

UNIVERSITÀ DEGLI STUDI DI PADOVA

---

CORSO DI LAUREA MAGISTRALE  
IN SCIENZE STATISTICHE

TESI DI LAUREA

# BOOTSTRAP PARAMETRICO IN MODELLI STRATIFICATI

RELATORE: PROF. NICOLA SARTORI

Dipartimento di Scienze Statistiche

CORRELATORE: PROF. RUGGERO BELLIO

Dipartimento di Scienze Economiche e Statistiche

Università degli Studi di Udine

LAUREANDO: LEONARDO MANFROTTO

---

ANNO ACCADEMICO 2011/2012



*A chi mi ha voluto qui, a chi non mi ha  
creduto qui.*



# Indice

<b>Introduzione</b>	VII
<b>1 L'inferenza di verosimiglianza</b>	1
1.1 Introduzione . . . . .	1
1.2 Specificazione del modello . . . . .	2
1.3 La funzione di verosimiglianza . . . . .	4
1.4 Parametri di disturbo . . . . .	6
1.5 Invarianza a riparametizzazioni . . . . .	7
1.6 Pseudo-verosimiglianze . . . . .	8
1.6.1 Verosimiglianza marginale e condizionata . . . . .	9
1.6.2 Verosimiglianza profilo . . . . .	11
1.6.3 Verosimiglianza profilo modificata . . . . .	14
1.7 Inferenza basata sulla verosimiglianza . . . . .	17
1.8 Bootstrap . . . . .	22
<b>2 Soluzioni per i problemi di Neyman e Scott</b>	27
2.1 Introduzione . . . . .	27
2.2 Analisi a due indici asintotici . . . . .	30
2.3 Miglioramenti via bootstrap . . . . .	34
2.3.1 Bootstrap per problemi con parametri di disturbo . . . . .	34
2.3.2 Prepivoting . . . . .	36
2.3.3 Discussione . . . . .	40
<b>3 Studi di simulazione</b>	41
3.1 Introduzione . . . . .	41

3.2	Modello gamma . . . . .	43
3.3	Modello normale . . . . .	52
3.4	Modello per dati binari . . . . .	62
3.4.1	Modello logit . . . . .	64
3.4.2	Modello probit . . . . .	72
	<b>Conclusioni</b>	<b>79</b>
<b>A</b>	<b>Ordini di successioni e in probabilità</b>	<b>81</b>

# Introduzione

Nell'ambito dell'inferenza statistica parametrica gli aspetti di interesse sono rappresentati, generalmente, da un numero di parametri inferiore al numero totale degli stessi, i quali vengono definiti parametri di interesse. I parametri restanti, detti parametri di disturbo, sono utili per rendere maggiormente completa e adeguata la specificazione del modello. Negli anni sono stati proposti molti metodi per attenuare l'incidenza dei parametri di disturbo nei risultati dell'inferenza statistica. Si tratta perlopiù di metodi derivati dal concetto di pseudo-verosimiglianza.

Esistono due situazioni diametralmente opposte per quanto riguarda la natura del parametro di disturbo: la sua dimensione può rimanere fissata o dipendere dall'aumento della numerosità campionaria  $n$ . Sebbene nel primo caso, in un contesto asintotico standard, l'aumento della numerosità campionaria comporta conseguentemente un aumento dell'informazione disponibile nel campione, ciò non risulta altrettanto efficace nel secondo caso.

In situazioni in cui la dimensione del parametro di disturbo è troppo elevata rispetto all'informazione disponibile, ci si riferirà alle sue componenti come a dei parametri incidentali. Per contro, la componente di interesse del parametro totale sarà definita come parametro strutturale. I problemi con presenza di parametri incidentali sono noti in letteratura come **problemi di Neyman e Scott**, in onore degli autori che per primi li hanno affrontati e che hanno messo in evidenza le problematiche nell'utilizzo delle usuali tecniche basate sulla verosimiglianza (Neyman e Scott, 1948).

In econometria si fa largo uso di modelli che possono presentare le caratteristiche dei problemi sopracitati. Si tratta di analizzare insiemi di dati che presentano un raggruppamento naturale, come possono essere le osservazioni di  $q$  diversi individui in  $m$  diversi istanti di tempo. Il parametro strutturale può riguardare l'insieme

globale dei dati, mentre i vari parametri incidentali possono riferirsi ad ogni singolo individuo. Si studiano quindi dei modelli per dati stratificati, dove ogni strato rappresenta un individuo. In questo contesto si parla anche di modelli per dati di panel.

Ricorrere alle usuali procedure di stima per studiare i problemi di Neyman e Scott, e in particolar modo alla verosimiglianza profilo, può portare a dei risultati totalmente inaffidabili. Nello specifico, una delle caratteristiche che rende la verosimiglianza profilo non paragonabile ad una verosimiglianza in senso proprio è la distorsione nella funzione *score* profilo. Se si abbina questo alla elevata dimensione del parametro di disturbo, allora la stima di massima verosimiglianza della profilo risulta essere inconsistente e rendere fuorvianti le conclusioni dello studio.

Tra i vari metodi proposti per correggere questo problema è stata introdotta la verosimiglianza profilo modificata ad opera di Barndorff-Nielsen (1980, 1983). Si è visto però come anche la verosimiglianza profilo modificata non sempre fornisca risultati affidabili, specialmente quando il numero di strati  $q$  cresce mentre il numero di repliche per strato  $m$  rimane fissato.

L'attenzione si è allora spostata in un contesto a due indici asintotici, in cui sia  $q$  che  $m$  sono liberi di crescere all'infinito. Uno studio pubblicato da Sartori (2003) ha mostrato che le condizioni di crescita dei due indici per ottenere risultati validi sono pari a  $1/m = o(q)$  per la verosimiglianza profilo e a  $1/m = o(q^{-1/3})$  per la verosimiglianza profilo modificata. Ne consegue che, in situazioni ancora più estreme, anche la verosimiglianza profilo modificata risulta essere non affidabile.

Una strada alternativa per l'inferenza sui parametri di interesse si basa sui metodi di simulazione, grazie al notevole sviluppo della potenza di calcolo nei calcolatori, e in particolare sul bootstrap parametrico. A tal proposito, sono state proposte in anni recenti delle idee che riguardano una modifica nello schema di simulazione e un procedimento chiamato **prepivotting**. Grazie alla semplicità di implementazione del bootstrap, si può pensare di utilizzarlo con queste modifiche anche nel contesto dei problemi di Neyman e Scott.

In questa tesi si vuole indagare sull'efficacia del *prepivotting* nella costruzione di intervalli di confidenza basati su statistiche test che fungono da quantità pivotali rispetto a quanto fatto dalla verosimiglianza profilo e dalla verosimiglianza profilo modificata.

Il Capitolo 1 è volto a richiamare i concetti di base dell'inferenza di verosimiglianza. Sarà dedicato ampio spazio alle caratteristiche dei parametri di disturbo e alle tecniche per fare inferenza in questo contesto, quali la verosimiglianza profilo e la verosimiglianza profilo modificata. Inoltre, si richiameranno anche le basi della metodologia bootstrap.

Nel Capitolo 2 l'attenzione si sposterà sui modelli stratificati. Verranno presentati i risultati già noti in letteratura per risolvere i problemi di Neyman e Scott basati sulle pseudo-verosimiglianze viste nel capitolo precedente e un riassunto sulle conclusioni a cui si è giunti nel contesto dei due indici asintotici. Nella speranza di superare le condizioni citate sopra, verranno introdotti tutti gli strumenti utili al contesto di simulazione bootstrap. Si presenterà in un primo momento una modificazione allo schema per il bootstrap parametrico e successivamente il procedimento di *prepivotting*.

Il Capitolo 3 presenterà degli studi di simulazione sia in modelli per dati continui che per dati discreti. In particolare, si considereranno dei modelli per dati gamma, normali e dei modelli lineari generalizzati per dati binari. Si verificheranno le coperture empiriche nella costruzione degli intervalli di confidenza in diversi tipi di modelli stratificati. Lo studio sarà effettuato sui metodi principali presentati nei capitoli precedenti: sulla verosimiglianza profilo, sulla verosimiglianza profilo modificata e sul *prepivotting* bootstrap.



# Capitolo 1

## L'inferenza di verosimiglianza

### 1.1 Introduzione

Il lavoro di questa tesi segue il paradigma dell'inferenza fisheriana, basato sulla funzione di verosimiglianza. In questo capitolo verranno quindi richiamati i concetti basilari sulla teoria di verosimiglianza, essenziali per comprendere gli argomenti esposti nei capitoli successivi.

Verranno presentati inizialmente i concetti di modello statistico parametrico, di funzione di verosimiglianza e di alcune proprietà ad essa associate. La funzione di verosimiglianza è lo strumento principale per l'analisi inferenziale e, come si vedrà, si può usare in un gran numero di modelli statistici parametrici. Un caso particolarmente interessante riguarda l'inferenza in presenza di parametri di disturbo. Per questo tipo di problema, solo alcune componenti del parametro globale sono d'interesse e quindi si cercherà di fare inferenza solo su di esse tramite una pseudo-verosimiglianza e, in particolare, si descriveranno in dettaglio la verosimiglianza profilo e la verosimiglianza profilo modificata. Per semplificare il problema di inferenza si può ricorrere eventualmente ad una riparametrizzazione del modello e, in tal caso, risulterà utile la proprietà di invarianza a riparametrizzazioni. L'inferenza sul parametro di interesse si può esprimere tramite stime intervallari o verifiche d'ipotesi. Verranno richiamati quindi in sintesi i concetti di regioni di confidenza per ottenere una stima intervallare del parametro e di statistiche test per le verifiche d'ipotesi.

Un'alternativa per l'inferenza, diversa dalle approssimazioni asintotiche di quantità di verosimiglianza, si basa sulla possibilità di utilizzare simulazioni per ottenere dei risultati ugualmente consistenti. Uno dei metodi principali è il bootstrap, di cui si darà un breve accenno sulla sua motivazione e il suo utilizzo.

Il supporto di base per la stesura di questo capitolo è stato dato da Pace e Salvan (1996, Capitoli 1-4), da Pace e Salvan (2001, Capitoli 2-3) e da Young e Smith (2005, Capitolo 11). Si rimanda a tali testi per l'approfondimento degli argomenti qui sintetizzati.

## 1.2 Specificazione del modello

L'inferenza statistica consiste nell'utilizzare i metodi più opportuni, al fine di raggiungere l'obiettivo di sintetizzare i dati  $y = (y_1, \dots, y_n)$  attraverso la ricostruzione della distribuzione teorica  $p^0(y)$  che li ha generati. Per quanto questo obiettivo sia di facile comprensione, la ricerca di una soluzione ai problemi che emergono nell'analisi statistica è tutt'altro che banale.

Fisher, nel suo celebre articolo del 1922, fornisce una suddivisione dei problemi statistici in tre categorie, oltre a presentare la prima nozione di modello statistico (Fisher, 1922). La suddivisione consiste in:

- **problemi di specificazione**, che riguardano la scelta del modello statistico  $\mathcal{F}$  da cui si pensa siano stati generati i dati;
- **problemi di stima**, ovvero il reperimento di una procedura statistica adatta a localizzare  $p^0(y)$  entro  $\mathcal{F}$ ;
- **problemi di distribuzione**, ossia verificare le proprietà della procedura statistica ottenuta.

Nello specifico, il **modello statistico**  $\mathcal{F}$  è una famiglia di distribuzioni di probabilità almeno qualitativamente compatibili con  $y$ , che contiene le forme possibili per il **modello probabilistico**  $p^0(y)$ , la vera distribuzione della variabile casuale di cui  $y$  rappresenta una realizzazione. Se  $p^0(y) \in \mathcal{F}$ , si dice che il modello statistico è **correttamente specificato**.

A seconda del volume dei dati disponibili, che rappresenta la quantità di informazione in possesso per l'analisi, si può dare un'indicazione sull'estensione del modello statistico  $\mathcal{F}$ . In ordine crescente di estensione, ma decrescente di assunzioni, si possono delineare i tre livelli di specificazione seguenti:

- **Specificazione parametrica.** La classe  $\mathcal{F}$  ha elementi indicizzabili attraverso un numero finito  $p$  di parametri reali:

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

dove  $\Theta$  rappresenta lo **spazio parametrico** in cui può variare il parametro  $\theta$ . Se il modello è correttamente specificato si ha  $p^0(y) = p(y; \theta_0)$  per un valore  $\theta_0 \in \Theta$ , dove  $\theta_0$  è il **vero valore del parametro**.

- **Specificazione semiparametrica.** Gli elementi di  $\mathcal{F}$  sono individuati tramite una componente parametrica e una non parametrica,

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

dove  $\theta = (\tau, h(\cdot))$ , con  $\tau \in T \subseteq \mathbb{R}^k$ , mentre  $h(\cdot)$  non è indicizzabile attraverso un insieme di parametri reali.

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

Nel seguito di questa tesi verranno considerati solo modelli statistici parametrici. La scelta di uno di questi per un determinato problema non è permanente, ma è variabile in base alle valutazioni fatte su di esso una volta applicate le procedure statistiche. Se esse non sono soddisfacenti si può riformulare il modello e ripetere l'inferenza. Questi passi possono ripetersi più volte in modo iterativo.

Infine verranno trattati modelli parametrici in cui il parametro  $\theta$  è **identificabile**, ovvero dove c'è corrispondenza biunivoca tra gli elementi di  $\Theta$  e quelli di  $\mathcal{F}$ .

### 1.3 La funzione di verosimiglianza

Una volta scelto il modello statistico parametrico  $\mathcal{F}$ , esso è completamente specificato dalla terna

$$(\mathcal{Y}, p(y; \theta), \Theta), \quad (1.1)$$

dove  $\mathcal{Y}$  è lo spazio campionario,  $p(y; \theta)$  rappresenta il tipo di funzione di densità nel modello e  $\Theta$  è, come già precedentemente detto, lo spazio parametrico, il quale è un sottoinsieme aperto non vuoto di  $\mathbb{R}^p$ .

Fissati i dati  $y = (y_1, \dots, y_n)$ , la funzione di densità si riduce ad una funzione dipendente solo dal parametro  $p$ -dimensionale  $\theta = (\theta_1, \dots, \theta_p)$ . Si definisce quindi la **funzione di verosimiglianza** di  $\theta$  basata sui dati  $y$  come la funzione  $L : \Theta \rightarrow \mathbb{R}$

$$L(\theta) = L(\theta; y) = c(y)p(y; \theta), \quad (1.2)$$

che rappresenta la sintesi informativa su  $\theta$  contenuta nei dati  $y$ . La costante moltiplicativa  $c(y)$  non dipende dal parametro  $\theta$  ed è trascurabile all'atto pratico. Infatti, siccome l'interesse è valutare il rapporto di verosimiglianza  $L(\theta_1)/L(\theta_2)$  per due parametri  $\theta_1$  e  $\theta_2$  diversi tra di loro, è evidente che i fattori che non dipendono dai parametri si possono semplificare. L'espressione  $L(\theta; y)$  serve a mettere in evidenza la dipendenza di  $L(\theta)$  dai dati  $y$ .

Nel caso particolare in cui le osservazioni sono realizzazioni indipendenti e identicamente distribuite (d'ora in poi i.i.d.) della variabile casuale  $Y$ , la funzione di verosimiglianza ha la seguente forma

$$L(\theta) = \prod_{i=1}^n p_Y(y_i; \theta), \quad (1.3)$$

dove  $p_Y(y; \theta)$  è la densità di  $Y$ .

Dal punto di vista pratico è più conveniente usare la versione logaritmica della funzione di verosimiglianza, ossia la **funzione di log-verosimiglianza**

$$l(\theta) = l(\theta; y) = c'(y) + \log p(\theta; y), \quad (1.4)$$

con  $c'(y) = \log c(y)$  e con la convenzione che  $l(\theta) = -\infty$  quando  $L(\theta) = 0$ . Se

le osservazioni sono realizzazioni i.i.d., anche in questo caso è possibile scrivere la funzione di log-verosimiglianza in una forma semplificata:

$$l(\theta) = \sum_{i=1}^n \log p_Y(y_i; \theta). \quad (1.5)$$

Visto che le costanti sono trascurabili ai fini dei calcoli delle quantità di interesse, possiamo definire due verosimiglianze (log-verosimiglianze) equivalenti quando differiscono solo a meno di costanti moltiplicative (additive).

Si assume nel seguito che la log-verosimiglianza sia una funzione regolare di  $\theta$ , ovvero che rispetti le **condizioni di regolarità** (Azzalini, 2001, Paragrafo 3.2.3). La log-verosimiglianza deve essere derivabile almeno tre volte e pure la sua inversa deve essere derivabile almeno tre volte, per  $l(\cdot)$  e le sue derivate deve esistere finito il valore atteso rispetto alla distribuzione nulla, il modello statistico parametrico sottostante deve avere supporto non dipendente da  $\theta$  e dev'essere identificabile. Sotto queste condizioni si possono definire le seguenti **quantità di verosimiglianza**:

- La **funzione score** è il vettore delle derivate parziali prime di  $l(\theta)$ , che verrà indicata con  $l_* = l_*(\theta; y) = (l_{\theta_1}, \dots, l_{\theta_p})$  e dove  $l_{\theta_r} = \frac{\partial}{\partial \theta_r} l(\theta)$ ;
- Verrà indicata con  $j(\theta)$  la **matrice di informazione osservata**, ossia la matrice contenente l'opposto delle derivate parziali seconde di  $l(\theta)$ . In notazione  $j = j(\theta) = [-l_{\theta_r \theta_s}]$ , con  $l_{\theta_r \theta_s} = \frac{\partial^2}{\partial \theta_r \partial \theta_s} l(\theta)$  elemento di posizione  $r, s$ ;
- La **matrice di informazione attesa** verrà indicata con  $i(\theta)$  e rappresenta il valore atteso della matrice di informazione osservata. In notazione  $i = i(\theta) = \mathbb{E}_\theta[j(\theta)]$ . L'elemento di posto  $r, s$  della matrice  $i$  sarà indicato con  $i_{rs}$ .

Grazie alle condizioni di regolarità c'è la possibilità di scambiare le operazioni di derivazione e di integrazione, il che permette di enunciare due proprietà utili relative alle quantità appena definite.

- Il valore atteso della funzione *score* è nullo:  $\mathbb{E}_\theta[l_*(\theta)] = 0, \quad \theta \in \Theta$ ;
- Vale l'**identità dell'informazione**, ovvero

$$\mathbb{E}_\theta[-l_{\theta_r \theta_s}(\theta)] = \mathbb{E}_\theta[l_{\theta_r}(\theta)l_{\theta_s}(\theta)], \quad \theta \in \Theta.$$

Sostanzialmente la matrice di informazione attesa è la matrice di varianze e covarianze della *score* e come tale è definita non negativa.

Un valore di  $\theta$  che massimizza  $L(\theta)$  su  $\Theta$ , ossia per cui  $L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta)$ , è detto **stima di massima verosimiglianza** (s.m.v.) di  $\theta$  ed è indicato con  $\hat{\theta}$ . Si usano le notazioni  $\hat{l} = l(\hat{\theta})$ ,  $\hat{j} = j(\hat{\theta})$  e  $\hat{i} = i(\hat{\theta})$  per indicare quantità di verosimiglianza valutate in  $\hat{\theta}$ . In un modello con funzione di verosimiglianza che rispetta le condizioni di regolarità, è spesso vantaggioso massimizzare  $l(\theta)$  anziché  $L(\theta)$  e cercare la s.m.v. tra le soluzioni dell'**equazione di verosimiglianza**

$$l_*(\theta) = 0. \tag{1.6}$$

Si assume che la stima di massima verosimiglianza sia unica e che sia soluzione della (1.6), la quale individua una **equazione di stima** nel caso si adotti il modello parametrico  $\mathcal{F}$ .

## 1.4 Parametri di disturbo

Il modello statistico  $\mathcal{F}$  contiene una serie di distribuzioni compatibili con i dati osservati  $y$  e, tipicamente, tanto più aumenta la complessità della realtà da rappresentare, tanto più elevata sarà la dimensione  $p$  del parametro  $\theta$ . Questo aumento di complessità fa in modo di cogliere tutti i vari aspetti di distribuzione che rendono il modello coerente con la realtà in esame.

Raramente però c'è interesse a studiare tutti gli aspetti di distribuzione e può essere che alcuni di questi siano considerati solamente per descrivere meglio la variabilità della popolazione. Sarà quindi necessario separare gli aspetti di interesse dagli aspetti di disturbo tramite un'opportuna partizione del parametro  $\theta$ .

Si consideri, eventualmente dopo un'opportuna riparametrizzazione, la partizione  $\theta = (\psi, \zeta)$ , dove  $\psi$  è il **parametro di interesse** di dimensione  $k$  che rappresenta gli aspetti di interesse, mentre  $\zeta$  è il **parametro di disturbo** di dimensione  $p - k$  che rappresenta gli aspetti di disturbo. D'ora in poi si supporrà che le due componenti di  $\theta$  siano a **variazione indipendente**, ovvero che, posto  $\psi \in \Theta_\psi$  e  $\zeta \in \Theta_\zeta$ , si abbia  $\Theta = \Theta_\psi \times \Theta_\zeta$ .

Finora si sono visti i termini generali del parametro di disturbo come elemento descrittivo della variabilità di  $Y$ , ma si possono individuare due casi estremi qualora si supponga, per semplicità, che i dati  $y$  siano derivanti da un campionamento casuale semplice di  $n$  elementi  $(y_1, \dots, y_n)$ . La densità congiunta prenderà allora la forma

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

dove  $p_{Y_i}(\cdot; \theta_i)$  è la densità di  $Y_i$ ,  $i = 1, \dots, n$ .

Il primo caso presenta una situazione tipica, in cui  $\theta_i = (\psi, \zeta)$ ,  $i = 1, \dots, n$ , ovvero con il parametro indipendente dalla dimensione  $n$  del campione. Nel secondo caso, invece, il parametro nell' $i$ -esima osservazione si presenta come  $\theta_i = (\psi, \zeta_i)$ , portando il parametro complessivo ad essere pari a  $\theta = (\psi, \zeta_1, \dots, \zeta_n)$ . Evidentemente la dimensione del parametro complessivo  $\theta$  aumenta con  $n$  ed, in particolare, solo il parametro di disturbo dipende dalla numerosità campionaria. In questo secondo contesto il parametro di interesse  $\psi$  viene definito **parametro strutturale**, mentre i parametri di disturbo  $(\zeta_1, \dots, \zeta_n)$  sono detti **parametri incidentali**.

I problemi con parametri incidentali vengono definiti **problemi di Neyman e Scott** (Neyman e Scott, 1948) e possono presentare problemi per le usuali tecniche inferenziali basate sulla verosimiglianza. Un esempio tipico è dato dalla seguente situazione: siano  $(Y_{i1}, Y_{i2})$ ,  $i = 1, \dots, n$ , variabili casuali i.i.d. con distribuzione  $N(\mu_i, \sigma^2)$ , con parametro strutturale  $\psi = \sigma^2$  e parametro incidentale  $\zeta = (\mu_1, \dots, \mu_n)$ . A causa dei parametri di disturbo, lo stimatore di massima verosimiglianza per  $\sigma^2$ ,  $\hat{\sigma}^2$ , non è consistente. Questo esempio e, più in generale, i problemi di Neyman e Scott verranno ripresi nel Capitolo 2.

## 1.5 Invarianza a riparametizzazioni

In un modello parametrico identificabile, il parametro  $p$ -dimensionale  $\theta$  e gli elementi  $p(y; \theta)$  del modello  $\mathcal{F}$  sono in relazione biunivoca tra di loro. Scegliere una parametrizzazione equivale a scegliere un determinato sistema di coordinate geometrico rispetto alle quali si possa sempre individuare univocamente gli elementi di  $\mathcal{F}$ .

Esiste anche la possibilità di riparametrizzare il modello: cambiando il sistema di coordinate non necessariamente è automatico che cambi anche il modello di riferimento. La scelta di riparametrizzare è dettata specialmente per avere maggiore convenienza nel manipolare gli elementi di  $\mathcal{F}$ , ma è desiderabile che la scelta di una determinata riparametrizzazione non cambi le conclusioni inferenziali delle analisi su quel modello.

