

Tecniche di indagine statistica

2 settembre 2019

1. Una banca desidera classificare i clienti che chiedono un prestito personale tra coloro che sono in grado di ripagare le rate ($Y=1$) e coloro che avranno dei ritardi o non saranno in grado di ripagare il debito ($Y=0$). A tal fine effettua un'analisi di regressione logistica su coloro che hanno ricevuto un prestito in passato. I risultati sono riportati nella tabella seguente:

Variabile	coeff.	s.e.	OR
Costante	-.5030	.6479	
Età (in anni)	.0423	.01603	1.5224
Abitazione (1= possesso)	-.4934	.2483	.6124
Cittadinanza (1= non italiana)	-1.3932	.5794	.2483
Stato civile (1= coniugato/a)	.3910	.1633	1.4785

- a. Sapendo che il valore della statistica G , basata sul rapporto di verosimiglianza, è pari a 12.59, valutare se complessivamente il modello può essere ritenuto ragionevole rispetto ai dati.
 - b. Interpretare i risultati. Per quali variabili si può ritenere una significatività pari all'1%?
 - c. Stimare, sulla base del modello, la probabilità di successo per un cliente straniero di 50 anni, coniugato e che possiede l'abitazione. Confrontarla con l'analoga di un cliente con le stesse caratteristiche ma celibe.
 - d. Dimostrare opportunamente quali assunzioni del modello di regressione $Y = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + e$ sono violate in questo caso in cui Y è una variabile dicotomica, X_1 è una variabile quantitativa e le altre X_2, X_3, X_4 sono variabili dicotomiche.
2. Un gruppo di 124 nazioni è stato suddiviso in 4 strati sulla base della popolazione di tre anni prima. Nei primi 3 strati sono stati estratti campioni indipendenti, rispettivamente di dimensione 6, 6, 11, mentre nell'ultimo sono state selezionate tutte le 7 nazioni dello strato, per una dimensione totale del campione stratificato $n = 30$. I dati osservati (in milioni) sono riportati in tabella:

Strato	Pop.ne 3 anni prima	N_h	n_h	$\sum_i y_i$	$\sum_i y_i^2$
1	0-9.99	70	6	31.1	234.95
2	10-29.99	29	6	119.6	2613.26
3	30-99.99	18	11	621.0	40122.36
4	100 e oltre	7	7	2671.4	1769052.90

- a. determinare rispettivamente la stima del totale di popolazione e la varianza in ogni strato e nelle 124 nazioni;
- b. costruire un intervallo di confidenza al 95% per il totale di popolazione delle 124 nazioni e commentare il risultato;
- c. quale sarebbero state le dimensioni dei campioni per strato nel caso di una allocazione proporzionale delle 124 nazioni?
- d. per quale motivo a vostro parere non è stata utilizzata l'allocazione proporzionale delle nazioni?

3. Motivando la risposta, quale delle seguenti affermazioni sul censimento della popolazione è da ritenersi corretta?
- a. il censimento è, a tutti gli effetti, una indagine campionaria poichè ci sono sempre persone che rifiutano di partecipare;
 - b. il censimento non è una indagine statistica poichè i dati raccolti sono usati principalmente per scopi amministrativi;
 - c. il censimento è una enumerazione completa della popolazione poichè, in linea di principio, ad ogni membro della popolazione sono richieste informazioni.
4. Con riferimento all'imputazione dei dati nella cosiddetta fase di *data preparation*:
- a. descrivere in quali situazioni è adottata tale procedura, specificando anche le ipotesi di riferimento;
 - b. fornire la formulazione del modello generale per l'imputazione singola;
 - c. quali valori devono assumere i parametri del modello descritto al punto [b.] per ottenere il caso particolare dell'imputazione della media?
5. Descrivere similarità e differenze dei *web panel* reclutati da società di mercato (per esempio SWG community di SWG) e i panel utilizzati dall'Istat nelle indagini longitudinali (per esempio l'indagine Istat sul reddito e le condizioni di vita EU-SILC).