UNIVERSITÀ DEGLI STUDI DI PADOVA
FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA IN SCIENZE STATISTICHE
DEMOGRAFICHE E SOCIALI

TESI DI LAUREA

EQUAZIONI DI STIMA E METODI DI SELEZIONE DEL MODELLO STATISTICO

RELATORE: PROF. LAURA VENTURA

LAUREANDO: NICOLA LUNARDON

ANNO ACCADEMICO 2007/2008

Indice

Introduzione

La dicitura "selezione del modello" può essere sintetizzata in un insieme di procedure inferenziali volte a valutare quale sia il miglior modello statistico per l'insieme di dati alla mano e per le osservazioni future. Questo modo di pensare è, presumibilmente, dovuto ad Akaike che, esportando il concetto di entropia dalla fisica e adattandolo alla statistica, nel 1973 propose il primo criterio di informazione su cui basare la scelta del modello statistico (Akaike, 1973). In rapida successione, all'*Akaike Information Criterion* (AIC) si sono aggiunti altri criteri di informazione direttamente collegati o totalmente svincolati dalle radici dell'AIC; questi sono, rispettivamente, il *Takeuchi Information Criterion* (Takeuchi, 1976) e il *Bayesian Information Criterion* (Schwarz, 1978). Negli anni si sono aggiunte numerose varianti di questi criteri, come ad esempio modificazioni dell'AIC per piccoli campioni. Nonostante tutto, la sostanza di questi criteri resta immutata da allora: sono criteri di selezione del modello in grado di confrontare solo modelli parametrici. Questa può sembrare una lacuna di poco conto in quanto le analisi di *routine* contemplano usualmente modelli parametrici. Tuttavia, non è raro che vi siano delle situazioni di interesse applicativo che obbligano l'analista a percorrere strade alterntive ai metodi parametrici.

L'argomento trattato in questa tesi prende spunto da uno studio condotto in ambito medico (si veda Fassina *et al.*, 2008) che ha posto il seguente problema: modellando dati di conteggio in possibile presenza di sovradispersione, si sono ottenuti due modelli soddisfacenti, uno era parametrico e l'altro semiparametrico. I criteri per la selezione del modello sopra enunciati non sono di nessun aiuto in questa circostanza. Scegliere tra questi due modelli

non è semplice in quanto è necessario ricorrere alla teoria delle equazioni di stima

L'argomento di questa tesi è pertanto focalizzato sui metodi di selezione del modello basati su equazioni di stima. Come è noto, le procedure inferenziali basate su equazioni di stima hanno il pregio di essere più flessibili, in quanto permettono di alleggerire gli assunti parametrici che una funzione di verosimiglianza richiede. Come conseguenza, i risultati delle procedure inferenziali basate su equazioni di stima continuano a valere sotto condizioni più ampie di quelle richieste dalle dalla teoria basata sulla verosimiglianza.

Purtroppo, il grande studio dedicato alle proprietà delle equazioni di stima non è stato altrettanto intenso e sistematico, come denunciato dagli stessi autori, nel fornire criteri di selezione del modello ampiamente studiati e validati. Allo stato attuale, i metodi di selezione del modello basati su equazioni di stima sono definiti mimando i metodi basati sulla minimizzazione di opportune distanze tra i modelli messi a confronto. Inoltre, non sono disponibili in letteratura studi di simulazione completi e intensivi che velidano tali criteri, almeno dal punto di vista empirico.

L'obiettivo principale di questa tesi consiste nel condurre studi di simulazione sui criteri di selezione del modello basati su equazioni di stima, per supplire alla carenza di studi presente in letteratura e investigare il loro grado di attendibilità. In particolare, la tesi concentrerà lo studio sul il criterio di selezione del modello discusso da Pan (Pan, 2001b) nell'ambito dei modelli lineari generalizzati e quello discusso da Ronchetti (Ronchetti, 1985) nell'ambito dell'inferenza robusta. Entrambi i metodi fanno uso di una funzione di verosimiglianza alternativa basata su equazioni di stima, ossia la quasi-verosimiglianza, che viene utilizzata al posto della verosimiglianza propria.

La struttura della tesi è la seguente.

Il Capitolo 1 introduce l'elemento principale su cui si basano l'AIC e il TIC, ovvero la distanza di Kullback-Leibler. Inoltre, verrà fatta una breve rassegna sui criteri di selezione del modello classici (AIC, TIC e BIC).

Il Capitolo 2 introduce le equazioni di stima e le loro principali propri-

età. Verranno anche introdotti i modelli lineari generalizzati (GLM), alcuni concetti di robustezza e la funzione di quasi-verosimiglianza in cui la teoria delle equazioni di stima gioca un ruolo fondamentale.

Il Capitolo 3 presenta i principali metodi di selezione del modello basati su equazioni di stima presenti in letteratura. In particolare, si darà più visibilità al criterio basato sulla quasi-verosimiglianza e alla versione robusta del TIC.

Il Capitolo 4 esporrà gli studi di simulazione e discuterà i risultati ottenuti. In particolare, si mostrerà perché alcuni criteri presentati al Capitolo 3 non sono degli strumenti affidabili nella selezione del modello quando si lavora, invece che con la funzione di verosimiglianza propria, con la pseudo-verosimiglianza.

Capitolo 1

Metodi tradizionali di selezione del modello statistico

In questo capitolo si presentano alcune note tecniche di selezione del modello basate sulla verosimiglianza, quali l'*AIC* (*Akaike Information Criterion*, Akaike, 1973), il *TIC* (*Takeuchi Information Criterion*, Takeuchi, 1976) e il *BIC* (*Bayesian Information Criterion*, Schwarz, 1978). Per la comprensione delle origini dei primi due metodi è essenziale, innanzitutto, introdurre il criterio di informazione di Kullback-Leibler (Kullback e Leibler, 1951, Kullback, 1997).

Qualche precisazione sulla notazione utilizzata è doverosa. Nel seguito si indicherà con $y = (y_1, \ldots, y_n)$ il campione costituito da n osservazioni indipendenti e identicamente distribuite (i.i.d.), realizzazione di una variabile casuale con modello probabilistico effettivo $f_0 = f_0(y)$. Si indicheranno inoltre con $f_\vartheta = f(y;\vartheta)$ e con $\vartheta \in \Theta \subseteq \mathbb{R}^k$, rispettivamente, gli elementi del modello statistico $\mathscr{F}$ specificato per y e il parametro che indicizza il modello. Infine, si indicherà con $E_{f_0}(\cdot)$ il valore atteso rispetto alla distribuzione f_0 di y.

Se la funzione di densità di probabilità che presiede alla generazione dei dati è un elemento di $\mathscr{F}$, ossia se $f_0 \in \mathscr{F}$, si dice che il modello statistico è correttamente specificato. In tal caso si ha $f_0(y) = f(y;\vartheta_{f_0})$ per un valore $\vartheta_{f_0} \in \Theta$, detto vero valore del parametro.

Se invece l'effettivo modello generatore dei dati, f_0, non appartiene al modello statistico $\mathscr{F}$, si dice che il modello è scorrettamente specificato, nel senso che, con probabilità positiva, $f_0(y) \neq f(y;\vartheta)$ per ogni $\vartheta \in \Theta$.

Con il 'termine selezione del modello' si intende segliere quel modello f_ϑ che meglio si adatta ai dati; ciò implica il dover scegliere da $\mathscr{F}$ un suo elemento, contraddistinto da un preciso valore del parametro, indicato con $\vartheta_0 \in \Theta$. In corrispondenza del campione osservato $y = (y_1 \dots y_n)$ è possibile ottenere la stima di massima verosimiglinaza (MLE) di ϑ, indicata con $\hat{\vartheta} = \hat{\vartheta}(y)$. Se la situazione è regolare (si veda Huber, 1967), $\hat{\vartheta}$ converge in probabilità al valore ϑ_0, per il quale

$$E_{f_0}\left[\log f(Y;\vartheta_0)\right] > E_{f_0}\left[\log f(Y;\vartheta)\right], \tag{1.1}$$

per ogni $\vartheta \in \Theta$, per cui $\vartheta \neq \vartheta_0$, ossia a quell'elemento della famiglia $\mathscr{F}$ che soddisfa la disuguaglianza di Wald quando il valore atteso è valutato rispetto alla vera distribuzione

1.1 La distanza di Kullback-Leibler

La teoria dell'informazione è stata sviluppata ed è presente in molte branche della scienza. Non è casuale che il criterio d'informazione (chiamata anche distanza o discrepanza) di Kullback-Leibler (KL) sia strettamente legato alla definizione di entropia di Boltzmann (Burnham e Anderson, 2002, cap.2; Akaike, 1995, cap.1), applicata alla fisica. Infatti, a posteriori, il criterio di KL si è mostrato essere l'entropia negativa di Boltzmann.

La distanza di KL tra le densità $f_0(y)$ e $f(y;\vartheta)$ è definita come

$$KL(f_0; f_\vartheta) = \int \log\left(\frac{f_0(y)}{f(y;\vartheta)}\right) f_0(y)dy = E_{f_0}\left[\log\left(\frac{f_0(y)}{f(y;\vartheta)}\right)\right]. \tag{1.2}$$

La (1.2), pur non essendo una distanza in senso proprio (non è simmetrica), è una misura di dissomiglianza tra i due modelli valutata sull'intero supporto e quantifica l'ammontare di informazione persa quando si utilizza $f(y;\vartheta)$ al posto di $f_0(y)$. In particolare, $KL(f_0; f_\vartheta) > 0$ a meno che f_0 e f_ϑ non siano essenzialmente equivalenti, ovvero $f(y;\vartheta) = f_0(y)$.

Si noti che la (1.2) può equivalentemente essere scritta come

$$KL(f_0; f_\vartheta) = c - \int \log f(y;\vartheta) f_0(y) dy = c - E_{f_0}\left[\log f(y;\vartheta)\right], \qquad (1.3)$$

con $c = \int \log f_0(y)\ f_0(y) dy$ termine che dipende solamente da f_0.

Nell'ambito di un problema di selezione del modello, la (1.2) suggerisce di scegliere quel modello per cui $KL(f_0; f_\vartheta)$ sia minima. Sia ϑ_0 il valore del parametro che, per un dato modello, minimizza la (1.2). Come si vede, la (1.1) caratterizza ϑ_0 come l'elemento della famiglia che ha minima distanza di KL da f_0.

1.2 Il criterio AIC e TIC

La distanza di Kullback-Leibler e il criterio di informazione di Akaike per la selezione del modello sono strettamente legati benchè, concettualmente, differiscono enormemente. Precedentemente, è stato rimarcato come nel calcolo della (1.2) non sia utilizzata alcuna osservazione campionaria, in quanto la distanza di KL necessita la conoscenza della sola forma funzionale dei due modelli confrontati. Tuttavia tale situazione è irreale, sostanzialmente in quanto sia f_0 sia ϑ sono ignoti. Per questo la distanza di KL è un criterio concettualmente corretto, ma che richiede qualche modifica per rispondere alla domanda: "quale tra i possibili modelli è una buona approssimazione dell'ignoto meccanismo generatore dei dati?".

Si supponga di disporre di due campioni, $y^* = (y_1^*, \ldots, y_n^*)$ e $y = (y_1, \ldots, y_n)$, provenienti da f_0. Allora è possibile scrivere il rapporto di verosimiglianza atteso per il confronto tra f_0 e $f(y;\vartheta)$, in $\vartheta = \hat{\vartheta}(y)$, come

$$E_{f_0}^{y^*}\left[\sum_{i=1}^{n} \log\left(\frac{f_0(y_i^*)}{f(y_i^*;\hat{\vartheta})}\right)\right] = nKL(f_0; f_{\hat{\vartheta}}) \geq nKL(f_0; f_{\vartheta_0}),$$

dove $E_{f_0}^{y^*}(\cdot)$ indica il valore atteso rispetto alla densità f_0 di y^*. Per rimuovere la dipendenza da $\hat{\vartheta} = \hat{\vartheta}(y)$, si può considerare il valore atteso rispetto a $f_0(y)$, che fornisce

$$E_{f_0}^{y}\left[E_{f_0}^{y^*}\left[\sum_{i=1}^{n} \log\left(\frac{f_0(y_i^*)}{f(y_i^*;\hat{\vartheta})}\right)\right]\right] = nE_{f_0}^{y}\left[KL(f_0; f_{\hat{\vartheta}})\right]. \qquad (1.4)$$

L'applicazione dello sviluppo in serie di Taylor alla funzione $\log f(y;\hat{\vartheta})$, attorno a ϑ_0, fornisce

$$\begin{aligned}\log f(y;\hat{\vartheta}) &= \log f(y;\vartheta_0) + (\hat{\vartheta}-\vartheta_0)^\top \left.\frac{\partial \log f(y;\vartheta)}{\partial \vartheta}\right|_{\vartheta=\vartheta_0} + \\ &\quad + \frac{1}{2}(\hat{\vartheta}-\vartheta_0)^\top \left.\frac{\partial^2 \log f(y;\vartheta)}{\partial\vartheta\partial\vartheta^\top}\right|_{\vartheta=\vartheta_0} (\hat{\vartheta}-\vartheta_0) + \dots\end{aligned}$$

e poiché ϑ_0 è quel valore del parametro che minimizza $KL(f_0; f_\vartheta)$, allora $E^y_{f_0}\left(\partial \log f(y;\vartheta_0)/\partial\vartheta\right) = 0$, ovvero la funzione score associata al modello $f(y;\vartheta_0)$ continua ad avere valore atteso nullo. Grazie a tali risultati, si ha che

$$\begin{aligned}nKL(f_0; f_{\hat{\vartheta}}) &= n\int \log\left(\frac{f_0(y)}{f(y;\hat{\vartheta})}\right) f_0(y)dy \\ &\doteq nKL(f_0; f_{\vartheta_{f_0}}) + \frac{1}{2} tr\left((\hat{\vartheta}-\vartheta_{f_0}) I_{f_0}(\vartheta_{f_0})(\hat{\vartheta}-\vartheta_{f_0})^\top\right)\end{aligned}$$

dove $I_{f_0}(\vartheta) = -n\int \frac{\partial^2 \log f(y;\vartheta)}{\partial\vartheta\partial\vartheta^\top} f_0(y)dy$. Si ricorda che, anche se il modello è scorrettamente specificato, lo stimatore di massima verosimiglianza $\hat{\vartheta}$ ha distribuzione asintotica normale con media ϑ_0 e matrice di varianza che risulta essere

$$I_{f_0}(\vartheta_0)^{-1} K(\vartheta_0) I_{f_0}(\vartheta_0)^{-1},$$

con $K(\vartheta) = n\int \frac{\partial \log f(y;\vartheta)}{\partial\vartheta}\frac{\partial \log f(y;\vartheta)}{\partial\vartheta^\top} f_0(y)dy$. In base a questo risultato è possibile scrivere la (1.4) come

$$nE_{f_0}\left[KL(f_0; f_{\hat{\vartheta}})\right] \doteq nKL(f_0; f_{\vartheta_0}) + \frac{1}{2} tr\left(K(\vartheta_0) I_{f_0}(\vartheta_0)^{-1}\right), \tag{1.5}$$

dove il secondo termine è un termine di penalizzazione che dipende dalla dimensione k del parametro ϑ.

