Analisi dati longitudinali

June 4, 2025

Contents

Chapter 1

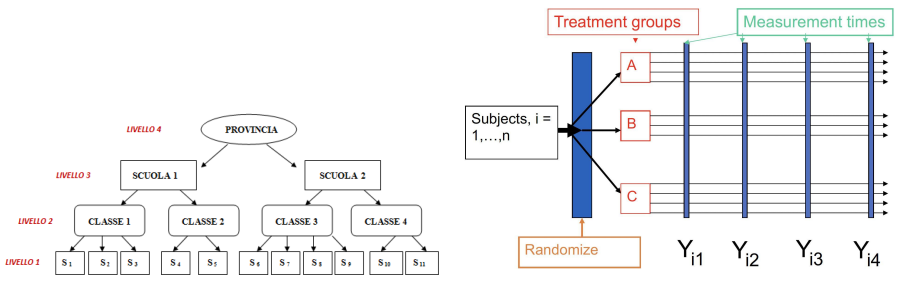
Introduzione

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

- **dato gerarchico**: in scienza dell'educazione lo studio dell'efficacia dei sistemi educativi scolastici analizza la struttura formata da alunni, classi, scuole.
  Le caratteristiche dei dati appartengono a diversi **livelli**, dove il livello 1 è quello più *disaggregato*
- **dato longitudinale**: in studi Epidemiologici un gruppo di soggetti sono seguiti nel tempo e rilevate periodicamente informazioni su fattori di rischio/protettivi ed esiti di salute.
  Un singolo soggetto i ha misurazioni di una variable a più istanti nel tempo, $Y_{i1}, \ldots, Y_{i4}$.



(a) (b)

Figure 1.1: Esempi dati gerarchici

Remark 1 (Ipotesi del modello di regressione lineare classico). Ricordando un generico modello lineare:

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

con $\beta_0, \beta_k,\ x_{1i}, \ldots x_{ki}$ fissi a livello di popolazione e ε_i variabile casuale della parte rimanente utile a spiegare y_i. Il modello si basa sulle seguenti ipotesi:

- Errori con valore atteso pari a 0

$$E\varepsilon_i | x_{1i}, x_{2i}, \ldots, x_{ki} = 0$$

- osservazioni IID ed errori **incorrelati**

$$\text{Cov}\left(\varepsilon_i, \varepsilon_j\right) = 0, \quad \forall i \neq j$$

- Errori omoschedastici (varianza del termine di errore indipendente dal livello delle covariate):

$$\text{Var}\left[\varepsilon_i\right] = \text{Var}\left[\varepsilon_i | x_{1i}, x_{2i}, \ldots, x_{ki}\right] = \sigma^2$$

- Normalità errore: $\varepsilon_i \sim \text{N}\left(0, \sigma^2\right)$

Important remark 1. Nel caso di dati gerarchici o longitudinali i dati non sono più indipendenti. Né nei modelli gerarchici, né nelle misure ripetutte le osservazioni sono in qualche modo correlate fra loro: cade dunque una delle ipotesi fondamentali del modello di regressione classico.

Important remark 2. Come trattare questi dati? Distinguiamo tra disegni:

- *cross-sectional* (raccolta dati in un punto e causalità più debole): nei dati cross-section ci può essere correlazione all'interno dei gruppi: la struttura di errore è complessa e rispecchia la gerarchia dei dati. Obiettivo è esplorare la variabilità tra gruppi: si utilizzano modelli multilevel (o ad effetti misti, a coefficienti casuali, ad effetti casuali, a componenti di varianza, tutti sinonimi).
  Occorre capire se la struttura dei dati permette o esige un'applicazione multilevel o si può anche fare con un modello più semplice

- *longitudinali* che portano a dati panel (misure ripetute): obiettivi qui sono misurare il cambiamento nel tempo dell'outcome intra-individuale (aumenta/diminuisce per un dato individuo) e tra individui (il cambiamento nel tempo varia a seconda dell'individuo o è omogeneo)

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:

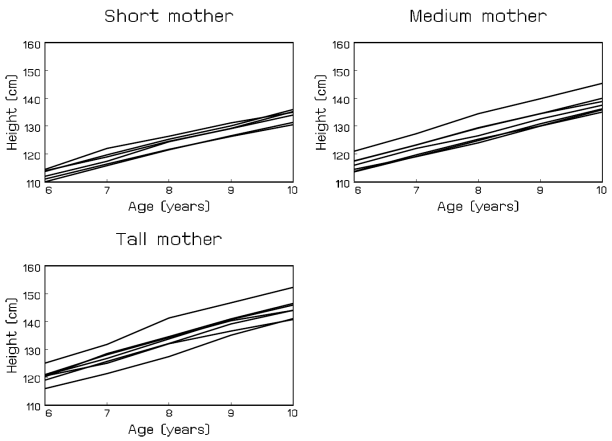


Figure 1.2: Esempio bambine

- 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; qui è semplice vedere perché le bambine sono poche
- prevedere i cambiamenti
- valutare l'effetto di covariate (qui solo 1)

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.
La visualizzazione in questo caso è resa semplice perché le bambine sono poche; se così non è, per le prime esplorazioni grafiche pre-modello, si consiglia di selezionare un subset casualmente.

Chapter 2

Multilevel models

2.1 Esempio scuole

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

- di **primo livello**: sono gli studenti
- di **secondo livello**: la scuola di appartenenza

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

```
library(tidyverse)
library(haven)
library(lattice)
library(lme4)

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

}
```

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

# Alcuni casi:
head(db)

##   schid stuid   ses    meanses homework white parented public ratio percmin
## 1     1     3 -0.13 -0.4826087        1     1        2      1    19       0
## 2     1     8 -0.39 -0.4826087        0     1        2      1    19       0
## 3     1    13 -0.80 -0.4826087        0     1        2      1    19       0
## 4     1    17 -0.72 -0.4826087        1     1        2      1    19       0
## 5     1    27 -0.74 -0.4826087        2     1        2      1    19       0
## 6     1    28 -0.58 -0.4826087        1     1        2      1    19       0
##   math sex race sctype cstr scsize urban region schnum
## 1   48   2    4      1    2      3     2      2      1
## 2   48   1    4      1    2      3     2      2      1
## 3   53   1    4      1    2      3     2      2      1
## 4   42   1    4      1    2      3     2      2      1
## 5   43   2    4      1    2      3     2      2      1
## 6   57   2    4      1    2      3     2      2      1
```

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)

##
##  1  2  3  4  5  6  7  8  9 10
## 23 20 24 22 22 20 67 21 21 20

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

## [1] 26

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

##
##   0   1
##  67 193

## 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
aggregate(db[c("math", "homework", "public")], by=list(db$schnum),
          mean,
          na.rm=TRUE)
```

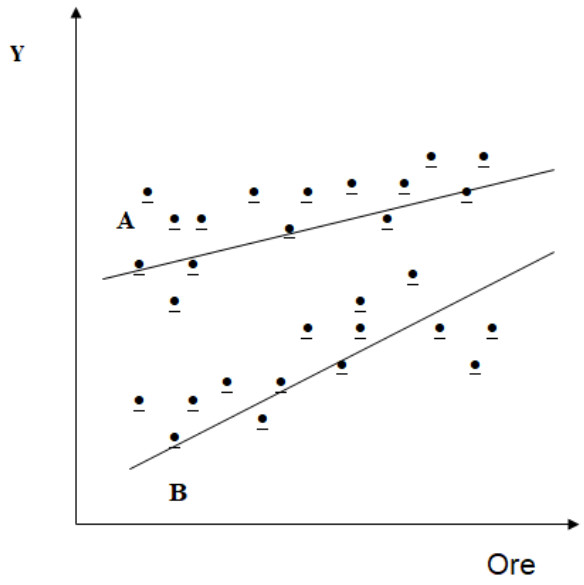


Figure 2.1: Relazione ore math ipotetica tra scuole

```
##    Group.1     math  homework public
## 1        1 45.73913 1.3913043      1
## 2        2 42.15000 2.3500000      1
## 3        3 53.25000 1.8333333      1
## 4        4 43.54545 1.6363636      1
## 5        5 49.86364 0.8636364      1
## 6        6 46.40000 1.1500000      1
## 7        7 62.82090 3.2985075      0
## 8        8 49.66667 2.0952381      1
## 9        9 46.33333 1.3333333      1
## 10      10 47.85000 1.6000000      1
```

Remark 2 (Classificazione variabili di livello > 1). 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))

Remark 3 (Obiettivo dello studio). Confronto tra le diverse scuole e nello specifico studiare l'effetto delle ore di studio (`homework`) sul punteggio `math`. Ad esempio se la situazione fosse quella mostrata in figura 2.1 abbiamo che

- variabilità tra scuole: vi è variabilità nel senso che scuola A in media più preparati in termini di performance su math

- correlazione tra fenomeni: in A la variabile ore di studio è meno predittiva dell'outcome

La prima analisi da condurre è la stima separata per scuola, e vedere la pendenze (ns)interesse la relazione tra ore di studio e performance di matematica

2.2 Problemi analisi classiche

Important remark 3. Si potrebbe di analizzare i dati con metodi "classici" ma questi danno luogo a problemi. Le alternative sarebbero

- **analisi disaggregata**: analisi a livello individuale (es. studente) ignorando l'appartenenza al gruppo. I problemi hanno a che fare con
  - **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, trattandoli come un campione casuale da una popolazione di gruppi, per capire ad esempio se vi è variabilità o si comportano diversamente.
    Si potrebbe al limite includere dummy per individuare i gruppi; il problema è che:
    - se i gruppi sono tanti si ha un numero elevato di parametri da stimare
    - non si dispone di una stima media/marginale, es si hanno tante intercette tante quanti i gruppi che però non sono troppo di interesse
  - vi è una errata dimensione campionaria delle variabili di livello 2, che hanno dimensione inferiore ad n (es pubblico/privato fa riferimento alla scuola, non all'individuo)
- **analisi aggregata** (pooled): si esegue una analisi su dati/medie a livello di gruppo (es. scuola). I problemi:
  - **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. Questo può portare ad **Ecological fallacy** (distorsione da aggregazione): le relazioni a livello di gruppo (cioè tra le medie di gruppo) possono essere diverse dalle corrispondenti relazioni a livello individuale;
  - **Interazione tra livelli**: l'analisi aggregata non consente di studiare le relazioni tra livelli gerarchici

Poi in generale appiattire e fare inferenza mediante modello lineare ad un solo livello (1 o 2) fa si che:

- si **ignori l'annidamento gerarchico**: quando gli individui sono annidati all'interno di gruppi, si hanno due sorgenti di variabilità che possono esser scomposte, all'interno dei gruppi e tra i gruppi;
- non si riesce a capire **l'interazione tra livelli**: si vogliono considerare interazioni tra variabili esplicative definite a differenti livelli di struttura gerarchica. Ad esempio: 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 riposta? 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?
- si possano commettere errori dovuti al non considerare il livello corretto:
  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): nell'esempio considerando K gruppi $Z = 1, \ldots, K$, la relazione entro gruppo tra una variabile e l'altra è negativa, ma se andiamo ad aggregare a livello di gruppo la relazione tra le due variabili diviene positiva.
     Questa fallacy 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

Important remark 4. Per queste motivazioni ricorriamo ad un approccio **multilevel** che permette di avere una stima corretta dell'errore standard e permettono di valutare la varianza tra gruppi.

2.2.1 Alcune stime classiche

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

##
## Call:
## lm(formula = math ~ homework, data = db)
##
## Residuals:
```

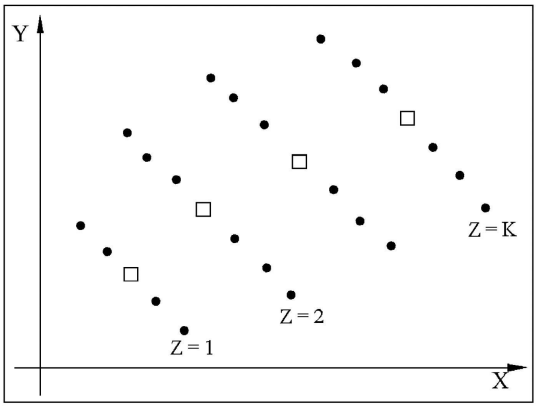
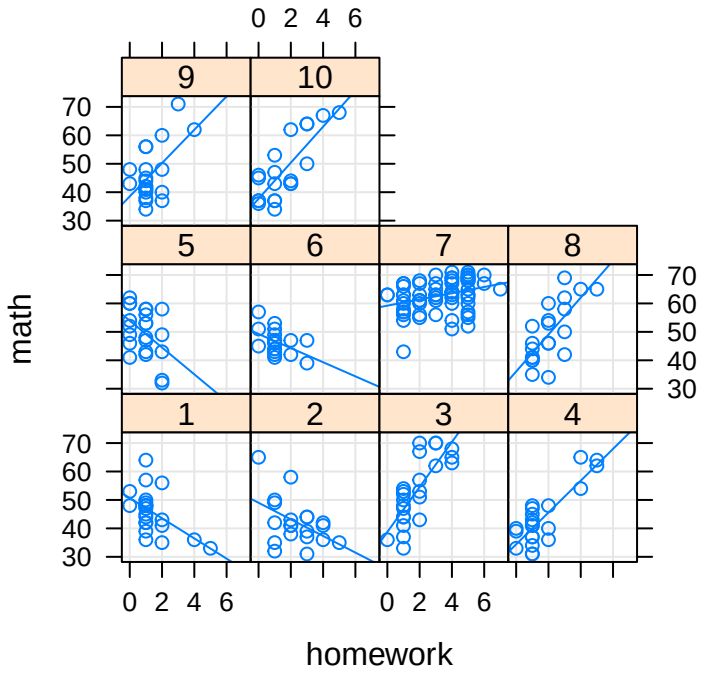


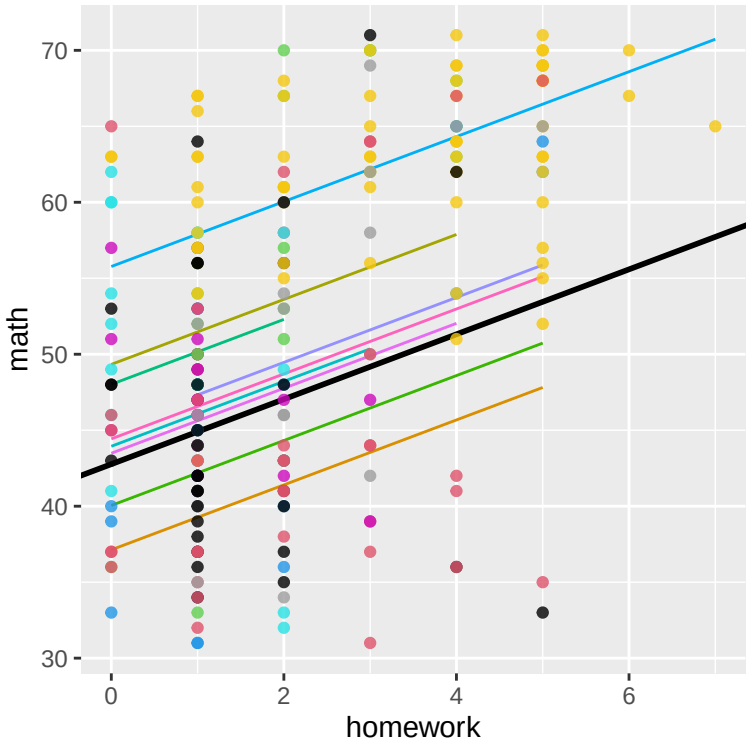
Figure 2.2: Ecological fallacy

```
##      Min       1Q   Median       3Q      Max
## -28.9331  -6.6457   0.3543   7.0669  20.9261
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  44.0739     0.9886   44.58   <2e-16 ***
## homework      3.5719     0.3882    9.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.682 on 258 degrees of freedom
## Multiple R-squared:  0.247,Adjusted R-squared:  0.2441
## F-statistic: 84.64 on 1 and 258 DF,  p-value: < 2.2e-16

