

METODI STATISTICI PER LA RICERCA SOCIALE

CAPITOLO 15. REGRESSIONE LOGISTICA: MODELLARE VARIABILI RISPOSTA CATEGORIALI

Alessandra Mattei

Dipartimento di Statistica, Informatica, Applicazioni (DiSIA)
Università degli Studi di Firenze
mattei@disia.unifi.it

LM-88 SOCIOLOGIA E RICERCA SOCIALE

Modelli per risposte binarie

Variabile risposta binaria

$$Y_i = \begin{cases} 1 & \text{successo} \\ 0 & \text{insuccesso} \end{cases}$$

- Pratica dell'autopalpazione
- Decisioni: Si/No
- Presenza/assenza di una certa caratteristica (Occupato/Non occupato)
- Sopravvivenza/Morte

Variabile esplicativa: X_i

Parametro di interesse = Probabilità di successo condizionata

$$Pr(Y_i = 1|X_i = x_i)$$

Esempio: Pozzi di acqua potabile in Bangladesh

- Obiettivo: studiare la decisione delle famiglie in Bangladesh di cambiare la loro fonte di acqua da bere
- Campione di 3020 famiglie
- Variabili

x_{i1} = Distanza (in ettometri) dal pozzo di acqua potabile più vicino

x_{i2} = Livello di arsenico nel pozzo di acqua potabile utilizzato dalla famiglia

y_i = Indicatore della decisione della famiglia

$$Y_i = \begin{cases} 1 & \text{La famiglia } i \text{ cambia fonte di acqua potabile} \\ 0 & \text{Altrimenti} \end{cases}$$

Esempio: Pozzi di acqua potabile in Bangladesh

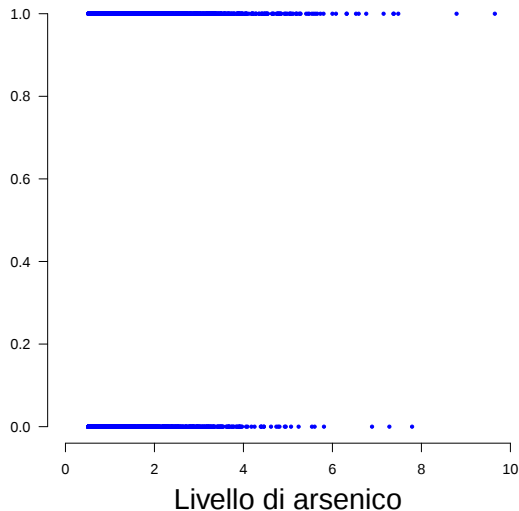
u_i	y_i	x_{i1}	x_{i2}
1	1	2.36	0.16826
2	1	0.71	0.47322
3	0	2.07	0.20967
4	1	1.15	0.21486
5	1	1.10	0.40874
$\vdots$	$\vdots$	$\vdots$	$\vdots$
3017	0	1.08	0.21386
3018	0	0.51	0.07708
3019	0	0.64	0.22842
3020	1	0.66	0.20844

Statistiche descrittive

Frequenze	$Y_i = 0$	$Y_i = 1$
Absolute	1283	1737
Relative	0.4248	0.5752

Variabile	Media
x_{i1}	0.4833
x_{i2}	1.6570
y_i	0.5752

Pozzi di acqua potabile in Bangladesh: Relazione tra decisione e livello di arsenico



Modello lineare nelle probabilità

- Le osservazioni Y_i sono variabili aleatorie indipendenti e hanno distribuzione Bernoulliana

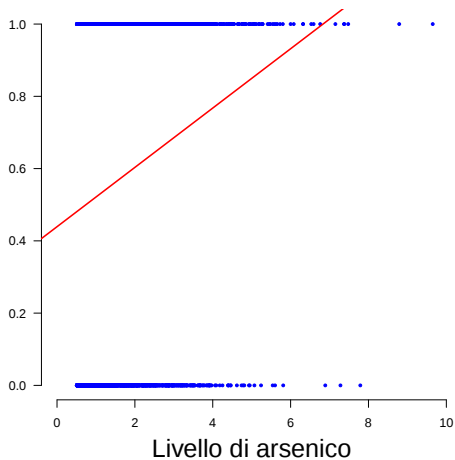
$$Y_i | X_i = x_i \sim \text{Bernoulli}(\pi_i)$$

con probabilità di successo $\pi_i = \text{Pr}(Y_i = 1 | X_i = x_i)$ che dipendono da x_i

- Modello lineare nelle probabilità

$$\pi_i = \text{Pr}(Y_i = 1 | X_i = x_i) = \mathbb{E}[Y_i | X_i = x_i] = \beta_0 + \beta_1 x_i$$