Una qualunque trasformazione  $\omega = \omega(\theta)$ , con  $\omega(\cdot)$  funzione biunivoca e regolare (infinitamente derivabile con inversa infinitamente derivabile) da  $\Theta \subseteq \mathbb{R}^p$  a  $\Omega \subseteq \mathbb{R}^p$  è una riparametrizzazione del modello. Si parla di **invarianza rispetto alla parametrizzazione** quando le conclusioni inferenziali ottenute nella parametrizzazione  $\omega$  sono le stesse a cui si è giunti attraverso la parametrizzazione  $\theta$  e che si possono esprimere in funzione di  $\theta = \theta(\omega)$ . Lo stimatore di massima verosimiglianza è l'esempio più evidente di quantità che rispetta questo principio, nella fattispecie chiamata proprietà di equivarianza.

È utile che anche in situazioni in cui sono presenti parametri di disturbo si possa riparametrizzare il modello senza stravolgere le conclusioni inferenziali. Generalmente una riparametrizzazione globale non mantiene la distinzione tra parametro di interesse e parametro di disturbo, per cui si consiglia di restringere il campo delle riparametrizzazioni possibili alle sole **riparametrizzazioni che non alterano l'interesse**. Con questo termine si intendono riparametrizzazioni  $\omega = \omega(\theta) = \omega(\psi, \zeta)$ , con  $\omega = (\rho, \xi)$ , tali che

$$\rho = \rho(\psi) \quad \text{e} \quad \xi = \xi(\psi, \zeta),$$

con  $\rho(\cdot)$  biunivoca. In corrispondenza con il caso di interesse globale, si considererà in questa situazione il **principio di invarianza rispetto alle parametrizzazioni che non alterano l'interesse**.

## 1.6 Pseudo-verosimiglianze

Nel caso in cui il modello preveda un parametro di interesse  $\psi \in \Theta_\psi \subseteq \mathbb{R}^k$  e un parametro di disturbo  $\zeta \in \Theta_\zeta \subseteq \mathbb{R}^{p-k}$ , appare utile, o quantomeno interessante, basare l'inferenza solo su  $\psi$ , specie se la struttura di  $\zeta$  è particolarmente complessa. Solitamente si cerca di ridurre il problema ad una funzione di verosimiglianza

dipendente solo da  $\psi$ , cercando di evitare o al più limitare perdite di informazioni sul parametro di interesse.

Si definisce **pseudo-verosimiglianza** una funzione dipendente, oltre che dai dati, solo dal parametro di interesse e che abbia una o più proprietà che la rendano simile ad una verosimiglianza in senso proprio. Sono aspetti che riguardano le condizioni di regolarità della funzione di verosimiglianza, come funzione *score* con valore atteso pari a zero, s.m.v. con distribuzione asintotica normale, eccetera.

Esistono situazioni in cui si ottiene una pseudo-verosimiglianza partendo da un modello statistico definito come riduzione del modello  $\mathcal{F}$  originale. Se le funzioni di densità appartenenti a questo modello indotto dipendono solo dal parametro di interesse  $\psi$ , allora la pseudo-verosimiglianza ottenuta è una verosimiglianza in senso proprio, dove l'assunzione fondamentale affinché l'inferenza su  $\psi$  rimanga valida è che nel modello ridotto l'informazione rimanga di ordine  $O(n)$ . Due esempi in cui la riduzione risulta essere così efficace sono dati dalla **verosimiglianza marginale** e dalla **verosimiglianza condizionata**.

D'altro canto esistono anche situazioni in cui non è possibile dedurre direttamente pseudo-verosimiglianze proprie. In questi casi si possono considerare pseudo-verosimiglianze più generali, le cui proprietà andranno indagate caso per caso. L'esempio più rilevante è quello della **verosimiglianza profilo** e delle sue modificazioni.

### 1.6.1 Verosimiglianza marginale e condizionata

Per indagare sulle possibili riduzioni del modello  $\mathcal{F}$ , è utile considerare la funzione di densità espressa nella seguente fattorizzazione:

$$p_Y(y; \psi, \zeta) = p_T(t; \psi, \zeta) p_{S|T=t}(s; t, \psi, \zeta) p_{Y|S=s, T=t}(y; s, t, \psi, \zeta), \quad (1.7)$$

dove sono state introdotte le statistiche  $t(y)$  e  $s(y)$ . Il primo passo consiste nel vedere se almeno un fattore della (1.7) dipende solo da  $\psi$ , in modo da capire se la pseudo-verosimiglianza cercata sia una verosimiglianza in senso proprio.

La fattorizzazione (1.7) possiede, come casi particolari, sia la riduzione per sufficienza che la riduzione per condizionamento a statistiche ancillari, se si pensa ad  $y$  come al risultato di una precedente riduzione per sufficienza.

Se si considera la statistica  $(t, s)$  sufficiente per tutto il parametro  $\theta$ , allora l'ultimo fattore della (1.7) diventa trascurabile in termini di verosimiglianza e ci si può limitare ai fattori

$$p_{T,S}(t, s; \psi, \zeta) = p_T(t; \psi, \zeta)p_{S|T=t}(s; t, \psi, \zeta), \quad (1.8)$$

da cui emergono dei casi particolarmente interessanti.

Il primo esempio dato dalla (1.8) riguarda un caso che si realizza raramente nelle applicazioni: se il parametro di interesse e il parametro di disturbo vengono divisi tra i fattori del secondo membro, ossia

$$p_{T,S}(t, s; \psi, \zeta) = p_T(t; \psi)p_{S|T=t}(s; t, \zeta), \quad (1.9)$$

la verosimiglianza che ne risulta viene detta **a parametri separabili** e si può scrivere come

$$L(\theta) = L_1(\psi)L_2(\zeta).$$

In questo caso l'informazione su  $\psi$  è catturata totalmente dal fattore di verosimiglianza  $L_1(\psi)$  e l'inferenza può procedere separatamente da quella su  $\zeta$ , come se fosse noto.

Ripartendo ancora dalla (1.8), si ottiene un secondo esempio rilevante. Supponendo si sia ottenuta la fattorizzazione

$$p_{T,S}(t, s; \psi, \zeta) = p_T(t; \psi)p_{S|T=t}(s; t, \psi, \zeta), \quad (1.10)$$

e supponendo che il fattore di verosimiglianza corrispondente a  $p_{S|T=t}(\cdot)$  sia trascurabile in termini di informazione, si può fare inferenza su  $\psi$  esclusivamente marginalizzando rispetto a  $T$ , cioè sul modello  $p_T(t; \psi)$ . Ciò porta alla funzione di verosimiglianza

$$L_M(\psi) = L_M(\psi; t) = p_T(t; \psi), \quad (1.11)$$

definita come **verosimiglianza marginale** basata su  $t$ .

Un'ulteriore possibilità è data dalla situazione duale della precedente, in cui il secondo fattore del membro di destra della (1.8) dipende solo dal parametro di

interesse, ovvero

$$p_{T,S}(t, s; \psi, \zeta) = p_T(t; \psi, \zeta) p_{S|T=t}(s; t, \psi). \quad (1.12)$$

Se  $p_T(t; \psi, \zeta)$  è trascurabile in termini di perdita di informazione, si può basare l'inferenza su  $\psi$  solo attraverso il modello ottenuto per condizionamento a  $T$ . La funzione di verosimiglianza corrispondente

$$L_C(\psi) = L_C(\psi; s) = p_{S|T=t}(s; t, \psi) \quad (1.13)$$

è detta **verosimiglianza condizionata** a  $T = t$ .

I fattori trascurabili nella (1.10) e nella (1.12) sono tali in quanto si suppone presentino **assenza di informazione sul parametro di interesse utilmente estraibile**, ovvero che la statistica  $T$  esaurisca il suo compito di estrarre informazione dal parametro di interesse  $\psi$  nella marginalizzazione (o condizionamento) del modello da ridurre. Per approfondire questo argomento si rimanda a Jørgensen (1993) e Pace e Salvan (1996, Capitolo 4).

I casi in cui è possibile ottenere funzioni di verosimiglianza come (1.11) e (1.13) sono essenzialmente limitati alle famiglie di gruppo e alle famiglie esponenziali, rispettivamente. Quando tali fattorizzazioni non sono ottenibili, si cercano altri tipi di funzioni di pseudo-verosimiglianza utili agli scopi inferenziali, sebbene non possano essere considerate strettamente verosimiglianze in senso proprio.

## 1.6.2 Verosimiglianza profilo

Un metodo ampiamente applicabile per la costruzione di una pseudo-verosimiglianza per il parametro di interesse  $\psi$  prevede la sostituzione del parametro di disturbo  $\zeta$  con una sua stima consistente. Nello specifico, sostituendo a  $\zeta$  la s.m.v. del parametro di disturbo per un fissato valore di  $\psi$ ,  $\hat{\zeta}_\psi$ , otteniamo la funzione di **verosimiglianza profilo**

$$L_P(\psi) = L(\psi, \hat{\zeta}_\psi). \quad (1.14)$$

Inoltre si indicherà con  $l_P(\psi) = \log L_P(\psi)$  la funzione di **log-verosimiglianza profilo**. Nei casi in cui vengono soddisfatte le condizioni di regolarità,  $\hat{\zeta}_\psi$  è soluzione in  $\zeta$  dell'equazione  $l_\zeta(\psi, \zeta) = 0$ .

La verosimiglianza profilo non può essere considerata una verosimiglianza in senso proprio, poiché non è stata dedotta direttamente da alcuna funzione di densità. Tuttavia la verosimiglianza profilo gode di alcune interessanti proprietà che la rendono simile ad una verosimiglianza propria:

- La stima di massima verosimiglianza profilo coincide con la s.m.v. di  $\psi$  basata su  $L(\psi, \zeta)$ ,

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

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

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

$$j_P(\psi) = -\frac{\partial^2}{\partial\psi\partial\psi^\top} l_P(\psi) = -\frac{\partial^2}{\partial\psi\partial\psi^\top} l(\psi, \hat{\zeta}_\psi).$$

L'inversa di  $j_P(\psi)$  è uguale al blocco  $(\psi, \psi)$  dell'inversa della matrice di informazione osservata complessiva. Infatti si osserva che

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

il secondo addendo di questa formula è uguale a zero poiché  $l_\zeta(\psi, \hat{\zeta}_\psi) = 0$ . Risulta pertanto

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

La matrice delle derivate seconde è data da

$$\frac{\partial^2}{\partial\psi\partial\psi^\top} l_P(\psi) = l_{\psi\psi}(\psi, \hat{\zeta}_\psi) + l_{\psi\zeta}(\psi, \hat{\zeta}_\psi) \frac{\partial}{\partial\psi} \hat{\zeta}_\psi.$$

Dato che  $l_\zeta(\psi, \hat{\zeta}_\psi) = 0$ , anche la sua derivata calcolata rispetto a  $\psi$  sarà pari a 0. Si ottiene così l'uguaglianza

$$l_{\zeta\psi}(\psi, \hat{\zeta}_\psi) + l_{\zeta\zeta}(\psi, \hat{\zeta}_\psi) \frac{\partial}{\partial\psi} \hat{\zeta}_\psi = 0,$$

da cui

$$\frac{\partial}{\partial\psi} \hat{\zeta}_\psi = -(l_{\zeta\zeta}(\psi, \hat{\zeta}_\psi))^{-1} l_{\zeta\psi}(\psi, \hat{\zeta}_\psi);$$

quindi segue che

$$j_P(\psi) = -(l_{\psi\psi} - l_{\psi\zeta}(l_{\zeta\zeta})^{-1}l_{\zeta\psi}),$$

dove tutte le derivate si intendono calcolate in  $(\psi, \hat{\zeta}_\psi)$ . Si ottiene quindi

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

dove  $j^{\psi\psi}(\psi, \hat{\zeta}_\psi)$  rappresenta il blocco  $(\psi, \psi)$  dell'inversa di  $j(\psi, \hat{\zeta}_\psi)$ ;

- il test del log-rapporto di verosimiglianza basato sulla verosimiglianza profilo è

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

ovvero coincide con l'usuale log-rapporto di verosimiglianza basato su  $L(\psi, \zeta)$  considerato per la verifica d'ipotesi per  $\psi$ , con  $\zeta$  ignoto. Sotto condizioni di regolarità  $W_P(\psi)$  ha distribuzione asintotica nulla pari a  $\chi_p^2$ . Per ulteriori dettagli sulle statistiche test di verosimiglianza si veda il paragrafo 1.7.

Grazie a queste proprietà la verosimiglianza profilo risulta essere uno strumento inferenziale notevolmente interessante, data la sua generalità. Ciò nonostante la verosimiglianza profilo non è assimilabile ad una verosimiglianza in senso proprio. In particolare, la funzione **score profilo**  $\frac{\partial}{\partial\psi}l_P(\psi)$  non ha media pari a zero; in generale, la *score* profilo ha valore atteso nullo di ordine  $O(1)$  nei casi in cui le condizioni di regolarità vengono rispettate, ma si raggiungono ordini più elevati in presenza di una struttura non regolare con parametri incidentali, fino ad un ordine  $O(n)$ . Si possono raggiungere risultati altamente insoddisfacenti poiché lavorare con la verosimiglianza profilo equivale a considerare  $\zeta$  noto e pari a  $\hat{\zeta}_\psi$ , trascurando così la variabilità della stima di  $\zeta$ . Questo problema è particolarmente severo quando  $\zeta$  è di grandi dimensioni e i dati non riescono ad avere informazioni sufficienti su di esso.

Per compensare questa mancanza di informazione, sono stati proposti alcuni metodi di modificazione alla verosimiglianza profilo.

### 1.6.3 Verosimiglianza profilo modificata

Esistono diverse possibilità di modificazioni alla verosimiglianza profilo. In questa tesi ci si concentrerà in modo particolare sulle due modificazioni principali: la **verosimiglianza condizionata approssimata** e la **verosimiglianza profilo modificata**.

Senza entrare troppo nel dettaglio, la **verosimiglianza condizionata approssimata** (Cox e Reid, 1987) punta a ricondursi ad una situazione simile alla verosimiglianza a parametri separabili. Si cerca quindi di sfruttare una parametrizzazione ortogonale, dove i parametri  $\psi$  e  $\zeta$  sono detti ortogonali quando i vettori *score*  $l_\psi$  e  $l_\zeta$  sono incorrelati. La log-verosimiglianza che ne risulta è a tutti gli effetti una versione penalizzata della log-verosimiglianza profilo:

$$l_{CA}(\psi) = l_P(\psi) - \frac{1}{2} \log |j_{\zeta\zeta}(\psi, \hat{\zeta}_\psi)|.$$

Tuttavia l'utilizzo della log-verosimiglianza condizionata approssimata presenta degli inconvenienti non indifferenti, come ad esempio l'esplicitazione di una parametrizzazione ortogonale e la limitazione al caso in cui  $\psi$  è unidimensionale. Inoltre la log-verosimiglianza approssimata non soddisfa il principio di invarianza rispetto a parametrizzazioni che non alterano l'interesse.

Un'altra importante proposta di modificazione è data dalla **verosimiglianza profilo modificata**, introdotta da Barndorff-Nielsen (1980, 1983). La verosimiglianza profilo modificata è definita come

$$L_M(\psi) = L_P(\psi)M(\psi), \quad (1.15)$$

dove il fattore di aggiustamento  $M(\psi)$  è pari a

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

dove nel primo determinante del fattore di modificazione compare la derivata mista

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

in cui è richiesta l'esplicitazione di una statistica ancillare  $a$ . La motivazione della presenza di questa statistica è data dal fatto che quando la s.m.v. di  $\theta$  è unica, allora è necessariamente funzione della statistica sufficiente minimale  $s$ , ma non è altrettanto certo che  $\hat{\theta}$  mantenga la sufficienza. Quindi una statistica  $a$  è detta **ancillare** rispetto a  $s$ , sufficiente minimale, se la relazione tra  $(\hat{\theta}, a)$  e  $s$  è biunivoca e se la distribuzione di  $a$  non dipende da  $\theta$ , cioè è costante in distribuzione. La statistica  $a$  esprime l'informazione mancante per ricostruire  $s$  da  $\hat{\theta}$ . Di conseguenza, la log-verosimiglianza viene scritta in funzione della statistica sufficiente minimale  $(\hat{\psi}, \hat{\zeta}, a)$ . Quantità come  $l_{\zeta;\hat{\zeta}}$ , che coinvolgono derivate rispetto alle componenti dello stimatore di massima verosimiglianza  $\hat{\theta}$ , sono dette **derivate nello spazio campionario**. Tale derivata è ovviamente in generale differente dalla derivata seconda

$$l_{\zeta\zeta}(\psi, \zeta) = \frac{\partial^2}{\partial \zeta \partial \zeta^\top} l(\psi, \zeta).$$

Partendo dalla verosimiglianza profilo modificata si può ricavare anche la **log-verosimiglianza profilo modificata**, ovvero

$$\begin{aligned} l_M(\psi) &= \log L_M(\psi) = l_P(\psi) + \log M(\psi) \\ &= l_P(\psi) - \log |l_{\zeta;\hat{\zeta}}(\psi, \hat{\zeta}_\psi; \hat{\psi}, \hat{\zeta}, a)| + \frac{1}{2} |j_{\zeta\zeta}(\psi, \hat{\zeta}_\psi; \hat{\psi}, \hat{\zeta}, a)|. \end{aligned} \quad (1.16)$$

Tra le proprietà della verosimiglianza profilo modificata, la più importante riguarda l'invarianza rispetto a riparametrizzazioni che non alterano l'interesse. Inoltre il fattore di modificazione  $M(\psi)$  è di ordine  $O_p(1)$ , rendendo di fatto  $L_P(\psi)$  e  $L_M(\psi)$  asintoticamente equivalenti al *primo ordine*, ma con una riduzione della distorsione della funzione *score* da  $O(1)$  a  $O(n^{-1})$ . Infine lo stimatore basato sulla massimizzazione di  $l_M(\psi)$  verrà indicato con  $\hat{\psi}_M$ .

La (1.15) può essere riscritta come

$$L_M(\psi) = L_P(\psi) D(\psi) |j_{\zeta\zeta}(\psi, \hat{\zeta}_\psi; \hat{\psi}, \hat{\zeta}, a)|^{-1/2}, \quad (1.17)$$

con

$$D(\psi) = \frac{|j_{\zeta\zeta}(\psi, \hat{\zeta}_\psi; \hat{\psi}, \hat{\zeta}, a)|}{|l_{\zeta; \hat{\zeta}}(\psi, \hat{\zeta}_\psi; \hat{\psi}, \hat{\zeta}, a)|}.$$

Derivando l'equazione di verosimiglianza  $l_\zeta(\psi, \zeta; \hat{\psi}, \hat{\zeta}, a) = 0$  rispetto a  $\hat{\zeta}$  è immediatamente evidente che

$$D(\psi) = \left| \frac{\partial \hat{\zeta}_\psi}{\partial \hat{\zeta}} \right|.$$

Data la (1.16), bastano dei semplici passaggi per arrivare alla scrittura alternativa della log-verosimiglianza profilo modificata:

$$l_M(\psi) = l_P(\psi) + \log D(\psi) - \frac{1}{2} \log |j_{\zeta\zeta}(\psi, \hat{\zeta}_\psi; \hat{\psi}, \hat{\zeta}, a)|. \quad (1.18)$$

Per il calcolo della verosimiglianza o log-verosimiglianza profilo modificata è necessario avere a disposizione la s.m.v. globale  $(\hat{\psi}, \hat{\zeta})$  e vincolata  $(\psi, \hat{\zeta}_\psi)$ , il blocco  $(\zeta, \zeta)$  della matrice di informazione osservata  $j$  e la derivata nello spazio campionario di  $l$  rispetto a  $\zeta$  e  $\hat{\zeta}$ . Quando esiste la verosimiglianza marginale o la verosimiglianza condizionata per  $\psi$ , la verosimiglianza profilo modificata è facile da calcolare ed è una buona approssimazione di esse. In generale, anche quando non esiste la verosimiglianza marginale o la verosimiglianza condizionata per  $\psi$ , è facile ottenere la verosimiglianza profilo modificata fintanto che il modello sottostante appartiene ad una famiglia esponenziale o ad una famiglia di gruppo.

Il calcolo della verosimiglianza profilo modificata diventa svantaggioso quando il modello sottostante non risponde a nessuno dei canoni descritti poco sopra. Esistono comunque diverse approssimazioni che mantengono le stesse proprietà asintotiche. Si veda Severini (2000, Paragrafo 9.3).

Tipicamente, esplicitare la (1.17) o la (1.18) richiede di specificare la statistica ancillare  $a$ . Quando però  $\hat{\zeta}_\psi = \hat{\zeta}$ , la matrice  $[\partial \hat{\zeta}_\psi / \partial \hat{\zeta}]$  corrisponde alla matrice identità e si ottiene

$$l_M(\psi) = l_P(\psi) - \frac{1}{2} \log |j_{\zeta\zeta}(\psi, \hat{\zeta}_\psi)|,$$

cioè

$$l_M(\psi) = l_{CA}(\psi).$$

La verosimiglianza profilo modificata equivale approssimativamente alla verosimiglianza condizionata approssimata anche quando i parametri  $\psi$  e  $\zeta$  sono ortogonali, tenendo comunque conto dei problemi nell'uso di  $L_{CA}(\psi)$ .

Un'altra proposta di approssimazione è stata proposta da Severini (1998), in cui si usano le derivate nello spazio campionario. Nel dettaglio si usa

$$M(\psi) = |I_{\zeta\zeta}(\hat{\psi}, \hat{\zeta}; \psi, \hat{\zeta}_\psi)|^{-1} |j_{\zeta\zeta}(\psi, \hat{\zeta}_\psi; \hat{\psi}, \hat{\zeta}, a)|^{1/2}, \quad (1.19)$$

dove

$$I_{\zeta\zeta}(\psi_0, \zeta_0; \psi_1, \zeta_1) = \mathbb{E}_{\psi_0, \zeta_0} [l_\zeta(\psi_0, \zeta_0) l_\zeta(\psi_1, \zeta_1)]. \quad (1.20)$$

I calcoli analitici della (1.20) sono abbastanza semplici per alcuni dei problemi che verranno esaminati in seguito.

Se le osservazioni sono tra loro indipendenti, una versione asintoticamente equivalente di questa correzione si ottiene sostituendo  $I_{\zeta\zeta}(\hat{\psi}, \hat{\zeta}; \psi, \hat{\zeta}_\psi)$  con la stima

$$\hat{I}_{\zeta\zeta}(\hat{\psi}, \hat{\zeta}; \psi, \hat{\zeta}_\psi) = \sum_{i=1}^n \frac{\partial}{\partial \zeta} \log f(y_i; \hat{\psi}, \hat{\zeta}) \frac{\partial}{\partial \zeta} \log f(y_i; \hat{\psi}, \hat{\zeta}_\psi);$$

vedi Severini (2000, Paragrafo 9.5.5). Una soluzione simile a questa è stata sviluppata da Arellano e Hahn (2007), partendo dal lavoro di DiCiccio e Stern (1993) e DiCiccio *et al.* (1996). Il fattore di modificazione prende la forma

$$M(\psi) = |j_{\zeta\zeta}(\psi, \hat{\zeta}_\psi; \hat{\psi}, \hat{\zeta}, a)|^{1/2} \left( \sum_{i=1}^n \left\{ \frac{\partial}{\partial \zeta} \log f(y_i; \psi, \hat{\zeta}_\psi) \right\}^2 \right)^{-1/2}.$$

## 1.7 Inferenza basata sulla verosimiglianza

Un metodo per ricostruire  $p^0(y; \theta)$  è sostituire la stima di massima verosimiglianza  $\hat{\theta}$  nella funzione di densità del modello  $\mathcal{F}$ . In questo senso la teoria asintotica è confortante: per  $n \rightarrow +\infty$ ,  $\hat{\theta}$  è consistente, ovvero  $\hat{\theta}$  converge in probabilità a  $\theta$ ; in notazione  $\hat{\theta} \xrightarrow{p} \theta$ , con  $\theta$  inteso come vero valore del parametro per non appesantire la notazione. Inoltre la distribuzione di  $\hat{\theta}$  per  $n$  divergente si può approssimare come segue:

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

dove  $p$  indica la dimensione del parametro. In tale espressione,  $i$  può essere sostituita dalla sua stima  $\hat{i}$ , da  $j$  o ancora da  $\hat{j}$ .

