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.2023.157

THUẬT TOÁN KIỂM SOÁT SỰ TÁCH BIỆT TRONG SỐ LIỆU MÔ PHỎNG THEO MÔ HÌNH HỒI QUY LOGISTIC

Phạm Thị Thu Hường* và Phạm Thị Thu Hoa
Bộ môn Toán, Khoa Sư phạm, Trường Đại học An Giang, Đại học Quốc gia Thành phố Hồ Chí Minh
**Người chịu trách nhiệm về bài viết: Phạm Thị Thu Hường (email: ptthuong@agu.edu.vn)*

Thông tin chung:
Ngày nhận bài: 26/01/2023
Ngày nhận bài sửa: 03/03/2023
Ngày duyệt đăng: 07/03/2023

Title:
Algorithms to control separation in simulated data for logistic regression model

Từ khóa:
Mô hình hồi quy logistic, nghiên cứu mô phỏng, số liệu tách biệt, số liệu hầu như tách biệt

Keywords:
Logistic regression model, simulation study, separation, quasi – separation

ABSTRACT

Separation appearing in logistic regression data greatly influences the estimated values of the regression parameters. In classical statistics, the maximum likelihood estimation will not exist when the data appear to be separated. In Bayesian statistics, the existence of a posterior mean depends on the prior distribution and the pattern of the data. Therefore, in the simulation study, it is meaningful to control the probability of separation occurrence in the data and study the impact of this type of data in statistical analysis. In this paper, we present algorithms to simulate data for a logistic regression model where separation occurrence in the data is controlled for any sample size and dimension of the independent variable. These algorithms are proven to be very effective through simulation results.

TÓM TẮT

Sự tách biệt xuất hiện trong số liệu theo mô hình hồi quy logistic có ảnh hưởng lớn đến giá trị ước lượng của các tham số hồi quy. Đối với thống kê cổ điển, ước lượng cực đại của hàm hợp lý sẽ không tồn tại khi số liệu xuất hiện sự tách biệt. Đối với thống kê Bayes, sự tồn tại của giá trị trung bình hàm hậu nghiệm phụ thuộc vào phân phối tiên nghiệm và kiểu dạng của số liệu. Do đó, trong nghiên cứu mô phỏng số liệu, việc kiểm soát xác suất xuất hiện của sự tách biệt là có ý nghĩa để nghiên cứu tác động của dạng số liệu này trong phân tích thống kê. Trong bài báo này, những thuật toán được trình bày để mô phỏng số liệu theo mô hình hồi quy logistic mà sự xuất hiện tách biệt trong số liệu được kiểm soát với bất kỳ cỡ mẫu và số chiều của biến độc lập. Những thuật toán này được kiểm chứng có hiệu quả rất tốt qua kết quả mô phỏng.

1. GIỚI THIỆU

Trong cả hai cách tiếp cận của thống kê cổ điển và thống kê Bayes, ước lượng tham số trong mô hình hồi quy logistic được xem là một vấn đề khó khăn do mức độ phức tạp của hàm hợp lý trong mô hình. Điều này có thể dẫn đến sự lặp lại nhiều lần của các vòng lặp tính toán để ước tính giá trị cực đại của hàm hợp lý với thống kê tần suất và trung bình của hàm mật độ hậu nghiệm đối với suy luận Bayes. Hơn nữa, các tính toán cho ước lượng của tham số thống kê trong mô hình hồi quy logistic sẽ chịu tác động quan trọng khi có sự tách biệt trong dữ liệu (Allison, 2008; Gelman, 2008; Wakefield, 2013; Atkinson & Woods, 2015).

Sự xuất hiện tách biệt trong số liệu của mô hình hồi quy logistic có ảnh hưởng rất lớn đến suy luận

thống kê theo cả hai cách tiếp cận là thống kê cổ điển và thống kê Bayes. Trong cách tiếp cận của thống kê cổ điển, ước lượng cho tham số của hàm hồi quy phụ thuộc hoàn toàn vào hàm hợp lý cực đại. Albert and Anderson (1984) đã xem xét ảnh hưởng của các dữ liệu có sự tách biệt đến ước lượng cực đại cho hàm hợp lý của các mô hình hồi quy logistic. Nhóm tác giả đã chứng minh rằng khi các quan sát trong mẫu có sự phân loại thì ước lượng cực đại cho hàm hợp lý sẽ không tồn tại. Cụ thể là nếu trong số liệu của mô hình có xuất hiện sự tách biệt hoặc hầu như tách biệt thì ước lượng hàm hợp lý cực đại cho các tham số sẽ tiến ra vô cùng.

Trong thống kê Bayes, hàm phân phối hậu nghiệm được xác định bởi hàm hợp lý cực đại và hàm phân phối tiên nghiệm. Sự tồn tại của ước lượng cho tham số trong mô hình hồi quy logistic có nhiều sự ảnh hưởng bởi sự hiện diện của sự tách biệt trong số liệu (Heinze, 2006; Speckman, 2009; Polson, 2013; Ghosh, 2018; Huong & Hoa, 2021). Polson (2013) đã chứng minh sự tồn tại của hàm phân phối hậu nghiệm thông qua chuỗi Markov và phân phối chuẩn cho phân phối tiên nghiệm. Ghosh (2018) đã chứng minh sự tồn tại cho giá trị trung bình của phân phối hậu nghiệm khi sử dụng phân phối tiên nghiệm tuân theo phân phối Cauchy. Huong & Hoa (2021) đã chứng minh cho trường hợp tổng quát như sau: i. Nếu có sự hiện diện của sự tách biệt trong dữ liệu và không có thông tin trước đối với tham số thì giá trị trung bình hậu nghiệm không tồn tại; ii. Nếu có sự hiện diện của sự phân tách trong dữ liệu và phân phối tiên nghiệm cho tham số có dạng phân phối tiêu chuẩn thì trung bình cho phân phối tiên nghiệm thực sự tồn tại; iii. Nếu có sự hiện diện của sự phân tách trong dữ liệu và phân phối thông tin tiên nghiệm là không tiêu chuẩn thì giá trị trung bình phân phối hậu nghiệm không tồn tại.

