

UNIVERSITÀ DEGLI STUDI DI PADOVA

CORSO DI LAUREA MAGISTRALE
IN SCIENZE STATISTICHE



TESI DI LAUREA

**VEROSIMIGLIANZE INTEGRATE IN
MODELLI PER DATI BINARI
STRATIFICATI**

RELATORE: Prof. Nicola Sartori

Dipartimento di Scienze Statistiche

CORRELATORE: Prof. Ruggero Bellio

Dipartimento di Scienze Economiche e Statistiche

Università di Udine

LAUREANDO: Enrico Gasparin

MATRICOLA N° 1065580

ANNO ACCADEMICO 2014/2015

Indice

Introduzione	5
1 L'inferenza di verosimiglianza	9
1.1 Introduzione	9
1.2 Specificazione del modello	9
1.3 La funzione di verosimiglianza	11
1.4 Problema regolare di stima	12
1.5 Quantità di verosimiglianza	13
1.6 Teoria esatta ed asintotica della verosimiglianza	14
1.6.1 Trasformazioni biettive	14
1.6.2 Diseguaglianza di Wald e identità di Bartlett	14
1.6.3 Proprietà asintotiche dello stimatore di massima verosimiglianza	15
1.7 Inferenza basata sulla verosimiglianza	16
1.8 Parametri di disturbo	17
1.9 Verosimiglianza profilo e sue modifiche	19
1.9.1 Verosimiglianza profilo modificata	21
1.10 Verosimiglianza integrata	22
1.10.1 Parametrizzazione Zero Score Expectation	23
1.10.2 Proprietà della verosimiglianza integrata	24
1.10.3 Verosimiglianza integrata per campioni stratificati	26
2 Modello logistico per campioni stratificati	33
2.1 Introduzione	33
2.2 Campioni stratificati	33

2.3	Modelli per dati binari	35
2.4	Verosimiglianza per dati binari	36
2.4.1	Funzione di verosimiglianza e log-verosimiglianza . . .	37
2.4.2	Verosimiglianza profilo	38
2.4.3	Verosimiglianza profilo modificata	38
2.4.4	Verosimiglianze integrate	40
3	Studi di simulazione	45
3.1	Introduzione	45
3.2	Aspetti computazionali	47
3.3	Risultati	49
	Conclusioni	59
	Bibliografia	63
	Appendice: Codice	65

Introduzione

Nei problemi di inferenza statistica accade spesso che l'interesse venga riposto solamente su alcuni aspetti del fenomeno oggetto di studio. In un contesto di modellazione di tipo parametrico, questo si traduce nella bipartizione del parametro globale in parametro di interesse e parametro di disturbo. Le procedure inferenziali basate sulla verosimiglianza mantengono comunque le usuali proprietà anche in presenza di un parametro di disturbo finché alcune condizioni di regolarità vengono soddisfatte. Tra queste vi è quella che afferma che la dimensione del parametro non deve dipendere dalla numerosità campionaria. Tale condizione è spesso violata quando si considerano campioni stratificati. Quando l'interesse è posto negli aspetti comuni a tutti gli strati della variabilità del fenomeno oggetto di studio, la modellazione di questo tipo di dati prevede, talvolta, l'introduzione di un parametro proprio di ogni strato per cogliere l'eterogeneità non osservata. Così facendo si introduce nel modello un parametro di disturbo la cui dimensione dipende dalla numerosità campionaria attraverso il numero degli strati. In questo contesto, è noto in letteratura che la verosimiglianza profilo cessa di essere uno strumento valido per l'inferenza. Ciò è dovuto al fatto che la distorsione della sua funzione punteggio aumenta con il crescere del numero degli strati, rendendo lo stimatore del parametro di interesse distorto e non consistente. Per far fronte a questo problema sono state proposte varie soluzioni. Un esempio è la verosimiglianza profilo modificata di Barndorff-Nielsen (1980,1983) che riduce la distorsione della funzione punteggio attraverso l'introduzione di un fattore di modificazione nella funzione di verosimiglianza profilo. Un'alternativa per l'inferenza in presenza di un parametro di disturbo è la verosimiglianza integrata, oggetto di questa tesi. Attraverso l'integrazione della funzione

di verosimiglianza rispetto ad una funzione peso costruita opportunamente per il parametro di disturbo, è possibile ottenere procedure inferenziali per il parametro di interesse che godono di buone proprietà.

Nel Capitolo 1 si mostreranno dapprima i concetti base dell'inferenza di verosimiglianza. Successivamente si introdurrà il problema dei parametri di disturbo incidentali e di come le procedure basate sulla verosimiglianza profilo falliscono per via della distorsione della sua funzione punteggio. Si accennerà brevemente alla verosimiglianza profilo modificata, soluzione per ovviare a tale problema. Infine, ampio spazio sarà dedicato alla verosimiglianza integrata. Si illustreranno le sue proprietà e di come esse dipendono dalla funzione peso associata al parametro di disturbo. In particolare si mostrerà come la verosimiglianza integrata possa essere approssimata dalla verosimiglianza profilo modificata e, quindi, di come goda delle stesse proprietà asintotiche. Questo accade quando si riparametrizza il modello costruendo un parametro di disturbo *almeno ortogonale* al parametro di interesse ed usando per esso una funzione peso uniforme. Per tale scopo, poichè la costruzione di parametri ortogonali è spesso difficoltosa, verrà illustrata la parametrizzazione "Zero-Score-Expectation" (ZSE): essa permette infatti di giungere ad un parametro di disturbo fortemente *disconnesso* dal parametro di interesse per il quale specificare la funzione peso uniforme. Infine verranno illustrate due alternative alla riparametrizzazione del modello attraverso il metodo ZSE. Esse specificano due funzioni peso costruite a partire dall'espressione della distorsione della funzione punteggio della verosimiglianza profilo con il fine di ridurre la distorsione.

Nel Capitolo 2 si introdurranno le principali caratteristiche dei campioni stratificati e di come usualmente si modella l'eterogeneità non osservata per questo tipo di dati. Inoltre verranno esplicitate tutte le quantità introdotte a livello teorico nel Capitolo 1 per il modello logistico per dati binari stratificati.

Infine, il Capitolo 3 sarà dedicato al confronto tra i metodi di stima basati sulla verosimiglianza profilo, sulla verosimiglianza profilo modificata, sulla verosimiglianza integrata costruita attraverso la parametrizzazione ZSE e sulla verosimiglianza integrata costruita con la *robust prior* e la *normal prior*. Verranno anche evidenziati alcuni elementi computazionali cruciali

per lo studio di simulazione condotto.

Capitolo 1

L'inferenza di verosimiglianza

1.1 Introduzione

In questo primo capitolo si introdurranno brevemente i concetti fondamentali della teoria della verosimiglianza fisheriana. Quindi, il punto di vista adottato per l'inferenza è quello frequentista: i dati sono delle realizzazioni di variabili casuali e le proprietà delle procedure inferenziali sono determinate sulla base del principio del campionamento ripetuto.

Nei primi paragrafi si dirà cos'è un modello statistico parametrico e si definiranno le principali quantità di verosimiglianza, le loro proprietà ed il loro utilizzo per fini inferenziali. Successivamente si introdurrà il problema dei parametri di disturbo e si illustreranno brevemente alcune soluzioni ad esso come la verosimiglianza profilo e la verosimiglianza profilo modificata. Infine si tratterà con maggior dettaglio come la verosimiglianza integrata possa essere un metodo alternativo alla risoluzione di tale problema.

1.2 Specificazione del modello

Il principale obiettivo dell'inferenza statistica è quello di individuare il vero ed ignoto modello probabilistico generatore dei dati, p_Y^0 , a partire dall'osservazione di un campione di osservazioni empiriche $y = (y_1, \dots, y_n) \in \mathcal{Y}$, dove \mathcal{Y} rappresenta lo spazio campionario.

Sia \mathcal{F} la famiglia di distribuzioni di probabilità coerenti qualitativamente con l'osservazione di y , realizzazione della variabile casuale Y . La famiglia \mathcal{F} è detta modello statistico e, nel momento in cui $p_Y^0 \in \mathcal{F}$, il modello è detto correttamente specificato. Fisher (1922) mette in evidenza come il processo di individuazione si articola in tre fasi:

1. specificazione del modello statistico in modo da cogliere le caratteristiche salienti dei dati;
2. individuazione tramite procedure statistiche di $p_Y^0(y) \in \mathcal{F}$;
3. studio della sensibilità della scelta del modello alla variabilità campionaria dei dati.

Nella specificazione del modello entra in gioco anche la quantità di informazione disponibile a priori riguardante il fenomeno oggetto di studio. Questo porta a diversi modi di specificare il modello:

- **specificazione parametrica:** si ha a disposizione una quantità di informazione che permette di effettuare delle ipotesi stringenti sul modello. Si ottiene che gli elementi di \mathcal{F} sono indicizzabili tramite un numero finito di parametri p reali, ossia:

$$\mathcal{F} = \{p_Y(y, \theta) : \theta \in \Theta \subseteq \mathbb{R}^p\},$$

dove Θ rappresenta lo spazio parametrico in cui può variare il parametro θ e $p_Y(y, \theta)$ è detta funzione del modello. Se il modello è correttamente specificato si ha $p_Y^0(y) = p_Y(y; \theta^0)$ per un valore $\theta^0 \in \Theta$, con θ^0 detto vero valore del parametro;

- **specificazione semi-parametrica:** si ha informazione riguardante solo alcuni aspetti del fenomeno. Gli elementi all'interno di \mathcal{F} vengono quindi individuati sia attraverso una componente parametrica che una non parametrica, ossia:

$$\mathcal{F} = \{p_Y(y, \theta) : \theta \in \Theta\},$$

dove $\theta = (\tau, h(\cdot))$, con $\tau \in T \subseteq \mathfrak{R}^k$, mentre l'insieme di possibili specificazioni della funzione $h(\cdot)$ non è indicizzabile con un numero finito di parametri reali;

- **specificazione non parametrica:** \mathcal{F} è costituita da tutte le distribuzioni di probabilità aventi supporto coerente con il fenomeno oggetto di studio.

Si vuol far notare come, al diminuire dell'informazione a priori, la specificazione non parametrica permette una maggior flessibilità rispetto a quella parametrica. Chiaramente è necessario tenere a mente il *trade-off* che esiste tra flessibilità e quantità di informazione che si riesce ad estrarre dai dati. Nel seguito si prenderà in considerazione la sola specificazione parametrica, non considerando problemi legati all'indentificabilità e alla non corretta specificazione del modello.

1.3 La funzione di verosimiglianza

Dato un modello statistico parametrico $\mathcal{F} = \{p_Y(y, \theta) : \theta \in \Theta \subseteq \mathfrak{R}^p\}$ per i dati $y \in \mathcal{Y}$, si può considerare $p_Y(y, \theta)$ come funzione solamente di θ , con y fissato al valore osservato. La funzione che ne deriva, $L : \Theta \rightarrow \mathfrak{R}^+$, definita come

$$L(\theta) \equiv p_Y(y; \theta) = c(y)p(y; \theta) \quad (1.1)$$

è detta **funzione di verosimiglianza**, dove il termine moltiplicativo ed indipendente da θ , indicato da $c(y)$ in (1.1), può essere trascurato, senza alterare le conclusioni inferenziali. Grazie alla sua interpretazione essa può essere utilizzata per l'inferenza sul parametro di interesse θ : presi due valori $\theta', \theta'' \in \Theta$, alla luce delle osservazioni y , il valore θ' è più *verosimile* di θ'' se $L(\theta') > L(\theta'')$. Detto in altro modo, è più verosimile che i dati siano stati generati dal modello individuato da θ' rispetto a quello individuato da θ'' .

La trasformazione logaritmica della funzione di verosimiglianza, più agevole da trattare, prende il nome di **funzione di log-verosimiglianza**:

$$l(\theta) = \log L(\theta) = c'(y) + \log p(y; \theta) \quad (1.2)$$

dove, se $L(\theta) = 0$, si definisce $l(\theta) = -\infty$. Come prima, il trascurare un termine additivo ed indipendente da θ , indicato da $c'(y)$ in (1.2), non altera le conclusioni inferenziali. Data la monotonicità della trasformazione, le conclusioni inferenziali che si ottengono a partire da (1.1) sono le stesse a cui si giunge usando (1.2).

Data l'interpretazione della funzione di verosimiglianza, risulta logico definire la **stima di massima verosimiglianza** come quel valore $\hat{\theta} \in \Theta$ tale che $L(\hat{\theta}) \geq L(\theta)$ per ogni $\theta \in \Theta$. Per la monotonicità della trasformazione logaritmica, $\hat{\theta}$ sarà un massimo anche della funzione di log-verosimiglianza.

1.4 Problema regolare di stima

Le proprietà delle procedure inferenziali basate sulla funzione di verosimiglianza si possono ottenere sotto alcune ipotesi, che definiscono il cosiddetto **problema regolare di stima** (Azzalini, 2001, Paragrafo 3.2.3). Preso un modello statistico parametrico, tali ipotesi sono:

- presenza di verosimiglianza regolare, ossia:
 - lo spazio campionario \mathcal{Y} non dipende dal parametro θ ;
 - lo spazio parametrico Θ è un insieme aperto;
 - la funzione di log-verosimiglianza è differenziabile almeno tre volte con continuità;
- il modello è identificabile (esiste cioè un'associazione biunivoca tra gli elementi di \mathcal{F} e Θ) e correttamente specificato;
- $\dim(\Theta)$ non dipende da n , numerosità campionaria;

- è possibile scambiare l'operazione di integrazione rispetto a y con quella di derivazione rispetto a θ .

1.5 Quantità di verosimiglianza

In un modello con verosimiglianza regolare, le informazioni riguardanti l'andamento della funzione di verosimiglianza sono contenute in $\hat{\theta}$ e nelle derivate parziali di $l(\theta)$ rispetto alle componenti di θ .

Si definisce quindi la **funzione punteggio** o **funzione score** il vettore delle derivate parziali prime della funzione di log-verosimiglianza e si indica con

$$l_{\theta}(\theta) = l_{\theta}(\theta; y) = \frac{\partial}{\partial \theta} l(\theta) = (l_{\theta_1}(\theta), \dots, l_{\theta_p}(\theta))^T,$$

dove $l_{\theta_r} = \frac{\partial}{\partial \theta_r} l(\theta)$. In un modello con verosimiglianza regolare, risulta logico pervenire alla stima di massima verosimiglianza individuando la soluzione delle **equazioni di verosimiglianza**, definite come:

$$l_{\theta}(\theta) = 0.$$

L'**informazione osservata** è definita come la matrice $p \times p$ delle derivate parziali seconde cambiate di segno e si indica come:

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

La sua interpretazione è la seguente: *maggiore è $j(\hat{\theta})$ in un senso matriciale, tanto più nettamente si distinguono nello spazio parametrico regioni con elevata verosimiglianza da regioni con modesta verosimiglianza* (Pace e Salvan, 2001, Paragrafo 3.3).

La matrice di **informazione attesa** è invece definita come

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

1.6 Teoria esatta ed asintotica della verosimiglianza

Si riportano nel seguito brevemente le principali proprietà della funzione di verosimiglianza. Per maggiori approfondimenti e per le dimostrazioni si rimanda a Pace e Salvan, 2001, Capitoli 5 e 6.

1.6.1 Trasformazioni biettive

La funzione di verosimiglianza è invariante rispetto a trasformazioni biettive dei dati. Ciò significa che, dato il modello statistico $\mathcal{F} = \{p_Y(y, \theta) : \theta \in \Theta \subseteq \mathbb{R}^p\}$, il modello statistico $\mathcal{F}_T = \{p_T(t, \theta) : \theta \in \Theta \subseteq \mathbb{R}^p\}$, ottenuto per $t = g(y)$, produce una verosimiglianza equivalente a quella ottenuta a partire da \mathcal{F} .

La funzione di verosimiglianza è invariante rispetto a riparametrizzazioni del modello: sia \mathcal{F} un modello statistico parametrico con parametro θ e sia $\psi = \psi(\theta)$ una trasformazione biettiva del parametro, con inversa $\theta = \theta(\psi)$. La funzione di verosimiglianza nella nuova parametrizzazione, $L^\Psi(\psi)$, è legata alla funzione di verosimiglianza nella vecchia parametrizzazione, $L^\Theta(\theta)$, dalla relazione

$$L^\Psi(\psi) = L^\Theta(\theta(\psi)).$$

Segue che la stima di massima verosimiglianza di ψ si può ottenere a partire da quella per θ secondo la seguente relazione:

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

Quest'ultima proprietà è nota con il nome di **equivarianza** della stima di massima verosimiglianza.

1.6.2 Diseguaglianza di Wald e identità di Bartlett

La diseguaglianza di Wald, sotto deboli condizioni di regolarità, permette di dimostrare che lo stimatore di massima verosimiglianza, $\hat{\theta}_n(Y)$, è consistente. Ciò torna utile nel momento in cui si vogliono determinare

le proprietà asintotiche delle procedure d'inferenza basate sulla funzione di verosimiglianza. La disuguaglianza di Wald afferma che:

$$E_{\theta}\{l(\theta; Y)\} > E_{\theta}\{l(\theta'; Y)\}, \quad \theta' \neq \theta.$$

Sotto le ipotesi di problema regolare di stima si ha che:

- **prima identità di Bartlett:**

$$E_{\theta}\{l_{\theta}(\theta; Y)\} = 0 \text{ per ogni } \theta \in \Theta;$$

- **seconda identità di Bartlett:**

$$\text{Var}_{\theta}\{l_{\theta}(\theta; Y)\} = E_{\theta}\{-l_{\theta\theta}(\theta; Y)\} = i(\theta) \text{ per ogni } \theta \in \Theta.$$

1.6.3 Proprietà asintotiche dello stimatore di massima verosimiglianza

Sfruttando i risultati illustrati in precedenza si elencano le seguenti proprietà relative allo stimatore di massima verosimiglianza per $n \rightarrow \infty$, valide sotto campionamento casuale semplice:

- **consistenza:**

$$\hat{\theta}_n(Y) \xrightarrow{p} \theta$$

sotto θ , vero valore del parametro;

- **normalità asintotica:**

$$\hat{\theta}_n(Y) \xrightarrow{d} \mathcal{N}_p(\theta, i(\theta)^{-1})$$

sotto θ , vero valore del parametro.

Per maggiori dettagli si veda Pace e Salvan, 2001, Paragrafo 6.2 e Pace e Salvan, 1997, Paragrafo 3.4.

