Università degli Studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in
Scienze Statistiche

COMBINAZIONE DI STIME TRAMITE CONFIDENCE DENSITIES NEL MODELLO DI REGRESSIONE LOGISTICA

Relatore Prof. Nicola Sartori
Dipartimento di Scienze Statistiche

Laureando Giovanni Romanò
Matricola 1223998

Anno Accademico 2020/2021

Indice

Introduzione

Nel recente passato l'applicazione all'inferenza statistica dell'approccio *divide et impera*, a volte chiamato anche *MapReduce*, è diventata sempre più diffusa, sia per la crescente disponibilità di moli di dati molto grandi, sia per la crescente possibilità di sfruttare architetture su cui potere distribuire le proprie operazioni. Esso consiste in (i) dividere i dati a disposizione in gruppi, se non sono già divisi, (ii) ottenere una stima in ciascun gruppo e (iii) combinare le stime per ottenere una stima finale.

Utilizzare metodi di tipo *MapReduce* comporta due scelte: la funzione *Map*, che viene eseguita in ciascuno dei gruppi in cui vengono divisi i dati, detti anche fonti, e la funzione *Reduce*, che combina gli output delle funzioni *Map*. La scelta della funzione *Reduce* è delicata, poiché non è garantito che si ottenga una buona stima combinando un insieme di buone stime. Zhou & Song (2017) propongono di utilizzare uno strumento chiamato *confidence distribution* per combinare stime ottenute come radici di equazioni di stima. Questo strumento risulta particolarmente adatto per questo compito poiché esso può essere costruito senza la necessità che dati relativi a ciascuna unità siano noti, ma è sufficiente che ciascuna fonte fornisca solo delle statistiche riassuntive. Inoltre, questo metodo, a differenza di molti proposti in passato, combina solo indirettamente le stime, proprio attraverso la combinazione di *confidence distribution*.

In questo lavoro di tesi si propone di adattare il lavoro di Zhou & Song (2017) alla regressione logistica binaria, utilizzando come metodo di stima nel passo *Map* la massimizzazione della verosimiglianza corretta proposta in Firth (1993). Tale correzione permette di ridurre la distorsione delle stime e garantisce l'esistenza di stime finite anche in casi in cui la stima di massima verosimiglianza è infinita. La combinazione delle stime viene effettuata attraverso *confidence distribution*, costruite sfruttando la normalità asintotica sia dell'equazione di stima che dello stimatore da questa ottenuto.

Nel primo capitolo viene presentato il concetto di *confidence distribution*, in particolare la definizione di questo strumento, come viene utilizzato per fare inferenza e come è

stato sfruttato in Zhou & Song (2017) per combinare un insieme di stime ottenute sulla base di diversi gruppi di dati.

Nel secondo capitolo viene descritto l'utilizzo della correzione di Firth (1993) nel modello di regressione logistica e viene mostrato come questo permetta di ridurre la distorsione.

Successivamente, nel terzo capitolo vengono mostrate alcune proposte di combinazione delle stime dei parametri di un modello di regressione logistica ottenute massimizzando la verosimiglianza e la verosimiglianza corretta di Firth (1993).

Infine, nel quarto capitolo le procedure definite nel capitolo precedente sono applicate a dati della piattaforma Spotify. Questa analisi permette di evidenziare i punti forti dei metodi di combinazione presentati in questo lavoro, ma anche di individuare possibili modifiche che potrebbero migliorarli o estenderli a situazioni diverse da quelle considerate in questa tesi.

Capitolo 1

Confidence distribution

Questo lavoro di tesi viene sviluppato assumendo per i dati $\boldsymbol{y}$ un modello statistico parametrico, definito (si veda ad esempio Pace & Salvan, 1997, Paragrafo 1.4) come una famiglia di distribuzioni di probabilità $\mathcal{F}$ i cui elementi possono essere indicizzati da un numero finito $p \geq 1$ di parametri reali, cioè

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

dove $y \in \mathcal{Y}$ è lo spazio campionario e Θ è detto spazio parametrico. Si assume inoltre che tale modello sia correttamente specificato, cioè che esiste un valore $\boldsymbol{\theta}_0 \in \Theta$ tale che $p(\boldsymbol{y}; \boldsymbol{\theta}_0)$ sia la vera distribuzione di probabilità.

1.1 Inferenza basata su *confidence distribution*

1.1.1 Definizione

Si riporta la definizione di *confidence distribution* data in Singh et al. (2007), nella quale $\boldsymbol{y} = [y_1, \ldots, y_n]$ e il parametro di interesse θ si assume essere scalare, cioè con $p = 1$.

Definizione 1.1.1 (*Confidence distribution*)**.** Una funzione $H(\theta, \boldsymbol{y}) : \mathcal{Y} \times \Theta \to [0, 1]$ è detta *confidence distribution* (CD) per un parametro scalare θ se (i) per ogni campione $\boldsymbol{y} \in \mathcal{Y}$, $H(\theta, \boldsymbol{y})$ è una funzione di ripartizione continua nello spazio parametrico Θ; (ii) valutata nel vero valore del parametro $\theta = \theta_0$, vista come funzione di $\boldsymbol{y}$, $H(\theta_0, \boldsymbol{y})$ si distribuisce come una variabile casuale $U(0, 1)$.

Se la funzione $H(\theta_0, \boldsymbol{y})$ è solo asintoticamente distribuita come una variabile casuale $U(0, 1)$, allora essa è una *confidence distribution* asintotica (aCD). Da qui in avanti si fa

riferimento anche ad aCD con il termine *confidence distribution*. Si riporta ora anche la definizione di *confidence density*, strettamente legato a quello di *confidence distribution*.

Definizione 1.1.2 (*Confidence density*)**.** Sia $H(\theta, \boldsymbol{y}) : \mathcal{Y} \times \Theta \to [0,1]$ una *confidence distribution* per θ, allora, se esiste, $h(\theta, \boldsymbol{y}) = \frac{d}{d\theta} H(\theta, \boldsymbol{y})$ è detta *confidence density*.

1.1.2 Costruzione di una CD

È possibile costruire una *confidence distribution* a partire da una funzione del campione e del parametro, sulla quale si pongono alcune restrizioni, se è nota la sua vera funzione di ripartizione.

Sia $t(\boldsymbol{y}, \theta)$ una quantità pivotale, che non contiene altri parametri ignoti oltre θ e, per ogni $\boldsymbol{y} \in \mathcal{Y}$, è una funzione di θ continua e monotona; sia $G(\cdot)$ la sua vera funzione di ripartizione. Allora una CD per θ è definita come

$$H(\boldsymbol{y}, \theta) = \begin{cases} G(t(\boldsymbol{y}, \theta)), & \text{se } t(\cdot) \text{ è crescente in } \theta \\ 1 - G(t(\boldsymbol{y}, \theta)), & \text{se } t(\cdot) \text{ è decrescente in } \theta. \end{cases}$$

Se $G(\cdot)$ è ignota, allora è possibile usare una sua approssimazione, come la distribuzione limite per dimensione campionaria divergente, oppure la distribuzione bootstrap.

Si mostra ora un esempio di come sia possibile costruire una *confidence distribution* per il valore atteso di una variabile casuale normale.

Esempio 1.1.1 (Valore atteso di una normale)**.** Sia $(y_1, \ldots, y_n)$ un campione casuale semplice da una variabile casuale $N(\mu, \sigma^2)$, con $\sigma = 2$ noto e μ ignoto parametro di interesse. Sia $\bar{Y}$ lo stimatore media campionaria. Allora $t(\boldsymbol{y}, \mu) = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$ è una quantità pivotale monotona decrescente in μ e distribuita come una variabile casuale Z, normale standard. Allora

$$H(\mu, \boldsymbol{y}) = 1 - P(Z \leq t(\boldsymbol{y}, \mu))$$

è una *confidence distribution* per μ per $\boldsymbol{y}$ fissato, rappresentata nella Figura 1.1 per un campione di 500 osservazioni con $\bar{y} = 3.01$. Se σ fosse ignoto, allora si potrebbe definire $t^*(\boldsymbol{y}, \mu) = \frac{\bar{Y} - \mu}{S/\sqrt{n}}$, dove S è lo stimatore deviazione standard corretta campionaria, che è una quantità pivotale decrescente in μ ed è distribuita come una variabile casuale T_{n-1}, t di Student con $n-1$ gradi di libertà. Allora,

$$H(\mu, \boldsymbol{y}) = 1 - P(T_{n-1} \leq t^*(\boldsymbol{y}, \mu))$$

è una *confidence distribution* per μ per $\boldsymbol{y}$ fissato e per qualsiasi valore di σ^2.

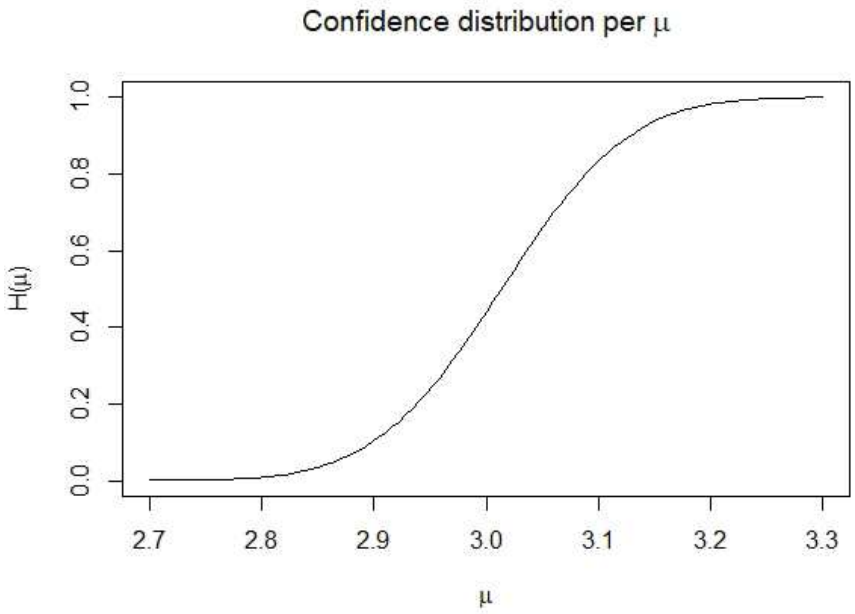

FIGURA 1.1: Grafico della *confidence distribution* per il valore atteso di una normale con varianza nota.

1.1.3 Stima intervallare

Lo strumento inferenziale più immediato da costruire sulla base di una *confidence distribution* è l'intervallo di confidenza. Infatti il punto (ii) della Definizione 1.1.1 implica che

$$P_{\theta_0}\{H^{-1}(\alpha/2, \boldsymbol{Y}) \leq \theta_0 \leq H^{-1}(1-\alpha/2, \boldsymbol{Y})\} = 1-\alpha$$

con $\alpha \in (0,1)$. Perciò un intervallo di confidenza con livello di confidenza $1-\alpha$ è dato da

$$\left(H^{-1}(\alpha/2, \boldsymbol{y}), H^{-1}(1-\alpha/2, \boldsymbol{y})\right).$$

Considerando nuovamente le ipotesi dell'Esempio 1.1.1, è possibile osservare in Figura 1.2 sull'asse delle ascisse l'intervallo di confidenza con livello 0.9 per il valore atteso della variabile casuale normale, corrispondente ai livelli 0.05 e 0.95 sulle ordinate, costruito sfruttando la *confidence distribution* costruita nell'esempio precedente.

1.1.4 Stima puntuale

La stima puntuale di un parametro sulla base di una *confidence distribution*, come spesso accade anche in altri contesti, può essere la mediana della CD, $M = H^{-1}(0.5, \boldsymbol{y})$, la media $\bar{\theta} = \int t\, h(t, \boldsymbol{y})\, dt$ o la moda $\hat{\theta} = \arg\max_{\theta} h(\theta, \boldsymbol{y})$ della *confidence density*. Sotto opportune condizioni, tutti e tre questi stimatori sono consistenti. Vengono riportate

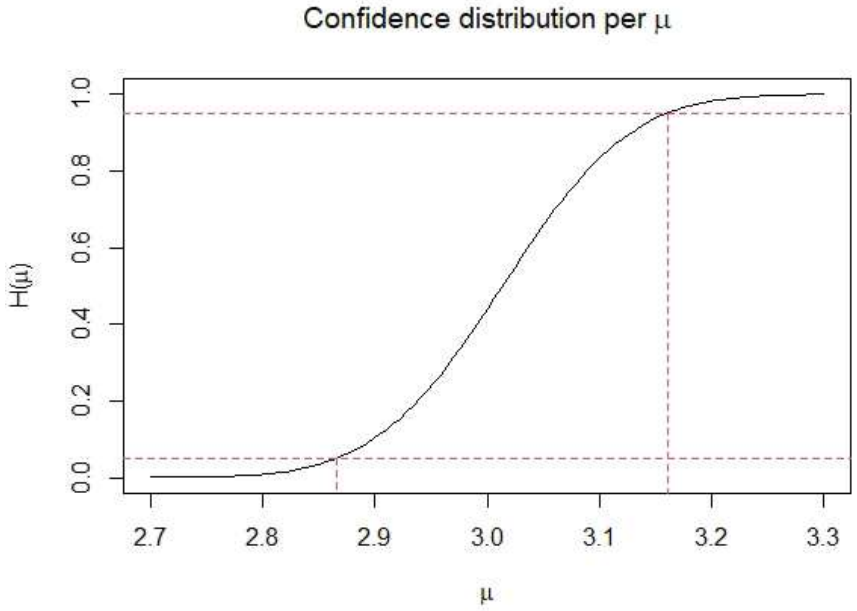

FIGURA 1.2: Grafico della *confidence distribution* per la media di una normale con varianza nota e relativo intervallo di confidenza con livello 0.9.

sotto forma di teoremi le ipotesi necessarie per la consistenza di ciascuno di questi tre stimatori, che vengono descritte in Singh et al. (2007). Nel seguito con il simbolo $\xrightarrow{p}$ si indica la convergenza in probabilità, al valore alla destra del simbolo, della successione di variabili casuali alla sinistra.

Teorema 1.1.1 (Consistenza della mediana)**.** Se per ogni ϵ e per ogni θ_0 fissato, $0 < \epsilon < \frac{1}{2}$, $H^{-1}(1-\epsilon, \boldsymbol{y}) - H^{-1}(\epsilon, \boldsymbol{y}) \xrightarrow{p} 0$ per $n \to \infty$, allora $M \xrightarrow{p} \theta_0$, per $n \to \infty$.

La condizione posta per la consistenza della mediana coincide con la richiesta che che la CD si concentri intorno a θ_0 quando n tende ad infinito.

Teorema 1.1.2 (Consistenza della media)**.** Se sono rispettate le ipotesi del Teorema 1.1.1 e inoltre il momento secondo $t = \int t^2 h(t, \boldsymbol{y})\, dt$ è limitato in probabilità, allora $\bar{\theta} \xrightarrow{p} \theta_0$, per $n \to \infty$.

Infine, la condizione per la consistenza della moda è data dal seguente teorema.

Teorema 1.1.3 (Consistenza della moda)**.** Sia ϵ_{inf} il valore definito come $\epsilon_{inf} = \inf_{0<\epsilon\leq 1/2} \Big\{\epsilon : \hat{\theta} \notin [H^{-1}(\epsilon, \boldsymbol{y}), H^{-1}(1-\epsilon, \boldsymbol{y})]\Big\}$. Se sono rispettate le ipotesi del Teorema 1.1.1 e inoltre esiste un $\epsilon^* > 0$ tale che $P(\epsilon_{inf} \geq \epsilon^*) \to 1$, allora $\hat{\theta} \xrightarrow{p} \theta_0$, per $n \to \infty$.

La condizione aggiuntiva, rispetto al Teorema 1.1.1, per la consistenza della media equivale a chiedere che, sotto θ_0, esista un valore ϵ_{iinf} per cui sia quasi certo che $\hat{\theta}$ non sia nelle code della CD con probabilità maggiore o uguale ad ϵ^*.

Inoltre, è diretta conseguenza della definizione di *confidence distribution* e della sua mediana, che quest'ultima, non solo è uno stimatore consistente, ma anche uno stimatore non distorto in mediana. Questo risultato e la maggior varietà di situazioni sotto la quale la mediana è consistente fanno propendere per la scelta della mediana come stimatore puntuale.

Nel prossimo paragrafo il concetto di *confidence distribution* viene adattato ad un parametro multidimensionale, cioè con $p > 1$, come una funzione dei dati e del parametro la cui distribuzione sotto il vero valore del parametro è un prodotto di variabili casuali uniformi, una per ciascuna componente del parametro. Nel contesto multidimensionale il concetto di mediana perde di significato e perciò la stima del parametro si otterrà come moda della *confidence density.*

1.2 Inferenza distribuita tramite *confidence distribution*

Nel precedente paragrafo è stata introdotta una *confidence distribution* e come può essere utilizzata per inferire sul valore di un parametro scalare. Nel seguito del capitolo viene presentato l'utilizzo di *confidence distribution*, in particolare della *confidence density* associata, per combinare stime relative ad un parametro, ottenute su insiemi di dati diversi. Zhou & Song (2017) propongono due diverse *confidence density* da costruire in ciascuna fonte, una funzione dell'equazione di stima e l'altra funzione dallo stimatore. In entrambi i casi ciò fatto viene sfruttando la normalità asintotica della quantità considerata, e quindi sono aCD, secondo la definizione precdente.

1.2.1 Costruzione di *confidence distribution*

Prima di procedere con la definizione delle procedure di combinazione di stime distribuite, in questo paragrafo si mostra come è possibile costruire *confidence distribution* a partire da un unico insieme di dati.

Si consideri un modello parametrico $\mathcal{F} = \{p(\boldsymbol{y}, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p\}$, con p fissato, e n campioni indipendenti $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$. Il parametro $\boldsymbol{\theta}$ può essere stimato dalla soluzione della seguente equazione di stima

$$\psi_{tot}(\boldsymbol{Y}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \psi(\boldsymbol{Y}_i; \boldsymbol{\theta}) = \mathbf{0},$$

che viene indicata con $\hat{\boldsymbol{\theta}}$.

È necessario definire le seguenti quantità relative all'equazione di stima: la matrice di variabilità

$$\boldsymbol{v}(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}}\{\psi_{tot}(\boldsymbol{Y};\boldsymbol{\theta})\},$$

la matrice di sensibilità

$$\boldsymbol{s}(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\left\{-\frac{\partial\psi_{tot}(\boldsymbol{Y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right\},$$

e la matrice di informazione di Godambe

$$\boldsymbol{g}(\boldsymbol{\theta}) = \boldsymbol{s}(\boldsymbol{\theta})^\top \boldsymbol{v}(\boldsymbol{\theta})^{-1}\boldsymbol{s}(\boldsymbol{\theta}).$$

È possibile stimare ciascuna di queste matrici tramite le loro controparti campionarie

$$\boldsymbol{V}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\left(\psi(\boldsymbol{y}_i;\boldsymbol{\theta})\right)^2, \qquad \boldsymbol{S}(\boldsymbol{\theta}) = -\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\psi(\boldsymbol{y}_i;\boldsymbol{\theta}),$$
$$\boldsymbol{G}(\boldsymbol{\theta}) = \boldsymbol{S}(\boldsymbol{\theta})^\top \boldsymbol{V}(\boldsymbol{\theta})^{-1}\boldsymbol{S}(\boldsymbol{\theta}).$$

Sotto opportune condizioni di regolarità (si veda ad esempio Song, 2007, Capitolo 3), tra cui la non distorsione dell'equazione di stima, sia $\psi_{tot}(\boldsymbol{Y};\boldsymbol{\theta})$ che $\hat{\boldsymbol{\theta}}$ convergono in distribuzione ad una variabile casuale normale. In particolare, si ha

$$\psi_{tot}(\boldsymbol{Y};\boldsymbol{\theta}_0) \xrightarrow{d} N_p(\boldsymbol{0}, \boldsymbol{v}(\boldsymbol{\theta_0}))$$
$$\hat{\boldsymbol{\theta}} \xrightarrow{d} N_p(\boldsymbol{\theta}_0, \boldsymbol{g}(\boldsymbol{\theta}_0)^{-1}),$$

dove $\xrightarrow{d}$ denota convergenza in distribuzione e $\boldsymbol{\theta}_0$ il vero valore del parametro.
Sfruttando questi risultati Zhou & Song (2017) definiscono due *confidence distribution*, una detta di tipo Rao, basata sull'equazione di stima $\psi_{tot}(\boldsymbol{y},\boldsymbol{\theta})$,

$$H_R(\boldsymbol{\theta}_0) = \Phi_p\Big(\boldsymbol{V}(\hat{\boldsymbol{\theta}})^{-1/2}\psi_{tot}(\boldsymbol{y};\boldsymbol{\theta}_0)\Big),$$

dove $\Phi_p(\cdot)$ è la funzione di ripartizione della variabile casuale normale standard p-variata, e una detta di tipo Wald, basata sullo stimatore $\hat{\boldsymbol{\theta}}$,

$$H_W(\boldsymbol{\theta}_0) = \Phi_p\Big(\boldsymbol{G}(\hat{\boldsymbol{\theta}})^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\Big),$$

dove, se $\boldsymbol{A}$ è una matrice quadrata, allora $\boldsymbol{A}^{1/2}$ è la matrice quadrata tale che $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$.

Visto che

$$\begin{aligned} P(H_R(\boldsymbol{\theta}_0) < x) &= P\left\{\Phi_p\Big(\boldsymbol{V}(\hat{\boldsymbol{\theta}})^{-1/2}\psi_{tot}(\boldsymbol{Y};\boldsymbol{\theta}_0)\Big) < x\right\} \\ &\dot{=} P\left\{\Phi_p(Z_1,\ldots,Z_p) < x\right\} \\ &\dot{=} P\left\{\Phi(Z_1)\times\ldots\times\Phi(Z_p) < x\right\} \\ &\dot{=} P\left\{U_1\times\ldots\times U_p < x\right\}, \end{aligned}$$

dove $Z_1,\ldots,Z_p \overset{iid}{\sim} N(0,1)$ e $U_1,\ldots,U_p \overset{iid}{\sim} U(0,1)$, si può affermare che $H_R(\boldsymbol{\theta}_0)$ si distribuisce come il prodotto di variabili casuali uniformi $U_1U_2\cdots U_p$. Si può mostrare analogamente che ciò vale anche per $H_W(\boldsymbol{\theta}_0)$.

1.2.2 Stime combinate alla Rao e alla Wald

Dopo aver visto come è possibile costruire su un unico insieme di dati una, *confidence distribution* sia di tipo Rao che di tipo Wald, in questo paragrafo viene presentato l'utilizzo di entrambi i tipi di *confidence distribution* per combinare stime di un comune parametro $\boldsymbol{\theta}$ ottenute in diversi insiemi di dati, così come fatto in Zhou & Song (2017).

Si consideri una partizione del campione $\boldsymbol{y}$ in K sottoinsiemi $\boldsymbol{y}^{(1)},\ldots,\boldsymbol{y}^{(K)}$, di dimensioni rispettivamente $n_1,\ldots,n_K$. Nel caso in cui il modello assunto ipotizzi una struttura di regressione che implica la presenza di una matrice del disegno $\boldsymbol{X}$, anch'essa viene partizionata allo stesso modo. Sia $\psi_k(\boldsymbol{y}^{(k)};\boldsymbol{\theta})$ la funzione di stima relativa al k-esimo sottoinsieme e $\hat{\boldsymbol{\theta}}_k$ la soluzione della corrispondente equazione di stima, $k = 1,\ldots,K$. Inoltre, siano $\boldsymbol{V}_k(\boldsymbol{\theta})$, $\boldsymbol{S}_k(\boldsymbol{\theta})$ e $\boldsymbol{G}_k(\boldsymbol{\theta})$ rispettivamente le matrici di variabilità, sensibilità e informazione di Godambe del k-esimo sottoinsieme.

Sfruttando la normalità approssimata delle K funzioni di stima è possibile costruire K *confidence density* di tipo Rao della seguente forma

$$h_{R,k}(\boldsymbol{\theta};\hat{\boldsymbol{V}}_k) \propto \phi_p\left\{\hat{\boldsymbol{V}}_k^{-1/2}\psi_k(\boldsymbol{y}^{(k)};\boldsymbol{\theta})\right\},$$

dove $\hat{\boldsymbol{V}}_k = \boldsymbol{V}_k(\hat{\boldsymbol{\theta}}_k)$ è la matrice di variabilità del k-esimo sottoinsieme valutata nella stima locale $\hat{\boldsymbol{\theta}}_k$ e $\phi_p(\cdot)$ è la densità della variabile casuale normale standard p-variata.

Analogamente, sfruttando la normalità approssimata dei K stimatori $\hat{\boldsymbol{\theta}}_1,\ldots,\hat{\boldsymbol{\theta}}_K$ è possibile costruire K *confidence density* di tipo Wald della seguente forma

$$h_{W,k}(\boldsymbol{\theta};\hat{\boldsymbol{G}}_k) \propto \phi_p\left\{\hat{\boldsymbol{G}}_k^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\right\},$$

dove $\hat{\boldsymbol{G}}_k = \boldsymbol{G}_k(\hat{\boldsymbol{\theta}}_k)$ è la matrice di informazione di Godambe del k-esimo sottoinsieme valutata nella stima locale $\hat{\boldsymbol{\theta}}_k$, $k = 1, \ldots, K$.

In entrambi i casi, assumendo l'indipendenza tra le variabili casuali $\boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(K)}$ è possibile combinare le *confidence density* considerandone il prodotto, ottenendo in questo modo

$$h_R(\boldsymbol{\theta}) = \prod_{k=1}^{K} h_{R,k}(\boldsymbol{\theta}; \hat{\boldsymbol{V}}_k),$$
$$h_W(\boldsymbol{\theta}) = \prod_{k=1}^{K} h_{W,k}(\boldsymbol{\theta}; \hat{\boldsymbol{G}}_k).$$

È possibile ottenere due stime del comune parametro $\boldsymbol{\theta}$ massimizzando le due quantità appena definite. Si hanno quindi la stima di tipo Rao

$$\hat{\boldsymbol{\theta}}_{rcd} = \arg\max_{\boldsymbol{\theta}} h_R(\boldsymbol{\theta})$$

e la stima di tipo Wald

$$\hat{\boldsymbol{\theta}}_{wcd} = \arg\max_{\boldsymbol{\theta}} h_W(\boldsymbol{\theta}).$$

1.2.3 Calcolo delle stime

Il problema di massimizzazione di $h_W(\boldsymbol{\theta})$ ha una soluzione esplicita, siccome si tratta di un problema di ottimo di una forma quadratica, e perciò il calcolo della stima $\hat{\boldsymbol{\theta}}_{wcd}$ è immediato. Infatti, siccome

$$h_W(\boldsymbol{\theta}) \propto \exp\left\{ -\frac{1}{2} \sum_{k=1}^{K} (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta})^\top \hat{\boldsymbol{G}}_k (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}) \right\},$$

la stima combinata alla Wald è

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{wcd} &= \arg\min_{\boldsymbol{\theta}} \left\{ \sum_{k=1}^{K} (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta})^\top \hat{\boldsymbol{G}}_k (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}) \right\} \\ &= \left\{ \sum_{k=1}^{K} \hat{\boldsymbol{G}}_k \right\}^{-1} \left\{ \sum_{k=1}^{K} \hat{\boldsymbol{G}}_k \hat{\boldsymbol{\theta}}_k \right\}. \end{aligned} \tag{1.1}$$

Invece, per quanto riguarda la stima combinata alla Rao, essa è

$$\hat{\boldsymbol{\theta}}_{rcd} = \arg\max_{\boldsymbol{\theta}} h_R(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{k=1}^{K} \log h_{R,k}(\boldsymbol{\theta}),$$

e perciò è data dalla radice dell'equazione di stima che pone il gradiente di $\sum_{k=1}^{K} \log h_{R,k}(\boldsymbol{\theta})$ uguale al vettore $\mathbf{0}$

$$\begin{aligned}\Psi_R(\boldsymbol{\theta}) &= \frac{\partial}{\partial \boldsymbol{\theta}} \log h_R(\boldsymbol{\theta}) \\ &= \sum_{k=1}^{K} \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ -\frac{1}{2} \left(\hat{\boldsymbol{V}}_k^{-1/2} \psi_k(\boldsymbol{y}^{(k)}; \boldsymbol{\theta}) \right)^\top \left(\hat{\boldsymbol{V}}_k^{-1/2} \psi_k(\boldsymbol{y}^{(k)}; \boldsymbol{\theta}) \right) \right\} \\ &= \sum_{k=1}^{K} \boldsymbol{S}_k(\boldsymbol{\theta})^\top \hat{\boldsymbol{V}}_k^{-1} \psi_k(\boldsymbol{y}^{(k)}; \boldsymbol{\theta}) = \mathbf{0}.\end{aligned} \tag{1.2}$$

In Zhou & Song (2017) una soluzione approssimata di (1.2) viene trovata tramite un'approssimazione con un passo Newton-Raphson attraverso cui viene aggiornata la stima alla Wald

$$\hat{\boldsymbol{\theta}}_{rcd} \approx \hat{\boldsymbol{\theta}}_{wcd} + \left\{ \sum_{k=1}^{K} \boldsymbol{S}_k(\hat{\boldsymbol{\theta}}_{wcd})^\top \hat{\boldsymbol{V}}_k^{-1} \boldsymbol{S}_k(\hat{\boldsymbol{\theta}}_{wcd}) \right\} \left\{ \sum_{k=1}^{K} \boldsymbol{S}_k(\hat{\boldsymbol{\theta}}_{wcd})^\top \hat{\boldsymbol{V}}_k^{-1} \psi_k(\boldsymbol{y}^{(k)}; \hat{\boldsymbol{\theta}}_{wcd}) \right\}.$$

È importante notare che le quantità necessarie per costruire la stima combinata alla Wald, cioè $\hat{\boldsymbol{\theta}}_k$ e $\hat{\boldsymbol{G}}_k$ per $k = 1, \ldots, K$, sono solo informazioni riassuntive, mentre la stima alla Rao necessita anche dei dati $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(K)}$.

Siccome il seguito di questo lavoro si concentra sul modello di regressione logistica, è utile osservare la struttura della stima combinata alla Wald quando le funzioni di stima $\psi_1(\boldsymbol{y}^{(1)}; \boldsymbol{\theta}), \ldots, \psi_K(\boldsymbol{y}^{(K)}; \boldsymbol{\theta})$ sono funzioni punteggio di un modello completamente specificato. In particolare, le matrici di variabilità $\boldsymbol{v}(\boldsymbol{\theta})$ e sensibilità $\boldsymbol{s}(\boldsymbol{\theta})$ coincidono tra di loro e con la matrice di informazione attesa e, siccome $\boldsymbol{g}(\boldsymbol{\theta}) = \boldsymbol{s}(\boldsymbol{\theta})\boldsymbol{v}(\boldsymbol{\theta})^{-1}\boldsymbol{s}(\boldsymbol{\theta})$, anche la matrice di informazione di Godambe è uguale alla matrice di informazione attesa. Perciò, la costruzione della stima alla Wald si semplifica in

$$\hat{\boldsymbol{\theta}}_{wcd} = \left\{ \sum_{k=1}^{K} \hat{\boldsymbol{s}}_k \right\}^{-1} \left\{ \sum_{k=1}^{K} \hat{\boldsymbol{s}}_k \hat{\boldsymbol{\theta}}_k \right\},$$

dove $\hat{\boldsymbol{s}}_k = \boldsymbol{s}_k(\hat{\boldsymbol{\theta}}_k)$.

Capitolo 2

Correzione di Firth

2.1 Stime di massima verosimiglianza infinite

Siccome il fulcro di questo lavoro è l'applicazione dell'approccio *divide et impera* al modello di regressione logistica binaria, è importante tenere in considerazione il comportamento delle stime di massima verosimiglianza in situazioni in cui il rapporto tra il numero di variabili e il numero di osservazioni cresce, fenomeno che accade se si ha a disposizione un'architettura che permette di dividere il compito di stima in un numero molto elevato di sottoinsiemi oppure se il numero di osservazioni complessivo non è molto alto.

In generale, l'esistenza di stime di massima verosimiglianza finite in un modello di regressione logistica non è sempre garantito. In particolare ciò non accade in situazioni di perfetta separazione o quasi-perfetta separazione, cioè quando esiste un iperpiano che individua due regioni di $\mathbb{R}^p$ (dove p è il numero di covariate) tali che in ciascuna delle due regioni si trovano osservazioni appartenenti a solo una delle due classi definite dalla risposta o al massimo alcune osservazioni sono situate esattamente sull'iperpiano (si veda Pace & Salvan, 1997, Paragrafo 6.5).

