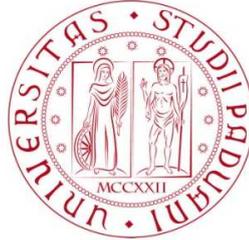


UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE
CORSO DI LAUREA TRIENNALE IN
STATISTICA PER L'ECONOMIA E L'IMPRESA



RELAZIONE FINALE

**L'uso del test F per valutare la significatività degli
effetti casuali in modelli lineari misti**

Relatore Prof. Alessandra Salvan
Dipartimento di Scienze Statistiche

Laureando Andrea Ongarato
Matricola 2003014

Anno Accademico 2022/2023

Indice

Introduzione	1
1 I modelli misti	3
1.1 Modelli per risposte correlate	3
1.1.1 Modelli marginali	4
1.1.2 Modelli con effetti individuali	4
1.1.3 Modelli con effetti casuali	4
1.2 Modelli con effetti casuali per risposte normali	5
2 Valutazione della significatività degli effetti casuali nei modelli misti	11
2.1 Richiami alla teoria asintotica della verosimiglianza	11
2.2 Limiti del test log-rapporto di verosimiglianza nei modelli misti	14
2.3 Soluzioni esistenti	16
2.3.1 Log-Rapporto di verosimiglianza ristretta	17
2.3.2 Tabella riassuntiva	19
3 Il test F basato su combinazioni lineari della risposta	21
3.1 Il test <i>FLC</i>	21
3.2 Implementazione	26
4 Studio di simulazione	29
4.1 Riproduzione dei risultati in Hui et al. (2019)	29
4.1.1 Studio di simulazione 1	30
4.1.2 Studio di simulazione 2	31
4.1.3 Studio di simulazione 3	33
4.2 Curve di potenza	35
5 Applicazione ai dati Stroke	41
5.1 Risultati	41

Conclusione	45
Appendice	47
A. Prodotto di Kronecker	47
B. Matrici di proiezione generate da matrici con colonne non linearmente indipendenti	47
C. Pseudo-inversa di Moore-Penrose	49
D. Distribuzione F non centrale	49
E. Funzione beta incompleta regolarizzata	50
F. Codice R	51
Bibliografia	65

Introduzione

In molte aree di ricerca, ci si può trovare in presenza di dati longitudinali, con più osservazioni ripetute per uno stesso soggetto nel corso di un arco temporale, oppure la variabile d'interesse potrebbe essere rilevata su gruppi di soggetti. In questi casi la variabile risposta è multivariata con componenti correlate, è quindi necessario un metodo appropriato per descriverne la struttura.

I modelli misti sono una classe di modelli usata per analizzare risposte multivariate. Per descrivere gli effetti che hanno determinate caratteristiche dei soggetti o dei gruppi di soggetti sulla risposta, vengono inclusi degli effetti casuali con l'obiettivo di spiegare una parte della variabilità che si può osservare nei dati. Spesso gli effetti casuali non sono di diretto interesse per l'analisi, ma vengono aggiunti per irrobustire le stime delle componenti di interesse. Ometterli porterebbe a distorsioni nel modello, ma includerli quando sono presenti poche osservazioni per stimarli potrebbe portare a situazioni ben peggiori: si tratta dell'usuale *trade-off* distorsione-varianza. Per esempio, può succedere che gli effetti casuali, inseriti per descrivere la struttura dei dati, siano talmente piccoli da rendere le stime, nel complesso, più accurate se vengono ignorati. Riuscire a distinguere in modo efficace se un effetto è significativo o meno diventa quindi fondamentale.

Introdurremo nel Capitolo 1 i modelli lineari misti per risposte normali e la notazione utilizzata. Il problema di valutare la significatività degli effetti casuali in questi modelli e le soluzioni esistenti verranno esposte nel Capitolo 2. I successivi capitoli sono principalmente basati sui contenuti dell'articolo di Hui et al. (2019). Nel Capitolo 3 viene descritto il test proposto dagli Autori, che verrà poi valutato con una serie di simulazioni nel Capitolo 4. Nel Capitolo 5 viene svolta una breve analisi sul dataset **Stroke** (Dobson & Barnett, 2018, paragrafo 11.2) con l'obiettivo di mostrare una semplice applicazione del test. I risultati presenti nel Capitolo 4 e nel Capitolo 5 sono prodotti utilizzando il *software* statistico R (R Core Team (2022)).

Capitolo 1

I modelli misti

Introduzione

In questo capitolo si introducono i modelli con effetti casuali, che rientrano nella classe dei modelli per risposte correlate, per poi concentrarsi sui modelli lineari misti per risposte normali, argomento principale di questa trattazione. I contenuti si basano su Salvani et al. (2020b, capitolo 7).

1.1 Modelli per risposte correlate

In molte situazioni la variabile risposta Y_i per ogni unità, $i = 1, \dots, n$, è multivariata. Alcuni esempi sono i dati *panel*, che sono risultato di studi osservazionali longitudinali, in cui si rileva una caratteristica di uno stesso soggetto in istanti temporali diversi (per esempio in ambito clinico, gli effetti di un trattamento su un paziente registrati ripetutamente a certi intervalli di tempo), oppure i casi in cui l'unità statistica è un gruppo (cosiddetto *cluster*) e si è interessati a una certa caratteristica per ogni membro del gruppo.

Siano quindi y_{ij} le osservazioni sulla risposta per l' i -esima unità statistica, $i = 1, \dots, n, j = 1, \dots, m_i$. Per ogni soggetto si hanno m_i osservazioni, quindi il numero totale di osservazioni è $N = \sum_{i=1}^n m_i$, oppure $N = nm$ se il numero di osservazioni è lo stesso per ogni unità.

Il vettore delle osservazioni sulla risposta dell' i -esima unità è indicato con:

$$y_i = (y_{i1}, \dots, y_{im_i})^\top$$

Il vettore riga p -dimensionale x_{ij} è quindi il vettore delle variabili esplicative per la variabile y_{ij} . Indichiamo con Y_{ij} la variabile casuale che ha generato y_{ij} e sia $\mu_{ij} = E(Y_{ij})$.

Le variabili casuali Y_{ij} e Y_{ih} , $j \neq k$, che descrivono osservazioni diverse appartenenti alla stessa unità saranno tipicamente correlate e i modelli statistici trattati in seguito tengono conto di questa possibile correlazione. Si assume invece indipendenza tra unità distinte.

Due approcci possibili per l'analisi di risposte correlate sono i **modelli marginali** e i **modelli con effetti individuali**.

1.1.1 Modelli marginali

I modelli marginali descrivono l'effetto delle variabili esplicative sul valore atteso μ_{ij} . La descrizione della correlazione tra osservazioni appartenenti ad una stessa unità non è di diretto interesse, ma è fondamentale tenerne conto per valutare correttamente gli *standard error* degli stimatori. Si assume $\mathbf{Y}_i \sim N_{m_i}(\mu_i, V_i)$ per $i = 1, \dots, n$, con $\mu_i = \mathbf{X}_i\beta$, dove \mathbf{X}_i è la matrice del modello $m_i \times p$ per l' i -esima unità e $\beta = (\beta_1, \dots, \beta_p)^\top$. Assumendo V_i nota, lo stimatore di massima verosimiglianza, ottenuto come soluzione delle equazioni normali, è lo stimatore dei minimi quadrati generalizzati. Quando la variabile risposta non è normale si adotta una specificazione semiparametrica basata sulle equazioni di stima generalizzate (**GEE**, *Generalized Estimating Equations*), si veda Salvan et al. (2020b, paragrafo 7.2.2) per approfondimenti.

1.1.2 Modelli con effetti individuali

L'assunzione principale di questo tipo di modelli è la presenza di caratteristiche non osservabili comuni a tutte le osservazioni di una stessa unità. Si parla di **modelli con effetti fissi** quando per ciascuna unità i , $i = 1, \dots, n$, si introduce un parametro, solitamente di intercetta. Questi parametri sono chiamati *parametri incidentali* e non sono di diretto interesse per l'analisi statistica. La dimensione dello spazio parametrico dipende dunque da n e pertanto, si perdono le condizioni di regolarità per la consistenza degli stimatori di massima verosimiglianza.

1.1.3 Modelli con effetti casuali

Si parla di modelli con effetti casuali quando, per descrivere le caratteristiche non osservabili comuni a tutte le osservazioni di una stessa unità, si utilizzano variabili casuali

dette **effetti casuali**. I modelli con effetti casuali sono anche chiamati **modelli misti**, proprio per la presenza congiunta di effetti fissi ed effetti casuali. Gli effetti casuali inducono correlazione tra le osservazioni appartenenti alla medesima unità e inoltre si estendono anche a dati con **struttura multilivello**, o in gergo *nested*. Nei dati con struttura multilivello esiste una scala gerarchica con k livelli, in cui osservazioni al livello 1 appartengono a cluster di livello 2, che a loro volta appartengono a cluster di livello più alto, e così dicendo fino al livello k . Inserendo quindi un effetto casuale ad ogni livello è possibile tenere conto della relazione tra osservazioni appartenenti allo stesso cluster.

I modelli misti possono essere usati per descrivere molteplici tipologie di risposte, si veda Bolker et al. (2009). In questa trattazione ci concentreremo sui **modelli lineari con effetti casuali per risposte normali**.

1.2 Modelli con effetti casuali per risposte normali

Immaginando, per iniziare, la struttura più semplice di correlazione, si pensi a una sperimentazione clinica per valutare l'efficacia di un farmaco per il diabete. Sia n il numero di soggetti. Per ogni soggetto si effettuano $m_i, i = 1, \dots, n$, misurazioni del livello di glicemia nel sangue dopo l'infusione del farmaco intervallate da un certo periodo di tempo (in questo caso per ogni soggetto si eseguono m misurazioni).

Per l'inferenza sul livello medio μ di glicemia per valutare l'effetto del farmaco, un modello adeguato può essere

$$Y_{ij} = \mu + a_i + \varepsilon_{ij}, \quad i = 1, \dots, n, j = 1, \dots, m,$$

dove a_i rappresenta l'effetto per l' i -esimo soggetto con una certa distribuzione nota e ε_{ij} è un errore casuale. Si parla in questo caso di modello lineare con effetti misti (LMM, *Linear Mixed Model*). L'assunzione che si fa in questi casi è quella di normalità sia per gli effetti casuali e sia per gli errori, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$, $a_i \sim N(0, \sigma_a^2)$, inoltre le ε_{ij} sono assunte indipendenti dalle a_i .

Ciò comporta che

$$E(Y_{ij}) = \mu_{ij} \text{ e } \text{Var}(Y_{ij}) = \sigma_a^2 + \sigma_\varepsilon^2.$$

Se si assume indipendenza tra le ε_{ij} , allora la covarianza entro le unità vale

$$\text{Cov}(Y_{ij}, Y_{ih}) = \text{Cov}(a_i + \varepsilon_{ij}, a_i + \varepsilon_{ih}) = E((a_i + \varepsilon_{ij})(a_i + \varepsilon_{ih})) = \sigma_a^2$$

e di conseguenza

$$\text{Cor}(Y_{ij}, Y_{ih}) = \frac{\text{Cov}(Y_{ij}, Y_{ih})}{\sqrt{\text{Var}(Y_{ij})}\sqrt{\text{Var}(Y_{ih})}} = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_\varepsilon^2} = \rho$$

per $j \neq h$.

La matrice di varianze e covarianze delle \mathbf{Y}_i diventa

$$\text{Var}(\mathbf{Y}_i) = V_i = (\sigma_a^2 + \sigma_\varepsilon^2) \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{pmatrix}$$

La matrice V_i ha struttura di **equicorrelazione**, con ρ , **coefficiente di correlazione intraclasse**, costante e positivo. Sempre riprendendo l'esempio iniziale, si può ipotizzare che i livelli di glicemia di uno stesso soggetto varino linearmente col tempo trascorso (passando da y_{ij} a $y_{ij'}$, con $j' = j + 1$) e che l'intensità, con la quale varii la glicemia, dipenda dal soggetto e accomuni tutte le osservazioni appartenenti a quella specifica unità. Il modello lineare avrebbe quindi la forma

$$Y_{ij} = \beta_1 + a_i + (\beta_2 + b_i)j + \varepsilon_{ij}. \quad (1.1)$$

I coefficienti β sono **effetti fissi**, mentre nel caso (1.1) a_i e b_i sono **effetti casuali**. Considerando per semplicità a_i , b_i e ε_{ij} realizzazioni di variabili casuali indipendenti con distribuzione $N(0, \sigma_a^2)$, $N(0, \sigma_b^2)$ e $N(0, \sigma_\varepsilon^2)$, rispettivamente, si ha

$$\begin{aligned} E(Y_{ij}) &= \beta_1 + \beta_2 j \\ \text{Var}(Y_{ij}) &= \sigma_a^2 + \sigma_b^2 + \sigma_\varepsilon^2 \\ \text{Cov}(Y_{ij}, Y_{ih}) &= \sigma_a^2 + \sigma_b^2 j h. \end{aligned}$$

In questo caso non si ha più equicorrelazione tra osservazioni entro le unità. Solitamente gli effetti fissi vengono distinti in componenti del primo tipo, ovvero **effetti fissi fra unità** (*between-subject*), e in effetti del secondo tipo detti **effetti fissi entro le unità** (*within-subject*). Nel modello (1.1) β_1 sarebbe un effetto del primo tipo mentre β_2 del secondo.

Una forma generale per i modelli misti lineari normali è data da

$$Y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{u}_i + \varepsilon_{ij}, \quad (1.2)$$

con \mathbf{x}_{ij} e \mathbf{z}_{ij} vettori riga. Si assume $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ indipendente da \mathbf{u}_i dove quest'ultimo è una realizzazione di una variabile casuale q -dimensionale con distribuzione $G(\cdot)$, chiamata *mixing distribution*, che è parametrizzata da ψ , vettore i cui elementi sono ignoti. L'ipotesi distributiva più comune è quella normale multivariata, con vettore delle medie uguale a 0 e matrice di varianze e covarianze Σ_u . Gli elementi ignoti di Σ_u sono detti **componenti di varianza** (*variance components*). La struttura più semplice è quella di indipendenza fra effetti casuali che prevede zeri al di fuori della diagonale principale di Σ_u , tuttavia modificandone la struttura è possibile tenere conto di casi più complessi, come nel caso di strutture gerarchiche.

In termini matriciali, assumendo $m_i = m$ per $i = 1, \dots, n$, la (1.2) equivale a

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{u}_i + \varepsilon_i \quad (1.3)$$

dove \mathbf{X}_i è la **matrice del modello** per gli effetti fissi con dimensione $n \times p$ avente \mathbf{x}_{ij} sulla generica riga. \mathbf{Z}_i è la matrice del modello $m \times q$ degli effetti casuali con generica riga \mathbf{z}_{ij} e $\varepsilon_i \sim N_m(0, \mathbf{R})$ indipendenti da \mathbf{u}_i e \mathbf{R} è la struttura di correlazione tra errori relativi alla medesima unità. Oltre alla già citata struttura d'equicorrelazione, vi sono altre possibili assunzioni per la matrice \mathbf{R} , tra cui la struttura autoregressiva, con $Cor(\varepsilon_{ij}, \varepsilon_{ih}) = \rho^{|j-h|}$, oppure una struttura di **Toeplitz** con $Cor(\varepsilon_{ij}, \varepsilon_{ih}) = \rho_{|j-h|}$, dipendente da $m - 1$ parametri.

Una struttura di correlazione non adeguata può comportare valutazioni errate degli *standard error* relativi agli effetti fissi entro le unità, per esempio β_2 nel modello (1.1). Possiamo ora per semplicità assumere indipendenza, $\mathbf{R} = \sigma_\varepsilon^2 I_m$. Inoltre β è il vettore degli effetti fissi e \mathbf{u}_i è il vettore degli effetti casuali. Il modello quindi comporta

$$E(\mathbf{Y}_i | \mathbf{u}_i) = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{u}_i$$

che ha la forma di un modello lineare con *offset* pari a $\mathbf{Z}_i\mathbf{u}_i$, e $Var(\mathbf{Y}_i | \mathbf{u}_i) = \sigma_\varepsilon^2 I_m$. Inoltre, assumendo la normalità degli effetti casuali $\mathbf{u}_i \sim N_q(0, \Sigma_u)$, otteniamo che marginalmente, \mathbf{Y}_i ha distribuzione normale con media $\mathbf{X}_i\beta$ e

$$Var(\mathbf{Y}_i) = V_i = \mathbf{Z}_i\Sigma_u\mathbf{Z}_i^\top + \sigma_\varepsilon^2 I_m. \quad (1.4)$$

Lo stimatore di massima verosimiglianza di β si ottiene a partire dal modello marginale per $\mathbf{y}_i = (\mathbf{y}_1^\top, \dots, \mathbf{y}_n^\top)^\top$. Posto $V_i = \mathbf{Z}_i\Sigma_u\mathbf{Z}_i^\top + \mathbf{R}$, assumendo V_1, \dots, V_n note, lo stimatore di massima verosimiglianza ha distribuzione normale, si rimanda a Salvani et al. (2020b, Paragrafo 7.3.1) per i dettagli.

Come nel modello normale per stimare la varianza degli errori \mathbf{R} si preferisce lo stimatore corretto per i gradi di libertà allo stimatore di massima verosimiglianza, in modo analogo la stima di V_i è basata su una **verosimiglianza marginale**. La verosimiglianza è ottenuta a partire dal modello statistico per una trasformazione di \mathbf{Y} , con distribuzione indipendente da β . I residui ottenuti dal modello per la trasformazione sono i residui linearmente indipendenti della regressione lineare di \mathbf{Y} su \mathbf{X} . Il metodo è chiamato **REML**, *restricted maximum likelihood estimation*, si veda Davison (2003, paragrafo 12.2) per approfondimenti. L'adattamento dei modelli lineari normali misti è ottenibile in \mathbf{R} con la funzione `lme` della libreria `nlme` (Pinheiro & Bates (2023)) oppure con la funzione `lmer` della libreria `lme4` (Bates et al. (2014)).

Previsione degli effetti casuali

In un modello lineare normale con effetti misti (1.3), è possibile ottenere una **previsione degli effetti casuali** \mathbf{u}_i . Si parla di previsione e non di stima poiché gli effetti casuali sono variabili casuali e non parametri. Il miglior predittore lineare non distorto (**BLUP**, *Best Linear Unbiased Predictor* si veda Henderson (1975)) $\tilde{\mathbf{u}}_i$ di \mathbf{u}_i è una funzione lineare di \mathbf{Y} con media zero e tale che, per qualsiasi combinazione lineare $\mathbf{a}^\top \mathbf{u}_i$, l'errore quadratico medio $E(\mathbf{a}^\top \mathbf{u}_i - \mathbf{a}^\top \tilde{\mathbf{u}}_i)^2$ è minimo fra i predittori lineari non distorti. Il previsore **BLUP** coincide con la stima di $E(\mathbf{u}_i | \mathbf{Y}_i = \mathbf{y}_i)$, ovvero il valore atteso condizionato ottenuto dalla distribuzione congiunta di \mathbf{Y}_i e \mathbf{u}_i . Se $G(\cdot)$ è normale, si ha che la distribuzione congiunta di \mathbf{Y}_i e \mathbf{u}_i è ancora normale, più precisamente

$$\begin{pmatrix} \mathbf{Y}_i \\ \mathbf{u}_i \end{pmatrix} \sim N_{m_i+q} \left(\begin{pmatrix} \mathbf{X}_i \beta \\ 0 \end{pmatrix}, \begin{pmatrix} V_i & \mathbf{Z}_i \Sigma_u \\ \mathbf{Z}_i \Sigma_u & \Sigma_u \end{pmatrix} \right)$$

dove $V_i = \mathbf{Z}_i \Sigma_u \mathbf{Z}_i^\top + \mathbf{R}$, cfr. (1.4) e

$$\text{Cov}(\mathbf{Y}_i, \mathbf{u}_i) = E[(\mathbf{X}_i \beta + \mathbf{Z}_i \mathbf{u}_i + \varepsilon_i) \mathbf{u}_i^\top] = \mathbf{Z}_i \Sigma_u$$

risultato dovuto all'indipendenza tra \mathbf{u}_i e ε_i . Per i risultati relativi alle distribuzioni condizionate in normali multivariate, risulta

$$E(\mathbf{u}_i | \mathbf{Y}_i = \mathbf{y}_i) = 0 + \Sigma_u \mathbf{Z}_i^\top V_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \beta)$$

e dunque se le matrici di covarianza sono note,

$$\tilde{\mathbf{u}}_i = \Sigma_u \mathbf{Z}_i^\top V_i^{-1} (\mathbf{y}_i - \mathbf{X}_i \beta).$$

Nel caso in cui V_i e Σ_u non sono note, verranno sostituite con le rispettive stime. Anche in modelli lineari generalizzati misti o nei modelli lineari normali misti quando \mathbf{u}_i non ha distribuzione normale, si possono ottenere previsioni degli effetti casuali basati su stime di $E(\mathbf{u}_i | \mathbf{Y}_i = \mathbf{y}_i)$, ottenute numericamente a partire dalla distribuzione congiunta $(\mathbf{Y}_i, \mathbf{u}_i)$. Stime BLUP possono essere ottenute in R con funzione `ranef` delle librerie `nlme` e `lme4` (Bates et al. (2014)), quest'ultima libreria tratta i modelli GLMM e le stime BLUP si basano sulle mode di $\mathbf{u}_i | \mathbf{Y}_i = \mathbf{y}_i$.

