Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in
Statistica per l'Economia e l'Impresa

Relazione Finale

Effetto di Hauck-Donner nei test basati sulla statistica di Wald

Relatore Prof. Alessandra Salvan
Dipartimento di Scienze Statistiche

Laureando Davide Bortoletto
Matricola 2045573

Anno Accademico 2023/2024

Indice

Introduzione

L'effetto di Hauck-Donner è un'anomalia di uno dei test più diffusi e utilizzati nelle applicazioni della Statistica, il test di Wald. Nonostante la prima esposizione del problema da parte di Hauck & Donner (1977), oramai quarant'anni fa, questo effetto non è stato oggetto di studio approfondito quanto altri difetti che caratterizzano il test Wald, pur essendo potenzialmente molto impattante, ad esempio portando a conclusioni inferenziali errate, come verrà mostrato in seguito.

La relazione introduce l'effetto di Hauck-Donner (*Hauck-Donner Effect*, HDE), espone alcuni metodi di diagnostica e di classificazione dell'HDE introdotti da Yee (2022), indaga sulle cause che portano al manifestarsi di questa anomalia e sulle conseguenze che può avere nell'ambito dei modelli lineari generalizzati.

Il Capitolo 1 richiama i concetti alla base della teoria di verosimiglianza, introduce i test statistici e la notazione che sarà adottata nel corso della relazione.

Il Capitolo 2 introduce la famiglia dei modelli lineari generalizzati (*Generalized Linear Models*, GLM), e offre una breve panoramica su una generalizzazione vettoriale dei GLM tratta da Yee (2015).

Il Capitolo 3 descrive l'effetto di Hauck-Donner, introducendo i metodi di rilevazione e classificazione di tale effetto e discutendo le possibili soluzioni per mitigarne l'impatto.

Il Capitolo 4 riporta uno studio di simulazione volto a indagare le cause e le conseguenze dell'effetto di Hauck-Donner, oltre a valutare i metodi proposti nella relazione.

Capitolo 1

Verosimiglianza

1.1 Introduzione

L'idealizzazione alla base dell'inferenza statistica consiste nel considerare i dati osservati y^{oss} una realizzazione di un vettore casuale Y, in breve si assume y^{oss} realizzazione di $Y \sim P^0$, con spazio campionario $\mathcal{Y}$. P^0 è una legge di probabilità almeno in parte ignota e che occorre ricostruire utilizzando l'informazione portata dai dati. Nel seguito si considererà sempre lo spazio campionario come un sottoinsieme di uno spazio euclideo, $\mathcal{Y} \subseteq \mathbb{R}^m$.

Inizialmente occorre delimitare le forme ritenute possibili per P^0, specificando una famiglia $\mathcal{F}$ di distribuzioni di probabilità almeno qualitativamente compatibili con la generazione di y^{oss}. La famiglia $\mathcal{F}$ è detta modello statistico, se $P^0 \in \mathcal{F}$, il modello è correttamente specificato.

Gli elementi di $\mathcal{F}$ sono spesso individuati tramite un parametro θ, che rappresenta una qualche quantità che caratterizza l'esperimento casuale che produce i dati e la popolazione da cui sono tratti come campione casuale.

Un modello statistico è una famiglia $\mathcal{F} = \{P_\theta \ : \ \theta \in \Theta\}$ dove Θ è detto spazio parametrico. Il modello può essere parametrico, non parametrico o semi-parametrico. Si ha un modello statistico parametrico se Θ è un sottoinsieme di uno spazio euclideo $\mathbb{R}^p$. In questa relazione si tratterà solamente di quest'ultimo tipo di modello. Si assumerà inoltre che il modello sia specificato in termini di funzione di densità $F = \{p_Y(y,\theta), y \in \mathcal{Y}, \theta \in \Theta \subseteq \mathbb{R}^p\}$ con $p_Y(y,\theta)$ densità di P_θ rispetto alla misura di Lebesgue nel caso continuo o rispetto alla misura contatore nel caso discreto. La funzione $p_Y(y,\theta)$ è anche detta funzione del modello.

1.2 Funzione di verosimiglianza

Sia $\mathcal{F}$ un modello statistico parametrico per i dati y con funzione del modello $p_Y(y, \theta)$, con $\theta = (\theta_1, \ldots, \theta_p) \in \Theta \subseteq \mathbb{R}^p$. Si consideri $p_Y(y, \theta)$ come funzione solamente di θ, con y fissato al valore osservato. La funzione $L : \Theta \to \mathbb{R}^+$ definita da

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

è detta funzione di verosimiglianza di θ basata sui dati y.

La funzione di verosimiglianza va interpretata come segue: alla luce dei dati osservati, $\theta^1 \in \Theta$ è più credibile di $\theta^2 \in \Theta$ come indice del modello probabilistico generatore dei dati se $L(\theta^1) > L(\theta^2)$. Due funzioni di verosimiglianza che differiscono per una costante moltiplicativa $c(y)$ si dicono equivalenti.

Spesso il modello statistico per i dati $y = (y_1, \ldots, y_n)$ assume che $Y = (Y_1, \ldots, Y_n)$ siano n variabili casuali indipendenti. in questo caso allora si può scrivere $L(\theta)$ come

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

Si parla di campionamento casuale semplice con numerosità n se le Y_i sono anche identicamente distribuite. Le procedure di inferenza basate su $L(\theta)$ sono espresse tramite la funzione di log-verosimiglianza per semplicità matematica,

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

dove, se $L(\theta) = 0$, si definisce $l(\theta) = -\infty$. Se le Y_i sono indipendenti $l(\cdot)$ assume la forma

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

Un valore $\hat{\theta} \in \Theta$ tale che $L(\hat{\theta}) \geq L(\theta)\ \forall\ \theta \in \Theta$ è detto stima di massima verosimiglianza di θ. $\hat{\theta}$ può essere trovato massimizzando la funzione di log-verosimiglianza, ma non è detto che esista o che sia unico. Se $\hat{\theta} = \hat{\theta}(y)$ esiste unico con probabilità uno, la variabile casuale $\hat{\theta} = \hat{\theta}(Y)$ è detta stimatore di massima verosimiglianza.

1.3 Quantità di verosimiglianza

Un modello si definisce a verosimiglianza regolare se:

- Θ è un sottoinsieme aperto dello spazio euclideo $\mathbb{R}^p,\ p \in \mathbb{N}^+$;
- $l(\theta)$ è una funzione differenziabile almeno 3 volte, con derivate parziali continue in Θ.

Nei modelli con verosimiglianza regolare, le principali informazioni sull'andamento della funzione di verosimiglianza sono contenute in $\hat{\theta}$ e nelle derivate parziali di $l(\theta)$ rispetto alle componenti di θ. Il vettore delle derivate parziali prime della funzione di log-verosimiglianza,

$$l_*(\theta) = \left(\frac{\partial l(\theta)}{\partial \theta_1}, \ldots, \frac{\partial l(\theta)}{\partial \theta_p} \right) = \left[\frac{\partial l(\theta)}{\partial \theta_r} \right] = [l_r(\theta)],$$

è detto funzione di punteggio o funzione score.

Se il modello ha verosimiglianza regolare, spesso la stima di massima verosimiglianza si individua come unica soluzione dell'equazione di verosimiglianza

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

Si tratta di un sistema di equazioni se $p > 1$, in generale $\hat{\theta}$ va determinato numericamente.

La matrice $p \times p$ delle derivate parziali seconde di $l(\theta)$ cambiate di segno,

$$j(\theta) = -l_{**}(\theta) = \left[-\frac{\partial^2 l(\theta)}{\partial \theta_r \partial \theta_s} \right],$$

è detta matrice di informazione osservata, ed è una misura dell'informazione che i dati forniscono sull'incognito parametro θ, in quanto più è grande $j(\hat{\theta})$, tanto più la verosimiglianza è concentrata attorno a $\hat{\theta}$.

Si dice informazione attesa o informazione di Fisher la quantità

$$i(\theta) = \mathbb{E}_\theta(j(\theta)) = \left[-\mathbb{E}_\theta \left(\frac{\partial^2 l(\theta)}{\partial \theta_r \partial \theta_s} \right) \right],$$

valore atteso dell'informazione osservata. Sotto condizioni di regolarità, valgono due proprietà note come prima e seconda identità di Bartlett:

- $\mathbb{E}_\theta(l_*(\theta, Y)) = 0,\ \theta \in \Theta$,
- $Var_\theta(l_*(\theta, Y)) = \mathbb{E}_\theta(-l_{**}(\theta, Y)) = i(\theta),\ \theta \in \Theta$.

1.4 Risultati asintotici e test

Sotto tenui condizioni di regolarità, lo stimatore di massima verosimiglianza $\hat{\theta}$ è consistente (cfr. Pace & Salvan (2001)), ossia sotto θ, vero valore del parametro,

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

dove $\xrightarrow{p}$ indica la convergenza in probabilità. Valgono inoltre i risultati di approssimazione in distribuzione per n grande, sotto θ, vero valore del parametro,

$$l_*(\theta) \dot{\sim} N_p(0, i(\theta)), \tag{1.1}$$

$$\hat{\theta} - \theta \dot{\sim} N_p(0, i(\theta)^{-1}). \tag{1.2}$$

Nella 1.2 è possibile sostituire $i(\theta)^{-1}$ con una sua stima consistente $j(\hat{\theta})^{-1}$. $N_p(\mu, \Sigma)$ indica la distribuzione normale p-variata con vettore delle medie μ e matrice di covarianza Σ. Inoltre

$$W_e(\theta) = (\hat{\theta} - \theta)^\top j(\hat{\theta})(\hat{\theta} - \theta) \dot{\sim} \chi^2_p, \tag{1.3}$$

$$W_u(\theta) = l_*(\theta)^\top i(\theta)^{-1} l_*(\theta) \dot{\sim} \chi^2_p, \tag{1.4}$$

$$W(\theta) = 2\{l(\hat{\theta}) - l(\theta)\} \dot{\sim} \chi^2_p, \tag{1.5}$$

dove χ^2_p indica la distribuzione chi-quadrato con p gradi di libertà. Le tre quantità basate sulla verosimiglianza $W_e(\theta), W_u(\theta), W(\theta)$ sono denominate rispettivamente di Wald, score (o di Rao), e del rapporto di verosimiglianza. Sono quantità approssimativamente pivotali per l'inferenza su θ, e possono essere impiegate per costruire test e regioni di confidenza per il parametro. Sono asintoticamente equivalenti, differendo per termini trascurabili al divergere di n sotto θ. Se θ è scalare ($p = 1$) si possono definire le versioni unilaterali di $W_e(\theta), W_u(\theta), W(\theta)$, sotto θ,

$$r_e(\theta) = (\hat{\theta} - \theta)\sqrt{j(\hat{\theta})} \dot{\sim} N(0, 1), \tag{1.6}$$

$$r_u(\theta) = l_*(\theta)/\sqrt{i(\theta)} \dot{\sim} N(0, 1), \tag{1.7}$$

$$r(\theta) = \mathrm{sgn}(\hat{\theta} - \theta)\sqrt{2\{l(\hat{\theta}) - l(\theta)\}} \dot{\sim} N(0, 1). \tag{1.8}$$

1.5 Verosimiglianza profilo

Si può essere interessati a fare inferenza, costruire intervalli e regioni di confidenza, o alla verifica d'ipotesi su un sottoinsieme del parametro θ, detto parametro d'interesse. Si suddivide θ in (ψ, λ) dove ψ è il blocco di componenti d'interesse e λ parametro di disturbo. Le quantità $\hat{\theta}, l_*(\theta), i(\theta), j(\theta)$ sono suddivise a blocchi di componenti corrispondenti: $\hat{\theta} = (\hat{\psi}, \hat{\lambda}), l_*(\theta)^\top = (l_\psi(\theta)^\top, l_\lambda(\theta)^\top)$,

$$i(\theta) = \begin{bmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{bmatrix}, \quad j(\theta) = \begin{bmatrix} j_{\psi\psi} & j_{\psi\lambda} \\ j_{\lambda\psi} & j_{\lambda\lambda} \end{bmatrix}.$$

Analogamente, sono suddivise in blocchi le matrici inverse $i(\theta)^{-1}$ e $j(\theta)^{-1}$

$$i(\theta) = \begin{bmatrix} i^{\psi\psi} & i^{\psi\lambda} \\ i^{\lambda\psi} & i^{\lambda\lambda} \end{bmatrix}, \quad j(\theta) = \begin{bmatrix} j^{\psi\psi} & j^{\psi\lambda} \\ j^{\lambda\psi} & j^{\lambda\lambda} \end{bmatrix},$$

con le relazioni, valide per le inverse di matrici a blocchi,

$$\begin{aligned} i^{\psi\psi} &= (i_{\psi\psi} - i_{\psi\lambda} i_{\lambda\lambda}^{-1} i_{\lambda\psi})^{-1}, \\ i^{\psi\lambda} &= -i^{\psi\psi} i_{\psi\lambda} i_{\lambda\lambda}^{-1}, \\ i^{\lambda\psi} &= -i^{\lambda\lambda} i_{\lambda\psi} i_{\psi\psi}^{-1}, \\ i^{\lambda\lambda} &= (i_{\lambda\lambda} - i_{\lambda\psi} i_{\psi\psi}^{-1} i_{\psi\lambda})^{-1}. \end{aligned}$$

Formule analoghe valgono per i blocchi di $j(\theta)^{-1}$.

Per l'inferenza su ψ valgono i risultati analoghi alle (1.1) – (1.8), utili per l'inferenza globale su θ. Si indichi con $\hat{\theta}_\psi$ la stima di massima verosimiglianza di θ nel sottomodello con ψ fissato. Si ha $\hat{\theta}_\psi = (\psi, \hat{\lambda}_\psi)$, con $\hat{\lambda}_\psi$ stima di massima verosimiglianza di λ per un fissato ψ, soluzione rispetto a λ dell'equazione di verosimiglianza parziale $l_\lambda(\psi, \lambda) = 0$. Valgono allora le approssimazioni

$$\begin{aligned} \hat{\psi} - \psi &\dot{\sim} N_{p_\psi}(0, i^{\psi\psi}(\hat{\theta})), \\ \hat{\psi} - \psi &\dot{\sim} N_{p_\psi}(0, j^{\psi\psi}(\hat{\theta})), \\ l_\psi(\hat{\theta}_\psi) &\dot{\sim} N_{p_\psi}(0, i^{\psi\psi}(\hat{\theta}_\psi)^{-1}), \\ W_{eP}(\psi) &= (\hat{\psi} - \psi)^\top (i^{\psi\psi}(\hat{\theta}))^{-1} (\hat{\psi} - \psi) \dot{\sim} \chi^2_{p_\psi}, \qquad (1.9) \\ W_{uP}(\psi) &= l_\psi(\hat{\theta}_\psi)^\top i^{\psi\psi}(\hat{\theta}_\psi) l_\psi(\hat{\theta}_\psi) \dot{\sim} \chi^2_{p_\psi}, \qquad (1.10) \\ W_P(\psi) &= 2\{l(\hat{\theta}) - l(\hat{\theta}_\psi)\} \dot{\sim} \chi^2_{p_\psi}, \qquad (1.11) \end{aligned}$$

dove p_ψ è la dimensione del parametro d'interesse ψ. In W_{eP} e W_{uP} le approssimazioni valgono sia con le matrici $j^{\psi\psi}(\cdot)$ e $i^{\psi\psi}(\cdot)$ calcolate in $\hat{\theta}$ sia con le stesse calcolate in $\hat{\theta}_\psi$.

