Modelli multilevel a intercetta variabile

Francesco Politano

2024-12-09

Indice

I dati multilevel

Nell'analisi della regressione affrontata finora nel corso, con approccio classico, è stato chiesto di estrarre informazioni dai dati, per poter condurre inferenze e descrivere la realtà attraverso l'utilizzo dei numeri.

Finora la variabilità degli outcome è stata studiata anche al variare di predittori di tipo categorico. Ad esempio, l'effetto del proprio campo di studi (STEM/nonSTEM) sui salari. Oppure si è risposto a domande ancora più avanzate con l'interazione, come l'effetto del campo di studi sul gap salariale tra donne e uomini. Per ora quindi le variabili categoriche sono state sempre ricodificate, a prescindere

dal numero di modalità che avevano, come delle variabili indicatrici. All'interno sempre però di un contesto in cui il campione è stato considerato come un unico blocco. Lo stesso è stato fatto per la matrice dei dati. La ricodifica in variabili dicotomiche verrà riesaminata ulteriormente più avanti.

Un nuovo utilizzo per alcune variabili categoriche

In molti casi di analisi di dati reali, è possibile che una colonna della nostra matrice permetta di raggruppare le unità statistiche. Per esempio:

- nel caso di una matrice di dati di studenti, la variabile che permetta di aggregarli potrebbe essere la scuola in cui studiano
- per dati su dei pazienti, la variabile potrebbe essere l'ospedale in cui sono ricoverati
- nel caso di dati su elettori, la variabile potrebbe essere il loro stato di residenza
- per dati sulle case negli Stati Uniti, la variabile potrebbe essere la contea in cui si trovano.

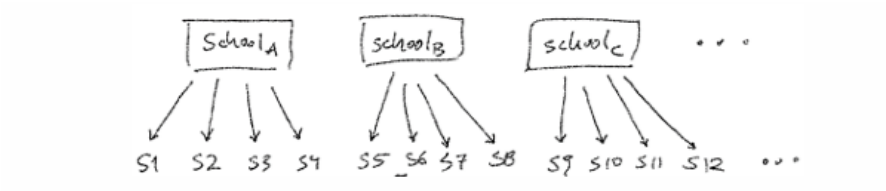


Figura 1: Visualizzazione della struttura dei dati multilevel gerarchici. Figura da Rosche (2022), Intro to multilevel modeling.

Gli identificativi dei gruppi (scuole, ospedali, stati, contee) vengono indicati solitamente dalla notazione come numeri progressivi da 1 a J (numero totale di gruppi). In un'aggregazione gerarchica semplice come quella appena presentata, ogni gruppo è un' *unità macro* (o di livello 2).

Inoltre, questi gruppi potrebbero avere anche delle loro caratteristiche, come riportato nel paragrafo *Le variabili Macro*.

Per poter capire l'effetto medio dei predittori sull'outcome, come si possono trattare i dati multilevel?

3 tipi di modelli classici

Le possibilità già note, utilizzando un approccio classico, sono principalmente 3:

- stimare un modello unico (senza indicatrici dei J gruppi) - modello *complete-pooling*
- stimare un modello unico con le indicatrici dei J gruppi
- stimare tanti modelli quanti sono i gruppi, separatamente - modello *no-pooling*

Ciascuno di questi 3 modelli, però, non è adatto.

Stimare un modello *complete-pooling* Il problema del modello di *complete-pooling* è che si perdono completamente le informazioni sui gruppi, ignorandone l'esistenza. Per esempio, non si considera che gli studenti appartengono a delle classi.

Il seguente esempio storico, legato proprio alla nascita dei modelli multilevel nelle scienze sociali, può aiutare a chiarire meglio quali siano le conseguenze di un approccio *complete-pooling*.

Bennet nel 1976 ha condotto un'analisi di tipo tradizionale, e i suoi risultati sono stati smentiti da Aitkin nel 1981, adottando un approccio multilevel.

La domanda di ricerca di entrambi riguardava l'istruzione di bambini negli ultimi anni della scuola elementare: per l'insegnamento della lettura, è meglio un approccio tradizionale oppure uno nuovo più informale? C'è effettivamente differenza tra i 2?

Bennett, stimando il suo modello classico ha concluso che c'era una differenza significativa nei risultati, preferendo la tecnica formale.

Aitkin ha affermato invece che tale differenza non era statisticamente significativa.

I risultati di Bennett non consideravano, erroneamente, il fallimento di una delle assunzioni fondamentali dei modelli classici: l'indipendenza degli errori. Va invece considerato il fatto che gli studenti, apprendendo nella stessa classe dallo stesso insegnante, apprendono in maniera più simile di coloro che sono inseriti in classi diverse.

Nella pratica, a cosa porta avere dei dati "dipendenti" tra loro? Ad avere meno informazioni che a confrontare studenti di classi diverse, con insegnanti diversi. Ai fini dell'analisi, avere più studenti oltre un certo numero provenienti dalla stessa classe non aiuta a stimare meglio l'effetto del metodo di insegnamento, perché quello migliora solo la precisione della stima dei risultati in una certa classe. Avere invece più classi e più insegnanti potrebbe aggiungere informazioni sulla differenza tra approccio formale e informale.

Bennett aveva considerato statisticamente significativo l'effetto del metodo di insegnamento perché sovrastimava la quantità di informazione presente nei dati. Di fatto, l'ampiezza reale del suo campione era inferiore al numero di studenti esaminati. E così facendo, gli standard error dei coefficienti erano sottostimati, conducendo a inferenze erronee.

Stimare un modello con indicatrici dei gruppi In questo caso l'informazione data dall'esistenza dei gruppi viene integrata nel modello. Purtroppo, però, vengono stimati $J-1$ coefficienti di variabili dicotomiche. Se un modello ha un vasto numero di gruppi (il che è auspicabile per i motivi visti nell'esempio di Bennett e Aitkin), l'interpretazione di così tanti coefficienti rischia di diventare un'impresa proibitiva.

Inoltre, se si volesse rispondere ad altre domande attraverso il modello, come l'effetto di alcune caratteristiche dei gruppi sull'outcome, si andrebbe a verificare il problema della collinearità tra i predittori. Non si tratta di un semplice difetto nelle stime come per i precedenti standard error sottostimati: nessun software restituirebbe dei risultati. Ad esempio su `R` il coefficiente della variabile macro risulterebbe `NA` o non comparirebbe proprio nell'outcome.

Stimare un modello *no-pooling* In questo caso si stima un modello a parte per ogni classe. Citando McElreath,[1] questo è un modello con *anterograde amnesia*, perché non ricorda ciò che succede negli altri gruppi quando stima ogni modello.

Nel caso dell'esempio, si andrebbe a stimare il livello dei progressi dei vari alunni, controllando per altri predittori che potrebbero intervenire all'interno di ogni classe (come ad esempio l'istruzione dei genitori). Otteniamo così delle intercette, che sono livelli di progressi a parità di altre condizioni per i bambini *baseline*, confrontabili tra i vari gruppi.

I problemi di un approccio del genere, in questo caso, sarebbero:

- la numerosità dei gruppi: quanto sarebbero affidabili delle stime ottenute con campioni di soli 10/20 studenti? E se per risparmiare si volesse evitare nella nostra indagine di effettuare praticamente un censimento classe per classe, intervistando magari meno di 10 bambini, si potrebbe non avere neanche un campione sufficiente ad avere le stime dei controlli o dei predittori (anche in questo caso `R` restituirebbe `NA`)
- l'amnesia (che per ogni gruppo che studiamo, si dimentica ciò che abbiamo appreso per il precedente) è inoltre un difetto, ed è proprio ciò che dà problemi sulle stime all'interno dei singoli gruppi.

