Analisi dei Dati

Regressione logistica

Domenico De Stefano

a.a. 2022/2023

Esempio: Elezioni presidenziali del 2016

Un'indagine campionaria per stabilire le caratteristiche dei votanti dei due candidati effettuata su un campione di 855 elettori americani ha restituito i seguenti risultati

- 531 (60%) si erano espressi a favore di Donald Trump e 354 (40%) a favore di Hilary Clinton
- Sul campione sono state rilevate altre variabili
  - X_1 denota l'identificazione di partito dell'intervistato (da 0 = democratico convinto a 6 =repubblicano convinto);
  - X_2 orientamento politico (da 1 = estrema sinistra a 7 = estrema destra);
  - X_3 variabile dicotomica che rappresentala etnia dell'intervistato (1: bianco, 0: altrimenti);
  - X_4 denota gli anni di istruzione formale (da 0 a 20).

Esempio: Elezioni presidenziali del 2016

Se si applica a questi dati un modello di regressione lineare in cui la variabile dipendente è il voto (Trump=1 e Clinton=0) otteniamo i seguenti risultati:

$$\hat{Y}_i = \underset{}{0,07} + \underset{(0,08)}{0,13}X_{1i} + \underset{(0,01)}{0,04}X_{2i} + \underset{(0,01)}{0,14}X_{3i} + \underset{(0,04)}{0,01}X_{4i} \quad \underset{(0,004)}{}$$

- 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 Trump (perchè è la modalità che assume valore =1)
- Ad esempio, poiché il coefficiente di regressione della variabile X_1, assume un valore pari a 0.13, possiamo affermare che ogni spostamento dell'identificazione di partito verso il polo repubblicano la probabilità di votare per Trump aumenta di 0.13 (13%);
- analogamente, i bianchi hanno il 14% di probabilità in più di votare per Trump, e così via.

Esempio: Elezioni presidenziali del 2016 I

Usando un modello di regressione lineare con una variabile dipendente di tipo dicotomico si violano due assunti fondamentali dell'analisi di regressione:

- gli errori non sono più distribuiti normalmente
  - La conseguenza di ciò è che sebbene le stime OLS dei parametri β, rimangano corrette (in media saranno uguali al parametro), esse cessano di essere le stime più efficienti (cioè quelle con la varianza campionaria più piccola). Dunque, i test di significatività (t-test) basati su queste stime e sui loro errori standard possono indurre il ricercatore a trarre conclusioni non valide, anche quando si analizzano campioni numerosi.
- I valori predetti dal modello possono essere privi di senso
  - Ad esempio se su un individuo si osservano valori estremi per tutte le variabili otteniamo una valore predetto dal modello di 1.27

Esempio: Elezioni presidenziali del 2016 II

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

$\Rightarrow$ questo individuo avrebbe una probabilità di votare Trump pari a 1.27! Pertanto il modello lineare in queste situazioni non è adeguato. Abbiamo bisogno di una valida alternativa.

Regressione logistica

- Si consideri una variabile risposta dicotomica Y. Essa assume valori 0 o 1
- Solitamente si parla, rispettivamente, di insuccesso e successo
- Ricordiamo che la media di una variabile dicotomica è pari alla proporzione di soggetti per i quali si osserva il successo, ossia $\pi = \sum 1/n$
- i modelli di regressione logistica stimano le proporzioni dei successi in una popolazione
- $P(Y = 1)$ rappresenta la probabilità di osservare un successo per ciascun soggetto della popolazione
- tuttavia anzichè usare la scala di probabilità (o frequenza relativa) si usa la trasformazione logistica della P(Y=1)
- inoltre per una Y dicotomica, si assume che la sua distribuzione sia una v.c. binomiale che ha media proprio uguale a $\pi = \sum 1/n$

*Alcune slides sono tratte da Nicola Tedesco - regressione logistica

Cos'è la trasformazione logistica? I

- Le % e le frequenze relative (o probabilità) non rappresentano gli unici modi per misurare una variabile dipendente dicotomica
- La trasformazione logistica è una valida alternativa e ha alcune proprietà interessanti utili nell'interpretazione di un modello di regressione
- l'unità di misura è il cosiddetto logit che si ottiene formando l'odds della probabilità di successo $P(Y = 1) = \pi$ rispetto al suo complementare $P(Y = 0) = 1 - \pi$ (tale rapporto sdi chiama odds) e calcolando il logaritmo naturale ln cioè il log base e di Eulero (per noi sarà sempre cosi anche se indicherò con il termine log) di tale rapporto.
- in breve il logit è il logaritmo naturale di un odds
  $logit_i = ln(\frac{\pi_i}{1-\pi_i})$