Candès & Sur (2020) individuano tre quantità fondamentali per determinare l'esistenza di stime finite quando p si assume crescere con il numero di osservazioni n: il limite κ del rapporto p/n, il segnale γ_0^2, definito come il limite per $n \to \infty$ della varianza del predittore lineare (intercetta esclusa), e il valore dell'intercetta θ_0. Si può mostrare (si veda ad esempio Candès & Sur, 2020, Teoremi 2.1 e 2.2 per modelli con e senza intercetta) che per ogni valore di (θ_0, γ_0) esiste una soglia $s(\theta_0, \gamma_0)$ tale che la probabilità che le stime di massima verosimiglianza siano finite converge a 1 se $\kappa < s(\theta_0, \gamma_0)$ e converge a 0 se $\kappa > s(\theta_0, \gamma_0)$.

In Sur & Candès (2019) vengono anche analizzati i comportamenti delle stime in campioni di dimensione finita. Dopo aver fissato la matrice del disegno con dimensione campionaria $n = 4000$, per ogni coppia di valori (κ, γ_0), con $\kappa \in (0, 0.6)$ e $\gamma_0 \in (0, 10)$, vengono simulati 50 insiemi di valori della risposta seguendo un modello logistico senza intercetta, con p che varia insieme a κ secondo $p = \kappa n$.

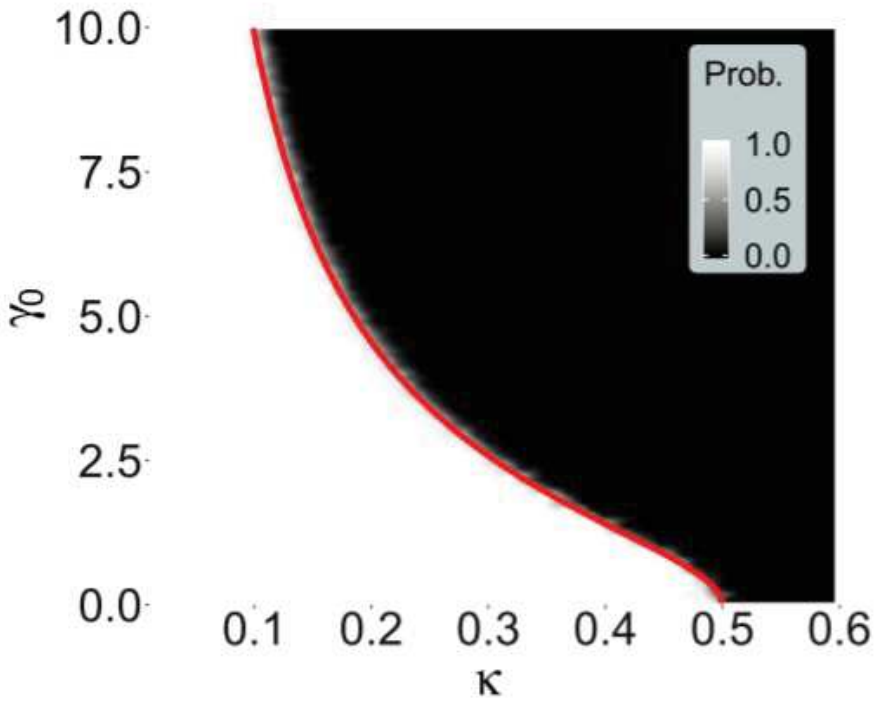

FIGURA 2.1: Probabilità dell'esistenza di stime finite al variare di κ e γ_0, in un modello con intercetta $\theta_0 = 0$ (Figura 2 (a) di Candès & Sur (2020)).

È evidente nella Figura 2.1 la presenza per ogni κ di un valore soglia di γ_0, rappresentato dalla riga rossa, che separa le situazioni in cui la regressione ammette stime finite da quelle in cui le stime sono infinite, come già dimostrato per p e n che divergono. Questo risultato motiva la scelta di un metodo di stima che garantisca sempre l'esistenza di stime finite o che porti a stime finite più frequentemente della stima di massima verosimiglianza e perciò all'interno di ciascuno dei K sottoinsiemi i parametri vengono stimati massimizzando non solo la verosimiglianza, ma anche la verosimiglianza corretta di Firth (1993). Nel prossimo paragrafo viene descritta tale correzione, in particolare nel modello di regressione logistica binaria, e nel prossimo capitolo saranno confrontate le stime ottenute combinando stime di massima verosimiglianza e quelle ottenute combinando stime di Firth.

2.2 Correzione di Firth nella regressione logistica

In un modello regolare la distorsione dell' r-sima componente dello stimatore di massima verosimiglianza $\hat{\boldsymbol{\theta}}$ per un parametro p-dimensionale $\boldsymbol{\theta}$ ha la seguente forma

$$\mathrm{E}\{(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})_r\} = \frac{1}{2} i^{rs} i^{rs} (\nu_{stu} + 2\nu_{st,u}) + O(n^{-2}), \tag{2.1}$$

dove il primo termine del membro di destra è di ordine $O(n^{-1})$. In dettaglio, indicando con $U_r(\boldsymbol{\theta})$ la componente r-esima della funzione *score* e con $U_{rs}(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta_s} U_r(\boldsymbol{\theta})$, $U_{rst}(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta_t} U_{rs}(\boldsymbol{\theta})$, dove θ_r è l'r-esima componente di $\boldsymbol{\theta}$, si ha che $\nu_{rst} = \mathrm{E}\{U_{rst}(\boldsymbol{\theta})\}$, $\nu_{rs,t} = \mathrm{E}\{U_{rs}(\boldsymbol{\theta}) U_t(\boldsymbol{\theta})\}$ e i^{rs} indica l'elemento (r, s) dell'inversa dell'informazione attesa $i(\boldsymbol{\theta})$. In (2.1) si è adottata la convenzione di somma di Einstein, per cui è sottintesa la sommatoria da 1 a p per tutti gli indici ripetuti (nella fattispecie s, t e u). Si veda Pace & Salvan (1997, Paragrafo 9.4.2) per dettagli.

Firth (1993) descrive un metodo preventivo che permette di eliminare il termine di ordine $O(n^{-1})$ della distorsione di $\hat{\boldsymbol{\theta}}$. In particolare, aggiungendo una quantità allo *score* (o equivalentemente alla log-verosimiglianza), si ha che $U_r^*(\boldsymbol{\theta}) = U_r(\boldsymbol{\theta}) + A_r(\boldsymbol{\theta})$ è la componente r-esima dello *score* modificato con una costante additiva. Siano inoltre $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\theta}}^*$ gli stimatori ottenuti come soluzioni rispettivamente di $U(\boldsymbol{\theta}) = \mathbf{0}$ e $U^*(\boldsymbol{\theta}) = \mathbf{0}$, dove $U(\boldsymbol{\theta}) = (U_1(\boldsymbol{\theta}), \ldots, U_p(\boldsymbol{\theta}))$ e $U^*(\boldsymbol{\theta}) = \big(U_1^*(\boldsymbol{\theta}), \ldots, U_p^*(\boldsymbol{\theta})\big)$. È possibile mostrare (si veda ad esempio Firth, 1993) che $A(\boldsymbol{\theta})$ deve essere tale che

$$\mathrm{E}(A_r(\boldsymbol{\theta})) = -\frac{1}{2} i^{tu} (\nu_{rtu} + 2\nu_{r,tu}) + O(n^{-1/2}), \tag{2.2}$$

per eliminare il termine $O(n^{-1})$ della distorsione di $\hat{\boldsymbol{\theta}}$. Anche se ci sono diverse forme asintoticamente equivalenti di $A(\boldsymbol{\theta})$, una possibilità è prendere il termine di ordine $O(1)$ del membro di destra della (2.2). In famiglie esponenziali con $\boldsymbol{\theta}$ parametro canonico, nelle quali inoltre $i(\boldsymbol{\theta}) = j(\boldsymbol{\theta})$, Firth (1993) dimostra che l'r-esima componente delle correzioni per lo *score* appena proposta è uguale a

$$\frac{1}{2} \frac{\partial}{\partial \theta_r} \log |i(\boldsymbol{\theta})|,$$

dove $|\boldsymbol{A}|$ indica il determinante della matrice $\boldsymbol{A}$, che equivale alla seguente modifica della verosimiglianza

$$L^*(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) |i(\boldsymbol{\theta})|^{1/2}.$$

È possibile notare che, in questo particolare caso, la stima di $\boldsymbol{\theta}$ ottenuta massimizzando tale verosimiglianza corretta coincide con la moda a posteriori di un'analisi bayesiana

in cui la priori scelta è quella di Jeffreys.

Per questo lavoro è di interesse la forma della correzione nel modello di regressione logistica binaria. Si assume che $Y_1, \ldots, Y_n$ siano indipendenti e che Y_i abbia distribuzione bernoulliana con probabilità π_i. Sia $\boldsymbol{X}$ la matrice del disegno, con i-esima riga $\boldsymbol{x}_i$. Si assume quindi $\pi_i = \left(1 + e^{-\boldsymbol{x}_i^\top \boldsymbol{\theta}}\right)^{-1}$. È possibile dimostrare che in un modello di regressione logistica (si veda ad esempio McCullagh & Nelder, 1989, Paragrafo 15.2) il termine di ordine $O(n^{-1})$ della distorsione è pari a

$$\left(\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{\xi},$$

dove $\boldsymbol{W}$ è la matrice diagonale $n \times n$ avente elementi diagonali pari a $\pi_i(1 - \pi_i)$ e il vettore $\boldsymbol{W}\boldsymbol{\xi}$ ha generico elemento $h_i(\pi_i - \frac{1}{2})$, con h_i l'i-esimo elemento diagonale della matrice

$$\boldsymbol{H} = \boldsymbol{W}^{1/2} \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{W}^{1/2}.$$

Di conseguenza la componente r-esima, $r = 1, \ldots, p$, dello *score* modificato di Firth è

$$\begin{aligned} U_r^*(\boldsymbol{\theta}) &= U_r(\boldsymbol{\theta}) + \frac{1}{2} \frac{\partial}{\partial \theta_r} \log |i(\boldsymbol{\theta})| \\ &= \sum_{i=1}^{n} \left\{ (y_i + h_i/2) - (m_i + h_i)\pi_i \right\} x_{ir}, \end{aligned}$$

dove m_i è l'indice binomiale, che nel nostro lavoro è sempre pari ad 1.

La correzione di Firth produce stime non solo meno distorte delle stime di massima verosimiglianza, ma anche sempre finite, cosa che non è sempre vero quando si massimizza la verosimiglianza, come visto in precedenza. Nel Paragrafo 2 di Kosmidis & Firth (2021) viene enunciato il seguente teorema, che garantisce la finitezza delle stime che si ottengono massimizzando una verosimiglianza modificata moltiplicandola per una potenza del determinante dell'informazione attesa.

Teorema 2.2.1. Sia

$$\ell_M(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + c \log |i(\boldsymbol{\theta})|$$

una log-verosimiglianza penalizzata, con c costante positiva fissata. Se il predittore lineare $\boldsymbol{x}_i\boldsymbol{\theta}$ è infinito per una qualsiasi osservazione, $i = 1, \ldots, n$, allora la log-verosimiglianza penalizzata $\ell_M(\boldsymbol{\theta})$ è pari a $-\infty$.

Conseguenza diretta di questo teorema è l'esistenza finita delle stime ottenute massimizzando la verosimiglianza modificata di Firth, che coincide con il caso $c = \frac{1}{2}$.

Capitolo 3

Combinazione di stime nella regressione logistica binaria

In questo lavoro di tesi si sono studiate le performance di stime calcolate combinando le stime di Firth ottenute nei sottoinsiemi in cui è stato diviso il dataset di partenza. Si è preso come punto di riferimento per i confronti stime combinate sulla base di stime di massima verosimiglianza nei sottoinsiemi e infine si sono confrontate le stime combinate con quelle che si otterrebbero sull'intero insieme dei dati, sia di massima verosimiglianza che di Firth.

In questo capitolo vengono mostrati i metodi con cui si sono combinate le stime di massima di verosimiglianza, le modifiche di tali metodi necessarie per ottenere buone performance a partire dalle stime di Firth e, infine, i confronti tra le stime ottenute in termini di distorsione e variabilità.

Sia i confronti intermedi che i confronti finali presentati in questo capitolo sono frutto di uno studio di simulazione, di cui viene ora presentata l'impostazione. Innanzitutto si è considerata una matrice del disegno $\boldsymbol{X}$ con $n = 10^4$ osservazioni i.i.d. di $p = 100$ variabili, simulate da una normale p-variata con vettore delle medie $\mathbf{0}$ e matrice di varianze e covarianze $\frac{1}{n}\boldsymbol{I}_p$, con $\boldsymbol{I}_p$ matrice identità $p \times p$. In seguito, analogamente a quanto fatto in Sur & Candès (2019), sono stati fissati metà dei parametri pari a 0 e l'altra metà sono stati simulati da una variabile causale normale con media 0 e varianza n/p; i valori dei parametri ottenuti nella simulazione sono riportati nella Tabella 3.1. Infine sono stati simulati 10^3 vettori risposta $\boldsymbol{y} = (y_1, \ldots, y_n)$, secondo il modello bernoulliano che assume $P(Y_i = 1|\boldsymbol{x_i}) = \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{\theta})}{1+\exp(\boldsymbol{x}_i^\top \boldsymbol{\theta})}$.

TABELLA 3.1: Veri valori dei parametri θ nella simulazione del modello di regressione logistica.

7.55	8.54	16.63	6.44	-7.05	6.85	-11.23	12.22	4.63	11.22
13.12	16.04	-10.28	17.43	-11.28	-19.49	-1.82	-16.60	8.31	-3.23
6.23	-8.69	-13.10	7.38	-1.84	-5.34	0.77	12.42	7.35	-4.19
-11.04	-8.22	-6.58	14.17	-7.42	7.19	12.46	-8.08	8.42	-0.51
-16.88	-4.71	-4.16	4.59	0.62	-2.90	-6.20	-1.82	-4.10	3.95
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

3.1 Stime di massima verosimiglianza

Sia $\hat{\boldsymbol{\theta}}_k$ la stima di massima verosimiglianza del k-esimo gruppo, ottenuta come soluzione dell'equazione di stima corrispondente allo *score* $U_k(\boldsymbol{\theta})$, e $i_k(\boldsymbol{\theta})$ la relativa matrice di informazione attesa.

Affinché sia possibile definire le *confidence density*, è necessario richiamare i seguenti risultati asintotici delle quantità di verosimiglianza (si veda ad esempio Pace & Salvan, 1997, Paragrafo 3.4.1).

Se la stima di massima verosimiglianza esiste finita, siccome essa è una soluzione consistente dell'equazione di verosimiglianza e il campione è composto da osservazioni indipendenti, al divergere di n valgono le seguenti convergenze per lo *score* e lo stimatore di massima verosimiglianza

$$
\begin{aligned}
i_k(\boldsymbol{\theta}_0)^{-1/2}\, U_k(\boldsymbol{\theta}_0) &\xrightarrow{d} N_p(\mathbf{0}, \boldsymbol{I}_p), \\
i_k(\boldsymbol{\theta}_0)^{1/2}\, (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0) &\xrightarrow{d} N_p(\mathbf{0}, \boldsymbol{I}_p),
\end{aligned}
$$

dove $\xrightarrow{d}$ indica convergenza in distribuzione e $\boldsymbol{\theta}_0$ è il vero valore del parametro. Nordberg (1980) e Gourieroux & Monfort (1981) individuano alcune ipotesi, riguardanti la matrice del disegno $\boldsymbol{X}$ e $\boldsymbol{X}^\top \boldsymbol{X}$, che garantiscono che al divergere della dimensione campionaria lo stimatore di massima verosimiglianza esiste, rispettivamente con probabilità 1 e quasi certamente, ed è consistente, rispettivamente in senso debole e forte, e dimostrano la sua normalità asintotica. È possibile sostituire $i_k(\boldsymbol{\theta}_0)$ con $i_k(\hat{\boldsymbol{\theta}}_k)$ e ottenere le seguenti

approssimazioni

$$i_k(\hat{\boldsymbol{\theta}}_k)^{-1/2}\, U_k(\boldsymbol{\theta}_0) \,\dot{\sim}\, N_p(\mathbf{0}, \boldsymbol{I}_p), \tag{3.1}$$

$$i_k(\hat{\boldsymbol{\theta}}_k)^{1/2}\, (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0) \,\dot{\sim}\, N_p(\mathbf{0}, \boldsymbol{I}_p). \tag{3.2}$$

3.1.1 Stime combinate alla Wald

Sulla base di (3.1) è possibile costruire una *confidence density* alla Wald, come fatto nel primo capitolo. Perciò si definisce per ogni sottoinsieme

$$h_{W,k}\,(\boldsymbol{\theta}) \propto \phi_p \left\{ i_k(\hat{\boldsymbol{\theta}}_k)^{1/2} (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}) \right\}$$

e unendo tali funzioni si ottiene la *confidence density* alla Wald complessiva

$$\begin{aligned} h_W\,(\boldsymbol{\theta}) &\propto \prod_{k=1}^{K} \phi_p \left\{ i_k(\hat{\boldsymbol{\theta}}_k)^{1/2} (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}) \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \sum_{k=1}^{K} (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta})^\top i_k(\hat{\boldsymbol{\theta}}_k) (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}) \right\}. \end{aligned}$$

La stima $\hat{\boldsymbol{\theta}}_{wcd}$ è la soluzione del problema di ottimo $\arg\max_{\boldsymbol{\theta}} h_W(\boldsymbol{\theta})$, che, come visto in (1.1), ha la seguente forma chiusa

$$\hat{\boldsymbol{\theta}}_{wcd} = \left\{ \sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}_k) \right\}^{-1} \left\{ \sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}_k) \hat{\boldsymbol{\theta}}_k \right\}.$$

3.1.2 Stime combinate alla Rao

Sulla base di (3.2) è invece possibile costruire una *confidence density* alla Rao. Analogamente a quanto fatto per la versione alla Wald, si definisce per ogni sottoinsieme

$$h_{R,k}\,(\boldsymbol{\theta}) \propto \phi_p \left\{ i_k(\hat{\boldsymbol{\theta}}_k)^{-1/2} U_k(\boldsymbol{\theta}) \right\},$$

e unendo tali funzioni si ottiene la *confidence density* alla Rao complessiva

$$\begin{aligned} h_R\,(\boldsymbol{\theta}) &\propto \prod_{k=1}^{K} \phi_p \left\{ i_k(\hat{\boldsymbol{\theta}}_k)^{-1/2} U_k(\boldsymbol{\theta}) \right\} \\ &\propto \exp\left\{ -\frac{1}{2} \sum_{k=1}^{K} U_k(\boldsymbol{\theta})^\top i_k(\hat{\boldsymbol{\theta}}_k)^{-1} U_k(\boldsymbol{\theta}) \right\}. \end{aligned}$$

Per trovare $\hat{\boldsymbol{\theta}}_{rcd}$ è necessario cercare i punti stazionari di $h_R(\boldsymbol{\theta})$. In questo caso l'equazione (1.2), la cui soluzione è la stima alla Rao, è

$$\Psi_R(\boldsymbol{\theta}) = \sum_{k=1}^{K} i_k(\boldsymbol{\theta}) i_k(\hat{\boldsymbol{\theta}}_k)^{-1} U_k(\boldsymbol{\theta}) = \mathbf{0}.$$

Anche trovare una soluzione approssimata tramite un passo Newton-Raphson è complesso, in quanto richiede il calcolo di $\frac{\partial}{\partial \boldsymbol{\theta}} i_k(\boldsymbol{\theta}) i_k(\hat{\boldsymbol{\theta}}_k)^{-1} U_k(\boldsymbol{\theta})$, che non è banale poiché anche il primo fattore $i_k(\boldsymbol{\theta})$ dipende da $\boldsymbol{\theta}$. Perciò si propone di semplificare tale compito aggiungendo un ulteriore grado di approssimazione, dato dalla sostituzione di $i_k(\boldsymbol{\theta})$ con $i_k(\hat{\boldsymbol{\theta}}_k)$. Si definisce perciò

$$\tilde{\Psi}_R(\boldsymbol{\theta}) = \sum_{k=1}^{K} U_k(\boldsymbol{\theta}),$$

che coincide quindi con lo *score* globale.

La linearizzazione di $\tilde{\Psi}_R(\boldsymbol{\theta})$ tramite il seguente sviluppo di Taylor attorno a $\hat{\boldsymbol{\theta}}_{wcd}$

$$\begin{aligned}\tilde{\Psi}_R(\boldsymbol{\theta}) &\doteq \tilde{\Psi}_R(\hat{\boldsymbol{\theta}}_{wcd}) + \left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \tilde{\Psi}_R(\boldsymbol{\theta}) \right\} \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{wcd}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{wcd}) \\ &\doteq \sum_{k=1}^{K} U_k(\hat{\boldsymbol{\theta}}_{wcd}) - \left\{ \sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}_{wcd}) \right\} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{wcd}),\end{aligned}$$

il cui secondo passaggio è garantito dall'identità di Bartlett, implica che

$$\begin{gathered}\tilde{\Psi}_R(\boldsymbol{\theta}) = \mathbf{0} \\ \Updownarrow \\ \sum_{k=1}^{K} U_k(\hat{\boldsymbol{\theta}}_{wcd}) - \left\{ \sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}_{wcd}) \right\} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{wcd}) = \mathbf{0} \\ \Updownarrow \\ \sum_{k=1}^{K} U_k(\hat{\boldsymbol{\theta}}_{wcd}) = \left\{ \sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}_{wcd}) \right\} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{wcd}) \\ \Updownarrow \\ \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{wcd} + \left\{ \sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}_{wcd}) \right\}^{-1} \left\{ \sum_{k=1}^{K} U_k(\hat{\boldsymbol{\theta}}_{wcd}) \right\}.\end{gathered}$$

Perciò la stima combinata alla Rao (approssimata) è

$$\hat{\boldsymbol{\theta}}_{rcd} = \hat{\boldsymbol{\theta}}_{wcd} + \left\{ \sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}_{wcd}) \right\}^{-1} \left\{ \sum_{k=1}^{K} U_k(\hat{\boldsymbol{\theta}}_{wcd}) \right\}.$$

3.2 Stime di Firth

Sia $\hat{\boldsymbol{\theta}}_k^*$ la stima ottenuta nel k-esimo gruppo massimizzando la verosimiglianza corretta di Firth (1993) e sia $U_k^*(\boldsymbol{\theta})$ lo *score* corrispondente alla verosimiglianza modificata.

È necessario mostrare risultati analoghi a (3.1) e (3.2) per le quantità riguardanti le stime di Firth. Lo *score* corretto è

$$U_k^*(\boldsymbol{\theta}) = U_k(\boldsymbol{\theta}) - \boldsymbol{X}_k^\top \boldsymbol{W}_k \boldsymbol{\xi}_k$$

con $\boldsymbol{X}_k$ matrice $n_k \times p$ del disegno del k-esimo gruppo, con i-esima riga $\boldsymbol{x}_i$, $\boldsymbol{W}_k$ matrice diagonale $n_k \times n_k$ avente elementi diagonali pari a $\pi_i(1-\pi_i)$ per $i = 1, \ldots, n_k$, $\pi_i = \left(1 + \exp(-\boldsymbol{x}_i^\top \boldsymbol{\theta})\right)^{-1}$ probabilità di successo per l'i-esima osservazione e $\boldsymbol{W}_k \boldsymbol{\xi}_k$ il vettore avente generico elemento $h_i^{(k)}(\pi_i - \frac{1}{2})$, dove $h_i^{(k)}$ è l'i-esimo elemento della diagonale della matrice

$$\boldsymbol{H}_k = \boldsymbol{W}_k^{1/2} \boldsymbol{X}_k \left(\boldsymbol{X}_k^\top \boldsymbol{W}_k \boldsymbol{X}_k \right)^{-1} \boldsymbol{X}_k^\top \boldsymbol{W}_k^{1/2}.$$

L'r-esima componente dello *score* modificato è pari a

$$U_{k,r}^*(\boldsymbol{\theta}) = U_{k,r}(\boldsymbol{\theta}) + \sum_{i=1}^{n_k} h_i^{(k)} \left(\frac{1}{2} - \pi_i \right) x_{ir},$$

dove $U_{k,r}(\boldsymbol{\theta})$ è l'r-esima componente dello *score* (non corretto).

È possibile mostrare che la correzione di Firth sommata allo *score* è di ordine $O(1)$ e quindi vale il seguente risultato asintotico

$$i_k(\boldsymbol{\theta}_0)^{-1/2} U_k^*(\boldsymbol{\theta}_0) \xrightarrow{d} N_p(\boldsymbol{0}, \boldsymbol{I}_p), \tag{3.3}$$

infatti, se la correzione di Firth è di ordine $O(1)$,

$$i_k(\boldsymbol{\theta}_0)^{-1/2} U_k^*(\boldsymbol{\theta}_0) = i_k(\boldsymbol{\theta}_0)^{-1/2} U_k(\boldsymbol{\theta}_0) + o_p(1)$$

e $i_k(\boldsymbol{\theta}_0)^{-1/2} U_k(\boldsymbol{\theta}_0)$ è asintoticamente normale per la (3.2).

Per confermare nel caso specifico che le componenti della correzione di Firth sono di ordine $O(1)$ è necessario mostrare che $h_i^{(k)} = O(n_k)$. Per alleggerire la notazione,

si considerino ora $\boldsymbol{X}$, $\boldsymbol{W}$, $i(\boldsymbol{\theta}_0)$, $\boldsymbol{\xi}$ definiti come $\boldsymbol{X}_k$, $\boldsymbol{W}_k$, $i_k(\boldsymbol{\theta}_0)$, $\boldsymbol{\xi}_k$ ma sull'intero insieme di dati, di dimensione campionaria n. In questo modo è possibile rimuovere i pedici/apici k che indicano l'appartenenza al k-esimo sottoinsieme. Il ragionamento che segue può essere replicato in ciascun gruppo.

L'elemento in posizione (r, s) della matrice $\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}$ è pari a $\sum_{i=1}^{n} w_i x_{ir} x_{is}$, con w_i i-esimo elemento diagonale di W, e perciò è di ordine $O(n)$. Per determinare l'ordine degli elementi di $\left(\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}\right)^{-1}$ è necessario scrivere l'inversa di una matrice tramite la matrice dei cofattori e valutare l'ordine del determinante di due matrici. Sia $\boldsymbol{Z}$ una matrice $p \times p$ invertibile, come $\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}$. Allora

$$\boldsymbol{Z}^{-1} = \frac{1}{|\boldsymbol{Z}|} (\text{cof}\boldsymbol{Z})^\top,$$

dove cof$\boldsymbol{Z}$ è la matrice dei cofattori di $\boldsymbol{Z}$, avente elemento in posizione (r, s) pari a $(-1)^{r+s} |\boldsymbol{Z}_{(-r),(-s)}|$ e $\boldsymbol{Z}_{(-r),(-s)}$ è la matrice ottenuta eliminando la r-esima riga e la s-esima colonna da $\boldsymbol{Z}$. Perciò, per valutare l'ordine degli elementi di $\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}$ è sufficiente essere in grado di valutare gli ordini dei determinanti di $\boldsymbol{Z}$ e $\boldsymbol{Z}_{(-r),(-s)}$. La formula di Leibniz per il calcolo del determinante stabilisce che

$$|\boldsymbol{Z}| = \sum_{\sigma \in S_p} \left(\text{sgn}(\sigma) \prod_{j=1}^{p} z_{j\sigma_j} \right),$$

dove z_{rs} è l'elemento in posizione (r, s) di $\boldsymbol{Z}$, S_p è l'insieme di tutte le permutazioni di $\{1, 2, \ldots, p\}$, σ è una generica permutazione di tale insieme e σ_i il suo i-esimo elemento e infine il segno di una permutazione sgn(σ) è $+1$ se σ può essere ottenuta a partire da $\{1, \ldots, p\}$ invertendo due entrate un numero pari di volte e -1 se tale numero è dispari. Nel caso in cui $\boldsymbol{Z} = \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}$, il prodotto di p suoi elementi è di ordine $O(n^p)$. Inoltre il numero di permutazioni in S_p dipende solo da p e non da n, perciò $|\boldsymbol{Z}| = O(n^p)$ e analogamente $|\boldsymbol{Z}_{(-r),(-s)}| = O(n^{p-1})$. Unendo tali informazioni si ha per il generico elemento di $i(\boldsymbol{\theta})^{-1} = (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1}$, per p fissato,

$$i^{rs} = (-1)^{r+s} \frac{|i(\boldsymbol{\theta})_{(-r),(-s)}|}{|i(\boldsymbol{\theta})|} = \frac{O(n^{p-1})}{O(n^p)} = O(n^{-1}).$$

Visto che le matrici $\boldsymbol{W^{1/2}} \boldsymbol{X}$ e $\boldsymbol{X}^\top \boldsymbol{W^{1/2}}$ hanno elementi in posizione (r, s) rispettivamente pari a $\sqrt{w_r} x_{rs}$ e pari a $\sqrt{w_s} x_{sr}$, la matrice $\boldsymbol{H}$ ha elemento in posizione (r, s) pari a

$$\sum_{t=1}^{p} \left(\sum_{u=1}^{p} \sqrt{w_r} x_{ru} i^{ut} \right) \sqrt{w_s} x_{st}.$$

Per p fissato, siccome $i^{ut} = O(n^{-1})$ e tutti gli altri elementi dell'ultima formula sono di ordine $O(1)$, $h_i = O(n^{-1})$, e di conseguenza $h_i^{(k)} = O(n_k^{-1})$. Questo risultato implica che la componente r-esima di $\boldsymbol{X}_k^\top \boldsymbol{W}_k \boldsymbol{\xi}_k$, che è uguale a $\sum_{i=1}^{n_k} x_{ir} h_i^{(k)} \left(\pi_i - \frac{1}{2}\right)$, sia di ordine $O(1)$, in quanto somma di n_k addendi, ciascuno di ordine $O(n_k^{-1})$.

Per quanto riguarda la consistenza e la normalità dello stimatore $\hat{\boldsymbol{\theta}}_k^*$, se valgono queste stesse proprietà per lo stimatore di verosimiglianza, è sufficiente notare che (si veda Kosmidis, 2007, Paragrafi 6.3, 6.4 e 6.5)

$$i(\boldsymbol{\theta}_0)^{1/2} \hat{\boldsymbol{\theta}}_k^* = i(\boldsymbol{\theta}_0)^{1/2} \{\hat{\boldsymbol{\theta}}_k + O_p(n^{-1})\} = i(\boldsymbol{\theta})^{1/2} \hat{\boldsymbol{\theta}}_k + o_p(1)$$

e perciò si può affermare che

$$i_k(\boldsymbol{\theta}_0)^{1/2} (\hat{\boldsymbol{\theta}}_k^* - \boldsymbol{\theta}_0) \xrightarrow{d} N_p(\mathbf{0}, \boldsymbol{I}_p). \tag{3.4}$$

Kosmidis (2007) afferma che è possibile dimostrare la consistenza e la normalità di $\hat{\boldsymbol{\theta}}_k^*$ anche indipendentemente dalla consistenza dello stimatore di massima verosimiglianza: la consistenza si dimostra sfruttando il risultato per gli stimatori di tipo Z (si veda ad esempio van der Vaart, 1998, Paragrafo 5.2) e la normalità grazie alla consistenza e ad uno sviluppo di Taylor dello *score* modificato.

