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

Relazione Finale

Stima della matrice di covarianza per equazioni di stima generalizzate con piccoli campioni

Relatore: Prof. Alessandra Salvan
Dipartimento di Scienze Statistiche

Laureando: Alessandro Fanesi
Matricola N° 1189570

Anno Accademico 2020/2021

Indice

Introduzione

Uno dei metodi statistici più utilizzati in studi di tipo longitudidale sono le equazioni di stima generalizzate (GEE). Questa metodologia ha il grande vantaggio di essere particolarmente flessibile nella modellazione, in quanto non è vincolata all'assuzione distributiva della varibiale risposta. Tuttavia è noto che lo stimatore della matrice di covarianza dello stimatore del parametro di regressione, usualmente lo stimatore "sandwich", tende ad avere scarsa accuratezza quando la numerostà campionaria è piccola. In generale, le varianze degli stimatori vengono sottostimate. Per ridurre questo problema, recentemente sono state proposte una serie di modifiche dello stimatore "sandwich", incentrate sulla riduzione della distorsione o sull'aumento dell'efficienza. L'obiettivo di questo lavoro è illustrare le modifiche proposte e proporre un confronto tramite simulazione. Vengono inoltre mostrate esemplificazioni con dati reali. In particolare per il confronto verranno presi in esame due test, rispettivamente di Wald e t, come guida per stabilire la dimensione campionaria necessaria ad ogni stimatore per preservare l'errore di primo tipo.

Questa relazione è una rassegna dei contenuti dell'articolo Wang et al. (2016a), aggiornata in merito ad alcuni rilievi presentati in da Silva & Colosimo (2016) e successive precisazioni in Wang et al. (2016b). Il primo capitolo contiene i richiami necessari sulla teoria della verosimiglianza con l'obiettivo di introdurre la notazione necessaria per lo sviluppo del lavoro. Nel capitolo 2 vengono presentate le diverse modifiche allo stimatore "sandwich" proposto da Liang & Zeger (1986), a cui segue una sintesi dei risultati relativi ai diversi confronti teorici in termini di efficienza tra i vari stimatori. Il capitolo 3 è dedicato invece ai confronti numerici effettuati tramite uno studio di simulazione e presenta due applicazioni a casi reali. Infine vengono discussi i risultati ottenuti nei precedenti capitoli.

Capitolo 1

Inferenza di verosimiglianza

Introduzione

L'obiettivo di questo capitolo è introdurre i concetti e la notazione sull'inferenza di verosimiglianza necessari alla stesura e comprensione del lavoro svolto. Per una trattazione più estesa si rimanda a Pace & Salvan (2001). In particolare, verranno richiamate: la funzione di verosimiglianza, la stima di massima verosimiglianza, le quantità di verosimiglianza e le rispettive proprietà campionarie. Successivamente verranno introdotti i modelli di regressione lineare, con particolare attenzione ai Modelli Lineari Generalizzati (MLG) sotto le più deboli ipotesi del secondo ordine. Quest'ultima generalizzazione fornirà le basi del modello di quasi-verosimiglianza e della metodologia delle Equazioni di Stima Generalizzate (GEE: *Generalized Estimating Equations*).

1.1 Ipotesi

Per l'intera trattazione si assumerà che il fenomeno di interesse sia descritto da una variabile casuale Y con supporto S. Y ha una distribuzione di probabilità o densità di probabilità $p_Y^0(y)$ non nota, ma di cui si assume l'appartenza ad un insieme di possibili distribuzioni denotato con $\mathcal{F}$ e chiamato **modello statistico**. L'obiettivo è di individuare $p_Y^0(y)$ vera distribuzione del fenomeno all'interno di $\mathcal{F}$. Per raggiungere lo scopo prefissato, si considera l'osservazione di un campione $(y_1, \ldots, y_n)$ realizzazione del vettore casuale $Y = (Y_1, \ldots, Y_n)^T$, con Y_i, $i = 1, \ldots, n$ indipendenti. Esistono vari livelli di specificazione di modelli statistici, questi possono essere parametrici, semiparametrici e non parametrici, ma per ora verranno considerati solamente modelli parametrici. In particolare si assume che valgano le seguenti ipotesi:

- gli elementi di $\mathcal{F}$ sono tutti dello stesso tipo;
- $\mathcal{F}$ sia correttamente specificato, cioè $p_Y^0(y) \in \mathcal{F}$;
- gli elementi di $\mathcal{F}$ differiscono solamente per il valore assunto da un parametro d-dimensionale $\theta \in \Theta \subseteq \mathbb{R}^d$ con Θ chiamato **spazio parametrico** e $d \in \mathbb{N}^+$;
- il parametro θ sia identificabile, cioè esista una corrispondenza biunivoca tra Θ e $\mathcal{F}$.

In tal caso si può scrivere

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

dove, se valgono le ipotesi precedenti, θ^0 è il valore del parametro che identifica la vera distribuzione di probabilità o densità di probabiltà $p_Y^0(y) = p_Y(y;\theta^0)$ ed è chiamato vero valore del parametro.

1.2 La funzione di verosimiglianza

Si definisce ora lo strumento matematico chiave per la ricerca, a posteriori (dopo la realizzazione di un campione), del vero valore del parametro θ^0: la **funzione di verosimiglianza**, introdotta da Fisher (1922). L'obiettivo è individuare una funzione che permetta di localizzare il valore del parametro θ che più verosimilmente ha generato i dati osservati y. Questa si denota con $L(\cdot) : \Theta \to [0, +\infty)$ definita da

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

con $c(y) > 0$ costante non dipendente da θ. Spesso si ricorre, per semplicità di calcolo, alla **funzione di log-verosimiglianza**,

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

dove se $L(\theta) = 0$ allora $l(\theta) = -\infty$. Essendo $Y = (Y_1, \ldots, Y_n)^T$ un vettore di variabili casuali indipendenti possiamo allora scrivere

$$L(\theta) = \prod_{i=1}^{n} L(\theta; y_i);$$

di conseguenza passando alla trasformata logaritmica si ha

$$l(\theta) = \sum_{i=1}^{n} l(\theta; y_i),$$

che può essere interpretata come somma dei contributi individuali delle osservazioni Y_i alla funzione di log-verosimiglianza.

Il paragrafo si conclude richiamando due importanti proprietà della funzione di verosimiglianza, la **proprietà di invarianza** rispetto a trasformazioni biiettive dei dati y e la **proprietà di equivarianza** rispetto a riparametrizzazioni cioè trasformazioni biiettive del parametro θ. Nel primo caso esprimere la funzionie di verosimiglianza come $L(\theta; y(t))$ dove $y(t)$ è l'inversa della funzione biiettiva $t(y)$, non modifica l'nferenza sul parametro θ di interesse. Nel secondo caso invece vale quanto segue: si consideri una riparametrizzazione $\psi = \psi(\theta)$ con inversa $\theta(\psi)$, questo permette di calcolare la funzione di verosmiglianza anzichè in θ, in $\theta(\psi)$ in quanto entrambi θ e $\theta(\psi)$ identificano la stessa distribuzione di probabilità in $\mathcal{F}$. In quest'ultimo caso la funzione di verosimiglianza per ψ è $L^{\Psi}(\psi) = L(\theta(\psi))$.

1.3 Quantità di verosimiglianza e proprietà campionarie

1.3.1 Problema regolare di stima

In questo sottoparagrafo vengono fornite le caratteristiche principali per definire un problema di stima **regolare**:

- lo spazio parametrico Θ è un insieme aperto in $\mathbb{R}^d$,
- tutte le densità di probabilità o distribuzioni di probabilità hanno lo stesso supporto S,
- la funzione di verosimiglianza è differenziabile almeno tre volte, con derivate parziali continue in Θ,
- la differenziazione rispetto a θ e l'integrazione rispetto ad y possono essere scambiate due volte.

Questi risultati garantiscono una serie di proprietà che verranno presentate nei paragrafi successivi.

1.3.2 Stima di massima verosimiglianza

Un valore $\hat{\theta} \in \Theta$ tale che

$$L(\hat{\theta}) \geq L(\theta), \qquad \theta \in \Theta,$$

è chiamato **stima di massima verosimiglianza** (SMV). Questo rappresenta il miglior candidato, alla luce dei dati osservati, per la stima del vero valore del parametro θ^0.

Sotto condizioni di regolarità di $L(\cdot)$, la procedura di ricerca del punto di massimo, ha inizio con il calcolo delle derivate parziali prime della funzione di log-verosimiglianza,

$$l_*(\theta) = l_*(\theta; y) = \left(\frac{\partial l(\theta)}{\partial \theta_1}, \ldots, \frac{\partial l(\theta)}{\partial \theta_d} \right)^T,$$

detta **funzione di punteggio** o **funzione score**, dove $l_*(\theta)$ ha generico elemento $l_r(\theta) = \partial l(\theta) / \partial \theta_r$, $r = 1, \ldots, d$. Segue poi la risoluzione del sistema di equazioni (se $d > 1$)

$$l_*(\theta) = 0, \tag{1.1}$$

chiamato **equazione di verosimiglianza**. Spesso la soluzione del sistema viene calcolata tramite metodi numerici (ad esempio Pace & Salvan, 2001, paragrafo 4.2).

Considerato θ come vettore colonna, la matrice $d \times d$ delle derivate parziali seconde di $l(\theta)$ o prime di $l_*(\theta)$,

$$j(\theta) = j(\theta; y) = -l_{**}(\theta) = -\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^T}$$

è chiamata matrice di **informazione osservata**. Questa è legata alla curvatura della funzione di log-verosimiglianza nel punto di massimo, permette quindi di rimarcare l'evidenza, sulla base dei dati osservati, del punto di massimo $\hat{\theta}$ rispetto a valori del parametro vicini alla SMV, appartenenti allo spazio parametrico Θ. Il generico elemento della matrice $j(\theta)$ è $j_{rs}(\theta) = -\partial^2 l(\theta) / \partial \theta_r \partial \theta_s$, $r, s = 1, \ldots, d$. Analogamente, si definisce la quantità

$$i(\theta) = E_\theta \left[j(\theta; Y) \right]$$

con generico elemento $i_{rs} = E_\theta \left[i_{rs}(\theta; Y) \right]$ come **informazione attesa**.

1.3.3 Proprietà esatte

Sotto condizioni di regolarità del prolema di stima, in particolare la non dipendenza del supporto di Y da θ e la possibilità di interscambiare il segno di integrale con quello di

derivata, valgono le seguenti proprietà:

$$E_\theta \left[l_*(\theta; Y) \right] = 0 \qquad \theta \in \Theta, \tag{1.2}$$

$$V_\theta \left[l_*(\theta; Y) \right] = E_\theta \left[l_*(\theta; Y)(l_*(\theta; Y)^T) \right] = i(\theta) \qquad \theta \in \Theta, \tag{1.3}$$

chiamate **identità di Bartlett**. La non distorsione delle equazioni di verosimiglianza (1.2) è di particolare interesse in quanto sottolinea la grande capacità della massima verosimiglianza di trattare problemi di stima. In pratica si osserva che in media la funzione di verosimiglianza ha punto di massimo nel vero valore del parametro, per ogni possibile ipotetico vero valore $\theta \in \Theta$.

1.3.4 Proprietà asintotiche

Fino a questo momento $\hat{\theta}(y_1, \ldots, y_n)$ è stato trattato come una stima del vero valore del parametro, ma per esprimerne le proprietà distributive è necessario riferirsi allo **stimatore di massima verosimiglianza** $\hat{\theta}(Y_1, \ldots, Y_n)$, la cui distribuzione dipende da quella del campione e $\hat{\theta}(y_1, \ldots, y_n)$ non è altro che una sua realizzazione. Si dimostra che lo stimatore di massima verosimiglianza, anche in casi di problemi di stima non regolari, per $n \to \infty$ è consistente

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

dove $\xrightarrow{p}$ indica covergenza in probabilità, e θ è il vero valore del parametro.

Ipotizzando quindi θ vero valore del parametro, valgono una serie di approssimazioni in distribuzione delle quantità legate alla funzione di verosimiglianza. Un risultato importante che esprime la validità di alternative procedure inferenziali come l'inferenza basata su equazioni di stima non distorte (trattata in seguito), è la convergenza in distribuzione alla normale della funzione score, che dà l'approssimazione

$$l_*(\theta) \dot{\sim} N_d(0, i(\theta)). \tag{1.4}$$

Sfruttando il risultato (1.4) si dimostra il risultato per eccellenza della teoria della verosimiglianza

$$(\hat{\theta} - \theta) \dot{\sim} N_d(0, i(\theta)^{-1}). \tag{1.5}$$

Essendo $j(\hat{\theta})$ e $i(\hat{\theta})$ stimatori consistenti di $i(\theta)$ è possibile riscrivere la (1.5) come

$$(\hat{\theta} - \theta) \dot{\sim} N_d(0, j(\hat{\theta})^{-1}). \tag{1.6}$$

Inoltre valgono anche le seguenti approssimazioni in distribuzione:

$$W_e(\theta) = (\hat{\theta} - \theta)^T j(\hat{\theta})(\hat{\theta} - \theta) \dot{\sim} \chi^2_d, \tag{1.7}$$

$$W_u(\theta) = l_*(\theta)^T i(\theta)^{-1} l_*(\theta) \dot{\sim} \chi^2_d, \tag{1.8}$$

$$W(\theta) = 2\{l(\hat{\theta}) - l(\theta)\} \dot{\sim} \chi^2_d; \tag{1.9}$$

dove le quantità $W_e(\theta)$, $W_u(\theta)$ e $W(\theta)$ sono rispettivamente denominate, quantità di **Wald**, **score** e **del rapporto di verosimiglianza**. Ognuna si distingue per alcune caratteristiche come l'equivarianza rispetto a riparametrizzazioni delle quantità score e del rapporto di verosimiglianza, la capacità di quest'ultima di adattarsi meglio alla funzione di verosimiglianza o la facilità di calcolo ed interpretazione della $W_e(\theta)$ e della $W_u(\theta)$. Tuttavia, al divergere di n le tre quantità pivotali approssimate sono equivalenti.

Spesso non si è interessati all'inferenza su tutte le componenti del vettore di parametri θ, ma solo ad un possibile sottoinsieme di queste, chiamato **parametro di interesse**. Sia $\theta = (\tau, \zeta) \in \Theta \subseteq \mathbb{R}^d$ con $\tau \in \mathbb{R}^{d1}$ parametro di interesse e $\zeta \in \mathbb{R}^{d-d1}$ **parametro di disturbo**, allora è possibile suddividere le quantità $\hat{\theta}$, $l_*(\theta)$, $i(\theta)$ e $j(\theta)$ nei blocchi relativi a τ e ζ: $\hat{\theta} = (\hat{\tau}, \hat{\zeta})$, $l_*(\theta)^T = (l_\tau(\theta)^T, l_\zeta(\theta)^T)$,

$$i(\theta) = \begin{pmatrix} i_{\tau\tau} & i_{\tau\zeta} \\ i_{\zeta\tau} & i_{\zeta\zeta} \end{pmatrix}, \qquad j(\theta) = \begin{pmatrix} j_{\tau\tau} & j_{\tau\zeta} \\ j_{\zeta\tau} & j_{\zeta\zeta} \end{pmatrix}.$$

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

$$i(\theta)^{-1} = \begin{pmatrix} i^{\tau\tau} & i^{\tau\zeta} \\ i^{\zeta\tau} & i^{\zeta\zeta} \end{pmatrix}, \qquad j(\theta)^{-1} = \begin{pmatrix} j^{\tau\tau} & j^{\tau\zeta} \\ j^{\zeta\tau} & j^{\zeta\zeta} \end{pmatrix},$$

dove valgono le seguenti relazioni,

$$\begin{aligned} i^{\tau\tau} &= (i_{\tau\tau} - i_{\tau\zeta} i_{\zeta\zeta}^{-1} i_{\zeta\tau})^{-1} \\ i^{\tau\zeta} &= -i^{\tau\tau} i_{\tau\zeta} i_{\zeta\zeta}^{-1} \\ i^{\zeta\tau} &= -i^{\zeta\zeta} i_{\zeta\tau} i_{\tau\tau}^{-1} \\ i^{\zeta\zeta} &= (i_{\zeta\zeta} - i_{\zeta\tau} i_{\tau\tau}^{-1} i_{\tau\zeta})^{-1}, \end{aligned}$$

analogamente per $j(\theta)^{-1}$.

Infine, come per l'inferenza su θ, valgono risultati analoghi ai (1.5)- (1.9) per l'inferenza sul parametro di interesse τ. Si indichi con $\hat{\theta}_\tau$ la stima di massima verosimiglianza

di θ nel sottomodello con τ fissato, dove $\hat{\theta}_\tau = (\tau, \hat{\zeta}_\tau)$, con $\hat{\zeta}_\tau$ stima di massima verosimiglianza di ζ per τ fissato, soluzione rispetto a ζ dell'equazione di verosimiglianza parziale $l_\zeta(\tau, \zeta) = 0$. Si ha

$$\begin{aligned}
\hat{\tau} - \tau &\dot{\sim} N_{d1}(0, i^{\tau\tau}(\hat{\theta})), \\
\hat{\tau} - \tau &\dot{\sim} N_{d1}(0, j^{\tau\tau}(\hat{\theta})), \\
l_\tau(\hat{\theta_\tau}) &\dot{\sim} N_{d1}(0, j^{\tau\tau}(\hat{\theta})), \\
W_{eP}(\tau) = (\hat{\tau} - \tau)^T (j^{\tau\tau}(\hat{\theta}))^{-1} (\hat{\tau} - \tau) &\dot{\sim} \chi^2_{d1}\,, \qquad (1.10) \\
W_{uP}(\tau) = l_\tau(\hat{\theta_\tau})^T i^{\tau\tau}(\hat{\theta_\tau}) l_\tau(\hat{\theta_\tau}) &\dot{\sim} \chi^2_{d1}\,, \qquad (1.11) \\
W_P(\tau) = 2\{l(\hat{\theta}) - l(\hat{\theta_\tau})\} &\dot{\sim} \chi^2_{d1}\,. \qquad (1.12)
\end{aligned}$$

I risultati di approssimazione delle quantità $W_{eP}(\tau)$ e $W_{uP}(\tau)$ valgono rispettivamente anche con le matrici $i^{\tau\tau}(\cdot)$ e $j^{\tau\tau}(\cdot)$ calcolate in $\hat{\theta}$ o in $\hat{\theta}_\tau$. La funzione $l(\hat{\theta}_\tau) = l(\tau, \hat{\zeta}_\tau)$ è chiamata **log-verosimiglianza profilo**.

1.4 Modelli lineari generalizzati

1.4.1 Introduzione ai modelli lineari generalizzati

Uno dei principali obiettivi della statistica è studiare la distribuzione di una variabile, denominata **variabile risposta**, condizionatamente ad altre variabili, denominate **esplicative** o **concomitanti**, le quali si pensa possano dare una qualche informazione sul comportamento della risposta. In questo contesto entrano quindi in gioco i modelli di regressione, dove Y viene indicata come variabile risposta e, differentemente da quanto detto in precedenza, questa è osservata contemporaneamente a p variabili concomitanti $x_{i1}, \ldots, x_{ip}$, $i = 1, \ldots, n$, non stocastiche, cioè di cui non si è interessati a studiare il comportamento in termini di variabili casuali. Ipotizzando quindi tale struttura, si nota che l'assunzione di identica distribuzione delle Y_i, $i = 1, \ldots, n$ è violata. In questa sede verranno trattati esclusivamente modelli di regressione specificati dalle assunzioni

$$\begin{aligned}
&Y_1, \ldots, Y_n \quad \text{v.c indipendenti}, &(1.13)\\
&Y_i \sim p(y_i; \theta_i) \quad \text{con} \quad E_{\theta_i}(Y_i) = \mu_i, &(1.14)\\
&g(E_{\theta_i}(Y_i)) = g(\mu_i) = \eta_i = x_i\beta = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, &(1.15)\\
&X \text{ matrice } n \times p,\ p < n, \text{con righe } x_1, \ldots, x_n \text{ e rango pieno } p, &(1.16)
\end{aligned}$$

con $x_i = (x_{i1}, \ldots, x_{ip})$ vettore riga delle variabili esplicative per l'i-esima osservazione, $i = 1, \ldots, n$, $\beta = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ vettore dei coefficienti di regressione, $p(y_i; \theta_i)$ distribuzione della risposta con $\theta_i = (\beta, \tau_i)$ vettore di parametri e $g(\cdot)$ **funzione di legame** (*link function*) liscia, invertibile e nota, che costituisce un legame univoco tra il valore atteso della risposta e il **predittore lineare** η_i. Il modello di regressione più frequentemente utilizzato è il modello lineare normale, dove $g(\cdot)$ è la funzione identità e la (1.14) è specificata con l'assunzione di normalità e omoschedasticità tale per cui $Y_i \sim N(x_i\beta, \sigma^2)$, $i = 1, \ldots, n$. La caratteristica principale di questo modello è la semplicità, sia per quanto riguarda la stima dei parametri $(\beta_1, \ldots, \beta_p, \sigma^2)$, sia per l'interpretazione dei coefficienti β. Tuttavia presenta due forti limitazioni, la prima dovuta proprio all'assunzione di omoschedasticità e la seconda alla poca flessibilità nel trattare risposte che non siano continue (ad esempio poisson o binomiale). I **modelli lineari generalizzati (GLM: *Generalized Linear Model*)** al contrario presentano una certa flessibilità nel trattare la distribuzione della risposta e nel modellare alcune forme di eteroschedasticità. Questi modelli in particolare mantengono le ipotesi (1.13), (1.16) e (1.15) mentre la (1.14) è specificata da

$$Y_i \sim DE_1(\mu_i, a_i(\phi)v(\mu_i)), \quad \mu_i \in M, \tag{1.17}$$

quindi Y_i appartiene alla **famiglia di dispersione esponenziale** (si veda l'appendice A per maggiori dettagli). Il parametro di un GLM è (β, ϕ) e la (1.15), con $g(\cdot)$ scelta opportunamente, permette di modellare diverse tipologie di variabile risposta, lasciando libero di variare in $\mathbb{R}^p$ il parametro β ed imponendo un legame tra lo spazio delle medie M e lo spazio $\mathbb{R}$ del predittore lineare η_i tramite la funzione di legame.

In particolare questa trattazione si concentra, oltre che su risposte di tipo continuo (normale), anche su altri due tipi di risposta: binaria e conteggio.

Risposta binaria

Nel caso di variabile risposta Y_i binaria, la cui media assume valori in $M = (0, 1)$ è necessario specificare una funzione di legame $g(\cdot)$ tale che $g : (0, 1) \to \mathbb{R}$. Si osserva che in generale l'inversa di qualsiasi funzione di ripartizione $F(\cdot)$ di variabili continue con supporto $\mathbb{R}$ è adatta a modellare questo tipo di rsposta, assumeremo quindi $g(\mu_i) = F^{-1}(\mu_i) = x_i\beta$, dove la funzione di legame **logistica** o **logit**

$$g(\mu_i) = \log\left(\frac{\mu_i}{1 - \mu_i}\right) = x_i\beta$$

è la più utilizzata.

Risposta conteggio

Per quanto riguarda una variabile risposta di tipo conteggio, dove $M = \mathbb{R}^+$, è neccessario definire una funzione di legame $g(\cdot)$ tale che $g : \mathbb{R}^+ \to \mathbb{R}$. La funzione di legame più utilizzata in questo caso è la **logaritmica**,

$$g(\mu_i) = \log(\mu_i) = x_i\beta.$$

1.4.2 Verosimiglianza e inferenza

Funzione punteggio ed equazioni di verosimiglianza

Sotto le assunzioni (1.13)- (1.16) e (1.17), la funzione di log-verosimiglianza risulta

$$l(\beta, \phi) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + \sum_{i=1}^{n} c(y_i, \phi) \tag{1.18}$$

con $\theta_i = \theta(\mu_i) = \theta(g^{-1}(x_i\beta))$. Assumendo ϕ noto e calcolando la derivata prima della log-verosimiglianza (1.18) si ottiene la funzione **score**

$$l_r = \sum_{i=1}^{n} \frac{1}{a_i(\phi)}(y_i - \mu_i)\frac{\partial\theta_i}{\partial\beta_r}, \quad r = 1, \ldots, p. \tag{1.19}$$

Sfruttando il risultato (1.19) è possibile scrivere le **equazioni di verosimiglianza per** β come

$$l_r = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{Var(Y_i)}\frac{\partial\eta_i}{\partial\beta_r} = 0, \quad r = 1, \ldots, p. \tag{1.20}$$

Inoltre la (1.20) può essere riscritta nella forma matriciale

$$D^T V^{-1}(y - \mu) = 0, \tag{1.21}$$

dove $y - \mu = (y_1 - \mu_1, \ldots, y_n - \mu_n)^T$, $V = \text{diag}[Var(Y_i)]$, $i =, \ldots, n$ e D è una matrice $n \times p$ con generico elemento $d_{ir} = \frac{\partial\mu_i}{\partial\beta_r} = \frac{\partial\mu_i}{\partial\eta_i}\frac{\partial\eta_i}{\partial\beta_i} = \frac{1}{g'(\mu_i)}x_{ir}$, $i = 1, \ldots, n$, $r = 1, \ldots, p$. Le equazioni (1.21) vanno risolte con metodi iterativi, tranne nel caso in cui il modello è lineare normale, dove presentano soluzione esplicita.

