Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche

Selezione di variabili in modelli lineari a effetti misti ad elevata dimensionalità: applicazione a dati reali per la previsione della risposta glicemica postprandiale

Relatrice Prof.ssa Manuela Cattelan
Dipartimento di Scienze Statistiche

Laureanda Carolina Balatroni
Matricola 2121692

Anno Accademico 2024/2025

Indice

Introduzione **1**

1 Modelli lineari a effetti misti **3**
1.1 Introduzione . . . 3
1.2 Modelli per risposte correlate . . . 3
1.3 Modelli con effetti casuali . . . 4
1.3.1 Modelli con effetti casuali per risposte normali . . . 4
1.3.2 Modelli con effetti casuali per risposte non normali . . . 9
1.3.3 Previsione degli effetti casuali . . . 10
1.3.4 Selezione del modello . . . 11

2 Modelli lineari a effetti misti in elevata dimensionalità **15**
2.1 Introduzione . . . 15
2.2 Limiti dell'approccio classico . . . 15
2.3 Funzioni di penalizzazione . . . 16
2.3.1 Penalità Lasso . . . 16
2.3.2 Penalità non convesse . . . 19
2.4 Penalizzazione nei modelli lineari a effetti misti . . . 21
2.4.1 Penalizzazione Lasso . . . 21
2.4.1.1 Risultati teorici . . . 22
2.4.1.2 Algoritmo computazionale . . . 25
2.4.2 Penalità non convesse . . . 33
2.4.2.1 Risultati teorici . . . 34
2.4.2.2 Algoritmo computazionale . . . 39

3 Approfondimento metodologico **43**
3.1 Introduzione . . . 43
3.2 Obiettivo I . . . 43
3.3 Obiettivo II . . . 45

4 Applicazione **51**
4.1 Descrizione dell'esperimento . . . 51
4.2 Contesto e motivazione scientifica . . . 52
4.3 Descrizione del dataset . . . 53
4.4 Analisi esplorativa . . . 54

4.4.1 Variabili demografiche 54
4.4.2 Variabili cliniche 55
4.4.3 Variabili sul microbioma intestinale 57
4.4.4 Letture CGM 57
4.4.5 Composizione dei pasti 58
4.5 Variabile risposta 59
4.5.1 iAUC - incremental Area Under the Curve 60
4.5.2 netAUC - net incremental Area Under the Curve 62
4.6 Preprocessing 62
4.7 Descrizione del dataset finale 66

5 Modellazione **71**
5.1 Introduzione 71
5.2 Struttura di correlazione 72
5.3 Predittori 74
5.4 Convalida incrociata a finestra espandibile 74
5.5 Modello lineare 76
5.6 Modelli lineari a effetti misti 80
5.6.1 Pendenza casuale 83
5.7 Selezione stepwise 86
5.8 Modelli lineari penalizzati 87
5.8.1 Penalizzazione Lasso 87
5.8.2 Penalizzazioni non convesse 88
5.9 Modelli lineari a effetti misti penalizzati 89
5.9.1 Pendenza casuale 93
5.10 Conclusione 97

Conclusione **99**

Appendice A **103**
A.1 Tabelle 103
A.2 Figure 116
A.3 Codice 116
A.3.1 Funzione 1: Selezione effetti casuali 116
A.3.1.1 Simulazione 1: Selezione effetti casuali 120
A.3.1.2 Applicazione ai dati 121
A.3.2 Funzione 2: Algoritmo per la selezione e la stima degli effetti fissi con penalità SCAD o Lasso 123

Bibliografia **137**

Introduzione

In numerose situazioni, i dati non costituiscono un insieme di osservazioni indipendenti, bensì sono organizzati in più livelli annidati: unità di livello inferiore risultano raggruppate in unità di livello superiore. Un esempio tipico è rappresentato dagli studi longitudinali, nei quali si dispone di osservazioni ripetute per ciascun soggetto. L'esistenza di tale struttura gerarchica comporta la necessità di adoperare metodi statistici specifici per l'analisi. Un possibile approccio per gestire dati di questo tipo è impiegare i modelli a effetti misti, che distinguono tra effetti fissi, comuni a tutta la popolazione, ed effetti casuali, specifici di gruppi o individui.

L'avvento di nuove tecnologie, specialmente in ambito clinico, ha reso sempre più frequente la disponibilità di dati gerarchici ad elevata dimensionalità, per i quali il numero di covariate p supera il numero totale di osservazioni N. Un esempio proviene dagli studi di espressione genica, dove per ciascun individuo vengono raccolte misure su migliaia di geni. In queste situazioni, i metodi tradizionali per la stima dei modelli lineari a effetti misti non possono essere impiegati. Pertanto, la gestione dei modelli lineari misti in contesti ad alta dimensionalità è diventata una problematica centrale e di fondamentale importanza nella statistica moderna. In particolare, l'attenzione della ricerca recente si è concentrata sullo sviluppo di metodi efficienti per la selezione delle variabili rilevanti in presenza di numerose covariate a effetti fissi. Al contrario, nella maggior parte delle applicazioni in ambito medico e biologico-clinico, il numero di effetti casuali è generalmente ridotto e può essere considerato prefissato.

Inizialmente, nel Capitolo 1, verranno presentati i modelli lineari a effetti misti nel caso standard in cui $p < N$, con particolare attenzione ai modelli per risposte normali. Si discuteranno le principali assunzioni e i metodi impiegati sia per la stima del modello, sia per la previsione degli effetti casuali. Si affronterà inoltre il tema della selezione del modello, richiamando i principali criteri e strumenti utilizzati a tale scopo.

Nel Capitolo 2 verranno trattati i modelli lineari a effetti misti in situazioni ad elevata dimensionalità, in cui $p \gg N$. Il numero di parametri di varianza e covarianza degli effetti casuali è ridotto ed è necessario prespecificare per quali variabili introdurre anche un effetto casuale. Per la selezione degli effetti fissi rilevanti verranno considerati due metodi di penalizzazione: la penalità Lasso e la penalità SCAD. In entrambi i casi sorgono difficoltà dovute alla non convessità della funzione obiettivo da ottimizzare, la quale rende l'analisi complessa sia da un punto di vista teorico, sia computazionale. Verranno presentate le proprietà degli stimatori e gli algoritmi implementati per la stima.

Si procederà poi nel Capitolo 3 con gli studi di simulazione. La prima questione che verrà affrontata è la selezione delle variabili a cui attribuire anche un effetto casuale. In particolare, si valuterà la capacità di uno specifico metodo proposto a tale scopo da Schelldorfer et al. (2011). Verranno adoperati diversi indicatori, tra cui ad esempio tasso di veri positivi e di falsi negativi. Successivamente, l'attenzione sarà rivolta a valutare la performance dell'algoritmo di selezione degli effetti fissi con penalità SCAD implementato. Nello specifico, si proporrà un'implementazione più generale dell'algoritmo presentato da Ghosh & Thoresen (2016) per consentire alla matrice di varianza e covarianza $\mathbf{\Psi}$ degli effetti casuali di assumere una struttura generica definita positiva. Si valuterà l'impatto di tale possibilità sulla selezione delle variabili rilevanti e sulla stima di $\mathbf{\Psi}$. Si confronteranno i risultati per valori crescenti di p e per diverse strutture della matrice del disegno $\mathbf{X}$.

Successivamente, si considererà l'applicazione ad un dataset reale. In particolare, nel Capitolo 4 verranno descritti i dati, relativi ad un esperimento condotto presso il *Sansum Diabetes Research Institute* in California, al quale hanno partecipato 45 soggetti adulti, suddivisi in tre categorie: sani, con pre-diabete e con diabete di tipo 2. Gli individui sono stati sottoposti a monitoraggio continuo della glicemia per circa una decina di giorni. A partire dalle curve glicemiche, si calcolerà l'indicatore netAUC per misurare la risposta glicemica postprandiale, cioè la variazione del livello glicemico successivamente ai pasti. Si discuteranno le principali azioni di esplorazione, pulizia e costruzione del dataset finale.

Infine, nel Capitolo 5 verranno stimati quattro modelli principali: modello lineare classico, modello lineare a effetti misti, modello lineare penalizzato e modello lineare a effetti misti penalizzato. I risultati saranno confrontati in termini di varianza degli stimatori, numero di variabili selezionate, struttura di correlazione del modello e performance predittiva.

Capitolo 1

Modelli lineari a effetti misti

1.1 Introduzione

In questo capitolo viene presentata una tipologia di modelli utilizzata per gestire risposte correlate: i modelli con effetti casuali, con particolare attenzione ai modelli lineari per risposte normali. I contenuti si basano su Salvan et al. (2020) e Laird & Ware (1982).

1.2 Modelli per risposte correlate

In molte situazioni, i dati presentano una struttura gerarchica o nidificata. Un esempio tipico è rappresentato dagli studi longitudinali, in cui, per ciascun soggetto, si dispone di una serie di misurazioni, ottenute a tempi diversi e potenzialmente sotto condizioni sperimentali variabili. Spesso, non è possibile controllare completamente le circostanze in cui le misurazioni vengono effettuate e può esserci una notevole variabilità tra individui sia nel numero, che nella tempistica delle osservazioni, conducendo a dataset sbilanciati. Nell'analisi di dati longitudinali si può pensare agli individui come unità di secondo livello e alle osservazioni ripetute come unità di primo livello o unità elementari. Siano y_{ij} le osservazioni sulla risposta per l'i-esima unità statistica, $i = 1, ..., n$ e $j = 1, ..., m_i$, realizzazioni delle variabili casuali Y_{ij}. Disponendo di m_i misurazioni per ogni soggetto, il numero totale di osservazioni risulta pari a $N = \sum_{i=1}^{n} m_i$. Il vettore delle osservazioni sulla risposta per l'i-esima unità è dato da

$$\mathbf{y}_i = (y_{i1}, ..., y_{im_i})^T. \tag{1.1}$$

Con $\mathbf{x}_{ij}$ si indica il vettore riga p-dimensionale delle variabili esplicative per l'osservazione j-esima sull'unità i-esima. Al variare di j, i valori delle variabili esplicative possono cambiare, è ciò che avviene tipicamente negli studi longitudinali.

Le osservazioni possono essere considerate indipendenti tra soggetti diversi, ma risultano tipicamente correlate all'interno dello stesso soggetto. I modelli per l'analisi di dati longitudinali devono riconoscere la relazione tra osservazioni successive su una stessa unità. Per modellare correttamente questa struttura di dipendenza e tener conto della variabilità tra unità, esistono diverse tipologie di modelli, tra cui i modelli con effetti casuali.

1.3 Modelli con effetti casuali

I modelli con effetti casuali assumono che vi siano delle caratteristiche non osservabili comuni a tutte le osservazioni relative alla stessa unità, descritte come realizzazione di una variabile casuale, detta per l'appunto effetto casuale. Gli effetti casuali implicano l'esistenza di correlazione tra misurazioni relative a una stessa unità, mentre si assume indipendenza tra gli effetti casuali relativi a unità diverse. I modelli con effetti casuali possono essere adattati anche a situazioni in cui i dati presentano una struttura multilivello, ovvero in cui si assume l'esistenza di k livelli annidati. In particolare, le osservazioni al livello 1 appartengono a gruppi di livello 2, che a loro volta appartengono a gruppi di livello superiore, e così via fino al livello k. Per considerare la relazione tra osservazioni appartenenti allo stesso gruppo, si inserisce un effetto casuale a ciascun livello.

1.3.1 Modelli con effetti casuali per risposte normali

Si consideri un campione di n soggetti, estratti casualmente da una popolazione di riferimento. Per ogni soggetto si dispone di $m_i,\ i = 1, ..., n$, osservazioni della variabile oggetto d'interesse, registrate in m_i istanti temporali. Per l'inferenza sul livello medio μ della variabile risposta, un modello adeguato può essere

$$Y_{ij} = \mu + a_i + \varepsilon_{ij}, \quad i = 1, ..., n, j = 1, ..., m_i \tag{1.2}$$

dove a_i rappresenta l'effetto dell'i-esimo soggetto e ε_{ij} è un errore casuale. Il modello è detto a effetti misti in quanto μ costituisce l'effetto fisso e a_i rappresenta l'effetto casuale. Si assume che gli effetti a_i e gli errori ε_{ij} siano realizzazioni di variabili casuali

indipendenti con distribuzione $N(0, \sigma_a^2)$ e $N(0, \sigma_\varepsilon^2)$, rispettivamente. Pertanto, si ha:

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

Le osservazioni relative a soggetti diversi sono indipendenti, mentre la covarianza intra-soggetto, cioè tra misurazioni diverse di uno stesso soggetto, risulta pari a

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

e di conseguenza,

$$Cor(Y_{ij}, Y_{ih}) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_\varepsilon^2} = \rho, \qquad j \neq h.$$

Il parametro ρ è detto coefficiente di correlazione intra-classe e costituisce la quota di varianza totale dovuta alla variabilità tra le osservazioni di uno stesso soggetto. Varia tra 0 e 1 e non dipende dagli indici j e h. Rappresenta una misura che giustifica il ricorso al modello con effetti casuali; infatti, un valore molto basso del coefficiente non segnala la presenza di correlazione all'interno dei gruppi, suggerendo dunque di evitare la modellazione a più livelli. Al contrario, un valore molto elevato di ρ, è indicazione del contributo esplicativo della struttura gerarchica. Sia $\boldsymbol{Y}_i$ il vettore di variabili casuali per l'i-esima unità. Allora, la matrice di varianze e covarianze di $\boldsymbol{Y}_i$ è data da

$$\mathrm{Var}(\boldsymbol{Y}_i) = \mathbf{V}_i = (\sigma_a^2 + \sigma_\varepsilon^2) \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}.$$

La matrice $\mathbf{V}_i$ ha struttura di equicorrelazione: tutti gli elementi non diagonali sono uguali al coefficiente di correlazione intra-classe ρ. Si sta dunque assumendo che ogni osservazione all'interno dello stesso gruppo, ad esempio lo stesso soggetto, sia ugualmente correlata alle altre. Si tratta di una semplificazione che riduce il numero di parametri da stimare. L'assunzione di equicorrelazione è spesso non realistica, specialmente per dati longitudinali, per i quali tendenzialmente si osserva che misurazioni vicine nel tempo sono più correlate rispetto a misurazioni lontane nel tempo.

Il modello (1.2) è il modello lineare gerarchico più semplice in quanto non include variabili esplicative. Si tratta di un modello a intercetta casuale (*random intercept*) poiché assume che ciascun soggetto abbia un proprio livello medio di risposta. Un'evoluzione naturale consiste nell'introdurre variabili esplicative sia a livello di unità di primo livello, che a livello di unità di secondo livello, con l'obiettivo di spiegare in modo

più completo la variabilità osservata all'interno e tra i gruppi. Inoltre, per modellare differenze individuali nell'effetto di una o più covariate, si può includere una pendenza casuale (*random slope*). Questa specificazione permette di considerare coefficienti di regressione che variano tra le unità di secondo livello, risultando spesso in un modello più flessibile. Un esempio di modello lineare con pendenza casuale è dato da

$$Y_{ij} = \beta_1 + a_i + (\beta_2 + b_i)j + \varepsilon_{ij} \quad i = 1, ..., n, \; j = 1, ..., m_i \tag{1.3}$$

e assume una relazione lineare tra tempo e risposta con coefficienti $\beta_1 + a_i$ e $\beta_2 + b_i$ variabili da unità a unità. I parametri β_1 e β_2 corrispondono agli effetti fissi e rappresentano intercetta e coefficiente angolare a livello di popolazione. I parametri a_i e b_i costituiscono gli effetti casuali e rappresentano la deviazione rispetto ai valori di popolazione per l'unità i-esima. Si definisce modello a pendenza casuale in quanto l'effetto della variabile j varia da soggetto a soggetto: ogni soggetto ha infatti la propria pendenza $\beta_2 + b_i$. In altre parole, al variare delle unità, le rette di regressione possono essere caratterizzate da diversa intercetta e diversa pendenza. Le quantità a_i, b_i e ε_{ij} sono 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 ricava

$$E(Y_{ij}) = \beta_1 + \beta_2 j,$$

$$Var(Y_{ij}) = \sigma_a^2 + \sigma_b^2 j^2 + \sigma_\varepsilon^2,$$

$$Cov(Y_{ij}, Y_{ih}) = \sigma_a^2 + \sigma_b^2 jh,$$

e quindi $Cor(Y_{ij}, Y_{ih})$ ha elementi che dipendono da j e da h. I modelli a intercetta e pendenza casuale (1.2) e (1.3) sono casi specifici del modello lineare normale con effetti misti, il quale nella sua formulazione generale diventa

$$Y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{u}_i + \varepsilon_{ij}, \tag{1.4}$$

dove $\boldsymbol{\beta}$ rappresenta il vettore p-dimensionale di effetti fissi, e $\mathbf{u}_i$ il vettore q-dimensionale di effetti casuali, avente distribuzione $N_q(\mathbf{0}, \boldsymbol{\Psi})$. Gli effetti casuali $\mathbf{u}_i$ sono indipendenti dagli errori casuali ε_{ij}, che hanno distribuzione $N(0, \sigma_\varepsilon^2)$. Generalmente le componenti di $\mathbf{z}_{ij}$ sono un sottoinsieme di quelle di $\mathbf{x}_{ij}$. Il termine $\mathbf{z}_{ij}\mathbf{u}_i$ descrive la variabilità tra unità, mentre ε_{ij} rappresenta la variabilità interna alle unità. Risulta che $E(Y_{ij}) = \mu_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta}$. Le componenti del vettore $\boldsymbol{\beta}$ possono essere associate a variabili esplicative

che dipendono solamente dall'i-esima unità e sono dette effetti fissi fra unità (*between-subject*), oppure possono essere associate a variabili esplicative che dipendono anche da j e sono in quel caso dette effetti fissi entro le unità (*within-subject*). Il modello (1.4) può essere scritto in forma più compatta tramite notazione matriciale. Sia $\mathbf{y}_i$ il vettore di risposte di dimensione $m_i \times 1$ osservato per ogni gruppo, $\mathbf{X}_i$ la matrice di effetti fissi di dimensione $m_i \times p$, $\boldsymbol{\beta}$ il vettore p-dimensionale di coefficienti di regressione fissi, $\mathbf{Z}_i$ la matrice di effetti casuali di dimensione $m_i \times q$ e $\mathbf{u}_i$ il vettore di coefficienti casuali specifico dell'i-esimo gruppo. Il modello può essere scritto come:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i \qquad i = 1, ..., n, \tag{1.5}$$

e assume che:

(i) $\boldsymbol{\varepsilon}_i \sim N_{m_i}(\mathbf{0}, \sigma_\varepsilon^2\mathbf{I}_{m_i})$ e incorrelati per $i = 1, ..., n$,

(ii) $\mathbf{u}_i \sim N_q(\mathbf{0}, \boldsymbol{\Psi})$ e incorrelati per $i = 1, ..., n$,

(iii) $\boldsymbol{\varepsilon}_1, ..., \boldsymbol{\varepsilon}_n, \mathbf{u}_1, ..., \mathbf{u}_n$ sono indipendenti.

L'assunzione (i) comporta che gli errori sono omoschedastici e indipendenti sia all'interno di ciascun gruppo che tra gruppi diversi, con varianza costante σ_ε^2 per tutte le osservazioni. Quest'ultima assunzione risulta piuttosto rigida in contesti reali e, in particolare, in studi longitudinali. Trascurare la correlazione tra ε_{ij} e ε_{ih} conduce solitamente a una errata valutazione degli standard error relativi a effetti fissi entro le unità. Per questo motivo, vi è la possibilità di estendere il modello (1.5) in maniera tale da modellare eteroschedasticità residua o correlazione residua intra-gruppo. In particolare, si può passare all'assunzione (i') $\boldsymbol{\varepsilon}_i \sim N_{m_i}(\mathbf{0}, \mathbf{R}_i)$ con $\mathbf{R}_i = \sigma_\varepsilon^2\boldsymbol{\Lambda}_i(\lambda)$ per un vettore di parametri λ. Ciò significa che gli errori possono avere varianze diverse o correlazioni tra loro all'interno dello stesso gruppo. Errori di gruppi diversi sono comunque assunti indipendenti, ma la struttura di covarianza può differire da gruppo a gruppo a seconda di $\boldsymbol{\Lambda}_i$. La matrice di covarianza $\mathbf{R}_i$ ha dimensione $m_i \times m_i$ ed è definita positiva; dipende da i attraverso la dimensione m_i, ma l'insieme dei parametri non noti in essa contenuti non dipende da i. In caso di dati longitudinali, si può ad esempio assumere una struttura autoregressiva con $Cor(\varepsilon_{ij}, \varepsilon_{ih}) = \rho^{|j-h|}$, o, più in generale, una struttura di Toeplitz con $Cor(\varepsilon_{ij}, \varepsilon_{ih}) = \rho_{|j-h|}$, che dipende da $m_i - 1$ parametri. Per semplicità di notazione, si procede considerando l'assunzione (i).

La matrice $\boldsymbol{\Psi} = \boldsymbol{\Psi}_\theta$ è una matrice di covarianza generale, dove $\boldsymbol{\theta}$ è un insieme non vincolato di parametri di dimensione q^* tale che $\boldsymbol{\Psi}_\theta$ sia definita positiva, ottenuta ad esempio tramite decomposizione di Cholesky. Le possibili strutture per $\boldsymbol{\Psi}$ possono essere

un multiplo dell'identità, una matrice diagonale oppure una matrice generale definita positiva. Si assume inoltre che le matrici del disegno $\mathbf{X}_i$ e $\mathbf{Z}_i$ siano deterministiche, cioè note e fisse. Il vettore di coefficienti $\boldsymbol{\beta}$ è comune a tutte le osservazioni, mentre il valore di $\mathbf{u}_i$ dipende dal gruppo a cui l'osservazione appartiene e rappresenta la deviazione specifica del gruppo dagli effetti generali $\boldsymbol{\beta}$.

Dal modello (1.5) si deduce che $\mathbf{y}_1, \ldots, \mathbf{y}_n$ sono indipendenti e che, condizionatamente a $\mathbf{X}_i$ e $\mathbf{Z}_i$, il vettore risposta per l'i-esimo gruppo $\mathbf{y}_i$ segue una distribuzione gaussiana, cioè $\mathbf{y}_i \sim N_{m_i}(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{V}_i(\boldsymbol{\theta}, \sigma_\varepsilon^2))$, con $\mathbf{V}_i(\boldsymbol{\theta}, \sigma_\varepsilon^2) = \mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i^\top + \sigma_\varepsilon^2\mathbf{I}_{m_i}$. Marginalmente quindi, $\boldsymbol{Y}_i$ si distribuisce come una normale di media $E(\boldsymbol{Y}_i) = \mathbf{X}_i\boldsymbol{\beta}$ e varianza $Var(\boldsymbol{Y}_i) = \mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i^T + \sigma_\varepsilon^2\mathbf{I}_{m_i}$, dove $\mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i^T$ è la componente attribuita alla variabilità tra unità, mentre $\sigma_\varepsilon^2\mathbf{I}_{m_i}$ rappresenta la variabilità entro le unità. Questo modello implica che le m_i risposte dell'unità i-esima sono indipendenti condizionatamente a $\mathbf{u}_i$ e $\boldsymbol{\beta}$. Condizionandosi agli effetti casuali si ottiene

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

che corrisponde ad un modello lineare avente $\mathbf{Z}_i\mathbf{u}_i$ come *offset*. Di conseguenza, si possono interpretare gli effetti fissi a livello di unità solo a parità di effetto casuale.

Si considerino i vettori $\mathbf{y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_n^T)^T$, $\mathbf{u} = (\mathbf{u}_1^T, ..., \mathbf{u}_n^T)^T$, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1^T, ..., \boldsymbol{\varepsilon}_n^T)^T$ e le matrici $\mathbf{X} = (\mathbf{X}_1^T, ..., \mathbf{X}_n^T)^T$, $\mathbf{Z} = diag(\mathbf{Z}_1, ..., \mathbf{Z}_n)$ e $\mathbf{V} = diag(\mathbf{V}_1, ..., \mathbf{V}_n)$. Allora, il modello (1.5) può essere scritto come:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}. \tag{1.6}$$

Sia $\boldsymbol{\phi} := (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T, \sigma_\varepsilon^2)^T$ il vettore completo di parametri con al più lunghezza $p + q(q+1)/2 + 1$. Allora, l'inferenza su $\boldsymbol{\phi}$ si basa sulla verosimiglianza del modello marginale per $\mathbf{y} = (\mathbf{y}_1^T, ..., \mathbf{y}_n^T)^T$. La log-verosimiglianza negativa è data da

$$-\ell(\boldsymbol{\phi}) = -\ell(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_\varepsilon^2) = \frac{1}{2}\left\{N\log(2\pi) + \log|\mathbf{V}| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}, \tag{1.7}$$

dove $|\mathbf{V}| = \det(\mathbf{V})$. Lo stimatore di massima verosimiglianza di $\boldsymbol{\beta}$ con $\mathbf{V}_1, ..., \mathbf{V}_n$ note si ottiene minimizzando la log-verosimiglianza negativa e ha la seguente forma

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\boldsymbol{Y} = \left(\sum_{i=1}^n \mathbf{X_i}^T\mathbf{V}_i^{-1}\mathbf{X_i}\right)^{-1}\left(\sum_{i=1}^n \mathbf{X_i}^T\mathbf{V}_i^{-1}\boldsymbol{Y_i}\right).$$

Lo stimatore si distribuisce come una normale di media $\boldsymbol{\beta}$ e matrice di covarianza

$$Var(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} = \left(\sum_{i=1}^{n} \mathbf{X_i}^T\mathbf{V}_i^{-1}\mathbf{X_i}\right)^{-1}.$$

Generalmente, lo stimatore di massima verosimiglianza dei parametri relativi alle matrici $\mathbf{V}_i$ presenta distorsione elevata. Pertanto, risulta preferibile ottenere la stima di $\mathbf{V}_i$ sulla base della verosimiglianza marginale. La verosimiglianza si ottiene a partire dal modello statistico per una trasformazione $\boldsymbol{AY}$ di $\boldsymbol{Y}$, con densità indipendente da $\boldsymbol{\beta}$. I residui ottenuti dal modello per la trasformazione sono i residui linearmente indipendenti della regressione lineare di $\boldsymbol{Y}$ su $\mathbf{X}$. Il metodo è detto della massima verosimiglianza ristretta o residua (REML, *restricted maximum likelihood estimation*). Si veda Davison (2003) per approfondimenti.

L'adattamento dei modelli lineari normali misti si può ottenere su `R` con la funzione `lme` della libreria `nlme` (Pinheiro et al., 2025) oppure con la funzione `lmer` della libreria `lme4` (Bates et al., 2015).

1.3.2 Modelli con effetti casuali per risposte non normali

Il modello lineare normale con effetti misti può essere generalizzato per trattare risposte con altre distribuzioni appartenenti alla famiglia esponenziale, tra cui ad esempio la binomiale e la Poisson. Prende il nome di modello lineare generalizzato con effetti misti (GLMM, *generalized linear mixed effects model*). Assume che, condizionatamente a $\mathbf{u}_i$, le osservazioni sulla risposta Y_{ij} siano indipendenti e distribuite secondo un modello lineare generalizzato con

$$g\left(E(Y_{ij}|\mathbf{u}_i)\right) = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{u}_i. \tag{1.8}$$

È necessario assegnare una distribuzione agli effetti casuali $\mathbf{u}_i$. Solitamente si assume che si tratti di realizzazioni indipendenti di una $N_q(\mathbf{0}, \boldsymbol{\Psi})$. Spesso si considera un effetto casuale unidimensionale, con distribuzione $N(0, \sigma_u^2)$, che introduce nel modello un'intercetta casuale. In un GLMM vale

$$E(Y_{ij}|\mathbf{u}_i) = g^{-1}(\mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{u}_i).$$

Di conseguenza, si ha che marginalmente

$$\mu_{ij} = E(Y_{ij}) = E(E(Y_{ij} \mid \mathbf{u}_j)) = \int g^{-1}(\mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{u}_i)\, p(\mathbf{u}_i; \boldsymbol{\Psi})\, \mathrm{d}\mathbf{u}_i,$$

dove $p(\mathbf{u}_i; \boldsymbol{\Psi})$ è la densità di una $N_q(\mathbf{0}, \boldsymbol{\Psi})$. L'inferenza su $(\boldsymbol{\beta}, \boldsymbol{\Psi})$ si basa sulla funzione di verosimiglianza marginale costruita dalla densità marginale di $\boldsymbol{Y}$

$$L(\boldsymbol{\beta}, \boldsymbol{\Psi}) = p(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\Psi}) = \int p(\mathbf{y} \mid \mathbf{u}; \boldsymbol{\beta})\, p(\mathbf{u}; \boldsymbol{\Psi})\, \mathrm{d}\mathbf{u}.$$

Per calcolare l'integrale si ricorre ad opportune approssimazioni, come l'integrazione numerica, i metodi Monte Carlo o le approssimazione di Laplace. L'adattamento dei modelli lineari generalizzati con effetti misti si può ottenere su `R` con la funzione `glmer` della libreria `lme4` (Bates et al., 2015).

1.3.3 Previsione degli effetti casuali

Gli effetti casuali sono predetti e non stimati in quanto si tratta di variabili casuali, non di parametri. In un modello lineare normale con effetti misti (1.5), la previsione di $\mathbf{u}_i$ è data dal miglior predittore lineare non distorto (BLUP, *Best Linear Unbiased Predictor*) $\tilde{\mathbf{u}}_i$. Consiste in una funzione lineare di $\boldsymbol{Y}_i$ con media zero e tale che, per qualunque combinazione lineare $\boldsymbol{a}^T\mathbf{u}_i$, l'errore quadratico $E(\boldsymbol{a}^T\mathbf{u}_i - \boldsymbol{a}^T\tilde{\mathbf{u}}_i)^2$ è minimo nella classe dei predittori lineari non distorti. Il previsore BLUP coincide con la stima di $E(\mathbf{u}_i|\boldsymbol{Y}_i = \mathbf{y}_i)$, ovvero il valore atteso condizionato che si calcola a partire dalla distribuzione congiunta

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

con $\mathbf{V}_i = \mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i^\top + \mathbf{R}$ e

$$Cov(\boldsymbol{Y}_i, \mathbf{u}_i) = E\left[(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i)\mathbf{u}_i^\top\right] = \mathbf{Z}_i\boldsymbol{\Psi}.$$

Per i risultati relativi alle distribuzioni condizionate in normali multivariate, risulta

$$E(\mathbf{u}_i \mid \boldsymbol{Y}_i = \mathbf{y}_i) = \mathbf{0} + \boldsymbol{\Psi}\mathbf{Z}_i^\top\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}).$$

Di conseguenza, se le matrici di covarianza sono note, si ha

$$\tilde{\mathbf{u}}_i = \boldsymbol{\Psi}\mathbf{Z}_i^\top\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}).$$

Mentre nel caso in cui $\mathbf{V}_i$ e $\boldsymbol{\Psi}$ non sono note, devono essere sostituite con le loro stime.

Anche nei GLMM si possono ottenere previsioni degli effetti casuali basate su stime di $E(\mathbf{u}_i|\boldsymbol{Y}_i = \mathbf{y}_i)$, ricavate numericamente a partire dalla distribuzione congiunta di

$(\boldsymbol{Y}_i, \mathbf{u}_i)$. Le previsioni degli effetti casuali sono ottenibili su `R` tramite la funzione `ranef` delle librerie `nlme` e `lme4`. In particolare, le previsioni restituite da `lme4` si basano sulle mode di $\mathbf{u}_i|\boldsymbol{Y}_i = \mathbf{y}_i$. Le funzioni `fitted` e `predict` permettono di ottenere i valori predetti della risposta $g^{-1}(\mathbf{x}_{ij}\hat{\boldsymbol{\beta}} + \mathbf{z}_{ij}\tilde{\mathbf{u}}_i)$.

1.3.4 Selezione del modello

Il principio della selezione del modello consiste nello scegliere, all'interno di una classe di modelli concorrenti caratterizzati da un diverso numero di parametri, il modello che meglio approssima la realtà sulla base di un opportuno criterio di selezione (Bozdogan, 1987). Idealmente, si vorrebbe individuare il "vero" modello, ovvero il modello sconosciuto sottostante il processo generatore dei dati osservati. Nella pratica, si cerca il modello più parsimonioso in grado di approssimare nella maniera più fedele possibile il vero modello. Si tratta di un obiettivo complesso nel contesto dei modelli lineari a effetti misti, a causa della presenza di due tipi di effetti con caratteristiche e ruoli diversi (Buscemi & Plaia, 2019). Nella maggior parte dei casi, si è interessati agli effetti fissi. Tuttavia, risulta cruciale identificare gli effetti casuali realmente importanti, cioè quelli i cui coefficienti variano tra soggetti. Infatti, l'esclusione di effetti rilevanti provoca una sottostima della matrice di varianza e covarianza degli effetti fissi, conducendo a stime distorte per le varianze dei coefficienti degli effetti fissi. D'altra parte, l'inclusione nel modello di effetti casuali irrilevanti fa sì che la matrice di varianza e covarianza degli effetti casuali sia singolare, determinando instabilità del modello. Pertanto, la selezione nei modelli lineari a effetti misti comporta l'individuazione sia dei "veri" parametri fissi, sia dei "veri" effetti casuali. A tale scopo sono stati proposti numerosi metodi, che si differenziano tra loro per diversi aspetti. Vi sono metodi specifici per la selezione degli effetti fissi e altri per la selezione degli effetti casuali. Inoltre, alcuni si basano sulla valutazione di un singolo modello, mentre altri prevedono il confronto tra due o più modelli. Un'ulteriore distinzione riguarda il tipo di modelli confrontati: alcuni criteri si applicano a modelli annidati, mentre altri consentono il confronto tra modelli non annidati.

Per quanto riguarda gli effetti casuali, si ha che un effetto casuale può essere considerato non rilevante se tutti i suoi elementi di varianza-covarianza, per tutte le osservazioni, risultano nulli (Ahn et al., 2012); pertanto, è sufficiente identificare correttamente le componenti diagonali non nulle e i relativi termini di covarianza (Wu et al., 2016). Un altro metodo generalmente utilizzato per valutare la struttura degli effetti casuali è il Test del rapporto di verosimiglianza (*Likelihood ratio test* - LRT), che prevede di confrontare modelli annidati in cui cambia la specificazione degli effetti casuali. Presentando la

medesima struttura per la parte a effetti fissi, i modelli possono essere stimati tramite REML. Sia L_2 la verosimiglianza del modello più generale, indicato come `modello`$_2$ e L_1 la verosimiglianza del modello ristretto, detto `modello`$_1$. Si ha che $L_2 > L_1$ e la statistica LRT

$$2 \log \left(\frac{L_2}{L_1} \right) = 2 \left[\log(L_2) - \log(L_1) \right]$$

è positiva. La distribuzione asintotica della statistica LRT, sotto l'ipotesi nulla che il modello ridotto è adeguato, corrisponde ad una χ^2 con $k_2 - k_1$ gradi di libertà, dove k_2 e k_1 sono il numero di parametri stimati nei modelli 2 e 1, rispettivamente. Sulla base del p-value ottenuto, si rifiuta o meno l'ipotesi nulla. Questo test può essere utilizzato anche per confrontare modelli annidati in cui ciò che cambia è la specificazione degli effetti fissi. In tal caso, i modelli devono essere stimati tramite massima verosimiglianza (ML). Il test LRT può essere ottenuto su `R` tramite la funzione `anova()` specificando i modelli da confrontare nel seguente modo: `anova(modello`$_2$`, modello`$_1$`)`. Per valutare la significatività dei termini degli effetti fissi, Pinheiro & Bates (2000) raccomandano l'utilizzo dei test condizionati, che permettono di eseguire test d'ipotesi sui coefficienti associati agli effetti fissi. In particolare, verificano la significatività marginale di ciascun coefficiente degli effetti fissi, dato che tutti gli altri effetti fissi sono presenti nel modello. Questo approccio prevede di condizionarsi alle stime dei parametri di varianza e covarianza degli effetti casuali $\hat{\boldsymbol{\theta}}$. Per un valore fissato di $\boldsymbol{\theta}$, le stime condizionate degli effetti fissi, $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ sono ottenute tramite minimi quadrati. La distribuzione delle stime degli effetti fissi, ottenute tramite ML o REML, è esatta per le stime condizionate $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})$. I test condizionati per la significatività di un termine nella specificazione degli effetti fissi sono dati dai test F o t per i modelli di regressione lineare, sostituendo con la stima REML condizionata della varianza residua $\hat{\sigma}^2_R(\boldsymbol{\theta})$. In pratica, il valore sconosciuto $\boldsymbol{\theta}$ viene sostituito con la sua stima ML o REML, e di conseguenza il test vale solo approssimativamente. In questo caso, la funzione `anova()` viene applicata ad un singolo modello. Sulla base dei test presentati è possibile adottare una strategia di selezione stepwise all'indietro (*backward*) per gli effetti fissi e gli effetti casuali. Si parte da un modello complesso e si rimuovono progressivamente le variabili non significative. Il metodo prevede di procedere in due step: semplificazione della struttura a effetti casuali e semplificazione della struttura a effetti fissi. È implementato nella libreria `lmerTest` di `R` e si applica a modelli stimati tramite le funzioni della libreria `lme4`. Per maggiori dettagli si veda Kuznetsova et al. (2017).

Un approccio alternativo per la selezione degli effetti fissi prevede di ricorrere ai criteri d'informazione, come il criterio di informazione di Akaike (AIC) e il criterio

d'informazione di Bayes (BIC), definiti come

$$AIC = -2l(\hat{\boldsymbol{\theta}}|y) + 2n_{par},$$

$$BIC = -2l(\hat{\boldsymbol{\theta}}|y) + n_{par}\log(N),$$

dove n_{par} indica il numero di parametri nel modello, N il numero di osservazioni e $l(\cdot)$ la log-verosimiglianza. Si preferisce il modello con il valore più basso di AIC o BIC. Entrambi i criteri contengono due termini che misurano la bontà d'adattamento e la complessità del modello. Nello specifico, il valore di verosimiglianza è usato per definire la bontà d'adattamento, mentre il numero di parametri quantifica la complessità. Nel caso del BIC si considera anche il numero di osservazioni, con la conseguenza che, in dataset di grandi dimensioni, la verosimiglianza deve crescere in maniera più evidente per restituire un modello migliore. La versione REML di AIC e BIC sostituisce $l(\hat{\boldsymbol{\theta}}|y)$ con $l_R(\hat{\boldsymbol{\theta}}|y)$ e N con $N - n_{par}$. Per questo motivo, un AIC basato su REML non può essere confrontato con un AIC ottenuto tramite ML e lo stesso vale per BIC (Zuur et al., 2009). Tale versione viene utilizzata per confrontare modelli con strutture annidate per gli effetti casuali e medesima specificazione per gli effetti fissi.

Zuur et al. (2009) discutono l'utilizzo di tali metodi all'interno di una strategia di tipo $top-down$ per affrontare il processo di selezione del modello. Per cominciare, è prevista la stima di un modello che include tutte le variabili esplicative e il maggior numero di interazioni. Successivamente, viene individuata la struttura ottimale per gli effetti casuali usando i criteri d'informazione o LRT. Infine, viene identificata la specificazione ottimale degli effetti fissi tramite i test condizionati. Questa strategia risulta talvolta non adeguata in presenza di numerose variabili esplicative, che rendono impraticabile la stima del modello completo.

Capitolo 2

Modelli lineari a effetti misti in elevata dimensionalità

2.1 Introduzione

Si discute ora un'estensione dei modelli lineari a effetti misti in situazioni ad elevata dimensionalità, in cui il numero di covariate può essere elevato, anche maggiore del numero totale di osservazioni. Si presentano dei metodi utilizzati per la selezione degli effetti fissi nei modelli lineari in contesti ad elevata dimensionalità e per la stima di massima verosimiglianza penalizzata sia degli effetti fissi che degli effetti casuali. In particolare, vengono discusse le proprietà teoriche degli stimatori e le implementazioni algoritmiche per diverse funzioni di penalità, tra cui il Lasso e una famiglia generale di penalizzazioni non convesse. I contenuti si basano su Schelldorfer et al. (2011) e Ghosh & Thoresen (2016).

2.2 Limiti dell'approccio classico

Si considera nuovamente il modello (1.5). Dato che il numero di covariate può essere elevato e soprattutto superiore al numero totale di osservazioni, per la stima dei parametri non si possono usare gli approcci classici di massima verosimiglianza o massima verosimiglianza ristretta (REML). In generale, quando $p \gg N$, il problema statistico è detto mal posto (*ill-posed*). Tuttavia, può essere risolto se il vero vettore dei coefficienti, indicato con $\boldsymbol{\beta}_0$, è sparso. La sparsità può essere misurata in termini di norma L_r, cioè $\|\boldsymbol{\beta}\|_r = (\sum_{j=1}^{p} |\beta_j|^r)^{1/r}$, $0 \leq r < \infty$. Solitamente, la sparsità viene misurata dal numero di coefficienti non nulli, cioè si ottiene calcolando la norma L_0 del vettore dei

coefficienti. Nel contesto dei modelli lineari a effetti misti, se si assume che i coefficienti di regressione fissi sono sparsi, cioè che molti parametri sono zero, si può applicare una penalità ai coefficienti di regressione fissi. Di seguito si assume che gli effetti casuali importanti sono pre-specificati e a dimensione ridotta ($q^* \ll N$).

2.3 Funzioni di penalizzazione

La scelta della funzione di penalità è un aspetto importante per la selezione regolarizzata delle variabili, in quanto determina in larga misura le proprietà degli stimatori penalizzati risultanti e della selezione delle variabili. La funzione di penalizzazione su un parametro scalare θ si indica come $p_\lambda(|\theta|)$, dove λ è il parametro di regolazione che controlla il grado di penalizzazione. Numerose funzioni di penalizzazione sono state introdotte per la regolarizzazione nel contesto dei modelli lineari a elevata dimensionalità.

2.3.1 Penalità Lasso

Un approccio molto popolare è il Lasso (*Least absolute shrinkage and selection operator*), proposto da Tibishirani (1996). Utilizza una penalità L_1 sui coefficienti di regressione, definita come $p_\lambda(|\theta|) = \lambda|\theta|$, che fa sì che tutti i coefficienti siano compressi verso lo zero e alcuni posti esattamente a zero, attuando una vera e propria selezione delle variabili. La log-verosimiglianza del modello $l(\boldsymbol{\beta})$ viene massimizzata e viene imposto un vincolo sulla norma L_1 del vettore di parametri $\boldsymbol{\beta}$. La stima lasso viene ottenuta risolvendo:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \ \ell(\boldsymbol{\beta}), \quad \text{sotto al vincolo } \|\boldsymbol{\beta}\|_1 \leq s, \tag{2.1}$$

con $s \geq 0$ e con $\|\cdot\|_1$ ad indicare la norma L_1. In maniera equivalente, la stima lasso $\hat{\boldsymbol{\beta}}$ può essere ricavata risolvendo il problema di ottimizzazione:

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \ [\ell(\boldsymbol{\beta}) - \lambda\|\boldsymbol{\beta}\|_1], \tag{2.2}$$

con $\lambda \geq 0$. I parametri s e λ sono parametri di regolazione che devono essere determinati, ad esempio tramite convalida incrociata. Nei contesti ad elevata dimensionalità può essere un'operazione molto onerosa e sono pertanto richiesti algoritmi efficienti per derivare le soluzioni delle equazioni (2.1) o (2.2). Il metodo di penalizzazione lasso è popolare in problemi di stima ad elevata dimensionalità perché risulta accurato per la previsione e la selezione di variabili, oltre ad essere computazionalmente adeguato perché coinvolge un problema di ottimizzazione convessa. Inoltre, gode di alcune importanti

proprietà nei problemi ad elevata dimensionalità (Hastie et al., 2015). Nel seguito si illustrano le principali.

Unicità della soluzione

Per $\lambda > 0$, la soluzione del problema di lasso è unica se le colonne di $\mathbf{X} \in \mathbb{R}^{n \times p}$ sono in posizione generica, cioè se ogni sotto-spazio affine $L \subset R^n$ di dimensione $k < n$ contiene al più $k + 1$ elementi dell'insieme $\{\pm x_1, \pm x_2, ..., \pm x_p\}$ escludendo coppie di punti agli antipodi. Tale proprietà risulta verificata con probabilità 1 se i dati provengono da una distribuzione di probabilità continua, mentre problemi di non unicità della soluzione si possono verificare se $\mathbf{X}$ contiene variabili discrete, ad esempio variabili dummy per predittori categoriali.

Consistenza in errore quadratico medio

Siano $\boldsymbol{\beta}_0$ e $\hat{\boldsymbol{\beta}}$ i parametri veri e quelli stimati dal lasso. Si dimostra che, per $p, n \to \infty$

$$\|\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\|_2^2/n = O_P\Big(\|\boldsymbol{\beta}_0\|_1 \sqrt{\log(p)/n}\ \Big),$$

con probabilità elevata, assumendo che la matrice del disegno $\mathbf{X}$ sia fissa e senza ulteriori condizioni su $\mathbf{X}$. Quindi, se $\|\boldsymbol{\beta}_0\|_1 = o\Big(\sqrt{\frac{n}{\log(p)}}\ \Big)$ il lasso è consistente per le previsioni. Ciò significa che, introducendo una penalità di tipo lasso, si prevede bene se la realtà è sparsa.

Consistenza nella selezione di variabili

Sia $S_0 = \{1 \leq k \leq p : \beta_{0,k} \neq 0\}$ il vero sottoinsieme di variabili attive, cioè l'insieme degli indici delle variabili con coefficienti non nulli nel vero modello sottostante. Siano $\hat{S}_\lambda = \{1 \leq k \leq p : \hat{\beta}_k(\lambda) \neq 0\}$ gli indici del sottoinsieme di variabili che il lasso include nel modello per un fissato valore λ. Sotto condizioni restrittive, con probabilità che tende a 1 quando $n \to \infty$, il lasso è consistente nella selezione di variabili, cioè $\exists \lambda \gg \sqrt{\log p/n}$, tale per cui

$$\mathbb{P}(\hat{S}_\lambda = S_0) \to 1 \quad (p \geq n \to \infty).$$

La consistenza nella selezione di variabili riguarda la capacità di selezionare il modello corretto, che significa da un lato spingere a zero i coefficienti delle variabili che non sono nell'insieme attivo, dall'altro recuperare con segno giusto le variabili nel modello vero. Una condizione necessaria affinché ciò avvenga riguarda la matrice del disegno $\mathbf{X}$ ed è

definita condizione di irrapresentabilità. Richiede che

$$\|(\mathbf{X}_{S_0}^T \mathbf{X}_{S_0})^{-1} \mathbf{X}_{S_0}^T \mathbf{X}_{S_0^c}\|_\infty \leq (1-\epsilon) \text{ per un qualche } \epsilon \in (0,1],$$

dove S_0^c è l'insieme complementare di S_0, cioè il sottoinsieme di variabili con coefficienti nulli nel vero modello sottostante. Significa che i coefficienti dei minimi quadrati per le colonne di $\mathbf{X}_{S_0^c}$ su $\mathbf{X}_{S_0}$ non devono essere troppo grandi, cioè che le variabili realmente attive non siano troppo correlate con le variabili di disturbo. Una conseguenza è che, se le variabili esplicative sono molto correlate, il lasso potrebbe non selezionare le variabili in modo corretto. Si tratta di una condizione molto restrittiva, che non viene rispettata da molte matrici del disegno $\mathbf{X}$. Per questo motivo, si rilassa la condizione, ottenendo la seguente proprietà:

$$\mathbb{P}(\hat{S}_\lambda \supseteq S_0) \to 1 \quad (p \geq n \to \infty),$$

la quale garantisce che tutte le variabili rilevanti sono incluse nel modello. Affinché valga la consistenza nella selezione di variabili è inoltre necessario che i coefficienti non nulli siano sufficientemente grandi, cioè che soddisfino la condizione $\inf_{k \in S_0} |\beta_{0,k}| \gg \sqrt{s \log(p)/n}$. Per la consistenza nella selezione di variabili, il valore di λ deve essere scelto accuratamente, in modo da bilanciare la sparsità con l'inclusione delle variabili realmente attive. Bisogna soddisfare simultaneamente $\lambda \to 0$ e $\sqrt{n}\lambda \to \infty$. Infatti, se $\lambda \to 0$, quando $n \to \infty$, allora il lasso seleziona tutte le variabili in S_0

$$P(\text{sign}(\hat{\beta}_j) = \text{sign}(\beta_{0j}) \;\; \forall j \in S_0) \to 1.$$

Se $\sqrt{n}\lambda \to \infty$ allora è possibile eliminare tutti i coefficienti in S_0^c con probabilità che tende a 1

$$P(\hat{\beta}_j = 0 \; \forall j \in S_0^c) \to 1.$$

Queste condizioni possono essere soddisfatte anche nel caso in cui $p > n$ purché p non cresca ad una velocità esponenziale rispetto a n. Si assume che la dimensione dell'insieme sparso $|S_0|$ rimanga fissata e la dimensione dell'insieme nullo cresca, in modo che $|S_0| \ll n$ e $|S_0^c| \gg n$. Al crescere di n e p viene mantenuta la sparsità, quindi se n cresce si dispone di abbastanza dati per modellare le variabili importanti. Inoltre, si ha che $|\hat{S}| \leq \min(n,p)$ e quindi, per $p \gg n$, si ottiene una forte riduzione della dimensionalità.