Cos'è la trasformazione logistica? II

- Il logit si distibuisce simmetricamente intorno a un valore centrale che si ha quando $\pi_i = 0.5$ (il suo complementare assume lo stesso valore $1 - \pi_i = 1 - 0.5 = 0.5$). Il logit (cioè il logaritmo naturale) di questo rapporto è $logit_i = ln(0.5/0.5) = ln(1) = 0$
- man mano che la P(Y=1) si allontana da 0.5 in una direzione o l'altra il logit si allontanano da 0 come mostra la seguente tabella

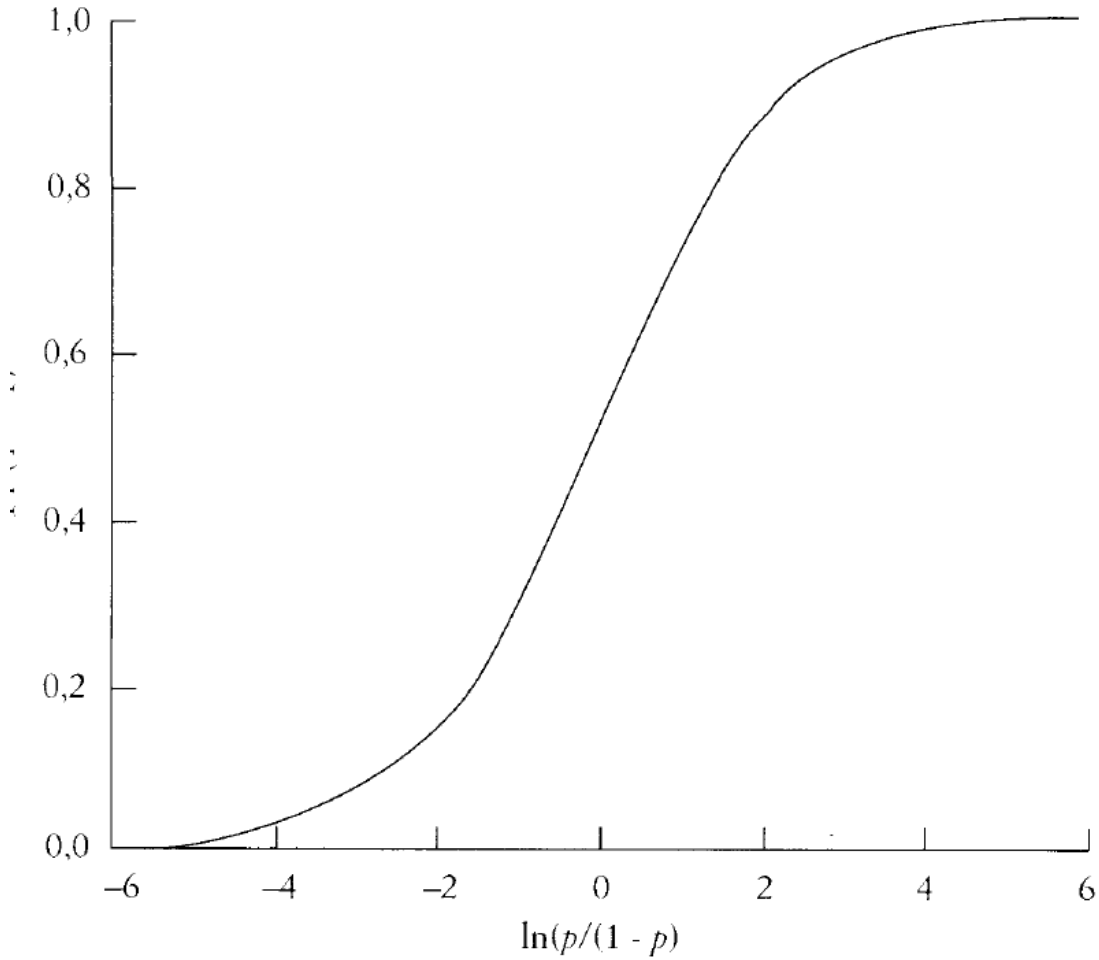
p_i	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90
$1 - p_i$	0,90	0,80	0,70	0,60	0,50	0,40	0,30	0,20	0,10
logit	–2,20	–1,39	–0,85	–0,41	0,00	0,41	0,85	1,39	2,20

- in virtù della sua natura simmetrica la trasformazione logistica di un odds $\pi/(1 - \pi)$ e del suo reciproco $(1 - \pi)/\pi$ da luogo a logit uguali in valore assoluto ma di segno opposto ad esempio $0.75/0.25 = 3$ e il suo logit a 1.099 per $0.25/0.75 = 0.3333$ e il suo logit è -1.099

Cos'è la trasformazione logistica? III

- Inoltre, sebbene non esista alcun limite inferiore o superiore per il logit, quando, è esattamente uguale a 1 o a 0 il logit risulta indefinito (ma questo equivale ad assenza di variabilità della Y!!!).

