Analisi dati longitudinali

May 8, 2025

Contents

Chapter 1

Introduzione

Nelle scienze sociali, psicologiche e biomediche è frequente che i dati abbiano una struttura gerarchica, anche su più livelli, dovute a

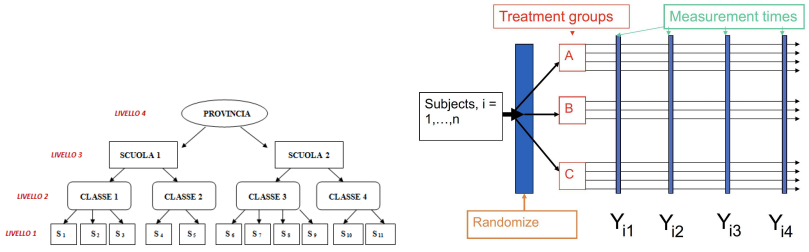
- **appartenenza ad un gruppo** (istituzionali, geografici, ...) o cluster (es rispondenti in un campionamento a più stadi)
- **Misure ripetute** per ogni individuo (es curve di crescita della glicemia in momenti diversi della giornata)

Example 1.0.1. Alcuni esempi 1.1

- Scienza dell'educazione: lo studio dell'efficacia dei sistemi educativi scolastici analizza la struttura formata da alunni, classi, scuole;
- Studi Epidemiologici: un gruppo di soggetti sono seguiti nel tempo e rilevate periodicamente informazioni su fattori di rischio/protettivi ed esiti di salute.

Important remark 1. I dati non sono più indipendenti. Come trattare questi dati?

Important remark 2. Quale tipo di disegno dello studio può generare dati con queste caratteristiche? Distinguiamo tra disegni



(a) (b)

Figure 1.1: Esempi dati gerarchici

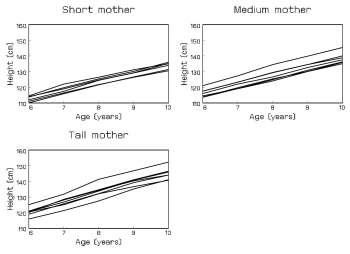


Figure 1.2: Esempio bambine

- *cross-sectional*: Nei dati cross-section ci può essere correlazione all'interno dei gruppi: la struttura di errore è complessa e rispecchia la gerarchia dei dati. Al fine di esplorare la variabilità tra gruppi si utilizzano modelli multilevel o - sinonimi - ad effetti misti, a coefficienti casuali, ad effetti casuali, a componenti di varianza.
  Occorre capire se la struttura dei dati permette un'applicazione multilevel oppure capire quando la struttura dati esige un'impostazione di tipo multilevel

- *longitudinali* che portano a dati Panel (misure ripetute): es una caratteristica è misurata più volte su una stessa unità. Es: pressione sistolica tutti i giorni per una settimana.
  Obiettivi qui sono misurare il cambiamento nel tempo intra-individuale e tra individui

Example 1.0.2. Statura di 20 bambine divise in base alla statura della madre (fig 1.2)

	Statura Madre & Numero bambina
bassa	<155 cm & 1->6
media	[115; 164] & 7->13
alta	>164 & 14->20

Domanda di ricerca:

- la crescita in statura della figlia è legata alla statura della madre?
- descrivere la traiettoria di ciascuna unità: come la variabile oggetto di studio (outcome) cambia nel tempo
- prevedere i cambiamenti
- valutare l'effetto di covariate

Dall'immagine (misure riferite a tempi fissi, non necessariamente cosi) si vede una relazione lineare quasi perfetta maggiore variabilità tra gruppi minor variabilità entro i gruppi stesso numero di misure per soggetto

Ipotesi del modello di regressione lineare classico

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + e_i \quad i = 1, \ldots, n$$

con ipotesi

- Errori con valore atteso pari a 0

$$E(e_i|x_{1i}, x_{2i}, \ldots, x_{ki}) = 0$$

- Errori incorrelati

$$cov(e_i, e_j) = 0, \quad \forall i \neq j$$

- Errori omoschedastici

$$var(e_i) = var(e_i|x_{1i}, x_{2i}, \ldots, x_{ki}) = \sigma^2$$

-

Chapter 2

Multilevel models

2.1 Esempio

Example 2.1.1. Dataset sulla performance in un test di matematica; le unità

- di **primo livello**: sono gli studenti. Livello micro/within
- di secondo livello: la scuola di appartenenza. Livello macro/between

```
library(tidyverse)

## - Attaching core tidyverse packages ------------------------------------------------------
tidyverse 2.0.0 -
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.4
## - Conflicts ------------------------------------------------------------------------------
tidyverse_conflicts() -
## x purrr::compose()     masks lbmisc::compose()
## x lubridate::day()     masks lbmisc::day()
## x dplyr::filter()      masks stats::filter()
## x purrr::flatten()     masks lbmisc::flatten()
## x lubridate::is.Date() masks lbmisc::is.Date()
## x dplyr::lag()         masks stats::lag()
## x lubridate::month()   masks lbmisc::month()
## x dplyr::recode()      masks lbmisc::recode()
## x tibble::view()       masks lbmisc::view()
## x readr::write_lines() masks lbmisc::write_lines()
## x lubridate::year()    masks lbmisc::year()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to
force all conflicts to become errors

library(haven)
```

```
##
## Caricamento pacchetto:  'haven'
##
## Il seguente oggetto è mascherato da 'package:lbmisc':
##
##     write_sas
```

```
library(lattice)
```

```
if (FALSE){
  fL <- "https://stats.idre.ucla.edu/stat/examples/imm/imm10.dta"
  dta <- read_dta(fL)
  write.csv(dta, file= "longitudinal_data_analysis/data/imm10.csv", row.names = FALSE
} else {
  db <- read.csv("longitudinal_data_analysis/data/imm10.csv")
}
```

```
## Warning in file(file, "rt"):  non è possibile aprire il file 'longitudinal_data_an
File o directory non esistente
## Error in file(file, "rt"):  non è possibile aprire la connessione
```

```
## id scuola
db$schid <- factor(db$schnum) # schnum è un progressivo scuola da 1 a 10
```

```
## Error in factor(db$schnum):  oggetto 'db' non trovato
```

Le variabili

- Livelli: id studente (livello 1) `stuid`; scuole (livello 2) `schnum`
- Variabile risposta Y: `math` punteggio in un test di matematica
- Covariata del livello 1: `homework` ore settimanali di studio a casa
- Covariata del livello 2 `public`: variabile binaria che indica scuola pubblica verso scuola private

In linea teorica le variabili del secondo livello (macro/gruppo) si possono distinguere in

- *Globali*: sono caratteristiche intrinseche delle unità del secondo livello (es. tipo di scuola, zona in cui si colloca la scuola)
- *Contestuali*: indicatori macro ottenuti come sintesi di valori individuali (es. proporzione di maschi/femmine, livello medio dello stato socio economico (SES))

Alcuni casi:

```
head(db)
```

```
## Error in head(db):  oggetto 'db' non trovato
```

Abbiamo 10 scuole (unità di livello 2) di differenti dimensioni (disegno non bilanciato). Il numero di studenti complessivo (unità di livello 1) è 260. Alcune statistiche descrittive

```
##  numerosità studenti nelle 10 scuole. La 7 è la più diversa
table(db$schnum)

## Error in table(db$schnum):  oggetto 'db' non trovato

## n medio studenti per scuola: molto influenzato da 67 di 7
mean(table(db$schnum))

## Error in table(db$schnum):  oggetto 'db' non trovato

## distributzione bambini per appartenenza a scuola pubblica o privata.
## bimbi in scuola privata sono 67, 193 in pubblica
table(db$public)

## Error in table(db$public):  oggetto 'db' non trovato

## statistiche descrittive raggruppando per scuola (n e media) su variabili
## math, homework e public
## si vede che la scuola privata ha un punteggio più alto e un umero di ore
## studiate a casa più alto
db %>%
  group_by(schnum) %>%
  summarise_at(vars(math, homework, public), funs(n(), mean(., na.rm=TRUE)))

## Error in group_by(., schnum):  oggetto 'db' non trovato

## come creare variabili di gruppo / scuola (usando dati individuali
## aggiungere il livello medio status socio economico per scuola
## (variabile aggiunta in fondo)
db %>%
  group_by(schnum) %>%
  mutate(mean_se=mean(ses)) %>%   #aggiunge variabili al dataset
  ungroup

## Error in group_by(., schnum):  oggetto 'db' non trovato

## invece come collassare il gruppo ad un unico valore: esse la scuola è pubblica o privata
## (faccio una media di una costante)
db %>%
  group_by(schnum) %>%
  summarize(mean(public)) %>%   #aggiunge variabili al dataset
  ungroup

## Error in group_by(., schnum):  oggetto 'db' non trovato
```

Obiettivo dello studio è il confronto tra le diverse scuole e nello specifico studiare l'effetto di valutando il punteggio `math` in funzione delle ore di studio `homework`. Ad esempio se la situazione fosse quella mostrata in figura 2.1 abbiamo che

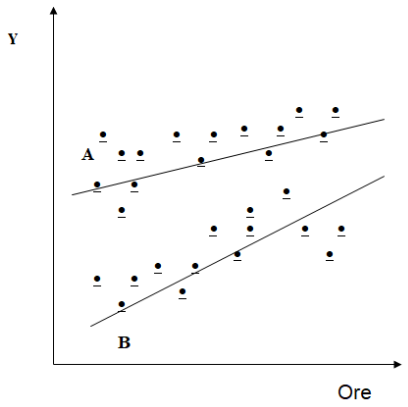


Figure 2.1: Relazione ore math ipotetica tra scuole

- variabilità tra scuole: scuola A in media meno preparati ha un livello superiore alla seconda in termini di performance su math
- correlazione tra fenomeni: in A la variabile ore di studio è meno predittiva dell'outcome

Remark 1.