Consistenza dello stimatore

In generale, uno stimatore è consistente quando, al crescere del numero di osservazioni nel campione, esso converge al valore vero del parametro che sta stimando, cioè $\hat{\theta} \xrightarrow{p} \theta$.

Affinché ciò accada, sono necessarie le seguenti due proprietà

$$\lim_{n \to \infty} E(\hat{\theta}_n) = \theta, \quad \lim_{n \to \infty} Var(\hat{\theta}_n) = 0.$$

Nel caso ortonormale, $\hat{\boldsymbol{\beta}}$ è uno stimatore consistente di $\boldsymbol{\beta}_0$ se $\lambda \to 0$, poiché converge allo stimatore OLS. Tuttavia, questa proprietà non fornisce informazioni sulla velocità di convergenza. Una condizione preferibile ma più impegnativa è la $\sqrt{n}$-consistency, che prevede che, aumentando n, l'errore tra lo stimatore e il parametro vero diminuisce proporzionalmente a $1/\sqrt{n}$. Ciò significa che lo stimatore converge al vero parametro alla velocità di $\sqrt{n}$. Inoltre, implica che la distribuzione asintotica dello stimatore, opportunamente normalizzato, è tipicamente normale. Pertanto, è possibile costruire intervalli di confidenza con ampiezza proporzionale a $1/\sqrt{n}$ ed effettuare test d'ipotesi standard. Nel caso ortonormale, $\boldsymbol{\beta}$ è uno stimatore $\sqrt{n}$-consistent di $\boldsymbol{\beta}_0$ se e solo se $\sqrt{n}\lambda \to c$, con $c < \infty$.

In sintesi, il lasso può essere sia consistente nella selezione, che $\sqrt{n}$-consistente per gli stimatori. Tuttavia, queste due condizioni non possono essere soddisfatte contemporaneamente e di conseguenza il lasso non gode della cosiddetta proprietà di oracolo. Infatti, la proprietà di oracolo richiede che $\hat{\boldsymbol{\beta}}$ soddisfi

- $\hat{\boldsymbol{\beta}}_{S_0^c} = \mathbf{0}$ con probabilità che tende a 1;
- $\hat{\boldsymbol{\beta}}_S$ sia $\sqrt{n}$-consistent per $\boldsymbol{\beta}_0 \in S_0$.

Un metodo possiede la proprietà di oracolo se seleziona correttamente le variabili rilevanti e fornisce stime efficienti e utilizzabili per l'inferenza come se conoscesse già il vero sottoinsieme di variabili rilevanti. Penalizzazioni non convesse, sotto opportune condizioni regolari, possono garantire la proprietà di oracolo.

2.3.2 Penalità non convesse

Le penalità non convesse sono caratterizzate dal fatto che la penalità non è né convessa, né concava, bensì è concava sia nella metà positiva sia in quella negativa della retta reale, ed è simmetrica per la presenza del valore assoluto. Sia $P(\beta|\lambda, \gamma)$ una generica penalità non convessa applicata a un singolo parametro β. A differenza del lasso, molte penalità non convesse dipendono da λ non in maniera moltiplicativa, ossia $P(\beta|\lambda) \neq \lambda P(\beta)$. Inoltre, dipendono spesso da un parametro di regolazione γ che controlla la concavità della penalità. Una delle più popolari penalità non convesse nei problemi ad elevata

dimensionalità è SCAD (*smoothly clipped absolute deviation*), definita come

$$P(\beta|\lambda,\gamma) = \begin{cases} \lambda|\beta| & \text{se } |\beta| \leq \lambda \\ \frac{2\gamma\lambda|\beta|-\beta^2-\lambda^2}{2(\gamma-1)} & \text{se } \lambda < |\beta| \leq \gamma\lambda \\ \frac{\lambda^2(\gamma+1)}{2} & \text{se } |\beta| \geq \gamma\lambda \end{cases}$$

per un qualche $\gamma > 2$. La sua derivata è

$$\begin{cases} \lambda & \text{se } |\beta| \leq \lambda \\ \frac{\gamma\lambda-|\beta|}{\gamma-1} & \text{se } \lambda < |\beta| \leq \gamma\lambda \\ 0 & \text{se } |\beta| \geq \gamma\lambda \end{cases}$$

Pertanto, la penalità SCAD comporta la stessa distorsione del lasso per i coefficienti piccoli, ma riduce progressivamente l'intensità della penalizzazione man mano che il valore del coefficiente aumenta. Nel caso ortonormale presenta soluzione analitica, nota come operatore SCAD thresholding, pari a

$$\hat{\beta}(z) = \lambda \begin{cases} sign(z)(|z|-\lambda)_+ & \text{se } |z| \leq 2\lambda, \\ \frac{1}{(\gamma-2)}\left[(\gamma-1)z - sign(z)\gamma\lambda\right] & \text{se } 2\lambda \leq |z| \leq \gamma\lambda, \\ z & \text{se } |z| > \gamma\lambda. \end{cases}$$

A differenza della penalità lasso, soddisfa tutte e tre le proprietà desiderabili per una funzione di penalità (Fan & Li, 2001):

1. non distorsione;
2. sparsità, al fine di ottenere una selezione automatica delle variabili importanti;
3. continuità dello stimatore risultante rispetto ai dati per evitare instabilità nelle previsioni del modello.

In particolare, è stato dimostrato che la penalizzazione SCAD riduce il numero di variabili erroneamente selezionate come significative rispetto al lasso e soddisfa sia la $\sqrt{n}$-consistenza, sia la proprietà di oracolo per la selezione delle variabili. Ci si aspetta che tali miglioramenti possano essere ottenuti per la selezione di variabili anche nei modelli lineari a effetti misti.

2.4 Penalizzazione nei modelli lineari a effetti misti

Le funzioni di penalità descritte sono state estese al contesto dei modelli lineari a effetti misti per la selezione degli effetti fissi in presenza di struttura gerarchica dei dati. In questa sezione, vengono descritti i principali approcci proposti in letteratura per integrare la penalizzazione sugli effetti fissi nei modelli a effetti misti, discutendone le formulazioni, le proprietà teoriche e i metodi di ottimizzazione utilizzati.

2.4.1 Penalizzazione Lasso

Schelldorfer et al. (2011) aggiungono una penalità in norma 1 ai coefficienti di regressione fissi e ottengono così una soluzione sparsa rispetto agli effetti fissi. Questo porta a considerare la seguente funzione obiettivo:

$$Q_\lambda(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_\varepsilon^2) = \frac{1}{2}\log|\mathbf{V}| + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{k=1}^{p} |\beta_k|, \tag{2.3}$$

dove λ è un parametro di regolazione non negativo. Si stimano $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ e σ_ε^2 minimizzando la funzione obiettivo.

$$\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}_\varepsilon^2) = \underset{\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_\varepsilon^2 > 0,\, \boldsymbol{\Psi} \geq 0}{\arg\min}\, Q_\lambda(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_\varepsilon^2). \tag{2.4}$$

Considerando come fissati i parametri di varianza $\boldsymbol{\theta}$ e σ_ε^2, la minimizzazione rispetto a $\boldsymbol{\beta}$ è un problema di ottimizzazione convesso. Tuttavia, rispetto al vettore di parametri $\boldsymbol{\phi}$, la funzione obiettivo è non convessa e quindi si ha a che fare con un problema non convesso. In altre parole, la penalità è convessa ma la funzione di perdita è non convessa.

I coefficienti $\mathbf{u}_i$ degli effetti casuali vengono predetti tramite il massimo a posteriori (MAP), cioè si determinano i valori $\mathbf{u}_i$ che massimizzano la densità della corrispondente variabile aleatoria gaussiana

$$\begin{aligned}
\bar{\mathbf{u}}_i &= \arg\max_{\mathbf{u}_i} f(\mathbf{u}_i \mid \mathbf{y}_1, \ldots, \mathbf{y}_n, \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_\varepsilon^2) = \arg\max_{\mathbf{u}_i} f(\mathbf{u}_i \mid \mathbf{y}_i; \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_\varepsilon^2) \\
&= \arg\max_{\mathbf{u}_i} \frac{f(\mathbf{y}_i \mid \mathbf{u}_i, \boldsymbol{\beta}, \sigma_\varepsilon^2) \cdot f(\mathbf{u}_i \mid \boldsymbol{\theta})}{f(\mathbf{y}_i \mid \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_\varepsilon^2)} \\
&= \arg\min_{\mathbf{u}_i} \left\{ \frac{1}{\sigma_\varepsilon^2} \|\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{Z}_i\mathbf{u}_i\|^2 + \mathbf{u}_i^T \boldsymbol{\Psi}_\theta^{-1} \mathbf{u}_i \right\}.
\end{aligned}$$

Per individuare il valore ottimale di λ si usa il BIC (*Bayesian information criterion*), definito da

$$\mathrm{BIC}_\lambda = -2l(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}, \hat{\sigma}_\varepsilon^2) + (\log\, N) \cdot \hat{df}_\lambda, \tag{2.5}$$

dove $\hat{df}_\lambda = |\{1 \leq k \leq p : \hat{\beta}_k \neq 0\}| + \dim(\boldsymbol{\theta})$ è la somma del numero di coefficienti di regressione fissi non nulli e il numero di parametri di varianza-covarianza. Tale formula deriva dal fatto che il numero di gradi di libertà per lo stimatore lasso in un modello lineare senza effetti misti è dato dal numero di coefficienti stimati diversi da zero, come dimostrato da Zou et al. (2007). Si aggiunge la dimensione dei parametri di varianza e covarianza degli effetti casuali per ottenere una misura equivalente dei gradi di libertà nel modello a effetti misti. Per selezionare il parametro di regolazione potrebbero essere usati anche altri metodi, come l'AIC e la convalida incrociata. Tuttavia, studi di simulazione mostrano che BIC sembra funzionare meglio nel scegliere il valore di λ (Schelldorfer et al., 2011). Inoltre, non è ben chiaro quale funzione obiettivo utilizzare per la convalida incrociata nel contesto dei modelli a effetti misti.

2.4.1.1 Risultati teorici

Si assume per semplicità di notazione che $m_i \equiv m > 1$, cioè che tutti i gruppi abbiano la stessa numerosità. Tutti i risultati teorici e computazionali rimangono validi nel caso in cui i gruppi presentano numerosità diverse. Sia $\mathbf{y}_i \in \mathcal{Y} \subset \mathbb{R}^m$ la variabile risposta. Sia $\mathbf{X}_i$ la matrice degli effetti fissi, appartenente ad un qualche spazio $\mathcal{X}^m \subset \mathbb{R}^{m \times p}$ e sia $\mathbf{Z}_i \subset \mathbf{X}_i$. Affinché quest'ultima condizione sia valida è sufficiente assegnare ad ogni variabile un effetto fisso pari a zero. Si definisce il vettore di parametri $\boldsymbol{\phi}^T = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T, 2\log\sigma_\varepsilon) = (\boldsymbol{\beta}^T, \boldsymbol{\theta}^T, \varrho) = (\boldsymbol{\beta}^T, \boldsymbol{\eta}^T) \in \mathbb{R}^{p+q^*+1}$ e sia $\boldsymbol{\phi}_0$ il vero vettore. Per una costante $0 < K < \infty$, si considera lo spazio parametrico

$$\boldsymbol{\Phi} = \left\{ \boldsymbol{\phi}^T = (\boldsymbol{\beta}^T, \boldsymbol{\eta}^T) : \sup_{\mathbf{x} \in \mathcal{X}} |\mathbf{x}^T \boldsymbol{\beta}| \leq K,\ \|\boldsymbol{\eta}\|_\infty \leq K,\ \boldsymbol{\Psi} > 0 \right\} \in \mathbb{R}^{p+q^*+1}, \tag{2.6}$$

dove $\|\boldsymbol{\eta}\|_\infty = \max_l |\eta_l|$. Si modifica lo stimatore nella formula (2.4) imponendo che la soluzione appartenga allo spazio parametrico $\boldsymbol{\Phi}$:

$$\hat{\boldsymbol{\phi}} = \arg\min_{\boldsymbol{\phi} \in \boldsymbol{\Phi}} \left\{ \frac{1}{2} \log|\mathbf{V}| + \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \sum_{k=1}^{p} |\beta_k| \right\}. \tag{2.7}$$

Una misura teorica della bontà di uno stimatore è data dal rischio in eccesso, che consiste nella differenza tra il rischio atteso dello stimatore considerato e il rischio atteso del miglior stimatore possibile. In formula:

$$\mathcal{E}(\hat{\boldsymbol{\phi}} | \boldsymbol{\phi}_0) = \mathbb{E}[\text{perdita}(\hat{\boldsymbol{\phi}})] - \mathbb{E}[\text{perdita}(\boldsymbol{\phi}_0)].$$

Se $\hat{\boldsymbol{\phi}} = \boldsymbol{\phi}_0$, il rischio in eccesso è pari a zero, conducendo ad una situazione ottimale. Se il rischio in eccesso va a zero per n che tende ad infinito, lo stimatore diventa ottimo al crescere della numerosità del campione ed è quindi consistente. Sia $f_{\boldsymbol{\phi},\mathbf{X},\mathbf{Z}}$ la densità gaussiana per $\mathbf{y}_i$, rispetto alla parametrizzazione assegnata. Dato che si utilizza la log-verosimiglianza negativa come funzione di perdita, il rischio in eccesso coincide con la distanza di Kullback–Leibler:

$$\mathcal{E}_{\mathbf{X},\mathbf{Z}}(\boldsymbol{\phi} \mid \boldsymbol{\phi}_0) = \int \log\left(\frac{f_{\boldsymbol{\phi}_0,\mathbf{X},\mathbf{Z}}}{f_{\boldsymbol{\phi},\mathbf{X},\mathbf{Z}}}\right) f_{\boldsymbol{\phi}_0,\mathbf{X},\mathbf{Z}}\, \mathrm{d}\mu,$$

dove μ denota la misura di Lebesgue. Si definisce il rischio medio in eccesso come:

$$\bar{\mathcal{E}}_{\mathbf{X}_1,\ldots,\mathbf{X}_n,\mathbf{Z}_1,\ldots,\mathbf{Z}_n}(\boldsymbol{\phi} \mid \boldsymbol{\phi}_0) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{E}_{\mathbf{X}_i,\mathbf{Z}_i}(\boldsymbol{\phi} \mid \boldsymbol{\phi}_0).$$

Nel seguito, si omettono gli indici $_{\mathbf{X},\mathbf{Z}}$ e $_{\mathbf{X}_1,\ldots,\mathbf{X}_n,\mathbf{Z}_1,\ldots,\mathbf{Z}_n}$. Vengono riportate tre assunzioni che devono essere verificate affinché valgano i risultati teorici presentati in seguito.

- Assunzione 1: riguarda le matrici del disegno $\mathbf{Z}_i$ degli effetti casuali. Richiede che gli autovalori di $\mathbf{Z}_i^T\mathbf{Z}_i$, indicati come $(v_j^{(i)})_{j=1}^q$, per $i = 1, ..., n$, siano limitati, cioè risultino inferiori ad un certo valore K dove K è una costante minore di infinito presente nella formula (2.6);

- Assunzione 2: richiede che

  1. Dati gli autovalori di $\mathbf{Z}_i\boldsymbol{\Psi}\mathbf{Z}_i^T$, indicati come $(\omega_j^{(i)})_{j=1}^m$, almeno due di essi siano diversi, cioè per ogni i, $\exists\ j_1 \neq j_2 \in \{1, ..., m\}$ tali che $\omega_{j_1}^{(i)} \neq \omega_{j_2}^{(i)}$.
  2. Per $i = 1, ..., n$, le matrici $\boldsymbol{\Omega}_i$ definite da $(\boldsymbol{\Omega}_i)_{r,s} = \mathrm{tr}\left(\mathbf{V}_i^{-1}\frac{\partial \mathbf{V}_i}{\partial \phi_{p+r}}\mathbf{V}_i^{-1}\frac{\partial \mathbf{V}_i}{\partial \phi_{p+s}}\right)$, $r, s = 1, \ldots, q^* + 1$, siano definite strettamente positive;

- Assunzione 3: è nota come condizione dell'autovalore ristretto ed è discussa in Bickel et al. (2009). Sia $S(\boldsymbol{\beta}) = \{1 \leq k \leq p : \beta_k \neq 0\}$ l'insieme di variabili attive in $\boldsymbol{\beta}$, cioè l'insieme di coefficienti non nulli, e $\boldsymbol{\beta}_K = \{\beta_k : k \in \mathcal{K}\}$ per $\mathcal{K} \subset \{1, \ldots, p\}$. Si indica con $S_0 = S(\boldsymbol{\beta}_0)$ il vero insieme di variabili attive e con $s_0 = |S_0|$ la sua cardinalità. Si scrive $\mathbf{X}_i^T = (\mathbf{x}_1^i, \ldots, \mathbf{x}_m^i)$ e si definisce:

$$\boldsymbol{\Sigma}_{n,m} = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m} \mathbf{x}_j^i(\mathbf{x}_j^i)^T \in \mathbb{R}^{p\times p}.$$

Allora, esiste una costante $k \geq 1$, tale che per tutti i $\boldsymbol{\beta} \in \mathbb{R}^p$ che soddisfano $|\boldsymbol{\beta}_{S_0^c}\|_1 \leq 6\|\boldsymbol{\beta}_{S_0}\|_1$, vale che $\|\boldsymbol{\beta}_{S_0}\|_2^2 \leq \kappa^2 \boldsymbol{\beta}^T \Sigma_{n,m} \boldsymbol{\beta}$.

Consistenza dello stimatore

Per la consistenza dello stimatore $\ell 1$-penalizzato è richiesta la validità dell'assunzione 1. Si considera uno schema triangolare di osservazioni a partire dal modello (1.5):

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta}_n + \mathbf{Z}_i \mathbf{u}_i + \boldsymbol{\varepsilon}_i \qquad i = 1, ..., n, \tag{2.8}$$

dove i parametri $\boldsymbol{\beta}_n$ e $\boldsymbol{\eta}_n$ possono dipendere da n. Si studia la consistenza al tendere di n ad infinito, cioè al crescere del numero di gruppi, mentre la dimensione m dei gruppi (più in generale $(m_i)_{i \geq 1}$) resta fissata. Inoltre, si definisce la notazione $a \vee b = \max\{a, b\}$. Schelldorfer et al. (2011) formulano il seguente teorema:

Teorema 1

Sotto l'assunzione 1 e assumendo che:

$$\|\boldsymbol{\beta}_{0,n}\|_1 = o\left(\sqrt{\frac{n}{\log^4(n)\log(p \vee n)}}\right), \quad \lambda_n = C\sqrt{\frac{\log^4(n)\log(p \vee n)}{n}}$$

per qualche $C > 0$, qualsiasi minimo globale $\hat{\boldsymbol{\phi}}$ come in (2.7), della funzione obiettivo (2.3), soddisfa

$$\bar{\mathcal{E}}(\hat{\boldsymbol{\phi}} \mid \boldsymbol{\phi}_0) = o_P(1) \quad \text{quando } n \to \infty.$$

Questo significa che l'errore in eccesso dello stimatore, cioè quanto si allontana dal vero valore, tende a zero in probabilità. In altre parole, lo stimatore è consistente. La condizione su $\|\boldsymbol{\beta}_{0,n}\|_1$ è una condizione di sparsità sui veri coefficienti degli effetti fissi: i veri parametri $\boldsymbol{\beta}_{0,n}$ non devono essere troppo grandi in norma L_1. Per quanto riguarda il parametro di penalizzazione, deve decrescere con n ma non troppo in fretta.

Disuguaglianza oracolo

Si ottiene per il lasso un risultato di ottimalità oracolo. La disuguaglianza oracolo (*oracle inequality*) è più debole della proprietà di oracolo e comporta che l'errore dello stimatore penalizzato è dell'ordine dell'errore migliore possibile più un termine asintoticamente trascurabile. Affinché sia valida tale disuguaglianza per lo stimatore penalizzato nel contesto dei modelli lineari a effetti misti, sono necessarie alcune condizioni specifiche. In particolare, è necessario ricorrere alle assunzioni 1, 2 e 3. Sia $\lambda_0 = M_n \log n \sqrt{\frac{\log(p \vee n)}{n}}$, dove M_n è dell'ordine di $\log n$. Sia inoltre $\mathcal{J}$ l'insieme definito dal processo empirico

sottostante

$$\mathcal{J} = \left\{ \sup_{\boldsymbol{\phi}=(\boldsymbol{\beta}^T,\boldsymbol{\eta}^T)\in\boldsymbol{\Phi}} \frac{|V_n(\boldsymbol{\phi}) - V_n(\boldsymbol{\phi}_0)|}{(\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1 + \|\boldsymbol{\eta}-\boldsymbol{\eta}_0\|_2) \vee \lambda_0} \leq T\lambda_0 \right\},$$

dove T è una costante e $V_n(\boldsymbol{\phi})$ rappresenta il processo empirico, ovvero la differenza tra la media campionaria della funzione di perdita $\ell_{\boldsymbol{\phi}}(\mathbf{y}_i)$ e la sua media attesa teorica.

$$V_n(\boldsymbol{\phi}) = \frac{1}{n}\sum_{i=1}^{n} \{\ell_{\boldsymbol{\phi}}(\mathbf{y}_i) - E[\ell_{\boldsymbol{\phi}}(\mathbf{y}_i)]\}.$$

Al crescere di n, $V_n(\boldsymbol{\phi})$ misura quanto la funzione calcolata sui dati differisce dal suo valore medio. La definizione di $\mathcal{J}$ impone che, uniformemente su tutti i possibili valori dei parametri $\boldsymbol{\phi}$, la deviazione di $V_n(\boldsymbol{\phi})$ rispetto a $V_n(\boldsymbol{\phi}_0)$ sia controllata. In sintesi, $\mathcal{J}$ è l'evento in cui le fluttuazioni stocastiche della funzione obiettivo attorno ai valori veri rimangono limitate. Si dimostra che $\mathcal{J}$ ha alta probabilità, ossia si verifica con probabilità tendente a 1 al crescere di n.

Teorema 2

Sia $\hat{\boldsymbol{\phi}}_{init} = (\hat{\boldsymbol{\beta}}_{init}, \hat{\boldsymbol{\theta}}_{init}, \hat{\varrho}_{init}) = \arg\min_{\boldsymbol{\phi}\in\boldsymbol{\Phi}} Q_{\lambda_{init}}(\boldsymbol{\beta}, \boldsymbol{\theta}, \varrho)$, lo stimatore iniziale in (2.7). Sotto le assunzioni 1, 2 e 3, e per $\lambda_{init} \geq 2T\lambda_0$, si ha che, sull'insieme $\mathcal{J}$,

$$\bar{\mathcal{E}}(\hat{\boldsymbol{\phi}}_{init}|\boldsymbol{\phi}_0) + 2(\lambda_{init} - T\lambda_0)\|\hat{\boldsymbol{\beta}}_{init} - \boldsymbol{\beta}_0\|_1 \leq 9(\lambda_{init} + T\lambda_0)^2 c_0^2 \kappa^2 s_0,$$

dove c_0 è una costante indipendente dal numero di gruppi n, dalla loro numerosità m, da p e dal disegno sperimentale.

2.4.1.2 Algoritmo computazionale

L'algoritmo per la stima dei parametri di regressione fissi e dei parametri di varianza–covarianza si basa sul metodo Block Coordinate Gradient Descent (BCGD) di Tseng & Yun (2009). Prevede di ciclare attraverso le coordinate e minimizzare la funzione obiettivo $Q_\lambda(\cdot)$ rispetto a una sola coordinata alla volta, mantenendo fissi gli altri parametri. La differenza tra il problema del lasso ordinario e la penalizzazione lasso in modelli a effetti misti è che nel primo caso il ciclo riguarda solamente i parametri di regressione, mentre nel secondo caso ci sono due tipi di parametri: coefficienti degli effetti fissi e parametri di varianza e covarianza. Ad ogni passo, si approssima $Q_\lambda(\cdot)$ con una funzione quadratica strettamente convessa. Successivamente, si calcola la direzione di discesa e si effettua una ricerca in linea inesatta per garantire una diminuzione della funzione obiettivo. La ricerca in linea consiste in un approccio iterativo per trovare

un minimo locale di una funzione non lineare multidimensionale utilizzando i gradienti della funzione stessa. Calcola una direzione di ricerca p_k e poi decide quanto lontano spostarsi lungo quella direzione. L'iterazione è data da $x_{k+1} = x_k + \alpha_k p_k$, dove α_k è uno scalare positivo detto lunghezza di passo. Tale valore deve garantire una riduzione sostanziale della funzione f, ma trovare il minimo esatto a ogni iterazione sarebbe computazionalmente oneroso e l'algoritmo richiederebbe molto tempo. Per questo motivo, si può sfruttare la ricerca in linea inesatta, che permette di identificare una lunghezza di passo che consente una riduzione adeguata per f a costo minimo. Per fare ciò, gli algoritmi provano una sequenza di diversi valori per α e accettano il primo per cui sono rispettate determinate condizioni, come le condizioni di Wolfe o le condizioni di Goldstein. Si veda Nocedal & Wright (2006) per approfondimenti. Per la scelta di α_k si utilizza la regola Armijo, che può essere definita nel seguente modo: si comincia con un valore iniziale $\alpha_k^{(0)}$ e si definisce α_k come il più grande elemento di $\{\alpha_k^{(0)}\delta^r\}_{r=0,1,2,...}$ che soddisfa

$$Q_{n,\lambda}(\boldsymbol{\beta} + \alpha_j d_j \boldsymbol{e}_j, \boldsymbol{\eta}) \leq Q_{n,\lambda}(\boldsymbol{\beta}, \boldsymbol{\eta}) + \alpha_j \rho \Delta_j,$$

dove $\Delta_j = \frac{\partial}{\partial \beta_j} L_n(\boldsymbol{\beta}, \boldsymbol{\eta}) d_j + \gamma d_j^2 h_j + P_{n,\lambda}(|\beta_j + d|) - P_{n,\lambda}(|\beta_j|)$. Scelte suggerite per le costanti sono $\delta = 0.1$, $\rho = 0.001$, $\gamma = 0$ e $\alpha_j^{(0)} = 1$ per tutti i j (Bertsekas, 2003).

Sia $\boldsymbol{\phi}^T = (\boldsymbol{\beta}^T, \boldsymbol{\eta}^T) \in \mathbb{R}^{p+q^*+1}$ la parametrizzazione introdotta in precedenza. Si definiscono le funzioni

$$P(\boldsymbol{\phi}) = \sum_{k=1}^{p} |\beta_k|, \quad g(\boldsymbol{\phi}) = \frac{1}{2} \log |\mathbf{V}(\boldsymbol{\eta})| + \frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}(\boldsymbol{\eta})^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

Allora, (2.4) può essere scritto come $\hat{\boldsymbol{\phi}}_\lambda = \arg\min_{\boldsymbol{\phi}} Q_\lambda(\boldsymbol{\phi}) = g(\boldsymbol{\phi}) + \lambda P(\boldsymbol{\phi})$. Sia $\boldsymbol{e}_j$ il j-esimo vettore unità. Il procedimento per la stima di $\boldsymbol{\phi}$ è riassunto nell'Algoritmo 1.

Algoritmo 1

(0) Sia $\boldsymbol{\phi}^0 \in \mathbb{R}^{p+q^*+1}$ un valore iniziale.
Per $\ell = 0, 1, 2, ...$, sia S^ℓ l'indice del ciclo attraverso le coordinate 1, 2,...,$p+q^*$, $p+q^*+1$.
Si ripetono i seguenti passi fino a convergenza:
(1) Si approssima la derivata seconda $\frac{\partial^2}{\partial(\phi_{S^\ell})^2} Q_\lambda(\boldsymbol{\phi}^\ell)$ tramite $h^\ell > 0$;
(2) Si calcola la direzione di discesa

$$d^\ell = \arg\min_{d \in \mathbb{R}} \left\{ g(\boldsymbol{\phi}^\ell) + \frac{\partial}{\partial \phi_{S^\ell}} g(\boldsymbol{\phi}^\ell) d + \frac{1}{2} d^2 h^\ell + \lambda P(\boldsymbol{\phi}^\ell + d\boldsymbol{e}_{S^\ell}) \right\};$$

(3) Si sceglie la lunghezza del passo $\alpha^\ell > 0$ e si fissa $\boldsymbol{\phi}^{\ell+1} = \boldsymbol{\phi}^\ell + \alpha^\ell d^\ell \boldsymbol{e}_{S^\ell}$ tale da garantire una riduzione nella funzione obiettivo.

Si può dimostrare che il problema di ottimizzazione raggiunge la convergenza numerica. Infatti, se $(\boldsymbol{\phi}^{\ell})_{\ell>0}$ è scelto sulla base dell'algoritmo precedente, allora ogni punto appartenente all'insieme $\{\boldsymbol{\phi}^{\ell}\}_{\ell>0}$ è un punto stazionario di $Q_{\lambda}(\boldsymbol{\phi})$. Tuttavia, dato che il problema di ottimizzazione è non convesso, l'algoritmo potrebbe non raggiungere l'ottimo globable. Il metodo prevede di non penalizzare le covariate che presentano sia un effetto fisso che un effetto casuale perché causerebbe problemi di distorsione. Si suppone che la k-esima covariata abbia sia un effetto fisso, sia un effetto casuale. Se il parametro di regolazione λ è grande e β_k è soggetto a penalizzazione, allora $\hat{\beta}_k$ è compresso verso lo zero, mentre la variabilità tra soggetti stimata $\hat{\theta}^2$ aumenterebbe. Di conseguenza, gli effetti casuali predetti risulterebbero troppo grandi e non centrati attorno a zero, ma attorno al vero effetto fisso. In altre parole, $(\hat{\mathbf{u}}_i)_k$ avrebbe una distorsione relativa alla quantità di shrinkage in $\hat{\beta}_k$. Per determinare le covariate che hanno sia un coefficiente di effetti fissi che di effetti casuali, Schelldorfer et al. (2011) propongono la seguente strategia:

(i) Si calcola una stima Lasso ordinaria $\hat{\boldsymbol{\beta}}^{\text{Lasso}}$ tramite convalida incrociata e si definisce l'insieme di variabili attive: $\hat{S}_{\text{init}} = \left\{1 \leq k \leq p : \hat{\beta}_k^{\text{Lasso}} \neq 0\right\}$.

(ii) Per ogni $l \in \hat{S}_{\text{init}}$, si stima un modello in cui solo la l-esima variabile ha un coefficiente ad effetto casuale. Si denota la stima della varianza corrispondente con $\hat{\theta}_l^2$.

(iii) Si ordinano le varianze stimate: $\hat{\theta}^2_{[1]} \geq \hat{\theta}^2_{[2]} \geq \cdots \geq \hat{\theta}^2_{[|\hat{S}_{\text{init}}|]}$, e, per un dato $\kappa > 0$, si definisce l'insieme:

$$\mathcal{R}_{\kappa} = \left\{l \in \hat{S}_{\text{init}} : \hat{\theta}_l^2 > \kappa\right\} \cap \left\{l \in \hat{S}_{\text{init}} : \text{BIC}_{\hat{\theta}_l^2} < \text{BIC}_0\right\},$$

dove BIC_0 è il valore del BIC del modello stimato al passo (i).

(iv) Si stima un modello con $\mathbf{Z}_i = \mathbf{X}_i^{\mathcal{R}_{\kappa}}$, dove $\mathbf{X}_i^{\mathcal{R}_{\kappa}}$ contiene solo le variabili in $\mathcal{R}_{\kappa}$, e $\boldsymbol{\Psi}$ è diagonale e mantiene solo gli elementi non nulli di $\hat{\boldsymbol{\Psi}}$.

Per maggiori dettagli, si veda l'Appendice nelle Supporting Information di Schelldorfer et al. (2011), dove sono riportate le dimostrazioni dei teoremi e approfondimenti sugli aspetti computazionali.

Algoritmo alternativo - GlmmLasso

Groll & Tutz (2014) hanno presentato un approccio alternativo per stimare i modelli lineari a effetti misti in elevata dimensionalità, includendo sempre un termine di penalità

L_1. Estendono la trattazione ai modelli lineari generalizzati, cioè assumono che la densità di $\mathbf{y}_i$, condizionatamente alle variabili esplicative e agli effetti casuali $\mathbf{u}_i$, appartenga alla famiglia esponenziale, cioè sia esprimibile nella forma

$$f(y_{ij}|\mathbf{x}_{ij}, \mathbf{u}_i) = \exp\left\{\frac{y_{ij}\theta_{ij} - \kappa(\theta_{ij})}{\phi} + c(y_{ij}, \phi)\right\}, \quad i = 1, ..., n, \ j = 1, ..., m_i.$$

Con $\theta_{ij} = \theta(\mu_{ij})$ si indica il parametro naturale, $\kappa(\theta_{ij})$ è una funzione specifica corrispondente al tipo di famiglia esponenziale, $c(\cdot)$ è la costante di log-normalizzazione e ϕ è il parametro di dispersione. Ad esempio, per la distribuzione normale si ha che $\theta = \mu$, $\phi = \sigma_\varepsilon^2$, $\kappa(\theta) = \frac{1}{2}\theta^2$ e $c(y, \phi) = -\frac{y^2}{2\phi} - \frac{1}{2}\log(2\pi\phi)$. Il modello GLMM assume che le osservazioni y_{ij} siano indipendenti condizionatamente agli effetti casuali del gruppo, con media $\mu_{ij} = E(y_{ij}|\mathbf{u}_i, \mathbf{x}_{ij}, \mathbf{z}_{ij})$ e varianza $var(y_{ij}|\mathbf{u}_i) = \phi v(\mu_{ij})$, dove $v(\cdot)$ è una funzione di varianza nota e ϕ è un parametro di scala. Inoltre, $\mathbf{y}_i^T$ è il vettore di osservazioni nell'i-esimo gruppo, $\mathbf{x}_{ij}^T = (1, x_{ij1}, ..., x_{ijp})$ è il vettore di covariate associate agli effetti fissi e $\mathbf{z}_{ij}^T = (1, z_{ij1}, ..., z_{ijq})$ è il vettore di covariate associate agli effetti casuali. Il modello GLMM può essere scritto nella forma

$$g(\mu_{ij}) = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \mathbf{z}_{ij}^T\mathbf{u}_i,$$

dove g è una funzione di legame monotona e continuamente differenziabile, $\boldsymbol{\beta}^T = (\beta_0, \beta_1, ..., \beta_p)$ è il vettore di parametri e $\mathbf{u}_i$ sono gli effetti casuali dell'i-esimo gruppo con matrice di covarianza $\boldsymbol{\Psi}$ di dimensione $q \times q$. In forma matriciale, il modello diventa

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u},$$

con $\mathbf{X}^T = [\mathbf{X}_1^T, ..., \mathbf{X}_n^T]$ e $\mathbf{Z} = diag(\mathbf{Z}_1, ..., \mathbf{Z}_n)$. Il vettore $\mathbf{u}^T = (\mathbf{u}_1^T, ..., \mathbf{u}_n^T)$ ha distribuzione normale con matrice di covarianza diagonale a blocchi $\boldsymbol{\Psi}_u = diag(\boldsymbol{\Psi}, ..., \boldsymbol{\Psi})$.

Un metodo utilizzato per massimizzare i GLMM è quello della quasi-verosimiglianza penalizzata. Si considerano i vettori $\boldsymbol{\delta}^T = (\boldsymbol{\beta}^T, \mathbf{u}^T)$ e $\boldsymbol{\gamma}^T = (\phi, \boldsymbol{\theta}^T)$ dove ϕ è il parametro di dispersione e $\boldsymbol{\theta}$ è il vettore di parametri della matrice di covarianza degli effetti casuali. Allora, la log-verosimiglianza congiunta è data da:

$$\ell(\boldsymbol{\delta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \log\left(\int f(\mathbf{y}_i|\boldsymbol{\delta}, \boldsymbol{\gamma})p(\mathbf{u}_i, \boldsymbol{\gamma})\,\mathrm{d}\mathbf{u}_i\right), \tag{2.9}$$

dove $p(\mathbf{u}_i, \boldsymbol{\gamma})$ fa riferimento alla densità degli effetti casuali. Breslow & Clayton (1993)

hanno ottenuto un'approssimazione basata sul metodo di Laplace

$$\ell^{app}(\boldsymbol{\delta}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \log\left(f(\mathbf{y}_i|\boldsymbol{\delta}, \boldsymbol{\gamma})\right) - \frac{1}{2}\mathbf{u}^T \boldsymbol{\Psi}^{-1}\mathbf{u}. \tag{2.10}$$

La quasi-verosimiglianza penalizzata generalmente sfrutta il passaggio alla verosimiglianza profilo. Distingue infatti tra la stima di $\boldsymbol{\delta}$, date le stime $\hat{\boldsymbol{\gamma}}$, restituendo la verosimiglianza profilo $\ell^{app}(\boldsymbol{\delta}, \hat{\boldsymbol{\gamma}})$, e la stima di $\boldsymbol{\gamma}$. Affinché sia possibile fare selezione delle variabili, viene aggiunto un termine di penalità in norma 1 ai coefficienti di regressione fissi, del tipo $\lambda \sum_{j=1}^{p} |\beta_j|$. La log-verosimiglianza penalizzata viene approssimata in maniera analoga a quanto specificato da Breslow & Clayton (1993), ottenendo

$$\ell^{\text{pen}}(\boldsymbol{\beta}, \mathbf{u}, \boldsymbol{\gamma}) = \ell^{\text{pen}}(\boldsymbol{\delta}, \boldsymbol{\gamma}) = \ell^{\text{app}}(\boldsymbol{\delta}, \boldsymbol{\gamma}) - \lambda \sum_{j=1}^{p} |\beta_j|. \tag{2.11}$$

Per un dato $\hat{\boldsymbol{\gamma}}$, il problema di ottimizzazione diventa

$$\hat{\boldsymbol{\delta}} = \arg\max_{\boldsymbol{\delta}} \ \ell^{\text{pen}}(\boldsymbol{\delta}, \hat{\boldsymbol{\gamma}}) = \arg\max_{\boldsymbol{\delta}} \left[\ell^{\text{app}}(\boldsymbol{\delta}, \hat{\boldsymbol{\gamma}}) - \lambda \sum_{j=1}^{p} |\beta_j|\right]. \tag{2.12}$$

Per massimizzare la log-verosimiglianza penalizzata (2.11), viene proposto un algoritmo che combina l'ottimizzazione per gradiente ascendente con l'algoritmo di Fisher scoring, sulla base dell'approccio di Goeman (2010). Il gradiente ascendente è un metodo iterativo utilizzato per individuare il massimo di una funzione. Rappresenta di fatto la controparte del metodo della discesa del gradiente, che cerca di minimizzare una funzione. Ad ogni iterazione, i parametri vengono aggiornati nella direzione del gradiente

$$\boldsymbol{w}_{n+1} = \boldsymbol{w}_n + \alpha \nabla_{\boldsymbol{w}} f(\boldsymbol{w}).$$

Il metodo del gradiente ascendente consente di spostarsi, a ogni iterazione, nella direzione del massimo più vicino. Non viene aggiornata una coordinata alla volta, ma si segue il gradiente della verosimiglianza a partire da un valore iniziale di $\boldsymbol{\delta}$ e si utilizza il gradiente completo a ogni passo. Si tratta di un algoritmo computazionalmente semplice in quanto non richiede inversione di matrici o altri calcoli computazionalmente onerosi, tuttavia è spesso necessario un gran numero di passi per la convergenza. In modo simile a Goeman (2010), l'algoritmo può passare automaticamente a una procedura di Fisher scoring quando si avvicina all'ottimo, evitando così la tendenza alla convergenza lenta tipica degli algoritmi di gradiente ascendente. Pertanto, le stime dei parametri vengono

aggiornate iterativamente mediante l'informazione di Fisher, invece della matrice Hessiana utilizzata nell'algoritmo di Newton Rapson. L'informazione di Fisher è la matrice delle varianze attese delle derivate prime ed è definita come:

$$\mathbf{I}(\boldsymbol{w}) = -E\left[\frac{\partial^2 l(\boldsymbol{w})}{\partial \boldsymbol{w}^2}\right]$$

e pertanto l'aggiornamento dei parametri risulta pari a

$$\boldsymbol{w}_{n+1} = \boldsymbol{w}_n + \mathbf{I}^{-1}\boldsymbol{w}_n \nabla_{\boldsymbol{w}} f(\boldsymbol{w}).$$

Il vantaggio è che l'informazione di Fisher è sempre definita positiva e garantisce maggiore stabilità e robustezza nella massimizzazione della verosimiglianza. È poi necessario un ulteriore passo per stimare le componenti di varianza-covarianza $\boldsymbol{\Psi}$ degli effetti casuali. Il procedimento è riportato nell'Algoritmo 2.

In presenza di predittori categoriali, che sono tipicamente codificati come dummy e quindi risultano in gruppi di variabili dummy, è possibile modificare il passaggio 1(a) dell'Algoritmo 2 in maniera tale da ricondursi alla strategia del Group Lasso. Il Group Lasso è stato introdotto da Yuan & Lin (2006) per superare il fatto che il Lasso non è in grado di selezionare interi fattori ma solamente variabili dummy distinte. Il Group Lasso permette di selezionare interi gruppi di variabili sfruttando una penalità di tipo lasso a livello di fattori e una penalità di tipo ridge all'interno del gruppo di coefficienti. Il termine di penalità lasso viene sostituito con $\sum_{g=1}^{G} \lambda_g \|\boldsymbol{\beta}_{I_g}\|_2$, dove I_g indica l'insieme degli indici del g-esimo gruppo di variabili, $g = 1, ..., G$. Il parametro di regolazione è $\lambda_g = \lambda\sqrt{\mathrm{df}_g}$, con df_g che rappresenta il numero di parametri del gruppo g, pari al numero di livelli del fattore meno uno per i predittori categoriali, mentre è uguale a 1 per i predittori continui. Il calcolo del gradiente della log-verosimiglianza viene modificato nel seguente modo:

(1a) Da $s(\boldsymbol{\delta}) = \partial \ell^{app}(\boldsymbol{\delta})/\partial \boldsymbol{\delta}$ si ottiene:

$$s_0^{pen}(\hat{\boldsymbol{\delta}}^{(l-1)}) = s_0(\hat{\boldsymbol{\delta}}^{(l-1)}), \quad s_i^{pen}(\hat{\boldsymbol{\delta}}^{(l-1)}) = s_i(\hat{\boldsymbol{\delta}}^{(l-1)}), \quad i = p+1, \ldots, p+nq.$$

Inoltre, per $g = 1, \ldots, G$, calcola:

Algoritmo 2

(0) Si calcolano i valori iniziali $\hat{\boldsymbol{\beta}}^{(0)}$, $\hat{\mathbf{u}}^{(0)}$, $\hat{\boldsymbol{\gamma}}^{(0)}$ tramite un modello con sola intercetta ed effetti casuali, del tipo $g(\mu_{ij}) = \beta_0 + \mathbf{z}_{ij}^T\mathbf{u}_i$. Si definisce $\hat{\boldsymbol{\eta}}^{(0)} = \mathbf{X}\hat{\boldsymbol{\beta}}^{(0)} + \mathbf{Z}\hat{\mathbf{u}}^{(0)}$
(1) Per l=1,2,... fino a convergenza si ripetono i seguenti passi:

(a) Si calcola il gradiente della log-verosimiglianza per un dato $\hat{\boldsymbol{\gamma}}^{(l-1)}$.
Da $s(\boldsymbol{\delta}) = \partial\ell^{app}(\boldsymbol{\delta})/\partial\boldsymbol{\delta}$ si ottiene:
$s_0^{\text{pen}}(\hat{\boldsymbol{\delta}}^{(l-1)}) = s_0(\hat{\boldsymbol{\delta}}^{(l-1)}), \quad s_i^{\text{pen}}(\hat{\boldsymbol{\delta}}^{(l-1)}) = s_i(\hat{\boldsymbol{\delta}}^{(l-1)}), \quad i = p+1, \ldots, p+nq$. Inoltre, per $i = 1, ..., p$ si calcola

$$s_i^{\text{pen}}(\hat{\boldsymbol{\delta}}^{(l-1)}) = \begin{cases} s_i(\hat{\boldsymbol{\delta}}^{(l-1)}) - \lambda\,\text{sign}(\hat{\beta}_i^{(l-1)}) & \text{se } \hat{\beta}_i^{(l-1)} \neq 0, \\ s_i(\hat{\boldsymbol{\delta}}^{(l-1)}) - \lambda\,\text{sign}(s_i(\hat{\boldsymbol{\delta}}^{(l-1)})) & \text{se } \hat{\beta}_i^{(l-1)} = 0 \text{ e } |s_i(\hat{\boldsymbol{\delta}}^{(l-1)})| > \lambda, \\ 0 & \text{altrimenti.} \end{cases}$$

(b) Calcolo della derivata seconda direzionale: sia $\mathbf{K} = \text{diag}(0, \ldots, 0, \boldsymbol{\Psi}^{-1}, \ldots, \boldsymbol{\Psi}^{-1})$ una matrice di penalità a blocchi diagonale, con una diagonale di $p+1$ zeri corrispondenti agli effetti fissi e n volte la matrice $\boldsymbol{\Psi}^{-1}$ e sia $\mathbf{A} = [\mathbf{X}, \mathbf{Z}]$. Allora, la matrice di Fisher in forma chiusa è definita come $\mathbf{F}^{pen}(\boldsymbol{\delta}) = \mathbf{A}^T\mathbf{W}(\boldsymbol{\delta})\mathbf{A} + \mathbf{K}$ con $\mathbf{W}(\boldsymbol{\delta}) = \mathbf{D}(\boldsymbol{\delta})\Sigma^{-1}(\boldsymbol{\delta})\mathbf{D}(\boldsymbol{\delta})^T$ e $\mathbf{D}(\boldsymbol{\delta}) = \partial h(\boldsymbol{\eta})/\partial\boldsymbol{\eta}$, $\Sigma(\boldsymbol{\delta}) = cov(\mathbf{y}|\boldsymbol{\delta})$. La derivata seconda direzionale è data, per ogni $\boldsymbol{\delta}$ e ogni vettore direzione $\mathbf{v} \in \mathbb{R}^{p+1+nq}$, da $\ell''_{pen}(\boldsymbol{\delta}; \mathbf{v}) = \mathbf{v}^T\mathbf{F}^{pen}(\boldsymbol{\delta})\mathbf{v}$

(c) Sulla base dell'approssimazione di Taylor usata da Goeman, si ottengono

$$t_{edge}^{(l-1)} = \min_i \left\{ -\frac{\hat{\boldsymbol{\delta}}_i^{(l-1)}}{s_i^{pen}(\hat{\boldsymbol{\delta}}^{(l-1)})} : \text{sign}(\hat{\boldsymbol{\delta}}_i^{(l-1)}) = -\text{sign}[s_i^{pen}(\hat{\boldsymbol{\delta}}^{(l-1)})] \neq 0 \right\},$$

$$t_{opt}^{(l-1)} = \frac{\|s^{pen}(\hat{\boldsymbol{\delta}}^{(l-1)})\|_2}{l''_{app}(\hat{\boldsymbol{\delta}}^{(l-1)}, s^{pen}(\hat{\boldsymbol{\delta}}^{(l-1)}))}$$

(d) Aggiornamento

$$\hat{\boldsymbol{\delta}}^{(l)} = \begin{cases} \hat{\boldsymbol{\delta}}^{(l-1)} + t_{edge}^{(l-1)} s^{pen}(\hat{\boldsymbol{\delta}}^{(l-1)}) & \text{se } t_{opt}^{(l-1)} \geq t_{edge}^{(l-1)}, \\ \hat{\boldsymbol{\delta}}_{NR}^{(l-1)} & \text{se } t_{opt}^{(l-1)} < t_{edge}^{(l-1)} \text{ e } \text{sign}(\hat{\boldsymbol{\delta}}_{NR}^{(l)}) = \text{sign}(\hat{\boldsymbol{\delta}}^{(l-1)}), \\ \hat{\boldsymbol{\delta}}^{(l-1)} + t_{opt}^{(l-1)} s^{pen}(\hat{\boldsymbol{\delta}}^{(l-1)}) & \text{altrimenti.} \end{cases}$$

dove $\hat{\boldsymbol{\delta}}_{NR}^{(l)}$ indica la stima ottenuta tramite Fisher scoring. Si ricava aggiungendo allo stimatore $\hat{\tilde{\boldsymbol{\delta}}}^{(l)}$ degli zeri per tutte le covariate non attive. $\hat{\tilde{\boldsymbol{\delta}}}^{(l)}$ corrisponde all'aggiornamento tramite algoritmo di Fisher scoring nel sottodominio vincolato in cui si considerano solo le variabile attive.

(e) Calcolo delle componenti di varianza-covarianza $\boldsymbol{\theta}^{(l)}$: le stime $\hat{\boldsymbol{\Psi}}^{(l)}$ sono ottenute come stime approssimate tramite un algoritmo EM o tramite metodi alternativi.

