JDI
JOURNAL OF DEVELOPMENT AND INTEGRATION

TẠP CHÍ PHÁT TRIỂN VÀ HỘI NHẬP
Số 86 (2026) 45-53 | jdi.uef.edu.vn

So sánh hồi quy nhị phân và hồi quy thứ bậc trong nghiên cứu giáo dục và hành vi người học

Nguyễn Ngọc Duy[1], Lê Phan Thanh Hòa[2,*]
[1]Viện Sau đại học, Trường Đại học Văn Lang, Thành phố Hồ Chí Minh, Việt Nam
[2]Khoa Kế toán - Kiểm toán, Trường Đại học Văn Lang, Thành phố Hồ Chí Minh, Việt Nam

TỪ KHÓA

Hành vi xã hội,
học sau đại học,
phân tích định lượng,
hồi quy nhị phân,
hồi quy thứ bậc.

TÓM TẮT

Nghiên cứu so sánh hiệu quả của hồi quy nhị phân và hồi quy thứ bậc trong phân tích mức độ sẵn sàng học sau đại học của nhân viên văn phòng. Dữ liệu khảo sát giả lập bởi ChatGPT gồm 400 quan sát với các yếu tố động lực, rào cản và hỗ trợ. Hai mô hình được áp dụng: hồi quy nhị phân (gom mức 4–5 thành "sẵn sàng", mức 1–3 thành "không sẵn sàng") và hồi quy thứ bậc (giữ nguyên 5 mức). Kết quả cho thấy hồi quy thứ bậc phát hiện rào cản gia đình có ảnh hưởng giảm mức độ sẵn sàng (β = -0,468, OR = 0,63, p = 0,021), trong khi mô hình nhị phân không phát hiện được. Độ phù hợp của cả hai mô hình còn thấp (Pseudo R^2 < 3%). Kết quả chỉ ra rằng hồi quy thứ bậc giúp khai thác đầy đủ thông tin thứ bậc, trong khi hồi quy nhị phân đơn giản và trực quan hơn. Nghiên cứu cũng khuyến nghị khảo sát thực địa để kiểm định kết quả.

1. Giới thiệu

Hồi quy nhị phân và hồi quy thứ bậc đều là các phương pháp thống kê dùng để dự đoán biến phụ thuộc phân loại. Hồi quy nhị phân được sử dụng khi biến phụ thuộc chỉ có hai trạng thái ("Có" hoặc "Không"), trong khi hồi quy thứ bậc áp dụng cho biến phụ thuộc dạng thứ bậc với nhiều mức thứ tự (mức 1 đến 5). Việc lựa chọn mô hình phù hợp ảnh hưởng đến khả năng tận dụng thông tin của dữ liệu và sức mạnh thống kê của phân tích (Fullerton, 2009). Nếu một biến thứ bậc bị phân đôi tùy ý thành biến nhị phân, mô hình nhị phân có thể làm mất thông tin và giảm độ chính xác so với mô hình thứ bậc (Williams, 2006). Ngược lại, mô hình thứ bậc phức tạp hơn, đòi hỏi giả định "tỷ số OR tương đương" – tức là tác động của biến độc lập được giả định là nhất quán trên các ngưỡng phân loại của biến phụ thuộc.

Phân tích dữ liệu thứ tự là một thách thức phổ biến trong khoa học xã hội và giáo dục, các biến kết quả thường đo lường mức độ, thái độ hoặc sự sẵn sàng dưới dạng thang đo thứ bậc. Một cách tiếp cận phổ biến nhưng gây tranh cãi là phân đôi biến thứ tự thành biến nhị phân để áp dụng hồi quy nhị phân. Nhiều nghiên cứu chỉ ra rằng việc phân đôi tùy tiện có thể làm mất mát thông tin, giảm sức mạnh thống kê, và gây sai lệch kết quả phân tích (Sankey, 1998). Ngược lại, hồi quy thứ bậc tận dụng toàn bộ thông tin thứ tự, giả định tỷ lệ odds tương đương trên các mức phân loại (Agresti, 2010; Fullerton, 2009), mô hình này cũng có những hạn chế nhất định, như yêu cầu giả định thống nhất ảnh hưởng ngưỡng cắt, và độ phức tạp trong kiểm định và diễn giải kết quả (Brant, 1990; Williams, 2006).

Nghiên cứu được thực hiện với mục tiêu so sánh hiệu quả phân tích của hồi quy nhị phân và hồi quy thứ bậc trong việc dự đoán biến phụ thuộc. Dữ liệu sử dụng

*Tác giả liên hệ. Email: *hoa.lpt@vlu.edu.vn*. Địa chỉ: *69/68 Đặng Thùy Trâm, Phường Bình Lợi Trung, Thành phố Hồ Chí Minh, Việt Nam*
https://doi.org/10.61602/jdi.2025.86.06
Ngày nộp bài: 02/5/2025; Ngày chỉnh sửa: 02/6/2025; Ngày duyệt đăng: 09/6/2025; Ngày online: 21/11/2025
ISSN (print): 1859-428X, ISSN (online): 2815-6234

trong nghiên cứu là dữ liệu giả lập, được xây dựng để mô phỏng hợp lý phân phối hành vi học tập, đảm bảo mục tiêu kiểm nghiệm kỹ thuật thay vì suy luận thực tiễn. Cách tiếp cận này phù hợp với định hướng nghiên cứu phương pháp học thuật, giúp minh họa những ưu điểm và hạn chế, những rủi ro tiềm ẩn của từng mô hình trong bối cảnh phân tích dữ liệu thứ tự (Hosmer, 2013).

Nghiên cứu này áp dụng cả hai mô hình để phân tích yếu tố ảnh hưởng đến mức độ sẵn sàng học sau đại học của 400 người, từ đó so sánh độ phù hợp mô hình, khả năng giải thích và ý nghĩa thực tiễn của mỗi phương pháp. Dữ liệu giả lập được khảo sát bởi ChatGPT cho 400 người với bảng câu hỏi dành cho nghiên cứu các yếu tố ảnh hưởng đến quyết định học sau đại học của nhân viên văn phòng gồm các biến ảnh hưởng là Động lực, Rào cản tác động đến quyết định học sau đại học. Biến phụ thuộc là mức độ sẵn sàng học được chia từ 1 đến 5 theo thang đo Likert (Fullerton, 2009) như sau: Không học (1 điểm), Có thể không học (2 điểm), Phân vân (3 điểm), Có thể học (4 điểm), Sẵn sàng học (5 điểm).