1.7 Inferenza basata sulla verosimiglianza

Data la consistenza dello stimatore di massima verosimiglianza, esso si pone come soluzione al problema di individuazione dell'elemento più *verosimile* alla luce dell'informazione campionaria all'interno di \mathcal{F} . Tale elemento sarà quello indicizzato da $\hat{\theta}$, stima di massima verosimiglianza. La teoria della verosimiglianza, quindi, fornisce strumenti per i problemi di **stima puntuale**, ma anche per i problemi di **stima intervallare** e di **verifica di ipotesi**. Questo è possibile in quanto, a partire dalle quantità di verosimiglianza, si possono ottenere delle quantità pivotali approssimate, ossia delle quantità che sono funzioni dei dati e del parametro, ma la cui distribuzione asintotica è nota indipendentemente dal valore del parametro.

La statistica **log-rapporto di verosimiglianza**

$$W(\theta) = 2\{l(\hat{\theta}) - l(\theta)\}$$

può essere utilizzata sia per l'individuazione di regioni di confidenza sia per la verifica di ipotesi. Nello specifico, poiché $W(\theta) \overset{\sim}{\sim} \chi_p^2$ sotto θ , una regione di confidenza per il parametro θ di livello nominale $1 - \alpha$ ha forma:

$$\hat{\Theta}(y) = \{\theta \in \Theta : W(\theta) < \chi_{p;1-\alpha}^2\}.$$

Per verificare invece il sistema di ipotesi $H_0 : \theta = \theta_0$ contro l'alternativa $H_1 : \theta \neq \theta_0$, si può sfruttare il fatto che valori grandi di $W(\theta_0)$ sono critici per l'ipotesi nulla. Il test $W(\theta_0)$ prende il nome di **test di Wilks** ed è distribuito asintoticamente come un chi-quadrato con p gradi di libertà, dove p è come al solito la dimensione del parametro.

Ulteriori statistiche basate su quantità di verosimiglianza, equivalenti asintoticamente a $W(\theta_0)$ sono il **test di Rao** e il **test di Wald** che rispettivamente sono definiti come:

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

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

Il loro utilizzo è il medesimo di $W(\theta)$.

Nel momento in cui $p = 1$, si possono condurre test di ipotesi per $H_0 : \theta = \theta_0$ contro alternative unilaterali $H_1^{dx} : \theta > \theta_0$ e $H_1^{sx} : \theta < \theta_0$. Le statistiche utili a questo scopo sono

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

$$r_u(\theta_0) = l_\theta(\theta_0) \sqrt{i(\theta_0)},$$

$$r_e(\theta_0) = (\hat{\theta} - \theta_0) \sqrt{i(\theta_0)}.$$

Tutte e tre hanno distribuzione asintotica nulla normale standard, sotto θ_0 .

Benchè le statistiche basate su $W_e(\theta)$ abbiano il pregio dell'immediata interpretabilità, spesso portano ad includere nelle regioni di confidenza elementi non appartenenti allo spazio parametrico. Inoltre tali statistiche non sono invarianti rispetto alla parametrizzazione del modello. Per quanto riguarda le statistiche basate su $W_u(\theta)$, esse non richiedono il calcolo della stima di massima verosimiglianza ma sono spesso instabili numericamente. Segue che le statistiche da preferire sono generalmente quelle basate sul log-rapporto di verosimiglianza.

1.8 Parametri di disturbo

Accade spesso nei modelli statistici parametrici che non tutte le componenti del parametro $\theta \in \Theta \subseteq \mathfrak{R}^p$ abbiano la stessa importanza da un punto di vista inferenziale. Capita quindi che il parametro θ possa essere scomposto in due componenti, cioè $\theta = (\psi, \lambda)$. L'interesse primario è focalizzato sul parametro ψ , k -dimensionale, con $1 \leq k \leq p$. Il parametro λ , con $p - k$ componenti comunque ignote, serve per garantire una maggior flessibilità al modello e prende il nome di **parametro di disturbo**. Se il vero valore del parametro di disturbo, λ^0 , fosse noto, per l'inferenza su ψ si utilizzerebbe la funzione di verosimiglianza $L(\psi) = L(\psi, \lambda^0)$, del sottomodello con λ noto. Il punto è che il parametro λ è ignoto e deve essere comunque stimato.

Una soluzione è quella di ricorrere ad una **pseudo-verosimiglianza** per ψ , cioè una funzione di ψ che possa svolgere il ruolo di $L(\psi, \lambda^0)$. Si vuole,

inoltre, che tale pseudo-verosimiglianza mantenga le proprietà usuali elencate in precedenza della funzione di verosimiglianza. Si possono verificare due situazioni:

1. la pseudo-verosimiglianza è una verosimiglianza propria, e ne mantiene le usuali proprietà, nel momento in cui è possibile ottenere un modello ridotto a partire da quello originale in cui le funzioni di densità dipendono solo dal parametro di interesse ψ . Si riesce cioè a giungere ad una fattorizzazione della verosimiglianza del tipo

$$L(\psi, \lambda) = L_1(\psi)L_2(\psi, \lambda),$$

e l'informazione su ψ in $L_2(\psi, \lambda)$ è, secondo opportune definizioni, trascurabile;

2. non esiste un modo per far dipendere le funzioni di densità dal solo parametro di interesse e le proprietà della verosimiglianza che ne deriva devono essere studiate di caso in caso. La verosimiglianza profilo e le sue modifiche rientrano in questo gruppo e verranno introdotte nel seguito.

Per approfondimenti sulle tecniche di riduzione si rimanda a Pace e Salvani (1997, Capitolo 4).

Un altro aspetto fondamentale del parametro di disturbo è la sua dimensione. Può accadere, infatti, che la sua dimensione sia legata alla numerosità campionaria. Si parla in questo caso di **parametri incidentali** (Neyman e Scott, 1948). Venendo a mancare una delle condizioni di regolarità del problema di stima, le proprietà delle procedure inferenziali devono essere studiate di caso in caso. In termini matematici, preso ad esempio un campione causale semplice $y = (y_1, \dots, y_n)$ con funzione di densità congiunta:

$$p_Y(y; \theta) = \prod_{i=1}^n p_{Y_i}(y_i; \theta_i),$$

si ha che $\theta_i = (\psi, \lambda_i)$, $i = 1, \dots, n$. Si può notare come la dimensione del parametro sia legato alla dimensione del campione n .

Nel Capitolo 2 si introdurranno i campioni stratificati. Si potrà notare come in essi sia presente un problema di parametri incidentali nel momento in cui si considera una specificazione ad effetti fissi per cogliere le caratteristiche proprie di ciascuno strato.

Si illustrano ora alcune funzioni di pseudo-verosimiglianza che permettono di stimare il parametro di interesse ψ in presenza di parametri di disturbo.

1.9 Verosimiglianza profilo e sue modifiche

Dato un modello statistico parametrico, sia $\theta = (\psi, \lambda) \in \Theta \subseteq \mathfrak{R}^p$ il parametro, con ψ parametro di interesse k -dimensionale e λ parametro di disturbo $p-k$ -dimensionale. L'inferenza su ψ può essere basata sulla **verosimiglianza profilo** definita come:

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

dove $\hat{\lambda}_\psi$ rappresenta la stima di massima verosimiglianza di λ per ψ fissato. Si indica invece con $l_P(\psi) = \log L_P(\psi)$ la funzione di **log-verosimiglianza profilo**. In linea teorica una qualsiasi stima di λ andrebbe bene ma vi sono delle ragioni per ritenere che la stima di massima verosimiglianza di λ per ψ fissato, $\hat{\lambda}_\psi$, sia da preferire (Pace e Salvani, 1997, Paragrafo 4.6). Sotto condizioni di regolarità, $\hat{\lambda}_\psi$ è soluzione dell'equazione

$$l_\lambda(\psi, \lambda) = \partial l(\psi, \lambda) / \partial \lambda = 0,$$

dove $l_\lambda(\psi, \lambda)$ è il vettore $p-k$ -dimensionale delle derivate parziali prime della funzione di log-verosimiglianza.

Poichè la pseudo-verosimiglianza $L_P(\psi)$ non deriva direttamente da una funzione di densità, essa non è una verosimiglianza propria. Gode comunque di alcune proprietà desiderabili come:

- la stima di massima verosimiglianza profilo è la stessa della stima di massima verosimiglianza di ψ ottenuta a partire da $L(\psi, \lambda)$, $\hat{\psi}$, ossia

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

dove $\hat{\theta} = (\hat{\psi}, \hat{\lambda})$;

- la matrice di informazione osservata profilo è definita come:

$$j_P(\psi) = -\frac{\partial^2}{\partial\psi\partial\psi^T}l_P(\psi) = -\frac{\partial^2}{\partial\psi\partial\psi^T}l(\psi, \hat{\lambda}_\psi);$$

- dato che

$$\frac{\partial}{\partial\psi}l_P(\psi) = l_\psi(\psi, \hat{\lambda}_\psi) + l_\lambda(\psi, \hat{\lambda}_\psi)\frac{\partial}{\partial\psi}\hat{\lambda}_\psi,$$

la **funzione punteggio profilo** è definita come

$$\frac{\partial}{\partial\psi}l_P(\psi) = l_\psi(\psi, \hat{\lambda}_\psi).$$

Questo segue dal fatto che $l_\lambda(\psi, \hat{\lambda}_\psi) = 0$;

- il test **log-rapporto di verosimiglianza profilo**,

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

coincide con il test log-rapporto di verosimiglianza ottenuto a partire da $L(\psi, \lambda)$. Ne mantiene quindi le stesse proprietà sotto le usuali condizioni di regolarità. In particolare ha distribuzione asintotica nulla χ_k^2 , dove k è la dimensione del parametro di interesse.

La verosimiglianza profilo presenta anche dei problemi. La funzione punteggio non ha valore atteso nullo e sotto condizioni di regolarità il suo valore atteso è di ordine $O(1)$ (Pace e Salvan, 1997, Paragrafo 9.4). Nel caso di parametri incidentali invece il valore atteso della funzione punteggio può essere di ordine $O(n)$: questo può far venire meno la consistenza dello stimatore di ψ . Inoltre, se i dati contengono poca informazione riguardante i parametri di disturbo, non è ragionevole usare la verosimiglianza profilo in quanto ciò corrisponde ad assumere che λ sia uguale a $\hat{\lambda}_\psi$.

Sono state proposte di conseguenza delle modifiche alla verosimiglianza profilo al fine di sopperire alla mancanza di informazione riguardante i

parametri di disturbo e di ottenere delle stime migliori per il parametro di interesse.

1.9.1 Verosimiglianza profilo modificata

La verosimiglianza profilo modificata, come definita da Barndorff-Nielsen (1980, 1983), ha espressione

$$L_M(\psi) = L_P(\psi)M(\psi). \quad (1.3)$$

dove il fattore di modificazione $M(\psi)$ è definito come

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

dove nel primo determinante compare la derivata mista

$$l_{\lambda; \hat{\lambda}}(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\lambda}, a) = \frac{\partial^2}{\partial \lambda \partial \hat{\lambda}^T} l(\psi, \hat{\lambda}_\psi; \hat{\psi}, \hat{\lambda}, a),$$

in cui è richiesta l'individuazione di una statistica ancillare a , sia essa esatta o approssimata, al fine di costruire la statistica sufficiente minimale $(\hat{\psi}, \hat{\lambda}, a)$, attraverso cui riformulare la verosimiglianza. Una statistica a è detta ancillare se la sua distribuzione è indipendente dal parametro θ e se la statistica sufficiente minimale è in relazione biunivoca con $(\hat{\theta}, a)$, dove $\hat{\theta}$ è la stima di massima verosimiglianza di θ .

La verosimiglianza profilo modificata si pone come un'approssimazione ad una verosimiglianza marginale o condizionata, qualora esistano (Barndorff-Nielsen e Cox 1994, Paragrafo 8.2). Nel caso in cui non esistesse nè una verosimiglianza marginale nè una condizionata, la verosimiglianza profilo modificata permette di giungere a delle equazioni di stima con proprietà migliori rispetto alla verosimiglianza profilo. Uno degli aspetti critici resta il reperimento di una statistica ancillare esatta o approssimata, che può essere individuata con semplicità solo se il modello sottostante appartiene ad una famiglia esponenziale o ad una famiglia di gruppo.

Una soluzione per ovviare al problema di reperimento di un'opportuna statistica ancillare è quello di definire il fattore di modificazione come fatto in Severini (1998):

$$M(\psi) = \frac{|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|^{1/2}}{|I_{\lambda\lambda}(\hat{\psi}, \hat{\lambda}; \psi, \hat{\lambda}_\psi)|} \quad (1.4)$$

dove $(\hat{\psi}, \hat{\lambda})$ è la stima di massima verosimiglianza dei parametri, $j_{\lambda\lambda}$ è il blocco- $\lambda\lambda$ della matrice di informazione osservata e $I_{\lambda\lambda}$ è definito come

$$I_{\lambda\lambda}(\hat{\psi}, \hat{\lambda}; \psi, \hat{\lambda}_\psi) = E_{\psi_0, \lambda_0} \{ l_\lambda(\psi_0, \lambda_0) l_\lambda(\psi_1, \lambda_1) \} \Big|_{(\psi_0=\hat{\psi}, \lambda_0=\hat{\lambda}, \psi_1=\psi, \lambda_1=\hat{\lambda}_\psi)}.$$

La costruzione di una verosimiglianza profilo modificata usando il fattore di modificazione sopra esposto porta ad approssimare le derivate nello spazio campionario tramite dei valori attesi. Rimangono tuttavia inalterate le proprietà asintotiche, che sono paragonabili a quelle della profilo modificata di Barndorff-Nielsen (1980, 1983).

1.10 Verosimiglianza integrata

La verosimiglianza integrata si pone come alternativa alla profilo e sue modifiche per risolvere il problema dei parametri di disturbo. Riprendendo la notazione precedente, l'inferenza sul parametro di interesse ψ tramite la funzione di verosimiglianza integrata è basata sull'eliminare il parametro di disturbo λ tramite integrazione utilizzando una qualche funzione peso per λ .

Sia $\theta = (\psi, \lambda) \in \Theta \subseteq \mathfrak{R}^p$ il parametro di un modello statistico parametrico. Denotiamo con $\pi(\lambda|\psi)$ una funzione non negativa su Λ , spazio parametrico del parametro di disturbo, che prende il nome di **funzione di densità condizionata a priori**. La funzione di verosimiglianza integrata rispetto alla a priori π è definita come:

$$L_I(\psi) = \int_{\Lambda} L(\psi, \lambda) \pi(\lambda|\psi) d\lambda. \quad (1.5)$$

Si vuol far notare come non è necessario che $\pi(\cdot)$ sia una funzione di densità propria in quanto l'ottica adottata non è quella bayesiana.

Uno dei principali vantaggi è che la funzione di verosimiglianza integrata è sempre disponibile, cosa che non accade sempre per le verosimiglianze marginali e condizionate. Inoltre essa è basata su una media e non su una massimizzazione, con una conseguente maggior stabilità numerica. Resta comunque il problema della scelta della densità a priori, la quale può influenzare le proprietà della verosimiglianza integrata (Severini, 2007).

Per lo studio delle proprietà della verosimiglianza integrata è necessario introdurre il concetto di **parametri “disconnessi”** e di come si possa costruire un parametro di disturbo disconnesso dal parametro di interesse.

1.10.1 Parametrizzazione Zero Score Expectation

Supponiamo che tramite un’opportuna riparametrizzazione siamo giunti ad ottenere γ , parametro disconnesso dal parametro di interesse ψ . Diremo che γ è debolmente disconnesso da ψ se

$$\hat{\gamma}_\psi = \hat{\gamma} + O(n^{-1})$$

per variazioni di ψ del tipo $\psi = \hat{\psi} + O(n^{-1/2})$, dove $\hat{\psi}$ e $\hat{\gamma}$ sono le stime di massima verosimiglianza dei parametri. Parametri debolmente disconnessi si possono ottenere a partire dalla parametrizzazione ortogonale: dato $\theta = (\psi, \lambda)$, ψ e λ sono detti ortognali se il blocco- $\psi\lambda$ della matrice di informazione attesa è nullo, cioè $i_{\psi\lambda} = 0$. Segue che $\hat{\psi}$ e $\hat{\lambda}$ sono asintoticamente indipendenti (Pace e Salvan, 1997, Paragrafo 4.7). L’individuazione dei parametri ortogonali richiede la risoluzione di equazioni differenziali e ciò può essere complicato. Inoltre, non sempre la soluzione esiste, a meno che ψ non sia scalare (De Bin, Sartori e Severini, 2015).

Una possibile soluzione è quella di ricorrere alla parametrizzazione Zero-Score-Expectation (ZSE), descritta successivamente. Oltre ad essere più agevole da calcolare rispetto alla parametrizzazione ortogonale, tale parametrizzazione garantisce che il parametro di disturbo individuato sia fortemente disconnesso dal parametro di interesse. Nello specifico, per fortemente

disconnesso si intende che

$$\hat{\gamma}_\psi = \hat{\gamma} + O(n^{-1/2}),$$

per variazioni di ψ del tipo $\psi - \hat{\psi} = O(1)$.

La costruzione di un parametro di disturbo fortemente disconnesso dal parametro di interesse ψ si può ottenere come segue. Sia $\phi \equiv \phi(\psi, \lambda; \hat{\psi})$ un parametro che dipende dai dati tramite la stima di massima verosimiglianza di ψ , $\hat{\psi}$. La soluzione dell'equazione

$$E\{l_\lambda(\psi, \lambda); \hat{\psi}, \phi\} \equiv E_{\psi_0, \lambda_0}\{l_\lambda(\psi, \lambda)\}_{(\psi_0, \lambda_0) = (\hat{\psi}, \phi)} = 0$$

in ϕ , per $(\psi, \lambda, \hat{\psi})$ fissato, dà $\phi(\psi, \lambda; \hat{\psi})$. Ques'ultima funzione è detta **parametrizzazione ZSE**. Si dimostra che $\hat{\phi} = \hat{\lambda}$ e che ϕ è fortemente disconnesso da ψ (Severini, 2007).

