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

Bootstrap parametrico in modelli con effetti fissi incrociati e dati discreti sparsi

Relatore: Prof.ssa Alessandra Salvan
Dipartimento di Scienze Statistiche

Correlatore: Prof. Nicola Sartori
Dipartimento di Scienze Statistiche

Laureando: Davide Benussi
Matricola 2071952

Anno Accademico 2023/2024

Indice

Introduzione

Nei problemi di inferenza nell'ambito di un modello statistico parametrico spesso si è interessati a fare inferenza solo su una componente dell'intero parametro, nota come **parametro di interesse**, mentre la componente complementare, a cui ci si riferisce convenzionalmente con il termine **parametro di disturbo**, ha la sola funzione di rendere il modello più realistico e flessibile, ma non è di diretto interesse. Poiché il parametro di disturbo viene usualmente stimato sulla base dei dati, occorre tenerne conto nella definizione delle procedure di inferenza sul parametro di interesse.

Esistono essenzialmente due tipologie di parametri di disturbo: parametri di disturbo con dimensione fissata o parametri di disturbo con dimensione che dipende dalla numerosità campionaria N. Il primo caso rappresenta la situazione standard, in quanto lo spazio parametrico ha dimensione che non dipende da N, soddisfacendo così una delle condizioni generali di regolarità per l'inferenza basata sulla verosimiglianza (Severini, 2000, §3.4). Ciò fa sì che, in un contesto asintotico standard, quando N aumenta, aumenta anche l'informazione a disposizione e, quindi, anche la precisione delle procedure inferenziali. Invece, nel secondo caso, quando la dimensione dello spazio parametrico cresce con la numerosità campionaria N, l'aumento di informazione potrebbe non essere altresì efficace, poiché, seppur l'informazione cresce, cresce anche la dimensione della componente di disturbo, e pertanto c'è il rischio concreto di trarre conclusioni poco affidabili. Quando il parametro di disturbo ha dimensione che non è fissata, esso prende il nome di **parametro incidentale**. Proprio per le difficoltà intrinseche per le procedure inferenziali in presenza di parametri incidentali, in letteratura ci si riferisce a tali problemi come **problemi di Neyman e Scott**, riferendosi ai due autori che in ambito econometrico hanno messo in luce le difficoltà per l'inferenza che possono sorgere in tali situazioni (Neyman & Scott, 1948). La letteratura recente, come Battey & Cox (2020, 2022), ha sottolineato come questi problemi fossero in realtà stati affrontati in precedenza anche da Bartlett (1937).

In presenza di parametri incidentali l'utilizzo di procedure inferenziali standard per il parametro di interesse, tipicamente basate sulla verosimiglianza profilo, potrebbe rivelarsi pertanto non adeguato (Sartori, 2003). Utilizzare la verosimiglianza profilo non è conveniente se i dati non contengono informazione a sufficienza sulla componente di disturbo, evenienza comune quando la dimensione del parametro di disturbo cresce con N. Per migliorare l'inferenza in presenza di parametri incidentali la letteratura (si veda, ad esempio, Severini, 2000, Capitolo 4) ha proposto diverse modifiche alla verosimiglianza profilo, volte tipicamente a correggere la distorsione della funzione punteggio profilo che peggiora all'aumentare della dimensione del parametro di disturbo.

Un' alternativa all'impiego di modifiche analitiche alle classiche quantità di verosimiglianza, come la statistica test radice con segno del rapporto di verosimiglianza, prevede l'utilizzo del *bootstrap* parametrico (Young, 2009). Infatti, grazie alle moderne risorse di calcolo, le procedure di simulazione basate sul *bootstrap* risultano non troppo computazionalmente dispendiose e possono essere utilizzate anche in quei contesti in cui le modifiche analitiche alle quantità di verosimiglianza non sono disponibili. Inoltre, in un contesto asintotico standard, è noto che il *bootstrap* parametrico permette di ottenere un livello di accuratezza superiore rispetto all'utilizzo delle procedure analitiche standard basate sulla verosimiglianza (DiCiccio, Martin & Stern, 2001).

I modelli stratificati rappresentano una delle situazioni classiche in cui può verificarsi il problema dei parametri incidentali e in cui le classiche procedure basate sulla verosimiglianza profilo potrebbero risultare fallimentari (Sartori, 2003). Nonostante ciò, tali problemi possono risultare ancora più gravi nel caso di alcuni specifici paradigmi asintotici: nel caso di modelli a due indici asintotici ed effetti fissi incrociati la dimensione del parametro di disturbo cresce in due direzioni e ciò fa sì che l'informazione a disposizione possa risultare insufficiente anche quando la numerosità campionaria è moderatamente elevata.

Il principale obiettivo di questa tesi è quello di confrontare le tecniche basate su modificazioni della verosimiglianza profilo e delle usuali quantità pivotali con l'approccio del *bootstrap* parametrico quando è di interesse fare inferenza su un parametro di interesse nel caso dei modelli a due indici asintotici con effetti fissi incrociati per dati discreti sparsi. Rispetto a quanto studiato da Bellio et al. (2023b), viene posta maggiore enfasi sui modelli per dati discreti, anziché solamente continui, e si analizza lo scenario degli effetti fissi incrociati, anziché stratificati, ipotizzando la presenza di sparsità nei dati.

Il **Capitolo 1** è volto a richiamare i concetti di base dell'inferenza basata sulla verosimiglianza, con particolare attenzione al problema dell'inferenza in presenza di parametri di disturbo e agli strumenti che possono essere utilizzati in tale contesto, quali la verosimiglianza profilo e la verosimiglianza profilo modificata. Inoltre, si presenta il *bootstrap* parametrico come possibile alternativa all'utilizzo delle classiche quantità pivotali basate sulla verosimiglianza.

Nel **Capitolo 2** l'attenzione si sposta sui modelli con effetti fissi stratificati. In particolare vengono presentate le difficoltà che sorgono per le procedure inferenziali in questo tipo di paradigma asintotico e illustrate alcune possibilità per aggiustare l'inferenza: verosimiglianza condizionata, marginale e verosimiglianza profilo modificata. Inoltre, particolare enfasi viene posta sui possibili vantaggi dell'utilizzo del *bootstrap* parametrico.

Nel **Capitolo 3** si introducono i modelli con effetti fissi incrociati, mostrando come le tecniche per aggiustare l'inferenza nel caso di effetti fissi stratificati possano essere utilizzate anche in questo contesto.

Il **Capitolo 4** è dedicato agli studi di simulazione. Nello specifico, l'attenzione è rivolta al caso dei modelli lineari generalizzati per dati discreti sparsi nel paradigma asintotico a due indici con effetti fissi incrociati, in quanto in tale scenario non sono ancora disponibili risultati generali né circa le proprietà delle procedure inferenziali né circa le proprietà del *bootstrap* parametrico.

Infine, nel paragrafo delle Conclusioni, vengono riassunti i risultati principali della tesi, enfatizzando quanto suggerito dagli studi di simulazione e si offrono alcuni spunti per studi o approfondimenti ulteriori.

Capitolo 1

Inferenza basata sulla verosimiglianza

1.1 Introduzione

In questo capitolo verranno richiamati alcuni concetti alla base della teoria dell'inferenza statistica. A partire dalla formalizzazione di un modello statistico parametrico, si introdurranno la funzione di verosimiglianza e le sue principali proprietà. La funzione di verosimiglianza può essere declinata nel suo utilizzo anche quando è di interesse fare inferenza solo su una componente del parametro del modello. In tale situazione, è necessario introdurre il concetto di pseudo-verosimiglianza. La verosimiglianza profilo è un esempio di funzione di pseudo-verosimiglianza.

Quando il modello statistico è caratterizzato da un numero elevato di parametri di disturbo, o quando la struttura della componente di disturbo dipende dalla dimensione campionaria, la verosimiglianza profilo potrebbe non essere la migliore pseudo-verosimiglianza da utilizzare per l'inferenza sul parametro di interesse. Dopo aver discusso dei problemi legati all'utilizzo della verosimiglianza profilo quando la dimensione del parametro di disturbo è grande rispetto alla dimensione campionaria, verranno illustrate alcune delle proposte che la letteratura ha introdotto per migliorare l'inferenza in questi scenari. In particolare, si introdurranno alcune possibili modifiche della verosimiglianza profilo, tipicamente basate su approssimazioni di ordine superiore.

Infine, verrà presentata una possibile alternativa all'utilizzo di modifiche analitiche alle quantità di verosimiglianza, ovvero la possibilità di ottenere le proprietà distributive

delle procedure standard di verosimiglianza tramite metodi di simulazione. In particolare, si discuterà dei possibili vantaggi dell'utilizzo del *bootstrap* parametrico e di alcune sue varianti.

Per una trattazione più approfondita su questi argomenti si rimanda a Pace & Salvan (1997) e Barndorff-Nielsen & Cox (1994) per i temi legati alla verosimiglianza e a Efron & Tibshirani (1993) e Davison & Hinkley (1997) per i temi legati al *bootstrap*.

1.2 Specificazione del modello

Lo scopo principale dell'inferenza statistica è quello di riuscire a spiegare aspetti di un fenomeno di interesse attraverso la specificazione di un modello che sia compatibile con i dati a disposizione su tale fenomeno. Ciò significa che, a partire da un insieme limitato di osservazioni sulla realtà di interesse, si desidera ricostruire il processo ignoto che ha generato i dati.

Formalizzando, si assume di avere a disposizione un campione $y = (y_1, \ldots, y_N)$ da una variabile casuale Y la cui legge $p^0(y)$ è compatibile con il fenomeno osservato, e si vuole ricostruire tale processo generatore a partire dai dati a disposizione e da eventuali informazioni ausiliarie.

Il punto di partenza è la specificazione di un modello statistico $\mathcal{F}$. Il modello statistico $\mathcal{F}$ è costituito da una famiglia di distribuzioni di probabilità che si ritengono compatibili con il fenomeno di interesse, e all'interno della quale dovrebbe ricadere la vera legge $p^0(y)$ da cui i dati sono stati generati. Se $p^0(y) \in \mathcal{F}$ si dice che il modello è correttamente specificato, altrimenti è $\mathcal{F}$ è mispecificato.

A seconda della tipologia dei dati a disposizione e del livello di informazione, è possibile specificare il modello statistico $\mathcal{F}$ con una diversa estensione. Esistono tre livelli generali di specificazione:

- **Specificazione parametrica**. Gli elementi all'interno del modello $\mathcal{F}$ possono essere indicizzati da un numero finito p di parametri reali, ossia

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

  dove $\mathcal{Y}$ è lo spazio campionario e Θ lo spazio parametrico per θ;

- **Specificazione semiparametrica**. Gli elementi all'interno del modello $\mathcal{F}$ sono individuati sia da una componente parametrica che da una non parametrica, ossia

$$\mathcal{F} = \{p(y;\theta), \quad y \in \mathcal{Y}, \quad \theta \in \Theta\},$$

dove $\theta = (\psi, h(\cdot))$, $\Psi \in T \subseteq \mathbb{R}^k$, mentre la funzione $h(\cdot)$ non è indicizzabile da un insieme finito di parametri reali;

- **Specificazione non parametrica**. Il modello $\mathcal{F}$ è una restrizione dell'insieme di tutte le distribuzioni di probabilità con supporto adeguato alla natura dei dati. Le assunzioni semplificatrici in tal senso sono globali e non individuano espressamente un numero finito di parametri per l'inferenza.

Sebbene la scelta del modello $\mathcal{F}$ sia molto importante ed influisca maggiormente sui risultati di un'analisi statistica rispetto alla scelta di uno specifico paradigma inferenziale, non vi sono indicazioni esplicite sul problema di specificazione nella teoria dell'inferenza statistica. Ne segue che la a specificazione del modello $\mathcal{F}$ è tipicamente il risultato di un processo iterativo, guidato da norme basate sul buon senso. Nel seguito verranno trattati esclusivamente specificazioni parametriche, in cui il parametro θ si assume essere identificabile, ovvero esiste una corrispondenza biunivoca tra gli elementi di Θ e di $\mathcal{F}$, nel senso che identificare il vero processo generatore dei dati $p^0(y)$ corrisponde a identificare il corrispondente vero valore del parametro θ.

1.3 Principi di riduzione del modello

Nelle procedure inferenziali spesso si è interessati a considerare delle **sintesi** dei dati $y = (y_1, \ldots, y_N)$ che siano in grado di estrarre tutta l'informazione possibile sul parametro θ, dato un modello $\mathcal{F}$. L'estrazione di informazione sul parametro θ dal campione $y = (y_1, \ldots, y_N)$ può essere attuata tramite statistiche. Una **statistica** è una trasformazione misurabile $s : \mathcal{Y} \to \mathcal{S}$, con s una funzione non iniettiva. In quanto trasformazione dei dati, una statistica è una variabile casuale a cui è associato il modello indotto $\mathcal{F}_S$, con generico elemento $p_S(s; \theta)$, ossia il modello per i dati trasformati $s = s(y)$. L'utilità di una statistica s per l'inferenza su θ dipende dalla riduzione inferenziale che essa opera, ossia dalla relazione che sussiste tra il modello statistico per i dati trasformati e quello per i dati originari. Nel seguito verranno richiamati brevemente le caratteristiche essenziali di tre possibili tipologie di statiche, utili per la comprensione dei temi successivi: le statistiche costanti in distribuzione, le statistiche sufficienti e le statistiche ancillari.

1.3.1 Statistiche costanti in distribuzione, sufficienti e ancillari

Dato un modello statistico $\mathcal{F}$, si dice che una statistica c è **costante in distribuzione** (rispetto ad $\mathcal{F}$) se il modello statistico indotto ha un solo elemento, ossia se la distribuzione di C non dipende da θ. In altri termini, la distribuzione di C dipende dal modello

$\mathcal{F}$ ma non dalla specifica collocazione di $p^0(y)$ entro $\mathcal{F}$.

Una statistica costante in distribuzione non contiene informazione sul parametro θ ma può consentire di effettuare una riduzione per condizionamento del problema inferenziale (Fisher, 1934, 1935). Si assuma che y sia in corrispondenza biunivoca con (t, c), dove c è costante in distribuzione e t è una statistica complementare a c. Sotto generali condizioni di regolarità, vale la fattorizzazione

$$p_{C,T}(c, t; \theta) = p_C(c) p_{T|C=c}(t; c, \theta).$$

Dunque, l'inferenza su θ può essere basata sul modello condizionato $p_{T|C=c}(t; c, \theta)$ al valore osservato della statistica costante in distribuzione. Questo porta a considerare il principio di condizionamento, secondo cui le valutazioni basate sul principio del campionamento ripetuto circa l'incertezza delle procedure inferenziali dovrebbero essere condizionate al valore osservato della statica costante in distribuzione. Tale principio, tuttavia, non è globalmente accettato nella teoria dell'inferenza statistica in quanto solleva alcune questioni complesse come la scelta del sottospazio campionario rilevante per il principio del campionamento ripetuto. Inoltre, sorgono diverse difficoltà pratiche, in quanto non esiste un criterio generale per trovare una statistica costante in distribuzione dato un modello $\mathcal{F}$, e potrebbe non esistere una statistica non banale costante in distribuzione.

Una situazione diametralmente opposta è quella che si verifica nel caso delle statiche sufficienti. Dato un modello statistico $\mathcal{F}$, si dice che una statistica s è **sufficiente** (rispetto ad $\mathcal{F}$) se la distribuzione di Y condizionata ad S non dipende da θ. In altri termini, sotto generali condizioni di regolarità, vale la fattorizzazione

$$p_Y(y, \theta) = p_S(s, \theta) p_{Y|S=S}(y; s),$$

per ogni y tale che $s(y) = s$. Dunque, l'inferenza su θ può essere basata sul modello indotto da S. Un modello statistico $\mathcal{F}$ ammette svariate statistiche sufficienti. Quanto più una statistica sufficiente è concisa, tanto più essa risulta adeguata allo scopo di riassumere tutta l'informazione portata dai dati sul parametro. Si definisce statistica sufficiente minimale per il modello $\mathcal{F}$ una statistica s che, oltre ad essere sufficiente, è funzione di ogni altra statistica sufficiente.

Nel modello indotto dalla statistica sufficiente minimale s potrebbero essere presenti ulteriori aspetti costanti in distribuzione. Quando ciò non avviene, e l'inferenza su θ può essere basata solo sulla riduzione per sufficienza, si dice che s è una statistica sufficiente completa, ovvero una statistica le cui sole funzioni costanti in distribuzione sono quelle

banali (costanti).

Infine, un'ultima classe di statistiche utili nell'inferenza di verosimiglianza è la classe delle statiche ancillari. La loro introduzione richiede alcuni definizioni legate alla verosimiglianza, in particolare quella di stima di massima verosimiglianza $\hat{\theta}$, pertanto si veda il paragrafo 1.4. Dato un modello statistico $\mathcal{F}$ e una statistica sufficiente minimale s per $\mathcal{F}$ si dice che una statistica a è **ancillare** (rispetto ad $\mathcal{F}$) se è ausiliaria, ossia se s è in relazione biunivoca con $(\hat{\theta}, a)$, e se a è costante in distribuzione. In altri termini, sotto generali condizioni di regolarità, vale la fattorizzazione

$$p_{\hat{\theta},A}(\hat{\theta}, a; \theta) = p_A(a) p_{\hat{\theta}|A=a}(\hat{\theta}; a, \theta),$$

dove $\hat{\theta}$ risulta statistica sufficiente minimale nel modello condizionato al valore osservato di a (Fisher, 1934). Una statistica ancillare ideale esprime l'informazione riguardante la precisione con cui la stima di massima verosimiglianza individua θ nello spazio parametrico Θ, che viene tipicamente persa nel passaggio da s a $\hat{\theta}$. Il principio di condizionamento, nella sua versione più debole, può pertanto essere ristretto alle sole statistiche ancillari. Il modello condizionato ad una statistica ancillare può risultare utile perché la stima di massima verosimiglianza, se unica, è funzione della statistica sufficiente minimale ma non conserva necessariamente la proprietà di sufficienza. Da un punto di vista pratico, il reperimento di una statistica ancillare a può non essere un compito facile. Qualora il passaggio dalla riduzione per sufficienza dei dati all'ulteriore sintesi rappresentata da $\hat{\theta}$ comporti perdita d'informazione e una statistica ancillare esatta non esista o non sia agevolmente individuabile, è possibile ricorrere a soluzioni di tipo approssimato, ovvero a statistiche ancillari asintoticamente costanti in distribuzione.

1.4 Verosimiglianza e quantità collegate

Un modello statistico parametrico $\mathcal{F}$ è identificato dalla terna

$$\{\mathcal{Y},\ p(y;\theta),\ \Theta\},$$

dove $\mathcal{Y}$ è lo spazio campionario per y, $\theta \in \Theta \subseteq \mathbb{R}^p$ è lo spazio parametrico per θ, mentre $p(y;\theta)$ è la densità di Y per θ fissato. Al contrario, fissato $y = (y_1, \ldots, y_N)$, $p(y;\theta)$ è interpretabile come funzione del solo parametro $\theta = (\theta_1, \ldots, \theta_p)$.

Si definisce la **funzione di verosimiglianza** (Fisher, 1922) per θ basata sui dati y

come la funzione $L : \Theta \to \mathbb{R}^+$

$$L = L(\theta) = L(\theta; y) = c(y)p_Y(y, \theta),$$

dove $\theta \in \Theta$ e $c(y)$ è una costante di proporzionalità arbitraria che non dipende da θ. La funzione di verosimiglianza rappresenta la sintesi più coincisa dei dati, basata sul modello $\mathcal{F}$, che non perde informazione sul parametro θ.

Nel caso particolare in cui le osservazioni $y = (y_1, \ldots, y_N)$ sono assunte essere realizzazioni indipendenti e identicamente distribuite (d'ora in poi i.i.d.) della variabile casuale Y_1, la funzione di verosimiglianza ha la seguente forma

$$L(\theta) = \prod_{i=1}^{N} p_{Y_1}(y_i, \theta),$$

dove $p_{Y_1}(y_i, \theta)$ è la distribuzione di Y_1. Invece che considerare la funzione di verosimiglianza, spesso è conveniente utilizzare la funzione di log-verosimiglianza, che è semplicemente la trasformazione logaritmica della funzione di verosimiglianza

$$\ell = \ell(\theta) = \ell(\theta; y) = c'(y) + \log p_Y(y, \theta),$$

dove $c'(y) = \log c(y)$, con la convenzione che $\ell(\theta) = -\infty$ quando $L(\theta) = 0$. Se le osservazioni $y = (y_1, \ldots, y_N)$ sono i.i.d. allora

$$\ell(\theta) = \sum_{i=1}^{N} \log p_{Y_1}(y_i, \theta).$$

La funzione di log-verosimiglianza è quindi definita a meno di costanti additive che non dipendono da θ. Due log-verosimiglianze che differiscono per una sola costante additiva che non dipende da θ sono dette equivalenti.

Se si adotta il principio del campionamento ripetuto come criterio per valutare la precisione delle procedure inferenziali, è necessario studiare la distribuzione di probabilità di $\ell(\theta; Y)$, e delle quantità ad essa connesse, considerando θ fissato, al variare di y nello spazio campionario $\mathcal{Y}$, con densità $p_Y(y, \tilde{\theta})$ in $\mathcal{F}$, dove $\tilde{\theta} \in \Theta$ è un parametro non necessariamente uguale a θ. In generale, si parla di distribuzione nulla quando $\theta = \tilde{\theta}$ e di distribuzione non nulla altrimenti.

Per le considerazioni successive è necessario introdurre delle condizioni di regolarità. Si dice che il **problema di stima è regolare** quando sono soddisfatte le seguenti condizioni (Azzalini, 2001, §3.2.3):

(i) il modello statistico $\mathcal{F}$ è identificabile, ovvero esiste una corrispondenza biunivoca tra gli elementi di Θ e di $\mathcal{F}$;

(ii) lo spazio parametrico Θ è un sottoinsieme aperto dello spazio euclideo $\mathbb{R}^p$, con p fissato e finito;

(iii) gli elementi di $\mathcal{F}$ sono caratterizzati dallo stesso supporto;

(iv) con riferimento agli elementi di $\mathcal{F}$, è possibile scambiare almeno due volte il segno di integrale rispetto ad y con quello di derivata rispetto a θ.

Se il modello statistico $\mathcal{F}$ rispetta le condizioni di regolarità appena descritte, allora è possibile introdurre le seguenti quantità di verosimiglianza:

- **Funzione punteggio** (*score*): è il vettore delle derivate parziali prime di $\ell(\theta)$, indicata con $\ell_\theta(\theta) = \ell_\theta(\theta; y) = (\ell_{\theta_1}(\theta), \ldots, \ell_{\theta_p}(\theta))^T$, con generico elemento $\ell_{\theta_r}(\theta) = \partial\ell(\theta)/\partial\theta_r$, $r = 1, \ldots, p$.

- **Matrice di informazione osservata**: è la matrice Hessiana della funzione di log-verosimiglianza cambiata di segno. È una matrice simmetrica di dimensione $p \times p$ con le derivate parziali seconde di $\ell(\theta)$ e rappresenta la curvatura della funzione di log-verosimiglianza. Si definisce come $j(\theta) = -\partial^2\ell(\theta)/\partial\theta\partial\theta^T = [-\ell_{\theta_r\theta_s}]$, con $\ell_{\theta_r\theta_s} = \partial^2\ell(\theta)/\partial\theta_r\partial\theta_s$ nell'elemento di posizione (r, s).

- **Matrice di informazione attesa** (informazione di Fisher): è il valore atteso nullo della matrice di informazione osservata. Viene indicata con $i(\theta)$, dove $i(\theta) = \mathbb{E}_\theta[j(\theta)] = \mathbb{E}_\theta[j(\theta; Y)]$. Il generico elemento di $i(\theta)$ di posto (r, s) viene indicato con i_{rs}.

Assumendo che siano soddisfatte le precedenti condizioni di regolarità, in particolare la possibilità di scambiare il segno di derivata con quello di integrale, le quantità di verosimiglianza godono di due proprietà fondamentali, note anche come prime due **identità di Bartlett**:

(1) La funzione punteggio $\ell_\theta(\theta; Y)$ ha valore atteso nullo pari a 0, ossia

$$\mathbb{E}_\theta[\ell_\theta(\theta; Y)] = 0, \quad \text{per ogni } \theta \in \Theta.$$

(2) Vale l'identità dell'informazione, ossia

$$\mathrm{Var}_\theta[\ell_\theta(\theta; Y)] = \mathbb{E}_\theta[\ell_\theta(\theta; Y)\ell_\theta(\theta; Y)^T] = i(\theta), \quad \text{per ogni } \theta \in \Theta.$$

Un valore di $\theta \in \Theta$ che massimizza $L(\theta; y)$ sullo spazio parametrico Θ, ossia un valore $\hat{\theta} = \hat{\theta}(y)$ tale che $L(\hat{\theta}) \geq L(\theta)$, per ogni $\theta \in \Theta$, è chiamato **stima di massima verosimiglianza** di θ. Se $L(\hat{\theta}) = \max_{\theta \in \Theta} L(\theta)$, allora vale anche che $\ell(\hat{\theta}) = \max_{\theta \in \Theta} \ell(\theta)$. In un modello con verosimiglianza regolare, la stima di massima verosimiglianza $\hat{\theta}$ può essere cercata tra le soluzioni dell'equazione di verosimiglianza

$$\ell_\theta(\theta) = 0,$$

che corrisponde a verificare le condizioni del primo ordine. Non è detto che la stima di massima verosimiglianza esista unica. Nel seguito si assume che $\hat{\theta}$ esista e sia l'unica soluzione dell'equazione di verosimiglianza.

La variabile casuale associata a $\hat{\theta}$, $\hat{\theta}(Y)$, prende il nome di **stimatore di massima verosimiglianza**. Nel seguito, quando non vi è ambiguità, il simbolo $\hat{\theta}$ verrà utilizzato sia per indicare la stima che lo stimatore, e il significato sarà chiaro a seconda del contesto. Tale stimatore gode di alcune importanti proprietà se il modello ha verosimiglianza regolare. In particolare, $\hat{\theta}$ è asintoticamente non distorto, ossia $\mathbb{E}_\theta(\hat{\theta}(Y)) \to \theta$, quando $N \to \infty$, e al divergere di N la sua varianza tende a 0, ossia $\mathrm{Var}_\theta(\hat{\theta}) \to 0$, $N \to \infty$. La non distorsione asintotica e il fatto che la varianza tenda asintoticamente a 0 implicano la convergenza in media quadratica di $\hat{\theta}$, e di conseguenza che $\hat{\theta}$ è consistente, ovvero al divergere di N, $\hat{\theta}$ converge in probabilità al vero valore θ del parametro, e si scrive $\hat{\theta} \xrightarrow{p} \theta$. In aggiunta, lo stimatore di massima verosimiglianza è asintoticamente efficiente, nel senso che la sua varianza raggiunge il reciproco dell'informazione attesa, $\mathrm{Var}_\theta(\hat{\theta}) = i(\theta)^{-1} + o(N^{-1})$. Infine, per N sufficientemente grande, $\hat{\theta}$ ha distribuzione nulla, ossia sotto θ, approssimata

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

dove nella varianza asintotica è possibile utilizzare $j(\theta)$ o le stime $i(\hat{\theta})$, $j(\hat{\theta})$ al posto di $i(\theta)$.

1.5 Invarianza rispetto alla parametrizzazione

Dato un modello statistico parametrico $\mathcal{F}$, se questo è identificabile, allora esiste una corrispondenza biunivoca tra gli elementi di Θ e gli elementi di $\mathcal{F}$, $p(y; \theta)$. Significa che il parametro $\theta \in \Theta$ rappresenta un' etichetta per gli elementi del modello $\mathcal{F}$. In altri termini, scegliere una parametrizzazione equivale ad assegnare un sistema di coordinate che permetta di individuare in modo univoco ciascun elemento di $\mathcal{F}$.

Talvolta, potrebbe essere di interesse, sia per motivi di interpretazione che di natura matematica e/o computazionale, cambiare tale sistema di coordinate e passare ad una parametrizzazione alternativa. Quando viene scelta una parametrizzazione alternativa, è naturale richiedere che le conclusioni inferenziali non debbano cambiare rispetto alla parametrizzazione originale, poiché è solo l'etichetta data agli elementi di $\mathcal{F}$ che viene modificata.

Più formalmente, si definisce **riparametrizzazione** una qualsiasi trasformazione $\varphi = \varphi(\theta)$, dove $\varphi : \Theta \subseteq \mathbb{R}^p \to \Phi \subseteq \mathbb{R}^p$ è una funzione biunivoca e regolare, infinitamente derivabile e con inversa infinitamente derivabile.

Si parla di **invarianza rispetto alla parametrizzazione** quando le conclusioni inferenziali nella parametrizzazione φ sono le stesse di quelle ottenute nella parametrizzazione θ, e possono essere espresse in termini $\theta = \theta(\varphi)$. Il modello statistico nella parametrizzazione originale $\mathcal{F}^{\Theta} = \{p(y;\theta),\ y \in \mathcal{Y},\ \theta \in \Theta \subseteq \mathbb{R}^p\}$ può essere riscritto come

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

dove $\Phi = \{\varphi \in \mathbb{R}^p : \varphi = \varphi(\theta),\ \theta \in \Theta\}$. Dunque il modello statistico e il problema inferenziale rimangono inalterati. Le conclusioni inferenziali nella nuova parametrizzazione sono una semplice traduzione delle conclusioni inferenziali nella parametrizzazione originale.

La funzione di verosimiglianza è invariante rispetto a parametrizzazioni del modello $\mathcal{F}$, in quanto θ e $\theta(\varphi)$ corrispondono allo stesso elemento di $\mathcal{F}$. La relazione che esiste tra funzione di verosimiglianza nella nuova parametrizzazione $L^{\Phi}(\varphi)$ e la verosimiglianza nella parametrizzazione originale $L^{\Theta}(\theta)$ è la seguente

$$L^{\Phi}(\varphi) = L^{\Theta}(\theta(\varphi)),$$

e per la la funzione di log-verosimiglianza, $\ell^{\Phi}(\varphi) = \ell^{\Theta}(\theta(\varphi))$. Ciò implica che lo stimatore di massima verosimiglianza, $\hat{\theta}$ è equivariante rispetto a parametrizzazioni, ossia

$$\hat{\varphi} = \varphi(\hat{\theta}).$$

È fondamentale, pertanto, che le procedure inferenziali si comportino coerentemente quando la parametrizzazione viene cambiata.

1.6 Procedure di inferenza di verosimiglianza

L'obiettivo principale dell'inferenza statistica è quello di utilizzare il campione y per ricostruire la legge ignota $p^0(y)$ che ha generato i dati. Se il modello specificato è parametrico, ciò equivale a saper individuare correttamente il vero valore del parametro θ. Dunque, i dati y vengono utilizzati per rispondere a specifiche domande sul vero valore del parametro θ, e per associare alle risposte un'adeguata valutazione dell'incertezza. I principali problemi inferenziali sono: problemi di stima puntuale, problemi di stima intervallare o per regioni, verifica di ipotesi e previsione.

Una stima puntuale $\hat{\theta}$ rappresenta la sintesi più naturale della conoscenza sul parametro θ. Tuttavia, in genere tale stima non coincide esattamente con il vero valore del parametro, e pertanto risulta più convincente, invece che fornire la sola stima puntuale, dare un insieme di valori plausibili per il parametro θ. Secondo l'approccio frequentista, lo scopo è determinare una regione di confidenza per il vero valore del parametro, ovvero un sottoinsieme dello spazio parametrico Θ, individuato sulla base dei dati, indicato con $\hat{\Theta}(Y)$, che soddisfi

$$\mathbb{P}_\theta(\theta \in \hat{\Theta}(Y)) \geq 1 - \alpha, \quad \text{per ogni } \theta \in \Theta,$$

dove $1 - \alpha$ indica il livello di confidenza nominale della regione $\hat{\Theta}(Y)$, e $\mathbb{P}_\theta(\theta \in \hat{\Theta}(Y))$ è detta probabilità di copertura. Quando il parametro θ è scalare, $\hat{\Theta}(y)$ è tipicamente un intervallo e si parla pertanto di stima intervallare.

Un modo possibile per costruire una regione di confidenza è utilizzare una quantità pivotale. Si definisce **quantità pivotale**, e si indica con $q(\theta; y)$, ogni funzione che dipende sia dai dati y che dal parametro θ e che abbia, sotto θ, distribuzione nota che non dipende da θ.

Le quantità pivotali sono utili anche per rispondere ai problemi di verifica di ipotesi. La verifica di ipotesi viene impiegata quando è di interesse valutare se una certa supposizione sul valore del parametro θ è supportata o meno dai dati y. Si dice ipotesi nulla l'ipotesi $H_0 : \theta \in \Theta_0 \subset \Theta$, che corrisponde all'ipotesi che $p_Y(y; \theta)$ appartenga al sottomodello $\mathcal{F}_0$ che ha spazio parametrico Θ_0. In aggiunta, si definisce **statistica test** una funzione dei dati y, $t : \mathcal{Y} \to \mathbb{R}$ che individua una bipartizione dello spazio campionario $\mathcal{Y}$ in due regioni:

- regione di accettazione A_{θ_0}: è l'insieme dei dati y per i quali non c'è evidenza empirica contro H_0;

- regione di rifiuto $\bar{A}_{\theta_0} = \mathcal{Y} \setminus A_{\theta_0}$: è l'insieme dei dati y per i quali c'è evidenza empirica contro H_0.

Se i dati $y \in \bar{A}_{\theta_0}$ si dice che il test basato su t è significativo contro H_0. Dopo aver calcolato il valore della statistica test $t(y)$ sulla base dei dati osservati, invece che scegliere se rifiutare H_0 sulla base di una soglia di discriminazione per $t(y)$ scelta con riferimento alla distribuzione nulla del test, è possibile calcolare il livello di significatività osservato del test (*p-value*). Se la regione di rifiuto è unilaterale destra il *p-value* è definito come

$$\alpha^{oss} = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(t(Y) \geq t(y)),$$

se invece la regione di rifiuto è unilaterale sinistra il *p-value* è definito come

$$\alpha^{oss} = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(t(Y) \leq t(y)),$$

infine, se la regione di rifiuto è bilaterale il *p-value* è definito come

$$\alpha^{oss} = 2 \sup_{\theta \in \Theta_0} \min \{\mathbb{P}_\theta(t(Y) \leq t(y)), \mathbb{P}_\theta(t(Y) \geq t(y))\}.$$

Il livello di significatività osservato del test può essere confrontato con un livello nominale α e sulla base del risultato del confronto si sceglie se rifiutare o meno l'ipotesi nulla H_0.

