Tạp chí Khoa học Đại học Cần Thơ
Phần A: Khoa học tự nhiên, Công nghệ và Môi trường
ISSN 1859-2333 | e-ISSN 2815-5599

DOI:10.22144/ctujos.2024.389

ƯỚC LƯỢNG THAM SỐ MÔ HÌNH HỒI QUY LOGISTIC VỚI HIỆP BIẾN THIẾU DỮ LIỆU NGẪU NHIÊN VÀ ỨNG DỤNG

Trần Phước Lộc*, Tạ Thị Thanh Thúy, Dương Thị Tuyền, Dương Thị Bé Ba, Lê Hoài Nhân và Lâm Hoàng Chương
Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ
**Tác giả liên hệ (Corresponding author): tploc@ctu.edu.vn*

Thông tin chung (Article Information)

Nhận bài (Received): 20/12/2023
Sửa bài (Revised): 19/01/2024
Duyệt đăng (Accepted): 20/02/2024

Title: *Estimating the parameters of logistic regression with covariates missing at random and application*

Author(s): *Tran Phuoc Loc*, Ta Thi Thanh Thuy, Duong Thi Tuyen, Duong Thi Be Ba, Le Hoai Nhan and Lam Hoang Chuong*

Affiliation(s): *Can Tho University*

TÓM TẮT

Nghiên cứu đề xuất phương pháp ước lượng hệ số của mô hình hồi quy logistic với hiệp biến thiếu dữ liệu ngẫu nhiên. Trước tiên, phương pháp thay thế lặp được sử dụng để thay thế các giá trị thiếu bằng các giá trị hợp lý thu được từ hàm phân phối thực nghiệm có điều kiện. Ước lượng các tham số của mô hình hồi quy và phương sai của nó sau đó thu được lần lượt bằng các phương trình ước lượng và phương sai tương ứng. Các tính chất cỡ mẫu lớn của ước lượng cũng được nghiên cứu. Hiệu quả tính toán của phương pháp đề xuất được nghiên cứu thông qua một số tình huống mô phỏng số và so sánh với các phương pháp khác. Kết quả cho thấy phương pháp đề xuất có hiệu quả tốt hơn các phương pháp xóa hàng, trọng số xác suất nghịch đảo bán tham số, hợp lý có điều kiện và thay thế lặp bằng phương pháp rừng ngẫu nhiên. Dữ liệu thực tế về y học được sử dụng để minh họa khả năng ứng dụng của phương pháp đề xuất.

Từ khoá: *Dữ liệu thiếu, hồi quy logistic, hợp lý cực đại, thay thế lặp*

ABSTRACT

This study proposes a method to estimate the coefficients of the logistic regression model with covariates missing at random. Firstly, the multiple imputation method replaces missing values with reasonable values obtained from an empirical conditional distribution function. The estimator of the parameters of the regression model and its variance are then obtained by the corresponding estimating and equations and estimated variance formulas, respectively. The large sample properties of the estimator are also studied. The estimator efficiency of the proposed method is studied through several numerical simulation scenarios and compared with other methods. The results show that the proposed method has outperformed the other methods, e.g., complete-case, semi-parametric inverse probability weighting, validation likelihood, and random forest multiple imputation methods. Real example data from medical research is used to illustrate the applicability of the proposed method.

Keywords: *Logistic regression, maximum likelihood, missing data, multiple imputation*

1. GIỚI THIỆU

Dữ liệu thiếu hoặc khuyết (missing data) khá phổ biến trong nghiên cứu và việc xử lý loại dữ liệu này đã và đang thu hút sự quan tâm của nhiều nhà khoa học trên thế giới. Dữ liệu thiếu có thể xảy ra do nhiều nguyên nhân: người được khảo sát quên hoặc không trả lời một số câu hỏi trong bảng khảo sát, sai sót trong quá trình nhập liệu, hoặc các thiết bị thu dữ liệu bị hỏng,… Có ba cơ chế thiếu dữ liệu cơ bản: thiếu hoàn toàn ngẫu nhiên (missing completely at random - MCAR), thiếu ngẫu nhiên (missing at random - MAR) và thiếu không ngẫu nhiên (missing not at random - MNAR), trong đó MAR là được nhiều sự quan tâm nghiên cứu (Rubin, 1976). Các lĩnh vực khác nhau như kinh tế, nông nghiệp, y học, sinh học, xã hội học, và giáo dục,… đều có thể gặp thách thức trong việc xử lý số liệu hoặc các vấn đề về độ chính xác của ước lượng hay dự báo do dữ liệu bị khuyết. Do đó, việc xử lý các giá trị khuyết và đưa ra phương pháp phân tích phù hợp là một trong những vấn đề trọng tâm để làm cho mô hình thống kê trở nên hợp lý và chính xác hơn.

Hồi quy logistic là một trong các mô hình phân loại đơn giản, hiệu quả và được sử dụng rộng rãi trong nhiều lĩnh vực, đặc biệt là y học, để tìm hiểu mối liên hệ giữa biến phụ thuộc dạng nhị phân: có/không có bệnh, sống/chết,… và các biến độc lập như tuổi, giới tính, chiều cao, cân nặng, các chỉ số cơ thể,… Tuy nhiên, các biến phụ thuộc có thể bị thiếu dữ liệu do các nguyên nhân khác nhau, điều này có thể dẫn đến giảm hiệu quả ước lượng, phân loại và dự báo của mô hình hồi quy logistic.

Để xử lý dữ liệu thiếu, một số phương pháp đã được đề xuất và áp dụng rộng rãi. Phương pháp xóa hàng (complete case - CC) thường được sử dụng, trong đó các hàng có dữ liệu thiếu sẽ bị xóa đi và việc phân tích dựa vào bộ dữ liệu đầy đủ còn lại. Tuy nhiên, khi tỷ lệ dữ liệu khuyết lớn hoặc cơ chế thiếu dữ liệu không phải là MCAR thì có thể dẫn đến sai lệch nhiều trong kết quả phân tích do thiếu thông tin từ các dòng dữ liệu bị xóa (Wang et al., 1997; Lee et al., 2012). Horvitz and Thompson (1952) là người đầu tiên đề xuất sử dụng trọng số xác suất nghịch đảo (inverse probability weighting - IPW) vào phương trình ước lượng của phương pháp CC để tạo ra ước lượng không chệch cho mô hình. Zhao and Lipsitz (1992) đã cải thiện phương pháp này khi xác suất lựa chọn đã biết. Khi xác suất lựa chọn chưa biết, Wang et al. (1997) và Wang and Wang (2001) đã đề xuất phương pháp IPW bán tham số (semi-parametric IPW - SIPW) để ước lượng tham số của mô hình. Mặc dù hiệu quả hơn phương pháp CC, phương pháp dựa trên trọng số vẫn bỏ qua một số thông tin của dữ liệu do xóa các dòng dữ liệu thiếu. Wang et al. (2002) đã đề xuất hai phương pháp sử dụng hàm hợp lý điều kiện (validation likelihood - VL) và kết hợp các hàm hợp lý có điều kiện (joint conditional likelihood - JCL) cho mô hình hồi quy logistic với hợp biến MAR. Hai phương pháp này sau đó đã được áp dụng trong các nghiên cứu của Lee et al. (2012), Hsieh et al. (2013), và Tran et al. (2023). Mặc dù phương pháp JCL cho kết quả tính toán khả quan và vượt trội so với các phương pháp khác, thuật toán của nó khá phức tạp và đòi hỏi thời gian tính toán dài (Lee et al., 2023). Jiang et al. (2020) đã sử dụng phương pháp xấp xỉ ngẫu nhiên của thuật toán cực đại kỳ vọng (SAEM) để ước lượng tham số cho mô hình hồi quy logistic với hiệp biến thiếu dữ liệu. Kết quả của họ cũng được so sánh với kết quả từ việc sử dụng phương pháp rừng ngẫu nhiên thay thế lặp (RFMI) từ gói *mice* trong phần mềm R (Buuren & Groothuis-Oudshoorn, 2011).

