UNIVERSITÁ DEGLI STUDI DI MILANO-BICOCCA

Scuola di Economia e Statistica

Corso di Laurea Magistrale in

SCIENZE STATISTICHE ED ECONOMICHE

- Clamses -

Distribuzione Secante Iperbolica: caratterizzazione, generalizzazione e applicazioni in modelli lineari generalizzati

Relatore: Dott. Tommaso Rigon

Correlatore: Prof.ssa Sonia Migliorati

Tesi di Laurea di:

Maria Regina Mucilli

Matr. N. 872352

Anno Accademico 2024/2025

Indice

Capitolo 1

Introduzione

L'elaborato nasce dall'approfondimento del lavoro di Carl N.Morris riguardo alla Famiglia di distribuzioni Esponenziale Naturale con particolare attenzione, tra queste, alle distribuzioni caratterizzate da funzione di varianza quadratica (Morris 1982). Le distribuzioni appartenenti alla Famiglia Esponenziale Naturale rivestono un ruolo centrale in statistica teorica e applicata, grazie alla loro struttura matematica elegante e versatile. Le distribuzioni di questa famiglia permettono di esprimere la funzione di densità in una forma standardizzata, caratterizzata da un parametro naturale e una funzione cumulante che semplifica il calcolo dei momenti e delle funzioni generatrici. Inoltre, le famiglie esponenziali costituiscono la base teorica dei modelli lineari generalizzati, in cui la scelta della distribuzione della variabile risposta determina la relazione tra media e varianza.

In Morris 1982 vengono identificate sei distribuzioni appartenenti alla famiglia esponenziale naturale, ognuna caratterizzata da una funzione di varianza che è una funzione specifica quadratica della media. Cinque delle sei distribuzioni sono ben note in letteratura statistica; se ne sono infatti studiate caratteristiche e proprietà e sono state ampiamente impiegate per lo sviluppo di modelli lineari generalizzati.

La sesta distribuzione, chiamata distribuzione Secante Iperbolica Generalizzata, oggetto

di approfondita analisi in questo elaborato, sembra essere meno conosciuta e approfondita nonostante sia caratterizzata da proprietà, come l'asimmetria e la relativa pesantezza delle code, che la rendono interessante in contesti di modellizzazione.

Ponendosi nel contesto descritto, questo elaborato si pone tre obiettivi principali.

In primo luogo, si intende caratterizzare la distribuzione secante iperbolica, derivandone la funzione di densità, i momenti (media e varianza), la funzione generatrice dei momenti, la funzione di ripartizione e la funzione quantile. Per approfondirne le caratteristiche si effettua un confronto dettagliato tra la distribuzione secante iperbolica nella sua forma standard e la distribuzione normale standard. Il confronto mira a evidenziare le peculiarità della distribuzione in analisi, come la sua leptocurtosi e le code più pesanti, che la rendono particolarmente adatta a modellare fenomeni caratterizzati dalla presenza di valori estremi.

Successivamente, si esplorano le applicazioni in modelli lineari generalizzati, basati sulla funzione di verosimiglianza della distribuzione Secante Iperbolica, valutandone le prestazioni in contesti simulativi. Infine, si esplorano le applicazioni di un modello di quasi-verosimiglianza, più flessibile e basato sui primi due momenti della distribuzione secante iperbolica, applicandolo a un dataset reale.

Per perseguire questi obiettivi, la tesi adotta un approccio combinato. L'analisi teorica si concentra sulla derivazione e sulla comprensione delle proprietà della distribuzione, mentre la parte applicativa utilizza implementazioni in `R` per condurre simulazioni, stimare parametri e verificare le prestazioni della distribuzione all'interno di modelli lineari generalizzati. In questo modo, si comprende la struttura matematica della distribuzione, e se ne valuta l'efficacia in contesti realistici di modellizzazione che deviano dalla normalità, presentando asimmetria e leptocurtosi.

La tesi si organizza nella seguente struttura.

Il capitolo 2 introduce il contesto teorico e metodologico di riferimento, fornendo gli stru-

menti necessari per comprendere le distribuzioni di probabilità appartenenti alla famiglia esponenziale naturale, i modelli lineari generalizzati e i modelli di quasi verosimiglianza.
Il capitolo 3 è dedicato alla caratterizzazione completa della distribuzione secante iperbolica, con un'analisi dettagliata delle sue proprietà. In particolare, si procede a un confronto della distribuzione in analisi nella sua forma standard e simmetrica con la distribuzione normale standard.
Nel capitolo 4 si sviluppa prima il modello di regressione, lineare generalizzato, tramite la distribuzione secante iperbolica e vengono sviluppate funzioni in `R` che ne consentono la stima. Poi, si deriva il modello di quasi-verosimiglianza, basato sui primi due momenti della distribuzione e, in particolare, sulla loro relazione.
Nel capitolo 5 si riportano i risultati ottenuti. I primi risultati nascono dall'applicazione del modello lineare generalizzato in un contesto simulativo; successivamente si riportano i risultati del modello di quasi-verosimiglianza basato sulla funzione di varianza della distribuzione secante iperbolica, applicato a un dataset reale. Infine, quest'ultimo modello viene confrontato con il corrispondente modello gaussiano classico applicato allo stesso dataset.
Nel capitolo 6 vengono presentate le conclusioni, con un riassunto degli obiettivi del progetto e dei risultati ottenuti, e un accenno ai possibili sviluppi futuri a seguito di questo lavoro.

Capitolo 2

Fondamenti Teorici

In questo capitolo si espongono i fondamenti teorici a supporto e a dimostrazione dell'elaborato. Il lavoro nasce a partire dall'articolo *Natural Exponential Families with Quadratic Variance Function* di Carl N. Morris (Morris 1982) in cui l'autore descrive formalmente la famiglia di distribuzioni esponenziali naturali e le loro caratteristiche.
Si descrive poi la formulazione del modello lineare generalizzato per le famiglie esponenziali e di un modello più flessibile di quasi-verosimiglianza.

2.1 Famiglia Esponenziale Naturale (NEF)

Le famiglie esponenziali naturali (NEF, *Natural Exponential Families*) costituiscono un'importante classe di distribuzioni di probabilità, utilizzata ampiamente in statistica teorica e applicata. Queste famiglie presentano una struttura algebrica e analitica che ne rende agevole l'utilizzo in contesti come la teoria della stima, la statistica bayesiana e i modelli lineari generalizzati. In particolare, le famiglie esponenziali naturali rappresentano un sottoinsieme delle famiglie esponenziali con specifiche caratteristiche.
Si parla di famiglia esponenziale nel caso di una classe di distribuzioni di probabilità che

può essere espressa in una specifica forma esponenziale all'interno della quale i parametri naturali e le statistiche naturali sono fondamentali per definire le proprietà della famiglia. In una famiglia esponenziale naturale, i parametri naturali e le statistiche naturali sono equivalenti alla funzione identità.

Una sottoclasse chiave delle famiglie esponenziali naturali è rappresenta dalle distribuzioni con funzione di varianza quadratica (NEF-QVF, *Natural Exponential Families with Quadratic Variance Function*), in cui la varianza è una funzione quadratica della media. Questa proprietà semplifica l'analisi e fornisce collegamenti a vari modelli statistici.

2.1.1 Specificazione

Famiglia Esponenziale (EF) di ordine 1

Sia Y una variabile aleatoria con distribuzione appartenente a una **Famiglia Esponenziale** (EF) di ordine 1 con parametro naturale θ nello spazio parametrico $\Theta \in \mathbb{R}$. Allora, la legge di probabilità di Y ammette tale rappresentazione:

$$P_\theta(Y \in A) = \int_A \exp\Big\{\theta T(y) - \psi(\theta)\Big\}\xi(dy), \tag{2.1}$$

dove

- θ è il parametro naturale;
- Θ, il più grande insieme per cui l'equazione è finita quando $A = \mathbb{R}$, è lo spazio del parametro naturale;
- ξ è una misura di riferimento che assorbe ogni fattore della densità di y che non dipende da $\theta \in \Theta$;
- $A \subset \mathbb{R}$ è un insieme misurabile;
- T è la statistica naturale, funzione misurabile a valori reali;

- $\psi(\theta)$ è la funzione cumulante (o *log-partition function*), funzione di normalizzazione scelta in modo tale che 2.1 ha probabilità unitaria se $A = \mathbb{R}$ (cioè definisce una probabilità).

Famiglia Esponenziale Naturale (NEF) di ordine 1

L'osservazione naturale in 2.1 è $X = T(Y)$. Allora, la distribuzione di X ha forma:

$$P_\theta(X \in A) = \int_A \exp\{\theta x - \psi(\theta)\} dF(x), \tag{2.2}$$

dove F si può sempre assumere essere una funzione di distribuzione cumulativa, senza perdita di generalità. Questa rappresentazione (2.2) definisce una **Famiglia Esponenziale Naturale** (NEF).

Allora, la famiglia parametrica esponenziale naturale di ordine 1 ha rappresentazione

$$NEF_1 = \Big\{ f(x,\theta) = \exp\Big\{ x\theta - \psi(\theta) \Big\} h(x), x \in S_X, \theta \in \Theta \Big\},$$

dove h è funzione nota di x e non dipende dal parametro θ mentre ψ è funziona nota di θ e indipendente dalla variabile x.

2.1.2 Proprietà della Famiglia Esponenziale Naturale

Si analizzano in questa sezione le caratteristiche fondamentali delle Famiglie Esponenziali Naturali, trattate in Morris 1983.

Funzione cumulante

Si definisce la funzione cumulante o *log-partition function* come

$$\psi(\theta) \equiv \log \int \exp\Big\{ \theta x \Big\} dF_0(x), \tag{2.3}$$

dove $\psi(\theta)$ è definita su Θ, cioè il più grande intervallo di valori per cui 2.3 esiste ed è finita.

La *log-partition function* svolge un duplice ruolo: da un lato assicura che la densità sia correttamente normalizzata, dall'altro costituisce lo strumento attraverso cui si possono caratterizzare i momenti della distribuzione.

Per costruzione, la funzione $\psi(\theta)$ assicura che $f(x, \theta)$ sia una densità di probabilità, ossia

$$\int_{S_X} f(x;\theta) dF_0(x) = 1.$$

Allora, le funzioni di distribuzione cumulative F_θ, $\theta \in \Theta$, definite dalle misure differenziali associate

$$dF_\theta(x) \equiv \exp\Big\{\theta x - \psi(\theta)\Big\} dF_0(x), \tag{2.4}$$

formano una famiglia esponenziale naturale, in cui ciascuna F_θ è una funzione di distribuzione cumulativa.

La funzione *log-partition* $\psi(\theta)$ non è soltanto un termine di normalizzazione della densità esponenziale, ma costituisce l'oggetto matematico che governa la struttura dei momenti, la geometria dei parametri e il comportamento della verosimiglianza.

Gode di alcune proprietà fondamentali:

- *Convessità* di $\psi(\theta)$ sull'insieme Θ: questa proprietà discende direttamente dall'ineguaglianza di Hölder e dalla definizione di ψ come logaritmo di una trasformata di Laplace (Brown 1986). La convessità garantisce l'unicità delle soluzioni nei problemi di stima basati sulla massimizzazione della verosimiglianza.

- *Differenziabilità* all'interno del dominio naturale: consente di caratterizzare i momenti della statistica sufficiente $T(x)$ attraverso le derivate della funzione cumulante.

- *Dualità naturale*: la funzione log-partition stabilisce una mappa biunivoca tra il dominio dei parametri naturali Θ e lo spazio dei momenti M. In particolare, la mappa tra θ e il primo momento μ (si deriva in 2.1.2) è liscia e uno-a-uno, e permette di introdurre una rappresentazione duale della distribuzione, sia in termini dei momenti

sia in termini dei parametri naturali. Questo legame tra Θ e M è alla base della geometria delle famiglie esponenziali (Barndorff-Nielsen 2014).

Funzione generatrice dei momenti e Funzione generatrice dei cumulanti

La funzione generatrice dei momenti (MGF) di una famiglia esponenziale naturale ha forma:

$$M_X(t) = E_\theta[e^{tX}] = \exp\{\psi(\theta + t) - \psi(\theta)\}. \tag{2.5}$$

La funzione generatrice dei cumulanti (CGF) si scrive allora come

$$\psi_\theta(t) = \log M_X(t) = \log E_\theta\left[e^{tx}\right] = \log \int \left[e^{tx}\right]\ dF_\theta(x) = \psi(t + \theta) - \psi(\theta). \tag{2.6}$$

Infatti, sostituendo nella definizione di CGF la forma della famiglia naturale esponenziale 2.4, si ottiene

$$\int e^{tX}\ dF_\theta(x) = \int e^{tX}\ \exp\Big\{\theta x - \psi(\theta)\Big\} dF_0(x) = e^{-\psi(\theta)} \int e^{(t+\theta)x} dF_0(x) = e^{-\psi(\theta)}\ e^{\psi(t+\theta)}.$$

Applicando ora il logaritmo:

$$\log\left(e^{-\psi(\theta)}\ e^{\psi(t+\theta)}\right) = \psi(t + \theta) - \psi(\theta).$$

La Funzione Generatrice dei Cumulanti $\psi_\theta(t)$ codifica tutti i cumulanti, dove l'r-esimo cumulante C_r di F_θ è definito come l'r-esima derivata della CGF in $t = 0$, cioè

$$C_r = \left.\frac{d^r \psi_\theta(t)}{dt^r}\right|_{t=0} = \psi^{(r)}(\theta). \tag{2.7}$$

Per $r = 1, 2$ si ottengono rispettivamente le formule di media μ (2.8) e varianza $V(\mu)$ (2.9), sostituendo $r = 3$ si calcola l'indice di asimmetria e con $r = 4$ si deriva la curtosi della famiglia esponenziale naturale.

Media

Dalla funzione cumulante definita in 2.3, si derivano i momenti di una famiglia esponenziale naturale.

Il valore atteso della famiglia esponenziale naturale è dato infatti da

$$E_\theta[X] = \mu = \int x \, dF_\theta(x) = \psi'(\theta), \quad \mu \in \Omega \subset R. \tag{2.8}$$

Si dimostra che derivando $\psi(\theta)$ si ottiene

$$\psi'(\theta) = \frac{\int x \exp\{\theta x\} dF_0(x)}{\int \exp\{\theta x\} dF_0(x)};$$

si nota che l'espressione

$$\frac{\exp\{\theta x\} dF_0(x)}{\int \exp\{\theta x\} dF_0(x)}$$

è esattamente la misura di probabilità $dF_\theta(x)$ (normalizzata) da cui

$$\psi'(\theta) = \int x \, dF_\theta(x) = E_\theta[X].$$

Funzione di varianza

La varianza della famiglia esponenziale naturale è funzione della media μ ed è

$$Var_\theta(X) = V(\mu) = \int (x-\mu)^2 \, dF_\theta(x) = \psi''(\theta) > 0. \tag{2.9}$$

Si dimostra che, derivando nuovamente la funzione cumulante (2.3), si ottiene

$$\psi''(\theta) = \int x^2 \, dF_\theta(x) - \left(\int x \, dF_\theta(x) \right)^2 = E_\theta[X^2] - (E_\theta[X])^2 = Var_\theta(X).$$

La varianza quindi dipende dalla media $\mu = \psi'(\theta)$; infatti, nella NEF ogni distribuzione è parametrizzata da θ e la media μ è funzione monotona di θ. Di conseguenza, ogni altra funzione di θ può essere espressa come funzione della media (Morris e Lock 2009).
Allora, $V(\mu) = \psi''(\theta)$ con il suo dominio Ω, (V, Ω), è detta **funzione di varianza** della famiglia esponenziale naturale.

Asimmetria e Curtosi

La skewness (asimmetria) indica mancanza di simmetria. In matematica, una figura si dice simmetrica se esiste un punto al suo interno tale che, tracciando per esso una perpendicolare all'asse X, la figura risulti divisa in due parti congruenti, cioè identiche in

ogni aspetto, oppure tali che una parte possa essere sovrapposta all'altra, ossia immagini speculari l'una dell'altra (Weyl 2015).

Si ottiene l'indice di asimmetria di una distribuzione tramite la derivazione del secondo e del terzo momento centrale. In particolare, si utilizzano i coefficienti β e γ definiti da Karl Pearson (Pearson 1905):

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

dove $\mu_2,\ \mu_3$ sono, rispettivamente, il secondo e il terzo momento centrale.

Riscrivendo nella notazione utilizzata finora e in Morris 1982, per il calcolo dell'asimmetria per la famiglia di distribuzione esponenziale naturale si deriva β_1 come

$$\beta_1 = \frac{C_3^2}{V(\mu)^3}$$

dove C_3 è il cumulante di ordine 3 (2.7) e $V(\mu)$ è la funzione di varianza.

Il coefficiente β_1 misura la magnitudine dell'asimmetria, ma non permette di stabilirne la direzione, positiva o negativa, in quanto assume sempre valori positivi.

Per questo si deriva il coefficiente γ_1:

$$\gamma_1 = \pm\sqrt{\beta_1} = \frac{C_3}{\sqrt{V(\mu)}}.$$

In questo caso, il segno dell'asimmetria dipende dal valore di C_3, che determina la direzione dell'asimmetria.

La conoscenza delle misure di tendenza centrale, di dispersione e di asimmetria non consente di ottenere una comprensione completa della distribuzione di una variabile. Per avere una descrizione esaustiva della forma della distribuzione, è necessario considerare un'ulteriore misura, che può essere analizzata attraverso la curtosi.

Karl Pearson (Pearson 1905) definisce la curtosi come la convessità di una curva, poiché essa quantifica il grado di concentrazione dei valori attorno alla media, fornendo quindi un'indicazione dell'appiattimento o del picco della distribuzione.

Il grado di curtosi di una distribuzione viene valutato rispetto a quello di una curva normale. In particolare:

- le distribuzioni con un picco più accentuato rispetto a quello della normale vengono definite leptocurtiche;
- le distribuzioni con un profilo più piatto rispetto alla normale vengono definite platicurtiche;
- la distribuzione normale, caratterizzata da un livello intermedio di curtosi, è definita mesocurtica.