1.10.2 Proprietà della verosimiglianza integrata

Come accennato in precedenza, le proprietà della verosimiglianza integrata dipendono dalla scelta della a priori. In particolare, nel momento in cui si riesce a costruire un parametro di disturbo γ , almeno debolmente disconnesso dal parametro di interesse ψ , la distribuzione a priori dovrebbe essere scelta in modo che i due parametri siano indipendenti. Questo permette di elencare le seguenti proprietà della verosimiglianza integrata:

- sia γ un parametro per cui vale la fattorizzazione $L(\theta) = L_1(\psi)L_2(\gamma)$. Se γ è un parametro almeno debolmente disconnesso da ψ , e la distribuzione a priori è scelta in modo tale per cui ψ e γ sono indipendenti, allora $L_1(\psi)$ può essere utilizzata per l'inferenza su ψ e coincide con la verosimiglianza integrata. Si può notare, inoltre, come una tale parametrizzazione e scelta della a priori rendano la verosimiglianza integrata indipendente dalla a priori utilizzata. Per maggiori dettagli si veda Severini (2007);

- si indichi con $l_I(\psi) = \log L_I(\psi)$ e con $l_{I\psi}(\psi) = \partial l_I(\psi)/\partial \psi$. In generale $E_\theta\{l_{I\psi}(\psi)\}$ e $E_\theta\{l_{I\psi}(\psi) + l_{I\psi}(\psi)l_{I\psi}(\psi)^T\}$ sono di ordine $O(1)$ per $n \rightarrow \infty$. Nel momento in cui si utilizza una parametrizzazione tale per cui γ è debolmente disconnesso da ψ , e la priori è tale per cui i parametri sono indipendenti, si ha che $E_\theta\{l_{I\psi}(\psi)\}$ è di ordine $O(n^{-1})$. Se invece γ e ψ sono fortemente disconnessi, e come al solito la priori è tale per cui i parametri sono indipendenti, si ha anche che $E_\theta\{l_{I\psi}(\psi) + l_{I\psi}(\psi)l_{I\psi}(\psi)^T\}$ è di ordine $O(n^{-1})$. Per maggiori dettagli si veda Severini (1998).

Tutto ciò suggerisce quindi di costruire un parametro di disturbo fortemente disconnesso dal parametro di interesse e successivamente di utilizzare una densità a priori tale per cui i due parametri siano indipendenti. Di conseguenza, ottenuto, ad esempio, con la parametrizzazione ZSE il parametro di disturbo ϕ , si può specificare per esso una a priori uniforme. Facendo così si può mettere in evidenza il legame esistente tra la verosimiglianza integrata, la profilo modificata e sue approssimazioni.

Sia $L_M(\psi)$ la verosimiglianza profilo modificata secondo l'espressione (1.3) e sia $\bar{L}_M(\psi)$ la verosimiglianza profilo modificata usando il fattore di modificazione (1.4). Nell'Appendice 2 di Severini (2007), si mostra che, presi due parametri ψ e γ fortemente disconnessi e fissata una a priori tale per cui essi siano indipendenti, vale

$$L_I(\psi) = c\bar{L}_M(\psi)\{1 + O(n^{-1/2})O(|\psi - \hat{\psi}|\}\}, \quad (1.6)$$

dove c è una costante indipendente da ψ . Di conseguenza, dalla relazione

$$\bar{L}_M(\psi) = cL_M(\psi)\{1 + O(n^{-1/2})O(|\psi - \hat{\psi}|\}\},$$

si ha che

$$L_I(\psi) = cL_M(\psi)\{1 + O(n^{-1/2})O(|\psi - \hat{\psi}|\}\},$$

dove c è una ancora una costante indipendente da ψ .

Poichè la verosimiglianza integrata può essere approssimata da \bar{L}_M , essa gode delle stesse proprietà asintotiche. Ad esempio, come evidenziato prima, la verosimiglianza integrata è caratterizzata dalla non distorsione della

funzione punteggio e della matrice di informazione attesa. L'uso di una a priori uniforme rende inoltre la verosimiglianza integrata invariante rispetto a parametrizzazioni. Il fatto poi che la priori sia scelta in modo tale per cui il parametro di interesse, ψ , ed il parametro di disturbo, ϕ , fortemente disconnessi, siano indipendenti, rende la verosimiglianza integrata poco sensibile alla scelta della a priori. Questo accade poichè l'approssimazione (1.6) vale per ogni scelta di $\pi(\phi)$. Per maggiori dettagli si rimanda a Severini (2007, Paragrafo 5).

1.10.3 Verosimiglianza integrata per campioni stratificati

Consideriamo ora un campione stratificato $y = (y_1, \dots, y_q)$, dove y_i è realizzazione di una variabile casuale Y_i di dimensione m_i con densità $p_i(y_i; \psi, \lambda_i)$. Supponiamo che Y_1, \dots, Y_q siano indipendenti ed indichiamo con ψ il parametro di interesse e con $\lambda = (\lambda_1, \dots, \lambda_q)$ il parametro di disturbo. Assumiamo inoltre che tutti i parametri di disturbo abbiano lo stesso spazio parametrico Λ .

In un contesto del genere, può accadere che il numero di parametri di disturbo sia molto più grande rispetto alla dimensione dello strato. Come anticipato, questo porta a risultati scadenti se si utilizzano i metodi di verosimiglianza legati alla semplice verosimiglianza profilo. Nel seguito si mostreranno le **proprietà asintotiche a due indici** della verosimiglianza integrata. Essa può essere quindi una valida soluzione al problema dei parametri incidentali o, più in generale, quando il numero degli strati è molto più grande della dimensione dello strato.

Proprietà asintotiche a due indici

Sia $L^{(i)}(\psi, \lambda_i) = p_i(y_i; \psi, \lambda_i)$ il contributo dell' i -esimo strato alla verosimiglianza. Poichè abbiamo assunto che gli strati sono tra loro indipendenti, si ha che

$$L(\psi, \lambda) = \prod_{i=1}^q L^{(i)}(\psi, \lambda_i).$$

Di conseguenza, la log-verosimiglianza diventa

$$l(\psi, \lambda) = \sum_{i=1}^q l^{(i)}(\psi, \lambda_i),$$

dove $l^{(i)}(\psi, \lambda_i) = \log L^{(i)}(\psi, \lambda_i)$. Indicando con $\hat{\lambda}_{i\psi}$ la stima di massima verosimiglianza di λ_i per ψ fissato, si può definire la log-verosimiglianza profilo come

$$l_p(\psi) = \sum_{i=1}^q l^{(i)}(\psi, \hat{\lambda}_{i\psi}).$$

Per semplicità di notazione si assume che il campione stratificato sia bilanciato, ossia $m_i = m$, per $i = 1, \dots, q$. I risultati di seguito elencati valgono anche se le dimensioni degli strati differiscono tra di loro, fintanto che $m_i = K_i m$, con $A \leq K_i \leq B$, dove A e B sono due costanti positive (Sartori, 2003). Questo fa sì che gli strati siano asintoticamente bilanciati, cioè che ogni m_i è di ordine $O(m)$.

Come anticipato, le proprietà della verosimiglianza integrata dipendono da come viene scelta la priori. Poiché abbiamo assunto che gli strati sono indipendenti, è logico assumere che anche i parametri di disturbo λ_i , propri di ogni strato, siano indipendenti. Questo porta a specificare una a priori per l'intero parametro di disturbo λ del tipo

$$\pi(\lambda|\psi) = \prod_{i=1}^q \pi(\lambda_i|\psi). \quad (1.7)$$

Di conseguenza la verosimiglianza integrata assume la forma

$$L_I(\psi) = \prod_{i=1}^q \int_{\Lambda} L^{(i)}(\psi, \lambda_i) \pi(\lambda_i|\psi) d\lambda_i. \quad (1.8)$$

L'utilizzo della parametrizzazione ZSE, illustrata in precedenza, permette di ridurre la distorsione della funzione punteggio. In particolare, si ottiene che

$$E_{\theta}\{l_{I\psi}(\psi)\} = O\left(\frac{q}{m}\right),$$

dove $l_{I\psi}(\psi) = \partial l_I(\psi)/\partial\psi$. Tale risultato è valido anche usando una parametrizzazione ortogonale. Viceversa, usando una qualsiasi altra parametrizzazione, la distorsione è di ordine $O(q)$. Questo è uno dei motivi per cui le proprietà asintotiche a due indici verranno ricavate ipotizzando l'utilizzo della parametrizzazione ZSE.

Siano $l_I^{(i)}(\psi)$ e $l_M^{(i)}(\psi)$ rispettivamente la log-verosimiglianza integrata e la log-verosimiglianza profilo modificata per lo strato i -esimo. Data l'indipendenza degli strati e dei parametri di disturbo segue che

$$l_I(\psi) = \sum_{i=1}^q l_I^{(i)}(\psi) \quad \text{e} \quad l_M(\psi) = \sum_{i=1}^q l_M^{(i)}(\psi).$$

Sia $\hat{\psi}_M$ il punto di massimo di $l_M(\psi)$. In Sartori (2003) si dimostra che, indicando con ψ il vero valore del parametro,

$$\hat{\psi}_M = \psi + O_p(1/\sqrt{mq}) \quad \text{se} \quad q/m^3 = o(1)$$

mentre $\hat{\psi}_M = \psi + O_p(1/m^2)$ altrimenti.

Sfruttando la relazione esistente tra la verosimiglianza integrata e la profilo modificata nel momento in cui si utilizza la parametrizzazione ZSE con una a priori uniforme, De Bin, Sartori e Severini (2015) determinano le seguenti proprietà della verosimiglianza integrata in un contesto asintotico a due indici.

- Per $\psi = \hat{\psi}_M + O_p(1/\sqrt{mq})$ e cioè per $q/m^3 = o(1)$ si hanno le seguenti approssimazioni:

$$l_I(\psi) = l_M(\psi) + O_p\left(\sqrt{\frac{q}{m^3}}\right) + O_p\left(\frac{1}{m}\right)$$

$$\frac{1}{\sqrt{mq}} l_{I\psi}(\psi) = \frac{1}{\sqrt{mq}} l_{M\psi}(\psi) + O_p\left(\sqrt{\frac{q}{m^3}}\right) + O_p\left(\frac{1}{m}\right)$$

$$\frac{1}{mq} l_{I\psi\psi}(\psi) = \frac{1}{mq} l_{M\psi\psi}(\psi) + O_p\left(\frac{1}{m}\right).$$

- Sia $\hat{\psi}_I$ il punto di massimo di $l_I(\psi)$. Si dimostra che lo stimatore $\hat{\psi}_I$ ha la stessa distribuzione asintotica di $\hat{\psi}_M$, se $q/m^3 = o(1)$ (De Bin, Sartori e Severini, 2015). In particolare, sfruttando le proprietà asintotiche di $\hat{\psi}_M$, descritte in Sartori (2003), e la relazione

$$\sqrt{mq}(\hat{\psi}_I - \psi) = \sqrt{mq}(\hat{\psi}_M - \psi) + O_p\left(\sqrt{\frac{q}{m^3}}\right) + O_p\left(\frac{1}{m}\right),$$

si ha che $\hat{\psi}_I = \psi + O_p(1/\sqrt{mq})$ se $q/m^3 = o(1)$ mentre $\hat{\psi}_I = \psi + O_p(1/m^2)$ altrimenti. Per confronto, lo stimatore di massima verosimiglianza, $\hat{\psi}$, soddisfa $\hat{\psi} = \psi + O_p(1/\sqrt{mq})$ se $q/m = o(1)$ e $\hat{\psi} = \psi + O_p(1/m)$ altrimenti. Questo mette in luce come, al fine di garantire la stessa precisione, lo stimatore di massima verosimiglianza necessita che la dimensione degli strati, m , cresca alla stessa velocità del numero degli strati, q . Lo stimatore $\hat{\psi}_I$ invece richiede che la dimensione degli strati cresca più velocemente di $q^{1/3}$. Quindi, in contesti in cui il numero degli strati è molto maggiore della dimensione degli strati, lo stimatore che deriva dalla verosimiglianza integrata è migliore dello stimatore di massima verosimiglianza e paragonabile asintoticamente a quello che deriva dalla profilo modificata.

- Il test log-rapporto di verosimiglianza

$$W_I(\psi) = 2\{l_I(\hat{\psi}_I) - l_I(\psi)\},$$

ha distribuzione asintotica chi-quadrato con p gradi di libertà, dove p è l'usuale dimensione del parametro di interesse nel momento in cui $q/m^3 = o(1)$. Se $p = 1$,

$$R_I(\psi) = \text{sgn}(\hat{\psi}_I - \psi)\sqrt{W_I}$$

ha distribuzione asintotica normale standard, sempre se $q/m^3 = o(1)$. Per confronto, il test log-rapporto di verosimiglianza basato sulla profilo ha distribuzione asintotica chi-quadrato con p gradi di libertà se $q/m = o(1)$. Ancora una volta, le procedure basate sulle verosimiglian-

za integrata forniscono risultati migliori rispetto alla profilo quando la dimensione del parametro di disturbo risulta essere molto più grande della dimensione degli strati.

Alternative alla parametrizzazione ZSE

Come abbiamo visto, l'utilizzo della parametrizzazione ZSE assieme ad una a priori uniforme permette di giungere a procedure statistiche basate sulla verosimiglianza integrata che godono di buone proprietà asintotiche. Esistono comunque delle alternative che non richiedono di dover riparametrizzare il modello.

Arellano e Bonhomme (2009) propongono infatti delle **distribuzioni a priori** che dipendono dai dati, le quali permettono di ridurre comunque la distorsione della funzione punteggio, senza ricorrere a riparametrazioni. I due autori mostrano come la distorsione della funzione punteggio per l' i -esimo strato, per una qualsiasi distribuzione a priori, abbia espressione:

$$b_i(\psi_0) = \frac{\partial}{\partial \psi} \Big|_{\psi_0} \log \pi_i(\hat{\lambda}_{i\psi} | \psi) - \frac{\partial}{\partial \psi} \Big|_{\psi_0} \log(E_{\theta_0} \{-l_{\lambda_i \lambda_i}^{(i)}(\psi, \hat{\lambda}_{i\psi})\}) \\ \times [E_{\theta_0} \{[l_{\lambda_i}^{(i)}(\psi, \hat{\lambda}_{i\psi})]^2\}]^{-1/2},$$

dove $\theta_0 = (\psi_0, \lambda_0)$ è il vero valore del parametro e $\pi_i(\cdot)$ è la priori per l' i -esimo strato. Di conseguenza, una a priori riduce la distorsione nel momento in cui:

$$\frac{\partial}{\partial \psi} \Big|_{\psi_0} \log \pi_i(\hat{\lambda}_{i\psi} | \psi) = \frac{\partial}{\partial \psi} \Big|_{\psi_0} \log(E_{\theta_0} \{-l_{\lambda_i \lambda_i}^{(i)}(\psi, \hat{\lambda}_{i\psi})\}) \\ \times [E_{\theta_0} \{[l_{\lambda_i}^{(i)}(\psi, \hat{\lambda}_{i\psi})]^2\}]^{-1/2} \\ + O\left(\frac{1}{m}\right).$$

Sostituendo, ad esempio, la stima di massima verosimiglianza al vero parametro ignoto θ_0 , si possono stimare consistentemente le quantità attese $E_{\theta_0} \{-l_{\lambda_i \lambda_i}^{(i)}(\psi, \lambda_i)\}$ e $E_{\theta_0} \{[l_{\lambda_i}^{(i)}(\psi, \lambda_i)]^2\}$ con $\hat{E} \{-l_{\lambda_i \lambda_i}^{(i)}(\psi, \lambda_i)\}$ e $\hat{E} \{[l_{\lambda_i}^{(i)}(\psi, \lambda_i)]^2\}$.

Si ottiene che la seguente a priori riduce la distorsione:

$$\pi_i^R(\lambda_i|\psi) \propto \hat{E}\{-l_{\lambda_i\lambda_i}^{(i)}(\psi, \lambda_i)\}[\hat{E}\{[l_{\lambda_i}^{(i)}(\psi, \lambda_i)]^2\}]^{-1/2}. \quad (1.9)$$

Sempre in Arellano e Bonhomme (2009, Paragrafo 3.2) viene illustrata un'altra distribuzione a priori che riduce la distorsione della funzione punteggio basata sulla densità di una normale. Un vantaggio di questa a priori è che, a differenza di (1.9), è una distribuzione propria e quindi, almeno tecnicamente, può essere utilizzata in un contesto bayesiano. Tale a priori ha la seguente espressione:

$$\pi_i^N(\lambda_i|\psi) = \mathcal{N}(\hat{\lambda}_{i\psi}, \widehat{Var}(\hat{\lambda}_{i\psi})), \quad (1.10)$$

dove

$$\widehat{Var}(\hat{\lambda}_{i\psi}) = I_{\lambda_i\lambda_i}^2(\hat{\psi}, \hat{\lambda}_i; \psi, \hat{\lambda}_{i\psi})j_{\lambda_i\lambda_i}^{-2}(\psi, \hat{\lambda}_{i\psi})j_{\lambda_i\lambda_i}^{-1}(\hat{\psi}, \hat{\lambda}_i),$$

in cui $I_{\lambda_i\lambda_i}$ è lo stesso usato per la costruzione della profilo modificata con fattore di modificazione pari a (1.4), solo che riferito allo strato i -esimo. L'espressione per la varianza della densità (1.10) si può ottenere a partire da Pace e Salvan (2006, Formula (9)).

Al momento, non esistono confronti in letteratura tra la diverse verosimiglianze integrate illustrate in precedenza. Nei capitoli seguenti, tale confronto verrà presentato per modelli stratificati per dati binari.

Capitolo 2

Modello logistico per campioni stratificati

2.1 Introduzione

In questo capitolo si introdurranno le principali caratteristiche di un campione stratificato e le problematiche ad esso associate. Successivamente l'attenzione verrà posta nei modelli per dati binari ed in particolare nel modello logistico per campioni stratificati. Quindi, per questo specifico modello, verranno illustrate tutte le quantità introdotte a livello teorico nel Capitolo 1.

2.2 Campioni stratificati

Il campionamento stratificato è un metodo che permette di campionare le unità statistiche di una popolazione. Tale metodo consiste nel suddividere la popolazione in strati omogenei al loro interno e *disgiunti* tra di loro. Successivamente le unità statistiche vengono campionate all'interno di ciascuno strato. In questo modo si riescono, in generale, ad ottenere stime più precise sfruttando il fatto che si evita il problema del sovra-campionamento di alcuni tipi di unità statistiche, riducendo così l'errore campionario. Da un punto di vista econometrico i **dati di panel** costituiscono un campione stratificato,

in cui ciascuno strato è costituito spesso da una singola unità statistica. Si pensi, ad esempio, ai macro-panel in cui le unità statistiche sono solitamente nazioni oppure ai micro-panel in cui le unità statistiche sono spesso degli individui o delle famiglie. Nulla vieta comunque di *stratificare* la popolazione e di considerare le osservazioni all'interno di ogni strato come riferite alla stessa unità statistica. Uno dei principali pregi di questo tipo di dati è che contengono un'enorme quantità di informazione, in quanto, potendo avere più osservazioni per la stessa unità statistica, si riesce ad unire l'informazione cross-sezionale a quella longitudinale. La modellazione statistica di questo tipo di dati prevede spesso la specificazione di un effetto proprio per ogni singolo strato o unità statistica, qualora essa venga osservata più volte, con il fine di cogliere le sue caratteristiche non direttamente osservabili. La specificazione di questo effetto può avvenire in due modi. Si può considerare l'effetto proprio di ciascuno strato come una variabile casuale ottenendo il cosiddetto **modello ad effetti casuali**. Siccome tale effetto non è osservabile direttamente, risulta dunque necessario fare delle assunzioni riguardanti la sua distribuzione. Inoltre, affinché le procedure statistiche mantengano buone proprietà (es. consistenza), si richiede che tale effetto sia indipendente dalle variabili usate per cogliere l'eterogeneità osservabile. Quest'ultima assunzione è spesso poco realistica e difficile da verificare tramite opportuni test di endogeneità. Si tende quindi a preferire la specificazione che considera l'effetto proprio di ciascuno strato come un parametro fisso, ottenendo il **modello ad effetti fissi**. Così facendo si permette alle variabili esplicative di dipendere dall'effetto non osservabile introdotto, aggirando l'assunzione di esogeneità. Sorge però il problema dei parametri incidentali in quanto, per ogni singolo strato, c'è un parametro che deve essere stimato. Di conseguenza, non risultano soddisfatte le condizioni di regolarità del problema di stima (Paragrafo 1.4) poichè la dimensione dello spazio parametrico viene a dipendere dalla numerosità campionaria attraverso il numero degli strati. Logicamente, tale problema è più evidente nel momento in cui il numero degli strati è molto maggiore rispetto al numero di osservazioni che si effettuano per ogni strato. Risulta quindi necessario ricorrere a procedure statistiche che mantengono buone proprietà asintotiche anche in questi scenari, come ad

esempio la verosimiglianza profilo modificata e la verosimiglianza integrata, introdotte nel capitolo precedente.