Cos'è la trasformazione logistica? IV



La distribuzione binomiale

- Per dati categoriali, possono verificarsi le seguenti condizioni:

1. Ciascuna osservazione cade in una di due categorie
2. Le probabilità per le due categorie sono le stesse per ciascuna osservazione. Indichiamo le probabilità con π per la categoria 1 e $(1 - \pi)$ per la categoria 2
3. I risultati di osservazioni successive sono indipendenti. Cioè, il risultato per una osservazione non dipende dal risultato delle altre osservazioni

- Un buon esempio è rappresentato dal lancio di una moneta: due possibili risultati (T o C), probabilità costante ad ogni lancio ($\pi = 0.50$) e il risultato ad ogni lancio è indipendente dai precedenti

La distribuzione binomiale

Probabilità per una Distribuzione Binomiale

Sia π la probabilità che un'osservazione assuma un valore della categoria 1. Nel caso di n osservazioni indipendenti, la probabilità di x successi per la categoria 1 è

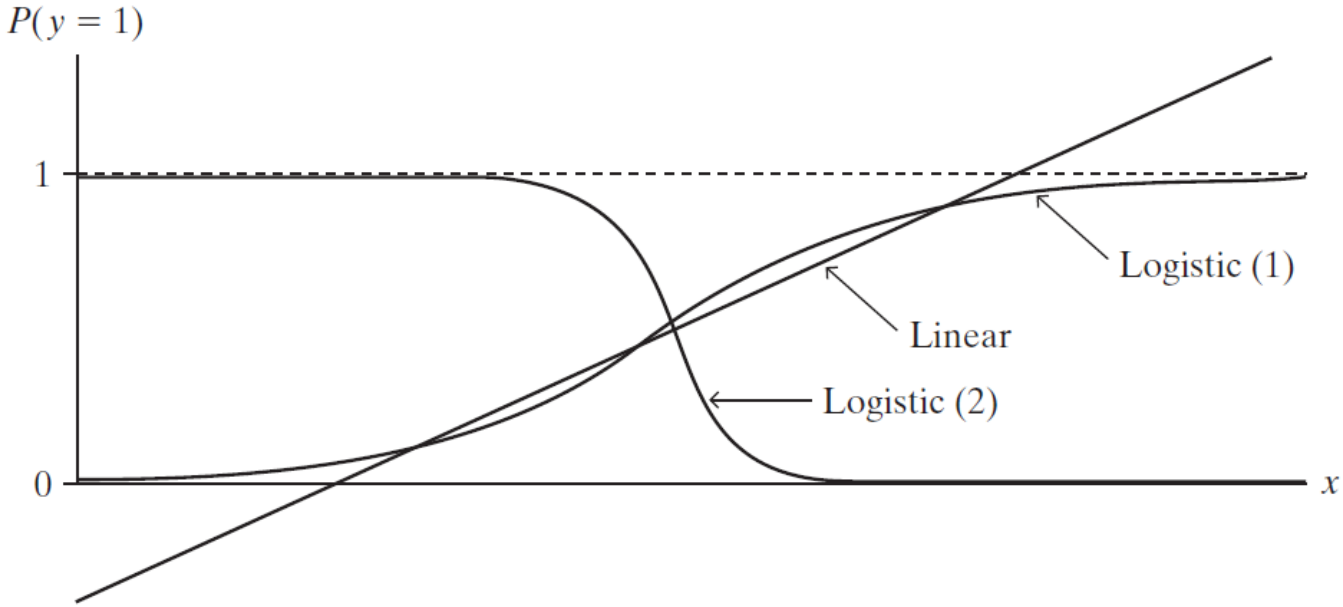
$$P(x) = \frac{n!}{x!(n-x)!}\pi^x(1-\pi)^{n-x}, \quad x = 0, 1, 2, \ldots, n.$$

Il simbolo $n!$ è chiamato **n fattoriale**. Rappresenta $n! = 1 \times 2 \times 3 \cdots \times n$. Ad esempio, $1! = 1$, $2! = 1 \times 2 = 2$, $3! = 1 \times 2 \times 3 = 6$, e così via. Per definizione, 0! è pari a 1.

