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

Relazione Finale

Criteri di selezione della matrice di correlazione in modelli basati su equazioni di stima generalizzate

Relatore Prof. Alessandra Salvan
Dipartimento di Scienze Statistiche

Laureando Bryan Patarini
Matricola 2051895

Anno Accademico 2023/2024

Indice

Introduzione

In diversi ambiti delle scienze applicate e sociali lo studio di dati longitudinali è sempre più frequente. In tale contesto la variabile risposta della singola unità statistica è multivariata, con osservazioni tra di loro correlate. Per variabili di questa natura si fa spesso ricorso a modelli marginali. Liang & Zeger (1986), propongono la metodologia delle equazioni di stima generalizzate come strumento per la specificazione e l'analisi di questi modelli. Lo stimatore per i parametri β che ne deriva rimane consistente anche se la matrice di correlazione non è correttamente specificata, ma viene formulata solamente come ipotesi di lavoro. La corretta specificazione impatta tuttavia sull'efficienza dello stimatore. Selezionare dunque un'appropriata struttura di correlazione, che rispecchi quella vera e ignota tra misure ripetute, porta ad un miglioramento dell'efficienza della stima dei parametri e dunque ad una inferenza più affidabile. Questa relazione si propone di rispondere a questa esigenza.

Nello specifico, l'obiettivo di questo elaborato è fornire una panoramica esaustiva sui criteri di selezione della struttura di correlazione presenti in letteratura, proponendo un confronto e valutandone l'efficacia via simulazione. Si illustrano in particolare i contenuti di cui agli articoli Pardo & Alonso (2019) e Carey & Wang (2011), integrandoli con criteri basati sulla verosimiglianza empirica (Chen & Lazar (2012)).

Nel primo capitolo si presenta una breve descrizione dei modelli lineari generalizzati, con particolare riferimento ai modelli per risposte correlate, e alla quasi-verosimiglianza. Nel Capitolo 2 vengono illustrati i criteri per la selezione della struttura di correlazione. Nel Capitolo 3 viene valutato il comportamento dei criteri di selezione attraverso un ampio studio di simulazione e l'applicazione a un caso reale. Infine, lo studio si conclude con un confronto dei vari criteri e la presentazione dei risultati ottenuti.

Capitolo 1

Modelli lineari generalizzati

1.1 Introduzione ai GLM

I modelli lineari generalizzati (*Generalized Linear Models*, in breve GLM) rappresentano un'estensione del modello lineare normale che permette di modellare variabili risposta con distribuzione non necessariamente normale e la cui media dipende dal predittore lineare anche attraverso funzioni diverse dalla funzione identità. Quanto esposto nel presente capitolo ricalca Salvan et al. (2020), mentre per una panoramica completa sull'argomento si veda Agresti (2015).

1.1.1 Famiglie di dispersione esponenziale

Siano $y_1, \ldots, y_n$, $i = 1, \ldots, n$ realizzazione delle variabili aleatorie $Y_1, \ldots, Y_n$ aventi distribuzione in una famiglia di distribuzione esponenziale univariata con funzione di densità

$$p(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}, \tag{1.1}$$

con $y_i \in S \subseteq \mathbb{R}$, $\theta_i \in \Theta \subseteq \mathbb{R}$, $a_i(\phi) > 0$. Il parametro θ_i viene definito parametro naturale, mentre ϕ è detto parametro di dispersione. Per risposte Y_i con densità descritta dalla (1.1), la funzione $b(\theta_i)$ determina tutti i momenti ed è definita generatore dei cumulanti. Il momento primo risulta allora

$$\mathbb{E}(Y_i) = \mu = b'(\theta_i) = \mu(\theta_i), \tag{1.2}$$

indipendente da ϕ. In aggiunta, si ottiene che

$$\mathbb{V}ar(Y_i) = a_i(\phi)b''(\theta_i)|_{\theta_i=\theta(\mu_i)} = a_i(\phi)v(\mu_i), \tag{1.3}$$

dove con b' e b'' si indicano le prime due derivate di $b(\theta_i)$, $\theta(\mu_i)$ è l'inversa della funzione $\mu(\theta_i)$ e $v(\mu_i) = b''(\theta_i)|_{\theta_i=\theta(\mu_i)}$ è detta funzione di varianza. Definendo lo spazio delle medie $\mathbb{M} = \mu(\text{int}(\Theta))$, dove $\text{int}(\Theta)$ è l'insieme dei punti interni di Θ, si introduce allora la notazione sintetica

$$Y_i \sim DE_1(\mu_i, a_i(\phi)v(\mu_i)), \quad \mu_i \in \mathbb{M} \tag{1.4}$$

per indicare la distribuzione di Y_i. Ferma restando l'ipotesi di indipendenza delle variabili casuali Y_i, $i = 1, \dots, n$, le componenti che caratterizzano un modello lineare generalizzato sono quelle di seguito elencate.

- Distribuzione della risposta della forma descritta nella (1.4).
- Predittore lineare: $\eta = X\beta$, con componenti $\eta_i = \mathbf{x}_i\beta$, dove X è una matrice $n \times p$ a rango pieno di costanti note con i-esima riga $\mathbf{x}_i$, $n < p$ e $\beta = [\beta_1, \dots, \beta_p]^\top$, vettore di coefficienti di regressione.
- Funzione di legame: detta *link function*, è la funzione $g(\cdot)$ che mette in relazione μ_i e η_i, ovvero tale che $g(\mu_i) = \eta_i = \mathbf{x}_i\beta = \sum_{i=1}^{n} x_{ir}\beta_r$.

La funzione di legame permette di modellare il legame tra covariate e variabile risposta in termini lineari, lasciando variare liberamente β in $\mathbb{R}^p$, e imponendo solamente un legame tra spazio delle medie $\mathbb{M}$ e lo spazio $\mathbb{R}$ del predittore lineare η_i. In generale si predilige la funzione $g(\cdot)$ tale che $g(\mu_i) = \theta(\mu_i)$. Quest'ultima è detta funzione di legame canonica. Viene infine indicata con $f(\cdot) = g^{-1}(\cdot)$ l'inversa della funzione di legame.

1.1.2 Verosimiglianza e inferenza

Nei modelli lineari generalizzati, al fine di ottenere stime puntuali e intervallari dei parametri di regressione e testarne la significatività statistica si ricorre all'inferenza di verosimiglianza. Per una trattazione dettagliata si veda Salvan et al. (2020), paragrafo 2.3. Una volta specificata una funzione di legame $g(\cdot)$, essendo le Y_i tra loro indipendenti e distribuite sotto l'ipotesi (1.4), risulta che la densità congiunta delle Y_i è data dal prodotto delle densità marginali (1.1) e la funzione di log-verosimiglianza per (β, ϕ) è

$$l(\beta, \phi) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + \sum_{i=1}^{n} c(y_i, \phi), \tag{1.5}$$

dove $\theta_i = \theta(\mu_i) = \theta(f(\mathbf{x}_i\beta))$. In generale, non esiste una statistica sufficiente con dimensione inferiore a n, anche supponendo ϕ noto. Ciononostante, se $g(\mu_i) = \theta(\mu_i)$, ovvero $g(\cdot)$ è la funzione di legame canonica, tale che $\theta_i = \mathbf{x}_i\beta$, per qualsiasi valore fissato di ϕ, esiste una statistica sufficiente p-dimensionale per l'inferenza su β. Nello specifico, si ha che la (1.5) si riduce alla forma

$$l(\beta, \phi) = \sum_{i=1}^{n} \frac{y_i \mathbf{x}_i\beta - b(\mathbf{x}_i\beta)}{a_i(\phi)} + \sum_{i=1}^{n} c(y_i, \phi). \tag{1.6}$$

Assumendo ϕ noto, la condizione del primo ordine per la ricerca di massimi impone l'uguaglianza del gradiente della (1.5), detta funzione *score*, a 0. Si ottengono dunque le equazioni di verosimiglianza per β

$$l_r = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mathbb{V}ar(Y_i)} \frac{\partial \mu_i}{\partial \beta_r} = 0, \quad r = 1, \ldots, p. \tag{1.7}$$

Se $g(\cdot)$ è il link canonico, le equazioni (1.7) si semplificano in

$$\sum_{i=1}^{n} \frac{1}{a_i(\phi)} y_i x_{ir} = \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \mu_i x_{ir}, \qquad r = 1, \ldots, p. \tag{1.8}$$

Le equazioni (1.7) possono essere espresse in forma matriciale, risultando

$$D^T V^{-1} (y - \mu) = 0 \tag{1.9}$$

dove $y - \mu = [y_1 - \mu_1, \ldots, y_n - \mu_n]^\top$, $V = \text{diag}[\mathbb{V}ar(Y_i)]$, $i = 1, \ldots, n$ e D è una matrice $n \times p$ in cui il generico elemento è

$$d_{ir} = \frac{\partial \mu_i}{\partial \beta_r} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_r} = \frac{1}{g\prime(\mu_i)} x_{ir}, \quad i = 1, \ldots, n \quad r = 1, \ldots, p. \tag{1.10}$$

Si dimostra che i parametri β e ϕ sono parametri ortogonali (si veda Salvan et al. (2020), paragrafo 2.3.3). Si ha inoltre che, qualora si utilizzi il legame canonico, la matrice di informazione osservata $j_{\beta\beta}$ coincide con il proprio valore atteso $i_{\beta\beta}$, che in forma matriciale risulta

$$i_{\beta\beta} = X^\top W X, \tag{1.11}$$

dove $W = \text{diag}(w_i)$, con $w_i = \frac{1}{(g'(\mu_i))^2 \mathbb{V}ar(Y_i)}$, $i = 1, \ldots, , n$. Se $g(\cdot)$ è la funzione di legame canonica allora i pesi w_i si riducono della forma

$$w_i = \frac{1}{(1/v(\mu_i))^2 a_i(\phi) v(\mu_i)} = \frac{v(\mu_i)}{a_i(\phi)}.$$

Infine, sfruttando la proprietà di normalità asintotica dello stimatore di massima verosimiglianza si ottiene l'approssimazione

$$\hat{\beta} \dot{\sim} N_p(\beta, (X^\top W X)^{-1}), \tag{1.12}$$

per n grande. Dalla (1.12) si può derivare una stima consistente della matrice di covarianza di β data da $(X^T \hat{W} X)^{-1}$, dove con $\hat{W}$ si indica la matrice W calcolata ponendo $\beta = \hat{\beta}$ e a ϕ, qualora fosse ignoto, una sua stima consistente. In particolare, per la stima di ϕ si fa ricorso a stimatori basati sul metodo dei momenti, quale lo stimatore con correzione

$$\tilde{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)}, \tag{1.13}$$

dove i valori attesi μ_i sono sostituiti dalle relative stime basate su $\hat{\beta}$.

1.2 Quasi-verosimiglianza

Si osservi per un momento la formula (1.7). Appare evidente come le equazioni di verosimiglianza per β dipendano dalla distribuzione della risposta unicamente attraverso $\mathbb{E}(Y_i) = \mu_i$ e da $\mathbb{V}ar(Y_i) = \phi v(\mu_i)$, una volta specificata la funzione di legame $g(\cdot)$. Risulta dunque interessante studiare le caratteristiche dello stimatore $\hat{\beta}$, soluzione della (1.7), sotto le più deboli ipotesi del secondo ordine di un GLM

$$\mathbb{E}(Y_i) = \mu(\mathbf{x}_i \beta) = f(\mathbf{x}_i \beta), \tag{1.14}$$

$$\mathbb{V}ar(Y_i) = \phi v(\mu_i), \tag{1.15}$$

$$Y_i \text{ e } Y_j \text{ sono indipendenti se } i \neq j, \tag{1.16}$$

dove ϕ è il parametro di dispersione, ignoto. Il modello statistico semiparametrico specificato dalle ipotesi (1.14)–(1.16) viene detto modello di quasi-verosimiglianza. Tale modello, formulabile sia per dati continui che per dati discreti, fornisce un aumento di flessibilità circa la modellazione della varianza della risposta rispetto alle più rigide specificazioni dei modelli lineari generalizzati. In diverse applicazioni può risultare più

opportuno assumere che la varianza sia, ad esempio, maggiore di quella prevista dal corrispondente GLM, ovvero che i dati presentino sovradispersione rispetto al modello parametrico ipotizzato, per il quale la varianza delle osservazioni è interamente determinata dal valore atteso. L'incremento per la varianza della risposta previsto rispetto a quella relativa al corrispettivo GLM è descritto dal parametro ϕ.

1.2.1 Definizione

Sotto le ipotesi (1.14)–(1.16) per la singola osservazione y_i è possibile formulare una funzione di quasi log-verosimiglianza, espressa dalla relazione

$$\frac{\partial Q(y_i, \mu_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{\phi v(\mu_i)}, \tag{1.17}$$

o equivalentemente

$$Q(y_i, \mu_i) = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi v(\mu_i)} dt + h(y_i), \tag{1.18}$$

dove $h(\cdot)$ è una funzione non meglio specificata dipendente solo dai dati. Per una trattazione approfondita si rimanda a Wedderburn (1974).

1.2.2 Equazioni di stima non distorte

La non distorsione delle equazioni di verosimiglianza (1.7) rappresenta l'assunto principale per dimostrare la consistenza dello stimatore di massima verosimiglianza. Tale ipotesi è dunque cruciale anche per modelli basati sulla quasi-verosimiglianza. Definendo la funzione $q(y; \beta)$

$$q(y; \beta) = [q(y_1; \beta), \ldots, q(y_n; \beta)]^\top = D^\top V^{-1}[y - \mu], \tag{1.19}$$

con $[y - \mu]^\top = [y_1 - \mu_1, \ldots, y_n - \mu_n]$, $V = \phi \, \mathrm{diag}[v(\mu_i)]$ e D matrice $n \times p$ con generico elemento espresso dalla (1.10), si dice che la (1.19) fornisce un'equazione di stima non distorta per β se

$$\mathbb{E}_\beta[q(Y; \beta)] = 0, \qquad \forall \, \beta \in \mathbb{R}^p. \tag{1.20}$$

Con riferimento alla (1.19), si noti che la funzione $q(y; \beta)$ rappresenta il gradiente della log-verosimiglianza di un GLM classico, e dunque

$$q_r(y; \beta) = l_r \ \text{ e } \ q_s(y; \beta) = l_s \qquad \text{con } r, s = 1, \ldots, p. \tag{1.21}$$

Si definiscono inoltre

$$J(\beta) = \mathbb{E}_\beta \left[q(Y;\beta) q(Y;\beta)^\top \right] \tag{1.22}$$

e

$$H(\beta) = -\mathbb{E}_\beta \left[\frac{\partial q(Y;\beta)}{\partial \beta^\top} \right] = X^\top W X, \tag{1.23}$$

assumendo $H(\beta)$ simmetrica. La (1.22) è esprimile nella forma

$$J(\beta) = \mathbb{V}ar(q(Y;\beta)) = X^\top W_* \mathbb{V}ar(Y) W_* X, \tag{1.24}$$

dove $W_* = \phi^{-1}\text{diag}[1/v(\mu_i)g'(\mu_i)]$.

1.2.3 Proprietà degli stimatori

Come mostrato più nel dettaglio in Salvan et al. (2020), anche sotto le più deboli ipotesi (1.14)–(1.16), continua a valere l'identità dell'informazione

$$\mathbb{E}_\beta(l_r l_s) = -\mathbb{E}_\beta(l_{rs}), \qquad \text{con } l_{rs} = \frac{\partial}{\partial \beta_s} l_r, \tag{1.25}$$

ovvero si ottiene che $J(\beta) = H(\beta)$. Attraverso uno sviluppo di Taylor al primo ordine di $q(y;\hat{\beta})$ e sfruttando una legge dei grandi numeri tale per cui

$$-\frac{\partial q(Y;\beta)}{\partial \beta^T} \doteq H(\beta), \tag{1.26}$$

al divergere di n, si ha che

$$\hat{\beta} - \beta \sim N_p(0, H(\beta)^{-1} J(\beta) H(\beta)^{-1}). \tag{1.27}$$

Risulta pertanto che

$$\mathbb{V}ar(\hat{\beta}) \doteq (X^\top W X)^{-1} X^\top W_* \mathbb{V}ar(Y) W_* X (X^\top W X)^{-1}. \tag{1.28}$$

Assumendo valida l'ipotesi sulla varianza (1.15), grazie all'identità dell'informazione tale per cui $J(\beta) = H(\beta)$ si ottiene $\mathbb{V}ar(\hat{\beta}) = (X^\top W X)^{-1}$ e dunque

$$\hat{\beta} \dot{\sim} N_p(\beta, (X^\top W X)^{-1}). \tag{1.29}$$

1.3 Modelli per risposte correlate

In studi longitudinali o con dati a gruppi (*cluster*), la risposta per l'unità i-esima è multivariata e, dunque, pensata come realizzazione di un vettore casuale a componenti dipendenti. La notazione comunemente utilizzata prevede che le osservazioni y_{ij} sulla risposta dell'i-esima unità statistica $i = 1, \ldots, n$, siano realizzazione di variabili casuali Y_{ij} correlate, $j = 1, \ldots, m$. In generale, per ciascun soggetto vengono rilevate m_i osservazioni, che diventano $m_i = m$ misurazioni, indipendenti da i, nel caso di disegni bilanciati. In questo caso il numero totale di osservazioni è pari a $N = n \times m$. Sia inoltre

$$y_i = [y_{i1}, \ldots, y_{im_i}]^\top$$

il vettore di osservazioni della risposta relative alla unità i-esima e $\mathbf{x}_{ij}$ il vettore riga p-dimensionale di variabili esplicative per l'osservazione j-esima sull'unità i-esima. Si indica infine con $\mathbb{E}(Y_{ij}) = \mu_{ij}$ il valore atteso marginale.

Una delle principali tipologie di modelli per lo studio di risposte correlate sono i modelli con effetti individuali. Alla base di questi modelli vi è l'assunzione che le osservazioni relative alla stessa unità condividano delle caratteristiche comuni non osservabili. Tali caratteristiche possono essere analizzate attraverso modelli a effetti fissi o mediante modelli con effetti casuali. Qualora siano presenti sia effetti fissi che effetti casuali si parla di modelli misti. Un'altra importante classe di modelli sono i modelli marginali, che descrivono l'effetto delle esplicative sui valori attesi marginali μ_{ij}, in cui la correlazione tra osservazioni della stessa unità viene tipicamente considerata un elemento di disturbo, di cui tuttavia bisogna tenere conto durante la modellazione.

1.3.1 Modelli con effetti casuali

Nei modelli con effetti casuali le osservazioni relative alla stessa unità condividono il valore di una stessa variabile aleatoria, detta effetto casuale, la cui presenza induce correlazione tra le misurazioni eseguite sullo stesso soggetto. Modelli di questo tipo permettono di modellare due effetti sulla variabile risposta, uno tra unità, che dipende solo da variabili esplicative relative all'unità i-esima e l'altro entro le unità.

La specificazione di un modello lineare normale con effetti misti risulta

$$Y_{ij} = \mathbf{x}_{ij}\beta + \mathbf{z}_{ij}\mathbf{u}_i + \epsilon_{ij}, \tag{1.30}$$

dove β è un vettore p-dimensionale di effetti fissi, $\mathbf{u}_i \sim N_q(0, \Sigma_u)$ è un vettore q-dimensionale di effetti casuali, mentre marginalmente $\epsilon_{ij} \sim N(0, \sigma_\epsilon)$, indipendente da

$\mathbf{u}_i$. Il modello (1.30) prevede dunque che $\mathbb{E}(Y_{ij}) = \mu_{ij} = \mathbf{x}_{ij}\beta$. Il vettore dei parametri β associato al vettore $\mathbf{x}_{ij}$, che comprende le variabili esplicative ed eventuale intercetta, permette di esprimere effetti fissi tra unità ed effetti fissi entro le unità. Inoltre il termine $\mathbf{z}_{ij}\mathbf{u}_i$ esprime la variabilità tra unità, mentre ϵ_{ij} descrive la variabilità delle osservazioni relative all'unità i-esima.

Un'estensione del modello lineare normale con effetti misti, formulabile anche qualora la variabile risposta sia dicotomica o di conteggio, è data dal modello lineare generalizzato con effetti misti (o GLMM, *generalized linear mixed effects*). Tale modello prevede che, condizionatamente a $\mathbf{u}_i$, le osservazioni sulla risposta Y_{ij} siano indipendenti e distribuite secondo un modello lineare generalizzato con

$$g\left(\mathbb{E}(Y_{ij}|\mathbf{u}_i)\right) = \mathbf{x}_{ij}\beta + \mathbf{z}_{ij}\mathbf{u}_i, \tag{1.31}$$

dove $g(\cdot)$ è la funzione di legame di un modello lineare generalizzato, mentre gli effetti casuali $\mathbf{u}_i$ si assume ticpicamente che siano realizzazioni indipendenti di una $N_q(0, \Sigma_u)$. In ultima battuta, rimangono valide le osservazioni fatte sui vettori $\mathbf{x}_{ij}$ e $\mathbf{z}_{ij}$ del modello lineare normale ad effetti misti di cui alla (1.30).

1.3.2 Modelli marginali

Si consideri inizialmente una variabile risposta continua. Sia $\mathbf{y}$ il vettore N-dimensionale delle risposte di n soggetti

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{bmatrix}.$$

Assumendo che $\mathbf{y}_1, \ldots, \mathbf{y}_n$ siano realizzazioni di vettori casuali indipendenti $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, data la tipologia della risposta è naturale formulare un modello lineare normale multivariato tale che $\mathbf{Y}_i \sim N_{m_i}(\boldsymbol{\mu}_i, V_i)$, $i = 1, \ldots, n$. Fissata l'unità i-esima, si assume $\boldsymbol{\mu}_i = \mathbf{X}_i\beta$, con β vettore di parametri p-dimensionale, $\mathbf{X}_i$ matrice del modello $m_i \times p$, dove la j-esima riga è pari a $\mathbf{x}_{ij}$, $j = 1, \ldots, m_i$. Si indica inoltre con $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_n]^\top$ la matrice di disegno complessiva di dimensione $N \times p$. Per quanto detto, il vettore aleatorio

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_n \end{bmatrix}$$

ha distribuzione normale multivariata N-dimensionale, con media $\boldsymbol{\mu} = \left[\boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_n^\top\right]^\top$ e matrice di covarianza

$$\mathbf{V} = \begin{bmatrix} V_1 & 0 & \cdots & 0 \\ 0 & V_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V_n \end{bmatrix}.$$

Si noti come, dalla struttura diagonale a blocchi di $\mathbf{V}$, rimanga ancora valida l'ipotesi di indipendenza tra Y_{ij} e Y_{ik} $i, k = 1, \ldots, n$, $i \neq k$. Inoltre se $V_1 = V_2 = \cdots = V_n$ si assume omoschedasticità. Si evidenzia tuttavia come il modello di regressione multivariata sopra specificato richieda di conoscere la struttura di V. Se così non fosse, si incorrerebbe nel problema computazionalmente non banale di dover stimare $\frac{m_i(m_i+1)}{2}$ parametri ignoti. Poichè nelle applicazioni reali la matrice di covarianza $\mathbf{V}$ non è nota, si ipotizza per essa una qualche struttura volta a ridurre il numero di parametri da stimare e che ne semplifichi la specificazione. Nel seguito, si assume per semplicità $m_i = m$ e $\mathbb{V}ar(Y_{ij}) = \sigma^2$, imponendo solo una struttura per le matrici V_i. La più semplice struttura che si può assumere per V_i è quella che prevede l'indipendenza tra osservazioni relative all'unità i-esima, ovvero

$$V_i = \sigma^2 \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}. \tag{1.32}$$

Nel seguito la struttura di correlazione (1.32) verrà indicata con I. Si vuole sottolineare che, nonostante porti notevoli semplificazioni dal punto di vista del calcolo, non è in generale ragionevole supporre una struttura di correlazione di questo tipo in quanto non considera la correlazione sussistente tra le osservazioni. Per ovviare a tale mancanza si possono ipotizzare altre strutture di cui, nel seguito, si presentano le principali specificazioni.

- **Equicorrelazione**,

$$V_i = \sigma^2 \begin{bmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{bmatrix}. \tag{1.33}$$

In questo caso tutti gli elementi fuori diagonale sono uguali e pari al parametro α, detto coefficiente di correlazione intraclasse, $\alpha \in (\frac{-1}{m-1}, 1)$. La struttura della matrice di correlazione (1.33) è detta scambiabile o sferica.