Quando il modello è corretto e regolare, allora $K(\vartheta_0) = I_{f_0}(\vartheta_0)$. Pertanto $tr(K(\vartheta_0) I_{f_0}(\vartheta_0)^{-1}) = k$ e

$$nE^y_{f_0}\left[KL(f_0; f_{\hat{\vartheta}})\right] \doteq nKL(f_0; f_{\vartheta_0}) + \frac{k}{2}. \tag{1.6}$$

Una stima della (1.5) è $-l(\vartheta) + c_1$, ove $l(\vartheta) = \log f(y;\vartheta)$ è la funzione di log-verosimiglianza per ϑ e c_1 è una stima del termine $tr\left(K\left(\vartheta_0\right) I_{f_0}(\vartheta_0)^{-1}\right)$.

Due possibili scelte per c_1 sono k, la dimensione di ϑ, e $tr(\widehat{J}\widehat{K}^{-1})$, con

$$\widehat{J} = \sum_{i=1}^{n} \frac{\partial \log f(y_i;\vartheta)}{\partial \vartheta} \frac{\partial \log f(y_i;\vartheta)}{\partial \vartheta^\top}\bigg|_{\vartheta=\hat{\vartheta}}$$

e

$$\widehat{K} = -\sum_{i=1}^{n} \frac{\partial^2 \log f(y_i;\vartheta)}{\partial \vartheta \partial \vartheta^\top}\bigg|_{\vartheta=\hat{\vartheta}}$$

Queste due scelte portano o all'indice AIC (Akaike, 1973), dato da

$$AIC = 2(-\log f(y;\hat{\vartheta}) + k) = -2l(\hat{\vartheta}) + 2k, \tag{1.7}$$

o all'indice TIC (Takeuchi, 1976), dato da

$$TIC = 2(-\log f(y;\hat{\vartheta}) + tr(\widehat{J}\widehat{K}^{-1})) = -2l(\hat{\vartheta}) + 2tr(\widehat{J}\widehat{K}^{-1}).$$

Il modello viene scelto in modo da rendere minimo o l'AIC o il TIC. L'AIC può essere interpretato come una verosimiglianza massimizzata penalizzata: infatti la minimizzazione della (1.3) è valutata in termini di caduta di verosimiglianza causata dall'ingresso di un nuovo parametro (ossia dall'ulteriore apporto di complessità al modello), concetto che rinvia direttamente al *principio di parsimonia* sottostante la selezione del modello. Il TIC nasce come un criterio che estende il risultato di Akaike ai casi in cui il modello è non correttamente specificato. Nonostante ciò, questo criterio di informazione è empiricamente poco usato, in quanto il termine $tr(\widehat{J}\widehat{K}^{-1})$, è suscettibile a grandi oscillazioni campionarie, a meno che non si disponga di elevate numerosità (Burnham e Anderson, 2002, cap. 8). Di riflesso questo mina l'affidabilità con cui si scelgono di volta in volta i vari modelli e si ritiene più semplice e sicuro usare lo "stimatore" più parsimonioso di c_1, ovvero k (si veda Shibata, 1989). Quest'ultima riflessione mostra come l'AIC sia un'approssimazione del TIC, precisa quando $f(y;\vartheta)$ è vicino a $f_0(y)$, scadente altrimenti.

1.3 Il criterio di informazione Bayesiano

Nella descrizione dell'AIC e del TIC si è fatto riferimento ad un approccio statistico di tipo frequentista, che si contrappone a quello bayesiano. Senza

entrare eccessivamente nel dettaglio, è possibile evidenziare coincisamente uno dei punti nodali che caraterizzano il dualismo tra le due scuole di pensiero: l'idea di fondo dell'approccio bayesiano consta nell'utilizzo di informazioni campionarie coadiuvate da altre informazioni, dette di tipo soggettivo. Quest'ultime informazioni vengono chiamate *a priori*. Esse si riferiscono alla fase che precede l'osservazione campionaria e si basano su conoscenze pregresse del fenomeno in esame. L'approccio bayesiano si avvale del teorema di Bayes per giungere alla distribuzione *a posteriori*, ovvero il risultato del ragionamento induttivo. La formula di Bayes si ricorda essere

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}, \tag{1.8}$$

$\forall B \in \Omega$ tale che $P(B) \neq 0$, con Ω generico spazio campionario. Se al posto di eventi A e B si considera il campione y e un modello, M, la (1.8) assume la forma

$$P(M|y) = \frac{P(M)P(y|M)}{P(y)}.$$

È possibile contestualizzare tale formula nell'ambito della selezione del modello. Infatti, $P(M)$ indica la probabilità a priori, o, equivalentemente, la veridicità che soggettivamente si assegna al modello M; $P(y)$ indica la probabilità marginale di y e $P(y|M)$ indica la probabilità di osservare i dati y dato il modello M. Si noti che questa formulazione, in un certo qual modo, concettualmente ricalca i due passi dell'AIC: dapprima si assegnano delle *a priori* al modello M, e poi, osservati i dati y, si calcolano le probabilità *a posteriori* $P(M|y)$. Il confronto tra due modelli, M_0 e M_1 rispettivamente, è valutabile tramite il rapporto (Racugno, 1998)

$$\frac{P(M_1|y)}{P(M_0|y)} = \frac{P(y|M_1)}{P(y|M_0)}\frac{P(M_1)}{P(M_0)}.$$

Il rapporto *a priori* $P(M_1)/P(M_0)$ viene aggiornato dai dati nel rapporto *a posteriori* $P(M_1|y)/P(M_0|y)$ tramite il fattore di Bayes, dato da

$$B_{10} = \frac{P(y|M_1)}{P(y|M_0)}.$$

Nelle situazioni più semplici, ovvero quando i due modelli M_0 e M_1 non dipendono da parametri ignoti, il fattore di Bayes B_{10} è pari al rapporto delle verosimiglianze $f_1(y)/f_0(y)$. Se invece i due modelli M_0 e M_1 dipendono da parametri ignoti, indicati con ϑ^0 e ϑ^1, allora il fattore di Bayes dipende dalle quantità

$$P(y|M_i) = \int f_i(y;\vartheta^i)\pi_i(\vartheta^i)d\vartheta^i, \quad i = 0,1\ , \tag{1.9}$$

dove $\pi_i(\vartheta^i)$ è la distribuzione a priori per ϑ^i sotto il modello M_i. In questo caso, il fattore di Bayes è un rapporto di verosimiglianze integrate. La quantità $2\log B_{10}$ è usata di frequente per riassumere l'evidenza per M_1 contro M_0.

Un problema che si incontra nell'attuare la via bayesiana è dovuto alla specificazione delle distribuzioni a priori $\pi_i(\vartheta^i)$ e al calcolo della (1.9) per $i = 0,1$. Per tale motivo, nel calcolo delle $P(y|M_i)$ viene utilizzata l'approssimazione di Laplace (si veda ad esempio Pace e Salvan, 1996, pag.92), che fornisce

$$\log P(y|M_i) \doteq \log f_i(y;\hat{\vartheta}^i) + \log\left[(det\widehat{J}_i)^{-\frac{1}{2}}(2\pi)^{\frac{k}{2}}\pi(\hat{\vartheta}^i)(\log n)^{\frac{k}{2}}\right],$$

dove $\hat{\vartheta}^i$ è la stima di massima verosimiglianza di ϑ^i e $det(\widehat{J}_i) = \left|-\frac{\partial \log f_i(y;\vartheta^i)}{\partial \vartheta_i}\frac{\partial \log f_i(y;\vartheta^i)}{\partial \vartheta_i^\top}\right|_{\vartheta^i=\hat{\vartheta}^i}$. Trascurando i termini di ordine $O(1)$, si ottiene il BIC, dato da

$$-2\log P(y|M_i) \doteq BIC = -2\log f_i(y;\hat{\vartheta}^i) + k\log(n),$$

detto criterio di informazione bayesiano, che può essere usato in maniera analoga all'AIC o al TIC per la scelta tra modelli. La forma funzionale di tale criterio ricalca quella dell'AIC tranne per il termine di penalizzazione, che in questo caso dipende anche dalla dimensione del campione. Quindi il BIC è un criterio che al crescere di n diventa sempre più conservativo a favore del modello con meno parametri. Sebbene il calcolo del BIC non sia poi molto diverso da quello dell'AIC esso ha, come si è visto, una genesi differente.

1.4 Conclusioni

In questo capitolo si sono presentati i principali criteri di selezione del modello presenti in letteratura. Per gli scopi di questa tesi, però, si farà sempre riferimento all'AIC e al TIC in quanto sono derivati a partire dalla distnaza di KL, che sarà anche il punto di partenza per i principali risultati che si presenteranno in seguito.

Capitolo 2

Inferenza basata su equazioni di stima

Nel capitolo precedente sono stati presentati alcuni metodi di selezione del modello, basati sulla specificazione di un modello statistico parametrico e sulla funzione di verosimiglianza. In molte situazioni di interesse applicativo, non è sempre possibile imporre le condizioni tali da consentire la costruzione di una funzione di verosimiglianza propria. Questo si verifica, ad esempio, nel contesto della teoria della robustezza, quando la stabilità rispetto alla specificazione scorretta o alla contaminazione è richiesta (si veda Hampel *et al.*, 1986), o nel contesto dei modelli lineari generalizzati, in presenza di sovradispersione o effetti casuali (si veda McCullagh e Nelder, 1989). In queste situazioni può essere preferibile basare l'inferenza su equazioni di stima. La teoria delle equazioni di stima ha avuto uno sviluppo continuativo a partire dal 1960, anno in cui Godambe (Godambe, 1960) portò alla luce una proprietà delle funzioni di stima.

In questo capitolo verranno presentate le equazioni di stima, le loro principali proprietà, che di volta in volta saranno contestualizzate a seconda del campo applicativo, e il loro legame con il concetto di quasi-verosimiglianza.

Ulteriori approfondimenti sulle equazioni di stima sono reperibili, per esempio, nei lavori di Godambe (1960, 1976, 1991).

2.1 Equazioni di stima: definizione e proprietà

Ai fini della specificazione di una funzione di stima non è tassativo invocare un modello parametrico del tipo $\mathscr{F} = \left\{ f(y;\vartheta);\ \vartheta \in \Theta \subseteq \mathbb{R}^k, k \geq 1 \right\}$, ma solamente richiedere l'esistenza delle quantità di interesse (per esempio, momento primo finito). Un'equazione di stima è una generica funzione dipendente dal campione osservato $y = (y_1, \ldots, y_n)$ e dal parametro oggetto di inferenza ϑ. In generale, un'equazione di stima è definita come

$$q(y;\vartheta) = \sum_{i=1}^{n} q(y_i;\vartheta) = 0, \tag{2.1}$$

ove $q(\cdot)$ è una funzione nota con valori in $\mathbb{R}^k$, che, risolta, fornisce una stima $\tilde{\vartheta}$ di ϑ. Si noti che per $k > 1$, $q(y;\vartheta) = 0$ rappresenta un sistema di k equazioni in ϑ. Nell'ambito del modello $\mathscr{F}$, un'equazione di stima è detta non distorta se

$$E_{f_\vartheta}\left[q(Y;\vartheta)\right] = 0, \quad \forall\, \vartheta \in \Theta.$$

Tale condizione non implica la non distorsione dello stimatore $\tilde{\vartheta}$ ottenuto come soluzione della (2.1) (Desmond, 1997), bensì fornisce l'argomento principale per mostrare la consistenza di $\tilde{\vartheta}$.

Per semplicità di esposizione si assuma ϑ scalare e sia $\ell_*(\vartheta) = \sum_{i=1}^{n} \frac{\partial \log f(y_i;\vartheta)}{\partial \vartheta}$ la *score* di verosimiglianza ottenuta dal generico elemento di $\mathscr{F}$. Si dimostra (Godambe, 1960) che, sotto $\mathscr{F}$, la funzione *score* gode di una proprietà di ottimalità fra tutte le funzioni di stima non distorte in valore atteso. Precisamente, vale

$$\frac{V_{f_\vartheta}(\ell_*(\vartheta))}{\left\{E_{f_\vartheta}\left[\partial \ell_*(\vartheta)/\partial\vartheta\right]\right\}^2} \leq \frac{V_{f_\vartheta}(q(Y;\vartheta))}{\left\{E_{f_\vartheta}\left[\partial q(Y;\vartheta)/\partial\vartheta\right]\right\}^2}, \quad \forall\, \vartheta \in \Theta,$$

con $V^y_{f_\vartheta}(\cdot)$ varianza calcolata rispetto a $f(y;\vartheta)$. Sia $Deff = \frac{V_{f_\vartheta}(q(Y;\vartheta))}{\left\{E_{f_\vartheta}\left[\partial q(Y;\vartheta)/\partial\vartheta\right]\right\}^2}$. Il numeratore del *Deff* è la varianza dell'equazione di stima, mentre il denominatore è un'indice di sensibilità della funzione nel discriminare i valori di ϑ: tanto più concentrata è una funzione attorno a un valore ϑ^*, allora tanto più potente sarà nel discriminare valori in un intorno di ϑ^*, e ciò implica un *Deff* piccolo (Desmond, 1997).

Al posto della funzione di stima $q(\cdot)$ si può lavorare con la funzione di stima standardizzata $q_s(\cdot)$, data da

$$q_s(y;\vartheta) = \frac{q(y;\vartheta)}{E_{f_\vartheta}\left[\frac{\partial q(Y;\vartheta)}{\partial \vartheta}\right]}.$$

La standardizzazione può essere richiesta poiché $q(\cdot)$ e $\lambda q(\cdot)$, per $\lambda \in \mathbb{R}$, definiscono lo stesso stimatore, ma vale $V_{f_\vartheta}[\lambda q(\cdot)] = \lambda^2 V_{f_\vartheta}[q(\cdot)]$. Ciò implica che la varianza di $\lambda q(\cdot)$ potrebbe essere ridotta, o aumentata, a piacere per opportune scelte della costante λ. Ne deriva che un'equazione di stima standardizzata ottima è unica al più di una costante moltiplicativa (Godambe e Kale, 1991). Quindi, per ottenere un'equazione di stima ottima tra quelle standardizzate è necessario minimizzare la varianza di $q_s(\cdot)$, ossia $V_{f_\vartheta}[q_s(Y;\vartheta)]$, essendo, quest'ultima, pari a *Deff*.

Nel caso in cui ϑ sia multidimensionale, i concetti di cui sopra sono facilmente generalizzabili (si veda Godambe e Kale, 1991). Sia $q_\vartheta(y;\vartheta) = \partial q(y;\vartheta)/\partial \vartheta^\top$ e $V_q(\vartheta)$ una matrice di ordine k definita positiva, data da

$$V_q(\vartheta) = {M_q}^{-1}(\vartheta) A_q(\vartheta) {M_q}^{-\top}(\vartheta), \tag{2.2}$$

ove $M_q(\vartheta) = E_\vartheta[q_\vartheta(Y;\vartheta)]$ è una matrice $(k \times k)$ non singolare e $A_q(\vartheta) = E_\vartheta\left[q(Y;\vartheta)q(Y;\vartheta)^\top\right]$. La funzione di stima $q^*(y;\vartheta)$ è una funzione di stima ottima tra le funzioni di stima non distorte se

$$tr(V_{q*}(\vartheta)) \leq tr(V_q(\vartheta)), \quad \forall\ q \in \Im,$$

dove $\Im$ indica l'insieme delle funzioni di stima non distorte. Vi sono altri criteri per determinare l'ottimalità nel caso multidimensionale, ma certamente questo è il più intuitivo visto che gli elementi diagonali di $V_{q*}(\cdot)$ sono le varianze delle funzioni di stima standardizzate (per ciascuna componente di ϑ).