Modello lineare nelle probabilità



$$\hat{\beta}_0 + \hat{\beta}_1 x_i = 0.44 + 0.08 x_i$$

Modello lineare nelle probabilità

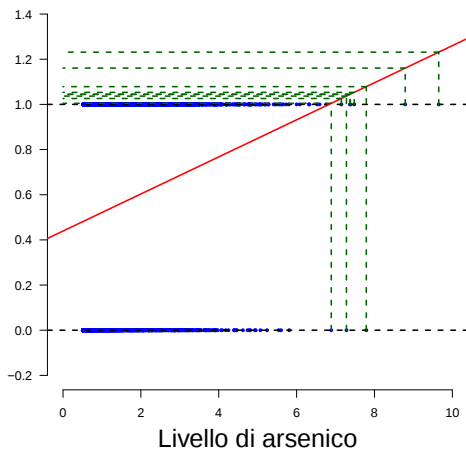
- Il modello lineare nelle probabilità può dar luogo a probabilità stimate non comprese tra 0 e 1
- Le varianze

$$\text{var}(Y_i|X_i = x_i) = \pi_i(1 - \pi_i)$$

non sono costanti, ma funzione delle variabili esplicative x

- Il modello di regressione multipla con assunzione di omoschedasticità non è correttamente specificato
- Le stime dei minimi quadrati sono inefficienti

Modello lineare nelle probabilità: Pozzi di acqua potabile in Bangladesh



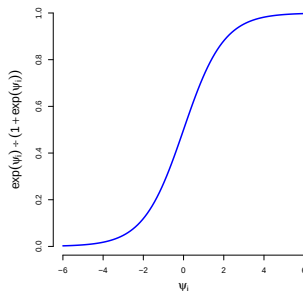
$$\hat{\beta}_0 + \hat{\beta}_1 x_i = 0.44 + 0.08 x_i$$

Il logit delle probabilità

$$\psi_i = \text{logit}(Pr(Y_i = 1|X_i = x_i)) = \log\left(\frac{Pr(Y_i = 1|X_i = x_i)}{1 - Pr(Y_i = 1|X_i = x_i)}\right)$$

- Il *logit* assume valori nell'insieme dei numeri reali
- Trasformazione inversa

$$\pi_i = Pr(Y_i = 1|X_i = x_i) = \frac{\exp(\psi_i)}{1 + \exp(\psi_i)}$$



Modello di regressione logistica lineare

- Le osservazioni Y_i sono variabili aleatorie indipendenti e hanno distribuzione Bernoulliana

$$Y_i|X_i = x_i \sim \text{Bernoulli}(\pi_i)$$

con probabilità di successo $\pi_i = P(Y_i = 1|X_i = x_i)$ che dipendono da x_i

- Modello logit

$$\text{logit}(\pi_i) = \log\left(\frac{\Pr(Y_i = 1|X_i = x_i)}{1 - \Pr(Y_i = 1|X_i = x_i)}\right) = \beta_0 + \beta_1 x_i$$

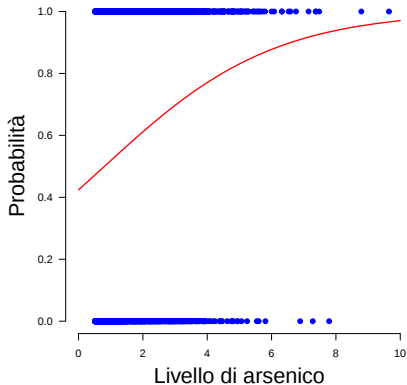
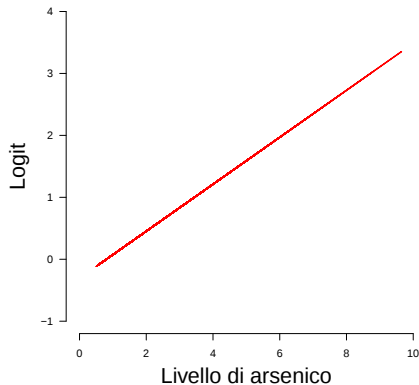


$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

Modello di regressione logistica lineare

Pozzi di acqua potabile in Bangladesh

$$\hat{\pi}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)} = \frac{\exp(-0.31 + 0.4x_i)}{1 + \exp(-0.31 + 0.38x_i)}$$



Modello di regressione logistica lineare

Funzione di regressione logistica

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

- $\beta_0 \mapsto \hat{\pi}(x = 0) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$
- È crescente se $\beta_1 > 0 \implies$ Probabilità di successo aumenta all'aumentare di x
- È decrescente se $\beta_1 < 0 \implies$ Probabilità di successo diminuisce all'aumentare di x
- È costante se $\beta_1 = 0 \implies$ Probabilità di successo costante
- Pendenza della retta tangente alla curva: $\beta_1 \pi_i (1 - \pi_i)$
- La massima pendenza della curva si ha nel punto $x = -\beta_0/\beta_1$ nel quale $\pi_i = 0.5$
- La pendenza della curva nel punto $x = -\beta_0/\beta_1$ è $\beta_1/4$: Un incremento unitario in x implica una variazione nella probabilità di successo di al massimo $\beta_1/4$

Modello di regressione logistica lineare: Il parametro β_1

$$\text{logit}(\pi(x)) = \beta_0 + \beta_1 x \implies \beta_1 = \text{logit}(\pi(x+1)) - \text{logit}(\pi(x))$$

per qualsiasi valore di x

- Si ricorda che

$$\text{logit}(\pi(x)) = \log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \log(\text{Odds}(x))$$