Se ψ è scalare, si possono definire le versioni unilaterali di W_{eP}, W_{uP} e W_P, sotto θ,

$$r_{eP}(\psi) = (\hat{\psi} - \psi)/\sqrt{i^{\psi\psi}(\hat{\theta})}, \tag{1.12}$$

$$r_{uP}(\psi) = l_\psi(\hat{\theta}_\psi)\sqrt{i^{\psi\psi}(\hat{\theta}_\psi)}, \tag{1.13}$$

$$r_P(\psi) = \mathrm{sgn}(\hat{\psi} - \psi)\sqrt{2\{l(\hat{\theta}) - l(\hat{\theta}_\psi)\}}. \tag{1.14}$$

La (1.12) è di centrale interesse per la relazione, in quanto è la quantità soggetta all'effetto di Hauck-Donner. Essa è utilizzata comunemente da `R` (`z value`) come statistica test per la verifica della nullità dei singoli parametri di un modello di regressione, ed è frequentemente adottata per la facilità di calcolo e di interpretazione. Di norma viene riportato inoltre il livello di significatività osservato approssimato (`Pr(>|z|)`), calcolato come

$$\alpha^{oss} = 2\left\{1 - \Phi\left(\ |\hat{\psi}|/\sqrt{i^{\psi\psi}(\hat{\theta})}\right)\right\}.$$

Due varianti della (1.12), che saranno utilizzate nel seguito sono definite come

$$r^*_{eP}(\psi) = (\hat{\psi} - \psi)/\sqrt{i^{\psi\psi}(\psi, \hat{\lambda}_\psi)}, \tag{1.15}$$

$$r^\dagger_{eP}(\psi) = (\hat{\psi} - \psi)/\sqrt{i^{\psi\psi}(\psi, \hat{\lambda})}. \tag{1.16}$$

Le quantità di cui alla (1.15) e (1.16) differiscono dalla versione originale (1.12) nel calcolo dell'informazione attesa. La matrice di informazione attesa in $r^*_{eP}(\psi)$ viene calcolata stimando la componente di disturbo λ attraverso una stima vincolata a ψ, ovvero $\hat{\lambda}_\psi$, mentre in $r^\dagger_{eP}(\psi)$ con la stima di massima verosimiglianza $\hat{\lambda}$.

Capitolo 2

Modelli lineari generalizzati

Il presente capitolo riassume definizioni e risultati di base relativi a modelli lineari generalizzati (GLM*: Generalized Linear Models*), un'estensione del classico modello di regressione lineare normale e i VGLM (*Vector Generalized Linear Models*), un'estensione multivariata dei GLM. Il materiale presentato è tratto da tratto da Salvan et al. (2020) e da Yee (2015).

2.1 Introduzione ai GLM

2.1.1 Famiglie di dispersione esponenziale

Siano $y_1 \dots y_n$ realizzazioni di variabili casuali $Y_1 \dots Y_n$, la densità Y_i appartiene alla famiglia di dispersione esponenziale univariata se si può esprimere nella forma

$$p(y_i; \theta_i; \phi) = \exp\left\{ \frac{\theta_i y_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\}, \tag{2.1}$$

con $y_i \in \mathcal{Y} \subseteq \mathbb{R}$, $\theta_i \in \Theta \subseteq \mathbb{R}$, $a_i(\phi) > 0$, ϕ è detto parametro di dispersione e in genere $a_i(\phi) = 1$, $a_i(\phi) = \phi$, oppure $a_i(\phi) = \phi/\omega_i$ con $\phi > 0$ e $\omega_i,\ i = 1 \dots n$ pesi noti nel caso di dati raggruppati.

Specificando le funzioni $a_i(\cdot)$, $b(\cdot)$, $c(\cdot)$ nella (2.1) è possibile ricondursi a distribuzioni note come la normale, Poisson, gamma, binomiale. È facile ottenere le espressioni di media e varianza di una generica Y_i con densità (2.1). Sia $l(\theta_i, \phi) = l(\theta_i, \phi; y_i)$ la funzione di log-verosimiglianza basata su y_i. Si ottiene

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

$$\frac{\partial l(\theta_i, \phi)}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a_i(\phi)},$$

$$\frac{\partial^2 l(\theta_i, \phi)}{\partial^2 \theta_i} = \frac{-b''(\theta_i)}{a_i(\phi)},$$

rispettivamente funzione di log-verosimiglianza, funzione score per θ_i e derivata parziale seconda.

Si possono ottenere delle espressioni generali per i primi due momenti di Y_i. Utilizzando le due identità di Bartlett,

$$\mathbb{E}_{\theta_i,\phi}\left(\frac{\partial l(\theta_i, \phi; Y_i)}{\partial \theta_i}\right) = 0$$

e

$$\mathbb{E}_{\theta_i,\phi}\left\{\left(\frac{\partial l(\theta_i, \phi; Y_i)}{\partial \theta_i}\right)^2\right\} = -\mathbb{E}_{\theta_i,\phi}\left(\frac{\partial^2 l(\theta_i, \phi; Y_i)}{\partial \theta_i^2}\right).$$

Si ottiene

$$\mathbb{E}_{\theta_i,\phi}\left(\frac{Y_i - b'(\theta_i)}{a_i(\phi)}\right) = 0, \quad \text{quindi} \quad \mathbb{E}(Y_i) = \mathbb{E}_{\theta_i,\phi}(Y_i) = b'(\theta_i), \tag{2.2}$$

indipendente da ϕ.

Per quanto riguarda la varianza, sapendo che,

$$\mathbb{E}_{\theta_i,\phi}\left\{\left(\frac{Y_i - b'(\theta_i)}{a_i(\phi)}\right)^2\right\} = \frac{Var_{\theta_i,\phi}(Y_i)}{[a_i(\phi)]^2} = \frac{b''(\theta_i)}{a_i(\phi)},$$

si ottiene

$$Var(Y_i) = Var_{\theta_i,\phi}(Y_i) = a_i(\phi)b''(\theta_i). \tag{2.3}$$

Dato che $Var_{\theta_i,\phi}(Y_i) > 0$ per ogni θ_i, la funzione $b(\theta_i)$ è convessa e $b'(\theta_i)$ è crescente in θ_i, perciò $E_{\theta_i,\phi}(Y_i)$ è funzione biiettiva di θ_i. Inoltre $b(\theta_i)$ determina tutti i momenti di Y_i ed è detta "generatore dei cumulanti" (Salvan et al., 2020, paragrafo 2.1.3).

Prima di introdurre le ipotesi del modello lineare generalizzato, si ricorre ad una parametrizzazione che introduce funzioni media e varianza. Posto

$$\mu_i = \mu(\theta_i) = \mathbb{E}_{\theta_i,\phi}(Y_i), \tag{2.4}$$

dalla (2.2) si ha

$$\mu_i = \mu(\theta_i) = b'(\theta_i). \tag{2.5}$$

Dalla (2.3) si ottiene

$$Var_{\theta_i,\phi}(Y_i) = a_i(\phi)\frac{d}{d\theta_i}\mu(\theta_i) = a_i(\phi)\mu'(\theta_i). \tag{2.6}$$

La funzione $\mu(\cdot)$ è monotona crescente con dominio Θ e codominio lo spazio delle medie $\mathcal{M} = \mu(\text{int}\Theta)$, con $\text{int}(\Theta)$ l'insieme dei punti interni di Θ. Dato che la media è compresa tra minimo e massimo valore di Y_i, lo spazio delle medie coincide con l'insieme dei punti interni dell'insieme generato dalle combinazioni lineari convesse dei punti di $\mathcal{Y}$, supporto di Y_i, che quindi non dipende da i. La (2.4) definisce una riparametrizzazione (μ_i, ϕ) di una famiglia di dispersione esponenziale. Sia $\theta(\mu_i)$ l'inversa di $\mu(\theta_i)$. La (2.6) nella parametrizzazione (μ_i, ϕ) è

$$Var_{\mu_i,\phi}(Y_i) = a_i(\phi)b''(\theta_i)\Big|_{\theta_i=\theta(\mu_i)} = a_i(\phi)v(\mu_i),$$

dove la funzione definita su $\mathcal{M}$

$$v(\mu_i) = b''(\theta_i)\Big|_{\theta_i=\theta(\mu_i)} \tag{2.7}$$

è detta funzione di varianza, essa insieme al suo dominio caratterizza uno specifico modello nella classe delle famiglie di dispersione esponenziale. Si introduce la notazione

$$Y_i \sim DE_1(\mu_i, a_i(\phi)v(\mu_i)), \quad \mu_i \in \mathcal{M} \tag{2.8}$$

per indicare la distribuzione di Y_i.

2.1.2 GLM

Un modello lineare generalizzato è specificato dalle seguenti ipotesi:

$$Y_1, \ldots, Y_n \quad \text{v.c. univariate indipendenti,} \tag{2.9}$$

$$g(\mathbb{E}(Y_i)) = g(\mu_i) = \eta_i = \mathbf{x}_i\beta, \tag{2.10}$$

$$Y_i \sim DE_1(\mu_i, a_i(\phi)v(\mu_i)), \quad \mu_i \in \mathcal{M}, \tag{2.11}$$

con $g(\cdot)$ funzione liscia invertibile e coerente (ossia $g : \mathcal{M} \to \mathbb{R}$) nota, detta funzione di legame (*link function*). Fra tutte le possibili funzioni di legame si dice funzione di legame canonica

$$g(\mu_i) = \theta(\mu_i) \tag{2.12}$$

quella per cui il parametro naturale θ_i della DE_1 risulta combinazione lineare delle variabili esplicative con coefficienti β, $\theta_i = \mathbf{x}_i\beta,\ i = 1, \dots, n$.

Siano $Y_1, \dots, Y_n$ variabili casuali distribuite secondo le assunzioni (2.9), (2.10), (2.11). Allora la densità congiunta di $(Y_1, \dots, Y_n)^\top$ è data dal prodotto delle marginali (2.1) e la funzione di log-verosimiglianza risulta

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

con $\theta_i = \theta(\mu_i) = \theta(g^{-1}(\mathbf{x}_i\beta))$. Se $g(\cdot)$ è la funzione di legame canonica, la (2.13) può essere semplificata come

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

Le equazioni di verosimiglianza per $\beta = (\beta_1, \dots, \beta_p)^\top$, assumendo ϕ noto sono

$$l_r = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{Var(Y_i)} \frac{\partial \mu_i}{\partial \beta_r} = 0, \quad r = 1, \dots, p. \tag{2.14}$$

Nel caso di funzione di legame canonica, le equazioni si semplificano,

$$\sum_{i=1}^{n} \frac{1}{a_i(\phi)} y_i x_{ir} = \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \mu_i x_{ir}, \quad r = 1, \dots, p. \tag{2.15}$$

Le (2.14) possono essere espresse nella notazione matriciale

$$D^\top V^{-1}(y - \mu) = 0, \tag{2.16}$$

dove $y - \mu = (y_1 - \mu_1, \dots, y_n - \mu_n)^\top$, $V = \text{diag}[Var(Y_i)], i = 1, \dots, n.$ e D è una matrice $n \times p$ con generico elemento

$$d_{ir} = \frac{\partial \mu_i}{\partial \beta_r} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_r} = \frac{1}{g'(\mu_i)} x_{ir}, \quad i = 1, \dots, n, \quad r = 1, \dots, p.$$

Le (2.16) vanno risolte con metodi iterativi, cfr. paragrafo 2.3.

Si può dimostrare che β e ϕ sono parametri ortogonali (Salvan et al. (2020), paragrafo 2.3.3), ovvero $i_{\beta\phi} = 0$. La conseguenza principale dell'ortogonalità è che gli stimatori di massima verosimiglianza per β e ϕ sono asintoticamente indipendenti, e quindi per fare inferenza su β è sufficiente disporre del blocco dell'informazione osservata o attesa relativa a β. Se il legame è canonico, $i_{\beta\beta} = j_{\beta\beta}$ e spesso viene riportata nella forma

matriciale

$$i_{\beta\beta} = X^\top W X, \tag{2.17}$$

dove $W = \text{diag}(\omega_i)$, con $\omega_i = \dfrac{1}{(g'(\mu_i))^2 Var(Y_i)}$, se il legame è canonico allora $\omega_i = \dfrac{v(\mu_i)}{a_i(\phi)}$, $i = 1, \ldots, n$. Per n grande, grazie al risultato generale di normalità asintotica dello stimatore di massima verosimiglianza vale l'approssimazione

$$\hat{\beta} \dot{\sim} N_p(\beta, (X^\top W X)^{-1}). \tag{2.18}$$

Una stima consistente della matrice di covarianza di β è pertanto $(X^\top \hat{W} X)^{-1})$, con $\hat{W}$ la matrice W calcolata per $\beta = \hat{\beta}$, e se ϕ è ignoto per ϕ pari a una sua stima consistente $\tilde{\phi}$ (Salvan et al. (2020), paragrafo 2.3.7).

2.2 Modelli per dati binari

Nella classe dei modelli lineari generalizzati, quando la risposta è dicotomica si adottano i modelli per dati binari. I dati possono essere presentati in due forme: dati raggruppati quando il vettore delle risposte rappresenta gli esiti individuali e dunque i valori della risposta sono 0 o 1; dati non raggruppati quando si hanno più risposte dicotomiche per ciascuna combinazione delle variabili concomitanti e vengono riportati il numero totale di successi per quella combinazione e il numero totale di osservazioni o di insuccessi.

