

Cách tiếp cận mới cho phương pháp Bayes để giảm sai số trong ước lượng tham số cho mô hình hồi quy logistic

Phạm Thị Thu Hoa*, Phạm Thị Thu Hường



Use your smartphone to scan this QR code and download this article

TÓM TẮT

Khi trong dữ liệu của mô hình hồi quy logistic xuất hiện nhiều quan sát có xác suất ước lượng rất gần 0 hoặc 1, phương pháp cực đại hàm hợp lý (Maximum Likelihood Estimation - MLE) thường gặp khó khăn trong việc ước lượng tham số và có thể tạo ra các ước lượng kém ổn định với sai số lớn. Trong bối cảnh đó, phương pháp Bayes được xem là một giải pháp hiệu quả nhờ khả năng kết hợp thông tin tiên nghiệm với dữ liệu quan sát. Tuy nhiên, hiệu quả của phương pháp Bayes phụ thuộc đáng kể vào việc lựa chọn phân phối tiên nghiệm phù hợp cho các tham số của mô hình. Trong bài báo này, chúng tôi đề xuất một phương pháp Bayes mới để làm giảm sai số trong ước lượng tham số của mô hình. Trước tiên, thuật toán loại bỏ và chấp nhận tham số được sử dụng dựa trên mức độ tương đồng giữa số liệu mô phỏng và số liệu quan sát, từ đó xây dựng thông tin tiên nghiệm cho tham số. Tiếp theo, phân phối hậu nghiệm được thiết lập dựa trên hàm hợp lý và thông tin tiên nghiệm vừa thu được. Thuật toán Metropolis-Hastings (MH) được áp dụng để tạo ra chuỗi Markov dừng và ước lượng tham số cho mô hình. Hiệu quả của phương pháp đề xuất được đánh giá thông qua các nghiên cứu mô phỏng và một bộ dữ liệu thực tế. Kết quả cho thấy phương pháp Bayes được đề xuất giúp cải thiện độ chính xác của các ước lượng tham số, giảm sai số ước lượng so với phương pháp MLE cũng như phương pháp Bayes sử dụng tiên nghiệm không chứa thông tin.

Từ khóa: Thống kê Bayes, thống kê tần số, mô hình hồi quy logistic, phương pháp loại bỏ, mô phỏng số liệu

GIỚI THIỆU

Hàm hợp lý của mô hình hồi quy logistic được thiết lập bằng tích của các xác suất tương ứng với giá trị thực tế của biến phụ thuộc tại từng giá trị quan sát¹⁻⁵. Trong trường hợp có nhiều quan sát mà xác suất dự đoán của biến phụ thuộc có giá trị gần về 0 hoặc 1, việc ước lượng tham số bằng phương pháp cực đại hàm hợp lý sẽ trở nên không ổn định^{1, 4, 6, 7}. Khi đó, giá trị của hàm hợp lý có thể trở nên cực kỳ nhỏ, tiến về 0, dẫn đến log hàm hợp lý tiến dần đến âm vô cùng, gây khó khăn cho quá trình ước lượng tham số và làm sai lệch kết quả. Khi xử lý số liệu có dạng này bằng hàm glm trong ngôn ngữ R với phương pháp cực đại hàm hợp lý, thường kết quả suy luận không đưa ra cảnh báo về độ sai lệch trong kết quả ước lượng do hiện tượng tách biệt trong số liệu chưa xảy ra. Với cách tiếp cận của thống kê tần số, khi hiện tượng tách biệt và gần tách xảy ra trong số liệu quan sát, Firth (1993)⁸ đã đề xuất phương pháp điều chỉnh hàm hợp lý để giúp cho xác định được hữu hạn và duy nhất giá trị cực đại của hàm hợp lý. Bên cạnh đó, với việc thêm ma trận thông tin vào hàm hợp lý sẽ làm giảm sai số cho ước lượng bằng phương pháp cực đại hàm hợp lý⁹.

Thống kê Bayes với phân phối tiên nghiệm có dạng phân phối xác định có thể giúp làm giảm sai số trong ước lượng tham số của mô hình hồi quy logistic, đặc biệt trong các trường hợp dữ liệu có nhiều quan sát mà xác suất thành công

Bộ môn Toán, Khoa Sư phạm, Trường Đại học An Giang, Đại học Quốc gia Thành phố Hồ Chí Minh

Liên hệ

Phạm Thị Thu Hoa, Bộ môn Toán, Khoa Sư phạm, Trường Đại học An Giang, Đại học Quốc gia Thành phố Hồ Chí Minh

Email: ptt Hoa@agu.edu.vn

Lịch sử

- Ngày nhận: 06-09-2025
- Ngày sửa đổi: 03-06-2026
- Ngày chấp nhận: 14-08-2026
- Ngày đăng: 17-09-2026

DOI: <https://doi.org/10.32508/vnuhcmj-arns.v10i3.1474>



Bản quyền

© Tạp chí ĐHQG Tp.HCM. Đây là bài báo công bố mở được phát hành theo các điều khoản của the Creative Commons Attribution 4.0 International license.

Trích dẫn bài báo này: Thu Hoa P T, Thu Hường P T. Cách tiếp cận mới cho phương pháp Bayes để giảm sai số trong ước lượng tham số cho mô hình hồi quy logistic. VNUHCM J. Adv. Res. Nat. Sci. 2026; 10(3): 3745-3766.

có giá trị gần 0 hoặc 1¹⁰⁻¹³. Khi không có thông tin về phân phối tiên nghiệm, Gelman (2008) đã đề xuất phương pháp thống kê Bayes với thông tin tiên nghiệm có phân phối Cauchy với tham số cho trước. Tuy nhiên với thông tin tiên nghiệm có độ trải giá trị rất rộng này, thì sai số trong ước lượng tham số có cải thiện nhưng không được cải thiện nhiều. Do đó, việc xác định thông tin cho phân phối tiên nghiệm trong phương pháp thống kê Bayes là yếu tố quan trọng để giúp cải thiện sai số trong ước lượng tham số của mô hình hồi quy logistic^{3, 10, 11, 14, 15}. Hiện nay, chưa có phương pháp Bayes hiệu quả hơn để ước lượng tham số cho mô hình trong trường hợp số liệu có dạng này và khi không có thông tin tiên nghiệm về tham số.

Trong bài báo này, một cách tiếp cận mới của phương pháp thống kê Bayes được đề xuất để cải thiện sai số trong ước lượng tham số của mô hình hồi quy logistic trong trường hợp khi có nhiều quan sát có xuất hiện xác suất thành công có giá trị gần 0 hoặc 1. Phương pháp này được thực hiện qua hai giai đoạn. Giai đoạn đầu, thông tin cho phân phối tiên nghiệm được rút ra dựa vào số liệu quan sát thông qua thuật toán loại bỏ và chấp nhận^{13, 16, 17}. Giai đoạn thứ hai chúng tôi tiến hành dùng thông tin phân phối tiên nghiệm có được để xác định phân phối hậu nghiệm và chạy chuỗi Markov để tìm ước lượng cho tham số.

Để có được thông tin tiên nghiệm cho tham số từ số liệu quan sát, chúng tôi bắt đầu bằng việc đề xuất ngẫu nhiên giá trị cho tham số. Từ tham số được đề xuất, số liệu từ mô hình hồi quy logistic được mô phỏng, sau đó, đem so sánh mức độ tương đồng giữa số liệu mô phỏng và số liệu quan sát để quyết định có chọn hay không giá trị của tham số được đề xuất. Với các giá trị tham số nhận được sẽ hình thành phân phối tiên nghiệm cho tham số. Tiếp theo, phân phối hậu nghiệm cho mô hình hồi quy logistic sẽ được thành lập từ hàm hợp lý và phân phối tiên nghiệm. Chúng tôi sử dụng thuật toán Metropolis-Hastings (MH) để tạo ra chuỗi Markov dừng và hội tụ để ước lượng cho tham số của mô hình.

Kết quả mô phỏng cho thấy, khi trong số liệu không xuất hiện nhiều xác suất dự đoán của biến phụ thuộc gần 0 hoặc 1, cả ba phương pháp cực đại hàm hợp lý, phương pháp Bayes khi không có thông tin tiên nghiệm và phương pháp Bayes được đề xuất có kết quả ước lượng chính xác và tương đương nhau. Tuy nhiên, khi trong số liệu xuất hiện nhiều xác suất dự đoán của biến phụ thuộc gần 0 hoặc 1, phương pháp Bayes được đề xuất có kết quả cải thiện hơn rất nhiều so với phương pháp cực đại hàm hợp lý và phương pháp Bayes khi không có thông tin tiên nghiệm. Phương pháp được đề xuất này cũng thể hiện kết quả ước lượng tham số tốt hơn khi được áp dụng cho số liệu thực tế của Racine et al. (1986)¹⁸.

Bài báo được trình bày bao gồm các nội dung sau: Phần 2 là kết quả lý thuyết. Ở phần này mô hình hồi quy logistic, phương pháp thống kê Bayes cải tiến để ước lượng tham số cho mô hình được trình bày. Nghiên cứu mô phỏng được trình bày ở phần 3. Ở phần này, nghiên cứu mô phỏng được tiến hành để kiểm tra độ chính xác cho ước lượng tham số từ ba phương pháp, phương pháp cực đại hàm hợp lý, phương pháp Bayes khi không có thông tin tiên nghiệm và

phương pháp Bayes được đề xuất. Bên cạnh đó, cả ba phương pháp được áp dụng cho số liệu Racine et al. (1986)¹⁸ để thấy được ưu điểm cải thiện sai số ước lượng của phương pháp Bayes được đề xuất. Phần 5 là phần kết luận của bài báo.