Khác với phần lớn các nghiên cứu hiện nay thường tập trung áp dụng một mô hình duy nhất để phân tích hành vi học tập, nghiên cứu này có đóng góp chính ở khía cạnh phương pháp luận bằng cách so sánh hai mô hình hồi quy phổ biến – hồi quy nhị phân và hồi quy thứ bậc – trong việc phân tích dữ liệu thứ tự. Qua đó, bài viết làm sáng tỏ tác động của việc phân đôi biến thứ tự đối với khả năng phát hiện các mối quan hệ thống kê; phân tích những điểm mạnh, điểm yếu của mô hình nhị phân so với thứ bậc; và đưa ra khuyến nghị về lựa chọn mô hình phù hợp khi xử lý dữ liệu thứ tự trong nghiên cứu hành vi giáo dục và xã hội học.

2. Cơ sở lý thuyết và tổng quan nghiên cứu trước

Hồi quy nhị phân dự đoán xác suất có quyết định sẵn sàng học sau đại học (Có = 1 và Không = 0) dựa trên 17 biến độc lập và Hồi quy thứ bậc dự đoán mức độ sẵn sàng (1 đến 5) dựa trên cùng các biến độc lập.

Mô hình hồi quy nhị phân được ước lượng bằng phương pháp tối đa hóa hàm hợp lý. Xác xuất để biến phụ thuộc Yi nhận được giá trị 1 (Có quyết định sẵn sàng học sau đại học) tương ứng với loạt biến Xi (17 biến) được xác định bởi công thức:

$$P(Y_i = 1 \mid X_i) = \frac{e^{x'_i \beta}}{1 + e^{x'_i \beta}} \quad (1)$$

Xác suất để biến phụ thuộc Yi nhận được giá trị 1 (Có quyết định không học sau đại học) được xác định bởi:

$$P(Y_i = 0 \mid X_i) = 1 - P(Y_i = 0 \mid X_i) \quad (2)$$

Hàm hợp lý (likelihood function) cho toàn bộ mẫu gồm n quan sát là:

$$l(\beta) = \prod_{i=1}^{n} \left(\frac{e^{x'_i \beta}}{1 + e^{x'_i \beta}} \right)^{Y_i} \left(\frac{1}{1 + e^{x'_i \beta}} \right)^{1 - Y_i} \quad (3)$$

và hàm log-hợp lý tương ứng:

$$l(\beta) = \sum_{i=1}^{n} \left[Y_i \left(X'_i \beta \right) - \log\left(1 + e^{X'_i \beta} \right) \right] \quad (4)$$

Mô hình hồi quy thứ bậc sử dụng hàm liên kết logit theo tỷ số OR tích lũy (proportional odds model). Trong mô hình này, xác suất tích lũy để Yi nằm ở mức nhỏ hơn hoặc bằng j được mô hình hóa như:

$$\log P(Y_i \le j \mid X_i) - \log P(Y_i > j \mid X_i) = \mu_j - X'_i \beta \quad (5)$$

với j = 1, 2, 3, 4 tương ứng với các ngưỡng phân loại giữa các mức quyết định Không học (1 điểm), Có thể không học (2 điểm), Phân vân (3 điểm), Có thể học (4 điểm), và Sẵn sàng học (5 điểm).

Trong đó, μ_j là các ngưỡng phân loại cần ước lượng và β là vector hệ số hồi quy chung cho tất cả các ngưỡng. Trong mô hình thứ bậc, bốn ngưỡng phân loại μ_1, μ_2, μ_3, μ_4 (giữa các mức 1-2, 2-3, 3-4, 4-5) được ước lượng đồng thời với các hệ số β.

Hàm log-hợp lý trong mô hình thứ bậc được xác định bởi:

$$l(\beta, \mu) = \sum_{i=1}^{n} \log\left(P\left(Y_i = j_i \right) \right) \quad (6)$$

Độ phù hợp chung của mô hình được đánh giá thông qua ba chỉ số sau:

1. Log-likelihood (l): đo lường mức độ khớp giữa mô hình và dữ liệu khảo sát.

2. Kiểm định Likelihood Ratio (LR):

$$\chi^2 = -2 \times \left(l_{null} - l_{model} \right) \quad (7)$$

Trong đó, l_{null} là log-likelihood của mô hình rút gọn (chỉ có hằng số) và l_{model} là log-likelihood của mô hình đầy đủ.

3. Chỉ số Pseudo R^2 của McFadden

$$R^2_{McFadden} = 1 - \frac{l_{model}}{l_{null}} \quad (8)$$

Kiểm tra tầm quan trọng của từng biến độc lập, các hệ số hồi quy được chuyển đổi sang tỷ số OR (odds ratio) theo công thức:

$$OR = e^{\beta} \quad (9)$$

Trong cả hai mô hình, OR > 1 nghĩa là biến độc lập tăng làm tăng tính sẵn sàng, OR < 1 nghĩa là có tác động giảm xác suất sẵn sàng.

Cuối cùng, chúng tôi so sánh độ phức tạp và khả năng ứng dụng thực tế của hai mô hình dựa trên việc diễn giải kết quả và yêu cầu giả định của mỗi mô hình.

Để thực hiện việc tính toán trên, các mã lập trình trong Python đã được sử dụng, cụ thể như sau: (i) Thư viện statsmodels.api.Logit được sử dụng để xây dựng mô hình hồi quy nhị phân; (ii) Thư viện statsmodels.miscmodels.ordinal_model.OrderedModel được sử dụng để xây dựng mô hình hồi quy thứ bậc; (iii) Thư viện statsmodels dùng để tính toán các kiểm định LR và Wald; (iv) Thư viện numpy dùng để tính toán OR.