Thay thế lặp (multiple imputation - MI) (Rubin, 1987, 1996) là một trong những phương pháp xử lý dữ liệu khuyết hiệu quả và phổ biến hiện nay. theo một cách tiếp cận khác, ý tưởng chính của MI là bổ sung các giá trị khuyết một cách ngẫu nhiên nhiều lần (M lần) từ các giá trị quan sát được của dữ liệu theo phân phối hay mô hình nào đó, chẳng hạn như giá trị trung bình, trung vị, hồi quy,… để tạo ra M dữ liệu đầy đủ mới. các giá trị thống kê thu được dựa vào trung bình tính toán theo M bộ dữ liệu đầy đủ được dùng làm kết quả phân tích cuối cùng. tuy nhiên, tùy vào sự phân bố của dữ liệu có giá trị khuyết và cơ chế của dữ liệu khuyết mà các phương pháp bổ sung này có thể cho hiệu quả cao hay không. vì vậy, việc tìm ra cơ chế để bổ sung giá trị khuyết của phương pháp mi sao cho tối ưu là một vấn đề thú vị và được nhiều người quan tâm. dựa trên ý tưởng của Fay (1996) và Wang and Chen (2009), Lee et al. (2016, 2020) đã đề xuất các phương pháp mi khác nhau sử dụng các hàm phân phối tích lũy kinh nghiệm (ECDF) để điền khuyết cho dữ liệu thiếu. phương pháp của họ đã được sử dụng cho một số nghiên cứu sau đó và thu được các kết quả rất khả quan so với các phương pháp đã đề cập ở trên (Lukusa et al., 2016; Lee et al., 2021, 2022, 2023). khả năng ứng dụng của các phương pháp mi này trong phân tích dữ liệu thiếu còn rất tiềm năng. do đó, phương pháp mi được tiếp tục nghiên cứu cho mô hình hồi quy logistic với hiệp biến mar và áp dụng chúng trong phân tích dữ liệu thực tế, đặc biệt là dữ liệu y học.

2. MÔ HÌNH VÀ PHƯƠNG PHÁP

2.1. Mô hình

Gọi Y là một biến nhị phân, nhận giá trị 1 nếu sự kiện quan tâm xảy ra và nhận giá trị 0 nếu ngược lại. Giả sử X và Z là các véctơ hiệp biến có số chiều lần lượt là r và s. Trong nghiên cứu này, giả sử X thiếu dữ liệu ngẫu nhiên (MAR) và Z là biến rời rạc và luôn quan sát được. Đặt $\mathcal{X} = (1, X^T, Z^T)^T$. Giả sử rằng cỡ mẫu là n và tất cả quan sát trong mẫu (Y_i, X_i, Z_i) $(i = 1,2, \dots, n)$ là độc lập và cùng phân phối. Xét mô hình hồi quy logistic được cho bởi biểu thức sau:

$$P(Y_i = 1|X_i, Z_i) = H(\beta_0 + \beta_1^T X_i + \beta_2^T Z_i) = H(\beta^T \mathcal{X}_i), \quad (1)$$

trong đó, $\beta = (\beta_0, \beta_1^T, \beta_2^T)^T$là một $r + s + 1$ véctơ hệ số của mô hình hồi quy logistic và

$$H(\beta^T \mathcal{X}_i) = \frac{1}{1 + e^{-\beta^T \mathcal{X}_i}}.$$

Gọi δ là biến đặc trưng cho trạng thái thiếu dữ liệu của X, trong đó $\delta_i = 1$ nếu X_i quan sát được và $\delta_i = 0$ nếu X_i không quan sát được. Gọi W là một biến rời rạc thay thế cho biến X, sao cho W có liên hệ với X nhưng độc lập với Y. Biến thay thế là biến có thể được đo lường dễ dàng hơn và sử dụng cho biến thiếu trong mô hình để cải thiện thông tin mất mát, và do đó tăng hiệu quả phân tích (Wang et al., 1997, 2002; Hsieh et al., 2013; Lee et al., 2012, 2020, 2023). Đặt $V = (Z^T, W^T)^T$. Dựa trên giả thiết rằng cơ chế thiếu dữ liệu của X là MAR, mô hình xác suất chọn được cho như sau:

$$P(\delta_i = 1|Y_i, X_i, Z_i, W_i) = \pi(Y_i, V_i), \quad (2)$$

nghĩa là xác suất của biến trạng thái thiếu (δ) phụ thuộc vào các biến quan sát được (V) nhưng độc lập với biến có dữ liệu thiếu (X). Mục tiêu chính là tìm ước lượng của hệ số β từ biểu thức (1) với các mức xác suất chọn khác nhau. Phần tiếp theo sẽ trình bày các phương pháp ước lượng phổ biến cho mô hình logistic được cho ở (1) khi X là MAR.

2.2. Tóm lược các phương pháp ước lượng

2.2.1. Phương pháp CC

Trước tiên, hàm hợp lý của mô hình hồi quy logistic được cho ở (1) có thể được viết như sau:

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} [H(\beta^T \mathcal{X}_i)]^{Y_i} [1 - H(\beta^T \mathcal{X}_i)]^{1-Y_i}.$$

Từ đó, hàm log-hợp lý được cho dưới đây:

$$\ell(\beta) = \sum_{i=1}^{n} \begin{Bmatrix} Y_i \ln H(\beta^T \mathcal{X}_i) \\ +(1 - Y_i) \ln[1 - H(\beta^T \mathcal{X}_i)] \end{Bmatrix}.$$

Khi một số biến trong dữ liệu bị thiếu giá trị, phương pháp CC thường được sử dụng. Các quan sát hay các dòng dữ liệu có giá trị thiếu sẽ bị xóa đi. Phương pháp CC chỉ sử dụng bộ dữ liệu đầy đủ còn lại để phân tích (dữ liệu CC). Ước lượng CC của β, kí hiệu $\hat{\beta}_C$, thu được từ việc giải phương trình ước lượng sau:

$$U_C(\beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \delta_i \phi_i(\beta) = 0, \quad (3)$$

với

$$\phi_i(\beta) = \mathcal{X}_i [Y_i - H(\beta^T \mathcal{X}_i)]. \quad (4)$$

Khi cơ chế thiếu dữ liệu của X là MAR, $\hat{\beta}_C$ là ước lượng chệch của β, và do đó nó có thể không đáng tin cậy cho dự báo (Wang et al., 1997; Lee et al., 2012).

2.2.2. Phương pháp SIPW

Một phương pháp khá phổ biến để giảm độ chệch khi xóa dữ liệu đó là sử dụng nghịch đảo $\pi(Y, V)$ như là trọng số trong phương trình ước lượng, gọi là phương pháp IPW (Horvitz & Thompson, 1952; Zhao & Lipsitz, 1992). Tuy nhiên, $\pi(Y, V)$ thường không xác định được nên cần phải được ước lượng trước. Wang et al. (1997) và Wang and Wang (2001) đã đề xuất phương pháp SIPW để ước lượng tham số của mô hình hồi quy với hiệp biến là MAR.

Gọi $\hat{\pi}(Y_i, V_i)$ là ước lượng của $\pi(Y_i, V_i)$ trong (2). Ước lượng SIPW của β, kí hiệu $\hat{\beta}_W$, thu được từ việc tìm nghiệm của phương trình ước lượng sau:

$$U_W(\beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}(Y_i, V_i)} \phi_i(\beta) = 0, \quad (5)$$

trong đó, giả sử V là biến rời rạc và cơ chế thiếu dữ liệu của X là MAR, $\hat{\pi}(Y_i, V_i)$ được viết dưới dạng:

$$\hat{\pi}(Y_i, V_i) = \frac{\sum_{j=1}^{n} \delta_j I(Y_j = Y_i, V_j = V_i)}{\sum_{k=1}^{n} I(Y_k = Y_i, V_k = V_i)}, \quad (6)$$

với $I(\cdot)$ là một hàm chỉ số với giá trị 1 hoặc 0.

