



UNIVERSITÀ DEGLI STUDI DI PADOVA
Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in Scienze Statistiche

Riduzione della distorsione in mediana
nella stima del parametro di dispersione

Relatore:

Ch.ma Prof.ssa
Alessandra Salvan

Laureanda:

Virginia Skerl
Matricola: 2160314

Indice

Introduzione	1
1 Modelli lineari generalizzati e parametro di dispersione	5
1.1 Inferenza statistica e verosimiglianza	5
1.1.1 Teoria dell'inferenza statistica	5
1.1.2 Modelli statistici parametrici e verosimiglianza	6
1.1.3 Presenza di parametri di disturbo e verosimiglianza profilo	8
1.1.4 Modelli statistici semi-parametrici e funzionali statistici	9
1.2 Famiglie di dispersione esponenziale	10
1.3 GLM: stima del parametro di dispersione	12
1.4 GLM: parametro di dispersione nel modello di quasi-verosimiglianza . . .	13
2 Riduzione della distorsione in mediana	17
2.1 Stimatori non distorti in media e in mediana	17
2.2 Metodi per la riduzione della distorsione in mediana	19
2.2.1 Riduzione della distorsione tramite funzione score modificata	20
2.2.2 Riduzione della distorsione tramite bootstrap	22
2.2.3 Riduzione analitica della distorsione tramite espansioni asintotiche	25
2.3 Applicazione ai GLM	28
2.3.1 Strategia di aggiustamento misto	28
2.3.2 Compromesso varianza-distorsione	29
2.4 Intervalli di confidenza	30
2.4.1 Intervalli di confidenza di Wald	30
2.4.2 Metodo HulC	31
3 Aspetti computazionali	33
3.1 Il metodo bootstrap: alcune considerazioni generali	33
3.2 Note nell'implementazione in R	36
3.2.1 Implementazione degli Algoritmi 2 e 3	36
3.2.2 Pacchetti e funzioni R	40

4	Applicazione a casi studio	43
4.1	Caso studio <code>medpar</code> : gamma	43
4.2	Caso studio <code>medpar</code> : quasi-Poisson	45
4.3	Caso studio <code>GasolineYield</code> : beta	47
5	Studi di simulazione	51
5.1	Misure di confronto empirico	51
5.2	Simulazioni nel caso parametrico: gamma	53
5.3	Confronto di metodi semi-parametrici: simulazione del modello binomiale negativo	57
5.4	Confronto di metodi semi-parametrici: simulazione del modello quasi-Poisson	60
	Conclusioni	61
	Appendice A	65
	Bibliografia	81

Introduzione

L'obiettivo dell'inferenza su componenti parametriche scalari di un modello statistico risiede nella stima puntuale, ma anche nell'associare a essa misure d'incertezza per valutare la significatività di effetti e costruire intervalli di confidenza. Di conseguenza, la presenza di distorsione, che sia in media o in mediana, causa l'invalidazione della procedura d'inferenza. Nel contesto dei modelli lineari generalizzati, l'interesse ricade spesso sui coefficienti di regressione, ma un ruolo altrettanto fondamentale è svolto dal parametro di dispersione: evidenze empiriche mostrano come la presenza di distorsione nella stima di tale parametro influenzi l'accuratezza delle distribuzioni utilizzate per test d'ipotesi e intervalli di confidenza sui coefficienti di regressione, rendendo per questi la copertura empirica non aderente al livello nominale.

Tipicamente, la stima del parametro di dispersione nei modelli lineari generalizzati e di quasi-verosimiglianza utilizza lo stimatore basato sul metodo dei momenti, mentre lo stimatore di massima verosimiglianza costituisce un'ulteriore opzione nel solo caso parametrico. Tuttavia, l'applicazione di tali metodi in campioni di dimensioni finite presenta difficoltà, legate alla presenza di distorsione, dovuta alla distribuzione campionaria asimmetrica, tipicamente a destra, del parametro di dispersione. A causa di questa asimmetria, la proprietà di non distorsione in mediana è preferibile rispetto a quella in media in quanto i valori estremi della coda destra influenzano la media campionaria dello stimatore. La non distorsione in mediana richiede che lo stimatore abbia uguale probabilità di sottostima e sovrastima del vero valore del parametro e dunque fornisce un'indicazione di centratura dello stimatore indipendentemente dalla forma della sua distribuzione campionaria.

La letteratura riguardante i metodi di riduzione della distorsione è ricca, a partire dall'approccio di correzione della funzione score di Firth (1993) per la distorsione in media e il corrispondente di Kenne Pagui et al. (2017) per la centratura in mediana.

In alternativa, Efron (1982) propone le procedure bootstrap per la correzione in media, mentre Cabrera & Watson (1997) introducono l'algoritmo per la riduzione della distorsione in mediana. Infine, Benussi et al. (2026) propongono un approccio analitico esplicito che sfrutta espansioni asintotiche per la riduzione della distorsione in mediana. Dal confronto tra le metodologie per la centratura in mediana emerge come, rispetto ai metodi basati sul bootstrap, le formulazioni di Kenne Pagui et al. (2017) e Benussi et al. (2026) richiedano un costo computazionale minimo. Tuttavia, esse sono soggette a condizioni di regolarità e differenziabilità del modello poiché si basano sul calcolo delle derivate, dei cumulanti e dei momenti di terzo ordine della log-verosimiglianza, al contrario della tecnica di Cabrera & Watson (1997). Infine, l'approccio di Benussi et al. (2026) è specificamente progettato per un parametro scalare, dunque qualora l'obiettivo inferenziale ricadesse sulla correzione in mediana simultanea di un vettore di parametri, l'approccio implicito di Kenne Pagui et al. (2017) risulta più appropriato.

L'obiettivo del presente lavoro consiste nella generalizzazione a contesti semi-parametrici del metodo proposto da Cabrera & Watson (1997). Nel caso in cui non sia disponibile una funzione di verosimiglianza per l'inferenza, poiché il modello è definito tramite equazioni di stima o relazioni sui momenti, come quelle di quasi-verosimiglianza, gli approcci analitici di correzione in mediana sopra menzionati risultano non applicabili. L'estensione dell'algoritmo all'utilizzo di bootstrap non parametrico permette di distaccarsi da vincoli di distribuzione ed assunzioni sui parametri. Inoltre, permette di raggiungere una maggiore robustezza rispetto a un'errata specificazione del modello e campioni di dimensioni ridotte. Per i temi trattati in precedenza, si è considerato il parametro di dispersione come parametro d'interesse, tuttavia gli algoritmi vengono presentati in una versione generalizzabile a diversi contesti.

Il Capitolo 1 introduce i concetti base dell'inferenza di verosimiglianza, dei modelli lineari generalizzati e della quasi-verosimiglianza, inoltre presenta i diversi metodi di stima del parametro di dispersione per i modelli statistici parametrici e semi-parametrici trattati. Il Capitolo 2 comprende la definizione di distorsione in mediana e gli approcci metodologici per la riduzione di questa presenti in letteratura. In aggiunta, si introduce un algoritmo di riduzione della distorsione in mediana per l'ambito semi-parametrico che sfrutta la procedura bootstrap non parametrica e una tecnica di calibrazione. Al fine di avere uno strumento con cui confrontare i risultati nei successivi capitoli, si introduce anche l'analogo semi-parametrico del metodo analitico esplicito con l'utilizzo dei cumulanti empirici. Infine, si approfondisce l'applicazione di tali metodi nei modelli

lineari generalizzati e si introducono gli intervalli di confidenza di Wald e con metodo HulC (Kuchibhotla et al., 2024). Infatti, un ulteriore aspetto per cui la riduzione della distorsione in mediana risulta d'interesse è la possibilità di costruzione di intervalli di confidenza tramite la procedura HulC, la quale porta a copertura empiricamente vicina a quella nominale in contesti in cui Wald fallisce, non richiede la stima della varianza e comporta un costo computazionale minimo. Il Capitolo 3 approfondisce gli aspetti computazionali del lavoro, prima focalizzandosi sul metodo bootstrap e successivamente trattando le componenti di implementazione degli algoritmi, tra cui i dettagli di calibrazione per i diversi parametri, una proposta di metodologia per la costruzione di intervalli di confidenza sulla base dell'algoritmo nel caso semi-parametrico e la correzione congiunta quando il parametro d'interesse è multidimensionale. Il Capitolo 4 include tre esempi di applicazione dei metodi presentati a dati reali, si impiegano i dataset `medpar` e `GasolineYield`. Data la natura di conteggio della variabile risposta del primo, si utilizzano due modelli per l'adattamento: quello parametrico gamma e quello semi-parametrico quasi-Poisson. Il secondo dataset è stato descritto tramite la regressione beta e si applicano le versioni dei metodi per la riduzione congiunta della distorsione in mediana per la stima di un vettore di parametri. Il Capitolo 5 riporta i risultati di un confronto empirico basato su studi di simulazioni, i quali si distinguono tra caso parametrico e semi-parametrico e seguono tre distinti schemi di simulazione. Infine, in Appendice si riporta il codice del programma **R** (versione 4.5.0).

Capitolo 1

Modelli lineari generalizzati e parametro di dispersione

Il presente capitolo raccoglie i concetti introduttivi per l'inferenza sul parametro di dispersione nell'ambito dei modelli lineari generalizzati (GLM). Inizialmente, si riassumono le nozioni di base della teoria dell'inferenza statistica, per poi presentare il contesto in cui si va a inserire il problema di inferenza su tale parametro. Infine, gli ultimi due paragrafi si propongono di riassumere gli approcci parametrico e semi-parametrico in rapporto alla modellazione e alla stima del parametro di dispersione, presentando i metodi più diffusi per la stima in entrambi.

I riferimenti principali per il paragrafo 1.1 sono Pace & Salvan (1996), mentre per la restante parte del capitolo Salvan et al. (2020). In aggiunta, il paragrafo 1.4 fa riferimento anche a McCullagh & Nelder (1989).

1.1 Inferenza statistica e verosimiglianza

1.1.1 Teoria dell'inferenza statistica

L'inferenza statistica si basa sull'osservazione di n unità statistiche, queste vanno a comporre un campione $y = (y_1, \dots, y_n)^\top$ che rappresenta una realizzazione del vettore casuale Y . Ciò può essere sintetizzato con la formulazione y realizzazione di $Y \sim p^0(y)$ con $y \in \mathcal{Y}$, dove $p^0(y)$ rappresenta una funzione di densità di probabilità non nota e $\mathcal{Y}$ lo spazio campionario.

La famiglia $\mathcal{F}$, detta modello statistico, rappresenta l'insieme di funzioni di densità compatibili con y . Le limitazioni su di essa identificano diversi livelli di specificazione, questi sono legati alla quantità di informazione ricavabile dai dati.

Il modello statistico parametrico rappresenta il livello con più informazione assunta, si presuppone che ciascun elemento che compone $\mathcal{F}$ sia identificabile da un parametro θ p -dimensionale:

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

Se $p^0(\cdot) \in \mathcal{F}$ si dice che il modello è correttamente specificato e si ha $p^0(y) = p(y; \theta_0)$ per il vero valore del parametro $\theta_0 \in \Theta$.

Il secondo livello di specificazione per estensione è quello semi-parametrico in cui gli elementi del modello statistico non sono individuabili interamente da una componente parametrica, in quanto è presente almeno una componente non indicizzabile da un numero finito di parametri.

Infine, nella specificazione non parametrica gli elementi del modello statistico $\mathcal{F}$ non sono individuabili tramite un numero finito di parametri reali.

1.1.2 Modelli statistici parametrici e verosimiglianza

Un modello parametrico $\mathcal{F}$ è specificato dalla formulazione $(\mathcal{Y}, p_Y(y; \theta), \Theta)$ con $\theta \subseteq \mathbb{R}^p$ con $p_Y(y; \theta)$ funzione di densità, o massa di probabilità nel caso discreto. La funzione di verosimiglianza, che raccoglie l'informazione su θ del modello statistico $\mathcal{F}$ e dal campione y , è definita come

$$L(\theta) = L(\theta; y) = c(y) p_Y(y; \theta),$$

dove $c(y) > 0$ è una costante di proporzionalità indipendente dal parametro e $p_Y(y; \theta)$ è la densità congiunta delle n osservazioni.

La funzione di log-verosimiglianza è ottenuta da quella di verosimiglianza come

$$l(\theta) = l(\theta; y) = \log L(\theta; y)$$

Nel caso in cui le osservazioni siano indipendenti la densità congiunta risulta calcolabile come prodotto delle singole densità, da cui

$$L(\theta) = c(y) \prod_{i=1}^n p_{Y_i}(y_i; \theta) \quad \text{e} \quad l(\theta) = \sum_{i=1}^n \log p_{Y_i}(y_i; \theta).$$

La funzione score (o di punteggio) è il vettore delle derivate parziali della funzione di log-verosimiglianza rispetto ai singoli parametri. Indicando con θ_r il generico elemento

r -esimo del vettore dei parametri θ , si ha

$$l_*(\theta) = l_*(\theta; y) = (l_1, \dots, l_p) \quad \text{con} \quad l_r = \frac{\partial l(\theta)}{\partial \theta_r}, \quad r = 1, \dots, p.$$

La stima di massima verosimiglianza $\hat{\theta}$ è il valore di $\theta \in \Theta$ che massimizza $L(\theta; y)$, ottenibile dalla soluzione dell'equazione di verosimiglianza

$$l_*(\theta) = 0.$$

Infine, la matrice d'informazione osservata ha come elementi le derivate parziali seconde di $l(\theta)$ cambiate di segno:

$$j(\theta) = j(\theta, y) = -l_{**}(\theta) = -\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^\top},$$

mentre si dice matrice d'informazione attesa (o di Fisher) il suo valore atteso

$$i(\theta) = E_\theta(j(\theta; Y)).$$

Qualora il vettore dei parametri sia partizionato in due componenti, $\theta = (\lambda, \psi)$, la matrice d'informazione attesa può essere espressa in forma a blocchi come:

$$i(\theta) = \begin{pmatrix} i_{\lambda\lambda}(\theta) & i_{\lambda\psi}(\theta) \\ i_{\psi\lambda}(\theta) & i_{\psi\psi}(\theta) \end{pmatrix}.$$

I due blocchi di parametri λ e ψ si dicono ortogonali nel senso di Fisher (Cox & Reid, 1987) se gli elementi al di fuori diagonale principale della matrice sono nulli, ossia se

$$i_{\lambda\psi}(\theta) = i_{\psi\lambda}(\theta)^\top = 0, \quad \forall \theta \in \Theta. \quad (1.1)$$

Sotto tale condizione, la matrice d'informazione attesa assume una struttura diagonale a blocchi, implicando l'indipendenza asintotica tra gli stimatori di massima verosimiglianza di λ e ψ .

In seguito, si utilizzeranno le notazioni $E_\theta(\cdot)$, $Var_\theta(\cdot)$, $Me_\theta(\cdot)$ e $\mathbb{P}_\theta(\cdot)$ per indicare, rispettivamente, il valore atteso, la varianza, la mediana e la probabilità calcolati in relazione a $p_Y(y; \theta)$.

Le identità di Bartlett stabiliscono delle relazioni tra i momenti della funzione score e le derivate successive della log-verosimiglianza, sotto condizioni di regolarità del modello. La prima identità dimostra che se il modello è correttamente specificato allora la funzione

score ha media nulla

$$E_\theta[l_*(\theta)] = 0.$$

La seconda identità di Bartlett lega la matrice di covarianza della score al valore atteso delle derivate seconde ossia la matrice d'informazione attesa

$$\text{Var}_\theta(l_*(\theta)) = E_\theta[l_*(\theta)l_*(\theta)^\top] = -E_\theta \left[\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^\top} \right] = i(\theta).$$

Adottando la notazione indiciale standard per i momenti delle derivate della log-verosimiglianza per il parametro θ (con indici $r, s, t, \dots$) allora si ha

$$\begin{aligned} \nu_{rs} &= E_\theta(l_{rs}(\theta)), \\ \nu_{r,s} &= E_\theta(l_r(\theta)l_s(\theta)), \\ \nu_{rst} &= E_\theta(l_{rst}(\theta)), \\ \nu_{r,st} &= E_\theta(l_r(\theta)l_{st}(\theta)), \\ \nu_{r,s,t} &= E_\theta(l_r(\theta)l_s(\theta)l_t(\theta)), \end{aligned} \tag{1.2}$$

dove $l_{rs}(\theta)$ e $l_{rst}(\theta)$ rappresentano le derivate d'ordine secondo e terzo di $l(\theta)$ rispetto a $\theta_r, \theta_s, \theta_t, \dots$. Inoltre, tramite tale notazione è possibile scrivere le identità di Bartlett come

$$\nu_r = 0 \quad \text{e} \quad \nu_{rs} + \nu_{r,s} = 0.$$

1.1.3 Presenza di parametri di disturbo e verosimiglianza profilo

Assumendo che il parametro θ sia tale che $\theta = (\psi, \lambda)$ dove ψ è il parametro d'interesse mentre λ la componente di disturbo, per l'inferenza su ψ si utilizza la funzione di verosimiglianza profilo

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

dove $\hat{\lambda}_\psi$ è la stima di massima verosimiglianza di λ con ψ fissato.

Le componenti identificano dei blocchi nelle quantità $l_*(\theta), j(\theta)$ e $i(\theta)$ nel seguente

modo

$$\begin{aligned} l_*(\theta) &= (l_\psi(\theta), l_\lambda(\theta)), \\ j(\theta) &= \begin{pmatrix} j_{\psi\psi} & j_{\psi\lambda} \\ j_{\lambda\psi} & j_{\lambda\lambda} \end{pmatrix}, \\ i(\theta) &= \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix}. \end{aligned}$$

Nel caso in cui ψ e λ siano ortogonali nel senso di Fisher, i blocchi fuori diagonale della matrice d'informazione attesa si annullano $i_{\psi\lambda} = i_{\lambda\psi}^\top = 0$. Tale proprietà geometrica implica l'indipendenza asintotica tra gli stimatori di ψ e λ , dunque la stima del parametro di disturbo non influisce asintoticamente sulla varianza dello stimatore del parametro d'interesse.

La funzione di log-verosimiglianza e la funzione di score profilo per il parametro d'interesse sono

$$l_P(\psi) = \log(L_P(\psi)) = l(\psi, \hat{\lambda}_\psi) \quad \text{e} \quad l_{*P}(\psi) = l_\psi(\psi, \hat{\lambda}_\psi).$$

Tuttavia, la verosimiglianza profilo non si comporta come una propria verosimiglianza in quanto la funzione score profilo, non avendo valore atteso nullo, non rispetta la prima identità di Bartlett.

1.1.4 Modelli statistici semi-parametrici e funzionali statistici

I funzionali statistici costituiscono uno strumento metodologico per formalizzare l'idea che la quantità che si desidera stimare debba possedere un'interpretazione coerente rispetto a qualsiasi modello di riferimento. Formalmente, un funzionale statistico $T(F)$ mappa lo spazio delle funzioni di ripartizione in uno spazio parametrico, permettendo di definire stime basate sulla funzione di ripartizione empirica $\hat{F}$ attraverso il cosiddetto metodo di sostituzione (*plug-in*), dove $T(\hat{F})$ rappresenta lo stimatore osservato (Pace & Salvan, 1997, paragrafo 3.6.3).

Nel dettaglio, dato un campione di osservazioni indipendenti e identicamente distribuite $X_1, \dots, X_n$, la funzione di ripartizione empirica $\hat{F}(x)$ è definita come la proporzione di dati campionari minori o uguali a un determinato valore x , ovvero:

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}(X_i \leq x).$$

Al variare di $T(\cdot)$ si distingue tra i funzionali statistici lineari e non lineari. Un funzionale $T(F)$ si definisce lineare se soddisfa la proprietà di linearità nello spazio delle funzioni di ripartizione, ossia se esprimibile come il valore atteso di una trasformazione $f(\cdot)$ della variabile casuale

$$T(F) = \int f(x) dF(x) = E_F(f(X_i)),$$

con controparte empirica $T(\hat{F}) = \frac{1}{n} \sum_{i=1}^n f(x_i)$. Nel caso in cui $f(x_i) = x_i$ si ottiene lo stimatore della media campionaria.

Al contrario, i funzionali non lineari non possono essere espressi come valore atteso di una funzione di X_i . La mediana campionaria è un esempio di funzionale statistico non lineare definito come $T(F) = F^{-1}(1/2)$ da cui $T(\hat{F}) = \hat{F}^{-1}(1/2)$.

Il concetto di non distorsione verrà approfondito nel prossimo capitolo, tuttavia in questo contesto per equazione di stima non distorta per i parametri θ si intende una funzione $q(\theta; X)$ tale che $E_\theta(q(\theta, X)) = 0, \forall \theta \in \Theta$. Sotto appropriate condizioni di regolarità, l'equazione di stima $E_{\hat{F}}\{q(\theta; X_i)\} = 0$ definisce implicitamente un funzionale statistico $\theta = T(F)$ come la soluzione rispetto a θ dell'equazione:

$$\int q(\theta; x) dF(x) = 0.$$

Nel contesto dei modelli semi-parametrici, dove non si assume una forma parametrica completa per la distribuzione, l'uso dei funzionali permette di definire stimatori $T(\hat{F})$ che rimangono interpretabili e consistenti anche quando il modello di riferimento è parzialmente specificato.

1.2 Famiglie di dispersione esponenziale

I modelli lineari generalizzati rappresentano un'estensione dei modelli di regressione lineare normale a modelli con distribuzione della risposta diversa dalla normale e introducono la funzione di legame che lega il valore atteso della risposta al predittore lineare.

Siano $Y_1, \dots, Y_n$ variabili casuali indipendenti. Si assume che la distribuzione della risposta Y_i sia parte della classe della famiglia di dispersione esponenziale, un'ampia classe a cui appartengono alcuni esempi notevoli di distribuzioni sia univariate che multivariate. In particolare, $Y_i, i = 1, \dots, n$, appartiene alla classe di famiglie di dispersione

esponenziale univariate se ha densità esprimibile nella forma

$$p(y_i; \theta_i, \phi) = \exp \left\{ \frac{\theta_i y_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\}, \quad (1.3)$$

dove $y_i \in S \subseteq \mathbb{R}$ con S supporto di Y_i , $\theta_i \in \Theta \subseteq \mathbb{R}$ con Θ spazio parametrico del parametro θ_i detto naturale. In aggiunta, sia $a_i(\phi) > 0$ con ϕ parametro di dispersione. Di frequente, $a_i(\phi)$ è di forma $a_i(\phi) = 1$, $a_i(\phi) = \phi$ o $a_i(\phi) = \phi/\omega_i$, con $\omega_i, i = 1, \dots, n$, pesi noti e $\phi > 0$. Da tale espressione generale, specificando le funzioni $a_i(\cdot)$, $b_i(\cdot)$ e $c(\cdot)$, si ottengono diversi modelli parametrici.

Sia X una matrice di dimensione $n \times p$, composta dai vettori riga p -dimensionali delle esplicative per ciascuna osservazione $\mathbf{x}_i$, e il vettore colonna p -dimensionale β dei coefficienti di regressione. Allora il predittore lineare è definito come $\eta = X\beta$ con componenti $\eta_i = \mathbf{x}_i \beta$. Infine, sia μ_i il valore atteso della variabile risposta Y_i , la funzione legame è la funzione $g(\cdot)$ nota, derivabile e monotona crescente tale che

$$g(\mu_i) = \eta_i.$$

La funzione $g(\cdot)$ permette il collegamento tra valore atteso e predittore lineare. Tuttavia, è necessario che risolva il problema di coerenza in cui si incorre quando lo spazio delle medie M è sottoinsieme di $\mathbb{R}$, ossia sia tale che $g: M \rightarrow \mathbb{R}$.