Informazione attesa e osservata

In un GLM si dimostra che i parametri β e ϕ sono ortogonali (si veda Salvan et al., 2020, paragrafo 1.5.3). Di conseguenza per l'inferenza su β è sufficiente il blocco di informazione osservata, o attesa relativa, a β. Derivando rispetto a β_s, $s = 1, \ldots, p$, la funzione (1.19), si ottiene

$$j_{rs} = -l_{rs} = \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \left\{ \frac{\partial \mu_i}{\partial \beta_i} \frac{\partial \theta_i}{\partial \beta_r} - (y_i - \mu_i) \frac{\partial^2 \theta_i}{\partial \beta_r \partial \beta_s} \right\}, \tag{1.22}$$

il cui valore atteso è

$$i_{rs} = E(j_{rs}) = \sum_{i=1}^{n} \frac{1}{a_i(\phi)} \frac{x_{ir} x_{is}}{(g'(\mu_i))^2 v(\mu_i)}. \tag{1.23}$$

La (1.23) si può riscrivere nella forma matriciale

$$i_{\beta\beta} = X^T W X, \tag{1.24}$$

dove $W = \text{diag}(w_i)$ con $w_i = \frac{1}{(g'(\mu_i))^2 Var(Y_i)}$, $i = 1 \ldots, n$.

Infine sfruttando il risultato generale della teoria asintotica della verosimiglianza, per $n \to \infty$

$$\hat{\beta} \dot{\sim} N_p(\beta, (X^T W X)^{-1}), \tag{1.25}$$

dove una stiama consistente della matrice si varianza e covarianza di β è $(X^T \hat{W} X)^{-1}$, con W calcolata in $\beta = \hat{\beta}$.

1.5 Modelli per risposte correlate

Viene ora introdotta una nuova classe di modelli frequentemete utilizzati in contesti dove per il fenomeno in analisi è naturale assumere una qualche forma di dipendenza. Alcuni esempi sono gli studi longitudinali in ambito clinico o economico, dove si presume una correlazione rispetto a rilevazioni sugli stessi soggetti o in generale sullo steso fenomeno di interesse (ad es. il prezzo di un titolo azionario) fatte ad istanti temporali differenti. Un altro caso riguarda gli studi dove le unità statistiche fanno parte di gruppi noti a priori, all'interno dei quali si assume una certa dipendenza. In questo contesto sono due le principali classi di modelli a cui si fa riferimento: i **modelli marginali** ed i **modelli con effetti idividuali**. Nel caso di modelli marginali l'interesse non è quello di studiare effettivamente l'entità della correlazione della risposta, ma al contrario questa

viene trattata come una componente di disturbo nello studio dell'effetto delle variabili esplicative. Diversamente, per i modelli con effetti individuali si è interessati a studiare l'effettiva correlazione della risposta rispetto a caratteristiche, assunte presenti ma non osservabili, specifiche delle unità.

In questo contesto, la variabile risposta Y_i, $i = 1, \ldots, n$, è un vettore casuale $Y_i = (Y_{i1}, \ldots, Y_{im_i})^T$ il quale rappresenta le m_i osservazioni sulla risposta per l'i-esima unità. Quindi il numero totale di osservazioni è $N = \sum_{i=1}^{n} m_i$ e y_{ij}, $j = 1, \ldots, m_i$ indica l'osservazione j-esima rilevata rispetto all'i-esima unità. Infine, per tener conto della correlazione si assume che le variabili casuali Y_{ij} e Y_{ih}, $j \neq h$ siano correlate, mentre viene mantenuta l'ipotesi di indipendenza rispetto alle rilevazioni fatte su unità differenti.

1.5.1 Quasi-verosimiglianza

Prima di entrare nello specifico della trattazione dei modelli per risposte correlate, è di particolare importanza introdurre il concetto di quasi-verosimiglianza ed equazioni di stima non distorta. Come per il modello lineare normale, in cui si dimostra che anche sotto le più deboli ipotesi del secondo ordine lo stimatore $\hat{\beta}$ risulta asintoticamente normale ed efficiente secondo il teorema di Gauss-Markov, è possibile estendere questo approccio anche ai modelli lineari generalizzati.

Analogamente a quanto fatto nel caso del modello lineare normale, è quindi possibile definire un GLM sotto ipotesi più deboli delle (1.13)- (1.14), in particolare viene omessa l'ipotesi fatta sulla distribuzione di Y_i,

$$E[Y_i] = \mu(x_i\beta) = g^{-1}(x_i\beta), \tag{1.26}$$

$$Var(Y_i) = \phi v(\mu_i), \tag{1.27}$$

$$Y_i \text{ e } Y_j \quad \text{indipendenti con } i \neq j \tag{1.28}$$

dove $\phi > 0$ è un parametro ignoto, detto **parametro di dispersione**. L'idea alla base di tale sviluppo deriva dalla struttura delle equazioni di verosimiglianza per β in un GLM (1.21). Poiché queste dipendono solamente dai primi due momenti della distribuzione di Y_i, $E[Y_i] = \mu_i$ e $Var(Y_i) = \phi v(\mu_i)$, è possibile studiare le proprietà dello stimatore $\hat{\beta}$ sotto le ipotesi (1.26)- (1.28). Il modello semi-parametrico specificato dalle assunzioni (1.26)- (1.28) è detto **modello di quasi-verosimiglianza**.

Equazioni di stima non distorte e proprietà degli stimatori

Una delle principali proprietà delle equazioni di verosimiglianza (1.2) è la loro non distorsione. È quindi importante che tale proprietà rimanga valida anche per un modello di quasi-verosimiglianza.

Definita la funzione $q(y;\beta)$ come

$$q(y;\beta) = D^T V^{-1}(y-\mu), \tag{1.29}$$

dove

$$E_\beta[q(Y;\beta)] = 0, \quad \beta \in \mathbb{R}^p, \tag{1.30}$$

con $q(y;\beta) = ((q_1(y;\beta),\ldots,q_p(y;\beta)))$ dove

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

quindi $q(y;\beta)$ è il vettore delle derivate parziali della log-verosimiglianza nel caso di un GLM. La (1.30) definisce un'**equazione di stima non distorta**. Siano inoltre

$$J(\beta) = E_\beta[q(Y;\beta)q(Y;\beta)^T] \quad \text{e} \quad H(\beta) = -E_\beta\left[\frac{\partial q(Y;\beta)}{\partial \beta^T}\right].$$

Si dimostra che lo stimatore $\hat{\beta}$ definito come soluzione di $q(y;\beta) = 0$ è consistente (si veda Liang & Zeger, 1986). Inoltre sotto le ipotesi (1.26)- (1.28), vale l'dentità dell'informazione

$$E[l_r l_s] = -E[l_{rs}],$$

ossia in questo caso,

$$J(\beta) = H(\beta),$$

quindi anche sotto il modello di quasi-verosimiglianza valgono le (1.2) e (1.3).

Tramite approssimazione di Taylor al primo ordine di $q(y;\hat{\beta})$ e sfruttando la legge dei grandi numeri tale per cui

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

si ottiene

$$\hat{\beta} - \beta \dot{\sim} N_p(0, H(\beta)^{-1}J(\beta)H(\beta)^{-1}). \tag{1.31}$$

Infine essendo

$$H(\beta) = -\frac{\partial(D^T V^{-1})}{\partial \beta} E[Y - \mu_i] + D^T V^{-1} \frac{\partial \mu(\beta)}{\partial \beta^T} = D^T V^{-1} D$$

e

$$J(\beta) = Var(q(T;\beta)) = D^T V^{-1} Var(Y) V^{-1} D,$$

risulta che

$$Var(\hat{\beta}) \doteq (D^T V^{-1} D)^{-1} D^T V^{-1} Var(Y) V^{-1} D (D^T V^{-1} D)^{-1}.$$

Sotto l'assunzione fatta sulla varianza del modello di quasi-verosimiglainza, $Var(Y_i) = \phi(v(\mu_i))$, allora per l'identità dell'informazione $J(\beta) = H(\beta)$ si ottiene $Var(\hat{\beta}) = (D^T V^{-1} D)^{-1}$ e conseguentemente

$$\hat{\beta} \dot{\sim} N_p(\beta, (D^T V^{-1} D)^{-1}). \tag{1.32}$$

Tuttavia il risultato asintotico (1.32) potrebbe risultare distorto nel caso in cui venga meno la validità dell'assuzione sulla varianza della risposta Y_i. Si preferisce quindi utilizzare una stima robusta di $Var(Y_i)$, definita come $\widehat{Var}(Y_i) = \text{diag}[(y_i - \hat{\mu}_i)^2]$, ottendo così

$$\widehat{Var}_{\text{R}}(\hat{\beta}) = (D^T V^{-1} D)^{-1} D^T V^{-1} \text{diag}[(y_i - \hat{\mu}_i)^2] V^{-1} D (D^T V^{-1} D)^{-1}. \tag{1.33}$$

1.5.2 Modelli marginali

Notazione

Si introduce ora la struttura alla base della formulazione dei modelli per risposte multivariate. Il vettore

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

dove $Y_i = (Y_{i1}, \ldots, Y_{im_i})$, $i = 1, \ldots, n$, segue una distribuzione multivariata con media $\mu = (\mu_1^T, \ldots, \mu_n^T)^T$, con $\mu_i = X_i \beta$, dove X_i è la matrice del modello $m_i \times p$ per l'i-esima

unità, $\beta = (\beta_1, \ldots, \beta_p)^T$ è un vettore di parametri e V tale che

$$V = \begin{bmatrix} V_1 & 0 & \ldots & 0 \\ 0 & V_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & V_n \end{bmatrix},$$

è la matrice di covarianza. Seguendo la struttura di V si nota che rimane l'assunzione di indipendenza tra Y_i e Y_j, $i, j = 1, \ldots, n$ e $i \neq j$. Inoltre vale anche l'assunzione di omoschedasticità per cui $V_1 = V_2 = \cdots = V_n$. Denotiamo infine con X la matrice di dimensioni $N \times p$ che contiene le matrici del modello X_i per l'unità i-esima

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}.$$

Data la complessità della formulazione di un modello di regressione multivariata, spesso si ricorre a delle specificazioni più semplici che riducono il numero di parametri da stimare e semplificano la struttura della matrice V di covarianza. In particolare, assumendo per semplicità m_i costanti e pari ad m, la $Var(Y_{ij}) = \sigma^2$ e per quanto rigurda la struttura di correlazione del modello si utilizzano le seguenti strutture di riferimento: equicorrelazione, autoregressione e non strutturata.

- **Equicorrelazione** (Interscambiabile)
  In questo caso tutti gli elementi fuori dalla diagonale sono uguali e pari a ρ, denomiato come **coefficiente di correlazione intra-classe** e $\rho \in (-1/(m-1), 1)$,

$$V_i = \sigma^2 \begin{bmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \vdots & \vdots \\ \rho & \rho & \ldots & 1 \end{bmatrix}.$$

- **Autoregressione**
  Questo tipo di struttura è molto utile per lo studio di dati longitudinali, in cui è importante considerare che osservazioni vicine nel tempo siano maggiormente correlate rispetto ad osservazioni lontane. La forma più utilizzata è quella autoregressiva del primo ordine, dove $Cor(Y_{ij}, Y_{ik}) = \rho^{|j-k|}$, con $\rho \in (-1, 1)$,

quindi

$$V_i = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{m-1} \\ \rho & 1 & \rho & \dots & \rho^{m-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho^{m-1} & \dots & \dots & \rho & 1 \end{bmatrix}.$$

- **Non strutturata**

  In questo caso non viene assunta nessuna struttura per V_i, difatti la stima della matrice di covarianza risulta molto complessa e spesso può portare a problemi di convergenza delle stime nei casi in cui m sia prossimo ad n, la struttura è la seguente

$$V_i = \sigma^2 \begin{bmatrix} 1 & \rho_{12} & \dots & \rho_{1m} \\ \rho_{21} & 1 & \dots & \rho_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ \rho_{m1} & \rho_{m2} & \dots & 1 \end{bmatrix}.$$

Equazioni di stima generalizzate (GEE)

Sfruttando la teoria della quasi-verosimiglianza introdotta precedentemente, è possibile estendere la teoria dei GLM. In particolare Liang & Zeger (1986) proposero una nuova metodologia per lo studio di risposte discrete correlate, quali di conteggio o binarie. Questo perchè la modellazione effettuata sfruttando la famiglia di dispersione esponenziale risulta essere poco flessibile, portando spesso a fenomeni di sovradispersione (varianza stimata dal modello minore di quela osservata) o sottodispersione (varianza stimata dal modello superiore di quella osservata), dove il primo caso in particolare è di particolare rilevanza, soprattutto in contesti finanziari.

Assumendo per semplicità $m_i = m$ è possibile generalizzare le ipotesi del secondo ordine (1.26)- (1.28) nel seguente modo

$$E[Y_i] = \mu_i, \qquad \text{con} \qquad g(\mu_i) = x_{ij}\beta, \tag{1.34}$$

$$Cov(Y_i) = V_i = \phi A_i^{1/2} R(\alpha) A_i^{1/2}, \tag{1.35}$$

$$Y_i \,\text{e}\, Y_h \quad \text{indipendenti con} \, i \neq h. \tag{1.36}$$

dove $A_i = \text{diag}(v(\mu_{ij}))$, con μ_{ij} generico elemento del vettore μ_i, $R(\alpha)$ matrice di correlazione di Y_i e ϕ parametro di dispersione positivo. Sfruttando la rappresentazione matriciale è possible scrivere le equazioni di stima generalizzate come

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

che sfruttando la partizione a blocchi delle matrici e le relative proprietà algebriche, equivale a

$$\sum_{i=1}^{n} D_i^T V_i^{-1}(y_i - \mu_i) = 0, \tag{1.37}$$

con D_i matrice $m \times p$ con generico elemento $\partial\mu_{ij}/\partial\beta_r$, $j = 1, \ldots, m$, $r = 1, \ldots, p$.

Infine si dimostra che sotto condizioni di regolarità, lo stimatore $\hat{\beta}$ ottenuto come soluzione delle (1.37) è asintoticamente normale con matrice di covarianza

$$V_{LZ} = Var(\hat{\beta})_{\text{LZ}} = \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i\right)^{-1} M_{LZ} \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i\right)^{-1}, \tag{1.38}$$

con

$$M_{LZ} = \left(\sum_{i=1}^{n} D_i^T V_i^{-1} Cov\big(Y_i\big) V_i^{-1} D_i\right).$$

LZ indica le iniziali di Liang e Zeger i quali proposero tale stimatore. Una stima robusta di $Var(\hat{\beta})$ si ottiene sostituendo le stime $\hat{\alpha}$ e $\hat{\phi}$ ottenute con il metodo dei momenti e stimando $Cov(Y_i)$ tramite $\hat{r}_i \hat{r}_i^T$, dove $\hat{r}_i = (y_i - \hat{\mu}_i)$. Per semplicità di notazione, successivamente si indicherà con V_{LZ} o M_{LZ} direttamente lo stimatore della matrice di covarianza denominato anche **stimatore sandwich**.

1.5.3 Modelli con effetti casuali

Nell'ambito dei modelli con effetti casuali è di particolare importanza distinguere due casi: i mdelli con effetti casuali per risposte normali ed i modelli con effetti casuali per risposte non normali.

Modelli con effetti casuali per risposte normali

Come sottolineato precedentemente, nel caso di modelli con effetti individuali è importante distinguere due tipi di effetti sulla variabile risposta. La distinzione che viene fatta è tra effetti fra le unità ed effetti entro le unità, dove entrambi possono essere casuali o fissi. Inoltre il primo tipo di effetto è di particolare rilevanza in quanto permette di indurre correlazione tra osservazione sulla stessa unità i-esima.

La formulazione generale di un **modello lineare normale con effetti misti** è la seguente

$$Y_{ij} = x_{ij}\beta + z_{ij}u_i + \varepsilon_{ij} \tag{1.39}$$

con $u_i \sim N_q(0, \Sigma_u)$ vettore q-dimensionale di effetti casuali, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ indipendente da u_i. Il termine u_i rappresenta l'effetto casuale dell i-esimo soggetto e combinato con il

vettore z_{ij} permette di esprimere le differenze specifiche di ogni unità, mentre ε_{ij} esprime la variabilità entro le unità. Infine, il vettore di parametri β combinato con il vettore x_{ij}, contenente le variabili esplicative ed eventuale intercetta, permette di esprimere effetti fissi fra le unità ed entro le unità.

Modelli con effetti casuali per risposte non normali

Nel caso in cui la variabile risposta non sia continua ma dicotomica o di conteggio, è possibile generalizzare il modello lineare normale con effetti misti al **modello lineare generalizzato con effetti misti (GLMM)** (*generalized linear mixed effects model*). Si assume che, condizionantamente ad u_i, le osservazioni sulla risposta Y_{ij} siano indipendenti e distribuite secondo un modello lineare generalizzato con

$$g(E[Y_{ij}|u_i]) = x_{ij}\beta + z_{ij}u_i, \tag{1.40}$$

dove $g(\cdot)$ è la funzioe di legame di un GLM e $u_i \sim N_q(0, \Sigma_u)$ vettore q-dimensionale di effetti casuali. Infine per i vettori x_{ij} e z_{ij} valgono le considerazioni fatte per il modello lineare normale in equazione (1.39).

Capitolo 2

Stimatori della matrice di covarianza per equazioni di stima generalizzate (GEE) con campioni di numerosità piccola

Introduzione

In questo capitolo vengono descritte le proprietà teoriche ed i concetti alla base degli stimatori della matrice di covarianza della SMV del parametro di regressione β p-dimensionale. Si riassumono inoltre i risultati disponibili relativamente alle proprietà teoriche in termini di efficienza, in base a proprietà asintotiche. Infine, si illustrerà un'approssimazione t di Student per la statistica di Wald su una singola componente di β, con gradi di libertà proporzionali al rapporto tra il valore atteso e la varianza dello stimatore della varianza di $\hat{\beta}$ (v. formula (2.16)). I test corrispondenti sono alla base dello studio di simulazione affrontato nel capitolo 3.

2.1 Modifiche dello stimatore sandwich

Uno dei motivi principali per cui V_{LZ} risulta inefficiente è dovuto al fatto che lo stimatore $\hat{r}\hat{r}^T$ è distorto per la stima di $E[(Y_i - \mu_i)^T(Y_i - \mu_i)]$. Questo accade specialmente in campioni di numerosità piccola dove le stime $\hat{\mu}_i$, tendono a "seguire" molto di più le osservazioni y_i. La conseguenza è quindi di sottostimare la matrice di covarianza della risposta Y e quindi la matrice di covarianza di $\hat{\beta}$, oltre che aumentare la variabilità

dello stimatore stesso. Segue un intervallo di confidenza per una generica componente di β più "stretto" e relativo test di ipotesi per la verifica della nullità più "liberale", ovvero l'ipotesi nulla viene rifiutata più spesso rispetto al livello nominale scelto a priori. Recentemente, sono state presentate diverse modifiche allo stimatore V_{LZ}. In questa sede se ne presentano alcune, le quali sono riassunte nella tabella 2.1.

TABELLA 2.1: Sintesi di otto modifiche dello stimatore V_{LZ} della matrice di covarianza di $\hat{\beta}$ nel caso di equazioni di stima generalizzate con piccoli campioni.

Stimatori	Modifiche	Riferimenti
V_{MK}	Aggiustamento per i gradi di libertà	MacKinnon (1985)
V_{KC}	Correzione della distorsione	Kauermann & Carroll (2001)
V_{PAN}	Aumento dell'efficienza	Pan (2001)
V_{GST}	Aumento dell'efficienza	Gosho et al. (2014)
V_{MD}	Correzione della distorsione	Mancl & DeRouen (2001)
V_{FG}	Correzione della distorsione	Fay & Graubard (2001)
V_{MBN}	Correzione della distorsione	Morel et al. (2003)
V_{WL}	Correzione della distorsione e aumento dell'efficienza	Wang & Long (2011)

Di seguito vengono descritte più nel dettaglio le varie proposte presentate: La notazione è la stessa utilizzata nella (1.38).

$\mathbf{V_{MK}}$ è lo stimatore della varianza "sandwich" corretto per i gradi di libertà proposto da MacKinnon (1985). Questo stimatore adotta un aggiustamento di un fattore pari a $\frac{n}{n-p}$, come segue

$$V_{MK} = \frac{n}{n-p} V_{LZ}, \tag{2.1}$$

dove, quando $n \to \infty$, V_{MK} è equivalente a V_{LZ}. V_{MK} riduce la distorsione, ma contemporaneamente incrementa la variabilità.

$\mathbf{V_{KC}}$ è uno stimatore "sandwich" corretto per la distorsione sotto l'assunzione di corretta specificazione della struttura di correlazione, proposto da Kauermann & Carroll (2001), definito come

$$V_{KC} = \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i \right)^{-1} M_{KC} \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i \right)^{-1} \tag{2.2}$$

con

$$M_{KC} = \sum_{i=1}^{n} D_i^T V_i^{-1} \big(I - H_{ii}\big)^{-1/2} \hat{r}_i \hat{r}_i^T \big(I - H_{ii}\big)^{-1/2} V_i^T D_i, \tag{2.3}$$

dove I è una matrice identità $m \times m$ e H_{ii} è una matrice diagonale che rappresenta il contributo dell'i-esima unità alla matrice di covarianza, la quale può essere calcolata come segue

$$H_{ii} = D_i \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i \right) D_i^T V_i^{-1}.$$

$\mathbf{V_{PAN}}$ questa modifica è stata proposta da Pan (2001) ed è valida sotto due assunzioni:

(A1) la varianza condizionata di Y_{ij} dato X_{ij} è correttamente specificata;

(A2) esiste una struttura di correlazione R_c comune a tutte le unità.

Lo stimatore proposto è

$$V_{PAN} = \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i \right)^{-1} M_{PAN} \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i \right)^{-1}, \tag{2.4}$$

con

$$M_{PAN} = \sum_{i=1}^{n} D_i^T V_i^{-1} \left\{ A_i^{1/2} \left(\frac{1}{n} \sum_{j=1}^{n} A_j^{-1/2} \hat{r}_j \hat{r}_j^T A_j^{-1/2} \right) A_j^{1/2} \right\} V_i^{-1} D_i. \tag{2.5}$$

Praticamente, V_{PAN} sfrutta tutta l'informazione disponibile per ogni unità per la stima di $Cov(Y_i)$, portando a migliori prestazioni in termini di efficienza.

$\mathbf{V_{GST}}$ è una modifica dello stimatore V_{PAN} che considera la distorsione di $A_i^{1/2} \left(\frac{1}{n} \sum_{i=1}^{n} A_i^{-1/2} \hat{r}_i \hat{r}_i^T A_i^{-1/2} \right) A_i^{1/2}$ con piccoli campioni, proposto da Gosho et al. (2014). Lo stimatore è il seguente

$$V_{GST} = \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i \right)^{-1} M_{GST} \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i \right)^{-1}, \tag{2.6}$$

dove

$$M_{GST} = \sum_{i=1}^{n} D_i^T V_i^{-1} \left\{ A_i^{1/2} \left(\frac{1}{n-p} \sum_{j=1}^{n} A_j^{-1/2} \hat{r}_j \hat{r}_j^T A_j^{-1/2} \right) A_i^{1/2} \right\} V_i^{-1} D_i. \tag{2.7}$$

Quindi anche V_{GST} sfrutta tutta l'informazione disponibile per ogni unità per la stima di $Cov(Y_i)$. Inoltre quando $n >> p$ e n è abbastanza grande, V_{GST} è approssimativamente uguale a V_{PAN}.

$\mathbf{V_{MD}}$ è un altro stimatore "sandwich" corretto per la distorsione proposto da Mancl & DeRouen (2001). A differenza di V_{KC}, lo stimatore non assume una corretta specificazione della struttura di correlazione ed è specificato come segue

$$V_{MD} = \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i\right)^{-1} M_{MD} \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i\right)^{-1}, \tag{2.8}$$

dove

$$M_{MD} = \sum_{i=1}^{n} D_i^T V_i^{-1} \big(I_i - H_{ii}\big)^{-1} \hat{r}_i \hat{r}_i^T \big(I_i - H_{ii}\big)^{-1} V_i^T D_i, \tag{2.9}$$

dove I_i e H_{ii} sono definite come per lo stimatore V_{KC}. Si fa notare inoltre che per correggere la distorsione è stato fatto riferimento alla seguente approssimazione in serie di Taylor

$$E\big(\hat{r}_i \hat{r}_i^T\big) \approx \big(I_i - H_{ii}\big) Cov(Y_i) \big(I_i - H_{ii}\big)^T,$$

dove è stato però ignorato il termine $\sum_{j \neq i} H_{ij} Cov(Y_i) H_{ij}^T$, apportando così una sovracorrezione.

$\mathbf{V_{FG}}$ è lo stimatore della varianza proposto da Fay & Graubard (2001), caratterizzato da un aggiustamento per un fattore di scala, lo stimatore è il seguente

$$V_{FG} = \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i\right)^{-1} M_{FG} \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i\right)^{-1}, \tag{2.10}$$

dove

$$M_{FG} = \sum_{i=1}^{n} \eta_i^{-1} D_i^T V_i^{-1} \hat{r}_i \hat{r}_i^T V_i^T D_i \eta_i^{T-1}, \tag{2.11}$$

dove $\eta_i = I_p - N_i$, con $N_i = D_i^T V_i^{-1} D_i \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i\right)^{-1}$.