Tuttavia fino ad ora si sono trattati i problemi di stima solo nell'ottica della stima puntuale. Altre procedure per individuare  $p^0(y; \theta)$  consistono nella **verifica d'ipotesi** e nella **stima intervallare**. La verifica d'ipotesi fa riferimento alla situazione in cui si desidera che un **test statistico** conduca a una tra due decisioni possibili: il parametro  $\theta$  appartiene al sottoinsieme  $\Theta_0$  di  $\Theta$  oppure  $\theta$  appartiene al complemento  $\Theta \setminus \Theta_0$ . La stima intervallare consiste nel determinare una **regione di confidenza**, la quale è un sottoinsieme di  $\Theta$  individuato in funzione dei dati osservati  $y$  e indicato con  $C(y)$ , tale che

$$\mathbb{P}_\theta[\theta \in C(y)] \geq 1 - \alpha, \quad \text{per ogni } \theta \in \Theta,$$

dove  $1 - \alpha$  indica il livello di confidenza nominale della regione. Nel caso in cui  $\theta$  sia scalare si parla di **intervalli di confidenza**. Per costruire gli intervalli di confidenza sono di grande utilità le quantità pivotali. Una **quantità pivotale**  $q(y; \theta)$  è una funzione dei dati  $y$  e del parametro  $\theta$ , la cui distribuzione è nota e indipendente dal parametro stesso. Fissando un livello di confidenza e risolvendo rispetto a  $\theta$  l'equazione che collega  $q(y; \theta)$  ai percentili opportuni della distribuzione nulla di  $q$  si ottiene l'intervallo di confidenza per il parametro, unilaterale o bilaterale. La quantità pivotale  $q(y; \theta)$  è alla base anche della costruzione di statistiche test per la verifica d'ipotesi. La verosimiglianza può essere usata per la costruzione di quantità pivotali approssimate. Ad esempio, l'ipotesi nulla del tipo  $H_0 : \theta = \theta_0$ , verrà confrontata la verosimiglianza per il valore  $\theta_0$ ,  $L(\theta_0)$ , con  $L(\hat{\theta})$ , valore massimo della verosimiglianza in tutto  $\Theta$ , costruendo così la statistica test **rapporto di verosimiglianza**

$$\frac{L(\hat{\theta})}{L(\theta_0)} \tag{1.21}$$

con regione di rifiuto unilaterale destra. Si può anche considerare la stessa statistica espressa in termini di log-verosimiglianza  $l(\hat{\theta}) - l(\theta_0)$ .

Per un sistema di ipotesi semplice  $H_0 : \theta = \theta_0$  contro l'alternativa bilaterale  $H_1 : \theta \neq \theta_0$ , si può usare una trasformazione monotona della (1.21), la statistica

**log-rapporto di verosimiglianza**, di forma

$$W(\theta_0) = 2\{l(\hat{\theta}) - l(\theta_0)\} = -2\{l(\theta_0) - l(\hat{\theta})\}. \quad (1.22)$$

Il test  $W(\theta_0)$  si distribuisce asintoticamente come un chi-quadrato con  $p$  gradi di libertà ed è quindi approssimativamente pivotale. Esistono anche versioni asintoticamente equivalenti di questo test, come lo **score test** o **test di Rao**

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

e il **test di Wald**

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

Quando  $p = 1$ , si possono considerare le versioni direzionali dei test appena presentati, valide per verifiche d'ipotesi con alternative unilaterali  $H_0 : \theta = \theta_0$  contro  $H_1 : \theta < \theta_0$  o  $H_1 : \theta > \theta_0$ . Le versioni direzionali hanno quindi forma:

$$\begin{aligned} r(\theta_0) &= \text{sgn}(\hat{\theta} - \theta_0) \sqrt{W(\theta_0)}, \\ r_u(\theta_0) &= l_*(\theta_0) \sqrt{i(\theta_0)}, \\ r_e(\theta_0) &= (\hat{\theta} - \theta_0) \sqrt{i(\theta_0)}, \end{aligned} \quad (1.23)$$

e tutte e tre hanno distribuzione asintotica nulla  $N(0,1)$ .

In caso di presenza di parametri di disturbo, grazie alle proprietà della verosimiglianza profilo, possiamo costruire dei test equivalenti in via asintotica a quelli costruiti con il parametro globale. Per l'ipotesi composita  $H_0 : \psi = \psi_0$ , il log-rapporto di verosimiglianza è

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

che coincide il **log-rapporto di verosimiglianza profilo**

$$W_P(\psi_0) = 2\{l_P(\hat{\psi}) - l_P(\psi_0)\}, \quad (1.24)$$

la cui distribuzione nulla asintotica è un chi-quadrato con  $k$  gradi di libertà, dove  $k$  è la dimensione del parametro di interesse  $\psi$ .

Se il parametro di interesse  $\psi$  è scalare, si può considerare anche la radice con

segno di  $W_P(\psi_0)$ ,

$$r_P(\psi_0) = \text{sgn}(\hat{\psi} - \psi_0) \sqrt{W_P(\psi_0)}, \quad (1.25)$$

ancora di distribuzione nulla asintotica  $N(0,1)$ .

Analogamente, si possono definire le versioni *score* e di Wald di  $W_P(\psi_0)$  e di  $r_P(\psi_0)$ , ovvero

$$\begin{aligned} W_{Pu}(\psi_0) &= l_\psi(\psi_0, \hat{\zeta}_{\psi_0})^\top i^{\psi\psi}(\psi_0, \hat{\zeta}_{\psi_0}) l_\psi(\psi_0, \hat{\zeta}_{\psi_0}), \\ r_{Pu}(\psi_0) &= l_\psi(\psi_0, \hat{\zeta}_{\psi_0}) \sqrt{i^{\psi\psi}(\psi_0, \hat{\zeta}_{\psi_0})} \end{aligned}$$

per quanto riguarda lo *score* test, e

$$\begin{aligned} W_{Pe}(\psi_0) &= (\hat{\psi} - \psi_0)^\top [i^{\psi\psi}(\psi_0, \hat{\zeta}_{\psi_0})]^{-1} (\hat{\psi} - \psi_0), \\ r_{Pe}(\psi_0) &= (\hat{\psi} - \psi_0) [i^{\psi\psi}(\psi_0, \hat{\zeta}_{\psi_0})]^{-1/2} \end{aligned}$$

per quanto riguarda il test di Wald. In entrambi i casi  $i^{\psi\psi}$  fa riferimento al blocco relativo a  $\psi$  di  $i^{-1}$ .

Nonostante il test di Wald  $W_e(\theta_0)$  sia spesso utilizzato per la sua semplicità di calcolo e il test di Rao  $W_u(\theta_0)$  abbia il vantaggio di non richiedere esplicitamente il valore di  $\hat{\theta}$ , vi sono ragioni teoriche per preferire l'uso del log-rapporto di verosimiglianza  $W(\theta_0)$ . In particolare il log-rapporto di verosimiglianza e i test *score* sono invarianti, mentre ciò non è vero per il test di Wald. Inoltre il test *score* è spesso instabile numericamente, rendendo  $W(\theta_0)$  la scelta preferibile. La stessa analisi vale anche per le versioni profilo e direzionali dei test: è raccomandato l'uso della (1.23), della (1.24) e della (1.25).

Una volta definita una quantità pivotale  $t(\cdot)$  e applicata ai dati per la verifica d'ipotesi, uno strumento per esprimere il contributo informativo del test è il **livello di significatività osservato** (*p-value*)

$$\alpha^{oss} = \sup_{\theta=\theta_0} \mathbb{P}_\theta[t(Y) \geq t(y)], \quad (1.26)$$

dove è importante notare la differenza tra la variabile aleatoria  $Y$  e la realizzazione osservata  $y$ . Un certo valore del *p-value* permette di accettare o rifiutare l'ipotesi nulla, confrontandolo con un livello  $\alpha$  fissato.

Anche per quanto riguarda la stima intervallare la verosimiglianza risulta essere uno strumento utile. Una regione di confidenza costituita dai valori di  $\theta$  che hanno elevata verosimiglianza può essere presentata come

$$\mathcal{I} \equiv \mathcal{I}(y) = \{\theta \in \Theta : L(\theta) \geq c L(\hat{\theta})\},$$

con  $c \in (0,1)$ , in base al livello di confidenza desiderato.

Si possono anche sfruttare le quantità approssimativamente pivotali viste precedentemente per costruire delle regioni di confidenza. Visto che la distribuzione asintotica di  $W(\theta_0)$  è  $\chi_p^2$ , e che quindi non dipende da  $\theta$ , si può usare  $W$  come quantità pivotale approssimata per calcolare delle regioni di confidenza per  $\theta$  con livello nominale  $1 - \alpha$ . Tali regioni avranno forma

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

La regione può essere scritta anche come

$$\mathcal{I}(y) = \{\theta \in \Theta : l(\theta) \geq l(\hat{\theta}) - \frac{1}{2} \chi_{p; 1-\alpha}^2\},$$

dove  $\chi_{p; 1-\alpha}^2$  è il quantile  $1 - \alpha$  della distribuzione  $\chi_p^2$ . Ciò vale anche nel caso di parametri di disturbo. Infatti, il test basato su  $W_P(\psi_0)$  è una quantità pivotale per  $\psi$  e ciò permette di costruire una regione di confidenza per  $\psi$  con livello nominale  $1 - \alpha$ , di forma

$$\mathcal{I}(y) = \{\psi \in \Theta_\psi : W_P(\psi) \leq \chi_{k; 1-\alpha}^2\},$$

dove  $\chi_{k; 1-\alpha}^2$  è il quantile  $1 - \alpha$  della distribuzione chi-quadrato. Nel caso  $\psi$  sia scalare questa regione è equivalente a

$$\mathcal{I}(y) = \{\psi \in \Theta_\psi : |r_P(\psi)| \leq z_{1-\frac{\alpha}{2}}\},$$

dove  $z_{1-\frac{\alpha}{2}}$  è il quantile di livello  $1 - \alpha/2$  della  $N(0,1)$ .

In certe condizioni (Pace e Salvan, 2001, Paragrafo 2.5.3) il livello di significatività osservato è esso stesso una quantità pivotale, rendendo evidente una dualità tra *p-value* e regioni di confidenza. La variabile casuale  $\alpha_\theta(Y)$ , ossia la 1.26 in cui  $\theta_0 = \theta$  per un  $\theta$  generico, sotto  $\theta$  ha legge di probabilità uniforme  $\mathcal{U}(0,1)$ . Ciò significa che

l'insieme dei valori  $\tilde{\theta}$  che non sono stati rifiutati per una classe di test  $t_{\tilde{\theta}}$  con livello  $\alpha$  per le ipotesi nulle  $H_0 : \theta = \tilde{\theta}$ , ovvero

$$\mathcal{I}(y) = \{\theta \in \Theta : \alpha_{\theta}(Y) \geq \alpha\}$$

definiscono una regione di confidenza di livello  $1 - \alpha$ . Infatti

$$\mathbb{P}_{\theta}[\theta \in \mathcal{I}(Y)] = \mathbb{P}_{\theta}[\alpha_{\theta}(Y) \geq \alpha] = 1 - \alpha$$

per ogni  $\theta \in \Theta$ . L'utilità di questa dualità risiede nel fatto che quindi è possibile costruire una regione di confidenza a partire dal *p-value* di un test e, viceversa, ottenere la regione di accettazione di un test trasformando una regione di confidenza con livello di confidenza  $1 - \alpha$ .

## 1.8 Bootstrap

Nel paragrafo precedente tutti i metodi presentati per l'inferenza di verosimiglianza riguardavano metodi analitici basati su approssimazioni asintotiche. Una valida alternativa riguarda l'implementazione di metodi basati su simulazione, tra i quali il bootstrap ricopre un ruolo predominante nell'inferenza frequentista.

Dalla sua introduzione ad opera di Efron (1979), il **bootstrap** è oggi diventato ancora di più un metodo per valutare empiricamente gli errori in un vasto ventaglio di procedure statistiche, molto spesso offrendo un alto livello di accuratezza.

A causa della sua versatilità, addirittura i libri sull'argomento non sembrano sufficienti a spiegare tutte le caratteristiche del bootstrap, ma si limitano a trattare solo certi aspetti. Nella redazione di questo paragrafo si è fatto affidamento soprattutto a Davison e Hinkley (1997, Capitoli 2, 4 e 5), a Davison *et al.* (2003) ed infine a Young e Smith (2005, Capitolo 11).

La metodologia del bootstrap comprende un grande insieme di idee, ma il punto di partenza è dato da due principi fondamentali:

1. il *substitution principle*, detto anche *plug-in principle*, che consiste nella sostituzione di una distribuzione di probabilità  $F$  ignota con una sua stima  $\hat{F}$  costruita a partire dal campione osservato;

2. la *simulazione Monte Carlo*, tramite la quale è possibile calcolare caratteristiche o proprietà di  $\hat{F}$  rimpiazzando il calcolo analitico, che può essere difficile o addirittura intrattabile.

Nel primo punto risiede la vera idea del bootstrap, seguendo il concetto di sostituzione di una quantità non conosciuta con una stima, tipico dell'inferenza frequentista. Questa impostazione può essere naturalmente applicata a qualsiasi tipo di specificazione del modello statistico.

In ambito non parametrico,  $\hat{F}$  è la funzione di ripartizione empirica stimata a partire dal campione casuale  $y = (y_1, \dots, y_n)$  di numerosità  $n$ . Ad ogni punto  $y_i$  viene associata una probabilità pari a  $1/n$  che rappresenta la probabilità di quel punto di essere estratto. Nel contesto parametrico invece, la distribuzione di probabilità dipende da un parametro  $\theta$  di dimensione fissata,  $F(y; \theta)$ . Si può rimpiazzare il parametro con una sua stima consistente, ad esempio con la sua stima di massima verosimiglianza per il modello  $F$ , ottenendo  $\hat{F} = F(y; \hat{\theta})$ . Successivamente la simulazione Monte Carlo produce un insieme di campioni di numerosità  $n$  indipendenti tra di loro, estraendo con reinserimento da  $\hat{F}$ .

Indipendentemente dal tipo di ambito in cui si classifica il problema, si fa inferenza considerando  $\hat{F}$  come la vera distribuzione che ha prodotto i dati  $Y$ . Sarà tramite la simulazione da  $\hat{F}$  che otterremo le proprietà di uno stimatore  $\hat{\tau}$  di una caratteristica  $\tau = \tau(F)$  dipendente dalla vera distribuzione di probabilità  $F$ . Infatti verranno generati  $B$  campioni bootstrap  $(Y_1^*, \dots, Y_n^*)$  da  $\hat{F}$  e tramite le stime di ogni campione  $(\hat{\tau}_1^*, \dots, \hat{\tau}_B^*)$  si potranno stimare le proprietà di  $\hat{\tau}$ . Ad esempio, se  $T(Y)$  fosse una statistica sotto campionamento da  $F$ , si potrebbe stimare la sua varianza  $\text{Var}_F[T(Y)]$  tramite la varianza della stessa statistica sotto campionamento da  $\hat{F}$ , cioè usando  $\text{Var}_{\hat{F}}[T(Y)]$ .

Una parte corposa delle metodologie bootstrap è dedicata alla costruzione di intervalli di confidenza. I due approcci principali sono basati sulla definizione di quantità pivotali studentizzate oppure sull'utilizzo diretto delle repliche bootstrap  $(\hat{\tau}_1^*, \dots, \hat{\tau}_B^*)$ .

Il primo approccio è stato ispirato dalla statistica  $t$  di Student. Oltre a  $\hat{\tau}^*$ , richiede il calcolo di una stima della varianza bootstrap  $V^*$  per  $\hat{\tau}^*$  basata sugli stessi

campioni bootstrap. In questo modo si può definire la statistica studentizzata

$$Z^* = \frac{(\hat{\tau}^* - \hat{\tau})}{V^{*1/2}},$$

i cui quantili sono stimatori consistenti dei quantili di

$$Z = \frac{(\hat{\tau} - \tau)}{V^{1/2}}$$

per una grande classe di stimatori  $\hat{\tau}$  e dove  $V$  è la varianza dello stimatore  $\hat{\tau}$ . Si useranno quindi i quantili bootstrap  $z_{(1)}^* < \dots < z_{(B)}^*$  per costruire l'**intervallo di confidenza studentizzato**. Un intervallo di confidenza con copertura nominale  $(1 - \alpha)$  per  $\tau$  è dato da

$$\mathcal{I} \equiv \mathcal{I}_{(1-\alpha)} = \left\{ \tau : \hat{\tau} - V^{1/2} z_{((1-\frac{\alpha}{2})(B+1))}^* < \tau < \hat{\tau} - V^{1/2} z_{(\frac{\alpha}{2}(B+1))}^* \right\},$$

dove  $z_{(q)}^*$  indica il  $q$ -esimo quantile ordinato di  $z^*$ .

Il secondo approccio considera i valori  $(\hat{\tau}_1^*, \dots, \hat{\tau}_n^*)$  come i quantili empirici di  $\hat{\tau}$ , senza alcun bisogno di costruire un pivot. Tale approccio fornisce un cosiddetto **intervallo basato sui percentili**. Per un intervallo di confidenza con copertura nominale  $(1 - \alpha)$  per  $\tau$  otteniamo

$$\mathcal{I}_{(1-\alpha)} = \left\{ \tau : \tau_{((1-\frac{\alpha}{2})(B+1))}^* < \tau < \tau_{(\frac{\alpha}{2}(B+1))}^* \right\}.$$

Esistono anche modifiche agli intervalli basati sui percentili in cui si può migliorare la copertura dell'intervallo stesso. L'intervallo ha la forma

$$\mathcal{I} = \left\{ \tau : \tau_{((1-\frac{\alpha_1}{2})(B+1))}^* < \tau < \tau_{(\frac{\alpha_2}{2}(B+1))}^* \right\},$$

dove si stimano empiricamente  $\alpha_1$  e  $\alpha_2$ . Questo porta ad intervalli con distorsione corretta e accelerati, chiamati anche **intervalli BC<sub>a</sub>**.

Tipicamente gli intervalli bootstrap coprono meno di quanto indica il valore nominale. Un intervallo basato sui percentili di livello nominale  $1 - \alpha$  ha in realtà copertura  $1 - \alpha + O(n^{-1/2})$ . Ciò significa che l'intervallo ha un ordine di errore pari a  $O(n^{-1/2})$ , cioè si dice che approssima al *primo ordine*.

Gli intervalli studentizzati e  $BC_a$  si comportano meglio, approssimando al *secondo ordine*, ovvero avendo un errore di ordine  $O(n^{-1})$ . Tuttavia gli intervalli  $BC_a$  sono invarianti rispetto a trasformazioni, proprietà che gli intervalli studentizzati non possiedono.

La scarsa copertura degli intervalli bootstrap può essere migliorata attraverso un procedimento di *prepivotting* che verrà esaminato in dettaglio nel Capitolo 2.

Le metodologie bootstrap possono essere utilizzate anche per le verifiche d'ipotesi. Supponiamo di avere un'ipotesi nulla  $H_0$  e una statistica test  $t(Y)$ . Il livello di incongruenza tra i dati osservati  $y$  e l'ipotesi  $H_0$  è misurato dal *p-value*, ovvero dalla probabilità, sotto l'ipotesi nulla, che il valore del test sia significativo. In notazione

$$\alpha^{oss} = \mathbb{P}_{H_0}(t(Y) \geq t(y)) = \mathbb{P}(t(Y) \geq t(y)|H_0),$$

dove  $t(y)$  è il valore osservato della statistica  $t(Y)$ .

A volte il calcolo di  $\alpha^{oss}$  è complesso e quindi si può stimare il *p-value* con  $\hat{\alpha}^{oss}$ , simulando da  $\hat{H}_0$ , a sua volta una stima di  $H_0$ . Vengono generati  $B$  campioni di  $y_b^*$ , che servono a costruire  $B$  statistiche  $t^* = (t(y_1^*), \dots, t(y_B^*))$ . Infine si stima  $\hat{\alpha}^{oss} = \mathbb{P}(t(Y) \geq t(y)|\hat{H}_0)$  con

$$\hat{\alpha} = \frac{\#\{t^* \geq t(y)\}}{B},$$

dove  $\#$  indica il numero di valori  $t^*$  maggiori o uguali a  $t(y)$ .

Un'alternativa per la costruzione di un test bootstrap è tramite un intervallo di confidenza, sfruttando la nota dualità tra i due metodi. Un'idea si basa sull'uso di una quantità pivotale, come può essere quella usata per costruire gli intervalli studentizzati, da usare come test pivot per verificare l'ipotesi nulla. Un approccio più semplice consiste eventualmente nell'utilizzo diretto di un intervallo di confidenza di livello  $1 - \alpha$  per  $\tau$  e vedendo se  $\alpha^{oss} \geq \alpha$ , purché la verifica d'ipotesi riguardi un parametro. Per avvantaggiarsi di queste possibilità, bisogna fare attenzione alla forma della verifica d'ipotesi: se l'ipotesi nulla è unilaterale, allora anche l'intervallo di confidenza usato dev'essere unilaterale.

Anche il procedimento inverso, ovvero partire dal risultato di un test per costruire un intervallo di confidenza è molto efficace. Tramite il procedimento di *prepivotting*

si parte dalla stima di un  $p$ -value per ottenere l'intervallo di confidenza invertendo la funzione  $p$ -value. Il *prepivoting* verrà discusso più approfonditamente nel Capitolo 2.

# Capitolo 2

## Soluzioni per i problemi di Neyman e Scott

### 2.1 Introduzione

Questo capitolo discute in modo approfondito il tipo di modelli principale della tesi, ovvero i modelli stratificati con parametri incidentali. Nel Paragrafo 1.4 è stata presentata la situazione in cui il modello possiede parametri incidentali, ovvero parametri di disturbo che dipendono dalla numerosità campionaria e che portano al cosiddetto problema di Neyman e Scott.

Sotto la denominazione di problemi con parametri incidentali rientrano nella letteratura econometrica anche i **modelli per dati di panel**. Le variabili aleatorie  $Y_{ij}$ , di cui  $y_{ij}$  sono realizzazioni, con  $i = 1, \dots, q$  e  $j = 1, \dots, m_i$ , si definiscono dati di panel se per  $q$  unità statistiche si rilevano  $m_i$  misure, che, ad esempio, possono rappresentare degli istanti di tempo. Qui  $q$  indica il numero degli strati e  $m_i$  il numero di repliche per lo strato  $i$ -esimo. La numerosità campionaria totale è  $n = \sum_{i=1}^q m_i$ . Per semplicità, ma senza perdita di generalità, si assumerà che ogni strato abbia lo stesso numero di repliche:  $m_i = m$  per ogni  $i = 1, \dots, q$ . La numerosità campionaria totale allora può essere calcolata semplicemente come  $n = mq$ . Nel seguito ci si riferirà a modelli con strati indipendenti tra di loro nella forma

$$Y_{ij} \sim f_{Y_{ij}}(y_{ij}; \psi, \zeta_i), \quad (2.1)$$

dove il parametro è nuovamente  $\theta = (\psi, \zeta)$ , con  $\psi$  parametro di interesse e  $\zeta = (\zeta_1, \dots, \zeta_q)$  parametro di disturbo di dimensione  $q$ . Eventualmente si possono considerare situazioni in cui sono presenti anche delle variabili esplicative  $x_{ij}$ . Il modello allora diventa

$$Y_{ij} \sim f_{Y_{ij}}(y_{ij}; \psi, \zeta_i, x_{ij}), \quad (2.2)$$

con le stesse considerazioni fatte sopra.

In ambito econometrico si usano tipicamente i modelli per dati di panel, in cui gli strati giocano il ruolo di individui e, per tenere conto dell'eterogeneità tra gli stessi, viene associato un parametro specifico ad ogni individuo. Ci saranno quindi  $q$  parametri individuali, che vengono supposti ad effetti fissi, ovvero senza il bisogno di definire una distribuzione sottostante ad essi. Tali parametri altro non sono che i parametri incidentali.

Come già accennato, i modelli con dati di panel possono incorrere in problemi quando si fa inferenza con le usuali tecniche di verosimiglianza. Si presenta nel dettaglio l'esempio appena accennato nel Paragrafo 1.4.