Modello lineare di probabilità

- abbiamo discusso dell'esempio di un variabile dicotomica e dei predittori quantitativi e qualitativi
- Nel caso di un modello con una sola variabile esplicativa (continua) avremo
  - $P(Y = 1) = \alpha + \beta X$
- si suppone che la probabilit'a di successo sia in funzione lineare con X
- Per questa ragione prende il nome di Modello lineare di probabilità
- Abbiamo già detto che tale modello sia inadeguato per prevedere la $P(Y = 1)$

- Sappiamo che la $P(y = 1)$ è un valore dell'intevallo $[0, 1]$ e non eccede tale intervallo
- La figura mostra che il Modello a Probabilità Lineare non rispetta questa condizione



- Al contrario, la funzione logistica (casi 1 e 2), si dimostra adatta allo scopo
- A questo punto il problema è: come faccio a linearizzare una funzione logistica?

- In primo luogo scriviamo l'equazione di regressione di Y su X, utilizzando la funzione logistica

$$P(y = 1) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}.$$

  In seguito riprenderemo questa formulazione. Per ora ci basti osservare come la relazione tra Y e X non sia lineare, bensì logistica
- Per semplificare la notazione e ottenere risultati più semplici, linearizziamo questa espressione

Dal modello lineare a quello logistico I

- I logit costituiscono la base per costruire un'alternativa al modello lineare di probabilità (che aveva i problemi illustrati sopra).
- secondo quest'ultimo la probabilità puo essere espressa come funzione lineare $P(Y_i = 1) = \pi_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij}$
- se indichiamo con possiamo $Z = \sum_{j=1}^{k} \beta_j X_{ij}$ scrivere la seguente forma funzionale di una trasformazione logistica come:
  $F(Z) = \frac{e^Z}{1+e^Z} = \frac{1}{1+e^{-Z}}$
- facendo le opportune sostituzioni, la probabilità che l'osservazione i-ma assuma valore 1 in corrispondenza della variabile dipendente può essere calcolata come segue:
  $\pi_i = \frac{1}{1+e^{-\beta_0 - \sum \beta_j X_{ij}}}$
- Poiché il logit associato all'osservazione i corrisponde al logaritmo naturale dell'odds si ha:
  $logit_i = \frac{\pi_i}{1-\pi_i} = ln(e^{\beta_0 + \sum \beta_j X_{ij}}) = \beta_0 + \sum \beta_j X_{ij}$

dettagli derivazione modello regressione logistico 1

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:

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

dettagli derivazione modello regressione logistico 2

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:

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

Regressione logistica sintesi specificazioni

$$\underbrace{\text{Ln}\left[\pi/(1-\pi)\right]}_{\text{Logit}} = \overbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}^{\text{Predittore lineare}}$$

Var.qualitative
e/o quantitative

$$\underbrace{\pi/(1-\pi)}_{\text{odds}} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)$$

$$\pi = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}$$

Stima dei parametri I

- Il modello di regressione logistica è molto simile al modello di regressione lineare
- Tuttavia la stima dei parametri β_j non si ottiene con il metodo dei minimi quadrati
- bisogna ricorrere a un metodo noto come stima di massima verosimiglianza (MLE, dall'inglese Maximum Likelihood Estimation).
  - In breve, la stima MLE si basa su una serie di approssimazioni successive ai valori incogniti dei veri parametri della popolazione β_j
  - L'obiettivo è ottenere stime dei parametri che massimizzino la probabilità di ottenere i valori campionari osservati.
  - A differenza del metodo OLS, 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 MLE individua quell'insieme dei valori dei parametri per cui la probabilità di osservare i dati campionari è massimizzata

Bontà di adattamento I

- Un modo per valutare la bontà di adattamento di un modello di regressione logistica è usare il cosiddetto rapporto di verosimiglianza
- il rapporto di verosimiglianza pone a confronto due modelli di regressione logistica concatenati, una delle quali è una versione ridotta dell'altro (cioè con meno variabili indipendenti)
- Il modello piu ampio ha K_1 variabili indipendenti e quello meno ampio ne ha K_0 (ognuna delle quali inclusa nel modello piu ampio)
- l'ipotesi nulla afferma che nessuno di $k_1 - K_0$ coefficienti di regressione è significativamente diverso da 0 nella popolazione
- in sostanza si calcolano i valori della funzione di verosimiglianza di entrambi i modelli e si effettua il rapporto (LLR = loglikelihood ratio). Indichiamo con L_1 il valore di tale funzione per il modello piu ampio e L_0 per quello ristretto la formula della statistica test è la seguente:
  - $\text{LLR} = -2ln(\frac{L_0}{L_1}) = (-2ln(L_0)) - (-2ln(L_1))$

