Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in
Statistica per le Tecnologie e le Scienze

RELAZIONE FINALE

METODI DI VEROSIMIGLIANZA PER LA STIMA DELLA SOVRADISPERSIONE NEL MODELLO BINOMIALE NEGATIVO

Relatore: Prof. Nicola Sartori
Dipartimento di Scienze Statistiche

Laureando: Federico Castellan
Matricola N° 1099794

Anno Accademico 2016/2017

Indice

Introduzione

La presente relazione si propone di presentare e affrontare un problema specifico che talvolta emerge nell'ambito dei modelli di regressione per dati di conteggio: la sovradispersione.

Viene a tal scopo esposto il modello binomiale negativo, una generalizzazione del modello di Poisson che prevede l'introduzione di un parametro di dispersione, sul quale si desidera fare inferenza tramite gli strumenti di verosimiglianza.

Si propongono inoltre alcuni strumenti che permettono di fare inferenza sul parametro di dispersione con accuratezza elevata anche in situazioni svantaggiate, dovute ad esempio alla presenza di parametri di disturbo.

Basandosi su alcuni insiemi di dati reali, si presenta infine l'applicazione dei metodi trattati attraverso l'utilizzo del *software* `R` e uno studio di simulazione votato alla verifica dell'accuratezza e correttezza di alcuni di tali metodi.

La trattazione si snoda in quattro capitoli, che cercano di coprire ordinatamente gli argomenti proposti. Nel primo capitolo si offre una panoramica contestualizzante la teoria di verosimiglianza, utilizzata estensivamente in tutta la relazione.

Nel secondo capitolo si introduce, nel contesto dei modelli lineari generalizzati, il modello di regressione di Poisson e si affronta il problema della sovradispersione proponendo come possibile soluzione il modello binomiale negativo.

Il terzo capitolo tratta invece il problema dell'accuratezza degli strumenti di inferenza utilizzati. Il quarto e ultimo capitolo contiene l'analisi di due *datasets* reali e lo studio di simulazione.

1. Inferenza di verosimiglianza

In questo capitolo vengono presentate in linea generale un'insieme di procedure d'inferenza mirate a trattare specificatamente i modelli statistici parametrici. Esse si basano sulla funzione di verosimiglianza, introdotta da Ronald Fisher. Si provvede a fornirne una esposizione di base, non certo esaustiva, bensì indirizzata a introdurre il contesto e le notazioni principali che saranno presupposti nel seguito. Per approfondimenti si veda, ad esempio, Pace e Salvan (2001) oppure Azzalini (2001), sui quali è altresì basata l'esposizione.

1.1. Funzione di verosimiglianza

Il modello statistico

Un modello statistico $\mathcal{F}$ per i dati y è rappresentato da un insieme di funzioni di probabilità o, nel caso continuo, da un insieme di funzioni di densità di probabilità $p(y;\theta)$. Si scrive $\mathcal{F} = \{p(y;\theta), \theta \in \Theta\}$, dove θ è un parametro con valori in Θ, detto spazio parametrico. Si ha un modello statistico parametrico se $\Theta \subseteq \mathbb{R}^d$, per un qualche $d \in \mathbb{N}^+$. Il parametro θ si dice identificabile se esiste una corrispondenza biunivoca tra Θ e $\mathcal{F}$. Si assume che i modelli considerati siano parametrici e parametrizzati da un parametro identificabile $\theta \in \mathbb{R}^d$, detto scalare nel caso in cui $d = 1$.

Il modello $\mathcal{F}$ è detto correttamente specificato se il parametro è identificabile e se $p^0(y) \in \mathcal{F}$, dove $p^0(y)$ è la vera e ignota densità di Y, variabile aleatoria di cui i dati y sono realizzazione. In tal caso, il valore θ^0 tale che $p(y;\theta^0) = p^0(y)$ si dice vero valore del parametro.

La funzione di verosimiglianza

La funzione di verosimiglianza di θ basata sui dati y è definita da:

$$L : \Theta \to \mathbb{R}^+ \qquad L(\theta) = c(y)p(y;\theta) \propto p(y;\theta).$$

Essa rappresenta la funzione del modello $p(y;\theta)$ vista come funzione di θ, per y fissato, definita a meno di costanti moltiplicative, riassumibili nel termine $c(y)$. Per mettere in evidenza la dipendenza di $L(\theta)$ dai dati si usa spesso la notazione $L(\theta;y)$, mentre per evidenziarne le proprietà distributive si scrive $L(\theta;Y)$. Spesso è più comodo trattare con la quantità:

$$l(\theta) = \log L(\theta), \tag{1.1}$$

detta funzione di log-verosimiglianza.

Se le n componenti di $(Y_1, Y_2, \ldots, Y_n)$ di Y sono indipendenti, con densità marginali $p_{Y_i}(y_i,\theta)$, $i = 1,\ldots,n$, allora la (1.1) diventa:

$$l(\theta) = \log \prod_{i=1}^{n} p_{Y_i}(y_i,\theta) = \sum_{i=1}^{n} \log p_{Y_i}(y_i,\theta).$$

La verosimiglianza per un valore del parametro θ', $L(\theta')$, può essere interpretata come il sostegno empirico che θ', come indice del modello generatore dei dati, riceve dall'osservazione di y.

La funzione di verosimiglianza gode delle seguenti proprietà di invarianza:

- La funzione $L(\theta)$ è invariante rispetto a trasformazioni biunivoche di y. Ciò significa che se invece di osservare y si è osservato $r = r(y)$, dove $r(\cdot)$ è trasformazione biunivoca con inversa $y = y(r)$, allora $L(\theta;y) = L(\theta;y(r))$.

- La funzione $L(\theta)$ è invariante rispetto a riparametrizzazioni di $\mathcal{F}$. La riparametrizzazione è definita come trasformazione biunivoca $\omega(\theta)$ del parametro del modello, con inversa $\theta(\omega)$. In questo caso la proprietà di invarianza si traduce nel fatto che le conclusioni inferenziali ottenute sotto la parametrizzazione $\theta \in \Theta$ sono le stesse di quelle ottenute sotto la parametrizzazione $\omega \in \Omega$, cioè $L^{\Omega}(\omega) = L^{\Theta}(\theta(\omega))$.

1.2. Stima di massima verosimiglianza

La stima di massima verosimiglianza di θ , è il valore $\hat{\theta}$ tale che $L(\hat{\theta}) \geq L(\theta), \forall\theta \in \Theta$, cioè che massimizza $L(\theta)$ o equivalentemente $l(\theta)$. Se $\hat{\theta} = \hat{\theta}(y)$ esiste ed è unico, allora la variabile aleatoria $\hat{\theta} = \hat{\theta}(Y)$ è detta stimatore di massima verosimiglianza. Aspetto importante della stima di massima verosimiglianza è la sua equivarianza rispetto alla parametrizzazione, nel senso che nella nuova riparametrizzazione $\omega(\theta), \omega \in \Omega$, si ha $\hat{\omega} = \omega(\hat{\theta})$.

Funzione di punteggio

La funzione di punteggio, o funzione *score* è definita come:

$$l_\theta(\theta) = \frac{\partial}{\partial\theta} l(\theta; y) = \left(\frac{\partial}{\partial\theta_1} l(\theta), \ldots, \frac{\partial}{\partial\theta_d} l(\theta) \right)^\top.$$

In un modello con verosimiglianza regolare (Azzalini, 2001, p.88), la stima di massima verosimiglianza si ottiene spesso come unica soluzione dell'equazione di verosimiglianza:

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

che in generale può non ammettere soluzione esplicita, la quale in tal caso va ricercata numericamente, ad esempio tramite il metodo di *Newton-Raphson*. Vale inoltre il seguente risultato esatto, detto prima identità di Bartlett:

$$E_\theta(l_\theta(\theta)) = 0, \forall\theta \in \Theta.$$

Informazione osservata e informazione attesa

Si dice informazione osservata la matrice $d \times d$ delle derivate parziali seconde di $l(\theta)$ cambiate di segno:

$$j(\theta) = -l_{\theta\theta}(\theta) = -\frac{\partial^2}{\partial\theta\partial\theta^\top} l(\theta) \tag{1.2}$$

con elementi $j_{rs}(\theta)$,$r, s = 1, \ldots, d$. Essa rappresenta la curvatura della verosimiglianza e dà un'indicazione di quanto nettamente si possono distinguere regioni di Θ a elevata verosimiglianza da regioni con verosimiglianza modesta.

Si dice informazione attesa o informazione di Fisher la quantità:

$$i(\theta) = E_\theta(j(\theta)), \tag{1.3}$$

con elementi $i_{rs}(\theta)$,$r, s = 1, \ldots, d$. È il valore atteso dell'informazione osservata (1.2), ed è una quantità importante per esprimere la distribuzione asintotica di $\hat{\theta}$. Vale il seguente risultato esatto, detto seconda identità di Bartlett:

$$Var_\theta(l_\theta(\theta)) = i(\theta), \forall\theta.$$

1.3. Test e regioni di confidenza

Risultati distributivi

Sotto determinate condizioni di regolarità (Pace e Salvan, 2001, Capitolo 6), valgono alcuni risultati distributivi asintotici del prim'ordine, ottenuti assumendo che $n \to \infty$, sia per lo stimatore $\hat{\theta}$, sia per altre importanti quantità, posto che θ sia il vero valore del parametro.

Vale il seguente risultato di convergenza in probabilità:

$$\hat{\theta} \xrightarrow{p} \theta,$$

che definisce la consistenza di $\hat{\theta}$, e i risultati distributivi asintotici:

$$l_\theta(\theta) \dot{\sim} \mathcal{N}_d(0, i(\theta)) \tag{1.4}$$

$$(\hat{\theta} - \theta) \dot{\sim} \mathcal{N}_d(0, i(\theta)^{-1}), \tag{1.5}$$

dove $i(\theta)^{-1}$ può essere sostituito da $i(\hat{\theta})^{-1}$ o anche $j(\hat{\theta})^{-1}$, suoi stimatori consistenti.

Si definiscono inoltre:

$$W_e(\theta) = (\hat{\theta} - \theta)^\top j(\hat{\theta})(\hat{\theta} - \theta), \tag{1.6}$$

$$W_u(\theta) = l_\theta(\theta)^\top i(\theta)^{-1} l_\theta(\theta), \tag{1.7}$$

$$W(\theta) = 2\left\{l(\hat{\theta}) - l(\theta)\right\}. \tag{1.8}$$

Le quantità $W_e(\theta)$,$W_u(\theta)$ e $W(\theta)$ hanno distribuzione asintotica χ^2_d e identificano tre quantità (approssimativamente) pivotali basate sulla verosimiglianza, chiamate rispettivamente statistica di *Wald* , *score* e del **rapporto di verosimiglianza**[1].

Nel caso particolare in cui θ sia un parametro scalare, si può definire anche la seguente quantità, derivata dalla (1.8):

$$r(\theta) = \text{sgn}(\hat{\theta} - \theta)\sqrt{W(\theta)}, \tag{1.9}$$

[1] Da qui in poi con rapporto di verosimiglianza si intende ciò che più precisamente è il log-rapporto di verosimiglianza.

che prende il nome di **radice con segno del rapporto di verosimiglianza** e ha distribuzione asintotica $\mathcal{N}(0,1)$.

Esistono anche le corrispettive versioni di $r(\theta)$, asintoticamente equivalenti, derivate dalle (1.6)-(1.7):

$$r_e(\theta) = \text{sgn}(\hat{\theta} - \theta)\sqrt{j(\hat{\theta})}, \tag{1.10}$$

$$r_u(\theta) = \text{sgn}(\hat{\theta} - \theta)\sqrt{i(\theta)^{-1}}, \tag{1.11}$$

entrambe con distribuzione asintotica $\mathcal{N}(0,1)$.

Test di verosimiglianza

Sia definito un sistema per la verifica di un'ipotesi semplice su θ:

$$H_0 : \theta = \theta_0,$$
$$H_1 : \theta \neq \theta_0$$

La statistica test naturale, basata sulla verosimiglianza, è $L(\hat{\theta})/L(\theta_0)$, con regione di rifiuto unilaterale destra. Tale statistica è funzione monotona di $W(\theta_0)$. Valori grandi di $W(\theta_0)$ sono critici per l'ipotesi nulla. Dai risultati distributivi per $W(\theta)$, si deduce:

$$W(\theta_0) \underset{H_0}{\dot{\sim}} \chi^2_d,$$

e analogamente vale per $W_e(\theta_0)$ e $W_u(\theta_0)$.

Il valore di significatività osservato associato al valore campionario delle statistica $W(\theta_0)$, cioè $W(\theta_0)^{oss}$, è:

$$\alpha^{oss} = Pr_{\theta_0}\left(W(\theta_0) \geq W(\theta_0)^{oss}\right) \doteq Pr\left(\mathscr{X} \geq W(\theta_0)^{oss}\right),$$

dove $\mathscr{X} \sim \chi^2_d$.

Dal risultato distributivo per $r(\theta)$, si deduce che:

$$r(\theta_0) \underset{H_0}{\dot{\sim}} \mathcal{N}(0,1),$$

e quindi $r(\theta_0)$ può essere utilizzato anche per la verifica di H_0 contro alternative unilaterali del tipo $H_1^{sx} : \theta < \theta_0$ e $H_1^{dx} : \theta > \theta_0$. Analogamente vale per $r_e(\theta)$ e $r_u(\theta)$. Effettuare il test $r(\theta_0)$ contro alternativa bilaterale equivale a utilizzare $W(\theta_0)$.

Il valore di significatività osservato associato al valore campionario delle statistica $r(\theta_0)$, cioè $r(\theta_0)^{oss}$, ad esempio in caso di alternativa unilaterale sinistra, è:

$$\alpha^{oss} = Pr_{\theta_0}\left(r(\theta_0) \leq r(\theta_0)^{oss}\right) \doteq Pr\left(\mathscr{Z} \leq r(\theta_0)^{oss}\right),$$

dove $\mathscr{Z} \sim \mathcal{N}(0,1)$.

Regioni di confidenza

Regioni di confidenza per il parametro θ possono essere ricavate dal fatto che le quantità (1.6)-(1.11) sono asintoticamente pivotali. Con una notazione sintetica e del tutto generale si possono scrivere le regioni di confidenza bilaterali con livello approssimato $(1-\alpha)$:

$$\hat{\Theta}(y) = \left\{\theta \in \Theta : \mathscr{W}(\theta) < \chi^2_{d;1-\alpha}\right\}, \tag{1.12}$$

dove con $\mathscr{W}(\theta)$ viene indicata una delle seguenti quantità: $W(\theta), W_r(\theta), W_u(\theta)$ oppure $r(\theta)^2, r(\theta)^2, r_u(\theta)^2$, nel caso in cui $d = 1$, e dove $\chi^2_{d;1-\alpha}$ rappresenta il quantile $(1-\alpha)$ della distribuzione χ^2_d.

In generale è da preferire l'utilizzo delle quantità basate sulla verosimiglianza $W(\theta)$ e $r(\theta)$, in quanto:

- sono invarianti rispetto alla parametrizzazione e quindi le corrispondenti regioni di confidenza godono della proprietà di equivarianza;
- forniscono intervalli di confidenza che rispettano sempre lo spazio parametrico ($\hat{\Theta}(y) \subset \Theta$), e che per costruzione includono tutti i punti con valore di verosimiglianza maggiore di una certa soglia, cioè $L(\theta') \geq L(\theta'')\ \forall \theta' \in \hat{\Theta}(y), \theta'' \notin \hat{\Theta}(y)$.

Tramite tali quantità si ottiene, nel caso in cui θ sia scalare:

$$\hat{\Theta}(y) = \left\{\theta \in \Theta : W(\theta) < \chi^2_{1;1-\alpha}\right\} = \left\{\theta \in \Theta : -z_{1-\alpha/2} < r(\theta) < z_{1-\alpha/2}\right\}, \tag{1.13}$$

dove $z_{1-\alpha/2}$ rappresenta il quantile $(1-\alpha/2)$ della distribuzione $\mathcal{N}(0,1)$.

1.4. Verosimiglianza profilo e criterio di Akaike

Verosimiglianza profilo

Nella maggior parte delle situazioni realistiche, i modelli in esame includono diversi parametri. Si suppone che il parametro complessivo θ sia espresso da due blocchi di componenti: $\theta = (\tau, \zeta)$. L'interesse è posto su τ , sul quale si vuole fare inferenza, ed è detto perciò **parametro di interesse**, mentre ζ rappresenta una componente residuale, di cui comunque bisogna tenere conto ma che assume un ruolo secondario ed è detto **parametro di disturbo**. I parametri τ e ζ sono rispettivamente q-dimensionale e $(d-q)$-dimensionale.

Un approccio abbastanza generale basato sulla verosimiglianza per trattare la natura residuale del parametro ζ è di sostituirlo con la stima vincolata di massima verosimiglianza $\hat{\zeta}_\tau$, ottenuta massimizzando $L(\tau, \zeta)$ rispetto a ζ per τ fissato, per poi riassumere la conoscenza su τ tramite la verosimiglianza profilo:

$$L_p(\tau) = \max_{\zeta} L(\tau, \zeta) = L(\tau, \hat{\zeta}_\tau),$$

o equivalentemente tramite la log-verosimiglianza profilo $l_p(\tau) = \log L_p(\tau)$. Si usa anche la notazione $\hat{\theta}_\tau = (\tau, \hat{\zeta}_\tau)$.

