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

VEROSIMIGLIANZA ASINTOTICA E PROBLEMI NON REGOLARI DI STIMA: IL COMPORTAMENTO DEL TEST RAPPORTO DI VEROSIMIGLIANZA

RELATORE: Prof. Alessandra Rosalba Brazzale
Dipartimento di Scienze Statistiche

LAUREANDA: Laura Ambrosi
MATRICOLA N ° 1034973

Anno Accademico 2013/2014

Indice

Introduzione

L'inferenza statistica consente di ricavare delle informazioni dai dati disponibili, ipotizzate osservazioni di variabili casuali. L'obiettivo dell'inferenza è arrivare a determinare stimatori puntuali, intervalli di confidenza e test d'ipotesi. I risultati che si traggono da queste procedure sono presi in considerazione solamente se sono associati ad una alta probabilità.

Un approccio molto diffuso per condurre una procedura di inferenza statistica è basato sulla funzione di verosimiglianza, introdotta nel 1922 da Fisher. Essa è un approccio immediato e semplice da implementare nei dati, che gode di buone proprietà a livello campionario e a livello asintotico. Proprio per questo motivo, si è cercato di ampliare il concetto di verosimiglianza anche alle situazioni in cui l'applicabilità di questo metodo risulta più difficoltosa, ad esempio in presenza di dipendenza dei dati, ricercando delle procedure con proprietà simili, come le pseudo-verosimiglianze. La verosimiglianza profilo cerca di semplificare il modello focalizzando l'attenzione solamente ai parametri di interesse, mentre la verosimiglianza ristretta non considera l'intera informazione disponibile dai dati, ma utilizza solo una parte della funzione di verosimiglianza costruita in modo che i parametri di disturbo non abbiano effetto.

Le proprietà della verosimiglianza che maggiormente interessano sono quelle asintotiche, come quelle del primo ordine, che permettono di avere una distribuzione per le stime di massima verosimiglianza e per le statistiche test.

La funzione di verosimiglianza profilo può sostituire quella di verosimiglianza standard solamente in casi particolari, e per questo motivo negli ultimi anni sono stati fatti degli studi per la ricerca di una versione modificata della verosimiglianza profilo, con delle approssimazioni asintotiche di ordine superiore al primo.

La teoria della verosimiglianza e i risultati asintotici connessi si basano su un principio fondamentale: il modello deve essere regolare. Quando ciò non accade, si hanno delle conseguenze particolari sulle stime e sulle procedure per i test, le cui distribuzioni limite non coincidono con quelle standard.

Il modello può non essere regolare sotto molti aspetti, dunque si è cercato di studiare il comportamento delle statistiche test sotto questa ipotesi, in particolare per la situazioni in cui il vero valore del parametro non è un punto interno allo spazio parametrico. Un problema inferenziale che prevede spesso un modello non regolare di questo tipo è la stima delle componenti di varianza, dove capita che il parametro di varianza assume il valore minimo consentitogli, cioè zero.

La tesi è suddivisa in quattro capitoli. Il Capitolo 1 sviluppa l'inferenza di verosimiglianza dal punto di vista frequentista, introducendo la verosimiglianza profilo. Il Capitolo 2 completa la verosimiglianza, sia standard che profilo, con le

teorie asintotiche, principalmente del primo ordine. In seguito, fornisce una versione modifica della verosimiglianza profilo. Il Capitolo 3 si concentra sui modelli non regolari, in particolare quelli con il vero valore del parametro sulla frontiera dello spazio parametrico, come nella stima delle componenti di varianza. Il Capitolo 4 presenta uno studio di simulazione, prendendo in esame il modello a componenti di varianza.

Capitolo 1

Teoria della verosimiglianza

L'obiettivo dell'inferenza statistica è quello di riuscire a determinare delle caratteristiche di una popolazione di riferimento attraverso lo studio di solo una parte di essa (campione), selezionato in modo casuale, effettuando stime puntuali e intervallari, verifiche di ipotesi e previsioni. Attraverso queste procedure si ottengono dei risultati riguardanti la distribuzione sottostante i dati. Le conclusioni ottenute dall'inferenza, tuttavia, non potranno mai essere definite certe, e per poterne giudicare l'affidabilità, ognuna deve essere accompagnata da una determinata misura di incertezza.

Durante l'utilizzo della procedura di inferenza possono sorgere tre tipi diversi di problemi:

- **problemi di specificazione**: sorgono in fase iniziale, quando si individua un modello statistico $\mathcal{F}$ per i dati osservati;
- **problemi di inferenza**: emergono quando si cerca di individuare la funzione di densità di probabilità associata al vero valore del parametro all'interno della famiglia di distribuzioni;
- **problemi di distribuzione**: nascono nella valutazione della statistica campionaria T.

In questo primo capitolo, verranno inizialmente presentati dei concetti fondamentali per un'analisi statistica, per poi passare all'introduzione della teoria della verosimiglianza, l'argomento su cui è incentrato. Nell'ultimo paragrafo viene descritta un'estensione della verosimiglianza, la verosimiglianza ristretta.

1.1 Modello statistico

L'idea che sta alla base dell'inferenza statistica è che i dati osservati $y^{oss} = (y_1^{oss}, ..., y_n^{oss})$ sono una realizzazione casuale di un vettore aleatorio Y. Più precisamente, $Y \sim P_0$, dove P_0 rappresenta una legge di probabilità ignota, che si cerca di ricostruire attraverso l'analisi dei dati, con la ricerca di forme per P_0 che sono compatibili con i dati generati y^{oss}, specificando una famiglia di distribuzioni $\mathcal{F}$.

Quindi, definito $\mathcal{Y}$ lo spazio campionario, un modello statistico è una famiglia di distribuzioni

$$\mathcal{F} = \{P_\theta : \theta \in \Theta\},$$

dove Θ è lo spazio parametrico, ossia tutti i possibili valori che può assumere il parametro θ. Il modello è *correttamente specificato* se $P_0 \in \mathcal{F}$, ovvero se la legge di probabilità che ha generato i dati appartiene al modello statistico ipotizzato.

Il modello statistico $\mathcal{F} = \{f(y;\theta), \theta \in \Theta\}$ è dunque anche rappresentabile come una collezione di funzioni di densità (o di probabilità, a seconda che si tratti di dati continui o di dati discreti).

Il parametro è detto *identificabile* se, definita $f_0(y)$ la funzione di densità (o probabilità) associata a P_0, si verifica che $f_0(y) = f(y, \theta^0)$ per un solo valore di $\theta^0 \in \Theta$, dove θ^0 è il vero valore del parametro. Un *modello parametrico* è una famiglia di distribuzioni che si può descrivere con un numero finito di parametri. In questo caso, l'insieme di tutti i possibili valori che θ può assumere, lo spazio parametrico Θ, è un sottoinsieme di $\mathbb{R}^p$. La dimensione p di parametri può essere 1 (in questo caso il parametro è scalare) o maggiore di 1 (in questo caso $\theta = (\theta_1, \cdots, \theta_p)$).

Un modello statistico parametrico può essere dunque espresso come:

$$\mathcal{F} = \{f(y;\theta), \theta \in \Theta \subseteq \mathbb{R}^p, y \in \mathcal{Y}\}.$$

Un' *ipotesi statistica* $H : f_0(y) \in \mathcal{F}$ è una congettura sulla distribuzione di probabilità, e può essere *semplice* o *composita* a seconda che specifici uno o più modelli probabilistici.

1.1.1 Modello statistico regolare

I modelli statistici parametrici con verosimiglianza regolare godono di molte proprietà, soprattutto a livello asintotico, che facilitano la ricerca di stimatori e statistiche test.

Siano dati uno spazio campionario $\mathcal{Y}$, uno spazio parametrico Θ e un modello statistico $\mathcal{F} = \{f(y;\theta), \theta \in \Theta\}$. Allora, le condizioni di regolarità richieste sono le seguenti:

1. il modello è identificabile, dunque esiste una relazione biunivoca tra lo spazio campionario $\mathcal{Y}$ e lo spazio parametrico Θ, e ad ogni $\theta \in \Theta$ è associato un solo modello probabilistico di $\mathcal{F}$;

2. il modello è correttamente specificato, quindi la legge di probabilità che ha generato i dati appartiene a $\mathcal{F}$ e $\theta^0 \in \Theta$;

3. lo spazio campionario Θ è un sottoinsieme aperto dello spazio euclideo $\mathbb{R}^p$, ovvero θ^0 deve essere un punto interno di Θ;

4. tutte le funzioni di probabilità specificate da $\mathcal{F}$ devono avere lo stesso supporto, e quest'ultimo deve essere indipendente da θ;

5. la funzione di log-verosimiglianza (definita nel §1.2.2) deve essere derivabile almeno fino al terzo ordine, con derivate parziali rispetto a θ continue; questa condizione assicura l'esistenza di un' approssimazione in serie di Taylor e la varianza finita delle derivate di $l(\theta)$.

Quando vengono soddisfatte le precedenti condizioni, si possono sfruttare dei risultati asintotici che verranno descritti dettagliatamente nel Capitolo 2, mentre nel Capitolo 3 si illustrerà quello che accade se non sono verificate queste condizioni.

1.1.2 Statistiche sufficienti

Spesso i risultati ottenuti dall'analisi eseguita sono riassunti in una funzione t che sintetizza i dati osservati. Per poter riassumere i dati, senza però perdere delle informazioni sul parametro di interesse θ, si utilizzano le così dette *statistiche sufficienti*.

Una *statistica* è una funzione del solo campione y. Dunque una statistica non dipende dal modello parametrico, ma solo dal campione osservato di dati y^{oss}. Con una statistica t viene indotta una partizione dello spazio campionario. La legge di probabilità $T = t(Y)$ è chiamata *distribuzione campionaria* della statistica. Una statistica $t(y)$ è detta sufficiente se esistono due funzioni $g(\cdot)$ e $h(\cdot)$ tali che:

$$f(y;\theta) = h(y)g(t(y);\theta), \qquad \forall\theta \in \Theta \quad \text{e} \quad y \in \mathcal{Y}. \tag{1.1}$$

Secondo la fattorizzazione di Neyman-Fisher, una statistica è definita sufficiente per θ se la distribuzione condizionata $f(Y|t(Y) = t)$ non dipende dal parametro, per ogni valore di t.

Inoltre, una statistica $t(y)$ si definisce *statistica sufficiente minimale (s.s.m.)* per θ se è funzione di ogni altra possibile statistica sufficiente, nel senso che può essere ottenuta da ogni altra statistica sufficiente per θ. Il termine *minimale* sta a indicare che non si può ridurre ulteriormente $t(y)$ senza perdere dell'informazione su θ. Diverse statistiche sufficienti inducono la stessa partizione dello spazio campionario $\mathcal{Y}$, e qualsiasi trasformazione biunivoca di una s.s.m. è a sua volta minimale. Per riuscire a riconoscere una s.s.m. bisogna controllare che:

$$\frac{f(y_1,\theta)}{f(y_2,\theta)} = c(y_1, y_2)$$

se e solo se $t(y_1) = t(y_2)$, per $y_1, y_2 \in \mathcal{Y}$, dove $c(y_1, y_2)$ è una quantità costante in θ che dipende esclusivamente da y_1 e y_2. Se questo rapporto è costante, e questo avviene se e solo se le due statistiche sufficienti sono uguali, allora $t(\cdot)$ è s.s.m. per θ.

1.2 Verosimiglianza

1.2.1 Concetti di base

Il metodo della massima verosimiglianza è stato introdotto da Fisher (1922), che ha presentato delle procedure di inferenza statistica.

Sia definito $\mathcal{F}$ un modello statistico parametrico, correttamente specificato, e con funzione di probabilità di densità $f(y;\theta)$ vista esclusivamente in funzione di θ, con y fissato ai dati osservati y^{oss}. La *funzione di verosimiglianza* per y è $L(\theta) = f(y;\theta)$, per $\theta \in \Theta$. Se si ha un campione casuale semplice (c.c.s.) $y = (y_1, \cdots, y_n)$ (assunzione che viene fatta molto spesso), di numerosità n, con distribuzioni marginali $f(y_i;\theta)$, la funzione di verosimiglianza $L(\theta)$ basata sui dati y che va da $\Theta \to \mathbb{R}^+$ è definita come

$$L(\theta) = \prod_{i=1}^{N} f(y_i,\theta).$$

L'obiettivo della funzione di verosimiglianza è di ottenere il maggior numero di informazione sul vero valore del parametro θ^0. La logica dietro la funzione di verosimiglianza è la seguente: in seguito ai dati osservati, $\theta_1 \in \Theta$ è più plausibile di $\theta_2 \in \Theta$ nel modello probabilistico generatore dei dati se $L(\theta_1) > L(\theta_2)$, ossia θ_1 ha più probabilità di essere il vero valore θ^0. Se la distribuzione con parametro θ_1 è più vicina alla distribuzione empirica dei dati rispetto alla distribuzione con θ_2, allora si avrà che la verosimiglianza valutata in θ_1 è maggiore di quella valutata in θ_2.

Il principio di verosimiglianza *debole* si può descrivere come segue: supponiamo di avere due osservazioni (y_1 e y_2) dal modello statistico $\{f(\cdot\,;\theta) : \theta \in \Theta\}$ e la funzione di verosimiglianza, $L(\theta; y)$ basata sull'osservazione di y; allora, se $L(\theta; y_1) = L(\theta; y_2)$, le conclusioni su θ basate sull'osservazione di $Y = y_1$ dovrebbero essere uguali a quelle ottenute osservando $Y = y_2$.

Un metodo di confronto tra la differenza nell'evidenza empirica dei dati y a favore di θ_1 rispetto a θ_2 è il rapporto $L(\theta_1)/L(\theta_2)$, detto *rapporto di verosimiglianza*. I fattori che non dipendono da θ in $L(\theta)$ possono essere eliminati, dato che non cambiano il valore del rapporto di verosimiglianza. Per questo motivo, le funzioni $L(\theta)$ e $cL(\theta)$, dove $c \in \mathbb{R}^+$, sono equivalenti.

Il rapporto di verosimiglianza aiuta a individuare una s.s.m. per θ, quindi una funzione del campione Y che riesce a sintetizzare le osservazioni, senza però perdere informazione sul parametro di interesse. Questo principio viene chiamato *criterio della partizione di verosimiglianza*, dato che y_1 e y_2 appartengono alla stessa curva di livello se e solo se hanno verosimiglianze equivalenti.

1.2.2 Log-verosimiglianza

In genere, a fini pratici, viene utilizzata la trasformazione logaritmica di $L(\theta)$:

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{N} f(y_i; \theta),$$

dove se $L(\theta) = 0$, $l(\theta) = -\infty$, per definizione.

Se si hanno due differenti insiemi di dati x e y, indipendenti tra loro, che contengono entrambi dell'informazione su θ, dato che la loro funzione di densità congiunta è il prodotto delle due marginali, allora la verosimiglianza per θ basata su x e y sarà:

$$L(\theta; x, y) = f(y, \theta) f(x, \theta) = L(\theta, y) L(\theta, x).$$

Una volta ottenuta la funzione di verosimiglianza, si può procedere ad applicare il metodo di inferenza con stime puntuali, intervallari e con i test d'ipotesi.

La *stima di massima verosimiglianza (SMV)*, è quel valore $\hat{\theta} \in \Theta$ che massimizza $l(\theta)$, tale che $L(\hat{\theta}) \geq L(\theta)\ \forall\ \theta \in \Theta$. Se $\hat{\theta} = \hat{\theta}(y)$ esiste ed è unico, $\hat{\theta} = \hat{\theta}(Y)$ è definito *stimatore di massima verosimiglianza*. Dato che il logaritmo è una funzione strettamente monotona, massimizzare $l(\theta)$ equivale a massimizzare $L(\theta)$. Per alcuni modelli, la SMV non è facile da calcolare analiticamente, e si deve ricorrere a delle procedure di calcolo numerico per riuscire a massimizzare $l(\theta)$.

Il vettore che contiene le derivate parziali di primo ordine della funzione di log-verosimiglianza viene chiamato *funzione punteggio*, o *funzione score*:

$$l_*(\theta) = \left(\frac{\partial l(\theta)}{\partial \theta_1}, \cdots, \frac{\partial l(\theta)}{\partial \theta_p} \right) = \left[\frac{\partial l(\theta)}{\partial \theta_r} \right] = [l_r(\theta)]\,. \tag{1.2}$$

Nella maggior parte dei casi, ossia nei modelli regolari, la SMV si trova dall'equazione $l_*(\theta) = 0$, che prende il nome di *equazione di verosimiglianza*, o sistema di equazioni, a seconda che $p = 1$ o $p > 1$ (dove p è il numero di parametri).

La matrice delle derivate seconde della funzione di log-verosimiglianza, cambiata di segno, viene chiamata *matrice di informazione osservata*:

$$j(\theta) = -l_{**}(\theta) = \left[-\frac{\partial^2 l(\theta)}{\partial \theta_r \partial \theta_s} \right]. \tag{1.3}$$

Si può provare che se la matrice delle derivate seconde

$$\left. \frac{\partial^2 l(\theta)}{\partial \theta_r \partial \theta_s} \right|_{\theta = \hat{\theta}} \tag{1.4}$$

è definita negativa $\forall \hat{\theta}$ soluzione di $l_*(\theta) = 0$ e siamo nel caso p=1, questa condizione è sufficiente ad assicurare l'unicità della SMV.

Il valore atteso dell'informazione osservata:

$$i(\theta) = E_\theta(j(\theta)) = \left[\frac{\partial^2 l(\theta)}{\partial \theta_r \partial \theta_s} \right], \tag{1.5}$$

viene chiamata *informazione attesa* o *informazione di Fisher*. Se siamo nel caso di c.c.s. , l'informazione attesa si semplifica a $i(\theta) = n i_1(\theta)$, dove $i_1(\theta)$ è l'informazione attesa per una singola osservazione.

1.3 Pseudo-verosimiglianza

Quando si lavora con un modello che ha un numero di parametri maggiore di 1 ($p > 1$), può capitare che l'interesse risieda solo in un sottovettore di θ o in un solo parametro. Allora $\theta = (\psi, \lambda)$, dove ψ è un vettore di parametri di interesse di lunghezza $1 \leq k < p$ su cui si desidera fare inferenza, mentre λ è un vettore di parametri di disturbo di lunghezza $p - k$, e lo spazio parametrico Θ può essere scritto come $\Psi \times \Lambda$. In genere, ψ è un parametro di dimensione 1 mentre λ è un vettore con dimensioni superiori, e maggiore è la dimensione di λ, maggiore è l'effetto potenziale sulle conclusioni riguardanti ψ.

Quando $\theta = (\psi, \lambda)$ la funzione score può essere suddivisa in due parti, dove la prima è la derivata calcolata rispetto a ψ mentre la seconda è la derivata rispetto a λ:

$$l_*(\theta) = \begin{bmatrix} \frac{\partial l(\theta)}{\partial \psi} \\ \frac{\partial l(\theta)}{\partial \lambda} \end{bmatrix},$$

e anche la matrice di informazione osservata può essere riscritta come matrice a blocchi:

$$j(\theta) = \begin{pmatrix} j_{\psi\psi}(\psi, \lambda) & j_{\psi\lambda}(\psi, \lambda) \\ j_{\lambda\psi}(\psi, \lambda) & j_{\lambda\lambda}(\psi, \lambda) \end{pmatrix}, \tag{1.6}$$

dove $j_{\psi\psi}(\psi,\lambda) = \left[-\frac{\partial^2 l(\psi,\lambda)}{\partial\psi\partial\psi^T}\right]$, e gli altri blocchi sono calcolati in modo del tutto analogo.

Se si fosse a conoscenza del vero valore di λ, λ^0, non si avrebbe problema ad ottenere la funzione di verosimiglianza propria $L(\psi,\lambda^0)$, ma dato che questo è ignoto, bisogna ricorrere alla *pseudo-verosimiglianza*. Quest'ultima è una funzione dei dati osservati e di ψ che può sostituire la verosimiglianza propria per l'inferenza statistica sul parametro di interesse.

