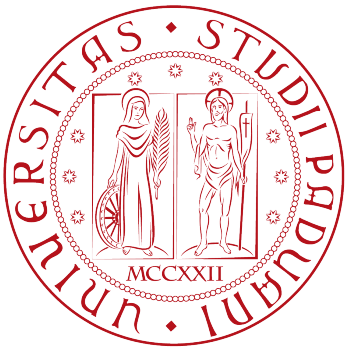
Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in

Statistica per le Tecnologie e le Scienze




RELAZIONE FINALE

# ROBUSTEZZA NEI MODELLI LINEARI NORMALI CON EFFETTI CASUALI

Relatrice Prof.ssa Laura Ventura

Dipartimento di Scienze Statistiche

Correlatore Dott. Erlis Ruli

Dipartimento di Scienze Statistiche

Laureando: Alberto Grassi

Matricola N 1147523

Anno Accademico 2018/2019

# Indice

# Introduzione

*All models are wrong, some models are useful.*

*- George E. P. Box*

La statistica dovrebbe offrire metodi non solo in grado di gestire l'incertezza dovuta al campionamento delle unità statistiche, ma anche quella dovuta al fatto che i modelli sono, nella migliore delle ipotesi, una buona approssimazione della realtà.

L'obiettivo di questo elaborato è confrontare, tramite uno studio di simulazione, il comportamento degli stimatori classici basati sulla funzione di verosimiglianza e quello degli stimatori robusti, in diversi contesti in cui il processo generatore dei dati non è esattamente quello specificato dal modello. Come campo applicativo per lo studio delle tecniche statistiche robuste, ci si focalizza su quello dei modelli lineari normali con effetti casuali. In molte applicazioni, la variabile risposta per ciascuna unità statistica è multivariata e va analizzata come un vettore casuale con componenti dipendenti. Si pensi al caso di dati longitudinali, in cui vengono rilevate misurazioni ripetute sugli stessi soggetti o, più in generale, al caso di diverse misurazioni sulle stesse unità di primo livello, siano esse soggetti o *cluster*, in diverse situazioni, non necessariamente scandite dal tempo. In un'indagine in ambito socio-economico, si può essere interessati a intervistare tutti i componenti di un nucleo famigliare, o in ambito medico a rilevare delle misurazioni sugli stessi pazienti sotto diverse condizioni sperimentali. In tutti questi casi è

importante tener conto nella scrittura del modello della dipendenza fra le misurazioni sulle stesse unità statistiche, ad esempio, ricorrendo a modelli con effetti casuali.

Lo schema di questo elaborato è strutturato nel seguente modo. Nel Capitolo 1 si presentano le tecniche classiche di stima (massima verosimiglianza e massima verosimiglianza ristretta) e di inferenza all'interno della classe dei modelli lineari normali con effetti casuali. Alla fine del capitolo si pone l'attenzione sui limiti di questo tipo di approccio in presenza di dati anomali, ovvero in presenza di una frazione di osservazioni che si discosta dal resto dei dati e si ritiene, pertanto, non rappresentativa della popolazione oggetto di studio. Nella prima parte del Capitolo 2 si presenta il concetto di robustezza attraverso la funzione di influenza e si presentano i principali stimatori robusti noti in letteratura per questa classe di modelli (stimatori di tipo S e di tipo MM). Nella seconda parte del capitolo si pone l'attenzione su stimatori robusti di tipo *trimmed*, innovativi per questa classe di modelli e per i quali viene presentata un'implementazione in R (`http://www.r-project.org/`). Questi stimatori sono basati sul concetto di massima verosimiglianza troncata e possono avere punti di rottura molto alti. Inoltre, risultano più generali rispetto agli stimatori robusti noti in letteratura essendo formulati sotto assunti meno stringenti. Il Capitolo 3 discute uno studio di simulazione condotto al fine di confrontare il comportamento dei diversi stimatori presentati nei capitoli precedenti sotto diversi scenari di numerosità campionaria e livello di contaminazione presente nei dati.

# Capitolo 1

# Modelli lineari normali con effetti casuali

## 1.1 Formulazione del modello

Esistono diverse tipologie di modelli per l'analisi di risposte correlate. In questo elaborato sono trattati in particolare i modelli con effetti casuali (si veda, ad esempio, Diggle *et al.*, 2002), nei quali si ipotizza che vi siano delle caratteristiche non osservabili comuni a tutte le osservazioni relative alla stessa unità statistica. Le caratteristiche comuni sono descritte come realizzazioni di variabili casuali, dette effetti casuali. La presenza degli effetti casuali nel modello comporta che vi sia correlazione fra le osservazioni relative alla stessa unità statistica, e quindi consente di descrivere la non indipendenza fra le misurazioni.

I modelli con effetti casuali consentono facilmente anche di trattare dati con struttura multilivello, nei quali le misurazioni al livello $k$ sono trattate come indipendenti condizionatamente alle osservazioni al livello $k+1$. Per approfondimenti si rimanda, tra gli altri, a Song (2007).

Indicato con $\boldsymbol{Y}$ il vettore $N$-dimensionale di tutte le misurazioni di tutte

le unità statistiche, il modello può essere espresso in forma matriciale come

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \sum_{j=1}^{r} \boldsymbol{Z}_j \boldsymbol{\gamma}_j + \boldsymbol{\varepsilon}, \tag{1.1}$$

dove $\boldsymbol{X}$ è la matrice di disegno degli effetti fissi di dimensione $N \times p$, $\boldsymbol{\beta} = (\beta_1 , \dots , \beta_p)^\mathsf{T}$ è il vettore $p$-dimensionale degli effetti fissi, $\boldsymbol{Z}_j$ sono le matrici di disegno degli effetti casuali, ciascuna di dimensione $N \times n$, $\boldsymbol{\gamma}_j = (\gamma_{j1} , \dots , \gamma_{jn})^\mathsf{T}$ sono i vettori aleatori $n$-dimensionali contenenti le variabili casuali per ogni unità statistica relative al $j$-esimo effetto casuale, $j = 1, \dots , r$, e $\boldsymbol{\varepsilon} = (\varepsilon_1 , \dots , \varepsilon_N)^\mathsf{T}$ è il vettore aleatorio $N$-dimensionale per il termine di errore. Il vettore degli effetti fissi $\boldsymbol{\beta}$ rappresenta la relazione media fra la risposta e l'insieme di variabili esplicative a livello di popolazione. Le sue componenti possono essere associate sia a variabili esplicative che dipendono unicamente dall'unità $i$-esima, e vengono dette effetti fissi fra le unità, sia a variabili esplicative che dipendono dalla specifica misurazione sulla medesima unità, e vengono dette effetti fissi entro le unità. L'eterogeneità fra le unità statistiche presenti nel campione, viene spiegata dall'inclusione degli effetti casuali che rappresentano, quindi, la variazione individuale rispetto agli effetti fissi. Risulta dunque ragionevole ipotizzare che gli effetti casuali siano vettori aleatori a media nulla. Definendo $\boldsymbol{Z} = (\boldsymbol{Z}_1 , \dots , \boldsymbol{Z}_r)$ la matrice di disegno per tutti gli effetti casuali di dimensione $N \times nr$, ottenuta affiancando tutte le matrici $\boldsymbol{Z}_j$, e $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\mathsf{T} , \dots , \boldsymbol{\gamma}_r^\mathsf{T})^\mathsf{T}$ il vettore $nr$-dimensionale ottenuto impilando tutti i vettori $\boldsymbol{\gamma}_j$ di ogni effetto casuale, il modello (1.1) è scrivibile in forma compatta come

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}. \tag{1.2}$$

Nella formulazione (1.2) è chiaro come i modelli con effetti casuali estendano i modelli di regressione lineari normali introducendo nel predittore lineare la componente $\boldsymbol{Z}\boldsymbol{\gamma}$, rendendo quindi la risposta una somma di effetti fissi e

di effetti casuali. Per questo motivo, questi modelli vengono spesso indicati anche con il termine modelli lineari con effetti misti (MLM, *mixed linear model*).

La specificazione più semplice del modello prevede $\boldsymbol{\gamma}_j \sim \mathcal{N}_n(\mathbf{0}, \sigma_j^2 \boldsymbol{I}_n)$, $j = 1, \ldots, r$, e $\boldsymbol{\varepsilon} \sim \mathcal{N}_N(\mathbf{0}, \sigma_\varepsilon^2 \boldsymbol{I}_N)$, con $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_r, \boldsymbol{\varepsilon}$ indipendenti. Tale formulazione assume quindi l'indipendenza tra effetti casuali ed errori casuali, nonché l'indipendenza fra effetti casuali sia relativi a diverse unità sia relativi alla medesima unità e fra errori casuali sia relativi a diverse unità sia relativi alla medesima unità. Queste ultime assunzioni di indipendenza sono piuttosto restrittive e, in alcuni casi, anche discutibili, in particolare con dati longitudinali. Una formulazione più generale del modello, che può mantenere per esempio l'indipendenza fra le unità statistiche ma permettere la correlazione fra effetti casuali e fra errori casuali relativi alla stessa unità, prevede $\boldsymbol{\gamma} \sim \mathcal{N}_{nr}(\mathbf{0}, \boldsymbol{D})$ e $\boldsymbol{\varepsilon} \sim \mathcal{N}_N(\mathbf{0}, \boldsymbol{R})$, con $\boldsymbol{D}$ e $\boldsymbol{R}$ matrici opportune. In questa relazione è trattata principalmente la formulazione più semplice del modello ma i concetti presentati sono estendibili anche a formulazioni meno restrittive.

Indicato con $Y_i = (Y_{i1}, \ldots, Y_{im_i})^\mathsf{T}$ il vettore casuale contenente le $m_i$ misurazioni dell'$i$-esima unità statistica, nella sua formulazione più semplice il modello prevede $\boldsymbol{Y}_i \sim \mathcal{N}_{m_i}(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$ indipendenti, con $\boldsymbol{X}_i$ matrice di disegno per l'$i$-esima unità statistica di dimensione $m_i \times p$ e corrispondente al blocco $i$-esimo della matrice $\boldsymbol{X}$ e $\boldsymbol{\Sigma}_i$ matrice di varianze e covarianze di dimensione $m_i \times m_i$, $i = 1, \ldots, n$. Quest'ultima è esprimibile come

$$\boldsymbol{\Sigma}_i = \sigma_\varepsilon^2 \boldsymbol{I}_{m_i} + \sum_{j=1}^{r} \sigma_j^2 [\boldsymbol{Z}_j \boldsymbol{Z}_j^\mathsf{T}]_{(ii)}, \tag{1.3}$$

dove $[\boldsymbol{Z}_j \boldsymbol{Z}_j^\mathsf{T}]_{(ii)}$ indica il blocco diagonale $i$-esimo della matrice $\boldsymbol{Z}_j \boldsymbol{Z}_j^\mathsf{T}$, $i = 1, \ldots, n$. Nella (1.3) si può vedere come la varianza del vettore delle misurazioni sull'$i$-esima unità statistica sia data dalla somma della variabilità

entro le unità, primo addendo, e quella tra le unità, secondo addendo. La matrice di varianze e covarianze per l'intero vettore $\boldsymbol{Y}$ è definita da

$$\boldsymbol{V} = \sigma^2_\varepsilon \boldsymbol{I}_N + \sum_{j=1}^{r} \sigma^2_j \boldsymbol{Z}_j \boldsymbol{Z}_j^\mathsf{T}$$

ed è una matrice $N \times N$ diagonale a blocchi per l'assunto di indipendenza fra le unità, avente sulla diagonale le matrici $\boldsymbol{\Sigma}_i$, $i = 1, \ldots, n$. Il vettore ignoto dei parametri di regressione sotto questi assunti è il vettore $\boldsymbol{\theta} = (\boldsymbol{\beta}^\mathsf{T}, \boldsymbol{\alpha}^\mathsf{T})^\mathsf{T}$ di dimensione $p + 1 + r$, dove $\boldsymbol{\beta}$ è il vettore degli effetti fissi e $\boldsymbol{\alpha} = (\sigma^2_\varepsilon, \sigma^2_1, \ldots, \sigma^2_r)^\mathsf{T}$ è il vettore contenente i parametri relativi alla varianza dell'errore casuale $\boldsymbol{\varepsilon}$ e degli effetti casuali $\boldsymbol{\gamma}_j, j = 1, \ldots, r$.

Nel seguito si assumerà che il disegno sperimentale sia bilanciato, ovvero che il numero di osservazioni, indicato con $m$, sia costante per ciascuna unità statistica. Con disegni bilanciati la dimensione del vettore $\boldsymbol{Y}$ è data semplicemente dal prodotto $mn$, dove $n$ è il numero di unità statistiche.

## 1.2 Procedure classiche di stima

### 1.2.1 Stima di massima verosimiglianza

Le procedure classiche per la stima del vettore dei parametri $\boldsymbol{\theta} = (\boldsymbol{\beta}^\mathsf{T}, \boldsymbol{\alpha}^\mathsf{T})^\mathsf{T}$ si basano sulla massimizzazione della funzione di verosimiglianza per le $n$ osservazioni $\boldsymbol{y}_i$, realizzazioni delle variabili casuali $\boldsymbol{Y}_i \sim \mathcal{N}_m(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}_i)$, $i = 1, \ldots, n$. La funzione di verosimiglianza (Agresti, 2015, Cap. 9) per $\boldsymbol{\theta}$ risulta

$$L(\boldsymbol{\theta}; \boldsymbol{y}) \propto \prod_{i=1}^{n} |\boldsymbol{\Sigma}_i|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^\mathsf{T} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \right\}, \tag{1.4}$$

dove con $|\boldsymbol{\Sigma}_i|$ viene indicato il determinante della matrice $\boldsymbol{\Sigma}_i$, $i = 1, \ldots, n$. La stima di massima verosimiglianza (MLE, *maximum likelihood estimation*), indicata con $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^\mathsf{T}, \hat{\boldsymbol{\alpha}}^\mathsf{T})^\mathsf{T}$, si ottiene massimizzando la (1.4) o,

equivalentemente, risolvendo in $\boldsymbol{\beta}$ l'equazione

$$\sum_{i=1}^{n} \boldsymbol{X}_i^{\mathsf{T}} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}) = \mathbf{0} \tag{1.5}$$

per il vettore degli effetti fissi e in $\sigma_j^2$ l'equazione

$$\sum_{i=1}^{n} \{ (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta})^{\mathsf{T}} \boldsymbol{\Sigma}_i^{-1} [\boldsymbol{Z}_j \boldsymbol{Z}_j^{\mathsf{T}}]_{(ii)} \boldsymbol{\Sigma}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}) - \operatorname{tr}(\boldsymbol{\Sigma}_i^{-1} [\boldsymbol{Z}_j \boldsymbol{Z}_j^{\mathsf{T}}]_{(ii)}) \} = 0 \tag{1.6}$$

per i parametri del vettore $\boldsymbol{\alpha}$, $j = 0, \ldots, r$. Nella (1.6) per convenienza di scrittura si è posto $\boldsymbol{Z}_0 = \boldsymbol{I}_N$, $\boldsymbol{\gamma}_0 = \boldsymbol{\varepsilon}$ e $\sigma_0^2 = \sigma_\varepsilon^2$.