È possibile sostituire in (3.3) e (3.4) $i_k(\boldsymbol{\theta}_0)$ con $i_k(\hat{\boldsymbol{\theta}}_k^*)$ e ottenere le seguenti approssimazioni

$$i_k(\hat{\boldsymbol{\theta}}_k^*)^{-1/2} \, U_k^*(\boldsymbol{\theta}) \dot{\sim} N_p(\mathbf{0}, \boldsymbol{I}_p), \tag{3.5}$$

$$i_k(\hat{\boldsymbol{\theta}}_k^*)^{1/2} \, (\hat{\boldsymbol{\theta}}_k^* - \boldsymbol{\theta}_0) \dot{\sim} N_p(\mathbf{0}, \boldsymbol{I}_p). \tag{3.6}$$

3.2.1 Stime combinate alla Wald

Sfruttando la (3.5) è possibile costruire una *confidence density* alla Wald, come fatto con le stime di massima verosimiglianza. In ciascuno dei K sottoinsiemi si definisce

$$h_{W,k}^*(\boldsymbol{\theta}) \propto \phi_p \left\{ i_k(\hat{\boldsymbol{\theta}}_k^*)^{1/2} (\hat{\boldsymbol{\theta}}_k^* - \boldsymbol{\theta}) \right\},$$

e unendo tali funzioni si ottiene la *confidence density* alla Wald complessiva

$$h^*_W(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \phi_p \left\{ i_k(\hat{\boldsymbol{\theta}}^*_k)^{1/2}(\hat{\boldsymbol{\theta}}^*_k - \boldsymbol{\theta}) \right\}$$
$$\propto \exp\left\{ -\frac{1}{2} \sum_{k=1}^{K} (\hat{\boldsymbol{\theta}}^*_k - \boldsymbol{\theta})^\top i_k(\hat{\boldsymbol{\theta}}^*_k)(\hat{\boldsymbol{\theta}}^*_k - \boldsymbol{\theta}) \right\}.$$

La stima $\hat{\boldsymbol{\theta}}^*_{wcd}$ è la soluzione del problema di ottimo $\arg\max_{\boldsymbol{\theta}} h^*_W$, che, come visto in precedenza, ha la seguente formulazione

$$\hat{\boldsymbol{\theta}}^*_{wcd} = \left\{ \sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}^*_k) \right\}^{-1} \left\{ \sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}^*_k)\hat{\boldsymbol{\theta}}^*_k \right\}.$$

3.2.2 Stime combinate alla Rao

Sfruttando la (3.6) è invece possibile costruire una *confidence density* alla Rao. Ancora una volta, si definisce per ogni sottoinsieme

$$h^*_{R,k}(\boldsymbol{\theta}) \propto \phi_p \left\{ i_k(\hat{\boldsymbol{\theta}}^*_k)^{-1/2} U^*_k(\boldsymbol{\theta}) \right\},$$

e unendo tali funzioni si ottiene la *confidence density* alla Rao complessiva

$$h^*_R(\boldsymbol{\theta}) \propto \prod_{k=1}^{K} \phi_p \left\{ i_k(\hat{\boldsymbol{\theta}}^*_k)^{-1/2} U^*_k(\boldsymbol{\theta}) \right\}$$
$$\propto \exp\left\{ -\frac{1}{2} \sum_{k=1}^{K} U^*_k(\boldsymbol{\theta})^\top i_k(\hat{\boldsymbol{\theta}}^*_k)^{-1} U^*_k(\boldsymbol{\theta}) \right\}.$$

Per trovare $\hat{\boldsymbol{\theta}}^*_{rcd}$ è necessario cercare i punti stazionari di $h^*_R(\boldsymbol{\theta})$. In questo caso l'equazione (1.2), la cui soluzione è la stima alla Rao, è

$$\Psi^*_R(\boldsymbol{\theta}) = \sum_{k=1}^{K} i_k(\boldsymbol{\theta}) i_k(\hat{\boldsymbol{\theta}}^*_k)^{-1} U^*_k(\boldsymbol{\theta}) = \mathbf{0}.$$

Ripetendo i passaggi fatti nel Paragrafo 3.1.2, si ottiene la versione approssimata di $\Psi^*_R(\boldsymbol{\theta})$

$$\tilde{\Psi}^*_R(\boldsymbol{\theta}) = \sum_{k=1}^{K} U^*_k(\boldsymbol{\theta}),$$

che, a differenza di quanto accaduto nel caso di stime di massima verosimiglianza, non

coincide con lo *score* corretto globale, $U^*(\boldsymbol{\theta})$. Tramite un passo Newton-Raphson derivante dallo sviluppo di Taylor attorno a $\hat{\boldsymbol{\theta}}^*_{wcd}$, in cui si approssima $\left\{\frac{\partial}{\partial \boldsymbol{\theta}}\tilde{\Psi}_R(\boldsymbol{\theta})\right\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{wcd}}$ con il suo termine dominante $\left\{\sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}^*_{wcd})\right\}$,

$$\tilde{\Psi}^*_R(\boldsymbol{\theta}) \doteq \sum_{k=1}^{K} U^*_k(\hat{\boldsymbol{\theta}}^*_{wcd}) - \left\{\sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}^*_{wcd})\right\}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^*_{wcd}),$$

si ottiene la seguente stima

$$\hat{\boldsymbol{\theta}}^*_{rcd} = \hat{\boldsymbol{\theta}}^*_{wcd} + \left\{\sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}^*_{wcd})\right\}^{-1}\left\{\sum_{k=1}^{K} U^*_k(\hat{\boldsymbol{\theta}}^*_{wcd})\right\}.$$

3.3 Confronti iniziali

In questo paragrafo si riportano alcuni confronti ottenuti tramite uno studio di simulazione, la cui impostazione è stata descritta nell'introduzione di questo capitolo. Visto l'alto numero di covariate utilizzate nel modello, risulta complicato monitorare i risultati relativi a tutti i parametri. Si ritiene che analizzare parametri sia positivi che negativi aventi valore assoluto su tre diverse scale (prossimo a 0, massimo e intermedio) possa costituire un'analisi sufficientemente completa. In particolare, i parametri tenuti in considerazione hanno i seguenti veri valori:

θ_{27}	θ_{40}	θ_1	θ_5	θ_{14}	θ_{16}
0.767	-0.508	7.553	-7.054	17.431	-19.494

Per ciascuna delle 10^3 iterazioni, $\boldsymbol{X}$ e $\boldsymbol{y}$ sono stati divisi in $K = 5, 10, 20, 40$ gruppi, rispettivamente di dimensione campionaria $2000, 1000, 500$ e 250. In ciascun gruppo sono state ottenute le stime di massima verosimiglianza e di Firth, per quest'ultime è stato usato il pacchetto `brglm2` (Kosmidis, 2020) del software R, e infine sono state combinate come mostrato nei precedenti paragrafi.

Con $K = 40$ il rapporto tra il numero di variabili, $p = 100$, e il numero di osservazioni in ciascun gruppo, $n_k = 250$, è 0.4 ed è piuttosto alto, tanto da non garantire stime di massima verosimiglianza finite in tutti i sottoinsiemi. Di seguito si presenta la Tabella 3.2, che riporta la frequenza del numero di sottogruppi in cui sono state individuate stime di massima verosimiglianza infinite. Tale controllo è stato effettuato con la funzione `detect_separation` implementata nella libreria R `detectseparation` (Kosmidis,

2021), in precedenza implementata in `brglm2` Kosmidis (2020), ma in disuso e non più mantenuta dalla versione 0.8.

TABELLA 3.2: Frequenza del numero di sottogruppi con stime di massima verosimiglianza infinite, al variare di K.

	0	1	2	3	4
$K = 5$	1000	0	0	0	0
$K = 10$	1000	0	0	0	0
$K = 20$	1000	0	0	0	0
$K = 40$	433	372	149	42	4

Con 5, 10 e 20 sottoinsiemi non sono state individuate stime infinite in nessun sottoinsieme in nessuno dei 10^3 campioni, mentre con $K = 40$ in più della metà dei campioni si è trovato almeno un sottogruppo con stime di massima verosimiglianza infinite. In questi casi tali gruppi sono stati esclusi dalla combinazione delle stime di massima verosimiglianza.

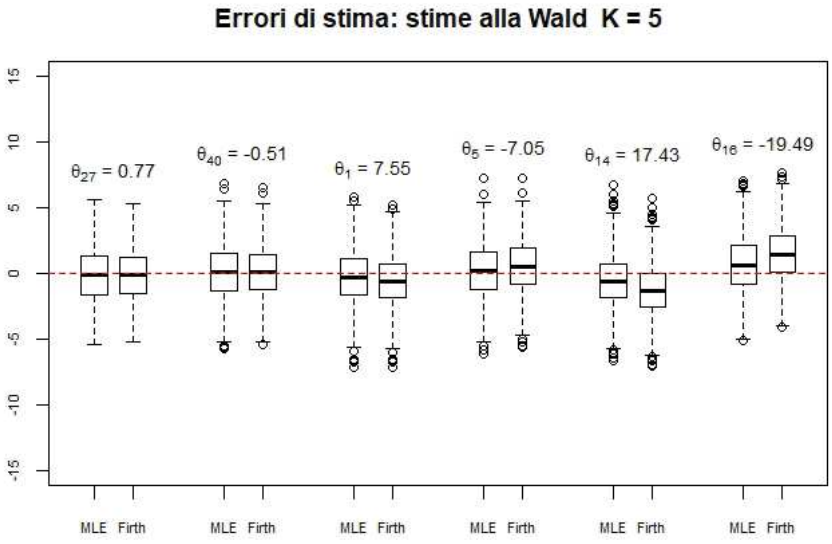

FIGURA 3.1: Boxplot degli errori di stima delle stime combinate alla Wald, $K = 5$.

Dalle Figure 3.1, 3.2, 3.3 e 3.4 è chiaro che le stime combinate alla Wald diventano sempre più distorte all'aumentare del numero di sottoinsiemi K. Bisogna anche notare però che la distorsione non interessa tutti i parametri allo stesso modo: fin da $K = 5$, ma diventa più evidente al crescere di K, va notato che è maggiore la distorsione dei parametri il cui vero valore è maggiore in modulo. In Cordeiro & McCullagh (1991) viene derivata, seppur sotto ipotesi piuttosto stringenti, un'espressione asintotica per la

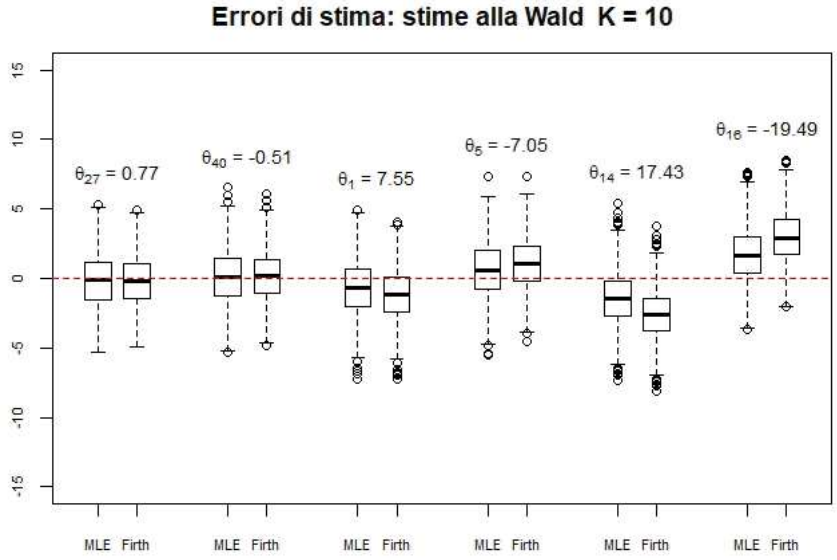

FIGURA 3.2: Boxplot degli errori di stima delle stime combinate alla Wald, $K = 10$.

distorsione degli stimatori di massima verosimiglianza nella regressione logistica binaria

$$\mathrm{E}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \simeq \frac{p}{n}\boldsymbol{\theta}_0.$$

Gli autori stessi consigliano di utilizzare questa formulazione solo come un'indicazione grezza, ma può essere comunque una spiegazione del fenomeno appena descritto. Anche Sur & Candès (2019) hanno verificato dal punto di vista numerico che la distorsione dipende dal rapporto p/n (si veda ad esempio la Figura 7A in Sur & Candès, 2019).

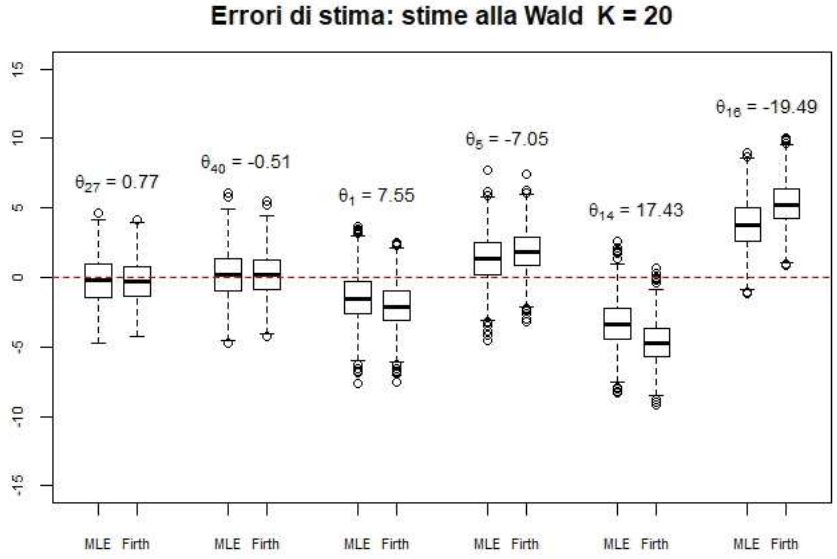

FIGURA 3.3: Boxplot degli errori di stima delle stime combinate alla Wald, $K = 20$.

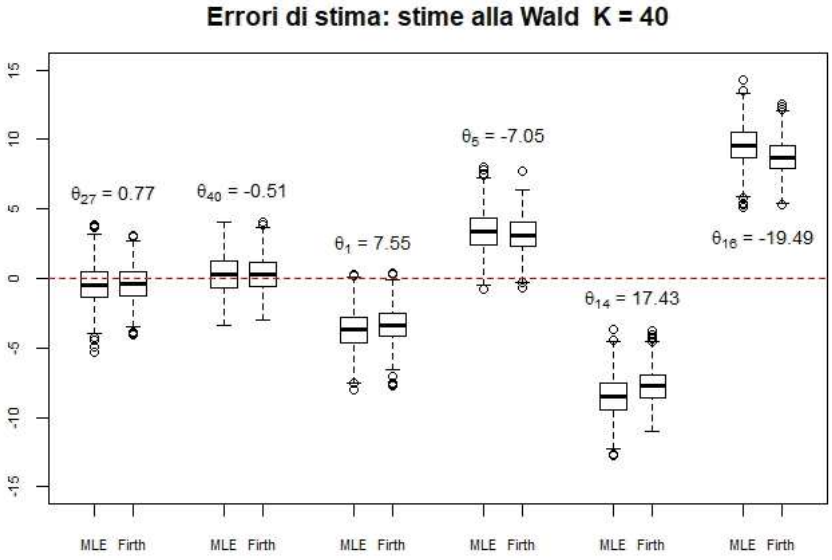

FIGURA 3.4: Boxplot degli errori di stima delle stime combinate alla Wald, $K = 40$.

Un aspetto apparentemente sorprendente, che è possibile evincere da queste figure, è l'andamento peggiore delle stime ottenute combinando le stime di Firth, rispetto a quelle ottenute combinando le stime di massima verosimiglianza, fino a $K = 20$.

TABELLA 3.3: Distorsione e distorsione relativa delle stime combinate alla Wald al variare di K.

Distorsione

	$K=5$		$K=10$		$K=20$		$K=40$	
	MLE	Firth	MLE	Firth	MLE	Firth	MLE	Firth
θ_{27}	-0.17	-0.19	-0.19	-0.22	-0.24	-0.28	-0.45	-0.40
θ_{40}	0.10	0.12	0.12	0.15	0.17	0.20	0.31	0.28
θ_1	-0.28	-0.60	-0.67	-1.16	-1.48	-2.05	-3.69	-3.37
θ_5	0.19	0.49	0.56	1.04	1.34	1.89	3.41	3.12
θ_{14}	-0.54	-1.28	-1.45	-2.60	-3.35	-4.68	-8.50	-7.77
θ_{16}	0.69	1.51	1.69	2.97	3.80	5.28	9.56	8.73

Distorsione relativa

	$K=5$		$K=10$		$K=20$		$K=40$	
	MLE	Firth	MLE	Firth	MLE	Firth	MLE	Firth
θ_{27}	-0.22	-0.24	-0.25	-0.29	-0.32	-0.37	-0.58	-0.52
θ_{40}	0.19	0.23	0.24	0.30	0.34	0.40	0.60	0.56
θ_1	-0.04	-0.08	-0.09	-0.15	-0.20	-0.27	-0.49	-0.45
θ_5	0.03	0.07	0.08	0.15	0.19	0.27	0.48	0.44
θ_{14}	-0.03	-0.07	-0.08	-0.15	-0.19	-0.27	-0.49	-0.45
θ_{16}	0.04	0.08	0.09	0.15	0.19	0.27	0.49	0.45

Nella Tabella 3.3 vengono riportate le distorsioni ottenute via simulazione e le distorsioni relative, cioè divise per il valore assoluto del vero valore del parametro corrispondente. L'andamento riscontrato nelle figure è confermato dalla tabella in alto, mentre la tabella in basso permette di notare, che rapportata al valore del parametro, la distorsione è maggiore sui parametri aventi vero valore piccolo (in modulo). Inoltre è interessante notare che a parametri il cui vero valore è positivo (θ_{27}, θ_1 e θ_{14}) corrispondono distorsioni negative, mentre a parametri il cui vero valore è negativo (θ_{40}, θ_5 e θ_{16}) corrispondono distorsioni positive.

Il peggiore andamento delle stime che combinano stime locali di Firth è in contrasto con i risultati teorici relativi alla riduzione della distorsione portata dalla correzione di Firth. Anche empiricamente si è verificato (si veda Tabella 3.4) che, tranne per i parametri prossimi a 0, la distorsione (relativa) delle stime di Firth nei sottogruppi è sensibilmente minore di quella delle stime di massima verosimiglianza, già da $K = 5$. I valori della Tabella 3.4 sono relativi alle stime nei sottoinsiemi che sono state combinate per ottenere le stime le cui caratteristiche sono state mostrate in precedenza; di conseguenza le distorsioni presentate sono basate su 5×10^3 stime per $K = 5$, su 10×10^3 per $K = 10$ e su 20×10^3 per $K = 20$.

Tabella 3.4: Distorsione relativa MLE e stime di Firth nei sottoinsiemi al variare di K.

	$K = 5$		$K = 10$		$K = 20$	
	MLE	Firth	MLE	Firth	MLE	Firth
θ_{27}	-0.18	-0.21	-0.15	-0.22	-0.10	-0.26
θ_{40}	-0.13	-0.18	-0.10	-0.20	-0.09	-0.29
θ_1	0.05	-0.00	0.11	-0.00	0.28	0.01
θ_5	0.06	0.01	0.12	0.01	0.30	0.02
θ_{14}	0.06	0.00	0.12	0.01	0.30	0.02
θ_{16}	0.05	-0.00	0.12	0.00	0.29	0.02

Nel Paragrafo 3.4.2 viene proposta una procedura alternativa per ottenere una stima combinata alla Wald per stime di Firth con migliori performance di quelle mostrate in questo paragrafo.

Per quanto riguarda le stime combinate alla Rao, è evidente che l'aggiornamento delle stime alla Wald tramite un passo Newton-Raphson è sufficiente per ottenere un netto miglioramento delle stime. Infatti, come si può vedere nelle Figure 3.5, 3.6, 3.7, 3.8, nel caso delle stime di massima verosimiglianza ciò porta ad avere ottime performance

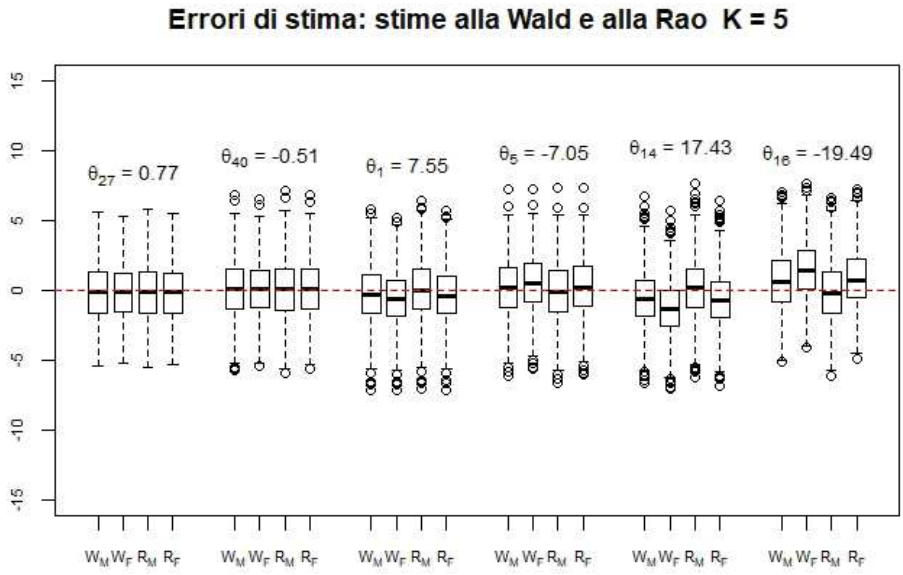

FIGURA 3.5: Boxplot degli errori di stima delle stime combinate alla Wald e alla Rao, $K = 5$ (W_M: Wald MLE, W_F: Wald Firth, R_M: Rao MLE, R_F: Rao Firth).

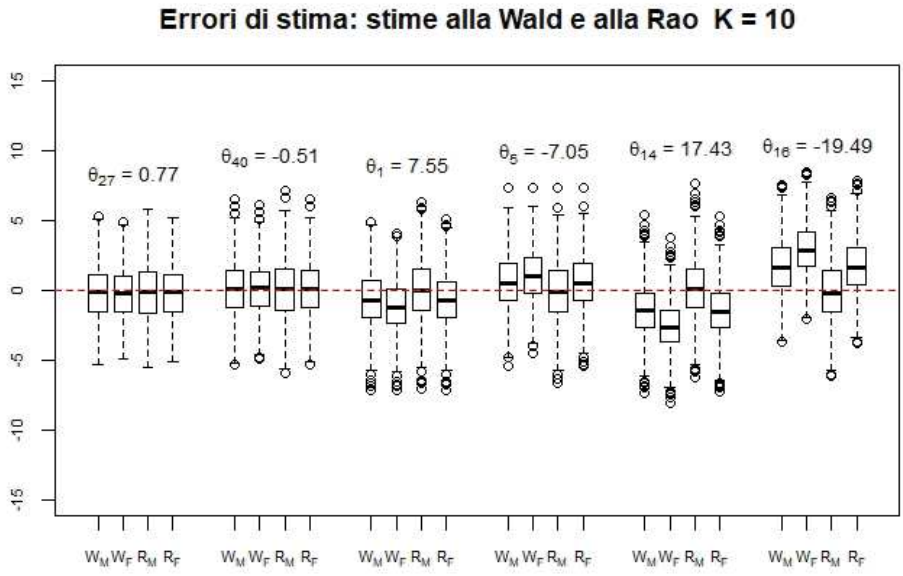

FIGURA 3.6: Boxplot degli errori di stima delle stime combinate alla Wald e alla Rao, $K = 10$ (W_M: Wald MLE, W_F: Wald Firth, R_M: Rao MLE, R_F: Rao Firth).

perfino con $K = 20$ (e comunque molto buone anche con $K = 40$), mentre nel caso delle stime di Firth il miglioramento non è sufficiente ad ottenere delle stime non distorte. Questi risultati sono notevoli da due punti di vista: non solo il semplice aggiornamento Newton-Raphson porta un significativo miglioramento, ma questo è tale da annullare

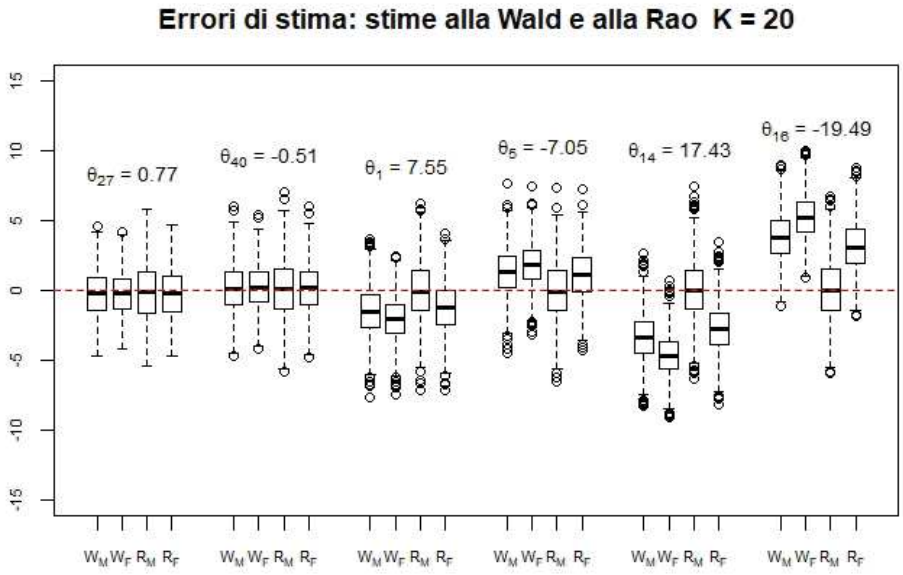

FIGURA 3.7: Boxplot degli errori di stima delle stime combinate alla Wald e alla Rao, $K = 20$ (W_M: Wald MLE, W_F: Wald Firth, R_M: Rao MLE, R_F: Rao Firth).

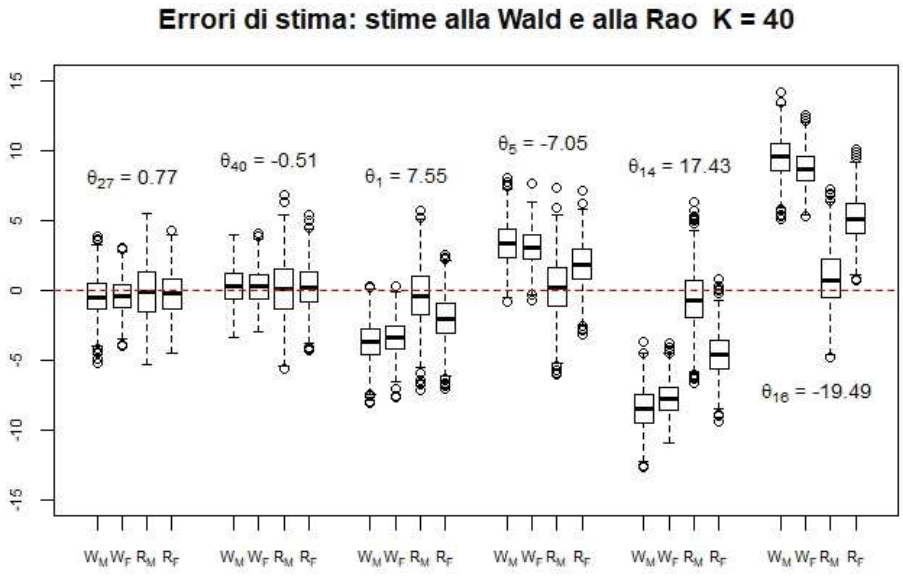

FIGURA 3.8: Boxplot degli errori di stima delle stime combinate alla Wald, $K = 40$ (W_M: Wald MLE, W_F: Wald Firth, R_M: Rao MLE, R_F: Rao Firth).

quasi completamente la distorsione delle stime di massima verosimiglianza. Nella Tabella 3.5 sono riportate le distorsioni e le distorsioni relative delle stime alla Rao, oltre che delle stime alla Wald, le quali sono già state mostrate anche nella Tabella 3.3. Si può vedere che le stime alla Rao di massima verosimiglianza sono quasi perfettamente

non distorte fino a $K = 20$, mentre le stime di massima verosimiglianza locali che vengono combinate hanno distorsione che cresce sensibilmente con $K = 20$ (si veda Tabella 3.4). Per i parametri in valore assoluto maggiori sembra addirittura che la stima alla Rao di verosimiglianza abbia distorsione decrescente al crescere di K, per un numero di sottoinsiemi minore o uguale a 20, cosa che non si può dire per nessuna delle altre stime.

Nel Paragrafo 5.7 di van der Vaart (1998) è dimostrato che, sotto alcune condizioni di regolarità della funzione di stima e delle sue derivate, lo stimatore ottenuto tramite un passo Newton-Raphson è un stimatore tanto buono quanto quello che si ottiene risolvendo esattamente l'equazione di stima. È inoltre necessario che sia disponibile, come punto di partenza dell'aggiornamento, uno stimatore che sia già un buono stimatore, nel senso che sia $\sqrt{n}-$consistente. Nel seguito viene mostrato un procedimento analogo a quello di van der Vaart (1998, Paragrafo 5.7), sfruttando risultati mostrati in quel paragrafo.

Nella dimostrazione del Teorema 5.42 di van der Vaart (1998), viene dimostrato che per ogni stimatore $\tilde{\boldsymbol{\theta}}$ consistente per $\boldsymbol{\theta}_0$ vale

$$\frac{1}{n} i(\tilde{\boldsymbol{\theta}}) \xrightarrow{p} \frac{1}{n} i(\boldsymbol{\theta}_0),$$

dove $\xrightarrow{p}$ denota convergenza in probabilità. Se $n_k = m = \frac{n}{k}$ per ogni $k = 1, \ldots, K$, vale $\frac{1}{m} i_k(\hat{\boldsymbol{\theta}}_k) \xrightarrow{p} \frac{1}{m} i_k(\boldsymbol{\theta}_0)$ e perciò

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{wcd} - \boldsymbol{\theta}_0 &= \left\{ \sum_{k=1}^{K} \frac{1}{m} i_k(\hat{\boldsymbol{\theta}}_k) \right\}^{-1} \left\{ \sum_{k=1}^{K} \frac{1}{m} i_k(\hat{\boldsymbol{\theta}}_k) \hat{\boldsymbol{\theta}}_k \right\} - \boldsymbol{\theta}_0 \\
&= \left\{ \sum_{k=1}^{K} \frac{1}{m} i_k(\hat{\boldsymbol{\theta}}_k) \right\}^{-1} \left\{ \sum_{k=1}^{K} \frac{1}{m} i_k(\hat{\boldsymbol{\theta}}_k) (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0) \right\} \\
&= \left\{ \sum_{k=1}^{K} \left(\frac{1}{m} i_k(\boldsymbol{\theta}_0) + o_p(1) \right) \right\}^{-1} \left\{ \sum_{k=1}^{K} \left(\frac{1}{m} i_k(\boldsymbol{\theta}_0) + o_p(1) \right) (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0) \right\}
\end{aligned}$$

che implica

$$\begin{gathered}
\left\{ \sum_{k=1}^{K} \left(\frac{1}{m} i_k(\boldsymbol{\theta}_0) + o_p(1) \right) \right\} \left(\hat{\boldsymbol{\theta}}_{wcd} - \boldsymbol{\theta}_0 \right) = \left\{ \sum_{k=1}^{K} \left(\frac{1}{m} i_k(\boldsymbol{\theta}_0) + o_p(1) \right) (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0) \right\} \\
\Updownarrow \\
O_p(1) \left(\hat{\boldsymbol{\theta}}_{wcd} - \boldsymbol{\theta}_0 \right) = \sum_{k=1}^{K} \{ O_p(1) o_p(1) \} = o_p(1)
\end{gathered}$$

TABELLA 3.5: Distorsione e distorsione relativa delle stime combinate alla Wald e alla Rao al variare di K (W_M: Wald MLE, W_F: Wald Firth, R_M: Rao MLE, R_F: Rao Firth).

Distorsione

		θ_{27}	θ_{40}	θ_1	θ_5	θ_{14}	θ_{16}
$K=5$	W_M	-0.17	0.10	-0.28	0.19	-0.54	0.69
	W_F	-0.19	0.12	-0.60	0.49	-1.28	1.51
	R_M	-0.15	0.08	0.03	-0.11	0.19	-0.12
	R_F	-0.17	0.10	-0.34	0.25	-0.68	0.84
$K=10$	W_M	-0.19	0.12	-0.67	0.56	-1.45	1.69
	W_F	-0.22	0.15	-1.16	1.04	-2.60	2.97
	R_M	-0.15	0.08	0.01	-0.10	0.16	-0.08
	R_F	-0.19	0.12	-0.67	0.56	-1.46	1.71
$K=20$	W_M	-0.24	0.17	-1.48	1.34	-3.35	3.80
	W_F	-0.28	0.20	-2.05	1.89	-4.68	5.28
	R_M	-0.15	0.08	-0.04	-0.04	0.02	0.07
	R_F	-0.22	0.16	-1.24	1.10	-2.78	3.17
$K=40$	W_M	-0.45	0.31	-3.69	3.41	-8.50	9.56
	W_F	-0.40	0.28	-3.37	3.12	-7.77	8.73
	R_M	-0.15	0.10	-0.34	0.22	-0.66	0.83
	R_F	-0.27	0.20	-2.01	1.81	-4.57	5.17

Distorsione relativa

		θ_{27}	θ_{40}	θ_1	θ_5	θ_{14}	θ_{16}
$K=5$	W_M	-0.22	0.19	-0.04	0.03	-0.03	0.04
	W_F	-0.24	0.23	-0.08	0.07	-0.07	0.08
	R_M	-0.19	0.15	0.00	-0.02	0.01	-0.01
	R_F	-0.22	0.20	-0.04	0.03	-0.04	0.04
$K=10$	W_M	-0.25	0.24	-0.09	0.08	-0.08	0.09
	W_F	-0.29	0.30	-0.15	0.15	-0.15	0.15
	R_M	-0.19	0.15	0.00	-0.01	0.01	-0.00
	R_F	-0.24	0.24	-0.09	0.08	-0.08	0.09
$K=20$	W_M	-0.32	0.34	-0.20	0.19	-0.19	0.19
	W_F	-0.37	0.40	-0.27	0.27	-0.27	0.27
	R_M	-0.19	0.16	-0.01	-0.01	0.00	0.00
	R_F	-0.29	0.31	-0.16	0.16	-0.16	0.16
$K=40$	W_M	-0.58	0.60	-0.49	0.48	-0.49	0.49
	W_M	-0.52	0.56	-0.45	0.44	-0.45	0.45
	W_M	-0.20	0.20	-0.04	0.03	-0.04	0.04
	W_M	-0.36	0.39	-0.27	0.26	-0.26	0.26

e quindi $\hat{\boldsymbol{\theta}}_{wcd}$ è consistente per $\boldsymbol{\theta}_0$.