Capitolo 2

Valutazione della significatività degli effetti casuali nei modelli misti

Introduzione

Nella pratica, ogniqualvolta si adatti un modello parametrico o semiparametrico, è d'interesse valutare la significatività dei coefficienti stimati.

Quando si ha a che fare con modelli normali lineari misti, è necessario un metodo per valutare se un certo effetto casuale o un gruppo di effetti casuali possono essere rimossi dal modello. Tuttavia, come si vedrà in seguito, questo tipo di problema ha delle caratteristiche per cui viene detto “non standard”.

2.1 Richiami alla teoria asintotica della verosimiglianza

Si assume che per i dati y sia specificato un modello statistico parametrico $\mathcal{F} = \{p_Y(y; \theta), \theta \in \Theta \subseteq \mathbb{R}^p\}$ con parametro θ , dove $y = (y_1, \dots, y_n)$ è un campione di realizzazioni dalla variabile d -dimensionale Y con supporto $\mathbf{Y} \subseteq \mathbb{R}^d$. Si assume che il modello \mathcal{F} sia **correttamente specificato** e il parametro θ sia **identificabile**. Si assume $Y \sim P^0$, con funzione di probabilità o di densità $p^0(y)$, con $p^0(y) = p(y; \theta^0)$, per uno e un solo valore $\theta^0 \in \Theta$, detto **vero valore del parametro**. Se le componenti di y sono indipendenti la **log-verosimiglianza** associata al modello \mathcal{F} è

$$l(\theta) = l(\theta; y) = \sum_{i=1}^n \log p_{Y_i}(y_i; \theta).$$

Il vettore delle derivate parziali prime della log-verosimiglianza, detto **funzione punteggio** (*score function*), viene indicato con

$$\begin{aligned} l_*(\theta) = l_*(\theta; y) &= \left(\frac{\partial l(\theta)}{\partial \theta_1}, \dots, \frac{\partial l(\theta)}{\partial \theta_p} \right)^\top \\ &= \left[\frac{\partial l(\theta)}{\partial \theta_r} \right] \\ &= [l_r(\theta)], \end{aligned}$$

con θ_r generico elemento di θ , $r = 1, \dots, p$. Si assumono inoltre valide le **condizioni di regolarità** del modello, che assicurano la validità dei risultati asintotici, si veda ad esempio Azzalini (1996, paragrafi 3.2.3-3.3.3).

Con $i(\theta)$ si indica l'**informazione attesa**, ovvero il valore atteso dell'**informazione osservata**, $i(\theta) = E_\theta(j(\theta))$ dove

$$j(\theta) = j(\theta; y) = - \frac{\partial^2 l(\theta)}{\partial \theta_r \partial \theta_s}.$$

Si richiamano di seguito i più importanti risultati asintotici, si rimanda a Pace & Salvan (2001, capitolo 6) per i dettagli.

Per n sufficientemente grande, la funzione punteggio valutata nel vero valore del parametro ha distribuzione approssimata

$$l_*(\theta) \sim N_p(0, i(\theta)) \quad (2.1)$$

qualunque sia il vero valore di θ in Θ . Inoltre sia $\hat{\theta}_n$ lo stimatore di massima verosimiglianza di θ , sotto le condizioni sopracitate e per n sufficientemente grande vale l'approssimazione

$$\hat{\theta}_n \sim N_p(\theta, i(\theta)^{-1}).$$

Come conseguenza di questi risultati, si possono definire diversi test statistici, per verificare $H_0 : \theta = \theta_0$ contro $H_1 : \theta \neq \theta_0$. Il primo, è detto **test di Wald** ed è definito come

$$W_e(\theta_0) = (\hat{\theta}_n - \theta_0)^\top i(\theta_0) (\hat{\theta}_n - \theta_0) \quad (2.2)$$

che sotto H_0 ha distribuzione asintotica χ_p^2 . Considerando invece il risultato (2.1) per la funzione punteggio, si ricava il **test di Rao** o *score test*, definito come

$$W_u(\theta_0) = l_*(\theta_0)^\top i(\theta_0)^{-1} l_*(\theta_0) \quad (2.3)$$

che ha asintoticamente la medesima distribuzione di $W_e(\theta_0)$, e sotto H_0 differisce da quest'ultimo per una quantità asintoticamente trascurabile. Per il risultato sulla distribuzione asintotica dei due test si veda Pace & Salvan (2001, capitolo 6).

Un'ulteriore quantità pivotale approssimata è il **log-rapporto di verosimiglianza**, è indicato con $W(\theta)$, e si definisce come

$$W(\theta_0) = 2(l(\hat{\theta}_n) - l(\theta_0)), \quad (2.4)$$

ed ha stessa distribuzione asintotica nulla approssimata di $W_e(\theta)$ e $W_u(\theta)$.

Spesso non si è interessati a fare inferenza su tutte le componenti di θ , ma su un sottoinsieme, detto **parametro di interesse**. Se $\theta = (\psi, \lambda)$, dove ψ è composto dalle componenti di interesse con dimensione d_1 e λ è il **parametro di disturbo** con dimensione d_2 , tale che $d_1 + d_2 = p$, allora le quantità definite in precedenza possono essere suddivise analogamente: $\hat{\theta}_n = (\hat{\psi}_n, \hat{\lambda}_n)$ e

$$l_*(\theta)^\top = (l_\psi(\psi, \lambda)^\top, l_\lambda(\psi, \lambda)^\top),$$

dove $l_\psi(\psi, \lambda) = \partial l(\psi, \lambda) / \partial \psi$ e $l_\lambda(\psi, \lambda) = \partial l(\psi, \lambda) / \partial \lambda$. L'informazione attesa può essere scomposta in blocchi

$$i(\theta) = \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix}, \quad j(\theta) = \begin{pmatrix} j^{\psi\psi} & j^{\psi\lambda} \\ j^{\lambda\psi} & j^{\lambda\lambda} \end{pmatrix},$$

così come $i(\theta)^{-1}$ e $j(\theta)^{-1}$:

$$i(\theta)^{-1} = \begin{pmatrix} i^{\psi\psi} & i^{\psi\lambda} \\ i^{\lambda\psi} & i^{\lambda\lambda} \end{pmatrix}, \quad j(\theta)^{-1} = \begin{pmatrix} j^{\psi\psi} & j^{\psi\lambda} \\ j^{\lambda\psi} & j^{\lambda\lambda} \end{pmatrix},$$

si veda Salvan et al. (2020b, paragrafo 1.5.3). Definendo $\hat{\theta}_{\psi_0}$, la stima di massima verosimiglianza di θ sotto $H_0 : \psi = \psi_0$, si ha che $\hat{\theta}_{\psi_0} = (\psi_0, \hat{\lambda}_{\psi_0})$, dove $\hat{\lambda}_{\psi_0}$ è la stima di massima verosimiglianza di λ con $\psi = \psi_0$, soluzione dell'equazione di verosimiglianza parziale $l_\lambda(\psi_0, \lambda) = 0$. Si ottengono quindi le versioni **profilo** delle quantità (2.2)-(2.3) definite come

$$W_{eP}(\psi_0) = (\hat{\psi}_n - \psi_0)^\top (j^{\psi\psi}(\hat{\theta}_n))^{-1} (\hat{\psi}_n - \psi_0),$$

$$W_{uP}(\psi_0) = l_\psi(\hat{\theta}_{\psi_0})^\top i^{\psi\psi}(\hat{\theta}_{\psi_0}) l_\psi(\hat{\theta}_{\psi_0}),$$

entrambe con distribuzione asintotica χ_{d1}^2 . Allo stesso modo si definisce il log-rapporto di verosimiglianza profilo

$$W_P(\psi_0) = 2(l(\hat{\theta}_n) - l(\hat{\theta}_{\psi_0})) \sim \chi_{d1}^2 \quad (2.5)$$

con $l(\hat{\theta}_{\psi_0}) = l(\psi_0, \hat{\lambda}_{\psi_0})$, detta **log-verosimiglianza profilo**. La distribuzione venne dimostrata da Wilks (1938).

Un importante risultato è che nel confronto di modelli annidati normali è possibile ricavare la distribuzione esatta di $W_P(\psi)$. Siano i modelli completo e ridotto rispettivamente

$$\mathbf{Y} = \begin{bmatrix} X_0 & X_1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \varepsilon = X\beta + \varepsilon$$

$$\mathbf{Y} = X_0\beta_0 + \varepsilon$$

dove X_0 ha dimensione $n \times p_0$ e X ha dimensione $n \times p$, con $p > p_0$. Valutare il modello ridotto contro il modello completo equivale a valutare la nullità della componente β_1 . Il test F è una trasformazione monotona crescente del test $W_P(\beta_1)$, è quindi una versione equivalente, ed ha la forma

$$\left(\frac{e_0^\top e_0 - e^\top e}{e^\top e} \right) \frac{(n-p)}{(p-p_0)} \quad (2.6)$$

dove con e si indica il vettore dei residui nel modello completo e e_0 quelli nel modello ridotto. Il test ha distribuzione esatta $F_{p-p_0, n-p}$, si rimanda a Grigoletto et al. (2017, paragrafo 4.8.2) per la dimostrazione.

2.2 Limiti del test log-rapporto di verosimiglianza nei modelli misti

Sia θ p -dimensionale, siano $\theta = (\psi, \lambda)$ parametri di interesse e disturbo, dove ψ ha dimensione p_0 e λ $p - p_0$. Il test del log-rapporto di verosimiglianza è spesso usato per verificare i ipotesi del tipo $H_0 : \psi_1 = \dots = \psi_{p_0} = 0$ contro $H_1 : \bar{H}_0$. Vogliamo quindi valutare la nullità della componente $p - p_0$ -dimensionale ψ . Il test ha espressione

$$W_P(\psi) = 2(l(\hat{\psi}_n, \hat{\lambda}_n) - l(\psi, \hat{\lambda}_\psi))|_{\psi=0},$$

come visto nella (2.5) il test ha distribuzione $\chi_{p_0}^2$.

Indicando Ψ lo spazio parametrico di ψ contenuto in \mathbb{R}^{p-p_0} , W_P verifica l'ipotesi che ψ si trovi in un sottoinsieme di Ψ , che indichiamo con Ψ_0 , contro l'alternativa che

ψ appartenga al complemento di Ψ_0 in Ψ , che indichiamo con Ψ_1 . La distribuzione asintotica di W_P è valida solo se ψ è un punto interno di Ψ . Nel caso del più semplice modello lineare misto in cui vi è un'intercetta casuale b_i con distribuzione normale di media 0 e varianza σ_b^2 , vogliamo valutare la significatività dell'effetto casuale e quindi confrontare i due modelli lineari normali

$$Y_{ij} = x_{ij}\beta + b_i + \varepsilon_{ij}$$

$$Y_{ij} = x_{ij}\beta + \varepsilon_{ij}.$$

Valutare la rimozione di b_i , equivale a verificare $H_0 : \sigma_b^2 = 0$ contro $H_1 : \sigma_b^2 > 0$. Lo spazio parametrico di σ_b^2 è \mathbb{R}^+ e dunque sotto H_0 il parametro si trova sulla frontiera del suo spazio parametrico, invalidando la distribuzione asintotica χ^2 del test log-rapporto di verosimiglianza. In particolare Self & Liang (1987) e Stram & Lee (1994) mostrano che in questo caso la distribuzione asintotica del test è una miscela 50 : 50 di un χ_1^2 e un χ_0^2 (con l'intera massa di probabilità in 0). Se si utilizza il test al livello nominale di significatività 5% assumendo la distribuzione nulla χ_1^2 del test in condizioni standard, avremmo ottenuto un test con probabilità effettiva di errore di primo tipo pari a 2.5%.

Più in generale, in un modello misto del tipo (1.3), sotto l'ipotesi che il vettore delle risposte y possa essere ripartito in n sottogruppi, y_i , realizzazioni di variabili indipendenti e identicamente distribuite da Y_i , indicata con Σ_u la matrice di varianze e covarianze degli effetti casuali, si vuole valutare

$$H_0 : \Sigma_u = \begin{pmatrix} D_{11} & 0 \\ 0 & 0 \end{pmatrix} \text{ contro } H_1 : \Sigma_u = \begin{pmatrix} D_{11} & D_{12} \\ D_{12}^\top & D_{22} \end{pmatrix}, \quad (2.7)$$

con D_{11} matrice $q \times q$ definita positiva, e D_{22} matrice $k \times k$, dove k è il numero di effetti casuali di cui si considera l'eliminazione. La distribuzione asintotica di W è in questo caso la miscela di q variabili χ^2 indipendenti con gradi di libertà pari a $kq, kq - 1, kq - 2, \dots, (k - 1)q$. I pesi delle q variabili χ^2 possono essere teoricamente calcolati a partire dalla matrice di informazione i per Σ_u sotto H_0 . Tuttavia, in Stram & Lee (1994) si deduce il risultato solo nello specifico caso in cui la matrice di informazione è uguale alla matrice identica, si preferiscono perciò metodi di simulazione per approssimare direttamente la distribuzione di W_P .

Gli articoli di Self & Liang (1987) e Stram & Lee (1994), mostrano per diverse specifiche assunzioni per la struttura della risposta che il test del log-rapporto di verosimiglianza è una miscela di χ^2 , si vedano gli articoli per i dettagli. Si rimanda inoltre a Verbeke & Molenberghs (2003) e Molenberghs & Verbeke (2007) per la distribuzione

asintotica dei test di Wald sotto la medesima condizione.

Quindi, se si utilizza la distribuzione nulla del test in condizioni standard, si ottengono test **conservativi**, che hanno probabilità effettiva di errore del primo tipo minore di quella nominale. Bisogna ricordare inoltre che per campioni limitati, fare riferimento alla mistura di χ^2 produce comunque test conservativi (Crainiceanu & Ruppert (2004)).

Nella pratica, la distribuzione χ_p^2 del test $W_P(\psi)$ è comunque spesso utilizzata, come nel caso della funzione `anova` del pacchetto `lme4` Bates et al. (2014) in R.

2.3 Soluzioni esistenti

Esistono diverse metodologie per valutare la significatività delle componenti di varianza. Tra le procedure statistiche più utilizzate vi sono i metodi di simulazione, che sfruttano lo sviluppo computazionale degli ultimi 20 anni per approssimare la distribuzione nulla di test statistici come il log-rapporto di verosimiglianza (2.4). Tuttavia, esistono anche test che sfruttano risultati asintotici e, come si vedrà nel prossimo capitolo, Hui et al. (2019) propongono un test con distribuzione nulla esatta qualunque sia la distribuzione degli effetti casuali. Di seguito si richiamano alcune metodologie. Ci si concentrerà in modo più dettagliato sul test **RLRT** (*Restricted likelihood ratio test*).

Un primo metodo proposto in Fitzmaurice et al. (2007) utilizzabile per valutare una singola componente di varianza, calcola il *p-value* del log-rapporto di verosimiglianza (o del test score) sotto l'ipotesi che, condizionatamente agli effetti casuali, tutte le osservazioni sono indipendenti. Permutando gli indici dei *cluster*, ricalcolano la statistica test per un gran numero di permutazioni. Il metodo riesce a produrre un test con probabilità di errore del primo tipo pari a quello indicato.

Un altro metodo di permutazione è il test **T** Drikvandi et al. (2013). Riprendendo la struttura dei modelli misti in (1.3) viene introdotta la statistica **T**

$$\mathbf{T} = \frac{1}{N} \text{tr}(Z_*(I \otimes \hat{D}_{22})Z_*^\top)$$

dove con $\text{tr}()$ si indica la funzione traccia. La matrice di esplicative degli effetti casuali Z_* ha una struttura diagonale a blocchi, composta dalle matrici Z_1, \dots, Z_n (ovvero le matrici del modello degli effetti casuali viste in (1.3)) di dimensione $n \times q$. L'operatore \otimes indica il **prodotto di Kronecker** (si veda Appendice) e \hat{D}_{22} è uno stimatore non distorto di D_{22} matrice di varianze e covarianze degli effetti valutati, secondo la medesima struttura di Σ_u in (2.7). Chiaramente sotto l'ipotesi $H_0 : D_{22} = 0$ vale $E(\mathbf{T}) = 0$, quindi si misurano gli allontanamenti dall'ipotesi nulla come allontanamenti dalla media. Si

procede permutando il campione e ricalcolando la statistica \mathbf{T} per ogni permutazione, approssimandone la distribuzione. Il metodo può essere utilizzato per valutare tutti gli effetti casuali oppure un sottogruppo degli stessi, tuttavia in quest'ultimo caso il metodo perde di efficacia in quanto sotto H_0 , le Y_{ij} non sono indipendenti.

Un approccio diverso di simulazione è presentato in Sinha (2009), si utilizza un *bootstrap parametrico* per simulare la distribuzione del test score ottenuto dalla verosimiglianza marginale della risposta sotto l'ipotesi di nullità degli effetti casuali. Il metodo è valido per varie tipologie di dati, per valutare la nullità di uno o più effetti casuali.

In Qu et al. (2013) viene introdotto un metodo di simulazione per l'eliminazione di uno o più effetti casuali basato sul test score derivato dalla log-verosimiglianza ristretta.

Ricordiamo inoltre il test introdotto in Li et al. (2014) pensato per i modelli **ANOVAMM**. Il test non usa metodi di simulazione ed è utilizzabile indipendentemente dalla numerosità campionaria e dal numero di componenti da valutare, tuttavia esso è applicabile solo in presenza di indipendenza degli effetti casuali. I modelli ANOVAMM (*Analysis of variance mixed model*) estendono ai modelli misti il problema dell'analisi della varianza (ANOVA). Le osservazioni sono quindi suddivise in gruppi secondo una o più variabili qualitative e si vogliono valutare le possibili differenze che intercorrono tra le osservazioni che presentano combinazioni differenti delle modalità dei fattori, tenendo comunque conto della correlazione che potrebbe esserci tra osservazioni appartenenti allo stesso gruppo. I modelli ANOVAMM hanno struttura del tipo

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}_1 a_1 + \cdots + \mathbf{Z}_s a_s + \varepsilon$$

dove \mathbf{Y} è il vettore delle risposte di dimensione $n \times 1$, \mathbf{X} è la matrice $n \times p$ delle esplicative, β il vettore $p \times 1$ degli effetti fissi, e \mathbf{Z}_k per $k = 1, \dots, s$ sono matrici note associate con il vettore di effetti casuali a_k . Ogni a_k ha lunghezza variabile m_k , media 0 e matrice di covarianza del tipo $\sigma_k^2 I_{m_k}$. In pratica le osservazioni appartengono a s gruppi (ogni gruppo di numerosità m_k), e per ogni gruppo vi è un effetto casuale differente.

