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

Maria Regina Mucilli
Relatore: Dott. Tommaso Rigon
Correlatore: Prof. Sonia Migliorati

Università degli Studi di Milano Bicocca
Corso di Laurea Magistrale in Scienze Statistiche ed Economiche

15 Ottobre 2025

Motivazione e Obiettivi

L'elaborato prende forma dal lavoro di **Morris 1982** sulle *sei distribuzioni* nella Famiglia Esponenziale Naturale con Funzione di Varianza Quadratica.

Obiettivi:

1. Caratterizzare la **distribuzione Secante Iperbolica (HS)**: forma, momenti, funzioni fondamentali e altre proprietà.
2. Sviluppare applicazioni in **Modelli Lineari Generalizzati**, basati sulla funzione di verosimiglianza della distribuzione e valutandone le performance tramite simulazioni.
3. Esplorare il **modello di Quasi-Verosimiglianza**, basato su ipotesi deboli e applicandolo a un dataset reale.
4. Confrontare il modello sviluppato con il modello gaussiano.

Contesto Teorico 1/2

- **Famiglia Esponenziale Naturale** (NEF):
  classe di distribuzioni di probabilità nella forma

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

  dove $\psi(\theta)$ è *funzione dei cumulanti*.

- **Proprietà delle NEF**:
  - Media: $E_\theta[X] = \mu = \psi'(\theta),\ \mu \in \Omega$
  - Funzione di varianza: $Var(\theta) = V(\mu) = \psi''(\theta)$

- **NEF con Funzione di varianza quadratica** (QVF):

$$V(\mu) = v_0 + v_1\mu + v_2\mu^2$$

Contesto Teorico 2/2

- **NEF-QVF note**

Distribuzione	Ω	$V(\mu)$
Normale	$\mathbb{R}$	σ^2
Poisson	$(0, \infty)$	μ
Binomiale (m)	$(0, m)$	$\mu\left(1 - \frac{\mu}{m}\right)$
Binomiale Negativa (r)	$(0, \infty)$	$\mu + \frac{\mu^2}{r}$
Gamma	$(0, \infty)$	$\frac{\mu^2}{\alpha}$

- **Distribuzione Secante Iperbolica**

Distribuzione	Ω	$V(\mu)$
Secante Iperbolica	$\mathbb{R}$	$1 + \mu^2$

Distribuzione Secante Iperbolica 1/3

Funzione di densità:

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

dove:

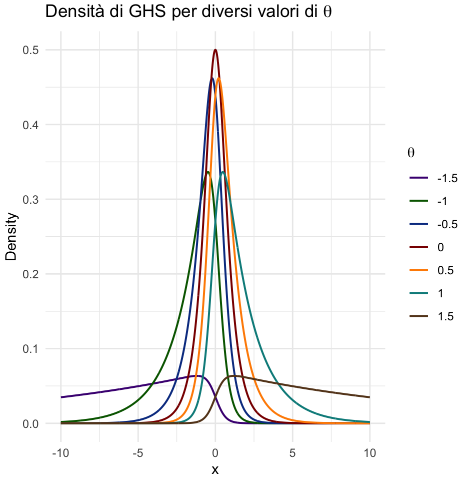
- $\psi(\theta) = -\log(\cos\theta)$ è la *log-partition function*;
- $\theta \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ è il parametro naturale.

Proprietà:

- $x = \frac{1}{\pi}\text{logit}(y)$ con $y \sim \text{Beta}\left(\frac{1}{2} + \frac{\theta}{\pi}, \frac{1}{2} - \frac{\theta}{\pi}\right)$

Distribuzione Secante Iperbolica 2/3

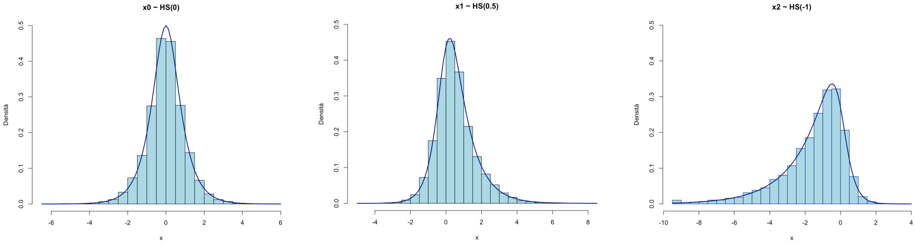
- Media:
  $\mu = \tan(\theta) \in \mathbb{R}$
- Varianza:
  $V(\mu) = 1 + \mu^2 = \sec^2(\theta)$
- Indice di Asimmetria:
  $\gamma_1 = \frac{2\mu}{\sqrt{V(\mu)}}$
- Indice di Curtosi:
  $\gamma_2 = \frac{5V(\mu)+4\mu^2}{V(\mu)}$