Per il calcolo della curtosi vengono utilizzati il secondo e il quarto momento centrale della variabile. A tal fine, si impiega la seguente formula proposta da Karl Pearson:

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{C_4 + 3C_2^2}{V(\mu)^2}$$

infatti per le proprietà delle NEF, i momenti centrali possono essere espressi tramite cumulanti e, in particolare, fino al terzo ordine coincidono. Alternativamente al coefficiente β_2, si può utilizzare l'indice γ_2 che descrive l'eccesso di curtosi della distribuzione in esame rispetto al valore di curtosi della distribuzione normale pari a 3:

$$\gamma_2 = \beta_2 - 3.$$

Interpretativamente, si ha

- se $\beta_2 = 3$ o $\gamma_2 = 0$ la distribuzione presenta un grado di picco simile a quello della curva normale, e viene definita mesocurtica;
- se $\beta_2 < 3$ o $\gamma_2 < 0$, la distribuzione risulta più piatta rispetto alla normale e viene definita platicurtica. In questo caso, i valori si distribuiscono più uniformemente attorno alla media, con code meno pronunciate e un picco centrale meno accentuato;
- se $\beta_2 > 3$ o γ_2,la distribuzione è caratterizzata da un picco più acuto rispetto alla normale, e viene definita leptocurtica. Ciò indica una maggiore concentrazione dei

valori vicino alla media e code più pesanti, che ne determina una maggiore probabilità di osservare valori estremi.

2.1.3 Famiglia Esponenziale Naturale con Funzione di Varianza Quadratica

Alcune famiglie esponenziali naturali godono di un'ulteriore proprietà relativa alla funzione di varianza (2.1.2), cioè hanno **funzione di varianza quadratica**.
Una funzione di varianza è detta quadratica se può essere scritta come funzione quadratica della media μ, cioè nella forma

$$V(\mu) = v_0 + v_1\mu + v_2\mu^2. \tag{2.10}$$

Proprietà

Le proprietà di questa specifica classe di distribuzioni sono trattate in Morris 1983.

- Le famiglie esponenziali naturali con funzione di varianza quadratica (NEF-QVF) sono chiuse rispetto alle convoluzioni di una trasformazione lineare (Morris e Lock 2009). Una convoluzione (Jacod e Protter 2004) di una trasformazione lineare di una NEF-QVF è anch'essa una NEF-QVF, anche se non necessariamente identica alla distribuzione originaria.

  Siano $X_1, \ldots, X_n$ variabili indipendenti e identicamente distribuite (i.i.d.) con distribuzione NEF-QVF, allora si considera la convoluzione di una trasformazione lineare di X:

  $$Y = \sum_{i=1}^{n} \frac{X_i - b}{c}, \quad \text{con } b,\, c \neq 0 \text{ costanti.}$$

  La media di Y è

  $$\mu^* = \frac{n(\mu - b)}{c}.$$

La varianza di Y può essere espressa in termini della funzione di varianza di X. Se

$$\text{Var}(X) = V(\mu) = \nu_0 + \nu_1\mu + \nu_2\mu^2,$$

allora la nuova NEF-QVF ha funzione di varianza

$$\text{Var}(Y) = V^*(\mu^*) = \nu_0^* + \nu_1^*\mu + \nu_2^*\mu^2,$$

dove

$$\nu_0^* = \frac{nV(b)}{c^2}, \quad \nu_1^* = \frac{V'(b)}{c}, \quad \frac{\nu_2^*}{n} = \frac{\nu_2}{n},$$

con $c \neq 0$ e b costanti.

- Siano X_1 e X_2 NEF indipendenti con lo stesso parametro θ e sia $Y = X_1 + X_2$. Allora la distribuzione condizionata di X_1 dato Y ha varianza quadratica in Y se e solo se X_1 e X_2 sono NEF-QVF.

2.1.4 NEF-QVF note

Solo sei famiglie univariate e a un parametro (e funzioni lineari di queste) sono NEF-QVF. Le prime famose cinque sono: Normale, Poisson, Gamma, Binomiale e Binomiale Negativa. Tutte le distribuzioni appartenenti alla classe di distribuzioni NEF-QVF possono essere caratterizzate tramite la propria funzione di varianza in forma quadratica 2.10.
La distribuzione Normale è l'unica distribuzione delle NEF-QVF ad avere funzione varianza quadratica costante cioè $V(\mu) = \sigma^2$, cioè $v_1 = v_2 = 0$.
La distribuzione di Poisson invece è l'unica con funzione di varianza lineare cioè $V(\mu) = v_0 + v_1\mu,\ v_1 \neq 0$.
Le distribuzioni Gamma, Binomiale e Binomiale Negativa presentano funzioni di varianza strettamente quadratiche cioè con $v_2 \neq 0$. In particolare, da 2.10, si definisce $X^* = aV'(X)$ (Morris 1982) cioè una funzione lineare di X con funzione di varianza data da

$$V^*(\mu^*) = s + v_2(\mu^*)^2, \quad s = -sgn(dv_2) \tag{2.11}$$

dove d è il discriminante del polinomio di secondo grado (Milne 2003) che definisce la funzione di varianza.

Le diverse possibili combinazioni per la scelta di $s = (0, \pm 1)$ e $v_2 \neq 0$ danno vita alle distribuzioni Gamma, Binomiale e Binomiale Negativa, ricordando che per definizione la funzione di varianza è positiva $V(\mu) > 0$.

La distribuzione Gamma è caratterizzata da una funzione di varianza con $v_2 > 0,\ s = 0$, la distribuzione Binomiale è nel caso $v_2 < 0,\ s = 1$, mentre la distribuzione Binomiale Negativa ha $v_2 > 0,\ s = 1$. L'ultima combinazione possibile, cioè $v_2 > 0,\ s = 1$, sarà l'oggetto di analisi nel capitolo 3.

2.1.5 Famiglia di dispersione esponenziale di ordine 1 (DEF)

La famiglia esponenziale naturale presentata nella sezione precedente (2.1) è un modello parametrico incluso in una classe più ampia di modelli parametrici, ovvero la classe delle **Famiglie di Dispersione Esponenziale** (Jørgensen 1987).

Sia X una variabile aleatoria con distribuzione appartenente a una Famiglia di Dispersione Esponenziale univariata con parametro naturale $\theta \in \Theta$ e parametro di dispersione ϕ tale che $a(\phi) > 0$. Allora, la densità di X è esprimibile nella forma

$$f_Y(y; \theta, \phi) = \exp\Big\{\frac{\theta x - \psi(\theta)}{a(\phi)}\Big\} h(x, \phi) \tag{2.12}$$

Se $a(\phi) = 1$ e $h(x, \phi) = h(x)$ si torna nel caso di una famiglia esponenziale naturale univariata, quindi con densità 2.2.

2.2 Modelli Lineari Generalizzati (GLM)

2.2.1 Specificazione del modello

I modelli lineari generalizzati (Agresti 2015) costituiscono un'estensione del modello lineare generale e sono impiegati per lo studio della dipendenza in media di una variabile

risposta da una o più variabili esplicative.

In un modello lineare generalizzato, le osservazioni della variabile dipendente Y si ipotizzano essere generate indipendentemente da una variabile casuale della famiglia esponenziale. Allora, le componenti fondamentali per la costruzione di un modello lineare generalizzato sono:

- la funzione di distribuzione della variabile dipendente $f(y_i, \theta_i),\ i = 1, ..., n$, appartenente a una famiglia esponenziale (2.12);
- il predittore lineare $\eta_i = X_i'\beta$, dove $\beta = (\beta_0, ..., \beta_p),\ p < n$ sono i coefficienti di regressione e $X_i,\ \forall i \in \{1, ..., n\}$ è il vettore delle p variabili esplicative per l'i-esimo individuo;
- una funzione g, detta *link*, tale che $\mu = g^{-1}(\eta)$, cioè una funzione liscia, invertibile e nota che collega una combinazione lineare degli elementi di β con il valore atteso $E(Y_i) = \mu_i,\ i = 1, ..., n$ dell'i-esima osservazione.

***Link* canonico**

Per completare la specificazione di un modello lineare generalizzato è necessario scegliere una funzione *link*. La funzione di legame permette di modellare in termini lineari la dipendenza in media della variabile riposta dalle variabili esplicative.

Si sceglie convenientemente una funzione monotona, in modo che sia invertibile e garantisca $\mu = g^{-1}(\eta)$, differenziabile e tale che:

$$g : \Omega \to \mathbb{R}$$

cioè g mappa l'intervallo dei valori atteso ($\mu \in \Omega$) sulla retta reale.

In questo modo,

$$g^{-1} : \mathbb{R} \to \Omega$$

permette, una volta calcolato il predittore lineare η sulla retta reale, di riportarlo a un valore valido di $\mu \in \Omega$. Quindi scegliere una funzione di *link* monotona e differenziabile garantisce che le predizioni siano univoche, valide e computazionalmente trattabili.
Per ciascuna specificazione del modello statistico esponenziale per una variabile risposta, fra tutte le possibili funzioni di legame si parla di *link canonico* facendo riferimento a una funzione g tale che:

$$\theta = \theta(\mu) = \eta = x'\beta$$

e implica che

$$g(\mu) = \theta = \theta(\mu)$$

con cui il parametro naturale della famiglia esponenziale θ coincide con il predittore lineare. L'adozione del *link* canonico conferisce al GLM diverse proprietà desiderabili (Hardin e Hilbe 2007):

- garantisce che la stima dei parametri tramite il metodo della massima verosimiglianza sia unica e ben definita;
- assicura che la statistica sufficiente per il modello sia una funzione lineare dei dati osservati, semplificando l'analisi e l'interpretazione dei risultati;
- preserva le proprietà di invarianza rispetto alle trasformazioni lineari delle variabili esplicative, facilitando l'interpretazione dei coefficienti stimati;
- semplifica la derivazione e l'implementazione degli algoritmi di stima rendendo l'ottimizzazione più efficiente e meno suscettibile a problemi numerici (Bates et al. 2015).

2.3 Stima dei coefficienti di regressione

Una volta specificato il modello, si procede alla stima dei relativi parametri.

In particolare, si considera una famiglia di dispersione esponenziale 2.12, per la quale i parametri da stimare sono i coefficienti di regressione β e il parametro di dispersione ϕ (Jørgensen 1987).

Si ricorda che la famiglia esponenziale naturale costituisce un caso particolare della famiglia di dispersione esponenziale, ottenuto ponendo $\phi = 1$.

Si utilizza il metodo della massima verosimiglianza, cioè si massimizza la funzione di verosimiglianza del modello al fine di stimare il vettore dei coefficienti $\beta = (\beta_1, ..., \beta_p)$.

Siano $Y_1, ..., Y_n$ variabili casuali distribuite indipendentemente come famiglia esponenziale. Allora la distribuzione congiunta di $\mathbf{Y} = (Y_1, ..., Y_n)'$ è pari al prodotto delle densità marginali

$$f(\mathbf{y}; \beta, \phi) = \prod_{i=1}^{n} f(y_i; \beta, \phi) = \prod_{i=1}^{n} \exp\left\{ \frac{\theta_i y_i - \psi(\theta_i)}{a_i(\phi)} + h(y_i, \phi) \right\}$$

dove $\theta_i = \theta(\mu_i) = \theta(g^{-1}(\mathbf{x}_i'\beta))$. Allora la funzione di log-verosimiglianza è:

$$l(\beta, \phi) = \sum_{i=1}^{n} \frac{y_i\theta_i - \psi(\theta_i)}{a_i(\phi)} + h(y_i, \phi). \tag{2.13}$$

Restringendo al campo delle famiglie naturali esponenziali (2.2), si ha

$$l(\beta) = \sum_{i=1}^{n} y_i\theta_i - \psi(\theta_i) + h(y_i).$$

Si calcolano la derivata prima e la derivata seconda della log-verosimiglianza 2.13 per l'inferenza sul modello di regressione lineare generalizzato.

Funzione di *score*

La derivata prima è la funzione di *score* e vale:

$$\begin{aligned} l_*(\beta,\phi) &= \frac{\delta}{\delta\beta_r} l(\beta,\phi) \\ &= \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \left(y_i \frac{\delta\theta_i}{\delta\beta_r} - \frac{\delta\psi(\theta_i)}{\delta\beta_r} \right) \qquad (2.14) \\ &= \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \left(y_i \frac{\delta\theta_i}{\delta\beta_r} - \psi'(\theta_i)\frac{\delta\theta_i}{\delta\beta_r} \right), \quad i = 1,\ldots,n, \quad r = 1,\ldots,p. \qquad (2.15) \end{aligned}$$

Ponendo $a_i(\phi) = \frac{1}{w_i}$ dove $\mathbf{w} = (w_1, ...w_n)$ è un sistema di pesi, si applica la regola della catena per la derivazione di funzioni composte (Apostol 1974):

$$l_*(\beta,\phi) = \sum_{i=1}^{n} \frac{w_i}{\phi} \left(y_i \frac{\delta\theta_i}{\delta\beta_r} - \psi'(\theta_i)\frac{\delta\theta_i}{\delta\mu_i}\frac{\delta\mu_i}{\delta\eta_i}\frac{\delta\eta_i}{\delta\beta_r} \right), \quad i = 1,\ldots,n, \quad r = 1,\ldots,p.$$

Ricordando inoltre che $g(\mu_i) = \mathbf{x}_i'\beta = \eta_i$ e $\theta_i = \theta(\mu_i)$ inversa di $\mu(\theta_i)$ e ancora che $\psi'(\theta) = \mu_i = g^{-1}(\eta_i)$, allora si ottiene:

$$l_*(\beta,\phi) = \sum_{i=1}^{n} w_i \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{x_{ir}}{g'(\mu_i)}. \qquad (2.16)$$

Lo stimatore di massima verosimiglianza per i coefficienti di regressione risolve il sistema di equazioni definito dalla funzione di *score* $l_*(\beta,\phi) = 0$.

In forma matriciale si scrive

$$D'V^{-1}(\mathbf{y} - \mu) = 0$$

dove $V = diag\left(\frac{V(\mu_1)}{w_1}, \ldots, \frac{V(\mu_n)}{w_n}\right)$ e D è una matrice $n \times p$ i cui elementi sono $d_{ir} = \frac{x_{ir}}{g'(\mu_i)}$.

Matrice di informazione osservata e attesa

Si considera la seconda derivata della log-verosimiglianza 2.13, presa con segno negativo:

$$\begin{aligned} j_{rs} &= -\frac{\delta}{\delta\beta_s}\left[\frac{\delta}{\delta\beta_r} l(\beta,\phi)\right] \qquad (2.17) \\ &= -\frac{\delta}{\delta\beta_s} \sum_{i=1}^{n} \frac{w_i}{\phi}\left[(y_i - \mu_i)\frac{\delta\theta_i}{\delta\beta_r}\right] \\ &= \sum_{i=1}^{n} \frac{w_i}{\phi}\left[\frac{\delta\mu_i}{\delta\beta_s}\frac{\delta\theta_i}{\delta\beta_r} - (y_i - \mu_i)\frac{\delta^2\theta_i}{\delta\beta_r\delta\beta_s}\right], \quad r,s = 1,...,p. \end{aligned}$$

sono gli elementi della matrice dell'osservazione osservata J.

Considerando ora il valore atteso della matrice definita dagli elementi 2.17, $E(J)$, si deriva la matrice dell'informazione attesa di Fisher I, i cui elementi hanno quindi forma:

$i_{rs} = E[j_{rs}]$.

Allora, dato che $E[Y_i] = \mu_i$, si ottiene

$$i_{rs} = \sum_{i=1}^{n} w_i \frac{x_{ir} x_{is}}{V(\mu_i) g'(\mu_i)^2}, \quad r, s = 1, ..., p, \quad i = 1, ..., n. \tag{2.18}$$

In forma matriciale si scrive l'informazione attesa come

$$I = X'WX$$

dove $W = diag(w_1, ..., w_n)$ è la matrice diagonale dei pesi

$$w_i = \frac{1}{V(\mu_i) g'(\mu_1)^2}, \quad i = 1, ..., n.$$

2.3.1 Stima con *link* canonico

Scegliere il *link* canonico porta a una semplificazione delle equazioni della funzione di *score* e della informazione attesa. Infatti, scegliere $\theta(\mu_i) = g(\mu_i)$, implica che

$\theta_i = \eta_i = x_{i1}\beta_1 + ... + x_{ip}\beta_p$.

Allora, la funzione di *score* 2.16 è:

$$l_*(\beta, \phi) = \sum_{i=1}^{n} w_i (y_i - \mu_i) x_{ir}$$

e l'informazione attesa di Fisher 2.18 vale

$$i_{rs} = w_i V(\mu_i) x_{ir} x_{is}$$

dato che $\frac{\delta^2 \theta_i}{\delta \beta_r \delta \beta_s} = 0$.

Scelto il *link* canonico e specificata la famiglia di distribuzioni, la stima dei coefficienti β si ottiene risolvendo le equazioni di *score* derivate dalla log-verosimiglianza:

$$\frac{\partial l(\beta)}{\partial \beta} = \sum_{i=1}^{n} (y_i - \mu_i) \frac{\partial \mu_i}{\partial \eta_i} X_i = 0.$$

Nel caso del *link* canonico, si ha $\frac{\partial \mu_i}{\partial \eta_i} = V(\mu_i)$, dove $V(\mu_i)$ è la funzione di varianza della distribuzione. Ciò semplifica notevolmente il calcolo e permette di applicare algoritmi iterativi come il metodo di Fisher scoring o l'IRLS (Iteratively Reweighted Least Squares) per ottenere le stime $\hat{\beta}$.

2.3.2 Algoritmo di Newton-Raphson per i GLM

La stima dei coefficienti di regressione β in un modello lineare generalizzato si basa sulla massimizzazione della funzione di log-verosimiglianza $l(\beta, \phi)$ (2.13). L'obiettivo è trovare $\hat{\beta}$ tale che la funzione *score* 2.16 si annulli:

$$l_*(\beta, \phi) = \frac{\partial l(\beta, \phi)}{\partial \beta} = 0.$$

In generale, questa equazione non ammette soluzione in forma chiusa e si ricorre quindi a metodi iterativi. Un approccio classico è l'algoritmo di **Newton-Raphson** (Wedderburn 1974), che aggiorna i parametri secondo la regola:

$$\beta^{(t+1)} = \beta^{(t)} - \Big[J(\beta^{(t)}, \phi)\Big]^{-1} \, l_*(\beta^{(t)}, \phi),$$

dove $J(\beta, \phi)$ è la matrice di informazione osservata.

La matrice J può essere sostituita dalla matrice di informazione attesa $I(\beta, \phi)$, allora l'iterazione di Newton-Raphson assume la forma:

$$\beta^{(t+1)} = \beta^{(t)} + (X'WX)^{-1} X'W(y - \mu).$$

In questo caso, se il *link* è canonico, l'algoritmo mantiene la sua convergenza e le espressioni si semplificano poiché la matrice di informazione osservata coincide con quella attesa. Questa procedura corrisponde a una regressione lineare pesata iterativa ed è nota come IRLS (*Iteratively Reweighted Least Squares*).