È possibile riscrivere lo stimatore alla Rao che combina le stime di massima verosimiglianza nel seguente modo

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{rcd} &= \hat{\boldsymbol{\theta}}_{wcd} + \left\{\sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}_{wcd})\right\}^{-1} \left\{\sum_{k=1}^{K} U_k(\hat{\boldsymbol{\theta}}_{wcd})\right\} \\ &= \hat{\boldsymbol{\theta}}_{wcd} + \left\{i(\hat{\boldsymbol{\theta}}_{wcd})\right\}^{-1} U(\hat{\boldsymbol{\theta}}_{wcd}) \\ &= \hat{\boldsymbol{\theta}}_{wcd} + \left\{\frac{1}{n} i(\hat{\boldsymbol{\theta}}_{wcd})\right\}^{-1} \left\{\frac{1}{n} U(\hat{\boldsymbol{\theta}}_{wcd})\right\} \\ &= \hat{\boldsymbol{\theta}}_{wcd} + \left\{\frac{1}{n} i(\boldsymbol{\theta}_0) + o_p(1)\right\}^{-1} \left\{\frac{1}{n} U(\hat{\boldsymbol{\theta}}_{wcd})\right\}. \end{aligned} \tag{3.7}$$

Lo sviluppo di Taylor dello *score* intorno alla stima di massima verosimiglianza su tutto l'insieme dei dati $\hat{\boldsymbol{\theta}}$ permette di approssimare lo *score* valutato nella stima alla Wald come

$$\begin{aligned} U(\hat{\boldsymbol{\theta}}_{wcd}) &\doteq U(\hat{\boldsymbol{\theta}}) + \left\{\frac{\partial}{\partial \boldsymbol{\theta}} U(\boldsymbol{\theta})\right\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} (\hat{\boldsymbol{\theta}}_{wcd} - \hat{\boldsymbol{\theta}}) \\ &\doteq -i(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{wcd} - \hat{\boldsymbol{\theta}}). \end{aligned}$$

Sostituendo tale approssimazione in (3.7) si ottiene

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{rcd} &\doteq \hat{\boldsymbol{\theta}}_{wcd} - \left\{\frac{1}{n} i(\boldsymbol{\theta}_0) + o_p(1)\right\}^{-1} \left\{\frac{1}{n} i(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{wcd} - \hat{\boldsymbol{\theta}})\right\} \\ &\doteq \hat{\boldsymbol{\theta}}_{wcd} - \left\{\frac{1}{n} i(\boldsymbol{\theta}_0) + o_p(1)\right\}^{-1} \left\{\frac{1}{n} i(\boldsymbol{\theta}_0) + o_p(1)\right\} (\hat{\boldsymbol{\theta}}_{wcd} - \hat{\boldsymbol{\theta}}) \\ &\doteq \hat{\boldsymbol{\theta}}. \end{aligned}$$

In sostanza si può approssimare la stima alla Rao aggiornando la stima $\hat{\boldsymbol{\theta}}_{wcd}$ con un passo Newton-Raphson che porta ad avere una stima asintoticamente equivalente alla stima di massima verosimiglianza globale. Questo risultato è coerente con l'osservazione del Paragrafo 3.1.2 in cui si è notato che l'approssimazione $\tilde{\Psi}_R(\boldsymbol{\theta})$ di $\Psi_R(\boldsymbol{\theta})$ coincide con lo *score* globale, che implica che l'approssimazione della stima alla Rao equivale ad approssimare la stima di massima verosimiglianza globale con un passo Newton-Raphson a partire dalla stima alla Wald. Chiaramente, per ottenere questo miglioramento di $\hat{\boldsymbol{\theta}}_{wcd}$ è necessario valutare gli *score* locali in $\hat{\boldsymbol{\theta}}_{wcd}$ e quindi riutilizzare i dati originali dei singoli sottocampioni.

Per quanto riguarda la stima alla Wald di massima verosimiglianza, è possibile mostrare che l'ordine del termine dominante della sua distorsione non dipende dalla dimensione campionaria complessiva n, ma dalla dimensione campionaria dei sottoinsiemi, giustificando in questo modo la maggiore distorsione rispetto alla stima alla Rao e la sua crescita all'aumentare del numero di sottogruppi (mantenendo la dimensione campionaria complessiva costante). Scrivendo la matrice di informazione attesa del k-esimo gruppo valutata nella stima di massima verosimiglianza locale come

$$i_k(\hat{\boldsymbol{\theta}}_k) \doteq i_k(\boldsymbol{\theta}_0) + \sum_{r=1}^{p} \frac{\partial\, i_k(\boldsymbol{\theta})}{\partial \theta_r}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\hat{\theta}_{k,r} - \theta_{0,r}),$$

dove $\theta_{0,r}$ e $\hat{\theta}_{k,r}$ sono le r-esime componenti rispettivamente di $\boldsymbol{\theta}_0$ e $\hat{\boldsymbol{\theta}}_k$, si può trovare la seguente espressione per $\sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}_k)$

$$\begin{aligned}\sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}_k) &\doteq \sum_{k=1}^{K} \left\{ i_k(\boldsymbol{\theta}_0) + \sum_{r=1}^{p} \frac{\partial\, i_k(\boldsymbol{\theta})}{\partial \theta_r}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\hat{\theta}_{k,r} - \theta_{0,r}) \right\} \\ &\doteq i(\boldsymbol{\theta}_0) + \sum_{k=1}^{K} \sum_{r=1}^{p} \frac{\partial\, i_k(\boldsymbol{\theta})}{\partial \theta_r}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\hat{\theta}_{k,r} - \theta_{0,r})\end{aligned}$$

e di conseguenza

$$\left\{ \sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}_k) \right\}^{-1} \doteq i(\boldsymbol{\theta}_0)^{-1} \left\{ \boldsymbol{I}_p + i(\boldsymbol{\theta}_0)^{-1} \sum_{k=1}^{K} \sum_{r=1}^{p} \frac{\partial\, i_k(\boldsymbol{\theta})}{\partial \theta_r}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\hat{\theta}_{k,r} - \theta_{0,r}) \right\}^{-1}$$

Se, come in precedenza, $n_k = m = \frac{n}{k}$ per ogni $k = 1, \ldots, K$, siccome l'addendo $i(\boldsymbol{\theta}_0)^{-1} \sum_{k=1}^{K} \sum_{r=1}^{p} \frac{\partial\, i_k(\boldsymbol{\theta})}{\partial \theta_r}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\hat{\theta}_{k,r} - \theta_{0,r})$ è di ordine $O_p(m^{-1/2})$, e quindi $o_p(1)$, si può sfruttare la seguente approssimazione al primo termine

$$\begin{aligned}&\left\{ \boldsymbol{I}_p + i(\boldsymbol{\theta}_0)^{-1} \sum_{k=1}^{K} \sum_{r=1}^{p} \frac{\partial\, i_k(\boldsymbol{\theta})}{\partial \theta_r}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\hat{\theta}_{k,r} - \theta_{0,r}) \right\}^{-1} \doteq \\ &\qquad \doteq \boldsymbol{I}_p - \left\{ i(\boldsymbol{\theta}_0)^{-1} \sum_{k=1}^{K} \sum_{r=1}^{p} \frac{\partial\, i_k(\boldsymbol{\theta})}{\partial \theta_r}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\hat{\theta}_{k,r} - \theta_{0,r}) \right\},\end{aligned}$$

poiché, se una matrice $\boldsymbol{A}$ è $\boldsymbol{A} = o_p(1)$, allora

$$\{\boldsymbol{I}_p + \boldsymbol{A}\}^{-1} = \boldsymbol{I}_p - \boldsymbol{A} + \boldsymbol{A}^2 - \boldsymbol{A}^3 + \ldots .$$

Mettendo assieme i termini, si ottiene

$$
\begin{aligned}
\hat{\boldsymbol{\theta}}_{wcd} - \boldsymbol{\theta}_0 &= \left\{ \sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}_k) \right\}^{-1} \left\{ \sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}_k)(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0) \right\} \\
&\doteq i(\boldsymbol{\theta}_0)^{-1} \left\{ \boldsymbol{I}_p + i(\boldsymbol{\theta}_0)^{-1} \sum_{k=1}^{K} \sum_{r=1}^{p} \left. \frac{\partial\, i_k(\boldsymbol{\theta})}{\partial \theta_r} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\hat{\theta}_{k,r} - \theta_{0,r}) \right\}^{-1} \times \\
&\qquad \times \left\{ \sum_{k=1}^{K} \left[i_k(\boldsymbol{\theta}_0) + \sum_{r=1}^{p} \left. \frac{\partial\, i_k(\boldsymbol{\theta})}{\partial \theta_r} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\hat{\theta}_{k,r} - \theta_{0,r}) \right] (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0) \right\} \\
&\doteq i(\boldsymbol{\theta}_0)^{-1} \left\{ \boldsymbol{I}_p + i(\boldsymbol{\theta}_0)^{-1} \sum_{k=1}^{K} \sum_{r=1}^{p} \left. \frac{\partial\, i_k(\boldsymbol{\theta})}{\partial \theta_r} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\hat{\theta}_{k,r} - \theta_{0,r}) \right\} \times \\
&\quad \times \left\{ \sum_{k=1}^{K} i_k(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0) + \sum_{k=1}^{K} \left[\sum_{r=1}^{p} \left. \frac{\partial\, i_k(\boldsymbol{\theta})}{\partial \theta_r} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\hat{\theta}_{k,r} - \theta_{0,r}) \right] (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0) \right\} \\
&\doteq \left\{ i(\boldsymbol{\theta}_0)^{-1} + i(\boldsymbol{\theta}_0)^{-2} \sum_{k=1}^{K} \sum_{r=1}^{p} \left. \frac{\partial\, i_k(\boldsymbol{\theta})}{\partial \theta_r} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\hat{\theta}_{k,r} - \theta_{0,r}) \right\} \times \\
&\qquad \times \left\{ \sum_{k=1}^{K} i_k(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0) + \sum_{k=1}^{K} \left[\sum_{r=1}^{p} \left. \frac{\partial\, i_k(\boldsymbol{\theta})}{\partial \theta_r} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\hat{\theta}_{k,r} - \theta_{0,r}) \right] (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0) \right\} \\
&\doteq i(\boldsymbol{\theta}_0)^{-1} \sum_{k=1}^{K} i_k(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0) + \\
&\qquad + i(\boldsymbol{\theta}_0)^{-1} \sum_{k=1}^{K} \left[\sum_{r=1}^{p} \left. \frac{\partial\, i_k(\boldsymbol{\theta})}{\partial \theta_r} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\hat{\theta}_{k,r} - \theta_{0,r}) \right] (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0) + \\
&\qquad + \left\{ i(\boldsymbol{\theta}_0)^{-2} \sum_{k=1}^{K} \sum_{r=1}^{p} \left. \frac{\partial\, i_k(\boldsymbol{\theta})}{\partial \theta_r} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\hat{\theta}_{k,r} - \theta_{0,r}) \right\} \left\{ \sum_{k=1}^{K} i_k(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0) \right\} + \\
&\qquad + \left\{ i(\boldsymbol{\theta}_0)^{-2} \sum_{k=1}^{K} \sum_{r=1}^{p} \left. \frac{\partial\, i_k(\boldsymbol{\theta})}{\partial \theta_r} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\hat{\theta}_{k,r} - \theta_{0,r}) \right\} \times \\
&\qquad\qquad \times \left\{ \sum_{k=1}^{K} \left[\sum_{r=1}^{p} \left. \frac{\partial\, i_k(\boldsymbol{\theta})}{\partial \theta_r} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\hat{\theta}_{k,r} - \theta_{0,r}) \right] (\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0) \right\}.
\end{aligned}
\tag{3.8}
$$

Per valutare l'ordine della distorsione della stima alla Wald occorre considerare il valore atteso dell'ultima espressione. Il primo addendo dell'approssimazione di $\mathrm{E}_{\boldsymbol{\theta}_0}(\hat{\boldsymbol{\theta}}_{wcd} - \boldsymbol{\theta}_0)$ risulta essere

$$
i(\boldsymbol{\theta}_0)^{-1} \sum_{k=1}^{K} i_k(\boldsymbol{\theta}_0) \mathrm{E}_{\boldsymbol{\theta}_0}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_0).
$$

Il secondo è

$$i(\boldsymbol{\theta}_0)^{-1}\sum_{k=1}^{K}\left[\sum_{r=1}^{p}\sum_{t=1}^{p}\left.\frac{\partial\, i_{k,st}(\boldsymbol{\theta})}{\partial\theta_r}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\mathrm{E}_{\boldsymbol{\theta}_0}\{(\hat{\theta}_{k,r}-\theta_{0,r})(\hat{\theta}_{k,t}-\theta_{0,t})\}\right]_{s=1,\ldots,p},$$

dove con $[a_s]_{s=1,\ldots,p}$ si indica il vettore $(a_1, a_2, \ldots, a_p)$ e con $i_{k,st}(\boldsymbol{\theta})$ l'elemento in posizione (s,t) della matrice $i_k(\boldsymbol{\theta})$. Il terzo addendo è

$$i(\boldsymbol{\theta}_0)^{-2}\sum_{k=1}^{K}\sum_{h=1}^{K}\left[\sum_{t=1}^{p}\sum_{r=1}^{p}\sum_{u=1}^{p}\left.\frac{\partial\, i_{k,st}(\boldsymbol{\theta})}{\partial\theta_r}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} i_{h,tu}(\boldsymbol{\theta}_0)\mathrm{E}_{\boldsymbol{\theta}_0}\{(\hat{\theta}_{k,r}-\theta_{0,r})(\hat{\theta}_{h,u}-\theta_{0,u})\}\right]_{s=1,\ldots,p}$$

e infine il quarto è

$$i(\boldsymbol{\theta}_0)^{-2}\sum_{k=1}^{K}\sum_{h=1}^{K}\left[\sum_{t=1}^{p}\sum_{r=1}^{p}\sum_{u=1}^{p}\sum_{v=1}^{p}\left.\frac{\partial\, i_{k,st}(\boldsymbol{\theta})}{\partial\theta_r}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\left.\frac{\partial\, i_{k,tu}(\boldsymbol{\theta})}{\partial\theta_v}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\times\right.$$
$$\left.\times\,\mathrm{E}_{\boldsymbol{\theta}_0}\{(\hat{\theta}_{k,r}-\theta_{0,r})(\hat{\theta}_{h,u}-\theta_{0,u})(\hat{\theta}_{h,v}-\theta_{0,v})\}\right]_{s=1,\ldots,p}.$$

Per quanto riguarda i valori attesi $\mathrm{E}_{\boldsymbol{\theta}_0}\{(\hat{\theta}_{k,r}-\theta_{0,r})(\hat{\theta}_{h,u}-\theta_{0,u})\}$ e $\mathrm{E}_{\boldsymbol{\theta}_0}\{(\hat{\theta}_{k,r}-\theta_{0,r})(\hat{\theta}_{h,u}-\theta_{0,u})(\hat{\theta}_{h,v}-\theta_{0,v})\}$ del terzo e quarto addendo, è comodo richiamare l'espansione di $\hat{\theta}_{k,r}-\theta_{0,r}$ in funzione delle componenti dello *score*. La Formula (9.61) di Pace & Salvan (1997) garantisce che

$$\hat{\theta}_{k,r}-\theta_{0,r}=\sum_{s=1}^{p} i_k^{rs}(\boldsymbol{\theta}_0)U_{k,s}(\boldsymbol{\theta}_0)+O(m^{-1}),$$

dove $i_k^{rs}(\boldsymbol{\theta}_0)$ è l'elemento in posizione (r,s) dell'inversa della matrice $i_k(\boldsymbol{\theta}_0)$. Perciò

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\theta}_0}\{(\hat{\theta}_{k,r}-\theta_{0,r})(\hat{\theta}_{h,u}-\theta_{0,u})\} &\doteq \mathrm{E}_{\boldsymbol{\theta}_0}\left\{\left(\sum_{s=1}^{p} i_k^{rs}(\boldsymbol{\theta}_0)U_{k,s}(\boldsymbol{\theta}_0)\right)\left(\sum_{t=1}^{p} i_k^{ut}(\boldsymbol{\theta}_0)U_{h,t}(\boldsymbol{\theta}_0)\right)\right\}\\ &\doteq \sum_{s=1}^{p}\sum_{t=1}^{p} i_k^{rs}(\boldsymbol{\theta}_0)i_k^{ut}(\boldsymbol{\theta}_0)\mathrm{E}_{\boldsymbol{\theta}_0}\left\{U_{k,s}(\boldsymbol{\theta}_0)U_{h,t}(\boldsymbol{\theta}_0)\right\},\end{aligned}$$

che è pari a 0 se $k \neq h$, poiché $U_{k,s}(\boldsymbol{\theta}_0)$ e $U_{h,t}(\boldsymbol{\theta}_0)$ sono variabili casuali indipendenti a media 0, ed è di ordine $O(m^{-1})$ altrimenti. Analogamente, per il quarto ed ultimo

addendo del valore atteso dell'ultima riga della (3.8) si ha che

$$\begin{aligned}
&\mathrm{E}_{\boldsymbol{\theta}_0}\{(\hat{\theta}_{k,r}-\theta_{0,r})(\hat{\theta}_{h,u}-\theta_{0,u})(\hat{\theta}_{h,v}-\theta_{0,v})\} \doteq \\
&\quad \doteq \mathrm{E}_{\boldsymbol{\theta}_0}\left\{\left(\sum_{s=1}^{p} i_k^{rs}(\boldsymbol{\theta}_0)U_{k,s}(\boldsymbol{\theta}_0)\right)\left(\sum_{t=1}^{p} i_k^{ut}(\boldsymbol{\theta}_0)U_{h,t}(\boldsymbol{\theta}_0)\right)\left(\sum_{W=1}^{p} i_k^{vw}(\boldsymbol{\theta}_0)U_{h,w}(\boldsymbol{\theta}_0)\right)\right\} \\
&\quad \doteq \sum_{s=1}^{p}\sum_{t=1}^{p} i_k^{rs}(\boldsymbol{\theta}_0)i_k^{ut}(\boldsymbol{\theta}_0)i_k^{vw}(\boldsymbol{\theta}_0)\mathrm{E}_{\boldsymbol{\theta}_0}\left\{U_{k,s}(\boldsymbol{\theta}_0)U_{h,t}(\boldsymbol{\theta}_0)U_{h,w}(\boldsymbol{\theta}_0)\right\}
\end{aligned}$$

è pari a 0 se $k \neq h$ ed è di ordine $O(m^{-2})$ altrimenti (si veda Pace & Salvan, 1997, Formula (9.74)).
Unendo i termini calcolati, si ottiene

$$\begin{aligned}
\mathrm{E}_{\boldsymbol{\theta}_0}(\hat{\boldsymbol{\theta}}_{wcd}-\boldsymbol{\theta}_0) &\doteq O(n^{-1})\sum_{k=1}^{K} O(m)O(m^{-1}) + O(n^{-1})\sum_{k=1}^{K} O(m)O(m^{-1}) + \\
&\quad + O(n^{-2})\sum_{k=1}^{K} O(m^2)O(m^{-1}) + O(n^{-2})\sum_{k=1}^{K} O(m^2)O(m^{-2}) \\
&\doteq O(m^{-1}) + O(m^{-1}) + O(n^{-1}) + O(n^{-1}m^{-1}) \\
&\doteq O(m^{-1}).
\end{aligned}$$

In conclusione, è stato mostrato che la stima alla Wald non può risultare in una buona stima finale per $\boldsymbol{\theta}$ poiché l'ordine della distorsione decresce con la dimensione campionaria dei sottogruppi, che è conveniente siano piccoli, e non con la dimensione campionaria complessiva, però è un buon punto di partenza per ottenere la stima alla Rao che asintoticamente si comporta come la stima di massima verosimiglianza ottenuta su tutti i dati.

3.4 Miglioramento delle stime combinate di Firth

3.4.1 Miglioramento delle stime alla Rao di Firth

Nonostante le stime di Firth nei sottogruppi siano meno distorte di quelle di massima verosimiglianza e la stessa procedura sia usata per combinare le stime di massima verosimiglianza e le stime di Firth, nell'ultimo paragrafo è stato mostrato che la distorsione è molto diversa nei due casi, soprattutto per quanto riguarda le stime combinate alla Rao. Perciò, in questo paragrafo si prova ad individuare il motivo di questa differenza e a correggere di conseguenza la combinazione.

La differenza principale nell'aggiornamento Newton-Raphson tra le stime di massima verosimiglianza e le stime di Firth è la possibilità di scrivere lo *score* esattamente come la somma degli *score* nei K sottoinsiemi, come già visto in (3.7)

$$\begin{aligned} U(\hat{\boldsymbol{\theta}}_{wcd}) &= \sum_{i=1}^{n} \left\{ y_i - \frac{\exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\theta}}_{wcd})}{1+\exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\theta}}_{wcd})} \right\} \boldsymbol{x}_i \\ &= \sum_{k=1}^{K} \sum_{i=1}^{n_k} \left\{ y_i - \frac{\exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\theta}}_{wcd})}{1+\exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\theta}}_{wcd})} \right\} \boldsymbol{x}_i \\ &= \sum_{k=1}^{K} U_k(\hat{\boldsymbol{\theta}}_{wcd}). \end{aligned}$$

Non si può fare lo stesso con la correzione di Firth, che è

$$\sum_{i=1}^{n} h_i \left\{ \frac{1}{2} - \frac{\exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\theta}}^*_{wcd})}{1+\exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\theta}}^*_{wcd})} \right\} \boldsymbol{x}_i,$$

dove h_i è l'i-esimo elemento diagonale della matrice $\boldsymbol{H}$ definita nel Capitolo 2. Nonostante apparentemente si possa riprodurre quanto fatto con $U(\hat{\boldsymbol{\theta}}_{wcd})$, la matrice $\boldsymbol{H}$ necessita di tutta la matrice $\boldsymbol{X}$ per essere calcolata e perciò

$$U^*(\hat{\boldsymbol{\theta}}^*_{wcd}) \neq \sum_{k=1}^{K} U_k^*(\hat{\boldsymbol{\theta}}^*_{wcd}) = U(\hat{\boldsymbol{\theta}}^*_{wcd}) + \sum_{k=1}^{K} \sum_{i=1}^{n_k} h_i^{(k)} \left(\frac{1}{2} - \frac{\exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\theta}}^*_{wcd})}{1+\exp(\boldsymbol{x}_i^\top \hat{\boldsymbol{\theta}}^*_{wcd})} \right) \boldsymbol{x}_i,$$

dove $h_i^{(k)}$ è l'elemento diagonale di $\boldsymbol{H}_k = \boldsymbol{W}_k^{1/2} \boldsymbol{X}_k \left(\boldsymbol{X}_k^\top \boldsymbol{W}_k \boldsymbol{X}_k \right)^{-1} \boldsymbol{X}_k^\top \boldsymbol{W}_k^{1/2}$.

Siano $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_K$ le matrici del disegno $n_k \times p$ dei K sottoinsiemi. Esse sono tali che

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_K. \end{bmatrix} \qquad \text{e} \qquad \boldsymbol{X}^\top = \begin{bmatrix} \boldsymbol{X}_1^\top \ \ldots \ \boldsymbol{X}_K^\top \end{bmatrix}.$$

Perciò la matrice $\boldsymbol{H}$ può essere scritta come segue

$$\begin{aligned} \boldsymbol{H} &= \begin{bmatrix} \boldsymbol{W}_1^{1/2} & & \\ & \ddots & \\ & & \boldsymbol{W}_K^{1/2} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_K. \end{bmatrix} \left(\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X} \right)^{-1} \begin{bmatrix} \boldsymbol{X}_1^\top \ \ldots \ \boldsymbol{X}_K^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{W}_1^{1/2} & & \\ & \ddots & \\ & & \boldsymbol{W}_K^{1/2} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{W}_1^{1/2} \boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{W}_K^{1/2} \boldsymbol{X}_K. \end{bmatrix} \left(\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X} \right)^{-1} \begin{bmatrix} \boldsymbol{X}_1^\top \boldsymbol{W}_1^{1/2} \ \ldots \ \boldsymbol{X}_K^\top \boldsymbol{W}_K^{1/2} \end{bmatrix}. \end{aligned}$$

Per la correzione di Firth sono necessari solo gli elementi diagonali di $\boldsymbol{H}$, perciò è sufficiente concentrarsi sui blocchi diagonali. Il k-esimo di questi blocchi è

$$\boldsymbol{W}_k^{1/2}\boldsymbol{X}_k\left(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}_k^\top\boldsymbol{W}_k^{1/2}.$$

Inoltre, analogamente,

$$\begin{aligned}\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X} &= \begin{bmatrix}\boldsymbol{X}_1^\top & \dots & \boldsymbol{X}_K^\top\end{bmatrix}\begin{bmatrix}\boldsymbol{W}_1 & & \\ & \ddots & \\ & & \boldsymbol{W}_K\end{bmatrix}\begin{bmatrix}\boldsymbol{X}_1 \\ \vdots \\ \boldsymbol{X}_K.\end{bmatrix} \\ &= \sum_{k=1}^{K}\boldsymbol{X}_k^\top\boldsymbol{W}_k\boldsymbol{X}_k.\end{aligned}$$

Questi risultati permettono di definire una struttura per ottenere, almeno in via approssimata, la matrice $\boldsymbol{H}$ su tutti i dati, che permette di aggiornare la stima $\hat{\boldsymbol{\theta}}^*_{wcd}$ con l'intero *score* modificato, e non con la somma degli *score* modificati dei sottoinsiemi. La scelta che permette di costruire, come verrà mostrato nella Paragrafo 3.5, lo stimatore con minore distorsione consiste nel valutare la matrice $\boldsymbol{H}$ in $\hat{\boldsymbol{\theta}}^*_{wcd}$. In particolare, indicando con $\hat{\boldsymbol{W}}^*_k$ la matrice $\boldsymbol{W}_k$ valutata in $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^*_{wcd}$, se in ogni sottoinsieme si salvano $\boldsymbol{X}_k^\top\hat{\boldsymbol{W}}^*_k\boldsymbol{X}_k$ e $\left(\hat{\boldsymbol{W}}^*_k\right)^{1/2}\boldsymbol{X}_k$ è possibile ottenere il k-esimo blocco diagonale della matrice $\boldsymbol{H}$

$$\left(\hat{\boldsymbol{W}}^*_k\right)^{1/2}\boldsymbol{X}_k\left(\sum_{k=1}^{K}\boldsymbol{X}_k^\top\hat{\boldsymbol{W}}^*_k\boldsymbol{X}_k\right)^{-1}\boldsymbol{X}_k^\top\left(\hat{\boldsymbol{W}}^*_k\right)^{1/2}, \tag{3.9}$$

per $k = 1, \dots, K$, e gli elementi diagonali di ciascuno di questi blocchi vengono indicati con $\hat{h}^*_i$, per $i = 1, \dots, n$.

Utilizzando tali valori è possibile ottenere la seguente stima alla Rao

$$\hat{\boldsymbol{\theta}}^*_{ct} = \hat{\boldsymbol{\theta}}^*_{wcd} + \left\{\sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}^*_{wcd})\right\}^{-1}\left\{\sum_{k=1}^{K} U_k(\hat{\boldsymbol{\theta}}^*_{wcd}) + \sum_{i=1}^{n}\hat{h}^*_i\left(\frac{1}{2} - \frac{\exp(\boldsymbol{x}_i^\top\hat{\boldsymbol{\theta}}^*_{wcd})}{1+\exp(\boldsymbol{x}_i^\top\hat{\boldsymbol{\theta}}^*_{wcd})}\right)\boldsymbol{x}_i\right\},$$

che da qui in avanti verrà chiamata stima alla Rao Correzione Totale (CT), per sottolineare che nell'aggiornamento Newton-Raphson viene utilizzata, la correzione di Firth su tutto l'insieme dei dati e non la somma delle correzioni nei K sottoinsiemi.

Finora in questo lavoro non è stata data importanza all'aspetto computazionale, ma può essere rilevante tenerne conto: questo tipo di procedure per la combinazione di stime risulta molto utile se si ha a disposizione un'architettura informatica distribuita, che permette di ottenere le stime dei K sottoinsiemi in K diversi nodi, che vengono

poi inviate al nodo centrale in cui vengono combinate. Da questo punto di vista, la scelta migliore è quella che ottimizza il trasferimento di informazioni tra i nodi. Una possibile approssimazione che va in questa direzione è data dalla sostituzione di $\hat{\boldsymbol{\theta}}^*_{wcd}$ con $\hat{\boldsymbol{\theta}}^*_k$ nel calcolo di ciascuna $\boldsymbol{W}_k$ per ottenere l'approssimazione di h_i e l'informazione attesa del k-esimo gruppo. In questo modo si utilizzerebbero solo elementi già necessari per la costruzione di $\hat{\boldsymbol{\theta}}^*_{wcd}$ e quindi sono oggetti che sono già stati trasferiti dai nodi distribuiti a quello centrale in cui avviene la combinazione. Va notato che anche in questo caso è comunque necessario valutare gli *score* locali in $\hat{\boldsymbol{\theta}}_{wcd}$ e quindi è necessario che il nodo centrale invii comunque $\hat{\boldsymbol{\theta}}^*_{wcd}$ ai nodi periferici, cosicché vengano calcolati gli *score* nella stima alla Wald, che infine vengono restituiti al nodo centrale. Perciò il calcolo di $i_k(\hat{\boldsymbol{\theta}}^*_{wcd})$ e degli $\hat{h}^*_i$ non richiede che venga stabilita un'ulteriore connessione, ma solo che ciascun nodo insieme a $U_k(\hat{\boldsymbol{\theta}}^*_{wcd})$ mandi al nodo centrale anche $i_k(\hat{\boldsymbol{\theta}}^*_{wcd})$ e $\left(\hat{\boldsymbol{W}}^*_k\right)^{1/2}\boldsymbol{X}_k$. Nel Paragrafo 3.5 verrà sottolineato come l'utilizzo delle quantità valutate in $\hat{\boldsymbol{\theta}}^*_{wcd}$ comporti un miglioramento nelle performance delle stime alla Rao (anche se probabilmente con un costo computazionale maggiore).

3.4.2 Miglioramento delle stime alla Wald di Firth

Nella precedente sezione è stato ipotizzato che il peggiore comportamento delle stime alla Rao di Firth sia attribuibile all'impossibilità di scrivere lo *score* corretto di Firth relativo all'intero insieme di dati come somma degli *score* corretti ottenuti nei K sottoinsiemi. Ciò può essere additato anche come causa della differenza nelle stime combinate alla Wald, in favore delle stime di massima verosimiglianza.

Innanzitutto è interessante evidenziare una seconda possibile via per ottenere le stime alla Wald, diversa da quella utilizzata in precedenza che sfrutta le *confidence density*. Sia

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n}\left\{y_i\boldsymbol{x}_i^\top\boldsymbol{\theta} - \log\left(1+e^{\boldsymbol{x}_i^\top\boldsymbol{\theta}}\right)\right\}\\ &= \sum_{k=1}^{K}\ell_k(\boldsymbol{\theta})\end{aligned}$$

la log-verosimiglianza per il parametro $\boldsymbol{\theta}$ del modello di regressione logistica binaria sull'intero insieme dei dati e analogamente $\ell_1(\boldsymbol{\theta}),\ldots,\ell_K(\boldsymbol{\theta})$ le log-verosimiglianze nei K sottoinsiemi. Lo sviluppo di Taylor di ciascuna log-verosimiglianza locale $\ell_k(\boldsymbol{\theta})$ intorno alla corrispondente stima di massima verosimiglianza locale $\hat{\boldsymbol{\theta}}_k$, permette di ottenere la

seguente forma quadratica

$$\begin{aligned}\ell_k(\boldsymbol{\theta}) &\doteq \ell_k(\hat{\boldsymbol{\theta}}_k) + U_k(\hat{\boldsymbol{\theta}}_k)^\top(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_k) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_k)^\top i_k(\hat{\boldsymbol{\theta}}_k)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_k) \\ &\doteq c - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_k)^\top i_k(\hat{\boldsymbol{\theta}}_k)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_k),\end{aligned}$$

dove $U_k(\hat{\boldsymbol{\theta}}_k) = \mathbf{0}$ e c è una costante indipendente da $\boldsymbol{\theta}$. Per lo *score* a questa approssimazione corrisponde

$$U_k(\boldsymbol{\theta}) \doteq i_k(\hat{\boldsymbol{\theta}}_k)(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}).$$

Siccome $\ell(\boldsymbol{\theta}) = \sum_{k=1}^K \ell_k(\boldsymbol{\theta})$, e quindi $U(\boldsymbol{\theta}) = \sum_{k=1}^K U_k(\boldsymbol{\theta})$, l'equazione di verosimiglianza può essere approssimata con

$$\mathbf{0} = U(\boldsymbol{\theta}) \doteq \sum_{k=1}^K i_k(\hat{\boldsymbol{\theta}}_k)(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}),$$

a cui corrisponde la seguente approssimazione della stima di massima verosimiglianza

$$\hat{\boldsymbol{\theta}} \doteq \left\{\sum_{k=1}^K i_k(\hat{\boldsymbol{\theta}}_k)\right\}^{-1} \left\{\sum_{k=1}^K i_k(\hat{\boldsymbol{\theta}}_k)(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta})\right\},$$

che coincide con $\hat{\boldsymbol{\theta}}_{wcd}$.

