Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

Statistica per le Tecnologie e le Scienze

RELAZIONE FINALE

REGRESSIONE PER DATI DI CONTEGGIO CON INDICI DI CENSURA MANCANTI

Relatore Prof. Matteo Grigoletto
Dipartimento di Scienze Statistiche

Laureando: Nicholas Piotto
Matricola N. 2073691

Anno Accademico 25/26

ABSTRACT

Il presente lavoro si pone l'obiettivo di analizzare diverse metodologie per l'inferenza statistica sui dati di conteggio nel contesto del modello di regressione di Poisson, con particolare attenzione alla problematica degli indici di censura mancanti. Partendo dallo studio di Bousselmi e Dupuy (2021), vengono presentati e confrontati tre approcci per la gestione dei dati mancanti: *regression calibration*, *multiple imputation* e *augmented inverse probability weighting*. Tramite l'approfondimento teorico, la simulazione su un dataset e l'applicazione a un dataset reale, si evidenziano le differenze, i vantaggi e i limiti dei diversi metodi.

Indice

INTRODUZIONE

Nel presente lavoro si analizzano alcuni metodi per l'inferenza statistica di dati di conteggio, con indici di censura mancanti, nel modello di regressione di Poisson. Si considera come riferimento l'articolo [1], *Censored count data regression with missing censoring information* (di Bousselmi e Dupuy, 2021), nel quale si sviluppano alcuni metodi per risolvere il problema degli indici di censura mancanti, proponendo i rispettivi stimatori e dimostrando le loro proprietà.

Il modello di regressione preso in esame studia la relazione tra la variabile di conteggio (variabile risposta) e l'insieme delle covariate (variabili predittive). Le osservazioni della variabile di conteggio appartengono all'insieme dei numeri naturali, in quanto generate da una variabile discreta; si assume che esse seguano una distribuzione di Poisson di parametro λ, dipendente dalle variabili predittive. Tali osservazioni possono essere censurate a destra: in questo caso si riscontrano delle complicazioni analitiche, poiché la presenza di un dato censurato indica che il valore reale è superiore a quello osservato. Ad esempio, in uno studio sulle abitudini dei fumatori, se una risposta è "almeno 20 sigarette", il dato risulta censurato al valore 20, sebbene il valore reale sia pari o superiore. Ignorare la censura produce stime distorte e inferenze errate; pertanto, si introduce la variabile indicatrice di censura δ, che assume valore 1 se il dato è osservato e 0 altrimenti.

Un ulteriore elemento di disturbo è rappresentato dalla presenza di dati incompleti, ovvero dall'assenza di osservazioni di una o più variabili su diversi soggetti. La gestione, l'analisi e l'inferenza in presenza di dati mancanti hanno generato una vasta letteratura statistica; a tal proposito, Dupuy [3] identifica tra le cause principali il malfunzionamento degli strumenti di misura, la perdita accidentale dei dati, gli errori di inserimento e le mancate risposte. I tre problemi critici derivanti dall'incompletezza dei dati sono la riduzione dell'efficienza, la complessità computazionale nell'analisi del dataset e l'introduzione di distorsioni sistematiche.

In questo lavoro, nel secondo capitolo, si illustra il modello preso in considerazione, evidenziandone le caratteristiche e le problematicità che si possono riscontrare. Nel terzo capitolo si illustrano tre metodi per affrontare il problema degli indici di censura mancanti: *regression calibration*, *multiple imputation* e *augmented inverse probability weighting*. Poiché la scelta dell'approccio analitico è determinante per l'affidabilità delle stime, si rimanda a [2, 3, 5] per un approfondimento sulla *regression calibration*, a [3, 4] per la *multiple imputation* e a [3, 6] per *l'augmented inverse probability weighting*. Nel quarto capitolo si procede al confronto degli stimatori con l'aiuto di un dataset simulato. Nel quinto capitolo si applicano i diversi metodi ad un caso di studio reale. Il sesto capitolo raccoglie le considerazioni finali emerse dalla ricerca.

IL MODELLO DI REGRESSIONE POISSON

Si presentano di seguito il modello, le variabili coinvolte, i parametri considerati e le rispettive notazioni.

Si rappresenta con Y la variabile risposta di interesse e con $X = \left(1, X_2, \dots, X_p\right)^T$ la costruzione della matrice di $p-1$ variabili esplicative. L'inclusione iniziale di un vettore di valori unitari (1) serve in fase di stima a calcolare l'intercetta, media di un caso base, in modo tale da poter valutare l'effetto delle singole modalità delle altre variabili esplicative. La prima assunzione del modello indica che, data la matrice di covariate X, la distribuzione condizionata di Y è generata da un modello di regressione di Poisson con parametro $\lambda = \exp(\beta^{\mathrm{T}} X)$. L'espressione $\beta^{\mathrm{T}} X$ rappresenta il predittore lineare del modello, che può includere trasformazioni o effetti interattivi tra le variabili originali, come $X \circ X$ e $\ln(X)$. Date n osservazioni e gli indici $j = 1, \dots, p$ (colonne di X) e $i = 1, \dots, n$ (righe di X), ognuno dei coefficienti β_j del vettore β è interpretato nel seguente modo: se la variabile esplicativa X_{ij} aumenta di una unità, a parità di tutte le altre variabili, allora la media di Y_i è moltiplicata per il termine e^{β_j}. Il segno di β_j determina, quindi, se la media di Y_i aumenta (se $\beta_j > 0$), o diminuisce (se $\beta_j < 0$).

La variabile Y può essere censurata a destra, quindi per alcuni soggetti è possibile osservare un valore minore rispetto al valore reale. Si introduce la variabile di censura C che, in termini generali, si considera come una funzione costante, il cui valore rappresenta la soglia superiore della variabile di interesse osservata. È ragionevole indicare la variabile risposta osservata come Y_i^* data da

$$Y_i^* = \min(Y_i, C_i).$$

Di conseguenza, per ogni individuo i si considerano due variabili, Y_i e C_i, ma l'osservazione di una non consente l'osservazione dell'altra. Se C_i è minore di Y_i allora C_i è osservata e Y_i è censurata, in caso contrario Y_i è osservata. Come accennato nell'introduzione, per conoscere la natura dell'i-esimo dato osservato, si introduce la variabile indice di censura δ_i come segue:

$$\delta_i = \begin{cases} 1, & Y_i < C_i \\ 0, & Y_i \geq C_i \end{cases} \quad \xrightarrow{allora} \quad Y^* = \begin{cases} C, & \delta = 1 \\ Y, & \delta = 0 \end{cases}.$$

In questo modo se $\delta_i = 0$, ovvero $Y_i \geq C_i$ e $Y_i^* = C_i$, si osserva la variabile di interesse censurata. Se invece $\delta_i = 1$, ovvero $Y_i < C_i$ e $Y_i^* = Y_i$, si conosce l'effettivo valore della realizzazione della variabile di interesse, che è interrotta da eventi esterni indipendenti dalla variabile C_i che, di conseguenza, non incide sul valore di Y_i.

In presenza di censura, vorremmo risultati dell'inferenza che siano vicini a quelli che si otterrebbero in caso di dati completi con assenza di censura. Risulta utile ipotizzare la censura non informativa, valida se le covariate sono stocasticamente indipendenti dalla variabile C. Per queste caratteristiche è necessario che la causa della censura sia dovuta alla conclusione del periodo di osservazione, oppure per cause che non dipendono dallo studio stesso, dall'evento di interesse o da eventi competitivi. Ne derivano le seguenti assunzioni per il modello; l'indipendenza condizionata di Y e C dato X e l'indipendenza della distribuzione di C da β.

Dati n soggetti indipendenti, si osserva per ognuno di essi (Y_i^*, X_i, δ_i), con $i \in \{1, \dots, n\}$. Il contributo alla funzione di verosimiglianza dato dalle osservazioni non censurate si può scrivere come: $P(Y_i = y_i^*, C_i > y_i^* | X_i)$, caso in cui è presente l'informazione del soggetto i non censurata, con Y_i che assume il valore di y_i^* e C_i è strettamente maggiore di y_i^*. Le osservazioni censurate, invece, forniscono un contributo pari a $P(Y_i \geq y_i^*, C_i = y_i^* | X_i)$. Si riporta la funzione di log-verosimiglianza:

$$l_n(\beta) = \sum_{i=1}^{n} \left\{ \delta_i \left(Y_i^* \beta^{\mathrm{T}} X_i - e^{\beta^{\mathrm{T}} X_i} - \log(Y_i^*!) \right) + (1 - \delta_i) \log \left(1 - \sum_{k=0}^{Y_i^* - 1} \frac{e^{-e^{\beta^{\mathrm{T}} X_i} + k\beta^{\mathrm{T}} X_i}}{k!} \right) \right\}.$$

Si fa presente che il primo addendo, all'interno delle parentesi della sommatoria, moltiplicato per un fattore δ_i, è associato al contributo delle osservazioni non censurate, mentre il secondo termine, moltiplicato per un fattore $(1 - \delta_i)$, è associato al contributo delle osservazioni censurate.