- Quindi

$$\begin{aligned}\beta_1 &= \text{logit}(\pi(x+1)) - \text{logit}(\pi(x)) = \log\left(\frac{\pi(x+1)}{1 - \pi(x+1)}\right) - \log\left(\frac{\pi(x)}{1 - \pi(x)}\right) \\ &= \log\left(\frac{\frac{\pi(x+1)}{1 - \pi(x+1)}}{\frac{\pi(x)}{1 - \pi(x)}}\right) = \log\left(\frac{\text{Odds}(x+1)}{\text{Odds}(x)}\right)\end{aligned}$$

- $e^{\beta_1} = \frac{\text{Odds}(x+1)}{\text{Odds}(x)}$: passando da x a $x+1$ gli odds sono moltiplicati per e^{β_1}
per qualsiasi valore di x

Modello di regressione logistica lineare

Pozzi di acqua potabile in Bangladesh

$$\hat{\pi}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)} = \frac{\exp(-0.31 + 0.38x_i)}{1 + \exp(-0.31 + 0.38x_i)}$$

- $\hat{\beta}_0 = -0.31 \mapsto \hat{\pi}(x = 0) = \frac{\exp(-0.31)}{1 + \exp(-0.31)} = 0.42$
- La funzione cresce ($\hat{\beta}_1 = 0.38$ è positivo) e le quote di successo crescono a un tasso di $e^{0.4} = 1.49$ ossia ogni incremento unitario nel livello di arsenico implica che gli odds siano moltiplicati per 1.49
- La massima pendenza della curva, uguale a

$$\hat{\beta}_1/4 = 0.38/4 = 0.09$$

si ha nel punto

$$x = -\hat{\beta}_0/\hat{\beta}_1 = -(-0.31)/(0.38) = 0.81$$

Modello di regressione logistica lineare

Pozzi di acqua potabile in Bangladesh

$$\hat{\pi}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)} = \frac{\exp(-0.31 + 0.38 x_i)}{1 + \exp(-0.31 + 0.38 x_i)}$$

$$\hat{\pi}(x = 0.75) = 0.49 \quad \hat{\pi}(x = 1.75) = 0.59 \quad \implies \quad \hat{\pi}(x = 1.75) - \hat{\pi}(x = 0.75) = -0.10$$

$$\hat{\pi}(x = 5) = 0.83 \quad \hat{\pi}(x = 6) = 0.88 \quad \implies \quad \hat{\pi}(x = 6) - \hat{\pi}(x = 5) = -0.05$$

- Incrementi non costanti
- Un incremento nella concentrazione di arsenico da 0.75 a 1.75 implica un incremento nella probabilità di cambiare fonte di acqua potabile da 0.49 a 0.59
- Un incremento nella concentrazione di arsenico da 5 a 6 implica un incremento nella probabilità di cambiare fonte di acqua potabile da 0.83 a 0.88
- Probabilità attesa per una concentrazione media di arsenico: $\bar{x} = 1.66$

$$\hat{\pi}(\bar{x}) = \frac{\exp(-0.31 + 0.38 \cdot 1.66)}{1 + \exp(-0.31 + 0.38 \cdot 1.66)} = 0.58$$

Stimare i parametri di un modello logistico

- Metodo della massima verosimiglianza
- La **funzione di verosimiglianza** da informazioni sulla “probabilità” di osservare i dati effettivamente osservati al variare del valore dei parametri: Ogni valore della funzione di verosimiglianza in corrispondenza di un dato valore dei parametri dice quanto è verosimile che i dati osservati provengano da una popolazione in cui i parametri assumono quel valore
- Determinare i valori di β_0 e β_1 che rendono massima la funzione di verosimiglianza (o equivalentemente la funzione di log-verosimiglianza)
- Le stime di massima verosimiglianza per i parametri del modello logit sono ottenute numericamente attraverso procedure iterative
- In generale non esistono espressioni in forma chiusa delle stime di massima verosimiglianza per i parametri del modello logit
- Funzione di verosimiglianza: $\mathcal{L}(\beta_0, \beta_1)$
- $\widehat{B}_0$ e $\widehat{B}_1$ =: stimatori di massima verosimiglianza dei parametri β_0 e β_1 , rispettivamente

Inferenza nel modello logistico: Verifica di ipotesi e IC per β_1

- Ipotesi: $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$
- Statistica test

$$Z = \frac{\widehat{B}_1}{\widehat{e.s.}(\widehat{B}_1)}$$

dove $\widehat{e.s.}(\widehat{B}_1)$ è la stima dell'errore standard di $\widehat{B}_1$

- Sotto l'ipotesi nulla la statistica test Z ha una distribuzione asintotica Normale standard: se il campione non è estremamente piccolo, si può approssimare la distribuzione della statistica test Z con la distribuzione normale
- Regione critica asintotica

$$RC_\alpha \approx Z \leq -z_{\alpha/2} \quad \text{oppure} \quad Z \geq z_{\alpha/2}$$

- Statistica test di Wald

$$W = \left(\frac{\widehat{B}_1}{\widehat{e.s.}(\widehat{B}_1)} \right)^2 = Z^2$$

Sotto l'ipotesi nulla H_0 la statistica test W ha una distribuzione asintotica χ_1^2

- Il test Z e di Wald sono equivalenti (stesso p -value)
- Intervallo di confidenza per $\exp(\beta_1)$ al livello di confidenza $1 - \alpha$

$$IC_{1-\alpha}(\exp(\beta_1)) = \left[\exp \left\{ \widehat{\beta}_1 - z_{\alpha/2} \cdot \widehat{e.s.}(\widehat{B}_1) \right\}; \exp \left\{ \widehat{\beta}_1 + z_{\alpha/2} \cdot \widehat{e.s.}(\widehat{B}_1) \right\} \right]$$