(2) Fase di ri-stima: il modello viene ristimato includendo solo le variabili corrispondenti ai coefficienti non nulli in $\hat{\boldsymbol{\beta}}$. Viene utilizzato l'algoritmo di Fisher scoring per ottenere le stime finali $\hat{\boldsymbol{\delta}}$ e $\hat{\boldsymbol{\Psi}}$.

$$s_{I_g}^{pen}(\hat{\boldsymbol{\delta}}^{(l-1)}) = \begin{cases} s_{I_g}(\hat{\boldsymbol{\delta}}^{(l-1)}) - \lambda_g \frac{\hat{\boldsymbol{\beta}}_{I_g}^{(l-1)}}{\|\hat{\boldsymbol{\beta}}_{I_g}^{(l-1)}\|_2} & \text{se } \|\hat{\boldsymbol{\beta}}_{I_g}^{(l-1)}\|_2 \neq 0, \\ \\ s_{I_g}(\hat{\boldsymbol{\delta}}^{(l-1)}) - \lambda_g \frac{s_{I_g}(\hat{\boldsymbol{\delta}}^{(l-1)})}{\|s_{I_g}(\hat{\boldsymbol{\delta}}^{(l-1)})\|_2} & \text{se } \|\hat{\boldsymbol{\beta}}_{I_g}^{(l-1)}\|_2 = 0 \text{ e } \|s_{I_g}(\hat{\boldsymbol{\delta}}^{(l-1)})\|_2 > \lambda_g, \\ \\ 0 & \text{altrimenti.} \end{cases}$$

Vengono suggeriti quattro diversi metodi per la scelta del valore ottimale del parametro di penalizzazione λ:

- 1° Metodo: prevede di utilizzare il BIC per determinare il valore ottimale del parametro λ. Si ottengono delle stime preliminari dell'intercetta fissa, dell'effetto casuale associato a ciascuna unità di 2° livello e della varianza dell'effetto casuale tramite un modello lineare a effetti misti con sola intercetta. Si definisce una griglia equispaziata di possibili valori del parametro di regolazione, e per ciascuno di essi si stima il modello, registrando il valore del BIC associato. Il valore ottimale di λ è quello che minimizza il BIC. Il modello viene ristimato sull'intero insieme di stima utilizzando il valore ottimale di λ;

- 2° Metodo: prevede di utilizzare una convalida incrociata a 5 folds per determinare il valore ottimale di λ. Le osservazioni vengono suddivise casualmente in 5 folds di dimensione approssimativamente uguale. A ogni iterazione, uno dei fold viene impiegato come insieme di verifica, mentre i restanti 4 sono utilizzati per la stima del modello. Per ogni combinazione di λ e fold, si stima il modello e si calcola la devianza sulle osservazioni dell'insieme di verifica. Si seleziona il valore λ che minimizza la somma delle devianze su tutti i folds e lo si utilizza per stimare un modello finale sull'intero insieme di stima;

- 3° Metodo: è una variante del primo metodo che prevede di sfruttare la stima corrente dei parametri come valore iniziale nell'iterazione successiva. Permette di migliorare la stabilità numerica e ridurre i tempi di convergenza. Il primo valore della griglia per λ deve essere fissato in maniera tale da "spingere" a zero tutti i coefficienti eccetto l'intercetta. Si sceglie il valore di λ che minimizza il BIC e si ristima il modello sull'intero insieme di stima;

- 4° Metodo: si tratta di una variante del secondo metodo che prevede di aggiornare iterativamente le stime dei coefficienti, utilizzando a ogni iterazione la stima dei

parametri ottenuta per il valore di λ precedente. Il valore ottimale del parametro di penalizzazione λ viene selezionato minimizzando la somma delle devianze calcolate sui dati di validazione dei 5 folds. Al termine, si ristima il modello sull'insieme di stima utilizzando il valore di λ ottimale.

Gli errori standard per gli effetti fissi e per le componenti di varianza degli effetti casuali possono essere ottenuti tramite una valutazione bootstrap parametrica basata sulla simulazione. In generale, esistono due possibili approcci bootstrap che vengono impiegati nel contesto dei GLMM. Il primo prevede di ricampionare in maniera non parametrica procedendo in due fasi: al primo step si campionano casualmente i gruppi con reinserimento, al secondo step si campionano le osservazioni all'interno dei gruppi con reinserimento. Tuttavia, se i dati mostrano una struttura longitudinale, campionare anche all'interno dei gruppi impedirebbe di preservare la correlazione temporale interna ai gruppi e l'ordine temporale delle osservazioni. Per questo motivo, è preferibile campionare solo i gruppi e mantenere tutte le osservazioni di ciascun gruppo. Il secondo approccio consiste nel simulare campioni bootstrap parametrici, seguendo la distribuzione parametrica della famiglia del modello sottostante. Sia $F_{\boldsymbol{\xi}} : \boldsymbol{\xi} \in \Xi$ la famiglia di distribuzioni parametriche del modello sottostante, dove $\boldsymbol{\xi}^T = (\boldsymbol{\beta}^T, \mathbf{u}^T, \text{vec}(\boldsymbol{\Psi})^T)$ è un vettore ignoto. Con $\text{vec}(\boldsymbol{\Psi})$ si indica la vettorizzazione colonna per colonna della matrice $\boldsymbol{\Psi}$. Sia $\hat{\boldsymbol{\xi}} = (\hat{\boldsymbol{\beta}}^T, \hat{\mathbf{u}}^T, \text{vec}(\hat{\boldsymbol{\Psi}})^T)$ la stima lasso di $\boldsymbol{\xi}$ per un dato valore di penalità λ su un certo dataset. Si simulano nuovi dataset bootstrap $(\mathbf{y}^*, \mathbf{u}^*)$ dalla distribuzione $F_{\hat{\boldsymbol{\xi}}}$, ovvero $(\mathbf{y}^*, \mathbf{u}^*) \sim F_{\hat{\boldsymbol{\xi}}}$. Questa procedura viene ripetuta un numero sufficiente di volte, ad esempio $B = 10000$, e si applica nuovamente l'algoritmo ad ogni nuovo dataset bootstrap $(\mathbf{y}^*_{(r)}, \mathbf{X}, \mathbf{W})$, per $r = 1, \ldots, B$. Le nuove stime $\hat{\boldsymbol{\xi}}^*_{(r)}$ corrispondenti all'r-esimo nuovo dataset fungono da stime bootstrap e possono essere utilizzate per derivare gli errori standard.

2.4.2 Penalità non convesse

Fan & Li (2012) applicano una penalizzazione non convessa nel contesto dei modelli a effetti misti per la selezione sequenziale delle variabili degli effetti fissi e casuali. Questo approccio presenta importanti limitazioni: non permette di stimare simultaneamente i parametri degli effetti fissi e casuali e utilizza una matrice proxy insieme alla verosimiglianza profilo penalizzata per evitare di stimare la matrice di covarianza degli effetti casuali. Sebbene propongano un criterio per scegliere la matrice proxy, è piuttosto difficile capire quale proxy utilizzare in una situazione pratica specifica. Questi limiti

vengono superati da Ghosh & Thoresen (2016), il cui approccio permette di evitare l'utilizzo di una matrice proxy e riduce lo sforzo computazionale stimando simultaneamente i coefficienti degli effetti fissi e i parametri di varianza e covarianza degli effetti casuali. Il metodo si basa sulla verosimiglianza penalizzata e utilizza penalizzazioni non convesse. In particolare, viene considerata la verosimiglianza penalizzata sia per i parametri degli effetti fissi, che per quelli degli effetti casuali $(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_\varepsilon^2)$ e viene massimizzata simultaneamente per ottenere i relativi stimatori di massima verosimiglianza penalizzata (MPL). Inoltre, sono dimostrate la consistenza e la proprietà di oracolo sia nei casi classici a bassa dimensionalità ($p < N$), sia nei casi ad alta dimensionalità, in cui il numero di parametri cresce in modo esponenziale rispetto alla dimensione del campione ($p \gg N$ con $\log P = O(N^\alpha)$ per un certo $\alpha \in (0, 1)$). Inoltre, Ghosh & Thoresen (2016) ricavano la distribuzione asintotica degli stimatori penalizzati nel contesto del modello lineare a effetti misti in alta dimensionalità.

Si considera la minimizzazione della funzione obiettivo (2.13), corrispondente alla log-verosimiglianza negativa penalizzata:

$$Q_{N,\lambda}(\boldsymbol{\beta}, \boldsymbol{\eta}) = -\ell_N(\boldsymbol{\beta}, \boldsymbol{\eta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|), \tag{2.13}$$

che, in forma estesa, è pari a

$$\frac{1}{2}\left[N \log(2\pi) + \log|\mathbf{V}(\boldsymbol{\theta}, \sigma_\varepsilon^2)| + (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}(\boldsymbol{\theta}, \sigma_\varepsilon^2)^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] + \sum_{j=1}^{p} P_{N,\lambda}(|\beta_j|).$$

$P_{N,\lambda}(\cdot) = N p_{\lambda_N}(\cdot)$ rappresenta una generale classe di funzioni di penalità non negative, che può dipendere dalla dimensione campionaria N oltre che da λ. Si tratta di un problema di ottimizzazione non convesso dato che la log-verosimiglianza è convessa solo rispetto a $\boldsymbol{\beta}$ e non convessa rispetto a $\boldsymbol{\eta}$. Pertanto, per ottenere le stime MLP servono tecniche di approssimazione quadratica e algoritmi iterativi.

2.4.2.1 Risultati teorici

Si considera inizialmente il caso con $p \leq N$ e si ri-etichettano le osservazioni originali $\{y_{ij}, \mathbf{x}_{ij}\}_{j=1,\ldots,m_i; i=1,\ldots,n}$ come $\{y_k, \mathbf{X}_k\}_{k=1,\ldots,N}$. Sia D_k l'identificativo del gruppo per la k-esima osservazione con D variabile casuale sottostante. Sia X una variabile stocastica con $Z \subset X$, tale per cui le osservazioni $V_k = (y_k, \mathbf{X}_k, D_k)$, $k = 1, ..., N$, sono N realizzazioni indipendenti e identicamente distribuite delle variabili $(Y, \mathbf{X}, D)$. Inoltre, per ogni distribuzione regolare delle covariate, devono valere le seguenti assunzioni sul modello:

1. Il modello è identificabile e il supporto di $f(V;\boldsymbol{\beta},\boldsymbol{\eta})$ è indipendente dai parametri $(\boldsymbol{\beta},\boldsymbol{\eta})$. Inoltre, la densità $f(V;\boldsymbol{\beta},\boldsymbol{\eta})$ possiede derivate di primo e secondo ordine che soddisfano
$$E\left[\frac{\partial \log f(V;\boldsymbol{\beta},\boldsymbol{\eta})}{\partial(\boldsymbol{\beta},\boldsymbol{\eta})}\right]=0,$$
e
$$\mathbf{I}(\boldsymbol{\beta},\boldsymbol{\eta})=E\left[\left(\frac{\partial \log f(V;\boldsymbol{\beta},\boldsymbol{\eta})}{\partial(\boldsymbol{\beta},\boldsymbol{\eta})}\right)\left(\frac{\partial \log f(V;\boldsymbol{\beta},\boldsymbol{\eta})}{\partial(\boldsymbol{\beta},\boldsymbol{\eta})}\right)^T\right]=E\left[-\frac{\partial^2 \log f(V;\boldsymbol{\beta},\boldsymbol{\eta})}{\partial(\boldsymbol{\beta},\boldsymbol{\eta})^2}\right].$$

2. Esiste un sottoinsieme aperto dello spazio dei parametri contenente i veri valori, su cui la densità $f(v;\boldsymbol{\beta},\boldsymbol{\eta})$ ammette tutte le derivate parziali fino al terzo ordine per quasi tutti i v che sono uniformemente limitati, su tale sottoinsieme aperto, da alcune funzioni aventi valore atteso finito sotto il vero valore dei parametri.

Se il valore atteso $E(\mathbf{X}^T\mathbf{X})$ è finito e definito positivo sotto la vera distribuzione, allora vale anche la seguente assunzione

3. La matrice d'informazione di Fisher $\mathbf{I}(\boldsymbol{\beta},\boldsymbol{\eta})$ è finita e definita positiva in $(\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)$.

Si definiscono le quantità

$$a_N=\max\{p'_{\lambda_N}(|\beta_{0j}|):\beta_{0j}\neq 0\},\quad b_N=\max\{p''_{\lambda_N}(|\beta_{0j}|):\beta_{0j}\neq 0\}$$

Si assume che il vero vettore di parametri $\boldsymbol{\beta}_0$ sia sparso e dato da $\boldsymbol{\beta}_0=(\beta_{01},...,\beta_{0p})^T=(\boldsymbol{\beta}_0^{(1)T},\boldsymbol{\beta}_0^{(2)T})^T$, dove $\boldsymbol{\beta}_0^{(1)}$ ha dimensione $s\ll p$ e $\boldsymbol{\beta}_0^{(2)}=\mathbf{0}_{p-s}$ è un vettore di tutti zeri. La medesima partizione può essere applicata al vettore di parametri stimati $\boldsymbol{\beta}$, restituendo $\boldsymbol{\beta}=(\boldsymbol{\beta}^{(1)T},\boldsymbol{\beta}^{(2)T})^T$. In questo caso, $\boldsymbol{\beta}^{(2)}$ non è necessariamente costituito da soli zeri.

Se valgono le assunzioni sul modello e $P_{N,\lambda}(\cdot)=Np_{\lambda_N}(\cdot)$, gli stimatori penalizzati possiedono le seguenti proprietà:

1. Se $b_N\to 0$ per $N\to\infty$, allora esiste un minimo locale $(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\eta}})=((\hat{\boldsymbol{\beta}}^{(1)T},\hat{\boldsymbol{\beta}}^{(2)T})^T,\hat{\boldsymbol{\eta}})$ di (2.13) che soddisfa
$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|=O_p(N^{-1/2}+a_N),\quad \|\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0\|=O_p(N^{-1/2}).$$
Ciò mostra l'esistenza di uno stimatore penalizzato di $(\boldsymbol{\beta},\boldsymbol{\eta})$ come minimo della funzione obiettivo (2.13) che converge al vero valore con velocità $O_p(N^{-1/2}+a_N)$ per $\boldsymbol{\beta}$ e $O_p(N^{-1/2})$ per $\boldsymbol{\eta}$. Pertanto, nel caso della penalità di tipo SCAD, se $\lambda_N\to 0$, $a_N\to 0$ e quindi lo stimatore è $\sqrt{N}$-consistente.

2. Se $\liminf_{N\to\infty} \liminf_{\beta\downarrow 0+} \frac{p'_{\lambda_N}(\beta)}{\lambda_N} > 0$ con $\lambda_N \to 0$ e $\sqrt{N}\lambda_N \to \infty$ per $N \to \infty$, si ha che $\hat{\boldsymbol{\beta}}^{(2)} = 0$ con probabilità che tende a 1 e le distribuzioni asintotiche di $\hat{\boldsymbol{\beta}}^{(1)}$ e $\hat{\boldsymbol{\eta}}$ sono date da

$$\sqrt{N}(\mathbf{I}_1(\boldsymbol{\beta}_0^{(1)}, \boldsymbol{\eta}_0) + \boldsymbol{\Sigma}) \left\{ \hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)} + (\mathbf{I}_1(\boldsymbol{\beta}_0^{(1)}, \boldsymbol{\eta}_0) + \boldsymbol{\Sigma})^{-1} \boldsymbol{\zeta} \right\} \xrightarrow{D} N_s(0, \mathbf{I}_1(\boldsymbol{\beta}_0^{(1)}, \boldsymbol{\eta}_0)),$$

$$\sqrt{N}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \xrightarrow{D} N_{q^*+1}(0, \mathbf{I}_2(\boldsymbol{\beta}_0^{(1)}, \boldsymbol{\eta}_0)^{-1}),$$

rispettivamente. Le quantità $\boldsymbol{\Sigma}$ e $\boldsymbol{\zeta}$ sono definite come

$$\boldsymbol{\Sigma} = \mathrm{diag}\{p''_{\lambda_N}(|\beta_{01}|), \ldots, p''_{\lambda_N}(|\beta_{0s}|)\},$$

$$\boldsymbol{\zeta} = (p'_{\lambda_N}(|\beta_{01}|)\,\mathrm{sign}(\beta_{01}), \ldots, p'_{\lambda_N}(|\beta_{0s}|)\,\mathrm{sign}(\beta_{0s}))^T.$$

Inoltre, $\mathbf{I}_1(\boldsymbol{\beta}_0^{(1)}, \boldsymbol{\eta}_0)$ e $\mathbf{I}_2(\boldsymbol{\beta}_0^{(1)}, \boldsymbol{\eta}_0)$ sono le matrici di informazione di Fisher corrispondenti a $\boldsymbol{\beta}^{(1)}$ e $\boldsymbol{\eta}$, assumendo $\boldsymbol{\beta}^{(2)} = 0$. In particolare,

$$\mathbf{I}_1(\boldsymbol{\beta}_0^{(1)}, \boldsymbol{\eta}_0) = \sum_{i=1}^{\dim(\boldsymbol{\beta}_0^{(1)})} E\left[\mathbf{X}_i^T \mathbf{V}_i(\boldsymbol{\eta}_0)^{-1} \mathbf{X}_i\right].$$

Anche $\mathbf{I}_2(\boldsymbol{\beta}_0^{(1)}, \boldsymbol{\eta}_0)$ può essere derivata esplicitamente sulla base della struttura assunta per $\boldsymbol{\Psi}$. Tale risultato ha una duplice utilità: da un lato mostra la proprietà di oracolo dello stimatore penalizzato, dall'altro presenta la distribuzione asintotica degli elementi non nulli di $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\eta}}$. In questo modo, è possibile ricavare gli errori standard della stima sparsa di $\boldsymbol{\beta}$ e dei parametri di varianza $\boldsymbol{\eta}$.

Si considera ora il caso ad elevata dimensionalità, in cui $p > N$. Si assume sempre che le osservazioni $V_k = (y_k, \mathbf{X}_k, D_k),\ k = 1, ..., N$, sono indipendenti e identicamente distribuite e si definisce $g(V_k; \boldsymbol{\beta}, \boldsymbol{\eta}) = \sum_{i=1}^n I(D_k = i) a_{ij_k}$ dove j_k denota l'indice j della k-esima osservazione e a_{ij} denota il j-esimo elemento del vettore $a_i = (\mathbf{y}_i - \mathbf{X}_i^T \boldsymbol{\beta}) \mathbf{V}_i(\boldsymbol{\eta})^{-1}$. Si assume che:

(A1) Esistono costanti $b_1, b_2 > 0$ e $r_1, r_2 > 0$ tali che, per ogni $t > 0$,

$$P\left(|g(Y, \mathbf{X}^T\boldsymbol{\beta}, \boldsymbol{\eta})| > t\right) \le \exp\left(-(t/b_1)^{r_1}\right), \quad \max_{l \le p} P\left(|X^{(l)}| > t\right) \le \exp\left(-(t/b_2)^{r_2}\right),$$

dove $X^{(l)}$ denota la l-esima coordinata del vettore di covariate $\mathbf{X}$.

(A2) $\min_{j\in S} Var\left(g(Y, \mathbf{X}^T\boldsymbol{\beta}, X^{(j)})\right)$ è limitata inferiormente lontano da zero per tutti i $j = 1, \ldots, p$.

(A3) $Var(X^{(j)})$ è limitata inferiormente lontano da zero e superiormente lontano da infinito in maniera uniforme per $j = 1, \ldots, p$.

(A4) Gli autovalori delle matrici $\mathbf{I}_1(\boldsymbol{\beta}_0^{(1)}, \boldsymbol{\eta}_0)$ e $\mathbf{I}_2(\boldsymbol{\beta}_0^{(1)}, \boldsymbol{\eta}_0)$ sono limitati in maniera tale da essere distanti sia da zero che da infinito.

Tali assunzioni valgono per il modello lineare a effetti misti con una scelta adeguata per la distribuzione delle covariate. Si fanno inoltre le seguenti assunzioni sulla funzione di penalità $P_{N,\lambda}(\cdot)$ e sulla funzione di perdita.

Assunzioni sulla penalità

(P1) $P_{N,\lambda}(0) = 0$;

(P2) $P_{N,\lambda}(t)$ è concava e non decrescente su $[0, \infty)$ e ha derivata continua $P'_{N,\lambda}(t)$ su $(0, \infty)$;

(P3) $\sqrt{s}P'_{N,\lambda}(d_N) = o(d_N)$, dove s è il numero di elementi non nulli di $\boldsymbol{\beta}$ e $d_N = \frac{1}{2}\min\{|\beta_{0j}| : \beta_{0j} \neq 0,\ j = 1, \ldots, p\}$ misura l'intensità del segnale;

(P4) Sia $\boldsymbol{\beta}_{S_0}$ il vero valore del vettore di parametri $\boldsymbol{\beta}_S$, corrispondenti alle sole variabili attive. Esiste una costante $c > 0$ tale che $\sup_{\boldsymbol{\beta} \in B(\boldsymbol{\beta}_{S_0}, cd_N)} \zeta(\boldsymbol{\beta}) = o(1)$, dove

$$\zeta(\boldsymbol{\beta}) = \limsup_{\epsilon \to 0+} \max_{j \leq s} \sup_{t_1 < t_2 : (t_1, t_2) \in (|\beta_j| - \epsilon, |\beta_j| + \epsilon)} - \left[\frac{P_{N,\lambda}(t_2) - P_{N,\lambda}(t_1)}{t_2 - t_1} \right].$$

(P5) $P'_{N,\lambda}(d_N) = o(1/\sqrt{Ns})$, $\quad P'_{N,\lambda}(d_N)s^2 = O(1)$, $\quad s\sqrt{\log p/N} = o(d_N)$, e $sP'_{N,\lambda}(d_N) + s\sqrt{\log p/N} + s^3 \log s/N = o(P'_{N,\lambda}(0+))$;

(P6) $\sup_{\|\boldsymbol{\beta} - \boldsymbol{\beta}_{S_0}\| \leq d_N/4} \zeta(\boldsymbol{\beta}) = o(1/\sqrt{s \log p})$;

(P7) $\max_{j \notin S} \|\mathbf{X}_S^T \mathbf{V}_S^{-1} X^{(j)}\| \sqrt{\log s/N} = o(P_N(0+))$, dove $\mathbf{X}_S$ si riferisce alle covariate corrispondenti a $\boldsymbol{\beta}_S$ e $\mathbf{V}_S$ rappresenta la somma delle matrici $\mathbf{V}_i$ associate.

Assunzione sulla funzione di perdita (L1)

$L_N(\boldsymbol{\beta}_S, \mathbf{0}; \boldsymbol{\eta})$ è differenziabile due volte rispetto a $\boldsymbol{\beta}_S$ e $\boldsymbol{\eta}$ in un intorno dei veri valori $(\boldsymbol{\beta}_{S_0}, \mathbf{0}; \boldsymbol{\eta}_0)$ ed esistono successioni di numeri reali positivi $a_N = o(d_N)$ e $c_N = o(1)$ tali che valgano le seguenti condizioni:

(i) $\|\nabla_S L_N(\boldsymbol{\beta}_{S_0}, \mathbf{0}; \boldsymbol{\eta}_0)\| = O_p(a_N)$ e $\|\nabla_{\boldsymbol{\eta}} L_N(\boldsymbol{\beta}_{S_0}, \mathbf{0}; \boldsymbol{\eta}_0)\| = O_p(c_N)$;

(ii) Per ogni $\epsilon > 0$, esiste una costante positiva C_ϵ tale che, per ogni N sufficientemente grande

$$P\left(\lambda_{min}(\nabla^2 L_N(\boldsymbol{\beta}_{S_0}, \mathbf{0}; \boldsymbol{\eta}_0)) > C_\epsilon\right) > 1 - \epsilon;$$

(iii) Per ogni $\epsilon > 0$, $\delta > 0$ e successioni non negative $\alpha_N = o(d_N)$ e $\gamma_N = o(1)$, esiste un intero N^* sufficientemente grande tale che, per ogni $N > N^*$

$$P\left(\sup_{\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_{S_0}\| \leq \alpha_n, \|\boldsymbol{\eta} - \boldsymbol{\eta}_0\| \leq \gamma_N} \|\nabla^2 L_N(\boldsymbol{\beta}_S, \mathbf{0}; \boldsymbol{\eta}) - \nabla^2 L_N(\boldsymbol{\beta}_{S_0}, \mathbf{0}; \boldsymbol{\eta}_0)\| \leq \delta\right) > 1 - \epsilon.$$

Teorema 3: Consistenza oracolo
Sotto le assunzioni (P1)-(P4) e (L1), esiste un minimo locale $\left(\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_S^T, \mathbf{0}^T), \hat{\boldsymbol{\eta}}^T\right)$ di

$$Q_{N,\lambda}(\boldsymbol{\beta}_S, \mathbf{0}; \boldsymbol{\eta}) = L_N(\boldsymbol{\beta}_S, \mathbf{0}; \boldsymbol{\eta}) + \sum_{j \in S} P_{N,\lambda}(|\beta_j|)$$

che soddisfa

$$\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{S_0}\| = O_p(a_N + \sqrt{s} P'_{N,\lambda}(d_N)), \quad \|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\| = O_p(c_N).$$

Inoltre, per ogni $\epsilon > 0$, il minimo locale $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}})$ è strettamente definito con probabilità almeno $1 - \epsilon$ per N sufficientemente grande.

Assunzione sulla funzione di perdita (L2)
Per il minimo locale $(\hat{\boldsymbol{\beta}}_S, \hat{\boldsymbol{\eta}})$ ottenuto nel Teorema 3, esiste un intorno $\mathcal{H} \subset \mathbb{R}^{p+q^*+1}$ di $(\hat{\boldsymbol{\beta}}_S^T, \mathbf{0}, \hat{\boldsymbol{\eta}}^T)^T$ tale che, con probabilità che tende a 1, si ha:

$$L_N(T\boldsymbol{\beta}, \boldsymbol{\eta}) - L_N(\boldsymbol{\beta}, \boldsymbol{\eta}) < \sum_{j \notin S} P_{N,\lambda}(|\beta_j|),$$

per ogni $\boldsymbol{\beta} = (\boldsymbol{\beta}_S^T, \boldsymbol{\beta}_N^T)^T$ con $(\boldsymbol{\beta}^T, \boldsymbol{\eta}^T) \in \mathcal{H}$ e $\boldsymbol{\beta}_N \neq \mathbf{0}$. La quantità $T\boldsymbol{\beta}$ rappresenta la proiezione di $\boldsymbol{\beta}$ nello spazio generato da S, ossia $T\boldsymbol{\beta} = (\beta'_1, \ldots, \beta'_p)^T$, con $\beta'_j = \beta_j$ se $j \in S$ e 0 altrimenti.
Le assunzioni (A1)-(A4) implicano l'assunzione (L1) con $a_N = \sqrt{s \log(p)/N}$ e $c_N = 1/\sqrt{N}$. Insieme a (P5), (P6) e (P7), implicano l'assunzione (L2).

Teorema 4: Ottimalità nella selezione di variabili
Sotto le Assunzioni (P), (L1) e (L2), valgono le seguenti affermazioni:

(i) $(\hat{\boldsymbol{\beta}}_S, \mathbf{0}, \hat{\boldsymbol{\eta}})$ ottenuto nel Teorema 3 è un minimo locale in $\mathbb{R}^{p+q^*+1}$ della funzione obiettivo (2.13), con probabilità che tende a uno.

(ii) Per ogni $\epsilon > 0$, il minimo locale $(\hat{\boldsymbol{\beta}}_S, \mathbf{0}, \hat{\boldsymbol{\eta}})$ è strettamente definito con probabilità almeno $1 - \epsilon$ per N sufficientemente grande.

In sintesi, se si assume che siano valide le assunzioni (A1)-(A4) e (P1)-(P7), allora esiste un minimo locale $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\eta}}) = ((\hat{\boldsymbol{\beta}}^{(1)T}, \hat{\boldsymbol{\beta}}^{(2)T})^T, \hat{\boldsymbol{\eta}})$ di (2.13) che soddisfa:

1. $\lim_{N\to\infty} P(\hat{\boldsymbol{\beta}}^{(2)} = \mathbf{0}) = 1$. Inoltre, il minimo locale è strettamente definito con probabilità vicina a uno per N sufficientemente grande.

2. Assumendo che $\hat{S} = \{j \leq p : \hat{\beta}_j \neq 0\}$ sia il set attivo stimato, si ottiene che $\lim_{N\to\infty} P(\hat{S} = S) = 1$.

3. Per ogni vettore unitario $\boldsymbol{\alpha} \in \mathbb{R}^s$,
$\sqrt{N}\boldsymbol{\alpha}^T \mathbf{I}_1(\boldsymbol{\beta}_0^{(1)}, \boldsymbol{\eta}_0)^{1/2}(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)}) \xrightarrow{D} N(0,1), \quad \sqrt{N}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0) \xrightarrow{D} N_{q^*+1}(\mathbf{0}, \mathbf{I}_2(\boldsymbol{\beta}_0^{(1)}, \boldsymbol{\eta}_0)^{-1})$.

Le assunzioni (P5)-(P7) possono essere considerate valide per la penalizzazione SCAD se il segnale d_N è sufficientemente forte e s è piccolo rispetto alla dimensione campionaria totale N, in modo tale che

$$s\frac{\log p}{N} + \frac{s^3 \log s}{N} \ll \lambda_N \ll d_N.$$

2.4.2.2 Algoritmo computazionale

La non convessità del problema di ottimizzazione richiede l'utilizzo di algoritmi iterativi per ottenere le stime MPL in maniera efficiente. In particolare, Ghosh & Thoresen (2016) utilizzano un'approssimazione locale quadratica della funzione obiettivo $Q_{N,\lambda}(\boldsymbol{\beta}, \boldsymbol{\eta})$ e ricorrono all'algoritmo di Newton-Raphson. Inoltre, per ottenere maggiore efficienza computazionale quando p è elevato, combinano questo approccio con una versione dell'algoritmo di coordinate descent. Ottenere un'approssimazione quadratica del termine relativo alla funzione di perdita, cioè $L_N(\boldsymbol{\beta}, \boldsymbol{\eta})$, è piuttosto semplice se si assume che sia una funzione liscia rispetto a entrambi i parametri, con derivate parziali di secondo ordine continue. L'approssimazione quadratica rispetto a $(\boldsymbol{\beta}, \boldsymbol{\eta})$ viene ottenuta usando l'approccio in serie di Taylor.

$$\begin{aligned} L_N(\boldsymbol{\beta}, \boldsymbol{\eta}) \approx\ & L_N(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0) + \nabla_{\boldsymbol{\beta}} L_N(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)^T(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \nabla_{\boldsymbol{\eta}} L_N(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)^T(\boldsymbol{\eta} - \boldsymbol{\eta}_0) \\ & + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \nabla^2_{\boldsymbol{\beta}} L_N(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)(\boldsymbol{\beta} - \boldsymbol{\beta}_0) + \frac{1}{2}(\boldsymbol{\eta} - \boldsymbol{\eta}_0)^T \nabla^2_{\boldsymbol{\eta}} L_N(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)(\boldsymbol{\eta} - \boldsymbol{\eta}_0) \\ & + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T \nabla_{\boldsymbol{\beta}\boldsymbol{\eta}} L_N(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)(\boldsymbol{\eta} - \boldsymbol{\eta}_0), \end{aligned}$$

dove $\nabla_{\boldsymbol{\beta}} L_N(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$ e $\nabla_{\boldsymbol{\eta}} L_N(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$ sono le derivate parziali di primo ordine rispetto a $\boldsymbol{\beta}$ e a $\boldsymbol{\eta}$ di $L_N(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$, mentre $\nabla^2_{\boldsymbol{\beta}} L_N(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$ e $\nabla^2_{\boldsymbol{\eta}} L_N(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$ sono le corrispondenti derivate parziali di secondo ordine e $\nabla_{\boldsymbol{\beta}\boldsymbol{\eta}} L_N(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)$ è la derivata parziale incrociata di secondo ordine rispetto a $\boldsymbol{\beta}$ e $\boldsymbol{\eta}$. Tuttavia, funzioni di penalità come Lasso e SCAD non hanno derivate di secondo ordine continue nell'origine e quindi non si può utilizzare l'approccio in serie di Taylor per ottenere l'approssimazione quadratica delle funzioni di penalità. Pertanto, viene utilizzata l'approssimazione locale quadratica delle penalità come descritta da Fan & Li (2001). L'approssimazione locale quadratica si ottiene combinando l'approssimazione locale con l'espansione in serie di Taylor. Dato un valore iniziale $\boldsymbol{\beta}_0$, vicino al minimo reale, se la sua j-esima componente β_{0j} non è molto vicina allo zero, si può usare l'approssimazione locale

$$[p_\lambda(|\beta_j|)]' = p'_\lambda(|\beta_j|)\,\mathrm{sign}(\beta_j) \approx \left\{ \frac{p'_\lambda(|\beta_j|)}{|\beta_j|} \right\} \beta_j$$

e fissare $\hat{\beta}_j = 0$ se β_{0j} è molto vicino a zero. Tale approssimazione viene combinata con l'espansione in serie di Taylor, ottenendo l'approssimazione locale quadratica per la funzione di penalità, definita come

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{0j}|) + \frac{1}{2} \left\{ \frac{p'_\lambda(|\beta_j|)}{|\beta_j|} \right\} (\beta_j^2 - \beta_{0j}^2),$$

per $\beta_j \approx \beta_{0j}$. Quindi, la funzione obiettivo generale può essere approssimata localmente da una funzione quadratica, la quale può essere minimizzata tramite il metodo di Newton-Raphson se il numero di parametri è piccolo. Si ottiene la seguente soluzione iterativa

$$\begin{aligned}
\boldsymbol{\beta}_1 &= \boldsymbol{\beta}_0 - [\nabla^2_{\boldsymbol{\beta}} L_N(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0) + \boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)]^{-1} \{\nabla_{\boldsymbol{\beta}} L_N(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0) + \boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}_0)\boldsymbol{\beta}_0\}, \\
\boldsymbol{\eta}_1 &= \boldsymbol{\eta}_0 - [\nabla^2_{\boldsymbol{\eta}} L_N(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)]^{-1} \{\nabla_{\boldsymbol{\eta}} L_N(\boldsymbol{\beta}_0, \boldsymbol{\eta}_0)\},
\end{aligned}$$

dove $\boldsymbol{\Sigma}_\lambda(\boldsymbol{\beta}) = \mathrm{diag}\left\{ \frac{p'_\lambda(|\beta_j|)}{|\beta_j|} \right\}_{j=1,\ldots,p}$. Per ottenere il valore che minimizza la funzione obiettivo (2.13) si può iterare sequenzialmente tra $\boldsymbol{\beta}$ e $\boldsymbol{\eta}$ fino a convergenza. L'algoritmo funziona quando il numero di parametri p è più piccolo della dimensione campionaria N e converge in maniera abbastanza efficiente per diverse funzioni di penalità. Tuttavia, nei contesti ad elevata dimensionalità, è richiesta l'inversione di una matrice molto grande, che comporta il fallimento dell'aggiornamento di $\boldsymbol{\beta}$, o, nei casi meno gravi, instabilità e inefficienza. Pertanto, si modifica l'algoritmo in maniera tale da impiegare una strategia di coordinate descent per l'aggiornamento di $\boldsymbol{\beta}$. In particolare, si ricorre all'approccio

di Schelldorfer et al. (2011), che prevede di implementare un algoritmo di coordinate gradient descent con penalità, incorporando l'ottimizzazione di parametri di varianza addizionali $\boldsymbol{\eta}$. Il procedimento è presentato nell'Algoritmo 3.

Algoritmo 3

(0) Si parte da un valore iniziale $\boldsymbol{\beta}^0$ e $\boldsymbol{\eta}^0$;
Si ripetono i seguenti passi fino a convergenza:
(1) Per $j = 1, 2, \ldots, p$,

(a) Si approssima la derivata seconda $\frac{\partial^2}{\partial \beta_j^2} Q_{N,\lambda}(\boldsymbol{\beta}, \boldsymbol{\eta})$ tramite il valore

$$h_j = \min\{\max\{I_{jj}, c_{\min}\}, c_{\max}\},$$

per alcune costanti opportune $c_{\min}$ e $c_{\max}$. Qui I_{jj} denota l'elemento diagonale j-esimo della matrice d'informazione di Fisher $\mathbf{I}(\boldsymbol{\beta}, \boldsymbol{\eta})$. (Schelldorfer et al. (2011) suggeriscono di impostare $c_{\min} = 10^{-6}$ e $c_{\max} = 10^8$);

(b) Si calcola la direzione del gradiente d_j minimizzando, rispetto a $d \in \mathbb{R}$, la seguente funzione

$$L_N(\boldsymbol{\beta}, \boldsymbol{\eta}) + \frac{\partial}{\partial \beta_j} L_N(\boldsymbol{\beta}, \boldsymbol{\eta}) d + \frac{1}{2} d^2 h_j + P_{N,\lambda}(|\beta_j + d|);$$

(c) Si sceglie un passo $\alpha_j > 0$ tale che $Q_{N,\lambda}(\boldsymbol{\beta} + \alpha_j d_j \boldsymbol{e}_j, \boldsymbol{\eta}) < Q_{N,\lambda}(\boldsymbol{\beta}, \boldsymbol{\eta})$, dove $\boldsymbol{e}_j$ è il vettore unitario j-esimo. Si aggiorna la componente j-esima di $\boldsymbol{\beta}$ come

$$\beta_j \leftarrow \beta_j + \alpha_j d_j.$$

Come in Schelldorfer et al. (2011), α_j può essere scelto con la regola di Armijo;

(2) Per $j = 1, 2, \ldots, q^* + 1$, si aggiorna η_j con il valore che minimizza $L_N(\boldsymbol{\beta}, \boldsymbol{\eta})$ rispetto a η_j, con $\boldsymbol{\beta}$ e $\boldsymbol{\eta}$ aggiornati dagli step precedenti.

L'Algoritmo 3 dipende fortemente dai valori di partenza e potrebbe non convergere all'ottimo globale a causa della non convessità della funzione obiettivo. Tuttavia, la convergenza dell'algoritmo a un ottimo locale è garantita. Una proposta è quella di utilizzare come stima iniziale $\boldsymbol{\beta}^0$ la soluzione al problema di lasso ordinario che non considera la struttura a effetti misti. Come valore iniziale $\boldsymbol{\eta}^0$ si può utilizzare la stima ottenuta tramite iterazione Gauss-Seidel sulla base del principio di massima verosimiglianza.

Per la penalità SCAD, la quantità d_j definita allo step 1(b) dell'algoritmo 3 può essere calcolata analiticamente. Ogni volta che β_j non è soggetta a penalità, ha la forma

$$d_j = -\frac{1}{h_j} \frac{\partial}{\partial \beta_j} L_N(\boldsymbol{\beta}, \boldsymbol{\eta}),$$

mentre se β_j è soggetto a penalizzazione tramite la penalità SCAD con parametri di regolarizzazione λ e γ, allora la soluzione d_j è data da

$$d_j = \begin{cases} -\frac{1}{h_j}\left[\lambda + \frac{\partial}{\partial \beta_j} L_N(\boldsymbol{\beta}, \boldsymbol{\eta})\right] & \text{se } \beta_j h_j - \frac{\partial}{\partial \beta_j} L_N(\boldsymbol{\beta}, \boldsymbol{\eta}) \leq \lambda(h_j + 1), \\ -\frac{1}{h_j}\frac{\partial}{\partial \beta_j} L_N(\boldsymbol{\beta}, \boldsymbol{\eta}) & \text{se } \beta_j h_j - \frac{\partial}{\partial \beta_j} L_N(\boldsymbol{\beta}, \boldsymbol{\eta}) > \lambda(\gamma h_j j), \\ -\frac{1}{(\gamma-1)h_j - 1}\left[(\gamma - 1)\frac{\partial}{\partial \beta_j} L_N(\boldsymbol{\beta}, \boldsymbol{\eta}) - (\gamma\lambda - \beta_j)\right] & \text{altrimenti.} \end{cases}$$

Inoltre, se $h_j = \mathcal{I}_{jj}$, ossia si usa la diagonale della matrice d'informazione di Fisher come approssimazione della derivata seconda, e non si applica il troncamento, si può ottenere una soluzione analitica per l'aggiornamento di β_j. Per i modelli lineari a effetti misti è dato da $h_j = \mathbf{x}_j^T \mathbf{V}^{-1} \mathbf{x}_j$ e $\hat{\gamma}((y - \tilde{y})\mathbf{V}^{-1}\mathbf{x}_j / h_j)$ dove $\tilde{y}$ è il valore marginale predetto di y sulla base di tutti gli effetti fissi eccetto il j-esimo.

La scelta del parametro di regolazione λ si basa sul criterio BIC. Anche il parametro γ può essere selezionato minimizzando il BIC; tuttavia, per semplicità, si preferisce fissare $\gamma = 3.7$, analogamente a quanto proposto da Fan & Li (2001) nel contesto del modello di regressione lineare.

Capitolo 3

Approfondimento metodologico

3.1 Introduzione

In questo capitolo, ci si concentra sul verificare empiricamente alcuni aspetti dei metodi discussi nel Capitolo 2. In particolare, vengono posti due obiettivi. Per prima cosa, si vuole analizzare l'affidabilità dell'approccio proposto da Schelldorfer et al. (2011) per individuare le variabili associate anche ad un effetto casuale. Il secondo interesse consiste nel verificare le prestazioni dell'algoritmo per la selezione e la stima degli effetti fissi con penalità SCAD.

3.2 Obiettivo I

Nella sezione 2.4.1.2 viene presentata una strategia per identificare le variabili che presentano anche un effetto casuale. Tale approccio viene qui implementato in `R`, e il relativo codice è riportato in Appendice A.3.1. La sua affidabilità viene verificata mediante uno studio di simulazione condotto su 100 repliche indipendenti. Per ogni replica si generano 40 gruppi sbilanciati di osservazioni, ciascuno con numerosità compresa tra 3 e 10, e si costruisce la matrice del disegno $\mathbf{X}$, formata da una prima colonna di 1, corrispondente all'intercetta, e da (p-1) colonne ottenute da una distribuzione normale multivariata con media nulla e matrice di covarianza pari all'identità, con $p = 100$. Le covariate sono pertanto indipendenti tra loro. Solamente quattro predittori, inclusa l'intercetta, sono associati a un effetto fisso diverso da zero; sono posizionati agli indici $\{1, 5, 10, 98\}$, e hanno coefficienti veri pari, rispettivamente, a (1, -3, 2.1, 0.8). Tra queste variabili, quelle in posizione 1 e 10 sono associate anche ad un effetto casuale. La matrice di

covarianza degli effetti casuali è caratterizzata da varianze $\theta_1^2 = 0.7$ e $\theta_2^2 = 1.3$, rispettivamente per l'intercetta e la pendenza casuale. Per ciascun gruppo e per ciascuna covariata con effetto casuale, viene ricavato un coefficiente casuale da una distribuzione normale con media nulla e varianza θ_j^2, $j = 1, 2$. La variabile risposta viene generata secondo il modello lineare a effetti misti, combinando tre componenti principali: la parte fissa, la parte casuale e il rumore. La componente fissa si ottiene moltiplicando la matrice $\mathbf{X}$ per il vettore dei veri coefficienti $\boldsymbol{\beta}_0$, mentre la componente casuale si ricava moltiplicando i coefficienti casuali per i corrispondenti predittori e sommando i contributi a livello di gruppo. Infine, il termine di rumore viene generato da una variabile normale con media nulla e varianza $\sigma_\varepsilon^2 = 0.5$. Su ciascun dataset così ottenuto viene applicata la funzione per individuare le variabili a cui attribuire un effetto casuale. Si calcola il numero di effetti casuali correttamente catturati (*Veri positivi* - TP), quelli erroneamente selezionati (*Falsi positivi* - FP) e quelli non individuati (*Falsi negativi* - FN). A partire da queste quantità, vengono ricavate le seguenti metriche:

$$\text{Precisione}=\frac{\text{TP}}{\text{TP+FP}}, \quad \text{Recupero}=\frac{\text{TP}}{\text{TP+FN}}.$$

La Precisione rappresenta una misura di affidabilità perché restituisce la proporzione di volte che un positivo è vero. Il Recupero misura quanti dei veri positivi sono stati individuati. A partire da queste due quantità si ottiene una misura riassuntiva, definita dalla media armonica di Precisione e Recupero, ovvero l'F_1-score

$$\text{F}_1=2\frac{\text{Precisione} \cdot \text{Recupero}}{\text{Precisione} + \text{Recupero}}.$$

Infine, viene calcolata la media delle metriche di valutazione calcolate nelle 100 repliche di simulazione. I risultati sono riportati in Tabella 3.1.

TP	FP	FN	Precisione	Recupero	F_1
1.91	0.05	0.09	0.983	0.955	0.960

Tabella 3.1: Metriche di valutazione mediate su 100 dataset simulati: Veri positivi (TP), Falsi Positivi (FP), Falsi Negativi (FN), Precisione, Recupero e F_1-score.

La Precisione è pari a 0.983, indicando che, quando ad una variabile viene attribuito un effetto casuale, si ha un'alta fiducia nel fatto che sia vero. Infatti, la proporzione di falsi positivi è limitata (FP=0.05). Il Recupero raggiunge un valore elevato (0.955), indicando una buona capacità del metodo di individuare i veri positivi, con un numero contenuto di falsi negativi (FN = 0.09). Infine, l'F_1-score risulta pari a 0.96, confermando una prestazione complessivamente adeguata, sia in termini di esattezza che di completezza.

3.3 Obiettivo II

Nella sezione 2.4.2.2 viene presentato un algoritmo per la selezione e la stima degli effetti fissi con penalità SCAD. L'algoritmo è stato implementato in `R` da Ghosh & Thoresen (2016), tuttavia gli autori si limitano alla situazione in cui gli effetti casuali vengono assunti indipendenti tra loro, cioè la matrice di varianze e covarianze $\boldsymbol{\Psi}$ ha elementi nulli al di fuori della diagonale principale. In particolare, richiedono di assumere per $\boldsymbol{\Psi}$ una struttura multiplo dell'identità o diagonale. Tale assunzione risulta spesso troppo restrittiva in contesti reali, in cui gli effetti casuali mostrano strutture di correlazione più complesse. Pertanto si decide di presentare una versione alternativa dell'algoritmo, in cui è possibile assumere per $\boldsymbol{\Psi}$ una struttura più generale. Nello specifico, è possibile considerare una matrice di covarianza generale $\boldsymbol{\Psi}$ tale che sia definita positiva. La funzione qui proposta e riportata interamente in appendice A.3.2 richiede di specificare la struttura per $\boldsymbol{\Psi}$. Nel caso in cui venga adottata la struttura più generale, l'algoritmo definisce $\boldsymbol{\Psi}$ tramite parametrizzazione di Cholesky, cioè $\boldsymbol{\Psi} = \mathbf{L}\mathbf{L}^T$, con $\mathbf{L}$ matrice triangolare inferiore. Inoltre, l'implementazione proposta consente di scegliere tra l'adozione della penalità Lasso e della penalità SCAD.

Vengono condotti studi di simulazione con un duplice scopo: da un lato, verificare se la funzione sviluppata consenta di risalire alla reale struttura di correlazione tra effetti casuali, dall'altro capire se riconoscere tale struttura permetta di selezionare in modo più accurato gli effetti fissi. Ogni analisi di simulazione è condotta su 100 repliche indipendenti, per ciascuna delle quali si mantengono fissi i gruppi, ma si generano i dati *ex novo* all'interno di ciascun gruppo. Su ciascuno dei 100 dataset così ottenuti si stima il modello lineare a effetti misti penalizzato, assumendo per $\boldsymbol{\Psi}$ tre possibili strutture: multiplo dell'identità ($\boldsymbol{\Psi}_I$), diagonale ($\boldsymbol{\Psi}_D$) e generica matrice definita positiva ($\boldsymbol{\Psi}_C$). Il valore ottimale del parametro di regolazione λ è individuato sulla base del BIC e, in corrispondenza di esso, si ricavano le seguenti quantità

- $|S(\hat{\boldsymbol{\beta}})|$: numero di variabili selezionate;
- TP: numero di effetti fissi correttamente identificati;
- Stime dei coefficienti corrispondenti agli effetti fissi non nulli;
- Media delle stime degli effetti fissi realmente nulli;
- PE: errore quadratico medio di previsione sui valori osservati;
- $\hat{\sigma}^2_\varepsilon$: stima finale del parametro di varianza del termine d'errore;

- $\hat{\mathbf{\Psi}}$: matrice di varianza e covarianza degli effetti casuali.

Per ciascuna configurazione di $\mathbf{\Psi}$, i risultati delle 100 repliche vengono sintetizzati calcolandone media e deviazione standard. Per le stime dei coefficienti degli effetti fissi e per i parametri di varianza e covarianza viene inoltre calcolato l'errore quadratico medio rispetto ai veri valori. L'analisi consente una valutazione congiunta della performance in termini di selezione e stima delle variabili, stima dei parametri di varianza e capacità predittiva del modello.