2.3 Modelli per dati binari

Sia y_{ij} realizzazione della variabile casuale bernoulliana

$$Y_{ij} \sim \text{Be}(\pi_{ij}) \quad i = 1, \dots, q \quad j = 1, \dots, m,$$

dove q è il numero degli strati ed m è il numero di osservazioni per strato, assunto essere comune a tutti gli strati per semplicità di notazione. L'interesse sta nel modellare la probabilità

$$\pi_{ij} = \Pr(Y_{ij} = 1 | X_{ij} = x_{ij}) = E\{Y_{ij} | X_{ij} = x_{ij}\} = g(x_{ij}; \theta_i) \quad (2.1)$$

come funzione di un insieme di variabili esplicative, indicate con x_{ij} , e di un parametro $\theta_i = (\psi, \lambda_i)$, in cui ψ è il parametro di interesse e λ_i è l'effetto non osservabile proprio di ogni strato.

Per determinare quale sia la funzione $g(\cdot)$ da preferire si può ricorrere alle proprietà dei **modelli lineari generalizzati**, poichè la distribuzione bernoulliana appartiene alla famiglia delle distribuzioni esponenziali. Sia $Y_{ij} \sim \text{Be}(\pi_{ij})$, la sua funzione di densità può essere espressa come

$$\begin{aligned} p_{Y_{ij}}(\pi_{ij}; y_{ij}) &= \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{1-y_{ij}} \\ &= \exp\{y_{ij} \log(\pi_{ij}) + (1 - y_{ij}) \log(1 - \pi_{ij})\} \\ &= \exp\{y_{ij} \log(\pi_{ij}/(1 - \pi_{ij})) + \log(1 - \pi_{ij})\} \\ &= \exp\{y_{ij} \gamma_{ij} - \log(1 + e^{\gamma_{ij}})\}, \end{aligned}$$

dove $\gamma_{ij} = \log(\pi_{ij}/(1 - \pi_{ij})) = \text{logit}(\pi_{ij})$ è il parametro naturale della distribuzione e $\mu(\gamma_{ij}) = \exp\{\gamma_{ij}\}/(1 + \exp\{\gamma_{ij}\}) = \pi_{ij}$ è la funzione media. Sia $\eta_{ij} = \lambda_i + x'_{ij}\psi$ il predittore lineare nei parametri e funzione delle esplicative. Nei modelli lineari generalizzati la funzione legame canonica è quella funzione

$h(\cdot)$ tale per cui (Pace e Salvan, 1997, Paragrafo 6.4):

$$\eta_{ij} = h(\pi_{ij}) = h(\mu(\gamma_{ij})) = \gamma_{ij}.$$

Segue che la funzione $h(\cdot)$ è l'inversa della funzione media. Di conseguenza, nel caso di dati binari, la funzione legame canonica è la funzione logit. Si ottiene quindi la seguente parametrizzazione:

$$\text{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \lambda_i + x'_{ij}\psi = \eta_{ij}. \quad (2.2)$$

Tale scelta per la funzione legame dà origine al **modello logistico**, o modello logit, ed è equivalente a prendere come funzione $g(\cdot)$ in (2.1) la funzione di ripartizione della variabile casuale logistica. Un'alternativa spesso utilizzata, ma qui non presa in considerazione, è quella di utilizzare al posto della funzione di ripartizione logistica quella della normale standard. Ciò dà origine al ben noto **modello probit**.

2.4 Verosimiglianza per dati binari

In questo paragrafo si mostreranno le funzioni di verosimiglianza, introdotte a livello teorico nel Capitolo 1, per la stima in presenza di parametri di disturbo incidentali.

Supponiamo di disporre di un campione stratificato di osservazioni dicotomiche. Sia y_{ij} realizzazione della variabile casuale bernoulliana

$$Y_{ij} \sim \text{Be}(\pi_{ij}) \quad i = 1, \dots, q \quad j = 1, \dots, m.$$

Sia $\theta = (\psi, \lambda)$ il parametro complessivo del modello. Esso è costituito da:

- ψ , parametro di interesse;
- $\lambda = (\lambda_1, \dots, \lambda_q)$, parametro di disturbo incidentale poichè la sua dimensione aumenta all'aumentare del numero degli strati. Come accennato nel Paragrafo 2.1, ciascun elemento di λ coglie le caratteristiche non

osservabili proprie di ogni strato. Si sta quindi facendo riferimento ad un modello ad effetti fissi.

Sfruttiamo ora le considerazioni fatte nel paragrafo precedente. Si indichi con $\eta_{ij} = \lambda_i + x'_{ij}\psi$, $i = 1, \dots, q$, $j = 1, \dots, m$, il predittore lineare nei parametri e funzione delle esplicative. Tale specificazione mette in evidenza un'intercetta variabile di strato in strato, la quale coglie l'eterogeneità non osservabile. La quantità di interesse è invece l'impatto delle variabili esplicative, rappresentato da ψ . Segue da (2.2) che il modello logit per dati binari implica la seguente riparametrizzazione:

$$\pi_{ij} = \frac{\exp\{\lambda_i + x'_{ij}\psi\}}{1 + \exp\{\lambda_i + x'_{ij}\psi\}} = \frac{\exp\{\eta_{ij}\}}{1 + \exp\{\eta_{ij}\}} = F(\eta_{ij}). \quad (2.3)$$

Si noti come la funzione $F(\cdot)$ rappresenti la funzione di ripartizione della distribuzione logistica.

2.4.1 Funzione di verosimiglianza e log-verosimiglianza

Assumendo l'indipendenza tra gli strati e tra le osservazioni all'interno di ogni strato, la **funzione di verosimiglianza** può essere scritta come:

$$\begin{aligned} L(\psi, \lambda) &= \prod_{i=1}^q L^{(i)}(\psi, \lambda_i) \\ &= \prod_{i=1}^q \prod_{j=1}^m p_{Y_{ij}}(y_{ij}; \pi_{ij}) \\ &= \prod_{i=1}^q \prod_{j=1}^m \pi_{ij}^{y_{ij}} (1 - \pi_{ij})^{1-y_{ij}} \\ &= \prod_{i=1}^q \prod_{j=1}^m F(\eta_{ij})^{y_{ij}} (1 - F(\eta_{ij}))^{1-y_{ij}}, \end{aligned} \quad (2.4)$$

dove $L^{(i)}(\psi, \lambda_i)$ rappresenta il contributo dello strato i -esimo alla verosimiglianza complessiva del modello. Indicando con $l^{(i)}(\psi, \lambda_i) = \log L^{(i)}(\psi, \lambda_i)$,

la **log-verosimiglianza** assume la forma:

$$\begin{aligned}
 l(\psi, \lambda) &= \log L(\psi, \lambda) \\
 &= \sum_{i=1}^q l^{(i)}(\psi, \lambda_i) \\
 &= \sum_{i=1}^q \sum_{j=1}^m y_{ij} \log \pi_{ij} + (1 - y_{ij}) \log(1 - \pi_{ij}) \\
 &= \sum_{i=1}^q \sum_{j=1}^m y_{ij} \log F(\eta_{ij}) + (1 - y_{ij}) \log(1 - F(\eta_{ij})).
 \end{aligned} \tag{2.5}$$

2.4.2 Verosimiglianza profilo

Poichè ciascun parametro di disturbo compare esclusivamente nel contributo alla verosimiglianza del proprio strato, la **log-verosimiglianza profilo** può essere scritta come:

$$l_P(\psi) = \sum_{i=1}^q l_P^{(i)}(\psi) = \sum_{i=1}^q l^{(i)}(\psi, \hat{\lambda}_{i\psi}), \tag{2.6}$$

dove, per ψ fissato, $\hat{\lambda}_{i\psi}$ è soluzione dell'equazione in λ_i :

$$\begin{aligned}
 l_{\lambda_i}^{(i)}(\psi, \lambda) &= \frac{\partial l(\psi, \lambda)}{\partial \lambda_i} \\
 &= \frac{\partial l^{(i)}(\psi, \lambda_i)}{\partial \lambda_i} \\
 &= \sum_{j=1}^m y_{ij} - F(\lambda_i + x'_{ij}\psi) = 0.
 \end{aligned} \tag{2.7}$$

2.4.3 Verosimiglianza profilo modificata

Con un ragionamento analogo a quello effettuato per la costruzione della verosimiglianza profilo, partendo da (1.3), si può scrivere la **log-verosimiglianza profilo modificata** come:

$$l_M(\psi) = l_P(\psi) + m(\psi), \tag{2.8}$$

dove $m(\psi) = \log M(\psi)$. Nel seguito, il fattore di modificazione considerato è quello definito in (1.4), poichè non richiede l'individuazione di una statistica ancillare. Inoltre, poichè ciascun parametro di disturbo compare solo nel rispettivo contributo alla verosimiglianza, si può scrivere tale fattore come:

$$\begin{aligned} m(\psi) &= \sum_{i=1}^q m_i(\psi) \\ &= \sum_{i=1}^q \frac{1}{2} \log |j_{\lambda_i \lambda_i}^{(i)}(\psi, \hat{\lambda}_{i\psi})| - \log |I_{\lambda_i \lambda_i}(\hat{\psi}, \hat{\lambda}_i; \psi, \hat{\lambda}_{i\psi})|, \end{aligned} \quad (2.9)$$

dove $(\hat{\psi}, \hat{\lambda}_i)$ rappresentano le stime di massima verosimiglianza dei parametri. Indicando con $l_{\lambda_i}^{(i)}(\psi, \lambda_i) = \partial l^{(i)}(\psi, \lambda_i) / \partial \lambda_i$ e con $f(\cdot) = F'(\cdot)$, si ha che

$$\begin{aligned} j_{\lambda_i \lambda_i}^{(i)}(\psi, \lambda_i) &= -l_{\lambda_i \lambda_i}^{(i)}(\psi, \lambda_i) \\ &= -\frac{\partial l_{\lambda_i}^{(i)}(\psi, \lambda_i)}{\partial \lambda_i} \\ &= \sum_{j=1}^m F(\lambda_i + x'_{ij}\psi)(1 - F(\lambda_i + x'_{ij}\psi)), \end{aligned} \quad (2.10)$$

e che

$$\begin{aligned} &I_{\lambda_i \lambda_i}(\psi_0, \lambda_{i0}; \psi_1, \lambda_{i1}) \\ &= E_{\psi_0, \lambda_{i0}} \{l_{\lambda_i}^{(i)}(\psi_0, \lambda_{i0}) l_{\lambda_i}^{(i)}(\psi_1, \lambda_{i1})\} \\ &= E_{\psi_0, \lambda_{i0}} \left\{ \left[\sum_{j=1}^m Y_{ij} - F(\lambda_{i0} + x'_{ij}\psi_0) \right] \left[\sum_{u=1}^m Y_{iu} - F(\lambda_{i1} + x'_{iu}\psi_1) \right] \right\} \quad (2.11) \\ &= \sum_{j=1}^m \frac{f(\lambda_{i1} + x'_{ij}\psi_1) f(\lambda_{i0} + x'_{ij}\psi_0)}{(1 - F(\lambda_{i1} + x'_{ij}\psi_1)) F(\lambda_{i1} + x'_{ij}\psi_1)}, \end{aligned}$$

in quanto, per $j \neq u$, l'indipendenza tra le osservazioni all'interno di ogni strato implica che

$$\begin{aligned} &E_{\psi_0, \lambda_{i0}} \{Y_{ij} Y_{iu} - Y_{ij} F(\lambda_{i1} + x'_{iu}\psi_1) + \\ &\quad - Y_{iu} F(\lambda_{i0} + x'_{ij}\psi_0) + F(\lambda_{i0} + x'_{ij}\psi_0) F(\lambda_{i1} + x'_{iu}\psi_1)\} = 0. \end{aligned}$$

L'espressione (2.11) deriva da Bellio e Sartori (2003) ed è valida in generale quanto si trattano dati binari. Proseguendo, poichè nel modello logit si ha che $f(\cdot) = F(\cdot)(1 - F(\cdot))$, si ottiene che

$$I_{\lambda_i \lambda_i}(\hat{\psi}, \hat{\lambda}_i; \psi, \hat{\lambda}_{i\psi}) = \sum_{j=1}^m F(\hat{\lambda}_i + x'_{ij}\hat{\psi})(1 - F(\hat{\lambda}_i + x'_{ij}\hat{\psi})).$$

Quindi, questo termine non dipende da ψ e può essere omissso nella costruzione della verosimiglianza profilo modificata. Ciò non è vero in generale se la funzione legame non è quella canonica. Ad esempio, nel modello probit tale termine non può essere trascurato.

2.4.4 Verosimiglianze integrate

Per definire la **verosimiglianza integrata** è innanzitutto necessario scegliere la distribuzione a priori per il parametro di disturbo. Sia $\gamma = (\gamma_1, \dots, \gamma_q)$ il parametro di disturbo o una sua eventuale riparametrizzazione. Data l'ipotesi di indipendenza tra gli strati, è lecito assumere che anche le componenti di γ siano considerate indipendenti nella specificazione della distribuzione del parametro di disturbo. Questo porta formulare una a priori del tipo:

$$\pi(\gamma|\psi) = \prod_{i=1}^q \pi(\gamma_i|\psi). \quad (2.12)$$

Di conseguenza, la verosimiglianza integrata assume la forma:

$$L_I(\psi) = \prod_{i=1}^q \int_{\Gamma} L^{(i)}(\psi, \gamma_i) \pi(\gamma_i|\psi) d\gamma_i.$$

Nel Capitolo 1 sono stati illustrati tre metodi per la costruzione della verosimiglianza integrata, i quali sono associati a tre diverse scelte della distribuzione a priori per il parametro di disturbo. Il primo metodo consiste nell'ottenere, per ogni strato, un parametro di disturbo fortemente disconnesso dal parametro di interesse tramite la parametrizzazione ZSE, specificando poi per esso una a priori uniforme. Gli altri due metodi, invece, non neces-

sitano di alcuna riparametrizzazione del modello e specificano le priori come illustrato in (1.9) e (1.10).

Per la costruzione della verosimiglianza integrata attraverso la **parametrizzazione ZSE**, è necessario risolvere l'equazione

$$E_{\psi_0, \lambda_{i0}} \{l_{\lambda_i}^{(i)}(\psi, \lambda_i)\} |_{(\psi_0, \lambda_{i0}) = (\hat{\psi}, \phi_i)} = \sum_{j=1}^m F(\phi_i + x'_{ij} \hat{\psi}) - F(\lambda_i + x'_{ij} \psi) = 0,$$

rispetto a ϕ_i , per $(\psi, \lambda_i, \hat{\psi})$ fissata. Si ottiene così un parametro di disturbo $\phi_i = \phi(\psi, \lambda_i, \hat{\psi})$ fortemente disconnesso da ψ (Severini, 2007). Successivamente, sfruttando la proprietà di invarianza della verosimiglianza rispetto a riparametrizzazioni del modello, si giunge alla seguente verosimiglianza integrata:

$$L_I(\psi) = \prod_{i=1}^q \int_{\Phi} L^{(i)}(\psi, \lambda(\phi_i)) d\phi_i, \quad (2.13)$$

dove $\lambda(\phi_i) \equiv \lambda(\psi, \phi_i, \hat{\psi}) = \phi^{-1}(\psi, \lambda_i, \hat{\psi})$. Quest'ultima funzione rappresenta la parametrizzazione inversa alla ZSE, che permette quindi di ottenere λ_i a partire da ϕ_i . Ciascun $\phi_i, i = 1, \dots, q$, rappresenta il nuovo parametro di disturbo proprio di ogni strato. Sfruttando l'assunzione di indipendenza tra gli strati è possibile quindi specificare una distribuzione a priori come quella mostrata in (2.12) per il parametro di disturbo complessivo $\phi = (\phi_1, \dots, \phi_q)$. Si noti, in questo caso, come la specificazione di una a priori uniforme per il parametro $\phi_i, i = 1, \dots, q$, introduca nella verosimiglianza una quantità positiva la quale è stata omessa in (2.13). La log-verosimiglianza integrata può essere quindi scritta come:

$$l_I(\psi) = \log L_I(\psi) = \sum_{i=1}^m \log \int_{\Phi} L^{(i)}(\psi, \phi_i) d\phi_i.$$

Qualora non si vogliano effettuare riparametrizzazioni del modello, si possono utilizzare le a priori proposte da Arellano e Bonhomme (2009). La prima distribuzione a priori per il parametro di disturbo considerata è la cosiddetta

robust prior. Essa ha la seguente espressione:

$$\pi_i^R(\lambda_i|\psi) \propto \hat{E}\{-l_{\lambda_i\lambda_i}^{(i)}(\psi, \lambda_i)\}[\hat{E}\{[l_{\lambda_i}^{(i)}(\psi, \lambda_i)]^2\}]^{-1/2}. \quad (2.14)$$