Stima di massima verosimiglianza nel modello logistico

Pozzi di acqua potabile in Bangladesh

<i>Variabile</i>	<i>Stima</i>	<i>Errore standard</i>	<i>z – value</i>	<i>p – value</i>
Costante	-0.31	0.0703	-4.3	0.0000
Arsenico	0.38	0.0385	9.8	0.0000

- La statistica test $z = 0.38/0.0385 = 9.8$ con $p - value = 0.0000$

$$RC_{\alpha=0.01} = Z \leq -2.58 \quad oppure \quad Z \geq 2.58$$

Evidenza contro l'ipotesi nulla al livello di significatività del 1%.

- Statistica di Wald: $w = (0.38/0.0385)^2 = (9.8)^2 = 96.04$ con $p - value = Pr(\chi_1^2 \geq 96.04) = 0.000$
- Intervallo di confidenza per $\exp(\beta_1)$ al livello di confidenza $1 - \alpha = 0.99$

$$\begin{aligned} IC_{0.99}(\exp(\beta_1)) &= [\exp(0.38 - 2.58 \cdot 0.0385); \exp(0.38 + 2.58 \cdot 0.0385)] \\ &= [\exp(0.28); \exp(0.48)] = [1.32; 1.61] \end{aligned}$$

Confronto tra modelli

Modello esteso: $\text{logit}(\pi_i) = \beta_0 + \beta_1 x_i \iff \pi_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$

Modello ridotto:
(modello nullo) $\text{logit}(\pi_i) = \beta_0 \iff \pi_i = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$

Test del rapporto di verosimiglianza

$$\begin{aligned} -2 \log \left(\frac{\max_{\beta_0} \mathcal{L}(\beta_0)}{\max_{\beta_0, \beta_1} \mathcal{L}(\beta_0, \beta_1)} \right) &= -2 \log \left(\frac{\mathcal{L}(\widehat{\beta}_0)}{\mathcal{L}(\widehat{\beta}_0, \widehat{\beta}_1)} \right) = \\ &= -2 \log \left(\frac{\mathcal{L}_r}{\mathcal{L}_e} \right) = -2 (\log \mathcal{L}_r - \log \mathcal{L}_e) \end{aligned}$$

Confronto tra modelli

- La statistica test del rapporto di verosimiglianza ha una distribuzione asintotica (sotto H_0) χ^2 con gdl uguale al numero di parametri posti a zero sotto l'ipotesi nulla (1 nel caso considerato)
- In presenza di una sola variabile esplicativa il test del rapporto di verosimiglianza per il confronto tra modello esteso e modello nullo usualmente conduce a risultati simili al test di Wald (e al test Z) per campioni di grandi dimensioni

Confronto tra modelli

Pozzi di acqua potabile in Bangladesh

- Modello completo: $\log \mathcal{L}(\widehat{\beta}_0, \widehat{\beta}_1) = -2004.337$

<i>Variabile</i>	<i>Stima</i>	<i>Errore standard</i>	<i>z - value</i>	<i>p - value</i>
Costante	-0.31	0.0703	-4.3	0.0000
Arsenico	0.38	0.0385	9.8	0.0000

- Modello nullo $\log \mathcal{L}(\widehat{\beta}_0) = -2059.05$

<i>Variabile</i>	<i>Stima</i>	<i>Errore standard</i>	<i>z - value</i>	<i>p - value</i>
Costante	0.30	0.0368	8.2	0.0000

- Statistica test del rapporto di verosimiglianza

$$-2 (\log \mathcal{L}_r - \log \mathcal{L}_e) = -2 \cdot (-2059.05 - (-2004.337)) = 109.426$$

- Fissato $\alpha = 0.01$, $\chi_1^2(0.01) = 6.635 \implies$ Evidenza contro l'ipotesi nulla al livello di significatività del 1%
- $p - value = Pr(\chi_1^2 > 109.426) = 0.000$

Modello di regressione logistica multiplo

- Le osservazioni sulle variabili esplicative $x_{i1}, \dots, x_{ik}$ sono fisse
- Le osservazioni Y_i sono variabili aleatorie indipendenti e hanno distribuzione Bernoulliana

$$Y_i | X_{i1} = x_{i1}, \dots, X_{ik} = x_{ik} \sim \text{Bernoulli}(\pi_i)$$

con probabilità di successo $\pi_i = P(Y_i = 1 | X_{i1} = x_{i1}, \dots, X_{ik} = x_{ik})$ che dipendono da $x_{i1}, \dots, x_{ik}$

- Modello logit

$$\begin{aligned} \text{logit}(\pi_i) &= \text{logit} \left(\frac{\Pr(Y_i = 1 | X_{i1} = x_{i1}, \dots, X_{ik} = x_{ik})}{1 - \Pr(Y_i = 1 | X_{i1} = x_{i1}, \dots, X_{ik} = x_{ik})} \right) \\ &= \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} \end{aligned}$$



$$\pi_i = \Pr(Y_i = 1 | X_{i1} = x_{i1}, \dots, X_{ik} = x_{ik}) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik})}$$