Le quantità $\hat{\theta}, l_\theta(\theta), i(\theta)$ e $j(\theta)$ del §1.2 possono essere riscritte considerando i blocchi corrispondenti alla struttura di θ: $\hat{\theta} = (\hat{\tau}, \hat{\zeta})$,

$$l_\theta(\theta)^\top = \left(l_\tau(\tau, \zeta)^\top, l_\zeta(\tau, \zeta)^\top\right),$$

dove $l_\tau(\tau, \zeta) = \partial l(\tau, \zeta)/\partial\tau$ e $l_\zeta(\tau, \zeta) = \partial l(\tau, \zeta)/\partial\zeta$. Le matrici di informazione (1.2) e (1.3) possono essere riscritte come:

$$i(\theta) = \begin{pmatrix} i_{\tau\tau} & i_{\tau\zeta} \\ i_{\zeta\tau} & i_{\zeta\zeta} \end{pmatrix}, j(\theta) = \begin{pmatrix} j_{\tau\tau} & j_{\tau\zeta} \\ j_{\zeta\tau} & j_{\zeta\zeta} \end{pmatrix},$$

e parimenti le loro inverse:

$$i(\theta)^{-1} = \begin{pmatrix} i^{\tau\tau} & i^{\tau\zeta} \\ i^{\zeta\tau} & i^{\zeta\zeta} \end{pmatrix}, j(\theta)^{-1} = \begin{pmatrix} j^{\tau\tau} & j^{\tau\zeta} \\ j^{\zeta\tau} & j^{\zeta\zeta} \end{pmatrix}. \tag{1.14}$$

L'informazione osservata per la log-verosimiglianza profilo può essere espressa in termini della verosimiglianza totale in base all'uguaglianza (Brazzale *e altri*, 2007, p.11):

$$j_p(\tau) = -\frac{\partial^2}{\partial\tau\partial\tau^\top} l_p(\tau) = \left\{j^{\tau\tau}(\hat{\theta}_\tau)\right\}^{-1}.$$

In condizioni di regolarità, similmente a quanto visto nel §1.2, la stima $\hat{\zeta}_\tau$ è soluzione in ζ dell'equazione di verosimiglianza parziale del sotto-modello con τ fissato , $l_\tau(\tau, \zeta) = 0$. La stima di massima verosimiglianza profilo, $\hat{\tau}$, coincide con la stima di massima verosimiglianza di τ.

Risultati distributivi del prim'ordine simili a quelli visti nel § 1.3 possono essere ottenuti considerando $l_p(\tau)$ come una verosimiglianza ordinaria, anche se a rigore non sarebbe il logaritmo di una funzione di densità di probabilità.

Si possono quindi costruire per la verosimiglianza profilo i corrispettivi delle quantità di verosimiglianza (1.4)-(1.8), con i corrispondenti risultati distributivi (si omettono per brevità le quantità *score*, che non verranno più utilizzate):

$$(\hat{\tau} - \tau) \dot{\sim} \mathcal{N}_q(0, j^{\tau\tau}(\hat{\theta}))$$

$$W_{ep}(\tau) = (\hat{\tau} - \tau)^\top (j^{\tau\tau}(\hat{\theta}))^{-1} (\hat{\tau} - \tau) \dot{\sim} \chi^2_q$$

$$W_p(\tau) = 2 \left\{ l(\hat{\theta}) - l(\hat{\theta}_\tau) \right\} \dot{\sim} \chi^2_q.$$

La quantità $W_p(\tau)$ è detta **rapporto di verosimiglianza profilo**.

Nel caso in cui τ sia un parametro scalare, si hanno anche:

$$r_{ep}(\tau) = \text{sgn}(\hat{\tau} - \tau) \sqrt{j_p(\hat{\tau})} \dot{\sim} \mathcal{N}(0, 1),$$

$$r_p(\tau) = \text{sgn}(\hat{\tau} - \tau) \sqrt{W_p(\tau)} \dot{\sim} \mathcal{N}(0, 1). \tag{1.15}$$

La quantità $r_p(\tau)$ è detta **radice con segno del rapporto di verosimiglianza profilo** o anche *directed deviance*.

Test di verosimiglianza e regioni di confidenza per τ si ottengono analogamente a quanto visto nel §1.3, tramite le quantità approssimativamente pivotali appena definite. Fissare τ è di particolare interesse, perché equivale a definire un sotto-modello $\mathcal{F}_0$ di $\mathcal{F}$. Un sotto-modello è definito infatti fissando dei vincoli su θ, che identificano una regione più ristretta dello spazio parametrico: $\Theta_0 \subseteq \Theta$. Si dice che $\mathcal{F}_0$ è annidato in $\mathcal{F}$, e anche che $\mathcal{F}_0$ è un modello ridotto rispetto al modello corrente $\mathcal{F}$.

In questo contesto più generale, considerando un'ipotesi composita $H_0 : \theta \in \Theta_0$, dove $\Theta_0 \subseteq \Theta$, la statistica test del rapporto di verosimiglianza è:

$$W_p^{H_0} = 2 \left\{ \sup_{\theta \in \Theta} l(\theta) - \sup_{\theta \in \Theta_0} l(\theta) \right\}, \tag{1.16}$$

la quale non richiede una specifica partizione di θ in parametro di interesse e parametro di disturbo. La distribuzione asintotica è χ^2_q, dove $q = \dim(\Theta) - \dim(\Theta_0)$ e valori grandi della statistica sono critici per l'ipotesi nulla.

Criterio di informazione di Akaike

Per la selezione di un modello statistico finale per i dati tra diversi modelli alternativi, anche non annidati, spesso è utile introdurre dei criteri basati su verosimiglianze penalizzate per il numero di parametri. Uno di questi è il criterio di Akaike (*AIC: Akaike Information Criterion*, si veda Akaike, 1973), definito da:

$$AIC(\mathcal{F}_{d^\star}) = 2p^\star - 2l\left(\hat{\theta}^{(d^\star)}; y\right), \tag{1.17}$$

Dove con $\mathcal{F}_{d^\star}$ si è indicato un generico modello di parametro $\theta^{(d^\star)}$, $p^\star$-dimensionale. Anche se tale criterio tende a selezionare modelli leggermente sovraparametrizzati (Pace e Salvan, 2001, p.182), per l'uso che ne verrà fatto esso offre un buon compromesso tra adattamento e parsimonia.

2. Modelli per dati di conteggio in presenza di sovradispersione

Spesso, all'interno del contesto generale dei modelli statistici parametrici, definiti nel Capitolo 1, è di interesse studiare la relazione tra una variabile risposta e altre variabili, dette esplicative, definendo un modello di regressione.

Il contesto basilare dal quale muove la discussione di questo capitolo è quello dei modelli di regressione lineari generalizzati, introdotti da Nelder e Wedderburn (1972), che costituiscono un'estensione del modello di regressione lineare normale multiplo.

Viene definito, come caso particolare, il modello di Poisson, adatto specificatamente a trattare insieme di dati nei quali la variabile risposta rappresenta il risultato di un conteggio. Semplici esempi sono il numero di visite mediche in un anno per ciascuno degli n studenti di una determinata classe, il numero di pesci catturati in una giornata da n pescatori partecipanti ad una gara sportiva, oppure anche il numero di visitatori per ciascuno degli n musei archeologici nazionali italiani in un determinato anno.

Il modello di Poisson assume che la varianza della risposta sia uguale alla sua media, ma nella pratica questa assunzione è spesso violata. Quando la varianza è superiore alla media si parla di sovradispersione rispetto al modello.

Esistono varie soluzioni per tenere in considerazione o per studiare la sovradispersione e una di queste è il modello binomiale negativo, il quale verrà analizzato in dettaglio.

2.1. Modelli lineari generalizzati

L'inferenza nel contesto dei modelli lineari generalizzati è basata sulla verosimiglianza e di norma sono valide le condizioni di regolarità che permettono la stima consistente dei parametri e l'utilizzo di risultati asintotici riconducibili a quelli visti nel §1.3 per l'ottenimento di regioni di confidenza o per la verifica di sistemi d'ipotesi. Per una trattazione completa dei modelli lineari generalizzati si veda la

monografia McCullagh e Nelder (1989), o anche Agresti (2015). Di seguito vengono presentate solo alcune principali caratteristiche di questa classe di modelli di regressione.

Definizione

Si indichi con $y = (y_1, \ldots, y_n)$ l'osservazione della variabile risposta su tutte le n unità statistiche. Ciò significa assumere che y_i, $i = 1, \ldots, n$, sia realizzazione di una variabile aleatoria Y_i, assunta univariata, componente relativa alla i-esima unità della variabile risposta $Y = (Y_1, \ldots, Y_n)$. Su ciascuna delle n unità, si rilevano inoltre i valori noti di p variabili dette esplicative o covariate, indicate con il vettore riga $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$, $i = 1, \ldots, n$.

Un modello lineare generalizzato viene definito dalle seguenti tre componenti:

Componente aleatoria: le variabili $Y_1, \ldots, Y_n$ sono variabili aleatorie indipendenti con media $E(Y_i) = \mu_i$, varianza $Var(Y_i) = a_i(\phi)v(\mu_i)$ per opportune funzioni a_i e v, e con distribuzione appartenente alla famiglia di dispersione esponenziale univariata (si veda l'Appendice A.1):

$$Y_i \sim DE_1\left(\mu_i, a_i(\phi)v(\mu_i)\right), i = 1, \ldots, n. \tag{2.1}$$

Predittore lineare: è definito come $\eta = X\beta$, dove $\beta = (\beta_1, \ldots, \beta_p)^\top$ è un parametro vettoriale di dimensione p e $X = \left(\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \ldots, \boldsymbol{x}_n^\top\right)^\top$ è una matrice $n \times p$, detta matrice di disegno, matrice del modello o anche matrice delle covariate. Il predittore è a sua volta un vettore $n \times 1$: $\eta = (\eta_1, \ldots, \eta_n)^\top = (\boldsymbol{x}_1\beta, \ldots, \boldsymbol{x}_n\beta)^\top$. Solitamente $\boldsymbol{x}_1 = (1, 1, \ldots, 1)$ e quindi β_1 rappresenta un parametro di intercetta.

Funzione di legame: è definita come la funzione nota e invertibile $g(\cdot)$ che collega la media μ_i al predittore η_i; $g(\mu_i) = \eta_i = \boldsymbol{x}_i\beta = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. La funzione di legame canonica o naturale è quella che pone il predittore lineare uguale al parametro naturale della distribuzione a dispersione esponenziale di Y_i. L'utilizzo della funzione legame canonica conduce spesso a semplificazioni analitiche della funzione di verosimiglianza.

Stima e parametro di dispersione

Le equazioni di verosimiglianza per un modello di regressione lineare generalizzato non ammettono quasi mai soluzione esplicita, quindi è necessario l'utilizzo di uno

specifico algoritmo numerico. Quest'ultimo è basato sul una particolare versione dell'algoritmo di *Newton-Raphson* ed è detto dei minimi quadrati pesati iterati (*IRLS*) per l'analogia esistente con il metodo dei minimi quadrati pesati utilizzato nel modello lineare normale. Le stime finali vengono ottenute tramite procedure di stima iterata, fino al raggiungimento della convergenza.

In seguito indicheremo con $\hat{\eta} = X\hat{\beta}$ il predittore lineare stimato e, posto $\mu = E(Y) = (\mu_1, \dots, \mu_n)^\top$, con $\hat{\mu} = (g^{-1}(\hat{\eta}_1), \dots, g^{-1}(\hat{\eta}_n))$ indicheremo il vettore dei valori predetti. Si porrà $\hat{y}{=}\hat{\mu}$.

Il parametro ϕ nella (2.1) è detto parametro di dispersione ed entra nella definizione della varianza della risposta; è un valore fissato nei casi in cui la variabile risposta sia discreta. Nel caso di risposta continua, alla stima di massima verosimiglianza $\hat{\phi}$ è solitamente preferita la stima, con correzione per i gradi di libertà, basata sul metodo dei momenti:

$$\tilde{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}. \tag{2.2}$$

Devianza, residui e scelta del modello

Per un modello lineare generalizzato si definisce la devianza residua:

$$D(y; \hat{\mu}) = 2\phi(l^\mu(y, \phi) - l^\mu(\hat{\mu}, \phi)), \tag{2.3}$$

dove l^μ indica la log-verosimiglianza del modello lineare generalizzato nella parametrizzazione (μ, ϕ). Il termine $l^\mu(y, \phi)$ rappresenta la log-verosimiglianza calcolata ponendo $\mu_i = y_i$, ossia adattando il modello di regressione saturo con $p = n$. Il termine $l^\mu(\hat{\mu}, \phi)$ rappresenta invece la log-verosimiglianza calcolata nella stima di massima verosimiglianza $\hat{\mu}$, ossia adattando il modello di regressione corrente con $p < n$ parametri. La (2.3) esprime una misura della diminuzione nella bontà di adattamento passando dal modello saturo al modello con p variabili esplicative: valori grandi esprimono un allontanamento dei valori predetti $\hat{\mu}_i$ dai valori osservati y_i. Può essere anche utilizzato, sotto certe condizioni, come test di bontà di adattamento.

Sia M lo spazio delle medie nel modello corrente, $\mu \in M$. Definita l'ipotesi composita $H_0 : \mu \in M_0$, con $M_0 \subseteq M$, il test di rapporto di verosimiglianza (1.16) si può riscrivere in termini della devianza come :

$$W_P = \frac{1}{\phi}(D(y; \hat{\mu}_{H_0}) - D(y; \hat{\mu})) \underset{H_0}{\dot{\sim}} \chi^2_{p-p_0}, n \to \infty, \tag{2.4}$$

dove ϕ, se ignoto e non fissato, va sostituito da una sua stima consistente, come $\tilde{\phi}$ ottenuta dalla (2.2).

Con modello nullo si intende il modello che assume $Y_1, \ldots, Y_n$ identicamente distribuite con media costante $E(Y_i) = \mu_i = g^{-1}(\beta_1), i = 1, \ldots, n$.

La scelta del modello finale tra diversi modelli si può basare sul test di devianza (2.4), se questi sono annidati, oppure anche su altri criteri come il criterio di Akaike (1.17), seguendo procedure di selezione all'indietro, in avanti o di tipo *stepwise*.

Definiamo infine due tipi principali di residui, rispettivamente di Pearson e di devianza:

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)}}, i = 1, \ldots, n$$

$$r_i^D = \text{sgn}(y_i - \hat{\mu}_i)\sqrt{D_i}, i = 1, \ldots, n,$$

dove D_i esprime il contributo di una singola osservazione alla devianza (2.3).

L'analisi grafica dei residui, insieme alla devianza ed ad altri tipi di strumenti, costituisce un metodo privilegiato per il controllo della corretta specificazione del modello.

2.2. Modello di Poisson

Si supponga ora che la variabile risposta rappresenti dei conteggi. Il modello di Poisson assume che $Y_1, \ldots, Y_n$ siano indipendenti con distribuzione di Poisson di media $E(Y_i) = \mu_i$: $Y_i \sim \text{Pois}(\mu_i) \equiv DE_1(\mu_i, \mu_i)$ (si veda l'Appendice A.1), e funzione di varianza:

$$Var(Y_i) = \mu_i. \tag{2.5}$$

Tale specificazione ricade in quella più generale dei modelli lineari generalizzati, con $\phi = 1$; si assume d'ora in poi la funzione di legame canonica: $g(\mu_i) = \log(\mu_i) = \eta_i$. Il modello così definito è detto modello di Poisson log-lineare, e tutte le procedure di stima e di inferenza applicabili nel contesto dei modelli lineari generalizzati rimangono valide. Per una trattazione monografica si veda ad esempio Cameron e Trivedi (2013).

La devianza (2.3) assume nel modello di Poisson la forma:

$$D(y;\hat{\mu}) = 2\sum_{i=1}^{n}\left(y_i \log(y_i/\hat{\mu}_i) - y_i + \hat{\mu}_i\right),$$

e in presenza di intercetta vale:

$$D(y;\hat{\mu}) = 2\sum_{i=1}^{n}\left(y_i \log(y_i/\hat{\mu}_i)\right) \doteq \sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i} = \sum_{i=1}^{n}\left(r_i^P\right)^2, \tag{2.6}$$

dove l'approssimazione è valida per n fissato se le medie μ_i divergono. Il membro a destra della (2.6) è detto statistica di Pearson e viene indicato con T^2. Si può dimostrare, sulla base del teorema del limite centrale, che $D(y;\hat{\mu})$ e dunque anche T^2 hanno distribuzione approssimata χ^2_{n-p} sotto l'ipotesi nulla del modello corrente, sempre per n fissato al divergere delle medie (Collings e Margolin, 1985). È naturale quindi considerare la statistica T^2 come test per valutare la bontà di adattamento del modello.

A tale scopo può essere utile anche l'analisi dei residui, che non devono presentare particolari sistematicità, oppure il confronto delle frequenze empiriche dei valori nel supporto della variabile risposta $\{0, 1, ...\}$ con le corrispondenti frequenze stimate tramite il modello: $\sum_{i=1}^{n} Pr(Y_i = h)/n$, $h = 0, 1, \ldots$.

2.3. Sovradispersione

Nell'utilizzo del modello di regressione di Poisson, nel quale il parametro di dispersione è fissato ($\phi = 1$), in ipotesi di corretta specificazione, ci si aspetterebbe di avere la seguente relazione tra devianza residua e gradi residui di libertà:

$$D(y;\hat{\mu}) \doteq n - p,$$

per quanto visto nel §2.2.

Come interpretare la situazione in cui $D(y;\hat{\mu}) \gg n-p$? Ci sono due principali insiemi di motivazioni, non mutuamente esclusive, da considerare:

- Si è ottenuto un modello con adattamento insoddisfacente, per esempio perché sono state omesse delle variabili esplicative o si sono inserite senza opportuna trasformazione, oppure perché la specificazione della funzione di legame è scorretta. Un'ulteriore ragione può essere la presenza di osservazioni anomale (*outliers*). Per approfondimenti si rimanda a McCullagh e Nelder (1989, Capitolo 12).

- La specificazione (2.5) per la varianza è troppo restrittiva o comunque scorretta, e quindi il modello attuale non è abbastanza flessibile. Può verificarsi la situazione in cui la varianza osservata sia minore di quella prevista dal modello, ma questa evenienza non è poi così frequente e non sarà ulteriormente discussa. Spesso invece la varianza osservata è maggiore di quella prevista e si parla in tal caso di **sovradispersione**. Si avrà dunque:

$$Var(Y_i) > \mu_i.$$

Per una discussione introduttiva alla sovradispersione nei modelli log-lineari si veda Breslow (1984).

Per effettuare un test sulla presenza di sovradispersione prima ancora di specificare dei modelli più avanzati, si può utilizzare la statistica di Pearson T^2, calcolata per il modello nullo:

$$T^2_{null} = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{\bar{y}},$$

con distribuzione nulla approssimata χ^2_{n-1}. L'indice di dispersione di Fisher è definito come $T^2_{null}/(n-1)$, che è il rapporto tra la varianza campionaria corretta e la media campionaria.

Origine e conseguenze

Ci sono diverse possibili cause per la sovradispersione, come ad esempio la presenza di correlazione tra le risposte individuali $Y_1, \ldots, Y_n$, contrariamente a quanto previsto con l'ipotesi di indipendenza. La sovradispersione può essere imputata anche alle peculiarità individuali delle unità statistiche, che provocano un aumento di variabilità in seguito alla necessaria casualizzazione sperimentale.

Altre ragioni possono essere l'utilizzo di tecniche particolari di campionamento, come ad esempio il campionamento a grappolo, oppure l'omissione di variabili esplicative rilevanti dal modello.