1.3.1 Verosimiglianza condizionata e marginale

Un primo tipo di pseudo-verosimiglianza viene ottenuto riducendo il modello originario $\mathcal{F}$ con l'eliminazione del parametro di disturbo dalla funzione di densità considerata, attraverso la marginalizzazione o il condizionamento.

Supponiamo esista la possibilità che la funzione di verosimiglianza possa essere riscritta come prodotto di due fattori, uno dipendente dal parametro di interesse e l'altro dal parametro di disturbo:

$$L(\theta) = L_*(\psi)L_{**}(\lambda).$$

In questo caso, si riesce a estrarre la funzione di verosimiglianza esatta per ψ, sulla quale viene compiuta l'inferenza, ma nella pratica accade raramente.

Un caso un po' meno raro rispetto al precedente si ha quando la funzione di probabilità di densità può essere fattorizzata come segue:

$$f(y;\psi,\lambda) = f(t|s;\psi)f(s;\psi,\lambda), \tag{1.7}$$

con la statistica (t,s) sufficiente per θ.

La statistica S non è sufficiente per λ nel modello generale, ma lo è se viene tenuto fisso ψ.

Una funzione di verosimiglianza per ψ può essere basata sul primo termine della (1.7), dato che non dipende da λ. Il termine $f(s;\psi,\lambda)$ può non essere considerato solo se la perdita di informazione su ψ è trascurabile. Allora, la funzione di verosimiglianza

$$L_C(\psi) = L_C(\psi,t) = f(t|s;\psi)$$

è chiamata *verosimiglianza condizionata* alla statistica $S = s$.

Se invece si suppone che esista una statistica T tale che la funzione di densità possa essere scritta come

$$f(y;\psi,\lambda) = f(t;\psi)f(y|t;\psi,\lambda), \tag{1.8}$$

la funzione di verosimiglianza per ψ può essere calcolata con la distribuzione marginale basata su t. In questo caso, l'eliminazione del termine $f(y|t;\psi,\lambda)$ non comporta una grossa perdita di informazione sul parametro di interesse, e quindi la funzione di verosimiglianza per ψ è

$$L_M(\psi) = L_M(\psi;t) = f(t;\psi),$$

ed è chiamata *verosimiglianza marginale*.

La costruzione di verosimiglianza marginale e condizionata è una procedura che riesce a semplificare il modello in presenza di parametri di disturbo; tuttavia può essere applicata solamente a modelli che hanno una particolare struttura. Inoltre, anche se si riesce a ricavare la verosimiglianza marginale o condizionata, il calcolo, in genere, è abbastanza complesso.

1.3.2 Verosimiglianza profilo

Un metodo ampiamente diffuso per ottenere una verosimiglianza per il parametro ψ è quello di sostituire il parametro di disturbo con una stima consistente di λ che non dipende da ψ.

Questa procedura viene chiamata *verosimiglianza profilo* per ψ, ed è definita come

$$L_p(\psi) = L(\psi, \hat{\lambda}_\psi),$$

dove $\hat{\lambda}_\psi$ è la SMV vincolata di λ, ottenuta fissando ψ, cioè $\hat{\lambda}_\psi = \max_\lambda L(\psi, \lambda)$.

Nonostante la verosimiglianza profilo non sia una verosimiglianza propria, può essere trattata come tale. Per questo motivo, si possono delineare le quantità introdotte precedentemente per la verosimiglianza propria.

La *log-verosimiglianza profilo* per ψ è

$$l_p(\psi) = \log(L_p(\psi))$$

e la *stima di massima verosimiglianza profilo* $\hat{\psi}$ coincide con la SMV di ψ basata su $L(\psi, \lambda)$. Questo risultato deriva direttamente dal fatto che $\hat{\lambda}_{\hat{\psi}} = \hat{\lambda}$.

La derivata prima di $l_p(\psi)$ è la funzione *score profilo*:

$$l_p^*(\psi) = \left[\frac{\partial l_p(\psi)}{\partial \psi}\right],$$

mentre l'*informazione osservata profilo* è

$$j_p(\psi) = \left[-\frac{\partial^2 l_p(\psi)}{\partial \psi \partial \psi^T}\right]$$

e *l'informazione attesa profilo* è il valore atteso di $j_p(\psi)$ calcolato rispetto a ψ. Si può dimostrare che l'inversa dell'informazione osservata profilo è uguale al blocco (ψ, ψ) della (1.6) calcolato in $(\psi, \hat{\lambda}_\psi)$.

La verosimiglianza profilo è largamente adottata perché può essere adoperata su quasi tutti i modelli, dato che non richiede l'estrazione del parametro di disturbo dalla funzione di densità, a differenza della verosimiglianza condizionata e marginale. Inoltre, gode di alcune proprietà che la rendono facilmente utilizzabile per l'inferenza su ψ, ma rimane comunque una verosimiglianza impropria: il valore atteso della quantità $l_p^*(\psi)$ non è pari a zero, come invece accade nella verosimiglianza originale (vedi (2.4)).

Nonostante questo, la verosimiglianza profilo si rivela una tecnica vantaggiosa in presenza di parametri di disturbo che può essere sfruttata per effettuare test e per costruire intervalli per il parametro di interesse ψ, come si vedrà nel Capitolo 2.

1.3.3 Verosimiglianza ristretta

La *verosimiglianza ristretta (REML)* è un'estensione della verosimiglianza che massimizza solo parte della verosimiglianza totale, e per questo si definisce *invariante localmente*. Viene spesso preferita alla massima verosimiglianza quando si stimano i parametri di covarianza nei modelli lineari perché ha il vantaggio di considerare la perdita dei gradi di libertà nello stimare la media, e produce delle equazioni di stima non distorte per i parametri di varianza. Proprio per questo motivo, la REML viene adoperata nella stima delle componenti di varianza nei modelli a effetti misti, come esposto più in dettaglio nel terzo Capitolo.

Inoltre, ha delle proprietà in piccoli campioni migliori rispetto alla verosimiglianza; essa può essere vista come un tipo di verosimiglianza marginale. La procedura REML ha anche più potenza nei test rispetto alla verosimiglianza, e il suo utilizzo non comporta nessuna perdita di informazione per il parametro di interesse.

La REML è considerata un'applicazione della verosimiglianza marginale ai modelli lineari misti. Supponiamo che il modello di partenza sia

$$y = X\beta + Zb + e, \tag{1.9}$$

dove y è un vettore di dimensioni $n \times 1$, X e Z sono due una matrici di dimensioni $n \times p$ e $n \times q$, rispettivamente; infine $b \sim \mathcal{N}_q(0, \Omega_b)$ e $e \sim \mathcal{N}_n(0, \sigma^2 I_n)$. La matrice di varianza $var(y) = Z\Omega_b Z^T + \sigma^2 I_n = \Sigma$, con $\Omega_b = \sigma_b^2 I_q$ che non dipende dal parametro fisso β. L'obiettivo è costruire una verosimiglianza per σ^2 e σ_b^2, non considerando il parametro β. La verosimiglianza REML per σ_b^2 e σ^2 è

$$l_R(\sigma_b^2, \sigma^2) = \frac{1}{2}\log|\Sigma^{-1}| - \frac{1}{2}\log|X^T\Sigma^{-1}X| - \frac{1}{2\sigma^2}(y - X\hat{\beta}_\psi)^T\Sigma^{-1}(y - X\hat{\beta}_\psi) - \frac{n-p}{2}\log\sigma^2, \tag{1.10}$$

dove $\hat{\beta}_{\sigma_b^2}$ è la stima vincolata di β tenuto σ_b^2 fisso.

Se nel modello non sono presenti gli effetti casuali, si ha che $\Sigma = \sigma^2 I_n$.

Smith e Verbyla (1996) hanno dimostrato che la REML può anche essere interpretata come una verosimiglianza condizionata a una determinata statistica sufficiente, per poter eliminare la dipendenza al parametro di disturbo.

La statistica sufficiente per il parametro di disturbo β è del tipo $t = AX^T\Sigma^{-1}y$, per σ_b^2 fisso. Allora, la funzione di verosimiglianza ristretta può essere vista come la verosimiglianza di y condizionata a t.

Se lo stimatore di massima verosimiglianza per λ è una funzione uno-a-uno della statistica t, non si ha nessuna perdita di informazione nello stimare σ_b^2 con la verosimiglianza condizionata rispetto a quella originale.

L'utilizzo delle verosimiglianza ristretta è da preferire alla verosimiglianza originale, soprattutto per la stima delle componenti di varianza nei modelli più complessi.

1.4 Considerazioni conclusive

In questo capitolo è stata introdotta la procedura più diffusa per compiere inferenza nei dati: la verosimiglianza. Sono state introdotte delle quantità fondamentali

collegate ad essa che in seguito si dimostrerà come utilizzare per effettuare stime puntuali e test.

Nell'ultimo paragrafo, si è mostrato come modificare la verosimiglianza in presenza di parametri di disturbo, arrivando a delineare una funzione di verosimiglianza che dipende esclusivamente dal parametro di interesse e che si presenta come una versione ridotta della verosimiglianza originale.

Nel prossimo capitolo, ci si concentrerà sulle proprietà asintotiche delle quantità qui introdotte, si descriveranno in dettaglio i test basati sulla funzione di verosimiglianza e si presenterà una versione modificata della verosimiglianza profilo.

Capitolo 2

Teoria asintotica della verosimiglianza

Nel precedente capitolo è stata introdotta la teoria della verosimiglianza. Un aspetto di importanza fondamentale per la verosimiglianza è lo studio del comportamento asintotico, quando la numerosità è molto alta. In questo capitolo verranno descritte le principali proprietà campionarie delle quantità di verosimiglianza, che valgono solamente sotto condizioni di regolarità.

Innanzitutto, nel primo paragrafo, verranno riportate le statistiche test basate sulla verosimiglianza. Nel secondo paragrafo verrà descritta la teoria asintotica del primo ordine, ossia la distribuzione delle statistiche test e delle quantità descritte nel primo Capitolo quando la numerosità è alta, e tende a $+\infty$. Il terzo paragrafo, invece, fornisce una versione modificata della verosimiglianza profilo da utilizzare in presenza di parametri di disturbo.

2.1 Test statistici

Se si vuole verificare un'ipotesi statistica, sulla base dei dati disponibili, la procedura che bisogna utilizzare è il *test statistico*, che verifica se i dati sono conformi a un sottomodello $\mathcal{F}_0$ di $\mathcal{F}$, ipotizzando che $\theta \in \Theta_0$ (ipotesi nulla), con $\Theta_0 \subset \Theta$, contro l'ipotesi alternativa: $H_1 : \theta \in \Theta \setminus \Theta_0$.

Ciò che permette di stabilire se è più ragionevole l'ipotesi nulla o quella alternativa è la *statistica test*, una funzione $t : \mathcal{Y} \to \mathbb{R}$ che divide lo spazio campionario in due sottoinsiemi disgiunti: R, la regione di rifiuto (o regione critica), e A, quella di accettazione. Se $y \in R$, si dice che il test è *significativo* contro H_0.

Il test statistico non è una procedura del tutto affidabile, perché può essere che il campione sorteggiato cada in R o in A per effetto del caso. Allora, si commette un *errore di I tipo* se si rifiuta H_0 quando questa è vera, e un *errore di II tipo* se si accetta H_0 quando questa è falsa. La massima probabilità di commettere un errore del I tipo è chiamata *livello di significatività*, e si indica con

$$\alpha = \sup_{\theta \in \Theta_0} \Pr_\theta(Y \in R).$$

Nella costruzione di un test, il criterio per determinare la regione di rifiuto R è fissare il livello di significatività α, data una statistica test t. La scelta più diffusa è tenere $\alpha = 0.05$.

Il test può avere regione critica *unilaterale destra*, *unilaterale sinistra* o *bilaterale*, se si rifiuta H_0 per valori grandi, piccoli o sia per valori grandi che piccoli di t, rispettivamente.

Una *regione di confidenza* per θ, basata sui dati y, si può rappresentare come

$$\hat{\Theta}(y) \subset \Theta, \tag{2.1}$$

con la quale si fa corrispondere ai dati y un sottoinsieme di Θ. Se il parametro è scalare, $\hat{\Theta}(y)$ è un intervallo di confidenza. Non sarà mai possibile sapere se il vero valore del parametro è contenuto nell'intervallo. Anche in questo caso, viene scelto il livello di confidenza $(1 - \alpha)$, tale che

$$\Pr_\theta(\theta \in \hat{\Theta}(Y)) = 1 - \alpha \qquad \forall \theta \in \Theta.$$

Si possono costruire delle regioni di confidenza con assegnato livello di significatività $(1 - \alpha)$ con dei test t_θ a livello α, e con ipotesi nulla $H_0 : \theta_0 = \theta$ al variare di $\theta \in \Theta$. Se A_θ è la regione di accettazione, allora

$$\hat{\Theta}(y) = \{\theta \in \Theta : y \in A_\theta\}.$$

2.1.1 Test basati sulla verosimiglianza

La procedura di verosimiglianza prevede la costruzione di statistiche test per verificare l'ipotesi $H_0 : \theta = \theta_0$ contro l'alternativa $H_1 : \theta \neq \theta_0$. Il test che maggiormente viene utilizzato è il *log-rapporto di verosimiglianza (LRT)*:

$$W(\theta) = 2 \log \left\{ \frac{L(\hat{\theta})}{L(\theta_0)} \right\} = 2\{l(\hat{\theta}) - l(\theta_0)\}. \tag{2.2}$$

Questa statistica calcola la distanza tra il valore della stima di massima verosimiglianza $\hat{\theta}$ più plausibile e il valore ipotizzato θ_0, attraverso la verosimiglianza. Se $W(\theta)$ è una funzione monotona crescente di una statistica $t(y)$, la cui distribuzione è nota, si riesce a calcolare facilmente il livello di significatività osservato (o p-value):

$$\alpha^{oss} = Pr(W(\theta_0) \geq W^{oss}(\theta_0)).$$

La maggior parte delle volte, però, la distribuzione esatta non è nota, e dunque si deve ricorrere a delle distribuzioni nulle approssimate a livello asintotico, che verranno descritte in dettaglio nel paragrafo successivo.

Se il parametro è scalare, si può anche verificare $H_0 : \theta = \theta_0$ contro le ipotesi alternative unilaterali $H_1 : \theta > \theta_0$ oppure $H_1 : \theta < \theta_0$, utilizzando la versione unilaterale del test rapporto di verosimiglianza:

$$r(\theta_0) = \text{sgn}(\hat{\theta} - \theta_0)\sqrt{W(\theta_0)}, \tag{2.3}$$

dove $\text{sgn}(\cdot)$ è la funzione segno, tale che $\text{sgn}(x) = 1$ se $x > 0$, $\text{sgn}(x) = -1$ se $x < 0$ e $\text{sgn}(x) = 0$ se $x = 0$.

Anche in questo caso si ricorre a delle approssimazioni asintotiche per la costruzione di regioni di confidenza e per il calcolo del livello di significatività osservato α.

Associate alla statistica $W(\theta)$, si possono calcolare anche altre quantità, come ad esempio la *statistica test di Wald* e il test *score*. La prima è definita come:

$$W_e(\theta) = (\hat{\theta}_n - \theta_0)^T i(\theta_0)(\hat{\theta}_n - \theta_0);$$

questa statistica misura la distanza tra la SMV e il valore che si vuole verificare θ_0 considerando anche l'errore di stima.

La seconda quantità, il test score, o *test di Rao*, si calcola come:

$$W_u = l_*(\theta_0)^T i(\theta_0)^{-1} l_*(\theta_0)$$

Queste due statistiche discendono dalla statistica LRT, e dato che differiscono da quest'ultima solo per delle quantità asintoticamente trascurabili, la loro distribuzione nulla approssimata è uguale a quella di $W(\theta)$.

Se il parametro è scalare, si può verificare anche la versione unilaterale del test, e dunque le due versioni di queste ultime due statistiche saranno:

$$r_e(\theta_0) = \sqrt{i(\theta_0)}(\hat{\theta}_n - \theta_0)$$
$$r_u(\theta_0) = l_*(\theta_0) i(\theta_0)^{-1/2}.$$

2.2 Teoria asintotica del primo ordine

Dal momento che è difficile riuscire ad ottenere la distribuzione esatta delle statistiche che derivano dalla verosimiglianza, si studia il loro comportamento quando la numerosità campionaria è molto alta. Il teorema del limite centrale e la legge dei grandi numeri permettono di ottenere una serie di risultati asintotici, che riguardano la SMV, le stime intervallari e le statistiche test. Questi risultati sono validi esclusivamente per modelli statistici parametrici regolari, quindi si suppone che tutte le condizioni di regolarità elencate nel §1.1 siano verificate.

2.2.1 Proprietà campionarie

Uno stimatore $\hat{\theta}_n$ è detto *non distorto* per θ se $E_\theta(\hat{\theta}_n) = \theta$, $\forall\ \theta \in \Theta$, ed è detto *efficiente* tra i non distorti se ha varianza minima tra tutti gli stimatori non distorti per θ. Inoltre, uno stimatore è detto *consistente* per θ se al divergere della numerosità campionaria, $\hat{\theta}_n \xrightarrow{p} \theta$ sotto θ, ovvero se $\forall\ \varepsilon > 0$ si ha che

$$\lim_{n \to +\infty} \Pr\{|\hat{\theta}_n - \theta| \geq \varepsilon\} = 0.$$

Per poter studiare le proprietà campionarie dello stimatore di massima verosimiglianza, bisogna conoscere alcuni risultati chiave.

Per quanto riguarda la stima di massima verosimiglianza, $\hat{\theta}_n$, si dimostra che è consistente, quindi converge in probabilità a θ, perché la differenza

$$E_{\theta^0}(l(\theta; Y_1)) - E_{\theta^0}(l(\theta^0; Y_1))$$

è negativa per $\theta \neq \theta^0$, dove $l(\theta, Y_1)$ è la log-verosimiglianza per una singola osservazione. Facendo ricorso alla legge dei grandi numeri, per n che tende a $+\infty$ si ha che $\frac{l(\theta)}{n} - \frac{l(\theta^0)}{n}$ converge in probabilità a un valore negativo per $\theta \neq \theta^0$, di conseguenza $l(\theta)$ è grande solo in un intorno di θ^0.

Lo stimatore di massima verosimiglianza viene anche definito asintoticamente efficiente perché ha varianza asintotica minima tra gli stimatori non distorti per θ.

Un'altra proprietà campionaria della verosimiglianza è che la funzione punteggio valutata nel vero valore del parametro ha, componente per componente, valori negativi e positivi che si compensano, dunque

$$E_\theta(l_*(\theta)) = 0 \qquad \forall \quad \theta \in \Theta. \tag{2.4}$$

Inoltre, vale l'identità

$$E_\theta(l_*(\theta) l_*(\theta)^T) = i(\theta) \qquad per \quad ogni \quad \theta \in \Theta,$$

pertanto l'informazione attesa è pari alla matrice di covarianza della funzione score.

2.2.2 Distribuzioni asintotiche

Sfruttando il teorema del limite centrale, per n grande, si possono ottenere una serie di distribuzioni asintotiche, utili nei test e nella costruzione di stime intervallari.

Verosimiglianza

Se vale che $l_*(\hat{\theta}_n) = 0$ e $\hat{\theta}_n - \theta \xrightarrow{p} 0$ sotto θ, allora

$$\hat{\theta}_n \dot{\sim} \mathcal{N}_p(\theta, i(\theta)^{-1}), \tag{2.5}$$

in cui $i(\theta)$ può essere sostituito dalle stime $i(\hat{\theta})$ o $j(\hat{\theta})$. Di conseguenza, si può ottenere la distribuzione dello stimatore di massima verosimiglianza normalizzata:

$$i(\theta)^{1/2}(\hat{\theta}_n - \theta) \dot{\sim} \mathcal{N}_p(0, I_p), \tag{2.6}$$

con I_p matrice identità di dimensione $p \times p$.