La funzione di verosimiglianza può essere utilizzata per rispondere a tutti i problemi principali dell'inferenza statistica nei modelli parametrici. Ad esempio, data la funzione di verosimiglianza $L(\theta)$, un modo naturale per definire una regione di confidenza di livello $1 - \alpha$ per il parametro θ è il seguente

$$\hat{\Theta}(Y) = \{\theta \in \Theta \ : \ L(\theta) \geq cL(\hat{\theta})\},$$

dove $c \in (0, 1)$ è una soglia che deve essere fissata sulla base del livello $1 - \alpha$ desiderato. La funzione di verosimiglianza può essere utilizzata anche per definire dei test statistici. Infatti, se è di interesse valutare l'ipotesi nulla $H_0 : \theta = \theta_0$, è possibile definire la statistica test rapporto di verosimiglianza come

$$\frac{L(\hat{\theta})}{L(\theta_0)},$$

e, poiché valori grandi di questo rapporto indicano evidenza contro H_0, la regione di rifiuto sarà costituita da valori grandi di tale rapporto. Questa statistica test può essere espressa anche in termini della funzione di log-verosimiglianza, $\ell(\hat{\theta}) - \ell(\theta_0)$ e, per motivi

di convenienza matematica, si definisce statistica test log-rapporto di verosimiglianza la trasformazione monotona

$$W(\theta_0) = 2\{\ell(\hat{\theta}) - \ell(\theta_0)\}.$$

La statistica test log-rapporto di verosimiglianza è nota anche come statistica di Wilks (Wilks, 1938), dal nome dello statistico che ha dimostrato per primo che, in condizioni di regolarità, $W(\theta_0)$ ha distribuzione nulla approssimata chi-quadro con p gradi di libertà, ossia

$$W(\theta_0) \dot{\sim} \ \chi^2_p.$$

Sulla base di $W(\theta_0)$ è possibile definire una regione di rifiuto con livello approssimato α

$$\bar{A}_{\theta_0} = \{y \in \mathcal{Y} \ : \ W(\theta_0) > \chi^2_{p;1-\alpha}\},$$

dove $\chi^2_{p;1-\alpha}$ indica il quantile di livello $1-\alpha$ della distribuzione chi-quadro con p gradi di libertà. In modo analogo, è possibile costruire delle regioni di confidenza con livello approssimato $1-\alpha$ come

$$\hat{\Theta}(y) = \{\theta \in \Theta \ : \ W(\theta) \leq \chi^2_{p;1-\alpha}\}, \tag{1.1}$$

che può essere riscritta equivalentemente come

$$\hat{\Theta}(y) = \{\theta \in \Theta \ : \ \ell(\theta) \geq \ell(\hat{\theta}) - \frac{1}{2}\chi^2_{p;1-\alpha}\}.$$

Esistono altri due test connessi alla funzione di verosimiglianza che risultano essere asintoticamente equivalenti a $W(\theta_0)$, e quindi con distribuzione approssimata nulla nota:

- il test di Wald

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

- il test di Rao (test score)

$$W_u(\theta_0) = \ell_\theta(\theta_0)^T i(\theta_0)^{-1} \ell_\theta(\theta_0).$$

In entrambi i casi è possibile sostituire $i(\theta_0)$ con $j(\hat{\theta})$ anche se tipicamente in $W_u(\theta_0)$ si preferisce utilizzare $i(\theta_0)$ in modo che la statistica test non richieda il calcolo dalla

stima $\hat{\theta}$.

Le tre quantità $W(\theta_0)$, $W_e(\theta_0)$ e $W_u(\theta_0)$ consentono di valutare la distanza tra $\hat{\theta}$ e θ secondo metriche diverse. Il test di Wald e il test score possono essere utilizzati al posto di $W(\theta_0)$ sia in problemi di verifica d'ipotesi che per la costruzione di regioni di confidenza. Tali quantità sono asintoticamente equivalenti e coincidono in genere solo quando $N \to \infty$. Al contrario, per N finito, possono manifestare comportamenti differenti e presentano alcune caratteristiche distintive. Solo $W(\theta_0)$ e $W_u(\theta_0)$ sono invarianti rispetto a riparametrizzazioni. D'altra parte, $W_e(\theta_0)$ è facile da calcolare ed interpretare. Tuttavia, poiché le regioni di confidenza costruite sulla base di $W_e(\theta_0)$ sono simmetriche rispetto a $\hat{\theta}$ ed ellittiche, l'insieme di valori di θ tali che $W_e(\theta_0) \leq c$ può includere valori non ammissibili dello spazio parametrico Θ. Il test di Rao, $W_u(\theta_0)$, non richiede il calcolo della stima $\hat{\theta}$ ma potrebbe risultare numericamente instabile. Infine, la distribuzione nulla approssimata χ^2_p risulta tipicamente più accurata per $W(\theta_0)$.

Quando il modello è caratterizzato da un parametro θ scalare, $p = 1$, potrebbe essere di interesse valutare ipotesi alternative unilaterali. È possibile quindi definire le versioni unilaterali delle statistiche $W(\theta_0)$, $W_e(\theta_0)$ e $W_u(\theta_0)$, rispettivamente $r(\theta_0)$, $r_e(\theta_0)$ e $r_u(\theta_0)$, per verificare ipotesi $H_0 : \ \theta = \theta_0$ contro ipotesi alternative $H_1 : \ \theta < \theta_0$ oppure $H_1 : \ \theta > \theta_0$. Le versioni unilaterali sono definite come

$$\begin{aligned} r(\theta_0) &= \text{sgn}(\hat{\theta} - \theta_0)\sqrt{W(\theta_0)}, \\ r_e(\theta_0) &= \sqrt{i(\theta_0)}(\hat{\theta} - \theta_0), \\ r_u(\theta_0) &= \frac{\ell_\theta(\theta_0)}{\sqrt{i(\theta_0)}}, \end{aligned} \tag{1.2}$$

e hanno tutte distribuzione nulla approssimata normale standard, $N(0, 1)$. La statistica $r(\theta_0)$ è nota come radice con segno del log-rapporto di verosimiglianza. Anche per le versioni unilaterali è possibile sostituire $i(\theta_0)$ con $j(\hat{\theta})$.

Ad esempio, il test di livello approssimato $1 - \alpha$ per $H_0 : \ \theta \leq \theta_0$ contro $H_1 : \ \theta > \theta_0$ basato sulla statistica radice con segno della statistica log-rapporto di verosimiglianza $r(\theta_0)$ porta a rifiutare H_0 se $r(\theta_0) > z_{1-\alpha}$, dove $z_{1-\alpha}$ è il quantile di livello $1 - \alpha$ di una normale standard, $N(0, 1)$. L'intervallo di confidenza con livello approssimato $1 - \alpha$ per θ è dato dalla (1.1), che si può esprimere in modo equivalente come

$$\hat{\Theta}(y) = \{\theta \in \Theta \ : \ |r(\theta)| \leq z_{1-\alpha/2}\}.$$

1.7 Inferenza di verosimiglianza in presenza di parametri di disturbo

1.7.1 Parametri di disturbo

Il modello statistico parametrico $\mathcal{F}$ dovrebbe rappresentare una semplificazione della realtà di interesse in grado di cogliere gli aspetti essenziali di quest'ultima. Tanto più il fenomeno di interesse è complesso, tanto più c'è il rischio di dover introdurre nel modello un numero elevato di parametri al fine di riuscire a mimare la complessità della realtà. Tuttavia, raramente l'interesse è rivolto a fare inferenza sull'intero parametro $\theta \in \Theta \subseteq \mathbb{R}^p$, quando la dimensione $p > 1$ del parametro è elevata. Tipicamente, solo alcuni degli aspetti sono di primario interesse, mentre la restante componente del parametro è necessaria affinché il modello sia in grado di catturare la complessità della realtà. Gli aspetti di primario interesse sono descritti dai **parametri di interesse**, mentre gli aspetti accessori sono descritti dai **parametri di disturbo**. L'uso del termine disturbo è convenzionale, ma sicuramente va detto che se tali parametri fossero noti, allora l'inferenza sulla componente di interesse sarebbe più efficace e semplice. Quando, come succede nella realtà, la componente di disturbo non è nota, è necessario che le procedure inferenziali risultino valide per diversi possibili valori dei parametri di disturbo.

Più formalmente, si consideri la partizione $\theta = (\psi, \lambda)$, dove ψ denota il parametro di interesse k-dimensionale, $1 \leq k < p$, mentre λ rappresenta il parametro di disturbo $(p-k)$-dimensionale. Si assume inoltre che i parametri ψ e λ siano a variazione indipendente, ossia $\Theta = \Psi \times \Lambda$, con $\psi \in \Psi \subseteq \mathbb{R}^k$ e $\lambda \in \Lambda \subseteq \mathbb{R}^{p-k}$.

Si assuma, per semplicità, che i dati $y = (y_1, \ldots, y_N)$ siano N realizzazioni indipendenti con funzione di densità congiunta

$$p_Y(y; \theta) = \prod_{i=1}^{N} p_{Y_1}(y_i; \theta_i).$$

Si considerino i due scenari estremi

- $\theta_i = (\psi, \lambda)$, per ogni $i = 1, \ldots, N$,
- $\theta_i = (\psi, \lambda_i)$, per ogni $i = 1, \ldots, N$.

Nel primo caso la distribuzione marginale di ogni osservazione è indicizzata in modo omogeneo dallo stesso parametro di disturbo e dunque la dimensione del parametro θ è fissata e non dipende dalla numerosità campionaria N. Al contrario, nel secondo caso,

il parametro di interesse rimane invariato per ciascuna osservazione, mentre si introduce un parametro di disturbo specifico per ciascuna osservazione, $\theta = (\psi, \lambda_1, \ldots, \lambda_N)$. Nel secondo caso quindi la dimensione del parametro θ è legata alla dimensione del parametro di disturbo e dipende dalla numerosità campionaria N. Il parametro di interesse ψ che riflette caratteristiche comuni a ciascuna osservazione è noto come **parametro strutturale**, mentre i parametri di disturbo λ_i prendono il nome di **parametri incidentali** (Neyman & Scott, 1948).

I problemi in cui il la dimensione dello spazio parametrico dipende dalla numerosità campionaria N sono problemi di stima non regolare, in quanto viene a mancare la seconda condizione di regolarità (ii) riportata a pagina 6, e sono noti in letteratura come **problemi di Neyman e Scott**. D'altra parte, sebbene il riferimento a tali problemi come problemi di Neyman e Scott sia ampiamente utilizzato in letteratura, queste tematiche in realtà erano già state affrontate in precedenza anche da Bartlett (1937), come sottolineato ad esempio in Battey & Cox (2020, 2022). La presenza di parametri incidentali pone nuovi problemi alle procedure inferenziali standard che si basano sull'assunzione che il parametro θ abbia dimensione fissata. In questi contesti, l'utilizzo delle usuali procedure inferenziali basate sulla verosimiglianza può portare a risultati fallimentari.

1.7.2 Riduzione del modello in presenza di parametri di disturbo

Si consideri il problema di riduzione del modello statistico $\mathcal{F}$ con parametro $\theta = (\psi, \lambda)$, in cui solamente ψ è di interesse per l'inferenza. Date due statistiche U e V, si dice che, dato $V = v$, u è parzialmente non informativa per ψ se la distribuzione di U condizionatamente a $V = v$ non dipende da ψ (sebbene possa dipendere da λ).

Per indagare sulle possibili riduzioni del modello $\mathcal{F}$ è conveniente considerare la seguente fattorizzazione della densità congiunta di $y = (y_1, \ldots, y_N)$

$$p_Y(y; \psi, \lambda) = p_V(v; \psi, \lambda) p_{U|V}(u; v, \psi, \lambda) p_{Y|U,V}(y; u, v, \psi, \lambda), \tag{1.3}$$

a cui corrisponde un'analoga fattorizzazione della funzione di verosimiglianza

$$L_Y(\psi, \lambda) = L_V(\psi, \lambda) L_{U|V}(\psi, \lambda) L_{Y|U,V}(\psi, \lambda).$$

Se (u, v) è in corrispondenza biunivoca con y, o se comunque è statistica sufficiente per θ, allora è equivalente ad y per l'inferenza su θ. Nel modello con densità $p_{U,V}(u, v; \psi, \lambda)$,

sotto generali condizioni di regolarità, vale la seguente fattorizzazione

$$p_{U,V}(u, v; \psi, \lambda) = p_V(v; \psi, \lambda) p_{U|V=v}(u; v, \psi, \lambda). \tag{1.4}$$

Esistono due casi speciali in cui è possibile ottenere un modello per riduzione che dipende solo da ψ:

- con riferimento alla fattorizzazione (1.4), se la densità marginale di V non dipende da λ, allora

  $$p_{U,V}(u, v; \psi, \lambda) = p_V(v; \psi) p_{U|V=v}(u; v, \psi, \lambda),$$

  dove il modello marginale per V, $p_V(v; \psi)$, è indicizzato solamente dal parametro di interesse ψ. La statistica v è detta **parzialmente costante in distribuzione** per λ;

- con riferimento alla fattorizzazione (1.4), se la densità condizionata di U dato $V = v$ non dipende da λ, allora

  $$p_{U,V}(u, v; \psi, \lambda) = p_V(v; \psi, \lambda) p_{U|V=v}(u; v, \psi),$$

  dove il modello condizionato $p_{U|V=v}(u; v, \psi)$ è indicizzato solamente dal parametro di interesse ψ. La statistica v è detta **parzialmente sufficiente** per λ.

In generale, data la fattorizzazione (1.4), è possibile che vi sia della perdita di informazione relativamente a ψ quando si trascurano i termini $p_{U|V=v}(u; v, \psi, \lambda)$ o $p_V(v; \psi, \lambda)$. Tuttavia, ciò non si verifica in due scenari speciali:

(a) quando v è sia parzialmente sufficiente per ψ che parzialmente costante in distribuzione per λ

$$p_{U,V}(u, v; \psi, \lambda) = p_V(v; \psi) p_{U|V=v}(u; v, \lambda), \tag{1.5}$$

(b) quando u è sia parzialmente costante in distribuzione per ψ che parzialmente sufficiente per λ

$$p_Y(u, v; \psi, \lambda) = p_U(u; \lambda) p_{Y|U=u}(y; u, \psi). \tag{1.6}$$

Nei casi speciali (1.5) e (1.6) la funzione di log-verosimiglianza $\ell(\theta) = \ell(\psi, \lambda)$ può essere scritta come

$$\ell(\theta) = \ell_1(\psi) + \ell_2(\lambda), \tag{1.7}$$

e si dice che $L(\theta)$ ha **parametri separabili**. In tal caso l'inferenza su ψ può essere condotta separatamente dall'inferenza su λ, come se questo fosse noto, con una notevole semplificazione. Infatti, nel caso di separazione, la stima vincolata di λ per ψ fissato, $\hat{\lambda}_\psi$, è equivalente alla stima globale per λ, $\hat{\lambda}$. Pertanto, in questo caso, la funzione di log-verosimiglianza profilo per ψ, $\ell_P(\psi)$, definita nel paragrafo 1.7.4, corrisponde al primo addendo, $\ell_1(\psi)$, della scomposizione (1.7).

Sfortunatamente, i casi in cui la verosimiglianza ha parametri separabili sono molto rari nella pratica. Pertanto, quando tale separazione non avviene, una possibilità è quella di basare l'inferenza su un fattore che dipende dal solo parametro di interesse ψ. Chiaramente è necessario che tale semplificazione avvenga al prezzo di una perdita trascurabile di informazione su ψ. Ciò si verifica solamente quando è possibile individuare una statistica parzialmente sufficiente per λ, tramite condizionamento, oppure quando è possibile individuare una statistica parzialmente costante in distribuzione per λ, tramite marginalizzazione.

1.7.3 Pseudo-verosimiglianze

Si consideri un modello statistico $\mathcal{F}$ con spazio parametrico Θ. Sia $\psi = \psi(\theta)$, con $\psi \in \Psi \subseteq \mathbb{R}^k$ il parametro di interesse. Più la struttura della componente del parametro θ complementare a ψ è complessa e più diventa interessante la possibilità di basare l'inferenza su una funzione di verosimiglianza che dipenda solo da ψ. Tale riduzione di complessità deve avvenire al prezzo di una perdita trascurabile di informazione sul parametro ψ.

Si definisce **pseudo-verosimiglianza** ogni funzione che dipende dal parametro di interesse e dai dati e che si comporta, almeno approssimativamente, come se fosse una verosimiglianza propria (ossia se rispetta le usuali proprietà: la funzione punteggio ha valore atteso nullo, vale l'identità dell'informazione, lo stimatore di massima verosimiglianza è asintoticamente normale, la statistica log rapporto di verosimiglianza ha distribuzione nulla approssimata chi-quadro, etc.).

Esistono essenzialmente due scenari:

(a) la funzione di pseudo-verosimiglianza è basata su un sottomodello di $\mathcal{F}$ ottenuto per riduzione in cui gli elementi dipendono solo dal parametro di interesse ψ. La

pseudo-verosimiglianza ottenuta in questo modo è una verosimiglianza propria e soddisfa le usuali proprietà. L'assunzione cruciale per le considerazioni asintotiche è che l'informazione nel modello ridotto rimanga di ordine $O(N)$;

(b) la funzione di pseudo-verosimiglianza non deriva da un sotto-modello di $\mathcal{F}$ ottenuto per riduzione. In altri termini, tale pseudo-verosimiglianza non discende da una fattorizzazione del tipo (1.3). Le proprietà inferenziali devono essere verificate caso per caso.

Ne consegue che, ove possibile, è preferibile individuare una pseudo-verosimiglianza propria. Questo si verifica, ad esempio, nel caso della verosimiglianza marginale e della verosimiglianza condizionata. Queste due pseudo-verosimiglianze sono definite nel modo seguente:

- sia y in corrispondenza biunivoca con (u, v), o più in generale, sia (u, v) una statistica sufficiente per θ per cui è soddisfatta la seguente fattorizzazione

$$p_{U,V}(u, v; \psi, \lambda) = p_V(v; \psi) p_{U|V=v}(u; v, \psi, \lambda),$$

dove v è parzialmente costante in distribuzione per λ. Se il contributo alla verosimiglianza che corrisponde al termine $p_{U|V=v}(\cdot)$ è trascurabile, l'inferenza su ψ può essere basata sul modello marginale per V con densità $p_V(v; \psi)$. La corrispondente funzione di verosimiglianza

$$L_M(\psi) = L_M(\psi; v) = p_V(v; \psi), \tag{1.8}$$

è detta funzione di **verosimiglianza marginale**, basata su v.

- sia u una statistica per cui è soddisfatta la fattorizzazione

$$p_Y(u, v; \psi, \lambda) = p_U(u; \psi, \lambda) p_{Y|U=u}(y; u, \psi),$$

dove u è parzialmente sufficiente per λ. Se il contributo alla verosimiglianza che corrisponde al termine $p_U(\cdot)$ è trascurabile, l'inferenza su ψ può essere basata sul modello condizionato con densità $p_{Y|U=u}(\cdot)$. La corrispondente funzione di verosimiglianza

$$L_C(\psi) = L_C(\psi; y|u) = p_{Y|U=u}(y; u, \psi), \tag{1.9}$$

è detta funzione di **verosimiglianza condizionata**, basata sul condizionamento ad u.

Si assume che nella (1.8) e nella (1.9) i fattori non coinvolti, $p_{U|V=v}(u; v, \psi, \lambda)$ e $p_U(u; \psi, \lambda)$ rispettivamente, siano trascurabili in quanto si suppone che rappresentino assenza di informazione sul parametro di interesse, ovvero che la statistica V esaurisca il suo compito di estrarre informazione dal parametro di interesse ψ nella marginalizzazione, e rispettivamente U nel condizionamento. Per una trattazione più approfondita di questo tema si rimanda a Pace & Salvan (1997, Capitolo 4) e Jørgensen (1993). Si veda anche Zhu & Reid (1994) e Jørgensen & Labouriau (2012, Capitolo 3). Le fattorizzazioni (1.8) e (1.9) che danno luogo rispettivamente alle pseudo-verosimiglianza marginale e condizionata si verificano essenzialmente soltanto nel caso delle famiglie di gruppo e delle famiglie esponenziali, per specifiche definizioni del parametro di interesse ψ. Qualora questo tipo di separazione nell'inferenza non dovesse essere disponibile, è necessario ricercare una pseudo-verosimiglianza al di fuori della classe delle verosimiglianza proprie. Ad esempio, un modo molto più generale per ottenere una funzione di pseudo-verosimiglianza in un modello parametrico è quello di sostituire il parametro di disturbo λ con una sua stima consistente nella funzione di verosimiglianza $L(\psi, \lambda)$. Una possibilità è quella di utilizzare la stima di massima verosimiglianza per λ con ψ fissato, $\hat{\lambda}_\psi$. In tal caso si parla di **verosimiglianza profilo**. Un'altra possibilità è quella di utilizzare la stima non vincolata $\hat{\lambda}$ al posto di λ, ottenendo così la pseudo-verosimiglianza di Gong e Samaniego (Gong & Samaniego, 1981), che tuttavia sovrastima l'informazione su ψ, a meno che ψ e λ non siano ortogonali.

1.7.4 Verosimiglianza profilo

Un metodo generale per la costruzione di una pseudo-verosimiglianza per il parametro di interesse ψ prevede di sostituire il parametro di disturbo λ con una sua stima consistente nella funzione di verosimiglianza $L(\psi, \lambda)$. Se viene utilizzata la stima di massima verosimiglianza per λ nel sotto-modello con ψ fissato, $\hat{\lambda}_\psi$, allora si ottiene la **verosimiglianza profilo**

$$L_P(\psi) = L(\psi, \hat{\lambda}_\psi). \tag{1.10}$$

La corrispondente funzione di log-verosimiglianza profilo è $\ell_P(\psi) = \log L(\psi, \hat{\lambda}_\psi)$. Se le condizioni di regolarità sono soddisfatte, la stima $\hat{\lambda}_\psi$ è ottenuta come soluzione in λ della funzione punteggio parziale relativa al sotto-modello con ψ fissato, ossia $\ell_\lambda(\psi, \lambda) = 0$. La verosimiglianza profilo non può essere considerata una verosimiglianza propria,

in quando non discende direttamente da una funzione di densità. Nonostante ciò, la verosimiglianza profilo gode di alcune interessanti proprietà che la rendono simile ad una verosimiglianza propria:

- la stima di massima verosimiglianza profilo coincide con la stima di massima verosimiglianza per ψ ottenuta con $L(\psi, \lambda)$, ossia

$$\sup_{\psi} L_P(\psi) = L(\hat{\psi});$$

- si può definire l'**informazione osservata profilo**

$$j_P(\psi) = -\frac{\partial^2}{\partial\psi\partial\psi^T}\ell_P(\psi),$$

che risulta espressa da

$$j_P(\psi) = j_{\psi\psi}(\psi, \hat{\lambda}_\psi) - j_{\psi\lambda}(\psi, \hat{\lambda}_\psi) j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)^{-1} j_{\lambda\psi}(\psi, \hat{\lambda}_\psi),$$

dove $j_{\psi\psi}$, $j_{\psi\lambda}$ e $j_{\lambda\lambda}$ sono i blocchi di $j(\theta)$

$$j(\theta) = \begin{bmatrix} j_{\psi\psi}(\theta) & j_{\psi\lambda}(\theta) \\ j_{\psi\lambda}(\theta) & j_{\lambda\lambda}(\theta) \end{bmatrix}.$$

L'inversa di $j_P(\psi)$ è uguale al blocco (ψ, ψ) dell'inversa della matrice di informazione osservata complessiva calcolata in $(\psi, \hat{\lambda}_\psi)$, ossia

$$[j_P(\psi)]^{-1} = j^{\psi\psi}(\psi, \hat{\lambda}_\psi),$$

dove $j^{\psi\psi}$ denota il blocco (ψ, ψ) di $j(\theta)^{-1}$;

- il log-rapporto di verosimiglianza profilo coincide con il log-rapporto di verosimiglianza basato su $L(\psi, \lambda)$ usato per la verifica d'ipotesi su ψ, con λ ignoto. Dunque, il **test del log-rapporto di verosimiglianza profilo** è

$$W_P(\psi) = 2\{\ell_P(\hat{\psi}) - \ell_P(\psi)\} = 2\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi, \hat{\lambda}_\psi)\}.$$

Sotto le usuali condizioni di regolarità $W_P(\psi)$ ha distribuzione asintotica nulla χ^2_k dove k è la dimensione del parametro d'interesse. Inoltre, come nel caso generale in cui è di interesse fare inferenza su tutto il parametro θ, è possibile definire due test asintoticamente equivalenti a $W_P(\psi)$, rispettivamente il test di Wald $W_{Pe}(\psi)$

e il test *score* $W_{Pu}(\psi)$

$$W_{Pe}(\psi) = (\hat{\psi} - \psi)^T [i^{\psi\psi}(\psi, \hat{\lambda}_\psi)]^{-1} (\hat{\psi} - \psi),$$
$$W_{Pu}(\psi) = \ell_\psi(\psi, \hat{\lambda}_\psi)^T i^{\psi\psi}(\psi, \hat{\lambda}_\psi) \ell_\psi(\psi, \hat{\lambda}_\psi),$$

dove $i^{\psi\psi}(\psi, \hat{\lambda}_\psi)$ denota il blocco (ψ, ψ) dell'inversa di $i(\theta)$, $i(\theta)^{-1}$, e può essere sostituito da $[j_P(\hat{\psi})]^{-1}$. Come $W_P(\psi)$, anche $W_{Pe}(\psi)$ e $W_{Pu}(\psi)$ hanno distribuzione asintotica nulla χ^2_k. Se il parametro di interesse ψ è scalare, è possibile definire le relative versioni unilaterali

$$r_P(\psi) = \text{sgn}(\hat{\psi} - \psi)\sqrt{W_P(\psi)}, \tag{1.11}$$
$$r_{Pe}(\psi) = (\hat{\psi} - \psi)[i^{\psi\psi}(\psi, \hat{\lambda}_\psi)]^{-1/2}, \tag{1.12}$$
$$r_{Pu}(\psi) = \ell_\psi(\psi, \hat{\lambda}_\psi)[i^{\psi\psi}(\psi, \hat{\lambda}_\psi)]^{1/2},$$

e tutte con distribuzione nulla approssimata normale standard. Come per le versioni bilaterali, $i^{\psi\psi}(\psi, \hat{\lambda}_\psi)$ può essere sostituito da $[j_P(\hat{\psi})]^{-1}$.

In modo simile a $W(\theta)$, il test basato su $W_P(\psi)$ è una quantità asintoticamente pivotale per ψ. Quindi è possibile costruire regioni di confidenza di livello approssimato $1 - \alpha$ per ψ, nel modo seguente

$$\hat{\Psi}(y) = \{\psi \in \Psi \ : \ W_p(\psi) \leq \chi^2_{k;1-\alpha}\}, \tag{1.13}$$

che può essere riscritta equivalentemente come

$$\hat{\Psi}(y) = \{\psi \in \Psi \ : \ \ell_P(\psi) \geq \ell_P(\hat{\psi}) - \frac{1}{2}\chi^2_{k;1-\alpha}\}.$$

Se ψ è scalare allora è possibile definire le regioni di confidenza di livello approssimato $1 - \alpha$ utilizzando anche le versioni unilaterali, ad esempio

$$\hat{\Psi}(y) = \{\psi \in \Psi \ : \ |r_p(\psi)| \leq z_{1-\alpha/2}\},$$

che è equivalente alla regione di confidenza (1.13).

Ovviamente anche in presenza di parametri di disturbo si potrebbe essere interessati ad un cambio di parametrizzazione, in particolare per il parametro di interesse. In tal caso si desidera che la nuova parametrizzazione non alteri in modo sostanziale le conclusioni inferenziali ottenute nella parametrizzazione originale. Tuttavia, poiché in generale una riparametrizzazione globale per $\theta = (\psi, \lambda)$ non mantiene la distinzione tra

parametro d'interesse e parametro di disturbo, tipicamente si restringe la richiesta di invarianza alla parametrizzazione alle sole riparametrizzazioni che non alterano l'interesse. Una riparametrizzazione che non altera l'interesse è una trasformazione del tipo $\varphi = \varphi(\psi, \lambda)$ con $\varphi = (\rho, \xi)$, tale che

$$\rho = \rho(\psi), \qquad \xi = \xi(\psi, \lambda),$$

dove $\rho(\cdot)$ è una funzione biunivoca che dipende solo da ψ. Si chiede quindi sia soddisfatto il principio di **invarianza rispetto alle riparametrizzazioni che non alterano l'interesse**. La verosimiglianza profilo è invariante a riparametrizzazioni che non alterano l'interesse, a differenza di altre pseudo-verosimiglianze.

Le proprietà appena descritte rendono la verosimiglianza profilo interessante. Ciò nonostante, la verosimiglianza profilo non è assimilabile ad una verosimiglianza propria e pertanto non gode di tutte le proprietà di una verosimiglianza in senso proprio. In particolare, vengono a mancare le proprietà relative al fatto che la funzione punteggio dovrebbe avere valore atteso nullo e soddisfare l'identità dell'informazione. Infatti, sotto campionamento casuale semplice con numerosità N,

$$\begin{aligned} \mathbb{E}_{\psi,\lambda}\Big[\frac{\partial}{\partial\psi}\ell_P(\psi)\Big] &= O(1), \\ \mathbb{E}_{\psi,\lambda}\Big[\frac{\partial^2}{\partial\psi\partial\psi^T}\ell_P(\psi)\Big] + \mathbb{E}_{\psi,\lambda}\Big[\Big\{\frac{\partial}{\partial\psi}\ell_P(\psi)\Big\}\Big\{\frac{\partial}{\partial\psi}\ell_P(\psi)\Big\}^T\Big] &= O(1), \end{aligned} \tag{1.14}$$

dove $O(1)$ denota una quantità asintoticamente limitata in probabilità, e si parla rispettivamente di *score bias* e *information bias*. Nelle situazioni standard, dove *score bias* e *information bias* sono effettivamente di ordine $O(1)$, lo stimatore di massima verosimiglianza per il parametro di interesse ψ rimane consistente ed è garantita l'usuale validità delle approssimazioni asintotiche per le distribuzioni di $W_P(\psi)$, $W_{Pe}(\psi)$ e $W_{Pu}(\psi)$.

Tuttavia, nel caso di modelli stratificati in presenza di parametri incidentali, o più in generale quando la dimensione del parametro di disturbo λ è grande rispetto alla numerosità campionaria N, i problemi relativi a *score bias* e *information bias* possono aggravarsi. Come verrà descritto nel paragrafo (2.2), nel caso di modelli stratificati, utilizzare la verosimiglianza profilo $L_P(\psi)$, trattando la componente di disturbo λ come se fosse nota e pari a $\hat{\lambda}_\psi$, non è ragionevole se i dati non contengono informazione a sufficienza sulla componente di disturbo. Lo stesso problema si presenta, in modo ancora più evidente, nel caso dei modelli con effetti fissi incrociati, discussi nel Capitolo 3.

Per compensare questa mancanza di informazione, in letteratura sono stati proposti

alcuni metodi di modificazione della verosimiglianza profilo.

1.8 Modificazioni della verosimiglianza profilo

Come illustrato nel paragrafo 1.7.4, in alcune situazioni, ad esempio quando la dimensione del parametro di disturbo λ è grande se confrontata con la numerosità campionaria N, le procedure inferenziali basate sulla verosimiglianza profilo potrebbero non essere accurate. Tra le alternative all'utilizzo della usuale verosimiglianza profilo, è possibile: apportare delle modifiche analitiche, in particolare ricorrere a funzioni di verosimiglianza profilo modificate, o a modificazioni delle statistiche test profilo, in modo tale che l'approssimazione normale per tali statistiche possa risultare accurata anche quando la numerosità campionaria è esigua. Nel seguito si discuteranno gli aspetti essenziali di entrambi gli approcci. Per una trattazione più approfondita in merito a questi argomenti si veda Barndorff-Nielsen & Cox (1994, Capitolo 8) e Severini (2000, Capitolo 9).

1.8.1 Verosimiglianza profilo modificata

Una prima alternativa all'usuale funzione di verosimiglianza profilo (1.10), nei contesti in cui la dimensione del parametro di disturbo λ è grande relativamente ad N, consiste nel considerare modificazioni analitiche di tale verosimiglianza. In letteratura, sono diverse le proposte che sono state fatte al fine di aggiustare il comportamento della verosimiglianza profilo. Tra queste, la verosimiglianza profilo modificata proposta da Barndorff-Nielsen (1980, 1983) e la verosimiglianza profilo condizionata approssimata, introdotta da Cox & Reid (1987).

Se $\theta = (\psi, \lambda)$ con $\psi \in \Psi \subseteq \mathbb{R}^k$ parametro di interesse e $\lambda \in \Lambda \subseteq \mathbb{R}^{p-k}$ parametro di disturbo, la **verosimiglianza profilo modificata** è definita come

$$L_M(\psi) = L_P(\psi) M(\psi),$$

dove $M(\psi)$ è un fattore di aggiustamento, dato da

$$M(\psi) = |\ell_{\lambda;\hat{\lambda}}(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\lambda}, a)|^{-1} |j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\lambda}, a)|^{1/2},$$

o, in forma più compatta, con $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$ stima vincolata,

$$M(\psi) = |\ell_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi)|^{-1} |j_{\lambda\lambda}(\hat{\theta}_\psi)|^{1/2}, \tag{1.15}$$

dove nella (1.15) nel primo determinante compare la derivata mista (*mixed derivative*)

$$\ell_{\lambda;\hat{\lambda}}(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\lambda}, a) = \frac{\partial}{\partial\lambda\partial\hat{\lambda}^T}\ell(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\lambda}, a) = \frac{\partial}{\partial\lambda\partial\hat{\lambda}^T}\ell(\hat{\theta}_\psi; \hat{\theta}, a),$$

e viene esplicitata la dipendenza dalla statistica ancillare a. La motivazione della presenza di questa statistica è data dal fatto che se la stima di massima verosimiglianza di θ è unica, allora è necessariamente funzione della statistica sufficiente minimale s, ma non è detto che $\hat{\theta}$ mantenga la sufficienza, come spiegato nel paragrafo 1.3.1. Di conseguenza, la funzione di log-verosimiglianza può essere scritta in funzione della statistica sufficiente minimale $(\hat{\theta}, a)$

$$\ell(\theta; y) = \ell(\theta; s) = \ell(\theta; \hat{\theta}, a) = \ell(\psi, \lambda; \hat{\psi}, \hat{\lambda}, a).$$

Ne segue che la funzione di log-verosimiglianza profilo modificata può essere definita come

$$\begin{aligned}\ell_M(\psi) &= \ell_P(\psi) + \log M(\psi) \\ &= \ell_P(\psi) + \frac{1}{2}\log|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\lambda}, a)| - \log|\ell_{\lambda;\hat{\lambda}}(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\lambda}, a)| \\ &= \ell_P(\psi) + \frac{1}{2}\log|j_{\lambda\lambda}(\hat{\theta}_\psi)| - \log|\ell_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi)|. \end{aligned} \tag{1.16}$$

Il fattore di modificazione $M(\psi)$ è di ordine $O(1)$, e quindi la verosimiglianza profilo $L_P(\psi)$ e la verosimiglianza profilo modificata $L_M(\psi)$ risultano asintoticamente equivalenti al primo ordine. Tuttavia, il fattore $M(\psi)$ permette di ottenere una riduzione della distorsione della funzione punteggio (1.14) da $O(1)$ a $O(N^{-1})$. Inoltre, come la verosimiglianza profilo $L_P(\psi)$, anche $L_M(\psi)$ è invariante rispetto a parametrizzazioni che non alterano l'interesse. In particolare, nel caso di parametri separabili, poiché vale la scomposizione (1.7), allora

$$\ell_M(\psi) = \ell_P(\psi) = \ell_1(\psi),$$

ossia, sia la log-verosimiglianza profilo che la log-verosimiglianza profilo modificata sono equivalenti all'addendo della funzione di log-verosimiglianza che dipende solo da ψ.