[1] dal libro *Statistical Rethinking*

- possibilità di rispondere alla domanda di ricerca. Non potrei avere nella stessa classe alunni che abbiano imparato da un insegnante con metodo formale e uno con insegnante con metodo informale. Dovrei confrontare le intercette dei vari gruppi, per concludere poi in quali classi gli individui *baseline* abbiano ottenuto i risultati migliori. Ma fare questo, e poter fare test e conclusioni sulla significatività non sarebbe possibile a meno di aggiungere ulteriori assunzioni.

Da qui in avanti, dunque, verrà adottato un nuovo approccio diverso da quello classico, che possa permettere di rispondere correttamente a domande di ricerca presenti quotidianamente nell'analisi di dati reali.

Dall'approccio single-level al partial-pooling

Il fallimento di assunzioni dei modelli classici quali:

1. indipendenza e identica distribuzione degli errori
2. omoschedasticità degli errori

nel caso dei dati che presentano una struttura multilevel rende necessaria l'introduzione di una nuova famiglia di modelli, che riproduca la gerarchia presente nei dati. Questi modelli, detti multilevel,[2] presentano una gerarchia nei loro parametri.

L'approccio multilevel è un compromesso tra i cosiddetti approcci

1. *no-pooling* (stimare J modelli separati, quanti sono i gruppi)
2. *complete-pooling* (stimare un modello unico ignorando la presenza raggruppamenti tra unità).

Modello nullo

Ciò si può formalizzare innanzitutto per un modello lineare nullo semplice.

Nel caso dell'approccio classico (*single-level*):

$$y_i = \alpha + \epsilon_i \tag{1}$$

con α costante (in questo caso sarà pari alla media delle y_i, cioè $\alpha = \frac{\sum_{i=1}^{n} y_i}{n}$) e $\epsilon_i \sim N(0, \sigma^2)$ i.i.d.

Nel caso dell'approccio multilevel, sono presenti J gruppi, e $j[i]$ rappresenta l'identificativo del gruppo a cui appartiene l'unità $i - esima$. Per questo primo modello si può innanzitutto dare la possibilità al coefficiente dell'intercetta di variare tra i vari gruppi $1, \ldots, J$.

$$y_i = \alpha_{j[i]} + \epsilon_i \tag{2}$$

con

- $\epsilon_i \sim N(0, \sigma_y^2)$
- $\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$

Questo modello è detto *varying-intercept model*[3]. μ_α e σ_α^2 sono iperparametri del modello, utili a descrivere la distribuzione di α_j nei vari gruppi $1, \ldots, J$. Un modello multilevel infatti:

- permette ai coefficienti della regressione (alcuni o tutti) di variare tra i gruppi
- assume un modello per la variabilità di questi coefficienti tra i gruppi. Questo è l'elemento che distingue i modelli multilevel da un modello semplice con variabili indicatrici.

[2] detti anche di *partial-pooling*, *mixed-effects models* o *random-effects models*
[3] modello a intercetta variabile

In questo caso si sta assumendo che le α_j possano variare tra i gruppi, e che guardando l'insieme delle intercette variabili dei gruppi esse seguano una distribuzione normale centrata in μ_α e con varianza σ^2_α.

Esempio: dati scolastici

Nel caso si stiano studiando i punteggi in un test sulla comprensione del testo[4] di un campione di studenti provenienti da diverse scuole, il modello nullo a intercette variabili più semplice [5], con il punteggio come *outcome*, si può scrivere nel seguente modo:

$$y_i = \alpha_{j[i]} + \epsilon_i \qquad \text{con} \qquad \epsilon_i \sim N(0, \sigma^2_y) \tag{3}$$

$$\alpha_j = \mu_\alpha + \eta_j \qquad \text{con} \qquad \eta_j \sim N(0, \sigma^2_\alpha) \quad \text{e} \quad Cov(\epsilon_i, \eta_{j[i]}) = 0 \tag{4}$$

Questo modello anziché stimare l'intercetta come media generale dei punteggi degli studenti come farebbe un modello *single-level*, produce J stime di intercette (una per ogni scuola), e una stima dell'iperparametro μ_α (intorno al quale variano, seguendo una normale, le intercette stimate per ogni scuola).

Cosa rappresentano le intercette

Nei modelli classici le intercette di un modello nullo rappresentano il livello medio del fenomeno nel campione. Nel caso del partial-pooling, le intercette non sono direttamente delle medie, ma il loro ruolo è molto simile.

Attraverso la stima di varie intercette (una per gruppo), in un modello a intercette casuali, stimiamo il livello medio del fenomeno in un gruppo. La presenza di gruppi di unità è una novità rispetto alla regressione *single-level*. Essi rappresentano infatti una nuova tipologia di fonte di informazione: per stimare il livello medio del fenomeno in un gruppo, non ha senso attribuire lo stesso peso a tutte le unità del campione.

Ha più senso, quindi, che il modello (che nel caso multilevel fa comunque tutto simultaneamente) dia maggior peso alle unità che appartengono al gruppo di cui ci interessa ottenere una stima. È poi anche utile tenere in considerazione anche il resto del campione, anche perché altrimenti si starebbe stimando un modello *no-pooling*, con tutti i problemi del caso.

Espressione approssimata dell'intercetta α_j

Per ottenere $\hat{\alpha_j}$ le fonti di informazione da sfruttare sono dunque:

1. $\bar{y}_j$ (media campionaria delle unità presenti nel gruppo j, anche detta stima *unpooled*)
2. $\bar{y}_{\text{all}}$ (media dell'intero campione, anche detta stima *completely-pooled*)

Queste fonti dovrebbero essere pesate nella maniera più utile ad avere delle stime affidabili e consistenti del livello medio del fenomeno.

Quindi se l'*outcome* è particolarmente eterogeneo all'interno dello stesso gruppo (alta variabilità *within*, per esempio se i punteggi degli studenti sono notevolmente variabili nella stessa scuola), allora la media campionaria di tale gruppo ha una varianza maggiore. Lo stesso si può dire nel caso in cui si abbiano poche osservazioni per un gruppo. Infatti che la varianza della media campionaria di y dipende dalla varianza di y e dalla numerosità del campione (anche nel caso non si assuma che la y sia gaussiana).

[4] dati PIRLS 2016

[5] senza predittori a livello macro

$$Var(\bar{y}_j) = \frac{\sigma_y^2}{n_j} \tag{5}$$

Nel caso in cui, invece, il fenomeno sia fortemente diverso tra un gruppo e l'altro (in cui, ad esempio, le scuole abbiano dei rendimenti degli studenti molto diversi tra loro), gli altri gruppi sono una fonte poco affidabile. Questo è il caso di un'elevata variabilità *between*: essa è descritta dall'iperparametro σ_α^2.

Sarebbe dunque utile pesare le 2 fonti di informazioni precedenti (media campionaria del gruppo j e media dell'intero campione) in modo che la loro rilevanza nella stima dell'intercetta α_j dipenda da:

- $\frac{\sigma_y^2}{n_j}$
- σ_α^2

In particolare, all'aumentare di queste varianze le rispettive fonti di informazioni diventano meno affidabili.

Ecco perché si può dire che la stima multilevel dell'intercetta $\hat{\alpha}_j^{\text{multilevel}}$ è approssimativamente pari a:

$$\hat{\alpha}_j^{\text{multilevel}} = \frac{\frac{n_j}{\sigma_y^2}\bar{y}_j + \frac{1}{\sigma_\alpha^2}\bar{y}_{\text{all}}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\alpha^2}} \tag{6}$$

La stima dell'intercetta quindi, a meno di casi estremi, non coincide con la media campionaria del singolo gruppo.