Due proprietà interessanti dello stimatore $\tilde{\vartheta}$ ottenuto da un'equazione di stima non distorta della forma (2.1) riguardano il loro comportamento asintotico, in analogia alla teoria degli stimatori di massima verosimiglianza. Infatti, sotto tenui condizioni di regolarità della funzione di stima $q(y;\vartheta)$, tra

le quali la non distorsione, si dimostra che lo stimatore $\tilde{\vartheta}$ è consistente (si veda, ad esempio, Pace e Salvan, 1996, Capitolo 3). Inoltre, sotto le stesse condizioni, vale l'approssimazione

$$\tilde{\vartheta} \dot{\sim} \mathcal{N}_k(\vartheta, V_q(\vartheta)).$$

Tali conclusioni continuano a valere se l'equazione di stima, anziché essere non distorta, è tale che $E_{f_\vartheta}[q(Y;\vartheta)] = O_p(1)$.

In generale, a differenza di quanto accade per una *score* propria, per l'equazione di stima (2.1) risulta $M_q(\vartheta) \neq A_q(\vartheta)$ e pertanto non vale l'identità, analoga a quella dell'informazione,

$$V_{f_\vartheta}(q(Y;\vartheta)) = -E_{f_\vartheta}(q_\vartheta(Y;\vartheta)). \tag{2.3}$$

Si noti tuttavia che l'equazione modificata

$$\tilde{q}(y;\vartheta) = B(\vartheta)q(y;\vartheta) \tag{2.4}$$

con $B(\vartheta)^\top = -A_q^{-1}(\vartheta)M_q(\vartheta)$, ha la stessa soluzione $\tilde{\vartheta}$ della (2.1), mantiene la non distorsione e soddisfa anche l'identità $M_q(\vartheta) = A_q(\vartheta)$, per ogni $\vartheta \in \Theta$. Di conseguenza lo stimatore $\tilde{\vartheta}$, ottenuto da $\tilde{q}(y;\vartheta) = 0$, avrà la distribuzione asintotica nulla

$$\tilde{\vartheta} \dot{\sim} \mathcal{N}_k(\vartheta, A_q(\vartheta)^{-1}).$$

Se la funzione $q(\cdot)$ non è la funzione *score* di un modello statistico parametrico, non si avrà in generale una funzione di log-verosimiglianza associata. Si veda tuttavia la nozione di quasi-verosimiglianza, richiamata nel §2.3.

2.2 Alcuni esempi notevoli

2.2.1 Equazioni di stima robuste

Un problema che s'incontra nell'analisi di dati reali concerne la corretta specificazione del modello statistico. Infatti, una scorretta specificazione, come ad esempio uno scostamento più o meno marcato dall'ipotesi di normalità, può portare, a seconda dei casi, a drammatiche conseguenze sulle

procedure inferenziali. Pertanto, nelle situazioni in cui non si abbiano sufficienti informazioni sul fenomeno d'interesse, provenienti dai dati o da altre fonti (ad esempio indagini precedenti), è auspicabile ricorrere a statistiche robuste rispetto alla specificazione scorretta (*misspecification*). La letteratura statistica propone l'utilizzo di equazioni di stima robuste, ovvero opportune equazioni di stima che devono sottostare a determinati vincoli. Un'importante classe di stimatori legata alle equazioni di stima robuste è quella degli stimatori di tipo M (*M-estimators*), introdotta da Huber (Huber, 1964, Hampel *et al.*, 1986). Il nome deriva da "*maximum likelihood type estimators*" essendo, gli stimatori di tipo M, una generalizzazione degli stimatori di massima verosimiglianza.

In questo paragrafo, tra i metodi proposti in letteratura per studiare le statistiche robuste e le loro proprietà, si privilegerà l'approccio infinitesimale, introdotto da Hampel (si veda, ad esempio, Hampel *et al.*, 1986), basato sulla funzione di influenza.

Sia $\mathscr{F} = \left\{ f(y;\vartheta), \vartheta \in \Theta \subseteq \mathbb{R}^k \right\}$ un modello statistico parametrico, $y = (y_1 \ldots, y_n)$ un campione a componenti i.i.d. proveniente da $f(y;\vartheta)$ e si ipotizzi che il vero modello $f_0(y)$ sia in un qualche intorno di $f(y;\vartheta) \in \mathscr{F}$, da cui può differire per aspetti che portano all'osservazione di una frazione di dati anomali. Obiettivo dell'inferenza resta comunque il parametro ϑ. Una procedura d'inferenza robusta sul parametro ϑ può essere guidata dall'utilizzo di uno stimatore di tipo M. Sia $\rho(\cdot) : \mathcal{Y} \times \Theta \longmapsto \mathbb{R}^k$ una generica funzione composta dalla somma di componenti indipendenti, ossia

$$\rho\,(y;\vartheta) = \left(\sum_{i=1}^n \rho_1(y_i;\vartheta), \quad \ldots \quad , \sum_{i=1}^n \rho_k(y_i;\vartheta) \right)^\top \tag{2.5}$$

alla quale si applica la condizione di primo ordine per la ricerca di un massimo, ottenendo

$$\begin{aligned} \psi(y;\vartheta) &= \frac{\partial \rho\,(y;\vartheta)}{\partial \vartheta} \\ &= \left(\sum_{i=1}^n \psi_1(y_i;\vartheta), \quad \ldots \quad , \sum_{i=1}^n \psi_k(y_i;\vartheta) \right)^\top = 0, \end{aligned} \tag{2.6}$$

con $\psi(\cdot)$ funzione nota con valori in $\mathbb{R}^k$, della quale lo stimatore $\tilde{\vartheta}$ di tipo M è soluzione. Come nel §2.1, anche in ambito robusto è richiesta la non

distorsione dell'equazione di stima, ovvero

$$E_{f_\vartheta}\left[\psi(Y;\vartheta)\right] = 0,$$

in quanto garantisce la consistenza secondo Fisher (*Fisher consistency*), ed è inoltre l'argomento principale per dimostrare le proprietà asintotiche dello stimatore ricavato da $\psi(\cdot)$.

In termini di robustezza è, invece, la funzione d'influenza (IF) per $\psi(\cdot)$ a giocare un ruolo fondamentale. La funzione d'influenza per $\tilde{\vartheta}$ rispetto al modello $\mathscr{F}$, calcolata nel punto c, indicata con $IF(c;\tilde{\vartheta})$, è definita come (Huber, 1981)

$$IF(c;\tilde{\vartheta}) = -M^{-1}(\vartheta)\psi(c;\vartheta), \quad c \in \mathbb{R}, \forall\vartheta \in \Theta. \tag{2.7}$$

Con $M(\vartheta) = E_{f_\vartheta}\left[\psi(Y;\vartheta)\right]$ e rappresenta l'effetto prodotto su $\tilde{\vartheta}$ da una contaminazione infinitesimale nel punto c. Tale interpretazione fu evidenziata da Hampel (Hampel, 1974) nell'ambito dello studio di stimatori robusti alla contaminazione.

La richiesta di robustezza si traduce nell'imposizione di opportune condizioni di limitatezza sulla IF, la più importante delle quali è rappresentata dalla richiesta che la quantità

$$\gamma = \sup_c \left\| IF(c;\tilde{\vartheta}) \right\|,$$

in cui $\|\cdot\|$ denota la norma Euclidea, sia finita. L'indice γ è noto nella letteratura robusta come l'indice di sensibilità ai grandi errori (*gross error sensivity*). Se γ risulta essere limitato, allora si ha la B-robustezza di $\tilde{\vartheta}$. Infatti, γ misura il più grande impatto che una contaminazione, di ammontare fissato, può avere sulla distorsione dello stimatore definito dalla (2.6). In virtù della (2.7), per uno stimatore di tipo M l'indice γ è finito se e solo se la funzione $\psi(y;\vartheta)$ è limitata.

Anche per un'equazione di stima robusta (2.6) vale la normalità asintotica dello stimatore da essa ricavato, essendo un caso particolare della (2.1). In questo caso la matrice di covarianza (2.2) può essere espressa per mezzo della IF, come

$$V(\vartheta) = \int IF(c;\tilde{\vartheta})IF(c;\tilde{\vartheta})^\top f(y;\vartheta)\, dy. \tag{2.8}$$

Se $V(\vartheta) < +\infty$, vale il risultato asintotico

$$\tilde{\vartheta} \dot{\sim} \mathcal{N}_k(\vartheta, V(\vartheta)).$$

Naturalmente un'equazione di stima robusta apporta un notevole beneficio in termini di affidabilità delle procedure inferenziali quando il modello è effettivamente contaminato, ad esempio è formato dalla mistura di più modelli. Ciò non vuol dire, però, che i benefci di un'equazione di stima robusta siano incondizionati: la ricerca di robustezza si sconta in termini di efficienza dello stimatore ottenuto. Questo è il compromesso che Huber (Huber, 1981) sintetizza con la dicitura *"two-person zero-sum game"*. Quindi un obiettivo da perseguire, nel momento in cui si decide di utilizzare un'equazione di stima robusta, può essere quello di ricercarne una che formisca lo stimatore B-robusto ottimo, ossia lo stimatore con varianza asintotica minima nella classe degli stimatori con indice γ limitato. Tale scelta segue criteri analoghi a quelli visti nel §2.1, con l'aggiunta di un ulteriore vincolo. Infatti, in ambito robusto i confronti tra equazioni di stima vanno effettuati a parità di γ. Dunque, definito un insieme $\xi_a = \{\psi(\cdot) |\ \gamma \leq a,\ a \in \mathbb{R}\}$, l'equazione di stima ottima in ξ_a sarà tale per cui

$$tr(V_*) < tr(V),\ \forall \psi \in \xi_a,$$

con $\psi_*(\cdot)$ equazione di stima B-robusta ottima tra le B-robuste.

Per $k = 1$, una stima della (2.8) è

$$\widehat{V}(\hat{\vartheta}) = \frac{\sum_{i=1}^{n} \psi(\hat{\vartheta}; y_i)^2}{\left[\sum_{i=1}^{n} \frac{\partial}{\partial \vartheta} \psi(\vartheta; y_i) \mid_{\vartheta=\hat{\vartheta}}\right]^2}.$$

Un intervallo di confidenza alla Wald di livello approssimato 1-α è allora

$$\hat{\vartheta} \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{V}(\hat{\vartheta})},$$

ove $z_{1-\frac{\alpha}{2}}$ indica il percentile di livello $1 - \frac{\alpha}{2}$ della distribuzione normale standard.

2.2.2 Modelli lineari generalizzati

I modelli lineari generalizzati (GLM) costituiscono un'estensione dei modelli lineari con errori normali (Nelder e Wedderburn, 1972, McCullagh e Nelder, 1989). Uno degli aspetti di generalizzazione rispetto a questi ultimi riguarda la distribuzione di riferimento, considerando i modelli provenienti dalla famiglia esponenziale, che includono come caso particolare quello normale. La necessità di tale generalizzazione nasce dal bisogno di modellare dati con risposte non continue, per esempio dati di conteggio o proporzioni. Sostanzialmente la generalizzazione avviene tramite la specificazione:

1. di un'elemento proveniente dalla famiglia esponenziale (componente casuale);
2. del predittore lineare (componente sistematica);
3. della funzione di legame tra predittore lineare e media della risposta.

Si consideri un campione $y = (y_1, \ldots, y_n)$ di osservazioni indipendenti proveniente da un elemento della famiglia esponenziale. Allora, la densità della singola osservazione assume la forma

$$f(y; \vartheta, \phi) = \exp\left\{\frac{y\vartheta - b(\vartheta)}{\phi} + c(y, \phi)\right\}, \tag{2.9}$$

ove ϑ è il parametro naturale, ϕ è il parametro di dispersione, mentre $b(\cdot)$ e $c(\cdot)$ sono due funzioni note per ogni elemento della famiglia esponenziale. Dalle assunzioni di indipendenza segue che la funzione di verosimiglianza per ϑ è data da $L(\vartheta, \phi) = \prod_{i=1}^{n} f(y_i; \vartheta, \phi)$. Si indichi inoltre la log-verosimiglianza con $\ell(\vartheta, \phi) = \sum_{i=1}^{n} \log L(\vartheta, \phi)$.

Per una specificazione del modello nella forma (2.9) il valore atteso di Y è dato da $b'(\vartheta) = \partial b(\vartheta)/\partial\vartheta$ ed è legato al predittore lineare $\eta_i = x_i^\top \beta$, per mezzo della funzione di legame $g(\cdot)$,

$$E_{f_\vartheta}[Y_i] = \mu_i = b'(\vartheta) = g(x_i^\top \beta), \quad i = 1, \ldots, n,$$

in cui x_i rappresenta il vettore noto delle k variabili esplicative rilevate presso l'i-esima unità, mentre β è il vettore k-dimensionale dei coefficienti di regressione.

I GLM, oltre a migliorare la capacità e possibilità di modellazione, vantano una trattazione matematica semplice basata sulla funzione di verosimiglianza. Risolvendo le equazioni di verosimiglianza, con il metodo dei *minimi quadrati pesati iterati*, si ottengono le stime dei parametri, indipendentemente dal valore del parametro di dispersione ϕ. Infatti, le equazioni di verosimiglianza per β sono date da

$$\frac{1}{\phi}\sum_{i=1}^{n}\frac{(y_i-\mu_i)}{V(\mu_i)g'(\mu_i)}x_{ij}=0, \quad j=1,\ldots,k, \tag{2.10}$$

ove $V(\mu_i)$ è la funzione di varianza tale per cui $V_{f_\vartheta}(Y_i)=\phi V(\mu_i)$. La componente $j-s$ della matrice di convarianze di β fornisce

$$I_{js}=\frac{1}{\phi}\sum_{i=1}^{n}\frac{x_{ij}x_{is}}{V(\mu_i)(g'(\mu_i))^2}, \quad j,s=1\ldots k. \tag{2.11}$$

La (2.11) viene spesso riportata nella forma matriciale

$$I_{\beta\beta}=\frac{1}{\phi}X^\top WX \tag{2.12}$$

con $W=diag(w_i)$ e $w_i=\frac{1}{V(\mu_i)(g'(\mu_i))^2}$. Le equazioni di verosimiglianza (2.10) per β rappresentano equazioni di stima non distorte purché sia $E_{f_\vartheta}[Y_i]=\mu_i=g^{-1}(\beta^\top x_i)$. In altri termini, l'assunzione parametrica sulla distribuzione della Y_i potrebbe anche non essere soddisfatta; ciò che risulta essenziale è l'ipotesi sulla sua media. Si noti anche che l'unica quantità necessaria per scrivere la (2.10) è la funzione di varianza $V(\mu)$. È possibile specificare quindi un GLM sotto le più deboli assunzioni del secondo ordine (Wedderburn, 1974):

1. $E(Y_i)=\mu_i=g^{-1}(\beta^\top x_i)$;
2. $Var(Y_i)=\phi V(\mu_i)$;
3. $Cov(Y_i,Y_j)=0$ se $i\neq j$.

Il modello semiparametrico specificato dalla assunzioni $1-3$ è detto modello di quasi-verosimiglianza. In tal caso, la stima di β, indicata con $\hat{\beta}$, è ancora

la soluzione della (2.10), mentre per la stima di ϕ si fa conveniente ricorso al metodo dei momenti, che fornisce

$$\tilde{\phi} = \frac{1}{n-k} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)},$$

dove $\hat{\mu}_i = g^{-1}(x_i^\top \hat{\beta})$. In generale, $\tilde{\phi}$ è consistente anche sotto il modello di quasi-verosimiglianza.