Per il calcolo della verosimiglianza o log-verosimiglianza profilo modificata è necessario calcolare: la stima di massima verosimiglianza globale $(\hat{\psi}, \hat{\lambda})$, la stima vincolata $(\psi, \hat{\lambda}_\psi)$, il blocco (λ, λ) della matrice di informazione osservata $j(\theta)$, relativo solo alla

componente di disturbo, e la derivata nello spazio campionario $\ell_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi)$. Inoltre, quando esiste una verosimiglianza marginale (1.8) o una verosimiglianza condizionata (1.9) per ψ, la verosimiglianza profilo modificata è tipicamente semplice da calcolare ed è una buona approssimazione di esse. In realtà, anche quando non esiste la verosimiglianza marginale o la verosimiglianza condizionata per ψ, non è difficile ottenere la verosimiglianza profilo modificata se il modello sottostante appartiene ad una famiglia esponenziale o ad una famiglia di gruppo. In particolare, se il modello è una famiglia esponenziale è ψ è una componente del parametro canonico, allora si può mostrare che la derivata nello spazio campionario $\ell_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi)$ non dipende da ψ e quindi il contributo dato da $|\ell_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi)|^{-1}$ nel fattore di modificazione (1.15) può essere trascurato. In tal caso la log-verosimiglianza profilo modificata può essere calcolata semplicemente come

$$\ell_P(\psi) + \frac{1}{2}\log|j_{\lambda\lambda}(\hat{\theta}_\psi)|,$$

che è un'approssimazione della funzione di log-verosimiglianza condizionata.

La **verosimiglianza condizionata approssimata** (Cox & Reid, 1987) è stata introdotta come possibile approccio per cercare di ricondursi ad una situazione semplificata in cui la verosimiglianza ha approssimativamente parametri separabili, come descritto nella (1.7). In particolare, la verosimiglianza condizionata approssimata è ottenibile a partire da una parametrizzazione in cui le due componenti di $\theta = (\psi, \lambda)$ sono ortogonali. Si dice i parametri ψ e λ sono ortogonali se le corrispondenti componenti della funzione punteggio ℓ_ψ e ℓ_λ sono incorrelate, ossia se $i_{\psi\lambda}(\psi, \lambda) = 0$. Sotto ortogonalità, come mostrato ad esempio in Severini (2000, §9.5), $\hat{\lambda}_\psi = \hat{\lambda} + O(N^{-1})$ e una possibile approssimazione della log-verosimiglianza profilo modificata è data da

$$\ell_{CA}(\psi) = \ell_P(\psi) - \frac{1}{2}\log|j_{\lambda\lambda}(\hat{\theta}_\psi)|, \tag{1.17}$$

che Cox & Reid (1987) hanno chiamato log-verosimiglianza condizionata approssimata. Si noti, infatti, che la (1.16) può essere scritta in modo alternativo come

$$L_M(\psi) = L_P(\psi)D(\psi)|j_{\lambda\lambda}(\hat{\theta}_\psi)|^{-1/2},$$

con

$$D(\psi) = \frac{|j_{\lambda\lambda}(\hat{\theta}_\psi)|}{|\ell_{\lambda;\hat{\lambda}}(\hat{\theta}_\psi)|} = \Big|\frac{\partial\hat{\lambda}_\psi}{\partial\hat{\lambda}}\Big|.$$

Dunque, la (1.16) può essere anche scritta nel seguente modo

$$\ell_M(\psi) = \ell_P(\psi) + \log D(\psi) - \frac{1}{2}|j_{\lambda\lambda}(\hat{\theta}_\psi)|,$$

ma, quando $\hat{\lambda}_\psi = \hat{\lambda}$, la matrice $[\partial\hat{\lambda}_\psi/\partial\hat{\lambda}]$ è la matrice identità, pertanto

$$\ell_M(\psi) = \ell_P(\psi) - \frac{1}{2}|j_{\lambda\lambda}(\hat{\theta}_\psi)|.$$

Questo è vero, almeno in modo approssimato, anche se $\hat{\lambda}_\psi$ dipende da ψ ma i parametri sono ortogonali. Infatti $\hat{\lambda}_\psi = \hat{\lambda} + O(N^{-1})$, e quindi, $D(\psi)$ è approssimativamente la matrice identità.

Tuttavia, la log-verosimiglianza condizionata approssimata $\ell_{CA}(\psi)$ presenta degli inconvenienti non indifferenti rispetto alla log-verosimiglianza profilo $\ell_P(\psi)$ e a alla log-verosimiglianza profilo modificata $\ell_M(\psi)$. In primis, la log-verosimiglianza condizionata approssimata richiede la specificazione di una parametrizzazione ortogonale, che esiste sicuramente solo quando ψ è scalare. Inoltre, a differenza di $\ell_P(\psi)$ e $\ell_M(\psi)$, $\ell_{CA}(\psi)$ non soddisfa il principio di invarianza rispetto a riparametrizzazioni che non alterano l'interesse.

Esistono altre proposte di aggiustamento alla funzione di verosimiglianza profilo in letteratura. Tutte queste portano ad una riduzione dello *score bias* da $O(1)$ a $O(N^{-1})$. Per ulteriori approfondimenti si rimanda a Severini (2000, Capitolo 9).

1.8.2 Modifiche di $r_P(\psi)$

Le procedure inferenziali basate sulla verosimiglianza profilo finora sono state basate su approssimazioni asintotiche del primo ordine. Tuttavia, quando la numerosità campionaria N è ridotta o quando la dimensione del parametro di disturbo λ è grande rispetto ad N potrebbe essere preferibile ricorrere ad approssimazioni di ordine superiore, che garantiscano un'accuratezza maggiore anche in scenari complessi.

Si supponga che il parametro di interesse ψ sia scalare. Se si considera il problema di verifica di ipotesi $H_0 : \psi = \psi_0$, contro una generica alternativa unidirezionale, è possibile utilizzare la statistica test radice con segno del log-rapporto di verosimiglianza profilo (1.11)

$$r_P(\psi_0) = \text{sgn}(\hat{\psi} - \psi_0)\sqrt{W_P(\psi_0)},$$

dove è noto che, sotto l'ipotesi nulla, $r_P(\psi_0)$ ha distribuzione approssimata normale standard con un errore di ordine $O(N^{-1/2})$. In particolare, per un alternativa unidirezionale sinistra, il livello di significatività osservato è

$$\alpha^{oss}(\psi_0) = \Pr_{\psi_0,\lambda}\{r_P(\psi_0) \leq r_P(\psi_0)^{oss}\} = \Phi(r_P(\psi_0)^{oss}) + O(N^{-1/2}),$$

dove $\Phi(\cdot)$ è la funzione di ripartizione di una normale standard e $r_P(\psi_0)^{oss}$ indica il valore osservato della statistica $r_P(\psi_0)$.

Il risultato principale alla base della teoria asintotica di ordine superiore è la formula p^* di Barndorff-Nielsen (1983), chiamata anche *magic formula* da Efron (1998). Questa formula consiste in un'approssimazione per la densità dello stimatore di massima verosimiglianza condizionata ad una statistica ancillare. Dato il modello $\mathcal{F}$, sia a una statistica ancillare, e dunque $(\hat{\theta}, a)$ statistica sufficiente minimale. Ne segue che, sotto campionamento casuale semplice con numerosità N

$$p_{\hat{\theta}|A=a}(\hat{\theta};\theta,a) = p^*(\hat{\theta};\theta,a)\{1 + O(N^{-3/2})\},$$

dove

$$p^*(\hat{\theta};\theta,a) = c(\theta,a)|j(\hat{\theta};\theta,a)|^{1/2} \exp\{\ell(\theta;\hat{\theta},a) - \ell(\hat{\theta};\hat{\theta},a)\},$$

con $c(\theta,a)$ costante di normalizzazione. Per ulteriori dettagli sulla qualità dell'approssimazione fornita dalla formula p^* si veda Skovgaard (1990).

Utilizzando la formula p^*, è possibile ottenere una modificazione dell'usuale statistica $r_P(\psi)$, che nel seguito verrà indicata con $r_P^*(\psi)$ (Barndorff-Nielsen, 1991), e che garantisce un'accuratezza del terzo ordine, ossia con un errore di ordine $O(N^{-3/2})$. Sia

$$\begin{aligned} r_P^*(\psi) &= r_P(\psi) + \frac{1}{r_P(\psi)} \log\left\{\frac{C(\psi)u_p(\psi)}{r_P(\psi)}\right\} \\ &= r_P(\psi) + \frac{1}{r_P(\psi)} \log\{C(\psi)\} + \frac{1}{r_P(\psi)} \log\{u_p(\psi)\}, \end{aligned} \tag{1.18}$$

dove

$$C(\psi) = \frac{|\ell_{\lambda;\hat{\lambda}}(\psi,\hat{\lambda}_\psi;\hat{\psi},\hat{\lambda},a)|}{\{|j_{\lambda\lambda}(\psi,\hat{\lambda}_\psi;\hat{\psi},\hat{\lambda},a)||j_{\lambda\lambda}(\hat{\psi},\hat{\lambda}_\psi;\hat{\psi},\hat{\lambda},a)|\}^{1/2}},$$

e

$$u_p(\psi) = j_P(\hat{\psi})^{-1/2} \frac{\partial}{\partial \hat{\psi}} \{\ell_P(\psi) - \ell_P(\hat{\psi})\}.$$

La statistica $r_P^*(\psi)$ così ottenuta ha ancora distribuzione asintotica nulla normale standard ma con errore $O(N^{-3/2})$. In particolare,

$$\alpha^{oss}(\psi_0) = \Pr_{\psi_0,\lambda}\{r_P^*(\psi_0) \leq r_P^*(\psi_0)^{oss}\} = \Phi(r_P^*(\psi_0)^{oss}) + O(N^{-3/2}),$$

Tale approssimazione risulta più accurata di quella usuale poiché l'errore va a zero più velocemente al crescere di N, o in altri termini, anche con numerosità campionarie modeste si ottiene un'accuratezza superiore.

Dalla scomposizione (1.18) si può notare che $r_P^*(\psi)$ può essere riscritta come

$$r_P^*(\psi) = r_P(\psi) + \mathrm{INF}(\psi) + \mathrm{NP}(\psi),$$

dove

$$\mathrm{INF}(\psi) = \frac{1}{r_P(\psi)} \log\{u_p(\psi)\},$$

è un fattore di correzione per la qualità dell'approssimazione normale, mentre

$$\mathrm{NP}(\psi) = \frac{1}{r_P(\psi)} \log\{C(\psi)\},$$

è un fattore di correzione per la presenza di parametri di disturbo. Il termine $\mathrm{INF}(\psi)$ comporta un aggiustamento usualmente di minore importanza, a patto che l'informazione profilo osservata $j_P(\hat{\psi})$ non sia eccessivamente piccola rispetto alla dimensione di λ. Al contrario, l'aggiustamento fornito dal termine $\mathrm{NP}(\psi)$, che discende direttamente dalla verosimiglianza profilo modificata (1.16) può risultare decisivo in presenza di un numero elevato di parametri di disturbo.

Si noti che il calcolo di $r_P^*(\psi)$ richiede il calcolo delle quantità $C(\psi)$ e $u_p(\psi)$, che in generale può risultare complesso, a causa del coinvolgimento delle derivate nello spazio campionario. Tuttavia, in alcuni casi specifici ci sono delle notevoli semplificazioni. Ad esempio, se il modello è una famiglia esponenziale e ψ è una componente del parametro canonico, allora $u_p(\psi)$ coincide con la statistica di Wald (1.12)

$$u_p(\psi) = r_{Pe}(\psi) = (\hat{\psi} - \psi)[j_P(\hat{\psi})]^{1/2},$$

mentre $C(\psi)$ si riduce alla quantità

$$C(\psi) = \left\{ \frac{j_{\lambda\lambda}(\hat{\theta})}{j_{\lambda\lambda}(\hat{\theta}_\psi)} \right\}^{1/2}.$$

Per maggiori dettagli sul calcolo di $r_P^*(\psi)$ nel caso delle famiglie esponenziali si veda anche Brazzale, Davison & Reid (2007, Capitolo 8). Nel caso generale la determinazione di $r_P^*(\psi)$ può risultare più ostica e per questo sono state proposte varie approssimazioni, che conducono a quantità note come versioni stabili di $r_P^*(\psi)$. Tra queste, Skovgaard (1996) ha suggerito di approssimare le derivate rispetto allo spazio campionario con apposite covarianze tra log-verosimiglianze e funzioni punteggio, in modo da evitare l'esplicitazione di una statistica ancillare necessaria per il calcolo delle derivate nello spazio campionario. La procedura richiede il calcolo di diversi valori attesi non banali che, tuttavia, in quanto integrali, possono essere facilmente approssimati mediante metodi di simulazione Monte Carlo. Utilizzando metodi Monte Carlo, le stime di massima verosimiglianza $\hat{\theta}$ e $\hat{\theta}_\psi$ sono calcolate una volta soltanto, ossia sono quelle ottenute sul campione originale osservato. Tale approccio differisce quindi dal *bootstrap* parametrico, descritto nel paragrafo 1.9.2, in cui ad ogni simulazione è necessario ripetere la massimizzazione della funzione di verosimiglianza e ottenere le nuove stime $\hat{\theta}$ e $\hat{\theta}_\psi$. Per un maggiore approfondimento in merito all'implementazione di queste procedure si rimanda al pacchetto `R likelihoodAsy` (Pierce & Bellio, 2017) che permette di calcolare agevolmente $r_P^*(\psi)$ e i relativi *p-value*, specificato, oltre alla log-verosimiglianza, un parametro di interesse scalare $\psi(\theta)$ e una funzione per generare dal modello.

1.9 Approssimazioni *bootstrap* della distribuzione di $r_P(\psi)$

A partire dalla fine anni '70, dopo che le tecniche *jackknife* erano state introdotte (Quenouille, 1956; Tukey, 1958), grazie all'aumento delle potenzialità di calcolo favorite dalla diffusione dei personal computer, è iniziata una rivoluzione anche nel mondo della statistica, in particolare per quanto riguarda l'utilizzo di metodi computazionalmente intensivi basati sulle simulazioni. Tale rivoluzione, "*electronic computation used for the extension of classic statistical inference*" (Efron & Hastie, 2016), ha avuto come apice l'introduzione del *bootstrap* (Efron, 1979). Questa metodologia permette di condurre verifiche di ipotesi e costruire intervalli di confidenza utilizzando simulazioni, e spesso

consentendo di evitare qualsiasi tipo di ipotesi distributiva riguardo alle variabili osservate relativamente al fenomeno oggetto di studio. Più in generale, il *bootstrap* è un metodo per valutare empiricamente gli aspetti distributivi in un ampio repertorio di procedure statistiche, offrendo spesso un elevato livello di accuratezza. Proprio a causa della sua versatilità, non è possibile riassumere qui tutte le caratteristiche del *bootstrap*, e nemmeno indicare un solo riferimento per una lettura più approfondita. Tuttavia, si consiglia di fare riferimento ai testi di Efron & Tibshirani (1993), Davison & Hinkley (1997) e Efron & Hastie (2016, Capitoli 10 e 11).

1.9.1 *Bootstrap*

Il *bootstrap* si basa essenzialmente su due concetti (Young & Smith, 2005, Capitolo 11):

- il principio del *plug-in*: consiste nella sostituzione di una distribuzione di probabilità F ignota con una sua stima $\hat{F}$ stimata a partire dal campione osservato;
- simulazioni Monte Carlo: si studiano le proprietà frequentiste delle quantità di interesse rispetto alla distribuzione di $\hat{F}$ (anziché di F), usando metodi Monte Carlo, ossia simulando campioni da $\hat{F}$. In questo modo si riescono ad evitare calcoli analitici, talvolta intrattabili.

È il concetto della sostituzione tramite il principio del *plug-in* che rappresenta la vera caratteristica del metodo *bootstrap*.

Il primo aspetto da considerare è quindi la scelta della stima $\hat{F}$. Quando sotto campionamento casuale semplice la stima $\hat{F}$ è semplicemente la funzione di ripartizione empirica di y si parla di *bootstrap* non parametrico, in quanto non viene fatta nessuna assunzione sulle distribuzioni che compongono il modello statistico $\mathcal{F}$ che descrive il fenomeno interesse. Se invece si ha qualche livello di conoscenza sul fenomeno di interesse e si è in grado di formulare un ipotesi parametrica, un'alternativa è quella di ottenere una stima $\hat{\theta}$ del parametro θ che indicizza $\mathcal{F}$, utilizzando, ad esempio, il metodo della massima verosimiglianza, e poi utilizzare $F(y;\hat{\theta})$ per generare i campioni *bootstrap*. Si parla in questo caso di *bootstrap* parametrico. Esiste, in realtà, anche una terza tipologia di *bootstrap*, quello semiparametrico, che si ha quando si formulano ipotesi parziali sul fenomeno di interesse, come può accadere, ad esempio, nell'ambito dei modelli di regressione.

Indipendentemente dal livello di conoscenza sul fenomeno di interesse e, dunque, dal tipo di *bootstrap* impiegato, l'inferenza basata sul *bootstrap* considera le proprietà frequentiste di una qualche procedura inferenziale assumendo che il modello che ha

generato i dati sia $\hat{F}$ invece che F. La giustificazione statistica di tale sostituzione è di natura asintotica, purché $\hat{F}$ sia una stima consistente di F.

Ad esempio, se $T = T(Y)$ è una statistica, sotto campionamento da F, è possibile stimare la varianza di T, $Var_F[T]$ tramite la varianza di T sotto campionamento da $\hat{F}$, ossia $Var_{\hat{F}}[T]$. Quest'ultima può essere valutata tramite metodi Monte Carlo, simulando B campioni indipendenti da $\hat{F}$.

1.9.2 *Bootstrap* parametrico senza parametri di disturbo

Dato il modello statistico parametrico $\mathcal{F}$, con parametro scalare θ, si supponga di essere interessati a fare inferenza su tutto il parametro θ, in assenza di parametri di disturbo. Se si considera il problema di verifica di ipotesi $H_0 : \theta = \theta_0$, contro una generica alternativa unidirezionale, è possibile utilizzare la statistica test log-rapporto di verosimiglianza (1.2), che, sotto l'ipotesi nulla, ha distribuzione approssimata normale standard con un errore di ordine $O(N^{-1/2})$. In particolare, per un alternativa unidirezionale sinistra, il livello di significatività osservato è

$$\alpha^{oss}(\theta_0) = \Pr_{\theta_0}\{r(\theta_0) \leq r(\theta_0)^{oss}\} = \Phi(r(\theta_0)^{oss}) + O(N^{-1/2}),$$

dove $r(\theta_0)^{oss}$ indica il valore osservato della statistica $r(\theta_0)$.

In assenza di parametri di disturbo, il *p-value* $\alpha^{oss}(\theta_0)$ può essere calcolato esattamente a meno dell'errore di simulazione Monte Carlo, senza ricorrere all'utilizzo del *bootstrap*. Infatti, poiché sotto l'ipotesi H_0 il parametro è completamente specificato, il **metodo Monte Carlo** prevede semplicemente di

(i) simulare y^{*b} da $p_Y(y; \theta_0)$ e calcolare $r(\theta_0)^b$,

(ii) ripetere il procedimento (i) B volte e calcolare

$$\hat{\alpha}^{oss}_{mc}(\theta_0) = \frac{1}{B}\sum_{i=1}^{B} \mathbb{1}(\ r(\theta_0)^b \leq r(\theta_0)^{oss}\ ).$$

dove $\mathbb{1}(\cdot)$ indica la funzione indicatrice, che assume valore 1 solo quando la condizione tra parentesi è rispettata, e 0 altrimenti. Con questo approccio è possibile ottenere un'estrema accuratezza semplicemente aumentando il numero di simulazioni B. Pertanto, se è di interesse solamente calcolare un *p-value*, il metodo Monte Carlo risulta preferibile rispetto a ricorrere al *bootstrap* in quanto il costo computazionale è ridotto. Se invece è di interesse ottenere anche degli intervalli di confidenza il *bootstrap* parametrico è computazionalmente più vantaggioso.

Indicando con $\hat{\theta}^{oss}$ la stima di θ ottenuta sul campione originale y, il ***bootstrap* parametrico** usa $\hat{\theta}^{oss}$ e prevede di

(i) simulare y^{*b} da $p_Y(y; \hat{\theta}^{oss})$ e calcolare $r(\hat{\theta}^{oss})^b$,

(ii) ripetere il procedimento (i) B volte e calcolare

$$\hat{\alpha}^{oss}(\theta_0) = \frac{1}{B}\sum_{i=1}^{B} \mathbb{1}(\ r(\hat{\theta}^{oss})^b \leq r(\theta_0)^{oss}\ ).$$

È possibile mostrare (Young & Smith, 2005, Capitolo 11) che quando $B \to \infty$, $\alpha^{oss} = \hat{\alpha}^{oss} + O(N^{-1})$. Pertanto, utilizzare $\hat{\alpha}^{oss}$ risulta più accurato rispetto ad utilizzare l'approssimazione del primo ordine $\Phi(r(\theta_0)^{oss})$.

1.9.3 *Bootstrap* parametrico con parametri di disturbo

Se si considera il problema di verifica di ipotesi $H_0 : \psi = \psi_0$, con ψ scalare, contro una generica alternativa unidirezionale, è possibile utilizzare la statistica test log-rapporto di verosimiglianza profilo (1.11) che, sotto l'ipotesi nulla, ha distribuzione approssimata normale standard, con un errore pari a $O(N^{-1/2})$. In particolare, per un'alternativa unidirezionale sinistra, il livello di significatività osservato è

$$\alpha^{oss}(\psi_0) = \sup_{\lambda\in\Lambda} \Pr_{\psi,\lambda}\{r_P(\psi_0) \leq r_P(\psi_0)^{oss}\} = \Phi(r_P(\psi_0)^{oss}) + O(N^{-1/2}),$$

dove $r_P(\psi_0)^{oss}$ indica il valore osservato della statistica $r_P(\psi_0)$.

In tale contesto, a differenza della situazione in assenza di parametri di disturbo, non è possibile simulare direttamente dall'ipotesi nulla H_0, poiché λ è ignoto. Il *bootstrap* parametrico offre due alternative: utilizzare la stima globale sul campione originale $\hat{\theta}^{oss} = (\hat{\psi}^{oss}, \hat{\lambda}^{oss})$ e in tal caso si parla di *bootstrap* non vincolato (***unconstrained bootstrap***); oppure utilizzare la stima vincolata sul campione originale $\hat{\theta}^{oss}_{\psi_0} = (\psi_0, \hat{\lambda}^{oss}_{\psi_0})$ e in tal caso si parla di *bootstrap* vincolato (***constrained bootstrap***). I due metodi funzionano nel seguente modo:

- l'***unconstrained bootstrap*** prevede di

  (i) simulare y^{*b} da $p_Y(y; \hat{\psi}^{oss}, \hat{\lambda}^{oss})$ e calcolare $r_p(\hat{\psi}^{oss})^b$,

  (ii) ripetere il procedimento (i) B volte e calcolare

$$\hat{\alpha}^{oss}_u(\psi_0) = \frac{1}{B}\sum_{i=1}^{B} \mathbb{1}(\ r_P(\hat{\psi}^{oss})^b \leq r_P(\psi_0)^{oss}\ ); \tag{1.19}$$

- il ***constrained bootstrap*** prevede di

  (i) simulare y^{*b} da $p_Y(y;\psi_0,\hat{\lambda}_{\psi_0}^{oss})$ e calcolare $r_p(\psi_0)^b$,

  (ii) ripetere il procedimento (i) B volte e calcolare

$$\hat{\alpha}_c^{oss}(\psi_0) = \frac{1}{B}\sum_{i=1}^{B} \mathbb{1}(\ r_P(\psi_0)^b \leq r_P(\psi_0)^{oss}\ ). \tag{1.20}$$

In particolare, il *constrained bootstrap* è in grado di garantire un'accuratezza superiore rispetto all'*unconstrained bootstrap*. Infatti, per $B \to \infty$, valgono le seguenti approssimazioni

$$\alpha^{oss}(\psi_0) = \hat{\alpha}_u^{oss}(\psi_0) + O(N^{-1}),$$
$$\alpha^{oss}(\psi_0) = \hat{\alpha}_c^{oss}(\psi_0) + O(N^{-3/2}).$$

Pertanto entrambi i metodi permettono di ottenere un guadagno in termini di accuratezza rispetto ad utilizzare l'approssimazione del primo ordine $\Phi(r_P(\psi_0)^{oss})$ quando $B \to \infty$, ma il *constrained bootstrap* permette di ottenere un'accuratezza asintotica del terzo ordine (Lee & Young, 2005), a differenza dell'*unconstrained bootstrap* che ha un'accuratezza asintotica del secondo ordine (DiCiccio & Romano, 1995) .

Nella pratica, in condizioni regolari, quando B è finito, le differenze numeriche tra i due approcci sono tipicamente trascurabili e l'*unconstrained bootstrap* risulta spesso accurato quanto quello *constrained* (Lee & Young, 2005; Young, 2009). Tuttavia, esistono delle situazioni in cui l'*unconstrained bootstrap* non è applicabile. Questo succede, ad esempio, quando il parametro ψ_0 è sulla frontiera dello spazio parametrico o quando la stima stessa si trova sulla frontiera. Per maggiori dettagli sulle condizioni che rendono la procedura del *bootstrap* consistente si rimanda a Bickel & Freedman (1981), o per esempi in cui il *bootstrap* non funziona correttamente a Beran (1997) e Davison et al. (2003). In generale, in presenza di parametri di disturbo, il *bootstrap* parametrico offre un'alternativa molto valida rispetto all'utilizzo delle classiche quantità pivotali basate sulla verosimiglianza. Uno degli aspetti negativi è il costo computazionale legato alle procedura basate sulle simulazioni che, tuttavia, può essere parzialmente compensato dal calcolo parallelo e dalle maggiori risorse computazionali oggi disponibili.

Capitolo 2

Modelli con effetti fissi stratificati

2.1 Introduzione

In questo capitolo vengono illustrate le principali caratteristiche dei modelli a due indici asintotici con effetti fissi stratificati. Successivamente, nel Capitolo 3, si discutono, invece, le caratteristiche dei modelli a due indici asintotici con effetti fissi incrociati.

Sia nel caso di effetti fissi stratificati che incrociati, si assume di disporre di un campione di osservazioni indipendenti y_{ij}, $i = 1, \ldots, R$, $j = 1, \ldots, C_i$, in cui in generale ogni strato potrebbe avere una diversa numerosità campionaria C_i, dove i è l'indice relativo agli strati, mentre j è l'indice relativo all'osservazione j-esima nello strato i-esimo. La numerosità campionaria complessiva è dunque $N = \sum_{i=1}^{R} C_i$. Senza perdita di generalità, nel seguito si assume che tutti gli strati abbiano la medesima numerosità campionaria C e, pertanto, che la numerosità campionaria complessiva sia semplicemente $N = RC$. In tal caso, si parla anche di schema di dati bilanciato.

Si supponga che la densità di ogni y_{ij}, $i = 1, \ldots, R$, $j = 1, \ldots, C$, sia specificata da

$$p(y_{ij}; \psi, \alpha_i),$$

dove il parametro strutturale ψ ha dimensione fissata k, mentre l'effetto fisso α_i rappresenta il parametro di disturbo specifico dello strato i, $i = 1, \ldots, R$. Il modello ha quindi parametro globale $\theta = (\psi, \alpha)$, con $\alpha = (\alpha_1, \ldots, \alpha_R)$, con dimensione $k + R$.

La densità di y_{ij} potrebbe essere condizionata ad un vettore di covariate x_{ij}, assunte note, come nel caso dei modelli lineari generalizzati. Per i modelli con effetti fissi

stratificati la funzione di log-verosimiglianza è

$$\ell(\psi, \alpha) = \sum_{i=1}^{R} \sum_{j=1}^{C} \log p(y_{ij}; \psi, \alpha_i).$$

In presenza di effetti fissi stratificati, come mostrato da Lancaster (2002), non sussistono problemi legati all'utilizzo della verosimiglianza profilo ogni qualvolta sia possibile scomporre la funzione di verosimiglianza in due componenti a parametri separabili, una relativa al parametro di interesse ψ e l'altra relativa al parametro di disturbo α. Questo tipo di situazione è quella che si verifica nello scenario descritto nell'esempio del paragrafo 2.3.2.

Più in generale, in questo tipo di contesto, in virtù dell'indipendenza tra i diversi strati, la funzione di log-verosimiglianza può essere scritta come la somma di contributi di ciascuno strato

$$\ell(\psi, \alpha) = \sum_{i=1}^{R} \ell^i(\psi, \alpha_i), \tag{2.1}$$

con

$$\ell^i(\psi, \alpha_i) = \sum_{j=1}^{C} \log p(y_{ij}; \psi, \alpha_i).$$

Dall'espressione (2.1) è immediato notare come la funzione di log-verosimiglianza, per ψ fissato, ammetta una separazione in termini dei parametri incidentali α_i, poiché nello strato i-esimo, $i = 1, \ldots, R$, è presente solamente il parametro α_i. La conseguenza di tale separazione è che è possibile ottenere il vettore delle stime vincolate $\hat{\alpha}_\psi = (\hat{\alpha}_{1\psi}, \ldots, \hat{\alpha}_{R\psi})$ tramite R equazioni di stima $\frac{\partial}{\partial \alpha_i} \ell^i(\psi, \alpha_i) = 0$. Ne segue che la funzione di log-verosimiglianza profilo per ψ risulta pari alla somma delle R log-verosimiglianze profilo di ciascuno strato, ovvero

$$\ell_P(\psi) = \ell(\psi, \hat{\alpha}_\psi) = \sum_{i=1}^{R} \ell^i(\psi, \hat{\alpha}_{i,\psi}) = \sum_{i=1}^{R} \ell_P^i(\psi).$$

In questo scenario, una possibile alternativa all'utilizzo della verosimiglianza profilo è rappresentata dall'utilizzo della verosimiglianza condizionata o marginale, che permettono di eliminare il problema dei parametri incidentali, ma la loro disponibilità è in generale garantita solo nel caso delle famiglie esponenziali e di gruppo, rispettivamente, per particolari scelte del parametro di interesse. Un'ulteriore possibilità è

quella di ricorrere alla verosimiglianza profilo modificata che consente di ripristinare, almeno approssimativamente, le identità di Bartlett, che non sono rispettate nel caso dell'usuale verosimiglianza profilo. Infine, Lunardon (2018) ha mostrato che opportune tecniche di *bias reduction* offrono una soluzione asintoticamente equivalente all'utilizzo della verosimiglianza profilo modificata.

I modelli con effetti fissi stratificati sono ampiamente adottati, in particolare in ambito economico, per l'analisi dei dati di *panel*, dove ciascun strato solitamente rappresenta una singola unità statistica (ad esempio, una nazione), con caratteristiche rilevate (ad esempio, una determinata variabile economica) in diversi momenti temporali. Per descrivere adeguatamente l'eterogeneità non osservabile tra le unità statistiche, si utilizza frequentemente l'approccio che prevede l'introduzione di effetti specifici per ciascuno strato, i quali catturano le caratteristiche invarianti nel tempo di ogni unità statistica. Questi effetti possono essere considerati come variabili casuali o parametri fissi. Nel primo caso, si parla di modelli a effetti casuali; nel secondo, di modelli a effetti fissi. In questa tesi, si considerano esclusivamente i modelli con effetti fissi.

È importante notare che la specificazione a effetti casuali richiede di assumere una distribuzione appropriata per tali effetti, sebbene essi non siano osservabili. Inoltre, una problematica rilevante di questo approccio è la necessità di ipotizzare l'assenza di correlazione tra gli effetti casuali e le variabili esplicative (Lancaster, 2000). Per l'implausibilità di questa assunzione, gli econometrici tendono spesso a non trattare gli effetti individuali come variabili casuali, preferendo i modelli a effetti fissi, che consentono la dipendenza delle caratteristiche individuali non osservabili dalle variabili esplicative. Tuttavia, anche questa scelta presenta svantaggi: adottare la specificazione a effetti fissi comporta l'introduzione del classico problema dei parametri incidentali.

In questo capitolo si discuteranno tali aspetti nel caso dei modelli a due indici con effetti fissi stratificati, soffermandosi maggiormente sulle proprietà della verosimiglianza profilo modificata e sull'utilizzo della verosimiglianza condizionata e marginale, illustrandone l'applicazione in alcuni semplici esempi.

2.2 Verosimiglianza profilo modificata in modelli con effetti fissi stratificati

L'utilizzo della verosimiglianza profilo in modelli stratificati in presenza di parametri incidentali porta a conclusioni non accurate a causa dei problemi relativi allo *score bias* ed *information bias*. Infatti, in modelli a due indici asintotici con effetti fissi stratificati,

come mostrato da Sartori (2003), è noto che

$$\mathbb{E}_{\psi,\lambda}\Big[\frac{\partial}{\partial\psi}\ell_P(\psi)\Big] = O(R),$$
$$\mathbb{E}_{\psi,\lambda}\Big[\frac{\partial^2}{\partial\psi\partial\psi^T}\ell_P(\psi)\Big] + \mathbb{E}_{\psi,\lambda}\Big[\Big\{\frac{\partial}{\partial\psi}\ell_P(\psi)\Big\}\Big\{\frac{\partial}{\partial\psi}\ell_P(\psi)\Big\}^T\Big] = O(R),$$

a differenza di quanto succede nello scenario standard, descritto nel paragrafo 1.7.4. Infatti, in questo caso, la necessità di dover stimare gli effetti fissi di strato α_i, per $i = 1, \ldots, R$, introduce una distorsione nella funzione punteggio profilo di ordine $O(R)$. Il vantaggio che si ottiene ricorrendo alla verosimiglianza profilo modificata sulla verosimiglianza profilo è la riduzione della distorsione della funzione punteggio profilo.

Infatti, quando il numero di strati R cresce con la numerosità di ciascuno strato C, in modo che $R = O(C^\upsilon)$, per $\upsilon > 0$, Sartori (2003) ha mostrato la funzione punteggio profilo $\ell_P^*(\psi)$ può essere scomposta come

$$\ell_P^*(\psi) = \ell_{\psi|\lambda} + B + Re, \tag{2.2}$$

dove

$$\ell_{\psi|\lambda} = \ell_\psi - i_{\psi\lambda}i_{\lambda\lambda}^{-1}\ell_\lambda = O(N^{1/2}) = O(C^{(\upsilon+1)/2}),$$

ha media 0 e varianza $i_{\psi\psi|\lambda}$. Il termine di distorsione B è di ordine $O(C^\upsilon)$ quando $\upsilon > 1$, mentre il termine residuale Re è di ordine $O(C^{\upsilon-1})$.

Come mostrato da Bellio et al. (2023b), quando $0 \leq \upsilon < 1$, i termini della scomposizione (2.2) sono in ordine decrescente. Invece, quando $1 \leq \upsilon < 3$, il termine dominante è il termine di distorsione B, seguito da $\ell_{\psi|\lambda}$. Infine, quando $\upsilon \geq 3$, il termine $\ell_{\psi|\lambda}$ viene dominato sia dal termine B che dal termine residuale Re.

Nel caso dei modelli con effetti fissi stratificati, dalla scomposizione (2.2) e in virtù dell'additività degli strati, segue che per lo strato i-esimo, $i = 1, \ldots, R$, vale che

$$\ell_P^i(\psi) = \ell_{\psi|\lambda}^i + B^i + Re^i,$$

e, dunque, ogni strato contribuisce alla distorsione totale della funzione punteggio associata alla log-verosimiglianza profilo con una quantità pari a

$$\mathbb{E}_{\psi,\lambda}\Big[\frac{\partial}{\partial\psi}\ell_P^i(\psi)\Big] = -\rho_\psi^i + O(C^{-1}),$$

dove $-\rho_\psi^i$, il valore atteso del termine di distorsione B^i dello strato i-esimo, è una

quantità di ordine $O(1)$. Di conseguenza, si può mostrare che la distorsione totale accumulata negli R strati è pari a $\sum_{i=1}^{R} \rho_\psi^i$, che è di ordine $O(R)$. Ne segue che la distorsione si aggrava all'aumentare del numero di strati. Quindi, seppur il problema dello *score bias*, in un contesto regolare senza parametri incidentali, non intacchi la consistenza dello stimatore per ψ basato sulla verosimiglianza profilo, quando si è in presenza di parametri incidentali, la distorsione accumulata di ordine $O(R)$ comporta tipicamente la potenziale perdita di consistenza dello stimatore per ψ, in particolare se C è limitato.