$\mathbf{V_{MBN}}$ è uno stimatore corretto per la distorsione raccomandato da Morel et al. (2003), il quale incorpora la correlazione dei residui e la numerosità campionaria, lo stimatore è il seguente

$$V_{MBN} = \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i\right)^{-1} M_{MBN} \left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i\right)^{-1}, \tag{2.12}$$

dove

$$M_{MBN} = \sum_{i=1}^{n} D_i^T V_i^{-1} \big(k \hat{r}_i \hat{r}_i^T + \delta_m \xi V_i\big) V_i^T D_i, \tag{2.13}$$

dove $k = \frac{N-1}{N-p}\frac{n}{n-1}$, $\delta_m = \begin{cases} \frac{p}{n-p} & n > (d+1)p \\ \frac{1}{p} & \text{altrimenti} \end{cases}$, e $\xi = \max\left(r, \frac{trace\left\{\left(\sum_{i=1}^n D_i^T V_i^{-1} D_i\right)^{-1} M_{LZ}\right\}}{p}\right)$ con $0 \leq r \leq 1$. Si sottolinea che il fattore k permette di correggere la distorsione dello stimatore empirico di $Cov(Y_i)$ e che δ_m può essere delimitato da $1/d$. I valori di default per d e r sono rispettivamente 2 e 1 come riportato in Morel et al. (2003).

$\mathbf{V_{WL}}$ è uno stimatore della varianza proposto da Wang & Long (2011), il quale considera i punti di forza dei due stimatori V_{PAN} e V_{MD}, dove il primo sfrutta tutta l'informazione disponibile per ogni unità per la stima di $Cov(Y_i)$, portando a migliori prestazioni in termini di efficienza, mentre il secondo riduce la distorsione dello stimatore $\hat{r}_i\hat{r}_i^T$. Lo stimatore è il seguente

$$V_{WL} = \left(\sum_{i=1}^n D_i^T V_i^{-1} D_i\right)^{-1} M_{WL} \left(\sum_{i=1}^n D_i^T V_i^{-1} D_i\right)^{-1}, \tag{2.14}$$

dove

$$M_{WL} = \sum_{i=1}^n D_i^T V_i^{-1} A_i^{1/2} \left\{\sum_{i=1}^n A_i^{-1/2}\left(I_i - H_{ii}\right)^{-1} \hat{r}_i\hat{r}_i^T \left(I_i - H_{ii}^T\right)^{-1} A_i^{-1/2}/n\right\} A_i^{1/2} V_i^T D_i. \tag{2.15}$$

Ci si attende che questo stimatore sia più accurato degli stimatori V_{PAN} e V_{MD}, ma devono comunque valere le assunzioni (A1) ed (A2) come per lo stimatore V_{PAN}.

Si procede ora con il confronto teorico tra i diversi stimatori della matrice di covarianza. La notazione utilizzata per presentare gli stimatori permetterà di focalizzare l'attenzione solamente sulla matrice M. Infatti, l'unico fattore di diversità tra le varie modifiche allo stimatore V_{LZ}, dipende dalla struttura della matrice M. Dovendo quindi calcolare la varianza dei vari stimatori per la matrice di covarianza di $\hat{\beta}$, è importante introdurre due operatori matematici che permetteranno di effettuare il confronto in maniera semplice e lineare:

vec$(\cdot)$ è l'operatore che denota l'operazione di vettorizzazione, la quale consiste in una trasformazione lineare che permette di convertire una matrice in un vettore colonna, ad esempio: se A è una matrice 2×2 definita come

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

allora l'operazione di vettorizzazione permette di ottere il seguente vettore colonna di dimensioni 4×1

$$vec(A) = \begin{bmatrix} a \\ c \\ b \\ d \end{bmatrix};$$

$\bigotimes$ è l'operatore che denota l'operazione prodotto di Kronecker, la quale è sempre applicabile indipendentemente dalle dimensioni delle matrici coinvolte nell'operazione. In pratica se A è una matrice $m \times n$ e B è una matrice $p \times q$ allora il prodotto di Kronecker $A \otimes B$ è una matrice $mp \times nq$ definita a blocchi come segue

$$A \otimes B = \begin{bmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \dots & a_{mn}B \end{bmatrix}.$$

Si sottolineano inoltre due importanti proprietà, una dell'operazione di vettorizzazione ed una del prodotto di Kronecker, che risulteranno ultili nei calcoli successivi. La prima sfrutta il prodotto di Kronecker ed è definita come segue: siano A, B e C tre matrici di dimensioni rispettivamente $k \times l$, $l \times m$ e $m \times n$, allora

$$vec(ABC) = \big(C^T \otimes A\big)vec(B).$$

La seconda invece: siano A e B due matrici invertibili allora anche il prodotto di Kronecker tra queste due matrici è invertibile, e si ha

$$(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}.$$

Calcolo della matrice di covarianza di vec(M)

Si riportano ora i risultati ed i calcoli intermendi relativi alla matrice di covarianza di $vec(M)$, per ognuna delle otto proposte più lo stimatore "sandwich" V_{LZ}:

$\mathbf{M_{LZ}}$ si procede inizialmente con la vettorizzazione della matrice M_{LZ},

$$vec(M_{LZ}) = vec\left(\sum_{i=1}^{n} D_i^T V_i^{-1} \hat{r}_i \hat{r}_i^T V_i^{-1} D_i\right),$$

dove sfruttando la proprietà di linearità dell'operatore $vec(\cdot)$ si ottiene

$$\begin{aligned} vec(M_{LZ}) &= \sum_{i=1}^{n} vec\Big(D_i^T V_i^{-1} \hat{r}_i \hat{r}_i^T V_i^{-1} D_i \Big) \\ &= \sum_{i=1}^{n} \underbrace{D_i^T V_i^{-1} \otimes D_i^T V_i^{-1}}_{S_i} vec(\hat{r}_i \hat{r}_i^T) \\ &= \sum_{i=1}^{n} S_i vec(\hat{r}_i \hat{r}_i^T). \end{aligned}$$

Si procede ora con il calcolo della covarianza di $vec(M_{LZ})$,

$$Cov(vec(M_{LZ})) = Cov\left(\sum_{i=1}^{n} S_i vec(\hat{r}_i \hat{r}_i^T) \right),$$

sfruttando l'indipendeza tra le unità si ottiene

$$\begin{aligned} Cov(vec(M_{LZ})) &= \sum_{i=1}^{n} Cov\Big(S_i vec(\hat{r}_i \hat{r}_i^T) \Big) \\ &= \sum_{i=1}^{n} S_i \underbrace{Cov(vec(\hat{r}_i \hat{r}_i^T))}_{T_i} S_i^T \\ &= \sum_{i=1}^{n} S_i T_i S_i^T; \end{aligned}$$

$\mathbf{M_{MK}}$ si procede con la vettorizzazione della matrice M_{MK},

$$\begin{aligned} vec(M_{MK}) &= \frac{n}{(n-p)} vec\left(\sum_{i=1}^{n} D_i^T V_i^{-1} \hat{r}_i \hat{r}_i^T V_i^{-1} D_i \right) \\ &= \frac{n}{(n-p)} \sum_{i=1}^{n} S_i vec(\hat{r}_i \hat{r}_i^T), \end{aligned}$$

da cui si ottiene

$$\begin{aligned} Cov(vec(M_{MK})) &= Cov\left(\frac{n}{(n-p)} \sum_{i=1}^{n} S_i vec(\hat{r}_i \hat{r}_i^T) \right) \\ &= \frac{n^2}{(n-p)^2} \sum_{i=1}^{n} Cov(S_i vec(\hat{r}_i \hat{r}_i^T)) \\ &= \frac{n^2}{(n-p)^2} \sum_{i=1}^{n} S_i T_i S_i^T; \end{aligned}$$

$\mathbf{M_{KC}}$ si procede con la vettorizzazione della matrice M_{KC},

$$\begin{aligned}
vec(M_{KC}) &= vec\Big(\sum_{i=1}^{n} D_i^T V_i^{-1}\big(I_i - H_{ii}\big)^{-1/2}\hat{r}_i\hat{r}_i^T\big(I_i - H_{ii}\big)^{-1/2}V_i^T D_i\Big)\\
&= \sum_{i=1}^{n} vec\Big(D_i^T V_i^{-1}\big(I_i - H_{ii}\big)^{-1/2}\hat{r}_i\hat{r}_i^T\big(I_i - H_{ii}\big)^{-1/2}V_i^T D_i\Big)\\
&= \sum_{i=1}^{n} S_i vec(\big(I_i - H_{ii}\big)^{-1/2}\hat{r}_i\hat{r}_i^T\big(I_i - H_{ii}\big)^{-1/2})\\
&= \sum_{i=1}^{n} S_i\big((I_i - H_{ii})^{-1/2}\big)^T \otimes (I_i - H_{ii})^{-1/2} vec(\hat{r}_i\hat{r}_i^T)\\
&= \sum_{i=1}^{n} S_i \underbrace{(I_i - H_{ii})^{-1/2} \otimes (I_i - H_{ii})^{-1/2}}_{F_i} vec(\hat{r}_i\hat{r}_i^T)\\
&= \sum_{i=1}^{n} S_i F_i vec(\hat{r}_i\hat{r}_i^T),
\end{aligned}$$

da cui si ottiene

$$\begin{aligned}
Cov(vec(M_{KC})) &= \sum_{i=1}^{n} Cov(S_i F_i vec(\hat{r}_i\hat{r}_i^T))\\
&= \sum_{i=1}^{n} S_i F_i T_i F_i^T S_i^T;
\end{aligned}$$

$\mathbf{M_{PAN}}$ si procede con la vettorizzazione della matrice M_{PAN},

$$\begin{aligned}
vec(M_{PAN}) &= \sum_{i=1}^{n} vec\left(D_i^T V_i^{-1}\left\{A_i^{1/2}\left(\frac{1}{n}\sum_{j=1}^{n} A_j^{-1/2}\hat{r}_j\hat{r}_j^T A_j^{-1/2}\right)A_i^{1/2}\right\}V_i^{-1}D_i\right)\\
&= \sum_{i=1}^{n} S_i vec\left(A_i^{1/2}\left(\frac{1}{n}\sum_{j=1}^{n} A_j^{-1/2}\hat{r}_j\hat{r}_j^T A_j^{-1/2}\right)A_i^{1/2}\right)\\
&= \sum_{i=1}^{n} S_i \underbrace{A_i^{1/2} \otimes A_i^{1/2}}_{E_i} vec\left(\frac{1}{n}\sum_{j=1}^{n} A_j^{-1/2}\hat{r}_j\hat{r}_j^T A_j^{-1/2}\right)\\
&= \sum_{i=1}^{n} S_i E_i \frac{1}{n}\sum_{j=1}^{n} A_j^{-1/2} \otimes A_j^{-1/2} vec(\hat{r}_j\hat{r}_j^T)\\
&= \sum_{i=1}^{n} S_i E_i \frac{1}{n}\sum_{j=1}^{n} E_j^{-1} vec(\hat{r}_j\hat{r}_j^T)
\end{aligned}$$

da cui si ottiene

$$\begin{aligned} Cov(vec(M_{PAN})) &= \sum_{i=1}^{n} Cov(S_i E_i \frac{1}{n} \sum_{j=1}^{n} E_j^{-1} vec(\hat{r}_j \hat{r}_j^T)) \\ &= \sum_{i=1}^{n} S_i E_i \frac{1}{n^2} Cov\left(\sum_{j=1}^{n} E_j^{-1} vec(\hat{r}_j \hat{r}_j^T) \right) E_i^T S_i^T \\ &= \sum_{i=1}^{n} S_i E_i \frac{1}{n^2} \left(\sum_{j=1}^{n} E_j^{-1} T_j (E_j^{-1})^T \right) E_i^T S_i^T \\ &= \sum_{i=1}^{n} S_i \left[E_i \frac{1}{n^2} \left(\sum_{j=1}^{n} E_j^{-1} T_j E_j^{-1} \right) E_i \right] S_i^T; \end{aligned}$$

$\mathbf{M_{GST}}$ si procede con la vettorizzazione della matrice M_{GST},

$$\begin{aligned} vec(M_{GST}) &= \sum_{i=1}^{n} vec\left(D_i^T V_i^{-1} \left\{ A_i^{1/2} \left(\frac{1}{n} \sum_{j=1}^{n} A_j^{-1/2} \hat{r}_j \hat{r}_j^T A_j^{1/2} \right) A_i^{-1/2} \right\} V_i^{-1} D_i \right) \\ &= \sum_{i=1}^{n} S_i vec\left(A_i^{1/2} \left(\frac{1}{n} \sum_{j=1}^{n} A_j^{1/2} \hat{r}_j \hat{r}_j^T A_j^{-1/2} \right) A_i^{-1/2} \right) \\ &= \sum_{i=1}^{n} S_i \underbrace{A_i^{1/2} \otimes A_i^{1/2}}_{E_i} vec\left(\frac{1}{n} \sum_{j=1}^{n} A_j^{-1/2} \hat{r}_j \hat{r}_j^T A_j^{-1/2} \right) \\ &= \sum_{i=1}^{n} S_i E_i \frac{1}{n} \sum_{j=1}^{n} A_j^{-1/2} \otimes A_j^{-1/2} vec(\hat{r}_j \hat{r}_j^T) \\ &= \sum_{i=1}^{n} S_i E_i \frac{1}{n} \sum_{j=1}^{n} E_j^{-1} vec(\hat{r}_j \hat{r}_j^T) \end{aligned}$$

da cui si ottiene

$$\begin{aligned} Cov(vec(M_{GST})) &= \sum_{i=1}^{n} Cov(S_i E_i \frac{1}{n} \sum_{j=1}^{n} E_j^{-1} vec(\hat{r}_j \hat{r}_j^T)) \\ &= \sum_{i=1}^{n} S_i E_i \frac{1}{n^2} Cov\left(\sum_{j=1}^{n} E_j^{-1} vec(\hat{r}_j \hat{r}_j^T) \right) E_i^T S_i^T \\ &= \sum_{i=1}^{n} S_i E_i \frac{1}{n^2} \left(\sum_{j=1}^{n} E_j^{-1} T_j (E_j^{-1})^T \right) E_i^T S_i^T \\ &= \sum_{i=1}^{n} S_i \left[E_i \frac{1}{n^2} \left(\sum_{j=1}^{n} E_j^{-1} T_j E_j^{-1} \right) E_i \right] S_i^T; \end{aligned}$$

$\mathbf{M_{MD}}$ si procede con la vettorizzazione della matrice M_{MD},

$$\begin{aligned} vec(M_{MD}) &= \sum_{i=1}^{n} vec\left(D_i^T V_i^{-1} \left(I_i - H_{ii}\right)^{-1} \hat{r}_i \hat{r}_i^T \left(I_i - H_{ii}\right)^{-1} V_i^T D_i \right) \\ &= \sum_{i=1}^{n} S_i vec\left(\left(I_i - H_{ii}\right)^{-1} \hat{r}_i \hat{r}_i^T \left(I_i - H_{ii}\right)^{-1} \right) \\ &= \sum_{i=1}^{n} S_i \underbrace{\left(I_i - H_{ii}\right)^{-1} \otimes \left(I_i - H_{ii}\right)^{-1}}_{G_i} vec(\hat{r}_i \hat{r}_i^T) \\ &= \sum_{i=1}^{n} S_i G_i vec(\hat{r}_i \hat{r}_i^T) \end{aligned}$$

da cui si ottiene

$$\begin{aligned} Cov(vec(M_{MD})) &= \sum_{i=1}^{n} Cov(S_i G_i vec(\hat{r}_i \hat{r}_i^T)) \\ &= \sum_{i=1}^{n} S_i G_i T_i G_i^T S_i^T; \end{aligned}$$

$\mathbf{M_{FG}}$ si procede con la vettorizzazione della matrice M_{FG},

$$\begin{aligned} vec(M_{FG}) &= \sum_{i=1}^{n} vec\left(\eta_i^{-1} D_i^T V_i^{-1} \hat{r}_i \hat{r}_i^T V_i^T D_i \eta_i^{T-1} \right) \\ &= \sum_{i=1}^{n} \underbrace{\left(\eta_i^{-1} D_i^T V_i^{-1}\right) \otimes \left(\eta_i^{-1} D_i^T V_i^{-1}\right)}_{H_i} vec(\hat{r}_i \hat{r}_i^T) \\ &= \sum_{i=1}^{n} H_i vec(\hat{r}_i \hat{r}_i^T) \end{aligned}$$

da cui si ottiene

$$\begin{aligned} Cov(vec(M_{FG})) &= \sum_{i=1}^{n} Cov(H_i vec(\hat{r}_i \hat{r}_i^T)) \\ &= \sum_{i=1}^{n} H_i T_i H_i^T; \end{aligned}$$

$\mathbf{M_{MBN}}$ si procede con la vettorizzazione della matrice M_{MBN},

$$\begin{aligned} vec(M_{MBN}) &= \sum_{i=1}^{n} vec\left(D_i^T V_i^{-1} \big(k\hat{r}_i\hat{r}_i^T + \delta_m \xi V_i\big) V_i^T D_i \right) \\ &= \sum_{i=1}^{n} S_i vec(k\hat{r}_i\hat{r}_i^T + \delta_m \xi V_i) \\ &= \sum_{i=1}^{n} S_i \underbrace{\big(kvec(\hat{r}_i\hat{r}_i^T) + vec(\delta_m \xi V_i)\big)}_{N_i} \end{aligned}$$

da cui si ottiene

$$Cov(vec(M_{MBN})) = \sum_{i=1}^{n} S_i N_i S_i^T;$$

$\mathbf{M_{WL}}$ si procede con la vettorizzazione della matrice M_{WL},

$$\begin{aligned} vec(M_{WL}) = \sum_{i=1}^{n} vec\Bigg(& D_i^T V_i^{-1} A_i^{1/2} \left\{ \sum_{j=1}^{n} A_j^{-1/2} \big(I_j - H_{jj}\big)^{-1} \hat{r}_j \hat{r}_j^T \big(I_j - H_{jj}^T\big)^{-1} A_j^{-1/2} / n \right\} \\ & A_i^{1/2} V_i^T D_i \Bigg) \\ = \sum_{i=1}^{n} S_i E_i \frac{1}{n} \sum_{j=1}^{n} E_j^{-1} G_j vec(\hat{r}_j \hat{r}_j^T) \end{aligned}$$

da cui si ottiene

$$\begin{aligned} Cov(vec(M_{WL})) &= \sum_{i=1}^{n} Cov\left(S_i E_i \frac{1}{n} \sum_{j=1}^{n} E_j^{-1} G_j vec(\hat{r}_j \hat{r}_j^T) \right) \\ &= \sum_{i=1}^{n} \left[E_i Cov\left(\sum_{j=1}^{n} \frac{1}{n} E_j^{-1} G_j vec(\hat{r}_j \hat{r}_j^T) \right) E_i \right] S_i^T \\ &= \sum_{i=1}^{n} S_i \left[E_i \frac{1}{n^2} \left(\sum_{j=1}^{n} E_j^{-1} G_j T_j G_j^T E_j^{-1} \right) E_i \right] S_i^T. \end{aligned}$$

Nella Tabella 2.2, vengono riassunti i risultati ottenuti precedentemente in corrispondenza di ogni stimatore proposto, dove $T = Cov(vec(\hat{r}_i\hat{r}_i^T))$; $S_i = (D_i^T V_i^{-1}) \otimes (D_i^T V_i^{-1})$; $F_i = (I_i - H_{ii})^{-\frac{1}{2}} \otimes (I_i - H_{ii})^{-\frac{1}{2}}$; $G_i = (I_i - H_{ii})^{-1} \otimes (I_i - H_{ii})^{-1}$; $E_i = A_i^{\frac{1}{2}} \otimes A_i^{\frac{1}{2}}$; $H_i = (\eta_i^{-1} D_i^T V_i^{-1}) \otimes (\eta_i^{-1} D_i^T V_i^{-1})$; $N_i = Cov(kvec(\hat{r}_i\hat{r}_i^T) + vec(\delta_m \xi V_i))$.

TABELLA 2.2: Matrici di covarianza della matrice M per nove stimatori della varianza nel caso di equazioni di stima generalizzate.

Matrice M	Matrice di covarianza di $vec(M)$
M_{LZ}	$\sum_{i=1}^{n} S_i T_i S_i^T$
M_{MK}	$\sum_{i=1}^{n} \frac{n^2}{(n-p)^2} S_i T_i S_i^T$
M_{KC}	$\sum_{i=1}^{n} S_i F_i T_i F_i^T S_i^T$
M_{PAN}	$\sum_{i=1}^{n} S_i \Big[E_i \Big(\sum_{j=1}^{n} \frac{1}{n^2} E_j^{-1} T_j E_j^{-1} \Big) E_i \Big] S_i^T$
M_{GST}	$\sum_{i=1}^{n} S_i \Big[E_i \Big(\sum_{j=1}^{n} \frac{1}{(n-p)^2} E_j^{-1} T_j E_j^{-1} \Big) E_i \Big] S_i^T$
M_{MD}	$\sum_{i=1}^{n} S_i G_i T_i G_i^T S_i^T$
M_{FG}	$\sum_{i=1}^{n} H_i T_i H_i^T$
M_{MBN}	$\sum_{i=1}^{n} S_i N_i S_i^T$
M_{WL}	$\sum_{i=1}^{n} S_i \Big[E_i \Big(\sum_{j=1}^{n} \frac{1}{n^2} E_j^{-1} G_j T_j G_j^T E_j^{-1} \Big) E_i \Big] S_i^T$

2.2 Valutazione teorica dell'efficienza degli stimatori proposti

Si riportano ora, i confronti in termini di efficienza, di ognuna delle otto modifiche allo stimatore "sandwich". Seguendo le dimostrazioni di Wang & Long (2011) (si veda l'appendice B), si ottiene che $Cov(vec(M_{LZ})) - Cov(vec(M_{WL}))$ e $Cov(vec(M_{MD})) - Cov(vec(M_{WL}))$ sono definite non negative con probabilità 1, mentre $Cov(vec(M_{PAN})) - Cov(vec(M_{WL}))$ converge a 0 con probabilità 1 per $n \to \infty$. Analoghi risultati valgono per le rimanenti alternative, dove

$$Cov(vec(M_{LZ})) - Cov(vec(M_{MK})) = \sum_{i=1}^{n} \left(1 - \frac{n^2}{(n-p)^2} \right) S_i T_i S_i^T,$$

$$Cov(vec(M_{LZ})) - Cov(vec(M_{KC})) = \sum_{i=1}^{n} S_i \Big(T_i - F_i T_i F_i^T \Big) S_i^T,$$

$$Cov(vec(M_{LZ})) - Cov(vec(M_{GST})) = \sum_{i=1}^{n} S_i \left(T_i - E_i \left(\sum_{j=1}^{n} \frac{1}{(n-p)^2} E_j^{-1} T_j E_j^{-1} \right) E_i \right) S_i^T,$$

$$Cov(vec(M_{LZ})) - Cov(vec(M_{FG})) = \sum_{i=1}^{n} \Big(1 - \eta^{-1} \otimes \eta^{-1} \Big) S_i T_i S_i^T \Big(1 - \eta_i^{T^{-1}} \otimes \eta_i^{T^{-1}} \Big),$$

$$Cov(vec(M_{LZ})) - Cov(vec(M_{MBN})) = \sum_{i=1}^{n} S_i \Big(T_i - N_i \Big) S_i^T.$$

Wang et al. (2016a) mostrano che sotto deboli condizoini di regolarità, $Cov(vec(M_{LZ}))-Cov(vec(M_{MK}))$, $Cov(vec(M_{LZ}))-Cov(vec(M_{KC}))$, $Cov(vec(M_{LZ}))-Cov(vec(M_{FG}))$, e $Cov(vec(M_{LZ}))-Cov(vec(M_{MBN}))$ convergono a 0 con probabilità 1 per $n \to \infty$, mentre $Cov(vec(M_{LZ}))-Cov(vec(M_{GST}))$ è una matrice definita non negativa con probabilità 1 per $n \to \infty$. Perciò asintoticamente gli stimatori sono equivalenti, ma ciò che interessa in questo contesto è valutarne l'efficienza con piccoli campioni e si osserva che tutte le nuove proposte hanno caratteristiche migliori dello stimtore V_{LZ}.

2.3 Proprietà asintotiche dei test di ipotesi

Nel caso di GEE, uno dei test più utilizzati per le procedure di verifica di ipotesi, è il test di Wald. Tuttavia, quando la numerosità campionaria è piccola, il test alla Wald porta tipicamente ad un errore di primo tipo inflazionato (maggiore del livello nominale scelto). Perciò sono state proposte varie modifiche per ottenere un migliore accordo tra livelli effettivi e nominali della statistica test per GEE. In questa trattazione si considera solamente il test-t, proposto da Pan (2001) ed estensivamente studiato da Wang & Long (2011).