La notazione compatta della classe delle famiglie di dispersione esponenziale univariate è

$$Y_i \sim DE_1(\mu_i, a_i(\phi) v(\mu_i)), \quad \text{con } \mu_i = g^{-1}(\eta_i) = g^{-1}(\mathbf{x}_i \beta).$$

Se Y_i ha densità (1.3) allora risulta che la funzione di log-verosimiglianza basata su un campione y_i sia

$$l(\theta_i, \phi) = l(\theta_i, \phi; y_i) = \frac{\theta_i y_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi). \quad (1.4)$$

Da questa sono ricavate le formulazioni per il valore atteso e la varianza

$$E_{\theta_i, \phi}(Y_i) = \mu_i = b'(\theta_i) \text{ e } Var_{\theta_i, \phi}(Y_i) = a_i(\phi) b''(\theta_i).$$

Inoltre, tramite la riparametrizzazione (μ_i, ϕ) definita da $\mu_i = \mu(\theta_i)$, con inversa $\theta_i = \theta(\mu_i)$, si ha

$$v(\mu_i) = b''(\theta_i) \Big|_{\theta_i = \theta(\mu_i)}$$

tale che

$$Var_{\mu_i, \phi}(Y_i) = a_i(\phi) v(\mu_i),$$

dove la funzione $v(\mu_i)$ è detta funzione di varianza. Entro la classe delle famiglie di dispersione esponenziale univariate, la funzione di varianza caratterizza, assieme a M , il modello parametrico specifico (Morris, 1982).

Una distinzione importante da evidenziare in relazione al parametro ϕ è la differenza tra modelli in cui esso è noto, come Poisson o binomiale dove $\phi = 1$, e modelli in cui deve essere stimato, come la Normale, Gamma o contesti di presenza di sovradisersione. Nel secondo contesto, l'errata stima di ϕ non influenza la non distorsione dell'equazione di stima per i coefficienti di regressione β , ma influenza la stima dei loro standard error, invalidando intervalli di confidenza e test d'ipotesi.

In questo contesto per equazione di stima non distorta per i parametri θ si intende una funzione $l_*(\theta) = l_*(\theta; Y)$ tale che $E_\theta(l_*(\theta)) = 0, \forall \theta \in \Theta$.

1.3 GLM: stima del parametro di dispersione

Dalla notazione compatta si comprende che i parametri d'interesse sono (β, ϕ) e il modello da loro identificato è un modello statistico parametrico.

Siano $Y_i, i = 1, \dots, n$, appartenenti alla famiglia DE_1 componenti di Y allora la funzione di log-verosimiglianza è data dalla sommatoria delle funzioni marginali (1.4)

$$l(\beta, \phi) = \sum_{i=1}^n \frac{\theta_i y_i - b(\theta_i)}{a_i(\phi)} + \sum_{i=1}^n c(y_i, \phi) \quad \text{con } \theta_i = \theta(\mu_i),$$

da cui

$$l(\beta, \phi) = \sum_{i=1}^n \frac{y_i \mathbf{x}_i \beta - b(\mathbf{x}_i \beta)}{a_i(\phi)} + \sum_{i=1}^n c(y_i, \phi).$$

La funzione punteggio $l_*(\beta, \phi; y)$ ha componenti

$$l_r = \frac{\partial l(\beta, \phi)}{\partial \beta_r} = \sum_{i=1}^n \frac{y_i - \mu_i}{a_i(\phi)} \frac{\partial \theta_i}{\partial \beta_r}, \quad r = 1, \dots, p,$$

$$l_\phi = \frac{\partial l(\beta, \phi)}{\partial \phi} = - \sum_{i=1}^n \frac{a'_i(\phi)}{(a_i(\phi))^2} (y_i \theta_i - b(\theta_i)) + \sum_{i=1}^n c'(y_i, \phi).$$

La stima di massima verosimiglianza per il parametro di dispersione si basa su quest'ultima sostituendo β con la sua stima di massima verosimiglianza $\hat{\beta}$.

Tuttavia, la stima $\hat{\phi}$ riscontra problematiche di robustezza e instabilità. Dunque un'alternativa consistente è la stima basata sul metodo dei momenti. Assumendo

$a_i(\phi) = \phi$, lo stimatore basato sul metodo dei momenti è

$$\tilde{\phi} = \frac{1}{n-p} \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}, \quad \text{con } \hat{\mu}_i = g^{-1}(\mathbf{x}_i \hat{\beta}).$$

La stima così ottenuta risulta essere consistente sotto opportune condizioni, si veda ad esempio Fahrmeir & Tutz (2001, paragrafo 2.2.1).

Utilizzando la verosimiglianza profilo la stima del parametro d'interesse tende a essere non consistente o poco accurata se il numero di parametri di disturbo cresce con la numerosità campionaria. Questo è il caso dei GLM quando ϕ è il parametro d'interesse e i parametri della media agiscono da parametri di disturbo rispetto al parametro di dispersione.

1.4 GLM: parametro di dispersione nel modello di quasi-verosimiglianza

Il modello statistico semi-parametrico di quasi-verosimiglianza si basa sulle ipotesi del secondo ordine dei modelli lineari generalizzati

$$\begin{aligned} E(Y_i) &= g^{-1}(\mathbf{x}_i \beta), \\ \text{Var}(Y_i) &= \phi v(\mu_i) \quad \text{dove } \phi > 0 \text{ ignoto,} \\ Y_i \text{ e } Y_j &\text{ indipendenti se } i \neq j. \end{aligned}$$

Esso è adatto a trattare risposte Y_i , $i = 1, \dots, n$, sia continue che discrete, in particolare per risposte binomiali e Poisson permette un incremento di flessibilità rispetto alla specificazione parametrica dei modelli lineari generalizzati e per tale motivo una comune applicazione è in presenza di sovradisersione.

La stima di β secondo il modello di quasi-verosimiglianza avviene sulla base di equazioni di stima non distorte, ponendo $U(\beta) = 0$, dove

$$U(\beta) = D^\top V^{-1} \frac{Y - \mu}{\phi}$$

dove D è una matrice $n \times p$ con elementi $D_{ir} = \partial \mu_i / \partial \beta_r$ e $V = \text{diag}(v(\mu_1), \dots, v(\mu_n))$. Analogamente all'approccio parametrico, la funzione $U(\beta)$ è detta funzione di quasi-score (o quasi-punteggio).

Anche in questo contesto, la stima di ϕ si basa sul metodo dei momenti, questa infatti rimane consistente anche sotto le ipotesi del modello di quasi-verosimiglianza (Liang & Zeger, 1995, paragrafo 2.3).

Tuttavia, tale stima risente della variabilità introdotta dalla stima dei coefficienti di regressione in campioni di ampiezza ridotta. Ciò può causare una varianza campionaria elevata e instabilità numerica. Farrington (1996) propone una modifica dello stimatore del metodo dei momenti col fine di ridurre la varianza asintotica di $\tilde{\phi}$ in presenza di dati sparsi o campioni limitati. Nello specifico, l'Autore inquadra il problema affiancando alle equazioni di quasi-verosimiglianza per β un'equazione di stima supplementare specifica per ϕ

$$\sum_{i=1}^n a_i(y_i - \mu_i) + \sum_{i=1}^n \left\{ \frac{(y_i - \mu_i)^2}{v(\mu_i)} - \phi \right\} = 0,$$

dove i coefficienti $a_i = a_i(\mu_i)$ rappresentano un termine di correzione del primo ordine, la scelta di $a_i = -\frac{\partial v(\mu_i)}{v(\mu_i) \partial \mu_i}$ garantisce l'ortogonalità locale tra il parametro di dispersione ϕ e i parametri di regressione β . Tale stimatore può essere riscritto in funzione dello stimatore originale secondo la formulazione

$$\tilde{\phi}_{Fa} = \tilde{\phi} - \frac{n\bar{s}}{n-p}, \quad \text{dove} \quad \bar{s} = \frac{1}{n} \sum_{i=1}^n \frac{\partial v(\hat{\mu}_i)}{\partial \hat{\mu}_i} \frac{1}{v(\hat{\mu}_i)} (y_i - \hat{\mu}_i). \quad (1.5)$$

Sebbene inizialmente proposto per verificare la bontà d'adattamento di un modello GLM, dunque testare l'ipotesi nulla che $\phi = 1$, esso può essere utilizzato come stimatore di ϕ sotto la condizione che il terzo cumulante della risposta $\kappa_{3i} = E((Y_i - \mu_i)^3)$ sia

$$\kappa_{3i} = \phi^2 v(\mu_i) \frac{\partial v(\mu_i)}{\partial \mu_i},$$

ossia che la relazione tra i primi tre cumulanti sia nella forma della famiglia *DE*.

In alternativa, Fletcher (2012) propone un'ulteriore estensione dei metodi basati sui momenti il cui fine è la riduzione della varianza. Sotto il modello di quasi-verosimiglianza, lo stimatore assume la forma

$$\tilde{\phi}_{Fl} = \frac{\tilde{\phi}}{1 + \bar{s}},$$

dove $\bar{s}$ ha forma (1.5). Si noti che nel caso di sola intercetta, il valore di $\bar{s}$ sarà nullo e dunque i tre stimatori saranno equivalenti. La stima così ottenuta gode di varianza ridotta rispetto a quella di $\tilde{\phi}$ se una condizione sul terzo momento, tipicamente soddisfatta nelle strutture di quasi-verosimiglianza notevoli quali la quasi-Poisson, risulta rispettata

$$\kappa_{3i} = \alpha v(\mu_i) \frac{\partial v(\mu_i)}{\partial \mu_i}, \quad \text{con } \alpha \geq \phi^2.$$

Sotto tale condizione lo stimatore $\tilde{\phi}_{Fl}$ ha varianza asintotica inferiore anche a quello di

Farrington, sotto condizioni più deboli. Rispetto allo stimatore basato sul metodo dei momenti, la formulazione introduce una modifica di contrazione attraverso un fattore moltiplicativo. Nel caso di dati sparsi tale fattore aumenta e la contrazione risulta più intensa.

In aggiunta alle correzioni basate sui momenti, sono presenti in letteratura approcci alternativi per la stima di ϕ in modelli per campioni caratterizzati da non continuità o asimmetria. Lo stimatore basato sulla devianza residua (Nelder & Pregibon, 1987, paragrafo 2) è definito come

$$\tilde{\phi}_D = \frac{D(Y; \hat{\mu})}{n - p},$$

dove $D(Y; \hat{\mu})$ indica la devianza totale del modello

$$D(Y; \hat{\mu}) = \sum_{i=1}^n d_i(y_i; \hat{\mu}_i) = 2 \sum_{i=1}^n \int_{\hat{\mu}_i}^{y_i} \frac{y_i - u}{v(u)} du.$$

A differenza degli stimatori basati sui momenti, le proprietà asintotiche di $\tilde{\phi}_D$ dipendono dalla distribuzione della risposta. Nel contesto di distribuzioni continue $\tilde{\phi}_D$ risulta asintoticamente non distorto e invariante rispetto a trasformazioni monotone della variabile risposta. Tuttavia, in presenza di dati discreti e sparsi la devianza residua non converge alla distribuzione asintotica χ_{n-p}^2 . In tali contesti, $\tilde{\phi}_D$ presenta una forte distorsione sistematica in media che persiste anche all'aumentare della numerosità campionaria n dunque $\tilde{\phi}_D$ risulta inaffidabile per campioni caratterizzati da elevata sparsità, rendendo necessario il ricorso a correzioni del primo ordine basate sui momenti oppure ad approcci di riduzione della distorsione.

Uno stimatore più robusto è quello basato sul MAD (*Median Absolute Deviation*) dei residui di Pearson r_i (Maronna et al., 2019, p. 36). Esso assume la forma

$$\tilde{\phi}_{MAD} = \left(\frac{\text{Me}_i |r_i - \text{Me}_j(r_j)|}{0.6745} \right)^2, \text{ dove } r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)}}$$

e dove l'operatore $\text{Me}_i(\cdot)$ indica la mediana calcolata sull'insieme indicizzato con i . La costante al denominatore, pari al terzo quartile della distribuzione normale standard, risulta necessario per correggere la consistenza e riportare la stima di ϕ sulla stessa grandezza dello standard error per la normale (σ). Successivamente, poiché nei modelli di quasi-verosimiglianza il parametro di dispersione ϕ esprime una variabilità su scala quadratica (pari a σ^2), l'espressione viene elevata al quadrato. Sostituendo gli operatori di media e quadrato dello stimatore del metodo dei momenti con la mediana degli scarti assoluti, tale stimatore permette di correggere l'effetto di salti discreti del campione o

valori estremi strutturali.

Il modello di quasi-verosimiglianza assume $Var(Y_i) = \phi v(\mu_i)$. Nel caso in cui $a_i(\phi) = \phi$ non sia valida per tutte le osservazioni, Nelder & Pregibon (1987) suggeriscono una generalizzazione denominata quasi-verosimiglianza estesa (*extended quasi-likelihood*) che permette la stima congiunta di (β, ϕ) .

Infine, la quasi-verosimiglianza è un esempio di approccio semi-parametrico dove non si specifica l'intera distribuzione, ma solo un'ipotesi sulla media, la relazione tra media e varianza e l'indipendenza tra osservazioni. Dunque, lo stimatore di quasi-verosimiglianza può essere interpretato come l'applicazione di un funzionale statistico.

Capitolo 2

Riduzione della distorsione in mediana

Il presente capitolo si pone l'obiettivo di introdurre il concetto di riduzione della distorsione in mediana per un generico parametro d'interesse. Inizialmente si definiscono gli stimatori non distorti in mediana. Successivamente, si affrontano tre metodologie per la riduzione della distorsione in mediana: la prima si basa sulla modifica della funzione score, dapprima per un generico parametro scalare, poi per un parametro vettoriale con una componente d'interesse scalare e infine per un parametro vettoriale. Il secondo metodo proposto è incentrato su una procedura di ricampionamento, di conseguenza generalizzabile al contesto semi-parametrico. Mentre, la terza metodologia si basa su espansioni asintotiche ed è anch'essa applicabile a scenari non parametrici. In conclusione, vengono evidenziate le ragioni per cui la riduzione della distorsione in mediana sia di particolare interesse nell'inferenza sul parametro di dispersione all'interno dei modelli lineari generalizzati e si richiamano le tecniche di inferenza intervallare.

I principali riferimenti bibliografici sono Kenne Pagui et al. (2017) per i paragrafi 2.1 e 2.2, per quest'ultimo si fa riferimento anche a Cabrera & Watson (1997) e Benussi et al. (2026). Infine, per il paragrafo 2.3, McCullagh & Nelder (1989) e Kosmidis et al. (2020), mentre per quello 2.4 Kuchibhotla et al. (2024).

2.1 Stimatori non distorti in media e in mediana

Sia $\hat{\theta}_n$ stimatore di un parametro scalare il cui valore vero è θ_0 , come introdotto nel paragrafo 1.1. Si dice che lo stimatore è non distorto in media se

$$E_{\theta_0}(\hat{\theta}_n(Y)) = \theta_0,$$

ossia lo stimatore è centrato in media sul vero valore del parametro, per ogni $\theta_0 \in \Theta$.

In maniera analoga, si dice che lo stimatore è non distorto in mediana se

$$Me_{\theta_0}(\hat{\theta}_n(Y)) = \theta_0,$$

ossia lo stimatore sovrastima e sottostima il vero valore del parametro con la stessa probabilità:

$$\mathbb{P}_{\theta_0}(\hat{\theta}_n \leq \theta_0) \geq \frac{1}{2} \quad \text{e} \quad \mathbb{P}_{\theta_0}(\hat{\theta}_n \geq \theta_0) \geq \frac{1}{2}. \quad (2.1)$$

Nel caso in cui ciò non sia verificato, è possibile ottenere un nuovo stimatore $\tilde{\theta}$ non distorto in mediana. Infatti, se la funzione $h_n(\theta)$ è monotona, la soluzione rispetto a θ_0 dell'equazione

$$Me_{\theta_0}(\hat{\theta}_n) = h_n(\theta_0) = \hat{\theta} \quad (2.2)$$

fornisce uno stimatore $\tilde{\theta}$ non distorto in mediana (Cabrera & Watson, 1997, Lemma 2.1).

I concetti di non distorsione in media e in mediana rappresentano due criteri distinti per valutare la centratura di uno stimatore.

La loro equivalenza, esplicita per distribuzioni di $\hat{\theta}_n$ simmetriche, si verifica anche in ambito asintotico. Infatti, sotto opportune condizioni di regolarità, all'aumentare della dimensione campionaria la distribuzione tende alla normalità per il teorema del limite centrale, dunque le discrepanze dovute all'asimmetria campionaria svaniscono al crescere di n , annullando la differenza tra i due approcci.

Dunque, quando la distribuzione di $\hat{\theta}_n$ è asimmetrica i due concetti si separano nettamente. In tal caso, a differenza della media, la non distorsione in mediana risulta più robusta rispetto a valori estremi o code pesanti della distribuzione in quanto fissa l'uguaglianza della probabilità di errore tra sovrastima e sottostima.

Nello studio delle proprietà degli stimatori, la riduzione della distorsione in mediana porta a vantaggi teorici ed empirici rispetto alla riduzione in media. Un primo aspetto è l'estensione della proprietà di equivarianza, citata in precedenza, per cui la non distorsione in mediana si preserva sotto qualsiasi trasformazione monotona dello stimatore. Ciò risolve le criticità legate al dominio dei parametri, in cui la correzione in media potrebbe generare stime non ammissibili o esterne allo spazio parametrico.

In aggiunta, basando la costruzione dello stimatore sull'equiprobabilità di sovrastima e sottostima, l'approccio in mediana permette di unificare nell'unica struttura formale delle curve di confidenza (*confidence curves*) la stima puntuale non distorta in mediana, i limiti e gli intervalli di confidenza (Birnbaum, 1961). Sotto opportune condizioni di

regolarità e in presenza di famiglie con rapporto di verosimiglianza monotono, la non distorsione in mediana consente di identificare stimatori ammissibili e uniformemente migliori, integrando la logica della massima verosimiglianza con la teoria dei test di Neyman-Pearson (Lehmann & Romano, 2005).

Inoltre, a livello inferenziale la disponibilità di stimatori non distorti in mediana permette l'utilizzo di metodologie come la procedura HulC (Kuchibhotla et al., 2024, *Hull-based Confidence*). Poiché la validità teorica e la copertura dei limiti d'intervallo HulC dipendono direttamente dal controllo della distorsione in mediana, la combinazione di tali stimatori con intervalli di confidenza di tipo HulC consente di raggiungere una copertura empirica molto vicina alla nominale, evitando la stima esplicita della varianza o della distribuzione limite (si veda il paragrafo 2.4.2).

Infine, nell'ambito della costruzione di intervalli di tipo Wald, sostituire lo stimatore di massima verosimiglianza con lo stimatore non distorto in mediana porta in genere a una copertura empirica degli intervalli di confidenza in campioni finiti più simile al livello nominale. Al contrario, gli intervalli Wald basati su stimatori non distorti in media o di massima verosimiglianza causano spesso fenomeni di sottocopertura.

2.2 Metodi per la riduzione della distorsione in mediana

I metodi in seguito presentati hanno l'obiettivo della riduzione della distorsione in mediana. Tuttavia seguono due metodologie diverse. I metodi impliciti riducono la distorsione correggendo la funzione obiettivo prima della stima, mentre i metodi espliciti intervengono a posteriori.

L'approccio di riduzione della distorsione basato sulla funzione score modificata rappresenta un metodo implicito sia in media che in mediana. Al contrario, il metodo basato sulla tecnica bootstrap è un esempio di metodo esplicito per la correzione in media, mentre in mediana è implicito e numerico. Un metodo alternativo esplicito è quello che si basa su correzioni analitiche a posteriori, presentato qui solo per la non distorsione in mediana.

Gli stimatori ottenuti da metodi impliciti godono di proprietà di contrazione (*shrinkage*) e sono robusti rispetto a stime sulla frontiera dello spazio parametrico, garantendo l'esistenza di stime finite anche in presenza di problemi di separazione dei dati nei modelli binomiali e multinomiali. Al contrario, i metodi espliciti apportano una correzione allo

stimatore di massima verosimiglianza e dunque, se lo stimatore di partenza è infinito, essi falliscono (si veda ad esempio Kosmidis, 2014b).

Tuttavia, il metodo della funzione score modificata richiede calcoli analitici complessi in quanto richiede noti e in forma chiusa la funzione score, l'informazione di Fisher e i termini di distorsione del primo ordine. Di conseguenza, risulta di difficile generalizzazione a modelli non standard. Viceversa, per quanto il bootstrap risulti intensivo dal punto di vista computazionale, è automatizzabile e applicabile anche quando le derivate della verosimiglianza non sono gestibili (per la correzione in media si veda Kosmidis, 2014a).

2.2.1 Riduzione della distorsione tramite funzione score modificata

I metodi impliciti per la riduzione della distorsione in media dello stimatore di massima verosimiglianza si basano sulla correzione del primo ordine della funzione score $l_*(\theta)$ (Firth, 1993). Nel caso di un parametro scalare, la funzione score modificata corrisponde a

$$l_*^F(\theta) = l_*(\theta) + (\nu_{\theta,\theta,\theta} + \nu_{\theta,\theta\theta})/(2i(\theta)),$$

dove $\nu_{\theta,\theta,\theta} = E_\theta(l_*(\theta)^3)$ e $\nu_{\theta,\theta\theta} = E_\theta(l_*(\theta)l_{**}(\theta))$, ossia i corrispondenti dei termini generali $\nu_{r,s,t}$ e $\nu_{r,st}$ (1.2) nel caso di un parametro di interesse scalare θ . Nel caso di una famiglia esponenziale monoparametrica, tale correzione risulta equivalente all'aggiunta di una penalizzazione alla verosimiglianza basata sulla a priori di Jeffreys.

Il corrispondente metodo di correzione della distorsione in mediana, proposto da Kenne Pagui et al. (2017), prevede la sottrazione dallo score, anziché della media, della sua mediana approssimata tramite sviluppo di Cornish-Fisher:

$$\tilde{l}_*(\theta) = l_*(\theta) + \frac{\nu_{\theta,\theta,\theta}}{6i(\theta)}.$$

Per le famiglie esponenziali ciò equivale a porre una penalizzazione della log-verosimiglianza di $1/6 \log(i(\theta))$.

A differenza dello stimatore ottenuto dalla correzione per la distribuzione in media ($\hat{\theta}^*$), lo score modificato in mediana porta ad uno stimatore $\tilde{\theta}^*$ equivariante rispetto riparametrizzazioni monotone.

Dato un parametro vettoriale $\theta = (\psi, \lambda^\top)^\top$, dove ψ è il parametro scalare d'interesse, gli Autori introducono una correzione della funzione score profilo. In questo contesto,

la funzione score profilo modificata per la non distorsione in mediana è

$$\tilde{l}_{*P}(\psi) = l_{*P}(\psi) - \kappa_{1\psi} + \frac{\kappa_{3\psi}}{6\kappa_{2\psi}}. \quad (2.3)$$

Indicizzando la componente di disturbo λ con $a, b, c, \dots$, le quantità sono ottenibili secondo le seguenti espressioni

$$\begin{aligned} \kappa_{1\psi} &= -\frac{1}{2} \nu^{ab} ((\nu_{\psi,ab} - \gamma_{\psi c} \nu_{c,ab}) + (\nu_{\psi,a,b} - \gamma_{\psi c} \nu_{a,b,c})), \\ \kappa_{2\psi} &= \nu_{\psi,\psi} - \gamma_{\psi a} \nu_{\psi,a}, \\ \kappa_{3\psi} &= \nu_{\psi,\psi,\psi} - 3\gamma_{\psi a} \nu_{\psi,\psi,a} + 3\gamma_{\psi a} \gamma_{\psi b} \nu_{\psi,a,b} - \gamma_{\psi a} \gamma_{\psi b} \gamma_{\psi c} \nu_{a,b,c}, \end{aligned}$$

dove ν^{ab} è l'elemento dell'inversa della matrice di elementi $\nu_{a,b}$ e, adottando la convenzione di somma di Einstein, $\gamma_{\psi a} = \nu^{ab} \nu_{\psi,b} = \sum_b \nu^{ab} \nu_{\psi,b}$.