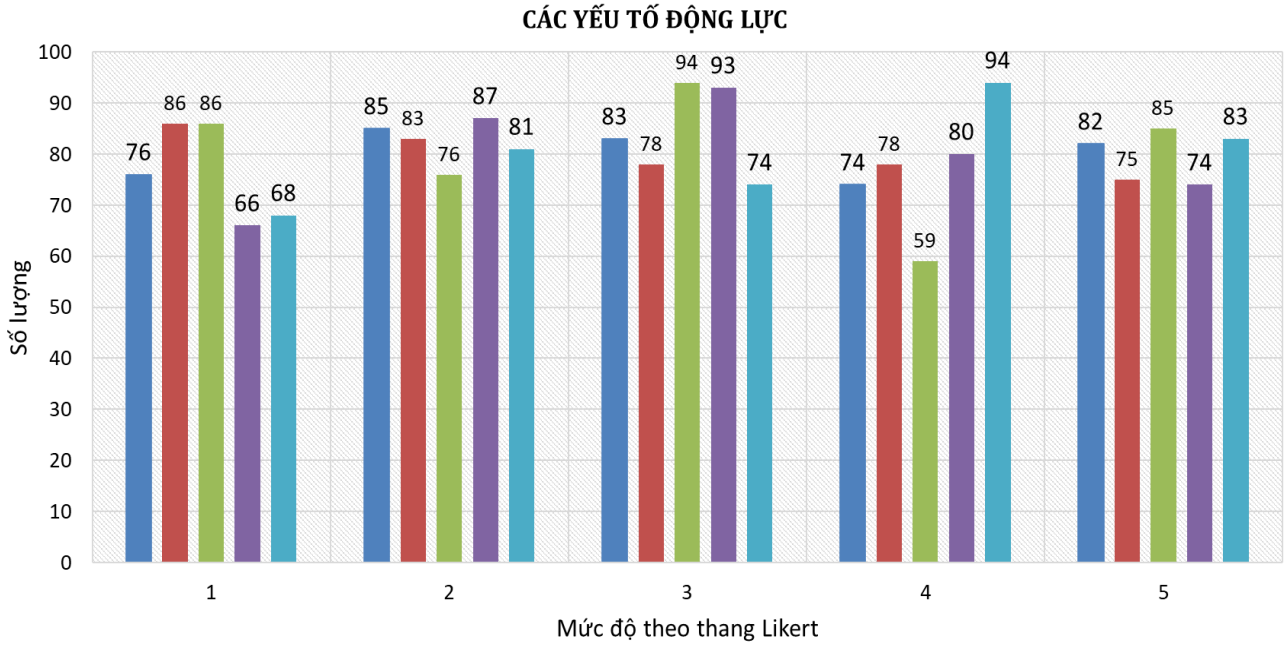
Trên thế giới, các nghiên cứu của Agresti (2010), Fullerton (2009), Liu và Agresti (2005) đã phân tích chi tiết các giả định, ưu điểm và giới hạn của từng loại mô hình. Bender và Grouven (1998) áp dụng các mô hình này vào các dữ liệu hành vi và học tập, chỉ ra rằng việc phân đôi tùy tiện dữ liệu thứ tự có thể làm suy giảm độ chính xác và gây ra sai lệch trong phân tích. Tại Việt Nam, Trần, V.H. và Nguyễn, T.T. (2020) phân tích các yếu tố ảnh hưởng đến ý định học tiếp của sinh viên ngành kinh tế tại TP.HCM. Nguyễn, T.N. và Phạm, M.Đ. (2021) nhấn mạnh vai trò động lực cá nhân và môi trường học tập trong quyết định học cao học. Phần lớn nghiên cứu trong nước vẫn dùng hồi quy nhị phân do dễ ước lượng nhưng chưa đánh giá đầy đủ tác động của việc phân loại dữ liệu thứ tự. Khoảng trống này chính là cơ hội để nghiên cứu đưa ra đóng góp bổ sung về lý luận và thực nghiệm.

3. Mô tả dữ liệu

Chúng tôi xem xét đến 05 yếu tố động lực (chuyển nghề, thăng tiến, được động viên, đam mê, nâng cao chuyên môn), 07 yếu tố có thể tác động (uy tín trường, lịch học linh hoạt, học phí, tính ứng dụng của chương trình, hình thức học online, học bổng, bạn bè tư vấn) và 06 yếu tố cản trở quyết định học sau đại học của nhân viên văn phòng. Số liệu thống kê về quyết định học sau đại học được trình bày ở Hình 1, Hình 2, Hình 3, Hình 4.