Theo Albert and Anderson (1984), khi cỡ mẫu nhỏ và số chiều của biến độc lập thấp thì sự xuất hiện tách biệt trong số liệu là thường gặp trong số liệu thực tế và số liệu mô phỏng. Tuy nhiên, khi cỡ mẫu lớn và số chiều của biến độc lập cao, để có sự tách biệt xuất hiện hay không xuất hiện trong số liệu mô phỏng là công việc khó và cần các kết quả lý thuyết cũng như thuật toán để thực hiện. Trong bài báo này, các thuật toán để mô phỏng ra mẫu ngẫu nhiên theo mô hình hồi quy logistic được xây dựng mà ở đó có thể kiểm soát được xác suất xuất hiện của dữ liệu tách biệt với bất kì cỡ mẫu và số chiều của biến độc lập được cho bất kỳ. Cụ thể, các điều kiện cần và đủ cho sự xuất hiện tách biệt trong số liệu được trình bày trong bài báo. Từ kết quả này, các bước và các kỹ thuật chọn tham số trong mô hình để mô phỏng số liệu được trình bày với số chiều của biến độc lập khác nhau và cỡ mẫu khác nhau mà ở đó xác suất xuất hiện sự tách biệt được kiểm soát. Thuật toán này đã được kiểm chứng có hiệu quả qua kết quả mô phỏng.

2. KẾT QUẢ NGHIÊN CỨU LÝ THUYẾT

2.1. Giới thiệu mô hình hồi quy logistic và số liệu tách biệt

Trong bài báo này, mô hình hồi quy logistic được sử dụng cho mô phỏng số liệu. Giả sử số liệu có 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 **x** là một vector có p chiều

$$\mathbf{x} = \begin{bmatrix} 1 & x_{12} & \dots & x_{1p} \\ 1 & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n2} & \dots & x_{np} \end{bmatrix}.$$

Ở đây, cột đầu tiên của ma trận là hệ số hằng và được giả sử có giá trị bằng 1. Biến phụ thuộc $\mathbf{y} = (y_1, ..., y_n)^T$ 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 $\mathbf{x}_i = \begin{bmatrix} 1 & x_{1i} & x_{2i} & \dots & x_{ni} \end{bmatrix}^T$ 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 mũ làm liên kết giữa biến độc lập và biến phụ thuộc có dạng sau:

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

Trong đó, vector $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_p)^T$ là tham số hồi quy trong mô hình hồi quy logistic. Khi đó

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

Tiếp theo, định nghĩa số liệu tách biệt và hầu như tách biệt trong mô hình hồi quy logistic được trình bày dựa theo bài báo của Albert and Anderson (1984).

Cho hai tập con rời nhau của tập $\{1,2,3,...,n\}$ dựa vào giá trị của biến ngẫu nhiên **y** được định nghĩa như sau: $A_0 = \{i : y_i = 0\}$ và $A_1 = \{i : y_i = 1\}$.

Định nghĩa 1: Sự tách biệt hoàn toàn xảy ra trên số liệu của mô hình hồi quy logistic khi tồn tại một vector $\boldsymbol{\alpha} = (\alpha_1,...,\alpha_p)^T \in \mathbb{R}^p$ sao cho:

$$\mathbf{x}_i^T\boldsymbol{\alpha} > 0 \ \ khi \ \ i \in A_1 \ \ và \ \ \mathbf{x}_i^T\boldsymbol{\alpha} < 0 \ \ khi \ \ i \in A_0,$$

cho tất cả các giá trị $i = 1,2,...,n$.

Định nghĩa 2: Sự hầu như tách biệt xảy ra trong số liệu của mô hình hồi quy logistic khi tồn tại một vector $\boldsymbol{\alpha} = (\alpha_1,...,\alpha_p)^T \in \mathbb{R}^p$ sao cho:

$$\mathbf{x}_i^T\boldsymbol{\alpha} \geq 0 \ \ khi \ \ i \in A_1 \ \ và \ \ \mathbf{x}_i^T\boldsymbol{\alpha} \leq 0 \ \ khi \ \ i \in A_0,$$

cho tất cả các giá trị $i = 1,2,...,n$.

2.2. Điều kiện cần và đủ để xuất hiện sự tách biệt trong mô hình hồi quy logistic với biến độc lập có hai chiều

Ta giả sử rằng trong mô hình hồi quy logistic chỉ có biến độc lập với hai chiều, vector cho biến độc lập có dạng sau:

$$\mathbf{x} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \cdots & \cdots \\ 1 & x_{n1} \end{bmatrix}.$$

Ở đây, giá trị ở cột thứ nhất được coi là hệ số hằng và chúng ta giả sử hệ số này có giá trị là một. Vector cho biến phụ thuộc $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ tuân theo phân phối Bernoulli. Khi đó, hàm liên kết giữa biến phụ thuộc **y** và biến độc lập **x** theo (2.1) được cho bởi:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i^T\boldsymbol{\alpha}.$$

Ta xác định hai tập hợp con của biến $\mathbf{x}_1 = (x_{11}, x_{21}, ..., x_{n1})^T$ được cho bởi

$$X_0 = \{x_{i1}, i \in A_0\} \text{ và } X_1 = \{x_{i1}, i \in A_1\}.$$

Điều kiện cần và đủ cho xuất hiện sự tách biệt và hầu như tách biệt trong dữ liệu tuân theo mô hình hồi quy logistic được trình bày trong Định lý 1.