2.3.1 Log-Rapporto di verosimiglianza ristretta

Si pensi a un modello lineare misto definito come in (1.3)

$$\mathbf{Y} = X\beta + Z\mathbf{b} + \varepsilon,$$

con

$$Var\left(\begin{matrix} \mathbf{b} \\ \varepsilon \end{matrix}\right) = \begin{pmatrix} \Sigma_b & 0 \\ 0 & \sigma_\varepsilon^2 I_n \end{pmatrix}$$

dove la matrice di covarianza degli effetti casuali è definita come $\Sigma_b = \sigma_b^2 D$, con D matrice simmetrica definita positiva $q \times q$. Come già visto in (1.4) segue

$$E(\mathbf{Y}) = X\beta,$$

$$Var(\mathbf{Y}) = \sigma_\epsilon^2 I_n + \sigma_b^2 Z D Z^\top = \sigma_\epsilon^2 V_\lambda,$$

con $\lambda = \sigma_b^2/\sigma_\epsilon^2$, $V_\lambda = I_n + \lambda Z D Z^\top$. Il parametro λ può essere interpretato come il rapporto tra segnale e rumore, ha spazio parametrico $[0, \infty)$. Si nota immediatamente che $\lambda = 0 \Leftrightarrow \sigma_b^2 = 0$ di conseguenza l'ipotesi nulla è $H_0 : \sigma_b^2 = 0$ ($\lambda = 0$) contro $H_1 : \sigma_b^2 > 0$ ($\lambda > 0$). Stiamo assumendo quindi indipendenza fra le osservazioni sotto H_0 .

Siano $\mu_{s,n}$ e $\xi_{s,n}$ i q autovalori delle matrici $q \times q$, $D^{1/2} Z^\top P_0 Z D^{1/2}$ e $D^{1/2} Z^\top Z D^{1/2}$, con $P_0 = I_n - X(X^\top X)^{-1} X^\top$. Secondo il Teorema 1 in Crainiceanu & Ruppert (2004) definiamo

$$N_n(\lambda) = \sum_{s=1}^q \frac{\lambda \mu_{s,n}}{1 + \lambda \mu_{s,n}} w_s^2$$

$$D_n(\lambda) = \sum_{s=1}^q \frac{w_s^2}{1 + \lambda \mu_{s,n}} + \sum_{s=q+1}^{n-p} w_s^2$$

dove w_s per $s = 1, \dots, n - p$ sono variabili normali standard indipendenti. Il **log-rapporto di verosimiglianza ristretta** è definito come

$$RLRT \stackrel{D}{=} \sup_{\lambda \geq 0} \left[(n - p) \log \left\{ 1 + \frac{N_n(\lambda)}{D_n(\lambda)} \right\} - \sum_{s=1}^q \log(1 + \lambda \mu_{s,n}) \right].$$

Per approssimare la distribuzione di $RLRT$ si sfrutta il metodo Monte-Carlo simulando le variabili normali standard w_s , si veda Crainiceanu & Ruppert (2004) per i dettagli.

Un vantaggio del metodo è che gli autovalori vengono calcolati una sola volta all'inizio della procedura, inoltre la velocità di calcolo dipende dal numero di componenti q di \mathbf{b} , e non dalla numerosità campionaria n . Il test può essere utilizzato in R con la funzione `exactRLRT` del pacchetto `RLRsim` Scheipl et al. (2008).

Una soluzione alternativa confronta la statistica test con la distribuzione asintotica ottenuta da Self & Liang (1987). Il metodo è implementato in R dalla funzione `reml.lrt.asreml` del pacchetto `asremlPlus` (Brien (2023)).

2.3.2 Tabella riassuntiva

La tabella presenta brevemente alcune metodologie esistenti. Si rimanda agli articoli per approfondimenti, e al pacchetto R che li implementa se presente.

Metodo	Modello	Caratteristiche
Fitzmaurice et al. (2007)	GLMM	un effetto casuale alla volta, approccio di permutazione
Drikvandi et al. (2013)	LMM	più effetti casuali, approccio di permutazione
Sinha (2009)	GLMM	più effetti casuali, approccio di simulazione, costi computazionali che aumentano se il modello nullo comprende effetti casuali
Qu et al. (2013)	LMM	più effetti casuali, distribuzione del test esatta se il modello nullo non comprende effetti casuali, approccio di simulazione in caso contrario
Li et al. (2014)	ANOVAMM	più effetti casuali, il test ha distribuzione asintotica nota
<i>PBLR</i>	GLMM	più effetti casuali, approccio di simulazione, pacchetto <code>pkbrtest</code> Halekoh & Højsgaard (2014)
<i>RLRT</i>	LMM	un effetto casuale alla volta, approccio di simulazione, pacchetto <code>RLRsim</code> Scheipl et al. (2008)

TABELLA 2.1

Capitolo 3

Il test F basato su combinazioni lineari della risposta

Introduzione

Nel capitolo precedente sono state presentate alcune soluzioni al problema in esame. Tutti gli approcci riassunti nella Tabella 2.1 sono di tipo computazionale e pongono dei limiti, in termini di tempo e di potenza di calcolo, alla complessità delle applicazioni. Uno strumento flessibile che offre buoni risultati in una ampia categoria di situazioni è il test **FLC** (*F-test based on linear combination of the responses*), che verrà descritto in dettaglio in questo capitolo. Il riferimento è Hui et al. (2019).

3.1 Il test FLC

Innanzitutto è necessario generalizzare la struttura (1.3). Sia $y = (y_1, \dots, y_N)^\top$ il vettore delle variabili risposta, \mathbf{X} la matrice $N \times p$ delle esplicative degli effetti fissi e \mathbf{Z} la matrice $N \times q$ delle covariate degli effetti casuali allora definiamo un modello normale misto come

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}b + \epsilon \quad (3.1)$$

dove β e b sono rispettivamente i vettori degli effetti casuali e misti. Gli errori ϵ hanno distribuzione normale $N_N(0, \mathbf{R})$, assumiamo che $\mathbf{R} = \sigma^2 \mathbf{G}$ dove \mathbf{G} è una matrice nota a rango pieno. Dato che premoltiplicare \mathbf{Y} per $\mathbf{G}^{-1/2}$ riconduce alla struttura di indipendenza $\mathbf{R} = \sigma^2 I_N$, assumeremo senza perdita di generalità $\mathbf{G} = I_N$. Chiaramente \mathbf{G} non è sempre nota ma potrebbe contenere dei parametri ignoti che devono essere

stimati. Gli Autori suggeriscono che \mathbf{G} debba essere stimata sotto il modello saturo e riconoscono che il metodo da utilizzare per la stima sia un quesito da approfondire.

Vale, come visto nel primo capitolo, $E(b) = 0$ e $Var(b) = \Sigma_b$, tuttavia non si fanno assunzioni distributive. Il modello definito in questo modo riesce a includere le più comuni applicazioni dei modelli misti.

Nel modello (3.1), $N = \sum_{i=1}^n m_i$ con n numero di gruppi e m_i numerosità dell' i -esimo gruppo, la matrice \mathbf{Z} ha una struttura diagonale con n blocchi

$$\mathbf{Z} = \begin{pmatrix} \mathbf{Z}_1 & 0 & \dots & 0 \\ 0 & \mathbf{Z}_2 & \dots & 0 \\ 0 & \ddots & \dots & 0 \\ 0 & 0 & \dots & \mathbf{Z}_n \end{pmatrix}$$

in cui il generico blocco \mathbf{Z}_i con $i = 1, \dots, n$ ha dimensione $m_i \times q_c$, con q_c tale che $q = nq_c$. Analogamente anche le matrici di covarianza \mathbf{R} e Σ_b degli errori e degli effetti casuali hanno struttura diagonale a blocchi, in particolare \mathbf{R} ha sulla diagonale le matrici \mathbf{R}_i , con dimensione $m_i \times m_i$, che sotto l'assunzione di indipendenza fra errori poniamo uguali a $\sigma^2 I_{m_i}$ per $i = 1, \dots, n$, di conseguenza come già visto $\mathbf{R} = \sigma^2 I_N$. Vale invece $\Sigma_b = I_n \otimes \mathbf{D}$ con dimensione $q \times q$, dove \otimes indica il prodotto di Kronecker, e \mathbf{D} ha dimensione $q_c \times q_c$. Per ogni *cluster* vi sono q_c effetti casuali, condizione che caratterizza modelli con gruppi indipendenti. Sotto il modello lineare misto appena definito y è realizzazione di

$$\begin{aligned} \mathbf{Y} &\sim N_N(\mathbf{X}\beta, V) \text{ con} \\ V &= \mathbf{Z}\Sigma_b\mathbf{Z}^\top + \sigma^2 I_N. \end{aligned}$$

Si considera quindi l'ipotesi $H_0 : \Sigma_b = \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & 0 \end{pmatrix}$ contro $H_1 : \Sigma_b = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^\top & \Sigma_{22} \end{pmatrix}$ dove Σ_{11} è una matrice $q_0 \times q_0$ definita positiva con $q_0 < q$ e Σ_{12} , Σ_{22} sono matrici diverse da zero con dimensione appropriata. Stiamo quindi valutando congiuntamente la nullità di $q - q_0$ effetti casuali, senza restrizioni distributive o vincoli sulla struttura delle matrici di covarianza, condizioni che nessuna procedura presente nel Capitolo 2 è in grado di assicurare.

Siano A_1 e A_2 due matrici con dimensioni $a \times b$ arbitrarie con $a > b$, indichiamo con $rk(A_1)$ il rango della matrice A_1 e con $\mathbf{P}_{A_1} = A_1(A_1^\top A_1)^- A_1^\top$ la corrispondente matrice di proiezione nel sottospazio generato dalle colonne di A_1 . In questo caso con $(A_1^\top A_1)^-$ indichiamo l'**inversa generalizzata** (si veda l'Appendice per un richiamo sulle matrici di proiezioni generate da matrici del modello \mathbf{X} con colonne non linearmente

indipendenti). Con $rk(A_1, A_2)$ e \mathbf{P}_{A_1, A_2} indicheremo il rango e la proiezione della matrice (A_1, A_2) formata dalla concatenazione delle colonne di A_1 e A_2 . Definiamo le partizioni $\mathbf{Z} = (\mathbf{Z}_0, \mathbf{Z}_1)$, in cui \mathbf{Z}_0 sono le prime q_0 colonne di \mathbf{Z} e \mathbf{Z}_1 le ultime $q - q_0$ e $b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}$ dove b_0 sono le prime q_0 componenti e b_1 le ultime le ultime $q - q_0$. Si introducono le matrici \mathbf{L}_0 e \mathbf{L}_1 , con dimensione $N \times rk(\mathbf{L}_0)$ e $N \times rk(\mathbf{L}_1)$, con $rk(\mathbf{L}_0) = N - rk(\mathbf{X}, \mathbf{Z}_0)$ e $rk(\mathbf{L}_1) = N - rk(\mathbf{X}, \mathbf{Z})$ rispettivamente, tali che

$$\begin{aligned}\mathbf{L}_0^\top \mathbf{L}_0 &= I_{N-rk(\mathbf{X}, \mathbf{Z}_0)} \\ \mathbf{L}_0 \mathbf{L}_0^\top &= I_N - \mathbf{P}_{\mathbf{X}, \mathbf{Z}_0}\end{aligned}$$

e allo stesso modo

$$\begin{aligned}\mathbf{L}_1^\top \mathbf{L}_1 &= I_{N-rk(\mathbf{X}, \mathbf{Z})} \\ \mathbf{L}_1 \mathbf{L}_1^\top &= I_N - \mathbf{P}_{\mathbf{X}, \mathbf{Z}}.\end{aligned}$$

Vale che $\mathbf{L}_0^\top \mathbf{X} = 0$, $\mathbf{L}_0^\top \mathbf{Z}_0 = 0$ e allo stesso modo $\mathbf{L}_1^\top \mathbf{X} = 0$, $\mathbf{L}_1^\top \mathbf{Z} = 0$, ciò si dimostra per le quattro identità come segue

$$\begin{aligned}\mathbf{L}_1^\top (\mathbf{X}, \mathbf{Z}) &= I_{N-rk(\mathbf{X}, \mathbf{Z})} \mathbf{L}_1^\top (\mathbf{X}, \mathbf{Z}) \\ &= \mathbf{L}_1^\top \mathbf{L}_1 \mathbf{L}_1^\top (\mathbf{X}, \mathbf{Z}) = \mathbf{L}_1^\top (I_N - \mathbf{P}_{\mathbf{X}, \mathbf{Z}}) (\mathbf{X}, \mathbf{Z}) \\ &= \mathbf{L}_1^\top ((\mathbf{X}, \mathbf{Z}) - (\mathbf{X}, \mathbf{Z})) = 0.\end{aligned}$$

Costruiamo quindi le combinazioni lineari

$$\begin{aligned}\mathbf{u}_0 &= \mathbf{L}_0^\top \mathbf{Y} = \mathbf{L}_0^\top \mathbf{Z}_1 b_1 + \mathbf{L}_0^\top \epsilon \\ \mathbf{u}_1 &= \mathbf{L}_1^\top \mathbf{Y} = \mathbf{L}_1^\top \epsilon.\end{aligned}$$

I modelli per le trasformazioni lineari della risposta \mathbf{u}_0 e \mathbf{u}_1 sono indipendenti dagli effetti fissi e da un sottoinsieme degli effetti casuali del modello iniziale, intuizione simile a quella che sta alla base della verosimiglianza ristretta (REML).

Per valutare l'ipotesi (3.1), la statistica test, che indicheremo con *FLC* è

$$\begin{aligned}FLC &= \frac{(\mathbf{u}_0^\top \mathbf{u}_0 - \mathbf{u}_1^\top \mathbf{u}_1) / [rk(\mathbf{X}, \mathbf{Z}) - rk(\mathbf{X}, \mathbf{Z}_0)]}{(\mathbf{u}_1^\top \mathbf{u}_1) / [N - rk(\mathbf{X}, \mathbf{Z})]} \\ &= \frac{y^\top (\mathbf{P}_{\mathbf{X}, \mathbf{Z}} - \mathbf{P}_{\mathbf{X}, \mathbf{Z}_0}) y / [rk(\mathbf{X}, \mathbf{Z}) - rk(\mathbf{X}, \mathbf{Z}_0)]}{y^\top (I_N - \mathbf{P}_{\mathbf{X}, \mathbf{Z}}) y / [N - rk(\mathbf{X}, \mathbf{Z})]}\end{aligned}\tag{3.2}$$

con distribuzione nulla esatta $F_{rk(\mathbf{X}, \mathbf{Z}) - rk(\mathbf{X}, \mathbf{Z}_0), N - rk(\mathbf{X}, \mathbf{Z})}$. Se l'ipotesi nulla è $H_0 : \Sigma_b = 0$

contro l'alternativa come in (3.1), il test FLC è

$$F = \frac{y^\top (\mathbf{P}_{\mathbf{X}, \mathbf{Z}} - \mathbf{P}_{\mathbf{X}}) y / [rk(\mathbf{X}, \mathbf{Z}) - p]}{y^\top (I_N - \mathbf{P}_{\mathbf{X}, \mathbf{Z}}) y / [N - rk(\mathbf{X}, \mathbf{Z})]}$$

con distribuzione nulla esatta $F_{rk(\mathbf{X}, \mathbf{Z}) - p, N - rk(\mathbf{X}, \mathbf{Z})}$. In pratica si applica il test F (2.6) per confrontare i due modelli condizionandoci ai valori di b_0 . Se la matrice (\mathbf{X}, \mathbf{Z}) presenta multicollinearità, la struttura del test rimane la stessa in quanto nella definizione delle matrici di proiezione $\mathbf{P}_{\mathbf{X}, \mathbf{Z}}$ e $\mathbf{P}_{\mathbf{X}, \mathbf{Z}_0}$ viene utilizzata l'inversa generalizzata. Un caso estremo si presenta quando il modello degli effetti fissi e quello degli effetti casuali includono le medesime covariate ($q_c = p$, $q = np$), come accade nel caso di gruppi indipendenti. Supponiamo di voler valutare l'ipotesi che per ogni gruppo solo i primi $q_{0c} < q_c$ effetti casuali siano diversi da 0, in questo caso i gradi di libertà del numeratore del test FLC sono pari a $rk(\mathbf{X}, \mathbf{Z}) - rk(\mathbf{X}, \mathbf{Z}_0) = np - (nq_{0c} + p - q_{0c}) = (n-1)(p - q_{0c})$. L'estremo opposto si ha quando le colonne \mathbf{X} e \mathbf{Z} sono ortogonali, il che comporta $rk(\mathbf{X}, \mathbf{Z}) - rk(\mathbf{X}, \mathbf{Z}_0) = (nq_c + p) - (nq_{0c} + p) = n(q_c - q_{0c})$.

Il test è applicabile solo se $N - rk(\mathbf{X}, \mathbf{Z}) > 0$, che in un modello con *cluster* indipendenti impone che $\min_{i=1, \dots, n} \{m_i\} - q_c > 0$. Questa condizione, per quanto comporti una forte limitazione dei casi d'applicazione del test, è rispettata da molte situazioni reali.

Si dimostra facilmente che la distribuzione esatta F è valida qualsiasi sia la distribuzione degli effetti casuali. Sotto l'ipotesi di indipendenza dei residui

$$\mathbf{Y}|b \sim N_N(\mathbf{X}\beta + \mathbf{Z}b, \sigma^2 I_N),$$

segue poi da $\mathbf{L}_1^\top \mathbf{L}_1 = I_{N - rk(\mathbf{X}, \mathbf{Z})}$ che

$$\mathbf{u}_1|b = \mathbf{L}_1^\top \mathbf{Y} \sim N_q(0, \sigma^2 I_{N - rk(\mathbf{X}, \mathbf{Z})}).$$

La distribuzione condizionata non dipende da b , quindi qualunque sia la distribuzione degli effetti casuali vale

$$\sigma^{-2} \mathbf{u}_1^\top \mathbf{u}_1 = \sigma^{-2} \mathbf{Y}^\top \mathbf{L}_1 \mathbf{L}_1^\top \mathbf{Y} = \sigma^{-2} \mathbf{Y}^\top (I_N - \mathbf{P}_{\mathbf{X}, \mathbf{Z}}) \mathbf{Y} \sim \chi_{N - rk(\mathbf{X}, \mathbf{Z})}^2. \quad (3.3)$$

Per quel che riguarda il numeratore della statistica FLC , dato che $(I_N - \mathbf{P}_{\mathbf{X}, \mathbf{Z}})\mathbf{X} = 0$ e $(I_N - \mathbf{P}_{\mathbf{X}, \mathbf{Z}_0})\mathbf{X} = 0$, vale

$$\mathbf{Y}^\top (\mathbf{P}_{\mathbf{X}, \mathbf{Z}} - \mathbf{P}_{\mathbf{X}, \mathbf{Z}_0}) \mathbf{Y}|b = \mathbf{Y}^\top (\mathbf{L}_0 \mathbf{L}_0^\top - \mathbf{L}_1 \mathbf{L}_1^\top) \mathbf{Y}|b = v^\top (\mathbf{L}_0 \mathbf{L}_0^\top - \mathbf{L}_1 \mathbf{L}_1^\top) v|b$$

con $v \sim N_N(\mathbf{Z}b, \sigma^2 I_N)$. Sotto l'ipotesi nulla (3.1), gli ultimi $q - q_0$ elementi di b devono

essere uguali a 0, in quanto $E(b) = 0$ (valore atteso e varianza nulle), di conseguenza $v|b \sim N_N(\mathbf{Z}_0 b_0, \sigma^2 I_N)$. Sfruttando inoltre $\mathbf{L}_0^\top \mathbf{Z}_0 = 0$ e $\mathbf{L}_1^\top \mathbf{Z}_0 = 0$, si ha

$$v^\top (\mathbf{L}_0 \mathbf{L}_0^\top - \mathbf{L}_1 \mathbf{L}_1^\top) v | b = w^\top (\mathbf{L}_0 \mathbf{L}_0^\top - \mathbf{L}_1 \mathbf{L}_1^\top) w | b$$

con $w \sim N_N(0, \sigma^2 I_N)$. Si ottiene quindi che, indipendentemente dalla distribuzione di b ,

$$\sigma^{-2} \mathbf{Y}^\top (\mathbf{L}_0 \mathbf{L}_0^\top - \mathbf{L}_1 \mathbf{L}_1^\top) \mathbf{Y} \sim \chi_{rk(\mathbf{X}, \mathbf{Z}) - rk(\mathbf{X}, \mathbf{Z}_0)}^2, \quad (3.4)$$

sotto l'ipotesi nulla, dove $rk(\mathbf{L}_0 \mathbf{L}_0^\top - \mathbf{L}_1 \mathbf{L}_1^\top) = rk(\mathbf{X}, \mathbf{Z}) - rk(\mathbf{X}, \mathbf{Z}_0)$. Infine, poichè $(\mathbf{L}_0 \mathbf{L}_0^\top - \mathbf{L}_1 \mathbf{L}_1^\top) \mathbf{L}_1 \mathbf{L}_1^\top = (\mathbf{P}_{\mathbf{X}, \mathbf{Z}} - \mathbf{P}_{\mathbf{X}, \mathbf{Z}_0})(I_N - \mathbf{P}_{\mathbf{X}, \mathbf{Z}}) = 0$, le forme quadratiche al numeratore (3.4) e al denominatore (3.3) sono indipendenti, e entrambe non dipendono dalla distribuzione di b . La distribuzione F del test (3.2) è quindi valida a prescindere dalla *mixing distribution* $G(\cdot)$.