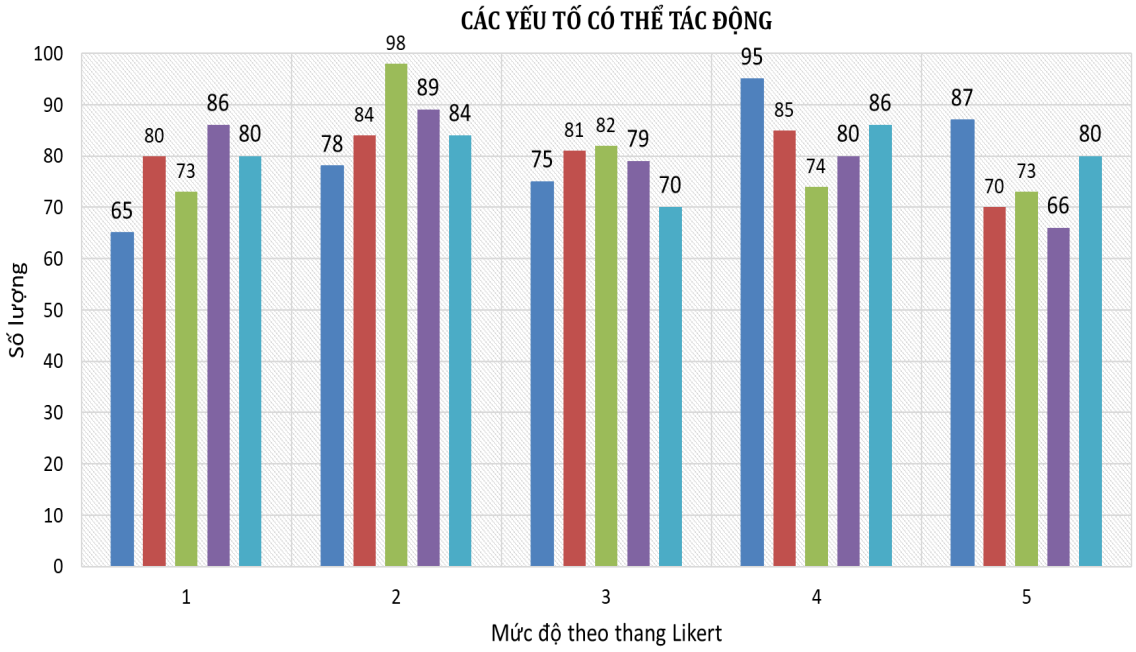
Kết quả trả lời cho các biến rào cản được mã hóa theo giá trị 0 (không là rào cản) và 1 (rào cản). Do mỗi người đều chọn đúng ba rào cản, tổng của sáu biến có giá trị bằng 3 với mọi quan sát, gây hiện tượng đa cộng tuyến hoàn hảo dẫn đến tình trạng không thể giải được bài toán ước lượng hồi quy (Gujarati, 2009). Để tránh đa cộng tuyến hoàn hảo, chúng tôi loại bỏ biến Không thấy nhu cầu học thêm. Như vậy, mô hình tổng cộng có 17 biến độc lập, bao gồm 05 biến động lực, 07 biến yếu tố có thể tác động và 05 biến giả rào cản.

Đối với biến phụ thuộc là Mức độ sẵn sàng học sau đại học, Hình 4 cho thấy tỷ lệ người chọn mức 1 (rất không sẵn sàng) là cao nhất (92 người, chiếm 23%), trong khi mức 4 (khá sẵn sàng) có ít người nhất (66 người, ~17%). Nhìn chung, có xu hướng nhiều người chưa sẵn sàng (mức 1–2–3 chiếm ~64%) hơn là sẵn sàng (mức 4–5 chiếm ~36%). Trong phân tích hồi



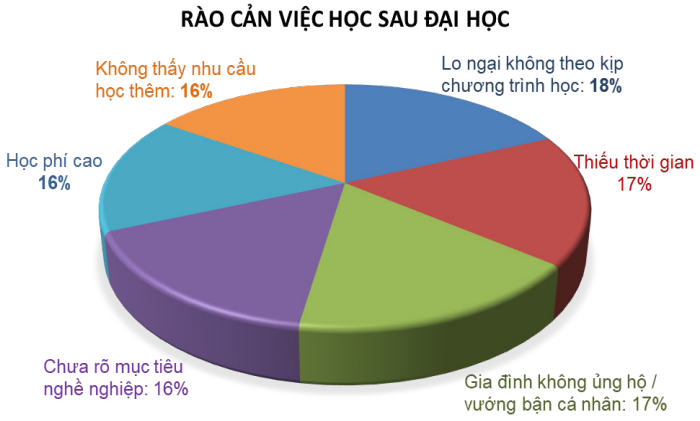
Hình 1. Biểu đồ thống kê các yếu tố động lực thúc đẩy quyết định học sau đại học của nhân viên văn phòng theo các mức độ của thang đo Likert

(Các cột từ trái qua phải tương ứng với các động lực: Chuyển nghề, Thăng tiến, Được động viên, Đam mê và Nâng cao chuyên môn)

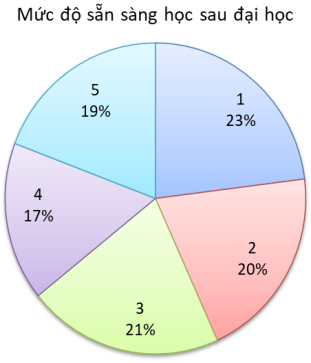


Hình 2. Biểu đồ thống kê các yếu tố có thể tác động đến quyết định học sau đại học của nhân viên văn phòng theo các mức độ của thang đo Likert

(Các cột từ trái qua phải tương ứng với các yếu tố: Uy tín trường, Lịch học linh hoạt, Học phí, Tính ứng dụng của chương trình, Hình thức học online, Học bổng và Bạn bè tư vấn)



Hình 3. Biểu đồ thống kê các yếu tố rào cản quyết định học sau đại học của nhân viên văn phòng



Hình 4. Biểu đồ thống kê mức độ độ sẵn sàng học sau đại học của nhân viên văn phòng

quy nhị phân, chúng tôi phân nhóm biến này thành hai mức: mức 4–5 được gộp thành nhóm "Có sẵn sàng" (mã hóa là 1), và mức 1–2–3 thành nhóm Không sẵn sàng (mã hóa 0). Như vậy, 35,8% mẫu (143 người) thuộc nhóm Có sẵn sàng, còn 64,2% (257 người) thuộc nhóm Không sẵn sàng. Đối với hồi quy thứ bậc, chúng tôi sử dụng trực tiếp biến thứ bậc 5 mức Likert làm biến phụ thuộc, bảo toàn toàn bộ thông tin thứ hạng.