## 1b) modello di regression classsico splittato per scuola
## ------------------------------------

## 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),
                       data = db,
                       type = c("p","g","r")))
```




```
# 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 scuola
## ------------------------------------------------------------------
modlin2 <- lm(math ~ homework + as.factor(schnum), data = db)
summary(modlin2)

##
## Call:
## lm(formula = math ~ homework + as.factor(schnum), data = db)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.4496  -5.0547  -0.1313   4.7138  27.8711
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          42.7664     1.7587  24.317  < 2e-16 ***
## homework              2.1366     0.3836   5.570 6.60e-08 ***
## as.factor(schnum)2   -5.6375     2.4845  -2.269  0.02412 *
```

```
## as.factor(schnum)3    6.5664     2.3512   2.793  0.00563 **
## as.factor(schnum)4   -2.7173     2.3985  -1.133  0.25834
## as.factor(schnum)5    5.2519     2.4052   2.184  0.02993 *
## as.factor(schnum)6    1.1764     2.4589   0.478  0.63275
## as.factor(schnum)7   13.0068     2.0754   6.267 1.61e-09 ***
## as.factor(schnum)8    2.4235     2.4406   0.993  0.32169
## as.factor(schnum)9    0.7181     2.4258   0.296  0.76746
## as.factor(schnum)10   1.6650     2.4585   0.677  0.49888
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.037 on 249 degrees of freedom
## Multiple R-squared:  0.4992,Adjusted R-squared:  0.4791
## F-statistic: 24.83 on 10 and 249 DF,  p-value: < 2.2e-16

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

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")

## Warning:  Using ‘size‘ aesthetic for lines was deprecated in ggplot2
3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning
was generated.

print(gg)
```




```
# 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
```

2.3 Modelli multilevel

Important remark 5. Un modello multilevel si ha quando vi è:

- struttura gerarchica nelle equazioni di regressione;
- una sola variable dipendente misurata al livello più *basso*;
- variabili esplicative a tutti i livelli.

Remark 4. Nel prosieguo ne vediamo alcuni di complessità crescente e alla base dei quali vi sono assunti diversi sulla relazione tra variabili.

2.3.1 Modello di regressione singolo per la media

Per motivi didattici partiamo da un modello NON gerarchico; un modello standard definito dalla sola intercetta, senza regressori, anche detto modello non condizionato o nullo. Quello che viene stimato è la media della variabile di

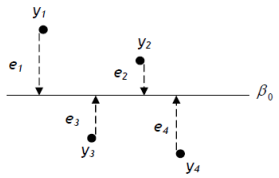


Figure 2.3: Modello di regressione singolo per la media

risposta.
Viene definito dall'equazione:

$$y_i = \beta_0 + \varepsilon_i$$

e rappresentato dal grafico 2.3 dove:

- 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

Questo è un modello ad effetti fissi: non considera l'appartenenza degli studenti a scuole diverse.

Example 2.3.1. `summary(lm_null <- lm(math ~ 1, data=db))`

```
##
## Call:
## lm(formula = math ~ 1, data = db)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -20.3   -9.3   -1.8   10.7   19.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  51.3000     0.6906   74.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.14 on 259 degrees of freedom
```

Notiamo che

- la stima dell'intercetta β_0 (media complessiva del punteggio di matematica) è 51.3
- l'errore standard conduce ad un t-value molto alto di 74.3
- la stima di $\sqrt{\sigma^2_\varepsilon} = 11.1$

2.3.2 Modello a effetti casuali non condizionato

Il primo modello gerarchico che vediamo è definito dall'equazione

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

dove

- Y: variabile risposta riferita unità di primo livello
- $j = 1, \ldots, J$ sono gli id delle unità di secondo livello (es scuole)
- $i = 1, \ldots n_j$ id delle unità di primo livello entro l'unità di secondo livello j (complessivamente le unità di primo livello sono $n_1 + \ldots + n_J = n$)
- β_{0j} è la media di Y nel gruppo j
- ε_{ij} è l' errore (legato alla variabilità individuale) residuo dell'unità i entro il gruppo j rispetto alla media di quest'ultimo

Le assunzioni di questo modello:

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

  $$\beta_{0j} = \gamma_0 + u_{0j}$$

  dove

  - γ_0 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

  La media di gruppo β_j ha una componente fissa γ_0 comune ed una casuale u_{0j} (una componente stocastica legata al secondo livello);

- alla luce di ciò, complessivamente il modello si può riscrivere come

  $$y_{ij} = \gamma_0 + u_{0j} + \varepsilon_{ij}$$

  con γ_0 componente fissa e $u_{0j} + \varepsilon_{ij}$ componenti casuali; il modello si può rappresentare come in figura 2.4

- le **assunzioni sulle componenti** sono

  $$\begin{aligned} &\varepsilon_{ij} \sim \mathrm{N}\left(0, \sigma^2_{\varepsilon}\right) \\ &u_{0j} \sim \mathrm{N}\left(0, \sigma^2_{u_0}\right) \\ &\varepsilon_{ij} \perp\!\!\!\perp u_{0j}, \quad \forall i, j \end{aligned}$$

  Si noti in particolare che per il termine ε_{ij}, da una variabilità entro le classi σ^2_{ε} costante per tutti/ indipendente dal gruppo di appartenenza (non dipendente da j).

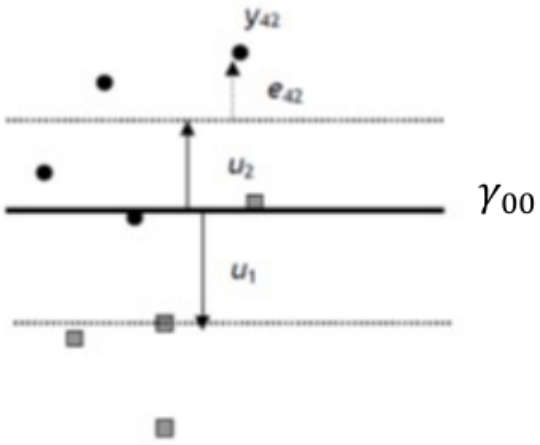


Figure 2.4: Modello ad effetti casuali non condizionato

Important remark 6. Per questo modello è possibile effettuare una **scomposizione di variabilità**

$$\text{Var}\left[y_{ij}\right] = \text{Var}\left[\gamma_0 + u_{0j} + \varepsilon_{ij}\right] = \sigma^2_{u_0} + \sigma^2_{\varepsilon}$$

Definition 2.3.1. Data la scomposizione di sopra possiamo definire il seguente indice, detto **coefficiente di correlazione intraclasse** (ICC), interpretabile come una percentuale di variabilità dovuta al raggruppamento

$$\rho = \text{Corr}\left(y_{ij}, y_{i'j}\right) = \frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_{\varepsilon}} = \frac{\text{cluster variance}}{\text{total variance}}, \quad \rho \in [0,1]$$

Important remark 7. 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/variabilità tra gruppi, suggerisce di evitare la modellizzazione a piu livelli e di ricorrere ai tradizionali modelli regressivi ad un solo livello.
  Indicativamente per applicare modelli multilevel vogliamo un ICC $> 0.2/0.3$

Example 2.3.2. Per la stima del modello multilevel non condizionato usiamo `lme4::lmer` dove:

- per indicare una intercetta casuale a livello di scuola aggiungiamo nella formula `(1 | schid)`

- `REML=FALSE`: REML sta per restricted ML e qui la 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. Si vedrà maggiormente in seguito

Per procedere alla stima del modello

```
summary(m0 <- lmer(math ~ (1 | schid), data=db, REML = FALSE))

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: math ~ (1 | schid)
##    Data: db
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##    1880.8    1891.5    -937.4    1874.8       257
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.33638 -0.65775 -0.08406  0.54767  2.87201
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  schid    (Intercept) 30.54    5.526
##  Residual             72.24    8.499
## Number of obs: 260, groups:  schid, 10
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   48.872      1.835   26.63
```

Notiamo che:

- la stima di $\widehat{\gamma_0} = 48.8$ e il t-value è 26.6: rispetto al modello semplice non gerarchico cambiano entrambe sia la stima dell'intercetta media e in particolare l'errore standard conduce ad un t-value inferiore (ed un p-value meno, se pur ancora largamente, significativo);

- la stima varianza residua $\widehat{\sigma^2_\varepsilon} = 72.2$ mentre quella associata all'intercetta casuale è $\widehat{\sigma^2_{u_0}} = 30.5$; la stima della varianza complessiva $\widehat{\text{Var}\,[y_{ij}]}$ è dunque

$$\widehat{\text{Var}\,[y_{ij}]} = 30.54 + 72.24 = 102.78$$

- per avere un confronto rispetto al modello non gerarchico abbiamo diminuito la variabilità dell'errore/residuo, portandola la sd da 11.1 a 8.5 (nel modello lineare riportata la sd, non la varianza). Nostro obiettivo nel prosieguo è diminuire quanto più la stima della variabilità residua $\widehat{\sigma^2_\varepsilon}$ (compatibilmente con la complessità del modello);

- i gradi di liberta residui son 257 (260 soggetti e tre parametri stimati, le due varianze e l'intecetta).

Per la stima dell'ICC $\widehat{\rho}$ abbiamo

```
(ICC <- 30.54 / (30.54 + 72.24)) # coefficiente di correlazione intraclasse

## [1] 0.2971395
```

quindi il 29,7% della varianza del punteggio in matematica è dovuta al raggruppamento nelle diverse scuole. Questo è un valore non ignorabile e che giustifica l'utilizzo di un modello multilevel.
In questo setting possiamo ottenere anche:

- la componente $\widehat{u_{0j}}$, residuo tra intercetta generale ($\widehat{\gamma_0}$) e l'intercette di ogni scuola

```
(u <- ranef(m0, condVar = TRUE))

## $schid
##    (Intercept)
## 1   -2.8408051
## 2   -6.0111939
## 3    3.9852031
## 4   -4.8095489
## 5    0.8953201
## 6   -2.2106384
## 7   13.4732174
## 8    0.7141693
## 9   -2.2817453
## 10  -0.9139783
##
## with conditional variances for "schid"
```

- le intercette per ogni scuola (stimate attraverso $\widehat{\gamma_0} + \widehat{u_{0j}}$)

```
coef(m0)

## $schid
##    (Intercept)
## 1     46.03126
## 2     42.86087
## 3     52.85727
## 4     44.06251
## 5     49.76738
## 6     46.66142
## 7     62.34528
## 8     49.58623
## 9     46.59032
## 10    47.95808
##
## attr(,"class")
## [1] "coef.mer"
```

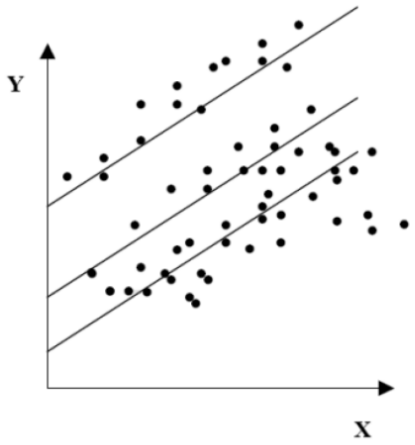


Figure 2.5: Intercetta casuale e pendenza fissa

2.3.3 Modello condizionato: introduzione di variabili indipendenti

Remark 5. Come nella regressione classica, l'introduzione di variabili esplicative (indipendenti) permette di spiegare meglio la variabilità del fenomeno oggetto di studio (Y); qui però in aggiunta possiamo tenere in considerazione la struttura gerarchica dei dati.

Remark 6. Considerare covariate fa sì che varianza "tra" ed "entro" gruppi si modificano e la differenza rispetto al modello non condizionato permette di valutare il contributo delle variabili esplicative.

Remark 7. Di solito si procede per passi introducendo variabili al primo livello (singolo individuo), poi al secondo (scuola) etc..

2.3.3.1 Intercetta casuale e pendenza fissa

In questo modello introduciamo un regressore di primo livello X ma ipotizziamo che il coefficiente di regressione (pendenza) sia costante per ogni gruppo.

- il modello di livello 1 ipotizzato (a livello di individuo) è

$$y_{ij} = \beta_{0j} + \beta_1 x_{ij} + \varepsilon_{ij}$$

- il modello di livello 2 (scuola) ipotizzato è

$$\begin{cases} \beta_{0j} = \gamma_0 + u_{0j} \\ \beta_1 = \gamma_1 \end{cases}$$

  Dove:

  - γ_0 è la 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_1 pendenza fissa: effetto medio della variabile X sulla variabile dipendente

- il **modello combinato** (rappresentato in figura 2.5) è

$$y_{ij} = \underbrace{\gamma_0 + \gamma_1 x_{ij}}_{\text{parte fissa}} + \underbrace{u_{0j} + \varepsilon_{ij}}_{\text{parte casuale}}$$

- Le **ipotesi sugli errori** di primo e secondo livello sono le stesse del modello non condizionato alla quale si aggiunge l'ipotesi di indipendenza di tutti i termini di errore delle variabili esplicative:

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

  Si ha che

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

  Lo stesso vale se ho più variabili indipendenti.

Example 2.3.3. La stima del modello con intercetta casuale ed effetto fisso avviene mediante

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

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: math ~ homework + (1 | schid)
##    Data: db
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##    1850.7    1864.9    -921.3    1842.7       256
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.6180 -0.6971 -0.0237  0.5993  3.3745
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  schid    (Intercept) 22.50    4.744
##  Residual             64.26    8.016
## Number of obs: 260, groups:  schid, 10
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  44.9784     1.7248  26.077
## homework      2.2143     0.3777   5.863
##
## Correlation of Fixed Effects:
##          (Intr)
## homework -0.387
```

- l'intercetta, media delle intercette nella popolazione delle scuole (γ_0), è 44.98
- il coefficiente di homework 2.21: ciascuna scuola ha lo stesso coefficiente
- la varianza totale è $22.5 + 64.26 = 86.76$
- i gradi liberta sono scesi perché abbiamo aggiunto la stima di un coefficiente (homework) in più
- AIC e BIC, comparati con il precedente dicono che sono meglio rispetto al precedente (ridotti): se i criteri sono discorsi considerare il BIC
- il modello ci restituisce anche una stima della correlazione tra coefficienti fissi (intercetta ed homework): è negativa suggerendo che chi parte da una intercetta più alta, ha poi un effetto delle ore di studio più basso

Per l'icc la stima è

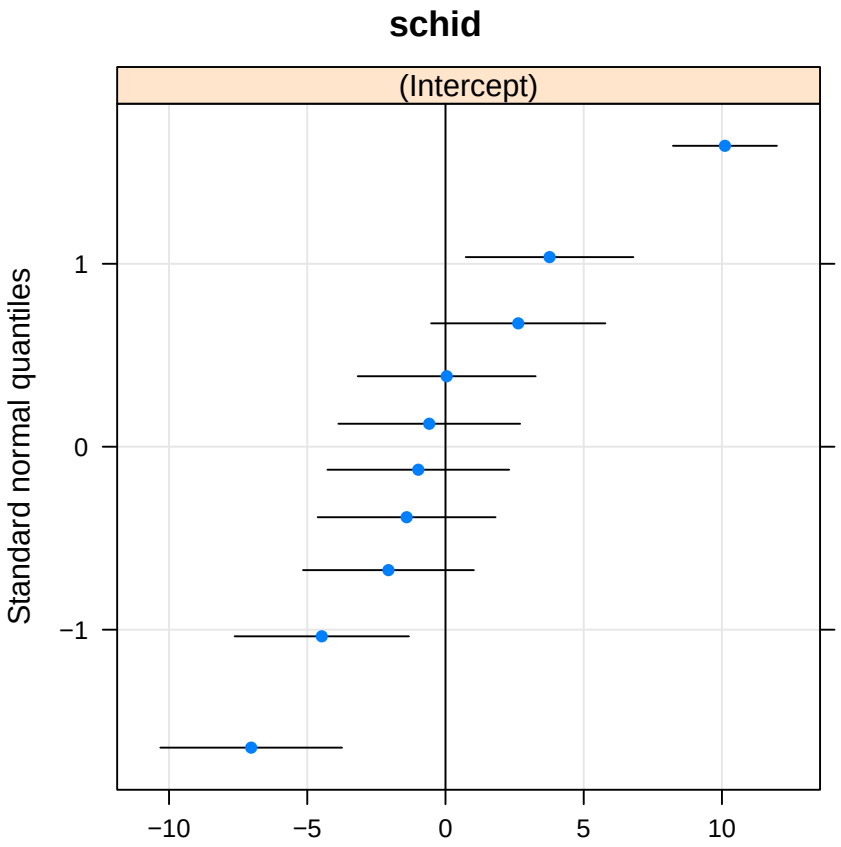
```
# coefficiente di correlazione intraclasse
(ICC = 22.50 / (22.50 + 64.26))

## [1] 0.2593361
```

