Statistica sociale - 14

Prof. Antonio Mussino

a. a. 2022-2023

SAPIENZA
UNIVERSITÀ DI ROMA

La Statistica multivariata: approccio inferenziale

La Regressione logistica - 1

- Si tratta di un particolare tipo di Regressione multipla, in cui la variabile dipendente è ***dicotomica*** e quelle indipendenti sono ***qualitative*** e ***quantitative***.

La Regressione logistica - 2

- Il modello della Regressione multipla è un'estensione di quello della Regressione bivariata.

- In questo caso la variabile dipendente (Y) è espressa in funzione di più variabili indipendenti, chiamate predittori (X_1, X_2, …X_i …, X_p) e si deve tener conto sia della relazione fra la Y e le singole X_i, sia delle interrelazioni fra di esse.

La Regressione logistica - 3

- L'equazione di riferimento nella popolazione è una funzione lineare che rappresenta un iperpiano:

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... \beta_i X_i + ... \beta_p X_p}$$

- di cui vanno stimati i parametri β_i, utilizzando i coefficienti (b_0, b_1, ….b_p), calcolati nel campione.

La Regressione logistica - 4

- Nel caso in cui la Y sia dicotomica, le premesse necessarie per testare ipotesi nel modello di Regressione multipla non più sono valide:
- pertanto la funzione precedente viene sostituita da una logistica, di cui vanno calcolati i coefficienti nel campione, da usare come stime dei parametri della popolazione o per testare ipotesi su di essi.

- La relazione sarà

$$Y = 1 / (1 + e^{-(\alpha + \beta x + \gamma z + \ldots.)}).$$

La Regressione logistica - 5

- Mentre la Y è dicotomica, le X_i possono essere quantitative o qualitative, ma in questo caso le modalità sono codificate come variabili binarie.
- Poiché la relazione fra la variabile dipendente e i predittori è espressa da una funzione logistica, essa assume valori con un range fra 0 e 1, quindi può essere assimilata a una probabilità:
- si può quindi pensare ad una misura del *rischio* (da 0 nessuno a 1 certezza) di passare dallo stato codificato con "0" a quello codificato con "1".

Il caso di studio - 1

- Come caso di studio si riprenderà in considerazione il file dell'indagine sulla partecipazione sportiva dei cittadini di Aracaju.
- La variabile dipendente, di cui vogliamo individuare le determinanti, è la pratica di almeno un'attività, sia essa sportiva, fisica o motoria vs. la non pratica, ovvero la sedentarietà;
- essa può così assumere due modalità: "0", che useremo per caratterizzare i sedentari e "1" per gli attivi.

Il caso di studio - 2

- Le determinanti, i *predittori*, che vogliamo mettere in relazione con la pratica sono:
  - l'età,
  - il livello di istruzione,
  - l'etnia (caratterizzata con il colore della pelle),
  - il genere e
  - il Body Mass Index (BMI2), variabile quantitativa i cui valori sono riaggregati in funzione dei livelli normalizzati di peso; le categorie aggregate sono: grave magrezza o sottopeso, regolare, sovrappeso, obeso.

Il caso di studio - 3

- Quindi il *rischio*, di cui sopra, è quello di passare da sedentari ad attivi: più è alto il coefficiente che misurerà la relazione (che vedremo poi), più è alta la probabilità che chi ha quella caratteristica (ad esempio "15-19 anni", "femmina", "bianco" e così via) sia attivo.
- Le informazioni sulle unità e sulle variabili utilizzate per l'analisi del modello di Regressione logistica sono presentate nella tabella che segue.

Tabella – Caratterizzazione delle variabili presenti nel modello di Regressione logistica.

Riepilogo dell'elaborazione dei casi			
Casi		N	Percentuale
Casi selezionati	Inclusi nell'analisi	1137	100,0
	Casi mancanti	0	,0
	Totale	1137	100,0
Totale		1137	100,0

Codifica variabile dipendente			
Valore originale	Valore interno		
Sedentari	0		
Attivi	1		

Codifiche variabili categoriali							
		Frequenza	Codifica del parametro				
			(1)	(2)	(3)	(4)	(5)
Età in classi	15-19 anni	135	1,000	,000	,000	,000	,000
	20-24 anni	171	,000	1,000	,000	,000	,000
	25-34 anni	304	,000	,000	1,000	,000	,000
	35-44 anni	229	,000	,000	,000	1,000	,000
	45-54 anni	175	,000	,000	,000	,000	1,000
	55-64 anni	123	,000	,000	,000	,000	,000
Scolarità ricodificata.	Inferiore	103	1,000	,000	,000	,000	
	Medio	147	,000	1,000	,000	,000	
	Superiore	474	,000	,000	1,000	,000	
	Universitario	213	,000	,000	,000	1,000	
	Post laurea	200	,000	,000	,000	,000	
BMI2	Sottopeso/grave magrezza	42	1,000	,000	,000		
	Regolare	607	,000	1,000	,000		
	Sovrappeso	389	,000	,000	1,000		
	Obeso	99	,000	,000	,000		
Colore della pelle	Bianca	237	1,000	,000	,000		
	Gialla	113	,000	1,000	,000		
	Marrone	564	,000	,000	1,000		
	Nera	223	,000	,000	,000		
Genere	Maschi	544	1,000				
	Femmine	593	,000				