4. Kết quả và biện luận

4.1. Mô hình hồi quy nhị phân

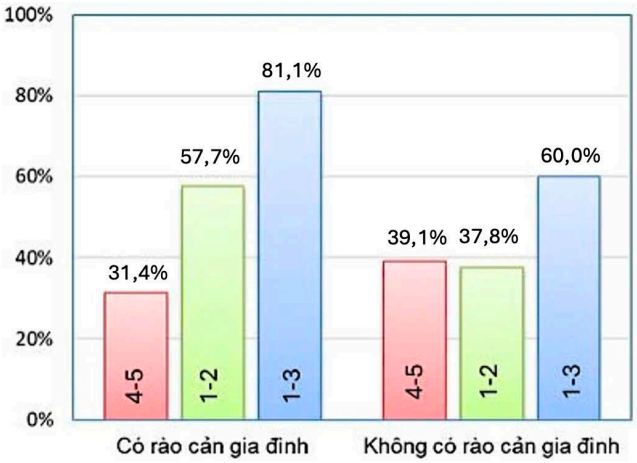
Kết quả hồi quy nhị phân được tóm tắt trong Bảng 1. Động lực Nâng cao năng lực có hệ số dương tương đối lớn (β = 0,094), cho thấy những người coi trọng

động lực nâng cao năng lực có xu hướng sẵn sàng học cao hơn, nhưng mức ảnh hưởng cũng chưa đủ mạnh do p-value lớn (p = 0,231). Các động lực khác như Thăng tiến, Được khuyến khích có hệ số dương rất nhỏ và không đáng kể (p > 0,3), còn Chuyển nghề và Đam mê có hệ số âm rất nhỏ và p > 0,65 nên các yếu tố này cũng không thúc đẩy mạnh quyết định học sau đại học. Các biến khác (Uy tín trường, Lịch học linh hoạt, Học phí, Tính ứng dụng của chương trình, Hình thức học online, Học bổng và Bạn bè tư vấn) ảnh hưởng không đáng kể đến biến phụ thuộc do p > 0,24.

Bảng 1. Kết quả hồi quy nhị phân

STT	Biến độc lập	β	OR	p
1	Thăng tiến	0,071	1,07	0,349
2	Chuyển nghề	–0,029	0,97	0,707
3	Được khuyến khích	0,033	1,03	0,658
4	Đam mê	–0,013	0,99	0,872
5	Nâng cao năng lực	0,094	1,10	0,231
6	Học phí cao	–0,289	0,75	0,244
7	Thiếu thời gian	–0,080	0,92	0,742
8	Lo ngại không theo kịp chương trình	0,073	1,08	0,754
9	Chưa rõ mục tiêu nghề nghiệp	0,120	1,13	0,611
10	Gia đình không ủng hộ/ vướng bận cá nhân	–0,316	0,73	0,183
11	Uy tín	–0,090	0,91	0,247
12	Lịch học linh hoạt	0,057	1,06	0,461
13	Học phí phù hợp	–0,033	0,97	0,671
14	Chương trình ứng dụng	0,042	1,04	0,595
15	Học online	–0,006	0,99	0,934
16	Học bổng	–0,070	0,93	0,376
17	Tư vấn bạn bè	–0,061	0,94	0,414

Đối với nhóm biến rào cản, không có biến nào ảnh hưởng nổi bật đến quyết định học sau đại học trong mô hình nhị phân. Tuy nhiên, một số biến có tác động lớn hơn so với các biến khác. Chẳng hạn như, biến Gia đình không ủng hộ (β = -0,316) có tác động mạnh hơn biến Học phí cao (β = -0,289) trong việc làm suy giảm khả năng sẵn sàng học. Điều này cũng được thể hiện qua chỉ số OR lần lượt là 0,75 và 0,73. Thật vậy, dữ liệu cho thấy khoảng 31% người có rào cản gia đình đạt mức sẵn sàng 4–5, so với 39% ở nhóm không có rào cản này (Hình 5). Các rào cản khác như Thiếu thời gian, Lo ngại không theo kịp, Chưa rõ mục tiêu có tính cản trở thấp hơn so với học phí cao và Gia đình không ủng hộ (|β| < 0,13 và p-value ~ 0,5 – 0,8).



Hình 5. Tỉ lệ chọn các mức độ sẵn sàng (4-5: mức độ sẵn sàng cao [có], 1-2: mức độ sẵn sàng rất thấp [không], 1-3: mức độ sẵn sàng thấp [không]) khi có/không rào cản gia đình

Khi đối chiếu với dữ liệu khảo sát, độ phù hợp của mô hình nhị phân nhìn chung không cao. Log-likelihood của mô hình là lmodel = -254,25 so với lnull = -260,79 của mô hình null. Kiểm định LR cho thấy mô hình không cải thiện đáng kể so với mô hình null (χ^2 = 13,08, p = 0,73). Giá trị Pseudo R^2 chỉ đạt 0,025, nghĩa là mô hình này chỉ giải thích được khoảng 2,5% thông tin so với mô hình đơn giản nhất. Giá trị Pseudo R^2 thấp như vậy khá phổ biến trong hồi quy cho dữ liệu hành vi phức tạp (Hosmer, 2013; McFadden, 1974). Điều này cho thấy các yếu tố động lực, rào cản và có thể tác động trong mô hình nhị phân chưa đủ để dự đoán tốt quyết định học sau đại học – có thể còn những nhân tố quan trọng khác (hoàn cảnh gia đình, tài chính, tính cách…) cần xem xét đưa vào mô hình.

4.2. Mô hình hồi quy thứ bậc

Bảng 2 trình bày kết quả hồi quy thứ bậc. Nhìn chung, độ lớn và dấu của các hệ số hồi quy thứ bậc có sự tương đồng với mô hình nhị phân, nhưng có một khác biệt quan trọng trong nhóm biến rào cản. Biến rào cản Gia đình không ủng hộ/vướng bận cá nhân có ảnh hưởng lớn đến quyết định học sau đại học, điều này không được tìm thấy trong mô hình nhị phân. Cụ thể, hệ số hồi quy đối với biến này là β = -0,468 (ứng với OR = 0,63 và p = 0,021). Kết quả này cho thấy những người gặp rào cản gia đình, mức sẵn sàng học sau đại học giảm rõ rệt (khoảng 37%) so với người không có rào cản này, nếu các yếu tố khác giống nhau. Kết quả này nhất quán với quan sát ở Hình 5.