Il termine $\kappa_{1\psi}$ rappresenta la distorsione del primo ordine della score profilo, mentre $\kappa_{2\psi}$ e $\kappa_{3\psi}$ sono le approssimazioni del secondo e terzo cumulante della score efficiente per ψ

$$\bar{l}_{\psi} = l_{\psi} - \gamma_{\psi a} l_a.$$

Risolvendo $\tilde{l}_{*P}(\psi) = 0$ con λ sostituito da $\hat{\lambda}_{\psi}$, lo stimatore ottenuto gode della proprietà di non distorsione in mediana. Di conseguenza, le relazioni (2.1) valgono in senso approssimato, con un errore di ordine $O(n^{-3/2})$ nel caso continuo, dove $\mathbb{P}_{\theta}$ indica la probabilità sotto θ con ψ fissato e λ qualunque.

In aggiunta, lo stimatore $\tilde{\psi}_P^*$ è equivariante rispetto trasformazioni monotone di ψ che non dipendono da λ .

Nel contesto in cui il parametro d'interesse sia θ , vettore p -dimensionale, non risulta definibile una correzione globale al vettore score poiché non esiste una definizione univoca di mediana multivariata. Dunque, Kenne Pagui et al. (2017) propongono un approccio componente per componente (*componentwise*) dove viene applicata la correzione per ciascuna componente del vettore, trattando le restanti come parametri di disturbo.

Il sistema di funzioni di score modificate è

$$\tilde{l}_r(\theta) = 0, \quad r = 1, \dots, p,$$

dove a ciascuna equazione corrisponde una formulazione della score profilo modificata per l' r -esima componente come (2.3). Lo stimatore congiunto $\tilde{\theta}^* = (\tilde{\theta}_1^*, \dots, \tilde{\theta}_p^*)^T$ garantisce che ogni singola componente sia individualmente approssimativamente non distorta in mediana. Inoltre, l'approccio preserva l'invarianza rispetto a trasformazioni

monotone applicate separatamente alle singole componenti del vettore.

2.2.2 Riduzione della distorsione tramite bootstrap

L'utilizzo di tecniche bootstrap per la riduzione della distorsione in media è stato introdotto da Efron (1982). Tale approccio è stato generalizzato alla distorsione in mediana da Cabrera & Watson (1997) per il caso parametrico, mentre l'estensione al bootstrap non parametrico costituisce una proposta originale del presente lavoro.

Nel bootstrap classico (Efron, 1982, paragrafo 5.5), l'obiettivo è stimare la distorsione intesa come valore atteso dello stimatore meno il vero valore del parametro. Dunque l'algoritmo segue i seguenti passaggi.

Algoritmo 1 Algoritmo di riduzione della distorsione in media

Input: Numero di replicazioni bootstrap R

Output: Stima con correzione per la distorsione in media $\hat{\theta}^{**}$

- (i) Calcolare la stima distorta $\hat{\theta}$ sul campione originale
 - (ii) Generare R campioni bootstrap
 - (iii) **for** $r = 1, \dots, R$ **do**
 - └ Calcolare la stima bootstrap $\hat{\theta}_r^{**}$
 - (iv) Approssimare il valore atteso della distribuzione bootstrap con la media campionaria delle repliche: $\bar{\theta}^{**} = \frac{1}{R} \sum_{r=1}^R \hat{\theta}_r^{**}$
 - (v) Ottenere la stima bootstrap della distorsione come $\bar{\theta}^{**} - \hat{\theta}$ e la stima corretta per la distorsione in media $\hat{\theta}^{**} = 2\hat{\theta} - \bar{\theta}^{**}$.
-

Il valore di R suggerito è compreso tra 100 e 400 al fine di ottenere una stima stabile. Al passo (iii) si distingue tra bootstrap parametrico e non parametrico: nel primo si simulano i R campioni dal modello sottostante con θ stimato, mentre nel secondo i R campioni si ottengono ricampionando dal campione originale.

Anche per il metodo di riduzione in mediana si effettua tale distinzione: se lo stimatore distorto in mediana è basato su un modello parametrico, allora le simulazioni utilizzeranno tale modello stimato, in caso contrario, dove si fa riferimento a un modello semi-parametrico, è necessario l'utilizzo di bootstrap non parametrico.

L'algoritmo introdotto da Cabrera & Watson (1997) per la riduzione della distorsione in mediana si focalizza sul primo caso, basandosi sulla risoluzione di (2.2) secondo i seguenti passi.

Algoritmo 2 Algoritmo di riduzione della distorsione in mediana:
caso parametrico

Input: Intervallo (θ_L, θ_U) , numero di punti N , numero di simulazioni m

Output: Stima non distorta in mediana $\tilde{\theta}^{**}$

- (i) Definire l'intervallo (θ_L, θ_U) contenente $\tilde{\theta}^{**}$
 - (ii) Selezionare N punti $(\theta_1, \dots, \theta_N)$ in (θ_L, θ_U)
 - (iii) **for** $i = 1, \dots, N$ **do**
 - Simulare il modello m volte
 - Ottenere i valori dello stimatore $\hat{\theta}_1, \dots, \hat{\theta}_m$
 - Calcolare la stima di $h_n(\theta_i)$ come $h_n^*(\theta_i) = Me(\hat{\theta}_1, \dots, \hat{\theta}_m)$
 - (iv) Ottenere la stima finale $\hat{h}_n(\theta)$ applicando una spline di lisciamiento ai punti $(\theta_i, h_n^*(\theta_i))$
 - (v) Ottenere $\tilde{\theta}^{**}$ tramite $\hat{h}_n^{-1}(\theta)$
-

Dunque, l'algoritmo segue uno schema di N valori di θ , θ_i con $i = 1, \dots, N$, e per ciascuno prevede di ottenere m valori dello stimatore. Gli Autori suggeriscono un valore di $N = 300$ e m variabile da 20 a 100 in funzione della variabilità dello stimatore.

Una stima $h_n^*(\theta_i)$ è ottenuta dalla mediana delle m replicazioni, mentre la stima finale $\hat{h}_n(\theta)$ di $h_n(\theta)$ applicando un lisciatore non parametrico, tipicamente una spline di lisciamiento, ai punti $(\theta_i, h_n^*(\theta_i))$. Per concludere, $\tilde{\theta}^{**}$ è ottenuta dall'inversione $\hat{h}_n^{-1}(\hat{\theta})$ tramite interpolazione, in alternativa si possono utilizzare stimatori equivalenti a quello di massima verosimiglianza come quello basato sul metodo dei momenti $\tilde{\theta}$.

Se la funzione $h_n(\theta)$ è monotona, $\tilde{\theta}^{**}$ è non distorto in mediana. Nel caso in cui la monotonia di $h_n(\theta)$ non sia nota a priori, risulta opportuno verificarla prima del passo (v).

La distinzione tra il caso parametrico e quello semi-parametrico si esplicita nel passo (iii) relativo alle simulazioni: nel primo caso si simula dal modello teorico sottostante m volte, altrimenti si applica una procedura di bootstrap non parametrico.

In particolare, in assenza di una formulazione esplicita per lo stimatore di θ e di un modello parametrico di riferimento da cui generare nuovi campioni, si utilizza il funzionale $T(F)$, le cui proprietà vengono analizzate attraverso ricampionamento dei dati originali. La distribuzione campionaria di $T(F)$ è approssimata empiricamente tramite bootstrap non parametrico e la funzione mediana $h_n(\theta)$ viene stimata ricostruendo la relazione tra il parametro θ e la distribuzione delle stime bootstrap.

Algoritmo 3 Algoritmo di riduzione della distorsione in mediana:
caso semi-parametrico

Input: Intervallo (θ_L, θ_U) , numero di punti N , numero di campioni bootstrap R

Output: Stima non distorta in mediana $\tilde{\theta}^{**}$

- (i) Definire l'intervallo (θ_L, θ_U) attorno $\tilde{\theta} = T(\hat{F})$, stima distorta in mediana sul campione originale
 - (ii) Selezionare N punti $(\theta_1, \dots, \theta_N)$ in (θ_L, θ_U)
 - (iii) Generare R campioni bootstrap dal campione originale e calcolare le stime bootstrap $T_1^*, \dots, T_R^*$
 - (iv) **for** $i = 1, \dots, N$ **do**
 - Calibrare le stime bootstrap sul punto θ_i tramite la funzione $t(\cdot)$:
$$T_{r,i}^* = t^{-1}(t(T_r^*) + (t(\theta_i) - t(\tilde{\theta}))), \quad r = 1, \dots, R$$
 - Stimare $h_n(\theta_i)$ come
$$h_n^*(\theta_i) = \text{Me}(T_{1,i}^*, \dots, T_{R,i}^*)$$
 - (v) Ottenere la stima finale $\hat{h}_n(\theta)$ applicando una spline di lisciamiento ai punti $(t(\theta_i), t(h_n^*(\theta_i)))$ e verificare che sia monotona
 - (vi) Ottenere $\tilde{\theta}^{**}$ tramite $\hat{h}_n^{-1}(\theta)$ in $\tilde{\theta}$ e l'applicazione della funzione inversa $t^{-1}(\cdot)$
-

Tuttavia, nel caso parametrico si svolge la simulazione dei m campioni per ciascuno degli N punti della griglia risultando in $N \times m$ realizzazioni dello stimatore, mentre nel caso semi-parametrico si generano R replicazioni bootstrap dal campione originale ottenendo R stime $T_r^*, r = 1, \dots, R$, che vanno riscalate sulla griglia di valori $\theta_1, \dots, \theta_N$.

Ciò è necessario solo nel caso del bootstrap non parametrico in quanto il ricampionamento avviene sempre dallo stesso campione originale, dunque si ottiene una distribuzione di stime che riflette la variabilità dello stimatore attorno alla stima empirica originale $\tilde{\theta}$, ma non comprende la variabilità rispetto agli N valori della griglia. Al contrario, nella controparte parametrica si ottengono m nuovi campioni simulandoli dal modello sottostante ponendo il vero valore del parametro θ pari a θ_i e dunque la distribuzione simulata dipende in maniera esplicita da $\theta_1, \dots, \theta_N$.

In assenza di un'opportuna calibrazione le stime bootstrap T_r^* non forniscono alcuna informazione sul comportamento dello stimatore nel caso in cui il vero parametro fosse θ_i . A tale scopo si applica un'operazione di riscalatura in maniera tale che la distribuzione bootstrap sia centrata rispetto al punto corrente $\theta_i, i = 1, \dots, N$. La tecnica di proiezione non è unica poiché deve rispettare le proprietà asintotiche dello stimatore e i

vincoli dello spazio parametrico. La formulazione unificata basata sulla funzione $t(\cdot)$ è

$$T_{r,i}^* = t^{-1}(t(T_r^*) + (t(\theta_i) - t(\tilde{\theta}))), \quad r = 1, \dots, R, \quad i = 1, \dots, N. \quad (2.4)$$

Rispetto alle stime originali T_r^* , le nuove stime $T_{r,i}^*$ simulano la distribuzione dello stimatore se il vero parametro fosse θ_i .

Tale approccio risulta vantaggioso in quanto, linearizzando lo spazio dei parametri tramite la trasformazione $t(\cdot)$ prima dell'adattamento della spline di liscio, questa opera su relazioni stabili, così garantendo stime coerenti indipendentemente dal parametro d'interesse. L'applicazione dell'inversa $t^{-1}(\cdot)$, invece, garantisce che le stime $T_{r,i}^*$ appartengano allo spazio originale. Nonostante ciò, la scelta opportuna della funzione $t(\cdot)$ è essenziale, in caso contrario la stima empirica della funzione mediana $\hat{h}_n(\theta)$ risulterebbe distorta.

La tecnica di riscalatura riprende il metodo utilizzato nei test d'ipotesi basati su procedure bootstrap non parametriche, in cui le stime ottenute dal ricampionamento vengono riscalate per imporre che l'ipotesi nulla sia verificata (Hall & Wilson, 1991; Efron & Tibshirani, 1993, paragrafo 16.4). Tuttavia, per l'utilizzo attuale è stata introdotta la linearizzazione dello spazio parametrico mediante la trasformata $t(\cdot)$.

Infine, in questo contesto il valore di R dovrebbe essere compreso tra 200 e 500, in modo tale da permettere di ottenere stime stabili senza aumentare eccessivamente il costo computazionale.

2.2.3 Riduzione analitica della distorsione tramite espansioni asintotiche

Una terza classe di metodologie affronta la riduzione della distorsione attraverso un approccio analitico esplicito (Benussi et al., 2026). A differenza dei metodi basati sulla modifica delle funzioni score, che richiedono la risoluzione numerica di equazioni modificate, e dei metodi basati sul bootstrap, che si basano su simulazioni computazionalmente costose, questo approccio corregge lo stimatore di massima verosimiglianza a posteriori tramite formulazioni in forma chiusa.

Sia $\theta \in \Theta \subset \mathbb{R}^p$ un parametro vettoriale la cui stima di massima verosimiglianza è $\hat{\theta}$, e sia $s : \Theta \rightarrow S \subset \mathbb{R}$ una funzione che definisce il parametro scalare d'interesse $\psi = s(\theta)$ per cui $\hat{\psi} = s(\hat{\theta})$. Si consideri una versione centrata e riscalata di $\hat{\psi}$

$$W = Var(\hat{\psi})^{-1/2} \{\hat{\psi} - E(\hat{\psi})\},$$

utilizzando lo sviluppo di Cornish-Fisher si ottiene Z , una quantità pivotale asintoticamente normale definita come

$$Z = W - \frac{\rho_3}{6}(W^2 - 1),$$

dove $\rho_3 = K_3/K_2^{3/2}$ è il terzo cumulante standardizzato ottenuto come funzione di $K_2 = \text{Var}(\hat{\psi})$ e $K_3 = E((\hat{\psi} - E(\hat{\psi}))^3)$. Dato che il pivot ha distribuzione approssimata normale standard, affinché lo stimatore sia non distorto in mediana si pone $Z = 0$.

Il primo stimatore proposto dagli Autori è

$$\tilde{\psi}^o = \hat{\psi} - B_{(\psi)} + \frac{1}{6} \frac{K_3}{K_2},$$

dove $B_{(\psi)} = E(\hat{\psi} - \psi)$ è la distorsione dello stimatore. Questo è detto stimatore *Oracle* in quanto assume che siano note quantità teoriche e dipendenti dal vero e ignoto valore del parametro θ , quali l'esatta distorsione $B_{(\psi)}$, la varianza teorica K_2 e il terzo cumulante teorico K_3 .

Tuttavia, se sono note le funzioni $B_{(\psi)}(\cdot)$, $K_2(\cdot)$ e $K_3(\cdot)$ allora tali quantità possono essere ottenute come funzione della stima $\hat{\theta}$ da cui

$$\check{\psi} = \hat{\psi} - \hat{B}_{(\psi)} + \frac{1}{6} \frac{\hat{K}_3}{\hat{K}_2}.$$

Nel caso in cui le funzioni non siano ottenibili in forma esplicita, queste possono essere sostituite da un'approssimazione asintotica del primo ordine calcolata in $\hat{\theta}$. Dunque, la distorsione, la varianza e il terzo cumulante vengono rispettivamente approssimati da

$$b_{(\psi)} = b_{(\theta)}^\top \frac{\partial s(\theta)}{\partial \theta} + \frac{1}{2} \text{Tr} \left(i(\theta)^{-1} \frac{\partial^2 s(\theta)}{\partial \theta \partial \theta^\top} \right), \quad \kappa_2 = \omega^\top \frac{\partial s(\theta)}{\partial \theta}$$

$$\text{e } \kappa_3 = -6 \sum_{t=1}^p \omega_t \omega_t^\top \left(\frac{1}{3} P_t + \frac{1}{2} Q_t \right) \omega + 3 \omega^\top \frac{\partial^2 s(\theta)}{\partial \theta \partial \theta^\top} \omega,$$

dove $b_{(\theta)} = -i(\theta)^{-1} A$, $\omega = i(\theta)^{-1} \frac{\partial s(\theta)}{\partial \theta}$, $P_t = E(l_*(\theta) l_*^\top(\theta) l_t(\theta))$, $Q_t = -E(j(\theta) l_t(\theta))$ e $A = \frac{1}{2} \text{Tr}(i(\theta)^{-1} (P_t + Q_t))$.

Lo stimatore finale ha formulazione

$$\check{\psi} = \hat{\psi} - \hat{b}_{(\psi)} + \frac{1}{6} \frac{\hat{\kappa}_3}{\hat{\kappa}_2}.$$

Nel caso in cui le componenti dello stimatore risultino di difficile elaborazione analitica, esse possono essere stimate tramite simulazioni Monte Carlo dal modello di riferimento.

Benussi et al. (2026) si limita al caso parametrico, tuttavia è possibile estendere l'utilizzo dello stimatore nel contesto semi-parametrico dove le quantità necessarie alla correzione sono stimate empiricamente tramite simulazione Monte Carlo basata su ricampionamento bootstrap non parametrico sul campione originale, sulla base del seguente algoritmo.

Algoritmo 4 Metodo esplicito analitico di riduzione della distorsione in mediana : estensione al caso semi-parametrico

Input: Stima originale $\hat{\psi} = s(\hat{\theta})$, numero di repliche M

Output: Stima non distorta in mediana $\check{\psi}$

(i) Generare M repliche dal campione originale tramite bootstrap non parametrico

for $j = 1, \dots, M$ **do**

Calcolare la stima bootstrap $\hat{\theta}_j$
 Ottenere la stima del parametro d'interesse tramite la trasformazione:
 $\hat{\psi}_j = s(\hat{\theta}_j)$

(ii) Calcolare la distorsione e cumulanti empirici per il parametro d'interesse

- la media Monte Carlo sulle repliche: $\bar{\psi} = \frac{1}{M} \sum_{j=1}^M \hat{\psi}_j$
- la distorsione empirica: $\hat{B}_{(\psi)} = \bar{\psi} - \hat{\psi}$
- la varianza empirica: $\hat{K}_2 = \frac{1}{M-1} \sum_{j=1}^M (\hat{\psi}_j - \bar{\psi})^2$
- il terzo cumulante empirico: $\hat{K}_3 = \frac{1}{M} \sum_{j=1}^M (\hat{\psi}_j - \bar{\psi})^3$

(iii) Calcolare la stima non distorta in mediana basata sulla correzione esplicita tramite espansione di Cornish-Fisher:

$$\check{\psi} = \hat{\psi} - \hat{B}_{(\psi)} + \frac{1}{6} \frac{\hat{K}_3}{\hat{K}_2}$$

L'utilizzo del denominatore M nel calcolo del terzo cumulante empirico $\hat{K}_3$ ha due ragioni: dapprima, se il numero di repliche M è elevato il fattore di correzione per campioni finiti $\frac{M^2}{(M-1)(M-2)}$ converge a 1. In secondo luogo, la formula con $1/M$ rappresenta lo stimatore per sostituzione dei momenti empirici rispetto alla distribuzione discreta uniforme delle repliche. Affinché la stima di K_3 sia accurata e stabile, si raccomanda di fissare un valore di M tra 500 e 2000, tale è superiore al numero di repliche R e m impiegati nel paragrafo precedente in quanto la stima dei momenti di ordine superiore al secondo richiede una dimensione campionaria più elevata.

Rispetto ai metodi basati sul bootstrap, questo metodo richiede un costo computazionale ridotto in quanto, laddove le formule siano note, si basa su un singolo calcolo analitico, evitando così le difficoltà di convergenza tipiche dei metodi espliciti. Inoltre,

permette il trattamento dei parametri di disturbo in maniera efficiente, poiché l'approccio si focalizza esclusivamente sulla funzione d'interesse $\psi = s(\theta)$. Per contro, la formula di Benussi et al. (2026) richiede il calcolo delle derivate e dei momenti di terzo ordine della log-verosimiglianza, esistono quindi limitazioni legate alla differenziabilità e alla regolarità delle funzioni del modello che non si riscontrano in Cabrera & Watson (1997). Infine, l'approccio è specificamente progettato per un parametro scalare, dunque se l'obiettivo inferenziale richiede la correzione in mediana simultanea di un intero vettore di parametri, l'approccio implicito di Kenne Pagui et al. (2017) risulta più appropriato.

2.3 Applicazione ai GLM

Nel contesto dei modelli lineari generalizzati con interesse al parametro di dispersione ϕ , i temi precedentemente trattati sono di particolare rilievo. Infatti, la proprietà di non distorsione in mediana nel caso di stimatori con distribuzione asimmetrica è preferibile rispetto a quella in media. Questo è il caso per il parametro di dispersione che ha tipicamente distribuzione campionaria asimmetrica a destra. Mentre la media campionaria dello stimatore viene influenzata dai valori estremi nella coda destra, la mediana fornisce un'indicazione più appropriata del valore centrale dello stimatore. La non distorsione in mediana garantisce che lo stimatore abbia la stessa probabilità di sovrastimare o sottostimare il valore vero del parametro.

Inoltre, la mediana gode della proprietà di equivarianza per trasformazioni monotone: ciò garantisce che se la stima del parametro ϕ è non distorta in mediana, lo rimarrà anche sue funzioni strettamente monotone. Al contrario, la correzione della distorsione in media non viene preservata sotto trasformazioni non lineari, rendendo necessario calcolare nuovamente le derivate al variare della parametrizzazione. Tale correzione in mediana, come già anticipato, è di particolare interesse quando il numero di parametri di disturbo è elevato rispetto alle osservazioni, configurando il problema di Neyman & Scott (1948; si veda anche Sartori, 2003), proprietà che vale anche per la correzione in media. Inoltre, evidenze empiriche mostrano come l'ottenimento di una stima di ϕ non distorta in mediana consenta la correzione dei test d'ipotesi e il raggiungimento di una copertura degli intervalli di confidenza per i coefficienti β più vicina al livello nominale.

2.3.1 Strategia di aggiustamento misto

Per i modelli lineari generalizzati Kosmidis et al. (2020) introducono una strategia di aggiustamento misto (*mixed adjustment strategy*). Tale procedura si propone di introdurre una metodologia unificata per la riduzione della distorsione in media e in mediana

tramite correzione della funzione score in maniera tale da mantenere l'invarianza rispetto a trasformazioni lineari dei parametri β dell'approccio di Firth (1993) e l'invarianza rispetto a trasformazioni monotone del metodo di Kenne Pagui et al. (2017). I risultati di tale metodologia differiscono a seconda del modello sottostante: per modelli con ϕ noto, come Poisson o binomiale, l'aggiustamento misto equivale alla correzione in media. Al contrario, per modelli normali lineari si ha che la strategia mista porta alla non distorsione in mediana.

A tale scopo, gli Autori si basano sull'ortogonalità di Fisher dei parametri (1.1) per trattare separatamente β e ϕ , rispettivamente con una correzione per distorsione in media e in mediana. Il calcolo di questi stimatori prevede la soluzione di un sistema di equazioni non lineari, tuttavia è applicabile l'algoritmo di Quasi-Fisher Scoring, equivalente alla metodologia classica per la stima nei GLM dell'algoritmo IWLS (*Iteratively Reweighted Least Squares*). L'algoritmo presenta vantaggi in termini di convergenza e stabilità rispetto alla massima verosimiglianza.