Ossia 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 (prima ra quasi 0.3: si è ridotto perché abbiamo aggiunto informazione con una variabile indipendente).
Infine stima di coefficienti (dove l'intercetta è a livello di scuola, mentre homework effetto fisso è costante)

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

## $schid
##    (Intercept) homework
## 1     42.91453 2.214345
## 2     37.94979 2.214345
## 3     48.74252 2.214345
## 4     40.50288 2.214345
## 5     47.60971 2.214345
## 6     43.99404 2.214345
## 7     55.08608 2.214345
## 8     45.02126 2.214345
## 9     43.57209 2.214345
## 10    44.39092 2.214345
##
## attr(,"class")
## [1] "coef.mer"
```

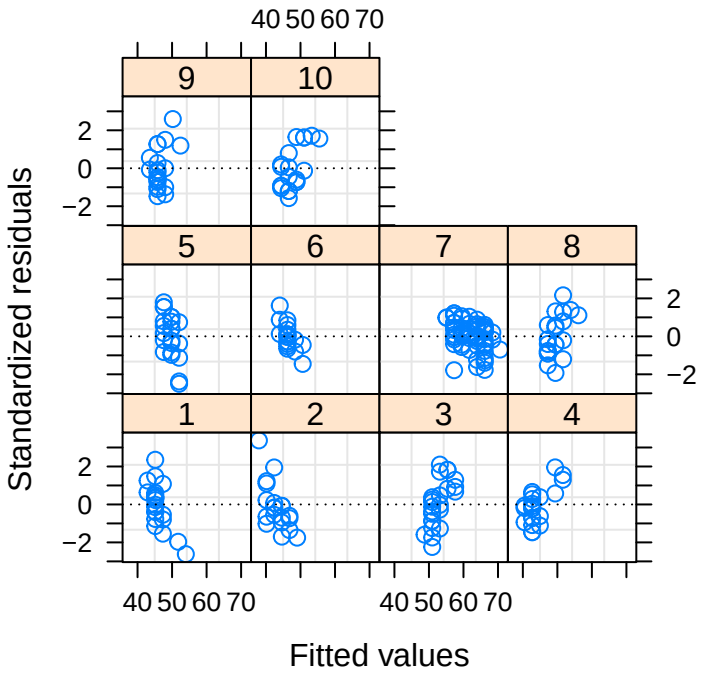
Stima e anche rappresentazione degli effetti random (in questo caso lo scostamento della sola intercetta dalla media globale). Per una rappresentazione:

```
(u <- ranef(m1, condVar = TRUE))

## $schid
##    (Intercept)
## 1  -2.06384967
## 2  -7.02859010
## 3   3.76413254
## 4  -4.47550382
## 5   2.63133101
## 6  -0.98434077
## 7  10.10769598
## 8   0.04287378
## 9  -1.40628841
## 10 -0.58746054
##
## with conditional variances for "schid"

qqmath(ranef(m1))

## $schid
```

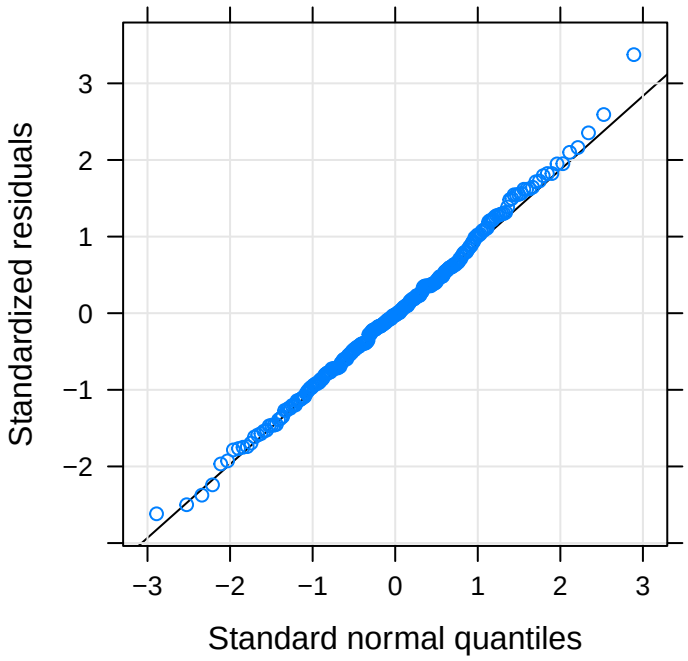


Per una rappresentazione grafica dei residui vediamo in ordine ...

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




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



Infine per una rappresentazione grafica del modello stimato

```
db$Predictions <- fitted(m1)
m1_est <- coef(summary(m1))[ , "Estimate"]
m1_se <- coef(summary(m1))[ , "Std. Error"]
palette(rainbow(10))
gg <- ggplot(db, aes(y = math, x = homework)) +
  geom_line(aes(y = Predictions, color=as.factor(schnum))) +
  geom_abline(slope = m1_est[2], intercept = m1_est[1], size=1) +
  geom_point(size = 1.5, alpha = 0.8, colour=factor(db$schnum)) +
  theme(legend.position="none")
print(gg)
```

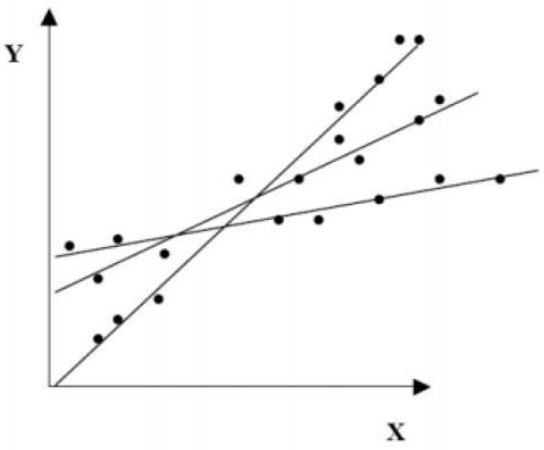
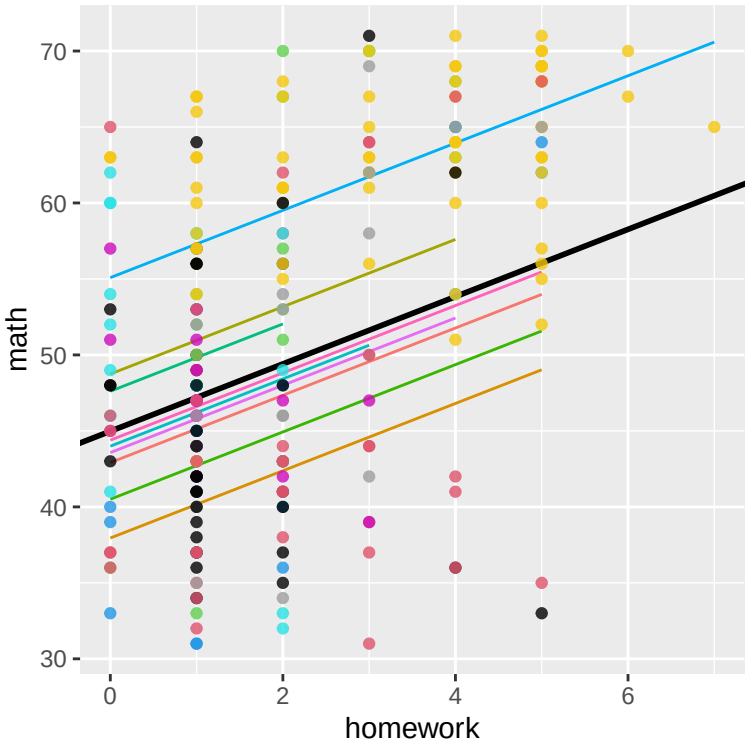


Figure 2.6: Intercetta e pendenza random



2.3.3.2 Intercetta e pendenza casuali

In questo caso consideriamo sia intercetta che pendenza come variabili aleatorie, non fisse. Si ha che

- il modello di livello 1 è

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

- il modello di livello 2

$$\begin{aligned} \beta_{0j} &= \gamma_0 + u_{0j} \\ \beta_{1j} &= \gamma_1 + u_{1j} \end{aligned}$$

- modello combinato diviene (fig 2.6)

$$y_{ij} = \underbrace{\gamma_0 + \gamma_1 x_{ij}}_{\text{parte fissa}} + \underbrace{u_{1j} x_{ij} + u_{0j} + \varepsilon_{ij}}_{\text{parte casuale}}$$

  rappresentanto in fig 2.6. In questo modello quindi abbiamo l'equivalente di un modello di regressione (separato) per ogni scuola: una intercetta β_{0j} per ogni scuola e un coefficiente β_{1j} per ogni scuola;

- gli errori di secondo livello (**effetti casuali**) sono caratterizzati dalle seguenti variabilità

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

  Gli *effetti casuali* consistono nella differenza del parametro del j-esimo cluster dalle corrispondenti medie (per intercetta e pendenza, separatamente) della popolazione. L'introduzione di covariate può aiutare a ridurre queste varianze.

- le assunzioni distributive che facciamo (normalità è una delle possibili assunzioni) per errori ed effetti random:

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

  notiamo che $\text{Var}\left[\varepsilon_{ij}\right] = \sigma^2_\varepsilon$ è costante e non dipendende dal j-th gruppo di appartenenza.
  Per quanto riguarda i parametri

$$\begin{bmatrix} \beta_{0j} \\ \beta_{1j} \end{bmatrix} iid\, \text{MVN}\left(\begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u_0} & \sigma_{u_{01}} \\ \sigma_{u_0 1} & \sigma^2_{u_1} \end{bmatrix} \right)$$

  Come **parametri fissi** abbiamo γ_0 (la media delle intercette) e γ_1 (la media dei coefficienti di regressione).
  Come **parametri causali**: $\sigma^2_{u_0}$ (varianza dell'intercetta), $\sigma^2_{u_1}$ (varianza della pendenza), σ^2_ε (varianza residua del livello 1).
  A questi si aggiunge $\sigma_{u_{01}}$ ossia la 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 "casuali" sono quantità *fisse*, se pur sconosciute, nella popolazione; la casualità rimane legata alla parte ε_{ij} casuale del modello

- la parte casuale totale è $u_{0j} + u_{1j}x_{ij} + \varepsilon_{ij}$ e implica **eteroschedasticità**, dato che la variabilità delle osservazioni (condizionate al valore assunto dalla covariate)

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

  non è costante ma dipende anche da x_{ij}, il valore assunto dalla covariata. In particolare la varianza tra le cluster è funzione quadratica di X. Da questo deriva il fatto che *in questo modello non si può calcolare l'ICC* (la varianza condizionata che sarebbe al denominatore non è costante). In generale nei modelli con pendenza random non si può calcolare l'ICC

- ulteriormente
  - La varianza dell'intercetta $\sigma^2_{u_0}$ e la covarianza intercetta-coefficiente $\sigma_{u_{01}}$ dipendono da X
  - la correlazione eterogenea entro i cluster

- Infine nel modello vi è incrocio delle rette, i cluster non sono ordinabili.

Important remark 8. Il modello a (sola) intercetta casuale senza covariata (fig 2.5)

$$y_{ij} = \gamma_0 + \gamma_1 x_{ij} + u_{0j} + \varepsilon_{ij}$$

è un caso speciale del modello generale dove:

- la varianza/covarianza degli effetti random è

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

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

$$\begin{aligned} &\text{Var}\left[y_{ij}|x_{ij}\right] = \sigma^2_{u_0} + \sigma^2_{\varepsilon} \\ &\text{Cov}\left(y_{ij}, y_{i'j}|x_{ij}, x_{i'j}\right) = \sigma^2_{u_0} \end{aligned}$$

- le rette sono parallele e i cluster "ordinabili"

Example 2.3.4. Vediamo anche qui il passaggio dal modello di regressione semplice (avente `homework` tra le covariate) all'equivale Nel caso di regressione standard ciascuna suola ha stessa intercetta e pendenza: una unica retta. La stima diviene

```
summary(lm_homework <- lm(math ~ homework, data=db))
```

```
##
## Call:
## lm(formula = math ~ homework, data = db)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.9331  -6.6457   0.3543   7.0669  20.9261
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  44.0739     0.9886   44.58   <2e-16 ***
## homework      3.5719     0.3882    9.20   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.682 on 258 degrees of freedom
## Multiple R-squared:  0.247,Adjusted R-squared:  0.2441
## F-statistic: 84.64 on 1 and 258 DF,  p-value: < 2.2e-16
```

L'intercetta è 44.07, la pendenza 3.57 e la varianza del termine di errore $9.68^2 = 93.7$. Passando al modello con intercetta e pendenza casuale la stima

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

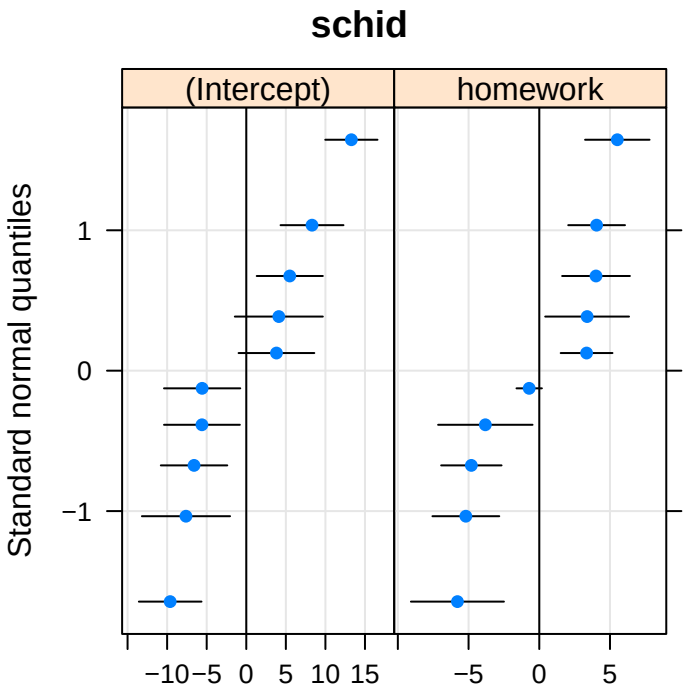
```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: math ~ homework + (homework | schid)
##    Data: db
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##    1781.4    1802.7    -884.7    1769.4       254
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.49760 -0.54240  0.02171  0.60686  2.57103
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  schid    (Intercept) 61.81    7.862
##           homework    19.98    4.470    -0.80
##  Residual             43.07    6.563
## Number of obs: 260, groups:  schid, 10
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   44.773      2.603  17.199
## homework       2.049      1.472   1.392
##
## Correlation of Fixed Effects:
##          (Intr)
## homework -0.803
```

dove

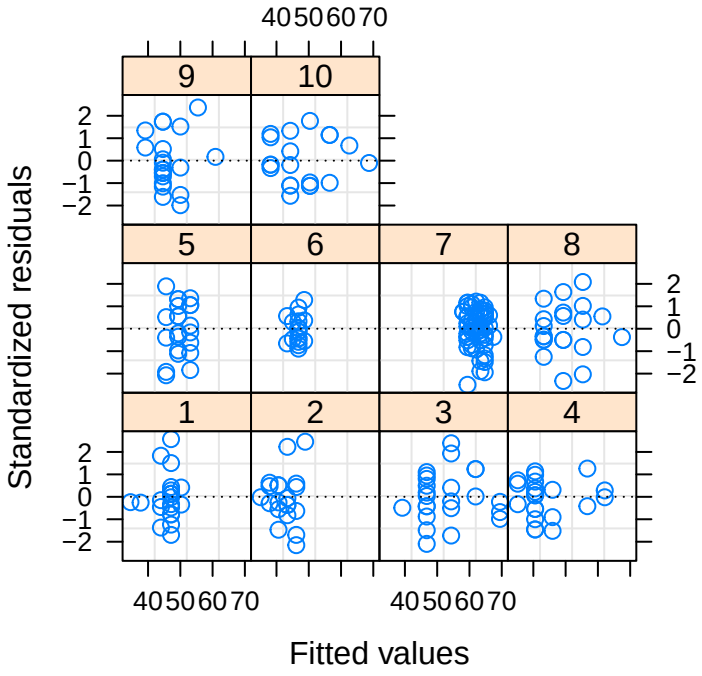
- la pendenza di homework 2.05 è una media del coefficiente nella popolazione
- l'intercept variance è 61.81
- l'homework variance è 19.98
- l'intercept-homework covariance è -28.26, corrisponde ad una correlazione di circa -0.8 (R non da la covarianza ma la correlazione dei fixed effects)
- la residual (lev 1) variance è 43.07: nel passaggio
- *non si può calcolare ICC*
- 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
##


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

## $schid
##    (Intercept)  homework
## 1     50.27092 -3.143413
## 2     48.88625 -2.754007
## 3     39.19525  7.566359
## 4     35.15660  5.394143
## 5     53.08113 -3.738260
## 6     48.58599 -1.765215
## 7     58.05543  1.335178
## 8     37.15228  6.060031
## 9     39.16965  5.430490
## 10    38.17276  6.101708
##
## attr(,"class")
## [1] "coef.mer"

# 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)
```