Important remark 3. Si può scegliere di analizzare i dati con metodi "classici"

- a livello individuale (es. studente), una analisi disaggregata
- a livello di gruppo (es. scuola); si svolge una analisi aggregata (pooled) su dati/medie di gruppo

Ma entrambe queste scelte danno luogo a dei problemi

Important remark 4 (Problemi analisi disaggregata). Si ha

- **Dipendenza**: viene violata l'ipotesi di indipendenza tipica dei metodi tradizionali: le osservazioni all'interno di un gruppo sono fra loro più simili rispetto a quelle di altri gruppi, per cui si ha una correlazione positiva all'interno dei gruppi.
  Se si adottano metodi tradizionale si ha un'errata stima degli errori standard (spesso si ha una sottostima degli errori standard: quindi errori del I tipo più alti del livello nominale α)
- **Inferenza sui gruppi**: non è possibile fare inferenza sui gruppi, cioè trattare i gruppi osservati come un campione casuale da una popolazione di gruppi

Errata dimensione campionaria delle variabili di livello 2 (inferiore ad n)

Important remark 5 (Problemi della analisi analisi aggregata). Si ha

- **Shift of meaning**: le variabili aggregate si riferiscono al gruppo e non all'individuo, per cui non possono nemmeno concettualmente essere usate per indagare le relazioni a livello di individuo

- **Ecological fallacy** (distorsione da aggregazione): le relazioni a livello di gruppo (cioè tra le medie di gruppo) sono diverse dalle corrispondenti relazioni a livello individuale
- **Interazione tra livelli**: l'analisi aggregata non consente di studiare le relazioni tra livelli gerarchici

Important remark 6 (Problemi nell'uso del modello classico di regressione lineare).

- **Dipendenza**: le osservazioni all'interno di un gruppo sono fra loro più simili rispetto a quelle di altri gruppi. La violazione dell'ipotesi di unità indipendenti porta ad errori del I tipo molto più alti del livello nominale α
- **Annidamento gerarchico**: quando gli individui sono annidati all'interno di gruppi. Si hanno due sorgenti di variabilità: all'interno dei gruppi e tra i gruppi.
- **Interazione tra livelli**: si vogliono considerare interazioni tra variabili esplicative definite a differenti livelli di struttura gerarchica.

Remark 2 (Integrazione delle dimensioni micro e macro della ricerca, dei livelli di osservazione e di analisi).

- Che parte della variabilità nei comportamenti è imputabile al contesto?
- Come agiscono le componenti macro sulle relazioni individuali?
- Quale effetto produce una struttura gerarchica nei dati sulla variabile risposta?
- Che relazione c'è tra l'intensità dei fenomeni a livello aggregato e i modelli di comportamento individuale che li determinano?
- Quali analogie si possono rintracciare nelle relazioni a livello aggregato e disaggregato tra le medesime variabili?

Important remark 7 (Diversi approcci).

- **Modello singolo**: si ignora la struttura Svantaggi: ignorare la dipendenza tra unità (le stime hanno errore standard più basso che determina conclusioni inferenziali non corrette $\Longrightarrow$ intervalli di confidenza più piccoli e rischio alto di errore di I tipo
- **Modello a effetti fissi**: inclusione di un insieme di variabili Dummy per individuare i gruppi Svantaggi: se i gruppi sono tanti si ha un numero elevato di parametri da stimare
- **Modello marginale**: la dipendenza è modellata direttamente Svantaggi: non si è in grado di valutare la variabilità tra gruppi (lo riprenderemo)

I **Modelli multilevel** permettono di avere una stima corretta dell'errore standard e permettono di valutare la varianza tra gruppi

Important remark 8. Ricorriamo ad un approccio multilevel:

1. perché la sola analisi individuale assume indipendenza tra le osservazioni - assunzione non più valida quando si considerano informazioni a livello aggregato (scolastico, territoriale ecc.) e dunque una implicita struttura in cluster - e ignora la struttura dei dati, portando problemi alle stime;

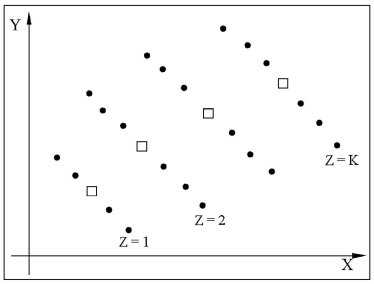


Figure 2.2: Ecological fallacy

2. perché la sola analisi a livello di gruppo fornisce risultati quasi mai verificabili a livello individuale; ignora la struttura dati e comporta perdita di informazione.

Problemi metodologici:

1. **Atomistic Fallacy**: problema in cui si incorre quando si formulano inferenze su un livello della gerarchia basandosi su analisi realizzate a un livello inferiore; si fanno ad esempio inferenze riguardanti associazioni a livello di gruppo mediante associazioni a livello individuale.
   In tal modo non si considera che i fattori che spiegano la variabilità tra individui all'interno dei gruppi non sono necessariamente gli stessi che spiegano la variabilità tra i gruppi, oppure non agiscono nel medesimo modo.
2. **Ecological Fallacy** (fig 2.2): consiste nell'interpretare dati aggregati come se fossero dati individuali. Si fanno inferenze riguardanti il livello individuale sulla base dei dati inerenti il livello di gruppo, considerando cioè aggregazioni a livello del gruppo cui gli individui appartengono; in tal modo si utilizza la correlazione tra variabili a livello di gruppo per fare affermazioni su relazioni di livello micro

2.2 Modelli multilevel

Assunzioni:

- struttura gerarchica
- una sola variable dipendente misurata al livello più *basso*
- variabili esplicative a tutti i livelli

Si implementa mediante un sistema gerarchico di equazioni di regressione

2.2.1 Modello di regressione singolo per la media

Definito dall'equazione

$$y_i = \beta_0 + \epsilon_i$$

e rappresentato dal grafico **??** dove

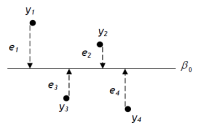


Figure 2.3: Modello di regressione singolo per la media

- i indice dell'unità ($i = 1, \ldots, n$)
- β_0: media di y nella popolazione
- residuo dell'i-esimo individuo, ossia differenza del valore di y con la media di popolazione

2.2.2 Modello a effetti casuali non condizionato

Definito dall'equazione

$$y_{ij} = \beta_{0j} + \epsilon_{ij}$$

dove

- Y: variabile risposta riferita unità di primo livello
- $i = 1, \ldots n_j$ unità di primo livello totale unità = N
- $j = 1, \ldots J$ sono gli id delle unità di secondo livello (es scuole)
- β_{0j} è la media di Y nel gruppo j
- ϵ_{ij} è l' errore residuo, errore entro gruppi legato alla variabilità individuale

Assunzioni:

- L'intercetta non è costante ma cambia da gruppo a gruppo pertanto è detta intercetta casuale. Si ha quindi

  $$\beta_{0j} = \gamma_{00} + u_{0j}$$

  dove

  - γ_{00} è la media generale (globale/complessiva)
  - u_{0j} è il termine d'errore riferito alla differenza che esiste tra media di gruppo e la media generale

- le assunzioni sulle componenti sono

  $$\begin{aligned} &\epsilon_{ij} \sim \mathrm{N}\left(0, \sigma_e^2\right) \\ &u_j \sim \mathrm{N}\left(0, \sigma_u^2\right) \\ &e_{ij} \perp\!\!\!\perp u_j, \quad foralli, j \end{aligned}$$

