UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA
IN STATISTICA E GESTIONE DELLE IMPRESE

TESI DI LAUREA

METODI PER LA RIDUZIONE DELLA DISTORSIONE DELLO STIMATORE DI MASSIMA VEROSIMIGLIANZA

RELATORE: PROF. Alessandra Salvan

LAUREANDA: ALICE GIRARDI

ANNO ACCADEMICO 2010/2011

Indice

Introduzione

Statistica

Può definirsi come quel complesso di metodi che presiede all'astrazione, dai dati osservati, di informazioni sintetiche che servono a caratterizzare il fenomeno studiato per la parte ritenuta essenziale a scopi particolari. La statistica trova pertanto largo campo di applicazione nello studio di tutti i fenomeni in cui si suppongono operanti, a fianco di fattori sistematici di cui si desidera mettere in luce gli effetti, dei fattori di disturbo; avviene così che sulla manifestazione dei primi fattori, che di per sé avrebbe potuto essere descritta con un 'legge matematica', viene a sovrapporsi una variabilità che trasforma detta legge in 'regolarità statistica'.

Per raggiungere lo scopo d'individuare le caratteristiche essenziali del fenomeno, la metodologia statistica ricorre ampiamente alle tecniche proprie del calcolo delle probabilità, specie quando le rilevazioni effettuate non si estendono a tutte le possibili manifestazioni del fenomeno in esame. Hanno così origine i problemi detti di 'inferenza statistica', in cui ci si prefigge d'indurre le caratteristiche di un aggregato dall'osservazione di una parte di esso. (…)

Quando l'aggregato di dati viene considerato come parte di un insieme ignoto (campione estratto a caso da una popolazione statistica), le informazioni desiderate possono trascendere l'aggregato stesso e riguardare l'insieme.

(A. Naddeo, 1963)

Fare inferenza significa utilizzare informazioni campionarie per ottenere informazioni riguardanti l'intera popolazione. La teoria della stima parametrica è un problema dell'inferenza statistica che ha come obiettivo la stima dei parametri, scalari o vettoriali, a partire da dati campionari per i quali si assume un modello di probabilità specificato a meno di tali parametri.
Nell'ambito della teoria dell'inferenza statistica, lo stimatore di massima verosimiglianza riveste un ruolo fondamentale. Esso gode di importanti proprietà tra cui la non distorsione asintotica e l'equivarianza. Viceversa, la non distorsione per campioni finiti non è in genere garantita dallo stimatore di massima verosimiglianza, essendo tale proprietà in conflitto con l'equivarianza.
L'obiettivo della tesi consiste nell'illustrare i principali metodi utilizzati per la riduzione della distorsione degli stimatori di massima verosimiglianza, soffermandosi in particolare sull'approccio proposto da Firth (1993) e sulle sue recenti applicazioni. Il vantaggio di tale metodo sembra andare oltre la riduzione della distorsione e va individuato soprattutto nel fornire, in talune situazioni problematiche, una funzione di verosimiglianza con massimo finito.
Nel primo capitolo, dopo aver richiamato l'importante differenza tra stima e stimatore di un parametro, verrà trattato lo stimatore di massima verosimiglianza, di cui saranno richiamate definizione e proprietà.
Il secondo capitolo si occuperà invece dei possibili metodi di riduzione della distorsione delle stime di massima verosimiglianza. Dopo aver presentato brevemente i due approcci tradizionali, 'correttivi' piuttosto che 'preventivi', studiati nella letteratura, verrà illustrato il metodo proposto da Firth (1993). Tale approccio per la riduzione della distorsione consiste in una correzione sistematica sviluppata per il meccanismo che produce la stima di massima verosimiglianza, cioè per l'equazione di verosimiglianza, piuttosto che per la stima stessa.
Nel terzo capitolo verrà discusso il metodo proposto adattandolo ad alcu-

ni esempi pratici.
Infine, il quarto capitolo si occuperà di fornire un quadro sintetico dei campi di applicazione del metodo di Firth.

Capitolo 1

Teoria della stima e verosimiglianza

Questo primo capitolo ha il solo obiettivo di richiamare alcuni concetti fondamentali dell'inferenza statistica; dopo una breve definizione di modello statistico parametrico, vengono richiamate le nozioni di stima e stimatore di un parametro, passando poi alla presentazione delle proprietà finite e asintotiche degli stimatori stessi. Nel quinto paragrafo viene analizzata la funzione di verosimiglianza, per giungere, nel sesto paragrafo, alla definizione dello stimatore di massima verosimiglianza e ad una sintesi delle sue proprietà. Infine, nel settimo paragrafo, viene introdotta la definizione di famiglia esponenziale. I testi di riferimento utilizzati sono Azzalini (2001), Pace e Salvan (2001) e Piccolo (2006).

1.1 Modello statistico parametrico

Un modello statistico $\mathcal{F}$ può essere rappresentato da un qualunque insieme di funzione di ripartizione. Esiste una situazione particolare che riveste un ruolo fondamentale dal punto di vista sia teorico che applicativo. Si tratta del caso in cui tutti gli elementi di $\mathcal{F}$ sono funzioni dello stesso tipo, differenti tra loro solo per quanto riguarda θ, valore libero di variare entro

l'insieme $\Theta \subseteq \mathbb{R}^p$. Allora è possibile scrivere

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

dove, al variare di θ, $P(\,\cdot\,;\theta)$ è una funzione di ripartizione su $\mathbb{R}^n$, con p e n numeri naturali.
Spesso tali funzioni di ripartizione corrispondono tutte ad una variabile discreta oppure ad una variabile continua; $\mathcal{F}$ può dunque essere specificata tramite le corrispondenti funzioni di probabilità (se la variabile è discreta) o di densità (se la variabile è continua). In tal caso si scrive

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

dove p è una funzione rispettivamente di probabilità o di densità. La classe $\mathcal{F}$ è definita **modello statistico parametrico**.
Nel seguito, si userà il termine generale 'funzione di densità', sia nel caso discreto sia in quello continuo, per non appesantire l'esposizione.

1.2 Stima e stimatore di un parametro

Sia $Y \sim p(y;\theta)$ una variabile casuale con funzione di densità $p \in \mathcal{F}$, con $\theta \in \Theta$ ignoto. Spesso y è un campione casuale di numerosità n estratto da una variabile casuale Y_1 con funzione di densità $p_1(y_1;\theta)$. Sia $\mathcal{Y}$ lo spazio campionario, ossia l'insieme dei valori possibili di y.
Nella teoria della stima puntuale (dall'inglese *point estimation*) l'obiettivo è quello di scegliere un valore di θ che meglio di altri spiega i dati osservati y. Si ricerca quindi un'opportuna applicazione $\hat{\theta} : \mathcal{Y} \to \Theta$, definita dallo spazio campionario allo spazio parametrico, che fa corrispondere ad ogni elemento $y \in \mathcal{Y}$ un valore in Θ, $\hat{\theta} = \hat{\theta}(y)$, detto **stima** di θ. La statistica $\hat{\theta}(Y)$ è detta **stimatore** di θ.
Quindi, la stima è un valore in $\mathbb{R}^p$, mentre lo stimatore è una variabile casuale.
La distribuzione campionaria di $\hat{\theta}(Y)$, sotto θ, fornisce informazioni sull'incertezza insita nel processo di stima. Infatti è opportuno specificarla,

sia per valutare la bontà di una particolare procedura di stima, sia per confrontare tra loro stimatori alternativi.

1.3 Proprietà finite di uno stimatore

- Uno stimatore $\hat{\theta}$ si dice **non distorto** (*unbiased*) per θ se $E_\theta(\hat{\theta}) = \theta$ per ogni θ. La distorsione (*bias*) di uno stimatore è definita come $b(\theta) = E_\theta(\hat{\theta}) - \theta$. Si parla di distorsione positiva se $E_\theta(\hat{\theta}) > \theta$ e di distorsione negativa se $E_\theta(\hat{\theta}) < \theta$. Uno stimatore non distorto ha distorsione nulla.
  La non distorsione indica che lo stimatore $\hat{\theta}$ ha una distribuzione centrata perfettamente sul parametro θ che si intende stimare. Poiché la varianza di uno stimatore misura la dispersione dello stimatore stesso attorno alla sua media, allora essa consente di valutare la bontà di uno stimatore $\hat{\theta}$ solo se esso è non distorto.

- Si definisce **errore quadratico medio** (*Mean Square Error*) di uno stimatore $\hat{\theta}$ il valore medio $MSE_\theta(\hat{\theta}) = E_\theta(\hat{\theta} - \theta)^2$. Esso tiene conto sia della varianza sia della distorsione dello stimatore, in particolare $MSE_\theta(\hat{\theta}) = Var_\theta(\hat{\theta}) + [b(\theta)]^2$. Quindi è ovvio che, se lo stimatore è non distorto, il suo MSE coincide con la sua varianza, poiché la distorsione è pari a zero.

  Siano $\hat{\theta}_1$ e $\hat{\theta}_2$ due stimatori del parametro θ. Lo stimatore $\hat{\theta}_1$ si dice **più efficiente** dello stimatore $\hat{\theta}_2$ se $MSE_\theta(\hat{\theta}_1) < MSE_\theta(\hat{\theta}_2)$ per ogni θ.
  Per confrontare due stimatori di θ si utilizzano quindi i rispettivi MSE; si misura l'efficienza relativa tramite l'indice:

$$\mathit{eff}_\theta(\hat{\theta}_1|\hat{\theta}_2) = \frac{\frac{1}{MSE_\theta(\hat{\theta}_1)}}{\frac{1}{MSE_\theta(\hat{\theta}_2)}} = \frac{MSE_\theta(\hat{\theta}_2)}{MSE_\theta(\hat{\theta}_1)}.$$

  1. Se $\mathit{eff}_\theta(\hat{\theta}_1|\hat{\theta}_2) < 1$ per ogni $\theta \in \Theta$, $\hat{\theta}_2$ è preferibile rispetto a $\hat{\theta}_1$;

2. se $\mathit{eff}_\theta(\hat{\theta}_1|\hat{\theta}_2) > 1$ per ogni $\theta \in \Theta$, $\hat{\theta}_1$ è preferibile rispetto a $\hat{\theta}_2$;
3. se $\mathit{eff}_\theta(\hat{\theta}_1|\hat{\theta}_2) = 1$ per ogni $\theta \in \Theta$, $\hat{\theta}_1$ è equivalente a $\hat{\theta}_2$ in termini di MSE.

Se entrambi gli stimatori sono non distorti, allora l'efficienza relativa può essere espressa mediante il confronto tra varianze:

$$E_\theta(\hat{\theta}_1) = E_\theta(\hat{\theta}_2) = \theta \Longrightarrow \mathit{eff}_\theta(\hat{\theta}_1|\hat{\theta}_2) = \frac{Var_\theta(\hat{\theta}_2)}{Var_\theta(\hat{\theta}_1)}.$$

- Uno stimatore $\hat{\theta}$ non distorto si dice **efficiente** per θ in un modello $\mathcal{F}$ che soddisfa determinate condizioni di regolarità se si ha che $Var_\theta(\hat{\theta}) < Var_\theta(\theta^*)$ rispetto ad un qualunque altro stimatore non distorto θ^* del medesimo parametro.

1.4 Proprietà asintotiche di uno stimatore

Si definiscono proprietà asintotiche di uno stimatore quelle proprietà che valgono quando $n \to \infty$, assumendo $y = (y_1, \ldots, y_n)$ campione casuale semplice con numerosità n. Infatti è ragionevole richiedere che le proprietà statistiche di uno stimatore migliorino con l'aumentare della numerosità campionaria.

- Uno stimatore $\hat{\theta}$ si dice **asintoticamente non distorto** per θ se il limite per n che tende all'infinito del valore atteso dello stimatore è uguale a θ, ovvero se il limite per n che tende all'infinito della distorsione dello stimatore è uguale a zero:

$$\lim_{n\to\infty} E_\theta(\hat{\theta}) = \theta \Longleftrightarrow \lim_{n\to\infty} b(\theta) = 0, \quad \theta \in \Theta.$$

- Uno stimatore $\hat{\theta}$ si dice **consistente in media quadratica** per θ se il limite per n che tende all'infinito dell'errore quadratico medio MSE dello stimatore stesso è uguale a zero:

$$\lim_{n\to\infty} MSE_\theta(\hat{\theta}) = \lim_{n\to\infty} E_\theta(\hat{\theta} - \theta)^2 = 0, \quad \theta \in \Theta.$$

La consistenza in media quadratica è molto importante in quanto il MSE misura la variabilità media dello stimatore attorno al parametro θ; se la dispersione media tende a zero per n che tende all'infinito, allora la distribuzione di $\hat{\theta}$ sarà asintoticamente concentrata in θ.

- Uno stimatore $\hat{\theta}$ si dice **consistente** per θ se, per ogni $\epsilon > 0$ fissato,

$$\lim_{n\to\infty} \Pr_\theta(|\hat{\theta} - \theta| < \epsilon) = 1, \quad \theta \in \Theta.$$

Se uno stimatore è consistente in media quadratica, allora è consistente.