Nello scenario in cui i coefficienti di regressione β vengano trattati come parametri di disturbo, l'approccio misto si basa non più sulla funzione score ma su quella di log-verosimiglianza profilo modificata. La proprietà di ortogonalità risulta ancora valida e permette di applicare il metodo di correzione per la distorsione in mediana alla funzione di score profilo. Tuttavia, non essendo più applicabile l'algoritmo di ottimizzazione simultanea, si utilizza un processo iterativo a due stadi nidificati: si stima ϕ non distorto in mediana e si aggiornano i valori $\hat{\beta}_\phi$ tramite l'algoritmo IWLS, così da isolare la stima della dispersione dall'incertezza legata alle covariate.

2.3.2 Compromesso varianza-distorsione

L'utilizzo di metodologie con l'obiettivo della riduzione della distorsione, che sia in media o in mediana, porta a riscontrare il problema del compromesso tra varianza e distorsione. In campioni di dimensioni limitate, la correzione della distorsione causa un incremento della varianza campionaria dello stimatore in confronto a quello di massima verosimiglianza. L'impatto complessivo per uno stimatore di ϕ generico (espresso come $\tilde{\phi}$) può essere misurato tramite l'errore quadratico medio (MSE), definito come

$$MSE(\tilde{\phi}) = Var(\tilde{\phi}) + (Dist(\tilde{\phi}))^2,$$

dove $Dist(\tilde{\phi}) = E(\tilde{\phi}) - \phi$ rappresenta la distorsione dello stimatore rispetto al vero valore del parametro ϕ .

In tale contesto, si noti come i metodi di riduzione della distorsione basati sulla funzione score modificata (Kenne Pagui et al., 2017) o sulla tecnica bootstrap (Cabrera & Watson, 1997) rappresentino il contrasto all'approccio proposto da Fletcher (2012). Quest'ultimo stabilizza la stima di ϕ riducendone la varianza; al contrario, la riduzione della distorsione in mediana permette un lieve incremento della variabilità campionaria a fronte di migliori proprietà inferenziali globali, come una maggiore stabilità degli intervalli di confidenza.

2.4 Intervalli di confidenza

Una stima intervallare è una procedura inferenziale che incorpora la stima di un parametro scalare con una misura d'incertezza, andando così a comporre un intervallo di confidenza $IC_{1-\alpha}$: un intervallo casuale che include il vero valore del parametro d'interesse con una probabilità pari a $1 - \alpha$, detto livello di confidenza, con α fissato a priori.

In questo contesto, la probabilità di copertura indica la probabilità che l'intervallo casuale $IC_{1-\alpha}$ contenga effettivamente il valore vero del parametro θ , l'obiettivo dell'inferenza è ottenere empiricamente una probabilità che coincida con il livello di confidenza nominale $1 - \alpha$, o vi si approssimi al crescere della dimensione campionaria.

In seguito, si richiama il metodo per la costruzione degli intervalli di confidenza basato sulla quantità approssimativamente pivotale di Wald, e si presenta l'approccio di costruzione di intervalli di tipo non parametrico ossia la procedura HulC (Kuchibhotla et al., 2024, *Hull-based Confidence*).

2.4.1 Intervalli di confidenza di Wald

Gli intervalli di confidenza sono usualmente costruiti sulla base di una quantità pivotale, una funzione $q: S \times \Theta \rightarrow \mathbb{R}$ tale che, per ogni $\theta \in \Theta$, $q(Y, \theta)$ è variabile casuale con distribuzione di probabilità che non dipende da θ .

La quantità di Wald è una quantità approssimativamente pivotale definita come

$$r_e(\theta) = \frac{\hat{\theta} - \theta}{\sqrt{\widehat{Var}(\hat{\theta})}}.$$

Essa rappresenta una standardizzazione dello stimatore di massima verosimiglianza ottenuta valutando la distanza tra stima e vero valore, e tenendo conto dell'errore standard.

L'intervallo di confidenza costruito sulla base di tale quantità è l'insieme di valori

$$\{\theta \in \Theta: |r_e(\theta)| \leq z_{1-\frac{\alpha}{2}}\},$$

esprimibile anche in forma esplicita

$$IC_{1-\alpha}(\theta) = \left\{ \theta \in \Theta: \theta \in \left(\hat{\theta} \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{Var}(\hat{\theta})} \right) \right\},$$

dove $z_{1-\frac{\alpha}{2}}$ è il quantile di livello $1 - \alpha/2$ della distribuzione normale standard $N(0, 1)$ e $\widehat{Var}(\cdot)$ è la varianza stimata, ottenuta dalla diagonale dell'inversa della matrice d'informazione osservata.

La statistica di Wald può essere calcolata in maniera analoga basandosi su stimatori asintoticamente equivalenti a quello di massima verosimiglianza, come ad esempio la stima ottenuta tramite il metodo dei momenti $\tilde{\theta}$.

2.4.2 Metodo HulC

Il metodo HulC (*Hull-based Confidence*) è una procedura per la costruzione di intervalli di confidenza proposta da Kuchibhotla et al. (2024). Questa si basa sull'assunzione che la distribuzione dello stimatore a cui affiancare la stima intervallare non sia estremamente asimmetrica, a tale scopo l'asimmetria è misurata tramite la distorsione in mediana dello stimatore considerato.

L'involucro (*hull*) convesso di H stime è l'intervallo delimitato dal valore minimo e da quello massimo. Per tale motivo gli intervalli così costruiti prendono il nome di *Hull-based Confidence* (HulC).

Sia $\tilde{\theta}$ uno stimatore non distorto in mediana per θ e siano $\tilde{\theta}_j$, $j = 1, \dots, H$, repliche indipendenti di $\tilde{\theta}$. Allora l'intervallo delimitato dal minimo $\tilde{\theta}_{min}$ e dal massimo $\tilde{\theta}_{max}$ dei H stimatori ha una probabilità di copertura almeno pari a $1 - 2^{-H+1}$. Affinché $1 - 2^{-H+1} \geq 1 - \alpha$ e dunque l'intervallo $(\tilde{\theta}_{min}, \tilde{\theta}_{max})$ raggiunga un livello di almeno $1 - \alpha$, H viene determinato come $H = H_\alpha = \lceil \log_2(2/\alpha) \rceil$.

Si noti che il numero di stimatori richiesti è dunque limitato, infatti posto $\alpha = 0.05$ si ottiene $H \geq 6$. Inoltre, stimatori indipendenti possono essere ricavati dividendo i dati in maniera tale che ogni stimatore si basi su circa n/H_α osservazioni; avere campioni di uguale numerosità non è necessario.

Nel caso in cui l'assunzione di non distorsione in mediana dello stimatore iniziale venga a mancare, il numero di suddivisioni H_α , per ottenere lo stesso livello α , aumenta

secondo la formulazione

$$H_\alpha = \left\lceil \frac{\log(2/\alpha)}{\log(2/(1+2\Delta))} \right\rceil,$$

dove $\Delta \in [0, 1/2)$ è la massima distorsione in mediana tra gli H stimatori indipendenti $\tilde{\theta}_j$, ovvero

$$\Delta = \max_{j=1,\dots,H} \left(\frac{1}{2} - \min(\mathbb{P}(\tilde{\theta}_j \leq \theta_0), \mathbb{P}(\tilde{\theta}_j \geq \theta_0)) \right)_+,$$

con $(k)_+ = \max(k, 0)$, $\forall k \in \mathbb{R}$.

In confronto ai metodi più comuni per la stima intervallare, la procedura HulC non richiede la stima della distribuzione dello stimatore, come nel caso degli intervalli alla Wald, né pone vincoli sul tasso di convergenza, come per tecniche basate sul sotto-campionamento. Per quanto riguarda i metodi di costruzione di intervalli basati su procedure bootstrap, la procedura HulC si basa su assunzioni più deboli, in quanto l'unica assunzione riguarda la conoscenza della distorsione in mediana massima Δ , e richiede solo il calcolo di H stime, dunque è meno computazionalmente impegnativo. Infatti, calcolare H stime su campioni di dimensione n/H è potenzialmente meno costoso che calcolarne uno su tutti i dati.

Il fatto che il metodo si incentri su stimatori, almeno asintoticamente, non distorti in mediana non costituisce un limite: se il parametro è identificabile allora la distorsione in mediana può essere ridotta e il metodo HulC applicato, si veda ad esempio Benussi et al. (2026) dove gli intervalli così ottenuti hanno copertura empirica in generale molto prossima a quella nominale.

Capitolo 3

Aspetti computazionali

Il presente capitolo è volto a introdurre la metodologia computazionale e i dettagli di implementazione dei metodi descritti al capitolo precedente. Nella prima parte si presentano le assunzioni fondamentali e le condizioni di regolarità che garantiscono la validità asintotica e la consistenza delle procedure bootstrap. Successivamente, si descrivono i dettagli specifici relativi al codice sviluppato per la tesi.

Il paragrafo 3.1 ha come riferimento bibliografico principale Davison & Hinkley (1997, Capitolo 2).

3.1 Il metodo bootstrap: alcune considerazioni generali

Il metodo bootstrap, introdotto da Efron (1979) e trattato approfonditamente ad esempio in Davison & Hinkley (1997), rappresenta una metodologia statistica parametrica e non parametrica basata su calcolo intensivo. Tale approccio risulta applicabile in contesti non trattabili con metodi tradizionali.

Il bootstrap si basa su un'applicazione del principio di sostituzione (*plug-in*) a due stadi. Si consideri un vettore di osservazioni indipendenti e identicamente distribuite $y = (y_1, \dots, y_n)$ ottenute da una popolazione con funzione di ripartizione F non nota. Sia $\theta = T(F)$ il parametro d'interesse, stimato mediante il funzionale statistico $\hat{\theta} = T(\hat{F})$, dove la funzione di ripartizione empirica $\hat{F}$ viene sostituita a F una prima volta. Le proprietà inferenziali di $\hat{\theta}$ dipendono da una quantità $p(\hat{\theta}, F)$, anch'essa ignota in quanto dipendente da F . Il principio di *plug-in* viene quindi applicato una seconda volta, approssimando la quantità d'interesse tramite la sua controparte empirica $p(\hat{\theta}, \hat{F})$.

Pertanto, si definisce il campione bootstrap come una simulazione del processo generativo dei dati originari da $\hat{F}$, dalla quale viene estratto un vettore $y^* = (y_1^*, y_2^*, \dots, y_n^*)$. Sul vettore ricampionato può essere calcolata la replica bootstrap dello stimatore:

$$\hat{\theta}^* = T(y^*)$$

e, ripetendo tale procedura un numero R elevato di volte, la distribuzione di frequenza empirica di $\hat{\theta}^*$ approssima la distribuzione campionaria reale di $\hat{\theta}$.

Nell'applicazione di tale metodo si distingue tra bootstrap parametrico e non parametrico, differenziati dalla modellazione della distribuzione F . Il bootstrap parametrico presuppone che F appartenga a una famiglia i cui elementi sono identificati da un vettore di parametri θ . Dunque, si stimano i parametri sul campione originale, ottenendo così $\hat{\theta}$, e si simulano campioni estraendo valori dalla funzione di ripartizione $F_{\hat{\theta}}$.

Al contrario, nel bootstrap non parametrico, il quale non vincola la distribuzione teorica della popolazione originale a una specifica famiglia, la stima $\hat{F}$ coincide con la funzione di ripartizione che assegna probabilità uniforme pari a $1/n$ a ciascuna osservazione del campione. Di conseguenza, il ricampionamento è svolto estraendo con reinserimento le osservazioni dal campione iniziale. Affinché tale procedura sia valida, risulta necessario che $\hat{F}$ sia uno stimatore consistente della vera F sotto adeguate condizioni di regolarità e di campionamento. Questo approccio risulta essere più robusto in confronto a quello parametrico che, per quanto più efficiente sotto corretta specificazione, porta a sistematiche distorsioni nel caso di non corretta specificazione.

Come descritto in Davison & Hinkley (1997, paragrafo 2.6.1), l'efficacia del metodo bootstrap è vincolata ad assunzioni e condizioni di regolarità: in particolare alla convergenza asintotica di $\hat{F}$ verso F . Si assume l'esistenza di un intorno $\mathcal{N}$ della vera F , al cui interno la stima $\hat{F}$ deve ricadere con probabilità pari a uno quando n tende ad infinito. Sotto queste condizioni, si indica con $G_{A,n}$ la distribuzione campionaria della statistica d'interesse, calcolata in rispetto a una legge probabilistica $A \in \mathcal{N}$. Affinché il principio di sostituzione sia valido e, dunque, il bootstrap sia consistente devono essere rispettate tre condizioni di regolarità $\forall A \in \mathcal{N}$: la prima è la convergenza debole della distribuzione campionaria $G_{A,n}$ a un limite asintotico stabile $G_{A,\infty}$ per $n \rightarrow \infty$. Le restanti condizioni impongono che tale convergenza sia uniforme in $\mathcal{N}$ e che la funzione da A a $G_{A,\infty}$ sia continua. Queste sono necessarie ad assicurare che $G_{\hat{F},n}$ tenda a $G_{\hat{F},\infty}$ qualunque sia la sequenza di $\hat{F}$ definita all'aumentare di n .

Le violazioni delle condizioni di regolarità causano problematiche di convergenza e instabilità all'interno delle simulazioni bootstrap. La non efficacia del metodo si riconduce

a quattro scenari: dapprima ci sono le statistiche d'ordine estremo, quando lo stimatore d'interesse risente in grande parte delle osservazioni poste agli estremi del supporto campionario. Una seconda criticità è quella in presenza di parametri sulla frontiera, ossia quando il vero valore del parametro d'interesse appartiene al limite estremo dello spazio parametrico. Il terzo caso riguarda le distribuzioni a code pesanti in quanto il bootstrap non parametrico fallisce se la popolazione originale presenta code pesanti tali da non garantire un momento di secondo ordine finito. Infine, un'ulteriore causa di instabilità è data da stimatori non regolari, definiti come statistiche che non possono essere espresse come funzioni continue e differenziabili del campione originale, il cui caso notevole è la mediana.

Mentre in letteratura le criticità del bootstrap sono spesso associate alla mancata regolarità della mediana campionaria (Davison & Hinkley, 1997, paragrafo 2.6.2), il funzionale $\hat{\theta}_n$ definito in (2.2) è una quantità regolare. Il motivo per cui l'equivalente della correzione bootstrap della distorsione in media di Efron (1982) non può essere applicato alla correzione per la distorsione in mediana è conseguenza della non linearità del funzionale: mentre la media campionaria è lineare e permette di ereditare le proprietà additive del valore atteso, la mediana costituisce un funzionale non lineare basato su statistiche d'ordine (come già mostrato al paragrafo 1.1.4). Di conseguenza, la semplice sottrazione della distorsione stimata non garantisce la rimozione della distorsione in mediana. L'ottenimento di uno stimatore non distorto in mediana richiede l'inversione della funzione di mediana teorica $h_n(\theta_0) = \text{Me}_{\theta_0}(\hat{\theta}_n) = \hat{\theta}$ poiché la funzione $h_n(\cdot)$ non ammette una forma esplicita, imponendo l'impiego di algoritmi numerici o bootstrap per la ricerca della radice.

Quando si estende la metodologia bootstrap ai modelli lineari generalizzati e ai contesti di quasi-verosimiglianza, la procedura richiede specifici adattamenti per gestire la relazione non lineare tra il predittore e il valore atteso, la funzione di legame $g(\cdot)$ e la relazione media-varianza. In questo ambito, l'applicazione del bootstrap parametrico viene effettuata generando le risposte simulate y_i^* direttamente dalla distribuzione della famiglia esponenziale specificata, fissando come parametro il valore atteso stimato $\hat{\mu}_i = g^{-1}(\mathbf{x}_i \hat{\beta})$ (Davison & Hinkley, 1997, Capitolo 7). Tuttavia questo approccio presenta criticità specifiche ai GLM: quando i valori stimati $\hat{\mu}_i$ si trovano in prossimità dei confini dello spazio parametrico, la simulazione parametrica tende a generare campioni bootstrap privi di variabilità. Tali campioni portano frequentemente al mancato raggiungimento della convergenza dell'algoritmo di stima IWLS o a fenomeni di separazione perfetta con stime dei coefficienti che divergono all'infinito. Inoltre, il bootstrap parametrico risulta inapplicabile nel caso del modello di quasi-verosimiglianza poiché

specifica unicamente la funzione di varianza $v(\mu)$, senza una famiglia di distribuzioni teoriche (Wedderburn, 1974).

Al contrario, il bootstrap non parametrico nei modelli lineari generalizzati viene formulato tramite il campionamento con reinserimento delle unità osservate, trattate come coppie di risposta e covariate $(\mathbf{x}_i, y_i)$. Inoltre, nonostante sia l'approccio più appropriato nel caso della quasi-verosimiglianza data la sua robustezza rispetto violazioni della struttura di varianza, anche il bootstrap non parametrico porta a difficoltà. Dapprima la minore efficienza in campioni di ridotta ampiezza; in aggiunta, risente della presenza di punti influenti, infatti l'estrazione casuale dei vettori $\mathbf{x}_i$ può produrre sotto-campioni con matrici del modello non invertibili (Efron & Tibshirani, 1993, Capitolo 9). Infine, qualora la funzione di varianza $v(\mu)$ ipotizzata nella quasi-verosimiglianza sia erroneamente specificata, ciò causa una discordanza tra la variabilità del campione generato e la matrice dei pesi impiegata nelle equazioni di stima a ogni replicazione bootstrap.

3.2 Note nell'implementazione in R

Il presente paragrafo ha l'obiettivo di fornire dettagli aggiuntivi inerenti alle funzioni che implementano in ambiente R i metodi presentati nel paragrafo 2.2. In un primo momento, si descrivono i punti critici e aspetti d'interesse dell'implementazione della funzione fornita nell'Appendice A in riferimento ai metodi del paragrafo 2.2.2. In seguito, si presenta una breve guida all'utilizzo dei pacchetti R già disponibili dedicati alla riduzione della distorsione in mediana tramite correzione della funzione score e un'analisi più dettagliata delle opzioni e parametri della funzione sviluppata.

3.2.1 Implementazione degli Algoritmi 2 e 3

L'implementazione del metodo basato sulla tecnica bootstrap per la riduzione della distorsione in mediana di Cabrera & Watson (1997) è fornito nell'Appendice A (funzione `MeBRboot`, per dettagli si veda 3.2.2) e comprende il metodo proposto sia nel caso parametrico che semi-parametrico. In aggiunta, si distingue rispetto al parametro a cui viene applicato: un singolo coefficiente di regressione β o parametro di dispersione ϕ . La distinzione del tipo di parametro in esame ha due conseguenze principali: dapprima permette di porre filtri che riflettano lo spazio parametrico d'appartenenza. Inoltre, nella versione semi-parametrica, si applica la calibrazione delle stime che dipende dal tipo di parametro.

Calibrazione per i parametri β e ϕ

Come descritto nell'Algoritmo 3 (paragrafo 2.2.2), applicato il bootstrap non parametrico, si ottengono le stime $T_r^*, r = 1, \dots, R$, che non dipendono dalla griglia di valori $\theta_i, i = 1, \dots, N$, e per tale motivo vanno riscalate. Tuttavia la calibrazione dipende dal tipo di parametro.

Dato che i coefficienti β possono assumere valori in $\mathbb{R}$ e tendono a una distribuzione simmetrica, la riscalatura viene trattata come una traslazione

$$T_r^* + (\theta_i - \tilde{\theta}),$$

in modo da preservarne la variabilità e la forma. Viceversa, per ϕ la riscalatura è moltiplicativa ossia una traslazione in scala logaritmica

$$\exp(\log(T_r^*) + (\log(\theta_i) - \log(\tilde{\theta}))),$$

in quanto ϕ può assumere solo valori positivi, ha distribuzione asimmetrica e la variabilità dello stimatore cresce con il valore del parametro stesso. Applicare una traslazione al parametro ϕ causerebbe la generazione di valori negativi (se $\tilde{\theta} > \theta_i$) o valori elevati del parametro con variabilità erroneamente ridotta (se $\tilde{\theta} < \theta_i$).

Ciò riflette la formulazione unificata (2.4) con la funzione $t(x) = \log(x)$ se l'interesse ricade sul parametro ϕ e $t(x) = x$ se β . L'utilizzo di tale formulazione unificata permette una semplice estensione all'inserimento di un ulteriore parametro d'interesse con spazio parametrico alternativo, ad esempio introducendo una probabilità ρ appartenente a $(0, 1)$ la funzione potrebbe essere $t(x) = \text{logit}(x)$.

Proposta di intervallo di confidenza per il caso semi-parametrico

L'algoritmo implementato offre, inoltre, la possibilità di calcolare un intervallo di confidenza per il parametro d'interesse al livello di confidenza nominale $1 - \alpha$, sebbene tale estensione sia vincolata all'utilizzo dell'approccio semi-parametrico. La costruzione di tali intervalli si basa sull'assunzione di invarianza rispetto alla traslazione applicata sulla scala definita dalla funzione $t(\cdot)$.

In particolare, il bootstrap non parametrico genera una distribuzione delle stime, $T_r^*, r = 1, \dots, R$, non condizionate alla griglia di valori θ_i . Tale distribuzione mantiene la variabilità empirica del campione originale, senza alcuna assunzione sulla distribuzione. Ottenuta la stima non distorta in mediana $\tilde{\theta}^{**}$, le replicazioni bootstrap possono essere traslate, all'interno dello spazio parametrico trasformato di una quantità pari

alla distorsione stimata, ossia pari alla differenza tra $t(\tilde{\theta})$ e $t(\tilde{\theta}^{**})$. Dunque, la formula unificata risulta essere

$$T_{r,grid}^* = t^{-1}(t(T_r^*) + (t(\tilde{\theta}^{**}) - t(\tilde{\theta})))$$

e l'intervallo di confidenza viene determinato estraendo i quantili empirici di livello $\alpha/2$ e $1 - \alpha/2$ dai valori $T_{r,grid}^*$, $r = 1, \dots, R$.

Tale metodo è stato ideato come alternativa alla costruzione di un intervallo di confidenza alla Wald che richiede il calcolo di una misura di incertezza come lo standard error della stima, a tale fine è necessario ricorrere a un *double bootstrap* (Davison & Hinkley, 1997, paragrafo 5.6), ossia un bootstrap a due livelli annidati che richiede elevato costo computazionale.

Inoltre, la procedura proposta si distingue dai tradizionali intervalli bootstrap percentili (Davison & Hinkley, 1997, paragrafo 5.3), i quali estraggono i quantili dalla distribuzione bootstrap originale non traslata così da ottenere intervalli centrati sulla stima $\tilde{\theta}$ distorta. Se lo stimatore di partenza presenta una forte distorsione, l'intervallo percentile classico risulterà avere una copertura reale lontana da quella nominale. In aggiunta, l'approccio proposto si differenzia dagli intervalli *Bias-Corrected* (Efron, 1987), questi correggono la distorsione modificando il livello dei quantili. Inoltre, tale metodo assume l'esistenza di una trasformazione che renda la distribuzione dello stimatore normale.

La limitazione dell'applicazione del metodo al caso semi-parametrico è dovuta al fatto che un intervallo di confidenza basato sulla distribuzione delle repliche parametriche non sarebbe rappresentativo della reale distribuzione campionaria dello stimatore. Anzi, rifletterebbe le assunzioni distributive così causando conclusioni inferenziali distorte, in particolare con ampiezza dipendente dal numero di replicazioni e dalla varianza teorica.