Non basandosi su risultati asintotici o procedure di simulazione, il test *FLC* ha dei chiari vantaggi computazionali. Inoltre, non vengono posti vincoli sulla struttura di Σ_b o di \mathbf{R} , rendendo il test applicabile a una ampia gamma di modelli lineari normali misti.

Potenza

In Hui et al. (2019) viene introdotta una formula generale per ricavare la potenza del test *FLC*. Può quindi essere ricavata la funzione che calcola la potenza (*power function*) al variare della matrice Σ_{22} sotto l'ipotesi alternativa, tramite metodi Monte-Carlo. Si espone qui il risultato. Sia $\lambda = b_1^\top \mathbf{Z}_1^\top (I_N - \mathbf{P}_{\mathbf{X}, \mathbf{Z}_0}) \mathbf{Z}_1 b_1$ e indicando con $power(\cdot)$ la funzione di potenza,

$$\begin{aligned} power(\Sigma_{22}) &= 1 - \int_{-\infty}^{\infty} \mathbf{F}(f_{1-\alpha} | \lambda, d_1, d_2) G(b_1 | \Sigma_{22}) db \\ &= 1 - E_b(\mathbf{F}(f_{1-\alpha} | \lambda, d_1, d_2) | \Sigma_{22}) \end{aligned} \quad (3.5)$$

con $d_1 = rk(\mathbf{X}, \mathbf{Z}) - rk(\mathbf{X}, \mathbf{Z}_0)$ e $d_2 = N - rk(\mathbf{X}, \mathbf{Z})$, $f_{1-\alpha}$ è il quantile di ordine $1 - \alpha$ di una variabile F con gradi di libertà pari a d_1 e d_2 e $\mathbf{F}(\cdot | \lambda, d_1, d_2)$ è la funzione di ripartizione di una variabile F non centrale (si veda Appendice) con parametro di non centralità pari a λ e gradi di libertà d_1 e d_2 . A differenza della distribuzione nulla che non dipende da $G(\cdot)$, la distribuzione del test sotto H_1 dipende dalla distribuzione degli effetti casuali.

Nel caso in cui b abbia distribuzione normale multivariata è possibile ricavare la

seguinte espressione della funzione di potenza

$$power(\Sigma_{22}) = 1 - \det(I_{q-q_0} + \Sigma_{22} \mathbf{Z}_1^\top (I_N - \mathbf{P}_{\mathbf{X}, \mathbf{Z}_0}) \mathbf{Z}_1)^{-1/2} \sum_{l=0}^{\infty} \frac{A_l}{l!} \mathbf{I} \left(\frac{d_1 f_{1-\alpha}}{d_1 f_{1-\alpha} + d_2} \middle| \frac{d_1}{2} + l, \frac{d_2}{2} \right),$$

in cui

$$A_l = \int (b_1^\top \mathbf{Z}_1^\top (I_N - \mathbf{P}_{\mathbf{X}, \mathbf{Z}_0}) \mathbf{Z}_1 b_1 / 2)^l G(b_1 | (\mathbf{Z}_1^\top (I_N - \mathbf{P}_{\mathbf{X}, \mathbf{Z}_0}) \mathbf{Z}_1 + \Sigma_{22}^{-1})^{-1}) db$$

e $\mathbf{I}(x|c_1, c_2)$ indica la **funzione beta incompleta regolarizzata** (si veda Appendice) valutata in x con parametri c_1, c_2 . Operativamente, è preferibile comunque l'espressione (3.5) anche in caso di normalità degli effetti casuali, in quanto è facilmente applicabile il metodo Monte-Carlo. A partire dall'espressione alternativa gli Autori mostrano che la potenza del test aumenta, all'aumentare di N e all'aumentare della variabilità degli effetti casuali. Al contrario quando il numero di effetti casuali q_c si avvicina alla numerosità dei *cluster* m_i , gli autori mostrano che il test, in termini di potenza si comporta peggio rispetto a metodi di simulazione come per esempio il *parametric bootstrap likelihood ratio test (PBLR)*, si veda Tabella 4.1.

3.2 Implementazione

Come suggerito in Hui et al. (2019), il test *FLC* può essere implementato in R utilizzando la funzione `anova`. L'implementazione è mostrata nell'Appendice ed è ripresa dal codice `flc.R` proposto insieme all'articolo.

Per sfruttare la funzione `anova` è necessario costruire due particolari modelli lineari, per il modello ridotto e il modello corrente. I due modelli hanno struttura del tipo

$$\mathbf{Y} = (\mathbf{X} \quad \mathbf{Z}) \begin{pmatrix} \beta \\ \eta \end{pmatrix} + \epsilon,$$

$$\epsilon \sim N_N(0, \sigma^2 \mathbf{G}),$$

con $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_n)$, \mathbf{X} la matrice delle esplicative relativa agli effetti fissi β . Se $\mathbf{G} = I_N$ imponiamo incorrelazione dei residui. La matrice \mathbf{Z} ha la medesima struttura vista in (3.1), inoltre il vettore di coefficienti η può essere suddiviso in (η_1, \dots, η_m) , dove ogni η_i per $i = 1, \dots, m$ ha dimensione $q_c \times 1$. In pratica il vettore di coefficienti η_i è composto dalle realizzazioni dei q effetti casuali b per l' i -esimo gruppo. Il modello così ottenuto ha struttura equivalente a un modello **SURE** (*seemingly unrelated regression*

equations) Zellner (1962), utilizzato in econometria. Valutare la nullità di $n(q_c - q_{0c})$ coefficienti η utilizzando il test F (2.6), corrisponde all'applicazione del test FLC . I modelli **SURE** possono essere stimati con la funzione `lm`, costruendo correttamente le matrici \mathbf{Z} e \mathbf{Z}_0 .

Nell'Appendice è riportata inoltre l'implementazione della funzione di potenza (3.5) con riferimento a `powercalc-template.R`. Il calcolo della matrice di proiezione $\mathbf{P}_{\mathbf{x}, \mathbf{z}_0}$ viene effettuato utilizzando la **pseudo-inversa di Moore-Penrose** (si veda Appendice) implementata dalla funzione `ginv` della libreria **MASS** (Venables & Ripley (2002)).

Capitolo 4

Studio di simulazione

Introduzione

Nel primo paragrafo di questo capitolo si andranno a presentare i risultati dello studio di simulazione, ricostruzione di quanto fatto da Hui et al. (2019) nel paragrafo 4 dell'articolo, con l'aggiunta della funzione di potenza (3.5). Successivamente, con le condizioni viste nel primo paragrafo, verranno costruite le curve di potenza. Il codice R utilizzato per la simulazione è presente nell'Appendice.

4.1 Riproduzione dei risultati in Hui et al. (2019)

Come già anticipato, in questo paragrafo verranno replicati i risultati ottenuti dagli Autori. In particolare per ogni test, si riporta in percentuale, il numero di volte in cui viene rifiutata l'ipotesi nulla, con livello di significatività pari a 5%, sul totale di ripetizioni indicato con n_{sim} . Viene riportato inoltre il tempo di calcolo medio per ogni test. I metodi che vengono confrontati sono

1. *FLC*-test (3.2);
2. *LRT*, log-rapporto di verosimiglianza standard (2.5);
3. *PBLRT*, *parametric bootstrap likelihood ratio* implementato dal pacchetto `pkbrtest` di Halekoh & Højsgaard (2014);
4. *Lin.score*, *linear score test* di Qu et al. (2013) implementato dal pacchetto `varComp` (non più supportato dall'attuale versione di R).

Per ogni configurazione vengono riportati in una prima Tabella i risultati presenti nell'articolo, con $n_{sim} = 1000$ e in una seconda Tabella i valori riguardanti il test FLC prodotti utilizzando il codice R presente in Appendice, con $n_{sim} = 10000$ (utilizzando il dispositivo "LAPTOP-NOAVMSUK" con processore "Intel(R) Core(TM) i5-8250U CPU @ 1.60GHz 1.80 GHz"). Viene anche indicato con $power$ il valore della potenza in percentuale calcolata con la funzione (3.5), attraverso metodo Monte-Carlo.

4.1.1 Studio di simulazione 1

La struttura per generare dati longitudinali è quella utilizzata in Drikvandi et al. (2013). Viene costruito un modello lineare misto con $cluster$ indipendenti, con $\mathbf{y}_{ij} = \mathbf{x}_{ij}^\top \beta + \mathbf{z}_{ij}^\top b_i + \varepsilon_{ij}$, con $i = 1, \dots, n$ e $j = 1, \dots, n_i$, dove $n_1 = \dots = n_n = m$. Inoltre, $\mathbf{x}_{ij} = \mathbf{z}_{ij} = (1, j)^\top$, $\beta = (1, 2)^\top$ e $\varepsilon_{ij} \sim N(0, 1)$. Il vettore degli effetti casuali b_i è invece generato da $N(0, \mathbf{D})$, con diverse strutture per \mathbf{D} . La numerosità campionaria n e la dimensione dei $cluster$ m vengono indicate insieme alla matrice \mathbf{D} nella tabella seguente. Si valuta la nullità della matrice \mathbf{D} e quindi la significatività di entrambi gli effetti casuali.

\mathbf{D}	n	m	FLC	LRT	$PBLRT$	$Lin.score$
$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$	10	3	4.6(< 0.01)	1.6(0.04)	4.8(39.36)	4.8(1.37)
		5	4.8(< 0.01)	0.8(0.04)	5.2(39.67)	4.2(1.51)
	15	3	5.0(< 0.01)	1.6(0.04)	3.6(40.91)	4.8(1.49)
		5	5.4(< 0.01)	1.2(0.04)	4.6(41.82)	5.4(1.82)
$\begin{pmatrix} 0.05 & 0.02 \\ 0.02 & 0.05 \end{pmatrix}$	10	3	15.6(< 0.01)	12.8(0.04)	22.6(39.18)	32.0(1.37)
		5	75.0(< 0.01)	64.4(0.05)	79.8(39.66)	85.0(1.51)
	15	3	24.2(< 0.01)	23.0(0.05)	33.8(40.91)	43.8(1.48)
		5	89.2(< 0.01)	83.8(0.06)	92.0(41.82)	93.4(1.83)
$\begin{pmatrix} 0.08 & 0.02 \\ 0.02 & 0.08 \end{pmatrix}$	10	3	28.4(< 0.01)	23.6(0.04)	35.0(39.33)	46.0(1.37)
		5	88.0(< 0.01)	81.4(0.05)	89.6(39.70)	92.2(1.51)
	15	3	33.6(< 0.01)	30.0(0.05)	44.4(40.57)	55.8(1.48)
		5	96.6(< 0.01)	94.4(0.06)	97.6(42.07)	98.2(1.83)
$\begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{pmatrix}$	10	3	35.6(< 0.01)	35.0(0.04)	48.0(39.34)	59.8(1.38)
		5	93.4(< 0.01)	91.2(0.05)	95.6(39.69)	97.2(1.51)
	15	3	55.7(< 0.01)	52.8(0.05)	65.4(40.56)	74.2(1.48)
		5	98.6(< 0.01)	97.4(0.05)	99.8(42.39)	99.8(1.84)

TABELLA 4.1

D	<i>n</i>	<i>m</i>	<i>FLC</i>	<i>power</i>
$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$	10	3	4.6(< 0.01)	5
		5	5.2(< 0.01)	5
	15	3	5.5(< 0.01)	5
		5	5.5(< 0.01)	5
$\begin{pmatrix} 0.05 & 0.02 \\ 0.02 & 0.05 \end{pmatrix}$	10	3	17.7(< 0.01)	17.7
		5	74.0(< 0.01)	74.6
	15	3	22.6(< 0.01)	22.3
		5	88.4(< 0.01)	88.2
$\begin{pmatrix} 0.08 & 0.02 \\ 0.02 & 0.08 \end{pmatrix}$	10	3	24.4(< 0.01)	24.1
		5	88.6(< 0.01)	88.2
	15	3	33.4(< 0.01)	32.9
		5	96.6(< 0.01)	96.4
$\begin{pmatrix} 0.1 & 0.05 \\ 0.05 & 0.1 \end{pmatrix}$	10	3	33.9(< 0.01)	33.9
		5	94.2(< 0.01)	94.4
	15	3	47.2(< 0.01)	47.1
		5	98.8(< 0.01)	98.9

TABELLA 4.2

La Tabella 4.1 mostra, come anticipato, che il log-rapporto di verosimiglianza standard si dimostra essere un test estremamente conservativo. D'altra parte, tutti gli altri test si avvicinano sotto l'ipotesi nulla al livello di significatività nominale. Il test *FLC* si dimostra essere poco competitivo rispetto al test *PBLRT* e *Lin.score* con $m = 3$, in quanto ci sono poche osservazioni per stimare efficacemente 2 effetti casuali. Con $m = 5$ invece il test si dimostra essere una valida alternativa, soprattutto considerando la velocità di calcolo. In questa situazione il *LRT* si comporta in modo simile al test *FLC* in termini di potenza. È importante notare (Tabella 4.2) come la potenza calcolata con la formula (3.5) e metodo Monte-Carlo abbia valori simili alla potenza empirica osservata. I valori si allontanano dalla potenza empirica per valori percentuali compresi tra -0.8 e $+0.8$, che con $n_{sim} = 10000$ corrisponde in pratica all'intervallo di confidenza al 95% per la media di una variabile binomiale.

4.1.2 Studio di simulazione 2

Anche in questo caso la struttura per generare dati longitudinali è presa da Drikvandi et al. (2013). Si ha un modello lineare misto con gruppi indipendenti con 2 effetti fissi e 3 effetti casuali. Sia \mathbf{x}_{ij} che \mathbf{z}_{ij} hanno come primo elemento il valore 1 per l'intercetta fissa e casuale, mentre i restanti sono stati generati da variabili uniformi indipendenti

nell'intervallo $[0, 1]$. I coefficienti fissi sono pari a $\beta = (1, 2)^\top$ e gli effetti casuali sono generati da $b_i \sim N_3(0, \mathbf{D})$, con diverse strutture per \mathbf{D} . La tabella è costruita in modo simile alla prima configurazione, in questo caso si fissa per tutti i casi $m = 10$. In questo caso si valuta l'eliminazione degli ultimi 2 effetti casuali, mantenendo però l'intercetta.

\mathbf{D}	n	FLC	LRT	$PBLRT$	$Lin.score$
$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	7	4.8(< 0.01)	1.5(0.03)	5(38.96)	5.2(0.07)
	15	5.4(< 0.01)	1.2(0.04)	5.8(41.87)	6.0(0.29)
	25	5.8(< 0.01)	1.(0.04)	6.0(47.52)	5.4(0.91)
	50	5.2(< 0.01)	2.1(0.06)	4.0(61.52)	3.8(5.66)
$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.2 & 0.1 \\ 0 & 0.1 & 0.2 \end{pmatrix}$	7	8.7(< 0.01)	3.1(0.03)	6.8(39.48)	9.6(0.08)
	15	12.8(< 0.01)	5.1(0.06)	11.8(43.38)	14.2(0.29)
	25	15.6(< 0.01)	6.5(0.07)	13.6(48.94)	17.2(0.90)
	50	22.6(< 0.01)	14.4(0.06)	18.4(65.77)	23.0(5.66)
$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.5 & 0.1 \\ 0 & 0.1 & 0.5 \end{pmatrix}$	7	20.8(< 0.01)	9.6(0.05)	18.0(39.74)	21.4(0.08)
	15	32.4(< 0.01)	17.6(0.06)	28.0(44.19)	33.6(0.28)
	25	43.4(< 0.01)	30.6(0.05)	37.8(49.51)	47.4(0.86)
	50	63.8(< 0.01)	56.1(0.07)	56.2(67.93)	66.4(5.49)
$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.2 \\ 0 & 0.2 & 1 \end{pmatrix}$	7	36.6(< 0.01)	26.2(0.03)	33.0(39.23)	38.4(0.07)
	15	64.4(< 0.01)	51.5(0.06)	58.2(44.28)	64.8(0.27)
	25	79.8(< 0.01)	74.2(0.07)	76.0(50.13)	81.0(0.84)
	50	96.2(< 0.01)	95.2(0.06)	92.4(66.91)	96.2(5.23)

TABELLA 4.3

\mathbf{D}	n	FLC	$power$
$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	7	5.2(< 0.01)	5
	15	5.2(< 0.01)	5
	25	4.6(< 0.01)	5
	50	5.2(< 0.01)	5
$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.2 & 0.1 \\ 0 & 0.1 & 0.2 \end{pmatrix}$	7	10.2(< 0.01)	9.2
	15	12.7(< 0.01)	13
	25	16.2(< 0.01)	14.5
	50	22.4(< 0.01)	22.8
$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0.5 & 0.1 \\ 0 & 0.1 & 0.5 \end{pmatrix}$	7	19.3(< 0.01)	22.0
	15	29.6(< 0.01)	31.1
	25	41.6(< 0.01)	43.9
	50	64.0(< 0.01)	65.0
$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.2 \\ 0 & 0.2 & 1 \end{pmatrix}$	7	38.8(< 0.01)	36.9
	15	61.9(< 0.01)	57.8
	25	78.7(< 0.01)	78.7
	50	96.2(< 0.01)	97.4

TABELLA 4.4

Anche in questo caso (Tabella 4.3) tutti i metodi tranne il test LRT si avvicinano sotto H_0 al livello di significatività nominale. Inoltre il test FLC ha comportamento simile a quello dei metodi computazionali come $PBLRT$ e $Lin.score$. Si noti come la velocità di calcolo rimanga praticamente invariata per il test FLC e si dimostri un metodo estremamente rapido.

4.1.3 Studio di simulazione 3

Nel terzo studio il modello viene ulteriormente ampliato. In questo caso, il modello lineare misto ha 8 effetti fissi e 4 effetti casuali, dove \mathbf{x}_{ij} comprende l'intercetta e i restanti 7 elementi sono generati da una normale multivariata con vettore delle medie pari a 0 e $Cov(x_{iju}, x_{ijv}) = 0.5^{|u-v|}$ per $u, v = 2, \dots, 8$. Il vettore \mathbf{z}_{ij} è posto pari ai primi 4 elementi di \mathbf{x}_{ij} . Inoltre $\varepsilon_{ij} \sim N(0, 1)$, $\beta = (1, -1, 1, \dots, -1)^\top$. Gli effetti casuali b_i si distribuiscono come $b_i \sim N_4(0, \mathbf{D})$, con \mathbf{D} matrice 4×4 , dove la sottomatrice 2×2 in alto a sinistra è generata da una matrice casuale Wishart con 3 gradi di libertà e matrice di scala diagonale con entrambi gli elementi pari a 0.5. La sottomatrice 2×2 in basso a destra ha invece sulla diagonale τ e fuori dalla diagonale $\tau/2$, con $\tau \geq 0$ parametro definito sotto. Per garantire positività della matrice \mathbf{D} le restanti celle sono

poste pari a $\tau/2$. In questo caso valutiamo la significatività degli ultimi due effetti casuali, mantenendo gli altri nel modello.