La derivazione del test-t nasce dall'idea di approssimare la distribuzione di $z = \frac{\hat{\beta}_r}{\sqrt{\hat{V}(\hat{\beta}_r)}}$, per la quale si fa di solito riferimento alla distribuzione approssimata $N(0,1)$ sotto l'ipotesi $H_0 : \beta_r = 0$. Utilizzando l'approssimazione $\hat{V}(\hat{\beta}_r) \dot{\sim} c\chi^2_d$ si ha

$$E[\hat{V}(\hat{\beta}_r)] \doteq cd,$$
$$Var(\hat{V}(\hat{\beta}_r)) \doteq c^2 2d,$$

quindi sfruttando il metodo dei momenti è possibile trovare uno stimatore per c e per d. Risolvendo il sistema

$$\begin{cases} E[\hat{V}(\hat{\beta}_r)] = cd \\ Var(\hat{V}(\hat{\beta}_r)) = c^2 2d \end{cases}$$

rispetto a c e d, si ottiene

$$c = \frac{Var(\hat{V}(\hat{\beta}_r))}{2E[\hat{V}(\hat{\beta}_r)]}, \quad d = \frac{2(E[\hat{V}(\hat{\beta}_r)])^2}{Var(\hat{V}(\hat{\beta}_r))}.$$

Si indicano rispettivamente con $\nu/2k$ e $2k^2/\nu$ la stima di c e d, dove ν indica la stima della varianza dello stimatore della varianza di $\hat{\beta}$, mentre k è la corrispondente stima

della media ($k \doteq cd$). Infine quindi si considera l'approssimazione

$$\frac{\hat{\beta}_r/\sqrt{k}}{\sqrt{Var(\hat{V}(\hat{\beta}_r))/cd}} \doteq \frac{\hat{\beta}_r}{\sqrt{Var(\hat{V}(\hat{\beta}_r))}} \dot{\sim} t_d, \tag{2.16}$$

con d stimato da $2\hat{V}(\hat{\beta}_r)^2/Var(\hat{V}(\hat{\beta}_r))$, dove $\hat{V}(\hat{\beta}_r)^2$ è un approssimazione di $(E[\hat{V}(\hat{\beta}_r)])^2$.

Capitolo 3

Risultati di simulazione ed applicazione

Introduzione

In questo capitolo viene analizzato numericamente il comportamento dei nove stimatori della matrice di covarianza presentati nel capitolo 2, incluso l'originale stimatore "sandwich" proposto da Liang & Zeger (1986). Tale studio è stato svolto utilizzando il *software* R ed il pacchetto "geesmv" sviluppato da Wang et al. (2016a) (il codice R utilizzato per le simulazioni e la costruzione dei grafici, è riportato nell'appendice E). I confronti verranno fatti considerando tre diversi scenari: risposta di tipo continuo, risposta di tipo conteggio e risposta binaria. La misura di bontà utilizzata per valutare i diversi stimatori è l'errore di primo tipo dei test di Wald e t. L'obiettivo è individuare gli stimatori che garantiscono il miglior avvicinamento al livello nominale del test, considerando anche diverse numerosità campionarie n ed m, in due casi: coefficienti di regressione pari a zero e coefficienti di regressione diversi da zero. Infine, verranno riportati due esempi con dati reali.

3.1 Studio di simulazione

I modelli utilizzati per la generazione dei dati sono

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + b_i + \epsilon_{ij}, \tag{3.1}$$

$$\log(E[Y_{ij}|b_i]) = \beta_0 + \beta_1 x_{ij} + b_i, \tag{3.2}$$

$$\text{logit}(E[Y_{ij}|b_i]) = \beta_0 + \beta_1 x_{ij} + b_i, \tag{3.3}$$

per risposte continua, di conteggio e binaria, rispettivamente. Inoltre $\beta_0 = 0$ e $\beta_1 = 0$, $i = 1, \ldots, n$, con numerosità campionaria $n = 10, 20, 30, 40, 50$ e $j = 1, \ldots, m$, quindi si ha lo stesso numero di osservazioni all'interno di ogni unità (dimensione del gruppo) per $m = 5, 10, 20$. La covariata x_{ij} si distribuisce come una $N(0, 1)$ i.i.d per $i = 1, \ldots, n$ e $j = 1, \ldots, m$. Gli effetti casuali di unità b_i si distribuiscono come $N(0, \sigma_b^2)$ i.i.d per $i = 1, \ldots, n$ con $\sigma_b^2 = 0.25$ e l'errore casuale ϵ_{ij} si distribuisce come una $N(0, \sigma_\epsilon^2)$ i.i.d per $i = 1, \ldots, n$, e $j = 1, \ldots, m$, con $\sigma_\epsilon^2 = 0.8$. Va sottolineato infine che ognuno dei modelli porta ad una struttura di correlazione interscambiabile (*exchangeable*) con diversi parametri di correlazione in corrispondenza del tipo di risposta. Segeundo i calcoli riportati nell'appendice C, si ottiene che nel caso di risposta continua il vero parametro di correlazione è $\alpha = \sigma_b^2/(\sigma_b^2 + \sigma_\epsilon^2) \approx 0.2$, nel caso di risposta conteggio $\alpha \approx \sigma_b^2/(1+\sigma_b^2) \approx 0.3$, mentre nel caso di risposta binaria $\alpha \approx \frac{\sigma_b^2/16}{E\left(\frac{1}{1+\exp(-b_i)}\right)\left[1-E\left(\frac{1}{1+\exp(-b_i)}\right)\right]} \approx 0.1$.

Per ogni scenario, sono stati generati 1000 data sets tramite generazione Monte Carlo, dove in ognuno è stato stimato il parametro β_1 e sono state calcolate le nove stime della matrice di covarianza. Il test di Wald ed il test t sono entrambi utilizzati per verificare l'ipotesi $H_0 : \beta_1 = 0$ contro l'alternativa $H_1 : \beta_1 \neq 0$, al livello di significatività 0.05 ed è stato calcolato il "vero" errore di primo tipo. L'obiettivo è quindi quello di studiare l'andamento dell'errore di primo tipo nei diversi contesti designati, per valutare quanto sono affidabili i due test di Wald e t. Si considerano inoltre tre tipi di matrici di lavoro (*working correlation matrix*): indipendenza, interscambiabile ed AR-1 (autoregressiva di ordine 1), questo per valutare quanto la misspecificazione della struttura di correlazione incida sui risultati della procedura di verifica di ipotesi.

I risultati di simulazione in forma di grafici sono riportati nella sezione D dell'appendice. In generale, si osserva che, come ci si aspettava, gli esiti basati sulla statistica di Wald mostrano che l'utilizzo dello stimatore V_{LZ} porta sempre ad un errore di primo tipo inflazionato quando la numerosità n è piccola, ad esempio ≤ 50. Tuttavia, anche nel caso si utilizzino altri stimatori della matrice di covarianza, si osserva ancora che in alcune circostanze l'errore di primo tipo risulta inflazionato. Lo stimatore che però generalmente ha migliore comportamento è V_{WL}. L'utilizzo del test t come è chiaro dai grafici, porta a migliori risultati in termini di preservare l'errore di primo tipo rispetto al test di Wald. Interessante è notare che lo stimatore V_{LZ}, quando la matrice di lavoro è correttamente specificata permette di ottenere risultati soddisfacenti anche quando la numerosità n è ≥ 30, specialmente considerando il test t. Inoltre V_{KC} ha prestazioni peggiori rispetto a V_{LZ} nel caso del test di Wald, ma migliora con l'aumento della dimensione dei gruppi. Va anche osservato che stimatori come V_{GST} e V_{MBN} sia nel caso

del test t che di Wald portano a test conservativi.

Concludendo, è possibile affermare che lo stimatore V_{WL} ha prestazioni migliori in una varietà di scenari differenti. Perciò questo è lo stimatore da preferire nel caso di equazioni di stima generalizzate (GEE) quando la numerosità n è inferiore a 10. Nella Tabella 3.1 vengono riportate le dimensioni campionarie n necessarie a preservare l'errore di primo tipo, rispettivamente per ognuno dei nove stimatori della matrice di covarianza di $\hat{\beta}$.

TABELLA 3.1: Guida per la numerosità n ottimale a preservare l'errore di primo tipo, per ciascuno dei nove stimatori della matrice di covarianza.

Simatori	Numerosità n
V_{LZ}	≥ 50
V_{MK}	≥ 40
V_{KC}	≥ 50
V_{PAN}	≥ 30
V_{GST}	≥ 20
V_{MD}	≥ 30
V_{FG}	≥ 40
V_{MBN}	≥ 50
V_{WL}	≥ 10

Si considera ora, il caso in cui, il parametro di regressione β abbia componenti diverse da zero. Questo ulteriore scenario, è di interesse, poichè spesso in casi di studio reali (ad esempio si veda il paragrafo 3.2), si osserva che i coefficienti di regressione sono diversi da zero. Inoltre, una conseguenza importante, è la possibile creazione di correlazione tra le unità appartenenti allo stesso gruppo. Tale fenomeno può causare un certo grado di imprecisione delle stime, portando a differenti conclusione per quanto riguarda il preservare l'errore di primo tipo dei test di Wald e t. Per motivi di natura computazionale, i casi che verranno analizzati sono: variabile risposta di tipo continuo e binaria. I modelli utilizzati per la generazione sono (3.1) e (3.3), ma con coefficienti di regressione $\beta_0 = 1$ e $\beta_1 = 1.5$. Inoltre, sono state considerate diverse numerosità n per i due scenari, rispettivamente $n = 10, 20, 30, 40, 50$ e $n = 30, 40, 50, 60, 70$. Questa distinzione è stata necessaria in quanto, nel caso di risposta binaria, considerando numerotià n troppo piccole si incorreva in problemi di non convergenza delle stime. Infine è importante sottolineare che l'ipotesi da verificare non è più l'ipotesi di nullità di β_1 ma l'ipotesi $H_0 : \beta_1 = 1.5$ contro l'alternativa $H_1 : \beta_1 \neq 1.5$, al livello di confidenza 0.05.

I risultati di simulazione in forma di grafici sono riportati nella sezione D dell'appendice. In generale valgono analoghe considerazioni al caso precedente per quanto

riguarda: il test di Wald e lo stimatore V_{LZ}, l'utilizzo di altri stimatori della matrice di covarianza, l'utilizzo del test t e le considerazioni fatte sull'accuratezza dei test nel caso di corretta specificazione della struttura di correlazione, soprattutto considerando il test t e la risposta di tipo continuo. Tuttavia, si osserva che il livello effettivo dei

TABELLA 3.2: Guida per la numerosità n ottimale a preservare l'errore di primo tipo, per ciascuno dei nove stimatori della matrice di covarianza, con risposta di tipo continuo e coefficienti di regressione diversi da zero.

Simatori	Numerosità n
V_{LZ}	> 50
V_{MK}	≥ 50
V_{KC}	> 50
V_{PAN}	≥ 30
V_{GST}	≥ 20
V_{MD}	≥ 40
V_{FG}	≥ 40
V_{MBN}	≥ 40
V_{WL}	≥ 10

test è generalmente più elevato del livello nominale. Questo dimostra che ponendo i coefficienti di regressione diversi da zero, cambia la capacità dei test di Wald e t di preservare l'errore di primo tipo. Nel caso di risposta binaria si osserva che l'aumento

TABELLA 3.3: Guida per le numerosità n ed m ottimali a preservare l'errore di primo tipo, per ciascuno dei nove stimatori della matrice di covarianza, con risposta binaria e coefficienti di regressione diversi da zero.

Simatori	Numerosità n	Numerosità m
V_{LZ}	≥ 70	≥ 10
V_{MK}	≥ 60	≥ 10
V_{KC}	≥ 70	≥ 20
V_{PAN}	≥ 70	≥ 10
V_{GST}	$\geq 50, \geq 60$	$\geq 10, \geq 5$
V_{MD}	≥ 60	≥ 10
V_{FG}	≥ 60	≥ 10
V_{MBN}	$\geq 50, \geq 60$	$\geq 10, \geq 5$
V_{WL}	$\geq 50, \geq 60$	$\geq 10, \geq 5$

della numerosità m all'interno dei gruppi, porta a test più vicini al livello nominale, inoltre si ottengono risultati soddisdacenti per $n \geq 50$ ed $m \geq 10$ nel caso di corretta

specificazione della struttura di correlazione. Infine, per la risposta di tipo continuo, lo stimatore V_{WL} sembrerebbe essere il migliore sempre in termini di coerenza con il livello nominale dei relativi test, mentre per la risposta binaria gli stimatori V_{WL}, V_{MBN} e V_{GST} sembrerebbero portare a test più accurati. Quindi, in generale per preservare l'errore di primo tipo, valgono i risultati nella Tabella 3.2, considerando le dimensioni campionarie n ed m e la variabile risposta risposta di tipo continuo, mentre per la risposta binaria, la guida è riportata nella Tabella 3.3.

3.2 Applicazione a casi reali

Si presentano ora due casi di studio per confrontare le performance dei diversi stimatori della matrice di covarianza nel caso di numerosità campionaria finita. Vengono considerate due tipi di risposta: continua e binaria. Per entrambe le analisi è stato utilizzato il pacchetto "geesmv" sviluppato da Wang et al. (2016a). Il primo data set contiene le

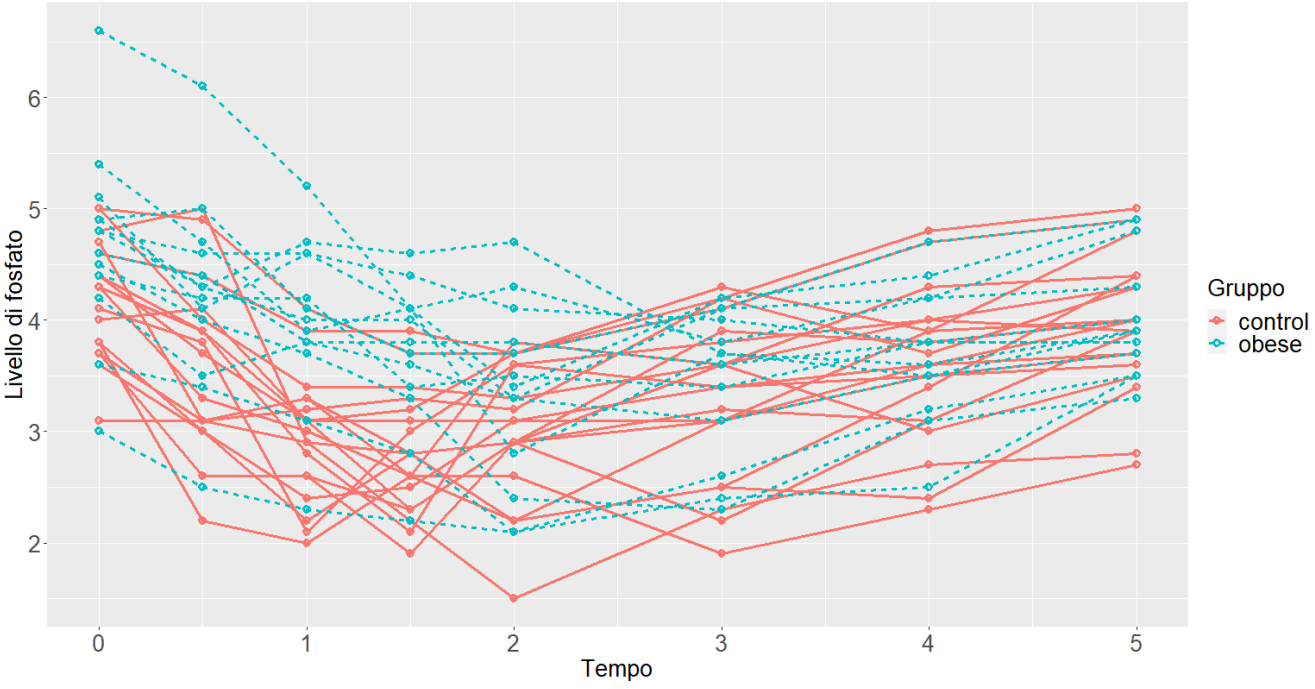

FIGURA 3.1: Misure del livello di fosfato inorganico nel plasma, andamenti individuali rispetto al tempo.

misurazioni del livello di fosfato inorganico nel plasma di 33 soggetti, 20 controlli e 13 obesi, a cui è stato somministrato un carico di glucosio. Le misurazioni sono state fatte a distanza di $0, 0.5, 1, 1.5, 2, 3, 4, 5$ ore dalla somministrazione. L'obiettivo è quello di valutare se e come i livelli di fosfato varino nel tempo e se vi sia differenza tra soggetti obesi e controlli.

In Figura 3.1 viene riportato il grafico degli andamenti indivudiali del livello di fosfato inorganico nel plasma rispetto al tempo, da cui si osserva che il tempo ha un effetto quadratico sulla risposta e che sembrerebbe esserci una differenza in media tra il gruppo di controllo ed il gruppo di obesi. Tramite alcune analisi preliminari è possibile stabilire che il miglior modello marginale per il livello di fosfato è

$$Y = \beta_0 + \beta_1 gruppo + \beta_2 tempo + \beta_3 tempo^2 + \beta_4 (gruppo \times tempo). \tag{3.4}$$

TABELLA 3.4: Stima della deviazione standard e risultati dei test di Wald e t, nel caso di studio: livello di fosfato nel plasma.

	$\hat{\beta}$	$\sqrt{V_{LZ}}$	$\sqrt{V_{MK}}$	$\sqrt{V_{KC}}$	$\sqrt{V_{PAN}}$	$\sqrt{V_{GST}}$	$\sqrt{V_{MD}}$	$\sqrt{V_{FG}}$	$\sqrt{V_{MBN}}$	$\sqrt{V_{WL}}$
					Indipendenza					
gruppo	0.84	0.24	0.26	0.25	0.22	0.24	0.26	0.26	0.26	0.23
tempo	-0.77	0.08	0.08	0.08	0.07	0.08	0.08	0.08	0.10	0.08
$tempo^2$	0.16	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01
$gr:t$	-0.16	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.05
					Interscambiabile					
gruppo	0.84	0.24	0.26	0.25	0.22	0.24	0.26	0.26	0.27	0.23
tempo	-0.77	0.08	0.08	0.08	0.07	0.08	0.08	0.08	0.08	0.08
$tempo^2$	0.16	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
$gr:t$	-0.16	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
					Autoregressiva					
gruppo	0.66	$0.24^{\ddagger}$	$0.26^{\dagger}$	$0.30^{\dagger}$	0.21	$0.23^{\ddagger}$	$0.26^{\dagger}$	$0.26^{\dagger}$	$0.26^{\dagger}$	$0.23^{\dagger}$
tempo	-0.76	0.06	0.07	0.12	0.06	0.07	0.06	0.06	0.07	0.06
$tempo^2$	0.15	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01
$gr:t$	-0.11	$0.04^{\ddagger}$	$0.05^{\dagger}$	$0.05^{\dagger}$	0.04	$0.04^{\dagger}$	$0.04^{\dagger}$	$0.04^{\dagger}$	$0.05^{\dagger}$	$0.04^{\ddagger}$

Tutti i test effettuati in corrispondenza di ciascuna modifica allo stimatore V_{LZ} (questo compreso) sono significativi sia al livello 0.01 che 0.05 utilizzando il test di Wald ed il test t, eccetto nel caso di presenza dell'apice, infatti:

† indica che il test non è significativo nè con la statistica di Wald nè con la statistica t al livello 0.01;

‡ indica che il test è significativo al livello 0.01 con il test di Wald ma non con il test t.

I risultati dell'analisi tramite equazioni di stima generalizzate, inclusa la stima dei parametri e delle rispetive varianze, considerando ciasuno dei nove stimatori della matrice di covarianza, sono riportati in Tabella 3.4. Entrambi i test di Wald e t sono applicati per la verifica di ipotesi di nullità dei parametri del modello (3.4), al livello 0.01 e

0.05. In generale, tutti gli stimatori portano risultati simili per quanto riguarda l'effetto di *tempo* e $tempo^2$, sia con il test di Wald che con il test t al livello di significatività 0.01. Tuttavia, specialmente nel caso di matrice di lavoro autoregressiva, utilizzando il test t si osservano differenti conclusioni per i parametri relativi alle variabili *gruppo* e $gruppo \times tempo$, indicando che la scelta di effettuare aggiustamenti della statistica di Wald nel caso di piccoli campioni influenza i risultati dei test.

Il secondo esempio è relativo ad una sperimentazione clinica, che ha coinvolto $n = 56$ pazienti, avente l'obiettivo di valutare l'effetto di due trattamenti sulla situazione respitatoria. I soggetti ricevono casualmente il trattamento (trt) attivo o il trattamento passivo (placebo). La situazione respiratoria del paziente è descritta da 0 (buona) e 1 (cattiva) in corrispondenza di 4 visite consecutive. Vengono riportate in aggiunta: la situazione del paziente in una visita precedente all'inizio del trattamento (base), il sesso e l'età (age).

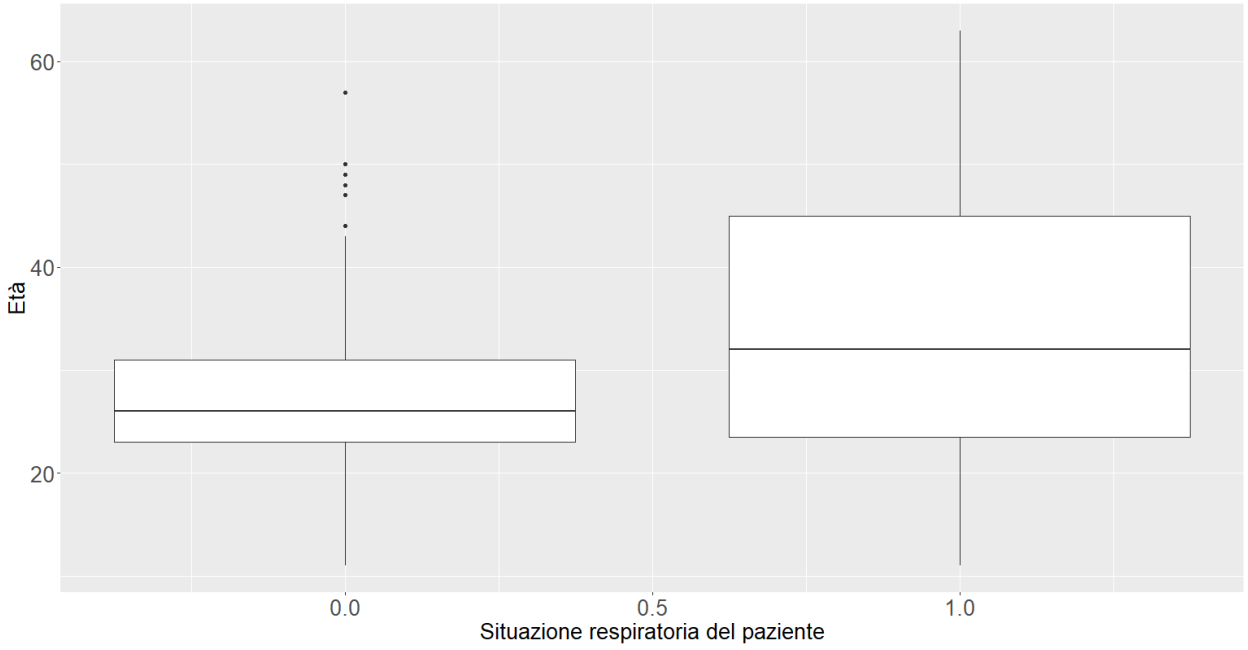

FIGURA 3.2: Incidenza dell'età sulla situazione respiratoria del paziente

Osservando il boxplot in Figura 3.2 per la valutazione dell'effetto dell'età sulla situazione respiratoria e le Tabelle di frequenza 3.5- 3.7, relative rispettivamente all'effetto sulla situazione respiratoria del tipo di trattamento, del sesso e della situazione respiratoria prima dell'inizio del trattamento (base), si nota che tutte le varibili sembrerebbero avere un effetto sull'andamento della situazione respiratoria del paziente nell'arco delle 4 visite, a meno della variabile base che invece non sembrerebbe avere alcun effetto. Infatti, successive analisi portano a preferire il seguente modello marginale logit-lineare

per la risposta

$$\text{logit}(E[Y]) = \beta_0 + \beta_1 visita + \beta_2 sesso + \beta_3 trt + \beta_4 (trt \times visita). \tag{3.5}$$

Tabella 3.5: Distribuzione di frequenza della situazione respiratoria, condizionata al tipo di trattamento.

Situazione respiratoria		
	0	1
Trt attivo	0.62	0.38
Placebo	0.36	0.64

Tabella 3.6: Distribuzione di frequenza della situazione respiratoria, condizionata al genere del paziente.

Situazione respiratoria		
	0	1
Femmina	0.11	0.89
Maschio	0.54	0.46

Tabella 3.7: Distribuzione di frequenza della situazione respiratoria, condizionata alla situazione respiratoria prima dell'inizio del trattamento (base).

Situazione respiratoria		
	0	1
Base 0	0.36	0.64
Base 1	0.51	0.49

I risultati dell'analisi sono riportati nella Tabella 3.8. Si nota che l'effetto delle varibili *sesso* e *trt* dipende fortemente dal tipo di test utilizzato. In paticolare utilizando il test t al livello di significatività 0.01 spesso l'ipotesi di nullità dei parametri β_2 e β_3 non viene rifiutata, dimostrando ancora che il test t risulta più conservativo rispetto al test di Wald. Inoltre è possibile osservare come gli stimatori V_{MBN}, V_{WL} , V_{MD} e V_{MK} influenzano la procedura di verifica di ipotesi, portando a test più conservativi.

TABELLA 3.8: Risultati di stima nel caso di studio: sperimentazione clinica di due trattamenti sulla situazione respiratoria.