Se si considera, invece, la log-verosimiglianza profilo modificata $\ell_M(\psi) = \ell_P(\psi) + \log M(\psi)$, per la separabilità dei parametri incidentali tipicamente vale che

$$M(\psi) = \sum_{i=1}^{R} M^i(\psi),$$

ed è possibile mostrare che il valore atteso del logaritmo del fattore di modificazione $\log M(\psi)$ nello strato i-esimo è pari a

$$\mathbb{E}_{\psi,\lambda}\Big[\log M^i(\psi)\Big] = -\mathbb{E}_{\psi,\lambda}\Big[\frac{\partial}{\partial\psi}\ell_P^i(\psi)\Big] + O(C^{-1}) = \rho_\psi^i + O(C^{-1}).$$

Di conseguenza, la modificazione della funzione di log-verosimiglianza profilo in ciascuno strato permette di eliminare il termine dominante della distorsione della funzione punteggio legata alla verosimiglianza profilo dello strato, a meno di una quantità di ordine $O(C^{-1})$.

Tuttavia, lo studio delle proprietà asintotiche della verosimiglianza profilo e della verosimiglianza profilo modificata nel caso di modello stratificati è in generale complesso, poiché bisogna tenere conto del comportamento delle varie quantità sia rispetto al numero di strati R che rispetto alla dimensione campionaria di ciascuno strato C. Nel seguito si consideri, per semplicità, la situazione in cui sia R che C possano tendere all'infinito. La condizione sufficiente affinché la verosimiglianza profilo presenti le usuali proprietà asintotiche è che C cresca più rapidamente di R, ossia in notazione $R = o(C)$. Sartori (2003) ha mostrato che la verosimiglianza profilo modificata riesce a ridurre la distorsione della funzione punteggio profilo consentendo quindi di ottenere proprietà asintotiche migliori, purché $R = o(C^3)$. Quindi, la condizione riguardante la verosimiglianza profilo modificata è più debole della condizione riguardante la verosimiglianza profilo.

In aggiunta, come nel caso della verosimiglianza profilo, è possibile definire analoghe quantità pivotali basate sulla verosimiglianza profilo modificata, ossia, $W_M(\psi)$, $W_{Me}(\psi)$

e $W_{Mu}(\psi)$

$$
\begin{aligned}
W_M(\psi) &= 2\{\ell_M(\hat{\psi}) - \ell_M(\psi)\},\\
W_{Me}(\psi) &= (\hat{\psi}_M - \psi)^T [j_M(\psi)](\hat{\psi}_M - \psi),\\
W_{Mu}(\psi) &= \ell_{M*}(\psi)^T [j_M(\psi)]^{-1} \ell_{M*}(\psi),
\end{aligned}
$$

e le rispettive versioni unilaterali $r_M(\psi)$, $r_{Me}(\psi)$ e $r_{Mu}(\psi)$

$$
\begin{aligned}
r_M(\psi) &= \mathrm{sgn}(\hat{\psi}_M - \psi)\sqrt{W_M(\psi)},\\
r_{Me}(\psi) &= (\hat{\psi}_M - \psi)[j_M(\psi)]^{1/2},\\
r_{Mu}(\psi) &= \ell_{M^*}(\psi)[j_M(\psi)]^{-1/2},
\end{aligned}
$$

dove $\ell_{M^*}(\psi)$ rappresenta la funzione punteggio relativa a $\ell_M(\psi)$ e $j_M(\psi)$ è la matrice di informazione osservata relativa a $\ell_M(\psi)$. In un contesto asintotico standard, è stato dimostrato che le quantità pivotali basate sulla verosimiglianza profilo modificata sono asintoticamente equivalenti tra di loro e, inoltre, sono asintoticamente equivalenti alle quantità pivotali basate sull'usuale verosimiglianza profilo, rispettivamente $W_P(\psi)$, $W_{Pe}(\psi)$, $W_{Pu}(\psi)$ e $r_P(\psi)$, $r_{Pe}(\psi)$, $r_{Pu}(\psi)$. Dunque, in condizioni regolari, anche le quantità pivotali basate sulla verosimiglianza profilo modificata, ad esempio le versioni unilaterali, sono ancora asintoticamente normali al primo ordine. Tuttavia, se si considera il paradigma a due indici asintotici in cui $R = o(C^3)$, ma non vale che $R = o(C)$, allora le versioni unilaterali delle quantità pivotali basate sulla verosimiglianza profilo modificata, ad esempio $r_M(\psi)$, sono ancora asintoticamente normali, mentre ciò non è vero per le versioni unilaterali delle quantità pivotali basate sulla verosimiglianza profilo, ad esempio $r_P(\psi)$.

In generale, inoltre, a prescindere dal comportamento di R ed C, Sartori (2003) ha mostrato che la distorsione dello stimatore basato sulla verosimiglianza profilo $\hat{\psi}$ è di ordine $O(C^{-1})$, mentre la distorsione dello stimatore basato sulla verosimiglianza profilo modificata $\hat{\psi}_M$ è inferiore, in quanto di ordine $O(C^{-2})$.

Dunque, quando la numerosità di ciascuno strato C non cresce più velocemente del numero degli strati R in modo che $R = o(C^3)$, si osserva un progressivo deterioramento delle procedure inferenziali, in particolare quando basate sulla verosimiglianza profilo, ma, almeno in situazioni estreme, anche quando basate sulla verosimiglianza profilo modificata.

2.3 Esempi di modelli con effetti fissi stratificati

Nel seguito si illustrano alcuni semplici esempi di modelli con effetti fissi stratificati, a partire dal modello di regressione normale con effetti fissi stratificati (si veda, ad esempio, Bartolucci et al., 2016), per proseguire con due modelli per dati discreti: il modello log-lineare Poisson con effetti fissi stratificati (si veda, ad esempio, Severini, 2000, Esempio 9.3) e il modello di Rasch ad un indice (Rasch, 1960). Per questi modelli, si discutono i problemi circa le usuali procedure inferenziali che possono sorgere in presenza di un numero elevato di parametri di disturbo, e come tali problemi possano essere in parte risolti utilizzando alcuni dei metodi introdotti in precedenza.

2.3.1 Modello normale con un effetti fissi stratificati

Un modello molto semplice è il modello di regressione normale con effetti fissi stratificati. Tale modello assume che y_{ij} siano realizzazioni di variabili casuali indipendenti Y_{ij} con distribuzione normale di media α_i e varianza ψ. La densità di Y_{ij} per questo modello è

$$p(y_{ij};\psi,\alpha)=\frac{1}{\sqrt{2\pi\psi}}\exp\Big\{-\frac{1}{2\psi}(y_{ij}-\alpha_i)^2\Big\},$$

per $i=1,\ldots,R$, $j=1,\ldots,C$. Il modello può essere scritto nella forma equivalente $Y_{ij}=\alpha_i+\epsilon_{ij}$, dove $\epsilon_{ij}\sim N(0,\psi)$ i.i.d., $i=1,\ldots,R$, $j=1,\ldots,C$. In questo modello il parametro di interesse ψ è scalare, mentre il parametro incidentale $\alpha=(\alpha_1,\ldots,\alpha_R)$ ha dimensione R.

La funzione di verosimiglianza per (ψ,α) basata sui dati y_{ij}, $i=1,\ldots,R$, $j=1,\ldots,C$, è

$$\begin{aligned}L(\psi,\alpha)&=\prod_{i=1}^{R}\prod_{j=1}^{C}p(y_{ij};\psi,\alpha_i)\\&=(\psi)^{-RC/2}\exp\Big\{-\frac{1}{2}\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(y_{ij}-\alpha_i)^2}{\psi}\Big\}.\end{aligned}$$

Ne segue che la funzione di log-verosimiglianza può essere scritta nel modo seguente

$$\ell(\psi,\alpha)=-\frac{RC}{2}\log\psi-\frac{1}{2}\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(y_{ij}-\alpha_i)^2}{\psi},\tag{2.3}$$

da cui è immediato verificare che la stima di massima verosimiglianza per α_i con ψ fissato non dipende da ψ, ed è pari alla media campionaria dell'i-esimo strato, ossia

$\hat{\alpha}_{i\psi} = \hat{\alpha}_i = \bar{y}_i$. La funzione di verosimiglianza profilo per ψ risulta quindi pari a

$$\ell_P(\psi) = -\frac{RC}{2}\log\psi - \frac{1}{2}\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(y_{ij}-\bar{y}_i)^2}{\psi},$$

da cui, derivando rispetto a ψ, si ottiene la funzione punteggio profilo per ψ

$$\frac{\partial}{\partial\psi}\ell_P(\psi) = -\frac{RC}{2\psi} + \frac{1}{2}\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(y_{ij}-\bar{y}_i)^2}{\psi^2}.$$

Calcolando $\mathbb{E}_\psi\left[\frac{\partial}{\partial\psi}\ell_P(\psi)\right]$ è immediato verificare che la funzione punteggio profilo è distorta, con distorsione pari a $-R/(2\psi)$, che aumenta all'aumentare del numero di strati. Ossia, al divergere di R, la distorsione della funzione punteggio parziale diventa di ordine $O(R)$. Dunque, lo stimatore di massima verosimiglianza per ψ che si ottiene risolvendo l'equazione punteggio parziale $\frac{\partial}{\partial\psi}\ell_P(\psi) = 0$

$$\hat{\psi} = \frac{1}{RC}\sum_{i=1}^{R}\sum_{j=1}^{C}(y_{ij}-\bar{y}_i)^2,$$

non è consistente per C fissato. Infatti

$$\hat{\psi} \xrightarrow{p} \frac{C-1}{C}\psi,$$

quando il numero di strati $R \to \infty$, e la numerosità C di ciascuno strato rimane fissata.

Tuttavia, poiché il modello in esame è una famiglia esponenziale piena, la stima di massima verosimiglianza è statistica sufficiente minimale. Questo permette di riscrivere la funzione di log-verosimiglianza (2.3) in funzione della stima di massima verosimiglianza. Ossia

$$\ell(\psi,\alpha;\hat{\psi},\hat{\alpha}) = -\frac{RC}{2}\log\psi - \frac{1}{2}\frac{RC\hat{\psi} + R\sum_{i=1}^{R}(\hat{\alpha}_i-\alpha_i)^2}{\psi},$$

dove, poiché

$$\sum_{i=1}^{R}\sum_{j=1}^{C}(y_{ij}-\alpha_i)^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}(y_{ij}-\hat{\alpha}_i)^2 + C\sum_{i=1}^{R}(\hat{\alpha}_i-\alpha_i)^2,$$

segue che la funzione punteggio parziale per α_i risulta pari a

$$\ell_{\alpha_i}(\psi, \alpha_i; \hat{\psi}, \hat{\alpha}_i) = \frac{\partial}{\partial \alpha_i} \ell(\psi, \alpha_i; \hat{\psi}, \hat{\alpha}_i) = \frac{C}{\psi}(\hat{\alpha}_i - \alpha_i).$$

Di conseguenza è semplice calcolare la derivata nello spazio campionario

$$\ell_{\alpha_i;\hat{\alpha}_i}(\psi, \alpha_i; \hat{\psi}, \hat{\alpha}_i) = \frac{\partial}{\partial \hat{\alpha}_i} \ell_{\alpha_i}(\psi, \alpha_i; \hat{\psi}, \hat{\alpha}_i) = \frac{C}{\psi},$$

che coincide con il blocco (α_i, α_i) dell'informazione osservata per (ψ, α_i), ossia

$$j_{\alpha_i,\alpha_i}(\psi, \alpha_i) = \frac{C}{\psi}.$$

Tale semplificazione rende immediato calcolare il fattore di modificazione $M(\psi)$ per la verosimiglianza profilo modificata. Infatti,

$$\log M(\psi) = -\frac{R}{2} \log \psi, \tag{2.4}$$

da cui

$$\begin{aligned} \ell_M(\psi) &= \ell_P(\psi) - \log M(\psi) \\ &= -R\frac{(C-1)}{2} \log \psi - \frac{1}{2} \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(y_{ij} - \bar{y}_i)^2}{\psi}. \end{aligned}$$

La funzione punteggio modificata risulta quindi uguale a

$$\frac{\partial}{\partial \psi} \ell_M(\psi) = -R\frac{(C-1)}{2\psi} - \frac{1}{2} \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(y_{ij} - \bar{y}_i)^2}{\psi^2},$$

che risulta essere esattamente non distorta. Infine, lo stimatore per ψ basato sulla verosimiglianza profilo modificata risulta pari a

$$\hat{\psi}_M = \frac{1}{R(C-1)} \sum_{i=1}^{R} \sum_{j=1}^{C} (y_{ij} - \bar{y}_i)^2,$$

che è non distorto e consistente quando il numero di strati $R \to \infty$, anche se e la numerosità C di ciascuno strato rimane fissata, a differenza dello stimatore basato sull'usuale verosimiglianza profilo. Inoltre, si può facilmente vedere che, in questo caso, la verosimiglianza profilo modificata $\ell_M(\psi)$ coincide sia con la verosimiglianza marginale sia con la verosimiglianza condizionata. Infatti, la funzione di log-verosimiglianza condizionata

può essere ottenuta condizionandosi alla statistica sufficiente per α data dal vettore R-dimensionale con generico elemento $y_{i+} = \sum_{j=1}^{C} y_{ij}$, $i = 1, \ldots, R$. Se $\ell^i(\psi, \alpha_i)$ denota il contributo i-esimo alla funzione di log-verosimiglianza, $i = 1, \ldots, R$, allora la funzione di log-verosimiglianza condizionata si può esprimere come somma rispetto ad i della seguente quantità

$$\ell^i(\psi, \alpha_i) - \left\{ -\frac{1}{2}\log\psi - \frac{1}{2}\frac{(y_{i+} - C\alpha_i)^2}{C\psi} \right\} = -\frac{(C-1)}{2}\log\psi - \frac{1}{2}\sum_{j=1}^{C}\frac{(y_{ij} - \bar{y}_i)^2}{\psi},$$

che coincide con la verosimiglianza profilo modificata $\ell_M(\psi)$. Infine, la funzione di log-verosimiglianza condizionata coincide con la funzione di log-verosimiglianza marginale basata sulla distribuzione della statistica $\sum_{i=1}^{R}\sum_{j=1}^{C}(y_{ij} - \bar{y}_i)^2$.

2.3.2 Modello log-lineare Poisson con effetti fissi stratificati

In diverse applicazioni i valori della variabile risposta y_{ij} possono rappresentare il risultato di un conteggio. Un modello statistico di base è quello che assume che le y_{ij} siano realizzazioni di variabili casuali indipendenti Y_{ij} con distribuzione Poisson di media μ_{ij}. La densità di Y_{ij} per questo modello è

$$p(y_{ij}; \psi, \alpha_i) = \frac{e^{-\mu_{ij}}(\mu_{ij})^{y_{ij}}}{y_{ij}!},$$

dove, per un'opportuna funzione di legame $g(\cdot)$, $g(\mu_{ij}) = \eta_{ij}$, con $\eta_{ij} = \alpha_i + \psi x_{ij}$, per $i = 1, \ldots, R$, $j = 1, \ldots, C$. Il vincolo di identificabilità può essere rispettato ponendo a zero uno degli α_i. Senza perdita di generalità si è assunto di avere a disposizione un'unica covariata x e che il parametro di interesse ψ sia scalare. Assumendo la funzione di legame canonica $g(\mu_{ij}) = \log\mu_{ij} = \eta_{ij}$, segue che

$$Y_{ij} \sim Po(e^{\alpha_i + \psi x_{ij}}),$$

con valore atteso $\mu_{ij} = e^{\eta_{ij}}$, $i = 1, \ldots, R$, $j = 1, \ldots, C$.

La funzione di verosimiglianza per (ψ, α) basata sui dati y_{ij}, $i = 1, \ldots, R$, $j =$

$1, \ldots, C$, è

$$\begin{aligned} L(\psi, \alpha) &= \prod_{i=1}^{R} \prod_{j=1}^{C} p(y_{ij}; \psi, \alpha_i) \\ &= \exp\Big\{ - \sum_{i=1}^{R} \sum_{j=1}^{C} e^{\alpha_i + \psi x_{ij}} + \sum_{i=1}^{R} \sum_{j=1}^{C} (\alpha_i + \psi x_{ij}) y_{ij} \Big\}. \end{aligned}$$

Ne segue che la funzione di log-verosimiglianza può essere scritta come

$$\begin{aligned} \ell(\psi, \alpha) &= - \sum_{i=1}^{R} \sum_{j=1}^{C} e^{\alpha_i + \psi x_{ij}} + \sum_{i=1}^{R} \sum_{j=1}^{C} (\alpha_i + \psi x_{ij}) y_{ij} \\ &= - \sum_{i=1}^{R} e^{\alpha_i} \sum_{j=1}^{C} e^{\psi x_{ij}} + \sum_{i=1}^{R} \alpha_i \sum_{j=1}^{C} y_{ij} + \psi \sum_{i=1}^{R} \sum_{j=1}^{C} x_{ij} y_{ij} \\ &= - \sum_{i=1}^{R} e^{\alpha_i} \sum_{j=1}^{C} e^{\psi x_{ij}} + \sum_{i=1}^{R} \alpha_i y_{i+} + \psi s, \end{aligned}$$

dove $y_{i+} = \sum_{j=1}^{C} y_{ij}$ è il totale di riga i-esimo e $s = \sum_{i=1}^{R} \sum_{j=1}^{C} x_{ij} y_{ij} = \sum_{i=1}^{R} s_i$. Il contributo i-esimo alla funzione di log-verosimiglianza è quindi

$$\ell^i(\psi, \alpha_i) = -e^{\alpha_i} \sum_{j=1}^{C} e^{\psi x_{ij}} + \alpha_i y_{i+} + \psi s_i,$$

e, derivando $\ell^i(\psi, \alpha_i)$ rispetto ad α_i, si ottiene

$$\ell_{\alpha_i}(\psi, \alpha_i) = -\frac{\partial}{\partial \alpha_i} \ell^i(\psi, \alpha_i) = e^{\alpha_i} \sum_{j=1}^{C} e^{\psi x_{ij}} + y_{i+}.$$

Risolvendo $\ell_{\alpha_i}(\psi, \alpha_i) = 0$ rispetto ad α_i, si ricava la stima di massima verosimiglianza di α_i per ψ fissato, $\hat{\alpha}_{i\psi}$

$$\hat{\alpha}_{i\psi} = \log \Big\{ \frac{y_{i+}}{\sum_{j=1}^{C} e^{\psi x_{ij}}} \Big\}.$$

Sostituendo le stime vincolate $\hat{\alpha}_{i\psi}$, $i = 1, \ldots, R$, al posto dei corrispondenti α_i è possibile

ottenere la funzione di log-verosimiglianza profilo per ψ

$$\begin{aligned}\ell_P(\psi) &= \ell(\psi, \hat{\alpha}_{1\psi}, \ldots, \hat{\alpha}_{R\psi}) \\ &= -\sum_{i=1}^{R} y_{i+} + \sum_{i=1}^{R}\Big\{ \log y_{i+} - \log \sum_{j=1}^{C} e^{\psi x_{ij}} \Big\} y_{i+} + \psi s \\ &= c(y) + \psi s - \sum_{i=1}^{R} y_{i+} \log \sum_{j=1}^{C} e^{\psi x_{ij}}. \end{aligned} \tag{2.5}$$

La statistica R-dimensionale con elementi $y_{i+} = \sum_{j=1}^{C} y_{ij}$, $i = 1, \ldots, R$, è parzialmente sufficiente per α ed ha distribuzione marginale nota. Pertanto, è immediato ottenere la verosimiglianza condizionata per ψ. Infatti, per l'assunzione di indipendenza

$$y_{i+} = \sum_{j=1}^{C} y_{ij} \sim P(e^{\alpha_i} \sum_{j=1}^{C} e^{\psi x_{ij}}),$$

e la corrispondente log-verosimiglianza basata su $y_{1+}, \ldots, y_{R+}$ risulta pari a

$$\ell(\psi, \lambda; y_{1+}, \ldots, y_{R+}) = -\sum_{i=1}^{R} e^{\alpha_i} \sum_{j=1}^{C} e^{\psi x_{ij}} + \sum_{i=1}^{R} \alpha_i y_{i+} + \sum_{i=1}^{R} y_{i+} \log \sum_{j=1}^{C} e^{\psi x_{ij}}.$$

Di conseguenza, la funzione di log-verosimiglianza condizionata per ψ può essere ottenuta come differenza tra la log-verosimiglianza basata sui dati originali y_{ij}, $i = 1, \ldots, R$, $j = 1, \ldots, C$, e la log-verosimiglianza basata su $y_{1+}, \ldots, y_{R+}$

$$\begin{aligned}\ell_C(\psi) &= \ell(\psi, \alpha) - \ell(\psi, \lambda; y_{1+}, \ldots, y_{R+}) \\ &= c(y) + \psi s - \sum_{i=1}^{R} y_{i+} \log \sum_{j=1}^{C} e^{\psi x_{ij}}. \end{aligned} \tag{2.6}$$

Dall'ultima espressione, si può osservare che per questo modello la funzione di log-verosimiglianza condizionata (2.6) per ψ coincide esattamente con la funzione di log-verosimiglianza profilo (2.5) per ψ.

In aggiunta, nella parametrizzazione mista $(\psi, \mu_{1+}, \ldots, \mu_{R+})$, con

$$\mu_{i+} = \sum_{j=1}^{C} \mu_{ij} = e^{\alpha_i} \sum_{j=1}^{C} e^{\psi x_{ij}},$$

si ottiene

$$\ell(\psi, \mu_{1+}, \ldots, \mu_{R+}) = \sum_{i=1}^{R}(y_{i+} \log \mu_{i+} - \mu_{i+}) + \psi s - \sum_{i=1}^{R} y_{i+} \log \sum_{j=1}^{C} e^{\psi x_{ij}}, \qquad (2.7)$$

con parametri separabili. Nello specifico, il secondo e terzo addendo nel membro di destra della (2.7)

$$\psi s - \sum_{i=1}^{R} y_{i+} \log \sum_{j=1}^{C} e^{\psi x_{ij}},$$

formano la log-verosimiglianza condizionata per ψ, $\ell_C(\psi)$, trovata nella (2.6).

Da un punto di vista teorico, l'equivalenza tra verosimiglianza profilo e condizionata giustifica il fatto che che i risultati ottenuti a partire da procedure inferenziali basate sulla log-verosimiglianza profilo siano accurati anche in caso di presenza di parametri incidentali nel modello. Pertanto, per l'inferenza su ψ è equivalente utilizzare le classiche quantità pivotali basate sulla verosimiglianza, come la radice con segno del log-rapporto di verosimiglianza profilo, o le loro versioni modificate, come la radice con segno del log-rapporto di verosimiglianza profilo modificata. Infatti, le modifiche alla verosimiglianza profilo proposte da Barndorff-Nielsen (1983) e Cox & Reid (1987) lasciano l'usuale verosimiglianza profilo inalterata. Si veda anche Severini (2000, Esempio 9.3).

Calcolando, infatti, la derivata parziale seconda della funzione di log-verosimiglianza rispetto ad α_i si ottiene che

$$\ell_{\alpha_i\alpha_i} = \frac{\partial}{\partial \alpha_i^2} \ell^i(\psi, \alpha_i) = -e^{\alpha_i} \sum_{j=1}^{C} e^{\psi x_{ij}},$$

pertanto il corrispondente elemento dell'informazione osservata è

$$j_{\alpha_i\alpha_i} = j_{\alpha_i\alpha_i}(\psi, \alpha_i) = e^{\alpha_i} \sum_{j=1}^{C} e^{\psi x_{ij}},$$

che, valutato nella stima vincolata $\hat{\theta}_\psi = (\psi, \hat{\alpha}_{i\psi})$, risulta uguale a

$$j_{\alpha_i\alpha_i}(\psi, \hat{\alpha}_{i\psi}) = e^{\hat{\alpha}_{i\psi}} \sum_{j=1}^{C} e^{\psi x_{ij}} = y_{i+}.$$

Poiché $j_{\alpha_i\alpha_i}(\psi, \hat{\alpha}_{i\psi})$ risulta indipendente da ψ, il fattore di correzione nella log-verosimiglianza

profilo modificata

$$\ell_M(\psi) = \ell_P(\psi) + \frac{1}{2}\log|j_{\alpha\alpha}(\hat{\theta}_\psi)|.$$

può essere trascurato, e di conseguenza, la log-verosimiglianza profilo modificata coincide con la log-verosimiglianza profilo.

2.3.3 Modello logistico di Rasch ad un indice

Un modello statistico molto utilizzato per dati binari, in particolare in diverse applicazioni nell'ambito della *item response analysis*, è il modello logistico di Rasch (Rasch, 1960). Si assuma che un gruppo di R soggetti venga sottoposto ad un totale di C test. Sia y_{ij} l'esito del test i-esimo per il j-esimo soggetto. Si supponga, ad esempio, che $y_{ij} = 1$ se la risposta al test è corretta, e $y_{ij} = 0$ se la risposta al test è sbagliata. Sia infine π_{ij} la probabilità che il soggetto i-esimo risponda correttamente al test j-esimo. Il modello base di Rasch è quello che assume che y_{ij} siano realizzazioni di variabili casuali indipendenti Y_{ij} con distribuzione bernoulliana con probabilità di successo π_{ij}. La densità di Y_{ij} per questo modello è

$$p(y_{ij}; \psi_j, \alpha_i) = {\pi_{ij}}^{y_{ij}}(1-\pi_{ij})^{1-y_{ij}},$$

dove $g(\pi_{ij}) = \eta_{ij}$, con $\eta_{ij} = \alpha_i - \psi_j$, per $i = 1, \ldots, R$, $j = 1, \ldots, C$. Il vincolo di identificabilità può essere rispettato, ad esempio, ponendo a zero uno degli α_i, o uno dei γ_j, o imponendo comunque un vincolo lineare sugli α_i o sui γ_j. In questo caso il parametro di interesse ψ ha dimensione pari a C, $\psi = (\psi_1, \ldots, \psi_C)$, ed ogni ψ_i può essere interpretato come la difficoltà del test j-esimo. Invece, il parametro incidentale α ha dimensione pari ad R, $\alpha = (\alpha_1, \ldots, \alpha_R)$, ed ogni α_i può essere interpretato come l'abilità del soggetto i-esimo. Assumendo la funzione di legame canonica $g(\pi_{ij}) = \text{logit}(\pi_{ij}) = \eta_{ij}$, con $\pi_{ij} = \mu_{ij}$, segue che

$$Y_{ij} \sim Bi\Big(1, \frac{e^{\alpha_i - \psi_j}}{1 + e^{\alpha_i - \psi_j}}\Big),$$

$i = 1, \ldots, R$, $j = 1, \ldots, C$. Si noti che all'aumentare di α_i aumenta $\text{logit}(\mu_{ij})$ e quindi la probabilità di successo, mentre all'aumentare di ψ_j diminuisce $\text{logit}(\mu_{ij})$, da cui segue l'interpretazione possibile di α_i e ψ_j rispettivamente come abilità del soggetto i-esimo e difficoltà del test j-esimo. Andersen (1980, Capitolo 6) ha dimostrato che lo stimatore di massima verosimiglianza per ψ non è consistente, se $R \to \infty$ e C è fissato.

Si consideri, senza perdita di generalità, il caso con $C = 2$, ossia solamente due test. Come vincolo di identificabilità si può usare $\sum_{j=1}^{C} \psi_j = 0$, da cui segue che $\psi_2 = -\psi_1$, quindi il parametro di interesse è ψ_1, scalare. La funzione di verosimiglianza per (ψ, α) basata sui dati y_{ij}, $i = 1, \ldots, R$, $j = 1, \ldots, C$, è

$$\begin{aligned} L(\psi, \alpha) &= \prod_{i=1}^{R} \prod_{j=1}^{2} p(y_{ij}; \psi_j, \alpha_i) \\ &= \prod_{i=1}^{R} \prod_{j=1}^{2} {\pi_{ij}}^{y_{ij}} (1 - \pi_{ij})^{1-y_{ij}} \\ &= \frac{\exp\left\{\sum_{i=1}^{R} \sum_{j=1}^{2} (\alpha_i - \psi_j) y_{ij}\right\}}{\prod_{i=1}^{R} \prod_{j=1}^{2} (1 + e^{\alpha_i - \psi_j})} \\ &= \frac{\exp\left\{\sum_{i=1}^{R} \sum_{j=1}^{2} \alpha_i y_{ij} - \sum_{i=1}^{R} \sum_{j=1}^{2} \psi_j y_{ij}\right\}}{\prod_{i=1}^{R} \prod_{j=1}^{2} (1 + e^{\alpha_i - \psi_j})}. \end{aligned}$$

Ne segue che la funzione di log-verosimiglianza può essere scritta nel modo seguente

$$\begin{aligned} \ell(\psi, \alpha) &= \sum_{i=1}^{R} \sum_{j=1}^{2} \alpha_i y_{ij} - \sum_{i=1}^{R} \sum_{j=1}^{2} \psi_j y_{ij} - \sum_{i=1}^{R} \sum_{j=1}^{2} \log\left(1 + e^{\alpha_i - \psi_j}\right) \\ &= \sum_{i=1}^{R} \alpha_i \sum_{j=1}^{R} y_{ij} - \sum_{j=1}^{2} \psi_j \sum_{i=1}^{R} y_{ij} - \sum_{i=1}^{R} \sum_{j=1}^{2} \log\left(1 + e^{\alpha_i - \psi_j}\right) \\ &= \sum_{i=1}^{R} \alpha_i y_{i+} - \sum_{j=1}^{2} \psi_j y_{+j} - \sum_{i=1}^{R} \sum_{j=1}^{2} \log\left(1 + e^{\alpha_i - \psi_j}\right) \\ &= \sum_{i=1}^{R} \alpha_i y_{i+} - \psi_1 y_{+1} - \psi_2 y_{+2} - \sum_{i=1}^{R} \log\left(1 + e^{\alpha_i - \psi_1}\right) - \sum_{i=1}^{R} \log\left(1 + e^{\alpha_i - \psi_2}\right), \end{aligned} \tag{2.8}$$

dove $y_{i+} = \sum_{j=1}^{2} y_{ij} = y_{i1} + y_{i2} \in \{0, 1, 2\}$, in quanto $y_{ij} \in \{0, 1\}$, $j = 1, 2$. Invece, y_{+1} rappresenta il numero totale di risposte corrette al primo test e y_{+2} rappresenta il numero totale di risposte corrette al secondo test.

Calcolando la derivata parziale di $\ell(\psi, \alpha)$ rispetto a ciascun α_i è possibile ottenere le stime vincolate per i parametri α_i, $\hat{\alpha}_{i\psi}$. Infatti,

$$\ell_{\alpha_i}(\psi, \alpha_i) = \frac{\partial}{\partial \alpha_i} \ell(\psi, \alpha_i) = y_{i+} - \frac{e^{\alpha_i - \psi_1}}{1 + e^{\alpha_i - \psi_1}} - \frac{e^{\alpha_i - \psi_2}}{1 + e^{\alpha_i - \psi_2}},$$

da cui si ottiene l'equazione di verosimiglianza per ψ fissato

$$\frac{e^{\alpha_i-\psi_1}}{1+e^{\alpha_i-\psi_1}} - \frac{e^{\alpha_i-\psi_2}}{1+e^{\alpha_i-\psi_2}} = y_{i+},$$

che usando il vincolo $\psi_2 = -\psi_1$ può essere riscritta come

$$\frac{e^{\alpha_i-\psi_1}}{1+e^{\alpha_i-\psi_1}} - \frac{e^{\alpha_i+\psi_1}}{1+e^{\alpha_i+\psi_1}} = y_{i+}.$$

La soluzione in α_i di tale equazione è immediata se si considera che $y_{i+} \in \{0,1,2\}$. Infatti,

$$\hat{\alpha}_{i\psi} = \begin{cases} -\infty & \text{se } \; y_{i+}=0 \\ 0 & \text{se } \; y_{i+}=1 \\ +\infty & \text{se } \; y_{i+}=2 \end{cases}.$$

La derivata parziale rispetto a ψ_1 porta invece all'equazione

$$\ell_{\psi_1}(\psi_1,\alpha) = y_{+1} + \sum_{i=1}^{R} \frac{e^{\alpha_i-\psi_1}}{1+e^{\alpha_i-\psi_1}} = 0 \Rightarrow y_{+1} = \sum_{i=1}^{R} \frac{e^{\alpha_i-\psi_1}}{1+e^{\alpha_i-\psi_1}}. \tag{2.9}$$

Sostituendo al posto degli α_i le corrispettive stime vincolate $\hat{\alpha}_{i\psi}$ nell'equazione (2.9), si ottiene la seguente semplificazione

$$\begin{aligned} y_{+1} &= \sum_{i=1}^{R} \frac{e^{-\infty}}{1+e^{-\infty}}\mathbb{1}(y_{i+}=0) + \sum_{i=1}^{R} \frac{e^{-\psi_1}}{1+e^{-\psi_1}}\mathbb{1}(y_{i+}=1) + \sum_{i=1}^{R} \frac{e^{+\infty}}{1+e^{+\infty}}\mathbb{1}(y_{i+}=2) \\ y_{+1} &= 0\sum_{i=1}^{R}\mathbb{1}(y_{i+}=0) + \frac{e^{-\psi_1}}{1+e^{-\psi_1}}\sum_{i=1}^{n}\mathbb{1}(y_{i+}=1) + 1\sum_{i=1}^{R}\mathbb{1}(y_{i+}=2) \\ y_{+1} &= R_0 + \frac{e^{-\psi_1}}{1+e^{-\psi_1}}R_1 + R_2 \\ y_{+1} &= \frac{e^{-\psi_1}}{1+e^{-\psi_1}}R_1 + R_2, \end{aligned} \tag{2.10}$$

dove si è indicato con $R_0 = \sum_{i=1}^{R}\mathbb{1}(y_{i+}=0)$, con $R_1 = \sum_{i=1}^{R}\mathbb{1}(y_{i+}=1)$ e con $R_2 = \sum_{i=1}^{R}\mathbb{1}(y_{i+}=2)$. Dall'espressione (2.10) è immediato ottenere che

$$\hat{\psi}_1 = \log\Big\{\frac{R_1 - y_{+1} + R_2}{y_{+1} - R_2}\Big\}, \tag{2.11}$$

dove $R_1 - y_{+1} + R_2$ rappresenta il numero di coppie $(0,1)$, ossia il numero di soggetti che hanno risposto erroneamente al primo test ma correttamente al secondo test, mentre

$y_{+1} - R_2$ è il numero soggetti che hanno risposto correttamente al primo test ma erroneamente al secondo test. La probabilità che il soggetto i-esimo risponda erroneamente al primo test ma correttamente al secondo test è

$$\begin{aligned}\Pr_{\psi_1,\alpha_i}(\{Y_{i1}=0, Y_{i2}=1\}) &= \Big(1 - \frac{e^{\alpha_i-\psi_1}}{1+e^{\alpha_i-\psi_1}}\Big)\Big(\frac{e^{\alpha_i+\psi_1}}{1+e^{\alpha_i+\psi_1}}\Big) \\ &= \frac{e^{\alpha_i+\psi_1}}{(1+e^{\alpha_i-\psi_1})(1+e^{\alpha_i+\psi_1})} = L_{1i},\end{aligned}$$

mentre la probabilità che il soggetto i-esimo risponda correttamente al primo test ma erroneamente al secondo test è

$$\begin{aligned}\Pr_{\psi_1,\alpha_i}(\{Y_{i1}=1, Y_{i2}=0\}) &= \Big(\frac{e^{\alpha_i-\psi_1}}{1+e^{\alpha_i-\psi_1}}\Big)\Big(1 - \frac{e^{\alpha_i+\psi_1}}{1+e^{\alpha_i+\psi_1}}\Big) \\ &= \frac{e^{\alpha_i-\psi_1}}{(1+e^{\alpha_i-\psi_1})(1+e^{\alpha_i+\psi_1})} = L_{2i}.\end{aligned}$$

Ne segue che, per la legge dei grandi numeri, quando R diverge

$$\frac{R_1 - y_{+1} + R_2}{R} \xrightarrow{p} \lim_{R\to\infty} \frac{1}{R}\sum_{i=1}^{R} L_{1i},$$

e

$$\frac{y_{+1} - R_2}{R} \xrightarrow{p} \lim_{R\to\infty} \frac{1}{R}\sum_{i=1}^{R} L_{2i},$$

pertanto

$$\frac{R_1 - y_{+1} + R_2}{y_{+1} - R_2} \xrightarrow{p} \frac{\lim_{R\to\infty}\sum_{i=1}^{R} L_{1i}}{\lim_{R\to\infty}\sum_{i=1}^{R} L_{2i}} = e^{2\psi_1}. \tag{2.12}$$

Di conseguenza, utilizzando l'ultimo risultato (2.12) nell'espressione dello stimatore di massima verosimiglianza (2.11), quando R diverge

$$\hat{\psi}_1 = \log\Big\{\frac{R_1 - y_{+1} + R_2}{y_{+1} - R_2}\Big\} \xrightarrow{p} \log\left(e^{2\psi_1}\right) = 2\psi_1.$$

Quindi si ha che lo stimatore di massima verosimiglianza per ψ_1 non è consistente.