Định lí 1:

Ta xác định khoảng cách giữa 2 tập hợp X_0 và X_1 như sau:

$$\begin{aligned} d_1 &= \min(X_0) - \max(X_1), \\ d_2 &= \min(X_1) - \max(X_0), \\ d &= \max[d_1, d_2]. \end{aligned} \qquad (2.3)$$

a) Số liệu có xuất hiện tách biệt khi và chỉ khi khoảng cách $d > 0$.

b) Số liệu hầu như có xuất hiện tách biệt khi và chỉ khi khoảng cách $d \geq 0$.

Chứng minh.

a) Trước tiên ta chứng minh điều kiện cần.

$(\Rightarrow)$ Ta giả sử rằng $d = \min(X_1) - \max(X_0)$ và $d > 0$. Theo định nghĩa 1, để chứng minh trong số liệu có sự tách biệt, ta cần xác định vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^T \in \mathbb{R}^2$ sao cho:

$$\mathbf{x}_i^T\boldsymbol{\alpha} > 0 \, khi \, i \in A_1 \, và \, \mathbf{x}_i^T\boldsymbol{\alpha} < 0 \, khi \, i \in A_0.$$

Trước hết ta định nghĩa hai tập hợp rời nhau E_0 và E_1 như sau:

$$E_0 = \{x_{i1} - \max(X_0) - d/2, i \in A_0\},$$

và

$$E_1 = \{x_{i1} - \max(X_0) - d/2, i \in A_1\}.$$

Ta nhận thấy rằng các phần tử thuộc tập hợp E_0 có giá trị âm và các phần tử thuộc tập hợp E_1 có giá trị dương. Do đó, với tham số $t > 0$ bất kì, ta luôn có:

$$\begin{cases} (x_{1i} - \max(X_0) - d/2)t < 0, & i \in A_0. \\ (x_{1i} - \max(X_0) - d/2)t > 0, & i \in A_1. \end{cases}$$

Với vector $\boldsymbol{\alpha} = (-t(\max(X_0) + d/2), t)$ và giá trị tham số $t > 0$ bất kì, ta có:

$$\mathbf{x}_i^T\boldsymbol{\alpha} = -t(\max(X_0) + d/2) + x_{1i}t$$
$$= (x_{1i} - \max(X_0) - d/2)t < 0,\ i \in A_0,$$
$$\mathbf{x}_i^T\boldsymbol{\alpha} = -t(\max(X_0) + d/2) + x_{1i}t$$
$$= (x_{1i} - \max(X_0) - d/2)t > 0,\ i \in A_1.$$

Như vậy, ta đã chứng minh được điều kiện cần cho Định lý 1.

$(\Leftarrow)$ Giả sử tồn tại vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)^T \in \mathbb{R}^2$ sao cho:

$$\mathbf{x}_i^T\boldsymbol{\alpha} > 0 \text{ khi } i \in A_1 \text{ và } \mathbf{x}_i^T\boldsymbol{\alpha} < 0 \text{ khi } i \in A_0.$$

Với $i \in A_0$, ta có

$$1\alpha_1 + x_{i1}\alpha_2 < 0$$
$$\Rightarrow x_{i1} < -\frac{\alpha_1}{\alpha_2}, i \in A_0.$$

Tương tự, ta có $x_{i1} > -\frac{\alpha_1}{\alpha_2}, i \in A_1$. Do đó,

$$d_1 = \min(X_0) - \max(X_1),$$
$$d_2 = \min(X_1) - \max(X_0),$$
$$d = \max[d_1, d_2] > 0. \quad \square$$

Kết quả chứng minh cho định lý vẫn đúng với trường hợp còn lại trong định nghĩa của khoảng cách d và kết quả cho số liệu hầu như tách biệt trong mục b) cũng được chứng minh tương tự.

2.3. Điều kiện cần và đủ xuất hiện tách biệt trong mô hình hồi quy logistic có biến độc lập với số chiều bất kỳ

Vector ma trận của biến độc lập được sử dụng có dạng sau:

$$\mathbf{x} = \begin{bmatrix} 1 & x_{12} & \dots & x_{1p} \\ 1 & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n2} & \dots & x_{np} \end{bmatrix}.$$

Ở đây, giá trị ở cột thứ nhất được coi là hệ số hằng và chúng ta giả sử hệ số này có giá trị là một. Vector cho biến phụ thuộc $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ tuân theo phân phối Bernoulli. Khi đó, hàm liên kết giữa biến phụ thuộc **y** và biến độc lập **x** theo (2.1) được cho bởi:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i^T\boldsymbol{\alpha}.$$

Ta tìm điều kiện cần và đủ để xuất hiện sự tách biệt trong số liệu của mô hình hồi quy logistic với số chiều của biến độc lập bất kỳ. Giả sử rằng có hai biến $\mathbf{x}_i$ và $\mathbf{x}_j$ là hai thành phần chính gây ra sự tách biệt trong số liệu. Hai tập hợp con của biến $x_i = (x_{1i}, x_{2i}, \dots, x_{ni})^T$ được cho bởi:

$$X_{0i} = \{x_{ki}, k \in A_0\} \text{ và } X_{1i} = \{x_{ki}, k \in A_1\}.$$

Hai tập hợp con của biến $x_j = (x_{1j}, x_{2j}, \dots, x_{nj})^T$ được cho bởi:

$$X_{0j} = \{x_{kj}, k \in A_0\} \text{ và } X_{1j} = \{x_{kj}, k \in A_1\}.$$

Các khoảng cách d_i và d_j được định nghĩa như sau:

$$d_{1i} = \min(X_{0i}) - \max(X_{1i}),$$
$$d_{2i} = \min(X_{1i}) - \max(X_{0i}), \qquad (2.4)$$
$$d_i = \max[d_{1i}, d_{2i}],$$

và

$$d_{1j} = \min(X_{0j}) - \max(X_{1j}),$$
$$d_{2j} = \min(X_{1j}) - \max(X_{0j}), \qquad (2.5)$$
$$d_j = \max[d_{1j}, d_{2j}].$$

Định lí 2: Nếu tồn tại giá trị bất kì $i, i \in \{1, \dots, p\}$ và $j, j \in \{1, \dots, p\}$ sao cho $d_i > 0$ và $d_j > 0$, khi đó, số liệu sẽ xuất hiện sự tách biệt.