In pratica, l'algoritmo viene implementato come segue:

1. Inizializzare $\beta^{(0)}$;

2. Calcolare $\eta^{(t)} = X\beta^{(t)}$ e $\mu^{(t)} = g^{-1}(\eta^{(t)})$;

3. Costruire la matrice dei pesi $W^{(t)} = diag(w_i)$ dove $w_i = \frac{\left(\frac{\delta\mu_i}{\delta\theta_i}\right)^2}{V(\mu_i)}$;

4. Costruire la variabile risposta aggiustata come $z_i^{(t)} = \theta_i^{(t)} + \frac{y_i - \mu_i^{(t)}}{\delta\mu_i/\delta\theta_i}$;

5. Aggiornare la stima dei coefficienti β risolvendo il problema di minimi quadrati pesati

$$\beta^{(t+1)} = \arg\min_{\beta}(z^{(t)} - X\beta)'W^{(t)}(z^{(t)} - X\beta)$$

6. Ripetere fino a convergenza, cioè fino a che

$$\|\beta^{(t+1)} - \beta^{(t)}\| < \epsilon$$

   per una soglia ϵ prefissata.

2.4 Definizione di una famiglia personalizzata in `R`

In `R`, la funzione `glm()` (R: Fitting GLM) consente di stimare modelli appartenenti alla classe dei *Generalized Linear Models* (GLM), i quali si basano sulla specificazione di una famiglia (`family`) che descrive la distribuzione della variabile risposta, la funzione legame (*link function*) e la funzione di varianza. Oltre alle famiglie predefinite (ad esempio `binomial`, `poisson`, `gaussian`), è possibile per l'utente definire famiglie personalizzate costruendo una funzione che restituisca un oggetto di classe `family` (R: Family Objects for Models). Tale funzione deve includere i principali elementi che caratterizzano la famiglia:

- la funzione legame diretta $g(\mu)$ e la sua inversa $g^{-1}(\mu)$;
- la derivata della funzione *link* inversa;
- la funzione di varianza $V(\mu)$, che determina la forma della varianza in funzione della media;
- la funzione dei residui di Pearson;

- la formula per il calcolo dell'AIC (Akaike Information Criteria) (Akaike 1998).

A questi elementi si possono aggiungere ulteriori componenti, come le condizioni di validità dei parametri se esistono vincoli su essi e l'espressione di inizializzazione che fornisce valori iniziali per la media. Una volta definite tali funzioni, esse vengono raccolte in una lista a cui viene attribuita la classe `family`, rendendo la nuova famiglia utilizzabile all'interno di `glm()` come qualsiasi famiglia standard.

2.5 Modello di Quasi-Verosimiglianza

Nell'applicazione a dati reali di modelli lineari generalizzati, è spesso utile rilassare le ipotesi del modello, rendendolo più flessibile e basato su relazioni di secondo ordine, che permettono una maggiore adattabilità a dati, continui o discreti. Questo permette di stimare un modello di regressione generalizzato anche quando il parametro di dispersione del modello ϕ non è noto o si discosta da quanto previsto dal modello teorico.
Le ipotesi più deboli alla base del modello di quasi-verosimiglianza (Salvan, Sartori e Pace 2020) sono:

- $E[Y_i] = \mu(x_i\beta) = g^{-1}(x_i\beta)$;
- $Var(Y_i) = \phi V(\mu_i)$;
- Y_i e Y_j indipendenti se $i \neq j$,

dove $\phi > 0$ è un parametro di dispersione, non noto.
A differenza dei modelli lineari generalizzati basati sulla verosimiglianza, non è necessario definire l'intera distribuzione condizionata della variabile risposta, ma solo la relazione tra media e varianza, permettendo così una maggiore flessibilità nell'analisi di dati reali. Il modello, così specificato, consente di tener conto della sovradispersione (o della sotto-dispersione) dei dati così da permettere un incremento (o decremento) della varianza,

rappresentato da ϕ, rispetto al corrispondente GLM, senza la necessità di una specificazione parametrica.

Allora, sotto le ipotesi di secondo ordine, è possibile specificare la funzione log-quasi-verosimiglianza (Wedderburn 1974):

$$l_Q(\beta; y, X, \phi) = \sum_{i=1}^{n} l_Q(\beta; y_i, x_i, \phi) = \sum_{i=1}^{n} \int_a^{\mu_i(\beta)} \frac{y_i - t}{\phi V(t)} dt$$

dove a è una costante indipendente da β.

La quasi-verosimiglianza mantiene molte proprietà della funzione di verosimiglianza; infatti le equazioni di stima rimangono non distorte e vale ancora l'identità dell'informazione:

$$E[\nabla l_Q(\beta; y, X, \phi)] = 0; \quad E[-\nabla^2 l_Q(\beta; y, X, \phi)] = E[\nabla l_Q \nabla l_Q']$$

con ∇ gradiente rispetto a β.

La stima dei parametri del modello β è ottenuta risolvendo le equazioni di *quasi-score*, analoghe alle equazioni di *score* della verosimiglianza, che in forma matriciale si scrivono:

$$U(\beta) = X'W(y - \mu) = 0,$$

dove $W = diag\left(\frac{1}{V(\mu_i)}\right)$.

Inoltre il parametro dispersione ϕ può essere stimato, corretto e consistente, tramite metodo dei momenti come

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

dove $\hat{\mu}_i$ sono le stime dei valori attesi basate sui coefficienti di regressione stimati dal modello $\hat{\beta}$, n il numero di osservazioni e p il numero di parametri del modello.

Gli stimatori $\hat{\beta}$ e $\hat{\phi}$ godono di importanti proprietà asintotiche: sono consistenti, asintoticamente normali e, sotto condizioni regolari, efficienti nella classe degli stimatori derivati da equazioni di *quasi-score* (McCullagh 1983). Tali proprietà garantiscono che l'inferenza sui parametri sia valida anche in assenza di una specifica distribuzione completa della variabile risposta.

Capitolo 3

Distribuzione Secante Iperbolica

Oltre alle cinque famiglie di distribuzioni più note (2.1.4), se ne considera una sesta, meno conosciuta ma con caratteristiche altrettanto interessanti. Si tratta della distribuzione Secante Iperbolica, anch'essa caratterizzata dalla forma esponenziale tipica delle famiglie esponenziali naturali e da una funzione di varianza quadratica che le permette di rientrare tra le NEF-QVF.

3.1 Derivazione della distribuzione Secante Iperbolica (HS)

3.1.1 Funzioni trigonometriche iperboliche

La distribuzione secante iperbolica è una distribuzione continua su $\mathbb{R}$ in cui, come evoca il nome della distribuzione stessa, la funzione secante iperbolica gioca un ruolo importante. Allora, si definiscono di seguito le funzioni trigonometriche iperboliche (McMahon 1896), utili per la comprensione dell'oggetto in analisi.

Le funzioni trigonometriche iperboliche sono definite a partire dalle funzioni esponenziali reali e sono strettamente legate all'iperbole equilatera $x^2 - y^2 = 1$.

Le principali funzioni trigonometriche iperboliche sono definite, per ogni $x \in \mathbb{R}$, come segue:

$$\begin{aligned}
\sinh(x) &= \frac{e^x - e^{-x}}{2} && \text{(seno iperbolico)},\\
\cosh(x) &= \frac{e^x + e^{-x}}{2} && \text{(coseno iperbolico)},\\
\tanh(x) &= \frac{\sinh(x)}{\cosh(x)} = \frac{e^x - e^{-x}}{e^x + e^{-x}} && \text{(tangente iperbolica)},\\
\operatorname{sech}(x) &= \frac{1}{\cosh(x)} = \frac{2}{e^x + e^{-x}} && \text{(secante iperbolica)}.
\end{aligned}$$

La funzione che caratterizza la distribuzione, la funzione trigonometrica secante iperbolica è definita dal reciproco della funzione cosecante iperbolica.

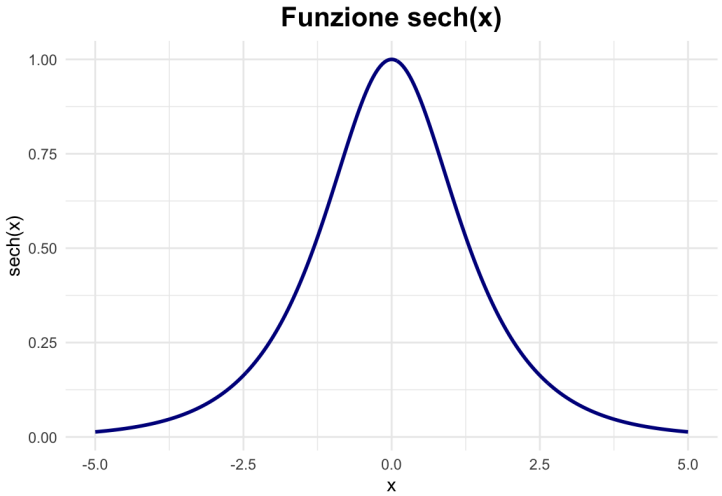

Figura 3.1: Grafico della funzione secante iperbolica

La funzione è definita su tutto $\mathbb{R}$, assume valori positivi compresi tra 0 e 1, è simmetrica rispetto all'asse delle ordinate e ha massimo pari a 1 per $x = 0$. Il suo grafico (3.1) ha come asintoto l'asse delle ascisse.

3.1.2 Derivazione

Una volta nota la funzione alla base della distribuzione in esame, si può procedere alla formulazione della distribuzione secante iperbolica (HS), fornita in Morris 1982 e alla derivazione delle sue quantità fondamentali.

Dall'analisi condotta in 2.1.4, le cinque famiglie di distribuzioni NEF-QVF ben note, corrispondono a cinque diverse combinazioni dei coefficienti in $V^*(\mu^*)$ (2.11).

In particolare, la distribuzione gamma, binomiale e binomiale negativa sono distribuzioni strettamente quadratiche a cui corrispondono tre diverse combinazioni di v_2 e s nella forma canonica $V^*(\mu^*)$.

Il caso mancante è il caso in cui $v_2 > 0$ e $s = 1$. Scegliendo $v_2 = 1$, si ha $V^*(\mu^*) = 1+(\mu^*)^2$, allora la distribuzione mancante è l'osservazione naturale di una famiglia esponenziale Beta.

Sia

$$y \sim Beta\left(0.5 + \frac{\theta}{\pi},\ 0.5 - \frac{\theta}{\pi}\right), \quad |\theta| < \frac{\pi}{2}$$

allora, la distribuzione di y può essere scritta in forma esponenziale (2.1).

Chiamando $\alpha = 0.5 + \theta/\pi$ e $\beta = 0.5 - \theta/\pi$, si ottiene

$$\begin{aligned} f_{\alpha,\beta}(y) &= \frac{y^{\alpha-1}\,(1-y)^{\beta-1}}{B(\alpha,\beta)} \\ &= \frac{1}{B(\alpha,\beta)}\, y^{\frac{\theta}{\pi}-\frac{1}{2}}\,(1-y)^{-\frac{\theta}{\pi}-\frac{1}{2}} \\ &= \frac{1}{B(\alpha,\beta)}\left(y^{-\frac{1}{2}}(1-y)^{-\frac{1}{2}}\right)\exp\Big\{\tfrac{\theta}{\pi}\Big(\log y - \log(1-y)\Big)\Big\} \\ &= \xi(y)\,\exp\left\{\theta\tfrac{1}{\pi}\log\left(\frac{y}{1-y}\right) - \log\left(B\Big(\tfrac{1}{2}+\tfrac{\theta}{\pi},\ \tfrac{1}{2}-\tfrac{\theta}{\pi}\Big)\right)\right\}. \end{aligned} \tag{3.1}$$

Allora, l'osservazione naturale $x = T(y)$ (2.1) corrisponde a

$$x \equiv \tfrac{1}{\pi}\log\left(\frac{y}{1-y}\right)$$

in 3.1, cioè x è una funzione logistica π-scalata, infatti è

$$x \equiv \frac{1}{\pi}\text{logit}(y).$$

Inoltre, da 3.1, si ha

$$h(y) = y^{-0.5}(1-y)^{-0.5}$$

e

$$\psi(\alpha, \beta) = \log\left(B\left(0.5 + \frac{\theta}{\pi},\ 0.5 - \frac{\theta}{\pi}\right)\right).$$

Si studia allora la densità dell'osservazione naturale x, calcolandola tramite cambio di variabile (Billingsley 1995).

Si ricava la relazione inversa

$$y = \frac{e^{\pi x}}{1+e^{\pi x}} = \frac{1}{1+e^{-\pi x}}$$

e se ne calcola la derivata

$$\frac{dy}{dx} = \frac{\pi e^{\pi x}}{(1+e^{\pi x})^2} = \pi y(1-y).$$

Allora, si dimostra la densità di x tramite cambio variabile come funzione logit π-scalata di y con distribuzione Beta:

$$\begin{aligned} f_X(x) &= f_Y(y(x))\left|\frac{dy}{dx}\right| \\ &= \frac{\pi}{B\left(\frac{1}{2}+\frac{\theta}{\pi},\ \frac{1}{2}-\frac{\theta}{\pi}\right)}\left(\frac{e^{\pi x}}{1+e^{\pi x}}\right)^{\frac{\theta}{\pi}+\frac{1}{2}}\left(\frac{1}{1+e^{\pi x}}\right)^{-\frac{\theta}{\pi}+\frac{1}{2}} \end{aligned}$$

dove $\alpha = \frac{\theta}{\pi} + \frac{1}{2}$ e $\beta = -\frac{\theta}{\pi} + \frac{1}{2}$ quindi $\alpha + \beta = 1$. Allora,

$$f_X(x) = \frac{\pi}{B\left(\frac{1}{2}+\frac{\theta}{\pi},\ \frac{1}{2}-\frac{\theta}{\pi}\right)} e^{\alpha\pi x}\left(1+e^{\pi x}\right)^{-(\alpha+\beta)}.$$

Per la formula di riflessione di Eulero per la funzione Gamma (Weisstein 2002) si ha che

$$\Gamma\left(\tfrac{1}{2}+u\right)\Gamma\left(\tfrac{1}{2}-u\right) = \frac{\pi}{\cos(\pi u)}, \quad \text{con } u = \frac{\theta}{\pi},$$

allora

$$B\left(\tfrac{1}{2}+\tfrac{\theta}{\pi},\ \tfrac{1}{2}-\tfrac{\theta}{\pi}\right) = \Gamma\left(\tfrac{1}{2}+\tfrac{\theta}{\pi}\right)\Gamma\left(\tfrac{1}{2}-\tfrac{\theta}{\pi}\right) = \frac{\pi}{\cos(\theta)}.$$

Sostituendo nell'equazione di $f_X(x)$ si ha

$$f_X(x) = \cos(\theta)\, \frac{e^{(\theta+\pi/2)x}}{1+e^{\pi x}}, \quad x \in \mathbb{R}.$$

Infine, sapendo che $\cosh(x) = \frac{e^x+e^{-x}}{2}$, allora $1+e^{\pi x} = 2\cosh\left(\frac{\pi x}{2}\right) e^{\pi x/2}$, da cui si ottiene la forma finale, presentata da Carl Morris (Morris 1982), della funzione di densità di X con distribuzione secante iperbolica:

$$f_X(x) = \frac{\exp\{\theta x + \log(\cos(\theta))\}}{2\cosh(\pi x/2)}, \quad \text{con } \theta \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right). \tag{3.2}$$

Si è quindi dimostrato che $f(x)$ appartiene alla famiglia esponenziale naturale con:

$$h(x) = \frac{1}{2\cosh(\frac{\pi x}{2})}, \quad x \in \mathbb{R} \quad \text{e}$$

$$\psi(\theta) = -\log(\cos(\theta)), \quad \theta \in \left(-\tfrac{\pi}{2}, \tfrac{\pi}{2}\right).$$

Si visualizza in figura 3.2 la forma della distribuzione secante iperbolica generalizzata per diversi valori di θ scelti all'interno dello spazio parametrico, $\Theta \equiv \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$.

In particolare, si visualizza la densità di distribuzione per il seguente insieme di valori: $\theta = (-1.5,\ -1,\ -0.5,\ 0,\ 0.5,\ 1,\ 1.5)$.

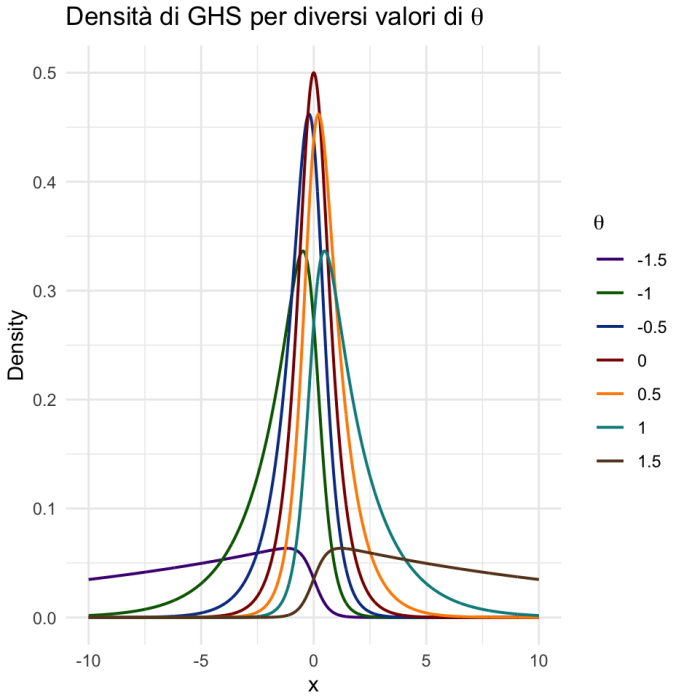

Figura 3.2: Densità della distribuzione secante iperbolica per diversi valori di θ

Dal grafico 3.2 si evidenzia la forma campanulare unimodale della distribuzione di interesse. Inoltre, per il valore $\theta = 0$ la distribuzione è simmetrica, mentre per i valori $\theta \neq 0$ la distribuzione è caratterizzata da asimmetria crescente all'aumentare del valore di $|\theta|$.

Moda

Si valuta la moda della distribuzione (Von Hippel 2005), cioè il valore di x che massimizza la funzione di densità della distribuzione (3.2).