Per dati non raggruppati, il modello statistico per l'i-esima osservazione è la binomiale elementare, $Bi(1, \pi_i)$, assumendo indipendenza tra le risposte per $i = 1, \ldots, n$. Nel caso di dati raggruppati si assume, per il totale di successi s_i nelle m_i osservazioni corrispondenti ad una combinazione di modalità delle variabili concomitanti, una distribuzione $Bi(m_i, \pi_i)$, assumendo le osservazioni entro ciascun gruppo indipendenti ed identicamente distribuite. Sia la risposta la proporzione di successi $y_i = s_i/m_i$, per l'i-esima combinazione di variabili concomitanti, $i = 1, \ldots, n$. Se i dati sono non raggruppati, $m_i = 1$ per ogni $i = 1, \ldots, n$. Si indicano con $y_1, \ldots, y_n$ le osservazioni sulla risposta e si assume per le corrispondenti variabili casuali $Y_1, \ldots, Y_n$

$$m_i Y_i \sim Bi(m_i, \pi_i), \quad \text{ossia} \quad Y_i \sim DE_1(\mu_i, \frac{1}{m_i}\mu_i(1-\mu_i)),$$

con $\mu_i = \pi_i$, $i = 1, \ldots, n$.

Si assume inoltre

$$g(\mu_i) = \eta_i = \mathbf{x}_i \beta = \sum_{r=1}^{p} \beta_r x_{ir},$$

con $g(\cdot)$ funzione di legame da specificare, la funzione di legame canonica è la funzione logistica o logit

$$g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \mathbf{x}_i\beta.$$

Altri tipi di funzioni di legame quali la funzione probit, log-log, log-log-complementare, cauchit, sono presentate in Salvan et al. (2020).

2.3 Metodo dei minimi quadrati pesati iterati

Per risolvere le equazioni di verosimiglianza (2.14) è necessario in genere procedere con metodi iterativi, come il metodo di Newton-Raphson (cfr. Pace & Salvan (2001), paragrafo 4.2). Sia l_β il vettore con elementi l_r e $j_{\beta\beta}$ il blocco della matrice di informazione osservata con elementi $-l_{rs}$. La $(m+1)$-esima iterazione fornisce l'approssimazione:

$$\hat{\beta}^{(m+1)} = \hat{\beta}^{(m)} + \left[j_{\beta\beta}(\hat{\beta}^{(m)})\right]^{-1} l_\beta(\hat{\beta}^{(m)}).$$

Se si sostituisce $j_{\beta\beta}$ con il suo valore atteso $i_{\beta\beta}$ si mantiene la convergenza dell'algoritmo e si semplificano le espressioni (metodo scoring di Fisher), ottenendo

$$\hat{\beta}^{(m+1)} = \hat{\beta}^{(m)} + \left[i_{\beta\beta}(\hat{\beta}^{(m)})\right]^{-1} l_\beta(\hat{\beta}^{(m)}),$$

da cui si ricava

$$i_{\beta\beta}(\hat{\beta}^{(m)})\hat{\beta}^{(m+1)} = i_{\beta\beta}(\hat{\beta}^{(m)})\hat{\beta}^{(m)} + l_\beta(\hat{\beta}^{(m)}). \tag{2.19}$$

Dopo opportune sostituzioni si può ottenere l'espressione

$$l_r = \sum_{i=1}^{n} x_{ir}(y_i - \mu_i)\omega_i g'(\mu_i),$$

e la relativa rappresentazione in forma matriciale

$$l_\beta = X^\top W u,$$

dove $u = ((y_1 - \mu_1)g'(\mu_1), \ldots, (y_n - \mu_n)g'(\mu_n))^\top$.

Applicando la 2.17, la 2.19 si può scrivere come

$$X^\top W X \hat{\beta}^{(m+1)} = X^\top W z^{(m)}, \tag{2.20}$$

dove

$$z^{(m)} = X\hat{\beta}^{(m)} + u.$$

$z^{(m)}$ è detta variabile dipendente aggiustata, la cui generica componente è

$$z_i^{(m)} = \mathbf{x}_i\hat{\beta}^{(m)} + (y_i + \mu_i)g'(\mu_i), \quad i = 1, \ldots, n. \tag{2.21}$$

Le quantità W e $z^{(m)}$ nella 2.20 si intendono valutate in $\hat{\beta}^{(m)}$. La $(m+1)$-esima iterazione dell'algoritmo calcola $\hat{\beta}^{(m+1)}$ come stima dei minimi quadrati generalizzati (Salvan et al. (2020) sezione 1.6.3) in un modello lineare avente come matrice di disegno X, come variabile risposta $z^{(m)}$ e come matrice dei pesi W, entrambe calcolate in calcolata in $\hat{\beta}^{(m)}$. Dato che la matrice dei pesi varia per ogni iterazione, l'algoritmo è detto dei minimi quadrati pesati iterati (*Iteratively Reweighted Least Squares, IRLS*). Raggiunta la convergenza dell'algoritmo si otterrà

$$\hat{\beta} = (X^\top \hat{W} X)^{-1} X^\top \hat{W} \hat{z},$$

dove $\hat{z} = X\hat{\beta} + \hat{u}$, con $\hat{u}$ la stima di u in $\hat{\beta}$.

2.4 VGLM

I *Vector Generalized Linear Models* (VGLM) rappresentano un'estensione multivariata dei GLM, e comprendono anche distribuzioni non incluse nella famiglia di dispersione esponenziale. Per approfondimenti, si rinvia a Yee (2015).

Generalizzando la (2.1) si ottiene,

$$p(y_i; \theta_i; \phi) = \exp\left\{\frac{y_i^\top \theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}, \tag{2.22}$$

dove $y_i = (y_{i1}, \ldots, y_{id})^\top \in \mathcal{Y} \subseteq \mathbb{R}^d$ è il vettore della variabile risposta realizzazione di variabili casuali indipendenti per $i = 1, \ldots, n$, $\theta_i \in \Theta \subseteq \mathbb{R}^d$, $\phi > 0$ è un parametro di dispersione e $a_i(\phi) > 0$. $x_i = (x_{i1}, \ldots, x_{ip})^\top$ è il vettore delle covariate, usualmente con un' intercetta $x_{i1} = 1$. Il j-esimo predittore lineare è

$$g_j(\mu_{ij}) = \eta_{ij} = \beta_j^\top x_i = \sum_{s=1}^{p} \beta_{js} x_{is}, \quad j = 1, \ldots, d, \quad i = 1, \ldots, n, \tag{2.23}$$

dove $\mu_i = \dfrac{\partial b(\theta_i)}{\partial \theta_i} = (\mu_{i1}, \ldots, \mu_{id})^\top$ vettore d-dimensionale con componenti $\mu_{ij} = \dfrac{\partial b(\theta_{ij})}{\partial \theta_{ij}}$. Le funzioni di legame $g_j(\cdot)$ soddisfano le usuali proprietà di monotonicità, coerenza e

differenziabilità. $W = (W_1, \ldots, W_n)$ è la matrice dei pesi detta *working matrix*, dove $W_i = -\mathbb{E}[\partial l_i/(\partial \eta_i \partial \eta_i^\top)]$, per ogni componente l_i della log-verosimiglianza $l = \sum_{i=1}^n l_i$. La matrice di covarianza stimata nei VGLM è della forma

$$\widehat{Var}(\hat{\beta}) = (X^\top W X)^{-1}, \tag{2.24}$$

valutata all'iterazione finale dell'algoritmo IRLS.

Capitolo 3

Effetto di Hauck-Donner

3.1 Introduzione

I contenuti di questo capitolo sono tratti da Yee (2022) e da Hauck & Donner (1977).

Nell'ambito della regressione logistica, il test di Wald può presentare un comportamento anomalo come evidenziato da Hauck & Donner (1977). Più in dettaglio, quando si manifesta l'effetto di Hauck-Donner (HDE) la statistica test non cresce in modo monotono al crescere della distanza della stima di massima verosimiglianza dal valore nullo del parametro, portando così a inferenze non affidabili e risultati che inducono conclusioni errate. In generale, si può affermare che il test di Wald risulta affidabile se la statistica sufficiente osservata è lontana dall'involucro convesso del suo supporto.

I metodi presentati in seguito sono basati su un approccio empirico e sono applicabili a tutti i GLM e VGLM. In questo capitolo sarà d'interesse l'ipotesi nulla del tipo $H_{0s} : \beta_s = \beta_{s0}$ per qualche valore fissato β_{s0}, solitamente 0, $s = 1, \ldots, p$. La versione unilaterale della statistica di Wald per β_s è

$$r_{eP}(\hat{\beta}_s) = \frac{\hat{\beta}_s - \beta_{s0}}{SE(\hat{\beta}_s)}, \quad \text{dove} \quad W_{eP}(\hat{\beta}_s) = r^2_{eP}(\hat{\beta}_s) \dot{\sim} \chi^2_1 \quad \text{sotto } H_0.$$

In questo Capitolo si indica con $r_{eP}(\hat{\beta}_s)$ la quantità $r_{eP}(\psi)$ definita dalla (1.12) vista come funzione della stima $\hat{\beta}_s$ avendo fissato il valore β_s ipotizzato da H_0, analoga notazione sarà mantenuta per tutti i test introdotti nel Capitolo 1. Se lo standard error al denominatore di $r_{eP}(\hat{\beta}_s)$ è calcolato nella stima di massima verosimiglianza, allora il pericolo di incorrere nell'HDE è sempre presente, mentre, come verrà presentato in seguito, calcolando lo standard error al valore fissato β_{s0} si ottiene un test privo dell'effetto, basato sulle quantità $r^*_{eP}(\psi)$ o $r^\dagger_{eP}(\psi)$ definite dalle (1.15) e (1.16). Per quanto

riguarda la notazione, l'elemento (s,t) della matrice $i_{\beta\beta} = X^\top WX$ è indicato come i_{st} mentre lo stesso elemento dell' inversa $i^{-1}_{\beta\beta}$ è i^{st}.

3.2 Descrizione dell'effetto

Hauck & Donner (1977) illustrano l'anomalia del test di Wald basandosi sul modello logistico elementare con $Y_i \sim Bi(\mu_i, 1)$ indipendenti per $i = 1, \ldots, n$, e

$$\text{logit } \mu_i = \beta_1 + \beta_2 x_{i2}, \qquad i = 1, \ldots, n, \tag{3.1}$$

con $x_{i2} = 0$ per $i = 1, \ldots, n_1$ $(n_1 < n)$ e $x_{i2} = 1$ per $i = n_1{+}1, \ldots, n$, variabile indicatrice, per due gruppi di osservazioni con numerosità n_1 e $n_2 = n - n_1$. Il valore atteso per le osservazioni risulta dunque $\mu_i = \mu_1 = \dfrac{e^{\beta_1}}{1 + e^{\beta_1}}$ se $x_{i2} = 0$, $\mu_i = \mu_2 = \dfrac{e^{\beta_1+\beta_2}}{1 + e^{\beta_1+\beta_2}}$ se $x_{i2} = 1$.

I dati sono riassunti dalla tabella di conteggi di dimensione 2×2:

	$y_i = 0$	$y_i = 1$	
$x_{i2} = 0$	$n_1 - s_1$	s_1	n_1
$x_{i2} = 1$	$n_2 - s_2$	s_2	n_2
	$n - s$	s	n

Tabella 3.1: Dataset di Hauck-Donner.

dove $s = \sum_{i=1}^n y_i$, $s_1 = \sum_{i=1}^n y_i(1 - x_{i2})$, $s_2 = \sum_{i=1}^n y_i x_{i2}$, sono rispettivamente il totale dei successi e il totale dei successi nei due gruppi. L'obiettivo è quello di testare l'uguaglianza delle proporzioni di successi nei due gruppi verificando l'ipotesi $H_0 : \mu_1 = \mu_2$, equivalente a saggiare l'ipotesi $H_0 : \beta_2 = 0$ contro l'alternativa bilaterale, in quanto β_2 corrisponde al logaritmo del rapporto di quote.
Come esempio si considera $n_1 = n_2 = 100$ e $s_1 = 25$ oppure $s_1 = 50$ per valutare il comportamento della statistica di Wald al variare del valore di s_2, che assume valori nell'intervallo discreto $\{1, \ldots, n_2 - 1\}$, evitando valori estremi dell'intervallo per non incorrere in perfetta separazione dei dati. Siano $\hat\pi_1 = s_1/n_1$ e $\hat\pi_2 = s_2/n_2$ le proporzioni di successi osservate nei due gruppi.
In questo paragrafo il test di Wald sarà confrontato con il test log-rapporto di verosimiglianza per valutare il diverso comportamento delle due statistiche test all'insorgere dell'effetto di Hauck-Donner. Si considerino le versioni bilaterali dei due test.

Dalla Tabella 3.2 (b) si evince che $W_{eP}(\hat{\beta}_2)$ cresce all'aumentare di $|\hat{\pi}_2 - \hat{\pi}_1|$ per $\hat{\pi}_2$ vicino a $\hat{\pi}_1$, ma ad una certa soglia comincia a diminuire, precisamente per $\hat{\pi}_2 < 0.03$ e per $\hat{\pi}_2 > 0.91$, la statistica log-rapporto di verosimiglianza $W_P(\hat{\beta}_2)$ non è invece oggetto dell'effetto risultando monotona crescente. La statistica di Wald presenta questa anomalia poichè per $\hat{\pi}_1$ sufficientemente grande, il denominatore di $W_{eP}(\hat{\beta}_2)$, che coincide con $i^{22}(\hat{\beta})$, cresce più velocemente del numeratore che nel test di nullità preso in esame coincide con $(\hat{\beta}_2)^2$. Come dimostrato in Hauck & Donner (1977), la statistica di Wald tende quindi a zero al divergere del valore di $\hat{\beta}$.

TABELLA 3.2: Test di Wald e log-rapporto di verosimiglianza per saggiare l'ipotesi $H_0 : \beta_2 = 0$ contro alternativa bilaterale, al variare di $\hat{\pi}_2$ con $\hat{\pi}_1$ fissato.