**Esempio 1** (Bartolucci *et al.*, 2012). *Si consideri il modello basato sull'assunzione*

$$Y_{ij} = \zeta_i + \varepsilon_{ij}, \quad i = 1, \dots, q, j = 1, \dots, m,$$

dove è stato posto  $m_i = m$  per ogni  $i$  e dove le  $\varepsilon_{ij}$  sono variabili casuali normali indipendenti con media 0 e varianza  $\psi$ , in notazione  $\varepsilon_{ij} \sim N(0, \psi)$ . Tralasciando termini additivamente costanti, la log-verosimiglianza è

$$l(\psi, \zeta) = -\frac{mq}{2} \log \psi - \frac{1}{2} \sum_i \sum_j \frac{(y_{ij} - \zeta_i)^2}{\psi}.$$

La stima di massima verosimiglianza per  $\zeta_i$  vincolata a  $\psi$  non dipende da  $\psi$  ed è pari alla media dell' $i$ -esimo strato,  $\hat{\zeta}_{i\psi} = \hat{\zeta}_i = \bar{y}_i$ . Si può scrivere direttamente la verosimiglianza profilo come

$$l_P(\psi) = -\frac{mq}{2} \log \psi - \frac{1}{2} \sum_i \sum_j \frac{(y_{ij} - \bar{y}_i)^2}{\psi}.$$

La score profilo risulta essere

$$\frac{\partial}{\partial \psi} l_P(\psi) = -\frac{mq}{2\psi} + \frac{1}{2} \sum_i \sum_j \frac{(y_{ij} - \bar{y}_i)^2}{\psi^2},$$

la cui distorsione non è nulla, ma bensì pari a  $-q/(2\psi)$ . Questo comporta che lo stimatore di massima verosimiglianza per  $\psi$ , pari a

$$\hat{\psi} = \frac{1}{mq} \sum_i \sum_j (y_{ij} - \bar{y}_i)^2,$$

sia non consistente. Nello specifico

$$\hat{\psi} \xrightarrow{p} \psi \frac{m-1}{m}$$

quando  $q \rightarrow \infty$  e  $m$  fissato.

△

Dall'esempio è evidente come al divergere di  $q$  la distorsione della funzione *score* sia di ordine  $O(q)$ . Infatti è possibile dimostrare che in un modello per dati stratificati, la distorsione della funzione *score* si accumula attraverso il numero di strati (McCullagh e Tibshirani, 1990). Per correggere questo problema e per studiare in generale situazioni in cui il numero dei parametri di disturbo è grande, è utile studiare gli strumenti per l'inferenza a due indici asintotici, in cui sia  $q$  che  $m$  possono andare all'infinito.

La scelta dei due indici è obbligata nel senso che se  $q$  è fissato e  $m$  può tendere all'infinito, si può applicare la teoria asintotica standard. Viceversa se  $m$  è fissato si ha che  $q = O(n)$  e quindi vengono estremizzate le caratteristiche dei problemi di Neyman e Scott (1948), come la non consistenza della stima di massima verosimiglianza. In tal caso il problema si può risolvere se per il modello è possibile separare dalla verosimiglianza globale le parti di verosimiglianza contenenti i parametri di disturbo, come accade nella verosimiglianza marginale e nella verosimiglianza condizionata. Tuttavia, come è noto, la definizione di tali pseudo-verosimiglianze è possibile solo per poche famiglie di modelli, quali le famiglie di gruppo e le famiglie esponenziali. Invece la verosimiglianza profilo e la verosimiglianza profilo modificata

sono strumenti generali per l'inferenza, ma non sempre questi metodi funzionano. Sartori (2003) ha mostrato le condizioni per cui soprattutto la verosimiglianza profilo modificata è affidabile nel contesto dei due indici asintotici. A questa rassegna è dedicato il Paragrafo 2.2

Quando l'affidabilità dei suddetti strumenti viene meno, è necessario ricercare una soluzione diversa. Una strada molto promettente consiste nell'implementare un metodo di simulazione via bootstrap. In particolare nel Paragrafo 2.3 si tratterà una fusione tra degli accorgimenti sul bootstrap parametrico in presenza di parametri di disturbo e un processo di correzione per le regioni di confidenza chiamato *prepivotting*.

## 2.2 Analisi a due indici asintotici

Con le indicazioni sul modello (2.1) fornite nel paragrafo precedente si può scrivere verosimiglianza come

$$L(\psi, \zeta) = \prod_{i=1}^q \prod_{j=1}^m f_{Y_{ij}}(y_{ij}; \psi, \zeta) = \prod_{i=1}^q L^i(\psi, \zeta_i),$$

dove  $L^i(\psi, \zeta_i)$  è la verosimiglianza relativa all' $i$ -esimo strato grazie all'indipendenza tra i vari strati. Si ottiene quindi la log-verosimiglianza

$$l(\psi, \zeta) = \sum_{i=1}^q l^i(\psi, \zeta_i) = \sum_{i=1}^q \log L^i(\psi, \zeta_i). \quad (2.3)$$

Come si può vedere dalla (2.3), la log-verosimiglianza, per  $\psi$  fissato, si può separare rispetto ai parametri incidentali, in quanto nello strato  $i$ -esimo è presente solo il parametro di disturbo  $\zeta_i$ . Il vantaggio maggiore in questa situazione è la possibilità di ricavare il vettore di stime vincolate  $\hat{\zeta}_\psi = (\hat{\zeta}_{1\psi}, \dots, \hat{\zeta}_{q\psi})$  attraverso  $q$  equazioni di stima  $\frac{\partial}{\partial \zeta_i} l^i(\psi, \zeta_i) = 0$  indipendenti tra di loro. Di conseguenza si può scrivere la log-verosimiglianza profilo per  $\psi$  come somma delle  $q$  log-verosimiglianze profilo di tutti gli strati

$$l_P(\psi) = l(\psi, \hat{\zeta}_\psi) = \sum_{i=1}^q l^i(\psi, \hat{\zeta}_{i\psi}) = \sum_{i=1}^q l_P^i(\psi). \quad (2.4)$$

Si può calcolare anche la log-verosimiglianza profilo modificata. Globalmente si presenta nella forma

$$l_M(\psi) = l_P(\psi) + \log M(\psi) = l_P(\psi) + M'(\psi),$$

in cui il logaritmo del fattore di modificazione è pari a

$$M'(\psi) = -\log |l_{\zeta; \hat{\zeta}}(\psi, \hat{\zeta}_\psi; \hat{\psi}, \hat{\zeta}, a)| + \frac{1}{2} \log |j_{\zeta\zeta}(\psi, \hat{\zeta}_\psi; \hat{\psi}, \hat{\zeta}, a)|,$$

una volta ottenuta la statistica ancillare  $a$  e aver calcolato la derivata nello spazio campionario  $l_{\zeta; \hat{\zeta}}(\theta)$ . A tal proposito si può vedere nel seguente esempio come ottenere una derivata nello spazio campionario partendo dal modello descritto nell'Esempio 1 per calcolare la log-verosimiglianza modificata.

**Esempio 2** (segue dall'Esempio 1). *Per lo stesso modello, si può riscrivere la log-verosimiglianza come*

$$l(\psi, \zeta; \hat{\psi}, \hat{\zeta}) = -\frac{mq}{2} \log \psi - \frac{1}{2} \frac{mq\hat{\psi} + q \sum_i (\hat{\zeta}_i - \zeta_i)^2}{\psi},$$

*poiché in questo modello, che è una famiglia esponenziale piena, la s.m.v. è una statistica sufficiente. In particolare:*

$$\sum_i \sum_j (y_{ij} - \zeta_i)^2 = \sum_i \sum_j (y_{ij} - \hat{\zeta}_i)^2 + m \sum_i (\hat{\zeta}_i - \zeta_i)^2.$$

*Questo implica che*

$$l_{\zeta_i}(\psi, \zeta_i; \hat{\psi}, \hat{\zeta}) = \frac{\partial}{\partial \zeta_i} l(\psi, \zeta_i; \hat{\psi}, \hat{\zeta}) = m \frac{\hat{\zeta}_i - \zeta_i}{\psi}$$

*e, di conseguenza, si ottiene la seguente derivata mista*

$$l_{\zeta_i; \hat{\zeta}_i}(\psi, \zeta_i; \hat{\psi}, \hat{\zeta}_i) = \frac{m}{\psi},$$

*che in questo modello coincide con la derivata seconda cambiata di segno calcolata*

rispetto a  $\zeta_i$ , ovvero

$$l_{\zeta_i \zeta_i}(\psi, \zeta_i) = -\frac{m}{\psi}.$$

A questo punto è facile calcolare il fattore di modificazione  $M(\psi)$ , tralasciando le parti che non dipendono da  $\psi$ . Si ottiene

$$M'(\psi) = -\frac{q}{2} \log \psi,$$

e quindi la verosimiglianza profilo modificata per  $\psi$  è

$$l_M(\psi) = -q \frac{m-1}{2} \log \psi - \frac{1}{2} \sum_i \sum_j \frac{(y_{ij} - \bar{y}_i)^2}{\psi}.$$

La funzione score corrispondente è uguale a

$$\frac{\partial}{\partial \psi} l_M(\psi) = -q \frac{m-1}{2\psi} + \frac{1}{2} \sum_i \sum_j \frac{(y_{ij} - \bar{y}_i)^2}{\psi^2},$$

che risulta essere non distorta. Infine, lo stimatore della verosimiglianza modificata è pari a

$$\hat{\psi}_M = \frac{1}{q(m-1)} \sum_i \sum_j (y_{ij} - \bar{y}_i)^2,$$

il quale, oltre ad essere non distorto, è anche consistente per  $q \rightarrow \infty$ , anche per  $m$  fissato.

△

Il vantaggio della verosimiglianza profilo modificata sulla verosimiglianza profilo, come già visto, è la riduzione della distorsione della funzione *score*. In presenza di un modello per dati di panel, ogni strato contribuisce alla distorsione totale della funzione *score* della log-verosimiglianza profilo di un valore pari a

$$\mathbb{E}_\theta \left[ \frac{\partial}{\partial \psi} l_P^i \right] = -\rho_\psi^i + O(m^{-1}),$$

dove  $\rho_\psi^i$  è una quantità di ordine  $O(1)$ . Si può mostrare che la distorsione totale accumulata è pari a  $-\sum_{i=1}^q \rho_\psi^i = O(q)$ . Il valore atteso del logaritmo della modificazione

per lo stesso strato è invece pari a

$$\mathbb{E}_\theta \left[ \frac{\partial}{\partial \psi} M'_i(\psi) \right] = -\mathbb{E}_\theta \left[ \frac{\partial}{\partial \psi} l_P^i \right] + O(m^{-1}) = \rho_\psi^i + O(m^{-1}),$$

dove  $M'_i(\psi)$  indica il logaritmo del fattore di modificazione nell' $i$ -esimo strato. Di conseguenza la modificazione in ogni strato elimina il termine dominante della distorsione della funzione *score* provocata dalla verosimiglianza profilo, a meno di un termine di ordine  $O(m^{-1})$ .

Tuttavia, iniziano a sorgere dei problemi quando si tenta di verificare le proprietà asintotiche della verosimiglianza profilo e della verosimiglianza profilo modificata, soprattutto in un modello con dati stratificati.

Nel seguito si consideri la situazione in cui sia  $m$  che  $q$  possano tendere all'infinito. La condizione sufficiente per cui la verosimiglianza profilo presenti le usuali proprietà asintotiche è che  $m$  cresca più rapidamente di  $q$ , in notazione  $1/m = o(q^{-1})$ . D'altro canto, Sartori (2003) ha dimostrato che il fatto che la verosimiglianza profilo modificata riduca la distorsione della funzione *score* permette di ottenere proprietà asintotiche migliori, a patto che  $m$  cresca più rapidamente di  $q^{1/3}$ , in notazione  $1/m = o(q^{-1/3})$ . Si noti quindi che la condizione riguardante la verosimiglianza profilo modificata è più debole della condizione riguardante la verosimiglianza profilo.

Questo risultato vale per qualsiasi statistica test si applichi ai dati. Si considerino le versioni basate sulla verosimiglianza profilo modificata dei test visti nel Paragrafo 1.7, ovvero

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

e le loro versioni direzionali  $r_M(\psi)$ ,  $r_{Mu}(\psi)$  e  $r_{Me}(\psi)$ , dove  $l_{M^*}(\psi)$  rappresenta il vettore *score* di  $l_M(\psi)$  e  $j_M(\psi)$  è la matrice di informazione osservata relativa a  $l_M(\psi)$ . È stato dimostrato che le varie versioni modificate sono asintoticamente equivalenti tra di loro e che inoltre sono asintoticamente equivalenti ai test calcolati attraverso la verosimiglianza profilo in un contesto asintotico standard. Quindi,

prendendo il caso in cui  $\psi$  è scalare, i test  $W$  si distribuiscono asintoticamente come una  $\chi_1^2$  e i test  $r$  come una  $N(0,1)$ . Come i test basati sulla profilo, anche i test basati sulla verosimiglianza profilo modificata mantengono un errore del *primo ordine*. Tuttavia in un contesto a due indici asintotici in cui  $1/m = o(q^{-1/3})$  ma  $1/m$  non è  $o(q^{-1})$  i test  $r_M$  si distribuiscono ancora asintoticamente una  $N(0,1)$ , mentre i test  $r_P$  non più.

In ogni caso, a prescindere da  $q$  e  $m$ , Sartori (2003) mostra che la distorsione di  $\hat{\psi}$  è di ordine  $O(m^{-1})$ , mentre la distorsione di  $\hat{\psi}_M$  è minore, e precisamente di ordine  $O(m^{-2})$ .

Quando la condizione sufficiente per ottenere gli usuali risultati asintotici non è valida, si assiste ad un progressivo peggioramento nel funzionamento dei metodi inferenziali, anche se sono basati sulla verosimiglianza profilo modificata. Ad esempio, un intervallo di confidenza costruito a partire da uno dei test sopra descritti ha una copertura minore di quanto sia indicato nominalmente. In questa tesi si proporrà un metodo per cercare di correggere questo problema.

## 2.3 Miglioramenti via bootstrap

La ricerca sulle metodologie bootstrap è proseguita proficuamente e in anni recenti sono stati proposti degli accorgimenti che potrebbero ben adattarsi al contesto dei problemi di Neyman e Scott.

Il primo accorgimento che verrà trattato riguarda il bootstrap parametrico, fondamentale per questa tesi. In particolare, si esaminerà il caso in cui sono presenti i parametri di disturbo. Un altro importante accorgimento riguarda invece un procedimento per migliorare la copertura degli intervalli di confidenza, facendo uso delle potenzialità del bootstrap. Tale processo viene chiamato *prepivotting*.

### 2.3.1 Bootstrap per problemi con parametri di disturbo

Sia  $Y = (Y_1, \dots, Y_n)$  un vettore casuale continuo con densità di probabilità  $f_{Y_i}(y_i; \theta)$  con  $\theta = (\psi, \zeta)$  dove  $\zeta$  è la componente di disturbo e  $\psi$  il parametro di interesse. Lo strumento principale qui presentato per questo problema è la log-verosimiglianza profilo  $l_P(\psi) = l(\psi, \hat{\zeta}_\psi)$ , dove si sostituisce con  $(\psi, \hat{\zeta}_\psi)$  il parametro

$(\psi, \zeta)$  nella funzione di log-verosimiglianza  $l(\psi, \zeta)$ . Successivamente l'inferenza sul parametro di interesse può essere basata, quando  $\psi$  è scalare, per esempio, sulla statistica  $r_P(\psi)$ , vista nel Paragrafo 1.7.

Visto che la distribuzione normale di  $r_P(\psi)$  è solo approssimata, si potrebbe pensare di impostare un bootstrap parametrico per migliorare l'accuratezza dell'approssimazione della distribuzione della statistica. Normalmente lo schema bootstrap per l'inferenza parametrica prevede di simulare i campioni  $y^*$  generando da  $f_Y(y; \hat{\theta})$ , con  $\hat{\theta} = (\hat{\psi}, \hat{\zeta})$  s.m.v. globale, e successivamente stimare la vera distribuzione di  $r_P(\psi)$  attraverso le statistiche  $r_P^*(\psi)$  generate. DiCiccio e Romano (1995) hanno mostrato che la distribuzione di  $r_P^*(\psi)$  approssima al *secondo ordine* la vera distribuzione di  $r_P(\psi)$ , ovvero con un errore di ordine  $O(n^{-1})$ . L'approssimazione risulta essere in generale migliore di quella basata sulle considerazioni asintotiche di  $r_P(\psi)$ , che si ferma al *primo ordine*.

In realtà, analiticamente, si potrebbe fare ancora meglio, utilizzando una versione aggiustata di  $r_P(\psi)$ , ideata da Barndorff-Nielsen (1986). La versione aggiustata ha forma

$$r_A(\psi) = r_P(\psi) + r_P(\psi)^{-1} \log \left( \frac{u_P(\psi; a)}{r_P(\psi)} \right) \quad (2.5)$$

e si distribuisce come una  $N(0,1)$  con un errore di ordine  $O(n^{-3/2})$ . Si noti che la quantità  $u_P$  dipende da una statistica ancillare  $a$  e coinvolge in generale derivate nello spazio campionario.

Si possono ottenere gli stessi risultati di approssimazione della versione aggiustata di  $r_P(\psi)$  cambiando lo schema bootstrap. Un'osservazione importante deriva dal fatto che, anche in un contesto con dei parametri di disturbo, si usa comunque  $(\hat{\psi}, \hat{\zeta})$  per generare campioni bootstrap. Una modifica di questo schema, che tenga conto della complessità del parametro, consiste nella generazione dei vari vettori  $y^*$  da  $f_Y(y; \psi, \hat{\zeta}_\psi)$ , ovvero utilizzando il parametro  $(\psi, \hat{\zeta}_\psi)$ , con la stima di  $\zeta$  vincolata a  $\psi$ . Seguendo questo schema, si calcolano le statistiche  $r_P^*(\psi)$  a partire dai campioni  $y^*$  e la distribuzione di  $r_P^*(\psi)$  approssima la vera distribuzione di  $r_P(\psi)$  con un errore di ordine  $O(n^{-3/2})$  (DiCiccio *et al.*, 2001).

In conclusione, sia lo schema modificato per il bootstrap parametrico che l'aggiustamento (2.5) offrono un'approssimazione per la distribuzione di  $r_P(\psi)$  al *terzo*

*ordine.* Ciò nonostante, la versione aggiustata di  $r_P(\psi)$  comporta ulteriori calcoli analitici e, per via della specificazione di una statistica ancillare, una scarsa applicabilità al di fuori di un insieme ristretto di classi parametriche.

Inoltre l'accuratezza derivata da  $r_A(\psi)$  rimane elevata fintanto che il parametro di disturbo è di dimensione ristretta. In una situazione con un parametro di disturbo di dimensione elevata, come può essere quella con un problema con dati di panel,  $r_A(\psi)$  è essenzialmente equivalente a  $r_M(\psi)$  (Sartori, 2003). Per questo motivo, ci si concentrerà su  $r_M(\psi)$ , tralasciando lo studio di  $r_A(\psi)$ .

L'implementazione dell'approccio bootstrap modificato resta inalterata indipendentemente dalla dimensione di  $\zeta$ . Per questi motivi, utilizzare lo schema modificato per il bootstrap in questo contesto risulta essere versatile e vantaggioso. Non esistono risultati teorici in un contesto a due indici asintotici relativi al bootstrap sia con schema standard che con schema modificato. In questa tesi verranno condotti esperimenti di simulazione per verificarne le proprietà in modelli per dati di panel. In particolare, si considererà come proposta un approccio basato sul prepivoting.

### 2.3.2 Prepivoting

Sia  $y = (y_1, \dots, y_n)$  un campione da  $Y$ , vettore casuale  $n$ -dimensionale, con funzione di densità dipendente da  $\theta$  parametro ignoto e tale che  $\theta \in \Theta$ . Quando si vuole costruire un intervallo di confidenza è difficile reperire una quantità pivotale con distribuzione esatta. Si supponga che  $u_n(\theta) = u_n(y, \theta)$  sia una quantità pivotale e che si sia in grado di costruire un intervallo di confidenza nella forma

$$\mathcal{I}(y) \equiv \mathcal{I}_{1-\alpha}(y) = \{\theta \in \Theta : u(y, \theta) \leq c_n(1 - \alpha)\},$$

dove  $c_n(1 - \alpha)$  rappresenta il quantile  $(1 - \alpha)$  della distribuzione di  $u_n(\theta)$ . Sotto campionamento ripetuto da  $Y$ , si ha che approssimativamente l' $(1 - \alpha)\%$  delle volte l'intervallo  $\mathcal{I}(y)$  conterrà il vero  $\theta$ . L'*errore di copertura* dell'intervallo  $\mathcal{I}(y)$  sarà quindi definito come

$$\mathbb{P}_\theta[\theta \in \mathcal{I}(y)] - (1 - \alpha).$$

Molto spesso  $u_n(\theta)$  sarà monotono in  $\theta$ , portando l'intervallo di confidenza ad essere in una forma semi-infinita  $(\hat{\theta}_l(y), \infty)$  o  $(-\infty, \hat{\theta}_u(y))$ , dove il limite inferiore

(superiore)  $\hat{\theta}$  si ottiene risolvendo l'equazione  $u_n(\theta) = c_n(1 - \alpha)$ .

Nel caso fosse necessario un intervallo in forma chiusa, una possibilità simmetrica è data dalla differenza di insiemi tra due intervalli di confidenza in forma semi-infinita. Quindi un intervallo  $\mathcal{I}_{1-\alpha}(y)$  di copertura nominale  $1 - \alpha$  è dato da

$$\mathcal{I}_{1-\alpha} = \mathcal{I}_{1-\alpha/2}(y) \setminus \mathcal{I}_{\alpha/2}(y).$$

Al di fuori di una piccola classe di problemi è molto difficile reperire una quantità pivotale per costruire intervalli o regioni di confidenza, sia che siano esatte o approssimate. Si supponga che ora  $u_n(y, \theta)$  non sia una quantità pivotale, ma che abbia una distribuzione che possa essere stimata usando il campione e sia  $c_n(1 - \alpha)$  il quantile  $(1 - \alpha)$  di questa distribuzione stimata. Allora si può usare  $u_n(y, \theta)$  per costruire un intervallo di confidenza, il cui errore di copertura diminuisce all'aumentare della dimensione del campione. Quando  $u_n(y, \theta)$  non è una vera quantità pivotale, verrà definita **confidence set root** (Beran, 1987).

Un procedimento generale per reperire un *confidence set root*  $u_n(y, \theta)$  è il seguente: sia  $s_n(\theta) \equiv s_n(Y, \theta)$  un pivot e sia  $H_n(\cdot; F)$  la funzione di ripartizione di  $s_n(\theta)$ , ovvero  $H_n(x; F) = \mathbb{P}[s_n(Y, \theta) \leq x | F]$ . Si supponga anche che asintoticamente  $s_n(Y, \theta)$  abbia funzione di ripartizione  $H(\cdot; F)$ . Sia  $\hat{F}_n$  una stima consistente della distribuzione  $F$ , ovvero tale che per una metrica  $d$  appartenente ad un opportuna famiglia di distribuzione, si abbia  $d(\hat{F}_n, F) \xrightarrow{p} 0$ . In un contesto parametrico  $\hat{F}_n$  potrebbe essere la stima ottenuta attraverso la massima verosimiglianza. Allora delle possibili radici dell'intervallo di confidenza possono essere

$$u_n(y, \theta) = \hat{H}\{s_n(\theta)\} \tag{2.6}$$

e

$$u_n(y, \theta) = \hat{H}_n\{s_n(\theta)\}, \tag{2.7}$$

dove  $\hat{H} = H(\cdot; \hat{F}_n)$  e  $\hat{H}_n = H_n(\cdot; \hat{F}_n)$  sono rispettivamente le stime di  $H(\cdot; F)$  e di  $H_n(\cdot; F)$ .

**Esempio 3** (vedi Esempio 11.3 di Young e Smith, 2005). *Sia  $y_1, \dots, y_n$  un campione casuale semplice di numerosità  $n$  da una  $N(\mu, \sigma^2)$ , con entrambi i parametri  $\mu$  e  $\sigma^2$*

ignoti. L'inferenza è richiesta su  $\mu$ . La s.m.v. per  $\mu$  è  $\hat{\mu} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ , mentre quella per  $\sigma^2$  è  $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$ .