- **Autoregressiva**,

$$V_i = \sigma^2 \begin{bmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{m-1} \\ \alpha & 1 & \alpha & \cdots & \alpha^{m-2} \\ \alpha^2 & \alpha & 1 & \cdots & \alpha^{m-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha^{m-1} & \cdots & \cdots & \alpha & 1 \end{bmatrix}. \tag{1.34}$$

Questa struttura trova ampia applicazione in studi di tipo longitudinale, in cui si presume che, ragionevolmente, la correlazione tra osservazioni relative alla stessa unità tenda a diminuire all'aumentare del *lag* temporale. La (1.34) corrisponde a una struttura autoregressiva del primo ordine, in breve AR(1), tale per cui, $Cor(Y_{ij}, Y_{ik}) = \alpha^{|j-k|}$, $\alpha \in (-1, 1)$.

- **Non strutturata**,

$$V_i = \sigma^2 \begin{bmatrix} 1 & \alpha_{12} & \cdots & \alpha_{1m} \\ \alpha_{21} & 1 & \cdots & \alpha_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{m1} & \cdots & \cdots & 1 \end{bmatrix}. \tag{1.35}$$

Se m risulta molto minore di n è possibile non specificare alcuna struttura per V_i, stimando $\frac{m(m-1)}{2}$ parametri di correlazione.

- **Stazionaria** o di **Toepliz**,

$$V_i = \sigma^2 \begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} & \cdots & \alpha_{1m} \\ \alpha_{21} & 1 & \alpha_{12} & \cdots & \vdots \\ \alpha_{31} & \alpha_{21} & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \alpha_{12} \\ \alpha_{m1} & \cdots & \cdots & \alpha_{21} & 1 \end{bmatrix}. \tag{1.36}$$

Una matrice siffatta rappresenta la generalizzazione di quelle specificate nella (1.33) e (1.34), nel senso che queste possono essere facilmente ricavate ponendo opportuni vincoli sui coefficienti α_{ij}. Ad esempio, imponendo il vincolo $\alpha_{12} = \cdots = \alpha_{1m} = \alpha$ si ottiene una struttura sferica.

Premettendo che nel caso in cui la risposta sia binaria o di conteggio, le corrispondenti distribuzioni multivariate risultano essere complesse e poco flessibili, Liang & Zeger (1986) introducono, generalizzando la teoria della quasi-verosimiglianza, la metodologia

delle equazioni di stima generalizzate (GEE) come estensione dei GLM per l'analisi di risposte multivariate correlate, sia continue che discrete. Nello specifico, considerando il vettore risposta $\mathbf{y}_i$, si possono generalizzare le ipotesi del secondo ordine (1.14)-(1.16) come segue

$$\mathbb{E}(\mathbf{Y}_i) = \boldsymbol{\mu}_i, \text{con } g(\mu_{ij}) = \mathbf{x}_{ij}\beta, \tag{1.37}$$

$$\mathbb{V}ar(\mathbf{Y}_i) = V_i = \phi A_i^{1/2} R(\alpha) A_i^{1/2}, \tag{1.38}$$

$$\mathbf{Y}_i \text{ e } \mathbf{Y}_h \text{ sono indipendenti se } i \neq h, \tag{1.39}$$

dove con μ_{ij} si indica il generico valore del vettore $\boldsymbol{\mu}_i$ e con $g(\cdot)$ la funzione di legame marginale. Inoltre, $A_i = \text{diag}(v(\mu_{ij}))$ e $R(\alpha)$ è la matrice di correlazione di $\mathbf{Y}_i$ dipendente dai parametri di correlazione α. Le equazioni di stima generalizzate, sfruttando il partizionamento a blocchi della (1.9), risultano

$$\sum_{i=1}^{n} D_i V_i^{-1} [\mathbf{y}_i - \boldsymbol{\mu}_i] = 0, \tag{1.40}$$

dove D_i è la matrice $m \times p$ con generico elemento dato da $\frac{\partial \mu_{ij}}{\partial \beta_r}$, $j = 1, \ldots, m$ e $r = 1, \ldots, p$. La soluzione alle (1.40) rispetto a β viene trovata attraverso algortimi iterativi, sostituendo ad α e ϕ le corrispondenti stime basate sul metodo dei momenti. Si dimostra inoltre che sotto condizioni di regolarità, tra le quali la consistenza degli stimatori di ϕ e α, lo stimatore β ottenuto come soluzione della (1.40) è asintoticamente normale con matrice di covarianza

$$\mathbb{V}ar(\beta)_{gee} = \left[\sum_{i=1}^{n} D_i^\top V_i^{-1} D_i\right]^{-1} \left[\sum_{i=1}^{n} D_i^\top V_i^{-1} \mathbb{V}ar(\mathbf{Y}_i) V_i^{-1} D_i\right] \left[\sum_{i=1}^{n} D_i^\top V_i^{-1} D_i\right]^{-1}. \tag{1.41}$$

con $\mathbb{V}ar(\mathbf{Y}_i)$ che può essere stimata in maniera robusta da $[\mathbf{y}_i - \boldsymbol{\mu}_i][\mathbf{y}_i - \boldsymbol{\mu}_i]^\top$. Lo stimatore della matrice (1.41), ottenuto sostituendo β con $\hat{\beta}$, α e ϕ con le relative stime, viene comunemente detto stimatore *sandwich*.

Capitolo 2

Criteri di selezione della struttura di correlazione per modelli GEE

2.1 Introduzione

Lo stimatore basato sulle equazioni di stima generalizzate risulterà più efficiente nel caso in cui la struttura di correlazione sia più vicina alla vera struttura sottostante tra un insieme di opzioni disponibili. In assenza di una verosimiglianza parametrica, i metodi tradizionali di selezione del modello basati sulla verosimiglianza non possono essere utilizzati per confrontare i modelli GEE. Allo stesso modo non possono essere utilizzati per scegliere la migliore *working correlation structure*. Recentemente si è cercato di ovviare a tale problema, sviluppando estensioni basate sulla quasi-verosimiglianza (Pan (2001) e Hin & Wang (2008)), sulla discrepanza tra matrici di covarianza (Gosho et al. (2011) e Pardo & Alonso (2019)) e su autovettori generalizzati (Jang (2011) e Carey & Wang (2011)). Chen & Lazar (2012) propongono infine un approccio basato sulla verosimiglianza empirica. L'obiettivo di questo capitolo è fornire una panoramica completa circa i criteri di selezione presenti in letteratura, introducendo dapprima un excursus teorico alla base degli stessi per poi entrare nel merito di ogni singolo criterio.

2.1.1 Divergenza di Kullback-Leibler

Dato un insieme di potenziali modelli $\mathbb{K}$, ciascuno indicizzato da un vettore di parametri β e da ϕ, nel seguito assunto noto, una misura della separazione tra il vero modello $\mathbf{M}_*$ e quello ipotizzato $\mathbf{M}_1$, rispettivamente con funzione di log-verosimiglianza $l(\beta_*, \phi)$ e $l(\beta, \phi)$, è data dalla divergenza di Kullback-Leibler, o cross-entropia. L'informazione di

Kullback-Leibler tra $\mathbf{M}_1$ e $\mathbf{M}_*$ risulta

$$\Delta_0(\beta, \beta_*) = \mathbb{E}_{\mathbf{M}_*}[-2l(\beta, \phi)], \tag{2.1}$$

dove il valore atteso viene calcolato rispetto alla vera distribuzione, ovvero sotto il vero modello $\mathbf{M}_*$. Intuitivamente rappresenta una misura dell'informazione persa nell'approssimare la vera distribuzione relativa a $\mathbf{M}_*$ mediante il modello corrente $\mathbf{M}_1$. Risulta dunque preferibile quel modello $\mathbf{M} \in \mathbb{K}$ che presenta il più piccolo valore di $\Delta_0(\beta, \beta_*)$. Si noti tuttavia come tale quantità vada stimata, essendo sia β che β_* ignoti. L'**AIC** (*Akaike Information Criterion*) nasce dal tentativo di trovare uno stimatore asintoticamente non distorto per $\mathbb{E}_{\mathbf{M}_*}[\Delta_0(\hat{\beta}, \beta_*)]$, dove $\hat{\beta}$ è lo stimatore di massima verosimiglianza per β. Akaike (1973) propone di utilizzare il criterio omonimo espresso dalla (2.2).

$$\mathbf{AIC} = -2l(\hat{\beta}, \phi) + 2p, \tag{2.2}$$

dove p è la dimensione del vettore β, per la selezione di modelli non annidati. Basandosi invece su un approccio di tipo bayesiano, è possibile definire il criterio del minimo **BIC** (*Bayesian Information Criterion*), dove

$$\mathbf{BIC} = -2l(\hat{\beta}, \phi) + p \log n. \tag{2.3}$$

Si fa presente che l'**AIC** può portare alla selezione di un modello sovraparametrizzato, mentre per n non elevato il **BIC** tende a selezionare un modello leggermente sottoparametrizzato. Per maggiori dettagli si rimanda a Salvan et al. (2020).

2.1.2 Verosimiglianza empirica

La metodologia basata sulla verosimiglianza empirica, introdotta da Owen (1990), è un metodo di inferenza non parametrico il cui principale vantaggio è utilizzare metodi di verosimiglianza senza specificare una forma per la distribuzione dei dati. In questo senso si ottengono dunque intervalli di confidenza che riflettono naturalmente la forma dei dati impiegati, assegnando un peso maggiore alle informazioni che attribuiscono maggiore plausibilità al parametro di interesse. Siano $Y_1, \ldots, Y_n$ variabili casuali indipendenti provenienti da una qualche distribuzione F_0, allora la funzione di ripartizione empirica F_n è la stima di massima verosimiglianza di F_0 in quanto massimizza la funzione di verosimiglianza

$$L(F) = \prod_{i=1}^{n} F(Y_i) - F(Y_i-) = \prod_{i=1}^{n} \mathbb{P}(Y_i = y) \tag{2.4}$$

rispetto a tutte le possibili funzioni di ripartizione F. Nella (2.4) si è usato $F(Y_i-)$ per indicare $\mathbb{P}(Y_i < y)$. Per una distribuzione F, si può inoltre definire il rapporto di verosimiglianza empirica come

$$R(F) = \frac{L(F)}{L(F_n)}. \tag{2.5}$$

É possibile applicare la metodologia della verosimiglianza non parametrica anche nel contesto di dati longitudinali, o più in generale, nel caso di dati correlati. Supponendo di avere come parametro di interesse β e definendo la funzione di stima

$$u([\mathbf{Y}_i, \mathbf{X}_i], \beta) = \sum_{j=1}^{m_i} w((Y_{ij}, X_{ij}), \beta), \;\; i = 1, \ldots, n, \tag{2.6}$$

con $\mathbb{E}\left[u([\mathbf{Y}_i, \mathbf{X}_i], \beta)\right] = 0$ e con $w(\cdot)$ funzione non meglio specificata dipendente dai dati e da β, il log rapporto di verosimiglianza empirica per β risulta

$$\mathcal{R}(\beta) = \sup \left\{ \prod_{i=1}^{n} np_i : p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i u([\mathbf{Y}_i, \mathbf{X}_i], \beta) = 0 \right\}. \tag{2.7}$$

La (2.7) consiste nell'assegnazione di p_i pesi di probabilità, a n osservazioni $(\mathbf{Y}_i, \mathbf{X}_i)$ indipendenti, $i = 1, \ldots, n$. Inoltre, poiché i pesi p_i non vengono attribuiti direttamente alla singola osservazione (Y_{ij}, X_{ij}) non viene imposta alcuna assunzione di indipendenza tra misurazioni afferenti alla stessa unità. D'altra parte essendo $u([\mathbf{Y}_i, \mathbf{X}_i]), \beta)$ una semplice somma, la (2.6) non tiene conto della correlazione tra osservazioni relative alla stessa unità, che quindi vengono trattate tutte allo stesso modo. Si può dunque estendere la (2.7), per definire un rapporto di verosimiglianza empirica nell'ambito delle GEE, includendo nella sua definizione una struttura di correlazione $R(\alpha)$ come segue

$$u([\mathbf{Y}_i, \mathbf{X}_i], \beta; R) = \left[\frac{\partial \mu_i}{\partial \beta^\top} \right] \phi^{-1} A_i^{-1/2} R^{-1}(\alpha) A_i^{-1/2} (\mathbf{Y}_i - \boldsymbol{\mu}_i). \tag{2.8}$$

Si noti che la quantità a destra dell'uguale è il classico addendo GEE del tipo (1.40), e dunque dipende dalla *working correlation matrix* specificata. Ad esempio scegliendo $R(\alpha) = I$, la (2.8) si riduce nella forma (2.6). Infine si definisce stimatore di massima verosimiglianza empirica per β

$$\hat{\beta}_E = \arg\max_{\beta \in \mathbb{R}^p} \mathcal{R}(\beta). \tag{2.9}$$

Nel caso in cui la dimensione del parametro β e quella della funzione $u([\mathbf{Y}_i, \mathbf{X}_i], \beta; R)$ nella (2.8) siano entrambe p è possibile dimostrare che $\hat{\beta}_E = \hat{\beta}_{gee}$ (per maggiori dettagli

si rimanda a Chen & Lazar (2012)). Qin & Lawless (1994) hanno dimostrato che la statistica log-rapporto di verosimiglianza empirica $W_E(\beta) = -2\log\mathcal{R}(\beta) - \left[-2\log\mathcal{R}(\hat{\beta}_E)\right]$ tende in distribuzione ad una χ^2_p, attraverso il quale è possibile costruire regioni di confidenza per β.

2.1.3 Autovettori e autovalori generalizzati

Per introdurre in maniera adeguata i criteri di selezione nel Paragrafo 2.2, è necessario richiamare alcuni concetti fondamentali di algebra lineare, che costituiranno la base teorica necessaria per comprendere appieno alcuni dei argomenti trattati. Gli autovettori e autovalori generalizzati rappresentano la naturale estensione degli "ordinari" autovettori e autovalori. Questi ultimi misurano le proprietà di matrici, catturandone i diversi aspetti sinteticamente. Sia A una matrice quadrata $p \times p$ e $\mathbf{x}$ un vettore p-dimensionale tale che $A\mathbf{x} = \lambda\mathbf{x}$. Allora λ rappresenta l'autovalore di A relativo all'autovettore $\mathbf{x}$. In altre parole $\mathbf{x}$ è un vettore che viene mappato in se stesso da A a meno di un fattore λ, ovvero che cambia al più scala e verso, ma non direzione. Se A è una matrice di covarianza, allora la somma degli autovalori, detta traccia e denotata con $\mathrm{tr}(\cdot)$, è interpretabile come misura totale di variabilità o di dispersione nello spazio multidimensionale. D'altra parte, il prodotto degli autovalori, cioè il determinante indicato con $|\cdot|$, viene solitamente ribattezzato varianza generalizzata e rappresenta anch'essa una misura complessiva di dispersione nello spazio p-dimensionale.

In generale, sia B una matrice $p \times p$ e $\mathbf{z}$ un vettore p-dimensionale con λ scalare che verifica l'equazione $A\mathbf{z} = \lambda B\mathbf{z}$. Allora si dice che $\mathbf{z}$ è un autovettore generalizzato di A rispetto a B, o nella metrica di B, relativo all'autovalore generalizzato λ. Gli autovettori e autovalori generalizzati non descrivono più solo una singola matrice, bensì permettono di operare confronti tra matrici diverse. Nello specifico, se A e B sono matrici di covarianza l'insieme degli autovalori generalizzati di A rispetto a B coglie la dispersione dei punti di $\mathbf{x}$ relativamente alla variabilità dei punti $\mathbf{z}$ nello spazio p-dimensionale: ciò significa che è possibile misurare la "similarità" tra matrici di covarianza guardando come la variabilità dei punti di una è espressa nella metrica dell'altra. Quanto più le matrici A e B saranno tra loro comparabili, tanto più gli autovalori generalizzati saranno vicini a uno, indicando forte "similarità" tra esse. Infine se λ è un autovalore generalizzato di A rispetto a B, si indica con

$$\kappa = \frac{\lambda}{1+\lambda} \tag{2.10}$$

l'autovalore di A rispetto a $A + B$ relativo all'autovettore $\mathbf{x}$.

2.2 Criteri di selezione

Nell'ambito delle equazioni di stima generalizzate, qualora si specifichi un modello con struttura di indipendenza tra osservazioni relative alla stessa unità, ovvero $R(\alpha) = I$, è possibile riscrivere la quasi-verosimiglianza (1.18) nella forma

$$Q(\beta, \phi, I, \mathcal{D}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} Q(\beta, \phi, I, (Y_{ij}, X_{ij})), \tag{2.11}$$

dove con $\mathcal{D}$ si indicano i dati in modo compatto. Poiché in modelli basati su GEE non vi è verosimiglianza, in quanto non si specifica una distribuzione congiunta delle osservazioni fissata l'unità, i classici metodi di selezione del modello o della struttura di correlazione non possono essere utilizzati. Pan (2001), nel tentativo di estendere l'**AIC**, sostituendo la veromiglianza nella (2.1) con la quasi-verosimiglianza di cui alla (2.11), propone una nuova discrepanza della forma

$$\Delta_0(\beta, \beta_*, I, \mathcal{D}) = \mathbb{E}_{\mathbf{M}_*}[-2Q(\beta, \phi, I, \mathcal{D})]. \tag{2.12}$$

Si noti che per la quasi-verosimiglianza (2.11), ponendo $\beta = \beta_*$, la (1.23) è esprimibile come segue

$$\Omega_I = -\mathbb{E}_{\beta_*}\left[\frac{\partial q(Y;\beta)}{\partial \beta^\top}\bigg|_{\beta=\beta_*}\right] = \sum_{i=1}^{n} D_i^\top V_i D_i\bigg|_{R(\alpha)=I}, \tag{2.13}$$

con Ω_I matrice semidefinita positiva è V_i della forma (1.38). Posto $\beta = \beta_*$, sfruttando la (1.19) e la convessità di Ω_I è possibile dimostare che β_* è minimizzatore locale della funzione $\Delta_0(\beta, \beta_*, I, \mathcal{D})$, ovvero che

$$\Delta_0(\beta, \beta_*, I, \mathcal{D}) \geq \Delta_0(\beta_*, \beta_*, I, \mathcal{D}). \tag{2.14}$$

$\Delta_0(\beta, \beta_*, I, \mathcal{D})$ risulta dunque ben definita per ogni β in un intorno di β_*. Considerando ora come β, lo stimatore $\hat{\beta}_{gee} = \hat{\beta}_{gee}(R(\alpha))$ ottenuto sotto una qualsiasi struttura di correlazione $R(\alpha)$, è possibile approssimare $\mathbb{E}_{M_*}\left[\Delta_0(\hat{\beta}_{gee}, \beta_*, I, \mathcal{D})\right]$ attraverso uno sviluppo in serie di Taylor al secondo ordine

$$\begin{aligned}\mathbb{E}_{M_*}\left[\Delta_0(\hat{\beta}_{gee}, \beta_*, I, \mathcal{D})\right] = &-2\mathbb{E}_{\mathbf{M}_*}[Q(\beta_{gee}, \phi, I, \mathcal{D})] \\ &+ 2\mathbb{E}_{\mathbf{M}_*}[(\hat{\beta}_{gee} - \beta_*)^\top S(\hat{\beta}_{gee}, I, \mathcal{D})] \\ &+ 2\mathrm{tr}(\Omega_I \mathbb{V}ar(\beta)_{gee}).\end{aligned} \tag{2.15}$$

Ω_I viene stimata dalla (2.13) ponendo $\beta = \hat{\beta}_{gee}$, mentre si utilizza lo stimatore *sandwich* per la stima di $\mathbb{V}ar(\beta)_{gee}$. Si noti che, per $\hat{\beta}_{gee} = \hat{\beta}_{gee}(R(\alpha))$, il gradiente $S(\hat{\beta}_{gee}, I, \mathcal{D})$ deve essere stimato, annullandosi solo nel caso in cui $R(\alpha) = I$. Essendo tuttavia difficile da stimare, viene proposto lo stimatore semplificato per $\mathbb{E}_{M_*}\left[\Delta_0(\hat{\beta}_{gee}, \beta_*, I, \mathcal{D})\right]$, in cui tale componente viene omessa, della forma

$$\mathbf{QIC}(R) \equiv -2Q(\hat{\beta}_{gee}(R), \tilde{\phi}, I, \mathcal{D}) + 2\text{tr}(\hat{\Omega}_I Var(\hat{\beta})_{gee}), \tag{2.16}$$

detto criterio di quasi-verosimiglianza sotto il modello di indipendenza. La scelta di formulare tale criterio sotto ipotesi di indipendenza è dettata da motivazioni squisitamente matematiche. E' infatti possibile dimostrare che solo qualora la matrice di correlazione $R(\alpha)$ sia l'identità, l'integrale (1.18) risulta unico, mentre tale risultato non è più garantito con una struttura $R(\alpha)$ qualsiasi (per approfondimenti si rimanda a (McCullagh & Nelder, 1989, p. 334-336)). Ignorare il secondo termine della (2.15) non influenza particolarmente l'efficacia del criterio, anche se è dimostrabile che il comportamento migliore si ottiene per modelli con struttura di indipendenza. Per una trattazione più dettagliata si rimanda a Pan (2001). Nel caso di indipendenza esso rappresenta anche uno stimatore asintoticamente non distorto per la (2.16). Il **QIC** può essere usato sia per la scelta del modello, che per la selezione di una matrice di correlazione. Nello specifico, tra diverse strutture candidate, si sceglierà quella matrice di correlazione che renderà minimo il valore del **QIC**.

Da uno sguardo alla (2.16) si nota che il primo addendo del **QIC** non dipende dalla struttura di correlazione ipotizzata, basandosi sulla quasi-verosimiglianza specificata per osservazioni indipendenti. Anzi, gli errori casuali dovuti alla stima del primo addendo introducono del "rumore" che può deteriorare la performance del criterio stesso (si veda Wang & Hin (2010)). Si noti d'altra parte che il secondo termine della (2.16) dipende dalla struttura $R(\alpha)$ specificata attraverso $\mathbb{V}ar(\beta)_{gee}$. Eliminato dunque il primo termine della (2.16), che non porta informazioni sulla struttura di correlazione, Hin & Wang (2008) propongono una modificazione del **QIC** che tenga conto solo del secondo termine, ovvero

$$\mathbf{CIC}(R) = \text{tr}(\hat{\Omega}_I \mathbb{V}ar(\hat{\beta})_{gee}), \tag{2.17}$$

detto criterio di informazione di correlazione. Attraverso la (2.17) è facile mostrare che $\mathbf{QIC}(R(\alpha)) = -2Q(\hat{\beta}_{gee}(R(\alpha)), \tilde{\phi}, I, D) + 2\mathbf{CIC}(R(\alpha))$. L'eliminazione della distorsione dovuta al primo addendo della (2.16) porta in generale, ad un aumento dell'efficacia del **CIC** rispetto al **QIC** nel selezionare la vera struttura di correlazione. Inoltre, per variabili risposta di tipo continuo, in cui $v(\mu_{ij}) = 1$ e $\phi = \tilde{\phi}$, si può mostrare che **QIC**

risulta essere una trasformazione affine di **CIC** i.e., $\mathbf{QIC}(R(\alpha)) = (N-p)+2\mathbf{CIC}(R(\alpha))$, il che significa che i due criteri sono equivalenti in termini di efficacia come si avrà modo di vedere nel Capitolo 3.

Nel caso in cui la matrice di correlazione $R(\alpha)$ venga correttamente specificata, ci si può aspettare che la somma dei quadrati dei residui pesati per l'inversa della matrice di covarianza $\mathbb{V}ar(\beta)_{gee}$ sia minima. Partendo da questa intuizione, Shults & Chaganty (1998) propongono quale criterio di selezione della struttura di correlazione quello che minimizza la funzione

$$\mathbf{SC}(R(\alpha)) = \sum_{i=1}^{n} [\mathbf{Y}_i - \boldsymbol{\mu}_i]^\top V_i^{-1} [\mathbf{Y}_i - \boldsymbol{\mu}_i]|_{\beta=\hat{\beta}(R),\, \phi=\tilde{\phi}(R(\alpha))}, \tag{2.18}$$

dove V_i assume la forma (1.38).