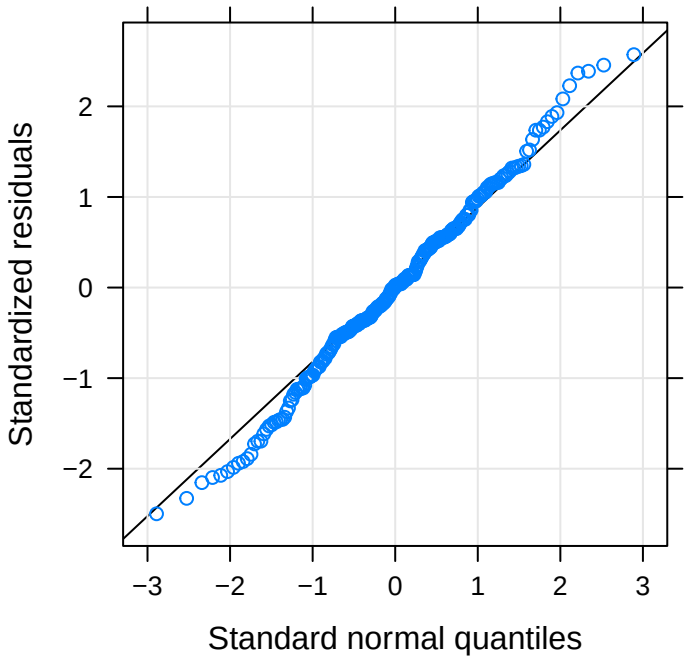
```
## Data: db
## Models:
## m1: math ~ homework + (1 | schid)
## m2: math ~ homework + (homework | schid)
##    npar    AIC    BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
## m1    4 1850.7 1864.9 -921.33    1842.7
## m2    6 1781.4 1802.8 -884.69    1769.4 73.272  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## -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))

## $schid
```



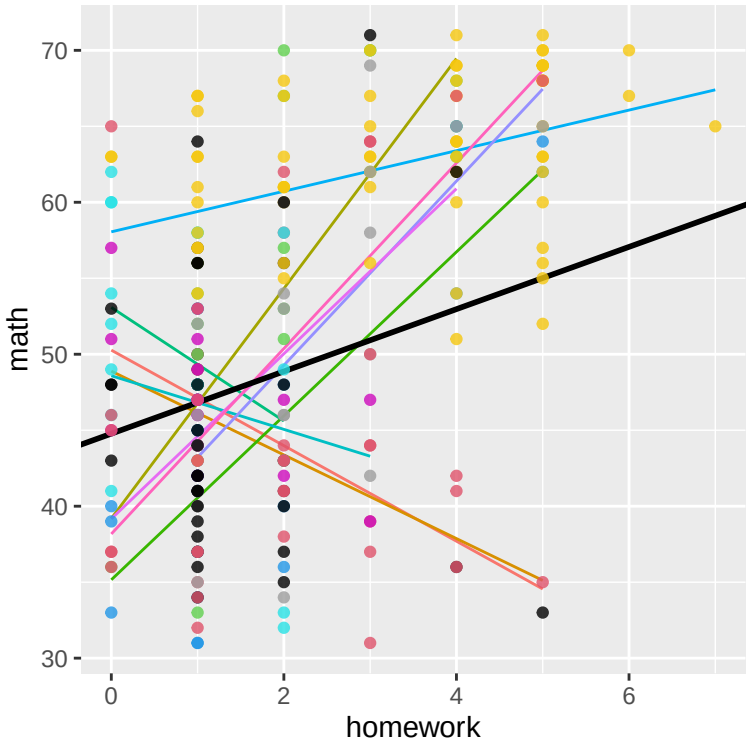
```
#analogo
# dotplot(ranef(model4, condVar=TRUE))

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




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




```
## rappresentazione grafica delle diverse (quella nera di riferimento)
db$Predictions <- fitted(m2)
m2_est <- coef(summary(m2))[ , "Estimate"]
m2_se <- coef(summary(m2))[ , "Std. Error"]
palette(rainbow(10))
gg <- ggplot(db, aes(y = math, x = homework)) +
  geom_line(aes(y = Predictions, color=as.factor(schnum))) +
  geom_abline(slope = m2_est[2], intercept = m2_est[1], size=1) +
  geom_point(size = 1.5, alpha = 0.8, colour=factor(db$schnum)) +
  theme(legend.position="none")
print(gg)
```



2.3.3.3 Modello con covariate di livello 1 e 2

L'introduzione di una covariata W al secondo livello fa si che si possa utilizzare caratteristiche dei cluster per:

- definire un modello (con covariata) anche per i parametri del livello 1 (β_{0j}, β_{1j})
- aumentare la spiegazione e ridurre la varianza del livello 2 $(\sigma^2_{u_0}, \sigma^2_{u_1})$

Si ha che il

- modello di livello 1 rimane:

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

- i modelli di livello 2 divengono:

$$\begin{cases} \beta_{0j} = \gamma_0 + \gamma_{01} w_j + u_{0j}, & \text{Var}\left[u_{0j}\right] = \sigma^2_{u_0} \\ \beta_{1j} = \gamma_1 + \gamma_{11} w_j + u_{1j}, & \text{Var}\left[u_{1j}\right] = \sigma^2_{u_1} \end{cases} \quad 0$$

- modello combinato

$$y_{ij} = \underbrace{\gamma_0 + \gamma_{01} w_j + \gamma_1 x_{ij} + \gamma_{11} w_j x_{ij}}_{\text{parte fissa}} + \underbrace{u_{0j} + u_{1j} x_{ij} + e_{ij}}_{\text{Parte casuale}}$$

dove:

- γ_0 è il valore medio di y quando sia x sia w sono pari a zero
- γ_{01} è l'effetto di w sulla intercetta e indica la variazione dell'intercetta media all'aumentare unitario di w
- γ_1 è effetto di X su Y quando W è uguale a zero
- γ_{11} (coefficiente di interazione cross-level) indica la variazione della pendenza all'aumentare di W $\implies$ effetto moderatore di W sulla relazione tra X e Y (ad esempio permettiamo che l'effetto di homework a livello individuale sia differente per scuole pubbliche e private)
- c'è una combinazione tra livelli $\gamma_{11} w_j x_{ij}$

Example 2.3.5. Per la stima impostiamo

```
summary(m3 <- lmer(math ~ homework + (1 + homework | schid) + public,
                   data = db, REML = FALSE))

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: math ~ homework + (1 + homework | schid) + public
##    Data: db
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##    1764.8    1789.8    -875.4    1750.8       253
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.63740 -0.56382 -0.05233  0.63932  2.61850
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  schid    (Intercept) 40.68    6.378
##           homework    21.68    4.657    -0.98
##  Residual             42.95    6.554
## Number of obs: 260, groups:  schid, 10
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   58.056      2.695  21.545
## homework       1.941      1.525   1.273
## public       -14.651      1.832  -7.999
##
## Correlation of Fixed Effects:
##          (Intr) homwrk
## homework -0.772
## public   -0.602  0.010
```

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)
- dato che sia intercetta che pendenza sono termini casuali abbiamo la stima della varianza delle intercette 40, la stima della varianza di homework 21 ed una alta correlazione negativa tra le due (quasi -1)
- infine la varianza residua è 43

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

Example 2.3.6. Se al modello precedente aggungiam una interazione tra public e homework abbiamo il seguente risultato (variabile esplicativa di secondo livello che agisce sulla intercetta e sul coefficiente di regressione della variabile esplicativa di primo livello)

```
summary(m4 <-lmer(math ~ homework + (homework| schid) + public + public:homework,
                  data = db,
                  REML = FALSE))

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: math ~ homework + (homework | schid) + public + public:homework
##    Data: db
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##    1766.8    1795.3    -875.4    1750.8       252
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.64037 -0.56364 -0.05374  0.63881  2.61814
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  schid    (Intercept) 40.51    6.364
##           homework    21.58    4.645    -0.98
##  Residual             42.95    6.554
## Number of obs: 260, groups:  schid, 10
##
## Fixed effects:
##                 Estimate Std. Error t value
## (Intercept)      59.2102     6.5978   8.974
## homework          1.0946     4.6688   0.234
## public          -15.9419     6.9778  -2.285
## homework:public   0.9472     4.9384   0.192
##
## Correlation of Fixed Effects:
##             (Intr) homwrk public
## homework    -0.966
## public      -0.946  0.913
## homwrk:pblc  0.913 -0.945 -0.965
```

Nella stima:

	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
M4	1766.8	1795.3	-875.4	1750.8	252

Table 2.1: Sintesi modelli

- 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)

2.3.4 Confronto tra diversi modelli

Riassumendo in tabella 2.1

- 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
- M4 covariata al livello 2 che agisce solo su intercetta e pendenza

AIC e BIC concordano nel dire che M3 è il modello complessivamente migliore e non c'è bisogno di aggiungere l'interazione tra public e homework.

2.4 Metodi di stima

I metodi di stima disponibili:

- basati su massima verosimiglianza: Full information (FIML) o Restricted (REML), vediamo questi
- Minimi quadrati generalizzati (Generalized Least square GLS): sono OLS pimped con metodi iterativi
- Equazioni di stima generalizzate (Generalize Estimating equation GEE)
- Inferenza bayesiana

2.4.1 Metodi di massima verosimiglianza

Il modello in forma generale/estesa

$$y_{ij} = \underbrace{\gamma_0 + \gamma_{01} w_j + \gamma_1 x_{ij} + \gamma_{11} w_j x_{ij}}_{\text{Parte fissa}} + \underbrace{u_{0j} + u_{1j} x_{ij} + \varepsilon_{ij}}_{\text{parte casuale}}$$

con i metodi di massima verosimiglianza non è possibile stimare tutti questi parametri ma si procede in due passi:

1. si inizia dalla stima del *parametri fissi* ($\gamma_0, \gamma_1, \gamma_{10}, \gamma_{11}$ e dei parametri di *varianza e covarianza*: al primo passo si stimano col metodo ols classico (ipotizzando indipendenza tra le unità); ($\sigma^2_\varepsilon, \sigma^2_{u0}, \sigma^2_{01}, \sigma^2_{u0}$)

2. si va alla *previsione* degli effetti casuali (u_{0j}, u_{1j}, $j = 1, \ldots, J$)

3. si reitera i due step di sopra fino ad arrivare a convergenza (e le stime si muovono di poco): a parte il primo passaggio dove si usa OLS, poi si usano i GLS che tengono conto della correlazione tra unità

2.4.1.1 FIML vs REML

Per quanto riguarda **FIML vs REML**:

- con FIML (verosimiglianza classica) vi può essere distorsione sulla stima dei param casuali (soprattutto su campioni non numerosi):
  - vi è una stima congiunta 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 invece i parametri casuali sono stimati
  - usando la verosimiglianza "ristretta", cioè basata sulla densità dei residui;
  - in modo appropriato anche in piccoli campioni

Se l'obiettivo è avere una stima delle componenti casuali meglio usare REML.

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

- Stime simili per σ^2
- Stime discordanti per i parametri 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

2.4.1.2 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.

2.5 Valutazione della bontà del modello

2.5.1 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.

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

2.5.2 Test sulla bontà d'adattamento

Si ha che:

- si definisce *Devianza* la quantità $-2\log(L)$, con L funzione di verosimiglianza. I modelli con devianza inferiore presentano un miglior adattamento ai dati;
- è possibile confrontare statisticamente (attraverso le devianze) due modelli *annidati*, ossia dove un modello specifico può essere derivato da un modello più generale rimuovendo uno o più parametri;
- 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.

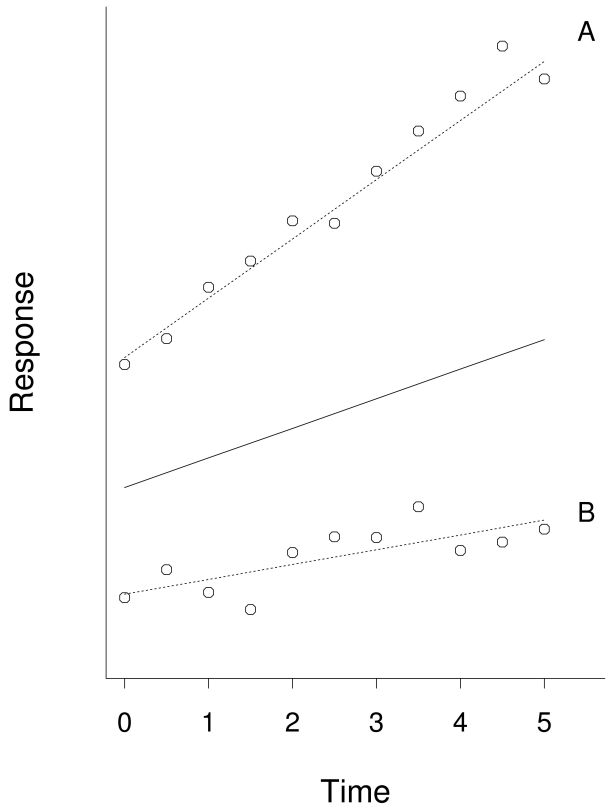


Figure 2.7: Conditional/marginal mean

2.6 Medie condizionate e marginali

Fino ad ora il focus è stato sulla stima delle componenti random (varianze in primis); ci interessa però la stima del valore atteso della variabile dipendente

Consideriamo come esempio il seguente modello ad effetti misti (pendenza e intercetta casuale, con covariata a livello 1):

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

C'è una importante distinzione tra:

- la **media condizionata** della variabile aleatoria Y

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

  la media condizionata è stima di un valore atteso per un individuo che appartiene a uno specifico gruppo (è una media condizionata al regressore e alle componente casuali. In figura 2.7 sono le rette tratteggiate A e B

- e la **media marginale**

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

e la stima per un soggetto "medio" senza considerare l'effetto/appartenenza di gruppo. Nella media sonopresenti di interesse gli effetti fissi. È la retta scura in 2.7

Remark 8. La media marginale è data dall'output standard mentre per la media condizionata abbiamo bisogno delle componenti previste u_{1j} e u_{0j}, vediamo come prevederle al prossimo punto; so far u_{0j}, u_{1j} non sono proprio stimate (stimiamo la rispettive variabilita/covarianza); andiamo a vedere come predirle è una predizione.

2.7 Previsione degli effetti casuali

Una volta stimate le componenti del modello (le varianze/covarianze degli effetti casuali in particolare) possiamo stimare ad esempio u_{0j}, u_{1j}.

In molte applicazioni (dove si vogliono le medie condizionate, I suppose) l'inferenza si concentra sui parametri fissi: ad esempio γ_0, β_1, in un modello ad intercetta casuale

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

Per quanto riguarda le componenti random, es u_{0j}, il focus della stima è stata la variabilità; ma può essere di interesse "prevedere" i fattori casuali specifici per gruppo.
Si può dimostrare che lo stimatore (BLUP) per u_{0j} è:

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

Dove:

$$\mu_{ij} = \gamma_0 + \beta_1 x_{ij}$$
$$\tau = \frac{n_j \sigma_u^2}{n_j \sigma_u^2 + \sigma_\varepsilon^2}$$

Si può notare che:

- μ_{ij} è la stima della media marginale (non considerando gruppi);
- lo stimatore $\widehat{u_{0j}}$ (considerando j fisso, ossia stiamo considerando uno specifico gruppo) è una media ponderata tra un residuo medio entro il gruppo j-esimo (differenza tra singolo valore osservato e media marginale) e il valore atteso di u_{0j} (ovvero 0, al secondo termine);
- il peso della media è dato da τ, detto *shrinkage factor*: questo è un numero compreso fra 0 e 1: τ aumenta se
  - aumenta la numerosità n_j del gruppo: per gruppi molto grandi si tenderà a shrinkare verso/utilizzare la media degli scarti dalla media marginale, per gruppi piccoli si tende a shrinkare verso 0. E' una info che tiene conto della quantità di info raccolte nei diversi gruppi;
  - all'aumentare della variabilità tra gruppi σ_u^2 (che sarebbe poi $\sigma_{u_0}^2$) rispetto a quella within σ_ε^2

Quindi τ fa spostare il residuo totale stimato medio $\widehat{u_{0j}}$ in modo differenziato a seconda della numerosità del gruppo n_j e del rapporto fra le componenti di varianza

- Lo shrinkage:
  - sarà più forte (verso 0) 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^2) è piccola rispetto a quella within;
  - 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;
  - 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.

2.8 Note finali

Remark 9. Alcuni tutorial multilevel (che utilizzano notazione differente) in ambito medico: il primo è introduttivo (modello a intercetta casuale), il secondo presenta un modello con covariate di 1 livello ed effetti casuali anche su coeff di regressione:

1. merlo et al 2005: a brief conceptual tutorial of multilevel analysis in social epidemiology: linking the statistical concept of clustering to the idea of contextual phenomemon
2. merlo et al 2005: a brief conceptual tutorial of multilevel analysis in social epidemiology: investigating contextual phenomena in different groups of people

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.

Chapter 3

Modelli per dati longitudinali

Remark 10. Libri di testo di riferimento per questa parte sono:

- J.D. Singer, J.B. Willet Applied Longitudinal Data Analysis, Oxford University Press, 2003 (più semplice)
- esempi tratti da G.M Fitzmaurice, N. M. Laird, J. H. Ware, Applied Longitudinal Analysis, Wiley, 2004 (più completo, fa vedere anche cose vecchie tipo l'anova per misure ripetute)

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*.
  In relazione alle occasioni temporali possiamo avere:
  - individui rilevati nelle stesse occasioni (*dataset* strutturati nel tempo) o no (occasioni di rilevazione sono diverse);
  - misurazioni *equispaziate o non* (non vi sono problemi su questo);
  - individui rilevati in un numero uguale di occasioni *disegno bilanciato* o non (numero diverso di rilevazioni per individuo);

	variabile				
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

Il *numero dei tempi* di rilevazione disponibili può essere variabile (es magari alcuni soggetti non rispondono o tempi di rilevazione proprio variabili), 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*.
Più valutazioni nel tempo intra soggetto vi sono e più possiamo applicare modelli funzionalmente elaborati (noi qui ci limiteremo a relazioni lineari del tempo, ma si potrebbe complicare dati permettendo);

- selezione di una opportuna *metrica del tempo* (es misuriamo il tempo su una scala originaria o su una trasformazione del tempo; rispettivamente in giorni, settimane etc). La scelta è chiaramente legata al fenomeno che stiamo analizzando [1]

Example 3.1.1 (Disegno bilanciato)**.** Considerando un disegno *bilanciato* con T tempi/osservazioni ripetute in n individui si genera tabella 3.1.

Important remark 10 (Notazione e assunzioni). 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;
- ogni individuo $\mathbf{y}_i = [y_{i1}, \ldots, y_{iT}]$ 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, $t = 1, \ldots, T$; le misurazioni entro individuo sono da considerarsi correlate

[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.

- 2 livello: *individui*, $i = 1, \ldots, n$; le misurazioni su soggetti diversi sono da considerarsi indipendenti

Example 3.1.2 (Esempio Tolerance (Fitzmaurice))**.** L'obiettivo è l'analisi della tolleranza verso comportamenti devianti/limite

- 590 individui, 5 occasioni di rilevazione da 11 a 15 anni di età, disegno bilanciato (tutte le rilevazioni per tutti)
- sono state rilevate una scala da 1 (comportamento ritenuto sbagliato) a 4 (non sbagliato) le seguenti variabili (giudizio morale su): 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_{it} con $i = 1, \ldots, 590$ individui e $t = 11, \ldots, 15$
- due covariate fisse/invarianti a livello individuale nel tempo:
  1. `sesso`: 1 = male, 0 = female
  2. `exposure`: indice di autovalutazione su esposizione a comportamenti limite alla età di 11 anni (scala 0-4)