Riduzione della distorsione in mediana congiunta di β e ϕ

Nel caso in cui fosse d'interesse correggere congiuntamente il vettore p -dimensionale β e il parametro scalare ϕ , in alternativa al metodo proposto da Kosmidis et al. (2020) o in contesti semi-parametrici, è possibile applicare l'Algoritmo 5 che segue uno schema in tre fasi. Sebbene, sotto condizioni di regolarità, i parametri β e ϕ dei GLM godano di ortogonalità asintotica, in campioni di numerosità finita si evidenzia una correlazione tra le distribuzioni campionarie degli stimatori. In particolare, la stima del parametro di dispersione ϕ dipende dai residui del modello e, di conseguenza, risente della distorsione dei coefficienti di regressione β , e viceversa. La soluzione ottimale sarebbe la correzione

congiunta dei parametri tramite un algoritmo multivariato che operi su (β, ϕ) , tuttavia ciò richiederebbe un elevato tempo di calcolo.

Con l'algoritmo alternativo proposto si ottiene la stima congiunta non distorta in mediana svolgendo tre passaggi: correzione dei coefficienti β , correzione del parametro ϕ e rifinitura delle stime di β .

Come dimostrato da MacKinnon & Smith (1998) e Guerrier et al. (2019), iterare la correzione dei due componenti fino a convergenza numerica esatta tenta di correggere residui di ordine superiore $O(n^{-2})$ che, nei campioni finiti, sono dominati dalla variabilità campionaria e dal rumore stocastico della simulazione. Dunque, proseguire con le iterazioni oltre il secondo passaggio alternato non porta alla riduzione della distorsione, ma causa un'inflazione della varianza dello stimatore e una marcata instabilità numerica nelle stime corrette. Risulta opportuno notare che, sebbene gli articoli citati siano stati sviluppati nel contesto della correzione della distorsione in media, le loro conclusioni si applicano alla centratura in mediana. Essendo lo stimatore non distorto in mediana ottenuto da tecnica bootstrap soggetto a una variabilità stocastica maggiore rispetto al valore atteso, l'iterazione dell'algoritmo alla ricerca del punto di convergenza porterebbe lo stimatore a un rischio ancora più elevato di instabilità numerica e inflazione della varianza.

Dunque, un singolo stadio correttivo, o una sequenza alternata a blocchi in presenza di vettori di parametri, è teoricamente sufficiente a eliminare il termine di distorsione dominante così riducendo la distorsione residua in mediana a un ordine di grandezza trascurabile.

Algoritmo 5 Algoritmo di riduzione della distorsione in mediana congiunta: approccio sequenziale a tre passi

Input: Stime iniziali $(\hat{\beta}, \hat{\phi})$

Output: Stime congiuntamente non distorte in mediana $(\tilde{\beta}^{**}, \tilde{\phi}^{**})$

(i) Fase di correzione del vettore β con $\phi = \hat{\phi}$ fisso:

Inizializzare il vettore $\beta^{(1)} \leftarrow \mathbf{0}_p$

for $k \leftarrow 1$ **to** p **do**

└ $\beta_k^{(1)} \leftarrow \text{MeBRboot}(\hat{\beta}_k, \dots, \text{parameter}='beta')$

(ii) Fase di correzione di ϕ con $\beta = \beta^{(1)}$ fisso:

$\tilde{\phi}^{**} \leftarrow \text{MeBRboot}(\hat{\phi}, \dots, \text{parameter}='phi')$

(iii) Fase di rifinitura del vettore β con $\phi = \tilde{\phi}^{**}$ fisso:

Inizializzare il vettore $\beta^{**} \leftarrow \mathbf{0}_p$

for $k \leftarrow 1$ **to** p **do**

└ $\tilde{\beta}_k^{**} \leftarrow \text{MeBRboot}(\beta_k^{(1)}, \dots, \text{parameter}='beta')$

return $(\tilde{\beta}^{**}, \tilde{\phi}^{**})$

L'utilizzo di tale algoritmo è di particolare interesse per ottenere una riduzione affidabile della distorsione in mediana per ϕ in campioni di dati distribuiti come una beta (per un'applicazione si veda il caso studio al paragrafo 4.3). Nella regressione beta, le realizzazioni della variabile casuale appartengono all'intervallo $(0, 1)$ e la varianza del modello è $Var(Y) = \frac{\mu(1-\mu)}{1+\phi}$. I coefficienti β stimati tramite massima verosimiglianza sono soggetti a una forte distorsione e dunque l'algoritmo univariato su ϕ fornirà risultati distorti.

La necessità di applicare l'algoritmo iterativamente può essere ricondotta al fatto che la distribuzione beta non appartenga ai modelli lineari generalizzati in quanto la sua densità non può essere espressa in forma DE_1 . La conseguenza principale risulta essere che, mentre nei GLM il parametro di dispersione ϕ è un fattore di scala che non ha effetti sulla stima dei parametri β , nella regressione beta il parametro ϕ è parte della funzione score per β , per cui viene a meno l'ortogonalità tra β e ϕ .

3.2.2 Pacchetti e funzioni R

Nell'ambiente R la riduzione della distorsione in mediana basata sulla correzione della funzione score si trova nel pacchetto **brglm2** (Kosmidis, 2026), tale libreria implementa i metodi impliciti di riduzione della distorsione per i modelli appartenenti ai GLM, la regressione logistica multinomiale e quella binomiale negativa. Le funzioni fornite

permettono la correzione della funzione score per media e mediana, oltre all'approccio di aggiustamento misto. In particolare, tale libreria permette di arricchire gli oggetti ottenuti da `glm()` tramite l'opzione `method="brglmFit"`. La scelta della correzione avviene con `type` che può assumere valore `"ML"` (massima verosimiglianza), `"AS_mean"` (aggiustamento in media), `"AS_median"` (aggiustamento in mediana) o `"AS_mixed"` (aggiustamento misto). Per esempi di utilizzo nel contesto dei modelli lineari generalizzati si rinvia a Kosmidis et al. (2020, Materiale Supplementare).

Nel caso di modelli di regressione beta il pacchetto `betareg` (Zeileis, 2025) comprende il metodo di riduzione della distorsione in media basato sulla modifica della funzione score (`type="BR"`) in alternativa al default della massima verosimiglianza (`type="ML"`), per applicazioni e maggiori informazioni si faccia riferimento a Grün et al. (2012). Infine, la funzione `mbrbetareg` per la riduzione della distorsione in mediana (`type="medianBR"`) è fornita da Kenne Pagui et al. (2017).

Per ciò che riguarda i metodi di riduzione della distorsione in mediana basati sulla tecnica bootstrap la funzione `MeBRboot`, fornita in Appendice A, implementa i dettagli discussi al paragrafo precedente. Inizialmente, si distingue tra il metodo parametrico, l'Algoritmo 2, e il metodo semi-parametrico, l'Algoritmo 3.

I parametri richiesti sono i seguenti:

- **Comuni:** la stima distorta iniziale (`theta_hat`), gli estremi dell'intervallo a cui appartiene la stima non distorta (`theta_L` e `theta_U`) e il valore di N (`N`) che di default è posto pari a 300;
- **Caso parametrico** (`method="model"`): il valore di m (`m` con default uguale a 100) e due funzioni, una da cui simulare i dati `sim_data_fn` e una di stima del parametro d'interesse `est_fn`;
- **Caso semi-parametrico** (`method="boot"`): il valore di R (`R` con default 350), il campione di dati originale (`datas`) e la funzione di stima del parametro `est_fn`.

Successivamente, all'interno della funzione `MeBRboot` si pone il parametro `parameter` pari a `"beta"` o `"phi"` in base a cui si seleziona automaticamente la funzione per la riscaturatura, e, solo per il metodo `boot`, è possibile porre l'opzione `ic = TRUE`, di default si otterrà un intervallo di confidenza di livello 0.95.

Infine, si è implementata una funzione per l'Algoritmo 4 che estende il metodo proposto da Benussi et al. (2026) al contesto semi-parametrico. Nello specifico nell'ambito dei GLM con interesse al parametro ϕ , se i parametri β sono noti o stimati si ha che

$\theta = \phi$ e la funzione $s(\cdot) = 1$. La funzione `MeBRben`, presente in Appendice, ha come argomenti in entrata la stima iniziale (`theta_hat`), il numero di replicazioni bootstrap (`M`), il campione originale (`datas`) e la funzione di stima del parametro d'interesse (`est_fn`).

Capitolo 4

Applicazione a casi studio

Definito il quadro teorico della riduzione della distorsione in mediana e i vantaggi dell'applicazione nei modelli lineari generalizzati, l'obiettivo del corrente capitolo è valutare il comportamento dei metodi impliciti ed espliciti in alcuni casi studio, rispetto alle stime di massima verosimiglianza e basata sul metodo dei momenti per il parametro di dispersione ϕ .

In particolare, si analizzano due casi studio: il primo utilizza i dati **medpar** (Hilbe, 2011) presenti nel pacchetto **COUNT** di R. Tale dataset verrà trattato nei primi due paragrafi, inizialmente analizzato tramite una distribuzione gamma e successivamente tramite una quasi-Poisson. Il secondo caso studio, invece, è volto ad analizzare i dati **GasolineYield** (Prater, 1956) del pacchetto **betareg**, questi verranno adattati attraverso una regressione beta. Il codice R relativo a ciascuno si trova nell'Appendice A.

4.1 Caso studio medpar: gamma

I dati **medpar** si riferiscono a 1495 pazienti di procedure chirurgiche cardiovascolari in Arizona nel 1991: la variabile risposta **los** (*length of stay*) rappresenta la durata di degenza. Le covariate utilizzate per il modello sono **hmo** (*Health Maintenance Organization*), una variabile binaria che indica se il soggetto fosse coperto da assicurazione sanitaria o abbia pagato privatamente, **white**, una variabile binaria che riflette se il paziente si identificasse come caucasico, e **type**, una variabile a 3 categorie che rappresenta il tipo di ammissione ospedaliera (programmata, urgente o di emergenza).

Il modello adattato a tali dati è quello gamma

$$Y_i \sim Ga(\alpha, \gamma) \text{ in cui } E(Y_i) = \alpha/\gamma \text{ e } Var(Y_i) = \alpha/\gamma^2,$$

dunque $\mu_i = g^{-1}(\mathbf{x}_i\beta) = \alpha/\gamma$ e $\phi = 1/\alpha$. Per tale modello GLM, la funzione di legame canonica è il reciproco, tuttavia nel presente caso studio si utilizza il logaritmo al fine di limitare μ_i a valori positivi, facilitare l'interpretazione e stabilizzare la varianza; la variabile risposta, infatti, presenta asimmetria positiva e dunque la trasformata logaritmica comprime la variabilità in modo tale che la distribuzione gamma sia più appropriata.

Dato l'interesse al parametro di dispersione ϕ , pari al reciproco del parametro di forma α , i coefficienti di regressione β sono fissati alla loro stima di massima verosimiglianza e vengono confrontati i seguenti stimatori di ϕ : massima verosimiglianza $\hat{\phi}$, non distorti in media $\hat{\phi}^*$ e in mediana $\tilde{\phi}^*$ basati sulla correzione della funzione score, e non distorto in mediana ottenuto da tecnica bootstrap $\tilde{\phi}^{**}$. Quest'ultimo è ottenuto applicando l'Algoritmo 2 con $N = 300$ e $m = 150$, e stima iniziale quella di massima verosimiglianza.

Oltre alle stime puntuali, i primi tre approcci consentono di ricavare gli standard error per via analitica. In aggiunta, per tutte e quattro le metodologie considerate gli standard error sono stati stimati anche tramite una procedura di *double bootstrap* basata su 200 replicazioni.

I risultati sono presentati nella Tabella 4.1. Si osserva come tutti i metodi di stima producano lo stesso risultato, con differenze che si evidenziano solo a partire dal terzo o quarto decimale. La stima puntuale del parametro di dispersione ϕ si stabilizza attorno al valore 0.58. Ciò indica la presenza di sottodispersione nei dati **medpar** rispetto al modello esponenziale, equivalente a quello gamma con $\phi = 1$. L'uguaglianza dei risultati da metodi differenti può essere ricondotta alla numerosità campionaria: se questa è elevata i termini di distorsione di ordine superiore si azzerano. Di conseguenza, l'effetto delle correzioni della distorsione in media e in mediana risultano trascurabili.

Dal confronto degli intervalli di confidenza alla Wald, emerge che quelli calcolati tramite l'approccio bootstrap risultano essere lievemente più ampi rispetto a quelli basati su standard error analitici. Tale differenza evidenzia il fatto che, sul campione osservato, lo stimatore bootstrap della variabilità rileva una maggiore incertezza attorno alle stime puntuali, catturando la variabilità campionaria senza dipendere dalle assunzioni distributive alla base delle approssimazioni analitiche asintotiche.

TABELLA 4.1: Confronto delle stime di ϕ e relativi intervalli di confidenza al 95% per i diversi metodi considerati per l'adattamento di **medpar** con gamma

Stimatore	Stima	IC alla Wald 95% (Analitico)		IC alla Wald 95% (Bootstrap)	
		Limite inferiore	Limite superiore	Limite inferiore	Limite superiore
$\hat{\phi}$	0.5806	0.5423	0.6189	0.5364	0.6248
$\hat{\phi}^*$	0.5810	0.5427	0.6193	0.5368	0.6252
$\tilde{\phi}^*$	0.5812	0.5429	0.6195	0.5369	0.6254
$\tilde{\phi}^{**}$	0.5808	—	—	0.5366	0.6250

4.2 Caso studio medpar: quasi-Poisson

Data la natura di conteggio della variabile risposta di **medpar** essa può essere analizzata sia come variabile continua tramite una distribuzione gamma, sia come discreta: in questo paragrafo si applica ai medesimi dati un modello quasi-Poisson, specificato tramite le prime due relazioni dei momenti

$$E(Y_i) = \mu_i \quad \text{e} \quad Var(Y_i) = \phi \mu_i,$$

dunque $\mu_i = g^{-1}(\mathbf{x}_i\beta)$, dove il legame canonico per la distribuzione Poisson, e quindi per quella quasi-Poisson, è il logaritmo $\mu_i = \exp(\mathbf{x}_i\beta)$. Rispetto alla modellazione precedente si noti come la relazione tra la variabilità della risposta e il valore atteso sia lineare, al contrario della distribuzione gamma in cui $Var(Y_i) = \phi \mu_i^2$.

In questo contesto, i coefficienti di regressione β vengono fissati alle stime ottenute mediante le equazioni di quasi-verosimiglianza. L'analisi si concentra sul confronto tra diversi stimatori del parametro di dispersione ϕ : il metodo dei momenti ($\tilde{\phi}$), le correzioni proposte da Farrington ($\tilde{\phi}_{Fa}$) e Fletcher ($\tilde{\phi}_{Fl}$), lo stimatore basato sulla devianza residua ($\tilde{\phi}_D$), lo stimatore robusto MAD ($\tilde{\phi}_{MAD}$), lo stimatore non distorto basato sul bootstrap secondo l'Algoritmo 3 ($\tilde{\phi}^{**}$) e, infine, l'Algoritmo 4 ($\hat{\phi}$). Per tali ultimi due si utilizza come stima iniziale quella basata sul metodo dei momenti, si fissano i parametri dell'approccio bootstrap a $N = 300$ e $R = 200$, e il numero di replicazioni per la correzione esplicita analitica pari a $M = 500$. Inoltre, al fine di ottenere una misura di incertezza per tutti i metodi si svolgono 200 replicazioni annidate mediante una procedura di *double bootstrap*.

I risultati così ottenuti sono raccolti nella Tabella 4.2. Si evidenzia la presenza di sovradisersione nei dati modellati. I metodi che non svolgono una correzione per la distorsione e si basano sui residui quadratici di Pearson convergono a valori simili,

attorno a 6.26. Al contrario, le procedure di riduzione della distorsione presentano uno spostamento verso valori maggiori. Tale comportamento è coerente con la presenza di distorsione negativa tipica del metodo dei momenti in campioni finiti. La marcata differenza tra gli stimatori dei momenti e gli approcci basati su devianza e MAD riflette la presenza di asimmetria e code pesanti nella distribuzione della variabile risposta. Lo stimatore MAD, robusto in quanto costruito mediante mediana, riduce l'influenza delle osservazioni estreme. Al contrario, lo stimatore della devianza risente dell'asimmetria della coda destra e delle discrepanze tra conteggio reale e predetto sui valori piccoli, fornendo un valore di dispersione più elevato.

Gli intervalli di confidenza al 95% risultano stabili. Quelli costruiti a partire da $\tilde{\phi}^{**}$ e $\check{\phi}$ mostrano una traslazione verso destra causata dalla correzione della stima puntuale, tuttavia non presentano un aumento dell'ampiezza rispetto quelli costruiti su $\tilde{\phi}$. In contrasto, si osserva come gli stimatori della devianza e MAD ottengano una varianza osservata inferiore rispetto agli altri approcci e dunque intervalli di confidenza più stretti. Infine, l'intervallo di confidenza costruito secondo la procedura descritta al paragrafo 3.2.1 ha estremi (5.5461, 7.2957), questo ha anche ampiezza inferiore rispetto a quello basato sulle replicazioni bootstrap di $\tilde{\phi}^{**}$ e non è centrato sulla stima puntuale.

TABELLA 4.2: Confronto delle stime di ϕ e relativi intervalli di confidenza al 95% per i diversi metodi considerati per l'adattamento di `medpar` con quasi-Poisson

Stimatore	Stima	IC alla Wald 95%	
		Limite inferiore	Limite superiore
$\tilde{\phi}$	6.2604	5.3305	7.1903
$\tilde{\phi}_{Fl}$	6.2578	5.3286	7.1871
$\tilde{\phi}_{Fa}$	6.2600	5.3302	7.1898
$\tilde{\phi}_D$	5.4649	4.8763	6.0534
$\tilde{\phi}_{MAD}$	3.9395	3.2777	4.6013
$\tilde{\phi}^{**}$	6.3429	5.3918	7.2940
$\check{\phi}$	6.3557	5.4109	7.3005

4.3 Caso studio GasolineYield: beta

I dati `GasolineYield` si riferiscono a 32 osservazioni la cui variabile risposta è `yield`, la percentuale di petrolio greggio trasformata in benzina dopo la distillazione e il frazionamento. Le covariate comprese nel modello sono `temp`, una variabile continua che indica la temperatura in gradi Fahrenheit alla quale tutta la benzina è evaporata, e `batch`, un fattore a 10 livelli che rappresenta la combinazione di condizioni relative alla densità del petrolio greggio (in gradi API), pressione di vaporizzazione del petrolio greggio (in lbf/in²) e temperatura a cui 10% della benzina è evaporata (in gradi F).

A tali dati si è adattata una regressione beta

$$Y_i \sim \text{Beta}(\phi\mu_i, \phi(1 - \mu_i)) \text{ in cui } \mu_i = g^{-1}(\mathbf{x}_i\beta),$$

dove $g(\cdot) = \text{logit}(\cdot)$ e $\beta = (\beta_0, \dots, \beta_{10})$.

La Tabella 4.3 raccoglie le stime dei parametri β e ϕ ottenute tramite massima verosimiglianza e quelle con riduzione della distorsione in media e in mediana basate sulla correzione della funzione score. Per questi si sono ottenuti gli intervalli alla Wald al livello 95% sia analitici che tramite una procedura di *double bootstrap* composta da 200 replicazioni. La Tabella 4.4 comprende le stime dei parametri con riduzione della distorsione tramite aggiustamento misto e tecnica bootstrap, affiancati dai relativi intervalli di confidenza calcolati tramite gli standard error ottenuti con *double bootstrap*.

I risultati presentati nell'articolo Kosmidis et al. (2020) riguardanti la strategia di aggiustamento misto si ereditano anche per la distribuzione beta. Tuttavia, essa non è compresa nella funzione `brglm2` e dunque si procede con un ciclo iterativo, inizializzato alle stime di massima verosimiglianza, in cui si corregge la distorsione in media dei coefficienti β e in mediana del parametro ϕ fino a convergenza.

Per ciò che riguarda la riduzione della distorsione in mediana tramite tecnica bootstrap, si applica l'Algoritmo 5 che, a partire dalle stime di massima verosimiglianza, ottiene una stima congiunta non distorta in mediana attraverso uno schema a tre fasi: correzione in mediana dei β , correzione in mediana di ϕ dato il nuovo vettore β e rifinitura di β dato ϕ corretto. Ciascuna di queste fasi viene svolta con l'utilizzo della funzione `MeBRboot` nella sua versione parametrica e opzioni $N = 100$ e $m = 80$.

Dal confronto dei risultati ottenuti con i cinque metodi emerge come le stime dei coefficienti di regressione risultano stabili al variare delle tecniche utilizzate, così come gli intervalli di confidenza. Ciò indica che, nonostante la numerosità limitata, la stima del valore atteso delle osservazioni non sia soggetta a distorsione importante né in media, né in mediana. Al contrario, la stima del parametro di dispersione presenta evidenti

differenze: la stima di massima verosimiglianza è quella con valore più alto, mentre i metodi di riduzione della distorsione mostrano una correzione concorde.