	$\hat{\beta}$	$\sqrt{V_{LZ}}$	$\sqrt{V_{MK}}$	$\sqrt{V_{KC}}$	$\sqrt{V_{PAN}}$	$\sqrt{V_{GST}}$	$\sqrt{V_{MD}}$	$\sqrt{V_{FG}}$	$\sqrt{V_{MBN}}$	$\sqrt{V_{WL}}$
					Indipendenza					
visita	-1.85	0.34	0.36	0.36	0.30	0.32	0.36	0.40	0.37	0.31
sesso	-2.64	0.69	0.73‡	0.87‡	0.92	0.97‡	0.81‡	0.84‡	0.77	0.98‡
trt	-2.61	0.97‡	1.03†	1.03	0.88	0.93	1.02†	1.14†	1.05†	0.91
age	0.06	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
trt : *vi*	1.63	0.38	0.40	0.41	0.35	0.37	0.40	0.45	0.41	0.36
					Interscambiabile					
visita	-1.90	0.34	0.36	0.24	0.29	0.31	0.35	0.36	0.36	0.31
sesso	-2.51	0.69	0.73‡	0.81	0.86‡	0.91‡	0.81‡	0.80‡	0.78‡	0.92‡
trt	-2.76	0.96	1.02	0.74	0.88	0.93	1.01	1.03‡	1.02‡	0.92
age	0.06	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
trt : *vi*	1.68	0.38	0.40	0.30	0.34	0.36	0.40	0.40	0.40	0.36
					Autoregressiva					
visita	-1.87	0.34	0.36	0.28	0.29	0.31	0.35	0.37	0.36	0.30
sesso	-2.50	0.72‡	0.77‡	0.82‡	0.89‡	0.94‡	0.85‡	0.84‡	0.80‡	0.95†
trt	-2.62	0.96	1.01‡	0.85	0.86	0.91	1.00‡	1.06†	1.03†	0.90
age	0.06	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
trt : *vi*	1.62	0.38	0.40	0.34	0.34	0.36	0.40	0.42	0.41	0.35
					Non strutturata					
visita	-1.87	0.33	0.35	0.28	0.29	0.31	0.35	0.35	0.35	0.30
sesso	-2.44	0.70	0.74‡	0.80‡	0.84‡	0.89‡	0.82‡	0.80‡	0.78‡	0.90‡
trt	-2.53	0.93‡	0.99†	0.90	0.86	0.92	0.97†	1.00†	1.00	0.90
age	0.07	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
trt : *vi*	1.60	0.37	0.39	0.33	0.34	0.36	0.38	0.39	0.39	0.35

Tutti i test effettuati in corrispondenza di ciascuna modifica allo stimatore V_{LZ} (questo compreso) sono significativi sia al livello 0.01 che 0.05 utilizzando il test di Wald ed il test t, eccetto nel caso di presenza dell'apice, infatti:

† indica che il test non è significativo nè con la statistica di Wald nè con la statistica t al livello 0.01;

‡ indica che il test è significativo al livello 0.01 con il test di Wald ma non con il test t.

Conclusione

In questa relazione, si è fornito un breve quadro dei recenti sviluppi in merito alle modifiche dello stimatore della matrice di covarianza per equazioni di stima generalizzate (GEE), allo scopo di migliorarne l'accuratezza in caso di piccoli campioni. Inoltre, sono stati discussi due tipi di test per GEE, quali il test di Wald ed il test t, di cui sono state valutate le capacità di preservare l'errore di primo tipo quando la numerosità campionaria è piccola. Tramite i diversi studi di simulazione e le applicazioni a casi reali, sono state confrontate le prestazioni di ogni modifica allo stimatore V_{LZ}. Come indicato nel capitolo relativo ai risultati di simulazione, si osserva che, in generale, il test t basato sullo stimatore V_{WL} (Wang & Long, 2011) ha migliori prestazioni in diversi scenari. Inoltre, è stata fornita una guida per la numerosità campionaria adeguata a controllare l'errore di primo tipo in corrisponza di ciascuno stimatore della matrice di covarianza, considerando il caso con coefficienti di regressione pari a zero e diversi da zero.

Appendice

A Famiglie di dispersione esponenziale

Una particolare famiglia di distribuzioni a cui si fa riferimento nei contesti dei modelli di regressione è la **famiglia di dispersione esponenziale** introdotta da Nelder & Wedderburn (1972), dove la densità di Y_i, $i = 1, \ldots, n$ è definita dalla formula

$$p(y_i; \theta, \phi) = \exp\left\{ \frac{\theta_i y_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \right\} \tag{A.1}$$

con $y_i \in S \subseteq \mathbb{R}$, $\theta_i \in \Theta \subseteq \mathbb{R}$, $a_i(\phi) > 0$. La grande particolarità di questa familgia di distribuzioni è la flessibilità e capacità di adattarsi a più contesti, dove la risposta non sia solo una variabile continua ma anche discreta come una Poisson o una binomiale. Inoltre, i modelli specificati tramite la (A.1) permettono di tenere in considerazione anche alcune forme di eteroschedasticità, a differenza del modello lineare normale classico. Il parametro θ_i è detto **parametro naturale**, mentre ϕ è detto **parametro di dispersione** e specificando le funzioni $a_i(\cdot)$, $b(\cdot)$ e $c(\cdot, \cdot)$ è possibile ottenere un particolare modello parametrico.

Si dimostrano inoltre delle importanti proprietà strutturali della famiglia di distribuzione esponenziale che ne sottolineano nuovamente la grande capacità di adattamento. In particolare tramite il calcolo della derivata prima e seconda rispetto al parametro θ_i della funzione di log-verosimigliaza

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

e grazie ai risultati (1.2) e (1.3), validi anche in questo contesto, si ottiene che

$$E(Y_i) = E_{\theta_i, \phi}(Y_i) = b^{'}(\theta_i), \tag{A.2}$$

$$Var(Y_i) = Var_{\theta_i, \phi}(Y_i) = a_i(\phi) b^{''}(\theta_i) \tag{A.3}$$

con $b^{'}(\theta_i)$ e $b^{''}(\theta_i)$ si indicano rispettivamente derivata prima e seconda di $b(\theta_i)$.

In generale la specificazione di una distrubuzione esponenziale avviene nella parametrizzazione (μ_i, ϕ) anziché (θ_i, ϕ). Definita la funzione $\mu(\cdot) : \Theta \longrightarrow M$ con $M = \mu(int\,\Theta)$ **spazio delle medie** e $int\,\Theta$ l'insieme dei punti interni di Θ. Seguono i risultati

$$\mu_i = \mu(\theta_i) = E_{\theta,\phi}(Y_i) = b'(\theta_i),$$

da cui, dalla (A.2), si ottiene

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

Nella parametrizzazione (μ_i, ϕ), definita con $\theta(\mu_i)$ la funzione inversa di $\mu(\theta_i)$, la varianza di Y_i diventa

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

con $v(\mu_i) = b''(\theta_i)\big|_{\theta_i=\theta(\mu_i)}$ definita in M e detta **funzione di varianza**.

Infine, la parametrizzazione (μ_i, ϕ), come in una famiglia di dispersione, permette di introdurre la notazione compatta per indicare la distribuzione esponenziale di una variabile casuale Y_i come

$$Y_i \sim DE_1(\mu_i, a_i(\phi)v(\mu_i)), \quad \mu_i \in M.$$

B Descrizione del confronto teorico di alcuni stimatori della matrice di covarianza

$\mathbf{M_{LZ}}$ e $\mathbf{M_{WL}}$ dopo il calcolo di $Cov(vec(M_{LZ})) - Cov(vec(M_{WL}))$, sviluppato come segue

$$\begin{aligned} Cov(vec(M_{LZ})) - Cov(vec(M_{WL})) &= \sum_{i=1}^{n} S_i T_i S_i^T - \sum_{i=1}^{n} S_i \Big[E_i \left(\sum_{j=1}^{n} \frac{1}{n^2} E_j^{-1} G_j T_j G_j^T E_j^{-1} \right) E_i \Big] S_i^T \\ &= \sum_{i=1}^{n} S_i \left[T_i - E_i \left(\sum_{j=1}^{n} \frac{1}{n^2} E_j^{-1} G_j T_j G_j^T E_j^{-1} \right) E_i \right] S_i^T, \end{aligned}$$

Wang & Long (2011) dimostrano che sotto tenui condizioni, per $n \to \infty$, $\frac{1}{n}\sum_{j=1}^{n}\Big(E_j^{-1}G_jT_jG_j^TE_j^{-1}\Big)E_i$ converge ad una generica matrice Υ con probabilità 1. Quindi $\frac{1}{n}\Upsilon$ tende a 0 con probabilità 1 per $n \to \infty$. Segue che $Cov(vec(M_{LZ})) - Cov(vec(M_{WL}))$ è definita non negativa con probailità 1 in quanto T_i (matrice di covarianza dello stimatore $\hat{r}_i$) è una matrice definita positiva.

$\mathbf{M_{MD}}$ e $\mathbf{M_{WL}}$ dopo il calcolo di $Cov(vec(M_{MD})) - Cov(vec(M_{WL}))$, sviluppato come segue

$$\begin{aligned} Cov(vec(M_{MD})) - Cov(vec(M_{WL})) &= \sum_{i=1}^{n} S_i G_i T_i G_i^T S_i^T - \\ &\sum_{i=1}^{n} S_i \left[E_i \frac{1}{n^2} \left(\sum_{j=1}^{n} E_j^{-1} G_j T_j G_j^T E_j^{-1} \right) E_i \right] S_i^T \\ &= \sum_{i=1}^{n} S_i \Bigg[G_i T_i G_i^T - \\ &E_i \left(\sum_{j=1}^{n} \frac{1}{n^2} E_j^{-1} G_j T_j G_j^T E_j^{-1} \right) E_i \Bigg] S_i^T, \end{aligned}$$

si dimostra che anche in questo caso per $n \to \infty$ il secondo addendo dell'equazione converge a 0 con probabilità 1 ed essendo T_i definita positiva allora $Cov(vec(M_{MD})) - Cov(vec(M_{WL}))$ è definita non negativa.

$\mathbf{M_{PAN}}$ e $\mathbf{M_{WL}}$ dopo il calcolo di $Cov(vec(M_{PAN})) - Cov(vec(M_{WL}))$, sviluppato come segue

$$\begin{aligned} Cov(vec(M_{PAN})) - Cov(vec(M_{WL})) &= \sum_{i=1}^{n} S_i \left[E_i \frac{1}{n^2} \left(\sum_{j=1}^{n} E_j^{-1} T_i E_j^{-1} \right) E_i \right] S_i^T - \\ &\sum_{i=1}^{n} S_i \left[E_i \frac{1}{n^2} \left(\sum_{j=1}^{n} E_j^{-1} G_j T_j G_j^T E_j^{-1} \right) E_i \right] S_i^T \\ &= \sum_{i=1}^{n} S_i \left[E_i \sum_{j=1}^{n} \left\{ \frac{1}{n^2} E_j^{-1} (T_j - G_j T_j G_j) E_j^{-1} \right\} E_i \right] S_i^T \end{aligned}$$

si dimostra che per $n \to \infty$ allora $\frac{1}{n} \sum_{j=1}^{n} E_j^{-1}(T_j - G_j T_j F_j^T) E_j^{-1}$ tende a 0 con probabilità 1. Quindi $Cov(vec(M_{PAN})) - Cov(vec(M_{WL}))$ converge a 0 con probailità 1 per $n \to \infty$.

C Derivazione coefficiente di correlazione

C.1 Risposta di tipo continuo

Per il calcolo del coefficiente di correlazione α si procede come segue,

$$Var(Y_{ij}) = \sigma_b^2 + \sigma_\epsilon^2 = 0.25 + 0.8 = 1.05$$

mentre

$$\begin{aligned} Cov(Y_{ij}, Y_{ik}) &= E[(b_i + \epsilon_{ij})(b_i + \epsilon_{ik})] \\ &= E[b_i^2 + b_i\epsilon_{ik} + \epsilon_{ij}b_i + \epsilon_{ij}\epsilon_{ik}] \\ &= E[b_i^2] = 0.25. \end{aligned}$$

per ogni $j \neq k$. Quindi

$$\alpha = \frac{Cov(Y_{ij}, Y_{ih})}{\sqrt{Var(Y_{ij})}\sqrt{VarY_{ih}}} = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_\epsilon^2} \approx 0.2.$$

C.2 Risposta binaria

Nel caso di risposta binaria, si dimostra in Guo et al. (2005) che se

$$Y_{ij}|b_i \sim Bin(1, \pi_i), \quad \pi_i = \frac{1}{1 + e^{-b_i}}, \quad b_i \sim N(0, \sigma^2),$$

allora sfruttando il metodo delta con approssimazione di Taylor al primo ordine in 0 per $\pi_i = \frac{1}{1+e^{-b_i}}$ tale per cui

$$\pi_i = \frac{1}{1 + e^{-b_i}} \approx \frac{1}{1 + e^{-b_i}}\bigg|_{b_i=0} - \frac{e^{-b_i}}{(1 + e^{-b_i})^2}\bigg|_{b_i=0} (b_i - 0) \approx \frac{1}{2} - \frac{b_i}{4},$$

si ottiene che

$$E[Y_{ij}] = E[E[Y_{ij}|b_i]] = E[\pi_i] = E\left[\frac{1}{1 + e^{-b_i}}\right] \approx \frac{1}{2}.$$

Si procede poi con il calcolo della varianza di Y_{ij}

$$\begin{aligned} Var(Y_{ij}) &= E[Var(Y_{ij}|b_i)] + Var(E[Y_{ij}|b_i]) \\ &= E[\pi_i(1 - \pi_i)] + Var(\pi_i) \\ &= E[\pi_i - \pi_i^2] + E[\pi_i^2] - (E[\pi_i])^2 \\ &= E[\pi_i] - (E[\pi_i])^2 \\ &= E[\pi_i](1 - E[\pi_i]) \approx \frac{1}{4}. \end{aligned}$$

Inoltre per ogni $j \neq k$ si ottine che

$$E[Y_{ij}Y_{ik}] = E[E[Y_{ij}Y_{ik}|b_i]] = E[\pi_i\pi_i] = E[\pi_i^2],$$

quindi

$$Cov(Y_{ij}, Y_{ik}) = E[Y_{ij}Y_{ik}] - E[Y_{ij}]E[Y_{ik}] = Var(\pi_i).$$

Sempre sfruttando il metodo delta è possibile ottenere unu'approssimazione di $Var(\pi_i)$ come segue

$$Var(\pi_i) = \left(\frac{e^{-b_i}}{(1+e^{-b_i})}\right)^2\Big|_{b_i=0} Var(b_i) = \frac{\sigma^2}{16}.$$

Si calcola infine il coefficiente di correlazione come

$$\alpha \approx \frac{\sigma^2/16}{1/4} = \sigma^2/4.$$

C.3 Risposta di tipo conteggio

Nel caso di risposta di tipo conteggio, si dimostra in Guo et al. (2005) che se

$$Y_{ij}|b_i \sim Pois(\mu_i), \quad \mu_i = e^{b_i}, \quad b_i \sim N(0, \sigma^2),$$

allora sfruttando il metodo delta con approssimazione di Taylor del primo ordine in 0 per $\pi_i = e^{b_i}$ tale per cui

$$\mu_i = e^{b_i} \approx e^{b_i}\big|_{b_i=0} + e^{b_i}\big|_{b_i=0}(b_i - 0) = 1 + b_i,$$

si ottiene che

$$E[Y_{ij}] = E[E[Y_{ij}|b_i]] = E[\mu_i] \approx 1.$$

Si procede poi con il calcolo della varianza di Y_{ij}

$$\begin{aligned} Var(Y_{ij}) &= E[Var(Y_{ij})] + Var(E[Y_{ij}|b_i]) \\ &= E[\mu_i] + Var(\mu_i) \\ &\approx 1 + Var(1 + b_i) = 1 + \sigma^2. \end{aligned}$$

Analogamente al caso con risposta binaria si ottiene

$$Cov(Y_{ij}Y_{ik}) = E[Y_{ij}Y_{ik}] - E[Y_{ij}]E[Y_{ik}] = E[\mu_i^2] - (E[\mu_i])^2 \approx \sigma^2,$$

per cui è possibile calcolare il coefficiente di correlazione

$$\alpha \approx \frac{\sigma^2}{1+\sigma^2}.$$

D Grafici

D.1 Caso di studio con coefficienti di regressione pari a zero

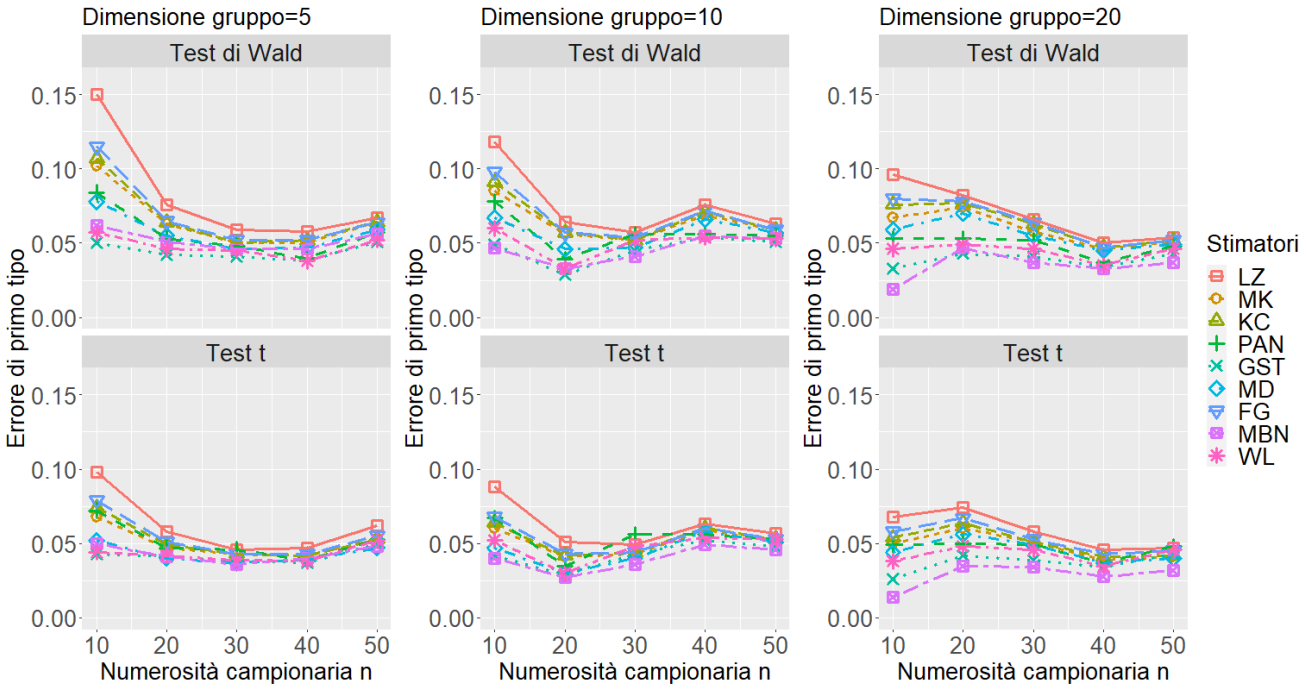

FIGURA 1: Risposta Continua, Indipendenza

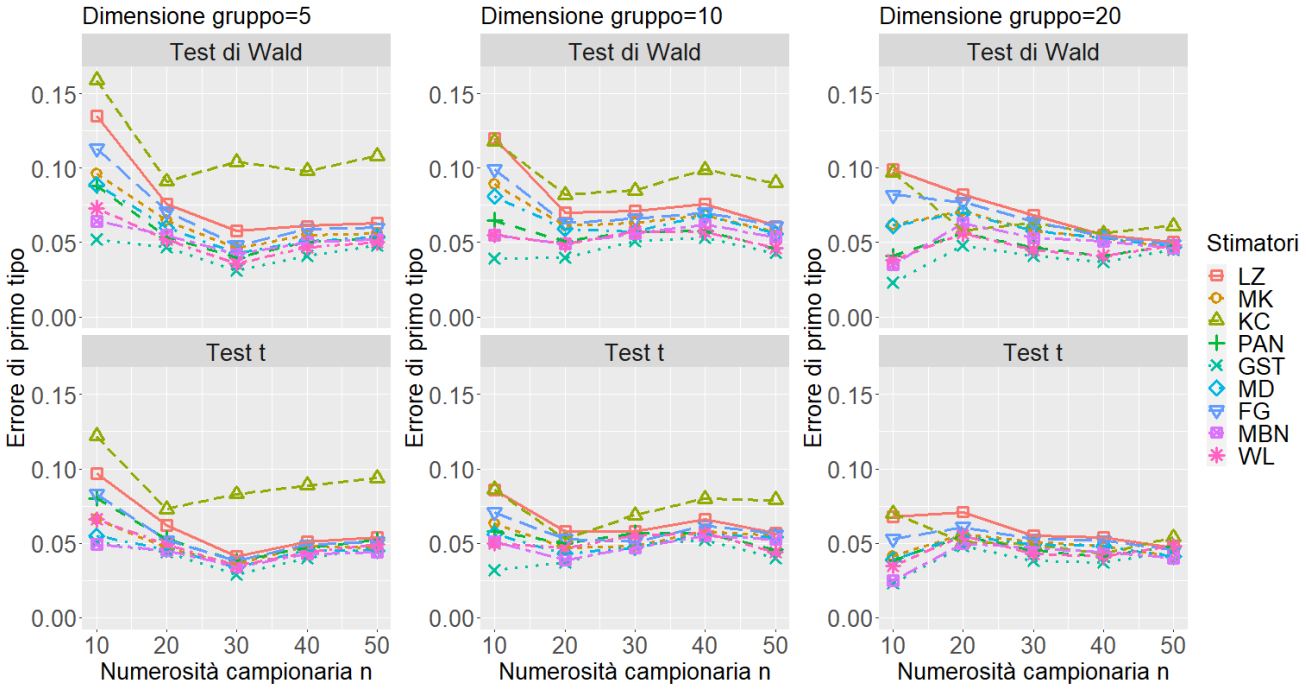

FIGURA 2: Risposta Continua, Interscambiabile (*exchangeable*)