Si presentano i risultati per uno studio di simulazione con le seguenti caratteristiche: si definiscono 20 gruppi con un numero variabile di osservazioni, compreso tra 5 e 12, conducendo ad una numerosità campionaria complessiva pari a N=165. Si confrontano scenari caratterizzati da dimensionalità crescente (p=10, p=100 e p=300). Il numero di effetti fissi è mantenuto costante pari a 5 e il vero valore dei coefficienti degli effetti fissi è $\boldsymbol{\beta}_0 = (1, -2, 3.5, 3, -0.5, 0, ..., 0)^T$. La matrice del disegno $\mathbf{X}$ è definita in maniera tale da avere una prima colonna di 1, corrispondente all'intercetta, e le successive $(p-1)$ colonne sono generate da una distribuzione normale multivariata con media nulla e matrice di covarianza avente in posizione (i, j) il valore $\rho_x^{|i-j|}$ per $i, j = 1, ..., p-1$. Si considera sia il caso in cui le variabili esplicative sono indipendenti ($\rho_x = 0$), sia il caso in cui sono correlate ($\rho_x = 0.5$). Il numero di effetti casuali è $q = 2$ e il vettore di coefficienti degli effetti casuali si distribuisce come una normale di media $\mathbf{0}$ e matrice di varianza e covarianza $\mathbf{\Psi}$. Le covariate con effetto casuale corrispondono alle prime due colonne della matrice del disegno degli effetti fissi $\mathbf{X}$, pertanto si considera un'intercetta casuale e una pendenza casuale. Il vero valore della varianza del termine d'errore è fissato a $\sigma_\varepsilon^2 = 0.25$. Inizialmente, si fissa come vera matrice $\mathbf{\Psi}$ una matrice caratterizzata da una correlazione moderata ($\rho = 0.3$) tra i due effetti casuali

$$\mathbf{\Psi} = \begin{pmatrix} 0.8 & 0.18 \\ 0.18 & 0.45 \end{pmatrix}.$$

Il vettore delle risposte viene generato secondo il modello lineare a effetti misti (1.6), combinando la componente fissa, il contributo degli effetti casuali e un termine di rumore gaussiano. I risultati relativi alla stima e alla selezione degli effetti fissi quando $\rho_x = 0$, riferiti alle 100 repliche di simulazione, sono presentati in Tabella 3.2 separatamente per ogni struttura assunta per $\mathbf{\Psi}$. Si osserva che, all'aumentare di p, l'insieme delle variabili attive rimane stabile, infatti $|S(\hat{\boldsymbol{\beta}})| \approx 5$ e TP=5. L'errore di previsione rimane contenuto al crescere di p e la sua deviazione standard (SD) diminuisce nel caso ad elevata dimensionalità (p=300). Le stime degli effetti fissi delle variabili realmente attive sono vicine ai veri valori. Si nota che la deviazione standard e l'MSE delle variabili a

cui è associato anche un effetto casuale sono maggiori. Soprattutto per le variabili x_1 e x_2, la SD cresce con p, indicando una maggiore instabilità delle stime al crescere della dimensionalità. Le variabili non attive hanno medie prossime a zero e MSE molto bassi, confermando un buon controllo dei falsi positivi. Le tre specificazioni assunte per $\boldsymbol{\Psi}$ portano a risultati molto simili in termini di selezione e stima degli effetti fissi, con MSE pressoché equivalenti.

Tabella 3.2: Risultati della procedura di stima e selezione degli effetti fissi quando $\boldsymbol{\Psi}$ è una matrice piena.

		$\lvert S(\hat{\boldsymbol{\beta}})\rvert$	TP	PE	β_1	β_2	β_3	β_4	β_5	β_N
					$p = 10$					
	Media	5.09	5.00	0.187	0.965	-1.903	3.498	3.000	-0.500	-0.001
Ψ_I	SD	0.35	0.00	0.026	0.208	0.167	0.039	0.047	0.048	0.006
	MSE				0.044	0.037	0.002	0.002	0.002	0.000
	Media	5.09	5.00	0.188	0.964	-1.902	3.498	3.000	-0.500	-0.001
Ψ_D	SD	0.35	0.00	0.026	0.208	0.167	0.039	0.047	0.048	0.006
	MSE				0.044	0.037	0.002	0.002	0.002	0.000
	Media	5.10	5.00	0.202	0.971	-1.904	3.499	3.002	-0.498	-0.001
Ψ_C	SD	0.36	0.00	0.069	0.208	0.169	0.042	0.049	0.052	0.006
	MSE				0.044	0.037	0.002	0.002	0.003	0.000
					$p = 100$					
	Media	5.12	5.00	0.186	0.970	-1.853	3.508	2.993	-0.486	0.000
Ψ_I	SD	0.41	0.00	0.025	0.217	0.185	0.048	0.042	0.051	0.000
	MSE				0.048	0.055	0.002	0.002	0.003	0.000
	Media	5.13	5.00	0.187	0.969	-1.853	3.508	2.993	-0.486	0.000
Ψ_D	SD	0.42	0.00	0.026	0.219	0.185	0.048	0.042	0.051	0.000
	MSE				0.048	0.055	0.002	0.002	0.003	0.000
	Media	5.12	5.00	0.195	0.974	-1.853	3.509	2.992	-0.482	0.000
Ψ_C	SD	0.46	0.00	0.047	0.226	0.185	0.048	0.044	0.056	0.000
	MSE				0.051	0.055	0.002	0.002	0.003	0.000
					$p = 300$					
	Media	5.12	5.00	0.189	0.958	-1.838	3.495	3.004	-0.483	0.000
Ψ_I	SD	0.38	0.00	0.023	0.223	0.211	0.045	0.045	0.056	0.000
	MSE				0.051	0.070	0.002	0.002	0.003	0.000
	Media	5.14	5.00	0.189	0.955	-1.838	3.496	3.004	-0.484	0.000
Ψ_D	SD	0.40	0.00	0.023	0.223	0.211	0.045	0.045	0.056	0.000
	MSE				0.051	0.070	0.002	0.002	0.003	0.000
	Media	5.17	5.00	0.191	0.965	-1.838	3.496	3.004	-0.485	0.000
Ψ_C	SD	0.47	0.00	0.024	0.222	0.212	0.045	0.046	0.057	0.000
	MSE				0.050	0.070	0.002	0.002	0.003	0.000

I risultati relativi alla stima dei parametri di varianza e covarianza sono riportati in Tabella 3.3, la quale mostra stime accurate in ogni scenario. Si osserva che, soprattutto per valori bassi di p, le configurazioni $\boldsymbol{\Psi}_I$ e $\boldsymbol{\Psi}_D$ tendono a sottostimare maggiormente la varianza residua, ma la stima è più stabile in quanto la deviazione standard è ridotta. Al crescere di p aumenta la precisione della stima. Per quanto riguarda la varianza e

Tabella 3.3: Stima dei parametri di varianza.

		σ^2_ε	θ^2_1	θ^2_2	θ_{12}
		$p = 10$			
Ψ_I	Media	0.241	0.615	0.615	0.000
	SD	0.032	0.163	0.163	0.000
	MSE	0.001	0.061	0.054	0.032
Ψ_D	Media	0.242	0.771	0.456	0.000
	SD	0.032	0.268	0.161	0.000
	MSE	0.001	0.072	0.026	0.032
Ψ_C	Media	0.255	0.762	0.457	0.184
	SD	0.072	0.288	0.182	0.146
	MSE	0.005	0.084	0.033	0.021
		$p = 100$			
Ψ_I	Media	0.241	0.596	0.596	0.000
	SD	0.032	0.148	0.148	0.000
	MSE	0.001	0.063	0.043	0.032
Ψ_D	Media	0.241	0.742	0.446	0.000
	SD	0.032	0.242	0.144	0.000
	MSE	0.001	0.061	0.020	0.032
Ψ_C	Media	0.248	0.735	0.447	0.176
	SD	0.050	0.250	0.149	0.157
	MSE	0.002	0.066	0.022	0.024
		$p = 300$			
Ψ_I	Media	0.244	0.634	0.634	0.000
	SD	0.029	0.162	0.162	0.000
	MSE	0.001	0.053	0.060	0.032
Ψ_D	Media	0.244	0.769	0.495	0.000
	SD	0.030	0.242	0.176	0.000
	MSE	0.001	0.059	0.033	0.032
Ψ_C	Media	0.245	0.768	0.503	0.171
	SD	0.030	0.250	0.203	0.146
	MSE	0.001	0.063	0.044	0.021

covarianza degli effetti casuali, si osserva che la configurazione $\mathbf{\Psi}_C$ riesce a recuperare bene il valore θ_{12}, con un MSE piuttosto basso. La configurazione $\mathbf{\Psi}_I$ restituisce un unico valore $\hat{\theta}$ da cui dipende l'intera matrice $\hat{\mathbf{\Psi}}$. Tale valore risulta sottostimare l'intercetta casuale e sovrastimare la pendenza casuale. La configurazione $\mathbf{\Psi}_D$ stima bene le varianze degli effetti casuali. Le impostazioni $\mathbf{\Psi}_I$ e $\mathbf{\Psi}_D$, pur non riuscendo a catturare la correlazione tra effetti casuali, permettono di ottenere stime precise degli effetti fissi. In sintesi, considerare una struttura più generale per $\mathbf{\Psi}$ permette di catturare la dipendenza esistente tra gli effetti casuali, mantenendo MSE competitivi.

I risultati per il caso $\rho_x = 0.5$, riferiti alle 100 repliche di simulazione, sono riportati in Appendice, in Tabelle A.1 e A.2. Risultano molto simili a quelli visti sopra, indicando che l'algoritmo è stabile anche quando vi è correlazione moderata tra predittori. Si osservano leggeri miglioramenti in termini di SD e MSE per le stime delle varianze e

covarianze degli effetti casuali, soprattutto quando si considera per $\boldsymbol{\Psi}$ la struttura più generale $\boldsymbol{\Psi}_C$. L'errore di previsione rimane stabile e si nota una leggera riduzione degli errori standard degli effetti fissi. Inoltre, il metodo rimane adeguato per l'obiettivo della selezione delle variabili realmente attive, infatti $|S(\hat{\boldsymbol{\beta}})| \approx 5$ per tutti gli scenari e TP=5.

Successivamente, si verifica se l'implementazione proposta sia adeguata anche nel caso in cui la matrice $\boldsymbol{\Psi}$ è un multiplo dell'identità, cioè $\boldsymbol{\Psi} = \theta\mathbf{I}$. A tale scopo, si mantengono i valori fissati in precedenza per $\boldsymbol{\beta}_0$ e σ^2_ε, e si definisce

$$\boldsymbol{\Psi} = \begin{pmatrix} 0.6 & 0 \\ 0 & 0.6 \end{pmatrix}.$$

I risultati relativi alla stima dei parametri di varianza sono riportati in Tabella 3.4. La media della covarianza stimata sulle 100 repliche quando si assume $\boldsymbol{\Psi} = \boldsymbol{\Psi}_C$ è pari a -0.008 per p=10 e a -0.010 per p=300, indicando che la scelta più generale per $\boldsymbol{\Psi}$ risulta adeguata anche quando la vera matrice sottostante è un multiplo dell'identità.

Tabella 3.4: Stima dei parametri di varianza.

		σ^2_ε	θ^2_1	θ^2_2	θ_{12}
		$p = 10$			
Ψ_I	Media	0.242	0.604	0.604	0.000
	SD	0.032	0.161	0.161	0.000
	MSE	0.001	0.026	0.026	0.000
Ψ_D	Media	0.242	0.611	0.597	0.000
	SD	0.032	0.208	0.226	0.000
	MSE	0.001	0.043	0.051	0.000
Ψ_C	Media	0.255	0.594	0.606	-0.008
	SD	0.076	0.223	0.268	0.166
	MSE	0.006	0.049	0.071	0.027
		$p = 300$			
Ψ_I	Media	0.245	0.606	0.606	0.000
	SD	0.030	0.155	0.155	0.000
	MSE	0.001	0.024	0.024	0.000
Ψ_D	Media	0.245	0.594	0.619	0.000
	SD	0.030	0.210	0.209	0.000
	MSE	0.001	0.044	0.044	0.000
Ψ_C	Media	0.259	0.581	0.624	-0.010
	SD	0.079	0.240	0.224	0.139
	MSE	0.006	0.057	0.050	0.019

Infine, è stato effettuato un confronto tra i risultati riportati nelle Tabelle 3.2 e 3.3 e quelli ottenuti tramite l'algoritmo di Groll & Tutz (2014), implementato nella libreria `glmmLasso` di `R`. I risultati numerici sono riportati in Appendice, nelle Tabelle A.3 e A.4. Il metodo dimostra di selezionare in maniera adeguata le variabili realmente attive (TP=5), con il vantaggio di un'elevata velocità computazionale e una buona accuratezza

nella stima dei coefficienti associati alle variabili attive. Tuttavia, tende a selezionare un numero maggiore di variabili ($|S(\hat{\boldsymbol{\beta}})|$=13.44), e a sottostimare i parametri di varianza e covarianza nella matrice $\boldsymbol{\Psi}$. Ciò comporta un leggero aumento dell'errore di previsione in-sample, che aumenta da 0.191 a 0.227. Poiché la stima della varianza residua σ_ε^2 non è restituita dalla funzione `glmmLasso` di `R`, si utilizza l'approssimazione

$$\sigma_\varepsilon^2 \approx \frac{RSS}{N - \hat{df}_\lambda}, \tag{3.1}$$

dove RSS corrisponde alla somma dei quadrati dei residui (*Residual Sum of Squares*), N è il numero complessivo di osservazioni e $\hat{df}_\lambda$ corrisponde ai gradi di libertà dello stimatore. Qui, $\hat{df}_\lambda = |\{1 \leq k \leq p : \hat{\beta}_k \neq 0\}| + \dim(\boldsymbol{\theta})$ è la somma del numero di coefficienti di regressioni degli effetti fissi risultati diversi da zero e del numero di parametri di varianza e covarianza. Tale formula conduce ad una varianza residua di 0.248, molto vicina al valore reale di 0.25 (MSE=0.001).

Capitolo 4

Applicazione

4.1 Descrizione dell'esperimento

I dati analizzati in questo capitolo provengono dal dataset CGMacros, reso disponibile pubblicamente su PhysioNet (Gutierrez-Osuna et al., 2025). Riguardano un esperimento condotto presso il *Sansum Diabetes Research Institute* a Santa Barbara, in California, al quale hanno partecipato 45 soggetti adulti, suddivisi in tre categorie: 15 soggetti sani, 16 con pre-diabete e 14 con diabete di tipo 2. A ciascun partecipante sono stati applicati due dispositivi per il monitoraggio continuo della glicemia (CGM), resi ciechi per evitare che la conoscenza dei livelli glicemici influenzasse le scelte alimentari. I dispositivi utilizzati sono: *Abbott FreeStyle Libre Pro CGM*, con campionamento ogni 15 minuti, e *Dexcom G6 Pro CGM*, con campionamento ogni 5 minuti, applicati rispettivamente sul braccio superiore e sull'addome. Per motivi di privacy e in conformità con le normative HIPAA sulle informazioni sanitarie protette (PHI), le date relative alle registrazioni CGM sono state spostate di ±N giorni, con N compreso tra 365 e 720. Per circa dieci giorni, ogni partecipante ha tracciato i propri pasti mediante l'applicazione *MyFitnessPal*, registrando le quantità (in grammi) di carboidrati, grassi, fibre e proteine, nonché la percentuale di pasto effettivamente consumata. La struttura dei pasti era così organizzata:

- Le colazioni consistevano in frullati proteici, con composizioni variabili di macronutrienti;
- I pranzi provenivano da una catena di ristoranti locali ed erano controllati a livello nutrizionale;
- Le cene erano a scelta libera dei partecipanti.

Per minimizzare le interferenze glicemiche tra i pasti, veniva richiesto un intervallo minimo di tre ore tra colazione e pranzo, e tra pranzo e cena, durante il quale era permesso consumare solo acqua o caffè non zuccherato.

Il primo giorno dello studio sono stati rilevati diversi parametri clinici e antropometrici, tra cui indice di massa corporea (BMI), emoglobina glicata (HbA1c), glicemia e insulina a digiuno, trigliceridi e colesterolo. Inoltre, sono state raccolte informazioni demografiche (età, sesso, etnia) e campioni fecali per l'analisi del microbioma intestinale, ovvero l'insieme dei microrganismi che popolano il tratto gastrointestinale umano. Il sequenziamento, effettuato tramite il Viome Gut Intelligence Test, ha permesso di definire la presenza o assenza di 1.979 batteri. A partire da queste informazioni sono stati derivati 22 punteggi compositi associati alla salute intestinale, classificati su una scala ordinale: "Good", "Average", "Not Optimal".

Infine, a ciascun partecipante è stato fornito uno smartwatch *Fitbit Sense* per misurare il battito cardiaco e le calorie consumate.

4.2 Contesto e motivazione scientifica

L'esperimento si colloca nel più ampio ambito della ricerca sulla risposta glicemica postprandiale (PPGR), ovvero la variazione della concentrazione del glucosio nel plasma dopo aver mangiato, e sul monitoraggio automatico della dieta. In particolare, l'obiettivo degli autori era quello di utilizzare la forma della PPGR per stimare la composizione in macronutrienti del pasto tramite modelli di apprendimento automatico. L'interesse della comunità scientifica per questo tema nasce dal fatto che un efficace controllo glicemico postprandiale è fondamentale per la prevenzione delle malattie croniche, in particolare quelle di origine metabolica. Infatti, un'elevata concentrazione di glucosio nel sangue dopo i pasti è stata associata a un maggior rischio di diabete di tipo 2, obesità e malattie cardiovascolari (Guizar et al., 2023).

Nonostante la crescente attenzione verso il monitoraggio della PPGR, fino a dieci anni fa non esisteva ancora un metodo efficace per prevedere come un individuo avrebbe risposto a un determinato pasto. Come riportato da Zeevi et al. (2015), la stima della PPGR si basava comunemente sul contenuto di carboidrati del pasto, nonostante fosse stato dimostrato già nella prima metà del Novecento che questo indicatore da solo predice molto male l'effettiva risposta glicemica. Anche metodi alternativi come l'indice glicemico, che quantifica la PPGR in risposta al consumo di un singolo alimento testato, e il carico glicemico, che ne è una derivazione, risultano poco adatti a rappresentare la risposta glicemica in contesti reali. I pasti consumati nella vita quotidiana sono

infatti composti da combinazioni complesse di alimenti, in quantità variabili, assunti in momenti diversi della giornata e a distanze diverse da altre attività, come esercizio fisico o altri pasti.

Studi recenti, come lo studio pioneristico di Zeevi et al. (2015), hanno mostrato miglioramenti nell'accuratezza della previsione della PPGR quando i modelli predittivi integrano caratteristiche individuali tra cui sesso, età, BMI, abitudini alimentari e stile di vita (es. attività fisica, orario dei pasti) e, in particolare, informazioni sul microbioma intestinale. In altre parole, hanno evidenziato che la risposta glicemica postprandiale è il risultato di una complessa interazione tra molteplici fattori, che comporta una notevole variabilità tra gli individui, indicando l'importanza di formulare raccomandazioni dietetiche su misura in base ai profili genetici dei singoli soggetti. Di conseguenza, negli ultimi anni si è affermato sempre più un approccio personalizzato alla predizione della PPGR. Sebbene la nutrizione personalizzata sia ancora nelle sue fasi iniziali e richieda ulteriori ricerche, in particolare in diverse popolazioni e contesti (Ordovas et al., 2018), questi studi dimostrano il suo potenziale per controllare le risposte glicemiche e affrontare le malattie metaboliche.

Tuttavia, come evidenziato da Mendes-Soares et al. (2019), la considerazione simultanea di un ampio numero di variabili individuali e nutrizionali può rappresentare una sfida e ostacolare l'adozione su larga scala di strategie efficaci per il controllo glicemico. Da un lato, l'attuale disponibilità commerciale di test biologici avanzati, come quello per misurare la composizione del microbioma, unitamente alla diffusione di dispositivi e app per l'auto-monitoraggio, consente una misurazione sempre più accurata dei dati personali, rappresentando una base promettente per la definizione di approcci personalizzati al controllo della glicemia. Dall'altro, affinché tali informazioni possano tradursi in raccomandazioni operative realmente utili, è necessario sviluppare strumenti integrati e facilmente accessibili. Raggiungere questo obiettivo rappresenterebbe un passo cruciale per contrastare la crescente diffusione dell'iperglicemia nella popolazione generale.

4.3 Descrizione del dataset

Il dataset originale è composto da 45 file principali in formato CSV (uno per ciascun partecipante) e da tre file supplementari. I file principali contengono i dati raccolti tramite monitoraggio continuo della glicemia (CGM) e fitness tracker, con una frequenza di campionamento pari a un minuto. Ogni file presenta una riga per ciascun timestamp e una colonna per ogni variabile registrata. Un aspetto rilevante riguarda l'interpolazione dei dati glicemici: per garantire la disponibilità di una misurazione per ogni minuto, gli

autori hanno interpolato le letture provenienti dai dispositivi CGM. Per questo motivo, in questo progetto di tesi si è scelto di utilizzare esclusivamente le misurazioni fornite dal dispositivo Dexcom G6 Pro, il quale offre una frequenza di campionamento più elevata (ogni 5 minuti) rispetto all'altro sensore utilizzato (Abbott FreeStyle Libre), risultando quindi più reattivo alle variazioni glicemiche.

I file includono inoltre informazioni dettagliate sui pasti consumati: per ciascun pasto vengono riportati il contenuto calorico, le quantità di carboidrati, proteine, grassi e fibre e il tipo di pasto (colazione, pranzo o cena). I tre file supplementari integrano i dati principali con ulteriori informazioni di carattere clinico e biologico, raccolte il primo giorno di osservazione. In particolare, includono:

- Dati demografici (età, genere, etnia);
- Misurazioni antropometriche (altezza, peso, BMI);
- Parametri biochimici: HbA1c, glucosio a digiuno, insulina, trigliceridi, colesterolo, lipoproteine ad alta densità (HDL), non-HDL, lipoproteine a bassa densità (LDL) e lipoproteine a densità molto bassa (VLDL);
- Tre misurazioni della glicemia da puntura del dito;
- Profilo del microbioma intestinale;
- Punteggi di salute intestinale Viome.

Nel complesso, il dataset rappresenta una risorsa multidimensionale per lo studio della risposta glicemica postprandiale e per lo sviluppo di metodi di monitoraggio automatico della dieta, in quanto combina dati fisiologici ad alta frequenza con informazioni nutrizionali, comportamentali e microbiologiche.

4.4 Analisi esplorativa

In questa sezione saranno effettuate alcune analisi esplorative relative al dataset, con particolare attenzione alle variabili raccolte prima dell'esperimento, al fine di fornire un primo quadro descrittivo dei soggetti coinvolti nello studio.

4.4.1 Variabili demografiche

Il campione dello studio è composto da 16 uomini e 29 donne, di età compresa tra i 18 e i 69 anni. La Figura 4.1 mostra la distribuzione dell'età per genere: si osserva che l'età

mediana è pari a 51 anni per le donne e 52 anni per gli uomini. Tra le partecipanti di sesso femminile, l'età risulta più concentrata, in particolare nella fascia compresa tra i 45 e i 60 anni, con la presenza di un outlier sotto i 30 anni. Al contrario, la distribuzione dell'età tra gli uomini appare più dispersa, con un intervallo interquartile più ampio.

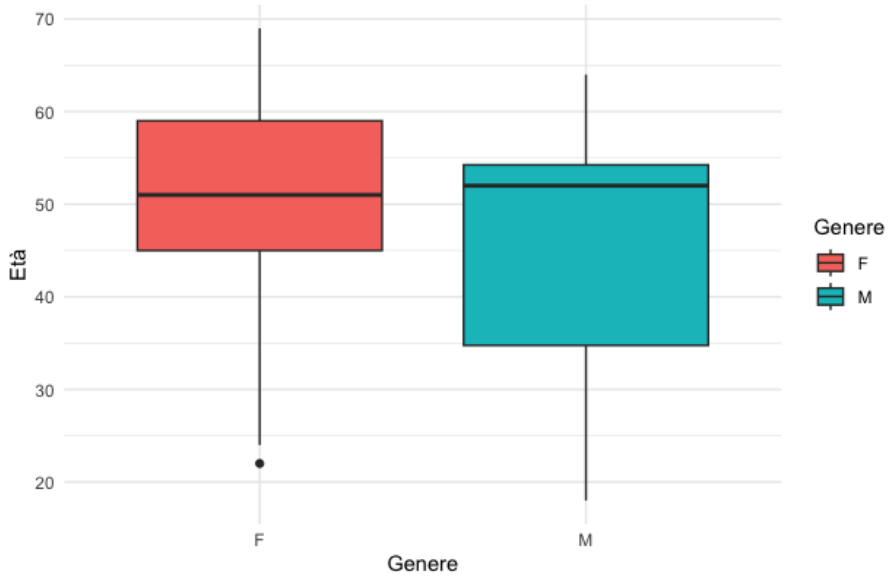

FIGURA 4.1: Distribuzione dell'età per genere.

Per quanto riguarda l'autoidentificazione etnica, emerge che:

- 3 partecipanti si identificano come afroamericani;
- 1 partecipante come nero, afroamericano;
- 34 partecipanti come ispanici/latini;
- 7 partecipanti come bianchi.

Risulta pertanto evidente una prevalenza di soggetti ispanici nel campione. Dato che la modalità "nero, afroamericano" è osservata per un solo soggetto e costituisce una specificazione della categoria più generale "afroamericano", si è deciso di aggregarla a quest'ultima.

4.4.2 Variabili cliniche

L'Indice di Massa Corporea (BMI) è una misura comunemente utilizzata per classificare lo stato ponderale degli adulti, calcolata come rapporto tra il peso (in kg) e il quadrato dell'altezza (in m^2). La Figura 4.2 mostra la ripartizione dei soggetti in tre categorie

sulla base del BMI, coerentemente con la classificazione dell'Organizzazione Mondiale della Sanità (World Health Organization & WHO Consultation on Obesity, 2000). Il 20% dei partecipanti allo studio è risultato normopeso, il 29% sovrappeso e il 51% in condizione di obesità.

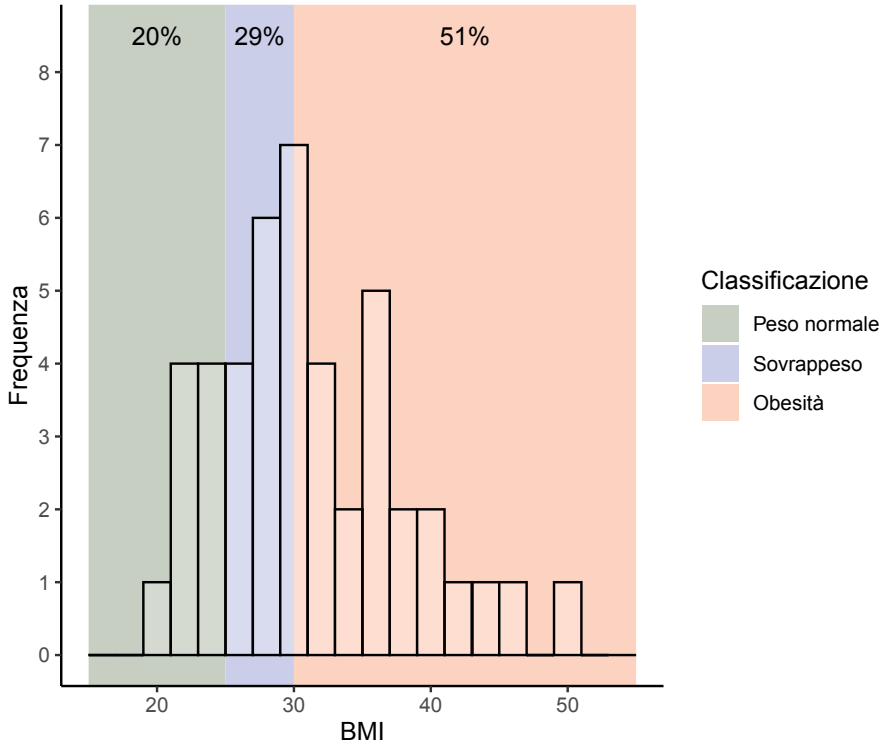

FIGURA 4.2: Ripartizione dei soggetti in 3 categorie sulla base dell'Indice di Massa Corporea.

L'HbA1c è un indicatore affidabile del controllo glicemico a lungo termine in quanto riflette la glicemia media nei 2–3 mesi precedenti (American Diabetes Association, 2001). Tuttavia, non rileva le variazioni a breve termine, come le escursioni glicemiche postprandiali (PPG). Sulla base del valore registrato di HbA1c, il campione può essere suddiviso in tre gruppi:

- 14 partecipanti sono non diabetici (HbA1c$<$5.7);
- 15 partecipante sono pre-diabetici (5.7 $\geq$ HbA1c$\leq$6.4);
- 13 partecipanti hanno diabete di tipo 2 (HbA1c $>$ 6.4).

Per fornire una panoramica del profilo metabolico del campione prima dell'esperimento, nella Tabella 4.1 sono riportate alcune statistiche descrittive (media, deviazione standard, minimo e massimo) delle principali variabili misurate a digiuno, tra cui insulina, colesterolo e altri parametri biochimici.

Tabella 4.1: Statistiche descrittive per le variabili biochimiche rilevate a digiuno prima dello studio.

Variabili	Media	Deviazione standard	Minimo	Massimo
Glucosio	120.69	30.23	79.00	218.00
Insulina	14.17	8.26	2.50	46.40
Trigliceridi	160.02	169.43	40.00	1150.00
Colesterolo	191.64	46.71	91.00	345.00
Colesterolo HDL	51.91	15.43	24.00	106.00
Colesterolo Non HDL	139.73	47.23	38.00	283.00
Colesterolo LDL	128.96	109.99	21.00	800.00
Colesterolo VLDL	35.78	57.61	8.00	400.00

4.4.3 Variabili sul microbioma intestinale

Dall'analisi preliminare sulle variabili relative alla presenza/assenza dei batteri, è emerso che la maggior parte di esse presenta una forte collinearità. Utilizzando la funzione `findLinearCombos()`, implementata nella libreria R `caret` (Kuhn, 2008), si è osservato che ben 1935 delle 1979 variabili batteriche risultano essere linearmente combinabili tra loro, ovvero ridondanti da un punto di vista informativo. A causa dell'elevato rischio di multicollinearità, che può compromettere la stima dei parametri e la stabilità dei modelli, si è scelto di non includere direttamente le singole variabili batteriche, ma di utilizzare i punteggi compositi forniti nel dataset. Questi rappresentano un riassunto più robusto e interpretabile del profilo microbiotico, e consentono di mantenere l'informazione biologica riducendo dimensionalità e collinearità. Tuttavia, questo comporta l'esclusione dall'analisi dei soggetti 26, 28 e 48, per i quali non si dispone delle informazioni sui punteggi di salute intestinale. Essendo tali punteggi espressi su scala ordinale, si è deciso di trattarli come fattori ordinali. Tuttavia, tra i 22 indicatori disponibili, 10 sono risultati ridondanti in quanto dipendenti linearmente dagli altri, come emerso utilizzando la funzione `findLinearCombos()`. Di conseguenza, solo un sottinsieme dei punteggi compositi viene incluso nelle analisi.

4.4.4 Letture CGM

La Figura 4.3 mostra le traiettorie individuali dei livelli glicemici di quattro partecipanti, per ogni giornata del periodo di monitoraggio. La curva nera rappresenta l'andamento medio dei valori di glucosio, mentre l'area grigia evidenzia l'intervallo di variabilità. Si nota che alcuni soggetti mostrano un andamento glicemico maggiormente regolare nel periodo dell'esperimento, riconducibile ipoteticamente a orari e composizione dei pasti più costanti (ad esempio il partecipante 12). Al contrario, altri soggetti, come il

partecipante 1, presentano fluttuazioni significative tra una giornata e l'altra, suggerendo un pattern alimentare e comportamentale più variabile.

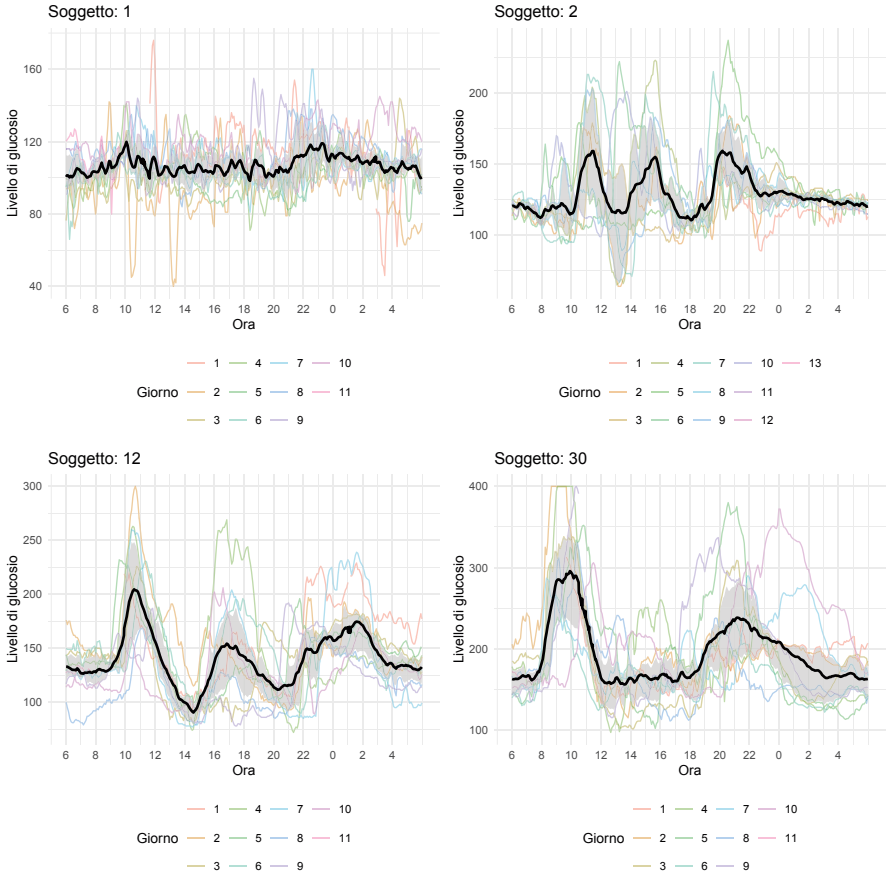

FIGURA 4.3: Curve glicemiche per 4 pazienti nel periodo d'osservazione.

4.4.5 Composizione dei pasti

La Figura 4.4 mostra, sull'asse delle ascisse, i quattro macronutrienti principali e, sull'asse delle ordinate, la quantità media (in grammi) contenuta nei pasti analizzati. I carboidrati risultano essere il macronutriente più presente, con una quantità media di poco inferiore a 60 grammi. Seguono le proteine, con circa 30 grammi, e i grassi, che si attestano attorno ai 20 grammi. Le fibre risultano essere il macronutriente meno rappresentato, con una quantità media di circa 10 grammi. Risulta essere un aspetto critico in quanto potrebbe indicare una scarsa presenza di frutta, verdura o cereali integrali, elementi essenziali per la salute intestinale e il controllo glicemico. Al contempo, l'elevata presenza di carboidrati suggerisce che i pasti analizzati sono tendenzialmente ricchi di zuccheri o amidi.

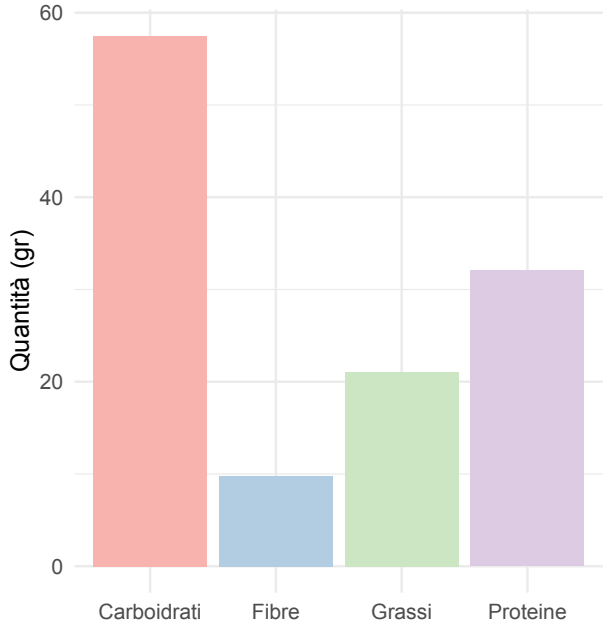

FIGURA 4.4: Grafico a barre della composizione media dei pasti a livello di macronutrienti.

4.5 Variabile risposta

A partire dalle letture del monitoraggio continuo del glucosio, è stata ricavata una misura della PPGR, il cui andamento è determinato da molti fattori, tra cui l'assorbimento dei carboidrati, la secrezione di insulina e glucagone, e i loro effetti coordinati sul metabolismo del glucosio a livello epatico e nei tessuti periferici (American Diabetes Association, 2001). Negli individui non diabetici, la glicemia a digiuno, cioè dopo un digiuno notturno di 8-10 ore, varia generalmente tra 70 e 110 mg/dl. Le concentrazioni di glucosio iniziano ad aumentare circa 10 minuti dopo l'inizio del pasto, come conseguenza dell'assorbimento dei carboidrati ingeriti. Il picco glicemico si verifica circa 60 minuti dopo l'inizio del pasto e solitamente non supera i 140 mg/dl, tornando ai livelli di partenza entro 2–3 ore. Nei pazienti affetti da diabete di tipo 1, l'assenza di insulina endogena rende i livelli glicemici dipendenti dalla quantità e modalità di somministrazione dell'insulina. Per chi soffre di diabete di tipo 2, la secrezione di insulina è ritardata e insufficiente, portando a escursioni glicemiche più alte e prolungate rispetto ai soggetti sani. Poiché l'assorbimento del cibo continua per 5–6 ore dopo un pasto, sia nei diabetici che nei non diabetici, è necessario determinare qual è il momento ottimale per misurare la glicemia postprandiale. In generale, una misurazione a 2 ore dall'inizio del pasto è considerata pratica e rappresentativa.

Numerosi studi recenti hanno evidenziato una forte variabilità della PPGR, sia intra-individuale (tra diversi momenti nello stesso soggetto) sia inter-individuale (tra soggetti

diversi). Considerare tale variabilità è fondamentale per migliorare la predizione della risposta glicemica e sviluppare strategie nutrizionali personalizzate volte alla prevenzione di risposte postprandiali elevate. Una rappresentazione grafica di questo aspetto si trova nella Figura 4.5, che mostra le curve glicemiche di nove partecipanti in risposta al medesimo pasto somministrato in due giornate diverse. Si tratta di una colazione così composta: 66 grammi di carboidrati, 22 di proteine e 10.5 di grassi, per un totale di 448 calorie. È evidente come la risposta glicemica possa variare sensibilmente da un individuo all'altro. Inoltre, alcuni soggetti presentano curve molto simili nelle due occasioni, indicando una buona riproducibilità, mentre altri mostrano risposte discordanti, riflettendo una maggiore variabilità interna.

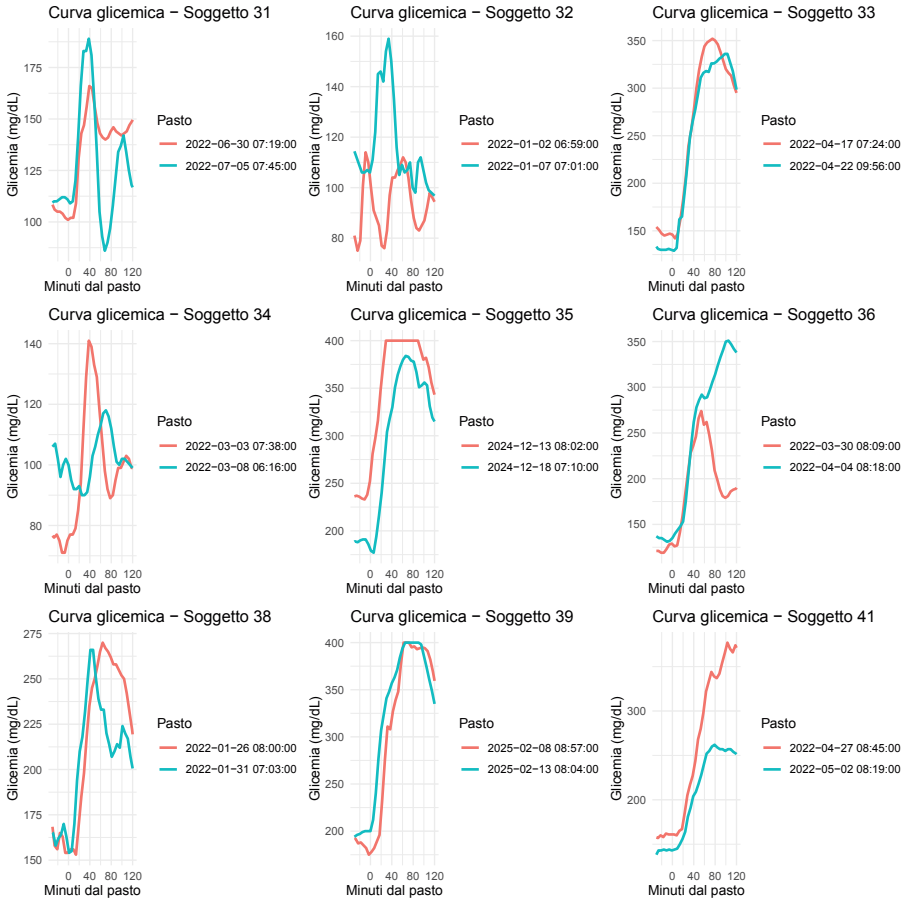

FIGURA 4.5: Curve glicemiche per 9 pazienti in risposta al medesimo pasto somministrato in due occasioni.

4.5.1 iAUC - incremental Area Under the Curve

Un indicatore frequentemente utilizzato in letteratura per calcolare la PPGR è l'Incremental Area Under the Curve (iAUC), applicato per la prima volta in questo contesto

da Wolever & Jenkins (1986). Prevede di misurare l'area sotto la curva glicemica nelle due ore successive all'inizio di un pasto, considerando solo i valori superiori al livello basale di glicemia e ignorando le porzioni della curva al di sotto di tale soglia. Il valore basale viene definito come la mediana delle misurazioni di glucosio registrate nei 10-30 minuti precedenti il pasto. Come fa notare Allison et al. (1995), in questo ambito vi è una certa ambiguità terminologica in quanto Wolever & Jenkins (1986) si riferiscono a questa quantità semplicemente come area incrementale, ma sarebbe in realtà l'area incrementale positiva. L'approccio di Wolever & Jenkins (1986) elimina le aree negative ed è talvolta criticato in letteratura, perché tale "troncamento" equivale a scartare gran parte della variabilità contenuta nelle letture inferiori al basale.

La funzione esatta che descrive la variazione della concentrazione di glucosio nel plasma non è nota e, di conseguenza, è necessario utilizzare un qualche metodo di approssimazione (Allison et al., 1995). Nella ricerca sul diabete si utilizza comunemente la regola del trapezio, un metodo di integrazione numerica che appartiene alla categoria più generale delle formule di Newton-Cotes. Consiste nel rappresentare i punti osservati in uno spazio cartesiano (cioè in un grafico x-y), collegare i punti adiacenti con segmenti e utilizzare la linea spezzata risultante come approssimazione della curva. Collegando questi segmenti all'asse delle ascisse si forma un poligono, e a quel punto è possibile calcolarne l'area. In questi termini, il problema si riduce a somma di aree di rettangoli e triangoli, e può essere risolto attraverso la geometria elementare. Per maggiori dettagli si veda Wolever & Jenkins (1986). L'accuratezza relativa di questo metodo può variare in funzione di diversi fattori, tra cui il numero di punti disponibili, la forma della curva reale e la presenza di errore di misura nei dati (Allison et al., 1995).

Per la stima dell'area, si considera una finestra temporale di 120 minuti post-pasto. Si calcola il contributo in ciascun intervallo temporale (t_i, t_{i+1}), con $i = 1, ..., 119$. I contributi dei singoli intervalli vengono sommati per ottenere l'area complessiva, definita come:

$$iAUC = \sum_{i=1}^{119} \left(\frac{\max(G_i - G_{\text{basale}}, 0) + \max(G_{i+1} - G_{\text{basale}}, 0)}{2} \right) \times (t_{i+1} - t_i),$$

dove G_i rappresenta la glicemia osservata al tempo t_i, e G_{basale} è il valore di riferimento calcolato prima del pasto. Un valore di iAUC elevato indica una risposta glicemica postprandiale più marcata, mentre un valore più basso suggerisce che l'impatto glicemico del pasto è ridotto.

4.5.2 netAUC - net incremental Area Under the Curve

Un indicatore alternativo utilizzato per quantificare la variazione della risposta glicemica postprandiale negli studi sul diabete è il netAUC, cioè l'area incrementale netta sotto la curva. A differenza dell'iAUC, il netAUC tiene conto anche delle aree in cui la glicemia scende al di sotto del valore basale, assegnando loro segno negativo, anziché troncarle a zero. Anch'esso si calcola applicando la regola del trapezio, che comporta una sottrazione delle aree al di sotto del livello basale da quelle al di sopra (Wolever, 2004). È una valida alternativa quando nei dati si osservano numerosi casi in cui il livello glicemico nelle due ore successive al pasto risulta inferiore al valore di partenza. Infatti, considerando anche le aree al di sotto del basale, il netAUC riesce a catturare l'intensità del calo glicemico e a recuperare informazione fisiologica importante.

4.6 Preprocessing

A partire dal dataset originale è stato ricavato un dataset di dimensione inferiore, in cui sono state mantenute solamente le righe corrispondenti ad un pasto. Il preprocessing e la costruzione del dataset su cui svolgere le analisi sono avvenuti sulla base delle indicazioni di Zeevi et al. (2015). In particolare, si è proceduto nel seguente modo:

(i) Sono stati uniti i pasti registrati a meno di 30 minuti di distanza l'uno dall'altro, a condizione che il primo pasto contenesse più di 50 Calorie. I valori associati al pasto unito sono stati ottenuti come somma di quelli relativi ai due pasti distinti. È stato mantenuto l'orario relativo alla prima osservazione;

(ii) Sono stati rimossi i pasti registrati entro 90 minuti da altri pasti contenenti più di 50 Calorie;

(iii) Sono stati rimossi i pasti con componenti dal peso superiore a 1 kg, a meno che non contenessero meno di 20 Calorie;

(iv) Sono stati rimossi i pasti con registrazioni incomplete;

(v) Sono stati rimossi i pasti con meno di 70 Calorie e meno di 15 grammi di peso totale;

(vi) Sono stati rimossi i pasti consumati nelle prime e ultime 12 ore del periodo di connessione al CGM;

(vii) Ad ogni pasto è stata associata la quantità totale di carboidrati assunti nelle 3, 6 e 12 ore precedenti il pasto;

(viii) Ad ogni pasto è stata associata la quantità totale di calorie assunte nelle 2, 3, 6 e 12 ore precedenti il pasto;

(ix) Ad ogni pasto è stata associata la quantità totale di fibre assunte nelle 12 e 24 ore precedenti il pasto;

(x) Ad ogni pasto è stato associato il valore della glicemia misurata 1 ora prima, 2 ore prima e 4 ore prima.

Inoltre, l'iAUC calcolato è espresso in mg/dL·min. Per riportare i valori in mg/dL·h, unità di misura più comunemente utilizzata in letteratura, è stata effettuata una conversione dividendo ciascun valore per 60. Lo stesso viene ripetuto per il netAUC.

In aggiunta, si eliminano i pasti per cui non si dispone delle informazioni rilevate tramite lo smartwatch *Fitbit*. Sono state effettuate alcune valutazioni sull'integrazione nel seguente studio delle misurazioni della frequenza cardiaca al minuto e le kilocalorie consumate, al fine di esplorare il potenziale contributo di queste variabili nella modellazione della risposta glicemica postprandiale. Studi precedenti si sono limitati a considerare, tra i predittori, lo stile di vita dei partecipanti (es. livello di attività fisica abituale), ma raramente disponevano di dati fisiologici così granulari come quelli forniti dal presente esperimento. Tenere conto di queste quantità è importante in quanto numerose evidenze suggeriscono che l'attività del sistema nervoso autonomo, valutata tramite la frequenza cardiaca e la sua variabilità (HRV), sia correlata alla regolazione della glicemia. Uno studio di Nickel et al. (2024) ha evidenziato un'associazione significativa, nei soggetti di sesso maschile, tra i parametri dell'HRV a riposo e la clearance del glucosio, cioè la velocità con cui il glucosio viene eliminato dal sangue dopo aver ingerito una certa quantità, misurata tramite Oral Glucose Tolerance Test (OGTT). Tali risultati suggeriscono che lo stato del sistema nervoso autonomo a riposo può essere predittivo della risposta glicemica in condizioni di carico glucidico. Al contrario, nelle partecipanti di sesso femminile non sono state riscontrate correlazioni significative, indicando la possibile esistenza di differenze fisiologiche di genere nel controllo della glicemia. Inoltre, Kang et al. (2023) hanno confermato che l'esercizio fisico postprandiale (PPE) riduce in modo significativo l'iperglicemia postprandiale, come mostrato dalla riduzione dell'area sotto la curva del glucosio e della glicemia media nelle 24 ore. Tali effetti sono attribuibili a molteplici meccanismi fisiologici: l'aumento dell'assorbimento muscolare del glucosio, mediato sia dall'insulina che dalla contrazione muscolare, la soppressione della produzione epatica di glucosio e l'incremento del flusso ematico periferico. Inoltre, il beneficio è più marcato nei pazienti con diabete di tipo 2, ma si osserva anche in soggetti sovrappeso o

obesi, indipendentemente dall'intensità dell'esercizio, purché la durata superi i 20 minuti. Questi risultati evidenziano l'importanza di considerare parametri cardiovascolari come predittori nella modellazione della risposta glicemica. A partire dai dati raccolti tramite il *Fitbit* sono state calcolate specifiche quantità per analizzare come si comporta la frequenza cardiaca (HR) prima e dopo il pasto, misurando medie, trend, aree sotto la curva, eventi anomali (picchi e salti) e tempi di reazione. In particolare:

- Si è ricavata la frequenza cardiaca media nei 30 minuti successivi al pasto;
- Si è calcolata la frequenza cardiaca media nei 30 minuti precedenti al pasto e nelle due ore successive. Si è poi ottenuta la differenza tra queste due quantità, che ha segno positivo se il battito cardiaco è stato mediamente superiore dopo il pasto, negativo altrimenti;
- Si è ottenuto il coefficiente angolare della regressione lineare tra la frequena cardiaca e il tempo nei 120 minuti dopo il pasto. Fornisce un'indicazione di quanto varia il battito cardiaco ogni minuto, in positivo o negativo;
- Si è calcolata l'incremental Area under the Curve in maniera analoga a quanto fatto per il livello di glucosio;
- Si è calcolato il numero di picchi, cioè il numero di volte in cui la frequenza cardiaca ha superato un valore soglia nella finestra temporale successiva al pranzo. La soglia è stata fissata a 70 battiti al minuto (bpm) sulla base del fatto che rappresenta la frequenza cardiaca a riposo di un uomo ed è di poco inferiore a quella delle donne, pari a 75 bpm. A partire dal numero di picchi è stato ricavato il numero di blocchi contigui, cioè di sequenze temporali consecutive in cui la frequenza cardiaca rimane costantemente sopra la soglia, al fine di valutare la durata e la frammentazione dell'attivazione cardiovascolare dopo il pasto;
- Se nelle due ore successive al pasto ci sono valori di frequenza cardiaca sopra soglia, è stata calcolata la media di questi valori. È stato ottenuto anche il valore massimo oltre la soglia;
- Sono state calcolate le differenze successive di frequenza cardiaca, cioè quanto cambia il battito da un minuto al successivo. Da queste è stato ricavato il numero di salti bruschi, definiti come variazioni di battito maggiori di 10 battiti/minuto in valore assoluto;
- Infine viene calcolato il tempo al primo picco, se questo è presente.