Tuttavia, è possibile ottenere uno stimatore consistente per ψ_1 basato sulla verosimiglianza condizionata. Infatti, la funzione di log-verosimiglianza (2.8), utilizzando il

vincolo $\psi_2 = -\psi_1$, può anche essere scritta come

$$\ell(\psi_1, \alpha) = \sum_{i=1}^{R} \alpha_i y_{i+} - \psi_1 \sum_{i=1}^{R} (y_{i1} - y_{i2}) - \sum_{i=1}^{R} \log\left(1 + e^{\alpha_i - \psi_1}\right) - \sum_{i=1}^{R} \log\left(1 + e^{\alpha_i + \psi_1}\right),$$

che è la log-verosimiglianza di una famiglia esponenziale di ordine $R+1$, con $(y_{1+}, \ldots, y_{R+})$ statistica parzialmente sufficiente per $\alpha = (\alpha_1, \ldots, \alpha_R)$. Ne consegue che il modello condizionato per $\sum_{i=1}^{R}(y_{i1} - y_{i2})$ dato $(y_{1+}, \ldots, y_{R+})$ è ancora una famiglia esponenziale di ordine 1, indipendente da α. Quindi, supponendo senza perdere di generalità che $y_{i+} = 1$ per le prime R_1 osservazioni ($R_1 < R$), allora

$$\ell_C(\psi_1) = \psi_1 \sum_{i=1}^{R_1} (y_{i1} - y_{i2}) - R_1 \log\left(1 + e^{-\psi_1}\right) - R_1 \log\left(1 + e^{\psi_1}\right), \tag{2.13}$$

e, come mostrato da Andersen (1980, Capitolo 6), lo stimatore basato sulla verosimiglianza condizionata $\ell_C(\psi_1)$ è consistente.

2.4 *Bootstrap* in modelli stratificati

Si consideri il contesto asintotico con due indici, con osservazioni indipendenti y_{ij}, con numerosità campionaria $N = RC$, dove i è l'indice relativo agli strati, mentre j è l'indice relativo all'osservazione j-esima nello strato i-esimo. Si assuma inoltre che il numero di strati R cresca con la numerosità C di ciascuno strato in modo che $R = O(C^{\upsilon})$, per $\upsilon > 0$. Il caso $\upsilon = 0$ corrisponde al contesto asintotico standard, in cui il numero di strati è limitato e si assume di avere un numero sufficiente di osservazioni per ciascuno strato.

Sartori (2003) ha mostrato che le classiche quantità pivotali basate sulla verosimiglianza, $r_P(\psi)$, $r_{Pe}(\psi)$ e $r_{Pu}(\psi)$ sono asintoticamente equivalenti con un errore di ordine $o(1)$, per $\upsilon \geq 0$. In particolare, quando $0 \leq \upsilon \leq 1$, le tre quantità pivotali sono asintoticamente equivalenti con un errore relativo di ordine $O(N^{-1/2}) = O(C^{-(\upsilon+1)/2})$, e asintoticamente normali standard. Tuttavia, quando $\upsilon \geq 1$, l'equivalenza asintotica tra le tre quantità pivotali è ancora valida, ma con errore di ordine $O(C^{-1})$, e le tre quantità non sono più asintoticamente distribuite come una normale standard, e pertanto, ad esempio, $\Phi(r_P(\psi))$ non risulta asintoticamente uniforme.

Lo studio delle proprietà asintotiche delle quantità di verosimiglianza nel caso dei modelli stratificati è più semplice se si considera la statistica $r_{Pu}(\psi)$, in quanto la funzione punteggio coincide con la somma delle funzioni punteggio dei singoli strati, in virtù

dell'indipendenza tra gli strati. Ad ogni modo, i risultati sono equivalenti per $r_P(\psi)$ e $r_{Pe}(\psi)$ in quanto asintoticamente equivalenti.

Si denoti con $F_\theta(\cdot)$ la funzione di ripartizione di $r_{Pu}(\psi)$ sotto θ. Ne consegue che $F_\theta(r_{Pu}(\psi))$ è esattamente uniforme. Bellio et al. (2023b) hanno dimostrato che sia l'*unconstrained bootstrap* che il *constrained bootstrap* rimangono procedure valide asintoticamente anche nel paradigma asintotico con due indici, sotto la condizione che $\upsilon < 3$, ossia quando $R = o(C^3)$. Questa condizione coincide con la condizione richiesta affinché la statistica $r_P^*(\psi)$ basata sulla verosimiglianza profilo modificata sia valida per l'inferenza su ψ, come mostrato da Sartori (2003).

Nello specifico, Bellio et al. (2023b) hanno dimostrato che, in modelli per dati continui, quando $0 < \upsilon < 1$, indicando con $\hat{\theta}$ la stima globale e con $\hat{\theta}_\psi$ la stima vincolata, allora

$$\Pr_\theta\Big\{F_{\hat{\theta}_\psi}(r_{Pu}(\psi) \leq u)\Big\} = u + O(C^{(\upsilon-3)/2}),$$

e

$$\Pr_\theta\Big\{F_{\hat{\theta}}(r_{Pu}(\psi) \leq u)\Big\} = u + O(C^{-1}).$$

Mentre, quando $1 \leq \upsilon < 3$

$$\Pr_\theta\Big\{F_{\hat{\theta}_\psi}(r_{Pu}(\psi) \leq u)\Big\} = u + O(C^{(\upsilon-3)/2}),$$

e

$$\Pr_\theta\Big\{F_{\hat{\theta}}(r_{Pu}(\psi) \leq u)\Big\} = u + O(C^{(\upsilon-3)/2}).$$

Pertanto, quando $1 \leq \upsilon < 3$, utilizzare l'*unconstrained bootstrap* o la versione *constrained* garantisce lo stesso livello di accuratezza, al contrario del caso $0 < \upsilon < 1$ dove utilizzare la versione *constrained* permette di ottenere un'accuratezza teorica maggiore.

Da un punto di vista intuitivo, il motivo per cui le due varianti *bootstrap* mostrano un comportamento analogo quando $\upsilon \geq 1$ è che l'effetto principale del *bootstrap* è quello di rimuovere il termine di distorsione che tende a divergere in scenari estremi. Di conseguenza, per $\upsilon \geq 1$ le differenze teoriche relativamente al comportamento delle due varianti *bootstrap* vengono mascherate dal termine di distorsione. Al contrario, per $0 \leq \upsilon < 1$, i due metodi presentano delle differenze teoriche non trascurabili.

Alla luce della scomposizione della funzione punteggio profilo (2.2) è immediato intuire da dove provengono i benefici nell'utilizzo del *bootstrap*. Nella pratica, infatti, per $1 \leq v < 3$, entrambe le varianti *bootstrap* (così come le modificazioni analitiche basate su approssimazioni di ordine superiore) risultano in grado di correggere il termine di distorsione B nella (2.2), facendo in modo che $\ell_{\psi|\lambda}$ sia il termine dominante nella scomposizione (2.2). Per una dimostrazione più dettagliata di tali risultati si veda Bellio et al. (2023b).

Dunque, nel caso di modelli stratificati per dati continui, Bellio et al. (2023b) hanno mostrato che, da un punto di vista teorico, il *bootstrap* parametrico permette di ripristinare la validità delle conclusioni inferenziali, con un comportamento asintoticamente leggermente migliore del *constrained bootstrap* in scenari meno estremi. Anche gli studi di simulazione in Bellio et al. (2023b), per modelli stratificati come i modelli di regressione beta o lineare troncato, confermano i risultati teorici. Tuttavia, le considerazioni teoriche sono limitate al caso continuo, dove sono soddisfatte le condizioni di regolarità per applicare le tecniche di approssimazione basate su sviluppi di Edgeworth (si veda, ad esempio, Severini, 2005, Capitolo 14). Nel caso di modelli per dati discreti, come un modello logistico per dati stratificati, le simulazioni in Bellio et al. (2023b) suggeriscono che l'equivalenza tra l'*unconstrained bootstrap* e la versione *constrained* nel migliorare l'accuratezza delle approssimazioni del primo ordine potrebbe non essere più valida. In particolare, seppur entrambe le varianti *bootstrap* mostrino una performance migliore rispetto all'utilizzo di approssimazioni del primo ordine, sembrerebbe che il *constrained bootstrap* sia in grado di portare a risultati sempre più accurati rispetto alla variante *unconstrained*, per diversi scenari in termini di valori di R e C.

2.5 *Bootstrap* in modelli con un numero elevato di parametri di disturbo

Zhao & Candès (2022) hanno studiato le proprietà del *bootstrap* parametrico nel caso del modello di regressione logistica in un regime asintotico a moderata dimensionalità, in uno scenario diverso da quello dei modelli stratificati discusso in precedenza, ma anch'esso caratterizzato dalla presenza di un numero elevato di parametri di disturbo. Nello specifico, nel modello considerato dagli Autori, si assume di disporre di N realizzazioni $y_1, \ldots, y_N$ da variabili casuali indipendenti $Y_1, \ldots, Y_N$, con $Y_i \in \{0, 1\}$ e $Y_i \sim$

$\text{Bi}(1, \pi_i)$, dove

$$\pi_i = \Pr(Y_i = 1) = g^{-1}(x_i^T \beta) = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}},$$

e $\Pr(Y_i = 0) = 1 - \Pr(Y_i = 1)$, $g(\pi_i) = \text{logit}(\pi_i)$, for $i = 1, \ldots, N$. Qui $\beta \in \mathbb{R}^p$ è il vettore dei parametri di regressione che devono essere stimati e $x_1, \ldots, x_N \in \mathbb{R}^p$ sono delle covariate. In particolare, Zhao & Candès (2022) hanno considerato lo scenario in cui le x_i sono indipendenti e con distribuzione marginale normale, ossia $x_i \sim N_p(0, n^{-1} I_p)$.

Gli Autori hanno evidenziato che entrambe le varianti del *bootstrap* parametrico non portano ad ottenere conclusioni accurate quando si utilizza l'usuale stima di massima verosimiglianza per generare i campioni *bootstrap*. Al contrario, se si penalizza la stima di massima verosimiglianza, introducendo uno *shrinkage* verso l'origine, è possibile ripristinare in alcuni scenari le usuali proprietà del *bootstrap* parametrico. Infatti, nella specificazione a moderata dimensionalità, quando sia $p \to \infty$ sia $N \to \infty$ e il rapporto $p/N \to \kappa$, con $\kappa \in (0, 1)$, lo stimatore di massima verosimiglianza esibisce una distorsione non trascurabile. Inoltre, il problema della separazione, e quindi della non esistenza della stima di massima verosimiglianza, si aggrava. Utilizzare una stima penalizzata permette di ridurre il problema della distorsione, di attenuare il problema delle stime infinite e consente al *bootstrap* parametrico di non sovrastimare la distorsione e la variabilità delle stime. Nel contesto ad elevata dimensionalità, Zhao et al. (2020) hanno infatti mostrato che, indicando con β_j il generico coefficiente di regressione e con τ_j la corrispondente deviazione standard, vale che

$$\frac{\sqrt{n}(\hat{\beta}_j - \alpha_* \beta_j)}{\sigma_* / \tau_j} \xrightarrow{d} N(0, 1),$$

dove α_* e σ_* sono fattori di correzione per la distorsione e la deviazione standard. Di conseguenza, la stima di massima verosimiglianza $\hat{\beta}_j$ non è centrata in β_j ma bensì in un valore traslato $\alpha_* \beta_j$.

Una possibile alternativa anche a quanto proposto da Zhao & Candès (2022) è quella di utilizzare, invece che l'usuale stima di massima verosimiglianza vincolata, una stima con distorsione ridotta in media o mediana, come proposto da Kosmidis et al. (2020). È ragionevole aspettarsi che utilizzare il *bootstrap* parametrico da una stima penalizzata con distorsione in media o in mediana ridotta possa portare a benefici anche nel caso di dati discreti.

In particolare, Firth (1993) ha mostrato che, nel caso di un modello di regressione logistica, utilizzare al posto dell'usuale funzione di verosimiglianza, la funzione di

verosimiglianza modificata

$$\ell_M(\beta) = \ell(\beta) + \frac{1}{2}\log|i(\beta)| = \ell(\beta) + \frac{1}{2}\log|X^T W X|, \tag{2.14}$$

dove $i(\beta) = j(\beta)$, interpretabile come verosimiglianza penalizzata, con termine di penalità dato dalla distribuzione a priori di Jefrreys per β, porta ad ottenere uno stimatore di massima verosimiglianza con distorsione asintoticamente inferiore rispetto all'usuale stimatore di massima verosimiglianza. Infatti, la penalità $\frac{1}{2}\log|i(\beta)|$ consente di ottenere una stimatore con distorsione media asintoticamente ridotta. Sebbene le procedure basate sulla penalizzazione della verosimiglianza fossero già ben note in letteratura (Good & Gaskins, 1971), dopo i risultati ottenuti da Firth (1993), tali metodi hanno avuto larga diffusione, mostrandosi efficaci anche in contesti diversi da quello per cui erano stati originariamente introdotti, come evidenziato da Kosmidis & Firth (2020). Lunardon (2018) ha mostrato i benefici dell'utilizzo di tecniche di riduzione della distorsione tramite penalizzazioni della verosimiglianza per l'inferenza su un parametro di interesse a bassa dimensionalità rispetto ad un parametro di disturbo ad elevata dimensionalità. Inoltre, Kosmidis & Firth (2020) hanno mostrato che nel contesto dei modelli di regressione logistica con specificazioni ad elevata dimensionalità, tali metodi di penalizzazione consentono di ottenere ottimi risultati, confrontabili con il nuovo approccio di Sur & Candès (2019) basato sulla teoria a moderata dimensionalità. Infine, di recente, Kosmidis & Firth (2020) hanno suggerito che in alcuni scenari può essere vantaggioso modificare il termine di penalità nella (2.14), considerando una verosimiglianza penalizzata più generale

$$\ell_M^{\dagger}(\beta) = \ell(\beta) + a\log|X^T W X|, \tag{2.15}$$

dove il coefficiente $a \geq 0$ può assumere valori diversi da $a = 1/2$, per cui si ottiene l'usuale verosimiglianza modificata (2.14), e da $a = 0$, che corrisponde alla verosimiglianza non penalizzata. In particolare, per $a = 1/6$, nel caso di un modello di regressione logistica con $p = 1$ in cui si è interessati a fare inferenza su uno solo dei coefficienti di regressione, utilizzare la verosimiglianza penalizzata $\ell_M^{\dagger}$ è equivalente all'approccio di riduzione della distorsione in mediana proposto da Kenne Pagui et al. (2017). Nel contesto dei modelli lineari generalizzati, la penalizzazione (2.15) è stata implementata nella libreria `brglm2` (Kosmidis et al., 2020), con il metodo `brglmFit`, specificando `MPL_Jeffreys` e il valore del coefficiente `a`.

Queste nuove metodologie di stima si prestano quindi ad essere facilmente combinate con le tecniche di *bootstrap* parametrico, e ulteriori approfondimenti sia teorici che basati

su metodi di simulazione sono necessari per valutare se e in che misura tali modifiche possano apportare miglioramenti all'usuale *bootstrap* parametrico in scenari complessi come quello dei modelli stratificati in presenza di effetti fissi stratificati o incrociati.

Capitolo 3

Modelli con effetti fissi incrociati

3.1 Introduzione

In questo capitolo vengono illustrate le principali caratteristiche dei modelli a due indici asintotici con effetti fissi incrociati.

Come descritto nel paragrafo 2.1 nel caso dei modelli con effetti stratificati, si consideri un campione di osservazioni indipendenti y_{ij}, $i = 1, \ldots, R$, $j = 1, \ldots, C$, in cui in generale ogni strato potrebbe avere una diversa numerosità campionaria C_i, dove i è l'indice relativo agli strati, mentre j è l'indice relativo all'osservazione j-esima nello strato i-esimo. La numerosità campionaria complessiva è dunque $N = \sum_{i=1}^{R} C_i$. Senza perdita di generalità, nel seguito si assume che tutti gli strati abbiano la medesima numerosità campionaria C e, pertanto, che la numerosità campionaria complessiva sia semplicemente $N = RC$. In tal caso, si parla anche di schema di dati bilanciato.

Si supponga che la densità di ogni y_{ij}, $i = 1, \ldots, R$, $j = 1, \ldots, C$, sia specificata da

$$p(y_{ij}; \psi, \alpha_i, \gamma_j),$$

dove il parametro strutturale ψ ha dimensione fissata k, mentre il parametro incidentale $\lambda = (\alpha, \gamma)$, con dimensione $R + C$, ha due componenti: l'effetto fisso $\alpha = (\alpha_1, \ldots, \alpha_R)$ e l'effetto fisso $\gamma = (\gamma_1, \ldots, \gamma_C)$. Il modello ha quindi parametro globale $\theta = (\psi, \alpha, \gamma)$, con dimensione $k + R + C$.

La densità di y_{ij} potrebbe essere condizionata ad un vettore di covariate x_{ij}, assunte note, come nel caso dei modelli lineari generalizzati. Per i modelli con effetti fissi

incrociati la funzione di log-verosimiglianza è

$$\ell(\psi, \lambda) = \ell(\psi, \alpha, \gamma) = \sum_{i=1}^{R} \sum_{j=1}^{C} \log p(y_{ij}; \psi, \alpha_i, \gamma_j).$$

In questo tipo di scenario, a differenza dei modelli a due indici con un unico effetto fisso, non si verifica più la semplice separazione in termini dei parametri incidentali α_i (o equivalentemente γ_j), poiché la verosimiglianza nello strato i-esimo non dipende più solamente dal parametro α_i ma anche da γ. Ciò fa sì che la trattazione teorica dei risultati inferenziali in questo tipo di modelli sia più complessa e priva di risultati generali come quelli presenti nel caso di un unico effetto fisso.

Questo paradigma asintotico, è stato investigato in particolare nella letteratura econometrica, dato l'ampio utilizzo di questi modelli per dati di *panel*. La modellazione statistica di questo tipo di dati prevede spesso la specificazione di un effetto specifico per ogni singolo strato, come spiegato nel paragrafo 2.1. Tuttavia, nel caso di effetti fissi incrociati, si introduce anche un effetto specifico per l'osservazione nello strato. Ciò permette di controllare anche le caratteristiche invarianti a livello cross-sezionale. Usualmente ci si riferisce a questi modelli come *two-way fixed effects models* oppure *large-R, C panel data models*. Per ulteriori dettagli sulla motivazione dell'impiego di questi modelli e per una trattazione più approfondita, si rimanda, ad esempio, a Jochmans & Otsu (2019), Fernàndez-Val & Weidner (2016), Fernàndez-Val & Weidner (2018) e Leng et al. (2023).

3.2 Inferenza in modelli con effetti fissi incrociati

In questo paragrafo si riassumo alcuni dei risultati disponibili circa le procedure inferenziali nello scenario dei modelli con effetti fissi incrociati, facendo particolare riferimento ai contributi provenienti dalla letteratura econometrica recente.

L'inferenza basata sulla verosimiglianza profilo non è accurata in presenza di un numero parametri di disturbo elevato rispetto alla numerosità campionaria N. Ciò risulta evidente nel paradigma dei modelli a due indici asintotici anche nel caso di un solo effetto fisso α, e dunque un parametro di disturbo specifico α_i per ciascuno strato $i = 1, \ldots, R$, dove la distorsione della funzione punteggio profilo è di ordine $O(R)$, come descritto nel paragrafo 2.2. La mancanza di accuratezza delle procedure inferenziali basate sulla verosimiglianza profilo rischia di aggravarsi nel caso di effetti fissi incrociati, $\alpha = (\alpha_1, \ldots, \alpha_R)$ e $\gamma = (\gamma_1, \ldots, \gamma_C)$, in quanto tipicamente c'è una maggiore

difficoltà nell'ottenere delle stime vincolate $\hat{\alpha}_\psi$ e $\hat{\gamma}_\psi$ accurate, se i dati non contengono informazione a sufficienza sulla componente di disturbo.

Lo studio delle procedure inferenziali e delle proprietà distributive delle quantità di verosimiglianza nello scenario dei modelli con effetti fissi incrociati è più complesso che nel caso dei modelli con effetti fissi stratificati. Infatti, almeno per le proprietà asintotiche, è cruciale effettuare una scelta del paradigma che si vuole considerare circa il comportamento di R e C. Motivata anche dalla tipologia dei dati oggi disponibili, la letteratura recente si è concentrata sullo studio delle proprietà asintotiche degli stimatori quando sia $R \to \infty$ sia $C \to \infty$. D'altra parte, nello scenario in cui C è fissato e $R \to \infty$, Chamberlain (2010) ha mostrato che, nel caso dei modelli per risposta binaria, non è possibile ottenere uno stimatore consistente per ψ, quando C è fissato.

Se $\hat{\alpha} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_R)$ e $\hat{\gamma} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_C)$ denotano le usuali stime di massima verosimiglianza per gli effetti fissi α e γ rispettivamente, allora, come mostrato, ad esempio, in Fernàndez-Val & Weidner (2018), tali stime risentono del problema dei parametri incidentali. Infatti, a meno che C non sia molto grande, la stima di ciascun parametro di disturbo α_i specifico dello strato i-esimo, $i = 1, \ldots, R$, risulta poco accurata, in quanto si dispone di sole C osservazioni che contengono informazioni su ciascun α_i. In modo simmetrico, a meno che R non sia molto grande, la stima di ciascun parametro di disturbo γ_j risulta poco accurata. La scarsa accuratezza nella stima degli effetti fissi incrociati α e γ si traduce quindi in una scarsa accuratezza nella stima del parametro di interesse ψ.

Leng et al. (2023) hanno mostrato che la distorsione nella stima di ψ diventa asintoticamente trascurabile qualora $R/C \to 0$, quando $R, C \to \infty$, se il modello include solo gli effetti fissi di riga α_i, e in modo speculare, qualora $R/C \to \infty$, quando $R, C \to \infty$, se il modello include solo gli effetti di colonna γ_j. Tuttavia, quando sono presenti entrambi gli effetti incrociati, non sembra possibile individuare, in modo generale per ogni possibile modello, un tasso relativo R/C per cui, anche quando $R, C \to \infty$, tale distorsione diventi asintoticamente trascurabile, almeno rispetto allo *standard error*.

Come suggerito da Fernàndez-Val & Weidner (2018), vi sono alcune eccezioni, come nel caso del modello log-lineare Poisson, approfondito nel paragrafo 3.3.2, in cui, il problema dei parametri incidentali diventa asintoticamente trascurabile al divergere di R e C. Infatti, nel paradigma asintotico in cui sia $R \to \infty$ sia $C \to \infty$, Fernàndez-Val & Weidner (2016) hanno mostrato che la distorsione nella stima di ψ è asintoticamente trascurabile, al divergere di R e C. Al contrario, nel caso del modello di regressione logistica, approfondito nel paragrafo 3.3.3, il termine di distorsione non risulta trascurabile nemmeno quando $R, C \to \infty$.

Più in generale, (Fernàndez-Val & Weidner, 2018, §3.3) hanno caratterizzato la distorsione della stima di massima verosimiglianza nel caso in cui $R, C \to \infty$, quando R/C tende ad una costante finita, mostrando che tale distorsione è di ordine $O(1/R)+O(1/C)$. Il termine di distorsione $O(1/C)$ deriva dalla stima degli effetti specifici di ciascuno strato, in quanto ci sono solamente C osservazioni a disposizione per ognuno di essi. In modo speculare, il termine di distorsione $O(1/R)$ deriva dalla stima degli effetti specifici all'interno di ciascuno strato, in quanto ci sono solamente R osservazioni per ognuno di essi.

Dunque, a differenza dei modelli con soli effetti fissi stratificati α_i, nel caso di effetti fissi incrociati, la necessità di dover stimare anche i parametri γ_j introduce un ulteriore termine di distorsione $O(1/R)$. In particolare, se R e C sono confrontabili, allora entrambi i termini di distorsione hanno un peso paragonabile nell'accuratezza della stima del parametro di interesse. Pertanto, è necessario utilizzare delle procedure che permettano di correggere la stima di ψ per ambedue i termini di distorsione. Fernàndez-Val & Weidner (2018) hanno proposto essenzialmente due possibili approcci: utilizzare delle correzioni analitiche per le quantità di distorsione, oppure fare ricorso ad uno stimatore di tipo *jackknife*. Altre possibili alternative all'utilizzo di tecniche standard come la verosimiglianza profilo modificata, sono state proposte da Jochmans & Otsu (2019), che hanno suggerito dei termini di modificazione della verosimiglianza profilo alternativi a quello di Barndorff-Nielsen (1980), sviluppati nello scenario specifico degli effetti fissi incrociati. Recentemente, anche Leng et al. (2023) hanno affrontato il problema dei parametri incidentali nel caso di effetti fissi incrociati, suggerendo un'altra possibilità per ridurre la distorsione nella stima del parametro di interesse, tramite una modifica dell'usuale funzione di verosimiglianza, nello scenario in cui $R, C \to \infty$, con R e C che crescono proporzionalmente. Il loro approccio differisce da quello considerato ad esempio in Fernàndez-Val & Weidner (2016), in quanto, la correzione proposta è applicata direttamente alla funzione di verosimiglianza, in modo simile a quanto suggerito da Jochmans & Otsu (2019).

3.3 Esempi di modelli con effetti fissi incrociati

Nel seguito si illustrano alcuni semplici esempi di modelli con effetti fissi incrociati, a partire dal modello di regressione normale con effetti fissi incrociati (si veda, ad esempio, Jochmans & Otsu, 2019), per proseguire con due modelli per dati discreti: il modello log-lineare Poisson con effetti fissi incrociati e il modello logistico con effetti

fissi incrociati (si veda, ad esempio, Fernàndez-Val & Weidner, 2018). Per questi modelli, si discutono i problemi circa le usuali procedure inferenziali che possono sorgere in presenza di un numero elevato di parametri di disturbo, e come tali problemi possano essere in parte risolti, seppur con maggiori difficoltà, rispetto allo scenario dei modelli stratificati, utilizzando alcuni dei metodi introdotti in precedenza.

È possibile formulare delle considerazioni di carattere più generale riguardo ai modelli log-lineare Poisson e logistico con effetti fissi incrociati, poiché essi condividono una struttura simile. Di conseguenza, molte delle conclusioni nei paragrafi 3.3.2 e 3.3.3, nei quali vengono discussi rispettivamente i due modelli, discendono immediatamente da questa trattazione più generale. Infatti, entrambi i modelli possono essere visti come un caso particolare del modello statistico che assume che y_{ij} siano realizzazioni di variabili casuali indipendenti Y_{ij} con densità

$$p(y_{ij}; \psi, \alpha_i, \gamma_j) = \exp\{\theta_{ij} y_{ij} - K(\theta_{ij})\}, \tag{3.1}$$

dove $\theta_{ij} = \alpha_i + \gamma_j + \psi x_{ij}$, per $i = 1, \ldots, R$, $j = 1, \ldots, C$. Il modello (3.1) appartiene ad una famiglia esponenziale di ordine $R + C + 1$, in parametrizzazione canonica, con $K(\theta_{ij})$ funzione generatrice dei cumulanti. Al solito, il vincolo di identificabilità può essere rispettato ponendo a zero uno degli α_i o dei γ_j.

La funzione di verosimiglianza per (ψ, α, γ) basata sui dati y_{ij}, $i = 1, \ldots, R$, $j = 1, \ldots, C$, è

$$\begin{aligned}
L(\psi, \alpha, \gamma) &= \prod_{i=1}^{R} \prod_{j=1}^{C} p(y_{ij}; \psi, \alpha_i) \\
&= \prod_{i=1}^{R} \prod_{j=1}^{C} \exp\Big\{(\alpha_i + \gamma_j + \psi x_{ij}) y_{ij} - K(\theta_{ij})\Big\} \\
&= \exp\Big\{\sum_{i=1}^{R}\sum_{j=1}^{C} \alpha_i y_{ij} + \sum_{i=1}^{R}\sum_{j=1}^{C} \gamma_j y_{ij} + \psi \sum_{i=1}^{R}\sum_{j=1}^{C} x_{ij} y_{ij} - \sum_{i=1}^{R}\sum_{j=1}^{C} K(\theta_{ij})\Big\} \\
&= \exp\Big\{\sum_{i=1}^{R} \alpha_i \sum_{j=1}^{C} y_{ij} + \sum_{j=1}^{C} \gamma_j \sum_{i=1}^{R} y_{ij} + \psi \sum_{i=1}^{R}\sum_{j=1}^{C} x_{ij} y_{ij} - \sum_{i=1}^{R}\sum_{j=1}^{C} K(\theta_{ij})\Big\} \\
&= \exp\Big\{\sum_{i=1}^{R} \alpha_i y_{i+} + \sum_{j=1}^{C} \gamma_j y_{+j} + \psi s - \sum_{i=1}^{R}\sum_{j=1}^{C} K(\theta_{ij})\Big\}
\end{aligned}$$

dove $y_{i+} = \sum_{j=1}^{C} y_{ij}$ è il totale di riga i-esimo, $y_{+j} = \sum_{i=1}^{R} y_{ij}$ è il totale di colonna

j-esimo e $s = \sum_{i=1}^{R} \sum_{j=1}^{C} x_{ij} y_{ij}$. La funzione di log-verosimiglianza è quindi

$$\ell(\psi, \alpha, \gamma) = \sum_{i=1}^{R} \alpha_i y_{i+} + \sum_{j=1}^{C} \gamma_j y_{+j} + \psi s - \sum_{i=1}^{R} \sum_{j=1}^{C} K(\theta_{ij})$$

e le statistiche y_{i+} e y_{+j}, $i = 1, \ldots, R$, $j = 1, \ldots, C$, sono parzialmente sufficienti per α e γ. Di conseguenza, la verosimiglianza condizionata per ψ si ottiene semplicemente condizionandosi sia al vettore dei totali di riga $(y_{1+}, \ldots, y_{R+})$ sia al vettore dei totali di colonna $(y_{+1}, \ldots, y_{+C})$. Tuttavia, tipicamente non esiste una forma chiusa a causa dell'espressione spesso complessa della funzione $K_+(\theta_{ij})$, dove il simbolo + a deponente indica che si fa riferimento alle statistiche relative ai totali di riga e di colonna. Inoltre, anche computazionalmente il calcolo della verosimiglianza condizionata può risultate oneroso, in quanto, almeno a livello teorico, richiede di considerare il condizionamento a tutti i possibili insieme di dati y_{ij}, $i = 1, \ldots, R$, $j = 1, \ldots, C$, che hanno gli stessi totali di riga e di colonna. Infatti, se Y denota la matrice casuale corrispondente a tutte le osservazioni y_{ij}, $i = 1, \ldots, R$, $j = 1, \ldots, C$, nel caso discreto l'espressione generale della funzione di probabilità condizionata è

$$\Pr(Y | (y_{1+}, \ldots, y_{R+}), (y_{+1}, \ldots, y_{+C}); \psi, \alpha, \gamma) = \frac{\Pr(Y; \psi, \alpha, \gamma)}{\sum_{\tilde{Y} \in Q} \Pr(Y = \tilde{Y}; \psi, \alpha, \gamma)},$$

dove Q è l'insieme delle possibile combinazioni di y_{ij} con stessi totali di riga e di colonna. Tuttavia, come anticipato, la somma che compare nel denominatore dell'ultima espressione non è tipicamente computazionalmente trattabile nella pratica.

Al contrario, la log-verosimiglianza profilo modificata è semplice da ottenere. Infatti, nel caso di una famiglia esponenziale in parametrizzazione canonica, l'informazione attesa coincide con l'informazione osservata. Questa ha la seguente struttura, nel caso di un modello con effetti fissi incrociati

$$j(\theta) = \left[\begin{array}{c|cccc|cccc} j_{\psi\psi} & j_{\psi\alpha_1} & j_{\psi\alpha_2} & \cdots & j_{\psi\alpha_{R-1}} & j_{\psi\gamma_1} & j_{\psi\gamma_2} & \cdots & j_{\psi\gamma_C} \\ \hline j_{\psi\alpha_1} & j_{\alpha_1\alpha_1} & j_{\alpha_1\alpha_2} & \cdots & j_{\alpha_1\alpha_{R-1}} & j_{\alpha_1\gamma_1} & j_{\alpha_1\gamma_2} & \cdots & j_{\alpha_1\gamma_C} \\ j_{\psi\alpha_2} & j_{\alpha_1\alpha_2} & \ddots & \cdots & \vdots & j_{\alpha_2\gamma_1} & \ddots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \ddots & \vdots & \vdots & \cdots & \ddots & \vdots \\ j_{\psi\alpha_{R-1}} & j_{\alpha_1\alpha_{R-1}} & \cdots & \cdots & j_{\alpha_{R-1}\alpha_{R-1}} & j_{\alpha_{R-1}\gamma_1} & \cdots & \cdots & j_{\alpha_{R-1}\gamma_C} \\ \hline j_{\psi\gamma_1} & j_{\alpha_1\gamma_1} & j_{\alpha_2\gamma_1} & \cdots & j_{\alpha_{R-1}\gamma_1} & j_{\gamma_1\gamma_1} & j_{\gamma_1\gamma_2} & \cdots & j_{\gamma_1\gamma_C} \\ j_{\psi\gamma_2} & j_{\alpha_1\gamma_2} & \ddots & \cdots & \vdots & j_{\gamma_2\gamma_1} & \ddots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \ddots & \vdots & \vdots & \cdots & \ddots & \vdots \\ j_{\psi\gamma_C} & j_{\alpha_1\gamma_C} & \cdots & \cdots & j_{\alpha_{R-1}\gamma_C} & j_{\gamma_1\gamma_C} & \cdots & \cdots & j_{\gamma_C\gamma_C} \end{array}\right],$$

che può essere riscritta come matrice a blocchi

$$j(\theta) = \left[\begin{array}{c|c|c} j_{\psi\psi} & j_{\psi\alpha}^T & j_{\psi\gamma}^T \\ \hline j_{\psi\alpha} & j_{\alpha\alpha} & j_{\alpha\gamma}^T \\ \hline j_{\psi\gamma} & j_{\alpha_\gamma} & j_{\gamma\gamma} \end{array}\right],$$

dove $j_{\psi\psi}$ è scalare, $j_{\alpha\alpha}$ è una matrice diagonale di dimensione $(R-1)\times(R-1)$, $j_{\gamma\gamma}$ è una matrice diagonale di dimensione $C \times C$, $j_{\alpha\gamma}$ è una matrice di dimensione $(R-1) \times C$, mentre $j_{\psi\alpha}$ è un vettore di lunghezza $(R-1)$ e $j_{\psi\gamma}$ è un vettore di lunghezza C.