- Una successione di variabili casuali $(Y_n)_{n\in\mathbb{N}}$ con funzioni di ripartizione P_n si dice **convergere in distribuzione** alla variabile casuale Y con funzione di ripartizione P, cioè $Y^n \xrightarrow{d} Y$ se il limite $\lim P_n(y) = P(y)$ esiste in ogni punto $y \in \mathbb{R}$ in cui P risulta continua. Uno stimatore $\hat{\theta}$ di θ si dice **asintoticamente normale** se al divergere di n lo stimatore standardizzato converge in distribuzione alla normale standard $N(0,1)$:

$$\frac{\hat{\theta} - E_\theta(\hat{\theta})}{\sqrt{Var_\theta(\hat{\theta})}} \xrightarrow{d} Z \sim N(0,1), \quad \theta \in \Theta.$$

1.5 Funzione di verosimiglianza

Sia $\mathcal{F}$ un modello statistico parametrico per i dati y con funzione del modello $p(y;\theta)$, con $\theta = (\theta_1, \ldots, \theta_p) \in \Theta \subseteq \mathbb{R}^p$. Una volta fissato il valore campionario $y, p(y;\theta)$ può essere considerata come funzione solo di θ. Si dice **funzione di verosimiglianza** (*likelihood function*) di θ basata sui dati y la funzione $L : \Theta \longrightarrow \mathbb{R}^+$ definita da

$$L(\theta) = p(y;\theta).$$

Tale scrittura viene sostituita da $L(\theta; y)$ quando si vuole evidenziare la dipendenza di $L(\theta)$ dai dati campionari y.

Siano θ^1 e $\theta^2 \in \Theta$. Allora, se $L(\theta^1) > L(\theta^2), \theta^1$ è più credibile come indice del modello probabilistico generatore dei dati. Mediante il rapporto $L(\theta^1)/L(\theta^2)$ il sostegno empirico che θ^1 riceve da y viene confrontato con quello ricevuto da θ^2.
La funzione di verosimiglianza gode di un'importante proprietà: due funzioni di verosimiglianza che differiscono solo per una costante moltiplicativa (fattore che non dipende dal parametro θ) sono tra loro equivalenti.
Se si dispone di n osservazioni campionarie indipendenti ed identicamente distribuite o di campionamento semplice, allora la funzione di verosimiglianza diventa

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

dove p_i indica la densità della singola osservazione. La verosimiglianza totale si ottiene dunque moltiplicando fra loro le funzioni di verosimiglianza ottenute nei singoli esperimenti.

Poiché $L(\theta)$ è una quantità non negativa, e spesso è anzi positiva quasi certamente su tutto Θ, le procedure di inferenza basate su $L(\theta)$ sono espresse tramite la **funzione di log-verosimiglianza** (*log-likelihood function*), definita come

$$l(\theta) = \log L(\theta),$$

con la convenzione che, se $L(\theta) = 0$, allora $l(\theta) = -\infty$. L'esecuzione dei calcoli e la derivazione dei risultati teorici risulta più semplice se, in luogo della funzione di verosimiglianza, si utilizza la sua trasformata monotona logaritmica.
I valori di θ caratterizzati da elevata verosimiglianza presentano anche elevata log-verosimiglianza. Così, tramite la differenza $l(\theta^1) - l(\theta^2)$ il sostegno empirico che $\theta^1 \in \Theta$ riceve da y viene confrontato con quello ricevuto da $\theta^2 \in \Theta$.
La funzione di log-verosimiglianza gode di un'importante proprietà: due funzioni di log-verosimiglianza che differiscono solo per una costante ad-

ditiva (che non dipende dal parametro θ) sono tra loro equivalenti.
Se si dispone di n osservazioni campionarie indipendenti ed identicamente distribuite o di campionamento semplice, allora la funzione di log-verosimiglianza diventa

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

Un eventuale fattore esponenziale (non pratico da studiare) presente in $p_i(y_i; \theta)$ viene abbattuto dall'applicazione della funzione logaritmica.

1.6 Stima di massima verosimiglianza

Il concetto di stima di massima verosimiglianza (abbreviato in SMV) è stato introdotto da Sir R. A. Fisher (1922, 1925), anche se in realtà si riscontrano esempi del suo uso già da parte di D. Bernoulli nel 1777.
Si definisce **stima di massima verosimiglianza** di θ un valore $\hat{\theta} \in \Theta$ che rende massima la funzione di verosimiglianza $L(\theta)$ sullo spazio parametrico Θ, cioè tale per cui

$$L(\hat{\theta}) = \sup_{\theta \in \Theta} L(\theta).$$

In altre parole, la stima di massima verosimiglianza è tale che $L(\hat{\theta}) \geq L(\theta)$ per ogni $\theta \in \Theta$. $\hat{\theta}$ si può determinare anche a partire dalla funzione di log-verosimiglianza, di cui costituisce un massimo.
In generale, è necessario sottolineare alcuni aspetti della stima di massima verosimiglianza:

1. non è detto che la SMV esista;

2. se ci sono diversi valori di θ per cui la funzione di verosimiglianza è massima, allora la SMV non è unica.

Tuttavia è vero che, in molti casi di interesse, la SMV esiste ed è unica.
Dato il modello statistico, ogni campione osservabile y dà luogo ad una

particolare funzione di verosimiglianza. Se $\hat{\theta} = \hat{\theta}(y)$ esiste unico con probabilità uno, allora la variabile casuale $\hat{\theta} = \hat{\theta}(Y)$ è definita **stimatore di massima verosimiglianza**. Essa gode di importanti proprietà tra cui la non distorsione asintotica e l'equivarianza. Viceversa, la non distorsione per campioni finiti non è in genere garantita dallo stimatore di massima verosimiglianza, essendo tale proprietà in conflitto con la proprietà di equivarianza di seguito richiamata.
Proprietà di equivarianza: sia $\psi(\cdot)$ una funzione biunivoca dallo spazio Θ nello spazio Ψ; allora la SMV di $\psi(\theta)$ è $\psi(\hat{\theta})$ se $\hat{\theta}$ è la SMV di θ relativa alla verosimiglianza $L(\theta)$.
Tale proprietà è importante sia dal punto di vista computazionale, sia dal punto di vista concettuale, in quanto evita incongruenze passando da una parametrizzazione ad un'altra. Spesso si sente parlare di 'invarianza' invece che di 'equivarianza', ma questa denominazione non è del tutto appropriata; significherebbe infatti che la stima non varia dopo la trasformazione di ψ.

1.7 Famiglie esponenziali

Le famiglie esponenziali rappresentano un'importante classe di modelli statistici parametrici. Esse si distinguono in famiglie esponenziali monoparametriche e famiglie esponenziali multiparametriche.

Si definisce **famiglia esponenziale monoparametrica** un modello statistico parametrico $\mathcal{F}$ per un'osservazione y, univariata o multivariata, con parametro $\theta \in \Theta \subseteq \mathbb{R}$, la cui funzione del modello è

$$p(y;\theta) = c(\theta)h(y)\exp\{\psi(\theta)t(y)\},$$

dove $h(\cdot) \geq 0$, $\psi(\theta)$ è una funzione reale con dominio Θ e $t(\cdot)$ è una statistica. Le distribuzioni in $\mathcal{F}$ sono o tutte discrete o tutte assolutamente continue. $c(\theta) \in (0, +\infty)$ è la costante di normalizzazione associata a

$h(y)\exp\{\psi(\theta)t(y)\}$ integrabile e non negativa.
Il supporto di Y sotto θ è lo stesso per tutte le leggi di probabilità di una data famiglia esponenziale monoparametrica; infatti, esso è la chiusura dell'insieme $\{y \in \mathbb{R}^p : h(y) > 0\}$.
Se Θ contiene almeno due elementi e $\psi(\cdot)$ è iniettiva, allora $\mathcal{F}$ è non banale e θ è un parametro identificabile. Si dice quindi che $\psi = \psi(\theta)$ è un **parametro canonico** di $\mathcal{F}$ e che $t = t(y)$ è una **statistica canonica** si $\mathcal{F}$. Capita spesso che Θ sia un intervallo in $\mathbb{R}$ (eventualmente illimitato) e che $\psi(\cdot)$ sia derivabile assieme alla sua inversa.
Sono famiglie esponenziali monoparametriche per un'osservazione univariata y la binomiale con indice m fissato e parametro $\pi \in (0, 1)$, la Poisson con media $\lambda > 0$, l'esponenziale con tasso di guasto $\lambda > 0$, la gamma con parametro di scala fissato e parametro di forma $\alpha > 0$ e le normali univariate, rispettivamente con varianza fissata e con $\mu \in \mathbb{R}$ oppure con media fissata e con $\sigma^2 > 0$.

Si definisce **famiglia esponenziale multiparametrica** un modello statistico parametrico $\mathcal{F}$ per un'osservazione y, univariata o multivariata, con parametro $\theta \in \Theta \subseteq \mathbb{R}^p$, $p > 1$, la cui funzione del modello è

$$p(y;\theta) = c(\theta)h(y)\exp\{\sum_{j=1}^{k}\psi_j(\theta)t_j(y)\},$$

dove $h(\cdot) \geq 0$ e $\psi(\theta) = (\psi_1(\theta), \ldots, \psi_k(\theta))$ è una funzione con dominio Θ e codominio $\Psi = \psi(\Theta) \subseteq \mathbb{R}^k$. Le distribuzioni in $\mathcal{F}$ sono o tutte discrete o tutte assolutamente continue. $c(\theta) \in (0, +\infty)$ è la costante di normalizzazione associata a $h(y)\exp\{\sum_{j=1}^{k}\psi_j(\theta)t_j(y_i)\}$ integrabile e non negativa.
Il supporto di Y sotto θ è lo stesso per tutte le leggi di probabilità di una data famiglia esponenziale multiparametrica; infatti, esso è la chiusura dell'insieme $\{y \in \mathbb{R}^p : h(y) > 0\}$.
Se $\psi(\cdot)$ è iniettiva, allora $\theta = (\theta_1, \ldots, \theta_p)$ è un parametro identificabile. Se Θ contiene almeno $k + 1$ elementi, le $k + 1$ funzioni reali $1, \psi_1(\theta), \ldots, \psi_k(\theta)$ sono linearmente indipendenti (se infatti, per ogni $\theta \in \Theta$, si avesse $\psi_k(\theta) =$

$c_0 + c_1\psi_1(\theta) + \ldots + c_{k-1}\psi_{k-1}(\theta)$, si potrebbe riscrivere la funzione del modello usando solo $\psi_j(\theta)$ con le statistiche associate $t_j^{'}(y) = t_j(y) + c_j t_k(y)$ per $j = 1, \ldots, k-1$) e se sono linearmente indipendenti anche le $k+1$ funzioni reali $1, t_1(y), \ldots, t_k(y)$ allora $p(y;\theta) = c(\theta)h(y)\exp\{\sum_{j=1}^{k}\psi_j(\theta)t_j(y)\}$ è una **rappresentazione minimale**, ossia coinvolge il minimo numero di funzioni $\psi_j(\theta)$ e di associate statistiche $t_j(\theta)$. In questo caso, k è l'**ordine** della famiglia e $t = t(y) = (t_1(y), \ldots, t_k(y))$ è una **statistica canonica** di $\mathcal{F}$.
Si dice che $\psi = \psi(\theta)$ è un **parametro canonico** se $k = p$, cioè se l'ordine della famiglia è uguale alla dimensione di Θ e $\psi(\theta)$ è una riparametrizzazione del modello, con $\psi(\cdot)$ differenziabile su $\mathrm{int}\Theta$, come pure l'inversa $\theta = \theta(\psi)$ su $\mathrm{int}\Psi$, dove $\mathrm{int}\Theta$ e $\mathrm{int}\Psi$ indicano rispettivamente l'insieme dei punti interni di Θ e l'insieme dei punti interni di Ψ.
Nella parametrizzazione canonica ψ, la statistica canonica t ha densità $p(t;\psi) = c(\theta(\psi))\tilde{h}(t)\exp\{\sum_{j=1}^{p}\psi_j t_j\}$ con $\psi \in \Psi$ e $\tilde{h}(t)$ opportuna. Se Ψ è un insieme aperto e contiene ogni $\psi \in \mathbb{R}^p$ per cui la funzione non negativa $\tilde{h}(t)\exp\{\sum_{j=1}^{p}\psi_j t_j\}$ è integrabile, allora la famiglia esponenziale con densità $p(t;\psi)$ è definita **regolare**.
Sono famiglie esponenziali multiparametriche di ordine due per un'osservazione univariata y la normale univariata con parametro $\theta = (\mu, \sigma^2)$, dove $\mu \in \mathbb{R}$ e $\sigma^2 > 0$, e la gamma con parametro $\theta = (\alpha, \lambda)$, dove $\alpha > 0$ e $\lambda > 0$. La multinomiale e la normale multivariata sono esempi di famiglie esponenziali multiparametriche di ordine maggiore di due.