Per facilità, si sceglie di derivare il logaritmo della funzione di densità, cioè

$$l(x;\theta) = \theta x + \log(\cos(\theta)) - \log(2) - \log(\cosh(\pi x/2))$$

da cui derivando si ottiene

$$\frac{\delta}{\delta x} l(x;\theta) = \theta - \frac{\pi}{2} \tanh\left(\frac{\pi x}{2}\right).$$

Allora, la moda si trova risolvendo per x, $\frac{\delta}{\delta x} l(x;\theta) = 0$:

$$x^* = \frac{2}{\pi} \text{arctanh}\left(\frac{2\theta}{\pi}\right).$$

Si nota che dato il vincolo su θ, cioè $|\theta| < \pi/2$, per l'argomento dell'arco tangente iperbolica vale $|\frac{2\theta}{\pi}| < 1$. Allora, la funzione arcotangente è definita e la moda della distribuzione corrisponde al valore:

$$x_{moda} = \frac{2}{\pi} \text{arctanh}\left(\frac{2\theta}{\pi}\right). \tag{3.3}$$

Momenti

Dalla funzione di densità ricavata, si evince che l'osservazione naturale x, funzione logistica π-scalata di y, è NEF-QVF, dove $\psi(\theta) = -\log(\cos(\theta))$ è la funzione cumulante (da 2.2). Allora, per le proprietà delle NEF [2.1.2], derivando la funzione cumulante $\psi(\theta)$, si ricavano il valore atteso e la funzione di varianza della distribuzione X come:

$$E[X] = \mu = \psi'(\theta) = \frac{\sin(\theta)}{\cos(\theta)} = \tan(\theta), \tag{3.4}$$

$$Var(X) = V(\mu) = \psi''(\theta) = \frac{\cos^2(\theta) + \sin^2(\theta)}{\cos^2(\theta)} = \frac{1}{\cos^2(\theta)} = \sec^2(\theta)$$
$$= 1 + \tan^2(\theta) = 1 + \mu^2. \tag{3.5}$$

Dal valore atteso di X (3.4) e per le proprietà delle NEF, per cui la media della distribuzione μ e il parametro naturale θ sono in relazione diretta, allora invertendo μ si ottiene $\theta = \mu^{-1} = \tan^{-1}(\mu) = \arctan(\mu)$.

Funzione Generatrice dei Momenti (MGF)

La funzione generatrice dei momenti di X è definita da

$$M_X(t) = \mathbb{E}[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx = \cos\theta \int_{-\infty}^{\infty} \frac{e^{(\theta+t)x}}{2\cosh\left(\frac{\pi x}{2}\right)}\, dx.$$

Data la seguente identità fondamentale (valida per ogni s con $|s| < \frac{\pi}{2}$):

$$\int_{-\infty}^{\infty} \frac{e^{sx}}{2\cosh\left(\frac{\pi x}{2}\right)}\, dx = \frac{1}{\cos s},$$

e applicandola con $s = \theta + t$, si ottiene che la MGF esiste per $|\theta + t| < \frac{\pi}{2}$, cioè per

$$t \in \Big(-\tfrac{\pi}{2} - \theta,\ \tfrac{\pi}{2} - \theta \Big),$$

e vale

$$M_X(t) = \cos\theta \cdot \frac{1}{\cos(\theta + t)} = \frac{\cos\theta}{\cos(\theta + t)}, \qquad |\theta + t| < \tfrac{\pi}{2}. \tag{3.6}$$

Qui la MGF esiste in un intervallo aperto contenente 0 perché $|\theta| < \frac{\pi}{2}$; più precisamente l'intervallo di esistenza è $t \in (-\frac{\pi}{2} - \theta, \frac{\pi}{2} - \theta)$.

Più semplicemente, per la proprietà delle NEF in 2.5, $M_X(t)$ si dimostra corrispondere all'esponenziale della funzione generatrice dei cumulanti:

$$M_X(t) = \exp\{-\log(\cos(\theta + t)) + \log(\cos(\theta))\} = \frac{\cos(\theta)}{\cos(\theta + t)}.$$

Asimmetria e Curtosi

Nella sezione 2.1.2 si sono descritti gli indici di asimmetria e curtosi derivati da Karl Pearson (Pearson 1905); nel caso della distribuzione secante iperbolica si ottengono i

seguenti risultati.

L'indice di asimmetria si deriva calcolando il cumulante di ordine 3 (2.7):

$$\psi^{(3)}(\theta) = \tfrac{\delta}{\delta\theta}(1+\tan^2(\theta)) = 2(1+\tan^2(\theta))\tan(\theta) = 2\mu V(\mu),$$

allora,

$$\beta_1 = \frac{(2\mu V(\mu))^2}{V(\mu)^3} \quad e \quad \gamma_1 = \frac{2\mu}{\sqrt{V(\mu)}}. \tag{3.7}$$

L'indice calcolato conferma l'idea già chiara dalla visualizzazione della forma della distribuzione in figura 3.2: la distribuzione secante iperbolica di parametro θ risulta essere non simmetrica dipendentemente al valore atteso della distribuzione e quindi al parametro θ. In particolare, la magnitudine dell'asimmetria è proporzionale alla grandezza del valore di θ nell'intervallo possibile ($|\theta| < \pi/2$) e la direzione dell'asimmetria dipende dal suo segno. Si calcola anche l'indice di curtosi per formalizzare l'andamento del picco e delle code della distribuzione; allora si calcola l'indice β_2 (2.1.2) e si riporta il risultato finale in termini di μ e $V(\mu)$:

$$\mu_4 = V(\mu)\left(5V(\mu) + 4\mu^2\right)$$

da cui,

$$\beta_2 = \frac{V(\mu)\left(5V(\mu) + 4\mu^2\right)}{V(\mu)^2} = \frac{5V(\mu) + 4\mu^2}{V(\mu)}. \tag{3.8}$$

Si possono quindi trarre alcune conclusioni sulla forma della distribuzione. Infatti, più la media μ si allontana da zero, più la distribuzione assume un andamento leptocurtico, con la curtosi che cresce da 5 a 9. Di conseguenza, scegliere valori di $|\theta|$lontani da zero produce distribuzioni con code più pesanti. Si osserva inoltre che la distribuzione presenta già una curtosi maggiore di 3 nel caso standard e simmetrico ($\theta = 0$), indicando code più pesanti rispetto a una normale standard (3.5).

3.2 Funzione di ripartizione

A partire dalla funzione di densità di X (3.2) si ricava la funzione di ripartizione, definita da $F(x) = \int_{-\infty}^{x} f(t)\,dt$ (Jacod e Protter 2004).

Tramite cambio di variabile: $u = \frac{\pi t}{2}$, con $dt = \frac{2}{\pi}du$ e $a = \frac{2\theta}{\pi}$, si ottiene

$$F(x) = \frac{\cos\theta}{\pi}\int_{-\infty}^{\pi x/2} \frac{e^{au}}{\cosh u}\,du.$$

Dalla sostituzione $w = e^{2u}$, $du = \frac{dw}{2w}$, $e^{au} = w^{a/2}$ e $\cosh u = \frac{w+1}{2\sqrt{w}}$, segue

$$F(x) = \frac{\cos\theta}{\pi}\int_{0}^{e^{\pi x}} \frac{w^{(a-1)/2}}{1+w}\,dw.$$

Ponendo infine $s = \frac{w}{1+w}$, da cui $w = \frac{s}{1-s}$ e $dw = \frac{ds}{(1-s)^2}$, si ottiene

$$F(x) = \frac{\cos\theta}{\pi}\int_{0}^{s(x)} s^{(a-1)/2}(1-s)^{-(a+1)/2}\,ds,$$

dove

$$s(x) = \frac{1}{1+e^{-\pi x}},$$

è l'inversa della funzione logit π-scalata.

Definendo $\alpha = \frac{a+1}{2} = \frac{1}{2} + \frac{\theta}{\pi}$, si riconosce la funzione Beta incompleta:

$$F(x) = \frac{\cos\theta}{\pi}\,B_{s(x)}(\alpha,\beta), \quad \text{con } \beta = 1-\alpha.$$

Infine, usando la relazione di riflessione di Eulero

$$B(\alpha, 1-\alpha) = \frac{\pi}{\sin(\pi\alpha)} = \frac{\pi}{\cos\theta},$$

segue

$$F(x) = \frac{B_{s(x)}(\alpha,1-\alpha)}{B(\alpha,1-\alpha)} = I_{s(x)}(\alpha,1-\alpha),$$

dove $I_z(a,b)$ è la Beta incompleta regolarizzata (Abramowitz e Stegun 1965).

Allora la funzione di ripartizione è

$$F(x) = I_{\frac{1}{1+e^{-\pi x}}}\left(\frac{1}{2}+\frac{\theta}{\pi},\,\frac{1}{2}-\frac{\theta}{\pi}\right). \tag{3.9}$$

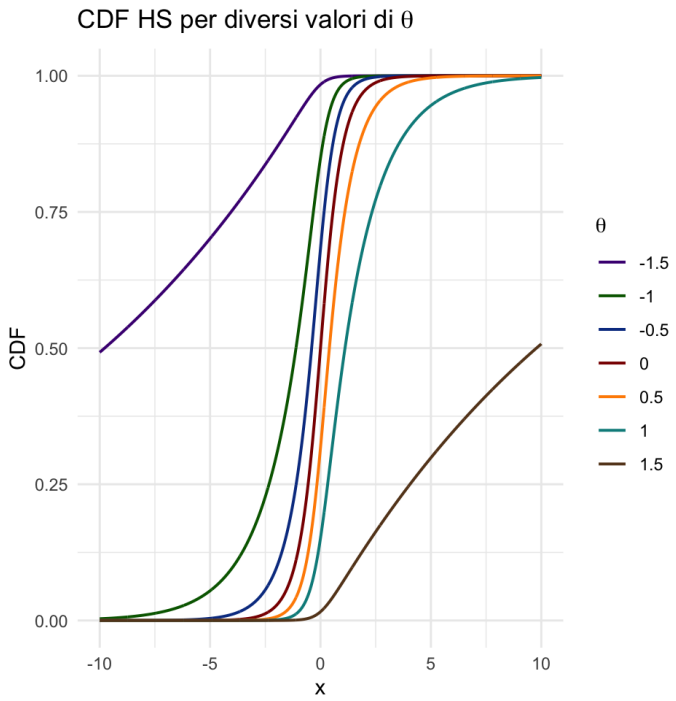

Figura 3.3: Funzione di ripartizione della distribuzione secante iperbolica per diversi valori di θ

3.3 Funzione quantile

Dalla funzione di ripartizione 3.9, si calcola la funzione quantile $Q(u)$, $u \in (0,1)$. Infatti, la funzione quantile corrisponde all'inversa della funzione di ripartizione:

$Q(u) = \inf\{x : F(x) \geq u\}$ (Jacod e Protter 2004).

Ponendo $F(x) = u$, si deriva la funzione quantile di X come:

$$I_{s(Q(u))}(\alpha, 1-\alpha) = u.$$

La funzione regolarizzata I_z è monotona crescente in $z \in (0,1)$ (Abramowitz e Stegun 1965), quindi invertibile rispetto a z, da cui segue

$$s(Q(u)) = qbeta(u;\, \alpha, 1-\alpha),$$

dove *qbeta* è la funzione quantile della distribuzione Beta$(\alpha, 1-\alpha)$.

Poiché $s(x) = \frac{1}{1+e^{-\pi x}}$, invertendo si ottiene

$$x = \frac{1}{\pi} \log \frac{s}{1-s}, \quad x \text{ è funzione logistica } \pi\text{-scalata di } s.$$

Allora,

$$Q(u) = \frac{1}{\pi} \log\left(\frac{qbeta(u;\, \alpha,\, 1-\alpha)}{1 - qbeta(u;\, \alpha,\, 1-\alpha)} \right), \qquad \alpha = \frac{1}{2} + \frac{\theta}{\pi} \tag{3.10}$$

è la funzione quantile della distribuzione secante iperbolica ricavata dalla trasformazione logistica π-scalata della distribuzione Beta.

3.4 Funzione per la generazione di numeri pseudocasuali

Metodo dell'inversione della funzione di ripartizione

Avendo ricavato in forma chiusa la formula della funzione di ripartizione in 3.9, è possibile implementare il metodo dell'inversione della funzione di ripartizione (ITM) per la generazione di numeri pseudocasuali dalla distribuzione continua di interesse.
L'algoritmo ITM viene allora implementato per generare dalla variabile X con distribuzione secante iperbolica di densità 3.2 definita su $x \in \mathbb{R}$. La procedura consiste in:

1. Generare casualmente un numero u da $U \sim U(0,1)$.
2. Risolvere rispetto ad x l'equazione $u = \int_{-\infty}^{x} f(t)dt$, ovvero $x = F^{-1}(u)$;
   in questo caso occorre risolvere $S = qbeta(U; \alpha, 1-\alpha) \sim \text{Beta}(\alpha, 1-\alpha)$.
3. Definire la trasformazione logit π-scalata per ottenere generazioni da X:

$$X = \frac{1}{\pi} \log \frac{S}{1-S}.$$

Si nota, in fase di utilizzo computazionale, che il metodo ITM per generare osservazioni pseudocasuali dalla distribuzione può portare a problemi di stabilità numerica nell'implementazione su `R`. Infatti, la distribuzione è definita su un intervallo dei parametri limitato a $\theta \in (-\pi/2, \pi/2)$ ed è definita come trasformazione di una distribuzione Beta. Quindi, quando i parametri della distribuzione Beta sottostante si avvicinano ai limiti estremi, la funzione π-logistica, che definisce la relazione tra la distribuzione secante iperbolica e la Beta, amplifica le deviazioni dei valori prossimi a 0 e 1. Questo può determinare instabilità numerica e la possibile generazione di valori estremi della distribuzione.

Nonostante le problematiche di stabilità numerica che può avere l'algoritmo e che saranno gestite computazionalmente, il metodo basato sulla trasformata della funzione di ripartizione sarà utilizzato in seguito per la simulazione di osservazioni da una variabile che segue distribuzione secante iperbolica.

3.5 Stima del parametro θ

Si considera una variabile casuale X con densità di probabilità parametrica:

$$f(x;\theta) = \frac{\exp(\theta x + \log\cos(\theta))}{2\cosh(\pi x/2)} = \frac{\cos(\theta)e^{\theta x}}{2\cosh(\pi x/2)}, \quad x \in \mathbb{R}, \quad \theta \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right). \tag{3.11}$$

Sia $X_1, X_2, \ldots, X_n$ un campione indipendente e identicamente distribuito da $f(x;\theta)$. Lo scopo è stimare θ tramite il metodo della massima verosimiglianza.

La funzione di verosimiglianza è definita come il prodotto delle densità valutate nei dati osservati:

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \frac{\cos(\theta)e^{\theta x_i}}{2\cosh(\pi x_i/2)}. \tag{3.12}$$

Poiché il termine $\prod_{i=1}^{n} 1/(2\cosh(\pi x_i/2))$ non dipende da θ, si considera costante ai fini della massimizzazione. La log-verosimiglianza, comoda per la derivazione, è:

$$\ell(\theta) = \log L(\theta) = n\log\cos(\theta) + \theta\sum_{i=1}^{n} x_i + h(\mathbf{x}). \tag{3.13}$$

Derivando $\ell(\theta)$ rispetto a θ:

$$\frac{d\ell}{d\theta} = -n\tan(\theta) + \sum_{i=1}^{n} x_i. \tag{3.14}$$

Ponendo la derivata prima uguale a zero si ottiene l'equazione di stima:

$$-n\tan(\hat{\theta}) + \sum_{i=1}^{n} x_i = 0 \quad \Rightarrow \quad \tan(\hat{\theta}) = \bar{X}, \tag{3.15}$$

dove $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ è la media campionaria.

$$\hat{\theta}_{\text{MLE}} = \arctan(\bar{X}). \tag{3.16}$$

Si verifica che 3.16 sia punto di massimo. La derivata seconda della log-verosimiglianza è:

$$\frac{d^2\ell}{d\theta^2} = -n\sec^2(\theta) < 0 \quad \forall\theta \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right), \tag{3.17}$$

quindi $\hat{\theta}_{\text{MLE}}$ corrisponde a un massimo globale, poiché la derivata seconda è negativa per ogni θ nell'intervallo ammissibile $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$.

L'informazione di Fisher per una singola osservazione è:

$$I_1(\theta) = -\mathbb{E}\left[\frac{d^2}{d\theta^2}\log f(X;\theta)\right] = \sec^2(\theta). \tag{3.18}$$

Per n osservazioni indipendenti:

$$I_n(\theta) = n\sec^2(\theta). \tag{3.19}$$

La varianza asintotica dello stimatore di massima verosimiglianza è quindi:

$$\text{Var}(\hat{\theta}_{\text{MLE}}) \approx \frac{1}{I_n(\theta)} = \frac{\cos^2(\theta)}{n}. \tag{3.20}$$

Questo risultato consente di quantificare la precisione dello stimatore in funzione della dimensione del campione e del reale valore del parametro.

Data la proprietà asintotica di normalità dello stimatore di massima verosimiglianza:

$$\hat{\theta}_{\text{MLE}} \overset{a}{\sim} N\left(\theta, \frac{\cos^2(\theta)}{n}\right) \quad \text{per } n \text{ grande}, \tag{3.21}$$

si può costruire un intervallo di confidenza al $100(1-\alpha)\%$ come:

$$\hat{\theta}_{\text{MLE}} \pm z_{1-\alpha/2}\sqrt{\frac{\cos^2(\hat{\theta}_{\text{MLE}})}{n}}, \tag{3.22}$$

dove $z_{1-\alpha/2}$ è il quantile della normale standard.

3.5.1 Proprietà dello stimatore

- **Consistenza:** poiché $\hat{\theta}_{\text{MLE}} = \arctan(\bar{X})$ e $\bar{X} \xrightarrow{p} \mathbb{E}[X]$ per $n \to \infty$, lo stimatore è consistente per θ.
- **Distorsione:** lo stimatore è asintoticamente non distorto, ovvero

$$\mathbb{E}[\hat{\theta}_{\text{MLE}}] = \theta + O(1/n).$$

- **Efficienza:** essendo uno stimatore di massima verosimiglianza, $\hat{\theta}_{\text{MLE}}$ è asintoticamente efficiente, raggiungendo il limite di Cramér-Rao.

3.6 Distribuzione Secante Iperbolica Standard

Si studiano ora le caratteristiche della distribuzione secante iperbolica quando il parametro θ (ricordando che $|\theta| < \pi/2$) è posto pari a 0. Per $\theta = 0$, si ha $\alpha = \frac{1}{2}$ e $\beta = \frac{1}{2}$ nella distribuzione di y, cioè $y \sim \text{Beta}(\frac{1}{2}, \frac{1}{2})$ in 3.1.

Ne deriva una distribuzione simmetrica con una classica forma unimodale, con media e mediana pari a 0 e con interessanti analogie con la distribuzione normale standard, analizzate nella sezione 3.6.4.