Per questo modello abbiamo una **scomposizione di variabilità** ed un indice derivante in quanto

$$\text{Var}\left[y_{ij}\right] = \text{Var}\left[\gamma_{00} + u_{0j} + \epsilon_{ij}\right] = \sigma^2_{u_0} + \sigma^2_{\epsilon}$$

Data la somma possiamo definire il seguente indice, interpretabile come una percentuale, detto **coefficiente di correlazione intraclasse** (ICC) o anche coefficiente di partizione della varianza (VPC)

$$\rho = \text{Corr}\left(y_{ij}, y_{i'j}\right) = \frac{\sigma^2_u}{\sigma^2_u + \sigma^2_{\epsilon}} = \frac{\text{cluster variance}}{\text{total variance}}, \quad \rho \in [0,1]$$

Alcune considerazioni:

- All'aumentare del coefficiente di correlazione intraclasse aumenta il contributo esplicativo dovuto alla strutturazione gerarchica. Questo coefficiente fornisce una misura dell'omogeneità all'interno di uno stesso gruppo e rappresenta la proporzione di varianza spiegata dal raggruppamento; misura quindi la parte di variabilità è dovuta all'effetto di raggruppamento e quella derivante dalla dipendenza tra osservazioni raggruppate in unita dello stesso livello.
- ρ rappresenta una misura che giustifica il ricorso al modello gerarchico. Un valore del coefficiente molto basso, infatti, non segnalando la presenza di correlazione all'interno dei gruppi, suggerisce di evitare la modellizzazione a piu livelli e di ricorrere ai tradizionali modelli regressivi ad un solo livello

Example 2.2.1. La varianza complessiva è

$$30.54 + 72.24 = 102.78$$

ICC è

$$ICC = 30.54/102.78 = 0.297$$

29,7% della varianza del punteggio in matematica è dovuta al raggruppamento nelle diverse scuole.

2.2.3 Modello condizionato: introduzione di variabili indipendenti

L'introduzione di variabili esplicative (indipendenti) per spiegare meglio la variabilità del fenomeno oggetto di studio (Y), tenendo in considerazione la struttura gerarchica dei dati, fa si che varianza "tra" e varianza "entro" si modificano e la differenza rispetto al modello non condizionato permette di valutare il contributo delle variabili esplicative.
Di solito si procede per passi introducendo variabili al primo livello, poi al secondo etc..

2.2.3.1 Intercetta casuale e pendenza fissa

Introduciamo un regressore di primo livello X e ipotizziamo che il coefficiente di regressione (pendenza) sia costante per ogni gruppo. I modelli di livello 1 e

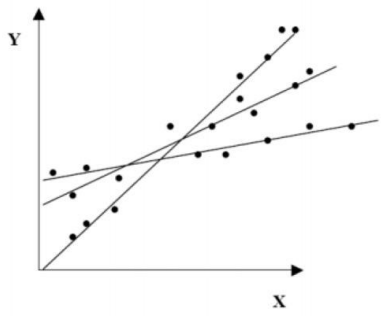


Figure 2.4: Intercetta e pendenza random

2 sono rispettivamente

$$y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \epsilon_{ij} \quad \text{mod.livello 1} \quad \begin{cases} \beta_{0j} = \gamma_{00} + u_{0j} \\ \beta_1 = \gamma_{10} \end{cases} \quad \text{mod.livello 2}$$

il primo è il modello di livello 1 (a livello di individuo), il secondo di livello 2 (a livello di gruppo), dove

- γ_{00} Parte fissa dell'intercetta: media complessiva della variabile Y a livello di popolazione
- u_{0j} Scostamento dell'intercetta del j-esimo gruppo dalla media generale
- γ_{10} Pendenza fissaàeffetto medio della variabile X sulla variabile dipendente

Per avere una unica equazione, il **modello combinato** è

$$y_{ij} = \underbrace{\gamma_{00} + \gamma_{10} x_{ij}}_{\text{parte fissa}} + \underbrace{u_{0j} + \epsilon_{ij}}_{\text{parte casuale}}$$

Le ipotesi sugli errori di primo e secondo livello sono le stesse del modello non condizionato (modello vuoto o nullo). Si aggiunge anche l'ipotesi di indipendenza di tutti i termini di errore delle variabili esplicative. Si ha quindi

$$\text{Var}\left[y_{ij}|x_{ij}\right] = \sigma^2_{u_0} + \sigma^2_{\epsilon}$$

Si ha che

- La varianza tra gruppi è costante rispetto ai valori della variabile esplicativa
- Si può calcolare ancora il coefficiente di correlazione intraclasse

2.2.3.2 Intercetta e pendenza casuali

Il

- modello livello 1

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \epsilon_{ij}$$

- modello livello 2

$$\beta_{0j} = \gamma_{00} + u_{0j}$$
$$\beta_{1j} = \gamma_{10} + u_{1j}$$

- modello combinato

$$y_{ij} = \underbrace{\gamma_{00} + \gamma_{10} x_{ij}}_{\text{parte fissa}} + \underbrace{u_{1j} x_{ij} + u_{0j} + \epsilon_{ij}}_{\text{parte casuale}}$$

rappresentanto in fig 2.4

- gli errori di secondo livello (**effetti casuali**):

$$\begin{cases} u_{0j} = \beta_{0j} - \gamma_{00} & \text{Var}\left[(\right] u_{0j}) = \sigma^2_{u_0} \\ u_{1j} = \beta_{1j} - \gamma_{10} & \text{Var}\left[(\right] u_{1j}) = \sigma^2_{u_1} \end{cases}$$

Gli *effetti casuali* consistono la differenza non spiegata del parametro dalla media di popolazione del j-esimo cluster.
L'introduzione di covariate può aiutare a ridurre queste varianze. Di solito le assunzioni distributive si riferiscono agli effetti casuali piuttosto che ai parametri intercetta e coefficiente

$$\begin{aligned} &\varepsilon_{ij} \sim iid\, \text{N}\left(0, \sigma^2_\epsilon\right) \\ &\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim iid\, \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u_0} & \sigma_{u_{01}} \\ \sigma_{u_0 1} & \sigma^2_{u_1} \end{bmatrix}\right) \\ &\begin{bmatrix} \beta_{0j} \\ \beta_{1j} \end{bmatrix} iid\, \text{MVN}\left(\begin{bmatrix} \gamma_{00} \\ \gamma_{10} \end{bmatrix}, \begin{bmatrix} \sigma^2_{u_0} & \sigma_{u_{01}} \\ \sigma_{u_0 1} & \sigma^2_{u_1} \end{bmatrix}\right) \\ &\epsilon_{ij} \perp\!\!\!\perp \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \end{aligned}$$

con, la normalità è una possibile assunzione e

  - parametri fissi
    - γ_{00} la media delle intercette
    - γ_{10} la media dei coefficienti di regressione
  - parametri causali
    - $\sigma^2_{u_0}$ varianza dell' intercetta
    - $\sigma^2_{u_1}$ varianza del coefficiente di regressione
    - σ^2_ϵ varianza residua del livello 1
    - $\sigma_{u_0 1}$ Covarianza intercetta-coefficiente:un valore positivo implica che gruppi con alto valore del residuo u_{0j} tendo ad avere alti valori dei residui per la pendenza u_{1j}

    questi parametri sono quantità fisse, la casualità è legata alla parte casuale del modello

TODO: quali

  - L'errore totale è $u_{0j} + u_{1j} x_{ij} + \epsilon_{ij}$ e implica **eteroschedasticità**: la varianza tra gruppi è funzione della variabile esplicativa

$$\text{Var}\left[(\right] y_{ij} | x_{ij}) = \sigma^2_{u_0} + 2\sigma_{u_{01}} x_{ij} + \sigma^2_{u_1} x^2_{ij} + \sigma^2_\epsilon$$

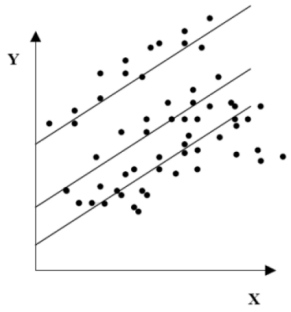


Figure 2.5: Intercetta casuale

Example 2.2.2 (un modello di regressione per ogni scuola)**.** Il modello è

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \epsilon_{ij}$$

il modello contiene:

- una intercetta β_{0j} per ogni scuola
- un coefficiente β_{1j} per ogni scuola
- termine di errore ϵ_{ij} con $\mathbb{E}[\epsilon_{ij}] = 0$ e $\text{Var}[\epsilon_{ij}] = \sigma_j$

assunzioni

- σ_j uguali per tutte le scuole
- intercette e coefficienti variano tra le scuol (coefficienti causali)

