logistica: $F(Z) = e^Z/(1 + e^Z)$, dove $Z = \alpha + \Sigma\beta_j X_{ji}$. Dunque, la probabilità attesa che una persona con le caratteristiche sopra indicate voti per Bush risulta uguale a:

$$P_{Y=1} = \frac{e^{5,24}}{1+e^{5,24}} = \frac{188,67}{1+188,67} = \frac{188,67}{189,67} = 0,995$$

Come si può vedere, alla combinazione di valori estremi pro-Bush delle variabili indipendenti corrisponde una probabilità del 99,5%, e non l'impossibile 127% predetto dal modello lineare di probabilità.

4.2. La misura della bontà di adattamento ai dati

Per valutare la bontà di adattamento ai dati di un modello di regressione logistica sono disponibili quattro misure. La prima, nota come **rapporto di verosimiglianza**, pone a confronto due equazioni di regressione logistica concatenate, una delle quali è una versione ristretta (cioè con meno variabili indipendenti) dell'altra. Se l'equazione più ampia comprende K_1 variabili indipendenti e quella ristretta ne comprende K_0, ognuna delle quali è inclusa anche nel modello più ampio, l'ipotesi nulla afferma che nessuno dei $K_1 - K_0$ coefficienti di regressione logistica supplementari è significativamente diverso da zero nella popolazione. In altri termini, per ogni β_j in cui $j > K_0$, l'ipotesi nulla è H_0: $\beta_{K_0+1} = \beta_{K_0+2} = \ldots = \beta_{K_1} = 0$. L'ipotesi alternativa afferma che almeno uno dei parametri β_j è significativamente diverso da zero.

L_1 indica la verosimiglianza massimizzata relativa all'equazione più ampia, quella con K_1 variabili indipendenti; ad essa sono associati $N - K_1 - 1$ gradi di libertà. L_0, invece, indica la verosimiglianza dell'equazione ristretta, quella con K_0 variabili indipendenti; ad essa sono associati $N - K_0 - 1$ gradi di libertà. La statistica del test G^2, che si basa sul rapporto fra queste due verosimiglianze, è distribuita come una variabile chi-quadrato con un numero di gradi di libertà pari alla differenza fra il numero di variabili indipendenti presenti nella prima equazione e il numero di variabili indipendenti incluse nella seconda equazione, cioè $gl = K_1 - K_0$. La formula della statistica del test è la seguente:

$$G^2 = -2\ln\left(\frac{L_0}{L_1}\right) = (-2\ln L_0) - (-2\ln L_1)$$

L'applicazione più comune del test del rapporto di verosimiglianza consiste nel verificare l'ipotesi che *tutti* i parametri di un'equazione di regressione logistica siano uguali a zero, ad eccezione dell'intercetta α che, in questo caso, equivale al logit della proporzione di casi che assumono valore 1 in corrispondenza della variabile dipendente. Nell'esempio del voto presidenziale, all'equazione che include solo l'intercetta ($K_0 = 0$) è associato il valore $-2\ln L_0 = 1.191,2$. All'equazione che

include tutte e quattro le variabili indipendenti prescelte ($K_1 = 4$), invece, è associato il valore $-2\ln L_1 = 712{,}1$. Di conseguenza, $G^2 = (-2\ln L_0) - (-2\ln L_1) = 1.191{,}2 - 712{,}1 = 479{,}1$, con $gl = 4$. Se poniamo $\alpha = 0{,}05$, il valore critico del test equivale a 9,49. Pertanto, possiamo tranquillamente rifiutare l'ipotesi nulla secondo la quale nella popolazione nessuna delle quattro variabili indipendenti prescelte esercita un effetto sulla variabile dipendente.

Il test del rapporto di verosimiglianza può essere utilizzato anche per verificare la significatività statistica di ciascun coefficiente di regressione logistica. In sostanza, si tratta di stimare un insieme di K equazioni ristrette, ognuna delle quali omette solo una delle variabili indipendenti. Ciascuna differenza fra la verosimiglianza associata al modello completo e quella associata al modello senza un predittore si distribuisce come una variabile chi-quadrato con 1 grado di libertà. Per $\alpha = 0{,}05$, il valore critico del test è pari a 3,84. Per esempio, quando escludiamo l'istruzione dal nostro modello sul voto presidenziale otteniamo $-2\ln L = 718{,}2$. Poiché, come abbiamo già visto, $-2\ln L_1 = 712{,}1$, ne deriva che $G^2 = 718{,}2 - 712{,}1 = 6{,}1$; questo valore risulta significativo per $\alpha = 0{,}05$.