Per risolvere il sistema di equazioni, si assume inizialmente che le varianze $\sigma_j^2$ siano fissate e, a partire dalla (1.5), con semplici passaggi si ottiene la soluzione di $\boldsymbol{\beta}$, in funzione dei parametri relativi alle varianze, data da

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}} &= \left( \sum_{i=1}^{n} \boldsymbol{X}_i^{\mathsf{T}} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i \right)^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i^{\mathsf{T}} \boldsymbol{\Sigma}_i^{-1} \boldsymbol{y}_i \\ &= (\boldsymbol{X}^{\mathsf{T}} \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^{\mathsf{T}} \boldsymbol{V}^{-1} \boldsymbol{y}. \end{aligned} \tag{1.7}$$

Si nota che se la matrice di varianze e covarianze $\boldsymbol{V}$ fosse nota o sostituita con una sua stima, dalla (1.7) si otterrebbe lo stimatore di massima verosimiglianza per la componente degli effetti fissi $\hat{\boldsymbol{\beta}}$, che coincide con lo stimatore ai minimi quadrati generalizzati. Nel caso in cui le varianze non siano note, per la stima della componente $\boldsymbol{\alpha}$ si risolvono le (1.6) sostituendo a $\boldsymbol{\beta}$ l'espressione della (1.7) in modo da ottenere un sistema di $r+1$ equazioni che dipendono solo da $\sigma_j^2$, $j = 0, \ldots, r$. Il risultato del sistema di equazioni fornisce la stima di massima verosimiglianza $\hat{\boldsymbol{\alpha}}$. Ottenuta questa, per la proprietà di equivarianza dello stimatore di massima verosimiglianza, si ottengono le stime delle matrici $\boldsymbol{\Sigma}_i$ e quindi di $\boldsymbol{V}$, indicate rispettivamente con $\hat{\boldsymbol{\Sigma}}_i(\hat{\boldsymbol{\alpha}})$ e $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\alpha}})$. Lo stimatore $\hat{\boldsymbol{\beta}}$, ottenuto dalla (1.7) calcolando $\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\alpha}}}$, ha asintoticamente distribuzione normale, è non distorto e per il teorema di Gauss-Markov

nella classe degli stimatori lineari non distorti è quello a varianza minima. La matrice di varianze e covarianze dello stimatore di massima verosimiglianza degli effetti fissi è pari a

$$Var(\hat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{n} \boldsymbol{X}_i^\top \boldsymbol{\Sigma}_i^{-1} \boldsymbol{X}_i\right)^{-1} = (\boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{X})^{-1}, \tag{1.8}$$

che può essere stimata sostituendo alle matrici di varianza le rispettive stime. Sotto condizioni regolari si può mostrare che $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\alpha}}$ sono asintoticamente incorrelati.

### 1.2.2 Stima di massima verosimiglianza ristretta

Sebbene lo stimatore di massima verosimiglianza sia asintoticamente consistente ed efficiente sotto gli assunti di normalità, gli stimatori delle varianze $\sigma_j^2$ sono solo asintoticamente non distorti (Agresti, 2015, Cap. 9). In piccoli campioni la distorsione può diventare elevata, in particolare quando la dimensione del vettore $\beta$ aumenta. Così come nel modello lineare normale, allo stimatore di massima verosimiglianza della varianza dell'errore si preferisce lo stimatore corretto per i gradi di libertà, in questo contesto allo stimatore di massima verosimiglianza si preferisce lo stimatore corretto basato su una verosimiglianza marginale (Diggle *et al.*, 2002, Cap. 4). Si tratta della verosimiglianza ottenuta a partire dal modello statistico per una trasformazione $\boldsymbol{LY}$ di $\boldsymbol{Y}$, con densità non dipendente da $\boldsymbol{\beta}$, dove $\boldsymbol{L}$ è una matrice di dimensione $(N-p) \times N$, di rango $N-p$. La trasformazione $\boldsymbol{LY}$ fornisce i residui linearmente indipendenti della regressione lineare di $\boldsymbol{Y}$ su $\boldsymbol{X}$. Per questo motivo, il metodo è detto metodo della massima verosimiglianza ristretta (REML, *restricted maximum likelihood estimation*). Per qualsiasi scelta di $\boldsymbol{L}$, la log-verosimiglianza marginale per $\boldsymbol{\alpha}$ corrispondente è

$$\ell_R(\boldsymbol{\alpha}; \boldsymbol{y}) \propto -\frac{1}{2}\left\{(\boldsymbol{Ly})^\top \boldsymbol{V}^{-1}(\boldsymbol{Ly}) + \log|\boldsymbol{V}| + \log|\boldsymbol{X}^\top \boldsymbol{V}^{-1} \boldsymbol{X}|\right\}. \tag{1.9}$$

La stima REML del vettore $\boldsymbol{\alpha}$ può essere ottenuta massimizzando la (1.9) o, equivalentemente, trovando la soluzione che annulla il vettore delle derivate parziali prime. Il sistema dato dalle equazioni $\partial \ell_R(\boldsymbol{\alpha}; \boldsymbol{y})/\partial \sigma_j^2 = 0$, $j = 0, \ldots, r$, può essere riscritto nella forma compatta

$$(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top \boldsymbol{V}^{-1} \boldsymbol{Z}_j \boldsymbol{Z}_j^\top \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - \operatorname{tr}(\boldsymbol{P}\boldsymbol{Z}_j \boldsymbol{Z}_j^\top) = 0, \tag{1.10}$$

per $j = 0, \ldots, r$, dove $\boldsymbol{P} = \boldsymbol{V}^{-1} - \boldsymbol{V}^{-1}\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{V}^{-1}$. La (1.10) è l'equivalente della (1.6) e permette di ottenere le stime non distorte $\hat{\boldsymbol{\alpha}}_{\mathrm{R}}$. Le stime REML per gli effetti fissi, indicate con $\hat{\boldsymbol{\beta}}_{\mathrm{R}}$, si ottengono dalla (1.7) calcolando $\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\alpha}}_{\mathrm{R}}}$. Gli stimatori $\hat{\boldsymbol{\beta}}_{\mathrm{R}}$ e $\hat{\boldsymbol{\alpha}}_{\mathrm{R}}$ sono asintoticamente normali e incorrelati e la varianza asintotica è data per entrambi dall'inversa della matrice di informazione di Fisher. In particolare, la varianza dello stimatore $\hat{\boldsymbol{\beta}}_{\mathrm{R}}$ segue la stessa struttura di quella dello stimatore di massima verosimiglianza, data nella (1.8), in cui la matrice di varianze e covarianze $\boldsymbol{\Sigma}_i$ viene stimata con $\hat{\boldsymbol{\Sigma}}_i(\hat{\boldsymbol{\alpha}}_{\mathrm{R}})$ anziché con $\hat{\boldsymbol{\Sigma}}_i(\hat{\boldsymbol{\alpha}})$.

## 1.3 Procedure classiche per la verifica d'ipotesi

Ottenuti gli stimatori per gli effetti fissi e per le varianze degli effetti casuali, è di interesse il problema della verifica di ipotesi, utile in questo contesto, ad esempio, per testare la significatività dei parametri del modello. L'interesse primario è tipicamente legato agli effetti fissi. La teoria propone tre test (Agresti, 2015) i quali, sotto l'ipotesi nulla, sono asintoticamente equivalenti: la statistica test di Wald, la statistica test *score* e la statistica test log-rapporto di verosimiglianza (LRT, *likelihood ratio test*). Nonostante sia possibile calcolarlo, il test *score* è poco utilizzato per questo tipo di modelli e quindi l'attenzione viene focalizzata sugli altri due.

Nella sua formulazione più generale, il test di Wald può essere utilizzato sia per verifiche d'ipotesi sui singoli coefficienti sia per problemi d'inferenza

globale e parziale utilizzando la sua versione quadratica. Indicando con $H_0$ l'ipotesi nulla e con $H_1$ l'ipotesi alternativa, il sistema di verifica d'ipotesi è scrivibile in generale come

$$\begin{cases} H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{\beta}_0 \\ H_1 : \overline{H_0} \end{cases},$$

dove $\boldsymbol{C}$ è una matrice $p \times p$ che dipende dall'ipotesi che si vuole verificare e $\boldsymbol{\beta}_0$ è il vettore contenente i valori fissati dall'ipotesi nulla, ad esempio $\boldsymbol{\beta}_0 = \mathbf{0}$. La distribuzione nulla approssimata del test è un chi quadrato con gradi di libertà pari al rango della matrice $\boldsymbol{C}$. La statistica test è

$$W_e = (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^{\mathsf{T}}[\boldsymbol{C}\,\hat{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{C}^{\mathsf{T}}]^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \tag{1.11}$$

e il suo valore osservato va confrontato con il quantile di livello $(1-\alpha)$ della distribuzione $\chi^2_{\text{rank}(\boldsymbol{C})}$ per condurre la verifica d'ipotesi al livello di significatività $\alpha$, tipicamente fissato al 5%.

Il test di Wald si può condurre sia utilizzando le stime MLE sia utilizzando le stime REML; dunque, nella (1.11) $\hat{\boldsymbol{\beta}}$ può essere sostituito con $\hat{\boldsymbol{\beta}}_{\text{R}}$. Per calcolare la varianza stimata dello stimatore $\hat{\boldsymbol{\beta}}$ si utilizzano le rispettive stime del vettore $\boldsymbol{\alpha}$. Si osserva che con questa formulazione generica del test di Wald, per verificare la significatività di un singolo parametro, ad esempio il primo senza perdita di generalità ($H_0 : \beta_1 = 0$), è sufficiente specificare

$$\boldsymbol{C} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \quad \text{e} \quad \boldsymbol{\beta}_0 = \mathbf{0}\,,$$

e in questo modo

$$W_e = \begin{pmatrix} \hat{\beta}_1 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} \widehat{Var(\hat{\beta}_1)} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}^{-1} \begin{pmatrix} \hat{\beta}_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \frac{\hat{\beta}_1^2}{\widehat{Var(\hat{\beta}_1)}}$$

va confrontato con il quantile della distribuzione $\chi^2_1$. Il valore osservato della statistica test corrisponde al quadrato di $z = \hat{\beta}_1 / SE(\hat{\beta}_1)$, dove $SE(\hat{\beta}_1)$ è la stima dello *standard error* di $\hat{\beta}_1$, ossia $SE(\hat{\beta}_1) = \sqrt{\widehat{Var(\hat{\beta}_1)}}$. Il test $z$ ha distribuzione asintotica nulla $\mathcal{N}(0, 1)$. Sebbene con verifiche d'ipotesi bilaterali i test siano equivalenti, nel caso in cui la verifica d'ipotesi riguardi un solo parametro o una sola combinazione lineare dei $\boldsymbol{\beta}$, utilizzare la statistica test $z$ permette di condurre anche test unilaterali.

Il test di Wald non tiene conto dell'incertezza delle stime del vettore $\boldsymbol{\alpha}$ utilizzate per stimare la varianza del vettore $\hat{\boldsymbol{\beta}}$. Questo porta ad avere tipicamente stime degli *standard error* degli effetti fissi troppo piccole. Un modo per risolvere questo problema è utilizzare al posto del test di Wald un opportuno test $F$, dato da (Heritier *et al.*, 2009, Cap. 4)

$$F = \frac{W_e}{\text{rank}(\boldsymbol{C})} \ , \tag{1.12}$$

con distribuzione nulla approssimata $F$ con al numeratore $\text{rank}(\boldsymbol{C})$ gradi di libertà e al denominatore $\nu$ gradi di libertà, con $\nu$ calcolato in base ai dati secondo la formula di Satterthwaite.

Il test log-rapporto di verosimiglianza, invece, è definito unicamente per le stime ottenute con il metodo della massima verosimiglianza e non per $\hat{\boldsymbol{\beta}}_{\text{R}}$. Questo è dovuto al fatto che le due verosimiglianze ristrette sotto $H_0$ e sotto $H_1$ non sono comparabili. In questo caso, è più conveniente scrivere il

sistema di verifica d'ipotesi nel seguente modo

$$\begin{cases} H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \\ H_1 : \overline{H_0} \end{cases},$$

dove $\boldsymbol{\beta}_0$ può essere immaginato suddiviso in due blocchi $(\boldsymbol{\beta}_{0(1)}^{\mathsf{T}} \; \boldsymbol{\beta}_{0(2)}^{\mathsf{T}})^{\mathsf{T}}$, con $\boldsymbol{\beta}_{0(1)}$ vettore fissato, ad esempio pari a $\mathbf{0}$, di dimensione $k$, $k \leq p$, e $\boldsymbol{\beta}_{0(2)}$ vettore non specificato di dimensione $p - k$. Il test log-rapporto di verosimiglianza può essere scritto come differenza delle log-verosimiglianze nel seguente modo

$$\text{LRT} = 2[\ell(\hat{\boldsymbol{\theta}}; y) - \ell(\tilde{\boldsymbol{\theta}}; y)]\,, \tag{1.13}$$

dove con $\tilde{\boldsymbol{\theta}}$ si è indicata la stima di massima verosimiglianza sotto $H_0$. Sotto l'ipotesi nulla, la statistica LRT ha distribuzione approssimata chi quadrato con $k$ gradi di libertà, pari al numero di vincoli imposti da $H_0$.

In un modello con effetti misti si può essere interessati anche a verificare la significatività delle stime dei parametri relativi alla varianza degli effetti casuali, ad esempio per scegliere fra un modello con intercetta casuale ed uno con intercetta e tendenza casuali. In generale, sovraparametrizzare la struttura di varianza comporta un'inefficienza anche nella stima degli effetti fissi e dei loro *standard error*. Risulta quindi opportuno introdurre una statistica test anche per la verifica della significatività delle componenti del vettore $\boldsymbol{\alpha}$. Come detto precedentemente, sia $\hat{\boldsymbol{\alpha}}$ che $\hat{\boldsymbol{\alpha}}_{\text{R}}$ hanno distribuzione asintotica normale con matrice di varianze e covarianze asintotica data dall'inversa della matrice d'informazione di Fisher. Tuttavia, l'ipotesi di interesse per la semplificazione del modello $H_0 : \sigma_j^2 = 0$, si trova sul confine dello spazio parametrico. Questo non rispetta le condizioni di regolarità richieste dalla distribuzione asintotica degli stimatori e quindi, sotto l'ipotesi nulla, $W_e$ e LRT non hanno più distribuzione asintotica chi quadrato. Quando il numero degli effetti fissi rimane costante e sotto l'assunto $Var(\boldsymbol{\varepsilon}) = \sigma_\varepsilon^2 \boldsymbol{I}_N$,

si può dimostrare che il LRT è distribuito come una mistura di chi quadrato. In particolare, se si vuole confrontare un modello con $q$ effetti casuali contro un modello con $q+1$ effetti casuali, la distribuzione nulla asintotica del LRT è data dalla mistura con pesi 0.5 e 0.5 di un $\chi^2_q$ e un $\chi^2_{q+1}$. Misture più complesse sono disponibili se si vuole testare contemporaneamente la significatività di più effetti casuali e quindi annullare contemporaneamente più elementi del vettore $\boldsymbol{\alpha}$. Il test log-rapporto di verosimiglianza per i parametri relativi alle varianze degli effetti casuali con questa distribuzione nulla può essere condotto utilizzando sia $\hat{\boldsymbol{\alpha}}$ che $\hat{\boldsymbol{\alpha}}_{\mathrm{R}}$.

## 1.4 Limiti delle procedure classiche

In ambito parametrico le procedure classiche risultano molto, ed a volte estremamente, sensibili a piccoli scostamenti dalla distribuzione dei dati dal modello assunto. Il modello potrebbe non rispecchiare esattamente la realtà, o per la presenza di dati anomali nel campione osservato o per il carattere approssimato del modello teorico stesso. Nonostante gli stimatori di massima verosimiglianza siano gli stimatori più efficienti all'interno della classe degli stimatori lineari non distorti, essi non sono robusti rispetto alla contaminazione o rispetto ad un'errata specificazione del modello (Huber e Ronchetti, 2009).