I due termini $\hat{E}\{-l_{\lambda_i\lambda_i}^{(i)}(\psi, \lambda_i)\}$ e $\hat{E}\{[l_{\lambda_i}^{(i)}(\psi, \lambda_i)]^2\}$ stimano rispettivamente le quantità attese $E_{\psi_0\lambda_{i0}}\{-l_{\lambda_i\lambda_i}^{(i)}(\psi, \lambda_i)\}$ e $E_{\psi_0\lambda_{i0}}\{[l_{\lambda_i}^{(i)}(\psi, \lambda_i)]^2\}$, in cui i veri valori dei parametri sono stati sostituiti con le stime di massima verosimiglianza. Sfruttando (2.10), si ottiene che il primo termine vale:

$$\begin{aligned} \hat{E}\{-l_{\lambda_i\lambda_i}^{(i)}(\psi, \lambda_i)\} &= E_{\psi_0\lambda_{i0}}\{-l_{\lambda_i\lambda_i}^{(i)}(\psi, \lambda_i)\}_{(\psi_0, \lambda_{i0})=(\hat{\psi}, \hat{\lambda}_i)} \\ &= E_{\psi_0\lambda_{i0}}\{j_{\lambda_i\lambda_i}^{(i)}(\psi, \lambda_i)\}_{(\psi_0, \lambda_{i0})=(\hat{\psi}, \hat{\lambda}_i)} \\ &= \sum_{j=1}^m F(\lambda_i + x'_{ij}\psi)(1 - F(\lambda_i + x'_{ij}\psi)). \end{aligned} \quad (2.15)$$

La quantità $\hat{E}\{[l_{\lambda_i}^{(i)}(\psi, \lambda_i)]^2\} = E_{\psi_0\lambda_{i0}}\{[l_{\lambda_i}^{(i)}(\psi, \lambda_i)]^2\}_{(\psi_0, \lambda_{i0})=(\hat{\psi}, \hat{\lambda}_i)}$, si può ottenere a partire dal fatto che

$$\begin{aligned} E_{\psi_0\lambda_{i0}}\{[l_{\lambda_i}^{(i)}(\psi, \lambda_i)]^2\} &= E_{\psi_0\lambda_{i0}}\left\{\left[\sum_{j=1}^m Y_{ij} - F(\lambda_i + x'_{ij}\psi)\right]\left[\sum_{u=1}^m Y_{iu} - F(\lambda_i + x'_{iu}\psi)\right]\right\} \\ &= \sum_{j=1}^m \sum_{u=1}^m E_{\psi_0\lambda_{i0}}\{[Y_{ij} - F(\lambda_i + x'_{ij}\psi)][Y_{iu} - F(\lambda_i + x'_{iu}\psi)]\}. \end{aligned} \quad (2.16)$$

Quindi, sostituendo ai veri parametri le stime di massima verosimiglianza, si ottiene che, per $j = u$:

$$\begin{aligned} E_{\psi_0\lambda_{i0}}\{[Y_{ij} - F(\lambda_i + x'_{ij}\psi)][Y_{iu} - F(\lambda_i + x'_{iu}\psi)]\}_{(\psi_0, \lambda_{i0})=(\hat{\psi}, \hat{\lambda}_i)} \\ = F(\hat{\lambda}_i + x'_{ij}\hat{\psi}) - 2F(\hat{\lambda}_i + x'_{ij}\hat{\psi})F(\lambda_i + x'_{ij}\psi) + F(\lambda_i + x'_{ij}\psi)^2, \end{aligned}$$

mentre, per $j \neq u$:

$$\begin{aligned} E_{\psi_0 \lambda_{i0}} \{ [Y_{ij} - F(\lambda_i + x'_{ij}\psi)][Y_{iu} - F(\lambda_i + x'_{iu}\psi)] \} |_{(\psi_0, \lambda_{i0}) = (\hat{\psi}, \hat{\lambda}_i)} \\ = F(\hat{\lambda}_i + x'_{ij}\hat{\psi})F(\hat{\lambda}_i + x'_{iu}\hat{\psi}) + F(\hat{\lambda}_i + x'_{ij}\hat{\psi})F(\lambda_i + x'_{iu}\psi) \\ + F(\lambda_i + x'_{ij}\psi)F(\hat{\lambda}_i + x'_{iu}\hat{\psi}) + F(\lambda_i + x'_{ij}\psi)F(\lambda_i + x'_{iu}\psi). \end{aligned}$$

La seconda distribuzione per il parametro di disturbo presa in considerazione è basata sulla densità della normale standard. Essa, come la *robust prior*, permette di ridurre la distorsione della funzione punteggio. Tale distribuzione, detta **normal prior**, assume la seguente forma:

$$\pi_i^N(\lambda_i | \psi) = \mathcal{N}(\hat{\lambda}_{i\psi}, \widehat{Var}(\hat{\lambda}_{i\psi})), \quad (2.17)$$

in cui $\hat{\lambda}_{i\psi}$ è soluzione dell'equazione (2.7). La varianza può essere scritta come

$$\widehat{Var}(\hat{\lambda}_{i\psi}) = I_{\lambda_i \lambda_i}^2(\hat{\psi}, \hat{\lambda}_i; \psi, \hat{\lambda}_{i\psi}) j_{\lambda_i \lambda_i}^{-2}(\psi, \hat{\lambda}_{i\psi}) j_{\lambda_i \lambda_i}^{-1}(\hat{\psi}, \hat{\lambda}_i),$$

dove, a partire da (2.10) e (2.11) si ha che:

$$\begin{aligned} I_{\lambda_i \lambda_i}(\hat{\psi}, \hat{\lambda}_i; \psi, \hat{\lambda}_{i\psi}) &= \sum_{j=1}^m F(\hat{\lambda}_i + x'_{ij}\hat{\psi})(1 - F(\hat{\lambda}_i + x'_{ij}\hat{\psi})) \\ j_{\lambda_i \lambda_i}(\psi, \hat{\lambda}_{i\psi}) &= \sum_{j=1}^m F(\hat{\lambda}_{i\psi} + x'_{ij}\psi)(1 - F(\hat{\lambda}_{i\psi} + x'_{ij}\psi)) \\ j_{\lambda_i \lambda_i}(\hat{\psi}, \hat{\lambda}_i) &= \sum_{j=1}^m F(\hat{\lambda}_i + x'_{ij}\hat{\psi})(1 - F(\hat{\lambda}_i + x'_{ij}\hat{\psi})). \end{aligned}$$

Capitolo 3

Studi di simulazione

3.1 Introduzione

In questo capitolo si confronteranno i metodi descritti precedentemente in un modello logistico stratificato. In particolare, tramite studi di simulazione, si valuteranno le proprietà dei seguenti metodi basati sulla funzione di verosimiglianza per la stima dei parametri di interesse:

1. verosimiglianza profilo;
2. verosimiglianza profilo modificata;
3. verosimiglianza integrata con parametrizzazione ZSE e distribuzione a priori uniforme;
4. verosimiglianza integrata con distribuzione a priori “robusta”, definita in (1.9);
5. verosimiglianza integrata con distribuzione a priori basata sulla densità normale, definita in (1.10).

Si vuol far notare come, nel modello logit, esiste una verosimiglianza condizionata esatta per ψ . Tuttavia, essa non è stata considerata nelle simulazioni poichè in letteratura è nota la relazione tra la verosimiglianza profilo modificata e tale verosimiglianza condizionata.

Quello che ci si attende a partire dalle considerazioni teoriche esposte nel Capitolo 1 è, innanzitutto, il fallimento del metodo basato sulla semplice verosimiglianza profilo. Infatti, in presenza di parametri di disturbo incidentali, vengono meno le condizioni di regolarità che ne garantiscono le buone proprietà asintotiche. Nei Paragrafi 1.10.2 e 1.10.3 si è messo in luce come la log-verosimiglianza integrata possa essere approssimata attraverso la log-verosimiglianza profilo modificata, fintanto che $q/m^3 = o(1)$, dove q è il numero degli strati e m è la dimensione degli strati. Di conseguenza, è lecito attendersi che i due metodi tendano a dare risultati molto simili. In particolare ci si attende che, all'aumentare della dimensione degli strati, le stime ottenute con la verosimiglianza integrata con parametrizzazione ZSE si avvicinino sempre più a quelle ottenute attraverso la profilo modificata. Per quanto riguarda le verosimiglianze integrate costruite tramite la *robust prior* e la *normal prior*, ci si attende che diano dei risultati almeno più accurati della semplice profilo. Questo perchè si conosce solamente il fatto che tali a priori riducono la distorsione della funzione punteggio.

Il modello considerato per i confronti è il modello logit per campioni stratificati illustrato nel Paragrafo 2.3. Le funzioni di verosimiglianza elencate in precedenza sono esplicitate invece nel Paragrafo 2.4. La notazione usata nel seguito è la stessa dei capitoli precedenti: si suppone cioè di disporre di un campione stratificato di osservazioni dicotomiche y_{ij} , realizzazioni delle variabili casuali

$$Y_{ij} \sim \text{Be}(\pi_{ij}) \quad i = 1, \dots, q \quad j = 1, \dots, m,$$

dove q è il numero degli strati ed m è il numero di osservazioni per strato. Si suppone che gli strati e che le osservazioni all'interno di ciascuno strato siano indipendenti. La probabilità di successo, $\pi_{ij} = \Pr(Y_{ij} = 1 | X_{ij} = x_{ij})$, dipende dalla esplicative secondo l'espressione:

$$\pi_{ij} = \frac{\exp\{\lambda_i + x'_{ij}\psi\}}{1 + \exp\{\lambda_i + x'_{ij}\psi\}}, \quad (3.1)$$

dove ψ costituisce il parametro di interesse e λ_i è il parametro di disturbo di natura incidentale, introdotto sotto forma di intercetta variabile. Segue che i metodi di stima vengono confrontati sulla base delle stime del parametro di interesse, per diverse configurazioni di q ed m .

3.2 Aspetti computazionali

L'implementazione delle funzioni necessarie per stimare attraverso la funzione di verosimiglianza integrata non è stata priva di problemi. Dato il largo impiego del costrutto *for*, l'idea di scrivere il codice solamente usando il linguaggio R è stata abbandonata fin da subito. Questo perchè tale linguaggio è notoriamente poco prestante nei cicli. In prima battuta, si sono posti come soluzione i linguaggi C e C++, quest'ultimo utilizzato tramite la libreria *Rcpp*. Si sono quindi implementate le funzioni di verosimiglianza integrata, costruite a partire dalle tre priori illustrate in precedenza. Si sono poi utilizzate le routines R per risolvere i problemi di massimizzazione. Così facendo, il carico computazionale resta comunque elevato. Nella *normal prior*, ad esempio, ad ogni iterazione per la ricerca del massimo è necessario determinare la stima di $\lambda_i, i = 1, \dots, q$ per ψ fissato. Questo richiede di individuare numericamente le radici di un'equazione usando un algoritmo tipo Newton-Raphson, rendendo quindi l'integrazione numerica per l'eliminazione del parametro di disturbo molto lenta. La stessa *robust prior*, sebbene coinvolge per lo più sommatorie, si è rivelata pesante computazionalmente. Infatti, i tempi necessari alla stima utilizzando queste due distribuzioni a priori erano dell'ordine degli 80¹ minuti e ciò non permette di sfruttare le funzioni così costruite per la simulazione. Il tempo necessario per la stima attraverso la funzione di verosimiglianza integrata costruita tramite la parametrizzazione ZSE, invece, è dell'ordine dei 5 minuti¹. Quindi, tale verosimiglianza è l'unica implementata attraverso il linguaggio C/C++. Inoltre, per rendere possibile l'integrazione numerica in C, è stata costruita un'apposita classe per il passaggio dei dati e dei parametri tra

¹Tempi riferiti a un computer Dell Inspiron, processore Intel Pentium Dual-Core CPU E5800, 3.20GHz, 4GB di RAM e ad un campione stratificato di dimensioni $q = 100$ e $m = 4$.

le varie funzioni. Infine, con lo scopo di rendere la simulazione più veloce si è sfruttato anche il calcolo parallelo attraverso le librerie `doMC` e `plyr`. Per fare ciò si è utilizzato il server di calcolo messo a disposizione dal Dipartimento di Scienze Statistiche.

Per rendere fruibile in ottica simulazione la verosimiglianza integrata costruita con la *robust prior* e la *normal prior*, si è ricorso alla libreria `TMB`. Tale libreria nasce per la stima di modelli ad effetti casuali e prevede un'integrazione numerica basata sull'approssimazione di Laplace. Per un approfondimento su tale tipo di integrazione si rimanda a Liu e Pierce (1994). La velocità garantita da questo metodo deriva dal fatto che le derivate vengono calcolate in modo analitico in fase di compilazione. Per adattare tale metodo al modello ad effetti fissi considerato è stato sufficiente costruire le due a priori ed indicare qual è il parametro di disturbo rispetto a cui effettuare l'integrazione. I tempi di calcolo sono stati notevolmente ridotti e sono attorno al secondo¹, sia per la *robust prior* che per la *normal prior*.

Un'ulteriore nota va fatta per la *normal prior*. Come detto in precedenza, essa prevede l'individuazione di $\lambda_i, i = 1, \dots, q$, per ψ fissato. Per diminuire il carico computazionale, la ricerca di $\hat{\lambda}_{i\psi}$, soluzione dell'equazione (2.7) per ψ fissato, è avvenuta attraverso un'approssimazione lineare. In particolare, approssimando la funzione punteggio per λ_i attorno alla stima di massima verosimiglianza $(\hat{\psi}, \hat{\lambda}_i)$ in ogni strato si ottiene che

$$l_{\lambda_i}^{(i)}(\psi, \hat{\lambda}_{i\psi}) \simeq l_{\lambda_i}^{(i)}(\hat{\psi}, \hat{\lambda}_i) + l_{\lambda_i\psi}^{(i)}(\hat{\psi}, \hat{\lambda}_i)(\psi - \hat{\psi}) + l_{\lambda_i\lambda_i}^{(i)}(\hat{\psi}, \hat{\lambda}_i)(\hat{\lambda}_{i\psi} - \hat{\lambda}_i),$$

dove $l_{\lambda_i\psi}^{(i)}(\psi, \lambda_i) = \partial l_{\lambda_i}^{(i)}(\psi, \lambda_i) / \partial \psi$. Poichè $l_{\lambda_i}^{(i)}(\hat{\psi}, \hat{\lambda}_i) = 0$, si ha che la relazione tra $\hat{\lambda}_{i\psi}$ e ψ può essere approssimata linearmente come

$$\hat{\lambda}_{i\psi} = \hat{\lambda}_i - \frac{l_{\lambda_i\psi}^{(i)}(\hat{\psi}, \hat{\lambda}_i)}{l_{\lambda_i\lambda_i}^{(i)}(\hat{\psi}, \hat{\lambda}_i)}(\psi - \hat{\psi}).$$

Così facendo, si riesce ad approssimare $\hat{\lambda}_{i\psi}$ utilizzando solo quantità calcolate nella stima non vincolata $(\hat{\psi}, \hat{\lambda}_i)$.

Infine, per ottenere le stime basate sulla verosimiglianza profilo e sulla profilo modificata, è stato sufficiente utilizzare in modo opportuno una serie di funzioni già implementate disponibili nella libreria `panelMPL2`.

3.3 Risultati

Nel seguito si confronteranno i metodi di stima elencati nel Paragrafo 3.1 tramite un approccio di tipo Monte Carlo. Il vero modello generatore dei dati utilizzato nella simulazione specifica (3.1) nel modo seguente:

$$\pi_{ij} = \frac{\exp\{\lambda_i - x_{1,ij} + 2x_{2,ij}\}}{1 + \exp\{\lambda_i - x_{1,ij} + 2x_{2,ij}\}}$$

dove le esplicative $x_{1,ij}$ e $x_{2,ij}$ sono realizzazioni di variabili casuali indipendenti normali standard per $i = 1, \dots, q$ e $j = 1, \dots, m$. L'effetto proprio di ogni strato, $\lambda_i, i = 1, \dots, q$, è stato ottenuto come media della prima variabile esplicativa a cui è stato aggiunto del rumore bianco. Il vero valore del parametro di interesse è quindi $\psi = (\psi_1, \psi_2) = (-1, 2)$.

Si sono quindi simulati $R = 1000$ campioni stratificati di dati binari, per varie configurazioni di q ed m , ottenendo, per ciascun metodo di stima considerato, i vettori $\psi_1^{sim} = (\psi_1^{(1)}, \dots, \psi_1^{(R)})$ e $\psi_2^{sim} = (\psi_2^{(1)}, \dots, \psi_2^{(R)})$. Si sono anche calcolati gli standard error delle stime, sempre per ciascun metodo di stima, a partire dal calcolo numerico dell'hessiano nel punto di massimo, ottenendo quindi i vettori $\sigma_1^{sim} = (\sigma_1^{(1)}, \dots, \sigma_1^{(R)})$ e $\sigma_2^{sim} = (\sigma_2^{(1)}, \dots, \sigma_2^{(R)})$. Per entrambi i vettori delle stime ψ_1^{sim} e ψ_2^{sim} , l'adeguatezza dei metodi è stata valutata calcolando la media (MEAN), la mediana (MED), la distorsione (BIAS), l'errore quadratico medio (MSE), la radice dell'errore quadratico medio (RMSE) e l'errore assoluto medio (MAE). Si è poi valutato il rapporto tra la media degli standard error ottenuti numericamente con lo standard error ottenuto a partire dal vettore delle stime (SD), ottenendo l'indicatore SE/SD. Infine si è valutata la copertura empirica degli intervalli di confidenza alla Wald (IC) ai livelli nominali 0.9, 0.95, 0.99.

Per quanto riguarda le dimensioni del campione stratificato, si sono considerati due scenari, in cui $q = 100$ e $q = 200$. Successivamente, per entrambi,

la dimensione degli strati, m , è stata fissata pari a 4, 8 e 12. In questo modo si possono cogliere due aspetti principalmente. In primo luogo, all'aumentare del numero degli strati ed a parità di dimensione degli strati, si vuole mostrare come i metodi di stima esposti continuino a funzionare nonostante la natura incidentale del parametro di disturbo. Alla luce delle considerazioni teoriche del Paragrafo 1.10, ci si aspetta che i metodi basati sulla profilo modificata e sulla verosimiglianza integrata siano migliori rispetto alla semplice profilo. In secondo luogo come, a parità del numero degli strati, la maggior informazione che si ottiene all'aumentare delle osservazioni per strato porti ad ottenere stimatori più precisi. In particolare, sempre alla luce delle considerazioni teoriche del Paragrafo 1.10, ci si attende una riduzione generale della distorsione degli stimatori e che le stime ottenute con la verosimiglianza integrata costruita attraverso la parametrizzazione ZSE si avvicinino a quelle ottenute a partire dalla profilo modificata.