Esempio: Pozzi di acqua potabile in Bangladesh

<i>Variable</i>	<i>Stima</i>	<i>Errore standard</i>	<i>z - value</i>	<i>p - value</i>
Costante	0.0027	0.0794	0.03	0.9724
Distanza	-0.8966	0.1043	-8.59	0.0000
Arsenico	0.4608	0.0414	11.13	0.0000

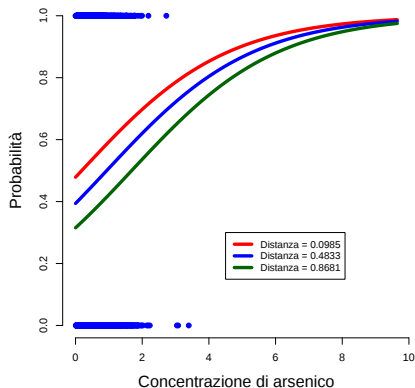
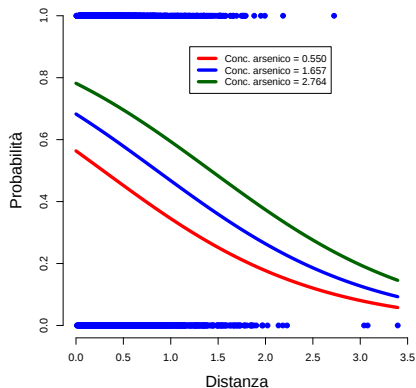
- $\hat{\beta}_0 = 0.0027 \mapsto \hat{\pi}(x_1 = 0, x_2 = 0) = \frac{\exp(0.0027)}{1 + \exp(0.0027)} = 0.5007$
- Confrontando famiglie che utilizzano pozzi di acqua caratterizzati dallo stesso livello di arsenico, ogni incremento di 100 metri nella distanza determina una riduzione di 0.90 nel *logit* della probabilità di cambiare fonte di acqua potabile $\iff$ gli odds di successo a favore del cambiamento si moltiplicano per $e^{-0.8966} = 0.4079$ per ogni incremento di 100 metri della distanza fissato il livello di arsenico
- Confrontando famiglie che si trovano alla stessa distanza dal pozzo di acqua utilizzato con fonte di acqua potabile, un incremento unitario nella concentrazione di arsenico determina un incremento di 0.46 nel *logit* della probabilità di cambiare fonte di acqua potabile $\iff$ gli odds di successo a favore del cambiamento si moltiplicano per $e^{0.46} = 1.5853$ per ogni incremento unitario della concentrazione di arsenico fissata la distanza

Esempio: Pozzi di acqua potabile in Bangladesh

<i>Variabile</i>	<i>Stima</i>	<i>Errore standard</i>	<i>z – value</i>	<i>p – value</i>
Costante	0.0027	0.0794	0.03	0.9724
Distanza	-0.8966	0.1043	-8.59	0.0000
Arsenico	0.4608	0.0414	11.13	0.0000

- Il segno dei coefficienti delle variabili esplicative sembrano ragionevoli: una famiglia ha maggiore probabilità di cambiare fonte di acqua potabile se si trova lontana dall'attuale fonte e se l'attuale fonte ha un'elevata concentrazione di arsenico
- Entrambi i coefficienti delle variabili esplicative sono statisticamente significativi

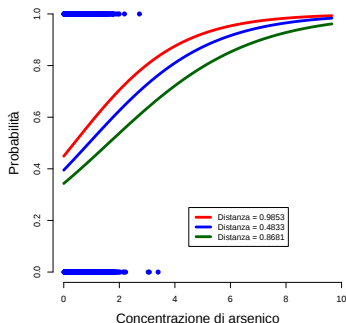
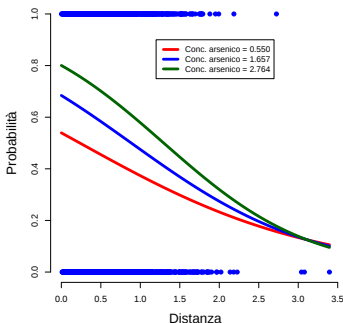
Esempio: Pozzi di acqua potabile in Bangladesh



Esempio: Pozzi di acqua potabile in Bangladesh

Modello con interazione

<i>Variable</i>	<i>Stima</i>	<i>Error standard</i>	<i>z - value</i>	<i>p - value</i>
Costante	-0.1479	0.1175	-1.26	0.2084
Distanza	-0.5772	0.2092	-2.76	0.0058
Arsenico	0.5560	0.0693	8.02	0.0000
Interazione	-0.1789	0.1023	-1.75	0.0804