3.6.1 Funzione di densità e funzione di ripartizione

La funzione di densità diventa

$$f_X(x) = \frac{1}{2\cosh(\frac{\pi x}{2})} = \frac{1}{2}\text{sech}\left(\frac{\pi x}{2}\right), \quad x \in \mathbb{R} \tag{3.23}$$

semplicemente sostituendo $\theta = 0$ in 3.2.

La funzione di ripartizione è pari a

$$F(x) = I_{\frac{1}{1+e^{-\pi x}}}\left(\tfrac{1}{2}, \tfrac{1}{2}\right) = \frac{2}{\pi}\arcsin\left(\sqrt{\frac{1}{1+e^{-\pi x}}}\right) = \frac{2}{\pi}\arctan\Big(\tanh\frac{\pi x}{4}\Big) + \frac{1}{2},$$

semplificando

$$F(x) = \frac{2}{\pi} \arctan\left(\frac{e^{\frac{\pi x}{4}} - e^{-\frac{\pi x}{4}}}{e^{\frac{\pi x}{4}} + e^{-\frac{\pi x}{4}}}\right) + \frac{1}{2}$$

$$= \frac{2}{\pi} \arctan\left(\frac{e^{\frac{\pi x}{2}} - 1}{e^{\frac{\pi x}{2}} + 1}\right) + \frac{1}{2}$$

dato $e^{\pi x/2} > 0$, si applica l'identità trigonometrica $\arctan(\frac{u-1}{u+1}) = \arctan(u) - \frac{\pi}{4}, \quad u > 0$:

$$F(x) = \frac{2}{\pi}\left(\arctan(e^{\frac{\pi x}{2}}) - \frac{\pi}{4}\right) + \frac{1}{2}$$

Allora, si ottiene

$$F(x) = \frac{2}{\pi}\left(\arctan\left(e^{\frac{\pi x}{2}}\right)\right). \tag{3.24}$$

Infine, si inverte la funzione di ripartizione e si ottiene la funzione quantile ponendo $u = F(x)$:

$$Q(u) = \frac{2}{\pi} \ln\left(\tan\left(\frac{\pi}{2}u\right)\right). \tag{3.25}$$

Ne segue che la mediana è $Q(\frac{1}{2}) = 0$ infatti $\tan(\pi/4) = 1$.

Si visualizzano nella figura 3.4 rispettivamente la funzione di densità, la funzione di ripartizione e la funzione quantile della distribuzione secante iperbolica standard:

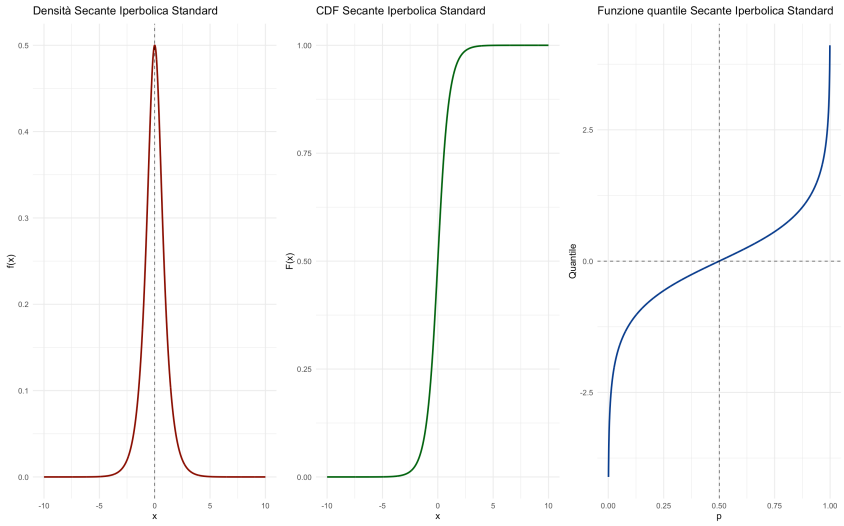

Figura 3.4: Funzione di densità, funzione di ripartizione e funzione quantile della distribuzione secante iperbolica standard

3.6.2 Momenti

Allora, i momenti di X (3.23) valgono, per $\theta = 0$:

$\mu = \tan(0) = 0$ e $V(\mu) = \sec^2(0) = 1$.

In questo caso si parla di distribuzione secante iperbolica standard, con forma campanulare unimodale e simmetrica in $\mu = 0$.

3.6.3 Funzione Generatrice dei momenti e Funzione caratteristica

La funzione generatrice dei momenti, calcolata sostituendo $\theta = 0$ in 3.6, è pari a

$$M_X(t) = \frac{1}{\cos(t)}, \quad |t| < \pi/2.$$

La funzione caratteristica, cioè $\chi_X(t) = E[e^{itX}]$, nel caso di famiglie esponenziali naturali si ottiene nella forma $\psi_X(t) = \exp\{\psi(\theta + it) - \psi(\theta)\}$.

Per la distribuzione secante iperbolica si ha funzione del cumulante pari a

$\psi(\theta) = -\log(\cos(\theta)) = \log(\sec(\theta))$, allora

$$\psi_X(t) = \exp\{\log(\sec(it)) - \log(\sec(0))\} = \sec(it) = \text{sech(t)}.$$

Si nota che la funzione di densità di probabilità può essere ottenuta dalla funzione caratteristica tramite una trasformazione di scala: $f(x) = \frac{1}{2}\chi(\frac{\pi}{2}x)$.

3.6.4 Confronto con normale standard

Dall'analisi condotta sulla distribuzione secante iperbolica standard, con parametro naturale $\theta = 0$, si evincono forti analogie con la distribuzione normale standard.

A partire dalla forma della distribuzione, entrambe le distribuzioni condividono la simmetria rispetto allo zero, la media nulla e la varianza unitaria ma differiscono nella modellazione delle code. In particolare, entrambe sono unimodali e presentano un picco centrale

in corrispondenza della media, ma differiscono nella forma e nella probabilità assegnata ai valori estremi: la normale standard possiede code più leggere che decadono secondo una legge gaussiana $e^{-\frac{x}{2}}$, mentre la distribuzione secante iperbolica presenta code più pesanti con decadimento esponenziale $e^{-\frac{\pi|x|}{2}}$.

Inoltre, la distribuzione secante iperbolica ha un picco centrale più pronunciato, riflesso di una curtosi maggiore. La curtosi descrive la concentrazione dei valori attorno alla media e la frequenza dei valori estremi, rispetto a una distribuzione normale che presenta curtosi pari a 3. Sostituendo $\theta = 0$ e quindi $\mu = 0$ e $V(\mu) = 1$ in 3.8 si ottiene $\beta_2 = 5$ che è indice di leptocurtosi cioè di una distribuzione con un picco alto con dati altamente concentrati attorno alla media e code pesanti con maggiore frequenza di valori estremi.

Dal confronto grafico della funzione di densità e della funzione di ripartizione delle due distribuzioni, riportato nella figura 3.5, si conferma la differenza in termini di leptocurtosi della distribuzione secante iperbolica:

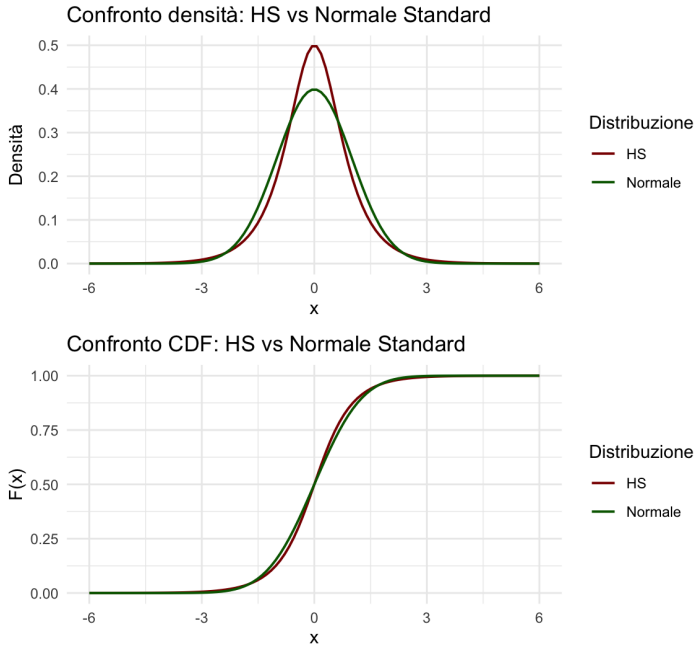

Figura 3.5: Funzione di densità e funzione di ripartizione della distribuzione secante iperbolica standard e della distribuzione normale standard a confronto

Simulazione di campionamento per il confronto

Si visualizza in questo paragrafo un confronto tra campioni: un campione generato tramite algoritmo ITM definendo il parametro $\theta = 0$ e un campione generato dalla funzione di `R` *rnorm* definendo i parametri `mean` $= 0$ e `sd` $= 1$.
Si riporta un confronto grafico (3.6, 3.7) tra un campione di osservazioni indipendenti dalla distribuzione secante iperbolica standard e un campione i.i.d. dalla normale standard. Il confronto evidenzia le analogie e le differenze fondamentali tra le due distribuzioni.

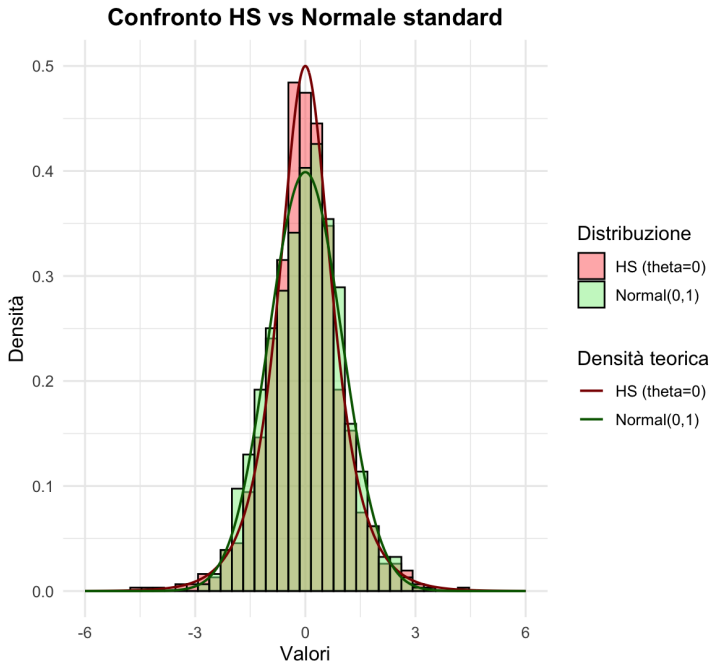

Figura 3.6: Istogrammi di distribuzione dei campioni generati da HS($\theta = 0$) e N(0,1)

L'istogramma 3.6 conferma che entrambe le distribuzioni condividono le proprietà fondamentali di simmetria rispetto alla media ($\mu = 0$) e mediana pari a 0. Tuttavia, la distribuzione iperbolica secante presenta un picco ragionevolmente più alto, con maggiore densità modale, rispetto alla distribuzione normale, indicando una maggiore concentrazione di massa probabilistica attorno alla media.

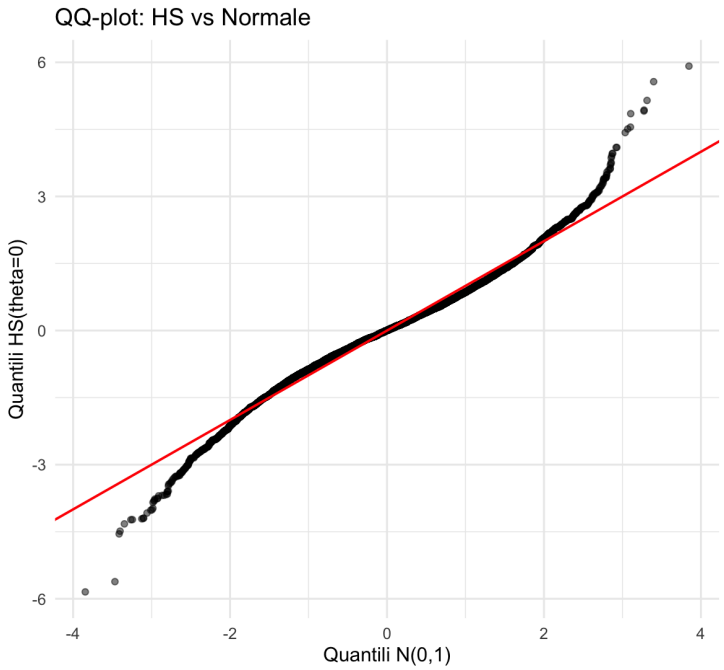

Figura 3.7: QQ-plot

Contemporaneamente, il diagramma quantile-quantile 3.7 mette a confronto i quantili teorici della normale standard (ascissa) con i quantili campionari della HS (se le due distribuzioni fossero identiche, i punti si disporrebbero lungo la bisettrice rossa). Il grafico fornisce informazioni sul fatto che la distribuzione secante iperbolica mostra code più pesanti della distribuzione normale: i punti giacciono approssimativamente sulla bisettrice nella regione centrale, che contiene la massa principale della distribuzione. Nelle code della distribuzione, invece, si osserva una sistematica deviazione al di sopra (a destra) e al di sotto (a sinistra) della bisettrice. Questo pattern conferma la presenza di code pesanti nella distribuzione iperbolica secante, con una probabilità significativamente maggiore di osservare valori estremi rispetto alla distribuzione normale.

Quindi, la distribuzione HS fornisce una modellizzazione più appropriata per fenomeni caratterizzati da eccesso di curtosi, capace di catturare sia l'elevata concentrazione centrale sia la maggiore massa nelle code rispetto alla distribuzione normale che potrebbe invece sottostimare simultaneamente la frequenza dei valori molto vicini alla media e la

probabilità di eventi estremi quando i dati provengono da una distribuzione leptocurtica.

3.7 Distribuzione Secante Iperbolica Generalizzata (GHS)

Dalla teoria di Morris, (Morris 1982), tramite convoluzioni della distribuzione secante iperbolica standard si genera la distribuzione Secante Iperbolica Generalizzata (GHS). Considerando la distribuzione secante iperbolica standard in 3.6, di parametro $\theta = 0$, cioè tale che $f_{1,0}(x) = \frac{1}{2\cosh\left(\frac{\pi x}{2}\right)}$, si vuole definire la densità della convoluzione di f con se stessa r volte:

$$f_{r,0}(x) = \underbrace{f_{1,0}(x) \cdots f_{1,0}(x)}_{r \text{ volte}}.$$

Il prodotto delle funzioni di densità della secante iperbolica standard definisce la funzione di densità di $X = X_1 + ... + X_r$ dove $X_1, ..., X_r$ sono variabili secanti iperboliche indipendenti con funzione caratteristica della forma

$$\chi(t) = \text{sech(t)}^r, \quad r > 0,$$

cioè la convoluzione r-esima di una variabile secante iperbolica.

In Fischer 2013 viene derivata la funzione di densità di X come sommatoria di variabili secanti iperboliche, ottenendo:

$$f(x; \theta, r) = \frac{2^{r-2}}{\pi\Gamma(r)} |\Gamma\left(\tfrac{r}{2} + i\tfrac{x}{2}\right)|^2 \exp\{\theta x + r \log\cos(\theta)\}, \quad |\theta| < \pi/2. \tag{3.26}$$

Una parametrizzazione più conveniente, suggerita in Morris 1983, è data da $\lambda \equiv \tan(\theta) \in \mathbb{R}$, da cui riscrivendo 3.26:

$$f(x; \lambda, r) = \left(\frac{1}{\sqrt{1+\lambda^2}}\right)^r \frac{2^{r-2}}{\pi\Gamma(r)} |\Gamma\left(\tfrac{r}{2} + i\tfrac{x}{2}\right)|^2 \exp\{\arctan(\lambda)x\}. \tag{3.27}$$

La funzione generatrice dei momenti nella forma generalizzata della secante iperbolica deriva da

$$M_X(t) = \int_{-\infty}^{\infty} e^{tX} \frac{e^{\theta x}}{\cos(\theta)^{-r}} |\Gamma\left(\tfrac{r}{2} + i\tfrac{x}{2}\right)|^2 dx,$$

e in Fischer 2013 si dimostra valere

$$M_X(t) = \left(\frac{(\sqrt{1+\lambda^2})^{-1}}{\cos(\arctan(\lambda)+t)} \right)^2 = \left(\frac{\cos(\theta)}{\cos(\theta+t)} \right)^2. \tag{3.28}$$

Anche nella forma generalizzata, tutti i momenti esistono finiti. In particolare:

$$E[X] = \mu = r \cdot \lambda = r \cdot \tan(\theta);$$

$$E[X^2] = V(\mu) = \mu^2/r + r.$$

La distribuzione descritta, nella parametrizzazione 3.27 è detta $NEF - GHS(r, \lambda)$ ed è dunque la distribuzione di $X_1 + ... + X_r$ per r intero se $X_i,\ \forall i \in (1, .., r)$ sono indipendenti e identicamente distribuite come $NEF - GHS(1, \lambda)$, cioè con $r = 1$, derivata nella sezione 3.1.

Capitolo 4

Regressione Secante Iperbolica

Nel Capitolo 3, si è introdotta e presentata la distribuzione secante iperbolica: una distribuzione continua su $\mathbb{R}$ della famiglia naturale esponenziale ricavata come trasformazione logit π-scalata di una distribuzione Beta. Se ne sono dimostrate le funzioni caratterizzanti e i momenti e si è scesi nel dettaglio del caso più semplice della distribuzione in esame: il caso della distribuzione Secante Iperbolica Standard con valore atteso pari a 0 e varianza pari a 1.

In questo capitolo, si presenta la metodologia utilizzata per la stima di un modello lineare generalizzato per la distribuzione NEF-GHS, implementato con due diverse funzioni di legame. Infine, si descrive la metodologia per la stima di un modello di quasi-verosimiglianza basato sulle condizioni di secondo ordine della distribuzione GHS.

4.1 Regressione Secante Iperbolica con *link* canonico

Nel contesto della distribuzione secante iperbolica generalizzata (3.7) con parametri $r = 1$ e $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2})$ che, come discusso nella sezione 2, appartiene alle famiglie esponenziali naturali (2.2), si implementa un modello lineare generalizzato (2.2).