Inoltre, le derivate nello spazio campionario $\ell_{\alpha_i;\hat{\alpha}_i}(\hat{\theta}_\psi)$ $\ell_{\alpha_i;\hat{\gamma}_j}(\hat{\theta}_\psi)$, e in modo speculare $\ell_{\gamma_j;\hat{\gamma}_j}(\hat{\theta}_\psi)$ $\ell_{\gamma_j;\hat{\alpha}_i}(\hat{\theta}_\psi)$ dipendono solo dai dati. Di conseguenza, si ha che

$$\ell_M(\psi) = \ell_P(\psi) + \frac{1}{2}\log|j_{\lambda\lambda}(\hat{\theta}_\psi)|,$$

dove $j_{\lambda\lambda}(\hat{\theta}_\psi)$ è il blocco relativo ai parametro di disturbo della la matrice di informazione osservata calcolata nella stima vincolata $\hat{\theta}_\psi = (\psi, \hat{\alpha}_\psi, \hat{\gamma}_\psi)$. Usualmente, per questi modelli, le stime vincolate $\hat{\alpha}_\psi$ e $\hat{\gamma}_\psi$ devono essere ottenute numericamente e non esiste una forma chiusa per la funzione di log-verosimiglianza profilo per ψ.

D'altra parte, è semplice ottenere la versione modificata $r_P^*(\psi)$ dell'usuale statistica test radice con segno del log-rapporto di verosimiglianza profilo $r_P(\psi)$. Infatti è sufficiente utilizzare l'usuale statistica di Wald $r_{Pe}(\psi)$ e il blocco $j_{\lambda\lambda}(\theta)$, calcolato nella stima globale $\hat{\theta}$ e nella stima vincolata $\hat{\theta}_\psi$, ossia

$$r_P^*(\psi) = r_P(\psi) + \frac{1}{r_P(\psi)}\log\left\{\frac{C(\psi)u_P(\psi)}{r_P(\psi)}\right\}$$

dove

$$u_P(\psi) = r_{Pe}(\psi), \qquad \text{e} \qquad C(\psi) = \left\{\frac{j_{\lambda\lambda}(\hat{\theta})}{j_{\lambda\lambda}(\hat{\theta}_\psi)}\right\}^{1/2}.$$

In alcune circostanze, per questi modelli, può essere utile prendere in considerazione la parametrizzazione mista $(\psi, \mu_{1+}, \ldots, \mu_{R+}, \mu_{+1}, \ldots, \mu_{+C})$, dove

$$\mu_{i+} = \sum_{j=1}^{C}\mu_{ij}, \qquad \text{e} \qquad \mu_{+j} = \sum_{i=1}^{R}\mu_{ij},$$

con il vincolo $\sum_{i=1}^{R}\mu_{i+} = \sum_{j=1}^{C}\mu_{+j}$.

Inoltre, per il modello log-lineare Poisson e logistico con effetti fissi incrociati, può essere utile considerare la seguente notazione matriciale, che permette di utilizzare risultati noti nel caso dei modelli lineare generalizzati e semplifica le procedure di simulazione da questi modelli quando è di interesse mantenere fissa la struttura degli effetti incrociati. In particolare, Y è rappresentato come un vettore N-dimensionale ottenuto scorrendo la matrice con elementi Y_{ij} nel senso delle righe. Inoltre, se X è la matrice del disegno, l'usuale predittore lineare $\eta = X\theta$ può essere scritto come

$$\eta = X\theta = \begin{bmatrix} x & Z_1 & Z_2 \end{bmatrix} \begin{bmatrix} \psi \\ \alpha \\ \gamma \end{bmatrix} = \psi x + Z_1\alpha + Z_2\gamma, \tag{3.2}$$

dove la matrice X può essere partizionata come $X = [x, Z_1, Z_2]$, dove in generale x è un vettore di lunghezza N, Z_1 è una matrice di dimensione $N \times R$ e Z_2 è una matrice di dimensione $N \times C$, dove $N = RC$. Le matrici Z_1 e Z_2, indicatrici degli effetti fissi incrociati, possono essere ottenute come

$$\underset{N\times R}{Z_1} = I_R \otimes 1_C, \qquad \underset{N\times C}{Z_2} = [1_R^T \otimes I_C]^T = 1_R \otimes I_C,$$

dove $\otimes$ indica il prodotto di Kronecker, I_R e I_C sono le matrici identità di dimensione R ed C rispettivamente, mentre 1_R e 1_C sono i vettori unitari di lunghezza R ed C rispettivamente. Il generico elemento di η è, quindi, dato dalla funzione di legame $g(\cdot)$ applicata al corrispondente elemento di $\mathbb{E}(Y)$.

L'informazione attesa, che per questi modelli coincide con $j(\theta)$, può essere scritta in forma matriciale come

$$j(\theta) = X^T W X,$$

dove $W = \text{diag}(w_{ij})$, con $1/w_{ij} = (g'(\mu_{ij}))^2 \text{Var}(Y_{ij})$, $i = 1, \ldots, R$, $j = 1, \ldots, C$. Sfruttando la notazione con le matrici Z_1 e Z_2 precedentemente introdotta, il blocco dell'informazione osservata relativo ai parametri di disturbo può essere scritto come

$$j_{\lambda\lambda}(\theta) = [Z_1, Z_2]^T \,\text{diag}(w_{ij})[Z_1, Z_2].$$

In generale le matrici Z_1 e Z_2 che svolgono ruolo di matrici indicatrici per gli effetti fissi α_i e γ_j sono assunte "piene", nel senso che per ogni unità statistica è presente sia l'effetto fisso α_i che γ_j. Tuttavia, in diversi contesti applicativi può essere ragionevole assumere che le matrici Z_1 e Z_2 siano "sparse", nel senso che non tutte le unità statistiche

siano caratterizzate dagli effetti fissi α_i e γ_j. In altri termini, si può assumere che vi sia sparsità nella struttura degli effetti fissi incrociati. Alcune possibili approcci per introdurre tale sparsità sono discussi nel Capitolo 4, dedicato agli studi di simulazione. Quando vi è sparsità nei dati e, quindi, nella struttura degli effetti fissi incrociati, le numerosità di riga e di colonna non sono più R e C, bensì r_i e c_j, con $r_i < R$ e $c_j < C$, rispettivamente. Pertanto, non sono più garantiti i risultati teorici descritti nel paragrafo 3.2.

3.3.1 Modello normale con effetti fissi incrociati

Nello stesso scenario introdotto nel paragrafo 2.3.1, un modello statistico alternativo è quello che assume che y_{ij} siano realizzazioni di variabili casuali indipendenti Y_{ij} con distribuzione normale di media $\alpha_i + \gamma_j$ e varianza ψ, per $i = 1, \ldots, R$, $j = 1, \ldots, C$. La densità di Y_{ij} per questo modello è

$$p(y_{ij}; \psi, \alpha_i, \gamma_j) = \frac{1}{\sqrt{2\pi\psi}} \exp\Big\{-\frac{1}{2\psi}(y_{ij} - \alpha_i - \gamma_j)^2\Big\},$$

per $i = 1, \ldots, R$, $j = 1, \ldots, C$. Il modello può essere scritto nella forma equivalente $Y_{ij} = \alpha_i + \gamma_j + \epsilon_{ij}$, dove $\epsilon_{ij} \sim N(0, \psi)$ i.i.d., $i = 1, \ldots, R$, $j = 1, \ldots, C$. In questo modello il parametro di interesse ψ è scalare, mentre il parametro incidentale $\lambda = (\alpha, \gamma) = (\alpha_1, \ldots, \alpha_R, \gamma_1, \ldots, \gamma_C)$ ha dimensione $R + C$. In realtà, il modello così specificato è sovraparametrizzato e il vincolo di identificabilità può essere rispettato ponendo, ad esempio, a zero uno degli α_i.

La funzione di verosimiglianza per (ψ, α, γ) basata sui dati y_{ij}, $i = 1, \ldots, R$, $j = 1, \ldots, C$, è

$$\begin{aligned} L(\psi, \alpha, \gamma) &= \prod_{i=1}^{R} \prod_{j=1}^{C} p(y_{ij}; \psi, \alpha_i) \\ &= (\psi)^{-RC/2} \exp\Big\{-\frac{1}{2} \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(y_{ij} - \alpha_i - \gamma_j)^2}{\psi}\Big\}. \end{aligned}$$

Ne segue che la funzione di log-verosimiglianza può essere scritta nel modo seguente

$$\ell(\psi, \alpha, \gamma) = -\frac{RC}{2} \log \psi - \frac{1}{2} \sum_{i=1}^{R} \sum_{j=1}^{C} \frac{(y_{ij} - \alpha_i - \gamma_j)^2}{\psi}.$$

Definite le quantità

$$\bar{y}_{i+} = \frac{1}{C}\sum_{j=1}^{C} y_{ij}, \quad i = 1, \ldots, R,$$
$$\bar{y}_{+j} = \frac{1}{R}\sum_{i=1}^{R} y_{ij}, \quad j = 1, \ldots, C,$$
$$\bar{y} = \frac{1}{RC}\sum_{i=1}^{R}\sum_{j=1}^{C} y_{ij},$$

come mostrato da Jochmans & Otsu (2019), la stima di massima verosimiglianza per ψ risulta pari a

$$\hat{\psi} = \frac{1}{RC}\sum_{i=1}^{R}\sum_{j=1}^{C}\Big\{(y_{ij} - \bar{y}) - (\bar{y}_i - \bar{y}) - (\bar{y}_j - \bar{y})\Big\}^2,$$

con distribuzione asintotica

$$\hat{\psi} \dot{\sim} \; N\Big(\psi - \frac{\psi}{R} - \frac{\psi}{C} + \frac{\psi}{RC}, \frac{2\psi^2}{RC}\Big).$$

Nella funzione di log-verosimiglianza è simmetrica rispetto i parametri di disturbo α_i e γ_j hanno un ruolo speculare, quindi le funzioni punteggio profilo rispetto ad α_i e γ_j hanno la stessa forma, ossia

$$\frac{\partial}{\partial\alpha_i}\ell(\psi, \alpha, \gamma) = \frac{1}{\psi}\sum_{j=1}^{C}(y_{ij} - \alpha_i - \gamma_j),$$
$$\frac{\partial}{\partial\gamma_j}\ell(\psi, \alpha, \gamma) = \frac{1}{\psi}\sum_{i=1}^{R}(y_{ij} - \alpha_i - \gamma_j).$$

Ponendo uguale a zero la funzione punteggio profilo per α_i si ottiene l'equazione

$$\sum_{j=1}^{C}(y_{ij} - \alpha_i - \gamma_j) = 0,$$

che, risolta rispetto ad α_i, per γ fissato, permette di ottenere la stima di α_i per ψ e γ fissati, $\hat{\alpha}_{i\psi\gamma}$, che in questo caso specifico non dipende da ψ

$$\hat{\alpha}_{i\gamma} = \bar{y}_{i+} - \frac{1}{C}\sum_{j=1}^{C}\gamma_j.$$

Sostituendo $\hat{\alpha}_{i\gamma}$ nella funzione punteggio profilo per γ_j si ottiene l'equazione

$$\sum_{i=1}^{R}(y_{ij} - \hat{\alpha}_{i\gamma} - \gamma_j) = 0$$
$$R\bar{y}_{+j} - \sum_{i=1}^{R}\bar{y}_{i+} + \frac{1}{C}\sum_{h=1}^{C}\gamma_h - \gamma_j = 0$$
$$R\bar{y}_{+j} - C^2 R\bar{y} + \frac{1}{C}\sum_{h=1}^{C}\gamma_h - \gamma_j = 0.$$

Per ottenere la stima $\hat{\gamma}_j$, che deve essere poi sostituita in $\hat{\alpha}_{i\gamma}$ per ricavare la stima $\hat{\alpha}_i$, bisogna risolvere il seguente sistema lineare in $\gamma = [\gamma_j]$

$$\Big(\frac{1}{C}1_C^T - I_C\Big)\gamma = C^2 R\bar{y} - R[\bar{y}_{+j}].$$

Quindi, in questo caso la stima vincolata del parametro di disturbo, $(\hat{\alpha}_\psi, \hat{\gamma}_\psi)$, non dipende da ψ.

La derivate seconde della log-verosimiglianza rispetto ai parametri di disturbo sono

$$\frac{\partial^2}{\partial\alpha_i\partial\alpha_{i'}}\ell(\psi,\alpha,\gamma) = \begin{cases} -\frac{C}{\psi} & i = i' \\ 0 & i \neq i' \end{cases}, \qquad \frac{\partial^2}{\partial\gamma_j\partial\gamma_{j'}}\ell(\psi,\alpha,\gamma) = \begin{cases} -\frac{R}{\psi} & j = j' \\ 0 & j \neq j' \end{cases},$$

$$\frac{\partial^2}{\partial\alpha_i\partial\gamma_j}\ell(\psi,\alpha,\gamma) = -\frac{1}{\psi},$$

da cui, consegue che il blocco relativo ai parametri di disturbo dell'informazione attesa, che coincide con l'informazione osservata, non dipende da λ. Dunque,

$$j_{\lambda\lambda}(\theta) = \begin{bmatrix} j_{\alpha\alpha} & j_{\alpha\gamma}^T \\ j_{\alpha\gamma} & j_{\gamma\gamma} \end{bmatrix} = \frac{1}{\psi}\begin{bmatrix} CI_{R-1} & 1_{R-1}1_C^T \\ 1_C 1_{R-1}^T & RI_C \end{bmatrix}.$$

La matrice $j_{\lambda\lambda}(\theta)$ ha dimensione $(R-1)+C\times(R-1)+C$, con $j_{\alpha\alpha}$ matrice diagonale di dimensione $(R-1)\times(R-1)$ e $j_{\gamma\gamma}$ matrice diagonale di dimensione $C\times C$.

Calcolando $j_{\lambda\lambda}(\theta)$ in corrispondenza della stima vincolata $\hat{\theta}_\psi = (\psi, \hat{\alpha}_\psi, \hat{\gamma}_\psi)$, è immediato ottenere il determinante di $j_{\lambda\lambda}(\hat{\theta}_\psi) = j_{\lambda\lambda}(\psi, \hat{\alpha}_\psi, \hat{\gamma}_\psi)$. Poiché il modello è una famiglia esponenziale e ψ è una componente del parametro canonico è inoltre immediato verificare che le derivate nello spazio campionario $\ell_{\alpha_i;\hat{\alpha}_{i'}}(\hat{\theta}_\psi)$ $\ell_{\alpha_i;\hat{\gamma}_j}(\hat{\theta}_\psi)$, e in modo

speculare $\ell_{\gamma_j;\hat{\gamma}_{j'}}(\hat{\theta}_\psi)\, \ell_{\gamma_j;\hat{\alpha}_i}(\hat{\theta}_\psi)$ dipendono solo dai dati. Dunque, data la struttura di famiglia esponenziale, la funzione di log-verosimiglianza profilo modificata è semplicemente esprimibile come

$$\ell_M(\psi) = \ell_P(\psi) + \frac{1}{2}\log|j_{\lambda\lambda}(\hat{\theta}_\psi)|.$$

3.3.2 Modello log-lineare Poisson con effetti fissi incrociati

Nello stesso scenario introdotto nel paragrafo 2.3.2, un modello statistico alternativo è quello che assume che y_{ij} siano realizzazioni di variabili casuali indipendenti Y_{ij} con distribuzione Poisson di media μ_{ij} dove $g(\mu_{ij}) = \eta_{ij}$, con $\eta_{ij} = \alpha_i + \gamma_j + \psi x_{ij}$, per $i = 1, \ldots, R$, $j = 1, \ldots, C$. Al solito, il vincolo di identificabilità può essere rispettato ponendo a zero uno degli α_i. Senza perdita di generalità, si è assunto di avere a disposizione un'unica covariata x e che il parametro di interesse ψ sia scalare. Invece, il parametro incidentale $\lambda = (\alpha, \gamma)$ ha dimensione pari ad $R + C$. Assumendo la funzione di legame canonica $g(\mu_{ij}) = \log \mu_{ij} = \eta_{ij}$, segue che

$$Y_{ij} \sim Po(e^{\eta_{ij}}),$$

$i = 1, \ldots, R$, $j = 1, \ldots, C$.

La funzione di verosimiglianza per (ψ, α, γ) basata sui dati y_{ij}, $i = 1, \ldots, R$, $j = 1, \ldots, C$, è

$$\begin{aligned} L(\psi, \alpha, \gamma) &= \prod_{i=1}^{R}\prod_{j=1}^{C} p(y_{ij}; \psi, \alpha_i, \gamma_j) \\ &= \exp\Big\{ -\sum_{i=1}^{R}\sum_{j=1}^{C} e^{\alpha_i+\gamma_j+\psi x_{ij}} + \sum_{i=1}^{R}\sum_{j=1}^{C}(\alpha_i + \gamma_j + \psi x_{ij})y_{ij}\Big\}. \end{aligned}$$

Ne segue che la funzione di log-verosimiglianza può essere scritta nel modo seguente

$$\begin{aligned} \ell(\psi, \alpha, \gamma) &= -\sum_{i=1}^{R}\sum_{j=1}^{C} e^{\alpha_i+\gamma_j+\psi x_{ij}} + \sum_{i=1}^{R}\sum_{j=1}^{C}(\alpha_i + \gamma_j + \psi x_{ij})y_{ij} \\ &= -\sum_{i=1}^{R} e^{\alpha_i}\sum_{j=1}^{C} e^{\gamma_j+\psi x_{ij}} + \sum_{i=1}^{R}\alpha_i\sum_{j=1}^{C} y_{ij} + \sum_{j=1}^{C}\gamma_j\sum_{i=1}^{R} y_{ij} + \psi\sum_{i=1}^{R}\sum_{j=1}^{C} x_{ij}y_{ij} \\ &= -\sum_{i=1}^{R} e^{\alpha_i}\sum_{j=1}^{C} e^{\gamma_j+\psi x_{ij}} + \sum_{i=1}^{R}\alpha_i y_{i+} + \sum_{j=1}^{C}\gamma_j y_{+j} + \psi s, \end{aligned}$$

dove $y_{i+} = \sum_{j=1}^{C} y_{ij}$ è il totale di riga i-esimo, $y_{+j} = \sum_{i=1}^{R} y_{ij}$ è il totale di colonna j-esimo e $s = \sum_{i=1}^{R}\sum_{j=1}^{C} x_{ij}y_{ij}$. Il contributo i-esimo alla funzione di log-verosimiglianza è quindi

$$\ell^i(\psi, \alpha_i, \gamma) = -e^{\alpha_i}\sum_{j=1}^{C} e^{\gamma_j+\psi x_{ij}} + \alpha_i y_{i+} + \sum_{j=1}^{C}\gamma_j y_{+j} + \psi s_i,$$

con $s_i = \sum_{j=1}^{C} x_{ij}y_{ij}$, $i = 1, \ldots, R$, che a differenza del caso di un solo effetto fisso, dipende non solo da α_i ma anche da γ. Derivando $\ell^i(\psi, \alpha_i, \gamma)$ rispetto a ciascun α_i si ottiene la funzione punteggio profilo per α_i

$$\ell_{\alpha_i}(\psi, \alpha_i, \gamma) = \frac{\partial}{\partial\alpha_i}\ell^i(\psi, \alpha_i, \gamma) = -e^{\alpha_i}\sum_{j=1}^{C} e^{\gamma_j+\psi x_{ij}} + y_{i+},$$

che, risolta rispetto ad α_i, permette di ottenere la stima di verosimiglianza di α_i per ψ e γ fissati, $\hat{\alpha}_{i\psi\gamma}$

$$\hat{\alpha}_{i\psi\gamma} = \log\Big\{\frac{y_{i+}}{\sum_{j=1}^{C} e^{\gamma_j+\psi x_{ij}}}\Big\}.$$

Dunque, a differenza del caso di un solo effetto fisso, ora essendo presenti sia α che γ, la stima vincolata di α_i per ψ fissato dipende anche da γ. La funzione di verosimiglianza profilo per ψ può essere ottenuta in due stadi. Innanzitutto, si sostituiscono le stime $\hat{\alpha}_{i\psi\gamma}$, $i = 1, \ldots, R$, al posto dei corrispondenti α_i nella funzione di log-verosimiglianza, in modo da ottenere una funzione che dipende solo da ψ e da γ

$$\begin{aligned}
\ell_P(\psi, \gamma) &= \ell(\psi, \hat{\alpha}_{1\psi\gamma}, \ldots, \hat{\alpha}_{R\psi\gamma}, \gamma_1, \ldots, \gamma_C) \\
&= -\sum_{i=1}^{R} y_{i+} + \sum_{i=1}^{R}\Big\{\log y_{i+} - \log\sum_{j=1}^{C} e^{\gamma_j+\psi x_{ij}}\Big\}y_{i+} + \sum_{j=1}^{C}\gamma_j y_{+j} + \psi s \\
&= c(y) + \psi s + \sum_{j=1}^{C}\gamma_j y_{+j} - \sum_{i=1}^{R} y_{i+}\log\sum_{j=1}^{C} e^{\gamma_j+\psi x_{ij}}.
\end{aligned}$$

Il secondo stadio prevede di derivare $\ell_P(\psi, \gamma)$ rispetto a γ_j in modo da ottenere la stima vincolata di γ_j per ψ fissato. La funzione punteggio profilo per γ_j con ψ fissato risulta pari a

$$\frac{\partial}{\partial\gamma_j}\ell_P(\psi, \gamma) = y_{+j} - \sum_{i=1}^{R} y_{i+}\frac{e^{\gamma_j+\psi x_{ij}}}{\sum_{h=1}^{C} e^{\gamma_h+\psi x_{ih}}},$$

la cui corrispondente equazione di stima

$$y_{+j} - \sum_{i=1}^{R} y_{i+} \frac{e^{\gamma_j + \psi x_{ij}}}{\sum_{h=1}^{C} e^{\gamma_h + \psi x_{ih}}} = 0, \tag{3.3}$$

non ammette una soluzione esplicita per γ_j dato ψ. Pertanto la stima $\hat{\gamma}_{j,\psi}$ rimane implicitamente definita dall'equazione (3.3). Ne segue che la funzione di log-verosimiglianza profilo per ψ deve essere scritta lasciando indicate le stime implicite $\hat{\gamma}_{j,\psi}$

$$\begin{aligned}\ell_P(\psi) &= \ell(\psi, \hat{\alpha}_{1\psi\hat{\gamma}_\psi}, \ldots, \hat{\alpha}_{R\psi\hat{\gamma}_\psi}, \hat{\gamma}_{1,\psi}, \ldots, \hat{\gamma}_{C,\psi}) \\ &= c(y) + \psi s + \sum_{j=1}^{C} \hat{\gamma}_{j,\psi} y_{+j} - \sum_{i=1}^{R} y_{i+} \log \sum_{j=1}^{C} e^{\hat{\gamma}_{j,\psi} + \psi x_{ij}}.\end{aligned}$$

Anche nel calcolo della log-verosimiglianza condizionata non è possibile ottenere le stesse semplificazioni come nel caso di un unico effetto fisso nei modelli stratificati, e quindi non è possibile mostrare allo stesso modo l'eventuale equivalenza tra la log-verosimiglianza profilo e la log-verosimiglianza profilo condizionata.

Tuttavia, le statistiche $y_{i+} = \sum_{j=1}^{C} y_{ij}$ e $y_{+j} = \sum_{i=1}^{R} y_{ij}$, $i = 1, \ldots, R$, $j = 1, \ldots, C$, rispettivamente il totale di riga i-esimo e di colonna j-esimo, sono parzialmente sufficienti per α e γ, data la struttura di famiglia esponenziale. Quindi, la distribuzione congiunta delle y_{ij} condizionata a (y_{i+}, y_{+j}) non dipende da α e γ.

Ad ogni modo, non è difficile ottenere la versione modificata della verosimiglianza profilo e derivare le approssimazioni di ordine superiore. Infatti, si noti che le funzioni punteggio profilo possono essere scritte come

$$\begin{aligned}\ell_\psi &= \frac{\partial}{\partial \psi} \ell(\psi, \alpha, \gamma) = s - \sum_{i=1}^{R} \sum_{j=1}^{C} x_{ij} e^{\alpha_i + \gamma_j + \psi_j x_{ij}} = s - \sum_{i=1}^{R} \sum_{j=1}^{C} x_{ij} \mu_{ij}, \\ \ell_{\alpha_i} &= \frac{\partial}{\partial \alpha_i} \ell(\psi, \alpha, \gamma) = y_{i+} - \sum_{j=1}^{C} e^{\alpha_i + \gamma_j + \psi_j x_{ij}} = y_{i+} - \sum_{j=1}^{C} \mu_{ij} = y_{i+} - \mu_{i+}, \\ \ell_{\gamma_j} &= \frac{\partial}{\partial \gamma_j} \ell(\psi, \alpha, \gamma) = y_{+j} - \sum_{i=1}^{R} e^{\alpha_i + \gamma_j + \psi_j x_{ij}} = y_{+j} - \sum_{i=1}^{R} \mu_{ij} = y_{+j} - \mu_{+j}.\end{aligned}$$

dove $\mu_{i+} = \sum_{j=1}^{C} \mu_{ij}$ e $\mu_{+j} = \sum_{i=1}^{R} \mu_{ij}$. Le derivate seconde della funzione di log-verosimiglianza risultano quindi pari a

$$\ell_{\psi\psi} = \frac{\partial^2}{\partial\psi\partial\psi}\ell(\psi,\alpha,\gamma) = -\sum_{i=1}^{R}\sum_{j=1}^{C} x_{ij}^2\mu_{ij},$$

$$\ell_{\psi\alpha_i} = \frac{\partial^2}{\partial\psi\partial\alpha_i}\ell(\psi,\alpha,\gamma) = -\sum_{j=1}^{C} x_{ij}\mu_{ij},$$

$$\ell_{\psi\gamma_j} = \frac{\partial^2}{\partial\psi\partial\gamma_j}\ell(\psi,\alpha,\gamma) = -\sum_{i=1}^{R} x_{ij}\mu_{ij},$$

$$\ell_{\alpha_i\alpha_{i'}} = \frac{\partial^2}{\partial\alpha_i\partial\alpha_{i'}}\ell(\psi,\alpha,\gamma) = \begin{cases} -\sum_{j=1}^{C}\mu_{ij} = -\mu_{i+} & i = i' \\ 0 & i \neq i' \end{cases},$$

$$\ell_{\gamma_j\gamma_{j'}} = \frac{\partial^2}{\partial\gamma_j\partial\gamma_{j'}}\ell(\psi,\alpha,\gamma) = \begin{cases} -\sum_{i=1}^{R}\mu_{ij} = -\mu_{+j} & j = j' \\ 0 & j \neq j' \end{cases},$$

$$\ell_{\alpha_i\gamma_j} = \frac{\partial^2}{\partial\alpha_i\partial\gamma_j}\ell(\psi,\alpha,\gamma) = -\mu_{ij},$$

per $i = 1,\ldots,R$, $j = 1,\ldots,C$. Cambiando il segno di tali derivate seconde, si ottengono gli elementi dell'informazione osservata $j(\theta) = j(\psi,\alpha,\gamma)$, che coincide con l'informazione attesa, $i(\theta) = j(\theta)$, poiché il modello è una famiglia esponenziale in parametrizzazione canonica. Il blocco di $j(\theta)$ relativo ai parametri di disturbo è quindi

$$j_{\lambda\lambda}(\theta) = \left[\begin{array}{c|c} j_{\alpha\alpha} & j_{\alpha\gamma}^T \\ \hline j_{\alpha\gamma} & j_{\gamma\gamma} \end{array}\right] = \begin{bmatrix} \mu_{i+}I_{R-1} & \mu_{ij}1_{R-1}1_C^T \\ \mu_{ij}1_C1_{R-1}^T & \mu_{+j}I_C \end{bmatrix},$$

che risulta essere una matrice di dimensione $(R-1)+C\times(R-1)+C$, in quanto α ha $R-1$ elementi. Calcolando $j(\theta)$ in corrispondenza della stima vincolata $\hat{\theta}_\psi = (\psi,\hat{\alpha}_\psi,\hat{\gamma}_\psi)$, è immediato ottenere il determinante di $j_{\lambda\lambda}(\hat{\theta}_\psi) = j_{\lambda\lambda}(\psi,\hat{\alpha}_\psi,\hat{\gamma}_\psi)$. Dunque, data la struttura di famiglia esponenziale, come descritto più in generale nel paragrafo 3.3, è semplice ottenere la funzione di log-verosimiglianza profilo modificata e la versione modificata $r_P^*(\psi)$ dell'usuale statistica test radice con segno del log-rapporto di verosimiglianza profilo $r_P(\psi)$.

Inoltre, poiché $W = \text{diag}\{\mu_{ij}\}$ nel caso del modello log-lineare Poisson, utilizzando la notazione con le matrici indicatrici Z_1 e Z_2, introdotta nel paragrafo 3.3, il blocco

dell'informazione osservata relativo ai parametri di disturbo può essere scritto come

$$j_{\lambda\lambda}(\theta) = [Z_1, Z_2]^T \operatorname{diag}\{\mu_{ij}\}[Z_1, Z_2].$$

3.3.3 Modello logistico con effetti fissi incrociati

Nello stesso scenario introdotto nel paragrafo 2.3.3, un modello statistico alternativo è quello che assume che y_{ij} siano realizzazioni di variabili casuali indipendenti Y_{ij} con distribuzione bernoulliana con probabilità di successo π_{ij}, $i = 1, \ldots, R$, $j = 1, \ldots, C$. La densità di Y_{ij} per questo modello è

$$p(y_{ij}; \psi, \alpha_i, \gamma_j) = {\pi_{ij}}^{y_{ij}}(1 - \pi_{ij})^{1-y_{ij}},$$

dove $g(\pi_{ij}) = \eta_{ij}$, con $\eta_{ij} = \alpha_i + \gamma_j + \psi x_{ij}$, per $i = 1, \ldots, R$, $j = 1, \ldots, C$. Al solito, il vincolo di identificabilità può essere rispettato ponendo a zero uno degli α_i. Senza perdita di generalità si è assunto di avere a disposizione un'unica covariata x e che il parametro di interesse ψ sia scalare. Invece, il parametro incidentale $\lambda = (\alpha, \gamma)$ ha dimensione pari ad $R + C$. Assumendo la funzione di legame canonica $g(\pi_{ij}) = \text{logit}(\pi_{ij}) = \eta_{ij}$, segue che

$$Y_{ij} \sim Bi\Big(1, \frac{e^{\eta_{ij}}}{1 + e^{\eta_{ij}}}\Big),$$

$i = 1, \ldots, R$, $j = 1, \ldots, C$.