Với một số điều kiện cụ thể, Wang et al. (1997) và Wang and Wang (2001) đã chỉ ra rằng $\hat{\beta}_W$ là một ước lượng không chệch và vững của β và đề xuất biểu thức ước lượng cho phương sai của $\hat{\beta}_W$.

2.2.3. Phương pháp VL

Wang et al. (2002) đã sử dụng hàm hợp lý điều kiện của Breslow and Cain (1988) để đề xuất phương pháp VL cho mô hình hồi quy logistic với dữ liệu thiếu. Cụ thể, khi X là MAR, ước lượng VL của β, kí hiệu $\hat{\beta}_V$, là nghiệm của phương trình:

$$U_V(\beta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \delta_i \mathcal{X}_i [Y_i - H_+(\beta^T \mathcal{X}_i)] = 0, \quad (7)$$

trong đó

$$H_+(\beta^T \mathcal{X}_i) = H\left(\beta^T \mathcal{X}_i + \ln \frac{\hat{\pi}(0, V_i)}{\hat{\pi}(1, V_i)}\right),$$

với $\hat{\pi}(Y_i, V_i)$ được định nghĩa trong (6). Wang et al. (2002) đã chỉ ra rằng $\hat{\beta}_V$ là một ước lượng không chệch của β.

2.3. Phương pháp MI

2.3.1. Thuật toán

Phần này trình bày thuật toán của phương pháp MI thông qua việc sử dụng hàm ECDF được đề xuất bởi Wang and Chen (2009) để lấy dữ liệu ngẫu nhiên thích hợp và điền vào các giá trị khuyết. Phương pháp MI này đã được chứng tỏ là hiệu quả tính toán cao và thời gian tính tương đối ngắn so với một số phương pháp MI khác được đề cập trong các nghiên cứu gần đây (Lee et al., 2016, 2020, 2021, 2022, 2023). Cụ thể, khi X thiếu dữ liệu ngẫu nhiên, hàm ECDF $\hat{F}(x|Y_i, V_i)$ sau đây được sử dụng:

$$\hat{F}(x|Y_i, V_i) = \sum_{k=1}^{n} \left(\frac{\delta_k I(Y_k = Y_i) I(V_k = V_i)}{\sum_{s=1}^{n} I(Y_s = Y_i) I(V_s = V_i)} \right) I(X_k \le x), \quad (8)$$

với $i = 1,2, \dots, n$, để tạo ra bộ số liệu thay thế giá trị khuyết của X.

Gọi M là số lần thay thế lặp. Thuật toán MI, dựa trên ý tưởng của Fay (1996), gồm 2 bước sau đây.

Bước 1. Tạo bộ dữ liệu đầy đủ thứ v ($v = 1,2, \dots, M$) dựa vào trạng thái thiếu dữ liệu của $X_i, i = 1,2, \dots, n$ như sau:

Nếu $\delta_i = 1$, khi X_i quan sát được, giữ nguyên giá trị của X_i. Đặt $\mathcal{X}_i = (1, X_i^T, Z_i^T)^T$ với tất cả v.

Nếu $\delta_i = 0$, khi X_i là MAR, tạo ra $\tilde{X}_{iv}$ từ bộ giá trị của hàm ECDF $\hat{F}(x|Y_i, V_i)$ để điền vào các giá trị khuyết của X_i. Đặt $\tilde{\mathcal{X}}_{iv} = \left(1, \tilde{X}_{iv}^T, Z_i^T\right)^T$.

Bước 2. Giải phương trình ước lượng sau:

$$U_M(\beta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \left[\delta_i \phi_i(\beta) + (1 - \delta_i)\tilde{\phi}_i(\beta)\right] = 0, \quad (9)$$

với $\phi_i(\beta)$ được cho ở (4), $\phi_{iv}(\beta) = \tilde{\mathcal{X}}_{iv}\left[Y_i - H\left(\beta^T \tilde{\mathcal{X}}_{iv}\right)\right]$, và $\tilde{\phi}_i(\beta) = \frac{1}{M}\sum_{v=1}^{M} \phi_{iv}(\beta)$.

Nghiệm thu được từ phương trình (9) được gọi là ước lượng MI của β và kí hiệu là $\hat{\beta}_M$. Một trong những thuận lợi cơ bản của phương pháp này là phương trình ước lượng chỉ được giải một lần thay vì M lần như các phương pháp MI truyền thống khác, điều này giúp cải thiện tốc độ tính toán của thuật toán. Tiếp theo, ước lượng phương sai của $\hat{\beta}_M$ có thể thu được theo phương pháp của Rubin (1987) và Lee et al. (2020) bằng cách thay thế β bởi $\hat{\beta}_M$ vào biểu thức sau:

$$G_M^{-1}(\beta)\mathcal{M}(\beta)[G_M^{-1}(\beta)]^T,$$

trong đó

$$\mathcal{M}(\beta) = \frac{1}{M}\sum_{v=1}^{M}\sum_{i=1}^{n} [U_{iv}(\beta)]^{\otimes 2} + \frac{M+1}{M(M-1)}\sum_{v=1}^{M} [U_v(\beta)]^{\otimes 2},$$

với

$$U_{iv}(\beta) = \frac{1}{\sqrt{n}}\left[\delta_i \phi_i(\beta) + (1 - \delta_i)\tilde{\phi}_i(\beta)\right],$$

$$U_v(\beta) = \sum_{i=1}^{n} U_{iv}(\beta),$$

$G_M^{-1}(\beta)$ là đạo hàm của $-U_M(\beta)$ theo β, và $a^{\otimes 2} = aa^T$ với mọi véctơ cột a.

2.3.2. Tính chất hội tụ

Phần này trình bày kết quả ngắn gọn về tính chất hội tụ của $\hat{\beta}_M$ với cỡ mẫu lớn. Trước tiên, giả sử một số điều kiện sau đây thỏa mãn:

(C1) Gọi $supp(V)$ là support của V. Giả sử $supp(V)$ không phụ thuộc vào β. Giả sử với mọi $y = 0, 1, v \in supp(V)$ thì $\pi(y, v) > 0$.

(C2) $E\left[[\phi_1(\beta)]^{\otimes 2}\right]$ là xác định dương trong lân cận của β.

(C3) $E\left[\frac{[\phi_1(\beta)]^{\otimes 2}}{\pi(Y_1, V_1)}\right]$ là hữu hạn và xác định dương trong lân cận của β.

(C4) Đạo hàm bậc nhất của $U_W(\beta)$ và $U_M(\beta)$ theo β tồn tại gần như chắc chắn trong một lân cận

của β. Ngoài ra, trong một lân cận như vậy, các đạo hàm này bị chặn trên bởi hàm của (Y, X, V), mà kỳ vọng của hàm này tồn tại.

Bổ đề 1. *Với các điều kiện (C1)-(C4), khi* $n, M \to \infty$, $U_M(\beta)$ *có thể biểu diễn như sau:*

$$U_M(\beta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i}{\pi(Y_i,V_i)}\phi_i(\beta) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(1-\frac{\delta_i}{\pi(Y_i,V_i)}\right)\phi_i^*(\beta) + O_p\left(M^{-1/2}\right) + o_p(1), \quad (10)$$

với $\phi_i^*(\beta) = E[\phi_i(\beta)|Y_i, V_i]$.

Chứng minh.