PHƯƠNG PHÁP

Mô hình hồi quy logistic

Trong nghiên cứu này, mô hình hồi quy logistic với hàm liên kết logit được sử dụng để mô tả mối quan hệ giữa biến độc lập và biến phụ thuộc^{1, 4, 5}. Giả sử số liệu gồm n quan sát, mô hình hồi quy logistic tổng quát bao gồm các biến ngẫu nhiên độc lập và phụ thuộc như sau: biến độc lập $\mathbf{x}$ là một vector có p chiều

$$\mathbf{x} = [\mathbf{x}_1 \quad \mathbf{x}_2 \quad \dots \quad \mathbf{x}_n]^T = \begin{bmatrix} 1 & x_{12} & \dots & x_{1p} \\ 1 & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n2} & \dots & x_{np} \end{bmatrix}.$$

Ở đây, cột đầu tiên của ma trận $\mathbf{x}$ là hệ số hằng, thường được gán giá trị bằng 1. Biến phụ thuộc $\mathbf{y} = (y_1, \dots, y_n)^T$ tuân theo phân phối Bernoulli với giá trị nhận được là 0 hoặc 1. Xác suất để biến ngẫu nhiên y_i nhận giá trị bằng 1 với điều kiện $\mathbf{x}_i = [1 \quad x_{i1} \quad x_{i2} \quad \dots \quad x_{ip}]^T$ được đặt là p_i . Khi đó, ta có:

$$\pi(\mathbf{x}_i) = p_i = \Pr(y_i = 1 | \mathbf{x}_i).$$

Mô hình hồi quy logistic với hàm liên kết logit để mô tả mối liên hệ giữa biến độc lập và biến phụ thuộc có dạng sau:

$$\ln \left(\frac{p_i}{1 - p_i} \right) = \mathbf{x}_i^T \boldsymbol{\beta}, i = 1, 2, \dots, n. \quad (2.1)$$

Trong đó, vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ là tham số hồi quy trong mô hình hồi quy logistic. Khi đó,

$$\pi(\mathbf{x}_i) = p_i = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}. \quad (2.2)$$

Hàm hợp lí cho mô hình hồi quy logistic có dạng sau:

$$\begin{aligned} l(\boldsymbol{\beta}) &= p(\mathbf{y} | \boldsymbol{\beta}) = \prod_{i=1}^n [\pi(\mathbf{x}_i)]^{y_i} [1 - \pi(\mathbf{x}_i)]^{1-y_i} \\ &= \prod_{i=1}^n \left[\frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \right]^{y_i} \left[\frac{1}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})} \right]^{1-y_i}. \end{aligned} \quad (2.3)$$

Lấy logarit cơ số e cho hai phía của phương trình (2.3), ta được:

$$\ln l(\boldsymbol{\beta}) = \sum_{i=1}^n y_i \ln [\pi(\mathbf{x}_i)] + \sum_{i=1}^n (1 - y_i) \ln [1 - \pi(\mathbf{x}_i)]. \quad (2.4)$$

Tại các giá trị $\mathbf{x}_i$ mà ở đó xác suất $\pi(\mathbf{x}_i)$ gần về 0 hoặc 1 được thêm vào hàm hợp lý cực đại sẽ làm cho giá trị của hàm này tiến dần về 0 hoặc giá trị hàm $\ln l(\boldsymbol{\beta})$ sẽ tiến về âm vô cùng. Do đó, nếu dùng phương pháp ước lượng cực đại cho hàm $\ln l(\boldsymbol{\beta})$ thì các tham số ước lượng tìm được có thể xuất hiện sai số đáng kể. Sự sai số trong các ước lượng này trong nhiều trường hợp thường không có sự cảnh báo nào theo cách tiếp cận của phương pháp thống kê cổ điển, đặc biệt khi trong số liệu quan sát sự tách biệt số liệu chưa xảy ra rõ ràng.

Phương pháp so sánh loại bỏ để lấy thông tin phân phối tiên nghiệm

Phương pháp so sánh được đề xuất trong phần này để trích thông tin tiên nghiệm cho tham số từ số liệu quan sát và không sử dụng thông tin từ hàm hợp lý^{13, 16, 17}. Chúng tôi bắt đầu bằng việc đề xuất ngẫu nhiên giá trị cho tham số từ một phân phối tổng quát. Từ tham số được đề xuất và mô hình hồi quy logistic trình bày tại (2.1), số liệu cho mô hình hồi quy được mô phỏng. Sau đó, mức độ trùng khớp giữa số liệu mô phỏng và số liệu quan sát được so sánh để quyết định có chọn hay không giá trị của tham số được đề xuất.

Các giá trị của tham số $\boldsymbol{\beta}$ được đề xuất ngẫu nhiên từ hàm tổng quát $g(\cdot)$ bao phủ không gian tham số và số liệu $\mathbf{y}^*$ được mô phỏng từ Phương trình (2.1). Khi đó, phân phối liên hợp giữa tham số $\boldsymbol{\beta}$ và số liệu mô phỏng được xác định như sau:

$$g(\boldsymbol{\beta}, \mathbf{y}^*) = g(\boldsymbol{\beta})p(\mathbf{y}^* | \boldsymbol{\beta}), \quad (2.5)$$

khi đó, $p(\mathbf{y}^* | \boldsymbol{\beta})$ là hàm hợp lý với số liệu mô phỏng $\mathbf{y}^*$.

Tiếp theo, phân phối hậu nghiệm kết hợp giữa tham số $\boldsymbol{\beta}$ và số liệu mô phỏng $\mathbf{y}^*$ có điều kiện trên số liệu quan sát $\mathbf{y}$ được xác định như sau:

$$\pi(\boldsymbol{\beta}, \mathbf{y}^* | \mathbf{y}) \propto I_\varepsilon[\rho(\mathbf{y}^*, \mathbf{y})]p(\mathbf{y}^* | \boldsymbol{\beta})\pi(\boldsymbol{\beta}). \quad (2.6)$$

Ở đây, $I_\varepsilon[\rho(\mathbf{y}^*, \mathbf{y})]$ là hàm chỉ thị đo mức độ gần giữa số liệu mô phỏng $\mathbf{y}^*$ và số liệu quan sát $\mathbf{y}$ với ngưỡng giá trị ε . Hàm $\pi(\boldsymbol{\beta})$ là phân phối tiên nghiệm cho tham số $\boldsymbol{\beta}$.

Để đảm bảo các tham số đề xuất ngẫu nhiên tuân theo phân phối trong phương trình (2.6), ta xác định giá trị trọng số như sau:

$$\omega_i = \frac{I_\varepsilon[\rho(\mathbf{y}^*, \mathbf{y})]p(\mathbf{y}^* | \boldsymbol{\beta})\pi(\boldsymbol{\beta})}{g(\boldsymbol{\beta})p(\mathbf{y}^* | \boldsymbol{\beta})}. \quad (2.7)$$

Nếu ta chọn phân phối $g(\boldsymbol{\beta})$ trong công thức (2.7) trùng với phân phối tiên nghiệm $\pi(\boldsymbol{\beta})$, khi đó, giá trị trọng số ω_i không phụ thuộc vào hàm hợp lý và có công thức:

$$\begin{aligned}\omega_i &= \frac{I_\varepsilon[\rho(\mathbf{y}^*, \mathbf{y})]p(\mathbf{y}^* | \boldsymbol{\beta})\pi(\boldsymbol{\beta})}{\pi(\boldsymbol{\beta})p(\mathbf{y}^* | \boldsymbol{\beta})} \\ &= I_\varepsilon[\rho(\mathbf{y}^*, \mathbf{y})].\end{aligned}\quad (2.8)$$

Trọng số ω_i sẽ nhận 2 giá trị là 0 và 1. Trọng số ω_i nhận giá trị 0 khi giá trị tham số $\boldsymbol{\beta}$ không được chấp nhận và ω_i nhận giá trị 1 khi giá trị tham số $\boldsymbol{\beta}$ được chấp nhận.

Số liệu của biến phụ thuộc trong mô hình hồi quy logistic nhận giá trị 0 hoặc 1, do đó việc sử dụng khoảng cách dựa trên trung bình và phương sai để so sánh giữa số liệu mô phỏng $\mathbf{y}^*$ và số liệu quan sát $\mathbf{y}$ là không hiệu quả. Bên cạnh đó, khi trong số liệu có nhiều quan sát mà xác suất dự đoán của biến phụ thuộc có giá trị gần về 0 hoặc 1, giá trị của hàm log-likelihood bị ảnh hưởng và tiến về âm vô cực. Vì vậy, trong trường hợp này khoảng cách log-likelihood cũng sẽ không phù hợp. Do đó, trong nghiên cứu này, chúng tôi đo lường mức độ tương đồng giữa số liệu mô phỏng và số liệu quan sát bằng số phần trăm trùng lặp giữa số liệu mô phỏng $\mathbf{y}^*$ và số liệu quan sát $\mathbf{y}$ bằng hàm $I_\varepsilon[\rho(\mathbf{y}^*, \mathbf{y})]$ được xác định như bên dưới.