Per quanto riguarda la misurazione, sempre fornita dal *Fitbit*, sulle calorie "bruciate", si è deciso di aggregare quest'informazione sommando i valori rilevati nelle due ore successive al pasto, in linea con il fatto che l'attività fisica postprandiale può influenzare la risposta glicemica. La somma delle calorie bruciate in questa finestra temporale consente di ottenere un indicatore complessivo dell'impegno energetico postprandiale. Per il soggetto 7 non era disponibile l'informazione sulle kilocalorie consumate ogni minuto, ma era presente l'informazione sui METs al minuto, moltiplicati per 10. I METs (acronimo di Metabolic Equivalent of Task) sono un'unità di misura utilizzata per esprimere l'intensità energetica di un'attività fisica rispetto al metabolismo basale, cioè al consumo energetico a riposo. Si può mostrare (si veda Institute of Medicine, 2005, pagina 882) che

$$kcal/min = \text{METs} \cdot \text{Peso}(kg) \cdot 0.0175\ kcal/kg/min.$$

Considerando che il peso dei soggetti è fornito in libre e i valori di METs sono stati moltiplicati per 10 (METs_10), la formula effettivamente utilizzata è la seguente:

$$kcal/min = \frac{\text{METs_10}}{10} \cdot \frac{\text{Peso in lib}}{2.20462} \cdot 0.0175.$$

L'orario del pasto è una variabile temporale che presenta una natura ciclica, in quanto si ripete ogni giorno con una struttura periodica (es. colazione, pranzo, cena), come emerge dalla Figura 4.6, che mostra l'andamento dell'iAUC nel periodo di osservazione per il Soggetto 1. Per modellare correttamente questa periodicità evitando discontinuità artificiali (ad esempio tra le 23:59 e le 00:00), è stata utilizzata una trasformazione in serie di Fourier, che consente di rappresentare variabili cicliche attraverso combinazioni di funzioni seno e coseno. Questa trasformazione ha il vantaggio di catturare in modo flessibile eventuali pattern non lineari nella relazione tra orario del pasto e risposta glicemica. Nello specifico, sono stati inclusi termini di primo, secondo e terzo ordine della trasformata, per modellare rispettivamente variazioni giornaliere di base (es. differenza mattino-sera), variazioni intermedie (es. distinzioni tra colazione, pranzo e cena) e, se presenti, effetti più complessi legati a orari meno frequenti. Questa scelta mira a preservare l'informazione temporale e a valorizzare la potenziale influenza dell'orario del pasto sulla risposta glicemica postprandiale, in linea con evidenze nella letteratura che indicano un diverso impatto del pasto sulla PPGR a seconda dell'ora del giorno (Shilo et al., 2022).

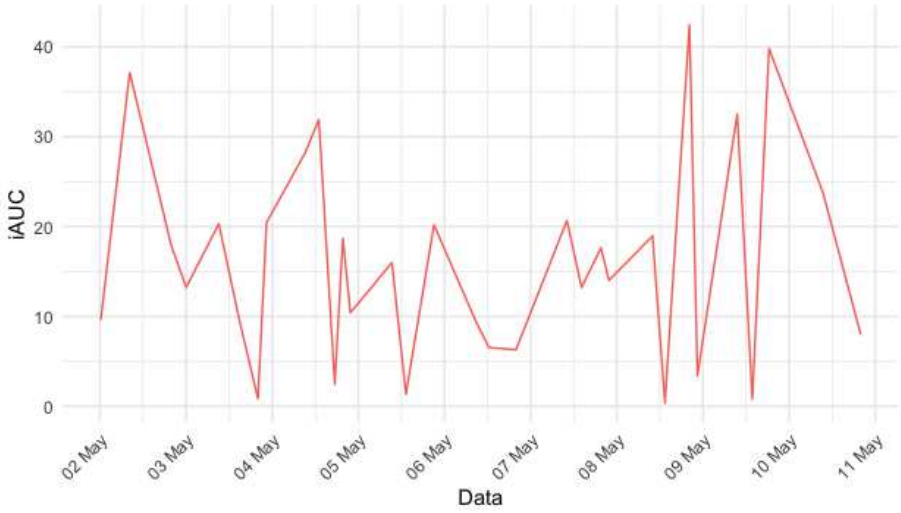

FIGURA 4.6: Andamento dell'iAUC per il Soggetto 1 nel periodo di osservazione.

4.7 Descrizione del dataset finale

Il processo di preprocessing ha portato a un totale di 1081 pasti. Il numero di osservazioni per ciascun soggetto varia all'interno del dataset, da un minimo di 8 a un massimo di 38, ed è riportato nella Tabella 4.2.

TABELLA 4.2: Numero di osservazioni per ogni soggetto.

Soggetto	Osservazioni	Soggetto	Osservazioni	Soggetto	Osservazioni
1	32	15	31	33	16
2	28	16	22	34	17
3	27	17	14	35	19
4	36	18	28	36	21
5	33	19	29	38	24
6	30	20	25	39	24
7	13	21	29	41	25
8	24	22	33	42	16
9	30	23	32	43	38
10	33	27	28	44	27
11	31	29	26	45	28
12	26	30	22	46	23
13	31	31	27	47	22
14	31	32	8	49	22

La distribuzione dell'iAUC è fortemente asimmetrica a destra, come mostrato dalla Figura 4.7. La maggior parte dei pasti ha valori di iAUC contenuti (sotto i 100 mg/dL·h), ma sono presenti alcuni pasti con valori molto elevati. Questo suggerisce che la risposta glicemica è generalmente bassa o moderata, ma ci sono situazioni in cui i pasti inducono

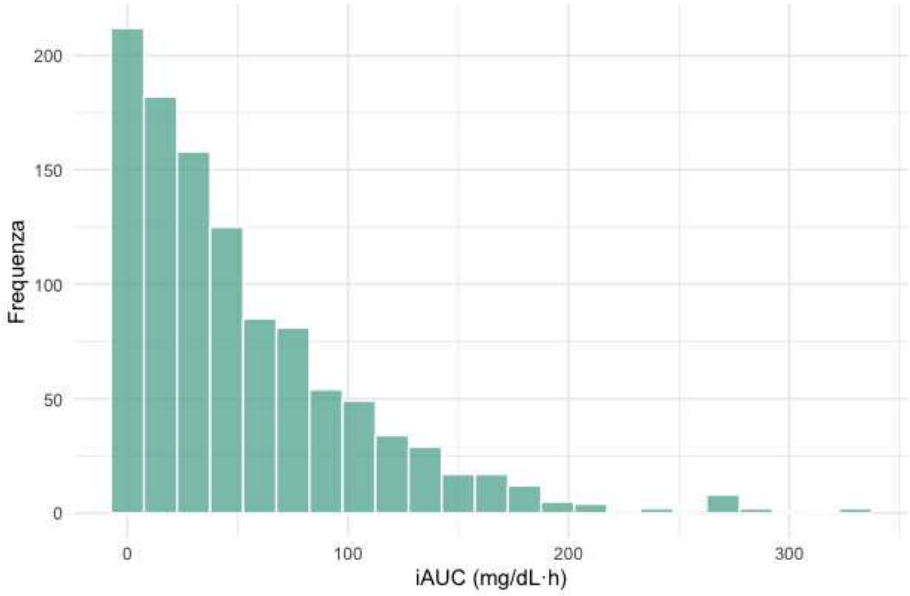

FIGURA 4.7: Distribuzione dell'iAUC.

una risposta significativamente più alta. Inoltre, è presente una porzione non trascurabile di pasti con iAUC pari a zero, a indicazione del fatto che la glicemia dopo l'ingestione risulta inferiore al valore di partenza. Questi valori nulli riflettono principalmente cause di natura fisiologica: sono infatti associati prevalentemente a soggetti con diabete di tipo 2, per i quali si osserva un'escursione glicemica ritardata, con un picco che può verificarsi oltre le tre ore dal pasto. Per questi individui, l'intervallo di osservazione standard potrebbe non essere sufficiente a catturare l'intera risposta glicemica, portando così a una sottostima dell'iAUC calcolato. Dato che l'iAUC riduce la capacità del modello di cogliere l'intera variabilità delle risposte fisiologiche dopo i pasti, si è deciso di utilizzare il netAUC come variabile d'interesse per quantificare la risposta glicemica postprandiale. In questo modo, anche le discese glicemiche rispetto al valore basale sono incluse nel calcolo dell'indicatore. La distribuzione del netAUC è mostrata in Figura 4.8 ed è caratterizzata da una forma complessivamente simile a una normale. Si nota la presenza di valori estremi (*outliers*), tuttavia si decide di non escluderli dall'analisi in quanto si ritiene che possano essere associati a soggetti con risposte atipiche ma caratteristiche della loro condizione e che quindi possano riflettere l'effettiva eterogeneità tra i soggetti.

La Figura 4.9 mostra le relazioni tra netAUC medio per soggetto e diverse variabili biologiche o antropometriche. Si nota una relazione positiva tra l'età del soggetto e la variabile risposta, infatti all'aumentare dell'età, il netAUC medio tende a crescere. Questo suggerisce che i soggetti più anziani potrebbero avere una maggiore risposta

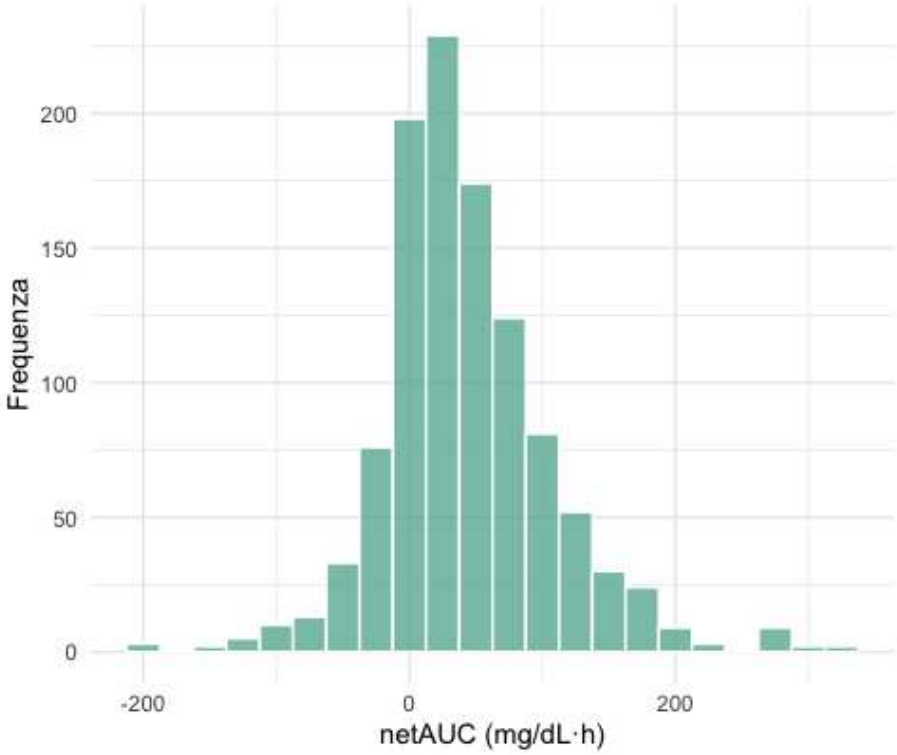

FIGURA 4.8: Distribuzione del netAUC.

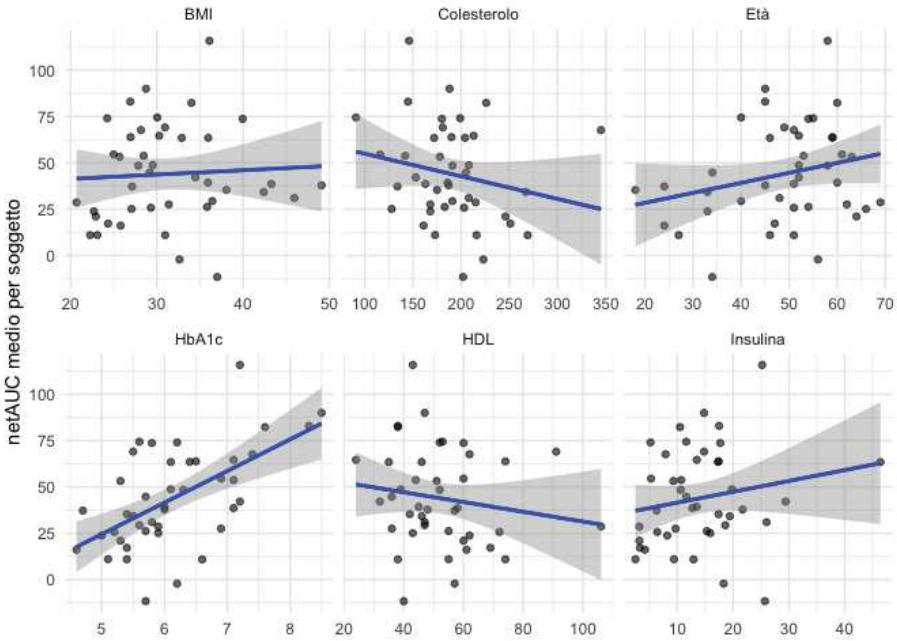

FIGURA 4.9: Relazione tra il netAUC medio calcolato per ogni soggetto e alcune variabili.

glicemica postprandiale. Vi è una relazione positiva anche tra l'indice HbA1c e il netAUC medio, tale per cui i soggetti con valori più alti di HbA1c mostrano una risposta glicemica media più elevata. Relazioni negative coinvolgono invece il colesterolo e il

colesterolo HDL: i soggetti con colesterolo più alto tendono ad avere una risposta glicemica media più bassa. Anche livelli più alti di HDL, cioè di colesterolo "buono", sono associati a valori più bassi della variabile risposta. Questo è coerente con quanto noto in letteratura, infatti un HDL alto è tipicamente associato a un metabolismo glucidico più efficiente. Per quanto riguarda l'insulina, vi è una relazione leggermente positiva, ma con alta variabilità. Potrebbe indicare che chi ha insulina a digiuno più alta tende ad avere un peggior controllo glicemico postprandiale, ma il segnale non è molto forte. Infine, anche per il BMI la relazione è debole e tendenzialmente piatta. Per comprendere meglio la relazione che lega la risposta glicemica postprandiale all'insieme delle variabili rilevate, si procede con l'adattamento dei modelli.

Capitolo 5

Modellazione

5.1 Introduzione

L'obiettivo di questo capitolo è valutare se tenere conto della struttura gerarchica dei dati e introdurre un termine di penalizzazione permettano di migliorare l'adattamento dei modelli e la loro capacità predittiva. A tale scopo, verranno adattati e confrontati quattro modelli principali:

- Modello lineare classico: assume che tutte le osservazioni siano indipendenti;
- Modello lineare a effetti misti: introduce degli effetti casuali per modellare la possibile correlazione tra osservazioni relative ad uno stesso soggetto;
- Modello lineare penalizzato: permette di fare selezione delle variabili, conducendo a soluzioni più parsimoniose. Non tiene conto del fatto di avere più misure per uno stesso soggetto;
- Modello lineare a effetti misti penalizzato: combina la selezione di variabili con l'inclusione di effetti casuali. Permette di valutare come agisce la penalizzazione in presenza di effetti casuali.

Il confronto tra i modelli si basa su diversi aspetti: varianza degli stimatori, numero di variabili selezionate, struttura di correlazione del modello e performance predittiva. Innanzitutto, si valuterà come cambia la varianza degli stimatori passando da modelli che non considerano la struttura gerarchica a quelli che la considerano. Si esaminerà poi il numero di variabili selezionate o risultate significative: infatti, non solo la penalizzazione tende a ridurre il numero di variabili nel modello, ma anche l'introduzione di effetti casuali può determinare la perdita di significatività di alcune covariate. Si

analizzerà anche la struttura di correlazione del modello, sia in termini di correlazione tra residui che in termini di correlazione tra le stime dei coefficienti dei predittori. A tale scopo, si procederà ad un confronto grafico per valutare se l'effetto casuale nei modelli a effetti misti sia in grado di cogliere l'eterogeneità tra soggetti. Saranno calcolati anche i coefficienti di correlazione intraclasse per quantificare la proporzione di varianza spiegata dalla componente tra soggetti. Si otterrà inoltre la matrice di correlazione delle stime dei coefficienti. Infine, si misurerà la capacità predittiva dei modelli in termini di RMSE.

5.2 Struttura di correlazione

Per prima cosa si procede a calcolare la matrice di correlazione tra le misurazioni della risposta glicemica (netAUC), calcolata aggregando i dati di tutti i 42 partecipanti. I valori numerici, relativi alle prime 10 misurazioni, sono riportati nella Tabella 5.1.

Tabella 5.1: Matrice di correlazione della risposta per tutti i soggetti per le prime 10 misurazioni.

	1	2	3	4	5	6	7	8	9	10
1	1.00	-0.03	-0.14	0.53	-0.12	0.19	0.42	0.24	0.17	0.19
2	-0.03	1.00	0.03	0.18	0.17	0.21	0.08	0.10	0.34	0.06
3	-0.14	0.03	1.00	-0.30	0.28	0.52	0.11	0.07	-0.26	-0.01
4	0.53	0.18	-0.30	1.00	-0.32	-0.10	0.27	0.14	0.25	0.06
5	-0.12	0.17	0.28	-0.32	1.00	0.23	-0.03	-0.02	-0.19	0.23
6	0.19	0.21	0.52	-0.10	0.23	1.00	0.13	0.10	-0.26	0.12
7	0.42	0.08	0.11	0.27	-0.03	0.13	1.00	-0.07	0.09	0.38
8	0.24	0.10	0.07	0.14	-0.02	0.10	-0.07	1.00	-0.21	0.01
9	0.17	0.34	-0.26	0.25	-0.19	-0.26	0.09	-0.21	1.00	0.15
10	0.19	0.06	-0.01	0.06	0.23	0.12	0.38	0.01	0.15	1.00

Si osserva una struttura non regolare, caratterizzata da coefficienti di correlazione che variano in modo sostanziale a seconda della distanza tra le misurazioni. Le correlazioni tra misurazioni consecutive risultano talvolta deboli, ad esempio tra la prima e la seconda osservazione (-0.03), talvolta moderatamente forti, ad esempio tra la terza e la quarta misurazione (-0.30). Questo risultato sembra suggerire che la risposta non è fortemente correlata tra misurazioni successive, ma potrebbe essere un artefatto dovuto ad un'alta variabilità tra soggetti. Il fatto che non emerga una chiara struttura di decadimento nella matrice di correlazione suggerisce che le misurazioni non sono temporalmente stazionarie o che ci sono forti variazioni legate al contesto dei pasti e al soggetto. Infatti, i pasti non sono stati consumati allo stesso orario da parte dei diversi soggetti o tra le varie giornate

di uno stesso individuo, facendo sì che la risposta glicemica possa variare in modo non sistematico. Inoltre, la struttura osservata nella matrice di correlazione potrebbe riflettere non solo la ciclicità dei pasti, ma anche la loro interazione con i ritmi circadiani dei soggetti, con risposte diverse al mattino rispetto alla sera. Per questo motivo si ritiene importante inserire l'orario del pasto come covariata nei successivi modelli. A distanze elevate tra misurazioni compaiono forti correlazioni; ciò è dovuto al fatto che solamente una porzione ridotta di soggetti ha un elevato numero di osservazioni e quindi le correlazioni finali sono calcolate su pochissime coppie di valori, risultando instabili e spurie.

Oltre al calcolo della matrice di correlazione tra misurazioni ordinate per ciascun soggetto, è stata valutata anche la correlazione tra misurazioni in funzione della distanza temporale. Sono stati definiti degli intervalli temporali di tre ore ciascuno e sono state calcolate le correlazioni empiriche tra le misurazioni di ciascun soggetto comprese in questi intervalli. L'andamento della correlazione media in funzione della distanza temporale è mostrato in Figura 5.1.

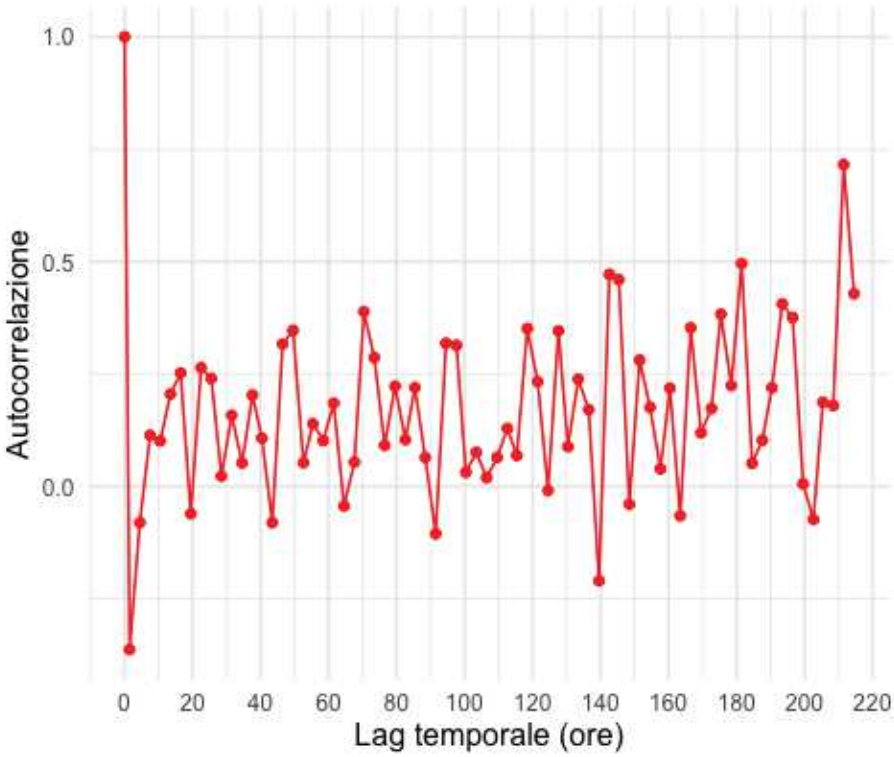

FIGURA 5.1: Funzione di autocorrelazione empirica rispetto alla distanza temporale tra misurazioni intra-soggetto.

L'andamento della funzione di autocorrelazione empirica risulta irregolare, con una forte instabilità delle correlazioni e l'assenza di un chiaro decadimento esponenziale, che sarebbe atteso in presenza di una struttura di tipo AR(1). Nella prima parte del grafico,

la maggior parte delle correlazioni assume valori nel range (0.1, 0.3), indicando una dipendenza temporale debole. Per lag più elevati si notano alcuni picchi, i quali però sono probabilmente dovuti alla bassa numerosità di coppie osservate in quegli intervalli. Da questa analisi non emerge dunque un pattern di autocorrelazione temporale sistematica che giustifichi l'introduzione di strutture di correlazione addizionali.

5.3 Predittori

Le covariate che vengono inserite nel modello includono le variabili demografiche, le variabili cliniche, i punteggi relativi al microbioma intestinale, le informazioni sulla composizione dei pasti e le variabili costruite in fase di pre-processing. Sono state inoltre costruite alcune interazioni, ritenute potenzialmente utili per descrivere la risposta. In particolare, si considera l'interazione tra tempo ciclico e carboidrati sulla base dell'idea che l'effetto del macronutriente possa variare a seconda dell'ora del pasto: a parità di quantità di carboidrati, la risposta glicemica può essere più alta al mattino che alla sera, o viceversa. Questo effetto non è catturato dalla variabile relativa ai carboidrati e all'ora del pasto presi singolarmente, quindi può essere utile l'interazione. Si inserisce inoltre l'interazione tra i carboidrati e l'indice HbA1c per considerare la possibilità che l'impatto di una certa quantità di carboidrati sulla glicemia postprandiale sia diverso per soggetti diabetici rispetto a soggetti sani. Infine, si include l'interazione tra i carboidrati e il livello glicemico un'ora prima del pasto per valutare se l'effetto di un pasto ricco di carboidrati può essere amplificato o attenuato a seconda della glicemia basale.

5.4 Convalida incrociata a finestra espandibile

Al fine di ottenere una valutazione onesta dell'efficacia predittiva di un modello si decide di sfruttare una procedura di convalida incrociata. In particolare, si ricorre ad una tecnica generalmente usata in ambito di serie storiche che permette di tenere conto dell'ordine temporale delle osservazioni. La differenza rispetto alla convalida incrociata standard è che i diversi insiemi di verifica (*test set*) ottenuti seguono temporalmente gli insiemi di stima (*training set*). Il metodo utilizzato è quella della convalida incrociata a finestra espandibile (*Expanding Window Cross-validation*). L'insieme di stima è inizialmente di dimensione ridotta, poi viene espanso gradualmente per includere un maggior numero di osservazioni. Il modello è adattato per ogni finestra temporale considerata e viene valutato sulle osservazioni escluse, ovvero quelle successive. Il meccanismo è mostrato in Figura 5.2.

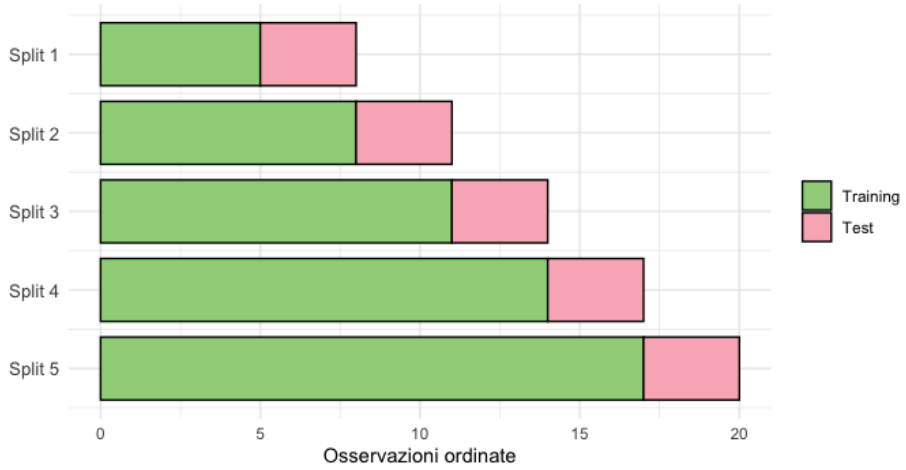

FIGURA 5.2: Schema di convalida incrociata a finestra espandibile.

Disponendo di un numero ridotto di osservazioni per ciascun soggetto, si procede nel seguente modo:

- Si escludono le ultime 3 osservazioni e si usano le precedenti per adattare il modello. Il training set sarà pertanto costituito dalle prime $n_i - 3$ misurazioni per ogni soggetto, mentre il test set conterrà la terzultima osservazione di ogni partecipante e avrà numerosità pari al numero di soggetti;

- Si eslcudono le ultime 2 osservazioni e si usano le precedenti per adattare il modello. Il training set sarà ora costituito dalle prime $n_i - 2$ misurazioni per ogni soggetto, mentre il test set includerà la penultima osservazione di ogni partecipante;

- Si eslcude l'ultima osservazione e si usano le precedenti per adattare il modello. Il training set sarà ora costituito dalle prime $n_i - 1$ rilevazioni per ogni soggetto, mentre il test set conterrà l'ultima osservazione di ogni partecipante.

La procedura di validazione predittiva descritta è utile in questo contesto di dati longitudinali in quanto le osservazioni hanno una bassa numerosità, specialmente per alcuni soggetti. Non richiede di dividere casualmente le osservazioni in training set e test set, ma permette di sfruttare tutti i dati disponibili per ogni previsione, compatibilmente con la loro posizione temporale. Al termine della procedura, si dispone di tre previsioni per ciascun soggetto, ottenute con modelli adattati sulla base dell'informazione disponibile fino a quel momento.

5.5 Modello lineare

Si procede adattando un modello lineare alla variabile risposta netAUC. Si adotta la procedura di convalida incrociata con finestra mobile specificata in precedenza. I risultati della procedura di stima sul training set costituito dalle prime $n_i - 1$ misurazioni per ciascun soggetto sono riportati in Tabella A.5 in Appendice. La variabile risposta risulta influenzata da una combinazione complessa di variabili nutrizionali, cliniche, microbiologiche e temporali. I carboidrati hanno un ruolo centrale e il loro effetto dipende sia da caratteristiche individuali che dal contesto del pasto. In particolare, emerge che:

- Tra le caratteristiche demografiche, risulta significativo il genere con coefficiente positivo, indicando che a parità delle altre variabili, essere uomini aumenta la risposta glicemica al pasto. L'etnia non mostra un effetto significativo sulla variabile risposta;
- I carboidrati sono associati a un aumento della risposta, mentre le proteine mostrano un effetto negativo, così come i grassi;
- La quantità di macronutrienti consumata in diversi intervalli temporali antecedenti al pasto non ha un effetto significativo sul netAUC;
- Le calorie bruciate nelle due ore successive al pasto sono associate negativamente alla variabile risposta, indicando una relazione inversa tra dispendio calorico postprandiale e risposta glicemica;
- Il glucosio misurato un'ora prima del pasto è fortemente associato a una riduzione della risposta;
- Il valore di HbA1c ha un effetto positivo fortemente significativo sul netAUC;
- Tra le variabili cliniche risulta che il livello di insulina e di colesterolo LDL sono associati negativamente alla risposta;
- L'età ha un effetto negativo sulla variabile risposta, mentre il BMI è associato positivamente;
- Tra le misure cardiovascolari, risultano significative il coefficiente angolare della regressione lineare tra la frequenza cardiaca e il tempo nei 120 minuti dopo il pasto e il numero di salti bruschi, con effetto positivo e negativo, rispettivamente;
- Le variabili relative ai punteggi intestinali risultano significative, eccetto quella relativa ai pathway coinvolti nella biosintesi dei lipopolisaccaridi, ad indicare un

ruolo attivo del microbioma intestinale nella regolazione della risposta postprandiale. Alcune variabili hanno effetto positivo, ad esempio la salute del rivestimento intestinale e i pathway coinvolti nell'assemblaggio dei flagelli, mentre altre negativo, ad esempio i pathway coinvolti nella produzione di butirrato o nella produzione di acido urico;

- L'interazione tra carboidrati e trasformata seno di Fourier dell'orario del pasto è significativa confermando che l'orario di assunzione del pasto influenza l'effetto dei carboidrati sulla risposta;

- Infine, anche l'interazione tra carboidrati e livello glicemico un'ora prima del pasto risulta significativa, indicando come l'effetto dei carboidrati sulla risposta sia influenzato dal livello glicemico individuale antecedente il pasto.

Il modello spiega circa il 57% della variabilità osservata nella variabile risposta, come indicato dal valore di R^2 pari a 0.566. Per valutare l'adattamento del modello lineare sull'insieme di stima si procede all'analisi dei residui. Si valutano i 5 numeri di Tukey per la distribuzione dei residui, ossia minimo, massimo e i quartili primo, secondo (mediana) e terzo. La valutazione di questi valori permette di comprendere se risulti ragionevole l'ipotesi di simmetria, che è una diretta conseguenza dell'ipotesi di normalità. In questo caso, i 5 numeri di Tukey risultano pari a (-191.34; -25.24; -1.20; 22.84; 190.30) e vi è pertanto una leggera asimmetria verso i valori negativi. I valori minimo e massimo sono grandi in valore assoluto, indicando la presenza di possibili *outliers*, ossia di valori della variabile risposta che si discostano dallo schema mostrato dalla maggior parte dei dati. Successivamente, si procede all'analisi grafica dei residui, riportata in Figura 5.3. L'analisi diagnostica del modello lineare evidenzia alcune lievi deviazioni dalle assunzioni classiche. In particolare, i residui mostrano una moderata eteroscedasticità, con una varianza crescente per valori previsti più elevati. Il Q-Q plot indica che la distribuzione dei residui si avvicina alla normalità, con alcune deviazioni nelle code. Inoltre, un'osservazione (n. 816) presenta una distanza di Cook elevata, suggerendo un possibile effetto influente sulla stima del modello. Nel complesso, il modello mostra un buon adattamento.

La Figura 5.4 mostra la distribuzione dei residui sui dati di stima separamente per ciascun soggetto. La maggior parte dei residui è centrata attorno allo zero, tuttavia si osservano differenze specifiche tra soggetti: per alcuni è evidente una maggiore dispersione, indicando che il modello in quei casi non risulta adeguato. Inoltre, per alcuni partecipanti si osserva una distribuzione dei residui concentrata su valori negativi, ad esempio per i soggetti 16 e 17, mentre per altri individui, tra cui il numero 18 e il numero

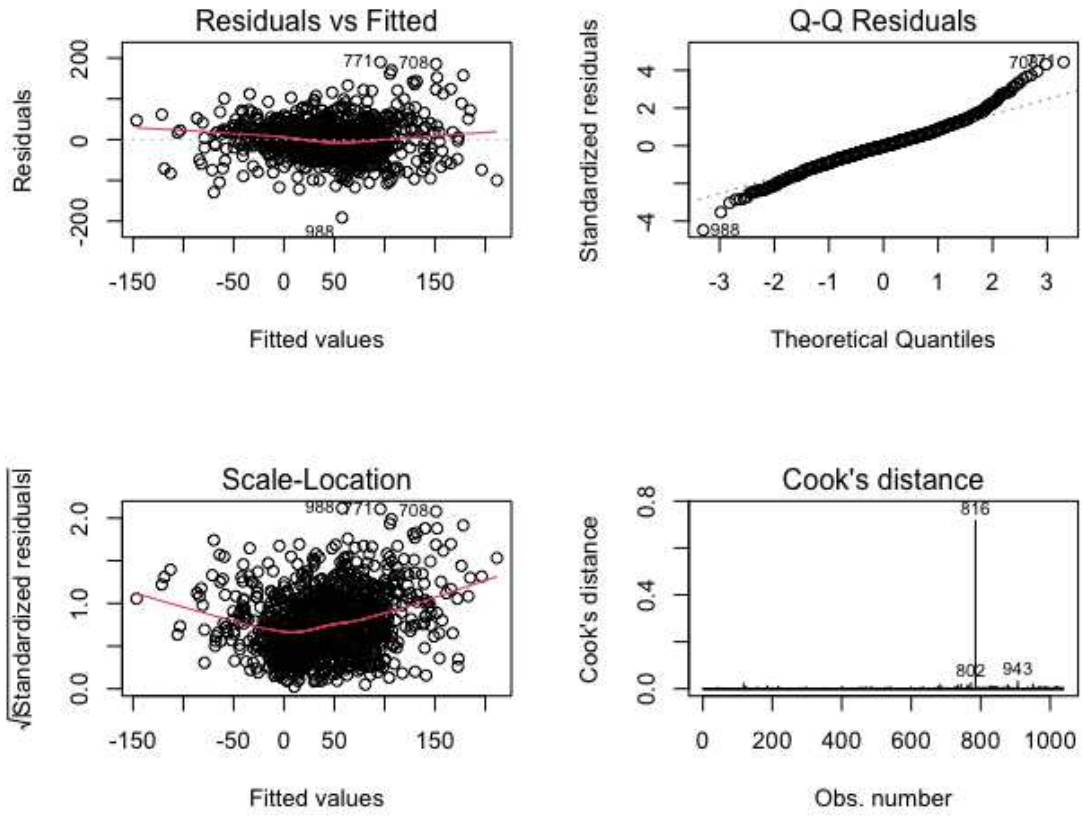

FIGURA 5.3: Analisi dei residui.

32, si nota una tendenza del modello a sottostimare i valori della risposta. La presenza di outliers è evidente in diversi soggetti.

Si procede ora valutando l'accuratezza predittiva del modello lineare sulla base delle tre osservazioni ottenute per ciascun partecipante tramite la procedura di convalida incrociata a finestra espandibile. Come metrica si utilizza la radice dell'errore quadratico medio (*RMSE*), risultata pari a 41.79. Per l'analisi dei valori previsti si impiegano anche strumenti grafici; il confronto tra valori osservati e previsti viene effettuato sia in forma aggregata (Figura 5.5), che separatamente per ciascun soggetto (Figura 5.6). Nel primo caso, si osserva una buona aderenza generale alla bisettrice, con la maggior parte dei punti distribuiti attorno alla diagonale. Questo suggerisce che il modello è in grado di catturare la tendenza della risposta glicemica. Tuttavia, si notano alcuni casi in cui vi è una differenza marcata tra valore osservato e previsto, specialmente quando il primo è sulle code della distribuzione.

Considerando invece le previsioni in forma disaggregata per ciascun soggetto, si nota che alcuni partecipanti sono caratterizzati da valori più regolari, che tendono a disporsi attorno alla bisettrice. Tuttavia, in diversi casi i valori previsti risultano distanti da

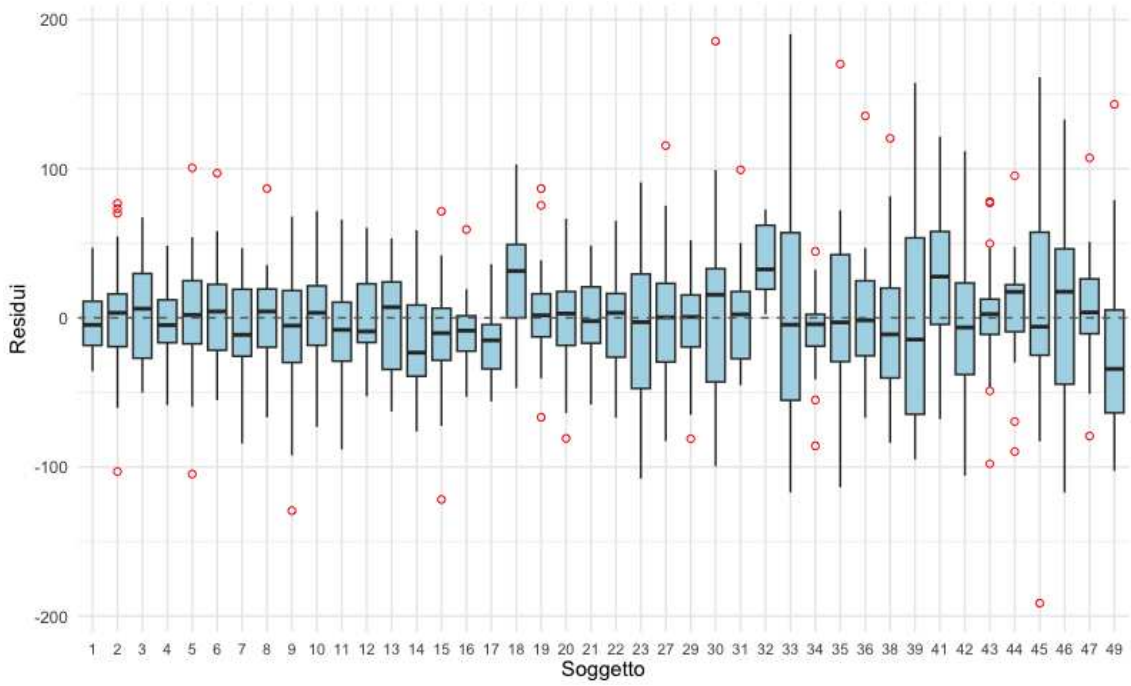

FIGURA 5.4: Distribuzione dei residui del modello lineare per soggetto.

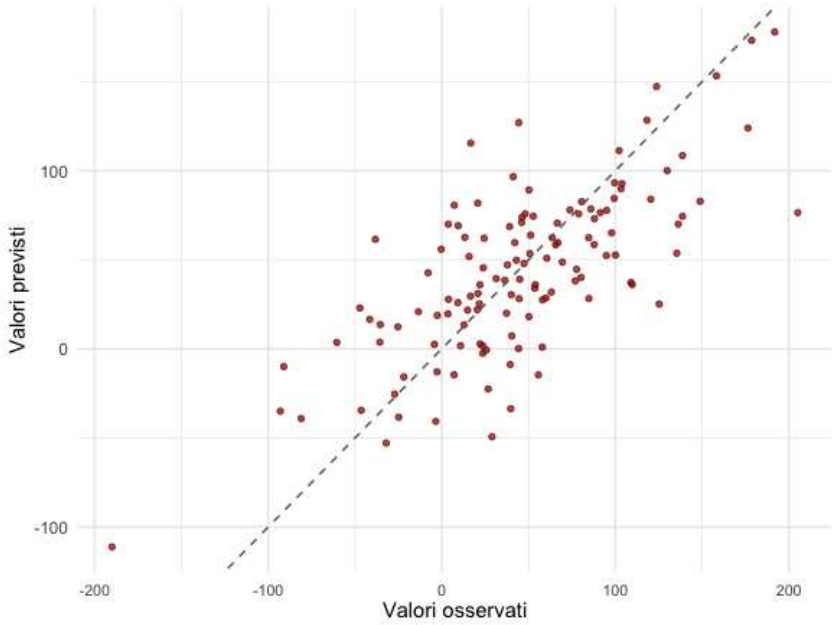

FIGURA 5.5: Valori previsti rispetto ai valori osservati.

quelli reali, suggerendo una variabilità residua non trascurabile a livello individuale, dovuta in parte anche al numero ridotto di osservazioni per soggetto. Questa visualizzazione consente di comprendere per quali soggetti il modello dimostra un'accuratezza predittiva maggiore e per quali invece commette errori sistematici, indicando la presenza di eterogeneità inter-individuale, che potrebbe giustificare l'inclusione di effetti casuali.

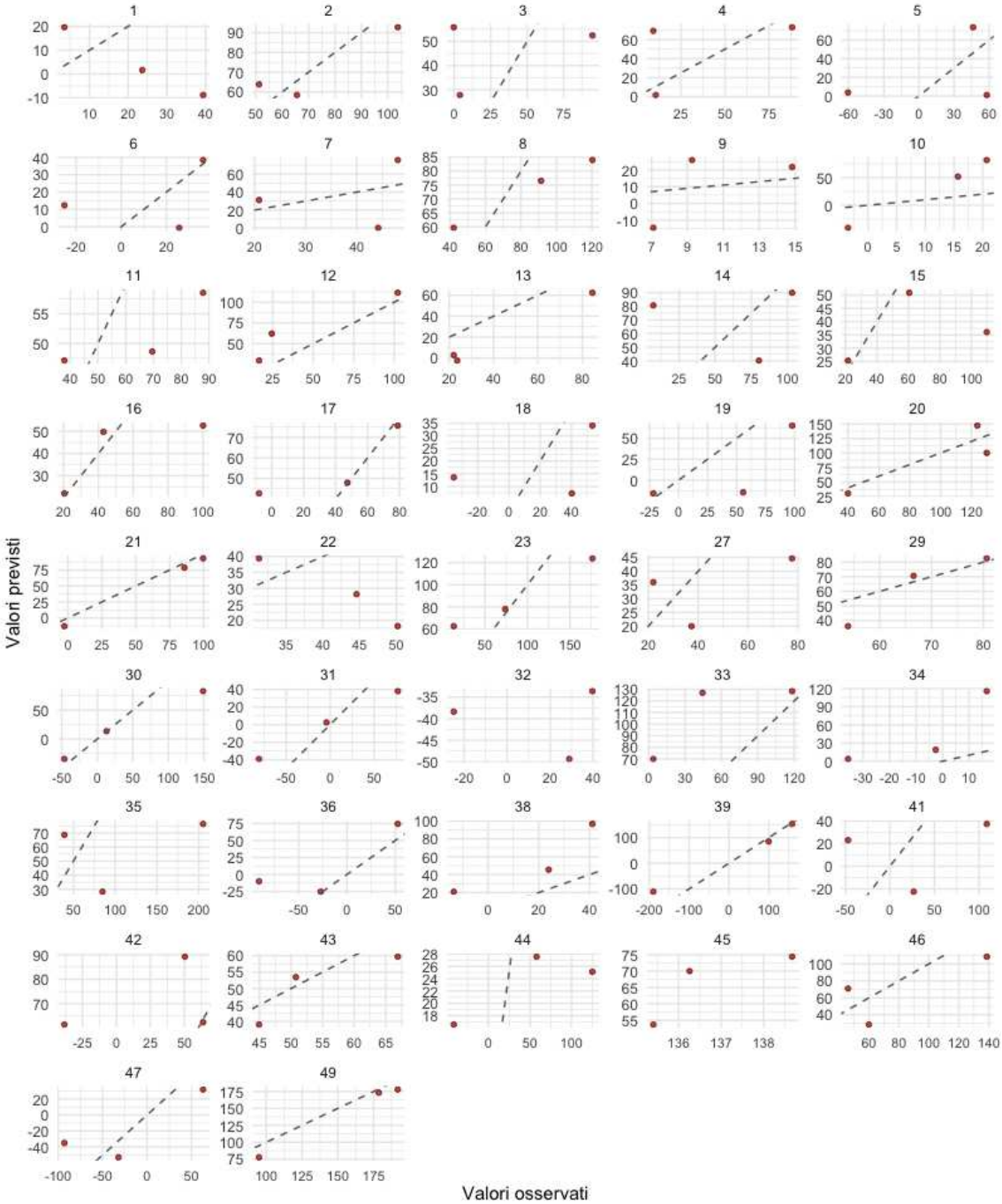

FIGURA 5.6: Valori previsti rispetto ai valori osservati.

5.6 Modelli lineari a effetti misti

Si valuta ora l'inclusione nel modello di effetti casuali. Inizialmente si inserisce un'intercetta casuale per soggetto, in modo da permettere ad ogni partecipante di avere un livello medio diverso per la variabile risposta. Si considerano come effetti fissi gli stessi valutati inizialmente per il modello lineare e si procede tramite convalida incrociata a finestra espandibile. L'RMSE risulta pari a 41.14, comparabile a quello ottenuto con il

modello lineare. Si procede all'analisi dell'adattamento del modello sull'insieme di stima all'ultimo step della procedura di convalida incrociata. La distribuzione dei residui sui dati di stima separamente per ciascun soggetto, mostrata in Figura 5.7, risulta più simmetrica rispetto a quella ottenuta in precedenza. Si osservano miglioramenti per la maggior parte dei soggetti, inclusi ad esempio il numero 7, 17 e 32.

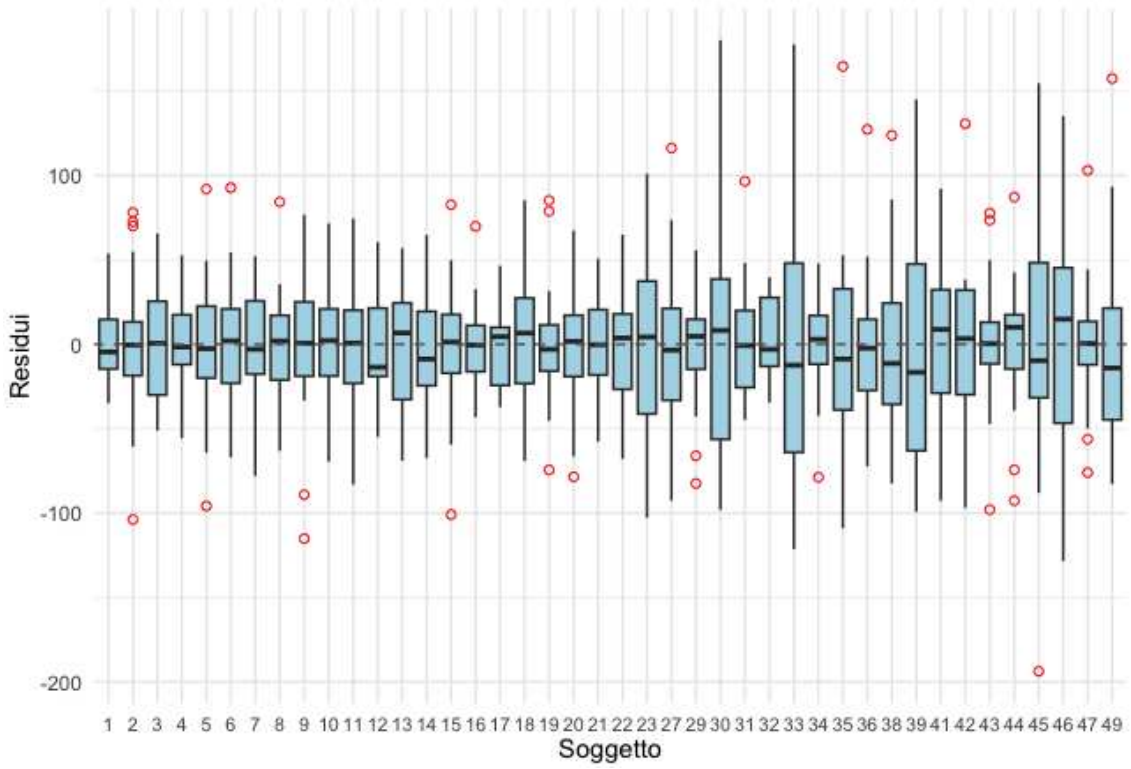

FIGURA 5.7: Distribuzione dei residui del modello lineare con intercetta casuale per soggetto.