Per quantificare in maniera più formale la sensibilità degli stimatori presentati al Paragrafo 1.2, nel prossimo capitolo viene introdotto il concetto di funzione di influenza. Si vedrà che gli stimatori $\hat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{\beta}}_{\mathrm{R}}$ non sono robusti, essendo basati su equazioni di stima non limitate. Anche gli stimatori $\hat{\boldsymbol{\alpha}}$ e $\hat{\boldsymbol{\alpha}}_{\mathrm{R}}$ non sono robusti; essi sono definiti infatti a partire dalla (1.6) e dalla (1.10), rispettivamente, funzioni non limitate che dipendono dai dati in modo quadratico.

# Capitolo 2

# Robustezza

## 2.1 Funzione di influenza

Una procedura statistica è detta robusta se è poco sensibile a piccoli scostamenti dal modello ipotizzato. I metodi robusti si collocano fra le procedure parametriche classiche, che fissano un modello parametrico $\mathcal{F}_{\boldsymbol{\theta}}$ assunto come vero processo generatore dei dati, e le procedure non parametriche, che non fissano alcun modello. Essi rappresentano dunque un compromesso fra efficienza e poca sensibilità a valori anomali, muovendosi in un intorno del modello $\mathcal{F}_{\boldsymbol{\theta}}$, considerato una ragionevole approssimazione del processo generatore. La notazione di intorno è formalizzabile come (si veda, ad esempio, Huber e Ronchetti, 2009, e i riferimenti qui citati)

$$\mathcal{F}_{\varepsilon} = (1 - \varepsilon)\mathcal{F}_{\boldsymbol{\theta}} + \varepsilon G, \tag{2.1}$$

dove $G$ è una distribuzione arbitraria e $0 \leq \varepsilon \leq 1$. In altre parole, i dati generati da $\mathcal{F}_{\varepsilon}$ provengono da $\mathcal{F}_{\boldsymbol{\theta}}$ con probabilità $(1-\varepsilon)$ e da $G$ con probabilità $\varepsilon$. Il valore assunto da $\varepsilon$ è una misura dell'errata specificazione del modello parametrico $\mathcal{F}_{\boldsymbol{\theta}}$ rispetto al vero processo generatore.

I metodi robusti possono essere considerati la scelta migliore quando vi

sono leggere controindicazioni rispetto agli assunti del modello parametrico in quanto (Farcomeni e Ventura, 2012):

1. sono più potenti dei metodi non parametrici basati sui ranghi;

2. permettono di essere calibrati affinché abbiano una piccola perdita di efficienza rispetto ai metodi parametrici classici, guadagnando per costruzione in robustezza rispetto a piccoli allontanamenti dal modello ipotizzato.

Un approccio generale allo studio della robustezza (Hampel *et al.*, 1986) è basato sulla funzione di influenza (IF, *influence function*). La funzione di influenza per $\hat{\boldsymbol{\theta}}$ a $\mathcal{F}_{\boldsymbol{\theta}}$ nel punto $x$ è definita come

$$IF(x; \hat{\boldsymbol{\theta}}, \mathcal{F}_{\boldsymbol{\theta}}) = \lim_{\varepsilon \to 0} \frac{\hat{\boldsymbol{\theta}}(\mathcal{F}_\varepsilon) - \hat{\boldsymbol{\theta}}(\mathcal{F}_{\boldsymbol{\theta}})}{\varepsilon}, \tag{2.2}$$

dove $\mathcal{F}_\varepsilon$ è dato nella (2.1) con $G$ uguale alla funzione di ripartizione di una variabile casuale degenere in $x$. La funzione di influenza misura la stabilità locale di una procedura statistica: rappresenta infatti l'effetto prodotto sullo stimatore $\hat{\boldsymbol{\theta}}(\mathcal{F}_\theta)$ da una contaminazione infinitesimale nel punto $x$, standardizzato per la massa $\varepsilon$ della contaminazione.

La richiesta di robustezza si traduce in opportune condizioni di limitatezza sulla funzione di influenza. Le principali misure di robustezza sono:

- Sensibilità rispetto ai grandi errori (GES, *gross error sensitivity*) di $\hat{\boldsymbol{\theta}}$ in $\mathcal{F}_{\boldsymbol{\theta}}$:

  $$\mathrm{GES}(\hat{\boldsymbol{\theta}}, \mathcal{F}_{\boldsymbol{\theta}}) = \sup_x \|IF(x; \hat{\boldsymbol{\theta}}\mathcal{F}_{\boldsymbol{\theta}})\|. \tag{2.3}$$

  Essa è la più importante misura di robustezza derivante dalla funzione di influenza e misura la peggiore influenza che può avere una contaminazione infinitesimale, ovvero la massima distorsione possibile dello stimatore. Se tale quantità è finita lo stimatore viene detto B-robusto.

- Sensibilità rispetto a fluttuazioni locali di $\hat{\boldsymbol{\theta}}$ in $\mathcal{F}_{\boldsymbol{\theta}}$:

$$\lambda^* = \sup_{x \neq y} \frac{\|IF(y; \hat{\boldsymbol{\theta}}, \mathcal{F}_{\boldsymbol{\theta}}) - IF(x; \hat{\boldsymbol{\theta}}, \mathcal{F}_{\boldsymbol{\theta}})\|}{y - x}. \tag{2.4}$$

  Essa misura l'effetto che si ottiene togliendo $x$ ed introducendo $y$ come nel caso di arrotondamento delle osservazioni. Se il valore di $\lambda^*$ è finito si ha la robustezza dello stimatore rispetto ad errori di arrotondamento.

- Punto di rifiuto:

$$\rho^* = \inf\{r > 0 \mid IF(x; \hat{\boldsymbol{\theta}}, \mathcal{F}_{\boldsymbol{\theta}}) = 0 \text{ per } d(x) > r\}, \tag{2.5}$$

  dove $d(\cdot)$ è un'opportuna misura di distanza. Si pone $\rho^* = \infty$ se non esiste un tale $r$. Tale misura rappresenta il punto di rigetto dei valori anomali da parte di $\hat{\boldsymbol{\theta}}$. Se tale valore esiste finito la contaminazione provocata da punti $x$ con $d(x) > \rho^*$ non esercita alcuna influenza sul valore dello stimatore.

- Punto di rottura:

$$\varepsilon^*(\hat{\boldsymbol{\theta}}, \mathcal{F}_{\boldsymbol{\theta}}) = \inf\{\varepsilon \mid \text{bias}(\hat{\boldsymbol{\theta}}, \mathcal{F}_{\boldsymbol{\theta}}, \varepsilon) = \infty\}, \tag{2.6}$$

  con $\text{bias}(\hat{\boldsymbol{\theta}}, \mathcal{F}_{\boldsymbol{\theta}}, \varepsilon) = \sup_G \|\hat{\boldsymbol{\theta}}(\mathcal{F}_\varepsilon) - \hat{\boldsymbol{\theta}}(\mathcal{F}_{\boldsymbol{\theta}})\|$. A differenza delle altre, questa misura non è collegata alla funzione di influenza ma riveste un ruolo molto importante. Essa rappresenta la percentuale massima di osservazioni anomale che lo stimatore può sopportare rimanendo affidabile. Una conseguenza è che se la sensibilità rispetto ai grandi errori (2.3) di $\hat{\boldsymbol{\theta}}$ non è finita, allora il suo punto di rottura è nullo. Uno stimatore robusto con funzione di influenza limitata, se non ha un punto di rottura sufficientemente elevato nella pratica può diventare inutile.

Gli stimatori da utilizzare dovrebbero avere buone proprietà di robustezza sia rispetto a contaminazioni infinitesimali sia in senso globale. La robustezza rispetto a contaminazioni infinitesimali è garantita qualora lo stimatore abbia una funzione di influenza limitata, mentre la robustezza in senso globale richiede che lo stimatore sia stabile in presenza di grandi allontanamenti dal modello ipotizzato ed è garantita qualora lo stimatore abbia un punto di rottura alto.

## 2.2 Classe generale degli stimatori di tipo M

Prima di trattare le due classi di stimatori utilizzate nei modelli lineari normali con effetti casuali (presentati nel Paragrafo 2.3), si forniscono alcune nozioni generali che saranno utili per evidenziare le differenze fra i vari stimatori e capire meglio i risultati successivi.

Una classe molto generale di stimatori è la classe degli stimatori di tipo M (si veda, ad esempio, Heritier *et al.*, 2009, Cap. 2), definiti come la soluzione $\hat{\boldsymbol{\theta}}_{\mathrm{M}}$ per il problema di minimo, espresso da

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \rho(\boldsymbol{y}_i; \boldsymbol{\theta}), \tag{2.7}$$

dove $\rho(\cdot; \boldsymbol{\theta})$ è una funzione opportuna. Questo equivale a trovare la soluzione in $\boldsymbol{\theta}$ delle equazioni di stima non distorte

$$\sum_{i=1}^{n} \boldsymbol{\Psi}(\boldsymbol{y}_i; \boldsymbol{\theta}) = 0, \tag{2.8}$$

dove $\boldsymbol{\Psi}(\cdot; \boldsymbol{\theta}) = \partial \rho(\cdot; \boldsymbol{\theta}) / \partial \boldsymbol{\theta}$. La condizione di non distorsione dell'equazione di stima prevede che $E_{\boldsymbol{\theta}}(\boldsymbol{\Psi}(\boldsymbol{Y}; \boldsymbol{\theta})) = 0$ e serve a garantire la consistenza dello stimatore.

Gli stimatori di tipo M (*maximum likelihood type*) sono una generalizzazione degli stimatori di massima verosimiglianza che sono ottenuti ponendo come

scelta particolare per la funzione $\rho(\boldsymbol{y};\boldsymbol{\theta})$ la funzione di log-verosimiglianza $\ell(\boldsymbol{\theta};\boldsymbol{y})$, cambiata di segno. Infatti, passando al problema duale di quello della (2.7), la soluzione $\hat{\boldsymbol{\theta}}_{\mathrm{M}}$ con $\rho(\cdot;\boldsymbol{\theta}) = -\ell(\boldsymbol{\theta};\cdot)$ è proprio il valore di $\boldsymbol{\theta}$ che massimizza la (log-)verosimiglianza. Nel caso degli stimatori di massima verosimiglianza la funzione $\boldsymbol{\Psi}(\boldsymbol{y};\boldsymbol{\theta})$ è dunque la funzione *score*, indicata con $\ell_*(\boldsymbol{\theta};\boldsymbol{y})$.

Per uno stimatore di tipo M la funzione di influenza è esprimibile come (Heritier *et al.*, 2009, Cap. 2)

$$IF(x;\boldsymbol{\theta},\mathcal{F}_{\boldsymbol{\theta}}) = M(\boldsymbol{\theta})^{-1}\boldsymbol{\Psi}(x;\boldsymbol{\theta}), \tag{2.9}$$

con

$$M(\boldsymbol{\theta}) = -\int \frac{\partial}{\partial\boldsymbol{\theta}}\boldsymbol{\Psi}(x;\boldsymbol{\theta})d\mathcal{F}_{\boldsymbol{\theta}}(x). \tag{2.10}$$

Si evidenzia che la *IF* di uno stimatore di tipo M è proporzionale alla funzione $\boldsymbol{\Psi}(\cdot;\boldsymbol{\theta})$ che lo definisce. Da questo deriva che per avere stimatori di tipo M con una sensibilità rispetto ai grandi errori (2.3) finita, la funzione $\boldsymbol{\Psi}(\cdot;\boldsymbol{\theta})$ deve essere limitata. Essendo generalmente la funzione *score* non limitata, si deduce che gli stimatori di massima verosimiglianza non sono in generale B-robusti e quindi non c'è limite alla distorsione che una singola osservazione anomala può comportare su di essi.

In generale, la *IF* degli stimatori di massima verosimiglianza è illimitata poiché tali stimatori attribuiscono peso pari ad uno a tutte le osservazioni, anche a quelle molto distanti dal modello stimato. Infatti, un altro modo di esprimere la funzione $\boldsymbol{\Psi}(\cdot;\boldsymbol{\theta})$ nell'equazione di stima è (Heritier *et al.*, 2009, Cap. 2)

$$\boldsymbol{\Psi}(\boldsymbol{y};\boldsymbol{\theta}) = w(d(\boldsymbol{y};\boldsymbol{\theta});\boldsymbol{\theta})d(\boldsymbol{y};\boldsymbol{\theta}), \tag{2.11}$$

dove $d(\boldsymbol{y};\boldsymbol{\theta})$ è una misura di distanza fra l'osservazione $\boldsymbol{y}$ e il modello, e $w(\cdot;\boldsymbol{\theta})$ è la funzione dei pesi attribuiti a ciascuna osservazione sulla base della loro distanza.

## 2.3 Stimatori robusti per i modelli con effetti casuali

### 2.3.1 Stimatori di tipo S

La classe degli stimatori robusti di tipo S (Heritier *et al.*, 2009, Cap. 2 e 4) è molto utilizzata nel contesto dei modelli di regressione e costituisce un'alternativa alla classe degli stimatori di tipo M. Gli stimatori di tipo S risolvono il problema principale degli stimatori di tipo M, ovvero quello di avere un punto di rottura troppo basso nei casi in cui il vettore $\boldsymbol{\theta}$ abbia dimensione relativamente alta.

In generale, gli stimatori di tipo S sono definiti come la soluzione $\hat{\boldsymbol{\beta}}_{\mathrm{S}}$ che minimizza una funzione di dispersione sotto il vincolo

$$\frac{1}{n}\sum_{i=1}^{n}\rho(d_i(\boldsymbol{y}_i;\boldsymbol{\theta});\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[\rho(d(\boldsymbol{y};\boldsymbol{\theta});\boldsymbol{\theta})], \tag{2.12}$$

con $d_i(\boldsymbol{y}_i;\boldsymbol{\theta})$ misura della distanza dell'$i$-esima osservazione e $\rho(\cdot;\boldsymbol{\theta})$ funzione simmetrica e limitata. Il valore atteso a destra dell'uguaglianza garantisce la consistenza dello stimatore. Imporre che la funzione $\rho(\cdot;\boldsymbol{\theta})$ sia simmetrica e limitata, ad esempio costante per valori del suo argomento grandi in modulo, equivale a richiedere che la funzione $\boldsymbol{\Psi}(\cdot;\boldsymbol{\theta})$ sia simmetrica e si annulli per valori del suo argomento grandi in modulo. Questo garantisce allo stimatore di tipo S di avere un punto di rifiuto finito e un punto di rottura alto.

Il punto di rottura di uno stimatore di tipo S può essere calcolato dal rapporto

$$\varepsilon^* = \frac{E_{\boldsymbol{\theta}}[\rho(d(\boldsymbol{y};\boldsymbol{\theta});\boldsymbol{\theta})]}{\max_{\boldsymbol{y}}\rho(\boldsymbol{y};\boldsymbol{\theta})}. \tag{2.13}$$

Nel contesto dei modelli lineari normali con effetti casuali questi stimatori possono essere applicati al solo caso di disegni bilanciati, ovvero disegni in cui il numero di osservazioni per ogni unità statistica è costante, $m_i = m\ \forall i$, e la matrice di varianze e covarianze per ogni unità è anch'essa costante,

$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}\ \forall i$. Per questi modelli, la funzione di dispersione da minimizzare è il determinante della matrice $\boldsymbol{\Sigma}$, la distanza $d_i(\cdot;\boldsymbol{\theta})$ del vincolo espresso nella (2.12) per l'$i$-esima unità è la distanza di Mahalanobis

$$d_i(\boldsymbol{y}_i;\boldsymbol{\theta}) = \sqrt{(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})}\,, \tag{2.14}$$

e una scelta tipica per la funzione $\rho(\cdot;\boldsymbol{\theta})$ è la funzione di Tukey (*Tukey's biweight*), definita da

$$\rho_{[bi]}(d(\cdot;\boldsymbol{\theta});\boldsymbol{\theta},c) = \begin{cases} 3\left(\frac{d}{c}\right)^2 - 3\left(\frac{d}{c}\right)^4 + \left(\frac{d}{c}\right)^6 & \text{se } |d| \le c \\ 1 & \text{se } |d| > c. \end{cases} \tag{2.15}$$

La costante $c$ viene scelta in modo da ottenere uno specifico punto di rottura, per esempio del 50%, sfruttando la relazione data nella (2.13) fra $\varepsilon^*$ e la funzione $\rho(\cdot;\boldsymbol{\theta},c)$.