Il caso di studio - 4

- Come si può notare, ogni variabile qualitativa con p modalità genera p-1 variabili dicotomiche;
- ad esempio il Colore della pelle ha quattro modalità (bianca, gialla, marrone e nera) e genera tre variabili dicotomiche: se in queste la codifica è 1,0,0 si individua la modalità “bianca”; se è 0,1,0 “gialla”; se 0,0,1 “marrone”; se, infine, si ha 0,0,0 allora si individua la “nera”.

Il caso di studio - 5

- I modelli di regressione, in genere, procedono con una sequenza *stepwise*, ovvero si inserisce o si elimina una variabile per volta e si continua aggiungendone o eliminandone un'altra, fino ad ottimizzare un criterio di capacità di predizione:
- questo può avvenire partendo da una per arrivare a tutte (*forward*), o partendo dal modello completo e eliminando le variabili non adeguatamente predittive (*backward*).

Il caso di studio - 6

- Il risultato finale può essere valutato con diversi test di significatività (p-value), ma in questo caso c'è anche la possibilità di analizzare la capacità predittiva del modello, utilizzando una tabella di contingenza.
- Questa tabella viene generata ricalcolando la probabilità di ogni unità di essere assegnata agli attivi o ai sedentari, avendo come valore di riferimento 0,50: attivi se >0,50; sedentari se <0,50.

Tabella di classificazione delle unità rispetto al rischio di essere "attivi"[a].

Osservato			Previsto		
			Attivi vs. sedentari		Percentuale corretta
			Sedentari	Attivi	
Passo 1	Attivi vs. sedentari	Sedentari	609	76	88,9
		Attivi	366	86	19,0
	Percentuale globale				61,1

a. Il valore di riferimento è ,500

Il caso di studio - 7

- L'assegnazione fatta dal modello (status previsto) viene poi confrontata con lo status osservato dell'unità:
  - si può vedere che nell'88,9% dei casi (609 volte su 685) un sedentario è stato classificato correttamente,
  - mentre nell'11,1% dei casi ciò non è avvenuto;
  - solo nel 19,0% dei casi (86 volte su 452) un attivo è stato classificato correttamente;
  - complessivamente il modello ha classificato correttamente solo nel 61,1% dei casi.
- Evidentemente le variabili utilizzate non sono complessivamente determinanti per predire uno stile di vita attiva.

Il caso di studio - 8

- Sempre complessivamente il modello è, però, significativo:
- ovvero le relazioni individuate sui dati campionari permettono di rifiutare l'ipotesi nulla di nessuna relazione fra la variabile dipendente e i predittori (p-value= 0,0000);
- in sintesi questa relazione c'è, anche se debole e relativa solo ad alcuni predittori.

- Nell'output dell'SPSS sono proposti due coefficienti tipo R^2: quello di Cox e Snell (0,069) e quello di Nagelkerte (0,092).

Il caso di studio - 9

- Quello che ci interessa maggiormente è di vedere, però, quali modalità siano significativamente predittive e quali siano i coefficienti che le mettono in relazione con la Y.
- L'output proposto (v.tabella) è piuttosto ricco di informazioni, che cerchiamo di analizzare nel dettaglio.

Variabili nell'equazione di Regressione logistica: 1^ modello.

		B	E.S.	Wald	Df	Sig.	Exp(B)
	Età ricodificata			**10,605**	**5**	**,060**	
	15-19 anni	,421	,261	2,613	1	,106	1,524
	20-24 anni	,349	,246	2,007	1	,157	1,417
	25-34 anni	-,047	,205	,052	1	,820	,954
	35-44 anni	,063	,207	,092	1	,761	1,065
	45-54 anni	-,259	,227	1,310	1	,252	,772
	Maschi	**,158**	**,124**	**1,630**	**1**	**,202**	**1,171**
	Colore pelle			**2,974**	**3**	**,396**	
	Bianca	-,172	,183	,878	1	,349	,842
	Gialla	-,377	,239	2,493	1	,114	,686
Passo 1[a]	Marrone	-,062	,152	,164	1	,686	,940
	Scolarità			**12,511**	**4**	**,014**	
	Inferiore	-,697	,253	7,570	1	,006	,498
	Medio	-,571	,226	6,409	1	,011	,565
	Superiore	-,191	,164	1,358	1	,244	,826
	Universitario	-,047	,207	,051	1	,822	,954
	BMI2			**2,860**	**3**	**,414**	
	Sottopeso/grave magrezza	-,118	,372	,100	1	,752	,889
	Regolare	-,221	,192	1,326	1	,249	,801
	Sovrappeso	-,326	,197	2,743	1	,098	,722

a. Variabili immesse al passo 1: Età ricodificata, Sesso, Colore della pelle, Scolarità, BMI2 (BMI ricodificato).