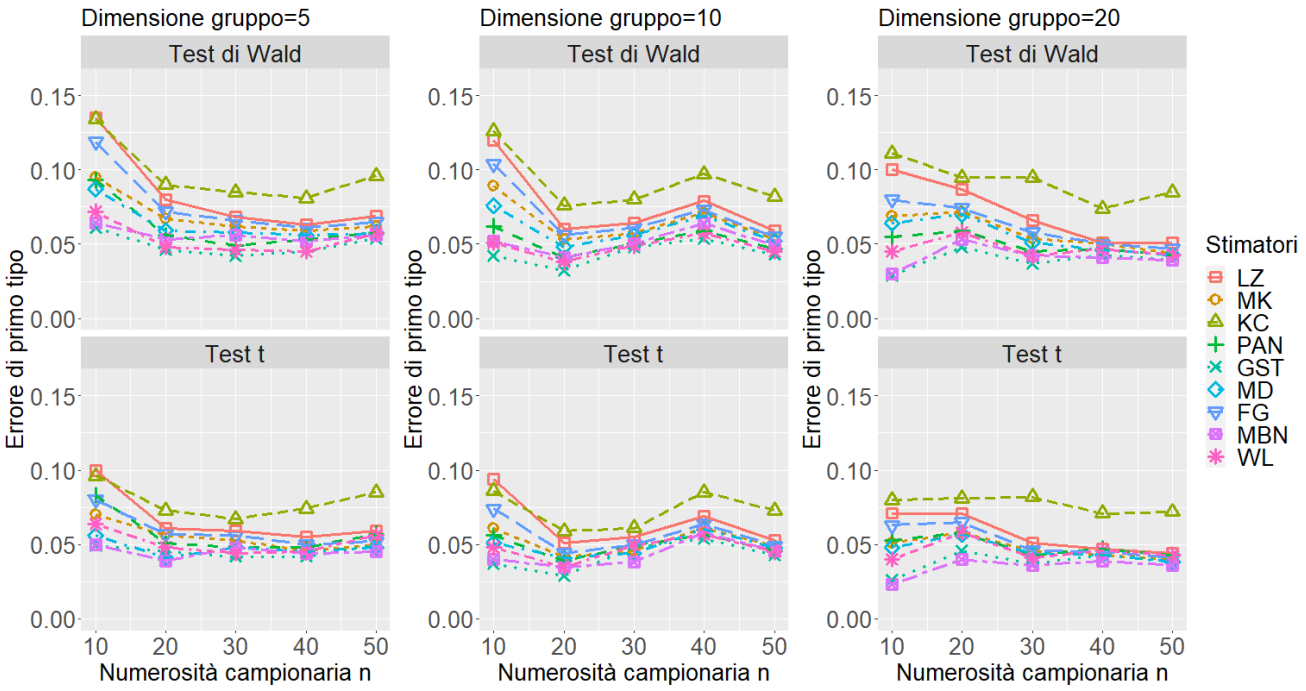

FIGURA 3: Risposta Continua, AR-1

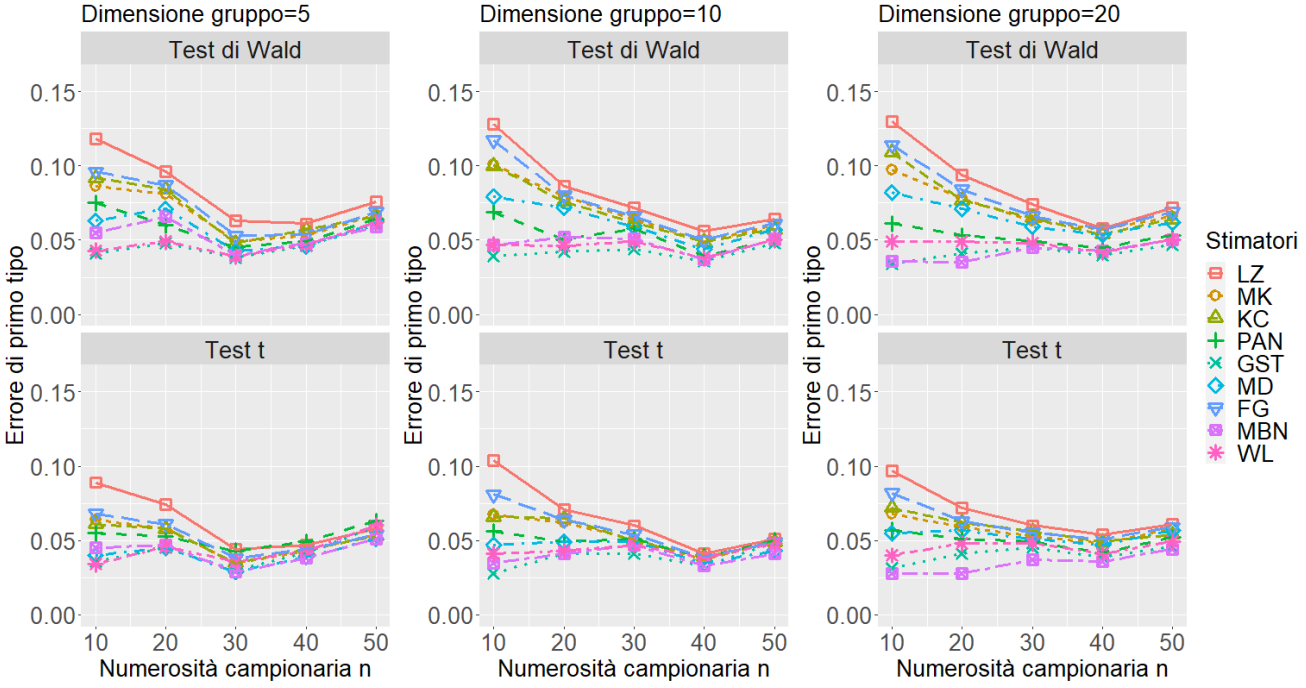

FIGURA 4: Risposta Conteggio, Indipendenza

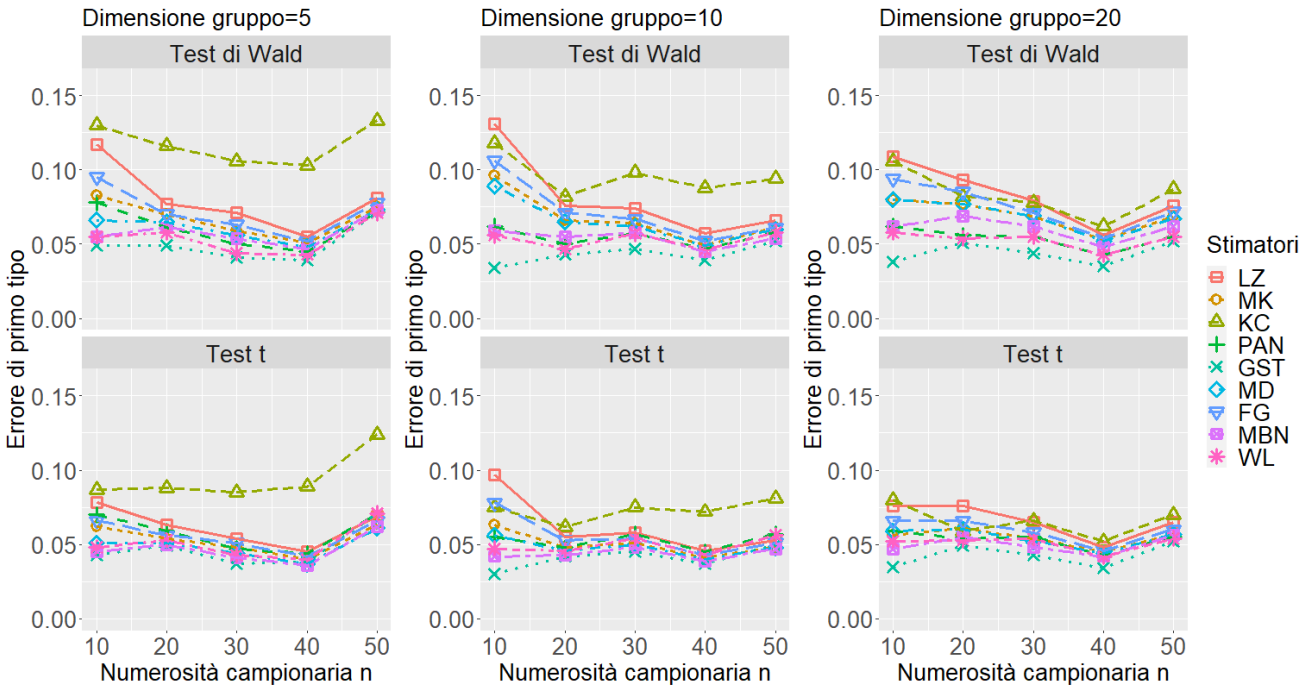

FIGURA 5: Risposta Conteggio, Interscambiabile (*exchangeable*)

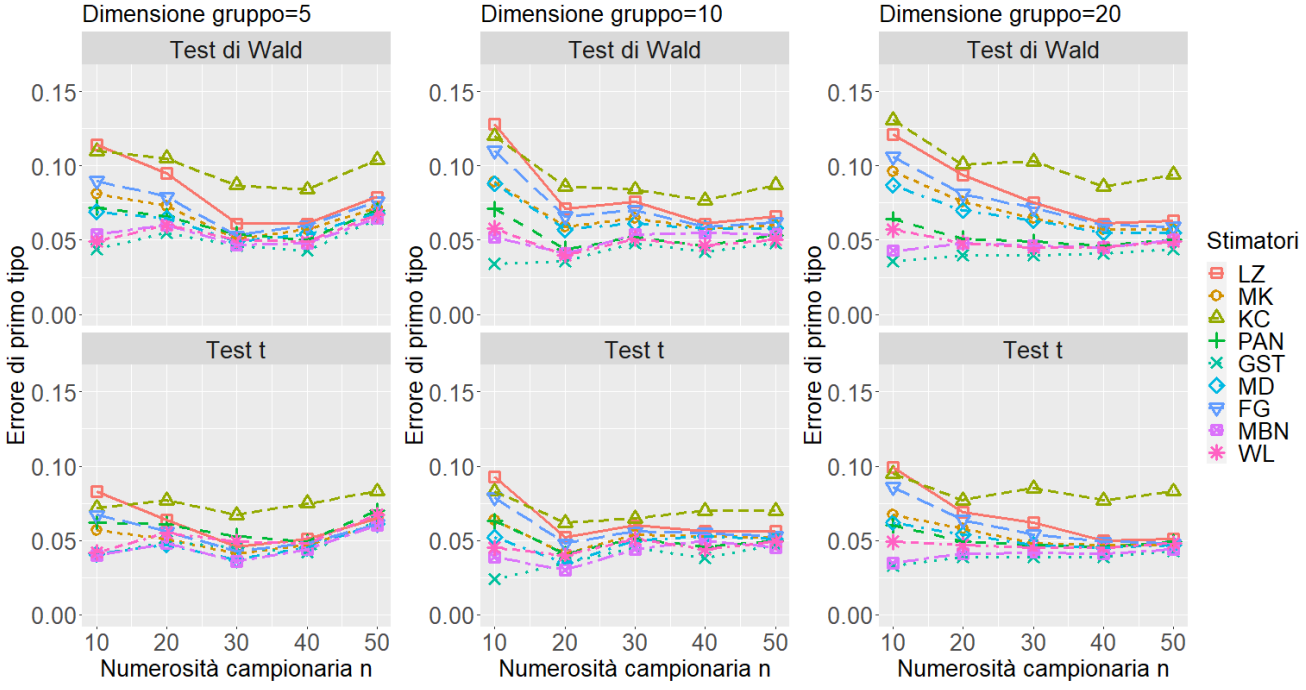

FIGURA 6: Risposta Conteggio, AR-1

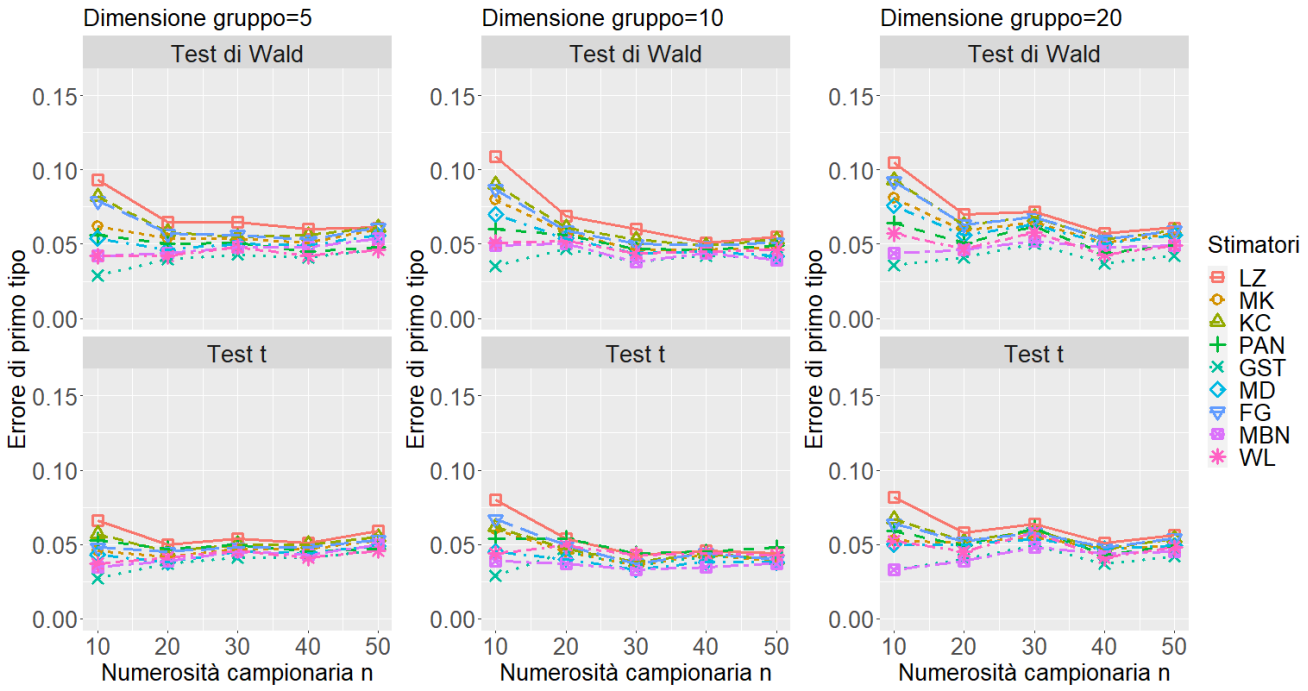

FIGURA 7: Risposta Binaria, Indipendenza

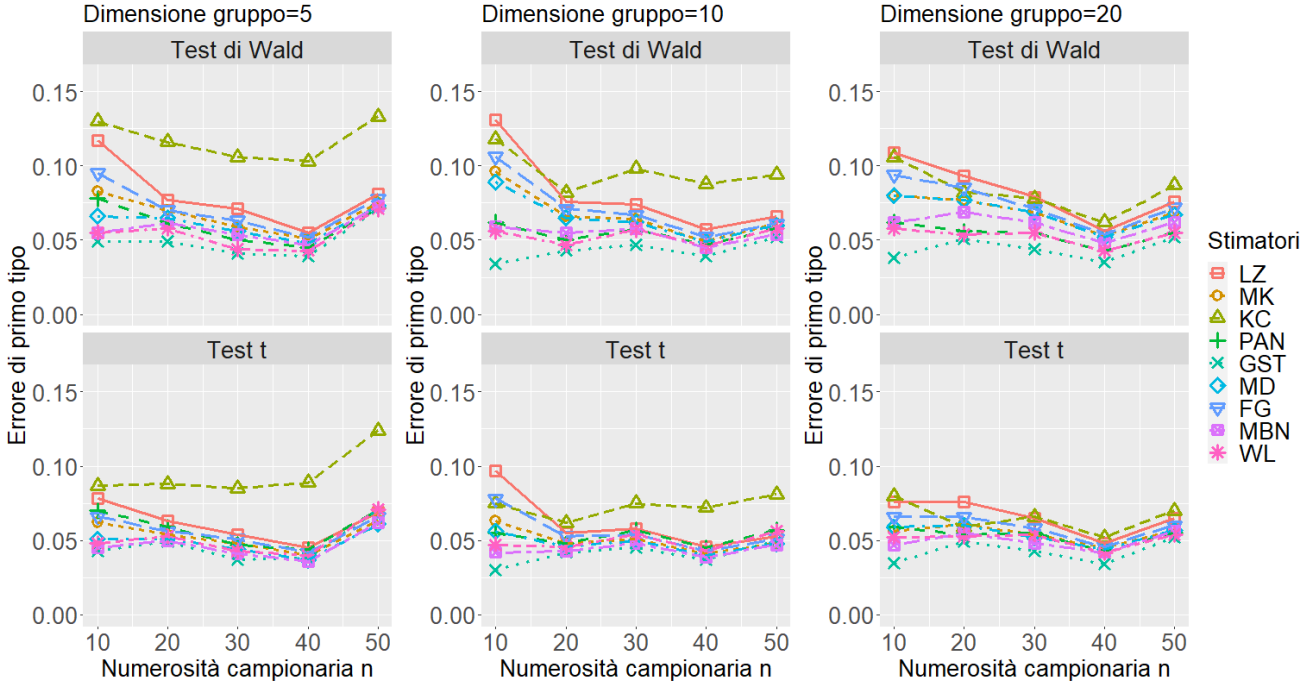

FIGURA 8: Risposta Binaria, Interscambiabile (*exchangeable*)

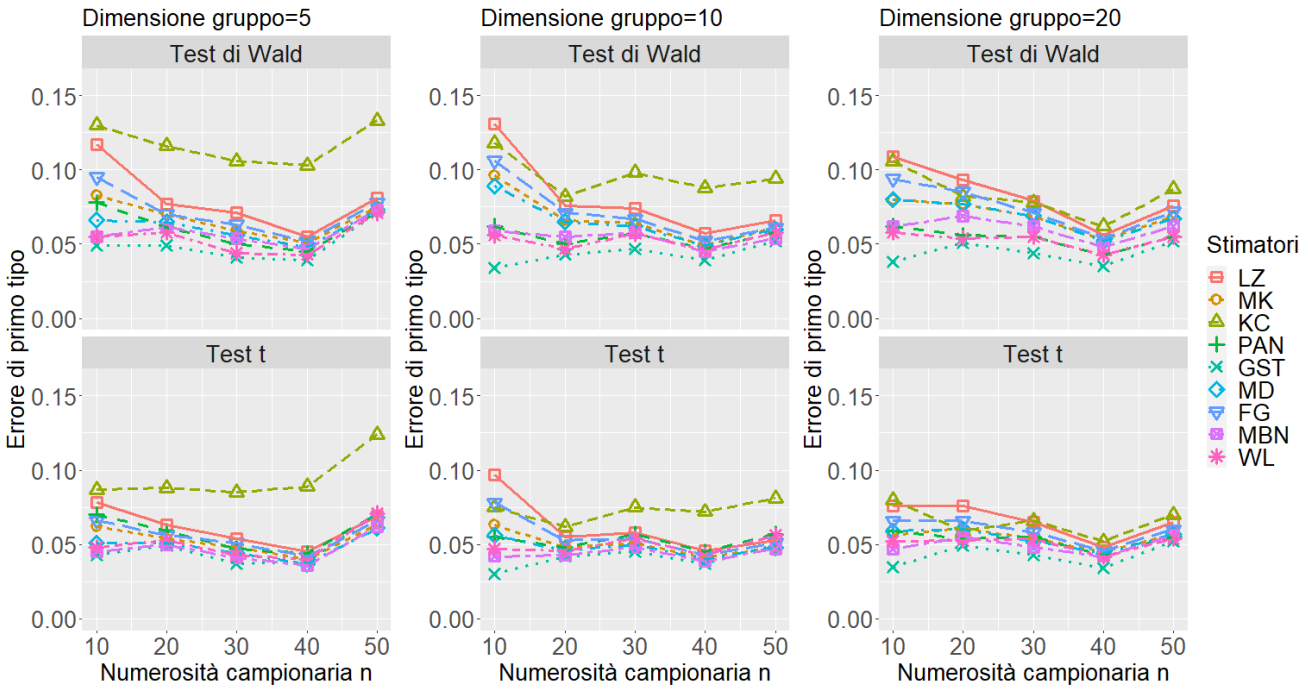

FIGURA 9: Risposta Binaria, AR-1

D.2 Caso di studio con coefficienti di regressione diversi da zero

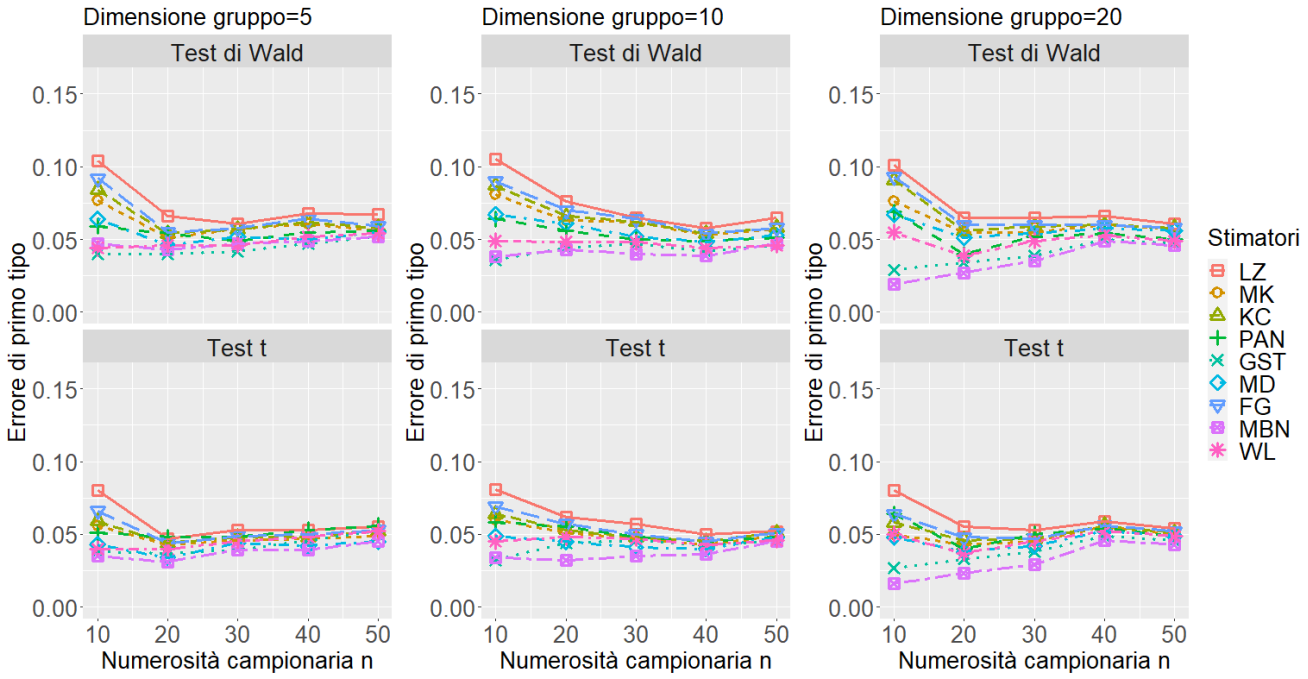

FIGURA 10: Risposta Continua, Indipendenza

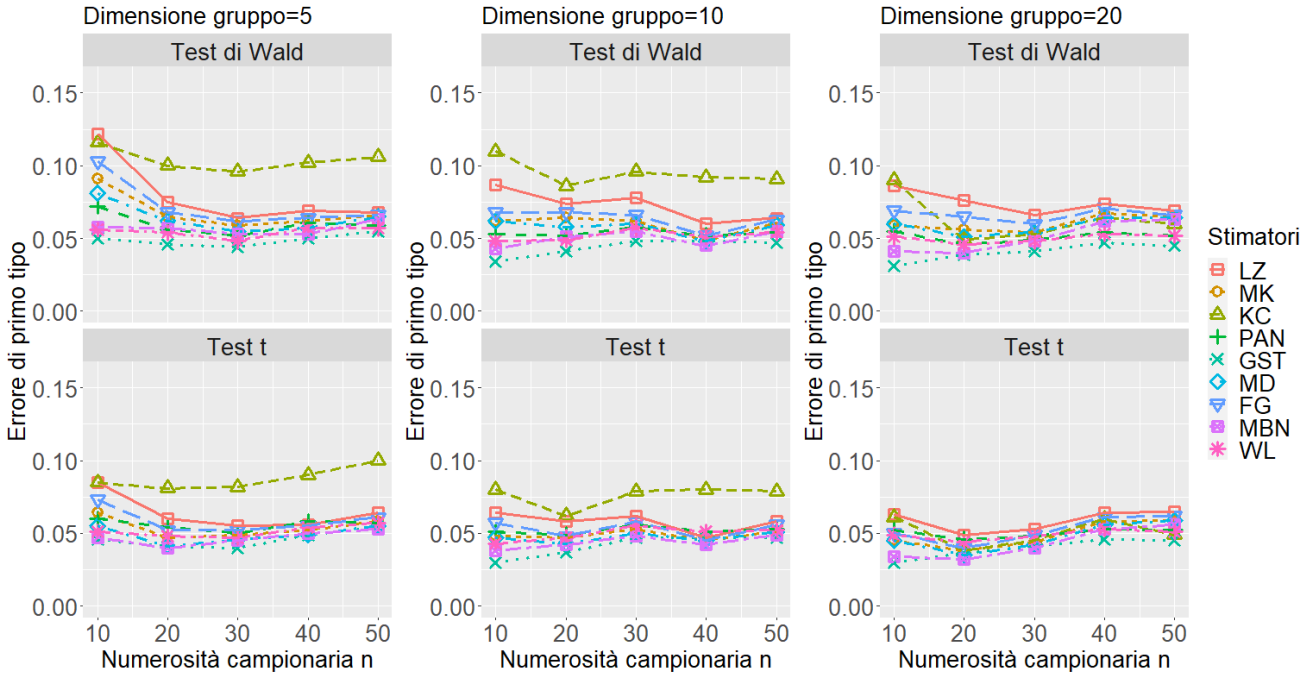

FIGURA 11: Risposta Continua, Interscambiabile (*exchangeable*)