TABELLA 4.3: Confronto delle stime di β e ϕ e relativi intervalli di confidenza al 95% per alcuni metodi considerati per l'adattamento di **GasolineYield** con beta

Metodo	Parametro	Stima	IC alla Wald 95% (Analitico)		IC alla Wald 95% (Bootstrap)	
			Limite inferiore	Limite superiore	Limite inferiore	Limite superiore
Massima	β_0	-6.1596	-6.5169	-5.8022	-6.5263	-5.7929
Verosimiglianza	β_1	1.7277	1.5293	1.9261	1.5137	1.9418
	β_2	1.3226	1.0915	1.5537	1.0725	1.5727
	β_3	1.5723	1.3447	1.7999	1.3404	1.8042
	β_4	1.0597	0.8591	1.2603	0.8662	1.2532
	β_5	1.1338	0.9308	1.3367	0.9316	1.3359
	β_6	1.0402	0.8323	1.2480	0.8268	1.2535
	β_7	0.5437	0.3298	0.7576	0.3227	0.7647
	β_8	0.4959	0.2824	0.7094	0.2750	0.7168
	β_9	0.3858	0.1534	0.6182	0.1506	0.6210
	β_{10}	0.0110	0.0102	0.0118	0.0101	0.0118
	ϕ	440.2784	224.6282	655.9286	—	—
Riduzione in media	β_0	-6.1417	-6.6040	-5.6794	-6.5076	-5.7758
	β_1	1.7233	1.4664	1.9801	1.5097	1.9368
	β_2	1.3186	1.0196	1.6176	1.0693	1.5680
	β_3	1.5673	1.2728	1.8619	1.3360	1.7987
	β_4	1.0568	0.7971	1.3164	0.8638	1.2498
	β_5	1.1302	0.8675	1.3929	0.9286	1.3319
	β_6	1.0371	0.7681	1.3062	0.8243	1.2500
	β_7	0.5424	0.2657	0.8191	0.3220	0.7628
	β_8	0.4945	0.2182	0.7707	0.2741	0.7148
	β_9	0.3846	0.0837	0.6855	0.1500	0.6192
	β_{10}	0.0109	0.0099	0.0120	0.0101	0.0117
	ϕ	261.0378	133.2136	388.8619	—	—
Riduzione in mediana	β_0	-6.1437	-6.5908	-5.6966	-6.5097	-5.7777
	β_1	1.7243	1.4759	1.9726	1.5106	1.9379
	β_2	1.3191	1.0299	1.6082	1.0695	1.5686
	β_3	1.5680	1.2832	1.8528	1.3365	1.7995
	β_4	1.0575	0.8064	1.3086	0.8644	1.2507
	β_5	1.1310	0.8770	1.3850	0.9292	1.3328
	β_6	1.0378	0.7776	1.2979	0.8248	1.2508
	β_7	0.5428	0.2752	0.8104	0.3222	0.7633
	β_8	0.4948	0.2277	0.7620	0.2743	0.7153
	β_9	0.3846	0.0936	0.6755	0.1498	0.6194
	β_{10}	0.0109	0.0099	0.0119	0.0101	0.0118
	ϕ	279.4087	142.5832	416.2342	—	—

TABELLA 4.4: Confronto delle stime di β e ϕ e relativi intervalli di confidenza al 95% per alcuni metodi considerati per l'adattamento di **GasolineYield** con beta

Metodo	Parametro	Stima	IC alla Wald 95% (Bootstrap)	
			Limite inferiore	Limite superiore
Aggiustamento misto	β_0	-6.1417	-6.5086	-5.7748
	β_1	1.7233	1.5092	1.9373
	β_2	1.3186	1.0684	1.5688
	β_3	1.5673	1.3354	1.7993
	β_4	1.0568	0.8633	1.2503
	β_5	1.1302	0.9281	1.3324
	β_6	1.0371	0.8238	1.2505
	β_7	0.5424	0.3214	0.7635
	β_8	0.4945	0.2736	0.7153
	β_9	0.3846	0.1494	0.6198
	β_{10}	0.0109	0.0101	0.0118
	ϕ	279.4087	—	—
Bootstrap	β_0	-6.1503	-6.5170	-5.7837
	β_1	1.7163	1.5023	1.9303
	β_2	1.3134	1.0633	1.5635
	β_3	1.5881	1.3562	1.8200
	β_4	1.0545	0.8610	1.2480
	β_5	1.1275	0.9254	1.3296
	β_6	1.0374	0.8241	1.2508
	β_7	0.5462	0.3251	0.7672
	β_8	0.4945	0.2736	0.7153
	β_9	0.3839	0.1487	0.6191
	β_{10}	0.0109	0.0101	0.0117
	ϕ	280.2154	—	—

Capitolo 5

Studi di simulazione

Il presente capitolo illustra tre studi di simulazione volti a valutare l'efficacia e le proprietà a campioni finiti delle metodologie presentate nei capitoli precedenti. In primo luogo si presentano le misure di confronto dei metodi analizzati, in seguito queste vengono applicate per l'analisi di tre scenari distinti che comprendono sia il contesto parametrico che semi-parametrico. Per ciascuno sono state svolte $S = 10000$ simulazioni e considerate diverse combinazioni del vero valore del parametro d'interesse e della numerosità campionaria. In Appendice A si trovano i relativi codici R.

5.1 Misure di confronto empirico

Dato un numero S di simulazioni, il valore vero del parametro scalare d'interesse θ e un vettore di S stime ottenute da campioni simulati $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_S)$, le misure di confronto utilizzate sono le seguenti:

1. *PU*: percentuale di sottostima (*Percentage of Underestimation*), indica la percentuale di simulazioni in cui lo stimatore ha prodotto un valore inferiore al parametro vero

$$PU = \frac{1}{S} \sum_{s=1}^S \mathbb{I}(\hat{\theta}_s < \theta) \times 100;$$

2. *B*: distorsione in mediana stimata (*Estimated Median Bias*), misura quanto lo stimatore si allontana in mediana dal vero valore del parametro

$$B = \text{Me}(\hat{\theta}_1, \dots, \hat{\theta}_S) - \theta;$$

3. *RMSE*: radice dell'errore quadratico medio (*Root Mean Squared Error*), rappresenta la radice quadrata della misura *MSE* e ne permette una maggiore interpretabilità

$$RMSE = \sqrt{\frac{1}{S} \sum_{s=1}^S (\hat{\theta}_s - \theta)^2};$$

4. *MAE*: errore assoluto medio (*Mean Absolute Error*), rappresenta la media delle deviazioni assolute delle stime rispetto al parametro vero ed è una misura di accuratezza robusta rispetto agli outlier

$$MAE = \frac{1}{S} \sum_{s=1}^S |\hat{\theta}_s - \theta|;$$

5. B_{mean}^2/SD^2 : incremento relativo di *MSE* a causa della distorsione in media, indica quanta parte dell'errore quadratico medio sia dovuta a questa rispetto alla varianza

$$\frac{B_{mean}^2}{SD^2} = \frac{\left(\frac{1}{S} \sum_{s=1}^S \hat{\theta}_s\right) - \theta}{\frac{1}{S-1} \sum_{s=1}^S (\hat{\theta}_s - \bar{\hat{\theta}})^2};$$

6. *C*: copertura empirica (*Coverage*), percentuale di simulazioni in cui l'intervallo di confidenza $(\theta_{inf}, \theta_{sup})$ contiene il valore vero del parametro

$$C = \frac{1}{S} \sum_{s=1}^S \mathbb{I}(\theta_{inf}^s \leq \theta \leq \theta_{sup}^s) \times 100.$$

Di particolare interesse nello studio degli stimatori non distorti in mediana sono le metriche *PU*, *B* e la copertura empirica *C* associata agli intervalli di confidenza costruiti tramite la procedura HulC, tuttavia l'insieme delle restanti misure consente di ottenere un quadro più completo.

Per quanto riguarda le metriche di errore (*B*, *RMSE* e *MAE*) lo stimatore preferibile sarà quello che minimizza tali valori. Per la percentuale di sottostima *PU*, la centratura in mediana ideale corrisponde a una convergenza verso 50% infatti la prossimità a tale valore è una misura della non distorsione in mediana. Per l'incremento relativo B^2/SD^2 il valore ottimale è 0, con una soglia limite solitamente fissata a 0.1, valori prossimi allo zero, infatti, indicano che l'errore dello stimatore è dominato dalla sola variabilità campionaria e l'effetto della distorsione è trascurabile. Infine, il metodo preferibile in base alla copertura empirica *C* è quello con valore simile al livello di confidenza nominale, fissato al 95% sia per gli intervalli di tipo Wald che per quelli HulC.

Infine, si tenga conto che in seguito i valori di B , $RMSE$ e MAE sono moltiplicati per 100 per facilitare l'interpretazione.

5.2 Simulazioni nel caso parametrico: gamma

Lo schema di generazione delle covariate fa riferimento ai dati `medpar`, si utilizzano le proporzioni empiriche congiunte delle variabili binarie `white` e `hmo` come parametri di una distribuzione bernulliana bivariata in modo tale che venga inclusa nel disegno anche la dipendenza tra queste. Dunque la funzione di probabilità congiunta per la coppia di variabili $\mathbf{x}_i = (x_{i1}, x_{i2})$, $i = 1, \dots, n$, con $x_{i1} \in \{0, 1\}$ (`white`) e $x_{i2} \in \{0, 1\}$ (`hmo`) è

$$p(\mathbf{x}_i) = \mathbb{P}(X_{i1} = x_{i1}, X_{i2} = x_{i2}) = p_{11}^{x_{i1}x_{i2}} p_{10}^{x_{i1}(1-x_{i2})} p_{01}^{(1-x_{i1})x_{i2}} p_{00}^{(1-x_{i1})(1-x_{i2})}$$

dove $p_{00} = 0.077$, $p_{01} = 0.008$, $p_{10} = 0.763$ e $p_{11} = 0.152$.

Per ciascuna delle dimensioni campionarie considerate ($n \in \{50, 100, 200\}$), viene generato un singolo scenario affinché la struttura così ottenuta sia mantenuta fissa all'interno dello studio per ciascun valore di n . Tramite tale struttura si definisce la matrice del disegno X e si calcola il vettore $\mu = \exp(X\beta)$.

Questo schema di simulazioni permette che la componente sistematica del modello rimanga stabile al variare delle replicazioni e che sia costruita in maniera coerente con i dati di riferimento. Inoltre, la variabile risposta è simulata dalla distribuzione gamma

$$Y \sim Ga\left(\frac{1}{\phi}, \frac{1}{\phi\mu}\right), \text{ con } \mu = g^{-1}(X\hat{\beta}) \text{ e } \phi \in \{0.3, 0.7\},$$

dove $g(\cdot) = \log(\cdot)$ e $\hat{\beta}$ sono le stime di massima verosimiglianza. La scelta di tali valori di ϕ è dovuta al fatto che la stima basata sul metodo dei momenti sui dati `medpar` originali è pari a 0.777.

La Tabella 5.1 raccoglie i risultati, gli stimatori del parametro di dispersione confrontati sono di massima verosimiglianza ($\hat{\phi}$, paragrafo 1.3), corretto per la distorsione in media ($\hat{\phi}^*$, paragrafo 2.2.1) e in mediana ($\tilde{\phi}^*$, paragrafo 2.2.1) basati sulla correzione della funzione score, e non distorto in mediana attraverso la procedura bootstrap ($\tilde{\phi}^{**}$, paragrafo 2.2.2). Per questo si è utilizzata la versione parametrica con stima iniziale la massima verosimiglianza e parametri $N = 200$ e $m = 80$. In aggiunta, per 2000 delle $S = 10000$ simulazioni si è svolto un *double bootstrap* per il calcolo degli intervalli di confidenza alla Wald empirici (*Wald E*) attraverso 50 simulazioni annidate; in questo

caso i parametri per il metodo di riduzione della distorsione in mediana basato sul bootstrap sono ridotti a $N = 50$ e $m = 30$. Gli altri intervalli di confidenza considerati sono quelli Wald basati sugli standard error analitici (*Wald A*), disponibili solo per le prime tre tecniche, e quelli ottenuti dalla procedura HulC.

I risultati mostrano come lo stimatore di massima verosimiglianza ($\hat{\phi}$) tenda a sottostimare sistematicamente il vero valore del parametro, tale distorsione in mediana negativa (B) risulta particolarmente evidente nei campioni di ridotta numerosità e cresce all'aumentare di ϕ . La presenza di distorsione è confermata anche dalla percentuale di sottostima (PU) che per $\hat{\phi}$ si ritrova sempre al di sopra del 50%, con un massimo di 57% nello scenario più critico. L'utilizzo di metodologie per la riduzione della distorsione risulta efficace nell'ottenere la corretta centratura in mediana. In particolare, lo stimatore $\tilde{\phi}^{**}$ garantisce una riduzione della distorsione rispetto allo stimatore $\hat{\phi}$, per entrambi i valori fissati di ϕ . Tuttavia, lo stimatore $\tilde{\phi}^*$ garantisce risultati migliori al crescere della dimensione campionaria in quanto la distorsione in mediana diminuisce fino a diventare trascurabile e il valore di PU ad esso associato converge alla soglia ottima di 50%. Inoltre emerge che lo stimatore corretto in media $\hat{\phi}^*$ mostra valori di percentuale di sottostima compresi tra 50.96% e 52.94% e dunque risulta complessivamente più soddisfacente rispetto alla correzione in mediana $\tilde{\phi}^{**}$ (PU tra 51.8% e 54.6%). Ciò potrebbe essere dovuto alla distribuzione campionaria dello stimatore di massima verosimiglianza per il parametro di dispersione nella famiglia gamma. Essendo questa asimmetrica a destra, la correzione della distorsione in media sposta lo stimatore a valori superiori così riequilibrando la massa di probabilità verso il vero valore, risultando dunque in PU prossime a 50%.

Per ciò che riguarda le misure di efficienza $RMSE$ e MAE , lo stimatore $\tilde{\phi}^{**}$ offre il miglior compromesso tra la riduzione della distorsione in mediana e la varianza campionaria. A differenza di altri stimatori non distorti che subiscono un lieve incremento dell'errore quadratico medio, $\tilde{\phi}^{**}$ mantiene valori di $RMSE$ e MAE inferiori rispetto a $\hat{\phi}^*$ e $\tilde{\phi}^*$ in tutte le combinazioni analizzate. L'indice B_{mean}^2/SD^2 si trova per tutti gli stimatori al di sotto della soglia critica di 0.10, confermando che per tutte le metodologie l'errore è dominato dalla variabilità campionaria.

Per quanto riguarda la validità dell'inferenza intervallare al livello di confidenza nominale del 95%, il confronto tra le metodologie evidenzia come gli intervalli di confidenza di Wald basati su standard error empirici ottenuti tramite bootstrap annidati mantengono una copertura empirica stabile (tra 94.50% e 95.65%) in tutti gli scenari studiati e per tutti i metodi di stima. Al contrario, gli intervalli di Wald calcolati con gli standard error analitici presentano problemi di sottocopertura in campioni di bassa numerosità,

particolarmente per $\phi = 0.3$. La copertura si avvicina al livello nominale solo al crescere della numerosità campionaria. Inoltre, l'applicazione della procedura HulC per la costruzione di intervalli di confidenza mostra un vantaggio dell'utilizzo dello stimatore $\tilde{\phi}^{**}$ per campioni limitati, mentre in combinazione con gli altri stimatori e al crescere di n la metodologia HulC tende ad essere conservativa.

Infine, l'analisi al variare della numerosità conferma il comportamento asintotico delle metodologie studiate infatti da $n = 50$ ad $n = 200$ si osserva un dimezzamento dei valori di $RMSE$ e MAE dovuto unicamente alla riduzione della distorsione residua in tutti gli stimatori.

TABELLA 5.1: Confronto dei metodi per la riduzione della distorsione in mediana di ϕ su simulazioni da distribuzione gamma

n	ϕ	Stimatore	PU	B	$RMSE$	MAE	B_{mean}^2/SD^2	C_{WaldE}	C_{WaldA}	C_{HulC}
50	0.3	$\hat{\phi}$	55.99	-0.8635	5.7108	4.5892	0.0095	94.75	90.70	91.00
		$\hat{\phi}^*$	52.01	-0.2739	5.7966	4.6318	0.0001	94.75	92.35	96.10
		$\tilde{\phi}^*$	49.50	0.0663	5.8637	4.6707	0.0043	94.50	93.25	96.95
		$\tilde{\phi}^{**}$	53.38	-0.4770	5.7548	4.6073	0.0009	94.95	—	94.95
	0.7	$\hat{\phi}$	56.95	-2.2588	12.7104	10.2143	0.0138	94.65	91.70	90.80
		$\hat{\phi}^*$	52.94	-0.9403	12.8480	10.2576	0.0001	94.80	93.00	96.60
		$\tilde{\phi}^*$	50.88	-0.2810	12.9488	10.3055	0.0016	94.80	93.50	97.25
		$\tilde{\phi}^{**}$	54.62	-1.4167	12.7807	10.2282	0.0027	94.90	—	95.10
100	0.3	$\hat{\phi}$	54.97	-0.4756	4.0426	3.2123	0.0080	94.85	93.15	93.95
		$\hat{\phi}^*$	51.68	-0.1802	4.0664	3.2241	0.0003	95.00	93.65	96.80
		$\tilde{\phi}^*$	50.13	-0.0122	4.0862	3.2364	0.0006	95.20	94.00	97.55
		$\tilde{\phi}^{**}$	52.82	-0.2813	4.0548	3.2162	0.0018	95.05	—	96.00
	0.7	$\hat{\phi}$	54.51	-0.9870	8.9679	7.1756	0.0068	94.90	94.15	94.90
		$\hat{\phi}^*$	51.59	-0.3239	9.0160	7.1977	0.0001	94.75	95.00	96.95
		$\tilde{\phi}^*$	50.00	0.0039	9.0511	7.2181	0.0008	94.70	94.85	96.65
		$\tilde{\phi}^{**}$	52.65	-0.5947	8.9911	7.1842	0.0013	94.85	—	96.30
200	0.3	$\hat{\phi}$	53.26	-0.2413	2.8359	2.2646	0.0020	95.20	94.75	96.20
		$\hat{\phi}^*$	51.37	-0.0933	2.8470	2.2685	0.0001	94.70	95.35	97.30
		$\tilde{\phi}^*$	50.14	-0.0097	2.8555	2.2724	0.0014	94.85	95.25	97.55
		$\tilde{\phi}^{**}$	52.00	-0.1456	2.8428	2.2669	0.0001	94.90	—	97.20
	0.7	$\hat{\phi}$	53.16	-0.4700	6.3112	5.0497	0.0032	95.65	95.20	96.30
		$\hat{\phi}^*$	50.96	-0.1380	6.3286	5.0577	0.0000	95.55	95.60	97.05
		$\tilde{\phi}^*$	49.88	0.0252	6.3412	5.0653	0.0005	95.55	95.75	97.15
		$\tilde{\phi}^{**}$	51.83	-0.2808	6.3180	5.0525	0.0006	95.60	—	97.15

5.3 Confronto di metodi semi-parametrici: simulazione del modello binomiale negativo

Il successivo studio di simulazione segue nuovamente lo schema di generazione della matrice X precedentemente descritto, tuttavia le numerosità campionarie prese in considerazione sono $n \in \{30, 50, 100\}$. In questo caso la variabile risposta è simulata come una binomiale negativa

$$Y \sim \text{Bineg}\left(\mu, \frac{\mu}{\phi - 1}\right),$$

al fine di ottenere una variabile casuale caratterizzata da $E(Y) = \mu$ e $\text{Var}(Y) = \phi\mu$ dove $\mu = \exp(X\hat{\beta})$, $\phi \in \{4, 7\}$ e $\hat{\beta}$ sono le stime di massima verosimiglianza. Tali valori di ϕ sono stati selezionati in quanto sul dataset originario la stima basata sul metodo dei momenti risulta pari a 7.740.

I metodi confrontati sono lo stimatore basato sul metodo dei momenti ($\tilde{\phi}$, paragrafo 1.3), le varianti di Fletcher ($\tilde{\phi}_{Fl}$, paragrafo 1.4) e Farrington ($\tilde{\phi}_{Fa}$, paragrafo 1.4), gli stimatori con riduzione della distorsione in mediana tramite procedura bootstrap ($\tilde{\phi}^{**}$, paragrafo 2.2.2) e metodologia esplicita ($\check{\phi}$, paragrafo 2.2.3). Per quelli con correzione in mediana si utilizza la versione semi-parametrica con stima iniziale quella del metodo dei momenti e parametri $N = 300$, $R = 200$ e $M = 500$. In maniera equivalente al paragrafo precedente si è svolto un *double bootstrap* di numerosità 50 per 2000 delle $S = 10000$ simulazioni al fine di ottenere gli standard error empirici di ciascun metodo per la costruzione degli intervalli di confidenza alla Wald. A questo scopo, i parametri sono fissati pari a $N = 100$ e $R = 80$ per lo stimatore $\tilde{\phi}^{**}$ e $M = 150$ per $\tilde{\phi}$. In aggiunta, gli intervalli sono stati costruiti tramite la procedura HulC e quella presentata al paragrafo 3.2.1, esclusivamente per $\tilde{\phi}^{**}$.

Dall'analisi dell'indice di distorsione in mediana B per gli stimatori $\tilde{\phi}$, $\tilde{\phi}_{Fl}$ e $\tilde{\phi}_{Fa}$ si osserva come tendano a sottostimare il vero valore del parametro, portando quindi a valori sempre negativi ed elevati. Allo stesso modo si evidenzia che la percentuale di sottostima PU sia sistematicamente oltre la soglia ottima di 50%, con un valore massimo che raggiunge il 63.28%. Al contrario, lo stimatore $\tilde{\phi}^{**}$ si dimostra efficace nell'ottenere la corretta centratura in mediana. Nello scenario più critico di campioni a dimensione ridotta ed elevata sovradisersione, mentre gli stimatori classici presentano una forte distorsione, $\tilde{\phi}^{**}$ riduce la distorsione a un valore quasi nullo ($B = 0.073$). Inoltre, per la stessa combinazione di parametri la misura PU risulta essere 48.77% e converge alla soglia ottima al crescere di n .

In aggiunta, lo stimatore $\tilde{\phi}^{**}$ tiene conto del compromesso varianza-distorsione in

media, infatti la metrica B_{mean}^2/SD^2 si mantiene largamente inferiore alla soglia limite (tra 0.015 e 0.074), confermando come la correzione applicata non introduca instabilità. In secondo luogo, tale correzione della distorsione non comporta aumenti della variabilità, ciò è dimostrato confrontando le metriche di errore globale $RMSE$ e MAE di $\tilde{\phi}^{**}$: i valori di $RMSE$ sono paragonabili a quelli dello stimatore senza correzione applicata $\tilde{\phi}$, specialmente all'aumento di n .

La valutazione della capacità di copertura empirica rispetto al livello nominale del 95% mostra come gli intervalli di confidenza alla Wald calcolati per $\tilde{\phi}^{**}$ ottengano una copertura stabile e precisa per ciascuna delle combinazioni in analisi, compresa tra 95.65% e 97.10%. Gli intervalli HulC per gli stimatori $\tilde{\phi}$, $\tilde{\phi}_{Fl}$, $\tilde{\phi}_{Fa}$ risultano essere soggetti a sottocopertura, in particolare in presenza di sovradisersione elevata, a causa dell'assunzione sottostante al metodo di non distorsione in mediana. Per tale motivo, la combinazione della procedura HulC con lo stimatore $\tilde{\phi}^{**}$ permette di ottenere una copertura simile alla nominale e stabile in qualsiasi scenario, sebbene lievemente conservativa (tra 96.15% e 97.35%). Infine, gli intervalli di confidenza calcolati sulla base della proposta del paragrafo 3.2.1, calcolabile solamente per $\tilde{\phi}^{**}$, ottengono una copertura ($C_{\tilde{\phi}^{**}}$) poco conservativa. Gli intervalli così costruiti risultano dunque stretti ma efficienti in quanto la costruzione è inclusa nell'algoritmo di stima e non richiede ulteriore costo computazionale. Al fine di aumentare la copertura empirica, una possibile soluzione è l'incremento del numero di replicazioni R .

Complessivamente, l'analisi delle metriche evidenzia un comportamento dello stimatore $\tilde{\phi}^{**}$ simile a quello dello stimatore $\tilde{\phi}$. L'evidenza empirica mostra quindi la bontà del nuovo metodo $\tilde{\phi}^{**}$, che si presenta come una alternativa soddisfacente per la riduzione della distorsione in mediana facilmente applicabile ai contesti semi-parametrici anche negli scenari considerati di numerosità limitata e sovradisersione elevata.