Si ottiene la stima di β tramite la massimizzazione della funzione di log-verosimiglianza, ottenendo lo stimatore $\hat{\beta}_n = \arg\max_{\beta} l_n(\beta)$. Lo stimatore risulta consistente e di distribuzione asintoticamente normale. Quando n è sufficientemente grande, $\hat{\beta}_n \xrightarrow{d} \mathrm{N}\left(\beta, \mathrm{I}_n\left(\hat{\beta}_n\right)^{-1}\right)$, con media il vettore dei coefficienti reali β e varianza $\mathrm{I}_n\left(\hat{\beta}_n\right)^{-1} = \frac{-\partial^2 l_n(\beta)}{\partial\beta\partial\beta^{\mathrm{T}}}$.

A causa di molteplici motivi, si può presentare la mancanza di alcuni dati. Si considera il caso in cui δ_i sia ignoto per uno o più soggetti. È necessario considerare una ulteriore variabile che indica se si conosce o meno l'indice di censura del soggetto i-esimo. In particolare, si definisce ξ_i come:

$$\xi_i = \begin{cases} 1, & se\ \delta_i \text{ è osservata} \\ 0, & se\ \delta_i \text{ è mancante} \end{cases}.$$

La notazione di δ_i e ξ_i risulta efficace in quanto il loro prodotto $\delta_i * \xi_i$ assume valore 1 solo se si conosce l'indice di censura e il dato non è censurato, e quindi sicuramente $Y_i = y_i^*$, mentre assume valore 0 nei casi in cui il valore è censurato oppure l'indice di censura è mancante.

Si assume che la natura di questa mancanza sia di tipo casuale e generata da un meccanismo di tipo MAR, *missing at random*. La probabilità di non conoscere δ_i non è uguale per tutti i soggetti, ma dipende unicamente dalle informazioni osservate. Ne deriva l'indipendenza condizionata di ξ e δ dato X. Questa assunzione è essenziale perché permette di ricostruire la distribuzione mancante degli indici di censura utilizzando i dati osservati nello studio. Infatti, a questo scopo si utilizza il valore atteso condizionato dell'indice di censura, date le osservazioni $\mathbb{E}[\delta_i|W_i]$, dove $W_i = \left(Y_i^*, X_i^{\mathrm{T}}, V_i^{\mathrm{T}}\right)^{\mathrm{T}}$ racchiude l'informazione della variabile d'interesse osservata, i valori delle covariate del soggetto i-esimo e le rispettive eventuali variabili surrogate V_i^{T} per δ_i. Quest'ultime hanno la funzione di sostituire o predire una variabile, per ottenere maggiori informazioni riguardo le osservazioni mancanti, e sono presenti soprattutto

in ambito medico tramite indicatori di stato di salute, risultati di test diagnostici intermedi…

Nel prossimo capitolo si propongono alcuni approcci ed i rispettivi stimatori per la stima di questo modello.

METODI DI STIMA

In questi casi non è opportuno utilizzare approcci naïve, come l'analisi dei dati completi, che consiste nello escludere gli individui con dati mancanti o incompleti, oppure l'analisi dei dati senza considerare l'indice di censura. Questi metodi, con la tipologia di dataset che si sta considerando, portano con sé gravi problemi di stima. Infatti, come riportato in [3, p.102], rinunciano ad una parte dell'informazione disponibile, perché riducono il dataset di analisi. In compenso si ottengono procedure di stima più semplici, ma stimatori poco precisi, con distorsioni considerevoli e varianze distorte. Si ottengono stime dei coefficienti del modello distorte rispetto a quelli reali, in quanto non trattano la censura con la giusta attenzione.
Gli stimatori illustrati di seguito, che sono riportati e descritti in [1] da Dupuy e Bousselmi, si basano sull'idea di sostituzione dei dati mancanti o di ponderazione dei dati osservati.

Regression calibration

La *regression calibration* si basa sull'idea di sostituire gli indici di censura mancanti con le previsioni date da un modello di regressione, modellando la distribuzione condizionata degli indici di censura osservati. Questo metodo riscuote successo soprattutto nei modelli lineari generalizzati, ma con approssimazioni non soddisfacenti in modelli altamente non lineari.
Per ottenere questo risultato, si propone prima un modello sui dati completi per catturare la distribuzione dell'indice di censura date le osservazioni delle altre variabili e, in seguito, si stima il valore mancante per ogni soggetto di cui non si conosce la realizzazione di δ. Si definisce quindi la versione approssimata per gli indici di censura come:

$$\hat{\delta}_i = \xi_i \delta_i + (1 - \xi_i)\mathbb{E}[\delta_i | W_i] \quad con\ W_i = \left(Y_i^*, X_i^{\mathrm{T}}, V_i^{\mathrm{T}}\right)^{\mathrm{T}}.$$

Per conoscere il valore atteso, si assume che $\mathbb{E}[\delta_i|\mathrm{W}_i]$ sia specificato da un modello parametrico $m(W_i, \vartheta)$, dove il parametro ϑ q-dimensionale è ignoto di vero valore ϑ_0, con stessa dimensione di W_i. Il modello parametrico utilizzato è il modello di regressione logistico, ma anche il modello probit può essere considerato come alternativa.

$$\mathbb{E}[\delta_i | W_i] = m(W_i, \vartheta) = logit^{-1}(\vartheta^{\mathrm{T}} W_i) = \frac{e^{\vartheta^{\mathrm{T}} W_i}}{1 + e^{\vartheta^{\mathrm{T}} W_i}}.$$

Lo stimatore aggiornato per l'indice di censura risulta $\hat{\delta}_i = \xi_i \delta_i + (1 - \xi_i) m\left(W_i, \hat{\vartheta}_n\right)$, dove $\hat{\vartheta}_n$ è la stima del vero parametro ϑ_0, ottenuto tramite massimizzazione della funzione di verosimiglianza del modello $m\left(W_i, \hat{\vartheta}_n\right)$.

Il contributo di ogni osservazione $u \in \mathbb{N}$ alla funzione di distribuzione cumulata di $Y \sim \mathrm{Poisson}(\lambda)$ è espresso da:

$$P(Y \leq u) = \sum_{k=0}^{u} e^{-\lambda} \frac{\lambda^k}{k!} = \frac{\Gamma(u+1, \lambda)}{u!},$$

dove $\Gamma(u + 1, \lambda) = \int_{\lambda}^{\infty} t^u e^{-t}\, dt$ è la funzione Gamma incompleta superiore. Derivando tale funzione rispetto a λ, si ottiene: $\frac{\partial \Gamma(u,\lambda)}{\partial \lambda} = -e^{\lambda} * \lambda^{u-1}$.

Ora che sono noti, oppure stimati, tutti i termini per la funzione di log-verosimiglianza del modello di regressione di Poisson, si può definire lo stimatore del *regression calibration* $\tilde{\beta}_n$ come

$$\tilde{\beta}_n = \arg\max_{\beta} \tilde{l}_n\,(\beta, \hat{\vartheta}_n),$$

dove

$$\tilde{l}_n\left(\beta, \hat{\vartheta}_n\right) = \sum_{i=1}^{n} \left\{ \hat{\delta}_i\left(\hat{\vartheta}_n\right)\left(Y_i^* \beta^{\mathrm{T}} X_i - e^{\beta^{\mathrm{T}} X_i} - \log(Y_i^*!)\right) + \left(1 - \hat{\delta}_i\left(\hat{\vartheta}_n\right)\right) \log\left(1 - \frac{\Gamma\left(Y_i^*, e^{\beta^{\mathrm{T}} X_i}\right)}{(Y_i^* - 1)!}\right) \right\}$$

Si riportano di seguito le condizioni di regolarità necessarie per la consistenza e le proprietà asintotiche sulla distribuzione dello stimatore $\tilde{\beta}_n$:

- *C1*: I vettori X_i e V_i sono limitati per ogni $i = 1, 2, \ldots$
- *C2*: I veri parametri su cui si basa l'inferenza β_0 e ϑ_0 sono contenuti rispettivamente nei due insiemi limitati $B \subseteq \mathbb{R}^p$ e $\Theta \subseteq \mathbb{R}^q$.
- *C3*: $P(\delta = 1) > 0$ e $P(Y^* \geq 1 | \xi\delta = 0) = 1$.
- *C4*: La funzione $m(w, \vartheta)$ è differenziabile rispetto a ϑ per ogni w. Per ogni $\vartheta, \tilde{\vartheta} \in \Theta$, $\left|m(w, \vartheta) - m\left(w, \tilde{\vartheta}\right)\right| \leq h(w)\left\|\vartheta - \tilde{\vartheta}\right\|$ per una qualche funzione limitata h tale che $\mathbb{E}[h(W)] = v$.