(a) $\hat{\pi}_1 = 0.50$

$\hat{\pi}_2$	$\hat{\beta}_2$	$W_{eP}(\hat{\beta}_2)$	$W_P(\hat{\beta}_2)$
0.60	0.405	2.01	2.02
0.70	0.847	8.19	8.40
0.80	1.386	18.75	20.27
0.90	2.197	31.95	40.70
0.93	2.587	34.56	49.69
0.94	2.752	34.85	53.16
0.95	2.944	34.61	56.94
0.99	4.595	20.11	77.27

(b) $\hat{\pi}_1 = 0.25$

$\hat{\pi}_2$	$\hat{\beta}_2$	$W_{eP}(\hat{\beta}_2)$	$W_P(\hat{\beta}_2)$
0.01	-3.497	11.50	30.89
0.02	-2.793	13.85	26.24
0.03	-2.377	14.24	22.57
0.04	-2.079	13.78	19.52
0.05	-1.846	12.91	16.91
0.10	-1.099	7.34	8.01
0.15	-0.636	3.07	3.15
0.35	0.480	2.37	2.39
0.45	0.898	8.60	8.88
0.55	1.299	18.01	19.11
0.65	1.718	30.33	33.30
0.75	2.197	45.27	52.32
0.85	2.833	60.92	78.25
0.90	3.296	66.05	95.26
0.91	3.412	66.37	99.14
0.92	3.541	66.28	103.23
0.95	4.043	61.95	117.03
0.99	5.694	30.48	141.96

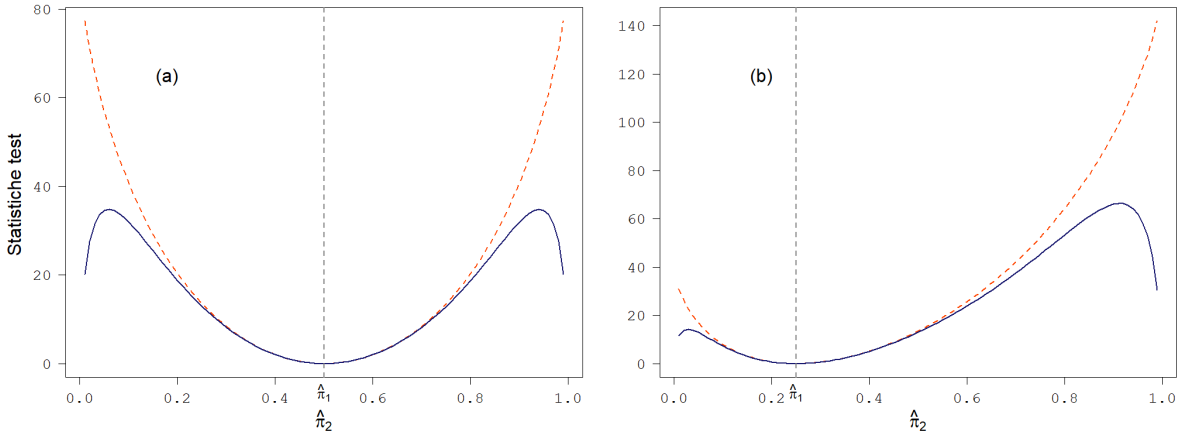

FIGURA 3.1: Test di Wald (linea continua) e log-rapporto di verosimiglianza (linea tratteggiata) per saggiare l'ipotesi $H_0 : \beta_2 = 0$ contro alternativa bilaterale, al variare di $\hat{\pi}_2$ per $\hat{\pi}_1 = 0.50$ (a) e per $\hat{\pi}_1 = 0.25$ (b).

3.3 Metodi di rilevazione dell'effetto nei VGLM

Yee (2022) espone una metodologia per rilevare l'effetto di Hauck-Donner basata sul calcolo della derivata della statistica di Wald per la vasta famiglia di modelli introdotta nel paragrafo 2.4.

Nei VGLM, la matrice di disegno è della forma $X_{VLM} = (X_1^\top, \ldots, X_n^\top)^\top$, per cui $i(\beta) = \sum_{i=1}^n X_i^\top W_i X_i$ e

$$\frac{\partial^\nu i}{\partial^\nu \hat{\beta}_{js}} = \sum_{i=1}^{n} X_i^\top \frac{\partial^\nu W_i}{\partial^\nu \hat{\beta}_{js}} X_i, \tag{3.2}$$

per $\nu \in \mathbb{N}^+$, $j \in \{1, \ldots, d\}$, $s \in \{1, \ldots, p\}$. Si può calcolare una derivata prima di (i^{ss}) come

$$\frac{\partial\, i^{-1}}{\partial \hat{\beta}_{js}} = -i^{-1} \frac{\partial\, i}{\partial \hat{\beta}_{js}} i^{-1}. \tag{3.3}$$

Per saggiare l'ipotesi nulla $H_0 : \beta_{js} = \beta_{js0}$ contro $H_1 : \beta_{js} \neq \beta_{js0}$, basandosi sulla statistica di Wald si può ottenere un test di rilevazione dell'effetto di Hauck-Donner, calcolando la derivata di $r_{eP}(\hat{\beta}_{js})$ rispetto al relativo $\hat{\beta}_{js}$

$$\frac{\partial r_{eP}(\hat{\beta}_{js})}{\partial \hat{\beta}_{js}} = \frac{\partial}{\partial \hat{\beta}_{js}} \frac{\hat{\beta}_{js} - \beta_{js0}}{\sqrt{i^{ss}}} = \frac{1}{\sqrt{i^{ss}}} \left[1 - \frac{\hat{\beta}_{js} - \beta_{js0}}{2} \frac{(i^{ss})'}{i^{ss}} \right]. \tag{3.4}$$

In un VGLM con risposta univariata ($d = 1$), dalla (3.4), dove i^{ss} e $(i^{ss})'$ sono valutate in $\hat{\beta}$, la condizione per la presenza dell'HDE è

$$\frac{1}{2}(\hat{\beta}_s - \beta_{s0}) \frac{d \log i^{ss}}{d \hat{\beta}_s} - 1 > 0. \tag{3.5}$$

In generale, se la (3.4) è definita negativa, si può concludere che l'HDE è presente nel test relativo a $\hat{\beta}_s$, pertanto l'affidabilità dello *standard error* e del p-value del test risultano compromesse. È fondamentale non trascurare questa anomalia per non incorrere in errori di valutazione dei parametri, e all'evenienza adottare delle contromisure, ad esempio utilizzare altri test non soggetti al problema come i test log-rapporto di verosimiglianza e di Rao, oppure, come verrà introdotto in seguito, optare per una versione del test di Wald priva di HDE. La funzione `hdeff` della libreria `VGAM` in `R`, implementa il metodo di diagnostica dell'effetto di Hauck-Donner presentato da Yee (2022). In Appendice sono riportati ulteriori risultati riguardanti l'effetto di Hauck-Donner nel contesto di distribuzioni monoparametriche.

3.4 Metodi di classificazione dell'effetto nei VGLM

In questo paragrafo viene introdotto un metodo per classificare l'effetto di Hauck-Donner in base alla sua gravità proposto da da Yee (2022). La classificazione dell'effetto è utile per non trascurare quei casi in cui la statistica di Wald inizia ad assumere un comportamento anomalo, ma senza presentare formalmente l'effetto, avendo derivata ancora positiva, ma una concavità negativa. A tal proposito si possono valutare le prime due derivate di $r_{eP}(\hat{\beta}_s)$ rispetto a $\hat{\beta}_s$, e sapendo che $r_{eP}(\hat{\beta}_s)$ è approssimabile asintoticamente con una parabola in un intorno dell'origine, si può assumere che la funzione $r_{eP}(\hat{\beta}_s)$ sia convessa-concava-convessa per $\hat{\beta}_s > 0$ e allo stesso modo per $\hat{\beta}_s < 0$. L'obiettivo suddividere lo spazio parametrico in base alla gravità dell'effetto, sfruttando la forma della funzione della statistica di Wald in funzione di $\hat{\beta}_s$.

Sia $\hat{\beta}_s$ l'asse delle ascisse, $r_{eP}(\hat{\beta}_s)$ l'asse delle ordinate, e $\zeta(\hat{\beta}_s)$ l'intersezione della retta perpendicolare alla tangente della funzione $r_{eP}(\hat{\beta}_s)$ nel punto $(\hat{\beta}_s, r_{eP}(\hat{\beta}_s))$, con l'asse $\hat{\beta}_s$ (si veda Figura 3.2). Il valore di $\zeta(\hat{\beta}_s)$ al variare di $\hat{\beta}$ è utile per determinare il comportamento della curva al di là dei punti di flesso; l'utilizzo di $\zeta(\hat{\beta}_s)$ rispetto ad altri metodi come la funzione tangente, risolve il problema di valori indeterminati in punti di discontinuità e di valori infiniti.

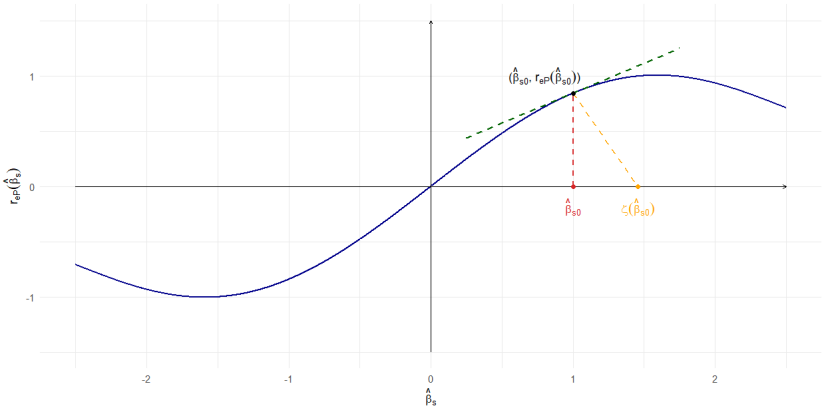

FIGURA 3.2: Grafico esemplificativo della procedura per ottenere $\zeta(\hat{\beta})$ per un generico valore $\hat{\beta}_{s0}$.

Per semplicità si indichino con $r'_{eP}(\hat{\beta}_s)$ e $r''_{eP}(\hat{\beta}_s)$ le derivate prima e seconda di $r_{eP}(\hat{\beta}_s)$ rispetto a $\hat{\beta}_s$ e in modo analogo per la funzione $\zeta(\hat{\beta}_s)$. Si cercano dei valori critici per determinare delle soglie di severità dell'effetto. In particolare si suddivide lo spazio parametrico in base alla gravità: *nessuna (n), molto debole (md), debole (d), moderata (m), forte (f), estrema (e).* I valori critici sono definiti come $r''_{eP}(\hat{\beta}_s^{n-md}) = 0, \zeta'(\hat{\beta}_s^{md-d}) = 0, r'_{eP}(\hat{\beta}_s^{d-m}) = 0, \zeta'(\hat{\beta}_s^{m-f}) = 0, r''_{eP}(\hat{\beta}_s^{f-e}) = 0$, $\hat{\beta}_s^{md-d}$ e $\hat{\beta}_s^{f-e}$ sono quindi punti di flesso. Per $\hat{\beta}_s > 0$

$$0 \leq \hat{\beta}_s^{n-md} \leq \widehat{\beta}_s^{md-d} \leq \widehat{\beta}_s^{d-m} \leq \widehat{\beta}_s^{m-f} \leq \widehat{\beta}_s^{f-e} \leq \infty,$$

e

$$\zeta(\hat{\beta}_s) = \hat{\beta}_s + r_{eP}(\hat{\beta}_s) \cdot r'_{eP}(\hat{\beta}_s),$$

$$\zeta'(\hat{\beta}_s) = 1 + \{r'_{eP}(\hat{\beta}_s)\}^2 + r_{eP}(\hat{\beta}_s) \cdot r''_{eP}(\hat{\beta}_s).$$

Per il dataset di Hauck-Donner riportato nella Tabella 3.1 la classificazione dell'effetto al variare di s_2 con $s_1 = 25$ risulta:

Gravità	Nessuna	Molto debole	Debole	Moderata	Forte	Estrema
s_2	(26,40)	$(11, 25) \cup (41, 69)$	$(3, 10) \cup (70, 91)$	$2; (92, 97)$	$1; 98$	99

TABELLA 3.3: Severità HDE al variare di s_2.

La determinazione dei livelli di gravità dell'effetto di Hauck-Donner permette il partizionamento dello spazio parametrico Θ in un sottoinsieme Θ^0 dove la gravità è al massimo *debole* e Θ^{HDE} con gravità almeno *moderata*, nel quale ricadono le stime dei parametri che causano il comportamento anomalo del test di Wald.
Il metodo di classificazione dell'HDE presentato è implementato nella funzione `hdeffsev`

della libreria `VGAM` in `R`.

3.5 Metodi di riduzione dell'effetto

Un metodo indicato da Yee (2022) per ovviare al problema dell'HDE è adottare una variante del test di Wald, nello specifico valutare lo *standard error* non più nella stima di massima verosimiglianza $\hat{\beta}$, ma nel suo valore fissato dall'ipotesi nulla. Si fa riferimento alla notazione introdotta nel paragrafo 1.5.

Si consideri il problema di verificare $H_{0s} : \beta_s = \beta_{0s}$ contro $H_{1s} : \beta_s \neq \beta_{0s}$, $s = 1, \ldots, p$, dove p è la dimensione del vettore dei parametri. Per ottenere il test HDE-*free*, nel calcolo dello *standard error* al denominatore della statistica di Wald si sostituisce $\hat{\beta}_s$ con il valore fissato nell'ipotesi nulla β_{0s}, in modo tale da annullare la derivata dello *standard error* rispetto a $\hat{\beta}_s$. Per quanto riguarda le stime degli altri coefficienti, si può decidere di procedere in due diversi modi, utilizzando o meno le stime vincolate al valore fissato β_{0s}. Sia $\tilde{\beta}_{-s}$ la stima vincolata per il vettore di parametri di disturbo e $\hat{\beta}_{-s}$ la stima non vincolata, ottenuta utilizzando la componente $\hat{\lambda}$ della stima di massima verosimiglianza. Si ottiene l'informazione attesa $i(\psi, \hat{\lambda}_\psi)$ nel primo caso, altrimenti $i(\psi, \hat{\lambda})$. Per ottenere lo standard error della statistica immune all'effetto per un generico β_s nell'ambito dei VGLM, occorre seguire i seguenti passaggi.

1. Ottenere $\hat{\beta}_{\psi_{-s}}$ tramite stima vincolata, altrimenti usare la stima di massima verosimiglianza $\hat{\beta}_{-s}$.
2. Calcolare η_i utilizzando $(\beta_{0s}, \tilde{\beta}_{-s})$ o $(\beta_{0s}, \hat{\beta}_{-s})$, $i = 1, \ldots, n$.
3. Aggiornare i generici valori stimati μ_i e i pesi W_i con i valori η_i.
4. Effettuare la decomposizione di Cholesky U_i dei W_i.
5. Calcolare $\text{diag}(U_1, \ldots, U_n)X$ e la relativa decomposizione QR.
6. Calcolare R^{-1} e successivamente $(R^{-1}R^{-\top})_{ss}^{1/2}$ ottenendo lo standard error desiderato.

I passaggi vanno effettuati per ogni β_s.

Yee (2022) ha riportato diversi risultati di simulazione sui costi computazionali delle procedure per trattare l'HDE. Il costo computazionale del test di rilevazione dell'effetto su tutti i coefficienti di regressione è circa un terzo rispetto a quello dei corrispondenti test di Wald HDE-free con stime vincolate. Questi ultimi sono il 30% più onerosi rispetto

ai test log-rapporto di verosimiglianza e presentano un costo simile ai test score. Senza stime vincolate, i test di Wald HDE-free risultano invece il 25% meno costosi rispetto ai test log-rapporto di verosimiglianza. Pertanto, i test log-rapporto di verosimiglianza e i test di Wald HDE-free sono comparabili in termini di costo computazionale.