Inoltre, la funzione punteggio, per n sufficientemente grande, valutata nel vero valore del parametro, ha distribuzione asintotica

$$l_*(\theta) \dot{\sim} \mathcal{N}_p(0, i(\theta)) \tag{2.7}$$

per p che rappresenta il numero di parametri.

Quando si verifica un'ipotesi, difficilmente si riesce ad recuperare la distribuzione esatta delle statistiche test. Nella maggior parte dei casi, si usufruisce della distribuzione asintotica delle statistiche test.

Il test rapporto di verosimiglianza in (2.2) sotto l'ipotesi nulla si distribuisce come una variabile Chi-quadrato:

$$W(\theta_0) \dot{\sim} \chi^2_p, \tag{2.8}$$

in cui p è sempre il numero di parametri. Si dice allora che $W(\theta)$ è una *quantità asintoticamente pivotale*, perché la sua distribuzione asintotica non dipende da θ. Allora si può costruire delle regioni di confidenza a livello approssimato $(1 - \alpha)$

$$\hat{\Theta}(y) = \{\theta \in \Theta : W(\theta) < \chi^2_{p,1-\alpha}\},$$

dove $\chi^2_{p,1-\alpha}$ è il quantile $(1 - \alpha)$ della distribuzione χ^2_p.

Verosimiglianza profilo

Se il parametro θ può essere suddiviso in (ψ, λ) e si vuole verificare l'ipotesi nulla $H_0 : \psi = \psi_0$ contro $H_1 : \psi \neq \psi_0$, il test log-rapporto di verosimiglianza diventa

$$2\{l(\hat{\psi}, \hat{\lambda}) - l(\psi_0, \hat{\lambda}_{\psi_0})\},$$

che coincide con il *test log-rapporto di verosimiglianza profilo*:

$$W_p(\psi) = 2\{l_p(\hat{\psi}) - l_p(\psi_0)\}. \tag{2.9}$$

La distribuzione asintotica in questo caso sarà

$$W_p(\psi) \dot{\sim} \chi^2_k, \qquad \text{sotto} \quad H_0$$

in cui k è la dimensione del vettore ψ.

Se $k = 1$, ovvero ψ è un parametro scalare, si può usufruire della statistica test *radice con segno profilo* per la verifica dell'ipotesi unilaterale $H_0 : \psi > \psi_0$ o $H_0 : \psi < \psi_0$:

$$r_p(\psi) = \text{sgn}(\hat{\psi} - \psi_0)\sqrt{W_p(\psi)}, \tag{2.10}$$

che si distribuisce invece come una variabile Normale:

$$r(\theta_0) \dot{\sim} \mathcal{N}(0, 1) \qquad \text{sotto} \quad H_0.$$

Queste distribuzioni asintotiche permettono la costruzione di regioni di confidenza per il parametro di interesse; in particolare

$$\hat{\Psi}(y) = \{\psi \in \Psi : W_p(\psi) < \chi^2_{k,1-\alpha}\},$$

è la regione di confidenza bilaterale per ψ a livello $(1 - \alpha)$ da utilizzare se $k > 1$, con $\chi^2_{k,1-\alpha}$ quantile $(1 - \alpha)$ di una Chi-quadrato con k g.d.l., mentre

$$\hat{\Psi}(y) = \{\psi \in \Psi : -z_{1-\frac{\alpha}{2}} < r_p(\psi) < z_{1-\frac{\alpha}{2}}\}$$

è la regione di confidenza a livello $(1 - \alpha)$ se ψ è un parametro scalare, con $z_{1-\frac{\alpha}{2}}$ quantile di $N(0, 1)$.

2.3 Verosimiglianza profilo modificata

La verosimiglianza profilo viene utilizzata per l'inferenza in campioni ad alta numerosità; nei piccoli campioni, invece, maggiore è l'informazione sul parametro di disturbo a disposizione e maggiormente $l_p(\psi)$ viene penalizzata. Emerge allora il bisogno di una quantità modificata di $l_p(\psi)$ da utilizzare nei campioni a bassa numerosità.

La *verosimiglianza profilo modificata* è una funzione del tipo

$$L_{mp}(\psi) = \exp\{l_{mp}(\psi)\} = M(\psi)L_p(\psi). \tag{2.11}$$

La funzione ideale $M(\psi)$ dovrebbe rendere l'inferenza basata su $L_{mp}(\psi)$ equivalente a seconda che si scelga di utilizzare la verosimiglianza marginale o condizionata per ψ. Una funzione che soddisfa questa proprietà è, ad esempio,

$$M(\psi) = |j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{-1/2} \left| \frac{\partial \hat{\lambda}}{\partial \hat{\lambda}_\psi^T} \right|, \tag{2.12}$$

dove $j_{\lambda\lambda}(\psi, \lambda)$ è il blocco (λ, λ) della matrice di informazione osservata (1.6). Il secondo termine della (2.12) è lo Jacobiano che garantisce l'invarianza a trasformazioni della verosimiglianza profilo.

La verosimiglianza ristretta rientra nella categoria della verosimiglianza profilo modificata.

Esempio (Davison (2003)) Nel classico modello lineare $y = X\beta + \varepsilon$, con $\varepsilon \sim N(0, \sigma^2)$ supponiamo che σ^2 sia il parametro di interesse, e che β sia quello di disturbo. La log-verosimiglianza è

$$l(\beta, \sigma^2) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta),$$

e la stima vincolata per β è $\hat{\beta}_{\sigma^2} = (X^T X)^{-1} X^T y$. Si ha che $\hat{\beta} = \hat{\beta}_{\sigma^2}$ perché $\hat{\beta}_{\sigma^2}$ è indipendente da σ^2. Le altre quantità necessarie al calcolo della stima modificata sono

$$j_{\beta\beta}(\sigma^2, \beta) = \sigma^{-2} X^T X, \quad \frac{\partial \hat{\beta}_{\sigma^2}^T}{\partial \hat{\beta}} = I_p, \quad M(\sigma^2) = (\sigma^2)^{p/2} |X^T X|^{-1/2}.$$

Da questo si ricava che

$$l_{mp}(\sigma^2) = -\frac{n-p}{2}(\log \sigma^2 - S^2/\sigma^2)$$

con S^2 stimatore non distorta di σ^2. In questo caso particolare, la log-verosimiglianza profilo modificata corrisponde alla verosimiglianza marginale per σ^2.

In genere è raro riuscire a calcolare il secondo termine della (2.12) e sono poche le volte in cui, come nell'esempio illustrato precedentemente, esso è pari a 1, dato che la stima vincolata di λ non dipende da ψ.

Una strategia che si può adottare è quella di ridurre la dipendenza di $\hat{\lambda}_\psi$ da ψ, per diminuire il peso dello Jacobiano nella determinazione di $M(\psi)$, attraverso la procedura di parametri ortogonali, descritta in Davison (2003), che è un'approssimazione di $L_{mp}(\psi)$.

La *funzione di verosimiglianza profilo aggiustata*, se i parametri ψ e λ sono ortogonali, è

$$L_a(\psi) = |j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{-1/2} L_p(\psi), \tag{2.13}$$

perché lo Jacobiano è pari a 1. Nei modelli in cui questo non accade, si cerca una parametrizzazione per i parametri in modo che $\hat{\lambda}_\psi \doteq \hat{\lambda}$. Allora, vale l'approssimazione di $L_{mp}(\psi)$ a $L_a(\psi)$, con un ordine di errore pari a $O(1^{-1/2})$.

La procedura di ortogonalità dei parametri ha degli svantaggi che non possono essere ignorati, perché l'ortogonalizzazione dei parametri è difficile da realizzare nella pratica.

L'inferenza su ψ viene eseguita trattando $l_{mp}(\psi)$ come una verosimiglianza propria. La *stima di massima verosimiglianza modificata* $\hat{\psi}$ si ottiene massimizzando la (2.11), e si possono costruire degli intervalli di confidenza incentrati sul parametro di interesse sfruttando le usuali approssimazioni alla Normale.

Inoltre, si possono costruire dei test profilo modificati per ψ_{mp} con la versione modificata del test rapporto di verosimiglianza

$$W_{mp}(\psi) = 2\{l_{mp}(\hat{\psi}_{mp}) - l_{mp}(\psi)\}$$

che segue l'usuale approssimazione alla variabile Chi-quadrato. È anche disponibile la versione modificata della statistica radice con segno profilo:

$$r_p^*(\psi) = r_p(\psi) + r_p(\psi)^{-1} \log \frac{q(\psi)}{r_p(\psi)},$$

con $q(\psi)$ quantità opportuna scelta a seconda dell'ordine di errore con cui si desidera che $r_p^*(\psi)$ si approssimi alla $N(0,1)$. Esempi di proposte per $q(\psi)$ sono presenti in Barndorff-Nielsen e Cox (1994) e in Severini (2000).

2.4 Considerazioni conclusive

In questo secondo capitolo, sono state introdotte le principali statistiche test per la verifica di ipotesi, costruite con le funzioni di verosimiglianza standard e profilo. Inoltre, sono stati illustrati i risultati asintotici che si sfruttano per fare inferenza sul parametro di interesse, dato che è molto difficile calcolare la distribuzione esatta delle statistiche test.

La verosimiglianza profilo in campioni esigui produce dei scarsi risultati, e per questo nell'ultimo paragrafo è stata presentata una versione modificata di $l_p(\psi)$, che sembra preferibile alla verosimiglianza profilo.

Nel capitolo successivo si tratteranno i modelli in cui alcune condizioni di regolarità non sono verificate e tutti i risultati presentati finora non sono più efficaci.

Capitolo 3

Problemi di stima non regolare

Nei precedenti capitoli sono stati presentati dei risultati e degli argomenti che valgono solamente se il modello di partenza è regolare. In questo capitolo si discuterà invece che cosa succede quando vengono a mancare delle condizioni di regolarità, soffermandosi soprattutto al caso in cui il parametro di interesse si trova sulla frontiera dello spazio parametrico.

L'ultimo paragrafo si concentrerà sulla stima delle componenti di varianza, specificando anche la stima REML, particolarmente utile nel caso di modelli non regolari.

Anche lo studio di simulazione, che sarà presentato nel prossimo capitolo, è stato impostato su modelli non regolari.

3.1 Modelli non regolari

Le condizioni di regolarità elencate nel §1.1 assicurano la validità delle approssimazioni asintotiche standard dello stimatore di massima verosimiglianza e delle statistiche test (presentate nel Capitolo 2).

Nella realtà, può capitare che non siano verificate alcune condizioni; se non c'è una corrispondenza tra il modello e lo spazio parametrico Θ è probabile che esistano più valori di θ^0 a cui converge $\hat{\theta}$ e il modello non è più identificabile. Ad esempio, siano $y_1, \cdots, y_n$ realizzazioni di n variabili di Poisson indipendenti $Y_1, ..., Y_n$ con medie positive

$$E(Y_j) = \begin{cases} \lambda_1 & j = 0, ..., \tau, \\ \lambda_2 & j = \tau + 1, ..., n. \end{cases}$$

Qui τ può assumere solo valori discreti $0, ..., n$. Se τ assume uno dei valori estremi (0 o n), nel modello rimane solo un λ. Se, invece, si pone $\lambda_1 = \lambda_2$ si ottiene lo stesso modello per qualsiasi valore assunto da τ, e la condizione 1 di regolarità non è più valida.

Rientrano in questa categoria anche i modelli *parametro ridondanti*, in cui non è possibile stimare tutti i parametri nel modello. Di conseguenza si ha un modello non identificabile, che può essere riscritto come funzione di un numero di parametri minore e la cui matrice di Informazione $i(\theta)$ è singolare, quindi non ammette l'inversa, e l'approssimazione (2.5) non è più valida.

Altro caso, è la mancata validità della condizione 5, che prevede che la log-verosimiglianza sia derivabile fino al terzo ordine, con derivate parziali, rispetto a θ, continue. Dato che sotto c.c.s. $i(\theta) = ni_1(\theta)$, l'informazione attesa cresce all'infinito per $n \to \infty$, e quindi si ha che $i(\theta) \to \infty$. Ad esempio, sia data una sequenza $Y_0, ..., Y_n$ in modo che, dati i valori di $Y_0, ..., Y_{j-1}$, la distribuzione di Y_j sia una Poisson di media θY_{j-1}, con $E\{Y_0\} = \theta$. Allora

$$l(\theta) = \sum_{j=0}^{n} Y_j log\theta - \theta \left(1 + \sum_{j=0}^{n-1} Y_j\right), \qquad J(\theta) = \theta^{-2} \sum_{j=0}^{n} Y_j.$$

e l'informazione attesa è $i(\theta) = \theta^{-2}(\theta + \cdots + \theta^{n+1})$. Se $\theta \geq 1$ si ha che $i(\theta) \to \infty$ per $n \to \infty$, altrimenti questo non accade, e la conseguenza è che lo stimatore di massima verosimiglianza non è consistente e neanche asintoticamente normale.

Se invece si ipotizza il modello sbagliato per i dati, vale a dire che si modella $f(y;\theta)$ ai dati quando il vero modello è $g(y)$, l'approssimazione asintotica dello stimatore di massima verosimiglianza è

$$\hat{\theta}_n \dot{\sim} \mathcal{N}_p(\theta_g, i_g(\theta_g)^{-1} K(\theta_g) i_g(\theta_g)^{-1}),$$

dove θ_g è il vero valore del parametro che minimizza la distanza di *Kullback-Leibler*, definita come

$$D(f_\theta, g) = \int \log \left\{ \frac{g(y)}{f(y;\theta)} \right\} g(y)\, dy$$

e che è una sorta di distanza tra la distribuzione ipotizzata e quella vera. Inoltre,

$$K(\theta_g) = n \int \frac{\partial l(\theta)}{\partial \theta} \frac{\partial l(\theta)}{\partial \theta^T} g(y)\, dy$$

$$i_g(\theta_g) = -n \int \frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^T} g(y)\, dy.$$

Ovviamente, se $g(y) = f(y;\theta)$, allora $\theta_g = \theta$, vero valore del parametro, e $J_g(\theta) = I_g(\theta_g) = I(\theta)$ e $\hat{\theta}$ si distribuisce come in (2.5).

Anche la distribuzione della statistica test rapporto di verosimiglianza è diversa dal risultato asintotico standard, dato che

$$W(\theta) \doteq n(\hat{\theta}_n - \theta_g)^T i_g(\theta_g)(\hat{\theta}_n - \theta_g),$$

ha distribuzione χ^2_p, ma con media $tr(i_g(\theta_g)^{-1} K(\theta_g))$.

Un'altra condizione di regolarità che può non essere rispettata è quella relativa allo spazio campionario Θ, cioè quando il vero valore del parametro non è un suo punto interno, ma risiede sulla frontiera. Lo stimatore di massima verosimiglianza non ha distribuzione limite Normale con media θ, e le statistiche test non hanno le approssimazioni asintotiche usuali. È di questo tipo il così detto problema delle *componenti di varianza*, in cui si vuole verificare la presenza o meno degli effetti casuali. Questo scenario verrà descritto in modo più dettagliato successivamente.

Un modello può anche non essere regolare se il supporto di Y dipende dal parametro (è violata la condizione 4). Un esempio molto diffuso di questa situazione

è la distribuzione Uniforme $U(0, \theta)$, con $\theta > 0$ ignoto:

$$f(y;\theta) = \begin{cases} \frac{1}{\theta} & \text{se} \quad 0 \leq y \leq \theta \\ 0 & \text{altrimenti} \end{cases}$$

Allora la funzione di verosimiglianza per θ è

$$L(\theta) = \prod_{i=1}^{n} \theta^{-1} I(0 < Y < \theta)$$

e la SMV non si può trovare derivando $l(\theta)$ e ponendo la funzione score pari a zero. In questo caso, il valore che massimizza la funzione di verosimiglianza è $\hat{\theta} = \max_{1 \leq i \leq n} Y_i$, e la distribuzione alla Normale non è più valida.

3.2 Modelli con il vero parametro sulla frontiera

Se il vero valore del parametro θ^0 non è un punto interno allo spazio parametrico Θ, ma si trova sulla frontiera, le usuali approssimazioni non valgono più.

Self e Liang (1987) hanno affrontato questo argomento in un'ottica generale, partendo dal lavoro di Moran (1971) e fornendo le basi per molti altri articoli pubblicati successivamente, che affrontano delle situazioni più specifiche (come quello di Crainiceanu e Ruppert (2004) e di Kopylev e Sinha (2010)).

Nel §3.3 viene presentata in dettaglio la situazione in cui si stimano le componenti di varianza e il vero valore del parametro si trova sulla frontiera.

3.2.1 Casi generali

Nel caso di modelli non regolari, i risultati inferenziali basati sul metodo della massima verosimiglianza non valgono più. In questi casi, la distribuzione asintotica dei test basati sulla verosimiglianza hanno ricevuto particolare interesse, e molti autori hanno dedicato studi e ricerche a questo argomento. Se si ipotizza una approssimazione asintotica non corretta per le statistiche test, probabilmente si otterranno dei p-value errati e delle procedure di inferenza sbagliate. Proprio per questo motivo, è molto importante recuperare la distribuzione asintotica di queste statistiche.

Ipotizziamo che $\theta = (\psi, \lambda)$ e ciò che si vuole verificare è $H_0 : \psi = 0$. Shapiro, nel 1988, ha dimostrato che la distribuzione nulla dei test di verosimiglianza, che sotto H_0 e per n grande differiscono solo per una quantità trascurabile, è una somma pesata di variabili Chi-quadrato, i cui pesi variano da caso a caso e vanno calcolati numericamente. Ad esempio, Stram e Lee (1994) hanno ottenuto che la distribuzione asintotica sotto l'ipotesi nulla quando si verifica la presenza di m verso $m + 1$ effetti casuali correlati in modelli lineari misti è

$$0.5\chi^2_m + 0.5\chi^2_{m+1}.$$

Oppure, se si vogliono verificare congiuntamente k parametri $\psi_j = 0$ contro $\psi_j > 0$, per $j = 1, ..., k$, la mistura sotto l'ipotesi nulla avrà la forma di una somma pesata

di variabili Chi-quadrato, del tipo

$$\sum_{j=0}^{k} 2^{-k} \binom{k}{j} \chi_j^2.$$

Questa distribuzione può essere anche calcolata come la somma pesata dei p-value di ogni variabile χ^2 che contribuisce alla somma.

Chen e Liang (2010) hanno esaminato il comportamento del test rapporto di pseudo-verosimiglianza con il vero valore del parametro sulla frontiera.

Ipotizzando sempre che lo spazio parametrico Θ possa essere partizionato come $\Psi \times \Lambda$ si può essere interessati a verificare

$$H_0 : \psi = \psi_0, \tag{3.1}$$

utilizzando la statistica test basata sulla verosimiglianza profilo $L_p(\psi, \hat{\lambda}_\psi)$. In questo caso la verosimiglianza profilo è molto utile se non si riesce a eliminare il parametro di disturbo dalla funzione tramite condizionamento o fattorizzazione.

La statistica test per verificare la (3.1) è la statistica di log-verosimiglianza profilo $W_p(\psi)$. Supponiamo che il vero valore del parametro ψ si trovi sulla frontiera dello spazio parametrico, mentre il parametro di disturbo sia un punto interno a Θ. Nell'articolo di Cheng e Liang (2010) viene dimostrato un risultato di consistenza per stimatore di massima pseudo-verosimiglianza per θ con ordine di errore $O(n^{-1/2})$.