Bontà di adattamento II

- L'applicazione piu comune è che si confronti il modello con solo l'intercetta contro il modello con tutti i parametri (in modo tale da testare la loro nullità congiunta)
- Esempio:
  - Nell'esempio dei votanti Clinton/Trump il modello che include solo l'intercetta β_0 ($K_0 = 0$) è associato il valore $-2ln(L_0) = 1191.2$
  - Il modello che include tutte e 4 le variabili indipendenti ($K_1 = 4$) invece ha associato il valore della funzione di verosimiglianza è $-2ln(L_1) = 712.1$
  - pertanto: la statistica test è $1191.2 - 712.1 = 479.1$ che si distribuisce com una variabila casuale chi-quadro con $K_1 - K_0 = 4 - 0 = 4$ gradi di libertà. Il p-value associato al valore 479.1 è molto basso (il valore critico associato alla soglia 0.05 è 9.49
  - Pertanto possiamo tranquillamente rifiutare l'ipotesi nulla secondo la quale nella popolazione nessuna delle 4 variabili indipendenti prescelte esercita un effetto sulla variabile dipendente.

Altro esempio

- Si consideri un campione di $n = 100$ adulti selezionati casualmente in Italia. Si è rilevato il reddito annuale e se possedevano o meno una carta di credito
- La variabile risposta è dicotomica (possesso CC: $1 =$ Sì, $0 =$ No). Il predittore è quantitativo
- A ciascun livello di X, si può calcolare la probabilità di possedere una CC, attraverso il rapporto tra i soggetti che posseggono una CC e il totale soggetti per quel valore di X

Tabella: Reddito Annuale (in Migliaia di Euro) e Possesso di una Carta di Credito.

Income	Number Cases	Credit Cards	Income	Number Cases	Credit Cards	Income	Number Cases	Credit Cards
12	1	0	21	2	0	34	3	3
13	1	0	22	1	1	35	5	3
14	8	2	24	2	0	39	1	0
15	14	2	25	10	2	40	1	0
16	9	0	26	1	0	42	1	0
17	8	2	29	1	0	47	1	0
19	5	1	30	5	2	60	6	6
20	7	0	32	6	6	65	1	1

Altro esempio

- Il modello stimato è

$$\text{logit}[\hat{P}(y = 1)] = -3.518 + 0.105x.$$

- Il valore di $\beta = 0.105 > 0$, indica che al crescere del Reddito Annuale cresce la probabilità di possedere una CC
- Il software fornisce il seguente prospetto

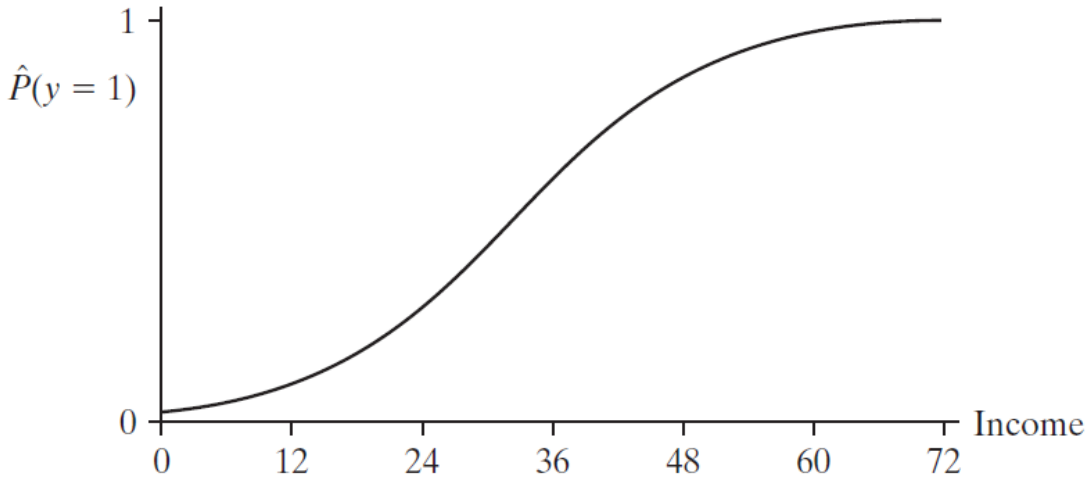
Tabella: Modello di Regressione Logistica sul Possesso della Carta di Credito in Italia