Chứng minh.

Ta giả sử rằng $d_i = \min\{X_{1i}\} - \max\{X_{0i}\}$ và $d_j = \min\{X_{1j}\} - \max\{X_{0j}\}$. Như định nghĩa sự tách biệt số liệu ta đi xác định vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_p)^T \in \mathbb{R}^p$ sao cho:

$$\mathbf{x}_i^T\boldsymbol{\alpha} > 0 \text{ khi } i \in A_1 \text{ và } \mathbf{x}_i^T\boldsymbol{\alpha} < 0 \text{ khi } i \in A_0.$$

Trước hết ta định nghĩa các cặp tập hợp rời nhau như sau:

$$E_{0i} = \{x_{ki} - \max\left(X_{0i}\right) - d_i/2, k \in A_0\};$$
$$E_{1i} = \{x_{ki} - \max\left(X_{0i}\right) - d_i/2, k \in A_1\}$$

và

$$E_{0j} = \{x_{kj} - \max\left(X_{0j}\right) - d_j/2, k \in A_0\};$$
$$E_{1j} = \{x_{kj} - \max\left(X_{0j}\right) - d_j/2, k \in A_1\}.$$

Ta có nhận xét sau, các phần tử thuộc các tập hợp E_{1i}, E_{1_j} có giá trị dương và các phần tử thuộc các tập hợp E_{0i}, E_{0j} có giá trị âm. Do đó, luôn tồn tại tham số $t_i > 0$ và $t_j > 0$ đủ lớn sao cho:

$$\begin{cases}(x_{ki} - \max\{X_{0i}\} - d_i/2)t_i \to -\infty, & k \in A_0.\\ (x_{ki} - \max\{X_{0i}\} - d_i/2)t_i \to +\infty, & k \in A_1.\end{cases}$$

Và

$$\begin{cases}(x_{kj} - \max\{X_{0j}\} - d_j/2)t_j \to -\infty, & k \in A_0.\\ (x_{kj} - \max\{X_{0j}\} - d_j/2)t_j \to +\infty, & k \in A_1.\end{cases}$$

Ta xác định vector $\boldsymbol{\alpha}$ như sau:

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_p), \qquad (2.6)$$

trong đó,

$$\alpha_1 = -t_i[\max\{E_{0i}\} + d_i/2] - t_j[\max\{E_{0j}\} + d_j/2];$$
$$\alpha_i = t_i;$$
$$\alpha_j = t_j.$$

Ta thấy, với giá trị tham số $t_i > 0, t_j > 0$ đủ lớn và $k \in A_0$, ta có:

$$\mathbf{x}_k^T\boldsymbol{\alpha} = -t_i[\max\left(E_{0i}\right) + d_i/2] - t_j[\max\left(E_{0j}\right) + d_j/2] + x_{k2}\alpha_2 + ... + x_{ki}t_i + ... + x_{kj}t_j + ... + x_{kp}\alpha_p$$
$$= \left(x_{ki} - \max\left\{E_{0i}\right\} - d_i/2\right)t_i + \left(x_{kj} - \max\left\{E_{0j}\right\} - d_j/2\right)t_j + x_{i2}\alpha_2 + ... + x_{i(j-1)}\alpha_{j-1} + x_{i(j+1)}\alpha_{j+1} + ... + x_{ip}\alpha_p \to -\infty, \text{ khi } t_i \to \infty, t_j \to \infty.$$

Tương tự, với giá trị $t_i > 0, t_j > 0$ đủ lớn và $k \in A_1$ ta có:

$$\mathbf{x}_k^T\boldsymbol{\alpha} = -t_i[\max\{E_{0i}\} + d_i/2] - t_j[\max\{E_{0j}\} + d_j/2] + x_{k2}\alpha_2 + ... + x_{ki}t_i + ... + x_{kj}t_j + ... + x_{kp}\alpha_p$$
$$= \left(x_{ki} - \max\left\{E_{0i}\right\} - d_i/2\right)t_i + \left(x_{kj} - \max\left\{E_{0j}\right\} - d_j/2\right)t_j + x_{i2}\alpha_2 + ... + x_{i(j-1)}\alpha_{j-1} + x_{i(j+1)}\alpha_{j+1} + ... + x_{ip}\alpha_p \to \infty, \text{ khi } t_i \to \infty, t_j \to \infty.$$

Như vậy, với giá trị $t_i > 0, t_j > 0$ đủ lớn, số liệu sẽ xuất hiện sự tách biệt. □

Ta cũng có kết quả tương tự như sau: nếu tồn tại giá trị bất kì $i, i \in \{1, ..., p\}$ và $j, j \in \{1, ..., p\}$ sao cho $d_i \geq 0$ và $d_j \geq 0$, khi đó, số liệu hầu như bị tách biệt .

Nhận xét: Nếu các thành phần gây ra số liệu tách biệt nhiều hơn 2 thì kết quả của định lí 2 vẫn đúng.