Alla luce di questi risultati, Yee (2022) suggerisce di effettuare in primo luogo i test di rilevazione dell'effetto. In caso di esito positivo, raccomanda di optare per il test log-rapporto di verosimiglianza qualora l'obiettivo sia ottenere un p-value affidabile. Se è d'interesse anche stimare l'errore standard, è preferibile applicare il test di Wald immune all'effetto. Per quest'ultimo, si consiglia di utilizzare la variante senza stime vincolate quando il costo computazionale rappresenta un problema, poiché quest'opzione comporta un costo circa dimezzato rispetto alla versione con stime vincolate.

Come verrà esposto nel Capitolo 4, l'effetto di Hauck-Donner può insorgere in caso di perfetta o quasi perfetta separazione nei dati, ossia quando una covariata discrimina perfettamente (o quasi perfettamente) la variabile risposta. In tal caso il problema può essere risolto utilizzando stimatori con riduzione della distorsione in media e mediana. Per maggiori approfondimenti si rimanda a Kosmidis et al. (2020). I metodi sono implementati in `R` nella libreria `brglm2` (Kosmidis, 2018).

Capitolo 4

Studio di simulazione

Il seguente Capitolo ha come obiettivo di testare i metodi proposti nella relazione e indagare sulle cause e conseguenze legate all'effetto di Hauck-Donner nell'ambito dei modelli lineari generalizzati, mediante uno studio di simulazione.

4.1 Introduzione

La simulazione si basa sul dataset `birthwt` analizzato da Kosmidis et al. (2020) e reperibile nella libreria `MASS` di `R`. Si tratta di $n = 100$ osservazioni del peso di neonati, la variabile risposta dicotomica `normwt` vale 1 se il peso è di almeno 2500 g e 0 altrimenti. Le variabili concomitanti sono :

- x_{i2} = `age`: variabile quantitativa discreta, età della madre in anni compiuti;
- x_{i3} = `race`: variabile dicotomica, vale 1 se la madre è di carnagione chiara e 0 altrimenti;
- x_{i4} = `smoke`: variabile dicotomica, vale 1 se la madre fumava durante la gravidanza e 0 altrimenti;
- x_{i5} = `ptl`: variabile dicotomica, vale 1 se la madre ha mai avuto parti prematuri e 0 altrimenti;
- x_{i6} = `ht`: variabile dicotomica, vale 1 se la madre ha mai sofferto di ipertensione e 0 altrimenti;
- x_{i7} = `loglwt`: variabile quantitativa continua, logaritmo del peso della madre.

Si adatta il modello di regressione logistica

$$\text{logit } \mu_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7}. \tag{4.1}$$

Il `summary` del modello adattato con la funzione `vglm` del pacchetto `VGAM`, produce un avvertimento riguardo il coefficiente β_7 relativo alla variabile `loglwt`, indicando la rilevazione dell'effetto di Hauck-Donner nella stima e oscurando la relativa statistica di Wald e il p-value corrispondente. Di seguito, è riportata una tabella riassuntiva con le quantità d'interesse nello studio dell'effetto di Hauck-Donner introdotte nei capitoli precedenti. I valori che vengono omessi di default nel summary del modello a causa dell'effetto sono qui riportati contrassegnandoli con un asterisco. Tutti i valori sono arrotondati alla terza cifra decimale.

TABELLA 4.1: Analisi della statistica di Wald nel modello (4.1), derivate prime e seconde di $r_{eP}(\hat{\beta}_s)$, derivate prime e seconde dello *standard error* e gravità HDE rilevata.

	Stima	*Standard error*	$r_{eP}(\hat{\beta}_s)$	$Pr(>\lvert r_{eP}(\hat{\beta}_s)\rvert)$	$r'_{eP}(\hat{\beta}_s)$	$r''_{eP}(\hat{\beta}_s)$	$(\sqrt{i^{ss}}\,)'$	$(\sqrt{i^{ss}}\,)''$	Gravità HDE
`Intercetta`	-8.496	5.826	-1.458	0.145	0.321	0.129	0.598	0.779	Nessuna
`age`	-0.067	0.053	-1.256	0.209	21.698	25.512	0.125	5.392	Debole
`racewhite`	0.690	0.566	1.219	0.223	1.630	-0.476	0.064	0.051	Nessuna
`smoke`	-0.560	0.576	-0.971	0.331	1.756	0.051	0.012	0.072	Nessuna
`ptl`	-1.603	0.697	-2.298	0.022	1.060	0.747	-0.114	0.122	Nessuna
`ht`	-1.211	0.924	-1.311	0.190	1.040	0.226	-0.029	0.113	Nessuna
`loglwt`	2.262	1.252	1.806*	0.071*	-0.241	-5.629	0.721	4.096	Estrema

L'attenzione ricade soprattutto sul coefficiente β_7 relativo alla variabile `loglwt` che presenta una gravità estrema dell'effetto di Hauck-Donner. Da una ulteriore analisi è emerso inoltre che se si usa la variabile originale senza la trasformazione logaritmica `lwt`, l'effetto non viene più rilevato (gravità = nessuna). Il motivo non è di semplice spiegazione poichè la causa dell'effetto in questo caso non è il problema di perfetta o quasi-separazione, una possibile causa sarà discussa in seguito. Si è verificato inoltre che la bontà di adattamento del modello rimane pressoché immutata utilizzando la variabile senza la trasformazione, pertanto in questo caso si potrebbe utilizzare la covariata originale. Tuttavia, in altre situazioni, la scelta di adottare una trasformazione logaritmica potrebbe essere dettata da esigenze interpretative e pertanto risultare importante ai fini dello studio.

Si confrontano ora i valori del test di Wald e della sua versione HDE-*free*, del test di Rao e del test log-rapporto di verosimiglianza, per $H_0 : \beta_s = 0$ contro l'ipotesi alternativa bilaterale, $s = 2, \ldots, 7$

	$r_{eP}(\hat{\beta}_s)$	$r^*_{eP}(\hat{\beta}_s)$	$r_P(\hat{\beta}_s)$	$r_{uP}(\hat{\beta}_s)$
`age`	-1.256	-1.245	-1.258	-1.256
`racewhite`	1.219	1.264	1.238	1.273
`smoke`	-0.971	-0.972	-0.974	-0.978
`ptl`	-2.298	-2.527	-2.422	-2.635
`ht`	-1.311	-1.295	-1.324	-1.331
`loglwt`	1.806	2.001	1.886	2.040

TABELLA 4.2: Confronto tra le statistiche di Wald, Wald HDE-free (contrassegnata da asterisco), test log-rapporto di verosimiglianza e test di Rao, s = 2,. . . ,7.

I vari test portano sostanzialmente alle medesime conclusioni inferenziali pur essendo stata rilevata una gravità HDE estrema sul coefficiente in questione. Questo risultato verrà indagato nello studio di simulazione, mettendo in discussione l'efficacia del metodo di classificazione presentato nel paragrafo 3.4.

4.2 Risultati

Vengono effettuate $N = 10000$ simulazioni a partire dal modello (4.1). In particolare, verrà simulato casualmente solamente il valore della variabile risposta dicotomica a partire da una variabile casuale binomiale con parametro $\pi = 0.5$. Le covariate rimangono fissate poichè per indagare sull'effetto di Hauck-Donner occorre mantenere la matrice di disegno originale da cui scaturisce l'effetto in almeno un coefficiente.

Di seguito sono riportate, per ogni coefficiente, le proporzioni di volte in cui l'HDE è risultato presente nelle 10000 simulazioni.

	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Proporzione HDE	0.0015	0.0001	0.0000	0.0002	0.0597	0.0027	0.6567

Per β_5 e β_7 si riportano le proporzioni delle severità degli effetti rilevate.

TABELLA 4.3

Proporzione severità HDE	β_5	β_7
Nessuna	0.9403	0.3433
Forte	0.0195	0.5278
Non determinata	0.0402	0.1289

La severità "non determinata" è dovuta alla mancata convergenza dell'algoritmo nella funzione `hdeffsev` in `R`, che implementa il metodo presentato nel paragrafo 3.4. In corrispondenza personale con l'Autore, si è avuta l'informazione che la funzione sarà aggiornata nel breve periodo risolvendo questo problema. Si specifica che, ogniqualvolta la gravità risulti "non determinata" l'effetto è comunque rilevato. Come previsto, l'unico coefficiente a manifestare problemi ricorrenti legati all'HDE è il β_7, con effetto rilevato in più del 65% delle simulazioni, confermando la alta gravità come si evince dalla Tabella (4.3).

Per quanto riguarda β_5, si riscontrano stime problematiche nel 6% circa dei casi e, per capire la natura dell'anomalia, è utile rappresentare graficamente le stime con un diagramma di dispersione. In una situazione regolare ci si aspetta un andamento monotono crescente di r_{eP} in funzione di $\hat{\beta}_5$.

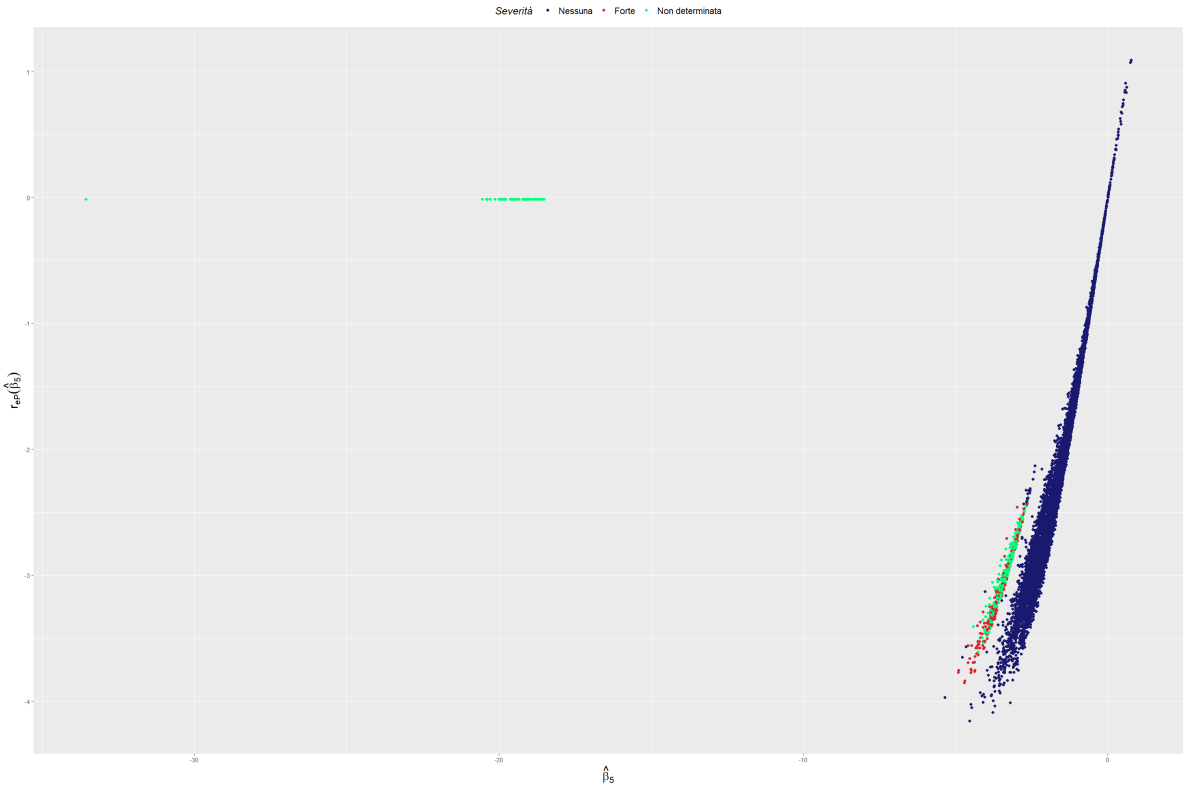

FIGURA 4.1: Diagramma di dispersione dei punti $(\hat{\beta}_5, r_{eP}(\hat{\beta}_5))$, distinti per severità dell'effetto.

Si può ipotizzare che la distorsione delle stime sia la causa dell'effetto di Hauck-Donner. Infatti, i punti più lontani dalla nuvola principale, con $\hat{\beta}_5$ dell'ordine di -20, sono caratterizzati da stime del coefficiente e da standard error che presentano valori anomali che portano a zero la statistica di Wald. La distorsione delle stime conduce quindi in questo caso all'insorgere dell'effetto.

In questi casi è ragionevole applicare dei metodi per ridurre la distorsione delle stime come quelli presentati da Kosmidis et al. (2020), per ottenere delle stime con riduzione della distorsione e per cui meno soggette all'HDE. Le funzioni per la rilevazione e la classificazione dell'effetto di Hauck-Donner in `R` sono però limitate alla classe dei modelli `VGLM`, pertanto è al momento possibile confrontare risultati ottenuti con diversi stimatori solo da un punto di vista descrittivo, riportando il grafico delle stime ottenute ad esempio con la riduzione della distorsione in media. Implementare le funzioni e i metodi introdotti nella relazione anche per altre classi di modelli contenuti ad esempio nella libreria `brglm2` (Kosmidis, 2018) sarebbe un passo in avanti per poter indagare e trattare ulteriormente cause, conseguenze e rimedi nell'ambito dell'effetto di Hauck-Donner. L'apice BR indica stime calcolate con il metodo di riduzione della distorsione in media, o statistiche test calcolate nelle stime corrette.