Si nota che la condizione *C3* richiede che vi sia un'informazione minima sulla risposta quando $\xi * \delta = 0$ $(P(Y^* \geq 1 | \xi\delta = 0) = 1)$. Questa condizione garantisce che uno zero osservato non può essere ricondotto a un dato censurato, oppure a uno per il quale non si conosce se è censurato o meno. Al contrario, se si osserva $Y^* = 0$ e l'osservazione è censurata, o si considera censurata in assenza dell'indice, essa non fornisce informazione utile all'analisi, poiché potrebbe corrispondere a qualunque valore della distribuzione di Poisson.

Sotto queste condizioni, gli autori di [1] dimostrano la consistenza e la normalità asintotica dello stimatore $\tilde{\beta}_n$. Prima di poter enunciare il teorema, è necessario introdurre alcune notazioni. Si definisce h_β come:

$$h_\beta = \frac{e^{-e^{\beta^{\mathrm{T}}x} + \beta^{\mathrm{T}}xy}}{(y-1)! - \Gamma\left(y, e^{\beta^{\mathrm{T}}x}\right)}, \text{ per ogni } \beta \in \mathbb{R}^p, x \in \mathbb{R}^p \text{ e } y \in \mathbb{N}\backslash\{0\}.$$

Si introducono anche i termini $\dot{m}(W_i, \vartheta) = \frac{\partial m(W_i, \vartheta)}{\partial \vartheta}$, $\pi(W) = P(\xi = 1 | W)$ e le matrici:

$$\Sigma_1(\beta) = \mathbb{E}\left[XX^{\mathrm{T}}\left(\delta e^{\beta^{\mathrm{T}}X} + (\delta - 1)\{Y^* - e^{\beta^{\mathrm{T}}X} - h_\beta(Y^*, X)\}h_\beta(Y^*, X)\right)\right],$$

$$\Sigma_2(\beta, \vartheta) = \mathbb{E}\left[X\dot{m}^{\mathrm{T}}(W, \vartheta)\left(Y^* - e^{\beta^{\mathrm{T}}X} - h_\beta(Y^*, X)\right)\left(1 - \pi(W)\right)\right],$$

$$\Sigma_3(\beta, \vartheta) = \mathbb{E}\left[X\dot{m}^{\mathrm{T}}(W, \vartheta)\left(Y^* - e^{\beta^{\mathrm{T}}X} - h_\beta(Y^*, X)\right)\right].$$

Dopo aver introdotto questi elementi, possiamo enunciare il primo teorema. Per la dimostrazione si rimanda all'articolo [1, p.4367].

Teorema 3.1.
Sotto le assunzioni C1-C4, lo stimatore $\tilde{\beta}_n$ *è consistente* $\tilde{\beta}_n \xrightarrow{P} \beta_0$*, con* $n \to \infty$*, ed ha distribuzione asintoticamente normale* $\sqrt{n}(\tilde{\beta}_n - \beta_0) \xrightarrow{d} N(0, \Sigma)$*, dove*

$$\Sigma = \Sigma_1^{-1}(\beta_0)\{\Sigma_1(\beta_0) + (2\Sigma_3(\beta_0, \vartheta_0) - \Sigma_2(\beta_0, \vartheta_0))\Theta^{-1}(\theta_0)\Sigma_2^{\mathrm{T}}(\beta_0, \vartheta_0)\}\Sigma_1^{-1}(\beta_0).$$

Il lettore può ottenere ulteriori informazioni riguardo l'approccio del *regression calibration* consultando i testi [2, 3, 5].
Le condizioni *C1-C4* si estendono anche ai prossimi metodi.

Multiple imputation

L'idea alla base del *multiple imputation* è di creare *M* dataset completi, sostituendo i dati mancanti con dei valori simulati casualmente dalla distribuzione di δ_i dato W_i.
Può essere definito come un processo di stima inferenziale a tre passi. Prima si genera l'insieme dei valori mancanti che riflette la distribuzione condizionata dell'indice di censura. Poi si sostituiscono i dati mancanti con i valori presenti nell'insieme appena generato per ognuno degli *M* dataset. Nel secondo passo, ogni dataset si analizza con i metodi usati per i dati completi. Infine, si combinano le *M* statistiche ottenute in una singola statistica. Come per la regression calibration, si specifica un modello parametrico $m(W_i, \vartheta_0)$ per catturare la distribuzione degli indici di censura date le osservazioni dei soggetti $\mathbb{E}[\delta_i, W_i]$.
Facendo un ulteriore passaggio e definendo con $\hat{\vartheta}_n$ la stima di massima verosimiglianza del modello, la procedura prevede di sostituire ogni δ_i mancante con la realizzazione di una distribuzione di Bernoulli $B\left(m(W_i, \hat{\vartheta}_n)\right)$, che ha come parametro il modello parametrico. Ripetendo la procedura *M* volte si ottengono *M*

dataset completi. Dato ϑ, si definisce con $D_{i,j} \sim B\big(m(W_i, \vartheta)\big)$ la realizzazione di δ_i nel j-esimo dataset. Si ottiene in questo modo lo stimatore del *multiple imputation* per l'indice di censura:

$$\delta_{i,j}^* = \xi_i \delta_i + (1 - \xi_i) D_{i,j}(\vartheta).$$

Si ottengono M stime del parametro β, quindi per il dataset j-esimo si ha lo stimatore $\hat{\beta}_{n,j}^*$ massimizzando la rispettiva funzione di log-verosimiglianza:

$$l_n^*(\beta, \hat{\vartheta}_n) = \sum_{i=1}^{n} \Bigg\{ \delta_{i,j}^*(\hat{\vartheta}_n)\big(Y_i^* \beta^{\mathrm{T}} X_i - e^{\beta^{\mathrm{T}} X_i} - \log(Y_i^*!)\big) + \big(1 - \delta_{i,j}^*(\hat{\vartheta}_n)\big) \log\left(1 - \frac{\Gamma\big(Y_i^*, e^{\beta^{\mathrm{T}} X_i}\big)}{(Y_i^* - 1)!}\right) \Bigg\}$$

Successivamente dalle M stime di β se ne ottiene una complessiva tramite una semplice media aritmetica $\hat{\beta}_n^* = \frac{1}{M} \sum_{j=1}^{M} \hat{\beta}_{n,j}^*$.

Prima di enunciare il teorema per le proprietà asintotiche dello stimatore $\hat{\beta}_n$, si introduce il termine O_i, al fine di rappresentare il vettore contenente le informazioni del i-esimo soggetto osservato in funzione della presenza ($O_i = \{Y_i^*, X_i, \delta_i, \xi_i\}$) o meno ($O_i = \{Y_i^*, X_i, \xi_i\}$) dell'indice di censura.

Teorema 3.2.

Per ogni $j = 1, \dots, M$, *si definisce* $f_{\beta,\vartheta,j}(O_i) = X_i\big\{\delta_{i,j}^*(\vartheta)\big[Y_i^* - e^{\beta^{\mathrm{T}} X_i} - h_\beta(Y_i^*, X_i)\big] + h_\beta(Y_i^*, X_i)\big\}$. *Si definisce inoltre* $\Sigma_1^*(\beta, \vartheta) = var\left(\frac{1}{M} \sum_{j=1}^{M} f_{\beta,\vartheta,j}(O_i)\right)$. *Sotto le assunzioni C1-C4, lo stimatore* $\hat{\beta}_n$ *è consistente* $\hat{\beta}_n \xrightarrow{P} \beta_0$, *con* $n \to \infty$, *ed ha distribuzione asintoticamente normale* $\sqrt{n}\big(\hat{\beta}_n^* - \beta_0\big) \xrightarrow{d} N(0, \Sigma^*)$, *dove*

$$\Sigma^* = \Sigma_1^{-1}(\beta_0)\big\{\Sigma_1^*(\beta_0, \vartheta_0) + \big(2\Sigma_3(\beta_0, \vartheta_0) - \Sigma_2(\beta_0, \vartheta_0)\big)\Theta^{-1}(\vartheta_0)\Sigma_2^{\mathrm{T}}(\beta_0, \vartheta_0)\big\}\Sigma_1^{-1}(\beta_0).$$

Per la dimostrazione si rimanda all'articolo [1, p.4370].