Se infine il processo stocastico sottostante alla osservazione dei dati di conteggio y è un processo composto, ad esempio un processo di Poisson composto (definito in seguito), il modello di Poisson rappresenta una semplificazione eccessiva.

Dopo aver identificato le possibili ragioni della sovradispersione, resta da capire a quali conseguenze possa portare il fatto di ignorarne la presenza.

Come affermano Milanzi *e altri* (2012), ignorare la presenza della sovradispersione ha generalmente un effetto blando sulla stime dei parametri β, mentre porta

invece a sottostimare in maniera rilevante gli errori standard degli stimatori $\hat{\beta}$ e quindi alla scorretta identificazione della significatività dei termini di β, che in molti casi è di primo interesse.

Se è vero che è importante non ignorare la sovradispersione, meno grave è adattare un modello che ne tiene conto quando in realtà essa non è presente.

2.4. Modelli per la sovradispersione

Seguendo l'esposizione di Hinde e Demétrio (1998), si proporranno un insieme di modelli di regressione adatti ad estendere il modello di Poisson nel caso sia presente sovradispersione.

Le diverse specificazioni di tali modelli discendono dai vari possibili meccanismi che caratterizzano il processo generatore dei dati. Si possono seguire principalmente due approcci, tramite specificazione della funzione di varianza (modelli media-varianza) oppure tramite costruzione di modelli gerarchici.

Modelli media-varianza

Questo tipo di modelli sostituiscono la relazione tra media e varianza (2.5) del modello originale con una forma più generale, che tipicamente coinvolge parametri addizionali. Modelli di questo tipo non garantiscono l'assegnazione di una distribuzione di probabilità valida per la risposta, ma possono comunque essere utilizzati grazie al metodo di quasi-verosimiglianza.

Si può assegnare al modello una funzione di varianza a dispersione costante, nel qual caso:

$$Var(Y_i) = \phi\mu_i. \tag{2.7}$$

Una funzione di varianza con questa forma può originarsi ad esempio quando il processo generatore dei dati è un processo di Poisson composto: si supponga che $N \sim \text{Pois}(\mu_N)$ e che $S = \sum_{i=1}^{N} Z_i$, dove le Z_i sono variabili discrete indipendenti e identicamente distribuite con valore atteso μ_Z e varianza σ_Z^2, allora:

$$\begin{aligned} E(S) = \mu_S &= E_N\left(E\left(S \mid N\right)\right) = E_N\left(N\mu_Z\right) = \mu_N\mu_Z \\ Var(S) &= E_N\left(Var\left(S \mid N\right)\right) + Var_N\left(E\left(S \mid N\right)\right) = \mu_S\left(\frac{\sigma_Z^2+\mu_Z^2}{\mu_Z}\right) \end{aligned} \tag{2.8}$$

cioè $Var(S) = \phi\mu_S$, dove ϕ dipende dai primi due momenti della variabile Z; la sovradispersione è presente se $E(Z^2) > E(Z)$.

Ci sono comunque altri processi che conducono ad una funzione di varianza a dispersione costante (2.7): per una interessante rassegna si può consultare Lindén e Mäntyniemi (2011), che presenta anche un'applicazione allo studio delle migrazioni di varie specie di uccelli.

Si può assumere una funzione di varianza ancora più generale:

$$Var(Y_i) = \mu_i \left(1 + \phi \mu_i^{\delta}\right), \tag{2.9}$$

dalla quale si ottiene la specificazione del modello Poisson originale (2.5) ponendo $\phi = 0$. Ponendo $\delta = 0$ si ritorna alla forma a dispersione costante (2.7) in una diversa parametrizzazione, mentre con $\delta = 1$ si ottiene la funzione di varianza del modello binomiale negativo, presentato nel §2.5.

Modelli gerarchici

I modelli gerarchici si ottengono assumendo che i parametri che caratterizzano la distribuzione della risposta siano a loro volta delle variabili casuali. Ciò conduce a delle distribuzioni di probabilità composta che talvolta non hanno forma esplicita e richiedono metodi di approssimazione numerica.

Oltre al modello binomiale negativo, un altro esempio è il modello Poisson-Normale a effetti misti, definito come:

$$Y_i \sim \text{Pois}(\lambda_i) \quad , \quad \log(\lambda_i) = \boldsymbol{x}_i\beta + Z_i \ ,$$

dove $Z_i \sim \mathcal{N}(0, \sigma^2)$, che implica, dopo derivazioni analoghe alle (2.8):

$$\begin{aligned} E(Y_i) &= e^{\boldsymbol{x}_i\beta + \frac{1}{2}\sigma^2} = \mu_i \\ Var(Y_i) &= \mu_i \left\{1 + \mu_i(e^{\sigma^2} - 1)\right\}. \end{aligned}$$

Si noti che la funzione di varianza ricade nella specificazione (2.9).

2.5. Modello binomiale negativo

Il modello binomiale negativo, sul quale si soffermerà l'attenzione, è un modello di regressione parametrico per dati di conteggio in presenza di sovradispersione, derivato da una definizione gerarchica che estende il modello di Poisson. Una monografia è Hilbe (2011).

Specificazione del modello

Sia $y = (y_1, \ldots, y_n)$ l'osservazione della variabile $Y = (Y_1, \ldots, Y_n)$. Si assume che le variabili Y_i siano indipendenti con distribuzione $\text{Pois}(\mu_i\lambda_i), i = 1, \ldots, n$. Siano i parametri λ_i a loro volta realizzazioni indipendenti di una variabile aleatoria $\Lambda \sim Ga(\kappa, \kappa)$; allora la densità di Y_i si può scrivere come :

$$p_{Y_i}(y_i; \mu_i, \kappa) = \int_0^{+\infty} p_{Y_i|\Lambda}(y_i; \mu_i, \kappa) p_\Lambda(\lambda_i; \mu_i, \kappa) d\lambda_i,$$

e quindi:

$$p_{Y_i}(y_i; \mu_i, \kappa) = \int_0^{+\infty} \frac{(\mu_i\lambda_i)^{y_i} e^{-\mu_i\lambda_i}}{y_i!} \kappa^\kappa \frac{\lambda_i^{\kappa-1} e^{-\kappa\lambda_i}}{\Gamma(\kappa)} d\lambda_i,$$

$$p_{Y_i}(y_i; \mu_i, \kappa) = \frac{\Gamma(y_i + \kappa)}{\Gamma(y_i + 1)\Gamma(\kappa)} \left(\frac{\mu_i/\kappa}{1 + \mu_i/\kappa}\right)^{y_i} \left(\frac{1}{1 + \mu_i/\kappa}\right)^{\kappa}, \tag{2.10}$$

per $y_i \in \mathbb{N}, \mu_i > 0, \kappa \geq 0, i = 1, \ldots, n$.

La distribuzione marginale per Y_i è quindi binomiale negativa con parametro di scala κ e parametro di probabilità $\pi_i = 1/(1 + \mu_i/\kappa)$:

$$Y_i \sim \text{Bineg}(\kappa, \pi_i), i = 1, \ldots, n.$$

Per le caratteristiche di tale distribuzione, oppure tramite derivazioni analoghe alle (2.8), si trova $E(Y_i) = \mu_i$ e:

$$Var(Y_i) = \mu_i\left(1 + \frac{\mu_i}{\kappa}\right). \tag{2.11}$$

Poiché (κ, π_i) è una riparametrizzazione di (κ, μ_i), si scriverà in seguito $Y_i \sim \text{Bineg}(\kappa, \mu_i)$, lasciando intendere che μ_i è la media, non il parametro di probabilità.

In presenza di p variabili esplicative, utilizzando la stessa notazione vista in precedenza, il modello di regressione binomiale negativo è allora definito semplicemente assumendo che le variabili risposta $Y_1, \ldots, Y_n$ siano indipendenti e abbiano distribuzione binomiale negativa con parametro di scala comune κ e parametro di media $\mu_i = g^{-1}(\boldsymbol{x}_i\beta) = g^{-1}(\eta_i), i = 1, \ldots, n$. Si utilizza di norma la funzione di legame logaritmica $g(\cdot) = \log(\cdot)$.

Si può inoltre dimostrare che, soltanto per κ fissato, la (2.10) appartiene alla famiglia di dispersione esponenziale e dunque il modello di regressione è un modello di regressione lineare generalizzato. Per tale dimostrazione e altri dettagli sulla distribuzione binomiale negativa si rinvia all'Appendice A.2.

Quantità di verosimiglianza

Considerata la funzione di varianza (2.11), il parametro scalare κ può essere anche chiamato, a ragione, **parametro di dispersione** per il modello binomiale negativo, analogamente a quanto avviene per ϕ nei modelli lineari generalizzati. Di norma κ non è fissato e quindi, non disponendo di risultati generali, diventa necessario ricavare tutte le quantità utili per l'inferenza.

Nel presente contesto il parametro del modello è (κ, μ), con $\mu = (\mu_1, \ldots, \mu_n)^\top$. Poiché si adatta un modello di regressione vi sono dei vincoli su μ e dunque $\theta = (\kappa, \beta)$, con $\beta = (\beta_1, \ldots, \beta_p)^\top$. Per fornire una corrispondenza con quanto visto nel §1.4, si consideri $\theta = (\tau, \zeta) = (\kappa, \beta)$.

Si noti che la densità (2.10) può essere riscritta come:

$$p(y_i; \mu_i, \kappa) = \frac{\Gamma(\kappa + y_i)}{\Gamma(y_i + 1)\Gamma(\kappa)} \frac{\mu_i^{y_i} \kappa^\kappa}{(\mu_i + \kappa)^{\kappa + y_i}},$$

da cui si ottiene la funzione di verosimiglianza per il modello binomiale negativo:

$$L(\theta; y) = L(\kappa, \mu; y) = \prod_{i=1}^{n} \left\{ \frac{\Gamma(y_i + \kappa)}{\Gamma(y_i + 1)\Gamma(\kappa)} \frac{\mu_i^{y_i} \kappa^\kappa}{(\mu_i + \kappa)^{\kappa + y_i}} \right\},$$

dove, come sempre, $\log(\mu_i) = \eta_i$.

La funzione di log-verosimiglianza è :

$$l(\theta; y) = \sum_{i=1}^{n} \left\{ y_i \log(\mu_i) + \kappa \log \kappa - (\kappa + y_i) \log(\kappa + \mu_i) + \mathcal{G}_i(\kappa) + 1 \right\},$$

dove $\mathcal{G}_i(\kappa) = \log \{\Gamma(y_i + \kappa)/\Gamma(y_i + 1)/\Gamma(\kappa)\}$.

La funzione di punteggio è data da :

$$l_\theta(\theta) = \left(\frac{\partial}{\partial \kappa} l(\theta), \frac{\partial}{\partial \beta} l(\theta)^\top \right)^\top = \left(\frac{\partial}{\partial \kappa} l(\theta), \frac{\partial}{\partial \beta_1} l(\theta), \ldots, \frac{\partial}{\partial \beta_p} l(\theta) \right)^\top,$$

con componenti:

$$l_\kappa(\theta) = \frac{\partial}{\partial \kappa} l(\theta) = \sum_{i=1}^{n} \left\{ \log \kappa - \frac{(\kappa + y_i)}{(\kappa + \mu_i)} - \log(\mu_i + \kappa) + 1 + \mathcal{D}_i(\kappa) \right\},$$

$$l_{\beta_r}(\theta) = \frac{\partial}{\partial \beta_r} l(\theta) = \frac{\partial l(\theta)}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_r} = \sum_{i=1}^{n} \left\{ \frac{y_i}{\mu_i} - \frac{(\kappa + y_i)}{(\kappa + \mu_i)} \right\} \mu_i x_{ir} \,, \; r = 1, \ldots, p,$$

dove $\mathcal{D}_i(\kappa) = \partial \mathcal{G}_i(\kappa)/\partial\kappa = \Psi(\kappa + y_i) - \Psi(\kappa)$, con $\Psi(x) = \partial \log \Gamma(x)/\partial x$ funzione digamma.

La matrici di informazione sono:

$$j(\theta) = \begin{pmatrix} j_{\kappa\kappa} & j_{\kappa\beta} \\ j_{\beta\kappa} & j_{\beta\beta} \end{pmatrix}, i(\theta) = \begin{pmatrix} i_{\kappa\kappa} & i_{\kappa\beta} \\ i_{\beta\kappa} & i_{\beta\beta} \end{pmatrix},$$

dove:

$$j_{\kappa\kappa} = -\frac{\partial^2}{\kappa^2} l(\theta) = -\sum_{i=1}^{n} \left\{ \frac{(y_i - \mu_i)}{(\kappa + \mu_i)^2} - \frac{1}{\mu_i + \kappa} + \frac{1}{\kappa} + \Psi'(\kappa + y_i) - \Psi'(\kappa) \right\};$$

$$j_{\kappa\beta} = j_{\beta\kappa}^{\top} = -\frac{\partial}{\partial\kappa} l_\beta(\theta)^{\top} = -\left(\frac{\partial}{\partial\kappa} l_{\beta_1}(\theta), \ldots, \frac{\partial}{\partial\kappa} l_{\beta_p}(\theta) \right), \text{con:}$$

$$\frac{\partial}{\partial\kappa} l_{\beta_r}(\theta) = \sum \left\{ \frac{(y_i - \mu_i)}{(\kappa + \mu_i)^2} \right\} \mu_i x_{ir}, r = 1, \ldots, p;$$

inoltre $j_{\beta\beta} = -l_{\beta\beta}(\theta)$, con elementi:

$$j_{\beta_r\beta_s} = -\frac{\partial}{\partial\beta_r \partial\beta_s} l(\theta) = \sum \left\{ \frac{(\kappa + y_i)}{(\kappa + \mu_i)^2} \right\} \kappa\mu_i x_{ir} x_{is}, r, s = 1, \ldots, p,$$

e infine:

$$i_{\kappa\kappa} = E_\theta(j_{\kappa\kappa}) \quad , \quad i_{\beta_r\beta_s} = E_\theta(i_{\beta_r\beta_s}) \quad = \sum_{i=1}^{n} \frac{\mu_i x_{ir} x_{is}}{(1 + \mu_i/\kappa)}, r, s = 1, \ldots, p,$$

$$i_{\kappa\beta} = E_\theta(j_{\kappa\beta}) = (0, \ldots, 0) = \underline{0}.$$

Le matrici $j^{-1}(\theta)$ e $i^{-1}(\theta)$ sono definite come nella (1.14), mentre con $\Psi'(x) = \partial\Psi(x)/\partial x$ si è indicata la funzione trigamma.

Si noti che l'uguaglianza $i_{\kappa\beta} = \underline{0}$ ha il significato di ortogonalità dei parametri β e κ e garantisce la semplificazione $i^{\kappa\kappa} = i_{\kappa\kappa}^{-1}, i^{\beta\beta} = i_{\beta\beta}^{-1}$ (Lawless, 1987).

Procedura di stima

Le equazioni di verosimiglianza del modello binomiale negativo non ammettono soluzioni in forma esplicita, che pertanto vanno ricercate tramite una opportuna tecnica numerica. Hinde e Demétrio (1998) propongono di ripetere sequenzialmente i seguenti passaggi fino al raggiungimento della convergenza:

- Per un fissato valore di κ (la stima corrente se disponibile) stimare β adattando il modello lineare generalizzato corrispondente, tramite *IRLS*. Si ricordi infatti che per κ fissato la distribuzione binomiale negativa appartiene alla famiglia di dispersione esponenziale.

- Fissando β al valore della stima corrente, stimare κ tramite il metodo di *Newton-Raphson*, fino a convergenza.

Per ricavare un valore iniziale di κ, $\kappa^{(0)}$, si può adattare il modello di Poisson, ottenendo i valori predetti $\tilde{\mu}_i$, e poi porre:

$$\kappa^{(0)} = \frac{\sum_{i=1}^{n} \tilde{\mu}_i(1 - h_i\tilde{\mu}_i)}{T^2 - (n-p)},$$

dove $h_i = Var(\boldsymbol{x}_i\tilde{\beta})|_{\mu_i=\tilde{\mu}_i} = Var(\tilde{\eta}_i)|_{\mu_i=\tilde{\mu}_i}$. Tale espressione è ricavata uguagliando la statistica di Pearson del modello di Poisson al suo valore atteso sotto l'ipotesi di validità del modello binomiale negativo (Breslow, 1984).

Inferenza di verosimiglianza

Per ottenere intervalli di confidenza per i parametri del modello binomiale negativo e per saggiare test di ipotesi sugli stessi si possono utilizzare i risultati generali ottenuti nei §1.3-1.4, a patto di assumere $\kappa > 0$ e altre tenui condizioni sulla matrice del modello X per assicurare che $i(\kappa, \beta)/n$ converga ad una quantità definita positiva al divergere di n (Lawless, 1987).

Si ottiene quindi la nota quantità pivotale di Wald, con relativa distribuzione asintotica, a partire dal risultato:

$$(\hat{\theta} - \theta) = \begin{pmatrix} \hat{\kappa} - \kappa \\ \hat{\beta} - \beta \end{pmatrix} \dot{\sim} \mathcal{N}_{p+1}\left(0, \begin{pmatrix} i^{\kappa\kappa}(\hat{\kappa}, \hat{\beta}) & \underline{0} \\ \underline{0} & i^{\beta\beta}(\hat{\kappa}, \hat{\beta}) \end{pmatrix}\right), \tag{2.12}$$

dove in luogo dei blocchi della quantità attesa $i(\hat{\kappa}, \hat{\beta})^{-1}$ si possono utilizzare quelli della quantità osservata $j(\hat{\kappa}, \hat{\beta})^{-1}$.

Una alternativa preferibile, soprattutto quando n non è molto elevato rispetto a p, è utilizzare la quantità basata sul rapporto di verosimiglianza:

$$W(\theta) = 2\left\{l(\hat{\kappa}, \hat{\beta}) - l(\kappa, \beta)\right\}.$$

Procedure di inferenza su singole componenti di interesse del parametro, ad esempio κ e β_r,$r = 1, \ldots, p$, si basano sulla (2.12) oppure sulla verosimiglianza profilo. Risulta spesso utile anche la statistica generale (1.16).

Solitamente, nel contesto della regressione, il parametro di interesse è il vettore dei coefficienti di regressione β. In questa sede, dove l'attenzione è concentrata prevalentemente sul problema della sovradispersione, il parametro di interesse è κ. Le condizioni di regolarità richiedono che il vero valore κ sia lontano dalla frontiera del suo spazio parametrico $K = [0, +\infty)$.