La funzione di verosimiglianza per (ψ, α, γ) basata sui dati y_{ij}, $i = 1, \ldots, R$, $j = 1, \ldots, C$, è

$$\begin{aligned}
L(\psi, \alpha, \gamma) &= \prod_{i=1}^{R}\prod_{j=1}^{C} p(y_{ij}; \psi, \alpha_i, \gamma_j) \\
&= \prod_{i=1}^{R}\prod_{j=1}^{C} {\pi_{ij}}^{y_{ij}}(1 - \pi_{ij})^{1-y_{ij}} \\
&= \frac{\exp\Big\{\sum_{i=1}^{R}\sum_{j=1}^{C}(\alpha_i + \gamma_j + \psi x_{ij})y_{ij}\Big\}}{\prod_{i=1}^{R}\prod_{j=1}^{C}(1 + e^{\alpha_i + \gamma_j + \psi x_{ij}})} \\
&= \frac{\exp\Big\{\sum_{i=1}^{R}\sum_{j=1}^{C}\alpha_i y_{ij} + \sum_{i=1}^{R}\sum_{j=1}^{C}\gamma_j y_{ij} + \psi\sum_{i=1}^{R}\sum_{j=1}^{C} x_{ij} y_{ij}\Big\}}{\prod_{i=1}^{R}\prod_{j=1}^{C}(1 + e^{\alpha_i + \gamma_j + \psi x_{ij}})}.
\end{aligned}$$

Ne segue che la funzione di log-verosimiglianza può essere scritta nel modo seguente

$$\begin{aligned}\ell(\psi,\alpha,\gamma) &= \sum_{i=1}^{R}\sum_{j=1}^{C}\alpha_i y_{ij} + \sum_{i=1}^{R}\sum_{j=1}^{C}\gamma_j y_{ij} + \psi\sum_{i=1}^{R}\sum_{j=1}^{C}x_{ij}y_{ij} - \sum_{i=1}^{R}\sum_{j=1}^{C}\log\left(1+e^{\alpha_i+\gamma_j+\psi x_{ij}}\right) = \\ &= \sum_{i=1}^{R}\alpha_i\sum_{j=1}^{C}y_{ij} + \sum_{j=1}^{C}\gamma_j\sum_{i=1}^{R}y_{ij} + \psi\sum_{i=1}^{R}\sum_{j=1}^{C}x_{ij}y_{ij} - \sum_{i=1}^{R}\sum_{j=1}^{C}\log\left(1+e^{\alpha_i+\gamma_j+\psi x_{ij}}\right) = \\ &= \sum_{i=1}^{R}\alpha_i y_{i+} + \sum_{j=1}^{C}\gamma_j y_{+j} + \psi s - \sum_{i=1}^{R}\sum_{j=1}^{C}\log\left(1+e^{\alpha_i+\gamma_j+\psi x_{ij}}\right),\end{aligned}$$

dove $y_{i+} = \sum_{j=1}^{C} y_{ij}$, $y_{+j} = \sum_{i=1}^{R} y_{ij}$ e $s = \sum_{i=1}^{R}\sum_{j=1}^{C} x_{ij}y_{ij}$. Dunque, le componenti della funzione punteggio risultano pari a

$$\begin{aligned}\ell_\psi &= \frac{\partial}{\partial\psi}\ell(\psi,\alpha,\gamma) = s - \sum_{i=1}^{R}\sum_{j=1}^{C}x_{ij}\frac{e^{\alpha_i+\gamma_j+\psi x_{ij}}}{1+e^{\alpha_i+\gamma_j+\psi x_{ij}}} = s - \sum_{i=1}^{R}\sum_{j=1}^{C}x_{ij}\mu_{ij}, \\ \ell_{\alpha_i} &= \frac{\partial}{\partial\alpha_i}\ell(\psi,\alpha,\gamma) = y_{i+} - \sum_{j=1}^{C}\frac{e^{\alpha_i+\gamma_j+\psi x_{ij}}}{1+e^{\alpha_i+\gamma_j+\psi x_{ij}}} = y_{i+} - \sum_{j=1}^{C}\mu_{ij} = y_{i+} - \mu_{i+}, \\ \ell_{\gamma_j} &= \frac{\partial}{\partial\gamma_j}\ell(\psi,\alpha,\gamma) = y_{+j} - \sum_{i=1}^{R}\frac{e^{\alpha_i+\gamma_j+\psi x_{ij}}}{1+e^{\alpha_i+\gamma_j+\psi x_{ij}}} = y_{+j} - \sum_{i=1}^{R}\mu_{ij} = y_{+j} - \mu_{+j}.\end{aligned}$$

dove $\mu_{i+} = \sum_{j=1}^{C}\mu_{ij}$ e $\mu_{+j} = \sum_{i=1}^{R}\mu_{ij}$. Osservando che

$$\frac{e^{\alpha_i+\gamma_j+\psi_j x_{ij}}}{(1+e^{\alpha_i+\gamma_j+\psi x_{ij}})^2} = \mu_{ij}(1-\mu_{ij}),$$

le derivate seconde della funzione di log-verosimiglianza risultano quindi pari a

$$\begin{aligned}\ell_{\psi\psi} &= \frac{\partial^2}{\partial\psi\partial\psi}\ell(\psi,\alpha,\gamma) = -\sum_{i=1}^{R}\sum_{j=1}^{C}x_{ij}^2\mu_{ij}(1-\mu_{ij}), \\ \ell_{\psi\alpha_i} &= \frac{\partial^2}{\partial\psi\partial\alpha_i}\ell(\psi,\alpha,\gamma) = -\sum_{j=1}^{C}x_{ij}\mu_{ij}(1-\mu_{ij}), \\ \ell_{\psi\gamma_j} &= \frac{\partial^2}{\partial\psi\partial\gamma_j}\ell(\psi,\alpha,\gamma) = -\sum_{i=1}^{R}x_{ij}\mu_{ij}(1-\mu_{ij}),\end{aligned}$$

$$\ell_{\alpha_i\alpha_{i'}} = \frac{\partial^2}{\partial\alpha_i\partial\alpha_{i'}}\ell(\psi,\alpha,\gamma) = \begin{cases} -\sum_{j=1}^{C}\mu_{ij}(1-\mu_{ij}) & i = i' \\ 0 & i \neq i' \end{cases},$$

$$\ell_{\gamma_j\gamma_{j'}} = \frac{\partial^2}{\partial\gamma_j\partial\gamma_{j'}}\ell(\psi,\alpha,\gamma) = \begin{cases} -\sum_{i=1}^{R}\mu_{ij}(1-\mu_{ij}) & j = j' \\ 0 & j \neq j' \end{cases},$$

$$\ell_{\alpha_i\gamma_j} = \frac{\partial^2}{\partial\alpha_i\partial\gamma_j}\ell(\psi,\alpha,\gamma) = -\mu_{ij}(1-\mu_{ij}),$$

per $i = 1,\ldots,R$, $j = 1,\ldots,C$. Cambiando il segno di tali derivate seconde, si ottengono gli elementi dell'informazione osservata $j(\theta) = j(\psi,\alpha,\gamma)$, che coincide con l'informazione attesa, $i(\theta) = j(\theta)$, poiché il modello è una famiglia esponenziale in parametrizzazione canonica. Il blocco di $j(\theta)$ relativo ai parametri di disturbo è quindi

$$j_{\lambda\lambda}(\theta) = \left[\begin{array}{c|c} j_{\alpha\alpha} & j_{\alpha\gamma}^T \\ \hline j_{\alpha\gamma} & j_{\gamma\gamma} \end{array}\right],$$

che risulta essere una matrice di dimensione $(R-1)+C\times(R-1)+C$. Calcolando $j(\theta)$ in corrispondenza della stima vincolata $\hat\theta_\psi = (\psi,\hat\alpha_\psi,\hat\gamma_\psi)$, è immediato ottenere il determinante di $j_{\lambda\lambda}(\hat\theta_\psi) = j_{\lambda\lambda}(\psi,\hat\alpha_\psi,\hat\gamma_\psi)$. Dunque, data la struttura di famiglia esponenziale, come descritto più in generale nel paragrafo 3.3, è semplice ottenere la funzione di log-verosimiglianza profilo modificata e la versione modificata $r_P^*(\psi)$ dell'usuale statistica test radice con segno del log-rapporto di verosimiglianza profilo $r_P(\psi)$.

Inoltre, poiché $W = \text{diag}\{\mu_{ij}(1-\mu_{ij})\}$ nel caso del modello logistico con funzione di legame canonica, utilizzando la notazione con le matrici indicatrici Z_1 e Z_2, introdotta nel paragrafo 3.3, il blocco dell'informazione osservata relativo ai parametri di disturbo può essere scritto come

$$j_{\lambda\lambda}(\theta) = [Z_1, Z_2]^T \,\text{diag}\,\{\mu_{ij}(1-\mu_{ij})\}[Z_1, Z_2].$$

Capitolo 4

Studi di simulazione

4.1 Introduzione

In questo capitolo, la teoria presentata in precedenza sarà esaminata mediante studi di simulazione. Le simulazioni sono ampiamente utilizzate in Statistica per valutare le proprietà delle procedure e delle metodologie adottate, soprattutto quando sono coinvolte approssimazioni difficili da comprendere pienamente con uno studio puramente analitico. Inoltre, tecniche come il *bootstrap* sono intrinsecamente basate sul concetto di simulazione, in quanto coinvolgono metodi di ricampionamento da un insieme di dati osservato per generare un numero elevato di campioni simulati.

I modelli considerati nelle simulazioni sono modelli con effetti fissi incrociati e dati discreti sparsi. L'attenzione verrà rivolta ai modelli lineari generalizzati per dati binari e per dati di conteggio, in particolare si considereranno rispettivamente il modello log-lineare Poisson, discusso nel paragrafo 3.3.2, e il modello di regressione logistica, discusso nel paragrafo 3.3.3.

Il parametro complessivo $\theta = (\psi, \lambda) = (\psi, \alpha, \gamma)$ è composto dal parametro di interesse ψ di dimensione k che nel seguito, senza perdita di generalità, verrà assunto essere scalare, $k = 1$, associato ad un unica covariata x, mentre il parametro di disturbo $\lambda = (\alpha, \gamma)$ è composto da due effetti fissi incrociati, con $\alpha = (\alpha_1, \ldots, \alpha_R)$ di dimensione R, e $\gamma = (\gamma_1, \ldots, \gamma_C)$ di dimensione C. Dunque, nel seguito θ è un parametro di dimensione $R + C + 1$ o, più precisamente, di dimensione $R + C$, se si considera il vincolo di identificabilità sugli α_i o sui γ_j.

4.2 Struttura delle simulazioni

Nelle simulazioni successive si è ipotizzata la presenza di sparsità nei dati, e di conseguenza nella struttura degli effetti fissi incrociati, in quanto ciò è coerente con le caratteristiche dei dati per cui i modelli considerati possono essere impiegati nelle applicazioni. Gli aspetti relativi alle assunzioni di sparsità vengono descritti nel dettaglio nel paragrafo successivo.

Per le simulazioni sono stati considerati diversi scenari per i valori di R ed C, ovvero per le dimensioni dei parametri di disturbo α e γ, tenendo conto anche della numerosità campionaria N, e del livello di sparsità desiderato. Poiché la procedura utilizzata per introdurre sparsità nei dati è caratterizzata da una numerosità campionaria N casuale, è stato necessario specificare la numerosità campionaria attesa desiderata $\mathbb{E}(N)$. Nello specifico, indicando con κ il rapporto nominale tra il numero di parametri di disturbo e la numerosità campionaria, $k = (R + C)/\mathbb{E}(N)$, per entrambi i modelli considerati, si sono considerati i seguenti 4 scenari

Scenari di simulazione 4.S:

(**S.1**) $R = 20$, $C = 20$, $\mathbb{E}(N) = 200$, $\kappa = 1/5$,

(**S.2**) $R = 20$, $C = 20$, $\mathbb{E}(N) = 120$, $\kappa = 1/3$,

(**S.3**) $R = 30$, $C = 30$, $\mathbb{E}(N) = 180$, $\kappa = 1/3$,

(**S.4**) $R = 50$, $C = 50$, $\mathbb{E}(N) = 300$, $\kappa = 1/3$.

Chiaramente, poiché la numerosità campionaria N effettiva è casuale, anche il rapporto effettivo tra numero di parametri di disturbo e numerosità campionaria risulta essere casuale. Pertanto, nel seguito, $\tilde{\kappa} = (R + C)/N$ denota il rapporto effettivo tra numero di parametri di disturbo e la numerosità campionaria effettiva. Come si può notare, ad eccezione dello scenario (S.1) in cui il rapporto nominale κ è pari ad 1/5, nei rimanenti casi è stato fissato ad 1/3. Dunque, gli scenari (S.2)-(S.4) permettono di studiare empiricamente le proprietà delle procedure inferenziali considerate quando κ è fissato ma la numerosità campionaria N aumenta. Invece, il confronto tra scenario (S.1) e scenario (S.2), dove in entrambi i casi $R = 20$ e $C = 20$, ma nello scenario (S.2) la numerosità campionaria N è inferiore, permette di capire cosa succede a parità di dimensione degli effetti fissi se si diminuisce l'informazione a disposizione. In particolare, è naturale aspettarsi che le procedure inferenziali godano di proprietà migliori nello scenario (S.1) rispetto allo scenario (S.2).

Gli effetti fissi incrociati α e γ sono stati generati da una distribuzione normale standard e considerati fissati in tutti i campioni generati. Poiché i modelli considerati sono modelli di regressione, senza perdita di generalità, si è assunto di disporre di un'unica covariata x, generata anch'essa da una distribuzione normale standard e considerata nota. Il parametro di interesse ψ associato alla covariata x, è stato fissato pari ad 1, $\psi = 1$, in tutti i casi considerati.

I risultati empirici vengono presentati in modo simile agli studi di simulazione in Bartolucci et al. (2016) e Bellio et al. (2023b). Per quanto riguarda il confronto tra verosimiglianza profilo e verosimiglianza profilo modificata, sulla base dei risultati delle simulazioni vengono riportate per gli stimatori d'interesse le seguenti quantità:

- la distorsione media (*bias*), ovvero la differenza tra la media delle stime ottenute nelle singole simulazioni e il vero valore del parametro;
- la probabilità di sottostima (*Probability of Underestimation*, PU), cioè la frequenza con cui le stime ottenute nelle simulazioni sono inferiori al vero valore del parametro d'interesse;
- la deviazione standard (*Standard Deviation*, SD), ovvero una stima della variabilità dello stimatore;
- l'errore standard (*Standard Error*, SE), cioè una media degli *standard error* stimati nelle simulazioni, utilizzando la varianza asintotica;
- la radice quadrata dell'errore quadratico medio (*Root Mean Squared Error*, RMSE) che rappresenta una misura di accuratezza dello stimatore;
- il rapporto SE/SD, come indicatore dell'accuratezza media degli standard error nello stimare la variabilità dello stimatore.

Alla luce della teoria presentata nei capitoli precedenti ci si aspetta che le stimatore del parametro d'interesse ψ ottenuto dalla massimizzazione della verosimiglianza profilo modificata goda di proprietà migliori rispetto allo stimatore ricavato dall'usuale massimizzazione della verosimiglianza profilo.

In aggiunta, per ognuno degli esperimenti di simulazione, sotto il modello considerato con il parametro $\theta = (\psi, \alpha, \gamma)$ fissato, si sono calcolate 8 statistiche test per la verifica di ipotesi sul parametro scalare ψ, considerando l'ipotesi nulla $H_0 : \psi = 1$ contro l'alternativa unilaterale sinistra $H_1 : \psi < 1$, e i corrispettivi 8 *p-value*. Tra queste 8 statistiche test (e corrispondentemente tra gli 8 *p-value*), 5 sono state calcolate tramite l'approccio del *bootstrap* parametrico. In particolare, il *bootstrap* permette di

calcolare il *p-value* associato alla verifica dell'ipotesi nulla H_0, e a partire da questo è possibile ottenere la corrispettiva statistica test, utilizzando la funzione inversa della funzione di ripartizione di una normale standard. Da un punto di vista teorico, ci si aspetta che le statistiche test dovrebbero avere una distribuzione approssimativamente vicina a quella di una normale standard, mentre i *p-value* dovrebbero avere una distribuzione approssimativamente uniforme nell'intervallo $(0, 1)$. Per valutare la vicinanza alla distribuzione di riferimento, normale standard, delle statistiche test calcolate nei diversi esperimenti di simulazione, verranno presentate delle tabelle relative alle coperture empiriche effettive delle statistiche test. I valori delle coperture empiriche ottenuti a partire dalle simulazioni possono essere confrontati con gli usuali livelli percentuali nominali $(1.0, 2.5, 5.0, 95.0, 97.7, 99.0)$, concentrandosi sul comportamento sulla coda destra e sinistra della distribuzione empirica.

La Tabella 4.1 riporta una una breve legenda dei simboli che verranno utilizzati per presentare i risultati delle simulazioni e una breve spiegazione circa il loro significato.

Statistica	Simbolo	Descrizione
$r_P(\psi)$	r_P	Radice con segno del log-rapporto di verosimiglianza profilo (1.11)
$r_M(\psi)$	r_M	Radice con segno del log-rapporto di verosimiglianza profilo modificata (1.16)
$r_P^*(\psi)$	r^*	Statistica basata sulla modifica della statistica r_P basata sulla formula p^* (1.18)
$\Phi^{-1}(\hat{\alpha}_u^{oss}(\psi))$	u. boot r_P	Statistica r_P ottenuta tramite *unconstrained bootstrap* (1.19)
$\Phi^{-1}(\hat{\alpha}_c^{oss}(\psi))$	c. boot r_P	Statistica r_P ottenuta tramite *constrained bootstrap* (1.20)
$\Phi^{-1}(\hat{\alpha}_u^{oss}(\psi)^\dagger)$	u. penalized $[a = 0.5]$ boot r_P	Statistica r_P ottenuta tramite *unconstrained bootstrap* con stime ricavate dalla verosimiglianza penalizzata (2.15), con $a = 0.5$
$\Phi^{-1}(\hat{\alpha}_c^{oss}(\psi)^\dagger)$	c. penalized $[a = 0.5]$ boot r_P	Statistica r_P ottenuta tramite *constrained bootstrap* con stime ricavate dalla verosimiglianza penalizzata (2.15), con $a = 0.5$
$\Phi^{-1}(\hat{\alpha}_c^{oss}(\psi)^\dagger)$	c. penalized $[a = 1]$ boot r_P	Statistica r_P ottenuta tramite *constrained bootstrap* con stime ricavate dalla verosimiglianza penalizzata (2.15), con $a = 1$

Tabella 4.1: Statistiche test confrontate nelle simulazioni.

In questa tesi, per gli studi di simulazione, tra i vari problemi inferenziali brevemente richiamati nel Capitolo 1, si è scelto di porre maggiore enfasi sul problema di verifica di ipotesi sul parametro di interesse ψ. Ad ogni modo, ci si aspetta che analoghe conclusioni possano essere tratte se si considerano le probabilità di copertura degli intervalli di confidenza basati sulle usuali quantità pivotali di verosimiglianza e le loro modificazioni.

Nelle simulazioni condotte, gli scenari considerati in termini di R ed C, le dimensioni dei parametri di disturbo, sono caratterizzati da un numero complessivo di parametri molto elevato. Di conseguenza, il calcolo basato sul *bootstrap* dei *p-value* e delle corrispondenti statistiche test risulta particolarmente oneroso in termini di tempo e risorse computazionali, anche sfruttando i vantaggi del calcolo parallelo su molteplici *cores*. Pertanto, il numero di campioni *bootstrap* in tutte le Nsim = 1000 simulazioni condotte è stato limitato a $B = 1000$. Ciò significa che, dato un modello, fissati i parametri e le matrici sparse Z_1 e Z_2, si sono generati Nsim = 1000 campioni, e per ognuno di questi campioni, per le procedure legate al *bootstrap*, è stato necessario simulare altri $B = 1000$ campioni. Poiché sono stai considerati 4 possibili scenari in termini di R ed C, e 5 statistiche *bootstrap*, in totale sono stati generati 20 milioni di campioni per ciascun modello per i calcoli basati sul *bootstrap*.

4.3 Sparsità negli effetti fissi incrociati

In questo paragrafo si discutono alcuni possibili approcci per introdurre sparsità nei dati, e conseguentemente, nella struttura degli effetti fissi incrociati $\lambda = (\alpha, \gamma) = (\alpha_1, \ldots, \alpha_R, \gamma_1, \ldots, \gamma_C)$, che deve rispettare la struttura dei dati a disposizione.

Si riconsideri la notazione introdotta nel paragrafo 3.3. Sia X la matrice del disegno e $\theta = (\psi, \alpha, \gamma)$ il vettore dei coefficienti di regressione, il predittore lineare $\eta = X\theta$ può essere scritto come

$$\eta = X\theta = \begin{bmatrix} x & Z_1 & Z_2 \end{bmatrix} \begin{bmatrix} \psi \\ \alpha \\ \gamma \end{bmatrix} = \psi x + Z_1\alpha + Z_2\gamma, \tag{4.1}$$

dove la matrice X può essere partizionata come $X = [x, Z_1, Z_2]$, dove in generale x è un vettore di lunghezza N, Z_1 è una matrice di dimensione $N \times R$ e Z_2 è una matrice di dimensione $N \times C$, dove $N = RC$. Le matrici Z_1 e Z_2, indicatrici degli effetti fissi

incrociati, possono essere definite come

$$\underset{N\times R}{Z_1} = I_R \otimes 1_C, \qquad \underset{N\times C}{Z_2} = [1_R^T \otimes I_C]^T = 1_R \otimes I_C.$$

Si consideri, ad esempio, il caso $R = 3$ ed $C = 4$, allora

$$\underset{12\times 3}{Z_1} = I_3 \otimes 1_4 = \left[\begin{array}{ccc} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ \hline 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{array}\right], \qquad \underset{12\times 4}{Z_2} = 1_3 \otimes I_4 = \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \hline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \hline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right].$$

Di conseguenza

$$Z_1\alpha = \left[\begin{array}{c} \alpha_1 \\ \alpha_1 \\ \alpha_1 \\ \alpha_1 \\ \hline \alpha_2 \\ \alpha_2 \\ \alpha_2 \\ \alpha_2 \\ \hline \alpha_3 \\ \alpha_3 \\ \alpha_3 \\ \alpha_3 \end{array}\right], \qquad Z_2\gamma = \left[\begin{array}{c} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \\ \hline \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \\ \hline \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{array}\right], \qquad \eta = \psi x + Z_1\alpha + Z_2\gamma = \left[\begin{array}{c} \psi x_1 + \alpha_1 + \gamma_1 \\ \psi x_2 + \alpha_1 + \gamma_2 \\ \psi x_3 + \alpha_1 + \gamma_3 \\ \psi x_4 + \alpha_1 + \gamma_4 \\ \hline \psi x_5 + \alpha_2 + \gamma_1 \\ \psi x_6 + \alpha_2 + \gamma_2 \\ \psi x_7 + \alpha_2 + \gamma_3 \\ \psi x_8 + \alpha_2 + \gamma_4 \\ \hline \psi x_9 + \alpha_3 + \gamma_1 \\ \psi x_{10} + \alpha_3 + \gamma_2 \\ \psi x_{11} + \alpha_3 + \gamma_3 \\ \psi x_{12} + \alpha_3 + \gamma_4 \end{array}\right].$$

In questo esempio, $R = 3$ è la dimensione dell'effetto fisso $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, $C = 4$ è la dimensione dell'effetto fisso $\gamma = (\gamma_1, \gamma_2, \gamma_3, \gamma_4)$. Ciascun α_i, $i = 1, 2, 3$, appare nel predittore η un numero di volte pari ad $C = 4$, mentre ciascun γ_j, $j = 1, 2, 3, 4$, appare nel predittore η un numero di volte pari ad $R = 3$.

Un modo per introdurre sparsità nella struttura degli effetti fissi incrociati (α, γ) è quello di limitare il numero di occorrenze di ciascun α_i e γ_j. Ad esempio, una possibilità è la seguente: date le matrici Z_1 e Z_2, si costruisce una matrice T di dimensione $R \times C$ con celle che assumo valori da 1 a RC, disposti per riga, ossia

$$T_{ij} = (i-1)C + j, \qquad i = 1, \ldots, R,\ j = 1, \ldots, C.$$

Si fissa il numero c, $c \leq C$ di occorrenze di ciascun α_i, si crea un vettore filtro f selezionando casualmente c elementi in ciascuna riga della matrice T e ordinandoli. Dunque il vettore f ha lunghezza Rc, in quanto per ciascuna riga di T si scelgono c elementi, e rappresenta la concatenazione degli indici selezionati in ciascuna riga in un unico vettore. Infine, si riduce il numero di righe della matrice Z_1, applicando il filtro f, ossia selezionando solo le righe di Z_1 specificate da f, creando una nuova matrice $\tilde{Z}_1 = Z_1[f, \cdot]$, in cui solo alcune delle righe di Z_1 saranno conservate. Si ripete la stessa riduzione per la matrice Z_2, ottenendo la matrice $\tilde{Z}_2 = Z_2[f, \cdot]$. Dunque, le matrici $\tilde{Z}_1$ e $\tilde{Z}_2$ hanno dimensione $Rc \times R$ e $Rc \times C$, rispettivamente, in quanto $\tilde{N} = Rc$ è la nuova numerosità campionaria.

Il nuovo predittore lineare è quindi $\tilde{\eta} = \psi\tilde{x} + \tilde{Z}_1\alpha + \tilde{Z}_2\gamma$. Questo approccio permette pertanto di controllare in modo omogeneo il numero di occorrenze di ciascun α_i, $i = 1, \ldots, R$, che risulterà esattamente pari a c. Al contrario, non garantisce un controllo omogeneo per quanto riguarda il numero di occorrenze di ciascun γ_j, $j = 1, \ldots, C$, che sarà però inferiore rispetto ad R, con la condizione che ciascun γ_j deve essere osservato almeno una volta. Ad ogni modo, non sembra rilevante per le analisi che sono condotte avere lo stesso numero di occorrenze per ciascun γ_j. Con la struttura sparsa, il numero totale di osservazioni in questo caso risulta pari a $\tilde{N} = Rc$, anziché $N = RC$, come si ha nel caso di una struttura piena per gli effetti fissi incrociati, dove $c \leq C$. Ne consegue che si passa dalla situazione in cui vi sono N osservazioni e $R+C$ parametri di disturbo, alla situazione in cui vi sono solo $\tilde{N}$ osservazioni per lo stesso numero di parametri.

Nell'esempio precedentemente introdotto, con $R = 3$ ed $C = 4$, il numero totale di osservazioni è $N = RC = 12$ e la dimensione del parametro di disturbo è $R + C = 7$, il numero di occorrenze di ciascun α_i è pari a 4, il numero di occorrenze di ciascun γ_j è pari a 3. Introducendo sparsità, scegliendo ad esempio $c = 3$, il nuovo numero di osservazioni risulta pari a $\tilde{N} = Rc = 9$, il numero di occorrenze di ciascun α_i è pari a $c = 3$, mentre il numero di occorrenze di ciascun γ_j sarà, seppur in modo non uniforme,

inferiore o uguale a 3. Le matrici $\tilde{Z}_1$ e $\tilde{Z}_2$ hanno infatti una struttura del tipo

$$\underset{9\times 3}{\tilde{Z}_1} = \left[\begin{array}{ccc} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ \hline 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ \hline 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{array}\right], \qquad \underset{9\times 4}{\tilde{Z}_2} = \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \hline 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \hline 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right],$$

e di conseguenza

$$\tilde{Z}_1\alpha = \left[\begin{array}{c} \alpha_1 \\ \alpha_1 \\ \alpha_1 \\ \hline \alpha_2 \\ \alpha_2 \\ \alpha_2 \\ \hline \alpha_3 \\ \alpha_3 \\ \alpha_3 \end{array}\right], \qquad \tilde{Z}_2\gamma = \left[\begin{array}{c} \gamma_1 \\ \gamma_3 \\ \gamma_4 \\ \hline \gamma_2 \\ \gamma_3 \\ \gamma_4 \\ \hline \gamma_1 \\ \gamma_2 \\ \gamma_4 \end{array}\right], \qquad \tilde{\eta} = \psi\tilde{x} + \tilde{Z}_1\alpha + \tilde{Z}_2\gamma = \left[\begin{array}{c} \psi x_1 + \alpha_1 + \gamma_1 \\ \psi x_2 + \alpha_1 + \gamma_3 \\ \psi x_3 + \alpha_1 + \gamma_4 \\ \hline \psi x_4 + \alpha_2 + \gamma_2 \\ \psi x_5 + \alpha_2 + \gamma_3 \\ \psi x_6 + \alpha_2 + \gamma_4 \\ \hline \psi x_7 + \alpha_3 + \gamma_1 \\ \psi x_8 + \alpha_3 + \gamma_2 \\ \psi x_9 + \alpha_3 + \gamma_4 \end{array}\right].$$

In tale esempio, mentre ciascun effetto fisso α_i compare esattamente $c = 3$ volte, si osserva che γ_1 e γ_2 compaiono 2 volte, mentre γ_4 registra 3 occorrenze.

Un approccio alternativo per indurre sparsità nella struttura degli effetti fissi incrociati $\lambda = (\alpha, \gamma) = (\alpha_1, \ldots, \alpha_R, \gamma_1, \ldots, \gamma_C)$ è il seguente. Si riconsideri il predittore lineare (4.1), con $X = [x, Z_1, Z_2]$, e Z_1, Z_2 matrici indicatrici degli effetti fissi incrociati precedentemente definite. Si consideri ora una matrice Z, di soli 0 ed 1, con dimensione $R \times C$ indicatrice delle osservazioni. Detta $\mathbb{E}(N)$ la numerosità campionaria complessiva attesa desiderata, la matrice Z può essere generata casualmente generando ciascun suo elemento in modo indipendente da una variabile casuale di Bernoulli con probabilità di successo $\mathbb{E}(N)/RC$, ossia

$$Z_{ij} \sim \mathrm{Bi}\Big(1, \frac{\mathbb{E}(N)}{RC}\Big), \quad i = 1, \ldots, R, \ j = 1, \ldots, C.$$

Ciò comporta che la numerosità campionaria effettiva $N = \sum_{i=1}^{R}\sum_{j=1}^{C} Z_{ij}$ sia casuale,

ma con un piccolo coefficiente di variazione.

Data la matrice Z, si costruisce un vettore filtro f di 0 ed 1, di lunghezza RC, semplicemente trasponendo la matrice Z e concatenando le colonne di Z^T in unico vettore, che registra se ogni elemento di Z corrisponde ad un'osservazione (1) oppure no (0). A questo punto, le matrici indicatrici piene Z_1 e Z_2 vengono trasformate nelle matrici sparse $\tilde{Z}_1 = Z_1[f, \cdot]$ e $\tilde{Z}_2 = Z_2[f, \cdot]$ applicando il filtro f, ossia selezionando solo le righe specificate da f. Dunque, il nuovo predittore lineare è $\tilde{\eta} = \psi\tilde{x} + \tilde{Z}_1\alpha + \tilde{Z}_2\gamma$. Questo metodo, a differenza del precedente, non permette di controllare in modo omogeneo la sparsità di uno dei due effetti fissi. Tuttavia, ha il vantaggio di permettere una migliore flessibilità in termini di sparsità in entrambi gli effetti fissi, controllando la probabilità di ciascuna osservazione. Inoltre, tale metodo è coerente con l'approccio considerato ad esempio in Ghosh et al. (2022b) e Bellio et al. (2023a).

Se R ed C vengono rispettivamente espressi come potenze della numerosità campionaria N, ossia $R = N^\rho$ e $C = N^\delta$, con $\rho, \delta > 0$, allora uno scenario bilanciato in termini di sparsità nelle righe e nelle colonne si ottiene quando $\rho = \delta$ e $\rho + \delta > 1$. Infatti, in tal caso, $N/RC = N/N^\rho N^\delta = N^{1-\rho-\delta} \to 0$, quando $N \to \infty$, e si ha una sparsità omogenea sia nelle righe che nelle colonne. Inoltre, nello scenario bilanciato, con $\rho = \delta$, si ha che il rapporto tra il numero di righe e di colonne $R/C = N^\rho/N^\delta$ è asintoticamente costante. Ciò significa che i due effetti fissi α e γ crescono asintoticamente con la stessa velocità. Al contrario, uno scenario sbilanciato si ottiene, ad esempio, quando $\rho > \delta > 0.5$. In tal caso il rapporto $R/C = N^\rho/N^\delta$ diverge quando $N \to \infty$, e dunque, uno degli effetti fissi, α, domina asintoticamente l'altro, γ.

Poiché da $R = N^\rho$ e $C = N^\delta$ segue che

$$\rho = \frac{\log R}{\log N}, \qquad \delta = \frac{\log C}{\log N},$$

gli scenari (S.1)-(S.4) considerati nelle simulazioni, descritti nella (4.S), tutti bilanciati, possono essere riassunti, utilizzando $\mathbb{E}(N)$ al posto di N, in termini di ρ e δ come

(**S.1**) $\rho = \delta \doteq 0.57$,

(**S.2**) $\rho = \delta \doteq 0.63$,

(**S.3**) $\rho = \delta \doteq 0.66$,

(**S.4**) $\rho = \delta \doteq 0.69$.

Sebbene entrambi i metodi precedentemente descritti siano stati considerati come possibili valide alternative per introdurre sparsità nei dati, nelle simulazioni di seguito riportate si è scelto di utilizzare il secondo schema, in quanto più coerente con la

letteratura recente. Nonostante ciò, risultati empirici, non riportati nel seguito, hanno mostrato che, a prescindere dal metodo scelto per costruire le matrici Z_1 e Z_2 in modo sparso, a parità di condizioni di sparsità, le procedure inferenziali esaminate si comportano in modo sostanzialmente analogo.

Sia nel caso non sparso che in quello sparso, il modello risulta sovraparametrizzato. Per rispettare i vincoli di identificabilità, è sufficiente imporre la condizione che uno dei parametri α_i sia uguale a zero (o equivalentemente uno dei γ_j). Questo corrisponde a rimuovere una delle colonne della matrice Z_1 (rispettivamente $\tilde{Z}_1$) o equivalentemente una delle colonne della matrice Z_2 (rispettivamente $\tilde{Z}_2$).

Questo tipo di rappresentazione sparsa degli effetti fissi è utile per modellare alcuni scenari in diversi contesti applicativi. Ad esempio, si consideri lo scenario in cui si dispone di una lista di R studenti al termine di un percorso di laurea e delle informazioni relative agli esami da essi svolti. Usualmente, gli studenti devono svolgere un numero prefissato di esami (ad esempio 19 in un corso di laurea triennale), di cui solamente alcuni sono obbligatori, mentre molti altri sono a scelta tra un paniere di C possibili esami. Ciò significa che non tutti gli studenti svolgono lo stesso insieme di esami a scelta e allo stesso tempo non tutti gli esami vengono fatti da tutti gli studenti. Pertanto, i dati a disposizione hanno una struttura del tipo

	Esame 1	Esame 2	Esame 3	...	Esame $C-1$	Esame C
Studente 1	26		21	...	25	20
Studente 2		28	23	...	21	
Studente 3	22	26		...		30
Studente 4	18	19	21	...	30	23
...	...	...	...	...	...	...
Studente $R-1$	30		24	...	29	
Studente R	19	27	29	...	22	23

In questo scenario gli effetti fissi α_i, $i = 1 \ldots, R$ potrebbero rappresentare l'abilità dello studente i-esimo, e gli effetti fissi γ_j, $j = 1, \ldots, C$ la difficoltà dell'esame j-esimo. Per questo tipo di dati, dunque, un modello di regressione log-lineare Poisson con effetti fissi incrociati sparsi potrebbe risultare adeguato.

Un altro esempio realistico, esaminato da Ghosh et al. (2022a,b), è quello relativo ai dati prodotti da un rivenditore di vestiti online americano (`Stitch Fix`). Questo rivenditore dà la possibilità di ricevere a casa delle scatole di vestiti e offre l'opportunità al soggetto che riceve la scatola di scegliere quali vestiti tenere e quali invece rendere. Assumendo che ai soggetti siano inviate scatole diverse, sulla base dei gusti espressi, gli

R soggetti sceglieranno di acquistare vestiti diversi tra tutti i C possibili capi d'abbigliamento. Date le liste complete dei soggetti e dei vestiti, se si registra solo l'informazione relativa a quali vestiti vengono tenuti (1) oppure no (0), i dati a disposizione hanno una struttura del tipo

	Vestito 1	Vestito 2	Vestito 3	...	Vestito $C-1$	Vestito C
Soggetto 1	1	0		...		0
Soggetto 2	0		1	...	0	0
Soggetto 3	1	1		...		0
Soggetto 4	0	0	0	...	1	1
...	...	...	...	...	...	...
Soggetto $R-1$	0	0	0	...		
Soggetto R	1	0	1	...	0	1

In questo scenario gli effetti fissi α_i, $i = 1 \ldots, R$, potrebbero essere relativi alle caratteristiche specifiche del soggetto i-esimo, e gli effetti fissi γ_j, $j = 1, \ldots, C$, essere relativi alle caratteristiche specifiche del vestito j-esimo. Per questo tipo di dati, dunque, un modello di regressione logistica con effetti fissi incrociati sparsi potrebbe risultare adeguato. In realtà, in Ghosh et al. (2022b) si assume che gli effetti incrociati α_i e γ_j siano a loro volta delle variabili casuali, ossia degli effetti casuali e il modello suggerito è un modello di regressione logistica con effetti casuali, che rientra nella classe dei GLMM (*Generalized Linear Mixed Models*). Nella presente trattazione, invece, si assume che α_i e γ_j siano effetti fissi e pertanto parametri fissati.

4.4 Simulazioni: modello log-lineare Poisson con effetti fissi incrociati

In questo paragrafo si presentano i risultati di simulazione relativamente al modello di regressione Poisson con funzione di legame canonica in presenza di effetti fissi incrociati e sparsità nei dati. Il modello, che viene brevemente richiamato in seguito, e le relative procedure inferenziali sono descritte in dettaglio nel paragrafo 3.3.2.

Si assuma che y_{ij} siano realizzazioni di variabili casuali indipendenti Y_{ij} con distribuzione Poisson di media μ_{ij} dove $g(\mu_{ij}) = \eta_{ij}$, con $\eta_{ij} = \alpha_i + \gamma_j + \psi x_{ij}$, per $i = 1, \ldots, R$, $j = 1, \ldots, C$. Al solito, il vincolo di identificabilità può essere rispettato ponendo a zero uno degli α_i. Senza perdita di generalità si è assunto di avere a disposizione un'unica covariata x e che il parametro di interesse ψ sia scalare. Invece, il parametro incidentale

$\lambda = (\alpha, \gamma)$ ha dimensione pari ad $R + C$. Assumendo la funzione di legame canonica $g(\mu_{ij}) = \log \mu_{ij} = \eta_{ij}$, segue che

$$Y_{ij} \sim Po(e^{\eta_{ij}}),$$

$i = 1, \ldots, R$, $j = 1, \ldots, C$.

Come spiegato nel paragrafo 4.2, la covariata x_{ij} e gli effetti fissi α_i e γ_j sono stati generati in modo indipendente da una distribuzione normale standard e mantenuti fissati in tutti i campioni simulati. Il parametro di interesse ψ è stato fissato pari ad 1, $\psi = 1$, e mantenuto fissato in tutti i campioni simulati.