Rotnitzky & Jewell (1990) propongono una statistica di Wald generalizzata per verificare l'ipotesi nulla $H_0 : \beta = \beta_0$ contro $H_1 : \beta \neq \beta_0$, con β_0 vettore di parametri fissati, basata sulla statistica $\Psi = \Psi_0^{-1}\Psi_1$ dove

$$\Psi_0 = \frac{1}{n} \sum_{i=1}^{n} D_i V_i^{-1} D_i, \tag{2.19}$$

$$\Psi_1 = \frac{1}{n} \sum_{i=1}^{n} D_i V_i^{-1} S_i S_i^\top V_i D_i, \tag{2.20}$$

con $S_i = [\mathbf{y}_i - \boldsymbol{\mu}_i]$. Nel caso in cui la struttura di correlazione venga correttamente specificata ci si aspetta che la (2.20) si riduca alla (2.19) e dunque che Ψ sia "vicina" alla matrice identità p-dimensionale. Sfruttando questa idea, Hin. et al. (2007) definiscono il criterio di Rotnitzky e Jewell basandosi sulla quantità

$$\mathbf{RJ}(R(\alpha)) = \sqrt{(1-C1)^2 + (1-C2)^2}|_{\beta=\hat{\beta}(R(\alpha)),\, \phi=\tilde{\phi}(R(\alpha))}, \tag{2.21}$$

dove $C1 = \mathrm{tr}(\Psi)/p$ e $C2 = \mathrm{tr}(\Psi^2)/p$. Un valore di $\mathbf{RJ}(R(\alpha))$ vicino a zero suggerisce che la matrice di correlazione è stata correttamente specificata. Inoltre, basandosi sulle quantità $C1$ e $C2$, Carey & Wang (2011) definiscono due criteri basati sulla distanza geodesica della forma

$$\Delta_1(R(\alpha)) = \sum_{i=1}^{p} \frac{(\lambda_i - 1)^2}{p}, \tag{2.22}$$

e

$$\Delta_2(R(\alpha)) = \sum_{i=1}^{p} \log(\lambda_i)^2, \tag{2.23}$$

dove λ_i sono gli autovalori della matrice Ψ. La ratio sottesa a questi due criteri è che nel caso in cui $R(\alpha)$ approssimi la vera struttura di correlazione, gli autovalori λ_i dovrebbero essere tutti vicini a 1, ovvero il valore di Δ_1 e Δ_2 dovrebbe avvicinarsi a zero. Si vuole infine sottolineare come ci si attenda che l'efficacia di **RJ**, Δ_1 e $\Delta 2$ sia simile essendo tutte basate sulla statistica Ψ.

Come si evince dalla (1.38), la matrice di covarianza V_i ipotizzata dipende dalla struttura di correlazione $R(\alpha)$ specificata. Pertanto, cercare la migliore struttura di correlazione tra diverse opzioni disponibili, è equivalente a specificare la matrice di covarianza V_i che sia il più simile possibile a quella stimata da $\mathbb{V}ar(\mathbf{Y}_i) = [\mathbf{y}_i - \boldsymbol{\mu}_i][\mathbf{y}_i - \boldsymbol{\mu}_i]^\top$. In quest'ottica, Gosho et al. (2011) propongono di scegliere quale migliore struttura di correlazione quella che minimizza la quantità

$$\mathbf{G}(R(\alpha)) = \text{tr}\left\{ \left[\frac{1}{n} \left(\sum_{i=1}^{n} S_i S_i^\top \right) \left(\frac{1}{n} \sum_{i=1}^{n} V_i \right)^{-1} - I_n \right]^2 \right\}, \tag{2.24}$$

dove $S_i = [\mathbf{y}_i - \boldsymbol{\mu}_i]$, mentre con I_n si è indicata la matrice identità n-dimensionale. La (2.24) viene detta criterio di Gosho-Hamada-Yoshimura.

Carey & Wang (2011) suggeriscono di scegliere la matrice di correlazione che porta a massimizzare la pseudo-verosimiglianza Gaussiana espressa da

$$L_{CW} = -\frac{1}{2} \sum_{i=1}^{n} \left\{ [\mathbf{Y}_i - \boldsymbol{\mu}_i]^\top V_i^{-1} [\mathbf{Y}_i - \boldsymbol{\mu}_i] + \log(|V_i|) \right\} |_{\beta = \hat{\beta}(R(\alpha)), \phi = \tilde{\phi}(R(\alpha))}. \tag{2.25}$$

In generale, tuttavia, il numero di parametri da stimare varia al variare della struttura di correlazione specificata. Per tenere conto della dimensione del modello, Zhu & Zhu (2013) propongono di sostituire la pseudo-verosimiglianza (2.25) nei criteri **AIC** e **BIC** di cui alle (2.2) e (2.3) e di definire i due nuovi criteri di pseudo-verosimiglianza Gaussiana

$$\mathbf{GAIC}(R(\alpha)) = -2L_{CW} + 2\text{dim}(\theta), \tag{2.26}$$

$$\mathbf{GBIC}(R(\alpha)) = -2L_{CW} + \log(n)\text{dim}(\theta), \tag{2.27}$$

dove $\theta = [\beta^\top, \alpha^\top]^\top$, $\beta \in \mathbb{R}^p$ e $\alpha \in \mathbb{R}^{m-1}$. Tra diverse opzioni disponibili, viene selezionata la matrice di correlazione che presenta il valore minimo di **GAIC** o **GBIC**.

Sfruttando la flessibilità della verosimiglianza empirica nelle GEE, contesto in cui non vi è verosimiglianza, Chen & Lazar (2012) hanno ampliato i possibili campi applicativi di questa metodologia per cercare di rispondere all'esigenza di selezionare la

matrice di correlazione di lavoro che risulti più adeguata. L'idea è stata quella di comparare modelli con diverse strutture di correlazione utilizzando come metrica unificata il rapporto di verosimiglianza empirica, attribuendo ad ogni modello un "punteggi". Si definisce dapprima modello completo quello basato su una struttura di correlazione stazionaria del tipo (1.36), in seguito definita con $R_F(\alpha)$. Tra C strutture candidate verrà selezionato il modello, e dunque la matrice di correlazione, che meglio si adatta alle osservazioni, ovvero che presenta il valore minore di rapporto di verosimiglianza empirica. Rimpiazzando $R(\alpha)$ con $R_F(\alpha)$ nella (2.8), è possibile definire una nuova funzione di stima $(p+m-1)$-dimensionale per $\theta = [\beta^\top, \alpha^\top]^\top$, della forma

$$g_F([\mathbf{Y}_i, \mathbf{X}_i], \beta, \alpha_1, \ldots, \alpha_{k-1}; R_F(\alpha)) = \begin{bmatrix} \left[\frac{\partial \mu_i}{\partial \beta^\top}\right]^\top A_i^{-\frac{1}{2}} R_F^{-1}(\alpha_1, \ldots, \alpha_{k-1}) A_i^{-\frac{1}{2}} [\mathbf{Y}_i - \boldsymbol{\mu}_i] \\ \sum_{i=1}^{m-1} e_{ij}(\beta) e_{i,j+1}(\beta) - \alpha_1 \hat{\phi}(\beta)(m-1-p/n) \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ \sum_{i=1}^{1} e_{ij}(\beta) e_{i,j+m-1}(\beta) - \alpha_{m-1} \hat{\phi}(\beta)(1-p/n) \end{bmatrix}, \tag{2.28}$$

dove con e_{ij} si sono indicati i residui di Pearson della forma

$$e_{ij}(\beta) = \frac{y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}}, \tag{2.29}$$

mentre ϕ viene stimato attraverso una generalizzazione della (1.13)

$$\hat{\phi} = \frac{1}{N-p} \sum_{i=1}^{n} \sum_{j=1}^{m} \left(\frac{y_{ij} - \mu_{ij}}{\sqrt{v(\mu_{ij})}} \right)^2 = \frac{1}{N-p} \sum_{i=1}^{n} \sum_{j=1}^{m} e_{ij}^2. \tag{2.30}$$

Le ultime $m-1$ righe di $g_F(\cdot)$ sono derivate dallo stimatore con metodo dei momenti per i coefficienti di correlazione di $R_F(\alpha)$,

$$\hat{\alpha}_t = \frac{1}{(n(m-t)-p)\phi} \sum_{i=1}^{n} \sum_{j=1}^{m-t} e_{ij} e_{i,j+t}. \tag{2.31}$$

La funzione $g_F(\cdot)$ di cui alla (2.28) è equivalente a $u(\cdot)$ nella (2.8) con $R(\alpha) = R_F(\alpha)$, nel senso che entrambe creano lo stesso stimatore $\hat{\beta}_{gee}$ per la stima del vettore di parametri β. Dunque, sostituendo $g_F(\cdot)$ al posto di $u(\cdot)$ nella (2.7) si definisce il rapporto di verosimiglianza empirica per il modello completo

$$\mathcal{R}^F(\beta) = \sup \left\{ \prod_{i=1}^{n} np_i : p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i g_F([\mathbf{Y}_i, \mathbf{X}_i], \beta, \alpha; R_F) = 0 \right\}, \tag{2.32}$$

dove $\alpha = [\alpha_1, \ldots, \alpha_{m-1}]$. La motivazione di definire un rapporto di verosimiglianza empirica per il modello completo è duplice. In prima battuta, l'informazione relativa alla struttura di correlazione tra le osservazioni viene costruita sotto le più deboli assunzioni che si possano specificare circa la struttura di correlazione, ovvero specificando la matrice stazionaria $R_F(\alpha)$ per la sottostante vera struttura di correlazione. Si ricorda inoltre che se $R(\alpha)$ è una delle matrici definite nel Paragrafo 1.3.2, allora è deducibile dalla più generale struttura stazionaria, presa in questo contesto come "struttura di riferimento". Sembra dunque ragionevole basare il confronto tra diversi modelli candidati, in termini di rapporto di verosimiglianza empirica per il modello completo. In effetti, guardando la definizione della (2.28), solo se $R(\alpha) = R_F(\alpha)$ il rapporto di verosimiglianza empirica vale 1, ovvero l'equazione $\sum_{i=1}^{n} p_i g_F([\mathbf{Y}_i, \mathbf{X}_i], \beta, \alpha; R_F) = 0$ è verificata, mentre in generale tale risultato non è garantito assumendo una struttura di correlazione diversa da $R_F(\alpha)$. Alla luce di ciò risulta dunque possibile discriminare modelli basati su matrici di correlazioni diverse confrontando i relativi "punteggi" di rapporto di verosimiglianza empirica. Chen & Lazar (2012), rifacendosi a quanto detto, propongono una modificazione dei noti **AIC** (2.2) e **BIC** (2.3) sostituendo la verosimiglianza empirica a quella parametrica, ottenendo

$$\mathbf{EAIC}(R) = -2\log \mathcal{R}^F(\hat{\theta}^c_{gee}) + 2\dim(\theta^c), \tag{2.33}$$

$$\mathbf{EBIC}(R) = -2\log \mathcal{R}^F(\hat{\theta}^c_{gee}) + \log(n)\dim(\theta^c), \tag{2.34}$$

dove c è l'indice del modello candidato parametrizzato da θ^c, $c = 1, \ldots, C$, e $\hat{\theta}^c_{gee}$ rappresenta la stima GEE basata sulla struttura di correlazione di lavoro $R^c(\alpha)$. Si selezionerà quella matrice di correlazione che renderà minimo i valori di **EAIC** e **EBIC**.

L'insieme degli autovalori e autovettori generalizzati fornisce informazioni circa la discrepanza in variabilità di un matrice di covarianza nei confronti di un'altra. Siano $\lambda_1, \ldots, \lambda_p$ gli autovalori relativi agli autovettori generalizzati $a_1, \ldots, a_p$ di $\mathbb{V}ar(\hat{\beta})_{gee}$ rispetto a $\hat{\Omega}_I$, dove lo stimatore sandwich è calcolato sotto la struttura di correlazione $R(\alpha)$ scelta in fase di specificazione del modello, mentre lo stimatore della matrice di covarianza naive è adattato sotto l'assunzione di indipendenza, ovvero con $R(\alpha) = I$. Se le due matrici sono comparabili in dimensione nella direzione a_j, allora λ_j è vicino a 1. Come mostrato in Jang (2011), la matrice $\hat{\Omega}_I$ tende ad presentare una distorsione positiva maggiore rispetto allo stimatore sandwich adattato sotto una qualsiasi struttura di correlazione di lavoro $R(\alpha)$ diversa dall'identità. Inoltre, nonostante rappresenti la più semplice matrice di covarianza che si possa assumere, la matrice $\hat{\Omega}_I$ è al tempo stesso la scelta peggiore che si possa fare, in quanto non tiene conto dell'associazione entro

le unità. D'altra parte, ci si aspetta che $\mathbb{V}ar(\hat{\beta})_{gee}$ colga bene la variabilità dei dati in quanto dipende da una matrice di correlazione $R(\alpha)$ diversa dalla struttura di indipendenza. Detto in altre parole si presume che la dimensione in termini di dispersione di $\mathbb{V}ar(\hat{\beta})_{gee}$ rispetto a $\hat{\Omega}_I$ sia minore, ovvero che il generico autovalore generalizzato λ_j sia piccolo indicando una grande disparità tra le due matrici nella direzione del corrispondente autovettore a_j. L'idea è dunque quella di selezionare la matrice di correlazione che renda massima la discrepanza tra $\mathbb{V}ar(\hat{\beta})_{gee}$ e $\hat{\Omega}_I$ o, equivalentemente, che minimizzi gli autovalori λ_j, $j = 1, \ldots, p$. In generale, un criterio di selezione della matrice di correlazione basato su autovettori generalizzati assume la forma $f_*(\lambda_1, \ldots, \lambda_p; a_j, \ldots, a_p)$. Per la scelta di $f_*(\cdot)$, ispirandosi alla analisi della varianza multivariata, Jang (2011) propone i seguenti criteri, esprimibili in funzione degli autovalori generalizzati κ_j (cfr. 2.10) della matrice $\mathbb{V}ar(\hat{\beta})_{gee}$ rispetto a $\mathbb{V}ar(\hat{\beta})_{gee} + \hat{\Omega}_I^{-1}$.

$$\mathbf{PT}(R(\alpha)) = \text{tr}\left(\mathbb{V}ar(\hat{\beta})_{gee}\left[\mathbb{V}ar(\hat{\beta})_{gee} + \hat{\Omega}_I^{-1}\right]^{-1}\right) = \sum_{j=1}^{p} \kappa_j, \tag{2.35}$$

definito criterio della traccia di Pillai.

$$\mathbf{WR}(R(\alpha)) = \det\left(\mathbb{V}ar(\hat{\beta})_{gee}\left[\mathbb{V}ar(\hat{\beta})_{gee} + \hat{\Omega}_I^{-1}\right]^{-1}\right) = \prod_{j=1}^{p} \kappa_j, \tag{2.36}$$

detto criterio del rapporto di Wilks, e infine

$$\mathbf{RMR}(R(\alpha)) = \max\left\{\kappa_j : j = 1, \ldots, p\right\} \ = \kappa_*, \tag{2.37}$$

denominato criterio della massima radice di Roy. Guardando la forma delle (2.35), (2.36) e (2.37), è facile intuire che verrà selezionato il modello, con struttura $R(\alpha)$, che presenterà il valore minimo di detti criteri. Infine, sfruttando il concetto di varianza generalizzata è possibile confrontare la matrice di covarianza robusta con quella basata sull'ipotesi (1.38). In particolare, utilizzando il rapporto dei relativi determinanti, che rappresenta la via più semplice per confrontare matrici simmetriche e definite positive, Pardo & Alonso (2019) propongono di selezionare la struttura di correlazione che renda tale rapporto il più possibile vicino a 1. Intuitivamente porre questa condizione equivale a minimizzare il cambiamento in volume degli elissoidi di confidenza associati alle due matrici di covarianza. Basato su questa idea, il criterio di Pardo e Alonso assume la forma

$$\mathbf{PAC}(R(\alpha)) = \left|\frac{\det\left(\frac{1}{n}\sum_{i=1}^{n} S_i S_i^\top\right)}{\det\left(\frac{1}{n}\sum_{i=1}^{n} V_i\right)} - 1\right|, \tag{2.38}$$

dove $S_i = [\mathbf{y}_i - \boldsymbol{\mu}_i]$ viene stimata da $\hat{S}_i$ e $\hat{V}_i$, sostituendo β con $\hat{\beta}_{gee}$ e ϕ con $\tilde{\phi}$.

Capitolo 3

Simulazione e applicazione a dati reali

3.1 Introduzione

In questo capitolo, viene testata l'efficacia dei vari criteri di selezione in diversi scenari. Il comportamento viene valutato in termini di frequenza relativa di volte in cui viene selezionata la vera struttura di correlazione con cui è stata generata la variabile risposta $\mathbf{Y}_i$. Lo studio sull'efficacia di selezione è stato applicato a risposte di tipo continuo, di conteggio e binarie. Studi simili in letteratura, circa il confronto via simulazione di alcuni dei criteri presentati, si possono trovare in Pardo & Alonso (2019), Gosho et al. (2011), Chen & Lazar (2012), Hin & Wang (2008) e Shults & Chaganty (1998). Lo studio presentato in questa sede propone invece un confronto più ampio, considerando tutti criteri presenti in letteratura. In particolare si è cercato di valutare se e come varia la frequenza di corretta selezione al variare del parametro di correlazione $\alpha \in [0.10, 0.15, 0.2, \ldots, 0.8]$ per risposte continue e di conteggio. Per evitare problemi di convergenza delle stime, nel caso di risposta binaria si è posto $\alpha \in [0.10, 0.15, 0.2, \ldots, 0.6]$. Si è inoltre investigato l'impatto che la numerosità campionaria n e il numero di osservazioni per l'i-esima unità m possono avere sull'efficacia dei criteri. A tal proposito, sono state effettuate 10000 simulazioni per ogni valore di α e per ogni possibile combinazione di n e m con $n = 50, 100, 200$ e $m = 4, 6, 12$. Un elemento di novità di questo studio è rappresentato dalle frequenze relative medie percentuali di corretta selezione della struttura di correlazione distinte per tipologia di risposta e sotto diverse configurazioni di n e m. Come matrici di correlazione $R(\alpha)$ si sono considerate le strutture di indipendenza, di equicorrelazione e autoregressiva del primo ordine. Tutte le simulazioni sono state svolte con

l'ausilio del *software* R di cui si riporta il codice integrale nell'Appendice. Il codice per il calcolo della verosimiglianza empirica è stato reso disponibile per questa relazione da Chen & Lazar (2012). Si riporta infine a titolo esemplificativo l'applicazione dei criteri a un dataset reale.

3.2 Studio di simulazione

Per risposte di tipo continuo il modello generatore scelto è

$$Y_{ij} = x_{ij1}\beta_1 + x_{ij2}\beta_2 + \epsilon_{ij}, \qquad i = 1, \dots, n \quad j = 1, \dots, m, \tag{3.1}$$

dove $\beta_1 = \beta_2 = 1$. Le covariate $x_{ij1}, \dots, x_{im1}$ sono generate da $Z \sim Bin(1, \pi = 0.5)$, mentre le variabili $x_{ij1}, \dots, x_{im1}$ vengono generate da una $N(0, 1)$. Gli errori $\epsilon_i = [\epsilon_{i1}, \dots, \epsilon_{im}]^\top$, indipendenti $i = 1, \dots, n$, sono generati da una distribuzione normale multivariata $N_m(0, R(\alpha))$. Nel modello (3.1) e in quelli a seguire si assume per semplicità che ogni soggetto abbia lo stesso numero di osservazioni, ovvero che $m_i = m$.

Per la generazione di risposte Poisson (di conteggio) correlate, si è utilizzato il modello

$$\log(\mu_{ij}) = \beta_0 + x_{ij1}\beta_1 + (j-1)\beta_2, \qquad i = 1, \dots, n, \quad j = 1, \dots, m, \tag{3.2}$$

dove $\beta_0 = 0.5$, $\beta_1 = -0.2$ e $\beta_2 = -0.2$. La variabile esplicativa x_{ij1} viene generata da una variabile casuale $Z \sim Bin(1, \pi = 0.5)$. Infine, per quanto concerne le risposte binarie correlate si è assunto un modello generatore del tipo

$$\text{logit}(\mu_{ij}) = \beta_0 + x_{ij1}\beta_1 + (j-1)\beta_2, \qquad i = 1, \dots, n, \quad j = 1, \dots, m, \tag{3.3}$$

con $\beta_0 = 0.5$, $\beta_1 = -0.2$ e $\beta_2 = -0.2$ e le covariate x_{ij1} dello stesso tipo del modello specificato alla (3.2).

Operativamente, l'i-esima risposta gaussiana correlata è stata creata attraverso la funzione `mvrnorm` appartenente alla libreria mvtnorm resa disponibile in R da Genz & Bretz (2009). Le risposte di conteggio correlate invece sono state create mediante la funzione `genPoisNor` del pacchetto PoisNor (Amatya et al. (2021)), mentre le risposte binarie correlate sono state generate tramite la funzione `SimCorrMultRes` del pacchetto omonimo creato da Touloumis (2016). Le stime dei parametri β, ϕ, α sono state ottenute tramite la funzione `gee` del pacchetto gee (Carey (2022)). Per quanto concerne il calcolo della verosimiglianza empirica si è utilizzata la libreria `emplik`, resa disponibile da Mai & Yang (2023). Nel seguito ci si limiterà a commentare i risultati principali, riportati

nelle tabelle, circa l'efficacia media percentuale di corretta selezione. Qualora si intenda approfondire il comportamento dei criteri per valori puntuali di α, n e m si rimanda ai grafici frutto della simulazione, inseriti in Appendice a corredo di quanto di seguito presentato. Ci si limita a commentare che in tutti i grafici per ciascuna combinazione di n e m, struttura di correlazione e tipologia di risposta la maggior parte dei criteri presentano un andamento crescente della frequenza stimata di corretta selezione all'aumentare del parametro di correlazione α. Inoltre, si nota come per la maggioranza dei criteri quanto più elevata è la numerosità campionaria e/o di *cluster*, tanto più velocemente la frequenza relativa di corrette selezioni converge a 1. Ad esempio, guardando il grafico A.1, il **PAC** raggiunge circa il 100% di efficacia per $m = 6$ e $\alpha = 0.6$, mentre con $m = 12$ tale livello viene raggiunto già per valori di correlazione pari a 0.20.

Per la risposta Gaussiana (si vedano i grafici A.1 e A.2), quando la matrice di correlazione $R(\alpha)$ è scambiabile o autoregressiva del primo ordine, la frequenza di selezione della corretta struttura di correlazione cresce al crescere del parametro di correlazione α per tutti i criteri. Tra i criteri meno efficaci (Tabelle 3.1 e 3.2), il criterio **SC** presenta una maggiore frequenza media di corretta selezione rispetto a Δ_1, Δ_2 e al **RJ**, nonostante rimanga sostanzialmente stabile al variare di numerosità campionaria e numerosità di *cluster*. Inoltre, nel caso in cui $R(\alpha)$ sia AR(1) i criteri **SC** e **RJ** tendono a selezionare scorrettamente una struttura di equicorrelazione rispettivamente nel 40% e nel 50% circa dei casi. Un comportamento analogo si registra per i criteri Δ_1 e Δ_2. Si vuole rimarcare il fatto che, come ci si attendeva, e il **QIC** e il **CIC** mostrano esattamente la stessa frequenza di selezione per ogni combinazione di n e m. Anche i criteri basati sugli autovalori generalizzati quali il **PT**, il **WR** e il **RMR** mostrano un trend medio equivalente, ovvero crescente all'aumentare di n e m. Per quanto concerne i criteri basati sulla verosimiglianza empirica, si nota come questi mostrino tra i più alti livelli di affidabilità nel selezionare la giusta struttura di correlazione. In particolare, i criteri **EAIC** e **EBIC** mostrano un'efficacia pressochè equivalente, mentre il **GAIC** mostra una frequenza di corretta selezione leggermente migliore ripetto all'**EAIC** per ogni configurazione di n e m. Anche nel caso in cui la "vera" struttura di correlazione sia l'indipendenza, il criterio migliore in termini di efficacia risulta essere il **GBIC**, mentre l'**EBIC** sembra comportarsi meglio dell'**EAIC**. D'altra parte i criteri **G**, **SC** e **PAC** presentano un'efficacia molto scarsa qualora la "vera" struttura sia l'indipendenza (Tabella 3.3). Nello specifico, per ogni valore di m, i criteri **G** e **PAC** selezionano la struttura di equicorrelazione o autoregressiva del primo ordine circa la metà delle volte ciascuna, mentre **SC** tende a selezionare la struttura di equicorrelazione la maggior parte delle volte (60% circa).