Capitolo 2

Riduzione della distorsione

In questo capitolo si affronta il problema della riduzione della distorsione degli stimatori di massima verosimiglianza. Dopo aver brevemente presentato i due metodi tradizionali, si passa all'approccio discusso da Firth (1993). Tale approccio per la riduzione della distorsione consiste in una correzione sistematica sviluppata per il meccanismo che produce la stima di massima verosimiglianza, cioè per l'equazione di verosimiglianza basata sulla funzione di punteggio, piuttosto che per la stima stessa.

2.1 Due metodi tradizionali

Se si dispone di un modello regolare con parametro $\theta = (\theta_1, \theta_2, \ldots, \theta_p)$ p-dimensionale, la distorsione asintotica dello stimatore di massima verosimiglianza $\hat{\theta}$ può essere scritta come

$$b(\theta) = \frac{b_1(\theta)}{n} + \frac{b_2(\theta)}{n^2} + \ldots,$$

dove n è la numerosità campionaria (Firth, 1993). Lo stimatore di θ basato su un campione di numerosità n è indicato con $\hat{\theta}(Y)$ e la sua media è $E_\theta(\hat{\theta}) = \theta + \frac{b_1(\theta)}{n} + \ldots$.

Sono stati ampiamente studiati e discussi due approcci tradizionali (Cox e Hinkley, 1974, § 8.4) per la riduzione della distorsione. Un aspetto comune

di questi due metodi è il fatto di essere 'correttivi' piuttosto che 'preventivi', cioè la stima di massima verosimiglianza $\hat{\theta}$ viene prima calcolata e poi corretta.

Il primo metodo (Quenouille, 1949, 1956) non richiede dettagliati calcoli numerici e risulta dunque particolarmente adatto a risolvere problemi complessi. Esso è chiamato *jackknife* (letteralmente **coltello a serramanico**) oppure *sample-splitting* e si basa sulla seguente idea. Sia y un campione casuale semplice con numerosità $n > 1$. Si supponga di avere a disposizione tutte le informazioni tranne la j-esima. Con queste $n-1$ osservazioni si può ancora stimare il valore del parametro della funzione. Tipicamente tale stima è leggermente diversa da quella ottenuta utilizzando tutte le n osservazioni e la differenza tra queste due stime fornisce proprio l'informazione con cui calcolare l'indeterminazione sulla stima del parametro. Dunque il metodo *jackknife* consiste nel ricalcolare più volte la grandezza statistica stimata lasciando fuori dal campione un'osservazione alla volta. Sia $\hat{\theta}_n$ una stima calcolata a partire da $Y_1, \ldots, Y_n$ e sia $\hat{\theta}_{n-1,j}$ la stessa stima calcolata a partire dall'insieme di $n-1$ variabili casuali ottenuto omettendo Y_j. Sia $\bar{\theta}_{n-1,\cdot}$ la media di $\hat{\theta}_{n-1,j}$ con $j = 1, \ldots, n$. Rispettivamente per $n-1$ e per n si ha

$$E_\theta(\hat{\theta}_{n-1}) = \theta + \frac{b_1(\theta)}{(n-1)} + \frac{b_2(\theta)}{(n-1)^2} + O((n-1)^{-3});$$

$$E_\theta(\bar{\theta}_{n-1,\cdot}) = \theta + \frac{b_1(\theta)}{(n-1)} + O(n^{-2});$$

$$E_\theta(\hat{\theta}_n) = \theta + \frac{b_1(\theta)}{n} + O(n^{-2}).$$

Dalle ultime due equazioni si può ottenere una combinazione lineare di $\bar{\theta}_{n-1,\cdot}$ e $\hat{\theta}_n$ con distorsione di ordine n^{-2}. Infatti, per

$$\hat{\theta}_n^J = n\hat{\theta}_n - (n-1)\bar{\theta}_{n-1,\cdot} = n\hat{\theta}_n - \frac{(n-1)}{n}\sum_{j=1}^{n}\hat{\theta}_{n-1,j}$$

risulta $E_\theta(\hat{\theta}_n^J) = \theta + O(n^{-2})$.

Se $\hat{\theta}_n$ è una media campionaria, allora $\hat{\theta}_n = \bar{\theta}_{n-1,\cdot} = \hat{\theta}_n^J$.

Il metodo *jackknife* richiede solo che il termine principale nella distorsione sia di ordine n^{-1}. Se il termine principale nello sviluppo di $E_\theta(\hat{\theta}_n)$ è di ordine n^{-2}, le modificazioni sopra descritte non sono necessarie; naturalmente, la distorsione di ordine uno non può essere rimossa.
Come già detto, il metodo *jackknife* è principalmente utilizzato nei problemi relativamente complicati come l'analisi dei dati di sopravvivenza, l'analisi delle serie temporali e l'analisi multivariata, in cui un diretto lavoro analitico non è possibile. Ciò nonostante, è utile osservare i risultati che si ottengono applicando il metodo a casi relativamente semplici. Un esempio elementare riguarda la stima della varianza a partire da

$$\hat{\theta}_n = \frac{\sum_{j=1}^n (Y_j - \bar{Y}_\cdot)^2}{n};$$

si dimostra che

$$\hat{\theta}_n^J = \frac{\sum_{j=1}^n (Y_j - \bar{Y}_\cdot)^2}{(n-1)}.$$

Il secondo approccio tradizionale è chiamato **riduzione della distorsione tramite sviluppo in serie** (*reduction of bias by series expansion*).
Sia $\hat{\theta}(Y)$ lo stimatore di massima verosimiglianza ottenuto da un campione casuale semplice con numerosità n. La media dello stimatore è

$$E_\theta[\hat{\theta}(Y)] = \theta + b(\theta) + O(n^{-2}),$$

dove $b(\theta)$ è di ordine $O(n^{-1})$.
Si può mostrare (Pace e Salvan, 1996, § 9.4.2) che nel caso univariato, cioè per $p = 1$,

$$E_\theta(\hat{\theta} - \theta) = \frac{1}{2} i(\theta)^{-2}(\nu_3(\theta) + 2\nu_{2,1}(\theta)) + O(n^{-2}) = b(\theta) + O(n^{-2}),$$

dove $i(\theta) = -E_\theta(\frac{\partial^2 l(\theta)}{\partial \theta^2})$, $\nu_3(\theta) = E_\theta(\frac{\partial^3 l(\theta)}{\partial \theta^3})$ e $\nu_{2,1}(\theta) = E_\theta(\frac{\partial^2 l(\theta)}{\partial \theta^2}\frac{\partial l\theta}{\partial \theta})$.
Semplicemente sostituendo $\hat{\theta}$ al posto del parametro ignoto θ in $\frac{b_1(\theta)}{n}$, la stima con distorsione 'corretta' risulta quindi essere

$$\hat{\theta}_I = \hat{\theta} - b(\hat{\theta}).$$

Si dimostra che essa ha distorsione $O(n^{-2})$: $E_\theta(\hat{\theta}_I(Y) - \theta) = O(n^{-2})$.
In linea di principio, tale metodo potrebbe essere esteso per fornire termini più elevati, ma nella pratica ciò è raramente necessario.

Entrambi questi metodi hanno successo nella rimozione del termine $\frac{b_1(\theta)}{n}$ dalla distorsione asintotica; il primo ha il vantaggio di non richiedere calcoli teorici (anche se ciò è solitamente controbilanciato da una perdita di precisione), mentre lo stimatore $\hat{\theta}_I$ è in generale efficiente (Firth, 1993).
Un requisito fondamentale per l'applicazione dell'uno o dell'altro metodo ad un campione di osservazioni è l'esistenza di $\hat{\theta}$ finito per tale campione (nel caso del *jackknife* $\hat{\theta}$ deve esistere finito anche per tutti i sotto-campioni del campione originario). Nel caso in cui $\hat{\theta}$ sia infinito, come può avvenire ad esempio nei modelli di regressione logistica, i due approcci tradizionali per la riduzione della distorsione non sono applicabili.

2.2 Il metodo di Firth (1993)

Firth (1993) mostra che nei modelli parametrici regolari, il termine dominante della distorsione asintotica dello stimatore di massima verosimiglianza può essere rimosso anche tramite un'appropriata modificazione della funzione di punteggio (*score function*).
Il metodo proposto da Firth (1993) non è vincolato alla finitezza di $\hat{\theta}$. Esso consiste in una correzione sistematica del meccanismo che produce la stima di massima verosimiglianza, cioè dell'equazione di verosimiglianza basata sulla funzione di punteggio, piuttosto che della stima stessa.

La stima di massima verosimiglianza $\hat{\theta}$ del parametro θ si ottiene come soluzione dell'equazione di verosimiglianza:

$$\nabla l(\theta) = U(\theta) = 0.$$

Per ridurre la distorsione asintotica dello stimatore di massima verosimiglianza si effettua una modificazione della funzione di punteggio.

La funzione di punteggio modificata è

$$U^*(\theta) = U(\theta) + A(\theta) = [U_r + A_r],$$

dove $[a_r]$ indica il vettore con generica componente a_r, per $r = 1, \ldots, p$.
La stima di massima verosimiglianza corretta θ^* si ottiene dunque come soluzione dell'equazione di verosimiglianza basata sulla funzione di punteggio modificata

$$U^*(\theta) = 0$$

e presenta distorsione asintotica $O(n^{-2})$, inferiore rispetto alla distorsione di $\hat{\theta}$.

Si consideri ora (Firth, 1993) un modello di una famiglia esponenziale con funzione di log-verosimiglianza $l(\theta) = t\theta - K(\theta)$ in cui $p = 1$. Si ottiene

$$U(\theta) = l^{'}(\theta) = t - K^{'}(\theta).$$

In questo caso, la statistica sufficiente t non influenza la forma di $U(\theta)$, ma solo la sua posizione. La distorsione di $\hat{\theta}$ deriva dalla combinazione di due fattori:

1. non distorsione della funzione di punteggio, $E_\theta\{U(\theta)\} = 0$ al vero valore di θ;

2. curvatura della funzione di punteggio, $l^{''}(\theta) \neq 0$.

Se $U(\theta)$ è lineare in θ, allora $E_\theta(\hat{\theta}) = \theta$, ma la curvatura e la non distorsione della funzione di punteggio si combinano provocando una distorsione nello stimatore di massima verosimiglianza $\hat{\theta}$.
Dunque la distorsione di $\hat{\theta}$ può essere ridotta attraverso l'introduzione di una piccola distorsione nella funzione di punteggio. La modificazione appropriata per $U(\theta)$ è data dalla semplice geometria del triangolo illustrata nella Fig. 2.1. Se $\hat{\theta}$ è soggetto ad una distorsione positiva (cioè se $E_\theta(\hat{\theta}) > \theta$), la funzione di punteggio deve essere spostata verso il basso in

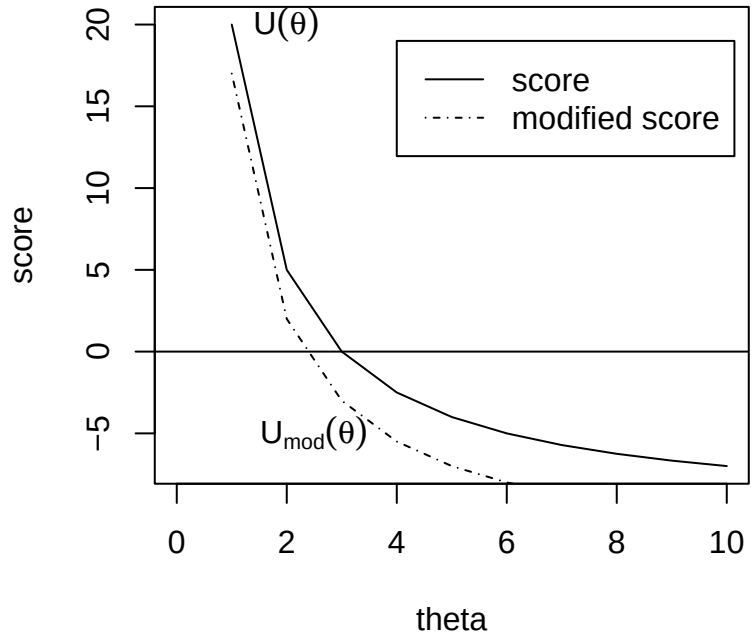

Figura 2.1: *Modificazione della funzione di punteggio non distorta, dove* $U_{mod}(\theta)$ *rappresenta* $U^*(\theta)$.

ogni punto θ di una quantità pari a $i(\theta)b(\theta)$, dove $b(\theta)$ indica la distorsione e $-i(\theta) = E_\theta(U^{'}(\theta))$ rappresenta il gradiente locale, con $U^{'}(\theta) = \frac{\partial U(\theta)}{\partial \theta}$. Si ha infatti

$$U(\theta^* + b(\theta^*)) \doteq U(\theta^*) + b(\theta^*)(-i(\theta^*)),$$

e dunque la traslazione verso il basso di $U(\theta)$ deve essere pari a $A(\theta) = -i(\theta)b(\theta)$.
Si ottiene quindi una nuova funzione di punteggio

$$U^*(\theta) = U(\theta) - i(\theta)b(\theta).$$

Essa può essere interpretata come un vettore di equazioni nel caso si disponga di un vettore di parametri e $i(\theta)$ è la matrice di informazione attesa (o informazione di Fisher).
Ponendo infine $U^*(\theta) = 0$, si ottiene la stima modificata θ^*.