Si riportano per comodità le quantità pivotali utili all'inferenza su κ, con relative distribuzioni asintotiche:

$$r_{ep}(\kappa) = \frac{(\hat{\kappa} - \kappa)}{\sqrt{i_{\kappa\kappa}(\hat{\kappa}, \hat{\beta})^{-1}}} \dot{\sim} \mathcal{N}(0, 1), \tag{2.13}$$

$$r_p(\kappa) = \text{sgn}(\hat{\kappa} - \kappa)\sqrt{W_p(\kappa)} \dot{\sim} \mathcal{N}(0, 1), \tag{2.14}$$

dove $W_p(\kappa) = 2\left\{l(\hat{\kappa}, \hat{\beta}) - l(\kappa, \hat{\beta}_\kappa)\right\}$.

Le statistiche test per l'ipotesi nulla $H_0 : \kappa = \kappa_0$ sono $r_{ep}(\kappa_0)$ e $r_p(\kappa_0)$, con opportune regioni di rifiuto in base all'ipotesi alternativa. Gli intervalli di confidenza per κ di livello approssimato $(1 - \alpha)$ sono:

$$\hat{C}_{ep}(y) = \left(\hat{\kappa} - z_{1-\alpha/2}\sqrt{i_{\kappa\kappa}(\hat{\kappa}, \hat{\beta})^{-1}}, \hat{\kappa} + z_{1-\alpha/2}\sqrt{i_{\kappa\kappa}(\hat{\kappa}, \hat{\beta})^{-1}}\right),$$

$$\hat{C}_p(y) = \left\{\kappa \in [0, +\infty) : -z_{1-\alpha/2} < r_p(\kappa) < z_{1-\alpha/2}\right\}.$$

Test per la sovradispersione

Spesso è d'interesse effettuare un test d'ipotesi per discriminare tra un modello per sovradispersione e il modello basilare di Poisson: nel caso del modello binomiale negativo ciò si riduce ad effettuare un test sul solo parametro addizionale κ. Si noti infatti che per $\kappa \to +\infty$ i parametri λ_i, coinvolti nella definizione del modello gerarchico, si riducono a costanti $\lambda_i = 1$ e dunque le variabili Y_i assumono distribuzione marginale di Poisson: $Y_i \sim \text{Pois}(\mu_i), i = 1, \ldots, n$. Poiché, come accennato sopra, le condizioni di regolarità sono violate quando $\kappa \to +\infty$, i risultati distributivi nelle (2.13)-(2.14) non sono più validi. Secondo Lawless (1987), la distribuzione nulla appropriata del test del rapporto di verosimiglianza profilo $W_p(+\infty)$ per l'ipotesi nulla $H_0 : \kappa = +\infty$ è una mistura con massa di probabilità 1/2 tra una variabile degenere in 0 e una distribuzione χ^2_1 su $\mathbb{R}^+$. Un approccio alternativo è l'utilizzo della distribuzione *bootstrap* delle statistiche (2.13)-(2.14), come presentato nel §3.4.

Selezione del modello e bontà di adattamento

La selezione delle covariate da inserire nel modello può essere effettuata tramite test locali sui parametri di regressione, nel caso di modelli annidati, oppure anche confrontando il criterio di informazione di Akaike (1.17).

Ottenere un test per valutazione della bontà di adattamento di un modello binomiale negativo non è in genere così semplice come nel caso del modello di Poisson, nel quale possono essere utilizzati la devianza residua o il test T^2 di Pearson, almeno come test approssimati. La stima del parametro addizionale di dispersione κ è infatti solitamente tale da rendere la devianza residua prossima a suoi gradi di libertà e quindi meno utile per la valutazione del modello.

Comunque, per κ fissato al valore di massima verosimiglianza $\hat{\kappa}$, si può definire la devianza residua:

$$D(y;\hat{\mu}) = 2\sum_{i=1}^{n}\left\{y_i \log(y_i/\hat{\mu}_i) - (y_i+\hat{\kappa})\log\left(\frac{\hat{\kappa}+y_i}{\hat{\kappa}+\hat{\mu}_i}\right)\right\},$$

che per $\hat{\kappa}\to+\infty$ ha la forma (2.6). Analogamente, posto $\kappa=\hat{\kappa}$, si definiscono i residui di devianza e di Pearson.

Come suggeriscono Hinde e Demétrio (1997), è possibile utilizzare a scopo diagnostico l'analisi dei residui tramite un grafico probabilità-probabilità che confronti i valori assoluti dei residui di devianza o di Pearson osservati con i corrispondenti quantili della distribuzione normale troncata ai valori positivi (*half-normal plots*). Si può anche aggiungere al grafico una regione di previsione per i residui ottenuta tramite simulazione dal modello corrente (*simulation envelope*): se il modello è correttamente specificato ci si aspetta che i residui osservati cadano all'interno della suddetta regione.

Infine è sempre possibile effettuare il confronto tra le frequenze empiriche dei valori nel supporto della variabile risposta con le corrispondenti frequenze attese, esattamente come avviene per il modello di Poisson (si veda §2.2).

3. Inferenza di ordine superiore

In questo capitolo vengono presentati alcuni sviluppi nella teoria dell'inferenza parametrica basata sulla verosimiglianza, che hanno come scopo il raggiungimento di una maggiore accuratezza nelle procedure di inferenza. Sono state ricercate in particolar modo delle distribuzioni approssimate che apportino un miglioramento alla teoria asintotica del prim'ordine nella misura di uno o due ordini di grandezza. Vengono presentate quindi le due principali vie, o modalità, che sono emerse per l'ottenimento di tale maggior accuratezza: le soluzioni analitiche e il *bootstrap*.

L'attenzione sarà concentrata prevalentemente sul caso in cui il parametro di interesse è scalare, in presenza di altri parametri di disturbo. Si è infatti interessati al parametro di dispersione del modello binomiale negativo.

Per una rassegna di questi argomenti si può consultare Young (2009).

3.1. Inferenza di verosimiglianza di ordine superiore

L'inferenza per un parametro scalare di interesse secondo la teoria della verosimiglianza vista nel Capitolo 1, è basata sulla normalità asintotica delle quantità r_{ep} e r_p, o sul corrispondente risultato asintotico delle quantità W_e e W_p. Spesso tale approssimazione risulta inaccurata, specialmente nel caso siano presenti numerosi parametri di disturbo (Sartori *e altri*, 2016), per esempio quando in un modello di regressione binomiale negativo è coinvolto un numero elevato di covariate e il parametro di interesse è quello di dispersione. Risultati di maggior accuratezza sono necessari anche quando la dimensione campionaria n è esigua. Si ricercano quindi delle procedure di inferenza che garantiscano la rapida convergenza a zero dell'errore di approssimazione all'aumentare di n.

La prima alternativa consiste nel basare le procedure di inferenza non più su r_{ep} o r_p, ma sulla normalità asintotica di una modificazione analitica di r_p, chiamata **radice con segno modificata** del rapporto di verosimiglianza profilo o *modified directed deviance* ed indicata con $r_p^\star$.

La seconda principale alternativa è utilizzare il *bootstrap*, una particolare tecnica statistica di ricampionamento, per stimare con maggior accuratezza la distribu-

zione di r_p. Sono state proposte anche altre alternative, qui non trattate, tra le quali anche un approccio bayesiano.

Ordine di approssimazione

Si chiarisce ora cosa si intende per ordine di approssimazione di un risultato distributivo. Sia di interesse $\psi = \psi(\theta)$, una funzione scalare dei parametri: ciò è equivalente a porre $\psi(\theta) = \tau$, dove $\theta = (\tau, \zeta)$ con τ parametro di interesse; nel modello binomiale negativo sarà $\tau = \kappa$ oppure $\tau = \varsigma(\kappa)$ con $\varsigma(\cdot)$ funzione biunivoca. Si consideri il risultato distributivo (1.15):

$$r_p(\tau) \dot{\sim} \mathcal{N}(0, 1).$$

In realtà si può più precisamente affermare, secondo la consueta teoria di verosimiglianza, che tale distribuzione approssimata è una distribuzione asintotica del primo ordine, che si esprime con:

$$s(\tau) = Pr_\tau \left(r_p(\tau) < r_p(\tau)^{oss} \right) = \Phi \left\{ r_p(\tau)^{oss} \right\} + O_p \left(n^{-1/2} \right), \tag{3.1}$$

dove $\Phi(\cdot)$ è la funzione di ripartizione della distribuzione normale standard, mentre il termine $O_p \left(n^{-1/2} \right)$ viene definito componente di errore del primo ordine. La notazione $r_p(\tau)^{oss}$ indica la quantità $r_p(\tau)$ calcolata sui dati osservati e la funzione $s(\tau)$ è detta **funzione di significatività**; infatti, fissato $\tau = \tau_0$, $s(\tau_0)$ esprime proprio un livello di significatività osservato per un test di ipotesi unilaterale sinistra su τ.

La notazione $O_p(\cdot)$ si riferisce invece alla limitatezza di una successione probabilistica. Sia Z_n una successione di variabili casuali e a_n una successione di costanti numeriche, $i = 1, \ldots, n$. Si scrive $Z_n = O_p(a_n)$ se $\forall \epsilon > 0$, $\exists\ n_\epsilon, c_\epsilon$:

$$Pr \left(\left| Z_n / a_n \right| > c_\epsilon \right) < \epsilon \quad , \quad \forall n > n_\epsilon \ .$$

Scrivere $Z_n = O_p \left(n^{-1/2} \right)$ significa in sostanza che la successione Z_n tende in probabilità allo zero con la stessa velocità di convergenza della successione $n^{-1/2}$.

Componenti di errore del secondo ordine $O_p \left(n^{-1} \right)$ e del terzo ordine $O_p \left(n^{-3/2} \right)$ sono quindi maggiormente trascurabili e se sostituite a $O_p \left(n^{-1/2} \right)$ nella (3.1) darebbero luogo ad un risultato asintotico più accurato.

3.2. Soluzioni analitiche

La strategia più diffusa per ottenimento di risultati asintotici con accuratezza di ordine più elevato è quella che utilizza metodi analitici basati sulla teoria asintotica di ordine elevato. Tali metodi prevedono di definire opportune quantità pivotali o statistiche test che sono modificazioni analitiche delle quantità basilari per l'inferenza di verosimiglianza del primo ordine.

Spesso è necessario utilizzare risultati analitici approssimati, poiché i risultati esatti richiedono elaborazioni oltremodo complesse.

Per una trattazione dettagliata dell'argomento si veda Brazzale *e altri* (2007) oppure Pierce e Bellio (2016).

Radice con segno modificata del rapporto di verosimiglianza profilo

Una delle quantità più importanti nell'approccio analitico è la **radice con segno modificata** del rapporto di verosimiglianza profilo, denotata con $r_p^\star$, della quale esistono diverse versioni. Quella proposta da Barndorff-Nielsen (1991) è:

$$r_p^\star(\tau) = r_p(\tau) + \frac{1}{r_p(\tau)} \log \left\{ \frac{u(\tau)}{r_p(\tau)} \right\},$$

La quantità $u(\tau)$ coinvolge i termini $j(\hat{\theta})$, $j_{\zeta\zeta}(\hat{\zeta}_\tau)$, una statistica ancillare[1] da specificare opportunamente e altri termini che contengono derivate parziali rispetto allo spazio campionario e sono pertanto difficili da ricavare esplicitamente e da trattare. Si può dimostrare che per $r_p^\star(\tau)$ vale il seguente risultato (Brazzale *e altri*, 2007, p.149):

$$s_p^\star(\tau) = Pr_\tau \left(r_p^\star(\tau) < r_p^\star(\tau)^{oss} \right) = \Phi \left\{ r_p^\star(\tau)^{oss} \right\} + O_p \left(n^{-3/2} \right), \tag{3.2}$$

che è un risultato distributivo per $r_p^\star(\tau)$ con accuratezza di terzo ordine.

La quantità $u(\tau)$ è generalmente calcolabile esattamente soltanto se la famiglia parametrica è esponenziale, ma esistono delle opportune approssimazioni di $u(\tau)$ che evitano la specificazione della statistica ancillare e le derivazioni rispetto allo spazio campionario. Una di tali approssimazioni, suggerita da Pierce e Bellio (2016) per la sua generalità, è quella di Skovgaard, che richiede il calcolo di alcuni valori attesi opportunamente approssimabili tramite il metodo Monte Carlo.

[1] Una statistica ancillare è una statistica la cui distribuzione non dipende dal parametro del modello e che, assieme a $\hat{\theta}$, risulta funzione biunivoca della statistica sufficiente.

Mantenendo la notazione $r_p^\star(\tau)$, nel seguito si sottintende che viene utilizzata l'approssimazione di Skovgaard, in base alla quale la (3.2) diventa:

$$s_p^\star(\tau) = Pr_\tau\left(r_p^\star(\tau) < r_p^\star(\tau)^{oss}\right) = \Phi\left\{r_p^\star(\tau)^{oss}\right\} + O_p\left(n^{-1}\right).$$

L'utilizzo dell'approssimazione provoca cioè un aumento dell'errore, che passa da $O_p\left(n^{-3/2}\right)$ a $O_p\left(n^{-1}\right)$. Ad ogni modo, per dati discreti l'accuratezza è sempre del secondo ordine.

Il fattore di correzione che porta alla definizione di $r_p^\star$ a partire da r_p si può scomporre in due parti:

$$r_p^\star(\tau) = r_p(\tau) + \mathrm{NP}_\tau + \mathrm{INF}_\tau,$$

dove NP_τ è una correzione che tiene conto del parametro di disturbo ζ, mentre INF_τ è una correzione di informazione che migliora l'approssimazione normale. Solitamente la correzione NP_τ è più rilevante di INF_τ.

Degno di nota, infine, è che la quantità $r_p^\star(\tau)$ mantiene la proprietà di invarianza rispetto alla parametrizzazione di τ.

Altre soluzioni

Esistono numerose soluzioni analitiche alternative alla radice con segno modificata. Se ne riportano due, sempre basate sulla modificazione delle usuali quantità r_p e W_p:

$$\widetilde{W}_p(\tau) = \frac{p}{E_\tau(W_p(\tau))} W_p(\tau),$$

$$\widetilde{r}_p(\tau) = \frac{r_p(\tau) - E_\tau\left(r_p(\tau)\right)}{\sqrt{Var\left(r_p(\tau)\right)}},$$

dette rapporto di verosimiglianza con correzione di Bartlett e versione standardizzata della radice con segno del rapporto di verosimiglianza. I valori attesi coinvolti vengono approssimati o analiticamente o attraverso *bootstrap* parametrico. Le loro distribuzioni sono rispettivamente χ^2_p con errore $O_p\left(n^{-2}\right)$ e $\mathcal{N}(0,1)$ con errore $O_p\left(n^{-3/2}\right)$.

3.3. Bootstrap parametrico

Il bootstrap parametrico si pone come alternativa alle soluzioni analitiche stimando la distribuzione della quantità di interesse $r_p(\tau)$ tramite campionamento

ripetuto da un modello statistico parametrico che ha per parametri le stime ottenute adattando il medesimo modello sui dati osservati. Questa soluzione si chiama anche *bootstrap prepivoting*, poiché vengono calcolate le distribuzioni bootstrap di una quantità approssimativamente pivotale. Ci sono due possibili strategie per applicare il bootstrap parametrico, che portano a risultati distributivi asintotici di accuratezza diversa: il bootstrap convenzionale e il bootstrap vincolato.

Bootstrap convenzionale

Il bootstrap convenzionale, o di massima verosimiglianza, prevede di simulare un numero elevato B di campioni dal modello statistico che ha per parametro la stima di massima verosimiglianza basata sui dati osservati, ovvero $\hat{\theta}^{oss} = \left(\hat{\tau}^{oss}, \hat{\zeta}^{oss}\right)$. Si indichino con y^b,$b = 1, \ldots, B$, i B campioni bootstrap simulati dal modello con parametro $\hat{\theta}^{oss}$, e con $r_p(\hat{\tau}^{oss})^b$ la quantità $r_p(\tau)$ calcolata per il b-esimo campione bootstrap, valutata in $\hat{\tau}^{oss}, b = 1, \ldots, B$. Allora la stima di $s(\tau)$ basata sul bootstrap convenzionale è:

$$\hat{s}_1(\tau) = \frac{1}{B} \sum_{b=1}^{B} I\left(r_p(\hat{\tau}^{oss})^b < r_p(\tau)^{oss}\right),$$

dove $I(\cdot)$ è la funzione indicatrice. In confronto al risultato (3.1), vale:

$$s(\tau) = Pr_\tau\left(r_p(\tau) < r_p(\tau)^{oss}\right) = \hat{s}_1(\tau) + O_p\left(n^{-1}\right). \tag{3.3}$$

Il bootstrap convenzionale quindi provvede un risultato distributivo con accuratezza del secondo ordine.

Bootstrap vincolato

Il bootstrap vincolato prevede di simulare B campioni dal modello statistico che ha per parametro la stima vincolata di massima verosimiglianza basata sui dati osservati, ovvero $\hat{\theta}_\tau^{oss} = (\tau, \hat{\zeta}_\tau^{oss})$, per τ fissato. Si indichino con y^b,$b = 1, \ldots, B$, i B campioni bootstrap simulati dal modello con parametro $\hat{\theta}_\tau^{oss}$ e con $r_p(\tau)^b$ la quantità $r_p(\tau)$ calcolata per il b-esimo campione bootstrap, valutata in τ, $b = 1, \ldots, B$. Allora la stima di $s(\tau)$ basata sul bootstrap vincolato è:

$$\hat{s}_2(\tau) = \frac{1}{B} \sum_{b=1}^{B} I\left(r_p(\tau)^b < r_p(\tau)^{oss}\right), \tag{3.4}$$

e, con la stessa notazione di sopra, vale:

$$s(\tau) = Pr_\tau \left(r_p(\tau) < r_p(\tau)^{oss} \right) = \hat{s}_2(\tau) + O_p\left(n^{-3/2}\right). \tag{3.5}$$

Il bootstrap vincolato provvede quindi un risultato distributivo con accuratezza del terzo ordine, maggiore di quella ottenuta tramite il bootstrap convenzionale. Si noti tuttavia che mentre nel bootstrap convenzionale la dipendenza da τ è presente solamente nella quantità $r_p(\tau)^{oss}$, nel bootstrap vincolato è presente anche nelle corrispondenti quantità bootstrap. Ciò significa che per l'ottenimento di intervalli di confidenza, che richiedono di calcolare $s(\tau)$ per un insieme di valori di τ, lo sforzo computazionale richiesto dal bootstrap vincolato è notevolmente maggiore.