Come visto in precedenza ciò non può essere fatto con la log-verosimiglianza e lo *score* modificato con la correzione di Firth. Per imitare con le stime di Firth la procedura appena mostrata per le stime di massima verosimiglianza si definisce

$$\tilde{\ell}_k(\boldsymbol{\theta}) = \ell_k(\boldsymbol{\theta}) + \frac{1}{K}M(\boldsymbol{\theta}),$$

dove $M(\boldsymbol{\theta}) = \frac{1}{2}\log|i(\boldsymbol{\theta})|$ è la correzione di Firth. In questo modo è possibile scrivere la log-verosimiglianza corretta di Firth come somma di K componenti

$$\ell^*(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) + M(\boldsymbol{\theta}) = \sum_{k=1}^K \tilde{\ell}_k(\boldsymbol{\theta}).$$

Lo sviluppo di Taylor di $\tilde{\ell}_k(\boldsymbol{\theta})$ attorno a $\hat{\boldsymbol{\theta}}_k^*$, che massimizza $\ell_k^*(\boldsymbol{\theta})$ e non $\tilde{\ell}_k(\boldsymbol{\theta})$, è

$$\tilde{\ell}_k(\boldsymbol{\theta}) \doteq \tilde{\ell}_k(\hat{\boldsymbol{\theta}}_k^*) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_k^*)^\top \left\{\frac{\partial}{\partial\boldsymbol{\theta}}\tilde{\ell}_k(\boldsymbol{\theta})\right\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_k^*} + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_k^*)^\top \left\{\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\tilde{\ell}_k(\boldsymbol{\theta})\right\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_k^*}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_k^*). \tag{3.10}$$

La derivata prima di $\tilde{\ell}_k(\boldsymbol{\theta})$ è

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\theta}}\tilde{\ell}_k(\boldsymbol{\theta}) &= U_k(\boldsymbol{\theta}) + \frac{1}{2K}\frac{\partial}{\partial\boldsymbol{\theta}}\log|i(\boldsymbol{\theta})| \\ &= U_k(\boldsymbol{\theta}) + \frac{1}{2K}\frac{\partial}{\partial\boldsymbol{\theta}}\log|i(\boldsymbol{\theta})| \pm \frac{1}{2}\frac{\partial}{\partial\boldsymbol{\theta}}\log|i_k(\boldsymbol{\theta})| \\ &= \underbrace{U_k(\boldsymbol{\theta}) + \frac{1}{2}\frac{\partial}{\partial\boldsymbol{\theta}}\log|i_k(\boldsymbol{\theta})|}_{U_k^*(\boldsymbol{\theta})} + \frac{1}{2}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{1}{K}\log|i(\boldsymbol{\theta})| - \log|i_k(\boldsymbol{\theta})|\right) \\ &= U_k^*(\boldsymbol{\theta}) + \frac{1}{2}\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{1}{K}\log|i(\boldsymbol{\theta})| - \log|i_k(\boldsymbol{\theta})|\right)\end{aligned}$$

e, siccome $U_k^*(\hat{\boldsymbol{\theta}}_k^*) = \mathbf{0}$,

$$\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\tilde{\ell}_k(\boldsymbol{\theta})\right\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_k^*} = \frac{1}{2}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{1}{K}\log|i(\boldsymbol{\theta})| - \log|i_k(\boldsymbol{\theta})|\right)\right\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_k^*}$$

Nel Paragrafo 3.2 è stato mostrato che la correzione di Firth sommata allo *score* è di ordine $O(1)$ e lo stesso vale per la sua derivata (mentre lo *score* e l'informazione attesa del k-esimo sottoinsieme sono di ordine $O(n_k)$). La derivata seconda di $\tilde{\ell}_k(\boldsymbol{\theta})$ è pari a

$$-i_k(\boldsymbol{\theta}) + \frac{1}{K}\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}M(\boldsymbol{\theta}) = -i_k(\boldsymbol{\theta}) + O(1).$$

Il calcolo di $\frac{1}{K}\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}M(\boldsymbol{\theta})$ coinvolge quantità, come $\frac{\partial}{\partial\boldsymbol{\theta}}\left(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}\right)^{-1}$, il cui calcolo richiede l'insieme di tutti i dati. Perciò si preferisce approssimare la derivata seconda presente nel terzo addendo dello sviluppo con $-i_k(\boldsymbol{\theta})$. Nel prossimo paragrafo verrà mostrato che l'approssimazione utilizzata non impedisce di ottenere delle stime alla Wald che combinano stime di Firth con performance almeno paragonabili a quelle che combinano le stime di Wald. In Appendice A vengono comunque presentati i calcoli necessari per ottenere $\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}M(\boldsymbol{\theta})$.

Sfruttando i risultati riguardanti $\frac{\partial}{\partial\boldsymbol{\theta}}\tilde{\ell}_k(\boldsymbol{\theta})$ e $\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\tilde{\ell}_k(\boldsymbol{\theta})$ è possibile riscrivere la (3.10) come

$$\begin{aligned}\tilde{\ell}_k(\boldsymbol{\theta}) \doteq \tilde{c} &+ \frac{1}{2}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_k^*\right)^\top\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{1}{K}\log|i(\boldsymbol{\theta})| - \log|i_k(\boldsymbol{\theta})|\right)\right\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_k^*} - \\ &- \frac{1}{2}\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_k^*\right)^\top i_k(\hat{\boldsymbol{\theta}}_k^*)\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_k^*\right)\end{aligned}$$

dove $\tilde{c}$ è una costante indipendente da $\boldsymbol{\theta}$, e di conseguenza si può approssimare $\frac{\partial}{\partial\boldsymbol{\theta}}\ell^*(\boldsymbol{\theta})$

con

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\theta}}\ell^*(\boldsymbol{\theta}) &= \frac{\partial}{\partial\boldsymbol{\theta}}\sum_{k=1}^{K}\tilde{\ell}_k(\boldsymbol{\theta}) \\ &\doteq \underbrace{\frac{1}{2}\sum_{k=1}^{K}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{1}{K}\log|i(\boldsymbol{\theta})| - \log|i_k(\boldsymbol{\theta})|\right)\right\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_k^*}}_{B(\hat{\boldsymbol{\theta}}_k^*)} - \sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}_k^*)\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_k^*\right).\end{aligned} \tag{3.11}$$

La stima di Firth su tutti i dati si trova massimizzando $\ell^*(\boldsymbol{\theta})$ e ponendo (3.11) pari a $\mathbf{0}$ se ne trova una sua approssimazione

$$\hat{\boldsymbol{\theta}}_{wm}^* = \left\{\sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}_k^*)\right\}^{-1} B(\hat{\boldsymbol{\theta}}_k^*) + \left\{\sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}_k^*)\right\}^{-1}\left\{\sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}_k^*)\hat{\boldsymbol{\theta}}_k^*\right\},$$

che da qui in avanti verrà chiamata stima alla Wald migliorata. È interessante notare come $B(\boldsymbol{\theta})$ sia la differenza tra la correzione che trasforma $U(\boldsymbol{\theta})$ in $U^*(\boldsymbol{\theta})$ e la somma delle correzioni che trasformano $U_k(\boldsymbol{\theta})$ in $U_k^*(\boldsymbol{\theta})$. Come già visto in precedenza, la correzione di Firth su tutto l'insieme dei dati è

$$\frac{1}{2}\frac{\partial}{\partial\boldsymbol{\theta}}\log|i(\boldsymbol{\theta})| = \sum_{i=1}^{n} h_i\left(\frac{1}{2}-\pi_i\right)\boldsymbol{x}_i$$

e quella nel k-esimo gruppo è

$$\frac{1}{2}\frac{\partial}{\partial\boldsymbol{\theta}}\log|i_k(\boldsymbol{\theta})| = \sum_{i=1}^{n_k} h_i^{(k)}\left(\frac{1}{2}-\pi_i\right)\boldsymbol{x}_i,$$

che differiscono perché h_i è l'elemento diagonale della matrice $\boldsymbol{H}$ relativa a tutte n le osservazioni, mentre $h_i^{(k)}$ è l'elemento diagonale della matrice $\boldsymbol{H}_k$ calcolata solo sulle n_k osservazioni del k-esimo gruppo. Per la prima delle due correzioni è necessario calcolare la matrice $\boldsymbol{H}$ e per fare ciò è possibile replicare quanto fatto in (3.9), ma utilizzando le stime locali $\hat{\boldsymbol{\theta}}_k^*$ al posto di $\hat{\boldsymbol{\theta}}_{wcd}$. Chiaramente tutte le stime alla Rao riguardanti le stime di Firth viste finora possono essere ripetute sostituendo l'originale stima alla Wald con quella appena definita.

Negli ultimi due paragrafi è stata evidenziata la differenza tra la correzione di Firth su tutti i dati, pari a $\frac{1}{2}\log|i(\boldsymbol{\theta})| = \frac{1}{2}\log|\sum_{k=1}^{K} i_k(\boldsymbol{\theta})|$, e la somma delle correzioni di Firth dei sottoinsiemi, pari a $\sum_{k=1}^{K}\frac{1}{2}\log|i_k(\boldsymbol{\theta})|$. Visto che la correzione di Firth può essere interpretata da un punto di vista bayesiano come una priori di Jeffreys, può

essere interessante anche osservare che la differenza sopracitata implica che per ottenere la stessa informazione a priori che si definisce con la priori di Jeffreys su tutti i dati, in ciascun sottogruppo non bisogna utilizzare la priori di Jeffreys relativa al sottogruppo, ma una distribuzione a priori pari alla priori di Jeffreys su tutti i dati elevata a $1/K$. Questa scelta per la priori può essere ritrovata anche in altri lavori, ad esempio in Jordan et al. (2019), in cui viene proposto un'altra procedura per eseguire inferenza distribuita su un parametro, sia dal punto di vista frequentista che bayesiano.

3.5 Confronti finali

Analogamente a quanto fatto nel Paragrafo 3.3 vengono presentati i comportamenti delle stime combinate al variare del numero di sottoinsiemi K (5, 10, 20 e 40). Siccome è già stata mostrata la netta differenza in termini di distorsione, le stime alla Rao sono analizzate separatamente da quelle alla Wald, che non hanno performance soddisfacenti come stime finali, ma sono importanti in quanto punto di partenza per il passo Newton-Raphson che porta alla stima alla Rao.

In aggiunta alla distorsione delle stime vengono monitorate anche la loro variabilità e la copertura empirica di intervalli di confidenza alla Wald (basati su stime alla Rao). Per costruire gli intervalli di confidenza è necessario definire una stima delle matrici di varianze e covarianze degli stimatori combinati. Come proposto in Zhou & Song (2017), è stato deciso di approssimare la matrice di varianze e covarianze utilizzando le matrici di informazione attesa dei K sottoinsiemi valutate nelle rispettive stime locali. Perciò la matrice di varianze e covarianze per le stime combinate basate sulle stime di massima verosimiglianza è

$$\hat{\boldsymbol{V}} = \left(\sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}_k) \right)^{-1},$$

mentre quella per le stime che combinano stime locali di Firth è

$$\hat{\boldsymbol{V}}^* = \left(\sum_{k=1}^{K} i_k(\hat{\boldsymbol{\theta}}_k^*) \right)^{-1}.$$

Siano $\hat{v}_{rr}$ e $\hat{v}^*_{rr}$ gli r-esimi elementi diagonali rispettivamente di $\hat{\boldsymbol{V}}$ e $\hat{\boldsymbol{V}}^*$ e $\hat{\theta}_{rcd,r}$ e $\hat{\theta}^*_{ct,r}$ le stime alla Rao di massima verosimiglianza e di Firth CT per θ_r, allora gli intervalli di confidenza di livello 1 - α alla Wald sono

$$\hat{\theta}_{rcd,r} \pm z_{1-\alpha/2}\sqrt{\hat{v}_{rr}} \qquad \text{e} \qquad \hat{\theta}^*_{ct,r} \pm z_{1-\alpha/2}\sqrt{\hat{v}^*_{rr}}, \tag{3.12}$$

dove $z_{1-\alpha/2}$ è il quantile $1-\alpha/2$ della variabile casuale normale standard. Analoghi intervalli di confidenza possono essere calcolati per le altre stime combinate, semplicemente sostituendo $\hat{\theta}_{rcd,r}$ o $\hat{\theta}^*_{ct,r}$ con la stima desiderata.

3.5.1 Confronti delle stime alla Wald

Le Figure 3.9, 3.10, 3.11 e 3.12 mostrano i confronti tra le stime alla Wald già analizzate nel Paragrafo 3.3 e la nuova stima alla Wald proposta per combinare le stime di Firth dei sottoinsiemi.

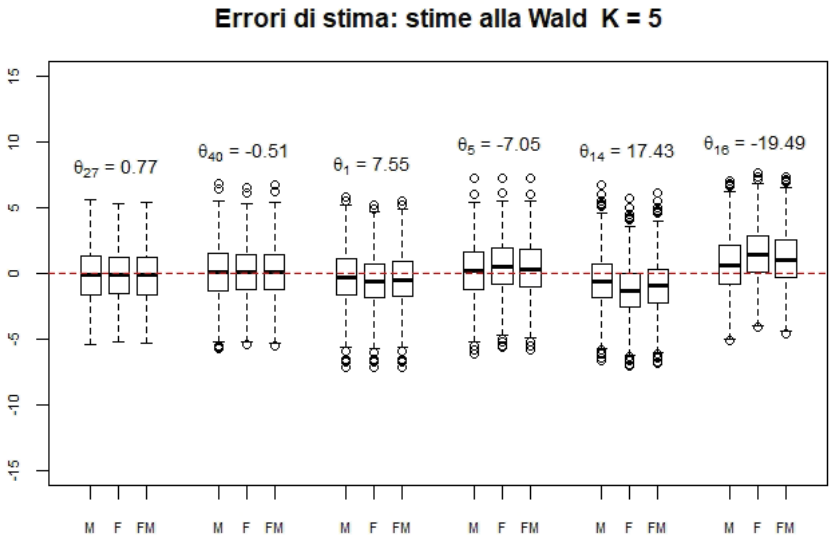

FIGURA 3.9: Boxplot degli errori di stima delle stime combinate alla Wald, $K = 5$ (M: MLE, F: Firth, FM: Firth Migliorato).

Da questi boxplot è possibile evincere che anche per le stime alla Wald migliorate la distorsione aumenta sia con il valore assoluto del vero parametro sia con il numero di sottoinsiemi in cui sono divise le osservazioni. È però interessante notare che, al contrario di quanto accadeva con le prime stime alla Wald di Firth, il comportamento di queste nuove stime è pressoché identico a quello delle stime che combinano stime locali di massima verosimiglianza per K tra 5 e 20 ed è nettamente migliore per K pari a 40, situazione in cui la distorsione aumenta notevolmente ed inizia a presentarsi il fenomeno di stime infinite nel caso di massima verosimiglianza. Questo risultato è importante, infatti l'iniziale proposta per combinare alla Wald le stime di Firth per $K = 40$ risultava migliore delle stime alla Wald che combinano le stime di massima verosimiglianza, ma al prezzo di un comportamento sensibilmente peggiore per $K < 40$. Questo andamento comportava l'obbligo di una diversa scelta per lo stimatore locale a seconda del numero

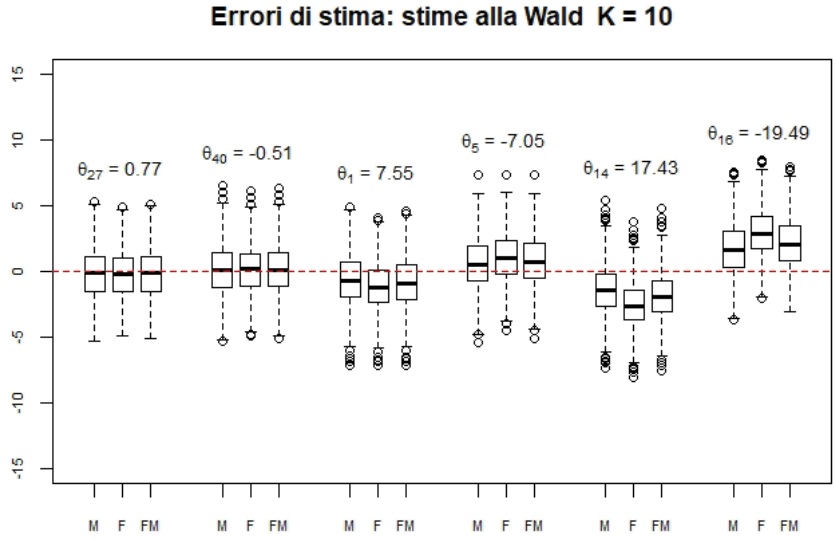

FIGURA 3.10: Boxplot degli errori di stima delle stime combinate alla Wald, $K = 10$ (M: MLE, F: Firth, FM: Firth Migliorato).

di sottoinsiemi, invece la nuova proposta, non solo migliora nettamente le stime per $K = 40$, ma permette anche di poter scegliere come stimatore locale quello di Firth indipendentemente dal numero di sottoinsiemi.

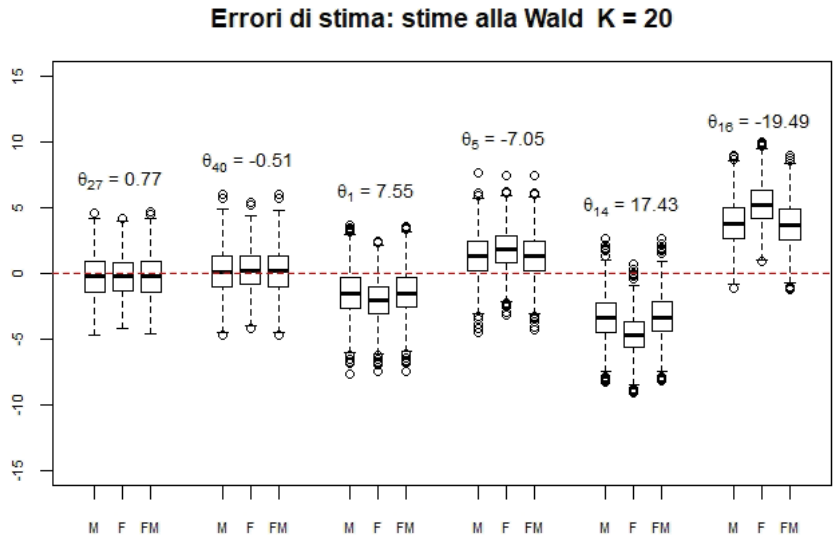

FIGURA 3.11: Boxplot degli errori di stima delle stime combinate alla Wald, $K = 20$ (M: MLE, F: Firth, FM: Firth Migliorato).

La Tabella 3.6 riporta le distorsioni, le distorsioni relative e gli standard error delle stime alla Wald. La tabella in alto riguardante la distorsione e quella in centro riguardante la distorsione relativa confermano che per $K < 40$ la stima alla Wald migliorata

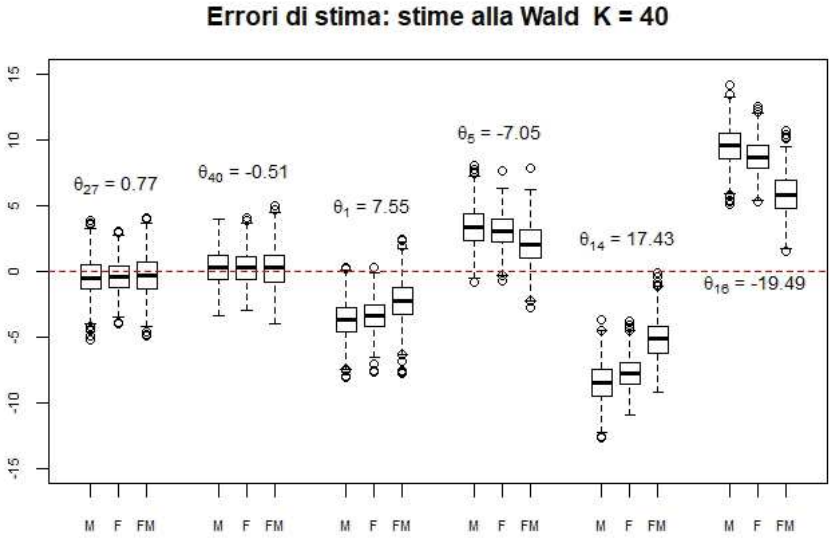

FIGURA 3.12: Boxplot degli errori di stima delle stime combinate alla Wald, $K = 40$ (M: MLE, F: Firth, FM: Firth Migliorato).

$\hat{\boldsymbol{\theta}}^*_{wm}$ ha distorsione leggermente maggiore, ma paragonabile, della stima combinata riguardante le stime di massima verosimiglianza, mentre quest'ultime quando $K = 40$ hanno distorsione pari a circa una volta e mezza quella della stima che combina le stime locali di Firth. La stima (non migliorata) che combina le stime di Firth dei sottoinsiemi $\hat{\boldsymbol{\theta}}^*_{wcd}$ ha distorsione maggiore delle altre due per ogni parametro per $K < 40$, mentre per $K = 40$ ha distorsione intermedia. La tabella in basso invece mostra che lo standard error di tutte le stime diminuisce al crescere di K, al contrario di quanto accade per la distorsione, ma non sembra esserci alcuna dipendenza dal valore del parametro. Lo conferma la Figura 3.13 (in cui è stata omessa $\hat{\boldsymbol{\theta}}^*_{wcd}$), dove si vede che la distribuzione dei punti è casuale per tutti i valori di K. Si nota in particolare la serie di punti corrispondenti ai parametri con vero valore pari a 0, che copre tutto il range delle ordinate nonostante il vero valore del parametro sia lo stesso; va comunque tenuto presente che la differenza tra il minimo e il massimo valore di standard error è piuttosto piccola.

3.5.2 Confronti delle stime alla Rao

In questo paragrafo vengono presentati i risultati delle simulazioni riguardanti le stime combinate alla Rao. Sono prese in considerazione quattro diverse stime: quella già vista che combina stime locali di massima verosimiglianza, quella già vista che combina stime locali di Firth, la stima alla Rao Correzione Totale di Firth e la versione approssimata di quest'ultima (in cui vengono utilizzate le stime locali $\hat{\boldsymbol{\theta}}^*_k$ al posto di $\hat{\boldsymbol{\theta}}^*_{wcd}$). Nei grafici sono rispettivamente indicate con M, F, CT e $\widetilde{\text{CT}}$. Siccome nel paragrafo precedente

TABELLA 3.6: Distorsione, distorsione relativa e standard error delle stime combinate alla Wald al variare di K (F: Firth, FM : Firth Migliorato).

Distorsione

	$K=5$			$K=10$			$K=20$			$K=40$		
	MLE	F	FM	MLE	F	FM	MLE	F	FM	MLE	F	FM
θ_{27}	-0.17	-0.19	-0.18	-0.19	-0.22	-0.20	-0.24	-0.28	-0.24	-0.45	-0.40	-0.31
θ_{40}	0.10	0.12	0.11	0.12	0.15	0.13	0.17	0.20	0.17	0.31	0.28	0.22
θ_1	-0.28	-0.60	-0.45	-0.67	-1.16	-0.86	-1.48	-2.05	-1.45	-3.69	-3.37	-2.27
θ_5	0.19	0.49	0.35	0.56	1.04	0.74	1.34	1.89	1.31	3.41	3.12	2.07
θ_{14}	-0.54	-1.28	-0.94	-1.45	-2.60	-1.88	-3.35	-4.68	-3.28	-8.50	-7.77	-5.19
θ_{16}	0.69	1.51	1.13	1.69	2.97	2.17	3.80	5.28	3.73	9.56	8.73	5.87

Distorsione relativa

	$K=5$			$K=10$			$K=20$			$K=40$		
	MLE	F	FM	MLE	F	FM	MLE	F	FM	MLE	F	FM
θ_{27}	-0.22	-0.24	-0.23	-0.25	-0.29	-0.26	-0.32	-0.37	-0.32	-0.58	-0.52	-0.41
θ_{40}	0.19	0.23	0.22	0.24	0.30	0.26	0.34	0.40	0.33	0.60	0.56	0.43
θ_1	-0.04	-0.08	-0.06	-0.09	-0.15	-0.11	-0.20	-0.27	-0.19	-0.49	-0.45	-0.30
θ_5	0.03	0.07	0.05	0.08	0.15	0.10	0.19	0.27	0.19	0.48	0.44	0.29
θ_{14}	-0.03	-0.07	-0.05	-0.08	-0.15	-0.11	-0.19	-0.27	-0.19	-0.49	-0.45	-0.30
θ_{16}	0.04	0.08	0.06	0.09	0.15	0.11	0.19	0.27	0.19	0.49	0.45	0.30

Standard error

	$K=5$			$K=10$			$K=20$			$K=40$		
	MLE	F	FM	MLE	F	FM	MLE	F	FM	MLE	F	FM
θ_{27}	2.06	1.97	2.01	1.95	1.81	1.90	1.72	1.55	1.72	1.37	1.21	1.53
θ_{40}	2.08	1.99	2.03	1.97	1.83	1.92	1.75	1.58	1.75	1.41	1.23	1.55
θ_1	2.10	2.00	2.05	1.99	1.84	1.93	1.78	1.59	1.77	1.44	1.24	1.57
θ_5	2.10	2.01	2.05	1.99	1.85	1.94	1.78	1.60	1.77	1.41	1.24	1.57
θ_{14}	2.08	1.99	2.03	1.97	1.83	1.91	1.75	1.57	1.74	1.44	1.23	1.55
θ_{16}	2.06	1.97	2.01	1.96	1.81	1.90	1.74	1.56	1.74	1.38	1.21	1.53

per le stime di Firth è stato mostrato il vantaggio della stima alla Wald migliorata rispetto a quella originale, i risultati di questo paragrafo sono relativi a stime alla Rao che aggiornano la stima alla Wald migliorata.

Nelle Figure 3.14, 3.15, 3.16 e 3.17 si nota che le prime stime alla Rao di Firth non solo non sono paragonabili a quelle che combinano le stime di massima verosimiglianza, ma neanche alle nuove stime proposte per combinare le stime locali di Firth. Infatti, quest'ultime hanno distorsioni basse, molto simili a quelle di $\hat{\boldsymbol{\theta}}_{rcd}$, in alcuni casi anche migliori. Per K piccoli, cioè 5 e 10, l'approssimazione della stima Rao CT è per alcuni parametri anche la stima migliore in termini di distorsione, come si può vedere nella Figura 3.18. Dallo stesso grafico, e dalle Tabelle 3.7 in alto e in centro, però si evince

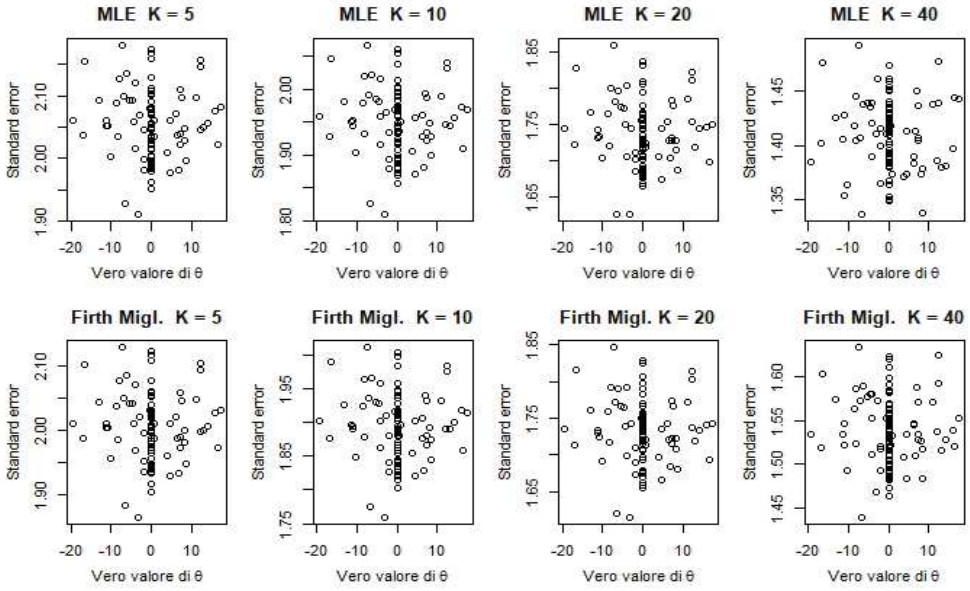

FIGURA 3.13: Relazione tra il vero valore del parametro e lo standard error.

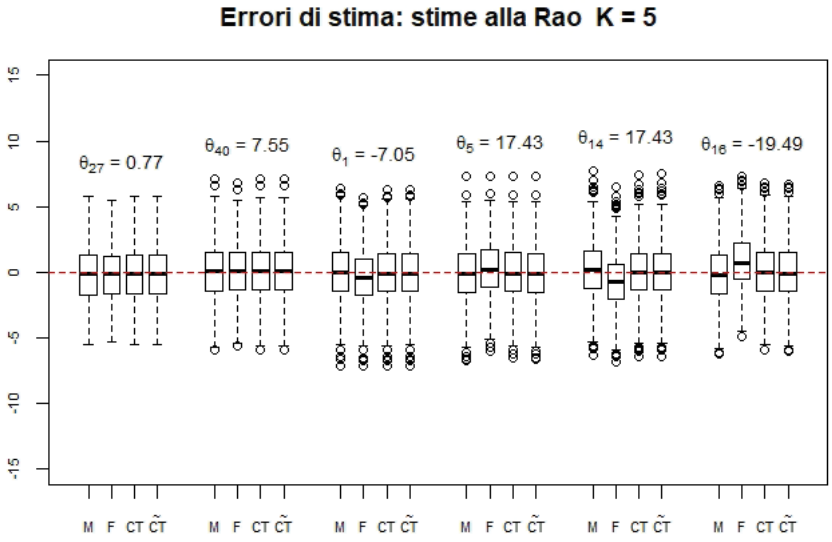

FIGURA 3.14: Boxplot degli errori di stima delle stime combinate alla Rao, $K = 5$ (M: MLE, F: Firth, CT: Correzione Totale, $\widetilde{\text{CT}}$: Correzione Totale con $\hat{\boldsymbol{\theta}}^*_k$).

anche che per K grandi, cioè 20 e 40, la stima $\widetilde{\text{CT}}$ ha distorsioni pari a più del doppio delle stime CT e di massima verosimiglianza. Invece le stime alla Rao di massima verosimiglianza e CT hanno distorsione pressoché costante fino a $K = 20$, e crescono solo per $K = 40$. Va comunque notato che le differenze tra le tre stime fino a $K = 10$ sono nell'ordine di uno o due decimi anche per parametri nell'ordine delle decine. Per $K = 40$ la stima CT è in assoluto la migliore in termini di distorsione, infatti per i due parametri prossimi a 0 (θ_{27} e θ_{40}) è pari a quella delle stime alla Rao di massima verosimiglianza e

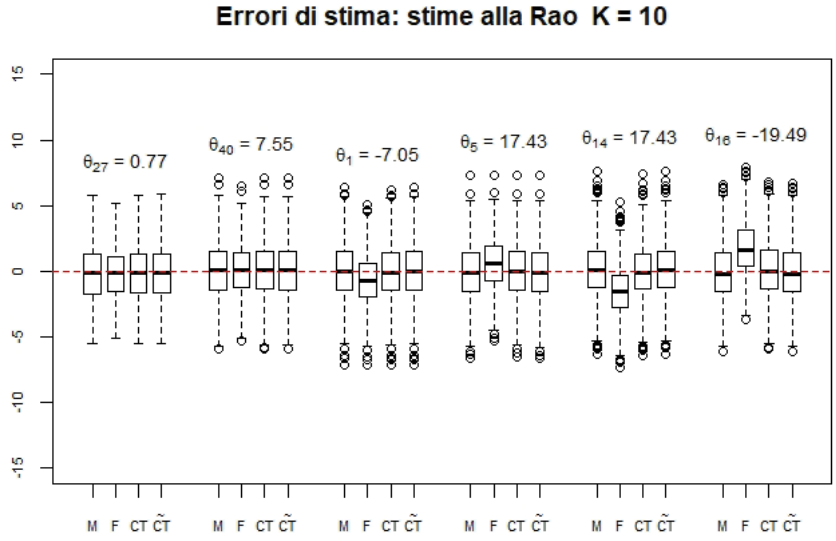

FIGURA 3.15: Boxplot degli errori di stima delle stime combinate alla Rao, $K = 10$ (M: MLE, F: Firth, CT: Correzione Totale, $\widetilde{\text{CT}}$: Correzione Totale con $\hat{\boldsymbol{\theta}}_k^*$).