Per formalizzare l'argomento fin qui discusso ed estenderlo anche ai problemi che non riguardano le famiglie esponenziali si utilizza una diversa notazione per le derivate della funzione di log-verosimiglianza e per i loro momenti nulli (McCullagh, 1987).
Le derivate della funzione di log-verosimiglianza sono date da

$$U_r(\theta) = \frac{\partial l(\theta)}{\partial \theta_r}, U_{rs}(\theta) = \frac{\partial^2 l(\theta)}{\partial \theta_r \partial \theta_s},$$

e così via, dove $\theta = (\theta_1, \ldots, \theta_p)$ è il vettore dei parametri.
I momenti nulli sono invece definiti come

$$\kappa_{r,s} = n^{-1}E_\theta\{U_r U_s\}, \kappa_{r,s,t} = n^{-1}E_\theta\{U_r U_s U_t\}, \kappa_{r,st} = n^{-1}E_\theta\{U_r U_{st}\},$$

e così via. Sono note (McCullagh, 1987) le seguenti relazioni:

$$\kappa_{rs} + \kappa_{r,s} = 0, \kappa_{rst} + \kappa_{r,st} + \kappa_{s,rt} + \kappa_{t,rs} + \kappa_{r,s,t} = 0.$$

Una funzione di punteggio modificata abbastanza generale è data da

$$U_r^*(\theta) = U_r(\theta) + A_r(\theta)$$

in cui il pedice r indica che la *score function* modificata segue le componenti di θ e A_r è scelta in modo da dipendere dai dati. A_r è $O_p(1)$ quando $n \to \infty$. Siano $\hat{\theta}$ e θ^* tali da soddisfare $U(\hat{\theta}) = 0$ e $U(\theta^*) = 0$ e sia $\hat{\gamma} = n^{\frac{1}{2}}(\theta^* - \theta)$. Allora la distorsione di θ^*, basata su uno sviluppo di $U_r^*(\theta^*)$, è

$$E_\theta(n^{-\frac{1}{2}}\hat{\gamma}^r) = n^{-1}\kappa^{r,s}\{-\kappa^{t,u}\frac{(\kappa_{s,t,u} + \kappa_{s,tu})}{2} + \alpha_s\} + O(n^{-\frac{3}{2}}),$$

dove $\kappa^{r,s}$ è l'inversa della matrice di informazione attesa $\kappa_{r,s}$ e α_s è il valore atteso nullo di A_s.
La distorsione di primo ordine di $\hat{\theta}$ è rappresentata dal termine

$$-n^{-1}\kappa^{r,s}\kappa^{t,u}\frac{(\kappa_{s,t,u} + \kappa_{s,tu})}{2} = n^{-1}b_1^r(\theta).$$

Quindi A_r rimuove il termine di primo ordine se soddisfa la seguente condizione:

$$\kappa^{r,s}\alpha_s = -b_1^r + O(n^{-\frac{1}{2}})$$

da cui si ottiene

$$\alpha_r = -\kappa_{r,s}b_1^s + O(n^{-\frac{1}{2}}).$$

In notazione matriciale, il vettore A dovrebbe essere tale che

$$E_\theta(A) = -i(\theta)\frac{b_1(\theta)}{n} + O(n^{-\frac{1}{2}}).$$

Usando rispettivamente l'informazione attesa e l'informazione osservata, ovvii candidati per una scelta di A che riduce la distorsione sono dunque $A^{(E)} = -i(\theta)\frac{b_1(\theta)}{n}$ e $A^{(O)} = -I(\theta)\frac{b_1(\theta)}{n}$, con $I(\theta) = -U'(\theta)$. Nel caso di famiglia esponenziale con parametrizzazione canonica $A^{(O)}$ e $A^{(E)}$ coincidono in quanto l'informazione osservata non dipende dai dati.
In generale, entrambe le modificazioni $A^{(E)}$ e $A^{(O)}$ rimuovono il termine di distorsione $O(n^{-1})$.

2.3 Applicazione del metodo alle famiglie esponenziali

2.3.1 La distribuzione a priori di Jeffreys come funzione di penalità per la riduzione della distorsione

Se θ è il parametro canonico di un modello di una famiglia esponenziale, $\kappa_{r,st} = 0$ per ogni r, s e t. Dunque, utilizzando la **convenzione per la somma** di Einstein (Pace e Salvan, 1996, § 9.1), in base alla quale, quando in un prodotto di elementi di matrici generalizzate un indice compare due o più volte, si sottintende la somma rispetto a quell'indice sopra il campo di variazione (sottinteso perché ovvio), l'elemento r-esimo di $A^{(E)}(\theta)$ (o equivalentemente di $A^{(O)}(\theta)$) è dato da

$$A_r = -n\kappa_{r,s}\frac{b_1^s}{n} = \frac{\kappa_{r,s}\kappa^{s,t}\kappa^{u,v}\kappa_{t,u,v}}{2} = \frac{\kappa^{u,v}\kappa_{r,u,v}}{2} = \frac{-\kappa^{u,v}\kappa_{ruv}}{2}$$

che, in notazione matriciale, diventa

$$A_r = \frac{1}{2}tr\{i^{-1}(\frac{\partial i}{\partial\theta_r})\} = \frac{\partial}{\partial\theta_r}\{\frac{1}{2}\log|i(\theta)|\}.$$

La soluzione di $U_r^* \equiv U_r + A_r = 0$, individua dunque un punto stazionario di

$$l^*(\theta) = l(\theta) + \frac{1}{2}\log|i(\theta)|$$

o, equivalentemente, della funzione di verosimiglianza penalizzata

$$L^*(\theta) = L(\theta)|i(\theta)|^{\frac{1}{2}}.$$

Il determinante dell'informazione di Fisher elevato alla $\frac{1}{2}$ si chiama **distribuzione a priori di Jeffreys**; per il parametro canonico di un modello di una famiglia esponenziale il termine di distorsione $O(n^{-1})$ è rimosso dallo stimatore definito come moda della distribuzione a posteriori basata su questa distribuzione a priori.

2.4 Applicazione del metodo agli altri modelli

Si consideri ora uno scenario più generale che include sia i modelli della famiglia esponenziale con parametrizzazione non canonica, sia i modelli non esponenziali.
In questo nuovo scenario, la funzione di punteggio modificata può essere scritta come

$$U_r^* = U_r + A_r,$$

dove $A_r(\theta)$ è basato o sull'informazione attesa, nel qual caso si ha

$$A_r = A_r^{(E)} = n\kappa_{r,s}\kappa^{s,t}\kappa^{u,v}\frac{(\kappa_{t,u,v} + \kappa_{t,uv})}{2n} = \kappa^{u,v}\frac{(\kappa_{r,u,v} + \kappa_{r,uv})}{2},$$

o sull'informazione osservata, per cui

$$A_r = A_r^{(O)} = -U_{rs}\kappa^{s,t}\kappa^{u,v}\frac{(\kappa_{t,u,v}\kappa_{t,uv})}{2n}.$$

Le stime derivate utilizzando $A_r^{(O)}$ risultano preferibili in termini di efficienza; per dimostrarlo si consideri uno sviluppo di $U^*(\theta^*)$. Dalla definizione,

$$0 = U_r^*(\theta^*) = U_r(\theta^*) + A_r(\theta^*).$$

Se $A_r(\theta) = A_r^{(O)}(\theta) = U_{rs}(\theta)\frac{b_1^s(\theta)}{n}$, si ha

$$(\theta^* - \hat{\theta})^r = -\frac{b_1^r(\hat{\theta})}{n} + O_p(n^{-2}),$$

mentre se $A_r(\theta) = A_r^{(E)}(\theta) = -i_{rs}(\theta)\frac{b_1^s(\theta)}{n}$, allora

$$(\theta^* - \hat{\theta})^r = -\frac{b_1^r(\hat{\theta})}{n} - i^{rs}(\hat{\theta})\{U_{st}(\hat{\theta}) + i_{st}(\hat{\theta})\}\frac{b_1^t(\hat{\theta})}{n} + O_p(n^{-2}).$$

La differenza $U_{st}(\hat{\theta}) + i_{st}(\hat{\theta})$ tra l'informazione attesa e l'informazione osservata nella stima di massima verosimiglianza è, in generale, $O_p(n^{-\frac{1}{2}})$, così che il termine supplementare in $(\theta^* - \hat{\theta})^r$ calcolato a partire da $A_r^{(E)}(\theta)$ è $O_p(n^{-\frac{3}{2}})$. Nel caso particolare di una famiglia esponenziale piena, tale termine scompare con qualche parametrizzazione.

Da $(\theta^* - \hat{\theta})^r = -\frac{b_1^r(\hat{\theta})}{n} + O_p(n^{-2})$ si può concludere (Firth, 1993) che, se U^* è calcolato utilizzando l'informazione osservata, θ^* coincide con $\hat{\theta}_I$ al secondo ordine, mentre non è così se si utilizza l'informazione attesa. In generale, l'uso della modificazione $A^{(E)}$ implica una perdita di precisione al secondo ordine rispetto all'uso di $A^{(O)}$.

2.5 Sintesi

Come mostrato, nei problemi regolari, la distorsione asintotica dello stimatore di massima verosimiglianza può essere ridotta mediante la rimozione del termine $O(n^{-1})$ ottenuta introducendo un appropriato termine di distorsione nella funzione di punteggio. Se il parametro di interesse è il parametro canonico di una famiglia esponenziale, ciò equivale semplicemente ad utilizzare la distribuzione a priori di Jeffreys come funzione di penalità per la verosimiglianza. Nel caso di altre parametrizzazioni, sono disponibili diversi tipi di correzioni, ottenute utilizzando l'informazione attesa o l'informazione osservata. Al di fuori dei modelli della famiglia esponenziale, l'uso dell'informazione attesa si traduce in una perdita di efficienza rispetto all'uso dell'informazione osservata. Non sempre la riduzione della distorsione dello stimatore di massima verosimiglianza risulta desiderabile.

Capitolo 3

Esempi

In questo capitolo vengono presentati alcuni esempi in cui si applica il metodo proposto da Firth (1993) per la riduzione della distorsione dello stimatore di massima verosimiglianza. Accanto agli esempi elaborati dallo stesso Firth (1993), ne sono presi in considerazione anche degli altri, con l'obiettivo di consolidare la teoria discussa nel capitolo precedente. In particolare, nel § 3.1 viene riproposto un esempio svolto interamente da Firth (1993), mentre nel § 3.2 e nel § 3.3, a partire da esempi di Firth (1993), si svolgono tutti i calcoli necessari per giungere ai risultati finali. Infine, nel § 3.4, viene proposto un esempio simile al precedente, ma nuovo, riguardante il tasso di guasto di una distribuzione esponenziale.

3.1 Regressione logistica binomiale

Il calcolo e la correzione della distorsione degli stimatori di massima verosimiglianza dei parametri della regressione logistica sono stati studiati da molti autori, tra cui Cordeiro e McCullagh (1991).
Se la probabilità di successo per l'i-esima osservazione è $\pi_i = \exp(\eta_i)/(1 + \exp(\eta_i))$, dove $\eta_i = \sum x_{ir}\beta_r$, le stime di massima verosimiglianza del parametro β sono distorte lontano dal punto $\beta = 0$. La correzione della distorsione richiede quindi un certo grado di 'riduzione' di $\hat{\beta}$ verso questo

punto.
Nella regressione logistica la matrice di informazione è $i(\beta) = I(\beta) = X^T W X$, dove $X = \{x_{ir}\}$ è la matrice di disegno, $W = \mathrm{diag}\{m_i \pi_i (1 - \pi_i)\}$ e m_i è l'indice binomiale per l'i-esimo livello. Il determinante è massimizzato per $\pi_i = \frac{1}{2}$, con $i = 1, \ldots, n$, che corrisponde a $\beta = 0$. Dunque, la distribuzione a priori di Jeffreys restringe le stime verso questo punto. Le argomentazioni del § 2.2 mostrano che l'entità della riduzione è esattamente quella necessaria per rimuovere il termine di distorsione $O(n^{-1})$. Ciò può essere mostrato anche dal fatto che, a condizione che X sia di rango pieno, il $\log |i(\beta)|$ è strettamente concavo ed illimitato inferiormente per $\beta \to \infty$; questo, assieme al fatto che la stessa $l(\beta)$ è strettamente concava e limitata superiormente, assicura che la stima di massima verosimiglianza penalizzata β^* esiste ed è unica.
A dimostrazione di quanto appena detto, si consideri ora una variante di un esempio con un solo parametro utilizzato da Copas (1988). Sia $\eta_i = x_i \beta$, con x_i che assume valori nell'insieme $\{-2, -1, 0, 1, 2\}$. Copas (1988) ha esaminato $\frac{b_1(\beta)}{n}$ quando dieci osservazioni binarie sono estratte da ciascuno dei cinque punti e si è accorto che, per valori piccoli di β, la distorsione asintotica (lontano da zero) è pari a circa il 3.4% del vero valore. Firth (1993) considera un esempio molto più semplice, in cui si estrae una sola osservazione binaria da ciascuno dei cinque punti, così che sia possibile la completa enumerazione. La statistica sufficiente $t = \sum_{i=1}^{n} y_i x_i$ può assumere solo sette valori (l'osservazione per $x_i = 0$ non contribuisce a t ed è quindi ridondante).
La Tabella 3.1 fornisce $\hat{\beta}$, $\hat{\beta}_I$ e β^* corrispondenti ai sette valori di t, e anche la distribuzione campionaria per due particolari valori di β. La media di β^* è pari a 0.46 quando $\beta = 0.5$ ed è pari a 0.82 quando $\beta = 1$. Ciò è soddisfacente date la dimensione molto piccola del campione e l'elevata probabilità che $\hat{\beta}$ sia infinito.
Il più semplice di tutti i modelli logistici è quello per cui, per una singola osservazione binomiale, il parametro di interesse è $\beta = \log\{\pi(1 - \pi)\}$.