Come capire la struttura della variabilità della y

Nel modello nullo esposto finora si ottengono altre stime oltre a quelle di μ_α e delle α_j. Vengono infatti anche simultaneamente stimati gli iperparametri σ_y^2 e σ_α^2, che hanno, come visto prima, anche un legame con la stima delle intercette dei gruppi.

Questi parametri descrivono la variabilità residua del fenomeno e, poiché il modello ha 2 livelli, anche essa ne presenta 2. La variabilità residua all'interno dei gruppi è descritta dall'iperparametro σ_y^2, mentre l'iperparametro σ_α^2 descrive la dispersione delle intercette α_j intorno alla loro media μ_α.

Considerare il rapporto tra le varianze può arricchire la lettura del nostro modello. Calcolando infatti il reciproco di $\hat{\sigma}_\alpha^2/\hat{\sigma}_y^2$ otteniamo una stima del numero di unità entro un gruppo che servono a dare la stessa quantità di informazione portata dagli altri gruppi.

Infatti affinché il peso di $\bar{y}$ e $\bar{y}_{\text{all}}$ sia lo stesso nella stima dell'intercetta, serve che:

$$\frac{\sigma_y^2}{n_j} = \sigma_\alpha^2 \tag{7}$$

$$\text{da cui} \tag{8}$$

$$n_j = \frac{\sigma_y^2}{\sigma_\alpha^2} \tag{9}$$

Perciò, in tutti quei gruppi che abbiano una numerosità maggiore della $\hat{n}_j$ stimata è la media delle unità appartenenti al gruppo a contare di più. Nei gruppi meno numerosi, invece, la fonte privilegiata di informazione sono gli altri gruppi.

Questo si può sintetizzare anche con un indice che assume esclusivamente valori tra 0 e 1: il coefficiente di correlazione intraclasse.

$$ICC = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_y} \tag{10}$$

- Quando assume valore 0 la divisione in gruppi delle unità non è informativa, in quanto essi hanno variabilità minima *tra* loro
- Quando assume valore 1 le unità dello stesso gruppo sono tutte identiche.

Esempio: i dati PIRLS 2016

Si analizzino i risultati di un test sulla comprensione del testo somministrato a studenti di quarta elementare in Italia nel 2016. I dati degli studenti e quelli delle scuole a cui sono iscritti sono presenti nel dataset `ita16` PIRLS2016.

```
load("ita_reduced.RData")
```

Per la stima dei modelli multilevel lineari e lineari generalizzati è necessario il pacchetto `lme4`

```
library(lme4)
```

La stima del modello nullo a intercetta variabile senza predittori di livello 2 (o macro) si ottiene attraverso la funzione `lmer`, nella quale, a destra del tilde ~ ci sono:

- la parte usuale della formula per i modelli lineari
- una nuova parte che introduce la variabilità dei coefficienti tra gruppi, posta tra parentesi. I termini prima della barra verticale definiscono quali coefficienti sono variabili tra i gruppi, mentre i termini dopo la barra definiscono quali sono i gruppi (identificativo dei gruppi).

La variabile `reading_score` è l'*outcome*.

```
fit0 <- lmer(data=ita16,formula=reading_score~1+(1|idschool))
```

Attraverso la funzione `display` della library `arm` possiamo ottenere una sintesi dell'output del modello.

```
library(arm)
display(fit0)
```

```
## lmer(formula = reading_score ~ 1 + (1 | idschool), data = ita16)
## coef.est  coef.se
##   548.93     2.22
##
## Error terms:
##  Groups   Name        Std.Dev.
##  idschool (Intercept) 24.03
##  Residual             59.76
## ---
## number of obs: 3940, groups: idschool, 149
## AIC = 43654.7, DIC = 43655.5
## deviance = 43652.1
```

Da cui:

- $\hat{\mu_\alpha} = 548.93$
- $\hat{\sigma}_y = 59.76$

- $\hat{\sigma}_\alpha = 24.03$

Si possono poi visualizzare i coefficienti dei primi 6 gruppi (stime delle medie dei punteggi nelle prime 6 scuole), e notare come non coincidono con le medie campionarie (Figura 1).

```
stime_intercette
```

```
##   intercette medie campionarie numerosità
## 1      529.7             524.1         21
## 2      536.0             529.3         12
## 3      532.6             522.4         10
## 4      547.0             546.4         22
## 5      540.5             538.5         25
## 6      556.9             559.1         22
```

Possiamo notare, sia dalla precedente tabella che dalla figura 1, come tutte le stime (prima colonna) tendano ad avvicinarsi, rispetto alla media campionaria del loro gruppo, all'intercetta $\hat{\mu_\alpha}$ (aumentando, nelle prime 5 scuole, e diminuendo nella scuola 6). Inoltre, la scuola 1, grazie alla numerosità doppia rispetto alla 3, ha un'intercetta che risente meno dell'influenza del fenomeno negli altri gruppi: a fronte di una media campionaria leggermente più alta della scuola 3, si avvicina meno al valore dell'intercetta generale.

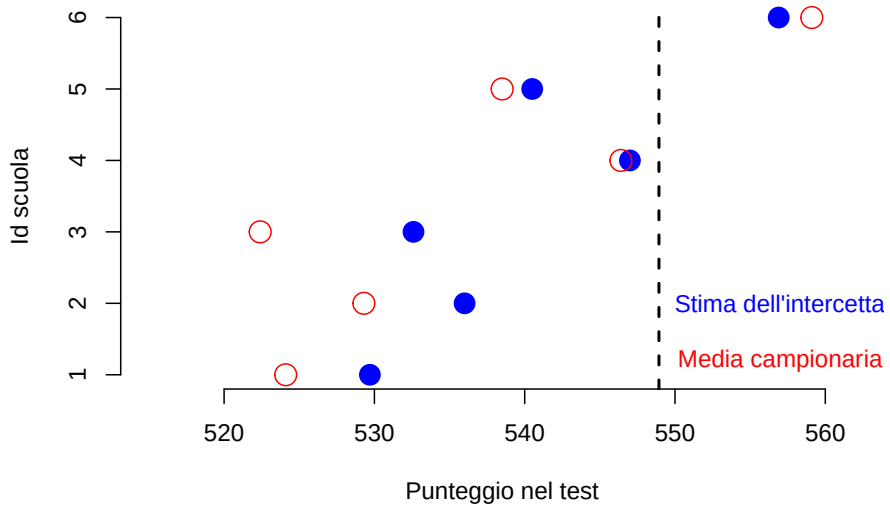


Figura 2: Stima delle intercette a seconda delle medie campionarie dei gruppi. Sono riportate le prime 6 scuole, nel modello nullo su dati PIRLS2016. La linea tratteggiata verticale rappresenta l'intercetta fissa

Il reciproco di $\hat{\sigma}^2_\alpha/\hat{\sigma}^2_y$, che ci dà informazioni sul numero minimo di unità in un gruppo per pareggiare la quantità di informazioni proveniente dagli altri gruppi, è pari a:

```
round(sigma_y^2/sigma_alpha^2,1)
```

```
## [1] 6.2
```

Quindi, anche se non sono presenti in questo dataset, i gruppi con meno di 6 unità hanno un'intercetta

che si avvicina più alla media generale y_{all} che alla media campionaria del loro gruppo. Oltre le 6 unità (ovvero per tutte le scuole di questo dataset) accade il contrario.