Il problema di minimizzare il determinante della matrice $\boldsymbol{\Sigma}$ può essere riformulato risolvendo l'equazione di stima non distorta

$$\sum_{i=1}^{n} \boldsymbol{\Psi}_\beta(\boldsymbol{y}_i, \boldsymbol{X}_i;\boldsymbol{\theta}) = \sum_{i=1}^{n} w(d_i(\boldsymbol{y}_i;\boldsymbol{\theta}))\boldsymbol{X}_i^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) = 0 \tag{2.16}$$

per i $\boldsymbol{\beta}$ e le equazioni di stima

$$\begin{aligned}\sum_{i=1}^{n} \boldsymbol{\Psi}_{\sigma_j^2}(\boldsymbol{y}_i, \boldsymbol{X}_i;\boldsymbol{\theta}) = \\ = \sum_{i=1}^{n}\{mw(d_i)(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^\mathsf{T}\boldsymbol{\Sigma}^{-1}\boldsymbol{z}_j\boldsymbol{z}_j^\mathsf{T}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) \\ - w(d_i)d_i^2\mathrm{tr}[\boldsymbol{\Sigma}^{-1}\boldsymbol{z}_j\boldsymbol{z}_j^\mathsf{T}]\} = 0\end{aligned} \tag{2.17}$$

per i $\sigma_j^2$, $j = 1,\ldots,r$, contenuti nel vettore $\boldsymbol{\alpha}$. Nella (2.17) si è indicato con $\boldsymbol{z}_j\boldsymbol{z}_j^\mathsf{T}$ il blocco $[\boldsymbol{Z}_j\boldsymbol{Z}_j^\mathsf{T}]_{(ii)}$, assunto uguale per tutte le unità statistiche.

Queste due equazioni possono essere scritte in forma compatta come

$$\sum_{i=1}^{n} \boldsymbol{\Psi}(\boldsymbol{y}_i, \boldsymbol{X}_i; \boldsymbol{\theta}) = 0, \tag{2.18}$$

dove $\boldsymbol{\Psi} = (\boldsymbol{\Psi}_{\beta}{}^{\mathsf{T}}, \boldsymbol{\Psi}_{\sigma_1^2}, \ldots, \boldsymbol{\Psi}_{\sigma_r^2})^{\mathsf{T}}$. Scegliendo nello specifico la funzione (2.15) di Tukey, il vettore che risolve l'equazione è lo stimatore robusto di tipo S indicato con $\hat{\boldsymbol{\theta}}_{[\mathrm{CBS}]} = (\hat{\boldsymbol{\beta}}^{\mathsf{T}}_{[\mathrm{CBS}]}, \hat{\boldsymbol{\alpha}}^{\mathsf{T}}_{[\mathrm{CBS}]})^{\mathsf{T}}$. Sotto deboli condizioni di regolarità, gli stimatori di tipo S sono consistenti e asintoticamente normali. Per la componente relativa agli effetti fissi $\hat{\boldsymbol{\beta}}_{\mathrm{S}}$, la varianza asintotica di uno stimatore di tipo S è data da

$$Var(\hat{\boldsymbol{\beta}}_{\mathrm{S}}) = \frac{e_1}{e_2^2}\Big(\sum_{i=1}^{n} \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{x}_i\Big)^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{x}_i \Big(\sum_{i=1}^{n} \boldsymbol{x}_i^{\mathsf{T}}\boldsymbol{x}_i\Big)^{-1}, \tag{2.19}$$

dove

$$e_1 = \frac{1}{m} E_{\Phi}[d^2 w(d)^2] \quad \text{e} \quad e_2 = E_{\Phi}\Big[w(d) + \frac{1}{m} d \frac{\partial}{\partial d} w(d)\Big].$$

Il difetto principale degli stimatori di tipo S è che tali stimatori non possono essere utilizzati per costruire una versione robusta del test log-rapporto di verosimiglianza. Come si vedrà nel Paragrafo 2.4, a causa del vincolo sotto il quale sono definiti, per tali stimatori il valore osservato della statistica test log-rapporto di verosimiglianza è sempre pari a zero (Heritier *et al.*, 2009, p. 106).

### 2.3.2 Stimatori di tipo MM

Per recuperare una procedura robusta basata sul test log-rapporto di verosimiglianza viene proposto, per la stima dei parametri relativi agli effetti fissi, uno stimatore appartenente alla classe degli stimatori di tipo MM (Heritier *et al.*, 2009, Cap. 4).

Gli stimatori di questa classe hanno punti di rottura alti anche in presenza

di valori leva*, buone proprietà in termini di efficienza e possono essere usati per costruire test robusti basati sul log-rapporto di verosimiglianza.

L'idea alla base della costruzione di tali stimatori è quella di dissociare la stima delle varianze degli effetti casuali dalla stima degli effetti fissi, dividendo la procedura in due fasi. Nella prima fase, si trovano le stime delle varianze degli effetti casuali utilizzando uno stimatore basato su una funzione $\rho_0(\cdot;\boldsymbol{\theta},c_0)$, scelta in modo da ottenere un punto di rottura alto. Nella seconda fase, si sceglie una seconda funzione $\rho_1(\cdot;\boldsymbol{\theta},c_1)$ per ottenere uno stimatore di tipo M per gli effetti fissi più efficiente. Nella prima fase in pratica una possibile scelta è utilizzare lo stimatore di tipo S descritto precedentemente, ottenendo la stima della matrice di varianze e covarianze indicata con $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}}_{[\mathrm{CBS}]})$. Si osserva che, per come è stato definito lo stimatore $\hat{\boldsymbol{\alpha}}_{[\mathrm{CBS}]}$, la funzione $\rho_0(\cdot;\boldsymbol{\theta},c_0)$ è la funzione di Tukey espressa nella (2.15). La seconda fase, invece, equivale a risolvere in $\boldsymbol{\beta}$ l'equazione di stima

$$\sum_{i=1}^{n}\boldsymbol{\Psi}_{\beta}(\boldsymbol{y}_i,\boldsymbol{X}_i;\boldsymbol{\beta})=\sum_{i=1}^{n}w_1(d_i(\boldsymbol{y}_i;\boldsymbol{\theta});\boldsymbol{\theta},c_1)\boldsymbol{X}_i^{\top}\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta})=0, \quad (2.20)$$

dove $w_1(\cdot;\boldsymbol{\theta},c_1)$ rappresenta la funzione dei pesi associata a $\rho_1(\cdot;\boldsymbol{\theta},c_1)$ e $\hat{\boldsymbol{\Sigma}}$ è la stima della matrice di varianze e covarianze ottenuta nella fase precedente, per esempio $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}}_{[\mathrm{CBS}]})$. Tale equazione di stima è analoga a quella fornita per gli effetti fissi di uno stimatore di tipo S nella (2.16), con l'unica differenza che qui l'ignota matrice di varianze e covarianze $\boldsymbol{\Sigma}$ viene sostituita dalla sua stima proprio perché la procedura di stima avviene in due fasi distinte. La soluzione della (2.20) è lo stimatore di tipo MM per gli effetti fissi indicato in generale con $\hat{\boldsymbol{\beta}}_{[\mathrm{MM}]}$.

Sotto deboli condizioni di regolarità, gli stimatori di tipo MM sono non

*Con valore leva si intende un'osservazione che ha un'anomalia nel valore della risposta ma non in quello delle covariate o viceversa. Queste osservazioni portano ad errori molto gravi nella stima del modello se non si utilizzano tecniche robuste adeguate.

distorti, hanno distribuzione asintotica normale e varianza asintotica pari a

$$Var(\hat{\boldsymbol{\beta}}_{[\mathrm{MM}]}) = \frac{1}{n} M^{-1} Q M^{-\mathsf{T}}, \tag{2.21}$$

dove

$$M = \int \boldsymbol{\Psi}(\boldsymbol{y};\boldsymbol{\theta})\ell_*(\boldsymbol{\theta};\boldsymbol{y})^{\mathsf{T}} d\mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{y}),$$
$$Q = \int \boldsymbol{\Psi}(\boldsymbol{y};\boldsymbol{\theta})\boldsymbol{\Psi}(\boldsymbol{y};\boldsymbol{\theta})^{\mathsf{T}} d\mathcal{F}_{\boldsymbol{\theta}}(\boldsymbol{y}).$$

La varianza dello stimatore può essere stimata, ad esempio, sostituendo a $M$ e $Q$ le rispettive versioni empiriche date da

$$M = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Psi}(\boldsymbol{y}_i;\boldsymbol{\theta})\ell_*(\boldsymbol{\theta};\boldsymbol{y}_i)^{\mathsf{T}} \quad \text{e} \quad Q = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\Psi}(\boldsymbol{y}_i;\boldsymbol{\theta})\boldsymbol{\Psi}(\boldsymbol{y}_i;\boldsymbol{\theta})^{\mathsf{T}}.$$

Due scelte comuni per la funzione $\rho_1(\cdot;\boldsymbol{\theta},c_1)$ sono la funzione $\rho$ di Huber, data da

$$\rho_{[Hub]}(d(\cdot;\boldsymbol{\theta});\boldsymbol{\theta},c) = \begin{cases} \frac{1}{2}d^2 & \text{se } |d| \le c \\ c|d| - \frac{1}{2}d^2 & \text{se } |d| > c \end{cases} \tag{2.22}$$

e la funzione $\rho$ di Tukey, utilizzata anche per lo stimatore di tipo S e definita nella (2.15). Queste danno luogo alle due corrispondenti funzioni $w_1(\cdot;\boldsymbol{\theta},c_1)$ contenenti i pesi assegnati a ciascuna osservazione, rispettivamente

$$w_{[Hub]}(d(\cdot;\boldsymbol{\theta});\boldsymbol{\theta},c) = \min\left(1, \frac{c}{|d|}\right) \tag{2.23}$$

per Huber e

$$w_{[bi]}(d(\cdot;\boldsymbol{\theta});\boldsymbol{\theta},c) = \begin{cases} \left(\left(\frac{d}{c}\right)^2 - 1\right)^2 & \text{se } |d| \le c \\ 0 & \text{se } |d| > c \end{cases} \tag{2.24}$$

per Tukey. Quest'ultima, per come è definita $\rho_{[bi]}(\cdot;\boldsymbol{\theta},c)$, si ottiene mol-

tiplicando per un fattore pari a $\frac{c^2}{6}$, ovvero calcolando $w_{[bi]}(d(\cdot;\boldsymbol{\theta});\boldsymbol{\theta},c) = (c^2/6)(\partial/\partial\boldsymbol{\theta})\rho_{[bi]}(d(\cdot;\boldsymbol{\theta});\boldsymbol{\theta},c)/d(\cdot;\boldsymbol{\theta})$. I due stimatori di tipo MM corrispondenti alle due scelte delle funzioni $\rho_1(\cdot;\boldsymbol{\theta})$, e quindi di $w_1(\cdot;\boldsymbol{\theta})$, vengono indicati rispettivamente con $\hat{\boldsymbol{\beta}}_{[Hub]}$ e $\hat{\boldsymbol{\beta}}_{[bi]}$. In generale, lo stimatore ottenuto con la funzione di Huber ha distorsione elevata in presenza di valori leva, mentre quello ottenuto con la funzione di Tukey è robusto sia in presenza di osservazioni anomale nelle covariate sia in presenza di osservazioni anomale nella variabile risposta.

Le costanti $c_0$ e $c_1$ nelle funzioni $\rho_0(\cdot;\boldsymbol{\theta},c_0)$ e $\rho_1(\cdot;\boldsymbol{\theta},c_1)$ vengono scelte con l'obiettivo di raggiungere uno specifico punto di rottura, attraverso $c_0$, e una specifica efficienza dello stimatore, attraverso $c_1$. Essendo la prima fase basata di fatto sullo stimatore $\hat{\boldsymbol{\alpha}}_{[\mathrm{CBS}]}$, la scelta della costante $c_0$ avviene allo stesso modo di quanto spiegato per gli stimatori di tipo S; fissato un certo punto di rottura che si vuole ottenere, ad esempio $\varepsilon^* = 50\%$, si risolve in $c = c_0$ l'equazione (2.13). Per determinare la costante $c_1$, invece, si basa il ragionamento sull'efficienza asintotica relativa dello stimatore (ARE, *asymptotic relative efficiency*). Essa è definita come (Huber e Ronchetti, 2009)

$$\mathrm{ARE} = \lim_{n\to\infty} \frac{Var(\hat{\boldsymbol{\theta}}_{[\mathrm{MM}]})/E(\hat{\boldsymbol{\theta}}_{[\mathrm{MM}]})^2}{Var(\hat{\boldsymbol{\theta}})/E(\hat{\boldsymbol{\theta}})^2}. \tag{2.25}$$

Mediante la costante $c_1$ si può scegliere quanto si è disposti a perdere in efficienza rispetto allo stimatore di massima verosimiglianza $\hat{\boldsymbol{\theta}}$, il più efficiente all'interno della classe degli stimatori lineari non distorti, nel caso in cui il processo generatore dei dati sia realmente $\mathcal{F}_{\boldsymbol{\theta}}$. Tipicamente si richiede che il livello di efficienza asintotica relativa sia del 95%.

I grafici in Figura 2.1 mostrano l'andamento delle funzioni $\rho(r;\boldsymbol{\theta})$, $\boldsymbol{\Psi}(r;\boldsymbol{\theta})$ e $w(r;\boldsymbol{\theta})$ per lo stimatore di massima verosimiglianza, per lo stimatore di Huber e per quello di Tukey. I grafici sono disegnati in funzione dei valori che possono assumere i residui, indicati con $r$, anziché della distanza $d(\cdot;\boldsymbol{\theta})$. Essi intendono evidenziare come la richiesta di B-robustezza si traduca in

opportuni vincoli di limitatezza della funzione $\boldsymbol{\Psi}(\cdot;\boldsymbol{\theta})$ (seconda riga), mentre la richiesta di punti di rottura alti in vincoli di limitatezza sulla funzione $\rho(\cdot;\boldsymbol{\theta})$ (prima riga). Nell'ultima riga della Figura 2.1 si vede il diverso modo dei tre stimatori di attribuire i pesi alle osservazioni, parimenti importanti per quello di massima verosimiglianza e con pesi decrescenti per gli stimatori robusti. Essendo le funzioni $w(r;\boldsymbol{\theta})$ simmetriche pari, esse possono essere studiate solo per valori positivi e interpretare l'andamento dei pesi in termini di distanza dal modello stimato.