Va notato che un vantaggio del *multiple imputation* come approccio analitico risiede nell'incorporare informazioni aggiuntive nel modello $m(W_i, \hat{\vartheta}_n)$. Queste informazioni ausiliarie potrebbero non essere di interesse nel modello di regressione, ma potrebbero rendere l'ipotesi della mancanza di dati di natura MAR sempre più plausibile, e queste informazioni sono semplici da aggiungere nel modello.

Confrontando i due metodi appena visti, essi si basano sulle stesse condizioni *C1-C4* e la funzione di verosimiglianza per il modello varia solo per la definizione dello stimatore dell'indice di censura. In particolare, anche se l'idea di base è la stessa, nel *regression calibration* lo stimatore $\hat{\delta}_i$ fornisce un valore compreso nell'intervallo $[0,1]$, a indicare la probabilità che l'i-esimo dato sia o meno censurato. Nel *multiple imputation*, dopo esser giunti allo stesso punto, si procede utilizzando la probabilità come parametro di una distribuzione di Bernoulli ottenendo lo stimatore $\delta^*_{i,j}$, che fornisce invece un valore nell'insieme $\{0,1\}$, a indicare la simulazione dell'i-esimo dato nel j-esimo dataset. L'imputazione per simulazione casuale può essere singola, con un solo dataset completo, o multipla, con M dataset completi. Si preferisce il *multiple imputation* per limitare le distorsioni alle stime causate dalla varianza delle variabili casuali, e si indicano come sufficienti $M = 50$ dataset.

Il lettore può ottenere ulteriori informazioni riguardo l'approccio del *multiple imputation* consultando i testi [3, 4].

Augmented inverse probability weighting (AIPW)

Il metodo *inverse probability weighting* ha un approccio diverso rispetto ai due metodi sopra visti. L'idea alla base è la specificazione di un modello di mancanza dei dati, cioè un modello per la probabilità che un soggetto sia un dato completo.

Si considera il dataset senza gli individui con dati mancanti e, sfruttando il peso campionario, si ottengono dei pesi a indicare la rappresentatività del rispettivo individuo all'interno del dataset completo. I pesi si calcolano tramite l'inverso della

probabilità di selezione (IPW per *inverse probability weighting*). La probabilità di selezione si definisce come

$$\pi(W_i) = P(\xi_i = 1|W_i).$$

Con questo metodo si nota una perdita di informazione perché le osservazioni delle variabili di soggetti con dati mancanti non sono pienamente utilizzate in fase di stima. Il metodo AIPW migliora l'IPW introducendo un termine aggiuntivo che coinvolge le osservazioni degli individui con alcuni dati mancanti.

Le quantità $\mathbb{E}[\delta_i, W_i]$ e $\pi(W_i)$ sono ignote; quindi, si stimano specificandole tramite due modelli parametrici, rispettivamente $m(W_i, \vartheta)$ e $\pi(W_i, \gamma)$. Dati $\vartheta, \gamma \in \mathbb{R}^q$, si definisce lo stimatore $\check{\delta}_i$ per gli indici di misura come:

$$\check{\delta}_i(\vartheta, \gamma) = \frac{\xi_i \delta_i}{\pi(W_i, \gamma)} + \left(1 - \frac{\xi_i}{\pi(W_i, \gamma)}\right) m(W_i, \vartheta).$$

Osservando lo stimatore si nota che, se non si conosce l'indice di censura, è uguale al modello parametrico $m(W_i, \vartheta)$ stimato. Le stime dei parametri ϑ e γ si ottengono tramite la massimizzazione della funzione di verosimiglianza dei rispettivi modelli, da cui i rispettivi stimatori $\hat{\vartheta}_n$ e $\hat{\gamma}_n$.

Si procede con la stima del modello di regressione di parametro β. In particolare, lo stimatore AIPW $\check{\beta}_n$ si ottiene risolvendo la funzione di log-verosimiglianza stimata $\check{l}_n(\beta, \hat{\vartheta}_n, \hat{\gamma}_n) = 0$, dove

$$\check{l}_n(\beta, \hat{\vartheta}_n, \hat{\gamma}_n) = \sum_{i=1}^{n} X_i \left[\check{\delta}_i(\hat{\vartheta}_n, \hat{\gamma}_n)\left(Y_i^* - e^{\beta^{\mathrm{T}} X_i} - h_\beta(Y_i^*, X_i)\right) + h_\beta(Y_i^*, X_i)\right],$$

con $h_\beta(Y_i^*, X_i) = \frac{\exp\left(-e^{\beta^{\mathrm{T}} X_i} + \beta^{\mathrm{T}} X_i Y_i^*\right)}{(Y_i^* - 1)! - \Gamma\left(Y_i^*, e^{\beta^{\mathrm{T}} X_i}\right)}$.

Si nota che l'utilizzo di un approccio diverso implica una funzione di verosimiglianza diversa rispetto a quelli precedenti. Questo metodo, infatti, utilizza un modello aggiuntivo, $\pi(W_i, \gamma)$, per il calcolo della probabilità di selezione. Per mantenere la consistenza dello stimatore $\check{\delta}_i(\vartheta, \gamma)$ e le sue proprietà asintotiche sulla distribuzione, oltre alle condizioni *C1-C4*, sono richieste altre ipotesi di regolarità:

- *C5*: lo spazio del parametro γ è contenuto nell'insieme limitato $G \subseteq \mathbb{R}^q$, dove si trova il vero parametro $\gamma_0 \in G$.
- *C6*: la funzione $\pi(\cdot,\cdot)$ soddisfa $\pi(w,\gamma) > 0$ per ogni $w \in W$ e $\gamma \in G$.
- *C7*: la funzione $\pi(w,\gamma)$ è differenziabile rispetto a γ per ogni w, e esiste una funzione limitata g tale che per ogni $\gamma, \tilde{\gamma} \in G, |\pi(w,\gamma) - \pi(w,\tilde{\gamma})| \leq g(w)\|\gamma - \tilde{\gamma}\|$ (si può indicare $\mathbb{E}[g(W) = u]$).

Le condizioni *C5* e *C7*, riferenti al modello per la probabilità di selezione, sono analoghe alle rispettive condizioni *C2* e *C4* per il modello sul valore atteso dell'indice di censura dato W_i. Dalla condizione *C6* si assume che la probabilità di selezione dei dati osservati non sia mai nulla.

Di seguito si enuncia il teorema per le proprietà asintotiche dello stimatore $\breve{\beta}_n$.

Teorema 3.3.

Sotto le assunzioni C1-C7, se almeno uno tra i modelli $m(W_i,\vartheta)$ *e* $\pi(W_i,\gamma)$ *è correttamente specificato, allora lo stimatore* $\breve{\beta}_n$ *è consistente* $\breve{\beta}_n \xrightarrow{P} \beta_0$, *con* $n \to \infty$. *Inoltre, sotto le assunzioni C1-C7, per* $n \to \infty$ *lo stimatore* $\breve{\beta}_n$ *ha la distribuzione asintotica di un vettore casuale Gaussiano* $\sqrt{n}(\breve{\beta}_n - \beta_0) \xrightarrow{d} N(0, \breve{\Sigma})$, *dove*

$$\breve{\Sigma} = \begin{cases} \Sigma_1^{-1}(\beta_0)\Sigma_7(\beta_0,\vartheta_0,\gamma^*)\Sigma_1^{-1}(\beta_0) \text{ se } m(W_i,\vartheta) \text{ è correttamente specificato,} \\ \Sigma_1^{-1}(\beta_0)\Sigma_8(\beta_0,\vartheta^*,\gamma_0)\Sigma_1^{-1}(\beta_0) \text{ se } \pi(W_i,\gamma) \text{ è correttamente specificato,} \\ \Sigma_1^{-1}(\beta_0) \text{ se entrambi } m(W_i,\vartheta) \text{ e } \pi(W_i,\gamma) \text{ sono correttamente specificati.} \end{cases}$$

Per la dimostrazione si rimanda all'articolo [1, p.4376-4379].

Il lettore può ottenere ulteriori informazioni riguardo l'approccio AIPW consultando i testi [3, 6].