Dal riepilogo dei risultati del modello si ricava che la mediana dei residui standardizzati è prossima a zero (0.0037), mentre i valori estremi sono elevati, pari a -4.48 e 4.17, indicando la presenza di outliers. La variabilità tra i soggetti nel livello medio della variabile risposta è racchiusa dalla deviazione standard associata all'intercetta casuale, pari a 30.622. La deviazione standard residua (σ_ε), che indica quanta variabilità all'interno di ciascun soggetto non è stata spiegata dal modello, è pari a 43.197. Per calcolare la proporzione di varianza spiegata dal soggetto si ricorre al coefficiente di correlazione intraclasse (ρ), che si ottiene tramite il seguente calcolo

$$\rho = \frac{\sigma^2_{casuale}}{\sigma^2_{casuale} + \sigma^2_\varepsilon}.$$

In questo caso, risulta pari a

$$\rho = \frac{30.622^2}{30.622^2 + 43.197^2} = 0.334.$$

Ciò significa che circa il 33% della varianza totale è attribuibile alle differenze tra soggetti, mentre il restante 67% è intra-soggetto, cioè tra i pasti dello stesso individuo. Questo risultato giustifica l'utilizzo di un modello a effetti misti con intercetta casuale per soggetto, ma suggerisce anche che la risposta glicemica postprandiale varia molto dentro allo stesso soggetto, in base alla composizione del pasto, l'attività cardiaca e l'orario del pasto. Le stime dei coefficienti e i relativi errori standard sono riportati in Appendice in Tabella A.6. Per quanto riguarda l'interpretazione dei coefficienti, risulta che:

- I carboidrati sono associati a un forte aumento del netAUC, mentre le proteine sono associate ad una riduzione e così anche i grassi;
- La somma delle calorie bruciate nelle due ore successive al pasto risulta influenzare negativamente la risposta;
- L'HbA1c risulta avere un'influenza positiva, seppur debole sulla riposta;
- Il livello di glucosio misurato un'ora prima del pasto è fortemente associato a una riduzione della risposta così come lo stesso valore registrato quattro ore prima del pasto;
- L'interazione tra i carboidrati e la trasformata seno di Fourier dell'orario del pasto risulta significativa, con coefficiente positivo.
- Anche l'interazione tra i carboidrati e il livello glicemico un'ora prima del pasto risulta significativa, con coefficiente positivo.

Molte variabili risultano non significative. Inoltre, alcune presentano errori standard (se) elevati, ad esempio il BMI (se = 19.80), il genere (se=34.90) e la maggior parte delle variabili associate ai punteggi intestinali. Il fatto che gli errori standard risultino tendenzialmente più elevati rispetto a quelli del modello lineare classico è atteso, in quanto il modello con intercetta casuale considera la dipendenza tra osservazioni relative ad uno stesso soggetto, mentre il modello lineare assume che tutte le osservazioni siano indipendenti e con stessa varianza, portando ad una sottostima della variabilità residua. L'aumento marcato degli errori standard per variabili come il Genere, il BMI o i punteggi intestinali è attribuile al fatto che queste variabili sono costanti per soggetto. Il modello

a effetti misti riconosce che vi sono meno informazioni indipendenti per stimare l'effetto di una variabile che non cambia tra i pasti dello stesso soggetto e restituisce errori standard più alti.

Gli errori standard elevati potrebbero essere dovuti anche ad altri fattori, tra cui la collinearità tra predittori, la bassa numerosità campionaria o il fatto di aver adattato un modello troppo complesso per i dati a disposizione. Se vi è collinearità tra le variabili esplicative originali, i coefficienti stimati possono risultare fortemente correlati, ad indicare ridondanza informativa. Pertanto, si procede ad un'analisi della correlazione tra le stime dei coefficienti, che mostra alcune forti associazioni. I risultati evidenziano la presenza di numerose coppie con correlazione superiore a $|0.90|$. In particolare, si osservano forti associazioni tra variabili microbiologiche (es. pathway coinvolti nella produzione di butirrato, metabolismo dell'acido biliare, produzione di ammoniaca), tra variabili nutrizionali (es. calorie assunte nelle 6 ore antecedenti il pasto e carboidrati assunti nella stessa finestra temporale), tra variabili cliniche (es. colesterolo LDL, colesterolo), e tra indicatori cardiovascolari (es. il coefficiente angolare della regressione lineare tra la frequenza cardiaca e il tempo nei 120 minuti dopo il pasto e l'iAUC del battito cardiaco). Anche l'etnia risulta fortemente correlata con punteggi intestinali e demografici. La conseguenza è che le stime dei coefficienti e i p-value potrebbero essere inaffidabili. Dal confronto tra le matrici di correlazione dei coefficienti stimati nel modello lineare semplice e in quello con intercetta casuale, non emergono differenze rilevanti. In particolare, si nota che la media delle correlazioni dei coefficienti in valore assoluto è leggermente più bassa nel modello con intercetta casuale (0.12), rispetto al modello lineare classico (0.141), ad indicare che l'inclusione della variabilità soggetto-specifica riduce la dipendenza tra i coefficienti fissi.

5.6.1 Pendenza casuale

Data l'elevata dimensionalità del dataset, non è possibile svolgere un confronto tra modelli a effetti misti con e senza pendenza casuale per ogni predittore. Tramite l'applicazione del metodo di analisi esplorativa proposto da Schelldorfer et al. (2011) e presentato al paragrafo 2.4.1.2, è stato ricavato un sottoinsieme di predittori potenzialmente associati a un effetto casuale. Come riportato in Appendice A.3.1.2, tra questi compaiono i carboidrati, le fibre, le calorie assunte nelle 6 ore antecedenti il pasto e alcune trasformazioni di Fourier dell'orario di assunzione del pasto. Tuttavia, l'inclusione simultanea di tutte queste variabili non risulta praticabile nel presente contesto, in quanto comporta un forte aumento della complessità e instabilità della stima. Per questo motivo, si è deciso di concentrarsi sui carboidrati, per valutare se il loro effetto vari tra i soggetti.

Tale scelta dipende dal fatto che, tra i diversi macronutrienti, i carboidrati rivestono un ruolo centrale nel regolare la variazione glicemica postprandiale. Pertanto, analizzare la variabilità inter-individuale dell'effetto dei carboidrati risulta cruciale per comprendere meglio la relazione tra le covariate e la risposta. La stima dei termini di varianza per gli effetti casuali indica una marcata eterogeneità tra soggetti, sia a livello dell'intercetta (deviazione standard = 32.096) sia nella sensibilità ai carboidrati (deviazione standard = 10.247). L'effetto legato all'intercetta e quello relativo alla pendenza casuale sono correlati positivamente ($\rho = 0.683$), suggerendo che i soggetti con livelli glicemici medi più elevati tendono a rispondere in maniera più marcata all'assunzione di carboidrati. Questi risultati confermano una rilevante variabilità tra partecipanti nell'effetto dei carboidrati. La Figura 5.8 mostra, per ciascun soggetto, la previsione dell'effetto casuale associato ai carboidrati, ossia quanto l'effetto individuale si discosta dall'effetto fisso stimato dal modello. La linea tratteggiata ad altezza zero corrisponde all'effetto fisso medio dei carboidrati. Si osserva una notevole variabilità delle previsioni, con valori che vanno da circa –14 a +14. I valori negativi sono relativi a individui per cui l'effetto dei carboidrati sulla risposta glicemica è più contenuto rispetto alla media, mentre valori positivi indicano un effetto più marcato, tipico dei soggetti maggiormente *sensibili* all'assunzione di carboidrati.

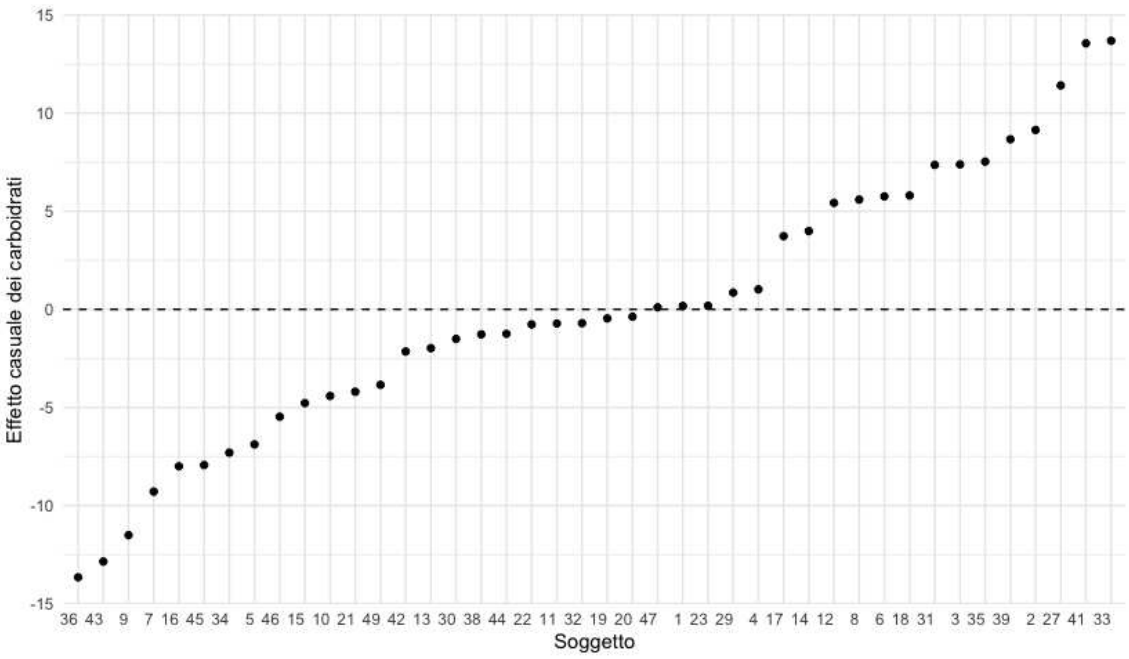

FIGURA 5.8: Previsione dello scostamento dell'effetto dei carboidrati dal valore medio, per ciascun soggetto.

Le stime dei coefficienti degli effetti fissi sono riportate in Appendice in Tabella A.7. I risultati mostrano che:

- I carboidrati hanno un effetto positivo significativo sulla variabile risposta, mentre proteine e grassi mostrano effetti negativi;
- La somma delle calorie bruciate nelle due ore successive al pasto è associata negativamente alla risposta glicemica;
- Il livello glicemico un'ora prima del pasto influenza negativamente la risposta;
- L'interazione tra i carboidrati e il valore HbA1c è significativa, con coefficiente positivo, suggerendo che l'effetto dei carboidrati è influenzato dalla condizione glicemica del soggetto;
- Anche l'interazione tra i carboidrati e la trasformata seno di Fourier dell'orario del pasto risulta avere un effetto significativo sul netAUC, con effetto simile a quello emerso nel modello precedente;
- Al contrario, molte variabili legate alla composizione del microbioma intestinale e alla frequenza cardiaca non raggiungono la significatività statistica.

Il confronto tra il modello con pendenza casuale per i carboidrati e quello con solo l'intercetta come effetto casuale avviene tramite ANOVA. Pertanto, è necessario stimare entrambi i modelli con il metodo di massima verosimiglianza (ML). Tuttavia, si riscontra un problema di convergenza nel modello con intercetta e pendenza casuale stimato con il metodo ML quando si includono tutte le variabili relative ai punteggi intestinali. La mancata convergenza potrebbe dipendere da diversi fattori, tra cui il fatto che l'utilizzo di ML, rispetto a REML, comporta una maggiore instabilità numerica in presenza di un elevato numero di predittori. Inoltre, potrebbero esserci problemi di collinearità o bassa variabilità di alcune variabili, tale per cui l'informazione disponibile per la stima è ridotta. Per questo motivo, il confronto viene condotto su un sottoinsieme di predittori. In particolare, si escludono le variabili relative alla salute del rivestimento intestinale e ai pathway coinvolti nella biosintesi dei lipopolisaccaridi. Il confronto ANOVA porta a preferire il modello con pendenza casuale per i carboidrati ($p\text{-}value = 0.033$), suggerendo quindi che la risposta glicemica postprandiale vari non solo tra soggetti ma anche nella sensibilità individuale ai carboidrati. La varianza degli stimatori dei coefficienti fissi risulta tendenzialmente maggiore nel modello a effetti misti con intercetta e pendenza casuale rispetto al modello con solo intercetta casuale. In particolare, l'effetto fisso medio dei carboidrati viene stimato tenendo in considerazione la variabilità tra soggetti nella sensibilità ai carboidrati. Di conseguenza, aumenta l'incertezza relativa al coefficiente fisso associato ai carboidrati e alle sue interazioni. Ad esempio, l'errore

standard associato all'interazione tra carboidrati e HbA1c cresce da 1.66 a 2.56.

Dall'adattamento del modello lineare e del modello lineare a effetti misti, si ricavano utili indicazioni. In particolare, emerge la possibilità di esplorare l'adozione di tecniche di selezione o penalizzazione delle variabili per migliorare la stabilità e l'interpretabilità del modello. Inoltre, risulta opportuno esplorare maggiormente l'inclusione nel modello di una pendenza casuale per i carboidrati, al fine di catturare eventuali differenze inter-individuali nella loro associazione con la risposta.

5.7 Selezione stepwise

Si applica una procedura di selezione stepwise ibrida delle variabili e si confrontano i risultati con quelli del modello lineare completo adattato in precedenza. Innanzitutto, il modello con selezione stepwise conferma il ruolo centrale che rivestono i carboidrati nell'incrementare la risposta glicemica postprandiale, così come l'effetto negativo delle proteine e dei grassi. Inoltre, rileva anch'esso che il genere maschile è associato a un aumento significativo della risposta. In questo caso, anche l'appartenenza al gruppo etnico degli ispanici/latini mostra un effetto positivo significativo sulla risposta. Vengono selezionate le variabili cardiovascolari risultate già significative nel modello lineare completo, cioè il coefficiente angolare della regressione tra frequenza cardiaca e tempo e il numero di salti bruschi. Inoltre, viene selezionata la media della frequenza cardiaca nei 30 minuti successivi al pasto, variabile che non era risultata significativa nel modello lineare completo. Il modello conferma l'effetto negativo delle calorie bruciate nelle due ore successive al pasto e del livello di glucosio un'ora prima del pasto, nonché l'effetto positivo dell'indicatore HbA1c. Anche l'età e il BMI mostrano lo stesso effetto emerso nel modello lineare completo, negativo e positivo, rispettivamente. Per quanto riguarda le variabili cliniche, entrambi i modelli identificano effetti negativi di insulina e colesterolo LDL. Una differenza importante riguarda la selezione delle variabili relative al microbioma intestinale. Infatti, mentre nel modello lineare completo la maggior parte dei punteggi intestinali erano risultati significativi, nel modello lineare stepwise viene effettuata una selezione più stringente. Infine, in entrambi i modelli risultano significative le interazioni tra carboidrati e trasformata seno di Fourier dell'orario del pasto e tra carboidrati e livello glicemico un'ora prima del pasto. La sintesi dei risultati del modello è presente in Appendice in Tabella A.8.

In termini di bontà d'adattamento, il modello lineare con selezione stepwise restituisce un indice R^2 pari a 0.5598, indicando che riesce a spiegare circa il 56% della variabilità

osservata nella variabile risposta. In termini di performance predittiva, presenta un RMSE pari a 41.15, paragonabile a quello del modello lineare. Complessivamente, la procedura di selezione stepwise ha permesso di ottenere un modello più parsimonioso, mantenendo i predittori già risultati importanti nel modello lineare completo.

5.8 Modelli lineari penalizzati

Si considera ora l'inclusione di un termine di penalità nella stima dei coefficienti, al fine di migliorare la stabilità del modello e gestire la complessità dei dati. L'approccio penalizzato può infatti essere utile nel gestire situazioni caratterizzate da elevata dimensionalità e collinearità tra predittori. Per questa ragione, in questa fase dell'analisi sono state reintrodotte le covariate precedentemente escluse dai modelli lineari per problemi di instabilità numerica o ridondanza informativa. Attenuando o annullando l'effetto delle variabili meno rilevanti, le tecniche di penalizzazione facilitano l'interpretazione del modello finale.

5.8.1 Penalizzazione Lasso

Si stima un modello di regressione lineare con penalità Lasso per selezionare le variabili rilevanti. La fase di individuazione del valore ottimale del parametro di regolazione λ viene inserita all'interno della procedura di convalida incrociata con finestra temporale espandibile. Si individuano due valori di λ: quello che minimizza l'errore di convalida incrociata (λ_{min}) e quello più parsimonioso entro uno standard error di distanza ($\lambda_{1\ se}$). Si calcolano le previsioni su ciascuna osservazione del test set usando i due valori di λ e si confrontano i risultati. In corrispondenza di $\lambda_{1\ se} = 1.283$, l'RMSE è risultato pari a 42.57, superiore rispetto a quello ottenuto per $\lambda_{min} = 0.125$, pari a 40.80. La prima soluzione seleziona 40 variabili, risultando più parsimoniosa, mentre la seconda ne include 74. Tra i predittori inclusi dal modello, vi sono i macronutrienti principali, come carboidrati e proteine, informazioni sul livello glicemico antecedente il pasto, l'orario del pasto, e numerose variabili relative al microbioma intestinale. Per tenere conto della presenza di variabili qualitative, si valuta la possibilità di adattare una penalizzazione di tipo Group Lasso. Il numero di variabili selezionate è elevato, pari a 80 e l'RMSE risulta pari a 40.64, ad indicazione del fatto che in questa situazione la penalizzazione di tipo Group Lasso non porta miglioramenti significativi in termini di performance predittiva. Il confronto tra i coefficienti stimati dal modello con penalizzazione Lasso e quello con penalizzazione Group Lasso mostra che le variabili che hanno un effetto maggiore sulla risposta, come i nutrienti principali e il livello glicemico un'ora prima

del pasto, sono selezionate in entrambi i modelli, con stime di intensità comparabile, a conferma della loro rilevanza predittiva. Il vantaggio del Group Lasso è che risulta migliore dal punto di vista interpretativo in quanto seleziona interi gruppi di variabili, caratteristica utile in questa analisi perché sono presenti molte covariate categoriali, codificate tramite variabili dummy, come i punteggi relativi al microbioma intestinale. Le stime dei coefficienti delle variabili selezionate da ciascun modello sono riportate in Tabella A.9 in Appendice.

5.8.2 Penalizzazioni non convesse

Oltre alle penalità Lasso e Group Lasso, sono stati considerati anche metodi di penalizzazione non convessa, come SCAD (*Smoothly Clipped Absolute Deviation*) e MCP (*Minimax Concave Penalty*) e la loro versione Group. Mentre la penalità di tipo Lasso comprime tutti i coefficienti verso zero, introducendo quindi distorsione per quelli associati alle variabili rilevanti, le penalizzazioni non convesse, come MCP e SCAD, applicano un tasso di penalizzazione decrescente all'aumentare della grandezza del coefficiente, comportandosi in modo simile alla penalità Lasso per coefficienti piccoli e in modo simile alla penalità Ridge per coefficienti grandi. In questo modo, favoriscono l'esclusione dal modello di variabili non rilevanti e al contempo evitano di comprimere eccessivamente i coefficienti delle variabili importanti.

Le penalizzazioni MCP e Group MCP hanno dato risultati simili in termini di performance predittiva, conducendo a un RMSE pari a 41.93 e 40.98, rispettivamente. Il valore ottimale del parametro di regolazione λ_{min} è risultato pari a 1.51 per MCP e 1.697 per Group MCP. I modelli sono molto parsimoniosi, infatti nel primo caso sono state selezionate 36 variabili, mentre nel secondo caso 35. Risultati simili sono ottenuti applicando le penalità SCAD e Group SCAD, che restituiscono un RMSE pari, rispettivamente, a 42.2 e 41.48. La procedura di convalida incrociata ha condotto ai seguenti valori del parametro di regolazione: $\lambda_{min} = 1.51$ per il modello lineare con penalità SCAD e $\lambda_{min} = 1.169$ per quello con penalità Group SCAD. I modelli includono 34 e 47 variabili, rispettivamente. Le variabili selezionate sono coerenti con quelle incluse dai modelli precedenti, anche in termini di direzione e magnitudine del loro effetto. Le penalizzazioni non convesse non apportano miglioramenti significativi in termini di RMSE. La Tabella A.9, riportata in Appendice, presenta le stime dei coefficienti delle variabili selezionate da ciascun modello.

5.9 Modelli lineari a effetti misti penalizzati

Per adattare un modello lineare a effetti misti con un termine di penalizzazione si utilizza la funzione `glmmLasso()` dell'omononima libreria `R`. Si specifica come famiglia di distribuzione quella gaussiana. I metodi presentati nella sezione 2.4.1.2 per la scelta del parametro di penalizzazione λ vengono inseriti all'interno della procedura di convalida incrociata a finestra espandibile. Una volta individuato il valore ottimale di λ il modello viene ristimato e viene utilizzato per il calcolo delle previsioni.

Tabella 5.2: Valori del parametro di penalizzazione ottenuti con 4 diversi metodi.

Insieme di stima	**Metodo 1**	**Metodo 2**	**Metodo 3**	**Metodo 4**
$n_i - 3$	0	30	696.8	150
$n_i - 2$	40	235	512.7	150
$n_i - 1$	0	455	462.8	100

Dalla Tabella 5.2, si osserva un comportamento sensibile alla quantità di dati di stima e al metodo utilizzato per la scelta di λ. Infatti, il valore ottimale di λ è fortemente influenzato dal numero di osservazioni incluse nell'insieme di stima, specialmente per i primi due metodi. Si nota che il primo metodo suggerisce di non penalizzare in due casi su tre e fa sì che il modello finale includa tutte le variabili. I metodi successivi attuano invece una selezione delle variabili. Per ottenere una valutazione della performance predittiva si è calcolato l'RMSE, riportato in Tabella 5.3.

Tabella 5.3: Valori del'RMSE ottenuti con 4 metodi.

Metodo 1	**Metodo 2**	**Metodo 3**	**Metodo 4**
45.20	48.43	40.74	40.64

I risultati suggeriscono di utilizzare come metodi per scegliere il parametro di penalizzazione quelli che, ad ogni iterazione, impiegano come valori di partenza le stime ottenute per il valore di λ precedente. Questo conduce sia ad una riduzione dell'RMSE, sia ad una riduzione del BIC e della devianza passando rispettivamente dal 1° al 3° metodo, e dal 2° al 4°. Si procede commentando i risultati del modello ottenuto utilizzando il terzo metodo per la scelta di λ ed escludendo dall'insieme di stima una sola osservazione. La Figura 5.9 mostra l'andamento del BIC in funzione di λ. Il valore ottimale di λ è stato selezionato in corrispondenza del minimo della curva.

La deviazione standard dell'intercetta casuale per soggetto è risultata pari a 44.319. Al fine di calcolare il coefficiente di correlazione intraclasse, è necessario ottenere una stima della varianza residua. Poiché la funzione `glmmLasso()` non restituisce direttamente la varianza residua, se ne calcola un'approssimazione tramite la formula (3.1). Si

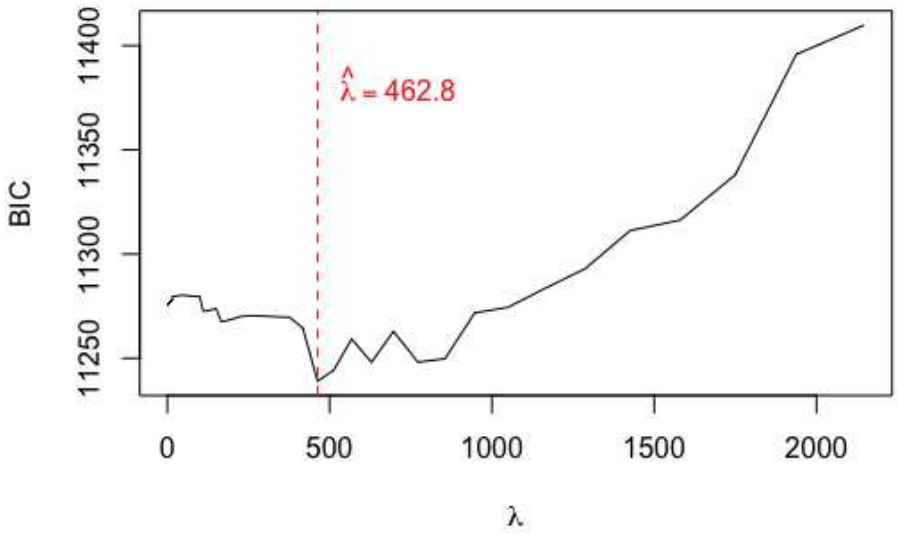

FIGURA 5.9: Selezione di λ tramite BIC.

ottiene che $\hat{df}_\lambda = 53 + 1 = 54$ e $\sigma_\varepsilon \approx 43.335$, da cui

$$\rho \approx \frac{44.319^2}{44.319^2 + 43.335^2} = 0.511.$$

Il valore del coefficiente di correlazione intraclasse è elevato, indicando la capacità del modello di catturare efficacemente la variabilità inter-individuale.

Le stime dei coefficienti degli effetti fissi selezionati sono riportate in Appendice, in Tabella A.11. Il modello include la maggior parte delle variabili considerate, risultando poco parsimonioso. Comprende i macronutrienti, l'età e alcune variabili cliniche, come HbA1c, insulina e colesterolo. Vengono selezionate anche diverse misure cardiovascolari, come la frequenza cardiaca media dopo il pasto e il numero di salti bruschi, i cui effetti hanno la stessa direzione già riscontrata nei modelli precedenti. Vengono selezionate anche variabili relative alla salute del microbioma intestinale. Inoltre, il modello mantiene le interazioni tra i carboidrati assunti e l'orario del pasto o il livello glicemico un'ora prima del pasto, suggerendo che l'effetto dei carboidrati sulla risposta postprandiale può essere modulato da queste covariate. Le variabili escluse dal modello, ovvero quelle con coefficiente stimato pari a zero a seguito della penalizzazione L_1, comprendono alcune covariate anagrafiche e demografiche come l'identificazione etnica, variabili nutrizionali specifiche, ad esempio carboidrati e fibre assunti in determinati intervalli temporali, misure antropometriche come BMI e peso corporeo, e alcuni punteggi di salute intestinale. È risultata non rilevante anche l'interazione tra carboidrati e HbA1c.

Si ristima nuovamente il modello considerando il valore di λ ottimale individuato, cioè λ pari a 462.8. In particolare, la stima finale avviene tramite procedura Fisher

scoring, con lo scopo di ottenere stime finali più accurate per i coefficienti delle variabili selezionate. Questo metodo consente di ottenere un p-value per valutare la significatività di ciascuna variabile. I risultati mostrano che le variabili fisse per ciascun soggetto, cioè il genere, l'etnia, i punteggi intestinali e le variabili cliniche, non sono significative.

Ora ci si concentra sul confronto della varianza dello stimatore tra il modello lineare con penalizzazione Group Lasso e il modello lineare a effetti misti con penalizzazione Group Lasso. Si stimano entrambi i modelli usando il pacchetto `glmmLasso` e specificando l'opzione `final.re=T`, che permette di ottenere una stima degli errori standard tramite l'applicazione dell'algoritmo di Fisher scoring. Nel caso del modello senza effetti casuali si specifica l'opzione `rnd=NULL`. Per il modello senza effetti casuali, il valore di λ che minimizza il BIC è pari a 980. Le stime dei coefficienti degli effetti fissi e dei relativi errori standard sono riportate in Appendice in Tabella A.10. Il confronto tra i due modelli mostra che aggiungere l'intercetta casuale comporta una penalizzazione leggermente più severa, infatti fa sì che vengano selezionate 53 variabili, contro le 57 variabili selezionate dal modello senza effetto casuale. Il modello con intercetta casuale non seleziona l'etnia autopercepita, BMI, Peso e altezza e alcune variabili relative a misurazioni cardiache e a pasti precedenti. Inoltre, diversi punteggi intestinali risultano esclusi, suggerendo che la loro informazione predittiva sia in gran parte sovrapposta alla componente soggetto-specifica, e dunque meno rilevante una volta modellata la gerarchia. Al contrario, variabili legate alla composizione del pasto, all'orario di assunzione e alla frequenza cardiaca postprandiale risultano selezionate in entrambi i modelli, con coefficienti simili in magnitudine e segno. Questo indica che la componente casuale non assorbe la variabilità delle covariate che variano tra i pasti, pertanto il loro contributo predittivo rimane stabile. Un comportamento diverso è associato alle variabili che rimangono costanti per ciascun individuo, per le quali si osservano spesso cambiamenti sia nella magnitudine, che nel segno del coefficiente stimato. Ad esempio, il coefficiente associato al colesterolo HDL passa da -5.39 nel modello con intercetta casuale a 11.59 nel modello senza. Tuttavia, nel primo caso lo standard error risulta maggiore della stima del coefficiente (se=10.90), facendo sì che la variabile, nonostante sia stata selezionata, non risulti significativa. Questo vale anche per le altre variabili che rimangono fisse tra i pasti di uno stesso soggetto, ad esempio l'errore standard associato all'effetto della variabile Età passa da 0.07 nel modello senza effetti casuali a 8.26 in quello con intercetta casuale. Questo incremento riflette il fatto che il modello misto riconosce la dipendenza intrinseca tra osservazioni dello stesso soggetto e aumenta la stima dell'incertezza di conseguenza. Per variabili con sufficiente variabilità intra-soggetto, come i

macronutrienti, il modello a effetti misti può ridurre l'incertezza, con errori standard comparabili o inferiori.

La distribuzione dei residui sui dati di stima separamente per ciascun soggetto viene riportata sia per il modello lineare con penalizzazione di tipo Group Lasso (Figura 5.10), che per il modello con intercetta casuale e analoga penalizzazione (Figura 5.11). Si osserva che l'aggiunta dell'intercetta casuale permette di centrare maggiormente la distribuzione dei residui attorno allo zero per ciascun soggetto. Inoltre, la variabilità intra-soggetto risulta leggermente ridotta rispetto al modello lineare senza effetti casuali.

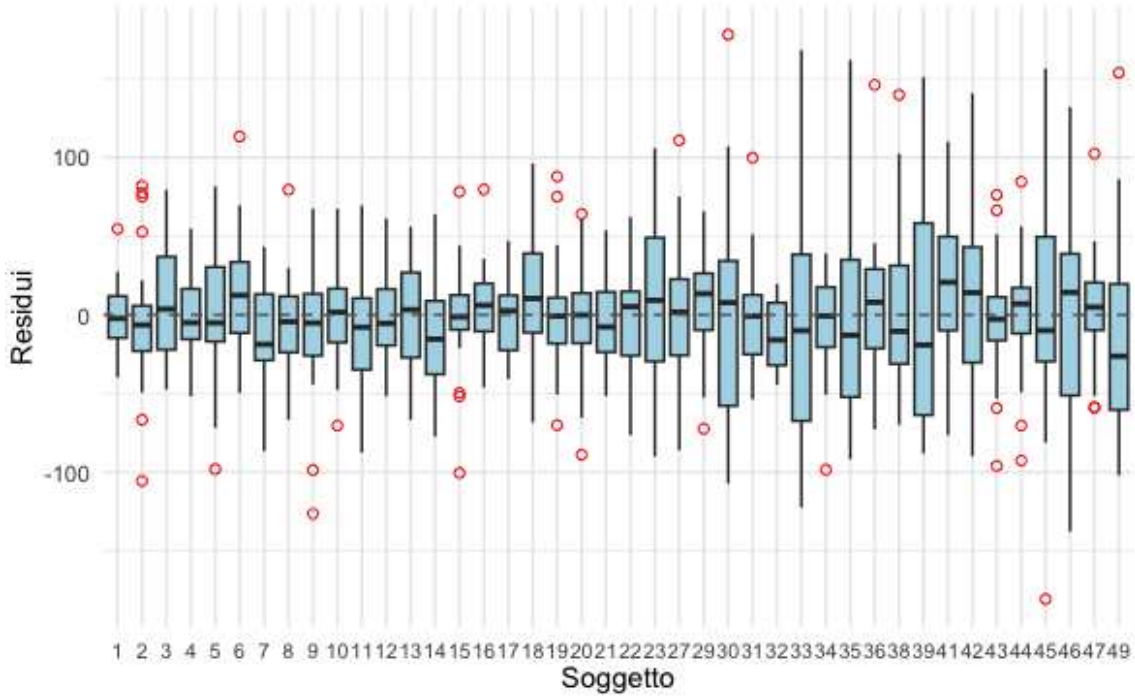

FIGURA 5.10: Distribuzione dei residui del modello lineare con penalizzazione Group Lasso per soggetto.

Per calcolare la matrice di correlazione tra le stime dei coefficienti relativi alle variabili selezionate da ciascuno dei due modelli, si implementa una procedura di tipo bootstrap. Ad ogni iterazione, i soggetti vengono estratti casualmente con reinserimento e, su ogni dataset bootstrap, si stima il modello includendo le variabili selezionate in corrispondenza del valore ottimale di λ. I coefficienti stimati sono stati aggregati in una matrice, a partire dalla quale è stata ricavata la matrice di correlazione. Dal confronto della correlazione media in valore assoluto tra i due modelli, emerge che aggiungere l'intercetta casuale ne comporta una riduzione, da 0.203 a 0.140. Il comportamento è simile a quello emerso nel confronto tra modello lineare e modello lineare a effetti misti, in assenza di penalizzazione. In pratica, l'assenza dell'effetto casuale costringe i coefficienti

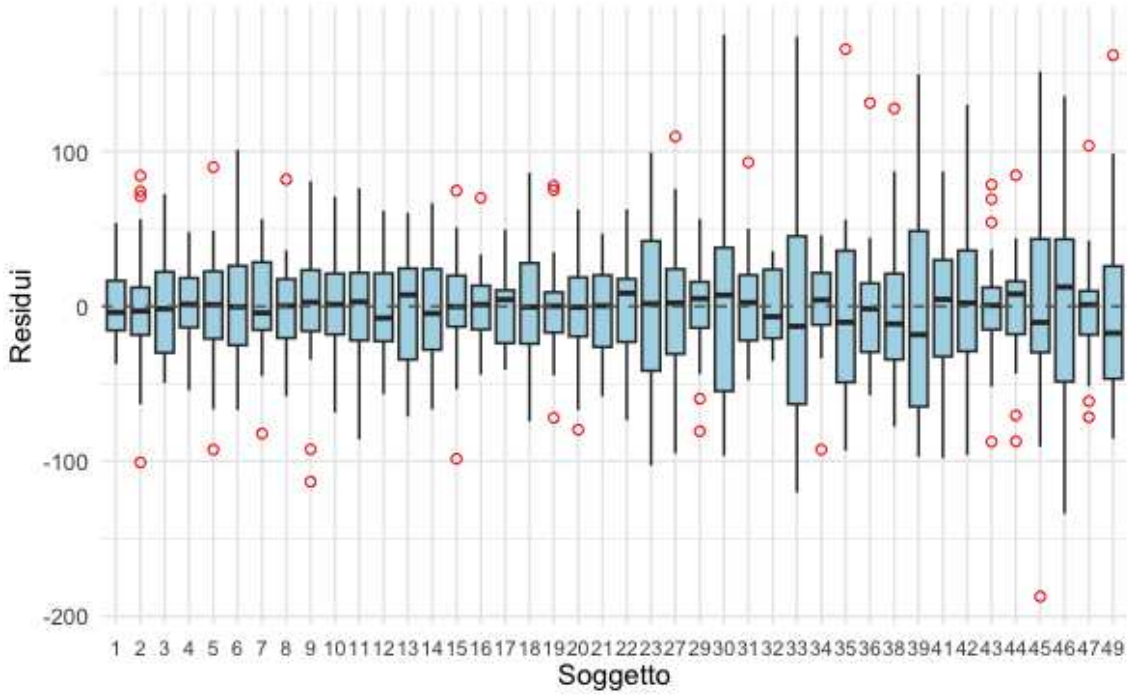

FIGURA 5.11: Distribuzione dei residui del modello lineare a effetti misti con penalizzazione Group Lasso per soggetto.

fissi ad assorbire anche la variabilità tra soggetti, aumentando la collinearità tra predittori. L'aggiunta dell'intercetta casuale consente di spiegare parte della varianza tra osservazioni correlate, conducendo ad una riduzione della collinearità spuria tra predittori fissi. Il confronto realizzato non tiene però conto del fatto che i due modelli hanno selezionato un sottoinsieme parzialmente diverso di variabili. Pertanto, si è deciso di procedere con il confronto delle correlazioni sul sottoinsieme comune di predittori, al fine di valutare l'impatto della specificazione del modello sulla dipendenza tra coefficienti stimati. Anche in questo caso, la correlazione risulta minore nel modello con intercetta casuale, pari a 0.146, mentre nel modello senza effetto casuale risulta uguale a 0.161. In sintesi, il confronto tra i modelli evidenzia come la presenza di un effetto casuale incida profondamente sia sulla selezione delle covariate, che sulla stima dell'incertezza.

5.9.1 Pendenza casuale

Si considera la possibilità di aggiungere una pendenza casuale al modello per valutare se l'effetto dei carboidrati vari da un soggetto all'altro. Per la selezione del parametro di penalizzazione λ, si utilizza il terzo metodo tra quelli proposti e si definisce una griglia di 100 valori equispaziati. Il primo valore della griglia deve essere sufficientemente elevato da spingere tutti i coefficienti a zero ($\lambda_{max} = 22000$). I coefficienti relativi alla quantità di carboidrati e alle sue interazioni non vengono penalizzati. L'andamento del BIC

rispetto al valore di λ è risultato instabile, con la presenza di vari minimi locali. La procedura seleziona $\lambda = 22000$, di conseguenza nessuna variabile è inclusa nel modello oltre a quelle non penalizzate. Rispetto al modello senza pendenza casuale, emerge un aumento del BIC e dell'RMSE, risultato pari a 59.01. Dato che la penalizzazione è molto severa, la variabilità della risposta è spiegata principalmente dagli effetti casuali, conducendo a stime elevate delle varianze casuali, sia per l'intercetta ($\sigma^2_{subject}$=629.69), che per la sensibilità ai carboidrati (σ^2_{carbs}=152.25). L'effetto legato all'intercetta e quello relativo alla pendenza casuale sono correlati positivamente ($\rho = 0.48$), suggerendo che i soggetti con livelli glicemici medi più elevati tendono a rispondere in maniera più marcata all'assunzione di carboidrati, in linea con quanto ottenuto nell'analogo modello in assenza di penalizzazione. La previsione dell'effetto casuale per i carboidrati è mostrata in Figura 5.12. Per ogni soggetto si osserva lo scostamento rispetto alla media globale dell'effetto fisso relativo ai carboidrati. Si nota una notevole variabilità tra individui, con valori che vanno da circa –21 a +24, tendenzialmente più marcati rispetto a quelli ottenuti nel modello senza termine di penalità. Per la maggior parte dei soggetti, si osservano risultati coerenti con quelli emersi dal modello con pendenza casuale in assenza di penalizzazione, ad esempio per i soggetti 9 e 36 è confermato un marcato scostamento negativo, mentre per i soggetti 2 e 33 positivo. Vi sono però dei soggetti per cui l'effetto casuale varia sostanzialmente tra i due modelli; ad esempio, il soggetto 47, nel modello non penalizzato non mostrava uno scostamento evidente dalla media globale per l'effetto dei carboidrati, mentre nel modello penalizzato risulta associato ad uno dei maggiori scostamenti negativi. Una possibile spiegazione di questo comportamento è che, essendo la penalizzazione molto severa, gli effetti casuali assorbono una parte maggiore della variabilità della risposta, comportando stime più estreme delle pendenze casuali rispetto al modello non penalizzato.

La Figura 5.13 mostra che l'aggiunta della pendenza casuale per i carboidrati assunti ha portato a una distribuzione dei residui più centrata attorno allo zero per la maggior parte dei soggetti, con una riduzione della variabilità intra-soggetto rispetto al modello con sola intercetta casuale. Questo risultato suggerisce che la relazione tra assunzione di carboidrati e risposta glicemica vari significativamente tra soggetti, e che l'inclusione di una pendenza casuale migliori la capacità del modello di adattarsi ai comportamenti. Si nota però che per alcuni soggetti le code risultano più lunghe (es. soggetti 30, 36, 39). Inoltre, la variabilità intra-soggetto non è uniforme: alcuni soggetti hanno distribuzioni più ampie di altri, suggerendo eterogeneità non completamente spiegata dal modello.
Sebbene la selezione automatica di λ individui come ottimale un valore pari a 22000,

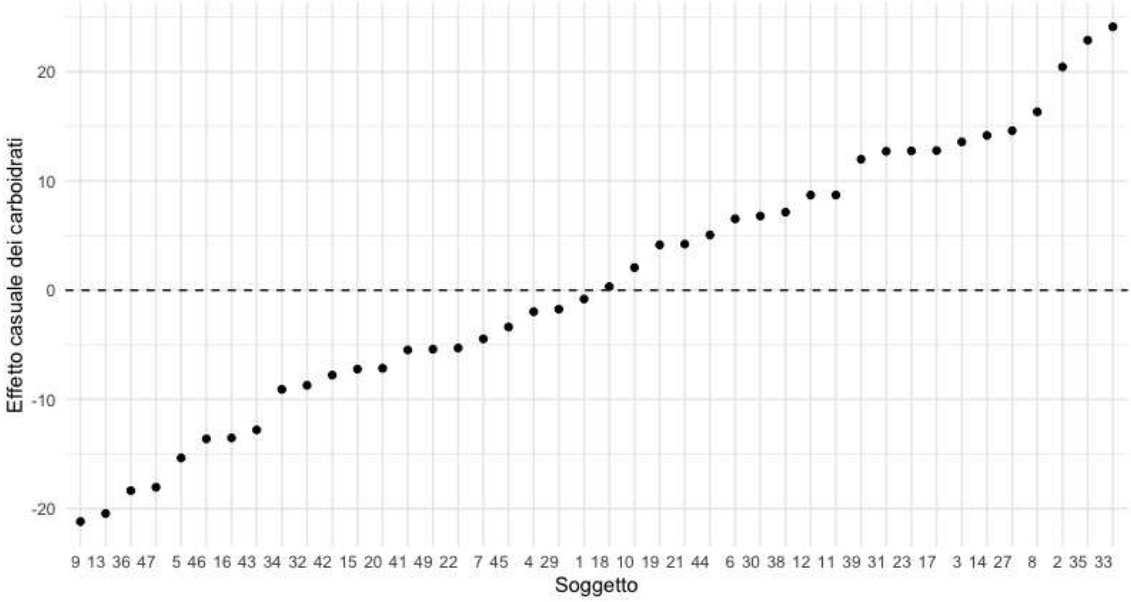

FIGURA 5.12: Previsione dello scostamento dell'effetto dei carboidrati dal valore medio, per ciascun soggetto.

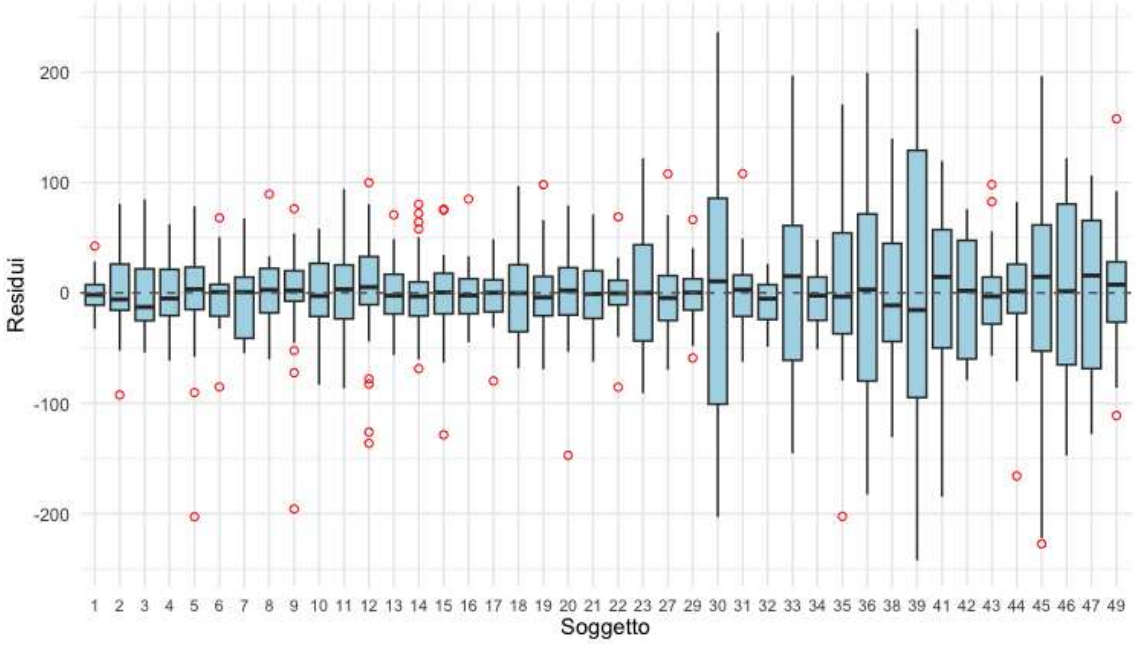

FIGURA 5.13: Distribuzione dei residui del modello lineare con intercetta e pendenza casuale con penalizzazione Group Lasso per soggetto.

determinando l'esclusione di tutte le variabili penalizzate, si è condotta un'analisi esplorativa fissando $\lambda = 5000$. L'obiettivo è comprendere se, riducendo la penalità, le variabili selezionate siano in linea con quelle emerse come rilevanti nei modelli adattati in precedenza. Tale analisi, pur non minimizzando il BIC, fornisce indicazioni utili sull'ordine di importanza relativa delle variabili secondo la penalizzazione Lasso. Il confronto tra i due livelli di penalizzazione mostra che, nel passaggio da $\lambda = 22000$ a $\lambda = 5000$, l'RMSE

diminuisce da 59 a 52, evidenziando un miglioramento della performance predittiva a fronte di un aumento della complessità del modello. Mentre $\lambda = 22000$ seleziona esclusivamente le variabili non penalizzate, $\lambda = 5000$ consente l'inclusione di ulteriori covariate nutrizionali, cliniche e microbiologiche. In particolare, il modello identifica i carboidrati, il livello glicemico antecedente il pasto, e i parametri legati alla frequenza cardiaca come principali predittori della risposta glicemica postprandiale. Inoltre, è emersa l'importanza di considerare le interazioni tra i carboidrati e il tempo. I risultati sono riportati in forma più estesa in Appendice (Tabella A.11). Essendo la penalizzazione meno severa, gli effetti fissi spiegano una porzione maggiore della variabilità della risposta, riducendo la quota di varianza attribuibile all'eterogeneità tra soggetti. Pertanto, le stime della varianza casuale risultano inferiori ($\sigma^2_{subject}$=1.78, σ^2_{carbs}=0.47), indicando una limitata eterogeneità tra individui, sia nei livelli glicemici medi, che nella risposta ai carboidrati, con una correlazione dimezzata rispetto al modello precedente ($\rho = 0.22$), confermando ad ogni modo che soggetti con glicemie basali più elevate tendono a presentare risposte ai carboidrati più marcate. Si ritiene che la riduzione delle varianze casuali sia indicativo del fatto che i dati non contengono sufficiente informazione per stimare separatamente intercetta e pendenza. Infatti si dispone di poche osservazioni, specialmente per alcuni soggetti, e le covariate presentano spesso una scarsa variabilità intra-soggetto. Di conseguenza, nel momento in cui vengono selezionate variabili costanti per soggetto, come le variabili cliniche e i punteggi del microbioma intestinale, il vantaggio derivante dalla modellazione gerarchica si riduce. Questa analisi esplorativa evidenzia come la riduzione della penalizzazione conduca a identificare un sottoinsieme più ampio e complesso di variabili esplicative, migliorando l'accuratezza predittiva, ma riducendo la varianza degli effetti casuali e la capacità di catturare la struttura gerarchica dei dati.