	B	S.E.	Exp(B)
reddito	.1054	.0262	1.111
costante	-3.5179	.7103	

- Il valore $exp(B) = exp(.1054) = 1.111$ consente di calcolare l'*odds ratio* (OR)

Altro esempio

- Il grafico mostra la funzione di previsione per il *logit*



- La probabilità di successo vale 0.50 per $x = -\hat{\alpha}/\hat{\beta} = (3.518)/(0.105) = 33.5$
- Cioè la probabilità stimata di possedere una CC è inferiore a 0.50 per redditi inferiori a 33.5 migliaia di euro, superiore a 0.50 per redditi superiori a quella soglia

Equazione di regressione per la probabilità

- Abbiamo visto a cosa corrisponde direttamente la $P(y = 1)$
- Si può, quindi, costruire un'equazione che stimi direttamente tale probabilità e non il suo *logit*, cioè

$$P(y = 1) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}. \qquad (1)$$

- In questa equazione la potenza di e rappresenta l'antilogaritmo
- Ricordiamo che il numero di Nepero e rappresenta la base del logaritmo naturale, cioè il numero a cui bisogna elevare la base per avere l'argomento. Ad es.,

$$log_e(1) = 0 \ \text{ in quanto } e^0 = 1$$

- Attraverso la formula (1) è possibile determinare la probabilità di successo per qualunque valore di x

Equazione di regressione per la probabilità

- Nel nostro caso, per un soggetto con Reddito Annuale $x = 12$ avremo

$$\hat{P}(y = 1) = \frac{e^{-3.52+0.105(12)}}{1 + e^{-3.52+0.105(12)}} = \frac{e^{-2.26}}{1 + e^{-2.26}} = \frac{0.104}{1.104} = 0.094.$$

- La probabilità stimata di possedere una CC è 0.094
- Per redditi pari a $x = 40$ e $x = 65$ avremo, rispettivamente

$$\hat{P}(y = 1) = \frac{e^{-3.52+0.105(40)}}{1 + e^{-3.52+0.105(40)}} = \frac{e^{0.68}}{1 + e^{0.68}} = \frac{1.974}{2.974} = 0.664.$$

$$\hat{P}(y = 1) = \frac{e^{-3.52+0.105(65)}}{1 + e^{-3.52+0.105(65)}} = \frac{e^{3.30}}{1 + e^{3.30}} = \frac{27.249}{28.249} = 0.970.$$

- In pratica chi ha un reddito di $x = 40$ ha una probabilità di possedere una CC pari a 0.664
- Per chi ha un reddito pari a $x = 65$, tale probabilità è 0.970

Interpretazione del modello di regressione logistica

- L'interpretazione di β non è semplice. Presenteremo due approcci distinti
- Abbiamo visto come il suo segno ci segnala una crescita ($\beta > 0$) o un decremento $\beta < 0$ nella $P(Y = 1)$, all'aumentare di x
- Tuttavia non siamo in grado di dire di quanto cresce/descresce $P(Y = 1)$
- Tale difficoltà deriva dalla particolare forma a S della funzione logistica

Interpretazione del modello di regressione logistica

- La soluzione che consente di interpretare adeguatamente β prevede di utilizzare l'*odds ratio*
- Applichiamo l'antilogaritmo ad ambo i membri dell'equazione

$$log\ \frac{P(y=1)}{1-P(y=1)} = \alpha + \beta x$$

- Otteniamo

$$\frac{P(y=1)}{1-P(y=1)} = e^{\alpha+\beta x} = e^{\alpha}(e^{\beta})^{x}.$$

- Si osserva chiaramente come il termine e^{β} indichi di quanto si modifichi l'*odds* (il rapporto tra le probabilità) in corrispondenza ad un incremento unitario in x

Interpretazione del modello di regressione logistica

- Sui dati dell'esempio sulle CC si ottiene $e^{\hat{\beta}} = e^{0.105} = 1.11$
- Cioè, all'incremento di 1000 euro di reddito, l'*odds* cresce di un fattore moltiplicativo pari a 1.11; in pratica cresce dell'11%
- L'*odds* per $x = 25$ è

$$\text{Odds Stimato} = \frac{\hat{P}(y=1)}{1-\hat{P}(y=1)} = e^{-3.518+0.105(25)} = 0.414$$