CONFRONTO DEGLI STIMATORI PER SIMULAZIONE

In questo capitolo si analizzano i comportamenti degli stimatori, ottenuti da *regression calibration* (RC), *multiple imputation* (MI) e AIPW, applicati a campioni simulati finiti e in diverse condizioni.

La differenza sostanziale tra i metodi descritti risiede nell'approccio AIPW, che ha bisogno di un altro modello per la probabilità di conoscere o meno l'indice di censura, mentre MI e RC hanno bisogno di un solo modello, per la distribuzione degli indici di censura δ_i mancanti data la matrice dei dati osservati.

Costruzione dei dataset simulati

Il campione è composto da n individui indipendenti, per ognuno si osserva la variabile risposta Y generata da un modello di regressione di Poisson con parametro λ definito come:

$$\lambda = \exp(\beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5),$$

dove $\beta = (0.2, -0.1, 0.4, 0.3, 0.5)$, $X_2 \sim N(0,1)$, $X_3 \sim \text{Bernoulli}(0.3)$, $X_4 \sim N(0,1.5)$, $X_5 \sim \text{Unif}[2,5]$.

Sono state realizzate varie simulazioni per osservare il comportamento degli stimatori in funzione della numerosità del dataset (n), del tasso di censura (TC) e del tasso di indici di censura mancanti (TM).

- Esperimento 1: si considera $n = 250, \text{TC} = 20\%, \text{TM} = 20\%$,
- Esperimento 2: si considera $n = 500, \text{TC} = 20\%, \text{TM} = 20\%$,
- Esperimento 3: si considera $n = 500, \text{TC} = 20\%, \text{TM} = 40\%$,
- Esperimento 4: si considera $n = 500, \text{TC} = 40\%, \text{TM} = 20\%$.

I meccanismi di censura e di assenza degli indici sono rispettivamente generati dai due modelli $\text{logit}\big(m(W, \vartheta)\big) = \vartheta_1 + \vartheta_2 X_2 + \vartheta_3 X_3 + \vartheta_4 X_4 + \vartheta_5 X_5 + \vartheta_6 Y$ e

$\text{logit}\big(\pi(W,\gamma)\big) = \gamma_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 X_4 + \gamma_5 X_5 + \gamma_6 Y^*$. I valori dei vettori ϑ e λ sono scelti in modo tale da ottenere le quattro situazioni presentate negli esperimenti.
Al fine di analizzare il comportamento degli stimatori, l'esperimento 2 viene utilizzato come riferimento. Il confronto con l'esperimento 1 consente di valutare l'impatto della numerosità del dataset, quello con l'esperimento 3 permette di isolare l'effetto di TM e quello con l'esperimento 4 di esaminare l'effetto di TC.
Si confrontano inoltre gli stimatori sotto tre scenari differenti: i) solo $m(W,\vartheta)$ è correttamente specificato, ii) solo $\pi(W,\gamma)$ è correttamente specificato e iii) sono entrambi correttamente specificati. Quando i modelli non sono correttamente specificati si prendono in esame $\text{logit}\big(\pi(W,\gamma)\big) = \gamma_1 + \gamma_2 X_2 + \gamma_3 X_3 + \gamma_4 Y^*$ per $\pi(W,\gamma)$ e $\text{logit}\big(m(W,\vartheta)\big) = \vartheta_1 + \vartheta_2 X_2 + \vartheta_3 X_3 + \vartheta_4 Y^*$ per $m(W,\vartheta)$. Infine, per quantificare la distorsione degli stimatori, si stabiliscono due stimatori di riferimento calcolati su un dataset di 1000 osservazioni. Il primo è lo stimatore FD (*full data*), che si basa sul dataset senza indici di censura mancanti, e il secondo è lo stimatore CC (*complete-case*), ottenuto tramite massimizzazione della log-verosimiglianza. Per l'approccio MI si ritengono sufficienti $M = 50$ dataset.
Tutte le stime sono state determinate utilizzando l'algoritmo Newton-Raphson, implementato in R.

Confronto dei risultati

Ci si sofferma ad analizzare lo stimatore AIPW $\check{\delta}_i(\vartheta,\gamma)$, in particolare si nota che: per $\xi_i = 0$ è uguale a $m(W_i,\vartheta)$ e per $\xi_i = 1$ è uguale a $m(W_i,\vartheta) + \frac{\delta_i - m(W_i,\vartheta)}{\pi(W_i,\gamma)}$. La forza del metodo AIPW risiede nella doppia robustezza dello stimatore perché utilizza sia $m(W,\vartheta)$ sia $\pi(W,\gamma)$. Si osserva, nel caso del secondo scenario, che ogni individuo contribuisce in modo errato alla funzione di log-verosimiglianza; quindi, lo stimatore $\check{\delta}_i(\vartheta,\gamma)$ rimane maggiormente sensibile all'errata specificazione di $m(W,\vartheta)$ rispetto

a $\pi(W,\gamma)$. Tuttavia, il termine di ponderazione $\frac{1}{\pi(W_i,\gamma)}$ permette di bilanciare la stima in modo sistematico, mantenendo lo stimatore consistente.

Si illustrano di seguito i risultati ottenuti dallo studio tramite simulazione. A fine capitolo si riportano le tabelle 1, 2, 3, 4 dell'articolo di riferimento [1, p.4361-4363]. I numeri delle tabelle corrispondono agli esperimenti, infatti la Tabella 2 è la tabella di riferimento ed i confronti sono effettuati come spiegato analogamente per gli esperimenti precedentemente.

Come si osserva nel confronto tra Tabella 1 e Tabella 2, in merito alla dimensione campionaria, gli stimatori offrono prestazioni migliori quando la numerosità del dataset è maggiore, come ci si poteva aspettare.

Riguardo il secondo scenario (seconda sezione di colonne nelle tabelle), ovvero quando solo il modello $\pi(W,\gamma)$ è correttamente specificato, lo stimatore AIPW $\check{\delta}_i(\vartheta,\gamma)$ è preferibile perché fornisce minori SE e RMSE rispetto agli stimatori RC e MI. Questa situazione si nota soprattutto quando si alza il tasso di censura, come riportano i risultati della Tabella 4. D'altro canto, quando la censura è moderata, (Tabella 1-Tabella 3) la distorsione della stima AIPW rimane moderata e dello stesso ordine di grandezza rispetto al primo scenario ($TC = 0{,}2$), mentre diviene considerevole quando si alza ($TC = 0{,}4$). Si può notare come il metodo AIPW, in generale, riporti prestazioni più simili alle stime di riferimento FD rispetto agli altri approcci. Gli stimatori $\hat{\delta}_i$ e $\delta^*_{i,j}$ non sono robusti in quanto si basano solo su $m(W,\vartheta)$, infatti, se questo modello non è correttamente specificato, non riescono a contenere la distorsione delle stime.

Nel primo scenario, come mostra la prima sezione di colonne nelle tabelle, tutti gli stimatori si comportano in modo simile. Nonostante la proprietà della doppia robustezza, lo stimatore AIPW non riporta prestazioni migliori. Dato che l'errore è asimmetrico, non riguardando tutti i soggetti, è più difficile da correggere. In modo analogo, nel terzo scenario, dove tutti i modelli sono correttamente specificati, tutti i metodi presentano prestazioni simili.

Riguardo la varianza degli stimatori, le tabelle non sono riportate ma confrontabili nell'articolo [1, p.4364]. Le loro stime tramite i metodi RC, MI e AIPW forniscono prestazioni migliori all'aumentare della numerosità campionaria e al diminuire di TC e TM. Nel primo scenario la varianza stimata di RC presenta prestazioni migliori in termini di errore relativo, leggermente inferiore quella di MI e ancora inferiori quella di AIPW. La varianza stimata di AIPW si rivela, in ogni scenario, con risultati migliori in termini di RMSE e, solo nel terzo scenario, si dimostra con prestazioni superiori rispetto RC e MI.

In conclusione, i metodi *regression calibration*, *multiple imputation* e *augmented inverse probability weighting* portano a risultati analoghi quando sono correttamente specificati entrambi i modelli $m(W,\vartheta)$ e $\pi(W,\gamma)$ o solo $m(W,\vartheta)$. Se invece il modello $m(W,\vartheta)$ non è correttamente specificato, allora l'approccio AIPW riporta prestazioni migliori rispetto RC e MI, in particolare in termini di stima puntuale. Lo stimatore CC ha ottenuto prestazioni migliori in tutti gli scenari.