Utilizzando la notazione di Self e Liang (1987), suddividiamo il vettore dei parametri in quattro categorie:

$$\theta = (\psi_1, ..., \psi_m; \psi_{m+1}, ...\psi_k, \lambda_1, ..., \lambda_q, \lambda_{q+1}, ..., \lambda_{p-k}), \tag{3.2}$$

dove i primi m parametri interesse hanno il vero valore sulla frontiera; i successivi $k - m$ parametri di interesse hanno il vero valore interno a Θ; le successive q coordinate di θ sono parametri di disturbo con il vero valore sulla frontiera; infine, gli ultimi $p - k - q$ parametri di disturbo hanno il vero valore interno allo spazio parametrico. A seconda dei valori assunti da m e q, la distribuzione della statistica rapporto di verosimiglianza varia, e diventa più complicata all'aumentare di questi due valori.

Se non si hanno vari valori di parametri sulla frontiera, e la configurazione di θ è $(0, k, 0, p - k)$, con $k - p$ parametri di disturbo e k parametri di interesse interni a Θ, la distribuzione del test rapporto di verosimiglianza profilo è quella usuale di χ_k^2.

Se la configurazione di θ è $(1, 0, 0, p-1)$, quindi si ha un parametro di interesse con il vero valore sulla frontiera, e i $p - 1$ parametri di disturbo con il vero valore interno a Λ. Allora, la distribuzione asintotica di $W_p(\psi)$ è una mistura di variabili Chi-quadrato: $W_p(\psi) \sim 0.5\chi_0^2 + 0.5\chi_1^2$.

Se siamo in presenza di un parametro di interesse con vero valore sulla frontiera, più un parametro di interesse con il vero valore interno a Ψ, come pure i parametri di disturbo $((1, 1, 0, p - 2))$, la distribuzione limite per il test è $W_p(\psi) \sim 0.5\chi_1^2 + 0.5\chi_2^2$. Nel problema delle componenti di varianza equivale a verificare la nullità di un effetto casuale e congiuntamente che la media sia pari a una costante μ_0.

Possono capitare anche delle situazioni in cui più di un vero valore del parametro (sia di interesse che di disturbo) si trovi sulla frontiera: la distribuzione di $W_p(\psi)$ risulta sempre una mistura di Chi-quadrato, ma più complessa da calcolare, perché lo spazio parametrico Θ viene suddiviso in più regioni.

3.3 Componenti di varianza

Per poter spiegare l'effetto di un fenomeno spesso si tiene conto di più fattori di variabilità, in modo da riuscire ad analizzare singolarmente il contributo di ognuno sulla varianza totale. Questi diversi tipi di variabilità vengono chiamati *componenti di varianza*. Le diverse modalità del fattore di interesse sono chiamate trattamenti, mentre i livelli del fattore secondario sono chiamati blocchi.

I modelli lineari a effetti misti (LMM) vengono utilizzati per stimare questo tipo di dati raggruppati e considerano sia effetti casuali che effetti fissi. Essi riescono a considerare più tipi di variabilità, sia quella all'interno dei gruppi che quella all'esterno.

Sono del tipo

$$\mathbf{Y} = X\mu + Z_1 b_1 + \cdots + Z_s b_s + \varepsilon, \tag{3.3}$$

dove $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I_N)$, $b_s \sim \mathcal{N}(0, \sigma_s^2 I_n)$ indipendenti tra loro e con μ vettore di p effetti fissi. Dunque i vettori b_s sono gli effetti casuali. Il vettore dei parametri da stimare è $\theta = (\mu; \sigma^2) = (\mu; \sigma_1^2, ..., \sigma_s^2, \sigma_\varepsilon^2)$. Il numero di gruppi è a, mentre la numerosità per gruppo è n (campione bilanciato); la numerosità totale è $N = a \times n$.

Allora $E(\mathbf{Y}) = X\mu$ e $Var(\mathbf{Y}) = V = \sum_{r=1}^{s+1} \sigma_r^2 J_r$, dove $J_r = Z_r^T Z_r$ per $r = 1, \cdots, s$ e $J_{s+1} = I_N$ dato che $\sigma_{s+1}^2 = \sigma_\varepsilon^2$.

In questo contesto, si può essere interessati a verificare la presenza o meno di uno o più effetti casuali:

$$H_0 : \sigma_s^2 = 0 \qquad \text{contro} \qquad H_1 : \sigma_s^2 > 0.$$

o comunque a ottenere delle stime per le componenti di varianza.

Spesso accade che quando si vuole verificare la presenza di un effetto casuale sul modello, il vero valore del parametro non è un punto interno di Θ, e dunque ci si ritrova a lavorare con un modello non regolare. In questo caso si può scegliere di utilizzare la massima verosimiglianza, anche se è maggiormente consigliato l'utilizzo della massima verosimiglianza ristretta, dato che, sotto H_1, è meno probabile che le stime si trovino sulla frontiera, e il test rapporto di verosimiglianza ristretta ha più potenza.

Si può dimostrare che la REML ha una probabilità di stima della varianza pari a zero minore rispetto al metodo di massima verosimiglianza, e questa affermazione è vera soprattutto sui campioni più piccoli, dato che asintoticamente le stime di massima verosimiglianza e quelle di verosimiglianza ristretta coincidono.

Se si è interessati a verificare solamente alcune componenti di σ^2, si può utilizzare la log-verosimiglianza profilo per i parametri di interesse. Allora $\sigma^2 = (\psi, \lambda)$, dove ψ sono le componenti di varianza di interesse e λ contiene la media μ e le restanti componenti di σ^2.

Considerando la stima vincolata $\tilde{\mu} = \mu(\psi, \tilde{\lambda}_\psi)$, la log-verosimiglianza profilo per ψ è

$$L_p(\psi) = L_p(\psi, \tilde{\lambda}_\psi) = -\frac{1}{2} y^T \tilde{P} y - \frac{1}{2} \log |\tilde{V}|, \tag{3.4}$$

con $\tilde{P} = P(\psi, \lambda)$, dove $P = V^{-1} - V^{-1}X(X^T V^{-1} X)^{-1} X^T V^{-1}$ e $\tilde{V} = V(\psi, \lambda)$ valutati in $\tilde{\lambda}$.

La versione REML della verosimiglianza profilo in (3.4), quindi la log-verosimiglianza profilo REML, è:

$$l_R(\psi) = -\frac{1}{2} Y^T \bar{P} Y - \frac{1}{2} \log |\bar{V}| - \frac{1}{2} \log |X^T \bar{V}^{-1} X|, \tag{3.5}$$

con $\bar{P}, \bar{V}$ valutati in $(\psi, \bar{\lambda}_(\psi))$.

Sotto il modello con un solo effetto casuale è possibile calcolare anche la probabilità esatta di stima con il parametro sulla frontiera: Stern e Welsh (2000) affermano che le stime REML hanno una minore probabilità di essere pari a zero rispetto a quelle di massima verosimiglianza.

Il test che viene utilizzato per la verifica delle ipotesi sulle componenti di varianza è il test rapporto di verosimiglianza (LRT) che, in questo modello, non segue più l'usuale distribuzione asintotica Chi-quadrato.

Di seguito, vengono presi in esame dei casi specifici di (3.3): il modello a una via e il modello a due vie.

3.3.1 Modello a una via

Sia definito il modello a un solo effetto casuale:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \qquad i = 1, ...a, \quad j = 1, ..., n \tag{3.6}$$

dove y_{ij} è la j-ma osservazione nella i-ma classe, α_i è l'effetto casuale sulla variabile y di essere osservata su un'unità che appartiene alla i-ma classe, e ε_{ij} è l'errore residuo. La scrittura matriciale della precedente equazione è

$$\mathbf{y} = \mathbf{X}\mu + Z\alpha + \boldsymbol{\varepsilon}, \tag{3.7}$$

con $\mathbf{X} = \mathbf{1}_N$, $\mathbf{Z} = (\mathbf{I}_n \otimes \mathbf{1}_a)$, dove $\otimes$ è il prodotto di Kroneker, che moltiplica la matrice a sinistra del prodotto per ogni elemento di quella che si trova a destra e $\mathbf{1}_N$ è un vettore colonna con tutti elementi pari a 1 di lunghezza N. Il vettore $\alpha = [\alpha_1, \cdots, \alpha_a]$ contiene tanti effetti casuali quante sono le classi. Nel modello (3.7) si ipotizza che

$$\begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_\alpha I_a & 0 \\ 0 & \sigma^2_\varepsilon I_N \end{bmatrix} \right).$$

Allora, la distribuzione di $\mathbf{y}$ è una Normale multivariata:

$$\mathbf{y} \sim \mathcal{N}(\mu \mathbf{1}_N, \mathbf{V}), \qquad \mathbf{V} = \text{diag}(\sigma^2_\alpha \mathbf{J}_n + \sigma^2_\varepsilon \mathbf{I}_n),$$

dove $\mathbf{J}_n = (\mathbf{1}_n \otimes \mathbf{1})_n$ con $\mathbf{1}_n$ colonna di 1 di lunghezza n.

La funzione di verosimiglianza del modello appena presentato è

$$L(\mu, \mathbf{V}|\mathbf{y}) = \frac{\exp[-\frac{1}{2}(\mathbf{y} - \mu\mathbf{1_n})^T\mathbf{V}^{-1}(\mathbf{y} - \mu\mathbf{1}_n)]}{(2\pi)^{\frac{1}{2}N}|\mathbf{V}|^{1/2}}. \tag{3.8}$$

Una volta fatto il logaritmo di $L(\mu, \mathbf{V}|\mathbf{y})$ e ipotizzando che il campione sia bilanciato si ottiene la funzione di log-verosimiglianza

$$\begin{aligned} l(\mu, \mathbf{V}|\mathbf{y}) = &-\frac{1}{2}N\log 2\pi - \frac{1}{2}a(n-1)\log\sigma_\varepsilon^2 - \frac{1}{2}a[\log(\sigma_\varepsilon^2 + n\sigma_\alpha^2)] \\ &-\frac{1}{2\sigma_\varepsilon^2}\left\{SSE + \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + n\sigma_\alpha^2}[SSA + N(\bar{y}_{..} - \mu)^2]\right\}, \end{aligned} \tag{3.9}$$

dove $SSA = n\sum_i(\bar{y}_{i\cdot} - \bar{y}_{..})^2$, $SSE = \sum_i\sum_j(y_{ij} - \bar{y}_{i\cdot})^2$ e $\bar{y}_{i\cdot} = \sum_j \frac{y_{ij}}{n}$, e con $\bar{y}_{..} = \sum_i\sum_j \frac{y_{ij}}{N}$ la media totale di tutte le osservazioni.

Derivando $l(\mu, \mathbf{V}|\mathbf{y})$ rispetto al parametro $\theta^T = (\mu, \sigma_\alpha^2, \sigma_\varepsilon^2)^T$ e ponendo le derivate uguali a zero si ottengono le soluzioni per le due varianze:

$$\dot{\sigma}_\varepsilon^2 = \frac{SSA}{a-1} = MSA \tag{3.10}$$

$$\dot{\sigma}_\alpha^2 = \frac{(1 - 1/a)MSA - MSE}{n}, \tag{3.11}$$

dove $MSE = \frac{SSE}{a(n-1)}$, mentre $\dot{\mu} = \bar{y}_{..}$ è la SMV che si ottiene dalla (3.9), facendone la derivata rispetto a μ.

Le stime di massima verosimiglianza non corrispondono esattamente alle soluzioni delle equazioni, perché possono dar luogo a valori negativi. In genere, ottenere delle stime negative per i parametri di varianza è un segnale di modello adattato ai dati errato, o sta a indicare che vero valore di σ_α^2 è nullo. La probabilità che questo accada è

$$\begin{aligned} \Pr\{\dot{\sigma}_\alpha^2 < 0\} &= \Pr\{MSA < MSE\} \\ &= \Pr\left\{(F_{a-1}^{a(n-1)} > \frac{(1 - 1/a)}{1 + n\tau})\right\}, \end{aligned}$$

dove $F_{a-1}^{a(n-1)}$ è una variabile F di Fisher con $a(n-1)$ g.d.l. al numeratore e $a-1$ g.d.l. al denominatore e $\tau = \sigma_\alpha^2/\sigma_\varepsilon^2$.

Allora, le stime di massima verosimiglianza sono le espressioni in (3.10) e (3.11), ma tenendo conto che σ_α^2 non può essere negativa:

$$\hat{\sigma}_\alpha^2 = \begin{cases} \dot{\sigma}_\alpha^2 & \text{se} \quad \dot{\sigma}_\alpha^2 \geq 0 \\ 0 & \text{se} \quad \dot{\sigma}_\alpha^2 < 0 \end{cases}$$

$$\hat{\sigma}_\varepsilon^2 = \begin{cases} MSE & \text{se} \quad \dot{\sigma}_\alpha^2 \geq 0 \\ \frac{(SSA+SSE)}{N} & \text{se} \quad \dot{\sigma}_\alpha^2 < 0 \end{cases}$$

Se si vuole verificare l'ipotesi nulla $H_0 : \sigma^2_\alpha = 0$ contro $H_1 : \sigma^2_\alpha > 0$, il test che si potrebbe impiegare è quello del log-rapporto di verosimiglianza:

$$LRT = 2(\sup l(\mu, \mathbf{V}|\mathbf{y}) - \sup_{H_0} l(\mu, \mathbf{V}|\mathbf{y})),$$

che nel modello a componenti di varianza non ha più la distribuzione asintotica standard.

Nel caso di modello a una via, le distribuzioni finite e asintotiche possono essere calcolate esplicitamente. Crainiceanu e Ruppert (2004) hanno ricavato che la distribuzione in campioni finiti della statistica log-rapporto di verosimiglianza (LRT) è

$$LRT \stackrel{\mathcal{D}}{=} N \log(X_{a-1} + X_{N-a}) - \inf_{d \geq 0} \left\{ N \log \left(\frac{X_{a-1}}{1+d} + X_{N-a} \right) + a \log(1+d) \right\},$$

dove X_{a-1} e X_{N-a} sono variabili casuali indipendenti con distribuzione χ^2_{a-1} e χ^2_{N-a}, rispettivamente, e che quella asintotica è

$$LRT \stackrel{\mathcal{D}}{\to} \{X_{a-1} - a - a \log(X_{a-1}/a)\} \mathbb{1}(X_{a-1} > a),$$

dove $\mathbb{1}(X_{a-1} > a)$ è la funzione indicatrice che vale 1 se $X_{a-1} > a$ e 0 altrimenti.

Questa distribuzione asintotica differisce da quella ricavata da Self e Liang (1987) che hanno ottenuto una approssimazione per LRT mistura di variabili Chi-quadrato

$$LRT \sim 1/2\chi^2_0 + 1/2\chi^2_1. \tag{3.12}$$

Le due approssimazioni non coincidono perché Self e Liang hanno posto l'assunzione restrittiva che la variabile risposta $\mathbf{Y}$ possa essere partizionata in sottovettori i.i.d. , con il numero di sottovettori che tende a $+\infty$. L'approssimazione determinata da Crainiceanu e Ruppert è quindi valida nei casi più generali, anche quando non vengono ipotizzati dati indipendenti e identicamente distribuiti.

Self e Liang hanno anche considerato il modello in cui si vuole verificare congiuntamente l'ipotesi che la media sia pari a μ_0 e che la varianza di un effetto casuale sia pari a zero, lasciando gli altri parametri (come la varianza dell'errore e quella degli altri effetti casuali) liberi da vincoli. In questo caso $H_0 : \mu = 0, \sigma^2_\alpha = 0$ e $H_1 : \mu \neq 0, \sigma^2_\alpha > 0$ e la distribuzione della statistica test risulta essere

$$LRT \sim 1/2\chi^2_1 + 1/2\chi^2_2.$$

REML

La procedura più consigliata per la stima delle componenti di varianza è quella REML, perché tiene conto della perdita di gradi di libertà dovuta alla stima delle componenti fisse, massimizzando quella parte di verosimiglianza che non dipende dagli effetti fissi. In altre parole, la REML stima le componenti di varianza basandosi sui residui calcolati modellando i minimi quadrati ordinari solo sulla parte fissa del modello.

Nel caso del modello a un effetto casuale, la REML si trova massimizzando quella parte di verosimiglianza che non dipende da μ. La verosimiglianza REML di (3.8) è

$$L(\mu, \sigma^2_\varepsilon, \sigma^2_\alpha|\mathbf{Y}) = L(\mu|\bar{y}_{..})L(\sigma^2_\varepsilon, \sigma^2_\alpha|SSA, SSE), \tag{3.13}$$

con

$$L(\mu|\bar{y}_{..}) = \frac{\exp\left[-\frac{(\bar{y}_{..}-\mu)^2}{2\lambda/N}\right]}{(2\pi)^{1/2}(\lambda/N)^{1/2}}.$$

La log-verosimiglianza si trova considerando solamente il secondo termine della (3.13):

$$l_R(\mu, \sigma^2_\varepsilon, \sigma^2_\alpha|\mathbf{Y}) = \log L(\sigma^2_\varepsilon, \sigma^2_\alpha|SSA, SSE) = -\frac{1}{2}(N-1)\log 2\pi - \frac{1}{2}\log N$$

$$-\frac{1}{2}a(n-1)\log\sigma^2_\varepsilon - \frac{1}{2}(a-1)\log\lambda - \frac{SSE}{2\sigma^2_\varepsilon} - \frac{SSA}{2\lambda}.$$

con $\lambda = \sigma^2_\varepsilon + n\sigma^2_\alpha$.

Le soluzioni delle equazioni REML portano a

$$\dot{\sigma}^2_{\varepsilon,R} = \frac{SSE}{a(n-1)} = MSE$$
$$\dot{\sigma}^2_{\alpha,R} = \frac{1}{n}(MSA - MSE),$$

con $MSA = SSA/(a-1)$.
La probabilità che σ^2_α sia negativa è

$$\begin{aligned}\Pr\{\dot{\sigma}^2_{\alpha,R} < 0\} &= \Pr\{MSA < MSE\} \\ &= \Pr(\mathcal{F}^{a(n-1)}_{a-1} > 1 + n\tau).\end{aligned}$$

Le stime REML invece, sempre tenendo conto che la varianza dell'effetto casuale non può essere nulla, sono:

$$\hat{\sigma}^2_{\varepsilon,R} = \begin{cases} MSE & \text{se} \quad \dot{\sigma}^2_{\alpha,R} > 0 \\ \frac{SSA+SSE}{N-1} & \text{se} \quad \dot{\sigma}^2_{\alpha,R} \leq 0 \end{cases}$$

$$\hat{\sigma}^2_{\alpha,R} = \begin{cases} \dot{\sigma}^2_{\alpha,R} & \text{se} \quad \dot{\sigma}^2_{\alpha,R} > 0 \\ 0 & \text{se} \quad \dot{\sigma}^2_{\alpha,R} \leq 0 \end{cases}$$

La distribuzione asintotica sotto l'ipotesi nulla della statistica *test rapporto di verosimiglianza ristretta (RLRT)* è la stessa di LRT in (3.12) (quando si ipotizzano dati i.i.d. per tutti i valori dei parametri), perché asintoticamente le approssimazioni REML e quelle di massima verosimiglianza non variano.

L'uso della statistica RLRT è appropriato solamente quando gli effetti fissi sono gli stessi sia sotto H_0 che sotto H_1, dato che nella funzione di verosimiglianza ristretta (3.13) compaiono solo nel termine che non viene considerato per il calcolo della log-verosimiglianza.

Quando i dati non sono i.i.d., nel modello con una sola componente di varianza, in cui si vuole verificare $H_0 : \sigma^2_\alpha = 0$, una statistica adatta è

$$RLRT = -2\{l_R(\mu, \sigma^2_\varepsilon|\mathbf{Y}) - l_R(\mu, \sigma^2_\varepsilon, \sigma^2_\alpha|\mathbf{Y})\}.$$

Crainiceanu e Ruppert hanno calcolato la distribuzione asintotica di questa statistica sotto l'ipotesi nulla:

$$RLRT \xrightarrow{\mathcal{D}} [X_{I-1} - (I-1) - (I-1)\log\{X_{I-1}/(I-1)\}]\, \mathbb{1}(X_{I-1} > I-1),$$

con $\mathbb{1}(X_{I-1} > I-1)$ funzione indicatrice che vale 1 se $X_{I-1} > I-1$. La probabilità asintotica di ottenere un valore pari a 0 è $\Pr(X_{I-1} < I-1)$.