Si specificano allora le quantità fondamentali del modello:

- $Y_1, ...Y_n \overset{ind}{\sim} GHS(1, \theta_i)$ variabile dipendente secante iperbolica, con funzione di densità
$$f(y_i; \theta_i) = \frac{\exp\{\theta_i y_i + \log(\cos(\theta_i))\}}{2\cosh(\frac{\pi y_i}{2})}, \quad \theta_i \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right), \ \forall i \ ;$$
- il predittore lineare $\eta_i = X_i'\beta$ con $\beta = (\beta_1, ..., \beta_p)$ coefficienti di regressione;
- funzione *link* canonico g tale che $\mu = g^{-1}(\theta)$ cioè $g(\mu) = \theta = \arctan(\mu)$.

Specificato il modello, si scrive la finzione di log-verosimiglianza del modello come:

$$l(\beta) = \sum_{i=1}^{n} \theta_i y_i + \log(\cos(\theta_i)) - 2\cosh\left(\tfrac{\pi y_i}{2}\right)$$

Funzione di *score* per la distribuzione GHS

Al fine di stimare i coefficienti di regressione β e di implementare poi l'algoritmo di massimizzazione Newton-Raphson, si derivano la funzione di *score*, come derivata prima della log-verosimiglianza, e l'informazione attesa di Fisher. Si noti che, avendo scelto il *link* canonico per l'implementazione del modello lineare generalizzato, la matrice di informazione attesa coincide con la matrice di informazione osservata.

La funzione di *score* 2.16 è

$$l_*(\beta) = \sum_{i=1}^{n} (y_i - \mu_i) \frac{\delta \theta_i}{\delta \beta_r} x_{ir}, \quad i = 1, ..., n, \quad r = 1, ..., p$$

Ricordando che $\mu_i = \tan(\theta_i)$ e $\theta_i = \eta_i = x_i'\beta$, si ha che $\frac{\delta \mu_i}{\delta \eta_i} = \frac{\delta}{\delta \eta_i} \tan(\eta_i)$. Allora,

$$l_*(\beta) = \sum_{i=1}^{n} (y_i - \mu_i) x_{ir}$$

e sostituendo,

$$l_*(\beta) = \sum_{i=1}^{n} (y_i - \mu_i)(1 + \mu_i^2) x_{ir}, \quad V(\mu_i) = 1 + \mu_i^2. \tag{4.1}$$

Allora, lo stimatore di massima verosimiglianza per i coefficienti di regressione di un modello lineare generalizzato per la distribuzione secante iperbolica generalizzata risolve

il sistema di equazioni definito, in forma matriciale, da

$$\underset{p\times n}{X}'(y-\mu)=0 \tag{4.2}$$

Matrice dell'informazione per la distribuzione GHS

Si considera ora la seconda derivata della log-verosimiglianza per la derivazione della matrice dell'informazione attesa 2.17:

$$i_{rs}=\sum_{i=1}^{n}\frac{x_{ir}x_{is}}{V(\mu_i)g'(\mu_i)^2}.$$

In questo caso $g(\mu_i)=\arctan(\mu_i)$ e derivando $g'(\mu_i)=\frac{1}{1+\mu_i^2}=\frac{1}{V(\mu_i)}$. Allora, si ottiene

$$i_{rs}=\sum_{i=1}^{n}x_{ir}x_{is}V(\mu_i). \tag{4.3}$$

In forma matriciale si scrive come $I=X'WX$ dove W è la matrice dei pesi con forma

$$\underset{n\times n}{W}=\begin{bmatrix} V(\mu_1) & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & V(\mu_n) \end{bmatrix}$$

Algoritmo di Newton–Raphson per la distribuzione GHS

La regola di aggiornamento per l'algoritmo di Newton-Raphson è:

$$\begin{aligned}\beta^{(t+1)} &= \beta^{(t)} - I[\beta^{(t)}]^{-1}l_*(\beta^{(t)}) \\ &= \beta^{(t)} - (X'WX)^{-1}X'W(y-\mu).\end{aligned} \tag{4.4}$$

Inserendo il passo di aggiornamento in un algoritmo iterativo (IRLS) si ottengono i seguenti oggetti a ogni t-esima iterazione:

$$\mu_i^{(t)}=\tan(x_i'\beta^{(t)}),$$

$$w_i^{(t)}=1+\mu_i^{2(t)},$$

$$z_i^{(t)}=\eta_i^{(t)}+\frac{y_i-\mu_i^{(t)}}{w_i^{(t)}}.$$

Sostituendo le quantità di interesse nel passo iterativo di un Newton Raphson, l'aggiornamento diventa:

$$\begin{aligned}\beta^{(t+1)} &= \beta^{(t)} - I[\beta^{(t)}]^{-1} l_*(\beta^{(t)}) \\ &= \beta^{(t)} - (X'WX)^{-1}X'W(y-\mu).\end{aligned} \tag{4.5}$$

In sintesi, l'algoritmo procede come segue:

1. Inizializzare $\beta^{(0)}$.
2. Calcolare $\theta_i^{(t)} = \eta_i^{(t)} = X_i'\beta^{(t)}$ e $\mu_i^{(t)} = \tan(\theta_i)$, da cui $V(\mu_i^{(t)}) = 1 + \mu_i^{2(t)}$
3. Calcolare i pesi $w_i^{(t)} = \frac{\left(V(\mu_i^{(t)})\right)^2}{V(\mu_i^{(t)})} = w_i^{(t)} = V(\mu_i^{(t)})$ e le variabili aggiustate $z_i^{(t)} = \theta_i^{(t)} + \frac{y_i - \mu_i^{(t)}}{1+\mu_i^{(t)2}}$.
4. Risolvere la regressione lineare pesata con pesi $w_i^{(t)}$ e ottenere quindi $\beta^{(t+1)} = (X'W^{(t)}X)^{-1}X'W^{(t)}z^{(t)}$.
5. Ripetere i passi da 2 a 5 fino a convergenza.

4.1.1 Limiti dell'utilizzo del *link* canonico

L'adozione della funzione di legame canonica per la stima di un modello lineare generalizzato (GLM) associato alla distribuzione Secante Iperbolica implica che valga la relazione $\theta_i = \eta_i,\ \forall i \in \{1, ..., n\}$ tra il parametro naturale θ e il predittore lineare η.
Dalla definizione della funzione di densità di $Y \sim GHS$ (3.2), lo spazio parametrico Θ in cui varia il parametro naturale θ è l'intervallo $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$.
Imponendo il *link* canonico si ottiene $\theta_i = \eta_i = X_i'\beta \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. In forma matriciale, le condizioni di ammissibilità possono essere espresse come $X\beta < \frac{\pi}{2}\mathbf{1}$ e $-X\beta < \frac{\pi}{2}\mathbf{1}$, cioè lo spazio dei parametri consentiti coincide con l'intersezione di $2n$ semispazi lineari, definendo un poliedro convesso aperto. Questo implica che, nella procedura di stima dei coefficienti del modello, non tutti i vettori di β sono ammissibili ma solo quelli che mantengono ogni i-esimo predittore all'interno dello spazio parametrico. In particolare, se l'r-esimo

predittore $X_r,\ r = 1, ..., p$, assume valori elevati o non disciplinati, lo spazio fattibile per il coefficiente associato può essere molto ridotto, riducendo la flessibilità del modello e complicando il processo di ottimizzazione numerica.

Per questo motivo, al fine di superare le restrizioni imposte dal *link* canonico e ampliare lo spazio dei coefficienti ammissibili, è opportuno ricorrere a una funzione di legame alternativa.

4.2 Regressione Secante Iperbolica con *link* identità

Un'alternativa naturale al *link* canonico, che consente di evitare le restrizioni sullo spazio parametrico, è il *link* identità, tramite cui il predittore lineare coincide direttamente con la media della variabile risposta.

La specificazione delle quantità fondamentali del modello è analoga alla specificazione in caso di *link* canonico (4.1), con la differenza che la funzione di legame assume ora la forma

$$g(\mu_i) = \mu_i = \tan(\theta_i) = \eta_i,\ \forall i \in \{1, ..., n\}.$$

Allora, a partire dalla funzione di log-verosimiglianza 2.13, che resta invariata, si derivano la funzione di score, la matrice dell'informazione osservata e la matrice dell'informazione attesa di Fisher. A differenza di quanto avviene con il *link* canonico, nel caso di una specificazione del modello lineare generalizzato con *link* alternativo, le due matrici di informazione non coincidono.

Funzione di score per la distribuzione GHS

Si derivano la funzione di score (2.16), come derivata prime della log-verosimiglianza $l(\theta) = \sum_{i=1}^{n} \theta_i y_i + \log(\cos(\theta_i))$, allora

$$l_*(\beta) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{V(\mu_i)} \frac{x_{ir}}{g'(\mu_i)},$$

dove $g'(\mu_i) = 1$ dato che $g(\mu_i) = \mu_i$. Si ottiene quindi

$$l_*(\beta) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{1 + \mu_i^2} x_{ir}$$

come r-esimo elemento della funzione di score.

Allora, lo stimatore di massima verosimiglianza per i coefficienti di regressione del modello risolve il sistema di equazioni definito, forma matriciale, da

$$\underset{p\times n}{X'} \underset{n\times n}{W} (y - \mu) = 0, \quad \text{con } W = diag\left(\tfrac{1}{1+\mu_i^2}\right). \tag{4.6}$$

Matrice dell'informazione osservata

Si considera la derivata seconda della log-verosimiglianza e si calcola:

$$j_{rs} = -\frac{\delta^2 l}{\delta\beta_r \delta\beta_s} = -\sum_{i=1}^{n} x_{ir} \frac{\delta}{\delta\beta_s} \left(\frac{y_i - \mu_i}{1 + \mu_i^2}\right),$$

dove

$$\frac{\delta}{\delta\beta_s} = \frac{d}{d\mu_i}\frac{\delta\mu_i}{\delta\beta_s} = \frac{d}{d\mu_i} x_{is};$$
$$\frac{d}{d\mu_i}\left(\frac{y_i - \mu_i}{1 + \mu_i^2}\right) = \frac{\mu_i^2 - 2\mu_i y_i - 1}{(1 + \mu_i^2)^2}.$$

Sostituendo si ottiene:

$$j_{rs} = \sum_{i=1}^{n} x_{ir} x_{is} w_i, \quad \text{con } w_i = -\frac{\mu_i^2 - 2\mu_i y_i - 1}{(1 + \mu_i^2)^2}.$$

Matrice dell'informazione attesa

Si considera ora il valore atteso di j_{rs}:

$$i_{rs} = E[j_{rs}] = \sum_{i=1}^{n} x_{ir} x_{is} E[w_i] = \sum_{i=1}^{n} x_{ir} x_{is} \frac{1 - \mu_i^2 + 2\mu_i E[y_i]}{(1 + \mu_i^2)^2},$$

dato che $E[y_i] = \mu_i$, si ottiene

$$i_{rs} = \sum_{i=1}^{n} \frac{x_{ir} x_{is}}{1 + \mu_i^2}.$$

In forma matriciale si può scrivere $I = X'WX$ con $w_i = \frac{1}{1+\mu_i^2}$.

Allora nell'implementazione dell'algoritmo IRLS 2.3.2 le quantità definite sono

$$\mu = X\beta, \quad V(\mu) = 1\mu^2, \quad \frac{\delta\mu}{\delta\eta} = 1;$$

$$w_i = \frac{1}{1+\mu_i^2} \quad \text{e} \quad z_i = \eta_i + \frac{y_i - \mu_i}{g'(\mu_i)} = y_i.$$

Implementazione su R dell'algoritmo IRLS

```r
irls <- function(X, y, beta_init = NULL, max_iter = 1000,
tol = 1e-8, verbose = TRUE){
n <- nrow(X)
p <- ncol(X)
eps <- 1e-6  #per evitare divergenza della tangente

# inizializzazione di beta
if(is.null(beta_init)){
  beta <- rep(0, p)
} else {
  beta <- beta_init
}

for(t in 1:max_iter){
  mu <- X %*% beta
  var_mu <- 1 + mu^2
  dmu <- 1

  # variabile aggiustata
  z <- mu + (y - mu)/dmu

  # pesi
  W <- diag(as.numeric(dmu^2 / var_mu))

  # passo di aggiornamento Newton-Raphson
  XtWX <- t(X) %*% W %*% X
  XtWz <- t(X) %*% W %*% z
  beta_new <- solve(XtWX, XtWz)

  # criterio di arresto
  if(norm(beta_new - beta, "2") < tol * (1 + norm(beta, "2"))){
    if(verbose){
      cat("Convergenza in", t, "iterazioni\n")
    }
    cov_beta <- solve(XtWX)
    se_beta <- sqrt(diag(cov_beta))
    return(list(beta = beta_new,
                se = se_beta,
                cov = cov_beta,
                converged = TRUE,
                iter = t))
  }

  beta <- beta_new
```

```
  }

  if(verbose){
    cat("Attenzione: massimo numero di iterazioni raggiunto\n")
  }
  return(list(beta = beta,
              se = NA,
              cov = NA,
              converged = FALSE,
              iter = max_iter))
}
```

4.3 Modello di Quasi-Verosimiglianza per dati reali

Nel contesto di una regressione secante iperbolica, le ipotesi deboli per l'implementazione di un modello di quasi verosimiglianza 2.5 sono:

- $E[Y_i] = \mu_i = \tan(\theta_i)$;
- $Var(Y_i) = \phi(1 + \mu_i^2)$, $\phi > 0$;
- Y_i e Y_j indipendenti se $i \neq j$.

Allora, le equazioni stima per i coefficienti β del modello sono non distorte. Inoltre, si mantiene la consistenza delle stime e la loro approssimazione normale per n elevato: $\hat{\beta} \sim N_p\left(\beta, (X'WX)^{-1}\right)$ dove $W = diag\left(\frac{1}{\phi(1+\mu_i^2)}\right)$.
Se ϕ è non noto, si stima consistentemente tramite $\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{1+\mu_i^2}$.
Allora, la matrice di covarianza asintotica di $\hat{\beta}$ si stima tramite

$$\widehat{\mathrm{Var}}(\hat{\beta}) = (X'\hat{W}X)^{-1},$$

con $\hat{W} = diag\left(\frac{1}{\hat{\phi}(1+\hat{\mu_i}^2)}\right)$.

4.3.1 Definizione della famiglia quasi-GHS in R

```
quasi.GHS <- function() {
  link_identity <- list(
    linkfun = function(mu) mu,
    linkinv = function(eta) eta,
```

```
    mu.eta = function(eta) rep(1, length(eta)),
    valideta = function(eta) TRUE,
    name = "identity"
  )

  structure(list(
    family = "quasi.GHS",
    link = "identity",
    linkfun = link_identity$linkfun,
    linkinv = link_identity$linkinv,
    mu.eta = link_identity$mu.eta,

    # Varianza come funzione di mu e phi
    variance = function(mu, phi = 1) phi * (1 + mu^2),

    # Residui di Pearson
    pears.resids = function(y, mu, wt, phi = 1) {
      2 * wt * ((y - mu)^2 / (phi * (1 + mu^2)))
    },

    aic = function(y, n, mu, wt, dev) 2 * sum(dev),

    initialize = expression({
      mustart <- y
    }),

    validmu = function(mu) TRUE,

    # Funzione per stimare phi
    phi.estimate = function(y, mu, wt, p) {
      sum(wt * (y - mu)^2 / (1 + mu^2)) / (sum(wt) - p)
    }

  ), class = "family")
}
```

SU R, si è così definita la funzione per l'utilizzo della famiglia di distribuzioni GHS come famiglia parametrica per la stima di un GLM, utilizzando la funzione di R `glm()`. Se ne mostra un esempio di utilizzo:

```
fit <- glm(y ~ ., family = quasi.GHS(), data=())
summary(fit)
```

In questo modo viene stimato un modello lineare generalizzato in cui il parametro di dispersione non è fissato a 1, ma viene stimato come media dei residui di Pearson al quadrato, così da poter modellare sovra o sotto dispersione.

Se si vuole impostare il valore di $\phi = 1$ fissato, si può specificare all'interno del comando `summary`:

```r
summary(fit, dispersion=1)
```

Così la `summary` mostra $\phi = 1$ e gli errori standard non vengono scalati dalla dispersione stimata. Non cambiano i coefficienti stimati, ma viene modificata solo la presentazione della dispersione e degli errori standard. Se ne visualizza l'*output* di `R` nell'applicazione al dataset reale in 5.7.

Capitolo 5

Risultati

In questo capitolo vengono presentati i risultati ottenuti dalla stima dei modelli lineari generalizzati, considerando due differenti funzioni di *link*, canonica e identità. I primi risultati si ottengono in un contesto simulativo mediante l'algoritmo di generazione di determinazioni pseudocasuali dalla distribuzione secante iperbolica generalizzata.
Successivamente, viene proposta un'applicazione a un dataset reale, in cui si stima un modello di quasi-verosimiglianza e si confrontano i risultati con quelli di un modello gaussiano.

5.1 Generazione di determinazioni pseudocasuali HS

Nella sezione 3.4 si è illustrato il metodo per la generazione di determinazioni pseudocasuali dalla distribuzione secante iperbolica. In questa sezione si simulano campioni dalla distribuzione di interesse e se ne analizzano i risultati per valutare l'adattamento dei dati simulati alla distribuzione teorica.
In `R` si definisce la funzione `rghs` che sviluppa l'algoritmo per generare da una variabile casuale con distribuzione secante iperbolica:

```r
rghs <- function(n = 1, theta = NULL, alpha = NULL, seed = NULL) {

  if (!is.null(seed)) {
    set.seed(seed)
  }

  if (is.null(alpha) && is.null(theta)) {
    stop("Specificare il valore di 'theta' o di 'alpha'.")
  }

  if (!is.null(alpha) && !is.null(theta)) {
    stop("pecificare il valore di 'theta' o di 'alpha', non entrambi.")
  }

  #Validazione dei parametri
  if (!is.null(alpha)) {
    if (!is.numeric(alpha) || length(alpha) != 1) {
      stop("Specificare 'alpha' scalare.")
    }
    if (alpha <= 0 || alpha >= 1) {
      stop("Specificare 'alpha' che soddisfi 0 < alpha < 1.")
    }
    shape_alpha <- alpha
  } else {
    if (!is.numeric(theta) || length(theta) != 1) {
      stop("Specificare 'theta' scalare.")
    }
    if (theta <= -pi/2 || theta >= pi/2) {
      stop("Specificare 'theta' che soddisfi -pi/2 < theta < pi/2.")
    }
    shape_alpha <- 1/2 + theta/pi
  }

  #Costante per stabilità numerica
  machine_epsilon <- .Machine$double.eps
  stability_threshold <- 1000 * machine_epsilon

  # Step 1: Generare da una v.c U(0,1)
  unif <- runif(n, min = 0, max = 1)

  unif <- pmax(unif, stability_threshold)
  unif <- pmin(unif, 1 - stability_threshold)

  # Step 2: Applicare la trasformazione inversa
  beta <- qbeta(
    p = unif,
    shape1 = shape_alpha,
    shape2 = 1 - shape_alpha
  )

  beta <- pmax(beta, stability_threshold)
  beta <- pmin(beta, 1 - stability_threshold)

  # Step 3: Applicare la trasformazione pi-logit
  ghs <- (1/pi) * log(beta / (1 - beta))

  return(ghs)
}
```

Si simulano in questo modo tre distinti campioni di osservazioni indipendenti e identicamente distribuite generate dalla densità di probabilità di una secante iperbolica. Si generano tre campioni di numerosità 10^4 per stime stabili e con diversi valori del parametro $\theta = (0,\ 0.5,\ -1)$ così da osservare il comportamento dei campioni nel caso simmetrico ($\theta = 0$), nel caso asimmetrico positivo ($\theta = 0.5$) e in un caso di asimmetria negativa più estremo in cui il valore del parametro si avvicina al valore limite dello spazio Θ ($\theta = 1$). Nel grafico 5.1, se ne confrontano le distribuzioni empiriche con quelle teoriche, definite da 3.2.

