



DOI:10.22144/ctu.jos.2026.180

KIỂM ĐỊNH KHI-BÌNH PHƯƠNG PEARSON CHO MÔ HÌNH HỒI QUY LOGISTIC CÓ HIỆP BIẾN THIẾU DỮ LIỆU NGẪU NHIÊN VÀ ỨNG DỤNG VÀO DỮ LIỆU TITANIC

Ngô Thảo Uyên^{1,2}, Dương Thị Tuyền¹ và Trần Phước Lộc^{1,*}

¹Trường Khoa học Tự nhiên, Đại học Cần Thơ, Việt Nam

²Trung tâm FPT Polytechnic Cần Thơ, Trường Đại học FPT Cần Thơ, Việt Nam

*Tác giả liên hệ (Corresponding author): tploc@ctu.edu.vn

Thông tin chung (Article Information)

Nhận bài (Received): 03/04/2026

Sửa bài (Revised): 25/04/2026

Duyệt đăng (Accepted): 22/08/2026

Title: Pearson's Chi-squared test for logistic regression model with covariates missing at random and application to the Titanic dataset

Author(s): Ngo Thao Uyen^{1,2}, Duong Thi Tuyen¹ and Tran Phuoc Loc^{1,*}

Affiliation(s): ¹College of Natural Sciences, Can Tho University, Viet Nam; ²FPT Polytechnic Can Tho Center, FPT University Can Tho, Viet Nam

TÓM TẮT

Nghiên cứu này được thực hiện nhằm đề xuất kiểm định Khi-bình phương Pearson (PCS) cho mô hình hồi quy logistic khi hiệp biến bị thiếu dữ liệu ngẫu nhiên. Một thống kê kiểm định có trọng số (*W*-Pearson) được xây dựng dựa trên phương pháp ước lượng trọng số xác suất nghịch đảo bán tham số (*SIPW*) nhằm hiệu chỉnh sai lệch do dữ liệu thiếu và đánh giá độ phù hợp (*GOF*) của mô hình hiệu quả hơn so với phương pháp sử dụng dữ liệu đầy đủ (*CC*). Kết quả lý thuyết cho thấy thống kê kiểm định có phân phối tiệm cận chuẩn sau khi chuẩn hóa. Việc mô phỏng cho thấy *W*-Pearson kiểm soát tốt sai lầm loại I và có công suất cao hơn *C*-Pearson, đặc biệt khi tỉ lệ thiếu lớn hoặc mô hình bị sai lệch, đồng thời tiệm cận với kiểm định Pearson khi dữ liệu đầy đủ. Bên cạnh đó, việc ứng dụng trên bộ dữ liệu Titanic cho thấy phương pháp *SIPW* cải thiện hiệu quả ước lượng và độ phù hợp của mô hình.

Từ khóa: Dữ liệu thiếu, hồi quy logistic, kiểm định độ phù hợp, kiểm định Pearson, trọng số xác suất nghịch đảo

ABSTRACT

This study proposes a Pearson's Chi-square (PCS) goodness-of-fit (*GOF*) test for a logistic regression model with covariates missing at random. A weighted test statistic (*W*-Pearson) is developed based on semiparametric inverse probability weighting (*SIPW*) to correct for bias induced by missing data and to provide a more reliable assessment of model adequacy compared to the complete case (*CC*) approach. Theoretical results show that the standardized test statistic is asymptotically normal. Simulation studies demonstrate that the *W*-Pearson test effectively controls type I error and achieves higher power than the *C*-Pearson test, particularly under high missing rates or model misspecification, while approaching the performance of the full-data Pearson test when no missingness is present. An application to the Titanic dataset indicates that the *SIPW* method improves estimation efficiency and model fit.

Keywords: Goodness-of-fit test, inverse probability weighting, logistic regression, missing data, Pearson Chi-square test

1. GIỚI THIỆU

Hồi quy logistic là một trong những phương pháp phổ biến trong phân tích dữ liệu có biến phân nhị phân. Phương pháp này được sử dụng rộng rãi trong nhiều lĩnh vực như y sinh học, khoa học xã hội, kinh tế và khoa học dữ liệu nhằm mô tả mối quan hệ giữa xác suất xảy ra của một biến phân hồi với các biến giải thích (Hosmer et al., 2013; Agresti, 2015). Trong mô hình hồi quy logistic, xác suất xảy ra của biến phân hồi thường được mô tả thông qua hàm liên kết logistic, cho phép ước lượng ảnh hưởng của các hiệp biến lên xác suất của sự kiện quan tâm.

Trong thực tế, dữ liệu thu thập được thường gặp tình trạng thiếu dữ liệu ở một số hiệp biến. Hiện tượng này có thể xuất phát từ nhiều nguyên nhân như lỗi ghi nhận, mất thông tin trong quá trình khảo sát hoặc người tham gia từ chối cung cấp thông tin. Rubin (1976) phân loại cơ chế thiếu dữ liệu thành ba nhóm chính gồm: thiếu hoàn toàn ngẫu nhiên (MCAR), thiếu ngẫu nhiên phụ thuộc vào dữ liệu quan sát được (MAR) và thiếu không ngẫu nhiên (NMAR). Trong đó, cơ chế thiếu dữ liệu MAR được xem là giả định hợp lý trong nhiều nghiên cứu thực nghiệm và là cơ sở cho nhiều phương pháp xử lý dữ liệu thiếu.

Khi tồn tại hiệp biến thiếu dữ liệu dưới giả định MAR, việc ước lượng và đánh giá mô hình hồi quy logistic trở nên phức tạp hơn so với trường hợp dữ liệu đầy đủ. Một số phương pháp đơn giản và thường được sử dụng để xử lý dữ liệu thiếu bao gồm phương pháp *sử dụng dữ liệu đầy đủ* (CC); trong đó, chỉ các quan sát có đầy đủ thông tin, hoặc các phương pháp *trọng số xác suất nghịch đảo* (IPW) được sử dụng nhằm hiệu chỉnh sai lệch do dữ liệu thiếu (Little & Rubin, 2019) và *IPW bán tham số* (SIPW, Wang et al., 1997; Wang & Wang, 2001), hoặc các phương pháp *thay thế lặp* (MI) nhằm điền những giá trị thích hợp vào các ô dữ liệu thiếu để tạo các bộ dữ liệu đầy đủ cho tính toán (Rubin, 1996; Lee et al., 2023; Tran et al., 2025), v.v.

Bên cạnh vấn đề ước lượng tham số, một bước quan trọng trong phân tích hồi quy logistic là đánh giá mức độ phù hợp của mô hình (goodness-of-fit – GOF). Kiểm định GOF cho phép đánh giá xem mô hình giả định có mô tả phù hợp mối quan hệ giữa biến phân hồi và các hiệp biến trong dữ liệu hay không. Nếu mô hình được chỉ định không đúng thì các kết luận suy ra từ mô hình có thể không đáng tin cậy. Trong trường hợp dữ liệu đầy đủ, nhiều phương pháp kiểm định GOF đã được phát triển cho mô hình hồi quy logistic, trong đó các kiểm định dựa trên

thống kê Khi-bình phương Pearson (PCS) là một trong những phương pháp được sử dụng phổ biến nhờ tính đơn giản và khả năng diễn giải rõ ràng (Hosmer et al., 2013). Tuy nhiên, khi tồn tại hiệp biến bị thiếu dữ liệu, đặc biệt dưới cơ chế MAR, việc xây dựng các kiểm định GOF thích hợp trở nên khó khăn hơn do cần phải xét đến cơ chế thiếu dữ liệu trong quá trình suy luận thống kê.