Tất cả các biến độc lập khác trong mô hình thứ bậc đều không có ảnh hưởng đáng kể (p > 0,05), kết quả này tương tự mô hình nhị phân. Các yếu tố động lực Thăng tiến có hệ số dương nhỏ (β = 0,068, p =

0,279), Nâng cao năng lực gần bằng 0 (β= 0,011, p = 0,863), trong khi Chuyển nghề và Được khuyến khích có hệ số âm rất nhỏ (p ~ 0,4–0,96). Các yếu tố Uy tín, Học phí phù hợp, Học bổng, Tư vấn bạn bè tiếp tục có hệ số âm nhẹ (0,05–0,07) nhưng không đáng kể (p ~ 0,25–0,74). Hai rào cản Học phí cao và Thiếu thời gian vẫn mang hệ số âm (~ -0,21 đến -0,24) tương tự mô hình nhị phân, nhưng p-value khá lớn nên chưa có tác động đáng kể (p ~ 0,24–0,30). Các rào cản Lo ngại không theo kịp và Chưa rõ mục tiêu có hệ số rất nhỏ (gần 0, p > 0,75).

Tóm lại, biến rào cản gia đình là yếu tố duy nhất thể hiện tác động có ý nghĩa trong việc phân tầng mức độ sẵn sàng. Các yếu tố động lực và hỗ trợ khác không cho thấy đóng góp rõ rệt trong mô hình hồi quy, có thể do mức độ sẵn sàng học chịu chi phối lớn bởi những yếu tố khác (như điều kiện cá nhân, gia đình) hơn là các động lực nội tại hay đặc điểm chương trình.

Bảng 2. Kết quả hồi quy thứ bậc

STT	Biến độc lập	β	OR	p
1	Thăng tiến	0,068	1,07	0,279
2	Chuyển nghề	–0,055	0,95	0,397
3	Được khuyến khích	–0,003	0,99	0,962
4	Đam mê	0,028	1,03	0,678
5	Nâng cao năng lực	0,011	1,01	0,863
6	Học phí cao	–0,241	0,79	0,238
7	Thiếu thời gian	–0,213	0,81	0,296
8	Lo ngại không theo kịp chương trình	–0,049	0,95	0,808
9	Chưa rõ mục tiêu nghề nghiệp	0,061	1,06	0,763
10	Gia đình không ủng hộ/ vướng bận cá nhân	–0,468	0,63	0,021
11	Uy tín	–0,075	0,93	0,254
12	Lịch học linh hoạt	0,055	1,06	0,391
13	Học phí phù hợp	–0,022	0,98	0,742
14	Chương trình ứng dụng	0,003	1,00	0,963
15	Học online	0,013	1,01	0,846
16	Học bổng	–0,073	0,93	0,271
17	Tư vấn bạn bè	–0,037	0,96	0,559

Về độ phù hợp, mô hình thứ bậc cũng chưa đáp ứng mức ý nghĩa thống kê chung. Log-likelihood của mô hình đầy đủ là lmodel = -633,57 so với lnull = -641,45 của mô hình rút gọn. Kiểm định LR cho kết quả (χ^2 = 15,76, p = 0,46), nghĩa là không có đủ bằng chứng để khẳng định mô hình thứ bậc giúp tăng độ vừa khớp so với mô hình gốc. Pseudo R^2 của mô hình thứ bậc rất thấp (≈ 0,012), thậm chí thấp hơn mô hình nhị phân. Điều này không mâu thuẫn, bởi lẽ mô hình rút gọn của hồi quy thứ bậc đã sử dụng nhiều thông tin hơn (phân phối 5 mức) so với mô hình rút gọn của hồi quy nhị phân chỉ dùng phân phối hai mức (có/không); do đó không gian để cải thiện log-likelihood của mô hình thứ bậc ít hơn. Nói cách khác, việc thêm các biến độc lập vào không cải thiện nhiều khả năng dự đoán mức độ sẵn sàng so với chỉ dùng phân phối quan sát. Kết quả dựa trên dữ liệu giả lập cho thấy các biến động lực/rào cản/yếu tố có thể chưa đủ mạnh để giải thích sự khác biệt mức sẵn sàng giữa các cá nhân.

Xem xét tỷ số OR của mô hình thứ bậc, kết quả cho thấy đa số các biến có ảnh hưởng rất nhỏ ở mọi mức phân loại, do đó việc gộp một hệ số chung là hợp lý. Riêng biến Rào cản gia đình, phân tích cho thấy hiệu ứng của nó tập trung mạnh ở ngưỡng phân biệt giữa nhóm rất không sẵn sàng với phần còn lại (mức 1-2 so với mức 4-5) hơn là ở ngưỡng cao (mức 4-5 so với mức 1-2). Tuy vậy, do mô hình thứ bậc gộp chung một hệ số trung bình nên vẫn phát hiện được ảnh hưởng tổng thể và cho kết luận thống kê nhất quán rằng Rào cản gia đình làm giảm khả năng đạt mức sẵn sàng cao (p < 0,05). Điều này chứng tỏ rằng kết quả hồi quy thứ bậc không bị phụ thuộc vào việc chọn điểm phân cắt (μi) cụ thể (như cách chia 1–2–3 [không] và 4–5 [có] ở mô hình nhị phân). Ngược lại, ở mô hình nhị phân, nếu chọn ngưỡng khác (ví dụ phân nhóm sẵn sàng là mức 3–4–5 so với mức 1–2), có thể một số biến khác sẽ có ý nghĩa. Chính vì vậy, mô hình thứ bậc giúp tận dụng toàn bộ thông tin thứ hạng, cung cấp cái nhìn toàn diện hơn về tác động của biến độc lập trên toàn dải kết quả.

Mô hình thứ bậc còn ước lượng các ngưỡng phân cách (μ_i) giữa 5 mức độ sẵn sàng. Các ngưỡng ước lượng được là μ_{1-2} = -1,929 (ngưỡng giữa mức 1 và 2), μ_{2-3} = -0,040, μ_{3-4} = -0,126, μ_{4-5} = -0,140. Ngưỡng đầu tiên μ_{1-2} mang giá trị âm khá lớn và khác 0 với p = 0,013. Điều này phản ánh việc rất nhiều người ở mức 1 (rất không sẵn sàng) so với các mức cao hơn. Các ngưỡng còn lại xấp xỉ 0 và không có ảnh hưởng đáng kể (p > 0,2), cho thấy xác suất ở các mức 2, 3, 4, 5 không chênh lệch nhiều một cách độc lập – điều này phù hợp với phân phối tương đối gần nhau của các mức 2–5. Nói cách khác, điểm khác biệt nổi bật nhất trong thang đo là giữa nhóm rất không sẵn sàng (mức 1) với nhóm còn lại; còn giữa các mức trung bình – khá – rất sẵn sàng, sự chuyển biến tương đối là như nhau.