L'output di `lmer`

- Gli effetti fissi (i coefficienti che non variano tra i gruppi, o la loro media, usata per scrivere l'equazione "media" della regressione)

```
fixef(fit0)
```

```
## (Intercept)
##    548.9318
```

- Gli effetti casuali (la parte dei coefficienti che varia tra i gruppi) η_j, con cui si possono ottenere le intercette variabili $\alpha_j = \mu_\alpha + \eta_j$

```
head(ranef(fit0)$idschool)
```

```
##   (Intercept)
## 1  -19.217160
## 2  -12.977028
## 3  -16.381285
## 4   -1.964269
## 5   -8.388138
## 6    7.932243
```

- Gli standard error degli effetti fissi, utili per gli intervalli di confidenza

```
se.fixef(fit0)
```

```
## (Intercept)
##    2.217175
```

```
fixef(fit0)+c(-2,2)*se.fixef(fit0)
```

```
## [1] 544.4975 553.3662
```

- Gli standard error degli effetti casuali, utili per gli intervalli di confidenza

```
head(se.ranef(fit0)$idschool)
```

```
##   (Intercept)
## 1    11.46184
## 2    14.01379
## 3    14.85435
## 4    11.25668
## 5    10.70158
## 6    11.25668
```

```
cbind(fixef(fit0)+head(ranef(fit0)$idschool)+
  c(-2)*head(se.ranef(fit0)$idschool), #estremo inferiore
  fixef(fit0)+head(ranef(fit0)$idschool)+
  c(2)*head(se.ranef(fit0)$idschool)) #estremo superiore
```

```
##   (Intercept) (Intercept)
## 1    506.7910    552.6383
## 2    507.9272    563.9824
## 3    502.8418    562.2592
## 4    524.4542    569.4809
```

```
## 5    519.1405    561.9468
## 6    534.3507    579.3774
```

Modello con predittori micro

Sempre restando all'interno dei modelli a sola intercetta variabile, è possibile aggiungere altri predittori a livello micro (ad esempio, variabili relativi agli studenti), che possano rispondere a delle domande di ricerca o fare da controlli per le nostre stime, garantendo una parità di condizioni che altrimenti altererebbe il confronto tra i vari gruppi a causa delle differenze presenti al loro interno.

Modelli del genere si possono stimare nel seguente modo:

```
fit1 <- lmer(data=ita16,formula=reading_score~male_stud+
               rescale(age_stud)+lang2_often_stud+
               books2_home_stud+
               (1|idschool))
```

mantenendo solamente 1 all'interno della parentesi riguardante i coefficienti variabili tra i gruppi.

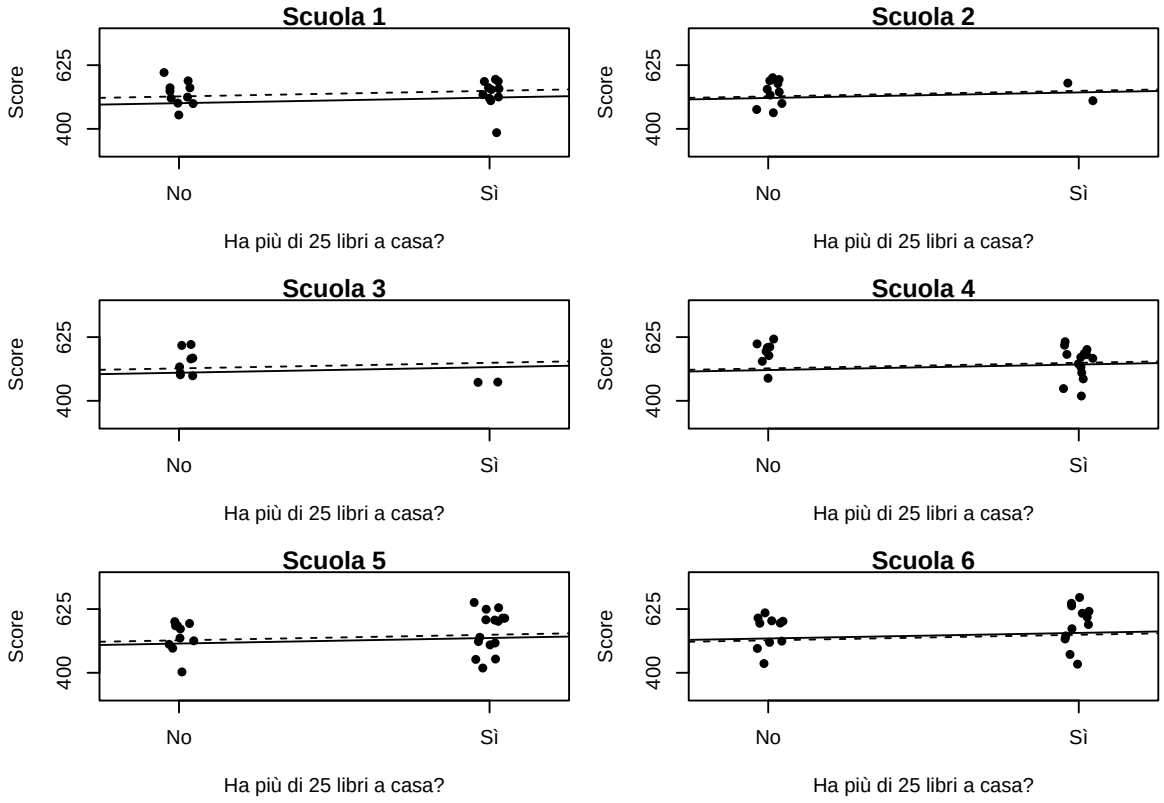
```
display(fit1)
```

```
## lmer(formula = reading_score ~ male_stud + rescale(age_stud) +
##     lang2_often_stud + books2_home_stud + (1 | idschool), data = ita16)
##                   coef.est coef.se
## (Intercept)       514.57     3.17
## male_stud          -7.80     1.86
## rescale(age_stud)   5.91     1.92
## lang2_often_stud   34.87     2.58
## books2_home_stud   19.64     1.93
##
## Error terms:
##  Groups   Name        Std.Dev.
##  idschool (Intercept) 20.99
##  Residual             57.34
## ---
## number of obs: 3897, groups: idschool, 149
## AIC = 42826.1, DIC = 42844.7
## deviance = 42828.4
```

Grafico per modello a intercetta variabile Si possono rappresentare con il seguente codice le rette di regressione baseline in funzione dei libri presenti a casa:

```
par(mfrow=c(3,2),mar=c(5,4,1,1))
for(scuola in 1:6){
plot(ita16$books2_home_stud,ita16$reading_score,type="n",
     xaxt="n",xlim=c(-0.2,1.2),xlab="Ha più di 25 libri a casa?",
     ylab="Score",main=paste("Scuola",scuola),yaxt="n")
axis(2,at=c(400,625),labels=c(400,625))
axis(1,at=c(0,1),labels=c("No","Sì"))
points(jitter(ita16$books2_home_stud[ita16$idschool==scuola],0.25),
       ita16$reading_score[ita16$idschool==scuola],pch=16)
intercetta <- coef(fit1)$idschool[scuola,1]
pendenza <- coef(fit1)$idschool[scuola,5]
abline(a=intercetta,b=pendenza)
```

```
intercetta_fissa <- fixef(fit1)[1]
abline(a=intercetta_fissa,b=pendenza,lty=2)
}
```



A variare sono le intercette nei vari gruppi, rispetto a quella fissa che è rappresentata nella retta tratteggiata (modello "medio" generale per tutti i gruppi). Si può notare come quindi l'unica scuola rappresentata qui con dei punteggi mediamente più alti della media dei gruppi è la 6.

Il fallimento dell'assunzione di errori i.i.d.