Gần đây, các phương pháp kiểm định GOF cho mô hình hồi quy logistic trong trường hợp có hiệp biến thiếu dữ liệu đã được đề xuất trong một số nghiên cứu. Các phương pháp này thường dựa trên việc kết hợp các kỹ thuật xử lý dữ liệu thiếu với các thống kê kiểm định truyền thống, nhằm xây dựng các thống kê kiểm định có tính chất tiệm cận tốt trong trường hợp cỡ mẫu lớn (Lee et al., 2022; Noughabi, 2025). Tuy nhiên, việc trình bày và áp dụng các kiểm định này trong bối cảnh dữ liệu thực tế vẫn còn hạn chế, đặc biệt trong các nghiên cứu minh họa bằng dữ liệu thực nghiệm.

Bên cạnh các nghiên cứu phương pháp luận, việc áp dụng các phương pháp thống kê vào các bộ dữ liệu thực tế đóng vai trò quan trọng trong việc minh họa và đánh giá hiệu quả của các phương pháp đề xuất. Trong lĩnh vực khoa học dữ liệu và học máy, bộ dữ liệu Titanic được xem là một bộ dữ liệu kinh điển thường được sử dụng để minh họa các mô hình phân loại nhị phân. Bộ dữ liệu này ghi nhận thông tin về các hành khách trên chuyến tàu RMS Titanic, bao gồm các biến như giới tính, hạng vé, tuổi, số lượng người thân đi cùng và tình trạng sống sót sau tai nạn. Một đặc điểm đáng chú ý của bộ dữ liệu này là biến tuổi (Age) có một tỉ lệ thiếu dữ liệu đáng kể, tạo ra một bối cảnh phù hợp để nghiên cứu các phương pháp phân tích hồi quy logistic với hiệp biến thiếu dữ liệu.

Xuất phát từ những vấn đề nêu trên, nghiên cứu này tập trung vào việc xây dựng và trình bày kiểm định PCS cho mô hình hồi quy logistic khi tồn tại hiệp biến thiếu dữ liệu dưới cơ chế MAR. Cụ thể, mô hình hồi quy logistic với hiệp biến bị thiếu dữ liệu đã được xem xét và phương pháp ước lượng tham số SIPW cũng đã được trình bày. Trên cơ sở đó, việc xây dựng thống kê kiểm định PCS đã được tiến hành để đánh giá mức độ phù hợp của mô hình và nghiên cứu các tính chất tiệm cận của kiểm định trong trường hợp cỡ mẫu lớn. Ngoài ra, thông qua các nghiên cứu mô phỏng và một ví dụ minh họa với bộ dữ liệu Titanic, nghiên cứu được thực hiện nhằm đánh giá hiệu quả của phương pháp được đề xuất trong thực hành phân tích dữ liệu.

2. PHƯƠNG PHÁP NGHIÊN CỨU

2.1. Mô hình

Giả sử Y là biến phản hồi nhị phân nhận giá trị 0 hoặc 1, ta quan tâm đến việc mô tả xác suất xảy ra của sự kiện $Y = 1$ thông qua một mô hình hồi quy logistic với các hiệp biến X và Z . Giả sử rằng ta có một mẫu độc lập gồm n quan sát (Y_i, X_i, Z_i) , $i = 1, 2, \dots, n$, trong đó $X_i = (X_{i1}, X_{i2}, \dots, X_{ir})^T$ có thể chứa giá trị thiếu và $Z_i = (Z_{i1}, Z_{i2}, \dots, Z_{is})^T$ là các biến rời rạc và luôn quan sát được, với $r + s = p$. Mô hình hồi quy logistic được định nghĩa

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

Trong đó $\beta = (\beta_0, \beta_1^T, \beta_2^T)^T$ với $\beta_1 \in \mathbb{R}^r, \beta_2 \in \mathbb{R}^s$ là các vector hệ số của X và Z , $\mathbf{X} = (1, X^T, Z^T)^T$ và $H(u) = (1 + e^{-u})^{-1}$ là hàm logistic. Khi dữ liệu đầy đủ, ước lượng $\hat{\beta}_F$ của β là nghiệm của hệ

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

với

$$\phi_i(\beta) = \mathbf{X}_i[Y_i - H(\beta^T \mathbf{X}_i)]. \quad (2)$$

Tuy nhiên, khi X_i bị thiếu đối với một số quan sát, vector hiệp biến X_i không thể được xác định đầy đủ. Khi đó, gọi δ là biến chỉ báo cho trạng thái quan sát của X , được định nghĩa bởi $\delta_i = 1$ nếu X_i được quan sát và $\delta_i = 0$ nếu X_i bị thiếu. Cơ chế thiếu dữ liệu của X được giả định là MAR. Giả sử W là biến rời rạc có tương quan với X nhưng độc lập với Y . Ta gọi W là biến thay thế cho X . Mô hình xác suất chọn quy ước cơ chế thiếu MAR của X (Lee et al., 2022; Lộc và ctv., 2024) được định nghĩa

$$P(\delta_i = 1|Y_i, X_i, Z_i, W_i) = P(\delta_i = 1|Y_i, Z_i, W_i) = \pi(Y_i, V_i), \quad (3)$$

trong đó $V_i = (Z_i^T, W_i^T)^T$. Giả định này cho phép xây dựng các phương pháp ước lượng và kiểm định GOF trong trường hợp tồn tại dữ liệu thiếu.

2.2. Phương pháp ước lượng

2.2.1. Phương pháp CC

Phương pháp CC là phương pháp đơn giản, trong đó chỉ sử dụng các quan sát có đầy đủ thông tin về hiệp biến X . Khi đó, phương trình ước lượng của tham số β có dạng

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \delta_i \phi_i(\beta) = 0. \quad (4)$$

Nghiệm của phương trình (4) cho ta ước lượng $\hat{\beta}_C$. Phương pháp CC có ưu điểm là đơn giản và dễ thực hiện khi chỉ sử dụng các quan sát đầy đủ. Tuy nhiên, việc loại bỏ các quan sát có dữ liệu thiếu có thể làm giảm kích thước mẫu và độ chính xác của ước lượng, đặc biệt khi tỉ lệ thiếu lớn. Ngoài ra, phương pháp này có thể gây sai lệch ước lượng dưới cơ chế MAR (Lee et al., 2023; Tran et al., 2025).

2.2.2. Phương pháp SIPW

Để khắc phục hạn chế của phương pháp CC, các phương pháp dựa trên trọng số đã được đề xuất nhằm tận dụng thông tin từ dữ liệu quan sát được. Một trong những phương pháp phổ biến là SIPW, trong đó mỗi quan sát được gán một trọng số tỉ lệ nghịch với xác suất quan sát của hiệp biến bị thiếu (Wang et al., 1997; Wang & Wang, 2001; Lee et al., 2023; Tran et al., 2023). Các trọng số này được ổn định nhằm giảm phương sai của các ước lượng. Phương trình ước lượng của tham số β khi đó có dạng:

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

trong đó, $\hat{\pi}_i = \hat{\pi}(Y_i, V_i)$ là một ước lượng của $\pi_i = \pi(Y_i, V_i)$ trong biểu thức (3) và được định nghĩa dạng

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

với $I(\cdot)$ là một hàm chỉ số.