Tabella 1. *Risultati della simulazione per* $n = 250$, $\mathrm{TC} = 20\%$, $\mathrm{TM} = 20\%$.

Bias: distorsione. SE: errore standard medio. RMSE: errore quadratico medio.

CP: probabilità di copertura empirica dell'intervallo di confidenza a un livello di significatività del 95%.

		correct $m(\mathbf{W},\theta)$ / incorrect $\pi(\mathbf{W},\gamma)$					incorrect $m(\mathbf{W},\theta)$ / correct $\pi(\mathbf{W},\gamma)$					both models correct				
estimator		β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
full data	bias	-0.0097	-0.0001	0.0005	0.0009	0.0021	-0.0097	-0.0001	0.0005	0.0009	0.0021	-0.0097	-0.0001	0.0005	0.0009	0.0021
	SE	0.1092	0.0213	0.0459	0.0164	0.0267	0.1092	0.0213	0.0459	0.0164	0.0267	0.1092	0.0213	0.0459	0.0164	0.0267
	RMSE	0.1549	0.0307	0.0643	0.0232	0.0380	0.1549	0.0307	0.0643	0.0232	0.0380	0.1549	0.0307	0.0643	0.0232	0.0380
	CP	0.9571	0.9397	0.9510	0.9540	0.9581	0.9571	0.9397	0.9510	0.9540	0.9581	0.9571	0.9397	0.9510	0.9540	0.9581
CC	bias	0.0624	-0.0033	-0.0062	-0.0096	-0.0082	0.0624	-0.0033	-0.0062	-0.0096	-0.0082	0.0624	-0.0033	-0.0062	-0.0096	-0.0082
	SE	0.1235	0.0233	0.0500	0.0187	0.0295	0.1235	0.0233	0.0500	0.0187	0.0295	0.1235	0.0233	0.0500	0.0187	0.0295
	RMSE	0.1842	0.0334	0.0698	0.0281	0.0423	0.1842	0.0334	0.0698	0.0281	0.0423	0.1842	0.0334	0.0698	0.0281	0.0423
	CP	0.9326	0.9418	0.9571	0.9142	0.9540	0.9326	0.9418	0.9571	0.9142	0.9540	0.9326	0.9418	0.9571	0.9142	0.9540
RC	bias	-0.0062	0.0001	-0.0003	0.0004	0.0014	0.0256	0.0026	-0.0069	-0.0042	-0.0061	-0.0062	0.0001	-0.0003	0.0004	0.0014
	SE	0.1111	0.0217	0.0469	0.0167	0.0273	0.1095	0.0217	0.0470	0.0164	0.0268	0.1111	0.0217	0.0469	0.0167	0.0273
	RMSE	0.1568	0.0311	0.0653	0.0235	0.0386	0.1615	0.0316	0.0666	0.0242	0.0396	0.1568	0.0311	0.0653	0.0235	0.0386
	CP	0.9540	0.9438	0.9510	0.9540	0.9540	0.9387	0.9336	0.9428	0.9234	0.9305	0.9540	0.9438	0.9510	0.9540	0.9540
AIPW	bias	-0.0119	-0.0002	0.0008	0.0012	0.0026	-0.0133	-0.0002	0.0014	0.0016	0.0029	-0.0100	-0.0001	0.0005	0.0010	0.0021
	SE	0.1089	0.0213	0.0458	0.0162	0.0267	0.1058	0.0211	0.0453	0.0159	0.0260	0.1092	0.0213	0.0460	0.0164	0.0268
	RMSE	0.1557	0.0308	0.0646	0.0232	0.0383	0.1609	0.0316	0.0660	0.0243	0.0395	0.1557	0.0308	0.0648	0.0233	0.0383
	CP	0.9510	0.9397	0.9428	0.9499	0.9459	0.9152	0.9183	0.9275	0.9122	0.9142	0.9520	0.9397	0.9459	0.9520	0.9489
MI	bias	-0.0069	0.0000	-0.0001	0.0005	0.0015	0.0241	0.0024	-0.0065	-0.0040	-0.0058	-0.0069	0.0000	-0.0001	0.0005	0.0015
	SE	0.1091	0.0212	0.0457	0.0163	0.0267	0.1124	0.0219	0.0476	0.0168	0.0274	0.1091	0.0212	0.0457	0.0163	0.0267
	RMSE	0.1556	0.0308	0.0646	0.0233	0.0383	0.1633	0.0317	0.0671	0.0245	0.0399	0.1556	0.0308	0.0646	0.0233	0.0383
	CP	0.9520	0.9418	0.9479	0.9489	0.9489	0.9459	0.9397	0.9510	0.9356	0.9408	0.9520	0.9418	0.9479	0.9489	0.9489

Tabella 2. *Risultati della simulazione per* $n = 500$, TC $= 20\%$, TM $= 20\%$.

		correct $m(\mathbf{W},\theta)$ / incorrect $\pi(\mathbf{W},\gamma)$					incorrect $m(\mathbf{W},\theta)$ / correct $\pi(\mathbf{W},\gamma)$					both models correct				
estimator		β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
full data	bias	-0.0060	-0.0008	0.0014	0.0005	0.0013	-0.0060	-0.0008	0.0014	0.0005	0.0013	-0.0060	-0.0008	0.0014	0.0005	0.0013
	SE	0.0766	0.0150	0.0323	0.0115	0.0188	0.0766	0.0150	0.0323	0.0115	0.0188	0.0766	0.0150	0.0323	0.0115	0.0188
	RMSE	0.1099	0.0215	0.0453	0.0162	0.0269	0.1099	0.0215	0.0453	0.0162	0.0269	0.1099	0.0215	0.0453	0.0162	0.0269
	CP	0.9460	0.9490	0.9500	0.9560	0.9420	0.9460	0.9490	0.9500	0.9560	0.9420	0.9460	0.9490	0.9500	0.9560	0.9420
CC	bias	0.0688	-0.0038	-0.0051	-0.0109	-0.0092	0.0688	-0.0038	-0.0051	-0.0109	-0.0092	0.0688	-0.0038	-0.0051	-0.0109	-0.0092
	SE	0.0866	0.0163	0.0350	0.0131	0.0207	0.0866	0.0163	0.0350	0.0131	0.0207	0.0866	0.0163	0.0350	0.0131	0.0207
	RMSE	0.1421	0.0238	0.0493	0.0213	0.0310	0.1421	0.0238	0.0493	0.0213	0.0310	0.1421	0.0238	0.0493	0.0213	0.0310
	CP	0.8676	0.9388	0.9519	0.8656	0.9188	0.8676	0.9388	0.9519	0.8656	0.9188	0.8676	0.9388	0.9519	0.8656	0.9188
RC	bias	-0.0022	-0.0006	0.0008	0.0000	0.0005	0.0311	0.0021	-0.0065	-0.0048	-0.0073	-0.0022	-0.0006	0.0008	0.0000	0.0005
	SE	0.0780	0.0153	0.0329	0.0117	0.0192	0.0769	0.0153	0.0330	0.0115	0.0188	0.0780	0.0153	0.0329	0.0117	0.0192
	RMSE	0.1114	0.0218	0.0463	0.0163	0.0273	0.1177	0.0220	0.0476	0.0174	0.0286	0.1114	0.0218	0.0463	0.0163	0.0273
	CP	0.9490	0.9490	0.9520	0.9570	0.9430	0.9100	0.9450	0.9330	0.9160	0.9120	0.9490	0.9490	0.9520	0.9570	0.9430
AIPW	bias	-0.0078	-0.0009	0.0020	0.0007	0.0017	-0.0069	-0.0008	0.0018	0.0009	0.0015	-0.0060	-0.0008	0.0017	0.0006	0.0013
	SE	0.0765	0.0150	0.0322	0.0114	0.0187	0.0747	0.0149	0.0319	0.0112	0.0183	0.0766	0.0150	0.0323	0.0115	0.0188
	RMSE	0.1106	0.0217	0.0459	0.0161	0.0271	0.1151	0.0221	0.0474	0.0170	0.0280	0.1106	0.0217	0.0459	0.0162	0.0271
	CP	0.9410	0.9450	0.9430	0.9500	0.9390	0.8990	0.9280	0.9290	0.9150	0.9070	0.9420	0.9420	0.9450	0.9520	0.9420
MI	bias	-0.0026	-0.0006	0.0009	0.0000	0.0006	0.0301	0.0019	-0.0062	-0.0047	-0.0071	-0.0026	-0.0006	0.0009	0.0000	0.0006
	SE	0.0772	0.0150	0.0324	0.0115	0.0189	0.0805	0.0155	0.0340	0.0120	0.0196	0.0772	0.0150	0.0324	0.0115	0.0189
	RMSE	0.1109	0.0216	0.0460	0.0162	0.0272	0.1199	0.0222	0.0482	0.0177	0.0290	0.1109	0.0216	0.0460	0.0162	0.0272
	CP	0.9410	0.9450	0.9430	0.9460	0.9380	0.9340	0.9500	0.9470	0.9330	0.9300	0.9410	0.9450	0.9430	0.9460	0.9380