2.3 Quasi-verosimiglianza

La quasi-verosimiglianza nasce dalla necessità di condurre procedure inferenziali basate su equazioni di stima non distorte per ϑ, quando per le osservazioni si considera un modello statistico semiparametrico o non parametrico, o un modello parametrico con la richiesta di robustezza. Ciò implica l'assenza della funzione di verosimiglianza propria $L(\vartheta)$ e quindi l'impossibilità di applicare le produre classiche di inferenza.

La genesi della funzione di log quasi-verosimilianza deriva dal quesisto se sia possibile reperire una funzione $\ell_Q(\vartheta) = \ell_Q(\vartheta; y)$ avente una generica equazione di stima della forma (2.1), o (2.6), come gradiente. Tale quesito ha una diretta risposta nel caso $k = 1$, data dalla relazione formale

$$\ell_Q(\vartheta) = \int_c^\vartheta q(y; t)dt,$$

ove c è una costante arbitraria. Se il parametro ϑ è multidimensionale e $q(\cdot)$ è derivabile con continuità rispetto a ϑ, una condizione necessaria per l'esistenza di $\ell_Q(\vartheta)$ è che la matrice $-J_Q(\vartheta) = \frac{\partial}{\partial \vartheta^\top} q(y; \vartheta)$ sia simmetrica (si veda, ad esempio, Pace e Salvan, 1996, §4.9). Si osservi che in generale, a differenza di quanto accade per una funzione *score* propria, non è detto che valga l'identità, analoga a quella dell'informazione (2.3). È tuttavia possibile, in alcuni casi, recuperare tale identità attraverso una trasformazione lineare di $q(\cdot)$ della forma (2.4).

Se una quasi-verosimiglianza soddisfa la (2.3), molte considerazioni asintotiche risultano semplificate. Ad esempio, la matrice $J_Q(\vartheta)$ di quasi-

informazione osservata risulta collegata nel modo usuale alla matrice di covarianza asintotica dello stimatore $\tilde{\vartheta}$ basato su $q(\cdot)$.

Un esempio importante in cui si perviene a una rappresentazione di $\ell_Q(\vartheta)$ è nell'ambito dei GLM. Infatti, nel caso particolare del §2.2.2 è possibile definire un'equazione di quasi-verosimiglianza a partire dalle assunzioni del secondo ordine. Poiché la matrice $\partial q(y;\beta)/\partial\beta^\top$ è simmetrica, esiste una funzione $l_Q(\beta)$, avente la (2.6) come gradiente, data da

$$\ell_Q(\beta) = \sum_{i=1}^{n} \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)} dt. \tag{2.13}$$

Per approfondimenti riguardanti la quasi-verosimiglianza e la sua applicazione nel campo dei GLM si rimanda a McCullagh e Nelder (1989).

Per un'applicazione della quasi-verosimiglianza a equazioni di stima robuste, della forma (2.6), si veda Adimari e Ventura (2001, 2002). In particolare, per $k = 1$, una quasi-verosimiglianza per ϑ con le usuali proprietà asintotiche è data da

$$\ell_Q(\vartheta) = \sum_{i=1}^{n} \int_{c}^{\vartheta} B(t)\psi(y_i;t)dt, \tag{2.14}$$

ove $B(\vartheta)$ è data nella (2.3)

2.4 Conclusioni

In questo Capitolo sono state presentate bervemente le equazioni di stima. Si è visto il loro collegamento con gli stimatori di massima verosimiglianza (limitatamente per le equazioni di stima che definiscono stimatori d tipo M) e i principali risultati distributivi degli stimatori ricavati a partire da equazioni di stima.

Nel prossimo capitolo verranno illustrati alcuni lavori di alcuni autori che si servono delle equazioni di stima e delle loro proprietà per ricavare criteri di selezioni del modello che tentano, anche, di sopperire ai limiti dell'AIC e del TIC.

Capitolo 3

Metodi di selezione del modello basati su equazioni di stima

Nel capitolo precedente è stato presentato l'approccio inferenziale basato sulle equazioni di stima, alternativo a quello basato sulla funzione di verosimiglianza. Esempi di notevoli applicazioni si hanno nell'ambito della robustezza o dei GLM con sovradispersione, in cui si ricorre all'utilizzo di opportune equazioni di stima che permettono di alleggerire le assunzioni parametriche essenziali per specificare la funzione di verosimiglianza.

In questo capitolo si illustra come, a partire da equazioni di stima non distorte, si possono definire criteri di selezione del modello, in maniera simile a quelli visti nel Capitolo 1, basati fondamentalmente su criteri di informazione che sfruttano la funzione di verosimiglianza. Nello specifico, si farà riferimento ai lavori pubblicati da Pan (2001a, 2001b) in cui si propongono, rispettivamente, un'indice per la selezione del modello basato su equazioni di stima del tipo (2.1) e uno basato sulla generalizzazione dell'AIC per mezzo di equazioni di stima generalizzate (GEE). Inoltre, molto sinteticamente, nell'ambito della regressione lineare si presenteranno metodi di selezione del modello robusti proposti da Ronchetti(1985), Ronchetti e Staudte (1994) e Ronchetti *et al.* (1997).

3.1 Metodi di selezione del modello di Pan

I metodi di selezione del modello discussi in letteratura da Pan, che saranno presentati in seguito, si basano su equazioni di stima e quindi possono esssere utilizzati sia nel caso in cui si consideri per le osservazioni y un modello statistico $\mathcal{F}$, parametrico, semiparametrico o non parametrico. Sia ϑ il parametro k-dimensionale d'interesse. Sotto queste condizioni l'inferenza su ϑ può essere basata su un'equazione di stima non distorta $\psi(\cdot) : \mathcal{Y} \times \Theta \longmapsto \mathbb{R}^k$, della forma

$$\psi(y; \vartheta) = \sum_{i=1}^{n} \psi(y_i; \vartheta) = 0. \tag{3.1}$$

Si noti che la struttura dell'equazione (3.1) dà luogo a uno stimatore $\tilde{\vartheta}$ di tipo M per ϑ (si veda il §2.1).

Pan (2001a, 2001b) discute metodi di selezione del modello basati su equazioni di stima, che hanno importanti applicazioni sia per modelli annidati (ad esempio nella selezione del predittore lineare in modelli di regressione) sia per modelli non annidati (ad esempio nella scelta della funzione di legame per i GLM).

3.1.1 La distorsione attesa di previsione

Sia $y = (y_1, \ldots, y_n)$ un campione a componenti i.i.d. distribuite secondo il modello $\mathcal{F}$ e sia, inoltre, y^* un altro campione distribuito secondo $\mathcal{F}$, indipendente da y. Pan (2001a) suggerisce di basare il criterio di selezione del modello a partire dalla distorsione attesa di previsione (*expected predictive bias, EPB*) dell'equazione di stima (3.1), definita come

$$EPB = E_{f_0}^{y} E_{f_0}^{y^*} \left[\left| \psi(Y^*; \tilde{\vartheta}(Y)) \right| \right], \tag{3.2}$$

dove $E_0^y(\cdot)$ e $E_{f_0}^{y^*}(\cdot)$ indicano, rispettivamente, il valore atteso rispetto a Y e rispetto a Y^* sotto $\mathcal{F}$. Evidentemente, data la definizione di $\psi(\cdot)$, EPB sarà un vettore di dimensione k. L'interpretazione della (3.2) è molto semplice: un buon modello per un'osservazione futura sarà quello per cui si ha un EPB

che, componente per componente, è prossimo a zero, essendo $\psi(y;\vartheta)$ una stima consistente di $E^{y}_{f_0}\left[\psi(Y;\vartheta)\right]$, che è nullo per costruzione.

Nell'ambito delle equazioni di stima, la (3.2) può essere considerata come una generalizzazione della statistica C_p (Mallows, 1973), data nella (3.13), e dell'AIC. La (3.2), generalizzando questi due criteri, sintetizza l'approccio della minimizzazione della somma dei quadrati dei residui aggiustata (C_p) e quello della minimizzazione della discrepanza di KL (AIC).

Un'espressione analitica della (3.2) è complicata da ricavare e deve, in pratica, essere stimata. Si può ricorrere alla *cross-validation (CV)*, congiuntamente al *bootstrap* (Efron, 1979) per ottenere una stima attendibile di EPB a partire dai dati a disposizione. In questo caso, alla CV si affianca il *bootstrap* per ridurre sia la variabilità di stima dell'EPB, che altrimenti si avrebbe utilizzando la sola CV, sia l'eccessiva sovrapposizione delle osservazioni nel campione per la stima di ϑ e in quello per la stima dell'EPB. Operativamente, si generano B campioni *bootstrap*, cioè campioni tratti dalla funzione di ripartizione empirica del campione y. Per il b-esimo campione *bootstrap* $y^{+}_{b} = (y^{+}_{1b}, \ldots, y^{+}_{nb})$, $b = 1, \ldots, B$, si calcola

$$\widehat{EPB}_b = \sum_{i=1}^{n} \left| \psi(y^{++}_{ib}, \tilde{\vartheta}(y^{+}_{ib})) \right|, \tag{3.3}$$

con $y^{++}_{b} = y \backslash \left\{ y^{+}_{b} \right\}$, applicando il principio della CV. In questo contesto il simbolo "\" indica l'operatore che rimuove le osservazioni che si trovano sia in y che in y^{+}_{b}. L'idea è stimare, in un primo momento, ϑ utilizzando y^{+}, in seguito calcolare la discrepanza tra y^{++} e $\tilde{\vartheta}(y^{+})$. Si noti che, per definizione, $\sum_{i=1}^{n} \left| \psi(y^{+}_{ib}, \tilde{\vartheta}(y^{+}_{ib})) \right| = 0$. Infine, la stima completa dell'EPB si ottiene semplicemente come

$$\widehat{EPB} = \frac{1}{B} \sum_{b=1}^{B} \widehat{EPB}_b. \tag{3.4}$$

Per una discussione più dettagliata sugli aspetti computazionali della (3.4) si rinvia all'articolo di Pan (2001a). Inoltre, in Pan (2001a) viene anche discusso come combinare le componenti dell'EPB per formare uno scalare.

Il metodo è generale e richiede la sola specificazione di un'equazione di stima non distorta per la quantità d'interesse.

3.1.2 Criterio di selezione del modello basato su equazioni di stima generalizzate

Pan (2001b) discute anche un criterio di selezione del modello nell'ambito della regressione quando le osservazioni sono correlate. In tale situazione le equazioni (3.1) possono essere le equazioni di stima generalizzate GEE (Liang e Zeger, 1986). Il contesto prevede che per ogni individuo si disponga di un vettore nella variabile risposta $y_i = (y_{i1}, \ldots, y_{it})^\top$ e di un vettore k-dimensionale nelle variabili esplicative $x_{it} = (x_{it1}, \ldots, x_{itk})^\top, t = 1, \ldots, T, i = 1, \ldots, n$. In generale, le componenti di Y_i sono correlate, mentre Y_i e Y_k sono indipendenti per ogni $i \neq k$.

Volendo modellare $E_{f_0}(Y_i) = \mu_i = g^{-1}(x_i^\top \beta)$, ove $g(\cdot)$ è la funzione di legame, si può ottenere una stima di β risolvendo le GEE, date da

$$\psi(y; \beta, \alpha, \phi) = \sum_{i=1}^{n} D_i^\top V_i^{-1}(y_i - \mu_i) = 0, \tag{3.5}$$

con $V_i = A_i^{\frac{1}{2}} R(\alpha) A_i^{\frac{1}{2}}$ matrice di covarianza di Y_i, A_i matrice diagonale con elementi $\phi V(\mu_{it}), t = 1, \ldots, T$, $D_i = \partial \mu_i(\beta)/\partial \beta^\top$ e $R(\alpha)$ matrice di correlazione dipendente dal vettore di parametri α, la cui stima $\tilde{\alpha}$ può essere ottenuta attraverso il metodo dei momenti o da un altro sistema di equazioni di stima. In questo contesto $V(\mu_{ij})$ indica la funzione di varianza per un GLM e ϕ rappresenta il parametro di dispersione, la cui stima $\tilde{\phi}$ può essere ottenuta, per esempio, con il metodo dei momenti (si veda §2.2.2).

La (3.5) è un'equazione di stima generalizzata ed include, per $R(\alpha) = I_k$, la situazione di indipendenza delle osservazioni. In particolare, nel caso in cui $R(\alpha) = I_k$, la (3.5) è equivalente al gradiente di una funzione di quasi-verosimiglianza per l'inferenza su β (si veda §2.3). Nel caso in cui $R(\alpha) \neq I_k$ allora devono essere soddisfatti alcuni vincoli perché valga la relazione di cui sopra (si veda McCullagh e Nelder, 1989, §9.3.2). Inoltre, anche se la quasi-verosimiglianza esiste, è in generale difficile da costruire.

Le GEE sono molto versatili e, nella trattazione di osservazioni dipendenti, permettono uno sviluppo teorico molto interessante. Infatti, è possibile specificare o stimare (tramite metodi numerici iterativi) la matrice di correlazione $R(\alpha)$, con lo scopo di ottenere delle stime di β più efficienti di quelle ottenibili sotto l'ipotesi di indipendenza. Va detto che lo stimatore di β rimane consistente anche se la matrice di correlazione non è correttamente specificata: è solo necessario che sia correttamente specificata la media delle osservazioni. Ciononostante, optare per una matrice di correlazione pari alla matrice identità, quando invece i dati sono correlati, non provoca una perdita di efficienza "drammatica" nella stima di β, a meno che la correlazione non sia elevata (Zeger, 1988; McDonald 1993).

Pan (2001b) discute un'indice di selezione del modello basato sulla funzione di quasi-verosimiglianza per β ottenuta sotto il modello di indipendenza. Il criterio proprosto è definito come

$$QIC = -2\ell_Q(\hat{\beta}_R, \tilde{\alpha}, \tilde{\phi}) + 2tr(\widehat{\Omega}_{I_k}\widehat{V}_R), \tag{3.6}$$

dove $\ell_Q(\beta, \alpha, \phi) = \sum_{i=1}^{n}\sum_{t=1}^{T}\int_{y_{ij}}^{\mu_{ij}} \frac{y_{ij}-s}{\phi V(s)}ds$ è la funzione di quasi-verosimiglianza per β, $\hat{\beta}_R$ è la stima di β ricavata dalla (3.5) con una matrice di correlazione $R(\alpha)$ diversa dalla matrice identità,

$$\widehat{\Omega}_{I_k} = -\frac{\partial^2 \ell_Q(\beta, \alpha, \phi)}{\partial\beta\partial\beta^\top}\bigg|_{\beta=\hat{\beta}_R, \alpha=\tilde{\alpha}, \phi=\tilde{\phi}}$$

e

$$\widehat{V}_R = \widehat{\Omega}_{I_k}^{-1}\widehat{J}\,\widehat{\Omega}_{I_k}^{-\top}$$

è una stima robusta o *sandwich* della matrice di covarianza di $\hat{\beta}$ (si veda Liang e Zeger, 1986), con $\widehat{J} = \sum_{i=1}^{n}\sum_{t=1}^{T}\left(\frac{\partial\ell_Q(\beta,\alpha,\phi,y_{it})}{\partial\beta}\right)\left(\frac{\partial\ell_Q(\beta,\alpha,\phi,y_{it})}{\partial\beta}\right)^\top\Bigg|_{\beta=\hat{\beta}_R, \alpha=\tilde{\alpha}, \phi=\tilde{\phi}}$, ϕ il parametro di dispersione. Si noti che, con opportune semplificazioni, il termine $tr(\widehat{\Omega}_{I_k}\widehat{V}_R)$ diventa $tr(\widehat{J}\,\widehat{\Omega}_{I_k}^{-\top})$, il termine di aggiustanto che compare nel TIC.