Ước lượng $\hat{\beta}_W$ của β từ phương trình (5) là một ước lượng nhất quán (consistent) dưới các điều kiện chuẩn (Wang & Wang, 2001), có trung bình 0 và phương sai được ước lượng bởi:

$$\hat{\Delta}_W = G^{-1}(\hat{\beta}_W, \hat{\pi}) J(\hat{\beta}_W, \hat{\pi}) [G^{-1}(\hat{\beta}_W, \hat{\pi})]^T, \quad (7)$$

trong đó:

$$G(\hat{\beta}_W, \hat{\pi}) = \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i} \mathbf{X}_i^{\otimes 2} H^{(1)}(\hat{\beta}_W^T \mathbf{X}_i),$$

$$J(\hat{\beta}_W, \hat{\pi}) = \frac{1}{n} \sum_{i=1}^n \left[\frac{\delta_i}{\hat{\pi}_i} \phi_i(\hat{\beta}_W) + \left(1 - \frac{\delta_i}{\hat{\pi}_i}\right) \hat{\phi}_i^*(\hat{\beta}_W) \right],$$

với $u^{\otimes 2} = uu^T$ với mọi vector cột u và $\hat{\phi}_i^*(\hat{\beta}_W)$ là một ước lượng của $\phi_i^*(\hat{\beta}_W) = E[\phi_1(\hat{\beta}_W)|Y_1 = Y_i, V_1 = V_i]$, được xác định bởi

$$\hat{\phi}_i^*(\hat{\beta}_W) = \frac{\sum_{j=1}^n \delta_j \phi_i(\hat{\beta}_W) I(Y_j = Y_i, V_j = V_i)}{\sum_{k=1}^n \delta_k I(Y_k = Y_i, V_k = V_i)}. \quad (8)$$

2.3. Kiểm định PCS

Sau khi ước lượng các tham số của mô hình hồi quy logistic, một bước quan trọng trong phân tích là đánh giá mức độ phù hợp của mô hình. Kiểm định GOF được thực hiện nhằm kiểm tra xem dạng hàm của mô hình logistic giả định có mô tả đúng mối quan hệ giữa biến phản hồi và các hiệp biến trong dữ liệu hay không. Nếu mô hình bị chỉ định sai, các suy luận thống kê và dự báo từ mô hình có thể không đáng tin cậy. Kiểm định PCS là một trong những kiểm định GOF phổ biến cho mô hình hồi quy logistic (Hosmer et al., 2013). Để thuận tiện cho việc suy luận tiệm cận, ta xét thống kê PCS dạng trung tâm (Lee et al., 2023) sau

$$G_{P,i}(\beta) = \frac{[Y_i - H(\beta^T \mathbf{X}_i)]^2}{H^{(1)}(\beta^T \mathbf{X}_i)} - 1. \quad (10)$$

Khi dữ liệu đầy đủ, thống kê PCS, gọi là F-Pearson, được định nghĩa là

$$G_{F,P}(\beta) = \sum_{i=1}^n G_{P,i}(\beta_F). \quad (11)$$

Khi X bị thiếu dữ liệu, đặc biệt dưới giả định MAR, thống kê kiểm định cần được điều chỉnh để phản ánh đúng cơ chế thiếu dữ liệu và phương pháp ước lượng tham số được sử dụng. Thống kê PCS cho ước lượng từ phương pháp CC và SIPW, gọi là C-Pearson và W-Pearson, lần lượt được định nghĩa dạng

$$G_{C,P}(\hat{\beta}_C) = \sum_{i=1}^n \delta_i G_{P,i}(\hat{\beta}_C), \quad (11)$$

$$G_{W,P}(\hat{\beta}_W, \hat{\pi}) = \sum_{i=1}^n \frac{\delta_i}{\hat{\pi}_i} G_{P,i}(\hat{\beta}_W). \quad (12)$$

So với thống kê $G_{C,P}(\hat{\beta}_C)$, thống kê $G_{W,P}(\hat{\beta}_W, \hat{\pi})$ sử dụng trọng số $\delta_i/\hat{\pi}_i$ nhằm hiệu chỉnh xác suất quan sát của hiệp biến bị thiếu, qua đó tận dụng thông tin từ các biến quan sát khác trong mẫu. Nhờ đó, W-Pearson có khả năng giảm sai lệch do cơ chế thiếu dữ liệu MAR và cải thiện hiệu quả thống kê so với phương pháp C-Pearson. Điều này đặc biệt quan trọng trong trường hợp tỉ lệ thiếu dữ liệu lớn hoặc phụ thuộc mạnh vào các biến quan sát được. Mục tiêu của bài toán kiểm định là kiểm tra giả thiết H_0 ở mức ý nghĩa α :

$$\begin{aligned} H_0: E(Y_i | \mathbf{X}_i) &= H(\beta^T \mathbf{X}_i), \\ H_1: E(Y_i | \mathbf{X}_i) &\neq H(\beta^T \mathbf{X}_i). \end{aligned} \quad (13)$$

2.4. Tính chất cỡ mẫu lớn

Trong phần này, ta nghiên cứu phân phối tiệm cận của thống kê kiểm định $G_{W,P}(\hat{\beta}_W, \hat{\pi})$ trong mô hình hồi quy logistic khi X thiếu dữ liệu theo cơ chế MAR. Kết quả này cho phép xây dựng quy tắc suy luận thống kê cho kiểm định GOF khi cỡ mẫu lớn.

Định lý 1. Giả sử các điều kiện trong nghiên cứu của Lee et al. (2023) cho mô hình hồi quy logistic (1) được thỏa mãn và mô hình logistic này được chỉ định dưới giả thiết H_0 . Khi kích thước mẫu lớn, hay $n \rightarrow \infty$, ta có kết quả sau

$$\frac{1}{\sqrt{n}} G_{P,W}(\hat{\beta}_W, \hat{\pi}) \xrightarrow{d} \mathcal{N}(0, \Delta_{P,W}(\beta, \pi)). \quad (14)$$

Trong đó,

$$\Delta_{P,W}(\beta, \pi) = \text{Var}(K_{P,W,1}(\beta, \pi_1)), \quad (15)$$

với

$$\begin{aligned} K_{P,W,i}(\beta, \pi) &= \frac{\delta_i}{\pi_i} [G_{P,i}(\beta) - A_P^T(\beta) F^{-1}(\beta) \phi_i(\beta)] \\ &+ \left(1 - \frac{\delta_i}{\pi_i}\right) [G_{P,i}^*(\beta) - A_P^T(\beta) F^{-1}(\beta) \phi_i^*(\beta)], \end{aligned} \quad (16)$$

và

$$\begin{aligned} A_P(\beta) &= E\{[1 - 2H(\beta^T \mathbf{X}_1)]\mathbf{X}_1\}, \\ F(\beta) &= E[X_1^{\otimes 2} H^{(1)}(\beta^T \mathbf{X}_1)], \\ G_{P,i}^*(\beta) &= E[G_{P,i}(\beta) | Y_1 = Y_i, V_1 = V_i]. \end{aligned} \quad (17)$$

Kết quả này cho thấy thống kê W-Pearson sau khi chuẩn hóa có phân phối tiệm cận chuẩn.

Chứng minh.