Tabella 3.1: Frequenza relativa media percentuale di selezione della corretta matrice di correlazione per risposta normale quando la vera struttura di correlazione è scambiabile, studio basato su 10000 simulazioni.

	$n = 50$			$n = 100$			$n = 200$		
	$m = 4$	$m = 6$	$m = 12$	$m = 4$	$m = 6$	$m = 12$	$m = 4$	$m = 6$	$m = 12$
QIC	82.564	91.139	98.384	90.395	98.384	99.774	95.434	98.809	99.982
CIC	82.564	91.139	98.384	90.395	98.384	99.774	95.434	98.809	99.982
SC	57.448	63.187	69.009	55.066	69.009	63.570	53.205	56.465	59.517
RJ	54.516	63.814	79.730	62.542	79.730	89.675	70.933	82.458	96.608
Δ_1	57.768	66.406	81.756	65.198	81.756	91.378	73.186	85.209	97.641
Δ_2	49.920	57.774	72.374	58.839	72.374	85.736	68.296	80.662	95.560
G	91.836	97.196	99.773	96.246	99.773	99.988	98.566	99.827	100
GAIC	89.939	96.708	99.818	95.507	99.818	99.992	98.442	99.835	100
GBIC	87.388	95.431	99.700	93.686	99.700	99.991	97.639	99.768	100
EAIC	89.284	96.438	99.461	95.336	99.461	99.982	98.391	99.826	99.999
EBIC	87.099	95.361	99.354	93.616	99.354	99.980	97.580	99.771	99.999
PT	81.730	90.280	97.827	89.908	97.827	99.645	95.179	98.662	99.968
WR	81.094	89.646	97.432	89.593	97.432	99.562	95.063	98.600	99.966
RMR	78.682	87.760	97.030	86.986	97.030	99.540	93.312	98.069	99.963
PAC	89.909	97.522	99.852	95.838	99.852	99.992	98.548	99.842	100

Tabella 3.2: Frequenza relativa media percentuale di selezione della corretta matrice di correlazione per risposta normale quando la vera struttura di correlazione è autoregressiva del primo ordine (AR-1), studio basato su 10000 simulazioni.

	$n = 50$			$n = 100$			$n = 200$		
	$m = 4$	$m = 6$	$m = 12$	$m = 4$	$m = 6$	$m = 12$	$m = 4$	$m = 6$	$m = 12$
QIC	82.424	88.169	91.902	89.045	92.996	95.398	93.829	96.263	97.842
CIC	82.424	88.169	91.902	89.045	92.996	95.398	93.829	96.263	97.842
SC	46.990	46.592	46.964	48.188	47.775	48.036	48.986	48.419	48.658
RJ	37.623	40.520	42.580	40.679	42.564	44.514	42.795	44.674	46.162
Δ_1	42.702	45.182	46.457	44.753	46.642	47.835	46.374	47.987	48.503
Δ_2	40.742	43.678	45.188	43.698	45.770	47.086	45.764	47.591	48.156
G	90.208	95.384	98.413	95.060	98.048	99.641	97.668	99.358	99.954
GAIC	86.594	93.366	98.050	93.094	97.142	99.581	96.987	99.185	99.960
GBIC	82.958	90.612	96.694	90.093	95.178	98.991	95.020	98.258	99.918
EAIC	86.074	93.202	97.331	92.908	97.053	99.438	96.908	99.126	99.942
EBIC	83.141	91.187	96.677	90.227	95.308	98.997	95.044	98.228	99.892
PT	82.202	87.825	91.661	88.765	92.886	95.361	93.662	96.180	97.824
WR	81.848	87.561	91.471	88.549	92.794	95.298	93.584	96.136	97.802
RMR	77.736	84.231	88.336	85.157	89.871	92.527	90.908	93.838	95.765
PAC	90.675	95.881	98.887	95.102	98.149	99.710	97.693	99.366	99.962

TABELLA 3.3: Frequenza relativa media percentuale di selezione della corretta matrice di correlazione per risposta normale quando la vera struttura di correlazione è l'indipendenza, studio basato su 10000 simulazioni al variare della numerosità campionaria $n = 50, 100, 200$.

	$m = 4$	$m = 6$	$m = 12$
QIC	17.606	17.060	16.853
CIC	17.606	17.060	16.853
SC	0.001	0.000	0.000
RJ	31.953	29.580	27.420
Δ_1	29.336	27.060	25.103
Δ_2	29.256	27.023	24.916
G	1.786	2.750	4.550
GAIC	75.450	73.463	72.256
GBIC	94.120	93.826	93.366
EAIC	70.983	67.736	61.393
EBIC	90.783	89.383	83.666
PT	17.756	17.413	17.023
WR	17.930	17.543	17.186
RMR	21.173	20.440	19.933
PAC	0.056	0.070	0.063

Per risposte Poisson ipotizzando $R(\alpha)$ scambiabile o AR(1), esattamente come nel caso normale, l'efficacia media nel selezionare la vera struttura di correlazione aumenta al crescere del parametro di correlazione α. Inoltre, per tutti i criteri si nota un'aumento medio dell'efficacia all'aumentare di n e m sia presi singolarmente che congiuntamente. L'unica eccezione riguarda il **QIC**. Con $R(\alpha)$ scambiabile si nota infatti, come l'incremento della corrispondente frequenza media di corretta selezione presenta un andamento non monotòno, quasi parabolico. Per n fissato, l'efficacia media del **QIC** cresce fino al massimo assunto con $m = 6$, per poi tornare a decrescere e stabilizzarsi attorno ai livelli iniziali. Risultato ancora più interessante è dato qualora si specifichi una struttura $R(\alpha)$ autoregressiva. In questo caso, l'efficacia media del criterio di quasi-verosimiglianza mostra un andamento decrescente all'aumentare di m, fissata la numerosità campionaria. Questo comportamento si spiega col fatto che, come anticipato nel secondo capitolo, il **QIC** dipende da un addendo che non porta informazioni riguardo la struttura di correlazione, ma introduce solo del "rumore" che ne deteriora l'efficacia. Con $R(\alpha)$ scambiabile, il **QIC** tende a selezionare scorrettamente la struttura AR(1) nel 30% dei casi, mentre con $R(\alpha)$ autoregressiva seleziona erroneamente la struttura di equicorrelazione circa il 25% delle volte. Si evidenzia inoltre come il criterio meno efficace risulta essere il **RJ** se la struttura specificata è scambiabile, mentre con $R(\alpha) = \text{AR}(1)$, il criterio meno prestazionale è **SC**, che seleziona invece una struttura scambiabile nel 40% dei

casi circa. Per ambo i criteri, la frequenza media di corretta selezione rimane pressoché costante all'aumentare di m per n fissato, mentre si apprezza un lieve aumento di efficacia all'aumentare della numerosità campionaria n fissato m. Con $R(\alpha)$ scambiabile, si ha inoltre che la frequenza media di corretta selezione di Δ_1 e Δ_2 è più elevata di quella relativa al criterio **RJ** con picchi di circa l'80 percento quando $m = 4$, mentre se $R(\alpha)$ ha struttura autoregressiva è il criterio **RJ** a presentare comportamento migliore di Δ_1, con un incremento massimo del 21.04 % che si registra con $m = 4$ e $n = 50$. Ad ogni modo, tali differenze tendono a ridursi all'aumentare di m indipendentemente dalla struttura di correlazione specificata, per scomparire completamente con $m = 12$. Il criterio che presenta generalmente la più alta efficacia è il **GAIC**. Anche l'**EAIC** presenta un alto tasso di corretta selezione, migliore seppur di poco all'**EBIC**. Alti tassi di corretta selezione sono infine quelli relativi ai criteri **PA** e **G**.

Tabella 3.4: Frequenza relativa media percentuale di selezione della corretta matrice di correlazione per risposta di conteggio quando la vera struttura di correlazione è scambiabile, studio basato su 10000 simulazioni.

	$n = 50$			$n = 100$			$n = 200$		
	$m = 4$	$m = 6$	$m = 12$	$m = 4$	$m = 6$	$m = 12$	$m = 4$	$m = 6$	$m = 12$
QIC	43.565	57.484	43.676	42.765	61.537	43.793	40.981	63.676	43.779
CIC	74.499	81.005	83.908	84.852	90.565	92.756	92.402	96.192	97.884
SC	61.698	69.402	75.926	57.682	64.013	69.401	55.167	59.629	63.786
RJ	50.732	70.114	99.024	54.513	81.036	99.812	58.569	90.564	99.991
Δ_1	88.512	96.586	99.895	94.872	98.989	99.99	97.877	99.776	100
Δ_2	92.074	97.304	99.875	96.359	99.11	99.989	98.342	99.776	100
G	90.47	96.69	99.687	95.569	98.973	99.982	98.166	99.764	100
GAIC	88.364	96.187	99.766	94.612	98.859	99.987	97.936	99.777	100
GBIC	85.335	94.564	99.607	92.448	98.097	99.98	96.776	99.663	100
EAIC	87.677	95.757	99.244	94.469	98.736	99.953	97.87	99.742	100
EBIC	85.18	94.46	99.116	92.436	98.055	99.95	96.786	99.616	100
PT	77.627	85.642	91.736	86.57	92.778	96.558	92.96	96.734	98.85
WR	77.154	85.236	91.335	86.361	92.611	96.465	92.866	96.69	98.814
RMR	40.041	44.373	65.168	44.952	48.614	71.708	49.947	53.203	78.633
PAC	88.831	97.178	99.826	95.056	99.092	99.986	97.808	99.773	100

Qualora la vera struttura di correlazione sia l'indipendenza i criteri che presentano le più alte frequenze di corretta selezione sono **GBIC** e **EBIC**, con un'efficacia che si attesta rispettivamente attorno al 93% e all' 85%. Inoltre, come si evince dalla Tabella 3.6, i criteri **G**, **PAC** e **SC** presentano una frequenza di corretta selezione molto bassa, prossima allo zero. In particolare, per ogni valore m considerato, **G** e **PAC** tendono a selezionare erroneamente le strutture di equicorrelazione o AR(1) poco meno del 50%

delle volte ciascuna. Inoltre, **SC** non riesce mai a cogliere la struttura di indipendenza, facendo selezionare circa il 60% delle volte la struttura di equicorrelazione.

TABELLA 3.5: Frequenza relativa media percentuale di selezione della corretta matrice di correlazione per risposta di conteggio quando la vera struttura di correlazione è autoregressiva del primo ordine (AR-1), studio basato su 10000 simulazioni.

	$n = 50$			$n = 100$			$n = 200$		
	$m = 4$	$m = 6$	$m = 12$	$m = 4$	$m = 6$	$m = 12$	$m = 4$	$m = 6$	$m = 12$
QIC	52.862	57.886	39.446	51.775	60.291	40.644	50.624	61.811	41.518
CIC	77.546	86.231	90.932	85.883	92.382	95.172	92.363	96.124	97.597
SC	43.902	42.951	44.071	46.176	45.17	45.889	47.656	46.756	47.176
RJ	78.919	86.794	89.814	86.575	91.692	93.716	91.77	95.094	96.393
Δ_1	78.977	86.98	90.089	87.203	92.144	94.072	92.824	95.598	97.001
Δ_2	65.201	77.729	83.566	80.056	87.948	91.031	89.66	93.751	95.72
G	85.801	93.88	98.001	93.15	97.324	99.505	96.805	99.033	99.922
GAIC	82.292	91.577	97.618	90.779	96.277	99.416	95.883	98.794	99.942
GBIC	78.239	88.578	96.076	87.446	93.934	98.686	93.496	97.524	99.864
EAIC	81.254	91.477	96.694	90.252	96.23	99.186	95.75	98.757	99.906
EBIC	78.036	89.238	95.971	87.226	94.229	98.568	93.482	97.57	99.799
PT	75.094	85.282	90.472	84.518	91.914	94.914	91.586	95.917	97.502
WR	74.22	84.696	90.056	84.010	91.614	94.769	91.332	95.809	97.438
RMR	63.269	69.798	72.688	70.405	77.119	78.727	78.923	84.72	84.722
PAC	87.251	94.789	98.691	93.045	97.565	99.632	96.332	99.074	99.946

TABELLA 3.6: Frequenza relativa media percentuale di selezione della corretta matrice di correlazione per risposta di conteggio quando la vera struttura di correlazione è l'indipendenza, studio basato su 10000 simulazioni al variare di $n = 50, 100, 200$.

	$m = 4$	$m = 6$	$m = 12$
QIC	22.280	21.113	22.926
CIC	19.846	19.796	20.250
SC	0.000	0.000	0.000
RJ	27.380	25.386	24.080
Δ_1	9.063	7.296	7.140
Δ_2	8.036	5.646	5.446
G	2.393	3.293	5.360
GAIC	75.516	73.373	72.290
GBIC	94.336	93.806	93.713
EAIC	69.530	65.220	56.633
EBIC	89.776	87.376	79.286
PT	20.033	19.960	20.003
WR	20.050	19.876	19.990
RMR	23.383	22.973	22.436
PAC	1.280	0.563	0.656

Per risposte binarie i criteri che presentano la maggiore frequenza media relativa percentuale di corretta selezione sono il Δ_2 e il **G** (si veda la Tabella 3.7). Anche i criteri Δ_1 e **PAC** presentano percentuali elevate, con oltre l'80 % di corrette selezioni già per $m = 4$ e $n = 50$. Mediamente i criteri basati sulla verosimiglianza empirica risultano equiparabili in termini di efficacia a **GAIC** e **GBIC**, anche se questi sembrano comportarsi leggermente meglio per basse numerosità e campionaria e di *cluster*. Con $R(\alpha)$ scambiabile, come nel caso di risposta Poisson, il **QIC** sembra essere meno efficace del **CIC** con una frequenza di corretta selezione mediamente inferiore del 20 %. D'altra parte, come si evince dalla Tabella 3.9, ipotizzando una struttura di indipendenza è il **QIC** a presentare un comportamento migliore del **CIC**. Come appare evidente dalla Tabella 3.7 gli unici criteri in controtendenza sono **SC** e **RMR**, che rimangono sostanzialmente stabili al variare di n e m. Eccettuati questi ultimi, tutti i criteri presentano un trend crescente all'aumentare di numerosità campionare e di *cluster*, sia prese singolarmente che congiuntamente.

TABELLA 3.7: Frequenza relativa media percentuale di selezione della corretta matrice di correlazione per risposta binaria quando la vera struttura di correlazione è scambiabile, studio basato su 10000 simulazioni.

	$n = 50$			$n = 100$			$n = 200$		
	$m = 4$	$m = 6$	$m = 12$	$m = 4$	$m = 6$	$m = 12$	$m = 4$	$m = 6$	$m = 12$
QIC	59.905	63.928	55.310	67.419	71.138	63.057	73.333	76.328	70.012
CIC	71.948	78.528	71.159	82.001	88.089	82.600	89.898	94.324	92.227
SC	44.865	48.572	45.036	41.830	44.170	43.282	38.399	39.287	41.882
RJ	41.345	66.516	96.370	43.846	77.200	98.599	46.390	86.986	99.589
Δ_1	80.491	92.146	98.992	89.020	96.850	99.815	94.856	98.894	99.986
Δ_2	84.405	93.266	98.899	91.194	97.154	99.800	95.637	98.982	99.984
G	81.157	91.250	97.609	89.219	96.460	99.576	94.610	98.774	99.967
GAIC	73.680	88.118	97.341	85.299	95.220	99.543	93.067	98.536	99.970
GBIC	67.210	83.981	95.890	79.826	92.113	99.076	89.202	97.117	99.950
EAIC	73.789	87.930	95.759	85.720	95.164	99.359	93.182	98.565	99.960
EBIC	67.554	84.110	94.807	80.198	92.171	98.882	89.332	97.141	99.933
PT	70.922	77.289	78.615	80.701	87.148	88.823	89.068	93.968	95.324
WR	70.633	76.838	78.351	80.454	86.941	88.680	88.940	93.897	95.306
RMR	42.601	42.033	42.787	46.860	45.207	42.580	51.417	47.963	41.740
PAC	80.793	91.623	98.118	89.098	96.540	99.562	94.647	98.810	99.958

Con la struttura di indipendenza, i criteri con le più alte proporzioni di corrette selezioni sono il **GAIC** e l'**EBIC** (Tabella 3.9). I criteri meno efficaci risultano **G** e **PAC**, che selezionano scorrettamente la struttura scambiabile o autoregressiva, con circa il 50% di selezioni ciascuna. Contrariamente a quanto visto per risposte normali o di conteggio, il criterio **SC** riesce a selezionare correttamente la struttura di indipendenza circa una volta su 4, tendendo ciononostante a scegliere con maggiore frequenza la struttura autoregressiva (circa il 40% delle volte).

TABELLA 3.8: Frequenza relativa media percentuale di selezione della corretta matrice di correlazione per risposta binaria quando la vera struttura di correlazione è autoregressiva del primo ordine (AR-1), studio basato su 10000 simulazioni.

	$n = 50$			$n = 100$			$n = 200$		
	$m = 4$	$m = 6$	$m = 12$	$m = 4$	$m = 6$	$m = 12$	$m = 4$	$m = 6$	$m = 12$
QIC	51.073	57.935	64.212	55.744	63.987	70.798	60.017	69.290	72.052
CIC	65.831	75.623	79.617	74.653	84.821	87.447	83.510	91.831	92.830
SC	37.804	42.402	44.234	39.002	43.490	45.255	39.739	44.285	45.789
RJ	70.658	78.019	79.602	77.795	84.981	86.240	84.892	90.517	91.174
Δ_1	66.508	74.277	77.630	74.860	82.799	85.402	83.123	89.917	90.983
Δ_2	56.380	64.844	68.350	68.261	77.424	80.428	78.756	87.193	88.533
G	71.481	83.270	91.801	80.311	91.124	96.662	88.315	96.040	98.861
GAIC	60.080	75.661	88.623	73.388	86.810	94.923	84.641	94.050	98.283
GBIC	52.390	68.999	83.694	66.172	80.936	91.004	78.746	89.757	95.906
EAIC	59.546	75.304	85.711	72.568	86.438	94.210	84.085	93.958	98.100
EBIC	52.248	69.600	83.177	65.521	80.917	90.908	78.283	89.670	95.782
PT	62.043	71.036	76.423	71.130	81.410	85.649	80.460	90.030	92.108
WR	61.191	70.168	75.616	70.447	80.750	85.167	79.887	89.653	91.853
RMR	62.076	67.340	66.052	70.320	74.800	72.480	78.493	82.304	79.417
PAC	71.310	83.387	92.634	79.913	90.882	96.676	87.900	95.829	98.780

TABELLA 3.9: Frequenza relativa media percentuale di selezione della corretta matrice di correlazione per risposta binaria quando la vera struttura di correlazione è l'indipendenza, studio basato su 10000 simulazioni al variare di $n = 50, 100, 200$.

	$m = 4$	$m = 6$	$m = 12$
QIC	22.543	22.390	23.346
CIC	16.333	17.160	19.953
SC	28.866	25.690	24.383
RJ	26.313	24.520	23.640
Δ_1	8.503	6.870	6.566
Δ_2	5.916	4.450	4.340
G	0.576	1.730	4.120
GAIC	74.923	73.496	71.793
GBIC	94.060	93.400	93.203
EAIC	73.526	69.946	59.206
EBIC	92.993	90.643	81.636
PT	16.213	17.513	20.060
WR	16.226	17.483	20.093
RMR	22.790	21.583	22.276
PAC	0.743	0.646	1.123

3.3 Applicazione a dati reali

In questo paragrafo si riportano, a titolo esemplificativo, i risultati dei criteri presentati applicati a un dataset reale. I dati scelti, tratti da Thall & Vail (1990), riguardano un *trial* clinico randomizzato effettuato su 59 pazienti affetti da epilessia. Inizialmente è stato rilevato il numero totale di crisi epilettiche registrate dall'i-esimo paziente in un periodo di 8 settimane, $i = 1, \ldots, 59$. Successivamente, il numero di eventi è stato misurato per quattro periodi successivi a intervalli di due settimane. Durante lo studio i pazienti sono stati assegnati in modo casuale a due gruppi, casi e controlli. Il gruppo dei controlli è stato trattato con un placebo, mentre i casi sono stati trattati con progabide. Il disegno sperimentale è bilanciato. L'intento è determinare se l'utilizzo di questo principio attivo fa diminuire significativamente il numero di crisi epilettiche. Le variabili rilevate sono:

- `base`: variabile quantitativa discreta che rappresenta il numero complessivo di crisi epilettiche registrate dopo 8 settimane;
- `trt`: variabile dicotomica, con valore 0 se il paziente afferisce al gruppo dei controlli e 1 altrimenti;
- `age`: variabile quantitativa discreta che rappresenta l'età in anni compiuti del paziente;
- `V4`: variabile dicotomica che rappresenta l'indicatrice per il quarto periodo, con valore 1 se `period` vale 4 e 0 altrimenti;
- `lbase`: variabile quantitativa continua che descrive il conteggio complessivo di crisi epilettiche in scala logaritmica e centrata;
- `lage`: variabile quantitativa continua che rappresenta l'età del paziente i-esimo in scala logaritmica e centrata.

Vengono infine rilevate le variabili `subject` e `period`, che raffigurano rispettivamente l'indicatore del soggetto e del periodo. La variabile risposta è y, ovvero il numero di crisi epilettiche del soggetto i-esimo registrate nel periodo k-esimo, $i = 1, \ldots, 59$ e $k = 1, \ldots, 4$. Nel seguito si assumerà un livello di significatività α pari al 5%.

Data la natura della risposta, viene specificato un modello marginale il quale prevede che la variabile risposta y_{ik} sia realizzazione della variabile casuale Y_{ik} $i = 1, \ldots, 59$ e $k = 1, \ldots, 4$. Il modello, basato sulle ipotesi (1.37)–(1.39), assume la forma

$$\log(y_{ik}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 k, \tag{3.4}$$

dove x_{i1} indica la variabile `lbase`, x_{i2} indica la variabile `trt`, x_{i3} indica la variabile `lage`, x_{i4} indica la variabile `V4`, mentre k indica la variabile `period`, con $k = 1, \ldots, 4$. Il modello (3.4) viene adattato assumendo tre diverse specificazioni per la *working correlation matrix*, rispettivamente indipendenza, scambiabile e autoregressiva del primo ordine. Si è ritenuto opportuno eliminare dal dataset il 49-esimo paziente in quanto presentava un numero di crisi pre e post-trattamento anomalo. Le variabili `V4` e `period` sono state eliminate in tutti e tre i modelli perché ampiamente non significative. Nei modelli con struttura scambiabile o AR(1), viene considerata la variabile `trt` seppur non significativa, perchè ritenuta importante per lo studio. Il modello stimato risulta dunque

$$\log(y_{ik}) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3}, \tag{3.5}$$

di cui si riportano le stime nella seguente tabella dividendole per tipologia di struttura di correlazione $R(\alpha)$ assunta.