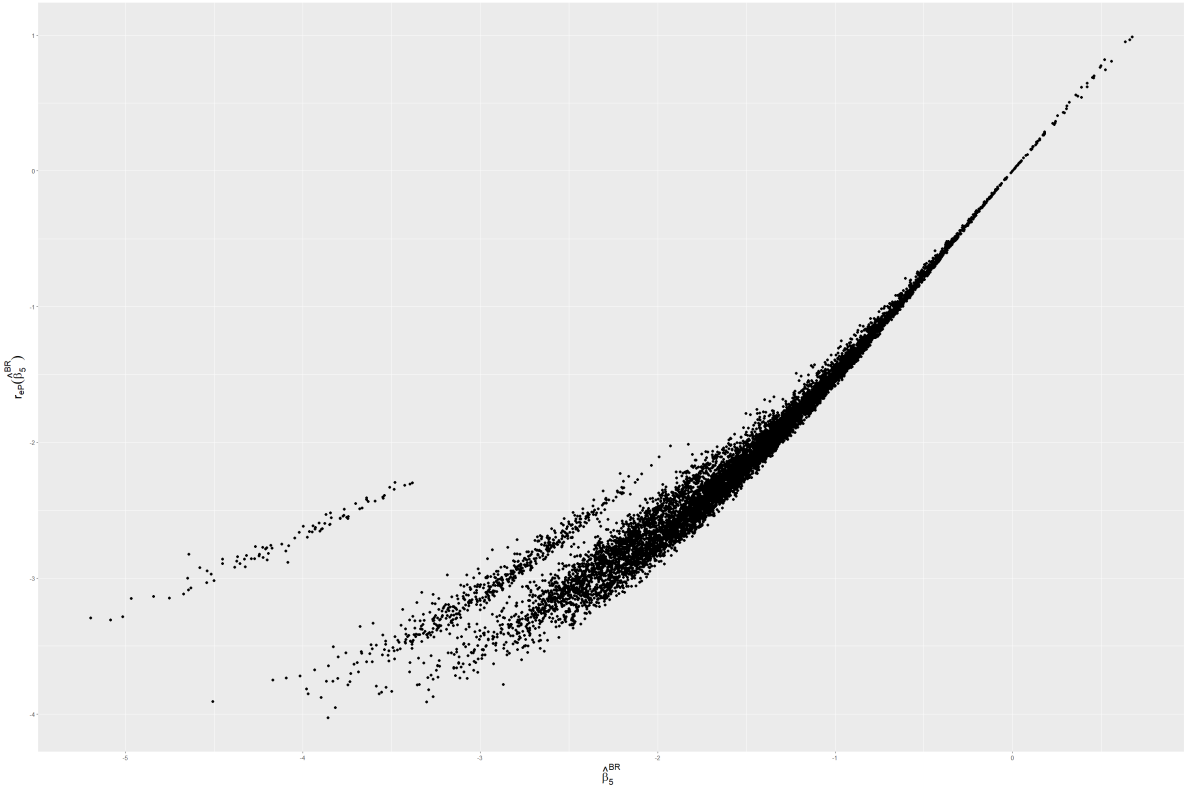

FIGURA 4.2: Diagramma di dispersione dei punti $(\hat{\beta}_5{}^{BR}, r_{eP}(\hat{\beta}_5{}^{BR}))$, dove $\hat{\beta}_5{}^{BR}$ è la stima di β_5 calcolata con metodo di riduzione del *bias* in media.

Come si può apprezzare, almeno graficamente la distribuzione delle statistiche di Wald è certamente più regolare rispetto a quello riportato dalla Figura 4.1. Nei campioni in cui valori anomali portavano a $r_{eP}(\hat{\beta}_5) \simeq 0$, l'uso dello stimatore con riduzione della distorsione ha prodotto risultati certamente più accettabili, tuttavia non è al momento

possibile ottenere dei test di rilevazione su modelli con stime corrette. Si rimanda a futuri studi verificare empiricamente la effettiva riduzione dell'effetto HDE.

Le stesse conclusioni si possono trarre analizzando i grafici relativi a β_6 di seguito riportati.

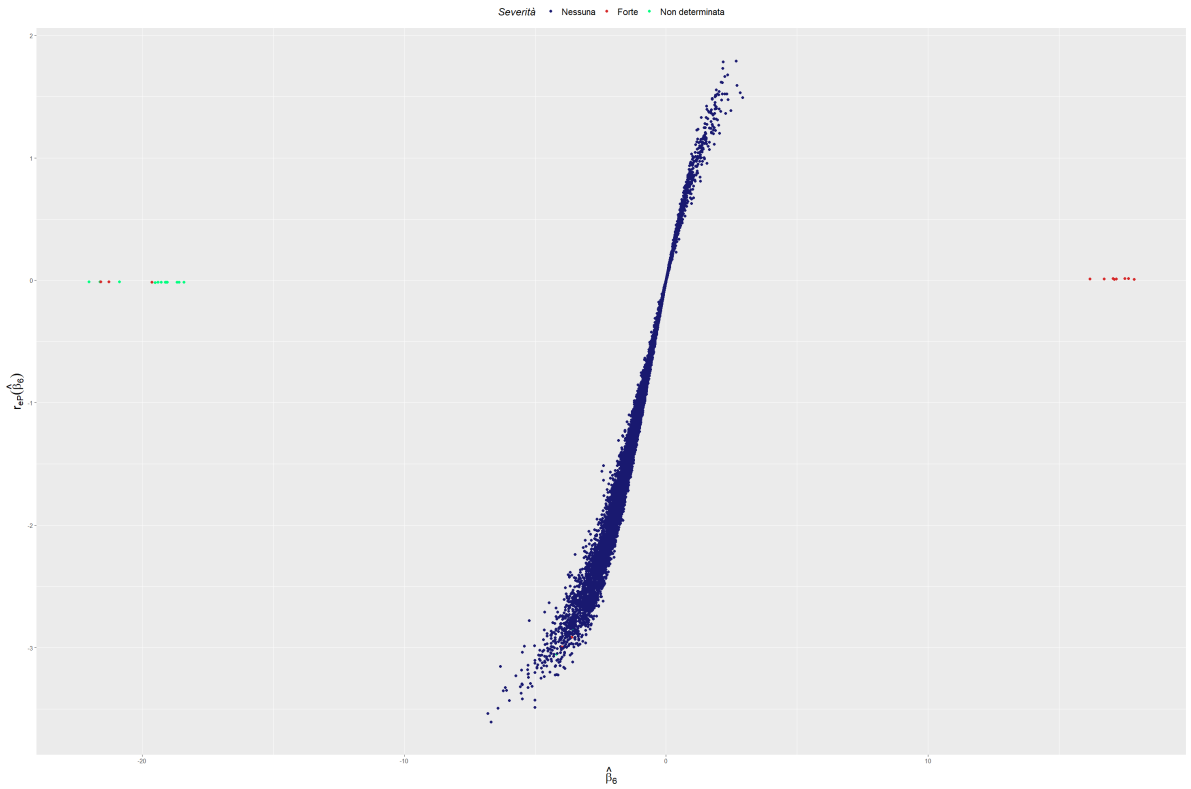

FIGURA 4.3: Diagramma di dispersione dei punti $(\hat{\beta}_6, r_{eP}(\hat{\beta}_6))$, distinti per severità dell'effetto.

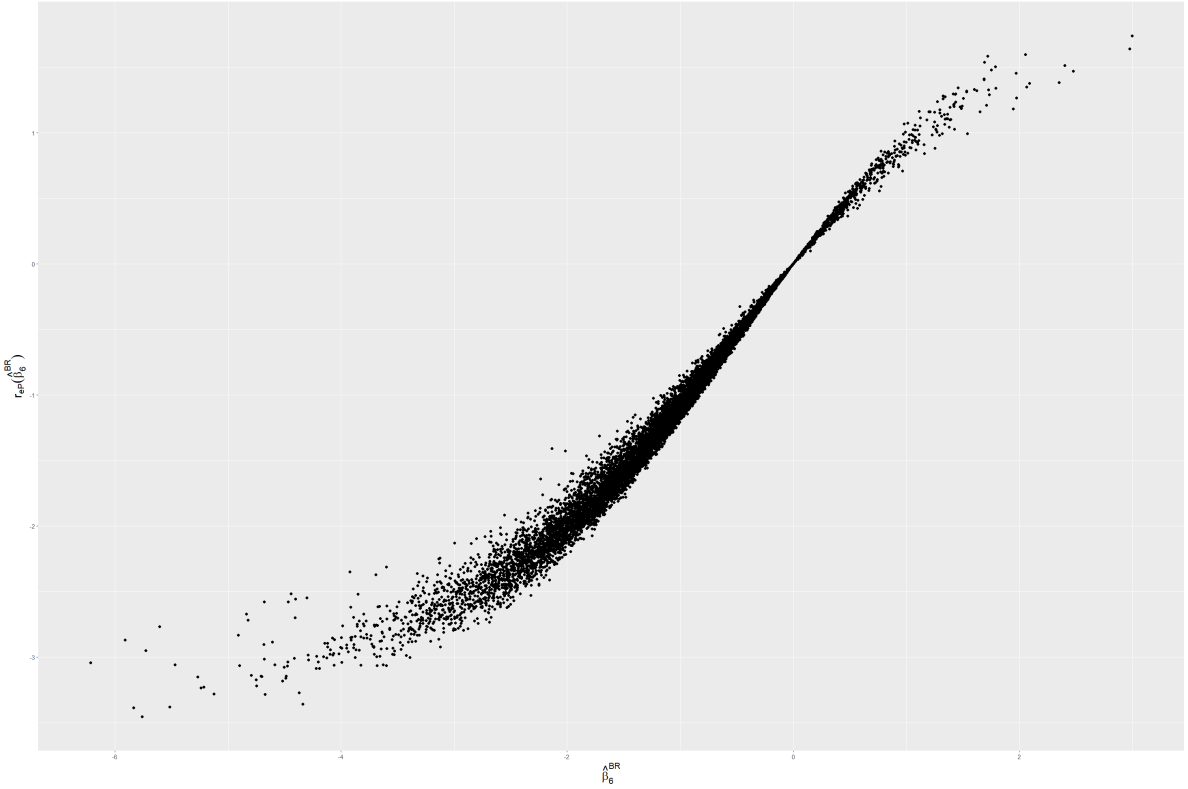

FIGURA 4.4: Diagramma di dispersione dei punti $(\hat{\beta}_6^{\,BR}, r_{eP}(\hat{\beta}_6^{\,BR}))$, dove $\hat{\beta}_6^{\,BR}$ è la stima di β_6 calcolata con metodo di riduzione del *bias* in media.

Si pone ora attenzione sul coefficiente β_7 e sull'andamento delle statistiche di Wald rispetto alle relative stime. Dalla Figura 4.5, con l'aiuto della curva non parametrica in grigio, si può apprezzare che la comparsa della diagnosi di presenza dell'effetto in forma forte coincide con il cambio di concavità di $r_{eP}(\hat{\beta}_7)$. Come introdotto nel paragrafo 3.4, l'effetto può essere diagnosticato anche per valori positivi della derivata prima, quando la concavità è negativa.

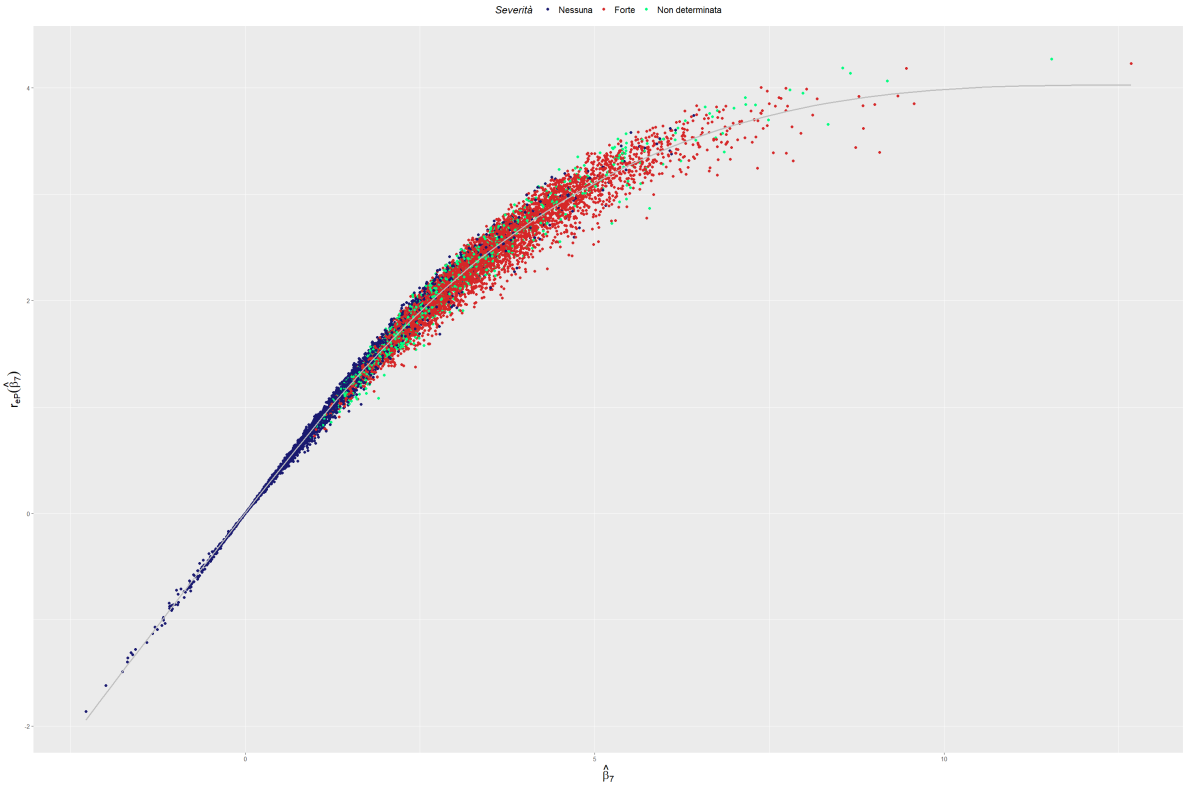

FIGURA 4.5: Diagramma di dispersione dei punti $(\hat{\beta}_7, r_{eP}(\hat{\beta}_7))$, distinti per severità dell'effetto, in grigio una curva non parametrica per approssimare la funzione $r_{eP}(\hat{\beta}_7)$.

I valori di $r_{eP}(\hat{\beta}_7)$ non sembrano giustificare la severità "forte" rilevata in più della metà delle simulazioni, dato che risulta monotona crescente in $\hat{\beta}_7$. Si vuole indagare l'efficacia del metodo di classificazione nel caso considerato.

A tal fine, si possono confrontare i vari test alternativi al test di Wald per verificare se essi portino a conclusioni sostanzialmente diverse.