Với giá trị đề xuất ngẫu nhiên tham số $\boldsymbol{\beta}$, ta tính được xác suất thành công tại từng giá trị quan sát theo công thức (2.2) và số liệu nhị thức $\mathbf{y}^*$ được mô phỏng từ Phương trình (2.1). Tỷ lệ trùng khớp giữa dữ liệu mô phỏng và dữ liệu quan sát được tính như sau:

$$\rho(\mathbf{y}^*, \mathbf{y}) = \frac{\sum_{i=1}^n I(y_i = y_i^*)}{n}.$$

Hàm chỉ thị $I_\varepsilon[\rho(\mathbf{y}^*, \mathbf{y})]$ với ngưỡng giá trị ε được xác định như sau:

$$I_\varepsilon[\rho(\mathbf{y}^*, \mathbf{y})] = \begin{cases} 1, & \rho(\mathbf{y}^*, \mathbf{y}) > \varepsilon \\ 0, & \rho(\mathbf{y}^*, \mathbf{y}) \leq \varepsilon \end{cases}.$$

Ta thấy được, khi $\varepsilon \rightarrow 1$ thì số liệu mô phỏng $\mathbf{y}^*$ và số liệu quan sát $\mathbf{y}$ gần như trùng nhau. Theo công thức (2.6) và (2.8), phân phối của các tham số đề xuất được chấp nhận sắp xỉ theo phân phối hậu nghiệm của tham số thực trong mô hình.

Để xác định ngưỡng loại bỏ, tiến trình so sánh và loại bỏ các giá trị của tham số $\boldsymbol{\beta}$ được tiến hành theo các bước trong Thuật toán 1. Phương pháp để xác định ngưỡng loại bỏ là khi tỷ lệ giá trị chấp nhận của các tham số đề xuất là $\alpha\%$ thì thuật toán dừng lại. Quá trình này được thực hiện như sau, một dãy ngưỡng giá trị được cho từ 0% đến 100% tại Bước 1. Với ngưỡng giá trị thứ i , từ tham số được đề xuất, tỷ lệ trùng khớp giữa số liệu mô phỏng và số liệu quan sát được sử dụng để quyết định việc chấp nhận tham số thông qua hàm chỉ thị, được tính toán từ Bước 3 đến Bước 5. Tiếp theo, tỷ lệ các tham số được chấp

nhận trong N vòng lặp được tính toán ở Bước 7. Nếu tỉ lệ chấp nhận của tham số nhỏ hơn hoặc bằng $\alpha\%$ thì ngưỡng giá trị này sẽ được chọn làm ngưỡng loại bỏ và thuật toán dừng lại. Trong bài báo này, chúng tôi chọn mức 5% của 100.000 vòng lặp để xác định ngưỡng loại bỏ.

Thuật toán 1:

1: Cho một dãy các ngưỡng giá trị $0 < \varepsilon_1 < \varepsilon_2 < \dots < \varepsilon_m < 1$,

Cho số lần thực hiện phép lặp N,

Cho giá trị ngưỡng chấp nhận $\alpha\%$.

2: Cho $i = 1$ đến m

Cho $j = 1$ đến N

3: Đề xuất ngẫu nhiên tham số β_j từ hàm tổng quát.

4: Mô phỏng số liệu cho mô hình hồi quy logistic từ tham số β_j .

5: Tính tỉ lệ trùng khớp giữa dữ liệu mô phỏng $\mathbf{y}^*$ và dữ liệu quan sát $\mathbf{y}$.

6: Tính hàm chỉ thị $I_{\varepsilon_i}[\rho(\mathbf{y}^*, \mathbf{y})]$.

7: Tính tỉ lệ chấp nhận các giá trị tham số r_β sau khi chạy N vòng lặp

8: Nếu $r_\beta < \alpha\%$:

9: Dừng vòng lặp cho j.

Dừng vòng lặp cho i.

Các tham số β được chấp nhận đảm bảo mức độ trùng với số liệu quan sát và có phân phối sắp sỡ với phân phối hậu nghiệm của tham số thực có trong mô hình. Do đó, phân phối tiên nghiệm được xây dựng có phân phối chuẩn nhiều chiều với giá trị trung bình và độ lệch chuẩn được tính dựa trên các giá trị tham số nhận được. Phân phối tiên nghiệm này phản ánh thông tin ban đầu về tham số trước khi kết hợp với dữ liệu trong bước suy luận Bayes. Ngoài ra, việc xây dựng phân phối tiên nghiệm dựa trên dữ liệu quan sát giúp tăng tính khách quan và phù hợp của mô hình so với việc sử dụng phân phối tiên nghiệm không có thông tin.

Gelman (2008)¹¹ và Hoa và Huong (2021)¹⁹ đã chứng minh rằng khi sử dụng phân bố tiên nghiệm có dạng tiêu chuẩn (tích phân của hàm mật độ trên toàn miền giá trị của tham số phải bằng 1), thì trung bình hậu nghiệm của tham số tồn tại trong trường hợp số liệu quan sát có xảy ra sự tách biệt và hơn nữa, kết quả này cũng đúng khi trong số liệu quan sát có xác suất của mô hình gần 0 hoặc 1. Do đó, với cách chọn phân phối tiên nghiệm tuân theo phân phối chuẩn có thông tin như trên sẽ đảm bảo sự hội tụ của phân phối hậu nghiệm.

Phương pháp thống kê Bayes để ước lượng tham số cho mô hình hồi quy logistic

Để áp dụng phương pháp Bayes, phân phối hậu nghiệm cho mô hình hồi quy logistic được thành lập như sau:

$$p(\beta | y) = \frac{p(y | \beta) p^*(\beta)}{p(y)}. \quad (2.9)$$

Ở đây, $p(y | \beta)$ là hàm hợp lý theo công thức (2.3) và $p^*(\beta)$ là phân phối tiên nghiệm có được sau khi tiến hành chạy Thuật toán 1. Do đó, phân phối hậu nghiệm có dạng sau:

$$\begin{aligned} p(\beta | y) &= \frac{p(y | \beta) p^*(\beta)}{p(y)} \\ &\propto \prod_{i=1}^n [\pi(\mathbf{x}_i)]^{y_i} [1 - \pi(\mathbf{x}_i)]^{1-y_i} p^*(\beta) \\ &\propto \prod_{i=1}^n \left[\frac{\exp(\mathbf{x}_i^T \beta)}{1 + \exp(\mathbf{x}_i^T \beta)} \right]^{y_i} \left[\frac{1}{1 + \exp(\mathbf{x}_i^T \beta)} \right]^{1-y_i} p^*(\beta). \end{aligned} \quad (2.10)$$

Do hàm phân phối hậu nghiệm không có dạng tiêu chuẩn, thuật toán Metropolis Hastings (MH) được sử dụng để tạo ra một chuỗi Markov dừng và hội tụ về phân phối hậu nghiệm. Để thực hiện thuật toán này tỉ số chấp nhận của tham số β được tính như sau:

$$r(\beta) = \frac{p(\beta^* | y) q(\beta^{(t-1)} | \beta^*)}{p(\beta^{(t-1)} | y) q(\beta^* | \beta^{(t-1)})}. \quad (2.11)$$

Ở đây, $\beta^{(t-1)}$ là giá trị tại vòng lặp t-1 của tham số β và β^* là giá trị đề xuất của tham số β . $p(\beta^* | y)$ là hàm phân phối hậu nghiệm theo công thức (2.10), $q(\beta^* | \beta^{(t-1)})$ là hàm mật độ xác suất cho tham số β^* . Chúng tôi sử dụng bước nhảy ngẫu nhiên MH từ phân phối chuẩn nhiều chiều với trung bình đặt tại giá trị $\beta^{(t-1)}$. Tham số điều chỉnh độ rộng của bước nhảy được lựa chọn nhằm đạt được tỷ lệ chấp nhận mong muốn, thường nằm trong khoảng từ 20% đến 50%. Do đó, ta có $q(\beta^* | \beta^{(t-1)}) = q(\beta^{(t-1)} | \beta^*)$. Tỉ số chấp nhận của tham số β^* của tham khi đó có công thức:

$$r(\beta) = \frac{\prod_{i=1}^n \left[\frac{\exp(\mathbf{x}_i^T \beta^*)}{1 + \exp(\mathbf{x}_i^T \beta^*)} \right]^{y_i} \left[\frac{1}{1 + \exp(\mathbf{x}_i^T \beta^*)} \right]^{1-y_i} p^*(\beta^*)}{\prod_{i=1}^n \left[\frac{\exp(\mathbf{x}_i^T \beta^{(t-1)})}{1 + \exp(\mathbf{x}_i^T \beta^{(t-1)})} \right]^{y_i} \left[\frac{1}{1 + \exp(\mathbf{x}_i^T \beta^{(t-1)})} \right]^{1-y_i} p^*(\beta^{(t-1)})}. \quad (2.12)$$

Thuật toán để tạo ra một chuỗi Markov hội tụ về hàm phân phối hậu nghiệm (2.10) được trình bày trong Thuật toán 2.

Thuật toán 2: Thuật toán Metropolis-Hasting

1: Cho giá trị hiện tại của tham số là $\beta^{(t-1)}$

2: Đề xuất giá trị mới cho tham số β^*

3: Tính tỉ lệ chấp nhận, $r(\beta)$, theo công thức (2.12)

4: Mô phỏng $u \sim \text{Uniform}(0,1)$

i. Nếu $r(\beta) > u$ thì ta đặt $\beta^{(t)} = \beta^*$,

ii. Nếu ngược lại thì ta đặt $\beta^{(t)} = \beta^{(t-1)}$.

