L'investimento in istruzione: come cambiano le opportunità dei laureati di ieri e di oggi

In questo lavoro viene utilizzato un modello multilivello o gerarchico con intercetta casuale, che permette di considerare la struttura gerarchica dei dati oggetto di studio, dove le unità di primo livello (i laureati) sono raggruppate in unità di secondo livello (l'ateneo presso il quale è stata conseguita la laurea). In generale le osservazioni di primo livello non sono completamente indipendenti: ad esempio, i laureati di uno stesso ateneo, avendo una storia comune condivisa frequentando lo stesso ambiente di studio, tendono ad essere più simili rispetto ai laureati di un diverso ateneo . Come conseguenza, la correlazione media tra le variabili misurate sui laureati di uno stesso ateneo tende ad essere maggiore della correlazione media tra le variabili misurate sui laureati di atenei diversi.

I modelli convenzionali poggiano sull'ipotesi di indipendenza delle osservazioni. La violazione di tale ipotesi porta ad una sottostima degli errori standard che porta a ritenere i risultati statisticamente significativi anche laddove ciò sia infondato.

I modelli multilivello sono dunque utili in caso di violazione dell'ipotesi di indipendenza delle osservazioni.

Le misure di interesse che si è inteso modellare sono la probabilità per un laureato di essere occupato a tre anni dal conseguimento del titolo e la probabilità di avere un'occupazione ottimale, ovvero

$$P_{ij} = Prob(Y_{ij} = 1),$$

dove Y_{ij} è pari a 1 se il laureato i dell'ateneo j è caratterizzato dall'evento di interesse (essere occupato oppure essere occupato ottimale) e 0 altrimenti.

Tale probabilità è funzione delle caratteristiche individuali del laureato, del suo percorso di studio e dell'ateneo di appartenenza.

Il modello gerarchico di regressione logistica a due livelli, ad intercetta e pendenze casuali, può essere rappresentato nel seguente modo:

$$logit(P_{ij}) = log\frac{P_{ij}}{1-P_{ii}} = \beta_0 + u_{0j} + \boldsymbol{\beta_k^T K_{ij}} + \boldsymbol{u_{wj}^T W_{ij}} + \boldsymbol{\beta_j^T R_j} + \varepsilon_{ij}$$

dove il vettore $\boldsymbol{K_{ij}}$ rappresenta le variabili esplicative di primo livello (con $\boldsymbol{W_{ij} \subseteq K_{ij}}$), il vettore $\boldsymbol{R_j}$ quelle di secondo; i parametri $\boldsymbol{\beta}$ sono i relativi coefficienti, e rappresentano gli effetti fissi del modello, costanti tra gli atenei; $\boldsymbol{u_{0j}}$ e $\boldsymbol{u_{wj}}$, caratteristica distintiva dei modelli gerarchici, rappresentano i residui di secondo livello (effetti casuali) associati all'intercetta e alle pendenze delle variabili esplicative.

In altri termini $\boldsymbol{u_{0j}}$ e $\boldsymbol{u_{wj}}$ rappresentano la parte variabile dell'intercetta e delle pendenze; essendo liberi di variare tra i gruppi, rappresentano gli effetti dovuti all'ateneo di appartenenza.

L'ipotesi sottostante al modello è che i residui di secondo livello si distribuiscano in modo normale con media nulla e varianza costante.

ε_{ij} rappresenta il residuo di primo livello, anch'esso con distribuzione normale, media nulla e varianza costante.

Nel caso specifico dell'analisi realizzata in questo lavoro, il modello utilizzato ha una forma semplificata in quanto non prevede effetti casuali associati alle pendenze delle variabili esplicative (il vettore $\boldsymbol{W_{ij}}$ è dunque nullo), così come non sono state incluse variabili di secondo livello legate all'ateneo ($\boldsymbol{R_j}$ è nullo).

Per il modello relativo alla probabilità di essere occupato a tre anni,le variabili esplicative di primo livello, ovvero gli elementi del vettore $\boldsymbol{K_{ij}}$, sono: il gruppo di afferenza del corso di laurea,

il conseguimento di una laurea in corso, la classe del voto di laurea, l'aver lavorato durante gli studi, il tipo di scuola secondaria frequentata, il genere, la classe di età al conseguimento della laurea, la professione dei genitori e l'area geografica di residenza. Per il modello sulla probabilità di avere un'occupazione ottimale, alle variabili suddette è stata aggiunta, per le annualità più recenti, anche l'aver trascorso periodi di studio all'estero.

Nella stima dei modelli multilivello, al fine di verificare se i dati abbiano una struttura gerarchica, si procede stimando inizialmente il modello 'nullo', ovvero senza variabili esplicative:

$$\text{logit}(P_{ij}) = \beta_0 + u_{0j} + u_{0j} + \varepsilon_{ij}$$

con $\varepsilon_{ij} \sim N(0, \sigma^2)$ e ε_{ij} indipendente da u_{0j} $\forall$ i,j

Tale modello presenta tre parametri: la media generale , la varianza di primo livello e la varianza di secondo livello τ.

Per il modello 'nullo' si ha che

$\text{Var}[\text{logit}(P_{ij})] = \text{Var}(u_{0j} + \varepsilon_{ij}) = \tau + \sigma^2$,

cioè la variabilità della trasformata logit della variabile dipendente può essere decomposta in una parte dovuta alla variabilità tra i gruppi (gli atenei) e una parte dovuta alla variabilità individuale.

Per verificare l'eventuale struttura gerarchica dei dati, si determina il coefficiente di correlazione intraclasse (ICC), dato da:

$$\text{ICC} = \rho = Corr\left[logit(P_{ij}), logit(P_{i'j'})\right] = \frac{\tau}{\tau + \sigma^2} = \frac{varianza\ dovuta\ ai\ gruppi}{varianza\ totale}$$

con $\rho \in [0,1]$

L'ICC fornisce una misura del grado di omogeneità tra osservazioni appartenenti allo stesso gruppo:quanto più elevato è il suo valore, tanto più importante diviene il ricorso ad una procedura di stima che tenga conto della dipendenza e della correlazione positiva tra le di unità di primo livello appartenenti ad una stessa unità di secondo livello.

Per saperne di più

Gelman A., Hill J. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge: Cambridge University Press.

Goldstein H. (1995). Multilevel Statistical Models. London: Edward Arnold.

J.Hox (2010) Multilevel Analysis Techniques and Applications, Lawrence Erlbaum Associates Publishers: London, Mahwah, New Jersey.