Đặt $\mathcal{D} = (Y, V^T)^T$. Kỳ vọng của $\phi_{iv}(\beta)$ theo hàm ECDF trong (8) có thể được viết sau đây

$$E_{\hat{F}}[\phi_{iv}(\beta)|\mathcal{D}_i] = \sum_{k=1}^{n}\left(\frac{\delta_k\phi_i(\beta)I(Y_k=Y_i)I(V_k=V_i)}{\sum_{s=1}^{n}I(Y_s=Y_i)I(V_s=V_i)}\right).$$

Từ đó ta có thể thu được

$$\sum_{k=1}^{n}(1-\delta_i)E_{\hat{F}}[\phi_{iv}(\beta)|\mathcal{D}_i] = \sum_{k=1}^{n}\delta_i\phi_i(\beta)\left(\frac{1}{\hat{\pi}(Y_i,V_i)}-1\right).$$

Do đó, ta có

$$E_{\hat{F}}[U_M(\beta)|\mathcal{D}_i] = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i\phi_i(\beta)}{\hat{\pi}(Y_i,V_i)} = U_W(\beta).$$

Tương tự, ta cũng có thể chứng minh

$$E_{\hat{F}}\left(\frac{\partial U_M(\beta)}{\partial\beta^T}\middle|\mathcal{D}_i\right) = \frac{\partial U_W(\beta)}{\partial\beta^T}.$$

$U_M(\beta)$ có thể viết dưới dạng sau:

$$U_M(\beta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_i\phi_i(\beta) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\delta_i)E_{\hat{F}}[\phi_{iv}(\beta)|\mathcal{D}_i] + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\delta_i)\{\tilde{\phi}_i(\beta)-E_{\hat{F}}[\phi_{iv}(\beta)|\mathcal{D}_i]\}.$$

Ta có thể chứng minh rằng

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\delta_i)\{\tilde{\phi}_i(\beta)-E_{\hat{F}}[\phi_{iv}(\beta)|\mathcal{D}_i]\} = O_p(M^{-1/2}),$$

nên theo kết quả của Lee et al. (2020, 2022), rằng

$$U_W(\beta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i\phi_i(\beta)}{\hat{\pi}(Y_i,V_i)} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{\delta_i}{\pi(Y_i,V_i)}\phi_i(\beta) + \left(1-\frac{\delta_i}{\pi(Y_i,V_i)}\right)\phi_i^*(\beta)\right] + o_p(1),$$

ta thu được điều phải chứng minh. ■

Định lý 1. Với các điều kiện (C1)-(C4), $\sqrt{n}(\hat{\beta}_M - \beta)$ hội tụ theo phân phối về phân phối chuẩn với trung bình 0 và phương sai $\Delta(\pi,\beta) = F^{-1}(\pi,\beta)Q(\pi,\beta)[F^{-1}(\pi,\beta)]^T$, với

$$Q(\pi,\beta) = E\left\{\left[\begin{array}{c}\frac{\delta_1}{\pi(Y_1,V_1)}\phi_1(\beta)\\ +\left(1-\frac{\delta_1}{\pi(Y_1,V_1)}\right)\phi_1^*(\beta)\end{array}\right]^{\otimes 2}\right\}, \quad (11)$$

$F^{-1}(\pi,\beta)$ là đạo hàm của $-U_W(\beta)$ theo β.

Chứng minh.

Tương tự **Bổ đề 1**, ta cũng chứng minh được

$$\frac{\partial U_M(\beta)}{\partial\beta^T} = \frac{\partial U_W(\beta)}{\partial\beta^T} + O_p\left(M^{-\frac{1}{2}}\right).$$

Vì $\hat{\beta}_M$ là nghiệm của phương trình (9), nên khai triển Taylor của $U_M(\hat{\beta}_M)$ tại β ta được

$$\begin{aligned}0 &= U_M(\hat{\beta}_M)\\ &= U_M(\beta) + \left(\frac{\partial U_M(\beta)}{\sqrt{n}\partial\beta^T}\right)\sqrt{n}(\hat{\beta}_M-\beta) + o_p(1)\\ &= U_M(\beta) + \left(\frac{\partial U_W(\beta)}{\sqrt{n}\partial\beta^T}\right)\sqrt{n}(\hat{\beta}_M-\beta) + o_p(1)\\ &= U_M(\beta) - F(\pi,\beta)\sqrt{n}(\hat{\beta}_M-\beta) + O_p\left(M^{-\frac{1}{2}}\right) + o_p(1),\end{aligned}$$

với $F(\pi,\beta) = \frac{\partial U_W(\beta)}{\sqrt{n}\partial\beta^T}$. Do đó,

$$\sqrt{n}(\hat{\beta}_M-\beta) = F^{-1}(\pi,\beta)U_M(\beta) + O_p\left(M^{-\frac{1}{2}}\right) + o_p(1).$$

Nên ta thu được

$$\sqrt{n}\left(\hat{\beta}_M - \beta\right) \xrightarrow{d} \mathcal{N}(0, \Delta(\pi,\beta)) \text{ khi } n, M \to \infty,$$

với $\Delta(\pi,\beta) = F^{-1}(\pi,\beta)Q(\pi,\beta)[F^{-1}(\pi,\beta)]^T$ và $Q(\pi,\beta)$ được cho trong (11). ■

3. MÔ PHỎNG

Phần này tập trung vào việc nghiên cứu hiệu quả của phương pháp đề xuất thông qua các tình huống mô phỏng số giả định. Các mô phỏng Monte Carlo bao gồm 1.000 lần lặp lại được áp dụng để đánh giá các ước lượng sau:

- $\hat{\beta}_F$: ước lượng hợp lý cực đại (ML) khi dữ liệu đầy đủ và dùng như một tiêu chuẩn để so sánh,
- $\hat{\beta}_C$: ước lượng CC từ phương trình (3),
- $\hat{\beta}_W$: ước lượng SIPW từ phương trình (5),
- $\hat{\beta}_V$: ước lượng VL từ phương trình (7),
- $\hat{\beta}_R$: ước lượng RFMI từ gói *mice* trong R,
- $\hat{\beta}_M$: ước lượng MI từ phương trình (9).

Các giá trị đặc trưng của 6 ước lượng này sẽ được tính toán, bao gồm độ chệch (Bias), độ lệch chuẩn (SD), xấp xỉ độ lệch chuẩn (ASE), tổng bình phương của bias và SD (MSE), và tỷ lệ các ước lượng nằm trong khoảng tin cậy 95% (CP). Ba trường hợp mô phỏng được xem xét. Trường hợp đầu tiên nghiên cứu sự ảnh hưởng của các cỡ mẫu khác nhau đến kết quả tính tính, trong khi trường hợp thứ hai xem xét các tỷ lệ thiếu dữ liệu khác nhau, và cuối cùng là số lần thay thế lặp M thay đổi.

Trong trường hợp đầu tiên, ba cỡ mẫu khác nhau được sử dụng, $n = \{500, 1.000, 1.500\}$, để nghiên cứu liệu có sự ảnh hưởng của cỡ mẫu đến hiệu quả của các phương pháp ước lượng hay không trên cùng mức độ thiếu dữ liệu và $M = 20$. Phân phối đều $U(-1{,}5, 1{,}5)$ được sử dụng để tạo dữ liệu cho biến X. Dữ liệu của biến Z được tạo ra từ phân phối Bernoulli $Ber(0{,}5)$. Gọi W là một biến thay thế của X sao cho $W = 1$ nếu $X \geq 0$ và $W = 0$ nếu $X < 0$. Phân phối $Ber(p)$ được dùng để tạo ra dữ liệu cho Y với $p = P(Y = 1|X, Z) = H(\beta_0 + \beta_1 X + \beta_2 Z)$, trong đó $\beta = (\beta_0, \beta_1, \beta_2)^T = (-\ln 2, \ln 3, \ln 3)^T$. Với giả sử X là MAR, dữ liệu của biến thiếu δ được tạo ra bằng cách sử dụng mô hình xác suất $P(\delta = 1|Y, Z, W) = \alpha_0 + \alpha_1 Y + \alpha_2 Z + \alpha_3 W$, với $\alpha = (\alpha_0, \alpha_1, \alpha_2, \alpha_3)^T = (1, -1, -0{,}5, 0{,}5)^T$. Từ đó, tỷ lệ thiếu dữ liệu của X là 38%, nghĩa là dữ liệu CC chiếm 62% dữ liệu ban đầu. Kết quả tính toán của trường hợp này được trình bày trong Bảng 1.