Khai triển Taylor của $\frac{1}{\sqrt{n}} G_{P,W}(\hat{\beta}_W, \hat{\pi})$ quanh β ta được

$$\begin{aligned} &G_{P,W}(\hat{\beta}_W, \hat{\pi}) \\ &= G_{P,W}(\beta, \pi) + \frac{\partial G_{P,W}(\beta, \pi)}{\partial \beta^T} (\hat{\beta}_W - \beta) \\ &+ \sum_{i=1}^n \frac{\partial G_{P,W}(\beta, \pi)}{\partial \pi_i} (\hat{\pi}_i - \pi_i) + o_p(1). \end{aligned}$$

Trong đó, kí hiệu $X_n = o_p(a_n)$ nghĩa là X_n/a_n bị chặn theo xác suất. Dưới giả thiết H_0 và theo luật yếu số lớn khi $n \rightarrow \infty$, ta có thể thu được các kết quả sau:

$$\begin{aligned} &\frac{\partial G_{P,W}(\beta, \pi)}{\partial \beta^T} \xrightarrow{p} -A_P(\beta), \\ &\sqrt{n}(\hat{\beta}_W - \beta) = F^{-1}(\beta) U_W(\beta, \hat{\pi}) + o_p(1), \end{aligned}$$

$$\begin{aligned} & \frac{\partial G_{P,W}(\beta, \pi)}{\partial \pi_i} (\hat{\pi}_i - \pi_i) \\ &= \left(1 - \frac{\delta_i}{\hat{\pi}_i}\right) G_{P,i}^*(\beta) + o_p(n^{-1}), \end{aligned}$$

và

$$\begin{aligned} & U_W(\beta, \hat{\pi}) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[\frac{\delta_i}{\hat{\pi}_i} \phi_i(\beta) + \left(1 - \frac{\delta_i}{\hat{\pi}_i}\right) \phi_i^*(\beta) \right] + o_p(1). \end{aligned}$$

Trong đó, $X_n = o_p(a_n)$ nghĩa là $X_n/a_n \xrightarrow{p} 0$. Từ đó, ta thu được:

$$\frac{1}{\sqrt{n}} G_{P,W}(\hat{\beta}_W, \hat{\pi}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n K_{P,W,i}(\beta, \pi) + o_p(1).$$

Mặt khác, theo định lý giới hạn trung tâm, ta có

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n K_{P,W,i}(\beta, \pi) \xrightarrow{d} \mathcal{N}(0, \Delta_{P,W}(\beta, \pi)).$$

Do đó, định lý Slutsky (Casella & Berger, 2002) được áp dụng, ta suy ra:

$$\frac{1}{\sqrt{n}} G_{P,W}(\hat{\beta}_W, \hat{\pi}) \xrightarrow{d} \mathcal{N}(0, \Delta_{P,W}(\beta, \pi)).$$

Điều này là do tổng của một đại lượng hội tụ theo phân phối và một đại lượng hội tụ về 0 theo xác suất sẽ hội tụ theo phân phối về cùng giới hạn. Định lý 1 đã được chứng minh xong.

Nhận xét 1. Trong thực hành, thống kê chuẩn hóa sau được sử dụng để tính p -value cho kiểm định PCS dưới giả thiết H_0

$$Z_{P,W} = \frac{1}{\sqrt{n}} \frac{G_{P,W}(\hat{\beta}_W, \hat{\pi})}{\sqrt{\hat{\Delta}_{P,W}(\hat{\beta}_W, \hat{\pi})}} \xrightarrow{d} \mathcal{N}(0,1), \quad (18)$$

khi $n \rightarrow \infty$, với $\hat{\Delta}_{P,W}(\hat{\beta}_W, \hat{\pi})$ là một ước lượng nhất quán của $\Delta_{P,W}(\beta, \pi)$. Khi đó, p -value của kiểm định được xác định bởi

$$p - \text{value} = 2[1 - \Phi(|Z_{P,W}|)], \quad (19)$$

với $\Phi(\cdot)$ là hàm phân phối tích lũy của $\mathcal{N}(0,1)$.

Nhận xét 2. Do $Z_{P,W} \xrightarrow{d} \mathcal{N}(0,1)$ nên bình phương của thống kê này thỏa

$$Z_{P,W}^2 \xrightarrow{d} \chi_1^2.$$

Trong đó, χ_1^2 là phân phối χ^2 với 1 bậc tự do. Vì vậy, một cách tương đương, p -value cũng có thể được tính như sau:

$$p_{\text{value}} = P(\chi_1^2 \geq Z_{P,W}^2) = 1 - F_{\chi_1^2}(Z_{P,W}^2), \quad (20)$$

với $F_{\chi_1^2}(\cdot)$ là hàm phân phối tích lũy của phân phối χ_1^2 .

Hai cách tính này là tương đương về mặt tiệm cận, bởi vì $P(|Z| \geq |z|) = P(Z^2 \geq z^2)$. Tuy nhiên, trong thực hành việc sử dụng phân phối χ^2 được hỗ trợ trực tiếp trong nhiều phần mềm thống kê. Trong bài báo này, công thức (19) được dùng để tính p -value cho các phương pháp nghiên cứu.

2.5. Thuật toán W-Pearson

Trong phần này, thuật toán kiểm định PCS cho mô hình hồi quy logistic khi hiệp biến bị thiếu dữ liệu dưới giả định MAR đã được trình bày, phương pháp ước lượng SIPW đã được sử dụng nhằm kiểm tra giả thiết H_0 trong (13).

- **Bước 1.** Thiết lập dữ liệu $(Y_i, X_i, Z_i, W_i, \delta_i)$, $i = 1, \dots, n$, và mô hình logistic dưới H_0 .
- **Bước 2.** Ước lượng hệ số mô hình hồi quy logistic theo SIPW:
 - i. Tìm ước lượng $\hat{\pi}_i$ theo công thức (6) và tính trọng số $\delta_i/\hat{\pi}_i$, $i = 1, \dots, n$.
 - ii. Tìm ước lượng $\hat{\beta}_W$ của β bằng cách giải hệ phương trình (5).
 - iii. Tìm ước lượng phương sai của $\hat{\beta}_W$ theo công thức (7).
- **Bước 3.** Tính thống kê W-Pearson:
 - i. Tính thống kê $G_{W,P}(\hat{\beta}_W, \hat{\pi})$ theo công thức (12).
 - ii. Tính phương sai $\Delta_{P,W}(\hat{\beta}_W, \hat{\pi})$ từ các công thức (15-17).
 - iii. Tính thống kê $Z_{P,W}$ theo công thức (18).
 - iv. Tính p -value theo công thức (19) và kết luận chấp nhận hay bác bỏ H_0 với mức ý nghĩa α cho trước.

3. MÔ PHÒNG

Để đánh giá hiệu năng của các kiểm định độ phù hợp được đề xuất trong các điều kiện dữ liệu khác nhau, một số thí nghiệm mô phỏng đã được tiến hành. Các kịch bản mô phỏng được thiết kế nhằm xem xét ảnh hưởng của kích thước mẫu, tỉ lệ thiếu dữ liệu và dạng sai lệch của mô hình đến mức ý nghĩa thực nghiệm và công suất kiểm định của các phương pháp F-Pearson (khi không có dữ liệu thiếu), C-Pearson và W-Pearson. Trong mỗi trường hợp, dữ liệu được sinh theo mô hình hồi quy logistic

với cơ chế thiếu dữ liệu MAR, và các kiểm định được áp dụng lặp lại 1.000 lần để ước lượng các đặc trưng thống kê tương ứng theo trung bình. Kết quả mô phỏng cung cấp cơ sở thực nghiệm để thực hiện việc so sánh và đánh giá tính ổn định cũng như hiệu quả của các phương pháp đề xuất trong các bối cảnh khác nhau.