4.3. So sánh và thảo luận

Độ phù hợp của mô hình: Cả hai mô hình đều cho giá trị Pseudo R^2 khá thấp (dưới 3%) và kiểm định LR không có ảnh hưởng đáng kể, chứng tỏ các biến độc lập chưa giải thích được nhiều sự biến thiên trong biến phụ thuộc. Nguyên nhân có thể do dữ liệu được giả lập ngẫu nhiên bằng ChatGPT, đòi hỏi cần có nghiên cứu thực địa để kiểm chứng. Mô hình nhị phân có R^2 ~ 0,025, nhỉnh hơn so với R^2 ~ 0,012 của mô hình thứ bậc. Sự khác biệt này phần lớn do cách định nghĩa trong hai mô hình khác nhau, mô hình thứ bậc có log-likelihood null thấp hơn (âm hơn) nhiều do sử dụng phân phối 5 mức nên khó đạt R^2 cao. Thực tế, log-likelihood mô hình thứ bậc (-633,57) đã cải thiện hơn mô hình null (-641,45)

một lượng (~7,88), thậm chí lớn hơn mức cải thiện của mô hình nhị phân (tăng ~6.54 so với null). Dù vậy, với số lượng biến khá lớn (17) so với cỡ mẫu 400, mức cải thiện này chưa đủ mạnh để đạt ý nghĩa. Kết quả này chỉ ra rằng mô hình có thể đang quá nhiều biến dự báo không cần thiết so với tín hiệu thực sự trong dữ liệu. Việc loại bớt biến ít ảnh hưởng hoặc thu thập thêm mẫu có thể cải thiện độ phù hợp trong tương lai.

Khả năng giải thích và ý nghĩa các biến: Hầu hết các yếu tố (động lực, rào cản,…) chưa chứng tỏ được vai trò rõ rệt trong việc dự đoán mức độ sẵn sàng học sau đại học. Chỉ có Rào cản gia đình ảnh hưởng tiêu cực có ý nghĩa (ở mô hình thứ bậc). Điều này cho thấy những khó khăn về gia đình (thiếu sự ủng hộ của người thân, bận rộn con cái, trách nhiệm gia đình) thực sự cản trở đáng kể nguyện vọng học tiếp của cá nhân. Ngược lại, các động lực tích cực như mong muốn thăng tiến, đam mê, hay nhu cầu nâng cao năng lực tuy có hiện hữu nhưng không đủ phân biệt nhóm sẵn sàng cao với nhóm thấp. Có thể những động lực này phổ biến ở nhiều người bất kể họ có học tiếp hay không, trong khi rào cản gia đình mang tính phân hóa quyết định hơn. Nhóm yếu tố về chương trình đào tạo (uy tín, lịch học, học phí,…) không có ảnh hưởng đáng kể – điều này có thể do nhận thức về chương trình chưa rõ ràng khi người được hỏi chưa thực sự trải nghiệm, hoặc họ xem những yếu tố này là điều kiện cần chứ không thúc đẩy quyết định cá nhân.

So sánh hai mô hình, ta thấy mô hình thứ bậc giúp phát hiện mối quan hệ có ý nghĩa mà mô hình nhị phân bỏ lỡ. Cụ thể, biến Rào cản gia đình có tác động thể hiện rõ khi xét trên toàn thang thứ bậc, nhưng khi chỉ xét cắt điểm 3-4 (mô hình nhị phân) thì lại không đủ mạnh để nổi bật ($p \sim 0,18$). Điều này minh họa nhược điểm của việc phân loại nhị phân tùy ý, biến độc lập có thể phụ thuộc vào ngưỡng phân nhóm. Theo Sankey và cộng sự (1998), nếu tùy tiện chia nhỏ thang đo thứ bậc thành hai nhóm, có thể làm sai lệch hoặc bỏ sót hiệu ứng, và tỷ số OR ước lượng được sẽ thay đổi theo điểm cắt. Mô hình thứ bậc tránh được vấn đề này bằng cách sử dụng toàn bộ thông tin thứ tự, cho một hệ số chung đại diện cho mọi phân cắt có thể của biến phụ thuộc. Thật vậy, hệ số -0.468 của Rào cản gia đình trong mô hình thứ bậc có thể hiểu là trung bình của các hệ số nếu chúng ta chạy các mô hình nhị phân tại mọi ngưỡng có thể. Do đó, mô hình thứ bậc tận dụng tốt hơn dữ liệu thứ tự và thường có sức mạnh thống kê cao hơn mô hình nhị phân tương ứng, điều này hoàn toàn phù hợp với các nghiên cứu của Agresti (2010), Liu (2005) và Fullerton (2009). Với cùng bộ dữ liệu, nghiên cứu cho thấy mô hình thứ bậc đã tìm thấy hiệu ứng có ý nghĩa của Rào cản gia đình, trong khi mô hình nhị phân thì không.

Mặc dù vậy, mô hình thứ bậc cũng có những thách thức riêng. Thứ nhất, mô hình này giả định các hệ số là như nhau cho mọi mức cắt (giả định "song song"). Nếu giả định này bị vi phạm, một số biến ảnh hưởng mạnh ở mức độ sẵn sàng thấp nhưng yếu dần ở mức cao (như trường hợp Rào cản gia đình), thì mô hình có thể không phù hợp hoàn toàn (Agresti, 2010). Có phương pháp thay thế như mô hình thứ bậc từng phần (partial proportional odds) cho phép một số hệ số thay đổi theo ngưỡng, nhưng mô hình đó phức tạp hơn nhiều (Fullerton, 2009). Thứ hai, trong bối cảnh thực tiễn, Rào cản gia đình làm giảm 37% mức độ sẵn sàng có thể khó hiểu đối với đối tượng không chuyên. Trong khi đó, mô hình nhị phân cho phép diễn giải trực tiếp xác suất "sẵn sàng hay không", thường trực quan hơn. Thứ ba, mô hình nhị phân phổ biến hơn và có nhiều công cụ đánh giá trực quan phục vụ mục tiêu dự báo kết quả dạng hai lựa chọn. Nếu mục tiêu nghiên cứu tập trung vào việc dự đoán một quyết định nhị phân cụ thể (có định học lên cao học hay không), thì mô hình nhị phân có thể đủ và phù hợp hơn để áp dụng.