Le assunzioni dei modelli di regressione classici sono, in ordine decrescente di importanza:

- *validità* - L'*outcome* deve essere una misura che rifletta effettivamente il fenomeno di interesse, e gli input sono i predittori "rilevanti"
- *additività e linearità*, in assenza delle quali sono necessarie trasformazioni dei dati (logaritmi, termini quadratici, aggiunta di interazioni, o altro ancora)
- *indipendenza degli errori*
- *errori omoschedastici* (a stessa varianza), in assenza dei quali è opportuno usare i minimi quadrati pesati (WLS) o i modelli multilevel[6]
- *normalità degli errori* (da non confondere con la normalità della distribuzione dell'outcome, esempio: misura dell'altezza di un campione di donne e uomini). Essa si verifica guardando ai residui, ma, anche se non si distribuiscono seguendo una normale, l'importante è che non mostrino dei *pattern* non casuali (per esempio con degli scatterplot che li rappresentino in funzione di qualche predittore x_i o dei valori stimati $X_i\hat{\beta}$)

[6] che in qualche modo, come vedremo, a loro volta pesano non tutte le osservazioni in maniera uguale

Quello che succede con i modelli multilevel, come dimostrato dal caso di Bennett e Aitkin, è legato all'indipendenza degli errori (per tutti) e alla loro varianza (nel caso dei *modelli a pendenza variabile*).

Questi modelli, in letteratura detti anche *mixed-effect models* o *random-effect models*,[7] possono essere anche espressi in un altro modo, che mostra le differenti assunzioni sulla variabilità rispetto a un modello *single-level* classico.

$$y_i = X_i\beta + \epsilon_i \tag{11}$$

con $X_i\beta$ costante (e in questo caso non coincide necessariamente né con la media generale, né con la media dei gruppi, come verrà mostrato più avanti) e $\epsilon \sim N(0, \Sigma)$. La struttura di covarianza degli errori stavolta non è necessariamente diagonale, a differenza del caso degli errori i.i.d. del modello *single-level*. Gli elementi di tale matrici sono pari a:

- $\Sigma_{ii} = Var(\epsilon_i) = \sigma_y^2 + \sigma_\alpha^2$
- $\Sigma_{ik} = Cov(\epsilon_i, \epsilon_k) = \sigma_\alpha^2$ (per unità nello stesso gruppo, $j[i] = j[k]$)
- $\Sigma_{ik} = Cov(\epsilon_i, \epsilon_k) = 0$ (per unità in differenti gruppi, $j[i] \neq j[k]$)

Si può riscrivere quindi anche:

$$corr(\epsilon_i, \epsilon_k) = \frac{\Sigma_{ik}}{\sqrt{\Sigma_{ii}\Sigma_{kk}}} = \begin{cases} \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_y^2} = ICC & \text{se} \quad j[i] = j[k] \\ 0 \quad \text{se} & j[i] \neq j[k] \end{cases} \tag{12}$$

Gli errori di osservazioni appartenenti allo stesso gruppo non sono quindi incorrelati, e anzi sono associati positivamente: infatti la loro covarianza è pari a σ_α^2. Ad esempio, gli studenti che hanno avuto lo stesso insegnante, che hanno frequentato la stessa scuola, gli elettori provenienti dallo stesso stato:[8] qualcosa li rende più simili di quelli che invece hanno avuto insegnanti, scuole o residenze diverse.

Avere delle osservazioni tra loro correlate (condizionatamente ai predittori) non fornisce la stessa quantità di informazione di avere delle osservazioni non associate tra loro. Questa perdita di informazione fa sì che, difatti, la numerosità campionaria "reale" non sia veramente pari a n, ma sia inferiore.

Di conseguenza, ignorare la struttura delle osservazioni in gruppi può portare a sottostimare l'incertezza all'interno delle stime. Gli standard error infatti sono maggiori se, a parità di altre condizioni, l'ampiezza del campione è minore. I modelli classici rischiano, come nel caso di Bennett, di portare a considerare statisticamente significative anche stime di coefficienti che non lo sono. In particolare, come vedremo più avanti, nei modelli multilevel, gli standard error delle variabili macro possono essere decisamente maggiori.

Come i modelli a pendenza variabile rilassano l'ipotesi di omoschedasticità Nei modelli classici la varianza degli errori è considerata costante per le unità, ed è pari a σ_y^2.

Per i modelli multilevel a sola intercetta variabile si assume che la varianza degli errori sia costante, ma per ciascun livello: nel caso del modello presentato finora, sia σ_y^2 che σ_α^2 devono essere costanti.

Esistono, d'altra parte, dei modelli multilevel più avanzati che rilassano l'ipotesi che la variabilità accidentale (quella non spiegata sistematicamente dalla combinazione lineare di coefficienti e predittori osservati) sia costante, come i modelli *varying-slope* ("a pendenza variabile").

Il seguente modello si definisce a *pendenza variabile*:

[7] soprattutto in ambito frequentista

[8] legge di Tobler, per cui le unità più vicine tra loro sono più correlate di quelle tra loro distanti

$$y_i = \alpha_{j[i]} + \beta_{1j[i]} X_i + \epsilon_i \qquad \text{con} \qquad \epsilon_i \sim N(0, \sigma_y^2) \tag{13}$$

$$\alpha_j = \mu_\alpha + \eta_{0j} \qquad \text{con} \qquad \eta_{0j} \sim N(0, \sigma_\alpha^2) \quad \text{e} \quad Cov(\epsilon_i, \eta_{0j[i]}) = 0 \tag{14}$$

$$\beta_j = \mu_{\beta_1} + \eta_{1j} \qquad \text{con} \qquad \eta_{1j} \sim N(0, \sigma_{\beta_1}^2) \quad \text{e} \quad Cov(\epsilon_i, \eta_{1j[i]}) = 0 \tag{15}$$

Che in una sola riga si può riscrivere più compattamente come:

$$y_i = \underbrace{\mu_\alpha + \mu_{\beta_1} X_i}_{\text{parte fissa}} + \underbrace{\eta_{0j} + \eta_{1j} X_i + \epsilon_i}_{\text{parte variabile}} \tag{16}$$

La parte variabile del modello *varying-intercepts* era pari a:

$$\eta_j + \epsilon_i \tag{17}$$

assumendo che la sua varianza fosse costante e pari a:

$$Var(\eta_j + \epsilon_i) = \sigma_\alpha^2 + \sigma_y^2 \tag{18}$$

Nel caso del modello *varying-intercepts and slopes* invece la parte variabile è pari a:

$$\eta_{0j} + \eta_{1j} X_i + \epsilon_i \tag{19}$$

Questa non è costante: è presente infatti il termine X_i. La varianza delle sue componenti è pari a:

$$Var(\epsilon_i) = \sigma_y^2 \tag{20}$$

$$Var(\eta_{0j[i]} + \eta_{1j[i]} X_i) = \sigma_\alpha^2 + 2\sigma_{10} X_i + \sigma_{\beta_1}^2 X_i^2 \tag{21}$$

dove

- σ_α^2 è la varianza di $\eta_{0j[i]}$ (variabilità dell'intercetta)
- $\sigma_{\beta_1}^2$ è la varianza di $\eta_{1j[i]}$ (variabilità della pendenza)
- σ_{10}^2 è la covarianza di $\eta_{0j[i]}$ e $\eta_{1j[i]}$

Di conseguenza, si può affermare che a seconda di vari tipi di funzioni di varianza degli errori (legami tra la X e la varianza degli errori) ci si può ritrovare di fronte alle seguenti 3 situazioni nel caso di una X *quantitativa*,[9] calcolabili attraverso l'equazione 21 di cui sopra:

1. Quando le pendenze sono tutte uguali ($\sigma_{\beta_1}^2 = 0$) e di conseguenza non c'è associazione tra intercette e pendenze ($\sigma_{10} = 0$), allora gli errori hanno varianza costante, pari a $\sigma_y^2 + \sigma_\alpha^2$
2. Quando c'è una covarianza positiva tra intercette e pendenze ($\sigma_{10} > 0$), allora la varianza dell'errore cresce all'aumentare della X
3. Quando c'è una covarianza negativa tra intercette e pendenze ($\sigma_{10} < 0$), allora la varianza degli errori decresce all'aumentare della X

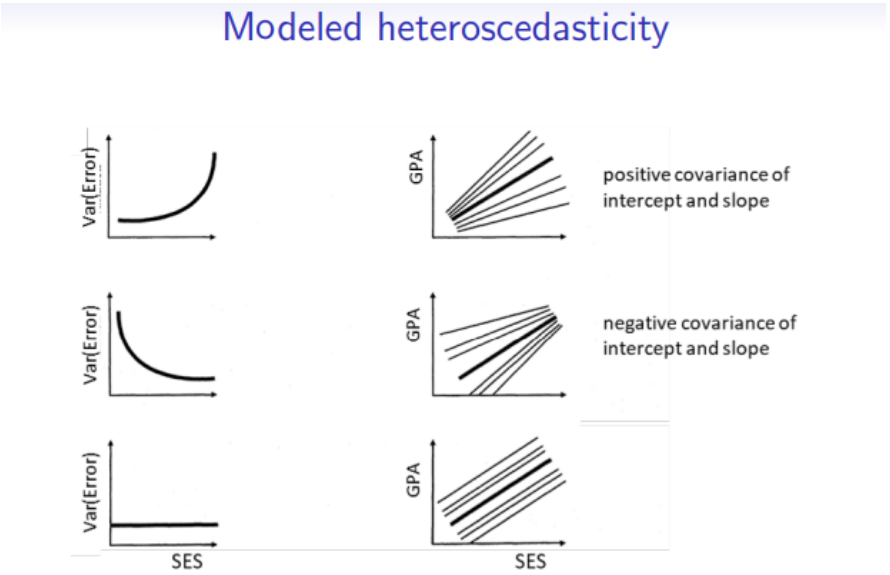


Figura 3: Differenti tipi di eteroschedasticità, e il loro legame con intercette e pendenze variabili. Figura da Rosche (2022), Intro to multilevel modeling. Adattata da Bullen, Jones e Duncan (1997)

Per concludere, il modello a intercetta variabile affrontato finora assume che le pendenze siano uguali (il modello non può restituire β_{1j} differenti per ogni gruppo). Questo fa sì che gli errori abbiano varianza costante, calcolata già prima.

Invece, i modelli a pendenza variabile rilassano l'ipotesi di omoschedasticità: la varianza degli errori può variare a seconda del valore dei predittori, anche più di uno, sia in maniera crescente che decrescente. In presenza di più pendenze variabili la funzione può risultare ancora più complessa, permettendo di essere ancora più flessibili riguardo le assunzioni sugli errori, superando la rigidità dei modelli classici. Questi predittori, inoltre, possono essere non solo quantitativi ma anche qualitativi.

Questo inoltre permette una maggiore varianza in alcuni gruppi piuttosto che in altri, a seconda delle caratteristiche dei predittori in quei gruppi.

Le variabili macro

Entrambi i modelli precedentemente visti (modello nullo a intercetta variabile e modello con predittori micro con la sola intercetta variabile) possono essere resi più complessi attraverso l'introduzione di variabili che descrivano le caratteristiche dei gruppi. Finora infatti il modello considera accidentali (e quindi non spiega) le differenti intercette α_j.

Finora infatti da un punto di vista formale $\alpha_j \sim N(\mu_\alpha, \sigma^2_\alpha)$. Poter spiegare le differenze esistenti tra i gruppi in maniera sistematica è una delle due caratteristiche fondamentali di un modello multilevel, oltre alla possibilità di far variare i coefficienti.

Per poter spiegare queste differenze è necessario aggiungere all'interno del modello un'ulteriore fonte di informazione: le variabili macro (o di livello 2, o in generale superiore a 1).

Le variabili macro sono, in molti casi di dati reali, reperibili ancora più facilmente di quelle micro. I risultati di un'elezione o il reddito medio in una certa area geografica possono non essere proprio disponibili a livello individuale. Si possono avere dei dati micro provenienti per esempio attraverso indagini

[9] assunzione utile a visualizzare meglio l'eteroschedasticità

campionarie o *survey*, dai quali si può cercare di risalire al comportamento dell'intera popolazione.

È per questi motivi che in molte ricerche vengono utilizzati anche direttamente i dati macro. Queste variabili macro possono essere:

- *globali*, quando rappresentano delle proprietà intrinseche dei gruppi (ad esempio, il tipo di una scuola, privata o pubblica, o la forma di governo in un Paese, il metodo di insegnamento, tradizionale o innovativo)
- *contestuali*, quando derivano dall'aggregazione dei dati degli individui in quel gruppo (ad esempio reddito, voto o punteggio medio)

Quando in un'analisi si utilizzano *solo* dati macro, bisogna essere cauti nelle proprie inferenze. In particolare, il rischio maggiore è quello di provare a fare:

- inferenza ecologica, in cui si cerca di sfruttare informazioni a livello macro sulle unità micro. Questo viene fatto guardando con analisi di correlazioni o regressioni[10] solamente a livello dei gruppi.

Nel caso delle elezioni americane questo è uno sbaglio in cui incorrono la maggior parte dei media: va di grande moda infatti dire che i Repubblicani ottengano il voto dei poveri, in quanto vincono negli stati mediamente più poveri. Questo è un esempio di fallacia ecologica in quanto, in realtà, in America sono i ricchi a votare di più per il partito Repubblicano, nonostante tutti i cambiamenti nel comportamento degli elettori degli ultimi decenni.

Modello con predittori macro

Un contesto in cui, invece di indurre in errore, le variabili macro arricchiscono i propri risultati, è quello dei modelli multilevel.

Come anticipato, l'obiettivo è quello di spiegare la variabilità tra i gruppi in maniera sistematica. Capire, nel caso dei dati PIRLS 2016, perché certe scuole abbiano in media dei risultati migliori di altre.

Introduciamo nella notazione la matrice dei predittori a livello macro:

- U, matrice di J righe (quante sono i gruppi). Nel caso il predittore macro sia solo uno, esso è chiamato u e assume anche in questo caso J valori.

Il modello nullo presentato inizialmente può essere quindi esteso nel seguente modo quando si vuole spiegare la variabilità delle intercette tra i gruppi:

$$y_i = \alpha_{j[i]} + \epsilon_i \qquad \text{con} \quad \epsilon_i \sim N(0, \sigma_y^2) \tag{22}$$

$$\alpha_j = \gamma_0 + \gamma_1 u_j + \eta_j \qquad \text{con} \quad \eta_j \sim N(0, \sigma_\alpha^2) \tag{23}$$

dove γ_0 rappresenta l'intercetta della regressione a livello 2 (macro) per descrivere il comportamento delle α_j, e in questo caso non rappresenta più necessariamente la media μ_α come era stata chiamata prima. γ_1 invece è interpretabile come l'incremento del valore atteso dell'intercetta in un gruppo per un incremento unitario del valore del predittore macro.

γ_1 è quindi la principale novità di questo modello: questo coefficiente serve a descrivere come e perché, in media, le intercette di alcuni gruppi siano maggiori di altri. Le α_j sono però delle variabili aleatorie, per cui possono anche discostarsi dal loro valore atteso, come già facevano rispetto a μ_α nel modello più semplice visto precedentemente.