$t(y)$	$\hat{\beta}$	$\hat{\beta}_I$	β^*	$p(t(y);0.5)$	$p(t(y);1)$
-3	$-\infty$	$-$	-1.38	0.010	0.001
-2	-1.01	-0.52	-0.68	0.034	0.006
-1	-0.42	-0.27	-0.31	0.084	0.023
0	0	0	0	0.185	0.083
1	0.42	0.27	0.31	0.229	0.168
2	1.01	0.52	0.68	0.251	0.305
3	∞	$-$	1.38	0.207	0.415

Tabella 3.1: *Distribuzione degli stimatori in un piccolo modello di regressione logistica.*

L'informazione è proporzionale a $\pi(1-\pi)$, così che (Cox e Snell, 1989, § 2.1.6) la verosimiglianza penalizzata è

$$L^* = \pi^{y+\frac{1}{2}}(1-\pi)^{m-y+\frac{1}{2}}.$$

La massimizzazione di L^* produce

$$\beta^* = \log(\frac{y+\frac{1}{2}}{m-y+\frac{1}{2}}).$$

Unicamente per questo semplice modello, β^* è la stima di massima verosimiglianza calcolata a partire dai dati aggiustati, ottenuti aggiungendo rispettivamente $\frac{1}{2}$ a y e 1 a m.
La forma generale del vettore di distorsione $O(n^{-1})$ è fornita da McCullagh e Nelder (1989):

$$\frac{b_1}{n} = (X^T W X)^{-1} X^T W \zeta,$$

dove $W\zeta$ ha come i-esimo elemento $h_i(\pi_i - \frac{1}{2})$ e h_i è l'i-esimo elemento diagonale della matrice 'hat'

$$H = W^{\frac{1}{2}} X (X^T W X)^{-1} X^T W^{\frac{1}{2}}.$$

Dunque $U^* = U - X^T W \zeta$, con r-esimo elemento componente

$$U_r^* = \sum_{i=1}^{n} \{(y_i + \frac{h_i}{2}) - (m_i + h_i)\pi_i\} x_{ir}, \quad (r = 1, \ldots, p).$$

La soluzione di $U^* = 0$ è quindi equivalente alla soluzione delle equazioni di massima verosimiglianza, basate sui dati aggiustati, ottenuti aggiungendo rispettivamente $\frac{h_i(\beta^*)}{2}$ a y_i e $h_i(\beta^*)$ a m_i. Ciò suggerisce un algoritmo iterativo in cui gli aggiustamenti $\{h_i\}$ vengono aggiornati ad ogni ciclo di una procedura standard dei minimi quadrati pesati iterati (Firth, 1992).

3.2 Distribuzione normale

Sia $(y_1, \ldots, y_n)$ un campione casuale da una distribuzione normale con media μ e varianza σ^2. Per semplicità, indichiamo σ^2 con ϕ.
La funzione di densità di Y_i è

$$p_i(y_i) = \frac{1}{\sqrt{2\pi\phi}} \exp\{-\frac{1}{2\phi}(y_i - \mu)^2\}$$

da cui si ottengono rispettivamente la funzione di verosimiglianza e la funzione di log-verosimiglianza:

$$L(\mu, \phi) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\phi}} \exp\{-\frac{1}{2\phi}(y_i - \mu)^2\} = \frac{1}{(2\pi\phi)^{\frac{n}{2}}} \exp\{-\frac{1}{2\phi}\sum_{i=1}^{n}(y_i - \mu)^2\};$$

$$l(\mu, \phi) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\phi - \frac{1}{2\phi}\sum_{i=1}^{n}(y_i - \mu)^2.$$

Il vettore di punteggio ha come componenti:

$$U_\mu = \frac{\partial l(\mu, \phi)}{\partial \mu} = \frac{1}{2\phi} 2 \sum_{i=1}^{n}(y_i - \mu) = \frac{1}{\phi}\sum_{i=1}^{n} y_i - \frac{n\mu}{\phi},$$

$$U_\phi = \frac{\partial l(\mu, \phi)}{\partial \phi} = -\frac{n}{2\phi} + \frac{2\sum_{i=1}^{n}(y_i - \mu)^2}{4\phi^2} = -\frac{n}{2\phi} + \frac{\sum_{i=1}^{n}(y_i - \mu)^2}{2\phi^2}.$$

Indicando rispettivamente $\sum_{i=1}^{n} y_i = y_.$ e $\sum_{i=1}^{n}(y_i - \mu)^2 = s(\mu)$ si ottiene $U_\mu = \frac{y_. - n\mu}{\phi}$ e $U_\phi = \frac{s(\mu)}{2\phi^2} - \frac{n}{2\phi}$.
Per la matrice di informazione osservata $I(\mu, \phi)$ è necessario calcolare le derivate seconde della funzione di log-verosimiglianza:

$$\frac{\partial^2 l(\mu, \phi)}{\partial \mu^2} = -\frac{n}{\phi};$$

$$\frac{\partial^2 l(\mu,\phi)}{\partial \phi^2} = \frac{2n}{4\phi^2} - \frac{4\phi s(\mu)}{4\phi^4} = \frac{n}{2\phi^2} - \frac{s(\mu)}{\phi^3};$$

$$\frac{\partial^2 l(\mu,\phi)}{\partial \mu \partial \phi} = \frac{\partial^2 l(\mu,\phi)}{\partial \phi \partial \mu} = -\frac{y_. - n\mu}{\phi^2}.$$

Così,

$$I(\mu,\phi) = -U^{'}(\mu,\phi) = \begin{pmatrix} \frac{n}{\phi} & \frac{y_. - n\mu}{\phi^2} \\ \frac{y_. - n\mu}{\phi^2} & \frac{s(\mu)}{\phi^3} - \frac{n}{2\phi^2} \end{pmatrix}.$$

L'informazione attesa corrisponde alla media dell'informazione osservata:

$$E_{\mu,\phi}(\frac{n}{\phi}) = \frac{n}{\phi};$$

$$E_{\mu,\phi}(\frac{y_. - n\mu}{\phi^2}) = \frac{1}{\phi^2} E_{\mu,\phi}\{\sum_{i=1}^{n} y_i\} - \frac{n\mu}{\phi^2} = \frac{1}{\phi^2}\sum_{i=1}^{n} E_{\mu,\phi} y_i - \frac{n\mu}{\phi^2} = \frac{n\mu}{\phi^2} - \frac{n\mu}{\phi^2} = 0;$$

$$\begin{aligned} E_{\mu,\phi}\{\frac{s(\mu)}{\phi^3} - \frac{n}{2\phi^2}\} &= \frac{1}{\phi^3} E_{\mu,\phi}\{\sum_{i=1}^{n}(y_i - \mu)^2\} - \frac{n}{2\phi^2} \\ &= \frac{1}{\phi^3} E_{\mu,\phi}\{\sum_{i=1}^{n}(y_i^2 + \mu^2 - 2\mu y_i)\} - \frac{n}{2\phi^2} \\ &= \frac{1}{\phi^3}\sum_{i=1}^{n} E_{\mu,\phi}\{y_i^2\} + \frac{n\mu^2}{\phi^3} - \frac{2\mu}{\phi^3}\sum_{i=1}^{n} E_{\mu,\phi}\{y_i\} - \frac{n}{2\phi^2} \\ &= \frac{n(\mu^2+\phi)}{\phi^3} + \frac{n\mu^2}{\phi^3} - \frac{2n\mu^2}{\phi^3} - \frac{n}{2\phi^2} \\ &= \frac{n\mu^2}{\phi^3} + \frac{n\phi}{\phi^3} - \frac{n\mu^2}{\phi^3} - \frac{n}{2\phi^2} \\ &= \frac{n}{\phi^2} - \frac{n}{2\phi^2} = \frac{2n-n}{2\phi^2} = \frac{n}{2\phi^2}, \end{aligned}$$

dove si è sfruttato $Var\{Y_i\} = \phi = E\{Y_i^2\} - [E\{Y_i\}]^2 = E\{Y_i^2\} - \mu^2$
da cui $E\{Y_i^2\} - \mu^2 = \phi \Rightarrow E\{Y_i^2\} = \mu^2 + \phi$.
Così,

$$i(\mu,\phi) = E\{I(\mu,\phi)\} = \begin{pmatrix} \frac{n}{\phi} & 0 \\ 0 & \frac{n}{2\phi^2} \end{pmatrix}.$$

Ricordando che

$$\sum (Y_i - \mu)^2 = \sum (Y_i - \bar{Y})^2 + n(\bar{Y} - \mu)^2 = (n-1)s^2 + n(\bar{Y} - \mu)^2;$$

$$E[(Y-\mu)^p] = \begin{cases} 0 & \text{se } p \text{ dispari} \\ (\sqrt{\phi})^p (p-1)!! & \text{se } p \text{ pari} \end{cases};$$

$$E[(\bar{Y}-\mu)^p] = \begin{cases} 0 & \text{se } p \text{ dispari} \\ (\frac{\phi}{n})^{\frac{p}{2}} (p-1)!! & \text{se } p \text{ pari} \end{cases};$$

$$n!! = \begin{cases} 1 & \text{se } n = 0 \text{ o } n = 1 \\ n[(n-2)!!] & \text{se } n \geq 2 \end{cases};$$

$$v \sim \chi^2_g \sim 2Ga(\frac{g}{2}, 1) \;\Rightarrow\; E(v^k) = 2^k \frac{\Gamma(\frac{g}{2} + k)}{\Gamma(\frac{g}{2})};$$

$$\frac{(n-1)s^2}{\phi} \sim \chi^2_{n-1} \;\Rightarrow\; E[(s^2)^h] = (\frac{\phi}{n-1})^h 2^h \frac{\Gamma(\frac{n-1}{2} + h)}{\Gamma(\frac{n-1}{2})},$$

le rimanenti quantità richieste per il calcolo di U^* sono:

$$\begin{aligned}
\kappa_{\mu,\mu,\phi} &= n^{-1} E(U_\mu^2 U_\phi) \\
&= \frac{1}{n} E\{(\frac{\sum Y_i - n\mu}{\phi})^2 [\frac{\sum (Y_i - \mu)^2}{2\phi^2}]\} \\
&= \frac{1}{n} E\{\frac{n^2(\bar{Y}-\mu)^2}{\phi^2} [\frac{\sum (Y_i - \bar{Y} + \bar{Y} - \mu)^2}{2\phi^2} - \frac{n}{2\phi}]\} \\
&= \frac{1}{n} E\{\frac{n^2(\bar{Y}-\mu)^2}{\phi^2} [\frac{(n-1)s^2 + n(\bar{Y}-\mu)^2}{2\phi^2} - \frac{n}{2\phi}]\} \\
&= \frac{1}{n} \{\frac{n^2(n-1)}{2\phi^4} E(\bar{Y}-\mu)^2 E(s^2) + \frac{n^3}{2\phi^4} E(\bar{Y}-\mu)^4 - \frac{n^3}{2\phi^3} E(\bar{Y}-\mu)^2\} \\
&= \frac{1}{n} \{\frac{n^2(n-1)}{2\phi^4} \frac{\phi}{n} \phi + \frac{n^3}{2\phi^4} \frac{3\phi^2}{n^2} - \frac{n^3}{2\phi^3} \frac{\phi}{n}\} \\
&= \frac{1}{n\phi^2} \{\frac{n^2 - n}{2} + \frac{3n}{2} - \frac{n^2}{2}\} \\
&= \frac{1}{n\phi^2} n = \frac{1}{\phi^2};
\end{aligned}$$