Nelle tabelle seguenti vengono presentati i risultati delle simulazioni Monte Carlo. Gli stimatori per il parametro $\psi_i, i = 1, 2$, vengono indicati come segue: $\hat{\psi}_i^P$ è lo stimatore basato sulla semplice verosimiglianza profilo mentre $\hat{\psi}_i^{MP}$ è quello basato sulla verosimiglianza profilo modificata; gli stimatori $\hat{\psi}_i^{ZSE}$, $\hat{\psi}_i^{RP}$ e $\hat{\psi}_i^{NP}$ sono basati sulla verosimiglianza integrata costruita rispettivamente con la parametrizzazione ZSE, la *robust prior* e la *normal prior*. Quello che emerge in generale è che, per q fissato, all'aumentare di m la distorsione degli stimatori proposti diminuisce, la copertura degli intervalli si avvicina al livello nominale e la media degli standard error stimati si avvicina alla deviazione standard delle stime. Il numero degli strati, q , sembra influire soprattutto sugli stimatori $\hat{\psi}_i^P$, $\hat{\psi}_i^{RP}$ e $\hat{\psi}_i^{NP}$, in particolare quando $m = 4$. Tale effetto sembra invece divenire trascurabile all'aumentare della dimensione degli strati. Queste prime considerazioni non implicano che tutti i metodi funzionino bene. La verosimiglianza profilo, come anticipato, fornisce stime altamente distorte per tutte le configurazioni di m e q considerate. Se a questo si aggiunge il fatto che tale metodo sottostima la variabilità dello stimatore (SE/SD inferiori a uno) è logico attendersi che le coperture degli intervalli di confidenza siano ben lontane dai rispettivi valori nominali. Segue che tale stimatore (standardizzato) difficilmente possa essere distribui-

to asintoticamente come una normale standard. Anche gli stimatori $\hat{\psi}_i^{RP}$ e $\hat{\psi}_i^{NP}$ presentano dei rapporti SE/SD molto bassi, spesso dello stesso ordine di grandezza di quelli che si ottengono a partire dalla verosimiglianza profilo. Inoltre, per $m = 4$, $\hat{\psi}_i^{RP}$ sembra risentire maggiormente la natura incidentale del parametro di disturbo rispetto agli altri stimatori. Infatti, all'aumentare di q , nelle Tabelle 3.1 e 3.2 si può notare come il rapporto SE/SD per $\hat{\psi}_i^{RP}$ diminuisca molto più di quello degli altri stimatori. Quindi, i risultati delle simulazioni confermano le considerazioni teoriche del Capitolo 1: i metodi basati sulla verosimiglianza integrata costruita attraverso la *robust prior* e la *normal prior*, sebbene insoddisfacenti, portano comunque a stimatori con proprietà migliori rispetto a quello che si ottiene a partire dalla profilo, per via della riduzione della distorsione della funzione punteggio. Tali miglioramenti sono però trascurabili se $\hat{\psi}_i^{RP}$ e $\hat{\psi}_i^{NP}$ sono confrontati con $\hat{\psi}_i^{MP}$ e $\hat{\psi}_i^{ZSE}$. Ad esempio, si può notare come la loro distorsione sia circa 7 volte più grande rispetto a quella di $\hat{\psi}_i^{MP}$ e $\hat{\psi}_i^{ZSE}$ nelle varie configurazioni di q e m considerate. Alla luce dei risultati delle simulazioni emerge che gli stimatori $\hat{\psi}_i^{MP}$ e $\hat{\psi}_i^{ZSE}$ sono quelli da preferire. Le proprietà dello stimatore basato sulla profilo modificata in un contesto asintotico a due indici sono esposte in Sartori (2003). Quello che si nota nelle simulazioni è che $\hat{\psi}_i^{MP}$ ha la minor distorsione, il rapporto SE/SD è sempre prossimo a uno e la copertuta degli intervalli di confidenza, vicina al valore nominale, rende l'assunzione di normalità asintotica verosimile. Ciò può essere supportato anche dal fatto che la mediana è molto vicino alla media, evidenziando una certa simmetria della distribuzione limite. Tale metodo è anche quello che risente meno dell'aumento della dimensione del parametro di disturbo per una fissata dimensione degli strati. Coerentemente con le considerazioni fatte nel Paragrafo 1.10, lo stimatore basato sulla verosimiglianza integrata costruita con la parametrizzazione ZSE tende ad avvicinarsi allo stimatore $\hat{\psi}_i^{MP}$ all'aumentare di m . Questo segue dal fatto che la verosimiglianza integrata può essere approssimata con la profilo modificata quando $q/m^3 = o(1)$. Di conseguenza, i due stimatori godono delle stesse proprietà asintotiche. Ad esempio, all'aumentare di m si osserva empiricamente che, per q fissato, la distorsione di $\hat{\psi}_i^{ZSE}$ diminuisce e tende a diventare dello stesso ordine di grandezza di quella di

$\hat{\psi}_i^{MP}$. Allo stesso modo si nota come i livelli di copertura degli intervalli di confidenza basati sulla verosimiglianza integrata con parametrizzazione ZSE siano sempre più simili a quelli degli intervalli costruiti a partire dalla profilo modificata. Considerazioni analoghe possono essere fatte sulle altre quantità riportate nelle tabelle per questi due stimatori.

	$\hat{\psi}_1^P$	$\hat{\psi}_1^{MP}$	$\hat{\psi}_1^{ZSE}$	$\hat{\psi}_1^{RP}$	$\hat{\psi}_1^{NP}$
MEAN	-1.6303	-0.9280	-1.1456	-1.2467	-1.2645
MED	-1.5897	-0.9142	-1.1133	-1.2288	-1.2370
BIAS	-0.6303	0.0720	-0.1456	-0.2467	-0.2645
MSE	0.5401	0.0349	0.0972	0.1356	0.1435
RMSE	0.7349	0.1869	0.3118	0.3682	0.3788
MAE	0.6351	0.1534	0.2261	0.2901	0.2984
SE/SD	0.7474	1.0898	0.8384	0.2948	0.8187
IC(0.90)	0.3230	0.9010	0.8430	0.3726	0.6890
IC(0.95)	0.4270	0.9520	0.9220	0.4079	0.7710
IC(0.99)	0.6720	0.9850	0.9840	0.4733	0.9000

Tabella 3.1: Confronto tra i cinque metodi di stima. R=1000 replicazioni Monte Carlo, parametro $\psi_1 = -1$. Dimensione del campione: $q = 100$, $m = 4$.

	$\hat{\psi}_1^P$	$\hat{\psi}_1^{MP}$	$\hat{\psi}_1^{ZSE}$	$\hat{\psi}_1^{RP}$	$\hat{\psi}_1^{NP}$
MEAN	-1.6027	-0.9068	-1.1452	-1.1706	-1.2372
MED	-1.5778	-0.9050	-1.1228	-1.1609	-1.2204
BIAS	-0.6027	0.0932	-0.1452	-0.1706	-0.2372
MSE	0.4395	0.0229	0.0592	0.0501	0.0945
RMSE	0.6630	0.1514	0.2434	0.2237	0.3074
MAE	0.6030	0.1240	0.1848	0.1846	0.2526
SE/SD	0.7242	1.0915	0.8322	0.1473	0.7910
IC(0.90)	0.1180	0.8270	0.7730	0.1081	0.5570
IC(0.95)	0.1800	0.8910	0.8640	0.1223	0.6640
IC(0.99)	0.3570	0.9670	0.9630	0.1509	0.8310

Tabella 3.2: Confronto tra i cinque metodi di stima. R=1000 replicazioni Monte Carlo, parametro $\psi_1 = -1$. Dimensione del campione: $q = 200$, $m = 4$.

	$\hat{\psi}_1^P$	$\hat{\psi}_1^{MP}$	$\hat{\psi}_1^{ZSE}$	$\hat{\psi}_1^{RP}$	$\hat{\psi}_1^{NP}$
MEAN	-1.2506	-1.0100	-1.0209	-1.1544	-1.1272
MED	-1.2336	-1.0012	-1.0107	-1.1366	-1.1148
BIAS	-0.2506	-0.0100	-0.0209	-0.1544	-0.1272
MSE	0.0925	0.0166	0.0182	0.0510	0.0386
RMSE	0.3042	0.1288	0.1347	0.2258	0.1965
MAE	0.2562	0.1028	0.1056	0.1786	0.1544
SE/SD	0.8553	0.9955	0.9717	0.8392	0.9137
IC(0.90)	0.5140	0.9200	0.9120	0.7187	0.7650
IC(0.95)	0.6270	0.9580	0.9540	0.7958	0.8440
IC(0.99)	0.8140	0.9930	0.9890	0.8719	0.9570

Tabella 3.3: Confronto tra i cinque metodi di stima. R=1000 replicazioni Monte Carlo, parametro $\psi_1 = -1$. Dimensione del campione: $q = 100$, $m = 8$.

	$\hat{\psi}_1^P$	$\hat{\psi}_1^{MP}$	$\hat{\psi}_1^{ZSE}$	$\hat{\psi}_1^{RP}$	$\hat{\psi}_1^{NP}$
MEAN	-1.2424	-1.0071	-1.0178	-1.1565	-1.1214
MED	-1.2380	-1.0034	-1.0136	-1.1518	-1.1173
BIAS	-0.2424	-0.0071	-0.0178	-0.1565	-0.1214
MSE	0.0728	0.0081	0.0089	0.0379	0.0255
RMSE	0.2698	0.0898	0.0942	0.1947	0.1596
MAE	0.2433	0.0723	0.0749	0.1643	0.1328
SE/SD	0.8549	0.9856	0.9633	0.8442	0.9103
IC(0.90)	0.2530	0.8890	0.8810	0.5340	0.6410
IC(0.95)	0.3440	0.9510	0.9370	0.6410	0.7340
IC(0.99)	0.5820	0.9940	0.9930	0.8030	0.8940

Tabella 3.4: Confronto tra i cinque metodi di stima. R=1000 replicazioni Monte Carlo, parametro $\psi_1 = -1$. Dimensione del campione: $q = 200$, $m = 8$.

	$\hat{\psi}_1^P$	$\hat{\psi}_1^{MP}$	$\hat{\psi}_1^{ZSE}$	$\hat{\psi}_1^{RP}$	$\hat{\psi}_1^{NP}$
MEAN	-1.1458	-1.0068	-1.0099	-1.0815	-1.0749
MED	-1.1340	-0.9974	-1.0006	-1.0706	-1.0642
BIAS	-0.1458	-0.0068	-0.0099	-0.0815	-0.0749
MSE	0.0337	0.0090	0.0091	0.0175	0.0162
RMSE	0.1836	0.0946	0.0956	0.1323	0.1272
MAE	0.1532	0.0732	0.0736	0.1028	0.0984
SE/SD	0.9133	0.9938	0.9900	0.9397	0.9501
IC(0.90)	0.6260	0.9020	0.9010	0.7970	0.8060
IC(0.95)	0.7370	0.9420	0.9390	0.8690	0.8820
IC(0.99)	0.8750	0.9880	0.9870	0.9500	0.9570

Tabella 3.5: Confronto tra i cinque metodi di stima. R=1000 replicazioni Monte Carlo, parametro $\psi_1 = -1$. Dimensione del campione: $q = 100$, $m = 12$.

	$\hat{\psi}_1^P$	$\hat{\psi}_1^{MP}$	$\hat{\psi}_1^{ZSE}$	$\hat{\psi}_1^{RP}$	$\hat{\psi}_1^{NP}$
MEAN	-1.1451	-1.0062	-1.0072	-1.0795	-1.0744
MED	-1.1430	-1.0048	-1.0061	-1.0781	-1.0732
BIAS	-0.1451	-0.0062	-0.0072	-0.0795	-0.0744
MSE	0.0275	0.0047	0.0047	0.0119	0.0111
RMSE	0.1660	0.0689	0.0689	0.1093	0.1053
MAE	0.1471	0.0547	0.0547	0.0900	0.0862
SE/SD	0.9303	1.0087	1.0107	0.9590	0.9662
IC(0.90)	0.4010	0.9000	0.9010	0.7080	0.7390
IC(0.95)	0.5180	0.9440	0.9480	0.8210	0.8390
IC(0.99)	0.7510	0.9910	0.9910	0.9310	0.9410

Tabella 3.6: Confronto tra i cinque metodi di stima. R=1000 replicazioni Monte Carlo, parametro $\psi_1 = -1$. Dimensione del campione: $q = 200$, $m = 12$.

	$\hat{\psi}_2^P$	$\hat{\psi}_2^{MP}$	$\hat{\psi}_2^{ZSE}$	$\hat{\psi}_2^{RP}$	$\hat{\psi}_2^{NP}$
MEAN	3.2788	1.8380	2.3250	2.4446	2.5302
MED	3.2216	1.8341	2.2635	2.3978	2.5010
BIAS	1.2788	-0.1620	0.3250	0.4446	0.5302
MSE	1.9594	0.0602	0.3273	0.3025	0.4255
RMSE	1.3998	0.2454	0.5721	0.5500	0.6523
MAE	1.2794	0.2011	0.3872	0.4551	0.5453
SE/SE	0.7347	1.2821	0.7180	0.1893	0.7615
IC(0.90)	0.0690	0.8590	0.8300	0.1456	0.4590
IC(0.95)	0.1200	0.9240	0.8990	0.1777	0.5790
IC(0.99)	0.3070	0.9790	0.9700	0.2420	0.7920

Tabella 3.7: Confronto tra i cinque metodi di stima. R=1000 repliche Monte Carlo, parametro $\psi_2 = 2$. Dimensione del campione: $q = 100$, $m = 4$.

	$\hat{\psi}_2^P$	$\hat{\psi}_2^{MP}$	$\hat{\psi}_2^{ZSE}$	$\hat{\psi}_2^{RP}$	$\hat{\psi}_2^{NP}$
MEAN	3.1926	1.8304	2.2969	2.3019	2.4765
MED	3.1580	1.8293	2.2503	2.2598	2.4555
BIAS	1.1926	-0.1696	0.2969	0.3019	0.4765
MSE	1.5775	0.0461	0.1841	0.1459	0.2962
RMSE	1.2560	0.2148	0.4291	0.3820	0.5442
MAE	1.1926	0.1826	0.3201	0.3096	0.4800
SE/SD	0.7284	1.2452	0.7407	0.0905	0.7616
IC(0.90)	0.0050	0.7530	0.6960	0.0356	0.2810
IC(0.95)	0.0120	0.8480	0.7890	0.0447	0.3810
IC(.99)	0.0500	0.9520	0.9150	0.0630	0.6040

Tabella 3.8: Confronto tra i cinque metodi di stima. R=1000 repliche Monte Carlo, parametro $\psi_2 = 2$. Dimensione del campione: $q = 200$, $m = 4$.

	$\hat{\psi}_2^P$	$\hat{\psi}_2^{MP}$	$\hat{\psi}_2^{ZSE}$	$\hat{\psi}_2^{RP}$	$\hat{\psi}_2^{NP}$
MEAN	2.5132	2.0289	2.0537	2.3016	2.2643
MED	2.5026	2.0258	2.0483	2.2698	2.2574
BIAS	0.5132	0.0289	0.0537	0.3016	0.2643
MSE	0.3264	0.0313	0.0351	0.1591	0.1148
RMSE	0.5713	0.1768	0.1873	0.3989	0.3388
MAE	0.5148	0.1387	0.1460	0.3193	0.2809
SE/SD	0.8149	0.9786	0.9682	0.7680	0.8682
IC(0.90)	0.2020	0.8970	0.8930	0.5960	0.5920
IC(0.95)	0.3090	0.9520	0.9420	0.7430	0.7110
IC(0.99)	0.5470	0.9860	0.9880	0.8950	0.8950

Tabella 3.9: Confronto tra i cinque metodi di stima. R=1000 replicazioni Monte Carlo, parametro $\psi_2 = 2$. Dimensione del campione: $q = 100$, $m = 8$.

	$\hat{\psi}_2^P$	$\hat{\psi}_2^{MP}$	$\hat{\psi}_2^{ZSE}$	$\hat{\psi}_2^{RP}$	$\hat{\psi}_2^{NP}$
MEAN	2.4891	2.0243	2.0479	2.3218	2.2491
MED	2.4835	2.0215	2.0432	2.3186	2.2442
BIAS	0.4891	0.0243	0.0479	0.3218	0.2491
MSE	0.2658	0.0139	0.0176	0.1287	0.0813
RMSE	0.5156	0.1179	0.1326	0.3587	0.2851
MAE	0.4891	0.0938	0.1032	0.3230	0.2518
SE/SD	0.8811	1.0524	0.9964	0.8625	0.9413
IC(0.90)	0.0390	0.9160	0.9090	0.2580	0.3890
IC(0.95)	0.0710	0.9680	0.9490	0.3660	0.5320
IC(0.99)	0.2100	0.9950	0.9910	0.6130	0.7670

Tabella 3.10: Confronto tra i cinque metodi di stima. R=1000 replicazioni Monte Carlo, parametro $\psi_2 = 2$. Dimensione del campione: $q = 200$, $m = 8$.

	$\hat{\psi}_2^P$	$\hat{\psi}_2^{MP}$	$\hat{\psi}_2^{ZSE}$	$\hat{\psi}_2^{RP}$	$\hat{\psi}_2^{NP}$
MEAN	2.2918	2.0160	2.0174	2.1587	2.1515
MED	2.2816	2.0083	2.0102	2.1480	2.1427
BIAS	0.2918	0.0160	0.0174	0.1587	0.1515
MSE	0.1114	0.0179	0.0179	0.0474	0.0447
RMSE	0.3338	0.1338	0.1339	0.2177	0.2114
MAE	0.2945	0.1058	0.1061	0.1757	0.1702
SE/SD	0.9067	1.0022	1.0053	0.9358	0.9419
IC(0.90)	0.3900	0.9040	0.9050	0.7140	0.7330
IC(0.95)	0.5130	0.9500	0.9480	0.8120	0.8210
IC(0.99)	0.7540	0.9950	0.9940	0.9340	0.9410

Tabella 3.11: Confronto tra i cinque metodi di stima. R=1000 replicazioni Monte Carlo, parametro $\psi_2 = 2$. Dimensione del campione: $q = 100$, $m = 12$.

	$\hat{\psi}_2^P$	$\hat{\psi}_2^{MP}$	$\hat{\psi}_2^{ZSE}$	$\hat{\psi}_2^{RP}$	$\hat{\psi}_2^{NP}$
MEAN	2.2868	2.0105	2.0136	2.1564	2.1460
MED	2.2840	2.0086	2.0117	2.1541	2.1445
BIAS	0.2868	0.0105	0.0136	0.1564	0.1460
MSE	0.0954	0.0090	0.0090	0.0356	0.0322
RMSE	0.3089	0.0948	0.0951	0.1887	0.1796
MAE	0.2868	0.0753	0.0755	0.1613	0.1522
SE/SD	0.9163	1.0118	1.0135	0.9456	0.9528
IC(0.90)	0.1450	0.9100	0.9070	0.5370	0.5740
IC(0.95)	0.2290	0.9610	0.9610	0.6610	0.7140
IC(0.99)	0.4460	0.9900	0.9910	0.8610	0.8850

Tabella 3.12: Confronto tra i cinque metodi di stima. R=1000 replicazioni Monte Carlo, parametro $\psi_2 = 2$. Dimensione del campione: $q = 200$, $m = 12$.