Esempio: Pozzi di acqua potabile in Bangladesh

Modello con interazione

- $\widehat{\beta}_0 = -0.1479 \mapsto \widehat{\pi}(x_1 = 0, x_2 = 0) = \frac{\exp(-0.1479)}{1 + \exp(-0.1479)} = 0.4631$
- Confrontando famiglie che utilizzano pozzi di acqua caratterizzati da una concentrazione nulla di arsenico, ogni incremento di 100 metri nella distanza determina una riduzione di 0.58 nel *logit* della probabilità di cambiare fonte di acqua potabile $\iff$ gli odds di successo a favore del cambiamento si moltiplicano per $e^{-0.5772} = 0.5615$ per ogni incremento di 100 metri della distanza fissato il livello di arsenico a zero
- Confrontando famiglie che si trovano distanza zero dal pozzo di acqua utilizzato con fonte di acqua potabile, un incremento unitario nella concentrazione di arsenico determina un incremento di 0.556 nel *logit* della probabilità di cambiare fonte di acqua potabile $\iff$ gli odds di successo a favore del cambiamento si moltiplicano per $e^{0.556} = 1.7436$ per ogni incremento unitario concentrazione di arsenico fissata la distanza a zero
- Interpretazione del coefficiente di interazione: Per ogni incremento unitario nella distanza (concentrazione di arsenico) un valore di -0.1789 è aggiunto al coefficiente della concentrazione di arsenico (distanza)

Esempio: Pozzi di acqua potabile in Bangladesh

Modello con interazione (variabili centrate rispetto alla media)

<i>Variabile</i>	<i>Stima</i>	<i>Errore standard</i>	<i>z – value</i>	<i>p – value</i>
Costante	0.3511	0.0399	8.81	0.0000
Distanza – 0.48	–0.8736	0.1048	–8.34	0.0000
Arsenico – 1.66	0.4695	0.0421	11.16	0.0000
Interazione	–0.1789	0.1023	–1.75	0.0804

Confronto tra modelli

- Modello esteso

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_h x_{ih} + \cdots + \beta_k x_{ik}$$

- Modello ridotto

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_h x_{ih}$$

- Test del rapporto di verosimiglianza

$$-2 \log \left(\frac{\mathcal{L}_r}{\mathcal{L}_e} \right) = -2 (\log \mathcal{L}_r - \log \mathcal{L}_e)$$

- La statistica test del rapporto di verosimiglianza ha una distribuzione asintotica (sotto H_0) χ^2 con gdl uguale al numero di parametri posti a zero sotto l'ipotesi nulla ($k - h$)

Confronto tra modelli

Pozzi di acqua potabile in Bangladesh

Modello	$\log \mathcal{L}$	$-2 (\log \mathcal{L}_r - \log \mathcal{L}_e)$	GdL	$p - value$
Modello con interazione (Modello esteso)	-1963.814			
Modello additivo: $H_0 : \beta_{Interazione} = 0$	-1965.334	3.040	1	0.08124
Solo Distanza: $H_0 : \beta_{Interazione} = 0$ e $\beta_A = 0$	-2038.119	148.610	2	0.00000
Solo Arsenico: $H_0 : \beta_{Interazione} = 0$ e $\beta_D = 0$	-2004.337	81.045	2	0.00000
Modello nullo: $H_0 : \beta_{Interazione} = \beta_D = \beta_A = 0$	-2059.050	190.470	3	0.00000

Confronto tra modelli

Pozzi di acqua potabile in Bangladesh

Modello	$\log \mathcal{L}$	$-2(\log \mathcal{L}_r - \log \mathcal{L}_e)$	GdL	$p - value$
Modello additivo (Modello esteso)	-1965.334			
Solo Distanza: $H_0 : \beta_A = 0$	-2038.119	145.570	1	0.00000
Solo Arsenico: $H_0 : \beta_D = 0$	-2004.337	78.005	1	0.00000
Modello nullo: $H_0 : \beta_D = \beta_A = 0$	-2059.050	187.431	2	0.00000

Information Criteria

Akaike (1973) Information Criteria (AIC)

$$AIC = -2 \log \mathcal{L} + 2 \cdot (k + 1)$$

dove $k + 1$ è il numero di parametri (inclusa l'intercetta)

- L'indice AIC si basa su un criterio di bilanciamento tra bontà di adattamento del modello e semplicità del modello (in termini di numero di variabili esplicative considerate)
- Il modello con l'indice AIC più basso è preferibile
- Può essere utilizzato per confrontare modelli non-nested

Bayes (1973) Information Criteria (BIC)

$$BIC = -2 \log \mathcal{L} + (k + 1) \cdot \log(n)$$

- Rispetto all'indice AIC, l'indice BIC attribuisce una maggiore penalità a modelli con un numero maggiore di parametri per $n > 8$ ($\log(8) > 2$)

Information Criteria

Pozzi di acqua potabile in Bangladesh

Modello	<i>AIC</i>	<i>BIC</i>
Modello con interazione	3935.63	3951.67
Modello additivo	3936.67	3943.65
Solo Distanza	4080.24	4016.69
Solo Arsenico	4012.67	4084.25
Modello nullo	4010.67	4076.24

Il modello di regressione logistica multinomiale

- Il modello di regressione logistica può essere esteso a situazioni in cui la variabile risposta assume più di due valori
- Esempio: Studio condotto da “Florida Game and Fresh Water Fish Commission” finalizzato a investigare i fattori che influenzano le scelte nutrizionale primaria degli alligatori
- Variabile risposta: Scelta nutrizionale primaria di alligatori determinata in base al contenuto trovato nello stomaco dell'alligatore

$$Y_i = \begin{cases} 1 & \text{Invertebrati} \\ 2 & \text{Pesce} \\ 3 & \text{Altro} \end{cases}$$

dove Invertebrati = Chioccioline, insetti acquatici, gamberi; Altro = rettili; anfibi, mammiferi, piante, sassi, detriti vari