Radice con segno bootstrap

Per comodità di trattazione, i risultati distributivi per $r_p(\tau)$ basati sulle tecniche bootstrap, cioè (3.3) e (3.5), possono equivalentemente essere espressi tramite la definizione di altre due quantità $r_p(\tau)$ modificate:

$$r_{p1}(\tau) = \Phi^{-1}\left(\hat{s}_1(\tau)\right),$$
$$r_{p2}(\tau) = \Phi^{-1}\left(\hat{s}_2(\tau)\right),$$

per le quali vale quindi l'approssimazione $\mathcal{N}(0,1)$, con accuratezza rispettivamente del secondo e del terzo ordine.

3.4. Test per la sovradispersione

Come già accennato, spesso è d'interesse effettuare un test d'ipotesi per discriminare tra il modello binomiale negativo e il modello basilare di Poisson, ovvero un test di ipotesi su κ. Si ricordi però che in tal caso $H_0 : \kappa = \kappa_0 = +\infty$, quindi si sta effettuando un test sulla frontiera dello spazio parametrico e non è possibile utilizzare il consueto risultato distributivo nullo per $r_p(\kappa_0)$ o equivalentemente per $W_p(\kappa_0)$.

Esiste invece un test bootstrap adatto a questa situazione, presentato a questo punto per pura convenienza espositiva, non essendo l'interesse concentrato sull'accuratezza di uno strumento inferenziale, ma piuttosto sulla sua correttezza.

Si adatti ai dati osservati un modello binomiale negativo, ottenendo la stima dei parametri $\hat{\theta}^{oss} = \left(\hat{\kappa}^{oss}, \hat{\beta}^{oss}\right)$, e si consideri il sistema di ipotesi:

$$H_0 : \kappa = \kappa_0 = +\infty,$$
$$H_1 : \kappa < \kappa_0$$

dove H_0 corrisponde al modello di Poisson, mentre H_1 al più generale modello binomiale negativo. Si adatti ai dati anche il modello di Poisson, ottenendo la stima dei parametri $\tilde{\beta}^{oss}$. Allora $\hat{\theta}^{oss}_{\kappa_0} = \left(\kappa_0, \tilde{\beta}^{oss}\right)$ e il valore osservato della statistica test di interesse, nel modello binomiale negativo, è $W_p(\kappa_0)^{oss} = 2\{l(\hat{\theta}^{oss}) - l(\hat{\theta}^{oss}_{\kappa_0})\}$.

Si indichino con y^b,$b = 1, \ldots, B$, i B campioni bootstrap simulati da un modello di Poisson con parametro $\tilde{\beta}^{oss}$ e con $W_p(\kappa_0)^b$ la quantità $W_p(\kappa_0)$ calcolata stimando sul b-esimo campione il modello binomiale negativo, $b = 1, \ldots, B$. La stima bootstrap della distribuzione nulla di $W_p(\kappa_0)$ è definita dalla funzione di ripartizione:

$$F_B(x) = \frac{1}{B} \sum_{b=1}^{B} I\left(W_p(\kappa_0)^b \leq x\right) \quad , \quad x > 0 \ , \tag{3.6}$$

quindi il livello di significatività osservato corrispondente all'osservazione di $W_p(\kappa_0)^{oss}$ è :

$$\alpha^{oss} \doteq \frac{1}{B} \sum_{b=1}^{B} I\left(W_p(\kappa_0)^b > W_p(\kappa_0)^{oss}\right) = 1 - F_B\left(W(\kappa_0)^{oss}\right), \tag{3.7}$$

in base al quale si decide se accettare o rifiutare l'ipotesi nulla per un fissato α, errore di primo tipo.

3.5. Confronto

Entrambe le soluzioni, analitiche e bootstrap, garantiscono un notevole miglioramento in accuratezza delle procedure di inferenza, specialmente in presenza di molti parametri di disturbo e quando la numerosità campionaria non è elevata. Le soluzioni bootstrap hanno il vantaggio di non dover richiedere l'esplicitazione di alcuna quantità aggiuntiva, come ad esempio $u(\tau)$; per contro richiedono un certo costo computazionale dovuto alle numerose replicazioni bootstrap ed erano perciò impraticabili nel passato. Per ottenere risultati accurati occorre infatti che B sia dell'ordine delle migliaia. Data l'odierna accessibilità a notevoli risorse di calcolo parallelo tale svantaggio è però in parte mitigato (Sartori *e altri*, 2016).

4. Applicazioni e studi di simulazione

Lo scopo di questo capitolo è di illustrare l'applicazione delle metodologie presentate nei precedenti capitoli, tramite l'utilizzo del software `R`, ed in particolare di alcuni suoi pacchetti, tra cui `likelihoodasy`.

Basandosi su alcuni insiemi di dati, sono presentate alcune applicazioni delle procedure di inferenza di verosimiglianza quando la variabile risposta rappresenta un conteggio. Sono di specifico interesse l'adattamento del modello binomiale negativo e la valutazione della sovradispersione.

Viene esposto infine uno studio di simulazione improntato alla verifica empirica dei miglioramenti asintotici ottenuti sia tramite soluzioni analitiche sia tramite *bootstrap*, quando il parametro di interesse è il parametro di dispersione del modello binomiale negativo.

Il codice utilizzato è riportato nell'Appendice B.

4.1. I dati `ships.dat`

Consideriamo i dati `ships.dat`, presentati da McCullagh e Nelder (1989, p.137). I dati provengono dal registro navale del *Lloyd's Register Group*, un'organizzazione di classificazione marittima, e riguardano uno specifico tipo di danno rilevato su un insieme di navi da carico. La variabile di interesse è il numero di navi danneggiate indicata con `incidents`, misurata su $n = 40$ categorie di imbarcazioni. Le categorie sono definite in base alle altre variabili rilevate:

- Il tipo di imbarcazione: `type`, variabile categoriale con cinque modalità: `A`, `B`, `C`, `D`, `E`.
- L'anno di costruzione: `year`, variabile quantitativa.
- Il periodo di funzionamento: `period`, variabile categoriale con possibili modalità: 1960-1974 e 1975-1979, indicate rispettivamente con `60` e `75`.

	type	year	period	service	incidents
1	A	60	60	127	0
2	A	60	75	63	0
3	A	65	60	1095	3
4	A	65	75	1095	4
5	A	70	60	1512	6
6	A	70	75	3353	18
7	A	75	60	0	0
8	A	75	75	2244	11
9	B	60	60	44882	39
10	B	60	75	17176	29
11	B	65	60	28609	58
12	B	65	75	20370	53
⋮	⋮	⋮	⋮	⋮	⋮

Tabella 4.1.: Struttura del dataset `ships.dat`

Media	Varianza	Mediana
8.9	223.84	2

Tabella 4.2.: Statistiche di sintesi per la variabile `incidents`

- Il numero complessivo di mesi di servizio delle imbarcazioni con uguali caratteristiche: `service`, variabile quantitativa.

La struttura della matrice dei dati, limitata alle prime osservazioni, è presentata nella Tabella 4.1. Una sintesi della variabile `incidents` è presentata nella Tabella 4.2. Si nota che la varianza campionaria è considerevolmente maggiore della media campionaria, e perciò l'ipotesi sulla funzione di varianza del modello di Poisson sembrerebbe inadeguata.

Si considera l'adattamento del modello nullo, ovvero senza l'introduzione di covariate, sia di tipo Poisson che binomiale negativo. Formalmente, il modello binomiale negativo di interesse è:

$$\begin{gathered} Y_i \sim \text{Bineg}\left(\kappa, \mu_i\right), i = 1, \ldots, n, \quad Y_i \perp Y_j, \forall i \neq j \\ \eta_i = \eta = \beta_1 \forall i \\ \log\left(E\left(Y_i\right)\right) = \log\left(\mu_i\right) = \log(\mu) = \eta, \end{gathered}$$

dove con Y si intende la variabile `incidents` e per $\kappa \to +\infty$ si ottiene il sotto-modello di Poisson corrispondente.

La funzione che permette di adattare un modello lineare generalizzato di Poisson è: `glm(·,family = poisson)`, mentre per il modello binomiale negativo la funzione è: `glm.nb`, della libreria `MASS`.

$\hat{\beta}_1$	2.186
$\sqrt{\widehat{Var}(\hat{\beta}_1)}$	0.053
T^2_{null}	980.854
Indice di dispersione di Fisher	25.15
Devianza residua	730.25
Gradi di libertà $(n-1)$	39
Test sulla devianza (α^{oss})	≈ 0
AIC	830.12

(a) Modello di Poisson

$\hat{\beta}_1$	2.186
$\sqrt{\widehat{Var}(\hat{\beta}_1)}$	0.29
$\hat{\kappa}$	0.308
$\sqrt{\widehat{Var}(\hat{\kappa})}$	0.077
Devianza residua	42.32
Gradi di libertà $(n-1)$	39
Devianza residua$/(n-1)$	1.085
AIC	236.73

(b) Modello binomiale negativo

Tabella 4.3.: Adattamento dei modelli

Incidenti	0	1	2	3	4	5	6	7	8-18	> 18
Frequenza	14	5	2	1	2	1	2	2	6	5

(a) Frequenze osservate

Incidenti	0	1	2	3	4	5	6	7	8-18	> 18
Frequenza	0.01	0.05	0.22	0.64	1.43	2.54	3.77	4.79	26.48	0.09

(b) Frequenze attese nel modello di Poisson

Incidenti	0	1	2	3	4	5	6	7	8-18	> 18
Frequenza	14.04	4.18	2.64	1.97	1.57	1.31	1.12	0.97	6.1	6.09

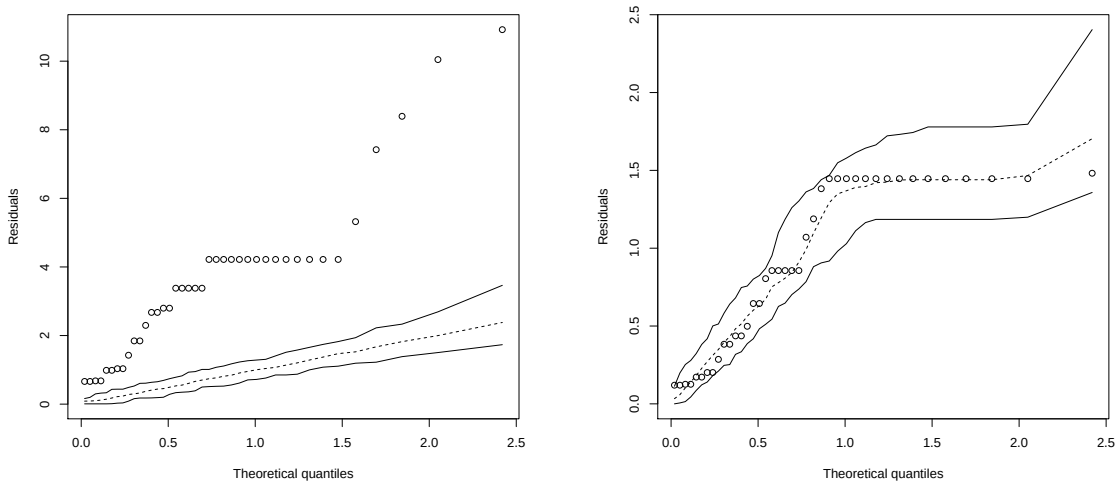
(c) Frequenze attese nel modello binomiale negativo

Tabella 4.4.: Frequenze osservate e attese

Tali funzioni utilizzano le stesse parametrizzazioni utilizzate nel Capitolo 2 e la funzione di legame è sempre quella logaritmica. Le principali quantità ottenute con l'adattamento del modello di Poisson e del modello binomiale negativo sono presentate in Tabella 4.3, mentre le frequenze attese stimate per ciascuno dei due modelli sono presentate, assieme alle frequenze osservate, in Tabella 4.4.

Questi risultati, sebbene relativi a due modelli molto semplici, offrono diversi spunti per la discussione di alcuni aspetti trattati precedentemente solo da un punto di vista teorico. Il valore osservato del test T^2_{null} di Pearson, così come la devianza residua del modello di Poisson, sono molto maggiori dei gradi di libertà residui, indicando uno scarso adattamento del modello di Poisson. Una indicazione analoga si ricava dall'indice di dispersione di Fisher, che indica la presenza di sovradispersione. Il modello binomiale negativo presenta un adattamento notevolmente migliore; per esempio si confrontino i valori del criterio AIC.

Come già accennato, l'introduzione del parametro di dispersione κ rende la



(a) Modello di Poisson

(b) Modello binomiale negativo

Figura 4.1.: *Half-normal plots* per i due modelli

devianza residua del modello binomiale negativo prossima ai suoi gradi di libertà: essa è quindi meno utile per valutare la bontà di adattamento del modello. Per valutare la bontà di adattamento del modello binomiale negativo, rispetto al modello di Poisson, si possono anche confrontare le rispettive frequenze attese dei valori della risposta con le frequenze osservate. Il modello di Poisson prevede una frequenza maggiore per i valori di `incidents` vicini alla media stimata $\hat{\mu} = 8.9$, mentre il modello binomiale negativo consente una maggiore flessibilità che porta ad un ottimo adattamento.

Un altro utile strumento di valutazione è il grafico quantile-quantile detto *half-normal plot*, implementato dalla funzione `hnp` dell'omonimo pacchetto, riportato in Figura 4.1 per entrambe i modelli stimati. L'indicazione che se ne ricava non si discosta dalle precedenti, in quanto i residui del modello binomiale negativo ricadono dentro gli intervalli di simulazione, indicando un buon adattamento, mentre ciò non avviene per il modello di Poisson.

Si noti infine che se si trascura la presenza della sovradispersione, adattando il modello di Poisson, si sottostima la varianza del parametro di regressione: ciò può portare a conclusioni inferenziali scorrette.

Finora sono state utilizzate funzioni generali già implementate in `R` per adattare il modello parametrico di interesse, ma è possibile anche illustrare nel dettaglio, tramite la definizione della funzione di verosimiglianza e l'utilizzo di alcune funzioni di ottimizzazione, come si ricavano tramite il *software* le stime di massima verosimiglianza e, ad esempio, degli intervalli di confidenza per κ.

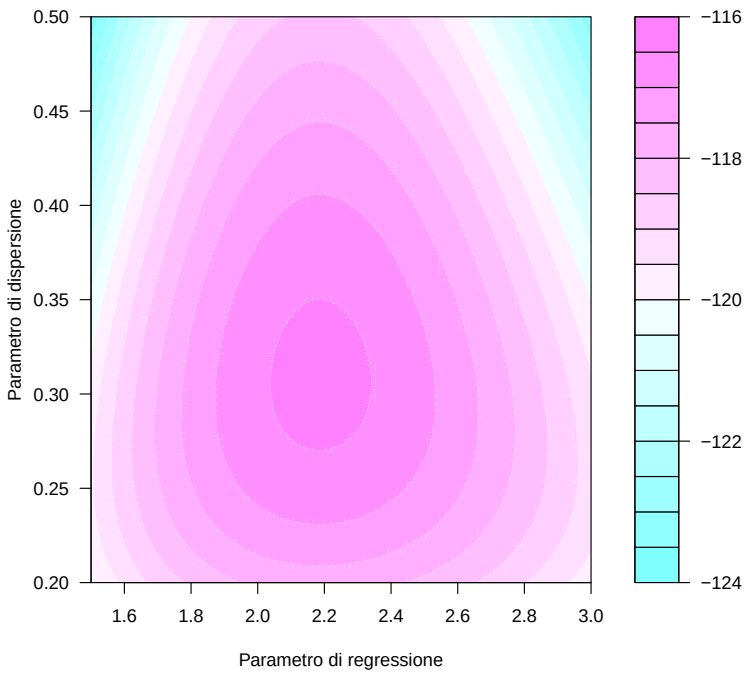


Figura 4.2.: Log-verosimiglianza per il modello binomiale negativo

$W_p(+\infty)^{oss}$	α^{oss}
545.4	≈ 0

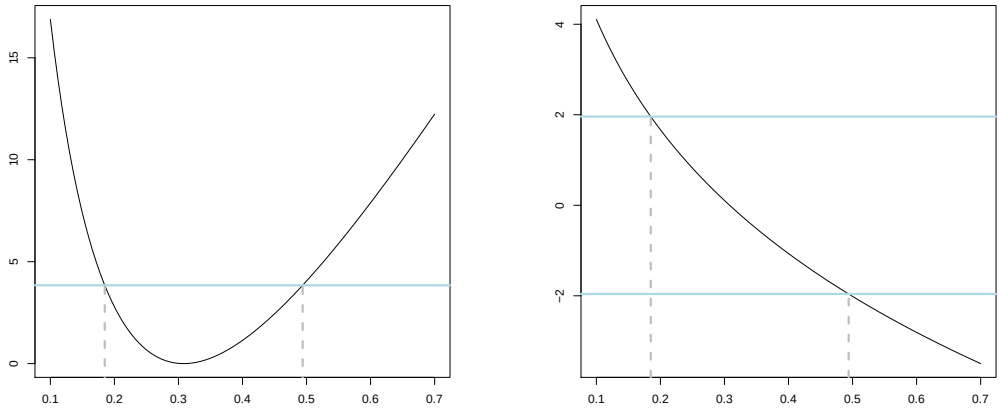
Tabella 4.5.: Test per la sovradispersione nei dati `ships.dat`

Considerando il modello binomiale negativo, una volta scritta la funzione di verosimiglianza, si può utilizzare la funzione `contour` per tracciare un grafico della funzione di log-verosimiglianza per una griglia di valori plausibili per i parametri β_1 e κ, come in Figura 4.2.

Dal grafico della log-verosimiglianza si possono desumere dei valori iniziali, ad esempio $\beta_1^{(0)} = 2.2$ e $\kappa^{(0)} = 0.3$ da utilizzare nella funzione `nlminb`, una funzione di minimizzazione che ha per argomenti l'opposto della funzione di log-verosimiglianza e i valori iniziali $\beta_1^{(0)}, \kappa^{(0)}$ e che permette di ottenere le medesime stime ottenute precedentemente.

Definendo la funzione di verosimiglianza profilo per κ si possono anche tracciare i grafici di $W_p(\kappa)$ e $r_p(\kappa)$ e desumere di conseguenza un intervallo di confidenza per κ, come mostrato in Figura 4.3.

Il test *bootstrap* per la sovradispersione (si veda §3.4) ha dato luogo ai risultati riportati in Tabella 4.5, confermando la preferibilità del modello binomiale negativo.