NGHIÊN CỨU MÔ PHỎNG

Trong nghiên cứu này, chúng tôi mô phỏng mẫu ngẫu nhiên tuân theo mô hình hồi quy logistic với 2 biến ngẫu nhiên độc lập. Hàm liên kết logit để mô tả mối liên hệ giữa biến độc lập và biến phụ thuộc cho mô hình hồi quy logistic được xác định theo (2.1) như sau:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2. \quad (3.1)$$

Để mô phỏng mẫu ngẫu nhiên không có nhiều quan sát với giá trị xác suất của biến phụ thuộc xuất hiện gần 0 hoặc 1 (Trường hợp 1), chúng tôi chọn biến độc lập $x_1 \sim N(0,1)$ và $x_2 \sim N(0,1)$. Theo công thức tính giá trị xác suất của biến phụ thuộc ở công thức (2.2), khi tích giữa giá trị tham số và giá trị của biến độc lập có giá trị cực bé hoặc cực lớn sẽ tác động đến giá trị của xác suất mô hình²⁰. Do đó, để tạo ra mẫu ngẫu nhiên có nhiều quan sát với giá trị xác suất của biến phụ thuộc xuất hiện gần 0 hoặc 1 (Trường hợp 2), chúng tôi chọn biến độc lập $x_1 \sim N(0,3)$ và $x_2 \sim N(0,1)$. Với mỗi mẫu ngẫu nhiên, chúng tôi mô phỏng số liệu với cỡ mẫu là 100 quan sát.

Chúng tôi chạy mô hình lặp lại 100 lần cho các giá trị khác nhau của tham số trong Trường hợp 1 và Trường hợp 2. Trong Trường hợp 1, khi tham số $\beta = (-1, 1, 1)$, biến độc lập $x_1 \sim N(0,1)$ và $x_2 \sim N(0,1)$, trung bình giá trị nhỏ nhất là 1.10^{-2} . Trong Trường hợp 2, khi tham số $\beta = (-1, 1, 1)$, biến độc lập $x_1 \sim N(0,3)$ và $x_2 \sim N(0,1)$, trung bình giá trị nhỏ nhất là 3.10^{-5} . Khi tham số $\beta = (-2, 1, 1)$, $\beta = (1, -2, 1)$ và $\beta = (2, -2, 1)$, biến độc lập $x_1 \sim N(0,3)$ và $x_2 \sim N(0,1)$, trung bình giá trị nhỏ nhất vào khoảng 8.10^{-9} . Phương pháp Bayes được đề xuất được áp dụng để làm giảm sai số ước lượng cho các trường hợp 2, đặc biệt cho các số liệu xuất hiện cảnh báo xác suất mô hình nhỏ (nhỏ hơn 1.10^{-15}) trong tính toán của hàm glm.

Tỉ lệ số liệu có xác suất gần 0 và 1 với các giá trị khác nhau của tham số

Với số liệu mô phỏng không xuất hiện nhiều giá trị xác suất của biến phụ thuộc gần 0 hoặc 1, tham số $\beta = (-1, 1, 1)$, biến độc lập $x_1 \sim N(0, 1)$ và $x_2 \sim N(0, 1)$ được đưa vào mô hình 3.1. Tỉ lệ các giá trị xác suất nằm trong khoảng nhỏ hơn 0.05, khoảng đóng từ 0.05 đến 0.95 và khoảng giá trị xác suất lớn hơn 0.95 được trình bày trong Bảng 1.

Bảng 1. Tỉ lệ các giá trị xác suất cho số liệu mô phỏng trong Trường hợp 1

$\beta = (-1, 1, 1)$	$p_i < 0.05$	$0.05 \leq p_i \leq 0.95$	$p_i > 0.95$	Tổng
Tỉ lệ	8.35%	91.39%	0.26%	100%

Từ Bảng 1, chúng ta thấy có 91.39% quan sát có giá trị xác suất thành công lớn hơn bằng 0.05 và nhỏ hơn bằng 0.95. Trong khi đó, trong mẫu mô phỏng chỉ có 8.35% xác suất thành công nhỏ hơn 0.05 và 0.26% quan sát có xác suất thành công lớn hơn 0.95.

Với số liệu mô phỏng xuất hiện nhiều giá trị xác suất của biến phụ thuộc gần 0 hoặc 1, các giá trị tham số β khác nhau, biến độc lập $x_1 \sim N(0, 3)$ và $x_2 \sim N(0, 1)$ được đưa vào mô hình 3.1. Tỉ lệ các giá trị xác suất nằm trong khoảng nhỏ hơn 0.05, khoảng đóng từ 0.05 đến 0.95 và khoảng giá trị xác suất lớn hơn 0.95 được trình bày trong Bảng 2.

Bảng 2. Tỉ lệ các giá trị xác suất cho số liệu mô phỏng trong Trường hợp 2

Tham số	$p_i < 0.05$	$0.05 \leq p_i \leq 0.95$	$p_i > 0.95$	Tổng
$\beta = (-1, 1, 1)$	26.49%	63.07%	10.44%	100%
$\beta = (-2, 1, 1)$	38.73%	55.19%	6.08%	100%
$\beta = (-1, 2, 1)$	37.67%	37.14%	25.19%	100%
$\beta = (-2, 2, 1)$	43.55%	35.15%	21.3%	100%

Trong Bảng 2, với các giá trị tham số khác nhau, tỉ lệ quan sát trong từng khoảng giá trị của xác suất thành công cũng khác nhau. Với giá trị tham số $\beta = (-1, 1, 1)$, có 63.07% quan sát có giá trị xác suất từ 0.05 đến 0.95 và có 36.93 % các quan sát có giá trị xác suất nhỏ hơn 0.05 và lớn hơn 0.95. Khi $\beta = (-2, 1, 1)$, tỉ lệ quan sát có giá trị xác suất từ 0.05 đến 0.95 chỉ còn 55.19%. Khi $\beta = (-1, 2, 1)$ và $\beta = (-2, 2, 1)$, các quan sát có giá trị xác suất nhỏ hơn 0.05 và lớn hơn 0.95 chiếm hơn 60% trong số mẫu mô phỏng.

Kết quả ước lượng tham số

Trong nghiên cứu này chúng tôi so sánh kết quả ước lượng tham số của ba phương pháp khác nhau cho mô hình hồi quy logistic. Các phương pháp đó là phương pháp cực đại hàm hợp lý, phương pháp Bayes khi không có thông tin tiên nghiệm, và phương pháp Bayes được đề xuất trong bài báo để ước lượng tham số cho mô hình hồi quy logistic. Với phương pháp cực đại hàm hợp lý, hàm glm trong phần mềm R được áp dụng. Khi không có thông tin tiên nghiệm, phương pháp Bayes dùng phân phối Cauchy với tham số 2.5 làm thông tin tiên nghiệm và hàm bayesglm trong phần mềm R được áp dụng. Với cách tiếp cận bằng thống kê Bayes được đề xuất, mỗi chuỗi Markov bao gồm 100,000 vòng lặp với 10,000 vòng đầu tiên được loại bỏ để đảm bảo sự hội tụ của chuỗi Markov về hàm hậu nghiệm. Trong cách tiếp cận này, các kỹ thuật tính toán được đòi hỏi cho việc tìm thông tin cho tham số ở Thuật toán 1, tuy nhiên, ở Thuật toán 2, việc chạy chuỗi Markov có kỹ thuật tính toán tương đương với cách tiếp cận Bayes dùng phân phối Cauchy.

Kết quả ước lượng cho tham số trong trường hợp không có nhiều quan sát có giá trị xác suất thành công của biến phụ thuộc gần 0 hoặc 1 gồm trung bình ước lượng, độ lệch chuẩn (SD), khoảng ước lượng tin cậy 95% cho phương pháp MLE (95% Sig. level) và khoảng ước lượng tin cậy 95% cho phương pháp Bayes (95% CrI) được trình bày trong Bảng 3.

Bảng 3. Ước lượng tham số mô hình trong Trường hợp 1

β	Phương pháp MLE			Phương pháp Bayes với phân phối Cauchy		
	Trung bình	SD	95% Sig. level	Trung bình	SD	95% CrI
-1	-1.0293	0.302	[-1.632;-0.504]	-0.9852	0.273	[-1.521;-0.452]
1	1.0963	0.359	[0.529;1.767]	1.0942	0.291	[0.439;1.568]
1	1.1442	0.245	[0.542;1.799]	1.1198	0.286	[0.449;1.465]
Phương pháp Bayes được đề xuất						
β	Trung bình	SD	95% CrI			
-1	-1.0274	0.247	[-1.375;-0.679]			
1	1.0971	0.215	[0.558;1.435]			
1	1.1174	0.237	[0.569;1.465]			

Trong Bảng 3, với giá trị tham số thật của tham số $\beta = (-1, 1, 1)$, các giá trị ước lượng tham số từ cả ba phương pháp đều rất chính xác và không có nhiều khác biệt. Hơn nữa, giá trị độ lệch chuẩn của cả ba phương pháp đều nhỏ hơn 0.5, điều này thể hiện khi trong số liệu không có quá nhiều quan sát có giá trị xác suất gần 0 hoặc 1, cả ba phương pháp này đều có thể áp dụng hiệu quả để ước lượng tham số cho mô hình. Cụ thể, giá trị ước lượng từ phương pháp cực đại hàm hợp lý là $\beta^* = (-1.0293; 1.0963; 1.1442)$, với phương pháp Bayes dùng phân phối Cauchy làm phân phối tiên nghiệm và phương pháp Bayes được đề xuất lần lượt có các giá trị ước lượng là $\beta^* = (-0.9852; 1.0942; 1.1198)$ và $\beta^* = (-1.0274; 1.0971; 1.1174)$.