Nel caso di regressione standard ciascuna suola ha stessa intercetta e pendenza: una unica retta. La stima diviene

```
# TODO
```

IL modello a intercetta casuale

$$y_{ij} = \gamma_{00} + \gamma_{10} x_{ij} + u_{0j} + \epsilon_{ij}$$

È un caso speciale del modello generale dove:

- la varianza covarianza de sta ceppa

$$\mathbf{\Sigma}_u = \begin{bmatrix} \sigma_{u_0} & 0 \\ 0 & 0 \end{bmatrix}$$

- la varianza del coefficeinte di regressione è nulla (e cosi anche la covarainza intercetta-coefficiente)
- la varianza dell'intercetta non dipende da x
- vi è omoschedasticità in quanto

$$\text{Var}[y_{ij}|x_{ij}] = \sigma^2_{u_0} + \sigma^2_{\epsilon}$$
$$\text{Cov}(y_{ij}, y_{i'j}|x_{ij}, x_{i'j}) = \sigma^2_{u_0}$$

- le rette sono parallele e i cluster ordinabili (fig 2.5)

La stima è
con

- l'intercetta, media delle intercette nella popolazione delle scuole
- il coefficiente di homework: ciascuna scuola ha lo stesso coefficiente
- varianza totale è $22.5 + 64.26 = 86.76$
- $ICC = 22.5/86.76 = 0.259$: il 25.9% della varianza del punteggio in matematica dopo aver tenuto conto dello studio a casa è dovuta al raggruppamento dei ragazzi nelle diverse scuole

Nel **modello con intercetta e pendenza casuali**, il caso generale si ha

- $$y_{ij} = \gamma_{00} + \gamma_{10}x_{ij} + u_{1j}x_{ij} + u_{0j} + \epsilon_{ij}$$
$$\mathbf{\Sigma}_u = \begin{bmatrix} \sigma^2_{u0} & \sigma_{u01} \\ sigma_{u01} & \sigma^2_{u1} \end{bmatrix}$$
- La varianza tra le scuole è funzione quadratica di X
- Correlazione eterogenea entro i cluster
- La varianza dell'intercetta e la covarianza intercetta-coefficiente dipendono da X

nel modello vi è incrocio delle rette, i cluster non sono ordinabili.
La stima
dove

- la pendenza di homework è una media del coefficiente nella popolazione
- l'intercept variance è 61.81
- l'homework variance è 19.98
- l'intercept-homework covariance è -28.26, corrisponde ad una correlazione di circa -0.8
- la residual lev 1 variance è 43.07
- *non si può calcolare ICC*

2.2.3.3 Introduzione covariata di secondo livello

L'introduzione covariata di secondo livello fa si che si introducano caratteristiche
Le covariate a livello 2 rappresentano caratteristiche dei cluster utili per:

- Definire un modello per i parametri del livello 1 (β_{0j}, β_{1j})
- e quindi ridurre la varianza del livello 2 ($\sigma^2_{u0}, \sigma^2_{u1}$)

Example 2.2.3. Nell'esempio

- 58.06 è l'intercetta media delle scuole private
- 1.94 è il coefficiente medio di homework
- -14.65 è la differenza nell'intercetta (public vs private)

La covariata di secondo livello agisce solo sulla intercetta (media)

2.2.3.4 Confrontoa tra diversi modelli

```
AIC & BIC & logLik & deviance & df.resid \\
M0 & 1880.8 & 1891.5 & -937.4 & 1874.8 & 257 \\
M1 & 1850.7 & 1864.9 & -921.3 & 1842.7 & 256 \\
M2 & 1781.4 & 1802.7 & -884.7 & 1769.4 & 254 \\
M3 & 1764.8 & 1789.8 & -875.4 & 1750.8 & 253 \\
```

dove:

-
- M0 modello non condizionato
- M1 modello con intercetta casuale
- M2 modello con intercetta e pendenza casuale
- M3 modello M2 con aggiunta di covariata al livello 2 che agisce solo sulla intercetta come da formulazione di sotto

Formulata in termini di:

- modello di livello 1:

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \epsilon_{ij}$$

- modelli di livello 2

$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01} w_j + u_{0j}, & \text{Var}\left[u_{0j}\right] = \sigma^2_{u0} \\ \beta_{1j} = \gamma_{10} + \gamma_{11} w_j + u_{1j}, & \text{Var}\left[u_{1j}\right] = \sigma^2_{u1} \end{cases} \quad 0$$

- modello combinato

$$y_{ij} = \underbrace{\gamma_{00} + \gamma_{01} w_j + \gamma_{10} x_{ij} + \gamma_{11} w_j x_{ij}}_{\text{parte fissa}} + \underbrace{u_{0j} + u_{1j} x_{ij} + e_{ij}}_{\text{Parte casuale}}$$

dove:

- γ_{00} è il valore medio di y quando sia x sia w sono pari a zero
- γ_{01} è l'effetto di effetto di w e indica la variazione dell'intercetta media all'aumentare unitario di w
- γ_{10} è effetto di X su Y quando W è uguale a zero

- γ_{11} coefficiente di interazione cross-level e indica la variazione della pendenza media all'aumentare di W $\implies$ effetto di X su Y all'aumentare di W $\implies$ effetto moderatoredi W sulla relazione tra X e Y
- C'è una combinazione tra livelli $\gamma_{11} w_j x_{ij}$

Example 2.2.4. Nella stima:

- 59.21 è Intercetta media scuole private
- 1.09 è Coefficiente medio (riferimento scuole private)
- -15.94 è Differenza nell'intercetta (pubblico vs. privato)
- 0.95 è Differenza nel coefficiente (pubblico vs. privato)

Qui la covariata di secondo livello agisce sulla intercetta (media) e sul coefficiente (medio)

2.3 Stima

I metodi di stima disponibili

- massima verosimiglianza: Full information (FIML) o Restricted (REML)
- Minimi quadrati generalizzati -Generalized Least square (GLS)
- Equazioni di stima generalizzatie- Generalize Estimating equation (GEE)
- Inferenza bayesiana

2.3.1 Metodi di massima verosimiglianza

Il modello in forma generale/estesa

$$y_{ij} = \underbrace{\gamma_{00} + \gamma_{01} w_j + \gamma_{10} x_{ij} + \gamma_{11} w_j x_{ij}}_{\text{Parte fissa}} + \underbrace{u_{0j} + u_{1j} x_{ij} + \epsilon_{ij}}_{\text{parte casuale}}$$

viene stimato mediante massima verosimiglianza in due passi:

1. si inizia dalla stima del *parametri fissi* ($\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11}$ e dei parametri di *varianza e covarianza* ($\sigma^2_\epsilon, \sigma^2_{u0}, \sigma^2_{01}, \sigma^2_{u0}$)
2. si va alla *previsione* degli effetti casuali (u_{0j}, u_{1j}, $j = 1, \ldots, J$)

Per quanto riguarda **FIML vs REML**:

- con FIML
  - vi è una stima contiunta dei parametri fissi e casuali
  - si ha una sottostima dei parametri casuali perchè quelli fissi vengono considerati come quantità nota (si ignorano i gradi di libertà)
- con REML

  - I parametri casuali sono stimati usando la verosimiglianza "ristretta", cioè basata sulla densità dei residui.
  - I parametri casuali sono stimati in modo appropriato anche in piccoli campioni

In un modello a due livelli REML and FIML portano a:

- Stime simili per σ^2
- Stime discordanti per i parametric della parte casuale se J (numero gruppi) è piccolo (in questo caso le stime FIML delle varianze sono più basse)

A meno che lo scopo principale sia la stima dei parametric casuali , FIML è da preferire perchè:

- gli stimatori FIML hanno una varianza campionaria più bassa
- il Likelihood Ratio Test può essere applicato sia per i parametri casuali sia per i fissi

Proprietà degli stimatori FIML Sotto deboli condizioni gli stimatori FIML hanno buone proprietà asintotiche

- Consistenza
- Normalità
- Efficienza

Attenzione: qui asintotico indica l'aumento del numero dei *clusters* (il solo aumento dell'ampiezza delle unità nei cluster non è suffiiciente), quindi J è la quantità fondamentale per parlare di asintotico.