(a) $W_p(\kappa)$ verso κ (b) $r_p(\kappa)$ verso κ

Figura 4.3.: Grafici di $W_p(\kappa)$ e $r_p(\kappa)$ per il modello binomiale negativo stimato, con intervallo di confidenza per κ di livello $(1-\alpha) = 0.95$

4.2. Il pacchetto `likelihoodasy`

Il pacchetto `likelihoodasy` è stato realizzato da Bellio e Pierce (2016) e contiene diverse funzioni che consentono di adattare un modello statistico parametrico, nonché di semplificare l'utilizzo di alcuni strumenti della teoria asintotica di verosimiglianza, tra i quali la radice con segno modificata. Per utilizzare le funzionalità del pacchetto, l'utente deve fornire autonomamente alcune funzioni.

La prima funzione da definire deve restituire il valore della log-verosimiglianza del modello statistico prescelto, mentre la seconda funzione deve restituire un insieme di dati generati casualmente dal suddetto modello. Entrambe le funzioni hanno come argomenti il parametro θ (`theta`) e una lista (`data`) che contiene i dati osservati y e la matrice del modello X. Una terza funzione restituisce il valore del gradiente della log-verosimiglianza. La definizione di quest'ultima funzione è opzionale, ma solitamente garantisce un risparmio di tempo in fase di calcolo. Si definisce infine la funzione scalare dei parametri $\psi = \psi(\theta)$ che è di interesse per l'inferenza.

Nell'Appendice B.1 sono esposte le funzioni specifiche per il modello binomiale negativo, in una diversa parametrizzazione. Si è utilizzata la parametrizzazione $\theta = (\log(\kappa), \beta) \in \Theta = \mathbb{R}^{p+1}$, perché più sicura per gli algoritmi di ottimizzazione numerica, visto che non impone vincoli di dominio. La funzione di interesse rimane sempre $\psi = \tau = \kappa$.

Il calcolo della statistica test $r_p^\star(\kappa_0)$ avviene attraverso la funzione `rstar`, men-

tre intervalli di confidenza per κ si possono ottenere tramite la funzione `rstar.ci`.

4.3. I dati `ants.dat`

I dati `ants.dat`, presentati da Mackisack (1994), si riferiscono ad un esperimento didattico sulle preferenze alimentari di una particolare specie di formiche onnivore, la *Iridomyrmex purpureus.* Per riprodurre ciò che accade in un parco australiano in presenza di escursionisti, sono stati preparati diversi tipi di esche costituite da combinazioni diverse di quattro tipi di pane, tre tipi di condimento e con la possibilità di aggiungere o meno del burro . Si sono poi effettuati due esperimenti per ogni possibile tipo di esca, per un totale di $n = 48$ esperimenti. L'esperimento, qui descritto sommariamente, consisteva nel contare il numero di formiche che raggiungeva l'esca in un lasso di tempo prefissato, allo scopo di valutare le loro preferenze alimentari. Le variabili presenti nel *dataset*, la cui struttura è riportata in Tabella 4.6, sono:

- Il tipo di pane: `Bread`, variabile categoriale con quattro modalità: `Rye`, `Wholemeal`, `Multi-grain`, `White`.
- Il tipo di condimento: `Filling`, variabile categoriale con tre modalità `Vegemite`, `PeanutButter`, `Ham`.
- L'utilizzo del burro: `Butter`, variabile dicotomica con possibili modalità: `yes`, `no`.
- Il numero complessivo di formiche che hanno raggiunto l'esca in cinque minuti: `Ants`, variabile quantitativa di conteggio.

Si consideri un modello di regressione binomiale negativo con intercetta che ha come variabile risposta `Ants` e include `Filling` e `Butter` come covariate. Dopo aver effettuato le dovute analisi esplorative come nel §4.1, si è selezionato il modello utilizzando gli appropriati test di verosimiglianza e il criterio `AIC`, forniti dalla funzione `glm.nb`, secondo un approccio *forward.* È stato inoltre adattato il sotto-modello di Poisson corrispondente.

Ci si propone ora di illustrare l'applicazione delle principali procedure di inferenza di verosimiglianza su κ, anche utilizzando risultati asintotici di ordine superiore. Non ci si sofferma ulteriormente invece sul coefficiente di regressione β.

Formalmente, il modello binomiale negativo di interesse è:

	Bread	Filling	Butter	Ants
1	Rye	Vegemite	no	22
2	Rye	Vegemite	yes	18
3	Rye	PeanutButter	no	27
4	Rye	PeanutButter	yes	43
5	Rye	Ham	no	68
6	Rye	Ham	yes	44
7	Wholemeal	Vegemite	no	57
8	Wholemeal	Vegemite	yes	29
9	Wholemeal	PeanutButter	no	42
10	Wholemeal	PeanutButter	yes	59
11	Wholemeal	Ham	no	58
12	Wholemeal	Ham	yes	34
⋮	⋮	⋮	⋮	⋮

Tabella 4.6.: Struttura del dataset `ants.dat`

$\hat{\kappa}$	$\sqrt{\widehat{Var}(\hat{\kappa})}$
11.303	3.231

Tabella 4.7.: Stima della dispersione nel modello (4.1)

$$\begin{gathered} Y_i \sim \text{Bineg}\left(\kappa, \mu_i\right), i = 1, \ldots, n, \quad Y_i \perp Y_j, \forall i \neq j \\ \eta_i = \boldsymbol{x}_i \beta = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}, i = 1, \ldots, n \\ \log\left(E\left(Y_i\right)\right) = \log\left(\mu_i\right) = \eta_i, \end{gathered} \tag{4.1}$$

dove con Y, x_2, x_3, x_4 si intendono rispettivamente le variabili `Ants` e le variabili indicatrici relative alle modalità di `Filling` e `Butter`. Per $\kappa \to +\infty$ si ottiene il sotto-modello di Poisson corrispondente. La stima di κ e del suo scarto quadratico medio sono riportate in Tabella 4.7.

Intervalli di confidenza approssimati per κ, basati sulle quantità approssimativamente pivotali $r_{ep}(\kappa), r_p(\kappa), r_p^\star(\kappa), r_{p1}(\tau), r_{p2}(\kappa)$, ottenute tramite la funzione `rstar.ci` e tramite *bootstrap*, sono presentati in Tabella 4.8. Sono state utilizzate $B = 2{\cdot}000$ replicazioni *bootstrap*.

Tali intervalli di confidenza si possono ricavare ad esempio dalle funzioni di significatività, o più precisamente dalle loro approssimazioni asintotiche. Si indichi con $\mathscr{R}(\kappa)$ una delle quantità pivotali appena menzionate; la funzione di significatività per $\mathscr{R}$ è, come già visto:

$$s_{\mathscr{R}}(\kappa) = Pr_\kappa\left(\mathscr{R}(\kappa) < \mathscr{R}(\kappa)^{oss}\right) \doteq \Phi\left(\mathscr{R}(\kappa)^{oss}\right) = \hat{s}_{\mathscr{R}}(\kappa).$$

Quantità	Livello di confidenza 90%	95%	99%
$r_{ep}(\kappa)$	(5.989,16.616)	(4.971,17.634)	(2.981,19.624)
$r_{p}(\kappa)$	(7.068,18.218)	(6.456,20.019)	(5.404,23.87)
$r^{\star}_{p}(\kappa)$	(6.02,15.651)	(5.49,17.185)	(4.578,20.7)
$r_{p1}(\tau)$	(6.02,15.651)	(5.49,17.185)	(4.578,20.7)
$r_{p2}(\kappa)$	(6.02,15.651)	(5.49,17.185)	(4.578,20.7)

Tabella 4.8.: Intervalli di confidenza per κ nel modello (4.1)

Un intervallo di confidenza per κ basato su $\mathscr{R}(\kappa)$ di livello approssimato $(1-\alpha)$ si ottiene allora come:

$$\begin{aligned}\hat{K}_{\mathscr{R}}(y) &= \left\{\kappa \in [0,+\infty) : \frac{\alpha}{2} < \hat{s}_{\mathscr{R}}(\kappa) < 1-\frac{\alpha}{2}\right\} \\ &= \left\{\kappa \in [0,+\infty) :\mid \mathscr{R}(\kappa) \mid < z_{1-\alpha/2}\right\}.\end{aligned}$$

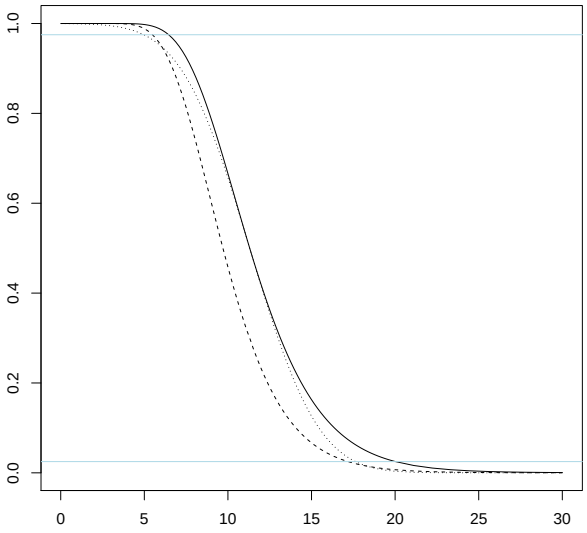
Definita l'ipotesi nulla $H_0 : \kappa = \kappa_0$, contro alternativa bilaterale o unilaterale, la funzione di significatività può parimenti essere utilizzata per ottenere il livello di significatività osservato del test $\mathscr{R}(\kappa_0)$, come sintetizzato di seguito.

$$\begin{aligned}H_1 : \kappa < \kappa_0 &\Rightarrow \alpha^{oss} \doteq \hat{s}_{\mathscr{R}}(\kappa_0) \\ H_1 : \kappa > \kappa_0 &\Rightarrow \alpha^{oss} \doteq 1 - \hat{s}_{\mathscr{R}}(\kappa_0) \\ H_1 : \kappa \neq \kappa_0 &\Rightarrow \alpha^{oss} \doteq 2\min\left\{\hat{s}_{\mathscr{R}}(\kappa_0), 1-\hat{s}_{\mathscr{R}}(\kappa_0)\right\}\end{aligned}$$

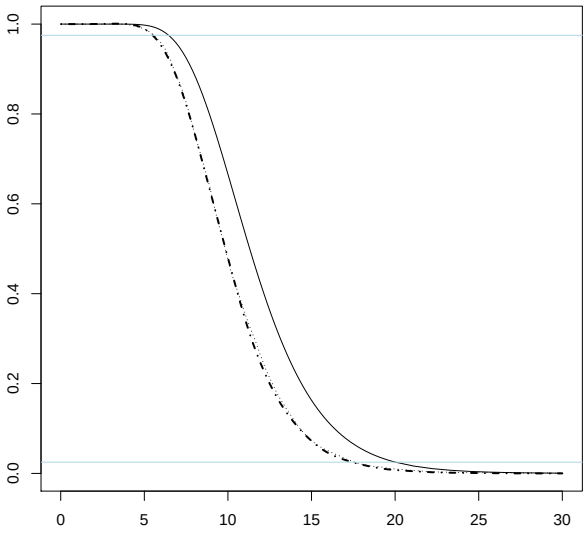
Una visualizzazione grafica delle funzioni di significatività approssimate $\hat{s}_{\mathscr{R}}(\kappa)$ per il modello stimato (4.1), è presentata in Figura 4.4, dove sono evidenziate anche le bande per ottenere gli intervalli di confidenza per κ con $(1-\alpha) = 0.95$. In Figura 4.5 viene riportata la rappresentazione corrispondente in scala $\mathscr{R}(\kappa)$.

Si noti che tutti gli intervalli di confidenza basati su approssimazioni asintotiche di ordine superiore sono sostanzialmente identici e più precisi, a parità di copertura nominale, degli intervalli di Wald e del rapporto di verosimiglianza. Tale differenza si osserva più generalmente confrontando l'andamento e la diversa pendenza delle funzioni di significatività.

Il comportamento anomalo di $r_{p2}(\kappa)$ in Figura 4.5 per $\kappa \to 0$ è legato al numero di replicazioni *bootstrap*, che sebbene elevato è pur sempre finito.

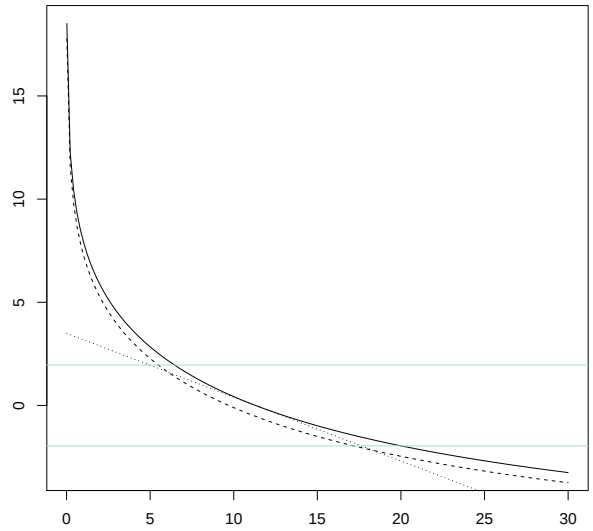


(a) $\hat{s}_p(\kappa)$ (linea continua), $\hat{s}_{ep}(\kappa)$ (linea punteggiata) e $\hat{s}^{\star}_p(\kappa)$ (linea tratteggiata), verso κ

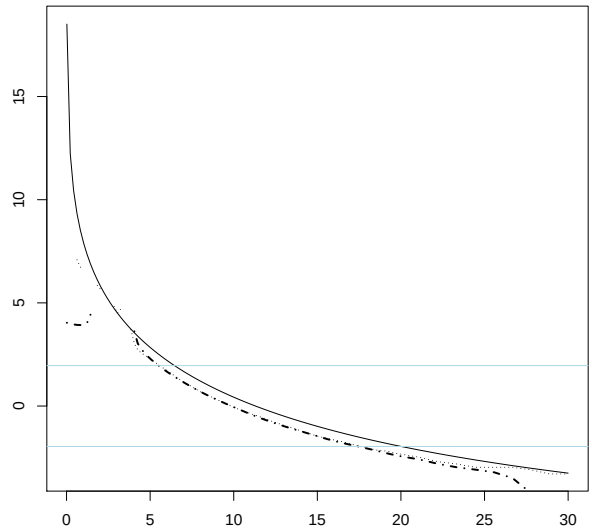


(b) $\hat{s}_p(\kappa)$ (linea continua), $\hat{s}_1(\kappa)$ (linea punteggiata) e $\hat{s}_2(\kappa)$ (linea mista), verso κ

Figura 4.4.: Funzioni di significatività



(a) $r_p(\kappa)$ (linea continua), $r_{ep}(\kappa)$ (linea punteggiata) e $r_p^\star(\kappa)$ (linea tratteggiata), verso κ



(b) $r_p(\kappa)$ (linea continua), $r_{p1}(\kappa)$ (linea punteggiata) e $r_{p2}(\kappa)$ (linea mista), verso κ

Figura 4.5.: Funzioni $\mathscr{R}(\kappa)$

4.4. Studi di simulazione

A partire dai *datasets* `ships.dat` e `ants.dat` si sono condotti degli studi di simulazione volti alla verifica empirica sia dei miglioramenti asintotici presentati per le procedure di inferenza sul parametro κ, sia della corretta utilizzazione del metodo *bootstrap* per la valutazione della sovradispersione.

A tale scopo sono state condotte, tramite `R`, un numero consistente di replicazioni Monte Carlo per stimare con sufficiente precisione, ad esempio, i livelli effettivi di copertura degli intervalli di confidenza per κ, nonché la correttezza della distribuzione nulla del test per la sovradispersione.

Dato il notevole costo computazionale è stato necessario ricorrere al *cluster* `Calculus` del Dipartimento di Scienze Statistiche.

Miglioramenti asintotici

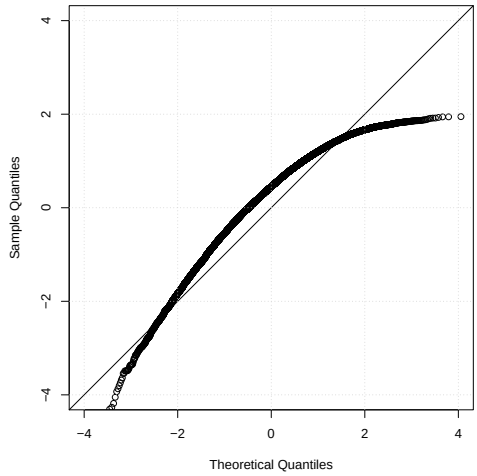
A partire dai dati `ants.dat`, sono stati generati $R = 20{\cdot}000$ campioni Monte Carlo indipendenti $y^r, r = 1, \ldots, R$ dal modello parametrico di regressione (4.1), avente per veri valori dei parametri le stime di massima verosimiglianza $(\kappa_0, \beta_0) = \left(\hat{\kappa}^{oss}, \hat{\beta}^{oss}\right)$ ottenute con i dati osservati. Per ogni campione y^r sono state calcolate le quantità $r_{ep}(\kappa_0)^r, r_p(\kappa_0)^r, r_p^\star(\kappa_0)^r, r_{p1}(\kappa_0)^r, r_{p2}(\kappa_0)^r$, $r = 1, \ldots, R$. Sono state utilizzate $B = 2{\cdot}000$ replicazioni *bootstrap*.

La distribuzione di $r_p(\kappa_0)$ è stimata dalla distribuzione empirica dell'insieme di osservazioni $\{r_p(\kappa_0)^r, r = 1, \ldots, R\}$, e analogamente avviene per le altre quantità. In Figura 4.6 sono riportati i grafici quantile-quantile di normalità per tali insiemi di osservazioni, al fine di valutarne la distribuzione: valori che si discostano dalla bisettrice indicano allontanamento dalla distribuzione normale standard.

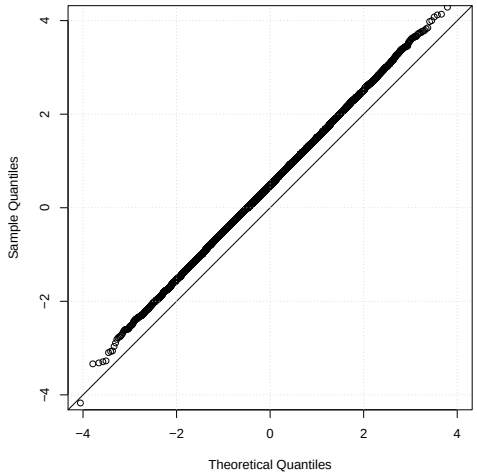
Nella Tabella 4.9 è riportata la percentuale di intervalli di confidenza di livello nominale $(1 - \alpha)$, costruiti a partire da $r_p(\kappa_0)^r$ e dalle altre quantità, che contengono il valore κ_0. Tale percentuale rappresenta una stima Monte Carlo della copertura effettiva degli intervalli di confidenza per κ basati su $r_p(\kappa)$ e le altre quantità, da confrontare con la copertura nominale $(1 - \alpha)$. Nella tabella sono riportate anche le stime degli errori di copertura (sinistro e destro) per i corrispondenti valori nominali di $(\alpha/2)$.