In QIC mima il TIC in quanto è basato sulla definizione di una nuova discrepanza, data da

$$E_{f_0}\left[\ell_Q(\beta, \alpha, \phi)\right] \tag{3.7}$$

che si tratterà più nel dettaglio nel capitolo successivo.

Nella definizione di $\ell_Q(\cdot)$ si lavora sotto il modello di indipendenza e tale assunzione ha come scopo principale quello di assicurare l'esistenza della funzione di quasi-verosimiglianza. Si noti, inoltre, che nella massimizzazione di $\ell_Q(\cdot)$ si usa $\hat{\beta}_R$, mentre $\tilde{\alpha}$ e $\tilde{\phi}$ sono delle stime di α e ϕ ottenetute a partire, per esempio, da altre equazioni di stima.

Il criterio (3.6) può essere utilizzato per scegliere la miglior matrice di correlazione $R(\alpha)$, per la selezione tra modelli non annidati o per la scelta della funzione di legame nei GLM. In particolare, quando nelle GEE la struttura di correlazione è correttamente specificata, allora asintoticamente $tr(\widehat{\Omega}_{I_k}\widehat{V}_R) \approx k$ e il criterio (3.6) diventa

$$QAIC = -2\ell_Q(\hat{\beta}_R, \tilde{\alpha}, \tilde{\phi}) + 2k. \tag{3.8}$$

Nell'ambito di un GLM classico in cui sussiste l'indipendenza delle osservazioni, il QIC (3.6) si riduce, invece, a

$$QIC = -2\ell_Q(\hat{\beta}, \tilde{\phi}) + 2tr(\widehat{J}_Q\widehat{K}_Q^{-1}),$$

in cui le matrici $\widehat{J}_Q = \sum_{i=1}^{n}(\frac{\partial \ell_Q(\beta,\phi;y_i)}{\partial \beta})(\frac{\partial \ell_Q(\beta,\phi;y_i)}{\partial \beta})^\top$ e $\widehat{K}_Q = -\sum_{i=1}^{n}\frac{\partial^2 \ell_Q(\beta,\phi;y_i)}{\partial \beta \partial \beta^\top}$ sono calcolate in $\hat{\beta}$ e $\tilde{\phi}$, sfruttando le proprietà della funzione di quasi-verosimiglianza.

L'utilizzo di una nuova discrepanza della forma (3.7) è stato discusso anche da Varin e Vidoni (2005), con la funzione di verosimiglianza composita in luogo della funzione di quasi-verosimiglianza. Infatti, la verosimiglianza composita è una particolare pseudo-verosimiglianza la cui *score* è un'equazione di stima non distorta e il cui pseudo-stimatore di massima verosimiglianza è un particolare stimatore di tipo M, con varianza asintotica data dalla (2.2) con $M(\vartheta) = E_{f_0}\left[\frac{\partial^2 \ell_c(\vartheta)}{\partial \vartheta \partial \vartheta^\top}\right]$ e $A(\vartheta) = E_{f_0}\left[\frac{\partial \ell(\vartheta)}{\partial \vartheta}(\frac{\partial \ell(\vartheta)}{\partial \vartheta})^\top\right]$, dove $\ell_c(\vartheta)$ indica la funzione di log-verosimiglinaza composita per ϑ. In particolare, Varin e Vidoni (2005) così come nella (3.6), suggeriscono il seguente criterio di selezione del modello dato da

$$\ell_c(\hat{\vartheta}_c) + tr(\widehat{J}_c\widehat{K}_c^{-1}), \tag{3.9}$$

in cui $\hat{\vartheta}_c$ è la stima di massima verosimiglianza composita e $\widehat{J}_c$ e $\widehat{K}_c$ sono, rispettivamente, i corrispettivi campionari di $M(\vartheta)$ e $A(\vartheta)$ calcolate entrambe in $\hat{\vartheta}_c$.

Nella derivazione della (3.9) si usano molte delle considerazioni già viste nel §1.2 per la derivazione del TIC associate alla definizione di una distanza di KL generalizzata del tipo 3.7.

3.2 Inferenza robusta e metodi di selezione del modello

In questo paragrafo si presentano brevemente alcuni metodi di selezione del modello basati su equazioni di stima robuste (si veda §2.1) presenti in letteratura. Un criterio robusto di selezione del modello può essere considerato legato a doppio filo con la teoria dell'inferenza robusta: da un lato la scarsa letteratura su metodi di selezione robusti può essere vista come una mera lacuna; dall'altro usare criteri di selezione del modello non robusti può vanificare tutti gli sforzi e i compromessi che la scelta di stimatori robusti comporta. Inoltre, i lavori proposti in letteratura propongono prevalentemente metodi di selezione robusti nell'ambito del modello di regressione lineare Ronchetti (1985), Ronchetti e Staudte (1994) e Ronchetti *et al.* (1997).

Data la popolarità del criterio di selezione del modello di Akaike e della statistica C_p, in maniera naturale i primi criteri robusti di selezione del modello hanno in qualche modo preso esempio dalle loro controparti non robuste.

3.2.1 Una versione robusta del TIC in modelli di regressione

Ronchetti (1985) ha proposto una versione robusta dell'TIC nell'ambito del modello di regressione lineare. In questo constesto il TIC (**??**) dipende direttamente dalla somma dei quadrati dei residui che, come è noto, non è una quantità robusta. La versione dell'AIC proposta da Ronchetti (1985)

prevede, in primo luogo, che gli errori non si distribuiscano necessariamente secondo la legge normale e, in secondo luogo, che la statistica proprosta sia una funzione limitata della somma dei residui riscalati. La versione robusta dell'TIC, indicata con $RTIC$, ha la seguente espressione

$$RTIC = 2\sum_{i=1}^{n} \rho(r_i;\beta) + 2\alpha, \tag{3.10}$$

con $\psi(y;\beta) = \partial\rho(y;\beta)/\partial\beta$ dove $r_i = (y_i - x_i^\top\tilde{\beta})/\hat{\sigma}$, k è la dimensione del parametro di regressione β, $\tilde{\beta}$ è una stima robusta dei coefficienti di regressione soluzione di $\psi(y;\beta)$, $\tilde{\sigma}$ è una stima robusta del parametro di scala, $\rho(\cdot)$ è una funzione limitata con $\psi(y;\beta) = \partial\rho(y;\beta)/\partial\beta$ e α è un termine di penalizzazione dato da

$$\alpha = E_{f_0}\left[\frac{\partial\psi(r)}{\partial r}\left(\frac{\partial\psi(r)}{\partial r}\right)^\top\right] E_{f_0}\left[\frac{\partial^2\psi(r)}{\partial r\partial r^\top}\right]^{-1}. \tag{3.11}$$

Una possibile scelta per $\psi(\cdot)$ può essere la funzione di Huber (si veda Hampel *et al.*, 1986, Capitolo **??**).

Si noti che il termine di penalizzazione α è direttamente riconducibile ai risultati del paragrafo precedente. Inoltre, quando le matrici (3.11) sono equivalenti si ottiene una versione robusta dell'AIC, indicata con $RAIC$, che ha la seguente espressione

$$RAIC = 2\sum_{i=1}^{n} \rho(r_i;\beta) + 2k, \tag{3.12}$$

in cui k è la dimensione di β.

3.2.2 Una versione robusta dell'indice C_p di Mallows

Un criterio non robusto per la selezione del modello in modelli di regressione è l'indice C_p di Mallows (1973), la cui espressione è

$$C_p = \frac{RSS_k}{\hat{\sigma}^2} + 2k - n, \tag{3.13}$$

dove RSS_k è la somma dei quadrati dei residui di regressione, ossia $RSS_k = \sum_{i=1}^{n}\left(y_i - x_i^\top\hat{\beta}\right)^2$, $\hat{\beta}$ la stima ai minimi quadrati di β e $\hat{\sigma}^2$ una stima non

robusta della varianza (per approfondimenti si veda Gorman e Toman, 1966). Poiché il calcolo di C_p è basato sul metodo dei minimi quadrati, esso è sensibile rispetto alla presenza di valori anomali nelle osservazioni e/o a piccoli allontanamenti dall'ipotesi di normalità degli errori.

Un primo passo per rendere robusta la statistica C_p è quello di reperire una stima robusta di β ricavata da un'equazione di stima robusta e perciò $\tilde{\beta}$ sarà soluzione di un'equazione della forma

$$\psi(y;\beta) = \sum_{i=1}^{n} \psi(x_i, y_i - x_i^\top \beta) x_i = 0.$$

Naturalmente, la scelta di uno stimatore robusto per β va accompagnata da un criterio di selezione del modello anche'esso robusto. Per fare fronte a questo problema (si veda Ronchetti e Staudte, 1994) si definisce l'errore quadratico di predizione riscalato (*rescaled mean squared weighted prediction error*), come

$$\Gamma_k = \frac{1}{\sigma^2} E_{f_\beta} \left[\sum_{i=1}^{n} \hat{w_i}^2 (\hat{y}_i - E_{f_\beta} y_i)^2 \right],$$

in cui $\hat{w}_i = \frac{\psi(x_i, y_i - x_i^\top \tilde{\beta})}{(y_i - x_i^\top \tilde{\beta})}$ sono i pesi associati alla funzione $\psi(\cdot)$, con $\hat{w}_i \in [0,1]$, $\hat{y}_i = x_i^\top \tilde{\beta}$ il valore predetto per l'i-esima unità e $E_{f_\beta}(\cdot)$ il valore atteso rispetto al modello $f(y;\beta)$. Il sistema di pesi è costruito in modo da attribuire maggior peso alle osservazioni conformi al modello e di limitare quello delle osservazioni anomale.

Posto $\delta_i = (\hat{y}_i - E_{f_\beta} y_i)$, è possibile scomporre $\sigma^2 \Gamma_k$ come

$$\sigma^2 \Gamma_k = \sum_{i=1}^{n} V_{f_\beta}(\hat{w}_i \delta_i) + \sum_{i=1}^{n} \left[E_{f_\beta}(\hat{w}_i \delta_i) \right]^2$$

essendo V_{f_β} la varianza calcolata rispetto a $f(y;\beta)$. Inoltre, scomponendo il secondo addendo a destra dell'uguaglianza, si ottiene

$$\sigma^2 \Gamma_k = V_k + B_k - 2(AB)_k + A_k,$$

dove $V_k = \sum_{i=1}^{n} V_{f_\beta}(\hat{w}_i \delta_i)$, $A_k = \sum_{i=1}^{n} a_i^2$, $B_k = \sum_{i=1}^{n} b_i^2$, $(AB)_k = \sum_{i=1}^{n} a_i b_i$, con $a_i = E_{f_\beta}(\hat{w}_i \epsilon_i)$, in cui $\epsilon_i = y_i - x_i^\top \beta$ e $b_i = E_{f_\beta}(\hat{w}_i (y_i - \hat{y}_i))$. Utilizzando

i risultati di Ronchetti e Staudte (1994) e trascurando opportuni termini, si arriva alla versione robusta dell'indice C_p, data da

$$RC_p = \frac{W_k}{\tilde{\sigma}^2} - (U_k - V_k),$$

con $W_k = \sum_{i=1}^{n} \hat{w_i}^2 (y_i - \hat{y}_i)^2$, $U_k = \sum_{i=1}^{n} V_{f_\beta} [\hat{w}_i (y_i - \hat{y}_i)]$ e $\tilde{\sigma}^2$ stima robusta di σ^2 fatta nel modello completo. I modelli privilegiati saranno quelli con RC_p vicino a V_k.

3.2.3 Selezione robusta del modello nel modello lineare

Ronchetti *et al.* (1997) propongono un crietrio di selezione del modello ottimo, dove l'ottimalità va intesa come massima efficienza dello stimatore del parametro, vincolato ad un certo livello di robustezza. Come visto per i criteri precedenti, il primo passo è quello di reperire una stima robusta di β. In questo caso, però, β sarà stimato con lo stimatore ottimo di Hampel (si veda Hampel *et al.*, 1986) piuttosto che con lo stimatore di Huber, in quanto il primo garantisce la massima efficienza dato un vincolo di robustezza. Pertanto le equazioni di stima utilizzate saranno sempre del tipo (3.1), ma con una scelta della $\psi(\cdot)$ tale da garantire l'ottimalità dello stimatore.

Ottenuta la stima di β, il passo successivo consiste nell'affiancare a questa stima un criterio di selezione del modello robusto basato sull'errore di predizione. Chiaramente si ricorre ad un criterio più robusto rispetto a

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \tilde{e}_i^2 = min, \tag{3.14}$$

in cui $\hat{y}_i$ è il valore predetto per l'i-esima unità, con la semplice adozione di una funzione degli errori di predizione, $\tilde{e}_i$, più robusta rispetto al quadrato e, chiaramente, a valori positivi. Se si indica con $\rho(\cdot)$ questa funzione, allora si sceglierà il modello per cui

$$\sum_{i=1}^{n} \rho(\tilde{e}_i) = min. \tag{3.15}$$

Molto spesso, non disponendo di ulteriori unità statistiche per valutare la versatilità/flessibilità del modello a spiegare osservazioni future, si ricorre

al *boostrap* (si veda Efron e Tibshirani, 1993). Quando la numerosità campionaria è sufficientemente elevata, il campione viene suddiviso in due parti: una parte viene chiamata *construction sample*, di numerosità n_c e l'altra, *validation sample*, di numerosità n_v; evidentemente $n_v = n - n_c$, se si fissa la numeroristà di n_c. I due campioni sono utilizzati congiuntamente nella fase di stima, mentre si usa solo il campione di verifica per il calcolo del criterio (3.15). Naturalmente, le suddivisioni del campione originario vengono effettuate più volte al fine di eliminare l'arbitrarierà di un'unica suddivisione; generalmente il numero di ripetizioni di tale suddivizione è effettuata per B, volte con $B \ll n!/(n-n_c)!n_c!$ e n_c fissato.

3.3 Conclusioni

In questo capitolo si sono presentati alcuni metodi di selezione del modello basati su equazioni di stima. In particolare, le proposte più interessanti sembrano essere quelle di Pan (2001a, 2001b) e di Ronchetti (1985). Se da un lato questi crietri di selezione del modello sono interessanti soprattutto se si considerano le possibili applicazioni, dall'altro non è stato dedicato un ampio studio per indagare il loro comportamento sia con dati reali sia con studi di simulazione. Quello che si farà nel capitolo successivo è di mettere alla prova, in situazioni molto semplici, questi criteri di selezione del modello tramite studi di simulazione.