La quantità  $s_n(\mu) = \sqrt{n}(\hat{\mu} - \mu)$  è quantità pivotale in quanto sia la distribuzione esatta che quella asintotica di  $s_n(\mu)$  è pari a  $N(0, \sigma^2)$ . Per questo motivo,  $H_n(x) \equiv H(x) = \Phi(x/\sigma)$  e, di conseguenza,  $\hat{H}_n(x) \equiv \hat{H}(x) = \Phi(x/\hat{\sigma})$ . Sfruttando la (2.6) e la (2.7) si arriva allo stesso confidence set root di forma

$$u_n(y, \mu) = \Phi \left\{ \frac{\sqrt{n}(\hat{\mu} - \mu)}{\hat{\sigma}} \right\} \quad (2.8)$$

In realtà, per un intervallo di confidenza più naturale si sfrutta il seguente noto risultato:

$$\frac{\sqrt{n-1}(\hat{\mu} - \mu)}{\hat{\sigma}} \sim t_{n-1},$$

ottenendo così il confidence set root

$$u_n(y, \mu) = T \left\{ \frac{\sqrt{n-1}(\hat{\mu} - \mu)}{\hat{\sigma}} \right\}, \quad (2.9)$$

dove  $T$  si riferisce alla distribuzione  $t$  di Student con  $n-1$  gradi di libertà,  $t_{n-1}$ . Gli intervalli di confidenza costruiti mediante la (2.9) hanno copertura esatta  $(1 - \alpha)$ , mentre l'uso della (2.8) porta ad un errore di ordine  $O(n^{-1/2})$ .

△

Una possibilità molto interessante è quella del **prepivoting**: dalla prospettiva del *prepivoting* si tenta di costruire o trasformare una radice  $u_n(\theta)$  in una variabile aleatoria uniforme  $\mathcal{U}(0,1)$ . Infatti se  $U = u_n(y, \theta) \sim \mathcal{U}(0,1)$ , allora l'intervallo di confidenza avrebbe esattamente la copertura desiderata pari a  $(1 - \alpha)$ :

$$\mathbb{P}_\theta[\theta \in \mathcal{I}(y)] = \mathbb{P}_\theta[u_n(y, \theta) \leq 1 - \alpha] = 1 - \alpha.$$

Generalmente però  $U$  non è distribuita uniformemente, cosicché l'errore di copertura di  $\mathcal{I}$  risulta essere maggiore di zero. Una buona soluzione allora è quella di ricavare una stima della funzione di ripartizione della radice  $u_n(\theta)$  attraverso il bootstrap. Tramite un metodo bootstrap, si cerca di costruire un nuovo *confidence*

set root  $u_{n,1}(\theta)$  tale che l'intervallo di confidenza

$$\mathcal{I}(y) = \{\theta \in \Theta : u_{n,1}(y, \theta) \leq 1 - \alpha\}$$

abbia un errore di copertura inferiore per  $\theta$ , che dipende da quanto  $U_1 = u_{n,1}(\theta)$  è vicina all'uniformità.

L'approccio bootstrap parametrico convenzionale prevede che si stimi la funzione di ripartizione di  $u_n(y, \theta)$ ,  $G(x; \theta)$ , con

$$\widehat{G}(x) = \mathbb{P}^*[u_n(y^*, \hat{\theta}) \leq x], \quad (2.10)$$

dove  $\hat{\theta}$  è lo stimatore di massima verosimiglianza di  $\theta$ , costruito a partire dal campione originale  $y_1, \dots, y_n$ . In questa situazione  $\mathbb{P}^*$  denota la probabilità sotto l'estrazione di campioni bootstrap  $y^*$  dalla funzione di densità stimata  $f_Y(y; \hat{\theta})$ . Si definisce quindi il *prepivoted confidence set root*

$$\hat{u}_{n,1}(y, \theta) = \widehat{G}(u_n(y, \theta)), \quad (2.11)$$

per ogni valore candidato del parametro.

Questo procedimento ricalca l'idea discussa poco sopra: se il metodo bootstrap riuscisse a stimare esattamente la distribuzione di  $u_n(y, \theta)$ , in modo che  $\widehat{G}$  sia la vera funzione di ripartizione  $G$ , allora il *prepivoted root*  $\hat{u}_{n,1}(y, \theta)$  si distribuirebbe esattamente come una  $\mathcal{U}(0,1)$ . Il risultato deriva dalla proprietà della trasformazione integrale di probabilità: se  $Z$  è una variabile casuale con funzione di ripartizione  $F_Z(\cdot)$ , allora  $F_Z(Z) \sim \mathcal{U}(0,1)$ . In questo modo l'intervallo di confidenza  $\{\theta \in \Theta : u_{n,1}(y, \theta) \leq 1 - \alpha\}$  avrebbe esattamente il livello di copertura desiderato.

Usare una stima di  $G$  produce un errore, anche se generalmente l'errore associato a  $u_{n,1}(\theta)$  è minore di quello ottenuto sfruttando  $u_n(\theta)$ , si veda DiCiccio *et al.* (2001).

Riprendendo l'Esempio 3, si vede che fare il *prepivoting* della radice (2.8) porta alla radice (2.9), producendo intervalli di confidenza di livello esatto  $(1 - \alpha)$ .

Per una trattazione più approfondita sul prepivoting si veda Beran (1987, 1988).

### 2.3.3 Discussione

L'approccio bootstrap usato per il *prepivotting* non elimina completamente l'errore. Oltre ad un irrinunciabile, per quanto controllabile, errore di simulazione, si possono ottenere risultati non soddisfacenti specialmente in situazioni non parametriche e in situazioni parametriche che prevedono dei parametri di disturbo.

Per quanto concerne il secondo caso, si può ottenere un miglioramento modificando lo schema bootstrap come visto nel Paragrafo 2.3.1. Sia ancora  $\theta = (\psi, \zeta)$ , con  $\psi$  parametro di interesse e  $\zeta$  parametro di disturbo. Con lo schema bootstrap modificato si ottiene il seguente *prepivotted root* al posto di  $u_{n,1}(\psi)$ , per costruire un intervallo di confidenza per il parametro di interesse:

$$\tilde{u}_{n,1}(y, \psi) = \tilde{G}\{u_n(y, \psi); \psi\}, \quad (2.12)$$

con

$$\tilde{G}(x; \psi) = \mathbb{P}^\dagger\{u_n(y^\dagger, \psi) \leq x\}. \quad (2.13)$$

In questo caso  $\mathbb{P}^\dagger$  rappresenta la probabilità per l'estrazione di campioni bootstrap  $y^\dagger$  da  $f_Y(y; \hat{\theta}_\psi)$ , dove si prende  $\hat{\theta}_\psi = (\psi, \hat{\zeta}_\psi)$ , ovvero con la stima di  $\zeta$  vincolata a  $\psi$ .

Un'analisi teorica, in un contesto asintotico standard, condotta da Lee e Young (2005) sugli errori delle regioni di confidenza mostra che, prendendo come riferimento l'originale radice  $u_n$ , l'uso del *prepivotted root*  $\hat{u}_{n,1}$  riduce l'errore di copertura di un fattore pari a  $O(n^{-1/2})$ , mentre l'uso del *prepivotted root* basato sull'approccio di bootstrap modificato  $\tilde{u}_{n,1}$  riduce l'errore di un fattore pari a  $O(n^{-1})$ .

# Capitolo 3

## Studi di simulazione

### 3.1 Introduzione

La discussione teorica impostata nel capitolo precedente verrà analizzata attraverso degli studi di simulazione. In questo capitolo quindi si considereranno dei modelli per dati stratificati e, di volta in volta, si analizzeranno i comportamenti di verosimiglianza profilo, verosimiglianza profilo modificata e del procedimento di *prepivotting* per quanto riguarda le coperture degli intervalli di confidenza per il parametro di interesse  $\psi$ .

Quello che ci si aspetta è che le coperture basate sulla verosimiglianza profilo modificata siano migliori rispetto a quelle basate sulla verosimiglianza profilo, come già mostrato in diversi studi di simulazione (si veda ad esempio Sartori, 2003 e Bartolucci *et al.*, 2012). Nei casi in cui anche la verosimiglianza profilo modificata risulta insoddisfacente, cioè quando non si verifica la condizione  $1/m = o(q^{-1/3})$  nel contesto dei due indici asintotici, il procedimento di *prepivotting* applicato alla verosimiglianza profilo potrebbe fornire delle coperture ancora migliori, ossia più vicine al livello nominale dell'intervallo di confidenza.

Nel seguito si considererà la notazione fissata nel Capitolo 2, in cui i dati sono realizzazioni delle variabili aleatorie indipendenti  $Y_{ij}$ , con  $i = 1, \dots, q$  ad indicare il numero di strati e  $j = 1, \dots, m$  ad indicare il numero di repliche per strato, supposto uguale per ognuno di essi. I modelli, che verranno visti in dettaglio negli esempi, sono supposti a strati indipendenti con parametro strutturale  $\psi$  e parametri

incidentalmente  $\zeta = (\zeta_1, \dots, \zeta_q)$ .

Il primo esempio che verrà trattato riguarda variabili aleatorie gamma, in cui il parametro di forma  $\psi$  è comune a tutti gli strati mentre il parametro di scala  $1/\zeta_i$  si riferisce solo allo strato  $i$ -esimo. Questo esempio viene sviluppato nel Paragrafo 3.2 ed estende l'Esempio 2 di Sartori (2003).

Il secondo esempio tratta un modello normale, con media comune a tutti gli strati e varianza dipendente dallo strato. Il parametro d'interesse è quindi la media delle variabili casuali  $Y_{ij} \sim N(\mu, \sigma_i^2)$ ,  $\psi = \mu$ . L'esempio è già stato trattato da Severini (2000, Esempio 7.15), ma solo per un numero piccolo di strati e senza implementare il bootstrap. Una prima estensione dell'esempio appena citato è ad opera di Davison *et al.* (2003, Esempio 2), in cui si è effettuata la simulazione bootstrap, ma sempre per un numero limitato di strati. Nel Paragrafo 3.3 si considererà un numero maggiore di strati.

Infine nel Paragrafo 3.4 si esaminerà un modello per dati binari. Il modello considerato appartiene alla categoria dei modelli lineari generalizzati. Nel predittore lineare associato al modello,  $\eta_i = \alpha_i + \beta x_{ij}$ , il parametro strutturale è rappresentato da  $\psi = \beta$ . Si esamineranno due casi distinti per la funzione legame: il primo riguarda il legame logit e il secondo il legame probit.

Per ognuno di questi esempi si considererà uno studio di simulazione per valutare l'accuratezza di diversi metodi per costruire intervalli di confidenza. Come si può notare, in tutti gli esempi il parametro di interesse è scalare. Ciò permette di calcolare i test basati sulla verosimiglianza nella loro versione direzionale. In particolare verranno calcolati il test  $r_P(\psi)$  per la verosimiglianza profilo e il test  $r_M(\psi)$  per la verosimiglianza profilo modificata. Successivamente, si sfrutterà la normalità approssimata sia di  $r_P(\psi)$  che di  $r_M(\psi)$  e si confronterà la copertura empirica degli intervalli di confidenza con quella nominale fissata per costruire degli intervalli di confidenza unilaterali.

Inoltre per ogni dataset verrà calcolato lo stimatore vincolato  $(\psi, \hat{\zeta}_\psi)$  per calcolare la verosimiglianza profilo e verrà impostato un bootstrap parametrico secondo quanto descritto nel Paragrafo 2.3.1. Per  $B$  repliche bootstrap, verrà calcolato il test  $r_P(\psi)$ . Sostanzialmente si applica il procedimento di *prepivotting*, in cui  $r_P(\psi)$  è una quantità pivotale di cui si conosce la distribuzione e che quindi permette di costruire il *confidence set root*  $u_n(\psi) = \hat{H}\{r_P(\psi)\}$ , con  $H$  vera funzione di ripartizione di

$r_P(\psi)$ . Attraverso i test bootstrap  $r_P^*(\psi)$ , stimare  $N$   $p$ -value per un'ipotesi  $H_0$  con

$$\hat{\alpha}_{oss} = \frac{\#\{r_P^*(\psi) \geq r_P(\psi)\}}{B}$$

equivale a ottenere la stima della distribuzione del *prepivoted root*

$$\hat{u}_{n,1}(\psi) = \widehat{G}\{u_n(\psi)\} \sim \mathcal{U}(0,1).$$

Gli intervalli di confidenza costruiti a partire da  $\hat{u}_{n,1}(\psi)$  dovrebbero avere un errore di copertura inferiore rispetto a quelli derivati da  $r_P(\psi)$ , come già visto nella discussione al Paragrafo 2.3.3, sia pur in un contesto asintotico standard.

Per ogni esempio verrà effettuata una simulazione per  $q = 250, 500$  strati e  $m = 4, 8$  repliche per strato, quando non indicato diversamente, e per ogni simulazione si prenderà  $N = 2000$  repliche e  $B = 2000$  repliche bootstrap per ogni replica. I risultati verranno presentati sotto forma di una tabella con indicati i vari livelli di copertura effettivi relativi alla verosimiglianza profilo, alla verosimiglianza profilo modificata e al metodo di *prepivoting* bootstrap.

## 3.2 Modello gamma

Si supponga che  $Y_{ij}$  siano delle variabili aleatorie indipendenti tra di loro, tali per cui

$$Y_{ij} \sim \text{Gamma}(\psi, 1/\zeta_i),$$

con parametro di forma  $\psi$  comune a tutte le variabili aleatorie e parametro di scala  $1/\zeta_i$  dipendente dallo strato  $i$ -esimo. La funzione di densità per la variabile  $Y_{ij}$  è pari a

$$f_{Y_{ij}}(y_{ij}; \psi, \zeta_i) = \frac{\zeta_i^{-\psi}}{\Gamma(\psi)} y_{ij}^{\psi-1} \exp\left\{-\frac{y_{ij}}{\zeta_i}\right\}.$$

Grazie all'indipendenza tra gli strati e tra le repliche di ogni strato, si può scrivere la verosimiglianza come

$$L(\psi, \zeta) = \prod_{i=1}^q \prod_{j=1}^m f_{Y_{ij}}(y_{ij}; \psi, \zeta_i) = \prod_{i=1}^q L^i(\psi, \zeta_i),$$

dove  $L^i(\psi, \zeta_i)$  fa riferimento alla verosimiglianza dell' $i$ -esimo strato. Tralasciando le costanti additive, si vede che la log-verosimiglianza dell' $i$ -esimo strato è proporzionale a

$$l^i(\psi, \zeta_i) \propto -m\psi \log(\zeta_i) - m \log \Gamma(\psi) + \psi \sum_{j=1}^m \log(y_{ij}) - \frac{1}{\zeta_i} \sum_{j=1}^m y_{ij}. \quad (3.1)$$

Siccome si desidera fare inferenza su  $\psi$ , è necessario calcolare la stima di  $\zeta_i$  vincolata a  $\psi$ , attraverso l'equazione di stima  $\frac{\partial}{\partial \zeta_i} l^i(\psi, \zeta_i) = 0$ . Si ottiene

$$\frac{\partial}{\partial \zeta_i} l^i(\psi, \zeta_i) = -\frac{m\psi}{\zeta_i} + \frac{1}{\zeta_i^2} \sum_{j=1}^m y_{ij} = 0,$$

e, di conseguenza si ricava la soluzione analitica

$$\hat{\zeta}_{i\psi} = \frac{\sum_{j=1}^m y_{ij}}{m\psi} = \frac{\bar{y}_i}{\psi} \quad (3.2)$$

dove  $\bar{y}_i$  è la media dell' $i$ -esimo strato. Si può quindi esplicitare la log-verosimiglianza profilo per l' $i$ -esimo strato, sostituendo a  $\zeta_i$  in  $l^i(\psi, \zeta_i)$  la stima vincolata appena calcolata:

$$\begin{aligned} l_P^i(\psi) &= l^i(\psi, \hat{\zeta}_{i\psi}) \\ &= m\psi \log(m\psi) - m \log \Gamma(\psi) - m\psi + \psi \left[ \sum_{j=1}^m \log(y_{ij}) - m \log \left( \sum_{j=1}^m y_{ij} \right) \right]. \end{aligned}$$

La log-verosimiglianza profilo totale è la somma delle  $q$  log-verosimiglianze relative ad ogni strato

$$l_P(\psi) = \sum_{i=1}^q l_P^i(\psi).$$

Essendo questa una famiglia esponenziale piena, si può scrivere la log-verosimiglianza profilo in funzione della statistica sufficiente (Sartori, 2003, Esempio 2) che ha componenti  $u_i = \sum_{j=1}^m \log y_{ij}$  e  $v_i = \sum_{j=1}^m y_{ij}$  in relazione all' $i$ -esimo strato. Ponendo  $s_i = u_i - m \log v_i$ , si può scrivere

$$l_P^i(\psi) = m\psi \log(m\psi) - m \log \Gamma(\psi) - m\psi + \psi s_i,$$

e quindi globalmente si ottiene

$$l_P(\psi) = \sum_{i=1}^q l_P^i(\psi) = mq\psi \log(m\psi) - mq \log \Gamma(\psi) - mq\psi + \psi s, \quad (3.3)$$

con  $u = \sum_{i=1}^q u_i$  e  $s = u - m \sum_{i=1}^q \log v_i$ .

Tramite la stessa fattorizzazione, Sartori (2003) esplicita anche la log-verosimiglianza profilo modificata globale per questo modello, ossia

$$l_M(\psi) = q \left( m\psi - \frac{1}{2} \right) \log(m\psi) - mq \log \Gamma(\psi) - mq\psi + \psi s. \quad (3.4)$$

Inoltre, in questo caso esiste anche la verosimiglianza condizionata, sempre per il fatto che il modello è una famiglia esponenziale piena e che  $\psi$  è un parametro canonico. Considerando ancora le componenti della statistica sufficiente  $u$  e  $v_i$  per  $i = 1, \dots, q$ , Sartori (2003) mostra che la log-verosimiglianza condizionata è pari a

$$l_C(\psi) = q \log \Gamma(m\psi) - mq \log \Gamma(\psi) + \psi s. \quad (3.5)$$

Il vantaggio nell'aver a disposizione una log-verosimiglianza condizionata è che si tratta di una verosimiglianza in senso proprio, e quindi può essere usata come metro di riferimento. In queste simulazioni sarà utile per avere un confronto con gli altri metodi. Infatti gli intervalli di confidenza costruiti a partire da  $r_C(\psi)$ , test basato sul log-rapporto di verosimiglianza per la log-verosimiglianza condizionata, dovrebbero avere una copertura empirica essenzialmente uguale a quanto indicato nominalmente, viste le elevate numerosità campionarie.

Infine si imposta un bootstrap parametrico per  $r_P(\psi)$  generando i  $B$  dataset dal modello gamma di parametri  $(\psi, \hat{\zeta}_\psi) = (\psi, \bar{y}_i/\psi)$ , secondo quanto calcolato alla (3.2).

Nella prova di simulazione si è posto che il vero valore del parametro di forma è pari a  $\psi_0 = 1$ , mentre i  $q$  parametri di disturbo  $\zeta_{i0}$  vengono generati casualmente da una variabile casuale uniforme  $\mathcal{U}(0,100)$ . Per quanto riguarda il bootstrap, si fa riferimento all'ipotesi nulla  $H_0 : \psi = \psi_0 = 1$ , generando quindi da  $(\psi_0, \hat{\zeta}_{i\psi_0}) = (1, \bar{y}_i)$ .

I risultati delle simulazioni per questo modello sono presentati nelle Tabelle 3.1-3.4 per quanto riguarda le coperture empiriche, mentre i grafici da Figura 3.1 a

Figura 3.4 si riferiscono ai Q-Q plot dei  $p$ -value stimati.

Nominale	$\Phi(r_P)$	$\Phi(r_C)$	$\Phi(r_M)$	Prepivoting
0.010	0.000	0.011	0.005	0.011
0.025	0.000	0.025	0.015	0.028
0.050	0.000	0.046	0.031	0.047
0.100	0.000	0.102	0.065	0.103
0.250	0.000	0.241	0.178	0.245
0.500	0.000	0.495	0.404	0.500
0.750	0.000	0.730	0.649	0.738
0.900	0.000	0.899	0.842	0.903
0.950	0.000	0.953	0.927	0.952
0.975	0.000	0.977	0.961	0.975
0.990	0.000	0.992	0.983	0.993

Tabella 3.1. Coperture empiriche per la simulazione con il modello gamma con  $q = 250$  strati e  $m = 4$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

Nominale	$\Phi(r_P)$	$\Phi(r_C)$	$\Phi(r_M)$	Prepivoting
0.010	0.000	0.011	0.009	0.011
0.025	0.000	0.025	0.021	0.027
0.050	0.000	0.046	0.042	0.046
0.100	0.000	0.097	0.081	0.100
0.250	0.000	0.255	0.223	0.256
0.500	0.000	0.507	0.476	0.509
0.750	0.001	0.761	0.732	0.760
0.900	0.011	0.891	0.878	0.893
0.950	0.023	0.954	0.943	0.953
0.975	0.043	0.974	0.968	0.976
0.990	0.086	0.993	0.989	0.992

Tabella 3.2. Coperture empiriche per la simulazione con il modello gamma con  $q = 250$  strati e  $m = 8$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

---

Nominale	$\Phi(r_P)$	$\Phi(r_C)$	$\Phi(r_M)$	Prepivoting
0.010	0.000	0.011	0.004	0.010
0.025	0.000	0.022	0.011	0.022
0.050	0.000	0.036	0.021	0.038
0.100	0.000	0.086	0.042	0.086
0.250	0.000	0.234	0.147	0.238
0.500	0.000	0.478	0.351	0.484
0.750	0.000	0.724	0.623	0.730
0.900	0.000	0.890	0.822	0.889
0.950	0.000	0.949	0.896	0.951
0.975	0.000	0.974	0.948	0.974
0.990	0.000	0.990	0.976	0.991

---

Tabella 3.3. *Coperture empiriche per la simulazione con il modello gamma con  $q = 500$  strati e  $m = 4$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.*

---

Nominale	$\Phi(r_P)$	$\Phi(r_C)$	$\Phi(r_M)$	Prepivoting
0.010	0.000	0.012	0.009	0.011
0.025	0.000	0.027	0.021	0.029
0.050	0.000	0.053	0.043	0.056
0.100	0.000	0.102	0.086	0.100
0.250	0.000	0.261	0.227	0.261
0.500	0.000	0.508	0.464	0.510
0.750	0.000	0.752	0.726	0.752
0.900	0.001	0.890	0.871	0.890
0.950	0.001	0.941	0.926	0.943
0.975	0.001	0.967	0.961	0.968
0.990	0.004	0.985	0.981	0.981

---

Tabella 3.4. *Coperture empiriche per la simulazione con il modello gamma con  $q = 500$  strati e  $m = 8$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.*

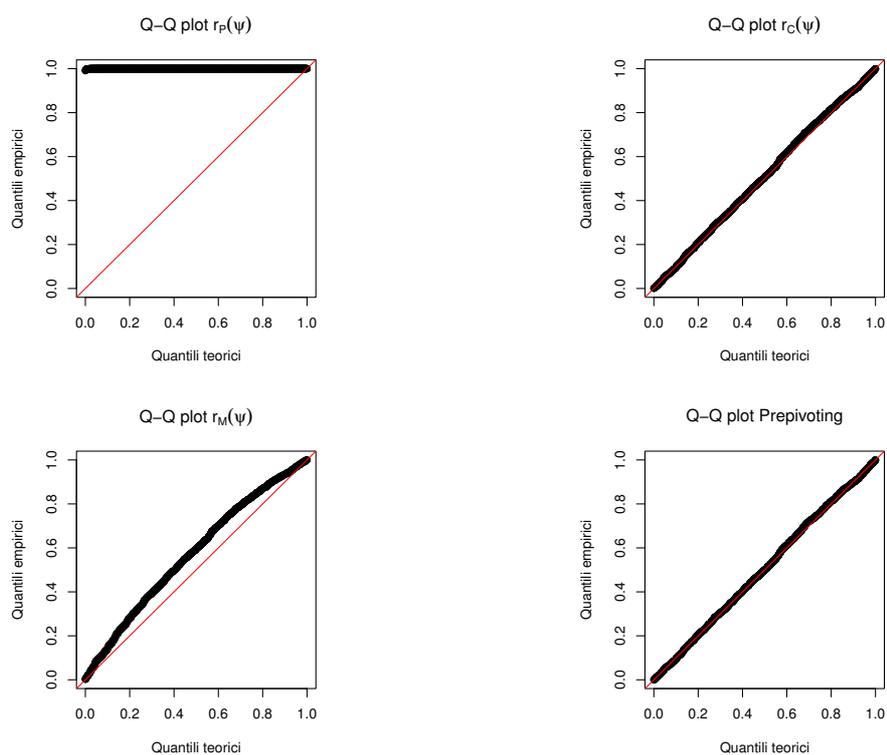


Figura 3.1. Q-Q plot per la simulazione con il modello gamma con  $q = 250$  strati e  $m = 4$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