Si nota come sia le verifiche di normalità che le stime della copertura degli intervalli di confidenza portino alle seguenti considerazioni:

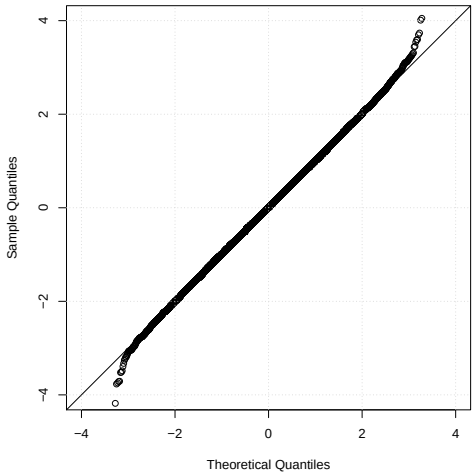
- I risultati asintotici di ordine superiore, sia analitici che basati sul *bootstrap*, portano ad un significativo aumento dell'accuratezza per le procedure di inferenza su κ, rispetto alle soluzioni del primo ordine.



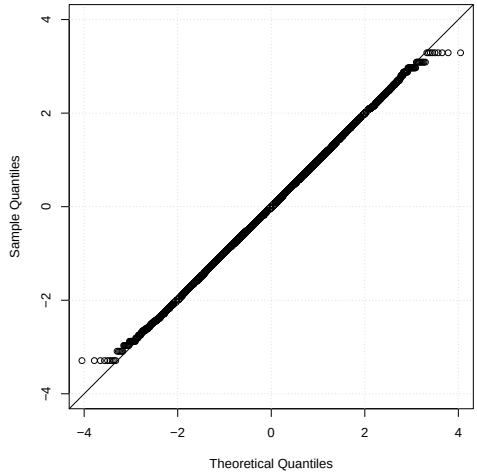
(a) $\{r_{ep}(\kappa_0)^r\}$



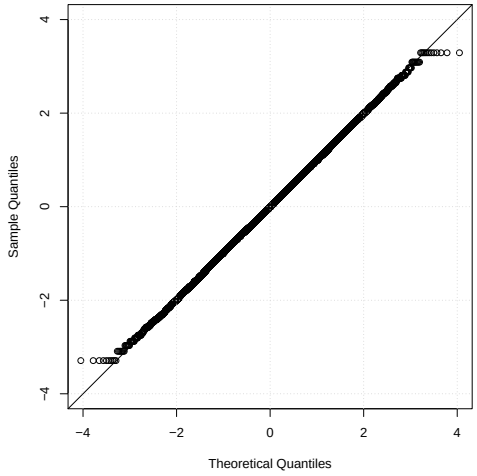
(b) $\{r_p(\kappa_0)^r\}$



(c) $\{r_p^\star(\kappa_0)^r\}$



(d) $\{r_{p1}(\kappa_0)^r\}$



(e) $\{r_{p2}(\kappa_0)^r\}$

Figura 4.6.: Grafici per la valutazione della normalità

Quantità di riferimento	Livello di copertura nominale $(1-\alpha)$ 90%	95%	99%
$r_{ep}(\kappa)$	94.2 (3.1; 2.7)	98.2 (1.8; 0)	99.4 (0.6; 0)
$r_p(\kappa)$	85.7 (1.8; 12.5)	91.8 (0.8; 7.4)	97.8 (0.1; 2.1)
$r_p^\star(\kappa)$	89.9 (4.9; 5.2)	94.9 (2.5; 2.6)	98.9 (0.6; 0.5)
$r_{p1}(\kappa)$	90.2 (4.9; 4.9)	95.1 (2.4; 2.5)	99.1 (0.4; 0.5)
$r_{p2}(\kappa)$	90.1 (5; 4.9)	95.1 (2.5; 2.4)	99.2 (0.4; 0.4)
Errore Monte Carlo	0.21 (0.15)	0.15 (0.15)	0.07 (0.15)

Tabella 4.9.: Stime dei livelli di copertura effettivi e degli errori di copertura (sinistro; destro) basati sulle diverse quantità

- Le procedure di tipo Wald sono caratterizzate da minore accuratezza, oltre che dalla mancanza della proprietà di equivarianza.
- Le procedure basate sulla teoria asintotica del primo ordine sono affette da un errore di asimmetria, oltre ad un problema di sottocopertura.
- Le procedure basate sulla teoria asintotica di ordine superiore, almeno per quanto riguardo lo scenario considerato, si equivalgono fra di loro in accuratezza.

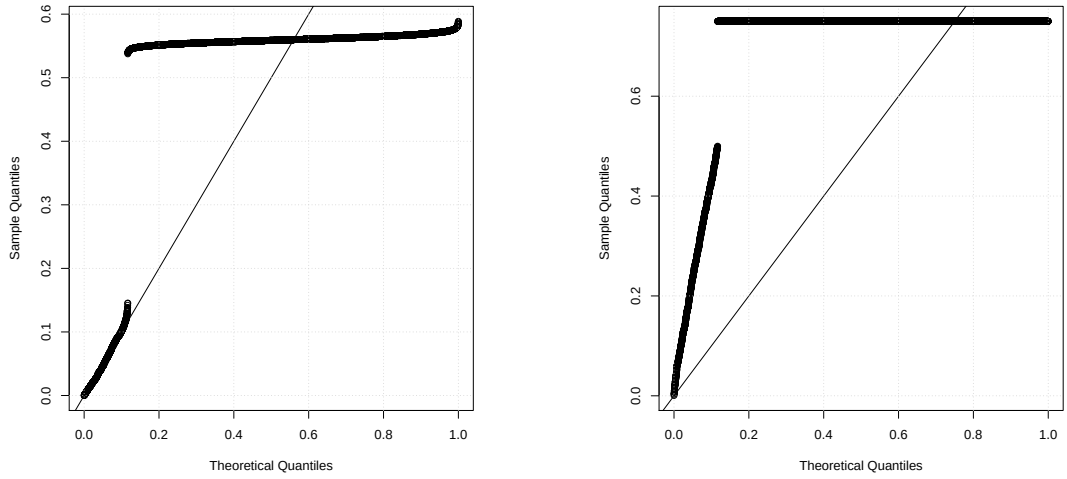
Test per la sovradispersione

Si considerino nuovamente i dati `ships.dat`, ai quali viene adattato il modello di regressione binomiale negativo che ha come variabile risposta `incidents` e come covariate `service`, `period` e `year`; si consideri in particolare la stima vincolata $\tilde{\beta}^{oss}$ sotto l'ipotesi $H_0 : \kappa = +\infty$. Sono stati generati $R = 10{\cdot}000$ campioni Monte Carlo indipendenti $y^r, r = 1, \ldots, R$ dal modello parametrico di regressione di Poisson avente le medesime covariate e $\tilde{\beta}^{oss}$ come vero valore del parametro. Per ogni campione y^r è stato adattato il modello di regressione binomiale negativo ed è stato calcolato il valore osservato della statistica test $W_p(+\infty)^r, r = 1, \ldots, R$, che saggia l'ipotesi nulla $H_0 : \kappa = +\infty$. Tramite *bootstrap*, ponendo $B = 2{\cdot}000$, sono stati quindi calcolati, per ogni campione, i valori di significatività osservata approssimati (3.7).

Se la stima *bootstrap* per la distribuzione nulla della statistica test per la sovradispersione è corretta, allora ci si aspetta che la proporzione dei casi in cui si rifiuta l'ipotesi nulla per α fissato negli R campioni coincida, a meno dell'errore di simulazione, con α, errore di primo tipo. Tale proporzione è:

Distribuzione nulla di riferimento	Errore α nominale 0.005	0.01	0.05	0.1
Bootstrap	0.005	0.01	0.053	0.099
Mistura	0	0.001	0.007	0.018
Errore Monte Carlo	0.001	0.001	0.0022	0.003

Tabella 4.10.: Stime monte carlo dell'errore di primo tipo effettivo



(a) Metodo Bootstrap

(b) Risultato asintotico

Figura 4.7.: Grafici per la valutazione dell'uniformità dei valori di significatività osservata

$$\hat{\alpha} = \sum_{i=1}^{R} \left(\alpha_i^{oss} < \alpha\right) / R.$$

Si noti che i campioni Monte Carlo devono essere necessariamente stati generati dal modello di Poisson, ovvero sotto l'ipotesi nulla. I risultati sono esposti in Tabella 4.10, per diversi valori nominali α, affiancati ai risultati ottenuti utilizzando la distribuzione asintotica mistura presentata nel §2.5 in luogo della distribuzione *bootstrap*. In Figura 4.7 è presentato il confronto tra i valori di significatività osservata e i corrispondenti quantili teorici della distribuzione uniforme.

Si può osservare che i valori di significatività osservata sono compatibili con la distribuzione uniforme soltanto per valori di α usuali ($\alpha < 0.1$). È doverosa a questo punto una precisazione.

Per una statistica con distribuzione nulla continua, la variabile aleatoria α^{oss} ha distribuzione nulla uniforme, mentre ciò non accade se la distribuzione nulla

del test è una mistura tra una distribuzione discreta e una continua, come per il test sulla sovradispersione.

In tal caso è buona norma applicare al livello di significatività osservata usuale la correzione suggerita da Franck (1986), ottenendo il cosiddetto *mid-Pvalue.* Tale correzione garantisce che il livello di significatività osservata abbia distribuzione nulla uniforme *per quanto possibile* (Blocker *e altri*, 2006), ed è stata applicata ai livelli di significatività calcolati nella simulazione in esame.

Sia ad esempio W la statistica di interesse, che porta al rifiuto dell'ipotesi nulla H_0 per valori elevati, e sia W^{oss} il suo valore osservato. Il *mid-Pvalue* è definito come:

$$\alpha_m^{oss} = Pr_{H_0}\left(W > W^{oss}\right) + \frac{1}{2} Pr_{H_0}\left(W = W^{oss}\right),$$

e garantisce la validità, piuttosto che l'esattezza, poiché vale $Pr_{H_0}\left(\alpha_m^{oss} \leq \alpha\right) \leq \alpha$ invece di $Pr_{H_0}\left(\alpha_m^{oss} \leq \alpha\right) = \alpha$.

Dalla simulazione possiamo infine trarre le seguenti conclusioni empiriche :

- La distribuzione nulla *bootstrap* (3.6) per il test $W_p(+\infty)^r$ è corretta, per valori di α usuali.
- La distribuzione mistura asintotica nulla suggerita da Lawless (1987), pur essendo corretta, non è sufficientemente accurata per la numerosità campionaria del *dataset* studiato.
- Per entrambe le soluzioni, l'utilizzo di valori di α più elevati di quelli usuali porta a decisioni conservative, a patto che sia stato utilizzato il *mid-Pvalue.*

Conclusioni

La relazione si è posta come obiettivo di affrontare il problema della sovradispersione nel contesto dei modelli di regressione per dati di conteggio. L'interesse si è concentrato sul modello binomiale negativo, per il quale sono stati utilizzati i metodi di verosimiglianza per l'inferenza.

Il modello binomiale negativo, grazie al parametro di dispersione, permette di trattare con sufficiente flessibilità i dati di conteggio, come è stato possibile verificare anche attraverso alcuni esempi basati su dati reali.

L'introduzione di alcune tecniche più avanzate per l'inferenza basata sulla verosimiglianza, quali approssimazioni analitiche di ordine superiore e metodi *bootstrap*, ha permesso di spingersi oltre sia in termini di precisione dell'inferenza che di varietà di ipotesi verificabili. Tali potenzialità sono emerse empiricamente anche dagli studi di simulazione effettuati.

È giusto ricordare infine che il modello binomiale negativo è soltanto una soluzione tra le tante possibili. Sono stati fatti perciò alcuni accenni a strade alternative per trattare la sovradispersione. Anche in tali modellazioni alternative sarebbe interessante valutare l'utilizzo di tecniche di verosimiglianza di ordine superiore.

A. Alcune distribuzioni

A.1. Famiglia di dispersione esponenziale

La funzione di probabilità, o di densità di probabilità di una variabilie aleatoria Y_i, sia essa discreta o continua, appartiene alla famiglia di dispersione esponenziale univariata se si può scrivere:

$$p(y_i; \vartheta_i, \phi) = \exp\left\{\frac{\vartheta_i y_i - b(\vartheta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}, \tag{A.1}$$

con $y_i \in S_{Y_i} \subseteq \mathbb{R}$, detto supporto di Y_i, $\vartheta_i \in \mathbb{R}$, detto parametro naturale, e $a_i(\phi) > 0$. Spesso $a_i(\phi) = \phi$, con ϕ parametro di dispersione. Anche se $a_i(\phi)$ può assumere diverse forme, in questa sede si considerano solo distribuzioni discrete con $a_i(\phi) = 1$.

Si può dimostrare che:

$$\mu_i(\vartheta_i) = E(Y_i) = b'(\vartheta_i),$$
$$Var(Y_i) = a_i(\phi)b''(\vartheta_i). = a_i(\phi)v(\mu_i),$$

dove $v(\mu_i) = b''(\vartheta_i)\,|_{\vartheta_i=\vartheta_i(\mu_i)}$ è detta funzione di varianza. Nella famiglia di dispersione esponenziale $(\mu_i, a_i(\phi))$ è una riparametrizzazione di $(\vartheta_i, a_i(\phi))$ e dunque è giustificata la seguente notazione:

$$Y_i \sim DE_1\left(\mu_i, a_i(\phi)v(\mu_i)\right).$$

La distribuzione di Poisson appartiene alla famiglia a dispersione esponenziale. Per dimostrare ciò si ponga $Y_i \sim \text{Pois}(\mu_i)$; allora la funzione di probabilità di Y_i è:

$$p(y_i; \mu_i) = \mu_i^{y_i} e^{-\mu_i}/y_i! = \exp\left\{y_i \log(\mu_i) - \mu_i - \log(y_i!)\right\},$$

che, posto $\vartheta_i = \log(\mu_i)$, $a_i(\phi) = 1$, $b(\vartheta_i) = \mu_i$, $c(y_i, \phi) = -\log(y_i!)$, soddisfa la (A.1). Si trova inoltre $v(\mu_i) = \mu_i$, e si utilizza perciò la notazione $Y_i \sim DE_1\left(\mu_i, \mu_i\right)$.

A.2. Distribuzione binomiale negativa

La distribuzione di Pascal è una distribuzione di probabilità discreta a due parametri che descrive il numero di fallimenti precedenti il successo κ-esimo in una successione di prove di Bernoulli indipendenti di parametro π.

Siano $X_1, X_2, \ldots, X_n$ una successione di variabili indipendenti con identica distribuzione di Bernoulli: $X_i \sim \text{Bi}(1, \pi), i = 1, \ldots, n$. La probabilità di successo di una singola prova è π mentre la probabilità di fallimento è $1 - \pi$. Allora la variabile $Y = \min\{y : X_1 + \cdots + X_{y+n} = \kappa\}$ segue una distribuzione di Pascal di parametri $\kappa \in \mathbb{N}$ e $\pi \in (0, 1)$: $Y \sim \text{Pasc}(\kappa, \pi)$ e la sua funzione di probabilità è:

$$p(y; \kappa, \pi) = \binom{y + \kappa - 1}{y} \pi^\kappa (1 - \pi)^y = \frac{(y + \kappa - 1)!}{y!(\kappa - 1)!} \pi^\kappa (1 - \pi)^y.$$

Si può generalizzare la distribuzione di Pascal al caso in cui κ possa variare nell'insieme dei numeri reali positivi, esprimendo il coefficiente binomiale tramite la funzione Gamma, per la quale vale la relazione $\Gamma(x + 1) = x!, \forall x \in \mathbb{N}$. In tal caso Y è definita dalla seguente funzione di probabilità:

$$p(y; \kappa, \pi) = \frac{\Gamma(y + \kappa)}{\Gamma(y + 1)\Gamma(\kappa)} \pi^\kappa (1 - \pi)^y, \tag{A.2}$$

dove $y \in \mathbb{N}, \pi \in (0, 1), \kappa \geq 0$, ed è detta distribuzione binomiale negativa: $Y \sim \text{Bineg}(\kappa, \pi)$. I parametri κ e π sono chiamati, rispettivamente, parametro di scala e di probabilità. La distribuzione binomiale negativa di parametri κ e π ha media $\mu = \kappa(1 - \pi)/\pi$ e varianza $\sigma^2 = \kappa(1 - \pi)/\pi^2 = \mu + \mu^2/\kappa$; vale inoltre la riparametrizzazione dove il parametro di probabilità è sostituito dalla media.

Si dimostra che, soltanto per κ fissato, la (A.2) appartiene alla famiglia di dispersione esponenziale (A.1). Sia $Y_i \sim \text{Bineg}(\kappa, \pi_i) \equiv \text{Bineg}(\kappa, \mu_i)$, allora:

$$p(y_i; \kappa, \mu_i) = \exp\left\{ y_i \log\left(\frac{\mu_i}{\kappa + \mu_i}\right) + \kappa \log\left(\frac{\kappa}{\mu_i}\right) + c(y_i, \kappa) \right\},$$

e il risultato si ottiene pondendo $\vartheta_i = \mu_i/(\kappa + \mu_i)$, $b(\vartheta_i) = -\kappa \log\{\kappa/(\kappa + \mu_i)\}$ e $v(\mu_i) = \mu_i(1 + \mu_i/\kappa)$. Per κ non fissato, invece, non si riesce ad esprimere la densità come richiesto nella (A.1).

Per ulteriori dettagli, ad esempio definizioni alternative della distribuzione oppure chiarimenti sul termine *binomiale negativa* si rimanda ad esempio a Bartko (1962).