- La variabile risposta è una variabile categorica con 3 modalità tra cui non esiste un ordine naturale
- Variabile esplicativa = Lunghezza dell'alligatore (in metri)
- Campione di 59 alligatori del Lago George in Florida (USA)

Scelta nutrizionale primaria di alligatori

x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
1.24	1	3.71	2	3.28	3	1.42	1
1.45	1	1.30	1	1.32	2	1.55	1
1.63	1	1.47	1	1.50	1	1.70	1
1.78	1	1.65	1	1.65	2	1.88	1
1.98	1	1.78	3	1.80	2	2.31	2
2.36	2	2.03	2	2.26	2	2.67	2
2.79	2	2.41	2	2.46	2	3.58	2
3.68	3	3.25	3	3.33	2	1.42	2
1.30	1	3.89	2	1.40	2	1.60	1
1.45	3	1.32	2	1.52	1	1.73	3
1.65	3	1.47	2	1.68	2	1.93	1
1.78	1	1.65	2	1.85	2	2.36	2
2.03	2	1.80	1	2.31	2	2.72	1
2.39	2	2.16	2	2.56	3	3.66	2
2.84	2	2.44	2	3.56	2		

Il modello di regressione logistica multinomiale

- Le osservazioni Y_i sono variabili aleatorie indipendenti categoriali con modalità codificate come $\{1, 2, \dots, J\}$
- Una modalità $m \in \{1, 2, \dots, J\}$ è considerata come categoria di riferimento (la scelta è arbitraria: spesso $m = 1$ o $m = J$.)
- Considerando la categoria m come categoria di riferimento si ha

$$\pi_{ij} = Pr(Y_i = j | X_{i1} = x_{i1}, \dots, X_{iK} = x_{iK})$$

per $j \neq m$ con $\pi_{im} = 1 - \sum_{j \neq m} \pi_{ij}$

Il modello di regressione logistica multinomiale

- Modello di regressione logistica multinomiale

$$\begin{aligned}\log\left(\frac{\pi_{ij}}{\pi_{im}}\right) &= \log\left(\frac{Pr(Y_i = j|X_{i1} = x_{i1}, \dots, X_{iK} = x_{iK})}{Pr(Y_i = m|X_{i1} = x_{i1}, \dots, X_{iK} = x_{iK})}\right) \\ &= \beta_{0j} + \beta_{1j}x_{i1} + \dots + \beta_{Kj}x_{iK}\end{aligned}$$

$i = 1, \dots, n$, $m =$ categoria di riferimento e $j \neq m$ (Per $m = 1$, $j = 2, 3, \dots, J$, per $m = J$, $j = 1, 2, \dots, J - 1$).

- Probabilità nel modello di regressione logistica multinomiale

$$\begin{aligned}\pi_{ij} &= Pr(Y_i = j|X_{i1} = x_{i1}, \dots, X_{iK} = x_{iK}) \\ &= \frac{\exp(\beta_{0j} + \beta_{1j}x_{i1} + \dots + \beta_{Kj}x_{iK})}{1 + \sum_{j \neq m} \exp(\beta_{0j} + \beta_{1j}x_{i1} + \dots + \beta_{Kj}x_{iK})}\end{aligned}$$

- I parametri del modello sono stimati utilizzando il metodo della massima verosimiglianza

Esempio: Scelta nutrizionale primaria di alligatori

- Modello di regressione logistica multinomiale: Per $i = 1, \dots, n = 59$

$$\log\left(\frac{\pi_{i1}}{\pi_{i3}}\right) = \log\left(\frac{Pr(Y_i = 1|X_{i1} = x_{i1})}{Pr(Y_i = 3|X_{i1} = x_{i1})}\right) = \beta_{0j=1} + \beta_{1j=1}x_{i1}$$

$$\log\left(\frac{\pi_{i2}}{\pi_{i3}}\right) = \log\left(\frac{Pr(Y_i = 2|X_{i1} = x_{i1})}{Pr(Y_i = 3|X_{i1} = x_{i1})}\right) = \beta_{0j=2} + \beta_{1j=2}x_{i1}$$

- Probabilità nel modello di regressione logistica multinomiale

$$\pi_{ij} = Pr(Y_i = j|X_{i1} = x_{i1}) = \frac{\exp(\beta_{0j} + \beta_{1j}x_{i1})}{1 + \exp(\beta_{0j=1} + \beta_{1j=1}x_{i1}) + \exp(\beta_{0j=2} + \beta_{1j=2}x_{i1})}$$

Esempio: Scelta nutrizionale primaria di alligatori

- Stime di massima verosimiglianza dei parametri

Logit Invertebrati vs Altro		
Variabile	Stima	E.S.
Costante	5.698	1.794
Lunghezza	-2.465	0.900

Logit Pesce vs Altro		
Variabile	Stima	E.S.
Costante	1.618	1.307
Lunghezza	-0.110	0.517

$$\log \mathcal{L} = -49.171$$

- Dalle stime dei parametri si ha :

$$\log \left(\frac{\hat{\pi}_{i1}}{\hat{\pi}_{i3}} \right) = 5.698 - 2.465 \cdot x_{i1} \qquad \log \left(\frac{\hat{\pi}_{i2}}{\hat{\pi}_{i3}} \right) = 1.618 - 0.110 \cdot x_{i1}$$