Khi số liệu xuất hiện sự tách biệt thì theo định nghĩa 1 tồn tại mặt phẳng $\mathbf{x}^T\boldsymbol{\alpha} = 0$ sao cho: $\mathbf{x}_i^T\boldsymbol{\alpha} > 0 \text{ khi } i \in A_1 \text{ và } \mathbf{x}_i^T\boldsymbol{\alpha} < 0 \text{ khi } i \in A_0,$ cho tất cả các giá trị $i = 1, 2, ..., n$. Ta gọi θ là góc giữa trục tọa độ Ox_i và mặt phẳng $\mathbf{x}^T\boldsymbol{\alpha} = 0$. Điều kiện đủ cho xuất hiện sự tách biệt trong mô hình được trình bày trong Định lý 3.

Định lý 3: Nếu số liệu được quay một góc $\frac{\pi}{2} - \theta$ sao cho mặt phẳng $\mathbf{x}^T\boldsymbol{\alpha} = 0$ vuông góc với trục tọa độ Ox_i thì khi đó

a) Tồn tại vector $\boldsymbol{\alpha}' = (\alpha'_1, ..., \alpha'_p)^T$ sao cho:

$\left(\mathbf{x}'_i\right)^T \boldsymbol{\alpha}' > 0 \; khi \; i \in A_1 \; và \left(\mathbf{x}'_i\right)^T \boldsymbol{\alpha}' < 0 \; khi \; i \in A_0,$

Ở đây, $\mathbf{x}'_i$ là ảnh của vector $\mathbf{x}_i$ sau khi quay và $\boldsymbol{\alpha}'$ là ảnh của vector $\boldsymbol{\alpha}$ sau khi quay.

b) Ta định nghĩa các tập sau: $E'_{0j} = \{x'_{ij}, i \in A_0\}$, $E'_{1j} = \{x'_{ij}, i \in A_1\}$ và

$$d'_{1i} = \min\left\{E'_{1i}\right\} - \max\left\{E'_{0i}\right\},$$
$$d'_{2i} = \min\left\{E'_{0i}\right\} - \max\left\{E'_{1i}\right\},$$
$$d'_i = \max\left\{d'_{1i}, d'_{2i}\right\}.$$

Khi đó, $d'_i > 0.$

Chứng minh:

Khi số liệu được quay một góc $\frac{\pi}{2} - \theta$ độ nên tồn tại một ma trận khả nghịch *A* sao cho:

$$\mathbf{x}' = \begin{bmatrix} x'_1 \\ x'_2 \\ \cdots \\ x'_p \end{bmatrix} = A.\mathbf{x} = A.\begin{bmatrix} x_1 \\ x_2 \\ \cdots \\ x_p \end{bmatrix}.$$

Mặt phẳng $\mathbf{x}^T\boldsymbol{\alpha} = 0$ có thể được trình bày lại như sau:

$$\mathbf{x}^T\boldsymbol{\alpha} = (A^{-1}.\mathbf{x}')^T\boldsymbol{\alpha} = \mathbf{x}'^T(A^{-1})^T\boldsymbol{\alpha} = 0,$$

trong đó, $\boldsymbol{\alpha}' = (A^{-1})^T\boldsymbol{\alpha}$.

Nếu $i \in A_1, \mathbf{x}_i^T\boldsymbol{\alpha} > 0$, ta có được kết quả

$$\left(\mathbf{x}'_i\right)^T \boldsymbol{\alpha}' = \left(A.\mathbf{x}_i\right)^T A\boldsymbol{\alpha} = \mathbf{x}_i^T A^T A\boldsymbol{\alpha}$$
$$= \left(A^T A\right)\mathbf{x}_i^T\boldsymbol{\alpha} > 0$$

Nếu $i \in A_0, \mathbf{x}_i^T\boldsymbol{\alpha} < 0$, ta có được kết quả

$$\left(\mathbf{x}'_i\right)^T \boldsymbol{\alpha}' = \left(A.\mathbf{x}_i\right)^T A\boldsymbol{\alpha} = \mathbf{x}_i^T A^T A\boldsymbol{\alpha}$$
$$= \left(A^T A\right)\mathbf{x}_i^T\boldsymbol{\alpha} < 0.$$

a) Khi mặt phẳng $\mathbf{x}'^T\boldsymbol{\alpha}' = 0$ vuông góc với trục trục tọa độ Ox_i nên phương trình của mặt phẳng này có dạng

$$\left(\mathbf{x}'_{(-j)}\right)^T \boldsymbol{\alpha}'_{(-j)} + k\alpha'_j = 0.$$

Ở đây, $x'_j = k.$ Ta định nghĩa $\boldsymbol{\alpha}_{(-j)}$ là vector α sau khi bỏ đi thành phần thứ *j*.

Nếu $i \in A_1$, ta có:

$$(\mathbf{x}'_i)^T\boldsymbol{\alpha}' > 0.$$

Do đó,

$$x'_{ij} > -\frac{(\mathbf{x}'_{i(-j)})^T\boldsymbol{\alpha}'_{i(-j)}}{\alpha'_j} > k.$$

Tương tự, nếu $i \in A_0$, ta có $(\mathbf{x}'_i)^T\boldsymbol{\alpha}' < 0$ và

$$x'_{ij} < -\frac{(\mathbf{x}'_{i(-j)})^T\boldsymbol{\alpha}'_{i(-j)}}{\alpha'_j} < k.$$

Theo định nghĩa của d'_j , ta có:

$$d'_j > 0. \qquad \square$$

3. KẾT QUẢ MÔ PHỎNG

3.1. Thuật toán mô phỏng số liệu trong mô hình hồi quy logistic kiểm soát sự tách biệt với biến độc lập có hai chiều

3.1.1. Thuật toán

Hàm liên kết giữa biến phụ thuộc **y** và biến độc lập **x** theo (2.1) có dạng mũ được sử dụng trong bài báo này:

$$\ln\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i^T\boldsymbol{\alpha}.$$