2.4 Valutazione della bontà del modello

Test sui parametri Si ha che:

- Uno dei test piu utilizzati per la verica di ipotesi nei modelli di regressione multilevel e il test di Wald, in cui statistica test, Z, viene calcolata rapportando la stima puntuale del parametro di interesse all'errore standard della stima stessa.
- La distribuzione di riferimento per la statistica Z è la normale standardizzata.
- Il test di Wald si basa sull'assunto che i parametri sottoposti a verifica di ipotesi abbiano una distribuzione campionaria normale.

$$\begin{cases} H_0 : \beta_k = 0 \\ H_1 : \beta_k \neq 0 \end{cases} \quad , \qquad z = \frac{\hat{\beta}_k}{SE(\hat{\beta_k})} \sim \mathrm{N}\,(0,1)$$

Test sulla bontà d'adattamento Si ha che:

- si definisce Devianza la quantità $-2log(L)$, con L funzione di verosimiglianza.
- E' possibile confrontare statisticamente due modelli annidati
- Annidati: modello specifico può essere derivato da un modello più generale rimuovendo uno o più parametri dal modello generale utilizzando le loro devianze. I modelli con devianza inferiore presentano un miglior adattamento ai dati.
- Infatti, la differenza tra le devianze di due modelli annidati, sotto l'ipotesi nulla di equivalenza tra i due modelli, si distribuisce come un Chi-quadrato con gradi di liberta pari alla differenza nel numero dei parametri stimati dai due modelli.

2.5 Medie condizionate e marginali

Nei modelli ad effetti misti, consideriamo come esempio il seguente:

$$y_{ij} = \gamma_0 + \gamma_{10} x_{ij} + u_{1j} x_{ij} + u_{0j} + \epsilon_{ij}$$

C'è una importante distinzione tra

- la **media condizionata** della v. a. Y

$$\mathbb{E}\left[y_{ij} | x_{ij}, u_{1j}, u_{0j}\right] = \gamma_0 + \gamma_1 x_{ij} + u_{1j} x_{ij} + u_{0j}$$

- e la **media marginale**

$$\mathbb{E}\left[y_i | x_{ij}\right] = \gamma_0 + \gamma_1 x_{ij}$$

in figura 2.6 la rappresentazione grafica del valore atteso marginale e specifico per soggetti appartenenti al medesimo gruppo, considerando anche l'errore residuo.

2.6 Previsione degli effetti casuali

In molte applicazioni l'inferenza si concentra sui parametri fissi: ad esempio γ_0, β_1, in un modello ad intercetta casuale

$$y_{ij} = \gamma_0 + \beta_1 x_{ij} + u_{0j} + \epsilon_{ij}$$

Ma può essere di interesse "prevedere" i fattori casuali specifici per gruppo; si può dimostrare che lo stimatore (BLUP) per u_{0j} è :

$$\hat{u_{0j}} = \tau \left(\frac{\sum_{i=1}^{n_j} (y_{ij} - \mu_{ij})}{n_j} \right) + (1 - \tau) \cdot 0$$

Dove:

$$\mu_{ij} = \gamma_0 + \beta_1 x_{ij}$$

$$\tau = \frac{n_j \sigma_u^2}{n_j \sigma_u^2 + \sigma_\epsilon^2}$$

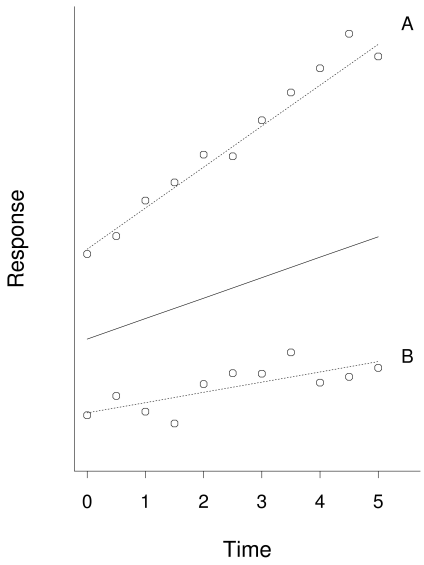


Figure 2.6: Conditional/marginal mean

τ è detto shrinkage factor.
Lo stimatore è una media ponderata del valore atteso di u_{0j} ,ovvero 0, ed il residuo medio di gruppo. Alcune note:

- Lo shrinkage factor è un numero compreso fra 0 e 1 che comprime il residuo totale stimato medio in modo differenziato a seconda della numerosità del gruppo n_j e del rapporto fra le componenti di varianza.
- Lo shrinkage sarà più forte nei gruppi poco numerosi che in quelli molto numerosi. A parità di numerosità, lo shrinkage sarà più forte quando la componente di varianza between (σ_u) è piccola rispetto a quella within.
- Lo shrinkage rende più affidabile la stima degli effetti casuali, poiché tende a riportare verso lo zero (cioè verso la media degli effetti casuali nella popolazione) la stima relativa ai gruppi poco numerosi, cioè che contengono poca informazione per la stima dell'effetto casuale.
- Occorre notare che lo shrinkage ha delle conseguenze indesiderate quando si vogliano confrontare due gruppi sulla base dei residui stimati. Può accadere che un gruppo con un elevato valore dell'effetto casuale ma di scarsa numerosità abbia lo stesso residuo stimato di un gruppo con un piccolo valore dell'effetto casuale ma di grande numerosità.

Remark 3. Alcune in ambito epidemiologico: paper

Important remark 9 (Ulteriori sviluppi di questi modelli?). si ha

- Considerare ulteriori livelli
- Inserire più variabili esplicative ad ogni livello
- Interazione tra variabili esplicative
- Considerare un modello non lineare: ad esempio logistico.