Với trường hợp thứ hai, ba mức độ thiếu dữ liệu khác nhau được sử dụng trên dữ liệu có cùng cỡ mẫu 1000 và $M = 20$. Dữ liệu của X và Z được tạo ra tương ứng từ phân phối chuẩn tắc $N(0, 1)$ và $Ber(0{,}5)$. Biến W và Y được tạo ra với cách tương tự như trong Trường hợp 1 và β không đổi. Giá trị của α được chọn lần lượt là $(1{,}5, -1, -0{,}5, 0{,}5)^T$, $(1, -1, -0{,}5, 0{,}5)^T$ và $(0{,}2, -1, -0{,}5, 0{,}5)^T$ để tạo tỷ lệ thiếu dữ liệu tương ứng là 28%, 38% và 56%. Các kết quả phân tích của trường hợp này được mô tả trong Bảng 2.

Trong trường hợp cuối, số lần thay thế khác nhau, $M = \{5, 15, 30\}$, được dùng cho phương pháp MI và RFMI trên dữ liệu có cỡ mẫu 1.000 và cùng mức độ thiếu dữ liệu. Các biến $\{X, Z, W, Y\}$ và tham số β được giữ như trong Trường hợp 2, trong khi giá trị của α là $(0{,}2, -1, -0{,}5, 0{,}5)^T$ để tạo tỷ lệ thiếu dữ liệu 56%. Các kết tính toán tương ứng được trình bày trong Bảng 3 và 4.

Nhìn chung, qua các kết quả phân tích trong Bảng 1-4 ta có thể nhận xét rằng hiệu quả ước lượng của $\widehat{\boldsymbol{\beta}}_F$ là tốt nhất, của $\widehat{\boldsymbol{\beta}}_C$ là tệ nhất và $\widehat{\boldsymbol{\beta}}_M$ tốt hơn các ước lượng còn lại. Tuy nhiên, ước lượng $\widehat{\boldsymbol{\beta}}_F$ sử dụng dữ liệu đầy đủ để tính toán, được dùng như một tiêu chuẩn để so sánh với các phương pháp khác. Ước lượng $\widehat{\boldsymbol{\beta}}_C$ sử dụng dữ liệu CC, dữ liệu đã xóa đi các dòng thiếu giá trị, do đó hiệu quả tính toán là rất thấp thông qua giá trị tuyệt đối của Bias rất lớn, các giá trị SD, ASE và MSE cao và giá trị CP thấp hơn mức xác suất danh nghĩa 95%, nghĩa là tỷ lệ các giá trị ước lượng trong 1.000 lần mô phỏng lọt vào khoảng tin cậy 95% là rất thấp. Ngoài ra, Bảng 1 cho thấy hiệu quả của các phương pháp đều tăng khi cỡ mẫu tăng từ 500 lên 1.500, thông qua các giá trị Bias, SD, ASE và MSE giảm. Xét riêng 4 cột cuối, ước lượng $\widehat{\boldsymbol{\beta}}_M$ từ phương pháp đề xuất nhìn chung hiệu quả hơn của $\widehat{\boldsymbol{\beta}}_W$, $\widehat{\boldsymbol{\beta}}_V$ và $\widehat{\boldsymbol{\beta}}_R$. Ước lượng $\widehat{\boldsymbol{\beta}}_R$ từ phương pháp RFMI có hiệu quả thấp nhất khi giá trị Bias cao và CP thấp cho trường hợp β_1, mặc dù giá trị SD của nó thấp hơn của 3 phương pháp còn lại. Bên cạnh đó, các nhận xét trong Bảng 1 cũng tương tự như trong Bảng 2 và 3. Bảng 2 cho thấy hiệu quả của các phương pháp dành cho dữ liệu thiếu, tức là 5 cột cuối cùng, giảm đi khi tỷ lệ dữ liệu thiếu tăng từ 28% lên 56%. Ngoài ra, khi số lần thay thế lặp M tăng từ 5 lên 30, kết quả từ Bảng 3 cho thấy hiệu quả ước lượng của phương pháp RFMI và phương pháp đề xuất MI tăng nhẹ. Cuối cùng, Bảng 4 cho thấy thời gian tính toán của RFMI nhiều hơn và tăng nhanh hơn so với của MI khi giá trị M tăng.

Bảng 1. Bảng tóm tắt kết quả mô phỏng của Trường hợp 1 với cỡ mẫu 500, 1.000 và 1.500, $M = 20$, và tỷ lệ thiếu dữ liệu là 38%

n			$\widehat{\boldsymbol{\beta}}_F$	$\widehat{\boldsymbol{\beta}}_C$	$\widehat{\boldsymbol{\beta}}_W$	$\widehat{\boldsymbol{\beta}}_V$	$\widehat{\boldsymbol{\beta}}_R$	$\widehat{\boldsymbol{\beta}}_M$
500	β_0	Bias	-0,0078	-0,3419	-0,0086	-0,0092	0,0032	-0,0085
		SD	0,1452	0,1832	0,1470	0,1491	0,1387	0,1471
		ASE	0,1477	0,1897	0,1504	0,1529	0,1517	0,1494
		MSE	0,0212	0,1504	0,0217	0,0223	0,0192	0,0217
		CP	0,9530	0,5930	0,9570	0,9520	0,9700	0,9540
	β_1	Bias	0,0075	0,0738	0,0090	0,0090	-0,1399	0,0087
		SD	0,1262	0,1724	0,1410	0,1405	0,1323	0,1419
		ASE	0,1289	0,1716	0,1440	0,1439	0,1701	0,1365
		MSE	0,0160	0,0352	0,0200	0,0198	0,0371	0,0202
		CP	0,9600	0,9390	0,9640	0,9670	0,8890	0,9490
	β_2	Bias	0,0081	-0,0968	0,0082	0,0096	-0,0813	0,0080
		SD	0,2067	0,2663	0,2104	0,2131	0,1950	0,2105
		ASE	0,2076	0,2709	0,2115	0,2150	0,2094	0,2100
		MSE	0,0428	0,0803	0,0443	0,0455	0,0446	0,0444
		CP	0,9480	0,9440	0,9520	0,9500	0,9460	0,9490
1000	β_0	Bias	-0,0018	-0,3358	-0,0021	-0,0023	0,0086	-0,0022
		SD	0,1056	0,1334	0,1064	0,1068	0,1006	0,1065
		ASE	0,1040	0,1332	0,1058	0,1071	0,1078	0,1051
		MSE	0,0112	0,1305	0,0113	0,0114	0,0102	0,0114
		CP	0,9460	0,2710	0,9540	0,9560	0,9690	0,9520
	β_1	Bias	0,0035	0,0687	0,0043	0,0049	-0,1432	0,0047
		SD	0,0929	0,1253	0,1064	0,1057	0,0967	0,1070
		ASE	0,0907	0,1205	0,1016	0,1011	0,1308	0,0963
		MSE	0,0086	0,0204	0,0113	0,0112	0,0299	0,0115
		CP	0,9480	0,9160	0,9450	0,9430	0,8280	0,9300
	β_2	Bias	0,0002	-0,1015	0,0002	0,0008	-0,0886	0,0003
		SD	0,1458	0,1866	0,1468	0,1485	0,1353	0,1471
		ASE	0,1462	0,1901	0,1488	0,1508	0,1476	0,1478
		MSE	0,0213	0,0451	0,0215	0,0220	0,0261	0,0216
		CP	0,9430	0,9220	0,9500	0,9520	0,9300	0,9490
1500	β_0	Bias	-0,0027	-0,3292	-0,0031	-0,0039	0,0076	-0,0032
		SD	0,0794	0,1039	0,0806	0,0815	0,0776	0,0806
		ASE	0,0849	0,1085	0,0864	0,0873	0,0889	0,0858
		MSE	0,0063	0,1192	0,0065	0,0067	0,0061	0,0065
		CP	0,9640	0,1090	0,9680	0,9650	0,9770	0,9670
	β_1	Bias	0,0050	0,0659	0,0057	0,0064	-0,1455	0,0058
		SD	0,0746	0,1015	0,0841	0,0839	0,0797	0,0846
		ASE	0,0740	0,0981	0,0829	0,0824	0,1155	0,0787
		MSE	0,0056	0,0147	0,0071	0,0071	0,0275	0,0072
		CP	0,9520	0,8970	0,9460	0,9470	0,8090	0,9360
	β_2	Bias	0,0031	-0,1068	0,0032	0,0043	-0,0832	0,0031
		SD	0,1167	0,1554	0,1188	0,1197	0,1119	0,1189
		ASE	0,1193	0,1550	0,1214	0,1229	0,1209	0,1207
		MSE	0,0136	0,0355	0,0141	0,0143	0,0194	0,0141
		CP	0,9590	0,8820	0,9610	0,9610	0,9050	0,9600