Để mô phỏng số liệu cho mô hình hồi quy logistic, biến ngẫu nhiên **y** được mô phỏng theo phân phối Bernoulli với tham số p_i và công thức tính xác suất p_i theo (2.2) được cho bởi:

$$p_i = \frac{exp(\mathbf{x}_i^T\boldsymbol{\alpha})}{1+exp(\mathbf{x}_i^T\boldsymbol{\alpha})}.$$

Như vậy, để đảm bảo sự xuất hiện của sự tách biệt trong số liệu mô phỏng, từ định lý 1 ta thấy,

Tham số t cần được điều chỉnh sao cho $\mathbf{x}_i^T\boldsymbol{\alpha} \to -\infty \; khi \; i \in A_0$ và $\mathbf{x}_i^T\boldsymbol{\alpha} \to \infty \; khi \; i \in A_1$. Khi đó,

$$\begin{cases} p_i \to 0, & i \in A_0. \\ p_i \to 1, & i \in A_1. \end{cases}$$

Xác suất để xuất hiện sự tách biệt đạt được rất cao và giá trị của biến $y_i \sim Bernoulli(p_i)$ sẽ nhận được là:

$$\begin{cases} y_i = 0, & i \in A_0. \\ y_i = 1, & i \in A_1. \end{cases}$$

Trong khi đó, nếu tham số t được điều chỉnh sao cho

$$\begin{cases} p_i \sim 0.5, & i \in A_0, \\ p_i \sim 0.5, & i \in A_1, \end{cases}$$

xác suất để xuất hiện sự tách biệt là gần như bằng không.

Do đó, với giá trị đề xuất trong Định lý 1, $\boldsymbol{\alpha} = (-t\left(\max\left(X_0\right) + d/2\right), t)$, khoảng cách giữa 2 tập X_0 và X_1 và độ lớn của tham số t sẽ ảnh hưởng trực tiếp đến độ lớn của xác suất $p_i, i = 1, 2, ..., n.$ Vì sự xuất hiện tách biệt trong mẫu mô phỏng phụ thuộc vào hai yếu tố này, thuật toán để kiểm soát sự tách biệt trong mẫu mô phỏng được đề xuất như sau:

Thuật toán 1:

Bước 1: Mô phỏng biến độc lập **x** thuộc vào hai nhóm X_0 và X_1.

Bước 2: Xác định vector $\boldsymbol{\alpha} = \left[-t\left(\max\left(X_0\right) + d/2\right), t\right].$

Bước 3: Tính giá trị $p_i, i = 1, 2, ..., n$ theo công thức:

$$p_i = \frac{exp(\mathbf{x}_i^T\boldsymbol{\alpha})}{1 + exp(\mathbf{x}_i^T\boldsymbol{\alpha})}.$$

Bước 4: Mô phỏng biến $y_i \sim Bernoulli(p_i)$.

Bước 5: Kết hợp giữa biến phụ thuộc **y** và biến độc lập **x** tạo thành mẫu ngẫu nhiên.

Lưu ý: Trong Bước 1, tập hợp tất cả các giá trị của biến ngẫu nhiên **x** là số quan sát n và số lượng phần tử của biến ngẫu nhiên **x** thuộc vào hai nhóm X_0 và X_1 không tác động đến xác suất xuất hiện sự tách biệt. Trong Bước 2, khoảng cách *d* càng lớn và giá trị tham số t càng lớn sẽ làm tăng xác suất xuất hiện của sự tách biệt trong số liệu mô phỏng.

3.1.2. Kết quả mô phỏng

Theo thuật toán 1, mô phỏng số liệu của biến ngẫu nhiên **x** được thực hiện theo phân phối đều với *n* = 100 quan sát. Hai nhóm của biến ngẫu nhiên x được mô phỏng như sau:

$$\begin{aligned} X_0 = & \quad \{x \sim U(a_0, b_0)\}, \\ X_1 = & \quad \{x \sim U(a_1, b_1)\}. \end{aligned}$$

Khoảng cách *d* giữa 2 nhóm X_0 và X_1 được xác định như sau:

$$\begin{aligned} d_1 &= \min\left(X_0\right) - \max\left(X_1\right), \\ d_2 &= \min\left(X_1\right) - \max\left(X_0\right), \\ d &= \max\left[d_1, d_2\right]. \end{aligned}$$

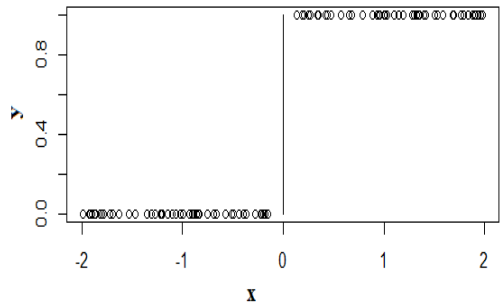
Bảng 1. Kết quả xác suất xuất hiện của sự tách biệt trong số liệu với các giá trị khoảng cách d và giá trị điều chỉnh t

d	***t***	**n**	**p**
2,0406	1	0	0%
2,0366	2	4	4%
2,0354	3	37	37%
2,0334	4	74	74%
2,0342	5	98	98%
2.0354	7	100	100%
0,5681	1	0	0%
0,5788	5	21	21%
0,5523	10	77	77%
0,5677	20	100	100%

Khi khoảng cách *d* và tham số *t* thay đổi sẽ tác động đến xác suất xuất hiện của sự tách biệt trong số liệu mô phỏng. Xác suất xuất hiện sự tách biệt được tính khi mô phỏng 100 mẫu ngẫu nhiên và đếm số xuất hiện của sự tách biệt khi thay đổi khoảng cách *d* và tham số *t*. Để kiểm tra sự xuất hiện của sự tách biệt trong số liệu mô phỏng, hàm gml trong

ngôn ngữ R được sử dụng. Kết quả của mô phỏng được thể hiện trong Bảng 1.