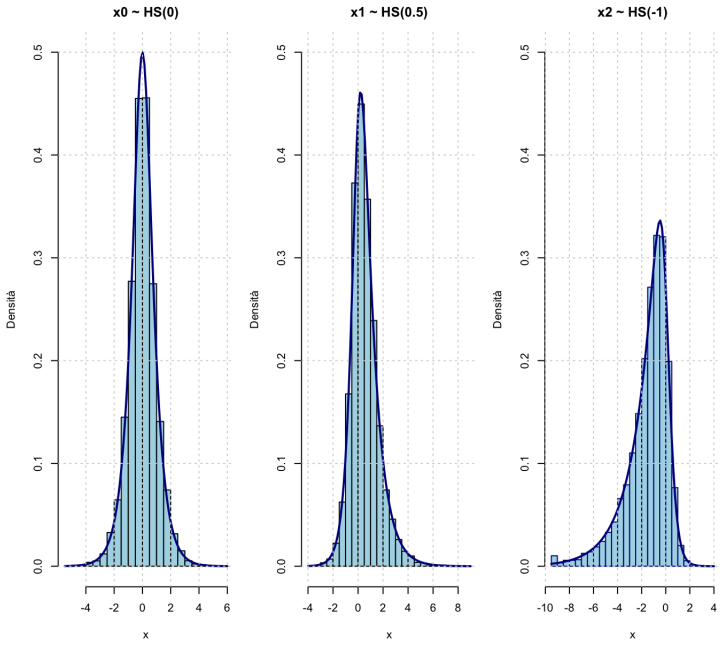

Figura 5.1: Confronto tra distribuzioni empiriche e teoriche per diversi valori di θ

Si evidenzia un comportamento delle distribuzioni dei campioni coerenti con le distribuzioni teoriche di riferimento (curve blu in 5.1). Nell'istogramma a destra, caratterizzato da un valore del parametro più elevato e vicino al *boundary* dello spazio parametrico Θ, si nota un picco di osservazioni sulla coda sinistra che può far presagire la presenza di valori estremi probabilmente dovuti alla forte asimmetria della distribuzione.

Calcolando le statistiche descrittive dei campioni e confrontandole con i valori teorici si

ottengono i risultati riportati nella tabella 5.1.

Tabella 5.1: Statistiche descrittive per 3 campioni: valori teorici Vs valori empirici

Campione	Statistica	Media	Varianza	Asimmetria	Curtosi
X_0	Teorica	0	1	0	2
	Empirica	0.006	1.009	0.015	1.781
X_1	Teorica	0.546	1.298	0.959	2.919
	Empirica	0.540	1.226	0.921	2.719
X_2	Teorica	-1.557	3.425	-1.682	4.832
	Empirica	-1.510	3.132	-1.480	2.978

I valori delle medie empiriche e dei coefficienti di asimmetria empirici coincidono bene con quelle teoriche per cui la forma della distribuzione dei campioni segue correttamente la distribuzione secante iperbolica, anche al variare del valore di θ. Nei valori della varianza e dell'indice di eccesso di curtosi si osservano leggere discrepanze, in generale ccomparabili con normali fluttuazioni campionarie, soprattutto nel caso di distribuzioni asimmetriche e con code lunghe. Comunque, i campioni simulati mantengono la forma leptocurtica tipica della distribuzione teorica di interesse.

Si implementa per un'ulteriore verifica, un test di casualità per testare l'ipotesi nulla H_0 per cui i dati seguono la distribuzione teorica $F(x)$. Il test utilizzato è l'Anderson-Darling test (Stephens 1974) che confronta la funzione di ripartizione empirica $F_n(x) = \frac{\{\#X_i \leq x\}}{n}$ con quella teorica $F(x)$ (3.9) dando peso agli scostamenti nelle code della distribuzione e usando una funzione di ponderazione per gli estremi. Si riportano i p-value del test per i tre campioni in esame nella tabella 5.2. I risultati mostrano che per i campioni X_0 ed

Tabella 5.2: P-value del test di Anderson-Darling per i tre campioni

	X_0	X_1	X_2
P-value	0.75	0.29	0.05

X_1 non c'è evidenza statistica sufficiente per rifiutare l'ipotesi nulla, indicando che questi campioni sono coerenti con la distribuzione teorica assunta. Per il campione X_2, il p-value risulta pari a 0.05, valore al limite convenzionale di significatività ($\alpha = 0.05$), suggerendo una possibile deviazione dalla distribuzione teorica.

Tuttavia, aumentando il numero di simulazioni nella generazione, il valore del p-value aumenta, correggendo la deviazione e producendo stime più stabili. Ad esempio, aumentando il numero di repliche a $5 \cdot 10^4$, il p-value del test aumenta ed è pari a 0.21.

Per le considerazioni fatte, si reputa l'algoritmo sviluppato tramite inversione della funzione di ripartizione sufficientemente accurato per la generazione di campioni pseudo casuali dalla distribuzione secante iperbolica generalizzata. Si utilizzerà in seguito la funzione definita `rghs` in fase di simulazione per lo sviluppo di un modello lineare generalizzato per una variabile target con distribuzione GHS.

5.2 Regressione Secante con *link* canonico

5.2.1 Stima del modello nullo con *link* canonico

Per introdurre l'approccio di stima, si implementa l'algoritmo di Newton–Raphson definito in 2.3.2 con simulazioni Monte Carlo al fine di stimare un modello lineare generalizzato. Come caso di studio preliminare si considera il modello nullo, ossia un modello che include unicamente l'intercetta. In particolare, si assume per la variabile risposta Y una distribuzione secante iperbolica generalizzata (GHS) e si specifica il coefficiente associato all'inter-

cetta pari a $\beta_0 = 0.5$. Il modello lineare generalizzato in esame risulta quindi: $g(\mu_i) = \beta_0$, dove la funzione di legame adottata è quella canonica nella forma $g(\mu_i) = \arctan(\mu_i)$.

Si genera un campione di dimensione $n = 100$ dalla distribuzione GHS e si stima il parametro β_0 mediante l'algoritmo di Newton–Raphson. Per valutare le proprietà delle stime, la procedura è stata replicata $R = 10^4$ volte attraverso simulazioni Monte Carlo.

Dall'analisi delle repliche si osserva che la media delle stime del coefficiente risulta pari a $\hat{\beta}_0 = 0.495$ con bias stimato pari a -0.0053 e la deviazione standard delle stime pari a 0.0877. Sulla base di tale risultato, si costruisce un intervallo di confidenza asintotico di livello 95% per β_0 utilizzando l'approssimazione normale, ottenendo $IC_{95\%}(\beta_0) = (0.323, 0.667)$.

Si visualizzano inoltre in 5.2 la distribuzione delle stime di β_0 nelle repliche Monte Carlo (prima tramite boxplot poi tramite istogramma) e gli intervalli di confidenza ansitotici.

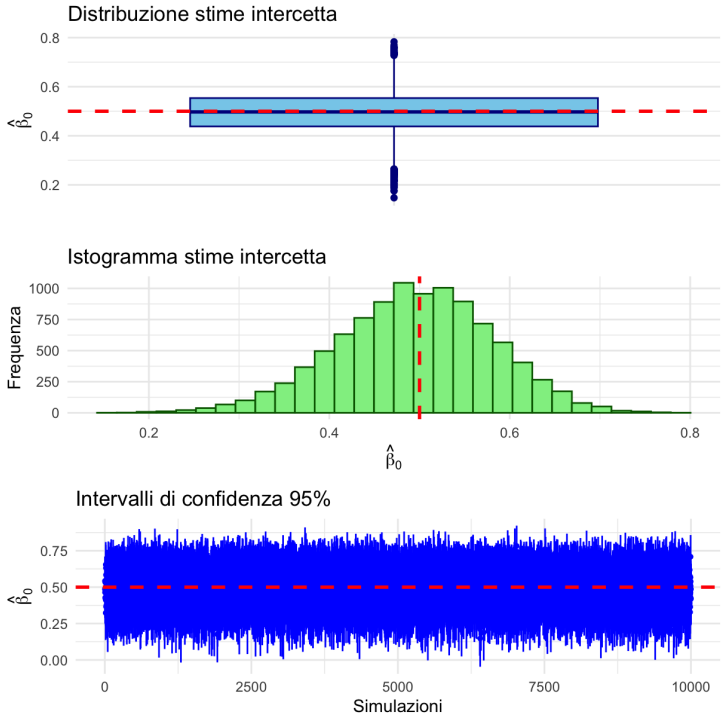

Figura 5.2: Stime del coefficiente β_0 e copertura del modello nullo

Infine, si conduce un'analisi di robustezza campionando per valori di n crescenti

ottenendo risultati molto simili a quelli riportati e visualizzati in 5.2.

5.2.2 Stima del modello con covariate

A partire dal modello nullo, si sviluppa un modello più complesso includendo covariate esplicative e se ne valutano i risultati ottenuti.

In particolare, si stima un modello ancora semplice con intercetta e due variabili esplicative continue, scelte con distribuzione normale standard.

La procedura iterativa è simile alla precedente, con la generazione di $n = 100$ osservazioni indipendenti dalla variabile target Y con distribuzione GHS e di $n = 100$ osservazioni indipendenti da due distribuzioni Normali standard. Si ripete il campionamento per $R = 10^4$ repliche Monte Carlo.

Si scelgono i parametri con valori $\beta = (\beta_0, \beta_1, \beta_2) = (0.2, 0.5, -0.3)$ e si stima il modello.

Le statistiche rilevate sono riportate nella tabella 5.3:

Tabella 5.3: Statistiche delle stime dei coefficienti di regressione

	Media	Bias	Standard Error	$IC_{95\%}$
$\hat{\beta}_0$	0.187	-0.013	0.080	(0.030, 0.344)
$\hat{\beta}_1$	0.440	-0.060	0.065	(0.313, 0.568)
$\hat{\beta}_2$	-0.266	0.034	0.074	(-0.410, -0.122)

Le stime dei coefficienti mostrano ancora bias contenuti e deviazioni standard compatibili con la variabilità attesa. Anche gli intervalli di confidenza includono i valori teorici dei parametri e questo comportamento sembra confermare l'accuratezza della procedura di stima. Nella figura 5.3 si mostrano gli andamenti delle stime dei tre coefficienti e degli intervalli di confidenza asintotici.

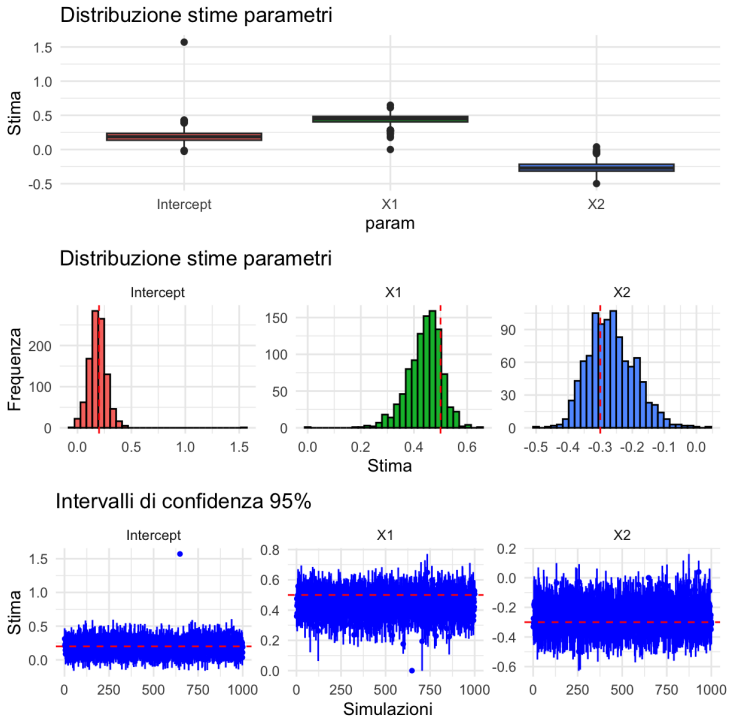

Figura 5.3: Stime dei coefficienti e copertura per il modello con covariate normali

5.3 Regressione Secante con *link* identità

L'impiego del *link* canonico impone forti limitazioni all'insieme dei valori ammissibili del predittore lineare (4.1.1), perciò i risultati stimati successivamente, su dati simulati più complessi, vedranno l'impiego del *link* alternativo scelto, il *link* identità.

5.3.1 Stima di un modello con covariate correlate

Per valutare le proprietà dello stimatore nel GLM con *link* identità, si generano 10^4 dataset simulati con covariata normale multivariata gaussiane con matrice di correlazione positiva e variabile risposta y dalla distribuzione GHS. Si scelgono valori reali dei coefficienti di regressione pari a $\beta = (-1.2,\ 0.7,\ 3.4,\ -0.2)$. Su ciascun dataset si stima il modello tramite la funzione `glm()` di `R` utilizzando la famiglia personalizzata `quasi.GHS` 4.3.1 in cui si fissa il parametro di dispersione pari a 1, ottenendo stime puntuali ed errori standard.

Dalle repliche Monte Carlo si calcolano bias, varianza, MSE e copertura degli intervalli di confidenza, ottenendo i risultati riportati nella tabella 5.4:

Tabella 5.4: Risultati Monte Carlo per il GLM con *link* identità

Parametro	Bias	Varianza	MSE	Copertura
Intercetta	0.132	0.043	0.060	0.897
β_1	-0.078	0.075	0.081	0.945
β_2	-0.375	0.187	0.328	0.852
β_3	0.022	0.074	0.075	0.955

Infine, i risultati sono stati sintetizzati graficamente in 5.4 attraverso boxplot delle stime, istogrammi della distribuzione delle stime con indicazione dei valori veri dei parametri, e copertura per ciascun parametro e per ciascuna simulazione.

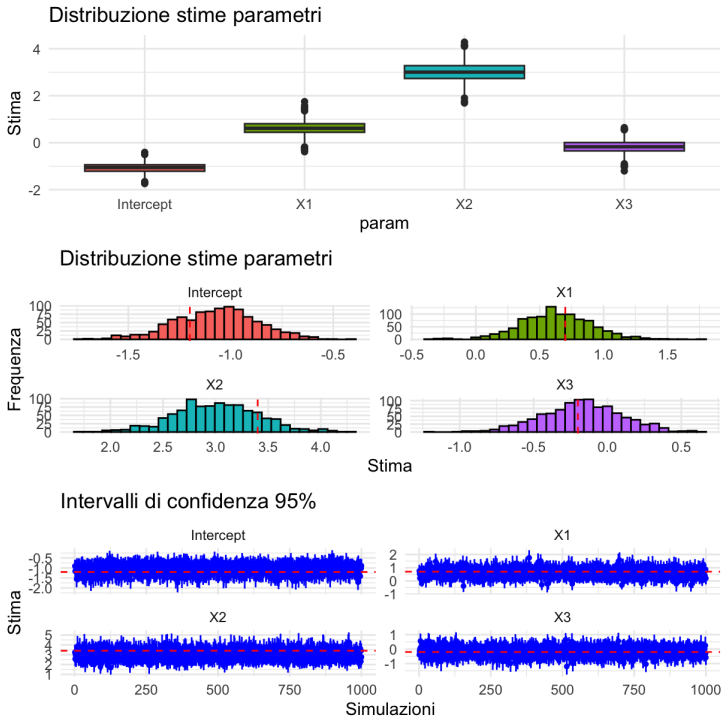

Figura 5.4: Stime dei coefficienti e copertura per il modello con covariate correlate

5.4 Applicazione del modello di quasi-verosimiglianza a dati reali

In questa sezione, il modello di quasi-verosimiglianza, tramite la funzione di `R` scritta in 4.3.1, viene utilizzato su un dataset reale, così da valutarne l'impiego in un contesto applicativo reale. Il dataset utilizzato è il dataset *"airquality"* (R: airquality) fornito da `R`, che registra misurazioni relative alla qualità dell'aria a New York tra il 1 maggio e il 30 settembre 1973. Il dataset è composto da 153 osservazioni e 6 variabili relative alla presenza di ozono, alle radiazioni solari, al vento e alla temperatura massima dell'aria per ogni giorno di rilevazione. Dopo aver trattato i dati, eliminando i valori mancanti, si ottiene un dataset composto da 111 osservazioni.

Si utilizza allora il dataset per la stima del livello di ozono nell'aria (*Ozone*), considerato quindi come variabile risposta del modello di quasi-verosimiglianza, sulla base delle variabili esplicative relative alla temperatura massima (*Temp*) e alla velocità del vento (*Wind*). Si assume che la relazione tra la media e la varianza della variabile target sia del tipo $V(\mu) = \phi(1+\mu^2)$, dove ϕ è il parametro di dispersione che verrà stimato dal modello come media corretta dei residui di Pearson al quadrato.

Allora, si implementa la funzione `glm` di `R` scegliendo `family=quasi.GHS()` (4.3.1) e si stima il modello lineare generalizzato di quasi-verosimiglianza:

```
1    glm <- glm(Ozone ~ Temp+Wind, family=quasi.GHS(), data=df_air)
2    summary(glm)
```

Si ottiene il seguente *output* (5.5):

```
Call:
glm(formula = Ozone ~ Temp + Wind, family = quasi.GHS(), data = df_air)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -52.7915    16.1302  -3.273  0.00143 **
Temp          1.3111     0.1824   7.189  8.9e-11 ***
Wind         -0.9755     0.3908  -2.496  0.01407 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasi.GHS family taken to be 0.3176476)

    Null deviance: 137.371  on 110  degrees of freedom
Residual deviance:  68.608  on 108  degrees of freedom
AIC: 143.22

Number of Fisher Scoring iterations: 25
```

Figura 5.5: *Output* R del modello di quasi-verosimiglianza

L'analisi dei coefficienti mostra che la variabile relativa alla temperatura ha un effetto positivo e fortemente significativo sul livello medio di ozono, mentre la velocità vento ha un effetto negativo ma comunque statisticamente significativo con un livello di significatività del 5% ($\alpha = 0.05$).