Capitolo 4

Alcune criticità sui metodi di selezione del modello basati su equazioni di stima

Nei due capitoli precedenti è stata messa in evidenza la flessibilità delle equazioni di stima nei problemi di inferenza, nel caso in cui il modello statistico $\mathcal{F}$ possa essere parametrico, semiparametrico o non parametrico. In particolare, è stato dato rilievo all'uso delle equazioni di stima per determinare alcuni criteri di selezione del modello. Mimando i risultati discussi nell'approccio tradizionale alla scelta del modello, avviata da Akaike nel 1973 (si veda Capitolo 1), nel Capitolo 3 si è visto come alcuni autori hanno proposto di minimizzare una generalizzazione della discrepanza di KL, che può essere definita come

$$KLG = c - E_{f_0}\left[\ell_{ps}(\vartheta)\right], \tag{4.1}$$

dove c è un'opportuna costante, $\ell_{ps}(\vartheta)$ indica una opportuna pseudo-verosimiglianza per ϑ derivata dall'equazione di stima e $E_{f_0}(\cdot)$ denota il valore atteso sotto il vero modello. Mentre usualmente la discrepanza di KL può essere interpretata come una misura di divergenza tra la distribuzione dei dati futuri generati dalla variabile Y e quella prevista dal modello, la KLG non costituisce necessariamente una distanza.

Il presente capitolo ha l'obiettivo di analizzare empiricamente i metodi

di selezione del modello discussi nel Capitolo 3, che si basano sulla minimizzazione di $-E_{f_0}[\ell_{ps}(\vartheta)]$ e di evidenziarne gli aspetti critici. Si mostrerà empiricamente, mediante alcuni studi di simulazione, perché basare la selezione del modello su una discrepanza generalizzata del tipo (4.1) può fornire risultati poco affidabili. Questo accade in quanto la discrepanza generalizzata (4.1) può essere, in alcuni casi, negativa. In questo capitolo sono riportati alcuni esempi.

Il primo esempio, affrontato nel §4.1, parte dalla KL classica e si concentra nella situazione della scelta della funzione di legame nell'ambito di un GLM di Poisson.

Nel secondo esempio, affrontato nel §4.2, si sviluppa l'esempio precedente, in quanto i modelli considerati sono sempre appartenenti alla categoria dei GLM, ma mette a confronto un modello di Poisson classico e un modello di quasi-verosimiglianza. Questa è una situazione di interesse pratico, quando si desidera modellare dati di conteggio in presenza di sospetta sovradispersione (si veda Fassina *et al.*, 2008).

Nel terzo esempio, affrontato nel §4.3, si fa riferimento alla teoria della robustezza e lo studio vuole mettere a confronto una verosimiglianza classica e una quasi-verosimiglianza derivata da equazioni di stima robuste.

4.1 Validità della distanza di KL in un esempio tradizionale

Ad illustrazione dell'esito prodotto dai tradizionali AIC e TIC nel caso di famiglie di modelli non annidati, si considera una prima applicazione nell'ambito di un problema della scelta della funzione di legame in un GLM. In particolare, si supponga di voler effettuare una regressione di Poisson nella situazione più semplice di una sola variabile esplicativa. Siano $y = (y_1, \ldots, y_n)$ i valori osservati sulla variabile risposta e $x = (x_1, \ldots, x_n)$ i valori della variabile esplicativa. Si assume $g(\mu_i) = \eta_i$, con $\mu_i = E_{f_0}(Y_i)$, $g(\cdot)$ funzione di legame, $\eta_i = \beta_0 + \beta_1 x_i$ predittore lineare e β_0 e β_1 i parametri di regressione.

Come è noto, per un GLM con distribuzione di Poisson, la funzione di

legame canonica è $g(\mu_i) = \log \mu_i$; in alternativa una possibile funzione di legame è la radice quadrata, ossia $g(\mu_i) = \sqrt{\mu_i}$. In questo esempio si desidera valutare, tramite simulazione, l'esito prodotto dall'AIC e dal TIC nella selezione tra due modelli di Poisson con diversa funzione di legame. Si noti che, in questo caso, entrambi i modelli appartengono alla classe parametrica $\mathscr{F} = \left\{ f(y;\mu) = e^{-\mu}\frac{\mu^y}{y!}, \mu \in \mathbb{R}^+, y \in \mathbb{N} \right\}$.

Nelle assunzioni fatte, l'espressione della verosimiglianza per i modelli considerati è data da

$$\ell(\beta_0, \beta_1) = \sum_{i=1}^{n} y_i \log \mu_i - \mu_i - \log y_i!. \tag{4.2}$$

Chiaramente, nel caso in cui il modello considerato sia quello di Poisson con funzione di legame radice quadrata si ha $\mu_i = (\beta_0 + \beta_1 x_i)^2$, mentre nel caso del modello con funzione di legame logaritmo si ha $\mu_i = \exp(\beta_0 + \beta_1 x_i)$, $i = 1 \dots n$. L'espressione dell'AIC è ricavabile immediatamente. Infatti, si ha

$$AIC = -2 \sum_{i=1}^{n} (y_i \log \hat{\mu}_i - \hat{\mu}_i - \log y_i!) + 4, \tag{4.3}$$

con $\hat{\mu}_i = g^{-1}(\hat{\beta}_0 + \hat{\beta}_1 x_i)$, $\hat{\beta}_0$ e $\hat{\beta}_1$ stime di massima verosimiglianza di β_0 e β_1. Per calcolare il TIC sono necessarie le matrici $\widehat{J}$ e $\widehat{K}$, definite nel Capitolo 1. Per il modello di Poisson con funzione di legame radice quadrata si ha

$$\widehat{J} = 4 \begin{pmatrix} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} & \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2 x_i}{\hat{\mu}_i} \\ \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2 x_i}{\hat{\mu}_i} & \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2 x_i^2}{\hat{\mu}_i} \end{pmatrix} \tag{4.4}$$

e

$$\widehat{K} = -2 \begin{pmatrix} \sum_{i=1}^{n} \frac{(y_i + \hat{\mu}_i)}{\hat{\mu}_i} & \sum_{i=1}^{n} \frac{(y_i + \hat{\mu}_i) x_i}{\hat{\mu}_i} \\ \sum_{i=1}^{n} \frac{(y_i + \hat{\mu}_i) x_i}{\hat{\mu}_i} & \sum_{i=1}^{n} \frac{(y_i + \hat{\mu}_i) x_i^2}{\hat{\mu}_i} \end{pmatrix}, \tag{4.5}$$

con $\hat{\mu}_i = (\hat{\beta}_0 + \hat{\beta}_1 x_i)^2$. Mentre per il modello di Poisson con funzione di

legame logaritmo si ha

$$\widehat{J} = \begin{pmatrix} \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 & \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 x_i \\ \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 x_i & \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 x_i^2 \end{pmatrix} \tag{4.6}$$

e

$$\widehat{K} = - \begin{pmatrix} \sum_{i=1}^{n} \hat{\mu}_i & \sum_{i=1}^{n} \hat{\mu}_i x_i \\ \sum_{i=1}^{n} \hat{\mu}_i x_i & \sum_{i=1}^{n} \hat{\mu}_i x_i^2 \end{pmatrix}, \tag{4.7}$$

con $\hat{\mu}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)$.

Per valutare il comportamento del TIC e dell'AIC si sono stimate le probabilità di corretta selezione, come $n^o corrette\ assegnazioni/n_r$, in cui n_r rappresenta il numero di replicazioni della simulazione.

La Tabella 4.1 riporta le probabilità di corretta selezione derivanti dallo studio di simulazione basato su $n_r = 5000$ replicazioni. Nella prima riga viene indicato il modello da cui si generano i dati. Nella simulazione si è posto $\beta_0 = 1$ e $\beta_1 = 3$, e la variabile esplicativa è stata generata da una variabile uniforme nell'intervallo $(0, 1)$.

Dalla Tabella 4.1 si nota come al crescere della numerosità campionaria n crescono anche le probabilità di corretta selezione, avvicinandosi al 100%. È, inoltre, possibile constatare che in questa applicazione AIC e TIC forniscono risultati molto simili. Il vantaggio dell'AIC è una maggiore semplicità di calcolo; inoltre il suo valore viene fornito in automatico da molti pacchetti statistici (vedi $glm(\cdot)$ in R).

Si ricorda che sia l'AIC che il TIC si basano sulla discrepanza di KL (1.2) che, per definizione, assume valori positivi o al più nulli. Ciò che garantisce la positività di tale discrepanza è la diseguaglianza (1.1) di Wald (Wald, 1949, Lemma 1). Per facilità di esposizione, si consideri la situazione più semplice in cui si confrontano due modelli di Poisson con differenti medie, indicate con μ_0 e μ_1 rispettivamente. Una rappresentazione grafica della distanza di KL, per $(\mu_0, \mu_1) \in (0, 30]^2$, è riportata in Figura 4.1.

Il grafico della KL mostra chiaramente che la distanza assume valori strettamente positivi, e nulli solo quando $\mu_0 = \mu_1$. Guardando la sezione

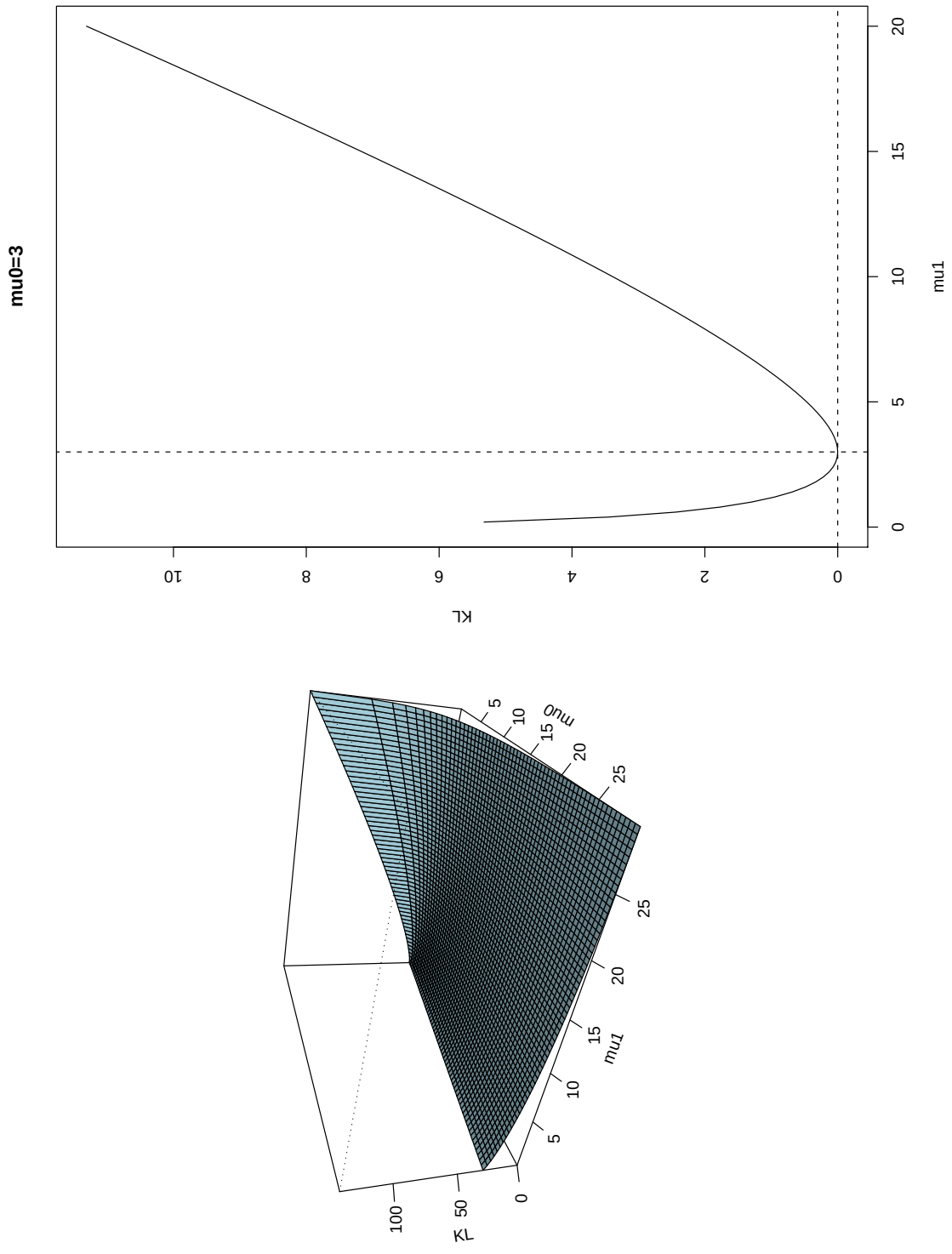

Figura 4.1: Discrepanza di KL per modelli di Poisson con diversa media.

	Legame $\sqrt{(\cdot)}$		**Legame** $\log(\cdot)$	
n	TIC	AIC	TIC	AIC
10	71.78	71.56	83.60	84.20
20	71.32	72.80	90.50	90.18
50	79.18	80.42	97.70	97.48
100	86.98	87.66	99.68	99.28

Tabella 4.1: Probabilità di corretta selezione nell'esempio 4.1

della distanza di KL per $\mu_0 = 3$, si ha una rappresentazione grafica della diseguaglianza di Wald.

4.2 Studio dei criteri basati sulle equazioni di stima

Come visto nel Capitolo 3, i metodi di selezione del modello di Pan (2001b) e Ronchetti (1985) sono essenzialmente basati sulla minimizzazione della

$$-E_{f_0}\left[\ell_{ps}(\vartheta)\right], \tag{4.8}$$

giustificata in base alla (4.1). Come appena ricordato, nel caso di una funzione di verosimiglianza genuina, basare la selezione del modello sulla minimizzazione della (4.8) deriva dalla discrepanza di KL, e tale minimizzazione è garantita dalla diseguaglianza di Wald. L'utilizzo da parte di Varin e Vidoni (2005) della (4.8) è giustificata dal fatto che la funzione di verosimiglianza composita è definita come un prodotto di contributi di verosimiglianza e che in tale caso la distanza KLG risulta combinazione lineare di discrepanze di KL (si sfrutta la proprietà di additività della discrepanza di KL), perciò somma di contributi positivi o al più nulli.

Quando, invece, si utilizzano altre pseudo-verosimiglianze derivate da equazioni di stima, come la quasi-verosimiglianza, non c'è alcuna garanzia che la discrepanza KLG (4.1) sia positiva. Questo accade in quanto non è più soddisfatta la diseguaglianza di Wald, non essendoci una densità a monte della funzione di pseudo-verosimiglianza.

La dimostrazione della non affidabilità dei metodi di selezione del modello di Pan (2001b) e Ronchetti (1985) basati sulla (4.8) e sulla quasi-verosimiglianza viene effettuata tramite studi di simulazione. Il risultato più importante è che vengono forniti almeno due esempi in cui il QIC, l'$RTIC$ e l'$RAIC$ basati sulla (4.1) non risultano affidabili, in quanto la KLG assume valori negativi. Questi esempi hanno ovviamente la messa in discussione dei criteri di selezione del modello basati sulla (4.1), e in particolare delle (3.6) e (3.10) di Pan e Ronchetti. Questo, è una conseguenza del fatto che la (4.1) non è, in generale, una discrepanza nel caso in cui si considerino pseudo-verosimiglianze basate su equazioni di stima.

4.2.1 Un'applicazione nei GLM

Come nell'esempio del §4.1, si supponga di voler confrontare due modelli appartenenti alla categoria dei GLM. In particolare, si confrontano un modello di Poisson, con funzione di legame radice quadrata, e un modello di quasi-verosimiglianza, con funzione di legame logaritmo. Per entrambi i modelli si assume la funzione di varianza pari a $V(\mu) = \mu$, ma vale $V_{f_0}(Y) = \mu$ per il modello di Poisson e $V_{f_0}(Y) = \phi V(\mu)$ per il modello di quasi-verosimiglianza, con $\phi > 0$ parametro di dispersione. La presenza del parametro di dispersione evidenzia la possibilità di maneggiare dati con sovra/sottodispersione, utilizzando il modello di quasi-verosimiglianza.