$$
\begin{aligned}
\kappa_{\phi,\phi,\phi} &= n^{-1}E(U_\phi^3) = \frac{1}{n}E\{(\frac{\sum(Y_i-\mu)^2}{2\phi^2} - \frac{n}{2\phi})^3\} \\
&= \frac{1}{n}E\{\frac{[\sum(Y_i-\mu)^2]^3}{8\phi^6} - \frac{n^3}{8\phi^3} - 3\frac{[\sum(Y_i-\mu)^2]^2}{4\phi^4}\frac{n}{2\phi} + 3\frac{\sum(Y_i-\mu)^2}{2\phi^2}\frac{n^2}{4\phi^2}\} \\
&= \frac{1}{8n\phi^6}E\{[\sum(Y_i-\mu)^2]^3\} - \frac{1}{n}\frac{n^3}{8\phi^3} - \frac{1}{n}\frac{3n}{8\phi^5}E\{[\sum(Y_i-\mu)^2]^2\} \\
&\quad + \frac{1}{n}\frac{3n^2}{8\phi^4}E\{\sum(Y_i-\mu)^2\} \\
&= \frac{1}{8n\phi^6}E\{[(n-1)s^2 + n(\bar{Y}-\mu)^2]^3\} - \frac{n^2}{8\phi^3} - \frac{3}{8\phi^5}E\{[(n-1)s^2 \\
&\quad + n(\bar{Y}-\mu)^2]^2\} + \frac{3n}{8\phi^4}E\{(n-1)s^2 + n(\bar{Y}-\mu)^2\} \\
&= \frac{1}{8n\phi^6}E\{(n-1)^3s^6 + n^3(\bar{Y}-\mu)^6 + 3(n-1)^2s^4n(\bar{Y}-\mu)^2 \\
&\quad + 3(n-1)s^2n^2(\bar{Y}-\mu)^4\} - \frac{n^2}{8\phi^3} - \frac{3}{8\phi^5}E\{(n-1)^2s^4 + n^2(\bar{Y}-\mu)^4 \\
&\quad + 2(n-1)s^2n(\bar{Y}-\mu)^2\} + \frac{3n}{8\phi^4}(n-1)E(s^2) + \frac{3n}{8\phi^4}nE(\bar{Y}-\mu)^2 \\
&= \frac{(n-1)^3}{8n\phi^6}E(s^6) + \frac{n^3}{8n\phi^6}E(\bar{Y}-\mu)^6 + \frac{3n(n-1)^2}{8n\phi^6}E(s^4)E(\bar{Y}-\mu)^2 \\
&\quad + \frac{3n^2(n-1)}{8n\phi^6}E(s^2)E(\bar{Y}-\mu)^4 - \frac{n^2}{8\phi^3} - \frac{3(n-1)^2}{8\phi^5}E(s^4) \\
&\quad - \frac{3n^2}{8\phi^5}E(\bar{Y}-\mu)^4 - \frac{6n(n-1)}{8\phi^5}E(s^2)E(\bar{Y}-\mu)^2 \\
&\quad + \frac{3n(n-1)}{8\phi^4}\phi + \frac{3n^2}{8\phi^4}\frac{\phi}{n} \\
&= \frac{(n-1)^3}{8n\phi^6}\frac{\phi^3}{(n-1)^2}(n+3)(n+1) + \frac{n^3}{8n\phi^6}\frac{15\phi^3}{n^3} \\
&\quad + \frac{3n(n-1)^2}{8n\phi^6}\frac{\phi^2}{(n-1)}(n+1)\frac{\phi}{n} + \frac{3n^2(n-1)}{8n\phi^6}\phi\frac{3\phi^2}{n^2} - \frac{n^2}{8\phi^3} \\
&\quad - \frac{3(n-1)^2}{8\phi^5}\frac{\phi^2}{(n-1)}(n+1) - \frac{3n^2}{8\phi^5}\frac{3\phi^2}{n^2} - \frac{6n(n-1)}{8\phi^5}\phi\frac{\phi}{n} + \frac{3n(n-1)}{8\phi^3} + \frac{3n}{8\phi^3} \\
&= \frac{(n^2-1)(n+3)+15}{8n\phi^3} + \frac{3(n^2-1)}{8n\phi^3} + \frac{9(n-1)}{8n\phi^3} - \frac{n^2}{8\phi^3} - \frac{3(n^2-1)}{8\phi^3} \\
&\quad - \frac{9}{8\phi^3} - \frac{6(n-1)}{8\phi^3} + \frac{3n^2-3n+3n}{8\phi^3} \\
&= \frac{n^3+3n^2-n-3+15+3n^2-3+9n-9}{8n\phi^3} + \frac{3n^2-n^2-3n^2+3-9-6n+6}{8\phi^3} \\
&= \frac{n^3+6n^2+8n+n(-n^2-6n)}{8n\phi^3} = \frac{n^3+6n^2+8n-n^3-6n^2}{8n\phi^3} = \frac{1}{\phi^3};
\end{aligned}
$$

$$\begin{aligned}
\kappa_{\mu,\mu\phi} &= n^{-1}E(U_\mu U_{\mu\phi}) \\
&= \frac{1}{n}E\{\frac{n(\bar{Y}-\mu)}{\phi}(-\frac{n(\bar{Y}-\mu)}{\phi^2})\} \\
&= \frac{1}{n}E\{-\frac{n^2(\bar{Y}-\mu)^2}{\phi^3}\} \\
&= -\frac{n^2}{n\phi^3}E(\bar{Y}-\mu)^2 \\
&= -\frac{n}{\phi^3}\frac{\phi}{n} = -\frac{1}{\phi^2};
\end{aligned}$$

$$\begin{aligned}
\kappa_{\phi,\phi\phi} &= n^{-1}E(U_\phi U_{\phi\phi}) \\
&= \frac{1}{n}E\{(\frac{\sum(Y_i-\mu)^2}{2\phi^2}-\frac{n}{2\phi})(\frac{n}{2\phi^2}-\frac{\sum(Y_i-\mu)^2}{\phi^3})\} \\
&= \frac{1}{n}E\{\frac{n\sum(Y_i-\mu)^2}{4\phi^4}-\frac{[\sum(Y_i-\mu)^2]^2}{2\phi^5}-\frac{n^2}{4\phi^3}+\frac{n\sum(Y_i-\mu)^2}{2\phi^4}\} \\
&= \frac{1}{n}E\{\frac{3n\sum(Y_i-\mu)^2}{4\phi^4}-\frac{[\sum(Y_i-\mu)^2]^2}{2\phi^5}-\frac{n^2}{4\phi^3}\} \\
&= \frac{1}{n}\frac{3n}{4\phi^4}E\{(n-1)s^2+n(\bar{Y}-\mu)^2\}-\frac{1}{2n\phi^5}E\{[(n-1)s^2+n(\bar{Y}-\mu)^2]^2\}-\frac{1}{n}\frac{n^2}{4\phi^3} \\
&= \frac{3(n-1)}{4\phi^4}E(s^2)+\frac{3n}{4\phi^4}E(\bar{Y}-\mu)^2-\frac{1}{2n\phi^5}E\{(n-1)^2s^4+n^2(\bar{Y}-\mu)^4 \\
&\quad +2(n-1)s^2n(\bar{Y}-\mu)^2\}-\frac{n}{4\phi^3} \\
&= \frac{3(n-1)}{4\phi^4}\phi+\frac{3n}{4\phi^4}\frac{\phi}{n}-\frac{(n-1)^2}{2n\phi^5}E(s^4)-\frac{n^2}{2n\phi^5}E(\bar{Y}-\mu)^4 \\
&\quad -\frac{2(n-1)n}{2n\phi^5}E(s^2)E(\bar{Y}-\mu)^2-\frac{n}{4\phi^3} \\
&= \frac{3(n-1)}{4\phi^3}+\frac{3}{4\phi^3}-\frac{(n-1)^2}{2n\phi^5}\frac{\phi^2}{n-1}(n+1)-\frac{n}{2\phi^5}\frac{3\phi^2}{n^2}-\frac{n-1}{\phi^5}\phi\frac{\phi}{n}-\frac{n}{4\phi^3} \\
&= \frac{3n-3+3-n}{4\phi^3}-\frac{(n-1)(n+1)}{2n\phi^3}-\frac{3}{2n\phi^3}-\frac{n-1}{n\phi^3} \\
&= \frac{2n}{4\phi^3}-\frac{n^2-1}{2n\phi^3}-\frac{3}{2n\phi^3}-\frac{n-1}{n\phi^3} \\
&= \frac{n^2-n^2+1-3-2(n-1)}{2n\phi^3} \\
&= \frac{-2-2n+2}{2n\phi^3} = -\frac{1}{\phi^3};
\end{aligned}$$

$$\begin{aligned}\kappa_{\mu,\mu,\mu} &= n^{-1}E(U_\mu^3)\\ &= \frac{1}{n}E\{(\frac{\sum Y_i - n\mu}{\phi})^3\}\\ &= \frac{1}{n}E\{\frac{n^3(\bar{Y}-\mu)^3}{\phi^3}\}\\ &= \frac{n^3}{n\phi^3}E(\bar{Y}-\mu)^3 = 0;\end{aligned}$$

$$\begin{aligned}\kappa_{\mu,\phi,\phi} &= n^{-1}E(U_\mu U_\phi^2)\\ &= \frac{1}{n}E\{\frac{n(\bar{Y}-\mu)}{\phi}(\frac{\sum(Y_i-\mu)^2}{2\phi^2} - \frac{n}{2\phi})^2\}\\ &= \frac{1}{n}E\{\frac{n(\bar{Y}-\mu)}{\phi}(\frac{[\sum(Y_i-\mu)^2]^2}{4\phi^4} + \frac{n^2}{4\phi^2} - 2\frac{n\sum(Y_i-\mu)^2}{4\phi^3})\}\\ &= \frac{1}{4\phi^5}E(\bar{Y}-\mu)E[\sum(Y_i-\mu)^2]^2 + \frac{n^2}{4\phi^3}E(\bar{Y}-\mu)\\ &\quad - \frac{n}{2\phi^4}E(\bar{Y}-\mu)E[\sum(Y_i-\mu)^2] = 0;\end{aligned}$$

$$\begin{aligned}\kappa_{\mu,\mu\mu} &= n^{-1}E(U_\mu U_{\mu\mu})\\ &= \frac{1}{n}E\{\frac{n(\bar{Y}-\mu)}{\phi}(-\frac{n}{\phi})\}\\ &= \frac{1}{n}E\{-\frac{n^2}{\phi^2}(\bar{Y}-\mu)\}\\ &= -\frac{n^2}{n\phi^2}E(\bar{Y}-\mu) = 0;\end{aligned}$$

$$\begin{aligned}\kappa_{\mu,\phi\phi} &= n^{-1}E(U_\mu U_{\phi\phi})\\ &= \frac{1}{n}E\{\frac{n(\bar{Y}-\mu)}{\phi}(\frac{n}{2\phi^2} - \frac{\sum(Y_i-\mu)^2}{\phi^3})\}\\ &= \frac{1}{n}E\{\frac{n^2(\bar{Y}-\mu)}{2\phi^3} - \frac{n(\bar{Y}-\mu)\sum(Y_i-\mu)^2}{\phi^4}\}\\ &= \frac{n^2}{2n\phi^3}E(\bar{Y}-\mu) - \frac{n}{n\phi^4}E(\bar{Y}-\mu)E\{\sum(Y_i-\mu)^2\} = 0;\end{aligned}$$

$$\begin{aligned}
\kappa_{\phi,\mu\mu} &= n^{-1}E(U_\phi U_{\mu\mu}) \\
&= \frac{1}{n}E\{(\frac{\sum(Y_i-\mu)^2}{2\phi^2}-\frac{n}{2\phi})(-\frac{n}{\phi})\} \\
&= \frac{1}{n}E\{-\frac{n}{2\phi^3}\sum(Y_i-\mu)^2+\frac{n^2}{2\phi^2}\} \\
&= \frac{1}{n}(-\frac{n}{2\phi^3})E\{\sum(Y_i-\mu)^2\}+\frac{1}{n}\frac{n^2}{2\phi^2} \\
&= -\frac{1}{2\phi^3}E\{(n-1)s^2+n(\bar{Y}-\mu)^2\}+\frac{n}{2\phi^2} \\
&= -\frac{n-1}{2\phi^3}E(s^2)-\frac{n}{2\phi^3}E(\bar{Y}-\mu)^2+\frac{n}{2\phi^2} \\
&= -\frac{(n-1)\phi}{2\phi^3}-\frac{n\phi}{2n\phi^3}+\frac{n}{2\phi^2} \\
&= \frac{1-n+1}{2\phi^2}+\frac{n}{2\phi^2}=0;
\end{aligned}$$

$$\begin{aligned}
\kappa_{\phi,\phi\mu} &= n^{-1}E(U_\phi U_{\phi\mu}) \\
&= \frac{1}{n}E\{(\frac{\sum(Y_i-\mu)^2}{2\phi^2}-\frac{n}{2\phi})(-\frac{n(\bar{Y}-\mu)}{\phi^2})\} \\
&= \frac{1}{n}(-\frac{n}{2\phi^4})E\{\sum(Y_i-\mu)^2\}E(\bar{Y}-\mu)+\frac{1}{n}\frac{n^2}{2\phi^3}E(\bar{Y}-\mu)=0.
\end{aligned}$$

Le due modificazioni alternative per $U(\mu,\phi)$ sono calcolate come

$$A_\mu^{(E)}=0, \quad A_\phi^{(E)}=\frac{1}{2\phi},$$

$$A_\mu^{(O)}=0, \quad A_\phi^{(O)}=\frac{s(\mu)}{n\phi^2}-\frac{1}{2\phi}.$$

La soluzione di $U+A^{(E)}=0$ è $\phi^*=\frac{s(\bar{y})}{n-1}$, mentre l'equazione alternativa $U+A^{(O)}=0$ produce $\phi^*=\frac{(n+2)s(\bar{y})}{n(n-1)}$. Il primo di questi stimatori è esattamente non distorto.