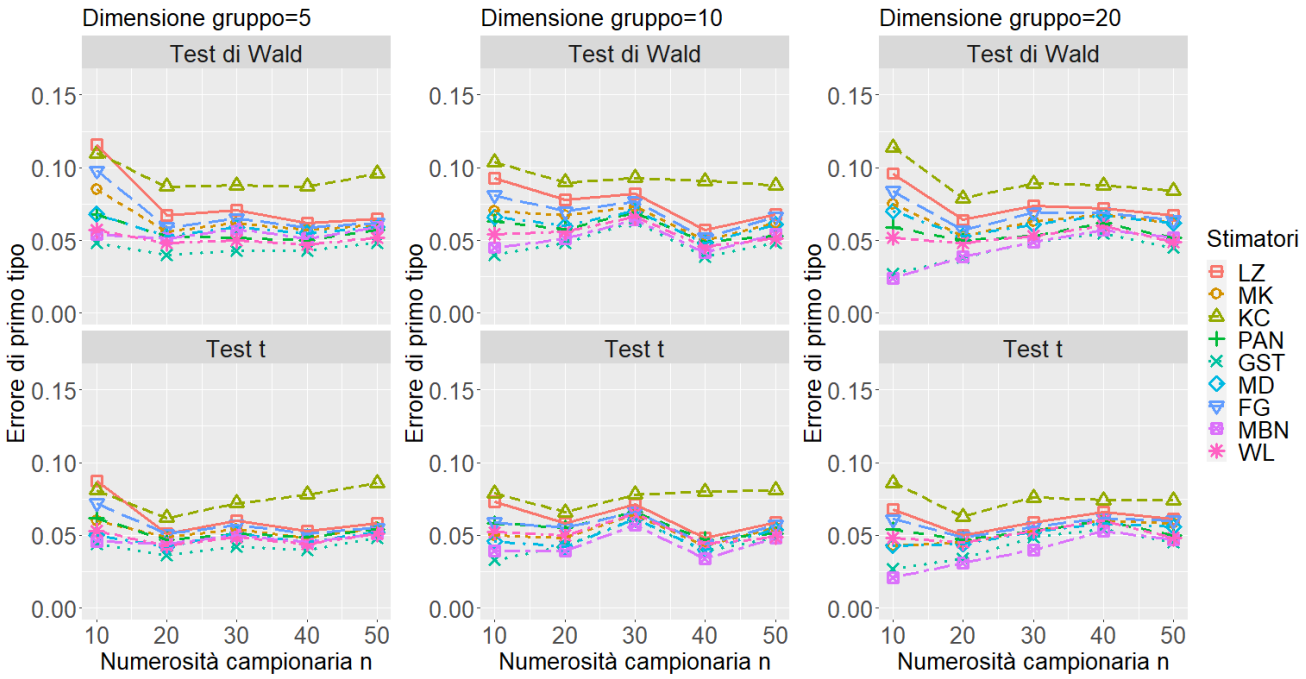

FIGURA 12: Risposta Continua, AR-1

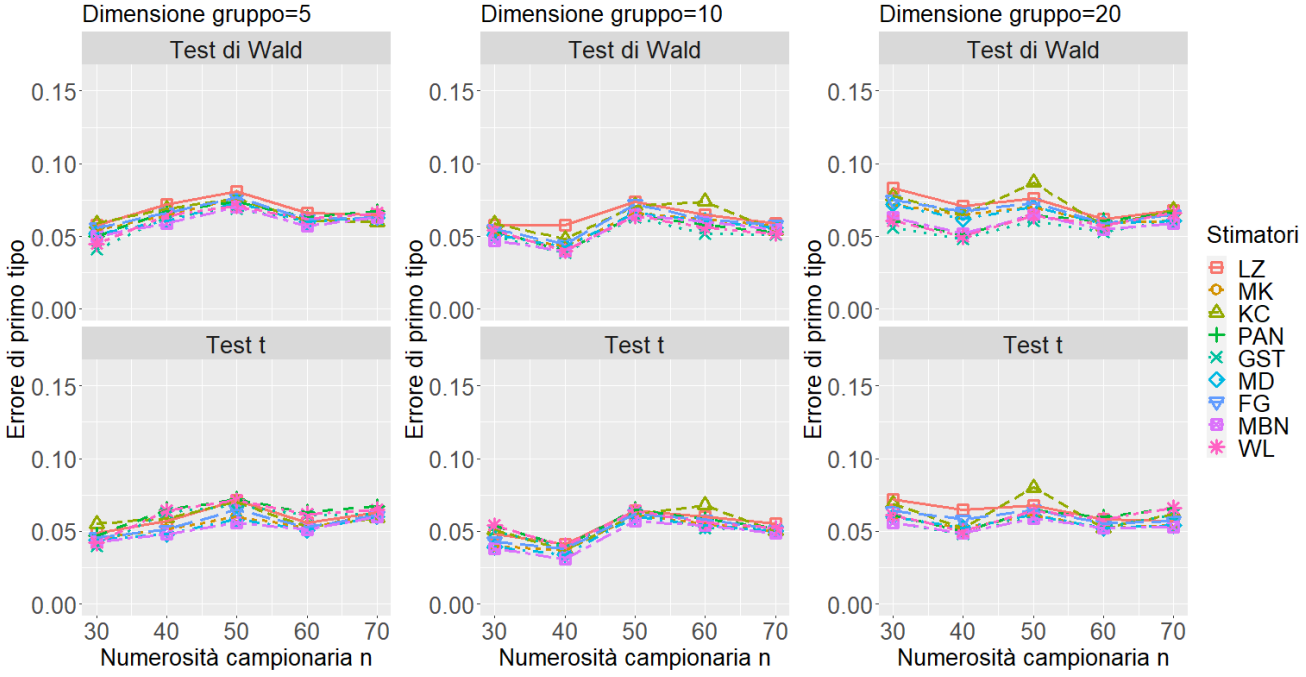

FIGURA 13: Risposta Binaria, Indipendenza

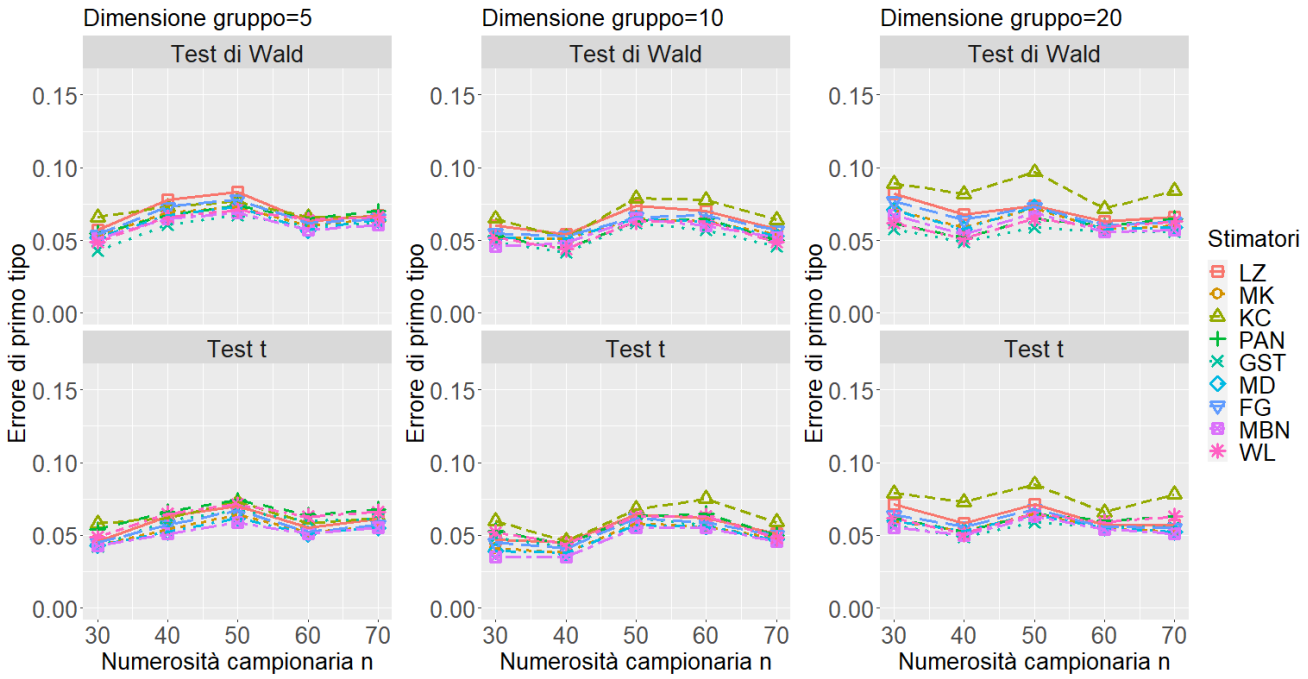

FIGURA 14: Risposta Binaria, Interscambiabile (*exchangeable*)