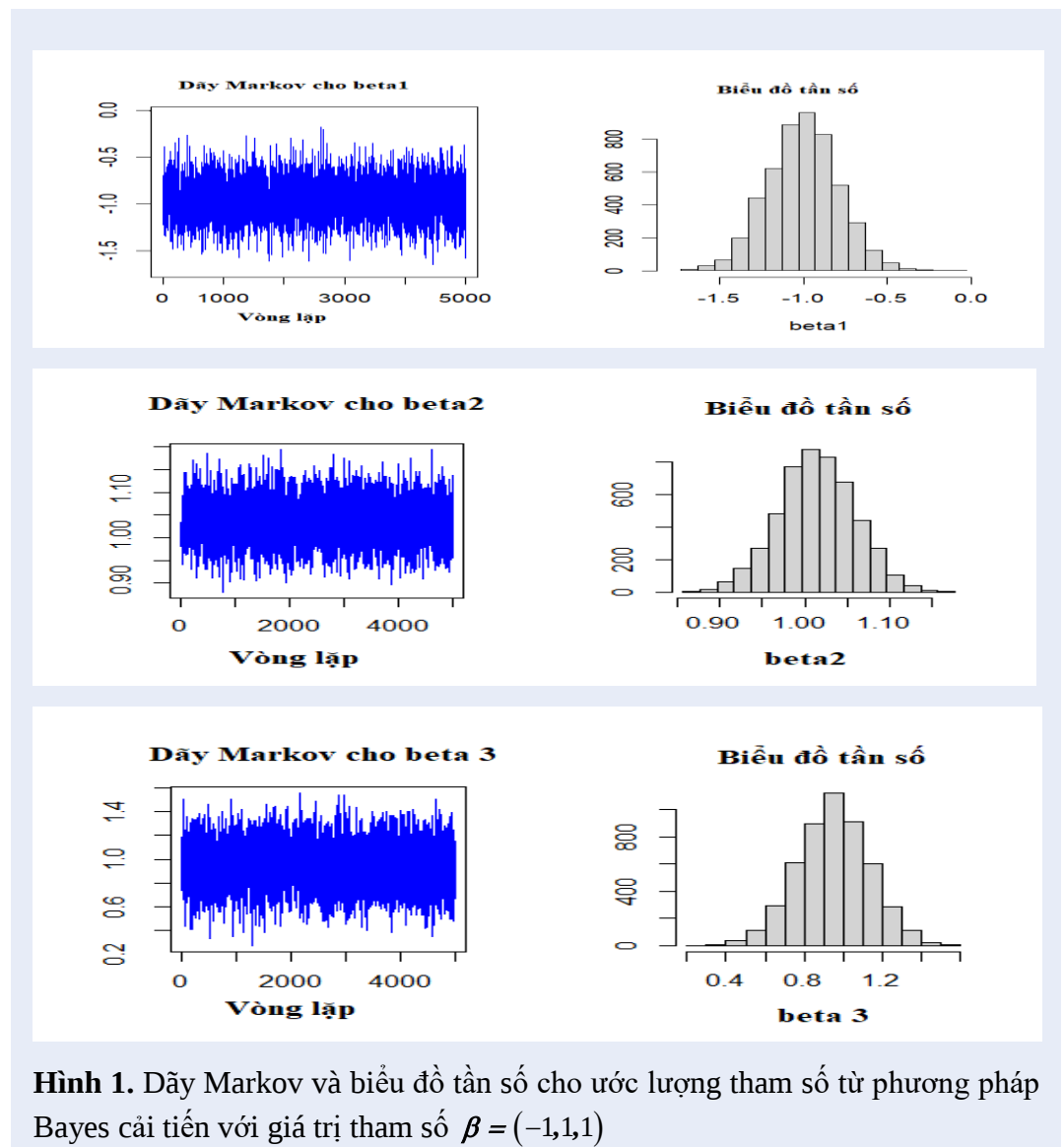
Trong trường hợp 2, khi mẫu mô phỏng xuất hiện nhiều giá trị xác suất thành công gần 0 và 1, cả ba phương pháp trên tiếp tục được áp dụng để ước lượng tham số của mô hình. Kết quả ước lượng được trình bày trong Bảng 4 và Bảng 5.

Bảng 4. Ước lượng tham số mô hình cho Trường hợp 2 khi $\beta = (-1, 1, 1)$

β	Phương pháp MLE			Phương pháp Bayes với phân phối Cauchy		
	Trung bình	SD	95% Sig. level	Trung bình	SD	95% CrI
-1	-1.0567	0.371	[-1.861;-0.384]	-0.9682	0.343	[-1.641;-0.297]
1	1.1032	0.237	[0.705;1.649]	0.9032	0.201	[0.598;1.388]
1	1.1657	0.391	[0.363;1.921]	0.9046	0.340	[0.243;1.575]
Phương pháp Bayes được đề xuất						
β	Trung bình	SD	95% CrI			
-1	-1.0479	0.220	[-1.480;-0.615]			
1	1.0835	0.204	[0.885;1.281]			
1	1.0393	0.304	[0.338;1.439]			

Với giá trị tham số thật của tham số $\beta = (-1, 1, 1)$ được đưa vào mô hình, theo kết quả từ Bảng 2, ta có 36.93 % các quan sát có giá trị xác suất nhỏ hơn 0.05 và lớn hơn 0.95. Trong trường hợp này, theo kết quả từ Bảng 4, các giá trị ước

lượng từ phương pháp MLE là $\beta^* = (-1.0567; 1.1032; 1.1657)$ có sai số lớn hơn so với hai phương pháp Bayes còn lại. Trong đó, ước lượng từ phương pháp Bayes không có thông tin tiên nghiệm là $\beta^* = (-0.9682; 0.9032; 0.9046)$ và từ phương pháp Bayes được đề xuất có thông tin tiên nghiệm là $\beta^* = (-1.0479; 1.0835; 1.0393)$. Ta thấy được kết quả ước lượng tham số có sai số nhỏ từ hai phương pháp Bayes sử dụng phân phối tiên nghiệm không có thông tin và phương pháp Bayes được đề xuất có thông tin tiên nghiệm. Tuy nhiên, độ lệch tiêu chuẩn từ phương pháp Bayes được đề xuất có giá trị nhỏ hơn. Điều này thể hiện, với thông tin tiên nghiệm có được từ số liệu quan sát, phương pháp Bayes được đề xuất tạo ra chuỗi Markov có độ chính xác cao hơn.



Chuỗi Markov với 5,000 bước nhảy cuối và biểu đồ tần số được trình bày trong Hình 1 ([Nguồn: nhóm tác giả]). Từ hình vẽ ta thấy được dãy Markov có giá trị rất ổn định và bao quanh giá trị ước lượng. Điều này thể hiện, chuỗi Markov là dãy dừng và hội tụ về hàm hậu nghiệm. Trong biểu đồ tần số, ta thấy được các giá trị đối xứng với nhau qua giá trị trung bình. Hơn nữa, độ phân tán dữ liệu

quanh giá trị trung bình của phương pháp này nhỏ. Điều này thể hiện độ chính xác, ổn định và hội tụ của chuỗi Markov.

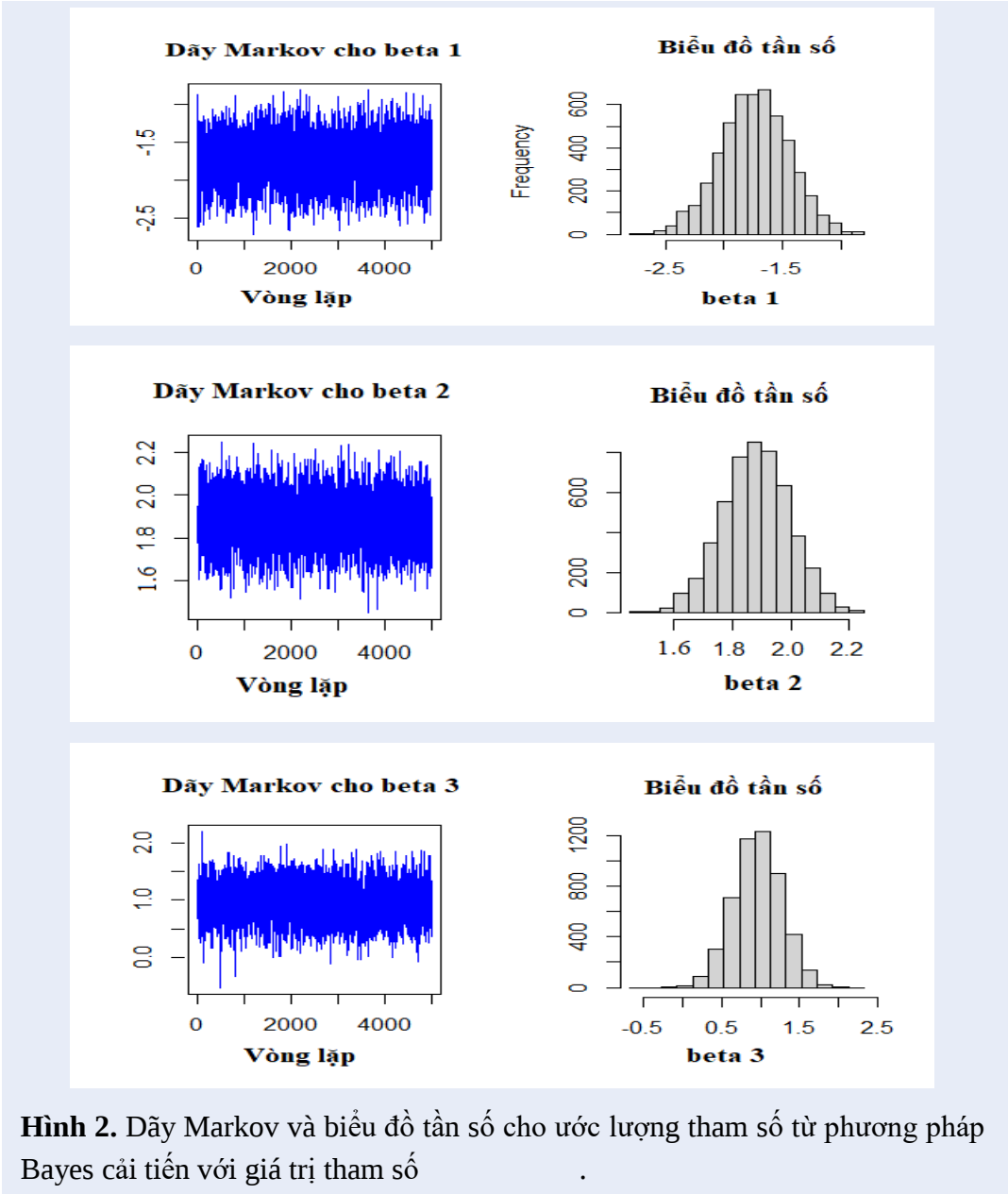
Chúng tôi dùng phương pháp Gelman-Rubin và Geweke để chẩn đoán cho sự hội tụ của chuỗi Markov cho các tham số trong mô hình^{21, 22}. Kết quả chẩn đoán cho mô hình hồi quy logistic Trường hợp 2, khi $\beta = (-1, 1, 1)$, được thể hiện trong phụ lục A.1 và A.2. Đối với các tham số, các phép chẩn đoán MCMC đều cho thấy quá trình hội tụ diễn ra nhanh chóng. Kết quả kiểm định Gelman-Rubin cho thấy ước lượng điểm của hệ số giảm quy mô tiềm năng (PSRF) đối với tất cả các tham số đều nhỏ hơn 1.15, thể hiện chuỗi đã hội tụ về phân phối hậu nghiệm mục tiêu. Bên cạnh đó, chẩn đoán Geweke cũng củng cố kết luận này, khi tất cả các giá trị Z-score chuẩn hóa đều có giá trị tuyệt đối nhỏ hơn 2.