Inoltre, hanno trovato la distribuzione in campioni finiti di $RLRT$ utilizzando la scomposizione spettrale, sempre per il modello con un effetto casuale. Un altro importante risultato descritto nel loro articolo è il calcolo della probabilità di ottenere un valore pari a 0 per $RLRT$, ossia di avere la varianza dell'effetto casuale sulla frontiera:

$$\Pr\left(\frac{\sum_{s=1}^{I} \mu_{s,N} w_s^2}{\sum_{s=1}^{N-p} w_s^2} \leq \frac{1}{N-p}\sum_{s=1}^{I} \mu_{s,N}\right),$$

dove $\mu_{s,N}$ sono gli autovalori della matrice $\Sigma^{1/2} Z^T P_0 Z \Sigma^{1/2}$ e con $P_0 = I_N - X(X^TX)^{-1}X^T$, $w_1, ..., w_s$ sono osservazioni indipendenti da $N(0,1)$.

3.3.2 Modello a due vie

Supponendo che i dati possano essere classificati da due fattori, il modello a due vie con effetti incrociati è rappresentabile come

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \tag{3.14}$$

dove y_{ijk} è la k-ma osservazione con l'effetto α i-mo e quello β j-mo, con $i = 1, ..., a$, $j = 1, ..., b$ e $k = 1, ...n$, e γ_{ij} interazione tra l'effetto α_i e quello β_j. L'interazione può esserci o non esserci nel modello; se non ci fosse, σ^2_γ non ci sarebbe e il modello risulta più semplice.

Tutte e tre gli effetti, compresa l'interazione, sono casuali, con media pari a 0 e varianza positiva:

$$\begin{aligned} E(\alpha_i) &= E(\beta_j) = E(\gamma_{ij}) = 0 \\ Var(\alpha_i) &= \sigma^2_\alpha, \quad Var(\beta_j) = \sigma^2_\beta, \quad Var(\gamma_{ij}) = \sigma^2_\gamma. \end{aligned}$$

Inoltre, vale che

$$\begin{aligned} cov(\alpha_i, \beta_j) &= cov(\alpha_i, \gamma_{ij}) = cov(\alpha_i, \varepsilon_{ij}) = 0 \\ cov(\beta_j, \gamma_j) &= cov(\beta_j, \varepsilon_{ij}) = 0 \\ cov(\gamma_{ij}, \varepsilon_{ij}) &= 0 \end{aligned}$$

e viene assunta normalità.

La funzione di log-verosimiglianza è uguale al logaritmo della (3.8), solo che $\mathbf{V} = Var(\mathbf{y})$ dipende anche dalla varianze $\sigma^2_\beta, \sigma^2_\gamma$, oltre che da $\sigma^2_\alpha, \sigma^2_\varepsilon$.

Il modello in (3.14) può essere di molti tipi, a seconda che ci sia o meno l'interazione, o a seconda che ci siano uno o due effetti casuali. Inoltre, il modello in (3.14) è definito a effetti incrociati, ma può anche esserci il modello a effetti *nidificati (nested)*:

$$y_{ijk} = \mu + \alpha_i + \beta_{ij} + \varepsilon_{ijk},$$

con β_{ij} nidificato in α_i.

In Searle *et al.* (1992) vengono riportate alcune stime della varianza in forma chiusa, anche per il modello a due vie. Queste differiscono a seconda che si tratti di un modello a effetti casuali misti, oppure di un modello con la presenza dell'interazione.

Per quanto riguarda il test rapporto di verosimiglianza, la configurazione del parametro cambia a seconda di H_0, e quindi pure la distribuzione asintotica di LRT, che diventa più complessa man mano che aumentano i parametri che si trovano nella frontiera dello spazio parametrico.

Se, ad esempio, si vuole verificare

$$H_0 : \sigma^2_\alpha = 0 \quad \text{contro} \quad H_1 : \sigma^2_\alpha > 0, \tag{3.15}$$

lasciando σ^2_β e σ^2_γ liberi da vincoli, la configurazione del parametro è (1,0,0,4), perché il parametro di interesse è σ^2_α con il vero valore sulla frontiera, e i parametri di disturbo sono $\sigma^2_\beta, \sigma^2_\gamma, \sigma^2_\varepsilon, \mu$ sono punti interni a Θ. Allora

$$LRT \dot{\sim} 0.5\chi^2_0 + 0.5\chi^2_1.$$

Se invece il test da verificare è

$$H_0 : \sigma^2_\alpha = 0, \mu = \mu_0 \quad H_1 : \sigma^2_\alpha > 0, \mu \neq \mu_0,$$

la configurazione del parametro è (1,1,0,3) e la statistica test ha distribuzione asintotica

$$LRT \dot{\sim} 0.5\chi^2_1 + 0.5\chi^2_2.$$

Nello studio di simulazione effettuato, verrà presa in considerazione il primo tipo di ipotesi.

3.4 Considerazioni conclusive

In questo capitolo, sono stati presentati i criteri di stima per i modelli non regolari, in particolare per quelli con il vero valore del parametro di interesse nella frontiera di Θ.

L'argomento focale di questo paragrafo è il test del log-rapporto di verosimiglianza, una statistica test che ha ricevuto molta attenzione e che negli anni è stato oggetto di studi approfonditi, soprattutto nella valutazione del suo comportamento in condizioni non standard. Il vantaggio dell'utilizzo di questo test è la facilità di implementazione, sempre se la funzione di verosimiglianza dei dati è agevole da calcolare.

In particolare, è stato esaminato il comportamento del test log-rapporto di verosimiglianza profilo, particolarmente utile quando si vuole semplificare il modello e restringere l'inferenza a solo il vettore di interesse.

L'ultimo paragrafo è dedicato alla stima delle componenti di varianza, considerando che nella pratica la maggior parte delle volte ci si trova a dover lavorare con dati multidimensionali e con modelli non regolari. Per questo modello sono stati descritti due metodi di stima, quello di massima verosimiglianza e quello REML, anche se il metodo di stima originale per le componenti di varianza è il metodo ANOVA (vedi Searle *et al.* (1992)).

Il prossimo capitolo espone lo studio di simulazione che è stato fatto partendo da un modello a effetti casuali, considerando proprio il caso della stima delle componenti di varianza e concentrandosi sul test log-rapporto di verosimiglianza. Verranno messe a confronto le due metodologie appena descritte, massima verosimiglianza classica e ristretta.

Capitolo 4

Studio di simulazione

4.1 Descrizione dello studio

Per riuscire a valutare il comportamento del test rapporto di verosimiglianza (LRT) e per poterlo poi confrontare con i risultati teorici standard, sono stati condotti due studi di simulazione di tipo Monte Carlo. La differenza tra le due simulazioni è nel tipo di modello generatore dei dati:

- nella prima simulazione, viene preso in esame solamente un fattore, con i trattamenti, che vengono assegnati casualmente alle unità sperimentali (*disegno completamente randomizzato*);

- nella seconda, vengono considerati due fattori differenti, uno con i trattamenti e l'altro con j blocchi, e ogni livello di un fattore è combinato con tutti i livelli dell'altro (*disegno fattoriale completo*).

Supponendo che $i = 1, ..., I$ e $j = 1, ..., J$, il totale delle osservazioni è $N = I * J$. Le simulazioni sono state condotte facendo aumentare di volta in volta la numerosità campionaria, cambiando sia il numero di trattamenti che il numero delle unità sperimentali.

Dopo aver ottenuto i dati, è stato calcolato il test LRT per verificare la presenza degli effetti casuali sul modello, in cui viene esaminata l'ipotesi nulla $H_0 : \sigma_a^2 = 0$ contro l'alternativa $H_1 : \sigma_a^2 > 0$. Il test LRT a cui si farà sempre riferimento in questo Capitolo è quello basato sulla verosimiglianza profilo, perché sia nella prima simulazione che nella seconda c'è la presenza di parametri di disturbo, come il valore atteso μ e la varianza dell'errore σ_ε^2.

Le simulazioni sono state fatte considerando 10.000 ripetizioni, ottenendo così un insieme di valori campionari sufficientemente grande per validare la distribuzione asintotica del test, e poi sono state confrontate con i quantili di una variabile χ_1^2. Il test è stato calcolato sia sotto l'ipotesi alternativa che sotto l'ipotesi nulla, quindi i dati sono stati generati da un modello sia con effetto casuale ($\sigma_a^2 > 0$), che senza ($\sigma_a^2 = 0$). Nel secondo caso, come descritto in precedenza, la distribuzione asintotica del test LRT è risultata una mistura di Chi-quadrato, ovvero $LRT \sim 0.5\chi_0^2 + 0.5\chi_1^2$.

4.2 Simulazioni

Gli studi di simulazione sono basati su 10.000 simulazioni, per $I = 10, 20, 50, 100$ e per $J = 5, 15, 20, 30$ rispettivamente, dunque per $N = 50, 300, 1000, 3000$, e sono stati fatti utilizzando l'ambiente di calcolo **R**.

Per la stima dei modelli si è scelto di utilizzare la procedura di massima verosimiglianza (*ML*), che verrà poi confrontata con i risultati di stima ottenuti con la procedura di verosimiglianza ristretta (*REML*) nel §4.3.

Di seguito, sono riportati i grafici di LRT, che facilitano la visualizzazione della distribuzione della statistica test. La simulazione in entrambi i casi è stata impostata generando dei dati da un modello, prima sotto l'ipotesi nulla e poi sotto l'alternativa. Una volta generati i dati, è stata verificata l'ipotesi di presenza dell'effetto casuale α_i, e sono stati raccolti i valori di LRT.

4.2.1 Un effetto casuale

Il primo tipo di simulazione è stato realizzato partendo dal modello con un solo effetto casuale, e il modello di riferimento è come quello riportato in (3.6):

$$y_{ij} = \mu + \alpha_i + e_{ij}, \qquad i = 1, ..., I, \quad j = 1, ..., J \tag{4.1}$$

dove μ è l'intercetta fissa per tutte le osservazioni, α_i sono gli effetti casuali del trattamento i, che si distribuiscono come $N(0, \sigma_a^2)$ tra loro indipendenti, e che sono indipendenti dagli errori e_{ij}, anch'essi distribuiti come $N(0, \sigma_e^2)$.

Si è deciso di fissare $\sigma_a^2 = 2.5$, $\sigma_e^2 = 4$ e $\mu = 5$. Per questo tipo di modello le stime sono disponibili in forma chiusa, come visto nel §3.3.1, e il calcolo delle stime delle componenti di varianza e della funzione di log-verosimiglianza è stato effettuato senza utilizzare alcuna procedura numerica.

Risultati sotto l'ipotesi nulla

I risultati ottenuti, per quanto riguarda lo scenario contemplato sotto l'ipotesi nulla, sono riportati nelle Figure 4.1–4.3 e nelle Tabelle 4.1–4.2.

Come si può osservare, sul grafico 4.1, si confronta la statistica LRT con i quantili di una χ_1^2: il modello a numerosità più alta (linea a puntini) si avvicina di più alla bisettrice del grafico, che rappresenta i quantili teorici di una χ_1^2. In questo caso, nei quantili di LRT sono considerati esclusivamente quelli positivi, ovvero quelli che si presume si distribuiscano come una χ_1^2.

Nelle Figura 4.2 sono riportati due tipi di grafici per ogni simulazione: sul primo c'è la rappresentazione della distribuzione di LRT, dove sulla sinistra è raffigurato l'istogramma, mentre sulla destra c'è la funzione di ripartizione empirica; nel secondo tipo di grafico c'è il confronto tra i quantili di LRT e quelli di una variabile χ_1^2 esclusi i valori nulli di LRT.

La funzione di ripartizione empirica riesce a mostrare in modo esaustivo come circa la metà dei valori siano nulli, mentre l'altra metà ha la tipica funzione di ripartizione di una χ_1^2.

Invece, la Figura 4.3 riporta il confronto tra i quantili di LRT (considerando solo i valori positivi) e i quantili di una variabile Chi-quadrato con un grado di libertà: c'è la conferma che la parte non nulla di LRT segue una distribuzione χ^2_1.

Nella Tabella 4.2, è riportato il numero di valori nulli di LRT ottenuti con le diverse numerosità. Come ci si poteva aspettare, man mano che la numerosità campionaria aumenta, il numero di zeri diminuisce, avvicinandosi sempre più al valore teorico (in questo caso 5.000, dato che le replicazioni è 10.000).

La Tabella 4.1 riporta i quantili principali della statistica LRT, confrontando quantili teorici ($0.5\chi^2_0 + 0.5\chi^2_1$) ed empirici (LRT).

Dopo aver simulato una distribuzione mistura di Chi-quadrato, di numerosità n=10.000, i quantili teorici sono stati calcolati partendo da

$$Pr(0.5\chi^2_0 + 0.5\chi^2_1 \leq u) = p,$$

dove u è il quantile che si vuole avere e $0 \leq p \leq 1$ è la probabilità di interesse. Attraverso qualche passaggio, si ottiene che

$$F^{-1}(2p - 1) = u,$$

dove $F^{-1}(x)$ è l'inversa della funzione di ripartizione di una χ^2_1.

La Tabella mostra come i quantili empirici si avvicinino a quelli teorici, e questo è proprio ciò che ci si aspettava, data l'alta numerosità del campione. In questo caso, è stato confrontato il campione con numerosità maggiore.

Tabella 4.1: Tabella di confronto tra quantili, per N=3.000 e con 10.000 replicazioni.

p	**0.5**	**0.75**	**0.9**	**0.95**	**0.975**	**0.99**
teorici	0	0.4729	1.6540	2.6275	3.6810	5.1755
empirici	0	0.319	1.364	2.383	3.568	4.913

Tabella 4.2: Numero di zeri ottenuti per LRT nella prima simulazione, suddivisi per la numerosità.

I	**J**	**N**	**N. ZERI**	**PROP.**
10	5	50	6216	62.16%
20	15	300	6027	60.27%
50	20	1000	5708	57.08%
100	30	3000	5460	54.60%

Risultati sotto l'ipotesi alternativa

Sotto l'ipotesi alternativa nel modello da cui vengono campionati i dati è presente la varianza della componente casuale ($\sigma^2_a > 0$). Quindi, la distribuzione in questo caso non è più una χ^2_1.

I grafici dei risultati ottenuti sono riportati nella Figura 4.4.

L'istogramma del primo campione mostra una distribuzione asimmetrica, con molti valori minori di 5; man mano che aumenta la numerosità le distribuzioni campionarie si simmetrizzano.

I valori di LRT aumentano al crescere della numerosità (l'intervallo di LRT passa da [0;50] a [600-1800] all'incirca), mentre la funzione di ripartizione empirica è abbastanza regolare, in tutti e quattro i casi.

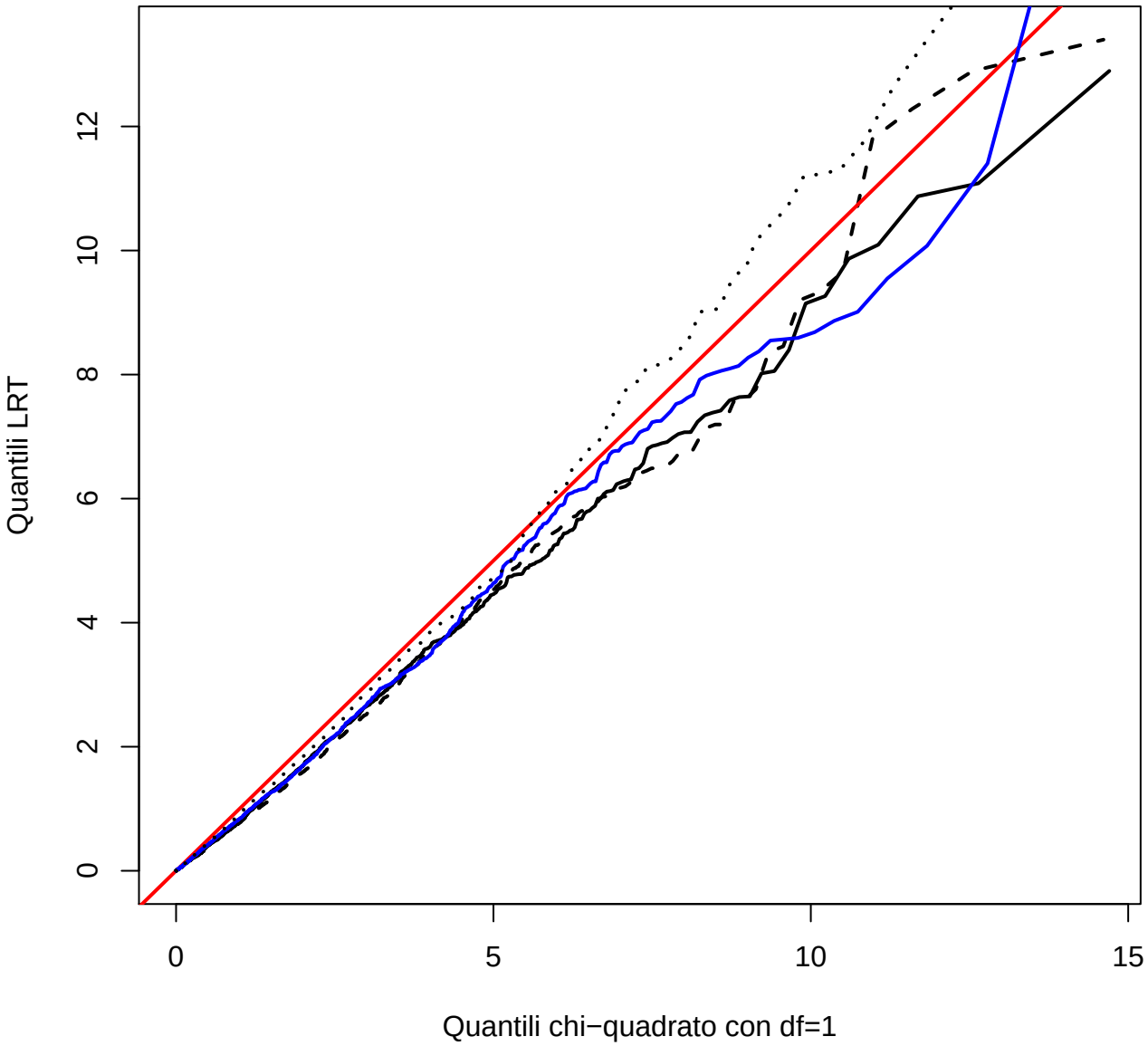

Figura 4.1: Grafico quantile-quantile per il modello ANOVA a una via. La linea rossa rappresenta i quantili teorici della χ^2_1. La linea tratteggiata è quella a numerosità più bassa ($N = 50$), la linea nera continua è quella che riguarda il modello a numerosità 300, quella blu rappresenta la statistica per il modello con 1.000 osservazioni, e infine la linea a puntini rappresenta la distribuzione per il modello a numerosità 3.000.