Una seconda misura, nota come **statistica della bontà di adattamento**, si basa sui residui del modello standardizzati. In sostanza, questa misura corrisponde alla somma dei residui al quadrato divisi per la loro deviazione standard stimata. Formalmente:

$$Z^2 = \sum_{i=1}^{N} \frac{(p_i - \hat{p}_i)^2}{(\hat{p}_i)(1 - \hat{p}_i)}$$

Minore è la differenza fra le probabilità predette dal modello e quelle osservate, migliore è l'adattamento del modello ai dati e, quindi, minore è il valore assunto dalla statistica Z^2. Quest'ultima si distribuisce come una variabile chi-quadrato con un numero di gradi di libertà approssimativamente uguale a $N - K - 1$. Nell'esempio del voto presidenziale, $Z^2 = 871{,}6$, con $gl = 880$. Questo valore ci porta a formulare le stesse conclusioni suggerite dal test del rapporto di verosimiglianza: almeno uno dei quattro coefficienti di regressione logistica è diverso da zero nella popolazione.

Una terza misura, chiamata **pseudo-R^2**, tiene conto del fatto che le distribuzioni chi-quadrato sono proporzionali alla dimensione del campione. Specificamente, tale statistica modifica il rapporto di verosimiglianza standardizzandolo per N e si configura come una misura della varianza spiegata dalle K variabili indipendenti. La sua formula è la seguente:

$$\text{pseudo} - R^2 = \frac{-2\ln L}{N + (-2\ln L)}$$

Questa formula non tiene conto dei gradi di libertà del modello; inoltre, per essa non è disponibile alcuna distribuzione campionaria. Pertanto, la statistica pseudo-R^2 non può essere sottoposta a test di significatività e dovrebbe essere vista solo come una misura descrittiva che esprime in modo approssimativo la proporzione di varianza della variabile dipendente «spiegata» dalle K variabili indipendenti incluse nel modello prescelto.

Nel nostro esempio sul voto presidenziale, pseudo-R^2 = (712,1)/(885 + 712,1) = 0,45; dunque, le quattro variabili indipendenti prescelte «spiegano» poco meno della metà della variazione osservata nella scelta di voto.

Infine, un'equazione di regressione logistica può essere utilizzata per classificare ogni osservazione secondo la sua categoria più probabile della variabile dipendente, data la particolare combinazione dei valori assunti dalle variabili indipendenti. Questi valori predetti possono essere poi confrontati con quelli osservati, utilizzando la percentuale di classificazioni corrette per valutare la misura in cui il modello «spiega» la variazione della variabile dipendente. Tornando ancora una volta al nostro esempio, l'equazione prescelta classifica correttamente 283 dei 354 elettori di Dukakis (80%) e 437 dei 531 elettori di Bush (82,3%), dando luogo a un'accuratezza di classificazione complessiva pari a 81,4%. Questo valore è meno rilevante di quanto può apparire a prima vista. Infatti, se assumessimo che tutti gli intervistati hanno votato per Bush otterremmo una predizione corretta nel 60% dei casi (pari alla percentuale osservata di voti per Bush). Comunque, la conoscenza dell'identificazione di partito, dell'orientamento politico, della razza e del livello di istruzione degli intervistati ci consente di ridurre gli errori di classificazione di oltre la metà, passando dal 40% al 18,6% di classificazioni scorrette.

5. L'analisi delle variabili dipendenti politomiche

La regressione logistica dicotomica rappresenta un caso speciale di un modello più generale per l'analisi delle variabili dipendenti di tipo discreto. È evidente che le variabili dicotomiche non esauriscono l'universo delle variabili discrete. Per fare solo un esempio, la condizione occupazionale degli individui potrebbe essere articolata nelle seguenti categorie: occupato a tempo pieno, occupato a tempo parziale, disoccupato, in cerca di prima occupazione, non forza di lavoro. Il modello di regressione logistica illustrato nel paragrafo precedente può essere facilmente esteso all'analisi delle variabili dipendenti politomiche, cioè di variabili discrete a M categorie ($M > 2$).

L'**equazione di regressione logistica politomica** può essere illustrata estendendo l'esempio precedente in modo tale da includere, fra le scelte di voto, anche l'astensione. Dunque, ai 531 elettori di Bush e ai