τ	n	m	<i>FLC</i>	<i>LRT</i>	<i>PBLRT</i>	<i>Lin.score</i>
0	10	10	5.3(< 0.01)	1.3(0.03)	5.3(150.07)	1.9(0.32)
		20	4.2(< 0.01)	1.3(0.04)	4.6(196.95)	3.3(1.44)
	20	10	5.0(< 0.01)	1.2(0.05)	4.3(193.42)	2.3(1.38)
		20	5.6(< 0.01)	1.2(0.04)	5.0(298.20)	3.0(9.54)
	40	10	4.7(< 0.01)	1.8(0.04)	4.4(294.68)	3.5(8.92)
		20	5.5(< 0.01)	1.1(0.04)	5.8(502.53)	2.7(64.03)
0.01	10	10	45.9(< 0.01)	30.2(0.04)	41.5(143.46)	48.1(0.27)
		20	85.2(< 0.01)	73.0(0.04)	79.1(194.38)	85.0(1.27)
	20	10	74.0(< 0.01)	58.4(0.04)	69.0(191.80)	76.8(1.19)
		20	98.9(< 0.01)	94.4(0.04)	96.5(299.36)	98.7(8.09)
	40	10	94.5(< 0.01)	87.9(0.04)	93.1(297.45)	96.8(7.69)
		20	100(< 0.01)	100(0.04)	100(521.04)	100(52.26)
0.02	10	10	75.1(< 0.01)	63.1(0.05)	72.6(142.81)	76.1(0.26)
		20	98.5(< 0.01)	95.1(0.04)	96.1(191.82)	98.2(1.23)
	20	10	95.8(< 0.01)	91.5(0.06)	94.9(189.24)	96.5(1.15)
		20	100(< 0.01)	99.7(0.04)	100(295.74)	100(8.06)
	40	10	99.8(< 0.01)	99.8(0.04)	99.3(295.85)	99.8(7.48)
		20	100(< 0.01)	100(0.07)	100(513.40)	100(52.52)

TABELLA 4.5

τ	n	m	FLC	power
0	10	10	5.2(< 0.01)	5
		20	5.2(< 0.01)	5
	20	10	4.5(< 0.01)	5
		20	5.2(< 0.01)	5
	40	10	4.8(< 0.01)	5
		20	5.1(< 0.01)	5
0.01	10	10	44.8(< 0.01)	41.7
		20	85.9(< 0.01)	83.6
	20	10	75.2(< 0.01)	82.2
		20	98.4(< 0.01)	99.4
	40	10	92.4(< 0.01)	94.4
		20	99.9(< 0.01)	99.9
0.02	10	10	74.0(< 0.01)	78.5
		20	98.(< 0.01)	97.2
	20	10	94.9(< 0.01)	97.6
		20	100(< 0.01)	99.9
	40	10	99.8(< 0.01)	99.9
		20	100(< 0.01)	100

TABELLA 4.6

Il test LRT (Tabella 4.5) continua ad essere un test estremamente conservativo, tuttavia si nota come anche il test Linear Score per questa struttura di simulazione non raggiunga il livello nominale di significatività, pur mantenendo potenza paragonabile al test FLC . Con $\tau \neq 0$ e numerosità elevate i 4 test hanno performance ottime, tuttavia il tempo di calcolo varia di molto tra i test. Il test FLC è il più veloce e la velocità di calcolo non è influenzata da N . Al contrario, un test di simulazione come $PBLRT$, con $n = 40$ e $m = 20$, richiede mediamente più di 8 minuti di elaborazione.

4.2 Curve di potenza

Nel paragrafo 4.1 abbiamo mostrato il comportamento del test FLC in situazioni diverse. Si evince anche come sia possibile calcolare una stima della potenza del test FLC , senza dover ricorrere a simulazioni. In questo capitolo verranno quindi costruite le curve di potenza del test, nei tre scenari considerati nel paragrafo precedente, al variare della matrice di covarianza degli effetti casuali. I grafici sono costruiti in R utilizzando il pacchetto `ggplot2` (Wickham (2016)).

Per i modelli considerati nel primo studio, vengono costruite le curve di potenza del test, nelle 4 configurazioni viste in precedenza, ovvero $(n = 10, m = 3)$, $(n = 10, m = 5)$, $(n = 15, m = 3)$ e $(n = 15, m = 5)$. La matrice \mathbf{D} ha struttura $\mathbf{D} = \begin{pmatrix} \sigma_b^2 & \rho \\ \rho & \sigma_b^2 \end{pmatrix}$. Per ogni valore di $\rho = (0, 0.1, 0.25, 0.5)$, σ_b^2 assume valori nell'intervallo $(\rho, 1]$, per i quali viene calcolata la funzione di potenza al variare di σ_b^2 , per ognuna delle quattro correlazioni previste.

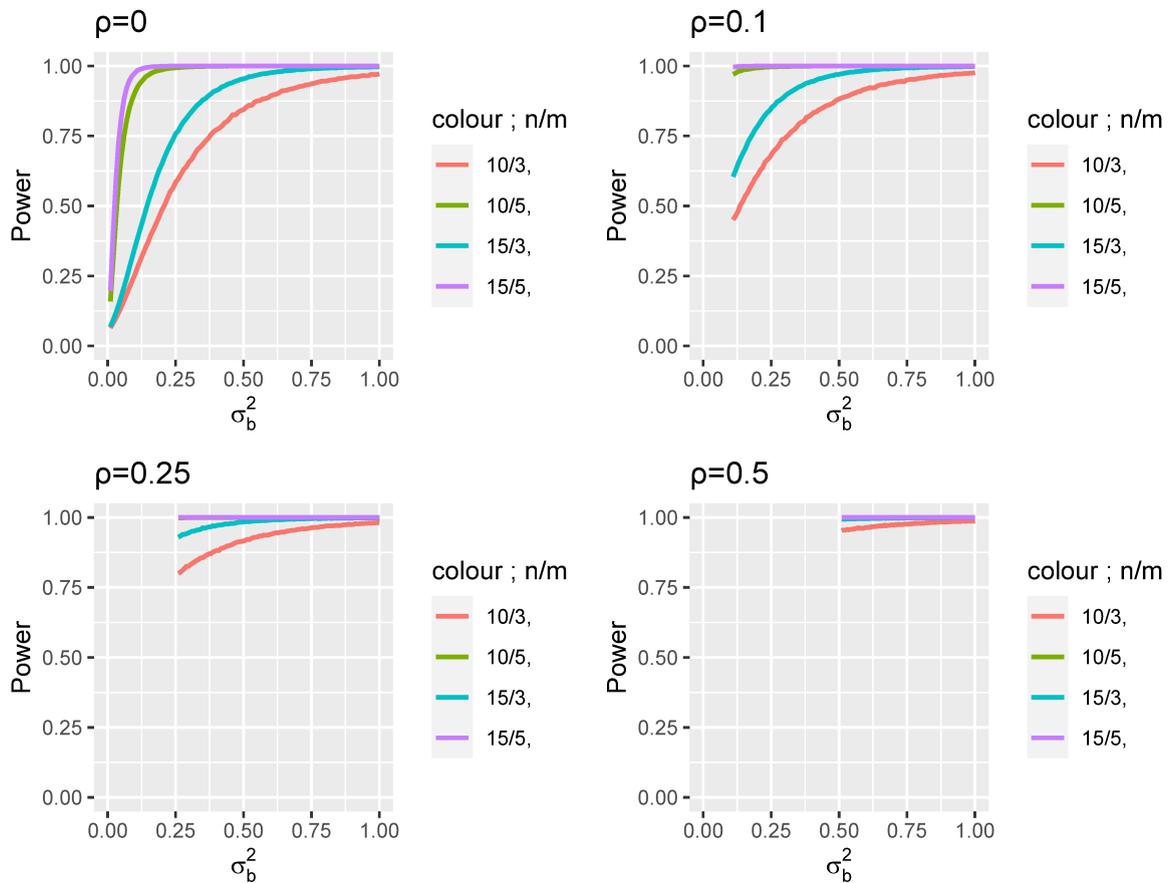


FIGURA 4.1: Studio di simulazione 1

Come già osservato quando $m = 5$ la potenza tende rapidamente a 1, nello specifico per $\sigma_b^2 \geq 0.15$ ($\sigma_b \geq 0.387$), indipendentemente da ρ . Per $m = 3$, la potenza tende a 1 per valori di σ_b^2 vicini a 1. Come già anticipato la potenza del test è bassa quando il numero di effetti casuali q_c , in questo caso 2, si avvicina alla numerosità dei cluster. In tutti e quattro i grafici, l'aumento di 2 osservazioni per cluster comporta un aumento più che doppio in termini di potenza, quando $\sigma_b^2 < 0.15$. La correlazione non sembra influenzare in modo determinante la potenza quando $m = 5$. Quando invece la numerosità di cluster diminuisce, l'aumento della correlazione, comporta un aumento della potenza, ci si allontana infatti sempre di più dall'ipotesi nulla $\mathbf{D} = 0$.

Le curve di potenza per la configurazione dello studio 2 vengono costruite con lo stesso

schema dello studio 1. In questo caso, $\mathbf{D} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sigma_b^2 & \rho \\ 0 & \rho & \sigma_b^2 \end{pmatrix}$, dove ρ e σ_b^2 assumono i medesimi valori visti sopra. La numerosità di *cluster* è fissata pari $m = 10$, mentre il numero di *cluster* assume i seguenti valori (7, 15, 25, 50).

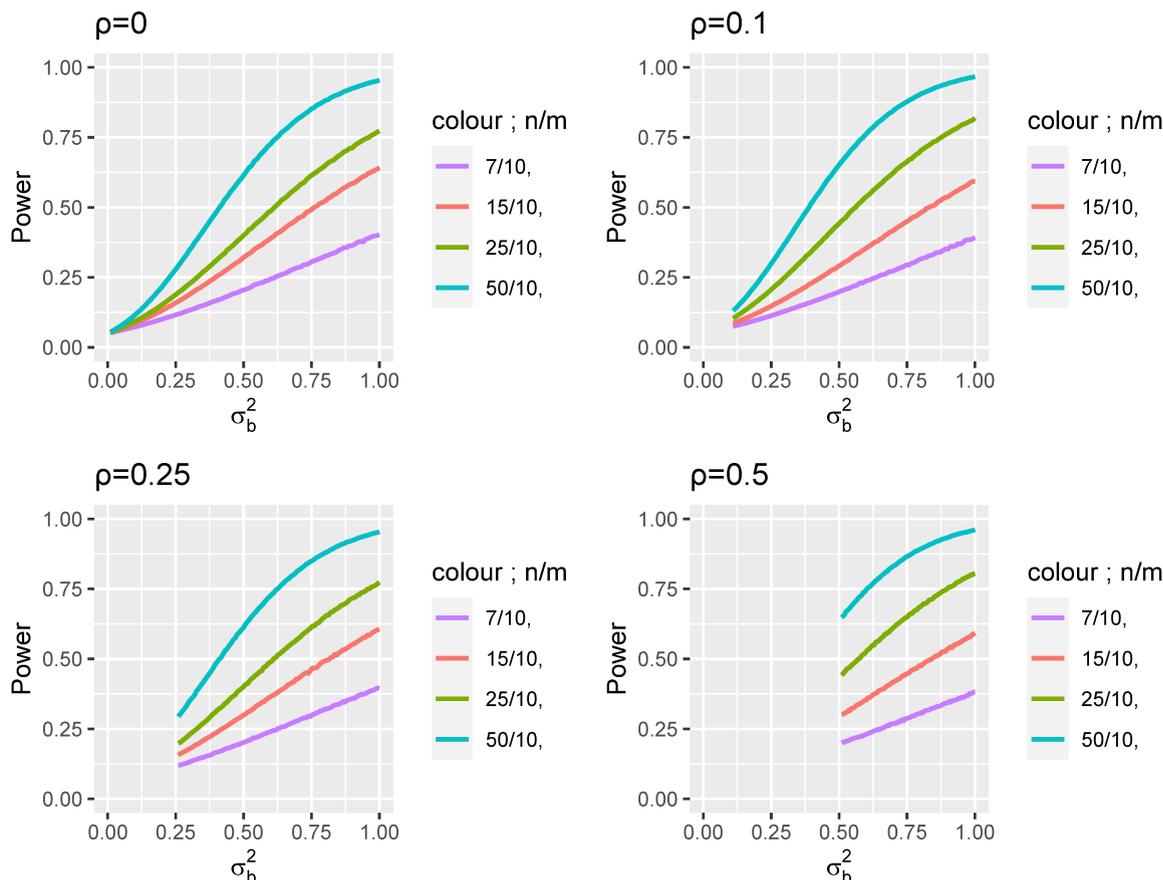


FIGURA 4.2: Studio di simulazione 2

La potenza si avvicina a 1, indipendentemente dalla correlazione tra effetti casuali, solo quando $n = 50$. Le variazioni in correlazione non mostrano differenze per quel che riguarda le curve. In questo caso sono presenti 10/3 osservazioni per effetto casuale, condizione comparabile a quella vista nel grafico precedente con $m = 5$ e $q_c = 2$. Tuttavia a differenza del primo grafico, la potenza rimane decisamente più contenuta. Ciò è dovuto al fatto che il modello nullo nello studio 1 non presenta effetti casuali, la correlazione presente entro i *cluster* non viene tenuta in considerazione. Nello studio 2, il modello nullo comprende invece intercetta casuale, in pratica assumendo equicorrelazione tra osservazioni appartenenti allo stesso *cluster*. Nel primo studio quindi il modello nullo non è in alcun modo in grado di descrivere i dati, mentre nel secondo studio la presenza di equicorrelazione sembra essere in grado di modellare, seppur limitatamente, i dati simulati.

Per quel che riguarda la terza configurazione la struttura della matrice \mathbf{D} rimane inalterata, con τ che “scorre” nell’intervallo $[0, 1]$. Il numero di *cluster*, e le numerosità degli stessi, sono le stesse viste nel paragrafo 4.1.3, ovvero $(n = 10, m = 10)$, $(n = 10, m = 20)$, $(n = 20, m = 10)$, $(n = 20, m = 20)$, $(n = 40, m = 10)$, $(n = 40, m = 20)$.

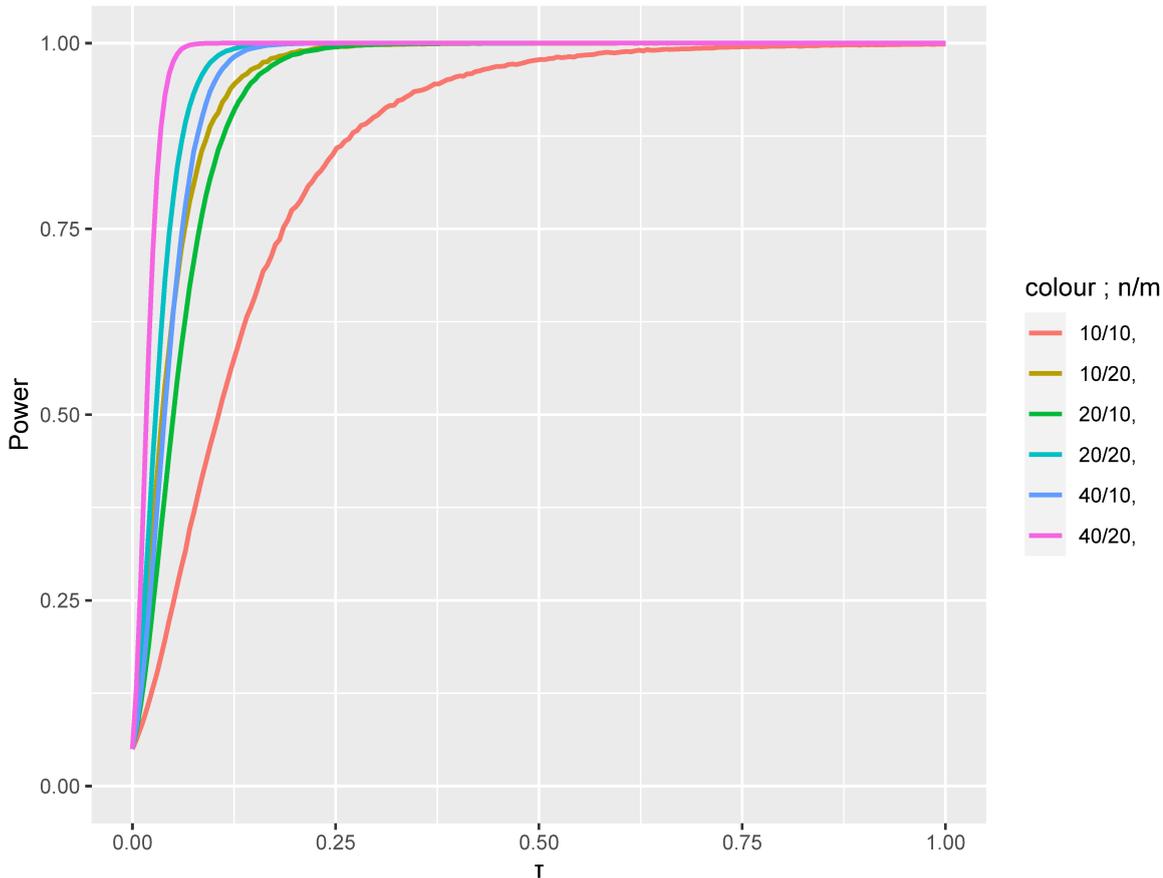


FIGURA 4.3: Studio di simulazione 3

In questo caso si nota immediatamente l’effetto che ha la numerosità di *cluster* sulla potenza del test, effetto che era già stato anticipato nel Capitolo 3. Il numero di effetti casuali è $q_c = 4$ e ciò significa che con $m = 10$ si hanno 2.5 osservazioni per stimare ogni effetto casuale. Con $m = 20$ si hanno 5 osservazioni per effetto casuale. Ricordiamo che gli autori indicano come valore limite dalle 5 alle 8 osservazioni per effetto casuale, per far sì che il test *FLC* abbia comportamento in termini di potenza paragonabile a metodi computazionali visti nello scorso paragrafo. Le tre curve “inferiori”, sono quelle con *cluster* più piccoli. È interessante notare come la curva con $n = 10$ e $m = 20$ sia sovrapponibile con la curva per $n = 40$ e $m = 10$. In pratica otteniamo con un campione totale di 200 osservazioni valori della potenza del test simili a quelli che otterremmo con uno di 400.

Il comportamento delle curve di potenza mostrate, conferma ciò che gli Autori preventivavano nell'articolo. I valori risultanti dalla formula (3.5) sono calcolati col metodo Monte-Carlo utilizzando 10000 repliche. Nella prima configurazione la differenza in valore assoluto tra potenza empirica e potenza approssimata (*power*) è sempre inferiore al punto percentuale. Tuttavia questi scostamenti aumentano in disegni più complessi come lo studio 2 e 3. Per problemi più complicati potrebbe essere necessario aumentare il numero di repliche. La possibilità di accompagnare i risultati del test *FLC*, con una stima della potenza del test, attraverso una stima di \mathbf{D} sotto l'ipotesi alternativa, è sicuramente un valore aggiunto nell'applicazione pratica del test. Possibilità che verrà esplorata nel prossimo capitolo con un caso pratico.