Tabella 3.10: Risultati di stima dei parametri di regressione del modello, distinti per tipologia di struttura di correlazione specificata.

	stime	*st. error*	*z-value*	*p-value*
Indipendenza				
$\hat{\beta}_0$	1.852	0.081	22.620	0.0000
$\hat{\beta}_1$	0.986	0.076	12.840	0.0000
$\hat{\beta}_2$	-0.231	0.107	-2.140	0.0324
$\hat{\beta}_3$	0.701	0.233	3.00	0.0027
Equicorrelazione				
$\hat{\beta}_0$	1.852	0.115	16.040	0.0000
$\hat{\beta}_1$	0.986	0.108	9.110	0.0000
$\hat{\beta}_2$	-0.231	0.152	-1.520	0.1290
$\hat{\beta}_3$	0.701	0.330	2.130	0.0330
Autoregressiva - AR(1)				
$\hat{\beta}_0$	1.850	0.113	16.42	0.0000
$\hat{\beta}_1$	0.986	0.106	9.330	0.0000
$\hat{\beta}_2$	-0.256	0.149	-1.71	0.0873
$\hat{\beta}_3$	0.770	0.323	2.380	0.0173

Inoltre i parametri di correlazione ρ stimati con struttura scambiabile e autoregressiva

sono nell'ordine $\hat{\rho} = 0.333$ e $\hat{\rho} = 0.446$. Come si evince dalla Tabella 3.11, la struttura autoregressiva del primo ordine è stata scelta da 9 criteri su 15, segue la struttura di equicorrelazione suggerita da 6 criteri, mentre nessun criterio seleziona una struttura di tipo indipendenza. Nello specifico, tutti i criteri basati su autovettori generalizzati, **PAC** e **G**, oltre che i criteri basati sulla verosimiglianza empirica (**EAIC** e **EBIC**) selezionano la struttura AR(1). Così anche il criterio di quasi-verosimiglianza, il **CIC**, il Δ_1 e il Δ_2. D'altra parte, i criteri basati sulla pseudo-verosimiglianza Gaussiana, il **GAIC** e il **GBIC** suggeriscono di utilizzare la struttura di equicorrelazione così come il criterio **RJ** e **SC**. Si consiglia di selezionare la struttura scambiabile suggerita dal **GBIC**, in quanto questo criterio presenta in generale la più alta efficacia di selezione per risposte di conteggio, indipendentemente dalla matrice di correlazione.

TABELLA 3.11: Calcolo dei criteri di selezione al variare della struttura di correlazione scelta per il modello. I valori minimi vengono sottolineati, indicando la corrispondente struttura selezionata dal criterio.

	Indipendenza	Equicorrelazione	AR(1)
QIC	-939.118	<u>-939.119</u>	-928.124
CIC	5.513	5.513	<u>4.826</u>
SC	228.012	<u>228.001</u>	240.000
RJ	2.769	<u>0.110</u>	0.242
Δ_1	0.946	0.072	<u>0.070</u>
Δ_2	1.246	0.342	<u>0.321</u>
G	7.100	<u>5.890</u>	7.160
GAIC	1392	<u>1364</u>	1366
GBIC	1406	<u>1381</u>	1383
EAIC	40.100	17.000	<u>11.600</u>
EBIC	53.900	34.300	<u>28.800</u>
PT	1.182	0.984	<u>0.9635</u>
WR	0.060	0.034	<u>0.032</u>
RMR	0.429	0.3654	<u>0.348</u>
PAC	0.526	0.210	<u>0.099</u>

Alla luce dei risultati di simulazione è possibile anche dare una misura dell'affidabilità del criterio proposto. Guardando il grafico A.8 che meglio rappresenta per numerosità campionaria e di *cluster* il dataset analizzato, è possibile ritenere che, fissando il parametro di correlazione ρ al valore stimato dal modello con struttura di equicorrelazione, $\rho = 0.33$, la conclusione a cui si è pervenuti sia affidabile al 92 % circa. Altri criteri

che mostrano un'alta affidabilità circa la selezione di $R(\alpha)$ sono **GAIC** e Δ_1. Guardando quest'ultimo si sarebbe tuttavia scelta la struttura autoregressiva. In definitiva, l'effetto del trattamento con progabide sul numero di crisi epilettiche non sembra essere significativo, sia scegliendo una struttura scambiabile che autoregressiva del primo ordine.

Conclusioni

In questo lavoro sono stati presentati i criteri di selezione per la scelta della matrice di correlazione di lavoro da definire in fase di specificazione del modello basato su GEE. Se ne è valutata l'efficacia via simulazione, confrontando il corrispondente comportamento sotto diverse configurazioni di numerosità campionaria e di *cluster*. I risultati ottenuti mostrano come tutti i criteri dipendano sostanzialmente dalla "vera" struttura di correlazione, dalla tipologia della risposta, dal valore di correlazione e infine dal numero di unità statistiche n e di osservazioni per unità m. In generale, si può asserire che quanto più alta è la correlazione tra osservazioni, fissato il soggetto, tanto meglio i criteri riusciranno a selezionare la "vera" struttura di correlazione sottostante. Si è inoltre riscontrata una maggiore frequenza di corretta selezione per numerosità campionarie elevate, fissato m, e per elevate numerosità di *cluster* fissato il numero di soggetti. L'incremento in efficacia è ancora più apprezzabile al crescere di n e m congiuntamente. Complessivamente, gli esiti di simulazione riportano come siano i criteri **GAIC**, **GBIC**, **EAIC**, **EBIC** a mostrare i risultati migliori, indipendentemente dalla "vera" struttura di correlazione. Nello specifico, la versione di **AIC** e **BIC** basata sulla pseudo-verosimiglianza Gaussiana sembrano mostrare una maggiore frequenza di corrette selezioni rispetto ai criteri di verosimiglianza empirica. Qualora $R(\alpha)$ sia scambiabile o AR(1) i criteri **GAIC** e **GBIC** sono mediamente equivalenti in termini di efficacia, così come lo sono **EAIC** e **EBIC**. L'unica sostanziale differenza si registra qualora la "vera" struttura sia l'indipendenza, nel qual caso il **GBIC** presenta in assoluto il comportamento migliore. Data l'alta affidabilità di quest'ultimo per tutte le strutture considerate, si consiglia di basare su di esso la scelta della *working correlation matrix*. Se $R(\alpha)$ ha una struttura di equicorrelazione o AR(1), altri criteri che presentano in media un buon andamento sono il **PAC** e il **G**, oltre che il **PT** e **WR**. Infine, i rimanenti criteri presentano frequenze di corretta selezione molto variabili, dipendendo in particolar modo dalla tipologia della risposta e dalla struttura di correlazione sottostante. In generale, data la scarsa affidabilità, ne viene sconsigliato l'utilizzo. In definitiva i criteri **GAIC**, **GBIC**, **EAIC**, **EBIC** rappresentano un'ottimo strumento su cui poter basare

la scelta della matrice di correlazione.

Appendice

FIGURA A.1: Andamento della frequenza di corretta selezione dei diversi criteri di selezione al variare di α, per $n = 50$ e $m = 4$, 6, 12 rispettivamente quando la risposta è Gaussiana e la vera struttura di correlazione è scambiabile.

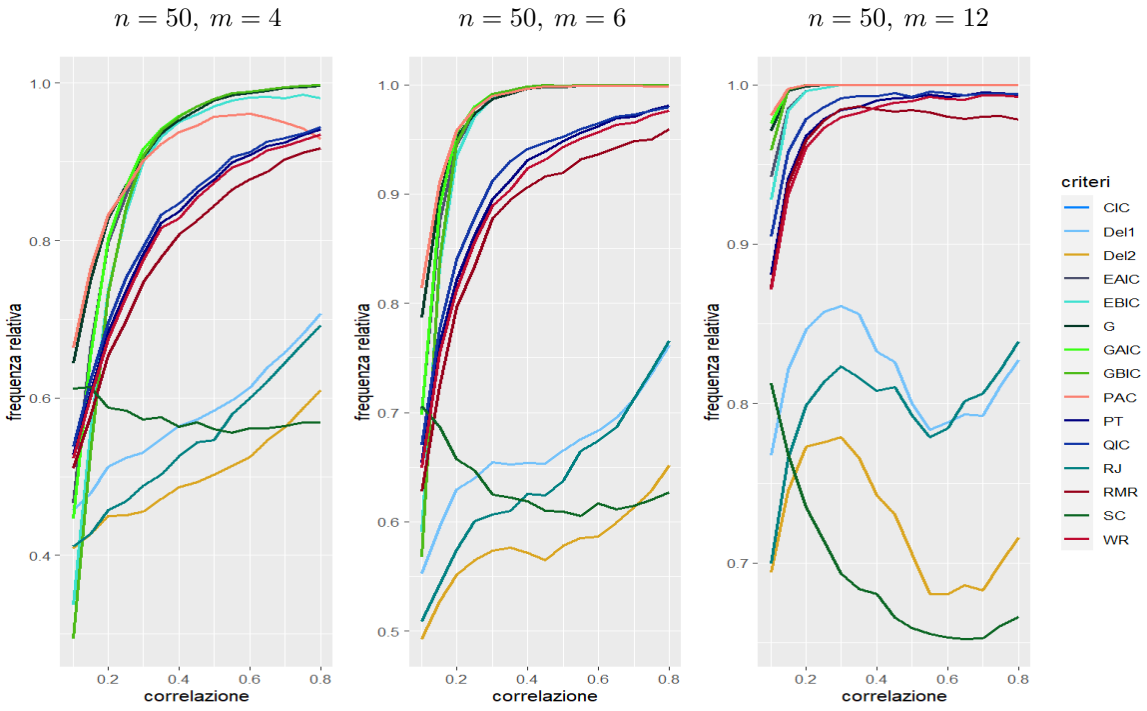


FIGURA A.2: Andamento della frequenza di corretta selezione dei diversi criteri di selezione al variare di α, per $n = 100$ e $m = 4$, 6, 12 rispettivamente quando la risposta è Gaussiana e la vera struttura di correlazione è scambiabile.

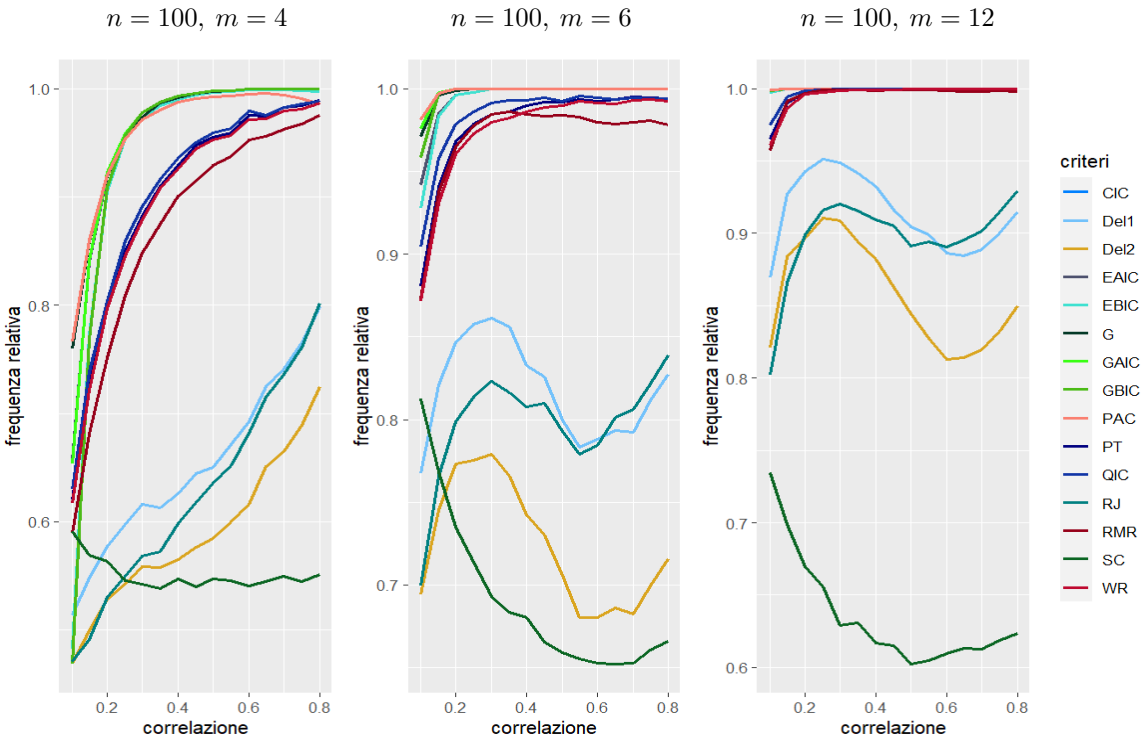


FIGURA A.3: Andamento della frequenza di corretta selezione dei diversi criteri di selezione al variare di α, per $n = 200$ e $m = 4$, 6, 12 rispettivamente quando la risposta è Gaussiana e la vera struttura di correlazione è scambiabile.

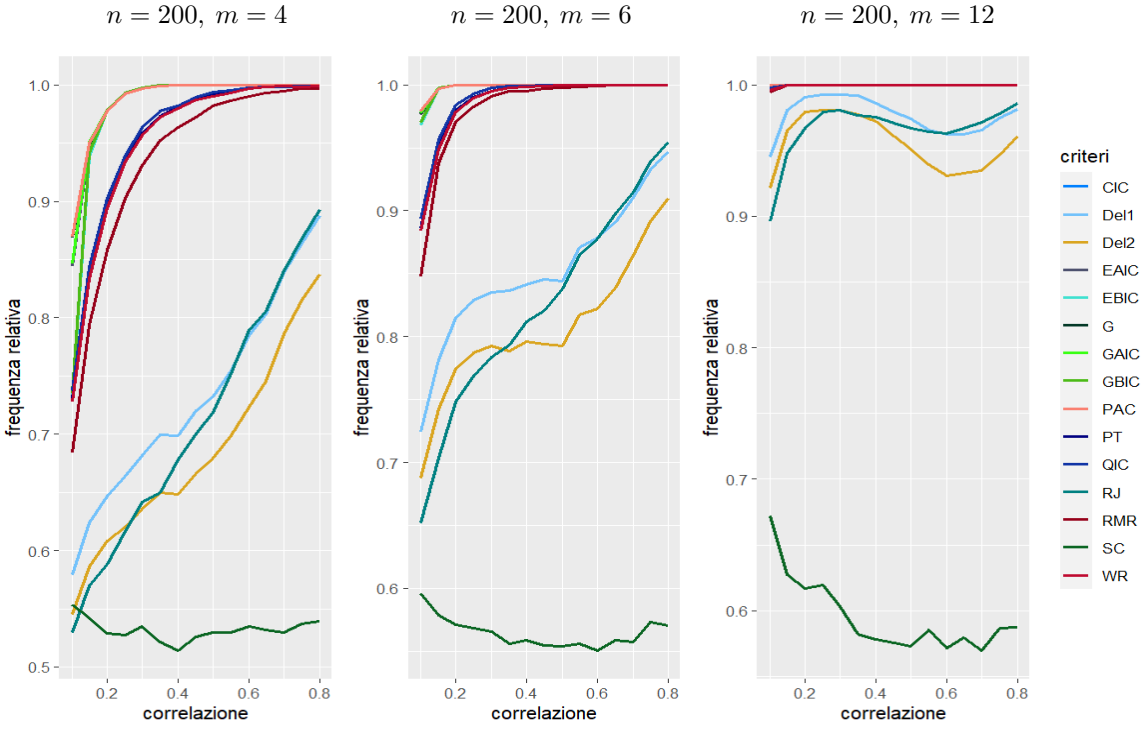


FIGURA A.4: Andamento della frequenza di corretta selezione dei diversi criteri di selezione al variare di α, per $n = 50$ e $m = 4$, 6, 12 rispettivamente quando la risposta è Gaussiana e la vera struttura di correlazione è AR(1).

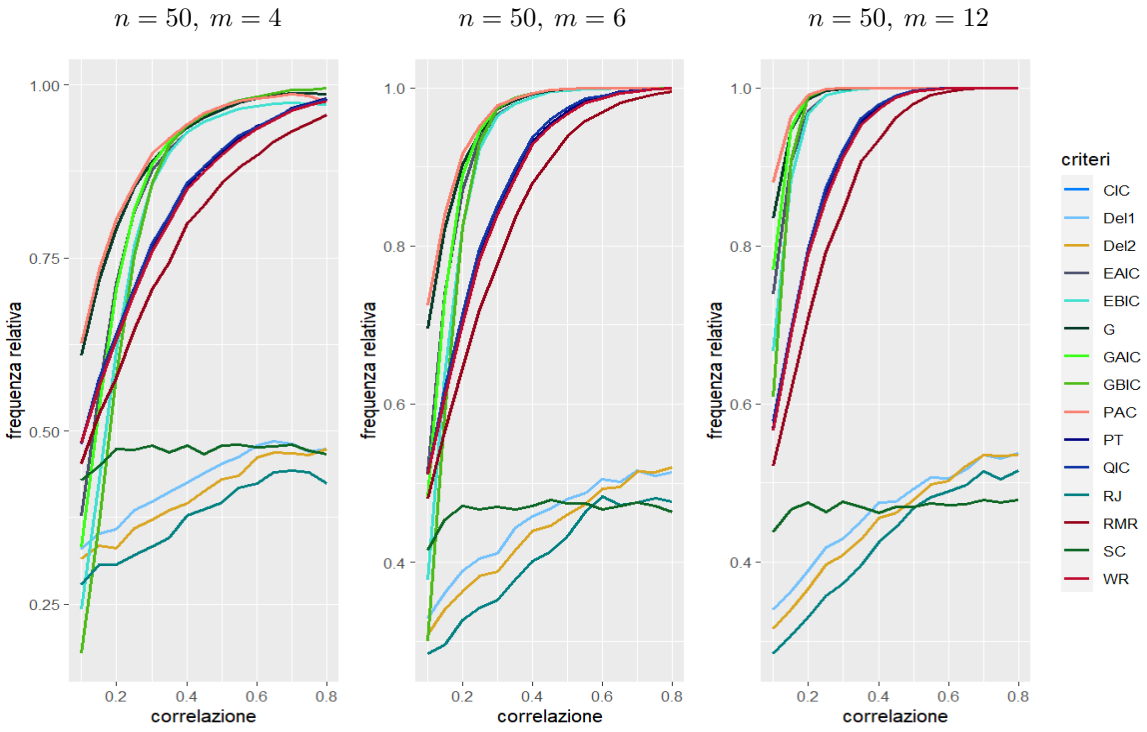


FIGURA A.5: Andamento della frequenza di corretta selezione dei diversi criteri di selezione al variare di α, per $n = 100$ e $m = 4$, 6, 12 rispettivamente quando la risposta è Gaussiana e la vera struttura di correlazione è AR(1).

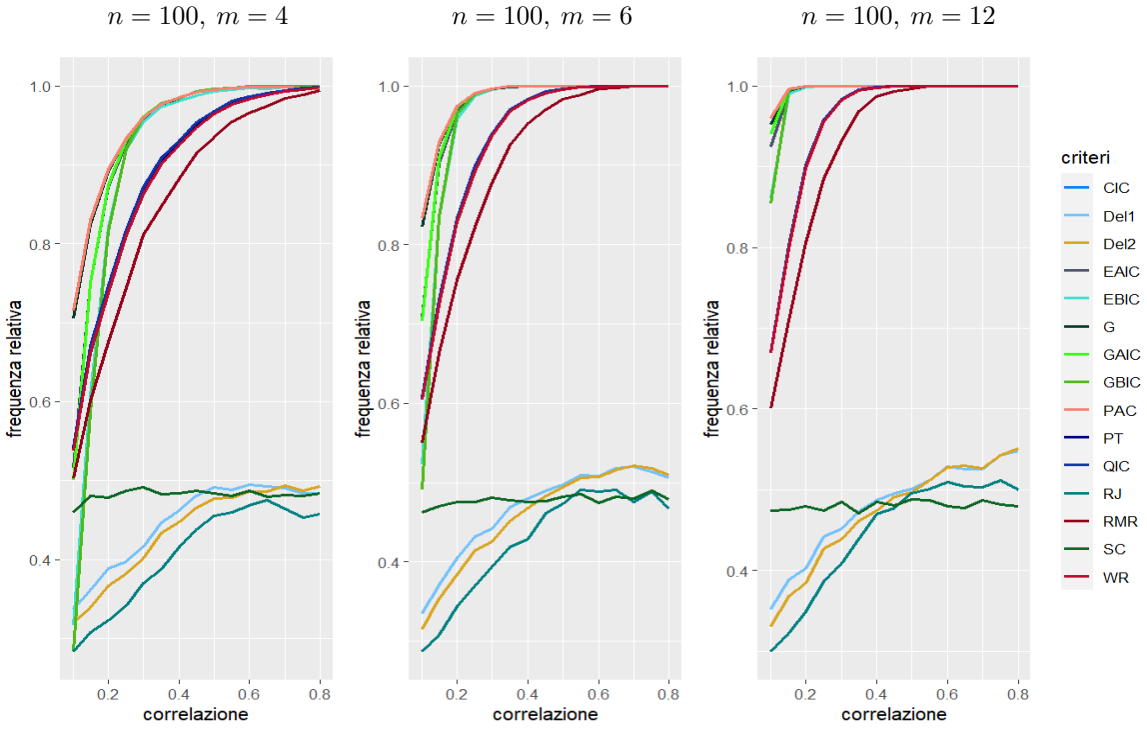


FIGURA A.6: Andamento della frequenza di corretta selezione dei diversi criteri di selezione al variare di α, per $n = 200$ e $m = 4$, 6, 12 rispettivamente quando la risposta è Gaussiana e la vera struttura di correlazione è AR(1).

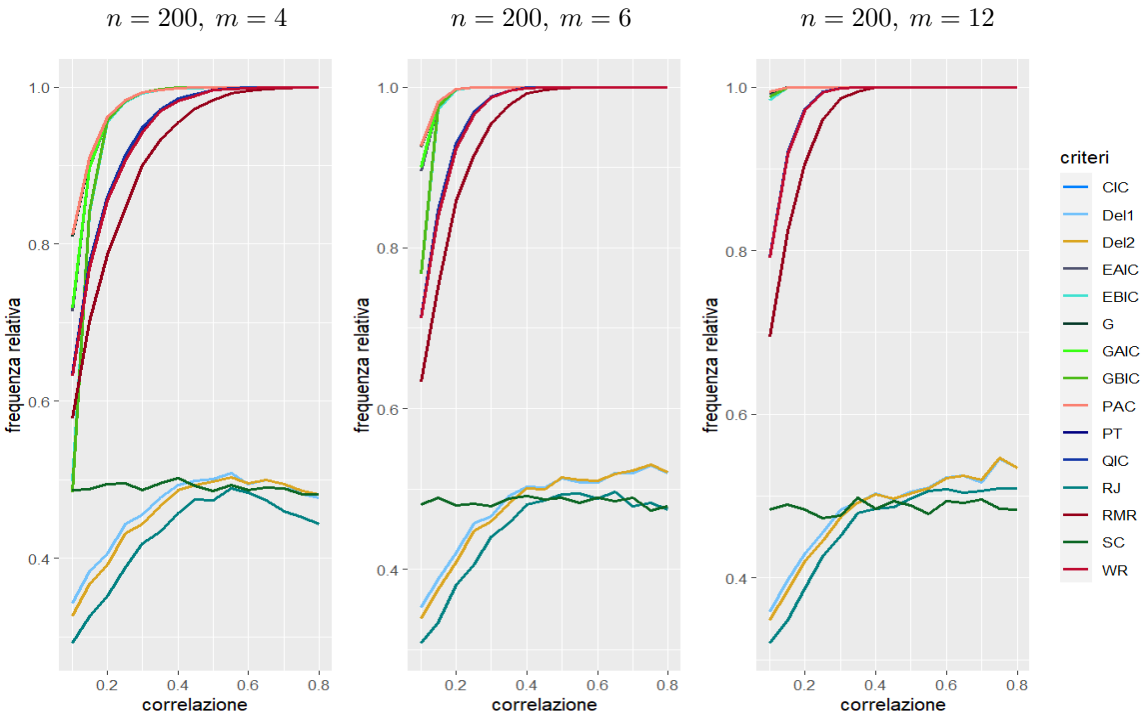


FIGURA A.7: Andamento della frequenza di corretta selezione dei diversi criteri di selezione al variare della numerosità campionaria n, per m fissato rispettivamente 4, 6 e 12, quando la risposta è Gaussiana e la vera struttura di correlazione è l'indipendenza.