2.7 Lab

```
## ------------------------------------------------------------------
## Modelli
## ------------------------------------------------------------------

## Si consideri il punteggio conseguito nella prova di matematica come variabile di r
# Si consideri il numero di ore settimanali di studio a casa come variabile indipende
## E' possibile ipotizzare una relazione di tipo lineare tra il punteggio in
## matematica (Y) e le ore di studio settimanali (X)?

# 1) modello di regressione classico
## ----------------------------------
modlin <- lm(math ~ homework, data = db)

## Error in is.data.frame(data):  oggetto 'db' non trovato

summary(modlin)

## Error in summary(modlin):  oggetto 'modlin' non trovato

# rappresentazione grafica con punti di colore diverso in base alla scuola
palette(rainbow(10))
gg <- ggplot(db, aes(y = math, x = homework)) +
  geom_smooth(method = lm, color = "black") +
  geom_point(size = 1.5, alpha = 0.8, colour=factor(db$schnum)) +
  theme(legend.position="none")

## Error in ggplot(db, aes(y = math, x = homework)):  oggetto 'db'
non trovato

print(gg)

## Error in print(gg):  oggetto 'gg' non trovato

## Proviamo a rappresentare le regressioni separate per scuola: con xyplot
## N.B. "p"=points; "g"=grid; "r"=regression line.
regressioni <- xyplot(math ~ homework | as.factor(schnum), type = c("p","g","r"))

## Error in is.factor(x):  oggetto 'schnum' non trovato

regressioni

## Error:  oggetto 'regressioni' non trovato

# maggior parte delle scuole ha coefficiente positivo anche scuola 7 ha il
# coefficiente più basso (parte da un livello più alto).
# quattro scuole hanno coefficiente negativo

# Un primo modo per considerare la non indipendenza aggiungiamo oltre ad
# homework le dummy per la schuola

# 2) modello di regressione lineare classico con coefficienti fissi per ciascuna scuo
## ------------------------------------------------------------------
modlin2 <- lm(math ~ homework + as.factor(schnum), data = db)
```

```
## Error in is.data.frame(data):  oggetto 'db' non trovato

summary(modlin2)

## Error in summary(modlin2):  oggetto 'modlin2' non trovato

## rispetto alla precedente (3.57 era il coefficiente di regressione associato
## ad homework, mentre ora è 2.12, in termini percentuali)
## homework si è ridotto perché teniamo conto dell'effetto della scuola
## l'intercetta è la media della prima scuola in assenza di compiti

# rappresentiamolo graficamente
db$FEPredictions <- fitted(modlin2) # predizione

## Error in fitted(modlin2):  oggetto 'modlin2' non trovato

ml_est <- coef(summary(modlin2))[ , "Estimate"] # stima

## Error in summary(modlin2):  oggetto 'modlin2' non trovato

ml_se <- coef(summary(modlin2))[ , "Std. Error"] # std error coefficienti

## Error in summary(modlin2):  oggetto 'modlin2' non trovato

palette(rainbow(10))
gg <- ggplot(db, aes(y = math, x = homework)) +
  geom_line(aes(y = FEPredictions, color=as.factor(schnum))) +
  geom_abline(slope = ml_est[2], intercept = ml_est[1], size=1) +
  geom_point(size = 1.5, alpha = 0.8, colour=factor(db$schnum)) +
  theme(legend.position="none")

## Error in ggplot(db, aes(y = math, x = homework)):  oggetto 'db'
non trovato

print(gg)

## Error in print(gg):  oggetto 'gg' non trovato

# in nero la scuola di baseline, in azzurro la numero 7

## è una scelta fare in questo modo o stimare un multilevel
## es se pensiamo che le scuole siano un campione
## fino ache abbiamo pochi livelli questo è praticabilea

## -------------------------------------
## Stima dei diversi modelli multilevel
## -------------------------------------

## install.packages("lme4")
## install.packages("lattice")

library(lme4)
```

```
## Caricamento del pacchetto richiesto:  Matrix
##
## Caricamento pacchetto:  'Matrix'
## I seguenti oggetti sono mascherati da 'package:tidyr':
##
##     expand, pack, unpack
```

```
library(lattice)
```

```
## Modello 0. Modello vuoto/non condizionato
## ----------------------------------------
## (1 | schid) indica una intercetta random (1) a livello di scuola schid
## REML: restricted ML (qui settiamo a FALSE, usiamo la verosimiglianza
## completa per avere risultati uguali a quelli presentati nelle slides)
## se si cambia REML i coefficienti sono lievemente diversi
m0 <- lmer(math ~ (1 | schid), data=db, REML = FALSE)
```

```
## Error:  bad 'data':  oggetto 'db' non trovato
```

```
summary(m0)
```

```
## Error in h(simpleError(msg, call)):  errore durante la valutazione
dell'argomento 'object' nella selezione di un metodo per la funzione
'summary':  oggetto 'm0' non trovato
```

```
## la varianza residua è 72.2 mentre quella associata all'interceta casuale è
## 20
## gradi di liberta residui son 257 (260 soggetti e tre parametri stimati, le
## due varianze e l'intecetta)

## La stima dell'ICC
(ICC <- 30.54 / (30.54+72.24)) # coefficiente di correlazione intraclasse
```

```
## [1] 0.2971395
```

```
## Intercette per ogni scuola
coef(m0)
```

```
## Error in coef(m0):  oggetto 'm0' non trovato
```

```
## residuo tra intercetta generale e intercette di ogni scuola (gamma00)
(u0 <- ranef(m0, condVar = TRUE))
```

```
## Error in ranef(m0, condVar = TRUE): oggetto 'm0' non trovato
```

```
## ------------------------------------------------------------------------
## Modello 1. Modello condizionato (1 variabile esplicativa di primo livello) con
## intercetta casuale
## ------------------------------------------------------------------------
#random intercept e aggiungo la covariata homework con coefficiente fisso

m1 <- lmer(math ~ homework + (1 | schid), data = db, REML = FALSE)
```

```
## Error:  bad 'data':  oggetto 'db' non trovato
```

```r
summary(m1)
```

```
## Error in h(simpleError(msg, call)): errore durante la valutazione
dell'argomento 'object' nella selezione di un metodo per la funzione
'summary': oggetto 'm1' non trovato
```

```r
## i gradi liberta sono scesi perché abbiamo aggiunto la stima id un
## coefficiente.
## AIC e BIC, comparati con il precedente dicono che sono meglio rispetto al
## precedente (ridotti)
## se i criteri sono discorsi considerare il BIC

# coefficiente di correlazione intraclasse
(ICC = 22.50 / (22.50 + 64.26))
```

```
## [1] 0.2593361
```

```r
## coefficienti: cambia intercetta (casuali) ma homework rimane costante
coef(m1)
```

```
## Error in coef(m1): oggetto 'm1' non trovato
```

```r
## residuo tra intercetta generale e intercette di ogni gruppo
u0_1 <- ranef(m1, condVar = TRUE)
```

```
## Error in ranef(m1, condVar = TRUE): oggetto 'm1' non trovato
```

```r
u0_1
```

```
## Error: oggetto 'u0_1' non trovato
```

```r
## reppresentazione grafica residui
qqmath(ranef(m1))
```

```
## Error in ranef(m1): oggetto 'm1' non trovato
```

```r
## per ciascuna scuola plot dei residui generali
plot(m1, resid(., scaled=TRUE) ~ fitted(.) | schid,
     xlab="Fitted values", ylab= "Standardized residuals",
     abline=0, lty=3)
```

```
## Error in plot(m1, resid(., scaled = TRUE) ~ fitted(.) | schid,
xlab = "Fitted values", : oggetto 'm1' non trovato
```

```r
## normality qqplot: ok
qqmath(m1, grid=TRUE)
```

```
## Error in qqmath(m1, grid = TRUE): oggetto 'm1' non trovato
```

```r
# rappresentazione grafica del modello stimato
db$Predictions <- fitted(m1)
```

```
## Error in fitted(m1): oggetto 'm1' non trovato
```

```r
m1_est <- coef(summary(m1))[ , "Estimate"]
```

```
## Error in h(simpleError(msg, call)): errore durante la valutazione
dell'argomento 'object' nella selezione di un metodo per la funzione
'summary': oggetto 'm1' non trovato
```

```
m1_se <- coef(summary(m1))[ , "Std. Error"]
```

```
## Error in h(simpleError(msg, call)): errore durante la valutazione
dell'argomento 'object' nella selezione di un metodo per la funzione
'summary': oggetto 'm1' non trovato
```

```
palette(rainbow(10))
gg <- ggplot(db, aes(y = math, x = homework)) +
  geom_line(aes(y = Predictions, color=as.factor(schnum))) +
  geom_abline(slope = m1_est[2], intercept = m1_est[1], size=1) +
  geom_point(size = 1.5, alpha = 0.8, colour=factor(db$schnum)) +
  theme(legend.position="none")
```

```
## Error in ggplot(db, aes(y = math, x = homework)): oggetto 'db'
non trovato
```

```
print(gg)
```

```
## Error in h(simpleError(msg, call)): errore durante la valutazione
dell'argomento 'x' nella selezione di un metodo per la funzione 'print':
oggetto 'gg' non trovato
```

```
## Modello 2. Modello condizionato a due livelli (1 variabile esplicativa di primo li
## random slope and random intercept
## ------------------------------------------------------------------
## aggiungiamo componente casuale per homework. di default non c'è necessità di
## aggiungere anche (1 + homework | schid), ce la mette lui

m2 <- lmer(math ~ homework + (homework | schid), data=db,REML = FALSE)
```

```
## Error: bad 'data': oggetto 'db' non trovato
```

```
summary(m2)
```

```
## Error in h(simpleError(msg, call)): errore durante la valutazione
dell'argomento 'object' nella selezione di un metodo per la funzione
'summary': oggetto 'm2' non trovato
```

```
## AIC e BIC scendono.
## i residui hanno un range ridotto
## calano ancora i gradi di libertà dei residui

## i coefficienti sono gamma_00 e gamma_10 medi CHECK
## R non da la covarianza ma la correlazione dei fixed effects


# diverse sia intercette che coeff di regressione (per alcuni negativi come
# visto nelle stime divise)
coef(m2)
```

```
## Error in coef(m2):  oggetto 'm2' non trovato

# qui facciamo un LRT per due modelli che differiscono nell'effetto casuale di
# homework
# i due modelli differiscono per questo e sono annidati quindi possiamo effettuarlo
anova(m1, m2)

## Error in anova(m1, m2):  oggetto 'm1' non trovato

## -2 log (verosim m1/ verosim m2)
## ampiamente significativo: sia con AIC/BIC che con LRT (appropriato in questo
## caso) otteniamo che c'è bisogno di intercetta casuale

## i gradi di liberta sono diminuiti di 2. nel modello piu cazzuto abbiamo
## dovuto anche stimare
## - la varianza associata al coefficiente di regressione di homework
## - la covarianza tra le due componenti casuali


## plot dei residui sia per intercetta che coefficiente di regressione
qqmath(ranef(m2))

## Error in ranef(m2):  oggetto 'm2' non trovato

#analogo
# dotplot(ranef(model4, condVar=TRUE))

#residual plot sulle singole scuole
plot(m2, resid(., scaled=TRUE) ~ fitted(.) | schid,
     xlab="Fitted values", ylab= "Standardized residuals",
     abline=0, lty=3)

## Error in plot(m2, resid(., scaled = TRUE) ~ fitted(.)  | schid,
xlab = "Fitted values", :  oggetto 'm2' non trovato

#normality qqplot; va unpo peggio ma tutto sommato ok
qqmath(m2, grid=TRUE)

## Error in qqmath(m2, grid = TRUE): oggetto 'm2' non trovato

## rappresentazione grafica delle diverse (quella nera di riferimento)
db$Predictions <- fitted(m2)

## Error in fitted(m2):  oggetto 'm2' non trovato

m2_est <- coef(summary(m2))[ , "Estimate"]

## Error in h(simpleError(msg, call)):  errore durante la valutazione
dell'argomento 'object' nella selezione di un metodo per la funzione
'summary':  oggetto 'm2' non trovato

m2_se <- coef(summary(m2))[ , "Std. Error"]

## Error in h(simpleError(msg, call)):  errore durante la valutazione
dell'argomento 'object' nella selezione di un metodo per la funzione
'summary':  oggetto 'm2' non trovato
```

```
palette(rainbow(10))
gg <- ggplot(db, aes(y = math, x = homework)) +
  geom_line(aes(y = Predictions, color=as.factor(schnum))) +
  geom_abline(slope = m2_est[2], intercept = m2_est[1], size=1) +
  geom_point(size = 1.5, alpha = 0.8, colour=factor(db$schnum)) +
  theme(legend.position="none")
```

```
## Error in ggplot(db, aes(y = math, x = homework)): oggetto 'db'
non trovato
```

```
print(gg)
```

```
## Error in h(simpleError(msg, call)): errore durante la valutazione
dell'argomento 'x' nella selezione di un metodo per la funzione 'print':
oggetto 'gg' non trovato
```

Chapter 3

Modelli per dati longitudinali

Remark 4. Libri di testo di riferimento per questa parte sono:

- J.D. Singer, J.B. Willet Applied Longitudinal Data Analysis, Oxford University Press, 2003
- esempi tratti da G.M Fitzmaurice, N. M. Laird, J. H. Ware, Applied Longitudinal Analysis, Wiley, 2004.