[10] che rientrano nella categoria della regressione *single-level* classica

Nel caso dei dati PIRLS, questo coefficiente può farci riconoscere dei predittori che a livello macro possono contribuire a dei risultati migliori nel test di comprensione del testo in alcune scuole.

Esempio: i dati PIRLS 2016

Sempre con l'utilizzo dei pacchetti `lme4` e `arm`, si vogliono analizzare le differenze sistematiche tra scuole con più e meno del 25% di studenti svantaggiati.

```
fit2 <- lmer(data=ita16,formula=reading_score~disadv_school25O+
               (1|idschool))
```

```
display(fit2)
```

```
## lmer(formula = reading_score ~ disadv_school25O + (1 | idschool),
##     data = ita16)
##                  coef.est coef.se
## (Intercept)      552.03     2.64
## disadv_school25O -10.48     5.31
##
## Error terms:
##  Groups   Name        Std.Dev.
##  idschool (Intercept) 23.73
##  Residual             59.91
## ---
## number of obs: 3608, groups: idschool, 137
## AIC = 39986.7, DIC = 39996
## deviance = 39987.4
```

Il coefficiente $\gamma_1 = -10.48$ evidenzia innanzitutto come, in media, il punteggio medio nei test sia ridotto di 10 punti nelle scuole con più studenti svantaggiati (quelle dove questa quota supera il 25%) rispetto alle scuole che ne hanno meno.

Il modello, inoltre, è stato stimato solamente per le 137 scuole (sulle 149 totali) per cui era disponibile l'informazione sullo status socioeconomico dei suoi studenti u_j. Questa informazione, come si può notare di seguito, non era disponibile ad esempio per la scuola con `id=3`.

Questo modello ha, rispetto a `fit0` (modello nullo a intercetta variabile non spiegata), un'intercetta "fissa" γ_0 maggiore. Per poter ricavare tutte le intercette dei gruppi purtroppo non è più sufficiente estrarre il vettore dei coefficienti.

```
head(coef(fit2)$idschool)
```

```
##   (Intercept) disadv_school25O
## 1    530.5673        -10.47559
## 2    544.0024        -10.47559
## 4    555.7978        -10.47559
## 5    541.2226        -10.47559
## 6    557.5066        -10.47559
## 7    565.6044        -10.47559
```

In questa matrice infatti sono riportati:

- nella prima colonna: $\gamma_0 + \eta_j$
- nella seconda colonna: γ_1

Per poter ricavare dunque α_j, poiché vale

$$\alpha_j = \gamma_0 + \gamma_1 u_j + \eta_j \qquad \text{con} \quad \eta_j \sim N(0, \sigma^2_\alpha) \tag{24}$$

si deve prendere, riga per riga (scuola per scuola):

- direttamente il valore della prima colonna per quelle scuole con meno del 25% di studenti svantaggiati ($u_j = 0$)
- la somma delle due colonne per quelle scuole con più del 25% di studenti svantaggiati ($u_j = 1$)

```
head(intercette)
```

```
##   gamma0+eta0    gamma1 u_j
## 1    530.5673 -10.47559   0
## 2    544.0024 -10.47559   1
## 4    555.7978 -10.47559   1
## 5    541.2226 -10.47559   0
## 6    557.5066 -10.47559   0
## 7    565.6044 -10.47559   0
```

Nel caso delle scuole 2 e 4 c'è una presenza notevole di studenti svantaggiati, al contrario delle scuole 1, 5, 6 e 7.

Il valore delle intercette per queste scuole è pari a:

```
##    alpha_j
## 1 530.5673
## 2 533.5268
## 4 545.3222
## 5 541.2226
## 6 557.5066
## 7 565.6044
```

La presenza dell'informazione data dal predittore macro può cambiare la stima di queste intercette: le stime sono diverse da `fit0`, e ciò non dipende soltanto dal fatto che a questo modello manchino 12 scuole che avevano il valore della variabile macro mancante.

L'esistenza di modelli più avanzati

Gli sviluppi di questi modelli possono andare anche oltre a quanto visto finora. È possibile infatti che:

- si voglia far variare anche gli altri coefficienti β a livello micro oltre all'intercetta
- si voglia utilizzare una variabile *outcome* che non rispetti l'assunzione di normalità degli errori, stimando così modelli lineari generalizzati, che utilizzano le stesse funzioni link dei modelli lineari generalizzati *single-level* (logit, probit, log, e così via)
- si vogliano stabilire delle interazioni tra i coefficienti micro variabili e i coefficienti macro per spiegare sistematicamente la variabilità dei coefficienti tra i gruppi, oltre la sola intercetta
- nei dati, e nel modello, siano presenti più di soli 2 livelli di dati
- che questi livelli di dati non siano tra loro annidati (*non-nested multilevel models*)

Raccontare una storia: i dati PIRLS 2016

Di generazione in generazione

Come riporta un articolo del 2019 di OpenPolis[11], le disuguaglianze presenti a livello socioeconomico vengono, attraverso la scuola, tramandate da una generazione all'altra. C'è, infatti una differenza importante nei risultati scolastici tra i ragazzi provenienti da famiglie benestanti e con genitori istruiti e quelli provenienti invece da situazioni di maggiore disagio.

Queste differenze possono addirittura non fermarsi al solo rendimento scolastico, ma possono anche incidere sulla scelta di molti ragazzi di interrompere il loro percorso di istruzione. Fermarsi così presto può creare poi una spirale negativa che da lavori sotto-pagati vada a impattare sullo status socio-economico anche dei propri figli.

In questo senso, l'educazione primaria può avere un impatto decisivo sul futuro successo accademico di uno studente. È compito dei decisori politici capire come investire per far sì che le barriere date dallo status socioeconomico familiare, le barriere linguistiche e altre differenze nel *background* degli studenti non vadano a perpetuare queste disuguaglianze, ponendo un freno importante alla mobilità sociale. Come riporta infatti l'articolo[12] certi interventi scolastici possono avere un effetto positivo sugli studenti svantaggiati.

L'impatto del contesto

Secondo Romeo et al. ,[13] inoltre, non sono soltanto lo status socio-economico individuale e l'esposizione alla lingua parlata a impattare positivamente o negativamente la propria *performance* scolastica, ma anche la qualità della scuola.

Gli studenti più svantaggiati hanno maggiori probabilità, infatti, di frequentare scuole con meno risorse e insegnanti con meno esperienza. Questo può amplificare gli svantaggi e le disuguaglianze nei loro risultati.

La *research question*

Ci si può chiedere, dunque:

Le scuole riflettono soltanto le condizioni individuali degli studenti?

Oppure amplificano o attenuano queste disparità a seconda delle risorse e del contesto?

Attraverso un modello multilevel infatti sarà possibile capire se, oltre all'effetto dello status socio-economico individuale, ha un effetto anche lo status della scuola. Gli studenti infatti, a parità di provenienza, potrebbero ottenere dei punteggi diversi a seconda della scuola.