```
## library(pastecs)
library(lattice) # per i grafici
library(ggplot2)
library(lme4)

## Analisi dati wide
## ------------------
tol_wide <- read.csv("data/tolerance1.txt") # in formato wide
if (FALSE) tol_wide <- read.csv("longitudinal_data_analysis/data/tolerance1.txt")
tol_wide # intero dataset, 16 righe, una per bambino (colonna id)

##       id tol11 tol12 tol13 tol14 tol15 male exposure
## 1      9  2.23  1.79  1.90  2.12  2.66    0     1.54
## 2     45  1.12  1.45  1.45  1.45  1.99    1     1.16
## 3    268  1.45  1.34  1.99  1.79  1.34    1     0.90
## 4    314  1.22  1.22  1.55  1.12  1.12    0     0.81
## 5    442  1.45  1.99  1.45  1.67  1.90    0     1.13
## 6    514  1.34  1.67  2.23  2.12  2.44    1     0.90
## 7    569  1.79  1.90  1.90  1.99  1.99    0     1.99
## 8    624  1.12  1.12  1.22  1.12  1.22    1     0.98
## 9    723  1.22  1.34  1.12  1.00  1.12    0     0.81
## 10   918  1.00  1.00  1.22  1.99  1.22    0     1.21
## 11   949  1.99  1.55  1.12  1.45  1.55    1     0.93
```

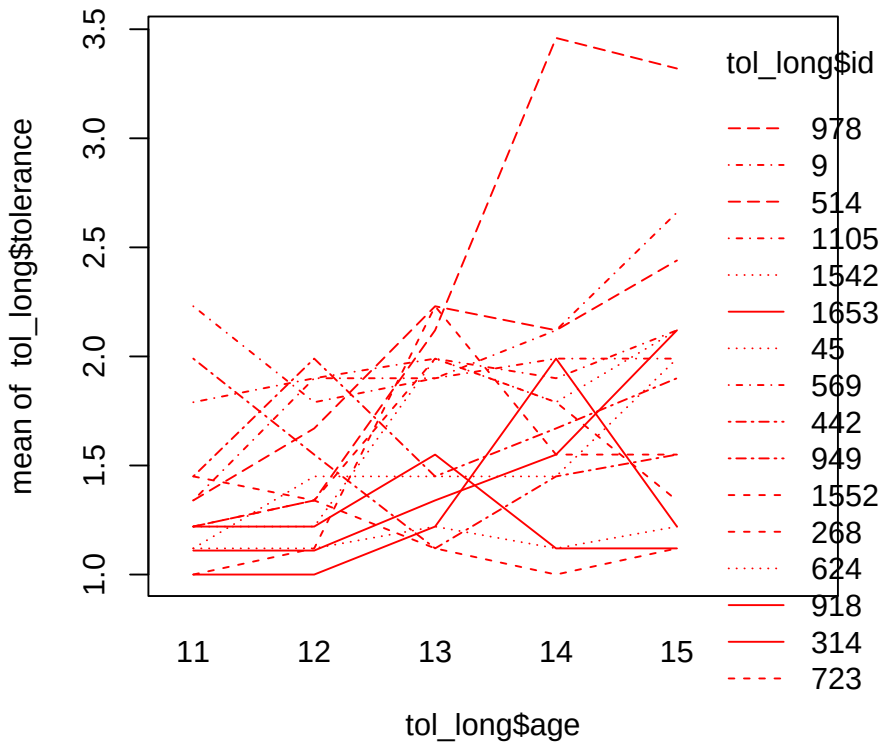
```
## 12  978  1.22  1.34  2.12  3.46  3.32    1     1.59
## 13 1105  1.34  1.90  1.99  1.90  2.12    1     1.38
## 14 1542  1.22  1.22  1.99  1.79  2.12    0     1.44
## 15 1552  1.00  1.12  2.23  1.55  1.55    0     1.04
## 16 1653  1.11  1.11  1.34  1.55  2.12    0     1.25

summary(tol_wide)

##        id             tol11           tol12           tol13
##  Min.   :   9.0   Min.   :1.000   Min.   :1.000   Min.   :1.120
##  1st Qu.: 410.0   1st Qu.:1.120   1st Qu.:1.195   1st Qu.:1.310
##  Median : 673.5   Median :1.220   Median :1.340   Median :1.725
##  Mean   : 762.8   Mean   :1.364   Mean   :1.441   Mean   :1.676
##  3rd Qu.:1009.8   3rd Qu.:1.450   3rd Qu.:1.700   3rd Qu.:1.990
##  Max.   :1653.0   Max.   :2.230   Max.   :1.990   Max.   :2.230
##      tol14           tol15            male           exposure
##  Min.   :1.000   Min.   :1.120   Min.   :0.0000   Min.   :0.8100
##  1st Qu.:1.450   1st Qu.:1.310   1st Qu.:0.0000   1st Qu.:0.9225
##  Median :1.730   Median :1.945   Median :0.0000   Median :1.1450
##  Mean   :1.754   Mean   :1.861   Mean   :0.4375   Mean   :1.1912
##  3rd Qu.:1.990   3rd Qu.:2.120   3rd Qu.:1.0000   3rd Qu.:1.3950
##  Max.   :3.460   Max.   :3.320   Max.   :1.0000   Max.   :1.9900

## il valore mediano di tol tende ad aumentare col tempo
## il valore mediano di exposure è 1.145 (che usiamo per dicotomizzare exposure)

## correlazione tra misurazione sulla stessa variabile a tempi successivi
cor(tol_wide[ , 2:6])

##            tol11     tol12      tol13     tol14     tol15
## tol11 1.00000000 0.6572937 0.06194915 0.1407631 0.2635371
## tol12 0.65729370 1.0000000 0.24755116 0.2056198 0.3922781
## tol13 0.06194915 0.2475512 1.00000000 0.5871742 0.5692116
## tol14 0.14076312 0.2056198 0.58717422 1.0000000 0.8254566
## tol15 0.26353705 0.3922781 0.56921163 0.8254566 1.0000000

## la correlazione tra tol al tempo 1 e 2 è forte poi tende a ridurre
## può essere una cosa ragionevole
## però tra 12 e 13 anni e 0.24, non è  cosi alta come le precedenti
## tra 13 e 14 0.57
## ecc
## questo dice che la matrice di correlazione non è composta da  valori tutti
## uguali tra loro


## Analisi dati long
## -----------------
tol_long <- read.csv("data/tolerance1_pp.txt") # formato long
if (FALSE) tol_long <- read.csv("longitudinal_data_analysis/data/tolerance1_pp.txt")
tol_long$id <- as.factor(tol_long$id)
head(tol_long)
```

```
##   id age tolerance male exposure time
## 1  9  11      2.23    0     1.54    0
## 2  9  12      1.79    0     1.54    1
## 3  9  13      1.90    0     1.54    2
## 4  9  14      2.12    0     1.54    3
## 5  9  15      2.66    0     1.54    4
## 6 45  11      1.12    1     1.16    0