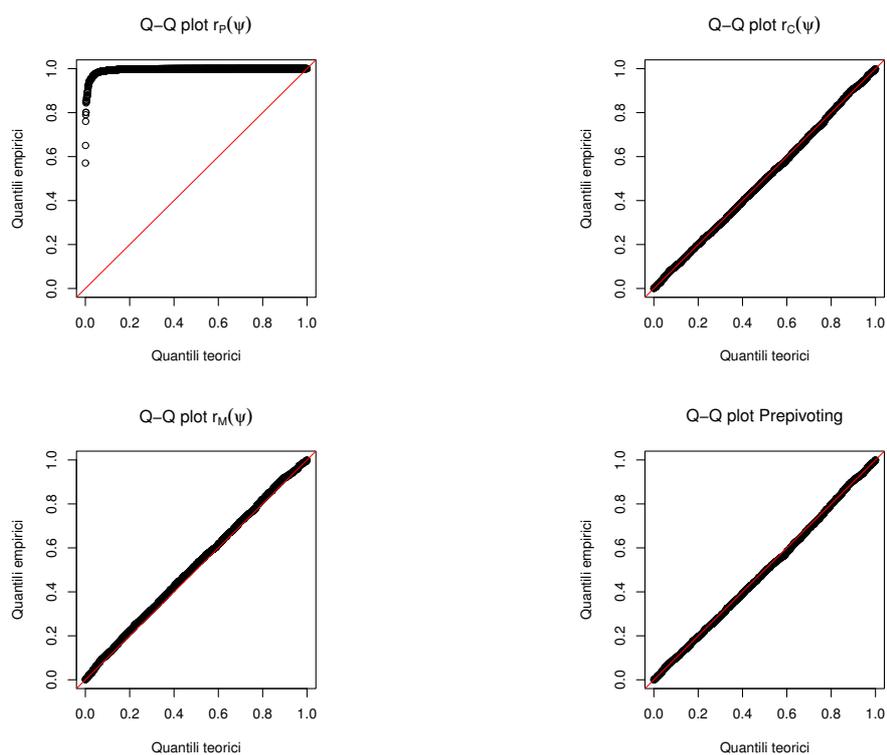


Figura 3.2. Q-Q plot per la simulazione con il modello gamma con  $q = 250$  strati e  $m = 8$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

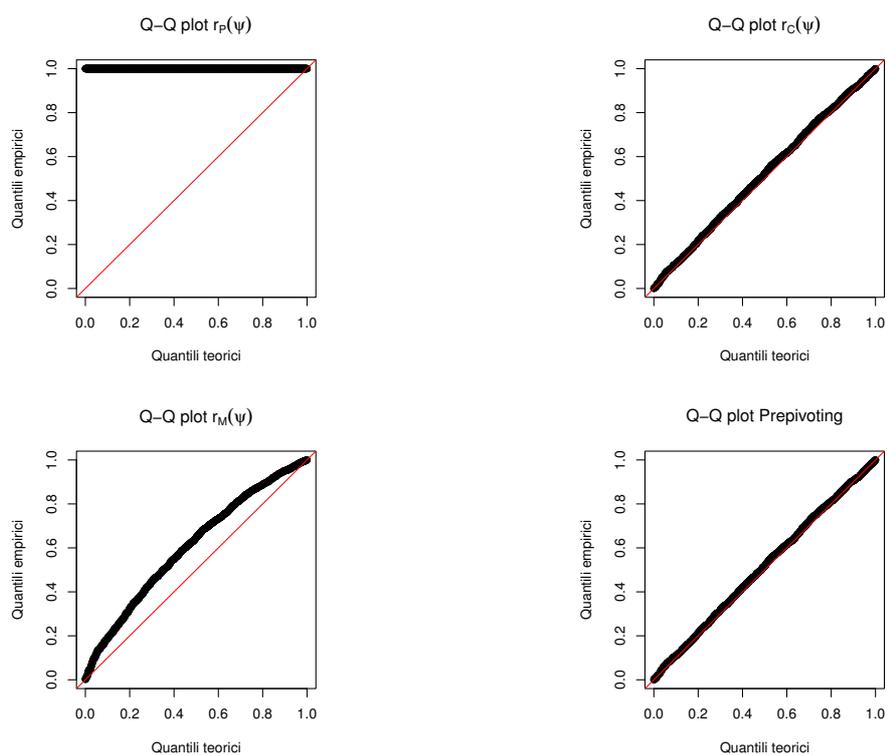


Figura 3.3. Q-Q plot per la simulazione con il modello gamma con  $q = 500$  strati e  $m = 4$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

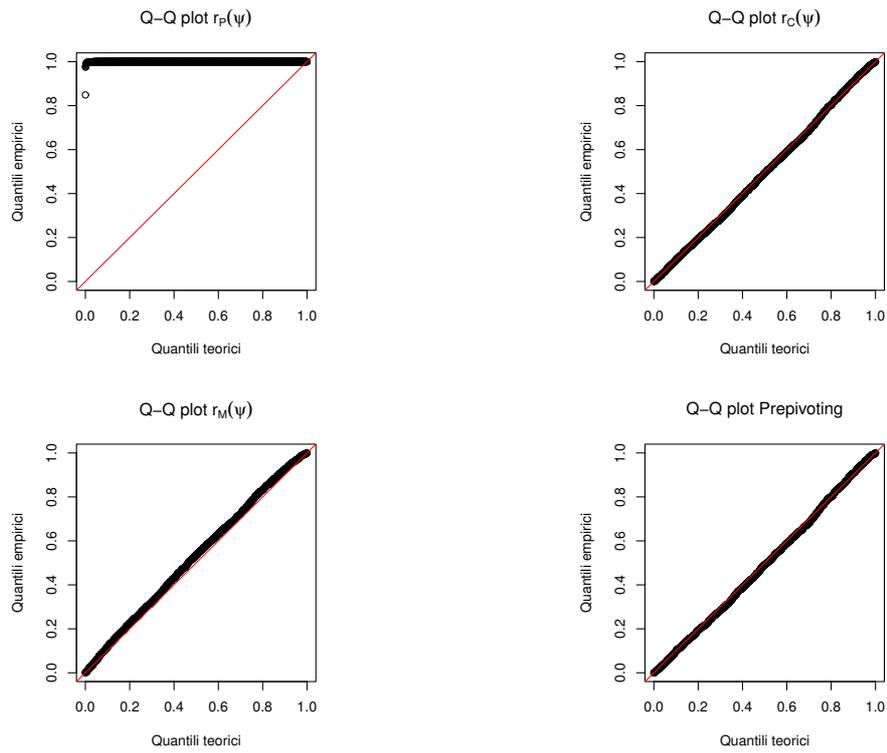


Figura 3.4. Q-Q plot per la simulazione con il modello gamma con  $q = 500$  strati e  $m = 8$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

## Osservazioni

Da tutte le tabelle emerge in modo lampante la totale inaffidabilità degli intervalli di confidenza costruiti a partire dalla log-verosimiglianza profilo. Il leggero miglioramento nella Tabella 3.2 e nella Tabella 3.4 è probabilmente dovuto al numero maggiore di repliche per strato ( $m = 8$ ) che permette di avere a disposizione più informazione sui parametri di disturbo  $\zeta_i$ . I grafici confermano questa tendenza, seppure si rimanga molto lontani dalle coperture nominali.

Come prevedibile, le probabilità di copertura stimate relative alla verosimiglianza profilo modificata sono molto più elevate delle probabilità di copertura empiriche relative alla verosimiglianza profilo. È evidente soprattutto nella differenza di uniformità dei  $p$ -value presentata nei Q-Q plot. Come per la verosimiglianza profilo, si ottengono probabilità di copertura migliori quando si considera un numero maggiore di repliche per strato, anche se generalmente tutti i risultati sono inferiori da quanto indicato nominalmente. Il caso peggiore è quello per  $q = 500$  e  $m = 4$  mostrato nella Tabella 3.3, in cui c'è un difetto di copertura assoluta che oscilla tra il 5 e il 10%. Inoltre, tutti i Q-Q plot mostrano un leggero scostamento sul centro della distribuzione. Le code sembrano più affidabili da questo punto di vista. Fa eccezione il Q-Q plot in Figura 3.2, in cui la condizione sui due indici asintotici sembra essere ancora valida.

Per quanto riguarda il *prepivotting*, si nota che in tutte le tabelle le coperture sono molto simili a quelle ottenute attraverso la statistica  $r_C(\psi)$ , che a loro volta sono essenzialmente pari ai livelli nominali, a meno di un lieve errore di simulazione. Tutti i Q-Q plot mostrano una corrispondenza pressoché perfetta tra i quantili teorici e quelli empirici.

## 3.3 Modello normale

Si considerino le variabili aleatorie  $Y_{ij}$ , indipendenti tra di loro e aventi distribuzione

$$Y_{ij} \sim N(\mu, \sigma_i^2),$$

per  $i = 1, \dots, q$  e  $j = 1, \dots, m$ . Il presente esempio estende quanto mostrato nell'Esempio 7.15 di Severini (2000), in cui le dimensioni erano state poste pari

a  $q = 6$  e  $m = 4$  considerando solo la verosimiglianza profilo e la verosimiglianza profilo modificata, e nell'Esempio 2 di Davison *et al.* (2003), in cui si mostrava anche il procedimento di *prepivoting* per  $r_P(\cdot)$  sempre per  $q$  e  $m$  limitati. In questa tesi si considera un modello con un numero molto maggiore di strati. Qui il parametro di interesse è scalare e pari a  $\psi = \mu$ , ovvero è la media delle distribuzioni, mentre i parametri di disturbo sono rappresentati dalle varianze di ogni strato  $\zeta = \sigma^2 = (\sigma_1^2, \dots, \sigma_q^2)$ . La funzione di densità per la variabile  $Y_{ij}$  è pari a

$$f_{Y_{ij}}(y_{ij}; \mu, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp \left\{ -\frac{1}{2\sigma_i^2} (y_{ij} - \mu)^2 \right\}.$$

Come nell'esempio precedente, si può scrivere la funzione di log-verosimiglianza totale per il modello in funzione delle funzioni di log-verosimiglianza relative ad ogni strato, ovvero

$$l(\mu, \sigma^2) = \sum_{i=1}^q l^i(\mu, \sigma_i^2),$$

con

$$l^i(\mu, \sigma_i^2) = -\frac{m}{2} \log \sigma_i^2 - \frac{1}{2\sigma_i^2} \sum_{j=1}^m (y_{ij} - \mu)^2, \quad (3.6)$$

a rappresentare la log-verosimiglianza dell' $i$ -esimo strato, tralasciando costanti additive.

Per calcolare la log-verosimiglianza profilo di questo modello, si procede calcolando la stima di  $\sigma_i^2$  vincolata alla media  $\mu$ . L'equazione di stima equivale a

$$\frac{\partial}{\partial \sigma_i^2} l^i(\mu, \sigma_i^2) = -\frac{m}{2\sigma_i^2} + \frac{1}{2(\sigma_i^2)^2} \sum_{j=1}^m (y_{ij} - \mu)^2 = 0,$$

ottenendo la soluzione analitica

$$\hat{\sigma}_{i\mu}^2 = \frac{1}{m} \sum_{j=1}^m (y_{ij} - \mu)^2, \quad (3.7)$$

che equivale alla varianza di ogni strato. Si può dimostrare che tale stima è non

distorta per ogni strato e per ogni valore di  $\mu$ . La log-verosimiglianza profilo dell' $i$ -esimo strato sarà quindi

$$l_P^i(\mu) = l^i(\mu, \hat{\sigma}_{i\mu}^2) = -\frac{m}{2} \log \hat{\sigma}_{i\mu}^2, \quad (3.8)$$

e, come sopra, la log-verosimiglianza profilo globale per  $\mu$  è data dalla somma delle varie quantità di ogni strato:  $l_P(\mu) = \sum_{i=1}^q l_P^i(\mu)$ .

Il calcolo del fattore di modificazione per costruire la log-verosimiglianza profilo modificata è stato fatto attraverso l'approssimazione vista in (1.19). Si ottiene

$$j_{\sigma_i^2 \sigma_i^2}(\mu; \hat{\sigma}_{i\mu}^2) = I_{\sigma_i^2 \sigma_i^2}(\hat{\mu}, \hat{\sigma}_i^2; \mu, \hat{\sigma}_{i\mu}^2) = \frac{\mu}{2(\hat{\sigma}_{i\mu}^2)^2},$$

per l' $i$ -esimo strato. Si può quindi calcolare la log-verosimiglianza profilo modificata per l' $i$ -esimo strato

$$\begin{aligned} l_M^i(\mu) &= l_P^i(\mu) + \frac{1}{2} \log j_{\sigma_i^2 \sigma_i^2}(\mu; \hat{\sigma}_{i\mu}^2) - \log I_{\sigma_i^2 \sigma_i^2}(\hat{\mu}, \hat{\sigma}_i^2; \mu, \hat{\sigma}_{i\mu}^2) \\ &= -\frac{m}{2} \log \hat{\sigma}_{i\mu}^2 + \log \hat{\sigma}_{i\mu}^2 \\ &= -\frac{m-2}{2} \log \hat{\sigma}_{i\mu}^2, \end{aligned} \quad (3.9)$$

e ancora una volta la verosimiglianza profilo modificata totale per  $\mu$  si ottiene

$$l_M(\mu) = \sum_{i=1}^q l_M^i(\mu).$$

Per le simulazioni si è posto che il vero valore della media è pari a  $\mu_0 = 1$ , mentre le varianze di ogni strato sono state generate, con un intervallo ampio, da una variabile casuale uniforme  $\mathcal{U}(0,100)$ . All'interno di ogni ciclo di simulazione sono stati poi generati i  $B$  dataset bootstrap, generando i dati dalla normale di parametri  $(\mu_0, \hat{\sigma}_{i\mu_0})$ , con riferimento all'ipotesi nulla  $H_0 : \mu = \mu_0$ .

Le probabilità di copertura empirica degli intervalli di confidenza per  $r_P(\mu)$ ,  $r_M(\mu)$  e *prepivoting* sono presentati nelle Tabelle 3.5-3.8, mentre i Q-Q plot per le varie simulazioni di questo modello sono presentati nelle Figure 3.5-3.8.

---

Nominale	$\Phi(r_P)$	$\Phi(r_M)$	Prepivoting
0.010	0.041	0.006	0.011
0.025	0.076	0.020	0.027
0.050	0.108	0.041	0.045
0.100	0.168	0.088	0.098
0.250	0.302	0.238	0.240
0.500	0.479	0.479	0.486
0.750	0.677	0.739	0.753
0.900	0.811	0.888	0.884
0.950	0.867	0.950	0.941
0.975	0.910	0.982	0.975
0.990	0.950	0.994	0.991

---

Tabella 3.5. Coperture empiriche per la simulazione con il modello normale con  $q = 250$  strati e  $m = 4$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

---

Nominale	$\Phi(r_P)$	$\Phi(r_M)$	Prepivoting
0.010	0.026	0.014	0.014
0.025	0.044	0.028	0.029
0.050	0.073	0.048	0.045
0.100	0.129	0.094	0.096
0.250	0.262	0.234	0.240
0.500	0.482	0.482	0.473
0.750	0.712	0.748	0.743
0.900	0.865	0.895	0.889
0.950	0.918	0.949	0.946
0.975	0.953	0.973	0.973
0.990	0.975	0.989	0.988

---

Tabella 3.6. Coperture empiriche per la simulazione con il modello normale con  $q = 250$  strati e  $m = 8$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

---

Nominale	$\Phi(r_P)$	$\Phi(r_M)$	Prepivoting
0.010	0.047	0.012	0.012
0.025	0.087	0.028	0.027
0.050	0.130	0.047	0.052
0.100	0.188	0.108	0.112
0.250	0.315	0.259	0.259
0.500	0.488	0.488	0.486
0.750	0.666	0.743	0.737
0.900	0.811	0.898	0.895
0.950	0.872	0.951	0.947
0.975	0.917	0.980	0.979
0.990	0.951	0.993	0.992

---

Tabella 3.7. Coperture empiriche per la simulazione con il modello normale con  $q = 500$  strati e  $m = 4$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

---

Nominale	$\Phi(r_P)$	$\Phi(r_M)$	Prepivoting
0.010	0.025	0.012	0.013
0.025	0.043	0.026	0.026
0.050	0.084	0.050	0.049
0.100	0.130	0.099	0.102
0.250	0.264	0.235	0.241
0.500	0.495	0.495	0.498
0.750	0.716	0.748	0.746
0.900	0.852	0.888	0.888
0.950	0.914	0.941	0.940
0.975	0.947	0.969	0.969
0.990	0.973	0.983	0.984

---

Tabella 3.8. Coperture empiriche per la simulazione con il modello normale con  $q = 500$  strati e  $m = 8$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

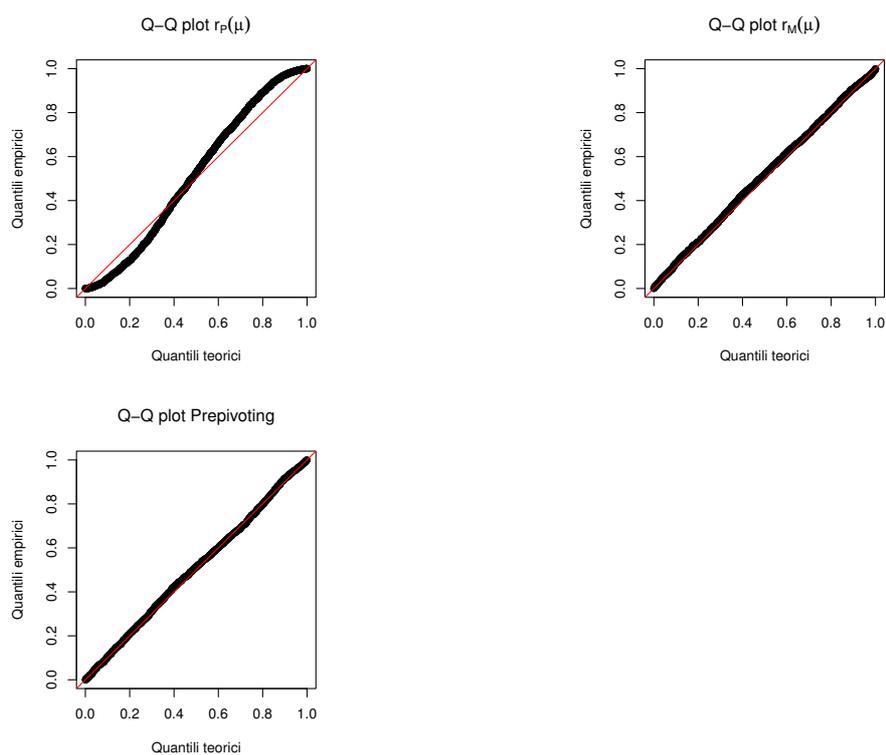


Figura 3.5. Q-Q plot per la simulazione con il modello normale con  $q = 250$  strati e  $m = 4$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

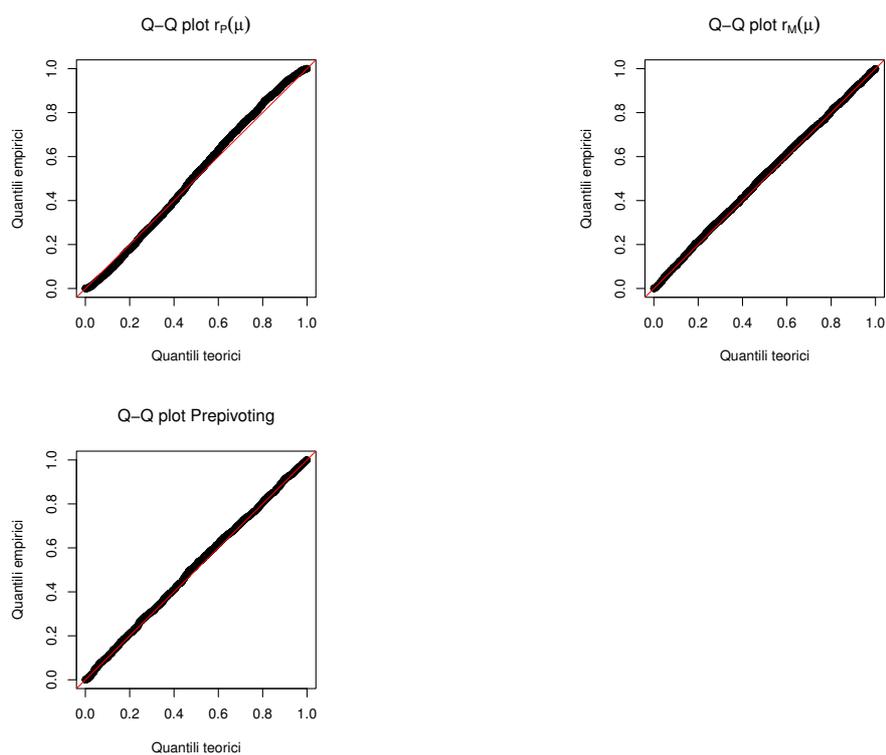


Figura 3.6. Q-Q plot per la simulazione con il modello normale con  $q = 250$  strati e  $m = 8$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

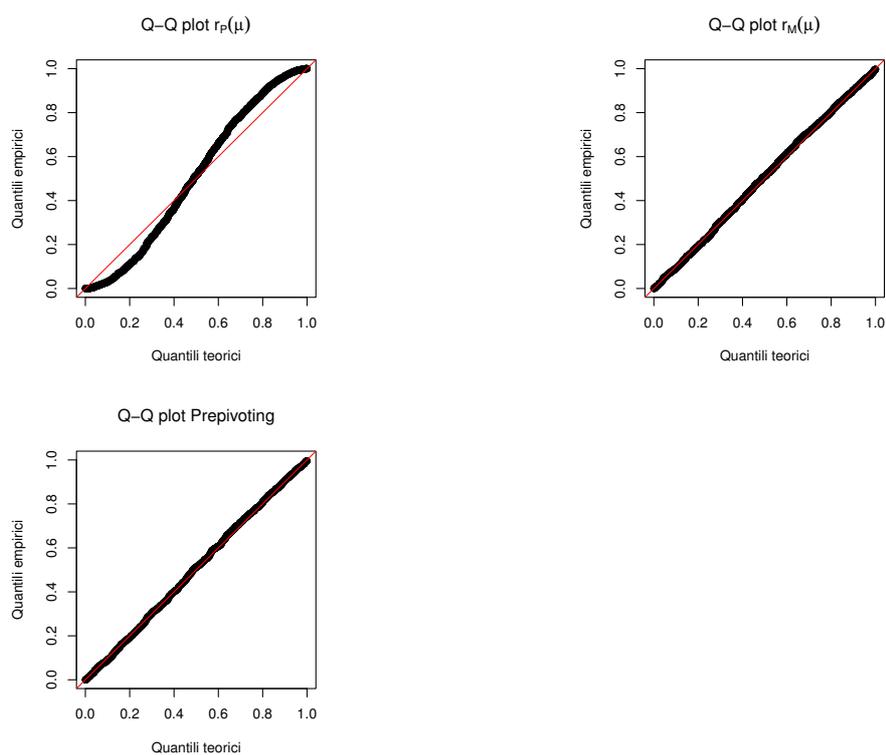


Figura 3.7. Q-Q plot per la simulazione con il modello normale con  $q = 500$  strati e  $m = 4$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

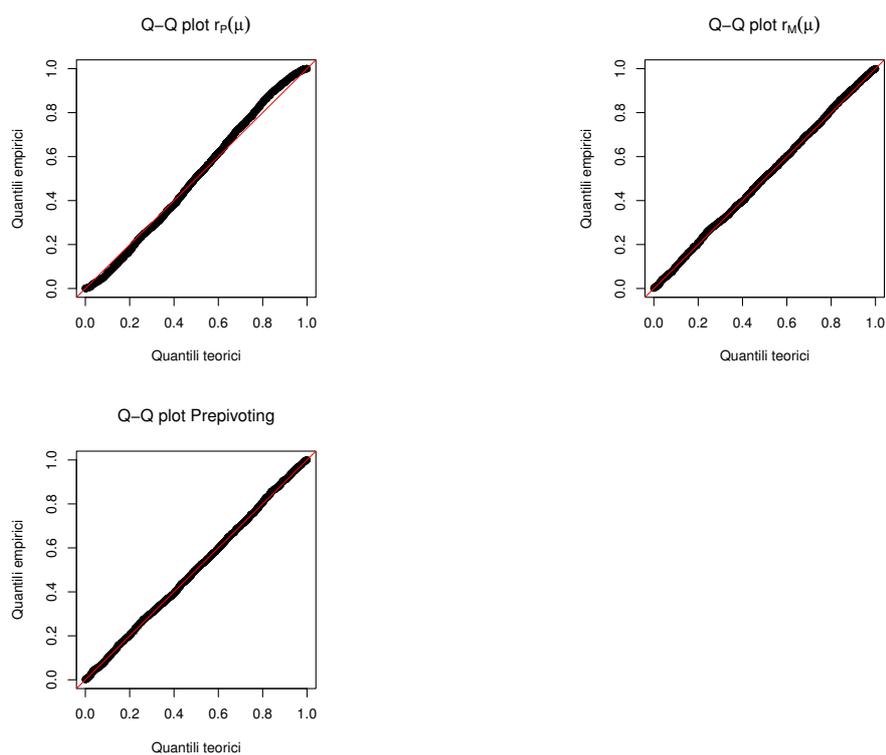


Figura 3.8. Q-Q plot per la simulazione con il modello normale con  $q = 500$  strati e  $m = 8$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

## Osservazioni

Dalle tabelle e dai grafici è evidente come anche in questo esempio le coperture basate su  $r_P(\mu)$  sono insoddisfacenti anche se in modo meno marcato di quanto visto nell'esempio precedente. Si nota però un livello di copertura molto vicino al livello nominale quando  $1 - \alpha = 0.5$ . Questo è dovuto alla non distorsione della score della log-verosimiglianza profilo, che centra la distribuzione della statistica test sullo zero. Il problema, in questo particolare esempio, riguarda l'ampiezza della funzione di verosimiglianza profilo per  $\mu$  che è troppo concentrata sulla stima  $\hat{\mu}$  e che quindi porta a costruire, generalmente, intervalli di confidenza con copertura inferiore a quanto desiderato.