Conclusioni

Si sono confrontate una serie di procedure inferenziali per la stima in presenza di parametri di disturbo incidentali in un contesto di dati binari stratificati. Come ci si attendeva, la verosimiglianza profilo non permette di fare inferenza in modo soddisfacente sul parametro di interesse per via della sua distorsione che rende gli stimatori inconsistenti. Inoltre, il fatto che la dimensione del parametro di disturbo sia legato al numero degli strati e quindi alla dimensione campionaria, costituisce una violazione delle condizioni di un problema di stima regolare e priva la verosimiglianza profilo delle usuali proprietà asintotiche. Una prima soluzione può essere quella di modificare la verosimiglianza profilo in modo da ridurre la distorsione, ottenendo la cosiddetta profilo modificata (Barndorff-Nielsen 1980, 1983). In questa tesi, tale verosimiglianza è stata utilizzata come metodo di riferimento per il confronto delle varie procedure di stima. I risultati del Capitolo 3 confermano quanto già mostrato in Sartori (2003): la profilo modificata permette di giungere a buone stime del parametro di interesse. Questo deriva dal fatto che tale verosimiglianza mantiene le usuali proprietà anche in contesti asintotici a due indici, fintanto che $q/m^3 = o(1)$, dove q indica il numero degli strati e m la dimensione dello strato. Un'alternativa alla profilo modificata per l'inferenza sul parametro di interesse in presenza di parametri di disturbo anche incidentali può essere la verosimiglianza integrata (Severini 2007, 2010), oggetto di questa tesi. Si sono visti più modi per costruire tale verosimiglianza: quello attraverso la parametrizzazione ZSE e una a priori uniforme, quello attraverso la *robust prior* e quello attraverso la *normal prior*. I tre metodi sono quindi stati valutati in un contesto di parametri incidentali e, alla luce dei risultati di simulazione del Capitolo 3, quello basato sulla parametrizzazione

ZSE sembra essere il migliore. Ciò può anche essere supportato dal fatto che la verosimiglianza integrata così costruita è l'unica approssimabile dalla profilo modificata, come accennato nel Capitolo 1. Inoltre, il costruire la verosimiglianza integrata attraverso la parametrizzazione ZSE e specificando una funzione peso per il parametro di interesse uniforme si è rivelato anche il metodo meno dispendioso da un punto di vista computazionale. Infatti, per le verosimiglianze integrate costruite con la *robust prior* e la *normal prior* si è dovuto ricorrere a metodi alternativi addirittura all'interfacciare R con l'ambiente C/C++. Per la loro implementazione è stata utilizzata la libreria TMB, la quale sfrutta ancora il linguaggio C/C++, ma permette di giungere alle derivate analitiche del modello. Questo, unito ad un'integrazione numerica basata sull'approssimazione di Laplace, diminuisce notevolmente i tempi necessari alla stima. Tale approssimazione, però, può essere anche uno dei motivi per cui le prestazioni della verosimiglianza integrata costruita con la *robust prior* e la *normal prior* sembrano essere inferiori a quelle offerte dalla parametrizzazione ZSE, specialmente per valori di m moderati. Infatti, gli standard error ottenuti numericamente da quei due metodi, ad esempio, sembrano sottostimare di molto la variabilità degli stimatori e la copertura degli intervalli alla Wald è distante dai livelli nominali considerati. Quindi, sembra che si tratti del classico *trade-off* tra costo computazionale ed efficienza del metodo e ciò può essere oggetto di studio. In particolare, sarebbe interessante provare a calcolare anche la verosimiglianza integrata costruita attraverso la parametrizzazione ZSE con TMB in modo da confrontare le due implementazioni. Così facendo, si potrebbe quindi far luce sulla questione della sottostima della variabilità degli stimatori.

Si vuole far notare anche come la *robust prior* per il modello logit usata in questa tesi è diversa da quella utilizzata da Arelanno e Bonhomme (2009), pur partendo dalla stessa espressione teorica data da (1.9). Questo può quindi essere oggetto di un ulteriore approfondimento da condurre. In più, si potrebbero anche delineare con maggior dettaglio le proprietà asintotiche della verosimiglianza integrata costruita attraverso la *robust prior* e la *normal prior*, sfruttando, ad esempio, una qualche sua possibile approssimazione, come fatto da De Bin, Sartori e Severini (2015) per l'integrata costruita tramite

parametrizzazione ZSE. Infine, gli studi di simulazione proposti potrebbero essere estesi anche ad altre funzioni legame, come, ad esempio, al modello probit. Purtroppo, per motivi di tempo, ciò non è stato fatto. Comunque, il codice è scritto in modo generale e questi ulteriori studi di simulazione si possono condurre facilmente.

Bibliografia

Arellano M.; Bonhomme S. (2009). Robust priors in nonlinear panel data models. *Econometrica*, **77**, 489-536.

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

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

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

Barndorff-Nielsen O. E.; Cox D. R. (1994). *Inference and Asymptotics*. CRC Press, Boca Raton.

Bellio R.; Sartori N. (2003). Extending conditional likelihood in models for stratified binary data. *Statistical Methods & Applications*, **12**, 121-132.

De Bin R.; Sartori N.; Severini T. (2015). Integrated likelihoods in models with stratum nuisance parameters. *Electronic Journal of Statistics*, **9**, 1474-1491.

Fisher R. A. (1922). On the mathematical foundation of theoretical statistics. *Phil. Trans. Roy. Soc., A*, 309–368.

Liu Q.; Pierce D. A. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, **81**, 624-629.

Neyman J.; Scott E. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, **16**, 33.

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

Pace L.; Salvan A. (2001). *Introduzione alla Statistica. II Inferenza, verosimiglianza, modelli*. CEDAM, Padova.

Pace L.; Salvan A. (2006). Adjustments of the profile likelihood from a new perspective. *Journal of Statistical Planning and Inference*, **136**, 3554-3564.

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

Severini T. A. (1998). An approximation to the modified profile likelihood function. *Biometrika*, **85**, 403-411.

Severini T. A. (2007). Integrated likelihood functions for non-Bayesian inference. *Biometrika*, **94**, 529-542.

Severini T. A. (2010). Likelihood ratio statistics based on an integrated likelihood. *Biometrika*, **97**, 481-496.

Codice

Generazione dati per la simulazione: R

Listing 1: Dati simulati

```
library(panelMPL2)
#numero strati; dimensione strati; replicazioni Monte Carlo; numero
  esplicative; vero parametro
q=100
m=4
R=1000
n_espl=2
beta0=c(-1,2)
#esplicative
panelInd <- rep(1:q,each=m)
set.seed(1991)
X <- matrix(rnorm(q * m * n_espl), nrow=q * m, ncol=n_espl)
alpha <- tapply(X[,1],panelInd,mean) + rnorm(q)
dati=matrix(0,nrow=m*q,ncol=R)
for(i in 1:R)
{
  y <- simu.panelMPL(X, panelInd, "statLogit", beta=beta0, alpha=alpha)
  dati[,i]=y
}
write.table(dati,"dati_4.csv",sep=" ",row.names=FALSE,col.names=FALSE)
write.table(X,"Xo_4.csv",sep=" ",row.names=FALSE,col.names=FALSE)
```

Verosimiglianza integrata: C/C++

Listing 2: Funzioni comuni alle tre Verosimiglianze integrate considerate

```

#include <Rcpp.h>
#include <R.h>
#include <Rmath.h>
#include <Rinternals.h>
#include <Rdefines.h>
#include <R_ext/Applic.h>
#include <R_ext/Lapack.h>
using namespace Rcpp;

//definisco classe per gestire dati
//della singola u.s. con funzioni
//per inserire ed accedere ad essi
class DATA_i
{
    NumericVector psi, yi, mle;
    NumericMatrix Xi;
    int mi,dimPsi;

    public:void insert(NumericVector, NumericVector, NumericMatrix, int, int
        , NumericVector);
    public:NumericVector get_psi();
    public:NumericVector get_yi();
    public:NumericMatrix get_Xi();
    public:int get_mi();
    public:int get_dimPsi();
    public:NumericVector get_mle();

};

void DATA_i::insert(NumericVector psii, NumericVector yii, NumericMatrix
    Xii, int mii, int dimPsii, NumericVector mlei)
{

```

```
    psi=psii;
    yi=yii;
    Xi=Xii;
    mi=mii;
    dimPsi=dimPsii;
    mle=mlei;
}

NumericVector DATA_i::get_psi()
{
    return(psi);
}

NumericVector DATA_i::get_yi()
{
    return(yi);
}

NumericMatrix DATA_i::get_Xi()
{
    return(Xi);
}

int DATA_i::get_mi()
{
    return(mi);
}

int DATA_i::get_dimPsi()
{
    return(dimPsi);
}

NumericVector DATA_i::get_mle()
{
```

```
    return(mle);
}

// [[Rcpp::export]]
double prange(double p)
{
    //funzione per evitare 0 e 1 esatti e di
    //conseguenza instabilit numeriche
    double eps=2.22e-15;
    double mu;
    mu=p;

    if (mu<eps)
        mu=eps;
    if (mu>(1-eps))
        mu=(1-eps);
    return(mu);
}

// [[Rcpp::export]]
double llik_logit_i(NumericVector psi, double lambdai, NumericVector yi,
    NumericMatrix Xi, int mi, int dimPsi)
{
    //parametrizzazione (psi,lambda)
    //i-esimo contributo alla log-verosimiglianza

    double out=0, eta=0, mu=0;
    int i=0, j=0;
    for(i=0; i<mi; i++)
    {
        eta=lambdai;
        for(j=0; j<dimPsi; j++)
        {
            eta+=Xi(i, j)*psi(j);
        }
    }
}
```

```
    mu=R::plogis(eta,0,1,1,0);
    mu=prange(mu);
    out+=yi(i)*log(mu)+(1-yi(i))*log(1-mu);
}
return(out);
}

// [[Rcpp::export]]
double score_logit_i(double lambdai, NumericVector psi, NumericVector yi,
    NumericMatrix Xi, int mi, int dimPsi)
{
    //derivata prima rispetto a lambda_i

    double out=0, eta=0, mu=0;
    int i=0,j=0;
    for(i=0;i<mi;i++)
    {
        eta=lambdai;
        for(j=0;j<dimPsi;j++)
        {
            eta+=Xi(i,j)*psi(j);
        }
        mu=R::plogis(eta,0,1,1,0);
        mu=prange(mu);
        out+=yi(i)-mu;
    }
    return(out);
}

// [[Rcpp::export]]
double hess_logit_i(double lambdai, NumericVector psi, NumericVector yi,
    NumericMatrix Xi, int mi, int dimPsi)
{
    //derivata seconda rispetto a lambda_i
```

```

double out=0, eta=0, mu=0;
int i=0,j=0;
for(i=0;i<mi;i++)
{
    eta=lambdai;
    for(j=0;j<dimPsi;j++)
    {
        eta+=Xi(i,j)*psi(j);
    }
    mu=R::plogis(eta,0,1,1,0);
    mu=prange(mu);
    out+=mu*(mu-1);
}
return(out);
}

// [[Rcpp::export]]
double NR_logit(double init, NumericVector psi, NumericVector yi,
    NumericMatrix Xi, int mi,
    int dimPsi)
{
    //Newton-Raphson per ottenere lambda_i(phi_i) con pseudo-dati
    //massimizzando la llik per psi fissato rispetto a lambda_i (ZSE)
    //Newton-Raphson per ottenere le SMV di lambda data SMV di psi
    //Newton-Raphson per ottenere le stime vincolate di lambda_i(psi)

    int maxiter=100,i=0,j=0;
    double lambdanew=0,lambdaold=0,tol=0.000001,diff=0,s1=0,s2=0,lam=0,f0=0,
        f1=0,g0=0,g1=0,delta=0,H=0;

    lambdaold=init;
    g0=score_logit_i(lambdaold, psi, yi, Xi, mi, dimPsi);
    f0=R_pow_di(g0,2) * 0.5;

    while(i<maxiter)

```

```
{
  i+=1;
  g1=score_logit_i(lambdaold, psi, yi, Xi, mi, dimPsi);
  H=1/hess_logit_i(lambdaold, psi, yi, Xi, mi, dimPsi);
  delta = H * g1;
  lambdanew=lambdaold-delta;

  f1= R_pow_di(g1,2) * 0.5;
  j=0;
  lam=-2.0*f0;

  while(f1>(f0+0.0001*lam))
  {
    j+=1;
    delta = delta * 0.5;
    lambdanew= lambdaold - delta;
    lam = lam * 0.5;

    g1=score_logit_i(lambdanew, psi, yi, Xi, mi, dimPsi);
    f1= R_pow_di(g1,2) * 0.5;

    if (j>100)
    {
      break;
    }
  }

  diff = lambdaold - lambdanew;
  s1 = fabs(g1);
  if (s1< tol)
  {
    break;
  }
  s1 = fabs(diff);
```

```

    s2 = fabs(lambdanew);
    if ( (s1/(1+s2))< tol)
    {
        break;
    }
    lambdaold = lambdanew;
    g0=g1;
    f0=f1;
}
return(lambdanew);
}

```

Listing 3: Verosimiglianza integrata con parametrizzazione ZSE

```

void ZSE_integrand_logit_i(double *x, int n, void *ex)
{

    double eta=0,out=0,mu=0,lambdai=0, phi=0;
    int i=0,j=0, r=0, mi=0, dimPsi=0;

    DATA_i *parp;
    DATA_i par;

    //trasformo ex come puntatore a dato di tipo DATA_i
    parp = (DATA_i *) ex;
    //assegno a par l'oggetto indicato dal puntatore
    par=*parp;

    mi=par.get_mi();
    dimPsi=par.get_dimPsi();

    NumericVector psi(dimPsi), yi(mi), mle(dimPsi);
    NumericMatrix Xi(mi,dimPsi);

    psi=par.get_psi();
    yi=par.get_yi();
}

```

```

Xi=par.get_Xi();
mle=par.get_mle();

NumericVector pseudo_y(mi);

for(r=0;r<n;r++)
{
  //pseudo-dati contenenti phi
  phi=x[r];
  for(i=0;i<mi;i++)
  {
    eta=phi;
    for(j=0;j<dimPsi;j++)
    {
      eta+=Xi(i,j)*mle(j);
    }
    mu=R::plogis(eta,0,1,1,0);
    mu=prange(mu);
    pseudo_y[i]=mu;
  }

  lambdai=NR_logit(0, psi, pseudo_y, Xi, mi, dimPsi);
  out=llik_logit_i(psi,lambdai,yi,Xi,mi,dimPsi);
  out=exp(out);
  x[r]=out;
}
}

extern "C" SEXP int_llik_logit_ZSE(SEXP psie, SEXP ye, SEXP Xe, SEXP qe,
                                SEXP me, SEXP dimPsie, SEXP mlee)
{
  NumericVector psi=as<NumericVector>(psie);
  NumericVector y=as<NumericVector>(ye);
  NumericMatrix X=as<NumericMatrix>(Xe);
  int q=as<int>(qe);

```

```

IntegerVector m=as<IntegerVector>(me);
int dimPsi=as<int>(dimPsie);
NumericVector mle=as<NumericVector>(mlee);

double inti=0,abserr=0,epsabs=0,epsrel=0,eps=2.22e-15,bound=0,llint=0;
int i=0,j=0,k=0,mi=0,mcum=0,neval=0,ier=0,limit=300,lenw=0,last=0,inf=2;
void *ex;
DATA_i par;

abserr=eps;
epsabs=eps;
epsrel=eps;

lenw = 4*limit;
int *iwork=new int[limit];
double *work=new double[lenw];

for(i=0;i<q;i++)
{
  //recupero dati i-esima u.s.
  mi=m(i);
  NumericVector yi(mi);
  NumericMatrix Xi(mi,dimPsi);

  for(j=0;j<mi;j++)
  {
    yi(j)=y(j+mcum);
    for(k=0;k<dimPsi;k++)
    {
      Xi(j,k)=X(j+mcum,k);
    }
  }
  mcum+=mi;

  par.insert(psi,yi,Xi,mi,dimPsi,mle);

```

```

//assegno a ex l'indirizzo di par
ex = (void *) &par;

Rdqagi(ZSE_integrand_logit_i,ex,&bound,&inf,&epsabs,&epsrel,&inti,&
      abserr,&neval,&ier,&limit,&lenw,&last,iwork,work);

if(inti<eps)
  inti=eps;

  llint+=log(inti);
}
return(wrap(llint));
}

```

Listing 4: Verosimiglianza integrata con robust prior

```

void RP_integrand_logit_i(double *x, int n, void *ex)
{
  double eta=0,out=0,mu=0,lambdai_mle=0,E_nhess=0,E_score2=0,etai=0,etau
    =0,etai_mle=0,etau_mle=0,
    mui=0,muu=0,mui_mle=0,muu_mle=0,rp=0,lambdai=0;
  int i=0,j=0,u=0,r=0,mi=0,dimPsi=0;

  DATA_i *parp;
  DATA_i par;

  //trasformo ex come puntatore a dato di tipo DATA_i
  parp = (DATA_i *) ex;
  //assegno a par l'oggetto indicato dal puntatore
  par=*parp;

  mi=par.get_mi();
  dimPsi=par.get_dimPsi();

  NumericVector psi(dimPsi), yi(mi), mle(dimPsi);

```

```
NumericMatrix Xi(mi,dimPsi);

psi=par.get_psi();
yi=par.get_yi();
Xi=par.get_Xi();
mle=par.get_mle();

for(r=0;r<n;r++)
{
  lambdai=x[r];
  //calcolo di E(-hess)
  for(i=0;i<mi;i++)
  {
    eta=lambdai;
    for(j=0;j<dimPsi;j++)
    {
      eta+=Xi(i,j)*psi(j);
    }
    mu=R::plogis(eta,0,1,1,0);
    mu=prange(mu);

    E_nhess+=mu*(1-mu);
  }

  //SMV di lambda
  lambdai_mle=NR_logit(0.0,mle,yi,Xi,mi,dimPsi);

  //calcolo di E(score^2)
  for(i=0;i<mi;i++)
  {
    etai_mle=lambdai_mle;
    etai=lambdai;
    for(j=0;j<dimPsi;j++)
    {
      etai_mle+=Xi(i,j)*mle(j);
    }
  }
}
```

```
    etai+=Xi(i,j)*psi(j);
}

mui_mle=R::plogis(etai_mle,0,1,1,0);
mui_mle=prange(mui_mle);
mui=R::plogis(etai,0,1,1,0);
mui=prange(mui);

for(u=0;u<mi;u++)
{
    etau_mle=lambdai_mle;
    etau=lambdai;
    for(j=0;j<dimPsi;j++)
    {
        etau_mle+=Xi(u,j)*mle(j);
        etau+=Xi(u,j)*psi(j);
    }

    muu_mle=R::plogis(etau_mle,0,1,1,0);
    muu_mle=prange(muu_mle);
    muu=R::plogis(etau,0,1,1,0);
    muu=prange(muu);

    if(u==i)
    {
        E_score2+=mui_mle-2*mui_mle*mui+mui*mui;
    }
    else
    {
        E_score2+=mui_mle*muu_mle-mui_mle*muu-
            muu_mle*mui+mui*muu;
    }
}
}
```

```

        //ritorna exp(l_i)
        rp=E_nhess*R_pow(E_score2,-0.5);
        out=llik_logit_i(psi,lambdai,yi,Xi,mi,dimPsi);
        out=exp(out)*rp;
        x[r]=out;
    }
}

extern "C" SEXP int_llik_logit_RP(SEXP psie, SEXP ye, SEXP Xe, SEXP qe,
                                SEXP me, SEXP dimPsie, SEXP mlee)
{
    NumericVector psi=as<NumericVector>(psie);
    NumericVector y=as<NumericVector>(ye);
    NumericMatrix X=as<NumericMatrix>(Xe);
    int q=as<int>(qe);
    IntegerVector m=as<IntegerVector>(me);
    int dimPsi=as<int>(dimPsie);
    NumericVector mle=as<NumericVector>(mlee);

    double inti=0,abserr=0,epsabs=0,epsrel=0,eps=2.22e-15,bound=0,llint=0;
    int i=0,j=0,k=0,mi=0,mcum=0,neval=0,ier=0,limit=300,lenw=0,last=0,inf=2;
    void *ex;
    DATA_i par;

    abserr=eps;
    epsabs=eps;
    epsrel=eps;

    lenw = 4*limit;
    int *iwork=new int[limit];
    double *work=new double[lenw];

    for(i=0;i<q;i++)
    {
        //recupero dati i-esima u.s.