Distribuzione Secante Iperbolica 3/3

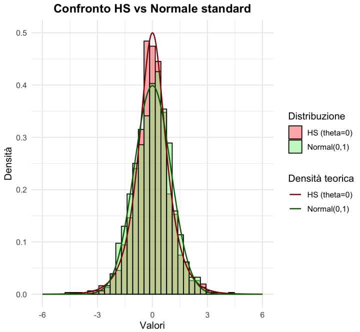
- **CDF** in forma Beta regolarizzata: $F(x) = \mathrm{I}_{\frac{1}{1+e^{-\pi x}}}\left(\frac{1}{2} + \frac{\theta}{\pi}, \frac{1}{2} - \frac{\theta}{\pi}\right)$

- **Generazione di numeri pseudocasuali** tramite algoritmo di *Inversione della funzione di ripartizione* (ITM):



Confronto tra distribuzioni simulate e teoriche con $\theta = (0,\ 0.5,\ 1)$.

Secante Iperbolica Standard vs Normale Standard



Distribuzione	Media	Varianza	Asimmetria	Curtosi
$HS(\theta = 0)$	0	1	0	5
$N(\mu = 0, \ \sigma = 1)$	0	1	0	3

Regressione Secante Iperbolica

1 Specificazione con link canonico

- Variabile dipendente: $Y_1, \ldots, Y_n \sim \text{HS}(\theta_i)$;
- Predittore lineare: $\eta_i = \mathbf{x}_i^T \beta$;
- *Link* canonico g: $g(\mu) = \theta = \arctan(\mu)$ tale che $\theta = \eta \in (-\frac{\pi}{2}, \frac{\pi}{2})$.

Limiti del link canonico

Vincolo sul predittore $\rightarrow$ link canonico incompatibile con il dominio delle medie Ω.
Si utilizza allora un *link* alternativo.

2 Specificazione con link identità

- Variabile dipendente: $Y_1, \ldots, Y_n \sim HS(\theta_i)$;
- Predittore lineare: $\eta_i = \mathbf{x}_i^T \beta$;
- *Link* identità g: $g(\mu) = \mu = \tan(\theta)$.

Modello di Quasi-Verosimiglianza

Il **modello di quasi-verosimiglianza** permette di rilassare le ipotesi del modello classico $\Rightarrow$ **Ipotesi di Secondo Ordine**

Specificazione del modello:

- $\mathbb{E}[Y_i] = \mu_i$
- $\mathrm{Var}(Y_i) = \phi(1 + \mu_i^2), \quad \phi > 0$
- Y_i e Y_j indipendenti per $i \neq j$

> - Il parametro di dispersione ϕ consente di modellare **sovradispersione** o **sottodispersione** nei dati.
> - La stima di ϕ è ottenuta come media corretta dei residui di Pearson al quadrato.

Implementazione in R

1. **Algoritmo IRLS** basato sul metodo Newton-Raphson per la stima dei coefficienti del modello;

2. Creazione di un **oggetto** *family* **personalizzato** da richiamare nella funzione `glm()` di R:

```
fit <- glm(y ~ X, family = quasi.HS(), data = data)
```

- Regressione secante iperbolica ($\phi = 1$ fissato):

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

- Modello di quasi-verosimiglianza (ϕ da stimare):

```
summary(fit)
```

Validazione dei modelli tramite simulazione

Regressione Secante Iperbolica con link canonico

- *Scenario 1*: Modello nullo con sola intercetta.
- *Scenario 2*: Modello con covariate generate da distribuzioni Normali standard.

Regressione Secante Iperbolica con link identità

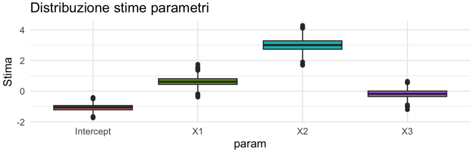
- *Scenario 3*: Modello con covariate correlate generate da una distribuzione gaussiana multivariata.