Tabella 3. *Risultati della simulazione per* $n = 500$, TC $= 20\%$, TM $= 40\%$.

		correct $m(\mathbf{W},\theta)$ / incorrect $\pi(\mathbf{W},\gamma)$					incorrect $m(\mathbf{W},\theta)$ / correct $\pi(\mathbf{W},\gamma)$					both models correct				
estimator		β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
full data	bias	-0.0015	-0.0001	-0.0001	0.0006	0.0003	-0.0015	-0.0001	-0.0001	0.0006	0.0003	-0.0015	-0.0001	-0.0001	0.0006	0.0003
	SE	0.0765	0.0150	0.0323	0.0115	0.0188	0.0765	0.0150	0.0323	0.0115	0.0188	0.0765	0.0150	0.0323	0.0115	0.0188
	RMSE	0.1100	0.0213	0.0447	0.0162	0.0270	0.1100	0.0213	0.0447	0.0162	0.0270	0.1100	0.0213	0.0447	0.0162	0.0270
	CP	0.9370	0.9450	0.9520	0.9540	0.9440	0.9370	0.9450	0.9520	0.9540	0.9440	0.9370	0.9450	0.9520	0.9540	0.9440
CC	bias	0.1224	0.0121	-0.0128	-0.0177	-0.0152	0.1224	0.0121	-0.0128	-0.0177	-0.0152	0.1224	0.0121	-0.0128	-0.0177	-0.0152
	SE	0.0993	0.0186	0.0391	0.0151	0.0233	0.0993	0.0186	0.0391	0.0151	0.0233	0.0993	0.0186	0.0391	0.0151	0.0233
	RMSE	0.1864	0.0289	0.0562	0.0275	0.0363	0.1864	0.0289	0.0562	0.0275	0.0363	0.1864	0.0289	0.0562	0.0275	0.0363
	CP	0.7500	0.8940	0.9400	0.7800	0.8920	0.7500	0.8940	0.9400	0.7800	0.8920	0.7500	0.8940	0.9400	0.7800	0.8920
RC	bias	0.0065	0.0004	-0.0021	-0.0006	-0.0013	0.0779	0.0010	-0.0173	-0.0113	-0.0181	0.0065	0.0004	-0.0021	-0.0006	-0.0013
	SE	0.0793	0.0154	0.0336	0.0119	0.0196	0.0764	0.0153	0.0335	0.0115	0.0187	0.0793	0.0154	0.0336	0.0119	0.0196
	RMSE	0.1134	0.0217	0.0465	0.0167	0.0281	0.1387	0.0220	0.0504	0.0204	0.0335	0.1134	0.0217	0.0465	0.0167	0.0281
	CP	0.9430	0.9480	0.9580	0.9490	0.9460	0.7750	0.9420	0.9200	0.8080	0.7920	0.9430	0.9480	0.9580	0.9490	0.9460
AIPW	bias	-0.0052	0.0000	0.0004	0.0010	0.0012	-0.0061	-0.0005	0.0008	0.0014	0.0013	-0.0017	0.0000	-0.0002	0.0007	0.0003
	SE	0.0765	0.0150	0.0322	0.0112	0.0188	0.0698	0.0149	0.0310	0.0105	0.0173	0.0765	0.0150	0.0323	0.0115	0.0188
	RMSE	0.1115	0.0215	0.0455	0.0163	0.0276	0.1185	0.0225	0.0480	0.0177	0.0291	0.1113	0.0216	0.0455	0.0164	0.0275
	CP	0.9390	0.9400	0.9410	0.9470	0.9370	0.8501	0.9080	0.8925	0.8273	0.8635	0.9410	0.9410	0.9430	0.9550	0.9420
MI	bias	0.0056	0.0004	-0.0018	-0.0005	-0.0011	0.0763	0.0010	-0.0168	-0.0110	-0.0177	0.0056	0.0004	-0.0018	-0.0005	-0.0011
	SE	0.0777	0.0150	0.0326	0.0116	0.0191	0.0822	0.0156	0.0352	0.0125	0.0200	0.0777	0.0150	0.0326	0.0116	0.0191
	RMSE	0.1124	0.0215	0.0458	0.0165	0.0278	0.1412	0.0222	0.0514	0.0209	0.0340	0.1124	0.0215	0.0458	0.0165	0.0278
	CP	0.9440	0.9420	0.9440	0.9470	0.9420	0.8330	0.9450	0.9390	0.8690	0.8440	0.9440	0.9420	0.9440	0.9470	0.9420

Tabella 4. *Risultati della simulazione per* $n = 500$, $\mathrm{TC} = 40\%$, $\mathrm{TM} = 20\%$.

		correct $m(\mathbf{W},\theta)$ / incorrect $\pi(\mathbf{W},\gamma)$					incorrect $m(\mathbf{W},\theta)$ / correct $\pi(\mathbf{W},\gamma)$					both models correct				
estimator		β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5	β_1	β_2	β_3	β_4	β_5
full data	bias	0.0039	0.0002	-0.0005	-0.0002	-0.0007	0.0039	0.0002	-0.0005	-0.0002	-0.0007	0.0039	0.0002	-0.0005	-0.0002	-0.0007
	SE	0.0907	0.0180	0.0402	0.0144	0.0233	0.0907	0.0180	0.0402	0.0144	0.0233	0.0907	0.0180	0.0402	0.0144	0.0233
	RMSE	0.1285	0.0256	0.0573	0.0199	0.0327	0.1285	0.0256	0.0573	0.0199	0.0327	0.1285	0.0256	0.0573	0.0199	0.0327
	CP	0.9587	0.9518	0.9420	0.9676	0.9538	0.9587	0.9518	0.9420	0.9676	0.9538	0.9587	0.9518	0.9420	0.9676	0.9538
CC	bias	0.0764	-0.0033	-0.0062	-0.0116	-0.0115	0.0764	-0.0033	-0.0062	-0.0116	-0.0115	0.0764	-0.0033	-0.0062	-0.0116	-0.0115
	SE	0.1049	0.0202	0.0451	0.0168	0.0263	0.1049	0.0202	0.0451	0.0168	0.0263	0.1049	0.0202	0.0451	0.0168	0.0263
	RMSE	0.1668	0.0289	0.0645	0.0261	0.0387	0.1668	0.0289	0.0645	0.0261	0.0387	0.1668	0.0289	0.0645	0.0261	0.0387
	CP	0.8702	0.9440	0.9479	0.8899	0.9272	0.8702	0.9440	0.9479	0.8899	0.9272	0.8702	0.9440	0.9479	0.8899	0.9272
RC	bias	0.0221	0.0012	-0.0037	-0.0027	-0.0051	0.1838	0.0123	-0.0361	-0.0271	-0.0464	0.0221	0.0012	-0.0037	-0.0027	-0.0051
	SE	0.0960	0.0190	0.0429	0.0152	0.0250	0.0903	0.0192	0.0441	0.0142	0.0231	0.0960	0.0190	0.0429	0.0152	0.0250
	RMSE	0.1356	0.0269	0.0605	0.0208	0.0349	0.2279	0.0299	0.0720	0.0346	0.0577	0.1356	0.0269	0.0605	0.0208	0.0349
	CP	0.9548	0.9548	0.9469	0.9676	0.9587	0.4808	0.9036	0.8673	0.5152	0.5034	0.9548	0.9548	0.9469	0.9676	0.9587
AIPW	bias	-0.0036	-0.0004	0.0022	0.0010	0.0011	0.0199	0.0017	-0.0038	-0.0028	-0.0047	0.0046	0.0002	0.0003	-0.0002	-0.0010
	SE	0.0899	0.0178	0.0398	0.0138	0.0231	0.0681	0.0169	0.0366	0.0110	0.0173	0.0907	0.0180	0.0402	0.0144	0.0233
	RMSE	0.1296	0.0263	0.0589	0.0198	0.0332	0.1332	0.0273	0.0617	0.0216	0.0340	0.1302	0.0264	0.0589	0.0201	0.0333
	CP	0.9508	0.9292	0.9272	0.9489	0.9489	0.7443	0.8741	0.8348	0.7443	0.7345	0.9479	0.9292	0.9361	0.9626	0.9459
MI	bias	0.0203	0.0010	-0.0031	-0.0025	-0.0046	0.1781	0.0117	-0.0344	-0.0262	-0.0449	0.0203	0.0010	-0.0031	-0.0025	-0.0046
	SE	0.0933	0.0183	0.0412	0.0147	0.0241	0.1011	0.0199	0.0467	0.0162	0.0258	0.0933	0.0183	0.0412	0.0147	0.0241
	RMSE	0.1337	0.0264	0.0594	0.0205	0.0343	0.2281	0.0301	0.0728	0.0348	0.0577	0.1337	0.0264	0.0594	0.0205	0.0343
	CP	0.9430	0.9390	0.9381	0.9587	0.9508	0.5821	0.9145	0.9046	0.6332	0.5929	0.9430	0.9390	0.9381	0.9587	0.9508