3.1. Trường hợp 1

Trường hợp này được tiến hành nhằm đánh giá ảnh hưởng của kích thước mẫu đến hiệu năng của các kiểm định PCS trong điều kiện dữ liệu thiếu 40%. Ba cỡ mẫu $N = \{500; 1.000; 1.500\}$ đã được xét. Hai hiệp biến $X \sim U[-2; 2]$ và $Z \sim$

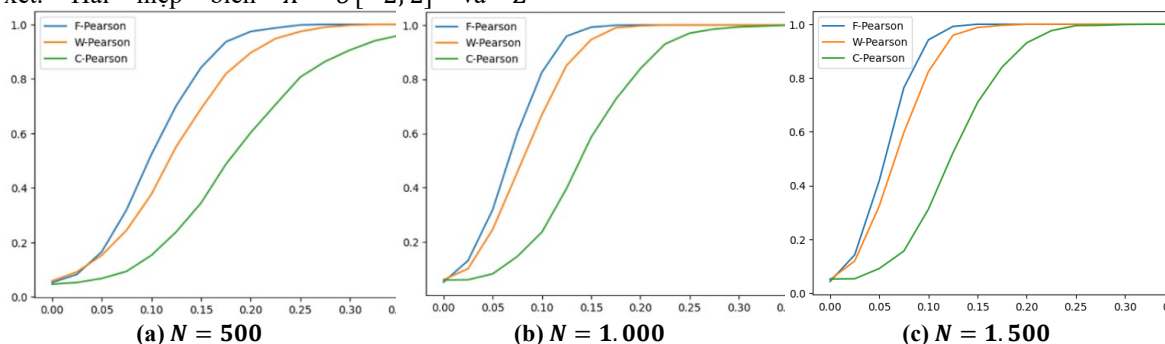
Bernoulli(0,5), trong đó X bị thiếu theo cơ chế MAR, còn Z luôn được quan sát. Biến phản hồi Y được sinh từ mô hình logistic có thành phần sai lệch $g(X; a)$:

$$P(Y = 1|X, Z) = H(\beta^T X + g(X; a)),$$

với $\beta = (\beta_0, \beta_1, \beta_2)^T = (0,3; 0,8; 0,05)^T$, $X = (1, X, Z)^T$ và $g(X; a) = a(X^4 - X^2)$, $a \geq 0$.

Các kiểm định F-Pearson, C-Pearson và W-Pearson được áp dụng để kiểm tra giả thuyết:

$$H_0 : E(Y |X, Z) = H(\beta^T X).$$



Hình 1. Hình mô tả công suất thực nghiệm của kiểm định PCS trường hợp 1

Ghi chú: Kích thước mẫu (N) lần lượt là 500, 1.000 và 1.500; tỉ lệ thiếu 40%

Bảng 1. Bảng tóm tắt kết quả mô phỏng của trường hợp 1

a	$N = 500$			$N = 1.000$			$N = 1.500$		
	F-Pearson	C-Pearson	W-Pearson	F-Pearson	C-Pearson	W-Pearson	F-Pearson	C-Pearson	W-Pearson
0,000	0,053	0,047	0,059	0,051	0,059	0,060	0,043	0,052	0,050
0,025	0,083	0,053	0,092	0,131	0,060	0,101	0,142	0,053	0,119
0,050	0,166	0,068	0,153	0,319	0,082	0,246	0,418	0,091	0,324
0,075	0,320	0,094	0,245	0,601	0,146	0,457	0,764	0,156	0,597
0,100	0,523	0,153	0,377	0,826	0,236	0,669	0,942	0,312	0,824
0,125	0,702	0,239	0,552	0,959	0,397	0,851	0,992	0,522	0,959
0,150	0,842	0,345	0,691	0,992	0,586	0,946	1,000	0,709	0,988
0,175	0,936	0,487	0,819	0,999	0,726	0,990	1,000	0,841	0,996
0,200	0,974	0,603	0,896	1,000	0,839	0,997	1,000	0,931	1,000
0,225	0,987	0,706	0,948	1,000	0,930	1,000	1,000	0,976	1,000
0,250	0,998	0,807	0,974	1,000	0,970	1,000	1,000	0,995	1,000
0,275	1,000	0,864	0,990	1,000	0,985	1,000	1,000	0,996	1,000
0,300	1,000	0,906	0,997	1,000	0,993	1,000	1,000	0,999	1,000
0,325	1,000	0,940	1,000	1,000	0,996	1,000	1,000	1,000	1,000
0,350	1,000	0,960	1,000	1,000	0,999	1,000	1,000	1,000	1,000

Ghi chú: Kích thước mẫu (N) lần lượt là 500, 1.000 và 1.500; tỉ lệ thiếu 40%

Khi $a = 0$, mô hình là đúng (giả thuyết H_0); khi $a \neq 0$, mô hình bị sai lệch.

Bên cạnh đó, để mô hình hóa cơ chế thiếu dữ liệu, ta xây dựng biến phụ W từ X như sau:

$$W = \begin{cases} 1, & \text{nếu } X \geq 0 \\ 0, & \text{nếu } X < 0. \end{cases}$$

Biến W được giả định là quan sát đầy đủ và được sử dụng trong việc mô hình hóa xác suất quan sát của X theo cơ chế MAR. Ngoài ra, để sinh biến chỉ

báo thiếu δ sao cho X bị thiếu theo cơ chế MAR, mô hình xác suất sau đây được sử dụng:

$$P(\delta = 1|Y, Z, W) = H(\gamma_0 + \gamma_1 Y + \gamma_2 Z + \gamma_3 W), \quad (21) \text{ với } \gamma = (\gamma_0, \gamma_1, \gamma_2, \gamma_3)^T = (0,75; -0,7; -0,4; 0,5)^T.$$

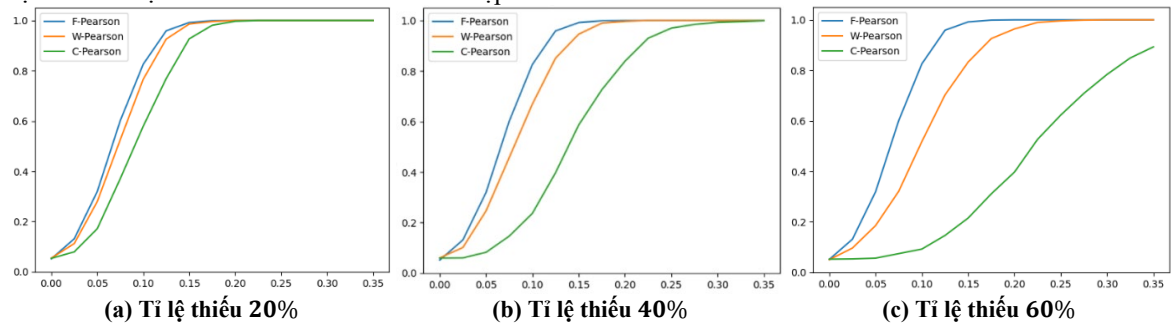
Kết quả trong Bảng 1 và Hình 1 cho thấy khi $a = 0$, các kiểm định đều kiểm soát tốt sai lầm loại I với mức ý nghĩa xấp xỉ $\alpha = 0,05$. Khi a tăng, công suất kiểm định của cả ba phương pháp đều tăng, phản ánh khả năng phát hiện sai lệch mô hình. Đáng chú ý, F-Pearson và W-Pearson đạt công suất cao nhanh hơn so với C-Pearson, đặc biệt ở cỡ mẫu nhỏ và trung bình. Khi kích thước mẫu tăng, hiệu năng của tất cả các kiểm định được cải thiện và sự khác biệt giữa các phương pháp giảm dần.

3.2. Trường hợp 2

Trường hợp này được thực hiện nhằm khảo sát ảnh hưởng của tỉ lệ thiếu dữ liệu đến hiệu năng kiểm định khi cố định cỡ mẫu $N = 1.000$. Hai hiệp biến

X và Z cùng với biến phụ W được tạo ra tương tự trong Trường hợp 1. Trong đó, Z luôn được quan sát và X bị thiếu theo cơ chế MAR với ba mức tỉ lệ thiếu 20%, 40% và 60%, dưới mô hình xác suất (21) có γ lần lượt là $(-1,1; -0,7; -0,4; 0,5)^T$, $(0,75; -0,7; -0,4; 0,5)^T$ và $(1,8; -0,7; -0,4; 0,5)^T$. Biến phản hồi Y được sinh từ mô hình logistic có thành phần sai lệch $g(X; a)$ và việc kiểm định giả thuyết $H_0: a = 0$ cũng được xác định tương tự trường hợp 1.