Capitolo 5

Applicazione ai dati Stroke

Descrizione dei dati

L'applicazione pratica del test *FLC* riguarda il *dataset Stroke* (Dobson & Barnett, 2018, paragrafo 11.2), contenuto nel pacchetto R *MLGdata* (Salvan, Sartori & Pace (2020a)). I dati derivano da una sperimentazione clinica per confrontare diverse terapie di riabilitazione post infarto. Il campione è composto da 24 pazienti ripartiti nelle 3 terapie, con 8 soggetti per ogni gruppo. I tre gruppi, codificati con le lettere A, B e C sono rispettivamente: il gruppo sottoposto a terapia innovativa, il gruppo sottoposto a terapia tradizionale nello stesso ospedale del gruppo A e il gruppo sottoposto a terapia tradizionale in un ospedale diverso. Si rileva per ogni paziente l'abilità funzionale attraverso l'indice di Barthel, con valori da 0 a 100, dove 100 indica autonomia nello svolgimento delle mansioni e 0 completa dipendenza da chi assiste. Ogni paziente viene valutato settimanalmente per un totale di 8 settimane. L'obiettivo dello studio era quello di verificare se il trattamento A fosse superiore agli altri due.

5.1 Risultati

Consideriamo un modello lineare misto per l'indice di Barthel, con coefficienti fissi relativi all'intercetta, il gruppo d'appartenenza, la settimana della misurazione e l'interazione fra le due. Il modello comprende poi un'intercetta casuale e un effetto casuale relativo alla settimana di rilevazione. Adattiamo il modello utilizzando la funzione `lmer` contenuta nel modello `lme4`. Valutiamo per prima cosa l'assunzione di normalità condizionata ai valori degli effetti casuali. Si veda il *normal QQ-plot*, calcolato sui residui condizionati (*conditional residuals*) $\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\mathbf{b}}$.

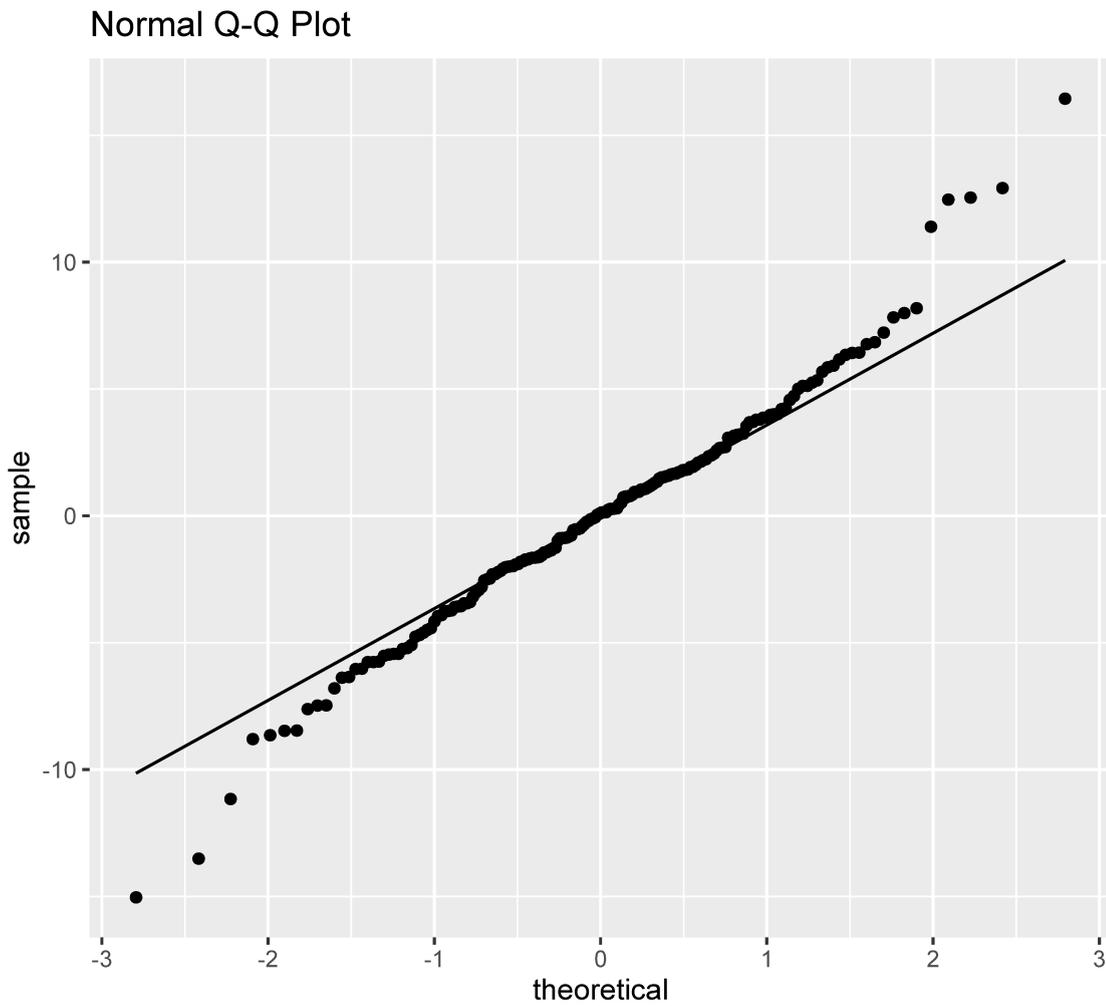


FIGURA 5.1

Sembrano esserci degli allontanamenti dalla distribuzione normale, il test Shapiro-Wilk rifiuta l'ipotesi nulla con livello di significatività 5%, ma con $\alpha = 1\%$ non rifiuta la normalità ($p\text{-value} = 0.013$). Seppur con qualche riserva, possiamo considerare valida l'ipotesi distributiva. Interessante valutare poi la distribuzione degli effetti casuali valutando le stime **BLUP**.

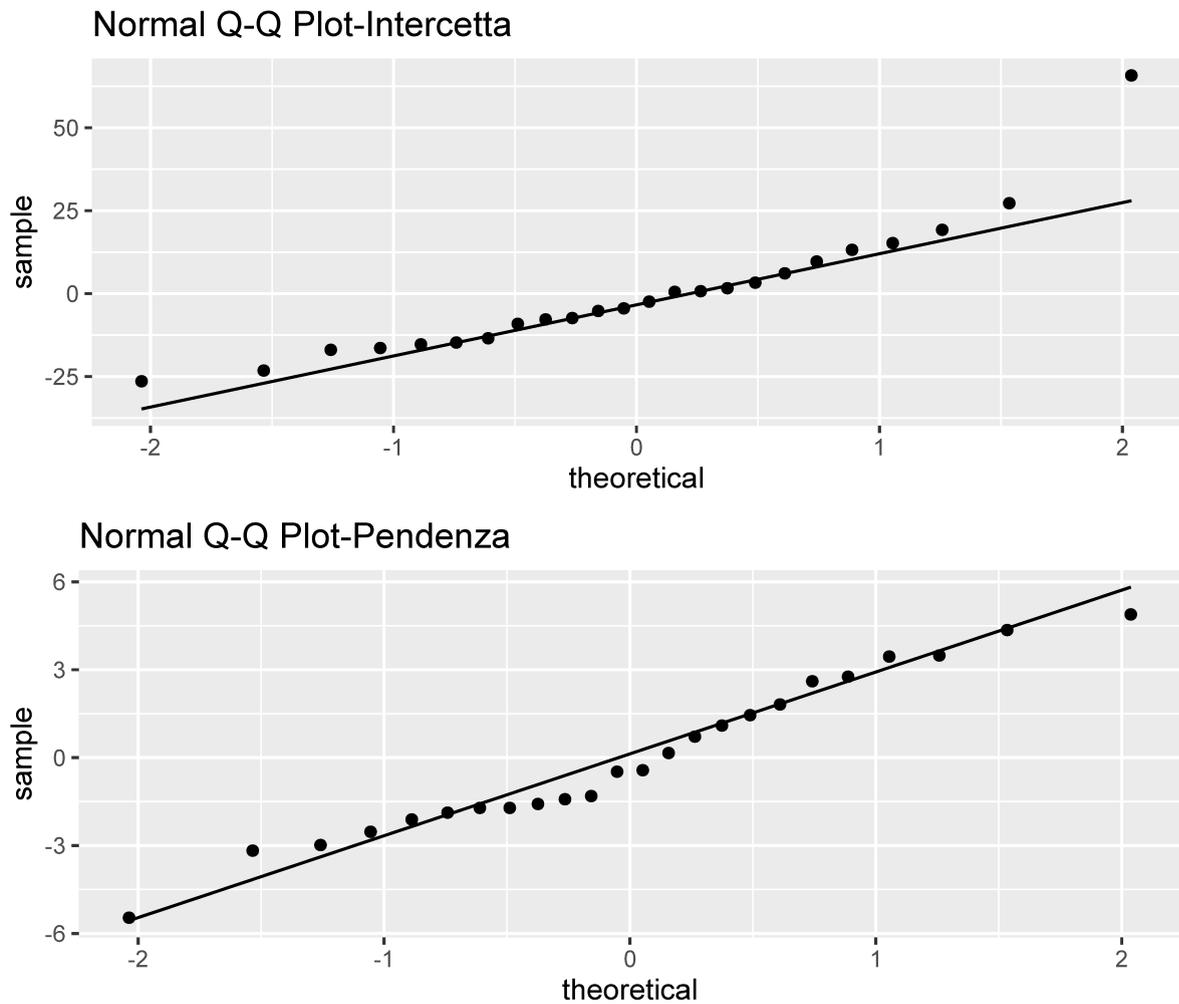


FIGURA 5.2

La distribuzione empirica delle intercette casuali è sicuramente non normale ($p\text{-value} < 0.006$). Rimuovendo il soggetto numero 5 il test di normalità per le intercette non rifiuta l'ipotesi nulla ($p\text{-value} > 0.5$). Tuttavia per l'analisi seguente l'unità statistica numero 5 non viene esclusa dal *dataset*. La distribuzione normale sembra essere coerente con la distribuzione empirica del secondo effetto casuale ($p\text{-value} > 0.5$). Quest'ultimo aspetto diventa importante successivamente nel calcolo della potenza del test. Valutiamo infatti il secondo effetto significativo, utilizzando il test *FLC* ($p\text{-value} < 0.001$). L'intervallo di confidenza al 95% per la varianza della pendenza casuale, calcolato con la funzione `confint`, è $[4.04, 13.98]$. Concludiamo quindi che il modello debba includere pendenza casuale e di conseguenza esista un effetto individuale, nella velocità di recupero da paziente a paziente. Medesimi risultati si ottengono con il test *LRT* ($p\text{-value} < 0.001$), *PBLRT* ($p\text{-value} < 0.003$) e *RLRT* ($p\text{-value} < 0.001$). Il problema trattato è relativamente poco complesso, il tempo di calcolo è stato inferiore al secondo per tutti i test, tranne che per *PBLRT*, che ha richiesto comunque 26 secondi. Possiamo utilizzare

la stima di \mathbf{D} ottenuta sotto l'ipotesi alternativa, per approssimare la potenza con la funzione (3.5), sotto l'assunzione che l'effetto valutato abbia distribuzione normale. In questo caso, la funzione restituisce un valore pari a 1, infatti la varianza dell'effetto sembra avere valori molto elevati. Inoltre, il numero di osservazioni per paziente $m = 8$, con $q_c = 2$ assicura buone stime dei coefficienti casuali. Il valore calcolato della potenza del test è quindi coerente con ciò che è stato visto nel capitolo precedente.

Conclusione

In questa relazione viene affrontato il problema legato alla valutazione della significatività degli effetti casuali in modelli normali misti. In primo luogo viene presentata una rassegna dei metodi esistenti. Ci si è concentrati poi sul test FLC , presentato da Hui et al. (2019) come una valida alternativa ad approcci computazionali. Lo studio di simulazione proposto riproduce ciò che è stato fatto dagli Autori, con l'aggiunta della potenza calcolata con la formula (3.5). Inoltre, per gli studi visti nel Paragrafo 4.1 vengono riportate le curve di potenza del test al variare della matrice di covarianza degli effetti casuali \mathbf{D} . Si è voluto poi presentare un semplice caso pratico in cui il test FLC mostra al meglio le sue potenzialità.

Concludiamo che il test FLC è sicuramente un valido strumento per risolvere il problema in esame. Come suggerito da Hui et al. (2019), esso può quantomeno essere utilizzato come metodo preliminare per valutare l'ipotesi nulla, considerando soprattutto l'esattezza della distribuzione nulla F e la velocità di calcolo. Indipendentemente dalla numerosità m e dalla struttura della matrice di covarianza degli effetti casuali, se FLC rifiuta H_0 si può ritenere che vi sia sufficiente evidenza empirica riguardo la significatività degli effetti casuali. In caso contrario, metodi più onerosi computazionalmente come $PBLRT$ potrebbero essere più idonei (potenza maggiore) nel rilevare effetti casuali diversi da zero. Come visto invece quando m è sufficientemente elevato (m/q_e vicino a 5 o superiore), allora il test FLC ha potenza paragonabile a quella di metodi computazionali.

La Tabella 2.1 mostra poi che per la maggior parte dei metodi presentati in questa relazione non vi è una libreria di riferimento in \mathbf{R} , o in altri *software*. È necessario che questi metodi, così come il test FLC , abbiano un'implementazione accessibile, preferibilmente supportando modelli stimati sia con il pacchetto `nlme` (Pinheiro & Bates (2023)) che con `lme4` (Bates et al. (2014)).

Appendice

A. Prodotto di Kronecker

Sia \mathbf{A} una matrice $m \times n$, con generico elemento a_{ij} $i = 1, \dots, m$ $j = 1, \dots, n$ e sia \mathbf{B} , con generico elemento b_{hk} $h = 1, \dots, p$ $k = 1, \dots, q$ una matrice $p \times q$, allora il prodotto di Kronecker $\mathbf{A} \otimes \mathbf{B}$ è la matrice a blocchi $pm \times qn$:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \dots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \dots & a_{mn}\mathbf{B} \end{bmatrix}$$

B. Matrici di proiezione generate da matrici con colonne non linearmente indipendenti

Sia \mathbf{Y} un vettore casuale n -dimensionale tale che

$$\begin{aligned} E(\mathbf{Y}) &= \mathbf{X}\beta \\ \text{Var}(\mathbf{Y}) &= \sigma^2 I_n \end{aligned}$$

dove \mathbf{X} è una matrice $n \times p$ con $n > p$, β vettore p -dimensionale e $\sigma^2 > 0$. Un modello con queste caratteristiche è detto **modello lineare con ipotesi del secondo ordine**. Le **equazioni normali** sono

$$\mathbf{X}^\top \mathbf{X} \hat{\beta} = \mathbf{X}^\top \mathbf{Y}$$

ottenute differenziando

$$(\mathbf{Y} - \mathbf{X}\hat{\beta})^\top (\mathbf{Y} - \mathbf{X}\hat{\beta}),$$

ovvero applicando i **minimi quadrati ordinari** al modello lineare di cui sopra. Le equazioni normali ammettono sempre almeno una soluzione, che è unica se, e solo se, $\mathbf{X}^\top \mathbf{X}$ è invertibile.

Sia \mathbf{Q} un vettore p -dimensionale, si definisce $\mathbf{Q}^\top \hat{\beta}$, **stimatore dei minimi quadrati** di una qualsiasi funzione parametrica $\mathbf{Q}^\top \beta$, dove $\hat{\beta}$ è una soluzione delle equazioni normali. $\mathbf{Q}^\top \hat{\beta}$ è stimabile solo se $\mathbf{Q}^\top \hat{\beta}$ è unico per ogni soluzione $\hat{\beta}$ delle equazioni normali, ovvero $\mathbf{Q} \in \mathcal{M}(\mathbf{X}^\top)$, dove con $\mathcal{M}(\mathbf{X}^\top)$ si indica lo spazio vettoriale generato dalle righe di \mathbf{X} e vale $\mathcal{M}(\mathbf{X}^\top) = \mathcal{M}(\mathbf{X}^\top \mathbf{X})$. La condizione precedente equivale all'esistenza di una qualche funzione lineare di \mathbf{Y} con valore atteso $\mathbf{Q}^\top \beta$. Inoltre se $\mathbf{Q}^\top \hat{\beta}$ è stimabile segue che è anche lo stimatore lineare non distorto con varianza minima (**BLUE**, *best linear unbiased estimator*) si veda Rao (1973, capitolo 4). Si dimostra inoltre che tutte le funzioni lineari parametriche sono stimabili se e solo se $rk(\mathbf{X}) = p$ (equivalentemente $rk(\mathbf{X}^\top \mathbf{X}) = p$), dove p è il numero di parametri da stimare β_j . Le equazioni normali hanno quindi un'unica soluzione ($\mathbf{X}^\top \mathbf{X}$ è invertibile) e $\mathbf{Q}^\top \hat{\beta}$ è unico per ogni \mathbf{Q} . Se $rk(\mathbf{X}) < p$, alcune funzioni parametriche non ammettono stimatori non distorti, che equivale in questo caso alla loro non stimabilità.

Introduciamo quindi l'**inversa generalizzata** (*g-inverse*) di una matrice A $n \times m$, indicata con A^- , tale che $X = A^-Y$ è soluzione dell'equazione $AX = Y$, per ogni $Y \in \mathcal{M}(A)$. La matrice A^- possiede le seguenti proprietà:

1. $AA^-A = A$;
2. sia $H = A^-A$, allora H è idempotente;
3. $AH = A$ e $rk(A) = rk(H) = tr(H)$;
4. una soluzione generale di $AX = 0$ è $(H - I_n)Z$ con Z vettore arbitrario n -dimensionale;
5. una soluzione generale di un'equazione consistente $AX = Y$ è $A^-Y + (H - I_n)Z$, con Z arbitrario;
6. $Q^\top X$ ha un unico valore per ogni X che soddisfa $AX = Y$, se e solo se $Q^\top H = Q^\top$.

Di conseguenza una soluzione generale delle equazioni normali è $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y} + (I_n - H)Z$ con $H = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}$ e Z vettore arbitrario p -dimensionale. Si dimostra che la condizione necessaria e sufficiente di stimabilità per $\mathbf{Q}^\top \beta$ è che $\mathbf{Q}^\top (I_n - H) = 0$.

Una soluzione delle equazioni normali $\mathbf{X}^\top \mathbf{X} \hat{\beta} = \mathbf{X}^\top \mathbf{Y}$ è $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$, ottenendo la corrispondente matrice di proiezione $\mathbf{P} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$.

C. Pseudo-inversa di Moore-Penrose

Sia $A \in \mathbb{R}^{n \times m}$ una matrice con $rk(A) \leq \min\{n, m\}$, utilizzando la **decomposizione ai valori singolari**, si ha

$$A = UDV^\top,$$

dove $U \in \mathbb{R}^{n \times n}$, $D \in \mathbb{R}^{n \times m}$, $V \in \mathbb{R}^{m \times m}$. Le matrici U e V sono matrici unitarie e in generale non uniche. Mentre D è una matrice rettangolare diagonale unica, che contiene sulla diagonale tutti i valori singolari della matrice A , ordinati in ordine crescente. Segue che la matrice pseudo-inversa è

$$A^- = VD^-U^\top,$$

con $D^- \in \mathbb{R}^{m \times n}$ (pseudo-inversa di D) può essere calcolata prendendo la trasposta di D e sostituendo i valori singolari non nulli col loro reciproco. La matrice A^- , ha dimensione $m \times n$ e soddisfa le seguenti proprietà:

$$\begin{aligned} AA^-A &= A \\ A^\top AA^\top &= A^- \\ (AA^-)^\top &= AA^- \\ (A^-A)^\top &= A^-A, \end{aligned}$$

che possono essere dimostrate per calcolo diretto. Inoltre la matrice pseudo-inversa di A^- è la matrice A , ergo $(A^-)^- = A$. Se poi A è quadrata con rango pieno la sua pseudo-inversa coincide con la matrice inversa standard, $A^- = A^{-1}$.