- Considerando un incremento unitario in X, per $x = 25$ si ha

$$\text{Odds Stimato} = \frac{\hat{P}(y=1)}{1-\hat{P}(y=1)} = e^{-3.518+0.105(26)} = 0.460,$$

- In pratica $0.460/0.414 = 1.11 = e^{0.105}$

Interpretazione del modello di regressione logistica

- La quantità $e^{0.105}$ non è altro che l'*Odds Ratio* (Rapporto tra Quote), dato appunto dal rapporto tra l'*odds* avendo un reddito $= 26$, diviso l'*odds* avendo un reddito $= 25$
- L'utilizzo dell'*Odds Ratio* (OR) risolve i problemi di rappresentazione e interpretazione di β
- Infatti, può variare tra 0 e 1 (se l'odds a numeratore è inferiore di quello a denominatore), oppure eccedere il valore 1 indefinitamente. Cioè

$$0 \leq OR \leq +\infty$$

- Supponiamo di confrontare due individui con, rispettivamente, reddito pari a $x = 20$ e $x = 30$
- Sarà sufficiente calcolare

$$e^{10\beta} = (e^{\beta})^{10} = (1.11)^{10} = 2.9$$

- L'*odds* per $x = 30$ è 2.9 volte l'*odds* per $x = 20$

Modello di regressione logistica multipla

- Naturalmente è possibile estendere il *Modello di Regressione Logistica* a k predittori

$$\text{logit}[P(y = 1)] = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

- Per calcolare la probabilità di successo avremo

$$P(y = 1) = \frac{e^{\alpha+\beta_1 x_1+\cdots+\beta_k x_k}}{1 + e^{\alpha+\beta_1 x_1+\cdots+\beta_k x_k}}.$$

- Ciascuna stima di β rappresenta l'effetto di quel predittore sul logit, controllando per le altre variabili
- Oppure indicherà l'effetto moltiplicativo di quel predittore sulla probabilità di successo, controllando per le altre variabili
- Quanto più $\beta > 0$, tanto più l'$OR > 1$ e rappresenterà un effetto più forte

Esempio

- Si tratta di un celebre caso-studio, molto noto in bibliografia. Si vuole determinare se la probabilità di essere condannato a morte dipenda (o meno) dalla razza dell'accusato e/o da quella della vittima

Tabella: Verdetti di Pena di Morte secondo la Razza dell'Imputato e della Vittima, nei Casi di Omicidi Plurimi in Florida

Razza imputato	Razza vittima	Pena di Morte Sì	Pena di Morte No	% Sì
Bianca	Bianca	53	414	11.3
	Nera	0	16	0.0
Nera	Bianca	11	37	22.9
	Nera	4	139	2.8

- La variabile risposta (Y) è la condanna (Pena di Morte Sì-No), i predittori sono la Razza della Vittima e quella dell'Imputato
- I predittori sono qualitativi categorici (Bianca-Nera)

Esempio

- Nell'ultima colonna è riportata la percentuale di imputati per ogni combinazione, che è stata condannata a morte
- Abbiamo i seguenti casi:

1. in caso di imputato bianco e vittima bianca, l'11.3% è stato condannato a morte;
2. in caso di imputato bianco e vittima nera, nessuno è stato condannato a morte;
3. in caso di imputato nero e vittima bianca, il 22.9% è stato condannato a morte;
4. in caso di imputato nero e vittima nera, il 2.8% è stato condannato a morte.

- In caso di vittima è bianca

a. la probabilità per un imputato bianco di essere condannato a morte è l'11.3% superiore rispetto al caso di una vittima nera
b. la probabilità per un imputato nero di essere condannato a morte è il 20.1% superiore rispetto al caso di una vittima nera (22.9% – 2.8%)

Esempio

- In buona sostanza, controllando per la razza dell'imputato, risulta evidente che la probabilità di essere condannato a morte è decisamente superiore se la vittima è bianca
- L'analisi con il controllo secondo la razza della vittima, evidenzia chiaramente come se la vittima è bianca, la probabilità di essere condannati a morte è superiore
- Tale probabilità è decisamente più elevata per un imputato nero (22.9 %) rispetto ad uno bianco (11.3 %)
- Quando la vittima è nera, le due probabilità sono quasi uguali (2,8% per un imputato nero, 0.0% per un imputato bianco), anche se il confronto merita un approfondimento
- Infatti, in termini relativi 2.8% rispetto a 0.0% può essere ritenuto molto più elevato