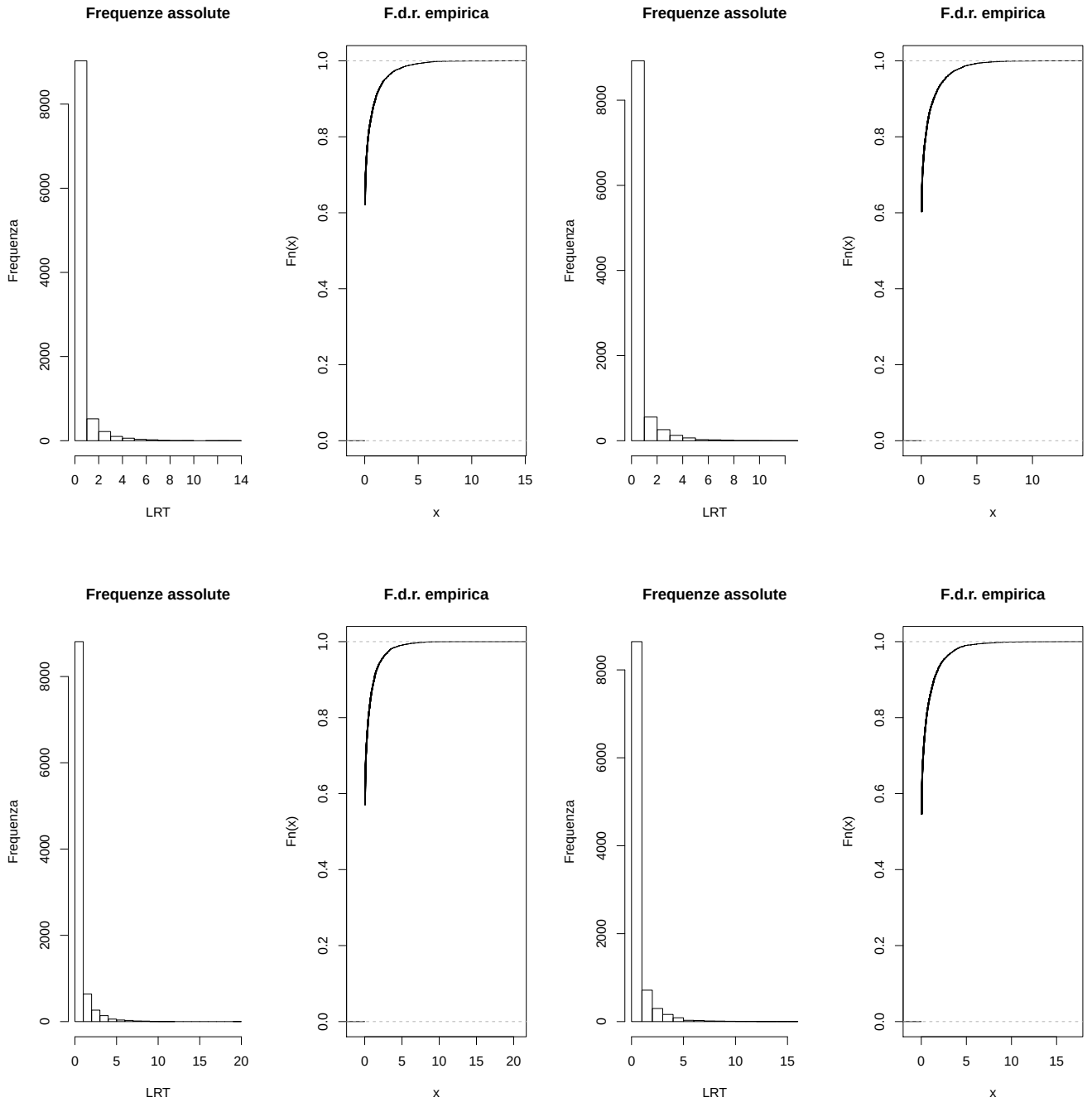

Figura 4.2: Simulazione per verificare la presenza di un effetto casuale: il grafico in alto a sinistra rappresenta il modello con $N = 50$, quello successivo è del modello con $N = 300$; il grafico in basso a sinistra rappresenta il modello per $N = 1.000$ e il grafico in basso a destra raffigura il modello con $N = 3.000$. Il numero di replicazioni è 10.000.

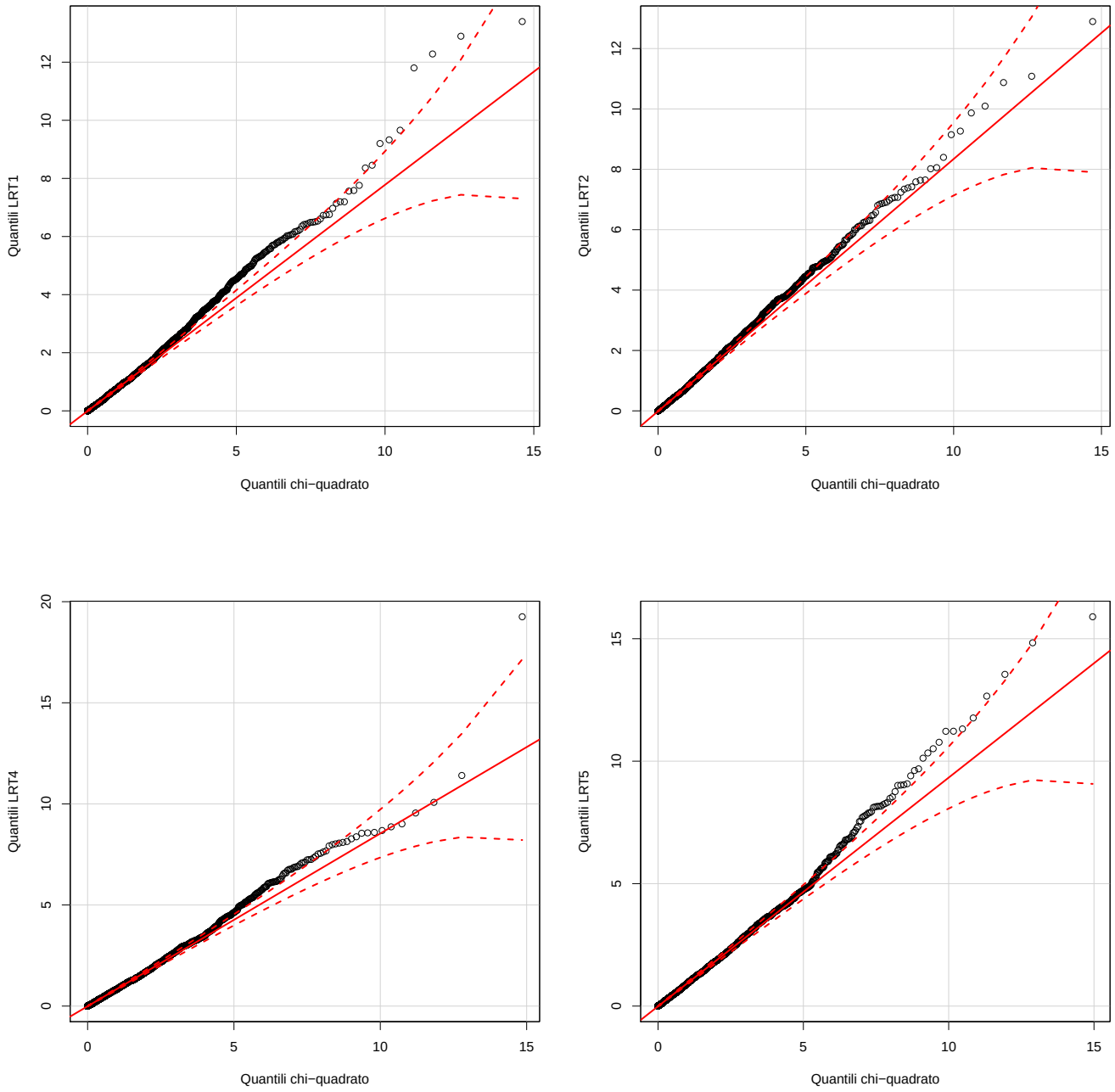

Figura 4.3: Confronto dei quantili di LRT e quantili di χ^2_1, per verificare la presenza di un effetto casuale:il grafico in alto a sinistra rappresenta il modello con $N = 50$, quello successivo è del modello con $N = 300$; il grafico in basso a sinistra rappresenta il modello per $N = 1.000$ e il grafico in basso a destra raffigura il modello con $N = 3.000$. Il numero di replicazioni è 10.000.

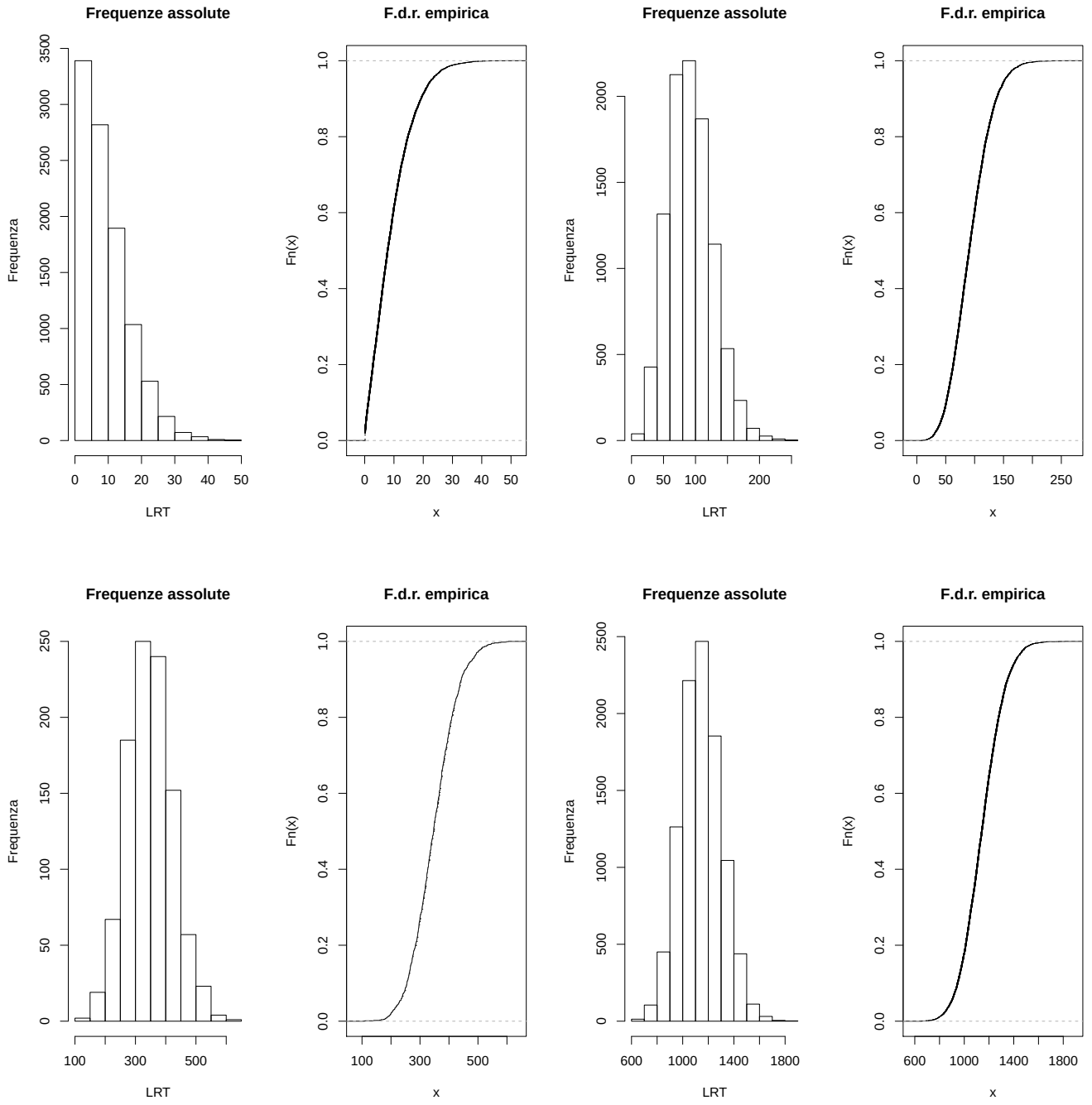

Figura 4.4: Simulazione per verificare la presenza di un effetto casuale sotto l'ipotesi alternativa: il grafico in alto a sinistra rappresenta il modello con $N = 50$, quello successivo è del modello con $N = 300$; il grafico in basso a sinistra rappresenta il modello per $N = 1.000$ e il grafico in basso a destra raffigura il modello con $N = 3.000$. Il numero di replicazioni è 10.000.

4.2.2 Due effetti casuali

Il secondo tipo di simulazione, invece, prevede due termini casuali nel modello di partenza, quindi come nel modello in (3.18), ma senza considerare l'interazione:

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}, \qquad i = 1, ..., I, \quad j = 1, ..., J \tag{4.2}$$

con μ l'intercetta, comune a tutte le osservazioni, α_i è l'effetto casuale del trattamento i che si distribuisce sempre come una $N(0, \sigma_a^2)$, mentre β_j è l'effetto casuale del trattamento j, che si distribuisce come $N(0, \sigma_b^2)$, ed è indipendente da α_i, $\forall i, j$. I due effetti casuali α_i e β_j sono indipendenti dall'errore e_{ij}, che si distribuisce anch'esso come una Normale di media 0 e varianza σ_e^2.

Nella seconda simulazione, sono stati fissati $\sigma_a^2 = 2.5$, $\sigma_b^2 = 3.1$, $\sigma_e^2 = 4$ e $\mu = 3$. L'ipotesi nulla rimane sempre la stessa: $H_0 : \sigma_a^2 = 0$, lasciando il parametro σ_b^2 senza vincoli, se non quello di essere maggiore di zero, essendo la varianza del secondo effetto casuale.

Dato che le formule della stima in forma chiusa sono più difficili ottenere, è stata utilizzata una libreria apposita per la stima del modello con effetti casuali: il pacchetto *lme4*.

Risultati sotto l'ipotesi nulla

Nel caso di simulazione sotto l'ipotesi nulla, dunque in assenza dell'effetto casuale, la stima del modello ha prodotto dei valori del test negativi molto piccoli, probabilmente dovuti all'approssimazione, perché il valore del test rapporto di verosimiglianza non può essere negativo, per definizione. La quantità di questi valori non è trascurabile, dato che rappresenta circa il $20 - 30\%$ del totale dei dati.

Nelle Figure 4.5–4.7 e nelle Tabelle 4.3–4.4, sono riportati i grafici ottenuti con il modello a due componenti casuali.

Nella Tabella 4.3, c'è il confronto tra quantili teorici ed empirici, per la simulazione con numerosità maggiore. In questo caso, la statistica LRT simulata ha valori molto più grandi verso la coda della distribuzione, perché i quantili tendono ad avvicinarsi alla distribuzione teorica a probabilità inferiori, mentre con l'aumento della probabilità i quantili di LRT sono molto alti.

La Figura 4.5 riflette questo andamento, perché illustra il confronto tra i quantili teorici e quelli empirici: la distribuzione di comparazione è sempre la χ_1^2, come riportato in Self e Yang (1987). Anche in questo caso, sulle ordinate la statistica LRT ha solamente i valori non nulli. A differenza del modello con un effetto casuale, LRT è più vicino alla bisettrice con i quantili teorici per valori più bassi, mentre tende a discostarsi per i valori più alti. In generale, tutte le simulazioni tendono ad avvicinarsi ai quantili teorici.

In Tabella 4.3 sono riportati il numero di valori nulli di LRT, sempre suddivisi per numerosità, con le percentuali sul totale delle ripetizioni. All'aumentare del numero si osservazioni, c'è la tendenza dell'abbassamento del numero di zeri, anche se in questo caso il campione con il numero minore non è quello con la più alta numerosità, bensì quello con $N = .1000$.

Nella Figura 4.6, sono riportati l'istogramma e la funzione di ripartizione empirica di LRT.

Questi grafici sono molto simili a quelli presentati per il modello con un effetto casuale. La funzione di ripartizione empirica riflette quella di una mistura di Chi-quadrato: per metà è zero e per l'altra metà è simile a una funzione di ripartizione di χ^2_1. Gli istogrammi di LRT sono caratterizzati da un numero elevato di zeri e di valori molto bassi .

Infine, nella Figura 4.7, sono riportati i valori di LRT diversi da zero, con i valori di una χ^2_1. La statistica rispecchia l'andamento di una variabile Chi-quadrato con 1 grado di libertà, dato che tutti i valori rientrano nelle bande di confidenza.

Tabella 4.3: Tabella di confronto tra quantili, per N=3000 e 10.000 replicazioni.

p	**0.5**	**0.75**	**0.9**	**0.95**	**0.975**	**0.99**
teorici	0	0.4729	1.6540	2.6275	3.6810	5.1755
empirici	0	0.394	1.571	2.591	3.546	5.114

Tabella 4.4: Numero di zeri ottenuti per LRT nella seconda simulazione, suddivisi per numerosità.

I	**J**	**N**	**N. ZERI**	**PROP.**
10	5	50	6524	65.24%
20	15	300	6168	61.68%
50	20	1000	5850	58.50%
100	30	3000	5943	59.43%

Risultati sotto l'ipotesi alternativa

Sotto l'ipotesi alternativa c'è la presenza dell'effetto casuale, e la distribuzione del test rapporto di verosimiglianza non è più una mistura di distribuzioni.

Nella Figura 4.8 sono illustrati l'istogramma e la funzione di ripartizione empirica della statistica test sotto l'ipotesi alternativa, nel modello a due vie.

L'intervallo in cui varia LRT aumenta con l'aumentare della numerosità campionaria (i valori di LRT passano da [0,40] a [600,1400]), e la distribuzione campionaria tende a diventare simmetrica, come nel modello con $N = 3.000$.

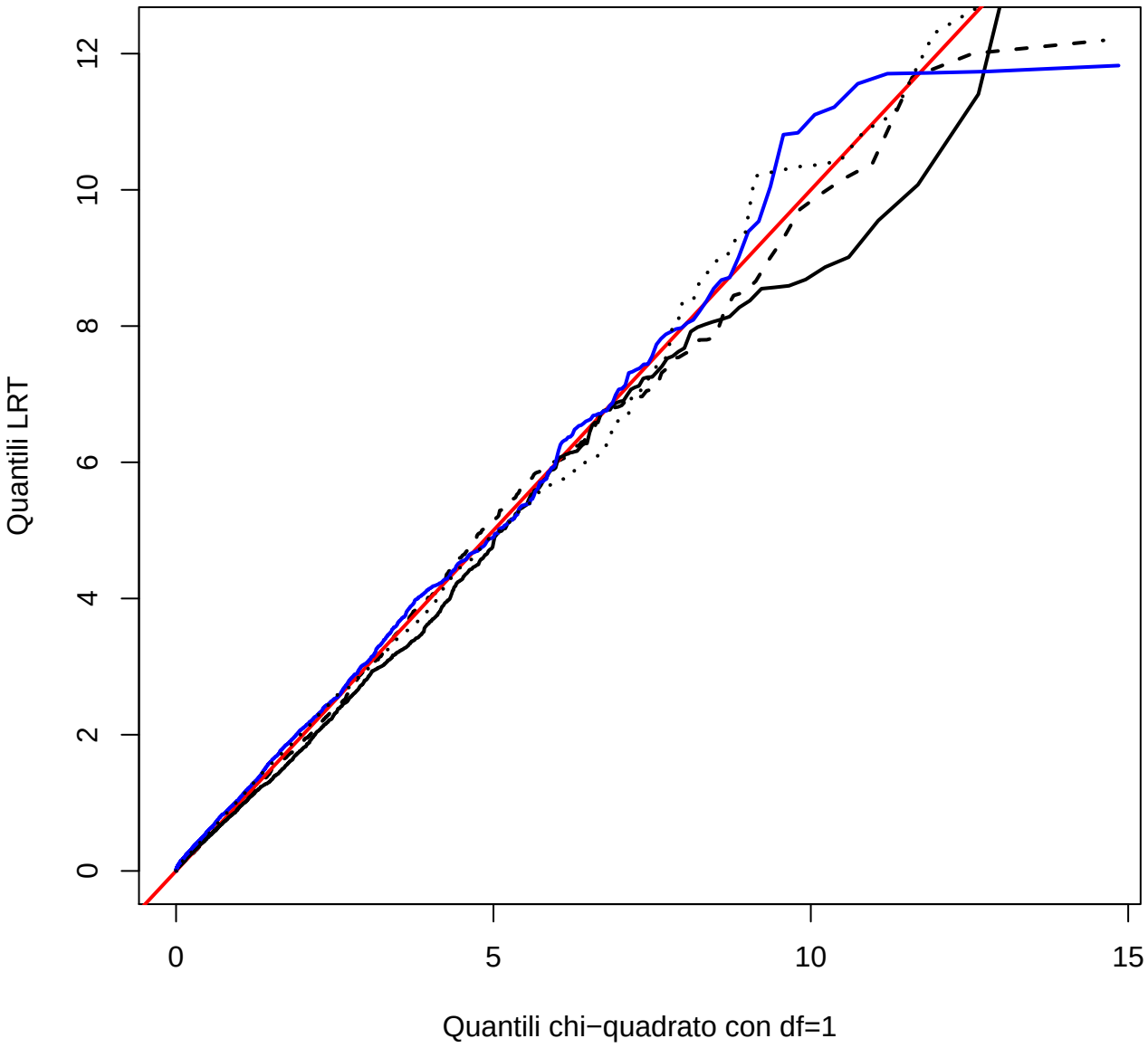

Figura 4.5: Grafico quantile-quantile per il modello ANOVA a due vie. La linea rossa rappresenta i quantili teorici della χ^2_1. La linea tratteggiata rappresenta la statistica test per il modello a numerosità più bassa, la linea nera continua è quella del modello a numerosità 300 mentre quella blu si rappresenta il modello con $N = 1.000$; infine la linea a puntini raffigura la distribuzione per il modello a numerosità 3.000.