Infine, è stato effettuato un confronto con i risultati ottenuti applicando l'algoritmo SCAD implementato in questo progetto di tesi, considerando il caso con intercetta e pendenza casuale per i carboidrati. Sono state analizzate due configurazioni per la matrice di varianza e covarianza degli effetti casuali $\mathbf{\Psi}$: una struttura multiplo dell'identità e, successivamente, una matrice piena definita positiva. Nel primo caso, in corrispondenza del valore ottimale del parametro di regolazione (λ=1.71), il numero di variabili selezionate, oltre all'intercetta, è pari a 45, con un RMSE di 47.51. La distribuzione dei residui sui dati di stima, analizzata separatamente per ciascun soggetto, mostra miglioramenti specialmente per alcuni individui, come il soggetto 39 (Figura A.1 in Appendice). I risultati numerici sono riportati in Appendice, in Tabella A.11, e si dimostrano coerenti con quelli ottenuti nel modello penalizzato con sola intercetta casuale in termini di variabili selezionate, segno e magnitudine dei coefficienti. Rispetto

all'algoritmo `glmmLasso`, l'approccio proposto mostra una maggiore stabilità all'aggiunta di una pendenza casuale, mantenendo una buona coerenza con le selezioni ottenute nei casi precedenti. Tuttavia, comporta tempi di esecuzione più lunghi e richiede spesso un numero elevato di iterazioni per raggiungere la convergenza.

Assumere una struttura piena per $\mathbf{\Psi}$ conduce a risultati molto simili in termini di selezione e stima dei coefficienti: vengono selezionate 32 variabili oltre all'intercetta e l'RMSE si riduce a 43.59. I risultati numerici sono riportati in Appendice, in Tabella A.11. Le stime delle varianze casuali risultano elevate, sia per l'intercetta ($\sigma^2_{subject} = 228.98$), sia per la sensibilità ai carboidrati ($\sigma^2_{carbs} = 155.65$). L'effetto legato ai carboidrati è in linea con quello restituito dall'algoritmo `glmmLasso` in corrispondenza del valore ottimale del parametro di regolazione. La correlazione tra i due effetti casuali è prossima a 1, condizione che può indicare che l'informazione disponibile non è sufficiente per stimare in modo affidabile un numero elevato di parametri di varianza e covarianza. In tali situazioni, è opportuno adottare strutture più parsimoniose per $\mathbf{\Psi}$, soprattutto se l'obiettivo principale è la selezione delle variabili maggiormente informative.

5.10 Conclusione

In questo capitolo sono stati stimati diversi modelli per valutare l'effetto dei macronutrienti, delle caratteristiche individuali e del microbioma intestinale sulla risposta glicemica postprandiale, misurata tramite l'indicatore netAUC. Gli obiettivi principali consistevano nel valutare quali variabili avessero un effetto significativo sulla risposta, esplorare la selezione delle variabili rilevanti attraverso l'inclusione di un termine di penalità e analizzare gli eventuali vantaggi derivanti dal considerare la struttura gerarchica dei dati per spiegare l'eterogeneità inter-individuale. I modelli adattati possono essere suddivisi in quattro categorie: modelli lineari, modelli lineari a effetti misti, modelli lineari penalizzati e modelli lineari a effetti misti penalizzati. Il modello lineare classico ha evidenziato un'associazione significativa tra i macronutrienti e la risposta glicemica, oltre a registrare un effetto significativo delle informazioni cliniche e di alcuni indicatori ottenuti a partire dalle misurazioni del *fitbit*. Tuttavia, l'analisi dei residui ha mostrato eterogeneità tra soggetti non spiegata, suggerendo la necessità di includere effetti casuali. L'introduzione di un'intercetta casuale per soggetto ha migliorato la simmetria dei residui e ha permesso di spiegare circa il 33% della varianza totale. L'inclusione di una pendenza casuale per i carboidrati ha ulteriormente evidenziato un'ampia eterogeneità inter-individuale nella risposta glicemica ai carboidrati, con una correlazione elevata tra

l'intercetta e la pendenza casuali ($\rho = 0.683$), suggerendo che soggetti con glicemie basali più alte tendono a essere più sensibili all'assunzione di carboidrati. La procedura di selezione stepwise delle variabili ha permesso di ricavare un sottoinsieme di covariate rilevanti, conducendo a un modello con una performance predittiva simile al modello lineare completo e confermando il contributo di nutrienti, orario del pasto, caratteristiche antropometriche, parametri clinici e microbioma intestinale. Il passaggio ai modelli penalizzati ha permesso di gestire l'alta dimensionalità e la collinearità tra predittori, selezionando modelli più parsimoniosi ma predittivamente simili. Infine, i modelli a effetti misti penalizzati hanno permesso di combinare la selezione delle variabili con la struttura gerarchica dei dati. L'aggiunta di effetti casuali ha migliorato la centratura dei residui per soggetto e ha permesso di cogliere in maniera efficace la variabilità inter-individuale. L'aggiunta di una pendenza casuale per i carboidrati ha confermato che i soggetti rispondono in maniera diversa all'assunzione di tali nutrienti. Tuttavia, la scelta di λ ha influito fortemente sulla selezione nel caso della penalità Group Lasso: il valore ottimale del parametro di regolazione ($\lambda = 22000$) ha portato ad escludere tutte le variabili penalizzate, mentre per valori inferiori ($\lambda = 5000$), sono state incluse variabili nutrizionali, cliniche e microbiologiche, migliorando la performance predittiva ma aumentando la complessità del modello. L'utilizzo di una penalità non convessa, in particolare la penalità SCAD, ha portato a risultati più stabili rispetto al Group Lasso in termini di selezione e stima degli effetti fissi. È inoltre emerso che, assumere una struttura piena per la matrice di varianza e covarianza degli effetti casuali, può condurre a correlazioni elevate tra effetti casuali e a stime meno affidabili; di conseguenza, in presenza di campioni poco numerosi e sbilanciati, risulta preferibile adottare strutture più semplici per $\mathbf{\Psi}$.

In sintesi, l'analisi ha dimostrato che la risposta glicemica postprandiale è influenzata da un insieme complesso di variabili nutrizionali, antropometriche e cliniche. Da un lato, i modelli a effetti misti permettono di considerare l'eterogeneità inter-individuale, dall'altro introdurre un termine di penalità consente di gestire l'alta dimensionalità, migliorare la stabilità delle stime e selezionare sottoinsiemi informativi di predittori.

Dall'analisi emergono diversi limiti di cui è importante tenere conto. Innanzittuto, in seguito alla fase di preprocessing, il numero di osservazioni si è ridotto notevolmente, sia in termini di soggetti inclusi nell'analisi, che di pasti associati a ciascun soggetto. Incrementare la numerosità campionaria permetterebbe di mitigare i problemi di instabilità numerica nei modelli ad alta dimensionalità, garantendo una maggiore potenza statistica. Una seconda problematica emersa da tutti i modelli impiegati riguarda la

scarsa performance predittiva. In parte, potrebbe essere attribuita al fatto che i modelli stimati assumono una relazione lineare tra predittori e risposta, ignorando la presenza di processi fisiologici e metabolici complessi e potenzialmente non lineari. Pertanto, per migliorare l'accuratezza predittiva, si potrebbero adattare algoritmi di machine learning, come Random Forests o Gradient Boosting, in grado di catturare relazioni non lineari e interazioni complesse tra variabili. Tuttavia, l'impiego di tali modelli avverrebbe a costo di una minore interpretabilità. Una strategia alternativa potrebbe essere quella di adottare un approccio di tipo Bayesiano per tenere conto di informazioni a priori nella stima dei parametri. In questo modo, si potrebbero ottenere le distribuzioni a posteriori per gli effetti fissi e casuali, consentendo di quantificare con maggiore precisione l'incertezza associata alla stima dei parametri. Infine, si osserva che gli approcci utilizzati considerano ogni pasto come unità di primo livello, modellando netAUC come variabile risposta. Tuttavia, si potrebbe valutare l'utilizzo di modelli non lineari a effetti misti (NLME) per modellare l'intera curva glicemica nel tempo per ciascun soggetto. Questa dinamica può essere descritta tramite funzioni non lineari e permette di includere parametri fisiologici interpretabili, come ad esempio la sensibilità, cioè quanto aumenta la glicemia dopo il pasto e la rapidità di assorbimento, cioè quanto velocemente la glicemia torna al valore basale. I NLME consentono di considerare effetti casuali per questi parametri, permettendo di ottenere, per ciascun soggetto, una previsione dello scostamento dalla media di tali caratteristiche nella popolazione. Una strategia alternativa consisterebbe nel modellare l'andamento temporale del livello glicemico tramite modelli di serie storiche multivariate. Questo metodo permetterebbe di inserire direttamente come predittori le informazioni rilevate dal *fitbit*, senza dover ricorrere a misure di sintesi. Tuttavia, questo approccio richiede un numero sufficiente di osservazioni per ciascun soggetto e una certa regolarità temporale nella rilevazione dei dati, condizioni che non sono soddisfatte nei dati analizzati. Infatti, vi sono individui per cui si dispone di meno di 15 misurazioni valide una volta terminata la procedura di preprocessing e i dati sono spesso caratterizzati da salti temporali a causa di periodi di inattività dei dispositivi di rilevazione della glicemia o del battito cardiaco.

Conclusione

In questo lavoro di tesi è stato affrontato il problema della selezione di variabili nei modelli lineari a effetti misti in situazioni ad elevata dimensionalità. L'interesse principale ha riguardato la selezione e la stima dei coefficienti associati agli effetti fissi tramite metodi di penalizzazione, assumendo come prespecificati gli effetti casuali e mantenendo ridotto il numero di parametri di varianza e covarianza. Sono stati presentati i principali approcci proposti in letteratura, con un focus sulle penalità Lasso e SCAD. Sono stati descritti gli algoritmi di stima e le proprietà degli stimatori nel caso $p > N$. Sotto opportune condizioni, lo stimatore Lasso è consistente e soddisfa la disuguaglianza oracolo, mentre lo stimatore SCAD gode della proprietà di oracolo e rappresenta un minimo locale ben definito della funzione obiettivo con probabilità tendente a 1, garantendo la validità e stabilità della soluzione al crescere della numerosità campionaria

Un contributo originale del lavoro è stato lo sviluppo di un'implementazione estesa dell'algoritmo SCAD, che consente di assumere per la matrice di varianza e covarianza degli effetti casuali $\mathbf{\Psi}$ una forma generica definita positiva, superando le limitazioni delle versioni esistenti, basate su strutture diagonali o multiple dell'identità. Tramite uno studio di simulazione, è stata valutata la capacità di tale estensione di individuare la reale struttura di correlazione tra effetti casuali, selezionare correttamente gli effetti fissi e stimarne in modo accurato i coefficienti. I risultati hanno mostrato che l'adozione di una struttura più flessibile per $\mathbf{\Psi}$ permette di migliorare la stima della matrice di varianza e covarianza, anche in scenari ad elevata dimensionalità, garantendo una buona selezione delle variabili e stime precise dei coefficienti. Tuttavia, ciò comporta un lieve aumento dell'errore di previsione, suggerendo che, quando l'obiettivo principale è la stima degli effetti fissi, può risultare più conveniente adottare strutture di $\mathbf{\Psi}$ più semplici.

Successivamente, i metodi discussi sono stati applicati a dati reali provenienti da un esperimento sul monitoraggio continuo della glicemia in risposta ai pasti. A partire dalle curve glicemiche, è stato calcolato l'indicatore netAUC, la cui distribuzione può essere considerata approssimativamente normale. Sono stati stimati quattro modelli principali: modello lineare classico, modello lineare a effetti misti, modello lineare

penalizzato e modello lineare a effetti misti penalizzato. Dall'analisi è emerso che la risposta glicemica postprandiale è influenzata da una combinazione complessa di variabili nutrizionali, antropometriche e cliniche. I modelli a effetti misti hanno permesso di considerare l'eterogeneità inter-individuale, cogliendo le differenze nella risposta media. Inoltre, l'utilizzo di tecniche di penalizzazione ha permesso di gestire l'alta dimensionalità, migliorare la stabilità delle stime e selezionare sottoinsiemi informativi di predittori. La penalità SCAD ha permesso di ottenere risultati coerenti anche quando il modello è stato reso più complesso, includendo una pendenza casuale per tenere conto della diversa sensibilità dei soggetti ai carboidrati. Al contrario, l'adozione della penalità Lasso ha determinato una scarsa stabilità nella selezione delle variabili, risultando eccessivamente severa in questo scenario e portando all'esclusione di tutte le variabili penalizzate.

In conclusione, il lavoro ha messo in evidenza come l'integrazione di metodi di penalizzazione nei modelli lineari a effetti misti costituisca una strategia efficace per l'analisi di dati complessi e ad alta dimensionalità, in grado di selezionare le variabili più rilevanti e descrivere le differenze individuali.

Appendice A

A.1 Tabelle

Tabella A.1: Risultati della procedura di stima e selezione degli effetti fissi quando le variabili esplicative sono correlate ($\rho_x = 0.5$), per valori crescenti di p e diverse assunzioni per $\mathbf{\Psi}$.

		$\lvert S(\hat{\beta})\rvert$	TP	PE	β_1	β_2	β_3	β_4	β_5	β_N
					$p = 10$					
	Media	5.09	5.00	0.187	0.964	-1.879	3.492	2.996	-0.494	0.000
Ψ_I	SD	0.32	0.00	0.025	0.210	0.162	0.054	0.061	0.053	0.006
	MSE				0.045	0.040	0.003	0.004	0.003	0.000
	Media	5.09	5.00	0.187	0.964	-1.881	3.490	2.996	-0.494	0.000
Ψ_D	SD	0.32	0.00	0.025	0.210	0.162	0.054	0.061	0.053	0.006
	MSE				0.045	0.040	0.003	0.004	0.003	0.000
	Media	5.06	5.00	0.190	0.972	-1.890	3.489	2.996	-0.493	0.000
Ψ_C	SD	0.28	0.00	0.031	0.214	0.162	0.057	0.061	0.053	0.005
	MSE				0.046	0.038	0.003	0.004	0.003	0.000
					$p = 100$					
	Media	5.17	5.00	0.185	0.966	-1.771	3.500	2.985	-0.487	0.000
Ψ_I	SD	0.24	0.00	0.023	0.201	0.216	0.055	0.053	0.059	0.000
	MSE				0.041	0.099	0.003	0.003	0.004	0.000
	Media	5.17	5.00	0.186	0.966	-1.771	3.498	2.985	-0.487	0.000
Ψ_D	SD	0.24	0.00	0.024	0.202	0.216	0.055	0.053	0.059	0.000
	MSE				0.041	0.099	0.003	0.003	0.004	0.000
	Media	5.24	5.00	0.205	0.976	-1.772	3.493	2.981	-0.485	0.000
Ψ_C	SD	0.75	0.00	0.079	0.211	0.217	0.056	0.056	0.060	0.000
	MSE				0.045	0.099	0.003	0.003	0.004	0.000
					$p = 300$					
	Media	5.30	5.00	0.187	0.952	-1.663	3.491	2.986	-0.477	0.000
Ψ_I	SD	0.67	0.00	0.024	0.225	0.228	0.057	0.067	0.067	0.000
	MSE				0.053	0.165	0.003	0.005	0.005	0.000
	Media	5.29	5.00	0.188	0.953	-1.663	3.488	2.986	-0.477	0.000
Ψ_D	SD	0.62	0.00	0.024	0.225	0.228	0.057	0.067	0.069	0.000
	MSE				0.052	0.165	0.003	0.005	0.005	0.000
	Media	5.31	5.00	0.195	0.974	-1.663	3.488	2.984	-0.475	0.000
Ψ_C	SD	0.60	0.00	0.055	0.239	0.229	0.058	0.067	0.069	0.000
	MSE				0.057	0.165	0.004	0.005	0.005	0.000

Tabella A.2: Stima dei parametri di varianza quando le variabili esplicative sono correlate ($\rho_x = 0.5$) per valori crescenti di p e diverse assunzioni per $\mathbf{\Psi}$.

		σ^2_ε	θ^2_1	θ^2_2	θ_{12}
		$p = 10$			
Ψ_I	Media	0.241	0.616	0.616	0.000
	SD	0.031	0.168	0.168	0.000
	MSE	0.001	0.062	0.055	0.032
Ψ_D	Media	0.241	0.775	0.451	0.000
	SD	0.031	0.268	0.152	0.000
	MSE	0.001	0.072	0.023	0.032
Ψ_C	Media	0.243	0.770	0.454	0.184
	SD	0.036	0.269	0.167	0.146
	MSE	0.005	0.072	0.027	0.021
		$p = 100$			
Ψ_I	Media	0.239	0.619	0.619	0.000
	SD	0.030	0.161	0.161	0.000
	MSE	0.001	0.058	0.054	0.032
Ψ_D	Media	0.240	0.755	0.479	0.000
	SD	0.030	0.238	0.157	0.000
	MSE	0.001	0.058	0.025	0.032
Ψ_C	Media	0.259	0.755	0.460	0.156
	SD	0.083	0.251	0.171	0.162
	MSE	0.007	0.064	0.029	0.027
		$p = 300$			
Ψ_I	Media	0.242	0.680	0.680	0.000
	SD	0.030	0.185	0.185	0.000
	MSE	0.001	0.048	0.087	0.032
Ψ_D	Media	0.243	0.769	0.588	0.000
	SD	0.030	0.240	0.214	0.000
	MSE	0.001	0.058	0.064	0.032
Ψ_C	Media	0.249	0.764	0.589	0.169
	SD	0.059	0.253	0.214	0.158
	MSE	0.003	0.065	0.065	0.025

Tabella A.3: Risultati della procedura di stima e selezione degli effetti fissi applicando l'algoritmo `glmmLasso` e fissando p=300.

	$\lvert S(\hat{\beta})\rvert$	TP	PE	β_1	β_2	β_3	β_4	β_5	β_N
Media	13.44	5.00	0.227	0.987	-2.028	3.471	2.975	-0.474	0.000
SD	8.69	0.00	0.039	0.174	0.177	0.062	0.064	0.055	0.001
MSE				0.030	0.032	0.005	0.005	0.004	0.000

Tabella A.4: Stima dei parametri di varianza applicando l'algoritmo `glmmLasso` e fissando p=300.

	σ^2_ε	θ^2_1	θ^2_2	θ_{12}
Media	0.248	0.553	0.276	0.124
SD	0.034	0.219	0.156	0.132
MSE	0.001	0.108	0.054	0.020

	Stima	Std. Error	p-value
(Intercetta)	5.760	21.444	0.788
Tipo pasto:cena	-16.594	15.524	0.285
Tipo pasto:pranzo	-18.103	12.097	0.135
Tipo pasto:snack	-20.323	14.973	0.175
Genere (M)	69.016	11.320	0.000
Identità:ispanici/latini	33.585	20.986	0.110
Identità:bianchi	7.820	14.621	0.593
Calorie	11.584	5.054	0.022
Carboidrati	18.559	3.319	0.000
Proteine	-13.066	2.403	0.000
Grassi	-7.327	3.122	0.019
Fibre	-1.892	1.599	0.237
Calorie assunte nelle 2h prima	2.140	2.041	0.295
Calorie assunte nelle 3h prima	0.091	4.771	0.985
Calorie assunte nelle 6h prima	-6.795	5.975	0.256
Calorie assunte nelle 12h prima	-2.680	6.421	0.676
Carboidrati assunti nelle 3h prima	-1.109	4.615	0.810
Carboidrati assunti nelle 6h prima	5.978	6.009	0.320
Carboidrati assunti nelle 12h prima	-5.933	5.767	0.304
Fibre assunte nelle 12h prima	-0.168	1.704	0.921
Fibre assunte nelle 24h prima	1.165	1.984	0.557
Media HR nei 30min successivi	8.763	8.018	0.275
Delta HR	6.968	5.387	0.196
Slope HR	6.136	2.999	0.041
AUC HR nelle 2h successive	-5.417	6.225	0.384
Numero spikes	-0.956	3.610	0.791
Numero blocchi contigui	-2.106	2.274	0.355
Media HR sopra soglia	1.289	7.162	0.857
Max HR sopra soglia	1.743	2.806	0.535
Numero salti grandi	-4.003	1.942	0.040
Tempo al primo spike	-0.505	1.539	0.743
Calorie bruciate nelle 2h successive	-18.570	2.849	0.000
Età	-10.690	3.875	0.006
BMI	20.794	6.706	0.002
Altezza	-2.191	4.570	0.632
HbA1c	36.056	5.961	0.000
Insulina	-15.884	4.810	0.001
Colesterolo	2.302	4.620	0.618
Colesterolo HDL	0.206	3.820	0.957
Colesterolo LDL	-12.767	7.545	0.091
Giorno studio	-0.332	1.473	0.822
Glucosio 1h prima	-38.851	2.608	0.000
Glucosio 2h prima	-1.501	2.788	0.591
Glucosio 4h prima	-2.015	1.985	0.310
sin_ora_pasto1	-0.410	5.913	0.945
cos_ora_pasto1	0.141	4.202	0.973
sin_ora_pasto2	-4.230	3.337	0.205
cos_ora_pasto2	-1.866	3.200	0.560
sin_ora_pasto3	-1.423	2.259	0.529
cos_ora_pasto3	3.483	2.461	0.157
Gut.Lining.Health.L	76.682	18.631	0.000
Gut.Lining.Health.Q	18.325	6.650	0.006
LPS.Biosynthesis.Pathways.L	4.318	14.256	0.762

LPS.Biosynthesis.Pathways.Q	3.322	5.810	0.568
Biofilm..Chemotaxis..and.Virulence.Pathways.L	-124.148	31.027	0.000
Biofilm..Chemotaxis..and.Virulence.Pathways.Q	-37.332	7.705	0.000
TMA.Production.Pathways.L	43.073	9.202	0.000
TMA.Production.Pathways.Q	24.916	10.576	0.019
Ammonia.Production.Pathways.L	43.032	20.056	0.032
Ammonia.Production.Pathways.Q	-3.012	8.566	0.725
Metabolic.Fitness.L	-34.112	6.921	0.000
Metabolic.Fitness.Q	28.957	7.439	0.000
Active.Microbial.Diversity.L	2.582	10.287	0.802
Active.Microbial.Diversity.Q	42.065	14.455	0.004
Butyrate.Production.Pathways.L	-129.817	31.003	0.000
Butyrate.Production.Pathways.Q	-32.526	13.933	0.020
Flagellar.Assembly.Pathways.L	47.881	9.726	0.000
Flagellar.Assembly.Pathways.Q	6.862	5.601	0.221
Putrescine.Production.Pathways.L	56.448	8.419	0.000
Putrescine.Production.Pathways.Q	-30.978	15.733	0.049
Uric.Acid.Production.Pathways.L	-13.169	18.630	0.480
Uric.Acid.Production.Pathways.Q	-19.306	7.805	0.014
Bile.Acid.Metabolism.Pathways.L	42.575	22.764	0.062
Bile.Acid.Metabolism.Pathways.Q	-5.191	5.070	0.306
Carboidrati:sin_ora_pasto1	12.027	2.564	0.000
Carboidrati:HbA1c	1.596	1.844	0.387
Carboidrati:glucosio_1h_prima	3.462	1.786	0.053
Carboidrati:cos_ora_pasto1	-2.627	1.920	0.172

Tabella A.5: Stime, errori standard e p-value dei coefficienti per il modello lineare.

	Stima	Std. Error	p-value
(Intercetta)	5.093	59.909	0.932
Tipo pasto:cena	-18.672	15.305	0.223
Tipo pasto:pranzo	-18.479	11.858	0.119
Tipo pasto:snack	-21.726	14.756	0.141
Genere (M)	72.007	34.898	0.085
Identità:ispanici/latini	33.002	62.043	0.614
Identità:bianchi	14.216	46.080	0.768
Calorie	12.686	4.941	0.010
Carboidrati	18.201	3.238	0.000
Proteine	-14.256	2.354	0.000
Grassi	-7.682	3.055	0.012
Fibre	-1.607	1.566	0.305
Calorie assunte nelle 2h prima	1.700	1.991	0.393
Calorie assunte nelle 3h prima	0.578	4.665	0.901
Calorie assunte nelle 6h prima	-7.426	5.819	0.202
Calorie assunte nelle 12h prima	-5.032	6.288	0.424
Carboidrati assunti nelle 3h prima	-1.381	4.510	0.759
Carboidrati assunti nelle 6h prima	7.302	5.859	0.213
Carboidrati assunti nelle 12h prima	-3.784	5.647	0.503
Fibre assunte nelle 12h prima	-0.402	1.660	0.809
Fibre assunte nelle 24h prima	1.981	1.945	0.309
Media HR nei 30min successivi	6.384	7.943	0.422
Delta HR	6.234	5.311	0.241
Slope HR	4.799	2.927	0.101
AUC HR nelle 2h successive	-5.278	6.141	0.390

Numero spikes	-2.293	3.557	0.519
Numero blocchi contigui	-1.963	2.235	0.380
Media HR sopra soglia	2.925	7.049	0.678
Max HR sopra soglia	0.676	2.744	0.806
Numero salti grandi	-3.260	1.952	0.095
Tempo al primo spike	-0.246	1.501	0.870
Calorie bruciate nelle 2h successive	-16.347	2.828	0.000
Età	-9.918	13.008	0.475
BMI	18.299	19.803	0.391
Altezza	-4.899	13.056	0.720
HbA1c	39.949	18.651	0.076
Insulina	-14.758	14.197	0.339
Colesterolo	4.166	14.422	0.782
Colesterolo HDL	-2.634	13.063	0.847
Colesterolo LDL	-13.783	22.812	0.568
Giorno studio	-0.196	1.436	0.891
Glucosio 1h prima	-42.035	2.585	0.000
Glucosio 2h prima	-3.259	2.742	0.235
Glucosio 4h prima	-4.866	1.990	0.015
sin_ora_pasto1	-0.010	5.810	0.999
cos_ora_pasto1	0.825	4.124	0.841
sin_ora_pasto2	-3.737	3.277	0.254
cos_ora_pasto2	-1.670	3.133	0.594
sin_ora_pasto3	-1.113	2.218	0.616
cos_ora_pasto3	2.731	2.417	0.259
Gut.Lining.Health.L	78.115	55.567	0.209
Gut.Lining.Health.Q	20.237	21.419	0.381
LPS.Biosynthesis.Pathways.L	3.732	45.649	0.937
LPS.Biosynthesis.Pathways.Q	2.838	19.434	0.889
Biofilm..Chemotaxis..and.Virulence.Pathways.L	-130.616	94.555	0.216
Biofilm..Chemotaxis..and.Virulence.Pathways.Q	-41.938	25.508	0.151
TMA.Production.Pathways.L	37.330	26.294	0.206
TMA.Production.Pathways.Q	23.415	33.425	0.510
Ammonia.Production.Pathways.L	41.834	61.871	0.524
Ammonia.Production.Pathways.Q	-5.013	26.331	0.855
Metabolic.Fitness.L	-33.396	22.919	0.195
Metabolic.Fitness.Q	25.356	22.813	0.309
Active.Microbial.Diversity.L	2.823	33.321	0.935
Active.Microbial.Diversity.Q	41.264	43.214	0.377
Butyrate.Production.Pathways.L	-125.641	91.754	0.220
Butyrate.Production.Pathways.Q	-26.681	43.060	0.558
Flagellar.Assembly.Pathways.L	49.133	31.457	0.169
Flagellar.Assembly.Pathways.Q	8.323	18.321	0.666
Putrescine.Production.Pathways.L	56.428	26.246	0.075
Putrescine.Production.Pathways.Q	-28.679	49.153	0.581
Uric.Acid.Production.Pathways.L	-10.043	58.111	0.868
Uric.Acid.Production.Pathways.Q	-17.137	23.973	0.502
Bile.Acid.Metabolism.Pathways.L	37.007	69.233	0.612
Bile.Acid.Metabolism.Pathways.Q	-6.049	17.298	0.739
Carboidrati:sin_ora_pasto1	12.380	2.502	0.000
Carboidrati:HbA1c	2.092	1.802	0.246
Carboidrati:glucosio_1h_prima	3.446	1.740	0.048
Carboidrati:cos_ora_pasto1	-1.936	1.886	0.305

Tabella A.6: Stime, errori standard e p-value dei coefficienti degli effetti fissi per il modello lineare a effetti misti con intercetta casuale.

	Stima	Std. Error	p-value
(Intercetta)	-9.908	56.487	0.861
Tipo pasto:cena	-18.207	15.226	0.232
Tipo pasto:pranzo	-18.895	11.710	0.107
Tipo pasto:snack	-21.208	14.666	0.148
Genere (M)	58.879	32.892	0.124
Identità:ispanici/latini	58.024	58.240	0.358
Identità:bianchi	28.995	42.849	0.524
Calorie	9.177	5.413	0.090
Carboidrati	22.440	4.049	0.000
Proteine	-13.592	2.390	0.000
Grassi	-6.753	3.120	0.031
Fibre	-1.769	1.554	0.255
Calorie assunte nelle 2h prima	1.220	1.966	0.535
Calorie assunte nelle 3h prima	0.391	4.607	0.932
Calorie assunte nelle 6h prima	-7.135	5.763	0.216
Calorie assunte nelle 12h prima	-4.357	6.208	0.483
Carboidrati assunti nelle 3h prima	-1.011	4.457	0.821
Carboidrati assunti nelle 6h prima	6.736	5.804	0.246
Carboidrati assunti nelle 12h prima	-4.406	5.565	0.429
Fibre assunte nelle 12h prima	-0.528	1.636	0.747
Fibre assunte nelle 24h prima	1.925	1.918	0.316
Media HR nei 30min successivi	8.007	7.885	0.310
Delta HR	4.858	5.283	0.358
Slope HR	4.849	2.893	0.094
AUC HR nelle 2h successive	-3.911	6.099	0.521
Numero spikes	-2.775	3.514	0.430
Numero blocchi contigui	-1.648	2.208	0.456
Media HR sopra soglia	3.044	6.987	0.663
Max HR sopra soglia	0.691	2.720	0.799
Numero salti grandi	-3.082	1.956	0.115
Tempo al primo spike	-0.279	1.480	0.850
Calorie bruciate nelle 2h successive	-17.160	2.804	0.000
Età	-3.735	12.146	0.769
BMI	16.257	18.623	0.416
Altezza	2.650	12.222	0.836
HbA1c	24.118	17.373	0.214
Insulina	-11.192	13.376	0.435
Colesterolo	6.117	13.585	0.668
Colesterolo HDL	-6.096	12.155	0.634
Colesterolo LDL	-18.816	21.296	0.411
Giorno studio	0.083	1.412	0.953
Glucosio 1h prima	-42.034	2.556	0.000
Glucosio 2h prima	-3.467	2.719	0.203
Glucosio 4h prima	-4.485	1.978	0.024
sin_ora_pasto1	-0.141	5.767	0.981
cos_ora_pasto1	1.497	3.961	0.706
sin_ora_pasto2	-3.900	3.218	0.226
cos_ora_pasto2	-0.938	3.085	0.761

sin_ora_pasto3	-1.473	2.175	0.498
cos_ora_pasto3	2.797	2.397	0.244
Gut.Lining.Health.L	77.315	52.250	0.189
Gut.Lining.Health.Q	16.200	20.166	0.452
LPS.Biosynthesis.Pathways.L	-15.688	42.309	0.724
LPS.Biosynthesis.Pathways.Q	-1.564	18.225	0.934
Biofilm..Chemotaxis..and.Virulence.Pathways.L	-110.789	89.020	0.260
Biofilm..Chemotaxis..and.Virulence.Pathways.Q	-32.920	24.060	0.220
TMA.Production.Pathways.L	35.666	24.811	0.201
TMA.Production.Pathways.Q	11.349	31.095	0.728
Ammonia.Production.Pathways.L	46.793	58.219	0.452
Ammonia.Production.Pathways.Q	-10.856	24.847	0.677
Metabolic.Fitness.L	-31.146	21.354	0.195
Metabolic.Fitness.Q	21.307	21.308	0.356
Active.Microbial.Diversity.L	-5.924	31.330	0.856
Active.Microbial.Diversity.Q	41.357	40.833	0.350
Butyrate.Production.Pathways.L	-105.351	86.134	0.267
Butyrate.Production.Pathways.Q	-19.793	40.273	0.641
Flagellar.Assembly.Pathways.L	30.284	29.839	0.349
Flagellar.Assembly.Pathways.Q	9.250	17.269	0.611
Putrescine.Production.Pathways.L	64.381	24.794	0.041
Putrescine.Production.Pathways.Q	-4.180	45.678	0.930
Uric.Acid.Production.Pathways.L	-23.016	54.256	0.686
Uric.Acid.Production.Pathways.Q	-11.177	22.407	0.636
Bile.Acid.Metabolism.Pathways.L	46.610	64.751	0.499
Bile.Acid.Metabolism.Pathways.Q	-5.657	16.129	0.738
Carboidrati:sin_ora_pasto1	12.669	2.356	0.000
Carboidrati:HbA1c	6.448	2.590	0.013
Carboidrati:glucosio_1h_prima	0.409	1.931	0.832

Tabella A.7: Stime, errori standard e p-value dei coefficienti degli effetti fissi del modello lineare a effetti misti con intercetta e pendenza casuale per i carboidrati.

	Stima	Std. Error	p-value
(Intercetta)	-6.930	11.361	0.542
Genere (M)	61.707	6.854	0.000
Identità:ispanici/latini	34.096	11.140	0.002
Identità:bianchi	5.988	10.737	0.577
Calorie	9.139	4.587	0.047
Carboidrati	19.862	3.107	0.000
Proteine	-12.095	2.255	0.000
Grassi	-5.966	2.965	0.045
Calorie assunte nelle 2h prima	2.154	1.540	0.162
Calorie assunte nelle 6h prima	-3.463	2.185	0.113
Carboidrati assunti nelle 12h prima	-7.109	2.350	0.003
Media HR nei 30min successivi	12.323	2.695	0.000
Delta HR	2.545	1.519	0.094
Slope HR	5.445	2.013	0.007
Numero salti grandi	-3.923	1.525	0.010
Calorie bruciate nelle 2h successive	-19.069	2.316	0.000
Età	-8.837	2.557	0.001
BMI	19.801	3.173	0.000
HbA1c	31.997	3.465	0.000
Insulina	-15.615	3.597	0.000

Colesterolo LDL	-9.675	3.100	0.002
Glucosio 1h prima	-39.487	1.822	0.000
sin_ora_pasto1	7.569	2.677	0.005
cos_ora_pasto1	0.439	2.565	0.864
sin_ora_pasto2	-3.861	2.313	0.095
cos_ora_pasto2	-3.786	2.308	0.101
cos_ora_pasto3	4.496	2.034	0.027
Gut.Lining.Health.L	69.607	8.575	0.000
Gut.Lining.Health.Q	18.721	4.687	0.000
Biofilm..Chemotaxis..and.Virulence.Pathways.L	-113.692	11.471	0.000
Biofilm..Chemotaxis..and.Virulence.Pathways.Q	-34.440	5.325	0.000
TMA.Production.Pathways.L	40.218	5.801	0.000
TMA.Production.Pathways.Q	22.432	4.698	0.000
Ammonia.Production.Pathways.L	36.394	7.730	0.000
Ammonia.Production.Pathways.Q	-1.217	5.428	0.823
Metabolic.Fitness.L	-32.815	5.911	0.000
Metabolic.Fitness.Q	27.600	5.569	0.000
Active.Microbial.Diversity.L	1.462	8.501	0.864
Active.Microbial.Diversity.Q	38.623	5.520	0.000
Butyrate.Production.Pathways.L	-119.734	12.510	0.000
Butyrate.Production.Pathways.Q	-29.111	5.939	0.000
Flagellar.Assembly.Pathways.L	42.674	7.012	0.000
Flagellar.Assembly.Pathways.Q	3.580	3.933	0.363
Putrescine.Production.Pathways.L	58.302	7.015	0.000
Putrescine.Production.Pathways.Q	-25.454	7.410	0.001
Uric.Acid.Production.Pathways.L	-8.556	8.086	0.290
Uric.Acid.Production.Pathways.Q	-19.535	5.940	0.001
Bile.Acid.Metabolism.Pathways.L	38.175	9.905	0.000
Bile.Acid.Metabolism.Pathways.Q	-6.206	4.451	0.164
Carboidrati:sin_ora_pasto1	12.140	2.418	0.000
Carboidrati:glucosio_1h_prima	4.030	1.682	0.017
Carboidrati:cos_ora_pasto1	-2.705	1.671	0.106

Tabella A.8: Stime, errori standard e p-value dei coefficienti per il modello lineare con selezione stepwise.

	Lasso	MCP	SCAD	GLasso	GMCP	GSCAD
(Intercetta)	39.647	47.639	46.974	47.717	47.317	28.973
Tipo pasto:cena	.	.	0.366	-13.725	.	-1.942
Tipo pasto:pranzo	-4.344	.	.	-16.597	.	-2.634
Tipo pasto:snack	-4.661	.	.	-18.498	.	-4.422
Genere (M)	9.553	.	.	17.770	.	25.843
Identità:ispanici/latini	3.399	.	.	0.324	.	.
Identità:bianchi	.	.	.	-0.065	.	.
Calorie	10.460	.	.	8.119	.	.
Carboidrati	19.203	24.877	25.403	19.679	25.563	24.546
Proteine	-13.157	-10.200	-9.710	-12.093	-9.751	-9.672
Grassi	-6.288	.	.	-4.805	.	.
Fibre	-1.081	.	.	-0.528	.	-0.116
Calorie assunte nelle 2h prima	1.569	0.402	.	1.329	.	0.874
Calorie assunte nelle 3h prima	.	.	.	.	.	.
Calorie assunte nelle 6h prima	-1.770	.	.	-1.013	.	.
Calorie assunte nelle 12h prima	-8.787	-10.056	.	-7.498	.	-11.032
Carboidrati assunti nelle 3h prima	.	.	.	.	.	.

Carboidrati assunti nelle 6h prima	0.613	.	.	.	.	.
Carboidrati assunti nelle 12h prima	-0.251	.	-8.358	-0.738	-7.764	.
Fibre assunte nelle 12h prima	.	.	.	.	.	.
Fibre assunte nelle 24h prima	1.853	1.195	0.480	1.892	.	.
Media HR nei 30min successivi	5.622	.	.	5.301	.	2.400
Delta HR	4.328	0.217	.	2.579	0.728	0.970
Slope HR	3.265	.	.	1.721	.	.
AUC HR nelle 2h successive	-2.841	.	.	-0.693	.	.
Numero spikes	-2.101	.	.	-1.644	.	.
Numero blocchi contigui	-1.804	.	.	-1.261	.	-0.867
Media HR sopra soglia	.	.	.	.	.	.
Max HR sopra soglia	.	.	.	.	.	.
Numero salti grandi	-2.077	-0.997	-0.171	-2.214	.	-0.157
Tempo al primo spike	-0.023	.	.	.	.	.
Calorie bruciate nelle 2h successive	-13.529	-13.054	-12.152	-13.088	-13.749	-12.746
Età	-8.302	-14.620	-10.559	-7.622	-14.287	-5.559
BMI	.	.	.	.	.	.
Peso	9.113	14.820	.	4.293	16.008	8.858
Altezza	4.667	.	10.408	.	.	0.829
HbA1c	0.731	.	.	2.011	.	.
Insulina	.	-13.281	.	.	-10.832	.
Colesterolo	-5.898	.	-1.262	-6.574	.	-5.476
Colesterolo HDL	.	.	.	.	.	.
Colesterolo LDL	-15.517	-7.821	-13.915	-9.820	-8.237	-19.057
Puntura dito glucosio 1	-7.970	.	.	.	.	.
Puntura dito glucosio 2	36.446	57.059	47.024	27.133	58.188	32.098
Puntura dito glucosio 3	30.050	.	0.544	21.822	.	23.445
Giorno studio	-0.168	.	.	-0.134	.	.
Glucosio 1h prima	-40.900	-41.893	-40.523	-39.904	-41.514	-42.977
Glucosio 2h prima	-2.731	.	.	-2.284	.	.
Glucosio 4h prima	-3.689	-1.684	.	-2.681	-1.071	-5.081
sin_ora_pasto1	6.002	11.095	12.887	0.959	12.743	0.371
cos_ora_pasto1	.	.	.	.	.	.
sin_ora_pasto2	-3.729	-0.518	.	-3.311	.	-6.780
cos_ora_pasto2	-2.927	-7.900	-8.028	-1.360	-7.713	.
sin_ora_pasto3	-0.871	.	.	-1.072	.	.
cos_ora_pasto3	3.838	.	0.208	3.456	0.037	7.018
Gut.Lining.Health.L	10.355	.	.	5.230	.	17.866
Gut.Lining.Health.Q	0.996	.	.	5.453	.	12.548
LPS.Biosynthesis.Pathways.L	-1.543	.	.	-8.763	11.867	-15.408
LPS.Biosynthesis.Pathways.Q	15.699	16.421	20.706	14.585	22.923	10.824
Biofilm..Chemotaxis..and.Virulence.Pathways.L	13.474	.	15.572	2.106	.	.
Biofilm..Chemotaxis..and.Virulence.Pathways.Q	10.279	.	.	1.864	.	.
TMA.Production.Pathways.L	.	.	.	2.664	2.373	1.865
TMA.Production.Pathways.Q	-36.430	-36.446	-40.127	-28.205	-28.287	-40.339
Ammonia.Production.Pathways.L	-3.378	.	.	-0.059	.	.
Ammonia.Production.Pathways.Q	.	.	.	0.005	.	.
Metabolic.Fitness.L	-14.872	-29.276	-25.612	-23.586	-27.889	-26.119
Metabolic.Fitness.Q	10.345	15.729	9.274	12.698	26.409	7.000
Active.Microbial.Diversity.L	.	.	.	0.887	-5.142	.
Active.Microbial.Diversity.Q	2.345	13.055	.	1.688	10.718	.
Butyrate.Production.Pathways.L	.	.	30.113	-4.795	.	0.044
Butyrate.Production.Pathways.Q	-3.738	.	.	-5.658	.	-0.534
Flagellar.Assembly.Pathways.L	10.223	39.806	21.459	21.754	43.581	19.673
Flagellar.Assembly.Pathways.Q	10.317	17.026	23.852	10.794	16.687	16.747

Putrescine.Production.Pathways.L	2.830	.	.	.	.	.
Putrescine.Production.Pathways.Q	.	16.010	.	.	.	.
Uric.Acid.Production.Pathways.L	6.997	21.557	.	14.658	20.448	.
Uric.Acid.Production.Pathways.Q	-11.089	-38.959	.	-6.734	-36.559	.
Bile.Acid.Metabolism.Pathways.L	.	.	.	-0.717	.	.
Bile.Acid.Metabolism.Pathways.Q	.	.	.	-2.902	.	.
Inflammatory.Activity.L	3.204	.	.	14.362	.	.
Inflammatory.Activity.Q	10.773	.	.	10.641	.	.
Gut.Microbiome.Health.L	.	.	.	2.903	.	8.836
Gut.Microbiome.Health.Q	13.234	.	.	10.515	.	14.477
Digestive.Efficiency.L	.	.	.	.	-22.511	.
Digestive.Efficiency.Q	.	.	.	.	-5.037	.
Protein.Fermentation.L	-10.344	-11.652	-21.897	-5.690	.	.
Protein.Fermentation.Q	.	.	.	1.828	.	.
Gas.Production.L	-20.373	.	.	-15.073	-11.743	-21.745
Gas.Production.Q	-3.379	7.215	.	-5.128	13.994	-3.291
Methane.Gas.Production.Pathways.L	.	.	-0.628	-4.233	.	.
Methane.Gas.Production.Pathways.Q	3.425	.	.	9.529	.	.
Sulfide.Gas.Production.Pathways.L	-2.405	.	-17.394	-7.277	.	.
Sulfide.Gas.Production.Pathways.Q	1.028	.	.	-0.646	.	.
Oxalate.Metabolism.Pathways.L	.	-0.042	.	.	0.082	.
Oxalate.Metabolism.Pathways.Q	.	13.777	0.972	.	0.227	.
Salt.Stress.Pathways.L	.	-18.268	.	-2.838	.	14.407
Salt.Stress.Pathways.Q	-24.475	.	-26.052	-15.199	.	-34.196
Microbiome.Induced.Stress.L	.	.	.	.	.	.
Microbiome.Induced.Stress.Q	0.920	.	21.218	.	.	.
Carboidrati:sin_ora_pasto1	12.731	15.405	16.053	12.261	15.997	15.370
Carboidrati:cos_ora_pasto1	-1.659	-0.137	-0.162	-2.193	.	-0.315
Carboidrati:glucosio_1h_prima	3.607	5.308	4.214	3.857	4.787	4.994
Carboidrati:HbA1c	1.345	.	.	0.855	.	.

Tabella A.9: Coefficienti stimati dai metodi penalizzati (Lasso, MCP, SCAD, Group Lasso, Group MCP e Group SCAD) in corrispondenza del relativo valore λ_{min}. Il simbolo "." indica le variabili non selezionate dal modello.

	Con intercetta casuale		Senza intercetta casuale	
	Stima	Std. Error	Stima	Std. Error
(Intercetta)	57.913	6.015	38.201	0.031
Tipo pasto:cena	-17.263	0.194	-22.717	0.190
Tipo pasto:pranzo	-18.738	0.223	-24.153	0.193
Tipo pasto:snack	-20.811	0.211	-27.758	0.200
Genere (M)	.		8.275	0.123
Identità:ispanici/latini	.		17.178	0.188
Identità:bianchi	.		4.219	0.317
Calorie	10.870	0.106	2.168	0.061
Carboidrati	19.553	0.071	22.765	0.065
Proteine	-13.664	0.052	-10.786	0.047
Grassi	-6.857	0.068	.	
Fibre	-1.263	0.036	.	
Calorie assunte nelle 2h prima	1.559	0.035	1.406	0.035
Calorie assunte nelle 3h prima	.		.	
Calorie assunte nelle 6h prima	-0.819	0.055	-3.278	0.123
Calorie assunte nelle 12h prima	-9.652	0.115	-4.658	0.139
Carboidrati assunti nelle 3h prima	.		.	

Carboidrati assunti nelle 6h prima	.		4.286	0.121
Carboidrati assunti nelle 12h prima	1.126	0.096	-4.338	0.127
Fibre assunte nelle 12h prima	.		.	
Fibre assunte nelle 24h prima	2.103	0.039	2.429	0.037
Media HR nei 30min successivi	6.803	0.058	.	
Delta HR	1.790	0.035	2.724	0.034
Slope HR	.		-0.089	0.037
AUC HR nelle 2h successive	.		.	
Numero spikes	-1.638	0.054	.	
Numero blocchi contigui	.		.	
Media HR sopra soglia	.		.	
Max HR sopra soglia	.		.	
Numero salti grandi	-2.226	0.035	-2.661	0.035
Tempo al primo spike	.		.	
Calorie bruciate nelle 2h successive	-13.099	0.046	-10.653	0.044
Età	-14.047	8.259	-9.709	0.065
BMI	.		.	
Peso	.		15.382	0.073
Altezza	.		4.308	0.073
Puntura dito glucosio 1	-17.725	20.516	.	
Puntura dito glucosio 2	33.131	19.206	48.214	0.134
Puntura dito glucosio 3	10.465	19.591	28.771	0.159
HbA1c	18.901	23.510	-23.159	0.168
Insulina	-6.950	11.035	-4.248	0.091
Colesterolo	-6.789	8.267	-2.139	0.058
Colesterolo HDL	-5.391	10.897	11.591	0.081
Colesterolo LDL	-9.550	9.227	-11.742	0.077
Giorno studio	.		.	
Glucosio 1h prima	-42.273	0.059	-40.173	0.058
Glucosio 2h prima	-3.332	0.062	-2.984	0.062
Glucosio 4h prima	-5.005	0.046	-3.029	0.044
sin_ora_pasto1	0.442	0.134	-2.902	0.125
cos_ora_pasto1	.		.	
sin_ora_pasto2	-3.146	0.059	-3.774	0.058
cos_ora_pasto2	-1.745	0.062	.	
sin_ora_pasto3	-1.245	0.038	-1.507	0.037
cos_ora_pasto3	2.527	0.055	4.287	0.053
Gut.Lining.Health.L	10.610	19.314	-4.075	0.093
Gut.Lining.Health.Q	12.242	15.281	4.346	0.125
LPS.Biosynthesis.Pathways.L	.		-14.110	0.150
LPS.Biosynthesis.Pathways.Q	.		9.249	0.124
Biofilm..Chemotaxis..and.Virulence.Pathways.L	-13.129	20.844	.	
Biofilm..Chemotaxis..and.Virulence.Pathways.Q	-2.098	14.487	.	
TMA.Production.Pathways.L	8.266	15.938	12.358	0.102
TMA.Production.Pathways.Q	-20.479	16.662	-37.630	0.123
Ammonia.Production.Pathways.L	.		-7.152	0.132
Ammonia.Production.Pathways.Q	.		-6.843	0.091
Metabolic.Fitness.L	-23.297	22.041	-35.331	0.142
Metabolic.Fitness.Q	2.949	14.193	21.382	0.103
Active.Microbial.Diversity.L	8.035	22.462	4.027	0.168
Active.Microbial.Diversity.Q	8.250	15.992	14.635	0.138
Butyrate.Production.Pathways.L	-8.920	26.763	.	
Butyrate.Production.Pathways.Q	-6.411	15.201	.	
Flagellar.Assembly.Pathways.L	.		20.965	0.137
Flagellar.Assembly.Pathways.Q	.		12.706	0.104

Putrescine.Production.Pathways.L	23.236	20.287	1.946	0.138
Putrescine.Production.Pathways.Q	11.557	16.050	28.637	0.178
Uric.Acid.Production.Pathways.L	.		14.466	0.165
Uric.Acid.Production.Pathways.Q	.		-38.635	0.131
Bile.Acid.Metabolism.Pathways.L	.		.	
Bile.Acid.Metabolism.Pathways.Q	.		.	
Carboidrati:sin_ora_pasto1	12.460	0.055	13.549	0.054
Carboidrati:cos_ora_pasto1	-2.757	0.038	-2.770	0.038
Carboidrati:HbA1c	.		.	
Carboidrati:glucosio_1h_prima	4.004	0.038	4.739	0.038

Tabella A.10: Confronto tra le stime dei coefficienti ottenute con penalità Group Lasso nei modelli con e senza intercetta casuale. Il simbolo "." indica le variabili non selezionate dal modello.