# variabile fondamentale è age, exposure e time sono time invariant
# per semplificare la nostra analisi assumiamo che la variabile tempo parta da
# 0 a 4 (definita sulla base di age)
```

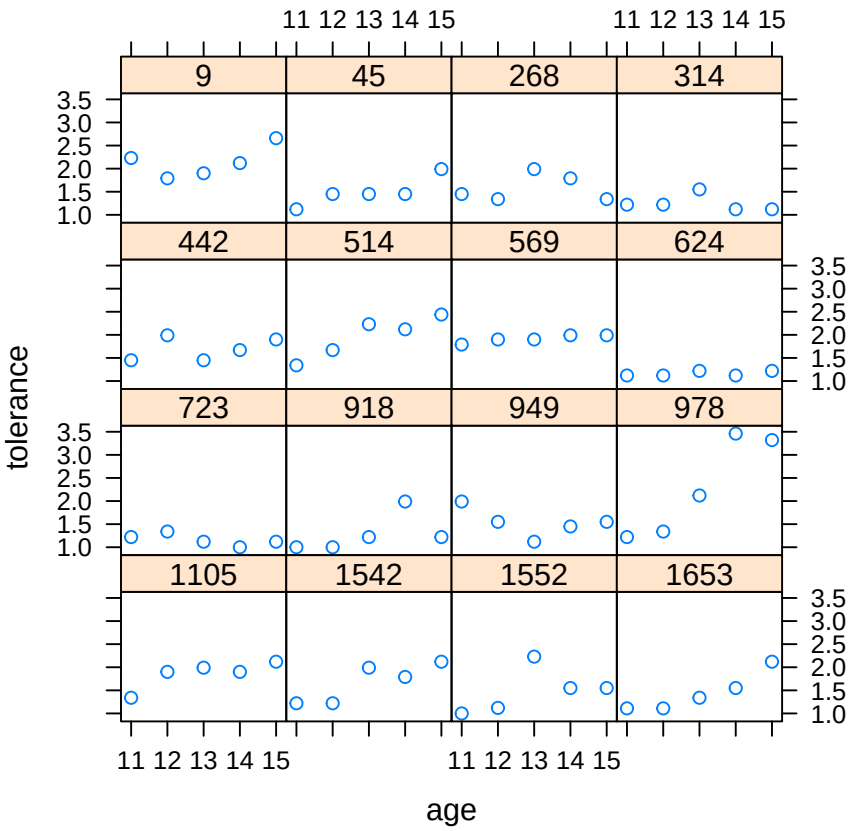
Una prima analisi fattibile è di descrivere i cambiamenti individuali nel tempo: riportare graficamente (scatterplot) la relazione tra variabile risposta e tempo. Alcune indicazioni:

- 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

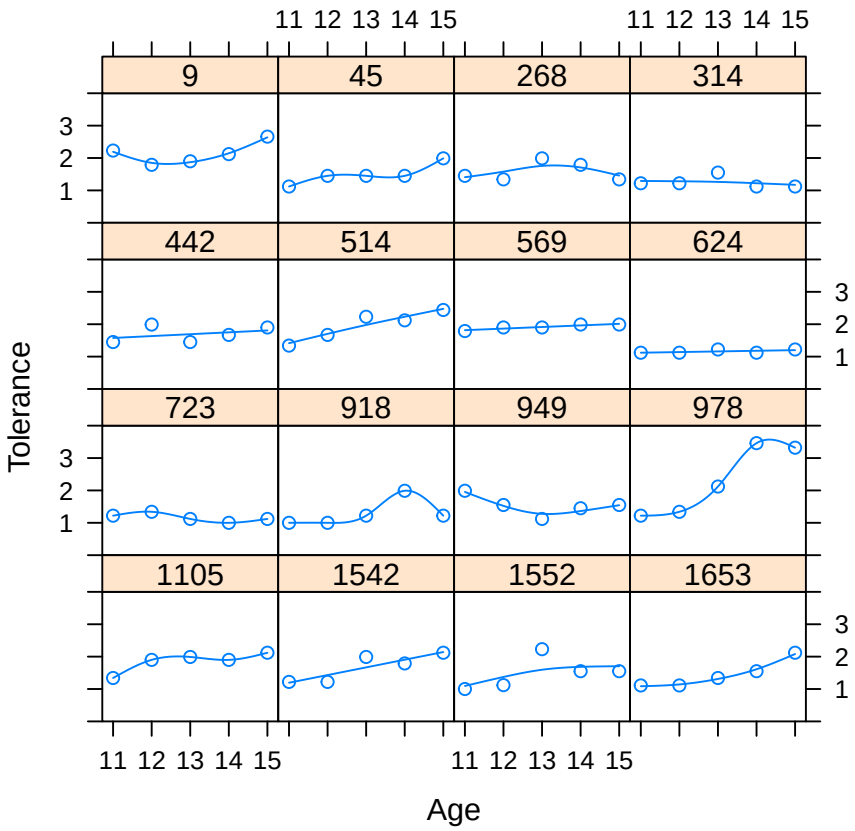
```
# spezzate per i dati
interaction.plot(tol_long$age, tol_long$id, tol_long$tolerance)
```



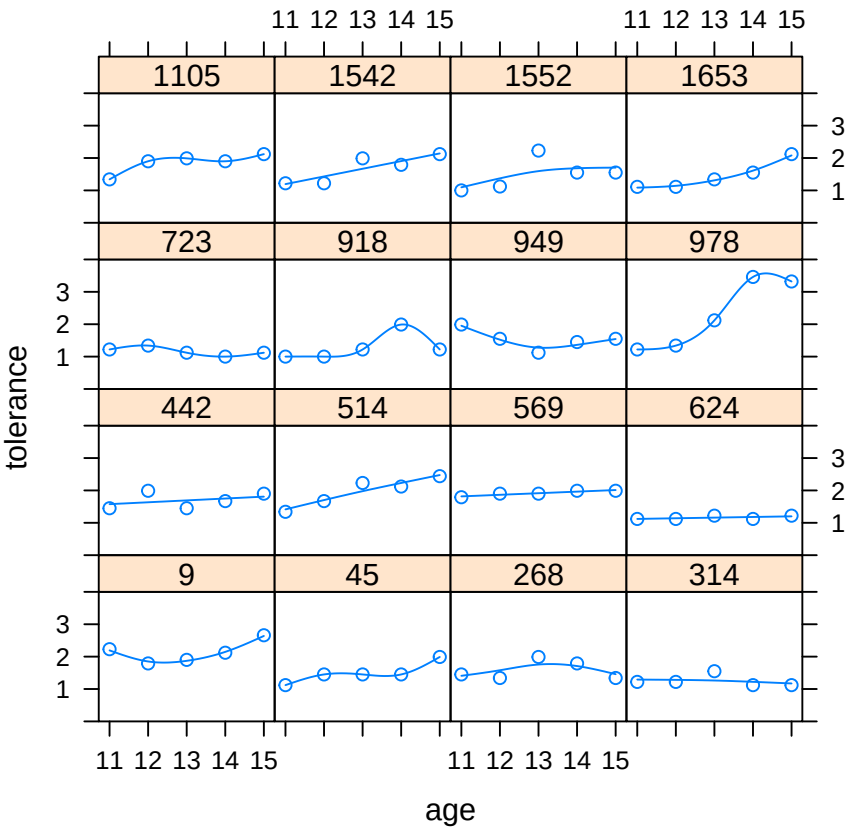

```
## andamento  per ciascun individuo per le 5 occasioni (età)
xyplot(tolerance ~ age | id, data=tol_long, as.table=T)
```



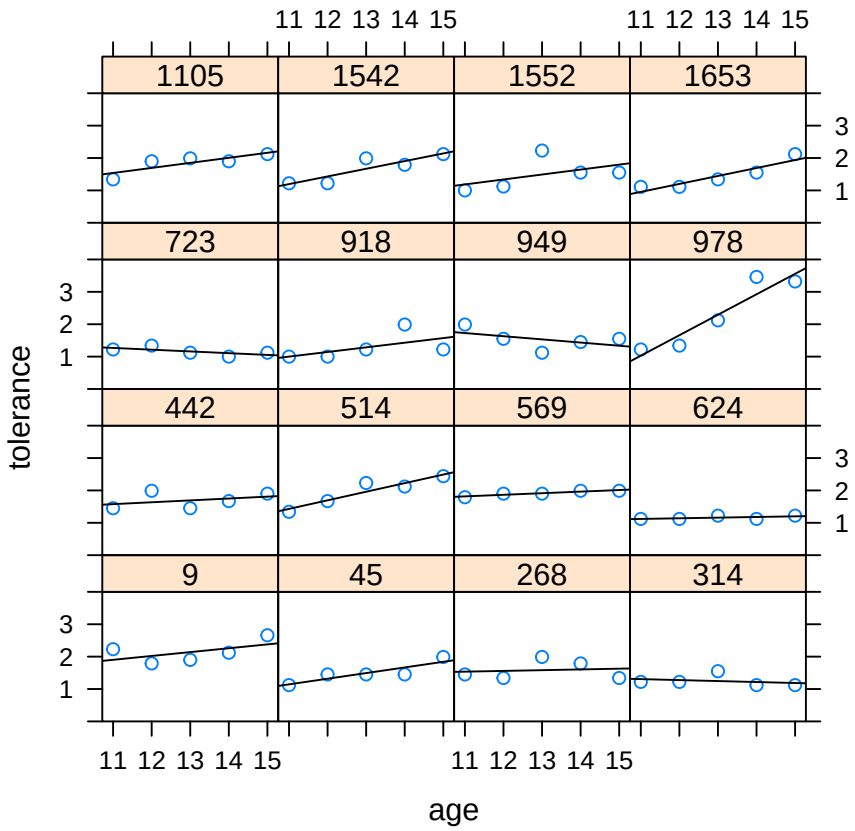

```
## interpolazione con stimatore non parametrico loess
xyplot(tolerance ~ age | id, ylim=c(0,4), data=tol_long,
       prepanel = function(x,y) prepanel.loess(x, y, family="gaussian"),
       xlab = "Age", ylab = "Tolerance", as.table=T,
       panel = function(x,y){
         panel.xyplot(x,y)
         panel.spline(x,y)} )
```




```
# interpolazione spline
xyplot(tolerance ~ age | id, ylim=c(0,4), data=tol_long,
       prepanel = function(x,y) prepanel.spline(x,y),
       xlab = "age", ylab = "tolerance",
       panel = function(x,y){
         panel.xyplot(x,y)
         panel.spline(x,y)} )
```




```
## non c'è una grossissima differenza tra loess e spline (grafici sono simili)

## interpolazione lineare
xyplot(tolerance ~ age | id, ylim=c(0,4), data=tol_long,
       prepanel = function(x,y) prepanel.lmline(x,y),
       xlab = "age", ylab = "tolerance",
       panel = function(x,y){
         panel.xyplot(x,y)
         panel.lmline(x,y)}
)
```




```
## stima dei modelli di regressione lineari per ciascun individuo
lm.summary <- by(tol_long, tol_long$id, function(x) summary(lm(tolerance ~ time, data
## lm.summary[1] # modello di regressione per primo soggetto
## lm.summary[[1]]$coefficients # i suoi coefficienti
## lm.summary[[1]]$coefficients[[1]] # intercetta

## salviamo per ogni modello: intercetta, slope ed R^2
## ---------------------------------------------------
## intercette e analisi
summary(all.intercept <- sapply(lm.summary, function(x) x$coefficients[1, 1]))

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.954   1.138   1.287   1.358   1.548   1.902

# intercetta sempre positiva, come ragionevole (al tempo 0 ossia a 11 anni)
stem(all.intercept, scale = 2)

##
##   The decimal point is 1 digit(s) to the left of the |
##
```

```
##    9 | 5
##   10 | 03
##   11 | 2489
##   12 | 7
##   13 | 1
##   14 | 3
##   15 | 448
##   16 |
##   17 | 3
##   18 | 2
##   19 | 0
```

```
## per interpretare lo stem, sort(all.intercept)
## nei grafici stem ci sono in ogni riga un numero su sx, una barra e un numero
## su dx
## come visto minimo 0.954 e max 1.90. lo stem prende le prime due cifre, dopo
## la barra viene riportato la terza cifra.
## nella seconda abbiamo .100 e .103
## nella terza abbiamo 1.12, 1.14, 1.18
##
## interpretazione guardando alla forma: c'e maggior concentrazione in 1.1 e 1.5

## slopes
summary(all.slopes <- sapply(lm.summary, function(x) x$coefficients[2,1]))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.09800  0.02225  0.13100  0.13081  0.18975  0.63200
```

```
## stem(all.slopes, scale=2) # qui qualche estremo, quello dello sbandato
## sort(all.slopes)

## R^2
summary(all.r2 <- sapply(lm.summary, function(x) x$r.squared))
```

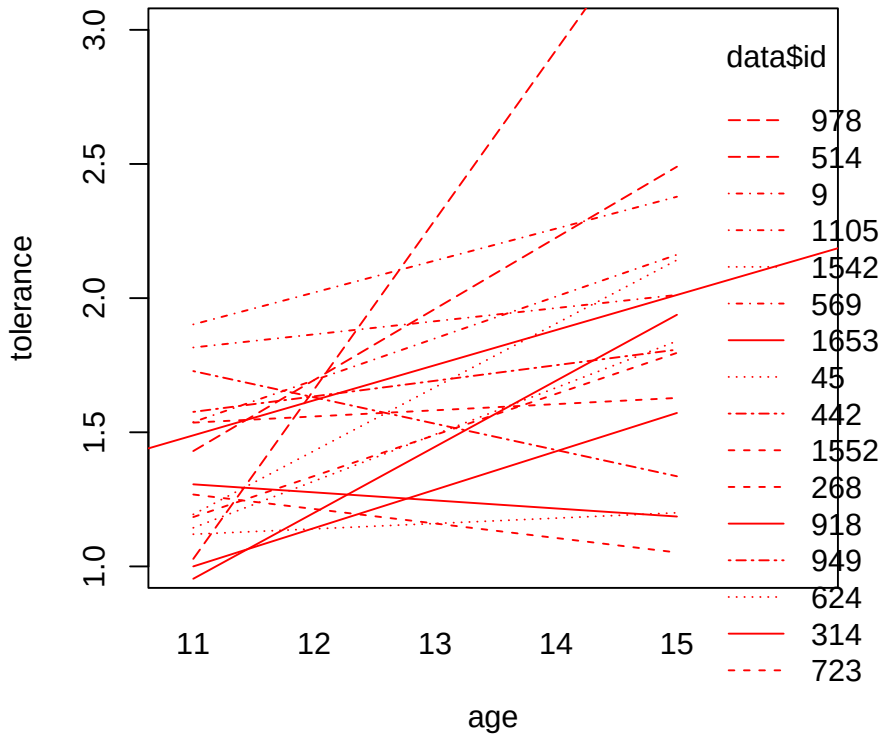
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01537 0.25048 0.39167 0.49101 0.79711 0.88641
```

```
## stem(all.r2, scale=2) # sort(all.r2)
# concentrazione di valori addoso a 0.8

## correlazione tra intercette e slopes
cor(all.intercept, all.slopes) #correlazione intercette pendenze: non altissima
```

```
## [1] -0.4481135
```

```
## grafico con tutte le regressioni (predizioni) e una media
## ------------------------------------------------------------
single_lm <- function(data, ylim = c(1,3)){
  ## browser()
  data$id <- droplevels(data$id)
  ## plot delle regressioni singole (via predizione)
```

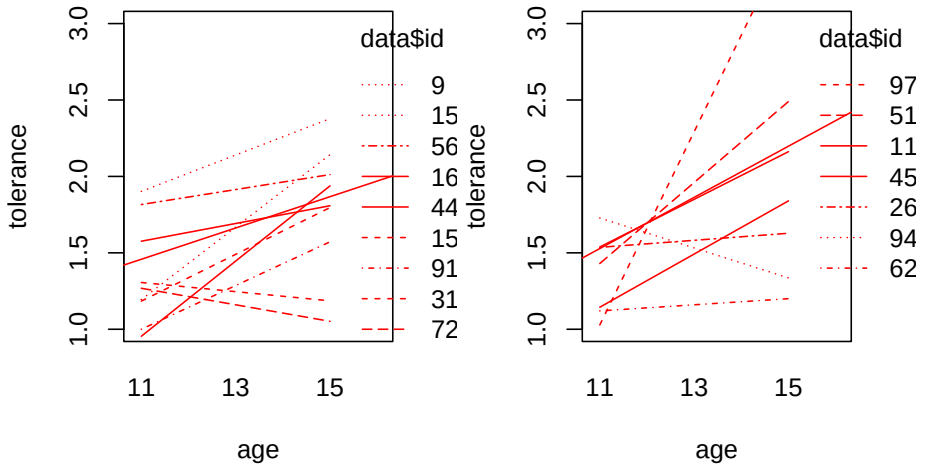
```
  fit <- by(data, data$id, function(x) fitted.values(lm(tolerance ~ time, data=x)))
  fit <- unlist(fit)
  interaction.plot(data$age, data$id, fit, xlab="age", ylab="tolerance",
                   ylim = ylim)
  ## aggiunta stima media
  mods <- by(data, data$id, function(x) summary(lm(tolerance ~ time, data=x)))
  intercepts <- unlist(lapply(mods, function(x) x$coefficients[1, 1]))
  slopes <- unlist(lapply(mods, function(x) x$coefficients[2,1]))
  abline(a = mean(intercepts, na.rm=TRUE), b=mean(slopes, na.rm=TRUE),
         col = 'red')
  invisible(list("intercepts"=intercepts, "slopes"=slopes))
}

## overall
res <- single_lm(tol_long)
```



consideriamo ora la covariate sesso

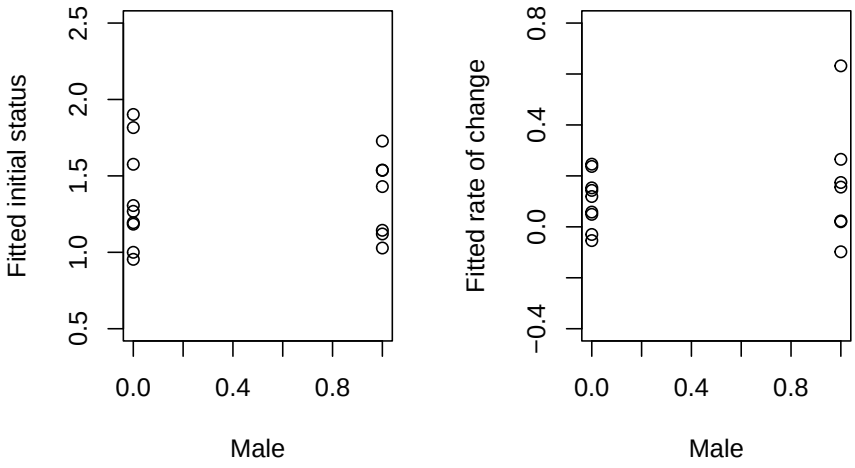
```
## splittato per genere
tol_male <- split(tol_long, tol_long$male)
par(mfrow=c(1,2))
```

```
lapply(tol_male, single_lm)
```




```
## $`0`
## $`0`$intercepts
##     9   314   442   569   723   918  1542  1552  1653
## 1.902 1.306 1.576 1.816 1.268 1.000 1.194 1.184 0.954
##
## $`0`$slopes
##      9    314    442    569    723    918   1542   1552   1653
##  0.119 -0.030  0.058  0.049 -0.054  0.143  0.237  0.153  0.246
##
##
## $`1`
## $`1`$intercepts
##    45   268   514   624   949   978  1105
## 1.144 1.536 1.430 1.120 1.728 1.028 1.538
##
## $`1`$slopes
##     45    268    514    624    949    978   1105
##  0.174  0.023  0.265  0.020 -0.098  0.632  0.156

## boh dicono che i maschi hanno coefficienti piu incliniati a me non sembrano cosi

# analisi intercette e pendenze
## variabilita delle intercette tra donne e uomini: molta piu variabilita nele
## donne rispetto agli uomini
par(mfrow=c(1,2))
plot(tol_wide$male, res$intercepts, xlab="Male", ylab="Fitted initial status",
     xlim=c(0, 1), ylim=c(0.5, 2.5))
```