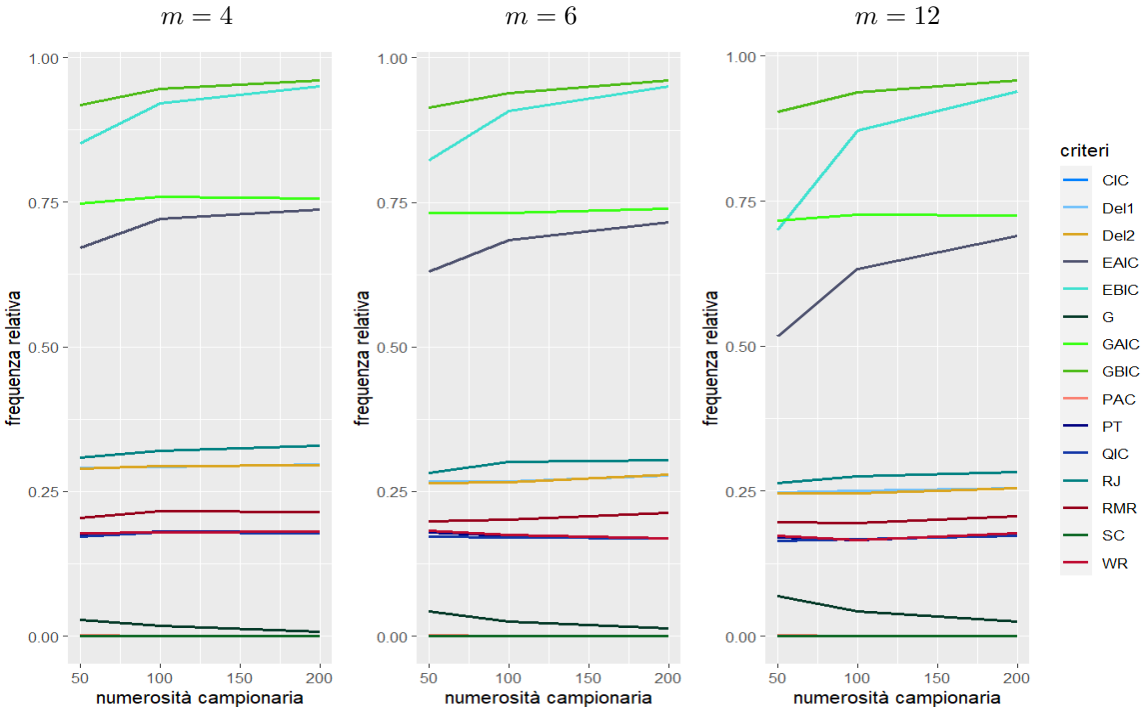


FIGURA A.8: Andamento della frequenza di corretta selezione dei diversi criteri di selezione al variare di α per $n = 50$ e $m = 4$, 6, 12 rispettivamente quando la risposta è Poisson e la vera struttura di correlazione è scambiabile.

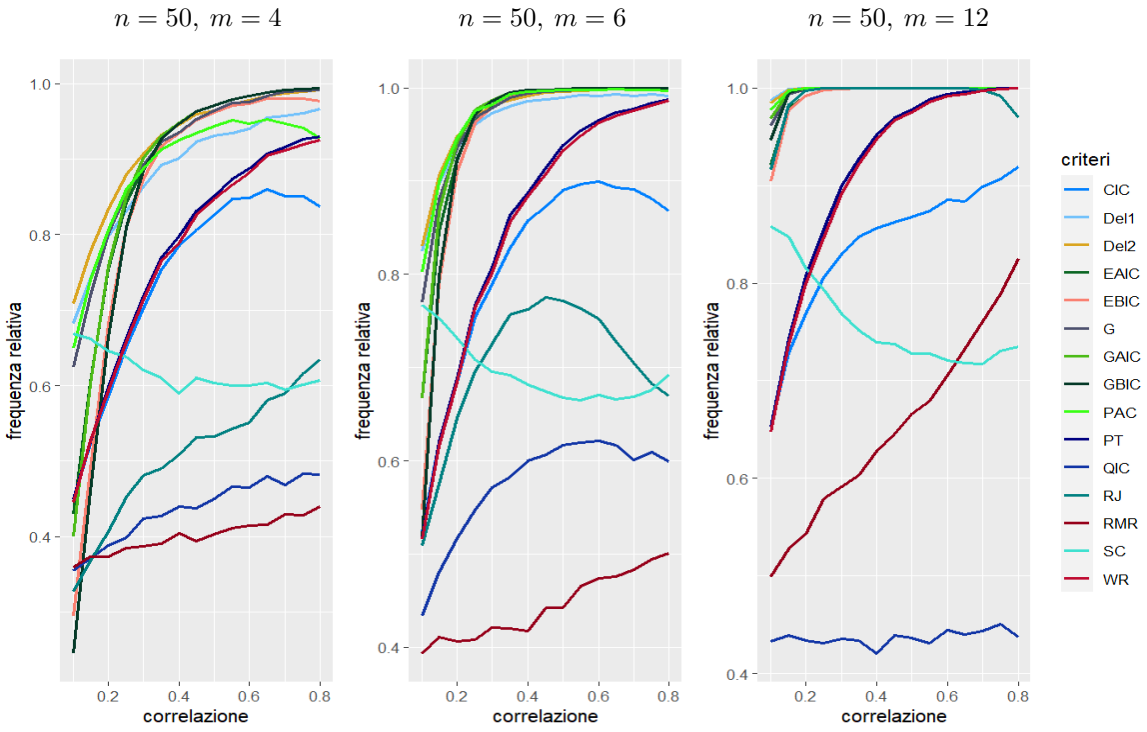


FIGURA A.9: Andamento della frequenza di corretta selezione dei diversi criteri di selezione al variare di α, per $n = 100$ e $m = 4$, 6, 12 rispettivamente quando la risposta è Poisson e la vera struttura di correlazione è scambiabile.

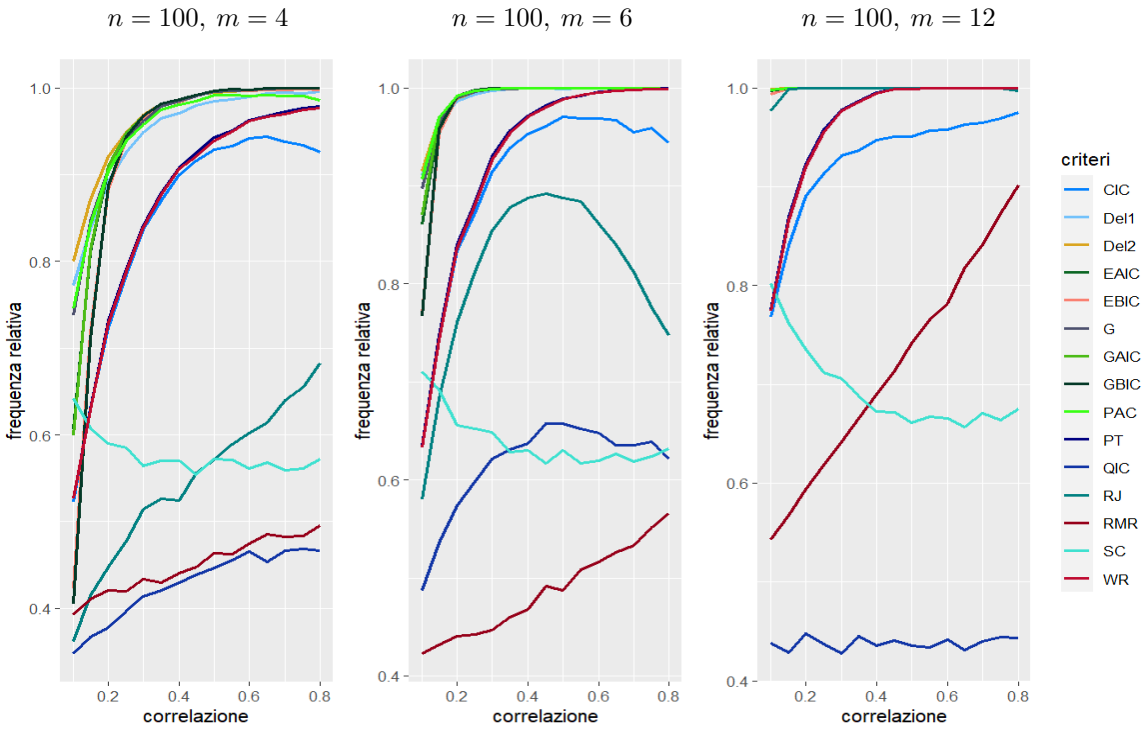


FIGURA A.10: Andamento della frequenza di corretta selezione dei diversi criteri di selezione al variare di α, per $n = 200$ e $m = 4$, 6, 12 rispettivamente quando la risposta è Poisson e la vera struttura di correlazione è scambiabile.

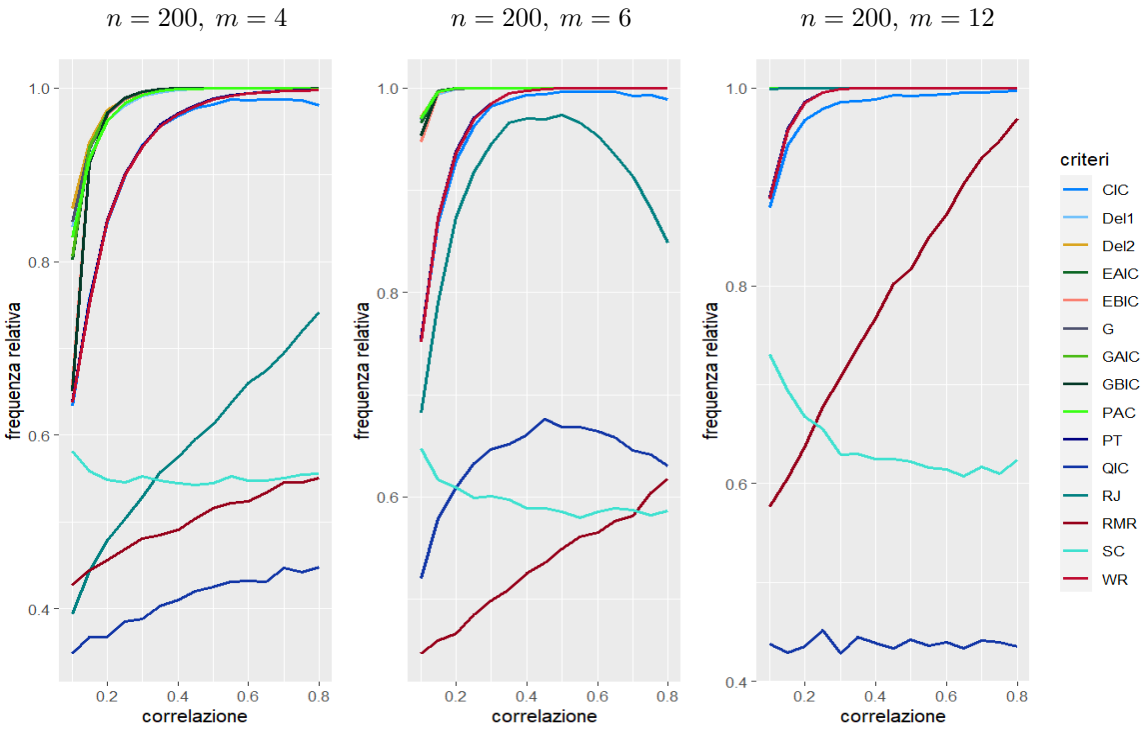


FIGURA A.11: Andamento della frequenza di corretta selezione dei diversi criteri di selezione al variare di α, per $n = 50$ e $m = 4$, 6, 12 rispettivamente quando la risposta è Poisson e la vera struttura di correlazione è AR(1).

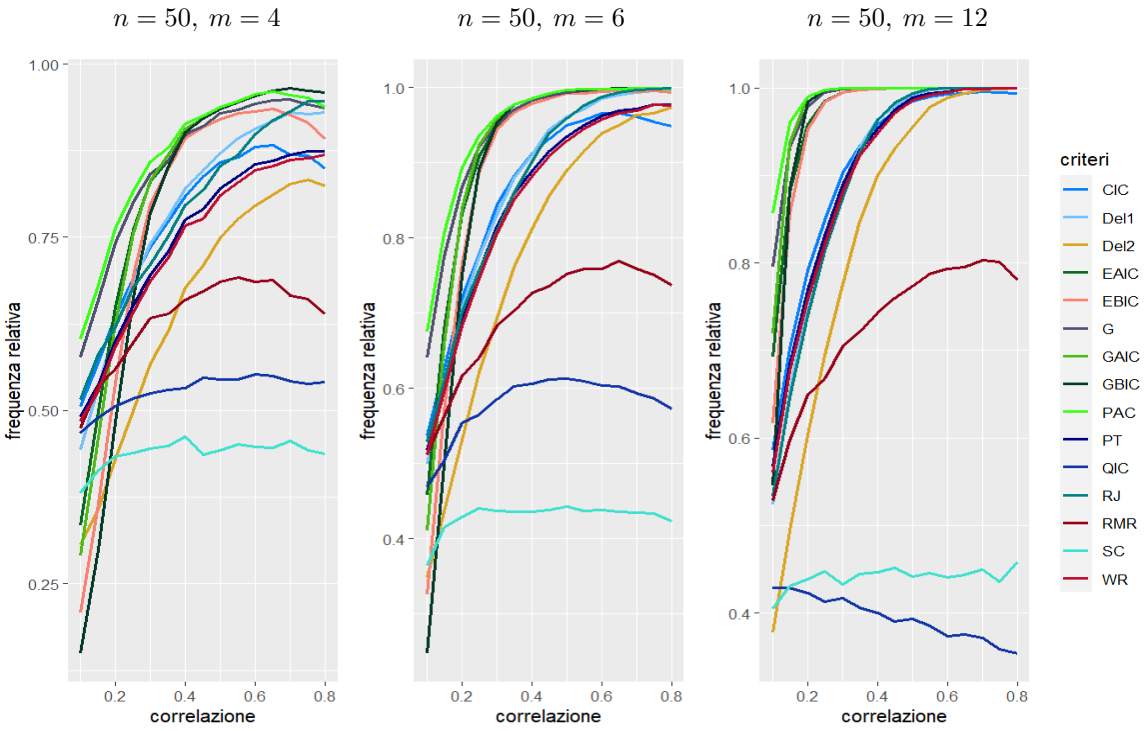


FIGURA A.12: Andamento della frequenza di corretta selezione dei diversi criteri di selezione al variare di α, per $n = 100$ e $m = 4$, 6, 12 rispettivamente quando la risposta è Poisson e la vera struttura di correlazione è AR(1).

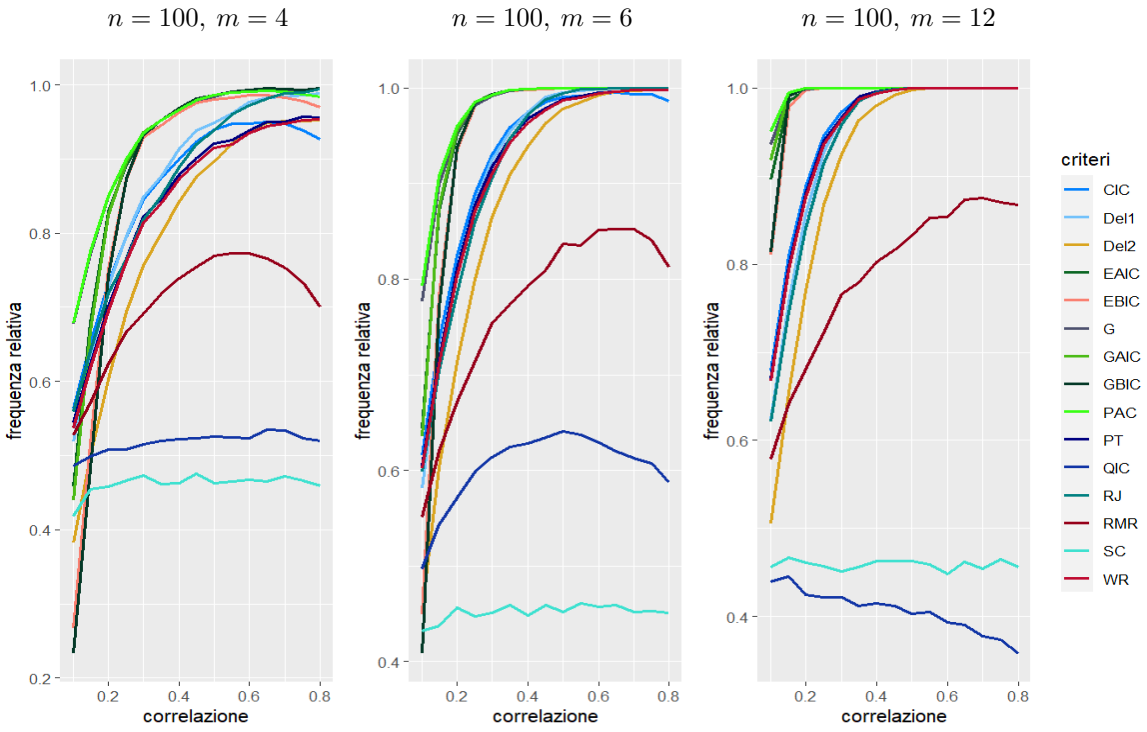


FIGURA A.13: Andamento della frequenza di corretta selezione dei diversi criteri di selezione al variare di α, per $n = 200$ e $m = 4$, 6, 12 rispettivamente quando la risposta è Poisson e la vera struttura di correlazione è AR(1).

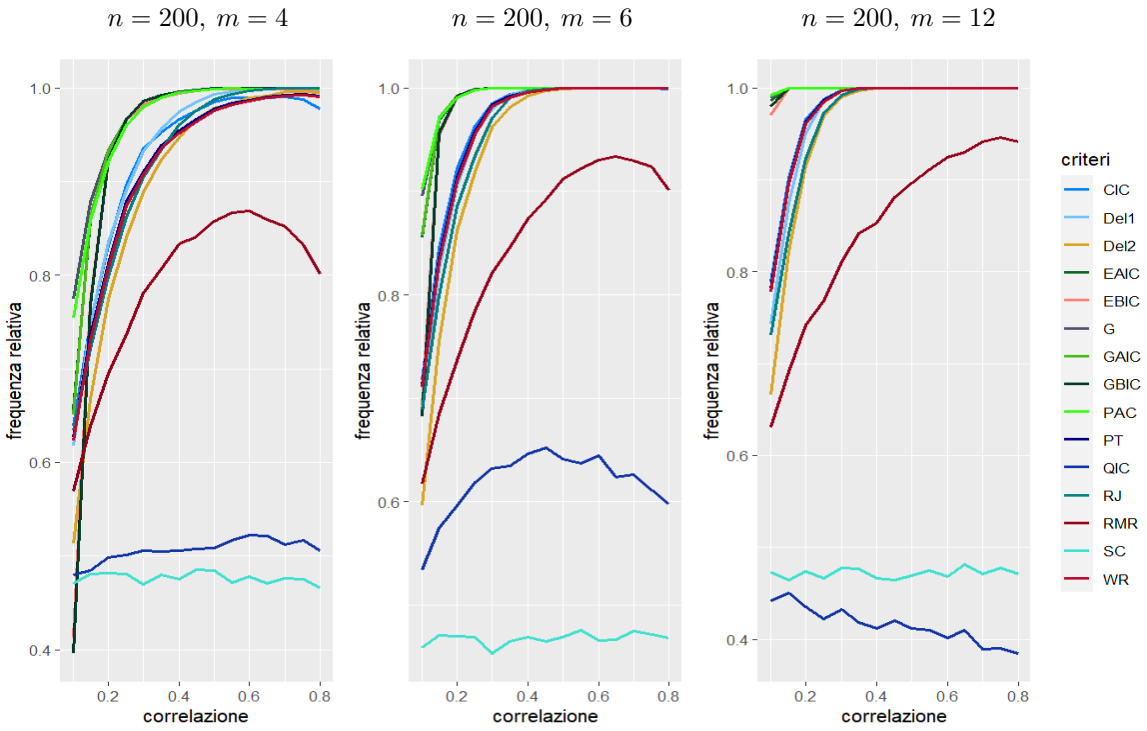


FIGURA A.14: Andamento della frequenza di corretta selezione dei diversi criteri di selezione al variare della numerosità campionaria n, per m fissato rispettivamente a 4, 6 e 12, quando la risposta è Poisson e la vera struttura di correlazione è l'indipendenza.

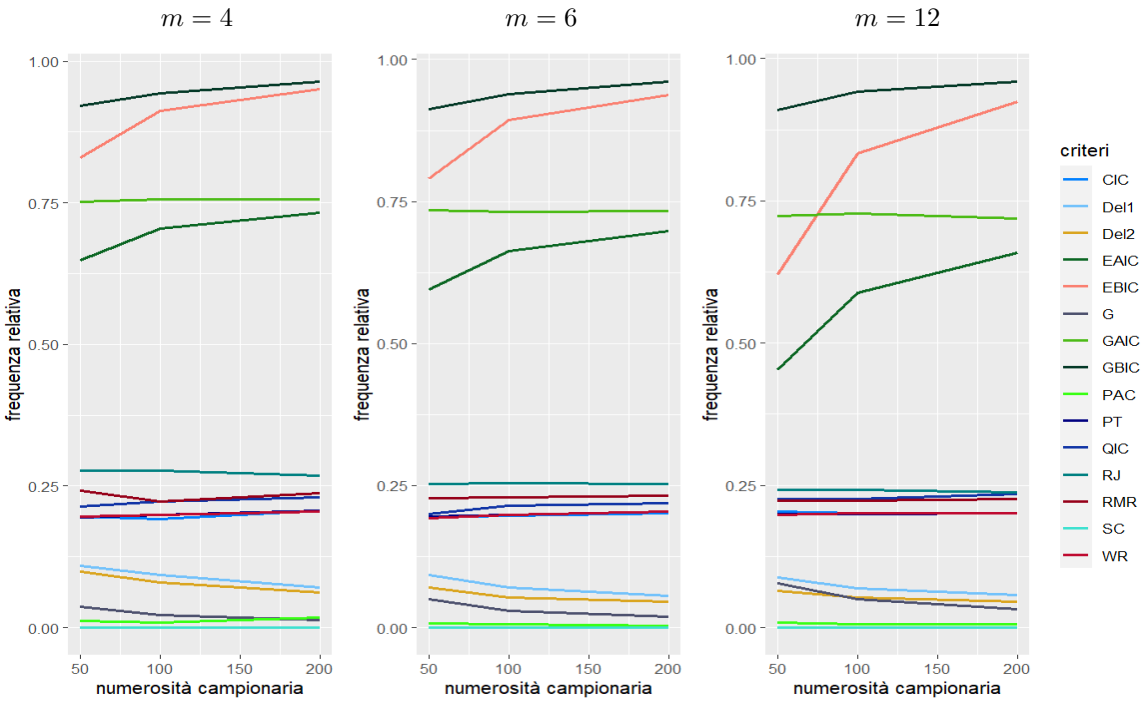


FIGURA A.15: Andamento della frequenza di corretta selezione dei diversi criteri di selezione al variare di α, per $n = 50$ e $m = 4$, 6, 12 rispettivamente quando la risposta è binaria e la vera struttura di correlazione è scambiabile.