```

```
mi=m(i);
NumericVector yi(mi);
NumericMatrix Xi(mi,dimPsi);

for(j=0;j<mi;j++)
{
  yi(j)=y(j+mcum);
  for(k=0;k<dimPsi;k++)
  {
    Xi(j,k)=X(j+mcum,k);
  }
}
mcum+=mi;

par.insert(psi,yi,Xi,mi,dimPsi,mle);

//assegno a ex l'indirizzo di par
ex = (void *) &par;

Rdqagi(RP_integrand_logit_i,ex,&bound,&inf,&epsabs,&epsrel,&inti,&
  abserr,&neval,&ier,&limit,&lenw,&last,iwork,work);

if(inti<eps)
{
  inti=eps;
}

llint+=log(inti);
}
return(wrap(llint));
}
```

Listing 5: Verosimiglianza integrata con normal prior

```
void N_integrand_logit_i(double *x, int n, void *ex)
{
```

```
int r=0,mi=0,dimPsi=0,k=0,j=0;
double rp=0,lambda_v=0,V=0,v1=0,v2=0,v3=0,lambda_i=0,lambda_mle=0,mu=0,
eta=0,temp=0;

DATA_i *parp;
DATA_i par;

//trasformo ex come puntatore a dato di tipo DATA_i
parp = (DATA_i *) ex;
//assegno a par l'oggetto indicato dal puntatore
par=*parp;

mi=par.get_mi();
dimPsi=par.get_dimPsi();

NumericVector psi(dimPsi), yi(mi), mle(dimPsi);
NumericMatrix Xi(mi,dimPsi);

psi=par.get_psi();
yi=par.get_yi();
Xi=par.get_Xi();
mle=par.get_mle();

for(r=0;r<n;r++)
{
    lambda_i=x[r];
    lambda_v=NR_logit(0.0,psi,yi,Xi,mi,dimPsi);
    lambda_mle=NR_logit(0,mle,yi,Xi,mi,dimPsi);

    for(k=0;k<mi;k++)
    {
        eta=lambda_mle;
        for(j=0;j<dimPsi;j++)
        {
            eta+=Xi(k,j)*mle(j);
```

```

    }
    mu=R::plogis(eta,0,1,1,0);
    mu=prange(mu);
    v1+=mu*(1-mu);
  }
  v1=R_pow_di(v1,2);

  v2=-hess_logit_i(lambda_v,psi,yi,Xi,mi,dimPsi);
  v2=R_pow_di(v2,-2);

  v3=-hess_logit_i(lambda_mle,mle,yi,Xi,mi,dimPsi);
  v3=R_pow_di(v3,-1);

  V=v1*v2*v3;
  V=R_pow(V,0.5);

  rp=R::dnorm(lambdai,lambda_v,V,0);

  temp=llik_logit_i(psi,lambdai,yi,Xi,mi,dimPsi);
  x[r]=exp(temp)*rp;
}
}

extern "C" SEXP int_llik_logit_N(SEXP psie, SEXP ye, SEXP Xe, SEXP qe,
                               SEXP me, SEXP dimPsie, SEXP mlee)
{
  NumericVector psi=as<NumericVector>(psie);
  NumericVector y=as<NumericVector>(ye);
  NumericMatrix X=as<NumericMatrix>(Xe);
  int q=as<int>(qe);
  IntegerVector m=as<IntegerVector>(me);
  int dimPsi=as<int>(dimPsie);
  NumericVector mle=as<NumericVector>(mlee);

  double inti=0,abserr=0,epsabs=0,epsrel=0,eps=2.22e-15,bound=0,llint=0;

```

```

int i=0,j=0,k=0,mi=0,mcum=0,neval=0,ier=0,limit=300,lenw=0,last=0,inf=2;
void *ex;
DATA_i par;

abserr=eps;
epsabs=eps;
epsrel=eps;

lenw = 4*limit;
int *iwork=new int[limit];
double *work=new double[lenw];

for(i=0;i<q;i++)
{
    //recupero dati i-esima u.s.
    mi=m(i);
    NumericVector yi(mi);
    NumericMatrix Xi(mi,dimPsi);

    for(j=0;j<mi;j++)
    {
        yi(j)=y(j+mcum);
        for(k=0;k<dimPsi;k++)
        {
            Xi(j,k)=X(j+mcum,k);
        }
    }
    mcum+=mi;

    par.insert(psi,yi,Xi,mi,dimPsi,mle);

    //assegno a ex l'indirizzo di par
    ex = (void *) &par;

    Rdqagi(N_integrand_logit_i,ex,&bound,&inf,&epsabs, &epsrel,&inti,&

```

```
    abserr,&neval,&ier,&limit,&lenw,&last,iwork,work);

    if(inti<eps)
        inti=eps;

    llint+=log(inti);
}
return(wrap(llint));
}
```

Verosimiglianza integrata con robust e normal prior: TMB

Listing 6: Verosimiglianza integrata con robust prior: TMB

```
#include <TMB.hpp>

template<class Type>
Type objective_function<Type>::operator() ()
{
    //dati
    DATA_INTEGER(q);
    DATA_IVECTOR(m);
    DATA_VECTOR(y);
    DATA_MATRIX(X);
    DATA_VECTOR(psi_mle);
    DATA_VECTOR(lambda_mle);

    //parametri
    PARAMETER_VECTOR(psi);
    PARAMETER_VECTOR(lambda);

    using namespace density;
```

```
Type nll=0.0;

vector<Type> Xpsi=X*psi;
vector<Type> Xpsi_mle=X*psi_mle;
int ind=0,u=0;

for(int i=0;i<q;i++)
{
    //calcolo della priori
    Type E_nhess=0.0;
    Type E_score2=0.0;
    for(int j=0;j<m(i);j++)
    {
        int k=j+ind;
        Type eta=Xpsi(k)+lambda(i);
        Type eta_mle=Xpsi_mle(k)+lambda_mle(i);

        Type prob=exp(eta)/(1+exp(eta));

        Type prob_mle=exp(eta_mle)/(1+exp(eta_mle));

        for(u=0;u<m(i);u++)
        {
            int k2=u+ind;
            Type etau=Xpsi(k2)+lambda(i);
            Type etau_mle=Xpsi_mle(k2)+lambda_mle(i);

            Type probu=exp(etau)/(1+exp(etau));

            Type probu_mle=exp(etau_mle)/(1+exp(etau_mle));

            if(u==j)
            {
                E_score2+=prob_mle-2*prob_mle*prob+prob*prob;
            }
        }
    }
}
```

```
    else
    {
        E_score2+=prob_mle*probu_mle-prob_mle*probu-probu_mle*prob+prob*
            probu;
    }
}

E_nhess+=prob*(1-prob);

nll-=dbinom(y(k),Type(1),prob,true);
}

Type prior=E_nhess*exp(-0.5*log(E_score2));
nll-=log(prior);
ind+=m(i);
}
return nll;
}
```

Listing 7: Verosimiglianza integrata con normal prior: TMB

```
#include <TMB.hpp>

template<class Type>
Type objective_function<Type>::operator() ()
{
    //dati
    DATA_INTEGER(q);
    DATA_IVECTOR(m);
    DATA_VECTOR(y);
    DATA_MATRIX(X);
    DATA_VECTOR(psi_mle);
    DATA_VECTOR(lambda_mle);

    //parametri
    PARAMETER_VECTOR(psi);
```

```
PARAMETER_VECTOR(lambda);

using namespace density;

Type nll=0.0;

vector<Type> Xpsi=X*psi;
vector<Type> Xpsi_mle=X*psi_mle;
int ind=0,u=0;

for(int i=0;i<q;i++)
{
    //calcolo della priori
    Type v1=0.0;
    Type v2=0.0;
    Type v3=0.0;
    matrix<Type> d_mix(1,X.cols()); //derivata mista per linearizzazione

    for(int j=0;j<m(i);j++)
    {
        int k=j+ind;
        Type eta=Xpsi(k)+lambda(i);
        Type eta_mle=Xpsi_mle(k)+lambda_mle(i);

        Type prob=exp(eta)/(1+exp(eta));
        Type prob_mle=exp(eta_mle)/(1+exp(eta_mle));

        v1+=prob_mle*(1-prob_mle);
        v3+=prob_mle*(prob_mle-1);

        d_mix=d_mix-(prob_mle*(1-prob_mle))*matrix<Type>(X.row(k));

        nll-=dbinom(y(k),Type(1),prob,true);
    }
}
```

```
matrix<Type> num(1,1);
matrix<Type> pmp(X.cols(),1);

pmp=matrix<Type>(psi-psi_mle);
num=d_mix*pmp;

//lambda per psi fissato con linearizzazione
Type lambda_v=Type(lambda_mle(i))-num(0,0)/v3;

//calcolo di v2
for(int j=0;j<m(i);j++)
{
    int k=j+ind;
    Type eta_v=Xpsi(k)+lambda_v;
    Type prob_v=exp(eta_v)/(1+exp(eta_v));
    v2+=prob_v*(prob_v-1);
}

v1=pow(v1,2.0);
v2=pow(-v2,-2);
v3=pow(-v3,-1);

//varianza della d. normale
Type V=v1*v2*v3;
V=pow(V,0.5);
Type prior=dnorm(lambda(i),lambda_v,V,true);

nll-=prior;
ind+=m(i);
}
return nll;
}
```

Simulazione: R

Listing 8: Simulazione per verosimiglianza profilo, profilo modificata e integrata con ZSE: calcolo in parallelo

```
#librerie per il calcolo in parallelo
library(plyr)
library(doMC)

library(Rcpp)
library(panelMPL2)
library(numDeriv)

#caricamento funzioni in C/C++
sourceCpp("Logit_v3.cpp")
dyn.load("Logit_v3.so")

q=100
m=4
R=1000
n_espl=2

#lettura dati simulati in precedenza
Xo=read.csv("Xo_4.csv",header=FALSE)
dati=read.csv("dati_4.csv",header=FALSE)

intlik_logit_ZSE=function(psi,y,X,panelInd,mle)
{
  #chiama funzione C e restituisce la log-verosimiglianza
  #integrata in psi cambiata di segno

  q=length(unique(panelInd))
  m=tapply(y,panelInd,length)
  dimPsi=ncol(X)

  out=.Call("int_llik_logit_ZSE",psi,y,X,q,m,dimPsi,mle)
```

```
    return(-out)
  }

logitPL<- function(psi,y,X,panelInd)
{
  #verosimiglianza profilo per std. error mle
  alpha<-dynLogitAlpha(X,panelInd,y,psi,rep(0,length(unique(panelInd))))
  eta<- X %*% psi+rep(alpha,rep(m,length(unique(panelInd))))
  p<- p.range(plogis(eta))
  out<- sum(y*log(p)+(1-y)*log(1-p))
  return(-out)
}

#creo dataset come lista avente y e Xo
dataset=as.list(numeric(R))

for(i in 1:R)
{
  dataset[[i]]=list(id=i,y=dati[,i])
}

fun_single=function(data,Xo,m,q)
{
  print(data$id)
  y=data$y
  panelInd <- rep(1:q,each=m)
  X=as.matrix(Xo)
  #tolgo strati non informativi
  mean.y=tapply(y,panelInd,mean)
  cond=(mean.y!=0)&(mean.y!=1)
  ind=match(panelInd,names(cond))
  sele=cond[ind]
  panelInd=panelInd[sele]
  y=y[sele]
```

```

X=X[sele,]

obj=panelMPL(y~X, panel=panelInd, model="statLogit")
mle=obj$mle
mp=obj$estimate
se_mp=sqrt(diag(solve((-obj$hessian))))
hes=numDeriv::hessian(logitPL,mle,y=y,X=X,panelInd=panelInd)
se_mle=sqrt(diag(solve((hes))))

est.ZSE=nlminb(mle, intlik_logit_ZSE,y=y,X=X,panelInd=panelInd,mle=mle,
  lower=mle-6*se_mp,upper=mle+6*se_mp)
zse=est.ZSE$par
hes=numDeriv::hessian(intlik_logit_ZSE,zse,y=y,X=X,panelInd=panelInd,mle
  =mle)
se_zse=sqrt(diag(solve(hes)))

data.frame(
  MLE_est1=mle[1],MLE_est2=mle[2],
  MLE_se1=se_mle[1],MLE_se2=se_mle[2],
  MP_est1=mp[1],MP_est2=mp[2],
  MP_se1=se_mp[1],MP_se2=se_mp[2],
  ZSE_est1=zse[1],ZSE_est2=zse[2],
  ZSE_se1=se_zse[1],ZSE_se2=se_zse[2]
)
}

registerDoMC(detectCores())
out=ldply(.data=dataset,.fun=fun_single,Xo=Xo,m=m,q=q,.parallel=TRUE)

```

Listing 9: Simulazione per verosimiglianza integrata con robust prior

```

library(panelMPL2)
library(TMB)
library(numDeriv)

q=100

```

```
m=4
R=1000
n_espl=2

#lettura dati simulati in precedenza
Xo=read.csv("Xo_4.csv",header=FALSE)
dati=read.csv("dati_4.csv",header=FALSE)

#compilazione del codice scritto con TMB
compile("RP_prior_v2.cpp")
dyn.load("RP_prior_v2")

RP_est=matrix(0,nrow=R,ncol=n_espl)
RP_se=matrix(0,nrow=R,ncol=n_espl)

for(i in 1:R)
{
  cat("Iterazione: ",i,"\n")
  #carico dati da .csv, stesso per tutte le priori
  y=dati[,i]
  panelInd <- rep(1:q,each=m)
  X=as.matrix(Xo)
  #tolgo strati non informativi
  mean.y=tapply(y,panelInd,mean)
  cond=(mean.y!=0)&(mean.y!=1)
  ind=match(panelInd,names(cond))
  sele=cond[ind]
  panelInd=panelInd[sele]
  y=y[sele]
  X=X[sele,]

  obj=panelMPL(y~X, panel=panelInd, model="statLogit")
  mle=obj$mle

  #SMV lambda per RP prior
```

```

lambda_mle=dynLogitAlpha(X,panelInd,y,mle,rep(0,length(unique(panelInd))
))

parameters=list(psi=rep(0,n_espl), lambda=rep(0,length(unique(panelInd))
))
data=list(q=length(unique(panelInd)),m=rep(m,length(unique(panelInd))),y
=as.vector(y),
        X=as.matrix(X),psi_mle=mle,lambda_mle=lambda_mle)

rp_tmb=MakeADFun(data=data,parameters=parameters,DLL="RP_prior_v2",
random=c("lambda"),silent=TRUE)

est.RP=nlminb(rp_tmb$par,rp_tmb$fn,rp_tmb$gr)
rp=est.RP$par
hes=numDeriv::hessian(rp_tmb$fn,rp)
se_rp=sqrt(diag(solve(hes)))

RP_est[i,]=est.RP$par
RP_se[i,]=se_rp

}

```

Listing 10: Simulazione per verosimiglianza integrata con normal prior

```

library(panelMPL2)
library(TMB)
library(numDeriv)

q=100
m=4
R=1000
n_espl=2

#lettura dati simulati in precedenza
Xo=read.csv("Xo_4.csv",header=FALSE)
dati=read.csv("dati_4.csv",header=FALSE)

```

```
#Compilazione del codice scritto con TMB
compile("N_prior.cpp")
dyn.load("N_prior")

N_est=matrix(0,nrow=R,ncol=n_espl)
N_se=matrix(0,nrow=R,ncol=n_espl)

for(i in 1:R)
{
  #carico dati da .csv, stesso per tutte le priori
  y=dati[,i]
  panelInd <- rep(1:q,each=m)
  X=as.matrix(Xo)
  #tolgo strati non informativi
  mean.y=tapply(y,panelInd,mean)
  cond=(mean.y!=0)&(mean.y!=1)
  ind=match(panelInd,names(cond))
  sele=cond[ind]
  panelInd=panelInd[sele]
  y=y[sele]
  X=X[sele,]

  obj=panelMPL(y~X, panel=panelInd, model="statLogit")
  mle=obj$mle

  #SMV lambda per N prior
  lambda_mle=dynLogitAlpha(X,panelInd,y,mle,rep(0,length(unique(panelInd)))
  )

  parameters=list(psi=rep(0,n_espl), lambda=rep(0,length(unique(panelInd)))
  )
  data=list(q=length(unique(panelInd)),m=rep(m,length(unique(panelInd))),y
  =as.vector(y),
  X=as.matrix(X),psi_mle=mle,lambda_mle=lambda_mle)
```

```
n_tmb=MakeADFun(data=data,parameters=parameters,DLL="N_prior",random=c("
  lambda"),silent=TRUE)

est.N=nlminb(n_tmb$par,n_tmb$fn,n_tmb$gr)
n=est.N$par
hes=numDeriv::hessian(n_tmb$fn,n)
se_n=sqrt(diag(solve(hes)))

N_est[i,]=est.N$par
N_se[i,]=se_n

}
```