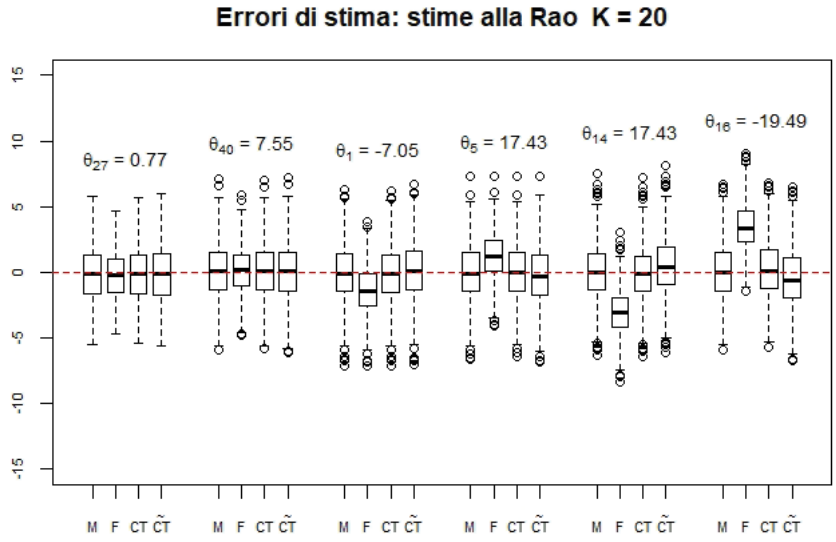

FIGURA 3.16: Boxplot degli errori di stima delle stime combinate alla Rao, $K = 20$ (M: MLE, F: Firth, CT: Correzione Totale, $\widetilde{\text{CT}}$: Correzione Totale con $\hat{\boldsymbol{\theta}}_k^*$).

per gli altri quattro parametri è sensibilmente minore. Infine un fenomeno interessante è la dimensione della distorsione rispetto al vero valore del parametro per $K < 40$, infatti si può notare in Figura 3.18 che la relazione non è monotona, ma la distorsione minore è quella relativa ai parametri con valore assoluto intermedio, mentre quella dei parametri con valore maggiore è simile a quella dei parametri circa pari a zero. Si nota che tale andamento è contrario a quello mostrato dal risultato di Cordeiro & McCullagh (1991),

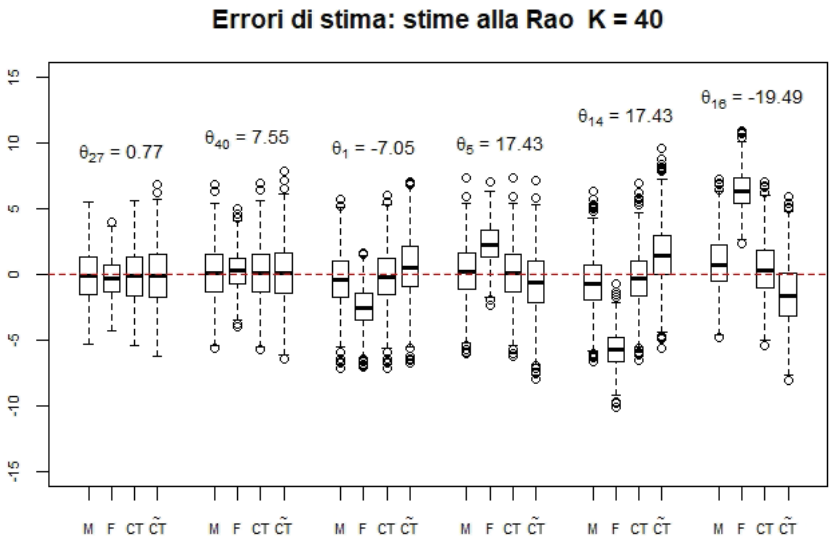

FIGURA 3.17: Boxplot degli errori di stima delle stime combinate alla Rao, $K = 40$ (M: MLE, F: Firth, CT: Correzione Totale, $\widetilde{\text{CT}}$: Correzione Totale con $\hat{\boldsymbol{\theta}}_k^*$).

che stabilisce la proporzionalità diretta approssimata della distorsione delle stime di massima verosimiglianza (non combinate) e del vero valore del parametro.

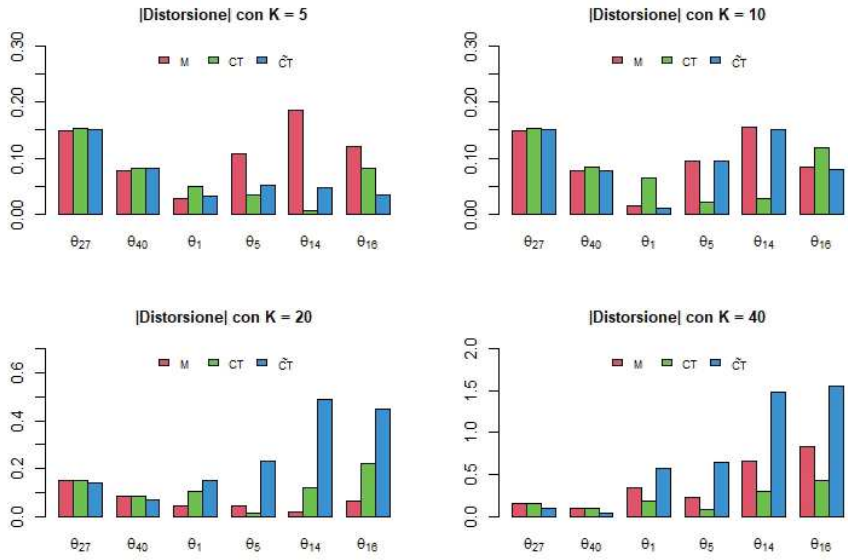

FIGURA 3.18: Distorsione in valore assoluto per stime alla Rao (M: MLE, CT: Correzione Totale, $\widetilde{\text{CT}}$: Correzione Totale con $\hat{\boldsymbol{\theta}}_k^*$).

Per quanto riguarda la variabilità (si veda la Tabella 3.7 in basso), le due stime migliori, cioè quella che combina stime locali di massima verosimiglianza e la Rao CT, hanno standard error pressoché costanti al variare di K. Per la stima $\widetilde{\text{CT}}$ il cambiamento è un po' più evidente e, a differenza di quanto accade per le altre due stime, la variabilità

cresce con il numero dei sottoinsiemi. Questo comportamento potrebbe essere attribuito all'utilizzo delle stime locali $\hat{\boldsymbol{\theta}}_k^*$ all'interno di tutti i termini che costituiscono $\widetilde{\text{CT}}$; infatti tali stime peggiorano notevolmente in termini di distorsione (come visto in Tabella 3.4) e variabilità, in misura molto maggiore delle stime $\hat{\boldsymbol{\theta}}_{wcd}$ e $\hat{\boldsymbol{\theta}}_{wm}^*$ che vengono utilizzate per costruire le altre due stime alla Rao.

TABELLA 3.7: Distorsione, distorsione relativa e standard error delle stime combinate alla Rao al variare di K (F: Firth, CT : Correzione Totale, $\widetilde{\text{CT}}$: Correzione Totale con $\hat{\boldsymbol{\theta}}_k^*$).

Distorsione in media

	$K=5$				$K=10$				$K=20$				$K=40$			
	MLE	F	CT	$\widetilde{\text{CT}}$	MLE	F	CT	$\widetilde{\text{CT}}$	MLE	F	CT	$\widetilde{\text{CT}}$	MLE	F	CT	$\widetilde{\text{CT}}$
θ_{27}	-0.15	-0.17	-0.15	-0.15	-0.15	-0.19	-0.15	-0.15	-0.15	-0.23	-0.15	-0.14	-0.15	-0.30	-0.15	-0.10
θ_{40}	0.08	0.10	0.08	0.08	0.08	0.12	0.08	0.08	0.08	0.16	0.09	0.07	0.10	0.23	0.09	0.04
θ_1	0.03	-0.35	-0.05	-0.03	0.01	-0.70	-0.06	0.01	-0.04	-1.36	-0.10	0.15	-0.34	-2.48	-0.19	0.57
θ_5	-0.11	0.25	-0.03	-0.05	-0.10	0.59	-0.02	-0.09	-0.04	1.22	0.01	-0.23	0.22	2.27	0.09	-0.64
θ_{14}	0.19	-0.69	0.01	0.05	0.16	-1.52	-0.03	0.15	0.02	-3.05	-0.12	0.49	-0.66	-5.69	-0.31	1.49
θ_{16}	-0.12	0.85	0.08	0.03	-0.08	1.78	0.12	-0.08	0.07	3.48	0.22	-0.45	0.83	6.41	0.43	-1.56

Distorsione relativa

	$K=5$				$K=10$				$K=20$				$K=40$			
	MLE	F	CT	$\widetilde{\text{CT}}$	MLE	F	CT	$\widetilde{\text{CT}}$	MLE	F	CT	$\widetilde{\text{CT}}$	MLE	F	CT	$\widetilde{\text{CT}}$
θ_{27}	-0.19	-0.22	-0.20	-0.20	-0.19	-0.24	-0.20	-0.20	-0.19	-0.30	-0.20	-0.18	-0.20	-0.39	-0.20	-0.13
θ_{40}	0.15	0.20	0.16	0.16	0.15	0.25	0.16	0.15	0.16	0.32	0.17	0.14	0.20	0.45	0.18	0.08
θ_1	0.00	-0.05	-0.01	-0.00	0.00	-0.09	-0.01	0.00	-0.01	-0.18	-0.01	0.02	-0.04	-0.33	-0.02	0.08
θ_5	-0.02	0.04	-0.00	-0.01	-0.01	0.08	-0.00	-0.01	-0.01	0.17	0.00	-0.03	0.03	0.32	0.01	-0.09
θ_{14}	0.01	-0.04	0.00	0.00	0.01	-0.09	-0.00	0.01	0.00	-0.18	-0.01	0.03	-0.04	-0.33	-0.02	0.09
θ_{16}	-0.01	0.04	0.00	0.00	-0.00	0.09	0.01	-0.00	0.00	0.18	0.01	-0.02	0.04	0.33	0.02	-0.08

Standard error

	$K=5$				$K=10$				$K=20$				$K=40$			
	MLE	F	CT	$\widetilde{\text{CT}}$	MLE	F	CT	$\widetilde{\text{CT}}$	MLE	F	CT	$\widetilde{\text{CT}}$	MLE	F	CT	$\widetilde{\text{CT}}$
θ_{27}	2.14	2.04	2.12	2.13	2.14	1.94	2.12	2.14	2.12	1.75	2.11	2.18	2.04	1.45	2.08	2.31
θ_{40}	2.17	2.06	2.15	2.16	2.17	1.96	2.15	2.17	2.15	1.77	2.13	2.21	2.07	1.46	2.11	2.34
θ_1	2.18	2.08	2.16	2.17	2.18	1.97	2.16	2.18	2.16	1.78	2.15	2.22	2.08	1.47	2.12	2.35
θ_5	2.19	2.08	2.17	2.17	2.19	1.98	2.16	2.19	2.17	1.79	2.15	2.23	2.09	1.47	2.13	2.37
θ_{14}	2.17	2.06	2.15	2.15	2.16	1.96	2.14	2.16	2.14	1.77	2.13	2.21	2.05	1.44	2.10	2.33
θ_{16}	2.15	2.04	2.13	2.13	2.15	1.94	2.12	2.15	2.13	1.75	2.11	2.19	2.03	1.43	2.08	2.32

Come ultimo confronto tra le stime alla Rao è stata paragonata la copertura empirica di intervalli di confidenza alla Wald. Nella Tabella 3.8 sono riportate le coperture empiriche per intervalli con livelli di confidenza con livello 0.95, costruiti come mostrato in (3.12). È evidente che per $\hat{\boldsymbol{\theta}}_{rcd}$ e $\hat{\boldsymbol{\theta}}_{ct}^*$ al crescere del numero di sottoinsiemi cresce anche la copertura, mentre per la versione approssimata di $\hat{\boldsymbol{\theta}}_{ct}^*$ il comportamento non è monotono. Per $K = 5$ la copertura empirica dista da quella nominale al massimo 0.008 per tutte e tre le stime, ma al crescere di K gli intervalli per la stima di massima verosimiglianza e CT diventano troppo conservativi, arrivando ad avere una copertura rispettivamente oltre al 99% e 97% per $K = 40$. Fino a $K = 20$ anche gli intervalli di

$\widetilde{\text{CT}}$ hanno copertura crescente, anche se sempre minore del 96.2%, mentre per $K = 40$ la copertura degli intervalli per i due parametri più grandi in valore assoluto (θ_{14} e θ_{16}) crolla. Siccome le stime delle varianze per CT e $\widetilde{\text{CT}}$ sono uguali, visto che sono basate sulle $\hat{\boldsymbol{\theta}}^*_k$, la differenza è dovuta alle stime dei parametri: è ragionevole pensare che la causa della forte diminuzione della copertura per $\widetilde{\text{CT}}$ sia dovuta alla sua notevole distorsione. Invece la differenza tra la stima di massima verosimiglianza e quella CT è da attribuire alla netta differenza dell'evoluzione della variabilità al crescere di K: le medie delle stime delle varianze, ottenute nei 10^3 diversi campioni, degli stimatori dei sei parametri considerati, per $\hat{\boldsymbol{\theta}}_{rcd}$ passano da $(4.77, 4.72, 4.77, 4.63, 4.84, 4.79)$ con $K = 5$ a $(8.85, 8.77, 8.86, 8.61, 8.97, 8.88)$ con $K = 40$, mentre per $\hat{\boldsymbol{\theta}}^*_{rcd}$ passano da $(4.71, 4.66, 4.71, 4.58, 4.77, 4.73)$ a $(5.90, 5.85, 5.91, 5.75, 5.95, 5.89)$, limitando così la crescita dell'ampiezza degli intervalli (e di conseguenza della copertura).

Tabella 3.8: Copertura empirica di intervalli alla Wald con stime alla Rao al variare di K (CT : Correzione Totale, $\widetilde{\text{CT}}$: Correzione Totale con $\hat{\boldsymbol{\theta}}^*_k$).

	$K=5$			$K=10$			$K=20$			$K=40$		
	MLE	CT	$\widetilde{\text{CT}}$	MLE	CT	$\widetilde{\text{CT}}$	MLE	CT	$\widetilde{\text{CT}}$	MLE	CT	$\widetilde{\text{CT}}$
θ_{27}	0.953	0.953	0.953	0.964	0.961	0.959	0.978	0.973	0.962	1.000	0.981	0.967
θ_{40}	0.954	0.956	0.955	0.960	0.959	0.956	0.974	0.966	0.957	0.998	0.978	0.957
θ_1	0.942	0.943	0.942	0.953	0.948	0.945	0.964	0.960	0.950	0.994	0.967	0.946
θ_5	0.950	0.949	0.950	0.955	0.957	0.952	0.976	0.966	0.955	0.996	0.978	0.951
θ_{14}	0.953	0.954	0.953	0.961	0.957	0.955	0.975	0.963	0.951	0.993	0.978	0.912
θ_{16}	0.954	0.956	0.956	0.961	0.959	0.957	0.976	0.969	0.959	0.993	0.977	0.907

Siccome il peggioramento delle performance degli intervalli di confidenza è attribuibile, almeno in parte, al deterioramento della stime della variabilità al crescere di K, si propone una diversa procedura per stimare la matrice di varianze e covarianze delle stime alla Rao. Per costruire la stima alla Rao CT è necessario calcolare $\hat{h}^*_i$, $i = 1, \ldots, n$, come mostrato in (3.9). In questo passaggio la matrice $\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}$ viene calcolata come la somma delle matrici $\boldsymbol{X}^\top_k \hat{\boldsymbol{W}}^*_k \boldsymbol{X}^\top_k$, che coincidono con le matrici di informazione attesa valutate in $\hat{\boldsymbol{\theta}}^*_{wcd}$ di ciascun sottoinsieme. Perciò la matrice di varianze e covarianze per le stime alla Rao, può essere approssimata da

$$\left(\sum_{k=1}^{K} \boldsymbol{X}^\top_k \hat{\boldsymbol{W}}^*_k \boldsymbol{X}^\top_k \right)^{-1}.$$

In questo modo le stime delle varianze rimangono pressoché costanti al variare di K, sia per lo stimatore alla Rao di verosimiglianza che per quello di Firth CT. Per il primo stimatore dei sei parametri considerati finora le varianze medie sono pari a (4.53, 4.49,

4.53, 4.40, 4.60, 4.55) con $K = 5$ e (4.23, 4.19, 4.24, 4.12, 4.25, 4.20) con $K = 40$. Per gli stimatori di Firth analogamente le medie sono (4.51, 4.47, 4.52, 4.39, 4.58, 4.53) e (4.33, 4.29, 4.34, 4.22, 4.37, 4.32) per K rispettivamente pari a 5 e 40. A differenza di quanto accadeva con le precedenti stime di variabilità, queste non crescono con il numero di sottoinsiemi, anzi c'è una diminuzione, seppure leggera.

Nella Tabella 3.9 vengono presentate le probabilità di copertura empiriche al variare di K per le stime alla Rao di massima verosimiglianza, Correzione Totale e Correzione Totale approssimata, ottenute utilizzando le nuove stime di variabilità. I risultati sono molto buoni per le prime due stime per tutti i valori di K, mentre per la stima $\widetilde{\text{CT}}$ il peggioramento per i K alti è ancora più evidente di prima, poiché le varianze più basse mettono ancora più in evidenza la distorsione che la caratterizza.

TABELLA 3.9: Copertura empirica di intervalli alla Wald con stime alla Rao al variare di K con stima delle varianze in $\hat{\boldsymbol{\theta}}^*_{wcd}$ (CT : Correzione Totale, $\widetilde{\text{CT}}$: Correzione Totale con $\hat{\boldsymbol{\theta}}^*_k$).

	$K=5$			$K=10$			$K=20$			$K=40$		
	MLE	CT	$\widetilde{\text{CT}}$	MLE	CT	$\widetilde{\text{CT}}$	MLE	CT	$\widetilde{\text{CT}}$	MLE	CT	$\widetilde{\text{CT}}$
θ_{27}	0.950	0.950	0.950	0.948	0.950	0.948	0.947	0.950	0.942	0.950	0.950	0.928
θ_{40}	0.948	0.949	0.949	0.945	0.948	0.946	0.944	0.948	0.939	0.952	0.948	0.918
θ_1	0.937	0.937	0.937	0.935	0.937	0.935	0.936	0.937	0.931	0.943	0.940	0.904
θ_5	0.941	0.943	0.942	0.940	0.942	0.940	0.939	0.941	0.932	0.946	0.943	0.887
θ_{14}	0.946	0.949	0.947	0.945	0.945	0.944	0.943	0.943	0.928	0.933	0.939	0.865
θ_{16}	0.953	0.952	0.952	0.952	0.951	0.952	0.950	0.951	0.940	0.934	0.948	0.859

Infine, dopo avere confrontato tra loro le stime combinate, è di interesse valutare le differenze tra le stime combinate e quelle che si otterrebbero sull'insieme dei dati intero. I confronti in termini di distorsione e variabilità sono positivi, infatti, come si può vedere dalla Tabella 3.10, fino a $K = 20$ le distorsioni delle stime combinate sono in linea con quelle delle stime sugli insiemi interi e per $K = 40$, a cui corrisponde un rapporto p/n pari a 0.4, solo le distorsioni delle stime combinate per i parametri di medio e alto valore assoluto si discostano e comunque non di molto, infatti la differenza in termini di distorsione relativa è molto piccola.

Per quanto riguarda la variabilità, nonostante le differenze siano molto piccole, i risultati sono ancora migliori, infatti per tutti i parametri e valori di K gli standard error delle stime combinate sono non maggiori di quelli delle corrispondenti stime intere.

Per quanto riguarda gli intervalli di confidenza, le coperture empiriche degli intervalli alla Wald con livello pari a 0.95 costruiti con le stime relative all'intero insieme dei dati sono riportate nella Tabella 3.11, dove sono confrontate con le coperture delle stime

TABELLA 3.10: Distorsione, distorsione relativa e standard error delle stime combinate alla Rao e delle stime sui dataset interi (F: Firth, CT: Correzione Totale).

Distorsione

	Interi		$K=5$		$K=10$		$K=20$		$K=40$	
	MLE	F	MLE	CT	MLE	CT	MLE	CT	MLE	CT
θ_{27}	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15	-0.15
θ_{40}	0.08	0.08	0.08	0.08	0.08	0.08	0.08	0.09	0.10	0.09
θ_1	0.03	-0.05	0.03	-0.05	0.01	-0.06	-0.04	-0.10	-0.34	-0.19
θ_5	-0.11	-0.04	-0.11	-0.03	-0.10	-0.02	-0.04	0.01	0.22	0.09
θ_{14}	0.20	0.01	0.19	0.01	0.16	-0.03	0.02	-0.12	-0.66	-0.31
θ_{16}	-0.13	0.07	-0.12	0.08	-0.08	0.12	0.07	0.22	0.83	0.43

Distorsione relativa

	Interi		$K=5$		$K=10$		$K=20$		$K=40$	
	MLE	F	MLE	CT	MLE	CT	MLE	CT	MLE	CT
θ_{27}	-0.19	-0.20	-0.19	-0.20	-0.19	-0.20	-0.19	-0.20	-0.20	-0.20
θ_{40}	0.15	0.16	0.15	0.16	0.15	0.16	0.16	0.17	0.20	0.18
θ_1	0.00	-0.01	0.00	-0.01	0.00	-0.01	-0.01	-0.01	-0.04	-0.02
θ_5	-0.02	-0.01	-0.02	-0.00	-0.01	-0.00	-0.01	0.00	0.03	0.01
θ_{14}	0.01	0.00	0.01	0.00	0.01	-0.00	0.00	-0.01	-0.04	-0.02
θ_{16}	-0.01	0.00	-0.01	0.00	-0.00	0.01	0.00	0.01	0.04	0.02

Standard error

	Interi		$K=5$		$K=10$		$K=20$		$K=40$	
	MLE	F	MLE	CT	MLE	CT	MLE	CT	MLE	CT
θ_{27}	2.15	2.12	2.14	2.12	2.14	2.12	2.12	2.11	2.04	2.08
θ_{40}	2.17	2.15	2.17	2.15	2.17	2.15	2.15	2.13	2.07	2.11
θ_1	2.19	2.16	2.18	2.16	2.18	2.16	2.16	2.15	2.08	2.12
θ_5	2.19	2.17	2.19	2.17	2.19	2.16	2.17	2.15	2.09	2.13
θ_{14}	2.17	2.15	2.17	2.15	2.16	2.14	2.14	2.13	2.05	2.10
θ_{16}	2.15	2.13	2.15	2.13	2.15	2.12	2.13	2.11	2.03	2.08

combinate alla Rao di massima verosimiglianza e Correzione Totale, che sono già state mostrate in precedenza. Anche in questo caso le stime combinate si comportano molto bene, avendo probabilità di copertura che differiscono da quelle delle stime su tutti i dati per non più di 3 millesimi nella maggior parte delle situazioni. Le differenze più ampie si osservano per $K = 40$: per i parametri θ_{14} e θ_{16}, che sono quelli maggiori in valore assoluto, la copertura della stima combinata di massima verosimiglianza dista da quelle sui dataset interi, e dal livello nominale, di quasi due centesimi, mentre la stima CT ha prestazioni peggiori solo per θ_{14}, mantenendo invece una probabilità di copertura molto prossima a 0.95 per θ_{16}.

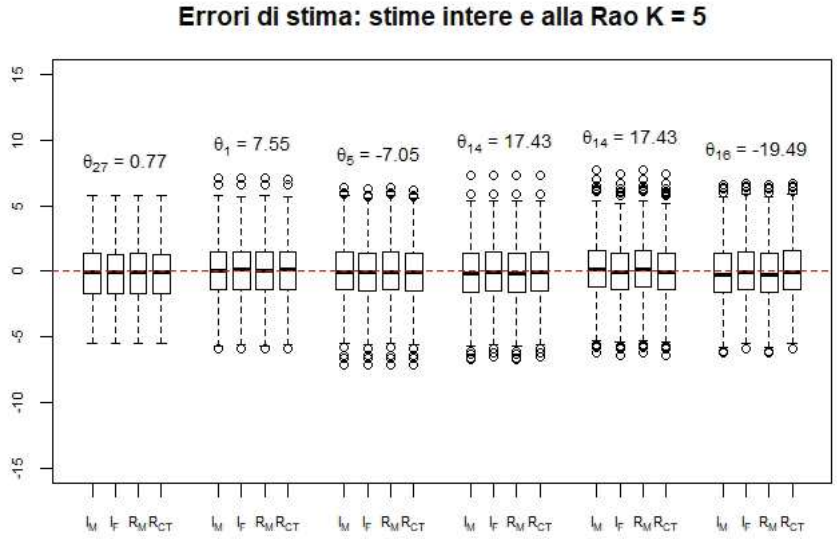

FIGURA 3.19: Boxplot degli errori di stima delle stime sui dataset interi e combinate alla Rao, $K = 5$ (I_M: MLE su dati interi, I_F: Firth su dati interi, R_M: Rao massima verosimiglianza, R_{CT}: Rao Correzione Totale).

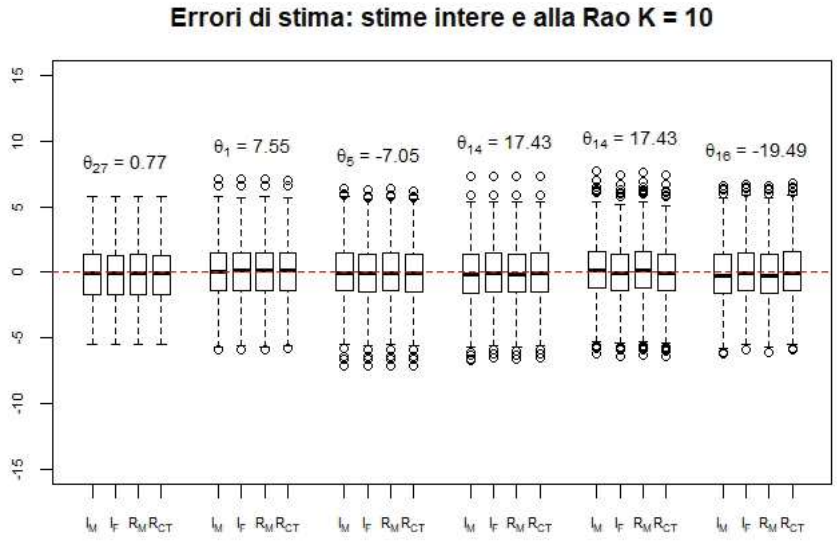

FIGURA 3.20: Boxplot degli errori di stima delle stime combinate alla Rao, $K = 10$ (I_M: MLE su dati interi, I_F: Firth su dati interi, R_M: Rao massima verosimiglianza, R_{CT}: Rao Correzione Totale).

Complessivamente le performance delle stime combinate si possono ritenere soddisfacenti, visto che, anche per sottoinsiemi con numerosità bassa rispetto al numero di variabili (ma sempre con $n > p$), hanno risultati paragonabili a quelli che si ottengono stimando i parametri nella situazione ottimale con un alto numero di osservazioni e un rapporto p/n pari a 0.01.

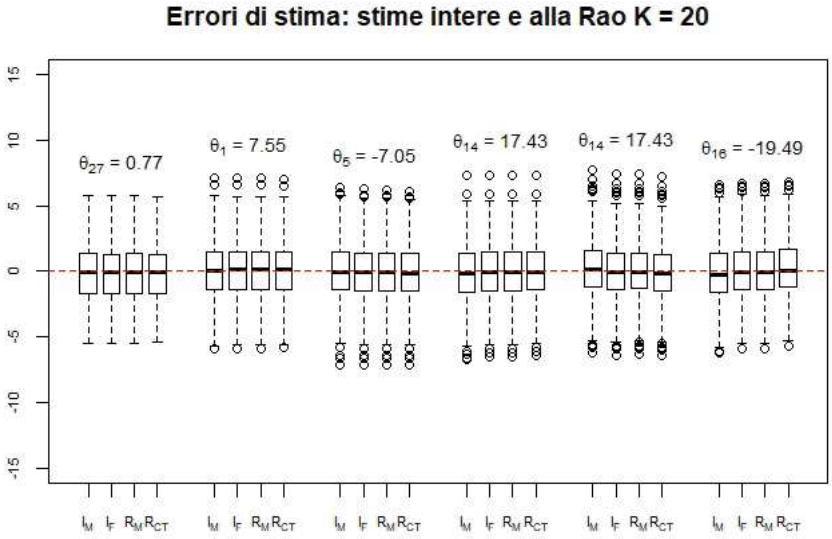

FIGURA 3.21: Boxplot degli errori di stima delle stime combinate alla Rao, $K = 20$ (I_M: MLE su dati interi, I_F: Firth su dati interi, R_M: Rao massima verosimiglianza, R_{CT}: Rao Correzione Totale).

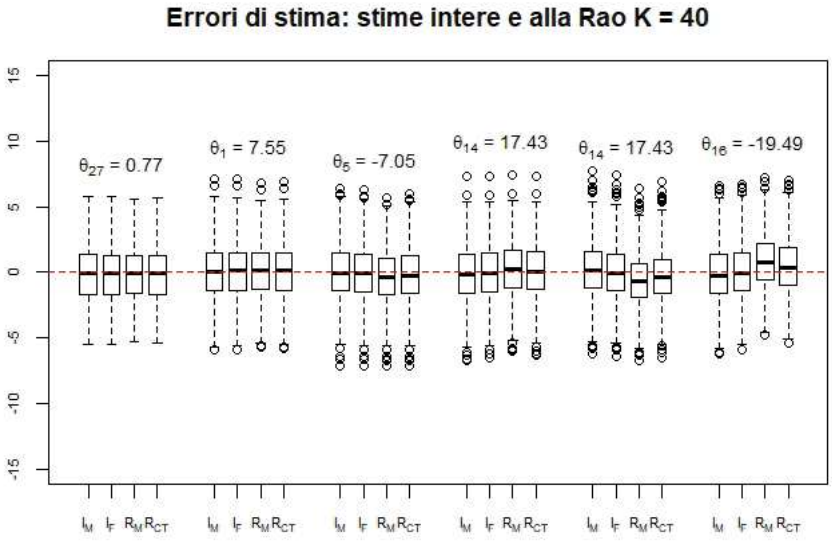

FIGURA 3.22: Boxplot degli errori di stima delle stime combinate alla Rao, $K = 40$ (I_M: MLE su dati interi, I_F: Firth su dati interi, R_{CT}: Rao massima verosimiglianza, R_{CT}: Rao Correzione Totale).

3.6 Interpretazione bayesiana delle stime combinate alla Wald

In Winkler (1981) viene proposta una procedura bayesiana per combinare stime distribuite di un comune parametro scalare θ. La metodologia proposta assume che in

TABELLA 3.11: Copertura empirica di intervalli alla Wald con stime su dataset interi e stime alla Rao al variare di K con stima delle varianze in $\hat{\boldsymbol{\theta}}^*_{wcd}$ (F: Firth, CT : Correzione Totale).

	Interi		$K=5$		$K=10$		$K=20$		$K=40$	
	MLE	F	MLE	CT	MLE	CT	MLE	CT	MLE	CT
θ_{27}	0.950	0.952	0.950	0.950	0.948	0.950	0.947	0.950	0.950	0.950
θ_{40}	0.948	0.952	0.948	0.949	0.945	0.948	0.944	0.948	0.952	0.948
θ_1	0.937	0.937	0.937	0.937	0.935	0.937	0.936	0.937	0.943	0.940
θ_5	0.941	0.945	0.941	0.943	0.940	0.942	0.939	0.941	0.946	0.943
θ_{14}	0.948	0.949	0.946	0.949	0.945	0.945	0.943	0.943	0.933	0.939
θ_{16}	0.953	0.952	0.953	0.952	0.952	0.951	0.950	0.951	0.934	0.948

ciascuna fonte sia disponibile una distribuzione di probabilità per il parametro di interesse e che le medie di queste distribuzioni, $\bar{\theta}_1, \ldots, \bar{\theta}_K$ vengano utilizzate come previsioni per il parametro. Sfruttando la "teoria normale degli errori" Winkler (1981) assume che gli errori di previsione $\bar{\theta}_1 - \theta, \ldots, \bar{\theta}_K - \theta$ si distribuiscano come una variabile casuale normale K-variata avente vettore delle medie $\mathbf{0}$ e generica matrice di varianze e covarianze $\boldsymbol{\Sigma}$. Infine, considerando la verosimiglianza proporzionale alla funzione di densità appena trovata, viene mostrato come trovare una distribuzione a posteriori per θ, che dipende da un'appropriata scelta di una distribuzione a priori per $(\theta, \boldsymbol{\Sigma})$.