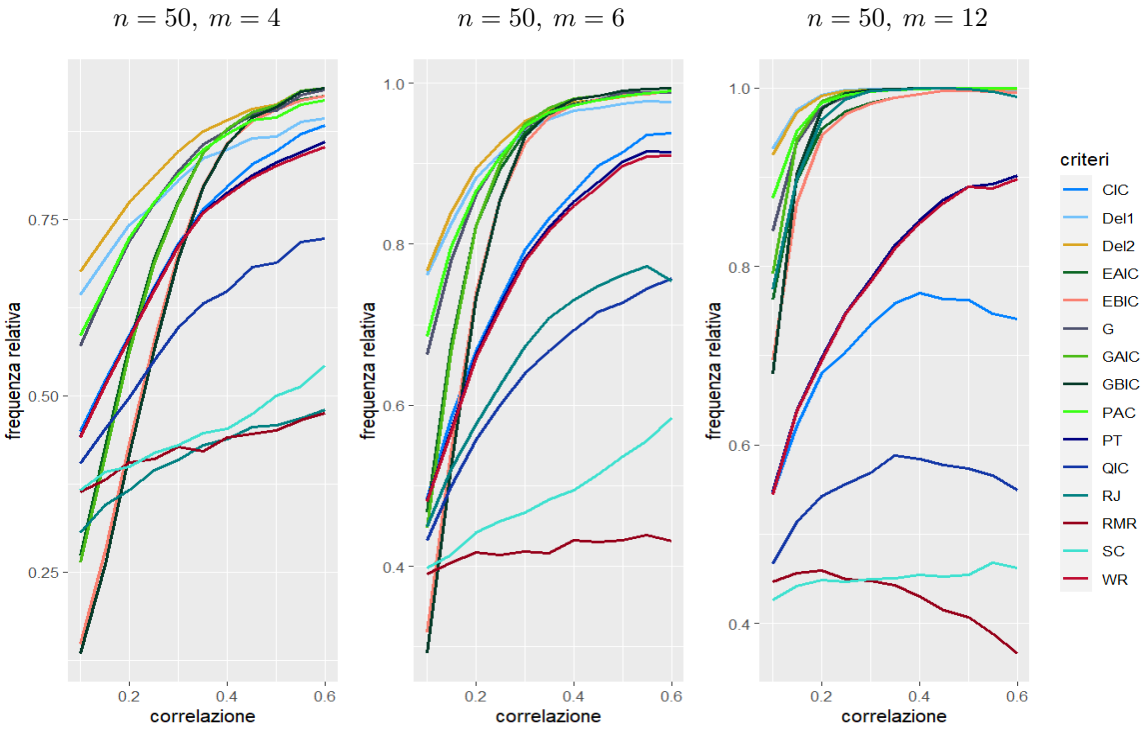


FIGURA A.16: Andamento della frequenza di corretta selezione dei diversi criteri di selezione al variare di α, per $n = 100$ e $m = 4$, 6, 12 rispettivamente quando la risposta è binaria e la vera struttura di correlazione è scambiabile.

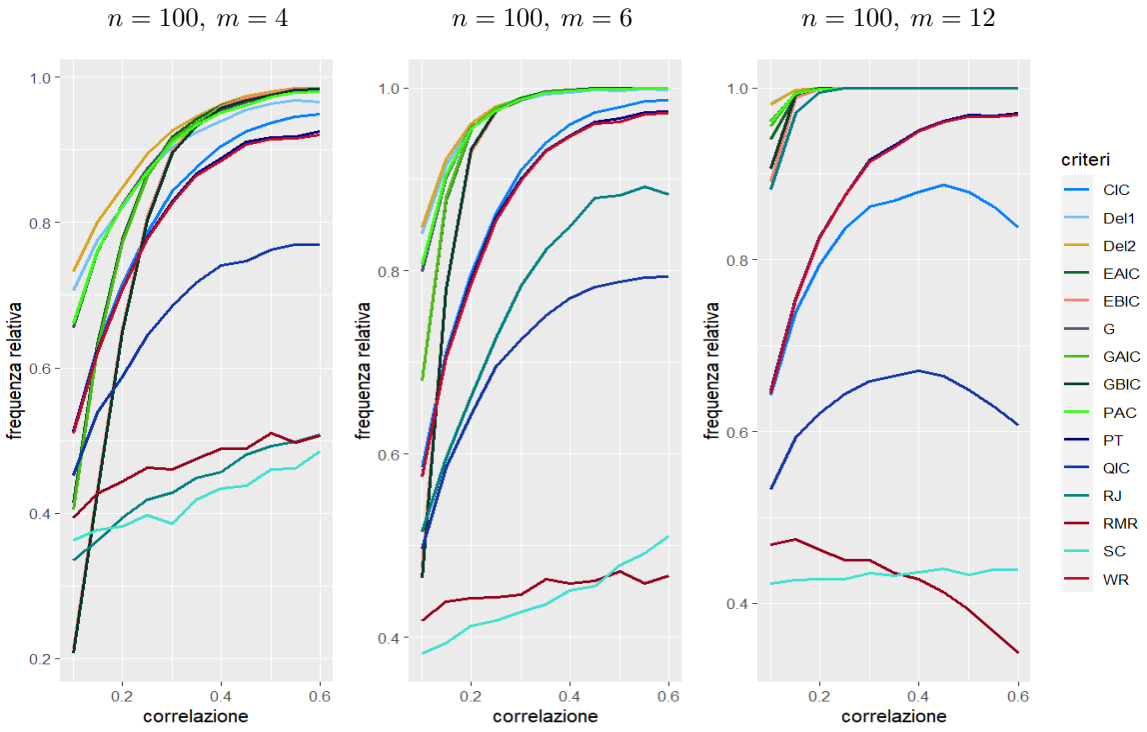


FIGURA A.17: Andamento della frequenza di corretta selezione dei diversi criteri di selezione al variare di α, per $n = 200$ e $m = 4$, 6, 12 rispettivamente quando la risposta è binaria e la vera struttura di correlazione è scambiabile.

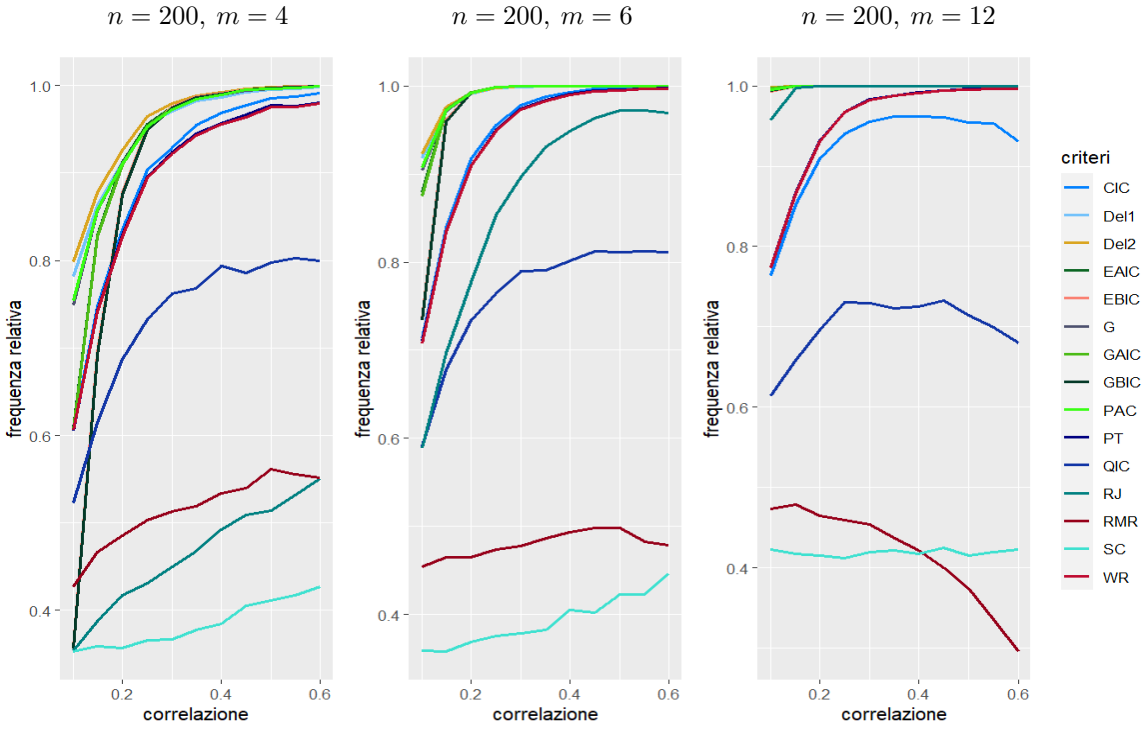


FIGURA A.18: Andamento della frequenza di corretta selezione dei diversi criteri di selezione al variare di α, per $n = 50$ e $m = 4$, 6, 12 rispettivamente quando la risposta è binaria e la vera struttura di correlazione è AR(1).

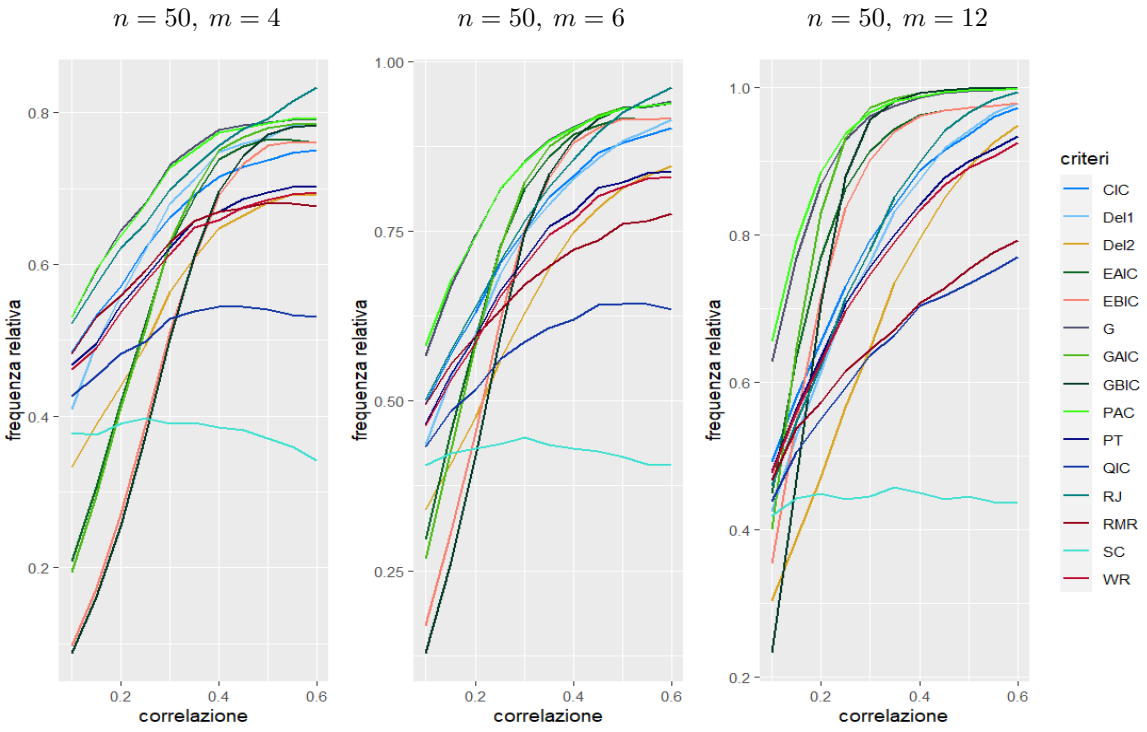


FIGURA A.19: Andamento della frequenza di corretta selezione dei diversi criteri di selezione al variare di α, per $n = 100$ e $m = 4$, 6, 12 rispettivamente quando la risposta è binaria e la vera struttura di correlazione è AR(1).

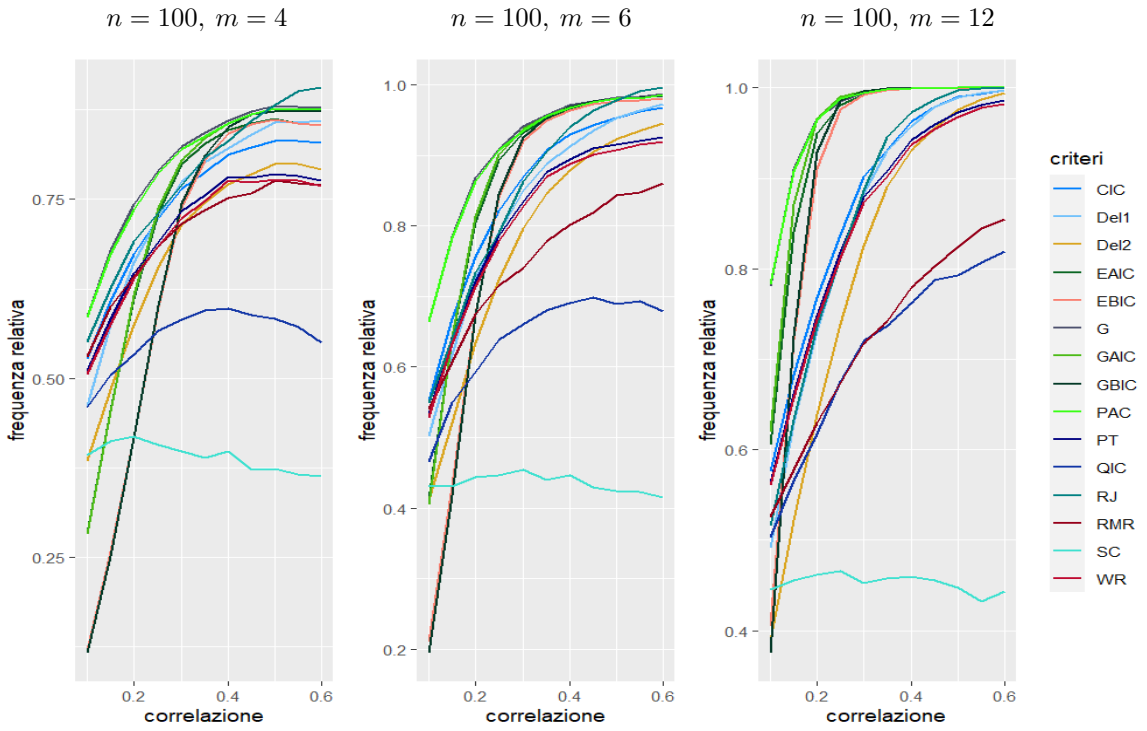


FIGURA A.20: Andamento della frequenza di corretta selezione dei diversi criteri di selezione al variare di α, per $n = 200$ e $m = 4$, 6, 12 rispettivamente quando la risposta è binaria e la vera struttura di correlazione è AR(1).

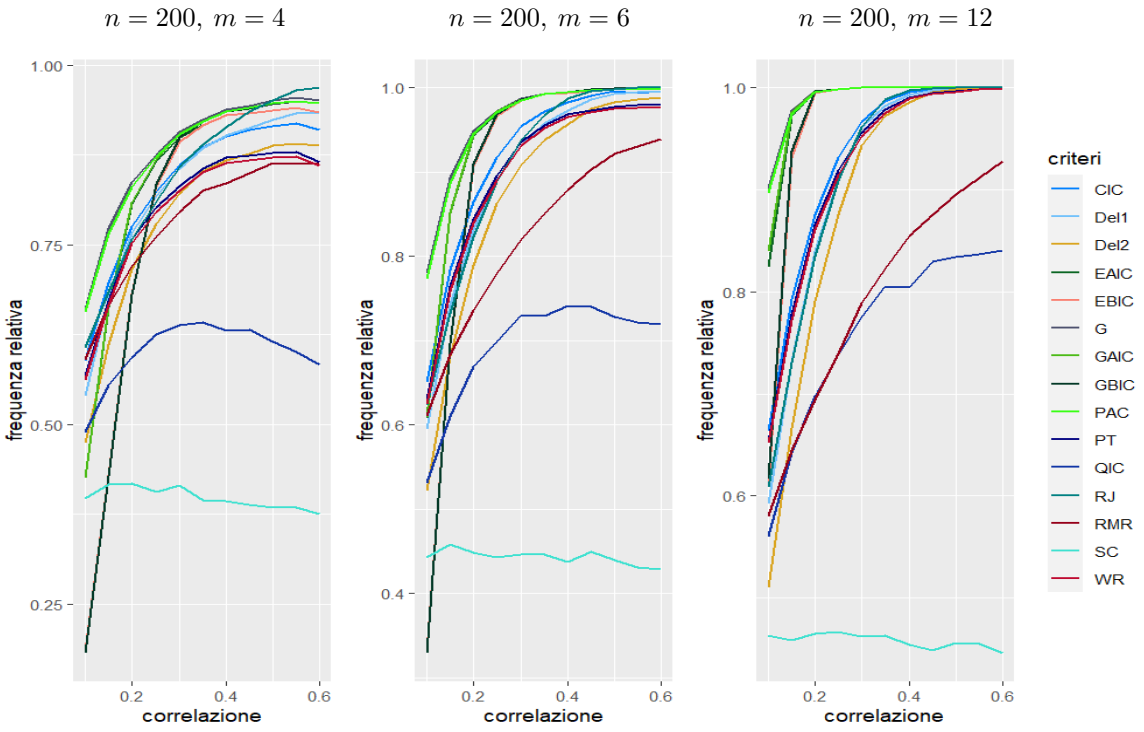
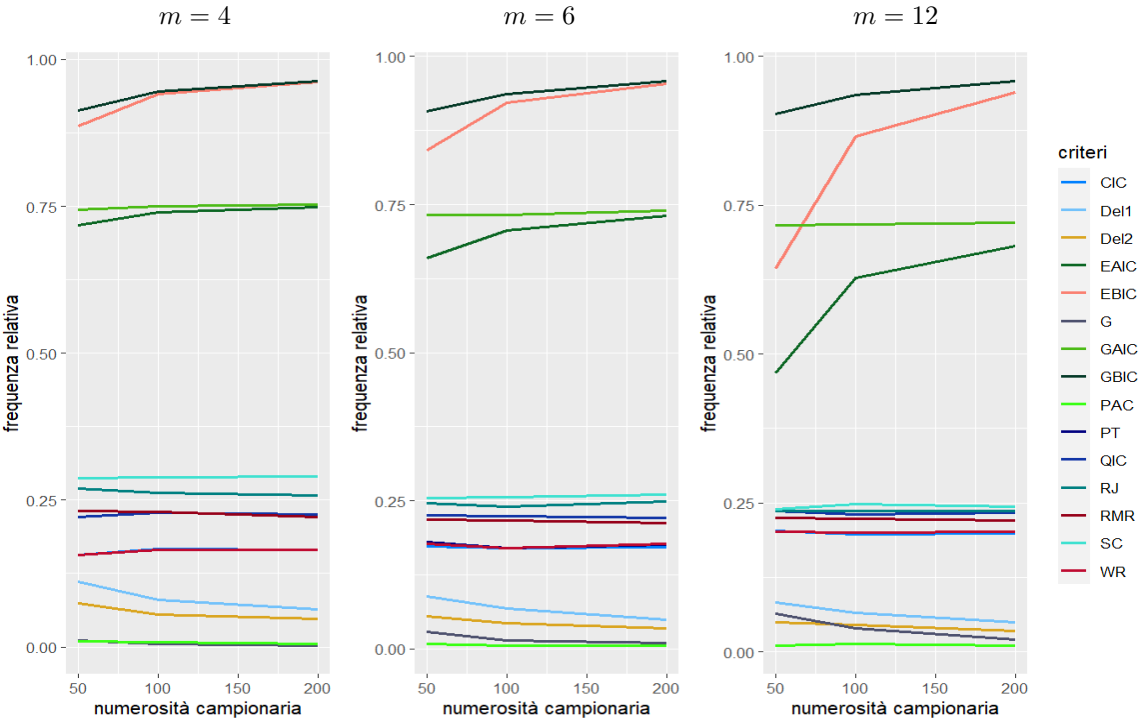


FIGURA A.21: Andamento della frequenza di corretta selezione dei diversi criteri di selezione al variare della numerosità campionaria n, per m fissato rispettivamente a 4, 6 e 12, quando la risposta è binaria e la vera struttura di correlazione è l'indipendenza.