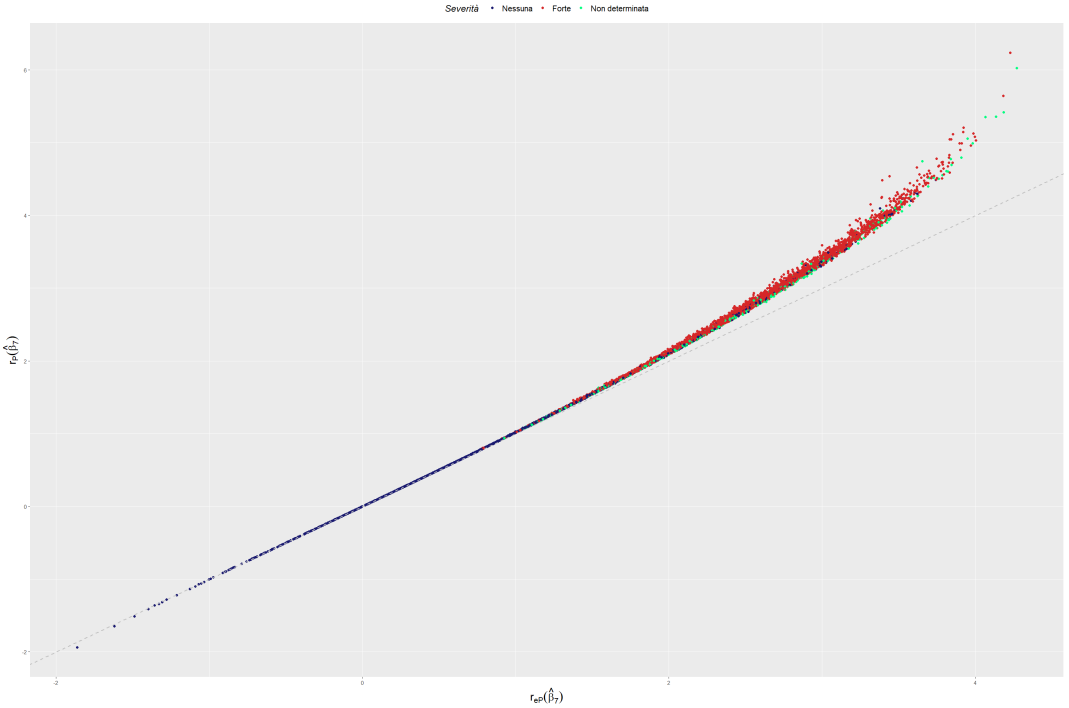

FIGURA 4.6: Diagramma di dispersione delle statistiche relative ai test di nullità di β_7, test di Wald in ascissa e test log-rapporto di verosimiglianza in ordinata, distinti per severità dell'effetto, in grigio la bisettrice del primo e terzo quadrante.

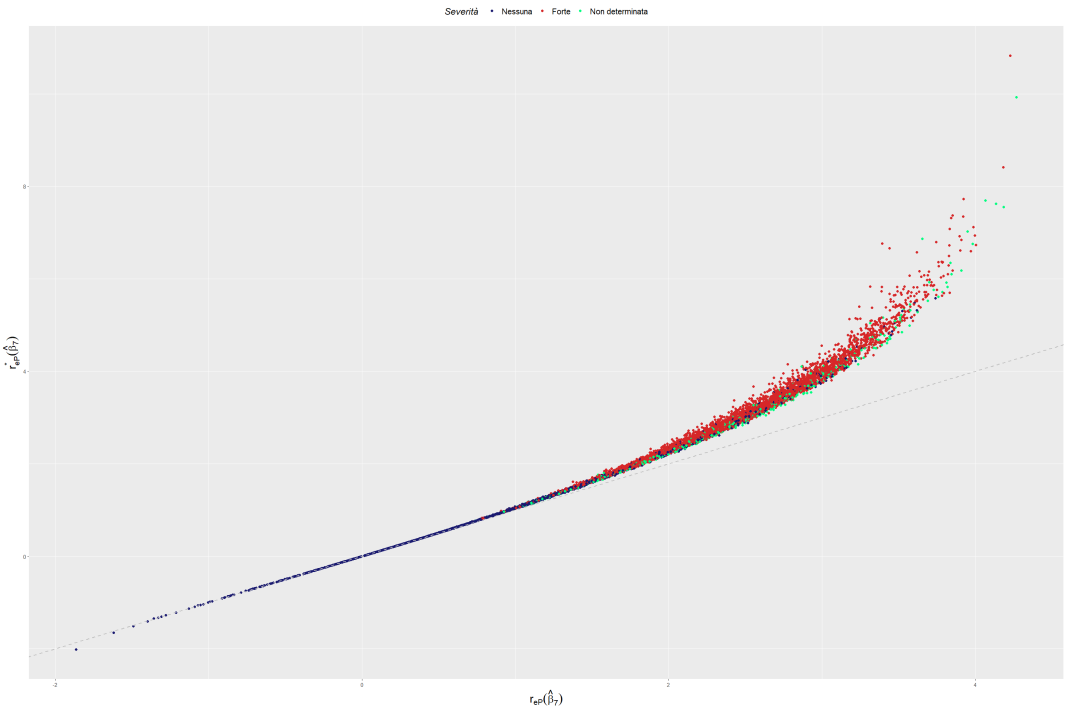

FIGURA 4.7: Diagramma di dispersione delle statistiche relative ai test di nullità di β_7, test di Wald in ascissa e test di Wald HDE-*free* in ordinata, distinti per severità dell'effetto, in grigio la bisettrice del primo e terzo quadrante.

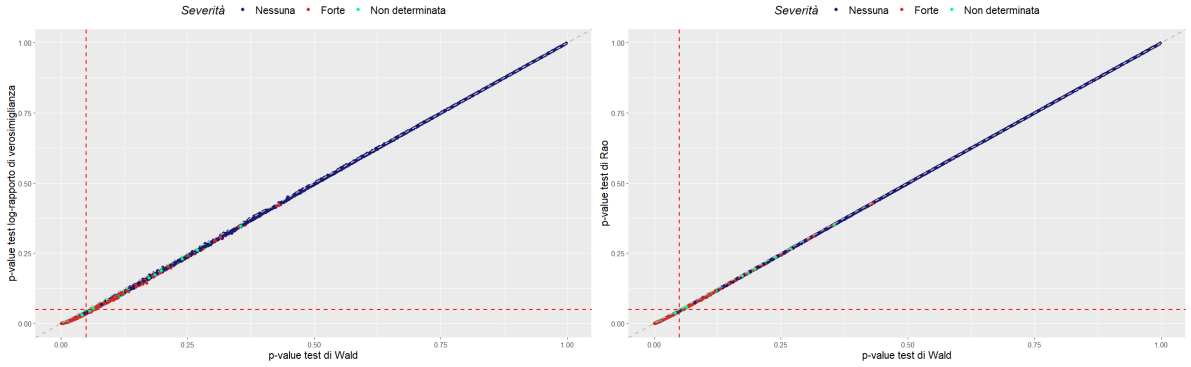

FIGURA 4.8: Confronto tra p-value del test di Wald con p-value del test log-rapporto di verosimiglianza e del test di Rao rispettivamente. In rosso le rette corrispondenti al valore usuale di significatività $\alpha = 0.05$, in grigio la bisettrice del primo e terzo quadrante.

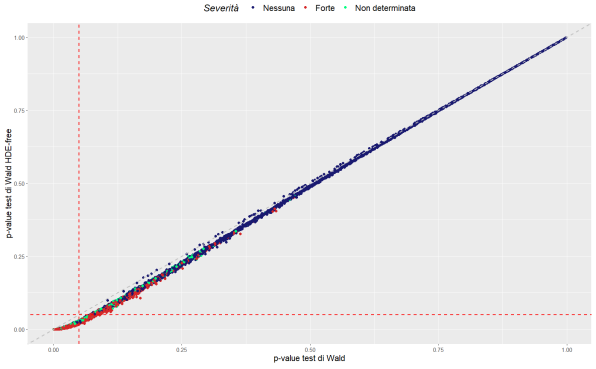

FIGURA 4.9: Confronto tra p-value dei test di Wald e Wald HDE-*free*. In rosso le rette corrispondenti al valore usuale di significatività $\alpha = 0.05$, in grigio la bisettrice del primo e terzo quadrante.

Come si evince dalla Figura 4.6, le statistiche test di Wald e log-rapporto di verosimiglianza si eguagliano quando nessun effetto è rilevato, mentre si nota un leggero scostamento tra le due statistiche quando è presente l'effetto di Hauck-Donner. Tuttavia come si può vedere dalla Figura 4.8, i p-value dei due test rimangono essenzialmente uguali indipendentemente dalla rilevazione o meno dell'effetto, in quanto le differenze tra le statistiche test si palesano oltre la soglia di 2, ove il p-value vale approssimativamente 0. Dalle Figure 4.7 e 4.9 si può apprezzare un analogo scostamento tra le statistiche di Wald e di Wald HDE-*free*, che in questo caso porta a conclusioni diverse in un limitato numero di simulazioni. Le conclusioni inferenziali a cui portano i vari test pertanto rimangono le stesse per questo coefficiente nonostante l'apparente allarme dovuto alla rilevazione dell'effetto con gravità "forte" in più della metà dei campioni simulati, facendo sorgere dunque alcune perplessità riguardo al metodo di diagnostica.

In secondo luogo, si è interessati a indagare la presenza dell'effetto di Hauck-Donner nel coefficiente β_7, che, come mostrato, non altera le conclusioni inferenziali, nonostante la alta gravità rilevata in maggior parte dei campioni simulati. Si pone l'attenzione sulla covariata `loglwt`. Infatti, come introdotto in precedenza, utilizzando la medesima variabile senza la trasformazione logaritmica, non viene rilevato l'effetto. Si ipotizza che il ristretto campo di variazione della covariata abbia un impatto sulla rilevazione del "falso" effetto, che, come visto in precedenza, non ha conseguenze su monotonicità della statistica di Wald, né sulle conclusioni inferenziali.

Al fine di indagare su questa ipotesi, si ricorre ad una simulazione ($N = 10000$) a partire da un modello logistico elementare, con dimensione campionaria $n = 100$.

$$\text{logit } \mu_i = \beta_1 + \beta_2 x_{i2}, \tag{4.2}$$

dove la variabile risposta y_i è simulata a partire da una variabile casuale binomiale $Bi(1, \mu_i)$, x_{i2} è una covariata simulata a partire da una variabile casuale normale $N(\mu, \sigma^2)$, con $\mu = 5$ fissato arbitrariamente e σ che sarà fatto variare in diverse simulazioni, al fine di confrontare i risultati con campi di variabilità di x_{i2} differenti.

σ	Proporzione HDE β_2
0.2	0.0057
0.1	0.0416
0.05	0.1184
0.01	0.3252
0.005	0.3882
0.001	0.4630
0.0005	0.4743

TABELLA 4.4: Proporzione di effetto rilevata sul coefficiente β_2, per ogni 10000 simulazioni effettuate con diversi valori di σ nella generazione di x_{i2}.

Come si può osservare dalla Tabella 4.4, al diminuire della variabilità della covariata x_{i2}, aumenta la proporzione di effetto rilevato nei 10.000 campioni simulati per ogni valore di σ. I risultati della simulazione confermano l'ipotesi formulata precedentemente, poiché il campo di variabilità ristretto della covariata sembra essere la causa della "falsa" rilevazione dell'effetto. In tutte le simulazioni, la statistica di Wald mantiene la monotonicità e il confronto con i test log-rapporto di verosimiglianza e di Rao porta alle medesime conclusioni inferenziali. Si può dunque affermare che il metodo implementato nella funzione `hdeff` non sia completamente affidabile in situazioni come quella descritta

in questa relazione. Questo ultimo esempio mostra che non è presente un comportamento anomalo della statistica di Wald. Per quanto riguarda la classificazione dell'effetto, negli studi di simulazione, oltre al problema della severità "non determinata", rilevata in un considerevole numero di casi, ogniqualvolta l'effetto viene rilevato, si sottolinea che l'unico livello di severità osservato è "forte"; pertanto, non è da considerarsi affidabile il livello di gravità dell'effetto restituito dalla funzione `hdeffsev`. Auspicabilmente, futuri aggiornamenti delle funzioni in `R` per la trattazione dell'effetto di Hauck-Donner risolveranno questi problemi emersi.

4.3 Conclusioni

Nella relazione è stato presentato l'effetto di Hauck-Donner nel test basato sulla statistica di Wald, spiegando in che modo l'anomalia può presentarsi e indurre a conclusioni inferenziali errate. Sono stati presentati metodi di diagnostica, di classificazione e di riduzione dell'effetto nell'ambito dei GLM e VGLM tratti da Yee (2022). Lo studio di simulazione, il cui obiettivo era di testare i metodi introdotti, ha portato a diversi risultati meritevoli di attenzione: in primo luogo si è constatata la potenziale inadeguatezza della diagnostica dell'effetto, in quanto, nell'esempio trattato, l'anomalia rilevata non portava a risultati differenti rispetto ad altri test immuni all'effetto, diversamente da quanto si poteva supporre inizialmente; in secondo luogo è emerso che la causa dell'errata diagnostica è riconducibile al campo di variabilità della covariata corrispondente al coefficiente esaminato. Per quanto riguarda le criticità legate alla relazione, la funzione `hdeffsev`, che implementa la metodologia di classificazione dell'HDE, in un considerevole numero di casi ritorna un valore indeterminato. La funzione dovrebbe venire aggiornata nel breve periodo, come da informazione ricevuta in corrispondenza personale con l'Autore. Un altro modo per contrastare l'effetto di Hauck-Donner, nell'evenienza in cui la causa scatenante sia riconducibile ad una situazione di perfetta o quasi perfetta separazione nei dati, è quello di utilizzare stimatori con riduzione del *bias* trattati in Kosmidis et al. (2020). Tuttavia, non è possibile verificare l'effettiva riduzione dell'effetto, in quanto i metodi di diagnostica e classificazione non sono ad oggi disponibili in `R` per modelli al di fuori della classe `vglm`. Per l'implementazione di metodi che permettano lo studio sull'HDE su una più vasta gamma di classi, si rinvia a progetti futuri. In conclusione, nel caso in cui venga rilevato l'effetto di Hauck-Donner nelle stime dei coefficienti di un GLM o VGLM, si consiglia comunque di optare per altri test immuni come il test log-rapporto di verosimiglianza, o perlomeno di confrontare i valori dei due test, per evitare di incorrere in errori nelle conclusioni inferenziali.

Appendice

Yee (2022) presenta alcuni risultati per grandi campioni riguardanti due punti critici nel caso particolare in cui β_s è l'unico componente del parametro θ. Si utilizza una notazione del tipo $W_e(\theta_0;\hat{\theta})$ quando è opportuno evidenziare la dipendenza dei dati tramite $\hat{\theta}$ delle quantità pivotali approssimate introdotte nel Capitolo 1.