Bảng 5. Ước lượng tham số mô hình cho Trường hợp 2 khi $\beta = (-2, 2, 1)$

β	Phương pháp MLE			Phương pháp Bayes với phân phối Cauchy		
	Trung bình	SD	95% Sig. level	Trung bình	SD	95% CrI
-2	-2.3911	0.778	[-4.302;-1.142]	-1.9674	0.589	[-3.121;-0.816]
2	2.3937	0.654	[1.400;4.058]	1.9410	0.451	[1.079;2.845]
1	1.2658	0.591	[0.253;2.647]	0.9253	0.439	[0.064;1.787]
Phương pháp Bayes được đề xuất						
β	Trung bình	SD	95% CrI			
-2	-1.9634	0.323	[-2.597;-1.331]			
2	1.9363	0.309	[1.481;2.612]			
1	0.9901	0.361	[0.278;1.503]			

Với giá trị tham số thật của tham số $\beta = (-2, 2, 1)$ được đưa vào mô hình, theo kết quả từ Bảng 2, có 64.85 % các quan sát có giá trị xác suất nhỏ hơn 0.05 và lớn hơn 0.95. Trong trường hợp này, theo kết quả từ Bảng 5, các giá trị ước lượng từ phương pháp MLE, $\beta^* = (-2.3911; 2.3937; 1.2658)$, có sai số lớn rõ rệt và lớn hơn nhiều so với hai phương pháp Bayes còn lại. Trong đó, ước lượng từ phương pháp Bayes không có thông tin tiên nghiệm là $\beta^* = (-1.9674; 1.9410; 0.9253)$ và từ phương pháp Bayes được đề xuất là $\beta^* = (-1.9634; 1.9363; 0.9901)$.

Ta thấy được kết quả ước lượng từ phương pháp Bayes sử dụng phân phối tiên nghiệm không có thông tin có sai số lớn hơn kết quả ước lượng từ phương pháp Bayes được đề xuất. Hơn nữa, độ lệch tiêu chuẩn từ phương pháp Bayes được đề xuất có giá trị nhỏ nhất trong cả ba phương pháp ước lượng. Điều này cho thấy, khi số liệu có khuynh hướng xuất hiện nhiều xác suất thành công tiến gần 0 hoặc 1 thì phương pháp Bayes được đề xuất thể hiện rất hiệu quả trong việc cải tiến độ chính xác ước lượng tham số.



Tương tự như trên, chuỗi Markov với 5,000 bước nhảy cuối và biểu đồ tần số từ phương pháp Bayes cải tiến với giá trị tham số $\beta = (-2, 2, 1)$ được trình bày trong Hình 2 ([Nguồn: nhóm tác giả]). Từ hình vẽ ta thấy được dãy Markov có các giá trị rất ổn định và bao quanh giá trị ước lượng trong trường hợp này. Hơn nữa, trong biểu đồ tần số, ta thấy được các giá trị đối xứng với nhau qua giá trị trung bình và độ phân tán dữ liệu cho giá trị chấp nhận của phương pháp

này nhỏ. Điều này thể hiện một lần nữa cho thấy chuỗi Markov là dãy dừng và hội tụ về hàm hậu nghiệm.

Phương pháp Gelman-Rubin và Geweke cũng được sử dụng để chẩn đoán cho sự hội tụ của chuỗi Markov cho các tham số trong mô hình. Kết quả chẩn đoán cho mô hình hồi quy logistic Trường hợp 2, khi $\beta = (-2, 2, 1)$, được thể hiện trong phụ lục B.1 và B.2. Kết quả kiểm định Gelman–Rubin cho thấy ước lượng điểm của hệ số giảm quy mô tiềm năng (PSRF) đối với tất cả các tham số đều nhỏ hơn 1.2. Bên cạnh đó, trong chẩn đoán Geweke, khi tất cả các giá trị Z-score chuẩn hóa đều có giá trị tuyệt đối nhỏ hơn 2 thể hiện chuỗi đã hội tụ về phân phối hậu nghiệm mục tiêu.

SỐ LIỆU THỰC TẾ

Số liệu xét nghiệm tính độc tính của một loại hóa chất được áp dụng trên chuột là kết quả nghiên cứu của Racine (1986)¹⁹ được trình bày ở Bảng 6. Trong nghiên cứu này, có 35 con chuột được tiêm một loại thuốc với các liều lượng khác nhau và sau một thời gian thì số lượng con chuột chết và số lượng con chuột còn sống sẽ được thống kê lại. Mô hình hồi quy logistic được áp dụng vào số liệu này với biến độc lập là lượng thuốc được tiêm vào chuột và biến phụ thuộc là biến nhị phân với hai giá trị thể hiện con chuột còn sống hay chết sau thí nghiệm. Phương trình hồi quy logistic sẽ có dạng

$$\ln\left(\frac{p}{1-p}\right) = \beta_1 + \beta_2 x.$$

Bảng 6. Số liệu nghiên cứu của Racine (1986)

Liều lượng (x) (log mg/ml)	Số lượng thí nghiệm	Số lượng chết (y)
1.699	5	1
2.301	10	4
2.477	5	2
2.602	10	6
3.301	5	4

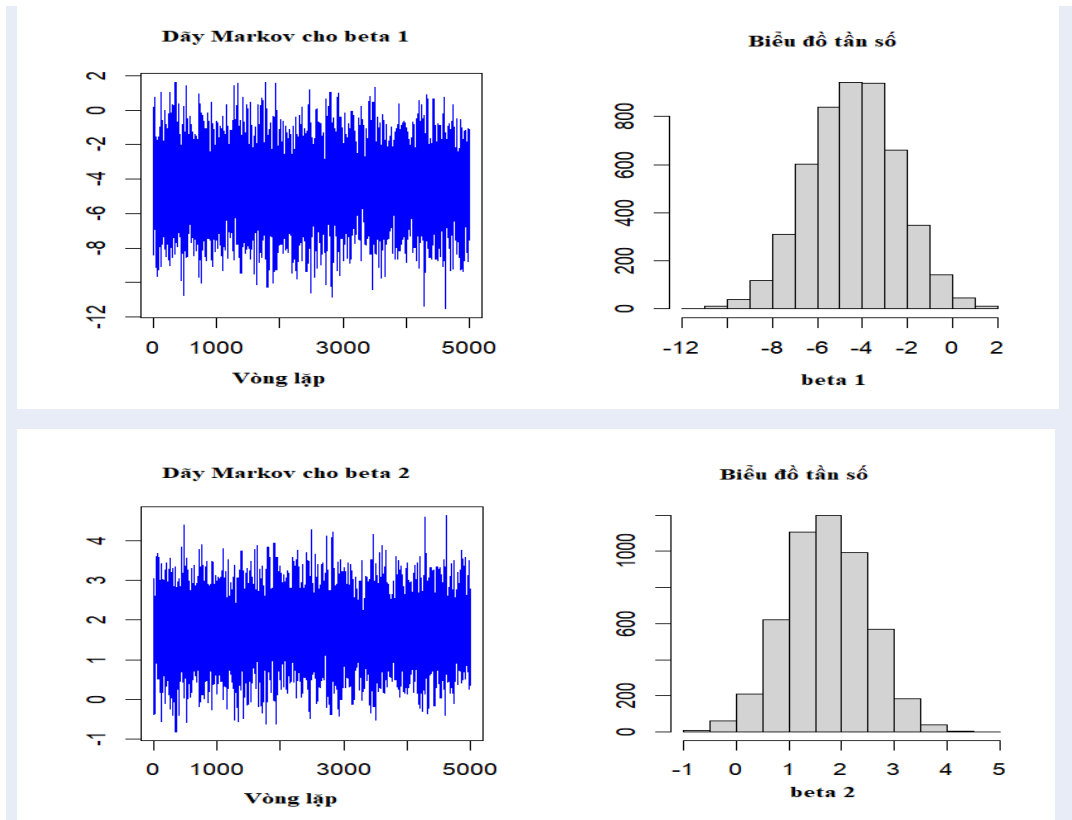
Ba phương pháp ước lượng cho mô hình được áp dụng bao gồm phương pháp cực đại hàm hợp lý, phương pháp Bayes với thông tin tiên nghiệm có phân phối Cauchy và phương pháp Bayes có thông tin tiên nghiệm thu được từ số liệu quan sát. Kết quả ước lượng được trình bày trong Bảng 7.