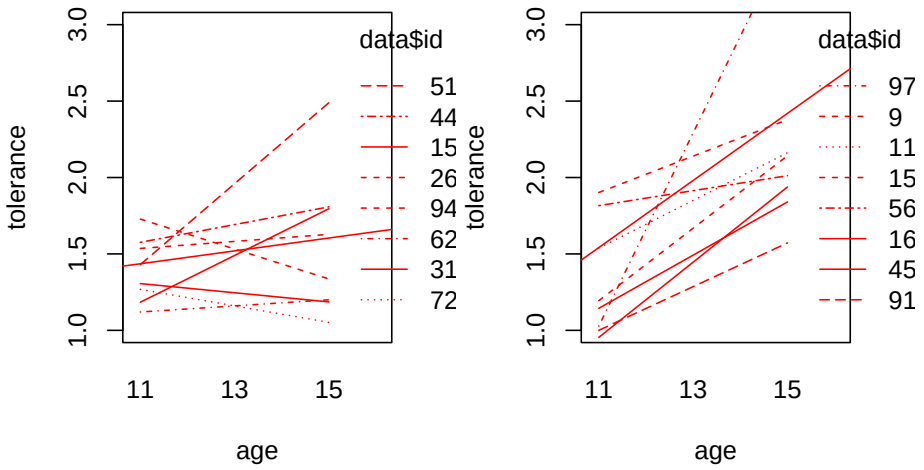
```
plot(tol_wide$male, res$slopes, xlab="Male", ylab="Fitted rate of change",
     xlim=c(0, 1), ylim=c(-0.4, .8)) # no grossa diff salvo un diverso punto dello sb
```



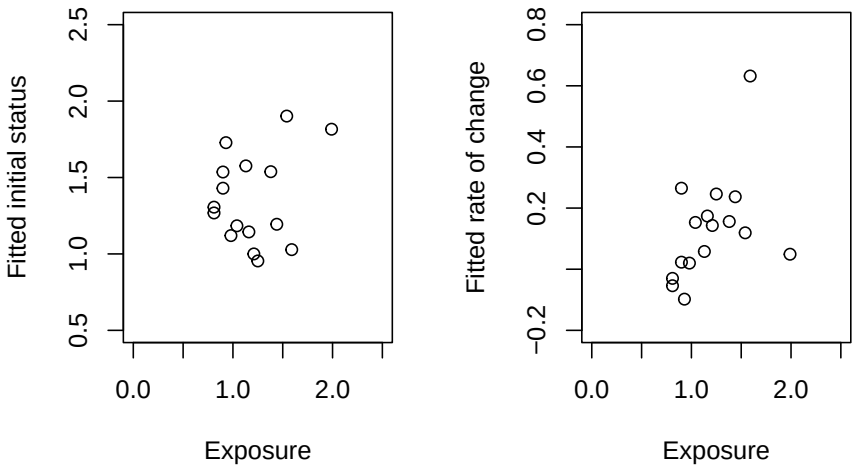

```
# giusto per avere una idea correlazione tra
# dicotomica e continua
cor(tol_wide$male, res$intercepts) # molto bassa

## [1] 0.008630119

cor(tol_wide$male, res$slopes) # cor strabassa

## [1] 0.1935703
```

Sia da analisi grafica che da correlazioni, il sesso non dovrebbe avere effetto sulla intercetta/pendenza.
Per quanto riguarda exposure procediamo a dicotomizzare in base alla mediana ed effettuare la stessa analisi

```
## splittato per exposure dicotomizzata
tol_long$high_exposure <- tol_long$exposure > median(tol_long$exposure)
tol_exp <- split(tol_long, tol_long$high_exposure)
par(mfrow=c(1,2))
lapply(tol_exp, single_lm)
```




```
## $`FALSE`
## $`FALSE`$intercepts
##   268   314   442   514   624   723   949  1552
## 1.536 1.306 1.576 1.430 1.120 1.268 1.728 1.184
##
## $`FALSE`$slopes
##    268    314    442    514    624    723    949   1552
##  0.023 -0.030  0.058  0.265  0.020 -0.054 -0.098  0.153
##
##
## $`TRUE`
## $`TRUE`$intercepts
##     9    45   569   918   978  1105  1542  1653
## 1.902 1.144 1.816 1.000 1.028 1.538 1.194 0.954
##
## $`TRUE`$slopes
##     9    45   569   918   978  1105  1542  1653
## 0.119 0.174 0.049 0.143 0.632 0.156 0.237 0.246

## nella low exposure non abbiamo il soggetto anomalo, ce sono solo 2 con
## inclunazione negativa
## qui si diomadonna che le pendenze sono piu alte nell'high exposure

## stessa cosa per intercetta e ed exposure
plot(tol_wide$exposure, res$intercepts, xlab="Exposure", ylab="Fitted initial status",
     xlim=c(0, 2.5), ylim=c(0.5, 2.5))
plot(tol_wide$exposure, all.slopes, xlab = "Exposure", ylab =
       "Fitted rate of change", xlim = c(0, 2.5), ylim = c(-0.2, 0.8))
```




```
cor(tol_wide$exposure, res$intercepts) # una quasi palla con correlazione non alta

## [1] 0.2324426

cor(tol_wide$exposure, all.slopes)

## [1] 0.4420925

## qui vi p una crescita al variare di exposure, correlazione non altissima ma
## schifo non fa
```

Per exposure invece c'è variabilita tra intercette e slope: questo suggerisce di utilizzare un mdello ad effetti misti: poi vedere se solo a intercetta casuale o altro, si vedrà.

3.2 Modelli vari

Remark 11 (Notazione). Considerando y_{it} con

- Y variabile casuale risposta dipendente
- pedici $i = 1, \dots n$ individui
- $t = 1, \dots T_i$ (se $T_i = T \forall i$ il disegno è bilanciato)
- l'osservazione su un soggetto $\mathbf{y}_i = [y_{i1}, \dots, y_{iT}]$ 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 $\mathbf{y}_i \perp\!\!\!\perp \mathbf{y}_j, \forall i \neq j$

Important remark 11. In questo setting per stime unconditional/non condizionate si intende stime senza effetti fissi ad esclusione del tempo

3.2.1 Non condizionato (intercetta casuale)

Ignorando anche l'effetto del tempo abbiamo un primo modello, da considerarsi nullo, dove a ciascun individuo è associata la sua media nel tempo

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

dove

$$\pi_{0i} = \gamma_0 + u_{0i}$$

quindi complessivamente

$$y_{it} = \underbrace{\gamma_0}_{\text{parte fissa}} + \underbrace{u_{0i} + \varepsilon_{it}}_{\text{parte casuale}}$$

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

$$\begin{aligned} &\varepsilon_{it} \sim \mathrm{N}\left(0, \sigma_\varepsilon^2\right) \\ &u_{0i} \sim \mathrm{N}\left(0, \sigma_u^2\right), \qquad\qquad \varepsilon_{it} \perp\!\!\!\perp u_{0i} \forall i, j \end{aligned}$$

Se il modello è unconditional con solo intercetta casuale studiamo valore atteso e varianza

$$\begin{aligned} &\mathbb{E}\left[y_{it}\right] = \gamma_0 \\ &\mathrm{Var}\left[y_{it}\right] = \sigma_{u_0}^2 + \sigma_\varepsilon^2 \end{aligned}$$

Example 3.2.1.
```
# Modello 0 (anche detto modello nullo). Modello non condizionato  -> solo inter
summary(m0 <- lmer(tolerance ~ (1 | id), data=tol_long, REML = FALSE))

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: tolerance ~ (1 | id)
##    Data: tol_long
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##     109.0     116.2     -51.5     103.0        77
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.1065 -0.5355 -0.1989  0.4641  3.3595
##
## 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
```

```
## qui abbiamo usato ML (stima distorta delle componenti casuali nostri
## modelli, distorsione che all'aumento della campione diminuisce
## se si è molto interessati alla componente casuale usare REML
## df resiudi sono i pz meno param stimati (3 qui, 2 casuali variabilita e l'intercet
(ICC=0.07465/(0.07465+0.16794)) # coefficiente di correlazione intraclasse

## [1] 0.3077208

coef(m0) # intercette per ogni bambino

## $id
##      (Intercept)
## 9       1.978436
## 45      1.531528
## 268     1.593598
## 314     1.361868
## 442     1.669463
## 514     1.854295
## 569     1.822570
## 624     1.302556
## 723     1.302556
## 918     1.389455
## 949     1.559115
## 978     2.083267
## 1105    1.778431
## 1542    1.652910
## 1552    1.530149
## 1653    1.499803
##
## attr(,"class")
## [1] "coef.mer"
```

Come stima di $\hat{\sigma}^2_\varepsilon = 0.16$ mentre per $\hat{\sigma}^2_u = 0.07465$. Il t-value del coefficiente è largamente significativo è l'ICC impone il considerare un modello gerarchico: tra diversi soggetti c'è una variabilità importante

3.2.2 Non condizionato (intercetta casuale, tempo lineare fisso)

In questo modello ipotizziamo una relazione lineare rispetto al tempo (tante rette parallele con stessa pendenza e diversa intercetta). Nell'esempio della tolerance abbiamo 5 osservazioni per individuo, può non essere troppo robusto andare su modelli più complessi. Il modello è:

$$y_{it} = \pi_{0i} + \pi_1 t + \varepsilon_{it}, \quad i = 1, \ldots, n, \quad t = 1, \ldots, T_i$$

dove

$$\pi_{0i} = \gamma_0 + u_{0i}$$

e complessivamente

$$y_{it} = \underbrace{\gamma_0 + \pi_1 t}_{\text{parte fissa}} + \underbrace{u_{0i} + \varepsilon_{it}}_{\text{parte casuale}}$$

In questo caso

$$\begin{aligned}\mathbb{E}[y_{it}] &= \gamma_0 + \pi_1 t \\ \text{Var}[y_{it}] &= \sigma^2_{u_0} + \sigma^2_{\varepsilon}\end{aligned}$$

Se qui studiamo la covarianza tra due misurazioni sullo stesso soggetto in tempi diversi

$$\begin{aligned}\text{Cov}(y_{it}, y_{it'}) &= \mathbb{E}[(y_{it} - \mathbb{E}[y_{it}])(y_{it'} - \mathbb{E}[y_{it'}])] \\ &\overset{(1)}{=} \mathbb{E}[(u_{0i} + \varepsilon_{it})(u_{0i} + \varepsilon_{it'})] \\ &= \mathbb{E}[u_{0i}u_{0i} + u_{0i}\varepsilon_{it} + u_{0i}\varepsilon_{it'} + \varepsilon_{it}\varepsilon_{it'}] \\ &\overset{(2)}{=} \mathbb{E}[u_{0i}u_{0i}] = \sigma^2_{u_0}\end{aligned}$$

dove in (1) rimangono solo le componenti casuali e in (2) $\mathbb{E}[u_{0i}\varepsilon_{it}] = \mathbb{E}[u_{0i}\varepsilon_{it'}] = \varepsilon_{it}\varepsilon_{it'} = 0$ è la covarianza di coppie di variabili indipendenti.
Quindi notiamo che sotto questo modello la covarianza non dipendende dal tempo t, è costante nel tempo

TODO: CHECK compound symmetric?

Example 3.2.2.

```
#Modello 1. Modello non condizionato (1 variabile esplicativa di primo livello)
# random intercept e aggiungo la covariata time con coefficiente fisso
summary(m1 <- lmer(tolerance ~ time + (1 | id), data=tol_long,REML = FALSE))

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: tolerance ~ time + (1 | id)
##    Data: tol_long
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##      92.2     101.7     -42.1      84.2        76
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.8814 -0.6410 -0.1174  0.5523  3.3714
##
## 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
```

```
##
## Correlation of Fixed Effects:
##      (Intr)
## time -0.562

(ICC=0.0832/(0.0832+0.1252)) # coefficiente di correlazione intraclasse aumenta

## [1] 0.3992322

# (maggiore variabilita tra individui una cvolta considerato tempo)
```

Alcune considerazioni

- nella tabella degli effetti casuali non cambia nulla (a livello di struttura) rispetto al precedente modello, proprio perché non cambiano gli effetti casuali considerati (intercette);
- 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 $\hat{\sigma}^2_\varepsilon$ di questo modello deve essere minore di quella del precedente. Ed effettivamente introducendo il tempo la varianza residua si è ridotta da 0.16 a 0.12;
- il coefficiente del tempo è lievemente positivo e statisticamente significativo
- una volta considerato il tempo emerge un aumento della variabilità tra individui (aumento dell'ICC)

Per confrontare questi primi due modelli possiamo usare il LRT essendo stimati entrambi con ML (se usassimo REML sarebbe piu indicativo/meno corretto) ed essendo modelli nested (il primo si ottiene da questo secondo imponendo $\pi_1 = 0$). Il test è

$$-2\log\left(\frac{\text{verosim. modello piu piccolo (M0)}}{\text{verosim. modello più grande (M1)}}\right)$$

se

- le due verosimiglianze sono uguali (variabile aggiuntiva non serve) il rapporto tra le verosimiglianze è 1 e il test (via log) diviene 0
- la variabile è utile la verosimiglianza al denominatore è più alta, argomento di log è <1, il log è negativo, e questo è motivo per cui si pone davanti il -2 (per avere un test positivo che abbia distribuzione nota)

Altro test che abbiamo altro test cha abbiamo è il test di Wald (il test è normale essendo ML, almeno asintoticamente), sotto ipotesi nulla con media 0 e un certo errore standard

```
anova(m0, m1)
```

```
## Data: tol_long
## Models:
## m0: tolerance ~ (1 | id)
## m1: tolerance ~ time + (1 | id)
##    npar     AIC    BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
## m0    3 109.022 116.17 -51.511   103.022
## m1    4  92.205 101.73 -42.103    84.205 18.816  1  1.439e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## deviance è -2log(verosimiglianza)
## possiamo risscrivere il chisq test come
## -2 ln(logM0 / logM1) = -2 (logLM0 - logLM1) =
## -2 log(LM0) - (-2 log(LM1))
## = 103 - 84.2 = 18.8
```

3.2.3 Non condizionato (intercetta casuale, tempo lineare fisso, slope casuale)

Al modello precedente aggiungiamo la possibilità che la traiettoria nel tempo (slope) sia diversa da individuo a individuo (oltre al livello di partenza) sia differente. Il modello è

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

dove

$$\begin{aligned} \pi_{0i} &= \gamma_0 + u_{0i} \\ \pi_{1i} &= \gamma_1 + u_{1i} \end{aligned}$$

e overall

$$y_{it} = \underbrace{\gamma_0 + \gamma_1 t}_{\text{parte fissa}} + \underbrace{u_{1i} t + u_{0i} + \varepsilon_{it}}_{\text{parte casuale}}$$

con assunzioni

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

In questo caso

$$\mathbb{E}\left[y_{it}\right] = \gamma_0 + \gamma_{10} t$$

$$\mathrm{Var}\left[y_{it}\right] = \underbrace{\sigma^2_{u_0} + \sigma^2_{u_1} t^2 + \sigma^2_\varepsilon}_{\text{var componenti casuali}} + \underbrace{2\sigma_{u_{01}} t + 0 + 0}_{\text{covs delle componenti casuali}}$$

e per la covarianza si ha

TODO: CHECK sta
varianza ...