TABELLA 5.2: Confronto dei metodi per la riduzione della distorsione in mediana di ϕ su simulazioni da binomiale negativa

n	ϕ	Stimatore	PU	B	$RMSE$	MAE	B_{mean}^2/SD^2	C_{WaldE}	C_{HulC}	$C_{\check{\phi}^{**}}$
30	4	$\check{\phi}$	58.60	-0.2530	1.1976	0.9414	0.0063	94.95	91.65	—
		$\check{\phi}_{Fl}$	58.63	-0.2528	1.1976	0.9414	0.0064	90.30	91.75	—
		$\check{\phi}_{Fa}$	58.62	-0.2527	1.1976	0.9414	0.0063	94.90	91.70	—
		$\check{\phi}^{**}$	44.23	0.1791	1.4069	1.0595	0.0742	96.40	97.35	86.76
		$\check{\phi}$	45.64	0.1431	1.4008	1.0486	0.0592	93.55	96.95	—
	7	$\check{\phi}$	63.24	-0.7015	2.2910	1.7949	0.0231	95.05	89.00	—
		$\check{\phi}_{Fl}$	63.28	-0.7028	2.2902	1.7947	0.0235	91.30	89.20	—
		$\check{\phi}_{Fa}$	63.25	-0.7024	2.2910	1.7950	0.0231	95.00	89.05	—
		$\check{\phi}^{**}$	48.77	0.0728	2.6027	1.9513	0.0342	97.10	96.80	81.37
		$\check{\phi}$	50.11	-0.0044	2.6515	1.9533	0.0265	94.05	95.35	—
	4	$\check{\phi}$	57.48	-0.1689	0.9359	0.7370	0.0047	95.35	94.15	—
		$\check{\phi}_{Fl}$	57.48	-0.1689	0.9359	0.7370	0.0047	92.90	94.15	—
		$\check{\phi}_{Fa}$	57.48	-0.1689	0.9359	0.7370	0.0047	95.35	94.15	—
		$\check{\phi}^{**}$	45.37	0.1126	1.0540	0.8018	0.0518	96.65	96.85	87.99
		$\check{\phi}$	46.14	0.0951	1.0503	0.7967	0.0449	94.35	97.00	—
50	7	$\check{\phi}$	59.82	-0.4674	1.8317	1.4457	0.0130	94.90	90.60	—
		$\check{\phi}_{Fl}$	59.82	-0.4674	1.8317	1.4457	0.0130	92.15	90.80	—
		$\check{\phi}_{Fa}$	59.82	-0.4674	1.8317	1.4457	0.0130	94.90	90.65	—
		$\check{\phi}^{**}$	49.22	0.0374	2.0445	1.5519	0.0282	96.45	96.45	84.15
		$\check{\phi}$	49.67	0.0115	2.0549	1.5517	0.0245	94.10	96.65	—
	4	$\check{\phi}$	56.60	-0.1099	0.6656	0.5287	0.0052	95.20	94.95	—
		$\check{\phi}_{Fl}$	56.61	-0.1101	0.6656	0.5287	0.0053	92.25	94.90	—
		$\check{\phi}_{Fa}$	56.61	-0.1099	0.6656	0.5287	0.0052	95.20	94.95	—
		$\check{\phi}^{**}$	47.36	0.0458	0.7151	0.5544	0.0250	95.65	96.85	89.34
		$\check{\phi}$	47.84	0.0385	0.7115	0.5520	0.0219	94.20	96.75	—
100	7	$\check{\phi}$	58.51	-0.2814	1.3054	1.0358	0.0113	95.30	92.90	—
		$\check{\phi}_{Fl}$	58.53	-0.2815	1.3054	1.0358	0.0113	92.95	92.85	—
		$\check{\phi}_{Fa}$	58.51	-0.2814	1.3054	1.0358	0.0113	95.30	92.90	—
		$\check{\phi}^{**}$	49.77	0.0057	1.4073	1.0854	0.0146	96.15	96.85	86.93
		$\check{\phi}$	50.07	-0.0019	1.4056	1.0828	0.0131	94.25	97.15	—

5.4 Confronto di metodi semi-parametrici: simulazione del modello quasi-Poisson

L'ultimo studio di simulazione utilizza lo stesso schema di generazione della matrice del disegno X del precedente.

La variabile risposta è simulata dalla distribuzione quasi-Poisson tramite l'utilizzo della variabile casuale Poisson moltiplicata per un fattore ϕ (Linero, 2026, paragrafo 5.1),

$$Y \sim \phi Po\left(\frac{\mu}{\phi}\right),$$

dove $\mu = \exp(X\hat{\beta})$ con $\hat{\beta}$ le stime di massima verosimiglianza, $n \in \{30, 50, 100\}$ e $\phi \in \{4, 7\}$. Sebbene si tratti di un metodo equivalente di simulazione dei campioni da distribuzione binomiale negativa in quanto portano agli stessi valore atteso e varianza, questa struttura comporta un supporto della variabile Y discreto e composto da multipli di ϕ . Per tale motivo i metodi di stima del parametro di dispersione di massima verosimiglianza e del metodo dei momenti non sono applicabili.

Dunque, gli stimatori confrontati sono quello basato sulla devianza ($\tilde{\phi}_D$, paragrafo 1.4), MAD ($\tilde{\phi}_{MAD}$, paragrafo 1.4), con riduzione della distorsione in mediana sia basata su tecniche bootstrap ($\tilde{\phi}^{**}$, paragrafo 2.2.2) che su espansioni asintotiche ($\check{\phi}$, paragrafo 2.2.3). Gli ultimi due sono utilizzati nella loro estensione al caso semi-parametrico con parametri $N = 300, R = 200$ per $\tilde{\phi}^{**}$ e $M = 500$ per $\check{\phi}$. In aggiunta, data l'impossibilità del calcolo dello stimatore $\tilde{\phi}$, utilizzano come stimatore iniziale quello basato sulla devianza.

Sono state svolte $S = 10000$ simulazioni e per 2000 di queste è stata utilizzata una struttura annidata di bootstrap per l'ottenimento degli standard error empirici sulla base di 50 replicazioni, in cui i metodi sono stati applicati con parametri $N = 100, R = 80$ e $M = 150$. Ciò ha permesso il calcolo delle coperture degli intervalli di confidenza alla Wald, oltre a quello basato sulla tecnica HulC e quello presentato al paragrafo 3.2.1, applicabile solo al metodo di stima corretto per la distorsione in mediana tramite procedure bootstrap.

Lo studio di simulazione ha portato a risultati empirici complessivamente non soddisfacenti, evidenziando criticità che intaccano la validità di tutti gli stimatori considerati in questo contesto.

La misura di centratura in mediana PU mostra come anche i metodi di correzione della distorsione in mediana ($\tilde{\phi}^{**}$ e $\check{\phi}$) non raggiungano il valore ottimo del 50%, in particolare essi tendono a sovrastimare il vero parametro ($PU < 50\%$), mostrando un

peggioramento in presenza di sovradisersione alta. Inoltre, si osserva un abbassamento della percentuale di sottostima al crescere della numerosità campionaria: nello scenario con $\phi = 7$, per $\tilde{\phi}^{**}$ il valore di PU scende dal 15.54% per $n = 30$ fino all'8.00% per $n = 100$.

Tale comportamento persiste anche nell'analisi del rapporto tra la distorsione in media al quadrato e la varianza (B_{mean}^2/SD^2). Nonostante la componente di distorsione risulti non trascurabile per tutti i metodi analizzati ($B_{mean}^2/SD^2 > 0.10$), per le stime $\tilde{\phi}^{**}$ e $\check{\phi}$ l'indice raddoppia al passaggio da $n = 30$ a $n = 100$, per entrambi i valori di ϕ considerati.

Infine, analizzando i risultati dell'inferenza intervallare con livello nominale fissato al 95%, si osserva come sia l'approccio di tipo Wald sia la procedura HulC non portino a esiti soddisfacenti, indipendentemente dallo stimatore considerato; essi infatti mostrano una sistematica sottocopertura, raggiungendo il 73.90% per il caso Wald e il 79% per quello HulC nello scenario $\phi = 7$ e $n = 100$ per lo stimatore $\check{\phi}$.

Il fatto che le proprietà asintotiche e l'efficacia dei metodi di correzione della distorsione in mediana peggiorino all'aumentare della dimensione campionaria n costituisce una criticità. Tale comportamento è riconducibile alla combinazione tra la natura discreta del supporto, vincolato a multipli di ϕ , e l'elevata frequenza di zeri, causata dalla riduzione del parametro μ/ϕ della distribuzione Poisson per valori elevati di ϕ .

Tuttavia, risulta importante evidenziare che tali limitazioni non sono imputabili allo stimatore $\tilde{\phi}^{**}$. La bontà del metodo è infatti confermata sia dai risultati positivi riscontrati nel paragrafo precedente, sia dalle prestazioni altrettanto non soddisfacenti degli altri stimatori considerati nello studio corrente. Ciò dimostra come le problematiche derivino dallo schema di generazione del modello quasi-Poisson, in quanto tale struttura impone vincoli che alterano la distribuzione campionaria degli stimatori.

Conclusioni

L'obiettivo di questa tesi è stato quello di generalizzare l'algoritmo di riduzione della distorsione in mediana basato su tecniche bootstrap proposto da Cabrera & Watson (1997) a contesti semi-parametrici. Infatti, in mancanza di un'esplicita assunzione distributiva e di una funzione di verosimiglianza di riferimento, i metodi analitici di correzione in mediana presenti in letteratura risultano inapplicabili. Tuttavia, la generalizzazione a contesti non strettamente parametrici è necessaria per superare le problematiche legate a errata specificazione del modello sottostante e campioni di dimensioni limitate. Per tale motivo, si è proposto un algoritmo basato su procedure bootstrap non parametriche e su una tecnica di calibrazione, ispirata a quella applicata nei problemi di verifica d'ipotesi. Inoltre, si è proposta una nuova metodologia per la costruzione di intervalli di confidenza basata sullo stimatore con riduzione di distorsione in mediana tramite bootstrap. Nonostante tale stimatore sia applicabile in diversi ambiti, il presente lavoro utilizza come parametro di studio quello di dispersione.

Nel Capitolo 4 le applicazioni a casi studio hanno valutato lo stimatore corretto in mediana basato su tecniche bootstrap ($\tilde{\phi}^{**}$) su dati reali sia nel contesto parametrico gamma che semi-parametrico quasi-Poisson. Per quanto riguarda il caso parametrico, poiché nei casi analizzati si ha un'elevata numerosità campionaria, l'effetto della correzione in mediana sulla stima puntuale risulta trascurabile in quanto tutti i metodi convergono nello stesso valore. Tuttavia, in confronto ai metodi analitici, il metodo di Cabrera & Watson (1997) sembra offrire un vantaggio nella valutazione dell'incertezza, ottenendo intervalli di confidenza alla Wald più ampi e che tengono conto della sottostima della variabilità tipica delle approssimazioni asintotiche. Nel contesto semi-parametrico, in cui si usano le sole relazioni tra i momenti, l'approccio di correzione basato sul bootstrap riduce la distorsione in mediana negativa tipica del metodo dei momenti applicato a campioni finiti. La stima del parametro di dispersione è traslata verso valori più alti, ottenendo dunque intervalli di confidenza aderenti ai dati reali.

Dagli studi di simulazione presentati nel Capitolo 5 si evidenziano le differenze

del comportamento campionario dello stimatore $\tilde{\phi}^{**}$ in contesti parametrici e semi-parametrici. Nel caso parametrico gamma, lo stimatore in esame si dimostra efficace nel bilanciare la correzione della distorsione in mediana con la variabilità, ottenendo indici di errore $RMSE$ e MAE inferiori allo stimatore con correzione analitica basata sulla funzione score, tuttavia quest'ultimo mostra una convergenza più rapida al valore ottimo di percentuale di sottostima PU , rendendo l'utilizzo dello stimatore $\tilde{\phi}^{**}$ di minore interesse per la stima puntuale nell'ambito parametrico.

Nei contesti semi-parametrici, il comportamento dello stimatore con riduzione della distorsione in mediana basato sul bootstrap varia a seconda dello schema di simulazione della variabile risposta: nel caso della binomiale negativa, $\tilde{\phi}^{**}$ risulta efficace. Al contrario, tale stimatore fallisce nel modello quasi-Poisson. Tuttavia, tale criticità è legata al supporto discreto e vincolato e all'elevata frequenza di zeri, in particolare all'aumentare di n , anziché essere un limite dello stimatore $\tilde{\phi}^{**}$. Nel caso della binomiale negativa, $\tilde{\phi}^{**}$ si dimostra efficace nella riduzione della distorsione negativa in mediana degli stimatori con cui è confrontato: nello scenario più critico ($n = 30, \phi = 7$), la correzione in mediana tramite bootstrap porta a PU prossima alla soglia ottima, mantenendo gli indici $RMSE$ e MAE a valori ridotti. Inoltre, $\tilde{\phi}^{**}$ garantisce intervalli di confidenza con coperture empiriche prossime al livello nominale; tale risultato si conferma anche per la procedura HulC che, basandosi sull'assunzione di non distorsione in mediana, avvalora l'efficacia dello stimatore nella riduzione di tale distorsione. In aggiunta, per quanto riguarda la nuova tecnica proposta per l'inferenza intervallare basata sullo stimatore $\tilde{\phi}^{**}$, nonostante questa produca intervalli stretti e caratterizzati da sottocopertura, essa presenta il vantaggio di non richiedere simulazioni annidate, risultando ottenibile direttamente dall'algoritmo di stima.

Infine, possibili sviluppi futuri potrebbero riguardare l'estensione dell'algoritmo di riduzione della distorsione in mediana tramite bootstrap a strutture di dati più complesse e l'applicazione della metodologia ad altri ambiti empirici.

Appendice A

Funzioni d'implementazione degli algoritmi

Stimatore non distorto in mediana basato sul bootstrap

L'implementazione è svolta seguendo gli Algoritmi 2 e 3, e i dettagli trattati nei capitoli precedenti. Si precisa che nell'implementazione della spline di lisciamento si utilizza il metodo della *Cross-Validation* (CV) per selezionare automaticamente il parametro di lisciamento, ciò risulta essere un metodo automatizzato e robusto in quanto decide il grado di lisciatura della funzione mediana basandosi sui dati ed evitando sovradattamento.

```
MeBRboot <- function(theta_hat, theta_L, theta_U, N = 300, R = 350, datas,
                      m = 100, sim_data_fn = NULL, est_fn = NULL,
                      method = c("model", "boot"), parameter = c("beta", "phi"),
                      ic = FALSE, conf.level = 0.95){
  method <- match.arg(method)
  parameter <- match.arg(parameter)

  if (parameter == "phi"){
    g <- function(x) log(x)
    g_inv <- function(x) exp(x)
    filter_fn <- function(x) x[!is.na(x) & x > 0.001]
  } else if (parameter == "beta"){
    g <- function(x) x
    g_inv <- function(x) x
    filter_fn <- function(x) x[!is.na(x)]
  }

  if (method == "boot"){
    if(is.null(datas)|is.null(est_fn))
      stop("Per il metodo \"boot\" necessario specificare \"datas\" e
           \"est_fn\"")
  }
```

```

    est_boot <- function(data_boot, k){return(est_fn(data_boot[k,]))}
    t_sim_nogrid <- boot::boot(datas, est_boot, R)$t
    t_sim_nogrid <- filter_fn(t_sim_nogrid)
  }

  theta_grid <- seq(from = theta_L, to = theta_U, length.out = N)
  h_star <- rep(NA, N)

  for (i in 1:N){
    theta_i <- theta_grid[i]
    if (method == "model"){
      if(is.null(sim_data_fn)|is.null(est_fn))
        stop("Per il metodo \"model\" necessario specificare \"sim_data_fn\" e
              \"est_fn\"")
      t_sim <- replicate(m, est_fn(sim_data_fn(theta_i)))
      t_sim <- filter_fn(t_sim)
      h_star[i] <- median(t_sim)
    }
    else if (method == "boot"){
      t_sim <- g_inv( g(t_sim_nogrid) + (g(theta_i) - g(theta_hat)) )
      h_star[i] <- median(t_sim)
    }
  }

  smooth_fit <- smooth.spline(x = g(theta_grid), y = g(h_star), cv = T)
  h_pred_g <- predict(smooth_fit, g(theta_grid))$y
  g_theta_new <- approx(x = h_pred_g, y = g(theta_grid), xout = g(theta_hat),
                       rule = 2)$y
  theta_out <- g_inv(g_theta_new)

  if (ic == T){
    limite_inf <- limite_sup <- NA
    do.ic <- T
    if (conf.level <= 0 || conf.level >= 1) {
      warning("Il parametro \"conf.level\" deve essere compreso tra 0 e 1")
      do.ic <- F
    }
    if(method == "model") {
      warning("Il calcolo dell'IC richiede il metodo \"boot\"")
      do.ic <- F
    }
  }

  if(do.ic){
    a <- 1 - conf.level
    t_sim_grid <- g_inv(g(t_sim_nogrid)+(g(theta_out)-g(theta_hat)) )
    limite_inf <- unname(quantile(t_sim_grid, probs = a / 2, na.rm = T))
    limite_sup <- unname(quantile(t_sim_grid, probs = 1 - a / 2, na.rm = T))
  }
  return(c(Stima = theta_out, Infe = limite_inf, Sup = limite_sup))
}

```

```

    return(theta_out)
}

```

Stimatore non distorto in mediana basato su espansioni asintotiche

Il codice seguente implementa l'Algoritmo 4, ossia il metodo introdotto da Benussi et al. (2026).

```

MeBRben <- function(theta_hat, M = 500, datas, est_fn){
  if(is.null(datas)|is.null(est_fn))
    stop("Necessario specificare \"datas\" e \"est_fn\"")
  n_oss <- nrow(datas)

  t_sim_hat <- replicate(M, {
    idx_boot <- sample(1:n_oss, size = n_oss, replace = TRUE)
    data_boot <- datas[idx_boot, ]
    est_fn(data_boot)
  })
  t_sim_hat <- t_sim_hat[!is.na(t_sim_hat) & t_sim_hat > 0.001]
  if (length(t_sim_hat) < (M * 0.5)) {
    warning("Più del 50% delle simulazioni ha prodotto valori non validi")
  }

  mu_emp <- mean(t_sim_hat)
  B_hat <- mu_emp - theta_hat
  K2_hat <- var(t_sim_hat)
  K3_hat <- mean((t_sim_hat - mu_emp)^3)

  theta_out <- theta_hat - B_hat + (1/6) * (K3_hat / K2_hat)
  return(theta_out)
}

```

Codice per le applicazioni a casi studio

Caso studio medpar: gamma

```
library(COUNT); library(brglm2); library(boot)
data("medpar", package = "COUNT")

mod_gamma <- glm(los ~ hmo+white+factor(type), data = medpar, family =
                Gamma(link = "log"), method = "brglmFit", type = "ML")
beta_smv <- coef(mod_gamma)
X_matrice <- model.matrix(mod_gamma)
mu_smv <- as.vector(exp(X_matrice %*% beta_smv))
n_oss <- length(mu_smv)

# Funzione per l'utilizzo di MeBRboot:
# -> sim_gamma simula nuovo dataset con covariate uguali e y simulata da gamma
# con parametri shape=1/phi e scale=mu*phi dove mu ottenuto dalle stime smv di
# beta e phi sarà il theta_i della grid
sim_gamma <- function(phi_corrente) {
  param_shape <- 1 / phi_corrente
  param_scale <- mu_smv * phi_corrente
  y_simulato <- rgamma(n = n_oss, shape = param_shape, scale = param_scale)
  y_simulato[y_simulato <= 0] <- 0.001
  df_simulato <- medpar
  df_simulato$los <- y_simulato
  return(df_simulato)
}

# -> est è funzione di stima di phi con massima verosimiglianza, si esplicita
# la funzione di log-verosimiglianza negativa per minimizzarla al fine di
# ridurre il tempo computazionale
est <- function(data_est) {
  y_curr <- data_est$los
  risultato <- optimize(f = nll_gamma, interval = c(0.001, 5), y = y_curr,
                       mu = mu_smv)
  return(risultato$minimum)
}

nll_gamma <- function(phi, y, mu) {
  if (phi <= 0) return(Inf)
  shape <- 1 / phi
  loglik <- (-length(y)*lgamma(shape))-(shape*sum(log(mu*phi)))+
    ((shape-1)*sum(log(y)))-sum(y/(mu*phi))
  return(-loglik)
}

# Funzione per la stima di phi con metodi di confronto (con standard error
# analitico)
est_gamma_phi <- function(data_est, metodo_phi = c("ML", "AS_mean", "AS_median")){
  metodo_phi <- match.arg(metodo_phi)
```

```

offset_covariate <- as.vector(X_matrice[, -1] %*% beta_smv[-1])
mod <- glm(los ~ 1 + offset(offset_covariate), family = Gamma(link = "log"),
          data = data_est, method = "brglmFit", type = metodo_phi)
se_as <- tryCatch({
  matrice_vcov <- suppressWarnings(vcov(mod, model = "full"))
  sqrt(matrice_vcov[nrow(matrice_vcov), ncol(matrice_vcov)])
}, error = function(e) return(NA))

return(c(stima = mod$dispersion[1], se = se_as))
}

phi_ML <- est_gamma_phi(medpar, "ML")
phi_mean <- est_gamma_phi(medpar, "AS_mean")
phi_median <- est_gamma_phi(medpar, "AS_median")
phi_MeBR <- MeBRboot(theta_hat = phi_ML[1], theta_L = 0.1, theta_U = 3, N = 300,
                    m = 150, datas = medpar, sim_data_fn = sim_gamma,
                    est_fn = est, method = "model", parameter = "phi")

```

Caso studio medpar: quasi-Poisson

```

data("medpar", package = "COUNT")

# Funzione est per la stima di phi basata sul metodo dei momenti per MeBRboot
est <- function(data_est){
  mod <- glm(los ~ hmo + white + factor(type), family = quasipoisson,
            data = data_est)
  return(summary(mod)$dispersion)
}

# Funzione per la stima di phi con metodi di Fletcher e Farrington
mod <- glm(los ~ hmo + white + factor(type), family = quasipoisson,
          data = medpar)
est_FF <- function(model) {
  n <- nobs(model)
  p <- length(coef(model))
  phi_P <- summary(model)$dispersion
  mu <- model$fitted.values
  y <- model$y
  s_bar <- mean((1 / mu) * (y - mu))
  phi_Fl <- phi_P / (1 + s_bar)
  phi_Fa <- phi_P - (n * s_bar) / (n - p)
  return(c(Fletcher = phi_Fl, Farrington = phi_Fa))
}

# Funzione per la stima di phi con MAD
mod2 <- glm(los ~ hmo + white + factor(type), family = poisson, data = medpar)

```

```

est_MAD <- function(model){
  mu_hat <- model$fitted.values
  res_p <- (model$y - mu_hat) / sqrt(mu_hat)
  me_res <- median(res_p, na.rm = T)
  phi_hat <- (median(abs(res_p - me_res), na.rm=T) / 0.6745)^2
  return(phi_hat)
}

phi_est <- est(medpar)
phi_Fl <- est_FF(mod) ["Fletcher"]
phi_Fa <- est_FF(mod) ["Farrington"]
phi_dev <- mod2$deviance / mod2$df.residual
phi_MAD <- est_MAD(mod2)
phi_boot <- MeBRboot(theta_hat = phi_est, theta_L = 1, theta_U = 10, N = 300,
                     R = 200, datas = medpar, est_fn = est, method="boot",
                     parameter="phi", ic=T)
phi_B <- MeBRben(theta_hat = phi_est, M = 500, datas = medpar, est_fn = est)

```

Caso studio GasolineYield: beta

```

library("betareg")
data("GasolineYield", package = "betareg")

# Stima di massima verosimiglianza con std error analitici
smv <- betareg(yield ~ batch + temp, data = GasolineYield)
ML <- smv$coefficients
ML_se <- sqrt(diag(smv$vcov))

# Stima con riduzione della distorsione in media con std error analitici
br <- betareg(yield ~ batch + temp, data = GasolineYield, type="BR")
BR <- coefficients(br)
BR_se <- sqrt(diag(br$vcov))

# Stima con riduzione della distorsione in mediana con std error analitici
source("mbrbeta.R")
mebr <- mbrbetareg(yield ~ batch + temp, data = GasolineYield, type="medianBR")
MeBR <- mebr$coefficients
MeBR_se <- sqrt(diag(mebr$vcov))

# Stima con algoritmo sequenziale per la riduzione della distorsione in
# mediana con MeBRboot
# -> Inizializzazione delle stime a smv
fit_base <- betareg(yield ~ temp + batch, data = GasolineYield)
beta_ml <- coef(fit_base, model = "mean")
phi_ml <- coef(fit_base, model = "precision")
n_beta <- length(beta_ml)