Codice R

```
library(mvtnorm)
library(gee)
library(emplik)
library(SimCorMultRes)
library(PoisNor)
##GENERAZIONE RISPOSTA GAUSSIANA MULTIVARIATA CORRELATA##
genera.risp.gauss = function(x,beta,cor,n,k){
  X = trasf.dati(x,n,k)
  mu.vet = NULL
  eps = c(t(rmvnorm(n,mean = rep(0,k), sigma = cor)))
  for (i in 1:n){
    mu.hat = X[,,i]%*% beta
    mu.vet = c(mu.vet,mu.hat)}
  y = mu.vet + eps
  return (y)
}
##GENERAZIONE RISPOSTA BINARIA MULTIVARIATA CORRELATA##
genera.risp.bin= function(x, beta, cor,n,k){
  beta1 = beta[1]
  beta2 = beta[2:3]
  dati.cov =  x[,c(2,3)]
  for (i in 1:n){
    y = rbin(12, intercepts = beta1, betas = beta2,
    xformula =~ dati.cov, link = 'logit', cor.matrix = cor)
  }
  return(y)
}
##GENERAZIONE RISPOSTA DI CONTEGGIO MULTIVARIATA CORRELATA##
genera.risp.poi = function(x, beta, cor,n,k){
  X = trasf.dati(x,n,k)
  Y=NULL
  for (i in 1:n){
    mu.hat = exp(X[,,i] %*% beta)
    y = c(genPoisNor(1,no.pois=k,no.norm =0,mean.vec= NULL,
    sd.vec=NULL,
    cmat.star=cor,lamvec=mu.hat))
    Y = c(Y, y)
```

```
38   }
39   return(Y)
40 }
41 ##MATRICE <-> ARRAY##
42 trasf.dati = function(mat,n,k){
43   p = length(beta)
44   a = array(NA,dim = c(k,p,n))
45   o = 1
46   while (TRUE){
47     if (o > n) return(a)
48     righe = 1:k
49     temp = as.matrix(mat[righe,],nrow=k,ncol=p)
50     a[,,o] = temp
51     mat = mat[-c(righe),]
52     o = o+1}
53 }
54 ##GENERATORI COVARIATE##
55 gen.covariate.norm <- function(tot.sim,n,k,p){
56   arr = array(NA, dim = c((n*k),p,tot.sim))
57   for (i in 1:tot.sim){
58     x = cbind(rbinom(n*k,1,0.5),rnorm((n*k),0,1))
59     arr[,,i] = x
60   }
61   return(arr)
62 }
63 gen.covariate.bin.poi <- function(tot.sim,n,k,p){
64   arr = array(NA, dim = c((n*k),p,tot.sim))
65   for (i in 1:tot.sim){
66     x = cbind(rep(1,n*k), rbinom(n*k,1,0.5), rep(seq(0,k-1),n))
67     arr[,,i] = x
68   }
69   return(arr)
70 }
71 ##QIC-CIC-RJ- Delta1- Delta2##
72 criteri = function(beta, a,phi, x,y,id,n,k, cor){
73   p= ncol(x)
74   omega = I0 = I1 = matrix(0,p,p)
75   ql = 0
76   if (cor=="indi") {R = diag(k)
```

```
77  }else if (cor=="exch"){ R = exch(a,k)
78  }else if(cor=="ar1"){R = ar1(a,k)}
79  for (i in 1:n){
80    xi=x[id==i,]
81    yi=y[id==i]
82    fitted = xi %*% beta
83    D = matrix(xi,nrow =k,ncol = p)
84    #se risposta binaria:
85    #fitted = plogis(xi%*%beta)
86    #D = matrix(rep(fitted*(1-fitted),p),k,p)*xi
87    #IA = diag(as.vector((fitted*(1-fitted))^(-1)))
88    #A.half = IA^(1/2)
89    #se risposta di conteggio:
90    #fitted = exp(xi %*% b)
91    #D = matrix(rep(fitted,p),k,p)*xi
92    #IA = diag(c((fitted)^(-1)))
93    #A.half = IA^(1/2)
94    res = yi-fitted
95    omega = omega + crossprod(D,D)/phi
96    I0 = I0 + crossprod(D, solve(R, D))/phi
97    I1.left = crossprod(D,solve(R,res))/phi
98    I1 = I1 + tcrossprod(I1.left, I1.left)
99    #se risposta binaria o di conteggio:
100   #omega = omega + crossprod(D, IA%*%D)
101   #I0 = I0 + crossprod(D, A.half %*% solve(R, A.half %*% D))/
    phi
102   #I1.left = crossprod(D, A.half %*% solve(R, A.half %*% res))
    /phi
103   #I1 = I1 + tcrossprod(I1.left, I1.left)
104   ql = ql - sum((res^2)/(2*phi))
105   #se risposta binaria:
106   #ql = ql + sum(yi*(log(fitted)-log(1-fitted))+log(1-fitted))
107   #se risposta di conteggio:
108   #ql = ql + sum(yi*log(fitted) -fitted)/phi
109 }
110 var.sand = solve(I0, I1)%*%solve(I0)
111 cic= tr(omega%*%var.sand)
112 qic= -2*ql+ 2*cic
113 Q = solve(I0)%*%I1
```

```
114   eig = fun(Q)
115   C1=tr(Q)/p
116   C2=tr(Q%*%Q)/p
117   det1=C2-2*C1+1
118   det2=sum((log(eig))^2)
119   RJ=sqrt((C1-1)^2+(C2-1)^2)
120   tot = c(qic,cic,RJ,det1,det2)
121   names(tot) = c('QIC','CIC','RJ','Delta1','Delta2')
122   return(tot)
123 }
124 ##Criterio di Schults e Chaganty##
125 sc = function(beta,alpha,phi,x,y,n,k,cor){
126   if (cor=="indi") {R = diag(k)}
127   else if (cor=="exch"){
128     R = exch(alpha,k)}
129   else if (cor=="ar1"){R = ar1(alpha,k)}
130   sc=0
131   for (i in 1:n){
132     xi=x[id==i,]
133     yi=y[id==i]
134     fitted = xi %*% beta
135     #se risposta binaria:
136     #fitted = plogis(xi%*%beta)
137     D = as.matrix(xi)
138     #se risposta binaria:
139     #A.sqrt = diag(c((fitted*(1-fitted))^(-1/2)))
140     #D = D = matrix(rep(fitted*(1-fitted),p),k,p)*xi
141     #se risposta di conteggio:
142     #fitted = exp(xi %*% beta)
143     #D = matrix(rep(fitted,p),k,p)*xi
144     #A.sqrt = diag(c((fitted)^(-1/2)))
145     res = yi-fitted
146     V.hat.inv= solve(R)/phi
147     #se risposta binaria o di conteggio:
148     #V.hat.inv= (A.sqrt%*%solve(R)%*%A.sqrt)/phi
149     sc=sc+t(res)%*%V.hat.inv%*%res
150   }
151   return(sc)
152 }
```

```
153 ##Criterio di Gosho-Hamada-Yoshimura##
154 gosho =function(beta,alpha,phi,x,y,n,k,cor){
155   if (cor=="indi")
156   {R = diag(k)}
157   else if (cor=="exch"){
158     R= exch(alpha,k)}
159   else if (cor=="ar1")
160   {R = ar1(alpha,k)}
161   cov=matrix(0,k,k)
162   var.tot= matrix(0,k,k)
163   for (i in 1:n){
164     xi=x[id==i,]
165     yi=y[id==i]
166     fitted = xi %*%beta
167     #se risposta binaria:
168     #fitted = plogis(xi%*%beta)
169     #se risposta di conteggio:
170     #fitted = exp(xi %*% beta)
171     res = yi-fitted
172     D = as.matrix(xi)
173     #se risposta binaria:
174     #A = diag(as.vector((fitted*(1-fitted))))
175     #Ai.sqrt = A^(1/2)
176     #D = D = matrix(rep(fitted*(1-fitted),p),k,p)*xi
177     #se risposta di conteggio:
178     #A = diag(as.vector((fitted)))
179     #D = matrix(rep(fitted,p),k,p)*xi
180     #Ai.sqrt = A^(1/2)
181     vi= R*phi
182     #se risposta binaria o di conteggio:
183     #vi= Ai.sqrt%*%R%*%Ai.sqrt*phi
184     cov = cov + res%*%t(res)
185     var.tot = var.tot+ vi
186   }
187   mat =cov %*% solve(var.tot) - diag(k)
188   CR= tr(mat%*%mat)
189   return(CR)}
190
191 ##Criterio di Pardo-Alonso
```

```
Pardo.Alonso = function(beta,alpha,phi,x,y,n,k,cor){
  if (cor=="indi")
  {R = diag(k)}
  else if (cor=="exch")
  {R = exch(alpha,k)}
  else if (cor=="ar1")
  {R = ar1(alpha,k)}
  S.tot = matrix(0,k,k)
  V.tot = matrix(0,k,k)
  for(i in 1:n){
    xi = x[id==i,]
    yi = y[id ==i]
    fitted = xi%*% beta
    #se risposta binaria:
    #fitted = plogis(xi%*%beta)
    #se risposta di conteggio:
    #fitted = exp(xi%*% beta)
    V.hat = R*phi
    #se risposta binaria:
    #A.sqrt = diag(c(fitted*(1-fitted))^(1/2))
    #V.hat = (A.sqrt %*% R %*% A.sqrt)*phi
    #se risposta di conteggio:
    #A.sqrt = diag(c(fitted)^(1/2))
    #V.hat = (A.sqrt %*% R %*% A.sqrt)*phi
    res = yi-fitted
    S.quad = res%*%t(res)
    S.tot = S.tot + S.quad
    V.tot = V.tot + V.hat
  }
  end = abs((det(S.tot/n)/(det(V.tot/n)))-1)
  return(end)
  }


##Calcolo pseudo-verosimiglianza gaussiana##
carey.wang = function(beta, alpha,x,y,n,k,cor,phi){
  if (cor == 'indi'){
    R = diag(k)}
  else if (cor =='exch'){
```

```
    R = exch(alpha,k)}
  else if (cor == 'ar1'){
    R = ar1(alpha,k)}
  L = 0
  for(i in 1:n){
    xi = x[id==i,]
    yi = y[id==i]
    mu.hat = xi%*%beta
    V = R*phi
    #se risposta binaria:
    #mu.hat = plogis(xi%*%beta)
    #A.sqrt = diag(c(mu.hat*(1-mu.hat))^(1/2))
    #V = (A.sqrt%*%(R)%*%A.sqrt)*phi
    #se risposta di conteggio:
    #mu.hat = exp(xi%*%beta)
    #res = yi -mu.hat
    #A.sqrt = diag(c(mu.hat)^(1/2))
    #V = (A.sqrt%*%(R)%*%A.sqrt)*phi
    res = yi -mu.hat
    L = L-(t(res)%*%solve(V)%*%res + log(det(V)))/2
  }
return(L)}

##Calcolo rapporto di verosimiglianza empirica##
emp.lik <-  function(beta,alpha,x,y,n,k,p) {
  X= trasf.dati(x,n,k)
  Y =  matrix(y,k,n)
  g = matrix(0, (p+k-1),n)
  R = toep(alpha)
  res= matrix(0,k,n)
  D = array(0,c(k,p,n))
  for (i in 1:n) {
    fitted = X[,,i] %*% beta
    D[,,i] = as.matrix(X[,,i])
    #se risposta binaria:
    #fitted = plogis( X[,,i] %*% beta )
    #pearson.res[,i] <- res[,i]*(fitted*(1-fitted))^(-1/2)
    #D[,,i] = matrix(rep(fitted*(1-fitted),p),k,p)*X[,,i]
    #A[,,i] = diag(as.vector((fitted*(1-fitted))^(-1)))
```

```
270     #se risposta di conteggio:
271     #fitted <- exp( X[,,i] %*% beta)
272     #pearson.res[,i] <- res[,i]*(fitted)^(-1/2)
273     #D[,,i] <- matrix(rep(fitted,p),k,p)*X[,,i]
274     #A[,,i] <- diag(as.vector((fitted)^(-1)))
275     res[,i] <- Y[,i]-fitted
276   }
277   phi.tilde <- sum(res^2)/(n*k-p)
278   for (i in 1:n) {
279     #se conteggio o binaria:
280     #A.half = A[,,i]^(1/2)
281     g[1:p,i] = crossprod(D[,,i],solve(R,res[,i]))/phi.tilde
282     #se risposta binaria o di conteggio:
283     g[1:p,i] = crossprod(D[,,i], A.half %*% solve(R, A.half %*%
      res[,i]))/phi.tilde
284     for(idx in 1:(k-1)){
285       sum=0
286       for(h in 1:(k-idx)){
287         sum=sum+ res[h,i]* res[h+idx,i]
288       }
289       g[p+idx,i]=sum - alpha[idx]*(k-idx-p/n)*phi.tilde
290     }
291   }
292   g <- t(g)
293   g.mu <- rep(0,(p+k-1))
294   el.test(g, g.mu,gradtol=1e-9)$"-2LLR"
295 }
296
297 ar1 <- function(cor,dim){
298   mat <- matrix(cor, nrow = dim, ncol = dim)
299   for (i in 1:dim){
300     for (j in 1:dim){
301       if (i!=j)
302       mat[i,j] <- cor^abs(i-j)
303       else mat[i,j] =1
304     }
305   }
306   return(mat)
307 }
```

```
308 ##Criteri basati su autovalori generalizzati##
309 gen.eigen = function(var.sand,var.ind,p){
310   b = solve(var.sand+var.ind)
311   a = var.sand%*%b
312   autovalori = eigen(a)$values
313   Pillai = Roy = 0
314   Wilks = 1
315   for (j in 1:p){
316     lambda = autovalori[j]
317     Pillai = Pillai + (lambda/(1+lambda))
318     Wilks = Wilks * (lambda/(1+lambda))
319     if (lambda/(1+lambda)>=Roy){
320       Roy = lambda/(1+lambda)
321     }
322   }
323   return(c(Pillai,Wilks,Roy))
324 }
325
326 exch <- function(cor,dim){
327   mat <- matrix(cor, nrow = dim, ncol = dim)
328   for (i in 1:dim){
329     for (j in 1:dim){
330       if (j == i)
331       mat[i,j] = 1
332     }
333   }
334   return (mat)
335 }
336
337 indi <- function(dim){
338   mat = diag(1,dim)
339   return(mat)
340 }
341
342 toep = function(rho){
343   a = c(1,rho)
344   toeplitz(a)
345   }
346
```

```
347   tr <-  function(mat){
348     traccia <- sum(diag(mat))
349     return(traccia)
350   }
351 fun = function(mat){
352   eig = eigen(mat)$values
353   for (i in 1:length(eig)){
354     if (eig[i] == 0)
355     eig[i]=1
356   }
357   return(eig)
358 }
359
360 ##MAIN##
361 set.seed(1)
362 #per conteggio
363 set.seed(2)
364 num.sim = 10
365 n = 50
366 k = 4
367 beta = c(1, 1)
368 #per risposta binaria e poisson
369 #beta = c(0.5, -0.2, -0.2)
370 p=length(beta)
371 rho.vettore=seq(from=0.1,to=0.8,by=0.05)
372 #rho.vettore =seq(from=0.1,to=0.6,by=0.05)
373 tot.oss= n*k
374 tot.sim = num.sim*length(rho.vettore)
375 param.mod = c(p,p+1,p+1)
376 #simulazione :)
377 rho.ex = matrix(NA,nr=num.sim,nc=length(rho.vettore))
378 rho.ar = matrix(NA,nr=num.sim,nc=length(rho.vettore))
379 stime.mod1 = array(NA,c(num.sim,p,length(rho.vettore)))
380 stime.mod2 = array(NA,c(num.sim,p,length(rho.vettore)))
381 stime.mod3 = array(NA,c(num.sim,p,length(rho.vettore)))
382 risul =array(NA,c(15,length(param.mod),length(rho.vettore)))
383 covariate = gen.covariate.norm(tot.sim,n,k,p)
384 #se risposta binaria o poisson:
385 #covariate = gen.covariate.bin.poi(tot.sim,n,k,p)
```

```
386 for (i in 1:length(rho.vettore)){
387   rho = rho.vettore[i]
388   QIC=CIC=RJ=Del1=Del2=EAIC =EBIC=GAIC=GBIC=SC=CR=PT=RMR=WR=PAC=
      rep(0,length(param.mod))
389   corr = exch(rho,k)
390   #corr = ar1(rho,k)
391   for (j in 1:num.sim){
392     cat('Simulazione n.',j, 'con valore di correlazione',rho)
393     x = covariate[,,j]
394     y = genera.risp.gauss(x,beta,corr,n,k)
395     #y = c(t(genera.risp.bin(x, beta, cor.teor,n,k)$Ysim))
396     #y = genera.risp.poi(x, beta, cor.teor,n,k)
397     id = rep(1:n,each = k)
398     dati = as.data.frame(cbind(y,x,id))
399     fit1 = gee(y ~ x-1, id = id, corstr = "independence", data =
      dati)
400     fit2 = gee(y ~ x-1, id = id, corstr = "exchangeable", data =
      dati)
401     fit3 = gee(y ~ x-1, id = id, corstr = "AR-M",data = dati)
402     #fit1 = gee(y ~ x-1, id=id, data=dati, corstr="independence
     ",family = binomial)
403     #fit2 = gee(y ~ x-1, id=id, data=dati, corstr="exchangeable
     ",family= binomial)
404     #fit3 = gee(y ~ x-1, id=id, corstr="AR-M", Mv=1,data=dati,
     family = binomial)
405     #fit1 = gee(y ~ x-1, id=id, data=dati, corstr="independence
     ",family = poisson)
406     #fit2 = gee(y ~ x-1, id=id, data=dati, corstr="exchangeable
     ",family = poisson)
407     #fit3 = gee(y ~ x-1, id=id, corstr="AR-M", Mv=1,family =
     poisson,data=dati)
408     stime.fit1 = fit1$coefficients
409     stime.fit2 = fit2$coefficients
410     stime.fit3 = fit3$coefficients
411     stime.mod1[j,,i] = stime.fit1
412     stime.mod2[j,,i] = stime.fit2
413     stime.mod3[j,,i]= stime.fit3
414     alpha.exch = fit2$working.correlation[1,2:k]
415     alpha.ar = fit3$working.correlation[1,2:k]
```

```
416     rho.ex[j,i] = alpha.exch[1]
417     rho.ar[j,i] = alpha.ar[1]
418     phi1 = fit1$scale
419     phi2 = fit2$scale
420     phi3 = fit3$scale
421     #se risposta binaria:
422     #phi = 1
423     #QIC etc..
424     crc.ind = criteri(stime.fit1,0, phi1,x,y,id,n,k, cor="indi")
425     crc.exc = criteri(stime.fit2,alpha.exch[1],phi2,x,y,id,n,k,
    cor="exch")
426     crc.ar = criteri(stime.fit3, alpha.ar[1],phi3,x,y,id,n,k,
    cor="ar1")
427     #crc.ind = criteri(stime.fit1,0, phi,x,y,id,n,k, cor="indi")
428     #per risposta binarie
429     #crc.exc = criteri(stime.fit2,alpha.exch[1],phi,x,y,id,n,k,
    cor="exch")
430     #crc.ar = criteri(stime.fit3, alpha.ar[1],phi,x,y,id,n,k,
    cor="ar1")
431     criteria= rbind(crc.ind,crc.exc,crc.ar)
432     qic = criteria[,1]
433     cic = criteria[,2]
434     rj=criteria[,3]
435     det1=criteria[,4]
436     det2=criteria[,5]
437     idx1=which.min(qic)
438     QIC[idx1] = QIC[idx1]+1
439     idx2=which.min(cic)
440     CIC[idx2] = CIC[idx2]+1
441     idx3=which.min(rj)
442     RJ[idx3]=RJ[idx3]+1
443     idx4=which.min(det1)
444     Del1[idx4]=Del1[idx4]+1
445     idx5=which.min(det2)
446     Del2[idx5]=Del2[idx5]+1
447     #Vero empirica, calcolo EAIC e EBIC
448     emplik1 = emp.lik(stime.fit1,rep(0,k-1),x,y,n,k,p)
449     emplik2 = emp.lik(stime.fit2,alpha.exch,x,y,n,k,p)
450     emplik3 = emp.lik(stime.fit3,alpha.ar,x,y,n,k,p)
```

```
451     emplik.vet = c(emplik1,emplik2,emplik3)
452     eaic = emplik.vet + 2*param.mod
453     ebic = emplik.vet + param.mod*log(n)
454     idx6 = which.min(eaic)
455     idx7 = which.min(ebic)
456     EAIC[idx6] = EAIC[idx6]+1
457     EBIC[idx7] = EBIC[idx7]+1
458     #Pardo Alonso
459     pa1 = Pardo.Alonso(stime.fit1,0,phi1,x,y,n,k,cor= 'indi')
460     pa2 = Pardo.Alonso(stime.fit2,alpha.exch[1],phi2,x,y,n,k,cor
    = 'exch')
461     pa3 = Pardo.Alonso(stime.fit3,alpha.ar[1],phi3,x,y,n,k,cor=
    'ar1')
462     #per risposta binarie
463     #pa1 = Pardo.Alonso(stime.fit1,0,phi,x,y,n,k,cor= 'indi')
464     #pa2 = Pardo.Alonso(stime.fit2,alpha.exch[1],phi,x,y,n,k,cor
    = 'exch')
465     #pa3 = Pardo.Alonso(stime.fit3,alpha.ar[1],phi,x,y,n,k,cor=
    'ar1')
466     pa1.vet = c(pa1,pa2,pa3)
467     idx8 = which.min(pa1.vet)
468     PAC[idx8]= PAC[idx8]+1
469     #pseudo-verosimgilianza gaussiana, calcolo GAIC e GBIC
470     cw1= carey.wang(stime.fit1,0,x,y,n,k, cor="indi",phi1)
471     cw2= carey.wang(stime.fit2, alpha.exch[1],x,y,n,k, cor="exch
    ",phi2)
472     cw3= carey.wang(stime.fit3, alpha.ar[1],x,y,n,k, cor="ar1",
    phi3)
473     #per risposta binaria
474     #cw1= carey.wang(stime.fit1,0,x,y,n,k, cor="indi",phi)
475     #cw2= carey.wang(stime.fit2, alpha.exch[1],x,y,n,k, cor="
    exch",phi)
476     #cw3= carey.wang(stime.fit3, alpha.ar[1],x,y,n,k, cor="ar1",
    phi)
477     gauss1 = -2*cw1 +2*p
478     gauss2 = -2*cw2 +(2*(p+1))
479     gauss3 = -2*cw3 + (2*(p+1))
480     gauss.tot = cbind(gauss1,gauss2,gauss3)
481     gbayes1 = -2*cw1 +log(n)*p
```

```
482     gbayes2 = -2*cw2 +log(n)*(p+1)
483     gbayes3 = -2*cw3 +log(n)*(p+1)
484     idx10 = which.min(gauss.tot)
485     gbayes.tot = c(gbayes1,gbayes2,gbayes3)
486     idx11 = which.min(gbayes.tot)
487     GAIC[idx10] = GAIC[idx10]+1
488     GBIC[idx11] = GBIC[idx11]+1
489     # Gosho Hamada Yoshimura
490     gosho1 = gosho(stime.fit1,0,phi1,x,y,n,k,cor = 'indi')
491     gosho2 = gosho(stime.fit2,alpha.exch[1],phi2,x,y,n,k,cor = '
    exch')
492     gosho3 = gosho(stime.fit3,alpha.ar[1],phi3,x,y,n,k,cor = '
    ar1')
493     #per risposta binaria
494     #gosho1 = gosho(stime.fit1,0,phi,x,y,n,k,cor = 'indi')
495     #gosho2 = gosho(stime.fit2,alpha.exch[1],phi,x,y,n,k,cor = '
    exch')
496     #gosho3 = gosho(stime.fit3,alpha.ar[1],phi,x,y,n,k,cor = '
    ar1')
497     gosho.vet = c(gosho1,gosho2,gosho3)
498     idx12 = which.min(gosho.vet)
499     CR[idx12] = CR[idx12]+1
500     #Schult - Chaganty
501     sc1 = sc(stime.fit1,0,phi1,x,y,n,k,cor = 'indi')
502     sc2 = sc(stime.fit2,alpha.exch[1],phi2,x,y,n,k,cor = 'exch')
503     sc3 = sc(stime.fit3,alpha.ar[1],phi3,x,y,n,k,cor = 'ar1')
504     #per risposta binaria
505     #sc1 = sc(stime.fit1,0,phi,x,y,n,k,cor = 'indi')
506     #sc2 = sc(stime.fit2,alpha.exch[1],phi,x,y,n,k,cor = 'exch')
507     #sc3 = sc(stime.fit3,alpha.ar[1],phi,x,y,n,k,cor = 'ar1')
508     sc.vet = c(sc1,sc2,sc3)
509     idx13 = which.min(sc.vet)
510     SC[idx13]= SC[idx13]+1
511     #Autovettori Generalizzati
512     eig.norm1 = gen.eigen(fit1$robust.variance,fit1$naive.
    variance,p)
513     eig.norm2 = gen.eigen(fit2$robust.variance,fit1$naive.
    variance,p)
```

```
514     eig.norm3 = gen.eigen(fit3$robust.variance,fit1$naive.
    variance,p)
515     pillai = c(eig.norm1[1],eig.norm2[1],eig.norm3[1])
516     wilks = c(eig.norm1[2],eig.norm2[2],eig.norm3[2])
517     roy =c(eig.norm1[3],eig.norm2[3],eig.norm3[3])
518     idx14 = which.min(pillai)
519     idx15 = which.min(wilks)
520     idx16 = which.min(roy)
521     PT[idx14] = PT[idx14]+1
522     WR[idx15]= WR[idx15]+1
523     RMR[idx16] = RMR[idx16]+1
524   }
525   risul[,,i]= rbind(QIC,CIC,RJ,Del1,Del2,EAIC,EBIC,GAIC,GBIC,CR,
     SC,PAC,PT,WR,RMR)/num.sim
526   }
527 colnames(risul)=c("Indipendenza","Sferica","Autoregressiva")
528 intest=c("QIC","CIC","RJ","Delta1","Delta2",'EAIC','EBIC','GAIC'
    ,'GBIC',"C(R)","SC",'PAC','PT','WR','RMR')
529 rownames(risul)= intest
```

Bibliografia

Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Wiley.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proceedings of the Second International Symposium on Information Theory* , 267–281.

Amatya, A., Demirtas, H. & Gao, R. (2021). *PoisNor: Simultaneous Generation of Multivariate Data with Poisson and Normal Marginals*. R package version 1.3.3.

Carey, V. & Wang, Y.-G. (2011). Working covariance model selection for generalized estimating equations. *Statistics in Medicine* **30**, 3117–3124.

Carey, V. J. (2022). *gee: Generalized Estimation Equation Solver*. R package version 4.3-25.

Chen, J. & Lazar, N. (2012). Selection of working correlation structure in generalized estimating equations via empirical likelihood. *Journal of Computational and Graphical Statistics* **21**, 18–41.

Genz, A. & Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Heidelberg: Springer-Verlag.

Gosho, M., Hamada, C. & Yoshimura, I. (2011). Criterion for the selection of a working correlation structure in the generalized estimating equation approach for longitudinal balanced data. *Communications in Statistics- Theory and Methods* **40**, 3839–3856.

Hin., L.-Y., Carey, V. J. & Wang, Y.-G. (2007). Criteria for working correlation-structure selection in gee: assessment via simulation. *The American Statistician* **61**, 360–364.

Hin, L.-Y. & Wang, Y.-G. (2008). Working-correlation-structure identification in generalized estimating equations. *Statistics in Medicine* **28**, 642–658.

Jang, M. J. (2011). *Working correlation selection in generalized estimating equations.* Phd thesis, University of Iowa.

Liang, K. Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized liner models. *Biometrika* **73**, 13–22.

Mai, Z. & Yang, Y. (2023). *emplik: Empirical Likelihood Ratio for Censored and Truncated Data*. R package version 1.3.

McCullagh, C. L. & Nelder, J. A. (1989). *Generalized Linear Models*. Monographs on Statistics and Applied Probability. London: Chapman and Hall.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics* **18**, 90–120.

Pan, W. (2001). Akaike information criterion in generalized estimating equations. *Biometrics* **57**, 120–125.

Pardo, M. C. & Alonso, R. (2019). Working correlation structure selection in gee analysis. *Statistical Papers* **60**, 1447–1467.

Qin, J. & Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics* **22**, 300–325.

Rotnitzky, A. & Jewell, N. P. (1990). Hypothesis testing of regression parameters in semi-parametric generalized linear models for cluster correlated data. *Biometrika* **77**, 485–497.

Salvan, A., Sartori, N. & Pace, L. (2020). *Modelli Lineari Generalizzati.* Milano: Springer-Verlag Italia.

Shults, J. & Chaganty, N. (1998). Analysis of serially correlated data using quasi-least square. *Biometrics* **54**, 1622–1630.

Thall, P. F. & Vail, S. C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics* **46**, 657–671.

Touloumis, A. (2016). Simulating correlated binary and multinomial responses under marginal model specification: The simcormultres package. *The R Journal* **8**, 79–91. R package version 1.9.0.

Wang, Y.-G. & Hin, L.-Y. (2010). Modelling strategies in longitudinal data analysis: Covariate, variance function and correlation structure selection. *Computational Statistics and Data Analysis* **52**, 3359–3370.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the gauss-newton method. *Biometrika* **61**, 439–447.

Zhu, X. & Zhu, Z. (2013). Comparison of criteria to select working correlation matrix in generalized estimating equations. *Chinese Journal of Applied Probability and Statistics* **29**, 515–530.