Bảng 2. Bảng tóm tắt kết quả mô phỏng của Trường hợp 2 với cỡ mẫu 1.000, $M = 20$, và tỷ lệ thiếu dữ liệu là 28%, 38% và 56%

% thiếu			$\widehat{\boldsymbol{\beta}}_F$	$\widehat{\boldsymbol{\beta}}_C$	$\widehat{\boldsymbol{\beta}}_W$	$\widehat{\boldsymbol{\beta}}_V$	$\widehat{\boldsymbol{\beta}}_R$	$\widehat{\boldsymbol{\beta}}_M$
28%	β_0	Bias	-0,0065	-0,2378	-0,0076	-0,0078	0,0069	-0,0077
		SD	0,1022	0,1222	0,1036	0,1045	0,0988	0,1034
		ASE	0,1054	0,1241	0,1070	0,1077	0,1082	0,1066
		MSE	0,0105	0,0715	0,0108	0,0110	0,0098	0,0107
		CP	0,9610	0,5180	0,9620	0,9590	0,9630	0,9610
	β_1	Bias	0,0049	0,0541	0,0064	0,0057	-0,1150	0,0064
		SD	0,0921	0,1093	0,1006	0,1002	0,0886	0,1005
		ASE	0,0889	0,1078	0,0973	0,0968	0,1190	0,0941
		MSE	0,0085	0,0149	0,0102	0,0101	0,0211	0,0101
		CP	0,9440	0,9240	0,9460	0,9450	0,8800	0,9340
	β_2	Bias	0,0096	-0,0871	0,0104	0,0105	-0,0653	0,0104
		SD	0,1413	0,1744	0,1436	0,1446	0,1359	0,1434
		ASE	0,1481	0,1768	0,1505	0,1515	0,1495	0,1499
		MSE	0,0201	0,0380	0,0207	0,0210	0,0227	0,0207
		CP	0,9570	0,9220	0,9580	0,9560	0,9400	0,9570
38%	β_0	Bias	-0,0065	-0,3389	-0,0086	-0,0087	0,0133	-0,0086
		SD	0,1022	0,1359	0,1066	0,1075	0,0994	0,1064
		ASE	0,1054	0,1349	0,1080	0,1095	0,1098	0,1070
		MSE	0,0105	0,1333	0,0114	0,0116	0,0101	0,0114
		CP	0,9610	0,2820	0,9580	0,9580	0,9700	0,9560
	β_1	Bias	0,0049	0,0664	0,0069	0,0059	-0,1684	0,0068
		SD	0,0921	0,1234	0,1063	0,1051	0,0905	0,1063
		ASE	0,0889	0,1180	0,1024	0,1015	0,1335	0,0958
		MSE	0,0085	0,0196	0,0113	0,0111	0,0365	0,0113
		CP	0,9440	0,9100	0,9520	0,9480	0,7900	0,9350
	β_2	Bias	0,0096	-0,0971	0,0117	0,0118	-0,0893	0,0117
		SD	0,1413	0,1881	0,1459	0,1465	0,1326	0,1457
		ASE	0,1481	0,1927	0,1520	0,1540	0,1499	0,1506
		MSE	0,0201	0,0448	0,0214	0,0216	0,0256	0,0214
		CP	0,9570	0,9180	0,9540	0,9530	0,9330	0,9540
56%	β_0	Bias	-0,0065	-0,5282	-0,0074	-0,0065	0,0324	-0,0075
		SD	0,1022	0,1624	0,1095	0,1120	0,0986	0,1095
		ASE	0,1054	0,1635	0,1112	0,1150	0,1120	0,1079
		MSE	0,0105	0,3054	0,0121	0,0126	0,0108	0,0120
		CP	0,9610	0,0680	0,9530	0,9550	0,9640	0,9480
	β_1	Bias	0,0049	0,0678	0,0123	0,0078	-0,2771	0,0126
		SD	0,0921	0,1469	0,1239	0,1197	0,0949	0,1243
		ASE	0,0889	0,1443	0,1165	0,1149	0,1579	0,0989
		MSE	0,0085	0,0262	0,0155	0,0144	0,0858	0,0156
		CP	0,9440	0,9320	0,9240	0,9330	0,5920	0,8800
	β_2	Bias	0,0096	-0,0979	0,0094	0,0083	-0,1292	0,0094
		SD	0,1413	0,2304	0,1507	0,1533	0,1307	0,1505
		ASE	0,1481	0,2348	0,1566	0,1608	0,1491	0,1518
		MSE	0,0201	0,0627	0,0228	0,0236	0,0338	0,0227
		CP	0,9570	0,9360	0,9520	0,9530	0,8910	0,9470

Bảng 3. Bảng tóm tắt kết quả mô phỏng của Trường hợp 3 với cỡ mẫu 1.000, $M = \{5, 15, 30\}$, và tỷ lệ thiếu dữ liệu là 56%

		$\widehat{\boldsymbol{\beta}}_F$	$\widehat{\boldsymbol{\beta}}_C$	$\widehat{\boldsymbol{\beta}}_W$	$\widehat{\boldsymbol{\beta}}_V$	$M = 5$		$M = 15$		$M = 30$	
						$\widehat{\boldsymbol{\beta}}_R$	$\widehat{\boldsymbol{\beta}}_M$	$\widehat{\boldsymbol{\beta}}_R$	$\widehat{\boldsymbol{\beta}}_M$	$\widehat{\boldsymbol{\beta}}_R$	$\widehat{\boldsymbol{\beta}}_M$
β_0	Bias	-0,0065	-0,5282	-0,0074	-0,0065	0,0324	-0,0072	0,0315	-0,0075	0,0319	-0,0073
	SD	0,1022	0,1624	0,1095	0,1120	0,0998	0,1096	0,0988	0,1094	0,0994	0,1095
	ASE	0,1054	0,1635	0,1112	0,1150	0,1128	0,1081	0,1121	0,1080	0,1118	0,1079
	MSE	0,0105	0,3054	0,0121	0,0126	0,0110	0,0121	0,0107	0,0120	0,0109	0,0120
	CP	0,9610	0,0680	0,9530	0,9550	0,9620	0,9470	0,9590	0,9490	0,9580	0,9470
β_1	Bias	0,0049	0,0678	0,0123	0,0078	-0,2756	0,0131	-0,2752	0,0130	-0,2748	0,0124
	SD	0,0921	0,1469	0,1239	0,1197	0,1037	0,1261	0,0959	0,1244	0,0928	0,1241
	ASE	0,0889	0,1443	0,1165	0,1149	0,1605	0,1001	0,1588	0,0990	0,1564	0,0987
	MSE	0,0085	0,0262	0,0155	0,0144	0,0867	0,0161	0,0849	0,0157	0,0841	0,0155
	CP	0,9440	0,9320	0,9240	0,9330	0,5610	0,8790	0,5900	0,8770	0,5910	0,8760
β_2	Bias	0,0096	-0,0979	0,0094	0,0083	-0,1290	0,0085	-0,1288	0,0093	-0,1290	0,0094
	SD	0,1413	0,2304	0,1507	0,1533	0,1325	0,1509	0,1307	0,1506	0,1314	0,1505
	ASE	0,1481	0,2348	0,1566	0,1608	0,1502	0,1521	0,1493	0,1519	0,1492	0,1518
	MSE	0,0201	0,0627	0,0228	0,0236	0,0342	0,0229	0,0337	0,0228	0,0339	0,0227
	CP	0,9570	0,9360	0,9520	0,9530	0,8830	0,9470	0,8910	0,9470	0,8910	0,9470