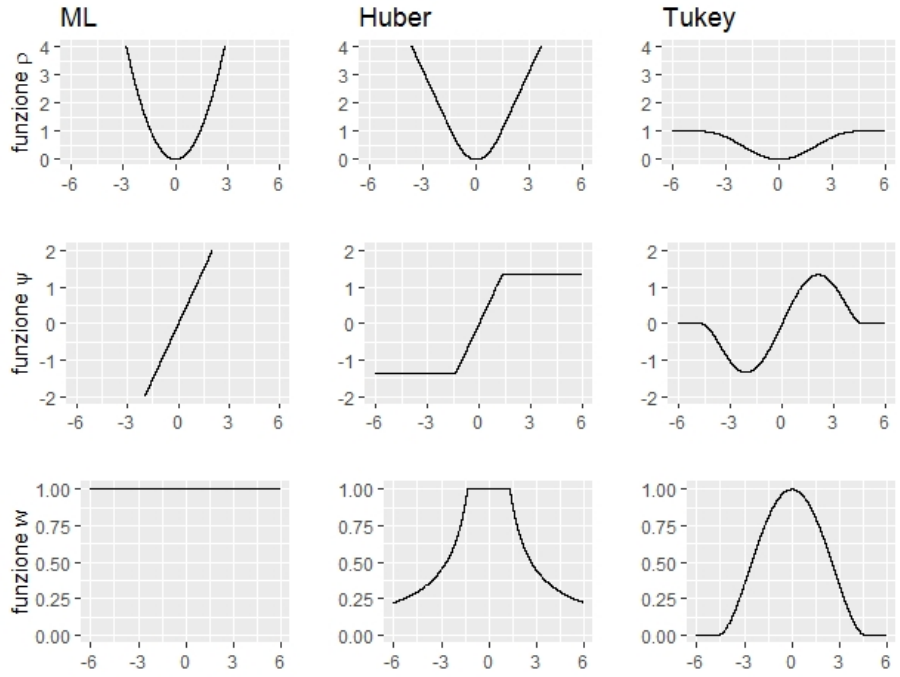



Figura 2.1: Grafici delle funzioni $\rho(r;\boldsymbol{\theta})$ (prima riga), $\boldsymbol{\Psi}(r;\boldsymbol{\theta})$ (seconda riga) e $w(r;\boldsymbol{\theta})$ (terza riga) per gli stimatori di massima verosimiglianza (prima colonna), di Huber con $c = 1.345$ (seconda colonna), e di Tukey con $c = 4.685$ (terza colonna). In ascissa residuo della generica osservazione.

## 2.4 Procedure robuste per la verifica d'ipotesi

Come nell'approccio classico, nelle verifiche d'ipotesi l'interesse principale è ristretto alla componente degli effetti fissi del parametro. Per ottenere una

statistica test di Wald robusta (Heritier *et al.*, 2009, Cap. 4) è sufficiente specificare nella (1.11) al posto di $\hat{\boldsymbol{\beta}}$ uno stimatore robusto, ad esempio $\hat{\boldsymbol{\beta}}_{[\mathrm{MM}]}$, e al posto della varianza stimata dello stimatore di massima verosimiglianza una stima della varianza dello stimatore robusto, ad esempio quella fornita nella (2.21), in cui $M$ e $Q$ vengono sostituiti dalle loro versioni empiriche. La distribuzione della statistica test sotto $H_0$ rimane quella di un chi quadrato con rank($\boldsymbol{C}$) gradi di libertà. Per il sistema di verifica d'ipotesi

$$\begin{cases} H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{\beta}_0 \\ H_1 : \overline{H_0}, \end{cases},$$

utilizzando lo stimatore robusto di tipo MM, si ottiene quindi

$$W_{e[\mathrm{MM}]} = (\boldsymbol{C}\hat{\boldsymbol{\beta}}_{[\mathrm{MM}]} - \boldsymbol{\beta}_0)^{\mathsf{T}}[\boldsymbol{C}\,\hat{Var}(\hat{\boldsymbol{\beta}}_{[\mathrm{MM}]})\boldsymbol{C}^{\mathsf{T}}]^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}}_{[\mathrm{MM}]} - \boldsymbol{\beta}_0). \quad (2.26)$$

In modo analogo, la statistica test $z$ robusta, valida anche per le verifiche d'ipotesi unilaterali su una combinazione lineare dei $\boldsymbol{\beta}$, si calcola semplicemente sostituendo allo stimatore di massima verosimiglianza uno stimatore robusto. Il sistema di verifica d'ipotesi può essere scritto in generale come

$$\begin{cases} H_0 : \boldsymbol{c}^{\mathsf{T}}\boldsymbol{\beta} = \beta_0 \\ H_1 : \overline{H_0} \end{cases},$$

dove $\boldsymbol{c}$ è un vettore $p$-dimensionale e $\beta_0$ è uno scalare, ad esempio zero. Utilizzando lo stimatore robusto di tipo MM, si ottiene

$$z_{[\mathrm{MM}]} = \frac{\boldsymbol{c}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{[\mathrm{MM}]} - \beta_0}{\sqrt{\widehat{Var}(\boldsymbol{c}^{\mathsf{T}}\hat{\boldsymbol{\beta}}_{[\mathrm{MM}]}\boldsymbol{c})}}. \quad (2.27)$$

La distribuzione asintotica nulla della statistica test $z_{[\mathrm{MM}]}$ rimane come nell'approccio classico la normale standard. Entrambi questi test di Wald

possono essere eseguiti utilizzando sia lo stimatore di tipo MM, con entrambe le scelte per la funzione $\rho_1(\cdot; \boldsymbol{\theta}, c_1)$, sia lo stimatore di tipo S con la rispettiva stima della varianza ottenuta a partire dalla (2.19).

L'altra statistica test molto utilizzata è la statistica test log-rapporto di verosimiglianza (Heritier *et al.*, 2009, Cap. 4). Essa è basata, come nell'approccio classico, sul sistema di verifica d'ipotesi

$$\begin{cases} H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \\ H_1 : \overline{H_0} \end{cases},$$

dove $\boldsymbol{\beta}_0$ può essere immaginato suddiviso in due blocchi $(\boldsymbol{\beta}_{0(1)}^{\mathsf{T}}, \boldsymbol{\beta}_{0(2)}^{\mathsf{T}})^{\mathsf{T}}$, con $\boldsymbol{\beta}_{0(1)}$ vettore fissato, ad esempio pari a $\mathbf{0}$, di dimensione $k$, $k \leq p$, e $\boldsymbol{\beta}_{0(2)}$ vettore non specificato di dimensione $p - k$. La statistica test log-rapporto di verosimiglianza robusta è

$$\mathrm{LRT}_{[\mathrm{MM}]} = 2 \sum_{i=1}^{n} [\rho(d_i(\boldsymbol{y}_i; \tilde{\boldsymbol{\beta}}_{[\mathrm{MM}]}, \hat{\boldsymbol{\alpha}})) - \rho(d_i(\boldsymbol{y}_i; \hat{\boldsymbol{\beta}}_{[\mathrm{MM}]}, \hat{\boldsymbol{\alpha}}))], \tag{2.28}$$

dove $\tilde{\boldsymbol{\beta}}_{[\mathrm{MM}]}$ è la stima ottenuta con lo stimatore robusto di tipo MM sotto il vincolo di $H_0$, $d_i(\boldsymbol{y}_i)$ è la distanza di Mahalanobis data nella (2.14) con $\boldsymbol{\Sigma}$ stimato, ad esempio, con $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}}_{[\mathrm{CBS}]})$ e $\hat{\boldsymbol{\alpha}}$ è una stima robusta della componente $\boldsymbol{\alpha}$, ad esempio $\hat{\boldsymbol{\alpha}}_{[\mathrm{CBS}]}$. Le funzioni $\rho$ sono le corrispondenti funzioni, o di Huber o di Tukey, utilizzate nel calcolo degli stimatori $\tilde{\boldsymbol{\beta}}_{[\mathrm{MM}]}$ e $\hat{\boldsymbol{\beta}}_{[\mathrm{MM}]}$.

A differenza del test classico, la statistica test log-rapporto di verosimiglianza robusta ha distribuzione nulla approssimata data dalla somma pesata di $k = \dim(\boldsymbol{\beta}_{0(1)})$ distribuzioni $\chi_1^2$ indipendenti. Nell'approccio robusto, le statistiche test di Wald e log-rapporto di verosimiglianza non sono quindi asintoticamente equivalenti. A differenza del test di Wald robusto, il test log-rapporto di verosimiglianza robusto può essere eseguito solo per gli stimatori di tipo MM. Infatti, per uno stimatore di tipo S il valore osservato della statistica test è pari a zero per costruzione, in quanto gli stimatori di

tipo S sono soggetti al vincolo (2.12). Il test log-rapporto di verosimiglianza robusto richiede di conoscere la funzione $\rho$ e di calcolare anche la stima sotto $H_0$, però presenta il vantaggio di essere più potente rispetto al test di Wald robusto.

Come discusso nell'approccio classico, per quanto riguarda i problemi di verifica d'ipotesi sulle varianze degli effetti casuali, l'ipotesi nulla del tipo $H_0 : \sigma_j^2 = 0$ si trova sul confine dello spazio parametrico e quindi la teoria generale dei test utilizzata per gli effetti fissi non è più valida. Nel Paragrafo 1.3 si era risolto il problema correggendo la distribuzione nulla del test; tuttavia, questo test risulta poco potente. Nell'approccio robusto una via percorribile per verificare questo tipo d'ipotesi è utilizzare tecniche bootstrap basate su stimatori robusti con alti punti di rottura ($\varepsilon^* = 50\%$). Per approfondimenti si rimanda a Heritier *et al.* (2009, Cap. 4).

## 2.5 Stimatore di massima verosimiglianza troncata

### 2.5.1 Definizione dello stimatore

Gli stimatori robusti noti in letteratura descritti nel Paragrafo 2.3 hanno un forte limite applicativo. Essi possono essere utilizzati solo in particolari casi di studio in cui il disegno sperimentale è bilanciato. In molte situazioni, invece, si può avere un numero diverso di osservazioni per ciascuna unità statistica: questo si può verificare perché alcune unità escono durante lo studio, si parla di *lost to follow up*, oppure perché le unità di primo livello hanno di loro natura dimensionalità diversa, si pensi all'esempio dell'intervista ai componenti di un nucleo famigliare. In tutte queste situazioni, lo stimatore di tipo S, e di conseguenza lo stimatore di tipo MM, non può essere applicato.

Una recente proposta alternativa è stata presentata nell'ambito dei modelli lineari multivariati con effetti casuali da Ruli *et al.* (2019). Il metodo proposto

si basa sul concetto di massima verosimiglianza troncata e risulta più flessibile delle alternative robuste note in letteratura, essendo applicabile anche in contesti in cui il disegno sperimentale non è bilanciato. L'interesse è di studiare lo stimatore di massima verosimiglianza troncata (TML, *Trimmed Maximum Likelihood*) applicandolo al caso dei modelli univariati trattati in questo elaborato.

L'idea alla base riprende quella del più noto stimatore *trimmed* della media che calcola una media aritmetica escludendo una porzione fra i valori minori e maggiori osservati. In questo modo lo stimatore *trimmed* utilizza i soli valori centrali, troncando i valori estremi ritenuti anomali. Questo rende robusto lo stimatore; basti pensare alla mediana come uno stimatore *trimmed* con una porzione di troncamento pari al 50% sia per i valori più piccoli sia per quelli più grandi. Troncare una porzione inferiore di valori rispetto alla mediana permette di ottenere un compromesso fra efficienza e robustezza.

Lo stimatore di massima verosimiglianza troncata utilizza i contributi alla log-verosimiglianza delle singole unità statistiche come misura per determinare quali osservazioni includere o meno nel calcolo. Per i modelli trattati in questo elaborato, lo stimatore TML si ottiene calcolando lo stimatore di massima verosimiglianza (Paragrafo 1.2.1) sul sottocampione $H^*$ di $n(1-\alpha)$ unità che massimizzano la log-verosimiglianza nel sottocampione. Lo stimatore è definito dunque come

$$\hat{\boldsymbol{\theta}}_{[\text{TML}]} = \underset{H \in \mathcal{H}}{\arg\max} \sum_{i \in H} \ell_i(\hat{\boldsymbol{\theta}}_H; \boldsymbol{y}_i)\,, \tag{2.29}$$

dove $\mathcal{H}$ è l'insieme di tutti i possibili sottocampioni di numerosità $n(1-\alpha)$, $\ell_i(\cdot)$ rappresenta l'$i$-esimo contributo alla log-verosimiglianza e

$$\hat{\boldsymbol{\theta}}_H = \underset{\boldsymbol{\theta} \in \Theta}{\arg\max} \sum_{i \in H} \ell_i(\boldsymbol{\theta}; \boldsymbol{y}_i) \tag{2.30}$$

è lo stimatore di massima verosimiglianza calcolato sul generico sottocampione $H$. Operando su $H^*$ si escludono dal calcolo le unità più distanti dal modello stimato, ritenendole anomale. Infatti, come si può osservare scrivendo l'$i$-esimo contributo alla log-verosimiglianza, data da

$$\ell_i(\hat{\boldsymbol{\theta}}_{[\text{TML}]}; \boldsymbol{y}_i) = -\frac{m}{2}\log(2\pi) - \frac{1}{2}\log(|\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}}_{[\text{TML}]})|) \\ - \frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_{[\text{TML}]})^\top \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}}_{[\text{TML}]})^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}_{[\text{TML}]}), \quad (2.31)$$

selezionare le unità statistiche con contributo maggiore equivale, in campioni bilanciati, a selezionare quelle che hanno il quadrato della distanza di Mahalanobis (2.14) inferiore.

Quello che si vorrebbe trovare è il sottocampione $H^* \in \mathcal{H}$, per cui il valore $\ell_i(\hat{\boldsymbol{\theta}}_{[\text{TML}]}; \boldsymbol{y}_i)$ è massimo, fra i $\binom{n}{n(1-\alpha)}$ sottocampioni di numerosità $n(1-\alpha)$ estraibili dal campione di numerosità $n$ (per la scelta della costante $\alpha$, si rimanda al Paragrafo 2.5.3). Data l'impossibilità computazionale di provare tutti i possibili sottocampioni $H$ per un $\alpha$ fissato, ad esempio pari a 0.25, anche in piccoli campioni, l'algoritmo segue una strada "sub-ottima" basata su una procedura iterativa (Rousseeuw e Van Driessen, 1999). Per ricercare il sottocampione su cui calcolare lo stimatore TML, ad ogni iterazione l'algoritmo calcola lo stimatore di massima verosimiglianza sul sottoinsieme delle osservazioni che danno un contributo alla log-verosimiglianza, calcolato al passo precedente, maggiore. Si osserva che in base al sottocampione di partenza scelto, l'algoritmo può convergere a stime differenti, per questo si provano diversi punti di partenza e si seleziona quello che converge ad un valore della log-verosimiglianza maggiore.

### 2.5.2 Algoritmo per il calcolo

Di seguito si elencano nel dettaglio i passaggi per il calcolo dello stimatore TML:

1. dall'insieme di dati a disposizione si generano 500 sottocampioni di dimensione $n(1-\alpha)$, ad esempio attraverso l'estrazione di numeri casuali o pseudo-casuali da un calcolatore. Questi rappresentano i diversi punti di partenza dell'algoritmo.

2. Su ciascun sottocampione si calcola lo stimatore di massima verosimiglianza $\hat{\boldsymbol{\theta}}_H$ basato sulle $n(1-\alpha)$ osservazioni.

3. Si calcolano i contributi alla log-verosimiglianza (2.31) usando la stima $\hat{\boldsymbol{\theta}}_H$ su ciascuna delle $n$ osservazioni. Sul sottocampione $H$ di questa iterazione, la log-verosimiglianza complessiva calcolata in $\hat{\boldsymbol{\theta}}_H$ è data da $\ell(\hat{\boldsymbol{\theta}}_H; \boldsymbol{y}) = \sum_{i=1}^{n} \ell_i(\hat{\boldsymbol{\theta}}_H; \boldsymbol{y}_i)$.