Esempio

- Costruiamo il modello. Ovviamente Y = Condanna a Morte. Se $Y = 1$ la risposta è Sì
- I due predittori sono dicotomici, per cui sono sufficienti 2 variabili dummy, d per la razza dell'imputato e v per la razza della vittima

$$d = 1, \text{ defendant} = \text{white}; d = 0, \text{defendant} = \text{black},$$

$$v = 1, \text{victims} = \text{white}; \quad v = 0, \text{victims} = \text{black}.$$

- Il modello di regressione logistica multivariato sarà

$$\text{logit}[P(y = 1)] = \alpha + \beta_1 d + \beta_2 v,$$

dove β_1 indica l'effetto della razza dell'imputato, controllando per quella della vittima e β_2 indica l'effetto della razza della vittima, controllando per quella dell'imputato

- La quantità e^{β_1} è l'OR tra Y e la razza dell'imputato, controllando per quella della vittima
- La quantità e^{β_2} è l'OR tra Y e la razza della vittima, controllando per quella dell'imputato

Esempio

- Il modello applicato ai dati è

$$\text{logit}[\hat{P}(y = 1)] = -3.596 - 0.868d + 2.404v.$$

- La stima di $\beta_1 = -0.868$ indica come essendo bianco ($d = 1$), si ha una probabilità di essere condannato a morte inferiore rispetto ad un nero ($d = 0$)
- La stima di $\beta_2 = 2.404$ indica come in caso di vittima bianca ($v = 1$), si ha una probabilità di essere condannato a morte superiore rispetto al caso di una vittima nera ($v = 0$)
- Vedremo tra poco di quantificare l'intensità di questi effetti

Esempio

Tabella: Stime dei Parametri del Modello Logistico sui Dati della Pena di Morte

	B	Std Error	Exp(β)
Intercetta	-3.596	.5069	.027
imputato=bianco	-.868	.3671	.420
imputato=nero	0	.	
vittima=bianca	2.404	.6006	11.072
vittima=nera	0	.	

- La stima della probabilità di essere condannati a morte è

$$\hat{P}(y=1) = \frac{e^{-3.596-0.868d+2.404v}}{1+e^{-3.596-0.868d+2.404v}}$$

- Nel caso di imputato nero ($d=0$) e vittima bianca ($v=1$) si ha

$$\hat{P}(y=1) = \frac{e^{-3.596-0.868(0)+2.404(1)}}{1+e^{-3.596-0.868(0)+2.404(1)}} = \frac{e^{-1.192}}{1+e^{-1.192}} = \frac{0.304}{1.304} = 0.233.$$

Esempio

- In pratica, se un imputato nero è riconosciuto colpevole di aver ucciso un bianco, ha il 23.3% di possibilità di essere condannato a morte
- Questa stima è molto simile al valore osservato nel campione, pari a 22.9%
- Proviamo a calcolare l'OR tra Y e razza dell'imputato. Si ottiene

$$e^{\hat{\beta}_1} = e^{-0.868} = 0.42 \quad OR = 0.42$$

- Un imputato bianco ha un odds pari a 0.42 volte quello di un imputato nero di essere condannato a morte
- Curiosità: se nel costruire la dummy per la razza dell'imputato invertissimo le categorie ($d = 1$ per la razza nera) otterremmo

$$e^{\hat{\beta}_1} = e^{0.868} = 2.38 \quad OR = 2.38$$

- Lavorando direttamente sull'OR, basta fare il reciproco del primo per ottenere il secondo $1/0.42 = 2.38$

Esempio

- Studiando la relazione tra Y e razza della vittima, la stima ottenuta mostra un fortissimo effetto del predittore su Y
- Infatti, $e^{2.404} = 11.1$, quindi OR=11.1
- Ciò sisgnifica che l'odds di essere condannato a morte in caso di vittima bianca è 11.1 volte quello nel caso di vittima nera
- In termini più semplicistici, potremmo sintetizzare i due OR ottenuti in questo modo:

1. Si registra una propensione (tendenza) a condannare a morte inferiore alla metà se l'imputato è bianco rispetto ad uno nero;
2. Si registra una propensione (tendenza) a condannare a morte pari a circa 11 volte se la vittima è bianca rispetto ad una nera.

- In buona sostanza esiste una discriminizione razziale nelle condanne a morte, più forte in relazione alla razza della vittima rispetto a quella dell'imputato, pur sempre presente