Bảng 2 và Hình 2 cho thấy khi tỉ lệ thiếu thấp (20%), ba kiểm định có hiệu năng khá gần nhau, trong đó F-Pearson và C-Pearson lần lượt là tốt nhất và thấp nhất. Tuy nhiên, khi tỉ lệ thiếu tăng, sự khác biệt trở nên rõ rệt: F-Pearson và W-Pearson vẫn duy trì công suất cao và tăng nhanh theo a , trong khi C-Pearson suy giảm đáng kể và tăng chậm hơn. Điều này cho thấy phương pháp W-Pearson ổn định hơn C-Pearson trong bối cảnh dữ liệu thiếu nhiều, nhờ cơ chế hiệu chỉnh trọng số.



Hình 2. Hình mô tả công suất thực nghiệm kiểm định PCS trường hợp 2

Ghi chú: Tỉ lệ thiếu 20%, 40% và 60%; Kích thước mẫu (N) 1.000

Bảng 2. Bảng tóm tắt kết quả mô phỏng của trường hợp 2

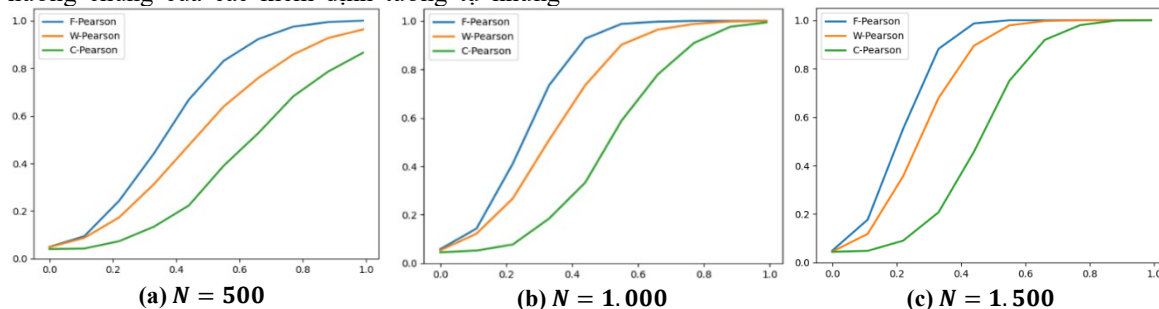
a	$MR \approx 20\%$			$MR \approx 40\%$			$MR \approx 60\%$		
	F-Pearson	C-Pearson	W-Pearson	F-Pearson	C-Pearson	W-Pearson	F-Pearson	C-Pearson	W-Pearson
0,000	0,051	0,053	0,054	0,051	0,059	0,060	0,051	0,059	0,058
0,025	0,131	0,079	0,112	0,131	0,060	0,101	0,131	0,059	0,096
0,050	0,319	0,171	0,279	0,319	0,082	0,246	0,319	0,056	0,185
0,075	0,601	0,370	0,525	0,601	0,146	0,457	0,601	0,074	0,321
0,100	0,826	0,579	0,766	0,826	0,236	0,669	0,826	0,092	0,518
0,125	0,959	0,769	0,926	0,959	0,397	0,851	0,959	0,146	0,703
0,150	0,992	0,927	0,986	0,992	0,586	0,946	0,992	0,215	0,833
0,175	0,999	0,981	0,996	0,999	0,726	0,990	0,999	0,311	0,926
0,200	1,000	0,997	1,000	1,000	0,839	0,997	1,000	0,397	0,964
0,225	1,000	1,000	1,000	1,000	0,930	1,000	1,000	0,527	0,990
0,250	1,000	1,000	1,000	1,000	0,970	1,000	1,000	0,623	0,996
0,275	1,000	1,000	1,000	1,000	0,985	1,000	1,000	0,709	0,999
0,300	1,000	1,000	1,000	1,000	0,993	1,000	1,000	0,784	1,000
0,325	1,000	1,000	1,000	1,000	0,996	1,000	1,000	0,849	1,000
0,350	1,000	1,000	1,000	1,000	0,999	1,000	1,000	0,893	1,000

Ghi chú: Tỉ lệ thiếu 20%, 40% và 60%; Kích thước mẫu (N) 1.000

3.3. Trường hợp 3

Trường hợp này được tiến hành nhằm đánh giá hiệu năng kiểm định khi dạng sai lệch của mô hình thay đổi phi tuyến với ba cỡ mẫu $N = 500, 1.000$ và 1.500 . Cụ thể, biến X được sinh từ phân phối đều trên khoảng $[-2; 2]$. Biến nhị phân Z luôn quan sát được và được sinh từ phân phối Bernoulli với xác suất thành công bằng 0.4 . Biến Y được tạo ra từ mô hình $P(Y = 1|X, Z) = H(\beta^T X + g(X; a))$, với $X = (1, X, Z)^T$, $\beta = (0,3; 0,81; 0,05)^T$ và $g(X; a) = a(0,5 - \cos 2X)$, $a \geq 0$. Tỷ lệ thiếu của X khoảng 50% thu được từ mô hình xác suất (21) với $\gamma = (0,25; -0,7; -0,4; 0,5)^T$.

Kết quả trong Bảng 3 và Hình 3 cho thấy xu hướng chung của các kiểm định tương tự nhưng



Hình 3. Hình mô tả công suất thực nghiệm của kiểm định PCS trường hợp 3

Ghi chú: Kích thước mẫu (N) lần lượt là 500, 1.000 và 1.500; tỉ lệ thiếu 50%

Bảng 3. Bảng tóm tắt kết quả mô phỏng của trường hợp 3

a	$N = 500$			$N = 1.000$			$N = 1.500$		
	F-Pearson	C-Pearson	W-Pearson	F-Pearson	C-Pearson	W-Pearson	F-Pearson	C-Pearson	W-Pearson
0,00	0,058	0,045	0,053	0,051	0,059	0,060	0,051	0,059	0,058
0,11	0,143	0,052	0,122	0,131	0,060	0,101	0,131	0,059	0,096
0,22	0,411	0,077	0,267	0,319	0,082	0,246	0,319	0,056	0,185
0,33	0,734	0,184	0,508	0,601	0,146	0,457	0,601	0,074	0,321
0,44	0,927	0,333	0,735	0,826	0,236	0,669	0,826	0,092	0,518
0,55	0,988	0,588	0,902	0,959	0,397	0,851	0,959	0,146	0,703
0,66	0,997	0,779	0,964	0,992	0,586	0,946	0,992	0,215	0,833
0,77	1,000	0,909	0,988	0,999	0,726	0,990	0,999	0,311	0,926
0,88	1,000	0,976	0,998	1,000	0,839	0,997	1,000	0,397	0,964
0,99	1,000	0,994	1,000	1,000	0,930	1,000	1,000	0,527	0,990

Ghi chú: Kích thước mẫu (N) lần lượt là 500, 1.000 và 1.500; tỉ lệ thiếu 50%

4. ỨNG DỤNG

4.1. Giới thiệu

Mục tiêu của phần này là minh họa khả năng ứng dụng của phương pháp đề xuất trong bối cảnh dữ liệu thực tế có hiệp biến bị thiếu. Thông qua bộ dữ liệu Titanic, hiệu quả của các phương pháp ước