3.1 Dati longitudinali

La caratteristica distintiva degli studi longitudinali è che le misurazioni sugli stessi individui vengono effettuate ripetutamente nel tempo. **Obiettivo**: descrivere il cambiamento della risposta nel tempo e i fattori che influenzano tale cambiamento.
Confrontando le risposte di ciascun individuo in due o più occasioni, un'analisi longitudinale può rimuovere fonti di variabilità estranee, ma inevitabili, tra gli individui. Ciò elimina le principali fonti di variabilità o "rumore" dalla stima del cambiamento nello stesso individuo. **Complicazioni**:

- le misurazioni ripetute sugli individui sono correlate,
- la variabilità è spesso eterogenea tra le diverse occasioni di misurazione.

Requisiti Fondamentali:

- La variabile (dipendente) rilevata deve evolvere sistematicamente nel tempo.
- I dati sono stati rilevati in più *occasioni temporali*.
  Il *numero dei tempi* di rilevazione è variabile, dipende dal fenomeno indagato ed influisce sulla possibilità di utilizzare modelli più o meno elaborati. Se disponiamo di soli 3 punti nel tempo per ciascun individuo, l'evoluzione temporale del fenomeno per ciascun individuo può essere descritta con una traiettoria di tipo lineare.
- Selezione di una opportuna *metrica del tempo*. La scelta della metrica influisce sul numero dei punti nel tempo osservati e sulla loro distanza (*equispaziati o no*). La scelta è chiaramente legata al fenomeno che stiamo

	asdasdasd				
Individual	1	2	3	...	T
1	y_{11}	y_{12}	y_{13}	$\cdots$	y_{1T}
2	y_{21}	y_{22}	y_{23}	$\cdots$	y_{2T}
... n	y_{n1}	y_{n2}	y_{n3}	$\cdots$	y_{nT}

Table 3.1: Disegno bilanciato

analizzando [1]
In relazione alla metrica possiamo distinguere:

- *occasioni di rilevazione*: equispaziate o non
- *dataset*: strutturati nel tempo (individui rilevati nelle stesse occasioni) o no (occasioni di rilevazione sono diverse)
- *disegno*: bilanciato (individui rilevati in un numero uguale di occasioni) o non (numero diverso di rilevazioni per individuo)

Considerando un disegno bilanciato con T tempi/osservazioni ripetute in n individui la seguente tabella si genera tabella 3.1. A livello di notazione usiamo y_{it}:

- Y la variabile di risposta
- $i = 1, \ldots, n$ individui
- $t = 1, \ldots, T_i$, dove se $T_i = T, \forall i$ il disegno è bilanciato
- $\mathbf{y}_i = [y_{i1}, \ldots, y_{iT}]$ è una serie storica e realizzazione finita di un processo stocastico, una successione di variabili aleatorie dipendenti e consideriamo le osservazioni relative ad individui diversi sono indipendenti

$$\mathbf{y}_i \perp\!\!\!\perp \mathbf{y}_j, \forall i \neq j$$

La struttura dati è ancora gerarchica:

- 1 livello: *osservazioni temporali* rilevate per ogni individuo unità di primo livello : tempo t, $t = 1, \ldots, T$
- 2 livello: *individui*, $i = 1, \ldots, n$

Nei modelli multilivello di crescita, al

- 1 livello: specifico la traiettoria individuale del fenomeno
- 2 livello: quali sono le principali differenze tra gli individui

Example 3.1.1 (Esempio tolerance)**.** abbiamo:

- 5 occasioni di rilevazione da 11 a 15 anni di età
- Obiettivo: analisi della tolleranza verso comportamenti devianti/limite

[1] Negli studi psicologici potrebbe essere registrato in termini di settimane o numero di sedute, in studi scolastici età o livello di istruzione, studi sulla genitorialità età dei genitori o del bambino.

- Sono state rilevate 9 variabili rilevate su una scala da 1 a 4 (1 comportamento ritenuto sbagliato – 4 comportamento ritenuto non sbagliato).
  Le variabili rilevate: Copiare ad un esame, Distruggere volontariamente la proprietà altrui, Fare uso di Marjuana, Rubare qualcosa che vale meno di 5 dollari, Picchiare o minacciare qualcuno senza motivo, Fare uso di Alchool, Intrufolarsi in un palazzo o auto per rubare, Spacciare droga pesante, Rubare qualcosa del valore superior a 5 dollari.

- Variabile risposta: "Tolerance" ottenuta come media dei punteggi delle 9 variabili: $y_i t$ con $i = 1, \ldots, 590$ individui e $t = 11, \ldots, 15$ (il disegno è bilanciato)

- Sono state rilevate due variabili a livello individuale e invarianti nel tempo:

  1. `sesso`: 1 = male 0= female
  2. `exposure` : un indice di autovalutazione circa l'esposizione a comportamenti limite alla età di 11 anni sempre su scala 0-4

Una prima analisi fattibile è di descrivere i cambiamenti individuali nel tempo: riportare graficamente (scatterplot) la relazione tra variabile risposta e tempo. Alcune indicazioni:

- è bene riportare le traiettorie per diversi individui in modo da analizzare la presenza di eventuali "pattern"

- usare la stessa scala tra i diversi individui

- se il dataset è molto grande, estrarre un campione di individui per effettuare questa analisi

- Successivamente la relazione tra variabile risposta e tempo in ciascun paziente deve essere analizzata osservando le traiettorie individuali mediante modelli non parametrici e parametrici.

- Cercheremo infine quindi di fare una prima valutazione della influenza che possono avere le covariate osservare sulla variabilità individuale nelle traiettorie: questo si farà ponendo sullo stesso grafico tutti i pazienti caratterizzati da un medesimo livello di covariata

3.2 Modelli vari

Considerando y_{it} con Y variabile casuale risposta dipendente dai pedici $i = 1, \ldots n$ individui $t = 1, \ldots T_i$ (se $T_i = T \forall i$ il disegno è bilanciato) si ha che $\mathbf{y}_i = [y_{i1}, \ldots, y_{iT}]$ è una serie storica e realizzazione finita di un processo stocastico, una successione di variabili aleatorie dipendenti.
Si ipotizza che le osservazioni relative ad individui diversi sono indipendenti $\mathbf{y}_i \perp\!\!\!\perp \mathbf{y}_j, \forall i \neq j$

3.2.1 Modello ad intercetta casuale – non condizionato (modello nullo)

Definito dall'equazione

$$y_{it} = \pi_{0i} + \epsilon_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T_i$$

dove

1. nel primo livello per ciascun soggetto il modello specifica la dipendenza tra le osservazioni
2. L'obiettivo è fare inferenza sul valore atteso del processo $\mathbb{E}\left[(\right] y_i t)$

$$\pi_{0i} = \gamma_{00} + u_{0i} y_{it} = \gamma_{00} + u_{0i} + \epsilon_{it}$$

quindi complessivamente

$$y_{it} = \underbrace{\gamma_{00}}_{\text{parte fissa}} + \underbrace{u_{0i} + \epsilon_{it}}_{\text{parte casuale}}$$

dove assumiamo che ϵ_{it} e u_{0i} siano iid e inoltre

$$\epsilon_{it} \sim \mathrm{N}\left(0, \sigma_\epsilon^2\right), \quad u_{0i} \sim \mathrm{N}\left(0, \sigma_u^2\right), \quad \epsilon_{it} \perp\!\!\!\perp u_{0i} \forall i, j$$

```
Modello ad intercetta casuale - non condizionato
(modello nullo)
AIC BIC logLik deviance df.resid
109.0 116.2 -51.5 103.0
77
Random effects:
Groups Name Variance Std.Dev.
id
(Intercept) 0.07465 0.2732
Residual
0.16794 0.4098
Number of obs: 80, groups: id, 16
Fixed effects:
Estimate Std. Error t value
(Intercept) 1.61937 0.08225 19.69 la media di tolerance è significativamente diversa
da 0
ICC=0.07465/(0.07465+0.16794) # coefficiente di correlazione intraclasse
-> 0.3077208
```