Trong đó, *d* là khoảng cách trung bình giữa 2 nhóm X_0 và X_1, *t* là giá trị tham số trong vector $\boldsymbol{\alpha}$, n là số lượng mẫu ngẫu nhiên xuất hiện sự tách biệt trong 100 mẫu mô phỏng và p là xác suất xuất hiện sự tách biệt. Từ Bảng 1, kết quả xác suất xuất hiện của sự tách biệt trong số liệu được trình bày với các giá trị khoảng cách *d* và giá trị điều chỉnh *t*, khi mô phỏng 100 số liệu ngẫu nhiên theo thuật toán 1. Chúng ta thấy, với khoảng cách cho trước, tham số *t* càng lớn thì xác suất xuất hiện của sự tách biệt càng lớn. Cụ thể, với *t* = 1, hầu như không có sự tách biệt trong số liệu mô phỏng. Khi *t* = 37 xác suất xuất hiện sự tách biệt trong số liệu là 37% và xác suất này tăng lên 100 khi *t* nhận giá trị bằng 7. Khi khoảng cách *d* giữa 2 tập X_0 và X_1 được chọn có giá trị trung bình khoảng 0,5, để số liệu có xác suất xuất hiện sự tách biệt trong số liệu là 100% giá trị tham số *t* được chọn cho thuật toán phải lớn hơn 20. Như vậy, với kỹ thuật chọn tham số *t* trong thuật toán 1, có thể kiểm soát được xác suất xuất hiện sự tách biệt trong số liệu với giá trị khoảng cách *d* giữa 2 tập X_0 và X_1 tùy ý. Hình vẽ của biến ngẫu nhiên **y** theo biến **x** khi số liệu có sự tách biệt với khoảng cách *d* = 0,2012 được trình bày trong Hình 1.



Hình 1. Hình vẽ của biến ngẫu nhiên y theo biến x với số liệu có tách biệt với khoảng cách *d* = 0,2012

3.2. Thuật toán mô phỏng số liệu mô hình hồi quy logistic có kiểm soát sự tách biệt với biến độc lập có nhiều chiều

3.2.1. Thuật toán

Ở đây, mô hình hồi quy logistic với biến ngẫu nhiên $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3]^T$ có số chiều là 3 được dùng để mô phỏng. Hai biến chính tạo ra sự tách biệt trong số liệu là biến $\mathbf{x}_2$ và $\mathbf{x}_3$. Chúng ta mô phỏng biến ngẫu nhiên $\mathbf{x}_2$ với các phần tử thuộc 2 tập hợp X_{02} và X_{12}. Tương tự, biến ngẫu nhiên $\mathbf{x}_3$ được mô phỏng với các phần tử thuộc 2 tập hợp X_{03} và X_{13}. Ta xác định vector $\boldsymbol{\alpha}$ theo (2.6) như sau:

$$\begin{aligned}\boldsymbol{\alpha} &= (\alpha_1, \alpha_2, \alpha_3)\\ &= (-t_2[\max\{E_{02}\} + d_2/2]\\ &\quad - t_3[\max\{E_{03}\} + d_3/2], t_2, t_3),\end{aligned}$$

với t_2 và t_3 là các tham số điều chỉnh để làm tăng giảm xác suất xuất hiện của sự tách biệt trong số liệu mô phỏng. Theo Định lí 2, khi giá trị t_2 và t_3 càng lớn thì xác suất xuất hiện sự tách biệt càng lớn. Theo kết quả Định lý 2, thuật toán mô phỏng được đề xuất như sau:

Thuật toán 2:

Bước 1: Mô phỏng biến độc lập **x** với hai biến $\mathbf{x}_2$ và $\mathbf{x}_3$ là yếu tố tạo ra sự tách biệt trong số liệu.

Bước 2: Xác định vector alpha

$$\begin{aligned}\boldsymbol{\alpha} &= (\alpha_1, \alpha_2, \alpha_3)\\ &= (-t_2[\max\{E_{02}\} + d_2/2]\\ &\quad - t_3[\max\{E_{03}\} + d_3/2], t_2, t_3).\end{aligned}$$

Bước 3: Tính giá trị p_i, $i = 1, 2, ..., n$ theo công thức:

$$p_i = \frac{exp(\mathbf{x}_i^T \boldsymbol{\alpha})}{1 + exp(\mathbf{x}_i^T \boldsymbol{\alpha})}.$$

Bước 4: Mô phỏng biến $y_i \sim Bernoulli(p_i)$.

Bước 5: Kết hợp giữa biến phụ thuộc **y** và biến độc lập **x** tạo thành mẫu ngẫu nhiên cần tìm.

3.2.2. Kết quả mô phỏng

Ta chọn 2 tập rời nhau A_0 và A_1 sao cho $A_0 \bigcup A_1 = \{1, 2, ..., n\}$. Ta mô phỏng

$$x_2 \sim Uniform(a, b) \text{ và}$$

$$x_3 \sim Norm(\mu_0, \sigma_0^2).$$

Ta xác định 2 tập hợp:

$X_{02} = \{x_{i2}, i \in A_0\}$ và $X_{12} = \{x_{i2}, i \in A_1\}$.

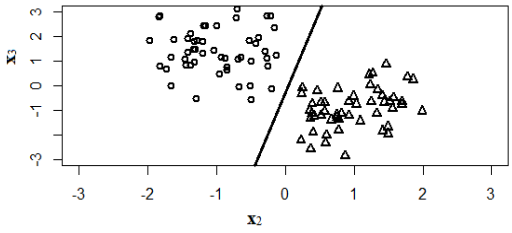
Tương tự

$X_{03} = \{x_{i3}, i \in A_0\}$ và $X_{13} = \{x_{i3}, i \in A_1\}$.