- Lo stima che il log degli odds a favore del pesce rispetto agli invertebrati è

$$\begin{aligned} \log \left(\frac{\hat{\pi}_{i2}}{\hat{\pi}_{i3}} \right) - \log \left(\frac{\hat{\pi}_{i1}}{\hat{\pi}_{i3}} \right) &= \log \left(\frac{\hat{\pi}_{i2}}{\hat{\pi}_{i1}} \right) = \\ (1.618 - 5.698) + (-0.110 - (-2.465)) \cdot x_{i1} &= -4.080 + 2.355 \cdot x_{i1} \end{aligned}$$

Esempio - Scelta nutrizionale primaria di alligatori: Interpretazione

- Per ogni logit, le stime dei parametri si interpretano come modelli di regressione logistica per variabili risposta binarie, focalizzando l'attenzione sulle due categorie coinvolte
- Esempio: Scelta nutrizionale primaria: pesce versus invertebrati

$$\log\left(\frac{\hat{\pi}_{i2}}{\hat{\pi}_{i1}}\right) = -4.080 + 2.355 \cdot x_{i1}$$

- ✓ Dato che la scelta nutrizionale primaria è pesce o invertebrati, la probabilità stimata che la scelta sia pesce cresce con la lunghezza dell'alligatore secondo una forma a S
- ✓ Per gli alligatori di lunghezza $x + 1$ metri gli odds stimati che la scelta nutrizionale primaria sia pesce piuttosto che invertebrati è $e^{2.355} = 10.54$ volte gli odds stimati per alligatori di lunghezza x

Esempio - Scelta nutrizionale primaria di alligatori: Probabilità

- Le probabilità stimate relative alla variabile risposta (Invertebrati, Pesce, Altro) sono

$$\hat{\pi}_{i1} = \frac{\exp(5.698 - 2.465 \cdot x_{i1})}{1 + \exp(5.698 - 2.465 \cdot x_{i1}) + \exp(1.618 - 0.110 \cdot x_{i1})}$$

$$\hat{\pi}_{i2} = \frac{\exp(1.618 - 0.110 \cdot x_{i1})}{1 + \exp(5.698 - 2.465 \cdot x_{i1}) + \exp(1.618 - 0.110 \cdot x_{i1})}$$

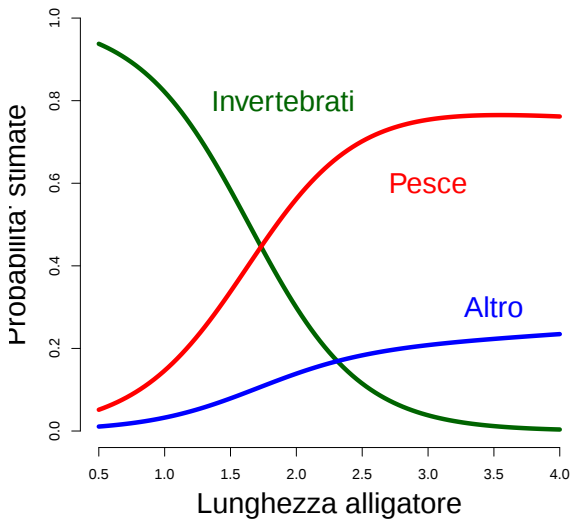
$$\hat{\pi}_{i3} = \frac{1}{1 + \exp(5.698 - 2.465 \cdot x_{i1}) + \exp(1.618 - 0.110 \cdot x_{i1})}$$

Il termine 1 in ogni denominatore e al numeratore della terza equazione rappresenta $\exp(\hat{\beta}_{0j=3} + \hat{\beta}_{1j=3} \cdot x_{i1})$ posto $\hat{\beta}_{0j=3} = \hat{\beta}_{1j=3} = 0$

- Per un alligatore di lunghezza $x^* = 3.9$ metri, la probabilità stimata che la scelta nutrizionale primaria sia “Altro” è

$$\hat{\pi}_3 = \frac{1}{1 + \exp(5.698 - 2.465 \cdot 3.9) + \exp(1.618 - 0.110 \cdot 3.9)} = 0.23$$

Esempio - Scelta nutrizionale primaria di alligatori: Probabilità



Esempio - Scelta nutrizionale primaria di alligatori

Confronto tra modello esteso e modello nullo: La scelta nutrizionale primaria è indipendente dalla lunghezza

$$H_0 : \beta_{11} = 0 \text{ e } \beta_{12} = 0 \quad \text{vs} \quad H_a : \beta_{11} \neq 0 \text{ o } \beta_{12} \neq 0$$

- Test del rapporto di verosimiglianza

$$-2(\log \mathcal{L}_0 - \log \mathcal{L}) = 16.80 \quad gdl = 2$$

- Regione critica ($\alpha = 0.01$)

$$RC_{0.01} : -2(\log \mathcal{L}_0 - \log \mathcal{L}) > \chi_2^2(0.01) = 9.21$$

- Il valore osservato della statistica test del rapporto di verosimiglianza cade nella regione critica: Evidenza contro l'ipotesi nulla
- $p\text{-value} < 0.00022$