```

```

beta_names <- names(beta_ml)
# -> Correzione vettore beta con phi fissato:
# si utilizzano funzioni sim_fn e est_fn specifiche a phi fissato ma che
# valgano per tutte le beta e si calcola un delta in maniera automatizzata
# pari a metà della stima di massima verosimiglianza del parametro per
# ottenere il range di MeBRboot
beta_ini <- numeric(n_beta)
names(beta_ini) <- beta_names
for (k in 1:n_beta){
  # Funzione sim_fn simula da beta con mu calcolato dove beta sono smv tranne
  # la k-esima che è theta_i dalla griglia
  sim_fn_beta <- function(theta_i){
    df_sim <- GasolineYield
    X <- model.matrix(fit_base)
    b_sim <- beta_ml
    b_sim[k] <- theta_i
    eta <- X %*% b_sim
    mu <- 1 / (1 + exp(-eta))
    df_sim$yield <- rbeta(length(mu), mu*phi_ml, (1-mu)*phi_ml)
    return(df_sim)
  }
  est_fn_beta <- function(df_sim){
    fit <- betareg(yield ~ temp + batch, data = df_sim)
    return(coef(fit, model = "mean")[k])
  }
  delta <- if (abs(beta_ml[k]) == 0) 0.5 else abs(beta_ml[k]) * 0.5
  beta_ini[k] <- MeBRboot(theta_hat = beta_ml[k], theta_L = beta_ml[k]-delta,
                        theta_U = beta_ml[k]+delta, N = 100, m = 80,
                        sim_data_fn = sim_fn_beta, est_fn = est_fn_beta,
                        method = "model", parameter = "beta")
}
# -> Correzione di phi con beta fissato a valori corretti
sim_fn_phi <- function(phi_i){
  df_sim <- GasolineYield
  X <- model.matrix(fit_base)
  eta <- X %*% beta_ini
  mu <- 1 / (1 + exp(-eta))
  df_sim$yield <- rbeta(length(mu), mu*phi_i, (1-mu)*phi_i)
  return(df_sim)
}
est_fn_phi <- function(df_sim){
  fit <- betareg(yield ~ temp + batch, data = df_sim)
  return(coef(fit, model = "precision"))
}
phi_final <- MeBRboot(theta_hat = phi_ml, theta_L = phi_ml*0.5, theta_U =
                    phi_ml*1.5, N = 100, m = 80, sim_data_fn = sim_fn_phi,
                    est_fn = est_fn_phi, method = "model", parameter = "phi")
# -> Ristima dei beta con phi fissato ai valori corretti:
# necessario riscrivere le funzioni ausiliarie di MeBRboot perchè phi variato
beta_final <- numeric(n_beta)

```

```

names(beta_final) <- beta_names
for (k in 1:n_beta) {
  sim_fn_beta_final <- function(theta_i){
    df_sim <- GasolineYield
    X <- model.matrix(fit_base)
    b_sim <- beta_ini
    b_sim[k] <- theta_i
    eta <- X %*% b_sim
    mu <- 1 / (1 + exp(-eta))
    df_sim$yield <- rbeta(length(mu), mu*phi_final, (1-mu)*phi_final)
    return(df_sim)
  }
  est_fn_beta <- function(df_sim){
    fit <- betareg(yield ~ temp + batch, data = df_sim)
    return(coef(fit, model = "mean")[k])
  }
  delta <- if (abs(beta_ini[k]) == 0) 0.5 else abs(beta_ini[k]) * 0.5
  beta_final[k] <- MeBRboot(theta_hat = beta_ini[k], theta_L =
                           beta_ini[k]-delta, theta_U = beta_ini[k]+delta,
                           N = 100, m = 80, sim_data_fn = sim_fn_beta_final,
                           est_fn = est_fn_beta, method = "model",
                           parameter = "beta")
}
mebrboot <- c(beta_final, phi_final)

```

Codice per gli studi di simulazione

Misure empiriche per il confronto

```
# Sia theta_sim i valori del parametro d'interesse ottenuti dalle S
# simulazioni per uno degli scenari studiati e sia phi_val il vero valore
B <- mean(theta_sim) - phi_val
RMSE <- sqrt(mean((theta_sim - phi_val)^2))
B2_SD2 <- (B^2) / var(theta_sim)
BMe <- median(theta_sim) - phi_val
PU <- mean(theta_sim < phi_val) * 100
MAE <- mean(abs(theta_sim - phi_val))
```

Caso parametrico: gamma

```
library(COUNT); library(brglm2)
data("medpar", package = "COUNT")

# Combinazioni di numerosità e phi valutate
vett_n    <- c(50, 100, 200)
vett_phi  <- c(0.3, 0.7)

# Estrazione delle stima di mu di beta e delle probabilità congiunte nel
# dataset originale
mod_vero  <- glm(los ~ hmo + white, family = Gamma(link = "log"),
                 data = medpar)
beta_vero <- coef(mod_vero)
n_00     <- 115
n_01     <- 12
n_10     <- 1141
n_11     <- 227
n_tot    <- n_00 + n_01 + n_10 + n_11
prob_c   <- c(n_00, n_01, n_10, n_11) / n_tot

# Creazione delle strutture per gli 6 scenari affinché si abbia dataset,
# vettore mu e matrice X data una n in vett_n
# Tuttavia rmultinom restituisce una matrice 4x1 con conteggi totali per
# cella dunque è necessario dividere i conteggi ripetendo i valori teorici
# delle covariate e mescolare l'ordine delle righe per non avere pattern
# ordinati, il set.seed assicura che i 3 campioni estratti siano sempre uguali
strutture_fisse <- list()
set.seed(12345)
for(n_struttura in vett_n){
  conteggi_celle <- rmultinom(n = 1, size = n_struttura, prob = prob_c)
```

```

vettore_white <- rep(c(0, 0, 1, 1), times = conteggi_celle)
vettore_hmo <- rep(c(0, 1, 0, 1), times = conteggi_celle)
ordine_casuale <- sample(1:n_struttura)
vettore_white <- vettore_white[ordine_casuale]
vettore_hmo <- vettore_hmo[ordine_casuale]
dati_fissi <- data.frame(los = rep(0, n_struttura), white = vettore_white,
                        hmo = vettore_hmo)
X_fissa <- model.matrix(los ~ hmo + white, data = dati_fissi)
mu_fissato <- as.vector(exp(X_fissa %*% beta_vero))
strutture_fisse[[n_struttura]] <- list(dati = dati_fissi, mu = mu_fissato,
                                       X = X_fissa)
}

# Simulazione di nuovo dataset con covariate da struttura e variabile risposta
# simulata tramite gamma con phi in vett_phi
sim_gamma <- function(n_target, phi_target, mu_scen, data_struttura){
  nuovo_data <- data_struttura
  y_sim <- rgamma(n = n_target, shape = 1/phi_target, scale =
                mu_scen*phi_target)
  y_sim[y_sim <= 0] <- 0.001
  nuovo_data$los <- y_sim
  return(nuovo_data)
}

# Funzione per la stima di phi secondo i metodi di brglm2
est_gamma <- function(data_est, metodo_phi, X_scen){
  offset_covariate <- as.vector(X_scen[,-1] %*% beta_vero[-1])
  mod <- glm(los ~ 1 + offset(offset_covariate), family = Gamma(link = "log"),
            data = data_est, method = "brglmFit", type = metodo_phi)
  return(mod$dispersion[1])
}

# Funzione ausiliaria di MeBRboot per la stima di phi con smu in maniera più
# efficiente
nll_gamma <- function(phi, y, mu){
  if (phi <= 0) return(Inf)
  loglik <- (-length(y) * lgamma(1/phi)) - ((1/phi) * sum(log(mu*phi))) +
    (((1/phi)-1) * sum(log(y))) - sum(y/(mu*phi))
  return(-loglik)
}

est_mebr <- function(data_est, mu_scen) {
  risultato <- optimize(f = nll_gamma, interval = c(0.001, 5), y =
                    data_est$los, mu = mu_scen)
  return(risultato$minimum)
}

# Applicazione per S simulazioni
scenari <- expand.grid(phi = vett_phi, n = vett_n)
matrice_risultati <- list()

```

```

for (row in 1:nrow(scenari)){
  n_corr <- scenari$n[row]
  phi_corr <- scenari$phi[row]

  struttura_scen <- strutture_fisse[[n_corr]]
  struttura_dati_fissa <- struttura_scen$dati
  mu_vettore_fissato <- struttura_scen$mu
  X_scen <- struttura_scen$X

  matrice_repliche <- matrix(NA, nrow = S, ncol = 4)
  colnames(matrice_repliche) <- c("ML", "AS_Mean", "AS_Median", "MeBRboot")

  for (s in 1:S){
    dati_sim <- sim_gamma(n_target = n_corr, phi_target = phi_corr, mu_scen =
                        mu_vettore_fissato, data_struttura =
                        struttura_dati_fissa)
    phi_ML <- est_gamma(dati_sim, "ML", X_scen)
    phi_mean <- est_gamma(dati_sim, "AS_mean", X_scen)
    phi_median <- est_gamma(dati_sim, "AS_median", X_scen)

    phi_MeBR <- MeBRboot(theta_hat = phi_ML, theta_L = 0.05, theta_U = 3.5,
                        N = 200, m = 80, method = "model", parameter = "phi",
                        sim_data_fn = function(theta_i){sim_gamma(n_target = n_corr,
                        phi_target = theta_i, mu_scen = mu_vettore_fissato,
                        data_struttura = struttura_dati_fissa)},
                        est_fn = function(dati_sim_mebr){est_mebr(dati_sim_mebr, mu_scen =
                        mu_vettore_fissato)})

    matrice_repliche[s,] <- c(phi_ML, phi_mean, phi_median, phi_MeBR)
  }

  nome_cella <- paste0("Scenario n ", n_corr, ", phi ", phi_corr)
  matrice_risultati[[nome_cella]] <- matrice_repliche
}

```

Caso semi-parametrico: quasi-Poisson

```

library(COUNT); library(boot)
data("medpar", package = "COUNT")

# Combinazioni di numerosità campionaria e valore di phi
vett_n <- c(30, 50, 100)
vett_phi <- c(4, 7)

# Stime di beta di massima verosimiglianza e probabilità congiunte da dati

```

```

# originali
mod_vero <- glm(los ~ hmo + white, family = quasipoisson, data = medpar)
beta_vero <- coef(mod_vero)
n_00 <- 115
n_01 <- 12
n_10 <- 1141
n_11 <- 227
n_tot <- n_00 + n_01 + n_10 + n_11
prob_c <- c(n_00, n_01, n_10, n_11) / n_tot

# Ottenimento di 6 strutture fisse che comprendano un dataset e vettore mu per
# ciascuno degli scenari
strutture_fisse <- list()
set.seed(12345)
for(n_struttura in vett_n){
  conteggi_celle <- rmultinom(n = 1, size = n_struttura, prob = prob_c)
  vettore_white <- rep(c(0, 0, 1, 1), times = conteggi_celle)
  vettore_hmo <- rep(c(0, 1, 0, 1), times = conteggi_celle)
  ordine_casuale <- sample(1:n_struttura)
  vettore_white <- vettore_white[ordine_casuale]
  vettore_hmo <- vettore_hmo[ordine_casuale]
  dati_fissi <- data.frame(los = rep(0, n_struttura), white = vettore_white,
                           hmo = vettore_hmo)
  X_fissa <- model.matrix(los ~ hmo + white, data = dati_fissi)
  mu_fissato <- as.vector(exp(X_fissa %*% beta_vero))
  strutture_fisse[[n_struttura]] <- list(dati = dati_fissi, mu = mu_fissato)
}

# Simulazione di dati con covariate fisse e risposta generata da  $\phi * Po(\mu/\phi)$ 
sim_data <- function(n_target, phi_target, mu_scen, data_struttura){
  nuovo_data <- data_struttura
  nuovo_data$los <- rpois(n_target, lambda = mu_scen/phi_target) * phi_target
  return(nuovo_data)
}

# Funzioni di stima di  $\phi$  basata su devianza residua e MAD
est_dev <- function(data_est){
  mod <- glm(los ~ hmo + white, family = poisson, data = data_est)
  phi_hat <- mod$deviance / mod$df.residual
  return(phi_hat)
}
est_MAD <- function(data_est){
  mod <- glm(los ~ hmo + white, family = poisson, data = data_est)
  mu_hat <- mod$fitted.values
  res_p <- (data_est$los - mu_hat) / sqrt(mu_hat)
  me_res <- median(res_p, na.rm = T)
  phi_hat <- (median(abs(res_p - me_res), na.rm = T) / 0.6745)^2
  return(phi_hat)
}

```

```
# Applicazione a S simulazioni: oltre alle 4 stime si hanno gli estremi
# dell'intervallo di confidenza interno a MeBRboot
matrice_risultati <- list()
for (n_corr in vett_n){
  for (phi_corr in vett_phi){
    struttura_scen <- strutture_fisse[[n_corr]]
    struttura_dati_fissa <- struttura_scen$dati
    mu_vettore_fissato <- struttura_scen$mu

    matrice_repliche <- matrix(NA, nrow = S, ncol = 6)
    colnames(matrice_repliche) <- c("MeBR_Boot", "Deviance", "MAD", "Benussi",
                                   "lower", "upper")

    for (s in 1:S){
      dati_sim <- sim_data(n_target = n_corr, phi_target = phi_corr,
                           mu_scen = mu_vettore_fissato, data_struttura =
                           struttura_dati_fissa)

      phi_dev <- est_dev(dati_sim)
      phi_MAD <- est_MAD(dati_sim)

      phi_M <- MeBRboot(theta_hat = phi_dev, theta_L = 1, theta_U =
                        phi_corr*2.5, N = 300, R = 200, datas = dati_sim,
                        est_fn = est_dev, method = "boot", parameter = "phi",
                        ic = T)

      phi_B <- MeBRben(theta_hat = phi_dev, M = 500, datas = dati_sim, est_fn =
                        est_dev)

      matrice_repliche[s,] <- c(phi_M["Stima"], phi_dev, phi_MAD, phi_B,
                               phi_M["Infe"], phi_M["Sup"])
    }

    nome_cella <- paste0("Scenario n ", n_corr, ", phi ", phi_corr)
    matrice_risultati[[nome_cella]] <- matrice_repliche
  }
}
```

Caso semi-parametrico: binomiale negativa

```
library(COUNT); library(boot)
data("medpar", package = "COUNT")

# Scenari presi in considerazione
vett_n <- c(30, 50, 100)
vett_phi <- c(4, 7)
```

```

# Estrazione di stime di beta di mv e probabilità congiunte dai dati originali
mod_vero <- glm(los ~ hmo + white, family = quasipoisson, data = medpar)
beta_vero <- coef(mod_vero)
n_00 <- 115
n_01 <- 12
n_10 <- 1141
n_11 <- 227
n_tot <- n_00 + n_01 + n_10 + n_11
prob_c <- c(n_00, n_01, n_10, n_11) / n_tot

# Creazione delle strutture che includano dati e vettore mu per ciascuna
# combinazione di n e phi
strutture_fisse <- list()
set.seed(12345)
for(n_struttura in vett_n){
  conteggi_celle <- rmultinom(n = 1, size = n_struttura, prob = prob_c)
  vettore_white <- rep(c(0, 0, 1, 1), times = conteggi_celle)
  vettore_hmo <- rep(c(0, 1, 0, 1), times = conteggi_celle)
  ordine_casuale <- sample(1:n_struttura)
  vettore_white <- vettore_white[ordine_casuale]
  vettore_hmo <- vettore_hmo[ordine_casuale]
  dati_fissi <- data.frame(los = rep(0, n_struttura), white = vettore_white,
                           hmo = vettore_hmo
  )
  X_fissa <- model.matrix(los ~ hmo + white, data = dati_fissi)
  mu_fissato <- as.vector(exp(X_fissa %*% beta_vero))
  strutture_fisse[[n_struttura]] <- list(dati = dati_fissi, mu = mu_fissato)
}

# Ottenere nuovo dataset con le covariate fissate dalla struttura e la
# variabile risposta simulata da binomiale negativa
sim_data_fissato <- function(n_target, phi_target, mu_scen, data_struttura){
  nuovo_data <- data_struttura
  nuovo_data$los <- rnbino(n_target, size = mu_scen/(phi_target-1),
                           mu = mu_scen)
  return(nuovo_data)
}

# Funzioni degli stimatori di confronto: metodo dei momenti, Fletcher e
# Farrington
est_pearson <- function(data_est){
  mod <- glm(los ~ hmo + white, family = quasipoisson, data = data_est)
  return(summary(mod)$dispersion)
}
est_fletcher_farrington <- function(data_est){
  mod <- glm(los ~ hmo + white, family = quasipoisson, data = data_est)
  n <- nobs(mod)
  p <- length(coef(mod))
  phi_pearson <- summary(mod)$dispersion

```

```

mu <- mod$fitted.values
y <- data_est$los
s_bar <- mean((1 / mu) * (y - mu))
phi_Fl <- phi_pearson / (1 + s_bar)
phi_Fa <- phi_pearson - (n * s_bar) / (n - p)
return(c(Fletcher = phi_Fl, Farrington = phi_Fa))
}

# Svolgimento delle S simulazioni: per ciascuna si estraggono 5 stime e gli
# estremi dell'intervallo di confidenza da MeBRboot
scenari <- expand.grid(phi = vett_phi, n = vett_n)
matrice_risultati <- list()

for (row in 1:nrow(scenari)) {
  n_corr <- scenari$n[row]
  phi_corr <- scenari$phi[row]

  struttura_scen <- strutture_fisse[[n_corr]]
  struttura_dati_fissa <- struttura_scen$dati
  mu_vettore_fissato <- struttura_scen$mu

  matrice_repliche <- matrix(NA, nrow = S, ncol = 7)
  colnames(matrice_repliche) <- c("Pearson", "MeBR_Boot", "Fletcher",
                                "Farrington", "Benussi", "lower", "upper")

  for (s in 1:S){
    dati_sim <- sim_data_fissato(n_target = n_corr, phi_target = phi_corr,
                                mu_scen = mu_vettore_fissato,
                                data_struttura = struttura_dati_fissa)
    phi_P <- est_pearson(dati_sim)
    phi_FF <- est_fletcher_farrington(dati_sim)

    phi_M <- MeBRboot(theta_hat = phi_P, theta_L = 1, theta_U = phi_corr*2.5,
                      N = 300, R = 200, datas = dati_sim, est_fn = est_pearson,
                      method = "boot", parameter = "phi", ic = T)

    phi_B <- MeBRben(theta_hat = phi_P, M = 500, datas = dati_sim, est_fn =
                    est_pearson)

    matrice_repliche[s,] <- c(phi_P, phi_M["Stima"], phi_FF["Fletcher"],
                             phi_FF["Farrington"], phi_B, phi_M["Infe"],
                             phi_M["Sup"])
  }

  nome_cella <- paste0("Scenario n ", n_corr, ", phi ", phi_corr)
  matrice_risultati[[nome_cella]] <- matrice_repliche
}

```


Bibliografia

- BENUSSI, D., KOSMIDIS, I., SALVAN, A. & SARTORI, N. (2026). Focused median bias reduction, ArXiv preprint arXiv:2606.28597 .
- BIRNBAUM, A. (1961). A unified theory of estimation, i. *Annals of Mathematical Statistics* **32**, 112–135.
- CABRERA, J. & WATSON, G. S. (1997). Simulation methods for mean and median bias reduction in parametric estimation. *Journal of Statistical Planning and Inference* **57**, 143–152.
- COX, D. R. & REID, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society. Series B (Methodological)* **49**, 1–39.
- DAVISON, A. C. & HINKLEY, D. V. (1997). *Bootstrap Methods and their Application*. Cambridge University Press.
- EFRON, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics* **7**, 1–26.
- EFRON, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.
- EFRON, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association* **82**, 171–185.
- EFRON, B. & TIBSHIRANI, R. J. (1993). *An Introduction to the Bootstrap*. Chapman Hall.
- FAHRMEIR, L. & TUTZ, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models, 2nd ed.* Springer.
- FARRINGTON, C. P. (1996). On assessing goodness of fit of generalized linear models to sparse data. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 349–360.

- FIRTH, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.
- FLETCHER, D. J. (2012). Estimating overdispersion when fitting a generalized linear model to sparse data. *Biometrika* **99**, 230–237.
- GRÜN, B., KOSMIDIS, I. & ZEILEIS, A. (2012). Extended beta regression in R: Shaken, stirred, mixed, and partitioned. *Journal of Statistical Software* **48**, 1–25.
- GUERRIER, S., DUPUIS-LOZERON, E., MA, Y. & VICTORIA-FESER, M.-P. (2019). Simulation-based bias correction methods for complex models. *Journal of the American Statistical Association* **114**, 146–157.
- HALL, P. & WILSON, S. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics* **47**, 757–762.
- HILBE, J. M. (2011). *Negative Binomial Regression*, 2nd ed. Cambridge University Press.
- KENNE PAGUI, E. C., SALVAN, A. & SARTORI, N. (2017). Median bias reduction of maximum likelihood estimates. *Biometrika* **104**, 923–938.
- KOSMIDIS, I. (2014a). Bias in parametric estimation: reduction and useful side-effects. *WIREs Computational Statistics* **6**, 155–209.
- KOSMIDIS, I. (2014b). Improved estimation in cumulative link models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **76**, 169–196.
- KOSMIDIS, I. (2026). *brglm2: Bias Reduction in Generalized Linear Models*. R package version 1.1.0.
- KOSMIDIS, I., KENNE PAGUI, E. & SARTORI, N. (2020). Mean and median bias reduction in generalized linear models. *Statistics and Computing* **30**, 43–59.
- KUCHIBHOTLA, A. K., BALAKRISHNAN, S. & L., W. (2024). The HulC: confidence regions from convex hulls. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **86**, 586–622.
- LEHMANN, E. L. & ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer.
- LIANG, K. & ZEGER, S. (1995). Inference based on estimating functions in the presence of nuisance parameters. *Statistical Science* **10**, 158–173.

- LINERO, A. R. (2026). Bayesian nonparametric quasi likelihood. *Journal of the American Statistical Association*, 1–14.
- MACKINNON, J. G. & SMITH, A. A. (1998). Approximate bias correction in econometrics. *Journal of Econometrics* **85**, 205–230.
- MARONNA, R., MARTIN, R., YOHAI, V. & SALIBIÁN-BARRERA, M. (2019). *Robust Statistics: Theory and Methods (with R)*. John Wiley Sons.
- MCCULLAGH, P. & NELDER, J. A. (1989). *Generalized Linear Models, 2nd ed.* Chapman and Hall.
- MORRIS, C. N. (1982). Natural exponential families with quadratic variance functions. *The Annals of Statistics* **10**, 65–80.
- NELDER, J. A. & PREGIBON, D. (1987). An extended quasi-likelihood function. *Biometrika* **74**, 221–232.
- NEYMAN, J. & SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32.
- PACE, L. & SALVAN, A. (1996). *Teoria della Statistica: Metodi, Modelli, Approssimazioni Asintotiche*. Cedam.
- PACE, L. & SALVAN, A. (1997). *Principles of Statistical Inference From a Neo-Fisherian Perspective*. World Scientific Publishing Company.
- PRATER, N. H. (1956). Estimate gasoline yields from crudes. *Petroleum Refiner* **35**, 236–238.
- SALVAN, A., SARTORI, N. & PACE, L. (2020). *Modelli Lineari Generalizzati*. Springer.
- SARTORI, N. (2003). Modified profile likelihoods in models with stratum nuisance parameters. *Biometrika* **90**, 533–549.
- WEDDERBURN, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika* **61**, 439–447.
- ZEILEIS, A. (2025). *betareg: Beta Regression*. R package version 3.6.0.