4. Si selezionano le $n(1-\alpha)$ unità statistiche a cui sono associati i contributi maggiori; tali unità andranno ad aggiornare quelle del sottocampione $H$ per l'iterazione successiva.

5. I passi 2 - 3 - 4 sono ripetuti finché la procedura non arriva a convergenza, ossia quando il gruppo di $n(1-\alpha)$ unità che costituiscono il sottocampione $H$ non cambia da una iterazione all'altra.

6. Arrivati a convergenza tutti e 500 i sottocampioni, si seleziona quello con $\ell(\hat{\boldsymbol{\theta}}_H; \boldsymbol{y})$ maggiore. Lo stimatore di massima verosimiglianza troncata, indicato con $\hat{\boldsymbol{\theta}}_{[\mathrm{TML}]}$, è quello calcolato su tale sottocampione.

Per l'implementazione in R dell'algoritmo di stima si veda l'Appendice A.

Tale algoritmo comporta il calcolo dello stimatore di massima verosimiglianza per 500 sottocampioni per il numero di iterazioni fatte su ciascuno. Tuttavia, da risultati empirici emerge che il singolo sottocampione converge molto rapidamente, solitamente in un paio di iterazioni al massimo, quindi la procedura non risulta eccessivamente lunga, considerando che l'algoritmo si presta ad una parallelizzazione esplicita naturale. In Figura 2.2 viene mostrata a titolo esemplificativo la velocità di convergenza delle log-verosimiglianze

calcolate in $\hat{\boldsymbol{\theta}}_H$ di 500 sottocampioni generati casualmente da un *dataset* di dati simulati. Quello che si nota è che, dopo poche iterazioni, il valore $\ell(\hat{\boldsymbol{\theta}}_H; \boldsymbol{y})$ rimane costante da una iterazione all'altra ad indicare che l'algoritmo è arrivato a convergenza su quel sottocampione. In altre parole, i passi 2 - 3 - 4 dell'algoritmo vengono eseguiti un paio di volte per ogni sottocampione. Si sottolinea che seguendo una strada "sub-ottima", non vi è garanzia che il sottocampione trovato dall'algoritmo corrisponda al sottocampione $H^*$ per cui vale la (2.29).


Figura 2.2: Andamento di $\ell(\hat{\boldsymbol{\theta}}_H; \boldsymbol{y})$ nei 500 sottocampioni in funzione del numero di iterazioni dell'algoritmo per il calcolo dello stimatore TML.

### 2.5.3 Scelta della cardinalità del sottocampione

La costante $\alpha$ rappresenta la percentuale di unità statistiche escluse dal calcolo dello stimatore TML e determina la cardinalità del sottocampione $H$. Tipicamente $\alpha$ varia nell'intervallo $(0\,,\,0.25)$, dove la scelta $\alpha = 0$ porta ovviamente a $\hat{\boldsymbol{\theta}}_{[\text{TML}]} = \hat{\boldsymbol{\theta}}$, ovvero lo stimatore di massima verosimiglianza calcolato su tutte le $n$ unità. La scelta di $\alpha$ è legata al punto di rottura che si desidera ottenere per lo stimatore. Il punto di rottura (2.6) per lo stimatore TML è infatti pari ad $\alpha$. Più si desidera uno stimatore con un

punto di rottura alto più si fissa un valore alto di $\alpha$. Bisogna tenere presente, tuttavia, che aumentando $\alpha$ lo stimatore perde sempre più in efficienza nel caso in cui il modello specificato coincida con il vero processo generatore, ovvero in assenza di osservazioni anomale.

### 2.5.4 Stima della matrice di varianze e covarianze

La matrice di varianze e covarianze $\boldsymbol{\Sigma}$ può essere stimata semplicemente sostituendo nella (1.3) le stime ottenute con lo stimatore TML per la componente relativa alle varianze degli effetti casuali $\hat{\boldsymbol{\alpha}}_{[\text{TML}]}$. Tuttavia si dimostra che $\boldsymbol{\Sigma}(\hat{\boldsymbol{\alpha}}_{[\text{TML}]})$ risulta non consistente (Ruli *et al.*, 2019). Per recuperare la consistenza, viene calcolato un fattore di correzione moltiplicativo per la stima $\boldsymbol{\Sigma}(\hat{\boldsymbol{\alpha}}_{[\text{TML}]})$ dato da

$$c_\alpha = \frac{1-\alpha}{F_{\chi^2_{2+m}}(\chi^2_{m;1-\alpha})}, \tag{2.32}$$

dove $F_{\chi^2_{2+m}}(\cdot)$ indica la funzione di ripartizione di un chi quadrato con $2+m$ gradi di libertà e $\chi^2_{m;1-\alpha}$ il quantile di livello $1-\alpha$ di un chi quadrato con $m$ gradi di libertà. Una stima consistente della matrice di varianze e covarianze si ottiene dunque calcolando

$$\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}}_{[\text{TML}]}) = c_\alpha \boldsymbol{\Sigma}(\hat{\boldsymbol{\alpha}}_{[\text{TML}]}). \tag{2.33}$$

## 2.6 Stimatore di massima verosimiglianza troncata ripesato

Lo stimatore *trimmed* presentato nel paragrafo precedente risulta inefficiente in situazioni in cui il modello ipotizzato coincide con il processo generatore dei dati. Questo accade perché lo stimatore TML esclude dal calcolo una percentuale costante di unità statistiche pari ad $\alpha$. Per superare questo problema, Ruli *et al.* (2019) propongono una versione ripesata dello

stimatore, che ne migliora l'efficienza, lasciando inalterate le proprietà di robustezza.

Una volta ottenuta la stima $\hat{\boldsymbol{\theta}}_{[\mathrm{TML}]}$, si calcolano i contributi $\ell_i(\hat{\boldsymbol{\theta}}_{[\mathrm{TML}]}; \boldsymbol{y}_i)$ sulle $n$ osservazioni usando la (2.31), in cui al posto della stima non consistente $\boldsymbol{\Sigma}(\hat{\boldsymbol{\alpha}}_{[\mathrm{TML}]})$ viene usata la stima consistente $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}}_{[\mathrm{TML}]})$ data nella (2.33). Successivamente, si selezionano le unità statistiche che danno un contributo alla log-verosimiglianza maggiore di una certa soglia. La soglia è fissata pari a

$$-\frac{1}{2}q_\delta - \frac{m}{2}\log(2\pi) - \frac{1}{2}\log(|\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}}_{[\mathrm{TML}]})|)\,, \tag{2.34}$$

dove $q_\delta = \chi^2_{m;1-\delta}$, con $0.01 \leq \delta \leq 0.025$. Si nota che in campioni bilanciati questo equivale a selezionare le unità che hanno il quadrato della distanza di Mahalanobis inferiore alla soglia $q_\delta$, ovvero le unità più vicine al modello stimato dallo stimatore $\hat{\boldsymbol{\theta}}_{[\mathrm{TML}]}$.

Una volta determinato il sottocampione di unità selezionate dalla soglia, lo stimatore di massima verosimiglianza troncata ripesato (RWTML, *Reweighted Trimmed Maximum Likelihood*), indicato con $\hat{\boldsymbol{\theta}}_{[\mathrm{RWTML}]}$, è ottenuto semplicemente calcolando lo stimatore di massima verosimiglianza su tale sottocampione. Per l'implementazione in R dell'algoritmo di stima si veda l'Appendice A.

A differenza dello stimatore $\hat{\boldsymbol{\theta}}_{[\mathrm{TML}]}$ che lavora con le $n(1-\alpha)$ unità più vicine indipendentemente dalla loro distanza dal modello stimato, lo stimatore $\hat{\boldsymbol{\theta}}_{[\mathrm{RWTML}]}$ lavora con le unità più vicine di una fissata soglia. Questo permette di escludere dal calcolo dello stimatore un numero inferiore di unità statistiche qualora il campione osservato contenga una percentuale di osservazioni anomale inferiore ad $\alpha$.

In campioni non contaminati ci si aspetta che in media lo stimatore RWTML tronchi una percentuale di unità statistiche pari a $\delta$ (Ruli *et al.*, 2019). Nel prossimo capitolo (in particolare nel Paragrafo 3.3.1), si vedrà che questo non è esattamente vero poiché la soglia (2.34) del *reweighted* è

basata sullo stimatore TML che, per campioni finiti di piccole dimensioni, si è dimostrato sottostimare la matrice di varianze e covarianze.

# Capitolo 3

# Studio di simulazione

## 3.1 Descrizione dello studio

In questo capitolo è d'interesse studiare il comportamento dei due stimatori robusti implementati in questo elaborato, in relazione agli altri stimatori noti in letteratura, per cui è già disponibile un'implementazione in R. In particolare, si confronteranno i due stimatori di tipo *trimmed* TML e RWTML con lo stimatore di massima verosimiglianza e lo stimatore di tipo S.

Il confronto è stato condotto attraverso uno studio di simulazione in cui si è considerato un modello con intercetta casuale, ovvero un modello del tipo

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \gamma_{i1} + \boldsymbol{\varepsilon}_i, \tag{3.1}$$

dove $\boldsymbol{X}_i$ è la matrice di disegno $m \times p$, $\gamma_{i1}$ la variabile aleatoria con distribuzione $\mathcal{N}(0, \sigma^2_{\gamma_1})$, che descrive l'effetto casuale, e $\boldsymbol{\varepsilon}_i$ il vettore aleatorio con distribuzione $\mathcal{N}_m(\boldsymbol{0}, \sigma^2_{\varepsilon}\boldsymbol{I}_m)$, che descrive il termine di errore sulle $m$ misurazioni dell'$i$-esima unità statistica, per $i = 1, \ldots, n$. Assumiamo inoltre $\gamma_{i1}$ e $\boldsymbol{\varepsilon}_i$ indipendenti fra loro e indipendenti per ogni $i = 1, \ldots, n$.

Si osserva che per effettuare il confronto con lo stimatore di tipo S, si è lavorato sempre con campioni bilanciati; tuttavia, si ricorda che il vantaggio

degli stimatori *trimmed* è quello di poter essere applicati anche in casi in cui il numero di osservazioni $m_i$, $i = 1, \ldots, n$, varia per ogni unità statistica.

Le simulazioni sono state condotte sotto diversi scenari, ovvero per diverse combinazioni di $n$ e di $m$, rispettivamente numero di unità statistiche e numero di osservazioni per ciascuna unità statistica. Il numero $p$ di effetti fissi è stato mantenuto costante, pari a tre. La frazione di dati contaminati presenti nel campione è stata fatta variare fra lo 0%, il 5% e il 10%, dove 0% rappresenta il caso di campione non contaminato. Le contaminazioni sono state costruite aggiungendo alla frazione di dati stabilita valori generati da una Uniforme$(10, 100)$.

Nell'implementazione degli stimatori *trimmed* adottata in queste simulazioni si è utilizzato $\alpha = 0.25$ per lo stimatore TML e $\delta = 0.025$ per lo stimatore RWTML. Lo stimatore di massima verosimiglianza è stato calcolato con la funzione `lme` del pacchetto `nlme` di R.

Il vero valore dei parametri da cui si sono generate le osservazioni $\boldsymbol{y}_i$ è:

- $\boldsymbol{\beta}^0 = (\beta_0^0,\ \beta_1^0,\ \beta_2^0)^\mathsf{T} = (0.75,\ -1.5,\ 2)^\mathsf{T}$,
- $\boldsymbol{\alpha}^0 = (\sigma_{\gamma_1}^{2\ 0},\ \sigma_\varepsilon^{2 0})^\mathsf{T} = (1.5,\ 1)^\mathsf{T}$.

## 3.2 Risultati

### 3.2.1 Metrica di valutazione

Per confrontare le *performance* dei diversi stimatori si è utilizzato l'errore quadratico medio (MSE, *mean squared error*), definito come

$$\mathrm{MSE}(\hat{\boldsymbol{\theta}}_{(k)}) = Var(\hat{\boldsymbol{\theta}}_{(k)}) + (\hat{\boldsymbol{\theta}}_{(k)} - \boldsymbol{\theta}^0_{(k)})^2, \tag{3.2}$$

dove $\hat{\boldsymbol{\theta}}_{(k)}$ indica il generico $k$-esimo elemento, per $k = 1, \ldots, \dim(\hat{\boldsymbol{\theta}}) = 5$, del generico stimatore $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^\mathsf{T}, \hat{\boldsymbol{\alpha}}^\mathsf{T})^\mathsf{T}$ e $\boldsymbol{\theta}^0_{(k)}$ il $k$-esimo elemento del vettore dei veri parametri $\boldsymbol{\theta}^0 = (\boldsymbol{\beta}^{0\mathsf{T}}, \boldsymbol{\alpha}^{0\mathsf{T}})^\mathsf{T}$.

### 3.2.2 Confronto fra i diversi stimatori

La Figura 3.1 mostra l'andamento dell'errore quadratico medio per i diversi stimatori di ciascun parametro al variare della percentuale di contaminazione presente nei dati, sotto diversi scenari. Ogni scenario è dato da una diversa combinazione dei valori di $m \in \{3, 5, 8\}$ e di $n \in \{30, 50\}$. L'andamento del MSE è misurato su scala logaritmica, per comodità di visualizzazione. La varianza e la distorsione degli stimatori sono state calcolate con tecniche Monte Carlo basate su 1005 replicazioni. In generale, si nota che mentre lo stimatore di massima verosimiglianza ha un MSE che cresce all'aumentare del livello di contaminazione presente nel campione, gli stimatori robusti hanno un MSE che rimane costante. Si evidenzia, però, che come atteso in campioni non contaminati, lo stimatore di massima verosimiglianza ha un MSE inferiore. Tuttavia, sono sufficienti piccolissimi livelli di contaminazione per preferire a quest'ultimo uno stimatore robusto.

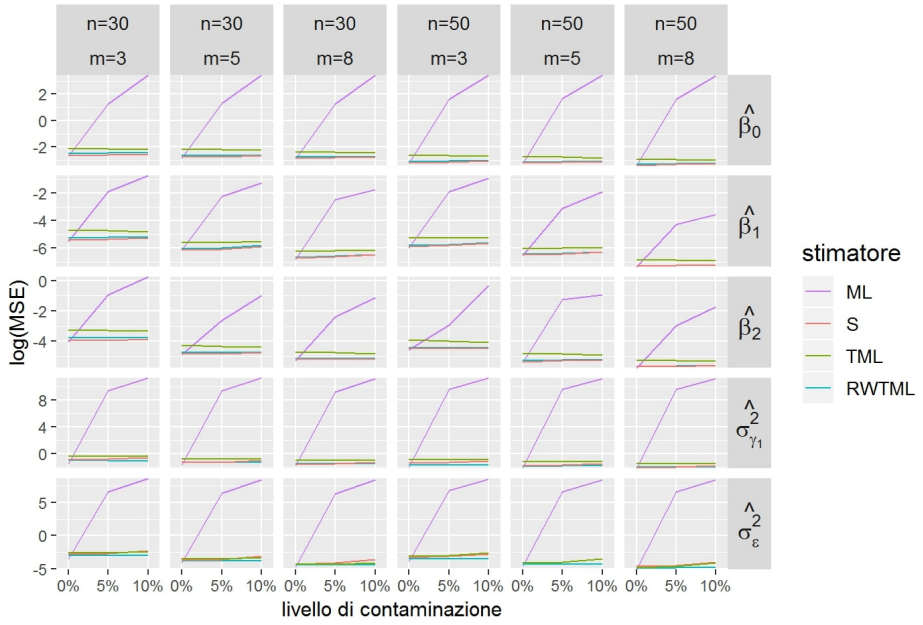



Figura 3.1: Confronto dell'andamento dell'errore quadratico medio (su scala logaritmica) per i diversi stimatori per ciascun parametro al variare della percentuale di dati contaminati presenti nel campione sotto diversi scenari.