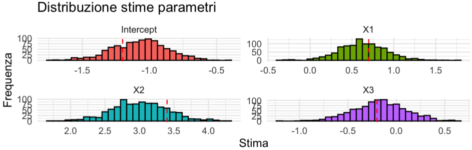
- Per ciascuno scenario, $R = 10^4$ repliche Monte Carlo.
- Metriche per la valutazione del modello:
  Bias, Errore Quadratico Medio e Copertura.

Regressione Secante Iperbolica con *link* identità

❸ *Scenario 3*: $(\beta_0, \beta_1, \beta_2, \beta_3) = (-1.2, 0.7, 3.4, -0.2)$

	Bias	MSE	Copertura
$\hat{\beta}_0$	0.13	0.06	90%
$\hat{\beta}_1$	-0.08	0.08	95%
$\hat{\beta}_2$	-0.38	0.33	85%
$\hat{\beta}_3$	0.02	0.08	95%





Applicazione a dati reali

- **Dati**: misurazioni giornaliere sulla qualità dell'aria a New York.
- **Obiettivo**: stimare i livelli di Ozono in funzione della temperatura massima e della velocità media del vento.

1. Stima del **modello di quasi-verosimiglianza** della distribuzione secante iperbolica.
2. Stima del **modello gaussiano**.
3. Confronto tra i modelli.

Modello Quasi-HS vs Modello Gaussiano

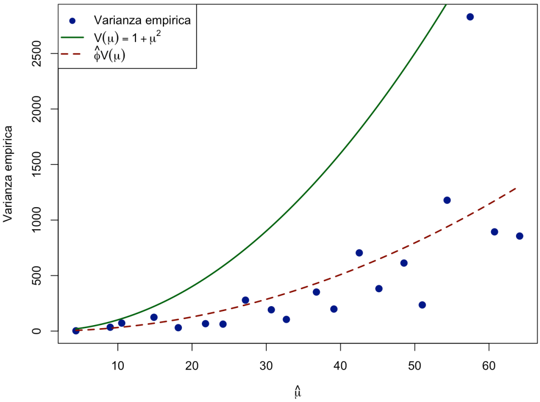
Modello quasi-HS			
Coefficiente	**Stima**	**SE**	**p-value**
Temperatura	1.31	0.18	$9 \cdot 10^{-11***}$
Vento	-0.98	0.39	0.01^{*}
Dispersione	$\hat{\phi} = 0.318$		

Modello gaussiano			
Coefficiente	**Stima**	**SE**	**p-value**
Temperatura	1.83	0.25	$6 \cdot 10^{-11***}$
Vento	-3.29	0.67	$3 \cdot 10^{-06***}$
Dispersione	$\hat{\sigma}^2 = 472.12$		

- Stima $\hat{\phi} < 1 \Rightarrow$ **sottodispersione**.
- Correzione degli SE tramite $\hat{\phi}$.

- Significatività dei coefficienti sovrastimata.
- $\hat{\sigma}^2$ elevato sotto **ipotesi di omoschedasticità**.

Analisi dei residui



Varianza empirica dei residui al quadrato del modello quasi-HS, nel confronto con la varianza teorica della HS e la stima scalata con $\hat{\phi}$.

Conclusioni e Sviluppi Futuri

Conclusioni e Contributi

- **Distribuzione Secante Iperbolica (HS)**: alternativa alla Normale per dati continui con **asimmetria** o **leptocurtosi**.
- Rispetto alla Normale, *HS* cattura meglio **eventi estremi** e nei GLM gestisce **eteroschedasticità**.
- L'approccio di **quasi-verosimiglianza** offre flessibilità, gestendo **dispersione** non specificata tramite la stima del parametro ϕ.

Sviluppi Futuri

- Estendere il modello con **interazioni**, **link non lineari** o **penalizzazioni** per dati ad alta dimensionalità.
- Confronti sistematici con altre distribuzioni (t-Student, Laplace).
- Applicazione di un **approccio Bayesiano**.

Grazie per l'attenzione!