D. Distribuzione F non centrale

La distribuzione F non centrale è una distribuzione di probabilità continua che generalizza la distribuzione F standard. Definiamo per prima cosa la variabile χ^2 non centrale. Siano X_1, \dots, X_k variabili normali indipendenti con media rispettivamente μ_i $i = 1, \dots, k$ e varianza unitaria, allora

$$\sum_{i=1}^k X_i^2 \sim \chi_{k,\lambda}^2,$$

dove $\chi_{k,\lambda}^2$ indica una distribuzione χ^2 non centrale con gradi di libertà pari a k e parametro di non centralità $\lambda = \sum_{i=1}^k \mu_i^2$. Con $\lambda = 0$ si ottiene la distribuzione χ_k^2 centrale.

Sia quindi X una variabile $\chi_{\nu_1, \lambda}^2$, e Y una variabile $\chi_{\nu_2}^2$, con X e Y indipendenti, allora

$$F = \frac{X/\nu_1}{Y/\nu_2},$$

ha distribuzione F non centrale con gradi di libertà pari a ν_1, ν_2 e parametro di non centralità λ . Anche in questo caso con $\lambda = 0$, si ottiene la distribuzione F centrale. La funzione di ripartizione di una variabile F non centrale è

$$\mathbf{F}(x|\nu_1, \nu_2, \lambda) = \sum_{j=0}^{\infty} \left(\frac{(\frac{1}{2}\lambda)^j}{j!} e^{-\lambda/2} \right) \mathbf{I} \left(\frac{\nu_1 x}{\nu_1 x + \nu_2} \middle| \frac{\nu_1}{2} + j, \frac{\nu_2}{2} \right),$$

con $\mathbf{I}(x|c_1, c_2)$ **funzione beta incompleta regolarizzata**.

E. Funzione beta incompleta regolarizzata

La **funzione beta di Eulero** è data da

$$\beta(x, y) = \int_0^1 t^{x-1} (1-t)^{y-1} dt.$$

L'integrale converge solo se x e y hanno parte reale positiva e non nulla. La funzione beta può essere generalizzata, sostituendo l'integrale definito con uno indefinito. Si definisce quindi la **funzione beta incompleta** come

$$\beta(x; a, b) = \int_0^x t^{a-1} (1-t)^{b-1} dt,$$

con $x \in [0, 1]$ e a, b numeri reali. La **funzione beta incompleta regolarizzata**, è definita in termini delle precedenti come:

$$\mathbf{I}(x|a, b) = \frac{\beta(x; a, b)}{\beta(a, b)}.$$

Le funzioni sopracitate, sono legate alla distribuzione Beta. Infatti se X è una variabile casuale con distribuzione Beta, definita sull'intervallo $[0, 1]$, di parametri (a, b) , con $a, b > 0$, allora la funzione di densità di X è

$$f(x|a, b) = \frac{x^{a-1} (1-x)^{b-1}}{\beta(a, b)},$$

mentre la funzione di ripartizione è la funzione beta incompleta regolarizzata $\mathbf{I}(x|a, b)$.

F. Codice R

```

library(mvtnorm)
library(Matrix)
library(MASS)
library(lme4)
library(tidyverse)
library(ggplot2)
library(nlme)
library(MLGdata)
library(pbkrtest)
library(RLRsim)
## Dataset generation for GLMM
## Intercept must be manually included in X and Z if desired
## D must be positive definite -- defined such as the random effects
#generated work on the first xxx columns of Z
gendat.glmm <- function(n, m, X, beta, Z, D = NULL, sigma2 = 1,
  trial.size = NULL, family = gaussian(),
  upper.count = Inf, custom.b = NULL) {
  n.ran <- nrow(D)
  X <- as.matrix(X); Z <- as.matrix(Z)
  if(sum(X[,1]) != nrow(X))
    warning("Has an intercept column being included in X?")
  if(!(family$family %in% c("gaussian","poisson","binomial")))
    stop("Current version does not permit specified family. Sorry!")
  if(family$family == "binomial" & is.null(trial.size))
    stop("trial.size must be provided for binomial family. Thanks!")

  if(is.null(colnames(X))) colnames(X) <- paste("x",1:ncol(X),sep="")
  if(is.null(colnames(Z))) colnames(Z) <- paste("z",1:ncol(Z),sep="")
  if(!is.null(custom.b)) {
    cat("Custom matrix of random effects supplied. Argument for D is ignored\n")
    true.b <- custom.b
    n.ran <- ncol(custom.b)
  }
  if(is.null(custom.b)) { true.b <- rmvnorm(n, rep(0,n.ran), as.matrix(D)) }
  if(ncol(Z) > n.ran) true.b <- cbind(true.b,matrix(0,n,ncol(Z)-n.ran))
  sim.y <- matrix(NA,n,m)
  for(i in 1:n) {
    eta <- X[(i*m-m+1):(i*m),]%%beta + Z[(i*m-m+1):(i*m),]%%true.b[i,]
    if(family$family == "gaussian") sim.y[i,] <- rnorm(m,eta,sd=sqrt(sigma2))
    if(family$family == "t")
      sim.y[i,] <- eta + rt(m, df = phi)/sqrt(phi/(phi-2))
    if(family$family == "poisson") {
      sim.y[i,] <- rpois(m,lambda=family$linkinv(eta))
      while(!all(is.finite(sim.y[i,])))
        sim.y[i,] <- rpois(m,lambda=family$linkinv(eta))
      while(any(sim.y[i,] > upper.count)) {
        new.true.b <- rmvnorm(1, rep(0,n.ran), as.matrix(D))
        if(ncol(Z)>n.ran)true.b[i,]<-
          c(new.true.b,rep(0,ncol(Z)-n.ran))
        eta <- X[(i*m-m+1):(i*m),]%%beta+
          Z[(i*m-m+1):(i*m),]%%true.b[i,]
        sim.y[i,] <- rpois(m,lambda=family$linkinv(eta))
        while(!all(is.finite(sim.y[i,])))

```

```

        sim.y[i,] <- rpois(m,lambda=family$linkinv(eta))
    } }
    if(family$family == "binomial")
        sim.y[i,] <- rbinom(m,size=trial.size,p=family$linkinv(eta))
}
out <- list(y = as.vector(t(sim.y)), n = n,
m = m, X = X, Z = Z, beta = beta, b = true.b,
sigma2 = sigma2, trial.size = trial.size,
family = family$family, nonzero.fixef = which(beta!=0),
nonzero.ranef = which(colSums(true.b^2) > 0),
id = rep(1:n,each=m))
if(is.null(custom.b)) out$D <- D
Z.mat <- vector("list",n);
for(i in 1:n) { sel.i <- which(out$id==i); Z.mat[[i]] <- Z[sel.i,] }
Z.mat <- as.matrix(bdiag(Z.mat))
out$Zdiag <- Z.mat
return(out)
}
#y = sim.y$y; X = sim.y$X; Z = list(batch = sim.y$Z, cask = sim.y$Z);
#id = id.list;
contr.loglik <- function(y, X, Z = NULL, id) {
    if(!is.vector(y)) stop("y must be a vector. Thanks!")
    X <- as.matrix(X); num.fixed <- ncol(X)
    n.uniids <- unlist(lapply(id, function(x) length(unique(x))))
    P <- diag(length(y)) - X%*%solve(t(X)%*%X)%*%t(X)

    if(is.null(Z)) {
        Z.mat <- new.b <- new.b.mat <- NULL
        RSS <- as.vector(t(y)%*%P%*(y))
        sigma2 <- RSS/(length(y)-num.fixed)
        logL <- -0.5*(length(y)-num.fixed)*log(2*pi*sigma2) -0.5*(length(y)-num.fixed)
        S <- Z.mat <- NULL
        num.ran <- 0
    }
    if(!is.null(Z)) {
        if(!is.list(Z) || !is.list(id))
            stop("Z and id must be list, each element of which
is the random effects covariates and corresponding ids. Thanks!")
        if(length(Z) != length(id))
            stop("Z and id must be lists of the same length. Thanks!")
        for(k in 1:length(Z)) { Z[[k]] <- as.matrix(Z[[k]]) }
        Z.full <- NULL
        for(k in 1:length(id)) {
            Z.mat <- matrix(0, nrow = length(y), ncol = ncol(Z[[k]])*n.uniids[k])
            Z.mat.colid <- rep(1:n.uniids[k], each = ncol(Z[[k]]))
            for(i in 1:n.uniids[k]) {
                sel.i <- which(id[[k]] == i)
                Z.mat[sel.i,Z.mat.colid==i] <- Z[[k]][sel.i,]
            }
            Z.full <- cbind(Z.full,Z.mat)
        }
    }
}
out.list <- list(Zdiag = Z.full)
return(out.list)
}

```

```

## sig22 is covariance matrix corresponding to random effects b_{-0}
## custom.bminus0 allows manual supply of the random effects b_{-0}
## which are used in the Monte-Carlo integration,
## when one wants to assume the true random effects distribution
## is NOT normally distributed
calc.power <- function(X, Z, Z0, Zminus0, sig22 = NULL,
custom.bminus0 = NULL, mc.samps = 8000,
seed = 123, alfa=0.05) {
  if(is.null(custom.bminus0)) {
    message("Since custom.bminus0 is not supplied,
then it is assumed the random effects are
multivariate normally distributed.")
    if(is.null(sig22)) {
      stop("sig22 much be given if the random
effects are multivariate normally distributed.")
    }
    bminus0 <- rmvnorm(mc.samps, sigma = sig22) ## b_{-0}
  }
  if(!is.null(custom.bminus0)) {
    mc.samps <- nrow(custom.bminus0)
    message("custom.bminus0 is supplied, so sig22 is ignored.")
    bminus0 <- custom.bminus0
    if(ncol(Zminus0) != ncol(bminus0))
      stop("The number of columns in Zminus0 and custom.bminus0 should match.
Please rectify.")
  }
  if(length(unique(nrow(X), nrow(Z), nrow(Z0), nrow(Zminus0))) != 1)
stop("The number of rows in X, Z, Z0 and Zminus0 should match.
Please reticfy.")
  if(ncol(Z) != (ncol(Z0) + ncol(Zminus0)))
stop("ncol(Z) = ncol(Z0) + ncol(Zminus0). Please reticfy.")
  old <- .Random.seed
  on.exit( { .Random.seed <<- old } )
  set.seed(123)
  fit.alt <- lm(rnorm(nrow(X)) ~ Z + X - 1)
  fit0 <- lm(rnorm(nrow(X)) ~ Z0 + X - 1)
  df1 <- fit.alt$rank - fit0$rank
  df2 <- nrow(X) - fit.alt$rank
  P <- model.matrix(fit0)
  P <- P%*%ginv(t(P)%*%P)%*%t(P)
  Q <- t(Zminus0)%*%(diag(nrow(X)) - P)%*%Zminus0
  rm(fit0, fit.alt, P)
  store.pows <- ncps <- numeric(mc.samps)
  crit.val <- qf(1-alfa, df1, df2)
  for(k in 1:mc.samps) {
    store.pows[k] <- 1-pf(crit.val, df1, df2,
ncp = t(bminus0[k,])%*%Q%*%bminus0[k,])
    ncps[k] <- t(bminus0[k,])%*%Q%*%bminus0[k,]
  }
  return(list(power = mean(store.pows), ncp = mean(ncps),
crit.val = crit.val, alfa=alfa))
}

#power function for eliminating all random effects
calc.powernull <- function(X, Z, Zminus0, sig22 = NULL,
custom.bminus0 = NULL, mc.samps = 8000,

```

```

seed = 123,alfa=0.05) {
  if(is.null(custom.bminus0)) {
    message("Since custom.bminus0 is not supplied,
            then it is assumed the random effects are multivariate normally distributed.")
    if(is.null(sig22)) {
      stop("sig22 much be given if the random effects are
            multivariate normally distributed.")
    }
    bminus0 <- rmvnorm(mc.samps, sigma = sig22) ## b_{-0}
  }
  if(!is.null(custom.bminus0)) {
    mc.samps <- nrow(custom.bminus0)
    message("custom.bminus0 is supplied, so sig22 is ignored.")
    bminus0 <- custom.bminus0
    if(ncol(Zminus0) != ncol(bminus0))
      stop("The number of columns in Zminus0 and custom.bminus0 should match.
            Please rectify.")
  }
  if(length(unique(nrow(X),nrow(Z),nrow(Zminus0))) != 1)
    stop("The number of rows in X, Z, and Zminus0 should match.
          Please reticfy.")
  if(ncol(Z) != ncol(Zminus0))
    stop("ncol(Z) = ncol(Zminus0). Please reticfy.")
  old <- .Random.seed
  on.exit( { .Random.seed <<- old } )
  set.seed(123)
  fit.alt <- lm(rnorm(nrow(X)) ~ Z + X - 1)
  fit0 <- lm(rnorm(nrow(X)) ~ X )
  df1 <- fit.alt$rank - fit0$rank
  df2 <- nrow(X) - fit.alt$rank
  P <- model.matrix(fit0)
  P <- P%*%ginv(t(P)%*%P)%*%t(P)
  Q <- t(Zminus0)%*%(diag(nrow(X)) - P)%*%Zminus0
  rm(fit0, fit.alt, P)

  store.pows <- ncps <- numeric(mc.samps)
  crit.val <- qf(1-alfa, df1, df2)
  pos=runif(mc.samps)
  for(k in 1:mc.samps) {
    store.pows[k] <-1-pf(crit.val, df1, df2,
                       ncp = t(bminus0[k,])%*%Q%*%bminus0[k,])
    ncps[k] <- t(bminus0[k,])%*%Q%*%bminus0[k,]
  }
  return(list(power = mean(store.pows), ncp = mean(ncps),
             crit.val = crit.val,minp=min(store.pows),alfa=alfa))
}

#####
#Simulations
#part one
true_D=matrix(c(0.1,0.05,0.05,0.1),nrow = 2,byrow = T)#covariance matrix
#of random effects
clus_size <- c(3,5)
num_clus <- c(10,15)
null_mod <- 0#number of random effects in nullmod
alt_mod <- 1:2

```

```

q_0=2#df
Q1=matrix(0,nrow = 4,3)
p=0
for (ni in 1:2){
  for (mi in 1:2){
    p=p+1
    pend=rep(1:clus_size[mi],num_clus[ni])
    X <- cbind(1,pend);
    Z <- cbind(1,pend)
    FLC=NULL
    T1=NULL
    #LRT=NULL
    for (i in 1:1000){
      simdat <- gendat.glmm(n = num_clus[ni], m = clus_size[mi], X = X,
        beta=c(1,2), Z = Z, D = true_D, sigma2 = 1,
        family = gaussian())
      #FLC
      fit_alt=contr.loglik(y = simdat$y,
        X = simdat$X,Z = list(inte2 = (simdat$Z[,1]),
        inte3 = (simdat$Z[,2])),
        id = list(cluster = simdat$id,cluster2 = simdat$id))
      ## To construct Zdiag for alt model
      fit_alt2 <- lm(simdat$y ~ fit_alt$Zdiag+ simdat$X - 1)
      fit_null2 <- lm(simdat$y ~ simdat$X )

      start.time <- Sys.time()
      a=anova(fit_null2, fit_alt2)
      end.time <- Sys.time()
      time.taken <- end.time - start.time
      T1=c(T1,time.taken)
      FLC=c(FLC,a$F[2])
    }
    df=c(a$Res.Df[1]-a$Res.Df[2],a$Res.Df[2])
    #power
    getp=calc.powernull(X=simdat$X,Z=simdat$Zdiag,
      Zminus0 = simdat$Zdiag,
      sig22 = kronecker(diag(num_clus[ni]), true_D),alfa=0.05)
    Q1[p,1]=sum(FLC>qf(0.95,df[1],df[2]))/1000
    Q1[p,3]=getp$power
    Q1[p,2]=mean(T1)
  }
}
rownames(Q1)=c("10:3","10:5","15:3","15:5")
colnames(Q1)=c("FLC","T1","pow")

Q1#matrix of the empirical power
#SECOND SETTING
D=matrix(c(1,0.2,0.2,1),nrow = 2,byrow = T)#covariance matrix
#of the tested effects
true_D=rbind(c(1,0,0),cbind(rep(0,2),D))
clus_size <- 10
num_clus <- c(7,15,25,50)

null_mod <- 1
alt_mod <- 1:3

```

```

Q2=matrix(0,nrow = 4,5)
p=1
for (ni in 1:4){
  FLC=NULL
  T1=NULL
  T2=NULL
  LRT=NULL
  for (i in 1:1000){
    X <- cbind(1,runif(num_clus[ni]*clus_size));
    Z <- cbind(1,runif(num_clus[ni]*clus_size),runif(num_clus[ni]*clus_size))
    simdat <- gendat.glmm(n = num_clus[ni], m = clus_size, X = X,beta=c(1,2),
    Z = Z, D = true_D, sigma2 = 1, family = gaussian())
    #FLC
    fit_null <- contr.loglik(y = simdat$y, X = simdat$X,
    Z = list(cluster = simdat$Z[,null_mod]),
    id = list(cluster = simdat$id))
    fit_alt=contr.loglik(y = simdat$y, X = simdat$X,
    Z = list(inte1 = (simdat$Z[,1]),
    z2 = (simdat$Z[,2]),z3=simdat$Z[,3]),
    id = list(cluster = simdat$id,cluster2 = simdat$id,
    cluster3 = simdat$id))
    ## To construct Zdiag for alt model
    fit_alt2 <- lm(simdat$y ~ fit_alt$Zdiag+ simdat$X - 1)
    fit_null2 <- lm(simdat$y ~ fit_null$Zdiag+ simdat$X -1 )

    start.time <- Sys.time()
    a=anova(fit_null2, fit_alt2)
    end.time <- Sys.time()
    time.taken <- end.time - start.time
    T1=c(T1,time.taken)
    FLC=c(FLC,a$F[2])
    #ANOVA
    y=as.matrix(simdat$y)
    id=as.matrix(simdat$id)
    x=simdat$X
    z=simdat$Z
    fit0=suppressMessages(lmer(y~x[,2]+ (1|id),REML = F))
    fit1=suppressMessages(lmer(y~x[,2]+(z[,2]+z[,3]|id ),REML = F))

    start.time <- Sys.time()
    lrt=anova(fit1,fit0)
    end.time <- Sys.time()
    time.taken <- end.time - start.time
    T2=c(T2,time.taken)
    LRT=c(LRT,lrt$Chisq[2])
  }
  df=c(a$Res.Df[1]-a$Res.Df[2],a$Res.Df[2])
  Q2[p,1]=sum(FLC>qf(0.95,df[1],df[2]))/1000
  fit_minus=contr.loglik(y = simdat$y, X = simdat$X,
  Z = list(z2 = (simdat$Z[,2]),z3=simdat$Z[,3]),
  id = list(cluster2 = simdat$id,cluster3 = simdat$id))
  ## To construct Zdiag for alt model
  getp=calc.power(X=simdat$X,Z=fit_alt$Zdiag,Z0=fit_null$Zdiag,
  Zminus0 = fit_minus$Zdiag,
  sig22 = kronecker(diag(num_clus[ni]),