In questo paragrafo viene descritta una procedura che si è provato a sviluppare analoga a quella appena descritta, che permette di interpretare le stime alla Wald anche da un punto di vista diverso da quello visto fino a qui. Innanzitutto, è importante evidenziare le differenze tra la seguente proposta e quella di Winkler (1981). La più evidente è la dimensione del parametro di interesse, che, come nei precedenti paragrafi, è $p > 1$. La seconda è relativa alla costruzione della verosimiglianza: si assume che in ciascun sottoinsieme sia disponibile lo stimatore $\tilde{\boldsymbol{\theta}}_k$, di massima verosimiglianza o di Firth, che è asintoticamente distribuito come una variabile casuale normale p-variata, e si assume che anche la distribuzione congiunta di $(\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_K)$ sia asintoticamente normale multivariata (pK-variata).

Siano $\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_K$, vettori $p \times 1$, gli stimatori di massima verosimiglianza o di Firth ottenuti nei K sottogruppi e sia $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_1^\top, \ldots, \tilde{\boldsymbol{\theta}}_K^\top)^\top$ il vettore $pK \times 1$ ottenuto concatenando i K vettori precedentemente definiti. Si assume che $\left((\tilde{\boldsymbol{\theta}}_1 - \boldsymbol{\theta})^\top, \ldots, (\tilde{\boldsymbol{\theta}}_K - \boldsymbol{\theta})^\top\right)^\top \dot{\sim} N_{pK}(\mathbf{0}, \boldsymbol{\Sigma})$ e vengono poste su $\boldsymbol{\theta}$ e $\boldsymbol{\Sigma}$ le seguenti priori indipendenti: una distribuzione impropria diffusa $\pi(\boldsymbol{\theta}) \propto 1$ per il vettore dei parametri e una distribuzione Wishart inversa $\boldsymbol{\Sigma} \sim W^{-1}(\nu, \boldsymbol{\Lambda}_0^{-1})$, con ν gradi di libertà e matrice di scala $\boldsymbol{\Lambda}_0^{-1}$, avente la seguente

densità

$$\pi(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(\nu+pK+1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Lambda}_0\boldsymbol{\Sigma}^{-1})\right\},$$

per la matrice di varianze e covarianze.

Sia $\boldsymbol{D}$ la matrice $p \times pK$ che si ottiene concatenando K matrici identità $\boldsymbol{I}_p$

$$\begin{aligned}\boldsymbol{D} &= \begin{pmatrix} 1 & & 1 & & \ldots\ldots & 1 & \\ & \ddots & & \ddots & \ldots\ldots & & \ddots \\ & 1 & & 1 & \ldots\ldots & & 1 \end{pmatrix} \\ &= [\boldsymbol{I}_p\, \boldsymbol{I}_p\, \ldots \boldsymbol{I}_p],\end{aligned}$$

allora si può scrivere

$$(\underbrace{\boldsymbol{\theta}^\top, \ldots, \boldsymbol{\theta}^\top}_{K \text{ volte}}) = \boldsymbol{\theta}^\top \boldsymbol{D},$$

che permette di avere la seguente forma compatta dell'assunzione di normalità

$$\tilde{\boldsymbol{\theta}} - \boldsymbol{D}^\top\boldsymbol{\theta} \dot{\sim} N_{pK}(\mathbf{0}, \boldsymbol{\Sigma}).$$

Sulla base della precedente ipotesi distributiva, si ha la seguente verosimiglianza

$$L(\boldsymbol{\theta}, \boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{D}^\top\boldsymbol{\theta}\right)^\top \boldsymbol{\Sigma}^{-1}\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{D}^\top\boldsymbol{\theta}\right)\right\},$$

che aggiorna le distribuzioni a priori scelte in precedenza, dando distribuzione a posteriori avente il seguente nucleo

$$\pi(\boldsymbol{\theta}, \boldsymbol{\Sigma} \,|\, \tilde{\boldsymbol{\theta}}) \propto |\boldsymbol{\Sigma}|^{-(\nu+pK+1)/2} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\left[\boldsymbol{\Lambda}_0 + \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{D}^\top\boldsymbol{\theta}\right)^\top\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{D}^\top\boldsymbol{\theta}\right)\right]\boldsymbol{\Sigma}^{-1}\right)\right\}.$$

Per ottenere la distribuzione a posteriori marginale per $\boldsymbol{\theta}$ è necessario integrare $\boldsymbol{\Sigma}$, che è piuttosto immediato, poiché è possibile riconoscere in $\pi(\boldsymbol{\theta}, \boldsymbol{\Sigma} \,|\, \tilde{\boldsymbol{\theta}})$ il nucleo di una distribuzione Wishart inversa in $\boldsymbol{\Sigma}$, con $(\nu + 1)$ gradi di libertà e matrice di scala $\boldsymbol{\Lambda}_*^{-1} = \left[\boldsymbol{\Lambda}_0 + \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{D}^\top\boldsymbol{\theta}\right)\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{D}^\top\boldsymbol{\theta}\right)^\top\right]^{-1}$. Perciò, completando il nucleo della Wishart inversa e tenendo solo i termini in $\boldsymbol{\theta}$, si ottiene la seguente distribuzione a posteriori

marginale per $\boldsymbol{\theta}$

$$\begin{aligned}\pi(\boldsymbol{\theta}\,|\,\tilde{\boldsymbol{\theta}}) &= \int \pi(\boldsymbol{\theta}, \boldsymbol{\Sigma}\,|\,\tilde{\boldsymbol{\theta}})\, d\boldsymbol{\Sigma} \\ &\propto \left|\boldsymbol{\Lambda}_0 + \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{D}^\top\boldsymbol{\theta}\right)\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{D}^\top\boldsymbol{\theta}\right)^\top\right|^{-(\nu+1)/2} \\ &\propto \left\{|\boldsymbol{\Lambda}_0|\left[1 + \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{D}^\top\boldsymbol{\theta}\right)^\top \boldsymbol{\Lambda}_0^{-1}\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{D}^\top\boldsymbol{\theta}\right)\right]\right\}^{-(\nu+1)/2} \\ &\propto \left[1 + \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{D}^\top\boldsymbol{\theta}\right)^\top \boldsymbol{\Lambda}_0^{-1}\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{D}^\top\boldsymbol{\theta}\right)\right]^{-(\nu+1)/2}, \end{aligned} \tag{3.13}$$

dove il penultimo passaggio è giustificato dal *Matrix determinant lemma*, che stabilisce che per una matrice invertibile $\boldsymbol{A}$ e due vettori $\boldsymbol{b}, \boldsymbol{c}$ vale

$$\left|\boldsymbol{A} + \boldsymbol{b}\boldsymbol{c}^\top\right| = |\boldsymbol{A}|\left(1 + \boldsymbol{c}^\top\boldsymbol{A}^{-1}\boldsymbol{b}\right).$$

Per riconoscere la distribuzione di $\boldsymbol{\theta}\,|\,\tilde{\boldsymbol{\theta}}$ è necessario definire le seguenti quantità

$$\boldsymbol{u} = \boldsymbol{D}\boldsymbol{\Lambda}_0^{-1}\tilde{\boldsymbol{\theta}}, \quad v = \tilde{\boldsymbol{\theta}}^\top\boldsymbol{\Lambda}_0^{-1}\tilde{\boldsymbol{\theta}}, \quad \boldsymbol{Z} = \boldsymbol{D}\boldsymbol{\Lambda}_0^{-1}\boldsymbol{D}^\top,$$

che sostituite in (3.13), permettono di riscrivere il nucleo della distribuzione a posteriori nel seguente modo

$$\begin{aligned}\pi(\boldsymbol{\theta}\,|\,\tilde{\boldsymbol{\theta}}) &\propto \left[1 + \left(\tilde{\boldsymbol{\theta}} - \boldsymbol{D}^\top\boldsymbol{\theta}\right)^\top \boldsymbol{\Lambda}_0^{-1}\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{D}^\top\boldsymbol{\theta}\right)\right]^{-(\nu+1)/2} \\ &\propto \left[1 + w + \boldsymbol{\theta}^\top\boldsymbol{Z}\boldsymbol{\theta} - 2\boldsymbol{\theta}\boldsymbol{u}\right]^{-(\nu+1)/2} \\ &\overset{\pm\boldsymbol{u}^\top\boldsymbol{Z}^{-1}\boldsymbol{u}}{\propto} \left[1 + w - \boldsymbol{u}^\top\boldsymbol{Z}^{-1}\boldsymbol{u} + \left(\boldsymbol{\theta} - \boldsymbol{Z}^{-1}\boldsymbol{u}\right)^\top \boldsymbol{Z}\left(\boldsymbol{\theta} - \boldsymbol{Z}^{-1}\boldsymbol{u}\right)\right]^{-(\nu+1)/2} \\ &\propto \left[1 + \frac{1}{\nu_*}\left(\boldsymbol{\theta} - \boldsymbol{Z}^{-1}\boldsymbol{u}\right)^\top \left(\frac{\lambda}{\nu_*}\boldsymbol{Z}^{-1}\right)^{-1}\left(\boldsymbol{\theta} - \boldsymbol{Z}^{-1}\boldsymbol{u}\right)\right]^{-(\nu_*+p)/2}, \end{aligned} \tag{3.14}$$

dove $\nu_* = \nu + 1 - p$ e $\lambda = 1 + w - \boldsymbol{u}^\top\boldsymbol{Z}^{-1}\boldsymbol{u}$. In (3.14) è possibile riconoscere il nucleo di una t di Student multivariata con ν_* gradi di liberta, parametro di posizione $\boldsymbol{Z}^{-1}\boldsymbol{u}$ e parametro di scala $\frac{\lambda}{\nu_*}\boldsymbol{Z}^{-1}$. Media, moda e mediana della distribuzione a posteriori appena trovata coincidono e perciò, qualsiasi sia la scelta, la stima puntuale a posteriori è

$$\hat{\boldsymbol{\theta}}^B = \boldsymbol{Z}^{-1}\boldsymbol{u} = \left(\boldsymbol{D}\boldsymbol{\Lambda}_0^{-1}\boldsymbol{D}^\top\right)^{-1}\boldsymbol{D}\boldsymbol{\Lambda}_0^{-1}\tilde{\boldsymbol{\theta}}.$$

Siano $\tilde{i}_1 = i_1(\tilde{\boldsymbol{\theta}}_1), \ldots, \tilde{i}_K = i_K(\tilde{\boldsymbol{\theta}}_K)$ le matrici di informazione attesa ottenute nei K gruppi e valutate nelle rispettive stime locali (di massima verosimiglianza o di Firth).

Se si considera l'iperparametro $\boldsymbol{\Lambda}_0^{-1}$ uguale alla matrice diagonale a blocchi avente sulla diagonale $\tilde{i}_1, \ldots, \tilde{i}_K$, allora

$$\begin{aligned}
\boldsymbol{D}\Lambda_0^{-1}\boldsymbol{D}^\top &= [\boldsymbol{I}_p\, \boldsymbol{I}_p\, \ldots \boldsymbol{I}_p] \begin{pmatrix} \tilde{i}_1 & & \\ & \ddots & \\ & & \tilde{i}_K \end{pmatrix} [\boldsymbol{I}_p\, \boldsymbol{I}_p\, \ldots \boldsymbol{I}_p]^\top \\
&= \left[\tilde{i}_1\, \ldots\, \tilde{i}_K\right] [\boldsymbol{I}_p\, \boldsymbol{I}_p\, \ldots \boldsymbol{I}_p]^\top \\
&= \sum_{k=1}^{K} \tilde{i}_k \boldsymbol{I}_p = \sum_{k=1}^{K} \tilde{i}_k,
\end{aligned}$$

dove $\left[\tilde{i}_1\, \ldots\, \tilde{i}_K\right]$ è la matrice $p \times pK$ ottenuta concatenando le matrici di informazione attesa, e analogamente

$$\boldsymbol{D}\boldsymbol{\Lambda}_0^{-1}\tilde{\boldsymbol{\theta}} = \sum_{k=1}^{K} \tilde{i}_k \tilde{\boldsymbol{\theta}}_k.$$

Perciò, mettendo assieme le varie componenti, si ottiene

$$\hat{\boldsymbol{\theta}}^B = \left(\sum_{k=1}^{K} \tilde{i}_k\right)^{-1} \left(\sum_{k=1}^{K} \tilde{i}_k \tilde{\boldsymbol{\theta}}_k\right),$$

che coincide con la stima combinata alla Wald utilizzando come stime locali le stime di massima verosimiglianza o quelle di Firth (la versione originale, non quella migliorata).

Questo risultato permette perciò di interpretare la stima alla Wald come moda (o media o mediana) a posteriori con una particolare scelta dell'iperparametro di scala della priori. Inoltre questa struttura bayesiana garantisce la possibilità di inserire interdipendenza tra le stime nelle diverse fonti, semplicemente scegliendo una matrice $\boldsymbol{\Lambda}_0$ non diagonale a blocchi. In particolare, la libertà di scelta degli elementi non diagonali (purché $\boldsymbol{\Lambda}_0^{-1}$ sia definita positiva) permette di rappresentare molte strutture di correlazione: se, ad esempio, i K sottoinsiemi sono ordinati temporalmente, o spazialmente, e si assume che la correlazione svanisca nel tempo, o nello spazio, allora è sufficiente specificare in $\boldsymbol{\Lambda}_0$ covarianze decrescenti all'aumentare della "distanza" tra i sottoinsiemi.

Zhou & Song (2017) propongono una procedura analoga per tenere conto della dipendenza tra le fonti, che, da un lato, ha una giustificazione puramente pratica, ma dall'altro, può essere utilizzata anche per le stime combinate alla Rao, cosa che non è possibile con quanto mostrato in questo paragrafo. Infatti le stime combinate alla Wald e alla Rao sono rispettivamente

$$\hat{\boldsymbol{\theta}}_{wcd} = \arg\min_{\theta} \left\{ \left((\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta})^\top, \ldots, (\hat{\boldsymbol{\theta}}_K - \boldsymbol{\theta})^\top \right) \hat{\mathbb{G}} \left((\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta})^\top, \ldots, (\hat{\boldsymbol{\theta}}_K - \boldsymbol{\theta})^\top \right)^\top \right\}$$

$$\hat{\boldsymbol{\theta}}_{rcd} = \arg\min_{\theta} \left\{ \psi_{conc}(\boldsymbol{W}; \boldsymbol{\theta})\, \hat{\mathbb{V}}\, \psi_{conc}(\boldsymbol{W}; \boldsymbol{\theta})^\top \right\},$$

dove $\psi_{conc}(\boldsymbol{W}; \boldsymbol{\theta}) = \left(\psi_1(\boldsymbol{W}^{(1)}; \boldsymbol{\theta})^\top, \ldots, \psi_K(\boldsymbol{W}^{(K)}; \boldsymbol{\theta})^\top \right)^\top$ è il vettore $pK \times 1$ ottenuti concatenando i K vettori di *score* e $\hat{\mathbb{G}}$ e $\hat{\mathbb{V}}$ sono matrici diagonali a blocchi contenenti rispettivamente le matrici di informazione di Godambe $\hat{\boldsymbol{G}}_k$ e le matrici di variabilità $\hat{\boldsymbol{V}}_k$. Se gli elementi non diagonali di $\hat{\mathbb{G}}$ e $\hat{\mathbb{V}}$ vengono imposti diversi da 0, analogamente a quanto fatto con $\boldsymbol{\Lambda_0}$, si può tenere conto dell'interdipendenza tra i sottoinsiemi.

Capitolo 4

Applicazione a dati Spotify

In questo capitolo viene presentata l'applicazione delle procedure definite nei capitoli precedenti a dati della piattaforma musicale Spotify, che permette di evidenziare le potenzialità di tali procedure, ma anche dare alcuni spunti per punti aperti da sviluppare in futuro.

4.1 Presentazione dei dati

I dati scelti sono relativi alle caratteristiche di canzoni presenti sulla piattaforma Spotify. L'obiettivo dell'analisi è di valutare ogni settimana quali sono gli attributi che favoriscono la presenza di canzoni pubblicate nell'ultimo mese nella playlist "Top 200 Globale", che raccoglie le canzoni più in voga del momento. In particolare è stata presa in considerazione la playlist sopracitata ogni sette giorni a partire dal giorno 01/01/2020, per un totale di 45 settimane. I dati utilizzati sono stati pubblicati sulla piattaforma Kaggle, dopo essere stati scaricati tramite l'API di Spotify (Spotify, 2021), sia quelli relativi alla playlist "Top 200 Globale" (Kaggle, 2021b) sia quelli relativi alle caratteristiche delle canzoni (Kaggle, 2021a).

La variabile risposta è una variabile casuale dicotomica che assume valore 1 se nella settimana in considerazione la canzone è nella playlist "Top 200 Globale" e 0 altrimenti. Le variabili esplicative del modello di regressione logistica riguardano caratteristiche audio della canzone, a partire dal volume della canzone in decibel fino ad arrivare ad una valutazione di quanto la canzone sia ballabile. Nella Tabella 4.1 sono elencate tutte le covariate e il loro significato; i nomi della maggior parte delle variabili sono lasciati in inglese vista la difficile traduzione in italiano.

Tabella 4.1: Variabili esplicative e loro significato.

Valence	Misura da 0 a 1 che descrive la positività comunicata dalla canzone.
Acousticness	Misura da 0 a 1 che descrive quanto la canzone sia acustica.
Danceability	Misura da 0 a 1 che descrive quanto la canzone sia adatta per ballare.
Energy	Misura da 0 a 1 di che descrive quanto la canzone sia percepita intensa.
Explicit	Definisce se la canzone è 'Parental Advisory Explicit Content' (1 : "Sì", 0 : "No").
Instrumentalness	Misura da 0 a 1 di quanto una canzone contenga contenuti vocali.
Key	Chiave stimata della canzone (0 = C, 1 = C#, ... 11 = B; 3 e 4 sono raggruppati in "Altro").
Liveness	Individua la presenza del pubblico nella traccia (assume valori da 0 a 1).
Loudness	Volume complessivo di una traccia in decibel (generalmente assume valori tra -60 e 0).
Mode Major	Definisce se il modo della canzone è maggiore (1) o minore (0).
Speechiness	Misura da 0 a 1 della presenza di parole parlate.
Tempo	Misura della velocità della canzone in BPM (Beats Per Minute).
Giorni dal rilascio	Numero di giorni dal rilascio.
Durata in sec.	Durata in secondi della traccia.

4.2 Modifiche della procedura

Tra le procedure definite nello scorso capitolo solo le stime alla Wald, sia di massima verosimiglianza che di Firth, sono applicabili senza subire modifiche. Infatti, i metodi presentati in questo lavoro sono pensati per essere applicati a sottogruppi di un insieme di dati che vengono raccolti nello stesso istante, mentre in questa analisi ciascun sottogruppo di dati è relativo ad un intervallo temporale diverso e in un dato momento sono note solo le osservazioni passate e non quelle future.

Affinché sia possibile costruire le stime alla Rao così come sono state definite in precedenza è necessario conservare i dati grezzi di tutti i sottoinsiemi passati, in modo tale da valutare le quantità necessarie nella nuova stima alla Wald aggiornata. Questa scelta però non è efficiente dal punto di vista computazionale e perciò la costruzione delle stime alla Rao è stata modificata nel seguente modo: le quantità della k-esima settimana vengono valutate nella stima alla Wald aggiornata alla k-esima settimana e cumulate a quelle delle settimane precedenti, che sono state valutate nelle corrispondenti versioni delle stime alla Wald.

Inoltre, anche la stima alla Wald migliorata di Firth richiede la conoscenza di tutti

i dati, poiché nella formula (3.11) è presente la matrice di informazione attesa $i(\boldsymbol{\theta})$. In questa applicazione il termine $\frac{1}{K}i(\boldsymbol{\theta})$ presente nella (3.11) viene calcolato con K pari al numero di settimane passate e $i(\boldsymbol{\theta})$ pari alla somma delle matrici di informazione attesa delle settimane passate.

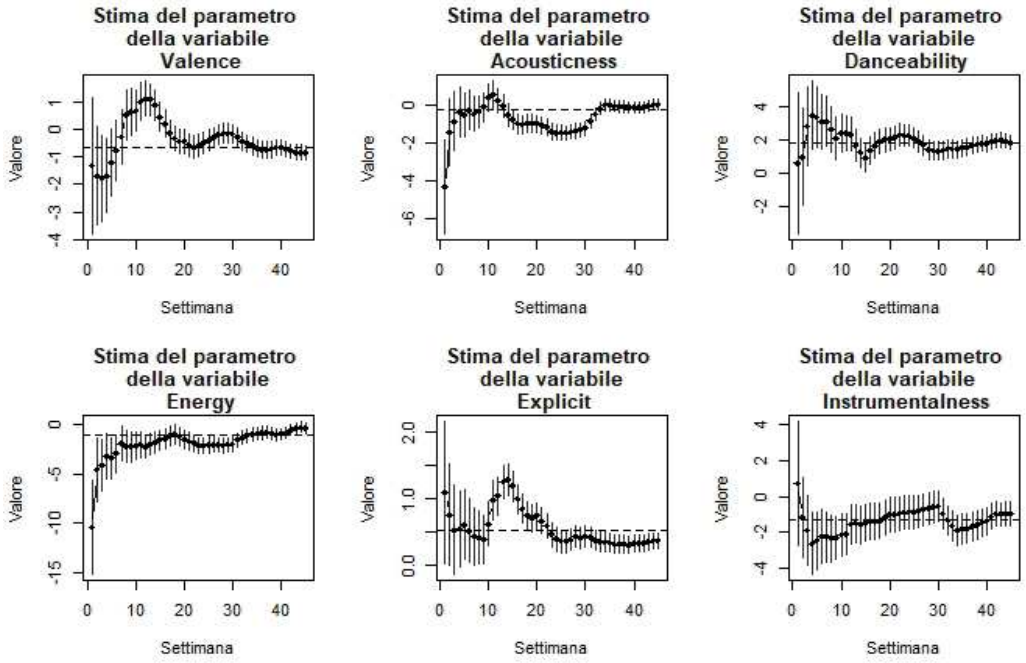

FIGURA 4.1: Evoluzione delle stime di Firth alla Rao CT dei parametri relativi alle variabili Valence, Acousticness, Danceability, Energy, Explicit e Instrumentalness.

4.3 Risultati e commenti

Per ciascuna settimana sono state calcolate le stime alla Wald e alla Rao di massima verosimiglianza e la stima alla Wald migliorata e alla Rao CT di Firth.

Per quanto riguarda le stime di massima verosimiglianza, i risultati sono instabili e incompleti, poiché il numero di settimane in cui le stime locali sono infinite è pari a 34. Ciò non solo implica che le stime alla quarantacinquesima settimana sono basate solo su 11 sottogruppi, ma anche che per le prime settimane non si ha a disposizione una stima, finché non si osservano stime locali finite e ciò avviene alla settimana numero 7.

I risultati relativi alle stime di Firth combinate alla Rao CT sono presentati nelle Figure 4.1, 4.2, 4.3, 4.4, in cui sono riportati i valori delle stime dei parametri ogni settimana, con relativo intervallo di confidenza alla Wald di livello 0.95, e il valore della stima di Firth alla Rao CT che si otterrebbe se si avessero tutti i dati contemporaneamente, rappresentato da una linea orizzontale tratteggiata. Le stime relative a gran parte dei

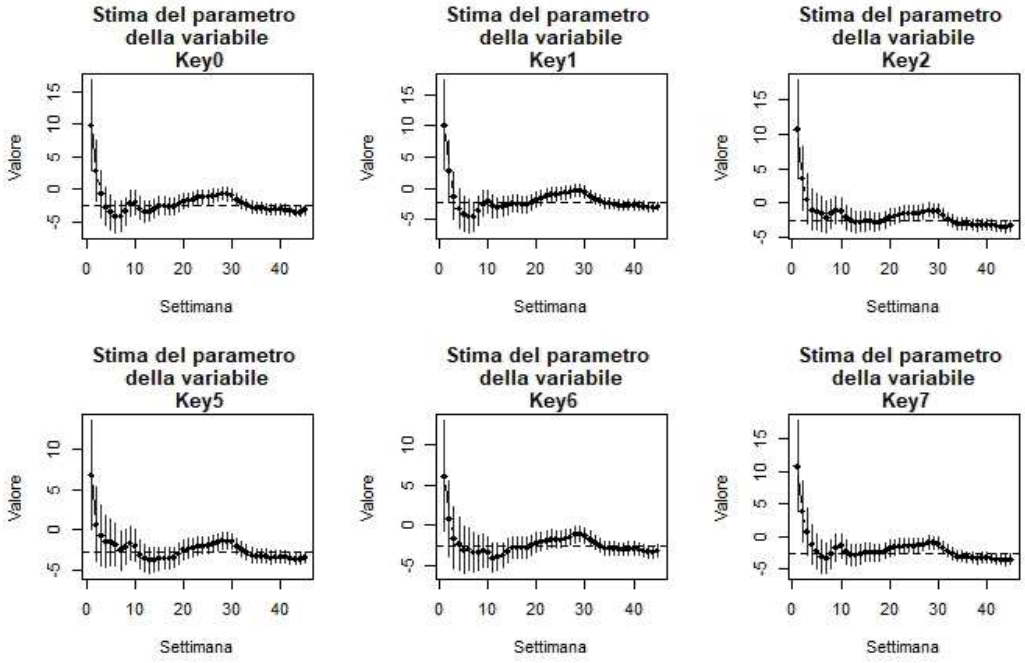

FIGURA 4.2: Evoluzione delle stime di Firth alla Rao CT dei parametri relativi ai livelli 0, 1, 2, 5, 6 e 7 della variabile Key.

parametri, dopo una fluttuazione iniziale, verso le ultime settimane si stabilizzano vicino al valore della stima alla Rao CT che si avrebbe se si osservassero tutti i dati in un solo momento.

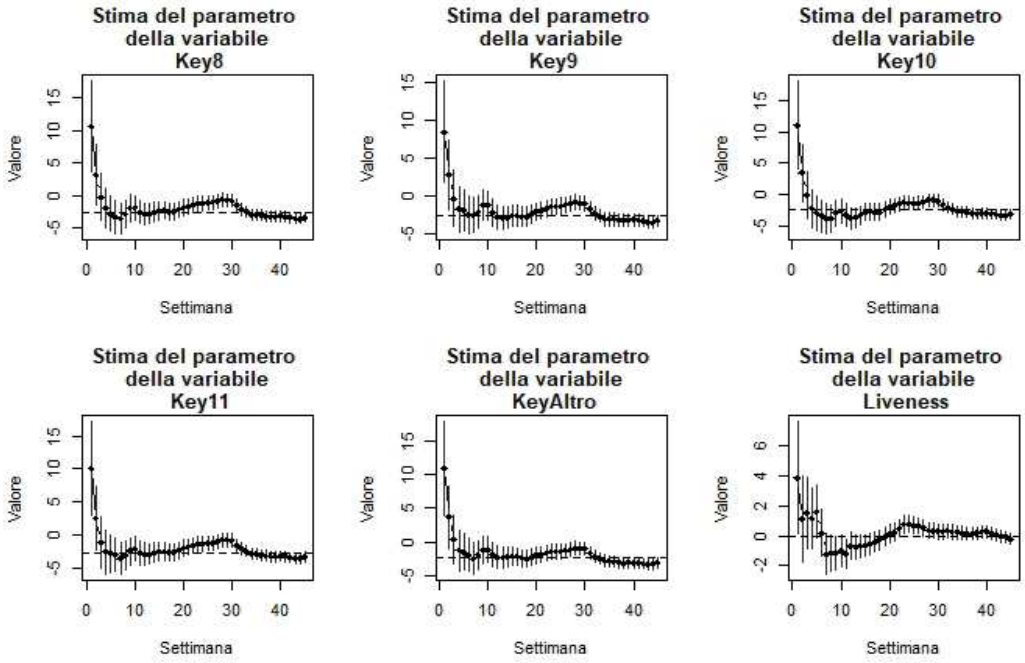

FIGURA 4.3: Evoluzione delle stime di Firth alla Rao CT dei parametri relativi ai livelli 8, 9, 10, 11 e Altro della variabile Key e alla variabile Liveness.

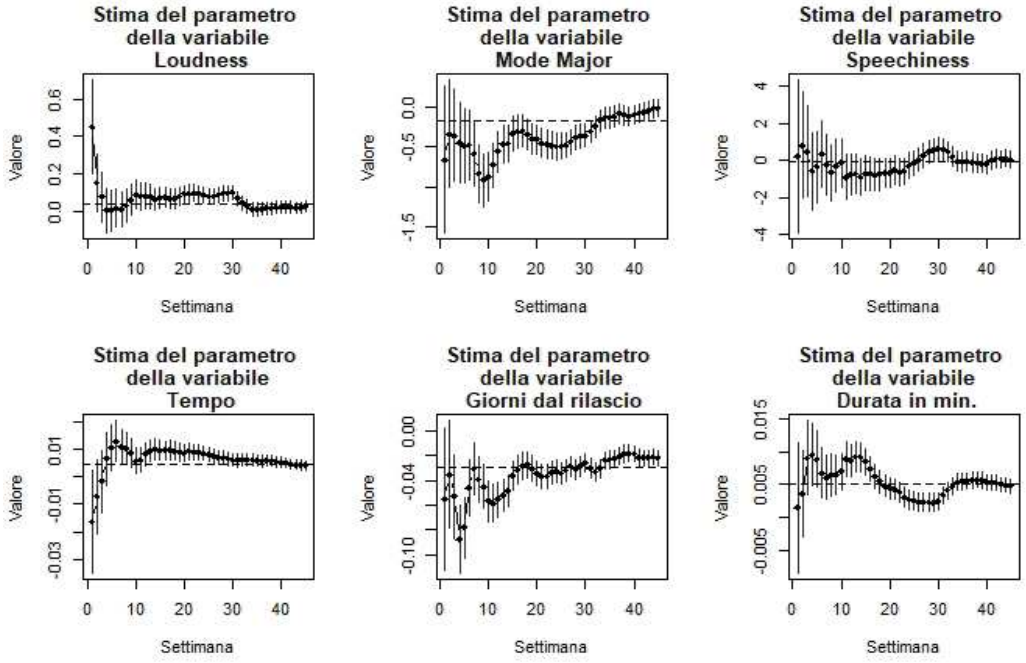

FIGURA 4.4: Evoluzione delle stime di Firth alla Rao CT dei parametri relativi alle variabili Loudness, Mode (livello Major), Speechiness, Tempo, Giorni dal rilascio e Durata in secondi.

Inoltre, da queste figure è possibile osservare alcuni cambi di tendenza che subisce l'importanza di alcune variabili. Ad esempio, la variabile *danceability* per quasi tutto l'anno è uno degli aspetti più importanti affinché una canzone finisca nella playlist "Top 200 Globale", però è possibile osservare una flessione della sua importanza tra la settimana 23 e la settimana 31, periodo che coincide con i mesi di giugno e luglio. Una possibile interpretazione che si può dare a tale fenomeno è che in un periodo, come quello estivo, in cui vengono pubblicate molte canzoni ad alta *danceability*, questo aspetto risulta meno importante, a discapito di altre caratteristiche, come la *speechiness*.

Questa applicazione mette in luce alcuni punti di forza dei metodi implementati. Innanzitutto, risulta fondamentale aver modificato la procedura per la costruzione delle stime alla Rao, ottenendo la stima alla Rao CT, in modo tale da ottenere stime con buone prestazione anche combinando stime di Firth. Infatti, utilizzando nei sottoinsiemi le stime di massima verosimiglianza non si avrebbero a disposizione valutazioni delle stime per quasi due mesi. In secondo luogo, aver definito una procedura che produce stime aventi performance paragonabili a quelle che si ottengono sull'intero insieme dei dati permette non solo di avere delle buone valutazioni all'ultima settimana, ma anche di poter osservare tali valutazioni anche in periodi intermedi e come esse evolvono.

D'altro canto è importante sottolineare l'utilizzo parzialmente improprio della combinazione di stime definita nei precedenti capitoli, poiché la combinazione delle *confidence*

density si basa sull'assunzione di indipendenza dei sottogruppi, che in questo caso è evidentemente violata. Infatti, ogni settimana vengono tenute in considerazione canzoni pubblicate nell'ultimo mese e perciò alcune canzoni sono in comune tra i dati di settimane che distano meno di quattro settimane.