APPLICAZIONE AD UN DATASET REALE

In questo capitolo si applicano i metodi di stima proposti a un caso di studio reale, analizzandone l'evoluzione dei risultati senza conoscere il processo generatore del fenomeno. Il dataset si riferisce ad un sondaggio riguardo il consumo giornaliero di frutta e verdura, condotto nel Regno Unito, presentato anche in [1].

Il numero totale di soggetti rispondenti è $n = 928$, dei quali 228 (il 24,6% del campione) con indice di censura mancante. Il 29,6% dei rispondenti di cui si conosce l'indice di censura presentano il valore del consumo giornaliero censurato a destra.

Le variabili esplicative raccolte sono il genere, l'età, lo stato civile (celibe/divorziato/separato o sposato), il titolo di studio (con tre livelli: "*General Certificate of Secondary Education (GCSE) or no qualification*", "*A-level or equivalent*", "*higher education")* e un valore per indicare l'apprezzamento del rispondente nel consumo giornaliero di frutta e di verdura (a tre livelli: "*enough*", "*not enough*", "*more than enough*"). Si utilizza il modello di regressione logistico per stimare il valore atteso condizionato dell'indice di censura $m(W, \vartheta)$ e la probabilità di selezione $\pi(W, \gamma)$.

I risultati sono riportati nella Tabella 5, a fine capitolo. In questa tabella le variabili assumono i seguenti significati: "*constant*" per l'intercetta; "*gender*" $= 1$ per i maschi e 0 per le femmine; "*age*" per l'età dei rispondenti; "*single*" $= 1$ riferito a rispondenti celibi/divorziati/separati e 0 per gli sposati; "*GCSE/no qualif.*" $= 1$ se il rispondente non possiede qualificazioni o ha conseguito il diploma di istruzione secondaria (obbligatorio nel Regno Unito, conseguito a 16 anni) e 0 altrimenti, "*A-level or equiv.*" $= 1$ se il rispondente ha ottenuto un diploma "*A-level*" o equivalente; "*more than enough*" $= 1$ se il rispondente considera abbondante il proprio consumo giornaliero di frutta e verdura e 0 altrimenti, "*not enough*" $= 1$ se il rispondente considera non sufficiente il proprio consumo giornaliero di frutta e verdura e 0 altrimenti.

I metodi di stima impiegati, RC, MI, AIPW e CC, hanno fornito un quadro coerente dei fattori che influenzano il consumo di frutta e verdura. Tutti gli approcci concordano

sul fatto che l'età è un fattore significativo, in particolare il consumo di frutta e verdura aumenta all'aumentare dell'età. Per quanto riguarda il genere, si osserva una discrepanza: la variabile risulta statisticamente significativa (al livello del 5%) esclusivamente nel modello basato sul metodo MI, mentre gli altri metodi non rilevano tale associazione. Nelle stime dei restanti fattori tutti i metodi riportano la presenza di significatività e la medesima tendenza. In particolare, lo stato civile "spostato" è associato a un aumento del consumo di frutta e verdura, un basso titolo di studio ne diminuisce il consumo e, come immaginabile, l'apprezzamento nell'assumere frutta e verdura esercita un'influenza positiva sul consumo.

Come ci si aspetterebbe, i metodi naïve, come l'implementazione di un modello di regressione di Poisson senza considerare la censura, risultano più semplici da applicare ma sottostimano molto il livello base del consumo di frutta e verdura.

Gli stimatori RC, MI e AIPW mantengono un comportamento analogo allo studio di simulazione. Le stime CC presentano solitamente errori standard maggiori, il che riflette la perdita di efficienza del metodo. Le tre stime della varianza AIPW ottenute sono uguali alle stime teoriche fino a tre cifre. Per questo motivo, alla luce dello studio di simulazione, con questo dataset, si consiglia l'utilizzo dell'approccio AIPW per la stima.

Tabella 5. *Risultati dello studio sul consumo giornaliero di frutta e verdura.*

	CC			RC			MI			AIPW		
	est	se	p-value	est	se	p-value	est	se	p-value	est	se	p-value
constant	1.6372	0.0897	0.0000	1.6042	0.0785	0.0000	1.5980	0.0632	0.0000	1.6187	0.0766	0.0000
gender	-0.0721	0.0452	0.1107	-0.0699	0.0402	0.0818	-0.0701	0.0353	0.0470	-0.0715	0.0385	0.0636
age	0.0027	0.0013	0.0343	0.0027	0.0011	0.0163	0.0028	0.0009	0.0033	0.0026	0.0011	0.0239
single	-0.1175	0.0444	0.0082	-0.0979	0.0382	0.0104	-0.0944	0.0329	0.0041	-0.1039	0.0382	0.0065
GCSE/no qualif.	-0.2600	0.0535	0.0000	-0.2392	0.0477	0.0000	-0.2389	0.0405	0.0000	-0.2389	0.0455	0.0000
A-level or equiv.	-0.1196	0.0797	0.1335	-0.1182	0.0710	0.0961	-0.1195	0.0630	0.0577	-0.1236	0.0675	0.0669
more than enough	0.3749	0.0679	0.0000	0.3746	0.0610	0.0000	0.3753	0.0582	0.0000	0.3712	0.0574	0.0000
not enough	-0.5611	0.0546	0.0000	-0.5045	0.0478	0.0000	-0.5029	0.0424	0.0000	-0.5072	0.0459	0.0000

CONCLUSIONI

Nel presente lavoro si presentano e si confrontano alcuni stimatori per la regressione parametrica di Poisson per i dati di conteggio, con indici di censura parzialmente mancanti. I metodi *regression calibration* e *multiple imputation* si basano sull'idea di stimare o sostituire gli indici di censura mancanti, tramite un modello che ne studia il valore atteso condizionato date le variabili osservate. Gli stimatori portano a inferenze affidabili quando questo modello è correttamente specificato. Il metodo *augmented inverse probability weighting* (AIPW), nonostante presenta distorsioni maggiori all'aumentare del tasso di censura, è preferibile in termini di errore standard e di errore quadratico medio. Inoltre, è robusto rispetto a scorrette specificazioni del modello in quanto gode della proprietà della doppia robustezza. Infatti, questo approccio utilizza un ulteriore modello, che studia la probabilità condizionata di conoscere o meno l'indice di censura date le variabili osservate.

In conclusione, se il modello che studia la distribuzione condizionata degli indici di censura è correttamente specificato, allora si ottengono prestazioni affidabili e simili da parte di tutti gli stimatori. Solo se quest'ultimo modello non è correttamente specificato il metodo AIPW risulta preferibile.

BIBLIOGRAFIA

[1] Bousselmi, B., Dupuy, J.-F., 2021. Censored count data regression with missing censoring information. Electronic Journal of Statistics, 15(2), 4343–4383.

[2] Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M., 2006. Measurement Error in Nonlinear Models: A Modern Perspective (2nd ed.). Chapman, Hall/CRC.

[3] Dupuy, J.-F, 2025. Generalized Linear Models: Problems with Censored, Missing, and Zero-inflated Data. ISTE, John Wiley & Sons.

[4] Horton, N.J., Lipsitz, S.R., 2001. Multiple imputation in practice: comparison of software packages for regression models with missing variables. The American Statistician 55(3), 244-254.

[5] Huang, S.Y.H., 2005. Regression calibration using response variables in linear models. Statistica Sinica 15, 685-696.

[6] Seaman, S.R., White, I.R., 2013. Review of inverse probability weighting for dealing with missing data. Statistical methods in medical research 22(3), 278-295.