Bảng 4. Bảng tóm tắt thời gian tính toán mô phỏng của Trường hợp 3 với cỡ mẫu 1.000, $M = \{5, 15, 30\}$, và tỷ lệ thiếu dữ liệu là 56%

Thời gian	$\widehat{\boldsymbol{\beta}}_F$	$\widehat{\boldsymbol{\beta}}_C$	$\widehat{\boldsymbol{\beta}}_W$	$\widehat{\boldsymbol{\beta}}_V$	$M = 5$		$M = 15$		$M = 30$	
					$\widehat{\boldsymbol{\beta}}_R$	$\widehat{\boldsymbol{\beta}}_M$	$\widehat{\boldsymbol{\beta}}_R$	$\widehat{\boldsymbol{\beta}}_M$	$\widehat{\boldsymbol{\beta}}_R$	$\widehat{\boldsymbol{\beta}}_M$
Min	0,01	0,01	0,32	0,08	0,36	0,07	1,07	0,17	2,08	0,34
Mean	0,03	0,03	0,67	0,19	0,77	0,18	2,19	0,46	4,32	0,90
Max	0,05	0,12	0,94	0,35	0,19	0,28	3,29	0,59	6,00	1,15

Bảng 5. Bảng tóm tắt kết quả phân tích mô hình hồi quy logistic cho dữ liệu GLOW500 với tỷ lệ thiếu dữ liệu là 44,2%

		$\widehat{\boldsymbol{\beta}}_F$	$\widehat{\boldsymbol{\beta}}_C$	$\widehat{\boldsymbol{\beta}}_W$	$\widehat{\boldsymbol{\beta}}_V$	$\widehat{\boldsymbol{\beta}}_R$	$\widehat{\boldsymbol{\beta}}_M$
β_0	Ước lượng	-0,6736	-1,0616	-0,6175	-0,5231	-0,6738	-0,6192
	ASE	0,2045	0,2994	0,2372	0,2542	0,2228	0,2339
	p-giá trị	0,0010	0,0004	0,0092	0,0396	0,0025	0,0081
$\beta_1\ (X)$	Ước lượng	-0,6824	-0,8663	-0,8026	-0,8004	-0,6699	-0,6920
	ASE	0,2496	0,3848	0,3565	0,3949	0,3487	0,3311
	p-giá trị	0,0063	0,0244	0,0244	0,0427	**0,0547**	0,0367
$\beta_2\ (Z_1)$	Ước lượng	0,7616	1,0243	0,7705	0,5775	0,7803	0,7705
	ASE	0,2349	0,3290	0,2523	0,2704	0,2372	0,2522
	p-giá trị	0,0012	0,0018	0,0023	0,0327	0,0010	0,0022
$\beta_3\ (Z_2)$	Ước lượng	-1,1788	-1,215	-1,2037	-1,3122	-1,1929	-1,2028
	ASE	0,3268	0,4947	0,3498	0,4117	0,3297	0,3481
	p-giá trị	0,0003	0,0141	0,0006	0,0014	0,0003	0,0005
$\beta_4\ (Z_3)$	Ước lượng	-0,6011	-0,7760	-0,6932	-0,5998	-0,6439	-0,6318
	ASE	0,2401	0,3720	0,2481	0,2853	0,2452	0,2422
	p-giá trị	0,0123	0,0370	0,0052	0,0355	0,0086	0,0091

4. ÁP DỤNG

Phần này nghiên cứu về khả năng áp dụng của phương pháp MI và các phương pháp ước lượng khác đã được trình bày ở các phần trên. Bộ dữ liệu y tế GLOW500 được sử dụng gồm 500 bệnh nhân được chọn ra từ *Nghiên cứu theo chiều dọc toàn cầu về bệnh loãng xương ở phụ nữ* từ 55 tuổi trở lên ở 10 quốc gia khác nhau, để nghiên cứu một số nguy cơ tác động đến việc gãy xương của phụ nữ lớn tuổi. Dữ liệu GLOW500 được công bố bởi Hosmer et al. (2013) và đã được sử dụng bởi một số nghiên cứu gần đây (Tran et al., 2023; Lee et al., 2023).

Trong nghiên cứu này, các biến trong mô hình hồi quy logistic được sử dụng như sau: Biến nhị

phân Y có giá trị 1 nếu có bất kỳ gãy xương nào trong năm đầu tiên ($fracture = 1$) và Y có giá trị 0 nếu không có bất kỳ gãy xương nào trong năm đầu tiên ($fracture = 0$). Biến X có giá trị 1 nếu nguy cơ gãy xương tự báo nhỏ hơn những người khác cùng tuổi ($raterisk = 1$) và X có giá trị 0 nếu nguy cơ gãy xương tự báo lớn hơn hoặc bằng những người khác cùng tuổi ($raterisk > 1$). Biến $Z_1 = 1$ nếu bệnh nhân có tiền sử gãy xương trước đó ($priorfrac = 1$) và $Z_1 = 0$ nếu ngược lại ($priorfrac = 0$). Các mức độ tuổi của bệnh nhân được sử dụng như là các biến độc lập dạng nhị phân, trong đó $Z_2 = 1$ nếu tuổi bệnh nhân không quá 60 ($age \leq 60$) và $Z_2 = 0$ nếu ngược lại, và $Z_3 = 1$ nếu tuổi bệnh nhân trên 60 và không quá 70 ($60 < age \leq 70$) và $Z_3 = 0$ nếu ngược lại. Cuối cùng, biến $W = 1$ nếu mẹ của bệnh nhân từng bị gãy xương ($momfrac = 1$) và $W = 0$ nếu ngược lại.

Chú ý rằng mặc dù các biến được sử dụng tương tự như trong Lee et al. (2023) nhưng nghiên cứu này chỉ sử dụng một biến X với dữ liệu thiếu thay vì hai biến như trong nghiên cứu của họ. Ngoài ra, để tăng tỷ lệ dữ liệu thiếu cho X và thỏa mãn cơ chế thiếu dữ liệu theo MAR, biến δ được tạo theo phân phối Bernoulli với xác suất thành công là $P(\delta = 1|Y, Z_1, Z_2, Z_3, W) = \eta_0 + \eta_1 Y + \eta_2 Z_1 + \eta_3 Z_2 + \eta_4 Z_3 + \eta_5 W$, với $(\eta_0, \eta_1, \eta_2, \eta_3, \eta_4, \eta_5)^T = (0{,}5, -0{,}5, 0{,}6, -0{,}5, -0{,}8, 0{,}1)^T$. Khi $\delta = 1$ thì X là quan sát được và $\delta = 0$ thì X là không quan sát được hoặc thiếu. Biến δ có trung bình là 0.558, nghĩa là tỷ lệ thiếu dữ liệu của X là 44,2%. Hai biến W và X có sự tương quan với nhau với hệ số tương quan Spearman là $-0{,}15$ và p-giá trị là 0,0114 (khi $\delta = 1$), điều này cho phép W được sử dụng như là một biến thay thế cho X trong mô hình hồi quy. Khi phân tích mô hình hồi quy logistic của biến δ theo các biến Y, Z_1, Z_2, Z_3, và W cho ta p-giá trị tương ứng theo từng biến là 0,001, $< 0{,}001$, 0,004, 0,149, $< 0{,}001$, và 0,014, chứng tỏ rằng cơ chế thiếu dữ liệu của X thỏa mãn điều kiện MAR. Do đó, ta sẽ sử dụng bộ dữ liệu với X là MAR để xây dựng mô hình logistic sau:

$$P(Y = 1|X, Z_1, Z_2, Z_3) = \beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2 + \beta_4 Z_3. \quad (12)$$