Conclusioni

Nel corso di questa tesi è stato inizialmente richiamato il concetto di *confidence distribution* e di *confidence density*, così da poter mostrare il loro utilizzo per combinare stime dello stesso parametro ottenuti su insieme di dati diversi. L'idea di questo lavoro inizialmente era adattare il lavoro di Zhou & Song (2017) al modello di regressione logistica binaria. In questo contesto si corre il rischio che il numero di osservazioni disponibili per ottenere le stime in ciascun sottoinsieme sia piccolo rispetto al numero dei parametri da stimare, pur rimanendo maggiore, e ciò può comportare la possibilità che emerga il fenomeno di stime infinite nella regressione logistica. Per sfruttare il più possibile procedure di combinazione di stime su architetture distribuite conviene dividere i dati in insiemi molto piccoli su diversi nodi, così che la loro gestione non sia computazionalmente pesante. In quest'ottica, per prevenire l'eventualità che le stime in alcuni sottoinsiemi siano infinite, si è pensato di implementare le stime basate sulla verosimiglianza corretta di Firth, che garantisce stime sempre finite e in grado di rimuovere il termine di ordine $O(n^{-1})$ della distorsione delle stime di massima verosimiglianza.

È stato eseguito uno studio di simulazione per confrontare le performance delle stime ottenute combinando stime di massima verosimiglianze e stime di Firth di 5, 10 e 20 insiemi di dati in cui è stato diviso il dataset di partenza. Innanzitutto, con questa impostazione non sono mai state individuate stime infinite nei sottoinsiemi, probabilmente poiché il rapporto tra il numero di variabili p e il numero di osservazioni nei sottoinsiemi n_k è piuttosto basso, al massimo 0.2, e inoltre i dati e i parametri simulati garantiscono una varianza del predittore bassa (che è la seconda condizione individuata da Sur & Candès (2019) per l'esistenza di stime finite). I risultati in termini di distorsione sono stati piuttosto stupefacenti: non solo le stime alla Rao di Firth hanno una distorsione già evidente per $K = 5$ e che diventa sempre più grande al crescere di K, ma le stime alla Rao di massima verosimiglianza hanno distorsione pressoché nulla anche per $K = 20$ e per parametri con vero valore alto (in modulo). Tutte le stime alla Wald hanno distorsione non banale in ogni situazione e non possono essere usate come stime finali, però si sono dimostrate essere un ottimo punto di partenza per il passo Newton-Raphson che

permette di approssimare le stime alla Rao.

Perciò si è proseguito in due direzioni: modificare la procedura per le stime alla Rao di Firth in modo che avessero migliori performance e aumentare il numero di sottoinsiemi a 40, per osservare il comportamento delle stime di massima verosimiglianza in una situazione più estrema (il rapporto p/n è così pari a 0.4). In questo caso per più della metà dei campioni simulati in almeno uno dei sottoinsiemi sono state trovate stime di massima verosimiglianza infinite. Questo fenomeno e la sempre crescente distorsione delle stime locali di massima verosimiglianza ha portato ad avere per la prima volta delle stime alla Rao di verosimiglianza leggermente distorte. Per quanto riguarda le stime di Firth, è stato importante notare che la log-verosimiglianza corretta, a differenza della log-verosimiglianza, non è pari alla somma delle K log-verosimiglianze corrette locali, a causa della presenza della matrice $\boldsymbol{H}$, che dipende da tutte le osservazioni, nella correzione di Firth. Grazie a questa osservazione è anche stato possibile definire una versione migliorata delle stime alla Wald di Firth. Grazie alla struttura degli elementi diagonali di $\boldsymbol{H}$, che sono gli unici di nostro interesse, sono state definite due nuove versioni di stima alla Rao di Firth: una, $\widetilde{\text{CT}}$, molto conveniente dal punto di vista computazionale che utilizza soprattutto quantità dipendenti dalle stime locali e una, CT, con proprietà statistiche migliori, che sostituisce le stime locali con quella combinata alla Wald (richiedendo così un maggiore trasferimento di informazioni tra i nodi in cui avvengono le funzioni *Map* e quello in cui avvengono le funzioni *Reduce*). Purtroppo la prima delle due nuove stime ha distorsione che aumenta notevolmente quando K raggiunge 40, mentre la seconda ha distorsione minori di quella di verosimiglianza in quasi ogni situazione, anche se ciò comporta un leggero aumento di variabilità.

Le performance delle due stime alla Rao migliori, cioè quella di massima verosimiglienza e quella CT di Firth, sono state confrontate con le stime che si ottengono utilizzando l'intero campione da 10^4 osservazioni. Al crescere di K la distorsione delle stime combinate si distanzia un po' dalle stime ottenute sui dati interi, ma soprattutto le stime CT hanno comportamento molto simile alle stime ottenute su tutti i dati: per $K = 40$ la distorsione è maggiore di pochi decimi per parametri con vero valore nell'ordine di una o due decine, la variabilità è minore per qualsiasi valore del parametro e la probabilità di copertura empirica di intervalli di confidenza alla Wald è molto vicina a quella nominale, pari a 0.95.

In conclusione, è stata verificata la sorprendente bontà delle stime alla Rao di massima verosimiglianza anche per un numero di sottoinsiemi alto, che causa stime locali sempre più distorte e variabili, e inoltre è stata proposta una procedura per combinare

stime di Firth, che, sebbene non sia sempre migliore di quella di massima verosimiglianza, è più robusta e garantisce buone performance anche in situazioni in cui le stime di massima verosimiglianza non sono sempre finite e potrebbe essere impiegata anche in situazioni più estreme in cui il rapporto p/n è ancora maggiore (pur restando inferiore a 1).

Infine, rimangono alcuni punti che potrebbero essere approfonditi in futuro. Innanzitutto, potrebbe valere la pena verificare le performance delle stime proposte in altre situazioni: non solo aumentando il numero di sottoinsiemi, ma anche variando il numero di osservazioni e variabili oppure simulando diversamente le variabili e/o i valori dei parametri, fatto che potrebbe causare il fenomeno di stime infinite anche con un rapporto tra il numero di variabili e di osservazioni più basso di quanto accaduto in questo lavoro. In secondo luogo, nella tesi si presenta brevemente come si può modificare la procedura per potere tenere conto della dipendenza tra sottoinsiemi e sarebbe interessante verificare il funzionamento di tale modifica e/o proporre ulteriori modifiche. Nel Capitolo 4 è risultato evidente che una procedura che possa tenere conto anche di dipendenze temporali tra i sottoinsiemi, permetterebbe l'applicazione delle stime alla Rao, non solo per gestire grandi moli dati difficilmente analizzabili uniti, ma anche per ottenere stime in diversi istanti temporali che si aggiornano tenendo in considerazione le stime degli istanti precedenti, senza dover però conservare tutte le osservazioni. L'ultimo punto di futuro interesse consiste nel confronto con altri metodi di combinazione di stime, come quello proposto in Jordan et al. (2019), che potrebbero anche dare spunti per integrare e migliorare la proposta di questo lavoro di tesi.

Appendice A

Derivata seconda della correzione di Firth

Sia A_r^i la componente relativa all'r-esimo parametro della correzione per lo score di Firth per l'i-esima unita. Essa è pari a

$$A_r^i(\boldsymbol{\theta}) = h_i(\boldsymbol{\theta}) \left(\frac{1}{2} - \pi_i(\boldsymbol{\theta}) \right) x_{ir}.$$

Dovendo derivare rispetto a $\boldsymbol{\theta}$ è stata sottolineata la dipendenza di h_i e π_i dal parametro.

Allora la derivata di $A_r^i(\boldsymbol{\theta})$ rispetto a $\boldsymbol{\theta}$ è

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\theta}} A_r^i(\boldsymbol{\theta}) &= x_{ir} \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ h_i(\boldsymbol{\theta}) \left(\frac{1}{2} - \pi_i(\boldsymbol{\theta}) \right) \right\} \\ &= x_{ir} \left\{ \frac{1}{2} \frac{\partial\, h_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \pi_i(\boldsymbol{\theta}) \frac{\partial\, h_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - h_i(\boldsymbol{\theta}) \frac{\partial\, \pi_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\}. \end{aligned}$$

Poiché gli $h_i(\boldsymbol{\theta})$ sono pari a

$$h_i(\boldsymbol{\theta}) = \pi_i(\boldsymbol{\theta}) \left(1 - \pi_i(\boldsymbol{\theta})\right) \boldsymbol{x}_i^\top \left(\boldsymbol{X}^\top \boldsymbol{W}_{\boldsymbol{\theta}} \boldsymbol{X} \right)^{-1} \boldsymbol{x}_i,$$

dove il pedice di $\boldsymbol{W}_{\boldsymbol{\theta}}$ sottolinea la dipendenza di questa matrice da $\boldsymbol{\theta}$, la loro derivata è

$$\begin{aligned} \frac{\partial\, h_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = &\left[\frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \pi_i(\boldsymbol{\theta}) \left(1 - \pi_i(\boldsymbol{\theta})\right) \right\} \right] \left[\boldsymbol{x}_i^\top \left(\boldsymbol{X}^\top \boldsymbol{W}_{\boldsymbol{\theta}} \boldsymbol{X} \right)^{-1} \boldsymbol{x}_i \right] + \\ &+ \left[\frac{\partial}{\partial \boldsymbol{\theta}} \left\{ \boldsymbol{x}_i^\top \left(\boldsymbol{X}^\top \boldsymbol{W}_{\boldsymbol{\theta}} \boldsymbol{X} \right)^{-1} \boldsymbol{x}_i \right\} \right] \left\{ \pi_i(\boldsymbol{\theta}) \left(1 - \pi_i(\boldsymbol{\theta})\right) \right\}. \end{aligned} \tag{A.1}$$

Si considera ora l'r-esima componente di $\frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{x}_i^\top \left(\boldsymbol{X}^\top \boldsymbol{W}_{\boldsymbol{\theta}} \boldsymbol{X}\right)^{-1} \boldsymbol{x}_i$, che è pari a

$$\begin{aligned}
\frac{\partial}{\partial \theta_r} \boldsymbol{x}_i^\top \left(\boldsymbol{X}^\top \boldsymbol{W}_{\boldsymbol{\theta}} \boldsymbol{X}\right)^{-1} \boldsymbol{x}_i &= \boldsymbol{x}_i^\top \frac{\partial}{\partial \theta_r} \left\{\left(\boldsymbol{X}^\top \boldsymbol{W}_{\boldsymbol{\theta}} \boldsymbol{X}\right)^{-1}\right\} \boldsymbol{x}_i \\
&= -\boldsymbol{x}_i^\top \left(\boldsymbol{X}^\top \boldsymbol{W}_{\boldsymbol{\theta}} \boldsymbol{X}\right)^{-1} \left[\frac{\partial}{\partial \theta_r} \left\{\boldsymbol{X}^\top \boldsymbol{W}_{\boldsymbol{\theta}} \boldsymbol{X}\right\}\right] \left(\boldsymbol{X}^\top \boldsymbol{W}_{\boldsymbol{\theta}} \boldsymbol{X}\right)^{-1} \boldsymbol{x}_i \\
&= -\boldsymbol{x}_i^\top \left(\boldsymbol{X}^\top \boldsymbol{W}_{\boldsymbol{\theta}} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \left\{\frac{\partial}{\partial \theta_r} \boldsymbol{W}_{\boldsymbol{\theta}}\right\} \boldsymbol{X} \left(\boldsymbol{X}^\top \boldsymbol{W}_{\boldsymbol{\theta}} \boldsymbol{X}\right)^{-1} \boldsymbol{x}_i.
\end{aligned}$$

È ora necessario calcolare $\frac{\partial}{\partial \theta_r} \boldsymbol{W}_{\boldsymbol{\theta}}$, che è la matrice diagonale avente elementi non nulli pari a

$$\frac{\partial}{\partial \boldsymbol{\theta}} \left\{\pi_i(\boldsymbol{\theta}) \left(1 - \pi_i(\boldsymbol{\theta})\right)\right\},$$

che è già necessario calcolare per il primo fattore del primo addendo di (A.1).

Il gradiente di $\pi_i(\boldsymbol{\theta})$ ha r-esima componente pari a

$$\begin{aligned}
\frac{\partial\, \pi_i(\boldsymbol{\theta})}{\partial \theta_r} &= \frac{\partial}{\partial \theta_r} \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\theta}}}{1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\theta}}} \\
&= \frac{x_{ir} e^{\boldsymbol{x}_i^\top \boldsymbol{\theta}} \left(1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\theta}}\right) - x_{ir} \left(e^{\boldsymbol{x}_i^\top \boldsymbol{\theta}}\right)^2}{\left(1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\theta}}\right)^2} \\
&= x_{ir} \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\theta}}}{\left(1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\theta}}\right)^2}
\end{aligned}$$

e perciò il gradiente è

$$\frac{\partial\, \pi_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{x}_i \frac{e^{\boldsymbol{x}_i^\top \boldsymbol{\theta}}}{\left(1 + e^{\boldsymbol{x}_i^\top \boldsymbol{\theta}}\right)^2} = \boldsymbol{x}_i \pi_i(\boldsymbol{\theta}) \left(1 - \pi_i(\boldsymbol{\theta})\right).$$

Ciò implica che

$$\begin{aligned}
\frac{\partial}{\partial \theta_r} \left\{\pi_i(\boldsymbol{\theta}) \left(1 - \pi_i(\boldsymbol{\theta})\right)\right\} = \frac{\partial}{\partial \theta_r} \left\{\pi_i(\boldsymbol{\theta}) - \pi_i(\boldsymbol{\theta})^2\right\} &= \frac{\partial\, \pi_i(\boldsymbol{\theta})}{\partial \theta_r} - 2\pi_i(\boldsymbol{\theta}) \frac{\partial\, \pi_i(\boldsymbol{\theta})}{\partial \theta_r} \\
&= x_{ir} \pi_i(\boldsymbol{\theta}) \left(1 - \pi_i(\boldsymbol{\theta})\right) \left(1 - 2\pi_i(\boldsymbol{\theta})\right)
\end{aligned}$$

e quindi

$$\frac{\partial}{\partial \boldsymbol{\theta}} \left\{\pi_i(\boldsymbol{\theta}) \left(1 - \pi_i(\boldsymbol{\theta})\right)\right\} = \boldsymbol{x}_i \pi_i(\boldsymbol{\theta}) \left(1 - \pi_i(\boldsymbol{\theta})\right) \left(1 - 2\pi_i(\boldsymbol{\theta})\right).$$

Definendo $\boldsymbol{V} = \boldsymbol{X}^\top \boldsymbol{W}_{\boldsymbol{\theta}} \boldsymbol{X}$ e unendo i vari termini calcolati finora si ottiene

$$\begin{aligned}
\frac{\partial}{\partial \theta_r} \frac{A_r^i(\boldsymbol{\theta})}{x_{ir}} &= x_{ir} \left\{ \left(\frac{1}{2} - \pi_i(\boldsymbol{\theta}) \right) \frac{\partial\, h_i(\boldsymbol{\theta})}{\partial \theta_r} - h_i(\boldsymbol{\theta}) \frac{\partial\, \pi_i(\boldsymbol{\theta})}{\partial \theta_r} \right\} \\
&= \left(\frac{1}{2} - \pi_i(\boldsymbol{\theta}) \right) \Big\{ x_{ir} \pi_i(\boldsymbol{\theta}) \left(1 - \pi_i(\boldsymbol{\theta})\right) \left(1 - 2\pi_i(\boldsymbol{\theta})\right) \boldsymbol{x}_i^\top \boldsymbol{V}^{-1} \boldsymbol{x}_i - \\
&\quad - \pi_i(\boldsymbol{\theta}) \left(1 - \pi_i(\boldsymbol{\theta})\right) \boldsymbol{x}_i^\top \boldsymbol{V}^{-1} \boldsymbol{X}^\top \mathrm{diag} \left\{ x_{ir} \pi_i(\boldsymbol{\theta}) \left(1 - \pi_i(\boldsymbol{\theta})\right) \left(1 - 2\pi_i(\boldsymbol{\theta})\right) \right\} \times \\
&\quad \times \boldsymbol{X} \boldsymbol{V}^{-1} \boldsymbol{x}_i \Big\} - x_{ir} \pi_i(\boldsymbol{\theta})^2 \left(-\pi_i(\boldsymbol{\theta}) \right)^2 \boldsymbol{x}_i^\top V^{-1} \boldsymbol{x}_i.
\end{aligned}$$

Appendice B

Codice R per le simulazioni

B.1: Funzione per simulare parametri e osservazioni

```
# p è il numero di variabili, n è il numero di unità e
# R il numero di campioni della risposta simulati.
sim_data <- function(p, n, R, seed = 23)
{
  set.seed(seed)

  # Matrice del disegno
  X <- matrix(rnorm(n * p, sd = sqrt(1 / n)),
              nrow = n, ncol = p)

  # Coefficienti di regressione
  betas0 <- rep(0, p)
  betas0[1:(p/2)] <- rnorm(p/2, mean = 0, sd = sqrt(n / p))
  pi0 <- plogis( X %*% betas0 )

  # Risposta
  all_y <- list()
  for (i in 1 : R)
  {
    y <- rbinom(n, size = 1, prob = pi0)
    all_y[[i]] <- y
  }

  list(all_y = all_y, X = X, betas0 = betas0, seed = seed)
}
```

B.2: Funzione per ottenere le quantità in ogni sottoinsieme

```
# L è per "Likelihood" e "F" per "Firth".

# est_ sono i vettori con le stime, pi_ i vettori con le probabilità
```

```
# i_ sono le informazioni attese/osservate, H_ le hat-matrix e A_
# la correzione di Firth per lo score.

single_map <- function(X_k, y_k)
{
  count_inf_est <- 0
  colnames(X_k) <- paste("X", 1:(ncol(X_k)), sep = "")

  # Verifica presenza di perfetta separazione (che implica
  # stime infinite).
  test <- detect_separation(X_k, y_k, family = binomial("logit"),
                            intercept = FALSE)$separation
  if (test)
  {
    count_inf_est <- count_inf_est + 1

    # Pongo pari a 0 sia le stime che la matrice di varianze e
    # covarianze così che non contribuiscano nelle stime combinate
    # dei parametri e delle relative matrice di varianze e covarianze.
    est_L <- rep(0, ncol(X_k))
    i_L <- matrix(0, nrow = ncol(X_k), ncol = ncol(X_k))
    fit_F <- glm(y_k ~ - 1 + X_k, family = binomial("logit"),
                 method = "brglmFit")
  } else
  {
    fit_L <- glm(y_k ~ -1 + X_k, family = binomial(link = "logit"))
    est_L <- unname(coef(fit_L))
    i_L <- t(X_k) %*%
      diag( predict(fit_L, type = "response") *
              (1 - predict(fit_L, type = "response")) ) %*% X_k
    fit_F <- update(fit_L, method = "brglmFit", type="AS_mean")
  }

  est_F <- unname(coef(fit_F))
  pi_F <- predict(fit_F, type = "response")

  W12 <- diag(pi_F * (1 - pi_F), nrow = nrow(X_k))^(1/2)
  W12_X <- W12 %*% X_k
  i_F <- t(W12_X) %*% W12_X
  H_F <- W12_X %*% solve(i_F) %*% t(W12_X)
  A_k <- 0.5 * colSums((diag(H_F) / 2 -  diag(H_F) * pi_F) * X_k)

  list(est_L = est_L, est_F = est_F, i_L = i_L, i_F = i_F,
       W12_X = W12_X, H_F = H_F, A_k = A_k, pi_F = pi_F,
```

```
      count_inf_est = count_inf_est)
}
```

B.3: Funzione per ottenere le quantità valutate nelle stime alla Wald

```
# wcd_L è la stima alla Wald di massima verosimiglianza, wcd_F è
# la stima alla Wald di Firth e wcd_F_imp è la stima alla Wald
# di Firth Migliorata.
wcd_operations <- function(X_k, y_k, wcd_L, wcd_F, wcd_F_imp)
{
  # Vettori con le probabilità pigreco
  pi_L <- as.vector( exp(X_k %*% wcd_L) / (1 + exp(X_k %*% wcd_L)) )
  pi_F <- as.vector( exp(X_k %*% wcd_F) / (1 + exp(X_k %*% wcd_F)) )
  pi_F_imp <- as.vector( exp(X_k %*% wcd_F_imp) /
                           (1 + exp(X_k %*% wcd_F_imp)) )

  # Informazioni attese (e matrici necessarie per calcolarle)
  W_F <- diag(pi_F * (1 - pi_F))
  W12_X <- W_F^(1/2) %*% X_k
  i_L <- t(X_k) %*% diag(pi_L * (1 - pi_L)) %*% X_k
  i_F <- t(W12_X) %*% W12_X
  W_F_imp <- diag(pi_F_imp * (1 - pi_F_imp))
  W12_X_imp <- W_F_imp^(1/2) %*% X_k
  i_F_imp <- t(W12_X_imp) %*% W12_X_imp

  ## Score
  score_L <- colSums((y_k - pi_L) * X_k)
  H_F <- W12_X %*% solve(i_F) %*% t(W12_X)
  score_F <- colSums( ((y_k + diag(H_F) / 2) -
                         (1 + diag(H_F)) * pi_F) * X_k)
  H_F_imp <- W12_X_imp %*% solve(i_F_imp) %*% t(W12_X_imp)
  score_F_imp <- colSums( ((y_k + diag(H_F_imp) / 2) -
                             (1 + diag(H_F_imp)) * pi_F_imp) * X_k)

  list(score_L = score_L, score_F = score_F,
       score_F_imp = score_F_imp,
       W12_X = W12_X, W12_X_imp = W12_X_imp,
       i_L = i_L, i_F = i_F, i_F_imp = i_F_imp,
       pi_F = pi_F, pi_F_imp = pi_F_imp)
}
```

B.4: Funzione per ottenere le stime combinate per un solo valore di K

```
get_est <- function(X, y, k)
{
```

```
library(brglm2); library(detectseparation)
n <- nrow(X); p <- ncol(X)

list_y <- list()
list_X <- list()

# Divisione delle X e delle y in sottoinsiemi.
# Siccome le osservazioni sono iid, si possono prendere le prime
# N, poi le seconde N e così via.
for (j in 1 : k)
{
  list_X[[j]] <- X[((j - 1) * (n / k) + 1) : (j * (n / k)), ]
  list_y[[j]] <- y[((j - 1) * (n / k) + 1) : (j * (n / k))]
}


# Stime locali
res_k <- lapply(1 : k,
                function(j) single_map(list_X[[j]], list_y[[j]]))


# Conteggio del numero di dataset in cui si sono trovate
# stime infinite.
count_inf <- sum(sapply(res_k, function(j) j$count_inf_est))


# Matrici di informazione ottentute come somma delle
# informazioni locali.
sum_i_L <- Reduce("+", lapply(res_k, function(x) x$i_L))
sum_i_F <- Reduce("+", lapply(res_k, function(x) x$i_F))


# Matrici di varianze e covarianze globali ottenute come inverse
# della somma delle matrici di informazione attesa locali. Per la
# massima verosimiglianza uso tryCatch per gestire le situazioni
# in cui tutti i sottoinsiemi restituiscono stime infinite.
vcov_L <- tryCatch(
  expr = {
    solve(sum_i_L)
  },
  error = function(e){
    matrix(NA, nrow = ncol(X), ncol = ncol(X))
  }
)
```

```
vcov_F <- solve(sum_i_F)


### Stime alla Wald
# Likelihood
wcd_L <- vcov_L %*%
  Reduce("+", lapply(res_k, function(x) x$i_L %*% x$est_L))

# Firth
wcd_F <- vcov_F %*%
  Reduce("+", lapply(res_k, function(x) x$i_F %*% x$est_F))

# Firth migliorato
all_pi_k <- do.call("c", lapply(res_k, function(x) x$pi_F))
rbind_W12_X_k <- do.call("rbind", lapply(res_k, function(x) x$W12_X))
dH_F_k <- diag(rbind_W12_X_k %*% vcov_F %*% t(rbind_W12_X_k))
A <- 0.5 * colSums( (dH_F_k / 2 - dH_F_k * all_pi_k) * X )
B <- A - Reduce("+", lapply(res_k, function(x) x$A_k))
wcd_F_imp <- solve(Reduce("+", lapply(res_k,
                                      function(x) x$i_F))) %*%
  (B + Reduce("+", lapply(res_k, function(x) x$i_F %*% x$est_F)))

for (j in 1:5) {gc()}


### Stime alla Rao
# Lista con gli oggetti valutati nelle stime alla Wald.
wcd_obj <- lapply(1 : k,
                  function(j) wcd_operations(list_X[[j]],
                                             list_y[[j]], wcd_L,
                                             wcd_F, wcd_F_imp))

# Somme delle informazioni attese locali valutate nelle stime
# alla Wald e correspondenti stime delle matrice di
# varianze e covarianze.
sum_i_L_wcd <- Reduce("+", lapply(wcd_obj, function(x) x$i_L))
sum_i_F_wcd <- Reduce("+", lapply(wcd_obj, function(x) x$i_F))
sum_i_F_imp_wcd <- Reduce("+", lapply(wcd_obj,
                                      function(x) x$i_F_imp))
vcov_L_wcd <- solve( sum_i_L_wcd )
vcov_F_wcd <- solve( sum_i_F_wcd )
vcov_F_imp_wcd <- solve( sum_i_F_imp_wcd )

# Likelihood
```

```
rcd_L <- wcd_L + vcov_L_wcd %*%
  Reduce("+", lapply(wcd_obj, function(x) x$score_L))


# Firth prima versione
rcd_F <- wcd_F + vcov_F_wcd %*%
  Reduce("+", lapply(wcd_obj, function(x) x$score_F ))

# Firth con Correzione Totale
all_pi_F <- do.call("c", lapply(wcd_obj, function(x) x$pi_F))
rbind_W12_X <- do.call("rbind",
                       lapply(wcd_obj, function(x) x$W12_X))
dH_F <- diag(rbind_W12_X %*% vcov_F_wcd %*% t(rbind_W12_X))
score_F <- colSums(((y + dH_F / 2) - (1 + dH_F) * all_pi_F) * X)
rcd_F1 <- wcd_F + vcov_F_wcd %*% score_F

# Approssimazione di Firth con la Correzione Totale (nella
# costruzione della matrice H e delle informazioni attese si
# utilizzano solo le stime locali e non quella alla Wald).
all_pi_F2 <- do.call("c", lapply(wcd_obj, function(x) x$pi_F))
rbind_W12_X2 <- rbind_W12_X_k
dH_F2 <- dH_F_k
score_F2 <- colSums(((y + dH_F2 / 2) - (1 + dH_F2) * all_pi_F2) * X)
rcd_F2 <- wcd_F + vcov_F %*% score_F2

# Firth prima versione partendo da Wald Migliorata
rcd_F_imp <- wcd_F_imp + vcov_F_imp_wcd %*%
  Reduce("+", lapply(wcd_obj, function(x) x$score_F_imp ))

# Firth con Correzione Totale partendo da Wald Migliorata
all_pi_F_imp <- do.call("c",
                        lapply(wcd_obj, function(x) x$pi_F_imp))
rbind_W12_X_imp <- do.call("rbind",
                           lapply(wcd_obj, function(x) x$W12_X_imp))
dH_F_imp <- diag(rbind_W12_X_imp %*%
                   vcov_F_imp_wcd %*% t(rbind_W12_X_imp))
score_F_imp <- colSums(((y + dH_F_imp / 2) -
                          (1 + dH_F_imp) * all_pi_F_imp) * X)
rcd_F1_imp <- wcd_F_imp + vcov_F_imp_wcd %*% score_F_imp

# Approssimazione di Firth con la Correzione Totale (nella
# costruzione della matrice H e delle informazioni attese si
# utilizzano solo le stime locali e non quella alla Wald) ma
# partendo da Wald Migliorata.
```

```
  score_F2_imp <- colSums( ((y + dH_F2 / 2) -
                              (1 + dH_F2) * all_pi_F_imp) * X)
  rcd_F2_imp <- wcd_F_imp + vcov_F %*% score_F2_imp


  for (el in ls())
  {
    if (!(el %in% c("wcd_L", "wcd_F", "wcd_F_imp",
                    "rcd_L", "rcd_F", "rcd_F1", "rcd_F2",
                    "rcd_F_imp", "rcd_F1_imp", "rcd_F2_imp",
                    "vcov_L", "vcov_F",
                    "vcov_L_wcd", "vcov_F_wcd", "vcov_F_imp_wcd",
                    "count_inf")))
    {
      rm(list = el)
    }
  }

  for (j in 1:5) {gc()}

  # Output
  list(wcd_L = drop(wcd_L), wcd_F = drop(wcd_F),
       wcd_F_imp = drop(wcd_F_imp),
       rcd_L = drop(rcd_L), rcd_F_vecchio = drop(rcd_F),
       rcd_F_unita = drop(rcd_F1), rcd_F_note = drop(rcd_F2),
       rcd_F_imp_vecchio = drop(rcd_F_imp),
       rcd_F_imp_unita = drop(rcd_F1_imp),
       rcd_F_imp_note = drop(rcd_F2_imp),
       vcov_L = vcov_L, vcov_F = vcov_F,
       vcov_L_wcd = vcov_L_wcd, vcov_F_wcd = vcov_F_wcd,
       vcov_F_imp_wcd = vcov_F_imp_wcd,
       count_inf = count_inf)
}
```

B.5: Funzione per ottenere le stime combinate per più valori di K

```
get_est_multi <- function(X, y, list_k, trace = FALSE)
{
  lapply(list_k, function(k) get_est(X, y, k, trace))
}
```

B.6: Esecuzione in parallelo sui 10^3 campioni

```
R <- 10^3 # Numero di campioni
p <- 100 # Numero di variabili
```

```
n <- 10^4 # Numero di osservazioni
data_R <- sim_data(p = p, n = n, R = R, seed = 23)

library(parallel)
library(doSNOW)
library(tcltk)
ncores <- detectCores()
cl <- makeSOCKcluster(ncores - 1)
registerDoSNOW(cl)
progress <- function(r) cat(sprintf("task %d is complete\n", r))
opts <- list(progress = progress)

out <- foreach(y_input = data_R$all_y, .options.snow=opts) %dopar%
  {
    get_est_multi(data_R$X, y = y_input, list_k = c(5, 10, 20, 40))
  }
```

Bibliografia

Candès, E. J. & Sur, P. (2020). The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *The Annals of Statistics* **48**, 27–42.

Cordeiro, G. M. & McCullagh, P. (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society. Series B (Methodological)* **53**, 629–643.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.

Gourieroux, C. & Monfort, A. (1981). Asymptotic properties of the maximum likelihood estimator in dichotomous logit models. *Journal of Econometrics* **17**, 83–97.

Jordan, M. I., Lee, J. D. & Yang, Y. (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association* **114**, 668–681.

Kaggle (2021a). Dataset caratteristiche canzoni. `https://www.kaggle.com/eduhdm/spotify-songs-dataset/metadata`.

Kaggle (2021b). Dataset classifiche giornaliere. `https://www.kaggle.com/pepepython/spotify-huge-database-daily-charts-over-3-years/metadata`.

Kosmidis, I. (2007). *Bias Reduction in Exponential Family Nonlinear Models*. Tesi di Dottorato, Department of Statistics, University of Warwick.

Kosmidis, I. (2020). *brglm2: Bias Reduction in Generalized Linear Models*. R package version 0.8.

Kosmidis, I. (2021). *detectseparation: Detect and Check for Separation and Infinite Maximum Likelihood Estimates*. R package version 0.2.

Kosmidis, I. & Firth, D. (2021). Jeffreys-prior penalty, finiteness and shrinkage in binomial-response generalized linear models. *Biometrika* **108**, 71–82.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman & Hall/CRC, 2nd ed.

Nordberg, L. (1980). Asymptotic normality of maximum likelihood estimators based on independent, unequally distributed observations in exponential family models. *Scandinavian journal of statistics* **7**, 27–32.

Pace, L. & Salvan, A. (1997). *Principles of Statistical Inference from a Neo-fisherian Perspective*. Singapore: World Scientific Publishing Company.

Singh, K., Xie, M. & Strawderman, W. E. (2007). Confidence distribution (cd): Distribution estimator of a parameter. *Lecture notes-monograph series* **54**, 132–150.

Song, P. X.-K. (2007). *Correlated Data Analysis: Modeling, Analytics, and Applications*. New York: Springer-Verlag, 1st ed.

Spotify (2021). API spotify. `https://developer.spotify.com/documentation/web-api/`.

Sur, P. & Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences - PNAS* **116**, 14516–14525.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge: Cambridge University Press.

Winkler, R. L. (1981). Combining probability distributions from dependent information sources. *Management Science* **27**, 479–488.

Zhou, L. & Song, P. X. K. (2017). Scalable and efficient statistical inference with estimating functions in the mapreduce paradigm for big data. `https://arxiv.org/abs/1709.04389`.