```

```

      true_D[-null_mod,-null_mod]), alfa=0.05)
      Q2[p,3]=getp$power
      Q2[p,2]=mean(T1)
      Q2[p,4]=sum(LRT>qchisq(0.95, lrt$Df [2]))/1000
      Q2[p,5]=mean(T2)
      p=p+1
    }
  rownames(Q2)=c("7", "15", "25", "50")
  colnames(Q2)=c("FLC", "T1", "pow", "LRT", "T2")

#####THIRD SETTINGS
library(mvtnorm)
num_clus=c(10,20,40)
clus_size=c(10,20)
phi=0.5#autoregressive parameter
cov=diag(7)#covariance matrix of the covariates
for (i in 1:7){
  for (j in 1:7){
    cov[i,j]=phi^(abs(i-j))
  }
}
beta=c(1,-1,1,-1,1,-1,1,-1)
tau=c(0,0.1,0.2)

null_mod=1:2
alt_mod=1:4
p=1
Q3=matrix(NA,nrow = 6,ncol=5)
rownames(Q3)=c(levels(as.factor(num_clus)[1]:as.factor(clus_size)[1]))
colnames(Q3)=c("FLC", "T1", "pow", "LRT", "T2")
tau=0.2
mat2=matrix(c(tau,tau/2,tau/2,tau),byrow = T,nrow = 2)
mat3=matrix(rep(tau/2,4),ncol = 2)

for (ni in 1:3){
  for (mi in 1:2){
    FLC=NULL
    T1=NULL
    LRT=NULL
    T2=NULL
    for (i in 1:1000){
      esp=rmvnorm(n=num_clus[ni]*clus_size[mi],sigma=cov)
      X=cbind(1,esp)
      Z=X[,1:4]
      mat1=rWishart(n=1,df=3,Sigma = diag(c(0.5,0.5)))[, ,1]
      true_D=rbind(cbind(mat1,mat3),cbind(mat3,mat2))
      simdat=gendat.glm(n=num_clus[ni],m=clus_size[mi],X=X,beta=beta,
      Z=Z,D=true_D)
      fit_null <- contr.loglik(y = simdat$y, X = simdat$X,
      Z = list(cluster = simdat$Z[,null_mod]),
      id = list(cluster = simdat$id))
      fit_alt=contr.loglik(y = simdat$y, X = simdat$X,
      Z = list(inte=simdat$Z[,alt_mod]),
      id = list(cluster = simdat$id))
      ## To construct Zdiag for alt model

```

```

fit_alt2 <- lm(simdat$y ~ fit_alt$Zdiag+ simdat$X - 1)
fit_null2 <- lm(simdat$y ~ fit_null$Zdiag+ simdat$X -1 )

start.time <- Sys.time()
a=anova(fit_null2, fit_alt2)
end.time <- Sys.time()
time.taken <- end.time - start.time
T1=c(T1,time.taken)

FLC=c(FLC,a$F [2])
#ANOVA
y=as.matrix(simdat$y)
id=as.matrix(simdat$id)
x=simdat$X
z=simdat$Z
fit0=suppressMessages(lmer(y~x[,1:8]+ (z[,2]|id),REML = F))
fit1=suppressMessages(lmer(y~x[,1:8]+(z[,2]+z[,3]+z[,4]|id ),REML = F))

start.time <- Sys.time()
lrt=anova(fit1,fit0)
end.time <- Sys.time()
time.taken <- end.time - start.time
T2=c(T2,time.taken)
LRT=c(LRT,lrt$Chisq [2])
}
df=c(a$Res.Df [1]-a$Res.Df [2],a$Res.Df [2])
Q3[p,1]=sum(FLC>qf(0.95,df [1],df [2]))/1000

fit_minus=contr.loglik(y = simdat$y, X = simdat$X,
Z = list(z2 = (simdat$Z[,-null_mod])),
id = list(cluster2 = simdat$id))
## To construct Zdiag for alt model
getp=calc.power(X=simdat$X,Z=fit_alt$Zdiag,Z0=fit_null$Zdiag,
Zminus0 = fit_minus$Zdiag,
sig22 = kronecker(diag(num_clus[ni]),
true_D[-null_mod,-null_mod]),alfa=0.05)
Q3[p,3]=getp$power
Q3[p,2]=mean(T1)
Q3[p,4]=sum(LRT>qchisq(0.95,lrt$Df [2]))/1000
Q3[p,5]=mean(T2)
p=p+1
}
}

#second part
#Simulation graphics code
multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
  library(grid)
  # Make a list from the ... arguments and plotlist
  plots <- c(list(...), plotlist)
  numPlots = length(plots)
  # If layout is NULL, then use 'cols' to determine layout
  if (is.null(layout)) {
    # Make the panel
    # ncol: Number of columns of plots

```

```

        # nrow: Number of rows needed, calculated from # of cols
        layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
            ncol = cols, nrow = ceiling(numPlots/cols))
    }
    if (numPlots==1) {
        print(plots[[1]])
    } else {
        # Set up the page
        grid.newpage()
        pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))
        # Make each plot, in the correct location
        for (i in 1:numPlots) {
            # Get the i,j matrix positions of the regions that contain this subplot
            matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))
            print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                layout.pos.col = matchidx$col))
        }
    }
}

#First design
powinull=function(sig){
    true_D[1,1]=true_D[2,2]=sig
    simdat <- gendat.glmm(n = num_clus[ni], m = clus_size[mi],
        X = X,beta=c(1,2), Z = Z, D = true_D, sigma2 = 1,
        family = gaussian())
    fit_alt=contr.loglik(y = simdat$y, X = simdat$X,
        Z = list(inte2 = (simdat$Z[,1]),inte3 = (simdat$Z[,2])),
        id = list(cluster = simdat$id,cluster2 = simdat$id))
    ## To construct Zdiag for alt model
    pow=calc.powernull(X=simdat$X,Z=simdat$Zdiag,Zminus0 = simdat$Zdiag,
        sig22 = kronecker(diag(num_clus[ni]), true_D),alfa=0.05)
    return(pow$power)
}

clus_size <- c(3,5)
num_clus <- c(10,15)
null_mod <- 0
alt_mod <- 1:2
q_0=2#number of random effects
K=100

ro=c(0,0.1,0.25,0.5)
p=1
Q1=matrix(NA,ncol=20,nrow = K)
colnames(Q1)=c("sig1","a","b","c","d","sig2","e","f","g","h","sig3",
    "i","j","k","l","sig4","m","n","o","p")
for (i in 1:4){
    true_D=matrix(c(1,0,0,1),nrow = 2,byrow = T)
    true_D[2,1]=true_D[1,2]=ro[i]
    sig=seq((ro[i]+0.01),1,length.out=K)
    Q1[,5*(i)-4]=sig
    for (ni in 1:2){
        for (mi in 1:2){
            pend=rep(1:clus_size[mi],num_clus[ni])

```

```

        X <- cbind(1,pend);
        Z <- cbind(1,pend)
        pp=sapply(sig, powinull)
        Q[,p+1]=pp
        p=p+1
    }
}
p=p+1
}

#Second design
pow2alt=function(sig){
  true_D[1,1]=true_D[2,2]=sig
  simdat <- gendat.glmm(n = num_clus[ni], m = clus_size, X = X,beta=c(1,2),
  Z = Z, D = D, sigma2 = 1, family = gaussian())
  fit_null=contr.loglik(y = simdat$y, X = simdat$X,
  Z = list(inte1 = (simdat$Z[,1])),
  id = list(cluster = simdat$id))
  ## To construct Zdiag for alt model
  fit_minus=contr.loglik(y = simdat$y, X = simdat$X,
  Z = list(inte2 = (simdat$Z[,2]),
  inte3 = (simdat$Z[,3])),
  id = list(cluster = simdat$id,cluster2 = simdat$id))
  ## To construct Zdiag for min model
  fit_alt=contr.loglik(y = simdat$y, X = simdat$X,
  Z = list(inte1 = (simdat$Z[,1]),inte2 = (simdat$Z[,2]),
  inte3 = (simdat$Z[,3])),
  id = list(cluster = simdat$id,cluster = simdat$id,
  cluster2 = simdat$id))
  ## To construct Zdiag for alt model
  pow=calc.power(X=simdat$X,Z=fit_alt$Zdiag,Zminus0 = fit_minus$Zdiag,
  Z0=fit_null$Zdiag,
  sig22 = kronecker(diag(num_clus[ni]),true_D),alfa=0.05)
  return(pow$power)
}

clus_size <- 10
num_clus <- c(7,15,25,50)

null_mod <- 1
alt_mod <- 1:3
q_0=2
ro=c(0,0.1,0.25,0.5)
K=100
p=1
Q2=matrix(NA,ncol=20,nrow = K)
colnames(Q2)=c("sig1","a","b","c","d","sig2","e","f","g","h","sig3",
"i","j","k","l","sig4","m","n","o","p")

for (i in 1:4){
  true_D=matrix(c(1,0,0,1),nrow = 2,byrow = T)
  true_D[2,1]=true_D[1,2]=ro[i]
  sig=seq((ro[i]+0.01),1,length.out=K)
  Q2[,5*(i)-4]=sig
  for (ni in 1:4){

```

```

        D=diag(3)
        D[2:3,2:3]=true_D
        X <- cbind(1,runif(clus_size*num_clus[ni]));
        Z <- cbind(1,runif(clus_size*num_clus[ni]),runif(clus_size*num_clus[ni]))
        pp=sapply(sig, pow2alt)

        Q2[,p+1]=pp
        p=p+1
    }
    p=p+1
}

#Third design
pow3alt=function(tau){
    mat2=matrix(c(tau,tau/2,tau/2,tau),byrow = T,nrow = 2)
    mat3=matrix(rep(tau/2,4),ncol = 2)
    mat1=rWishart(n=1,df=3,Sigma = diag(c(0.5,0.5)))[, ,1]
    true_D=rbind(cbind(mat1,mat3),cbind(mat3,mat2))
    simdat=gendat.glm(n=num_clus[ni],m=clus_size[ni],X=X,beta=beta,
    Z=Z,D=true_D)
    fit_null <- contr.loglik(y = simdat$y, X = simdat$X,
    Z = list(cluster = simdat$Z[,null_mod]),
    id = list(cluster = simdat$id))
    fit_alt=contr.loglik(y = simdat$y, X = simdat$X,
    Z = list(inte=simdat$Z[,alt_mod]),
    id = list(cluster = simdat$id))
    ## To construct Zdiag for alt model
    fit_minus=contr.loglik(y = simdat$y, X = simdat$X,
    Z = list(z2 = (simdat$Z[,-null_mod])),
    id = list(cluster2 = simdat$id))
    ## To construct Zdiag for alt model
    getp=calc.power(X=simdat$X,Z=fit_alt$Zdiag,Z0=fit_null$Zdiag,
    Zminus0 = fit_minus$Zdiag,
    sig22 = kronecker(diag(num_clus[ni]),
    true_D[-null_mod,-null_mod]),alfa=0.05)
    return(getp$pow)
}

num_clus=c(10,20,40)
clus_size=c(10,20)

phi=0.5
cov=diag(7)
for (i in 1:7){
    for (j in 1:7){
        cov[i,j]=phi^(abs(i-j))
    }
}

K=200
beta=c(1,-1,1,-1,1,-1,1,-1)
null_mod=1:2
alt_mod=1:4

p=1

```

```

Q3=matrix(NA,nrow = K,ncol=7)
colnames(Q3)=c("tau","a","b","c","d","e","f")

#matrice di covar degli effetti casuali
tau=seq(0,1,length.out=K)
for (ni in 1:3){
  for (mi in 1:2){
    esp=rmvnorm(n=num_cclus[ni]*clus_size[mi],sigma=cov)
    X=cbind(1,esp)
    Z=X[,1:4]
    Q3[,p+1]=sapply(tau,pow3alt)
    p=p+1
  }
}
Q3[,1]=tau

#graphics code example for the first design
Q1=data.frame(Q1)
p1=Q1%>%ggplot(aes(x=sig1))+geom_line(aes(y=a,colour="10:3"),lwd=0.9)+
xlim(0,1)+ylim(0,1)+
geom_line(aes(y=b,colour="10:5"),lwd=0.9)+
geom_line(aes(y=c,colour="15:3"),lwd=0.9)+
geom_line(aes(y=d,colour="15:5"),lwd=0.9)+
labs(title="\u03C1=0",x=expression(sigma^2),y="Power")

p2=Q1%>%ggplot(aes(x=sig2))+geom_line(aes(y=e,colour="10:3"),lwd=0.9)+
xlim(0,1)+ylim(0,1)+
geom_line(aes(y=f,colour="10:5"),lwd=0.9)+
geom_line(aes(y=g,colour="15:3"),lwd=0.9)+
geom_line(aes(y=h,colour="15:5"),lwd=0.9)+
labs(title="\u03C1=0.1",x=expression(sigma^2),y="Power")

p3=Q1%>%ggplot(aes(x=sig3))+geom_line(aes(y=i,colour="10:3"),lwd=0.9)+
xlim(0,1)+ylim(0,1)+
geom_line(aes(y=j,colour="10:5"),lwd=0.9)+
geom_line(aes(y=k,colour="15:3"),lwd=0.9)+
geom_line(aes(y=l,colour="15:5"),lwd=0.9)+
labs(title="\u03C1=0.25",x=expression(sigma^2),y="Power")

p4=Q1%>%ggplot(aes(x=sig4))+geom_line(aes(y=m,colour="10:3"),lwd=0.9)+
xlim(0,1)+ylim(0,1)+
geom_line(aes(y=n,colour="10:5"),lwd=0.9)+
geom_line(aes(y=o,colour="15:3"),lwd=0.9)+
geom_line(aes(y=p,colour="15:5"),lwd=0.9)+
labs(title="\u03C1=0.5",x=expression(sigma^2),y="Power")
multiplot(p1, p3, p2, p4, cols=2)
#Dataset Stroke
stroke.lmes=lmer(y~Group*Week+(1|Subject)+(0+Week|Subject),data=Stroke1)
summary(stroke.lmes)

stroke.lme=update(stroke.lmes,~.- (0+Week|Subject))
summary(stroke.lme)
####
res=residuals(stroke.lmes,type="response")
ggplot(data.frame(res), aes(sample =res)) +

```

```

stat_qq(distribution = qnorm )+
stat_qq_line(distribution =qnorm)+
labs(title="Normal Q-Q Plot",x="theoretical",y="sample")

shapiro.test(res)

h=ranef(stroke.lmes)$Subject

p1=ggplot(data.frame(h[,1]), aes(sample =h[,1])) +
stat_qq(distribution = qnorm )+
stat_qq_line(distribution =qnorm)+
labs(title="Normal Q-Q Plot-Intercetta",x="theoretical",y="sample")

shapiro.test(h[,1])
p2=ggplot(data.frame(h[,2]), aes(sample =h[,2])) +
stat_qq(distribution = qnorm )+
stat_qq_line(distribution =qnorm)+
labs(title="Normal Q-Q Plot-Pendenza",x="theoretical",y="sample")

shapiro.test(h[,2])

multiplot(p1,p2,cols=1)
anova(stroke.lme,stroke.lmes)
gruppo=c(rep(1,64),rep(2,64),rep(3,64))
fitn=contr.loglik(y = Stroke1$y, X = cbind(1,Stroke1$Group,Stroke1$Week),
Z =list(inte=rep(1,192)) ,
id = list(cluster = Stroke1$Subject))
fitb=contr.loglik(y = Stroke1$y, X = cbind(1,Stroke1$Group,Stroke1$Week),
Z =list(inte=rep(1,192),sett=Stroke1$Week) ,
id = list(cluster=Stroke1$Subject,c1=Stroke1$Subject))
(fitn$Zdiag)

f1=lm(Stroke1$y~fitn$Zdiag+cbind(1,Stroke1$Group,Stroke1$Week)-1)
summary(f1)

f2=lm(Stroke1$y~cbind(1,Stroke1$Group,Stroke1$Week)+fitb$Zdiag-1)
summary(f2)

anova(f1,f2)
anova(stroke.lme,stroke.lmes)

####power function
X=cbind(1,Stroke1$Group,Stroke1$Week)
num_clus=length(unique(Stroke1$Subject))
num_eff=2

fitminus=contr.loglik(y = Stroke1$y, X = cbind(1,Stroke1$Group,Stroke1$Week),
Z =list(sett=Stroke1$Week) ,
id = list(cluster=Stroke1$Subject))
q=summary(stroke.lmes)
D=diag(c(q$varcor$Subject[1],q$varcor$Subject.1[1]))

D
calc.power(X=X,Z=fitb$Zdiag,Z0=fitn$Zdiag, Zminus0 = fitminus$Zdiag,
sig22 =kronecker(diag(num_clus),D[-1,-1]))

```

```
#parametric bootstrap
Q=PBmodcomp(stroke.lmes,stroke.lme)
Q$test
Q$time

#Restricted likelihood ratio
stroke.lmess=update(stroke.lmes,~.(1|Subject))
exactRLRT(m=stroke.lmess,m0=stroke.lme,mA=stroke.lmes)
```

Bibliografia

- AZZALINI, A. (1996). *Statistical Inference based on the Likelihood*. Chapman & Hall.
- BATES, D., MAECHLER, M., BOLKER, B., WALKER, S. et al. (2014). lme4: Linear mixed-effects models using eigen and s4. *R package version 1.1-7* .
- BOLKER, B. M., BROOKS, M. E., CLARK, C. J., GEANGE, S. W., POULSEN, J. R., STEVENS, M. H. H. & WHITE, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution* **24**, 127–135.
- BRIEN, C. (2023). asremlplus: Augments 'asreml-r' in fitting mixed models and packages generally in exploring prediction differences. *R package version 4.3.55* .
- CRAINICEANU, C. M. & RUPPERT, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**, 165–185.
- DAVISON, A. C. (2003). *Statistical Models*. Cambridge University Press.
- DOBSON, A. J. & BARNETT, A. G. (2018). *An Introduction to Generalized Linear Models*. CRC Press.
- DRIKVANDI, R., VERBEKE, G., KHODADADI, A. & PARTOVI NIA, V. (2013). Testing multiple variance components in linear mixed-effects models. *Biostatistics* **14**, 144–159.
- FITZMAURICE, G. M., LIPSITZ, S. R. & IBRAHIM, J. G. (2007). A note on permutation tests for variance components in multilevel generalized linear mixed models. *Biometrics* **63**, 942–946.
- GRIGOLETTO, M., VENTURA, L. & PAULI, F. (2017). *Modello Lineare: Teoria e Applicazioni con R*. G Giappichelli Editore.

- HALEKOH, U. & HØJSGAARD, S. (2014). A kenward-roger approximation and parametric bootstrap methods for tests in linear mixed models—the r package pbkrtest. *Journal of Statistical Software* **59**, 1–32.
- HENDERSON, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423–447.
- HUI, F., MÜLLER, S. & WELSH, A. (2019). Testing random effects in linear mixed models: another look at the f-test (with discussion). *Australian & New Zealand Journal of Statistics* **61**, 61–84.
- LI, Z., CHEN, F. & ZHU, L. (2014). Variance components testing in anova-type mixed models. *Scandinavian Journal of Statistics* **41**, 482–496.
- MOLENBERGHS, G. & VERBEKE, G. (2007). Likelihood ratio, score, and wald tests in a constrained parameter space. *The American Statistician* **61**, 22–27.
- PACE, L. & SALVAN, A. (2001). *Introduzione alla Statistica-II. Inferenza, Verosimiglianza, Modelli*. Cedam.
- PINHEIRO, J. & BATES, D. (2023). nlme: Linear and nonlinear mixed effects models. *R package version 3.1-162* .
- QU, L., GUENNEL, T. & MARSHALL, S. L. (2013). Linear score tests for variance components in linear mixed models and applications to genetic association studies. *Biometrics* **69**, 883–892.
- R CORE TEAM (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.
- RAO, C. R. (1973). *Linear statistical inference and its applications*, vol. 2. Wiley Online Library.
- SALVAN, A., SARTORI, N. & PACE, L. (2020a). Mlgdata: Datasets for use with salvan, sartori and pace (2020). *R package version 0.1.0* .
- SALVAN, A., SARTORI, N. & PACE, L. (2020b). *Modelli Lineari Generalizzati*. Springer.
- SCHEIPL, F., GREVEN, S. & KUECHENHOFF, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis* **52**, 3283–3299.

- SELF, S. G. & LIANG, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association* **82**, 605–610.
- SINHA, S. K. (2009). Bootstrap tests for variance components in generalized linear mixed models. *Canadian Journal of Statistics* **37**, 219–234.
- STRAM, D. O. & LEE, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* , 1171–1177.
- VENABLES, W. N. & RIPLEY, B. D. (2002). *Modern Applied Statistics with S*. Springer, 4th ed. ISBN 0-387-95457-0.
- VERBEKE, G. & MOLENBERGHS, G. (2003). The use of score tests for inference on variance components. *Biometrics* **59**, 254–262.
- WICKHAM, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag.
- WILKS, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics* **9**, 60–62.
- ZELLNER, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. *Journal of the American Statistical Association* **57**, 348–368.