Kết quả phân tích mô hình (12) được cho trong Bảng 5, bao gồm các giá trị ước lượng, ASE và p-giá trị cho các tham số β_j ($j = 0{,}1, \ldots, 4$) lần lượt tương ứng là hệ số chặn và tham số của các biến X, Z_1, Z_2 và Z_3, thu được từ: $\widehat{\boldsymbol{\beta}}_F$ là kết quả của phương pháp ước ước lượng cực đại (ML) khi đầy đủ dữ liệu, $\widehat{\boldsymbol{\beta}}_C$, $\widehat{\boldsymbol{\beta}}_W$, $\widehat{\boldsymbol{\beta}}_V$, $\widehat{\boldsymbol{\beta}}_R$, và $\widehat{\boldsymbol{\beta}}_M$ lần lượt là kết quả của các phương pháp SIPW, VL, RFMI, và MI, khi tỷ lệ thiếu dữ liệu của X là 44,2%. Qua Bảng 5, ta nhận thấy rằng các tham số β_j ($j = 0{,}1, \ldots, 4$) đều khác 0 có ý nghĩa thống kê trong mô hình với mức ý nghĩa 5% cho tất cả các phương pháp ước lượng, ngoại trừ β_1 từ phương pháp RFMI. Kết quả ước tính $\beta_2 > 0$ của tất cả các phương pháp cho thấy tỷ lệ phụ nữ có tiền sử gãy xương trước đó sẽ có nhiều khả năng bị gãy xương trong năm đầu tiên của nghiên cứu hơn những người không có tiền sử này. Kết quả ước tính âm của β_1, β_3 và β_4 lần lượt cho thấy rằng những phụ nữ có nguy cơ gãy xương tự báo nhỏ hơn những người khác cùng tuổi và những phụ nữ có tuổi từ 70 trở xuống ít có khả năng bị gãy xương trong năm đầu của nghiên cứu. Ngoài ra, nếu lấy kết quả từ $\widehat{\boldsymbol{\beta}}_F$ làm mốc thì $\widehat{\boldsymbol{\beta}}_C$ có kết quả ước lượng sai lệch nhiều nhất và ASE là cao nhất so với các phương pháp còn lại. Kết quả ước lượng của β_1, tức là của biến thiếu dữ liệu X, của 3 phương pháp CC, SIPW và VL là khá cao so với ML và MI. Giá trị ước lượng của MI tương đối gần với ML và giá trị ASE của MI cũng tương đối thấp hơn so với CC, SIPW và VL, điều đó chứng tỏ hiệu quả tính toán của phương pháp đề xuất khá tốt và khớp với các kết quả trong phần nghiên cứu mô phỏng.

5. KẾT LUẬN

Nghiên cứu đã đề xuất các bước thuật toán và các tính chất cỡ mẫu lớn của phương pháp MI cho mô hình hồi quy logistic với hiệp biến thiếu dữ liệu ngẫu nhiên. Thông qua các tình huống tính toán mô phỏng, thuật toán đề xuất cho thấy hiệu quả tốt hơn so với các phương pháp ước lượng khác. Một bộ dữ liệu thực tế về y học cũng được sử dụng để minh họa tính ứng dụng của phương pháp đề xuất. Vì hồi quy logistic là một dạng của mô hình phân loại, nên nghiên cứu này có nhiều tiềm năng để áp dụng vào phân loại các dữ liệu thiếu trong các lĩnh vực khác nhau.

LỜI CẢM TẠ

Đề tài này được tài trợ bởi Trường Đại học Cần Thơ, Mã số: T2023-18.

TÀI LIỆU THAM KHẢO

Breslow, N. E., & Cain, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika*, *75*(1), 11-20. https://doi.org/10.1093/biomet/75.1.11

Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1-67.

Fay, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, *91*(434), 490-498. https://doi.org/10.1080/01621459.1996.10476909

Hosmer, D. W., Lemeshow S., & Sturdivant R. X. (2013). *Applied logistic regression*. John Wiley & Sons. https://doi.org/10.1002/9781118548387

Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 66-685.

Hsieh, S. H., Li, C. S., & Lee, S. M. (2013). Logistic regression with outcome and covariates missing separately or simultaneously. *Computational Statistics and Data Analysis*, *66*, 32-54. https://doi.org/10.1016/j.csda.2013.03.007

Jiang, W., Josse, J., Lavielle, M., & Group, T. (2020). Logistic regression with missing covariates - Parameter estimation, model selection and prediction within a joint modeling framework. *Computational Statistics and Data Analysis*, *145*, 106907. https://doi.org/10.1016/j.csda.2019.106907

Lee, S. M., Li, C. S., Hsieh, S. H., & Huang, L. H. (2012). Semiparametric estimation of logistic regression model with missing covariates and outcome. *Metrika*, *75*, 621-653. https://doi.org/10.1007/s00184-011-0345-9

Lee, S. M., Lukusa, T. M., & Li, C. S. (2020). Estimation of a zero-inflated Poisson regression model with missing covariates via nonparametric multiple imputation methods. *Computational Statistics*, *35*, 725-754. https://doi.org/10.1007/s00180-019-00930-x

Lee, S. M., Tran, P. L., & Li, C. S. (2022). Goodness-of-fit tests for a logistic regression model with missing covariates. *Statistical Methods in Medical Research*, *31*(6), 1031-1050. https://doi.org/10.1177/09622802221079350

Lee, S. M., Le, T. N., Tran, P. L., & Li, C. S. (2023). Estimation of logistic regression with covariates missing separately or simultaneously via multiple imputation methods. *Computational Statistics*, *38*, 899-934. https://doi.org/10.1007/s00180-022-01250-3

Lukusa, T. M., Lee, S. M., & Li, C. S. (2016). Semiparametric estimation of a zero-inflated Poisson regression model with missing covariates. *Metrika*, *79*(4), 457-483. https://doi.org/10.1007/s00184-015-0563-7

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Statistical analysis with missing data*. John Wiley & Sons.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473-489. https://doi.org/10.1080/01621459.1996.10476908

Tran, P. L., Le, T. N., Lee, S. M., & Li, C. S. (2023). Estimation of parameters of logistic regression with covariates missing separately or simultaneously. *Communications in Statistics - Theory and Methods*, *52*(6), 1981-2009. https://doi.org/10.1080/03610926.2021.1943443

Wang, S., & Wang, C. Y. (2001). A note on kernel assisted estimators in missing covariate regression. *Statistics and Probability Letters*, *55*(4), 439-449. https://doi.org/10.1016/S0167-7152(01)00167-5

Wang, D., & Chen, S. X. (2009). Empirical likelihood for estimating equations with missing values. *The Annals of Statistics*, *37*(1), 490-517. https://doi.org/10.1214/07-AOS585

Wang, C. Y., Wang, S., Zhao, L. P., & Ou, S. T. (1997). Weighted semiparametric estimation in regression analysis with missing covariate data. *Journal of the American Statistical Association*, *92*(438), 512-525. https://doi.org/10.1080/01621459.1997.10474004

Wang, C. Y., Chen, J. C., Lee, S. M., & Ou, S. T. (2002). Joint conditional likelihood estimator in logistic regression with missing covariate data. *Statistica Sinica*, *12*(2), 555-574.

Zhao, L. P., & Lipsitz, S. (1992). Designs and analysis of two-stage studies. *Statistics in Medicine*, *11*(6), 769-782. https://doi.org/10.1002/sim.4780110608