B. Codice utilizzato

B.1. Funzioni per il pacchetto `likelihoodasy`

Funzione di log-verosimiglianza

```
loglik=function(theta,data)     # theta : il vettore di parametri
{
  y=data$y
  X=data$X       # X matrice del modello
  logk=theta[1]
  beta=theta[-1]
  k=exp(logk) # Riparametrizzazione conveniente per gli algoritmi
  mu=exp(X%*%beta)    # vettore delle medie definite dal modello

  # verosimiglianza:
  result=sum(dnbinom(y,k,mu=mu,log = T))
  return(result)
}
```

Funzione generatrice

```
gendata=function(theta,data)
{
  X=data$X       # X matrice del modello
  n=nrow(X)

  logk=theta[1]
  beta=theta[-1]
  k=exp(logk)
  mu=exp(X%*%beta)
```

```
  data$y=rnegbin(n,mu=mu,k)
  return(data)
}
```

Funzione gradiente

```
gradiente=function(theta,data)

{
  y=data$y
  X=data$X      # X matrice del modello
  logk=theta[1]
  beta=theta[-1]
  k=exp(logk)
  mu=exp(X%*%beta)

  #funzione punteggio per beta:
  score.beta=t(X)%*%((y-mu)/(1+mu/k))

  # funzione punteggio per logk:
  score.logk=sum( k*(-(k+y)/(mu+k)+logk+1-log(mu+k)+
                  digamma(k+y)-digamma(k)) )

  result=c(score.logk,score.beta)
  return(result)
}
```

Funzione di interesse

```
interesse=function(theta)
{
  k=exp(theta[1]) # parametrizzazione originale
  return(k)
}
```

B.2. Analisi di ships.dat e ants.dat

Dati ships.dat

```
library(MASS);source("funz_ver.R");library(hnp)
library(parallel);data("ships")
ships$period=as.factor(ships$period)
attach(ships);table(incidents)
M1=glm(incidents~service+period+year,family="poisson")
summary(M1)
M2=glm.nb(incidents~service+period+year)
summary(M2)
y=incidents;ships2=ships[,2:4]
####################################
## Funzione di log-verosimiglianza
# qui assumiamo log(mu)=b (modello nullo)
loglik=function(th,y)  # th è il parametro c(b,log(k))
{
  b=th[1];k=th[2];mu=exp(b)
  loglik1=function(x)
  {
    x*log(mu)+k*log(k)-(k+x)*log(mu+k)+
    log(gamma(k+x)/gamma(k)/gamma(x+1))
  }
  val=sum(sapply(y,loglik1));return(val)
}
nloglik=function(th,y) -loglik(th=th,y=y)
###### esempio
M0pois=glm(y~1,family = poisson);M0=glm.nb(y~1);k0=M0$theta
b0=coef(M0);logLik(M0)  # Verosimiglianza definita in R
loglik(c(b0,k0),y) # Verosimiglianza da noi costruita
#### Esempio di ricerca grafica
ss1=seq(1,3.5,length.out =80);ss2=seq(0.001,2,length.out =80)
tab=expand.grid(ss1,ss2)
val=apply(tab,1,function(x) loglik(x,y=y))
val=matrix(val,80,80);par(pty="m")
filled.contour(ss1,ss2,val,xlab="Parametro di regressione",
```

```
                ylab="Parametro di dispersione")
fit=nlminb(c(2.2,0.3),nloglik,y=y,
lower = c(0.9,0.01),upper = c(3.5,0.7));fit$par
#### Calcolo di intervalli di confidenza per k
#### basati sulla verosimiglianza profilo
nprofk=function(b,k){return(-loglik(c(b,k),y=y))}
Wp=function(k)
{
  fWp=function(x)
  { return( 2*(loglik(fit$par,y=y)+
    nlm(nprofk,b0,x)$minimum))
  }
  sapply(k,fWp)
}
rp(ICk);Low=optimize(function(x) abs(Wp(x)-qchisq(0.95,1)),
interval = c(0.1,k0))$minimum
Up=optimize(function(x) abs(Wp(x)-qchisq(0.95,1)),
interval = c(k0,1.4))$minimum
ICk=c(Low,Up)
ICk ## intervallo di confidenza per k basato su Wp
rp=function(k) sign(k0-k)*sqrt(Wp(k))
curve(rp(x),0.1,0.7,xlab="k",ylab="rp(k)")
## interallo per k alla Wald:
M0$theta+c(-1,1)*qnorm(0.975)*M0$SE.theta
```

Dati `ants.dat`

```
library(MASS);library(boot);library(parallel);library(pspline)
source("funz_ver.R") # funzioni di verosimiglianza
ncpus=min(10,detectCores())
ants=read.table("ants.dat",header=T)
ants$Bread=as.factor(ants$Bread)
ants$Butter=as.factor(ants$Butter)
ants$Filling=as.factor(ants$Filling)
y=ants$Ant_count;attach(ants)
## Glm
```

```
M1=glm(y~Filling+Butter,family = poisson);summary(M1)
## Bin Neg
M2=glm.nb(y~Filling+Butter);summary(M2)
Mle=c(log(M2$theta),coef(M2))cash=0
rsci= rstar.ci(data=list(y=M2$y,X=model.matrix(M2)),
       thetainit =Mle,floglik = loglik,
       fpsi = interesse,datagen = gendata,
       fscore = gradiente,R=1000,
       trace=F,seed=10,psidesc="K, parameter of dispersion")
r_val=function(k,dati,thetainit,ronly=F)
{rs= rstar(data=dati,thetainit =thetainit,floglik = loglik,
       fpsi = interesse,psival =k,datagen = gendata,
       fscore = gradiente,R=1000,
       trace=F,seed=10,psidesc="K, parameter of dispersion",
       ronly =ronly)
 c(rs$r,rs$rs)
}
kseq=seq(0.02,30,length=100)    # valori di k
rvals=sapply(kseq,r_val,dati=list(y=M2$y,X=model.matrix(M2)),
thetainit=Mle) # valori di r
wvals=(M2$theta-kseq)/M2$SE.theta;par(pty="s")
plot(kseq,pnorm(rvals[1,]));points(kseq,pnorm(rvals[2,]))
points(kseq,pnorm(wvals))
######## Bootstrap convenzionale (prepivot Mle)
B=2000
boot_Mle=replicate(B,gendata(Mle,data=
list(y=NULL,X=model.matrix(M2))))
boot_Mle_r=apply(boot_Mle,2,
function(x) r_val(exp(Mle[1]),x,thetainit = Mle,ronly=T))
phat_Mle=sapply(rvals[1,], function(x) mean(boot_Mle_r<=x))
######## Bootstrap vincolato  (prepivot da H0) (terz'ordine)
cash=0 #contatore
boot_vincolato=function(k,dati,thetainit,ronly=T)
{
  rs= rstar(data=dati,thetainit =thetainit,floglik = loglik,
        fpsi = interesse,psival =k,datagen = gendata,
        fscore = gradiente,R=1000,
```

```
        trace=F,seed=10,psidesc="K,parameter of dispersion",
        ronly =ronly)
 stima_vincolata=rs$theta.hyp
 gg=function(x,mle)
{gendata(stima_vincolata,data=list(y=NULL,X=dati$X))}
 ff=function(x)
{r_val(exp(stima_vincolata[1]),x,thetainit = Mle,ronly=T)}
 dist_boot_cons_r=boot(dati,ff,R=B,sim="parametric",ran.gen = gg,
 parallel = "multicore",ncpus=ncpus)$t
 cat("-------BOOT",cash,"\n")
 cash<<-cash+1
 return(mean(dist_boot_cons_r<=rs$r))
}
phat_cons=sapply(kseq,function(x) boot_vincolato(x,
        dati =list(y=M2$y,X=model.matrix(M2)),thetainit = Mle ))
######## Intervalli bootstrap per k
phatINVB1=sm.spline(x=phat_Mle_new,y=kseq)
phatINVB2=sm.spline(x=phat_cons_new,y=kseq)
phatB1=sm.spline(y=phat_Mle_new,x=kseq)$ysmth
phatB2=sm.spline(y=phat_cons_new,x=kseq)$ysmth
B1cif=function(a) predict(phatINVB1,x=c(1-a/2,a/2))
B2cif=function(a) predict(phatINVB2,x=c(1-a/2,a/2))
B1ci=t(sapply(c(0.1,0.05,0.01),B1cif))
B2ci=t(sapply(c(0.1,0.05,0.01),B2cif))
```

B.3. Studi di simulazione

Miglioramenti asintotici

```
## Una volta caricate le analisi del
## dataset ants.dat, precedentemente svolte...
ncpus=min(10,detectCores());IpNulla=c(log(M2$theta),coef(M2))
simulation=function(theta0=IpNulla,X0,Nsim=1000,B=2000)
{
  Co=0  # contatore
  MonteCarlo=function(data_i) ## funzione che calcola i test
```

```
 {
   Rst=invisible(rstar(data=list(y=data_i,X=X0),
         thetainit =theta0,floglik = loglik,
         fpsi = interesse,psival =exp(theta0[1]),
     datagen = gendata, fscore = gradiente,R=1000,trace=F ) )
   hat=Rst$theta.hat;hat0=Rst$theta.hyp
   stime=cbind(hat,hat0)
  testboot=function(databoot,choose)
## choose serve per scegliere tra boot mle e vincolato
    { invisible(rstar(data=list(y=databoot,X=X0),
            thetainit =stime[,choose],floglik = loglik,
            fpsi = interesse,psival =exp(stime[1,choose]),
            datagen = gendata,
            fscore = gradiente,R=1000,trace=F,ronly = T))$r
     }
 ####### BOOTSTRAP MLE
   gendata_boot_MLE=function(data,mle)
       { gendata(hat,list(y=NULL,X=X0))$y}
  # funzione che calcola la distribuzione bootstrap
  # (prepivot di mle) è il bootstrap convenzionale(non vincolato)
    dist_boot_MLE=boot(data_i,testboot,sim="parametric",R=B,
      ran.gen = gendata_boot_MLE,
      ncpus = ncpus,parallel="multicore",choose=1)$t
 ####### BOOTSTRAP VINCOLATO
    gendata_boot_VINC=function(data,mle)
      {gendata(hat0,list(y=NULL,X=X0))$y}
     # funzione che calcola la distribuzione bootstrap prepivot
     # è il bootstrap vincolato
     dist_boot_VINC=boot(data_i,testboot,sim="parametric",
       R=B,ran.gen = gendata_boot_VINC,
       ncpus = ncpus,parallel="multicore",choose=2)$t
 ############################
 rBmle=qnorm(mean(dist_boot_MLE<=Rst$r))
 rBvinc=qnorm(mean(dist_boot_VINC<=Rst$r))
 Co<<-Co+1
 cat(Co,"/",Nsim,"\n")
 Wald=(Rst$psi.hat-Rst$psi.hyp)/Rst$se.psi.hat
```

```
  return(list(r=Rst$r,rs=Rst$rs,wa=Wald,rBmle=rBmle,rBvinc=rBvinc))
 }
# funzione che simula i campioni (con theta0 come parametro)
samples=replicate(Nsim,gendata(theta0,data=list(y=NULL,X=X0))$y)
## Valori delle statistiche
result=apply(samples,2,MonteCarlo)
vec=1:Nsim
r_values=sapply(vec, function(x) result[[x]]$r)
rs_values=sapply(vec, function(x) result[[x]]$rs)
wald_values=sapply(vec, function(x) result[[x]]$wa)
rBootmle_values=sapply(vec, function(x) result[[x]]$rBmle)
rBootvinc_values=sapply(vec, function(x) result[[x]]$rBvinc)
return(list(r=r_values,rs=rs_values,wa=wald_values,
rBm=rBootmle_values,rBV=rBootvinc_values))
}
sim=simulation(X0=model.matrix(M2),Nsim = 20000)
```

Test per la sovradispersione

```
library(MASS);library(boot);source("funz_ver.R")
library(parallel);data("ships");ncpus=min(10,detectCores())
ships$period=as.factor(ships$period);attach(ships)
M1=glm(incidents~service+period+year,family="poisson");summary(M1)
M2=glm.nb(incidents~service+period+year);summary(M2)
mydata=list(y=M2$y,X=model.matrix(M2))
theta_mle=c(log(M2$theta),coef(M2))
ships2=ships[,2:4]
myr=function(y=mydata$y,covariate)
{
  Mneg= glm.nb(y~.,data = covariate)
  Mpois= glm(y~.,data = covariate,family = poisson)
  Wp_er= as.numeric(2*(logLik(Mneg)-logLik(Mpois) ))
  ## alcuni valori sono negativi per l'approssimazione
  ## delle verosimiglianze (es: -0.0002)
  Wp=max(0,Wp_er);Wp
}
```

```
#### Bootstrap da H0 stimata
ran_poisson=function(data,mle){rpois(length(data),lambda =mle)}
BB=boot(mydata$y,myr,R=2000,ran.gen = ran_poisson,
sim="parametric",mle=fitted(M1),parallel = "multicore",
ncpus = ncpus,covariate=ships2)
############ Simulazione
simulation=function(M1,covariate,Nsim=1000,B=2000)
{  Co=0  # contatore
   MonteCarlo=function(y)
  {
    Mneg= glm.nb(y~.,data = covariate)
    Mpois= glm(y~.,data = covariate,family = poisson)
    Wp_er= as.numeric(2*(logLik(Mneg)-logLik(Mpois) ))
    ## alcuni valori sono negativi per l'approssimazione
    ## delle verosimiglianze (es: -0.0002)
    Wp=max(0,Wp_er);fit0=fitted(Mpois)
    testboot=function(databoot)
    {
      MnegB= glm.nb(databoot~.,data = covariate)
      MpoisB= glm(databoot~.,data = covariate,family = poisson)
      Wp_erB= as.numeric(2*(logLik(MnegB)-logLik(MpoisB) ))
      ## alcuni valori sono negativi per l'approssimazione
      ## delle verosimiglianze (es: -0.0002)
      WpB=max(0,Wp_erB);WpB
    }
   # funzione che calcola la distribuzione bootstrap prepivot
   # è il bootstrap vincolato
   dist_boot=boot(y,testboot,sim="parametric",R=B,
            ran.gen = ran_poisson,
            ncpus = ncpus,parallel="multicore",mle=fit0)$t
   #####
    Co<<-Co+1;cat(Co,"/",Nsim,"\n");aoss=mean(dist_boot>Wp)
    return(list(aoss=aoss,Wp=Wp,max=max(dist_boot)))
  }
# funzione che simula i campioni da Poisson
  samples=replicate(Nsim,ran_poisson(data=M1$y,mle=fitted(M1)))
  result=apply(samples,2,MonteCarlo)
```

```
  result
}
sim=simulation(M1,covariate=ships2,Nsim = 10000,B=2000)
AossB=sapply(1:10000,function(x) sim[[x]]$aoss)
AossMis=sapply(1:10000,function(x) pchisq(sim[[x]]$Wp,1,lower=F)/2)
accBf=function(x) sapply(x,function(j) mean(AossB<j))
accMisf=function(x) sapply(x,function(j) mean(AossMis<j))
alpha=c(0.005,0.01,0.05,0.1);accB=accBf(alpha)
accMis=accMisf(alpha);errore=sqrt(alpha*(1-alpha)/10000)
```

Bibliografia

Agresti A. (2015). *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons.

Akaike H. (1973). Information theory and an extension of the maximum likelihood principle. *Proceedings of the 2nd International Symposium on Information,Czaki, Akademiai Kiado, Budapest*.

Azzalini A. (2001). *Inferenza Statistica: una Presentazione Basata sul Concetto di Verosimiglianza*. Springer Verlag.

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

Bartko J. J. (1962). A note on the negative binomial distribution. *Technometrics*, **4**, 609–610.

Bellio R.; Pierce D. A. (2016). *likelihoodAsy: Functions for Likelihood Asymptotics*. R package version 0.45.

Berk R.; MacDonald J. M. (2008). Overdispersion and Poisson regression. *Journal of Quantitative Criminology*, **24**, 269–284.

Blocker C.; Conway J.; Demortier L.; Heinrich J.; Junk T.; Lyons L.; Punzig G. (2006). Simple facts about p-values. *Technical internal note 8023, CDF Statistics Commitee, Rockefeller University*.

Brazzale A. R.; Davison A. C.; Reid N. (2007). *Applied asymptotics: Case Studies in Small-Sample Statistics*. Cambridge University Press.

Breslow N. E. (1984). Extra-Poisson variation in log-linear models. *Applied Statistics*, **33**, 38–44.

Cameron C.; Trivedi P. (2013). *Regression Analysis of Count Data*. Cambridge University Press.

Collings B. J.; Margolin B. H. (1985). Testing goodness of fit for the Poisson assumption when observations are not identically distributed. *Journal of the American Statistical Association*, **80**, 411–418.

Dean C.; Lawless J. F. (1989). Tests for detecting overdispersion in Poisson regression models. *Journal of the American Statistical Association*, **84**, 467–472.

Franck W. E. (1986). P-values for discrete test statistics. *Biometrical Journal*, **28**, 403–406.

Hilbe J. M. (2011). *Negative Binomial Regression*. Cambridge University Press.

Hinde J.; Demétrio C. (1997). Half-normal plots and overdispersion. *GLIM newsletter*, **27**, 19–26.

Hinde J.; Demétrio C. (1998). Overdispersion: models and estimation. *Computational Statistics & Data Analysis*, **27**, 151–170.

Lawless J. F. (1987). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*, **15**, 209–225.

Lindén A.; Mäntyniemi S. (2011). Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology*, **92**, 1414–1421.

Mackisack M. (1994). What is the use of experiments conducted by statistics students. *Journal of Statistics Education*, **2**, 1–15.

McCullagh P.; Nelder J. A. (1989). *Generalized Linear Models*. Chapman & Hall.

Milanzi E.; Alonso A.; Molenberghs G. (2012). Ignoring overdispersion in hierarchical loglinear models: Possible problems and solutions. *Statistics in medicine*, **31**, 1475–1482.

Moore D. F. (1986). Asymptotic properties of moment estimators for overdispersed counts and proportions. *Biometrika*, **73**, 583–588.

Nelder J. A.; Wedderburn R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135**, 370–384.

Pace L.; Salvan A. (2001). *Introduzione alla statistica*, volume II: Inferenza, Verosimiglianza, Modelli. Cedam.

Pierce D. A.; Bellio R. (2016). Modern likelihood-frequentist inference. *Sottoposto per pubblicazione*.

Sartori N.; Bellio R.; Kosmidis I.; Salvan A. (2016). Bootstrap prepivoting in the presence of many nuisance parameters. *Proceedings of the 48th SIS Scientific Meeting of the Italian Statistical Society, University of Salerno, June 8-10*.

Young G. A. (2009). Routes to higher-order accuracy in parametric inference. *Australian & New Zealand Journal of Statistics*, **51**, 115–126.