$$\begin{aligned}\text{Cov}\,(y_{it}, y_{it'}) &= \mathbb{E}\,[(y_{it} - \mathbb{E}\,[y_{it}])(y_{it'} - \mathbb{E}\,[y_{it'}])] \\ &= \mathbb{E}\,[(u_{0i} + u_{01}t + \varepsilon_{it})(u_{01} + u_{1i}t' + \varepsilon_{it'})] \\ &= \mathbb{E}\,[u_{0i}u_{0i} + u_{1i}tu_{0i} + \ldots] \\ &= \ldots \\ &= \sigma^2_{u_0} + t\sigma_{u_{01}} + t'\sigma_{u_{01}} + t \cdot t'\sigma^2_{u_1} \\ &= \sigma^2_{u_0} + (t + t')\sigma_{u_{01}} + t \cdot t'\sigma^2_{u_1}\end{aligned}$$

con due componenti casuali la covarianza non è piu costante ma dipende dai tempi con cui la pongo a confronto. covarianza e correlazione cambia in funzione dei tempi

Example 3.2.3. Per questo modello la stima porta a

```
#random intercept e aggiungo la covariata time con coefficiente casuale
summary(m2 <- lmer(tolerance ~ time + (time | id), data=tol_long, REML = FALSE))

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: tolerance ~ time + (time | id)
##    Data: tol_long
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##      76.0      90.3     -32.0      64.0        74
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5885 -0.5127 -0.2328  0.2798  2.5899
##
## 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  18.837
## time         0.13081    0.04171   3.137
##
## Correlation of Fixed Effects:
##      (Intr)
## time -0.448

## coppie di coefficienti (che oscillano attorno a 1.35 e 0.1308)
coef(m2)

## $id
##      (Intercept)         time
```

```
## 9       1.630003  0.195303579
## 45      1.256532  0.137848120
## 268     1.431527  0.064666103
## 314     1.307887 -0.011276216
## 442     1.456847  0.099519805
## 514     1.413992  0.253350335
## 569     1.576342  0.124775870
## 624     1.221648  0.005272896
## 723     1.285202 -0.036621197
## 918     1.179446  0.092134080
## 949     1.510274 -0.010672276
## 978     1.265943  0.506512610
## 1105    1.452226  0.176602541
## 1542    1.291032  0.197576997
## 1552    1.273561  0.125685820
## 1653    1.171537  0.172320933
##
## attr(,"class")
## [1] "coef.mer"
```

```
## faccamo una LRT: in uno slope fissa e nell'altro slope casuale (possiamo
## farlo perché i due modelli sono nested)
anova(m1, m2)
```

```
## Data: tol_long
## Models:
## m1: tolerance ~ time + (1 | id)
## m2: tolerance ~ time + (time | id)
##    npar    AIC     BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
## m1    4 92.205 101.734 -42.103    84.205
## m2    6 76.031  90.323 -32.015    64.031 20.175  2   4.16e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## qui due parametri in piu vafrianza della slope e covarianza di intercetta e slope
## interpre
```

Nella stima

- $\hat{\sigma}^2_{u_0} = 0.038$, $\hat{\sigma}^2_{u_1} = 0.0204$, $\hat{\sigma}^2_{\varepsilon} = 0.074$ e
- invece della covarianza $\hat{\sigma}_{u_{01}}$ R riporta la correlazione $\hat{\rho}_{u_{01}} = -0.24$ (legata alla prima): chi ha intercetta piu alta tende nel tempo ad avere una crescita piu bassa. 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.
- l'effetto del tempo è ancora statisticamente significativo
- Dal momento in cui usiamo la slope casuale non possiamo calcolare più il coefficiente di correlazione intracluster

3.2.4 Modello condizionato (covariate a livello di individuo), intercetta e slope casuali

Con il modello seguente introduciamo una covariata W_i che ha effetto sia sulla intercetta che la pendenza nel tempo

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

con

$$\begin{aligned} \pi_{0i} &= \gamma_0 + \gamma_{01}W_i + u_{0i} \\ \pi_{1i} &= \gamma_1 + \gamma_{11}W_i + u_{1i} \end{aligned}$$

e mettendo tutto assieme

$$\begin{aligned} y_{it} &= \gamma_0 + u_{0i} + (\gamma_1 + \gamma_{11}W_i + u_{1i})t + \varepsilon_{it} \\ &= \underbrace{\gamma_0 + \gamma_{01}W_i + \gamma_1 t + \gamma_{11}W_i t}_{\text{Parte fissa}} + \underbrace{u_{1it} + u_{0i} + \varepsilon_{it}}_{Partecasuale} \end{aligned}$$

se ad esempio W_i è `male`, γ_0 è il valore intercetta delle femmine mentre γ_{01} è la differenza tra maschi e femmine. Allo stesso modo γ_1 è il coefficiente angolare delle femmine mentre γ_{11} la differenza di coefficiente angolare tra maschi e femmine.
Valgono le seguenti assunzioni:

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

Remark 12. sopra i modelli sono unconditional: la covariata è solo il tempo. Si possono aggiungere covariate (genere ed exposure), queste solitamente indicate con lettere z o w; l'effetto puo essere su intercetta e o pendenza.
In questo caso la covarianza tra due osservazioni nel tempo $\mathrm{Cov}\left(y_{it}, y_{it'}\right)$ nello stesso soggetto dipende dal tempo. Qui vi è eteroschedasticità (not shown)

Example 3.2.4. la stima con la dummy male porta ai segguenti risultati

```
##Modello 3. Modello condizionato (1 variabile esplicativa di primo livello ed
## 1 di secondo livello) -> intercetta casuale e slope casuale random intercept
## e aggiungo la covariata time con coefficiente casuale
summary(m3 <- lmer(tolerance ~ time + (time | id) + male , data=tol_long, REML = FALS

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: tolerance ~ time + (time | id) + male
##    Data: tol_long
```

```
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##      77.8      94.5     -31.9      63.8        73
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.6093 -0.5156 -0.2041  0.2995  2.6098
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept) 0.03934  0.1983
##           time        0.02042  0.1429   -0.27
##  Residual             0.07412  0.2722
## Number of obs: 80, groups:  id, 16
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  1.33255    0.09181  14.514
## time         0.13081    0.04171   3.137
## male         0.05759    0.12913   0.446
##
## Correlation of Fixed Effects:
##      (Intr) time
## time -0.367
## male -0.615  0.000

anova(m2, m3) # anche qui

## Data: tol_long
## Models:
## m2: tolerance ~ time + (time | id)
## m3: tolerance ~ time + (time | id) + male
##    npar    AIC    BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
## m2    6 76.031 90.323 -32.015    64.031
## m3    7 77.840 94.515 -31.920    63.840 0.1903  1     0.6627

## aggiungiamo l'interazione (ottenendo di fatto il modello di sopra direi)

#Modello 4. Modello condizionato (1 variabile esplicativa di primo livello ed 1 di secondo live
#random intercept e aggiungo la covariata time con coefficiente casuale ed interazione tra time

summary(m4 <- lmer(tolerance ~ time + (time | id) + male + male:time, data=tol_long, REML = FAL

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: tolerance ~ time + (time | id) + male + male:time
##    Data: tol_long
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##      79.2      98.3     -31.6      63.2        72
##
```

```
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5911 -0.4875 -0.2427  0.2403  2.6098
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept) 0.03865  0.1966
##           time        0.01938  0.1392   -0.25
##  Residual             0.07412  0.2722
## Number of obs: 80, groups:  id, 16
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 1.355556   0.096104  14.105
## time        0.102333   0.054557   1.876
## male        0.005016   0.145296   0.035
## time:male   0.065095   0.082482   0.789
##
## Correlation of Fixed Effects:
##           (Intr) time   male
## time      -0.458
## male      -0.661  0.303
## time:male  0.303 -0.661 -0.458
```

```
anova(m2, m4) # m2 era il migliore sino a questo punto, proprio non vi è utilita in m
```

```
## Data: tol_long
## Models:
## m2: tolerance ~ time + (time | id)
## m4: tolerance ~ time + (time | id) + male + male:time
##    npar    AIC    BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
## m2    6 76.031 90.323 -32.015    64.031
## m4    8 79.229 98.286 -31.615    63.229 0.8013  2     0.6699
```

Guardando all'ultimo modello su sex

- i valori di AIC e BIC sono aumentati
- 1.355 è il valore di intercetta per le donne
- 0.005 è il differenziale nel valore intercetta tra ragazzi e ragazze
- 1.876 è la slope annuale per ragazze
- 0.065095 è il differenziale nel tasso di crescita tra ragazzi e ragazze
- confrontando via anova, `male` non è così rilevante e il sesso si può ignorare

Effettuiamo stessa cosa per la exposure in continuo (senza e con interazione col tempo)

```r
#Modello 5. Modello condizionato (1 variabile esplicativa di primo livello ed 1 di secondo live
#random intercept e aggiungo la covariata time con coefficiente casuale
# exposure continua

summary(m5<-lmer(tolerance ~ time + (time | id)+ exposure , data=tol_long, REML = FALSE))
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: tolerance ~ time + (time | id) + exposure
##    Data: tol_long
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##      72.6      89.3     -29.3      58.6        73
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5521 -0.5336 -0.1396  0.2965  2.6328
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept) 0.04179  0.2044
##           time        0.02042  0.1429   -0.62
##  Residual             0.07412  0.2722
## Number of obs: 80, groups:  id, 16
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  0.78025    0.21631   3.607
## time         0.13081    0.04171   3.137
## exposure     0.48478    0.17080   2.838
##
## Correlation of Fixed Effects:
##          (Intr) time
## time     -0.228
## exposure -0.941  0.000
```

```r
## AIC e BIC ridotti
anova(m2, m5) # vi è utilita nella variabile introdotta
```

```
## Data: tol_long
## Models:
## m2: tolerance ~ time + (time | id)
## m5: tolerance ~ time + (time | id) + exposure
##    npar    AIC    BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
## m2    6 76.031 90.323 -32.015    64.031
## m5    7 72.589 89.263 -29.295    58.589 5.4415  1    0.01966 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Modello 6. Modello condizionato (1 variabile esplicativa di primo livello ed 1 di secondo live
#random intercept e aggiungo la covariata time con coefficiente casuale ed interazione tra time
```

```
# exposure continua
summary(m6<-lmer(tolerance ~ time + (time | id)+ exposure + exposure:time, data=tol_l

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: tolerance ~ time + (time | id) + exposure + exposure:time
##    Data: tol_long
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##      71.1      90.2     -27.6      55.1        72
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.5589 -0.4896 -0.2054  0.3811  2.6328
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept) 0.03417  0.1848
##           time        0.01498  0.1224   -0.52
##  Residual             0.07412  0.2722
## Number of obs: 80, groups:  id, 16
##
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept)     1.1069     0.2716   4.075
## time           -0.1452     0.1449  -1.002
## exposure        0.2106     0.2203   0.956
## time:exposure   0.2317     0.1175   1.971
##
## Correlation of Fixed Effects:
##             (Intr) time   exposr
## time        -0.631
## exposure    -0.966  0.610
## time:exposr  0.610 -0.966 -0.631

anova(m5, m6) # siamo borderini, basati su interazione

## Data: tol_long
## Models:
## m5: tolerance ~ time + (time | id) + exposure
## m6: tolerance ~ time + (time | id) + exposure + exposure:time
##    npar    AIC    BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
## m5    7 72.589 89.263 -29.295    58.589
## m6    8 71.110 90.166 -27.555    55.110 3.4795  1    0.06213 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## quando numerosita è grande i due test danno esito simile: con numerosita
#piccola possono dare risultati diversi (come). IL lrt è meglio per piccoli campioni
```

Exposure dicotomizzata in base alla mediana (e direttamente interazione)

```
#Modello 7. Modello condizionato (1 variabile esplicativa di primo livello ed 1 di secondo live
#random intercept e aggiungo la covariata time con coefficiente casuale ed interazione tra time
# exposure dicotomica

summary(m7<-lmer(tolerance ~ time + (time | id) + high_exposure +
                   high_exposure:time,
                 data = tol_long, REML = FALSE))

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: tolerance ~ time + (time | id) + high_exposure + high_exposure:time
##    Data: tol_long
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##      74.3      93.4     -29.2      58.3        72
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.6271 -0.5120 -0.2140  0.3737  2.6281
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept) 0.03738  0.1933
##           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                   0.04212    0.04996   0.843
## high_exposureTRUE     -0.07150    0.14305  -0.500
## time:high_exposureTRUE  0.17738    0.07065   2.511
##
## Correlation of Fixed Effects:
##             (Intr) time   h_TRUE
## time        -0.455
## hgh_xpsTRUE -0.707  0.322
## tm:hgh_TRUE  0.322 -0.707 -0.455

##
```

Terzili di exposure

```
#Modello 8. Modello condizionato (1 variabile esplicativa di primo livello ed 1 di secondo live
#random intercept e aggiungo la covariata time con coefficiente casuale ed interazione tra time
# exposure su tre livelli


terzili <- quantile(tol_wide$exposure, probs = c(0, 1/3, 2/3, 1), na.rm = TRUE)
tol_long$exp_tertiles <- cut(tol_long$exposure,
```

```
                            breaks = terzili,
                            include.lowest = TRUE,
                            labels = c("Primo", "Secondo", "Terzo"))

summary(m8 <- lmer(tolerance ~ time + (time | id)+ exp_tertiles +
                     exp_tertiles:time,
                   data=tol_long, REML = FALSE))

## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: tolerance ~ time + (time | id) + exp_tertiles + exp_tertiles:time
##    Data: tol_long
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##      69.5      93.3     -24.8      49.5        70
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.7260 -0.5562 -0.1797  0.4623  2.7328
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept) 0.02129  0.1459
##           time        0.01209  0.1099   -0.52
##  Residual             0.07412  0.2722
## Number of obs: 80, groups:  id, 16
##
## Fixed effects:
##                            Estimate Std. Error t value
## (Intercept)                 1.39800    0.10469  13.354
## time                        0.02100    0.05701   0.368
## exp_tertilesSecondo        -0.22640    0.15528  -1.458
## exp_tertilesTerzo           0.09760    0.15528   0.629
## time:exp_tertilesSecondo    0.13380    0.08455   1.582
## time:exp_tertilesTerzo      0.21760    0.08455   2.574
##
## Correlation of Fixed Effects:
##             (Intr) time   exp_tS exp_tT tm:x_S
## time        -0.646
## exp_trtlsSc -0.674  0.436
## exp_trtlsTr -0.674  0.436  0.455
## tm:xp_trtlS  0.436 -0.674 -0.646 -0.294
## tm:xp_trtlT  0.436 -0.674 -0.294 -0.646  0.455
```

Final comparison

```
do.call(rbind, lapply(list("m0" = m0, "m1" = m1, "m2" = m2, "m3" = m3, "m4" = m4,
                           "m5" = m5, "m6" = m6, "m7" = m7, "m8" = m8),
                      function(x) summary(x)$AIC))

##          AIC       BIC    logLik -2*log(L) df.resid
```

```
## m0 109.02185 116.16793 -51.51093 103.02185       77
## m1  92.20549 101.73359 -42.10274  84.20549       76
## m2  76.03074  90.32290 -32.01537  64.03074       74
## m3  77.84042  94.51461 -31.92021  63.84042       73
## m4  79.22939  98.28560 -31.61470  63.22939       72
## m5  72.58922  89.26341 -29.29461  58.58922       73
## m6  71.10975  90.16597 -27.55488  55.10975       72
## m7  74.34796  93.40417 -29.17398  58.34796       72
## m8  69.51755  93.33782 -24.75878  49.51755       70
```

Stando all'AIC dovremmo scegliere il modello con terzili di exposure mentre col BIC (piu conservativo il modello m5)