Bảng 7. Kết quả ước lượng tham số từ phương pháp MLE, phương pháp Bayes với phân phối tiên nghiệm Cauchy và phương pháp Bayes được đề xuất

Tham số	MLE			
	Trung bình	SD	z-value	p-value
β_1	-4.5420	2.3811	-1.9075	0.0565
β_2	1.8159	0.9541	1.9032	0.0570
Tham số	Bayes có phân phối tiên nghiệm là Cauchy			
	Trung bình	SD	z-value	p-value
β_1	-3.9299	2.1117	-1.8610	0.0628
β_2	1.5677	0.8433	1.8590	0.0631
Tham số	Bayes có phân phối tiên nghiệm thu được từ số liệu quan sát			
	Trung bình	SD	z-value	p-value
β_1	-4.4698	1.9653	-2.2743	0.0229
β_2	1.7409	0.7796	2.2330	0.0255

Từ Bảng 7, ta thấy ước lượng tham số $(\beta_1, \beta_2) = (-4.5420; 1.8159)$ từ phương pháp MLE, $(\beta_1, \beta_2) = (-3.9299; 1.5677)$ từ phương pháp Bayes với phân phối tiên nghiệm Cauchy và $(\beta_1, \beta_2) = (-4.4698; 1.7409)$ từ phương pháp Bayes được đề xuất. Độ lệch chuẩn (SD) cho các ước lượng từ phương pháp Bayes có thông tin tiên nghiệm nhỏ hơn so với phương pháp MLE và phương pháp Bayes không có thông tin tiên nghiệm. Điều này cho thấy, thông tin tiên nghiệm của tham số có được từ số liệu quan sát giúp các ước lượng tham số được tập trung hơn và chính xác hơn.

Với giá trị kiểm định z-value trong Bảng 7, p-value được tính bằng công thức $P(|Z| > |z\text{-value}|)$, với $Z \sim N(0,1)$. Kết quả cho thấy, hai phương pháp MLE và phương pháp Bayes với phân phối tiên nghiệm Cauchy có giá trị p-value cho từng tham số trong mô hình đều lớn hơn 0.05, nên chưa đủ bằng chứng để kết luận các tham số có ý nghĩa thống kê ở mức ý nghĩa 5%. Trong khi đó, với phương pháp Bayes sử dụng phân phối tiên nghiệm đề xuất, các giá trị p-value đều nhỏ hơn 0.05, cho thấy các tham số có ý nghĩa thống kê. Điều này thể hiện biến liều lượng thuốc có ảnh hưởng đáng kể đến tỷ lệ chết của chuột trong mô hình hồi quy logistic từ phương pháp được đề xuất. Hơn nữa, hình vẽ của dãy Markov cho các giá trị ước lượng cũng rất ổn định và bao quanh giá trị ước lượng trung bình ở Hình 3 ([Nguồn: nhóm tác giả]).



Hình 3. Dãy Markov và biểu đồ tần số cho ước lượng tham số từ phương pháp Bayes cải tiến

KẾT LUẬN

Khi số liệu quan sát trong mô hình hồi quy logistic có xác suất dự đoán của biến phụ thuộc gần chắc chắn xảy ra hoặc không xảy ra, đặc biệt khi có cảnh báo xác suất gần 0 hoặc 1 của hàm glm, thì giá trị của hàm hợp lý sẽ bị ảnh hưởng và tiến gần về không. Do đó, ước lượng tham số của mô hình với phương pháp cực đại hàm hợp lý sẽ xuất hiện nhiều sai số.

Qua kết quả mô phỏng, với thông tin tiên nghiệm được xác định từ số liệu quan sát, phương pháp Bayes được đề xuất giúp làm giảm sai số trong ước lượng tham số so với phương pháp cực đại hàm hợp lý và phương pháp Bayes không có thông tin tiên nghiệm. Do đó, khi thông tin tiên nghiệm càng chính xác thì phương pháp Bayes càng nâng cao độ chính xác trong việc ước lượng tham số của mô hình hồi quy logistic.

Trong phạm vi bài báo này, chúng tôi chỉ xem xét mô hình hồi quy logistic với các biến độc lập và số chiều của mô hình nhỏ. Cần có những nghiên cứu thêm để xem xét sự hiệu quả và chi phí tính toán của phương pháp được đề xuất cho trường hợp có sự tương quan giữa các biến, số chiều của mô hình lớn cũng như độ nhạy của phương pháp đề xuất với cấu trúc của mô hình sinh dữ liệu.

Trong tương lai, các phương pháp hiện đại hơn như Approximate Bayesian computation - sequential Monte Carlo (ABC-SMC) sẽ được nghiên cứu để tìm điều kiện xác định ngưỡng chấp nhận của tham số đề xuất. Khi đó, thuật toán này kỳ vọng sẽ giúp giảm sai số trong ước lượng tham số, đặc biệt trong trường hợp số liệu xuất hiện hiện tượng tách biệt hoặc gần tách biệt. Bên cạnh đó, cần

có nghiên cứu giúp giảm hệ số quy mô tiềm năng (PSRF) trong kiểm định Gelman -Rubin của nghiên cứu này về mức an toàn hơn.

LỜI CẢM ƠN

Nghiên cứu được tài trợ bởi Đại học Quốc gia Thành phố Hồ Chí Minh (ĐHQG-HCM) trong khuôn khổ đề tài mã số DS.C2025-16-02.

DANH MỤC TỪ VIẾT TẮT

MLE: cực đại hàm hợp lý

MH: Metropolis-Hastings

TUYÊN BỐ XUNG ĐỘT LỢI ÍCH

Bản thảo này không có xung đột lợi ích.

TUYÊN BỐ ĐÓNG GÓP TÁC GIẢ

Công trình này là thành quả của nhóm nghiên cứu và không trùng lặp ý tưởng đề xuất với bất kỳ công trình nào khác đã công bố.

TÀI LIỆU THAM KHẢO

1. Albert A, Anderson JA. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*. 1984;71(1):1–10. Available from: <https://doi.org/10.1093/biomet/71.1.1>.
2. Atkinson AC, Woods DC. Designs for generalized linear models. In: *Handbook of design and analysis of experiments*. Chapter 7. 2015. Available from: <https://doi.org/10.48550/arXiv.1510.05253>.
3. Harrell F. Regression modeling strategies: with applications to linear models, logistic and ordinal regression and survival analysis, New York: Springer; 2015. Available from: <https://doi.org/10.18637/jss.v070.b02>.
4. Speckman PL, Lee J, Sun D. Existence of the MLE and propriety of posteriors for a general multinomial choice model. *Statistica Sinica*. 2009;19(1):731–748. Available from: <https://www.jstor.org/stable/24308853>
5. Wakefield J. Bayesian and Frequentist Regression Methods. New York: Springer Science & Business Media; 2013. Available from: <http://faculty.washington.edu/jonno/regression-methods.html>
6. Allison PD. Convergence failures in logistic regression. *SAS Glob Forum*. 2008;360(1):1-11. Available from: <https://support.sas.com/resources/papers/proceedings/pdfs/sgf2008/360-2008.pdf>.
7. Cabras S, Castellanos ME, Ratmann, O. Goodness of fit for models with intractable likelihood. *TEST*. 2020; 30(3):713-736. Available from: <https://doi.org/10.1007/s11749-020-00747-7>.
8. Firth D. Bias reduction of maximum likelihood estimates. *Biometrika*. 1993;80(1):27-38. Available from: <http://www.jstor.org/stable/2336755>.