Vista l'inadeguatezza dello stimatore di massima verosimiglianza nel caso in cui il campione contenga osservazioni anomale, si concentra il confronto unicamente fra i tre stimatori robusti. In Figura 3.2 viene riportato l'errore quadratico medio, condizionatamente a fissati valori di numerosità campionaria a disposizione e livello di contaminazione presente nei dati.

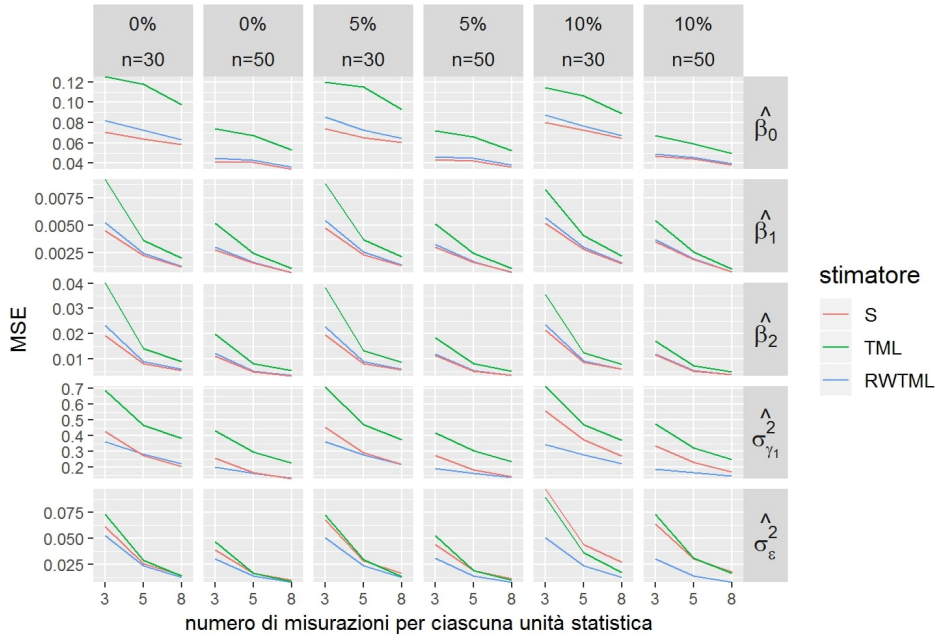



Figura 3.2: Confronto dell'andamento dell'errore quadratico medio per i diversi stimatori per ciascun parametro al variare del numero $m$ di misurazioni per unità statistica, condizionatamente al numero $n$ di unità e al livello di contaminazione presente nei dati.

Si può notare che, aumentando il numero $m$ di misurazioni per ciascuna unità statistica, tutti gli stimatori tendono a migliorare in termini di MSE, aumentando l'informazione proveniente dai dati. Come ci si aspettava, lo stimatore RWTML è sempre preferibile in termini di errore quadratico medio rispetto allo stimatore *trimmed* non ripesato. In particolare, si nota che il *reweighted* per la componente relativa agli effetti fissi ha un MSE di poco superiore rispetto a quello ottenuto dallo stimatore di tipo S, mentre per la componente relativa agli effetti casuali risulta nella maggior parte dei casi lo

stimatore migliore.

## 3.3 Correzione dello stimatore trimmed

### 3.3.1 Presentazione del problema

Lo stimatore di massima verosimiglianza troncata, nel contesto dei modelli con effetti casuali, risente dello stesso problema evidenziato da Pison *et al.* (2002) nel caso dei modelli lineari normali, ovvero sottostimare la matrice di varianze e covarianze $\boldsymbol{\Sigma}$. In Figura 3.3 viene mostrato, infatti, che nella maggior parte dei casi lo stimatore *trimmed* sottostima i valori di $\sigma^2_{\gamma_1}$ e di $\sigma^2_\varepsilon$, in particolare nei campioni di piccole dimensioni. Il fattore di correzione moltiplicativo $c_\alpha$ della (2.32) non è dunque sufficiente a correggere lo stimatore $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}}_{[\mathrm{TML}]})$.

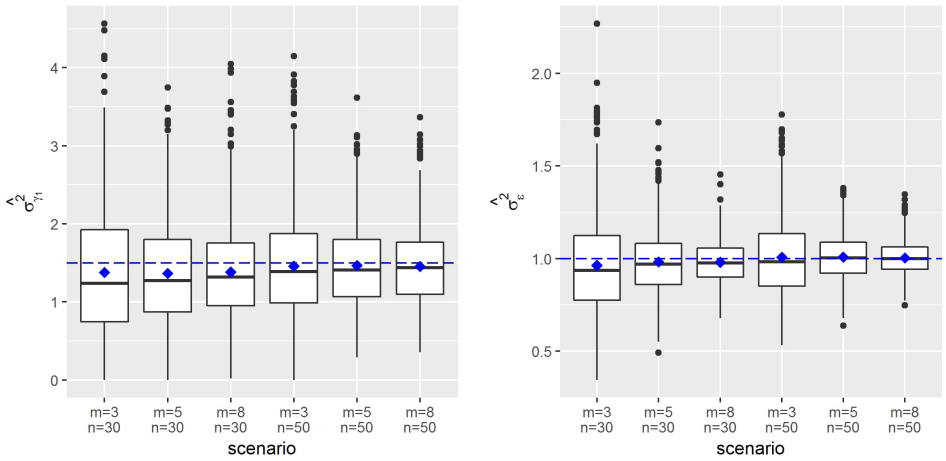



Figura 3.3: Boxplot degli stimatori *trimmed* per i parametri $\sigma^2_{\gamma_1}$ (a sinistra) e $\sigma^2_\varepsilon$ (a destra) in campioni non contaminati sotto diversi scenari. Puntino blu: media Monte Carlo, linea blu trattaggiata: vero valore del parametro.

In campioni non contaminati, i quadrati della distanza di Mahalanobis, dati da

$$(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\,, \tag{3.3}$$

seguono la distribuzione $\chi^2_m$. Sottostimando la matrice di varianze e covarianze, i quadrati delle distanze di Mahalanobis stimate risultano maggiori e le loro distribuzioni empiriche, in media, stocasticamente maggiori della distribuzione teorica.

Si osserva che tale sottostima si ripercuote negativamente anche sullo stimatore RWTML che, utilizzando la soglia $q_\delta$, tronca mediamente troppe osservazioni non anomale considerandole tali. Infatti, se le stime della (3.3) seguissero la distribuzione $\chi^2_m$, ci si aspetterebbe mediamente una frazione di osservazioni con quadrato della distanza di Mahalanobis maggiore di $q_\delta$ pari a $\delta = 2.5\%$. Dalle simulazioni svolte (Figura 3.4) si è notato che tale frazione è mediamente superiore. La correzione dello stimatore TML può avere dunque un duplice vantaggio.

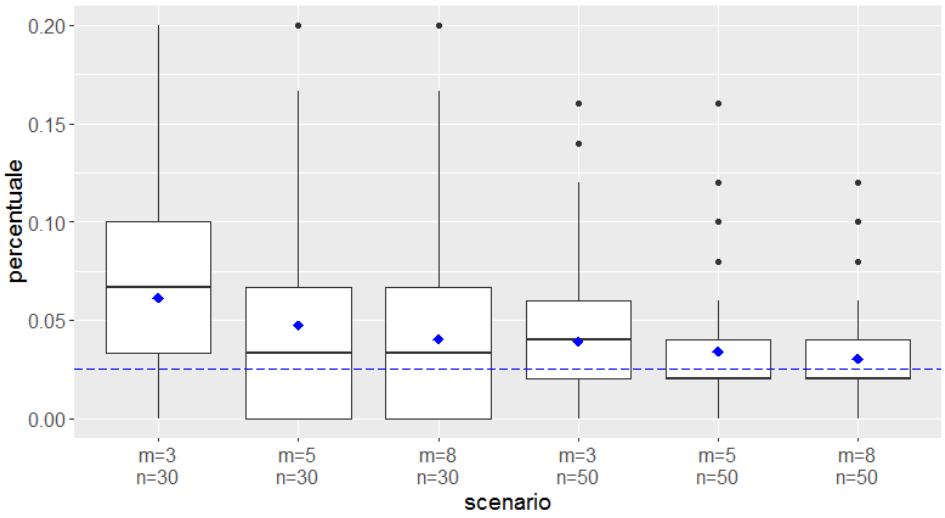



Figura 3.4: Boxplot delle percentuali di osservazioni con quadrato della distanza di Mahalanobis, stimata con il TML, superiore a $q_\delta$ in campioni non contaminati nei diversi scenari. Puntino blu: media Monte Carlo, linea blu trattaggiata: valore atteso teorico.

### 3.3.2 Un possibile approccio risolutivo

In analogia al lavoro svolto da Pison *et al.* (2002), si focalizza l'attenzione sulla stima $\hat{\mathbf{\Sigma}}(\hat{\boldsymbol{\alpha}}_{[\mathrm{TML}]})$, data nella (2.33), e si sceglie di correggere tale matrice

per un opportuno fattore moltiplicativo $c$, ottenendo lo stimatore

$$\hat{\boldsymbol{\Sigma}}_c(\hat{\boldsymbol{\alpha}}_{[\mathrm{TML}]}) = c\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}}_{[\mathrm{TML}]})\,. \tag{3.4}$$

Si osserva che correggere la stima della matrice $\boldsymbol{\Sigma}$ per un generico fattore moltiplicativo $c$ è equivalente a correggere le stime della componente $\boldsymbol{\alpha}$ moltiplicandole per lo stesso fattore $c$.

Il fattore $c$ viene scelto in modo da minimizzare la distanza fra la distribuzione empirica dei quadrati delle distanze di Mahalanobis stimate e la distribuzione teorica $\chi^2_m$. Per misurare la distanza fra le due distribuzioni si è utilizzata la distanza di Wasserstein di ordine uno, data da

$$W_1(F, G) = \int_{-\infty}^{\infty} |F(x) - G(x)|\, dx\,, \tag{3.5}$$

che confronta le due funzioni di ripartizione, $F(x)$ e $G(x)$. Tale distanza può essere utilizzata come alternativa alla distanza di Kolmogorov-Smirnov. Viene indicato con $c^*$ il fattore $c$ per cui tale distanza è minima, ovvero

$$c^* = \arg\min_c W_1(T(c), F_{\chi^2_m})\,, \tag{3.6}$$

dove con $T(c;\cdot)$ si è indicata la funzione di ripartizione dei quadrati delle distanze di Mahalanobis stimate con $\hat{\boldsymbol{\beta}}_{[\mathrm{TML}]}$ e $\hat{\boldsymbol{\Sigma}}_c(\hat{\boldsymbol{\alpha}}_{[\mathrm{TML}]})$ al variare di $c$, e con $F_{\chi^2_m}(\cdot)$ la funzione di ripartizione di un $\chi^2_m$. La (3.6) è stata risolta per via numerica seguendo un approccio di tipo *grid search* su una sequenza di valori. Trovato il valore $c^*$, si corregge la stima della matrice di varianze e covarianze calcolando

$$\hat{\boldsymbol{\Sigma}}_{c^*}(\hat{\boldsymbol{\alpha}}_{[\mathrm{TML}]}) = c^*\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\alpha}}_{[\mathrm{TML}]}). \tag{3.7}$$

Chiaramente, il valore di $c^*$ dipende dal singolo campione e può essere calcolato solo su campioni non contaminati. In presenza di osservazioni

anomale, non ha senso richiedere che i quadrati delle distanze di Mahalanobis seguano la distribuzione $\chi^2_m$. L'obiettivo è quello di studiare la distribuzione dei $c^*$ in campioni non contaminati e poi, una volta trovato un valore opportuno per ogni scenario, utilizzare tale correzione anche per i campioni contaminati.

## 3.4 Risultati dopo la correzione

La Figura 3.5 mostra il confronto fra le distribuzioni dello stimatore *trimmed* per i due parametri della matrice $\mathbf{\Sigma}$ con e senza la correzione per il fattore $c^*$, nei diversi scenari.

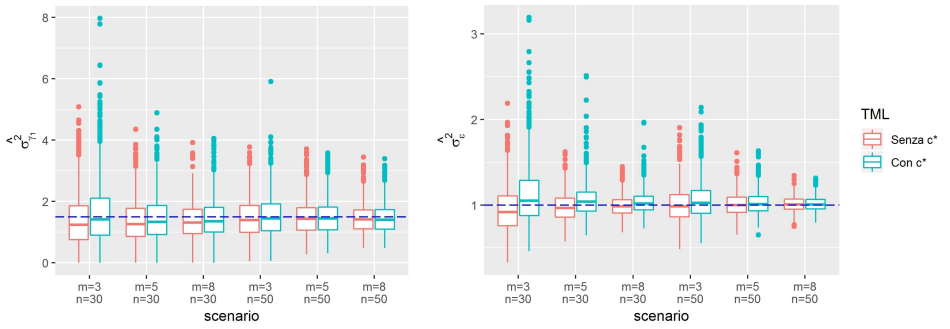



Figura 3.5: Confronto fra i boxplot degli stimatori *trimmed* con e senza la correzione per $c^*$ per i parametri $\sigma^2_{\gamma_1}$ (a sinistra) e $\sigma^2_{\varepsilon}$ (a destra) in campioni non contaminati sotto diversi scenari. Linea blu tratteggiata: vero valore del parametro.

Si osserva che la correzione per $c^*$ fa diminuire la distorsione dello stimatore, in particolare, si sottostimano di meno entrambi i parametri della matrice di varianze e covarianze. Tuttavia, tale miglioramento non compensa l'aumento della varianza dello stimatore che ne segue. Se si utilizza l'errore quadratico medio come metrica di valutazione, la correzione per $c^*$ peggiora le *performance* dello stimatore. Ottenere la distribuzione teorica $\chi^2_m$ per le stime dei quadrati delle distanze di Mahalanobis è una condizione necessaria ma non sufficiente.

Al fine di trovare un fattore di correzione comune per ogni scenario, risulta di interesse studiare la distribuzione dei $c^*$ ottimizzati su ogni singolo *dataset* nei diversi scenari (Figura 3.6). Le distribuzioni sono ottenute sulla base di 1005 replicazioni Monte Carlo per ogni scenario.

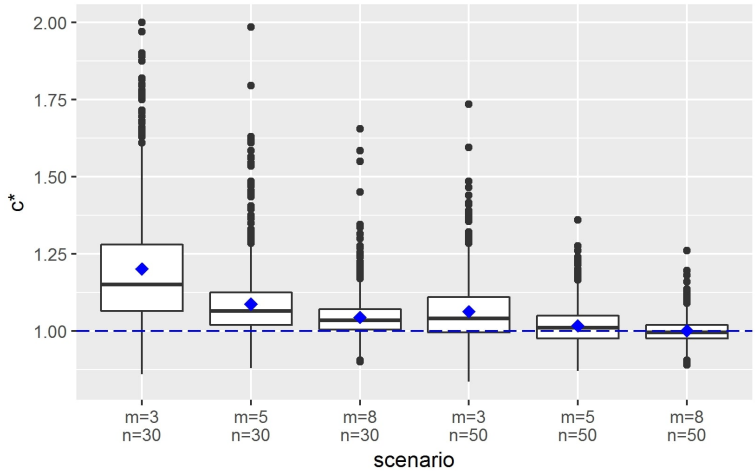



Figura 3.6: Boxplot dei valori assunti da $c^*$ in 1005 campioni non contaminati nei diversi scenari.