Il confronto di questi due modelli non nasce dalla sola esigenza di fornire un'esempio in questa tesi. Infatti, in situazioni di interesse applicativo (si veda, ad esempio, Fassina *et al.*, 2008) può accadere di dover modellare dati di conteggio che presentano una lieve sovradispersione. In questo caso, il modello di Poisson con legame radice quadrata potrebbe rappresentare una valida alternativa al modello di quasi-verosimiglianza, in quanto la funzione di legame radice quadrata ha la proprietà di stabilizzare la varianza. Si tratta di scegliere non solo tra due modelli non annidati, ma di scegliere anche tra un modello parametrico e un modello semi-parametrico. In particolare il modello di Poisson appartiene alla classe parametrica $\mathscr{F}$, vista nel §4.1, perciò la funzione di log-verosimiglianza per $\beta = (\beta_0, \beta_1)$ è analoga alla (4.2).

Per il modello di quasi-verosimiglianza l'espressione della funzione di quasi-verosimiglianza è data da

$$\begin{aligned} \ell_Q(\beta_0, \beta_1) &= \sum_{i=1}^{n} \int_c^{\mu_i} \frac{y_i - t}{\phi V(t)} dt \\ &= \sum_{i=1}^{n} \frac{1}{\phi}(y_i \log \mu_i) - \mu_i, \end{aligned}$$

con $\mu_i = g^{-1}(\beta_0 + \beta_1 x_i)$. Con le espressioni delle verosimiglianze fornite è immediato calcolare l'AIC e l'$RAIC$ per i due modelli. Infatti, per il modello di Poisson si ha

$$AIC = -2 \sum_{i=1}^{n} (y_i \log \hat{\mu}_i - \hat{\mu}_i - \log y_i!) + 4,$$

con $\hat{\mu}_i = (\hat{\beta}_0 + \hat{\beta}_1 x_i)^2$, mentre per il modello di quasi-verosimiglianza si ha

$$QAIC = -\frac{2}{\phi} \sum_{i=1}^{n} (y_i \log \hat{\mu}_i - \hat{\mu}_i) + 4,$$

con $\hat{\mu}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)$. Per entrambi i modelli $\hat{\beta}_0$ e $\hat{\beta}_1$ sono le stime di β_0 e β_1. Si noti che il $QAIC$ è un crietrio di selezione del modello che ingloba l'AIC nel caso particolare in cui l'equazione di stima utilizzata è la *score* di verosimiglianza.

Per il calcolo del TIC e del QIC sono necessarie le matrici, rispettivamente, $\widehat{J}$, $\widehat{K}$ e $\widehat{J}_Q$, $\widehat{K}_Q$ definite nei Capitoli 1 e 3. Per il modello di Poisson le matrici sono

$$\widehat{J}_Q = \frac{1}{\phi} \begin{pmatrix} \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 & \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 x_i \\ \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 x_i & \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 x_i^2 \end{pmatrix} \tag{4.9}$$

e

$$\widehat{K}_Q = -\frac{1}{\phi} \begin{pmatrix} \sum_{i=1}^n \hat{\mu}_i & \sum_{i=1}^n \hat{\mu}_i x_i \\ \sum_{i=1}^n \hat{\mu}_i x_i & \sum_{i=1}^n \hat{\mu}_i x_i^2 \end{pmatrix}, \tag{4.10}$$

con $\hat{\mu}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_i)$.

Per il modello di quasi-verosimiglianza, invece, le matrici si ottengono riscalando la (4.10) e la (**??**) per il parametro di dispersione, ossia

$$\widehat{J}_Q = \frac{1}{\phi} \begin{pmatrix} \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 & \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 x_i \\ \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 x_i & \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 x_i^2 \end{pmatrix}$$

e

$$\widehat{K}_Q = \frac{1}{\phi} \begin{pmatrix} \sum_{i=1}^n \hat{\mu}_i & \sum_{i=1}^n \hat{\mu}_i x_i \\ \sum_{i=1}^n \hat{\mu}_i x_i & \sum_{i=1}^n \hat{\mu}_i x_i^2 \end{pmatrix},$$

Anche in questo esempio, come nel precedente, si vuole valutare il comportamento dei metodi di selezione del modello. Nello specifico i criteri messi a confronto sono il TIC contro il QIC e l'AIC contro il $QAIC$.

Nella simulazione i valori della variabile esplicativa sono stati generati da un'uniforme su $(0, 1)$ e si sono fissati $\beta_0 = 1$ e $\beta_1 = 3$. La generazione di dati dal modello di Poisson con funzione di legame radice quadrata è stata effettuata come nell'esempio precedente, ovvero generando da un modello di Poisson di media $\mu_i = (\hat{\beta}_0 + \hat{\beta}_1 x_i)^2$. Invece, per generare dati dal modello di quasi-verosimiglianza si può far riferimento dalla distribuzione binomiale negativa con opportuni parametri (si veda, per esempio, McCullagh e Nelder, 1988, §6.2.2). Si noti, inoltre, che si devono stimare non solo β_0 e β_1, ma anche il parametro di dispersione ϕ. Questo può essere stimato in maniera consistente tramite $\hat{\phi} = \frac{1}{n-k} \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$, e tale valore viene poi sostituito nel $QAIC$ e nel QIC (si veda Pan, 2001b).

Le Tabelle 4.2 e 4.3 riportano le probabilità di corretta selezione basate sulla discrepanza di KL generalizzata (4.1). Per la Tabella 4.3 si riportano vari scenari al variare del parametro di dispersione ϕ.

La Tabella 4.2 mostra il confronto tra il modello di Poisson e quello di quasi-verosimiglianza quando il vero modello è quello di Poisson. Si nota subito un'anomalia in questo confronto, in quanto al crescere di n la probabilità di corretta assegnazione diminuisce. Il fatto che le probabilità non siano eccessivamente alte lo si può giustificare dal fatto che i modelli sono molto

	Poisson	
n	TIC	AIC
10	53.3	53.3
20	51.0	50.8
50	47.9	47.7
100	45.9	45.9

Tabella 4.2: Probabilità di corretta assegnazione nell'esempio 4.2.1. I dati provengono dal modello di Poisson

simili. Ciò che invece appare anomalo è il peggioramento nelle probabilità di corretta assegnazione al crescere dell'informazione campionaria.

La Tabella 4.3 riporta invece risultati coerenti con quello che ragionevolmente ci si aspetta dal criterio di selezione del modello. Si nota che nei casi estremi di sovradispersione o sottodisperione, rispettivamente $\phi = 1.8$ e $\phi = 0.2$, il criterio non mostra uno stesso comportamento in quanto, nel primo caso, risulta essere più selettivo già per n basso. Analogamente, i due scenari centrali presentano anomalie simili: il criterio sceglie meglio nel caso di lieve sottodisperione e peggio in quello di lieve sovradispersione.

Il confronto di modelli di verosimiglianza e quasi-verosimiglianza mediante la (4.1) non risulta soddisfacente in questo esempio, in particolare per i risultati visti nella Tabella 4.2. La giustificazione è che basare la selezione del modello sulla (4.8) vuol dire affidarsi a una distanza che non è tale, poiché può assumere valori in $\mathbb{R}$, e non solo positivi. Questo accade in quanto non è più soddisfatta la diseguaglianza di Wald che assicura l'ordinamento crescente, rispetto al vero modello, dei valori attesi delle log-verosimiglianze, ottenute, però, come il logaritmo di una densità.

Il calcolo e la rappresentazione della KLG sono d'aiuto per confermare le affermazioni di cui sopra. Per facilità di esposizione, si consideri la situazione in cui il vero modello generatore dei dati, f_0, sia quello di Poisson di media μ_0 e sia $\ell_Q(\cdot)$ il modello di quasi-verosimiglianza con media μ_1. La distanza

	Quasi-verosimiglianza	
	$\phi = 0.2$	
n	QIC	QAIC
10	62.1	60.6
20	71.2	70.5
50	84.7	83.9
100	95.0	94.7
	$\phi = 0.9$	
10	88.3	86.9
20	95.2	94.9
50	99.7	99.7
100	100.0	100.0
	$\phi = 1.1$	
10	53.8	53.2
20	66.1	64.8
50	75.2	74.2
100	83.0	82.6
	$\phi = 1.8$	
10	82.2	81.9
20	94.6	94.2
50	99.8	99.8
100	99.9	99.9

Tabella 4.3: Probabilità di corretta assegnazione nell'esempio 4.2.1. I dati provengono dal modello di quasi-verosimiglianza.

KLG (4.1) tra f_0 e ℓ_Q, per una sola osservazione, può essere scritta come

$$\begin{aligned} KLG &= E_{f_0}\left[\log f(y;\mu_0)\right] - E_{f_0}\left[\ell_Q(\mu_1)\right] = \\ &= -\mu_0 + \mu_0 \log \mu_0 - E_{f_0}\left[\log y!\right] + \frac{1}{\phi}(\mu_1 - \mu_0 \log \mu_1), \end{aligned} \tag{4.11}$$

per μ_0, μ_1 e ϕ appartenenti a $\mathbb{R}^+$. Nella Figura 4.2 è rappresentata la (4.11) e si nota come questa assuma valori sia negativi che positivi, per $(\mu_0, \mu_1) \in (0, 30]^2$ e diversi valori di ϕ.

Anche se si scambiano i ruoli tra f_0 e ℓ_Q, la KLG continua ad assumere sia valori positivi che negativi. Per questo caso si omettono i calcoli e si riporta solamente il grafico, nella Figura 4.3, della KLG tra il modello basato sulla quasi-verosimiglianza e quello basato sulla funzione di verosimiglianza. Questo andamento della KLG conferma che i risultati mostrati nelle Tabelle 4.2 e 4.3 sono poco affidabili

Questo secondo esempio mostra, quindi, come con i mezzi attuali a disposizione, nello specifico $QAIC$ e QIC, non sia possibile effettuare una scelta tra un modello di regressionedi Poisson e un modello di quasi-verosimiglianza.

4.2.2 Inferenza robusta

Nella sezione precedente si è fornito un esempio sulla non adeguatezza della (4.8) ai fini della selezione del modello, quando si utilizza una funzione di quasi-verosimiglianza derivante da un'equazione di stima. In questo paragrafo si considera un secondo esempio, in cui la quasi-verosimiglianza è basata su un'equazione di stima robusta (si veda Ronchetti, 1985).

L'esempio riguarda il confronto tra un modello più semplice, quello esponenziale per il parametro di scala ϑ, che viene assunto come modello centrale per la costruzione della $\ell_Q(\vartheta)$ e un modello più complesso, che considera un possibile allontanamento dal modello esponenziale. Questo confronto può nascere, in un problema applicativo, quando sono presenti alcune osservazioni anomale nella coda destra della distribuzione empirica delle osservazioni (si veda, ad esempio, Adimari e Ventura, 2002). Volendosi tutelare dalla scorretta specificazione del modello, si sceglie di basare l'inferenza su ϑ tramite