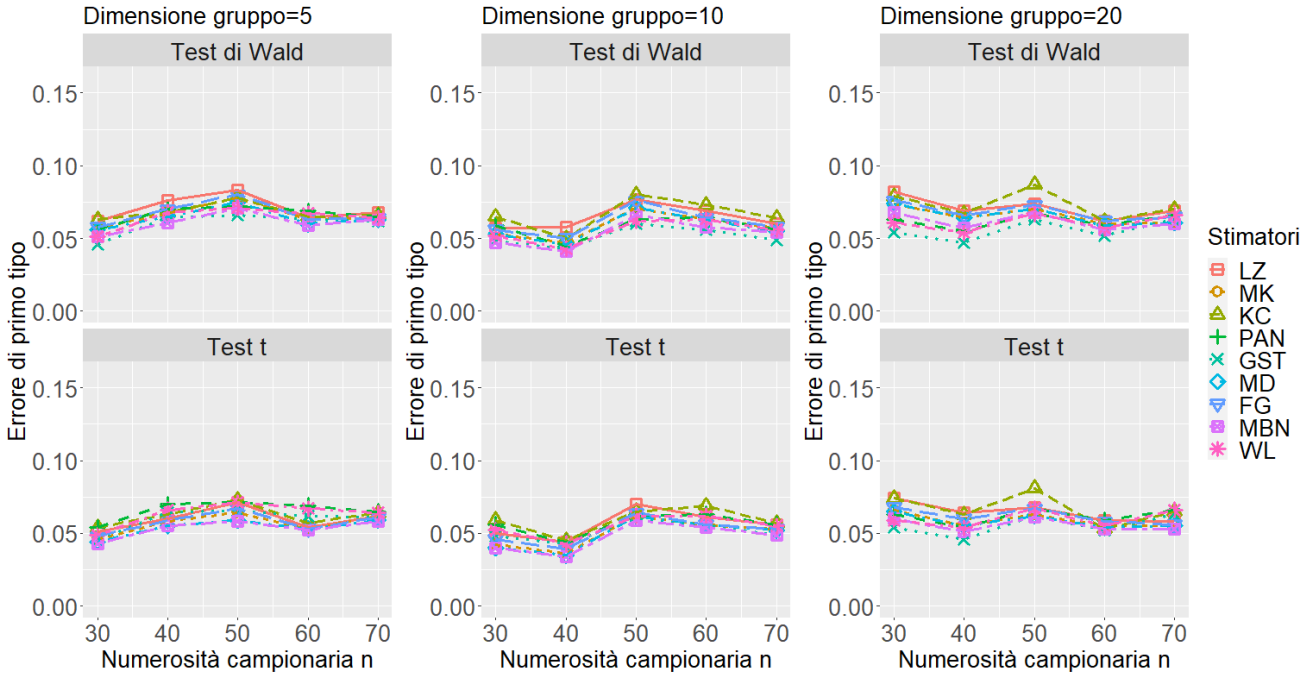

FIGURA 15: Risposta Binaria, AR-1

E Codice R

```
library(geesmv) # libreria sviluppata dagli autori dell'articolo
library(gee)
library(tidyverse)
library(patchwork) # libreria utile per fare i grafici
```

```
##### GENERAZIONE RISPOSTA CONTINUA ######

genera.exch <- function(n, m, B=1000) {
  pseudo <- matrix(NA, n*m, B)
  for(i in 1:B) {
    b <- rnorm(n,mean=0, sd=sqrt(0.25))
    mu <- b
    normale <- rnorm(m*n, mu, sd=sqrt(0.8))
    y <- matrix(normale, nrow=n)
    z <- y[1,]
    for(j in 2:n) {
      z <- c(z, y[j,])
    }
    pseudo[,i] <- z
  }
  pseudo
}

genera.exch.beta <- function(n, m, B=1000, x) {
  pseudo <- matrix(NA, n*m, B)
  beta0 <- 1
  beta1 <- 1.5
  for(i in 1:B) {
    b <- rnorm(n,mean=0, sd=sqrt(0.25))
    mu <- b
    mu.beta <- beta0 + beta1*x
    normale <- rnorm(m*n, mu, sd=sqrt(0.8))
    y <- matrix(normale, nrow=n) + matrix(mu.beta, nrow=n)
    z <- y[1,]
    for(j in 2:n) {
      z <- c(z, y[j,])
    }
    pseudo[,i] <- z
  }
  pseudo
}

###### GENERAZIONE RISPOSTA CONTEGGIO ########

genera.count.exch <- function(n, m, B=1000) {
  pseudo <- matrix(NA, n*m, B)
  for(i in 1:B) {
    b <- rnorm(n,mean=0,sd=sqrt(0.25))
    mu <- exp(b)
    poi <- rpois(m*n, mu)
    y <- matrix(poi, nrow=n)
    z <- y[1,]
    for(j in 2:n) {
      z <- c(z, y[j,])
    }
    pseudo[,i] <- z
  }
  pseudo
}

genera.count.exch.beta <- function(n, m, B=1000, x) {
  pseudo <- matrix(NA, n*m, B)
  beta0 <- 1
  beta1 <- 1.5
```

```
  for(i in 1:B) {
    b <- rnorm(n,mean=0,sd=sqrt(0.25))
    b.mat <- matrix(rep(b,each=m), nrow = n,byrow = T)
    mu <- exp(b.mat + matrix(beta0 + beta1*x, nrow=n))
    poi <- rpois(m*n, mu)
    y <- matrix(poi, nrow=n)
    z <- y[1,]
    for(j in 2:n) {
      z <- c(z, y[j,])
    }
    pseudo[,i] <- z
  }
  pseudo
}

###### GENERAZIONE RISPOSTA BINARIA ########

genera.bin.exch <- function(n, m, B=1000) {
  pseudo <- matrix(NA, n*m, B)
  for(i in 1:B) {
    b <- rnorm(n,mean=0,sd=sqrt(0.25))
    mu <- plogis(b)
    poi <- rbinom(m*n,1, mu)
    y <- matrix(poi, nrow=n)
    z <- y[1,]
    for(j in 2:n) {
      z <- c(z, y[j,])
    }
    pseudo[,i] <- z
  }
  pseudo
}

genera.bin.exch.beta <- function(n, m, B=1000,x) {
  pseudo <- matrix(NA, n*m, B)
  beta0 <- 1
  beta1 <- 1.5
  for(i in 1:B) {
    b <- rnorm(n,mean=0,sd=sqrt(0.1))
    b.mat <- matrix(rep(b,each=m), nrow = n,byrow = T)
    mu <- plogis(b.mat + matrix(beta0 + beta1*x, nrow=n))
    bin <- rbinom(m*n,1, mu)
    y <- matrix(bin, nrow=n)
    z <- y[1,]
    for(j in 2:n) {
      z <- c(z, y[j,])
    }
    pseudo[,i] <- z
  }
  pseudo
}

## funzione che restituisce gli indici delle unità
id <- function(n,m) {
  idx <- NULL
  for(i in 1:n) {
    idx <- c(idx, rep(i,m))
  }
  idx
```

```
}

#### STIMA DEI MODELLI E CALCOLO DELLE STATISTICHE TEST #####

## risposta continua

fit.cont.exch <- function(y, n, m, x) {
  idx <- id(n,m)
  wald <- matrix(NA, 9, ncol(y))
  d <- matrix(NA, 9, ncol(y))
  x <- matrix(x, nrow=n)
  z <- x[1,]
  for(j in 2:n) z <- c(z,x[j,])
  x <- z

  for(i in 1:ncol(y)) {
    simdata <- data.frame(y=y[,i], x=x, subject=idx)
    m.gee <- gee(y~x, id=subject, family=gaussian,
                 corstr="exchangeable", data=simdata)
    beta <- coefficients(m.gee)[2]

    m.lz <- GEE.var.lz(y~x, id="subject", family=gaussian,
                       corstr = "exchangeable", data=simdata)
    m.mk <- GEE.var.mk(y~x, id="subject", family=gaussian,
                       corstr = "exchangeable", data=simdata)
    m.kc <- GEE.var.kc(y~x, id="subject", family=gaussian,
                       corstr = "exchangeable", data=simdata)
    m.pan <- GEE.var.pan(y~x, id="subject", family=gaussian,
                         corstr = "exchangeable", data=simdata)
    m.gst <- GEE.var.gst(y~x, id="subject", family=gaussian,
                         corstr = "exchangeable", data=simdata)
    m.md <- GEE.var.md(y~x, id="subject", family=gaussian,
                       corstr = "exchangeable", data=simdata)
    m.fg <- GEE.var.fg(y~x, id="subject", family=gaussian,
                       corstr = "exchangeable", data=simdata)
    m.mbn <- GEE.var.mbn(y~x, id="subject", family=gaussian,
                         corstr = "exchangeable", data=simdata)
    m.wl <- GEE.var.wl(y~x, id="subject", family=gaussian,
                       corstr = "exchangeable", data=simdata)

    # estrazione della stima della varianza di \hat{\beta}_1
    var.beta <- c(m.lz$cov.beta[2], m.mk$cov.beta[2],
                  m.kc$cov.beta[2], m.pan$cov.beta[2],
                 m.gst$cov.beta[2], m.md$cov.beta[2],
                 m.fg$cov.beta[2], m.mbn$cov.beta[2],
                  m.wl$cov.beta[2])

    # estrazione della stima della varianza dello stimatore
    #della varianza di \hat{\beta}_1
    var.var.beta <- c(m.lz$cov.var[4,4], m.mk$cov.var[4,4],
                      m.kc$cov.var[4,4], m.pan$cov.var[4,4],
                      m.gst$cov.var[4,4], m.md$cov.var[4,4],
                      m.fg$cov.var[4,4], m.mbn$cov.var[4,4],
                      m.wl$cov.var[4,4])

    wald[,i] <- beta/sqrt(var.beta) # al numeratore (beta - 1.5) nel caso di
    # verifica di ipotesi H_0: beta_1=1.5 contro H_1: beta_1 \ne 1.5
    d[,i] <- 2*var.beta^2/var.var.beta
```

```
    cat("\n\n------------------------- ",i," -------------------------\n\n")
  }

  phi.w <- rep(NA,9)
  phi.t <- rep(NA, 9)
  for(i in 1:9) {
    phi.w[i] <- mean(wald[i,] <= qnorm(0.025) | wald[i,] >= qnorm(0.975))
    phi.t[i] <- mean(wald[i,] <= qt(0.025,d[i,]) | wald[i,] >= qt(0.975,d[i,]))
  }
  return(c(phi.w, phi.t))
}

fit.cont.ar1 <- function(y, n, m, x) {
  idx <- id(n,m)
  wald <- matrix(NA, 9, ncol(y))
  d <- matrix(NA, 9, ncol(y))
  x <- matrix(x, nrow=n)
  z <- x[1,]
  for(j in 2:n) z <- c(z,x[j,])
  x <- z

  for(i in 1:ncol(y)) {
    simdata <- data.frame(y=y[,i], x=x, subject=idx)
    m.gee <- gee(y~x, id=subject, family=gaussian,
                 corstr="AR-M", data=simdata)
    beta <- coefficients(m.gee)[2]
    m.lz <- GEE.var.lz(y~x, id="subject", family=gaussian,
                       corstr = "AR-M", data=simdata)
    m.mk <- GEE.var.mk(y~x, id="subject", family=gaussian,
                       corstr = "AR-M", data=simdata)
    m.kc <- GEE.var.kc(y~x, id="subject", family=gaussian,
                       corstr = "AR-M", data=simdata)
    m.pan <- GEE.var.pan(y~x, id="subject", family=gaussian,
                         corstr = "AR-M", data=simdata)
    m.gst <- GEE.var.gst(y~x, id="subject", family=gaussian,
                         corstr = "AR-M", data=simdata)
    m.md <- GEE.var.md(y~x, id="subject", family=gaussian,
                       corstr = "AR-M", data=simdata)
    m.fg <- GEE.var.fg(y~x, id="subject", family=gaussian,
                       corstr = "AR-M", data=simdata)
    m.mbn <- GEE.var.mbn(y~x, id="subject", family=gaussian,
                         corstr = "AR-M", data=simdata)
    m.wl <- GEE.var.wl(y~x, id="subject", family=gaussian,
                       corstr = "AR-M", data=simdata)

    # estrazione della stima della varianza di \hat{\beta}_1
    var.beta <- c(m.lz$cov.beta[2], m.mk$cov.beta[2],
                  m.kc$cov.beta[2], m.pan$cov.beta[2],
                  m.gst$cov.beta[2], m.md$cov.beta[2],
                  m.fg$cov.beta[2], m.mbn$cov.beta[2],
                  m.wl$cov.beta[2])

    # estrazione della stima della varianza dello stimatore
    #della varianza di \hat{\beta}_1
    var.var.beta <- c(m.lz$cov.var[4,4], m.mk$cov.var[4,4],
                      m.kc$cov.var[4,4], m.pan$cov.var[4,4],
                      m.gst$cov.var[4,4], m.md$cov.var[4,4],
                      m.fg$cov.var[4,4], m.mbn$cov.var[4,4],
                      m.wl$cov.var[4,4])
```

```
    wald[,i] <- beta/sqrt(var.beta) # al numeratore (beta - 1.5) nel caso di
    # verifica di ipotesi H_0: beta_1=1.5 contro H_1: beta_1 \ne 1.5
    d[,i] <- 2*var.beta^2/var.var.beta

    cat("\n\n------------------------- ",i," -------------------------\n\n")
  }

  phi.w <- rep(NA,9)
  phi.t <- rep(NA, 9)

  for(i in 1:9) {
    phi.w[i] <- mean(wald[i,] <= qnorm(0.025) | wald[i,] >= qnorm(0.975))
    phi.t[i] <- mean(wald[i,] <= qt(0.025,d[i,]) | wald[i,] >= qt(0.975,d[i,]))
  }
  return(c(phi.w, phi.t))
}

fit.cont.ind <- function(y, n, m, x) {
  idx <- id(n,m)
  wald <- matrix(NA, 9, ncol(y))
  d <- matrix(NA, 9, ncol(y))
  x <- matrix(x, nrow=n)
  z <- x[1,]
  for(j in 2:n) z <- c(z,x[j,])
  x <- z

  for(i in 1:ncol(y)) {
    simdata <- data.frame(y=y[,i], x=x, subject=idx)
    m.gee <- gee(y~x, id=subject, family=gaussian,
                 corstr="independence", data=simdata)
    beta <- coefficients(m.gee)[2]
    m.lz <- GEE.var.lz(y~x, id="subject", family=gaussian,
                       corstr = "independence", data=simdata)
    m.mk <- GEE.var.mk(y~x, id="subject", family=gaussian,
                       corstr = "independence", data=simdata)
    m.kc <- GEE.var.kc(y~x, id="subject", family=gaussian,
                       corstr = "independence", data=simdata)
    m.pan <- GEE.var.pan(y~x, id="subject", family=gaussian,
                         corstr = "independence", data=simdata)
    m.gst <- GEE.var.gst(y~x, id="subject", family=gaussian,
                         corstr = "independence", data=simdata)
    m.md <- GEE.var.md(y~x, id="subject", family=gaussian,
                       corstr = "independence", data=simdata)
    m.fg <- GEE.var.fg(y~x, id="subject", family=gaussian,
                       corstr = "independence", data=simdata)
    m.mbn <- GEE.var.mbn(y~x, id="subject", family=gaussian,
                         corstr = "independence", data=simdata)
    m.wl <- GEE.var.wl(y~x, id="subject", family=gaussian,
                       corstr = "independence", data=simdata)

    # estrazione della stima della varianza di \hat{\beta}_1
    var.beta <- c(m.lz$cov.beta[2], m.mk$cov.beta[2],
                  m.kc$cov.beta[2], m.pan$cov.beta[2],
                  m.gst$cov.beta[2], m.md$cov.beta[2],
                  m.fg$cov.beta[2], m.mbn$cov.beta[2],
                  m.wl$cov.beta[2])

    # estrazione della stima della varianza dello stimatore
```

```
    #della varianza di \hat{\beta}_1
    var.var.beta <- c(m.lz$cov.var[4,4], m.mk$cov.var[4,4],
                      m.kc$cov.var[4,4], m.pan$cov.var[4,4],
                      m.gst$cov.var[4,4], m.md$cov.var[4,4],
                      m.fg$cov.var[4,4], m.mbn$cov.var[4,4],
                      m.wl$cov.var[4,4])

    wald[,i] <- beta/sqrt(var.beta) # al numeratore (beta - 1.5) nel caso di
    # verifica di ipotesi H_0: beta_1=1.5 contro H_1: beta_1 \ne 1.5
    d[,i] <- 2*var.beta^2/var.var.beta

    cat("\n\n------------------------- ",i," -------------------------\n\n")
  }

  phi.w <- rep(NA,9)
  phi.t <- rep(NA, 9)

  for(i in 1:9) {
    phi.w[i] <- mean(wald[i,] <= qnorm(0.025) | wald[i,] >= qnorm(0.975))
    phi.t[i] <- mean(wald[i,] <= qt(0.025,d[i,]) | wald[i,] >= qt(0.975,d[i,]))
  }
  return(c(phi.w, phi.t))
}

## risposta conteggio

fit.count.exch <- function(y, n, m, x) {
  idx <- id(n,m)
  wald <- matrix(NA, 9, ncol(y))
  d <- matrix(NA, 9, ncol(y))
  x <- matrix(x, nrow=n)
  z <- x[1,]
  for(j in 2:n) z <- c(z,x[j,])
  x <- z

  for(i in 1:ncol(y)) {
    simdata <- data.frame(y=y[,i], x=x, subject=idx)
    m.gee <- gee(y~x, id=subject, family=poisson,
                 corstr="exchangeable", data=simdata)
    beta <- coefficients(m.gee)[2]
    beta0 <- coefficients(m.gee)[1]
    m.lz <- GEE.var.lz(y~x, id="subject", family=poisson,
                       corstr = "exchangeable", data=simdata)
    m.mk <- GEE.var.mk(y~x, id="subject", family=poisson,
                       corstr = "exchangeable", data=simdata)
    m.kc <- GEE.var.kc(y~x, id="subject", family=poisson,
                       corstr = "exchangeable", data=simdata)
    m.pan <- GEE.var.pan(y~x, id="subject", family=poisson,
                         corstr = "exchangeable", data=simdata)
    m.gst <- GEE.var.gst(y~x, id="subject", family=poisson,
                         corstr = "exchangeable", data=simdata)
    m.md <- GEE.var.md(y~x, id="subject", family=poisson,
                       corstr = "exchangeable", data=simdata)
    m.fg <- GEE.var.fg(y~x, id="subject", family=poisson,
                       corstr = "exchangeable", data=simdata)
    m.mbn <- GEE.var.mbn(y~x, id="subject", family=poisson,
                         corstr = "exchangeable", data=simdata)
    m.wl <- GEE.var.wl(y~x, id="subject", family=poisson,
                       corstr = "exchangeable", data=simdata)
```

```
    var.beta <- c(m.lz$cov.beta[2], m.mk$cov.beta[2],
                  m.kc$cov.beta[2], m.pan$cov.beta[2],
                  m.gst$cov.beta[2], m.md$cov.beta[2],
                  m.fg$cov.beta[2], m.mbn$cov.beta[2],
                  m.wl$cov.beta[2])

    var.var.beta <- c(m.lz$cov.var[4,4], m.mk$cov.var[4,4],
                      m.kc$cov.var[4,4], m.pan$cov.var[4,4],
                      m.gst$cov.var[4,4], m.md$cov.var[4,4],
                      m.fg$cov.var[4,4], m.mbn$cov.var[4,4],
                      m.wl$cov.var[4,4])

    wald[,i] <- beta/sqrt(var.beta) # al numeratore (beta - 1.5) nel caso di
    #  verifica di ipotesi H_0: beta_1=1.5 contro H_1: beta_1 \ne 1.5
    d[,i] <- 2*var.beta^2/var.var.beta

    cat("\n\n------------------------- ",i," -------------------------\n\n")
  }

  phi.w <- rep(NA,9)
  phi.t <- rep(NA, 9)

  for(i in 1:9) {
    phi.w[i] <- mean(wald[i,] <= qnorm(0.025) | wald[i,] >= qnorm(0.975))
    phi.t[i] <- mean(wald[i,] <= qt(0.025,d[i,]) | wald[i,] >= qt(0.975,d[i,]))
  }
  return(c(phi.w, phi.t))
}

fit.count.ar1 <- function(y, n, m, x) {
  idx <- id(n,m)
  wald <- matrix(NA, 9, ncol(y))
  d <- matrix(NA, 9, ncol(y))
  x <- matrix(x, nrow=n)
  z <- x[1,]
  for(j in 2:n) z <- c(z,x[j,])
  x <- z

  for(i in 1:ncol(y)) {
    simdata <- data.frame(y=y[,i], x=x, subject=idx)
    m.gee <- gee(y~x, id=subject, family=poisson,
                 corstr="AR-M", data=simdata)
    beta <- coefficients(m.gee)[2]
    m.lz <- GEE.var.lz(y~x, id="subject", family=poisson,
                       corstr = "AR-M", data=simdata)
    m.mk <- GEE.var.mk(y~x, id="subject", family=poisson,
                       corstr = "AR-M", data=simdata)
    m.kc <- GEE.var.kc(y~x, id="subject", family=poisson,
                       corstr = "AR-M", data=simdata)
    m.pan <- GEE.var.pan(y~x, id="subject", family=poisson,
                         corstr = "AR-M", data=simdata)
    m.gst <- GEE.var.gst(y~x, id="subject", family=poisson,
                         corstr = "AR-M", data=simdata)
    m.md <- GEE.var.md(y~x, id="subject", family=poisson,
                       corstr = "AR-M", data=simdata)
    m.fg <- GEE.var.fg(y~x, id="subject", family=poisson,
                       corstr = "AR-M", data=simdata)
    m.mbn <- GEE.var.mbn(y~x, id="subject", family=poisson,
```

```
                          corstr = "AR-M", data=simdata)
    m.wl <- GEE.var.wl(y~x, id="subject", family=poisson,
                       corstr = "AR-M", data=simdata)

    var.beta <- c(m.lz$cov.beta[2], m.mk$cov.beta[2],
                  m.kc$cov.beta[2], m.pan$cov.beta[2],
                  m.gst$cov.beta[2], m.md$cov.beta[2],
                  m.fg$cov.beta[2], m.mbn$cov.beta[2],
                  m.wl$cov.beta[2])

    var.var.beta <- c(m.lz$cov.var[4,4], m.mk$cov.var[4,4],
                      m.kc$cov.var[4,4], m.pan$cov.var[4,4],
                      m.gst$cov.var[4,4], m.md$cov.var[4,4],
                      m.fg$cov.var[4,4], m.mbn$cov.var[4,4],
                      m.wl$cov.var[4,4])

    wald[,i] <- beta/sqrt(var.beta) # al numeratore (beta - 1.5) nel caso di
    # verifica di ipotesi H_0: beta_1=1.5 contro H_1: beta_1 \ne 1.5
    d[,i] <- 2*var.beta^2/var.var.beta

    cat("\n\n------------------------- ",i," -------------------------\n\n")
  }

  phi.w <- rep(NA,9)
  phi.t <- rep(NA, 9)

  for(i in 1:9) {
    phi.w[i] <- mean(wald[i,] <= qnorm(0.025) | wald[i,] >= qnorm(0.975))
    phi.t[i] <- mean(wald[i,] <= qt(0.025,d[i,]) | wald[i,] >= qt(0.975,d[i,]))
  }
  return(c(phi.w, phi.t))
}

fit.count.ind <- function(y, n, m, x) {
  idx <- id(n,m)
  wald <- matrix(NA, 9, ncol(y))
  d <- matrix(NA, 9, ncol(y))
  x <- matrix(x, nrow=n)
  z <- x[1,]
  for(j in 2:n) z <- c(z,x[j,])
  x <- z

  for(i in 1:ncol(y)) {
    simdata <- data.frame(y=y[,i], x=x, subject=idx)
    m.gee <- gee(y~x, id=subject, family=poisson,
                 corstr="independence", data=simdata)
    beta <- coefficients(m.gee)[2]
    m.lz <- GEE.var.lz(y~x, id="subject", family=poisson,
                       corstr = "independence", data=simdata)
    m.mk <- GEE.var.mk(y~x, id="subject", family=poisson,
                       corstr = "independence", data=simdata)
    m.kc <- GEE.var.kc(y~x, id="subject", family=poisson,
                       corstr = "independence", data=simdata)
    m.pan <- GEE.var.pan(y~x, id="subject", family=poisson,
                         corstr = "independence", data=simdata)
    m.gst <- GEE.var.gst(y~x, id="subject", family=poisson,
                         corstr = "independence", data=simdata)
    m.md <- GEE.var.md(y~x, id="subject", family=poisson,
                       corstr = "independence", data=simdata)
```

```
    m.fg <- GEE.var.fg(y~x, id="subject", family=poisson,
                       corstr = "independence", data=simdata)
    m.mbn <- GEE.var.mbn(y~x, id="subject", family=poisson,
                         corstr = "independence", data=simdata)
    m.wl <- GEE.var.wl(y~x, id="subject", family=poisson,
                       corstr = "independence", data=simdata)

    var.beta <- c(m.lz$cov.beta[2], m.mk$cov.beta[2],
                  m.kc$cov.beta[2], m.pan$cov.beta[2],
                  m.gst$cov.beta[2], m.md$cov.beta[2],
                  m.fg$cov.beta[2], m.mbn$cov.beta[2],
                  m.wl$cov.beta[2])

    var.var.beta <- c(m.lz$cov.var[4,4], m.mk$cov.var[4,4],
                      m.kc$cov.var[4,4], m.pan$cov.var[4,4],
                      m.gst$cov.var[4,4], m.md$cov.var[4,4],
                      m.fg$cov.var[4,4], m.mbn$cov.var[4,4],
                      m.wl$cov.var[4,4])

    wald[,i] <- beta/sqrt(var.beta) # al numeratore (beta - 1.5) nel caso di
    #  verifica di ipotesi H_0: beta_1=1.5 contro H_1: beta_1 \ne 1.5
    d[,i] <- 2*var.beta^2/var.var.beta

    cat("\n\n------------------------ ",i," ------------------------\n\n")
  }

  phi.w <- rep(NA,9)
  phi.t <- rep(NA, 9)

  for(i in 1:9) {
    phi.w[i] <- mean(wald[i,] <= qnorm(0.025) | wald[i,] >= qnorm(0.975))
    phi.t[i] <- mean(wald[i,] <= qt(0.025,d[i,]) | wald[i,] >= qt(0.975,d[i,]))
  }
  return(c(phi.w, phi.t))
}

## risposta binaria

fit.bin.exch <- function(y, n, m, x) {
  idx <- id(n,m)
  wald <- matrix(NA, 9, ncol(y))
  d <- matrix(NA, 9, ncol(y))
  x <- matrix(x, nrow=n)
  z <- x[1,]
  for(j in 2:n) z <- c(z,x[j,])
  x <- z

  for(i in 1:ncol(y)) {
    simdata <- data.frame(y=y[,i], x=x, subject=idx)
    m.gee <- gee(y~x, id=subject, family=binomial,
                 corstr="exchangeable", data=simdata)
    beta <- coefficients(m.gee)[2]
    beta0 <- coefficients(m.gee)[1]
    m.lz <- GEE.var.lz(y~x, id="subject", family=binomial,
                       corstr = "exchangeable", data=simdata)
    m.mk <- GEE.var.mk(y~x, id="subject", family=binomial,
                       corstr = "exchangeable", data=simdata)
    m.kc <- GEE.var.kc(y~x, id="subject", family=binomial,
                       corstr = "exchangeable", data=simdata)
```

```
    m.pan <- GEE.var.pan(y~x, id="subject", family=binomial,
                         corstr = "exchangeable", data=simdata)
    m.gst <- GEE.var.gst(y~x, id="subject", family=binomial,
                         corstr = "exchangeable", data=simdata)
    m.md <- GEE.var.md(y~x, id="subject", family=binomial,
                       corstr = "exchangeable", data=simdata)
    m.fg <- GEE.var.fg(y~x, id="subject", family=binomial,
                       corstr = "exchangeable", data=simdata)
    m.mbn <- GEE.var.mbn(y~x, id="subject", family=binomial,
                         corstr = "exchangeable", data=simdata)
    m.wl <- GEE.var.wl(y~x, id="subject", family=binomial,
                       corstr = "exchangeable", data=simdata)

    var.beta <- c(m.lz$cov.beta[2], m.mk$cov.beta[2],
                  m.kc$cov.beta[2], m.pan$cov.beta[2],
                  m.gst$cov.beta[2], m.md$cov.beta[2],
                  m.fg$cov.beta[2], m.mbn$cov.beta[2],
                  m.wl$cov.beta[2])

    var.var.beta <- c(m.lz$cov.var[4,4], m.mk$cov.var[4,4],
                      m.kc$cov.var[4,4], m.pan$cov.var[4,4],
                      m.gst$cov.var[4,4], m.md$cov.var[4,4],
                      m.fg$cov.var[4,4], m.mbn$cov.var[4,4],
                      m.wl$cov.var[4,4])

    wald[,i] <- beta/sqrt(var.beta) # al numeratore (beta - 1.5) nel caso di
    #  verifica di ipotesi H_0: beta_1=1.5 contro H_1: beta_1 \ne 1.5
    d[,i] <- 2*var.beta^2/var.var.beta

    cat("\n\n------------------------- ",i," -------------------------\n\n")
  }

  phi.w <- rep(NA,9)
  phi.t <- rep(NA, 9)

  for(i in 1:9) {
    phi.w[i] <- mean(wald[i,] <= qnorm(0.025) | wald[i,] >= qnorm(0.975))
    phi.t[i] <- mean(wald[i,] <= qt(0.025,d[i,]) | wald[i,] >= qt(0.975,d[i,]))
  }
  return(c(phi.w, phi.t))
}

fit.bin.ar1 <- function(y, n, m, x) {
  idx <- id(n,m)
  wald <- matrix(NA, 9, ncol(y))
  d <- matrix(NA, 9, ncol(y))
  x <- matrix(x, nrow=n)
  z <- x[1,]
  for(j in 2:n) z <- c(z,x[j,])
  x <- z

  for(i in 1:ncol(y)) {
    simdata <- data.frame(y=y[,i], x=x, subject=idx)
    m.gee <- gee(y~x, id=subject, family=binomial,
                 corstr="AR-M", data=simdata)
    beta <- coefficients(m.gee)[2]
    m.lz <- GEE.var.lz(y~x, id="subject", family=binomial,
                       corstr = "AR-M", data=simdata)
    m.mk <- GEE.var.mk(y~x, id="subject", family=binomial,
```

```
                        corstr = "AR-M", data=simdata)
    m.kc <- GEE.var.kc(y~x, id="subject", family=binomial,
                        corstr = "AR-M", data=simdata)
    m.pan <- GEE.var.pan(y~x, id="subject", family=binomial,
                         corstr = "AR-M", data=simdata)
    m.gst <- GEE.var.gst(y~x, id="subject", family=binomial,
                         corstr = "AR-M", data=simdata)
    m.md <- GEE.var.md(y~x, id="subject", family=binomial,
                        corstr = "AR-M", data=simdata)
    m.fg <- GEE.var.fg(y~x, id="subject", family=binomial,
                        corstr = "AR-M", data=simdata)
    m.mbn <- GEE.var.mbn(y~x, id="subject", family=binomial,
                         corstr = "AR-M", data=simdata)
    m.wl <- GEE.var.wl(y~x, id="subject", family=binomial,
                        corstr = "AR-M", data=simdata)

    var.beta <- c(m.lz$cov.beta[2], m.mk$cov.beta[2],
                  m.kc$cov.beta[2], m.pan$cov.beta[2],
                  m.gst$cov.beta[2], m.md$cov.beta[2],
                  m.fg$cov.beta[2], m.mbn$cov.beta[2],
                  m.wl$cov.beta[2])

    var.var.beta <- c(m.lz$cov.var[4,4], m.mk$cov.var[4,4],
                      m.kc$cov.var[4,4], m.pan$cov.var[4,4],
                      m.gst$cov.var[4,4], m.md$cov.var[4,4],
                      m.fg$cov.var[4,4], m.mbn$cov.var[4,4],
                      m.wl$cov.var[4,4])

    wald[,i] <- beta/sqrt(var.beta) # al numeratore (beta - 1.5) nel caso di
    #  verifica di ipotesi H_0: beta_1=1.5 contro H_1: beta_1 \ne 1.5
    d[,i] <- 2*var.beta^2/var.var.beta

    cat("\n\n------------------------- ",i," -------------------------\n\n")
  }

  phi.w <- rep(NA,9)
  phi.t <- rep(NA, 9)

  for(i in 1:9) {
    phi.w[i] <- mean(wald[i,] <= qnorm(0.025) | wald[i,] >= qnorm(0.975))
    phi.t[i] <- mean(wald[i,] <= qt(0.025,d[i,]) | wald[i,] >= qt(0.975,d[i,]))
  }
  return(c(phi.w, phi.t))
}

fit.bin.ind <- function(y, n, m, x) {
  idx <- id(n,m)
  wald <- matrix(NA, 9, ncol(y))
  d <- matrix(NA, 9, ncol(y))
  x <- matrix(x, nrow=n)
  z <- x[1,]
  for(j in 2:n) z <- c(z,x[j,])
  x <- z

  for(i in 1:ncol(y)) {
    simdata <- data.frame(y=y[,i], x=x, subject=idx)
    m.gee <- gee(y~x, id=subject, family=binomial,
                 corstr="independence", data=simdata)
    beta <- coefficients(m.gee)[2]
```

```
    m.lz <- GEE.var.lz(y~x, id="subject", family=binomial,
                       corstr = "independence", data=simdata)
    m.mk <- GEE.var.mk(y~x, id="subject", family=binomial,
                       corstr = "independence", data=simdata)
    m.kc <- GEE.var.kc(y~x, id="subject", family=binomial,
                       corstr = "independence", data=simdata)
    m.pan <- GEE.var.pan(y~x, id="subject", family=binomial,
                         corstr = "independence", data=simdata)
    m.gst <- GEE.var.gst(y~x, id="subject", family=binomial,
                         corstr = "independence", data=simdata)
    m.md <- GEE.var.md(y~x, id="subject", family=binomial,
                       corstr = "independence", data=simdata)
    m.fg <- GEE.var.fg(y~x, id="subject", family=binomial,
                       corstr = "independence", data=simdata)
    m.mbn <- GEE.var.mbn(y~x, id="subject", family=binomial,
                         corstr = "independence", data=simdata)
    m.wl <- GEE.var.wl(y~x, id="subject", family=binomial,
                       corstr = "independence", data=simdata)

    var.beta <- c(m.lz$cov.beta[2], m.mk$cov.beta[2],
                  m.kc$cov.beta[2], m.pan$cov.beta[2],
                  m.gst$cov.beta[2], m.md$cov.beta[2],
                  m.fg$cov.beta[2], m.mbn$cov.beta[2],
                  m.wl$cov.beta[2])

    var.var.beta <- c(m.lz$cov.var[4,4], m.mk$cov.var[4,4],
                      m.kc$cov.var[4,4], m.pan$cov.var[4,4],
                      m.gst$cov.var[4,4], m.md$cov.var[4,4],
                      m.fg$cov.var[4,4], m.mbn$cov.var[4,4],
                      m.wl$cov.var[4,4])

    wald[,i] <- beta/sqrt(var.beta) # al numeratore (beta - 1.5) nel caso di
    # verifica di ipotesi H_0: beta_1=1.5 contro H_1: beta_1 \ne 1.5
    d[,i] <- 2*var.beta^2/var.var.beta

    cat("\n\n------------------------ ",i," ------------------------\n\n")
  }

  phi.w <- rep(NA,9)
  phi.t <- rep(NA, 9)

  for(i in 1:9) {
    phi.w[i] <- mean(wald[i,] <= qnorm(0.025) | wald[i,] >= qnorm(0.975))
    phi.t[i] <- mean(wald[i,] <= qt(0.025,d[i,]) | wald[i,] >= qt(0.975,d[i,]))
  }
  return(c(phi.w, phi.t))
}

########## APPLICAZIONE ##########

app <- function(genera, m, fit) {
  n <-  c(10,20,30,40,50) # n <- (30,40,50,60,70) nel caso
  # di risposta binaria e coefficienti di regressione diversi da zero
  set.seed(123456)
  phi <- as.vector(sapply(n, function(k) {
    x <- rnorm(k*m)
    fit(genera(k,m),k,m,x) # genera(k,m, x) nel caso di verifica di ipotesi
    #H_0: beta_1=1.5 contro l'alternativa H_1: beta_1 \ne 1.5
  }))
```

```
  return(phi)
}


phi.cont.exch <- matrix(NA, 90, 3)
phi.cont.ar1 <- matrix(NA, 90, 3)
phi.cont.ind <- matrix(NA, 90, 3)
phi.count.exch <- matrix(NA, 90, 3)
phi.count.ar1 <- matrix(NA, 90, 3)
phi.count.ind <- matrix(NA, 90, 3)
phi.bin.exch <- matrix(NA, 90, 3)
phi.bin.ar1 <- matrix(NA, 90, 3)
phi.bin.ind <- matrix(NA, 90, 3)

m<- c(5,10,20)
done <- FALSE
repeat {
  for(i in 1:length(m)) phi.cont.exch[,i] <- app(genera.exch,
                                                 m[i], fit.cont.exch)
  for(i in 1:length(m)) phi.cont.ar1[,i] <- app(genera.exch,
                                                m[i], fit.cont.ar1)
  for(i in 1:length(m)) phi.cont.ind[,i] <- app(genera.exch,
                                                m[i], fit.cont.ind)
  for(i in 1:length(m)) phi.cont.exch[,i] <- app(genera.count.exch,
                                                 m[i], fit.count.exch)
  for(i in 1:length(m)) phi.cont.ar1[,i] <- app(genera.count.exch,
                                                m[i], fit.count.ar1)
  for(i in 1:length(m)) phi.cont.ind[,i] <- app(genera.count.exch,
                                                m[i], fit.count.ind)
  for(i in 1:length(m)) phi.cont.exch[,i] <- app(genera.bin.exch,
                                                 m[i], fit.bin.exch)
  for(i in 1:length(m)) phi.cont.ar1[,i] <- app(genera.bin.exch,
                                                m[i], fit.bin.ar1)
  for(i in 1:length(m)) phi.cont.ind[,i] <- app(genera.bin.exch,
                                                m[i], fit.bin.ind)
  done <- TRUE
  if(done) break
}
# nel caso in cui si considerano i coefficienti di regressione diversi da zero,
# ed è quindi necessario verificare l'ipotesi H_0: beta_1=1.5
# contro l'alternativa H_1: beta_1 \ne 1.5,
# allora è sufficiente sostituire le funzioni di generazione
# con le funzioni di generazione che terminano con .beta

################ GRAFICI ######################

# per ottenere i grafici nel caso di risposta di tipo conteggio
# e binaria, è sufficiente sostituire phi.cont.exch con
# rispettivamente phi.count.exch e phi.bin.exch

mat <- matrix(phi.cont.exch[,1],18,5)
phi.b <- as.vector(mat)

d <- tibble(Phi=phi.b, X=x,Stimatori=est, Test=test)
bilat5 <- ggplot(d, aes(x=X,y=Phi,group=Stimatori,color=Stimatori)) +
  geom_point(aes(shape=Stimatori),show.legend = FALSE, size=3, stroke=1.8) +
  geom_line(aes(linetype=Stimatori), show.legend = FALSE, size=1.1) +
  facet_wrap(~Test, nrow=2) +
  coord_cartesian(ylim=c(0,0.16)) +
```

```
  scale_shape_manual(values=seq(0,9)) +
  labs(title = "Dimensione gruppo=5", x="Numerosità campionaria n",
       y="Errore di primo tipo") +
  theme(legend.text = element_text(size=20),
        legend.title = element_text(size=20),
        axis.title = element_text(size=20),
        axis.text = element_text(size=20),
        plot.title = element_text(size=20),
        strip.text = element_text(size=20))
bilat5

mat <- matrix(phi.cont.exch[,2],18,5)
phi.b <- as.vector(mat)

d <- tibble(Phi=phi.b, X=x,Stimatori=est, Test=test)
bilat10 <- ggplot(d, aes(x=X,y=Phi,group=Stimatori,color=Stimatori)) +
  geom_point(aes(shape=Stimatori),show.legend = FALSE, size=3, stroke=1.8) +
  geom_line(aes(linetype=Stimatori), show.legend = FALSE, size=1.1) +
  facet_wrap(~Test, nrow=2) +
  coord_cartesian(ylim=c(0,0.16)) +
  scale_shape_manual(values=seq(0,9)) +
  labs(title = "Dimensione gruppo=10", x="Numerosità campionaria n",
       y="Errore di primo tipo") +
  theme(legend.text = element_text(size=20),
        legend.title = element_text(size=20),
        axis.title = element_text(size=20),
        axis.text = element_text(size=20),
        plot.title = element_text(size=20),
        strip.text = element_text(size=20))
bilat10

mat <- matrix(phi.cont.exch[,3],18,5)
phi.b <- as.vector(mat)

d <- tibble(Phi=phi.b, X=x,Stimatori=est, Test=test)
bilat20 <- ggplot(d, aes(x=X,y=Phi,group=Stimatori,color=Stimatori)) +
  geom_point(aes(shape=Stimatori), size=3, stroke=1.8) +
  geom_line(aes(linetype=Stimatori), size=1.1) +
  facet_wrap(~Test, nrow=2) +
  coord_cartesian(ylim=c(0,0.16)) +
  scale_shape_manual(values=seq(0,9)) +
  labs(title = "Dimensione gruppo=20", x="Numerosità campionaria n",
       y="Errore di primo tipo") +
  theme(legend.text = element_text(size=20),
        legend.title = element_text(size=20),
        axis.title = element_text(size=20),
        axis.text = element_text(size=20),
        plot.title = element_text(size=20),
        strip.text = element_text(size=20))
bilat20

# permette l'inserimento dei grafici in unica finestra,
# come riportati nella relazione
bilat5 + bilat10 + bilat20
```

Bibliografia

da Silva, J. L. P. & Colosimo, E. A. (2016). Comments on 'covariance estimators for generalized estimating equations (GEE) in longitudinal analysis with small samples'. *Statistics in Medicine* **35**, 5315–5317.

Fay, M. P. & Graubard, B. I. (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics* **57**, 1198–1206.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London A* **222**, 309–368.

Gosho, M., Sato, Y. & Takeuchi, H. (2014). Robust covariance estimator for small-sample adjustment in the generalized estimating equations: a simulation study. *Science Journal of Applied Mathematics and Statistics* **2**, 20–25.

Guo, X., Pan, W., Connett, J. E., Hannan, P. J. & French, S. A. (2005). Small-sample performance of the robust score test and its modifications in generalized estimating equations. *Statistics in Medicine* **24**, 3479–3495.

Kauermann, G. & Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* **96**, 1387–1398.

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

MacKinnon, J. G. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics* **29**, 305–325.

Mancl, L. A. & DeRouen, T. A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics* **57**, 126–134.

Morel, J. G., Bokossa, M. C. & Neerchal, N. K. (2003). Small sample correction for the variance of GEE estimators. *Biometrical Journal* **45**, 395–409.

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society A* **135**, 370–384.

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

Pan, W. (2001). On the robust variance estimator in generalized estimating equations. *Biometrika* **88**, 901–906.

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

Wang, M., Kong, L., Li, Z. & Zhang, L. (2016a). Covariance estimators for generalized estimating equations (GEE) in longitudinal analysis with small samples. *Statistics in Medicine* **35**, 1706–1721.

Wang, M., Kong, L., Li, Z. & Zhang, L. (2016b). Authors' reply: Covariance estimators for generalized estimating equations (GEE) in longitudinal analysis with small samples. *Statistics in Medicine* **35**, 5318–5319.

Wang, M. & Long, Q. (2011). Modified robust variance estimator for generalized estimating equations with improved small-sample performance. *Statistics in Medicine* **30**, 1278–1291.