Bibliografia I

airquality: New York Air Quality Measurements (n.d.). R Documentation, datasets package. R Foundation for Statistical Computing. URL: https://www.rdocumentation.org/packages/datasets/versions/3.6.2/topics/airquality.

Fischer, Matthias J (2013). “Hyperbolic Secant Distributions”. In: *Generalized Hyperbolic Secant Distributions: With Applications to Finance*. Springer, pp. 1–13.

Liu, Jun S and Jun S Liu (2001). *Monte Carlo strategies in scientific computing*. Vol. 10. Springer.

McCullagh, Peter (1983). “Quasi-likelihood functions”. In: *The Annals of Statistics*, pp. 59–67.

Morris, Carl N (1982). “Natural exponential families with quadratic variance functions”. In: *The Annals of Statistics*, pp. 65–80.

Bibliografia II

- Morris, Carl N and Kari F Lock (2009). "Unifying the named natural exponential families and their relatives". In: *The American Statistician* 63.3, pp. 247–253.
- R Core Team (n.d.[a]). *R: Family Objects for Models*. R Documentation, stats package. R Foundation for Statistical Computing. URL: https://search.r-project.org/R/refmans/stats/html/glm.html.
- — (n.d.[b]). *R: Fitting Generalized Linear Models*. R Documentation, stats package. R Foundation for Statistical Computing. URL: https://search.r-project.org/R/refmans/stats/html/glm.html.
- Salvan, Alessandra, Nicola Sartori, and Luigi Pace (2020). "Modelli lineari generalizzati". In: *Modelli Lineari Generalizzati*. Springer, pp. 67–119.
- Wedderburn, Robert WM (1974). "Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method". In: *Biometrika* 61.3, pp. 439–447.

Algoritmo ITM (`rghs`)

1. Specificare il parametro della distribuzione θ oppure $\alpha = \frac{1}{2} + \theta/\pi$.
2. Generare casualmente u da $U \sim \text{Uniform}(0, 1)$.
3. Applicare la trasformazione inversa della Beta: $B = qbeta(U; \alpha, 1 - \alpha)$.
4. Applicare la trasformazione π-*logit* per ottenere le variabili casuali dalla distribuzione Secante Iperbolica: $x = \frac{1}{\pi} \log \left(\frac{B}{1-B} \right)$.

Algoritmo IRLS

1. Calcolo di media $\mu = X\beta$ e varianza $V(\mu) = 1 + \mu^2$.
2. Aggiornamento dei pesi $W = diag(\frac{1}{1+\mu_i^2})$.
3. Aggiornamento della stima dei coefficienti β_{new} tale che: $(X^T WX)\beta_{new} = X^T Wy$.
4. Iterazione fino alla convergenza o al raggiungimento del numero massimo di iterazioni.

Risultati delle Simulazioni

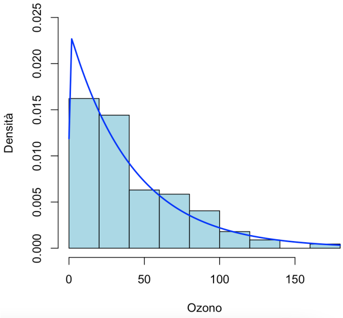
1. *Scenario 1*: $\beta_0 = 0.5$

	Bias	MSE	Copertura
$\hat{\beta}_0$	-0.005	0.008	95%

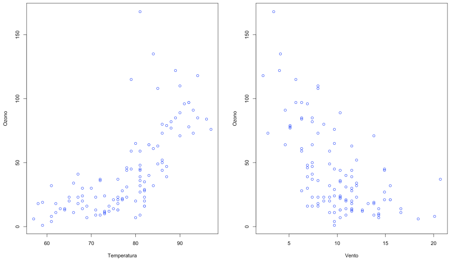
2. *Scenario 2*: $(\beta_0, \beta_1, \beta_2) = (0.2, 0.5, -0.3)$

	Bias	MSE	Copertura
$\hat{\beta}_0$	-0.01	0.007	92%
$\hat{\beta}_1$	-0.06	0.008	70%
$\hat{\beta}_2$	-0.03	0.006	85%

Distribuzione dei dati



Distribuzione della variabile Ozono



Distribuzione del livello di Ozono rispetto alle covariate Temperatura e Vento

Output R

Modello quasi-HS

```
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
```

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
```