	Stima	Stima	Stima	Stima
(Intercetta)	52.993	41.814	47.973	47.364
Tipo pasto:cena	-7.727	-1.568	-7.927	-8.006
Tipo pasto:pranzo	-12.184	-3.214	-11.583	-11.583
Tipo pasto:snack	-12.942	-2.952	-12.764	-13.408
Genere (M)	.	.	11.876	9.051
Identità:ispanici/latini	.	-0.071	.	.
Identità:bianchi	.	-0.367	.	-26.064
Calorie	2.293	0.021	.	.
Carboidrati	21.334	-0.137	19.739	19.544
Proteine	-9.543	0.032	-10.687	-6.923
Grassi	-1.069	.	-0.688	.
Fibre	-0.163	-0.029	.	.
Calorie assunte nelle 2h prima	1.285	-0.028	.	.
Calorie assunte nelle 3h prima	.	-1.845	.	.
Calorie assunte nelle 6h prima	-1.242	-1.069	.	-1.344
Calorie assunte nelle 12h prima	-4.827	-0.001	-9.933	-1.268
Carboidrati assunti nelle 3h prima	.	-2.730	.	.
Carboidrati assunti nelle 6h prima	.	-1.839	.	.
Carboidrati assunti nelle 12h prima	-3.346	.	.	-2.881
Fibre assunte nelle 12h prima	.	.	.	.
Fibre assunte nelle 24h prima	1.176	.	0.36	.
Media HR nei 30min successivi	3.449	0.057	7.935	.
Delta HR	2.110	-0.047	.	.
Slope HR	.	0.047	0.213	.
AUC HR nelle 2h successive	.	-0.001	.	.
Numero spikes	-0.222	.	.	.
Numero blocchi contigui	.	.	.	.
Media HR sopra soglia	.	.	.	.
Max HR sopra soglia	.	.	.	.
Numero salti grandi	-1.688	.	.	.
Tempo al primo spike	.	-0.604	.	.
Calorie bruciate nelle 2h successive	-11.659	0.053	-14.208	-9.085
Età	-3.792	.	-17.552	.
BMI	.	.	.	.
Peso	.	.	21.006	.
Altezza	.	4.037	.	.
Puntura dito glucosio 1	3.799	.	.	.
Puntura dito glucosio 2	2.136	-1.592	58.195	20.212

Puntura dito glucosio 3	6.355	.	14.478	13.206
HbA1c	8.605	.	.	.
Insulina	-0.180	2.179	-15.522	.
Colesterolo	-0.704	6.305	.	-0.987
Colesterolo HDL	-2.840	4.837	13.782	.
Colesterolo LDL	-4.985	.	-15.493	-1.365
Giorno studio	.	-14.216	.	.
Glucosio 1h prima	-37.775	-2.875	-43.584	-40.342
Glucosio 2h prima	-2.058	1.005	-1.164	.
Glucosio 4h prima	-1.159	0.980	-3.41	.
sin_ora_pasto1	2.901	0.007	.	2.6
cos_ora_pasto1	.	-1.049	.	.
sin_ora_pasto2	-3.179	-0.013	-5.6	-1.148
cos_ora_pasto2	-1.230	.	.	.
sin_ora_pasto3	-0.322	3.629	-0.271	.
cos_ora_pasto3	4.808	0.031	4.969	3.191
Gut.Lining.Health.L	0.078	0.149	.	.
Gut.Lining.Health.Q	1.755	0.070	.	16.655
LPS.Biosynthesis.Pathways.L	.	0.124	.	-1.149
LPS.Biosynthesis.Pathways.Q	.	.	12.015	12.899
Biofilm..Chemotaxis..and.Virulence.Pathways.L	0.071	.	.	.
Biofilm..Chemotaxis..and.Virulence.Pathways.Q	-0.113	.	10.966	.
TMA.Production.Pathways.L	0.303	.	19.713	9.932
TMA.Production.Pathways.Q	-0.950	.	-23.576	-12.973
Ammonia.Production.Pathways.L	.	.	.	.
Ammonia.Production.Pathways.Q	.	1.345	-1.469	.
Metabolic.Fitness.L	-0.663	1.983	-40.933	-31.722
Metabolic.Fitness.Q	1.395	.	37.259	22.631
Active.Microbial.Diversity.L	-1.660	.	25.654	14.325
Active.Microbial.Diversity.Q	1.449	.	23.119	.
Butyrate.Production.Pathways.L	-0.925	.	-41.302	.
Butyrate.Production.Pathways.Q	0.842	.	-11.384	.
Flagellar.Assembly.Pathways.L	.	.	17.452	29.129
Flagellar.Assembly.Pathways.Q	.	0.789	18.077	1.522
Putrescine.Production.Pathways.L	0.057	3.335	10.77	.
Putrescine.Production.Pathways.Q	0.023	.	.	2.117
Uric.Acid.Production.Pathways.L	.	.	.	.
Uric.Acid.Production.Pathways.Q	.	.	-51.518	-17.419
Bile.Acid.Metabolism.Pathways.L	.	.	10.466	.
Bile.Acid.Metabolism.Pathways.Q	.	13.806	-8.031	.
Carboidrati:sin_ora_pasto1	12.326	3.830	12.998	14.756
Carboidrati:cos_ora_pasto1	-2.935	-4.659	-0.805	.
Carboidrati:HbA1c	.	0.909	7.442	.
Carboidrati:glucosio_1h_prima	4.342	2.607	.	.

Tabella A.11: Da sinistra a destra: coefficienti stimati dal modello con intercetta casuale ($\lambda = \lambda_{opt}$ - algoritmo `glmmLasso`), coefficienti stimati dal modello con intercetta e pendenza casuale ($\lambda = 5000$ - algoritmo `glmmLasso`), coefficienti stimati dal modello con intercetta e pendenza casuale ($\boldsymbol{\Psi}$ multiplo dell'identità), coefficienti stimati dal modello con intercetta e pendenza casuale ($\boldsymbol{\Psi}$ piena). Il simbolo "`.`" indica le variabili non selezionate dal modello.

A.2 Figure

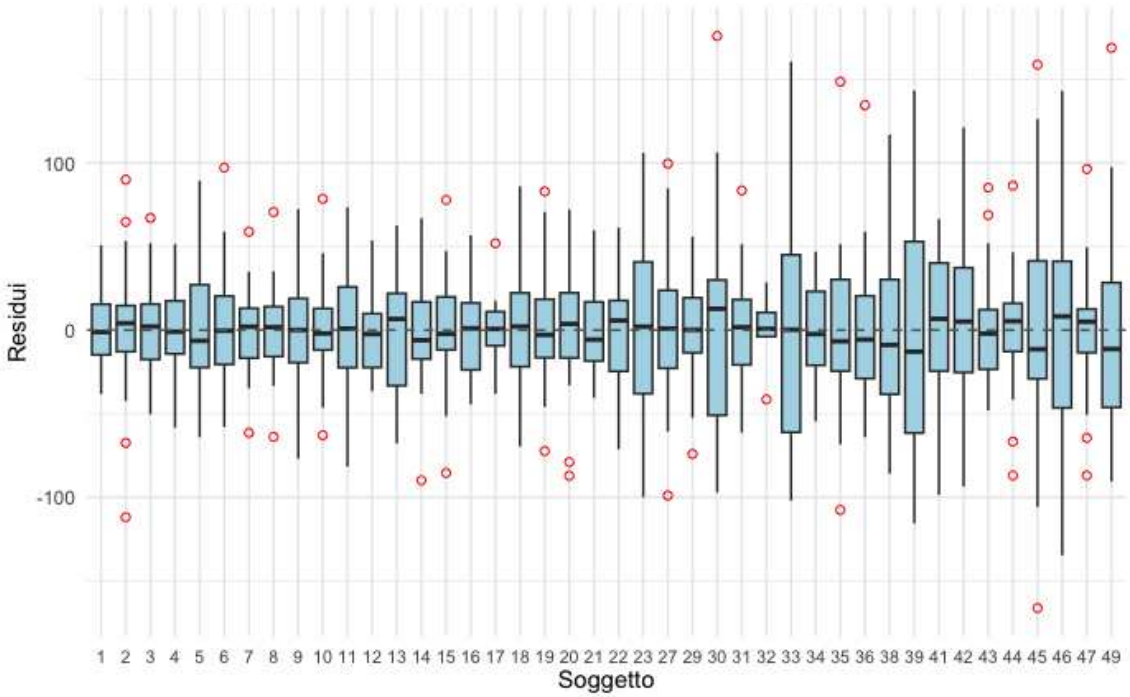

FIGURA A.1: Distribuzione dei residui del modello lineare con intercetta e pendenza casuale per i carboidrati, con penalizzazione SCAD, separatamente per ciascun soggetto.

A.3 Codice

A.3.1 Funzione 1: Selezione effetti casuali

```
# Funzione per la selezione di effetti casuali
selection_strategy <- function(y, X, group, kappa = 0.05,
                               true_random_slopes_idx = NULL,
                               comparison = TRUE) {
  n <- length(y)
  p <- ncol(X)

  # Step (i)
  cvfit <- cv.glmnet(X[,-1], y, alpha = 1)
  beta_lasso <- coef(cvfit, s = "lambda.min")
  S_init <- which(beta_lasso != 0)
  data_active <- data.frame(y = y, X[,S_init, drop=FALSE], group=factor(group))
  colnames(data_active)[2:(length(data_active)-1)] <- paste0("X",S_init)
  formula_active <- as.formula(paste("y ~-1+", paste(paste0("X", S_init),
                    collapse = "+")))
```

```
model_active <- lm(formula_active, data = data_active)
logLik_ml <- logLik(model_active)
df_lambda <- length(S_init)
BIC0 <- -2 * as.numeric(logLik_ml) + df_lambda * log(n)

# Step (ii):
theta2_hat <- numeric(length(S_init))
BICtheta <- numeric(length(S_init))
names(theta2_hat) <- names(BICtheta) <- paste0("X", S_init)

for (i in seq_along(S_init)) {
  k <- S_init[i]

  # Formula: formula_active + random effect for X_k
  f_fixed <- paste(paste0("X", S_init), collapse = " + ")
  f_random <- paste0("(0 + X", k, " | group)")
  formula_k <- as.formula(paste("y ~ -1+", f_fixed, "+", f_random))

  model_k <- lmer(formula_k, data = data_active, REML = FALSE,
                  control = lmerControl(check.conv.singular = "ignore"))

  # variance estimate for the random effect coefficient
  theta2_hat[i] <- as.numeric(VarCorr(model_k)$group[1])

  # BIC
  BICtheta[i] <- BIC(model_k)
}

# Step (iii):
R_kappa <- S_init[which(theta2_hat > kappa & BICtheta <= BIC0)]

# Step (iv):
# Estimate of final_model
if (length(R_kappa) > 0) {

  # Formula:
  f_fixed <- paste(paste0("X", S_init), collapse = " + ")
  f_random <- paste0("(0 + X", R_kappa, " | group)", collapse = " + ")
  formula_final <- as.formula(paste("y ~ -1 +", f_fixed, "+", f_random))

  model_final <- lmer(formula_final, data = data_active, REML = FALSE,
                      control = lmerControl(check.conv.singular = "ignore"))
} else {
  model_final <- NULL}
```

```
  Psi_hat <- NULL
  nonzero_Psi <- NULL

  if (!is.null(model_final)){
    Psi_hat <- VarCorr(model_final)
    group_terms <- Psi_hat[grepl("^group", names(Psi_hat))]
    variances <- sapply(group_terms, function(x) diag(x)[1])
    names(variances) <- sapply(group_terms, function(x) colnames(x)[1])
    nonzero_Psi <- variances[variances > 1e-4]
    R_kappa2 <- as.numeric(sub("X", "", names(nonzero_Psi)))
  }

  # Step (iv):
  if (!is.null(true_random_slopes_idx)) {
    TP <- sum(R_kappa2 %in% true_random_slopes_idx)
    FP <- sum(!(R_kappa2 %in% true_random_slopes_idx))
    FN <- sum(!(true_random_slopes_idx %in% R_kappa2))
  }else{
    TP <- NULL
    FP <- NULL
    FN <- NULL
  }

  if (comparison) {
      cat("TP:", TP, "| FP:", FP, "| FN:", FN, "\n")
      metrics <- list(TP = TP, FP = FP, FN = FN)
    } else {
      metrics <- NULL
  }

  return(list(
    S_init = S_init,
    theta2_hat = theta2_hat,
    BICtheta = BICtheta,
    BIC0 = BIC0,
    R_kappa = R_kappa,
    R_kappa2 = R_kappa2,
    metrics=metrics))
}

###### Funzione generale per simulare dati gerarchici ######
simulate_mixed_data <- function(N, p, beta_nonzero_idx, beta_true,
                                random_slopes_idx, sigma2_eps,
                                theta, n_min=3, n_max=10) {
```

```
  # Gruppi sbilanciati
  ni_vec <- sample(n_min:n_max, N, replace = TRUE)
  group <- rep(1:N, ni_vec)
  n <- sum(ni_vec)

  # Matrice X
  X <- cbind(1, matrix(rnorm(n * (p-1)), nrow = n))
  colnames(X) <- paste0("X", 1:p)

  # Vettore beta
  beta <- rep(0, p)
  beta[beta_nonzero_idx] <- beta_true

  # Parte fissa
  eta_fixed <- X %*% beta

  # Effetti casuali
  b_mat <- matrix(0, nrow = N, ncol = length(random_slopes_idx))
  for (j in seq_along(random_slopes_idx)) {
    b_mat[, j] <- rnorm(N, 0, sqrt(theta[j]))
  }

  random_effect <- numeric(n)
  for (i in 1:N) {
    idx_i <- which(group == i)
    for (j in seq_along(random_slopes_idx)) {
      x_col <- X[idx_i, random_slopes_idx[j]]
      random_effect[idx_i] <- random_effect[idx_i] + b_mat[i, j] * x_col
    }}

  # Rumore
  epsilon <- rnorm(n, 0, sqrt(sigma2_eps))

  # Risposta
  y <- eta_fixed + random_effect + epsilon

  list(y = y,
       X = X,
       group = group,
       beta = beta,
       random_slopes_idx = random_slopes_idx,
       theta = theta,
       ni_vec = ni_vec)
}
```

A.3.1.1 Simulazione 1: Selezione effetti casuali

```
# Funzione per il calcolo delle metriche
compute_metrics <- function(tp, fp, fn) {
  precision <- if ((tp + fp) > 0) tp / (tp + fp) else NA
  recall <- if ((tp + fn) > 0) tp / (tp + fn) else NA
  f1 <- if (!is.na(precision) & !is.na(recall) & (precision + recall > 0)) {
    2 * precision * recall / (precision + recall)
  } else NA
  c(Precision = precision, Recall = recall, F1 = f1)
}

###### SIMULAZIONE ######
selection_simulation <- function(nsim = 100, N, p, beta_nonzero_idx,
    beta_true, random_slopes_idx, sigma2, theta, kappa = 0.05,
    comparison = FALSE) {

  # Risultati
  results <- data.frame(TP = numeric(nsim), FP = numeric(nsim),
    FN = numeric(nsim), Precision = numeric(nsim),
    Recall = numeric(nsim), F1 = numeric(nsim))

  selection_mat <- matrix(0, nrow = nsim, ncol = length(random_slopes_idx))
  colnames(selection_mat) <- paste0("X", random_slopes_idx)

  for (rep in 1:nsim) {
    cat("rep: ", rep, "\n")
    sim <- simulate_mixed_data(N = N, p = p,
                               beta_nonzero_idx = beta_nonzero_idx,
                               beta_true = beta_true,
                               random_slopes_idx = random_slopes_idx,
                               sigma2 = sigma2, theta = theta)

    res <- selection_strategy(y = sim$y, X = sim$X, group = sim$group,
                              kappa = kappa,
                              true_random_slopes_idx = sim$random_slopes_idx,
                              comparison = comparison)

    if (!is.null(res$metrics)) {
      tp <- res$metrics$TP
      fp <- res$metrics$FP
      fn <- res$metrics$FN
      met <- compute_metrics(tp, fp, fn)
      results[rep, ] <- c(tp, fp, fn, met)
```

```
    } else {
      results[rep, ] <- NA
    }

    if (!is.null(res$R_kappa2)) {
      selected_vars <- res$R_kappa2
      selection_mat[rep, ] <- as.numeric(random_slopes_idx %in% selected_vars)
    }
  }

  summary <- colMeans(results, na.rm = TRUE)
  per_variable_active <- colMeans(selection_mat, na.rm = TRUE)

  return(list(results = results, summary = summary, selection_mat = selection_mat,
              per_variable_active = per_variable_active))
}

set.seed(123)
sim_results <- selection_simulation(nsim = 100, N = 40, p = 100,
                                    beta_nonzero_idx = c(1, 5, 10, 98),
                                    beta_true = c(1, -3, 2.1, 0.8),
                                    random_slopes_idx = c(1, 10),
                                    sigma2 = 0.5, theta= c(0.7, 1.3),
                                    kappa = 0.05,
                                    comparison = TRUE)

# Risultati
sim_results$summary
sim_results$per_variable_active
```

A.3.1.2 Applicazione ai dati

```
# Lettura dati
dati<- read.csv("dati.csv")
dati$subject <- as.factor(dati$subject)
dati$Meal.Type <- as.factor(dati$Meal.Type)
dati$Gender <- as.factor(dati$Gender)
dati$Self.identify <- as.factor(dati$Self.identify)

# Trasformazione dei punteggi in fattori ordinati
dati <- dati %>% mutate(across(
    c(Gut.Lining.Health, LPS.Biosynthesis.Pathways,
      Biofilm..Chemotaxis..and.Virulence.Pathways, TMA.Production.Pathways,
```

```
      Ammonia.Production.Pathways, Metabolic.Fitness, Active.Microbial.Diversity,
      Butyrate.Production.Pathways, Flagellar.Assembly.Pathways,
      Putrescine.Production.Pathways, Uric.Acid.Production.Pathways,
      Bile.Acid.Metabolism.Pathways, Inflammatory.Activity, Gut.Microbiome.Health,
      Digestive.Efficiency, Protein.Fermentation, Gas.Production,
      Methane.Gas.Production.Pathways, Sulfide.Gas.Production.Pathways,
      Oxalate.Metabolism.Pathways, Salt.Stress.Pathways, Microbiome.Induced.Stress),
    as.ordered))

# Vettore con tutti i soggetti unici
soggetti <- unique(dati$subject)

# Variabili numeriche da standardizzare
numeriche <- names(dati)[sapply(dati, is.numeric)]
numeriche <- setdiff(numeriche, "y")

# Train set escludendo l'ultima osservazione per ciascun soggetto
train <- do.call(rbind, lapply(soggetti, function(s) {
  dati_s <- dati[dati$subject == s, ]
  if (nrow(dati_s) > 1) dati_s[1:(nrow(dati_s) - 1), ] else NULL}))

# Media e sd sul train set
means <- sapply(train[, numeriche], mean, na.rm = TRUE)
sds <- sapply(train[, numeriche], sd, na.rm = TRUE)

# Standardizzazione
train[, numeriche] <- scale(train[, numeriche], center = means, scale = sds)

# Matrice del disegno
X <- model.matrix(y~Meal.Type + Gender + Self.identify +
                    Calories + Carbs + Protein + Fat + Fiber +
                    Cal2 + Cal3 + Cal6 + Cal12 + Carbs3 + Carbs6 +
                    Carbs12 + Fiber12 + Fiber24 + mean_HR_post30 + delta_HR +
                    slope_HR + AUC_HR_2h + n_spikes + n_blocchi_contigui +
                    HR_mean_over_soglia + HR_max_over_soglia + n_salti_grandi +
                    tempo_al_primo_spike + sum_calories_2h + Age + BMI + Body.weight
                    + Height + X.1.Contour.Fingerstick.GLU
                    + X.2.Contour.Fingerstick.GLU + X.3.Contour.Fingerstick.GLU +
                    A1c.PDL..Lab. + Insulin + Cholesterol + HDL + LDL..Cal. +
                    giorno_studio + gluc_1h_prima + gluc_2h_prima + gluc_4h_prima +
                    sin_ora_pasto1 + cos_ora_pasto1 +  sin_ora_pasto2 +
                    cos_ora_pasto2 + sin_ora_pasto3 + cos_ora_pasto3 +
                    Gut.Lining.Health + LPS.Biosynthesis.Pathways +
                    Biofilm..Chemotaxis..and.Virulence.Pathways +
```

```
                    TMA.Production.Pathways + Ammonia.Production.Pathways +
                    Metabolic.Fitness + Active.Microbial.Diversity +
                    Butyrate.Production.Pathways + Flagellar.Assembly.Pathways +
                    Putrescine.Production.Pathways + Uric.Acid.Production.Pathways +
                    Bile.Acid.Metabolism.Pathways, data=train)

# Dataframe
dati_uniti <- data.frame(y = train$y, gruppo = train$subject, X)

# Applicazione del metodo
selection_dati <- selection_strategy(dati_uniti$y, X,  dati_uniti$gruppo,
        kappa=0.05, comparison=FALSE)
selection_dati
colnames(X)[selection_dati$R_kappa2]
```

A.3.2 Funzione 2: Algoritmo per la selezione e la stima degli effetti fissi con penalità SCAD o Lasso

```
# Librerie
library(glmnet)
library(MASS)
library(dplyr)
library(lme4)

# Paramatri
cmin = 1e-6
cmax = 1e8

## Costruzione Psi
# multiplo diagonale/ identità: valori originali
# altrimenti: parametrizzazione di Cholesky
build_Psi_single <- function(theta, q) {
  if (length(theta)==1){
    Psi <- theta * diag(q)
  }else if (length(theta)>1 & length(theta)==q){
    Psi <- diag(theta)
  }else{
    L <- matrix(0, nrow = q, ncol = q)
    idx <- 1
    for (i in 1:q) {
      for (j in 1:i) {
        if (i == j) {
```

```
          L[i, j] <- exp(theta[idx])
        } else {
          L[i, j] <- theta[idx]
        }
        idx <- idx + 1
      }
    }
    Psi = L %*% t(L)
  }
  return(Psi)
}

## Costruzione Psi come matrice a blocchi
build_Psi_block_d <- function(theta, I, q) {
  Psi_single <- build_Psi_single(theta, q)
  Psi_list <- replicate(I, Psi_single, simplify = FALSE)
  Psi_block <- Matrix::bdiag(Psi_list)
  return(as.matrix(Psi_block))
}

## Calcolo della log-verosimiglianza negativa
NeglogLik<- function(eta, y, X, Z, beta, I, group, q) {
  theta <- (eta[-length(eta)])
  sigma2 <- (eta[length(eta)])
  Psi <- build_Psi_single(theta, q)
  nll <- 0
  group_lengths <- as.numeric(table(group))
  group_start <- cumsum(c(1, head(group_lengths, -1)))
  group_end <- cumsum(group_lengths)

  for (i in seq_along(group_lengths)) {
    idx <- group_start[i]:group_end[i]
    s <- seq_along(group_lengths)[i]

    y_i <- y[idx]
    X_i <- as.matrix(X[idx, ])
    Z_i <- as.matrix(Z[idx, ((s - 1) * q + 1):(s * q)])
    n_i <- length(y_i)
    V_i <- Z_i %*% Psi %*% t(Z_i) + sigma2 * diag(n_i)
    logdetV_i <- determinant(V_i, logarithm=TRUE)$modulus

    if (rcond(V_i) < 1e-12 || any(!is.finite(V_i))) {
      return(1e10)
    }
```

```
    cholV_i <- tryCatch(chol(V_i), error = function(e) return(NULL))
    if (is.null(cholV_i)) return(1e10)
    Vinv_i <- chol2inv(cholV_i)

    res_i <- y_i - X_i %*% beta
    quad_i <- t(res_i) %*% Vinv_i %*% res_i

    nll <- nll + 0.5 * (n_i * log(2*pi) + logdetV_i + quad_i)
  }
  return(nll)
}

## Aggiornamento vettore di parametri eta
update_eta_coordwise <- function(eta_old, y, X, Z, beta, I,
                                 group, q, struttura) {
  d <- length(eta_old)
  eta_new <- eta_old

  for (j in 1:d) {
    obj_single_eta <- function(eta_j) {
      eta_temp <- eta_new
      eta_temp[j] <- eta_j
      NeglogLik(eta_temp, y, X, Z, beta, I, group,q)
    }

    # Intervalli
    if (j <= q) {
      if (struttura=="iden"|struttura=="diago"){
        lower_j <- 1e-4
        upper_j <- 100
      } else if (struttura=="corr"){
        # Posizioni corrispondenti a log(l_kk)
        chol_indices <- cumsum(1:q)
        if (j %in% chol_indices) {
          lower_j <- log(1e-4)
          upper_j <- log(1e09)
        }else {
          lower_j <- -100
          upper_j <- 100
        }
      }
    } else if (j == d) {  # sigma^2
      lower_j <- 1e-3
      upper_j <- var(y)
```

```
    }
    res <- nlminb(start = eta_new[j], objective = obj_single_eta,
                  lower = lower_j, upper = upper_j)
    eta_new[j] <- res$par
  }
  return(eta_new)
}

## Calcolo direzione SCAD
compute_dj_SCAD <- function(j, beta_j, grad_j, hj, lambda, a = 3.7, penalize = TRUE){
  if (!penalize) {
    dj <- - grad_j / hj
  } else {
    condition1 <- (beta_j * hj - grad_j) <= lambda * (hj + 1)
    condition2 <- (beta_j * hj - grad_j) > lambda * (a * hj*j)
    if (condition1) {
      dj <- - (lambda + grad_j) / hj
    } else if (condition2) {
      dj <- - grad_j/hj
    } else {
      dj <- - ((a - 1) * grad_j - (a * lambda - beta_j)) / ((a - 1) * hj - 1)
    }
  }
  return(dj)
}

## Formula esplicita scad
scad_shrinkage <- function(z, lambda, a = 3.7) {
  abs_z <- abs(z)
  sgn_z <- sign(z)

  if (abs_z <= 2 * lambda) {
    res <- sgn_z * pmax(abs_z - lambda, 0)
  } else if (abs_z <= a * lambda) {
    res <- ((a - 1) * z - sgn_z * a * lambda) / (a - 2)
  } else {
    res <- z
  }
  return(res)
}

## Formula esplicita Lasso
soft_threshold <- function(x, t) {
  sign(x) * pmax(abs(x) - t, 0)}
```

```
## Calcolo direzione Lasso
compute_dj_LASSO <- function(j, beta_j, grad_j, hj, lambda, penalize = TRUE) {
  if (!penalize) {
    dj <- - grad_j / hj
  } else {
    u_j <- beta_j - grad_j / hj
    dj <- soft_threshold(u_j, lambda / hj) - beta_j
  }
  return(dj)
}

## Penalità SCAD
scad_penalty <- function(beta, lambda, a = 3.7, penalize = rep(TRUE, length(beta))){
  abs_beta <- abs(beta)
  pen <- numeric(length(beta))

  for (j in seq_along(beta)) {
    if (!penalize[j]) {
      pen[j] <- 0
    } else if (abs_beta[j] <= lambda) {
      pen[j] <- lambda * abs_beta[j]
    } else if (abs_beta[j] <= a * lambda) {
      pen[j] <- (-abs_beta[j]^2 + 2 * a * lambda * abs_beta[j] - lambda^2) /
      (2 * (a - 1))
    } else {
      pen[j] <- (lambda^2 * (a + 1)) / 2
    }
  }
  return(sum(pen))
}

## Penalità LASSO
lasso_penalty <- function(beta, lambda, penalize = rep(TRUE, length(beta))) {
  return(sum(lambda * abs(beta[penalize])))
}

## Funzione Q
Q_fun <- function(beta, y, X, Z, group, theta, sigma2,
                  lambda, a, penalize, penalty, q) {
  Psi <- build_Psi_single(theta, q)
  nll <- 0
  group_lengths <- as.numeric(table(group))
  group_start <- cumsum(c(1, head(group_lengths, -1)))
  group_end <- cumsum(group_lengths)
```

```
  for (i in seq_along(group_lengths)) {
    idx <- group_start[i]:group_end[i]
    s = seq_along(group_lengths)[i]

    y_i <- y[idx]
    X_i <- as.matrix(X[idx, ])
    Z_i <- as.matrix(Z[idx, ((s - 1) * q + 1):(s * q)])
    n_i <- length(y_i)

    V_i <- Z_i %*% Psi %*% t(Z_i) + sigma2 * diag(n_i)
    if (rcond(V_i) < 1e-12 || any(!is.finite(V_i))) {
      return(1e10)
    }
    cholV_i <- tryCatch(chol(V_i), error = function(e) return(NULL))
    if (is.null(cholV_i)) return(1e10)
    Vinv_i <- chol2inv(cholV_i)
    res_i <- y_i - X_i %*% beta
    quad_i <- t(res_i) %*% Vinv_i %*% res_i
    logdetV_i <- determinant(V_i, logarithm = TRUE)$modulus

    nll <- nll + 0.5 * (n_i * log(2 * pi) + logdetV_i + quad_i)
  }

  # Penalità
  if (penalty == "scad") {
    pen <- scad_penalty(beta, lambda, a, penalize)
  } else if (penalty == "lasso") {
    pen <- lasso_penalty(beta, lambda, penalize)
  }
  return(as.numeric(nll + pen))
}

## Armijo line search
armijo_line_search <- function(beta, j, dj, grad_j, hj, Q_fun, penalty_fun,lambda,
                               rho = 0.001, gamma = 0, delta = 0.1,
                               max_ls_iter = 20) {
  alpha <- 1
  Q0 <- Q_fun(beta)

  penalty_diff <- function(beta_j_old, beta_j_new) {
    penalty_fun(beta_j_new, lambda) - penalty_fun(beta_j_old, lambda)
  }

  delta_j <- grad_j * dj + gamma * dj^2 * hj + penalty_diff(beta[j], beta[j] + dj)
```

```
  for (i in 1:max_ls_iter) {
    beta_new <- beta
    beta_new[j] <- beta_new[j] + alpha * dj
    Q_new <- Q_fun(beta_new)

    if (Q_new <= Q0 + rho * alpha * delta_j) {
      break
    } else {
      alpha <- alpha * delta
    }
  }
  return(alpha)
}

## Riportare theta su scala originale
expand_theta <- function(theta, q) {
  idx_diag <- cumsum(1:q)
  theta_exp <- theta
  theta_exp[idx_diag] <- exp(theta[idx_diag])
  return(theta_exp)
}

## Calcolare Psi
get_variances_from_theta_exp <- function(theta_expo, q) {
  L <- matrix(0, q, q)
  idx <- 1
  for (i in 1:q) {
    for (j in 1:i) {
      L[i, j] <- theta_expo[idx]
      idx <- idx + 1
    }
  }
  cov_matrix <- L %*% t(L)
  return(cov_matrix)
}

## Calcolo dei valori iniziali di eta
get_eta_init_from_lmer <- function(y, X, groups, random_vars = NULL,
                                   default_theta = 0.1, default_sigma2 = 0.1,
                                   struttura,q) {
  df <- as.data.frame(X)
  df$y <- y
  df$groups <- groups
```

```
if (is.null(random_vars)) {
  stop("Specifica le variabili da includere come effetti casuali.")
}

if (struttura=="iden"){
  random_formula <- paste("1|groups")
  fixed_formula <- paste("y ~ 1")
  full_formula <- as.formula(paste(fixed_formula, "+ (", random_formula, ")"))
}else if(struttura=="diago"|struttura=="corr"){
  if (length(random_vars) > 1) {
    random_formula <- paste("1 +", paste(random_vars[-1], collapse = " + "),
    "|groups")
  } else {
    random_formula <- "1|groups"
  }
  fixed_formula <- paste("y ~ 1")
  full_formula <- as.formula(paste(fixed_formula, "+ (", random_formula, ")"))
}

# Tentativo stima lmer
try_result <- tryCatch({
  fit_init <- lmer(full_formula, data = df, REML = FALSE,
  control = lmerControl(check.conv.singular = "ignore"))

  vc <- VarCorr(fit_init)
  if(struttura=="iden"){
    Psi_hat <- build_Psi_single(vc$groups[1],q)
    theta_init <- vc$groups[1]
  }else if(struttura=="diago"){
    Psi_hat <- build_Psi_single(diag(vc$groups),q)
    theta_init <- diag(vc$groups)
  }else if(struttura=="corr"){
    Psi_hat <- as.matrix(vc$groups)

    # Cholesky
    L <- tryCatch(t(chol(Psi_hat)), error = function(e) NULL)
    if (is.null(L) || any(!is.finite(L))) stop("Cholesky fallita.")
    q <- nrow(L)
    theta_init <- c()
    for (i in 1:q) {
      for (j in 1:i) {
        if (i == j) {
          theta_init <- c(theta_init, log(L[i, j]))
        } else {
```

```
            theta_init <- c(theta_init, L[i, j])
    }}}}

    sigma2_init <- attr(vc, "sc")^2

    # Verifica finitezza
    if (any(!is.finite(theta_init)) ||
    !is.finite(sigma2_init)) stop("Valori non finiti")

    list(theta_init = theta_init, sigma2_init = sigma2_init)
  }, error = function(e) {
    warning("Problema nella stima iniziale con lmer. Uso valori di default.")

    q <- length(random_vars)
    if(struttura=="iden"){
      theta_init <- default_theta
    }else if(struttura=="diago"){
      theta_init <- rep(default_theta, q)
    }else if(struttura=="corr"){
      theta_init <- c()
      for (i in 1:q) {
        for (j in 1:i) {
          if (i == j) {
            theta_init <- c(theta_init, log(default_theta))
          } else {
            theta_init <- c(theta_init, 0.05)
          }}}
    }
    list(theta_init = theta_init, sigma2_init = default_sigma2)
  })
  return(try_result)
}

## Previsione degli effetti casuali
compute_random_effects <- function(y, X, Z, groups,
                                   beta_hat, theta_hat, sigma2_hat, q) {
  I <- length(unique(groups))
  Psi <- build_Psi_single(theta_hat, q)

  b_hat <- matrix(NA, I, q)
  group_lengths <- as.numeric(table(groups))
  group_start <- cumsum(c(1, head(group_lengths, -1)))
  group_end <- cumsum(group_lengths)
```

```
  for (i in 1:I) {
    idx <- group_start[i]:group_end[i]
    y_i <- y[idx]
    X_i <- X[idx, , drop = FALSE]
    Z_i <- Z[idx, ((i - 1) * q + 1):(i * q), drop = FALSE]
    n_i <- length(y_i)

    V_i <- Z_i %*% Psi %*% t(Z_i) + sigma2_hat * diag(n_i)
    cholV <- tryCatch(chol(V_i), error = function(e) return(NULL))
    if (is.null(cholV)) next
    Vinv_i <- chol2inv(cholV)

    b_hat_i <- Psi %*% t(Z_i) %*% Vinv_i %*% (y_i - X_i %*% beta_hat)
    b_hat[i, ] <- as.numeric(b_hat_i)
  }
  return(b_hat)
}

## Calcolo valori predetti
compute_predictions_from_bhat <- function(X, Z, groups,
                                          beta_hat, b_hat, q) {
  n <- nrow(X)
  I <- length(unique(groups))
  pred <- numeric(n)

  group_lengths <- as.numeric(table(groups))
  group_start <- cumsum(c(1, head(group_lengths, -1)))
  group_end <- cumsum(group_lengths)

  for (i in 1:I) {
    idx <- group_start[i]:group_end[i]
    X_i <- X[idx, , drop = FALSE]
    Z_i <- Z[idx, ((i - 1) * q + 1):(i * q), drop = FALSE]
    pred[idx] <- X_i %*% beta_hat + Z_i %*% b_hat[i, ]
  }
  return(pred)
}

## Funzione per la selezione degli effetti fissi
funzione_lmm <- function(y, X, groups, lambda, a = 3.7,
                         max_iter = 100, tol = 1e-4, penalize = rep(TRUE, ncol(X)),
                         beta_init = NULL, theta_init = NULL, sigma2_init = NULL,
                         penalty = c("scad", "lasso"), random_vars = NULL,
                         struttura) {
```

```
penalty <- match.arg(penalty)

I <- length(unique(groups))
p <- ncol(X)
q = length(random_vars)

# Assegnare nomi a X se mancano
if (is.null(colnames(X))) {
  colnames(X) <- paste0("X", 1:ncol(X))
}

# Inizializzare beta se non fornito
if (is.null(beta_init)) {
  cvfit <- cv.glmnet(X[,-1], y, alpha = 1, nfolds = 10)
  beta <- as.numeric(coef(cvfit, s = "lambda.min"))
} else {
  beta <- beta_init
}

# Inizializzare theta e sigma2 se non forniti
if (is.null(theta_init) || is.null(sigma2_init)) {
  init_vals <- get_eta_init_from_lmer(y, X, groups,
                                      random_vars = random_vars,
                                      struttura = struttura, q=q)
  theta_init <- init_vals$theta_init
  sigma2_init <- init_vals$sigma2_init
}
theta <- theta_init
sigma2 <- sigma2_init

# Estrarre nomi degli effetti casuali
if (is.null(random_vars)) stop("Devi specificare random_vars.")

# Estrarre le colonne da X per costruire Z
Z_raw <- as.matrix(X[, random_vars, drop = FALSE])
q <- ncol(Z_raw)

# Creare Z
n <- length(y)
Z <- matrix(0, nrow = n, ncol = I * q)
for (i in 1:n) {
  id <- groups[i]
  Z[i, ((id - 1) * q + 1):(id * q)] <- Z_raw[i, ]
}
```

```
for (iter in 1:max_iter) {
  beta_old <- beta
  Psi <- build_Psi_block_d(theta, I, q)
  V <- Z %*% Psi %*% t(Z) + sigma2 * diag(nrow(Z))
  cholV <- tryCatch(chol(V), error = function(e) return(NULL))
  if (is.null(cholV)) stop("Matrice V non SPD")
  Vinv <- chol2inv(cholV)

  for (j in 1:p) {
    r <- y - X %*% beta
    grad_j <- - t(X[, j]) %*% Vinv %*% r
    Ijj <- t(X[, j]) %*% Vinv %*% X[, j]
    hj <- min(max(Ijj, cmin), cmax)

    if (penalty == "scad") {
      if (Ijj==hj){
        # previsione marginale
        y_tilde <- X[, -j, drop=FALSE] %*% beta[-j]

        if (!penalize[j]) {
          # Aggiornamento non penalizzato classico
          grad_j <- - t(X[, j]) %*% Vinv %*% (y - y_tilde)
          dj <- as.numeric(grad_j) / hj
          Q_current <- function(beta_input)
            Q_fun(beta_input, y, X, Z, group = groups, theta, sigma2,
                  lambda, a, penalize, penalty,q)
          alpha_j <- armijo_line_search(beta, j, dj, grad_j, hj,
                                        Q_current,scad_penalty, lambda)
          beta[j] <- beta[j] + alpha_j * dj
        } else {
          # Penalizzato
          z <- as.numeric(t(y - y_tilde) %*% Vinv %*% X[, j]) / hj
          beta[j] <- scad_shrinkage(z, lambda, a)
        }
      }else{
        dj <- compute_dj_SCAD(j, beta[j], grad_j, hj, lambda, a, penalize[j])
        Q_current <- function(beta_input)
          Q_fun(beta_input, y, X, Z, group = groups, theta, sigma2,
                lambda, a, penalize, penalty,q)
        alpha_j <- armijo_line_search(beta, j, dj, grad_j, hj,
                                      Q_current,scad_penalty, lambda)
        beta[j] <- beta[j] + alpha_j * dj
        }
    } else if (penalty == "lasso") {
```

```
      if (Ijj==hj){
        # previsione marginale
        y_tilde <- X[, -j, drop=FALSE] %*% beta[-j]

        if (!penalize[j]) {
          # Aggiornamento non penalizzato classico
          grad_j <- - t(X[, j]) %*% Vinv %*% (y - y_tilde)
          dj <- as.numeric(grad_j) / hj
          Q_current <- function(beta_input)
            Q_fun(beta_input, y, X, Z, group = groups, theta, sigma2,
                  lambda, a, penalize, penalty,q)
          alpha_j <- armijo_line_search(beta, j, dj, grad_j, hj,
                                        Q_current,lasso_penalty, lambda)
          beta[j] <- beta[j] + alpha_j * dj
        } else {
          # Penalizzato
          z <- as.numeric(t(y - y_tilde) %*% Vinv %*% X[, j]) / hj
          beta[j] <- sign(z) * pmax(abs(z) - lambda / hj, 0)
        }
      }else{
        dj <- compute_dj_LASSO(j, beta[j], grad_j, hj, lambda, penalize[j])
      Q_current <- function(beta_input)
        Q_fun(beta_input, y, X, Z, group = groups, theta, sigma2,
              lambda, a, penalize, penalty,q)
      alpha_j <- armijo_line_search(beta, j, dj, grad_j, hj,
                                    Q_current, lasso_penalty, lambda)
      beta[j] <- beta[j] + alpha_j * dj
  }}}

  eta <- c(theta, sigma2)
  eta_new <- update_eta_coordwise(eta, y, X, Z, beta, I,
                                  group = groups,q, struttura=struttura)
  theta <- eta_new[1:length(theta)]
  sigma2 <- eta_new[length(eta_new)]

  if (sqrt(sum((beta - beta_old)^2)) < tol) {
    cat("Convergenza raggiunta in", iter, "iterazioni\n")
    break
  }
}

u <- compute_random_effects(y, X, Z, groups=groups,
                            beta_hat=beta, theta_hat=theta,
                            sigma2_hat=sigma2, q=q)
```

```
  if (struttura=="corr"){
    theta_exp <- expand_theta(theta, q)
    cov_matrix <- get_variances_from_theta_exp(theta_exp, q)
  }else{
    theta_exp <- theta
    if (struttura=="iden"){
      cov_matrix <- theta_exp * diag(q)
    }else{
      cov_matrix <- diag(theta_exp)
    }
  }
  return(list(beta = beta, theta = theta_exp, sigma2 = sigma2, mat=cov_matrix,
              u=u))
}

## Calcolo BIC
compute_BIC <- function(logLik, beta_hat, eta_hat, y) {
  n <- length(y)
  df_beta <- sum(beta_hat != 0)
  df_eta <- length(eta_hat)
  df_total <- df_beta + df_eta
  BIC <- 2 * logLik + log(n) * df_total
  return(BIC)
}
```

Bibliografia

Ahn, M., Zhang, H. H. & Lu, W. (2012). Moment-based method for random effects selection in linear mixed models. *Statistica Sinica* **22(4)**, 1539–1562.

Allison, D. B., Paultre, F., Maggio, C., Mezzitis, N. & Pi-Sunyer, F. X. (1995). The use of areas under curves in diabetes research. *Diabetes Care* **18**, 245–250.

American Diabetes Association (2001). Postprandial blood glucose. *Diabetes Care 1* **24**, 775–778.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67(1)**, 1–48.

Bertsekas, D. (2003). *Nonlinear Programming.* Athena Scientific.

Bickel, P. J., Ritov, Y. & Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* **37(4)**, 1705 – 1732.

Bozdogan, H. (1987). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika* **52(3)**, 345–370.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88(421)**, 9–25.

Buscemi, S. & Plaia, A. (2019). Model selection in linear mixed-effect models. *AStA Advances in Statistical Analysis* **104**.

Davison, A. (2003). *Statistical Models.* Cambridge University Press, Cambridge.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96(456)**, 1348–1360.

Fan, Y. & Li, R. (2012). Variable selection in linear mixed effects models. *The Annals of Statistics* **40(4)**, 2043 – 2068.

Ghosh, A. & Thoresen, M. (2016). Non-concave penalization in linear mixed-effects models and regularized selection of fixed effects. *AStA Advances in Statistical Analysis* **102(4)**.

Goeman, J. J. (2010). L1 penalized estimation in the cox proportional hazards model. *Biometrical Journal. Biometrische Zeitschrift* **52(1)**, 70–84.

Groll, A. & Tutz, G. (2014). Variable selection for generalized linear mixed models by l1-penalized estimation. *Statistics and Computing* **24**, 137–154.

Guizar, R., Noriega, L., Rivera, A. L., Guevara, M., Torres, N. & Tovar, A. (2023). A new approach to personalized nutrition: Postprandial glycemic response and its relationship to gut microbiota. *Archives of Medical Research* **54**.

Gutierrez-Osuna, R., Kerr, D., Mortazavi, B. & Das, A. (2025). Cgmacros: a scientific dataset for personalized nutrition and diet monitoring (version 1.0.0). `https://doi.org/10.13026/3z8q-x658`. PhysioNet. RRID:SCR_007345.

Hastie, T., Tibshirani, R. & Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations (1st ed.)*. Chapman and Hall/CRC.

Institute of Medicine (2005). *Dietary Reference Intakes for Energy, Carbohydrate, Fiber, Fat, Fatty Acids, Cholesterol, Protein, and Amino Acids.* Washington, DC: The National Academies Press.

Kang, J., Fardman, B., Ratamess, N., Faigenbaum, A. & Bush, J. (2023). Postprandial exercise in mitigating glycemic responses in overweight individuals and individuals with obesity and type 2 diabetes-a systematic review and meta-analysis. *Nutrients* **15**.

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software* **28**, 1–26.

Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. (2017). lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software* **82(13)**, 1–26.

Laird, N. M. & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics* **38(4)**, 963–974.

Mendes-Soares, H., Raveh-Sadka, T., Azulay, S., Edens, K., Ben-Shlomo, Y., Cohen, Y., Ofek, T., Bachrach, D., Stevens, J., Colibaseanu, D., Segal, L. & Kashyap, P. a. H. (2019). Assessment of a personalized approach to predicting postprandial glycemic responses to food among individuals without diabetes. *JAMA Network Open* **2**.

Nickel, A., Buresh, R., McLester, C., Canino, A., Wilner, G., Vaughan, K., Chung, P. & Kliszczewicz, B. (2024). The relationship between heart rate variability and glucose clearance in healthy men and women. *PLoS One* **19**.

Nocedal, J. & Wright, S. (2006). *Numerical Optimization*. Springer Science and Business Media.

Ordovas, J., Ferguson, L., Tai, E. & Mathers, J. (2018). Personalised nutrition and health. *BMJ* **361**.

Pinheiro, J., Bates, D. & R Core Team (2025). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-168.

Pinheiro, J. C. & Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer New York, NY.

Salvan, A., Sartori, N. & Pace, L. (2020). *Modelli Lineari Generalizzati*. Springer Milano.

Schelldorfer, J., Bühlmann, P. & Van De Geer, S. (2011). Estimation for high-dimensional linear mixed-effects models using 1-penalization. *Scandinavian Journal of Statistics* **38(2)**, 197–214.

Shilo, S., Godneva, A., Rachmiel, M., Korem, T., Kolobkov, D., Karady, T., Bar, N., Wolf, B., Glantz-Gashai, Y., Cohen, M., Zuckerman, L. N., Shehadeh, N., Gruber, N., Levran, N., Koren, S., Weinberger, A., Pinhas-Hamiel, O. & E., S. (2022). Prediction of personal glycemic responses to food for individuals with type 1 diabetes through integration of clinical and microbial data. *Diabetes Care* **45**, 502–511.

Tibishirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58(1)**, 267–288.

Tseng, P. & Yun, S. (2009). A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming* **117**, 387–423.

Wolever, T. M. (2004). Effect of blood sampling schedule and method of calculating the area under the curve on validity and precision of glycaemic index values. *The British journal of nutrition* **91(2)**, 295–301.

Wolever, T. M. & Jenkins, D. J. (1986). The use of the glycemic index in predicting the blood glucose response to mixed meals. *The American Journal of Clinical Nutrition* **43**, 167–172.

World Health Organization & WHO Consultation on Obesity (2000). Obesity: Preventing and managing the global epidemic. report of a who consultation. WHO Technical Report Series 894, World Health Organization, Geneva, Switzerland.

Wu, P., Luo, X., Xu, P. & Zhu, L. (2016). New variable selection for linear mixed-effects models. *Annals of the Institute of Statistical Mathematics* **69**, 627–646.

Yuan, M. & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society* **68**, 49–67.

Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., Suez, J., Mahdi, J., Matot, E., Malka, G., Kosower, N., Rein, M., Zilberman-Schapira, G., Dohnalová, L., Pevsner-Fischer, M., Bikovsky, R., Halpern, Z., Elinav, E. & Segal, E. (2015). Personalized nutrition by prediction of glycemic responses. *Cell* **163**, 1079–1094.

Zou, H., Hastie, T. & Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics* **35(5)**, 2173 – 2192.

Zuur, A., Ieno, E. N., Walker, N. & Saveliev, A. A. Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R.* Springer.