đường cong công suất hội tụ về 1 chậm hơn trong Trường hợp 1. Khi $a = 0$, cả ba kiểm định đều kiểm soát tốt sai lầm loại I với mức ý nghĩa xấp xỉ $\alpha = 5\%$. Khi a tăng, công suất kiểm định của F-Pearson, C-Pearson và W-Pearson đều tăng dần, phản ánh khả năng phát hiện sự sai lệch của mô hình. Tuy nhiên, so với trường hợp 1, tốc độ gia tăng công suất chậm hơn đáng kể, đặc biệt ở các mức sai lệch nhỏ và trung bình. Điều này có thể được giải thích bởi dạng sai lệch phi tuyến phức tạp của hàm lượng giác $g(X; a) = a(0,5 - \cos 2X)$, khiến mức độ sai lệch khỏi mô hình tuyến tính khó được phát hiện hơn. Trong tất cả các kịch bản, W-Pearson vẫn cho hiệu năng tốt hơn và đạt công suất cao sớm hơn so với C-Pearson, trong khi W-Pearson tiếp tục cho kết quả tiệm cận với F-Pearson khi kích thước mẫu tăng.

lượng đã được đánh giá và việc kiểm định độ phù hợp của mô hình trong điều kiện dữ liệu không đầy đủ đã được tiến hành, từ đó làm rõ tính thực tiễn của kiểm định PCS đã đề xuất.

Bộ dữ liệu Titanic (Nguồn: Kaggle) bao gồm 891 quan sát về hành khách trên chuyến tàu RMS Titanic, với các thông tin nhân khẩu học và điều kiện

di chuyển, cùng với trạng thái sống sót. Đây là một bài toán phân loại nhị phân điển hình, trong đó xác suất sống sót không mang tính ngẫu nhiên mà phụ thuộc vào các đặc trưng của hành khách. Đáng chú ý, biến tuổi (Age) có chứa nhiều giá trị bị thiếu, tạo ra bối cảnh phù hợp để nghiên cứu mô hình hồi quy logistic với dữ liệu thiếu.

Để xây dựng mô hình hồi quy logistic, các biến được định nghĩa sau đây. Biến phản hồi nhị phân Y biểu thị trạng thái sống sót (Survived) của hành khách với $Y = 1$ nếu sống sót và $Y = 0$ nếu tử vong. Hiệp biến chính X là tuổi (Age), có giá trị thiếu. Hai biến giả Z_1 và Z_2 đại diện cho hạng vé (Pclass): $Z_1 = 1$ nếu hành khách thuộc hạng vé thứ hai, $Z_2 = 1$ nếu hành khách thuộc hạng vé thứ ba và hạng 1 làm nhóm tham chiếu). Biến Z_3 biểu thị giới tính (Sex) trong đó $Z_3 = 1$ nếu nam và 0 nếu nữ. Biến Z_4 là số người thân đi cùng (SibSp), có giá trị từ 0 đến 8. Ngoài ra, biến W được xây dựng từ Parch như một biến thay thế cho X , với $W = 1$ nếu hành khách có ít nhất một người thân đi cùng và $W = 0$ ngược lại, có giá trị từ 0 đến 6.

Kết quả thống kê mô tả trong Bảng 4 cho thấy trong tổng số 891 hành khách, có 549 người tử vong (61,62%) và 342 người sống sót (38,38%), phản ánh tỉ lệ tử vong cao. Biến tuổi X có 177 giá trị bị thiếu (19,87%), đây là mức thiếu đáng kể và có thể ảnh hưởng đến độ tin cậy của phân tích nếu không được xử lý phù hợp. Khi chỉ các quan sát đầy đủ được sử dụng ($n = 714$), tuổi trung bình khoảng 29,7 và độ lệch chuẩn 14,53 cho thấy sự phân tán lớn. Về các biến phân loại, hành khách hạng 3 chiếm tỉ lệ cao nhất (55,11%), nam giới chiếm ưu thế (64,76%). Số người đi cùng tương đối thấp với Z_4 có giá trị trung bình 0,52 và độ lệch chuẩn là 1,1 và biến W cho thấy 23,91% hành khách có ít nhất một người thân đi cùng.

Để đánh giá mối liên hệ giữa X và biến thay thế W , hệ số tương quan hạng Spearman được tính trên các quan sát đầy đủ, cho kết quả $\rho = -0,2549$ với $p < 0,05$, cho thấy mối liên hệ nghịch có ý nghĩa thống kê nhưng không mạnh. Điều này hàm ý rằng hành khách có người thân đi cùng (giá trị W cao) có xu hướng trẻ hơn. Ngoài ra, W không tương quan với Y , do đó có thể được sử dụng như biến phụ hỗ trợ cho X trong xử lý dữ liệu thiếu.

Gọi δ là biến chỉ báo quan sát của X ($\delta = 1$, nếu quan sát và $\delta = 0$ nếu thiếu). Xác suất quan sát của X được mô hình hóa bởi hồi quy logistic:

$$P(\delta = 1|Y, Z_1, Z_2, Z_3, Z_4, W) = H(\gamma_0 + \gamma_1 Y + \gamma_2 Z_1 + \gamma_3 Z_2 + \gamma_4 Z_3 + \gamma_5 Z_4 + \gamma_6 W).$$

Kết quả (Bảng 5) cho thấy các biến Z_1, Z_2, Z_4 và W có ý nghĩa thống kê, trong khi Y và Z_3 không có ý nghĩa. Dấu của các hệ số cho thấy Z_1 làm tăng xác suất quan sát của X , trong khi Z_2 và Z_4 làm giảm xác suất này; đặc biệt, số người thân đi cùng càng nhiều thì khả năng thiếu tuổi càng cao. Ngoài ra, kiểm định tổng thể (Bảng 6) cho thấy mô hình đầy đủ vượt trội so với mô hình rỗng ($\chi^2_6 = 80,64$, $p < 0,001$), với độ lệch (deviance) giảm đáng kể. Do đó, xác suất quan sát của X phụ thuộc vào các biến quan sát được nhưng không phụ thuộc đáng kể vào Y , ủng hộ giả định MAR thay vì MCAR.

Bảng 4. Thống kê mô tả các biến được mã hóa trong bộ dữ liệu Titanic

Biến	Thống kê mô tả
Y	891 (100%)
$Y = 0$	549 (61,62%)
$Y = 1$	342 (38,38%)
X	$\bar{X} = 29,7$ ($s_X = 14,53$)
$X = NA$	177 (19,87%)
$Z_1 = 1$	184 (20,65%)
$Z_2 = 1$	491 (55,11%)
$Z_3 = 0$	314 (35,24%)
$Z_3 = 1$	577 (64,76%)
Z_4	$\bar{Z}_4 = 0,52$ ($s_{Z_4} = 1,10$)
W	891 (100%)
$W = 0$	678 (76,09%)
$W = 1$	213 (23,91%)

Bảng 5. Kết quả mô hình hồi quy logistic cho cơ chế thiếu của biến X

Biến	Tham số	Est	SE	z	p
	γ_0	1,5453	0,3163	4,885	< 0,001
Y	γ_1	0,1306	0,2331	0,560	0,575
Z_1	γ_2	0,9281	0,3742	2,480	0,013
Z_2	γ_3	-0,8171	0,2384	-3,428	< 0,001
Z_3	γ_4	0,1001	0,2267	0,442	0,659
Z_4	γ_5	-0,2581	0,0892	-2,894	0,004
W	γ_6	1,5695	0,3372	4,655	< 0,001

Bảng 6. Kiểm định tổng thể mô hình logistic cho cơ chế thiếu

Model	Residual df	Deviance	p -value
Null model	890	888,38	
Full model	884	807,74	$2,63 \times 10^{-15}$

4.2. Kết quả phân tích

Để nghiên cứu mối quan hệ giữa xác suất sống sót và các biến giải thích trong bộ dữ liệu Titanic, mô hình hồi quy logistic được sử dụng:

$$P(Y = 1|X, Z_1, Z_2, Z_3, Z_4) = H(\beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2 + \beta_4 Z_3 + \beta_5 Z_4). \quad (22)$$

Tham số β được ước lượng theo hai phương pháp CC và SIPW, cùng với sai số chuẩn (ASE) và p -value, được trình bày trong Bảng 7. Kết quả cho thấy tất cả các hệ số đều có ý nghĩa thống kê ở mức 5% đối với cả hai phương pháp. Các ước lượng giữa hai phương pháp khá tương đồng, tuy nhiên SIPW có xu hướng cho giá trị nhỏ hơn và sai số chuẩn thấp hơn (đặc biệt từ β_2 đến β_5), cho thấy hiệu quả ước lượng được cải thiện. Hệ số của biến tuổi mang dấu âm, cho thấy xác suất sống sót giảm khi tuổi tăng,

trong khi các biến hạng vé và giới tính cũng có ảnh hưởng đáng kể.