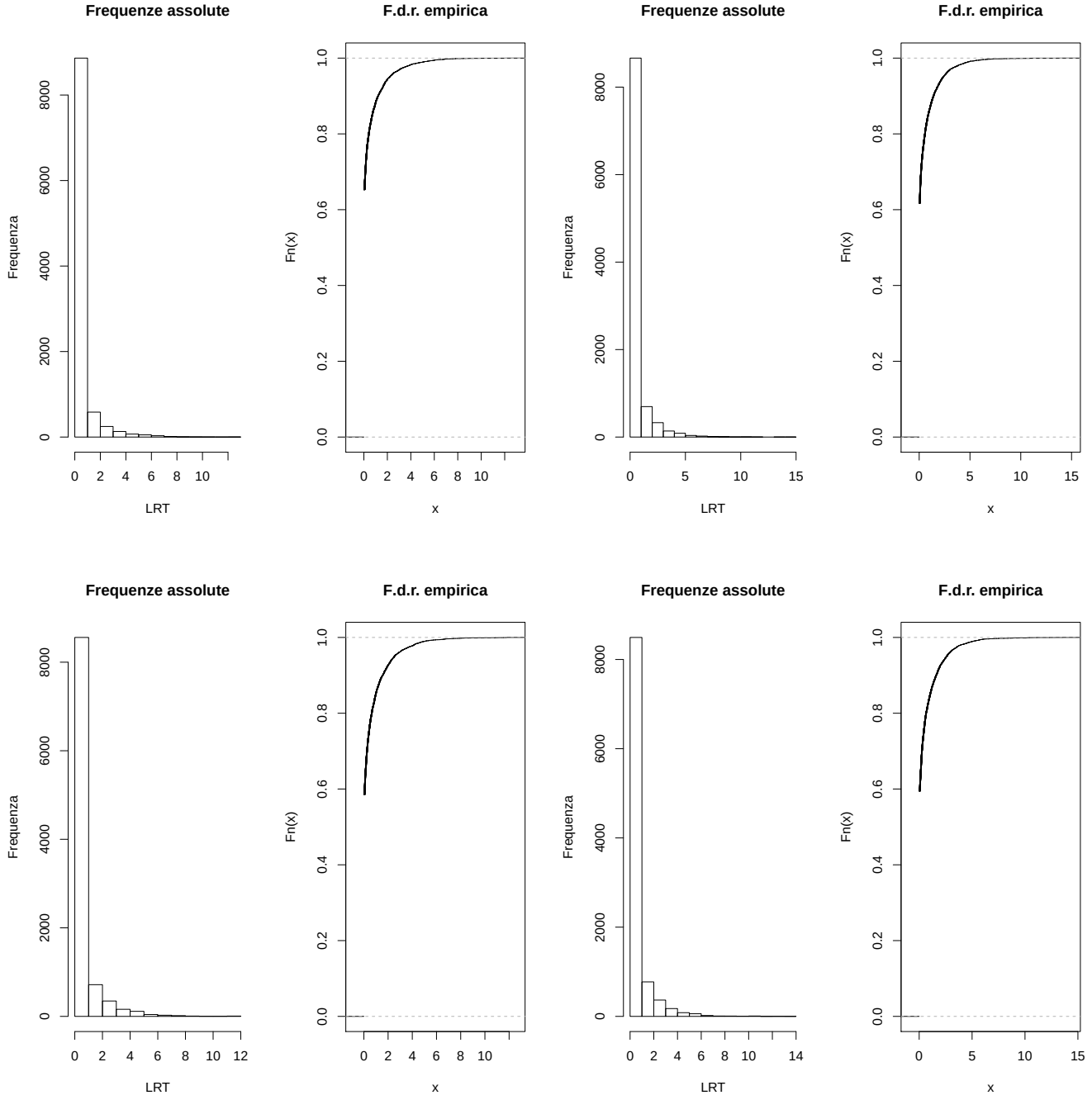

Figura 4.6: Simulazione per verificare la presenza di un effetto casuale per il modello ANOVA a due vie: il grafico in alto a sinistra rappresenta il modello con $N = 50$, quello successivo è del modello con $N = 300$; il grafico in basso a sinistra rappresenta il modello per $N = 1.000$ e il grafico in basso a destra raffigura il modello con $N = 3.000$. Il numero di replicazioni è 10.000.

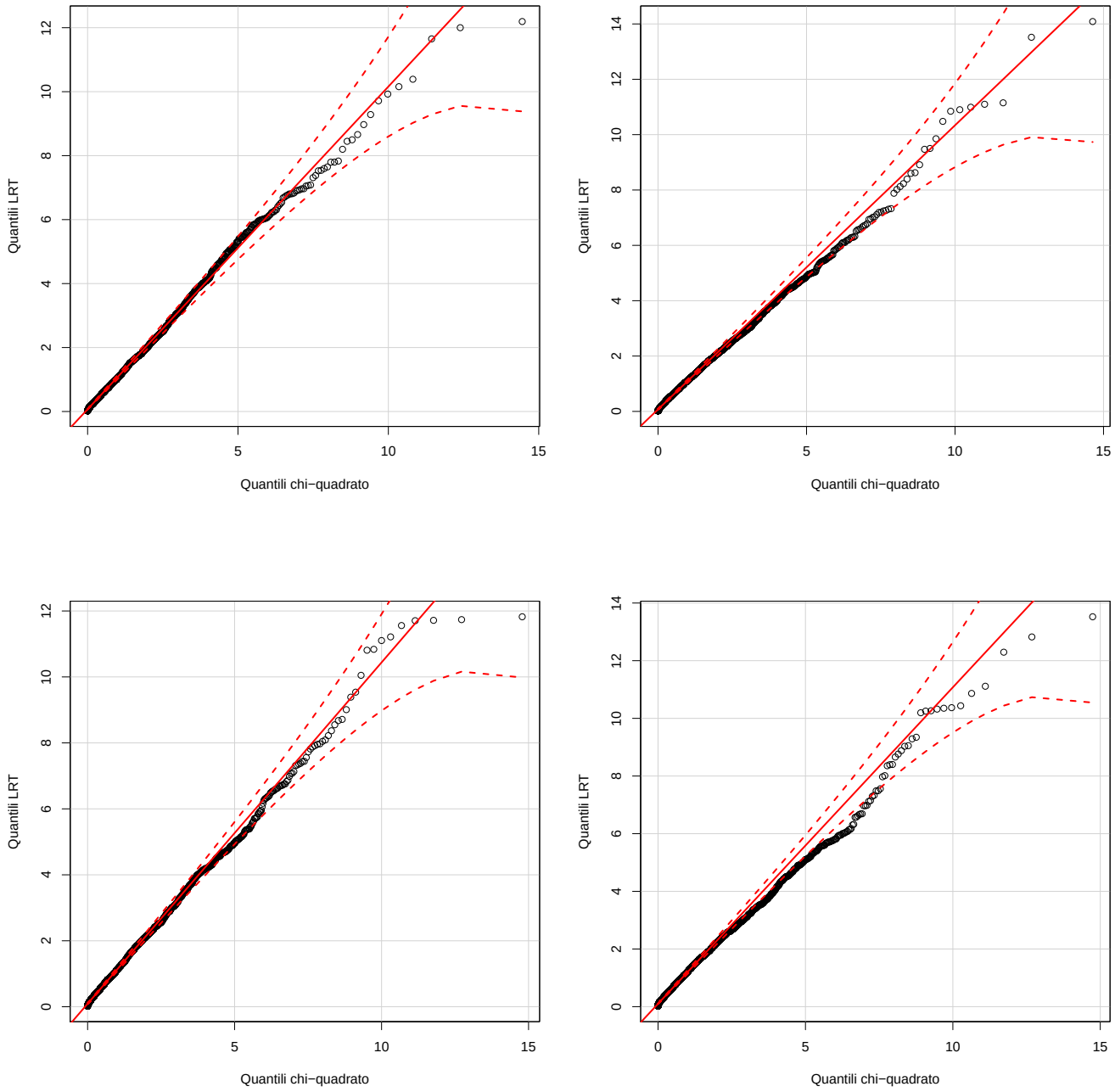

Figura 4.7: Confronto dei quantili di LRT e quantili di χ^2_1, per verificare la presenza di un effetto casuale nel modello ANOVA a due vie:il grafico in alto a sinistra rappresenta il modello con $N = 50$, quello successivo è del modello con $N = 300$; il grafico in basso a sinistra rappresenta il modello per $N = 1.000$ e il grafico in basso a destra raffigura il modello con $N = 3.000$. Il numero di replicazioni è 10.000.

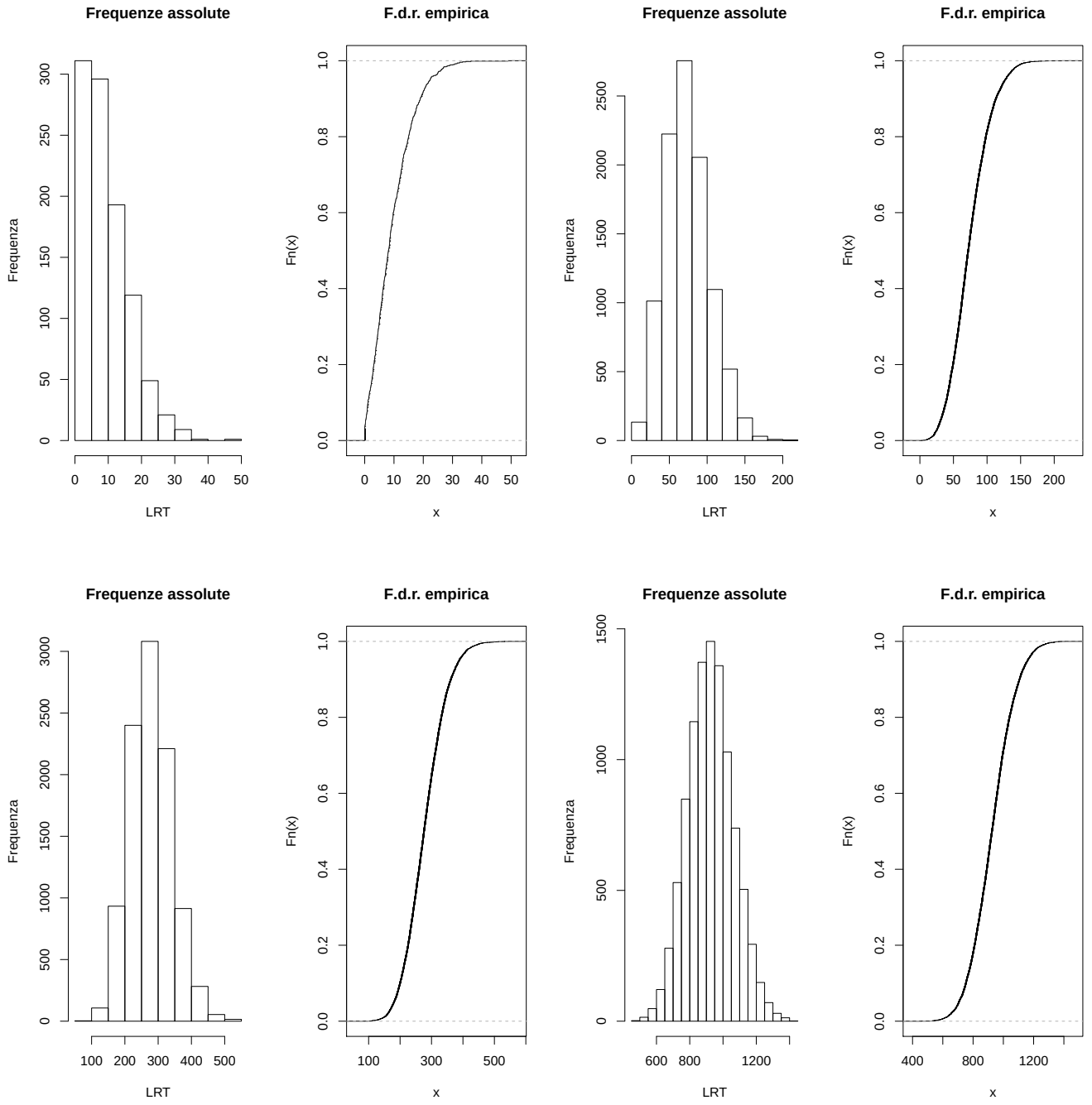

Figura 4.8: Simulazione per verificare la presenza di un effetto casuale per il modello ANOVA a due vie, sotto l'ipotesi nulla: il grafico in alto a sinistra rappresenta il modello con $N = 50$, quello successivo è del modello con $N = 300$; il grafico in basso a sinistra rappresenta il modello per $N = 1.000$ e il grafico in basso a destra raffigura il modello con $N = 3.000$. Il numero di replicazioni è 10.000.

4.3 REML

Nelle simulazioni precedenti è stato utilizzata la stima di massima verosimiglianza, anche se in realtà può essere utilizzata anche la stima REML. Come descritto nel Capitolo 3, la stima REML dovrebbe portare a una minore probabilità che la stima del del parametro si trovi sulla frontiera sotto l'ipotesi alternativa.

L'ipotesi verificata è sempre la stessa: $H_0 : \sigma^2_\alpha = 0$ contro l'alternativa $H_0 : \sigma^2_\alpha \neq 0$, lasciando gli altri parametri liberi.

Il modello di partenza per la stima REML con un effetto casuale è quello in (4.1). Come nel caso di massima verosimiglianza, le stime REML sono disponibili in forma esplicita in Searle *et al.* (1992), e non è stato utilizzato alcuna procedura numerica per la stima.

Nella Figura 4.9 e nelle Tabelle 4.5–4.6 sono riportati i risultati ottenuti per la simulazione sotto l'ipotesi nulla.

La Tabella 4.5 riporta il confronto tra quantili empirici e teorici, per il modello con numerosità maggiore. Come si può vedere, i quantili del test rapporto di verosimiglianza ristretta (LRTR) si avvicinano a quelli di una variabile $0.5\chi^2_0 + 0.5\chi^2_1$.

La Tabella 4.6, riporta il numero di zeri ottenuti nella simulazione. Come annunciato, il numero di zeri ottenuti risulta inferiore: già dalla simulazione a numerosità $N = 300$ il numero di zeri si avvicina al 50% previsto dalla teoria.

La Figura 4.9 mostra il confronto con i quantili della variabile χ^2_1, non considerando i valori nulli di LRT. Anche in questo caso, la simulazione con numerosità maggiore si avvicina meglio alla linea che rappresenta i quantili teorici. Come accadeva per il caso a due vie, verso i valori più alti, LRTR tende a discostarsi dai quantili teorici.

Tabella 4.5: Tabella di confronto tra quantili, per N=3000 e 10.000 replicazioni.

p	**0.5**	**0.75**	**0.9**	**0.95**	**0.975**	**0.99**
teorici	0	0.4729	1.6540	2.6275	3.6810	5.1755
empirici	0	0.394	1.548	2.596	3.767	5.377

Tabella 4.6: Numero di zeri ottenuti per LRT ottenuti nella prima simulazione, con stima REML, suddivisi per numerosità.

I	**J**	**N**	**N. ZERI**	**PROP.**
10	5	50	6216	65.24%
20	15	300	5295	52.95%
50	20	1000	5300	53.00%
100	30	3000	5124	51.24%

Per quanto riguarda la seconda simulazione, la stima è stata eseguita con il pacchetto *lme4*, che prevede anche la possibilità di stimare con la verosimiglianza ristretta.

Nella Figura 4.10 e nella Tabella 4.7 sono riportati i risultati ottenuti per la simulazione sotto l'ipotesi nulla.

A differenza del caso con un effetto casuale, il numero di zeri si avvicina al 50% più lentamente, e come era accaduto per la stima di massima verosimiglianza a due vie, l'ultima simulazione non è quella che genera meno valori nulli per LRT. Nel modello a due effetti casuali, non c'è molta differenza tra la stima con REML e quella con la massima verosimiglianza.

Il Grafico 4.10 riporta il confronto tra i quantili teorici e i quantili empirici. Questo mostra che il quantili di LRT (sono stati considerati solo valori non nulli) si avvicinano molto ai quantili teorici, soprattutto per valori più piccoli di LRT, che comunque sono la maggioranza.

Tabella 4.7: Numero di zeri ottenuti per LRT nella seconda simulazione, con la stima REML, suddivisi per numerosità.