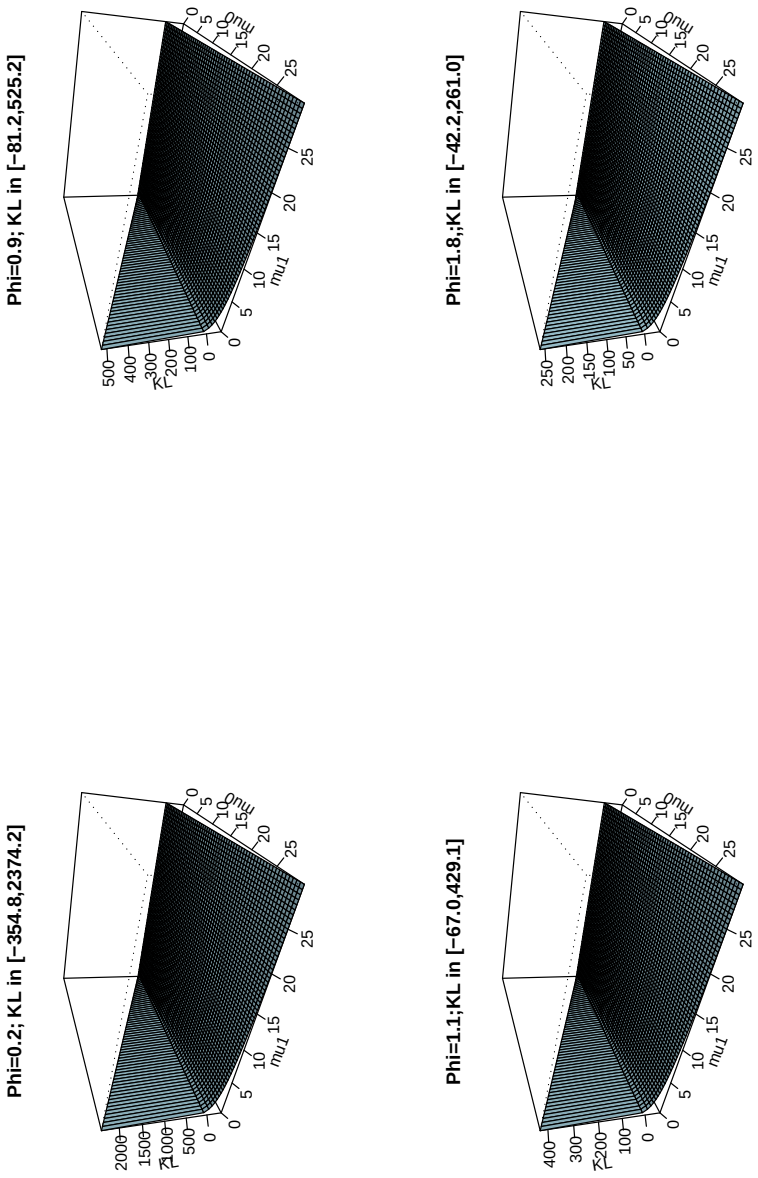

Figura 4.2: KLG tra f_0 e ℓ_Q per quattro scelte di ϕ nell'esempio 4.2.1

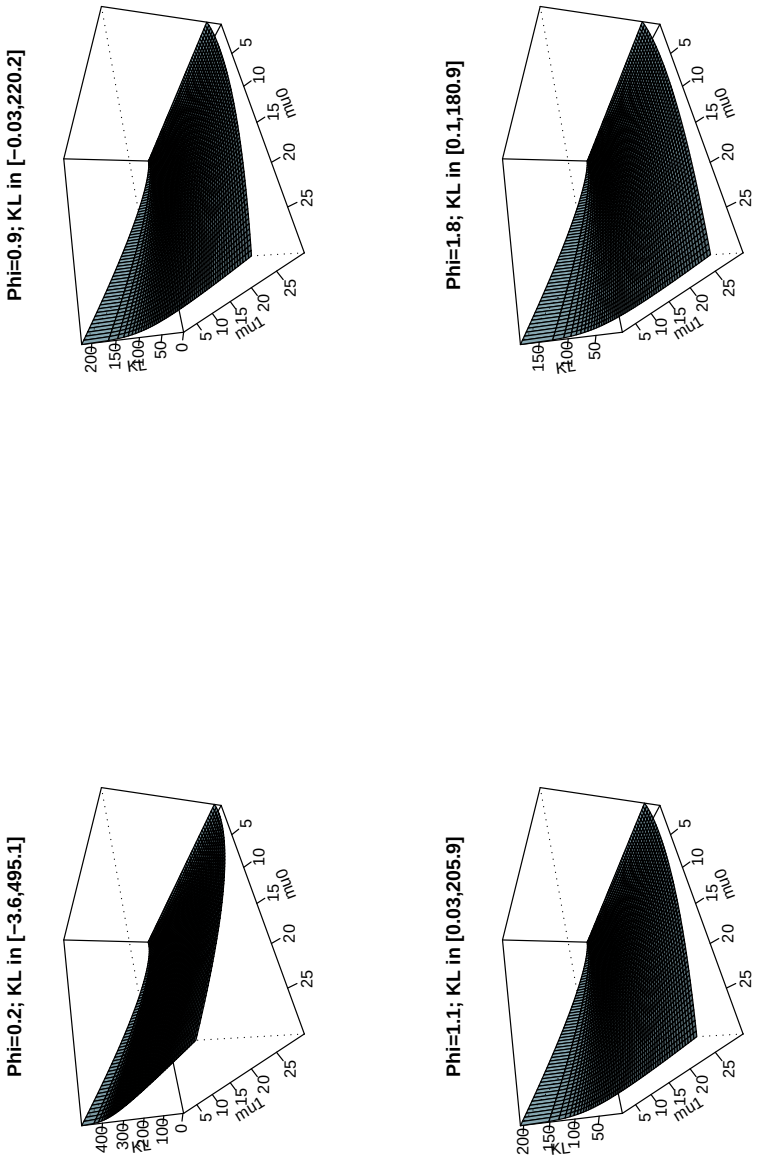

Figura 4.3: KLG tra ℓ_Q e f_0 per quattro scelte di ϕ nell'esempio 4.2.1

l'equazione di stima robusta proposta da Hampel (Hampel, 1986) anzichè per mezzo della consueta *score* di verosimiglianza. In alternativa si considera il modello gamma con parametro di forma di poco superiore a 1. Pertanto, si desidera fare il confronto tra un modello di quasi-verosimiglianza derivante dall'equazione di stima ottimale di Hampel e basata sul modello esponenziale, e il modello gamma (con parametro di forma α fissato), basato invece sulla funzione di verosimiglianza propria. Come nell'esempio precedente si confrontano l'$RTIC$ e l'$RAIC$ proposti da Ronchetti (1985) contro, rispettivamente, il TIC e l'AIC. Si ricorda che il TIC e l'AIC possono essere considerati casi particolari dell'$RTIC$ e dell'$RAIC$ quando le equazioni di stima utilizzate sono quelle basate sulla funzione di verosimiglianza.

L'equazione di stima robustta per ϑ è definita come

$$\psi(\vartheta; y) = \sum_{i=1}^{n} \psi(\vartheta; y_i), \tag{4.12}$$

dove $\psi(\vartheta; y) = \max\{-b, min\{b, a - y\}\}$, con a e b opportune costanti che garantiscono la robustezza dell'equazione di stima. In questa applicazione si considera $a = 0.8636$ e $b = 1.129$. La funzione di quasi-verosimiglianza per ϑ è data da (si veda Adimari e Ventura, 2002)

$$\begin{aligned}
\ell_Q(\vartheta) &= \sum_{i=1}^{n} \int \psi(\vartheta; t) dt \\
&= A \sum_{i=1}^{n} \left[(b \log \vartheta + c_{1i}) I_{\vartheta < (a-b)/y_i} \right. \\
&+ (a \log \vartheta - \vartheta y_i) I_{(a-b)/y_i \leq \vartheta < (a+b)/y_i} \\
&- \left. (b \log \vartheta + c_{2i}) I_{\vartheta \geq (a+b)/y_i} \right],
\end{aligned}$$

con

$$\begin{aligned}
c_{1i} &= (a - b) \left[\log \frac{a - b}{y_i} - 1 \right] \\
c_{2i} &= (a - b) \left[\log \frac{a + b}{y_i} - 1 \right].
\end{aligned}$$

Nella funzione di quasi-verosimiglianza il termine A è una quantità legata al valore della differenza $(a-b)$ che, essendo nel caso specifico minore di zero, risulta definita come

$$A = \frac{(b+a+1)exp(-a-b)-1}{a(a-2)-2(b+1)exp(-a-b)+2}.$$

La funzione di log-verosimiglianza per ϑ, per $\alpha = \alpha_0$ fissato, è

$$\ell(\vartheta) = \sum_{i=1}^{n}(\alpha_0 \log \vartheta - \vartheta y_i + (\alpha_0 - 1)\log y_i - \log \Gamma(\alpha_0)),$$

con funzione *score*

$$\ell_*(\vartheta) = \sum_{i=1}^{n}(\frac{\alpha_0}{\vartheta} - y_i),$$

e l'informazione osservata risulta

$$j(\vartheta) = \frac{n\alpha_0}{\vartheta}.$$

Il $RTIC$ per questo esempio ha la seguente espressione

$$RTIC = -2\ell_Q(\tilde{\vartheta}) + 2\frac{\tilde{\vartheta}}{A},$$

in cui $\tilde{\vartheta}$ è la stima robusta di ϑ, soluzione di $\sum_{i=1}^{n} \max\{-b, min\{b, a-y_i\}\} = 0$.

Mentre il TIC per il modello gamma è dato da

$$TIC = -2\ell(\hat{\vartheta}) + \frac{2n}{\alpha_0}\left(\alpha_0 - \hat{\vartheta}\overline{y}\right).$$

in cui $\hat{\vartheta}$ è la stima di massima verosimiglianza di ϑ e $\overline{y}$ è la media campionaria.

Nello studio di simulazione condotto, i dati sono generati dalla distribuzione esponenziale con $\vartheta = 1$ e dalla distribuzione gamma con $\vartheta = 1$ e α prossimo a 1. Si noti che fissare $\alpha \approx 1$ significa mantenersi in un intorno del modello esponenziale, dove le procedure robuste continuano a valere. Fissare α grande implicherebbe, di fatto, di optare senza alcun dubbio per il modello gamma.

La Tabella 4.4 fornisce le probabilità di corretta selezione per il confonto del modello di quasi-verosimiglianza versus quello Gamma. Quando il vero

	Gamma		**Quasi-verosimiglianza**	
	$\alpha = 1.03$		$\alpha = 1.03$	
n	TIC	AIC	RTIC	RAIC
10	0.0	2.3	99.2	96.8
20	0.0	0.0	100.0	99.7
50	0.0	0.0	100.0	100.0
100	0.0	0.0	100.0	100.0
	$\alpha = 1.5$		$\alpha = 1.5$	
10	19.9	35.8	95.2	79.8
20	12.5	20.1	99.4	96.2
50	3.5	5.3	100.0	100.0
100	0	0	100.0	100.0
	$\alpha = 2$		$\alpha = 2$	
10	95.1	99.6	43.3	8.0
20	99.0	100.0	31.8	7.5
50	100.0	100.0	18.0	4.7
100	100.0	100.0	7.4	2.9

Tabella 4.4: Probabilità di corretta assegnazione nell'esempio 4.2.2

modello è quello esponenziale, si nota che per $\alpha = 1.03$ e $\alpha = 1.5$ il modello basato sull'equazione di stima robusta riesce a discriminare bene il modello esponenziale da quello gamma per tutte le numerosità campionarie. Chiaramente, per $\alpha = 2$ sia l'$RTIC$ che l'$RAIC$ falliscono, presumibilmente perché ci si allontana troppo dal vero modello. Quando però, si analizza l'altro scenario, cioè quando si genera dalla distribuzione gamma, succede che i criteri tradizionali di selezione non riescano a discriminare tra i due modelli. La situazione resta pressocché inalterata fino a quando α non è pari a 2: in questo caso la diversità tra i due viene colta.

I risultati di cui sopra sono un chiaro sintomo della non adeguatezza dell'$RAIC$ e dell'$RTIC$ nella valutazione tra modelli di verosimiglianza e quasi-verosimiglianza. Infatti, una valutazione numerica della (4.1) fornisce sia valori negativi che positivi. La Tabella 4.5 riporta i valori stimati, per $n_r = 5000$, della KLG per questo esempio al variare di α.

	Gamma	**Quasi-verosimiglianza**
	$n = 100$	$n = 100$
α	KLG	KLG
1.03	-132.3	131.7
1.5	-121.8	158.8
2	-84.4	187.5
3	114.3	245.9

Tabella 4.5: Stima della KLG per alcune scelte di α

Come nel §4.2 si ha che la KLG assume valori negativi e positivi. In particolare, laddove il TIC e l'AIC sembrano funzionare peggio quando invece non dovrebbero, ovvero quando i dati sono generati da una gamma, la KLG tende ad assumere valori sempre più grandi al crescere di α.

4.3 Conclusioni

A completamento di quanto presentato nel Capitolo 3 sul criterio $RTIC$ e QIC, questo capitolo ha presentato due applicazioni volte a supplire la quasi

totale assenza di risultati empirici in letteratura, riguardo il comportamento di questi criteri di selezione del modello.

Entrambi gli esempi, uno nel contesto della quasi-verosimiglianza per i GLM, per saggiare il QIC, e uno nel contesto della quasi-verosimiglianza derivata da equazioni di stima robuste, per studiare l'$RTIC$, ne hanno evidenziato le carenze e la non affidabilità. Come è stato più volte evidenziato lungo questo capitolo, i due criteri sono da un punto di vista pratico carenti poiché mimano i criteri tradizionali, senza prendere in considerazione la diversa natura dei modelli trattati da tali criteri. Questo induce a basare la selezione del modello sulla KLG che, in questi casi, non èuno strumento coerente per il semplice fatto che non è una discrepanza.

Conclusioni

Obiettivo principale della tesi è indagare il comportamento dei metodi di selezione del modello basati su equazioni di stima proposti in letteratura, come generalizzazione dell'AIC e del TIC classici. Dagli studi empirici effettuati nel Capitolo 4 emerge che vi sono situazioni in cui l'$RAIC$ e il QIC non sono affidabili. Ciò è dovuto al fatto che la KLG non costituisce, in generale, una distanza.

Ciò che rende scorrettamente definiti il QIC (§3.1.2) e il $RTIC$ (§3.2.1) è il loro mero tentativo di mimare il TIC. Tale tentativo porta con sè, facendo diventare il cuore del criterio di selezione del modello basato su equazioni di stima, uno strumento pensato per il confronto di modelli parametrici, ovvero la discrepanza di KL. Si capisce che così facendo non è possibile slegarsi totalmente dai modelli parametrici, così come invece si vorrebbe, e questo porta a definire criteri di selezione del modello fallaci.

Da questi primi e semplici studi di simulazione appare evidente che se si vogliono fare progressi bisogna abbandonare la discrepanza di KL, o almeno così come originariamente pensata, e pensare a una qualche forma di discrepanza che sfrutti le proprietà delle equazioni di stima utilizzate.

Non ultimo problema, anzi, è quello di definire il concetto di selezione del modello in ambito semiparametrico, dove l'assunto distributivo è il primo a venire meno e quindi la possibilità di fare riferimento a un vero modello. Si noti che questo interrogativo può avere risposte ben diverse in relazione all'ambito di utilizzo del criterio di selezione del modello. Già in ambito robusto questo problema può venire meno, in quanto le equazioni di stima utilizzate fanno chiaro riferimento a un modello statistico parametrico. D'al-

tro canto, però, bisogna tener conto che le procedure inferenziali robuste sono tali da rimanere valide anche in un'intorno del modello parametrico assunto per i dati. Si capisce, in questo ambito, come la definizione di modello parametrico di riferimento ha contorni, letteralmente, sfumati.

Bibliografia

Adimari, G., Ventura, L. (2001). Robust inference for generalized linear models with application to logistic regression. *Statistics and Probability Letters*, **55**, 413-419.

Adimari, G., Ventura, L. (2002). Quasi-likelihood from M-estimators: A numerical comparison with empirical likelihood. *Statistical Methods & Applications*, **11**, 175-185.

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle.* Second International Symposium on Information Theory. Petrov B.N. and Csaki F., Budapest.

Akaike, H. (1995). Prediction and entropy. In: *A celebration of statistics*, Springer-Verlag, 1-24.

Burnham, K.P., Anderson, D.A. (2002). *Model selection and multimodel inference.* Springer, New York.

Desmond, A.F. (1997). Optimal estimating functions, quasi-likelihood and statistical modelling. *Journal of statistical planning and inference*, **60**, 77-104.

Davison, A.C. (2003). *Statistical models.* Cambridge University Press, Cambridge.

Efron, B., Tibshirani, R.J. (1993). *An introduction to the bootstrap.* Chapman & Hall, London.

Fassina A., Corradin M., El Mazloum R., Murer B., Furlan C., Montisci M., Ventura L. (2008). Silica levels and cancer: results of an ESEM study, Inalation Toxicology, in corso di pubblicazione.

Godambe, V.P. (1960). An Optimum Property of Regular Maximum Likelihood Estimation. *The Annals of Mathematical Statistics*, **31**, 1208-1211.

Godambe, V.P. (1991). *Estimating functions.* Clarendon Press, Oxford.

Gorman, J.W., Toman R.J. (1966). Selection of variables for fitting equations to data. *Technometrics*, **8**, 27-51.

Huber, P.J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, **35**, 73-101.

Huber, P.J. (1967). *The behaviour of maxim likelihood estimates under nonstandard conditions.* Proceedings of the Fith Berkley Symposium, University of California Press, Berkley.

Kullback, S.(1997). *Information Theory and Statistics.* Dover, Mineola.

Liang, K.Y, Zeger, S.L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, **73**,13-22.

Linhart, H., Zucchini, W. (1986). *Model selection.* John Wiley & Sons, New York.

Mallows, C.L. (1973). Some comments on CP. *Technometrics*, **15**, 661-675.

McCullagh, P., Nelder, J.A (1989). *Generalized linear models.* Chapman & Hall, London.

McDonald, B.W. (1993). Estimating logistic regression parameter for bivariate binary data. *Journal of the Royal Statistical Society.* B, **55**, 391-397.

Nelder, J.A., Wedderburn, R.W.M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society*. A, **135**, 370-384.

Pan, W. (2001a). Akaike's Information Criterion in Generalized Estimating Equations. *Biometrics*, **57**, 120-125.

Pan, W. (2001b). Model selection in estimating equations. *Biometrics*, **57**, 529-534.

Parzen, E., Tanabe K., Kitagawa G. (1998). *Selected papers of Hirotugu Akaike*. Springer, New York.**da citare nel testo**

Racugno, W. (1998). La selezione del modello statistico. Atti della XXXIX Riunione Scientifica della Società Italiana di Statistica, Sorrento, 14-17 Aprile 1998, vol.1,.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461-464.

Shibata, R. (1989). *From data to model*. Springer-Verlag, London.

Takeuchi, K. (1976). Distribution of information statistics and validity criteria of models. *Mathematical Science*, **153**, 12-18.

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

Wald, A. (1949). Note on the consistency of the maximum likilihood estimate. *The Annals of Mathematical Statistics*, **4**, 595-601.

Wedderburn, R.W.M. (1974). Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method. *Biometrika*, **61**, 439-447.

Zeger, S.L. (1988). The analysis of discrete longitudinal data: Commentary. *Statistics in Medicine*, **7**, 161-168.

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology*, **44**, 41-61.