Khi dữ liệu thiếu theo cơ chế MAR, phương pháp SIPW cho kết quả ổn định và phù hợp với các kết quả mô phỏng trước đó. Để đánh giá mức độ phù hợp của mô hình (22), kiểm định GOF được thực hiện với giả thuyết:

$$E(Y|X, Z_1, Z_2, Z_3, Z_4) = H(\beta_0 + \beta_1 X + \beta_2 Z_1 + \beta_3 Z_2 + \beta_4 Z_3 + \beta_5 Z_4).$$

Kết quả trong Bảng 8 cho thấy p -value của cả hai kiểm định C-Pearson (0,2654) và W-Pearson (0,2931) đều lớn hơn 0,05, do đó không bác bỏ H_0 . Điều này cho thấy mô hình logistic là phù hợp với dữ liệu. Đồng thời, giá trị thống kê của W-Pearson (1,0513) nhỏ hơn C-Pearson (1,1136), cho thấy sự cải thiện về độ phù hợp khi áp dụng phương pháp SIPW.

Bảng 7. Ước lượng tham số của mô hình hồi quy logistic

Biến	Tham số	CC			SIPW		
		Est	ASE	p -value	Est	ASE	p -value
X	β_0	4,3342	0,4507	< 0,001	4,2690	0,4517	< 0,001
	β_1	-0,0448	0,0082	< 0,001	-0,0433	0,0088	< 0,001
Z_1	β_2	-1,4144	0,2847	< 0,001	-1,3159	0,2495	< 0,001
Z_2	β_3	-2,6526	0,2858	< 0,001	-2,5721	0,2770	< 0,001
Z_3	β_4	-2,6277	0,2148	< 0,001	-2,7237	0,1918	< 0,001
Z_4	β_5	-0,3802	0,1215	0,0018	-0,3987	0,1146	< 0,001

Bảng 8. Kết quả kiểm định độ phù hợp của mô hình hồi quy logistic

Kiểm định	Thống kê kiểm định	p -value
C-Pearson	1,1136	0,2654
W-Pearson	1,0513	0,2931

5. KẾT LUẬN

Nghiên cứu này được thực hiện đã đề xuất và phát triển kiểm định PCS cho mô hình hồi quy logistic trong trường hợp tồn tại hiệp biến bị thiếu dữ liệu dưới giả định MAR. Trên cơ sở phương pháp ước lượng SIPW, thống kê kiểm định W-Pearson được xây dựng nhằm hiệu chỉnh ảnh hưởng của cơ chế thiếu dữ liệu, từ đó cho phép đánh giá mức độ phù hợp của mô hình một cách đáng tin cậy hơn so với phương pháp CC truyền thống.

Kết quả lý thuyết cho thấy thống kê kiểm định đề xuất có phân phối tiệm cận chuẩn sau khi chuẩn hóa, tạo cơ sở cho việc suy luận thống kê trong mẫu lớn. Các nghiên cứu mô phỏng đã chứng minh rằng kiểm định W-Pearson kiểm soát tốt sai lầm loại I và có công suất kiểm định cao, đặc biệt trong các trường hợp tỉ lệ thiếu dữ liệu lớn hoặc mô hình có sai lệch phi tuyến. So với kiểm định C-Pearson,

phương pháp đề xuất cho thấy tính ổn định và hiệu quả vượt trội hơn, đồng thời tiệm cận với kiểm định F-Pearson trong trường hợp dữ liệu đầy đủ.

Việc ứng dụng trên bộ dữ liệu Titanic cho thấy phương pháp SIPW giúp cải thiện độ chính xác của ước lượng và nâng cao độ phù hợp của mô hình so với phương pháp CC. Kết quả kiểm định GOF cho thấy mô hình hồi quy logistic được giả định là phù hợp với dữ liệu, đồng thời minh họa tính khả thi và hiệu quả của phương pháp đề xuất trong thực hành.

Bên cạnh đó, so với các phương pháp MI, cách tiếp cận dựa trên SIPW cho phép xây dựng trực tiếp thống kê kiểm định mà không cần tạo và tổng hợp nhiều bộ dữ liệu hoàn chỉnh, từ đó giúp đơn giản hóa quá trình tính toán và suy luận thống kê. Đây là một ưu điểm thực tiễn quan trọng của phương pháp đề xuất trong các bài toán có dữ liệu thiếu.

Trong các nghiên cứu tiếp theo, phương pháp này có thể được mở rộng cho các mô hình hồi quy tổng quát khác hoặc các trường hợp dữ liệu thiếu phức tạp hơn, chẳng hạn như nhiều hiệp biến bị thiếu đồng thời hoặc cơ chế thiếu NMAR. Ngoài ra, việc kết hợp với các phương pháp học máy hiện đại cũng

là một hướng nghiên cứu tiềm năng nhằm nâng cao khả năng ứng dụng trong các bài toán dữ liệu lớn.

LỜI CẢM ƠN

Đề tài này được tài trợ bởi Đại học Cần Thơ, mã số: T2025-13.

TÀI LIỆU THAM KHẢO

- Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons.
- Casella, G., & Berger, R. (2002). *Statistical inference*. Chapman and Hall/CRC.
- Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.
<https://doi.org/10.1002/9781118548387>
- Little, R. J., and Rubin, D. B. (2019). *Statistical analysis with missing data*. John Wiley and Sons.
<https://doi.org/10.1007/s11336-022-09856-8>
- 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/10.1177/09622802211079350>
- 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>
- Lộc, T. P., Thúy, T. T. T., Tuyền, D. T., Ba, D. T. B., Nhân, L. H., & Chương, L. H. (2024). Ước lượng tham số mô hình hồi quy logistic với hiệp biến thiếu dữ liệu ngẫu nhiên và ứng dụng. *Tạp chí Khoa học Đại học Cần Thơ*, 60(4), 88-98.
<https://doi.org/10.22144/ctujos.2024.389>
- Noughabi, H. A. (2025). A novel goodness-of-fit test for lifetime models based on Type II censored data. *International Journal of Reliability, Quality and Safety Engineering*, 32(04), 2450055.
<https://doi.org/10.1142/S0218539324500554>
- 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. (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>
- Tran, P. L., Lee, S. M., Le, T. N., & Li, C. S. (2025). Large-sample properties of multiple imputation estimators for parameters of logistic regression with covariates missing at random separately or simultaneously. *Annals of the Institute of Statistical Mathematics*, 77(2), 251-287.
<https://doi.org/10.1007/s10463-024-00914-9>
- 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, 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](https://doi.org/10.1016/S0167-7152(01)00167-5)