3.2.2 Modello ad intercetta casuale – non condizionato (effetto del tempo fisso)

Espresso da

$$y_{it} = \pi_{oi} + \pi_1 t + \epsilon_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T_i$$

dove

1. Nel 1° livello per ciascun soggetto il modello specifica la dipendenza tra le osservazioni
2. L'obiettivo è fare inferenza sul valore atteso del processo E(yit)
3. L'ipotesi iniziale è che la traiettoria del processo sia lineare nel tempo

$$\pi_{0i} = \gamma_{00} + u_{0i}$$
$$y_{it} = \gamma_{00} + u_{0i} + \pi_1 t + \epsilon_{it}$$

and overall

$$y_{it} = \underbrace{\gamma_{00} + \pi_1 t}_{\text{parte fissa}} + \underbrace{u_{0i} + \epsilon_{it}}_{\text{parte casuale}}$$

La stima porta a

```
AIC BIC logLik deviance df.resid
92.2 101.7 -42.1 84.2
76
Random effects:
Groups Name
Variance Std.Dev.
id
(Intercept) 0.0832 0.2885
Residual
0.1252 0.3538
Number of obs: 80, groups: id, 16
Fixed effects:
Estimate Std. Error t value
(Intercept) 1.35775 0.09947 13.650
time
0.13081 0.02797 4.677
```

La varianza residua (livello 1) sintetizza la dispersione dei valori individuali rispetto alla propria traiettoria. Se il modello scelto è adeguato, c'è un effetto tempo allora la varianza residua di questo modello deve essere minore di quella del precedente.

```
ICC=0.0832/(0.0832+0.1252) # coefficiente di correlazione intraclasse
-> 0.3992322
```

3.2.3 Modello ad intercetta casuale – non condizionato intercetta e slope casuali

Definito da

$$y_{it} = \pi_{0i} + \pi_{1i} t + \epsilon_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots T_i$$

where

$$\pi_{0i} = \gamma_{00} + u_{0i} \pi_{1i} = \gamma_{10} + u_{1i}$$

and overall

$$y_{it} = \gamma_{00} + u_{0i} + (\gamma_{10} + u_{1i})t + \epsilon_{it}$$
$$y_{it} = \underbrace{\gamma_{00} + \gamma_{10}t}_{\text{parte fissa}} + \underbrace{u_{1i}t + u_{0i} + \epsilon_{it}}_{\text{parte casuale}}$$

and assumptions

$$\epsilon_{ij} \sim \mathrm{N}\left(0, \sigma^2_\epsilon\right)$$
$$\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u_0} & \sigma_{u_{01}} \\ \sigma_{u_{01}} & \sigma^2_{u_1} \end{bmatrix}\right)$$
$$\epsilon_{ij} \perp\!\!\!\perp \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix}$$

e cosi poi de botto si ha che

$$\begin{bmatrix} \pi_{0i} & \pi_{1i} \end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix} \gamma_{00} \\ \gamma_{10} \end{bmatrix}, \begin{bmatrix} \sigma^2_{u_0} & \sigma_{u_{01}} \\ \sigma_{u_{01}} & \sigma^2_{u_1} \end{bmatrix}\right)$$

La stima porta a

```
AIC
76.0

BIC logLik deviance df.resid
90.3 -32.0 64.0
74

Random effects:
Random effects:
Groups Name
Variance Std.Dev. Corr
id
(Intercept) 0.03866 0.1966
time
0.02042 0.1429 -0.24
Residual
0.07412 0.2722
Number of obs: 80, groups: id, 16
Fixed effects:
Estimate Std. Error t value
(Intercept) 1.35775 0.07208 18.836
time
0.13081 0.04171 3.137
```

Remark 5. La correlazione di popolazione tra lo stato inziale e il tasso di crescita è -0.24. Ci dice che gli adolescenti che avevano una tolerance maggiore a 11 anni aumentano il valore di questa variabile meno rapidamente nel tempo.

3.2.4 Modello ad intercetta casuale – condizionato intercetta e slope casuali con covariate a livello di individuo

Con il modello

$$y_{it} = \pi_{0i} + \pi_{1i}t + \epsilon_{it} \quad i = 1, \ldots, n \quad t = 1, \ldots, T_i$$

con

$$\begin{aligned} \pi_{0i} &= \gamma_{00} + \gamma_{01}W_i + u_{0i} \\ \pi_{1i} &= \gamma_{10} + \gamma_{11}W_i + u_{1i} \end{aligned}$$

and putting all togheter

$$\begin{aligned} y_{it} &= \gamma_{00} + u_{0i} + (\gamma_{10} + \gamma_{11}W_i + u_{1i})t + \varepsilon_{it} \\ &= \underbrace{\gamma_{00} + \gamma_{01}W_i + \gamma_{10}t + \gamma_{11}W_i t}_{\text{Parte fissa}} + \underbrace{u_{1it} + u_{0i} + \varepsilon_{it}}_{Partecasuale} \end{aligned}$$

E le seguenti assunzioni

$$\begin{aligned} &\varepsilon_{ij} \sim \mathrm{N}\left(0, \sigma_\varepsilon^2\right) \\ &\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_{u_0}^2 & \sigma_{u_{01}} \\ \sigma_{u_{01}} & \sigma_{u_1}^2 \end{bmatrix}\right) \\ &\varepsilon_{ij} \perp\!\!\!\perp \begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix} \\ &\begin{bmatrix} \pi_{0i} \\ \pi_{1i} \end{bmatrix} \sim \mathrm{MVN}\left(\begin{bmatrix} \gamma_{00} \\ \gamma_{10} \end{bmatrix}, \begin{bmatrix} \sigma_{u_0}^2 & \sigma_{u_{01}} \\ \sigma_{u_{01}} & \sigma_{u_1}^2 \end{bmatrix}\right) \end{aligned}$$

la stima con la dummy male porta ai segguenti risultati

```
AIC
79.2

BIC logLik deviance df.resid
98.3 -31.6 63.2
72


Random effects:
Random effects:
Groups Name
Variance Std.Dev. Corr
id
(Intercept) 0.03864 0.1966
time
0.01938 0.1392 -0.25
Residual
0.07411 0.2722
```

```
Number of obs: 80, groups: id, 16
Fixed effects:
Estimate Std. Error t value
(Intercept) 1.355556 0.096096 14.106

18

male

0.005016 0.145284 0.035

time

0.102333 0.054556 1.876

time:male 0.065095 0.082481 0.789
```

Dove

- i valori di AIC e BIC sono aumentati
- 14.106 è il valore di intercetta per le donne
- 0.035 è il differenziale nel valore intercetta tra ragazzi e ragazze
- 1.876 è la slope (tasso di crescita per ragazze)
- 0.789 è il differenziale nel tasso di crescita tra ragazzi e ragazze

La stima con exposure (dicotomizzata) e interazione porta a

```
AIC
74.3

BIC logLik deviance df.resid
93.4 -29.2 58.3
72

Random effects:
Random effects:
Groups Name
Variance Std.Dev. Corr
id
(Intercept) 0.03738 0.1934
time
0.01255 0.1120 -0.16
Residual
0.07412 0.2722
Number of obs: 80, groups: id, 16
Fixed effects:
Estimate Std. Error t value
(Intercept)
1.39350 0.10115 13.776
```

```
time
0.04213 0.04996 0.843
expdic
-0.07150 0.14305 -0.500
time:expdic 0.17737 0.07065 2.511
```

Dove notare BIC e l'interazione tra tempo ed esposizione
Infine per exposure in continuo abbiamo

```
Modello ad intercetta casuale - condizionato
intercetta e slope casuali con covariata - exposure(continua)
AIC
71.1

BIC logLik deviance df.resid
90.2 -27.6 55.1
72

Random effects:
Random effects:
Groups Name
Variance Std.Dev. Corr
id
(Intercept) 0.03417 0.1848
time
0.01498 0.1224 -0.52
Residual
0.07411 0.2722
Number of obs: 80, groups: id, 16
Fixed effects:
Estimate Std. Error t value
(Intercept)
1.1069 0.2716 4.076
time
-0.1452 0.1449 -1.002
exposure
0.2106 0.2203 0.956
time:exposure 0.2317 0.1175 1.971

20
```