La stima del coefficiente di dispersione per il modello è pari a $\hat{\phi} = 0.318$, il dataset è quindi caratterizzato da sottodispersione, cioè la variabilità reale dei dati sull'ozono è minore di quella attesa dalla distribuzione teorica secante iperbolica ($\phi = 1$).

Il grafico 5.6 mostra la relazione tra i valori stimati della media del modello, $\hat{\mu}_i$, e la varianza empirica dei residui al quadrato, $(y_i - \hat{\mu}_i)^2$, calcolata in 20 intervalli dei valori predetti. I punti blu rappresentano la varianza empirica osservata per ciascun intervallo, mentre la curva verde indica la forma teorica della funzione di varianza della famiglia di distribuzioni secante iperbolica, cioè $V(\mu) = 1 + \mu^2$. La linea rossa tratteggiata invece corrisponde alla varianza predetta dal modello, scalata dal parametro di dispersione stimato $\hat{\phi}$. La distanza tra le due curve evidenzia la sottodispersione della variabilità del dataset rispetto alla variabilità prevista dal modello basato sulla distribuzione secante iperbolica (linea verde).

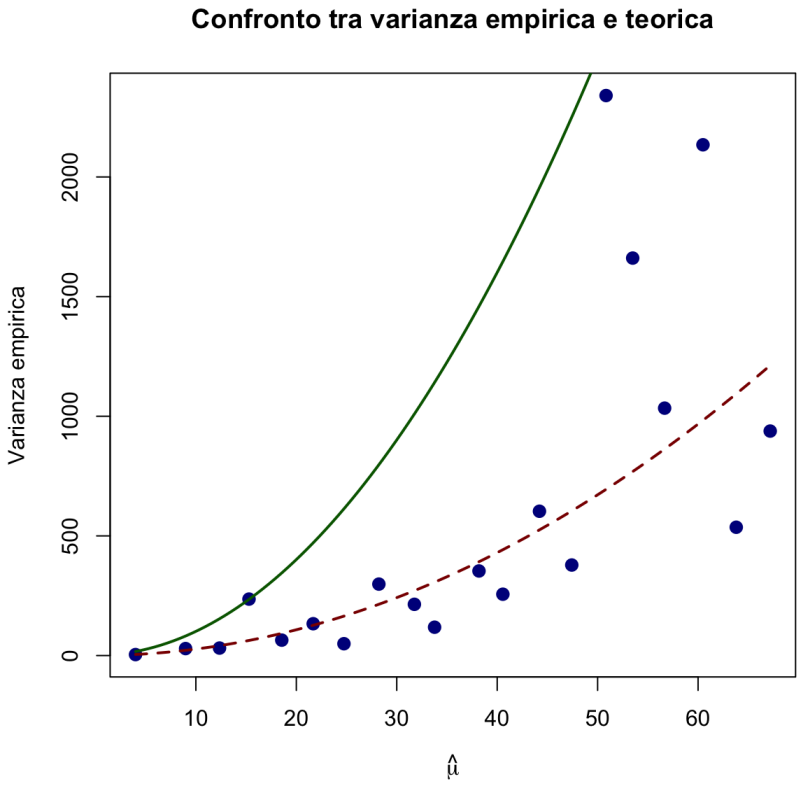

Figura 5.6: Confronto tra varianza empirica e teorica del modello di quasi-verosimiglianza

Infine, si confronta l'*output* ottenuto dal modello di quasi-verosimiglianza (5.5) con quello ottenuto stimando il medesimo modello ma forzando il parametro di dispersione $\phi = 1$ (5.7). Si stima cioè un modello lineare generalizzato assumendo che la variabile target abbia distribuzione secante iperbolica generalizzata, quindi con funzione di varianza pari esattamente a $V(\mu) = 1 + \mu^2$ e tramite

```
1    glm.ghs <- glm(Ozone~Temp+Wind, family=quasi.GHS(), data=df_ air)
2    summary(glm.ghs, dispersion=1)
```

si ottiene:

```
Call:
glm(formula = Ozone ~ Temp + Wind, family = quasi.GHS(), data = df_air)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -52.7915    28.6199  -1.845   0.0651 .
Temp          1.3111     0.3236   4.052 5.08e-05 ***
Wind         -0.9755     0.6934  -1.407   0.1595
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasi.GHS family taken to be 1)

    Null deviance: 137.371  on 110  degrees of freedom
Residual deviance:  68.608  on 108  degrees of freedom
AIC: 143.22

Number of Fisher Scoring iterations: 25
```

Figura 5.7: *Output* R del modello GHS con $\phi = 1$

I due *output*, 5.5 e 5.7, sono sovrapponibili nelle stime dei coefficienti di regressione e nei valori delle devianze ma differiscono nel parametro di dispersione e negli *standard error* associati alle stime.

Nella tabella 5.5 si visualizzano i valori degli errori standard associati all'intercetta e alle due variabili esplicative.

Regressore	**Std. Error** ($\hat{\phi} = 0.269$)	**Std. Error** ($\phi = 1$)
Intercept	16.130	28.620
Temp	0.182	0.324
Wind	0.391	0.693

Tabella 5.5: Confronto degli standard error dei coefficienti tra i due modelli

Il fattore di dispersione stimato nel modello di quasi-verosimiglianza modifica gli standard error che vengono moltiplicati per $\sqrt{\hat{\phi}}$, risultando così più piccoli e indicando la presenza di sottodispersione.

I valori di standard error più piccoli nel modello di quasi-verosimiglianza rendono gli

intervalli di confidenza associati ai coefficienti di regressione più stretti e con p-value più piccoli. Nel caso di dataset con sottodispersione infatti, utilizzare un modello che non ne tiene conto (come in 5.7) tende a sovrastimare gli errori standard dei parametri e a concludere che alcuni effetti non sono significativi quando in realtà lo sono.

5.4.1 Confronto con il modello gaussiano

Si stima un modello normale sui dati *"airquality"* e se ne confronta la performance con il modello di quasi-verosimiglianza stimato precedentemente (5.5). Il modello Gaussiano assume omoschedasticità dei residui, ipotizzando cioè che la varianza sia costante lungo l'intero supporto della variabile risposta, $Var(Y_i) = \sigma^2$, $\forall i$, indipendentemente dal valore del predittore lineare.

La stima del modello lineare normale è condotta mediante la funzione `glm`, così da consentire un confronto diretto con i risultati derivanti dal modello di quasi-verosimiglianza GHS. È opportuno osservare che, nel caso di un `glm` con famiglia gaussiana, il parametro di dispersione stimato coincide con la varianza dei residui, σ^2 e la funzione *link* di *default* di `R` è la funzione identità, $g(\mu) = \mu$.

In figura 5.8 è riportato l'*output* ottenuto dal codice utilizzato per la stima del modello gaussiano su `R`:

```
1    glm.gauss <- glm(Ozone~Temp+Wind, family=gaussian, data=df_ air)
2    summary(glm.gauss)
```

Si possono allora confrontare i risultati ottenuti in 5.8 con l'*output* del modello quasi-GHS (5.5) evidenziando così differenze rilevanti e sistematiche tra i due approcci, riconducibili a una violazione dei presupposti teorici del modello gaussiano.

```
Call:
glm(formula = Ozone ~ Temp + Wind, family = gaussian, data = df_air)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -67.3220    23.6210  -2.850  0.00524 **
Temp          1.8276     0.2506   7.294 5.29e-11 ***
Wind         -3.2948     0.6711  -4.909 3.26e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 472.12)

    Null deviance: 121802  on 110  degrees of freedom
Residual deviance:  50989  on 108  degrees of freedom
AIC: 1003.4

Number of Fisher Scoring iterations: 2
```

Figura 5.8: *Output* R del modello gaussiano

Il pattern eteroschedastico presente nei dati, per cui la variabilità dei livelli di ozono aumenta proporzionalmente al loro valore atteso, invalida l'ipotesi di omoschedasticità su cui si basa il modello gaussiano. Il modello gaussiano infatti attribuisce al vento un effetto molto marcato, ($\hat{\beta}_{Wind} = -3.29$), e altamente significativo, mentre il modello di quasi-verosimiglianza, che utilizza la struttura di varianza nella forma $Var(Y_i) = \phi(1+\mu_i^2)$, ridimensiona fortemente tale effetto ($\hat{\beta}_{Wind} = -0.98$), pur mantenendolo statisticamente significativo. Analogamente, l'effetto della temperatura risulta sovrastimato nel modello gaussiano rispetto al modello di quasi-verosimiglianza. Dal punto di vista teorico, gli stimatori dei coefficienti (*Temp*, *Wind*) nel modello normale, *OLS* (*Ordinary Least Squares*), rimangono non distorti, ma non sono più efficienti e, soprattutto, gli errori standard risultano calcolati in modo non corretto. Essi infatti derivano da una stima della varianza residua omoschedastica che non tiene conto del fatto che la dispersione degli errori cresce con μ_i. Ignorare la varianza residua non costante comporta, in questo caso, p-value poco attendibili e una tendenza a sovrastimare la significatività delle covariate, enfatizzando artificialmente il loro effetto. Inoltre, il parametro di dispersione stimato costante dal

modello gaussiano e pari a $\hat{\sigma^2} = 472.12$, risulta elevato e non rappresentativo della reale struttura di varianza, confermando la non sostenibilità empirica dell'assunzione di omoschedasticità.

La superiorità diagnostica del modello basato sulla funzione di quasi-verosimiglianza della distribuzione Secante Iperbolica riflette la sua capacità di modellare appropriatamente l'eteroschedasticità intrinseca dei dati relativi alla concentrazione media di ozono nell'aria, in cui la variabilità delle misurazioni tende ad aumentare con l'intensità del fenomeno osservato, rendendo il modello più appropriato sia dal punto di vista statistico che interpretativo per questo tipo di dati ambientali.

Capitolo 6

Discussione e conclusioni

L'elaborato ha avuto come obiettivi principali la definizione e la caratterizzazione della distribuzione secante iperbolica generalizzata, appartenente alle distribuzioni della Famiglia Esponenziale Naturale con funzione di varianza quadratica. A partire dalla distribuzione di interesse, si è sviluppato un modello lineare generalizzato e infine un modello di quasi-verosimiglianza, basato sulla relazione quadratica tra la media e la varianza della distribuzione.

6.1 Discussione dei risultati

Partendo dal lavoro di Carl N. Morris (Morris 1982) è stata presentata la Famiglia Esponenziale Naturale con particolare interesse per le distribuzioni caratterizzate da funzione di varianza quadratica (NEF-QVF): ne è stata descritta la struttura algebrica e analitica che ne facilita l'utilizzo nella teoria della stima e nei modelli lineari generalizzati.
Tra le distribuzioni che rientrano in questa classe, si è approfondita la distribuzione Secante Iperbolica. La distribuzione secante iperbolica è stata derivata a partire da una trasformazione logit π-scalata di una distribuzione Beta, ottenendo la funzione di densità

caratterizzata dal parametro naturale $\theta \in (-\pi/2, \pi/2)$.

Studiandone le proprietà fondamentali, si è dimostrato che la distribuzione ha media $\mu = \tan(\theta)$ e funzione di varianza pari a $V(\mu) = 1 + \mu^2$, evidenziando la relazione quadratica tra media e varianza.

Dalla funzione di densità, sono state ricavate la funzione di ripartizione e la funzione quantile come sua inversa; infine, a partire da quest'ultima, tramite inversione della funzione di ripartizione (ITM), si è derivata su `R` la funzione per la generazione di osservazioni pseudocasuali.

Analizzandone la forma per vari valori del parametro θ, gli indici di asimmetria e la curtosi della distribuzione, si è potuto concludere che la secante iperbolica ha forma campanulare e unimodale con un indice di curtosi elevato che ne caratterizza le code, più pesanti rispetto a quelle di una distribuzione normale. Inoltre, la distribuzione è asimmetrica: la grandezza del valore del parametro ne determina la magnitudine e la direzione dipende invece dal suo segno.

Il confronto della distribuzione secante iperbolica nel caso standard, cioè con $\theta = 0$, con la normale standard ha rivelato che, sebbene entrambe siano unimodali, simmetriche rispetto allo zero, con media nulla e varianza unitaria, la distribuzione secante iperbolica presenta code più pesanti e un picco centrale più pronunciato. Questo la rende più appropriata per modellare fenomeni caratterizzati da un eccesso di curtosi, in cui si osserva una maggiore concentrazione di valori attorno alla media e una più alta probabilità di valori estremi rispetto alla distribuzione normale.

In seguito, nel capitolo 4, l'elaborato si è concentrato sull'applicazione della distribuzione secante iperbolica generalizzata nei modelli lineari generalizzati, esplorando l'uso di diverse funzioni *link*, in particolare il *link* canonico e il *link* identità, e infine sul modello di quasi-verosimiglianza per l'analisi di dati reali.

La stima dei coefficienti è stata effettuata con due approcci computazionali su `R`: prima

tramite l'algoritmo di Newton-Raphson (e IRLS), poi tramite la definizione personalizzata di un oggetto di tipo `family` da utilizzare direttamente nella funzione `glm` di `R`.

Le prestazioni dei modelli lineari generalizzati stimati sono state valutate su dati simulati tramite iterazioni Monte Carlo: la procedura è iniziata con la stima del modello nullo con *link* canonico e con l'aggiunta poi di due covariate numeriche al modello ma rilevando forti limitazioni allo spazio dei valori ammissibili del predittore lineare dovute alla scelta del *link* canonico.

Per superare le restrizioni del *link* canonico, è stato adottato il *link* identità per stimare un modello più complesso con covariate simulate con matrice di correlazione positiva. Dall'analisi dei risultati ottenuti dalle stime Monte Carlo, i modelli hanno presentato statistiche descrittive soddisfacenti, includendo in tutti i casi il parametro reale e con tassi di copertura desiderabili.

Infine, per affrontare la modellazione di dati reali e rilassare le ipotesi sulla distribuzione completa, è stato introdotto il modello di quasi-verosimiglianza. Questo approccio richiede solo la specificazione della relazione tra media e varianza, consentendo di stimare i parametri del modello e un parametro di dispersione ϕ, che gestisce sovra o sotto-dispersione. I risultati ottenuti mostrano che la distribuzione secante iperbolica è una valida alternativa per la modellazione di dati continui che presentano asimmetria e leptocurtosi, caratteristiche che la distinguono dalla più comune distribuzione normale. Sebbene il *link* canonico offra vantaggi computazionali e teorici, le sue restrizioni sullo spazio parametrico lo rendono meno flessibile in scenari più complessi o con valori estremi del predittore lineare. L'adozione del *link* identità e, in particolare, l'utilizzo del modello di quasi-verosimiglianza, offrono maggiore flessibilità e robustezza nell'analisi di dati reali, permettendo di gestire la dispersione non specificata dalla distribuzione teorica.

6.2 Limiti e sviluppi futuri

Il presente lavoro ha mostrato come la regressione con distribuzione secante iperbolica generalizzata (GHS) possa rappresentare uno strumento efficace per stimare relazioni tra variabili, sia in contesti simulativi controllati sia in applicazioni su dati reali. Tuttavia, alcuni limiti emergono naturalmente dall'analisi condotta.

In particolare, l'uso di campioni di dimensioni relativamente ridotte e di un numero limitato di covariate potrebbe non cogliere pienamente la complessità dei fenomeni reali, mentre l'analisi dei dati reali ha evidenziato sottodispersione, suggerendo che la funzione di varianza teorica del modello di regressione secante iperbolica potrebbe non adattarsi perfettamente a tutte le situazioni, con il rischio di non stimare correttamente la significatività di alcune stime. Infine, l'applicazione empirica è stata condotta su un dataset reale circoscritto. Sebbene abbia fornito indicazioni interessanti, essa non consente di trarre conclusioni generali sull'efficacia del modello in contesti applicativi più ampi. Sarebbe quindi opportuno testare il modello su una varietà maggiore di dati, sia simulati che reali, appartenenti a diversi ambiti economici e statistici.

Alla luce di questi limiti, diversi possibili sviluppi si prospettano per ampliare e consolidare i risultati. Dal punto di vista metodologico, un'estensione naturale riguarda la costruzione di modelli multivariati basati sulla distribuzione secante iperbolica, che permettano di affrontare problemi di regressione con più variabili risposta o applicazioni a dati panel e serie temporali. Questo ampliamento consentirebbe di valutare la flessibilità della distribuzione anche in contesti dinamici e ad alta dimensionalità. Inoltre, si potrebbero ampliare i risultati relativamente ai modelli con famiglia secante iperbolica sviluppando GLM con *link* alternativi, non lineari, o introducendo fattori di penalizzazione.

Un'altra direzione interessante consiste nel confronto sistematico con altre distribuzioni capaci di gestire leptocurtosi e code pesanti, come la distribuzione *t di Student* o *Laplace*. Tale confronto permetterebbe di mettere meglio in evidenza i vantaggi e le peculiarità

della secante iperbolica, chiarendo in quali situazioni essa risulti preferibile.

Inoltre, le applicazioni pratiche meritano un ulteriore approfondimento, in particolare in ambito economico-finanziario. La capacità della HS di modellizzare fenomeni con presenza di valori estremi la rende infatti particolarmente adatta per analizzare rendimenti, rischi e variabili economiche caratterizzate da distribuzioni non gaussiane.

Infine, l'approccio bayesiano offre prospettive stimolanti: attraverso l'integrazione di informazioni a priori e la valutazione completa delle distribuzioni a posteriori dei parametri, è possibile ottenere stime più robuste e intervalli di incertezza più realistici, soprattutto in presenza di campioni piccoli o dati rumorosi. L'adozione di metodi bayesiani potrebbe quindi arricchire significativamente l'analisi, permettendo di combinare la potenza dei modelli basati sulla distribuzione secante iperbolica con un quadro inferenziale più flessibile e interpretabile.

In sintesi, i risultati raggiunti hanno confermato la validità della distribuzione secante iperbolica come strumento di modellizzazione e hanno posto le basi per ulteriori sviluppi. Le direzioni future individuate mostrano come questo lavoro possa costituire un punto di partenza per ricerche più ampie, capaci di integrare rigore teorico e applicazioni concrete in contesti complessi.

Elenco delle figure

Elenco delle tabelle