Riguardo le probabilità di copertura empiriche basate su  $r_M(\mu)$  e sul *pre-pivoted root* non si colgono particolari differenze. Si ottengono dei valori vicini ai valori nominali con entrambi i metodi. Nelle simulazioni in cui si è posto  $m = 4$  si nota comunque un leggero peggioramento nelle code di  $r_M(\mu)$  rispetto ai risultati di *pre-pivoting*, visibile nella Tabella 3.5 e nella Tabella 3.7, anche se tale diversità è minore dell'errore di simulazione. Vale infatti il discorso fatto per la verosimiglianza profilo: la distribuzione della statistica  $r_M(\mu)$  è centrata sullo zero, ma oltre a questo è presente un miglioramento dovuto ad una funzione di verosimiglianza con ampiezza maggiore per la verosimiglianza profilo modificata rispetto alla versione più concentrata della verosimiglianza profilo. Di conseguenza gli intervalli di confidenza costruiti a partire dalla verosimiglianza profilo modificata offrono una copertura maggiore, come effettivamente è stato ottenuto dalle simulazioni.

Nel complesso il metodo basato sul *pre-pivoted root* sembra funzionare indipendentemente dalla dimensione degli strati del modello e dal numero di repliche per ogni strato. Si vede sia nelle tabelle, come già affermato, sia sui Q-Q plot, in cui i *p-value* bootstrap offrono un buon livello di uniformità. In questo caso, rispetto alla simulazione vista nel Paragrafo 3.2, non sembra esserci un sostanziale miglioramento rispetto alle probabilità di copertura empiriche relative alla verosimiglianza profilo modificata, rendendo preferibile la scelta di quest'ultimo metodo per via di un minor carico computazionale di calcolo.

### 3.4 Modello per dati binari

Siano  $Y_{ij}$  delle variabili aleatorie indipendenti con distribuzione

$$Y_{ij} \sim Bi(1, \pi_{ij}),$$

per  $i = 1, \dots, q$  e  $j = 1, \dots, m$ , che danno origine al campione stratificato  $y_{ij}$ . Si supponga di essere interessati a formulare un modello che esprima la probabilità di successo  $\pi_{ij} = \mathbb{P}[Y_{ij} = 1]$ , in funzione di una variabile esplicativa. Ponendo

$$g(\pi_{ij}) = \eta_{ij} = \alpha_i + \beta x_{ij},$$

con  $\alpha_i, \beta \in \mathbb{R}$  e  $g(\cdot)$  funzione legame, si ottiene un modello lineare generalizzato per dati stratificati discreti. L'intercetta  $\alpha_i$  è specifica di ogni strato, mentre il coefficiente angolare  $\beta$  è comune a tutti gli strati. Quindi in questa situazione il parametro  $\psi = \beta$  rappresenta il parametro strutturale del modello, mentre  $\zeta = \alpha = (\alpha_1, \dots, \alpha_q)$  è la componente di disturbo.

Per basare l'inferenza solo sul parametro di interesse  $\beta$ , si può ricorrere alla verosimiglianza condizionata, ma solo nel caso in cui la funzione legame  $g(\cdot)$  sia quella canonica. In tal caso lo stimatore di  $\beta$  per la verosimiglianza condizionata ha proprietà migliori rispetto allo stimatore  $\hat{\beta}$  non condizionato, soprattutto in un modello per dati stratificati, in cui  $q$  ha dimensione maggiore rispetto a  $m$ . Durante la simulazione verrà tralasciato lo studio della verosimiglianza condizionata. Ci si concentrerà sullo studio di altre pseudo-verosimiglianze, quali la verosimiglianza profilo e la verosimiglianza profilo modificata, anche per eventualmente studiare funzioni legame diverse da quelle canoniche. Il confronto tra la verosimiglianza profilo modificata e la verosimiglianza condizionata per il modello logit è già stato studiato in Bartolucci *et al.* (2012), mostrando di fatto una sostanziale uguaglianza nei risultati tra i due metodi.

Si ponga  $\pi = G(\eta)$  con  $G(\cdot)$  funzione di ripartizione assegnata. La funzione di log-verosimiglianza, a meno di costanti additive, è pari a

$$l(\alpha, \beta) = \sum_{i=1}^q l^i(\alpha_i, \beta),$$

dove la verosimiglianza dell' $i$ -esimo strato è pari a

$$l^i(\alpha_i, \beta) = \sum_{j=1}^m \left[ y_{ij} \log G(\eta_{ij}) + (1 - y_{ij}) \log(1 - G(\eta_{ij})) \right]. \quad (3.10)$$

Per calcolare a log-verosimiglianza profilo dell' $i$ -esimo strato, è necessario calcolare la stima di  $\alpha_i$  vincolata a  $\beta$ ,  $\hat{\alpha}_{i\beta}$ . Generalmente non esistono soluzioni analitiche e quindi si ricorre ad algoritmi di ottimizzazione numerica. La verosimiglianza profilo globale per il parametro di interesse  $\beta$  è

$$l_P(\beta) = \sum_{i=1}^q l_P^i(\beta) = \sum_{i=1}^q l^i(\hat{\alpha}_{i\beta}, \beta),$$

con

$$l_P^i(\beta) = \sum_{j=1}^m \left[ y_{ij} \log G(\tilde{\eta}_{ij}) + (1 - y_{ij}) \log(1 - G(\tilde{\eta}_{ij})) \right] \quad (3.11)$$

a rappresentare la log-verosimiglianza profilo per l' $i$ -esimo strato e dove  $\tilde{\eta}_{ij} = \hat{\alpha}_{i\beta} + \beta x_{ij}$ .

La log-verosimiglianza profilo modificata per l' $i$ -esimo strato viene calcolata a partire dalla (3.11) con l'approssimazione (1.19) e avrà forma

$$l_M^i(\beta) = l_P^i(\beta) + M_i'(\beta).$$

Il fattore di modificazione  $M_i'(\beta)$  per l' $i$ -esimo strato è

$$M_i'(\beta) = \frac{1}{2} \log |j_{\alpha_i \alpha_i}(\hat{\alpha}_{i\beta}, \beta)| - \log |I_{\alpha_i \alpha_i}^i(\hat{\alpha}_{i\beta}, \beta; \hat{\alpha}_i, \hat{\beta})|, \quad (3.12)$$

ottenendo il fattore di modificazione globale con  $M'(\beta) = \sum_{i=1}^q M_i'(\beta)$ . Nella (3.12) la varie quantità sono pari a

$$I_{\alpha_i \alpha_i}^i(\hat{\alpha}_{i\beta}, \beta; \hat{\alpha}_i, \hat{\beta}) = \sum_{j=1}^m \frac{G'(\hat{\eta}_{ij})G'(\tilde{\eta}_{ij})}{\{1 - G(\tilde{\eta}_{ij})\}G(\tilde{\eta}_{ij})},$$

$$j_{\alpha_i \alpha_i}(\hat{\alpha}_{i\beta}, \beta) = I_{\alpha; \alpha}^i(\hat{\alpha}_{i\beta}, \beta; \hat{\alpha}_{i\beta}, \beta) - \sum_{j=1}^m \{y_{ij} - G(\tilde{\eta}_{ij})\} c_{ij},$$

$$c_{ij} = \left[ \frac{G''(\tilde{\eta}_{ij})}{\{1 - G(\tilde{\eta}_{ij})\}G(\tilde{\eta}_{ij})} - \frac{[G'(\tilde{\eta}_{ij})]^2 \{1 - 2G(\tilde{\eta}_{ij})\}}{\{1 - G(\tilde{\eta}_{ij})\}^2 [G(\tilde{\eta}_{ij})]^2} \right],$$

con  $\hat{\eta}_{ij} = \hat{\alpha}_i + \hat{\beta}x_{ij}$  predittore lineare calcolato nelle stime di massima verosimiglianza e  $G'(\cdot)$  e  $G''(\cdot)$  rispettivamente derivate di ordine uno e due di  $G(\cdot)$ . Una volta calcolata la log-verosimiglianza profilo modificata globale  $l_M(\beta) = \sum_{i=1}^q l_M^i(\beta)$  si procederà al calcolo della stima di massima verosimiglianza modificata  $\hat{\beta}_M$  tramite massimizzazione numerica (Bartolucci *et al.*, 2012).

Gli studi di simulazione effettuati in questo contesto riguardano due diverse funzioni legame. Si studierà in prima istanza la funzione legame canonica per questo modello, ovvero la funzione logit. Successivamente si considererà la funzione legame probit. Come nei precedenti studi di simulazione, si vogliono calcolare le coperture per degli intervalli di confidenza su  $\beta$  basati sulle statistiche test  $r_P(\beta)$  e  $r_M(\beta)$ . Inoltre si vedranno anche le coperture tramite il procedimento di *prepivotting* per  $r_P(\beta)$ , impostando un bootstrap parametrico con schema modificato, ovvero generando i campioni bootstrap da  $(\hat{\alpha}_{i\beta}, \beta)$ . Per entrambi gli studi ci si è avvalsi dell'aiuto del pacchetto R `panelMPL` (Bellio e Sartori, 2012), che, tra i vari modelli che può trattare, permette di calcolare le verosimiglianze e le s.m.v. nel modello logit e nel modello probit.

### 3.4.1 Modello logit

Siano  $Y_{ij}$  delle variabili casuali indipendenti con distribuzione

$$Y_{ij} \sim Bi(1, \pi_{ij})$$

con funzione legame logistica, ovvero tale che

$$g(\pi_{ij}) = \log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \eta_{ij} = \alpha + \beta x_{ij}, \quad (3.13)$$

con inversa

$$\pi_{ij} = G(\eta_{ij}) = \frac{e^{\eta_{ij}}}{1 + e^{\eta_{ij}}}. \quad (3.14)$$

Tale modello è allora detto modello di regressione logistica o modello logit.

Visto che il modello per dati discreti basato sulla funzione legame logistica è il modello con funzione legame canonica, sono disponibili alcune semplificazioni nella specificazione del modello, in particolare per la verosimiglianza profilo modificata.

Si consideri la probabilità di successo associata alla variabile  $Y_{ij}$  definita alla (3.13). Si può calcolare la derivata prima rispetto a  $\eta_{ij}$ , ottenendo la funzione di densità

$$G'(\eta_{ij}) = \frac{e^{\eta_{ij}}}{(1 + e^{\eta_{ij}})^2},$$

di cui successivamente si può calcolarne a sua volta la derivata, ottenendo

$$G''(\eta_{ij}) = \frac{e^{\eta_{ij}}(1 - e^{\eta_{ij}})}{(1 + e^{\eta_{ij}})^3} = \frac{1 - e^{\eta_{ij}}}{1 + e^{\eta_{ij}}} G'(\eta_{ij}).$$

Impiegando questi risultati nelle quantità della (3.12) si può notare che il termine  $I_{\alpha;\alpha}^i(\hat{\alpha}_{i\beta}, \beta; \hat{\alpha}_i, \hat{\beta})$  non dipende da  $\tilde{\eta}_{ij}$ , rendendolo trascurabile ai fini dell'inferenza su  $\beta$ , mentre la quantità  $c_{ij}$  si annulla. Si può quindi scrivere la log-verosimiglianza profilo modificata nella versione modificata

$$l_M(\beta) = l_P(\beta) + \frac{1}{2} \sum_{i=1}^q \log |j_{\alpha_i \alpha_i}(\hat{\alpha}_{i\beta}, \beta)|, \quad (3.15)$$

con

$$j_{\alpha_i \alpha_i}(\hat{\alpha}_{i\beta}, \beta) = I_{\alpha;\alpha}^i(\hat{\alpha}_{i\beta}, \beta; \hat{\alpha}_{i\beta}, \beta).$$

Nelle simulazioni si è posto  $\beta_0 = 1$ , mentre i valori delle variabili esplicative  $x_{ij}$ , con  $i = 1, \dots, q$  e  $j = 1, \dots, m$  sono stati generati casualmente da una normale  $N(0,1)$ . I valori delle intercette  $\alpha_{i0}$  sono stati invece generati da una normale  $N(\bar{x}_i, 1)$ , dove con  $\bar{x}_i$  si è indicata la media dei valori delle  $x_{ij}$  nello strato  $i$ -esimo. Per ognuna delle  $N$  iterazioni nella simulazione, si sono generati i valori  $y_{ij}$  tenendo fissati sia le  $x_{ij}$  che gli  $\alpha_{i0}$ . Infine, si sono generati i campioni bootstrap da  $(\hat{\alpha}_{i\beta_0}, \beta_0)$ , riferendosi all'ipotesi nulla  $H_0 : \beta = \beta_0$ .

I risultati delle simulazioni consistono nei valori delle probabilità di copertura empiriche mostrati nelle Tabelle 3.9 - 3.12 e nei Q-Q plot nelle Figure 3.9 - 3.12.

---

Nominale	$\Phi(r_P)$	$\Phi(r_M)$	Prepivoting
0.010	0.000	0.002	0.004
0.025	0.000	0.006	0.013
0.050	0.000	0.017	0.029
0.100	0.000	0.041	0.064
0.250	0.001	0.116	0.164
0.500	0.004	0.302	0.386
0.750	0.020	0.565	0.654
0.900	0.067	0.770	0.843
0.950	0.107	0.864	0.910
0.975	0.166	0.921	0.948
0.990	0.252	0.958	0.979

---

Tabella 3.9. Coperture empiriche per la simulazione con il modello logit con  $q = 250$  strati e  $m = 4$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

---

Nominale	$\Phi(r_P)$	$\Phi(r_M)$	Prepivoting
0.010	0.000	0.007	0.009
0.025	0.001	0.018	0.024
0.050	0.001	0.034	0.043
0.100	0.001	0.076	0.093
0.250	0.004	0.189	0.221
0.500	0.021	0.411	0.458
0.750	0.081	0.677	0.726
0.900	0.184	0.866	0.891
0.950	0.275	0.927	0.944
0.975	0.386	0.964	0.972
0.990	0.523	0.985	0.990

---

Tabella 3.10. Coperture empiriche per la simulazione con il modello logit con  $q = 250$  strati e  $m = 8$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

Nominale	$\Phi(r_P)$	$\Phi(r_M)$	Prepivoting
0.010	0.000	0.002	0.004
0.025	0.000	0.004	0.008
0.050	0.000	0.008	0.015
0.100	0.000	0.021	0.046
0.250	0.000	0.073	0.133
0.500	0.000	0.217	0.326
0.750	0.001	0.447	0.579
0.900	0.004	0.679	0.802
0.950	0.009	0.805	0.888
0.975	0.016	0.884	0.937
0.990	0.042	0.940	0.971

Tabella 3.11. Coperture empiriche per la simulazione con il modello logit con  $q = 500$  strati e  $m = 4$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

Nominale	$\Phi(r_P)$	$\Phi(r_M)$	Prepivoting
0.010	0.000	0.006	0.007
0.025	0.000	0.013	0.024
0.050	0.000	0.029	0.036
0.100	0.000	0.059	0.084
0.250	0.000	0.169	0.214
0.500	0.003	0.386	0.453
0.750	0.010	0.625	0.693
0.900	0.037	0.809	0.854
0.950	0.081	0.890	0.919
0.975	0.131	0.937	0.957
0.990	0.217	0.973	0.984

Tabella 3.12. Coperture empiriche per la simulazione con il modello logit con  $q = 500$  strati e  $m = 8$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

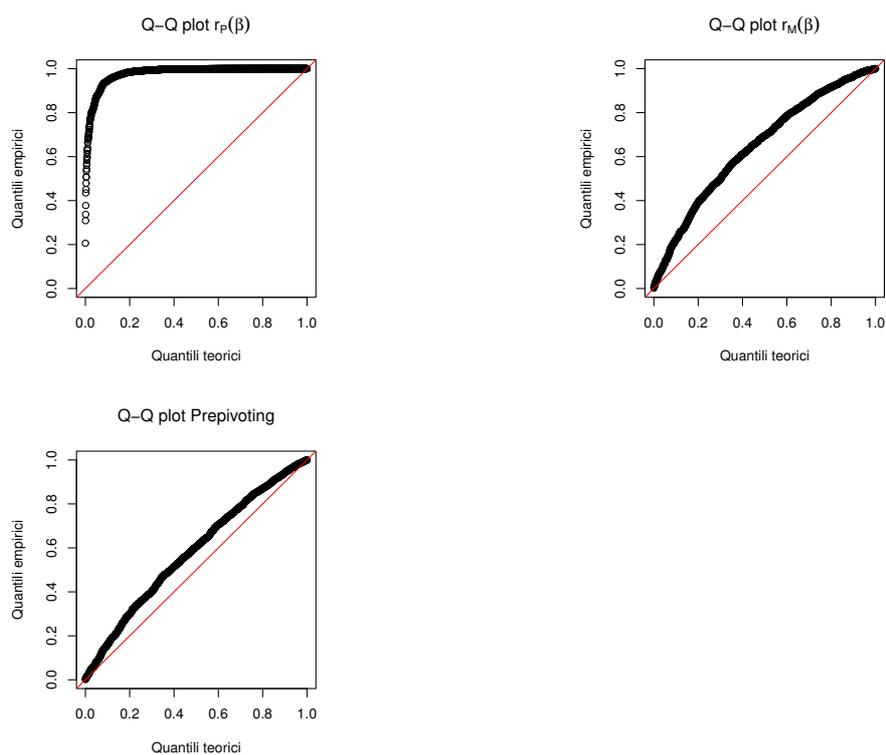


Figura 3.9. Q-Q plot per la simulazione con il modello logit con  $q = 250$  strati e  $m = 4$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

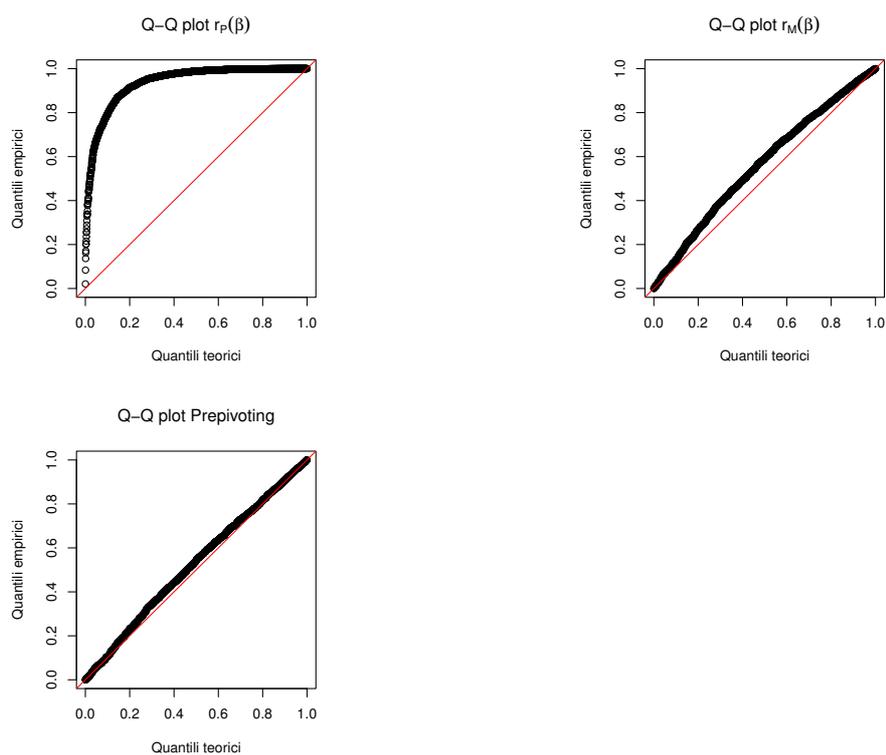


Figura 3.10. Q-Q plot per la simulazione con il modello logit con  $q = 250$  strati e  $m = 8$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

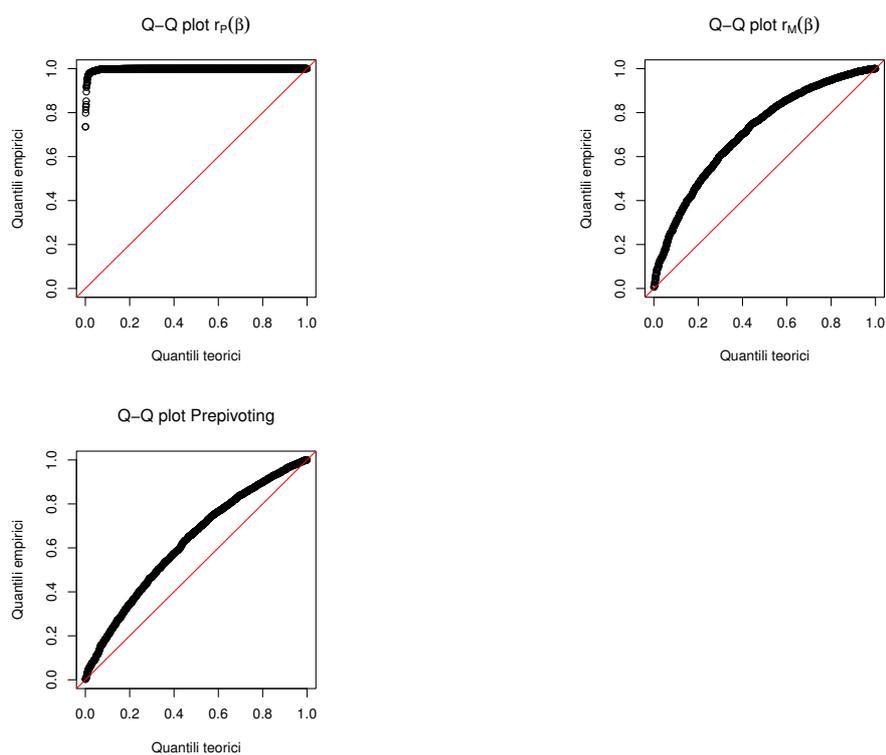


Figura 3.11. Q-Q plot per la simulazione con il modello logit con  $q = 500$  strati e  $m = 4$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

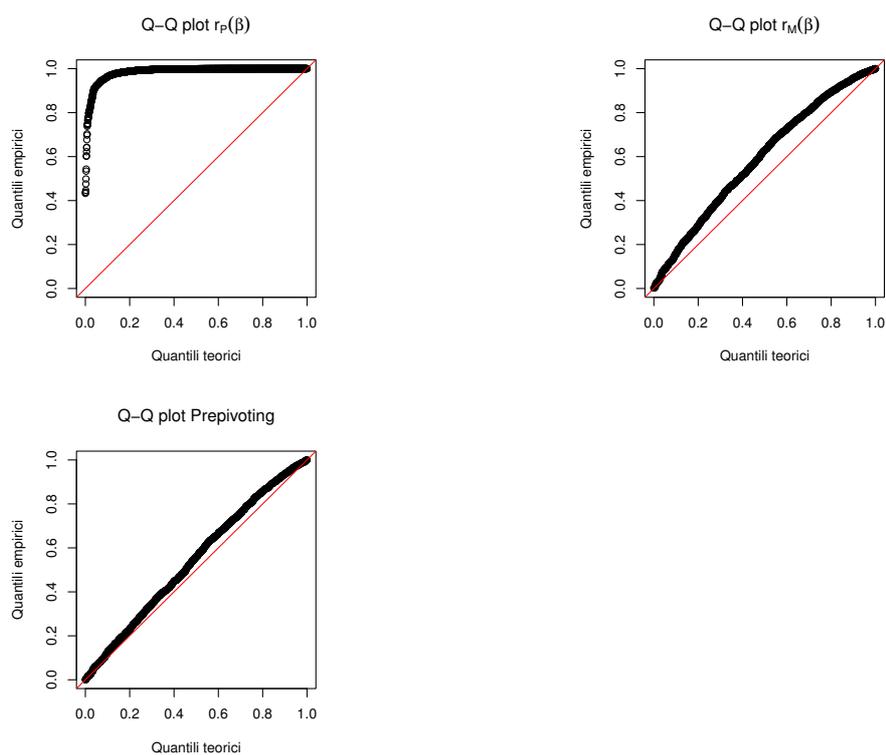


Figura 3.12. Q-Q plot per la simulazione con il modello logit con  $q = 500$  strati e  $m = 8$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

## Osservazioni

Le probabilità di copertura empirica relative alla verosimiglianza profilo si confermano ancora una volta altamente insoddisfacenti. In tutte le simulazioni effettuate con il modello logit, i  $p$  – *value* sono totalmente non uniformi.