Entrambi gli stimatori sono efficienti al secondo ordine (Pace e Salvan, 1996, § 9.4.3) e presentano varianza pari a $\frac{2\phi^2(n+1)}{n^2}+O(n^{-3})$.

3.3 Reciproco della media di una distribuzione di Poisson

Sia $(y_1, \ldots, y_n)$ un campione casuale di osservazioni estratte indipendentemente da una distribuzione di Poisson con media μ, e si ponga l'attenzione in $\phi = \frac{1}{\mu}$; ciò potrebbe presentarsi, ad esempio, in connessione con l'analisi di dati di conteggio da un processo di Poisson dove ϕ è la media del tempo tra gli eventi.
La funzione di probabilità di Y_i è

$$p_i(y_i) = \frac{e^{-\mu}\mu^{y_i}}{y_i!}$$

da cui si ottengono rispettivamente la funzione di verosimiglianza e la funzione di log-verosimiglianza:

$$L(\mu) = \prod_{i=1}^{n} \frac{e^{-\mu}\mu^{y_i}}{y_i!} = e^{-n\mu}\mu^{\sum_{i=1}^{n} y_i} \prod_{i=1}^{n} \frac{1}{y_i!};$$

$$l(\mu) = -n\mu + \log\mu \sum_{i=1}^{n} y_i.$$

Usando ϕ:

$$l(\phi) = -\frac{n}{\phi} + \log\frac{1}{\phi} \sum_{i=1}^{n} y_i = -\frac{n}{\phi} - \log\phi \sum_{i=1}^{n} y_i,$$

e, sapendo che $\sum_{i=1}^{n} y_i = y_.$, la funzione di punteggio è

$$U(\phi) = \frac{\partial l(\phi)}{\partial \phi} = \frac{n}{\phi^2} - \frac{y_.}{\phi}.$$

L'informazione osservata è

$$I(\phi) = -U'(\phi) = -(-\frac{2n\phi}{\phi^4} + \frac{y_.}{\phi^2}) = \frac{2n}{\phi^3} - \frac{y_.}{\phi^2},$$

mentre l'informazione attesa è

$$i(\phi) = E[I(\phi)] = E(\frac{2n}{\phi^3} - \frac{y_.}{\phi^2}) = \frac{2n}{\phi^3} - \frac{1}{\phi^2}\sum_{i=1}^{n} E(y_i) = \frac{2n}{\phi^3} - \frac{n}{\phi^3} = \frac{n}{\phi^3}.$$

La stima di massima verosimiglianza di ϕ, indicata con $\hat{\phi}$, si ottiene ponendo la funzione di punteggio uguale a zero:

$$U(\phi) = 0 \ \rightarrow \ \frac{n - y_{.}\phi}{\phi^2} = 0 \ \rightarrow \ \hat{\phi} = \frac{n}{y_{.}}.$$

Ricordando che $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n} y_{.}$ si ha $y_{.} = n\bar{y}$, da cui

$$\hat{\phi} = \frac{1}{\bar{y}}.$$

La formula generale per il calcolo della distorsione asintotica è

$$E[g(\bar{Y})] = E[g(\mu) + g^{'}(\mu)(\bar{Y} - \mu) + \frac{1}{2} g^{''}(\mu)(\bar{Y} - \mu)^2 + \ldots] \doteq g(\mu) + \frac{1}{2} g^{''}(\mu)\frac{\sigma^2}{n}.$$

Dunque, se $g(x) = \frac{1}{x}$, $g^{'}(x) = -\frac{1}{x^2}$ e $g^{''}(x) = \frac{2x}{x^4} = \frac{2}{x^3}$.
Nel caso della distribuzione Poisson la distorsione asintotica è

$$E(\hat{\phi}) = \frac{1}{\mu} + \frac{1}{2}\frac{2}{\mu^3}\frac{\mu}{n} = \phi + \frac{\phi^2}{n} + O(n^{-2})$$

con $\frac{b_1(\phi)}{n} = \frac{\phi^2}{n}$.
Le due modificazioni alternative per $U(\phi)$ sono calcolate come

$$A^{(E)}(\phi) = -i(\phi)\frac{b_1(\phi)}{n} = -\frac{n}{\phi^3}\frac{\phi^2}{n} = -\frac{1}{\phi};$$

$$A^{(O)}(\phi) = -I(\phi)\frac{b_1(\phi)}{n} = (\frac{y_{.}}{\phi^2} - \frac{2n}{\phi^3})\frac{\phi^2}{n} = \frac{y_{.}}{n} - \frac{2}{\phi} = \bar{y} - \frac{2}{\phi}.$$

Con $A^{(E)}(\phi)$:

$$U^*(\phi) = U(\phi) + A^{(E)}(\phi) = \frac{n}{\phi^2} - \frac{n\bar{y}}{\phi} - \frac{1}{\phi} = \frac{n}{\phi^2} - \frac{n\bar{y} + 1}{\phi},$$

quindi, ponendo $U^*(\phi) = 0$, si ha

$$\frac{n}{\phi^2} - \frac{n\bar{y} + 1}{\phi} = 0$$

da cui

$$\phi(n\bar{y} + 1) = n \ \rightarrow \ \phi^* = \frac{n}{n\bar{y} + 1} = \frac{1}{\bar{y} + \frac{1}{n}}.$$

Con $A^{(O)}(\phi)$:

$$U^*(\phi) = U(\phi) + A^{(O)}(\phi) = \frac{n}{\phi^2} - \frac{n\bar{y}}{\phi} + \bar{y} - \frac{2}{\phi} = \frac{n - n\bar{y}\phi + \bar{y}\phi^2 - 2\phi}{\phi^2},$$

quindi, ponendo $U^*(\phi) = 0$, si ha

$$\frac{n - n\bar{y}\phi + \bar{y}\phi^2 - 2\phi}{\phi^2} = 0$$

da cui

$$\phi^* = \begin{cases} n\{\bar{y} + \frac{2}{n} - \sqrt{\bar{y}^2 + \frac{4}{n^2}}\}/(2\bar{y}) & \text{se } \bar{y} > 0 \\ \frac{n}{2} & \text{se } \bar{y} = 0 \end{cases}.$$

Entrambi questi stimatori sono finiti per tutti i campioni, hanno distorsione pari a $O(n^{-2})$ e sono efficienti al secondo ordine (Pace e Salvan, 1996, § 9.4.3) con varianza $\frac{1}{n\mu^3} + 2(n^2\mu^4) + O(n^{-3})$.

3.4 Tasso di guasto di una distribuzione esponenziale

Siano $y_1, \ldots, y_n$ realizzazioni di variabili casuali indipendenti $Y_i \sim Esp(\lambda)$, con tasso di guasto $\lambda > 0$.
La funzione di densità di Y_i è

$$p_i(y_i) = \lambda e^{-\lambda y_i}, \quad y \geq 0$$

da cui si ottengono rispettivamente la funzione di verosimiglianza e la funzione di log-verosimiglianza:

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda y_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} y_i};$$

$$l(\lambda) = n \log \lambda - \lambda \sum_{i=1}^{n} y_i.$$

Derivando rispetto a λ la funzione di log-verosimiglianza si ottiene la funzione di punteggio

$$U(\lambda) = \frac{\partial l(\lambda)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i$$

e, ricordando che $\frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$, da cui $\sum_{i=1}^{n} y_i = n\bar{y}$, la funzione di punteggio diventa

$$U(\lambda) = \frac{n}{\lambda} - n\bar{y}.$$

L'informazione osservata e l'informazione attesa coincidono:

$$I(\lambda) = -U^{'}(\lambda) = -(-\frac{n}{\lambda^2}) = \frac{n}{\lambda^2};$$

$$i(\lambda) = E[I(\lambda)] = E(\frac{n}{\lambda^2}) = \frac{n}{\lambda^2}.$$

Ponendo $U(\lambda) = 0$ si ottiene la stima di massima verosimiglianza $\hat{\lambda}$ di λ:

$$\frac{n}{\lambda} - n\bar{y} = 0 \ \rightarrow \ n\lambda\bar{y} = n \ \rightarrow \ \hat{\lambda} = \frac{1}{\bar{y}}.$$

La formula generale per il calcolo della distorsione asintotica è

$$E[g(\bar{Y})] = E[g(\mu) + g^{'}(\mu)(\bar{Y} - \mu) + \frac{1}{2}g^{''}(\mu)(\bar{Y} - \mu)^2 + \ldots] \doteq g(\mu) + \frac{1}{2}g^{''}(\mu)\frac{\sigma^2}{n}.$$

Dunque, se $g(x) = \frac{1}{x}$, $g^{'}(x) = -\frac{1}{x^2}$ e $g^{''}(x) = \frac{2x}{x^4} = \frac{2}{x^3}$.
Nel caso della distribuzione esponenziale la distorsione asintotica è

$$E(\hat{\lambda}) = \frac{1}{\mu} + \frac{1}{2}\frac{2}{\mu^3}\frac{\sigma^2}{n},$$

ma è noto che $\mu = \frac{1}{\lambda}$ e $\sigma^2 = \frac{1}{\lambda^2}$, quindi

$$E(\hat{\lambda}) = \lambda + \lambda^3\frac{\frac{1}{\lambda^2}}{n} = \lambda + \frac{\lambda}{n} + O(n^{-2})$$

con $\frac{b_1(\lambda)}{n} = \frac{\lambda}{n}$.
Poiché l'informazione osservata e l'informazione attesa coincidono, anche le due modificazioni $A^{(E)}(\lambda)$ e $A^{(O)}(\lambda)$ sono uguali:

$$A^{(E)}(\lambda) = -i(\lambda)\frac{b_1(\lambda)}{n} = -\frac{n}{\lambda^2}\frac{\lambda}{n} = -\frac{1}{\lambda};$$

$$A^{(O)}(\lambda) = -I(\lambda)\frac{b_1(\lambda)}{n} = -\frac{n}{\lambda^2}\frac{\lambda}{n} = -\frac{1}{\lambda}.$$

Indicando $A^{(E)}(\lambda) = A^{(O)}(\lambda) = A(\lambda)$,

$$U^*(\lambda) = U(\lambda) + A(\lambda) = \frac{n}{\lambda} - n\bar{y} - \frac{1}{\lambda} = \frac{n-1}{\lambda} - n\bar{y},$$

da cui si ottiene λ^* ponendo $U^*(\lambda) = 0$:

$$\frac{n-1}{\lambda} - n\bar{y} = 0$$

$$n - 1 - n\bar{y}\lambda = 0$$

$$n\bar{y}\lambda = n - 1$$

$$\lambda^* = \frac{n-1}{n\bar{y}} = \frac{1 - \frac{1}{n}}{\bar{y}}.$$

Lo stimatore λ^* ha distorsione $O(n^{-2})$ (è stato quindi rimosso il termine $\frac{b_1(\lambda)}{n}$) ed è efficiente al secondo ordine (Pace e Salvan, 1996, § 9.4.3).

Capitolo 4

Applicazioni recenti

In quest'ultimo capitolo vengono dapprima descritti i principali campi di applicazione del metodo proposto da Firth (1993) per la riduzione della distorsione dello stimatore di massima verosimiglianza. Successivamente vengono descritti alcuni esempi di utilizzo del metodo nella letteratura dell'ultimo decennio.

4.1 Panoramica d'insieme

I campi di applicazione del metodo di Firth (1993) sono molti e molto diversi tra loro.
Primo fra tutti, la statistica metodologica, che se ne serve per l'analisi dei modelli lineari e dei modelli lineari generalizzati e per l'analisi della regressione.
In secondo luogo, la statistica medica e l'epidemiologia che ad esempio utilizzano il metodo discusso nel Capitolo 2 in studi di particolari patologie tra cui il delirio (Voyer et al., 2008) e per comprendere le interazioni tra parassiti (Johnson e Buller, 2011).
Altre applicazioni del metodo di Firth (1993) si trovano in sociologia, per lo studio delle origini sociali e l'istruzione (Van de Werfhorst e Luijkx, 2010).
Ancora, il metodo di Firth (1993) è stato utilizzato per l'analisi di dati sul-

l'ambiente, come ad esempio gli effetti delle variazioni climatiche (Foufopoulos, Kilpatrick e Ives, 2010).
Il metodo di Firth (1993) viene utilizzato per lo più, non tanto per ottenere una riduzione della distorsione dello stimatore di massima verosimiglianza, quanto piuttosto come soluzione al problema della separazione. Si tratta di un problema comunemente riscontrato nei modelli con variabili dipendenti dicotomiche. Il fenomeno della separazione si manifesta con stime dei parametri non finite (si veda ad esempio, Heinze e Schemper, 2002).
L'approccio di Firth (1993) produce stime finite dei parametri, equivalendo all'impiego di una penalizzazione della verosimiglianza, e rappresenta dunque una soluzione ideale al problema della separazione.