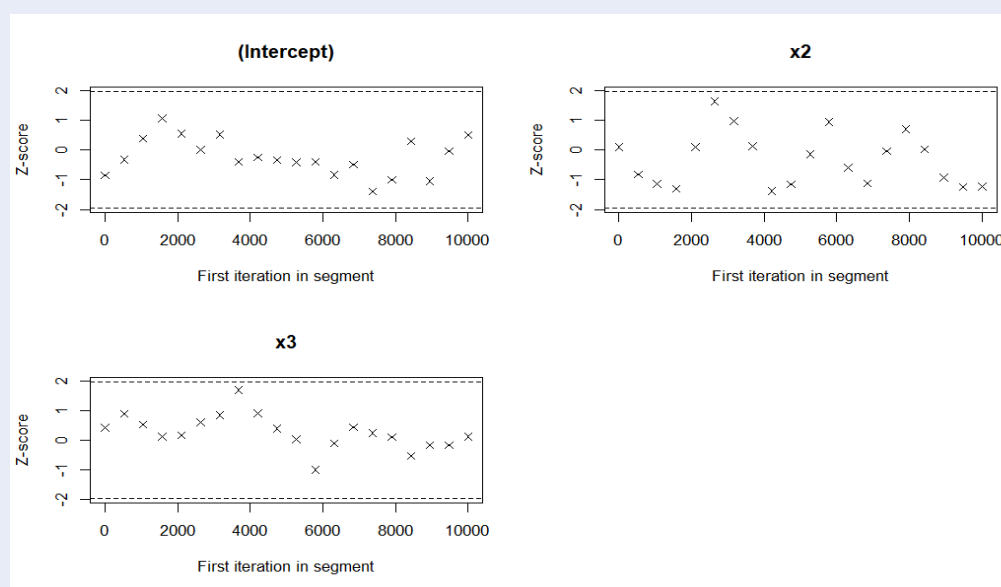
9. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. *Statistics in medicine*. 2002;21:2409-2419. Available from: <https://doi.org/10.1002/sim.1047>.
10. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. Chapman & Hall; 1995. Available from: <https://doi.org/10.1201/b16018>.
11. Gelman A, Jakulin A, Pittau MG, Su Y. A weakly informative default prior distribution for logistic and other regression models. *The annals of applied statistics*. 2008;2(4):1360–1383. Available from: <https://doi.org/10.48550/arXiv.0901.4011>.
12. Huong PTT, Hoa P, Kai SY. Applying non-informative G-prior for logistic regression models with different patterns of data points. *Monte Carlo Methods and Applications*. 2025. Available from: <https://doi.org/10.1515/mcma-2025-2016>.
13. Sisson SA, Fan Y, Beaumont M. *Handbook of approximate Bayesian computation*. CRC Press; 2018. Available from: <https://doi.org/10.1080/01621459.2020.1846973>.
14. Ghosh Y, Li J, Mitra R. On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis*. 2018;13(2):359–383. Available from: <https://doi.org/10.1214/17-BA1051>.
15. Heinze G. A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in medicine*. 2006;25(24):4216–4226. Available from: <https://doi.org/10.1002/sim.2687>.
16. Dennis P. Adapting the ABC Distance Function. *Bayesian Anal*. 2017;12(1):289-309. Available from: <https://doi.org/10.1214/16-BA1002>.
17. Drovandi CC., Pettitt AN, Faddy MJ. Approximate Bayesian computation using indirect inference. *Journal of the Royal Statistical Society Series C: Applied Statistics*. 2011;60(3): 317–337. Available from: <https://doi.org/10.1111/j.1467-9876.2010.00747.x>.
18. Racine A, Grieve AP, Flühler H, Smith AFM. Bayesian methods in practice: experiences in the pharmaceutical industry. *Applied Statistics*. 1986;35(2):93-150. Available from: <https://doi.org/10.2307/2347264>.
19. Huong PTT, Hoa P. On the existence of posterior mean for Bayesian logistic regression. *Monte Carlo Methods and Applications*. 2021;7(3):277–288. Available from: <https://doi.org/10.1515/mcma-2021-2089>.
20. Huong PTT, Hoa P. Controlling Separation in Generating Samples for Logistic Regression Models. *Mathematical methods of Statistics*. 2024;33(1):1–11. Available from: <https://doi.org/10.3103/S1066530724700017>.

21. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statistical Science. 1992;4 (7): 457–472. Available from: <https://doi.org/10.1214/ss/1177011136>.
22. Geweke J. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments Staff Report 148, Federal Reserve Bank of Minneapolis, 1991. Available from: <https://doi.org/10.21034/sr.148>.

PHỤ LỤC

A. Kết quả chẩn đoán cho mô hình hồi quy logistic Trường hợp 2 với $\beta = (-1, 1, 1)$

Phụ lục A.1. Chẩn đoán Geweke cho trường hợp $\beta = (-1, 1, 1)$



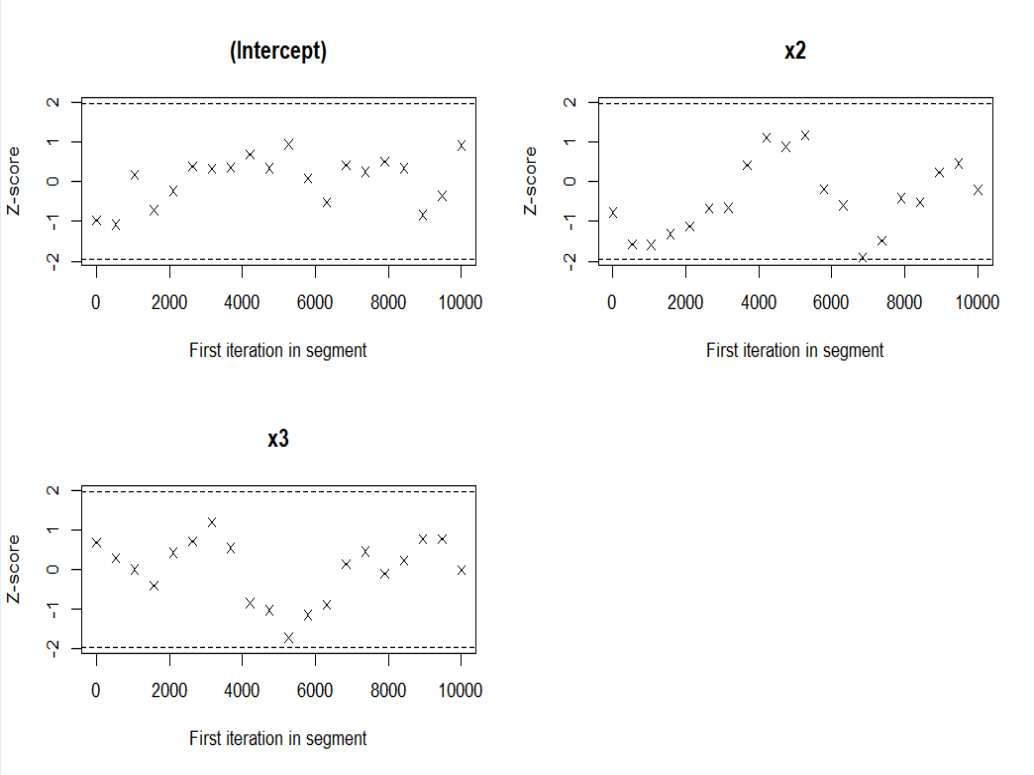
Phụ lục A.2. Chẩn đoán Gelman và Rubin cho trường hợp $\beta = (-1, 1, 1)$

	Ước lượng điểm	Cận trên khoảng tin cậy
Hệ số tự do	1.19	1.35
	1.05	1.24
	1.11	1.20
Hệ số PSRF x_2	1.15	

x_3

B. Kết quả chẩn đoán cho mô hình hồi quy logistic Trường hợp 2 với $\beta = (-2, 2, 1)$

Phụ lục B.1. Chẩn đoán Geweke cho trường hợp



Phụ lục B.2. Chẩn đoán Gelman và Rubin cho trường hợp $\beta = (-2, 2, 1)$

	Ước lượng điểm	Cận trên khoảng tin cậy
Hệ số tự do	1.18	1.45
x_2	1.09	1.24
x_3	1.16	1.31
Hệ số PSRF	1.17	

A new Bayesian approach to reduce estimation errors in logistic regression models

Pham Thi Thu Hoa*, Pham Thi Thu Huong



Use your smartphone to scan this QR code and download this article

ABSTRACT

When a logistic regression dataset contains a large number of observations with estimated probabilities very close to 0 or 1, the Maximum Likelihood Estimation (MLE) method often encounters difficulties in parameter estimation and may produce unstable estimates with large errors. In this context, the Bayesian approach is considered an effective alternative because it allows prior information to be combined with the observed data. However, the performance of Bayesian methods depends substantially on the selection of appropriate prior distributions for the model parameters. In this paper, we propose a new Bayesian method to reduce estimation errors in logistic regression models. First, a parameter acceptance–rejection algorithm is employed to construct prior information based on the similarity between simulated data and observed data. The resulting prior information is then combined with the likelihood function to establish the posterior distribution of the parameters. Subsequently, the Metropolis–Hastings (MH) algorithm is applied to generate a stationary Markov chain and estimate the model parameters. The performance of the proposed method is evaluated through simulation studies and a real-world dataset. The results demonstrate that the proposed Bayesian approach improves the accuracy of parameter estimates and reduces estimation errors compared with both the MLE method and Bayesian methods that employ non-informative priors.

Key words: Bayesian statistics, frequentist statistics, logistic regression model, reject-accept method, simulation study

Department of Mathematics, An Giang University, Vietnam National University - Ho Chi Minh City, Ho Chi Minh, Vietnam.

Correspondence

Pham Thi Thu Hoa, Department of Mathematics, An Giang University, Vietnam National University - Ho Chi Minh City, Ho Chi Minh, Vietnam.

Email: pthoa@agu.edu.vn

History

- Received: 06-09-2025
- Revised: 03-06-2026
- Accepted: 14-08-2026
- Published Online: 17-09-2026

DOI : <https://doi.org/10.32508/vnuhcmj-arns.v10i3.1474>



Copyright

© VNUHCM Journal. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International license.

Cite this article : Thu Hoa P T, Thu Huong P T. **A new Bayesian approach to reduce estimation errors in logistic regression models.** *VNUHCM J. Adv. Res. Nat. Sci.* 2026; 10(3): 3745-3766.