Nella Tabella 4.2 si presentano i risultati delle simulazioni per quanto riguarda il confronto tra l'usuale stimatore di massima verosimiglianza $\hat{\psi}$, denotato con MLE, e lo stimatore basato sulla verosimiglianza profilo modificata (1.16), $\hat{\psi}_M$, denotato con MPL. Globalmente, i risultati mostrano come entrambi gli stimatori si comportino in modo soddisfacente in tutti e 4 gli scenari considerati. Tuttavia, si osserva che la distorsione dello stimatore $\hat{\psi}_M$ è sempre inferiore, seppur di poco, alla distorsione dell'usuale stimatore di massima verosimiglianza. Questo risulta meno evidente nel primo scenario (S.1), che è confrontabile con lo scenario (S.2) in termini di dimensione di R e C, ma che è caratterizzato da una sparsità inferiore. Ciò è dovuto al fatto che, a parità di R e C, nello scenario (S.1) c'è più informazione a disposizione per stimare il parametro di disturbo e quindi anche l'inferenza sul parametro di interesse ψ è più accurata. Dall'altra parte, se si confrontano gli scenari (S.2)-(S.4), si osserva come, tenuto praticamente fissato il rapporto $\tilde{k}$ tra il numero effettivo di parametri di disturbo e numerosità campionaria effettiva N, le differenze tra i due stimatori tendono a ridursi all'aumentare di R e C, in uno schema bilanciato. Infatti, le differenze maggiori tra i due metodi si osservano in corrispondenza dello scenario (S.2). Negli scenari (S.2)-(S.4), entrambi i metodi considerati mostrano una lieve tendenza a sottostimare la variabilità dello stimatore, poiché il rapporto SD/SE è sempre inferiore ad 1, seppur di poco.

Nella Tabella 4.3 si riassumono, invece, i risultati delle simulazioni per quanto riguarda le 8 statistiche test che sono state calcolate per la verifica di ipotesi $H_0 : \psi = 1$ contro $H_1 : \psi < 1$, per i 4 diversi scenari (S.1)-(S.4). I simboli utilizzati sono spiegati più nel dettaglio nella Tabella 4.1. Si evince che, nonostante i livelli empirici dell'usuale statistica test radice con segno del log-rapporto di verosimiglianza profilo $r_P(\psi)$ non si discostino in modo evidente da quelle nominali (i valori di riferimento sono riportati nell'intestazione della Tabella 4.3, e rappresentano le probabilità di copertura attese in ipotesi di normalità), i metodi basati sulla verosimiglianza profilo modificata e sulle

	Estimate	Bias	PU	SD	SE	RMSE	SE/SD
(S.1)	$\boldsymbol{R = 20,\ C = 20,\ \mathbb{E}(N) = 200,\ N = 189,\ \tilde{k} = 0.21}$						
MLE	1.00889	0.00889	0.458	0.06966	0.07059	0.07019	1.01331
MPL	1.00119	0.00119	0.508	0.06884	0.06998	0.06882	1.01652
(S.2)	$\boldsymbol{R = 20,\ C = 20,\ \mathbb{E}(N) = 120,\ N = 115,\ \tilde{k} = 0.35}$						
MLE	1.02348	0.02348	0.433	0.12246	0.11872	0.12463	0.96945
MPL	1.00816	0.00816	0.473	0.12012	0.11680	0.12034	0.97239
(S.3)	$\boldsymbol{R = 30,\ C = 30,\ \mathbb{E}(N) = 180,\ N = 177,\ \tilde{k} = 0.34}$						
MLE	1.01789	0.01789	0.409	0.07220	0.07020	0.07435	0.97236
MPL	1.00558	0.00558	0.470	0.07081	0.06879	0.07100	0.97149
(S.4)	$\boldsymbol{R = 50,\ C = 50,\ \mathbb{E}(N) = 300,\ N = 284,\ \tilde{k} = 0.35}$						
MLE	1.01328	0.01328	0.434	0.06625	0.06391	0.06754	0.96460
MPL	1.00154	0.00154	0.504	0.06508	0.06250	0.06507	0.96040

TABELLA 4.2: Modello Poisson: inferenza su $\psi = 1$. Confronto tra verosimiglianza profilo (MLE) e verosimiglianza profilo modificata (MPL), per i 4 diversi scenari (S.1)-(S.4) descritti nella (4.S).

modifiche di $r_P(\psi)$ registrano delle coperture globalmente più vicine a quelle attese. In modo simile, anche le procedure basate sul *bootstrap* sembrano apportare dei miglioramenti rispetto all'utilizzo della statistica $r_P(\psi)$. Ad ogni modo, sia le modificazioni analitiche che gli approcci basati sul *bootstrap* non risultano apportare miglioramenti marcatamente evidenti. Se ci si focalizza sui metodi *bootstrap*, risulta che l'*unconstrained bootstrap* porti ad ottenere coperture empiriche leggermente peggiori rispetto al *constrained bootstrap*, in modo più marcato nello scenario (S.4) e quando non si considerano i metodi penalizzati. Al contrario, nel confronto tra diverse procedure *bootstrap* basate su stime penalizzate, non si evidenziano differenze evidenti tra le due varianti del *bootstrap* e tra le diverse scelte del parametro di penalizzazione a. D'altra parte, le procedure penalizzate sembrano funzionare sempre leggermente meglio rispetto all'usuale *unconstrained bootstrap*, mentre producono risultati molto simili al *constrained bootstrap*, con un'unica lieve eccezione per lo scenario (S.1), dove il *constrained bootstrap* da stime penalizzate porta ad ottenere risultati leggermente migliori. Nel confronto tra i diversi scenari (S.1)-(S.4), sembra che tutti i metodi portino ad ottenere risultati più vicini a quelli attesi nel caso dello scenario meno sparso (S.1), soprattutto se ci si concentra sulla coda destra della distribuzione. Al contrario, non sembra emergere una direzione evidente nel comportamento delle diverse procedure quando si confrontano gli scenari (S.2)-(S.4), a parità di sparsità rappresentata da $\tilde{k}$.

	1.0	2.5	5.0	95.0	97.5	99.0
(S.1) $\boldsymbol{R = 20,\ C = 20,\ \mathbb{E}(N) = 200,\ N = 189,\ \tilde{k} = 0.21}$						
r_p	0.6	2.0	4.1	94.8	97.0	99.0
r_M	0.8	2.6	4.7	96.0	98.0	99.5
r^*	1.2	3.1	5.3	95.9	98.0	99.3
u. boot r_p	0.5	2.1	4.0	95.2	97.4	99.0
c. boot r_p	0.5	1.9	4.0	94.9	97.0	99.0
u. penalized $[a = 0.5]$ boot r_p	0.7	1.7	3.8	94.9	97.4	99.0
c. penalized $[a = 0.5]$ boot r_p	0.6	2.2	3.9	95.2	97.5	99.0
c. penalized $[a = 1]$ boot r_p	0.8	2.1	4.3	95.2	97.4	99.1
(S.2) $\boldsymbol{R = 20,\ C = 20,\ \mathbb{E}(N) = 120,\ N = 115,\ \tilde{k} = 0.35}$						
r_p	0.8	2.2	4.9	92.8	96.0	98.4
r_M	1.9	4.2	7.0	94.9	97.8	99.5
r^*	1.2	3.0	5.9	94.6	97.6	99.4
u. boot r_p	1.8	3.5	6.4	94.5	97.5	99.1
c. boot r_p	0.2	2.1	4.9	94.4	97.0	98.8
u. penalized $[a = 0.5]$ boot r_p	0.6	2.5	5.4	94.1	96.4	98.7
c. penalized $[a = 0.5]$ boot r_p	0.7	2.5	5.2	94.3	97.1	98.9
c. penalized $[a = 1]$ boot r_p	0.9	2.6	5.5	94.3	96.7	98.5
(S.3) $\boldsymbol{R = 30,\ C = 30,\ \mathbb{E}(N) = 180,\ N = 177,\ \tilde{k} = 0.34}$						
r_p	0.7	1.0	2.9	91.7	95.6	98.3
r_M	0.9	2.3	6.0	95.1	97.1	98.9
r^*	0.9	1.8	4.9	94.9	96.6	98.9
u. boot r_p	1.7	2.6	5.5	94.5	96.4	98.6
c. boot r_p	1.0	2.8	5.1	95.4	98.3	99.2
u. penalized $[a = 0.5]$ boot r_p	1.1	2.9	5.3	94.9	98.3	99.1
c. penalized $[a = 0.5]$ boot r_p	1.2	3.0	5.0	95.2	98.3	99.3
c. penalized $[a = 1]$ boot r_p	1.1	2.9	5.1	95.1	98.3	99.2
(S.4) $\boldsymbol{R = 50,\ C = 50,\ \mathbb{E}(N) = 300,\ N = 284,\ \tilde{k} = 0.35}$						
r_p	1.0	2.5	3.9	92.2	95.3	97.8
r_M	2.5	4.0	7.6	94.7	97.1	98.9
r^*	2.2	4.0	7.3	94.4	96.8	98.5
u. boot r_p	3.6	5.4	8.8	93.1	95.5	97.4
c. boot r_p	0.9	2.5	5.1	93.2	96.8	98.6
u. penalized $[a = 0.5]$ boot r_p	1.1	2.9	5.1	93.1	97.0	98.3
c. penalized $[a = 0.5]$ boot r_p	1.1	2.3	4.8	93.4	96.8	98.3
c. penalized $[a = 1]$ boot r_p	1.2	3.0	5.5	93.1	96.9	98.3

TABELLA 4.3: Modello Poisson: coperture empiriche x 100. Confronto tra le statistiche test per la verifica di ipotesi $H_0 : \psi = 1$ contro $H_1 : \psi < 1$, per i 4 diversi scenari (S.1)-(S.4) descritti nella (4.S).

4.5 Simulazioni: modello logistico con effetti fissi incrociati

In questo paragrafo si presentano i risultati di simulazione relativamente al modello di regressione logistico con funzione di legame canonica in presenza di effetti fissi incrociati e sparsità nei dati. Il modello, che viene brevemente richiamato in seguito, e le relative procedure inferenziali sono descritte in dettaglio nel paragrafo 3.3.3.

Si assuma che y_{ij} siano realizzazioni di variabili casuali indipendenti Y_{ij} con distribuzione bernoulliana con probabilità di successo π_{ij}, $i = 1, \ldots, R$, $j = 1, \ldots, C$, dove $g(\pi_{ij}) = \eta_{ij}$, con $\eta_{ij} = \alpha_i + \gamma_j + \psi x_{ij}$, per $i = 1, \ldots, R$, $j = 1, \ldots, C$. Al solito, il vincolo di identificabilità può essere rispettato ponendo a zero uno degli α_i. Senza perdita di generalità si è assunto di avere a disposizione un'unica covariata x e che il parametro di interesse ψ sia scalare. Invece, il parametro incidentale $\lambda = (\alpha, \gamma)$ ha dimensione pari ad $R + C$. Assumendo la funzione di legame canonica $g(\pi_{ij}) = \text{logit}(\pi_{ij}) = \eta_{ij}$, segue che

$$Y_{ij} \sim Bi\Big(1, \frac{e^{\eta_{ij}}}{1 + e^{\eta_{ij}}}\Big),$$

$i = 1, \ldots, R$, $j = 1, \ldots, C$.

Come spiegato nel paragrafo 4.2, la covariata x_{ij} e gli effetti fissi α_i e γ_j sono stati generati in modo indipendente da una distribuzione normale standard e mantenuti fissati in tutti i campioni simulati. Il parametro di interesse ψ è stato fissato pari ad 1, $\psi = 1$, e mantenuto fissato in tutti i campioni simulati.

In modo analogo alla Tabella 4.2, nella Tabella 4.4 si presentano i risultati delle simulazioni per quanto riguarda il confronto tra l'usuale stimatore di massima verosimiglianza $\hat{\psi}$, denotato con MLE, e lo stimatore basato sulla verosimiglianza profilo modificata (1.16), $\hat{\psi}_M$, denotato con MPL. Globalmente, i risultati mostrano come, a differenza di quanto emerso nel caso del modello log-lineare Poisson, entrambi gli stimatori si comportino in modo insoddisfacente nei 4 scenari considerati. I risultati migliori si hanno in corrispondenza dello scenario (S.1) se si considera lo stimatore basato sulla verosimiglianza profilo modificata. In questo scenario, con minore sparsità, la presenza di maggiore informazione permette allo stimatore basato sulla verosimiglianza profilo modificata di avere una distorsione quasi trascurabile. In tutti e 4 gli scenari è evidente che lo stimatore di massima verosimiglianza $\hat{\psi}$ risulta inaccurato per l'inferenza sul parametro di interesse ψ, esibendo sempre una grave distorsione, in particolare nello

	Estimate	Bias	PU	SD	SE	RMSE	SE/SD
(S.1)	$\boldsymbol{R = 20,\ C = 20,\ \mathbb{E}(N) = 200,\ N = 189,\ \tilde{k} = 0.21}$						
MLE	1.42866	0.42866	0.116	0.38990	0.29952	0.57933	0.76819
MPL	1.07833	0.07833	0.401	0.26019	0.24352	0.27160	0.93593
(S.2)	$\boldsymbol{R = 20,\ C = 20,\ \mathbb{E}(N) = 120,\ N = 115,\ \tilde{k} = 0.35}$						
MLE	8.37514	7.37514	0.083	45.12041	944.90591	45.69680	20.94188
MPL	1.30951	0.30951	0.294	0.56178	0.42680	0.64115	0.75974
(S.3)	$\boldsymbol{R = 30,\ C = 30,\ \mathbb{E}(N) = 180,\ N = 177,\ \tilde{k} = 0.34}$						
MLE	2.38542	1.38542	0.030	2.34456	0.63357	2.72229	0.27023
MPL	1.23561	0.23561	0.256	0.38013	0.30457	0.44707	0.80122
(S.4) [1]	$\boldsymbol{R = 50,\ C = 50,\ \mathbb{E}(N) = 300,\ N = 284,\ \tilde{k} = 0.35}$						
MLE	2.30308	1.30308	0.016	0.88971	0.48037	1.57735	0.53992
MPL	1.22981	0.22981	0.222	0.33042	0.25589	0.40220	0.77444

TABELLA 4.4: Modello logistico: inferenza su $\psi = 1$. Confronto tra verosimiglianza profilo (MLE) e verosimiglianza profilo modificata (MPL), per i 4 diversi scenari (S.1)-(S.4) descritti nella (4.S).

scenario (S.2), caratterizzato dalla maggiore sparsità. Se si confrontano gli scenari (S.2)-(S.4), si osserva come, tenuto praticamente fissato il rapporto $\tilde{k}$ tra il numero effettivo di parametri di disturbo e numerosità campionaria effettiva N, le differenze tra i due stimatori tendono a ridursi all'aumentare di R e C, in uno schema bilanciato. Infatti, le differenze maggiori tra i due metodi si osservano in corrispondenza dello scenario (S.2). Tuttavia, a differenza del caso del modello log-lineare Poisson, la distorsione di entrambi gli stimatori non sembra diventare trascurabile all'aumentare di R e C.

In modo analogo alla Tabella 4.4, nella Tabella 4.3 si riassumono, invece, i risultati delle simulazioni per quanto riguarda le 8 statistiche test che sono state calcolate per la verifica di ipotesi $H_0 : \psi = 1$ contro $H_1 : \psi < 1$, per i 4 diversi scenari (S.1)-(S.4). I simboli utilizzati sono spiegati più nel dettaglio nella Tabella 4.1. Si evince che i metodi basati sull'usuale verosimiglianza profilo conducono a conclusioni inferenziali poco affidabili in tutti e 4 gli scenari considerati. Infatti, i livelli empirici dell'usuale statistica test radice con segno del log-rapporto di verosimiglianza profilo $r_P(\psi)$ si discostano in modo evidente da quelli nominali in tutti gli scenari. I metodi basati sulla verosimiglianza profilo modificata e sulle modifiche di $r_P(\psi)$ permettono in generale di ottenere risultati migliori rispetto all'usuale verosimiglianza profilo, sebbene i corrispettivi livelli empirici delle statistiche test si discostino in alcuni casi in modo un po' più evidente dai livelli nominali. In particolare, contrariamente a quanto atteso, la statistica $r_M(\psi)$

[1] Per motivi legati ai tempi computazionali, nella Tabella 4.4 le simulazioni relative allo scenario (S.4) sono state basate solo su Nsim = 500 replicazioni, anziché 1000.

sembra mostrare un comportamento migliore rispetto alla statistica $r_P^*(\psi)$, nei diversi scenari considerati, ad eccezione dello scenario (S.1) con minore sparsità. Ciò, probabilmente, è legato a motivi di instabilità numerica nel calcolo di $r_P^*(\psi)$. In modo simile, anche le procedure basate sul *bootstrap* apportano dei miglioramenti rispetto all'utilizzo della statistica $r_P(\psi)$. A differenza del caso del modello log-lineare Poisson, sia le modificazioni analitiche che gli approcci basati sul *bootstrap* consentono di ottenere miglioramenti marcatamente evidenti rispetto all'utilizzo dell'usuale verosimiglianza profilo. Se ci si focalizza sui metodi *bootstrap*, risulta che l'*unconstrained bootstrap* porti ad ottenere coperture empiriche leggermente peggiori rispetto al *constrained bootstrap*, in modo più marcato nello scenario (S.4) e quando non si considerano i metodi penalizzati. Nel confronto tra diverse procedure *bootstrap* basate su stime penalizzate, a differenza del caso del modello log-lineare Poisson, si evidenziano alcune differenze tra le due varianti del *bootstrap* e tra le diverse scelte del parametro di penalizzazione a. Nello specifico, utilizzare un valore del parametro di penalizzazione a nella (2.15) più grande rispetto all'usuale coefficiente $a = 1/2$ della (2.14), in particolare $a = 1$, consente di registrare delle coperture empiriche sempre migliori in presenza di moderata sparsità, ossia negli scenari (S.2)-(S.4), mentre tale differenza non risulta evidente in presenza di minore sparsità, come nello scenario (S.1). Inoltre, le procedure penalizzate funzionano sempre meglio rispetto all'usuale *unconstrained bootstrap*, e producono risultati migliori rispetto al *constrained bootstrap* (ad eccezione per lo scenario (S.1)), a differenza del caso del modello log-lineare Poisson, dove le differenza tra procedure *bootstrap* penalizzate e *constrained bootstrap* sono meno evidenti.

Pertanto, almeno per il modello logistico, fare ricorso all'utilizzo di stime penalizzate nel *constrained bootstrap* sembra permettere di correggere l'inferenza sul parametro di interesse in modo migliore rispetto all'usuale *constrained bootstrap*, con risultati simili all'utilizzo delle quantità pivotali basate sulla verosimiglianza profilo modificata. Dunque, mentre come emerso in Bellio et al. (2023b) nel caso di modelli stratificati per dati continui, entrambe le varianti del *bootstrap* consentono di raggiungere il medesimo livello di accuratezza delle modificazioni analitiche della funzione di verosimiglianza, nel caso dei modelli con effetti fissi incrociati e dati discreti sparsi, sembra che tale equivalenza sia rispettata solamente affidandosi ai metodi *bootstrap* penalizzati. Considerazioni teoriche ulteriori dovrebbero essere prese in considerazione per stabilire se effettivamente esista questo tipo di equivalenza. Allo stesso modo, la scelta del parametro di penalizzazione a nella (2.15) rimane una questione delicata, che dev'essere ulteriormente approfondita sia da un punto di vista teorico che empirico.

	1.0	2.5	5.0	95.0	97.5	99.0
(S.1) $R = 20,\ C = 20,\ \mathbb{E}(N) = 200,\ N = 189,\ \tilde{k} = 0.21$						
r_p	0.0	0.1	0.3	58.3	68.6	78.7
r_M	0.9	1.8	3.5	91.4	95.1	97.9
r^*	1.4	2.7	4.4	92.8	96.1	98.6
u. boot r_p	0.1	0.1	0.3	96.1	98.2	99.7
c. boot r_p	0.3	1.7	4.0	95.1	96.9	98.5
u. penalized [$a = 0.5$] boot r_p	0.0	0.1	0.2	98.0	99.0	99.9
c. penalized [$a = 0.5$] boot r_p	0.7	2.5	5.1	95.3	97.0	98.9
c. penalized [$a = 1$] boot r_p	0.6	3.0	6.2	95.7	96.8	98.7
(S.2) $R = 20,\ C = 20,\ \mathbb{E}(N) = 120,\ N = 115,\ \tilde{k} = 0.35$						
r_p	0.0	0.4	0.4	41.8	51.2	60.5
r_M	0.6	2.2	4.1	95.2	98.0	99.7
r^*	2.9	3.8	4.9	85.3	91.1	95.5
u. boot r_p	8.3	10.8	13.4	90.8	92.1	93.2
c. boot r_p	4.7	7.6	11.3	99.9	100.0	100.0
u. penalized [$a = 0.5$] boot r_p	5.8	8.5	11.5	97.0	99.3	99.8
c. penalized [$a = 0.5$] boot r_p	3.8	5.2	6.9	100.0	100.0	100.0
c. penalized [$a = 1$] boot r_p	1.0	3.2	5.9	95.4	97.8	99.4
(S.3) $R = 30,\ C = 30,\ \mathbb{E}(N) = 180,\ N = 177,\ \tilde{k} = 0.34$						
r_p	0.0	0.0	0.2	31.7	41.3	52.3
r_M	0.9	1.3	2.3	93.8	96.5	98.5
r^*	3.2	3.8	4.1	82.8	88.9	94.1
u. boot r_p	5.5	8.1	10.3	93.9	95.4	96.2
c. boot r_p	4.6	7.4	9.0	97.9	99.7	100.0
u. penalized [$a = 0.5$] boot r_p	2.8	4.8	7.3	97.7	98.6	99.4
c. penalized [$a = 0.5$] boot r_p	2.8	3.5	4.0	99.4	100.0	100.0
c. penalized [$a = 1$] boot r_p	1.5	3.3	6.7	96.8	98.3	98.9
(S.4) [2] **$R = 50,\ C = 50,\ \mathbb{E}(N) = 300,\ N = 284,\ \tilde{k} = 0.35$**						
r_p	0.0	0.2	0.2	16.0	23.2	31.9
r_M	1.4	3.4	6.0	94.6	98.4	99.4
r^*	4.2	4.4	4.4	75.8	85.2	92.6
u. boot r_p	9.8	11.4	14.4	92.0	93.6	96.0
c. boot r_p	1.6	4.6	8.4	84.4	97.2	99.8
u. penalized [a=0.5] boot r_p	1.6	3.2	6.4	95.8	98.4	99.2
c. penalized [a=0.5] boot r_p	0.8	1.4	1.8	99.2	99.6	100.0
c. penalized [a=1] boot r_p	0.6	1.8	5.2	94.2	97.2	98.8

TABELLA 4.5: Modello logistico: coperture empiriche x 100. Confronto tra le statistiche test per la verifica di ipotesi $H_0 : \psi = 1$ contro $H_1 : \psi < 1$, per i 4 diversi scenari (S.1)-(S.4) descritti nella (4.S).

[2]Per motivi legati ai tempi computazionali, nella Tabella 4.5 le simulazioni relative allo scenario (S.4) sono state basate solo su Nsim = 500 replicazioni, anziché 1000.

Conclusioni

In questa tesi sono stati presentati diversi approcci per migliorare l'inferenza in presenza di un elevato numero di parametri di disturbo. Particolare attenzione è stata dedicata ai modelli a due indici asintotici con effetti fissi stratificati e incrociati.

Dopo aver richiamato alcune alternative all'utilizzo della verosimiglianza profilo, che non consente un'inferenza accurata nel caso dei parametri incidentali, come opportune pseudo-verosimiglianze, modificazioni della consueta funzione di log-verosimiglianza profilo e della statistica radice con segno del log-rapporto di verosimiglianza profilo, è stata considerata anche la possibilità di ricorrere a metodi *bootstrap*. L'obiettivo della tesi era confrontare l'efficacia dei metodi analitici e dei metodi *bootstrap* rispetto all'utilizzo della verosimiglianza profilo.

In particolare, si è visto che, nella letteratura relativa ai modelli con effetti fissi stratificati, l'inferenza si basa sull'utilizzo di una verosimiglianza marginale o condizionata, che permettono spesso di ripristinare le proprietà usuali dello stimatore di massima verosimiglianza, o in alternativa, sul ricorso a modifiche della funzione di verosimiglianza profilo o alle tecniche *bootstrap*. Successivamente, l'interesse si è concentrato sui modelli con effetti fissi incrociati, per i quali non esistono ancora risultati teorici generali riguardo alle proprietà delle procedure inferenziali, in particolare per dati discreti sparsi. Si è posto un particolare accento sui modelli per dati discreti, come il modello log-lineare Poisson e il modello di regressione logistica, poiché in questo contesto non sono disponibili nemmeno risultati generali sulle proprietà dei metodi *bootstrap*, che sono dimostrati nel caso continuo. Dopo aver esaminato teoricamente le procedure inferenziali per questi modelli, queste ultime sono state valutate empiricamente attraverso alcuni studi di simulazione.

Nelle simulazioni condotte per confrontare i metodi analitici e l'approccio *bootstrap*, è stata considerata la presenza di sparsità nei dati. In particolare, sono state discusse diverse possibilità per introdurre tale sparsità e come questa possa essere controllata.

Coerentemente con la letteratura recente, si veda ad esempio Fernàndez-Val & Weidner (2018) e Leng et al. (2023), come approfondito nel paragrafo 3.2, gli studi di simulazione hanno mostrato che nel caso di effetti fissi incrociati e dati discreti sparsi, quando è di interesse fare inferenza su un parametro scalare, la distorsione dello stimatore di massima verosimiglianza è asintoticamente trascurabile, almeno rispetto allo *standard error*, all'aumentare della dimensione degli effetti fissi nel caso del modello log-lineare Poisson, ma non nel caso del modello logistico. In entrambi i casi, lo stimatore basato sulla verosimiglianza profilo modificata esibisce sempre una distorsione inferiore, al variare degli scenari considerati, portando a benefici significativi soprattutto nel caso del modello logistico. Sebbene, tra i vari problemi inferenziali, nelle considerazioni teoriche e negli studi di simulazione in questa tesi, si sia posto maggiore accento sul problema della verifica di ipotesi, è naturale attendersi che i risultati siano equivalenti per quanto concerne le probabilità di copertura degli intervalli di confidenza basati sulle usuali quantità pivotali di verosimiglianza, come la statistica test radice con segno del log-rapporto di verosimiglianza, o su loro modificazioni. Dunque, se ci si concentra sul problema di verifica di ipotesi sul parametro di interesse, mentre nel caso del modello log-lineare Poisson l'inferenza basata sul test del log-rapporto di verosimiglianza risulta accurata, e i miglioramenti derivanti dall'utilizzo di modificazioni analitiche o il ricorso ai metodi *bootstrap* non sono così evidenti, nel caso del modello logistico, ricorrere a metodi alternativi all'usuale verosimiglianza profilo sembra essere cruciale. Difatti, per tale modello, le usuali procedure di verosimiglianza portano a risultati scarsamente accurati e decisamente inaffidabili in condizioni di elevata sparsità. D'altro canto, i risultati numerici suggeriscono che nemmeno per il modello Poisson con effetti fissi incrociati esista una perfetta equivalenza tra l'usuale funzione di verosimiglianza profilo e verosimiglianza profilo modificata.

Coerentemente con le intuizioni legate agli studi empirici condotti da Bellio et al. (2023b), nel caso del modello logistico, gli studi di simulazione condotti in questa tesi confermano che, a differenza di quanto succede per modelli stratificati per dati continui, non sembra valida l'equivalenza tra *unconstrained* e *constrained bootstrap* nel recuperare l'accuratezza del primo ordine delle procedure inferenziali, con la variante *constrained* che porta in generale ad ottenere risultati migliori. Non solo, il ricorso a procedure *bootstrap* da stime penalizzate sembra risultare preferibile all'usuale *constrained bootstrap*, in particolare in situazioni di elevata sparsità. D'altra parte, la scelta dell'opportuno parametro di penalizzazione nella verosimiglianza modificata (2.15), e come tale scelta dipenda dal livello di sparsità considerato, rimangono aspetti che necessitano di ulteriori approfondimenti e che potrebbero essere oggetto di ricerche future.

I risultati degli studi di simulazione condotti in questa tesi si prestano facilmente ad essere resi più accurati aumentando sia il numero di simulazioni sia il numero di campioni *bootstrap*. Inoltre, possono essere altresì estesi per valutare diversi scenari in termini di sparsità e altri modelli per dati discreti.

Bibliografia

Andersen, E. (1980). *Discrete Statistical Models with Social Science Applications*. North-Holland, Amsterdam.

Azzalini, A. (2001). *Inferenza Statistica. Una Presentazione Basata sul Concetto di Verosimiglianza*. Milano: Springer-Verlag Italia.

Barndorff-Nielsen, O. (1980). Conditionality resolutions. *Biometrika* **67**, 293–310.

Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika* **70**, 343–365.

Barndorff-Nielsen, O. (1991). Modified signed log likelihood ratio. *Biometrika* **78**, 557–563.

Barndorff-Nielsen, O. & Cox, D. (1994). *Inference and Asymptotics*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* **160**, 268–282.

Bartolucci, F., Bellio, R., Salvan, A. & Sartori, N. (2016). Modified profile likelihood for fixed-effects panel data models. *Econometric Reviews* **35**, 1271–1289.

Battey, H. S. & Cox, D. R. (2020). High dimensional nuisance parameters: an example from parametric survival analysis. *Information Geometry* **3**, 119–148.

Battey, H. S. & Cox, D. R. (2022). Some Perspectives on Inference in High Dimensions. *Statistical Science* **37**, 110 – 122.

Bellio, R., Ghosh, S., Owen, A. B. & Varin, C. (2023a). Scalable estimation of probit models with crossed random effects.

Bellio, R., Kosmidis, I., Salvan, A. & Sartori, N. (2023b). Parametric bootstrap inference for stratified models with high-dimensional nuisance specifications. *Statistica Sinica* **33**, 1069–1091.

Beran, R. (1997). Diagnosing bootstrap success. *Annals of the Institute of Statistical Mathematics* **49**, 1–24.

Bickel, P. J. & Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics* **9**, 1196–1217.

Brazzale, A. R., Davison, A. C. & Reid, N. (2007). *Applied Asymptotics: Case Studies in Small-Sample Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Chamberlain, G. (2010). Binary response models for panel data: Identification and information. *Econometrica* **78**, 159–168.

Cox, D. R. & Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)* **49**, 1–39.

Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press.

Davison, A. C., Hinkley, D. V. & Young, G. A. (2003). Recent developments in bootstrap methodology. *Statistical Science* **18**, 141–157.

DiCiccio, T. J., Martin, M. A. & Stern, S. E. (2001). Simple and accurate one-sided inference from signed roots of likelihood ratios. *The Canadian Journal of Statistics* **29**, 67–76.

DiCiccio, T. J. & Romano, J. P. (1995). On bootstrap procedures for second-order accurate confidence limits in parametric models. *Statistica Sinica* **5**, 141–160.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* **7**, 1–26.

Efron, B. (1998). R. A. Fisher in the 21st Century. *Statistical Science* **13**, 95–114.

Efron, B. & Hastie, T. (2016). *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Institute of Mathematical Statistics Monographs. Cambridge University Press.

Efron, B. & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

Fernàndez-Val, I. & Weidner, M. (2016). Individual and time effects in nonlinear panel models with large n, t. *Journal of Econometrics* **192**, 291–312.

Fernàndez-Val, I. & Weidner, M. (2018). Fixed effects estimation of large-t panel data models. *Annual Review of Economics* **10**, 109–138.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A* **222**, 309–368.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **144**, 285–307.

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society* **98**, 39–54.

Ghosh, S., Hastie, T. & Owen, A. B. (2022a). Backfitting for large scale crossed random effects regressions. *The Annals of Statistics* **50**, 560 – 583.

Ghosh, S., Hastie, T. & Owen, A. B. (2022b). Scalable logistic regression with crossed random effects. *Electronic Journal of Statistics* **16**, 4604 – 4635.

Gong, G. & Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: Theory and applications. *The Annals of Statistics* **9**, 861–869.

Good, I. J. & Gaskins, R. A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58**, 255–277.

Jochmans, K. & Otsu, T. (2019). Likelihood corrections for two-way models. *Annals of Economics and Statistics* , 227–242.

Jørgensen, B. (1993). The rules of conditional inference: Is there a universal definition of nonformation?,. *Bull. Int. Statist. Inst. 55* **2**, 323 – 340.

Jørgensen, B. & Labouriau, R. (2012). *Exponential Families and Theoretical Inference*, vol. 52 of *Monografías de Matemática*. Rio de Janeiro: Instituto de Matematica Pura e Aplicada (IMPA).

Kenne Pagui, E. C., Salvan, A. & Sartori, N. (2017). Median bias reduction of maximum likelihood estimates. *Biometrika* **104**, 923–938.

Kosmidis, I. & Firth, D. (2020). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika* **108**, 71–82.

Kosmidis, I., Kenne Pagui, E. C. & Sartori, N. (2020). Mean and median bias reduction in generalized linear models. *Statistics and Computing* **30**, 43–59.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics* **95**, 391–413.

Lancaster, T. (2002). Orthogonal parameters and panel data. *The Review of Economic Studies* **69**, 647–666.

Lee, S. M. S. & Young, G. A. (2005). Parametric bootstrapping with nuisance parameters. *Statist. Probab. Lett.* **71**, 143–153.

Leng, X., Mao, J. & Sun, Y. (2023). Debiased inference for dynamic nonlinear models with two-way fixed effects. ArXiv:2305.03134.

Lunardon, N. (2018). On bias reduction and incidental parameters. *Biometrika* **105**, 233–238.

Neyman, J. & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32.

Pace, L. & Salvan, A. (1997). *Principles of Statistical Inference from a neo-Fisherian Perspective*. Singapore: World Scientific.

Pierce, D. A. & Bellio, R. (2017). Modern likelihood-frequentist inference. *International Statistical Review* **85**, 519–541.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika* **43**, 353–360.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Studies in mathematical psychology. Danmarks Paedagogiske Institut.

Sartori, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika* **90**, 533–549.

Severini, T. (2000). *Likelihood Methods in Statistics*. Oxford: Oxford University Press.

Severini, T. A. (2005). *Elements of Distribution Theory*. Cambridge: Cambridge University Press.

Skovgaard, I. M. (1990). On the density of minimum contrast estimators. *The Annals of Statistics* **18**, 779–789.

Skovgaard, I. M. (1996). An explicit large-deviation approximation to one-parameter tests. *Bernoulli* **2**, 145 – 165.

Sur, P. & Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences* **116**, 14516–14525.

Tukey, J. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics* **29**, 614.

Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics* **9**, 60 – 62.

Young, G. A. (2009). Routes to higher-order accuracy in parametric inference. *Australian & New Zealand Journal of Statistics* **51**, 115–126.

Young, G. A. & Smith, R. L. (2005). *Essentials of Statistical Inference*. Cambridge: Cambridge University Press.

Zhao, Q. & Candès, E. J. (2022). An adaptively resized parametric bootstrap for inference in high-dimensional generalized linear models. Preprint, No: SS-2022-0296.

Zhao, Q., Sur, P. & Candès, E. J. (2020). The asymptotic distribution of the mle in high-dimensional logistic models: Arbitrary covariance. *Bernoulli* **28**, 1835–1861.

Zhu, Y. & Reid, N. (1994). Information, ancillarity, and sufficiency in the presence of nuisance parameters. *The Canadian Journal of Statistics* **22**, 111–123.