***Teorema* 1.** *Data l'ipotesi nulla $H_0 : \theta = \theta_0$ contro $H_1 : \theta \neq \theta_0$, quando l'informazione attesa è valutata in $\hat{\theta}$, se l'effetto di Hauck-Donner è presente allora*

1. *il rapporto tra la statistica di Wald e la statistica LRT soddisfa*

$$\frac{W_e(\theta_0;\hat{\theta})}{W(\theta_0;\hat{\theta})} < \frac{3}{5} + O_p(n^{-1}), \tag{A.1}$$

2. *il rapporto tra la statistica di Wald e la statistica di Rao soddisfa*

$$\frac{W_e(\theta_0;\hat{\theta})}{W_u(\theta_0;\hat{\theta})} < \frac{1}{4} + O_p(n^{-3/2}). \tag{A.2}$$

Dimostrazione. 1. Supponendo che l'informazione attesa sia uguale a quella osservata, si ha $i(\hat{\theta})^{-1} = [-l_{**}(\hat{\theta})]^{-1}$, l'HDE è presente se e solo se (dalla (3.5))

$$\frac{\hat{\theta}-\theta_0}{2} \cdot \frac{\frac{d}{d\theta}[-l_{**}(\hat{\theta})]^{-1}}{[-l_{**}(\hat{\theta})]^{-1}} = \frac{\hat{\theta}-\theta_0}{2}\frac{l_{***}(\hat{\theta})}{[-l_{**}(\hat{\theta})]} > 1. \tag{A.3}$$

Per lo sviluppo in serie di Taylor di $l(\theta_0)$ in un intorno di $\hat{\theta}$ si ottiene

$$\begin{aligned} W(\theta_0;\hat{\theta}) &= -l_{**}(\hat{\theta})(\hat{\theta}-\theta_0)^2 + \frac{1}{3}l_{***}(\hat{\theta})(\hat{\theta}-\theta_0)^3 + O_p(n^{-1}) \\ &= [-l_{**}(\hat{\theta})](\hat{\theta}-\theta_0)^2\left\{1 + \frac{2}{3}\cdot\frac{1}{2}\frac{l_{***}(\hat{\theta})(\hat{\theta}-\theta_0)}{[-l_{**}(\hat{\theta})]}\right\} \\ &> [-l_{**}(\hat{\theta})](\hat{\theta}-\theta_0)^2\left[1 + \frac{2}{3}\cdot 1\right] + O_p(n^{-1}) \text{ per la } (A.3) \\ &= \frac{5}{3}\cdot W_e(\theta_0;\hat{\theta}) + O_p(n^{-1}). \end{aligned}$$

La (A.1) si ricava dalla proprietà $W = O_p(1)$.

2. Sviluppando in serie di Taylor del numeratore di $r_u = l_*(\theta_0)/[-l_{**}(\hat{\theta})]^{1/2}$ si ottiene

$$\begin{aligned} r_u(\theta_0;\hat{\theta}) &= \frac{\hat{\theta}-\theta_0}{\sqrt{-l_{**}(\hat{\theta})}}\left[-l_{**}(\hat{\theta}) + \frac{1}{2}(\hat{\theta}-\theta_0)l_{***}(\hat{\theta}) + O_p(n^{-1/2})\right] \\ &= \frac{(\hat{\theta}-\theta_0)\sqrt{-l_{**}(\hat{\theta})}}{-l_{**}(\hat{\theta})}\left[-l_{**}(\hat{\theta}) + \frac{1}{2}(\hat{\theta}-\theta_0)l_{***}(\hat{\theta})\right] + O_p(n^{-3/2}) \\ &= r_e(\theta_0;\hat{\theta})\left[1 + \frac{1}{2}(\hat{\theta}-\theta_0)\frac{l_{***}(\hat{\theta})}{-l_{**}(\hat{\theta})} + O_p(n^{-3/2})\right]. \end{aligned}$$

Dato che $W_u = O_p(1)$ e dalla (3.5)

$$r'_e(\theta_0;\hat{\theta}) = \sqrt{-l_{**}(\hat{\theta})}\left[1 + \frac{\hat{\theta}-\theta_0}{2}\cdot\frac{l_{***}(\hat{\theta})}{l_{**}(\hat{\theta})}\right], \tag{A.4}$$

allora

$$\frac{r_u(\theta_0;\hat{\theta})}{r_e(\theta_0;\hat{\theta})} = 1 + 1 - \frac{r'_e}{\sqrt{-l_{**}(\hat{\theta})}} + O_p(n^{-3/2}) > 2 + O_p(n^{-3/2})$$

quando l'effetto è presente. La (A.2) è ottenuta elevando al quadrato il reciproco dell'ultima espressione. □

In conclusione, dalla (A.3) si può definire un vincolo da applicare oltre alle usuali condizioni di regolarità in modo tale da escludere preventivamente il sottoinsieme dello spazio parametrico da cui ha origine l'HDE, sotto H_0:

$$\Theta_* = \left\{\theta : \frac{\theta-\theta_0}{-2}\cdot\frac{l_{***}(\theta)}{l_{**}(\theta)} < 1\right\}. \tag{A.5}$$

Per ulteriori proprietà e considerazioni riguardanti l'effetto di Hauck-Donner per i VGLM con dimensione $d = 1$, si rimanda a Yee (2022).

Codice R

```
library(VGAM)
library("MASS")
library("brglm2")
#creazione dataset utilizzato in Kosmidis et al.(2020)
bwt <- with(birthwt, {
  age <- age
  racewhite <- ifelse(race==1,1,0)
  smoke <- smoke
  ptl <- ifelse(ptl>0,1,0)
  ptd <- factor(ptl > 0)
  ht <- ht
  loglwt <- log(lwt)
  data.frame(normwt = 1-low, age, racewhite, smoke, ptl,ht,
 loglwt,ftv)
})
bwt <- subset(bwt, subset = (ftv==0), select= -c(ftv))

## adattamento modello (4.1)
bwt_ml <- vglm(normwt ~ ., family = binomialff, data = bwt)
beta <- coef(bwt_ml)
summary(bwt_ml)
summary(bwt_ml, hde.NA = F)


# Analisi della statistica di Wald tabella 4.1
hdstatn <- hdeff(bwt_ml, deriv = 2, se = TRUE)
cfit <- coef(summary(bwt_ml, hd.NA = FALSE))
ans  <- cbind(cfit, hdstatn)
rownames(ans) <- NULL
round(ans, digits = 3)

# Confronto tra le statistiche:
summary(bwt_ml, lrt0 = TRUE, score0 = TRUE, wald0 = TRUE)

# funzioni per la trattazione dell'HDE
hdeff(bwt_ml)                                  #diagnostica HDE
deriv = hdeff.vglm(bwt_ml, derivative = 2)
```

```
hdeffsev(sort(beta),                          #severita' HDE
wald.stat(bwt_ml, orig.SE = F, omit1s = FALSE),
dy = deriv[,1],
ddy = deriv[,2],
allofit = T)

#studio di simulazione pag.27
set.seed(123)
Nsim = 10000
simulazione = simulate(bwt_ml, nsim = Nsim)

#creazione matrici per salvare i risultati
ml <- ml_se <- HDE_detection <- HDE_severity <-
Wald <- Pval <-  matrix(NA, nrow = Nsim, ncol = length(beta))
br <- br_se <- br_Wald <- br_Pval <-matrix(NA, nrow = Nsim,
 ncol = length(beta))

beta1.DERIV <-beta2.DERIV <-beta3.DERIV <-beta4.DERIV <-beta5.
 DERIV <-
beta6.DERIV <-beta7.DERIV <- matrix(NA,nrow = Nsim, ncol = 2)
SR_LRT <- LRT <- pvalueLRT <- Rao <- Raopval <-
HDEfreeWALD <- HDEfreeWALDpval <- matrix(NA, nrow = Nsim ,
 ncol = 6)

colnames(ml) <- colnames(ml_se) <- colnames(HDE_detection) <-
 colnames(HDE_severity) <-
colnames(Wald) <- colnames(Pval) <- c("Intercetta", "Beta2", "
 Beta3",
"Beta4", "Beta5", "Beta6", "Beta7")
colnames(SR_LRT) <- colnames(LRT) <- colnames(pvalueLRT) <-
colnames(Rao) <-  colnames(HDEfreeWALD) <- c("Beta2", "Beta3",
"Beta4", "Beta5", "Beta6", "Beta7")

#simulazione
for (i in 1:Nsim){
  current_data <- within(bwt, { normwt <- simulazione[[i]] })
  if(i%%100 == 0) print(i)
  ml_fit <- update(bwt_ml, data = current_data)
```

```
70    glm.fit <- glm(normwt ~ ., family = binomial, data = current
      _data)
71    br_fit <- update(glm.fit, method = "brglmFit", type = "AS_
      mean", data = current_data)
72
73    current_beta <- coef(ml_fit)
74    current_deriv <- hdeff.vglm(ml_fit, derivative = 2)
75    sum_ml <- summary(ml_fit)
76    ml[i,] <- sum_ml@coef3[,1]        #stime dei coefficenti
77    ml_se[i,] <- sum_ml@coef3[,2]     #standard error
78    Wald[i,] <- sum_ml@coef3[,3]      #statistica Wald osservata
79    Pval[i,] <- sum_ml@coef3[,4]      #p-value
80    #derivate prime e seconde della statistica di Wald
81    #per ogni coefficiente
82    beta1.DERIV[i,] <- current_deriv[1,]
83    beta2.DERIV[i,] <- current_deriv[2,]
84    beta3.DERIV[i,] <- current_deriv[3,]
85    beta4.DERIV[i,] <- current_deriv[4,]
86    beta5.DERIV[i,] <- current_deriv[5,]
87    beta6.DERIV[i,] <- current_deriv[6,]
88    beta7.DERIV[i,] <- current_deriv[7,]
89
90    sum_br <- summary(br_fit)
91    br[i,] <- sum_br$coef[,1]
92    br_se[i,] <- sum_br$coef[,2]
93    br_Wald[i,] <- sum_br$coef[,3]
94    br_Pval[i,] <- sum_br$coef[,4]
95
96    #statistica r:
97    SR_LRT[i,] <- lrt.stat(ml_fit, all.out = TRUE)$lrt.stat
98    #statistica W:
99    LRT[i,] <- lrt.stat(ml_fit, all.out = TRUE)$Lrt.stat2
100   pvalueLRT[i,] <-  lrt.stat(ml_fit, all.out = TRUE)$pvalues
101   HDEfreeWALD[i,] <- wald.stat(ml_fit, orig.SE = F)       #re*
102   HDEfreeWALDpval[i,] <- summary(ml_fit, lrt0 = TRUE, score0 =
      TRUE, wald0 = TRUE)@coef4wald0[,4]
103   Rao[i,] <- summary(ml_fit, lrt0 = TRUE, score0 = TRUE, wald0
      = TRUE)@coef4score0[,3] #statistica di Rao
```

```
104     Raopval[i,] <- summary(ml_fit, lrt0 = TRUE, score0 = TRUE,
    wald0 = TRUE)@coef4score0[,4]
105     #HDE
106     HDE_detection[i,] <- hdeff(ml_fit)
107     HDE_severity[i,] <- hdeffsev(sort(current_beta),
108     wald.stat(ml_fit, orig.SE <- TRUE, omit1s <- FALSE),
109     dy <- current_deriv[,1],
110     ddy <- current_deriv[,2])
111
112   }
113   #analisi dell'effetto sui coefficienti nelle simulazioni
114   detection_vector <- rep(NA, 7)
115   for (i in 1:length(detection_vector)){
116     detection_vector[i] <- sum(HDE_detection[,i])
117   }
118   prop.detection_vector <- detection_vector/Nsim
119   names(prop.detection_vector) <- c("Beta1", "Beta2", "Beta3",
120   "Beta4", "Beta5", "Beta6", "Beta7")
121   prop.detection_vector
122
123   table(HDE_severity[,7])
124
125   #esempio di grafico
126   library(ggplot2)
127   library(tidyverse)
128
129   grafico <- cbind(ml[,7],Wald[,7])
130   ggplot(as_tibble(grafico[,1:2]), aes(x = grafico[,1],y =
    grafico[,2])) +
131   geom_point(aes(col = HDE_severity[,7]))+
132   geom_smooth(method = "auto", se = FALSE, col = "grey")+
133   labs(x = expression(hat(beta)[7]), y = expression(r[eP](hat(
    beta)[7])),color = "Severita")+
134   scale_color_manual(values = c("None" = "#191970" , "Strong" =
    "#d62728", "Undetermined" = "springgreen"),
135   labels = c("Nessuna", "Forte", "Non determinata")) +
136   theme(legend.position = "top",
137   legend.title = element_text(size = 15, face = "italic"),
138   legend.text = element_text(size = 13),
```

```
  legend.background = element_rect(fill = "white"),
  legend.key = element_rect(fill = "white"),
  axis.title.x = element_text(size = 18),
  axis.title.y = element_text(size = 18) )
```

Codice R II parte

```
set.seed(33)
n <- 100
Nsim <- 10000
y <- list(data = NA, dim = c(Nsim,n))
x1 <- list(data = NA, dim = c(Nsim,n))

#generazione casuale di y e x1
for (i in 1:Nsim){
  y[[i]] <- rbinom(n, size = 1, prob = 0.5)
  x1[[i]] <- rnorm(n, mean = 5, sd = 0.05)
}
#creazione matrici per salvare i risultati delle simulazioni
mle <- mle_se <- HDE_det <- HDE_sev <-
Waldstat <- PvalWald <-  matrix(NA, nrow = Nsim, ncol =  2)

#adattamento dei modelli e salvataggio dei dati
for(i in 1:Nsim){
  modello <- vglm(y[[i]] ~ x1[[i]],
  family = binomialff)
  current_beta <- coef(modello)
  current_deriv <- hdeff.vglm(modello, derivative = 2)
  sum_modello <- summary(modello)
  mle[i,] <- sum_modello@coef3[,1]            #stime dei
 coefficenti
  mle_se[i,] <- sum_modello@coef3[,2]         #standard error
  Waldstat[i,] <- sum_modello@coef3[,3]       #statistica Wald
  osservata
  PvalWald[i,] <- sum_modello@coef3[,4]       #p-value

  HDE_det[i,] <- hdeff(modello)               #diagnostica HDE
  HDE_sev[i,] <- hdeffsev(sort(current_beta), #severita' HDE
```

```
    wald.stat(modello, orig.SE <- TRUE, omit1s <- FALSE),
    dy <- current_deriv[,1],
    ddy <- current_deriv[,2])
  }
  detection_vector <- rep(NA, 2)
  for (i in 1:length(detection_vector)){
    detection_vector[i] <- sum(HDE_det[,i])
  }
  prop.detection_vector <- detection_vector/Nsim
  names(prop.detection_vector) <- c("Beta1", "Beta2")
  prop.detection_vector
  table(HDE_sev[,2])/Nsim
```

Bibliografia

Hauck, W. W. & Donner, A. (1977). Wald's test as applied to hypoteses in logit analysis. *Journal of the American Statistical Association* **72**, 851–853.

Kosmidis, I., Kenne Pagui, E. C. & Sartori, N. (2020). Mean and median bias reduction in generalized linear models. *Statistics and Computing* **30**, 43–59.

Pace, L. & Salvan, A. (2001). *Introduzione alla Statistica - II. Inferenza, Verosimiglianza, Modelli.* Cedam, Padova.

Salvan, A., Sartori, N. & Pace, L. (2020). *Modelli Lineari Generalizzati*. Springer.

Yee, T. W. (2015). *Vector Generalized Linear and Additive Models*. Springer.

Yee, T. W. (2022). On the Hauck–Donner effect in Wald tests: detection, tipping points, and parameter space characterization. *Journal of the American Statistical Association* **117**, 1763–1774.