Si osserva che le distribuzioni sono fortemente asimmetriche a destra e, come ci si aspettava, convergono ad uno all'aumentare sia di $m$ che di $n$. Come già evidenziato, il fattore moltiplicativo $c^*$ è utile per correggere la stima della matrice di varianze e covarianze in campioni finiti di piccole dimensioni.

Si mostra, infine, il confronto con le distribuzioni dello stimatore *trimmed* in cui si è adottato un fattore di correzione $c^*_{m,n}$ comune in ogni scenario. Risulta di interesse valutare il confronto fra lo stimatore *trimmed* non corretto e quello corretto per $c^*_{m,n}$ nell'ottica che questa correzione possa essere estesa anche al caso di campioni contaminati. Osservando le distribuzioni asimmetriche di Figura 3.6, come valore comune $c^*_{m,n}$ si è provato ad utilizzare sia il valore mediano (Figura 3.7) che il valore medio (Figura 3.8) delle distribuzioni di ciascuno scenario.

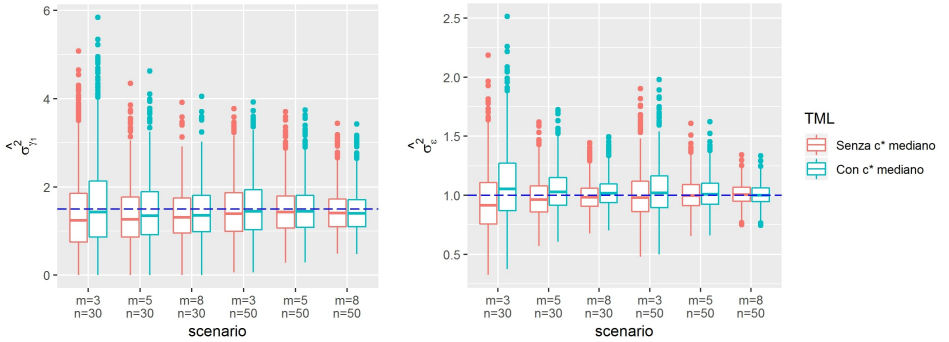



Figura 3.7: Confronto fra i boxplot degli stimatori *trimmed* per i parametri $\sigma^2_{\gamma_1}$ (a sinistra) e $\sigma^2_\varepsilon$ (a destra) in campioni non contaminati, con e senza la correzione, usando un $c^*_{m,n}$ mediano comune in ogni scenario. Linea blu tratteggiata: vero valore del parametro.

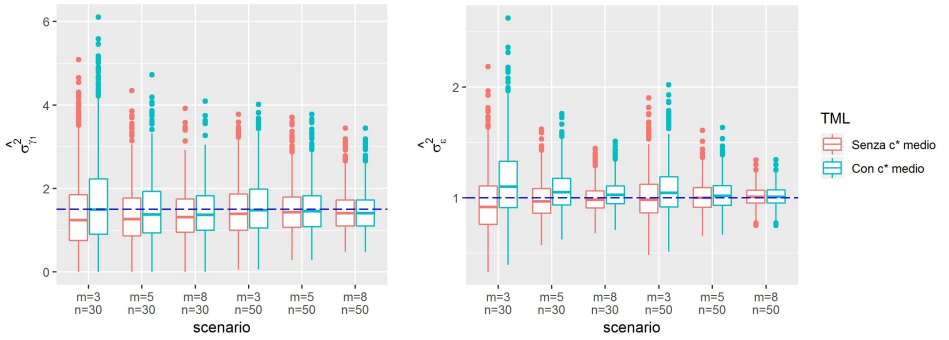



Figura 3.8: Confronto fra i boxplot degli stimatori *trimmed* per i parametri $\sigma^2_{\gamma_1}$ (a sinistra) e $\sigma^2_\varepsilon$ (a destra) in campioni non contaminati, con e senza la correzione, usando un $c^*_{m,n}$ medio comune in ogni scenario. Linea blu tratteggiata: vero valore del parametro.

Analogamente al caso discusso con i $c^*$ ottimizzati su ogni singolo *dataset*, questo tipo di correzione, benché migliori leggermente la distorsione, provoca un aumento della varianza. Usando l'errore quadratico medio si preferisce lo stimatore *trimmed* senza correzione. Essendo le distribuzioni dei $c^*$ asimmetriche, risulta preferibile lavorare con $c^*_{m,n}$ pari al valore mediano; utilizzare la media può aumentare la distorsione anziché ridurla e portare ad una sovrastima.

Dai risultati delle simulazioni si è visto che questo tipo di correzione non aiuta a migliorare l'errore quadratico medio dello stimatore *trimmed*. Tuttavia, si può dimostrare che avvicinando la distribuzione empirica dei quadrati delle distanze di Mahalanobis stimate con il TML alla distribuzione teorica $\chi^2_m$ e, in particolare, non sottostimando la matrice di varianze e covarianze $\mathbf{\Sigma}$, lo stimatore RWTML guadagna in efficienza. Correggendo lo stimatore TML con il valore $c^*_{m,n}$ mediano in ogni scenario, lo stimatore RWTML considera erroneamente anomale una percentuale inferiore di osservazioni (Figura 3.9).

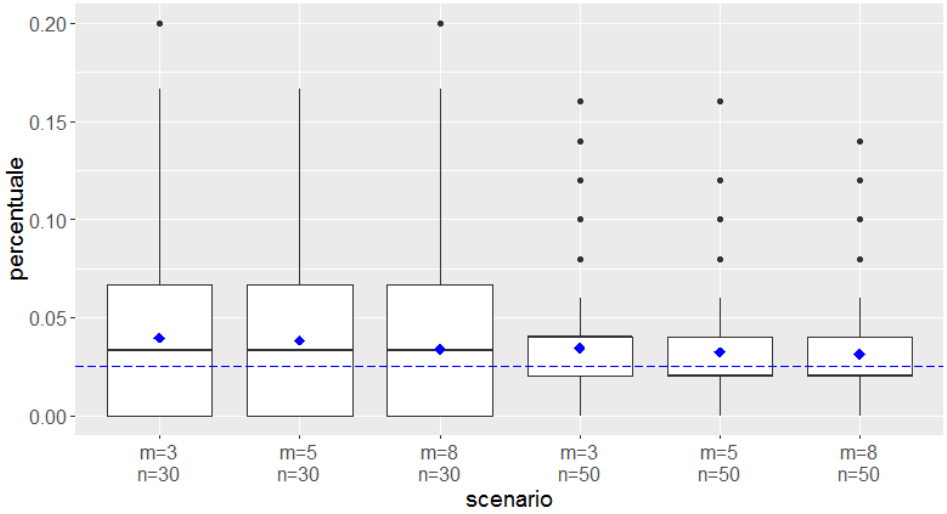



Figura 3.9: Boxplot delle percentuali di osservazioni con quadrato della distanza di Mahalanobis, stimata con il TML con la correzione data dal $c^*_{m,n}$ mediano, superiore a $q_\delta$ in campioni non contaminati nei diversi scenari. Puntino blu: media Monte Carlo, linea blu trattaggiata: valore atteso teorico.

# Conclusioni

L'interesse principale in questo elaborato si focalizza sullo studio di diverse tecniche statistiche robuste applicate all'ambito dei modelli lineari normali con effetti casuali. Questo ha compreso lo studio e l'implementazione di stimatori robusti di tipo *trimmed*, innovativi per questa classe di modelli, e la ricerca di una valida metodologia per la correzione di questi ultimi, a cui si è dedicata tutta la seconda parte dell'ultimo capitolo. Vantaggio rilevante degli stimatori di tipo *trimmed* è che non richiedono un disegno sperimentale bilanciato e risultano, pertanto, applicabili in un numero molto maggiore di situazioni.

Un possibile sviluppo di questo lavoro è la continuazione della ricerca di un opportuno fattore di correzione per lo stimatore *trimmed*, in particolare per la stima della matrice di varianze e covarianze. In questo elaborato si è seguito un approccio che prevede la correzione dell'intera matrice $\mathbf{\Sigma}$ tramite un unico fattore di correzione moltiplicativo. Un altro possibile approccio è quello di lavorare direttamente sulle singole stime $\hat{\sigma}^2_{\gamma_1\,[\mathrm{TML}]}$ e $\hat{\sigma}^2_{\varepsilon\,[\mathrm{TML}]}$, utilizzando due fattori di correzione non necessariamente moltiplicativi, rispettivamente $c^*_{\gamma_1}$ e $c^*_{\varepsilon}$. Una volta determinati dei buoni fattori di correzione per alcuni scenari, può essere di interesse modellare l'andamento di tali valori al variare di $m$ e di $n$ e ricostruire una mappa di fattori di correzione per ogni possibile combinazione di tali valori.

# Appendice A

# Codice R

## A.1 Funzioni ausiliarie

```
#estrazione indici pseudo-casuali per la prima iterazione
indici_casuali=function(alpha,n){
  sample(1:n,round(n*(1-alpha)))
}

#estrae il sottocampione di numerosità n*(1-alpha)
sotto_campione=function(idx,m,n,dati){
  flag=NULL
  for(i in 1:n){
    flag=c(flag,rep(any(i==idx),m[i]))
  }
  dati[flag,]
}

#valore per la soglia dello stimatore RWTML nella (2.34)
q=function(delta,m_i){
  qchisq(1-delta,df=m_i)
```

```
}

#fattore di correzione della (2.32) per la matrice Sigma
c_alpha=function(alpha,m_i){
  (1-alpha)/(pchisq(qchisq(1-alpha,df=m_i),df=2+m_i))
}

#stima matrice Sigma_hat della (2.33)
Sigma_corretta=function(alpha,eff_casuali_hat,m_i){
  c=c_alpha(alpha,m_i)
  Sigma_tilde=eff_casuali_hat[1]*matrix(1,m_i,m_i) +
    eff_casuali_hat[2]*diag(m_i)
  c*Sigma_tilde
}

#contributi alla log-verosimiglianza, dati nella (2.31)
ll_contributi=function(dati,y_hat,eff_casuali_hat,m,n,alpha=0){
  lli=rep(NA,n)
  for(i in 1:n){
    yi=dati$y[dati$sogg==i]
    yi_hat=y_hat[dati$sogg==i]
    Sigma_i_hat=Sigma_corretta(alpha,eff_casuali_hat,m[i])
    lli[i] = -m[i]/2*log(2*pi)-1/2*log(det(Sigma_i_hat))-
      1/2*t(yi-yi_hat)%*%solve(Sigma_i_hat)%*%(yi-yi_hat)
  }
  lli
}
```

## A.2 Stimatore trimmed

```
tml=function(data,alpha=0.25){
#-----------------------------------------------------------
# data = oggetto di tipo data.frame con colonne
#        y (risposta), x1 e x2 (esplicative),
#        sogg (identificativo del soggetto)
# alpha = parametro alpha dello stimatore
#         (vedi Paragrafo 2.5.3)
#
# output: oggetto di tipo lista con
#            beta (stime degli effetti fissi),
#            se_beta (standard error dei beta) e
#            sigma2_j (stime varianza dell'effetto
#                      casuale e dell'errore senza correzione)
#-----------------------------------------------------------

  m=table(dati$sogg)  #vettore con gli m_i per ogni soggetto
  n=length(unique(dati$sogg))
  migliore_ll=-Inf
  migliore_idx=rep(NA,round(n*(1-alpha)))

  for(i in 1:500){
    idx=indici_casuali(alpha,n)
    for(j in 1:15){

      #stime calcolate sulle n*(1-alpha) osservazioni
      sub_dati=sotto_campione(idx,m,n,dati)
      fit=lme(y~x1+x2,random=~1|sogg,data=sub_dati,method="ML")
```

```
      #contributi su tutte le n osservazioni
      lli=ll_contributi(dati,
        cbind(1,dati$x1,dati$x2)%*%summary(fit)$coefficients$fixed,
        as.numeric(VarCorr(fit)[,1]),m,n)

      #continua se non è arrivato a convergenza
      if(all(sort(idx) ==
             sort(order(lli,decreasing=T)[1:round(n*(1-alpha))]))
         ) break else idx=order(lli,decreasing=T)[1:round(n*(1-alpha))]
    }
    #aggiorna il sottocampione H^*
    if(sum(lli[idx])>migliore_ll){
      migliore_ll=sum(lli[idx])
      migliore_idx=idx
    }
  }
  sub_dati=sotto_campione(migliore_idx,m,n,dati)
  fit=lme(y~x1+x2,random=~1|sogg,data=sub_dati,method="ML")

  list(beta=summary(fit)$coefficients$fixed,
       se_beta=summary(fit)$tTable[,2],
       sigma2_j=as.numeric(VarCorr(fit)[,1]))
}
```

## A.3 Stimatore reweighted

```
rwtml=function(tml_hat,dati,alpha=0.25,delta=0.025){
#------------------------------------------------------------
# tml_hat = stimatore TML ottenuto con la funzione tml
# data = oggetto di tipo data.frame() con colonne
```

```
#        y (risposta), x1 e x2 (esplicative),
#        sogg (identificativo del soggetto)
# alpha = parametro alpha dello stimatore TML
# delta = parametro delta per la soglia del RWTML
#
# output: oggetto di tipo lista come nella funzione tml
#-------------------------------------------------------------

  m=table(dati$sogg)  #vettore con gli m_i per ogni soggetto
  n=length(unique(dati$sogg))

  #contributi (basati sul TML) usando la matrice Sigma corretta
  lli=ll_contributi(
    dati,cbind(1,dati$x1,dati$x2)%*%tml_hat$beta,
    tml_hat$sigma2_j,m,n,alpha)

  #seleziona le unità che hanno contributo maggiore della soglia
  soglia=NULL
  for(i in 1:2){
    Sigma_i_hat=Sigma_corretta(alpha,tml_hat$sigma2_j,m[i])
    soglia=c(soglia,-1/2*q(delta,m[i])-
               m[i]/2*log(2*pi)-1/2*log(det(Sigma_i_hat)))
  }
  idx=na.omit(c(1:n)[lli>=soglia])
  sub_dati=sotto_campione(idx,m,n,dati)
  fit=lme(y~x1+x2,random=~1|sogg,data=sub_dati,method="ML")

  list(beta=summary(fit)$coefficients$fixed,
       se_beta=summary(fit)$tTable[,2],
```

```
        sigma2_j=as.numeric(VarCorr(fit)[,1]))

}
```

# Bibliografia


Agresti A. (2015). *Foundations of Linear and Generalized Linear Models*. Wiley, Hoboken.

Diggle P. J., Heagerty P., Liang K.-Y. e Zeger S. L. (2002). *Analysis of Longitudinal Data*. Oxford Univerity Press, Oxford.

Farcomeni A. e Ventura L.(2012). An overview of robust methods in medical research. *Statistical Methods in Medical Research*. **21**, 111–133.

Hampel F. R., Ronchetti E. M., Rousseeuw P. J. e Stahel W. A. (1986). *Robust Statistics. The Approach Based on Influence Functions*. Wiley, New York.

Heritier S., Cantoni E., Copt S. e Victoria-Feser M.-P. (2009). *Robust Methods in Biostatistics*. Wiley, Chichester UK.

Huber P. J. e Ronchetti E. M. (2009). *Robust Statistics*. Wiley, New York.

Pison G., Aelst S. V. e Willems G.(2002). Small sample corrections for lts and mcd. *Metrika*. **55**, 111–123.

Rousseeuw P. J. e Van Driessen K.(1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*. **41**, 212–223.

Ruli E., Farcomeni A. e Ventura L.(2019). Trimmed-maximum likelihood estimation of multivariate linear mixed models. Manoscritto.

Song P. X.-K. (2007). *Correlated Data Analysis: Modeling, Analytics and Applications*. Springer, New York.