I dati

Dopo aver esplorato i dati, si può passare al modello multilevel. Si decide di prendere come misura dello status socio-economico degli studenti di quarta elementare a livello individuale alcune variabili, tra cui

- la quantità di libri presenti a casa loro X. La variabile X, `books2_home_stud`, è dicotomica, e assume valore 1 se gli studenti hanno a casa loro almeno 26 libri (abbastanza per riempire una libreria)
- la presenza di Internet a casa per studiare o meno (`internet_stud`)

[11] https://www.openpolis.it/gli-studenti-svantaggiati-e-le-disuguaglianze-educative-a-scuola/
[12] https://www.tandfonline.com/doi/epdf/10.1080/01443410420002717 46?needAccess=true
[13] https://pmc.ncbi.nlm.nih.gov/articles/PMC9588575/pdf/nihms-1826606.pdf

- se a casa l'italiano si parla sempre o quasi sempre (`lang2_often_stud`)

A livello aggregato, invece, si raggruppano le scuole in 3 gruppi, sfruttando 2 variabili presenti nel dataset originale:

- le scuole con più del 25% di studenti svantaggiati e meno del 25% di benestanti (scuole con più svantaggiati, `most_disadvantaged`)
- le scuole con la stessa quantità di studenti svantaggiati e benestanti, entrambe al di sopra del 25%, `equal_extremes`
- le scuole con la stessa quantità di studenti svantaggiati e benestanti, entrambe al di sotto del 25%, `balanced`
- le scuole con meno del 25% di studenti svantaggiati e più del 25% di benestanti (scuole con più benestanti, `most_affluent`)

```
##
##      most_affluent           balanced     equal_extremes most_disadvantaged
##               1300               1423                269                595
```

Sono risultati quindi 4 tipi di contesti scolastici. Ci possiamo chiedere come incidano sui punteggi medi all'interno delle varie scuole. Per rispondere a questa domanda introduciamo la variabile `ses_school` all'interno del modello a intercetta variabile.

Modello multilevel con soli predittori macro

```
fit.ineq1 <- lmer(data=ita16,
                  formula=reading_score~ses_school+(1|idschool))
display(fit.ineq1)
```

```
## lmer(formula = reading_score ~ ses_school + (1 | idschool), data = ita16)
##                              coef.est coef.se
## (Intercept)                  555.11     3.99
## ses_schoolbalanced            -5.44     5.29
## ses_schoolequal_extremes     -17.75     9.33
## ses_schoolmost_disadvantaged -13.96     6.84
##
## Error terms:
##  Groups   Name        Std.Dev.
##  idschool (Intercept) 23.50
##  Residual             59.98
## ---
## number of obs: 3587, groups: idschool, 136
## AIC = 39751, DIC = 39779.5
## deviance = 39759.2
```

Notiamo come le scuole con più individui provenienti da contesti svantaggiati (sia che abbiano o che non abbiano molti studenti benestanti) hanno dei punteggi medi più bassi di oltre 13 punti rispetto alle scuole con uno status socio-economico medio migliore. Tuttavia, senza controlli individuali, queste differenze non sono ancora attribuibili all'effetto del contesto scolastico, ma potrebbero anche dipendere dai differenti background individuali anche all'interno delle stesse scuole.

Modello multilevel completo

Nelle scuole più svantaggiate, infatti, ci saranno sicuramente studenti con uno status socio-economico più basso rispetto alle scuole `most_affluent`, prese come riferimento dal modello.

Inserendo come controlli individuali:

- l'età dello studente (che ha un leggero effetto sui risultati, ma che è al di fuori della nostra domanda di ricerca)
- il genere dello studente,

come misure della provenienza individuale degli studenti:

- la presenza di almeno 26 libri in casa
- la disponibilità di Internet per studiare a casa
- il parlare sempre o quasi sempre in italiano a casa

e come controlli macro:

- la zona in cui si trova la scuola (rurale o meno)

risulta il seguente modello

```
fit.ineq2 <- lmer(data=ita16,
                  formula=reading_score~rescale(age_stud)+male_stud+
                    books2_home_stud+internet_stud+lang2_often_stud+
                    ses_school+
                    rural+(1|idschool))
display(fit.ineq2)
```

```
## lmer(formula = reading_score ~ rescale(age_stud) + male_stud +
##     books2_home_stud + internet_stud + lang2_often_stud + ses_school +
##     rural + (1 | idschool), data = ita16)
##                               coef.est coef.se
## (Intercept)                   511.64     5.43
## rescale(age_stud)               6.71     2.02
## male_stud                      -7.63     1.96
## books2_home_stud               20.60     2.04
## internet_stud                   7.54     2.70
## lang2_often_stud               34.13     2.71
## ses_schoolbalanced             -4.53     4.83
## ses_schoolequal_extremes       -7.50     8.37
## ses_schoolmost_disadvantaged -10.40     6.07
## rural                           1.94     4.42
##
## Error terms:
##  Groups   Name        Std.Dev.
##  idschool (Intercept) 20.15
##  Residual             57.47
## ---
## number of obs: 3526, groups: idschool, 136
## AIC = 38737.1, DIC = 38795.8
## deviance = 38754.5
```

I coefficienti macro sullo status socio-economico della scuola sono cambiati notevolmente in questo caso. Attraverso il modello si può riconoscere che parte delle differenze tra le scuole in contesti più svantaggiati e quelle in contesti più benestanti dipendeva dai predittori individuali. Il punteggio atteso di un individuo baseline è più basso di circa 10 punti nella scuola nel contesto peggiore.

Inoltre, le scuole con più del 25% di studenti svantaggiati e più del 25% di studenti benestanti non sono più mediamente le peggiori, ma in media gli individui baseline che studiano lì hanno dei punteggi

migliori di soli 3 punti delle scuole peggiori.

Le caratteristiche del proprio background familiare incidono comunque in maniera più forte (libri, Internet e lingua parlata a casa), *ceteris paribus*. Tuttavia, tali differenze, che sarebbero già notevoli tra individui di una stessa scuola, vengono addirittura ampliate se gli studenti provengono da scuole in contesti differenti: gli studenti individualmente svantaggiati che frequentano scuole svantaggiate possono avere dei risultati ancora peggiori di quelli che hanno individui identici a loro per gli altri predittori ma che frequentano scuole dove vanno molti ragazzi benestanti.

Conclusioni

Dare accesso, anche nei contesti più svantaggiati, a biblioteche, connessione a Internet per studiare, predisporre e finanziare dei programmi di insegnamento della lettura, sono tutte iniziative che possono essere fondamentali nello spezzare quella spirale negativa delle disuguaglianze e di mancanza di opportunità per gli individui che, oltre a subire l'effetto negativo del basso status socio-economico della propria famiglia, frequentano anche dei contesti che amplificano le loro difficoltà.

Bibliografia

Gelman, Andrew. Data analysis using regression and multilevel/hierarchical models. Cambridge university press, 2007.

SOURCE: PIRLS 2016 Assessment Frameworks. Copyright © 2015 International Association for the Evaluation of Educational Achievement (IEA). Publisher: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Bennett, Neville, et al. "Teaching styles and pupil progress." (1976): 335-350.

Aitkin, Murray, Dorothy Anderson, and John Hinde. "Statistical modelling of data on teaching styles." Journal of the Royal Statistical Society Series A: Statistics in Society 144.4 (1981): 419-448.

Blanchard, M. (2023, January 5). The relationship between socioeconomic status and literacy: How literacy is influenced by and influences SES. Michigan Journal of Economics. https://sites.lsa.umich.edu/mje/2023/01/05/the-relationship-between-socioeconomic-status-and-literacy-how-literacy-is-influenced-by-and-influences-ses/

Redazione. (2022, January 26). Gli studenti svantaggiati e le disuguaglianze educative a Scuola. Openpolis. https://www.openpolis.it/gli-studenti-svantaggiati-e-le-disuguaglianze-educative-a-scuola/

D'Angiulli *, A., Siegel, L. S., & Hertzman, C. (2004). Schooling, Socioeconomic Context and Literacy Development. Educational Psychology, 24(6), 867–883. https://doi.org/10.1080/0144341042000271746

Rosche Benjamin (2022). Multilevel models: when to use them, how they differ from OLS regression, and how to implement them in Stata and R benrosche.com

McElreath, Richard. Statistical rethinking: A Bayesian course with examples in R and Stan. Chapman and Hall/CRC, 2018.