I	**J**	**N**	**N. ZERI**	**PROP.**
10	5	50	6524	65.24%
20	15	300	6151	61.51%
50	20	1000	5902	59.02%
100	30	3000	5985	59.85%

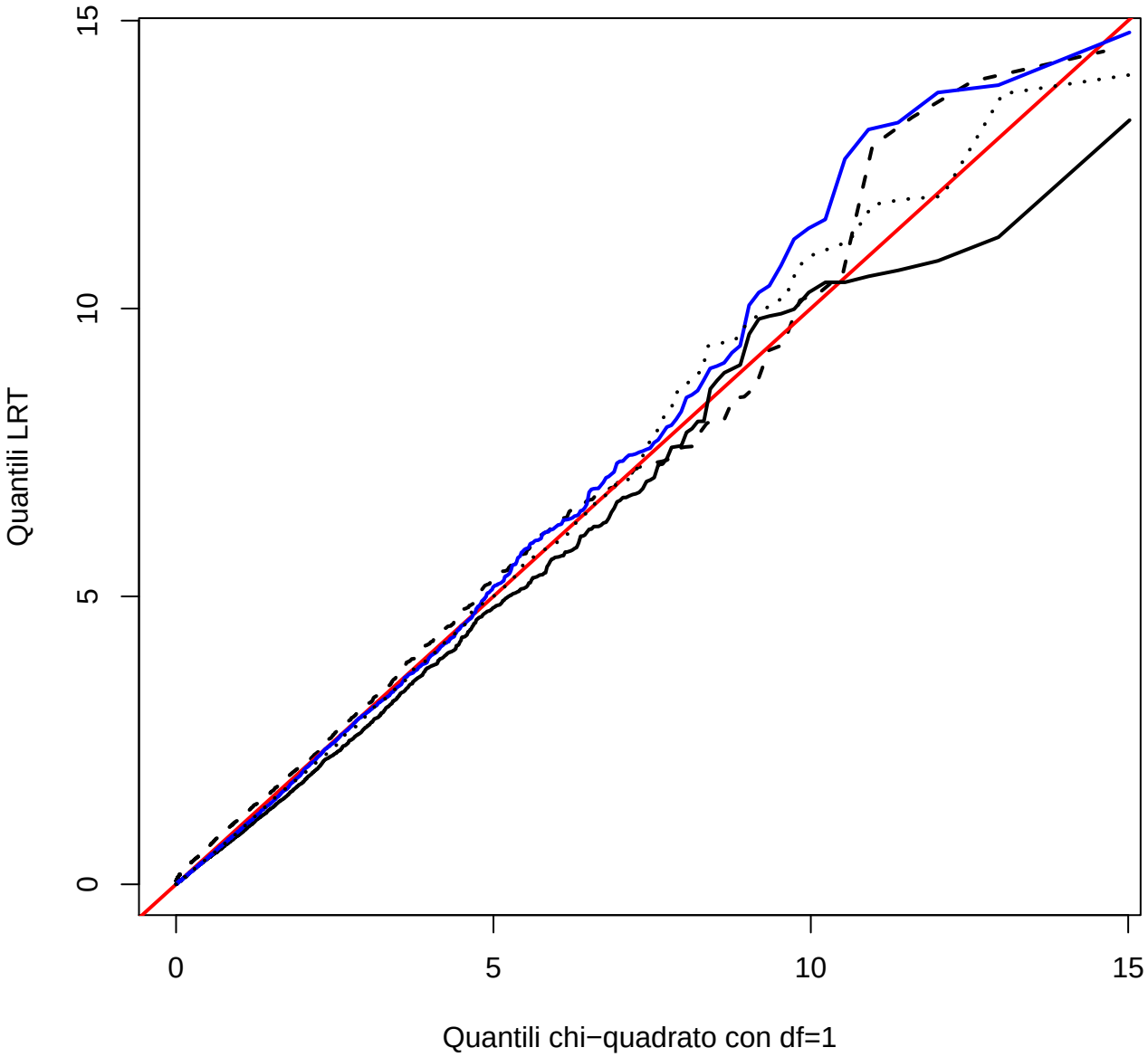

Figura 4.9: Grafico quantile-quantile per il modello ANOVA a una vie con stima REML. La linea rossa rappresenta i quantili teorici della χ^2_1. La linea tratteggiata rappresenta la statistica test per il modello a numerosità più bassa, la linea nera continua è quella del modello a numerosità 300 mentre quella blu si riferisce al modello con N=1.000; infine la linea a puntini rappresenta la distribuzione per il modello a numerosità 3.000.

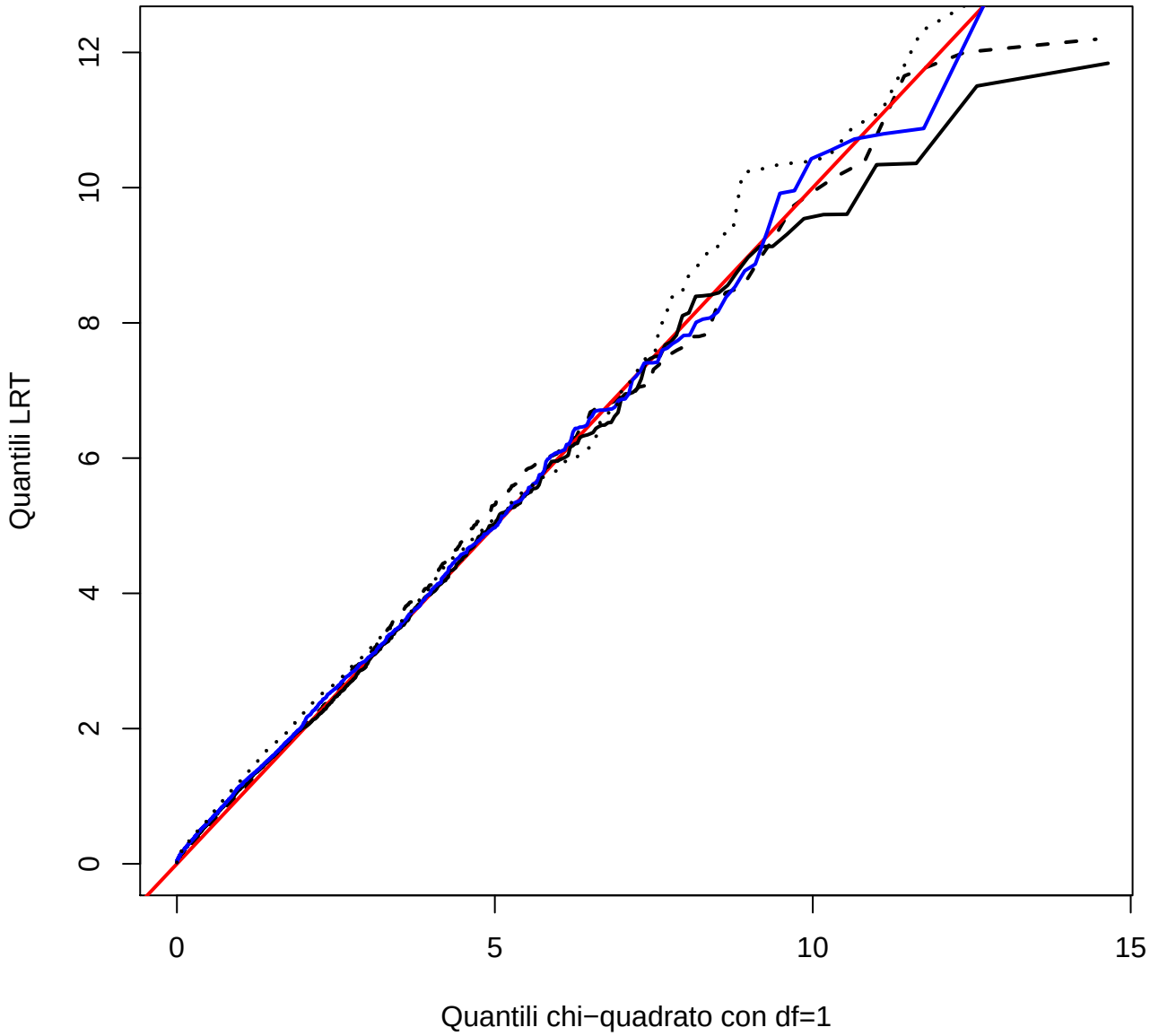

Figura 4.10: Grafico quantile-quantile per il modello ANOVA a due vie con stima REML. La linea rossa rappresenta i quantili teorici della χ^2_1. La linea tratteggiata è la statistica test per il modello a numerosità più bassa, la linea nera continua è quella del modello a numerosità 300 mentre quella blu si riferisce al modello con $N = 1.000$; infine la linea a puntini rappresenta la distribuzione per il modello a numerosità 3.000.

Conclusioni

L'argomento centrale di questa tesi è il comportamento del test rapporto di verosimiglianza nei modelli non regolari e in presenza di parametri di disturbo.

Il test rapporto di verosimiglianza è una statistica semplice da determinare. Inoltre, il test basato sul rapporto di verosimiglianza è frequentemente adottato nell'inferenza per le sue desiderabili proprietà asintotiche, che però vengono a mancare se non sono rispettate alcune condizioni di regolarità riguardanti il modello. In particolare, quando il vero valore del parametro non è un punto interno allo spazio parametro, il modello è non regolare. Questa situazione si verifica spesso nei modelli di regressione lineari in cui l'obiettivo dell'analisi non sono più i parametri di regressione, ma le varianze.

Nei modelli a componenti di varianza interessa capire quanto le diverse fonti di variabilità, che entrano in gioco quando si vuole spiegare l'effetto di un fenomeno, influiscono sulla varianza totale, riuscendo a stimare il contributo di variabilità di ogni fattore singolarmente. In questo contesto, il metodo di stima che viene utilizzato è la massima verosimiglianza e una sua estensione, la massima verosimiglianza ristretta.

Oltre alla stima delle componenti di varianza, si possono applicare delle procedure inferenziali per verificare se la variabilità di alcuni fattori abbia un contributo talmente piccolo su quella totale da poter essere ignorata. Il test adottato per la verifica non è più basato sulla funzione di verosimiglianza propria, per la presenza di parametri di disturbo, come l'intercetta e le varianze degli altri fattori. C'è la necessità di una pseudo-verosimiglianza, come la verosimiglianza profilo, che riesce incentrare la verosimiglianza sul parametro di interesse.

In questa tesi sono illustrati i risultati di uno studio di simulazione per un modello lineare con uno e due effetti casuali. In particolare, è stato analizzato il comportamento asintotico del test rapporto di verosimiglianza quando si verifica la presenza degli effetti casuali, sia standard che basato sulla verosimiglianza ristretta. I risultati ottenuti confermano quelli descritti nell'articolo di Self e Liang (1987).

Uno sviluppo interessante in questo ambito è l'utilizzo della verosimiglianza composita per i modelli più complessi, come nell'articolo elaborato da Bellio e Varin (2005) (vedi Appendice).

Appendice A

Verosimiglianza composita

La verosimiglianza è un metodo immediato e semplice da calcolare per riuscire a ottenere una serie di informazioni riguardanti il parametro di interesse, e per riuscire ad avere delle informazioni sul processo che ha generato i dati.

Quest'affermazione però non considera le situazioni più complesse: molte volte risulta difficoltoso calcolare la funzione di verosimiglianza, come pure quantità connesse ad essa, a causa della presenza di integrali complicati da risolvere, di insieme di dati molto grandi e poco maneggevoli da analizzare o per la presenza di matrici con dimensioni che crescono all'aumentare della numerosità campionaria e che devono essere invertite.

Per superare questo problema computazionale, è stata proposta una procedure che cerca di oltrepassare i limiti della funzione di verosimiglianza, ma basata su di essa: la *verosimiglianza composita*, che rientra nella categoria delle pseudo-verosimiglianze.

Sia data una variabile casuale $Y = (Y_1, ..., Y_n)^T$ con densità congiunta $f(y; \theta)$; inoltre, sia dato un modello statistico parametrico $\mathcal{F} = \{f(y; \theta), y \in \mathcal{Y} \subseteq \mathbb{R}^n, \theta \in \Theta \subseteq \mathbb{R}^p\}$ e un insieme di eventi misurabili $\{\mathcal{A}_i; i = 1, ..., m\}$. Supponiamo che $f(y; \theta)$ sia difficile da calcolare, ma che per qualche sottoinsieme di dati le verosimiglianze siano facilmente ottenibili. Allora, una *verosimiglianza composita (CL)* è un prodotto pesato delle verosimiglianze corrispondenti a ogni singolo evento:

$$CL(\theta; y) = \prod_{i=1}^{m} f(y \in \mathcal{A}_i; \theta)^{w_i}, \tag{A.1}$$

dove $w_i, i = 1, ..., m$ sono dei pesi positivi.

Le verosimiglianze composite possono essere raggruppate in due classi: verosimiglianze composite *marginali*, se sono costruite partendo da densità marginali, e verosimiglianze composite *condizionali*, se sono costruite partendo da densità condizionate.

Un buon motivo per utilizzare questo tipo di verosimiglianza è molto più semplice modellare dipendenze univariate e bivariate piuttosto che la totale dipendenza congiunta dei dati.

Un esempio di utilizzo di verosimiglianza composita è quello descritto in Bellio e Varin (2005). Utilizzando la verosimiglianza a coppie su un modello lineare generalizzato con effetti casuali, hanno ridotto la complessità del calcolo di integrali

da più dimensioni a integrali bivariati. L'insieme di dati utilizzato è quello sull'accoppiamento delle salamandre, disponibile nel sito *http://stat.uibk.ac.at/SMIJ*.

Il vantaggio di questo metodo è che non c'è bisogno di utilizzare delle simulazioni per la stima, ma soprattutto, produce degli stimatori consistenti e asintoticamente normali.

I dati a disposizione sono discreti: $\mathbf{y} = \{y_{ij}\}$ e $\{x_{ij}\}$ per $i = 1, ..., q_1, j = 1, ..., q_2$, e il modello è a effetti misti a due vie, con struttura incrociata ma senza interazione. La media condizionale è $g\{E(Y_{ij})|u_j, \nu_i\} = x_{ij}^t\beta + u_j + \nu_i$, dove β è un vettore a p dimensioni di effetti fissi, $g(\cdot)$ è la funzione legame, $u_i \sim \mathcal{N}(0, \sigma_u^2)$ e $\nu_j \sim \mathcal{N}(0, \sigma_\nu^2)$ sono i due effetti casuali indipendenti tra loro. La funzione di verosimiglianza completa è un integrale di dimensioni $q_1 \times q_2$, in genere difficile da calcolare, e quindi si passa alla verosimiglianza a coppie:

$$L_2(\theta; y) = \prod_{i=1}^{q_1}\prod_{j<j'}^{q_2} P(Y_{ij} = y_{ij}, Y_{ij'} = y_{ij'}; \theta) \prod_{i<i'}^{q_1}\prod_{j=1}^{q_2} P(Y_{ij} = y_{ij}, Y_{i'j} = y_{i'j}; \theta).$$

Se il legame è la funzione probit, quindi $g(p) = \Phi^{-1}(p)$, dove Φ è la funzione di ripartizione della Normale standard, si avrà

$$P(Y_{ij} = 1, Y_{ij'} = 1; \theta) = \Phi_2\left(\frac{x_{ij}^t\beta}{\sqrt{1+\sigma_u^2+\sigma_\nu^2}}, \frac{x_{ij'}^t\beta}{\sqrt{1+\sigma_u^2+\sigma_\nu^2}}; \frac{\sigma_u^2}{1+\sigma_u^2+\sigma_\nu^2}\right),$$

dove $\Phi_2(a, b; \rho)$ è la funzione di ripartizione di una Normale standard bivariata con correlazione ρ calcolata in $(a, b)^T$.

Nello specifico insieme di dati delle salamandre, è stata modellata la probabilità di accoppiamento tra una femmina della popolazione R con un maschio della popolazione W:

$$\pi_{R/W} = P(Y = 1|X_{R/R} = 0, X_{R/W} = 1, X_{W/R} = 0, X_{W/W} = 0; \boldsymbol{\theta}),$$

dove $X_{i/j}$ indica se c'è stato l'accoppiamento tra una femmina della popolazione i con un maschio della popolazione j, per $i, j = R, W$, e le altre probabilità $\pi_{i/j}$, $i, j = R, W$ sono state trovate allo stesso modo. Per il calcolo di intervalli di confidenza per le probabilità π_{ij} sono state utilizzate tecniche di boostrap, gli effetti casuali sono stati verificati utilizzando la statistica test basata sul rapporto di verosimiglianza composita

$$LRT_2(\theta; y) = 2\{l_2(\hat{\theta}; y) - l_2(\hat{\theta}^0; y)\},$$

dove $l_2(\theta; y) = \log L_2(\theta; y)$, $\hat{\theta}$ è la stima di massima verosimiglianza composita e $\hat{\theta}^0$ è la stima di massima verosimiglianza composita sotto l'ipotesi nulla che alcuni componenti di varianza possano essere esclusi dal modello. I risultati ottenuti con $LRT_2(\theta; y)$ confermano la presenza di entrambi gli effetti casuali. Inoltre, il metodo di verosimiglianza composita a coppie è risultato essere il migliore in termini di performance, anche rispetto allo stimatore calcolato con la REML.

Bibliografia

[1] Barndorff-Nielsen, O.E., Cox, D.R. (1994). *Inference and Asymptotics*. CHAPMAN & HALL, London.

[2] Bellio, R., Varin, C.(2005). A pairwise likelihood approach to generalized models with crossed random effects. *Statistical Modelling* **5**,217-227.

[3] Boente, G., Fraiman, R.(1988). On the asymptotic behaviour of general maximum likelihood estimates for the nonregular case under nonstandard conditions.*Biometrika* **75**, 45-56.

[4] Brazzale, A.R., Davison, A.C., Reid, N.(2007).*Applied Asymptotics: Case Studies in Small-Sample Statistics*. Cambridge University Press, New York.

[5] Chen, Y., Liang, K.Y.(2010). On the asymptotic behaviour of the pseudolikelihood ratio test statistic with boundary problems.*Biometrika* **97**, 603-620.

[6] Crainicenau, C.M., Ruppert, D.(2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society: Series B* **66**, 165-185.

[7] Davison, A.C.(2003).*Statsitical Models*. Cambridge Aeries in Stataitical and Probabilistic Mathematics.

[8] Feng, Z., McCulloch, C.E.(1992). Statsitical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space.*Statsitics and Probability Letters* **13**, 325-332.

[9] Fisher, R.A.(1922). On the mathematical foundations of theorical statistics.*Philosophical Transactions of the Royal Society of London. Series A*,**222**, 309-368.

[10] Fraser, D.A.S.(1991). Statistical inference: likelihood to significance.*Journal of the American Statistical Association* **86**, 258-265.

[11] Kopylev, L., Sinha, B.(2011). On the asymptotic distribution of likelihood ratio test when parameters lie on the boundary.*Sankhya B* **73**, 20-41.

[12] Le Cessie, S., Van Houwelingen, J.C.(1994). Logistic regression for correlated binary data.*Appl. Stat.* **43**, 95-108.

[13] Miller, J.J.(1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *The Annals of Statistics* **5**,746-762.

[14] Molenberghs, G., Veberke,G.(2007). Likelihood ratio, score and Wald tests in a constrained parameter space. *The American Statistician* **61**, 22-27.

[15] Moran, P.A.P.(1971). Maximum-likelihood estimation in non-standard conditions.*Mathematical Proceedings of the Cambridge Philosophical Society* **70**, 441-450.

[16] Pace, L., Salvan, A.(2001). *Introduzione alla statsitica. II Inferenza, verosimiglianza, modelli*. CEDAM, Padova.

[17] Russel, T., Bradley, R.A.(1958). One-way variances in two-way classification.*Biometrika* **45**, 111-129.

[18] Satterthwaites, F.E.(1946). An approximate distribution of estimates of variance components.*Biometrics*,**2**,110-114.

[19] Searle, S.R., Casella, G., McCulloch, C.E.(1992). *Variance Components*. Wiley, New York.

[20] Self, G.S., Liang, K.Y.(1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions.*Journal of American Statsitical Association* **82**, 605-610.

[21] Severini, T.A.(2000). *Likelihood Methods in Statistics*. OXFORD, New York.

[22] Shapiro, A.(1988). Towars a unified theory of inequality constrained testing in multivariate analysis.*International statistical review* **56**, 49-62.

[23] Smyth, G.K., Verbyla, A.P. (1996). A conditional approach to residual maximum likelihood estimation in generalized linear models. *J.R. Static. Soc. B* **58**, 565-572.

[24] Stein, M.L., Chi Z., Welty, L.J.(2003). Approximating likelihoods for large spatial data sets. *J.R. Static. Soc. B* **66**, 275-296.

[25] Stern, S.E., Welsh, A.H.(2000). Likelihood inference for small variance components. *The Canadian Journal of Statistics* **28**, 517-532.

[26] Stram, D.O., Lee, J.W.(1994). Variance components testing in the longitudinal mixed effects model.*Biometrics* **50**, 1171-1177.

[27] Varin, C., Vidoni, P.(2005). A note on composite likelihood inference and model selection.*Biometrika* **92**, 519-528.

[28] Varin, C., Reid, N., Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* **21**, 5-42.

[29] Varin, C.(2008). On composite marginal likelihoods. *AStA: Advances in Statistical Analysis*,**92**, 1-28.

[30] Visscher, P.M.(2006). A note on the asymptotic distribution of likelihood ratio tests to test variance components. *Twin Research and Human Genetics* **9**, 490-495.

[31] Vu, H.T.V., Zhou, S.(1997). Generalization of likelihood ratio tests under non standard conditions. *The Annals of Statistics* **25**, 897-916.