Il caso di studio - 10

- L'informazione più interessante è posta nell'ultima colonna: si tratta di *odds ratio*, utili a valutare l'impatto delle modalità predittive in modo più facile da comunicare.
- L'odds ratio di un evento che chiamiamo "favorevole" è il rapporto della probabilità che si verifichi su quella che non si verifichi, ma si verifichi il caso "contrario":
- in caso di equiprobabilità il rapporto è 1; un odds ratio minore di 1 indica una probabilità di successo minore di quella di fallimento e viceversa.

Il caso di studio - 11

- Prendiamo il caso dell'età: l'odds ratio della modalità 55-64 anni, che non appare nell'output, è convenzionalmente posto uguale ad 1 (exp(0)=1);
- sensibilmente più alto è quello della modalità 15-19 anni, ovvero 1,524;
- rozzamente potremmo dire che, se abbiamo 100 attivi tra coloro che hanno 55-64 anni, tra gli adolescenti ce ne sono 152;
- al contrario tra chi ha tra 45 e 54 anni ce ne sono 77;
- ogni 100 attive tra le donne ci sono 117 attivi fra gli uomini e così via.

Il caso di studio - 12

- Il coefficiente B, assimilabile al classico coefficiente di regressione, indica quanto le modalità delle variabili indipendenti contribuiscono, nel calcolo della funzione logaritmica, alla predizione della variabile dipendente;
- per ogni variabile sono presenti ovviamente p-1 modalità, l'ultima modalità ha un coefficiente convenzionalmente posto a 0.

Il caso di studio - 13

- Questi coefficienti sono calcolati con i dati del campione;
- la colonna E.S. ci dà l'errore standard della stima dei corrispondenti coefficienti β nella popolazione di riferimento;
- le colonne successive servono a testare l'ipotesi nulla di indipendenza della variabile dipendente dalla variabile indipendente complessivamente presa e dalle sue singole modalità:
  - il test statistico (Wald), i gradi di libertà (df) e il p-value (Sig.);
  - l'ipotesi nulla è rifiutata se Sig.<0,05, o <0,01 a seconda del livello di accuratezza desiderato.

Il caso di studio - 14

- Le relazioni proposte nella prima tavola sono facilmente leggibili;
- in realtà solo poche sono significative, ovvero permettono di rifiutare l'ipotesi nulla di indipendenza della variabile sedentarietà/attività dalla modalità predittiva: questo dipende fondamentalmente dalla ridotta dimensione del campione.
- In particolare risalta la relazione fra stile di vita attiva e scolarità, dove il caso più favorevole all'attività è quello di chi ha un diploma post laurea (1), poi universitario (0,954) e così via fino al minimo (0,498) per chi ha un titolo a livello inferiore.

Il caso di studio - 15

- Il modello proposto nella prima tavola è molto complesso e introduce relazioni con variabili che sembrano non essere determinanti per la opzione sedentario vs. attivo (colore della pelle, Body Mass Index);
- pertanto si è proposto un secondo modello (tabella che segue) con sole tre variabili, per le quali risultano, ovviamente, le stesse relazioni, ma i livelli dei p-value sono più bassi e, pertanto, le relazioni significative aumentano di numero.

Variabili nell'equazione di Regressione logistica: 2^ modello.

		B	E.S.	Wald	df	Sig.	Exp(B)
Passo 1[a]	**Maschi**	**,106**	**,119**	**,789**	**1**	**,374**	**1,111**
	Età ricodificata			**15,198**	**5**	**,010**	
	15-19 anni	,236	,224	1,108	1	,293	1,266
	20-24 anni	,136	,211	,413	1	,521	1,145
	25-34 anni	-,293	,163	3,242	1	,072	,746
	35-44 anni	-,163	,172	,900	1	,343	,849
	45-54 anni	-,468	,200	5,473	1	,019	,626
	Scuola			**20,905**	**4**	**,000**	
	Inferiore	-,844	,237	12,663	1	,000	,430
	Medio	-,696	,215	10,504	1	,001	,499
	Superiore	-,306	,152	4,076	1	,043	,736
	Universitario	-,125	,200	,386	1	,534	,883

a. Variabili immesse al passo 1: Sesso, Età ricodificata, Scolarità.

Il caso di studio - 16

- È come se avessimo eliminato il *rumore di fondo* che rende più difficile enucleare le relazioni significative esistenti, individuando un modello più *parsimonioso* per analizzarle!