4.2 Comprendere le interazioni tra parassiti

In natura, gli animali ospiti sono esposti ad una 'mistura' di diversi parassiti che costituiscono una comunità dinamica all'interno dell'ospite (Pedersen e Fenton, 2007). Le interazioni tra parassiti contemporaneamente presenti possono avere effetti significativi sia sugli altri parassiti sia sull'individuo ospite, dimostrando così l'importanza di conoscere le relazioni tra agenti patogeni in medicina veterinaria e in discipline attinenti (Johnson e Buller, 2011).
Per analizzare i meccanismi di interazione tra parassiti, Johnson e Buller (2011) hanno condotto un esperimento di laboratorio, in cui le larve degli anfibi sono state esposte al numero realistico di ogni parassita, prima individualmente e poi congiuntamente. Gli obiettivi erano quelli di confrontare gli effetti di ogni parassita e di esaminare come le loro interazioni influenzano il successo dell'infezione e lo sviluppo della patologia nell'ospite. Tramite la combinazione di metodi ed approcci sperimentali, Johnson e Buller (2011) hanno puntato inoltre ad incorporare dati provenienti da fonti diverse per comprendere le interazioni tra parassiti.

Utilizzando l'analisi parametrica della sopravvivenza, Johnson e Buller (2011) hanno analizzato dati relativi alla sopravvivenza dell'ospite, con individui che sono sopravvissuti per meno di dieci giorni (esperimento 1) e altri che sono sopravvissuti fino alla metamorfosi (esperimento 2), classificati come 'censurati'. Nel secondo esperimento, dati riguardanti la presenza di malformazioni (si o no) sono stati analizzati (Johnson e Buller, 2011) tramite l'uso di modelli lineari generalizzati, tenendo conto della correzione introdotta da Firth (1993) per la separazione. Con l'analisi della varianza ANOVA a due criteri di classificazione, Johnson e Buller (2011) hanno valutato gli effetti del trattamento in base alla grandezza, alla massa e al tempo necessario per la metamorfosi dell'ospite. Poiché non hanno riscontrato differenze tra basse esposizioni (40 esemplari) ed alte esposizioni (160 esemplari) dell'individuo ospite agli Echinostomi, le hanno considerate come appartenenti ad un unico gruppo di analisi. Per studiare infine la guarigione dai parassiti, Johnson e Buller (2011) hanno applicato due metodi: modelli lineari generalizzati con distribuzione di tipo Poisson e ANOVA.
I risultati ottenuti in questo studio rinforzano l'importanza delle interazioni tra parassiti nel determinare l'infezione, ma mettono anche in evidenza l'importanza del metodo nell'influire sul risultato di tali interazioni. Tramite numerosi esperimenti di manipolazione sono stati scoperti (Johnson e Buller, 2011) modelli robusti di associazione tra gli agenti patogeni Ribeiroia ed Echinostoma.

4.3 Casi di delirio tra gli infermieri

Il delirio è un problema diffuso tra gli infermieri che operano in strutture di assistenza in cui l'età avanzata ed il deficit cognitivo dei pazienti rappresentano due importanti fattori di rischio per tale patologia. Il delirio è spesso associato ad esiti negativi, tra cui l'aumento della morbilità e, in casi estremi, della mortalità. Gli obiettivi dello studio (Voyer et al., 2008)

sono quelli di determinare i sintomi della patologia e di identificare i fattori associati ai casi di delirio riscontrati tra gli infermieri.
In primo luogo, Voyer et al. (2008) hanno utilizzato delle semplici analisi descrittive della popolazione oggetto di studio (pazienti ed infermieri). Successivamente, hanno confrontato i tassi di delirio riscontrati tra gli infermieri con tassi di delirio standard, presi come riferimento. Infine, per stabilire se ci sono fattori effettivamente associati al delirio, hanno utilizzato l'analisi della regressione logistica.
Vale la pena notare che per le due variabili, tipo di delirio ed età degli infermieri, è stata osservata la quasi completa separazione. Per questo motivo, Voyer et al. (2008) hanno scelto di adattare ai dati un modello di regressione logistica con verosimiglianza penalizzata di Firth (1993). Tuttavia, gli intervalli di confidenza ottenuti hanno livello approssimato 0.95 e le loro probabilità di copertura possono essere diverse da quelle attese.

4.4 Origini sociali ed istruzione

In questo esempio, Van de Werfhorst e Luijkx (2010) esaminano la relazione esistente tra l'origine sociale e l'istruzione. Utilizzando dati della ricerca olandese relativi agli uomini, hanno trovato che i figli spesso scelgono campi di studio in cui riscontrano un'affinità con la classe sociale del padre. In questo modo, la selezione sociale nei campi di studio è guidata dall'ambito occupazionale del padre. Cosa molto importante, l'affinità degli ambiti occupazionali tra le generazioni ostacola la mobilità sociale intergenerazionale.
Poiché i dati utilizzati per lo studio sono piuttosto sparsi, Van de Werfhorst e Luijkx (2010) hanno seguito il metodo di Firth (1993) usando un aggiustamento per la riduzione della distorsione delle stime nei modelli log-lineari.

4.5 Effetti delle variazioni climatiche

I recenti cambiamenti climatici hanno costretto numerose specie a spostarsi verso i poli, ma solo pochi studi empirici hanno stabilito quali specie risulteranno maggiormente vulnerabili alle variazioni climatiche a lungo termine.
Per studiare le passate conseguenze dei cambiamenti climatici, Foufopoulos, Kilpatrick e Ives (2010) hanno calcolato i tassi di estinzione relativi agli ultimi 16 mila anni di 35 specie di rettili in 87 isole greche del Mediterraneo. I tassi di estinzione sono risultati maggiori per le specie che oggi presentano distribuzioni più settentrionali. Foufopoulos, Kilpatrick e Ives (2010) hanno inoltre trovato che le specie del nord avevano a disposizione meno habitat idonei alle loro esigenze nelle isole sopra citate, il che ha contibuito a determinare elevati tassi di estinzione. Le estinzioni sono avvenute in un contesto di crescente frammentazione degli habitat, riduzione delle dimensioni delle isole ed innalzamento del livello del mare. Così, le circostanze affrontate dai rettili sulle isole greche sono molto simili alle sfide che numerose specie di oggi devono affrontare quando avvengono cambiamenti climatici.
Per condurre questo studio, piuttosto che utilizzare una funzione con legame logit, comunemente utilizzata nella regressione logistica (McCullagh e Nelder, 1989), Foufopoulos, Kilpatrick e Ives (2010) hanno scelto una funzione con legame logaritmico che ben si adatta alla forma esponenziale dei dati. Per ridurre la distorsione delle stime dei coefficienti hanno adattato al caso il metodo di Firth (1993).

Conclusioni

In questa tesi è stato considerato il metodo proposto da Firth (1993) utilizzato per ridurre la distorsione dello stimatore di massima verosimiglianza tramite un'opportuna modificazione della funzione di punteggio.
L'obiettivo principale era quello di analizzare la validità di tale metodo, evidenziando le differenze rispetto ai due tradizionali approcci per la riduzione della distorsione (*jackknife* e riduzione della distorsione tramite sviluppo in serie). In particolare, mentre i due metodi tradizionali risultano essere 'correttivi' anziché 'preventivi', il metodo di Firth (1993) procede attraverso una modificazione del meccanismo che produce la stima di massima verosimiglianza, cioè dell'equazione di verosimiglianza basata sulla funzione di punteggio, piuttosto che della stima stessa. Il vantaggio di tale approccio sembra spingersi oltre la riduzione della distorsione degli stimatori di massima verosimiglianza e va individuato soprattutto nel fornire, in talune situazioni problematiche, una funzione di verosimiglianza con massimo finito.
Nel Capitolo 1 sono stati richiamati alcuni concetti fondamentali dell'inferenza statistica, quali le definizioni di modello statistico parametrico, di stima e stimatore di un parametro e della funzione di verosimiglianza, soffermandosi in particolare sullo stimatore di massima verosimiglianza e sulle sue proprietà. Infine sono state introdotte le famiglie esponenziali monoparametriche e multiparametriche.
Nel Capitolo 2, dopo una descrizione dei due approcci tradizionali per la riduzione della distorsione dello stimatore di massima verosimiglian-

za, si è trattato il metodo proposto da Firth (1993). Come mostrato, nei problemi regolari, la distorsione asintotica dello stimatore di massima verosimiglianza può essere ridotta mediante la rimozione del termine dominante, ottenuta introducendo un'opportuna distorsione nella funzione di punteggio. Se il parametro di interesse è il parametro canonico di una famiglia esponenziale, ciò equivale ad utilizzare la distribuzione a priori di Jeffreys come funzione di penalità per la verosimiglianza. Nel caso di altre parametrizzazioni, sono disponibili varie correzioni, ottenute utilizzando l'informazione attesa o l'informazione osservata. Al di fuori della famiglia esponenziale, l'uso dell'informazione attesa rispetto all'uso dell'informazione osservata comporta una perdita di efficienza.
Nel Capitolo 3 la teoria esposta nei capitoli precedenti è stata adattata ad alcuni esempi. I risultati ottenuti nei §§ 3.2, 3.3 e 3.4 dimostrano che il metodo di Firth (1993) ha successo nella riduzione della distorsione dello stimatore di massima verosimiglianza, rimuovendo il termine $O(n^{-1})$ tramite l'introduzione di una piccola distorsione nella funzione di punteggio.
Nel Capitolo 4, infine, si è fornito un quadro generale dei principali campi di applicazione del metodo di Firth (1993).

Riferimenti bibliografici

Azzalini, A. (2001). *Inferenza Statistica - Una Presentazione Basata sul Concetto di Verosimiglianza*. 2nd ed. Springer Verlag, Milano.

Copas, J.B. (1988). Binary regression models for contaminated data. *J. R. Statist. Soc.*, **B 50**, 225-265.

Cordeiro, G.M. e McCullagh, P. (1991). Bias correction in generalized linear models. *J. R. Statist. Soc.*, **B 53**, 629-643.

Cox, D.R. e Hinkley, D.V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

Cox, D.R. e Snell, E.J. (1989). *Analysis of Binary Data*. 2nd ed. London: Chapman and Hall.

Firth, D. (1992). Bias reduction, the Jeffreys prior and GLIM. In *Advances in GLIM and Statistical Modelling*, Ed. L. Fahrmeir, B. Francis, R. Gilchrist and G. Tutz, 91-100. New York: Springer-Verlag.

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

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc. London, Series A*, **222**, 309-368.

Fisher, R.A. (1925). Theory of statistical estimation. *Proc. Cambrige Phil. Soc.*, **22**, Pt. 5, 309-368.

Foufopoulos, J., Kilpatrick, A.M. e Ives, A.R. (2011). Climate Change and Elevated Extinction Rates of Reptiles from Mediterranean Islands. *The American Naturalist*, **177.1**, 119-129.

Heinze, G. e Schemper, M. (2002). A solution to the problem of separation in logistic regression. *Statist. Med.*, **21**, 2409-2419.

Johnson, P.T.J. e Buller, I.D. (2011). Parasite competition hidden by correlated coinfection: using surveys and experiments to understand parasite interactions. *Ecology*, **92(3)**, 535-541.

McCullagh, P. (1987). *Tensor Methods in Statistics*. London: Chapman and Hall.

McCullagh, P. e Nelder, J.A. (1989). *Generalized Linear Models*. 2nd ed. London: Chapman and Hall.

Naddeo, A. (1963). *La Teoria dei Test Statistici*. Giuffrè, Milano.

Pace, L. e Salvan, A. (2001). *Introduzione alla Statistica II - Inferenza, Verosimiglianza, Modelli*. Cedam, Padova.

Pace, L. e Salvan, A. (1996). *Teoria della Statistica: Metodi, Modelli, Approssimazioni Asintotiche*. Cedam, Padova.

Pedersen, A.B. e Fenton, A. (2007). Emphasizing the ecology in parasite community ecology. *Trends in Ecology and Evolution*, **21**, 133-139.

Piccolo, D. (2006). *Statistica per le Decisioni - La Conoscenza Umana Sostenuta dall'Evidenza Empirica*. Bologna, Il Mulino.

Quenouille, M.H. (1949). Approximate tests of correlation in time-series. *J. R. Statist. Soc.*, **B 11**, 68-84.

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

Van de Werfhorst, H.G. e Luijkx, R. (2010). Educational Field of Study and Social Mobility: Disaggregating Social Origin and Education. *Sociology*, **44**, 695-715.

Voyer, P., Richard, S., Doucet, L., Danjou, C. e Carmichael, P.H. (2008). Detection of delirium by nurses among long-term care residents with dementia. *BMC Nursing*, **7**, 4.