Ta tìm khoảng cách d_2 giữa 2 tập hợp X_{02} và X_{12}, khoảng cách d_3 giữa 2 tập hợp X_{03} và X_{13}. Kết quả mô phỏng được trình bày trong Bảng 2.

Bảng 2. Kết quả xác suất xuất hiện của sự tách biệt trong số liệu với các giá trị khoảng cách d_2, d_3 và giá trị điều chỉnh t_2 và t_3

d_2	d_3	t_2	t_3	**n**	**p**
2,0440	0,0082	0,5	0,5	0	0%
2,0314	0,0105	1	1	21	21%
2,0375	0,0153	2	2	93	93%
2,0512	0,0084	3	3	98	98%
2,0414	0,0129	5	5	100	100%



Hình 2. Hình vẽ của biến y theo 2 hai biến $\mathbf{x}_2$ và $\mathbf{x}_3$ với số liệu có tách biệt

Bảng 2 ta thấy với mô hình có hai biến tác động đến sự tách biệt trong số liệu, khoảng cách d_2, d_3 và tham số t_2, t_3 tác động đến xác suất p_i. Với khoảng cách d_2 gần bằng 2 và d_3 gần bằng 0,01 được xác định, giá trị tham số t_2, t_3 càng lớn dẫn đến xác suất xuất hiện sự tách biệt càng lớn. Cụ thể, với khi $t_2 = 0,5$ và $t_3 = 0,5$, trong mẫu mô phỏng không xuất hiện sự tách biệt. Khi $t_2 = 1$ và $t_3 = 1$, xác suất xuất hiện sự tách biệt là 21% và hơn nữa, khi $t_2 = 5$ và $t_3 = 5$, sự tách biệt hầu như xảy ra trong mọi mẫu ngẫu nhiên. Hình vẽ của biến **y** theo 2 hai biến $\mathbf{x}_2$ và $\mathbf{x}_3$ với số liệu có tách biệt được trình trong Hình 2. Với kí hiệu (o) là tại vị trí quan sát tại đó biến ngẫu nhiên **y** nhận giá trị 0 và kí hiệu (Δ) là tại vị trí quan sát tại đó biến ngẫu nhiên **y** nhận giá trị 1.

4. KẾT LUẬN

Khi sự tách biệt xuất hiện trong mô hình hồi quy logistic thì ước lượng các tham số trong mô hình sẽ không tồn tại giá trị ước lượng đối với thống kê tần suất và có ảnh hưởng nghiêm trọng đối với thống kê Bayes. Do đó, trong nghiên cứu mô phỏng, việc mô phỏng số liệu theo mô hình hồi quy logistic với cỡ mẫu và số chiều của biến độc lập tùy ý và kiểm soát được xác suất xuất hiện sự tách biệt trong số liệu là kết quả có ý nghĩa.

Trong bài báo này các thuật toán để mô phỏng số liệu theo mô hình hồi quy logistic được trình bày với sự kiểm soát xác suất xuất hiện sự tách biệt trong số liệu. Với cách chọn tham số cho phương trình hồi quy logistic trong thuật toán, sự tách biệt xuất hiện trong mẫu ngẫu nhiên được mô phỏng không phụ thuộc vào cỡ mẫu, giá trị và số chiều của biến độc lập trong mô hình.

Khi trung bình của phân phối hậu nghiệm trong phân tích Bayes tồn tại, tốc độ hội tụ về phân phối mục tiêu và độ lớn của các mẫu cần phải được nghiên cứu. Những yếu tố này có thể bị ảnh hưởng bởi các phân phối tiên nghiệm hoặc sự hiện diện của sự tách biệt trong quan sát. Kết quả của bài báo này sẽ hỗ trợ trong các nghiên cứu tương lai về tốc độ hội tụ của phân phối hậu nghiệm dưới sự lựa chọn khác nhau của các phân phối tiên nghiệm và các kiểu dữ liệu khác nhau trong mô hình hồi quy logistic.

TÀI LIỆU THAM KHẢO

Albert, A., & Anderson. J. A. (1984*). On the existence of maximum likelihood estimates in logistic regression models*. *Biometrika, 71*(1), 1–10. https://doi.org/10.1093/biomet/71.1.1

Allison, P. D. (2008). Convergence failures in logistic regression. *SAS Global Forum, 360*(1), 11.

Atkinson, A. C, & Woods, D. C. (2015). Designs for generalized linear models. *Handbook of Design and Analysis of Experiments, (7),* 471–514.

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. (2008). A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics, 2*(4), 1360–1383. https://doi.org/10.1214/08-AOAS191

Heinze, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in medicine*, *25*(24), 4216–4226. https://doi.org/10.1002/sim.2687

Huong, P. T. T., & Hoa, P. (2021). On the existence of posterior mean for bayesian logistic regression. *Monte Carlo Methods and Applications*, *7*(3277-288), 277–288. https://doi.org/10.1515/mcma-2021-2089

Ghosh, Y., Li, J., & Mitra, R. (2018). On the use of cauchy prior distributions for bayesian logistic regression. *Bayesian Analysis*, *13*(2), 359–383. https://doi.org/10.1214/17-BA1051

Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using polya-gamma latent variables. *Journal of the American statistical Association*, *108*(504), 1339–1349. https://doi.org/10.1080/01621459.2013.829001

Speckman, P. L., Lee, J., & Sun, D. (2009). Existence of the mle and propriety of posteriors for a general multinomial choice model. *Statistica Sinica*, 731–748.

Wakefield, J. (2013). Bayesian and Frequentist Regression Methods. *Springer Science & Business Media, New York*. https://doi.org/10.1007/978-1-4419-0925-1