Tuttavia, in questa serie di simulazioni anche le coperture empiriche ottenute a partire dalla log-verosimiglianza profilo modificata e dal *prepivotated root* presentano un sensibile difetto di copertura rispetto a quanto indicato nei livelli nominali.

Come ci si aspetta, avere un maggior numero di repliche per strato porta una maggiore quantità di informazione per stimare l'intercetta di ogni strato. Questo si riflette generalmente sulle probabilità di copertura e sui Q-Q plot, con un vistoso miglioramento sia sugli intervalli costruiti con  $r_M(\beta)$  che su quelli di *prepivotating*. Si vedano in particolare le differenze tra Figura 3.9 e Figura 3.10 e tra Figura 3.11 e Figura 3.12 per verosimiglianza profilo modificata e *prepivotating*.

Ciò nonostante, il metodo basato sul bootstrap sembra essere più efficace della log-verosimiglianza profilo modificata in ogni singola situazione. Nella situazione più critica per la verosimiglianza profilo modificata, ossia con  $q = 500$  e  $m = 4$ , ci sono picchi di miglioramento che raggiungono il 13% per le probabilità assolute di copertura empirica rispetto a quelle basate su  $r_M(\beta)$ , come si può vedere nella Tabella 3.11. Il miglioramento è più evidente nelle zone centrali rispetto alle code.

### 3.4.2 Modello probit

Rispetto al modello logistico impostato nel Paragrafo 3.4.1, ora si concentra l'analisi sul modello probit. Siano ancora  $Y_{ij}$  delle variabili casuali indipendenti con distribuzione

$$Y_{ij} \sim Bi(1, \pi_{ij}),$$

ma questa volta la funzione legame è la funzione legame probit, ovvero

$$g(\pi_{ij}) = \Phi^{-1}(\pi_{ij}) = \eta_{ij} = \alpha_i + \beta x_{ij}, \quad (3.16)$$

con probabilità di successo

$$\pi_{ij} = \Phi(\eta_{ij}), \quad (3.17)$$

dove  $\Phi(\cdot)$  è la funzione di ripartizione della normale standard. Si sono quindi usate le formule di pagina 63 con  $G(\cdot) = \Phi(\cdot)$ .

Come nelle simulazioni effettuate nel paragrafo precedente, si è posto il parametro di interesse  $\beta_0 = 1$  e la costruzione delle  $x_{ij}$  e degli  $\alpha_{i0}$  e la generazione delle  $y_{ij}$  è avvenuta allo stesso modo. I campioni bootstrap sono stati generati ancora una volta riferendosi all'ipotesi nulla  $H_0 : \beta = \beta_0$ , ovvero da  $(\hat{\alpha}_{i\beta_0}, \beta_0)$ . L'unica variante riguarda la funzione legame.

A causa dei lunghi tempi di calcolo, si sono effettuate le simulazioni solo per  $q = 250$  e  $m = 4, 8$ . I risultati delle simulazioni sono presentati in Tabella 3.13 e in Tabella 3.14 per quanto riguarda le coperture, mentre in Figura 3.13 e in Figura 3.14 per quanto riguarda i Q-Q plot.

---

Nominale	$\Phi(r_P)$	$\Phi(r_M)$	Prepivoting
0.010	0.000	0.002	0.002
0.025	0.000	0.004	0.004
0.050	0.000	0.015	0.011
0.100	0.000	0.041	0.035
0.250	0.000	0.154	0.124
0.500	0.000	0.381	0.321
0.750	0.000	0.674	0.579
0.900	0.002	0.875	0.784
0.950	0.003	0.943	0.874
0.975	0.007	0.977	0.932
0.990	0.021	0.993	0.966

---

Tabella 3.13. Coperture empiriche per la simulazione con il modello probit con  $q = 250$  strati e  $m = 4$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

---

Nominale	$\Phi(r_P)$	$\Phi(r_M)$	Prepivoting
0.010	0.000	0.003	0.005
0.025	0.000	0.009	0.013
0.050	0.000	0.022	0.033
0.100	0.000	0.051	0.064
0.250	0.000	0.160	0.188
0.500	0.001	0.386	0.429
0.750	0.004	0.650	0.686
0.900	0.019	0.851	0.872
0.950	0.042	0.924	0.935
0.975	0.076	0.956	0.965
0.990	0.137	0.983	0.985

---

Tabella 3.14. Coperture empiriche empiriche per la simulazione con il modello probit con  $q = 250$  strati e  $m = 8$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

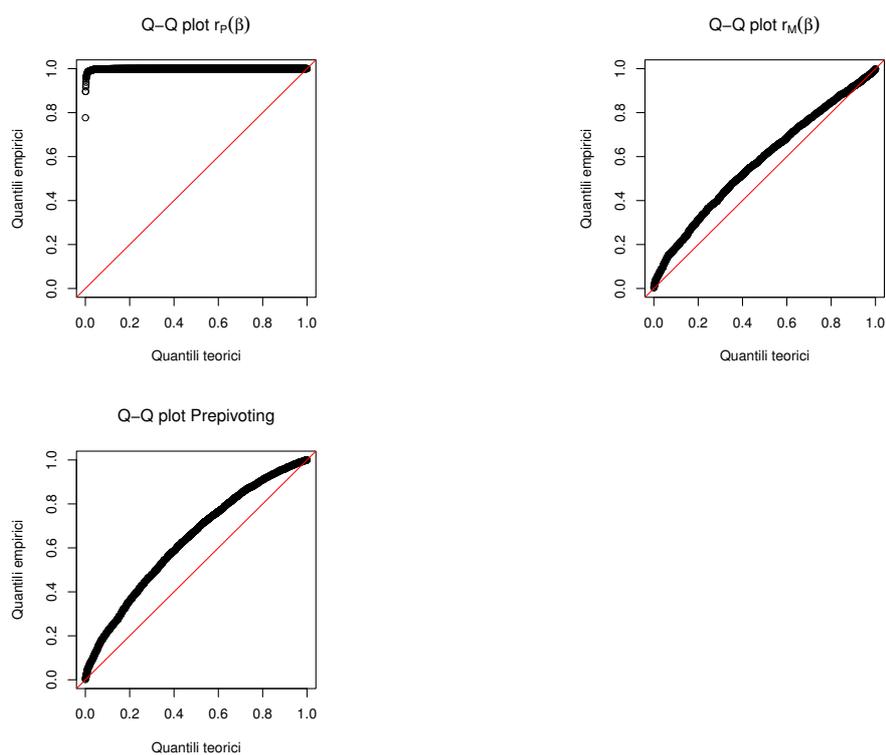


Figura 3.13. Q-Q plot per la simulazione con il modello probit con  $q = 250$  strati e  $m = 4$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

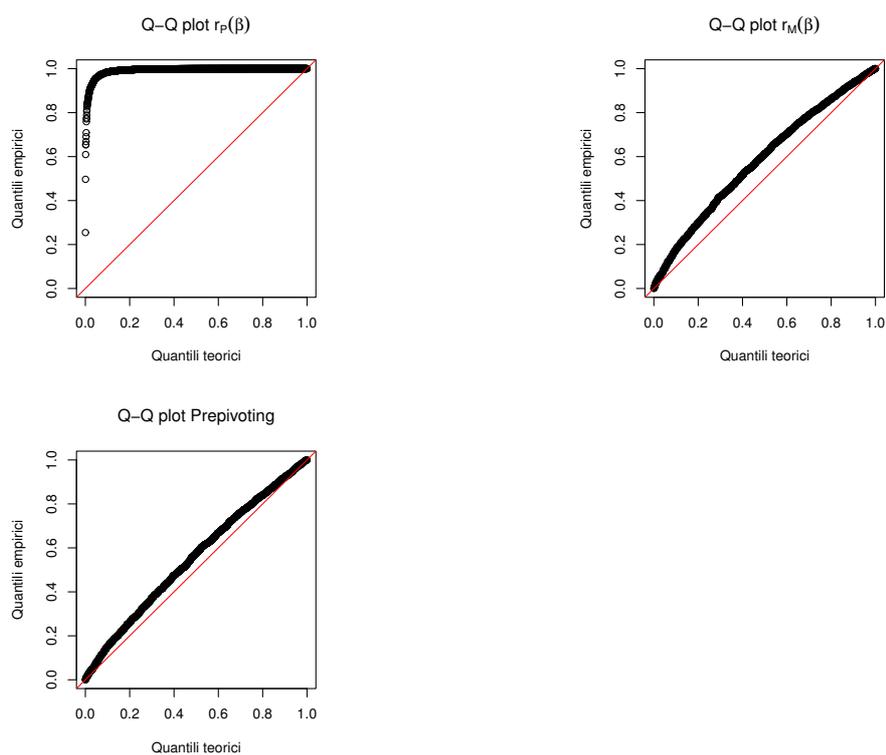


Figura 3.14. Q-Q plot per la simulazione con il modello probit con  $q = 250$  strati e  $m = 8$  repliche per strato. I risultati sono stati ottenuti attraverso  $N = 2000$  iterazioni e  $B = 2000$  repliche bootstrap.

## Osservazioni

Si confermano ancora i problemi della log-verosimiglianza profilo. Nelle situazioni proposte nelle simulazioni c'è troppa poca informazione per avere degli intervalli di confidenza affidabili.

Dalle tabelle e dai grafici risulta una sostanziale uguaglianza nelle simulazioni della log-verosimiglianza profilo modificata. Gli intervalli basati su  $r_M(\beta)$  forniscono gli stessi livelli di copertura sia per  $m = 4$  che per  $m = 8$ , come si vede dalla Tabella 3.13 e dalla Tabella 3.14. Si nota invece un miglioramento nelle coperture degli intervalli basati sul *prepivotated root*, soprattutto nella parte centrale della distribuzione in Figura 3.14.

In generale, il comportamento in questa simulazione va in controtendenza rispetto a quanto visto nei precedenti esempi, soprattutto per quanto riguarda gli intervalli basati su  $r_M(\beta)$ . In tutti gli esempi c'è stato un miglioramento nelle coperture all'aumentare del numero di repliche per strato  $m$ , mentre qui i risultati sono pressoché identici. Questo ha provocato delle coperture peggiori negli intervalli di *prepivotating* rispetto alla log-verosimiglianza profilo modificata per  $m = 4$ , ma delle coperture migliori dei primi rispetto ai secondi quando  $m = 8$ .



# Conclusioni

Nel corso della tesi sono stati proposti diversi metodi per fare inferenza nel contesto dei problemi di Neyman e Scott in un contesto a due indici asintotici. La verosimiglianza profilo fornisce dei risultati altamente inattendibili e, in situazioni estreme con molti strati di bassa numerosità, anche la verosimiglianza profilo modificata tende a presentare un'accuratezza poco soddisfacente. Si è proposto quindi un metodo basato sulla simulazione bootstrap, tramite alcuni aggiustamenti che potessero adattarsi al contesto con parametri di disturbo. Tali aggiustamenti riguardano uno schema modificato per il bootstrap e il procedimento di *prepivoting*.

L'obiettivo della tesi era quello di analizzare la bontà dell'approccio bootstrap così specificato per verificare le probabilità di copertura empiriche in diversi modelli per dati stratificati tramite studi di simulazione. Nel Capitolo 3 sono stati confermati i vantaggi già noti della verosimiglianza profilo modificata sulla verosimiglianza profilo. Infatti, in tutti gli esempi si sono ottenute coperture empiriche per gli intervalli di confidenza unilaterali più vicini al livello nominale usando la verosimiglianza profilo modificata. Ciò nonostante, questi risultati sembrano essere stati migliorati dalle probabilità di copertura empirica basati sul procedimento di *prepivoting*, con l'eccezione dell'esempio relativo al modello normale in cui sostanzialmente i risultati si equivalgono.

In particolare, si è visto che la verosimiglianza profilo modificata si rivela affidabile fintanto che viene rispettata la condizione per i due indici asintotici  $1/m = o(q^{-1/3})$ , già descritta da Sartori (2003). Al contrario, i risultati ottenuti basandosi sull'utilizzo di un *prepivoted root* sembrano risentire meno della mancanza di informazione disponibile sugli strati del modello dovuta al basso numero di repliche per ogni strato.

I risultati migliori si sono ottenuti per i campioni gamma, in cui le probabilità di

copertura empirica erano molto vicine ai valori nominali e grazie alla disponibilità di stime esplicite per i parametri incidentali il carico computazionale è risultato poco importante ai fini delle analisi. Lo stesso non si può dire per i modelli per dati discreti, in cui era necessario ottenere le stime dei parametri di disturbo attraverso algoritmi numerici, aumentando di molto lo sforzo computazionale. A tal proposito, il vantaggio nell'utilizzo di un metodo basato sulla simulazione risiede nelle notevoli potenze di calcolo dei calcolatori, che permettono un maggior numero di iterazioni per ottenere risultati accurati. Per migliorare ulteriormente questo aspetto e accorciare i tempi di attesa, si può pensare di implementare le funzioni usate per le simulazioni nel linguaggio di programmazione **C** piuttosto che in **R**, il quale generalmente è poco prestante sui cicli. Inoltre, vista la natura del bootstrap, un'efficiente utilizzo del calcolo parallelo potrebbe diminuire notevolmente i tempi di calcolo. Il miglioramento in termini di tempo di simulazione sarebbe più evidente per i modelli per dati discreti. Questo ragionamento è giustificato dal miglioramento dei risultati nelle coperture empiriche degli intervalli di confidenza unilaterali. Se così non avvenisse, come nella già accennata sostanziale uguaglianza dei risultati nel modello normale, allora è consigliabile procedere con la verosimiglianza profilo modificata, che risulta indubbiamente più rapida nei calcoli.

Tuttavia, le conclusioni a cui si è giunti in questa tesi sono solamente di carattere empirico e limitate agli esempi del Capitolo 3. Non sono ancora disponibili valutazioni teoriche sul miglioramento effettivo del procedimento di *prepivoting* nei modelli stratificati, ma potranno essere oggetto di studio in futuro. Un'ulteriore aspetto che potrebbe migliorare l'approccio tramite bootstrap è la sostituzione delle stime vincolate dei parametri di disturbo, da cui si simula, con delle stime migliorate, come ad esempio quelle proposte da Firth (1993) o da Giummolè e Ventura (2002).

# Appendice A

## Ordini di successioni e in probabilità

Esiste un'opportuna notazione per gli ordini d'errore per facilitare lo studio negli sviluppi asintotici, costituita dai simboli  $O(\cdot)$  e  $o(\cdot)$  e introdotta da Landau per quantità non stocastiche. La notazione è stata successivamente estesa al caso stocastico da Mann e Wald. Per approfondire gli argomenti esposti si può fare riferimento a Pace e Salvan (2001, Paragrafo 3.4) o a Severini (2005, Paragrafo 11.6). I simboli  $O(\cdot)$  e  $o(\cdot)$  rappresentano una relazione d'ordine tra le quantità studiate e delle quantità di riferimento.

Viene considerato inizialmente il caso non stocastico. Una successione di reali  $a_n$  è detta **di ordine**  $\mathbf{o}(n^\alpha)$  se la successione  $a_n/n^\alpha$  è infinitesima, ossia se

$$\lim_{n \rightarrow +\infty} \frac{a_n}{n^\alpha} = 0.$$

Nel caso in cui  $a_n$  sia infinitesima, la successione è di ordine  $o(1)$ .

Non è necessario che la successione  $a_n$  abbia limite e non è automatico il fatto che la relazione d'ordine indicata sia la più stretta possibile: ad esempio si può dire che  $n^{-1}$  è sia una successione di ordine  $o(n)$  che di ordine  $o(n^{-1/2})$ . Nonostante queste possibilità, risulta essere maggiormente informativa la potenza  $\alpha$  più piccola per cui  $a_n/n^\alpha$  è infinitesima. Si noti che usualmente si usano le potenze intere di  $n$  e di  $n^{1/2}$ .

Una successione di reali  $a_n$  è detta **di ordine**  $\mathbf{O}(n^\alpha)$  se la successione  $a_n/n^\alpha$  è

limitata, ossia se esiste un reale  $A$  tale che per ogni  $n \in \mathbb{N}$  sia

$$\left| \frac{a_n}{n^\alpha} \right| < A,$$

dove non è richiesta la convergenza di  $a_n$  né di  $a_n/n^\alpha$ . Se una successione  $a_n$  converge ad un limite finito diverso da zero, allora è di ordine  $O(1)$  (ma non di ordine  $o(1)$ ).

Nel caso stocastico si fa riferimento specialmente alla nozione di convergenza in probabilità. Una successione di variabili casuali  $Y_n$  è detta **di ordine  $o(n^\alpha)$  in probabilità**, e si indica con il simbolo  $o_p(n^\alpha)$ , se  $Y_n/n^\alpha \xrightarrow{p} 0$ , ovvero se, per ogni  $\delta > 0$ ,

$$\lim_{n \rightarrow +\infty} P \left\{ \left| \frac{Y_n}{n^\alpha} \right| > \delta \right\} = 0.$$

Una successione di variabili casuali  $Y_n$  è invece detta **di ordine  $O(n^\alpha)$  in probabilità**, e si scrive con il simbolo  $O_p(n^\alpha)$ , se per ogni  $\epsilon > 0$  esistono un reale  $A = A_\epsilon > 0$  e un naturale  $\bar{n} = \bar{n}_\epsilon$  tali che per ogni  $n \geq \bar{n}$

$$\lim_{n \rightarrow +\infty} P \left\{ \left| \frac{Y_n}{n^\alpha} \right| < A \right\} = 1 - \epsilon.$$

Se per una successione vale che  $Y_n \xrightarrow{d} Y$ , allora questa è detta **limitata in probabilità** ed equivalentemente è di ordine  $O_p(1)$ .

Esistono anche un certo numero di proprietà che rendono gli ordini in probabilità appartenenti ad una sorta di algebra. Si enuncia il seguente teorema che fornisce le linee guida.

**Teorema 1.** *Siano  $X_n, Y_n, W_n$  e  $Z_n$  delle successioni di variabili aleatorie a valori reali tali che  $Y_n > 0$  e  $Z_n > 0$ ,  $n = 1, 2, \dots$*

1. *Se  $X_n = o_p(1)$ , allora  $X_n = O_p(1)$ .*
2. *Se  $W_n = O_p(1)$  e  $X_n = O_p(1)$ , allora  $W_n + X_n = O_p(1)$  e  $W_n X_n = O_p(1)$ ; quindi  $O_p(1) + O_p(1) = O_p(1)$  e  $O_p(1)O_p(1) = O_p(1)$ .*
3. *Se  $W_n = O_p(1)$  e  $X_n = o_p(1)$ , allora  $W_n + X_n = O_p(1)$  e  $W_n X_n = o_p(1)$ ; quindi  $O_p(1) + o_p(1) = O_p(1)$  e  $O_p(1)o_p(1) = o_p(1)$ .*

---

4. Se  $X_n = O_p(Y_n)$  e  $W_n = O_p(Z_n)$ , allora

$$X_n W_n = O_p(Y_n Z_n) \quad e \quad X_n + W_n = O_p(\max(Y_n, Z_n));$$

quindi

$$O_p(Y_n)O_p(Z_n) = O_p(Y_n Z_n) \quad e \quad O_p(Y_n) + O_p(Z_n) = O_p(\max(Y_n, Z_n)).$$

Per la dimostrazione del teorema appena enunciato si rimanda a Severini (2005, Paragrafo 11.6).



# Bibliografia

- Arellano, M. e Hahn, J. (2007). Understanding bias in nonlinear panel models: some recent developments In *Advances in Economics and Econometrics*. A cura di Blundell R., Newey W., Persson T., volume 3. Cambridge University Press.
- 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. (1986). Inference on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika*, **73**, 307–322.
- Bartolucci, F., Bellio, R., Salvan, A. e Sartori, N. (2012). Modified profile likelihoods for panel data models. Working paper, sottoposto per pubblicazione.
- Bellio, R. e Sartori, N. (2012). panelMPL: Modified profile likelihood estimation for fixed-effects panel data models. R package version 0.12. Available at: <http://www.dies.uniud.it/software-bellio.html>.
- Beran, R. (1987). Prepivoting to reduce level error. *Biometrika*, **74**, 457–468.
- Beran, R. (1988). Prepivoting test statistics: A bootstrap view of asymptotic refinements. *J. Amer. Statist. Assoc.*, **83**, 687–697.

- Cox, D. R. e Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. Roy. Statist. Soc. B*, **49**, 1–39.
- Davison, A. C. e Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge.
- Davison, A. C., Hinkley, D. V. e Young, G. A. (2003). Recent developments in bootstrap methodology. *Statist. Sci.*, **18**, 141–157.
- DiCiccio, T. J. e Romano, J. P. (1995). On bootstrap procedures for second-order accurate confidence limits in parametric models. *Statist. Sinica*, **5**, 141–160.
- DiCiccio, T. J. e Stern, S. E. (1993). An adjustment to profile likelihood based on observed information. Relazione tecnica, Department of Statistics, Stanford University.
- DiCiccio, T. J., Martin, M. A., Stern, S. E. e Young, G. A. (1996). Information bias and adjusted profile likelihoods. *J. Roy. Statist. Soc. B*, **58**, 189–203.
- DiCiccio, T. J., Martin, M. A. e Stern, S. E. (2001). Simple and accurate one-sided inference from signed roots of likelihoods ratios. *Canad. J. Statist.*, **29**, 67–79.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, **7**, 1–26.
- Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, **80**, 27–38.
- Fisher, R. A. (1922). On the mathematical foundation of theoretical statistics. *Phil. Trans. Roy. Soc.*, **A**, 309–368.
- Giummolè, F. e Ventura, L. (2002). Pratical point estimation from higher-order pivots. *J. Statist. Comput. Simul.*, **72**, 419–430.
- Jørgensen, B. (1993). A review of conditional inference: is there a universal definition of noninformation? *Bull. Int. Statist. Inst.*, **55**, 323–340.
- Lee, S. M. S. e Young, G. A. (2005). Parametric bootstrapping with nuisance parameters. *Stat. Probabil. Lett.*, **71**, 143–153.

- McCullagh, P. e Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *J. Roy. Statist. Soc. B*, **52**, 325–344.
- Neyman, J. e Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, **16**, 33.
- Pace, L. e Salvan, A. (1996). *Teoria della statistica*. CEDAM, Padova.
- Pace, L. e Salvan, A. (2001). *Introduzione alla Statistica*. CEDAM, Padova.
- 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. (2000). *Likelihood Methods in Statistics*. Oxford University Press, Oxford.
- Severini, T. A. (2005). *Elements of Distribution Theory*. Cambridge University Press.
- Young, G. A. e Smith, R. L. (2005). *Essentials in Statistical Inference*. Cambridge University Press.