Những phát hiện của nghiên cứu này cũng có thể được lý giải qua các lý thuyết hành vi và kinh tế xã hội. Theo lý thuyết hành vi dự định (Ajzen, 1991), sự thiếu ủng hộ từ gia đình làm suy yếu chuẩn mực chủ quan, từ đó làm giảm ý định học tập. Thuyết nhu cầu của Maslow (1943) cho thấy các nhu cầu cơ bản như sự ổn định gia đình cần được thỏa mãn trước khi cá nhân theo đuổi nhu cầu tự hoàn thiện thông qua học tập. Đồng thời, theo lý thuyết vốn con người (Becker, 1964), rào cản gia đình làm tăng chi phí đầu tư giáo dục, khiến cá nhân cân nhắc lại lợi ích và có thể từ bỏ ý định học tiếp. Những lý thuyết này củng cố giải thích cho kết quả phân tích hồi quy thứ bậc về vai trò tiêu cực của rào cản gia đình đối với mức độ sẵn sàng học sau đại học.

Độ phức tạp của mô hình và ứng dụng thực tế: Mô hình nhị phân đơn giản hơn về mặt tính toán và diễn giải, do đó dễ triển khai trong thực tế. Mô hình thứ bậc phức tạp hơn, yêu cầu kiểm tra giả định, nhưng lại cung cấp cái nhìn sâu hơn về dữ liệu thứ tự. Nếu nhà quản lý muốn nhận diện những yếu tố thật sự cản trở nguyện vọng học tiếp, mô hình thứ bậc cung cấp thông tin Rào cản gia đình là điểm nghẽn quan trọng cần được hỗ trợ (tạo điều kiện về thời gian, khuyến khích từ gia đình). Ngược lại, nếu chỉ dùng mô hình nhị phân thì có thể bỏ qua yếu tố này vì nó không nổi bật ở ngưỡng đã chọn. Tuy nhiên, với mục đích dự đoán hành vi sẽ học tiếp hay không, mô hình nhị phân có ưu thế là trực tiếp và dễ áp dụng (tính xác suất mỗi người và phân loại nhóm rủi ro).

Từ kết quả trên có thể thấy mỗi mô hình có ưu và nhược điểm. Hồi quy nhị phân gọn nhẹ, trực quan nhưng có nguy cơ mất thông tin khi biến kết quả có nhiều mức. Hồi quy thứ bậc tận dụng tốt dữ liệu thứ tự và có thể phát hiện hiệu ứng tinh tế hơn, nhưng lại phức tạp và cần thận trọng với giả định. Việc lựa chọn mô hình nên tùy thuộc vào mục tiêu nghiên cứu, nếu mục tiêu là hiểu rõ toàn bộ phổ mức độ sẵn sàng và tận dụng dữ liệu khảo sát chi tiết, mô hình thứ bậc phù hợp hơn. Ngược lại, nếu mục tiêu tập trung vào quyết định cuối cùng (sẵn sàng hay không) và cần một mô hình

đơn giản để dự đoán, mô hình nhị phân có thể được ưu tiên.

Giới hạn của nghiên cứu là sử dụng dữ liệu giả lập thay vì dữ liệu khảo sát thực địa. Việc công bố rõ ràng giới hạn này nhằm đảm bảo tính minh bạch học thuật và phù hợp với khuyến nghị khi sử dụng dữ liệu mô phỏng trong nghiên cứu phương pháp luận (Sankey & Weissfeld, 1998). Bộ dữ liệu được xây dựng dựa trên các giả định hợp lý về phân phối thang đo mức độ sẵn sàng, các biến động lực, rào cản, và yếu tố hỗ trợ, với mục tiêu mô phỏng đặc điểm hành vi điển hình của đối tượng nghiên cứu. Dữ liệu được thiết kế sao cho phân phối các biến độc lập dạng Likert gần với chuẩn phân phối thực tế (trải đều từ 1 đến 5) và mối tương quan giữa các biến giữ ở mức thấp (đa phần hệ số tương quan Spearman $|r| < 0,1$), đảm bảo tính độc lập cần thiết cho mô hình hồi quy (Fullerton, 2009).

5. Kết luận và khuyến nghị

Nghiên cứu phân tích dữ liệu giả lập khảo sát 400 người để dự đoán mức độ sẵn sàng học sau đại học dựa trên các biến động lực, rào cản và yếu tố hỗ trợ. Hai phương pháp hồi quy - nhị phân và thứ bậc - được áp dụng và so sánh. Kết quả cho thấy mô hình thứ bậc nhạy hơn trong việc phát hiện yếu tố quan trọng, khi xác định được ảnh hưởng có ý nghĩa của rào cản gia đình mà mô hình nhị phân không tìm thấy. Cả hai mô hình đều có độ phù hợp tổng thể chưa cao, chỉ ra rằng những biến đưa vào chưa đủ mạnh để dự báo quyết định học sau đại học. Về so sánh, mô hình nhị phân đơn giản, dễ hiểu nhưng có thể bỏ sót thông tin thứ bậc; mô hình thứ bậc cung cấp bức tranh đầy đủ hơn về dữ liệu thứ tự, đổi lại yêu cầu giả định và diễn giải phức tạp hơn. Các nhà nghiên cứu và quản lý giáo dục cần cân nhắc mục tiêu sử dụng mô hình để lựa chọn phương pháp phù hợp. Nếu muốn hiểu sâu và khai thác tối đa dữ liệu khảo sát mức độ sẵn sàng, nên dùng hồi quy thứ bậc. Nếu cần công cụ dự báo hoặc phân loại đơn giản về quyết định học, hồi quy nhị phân có thể đáp ứng tốt hơn. Cuối cùng, nghiên cứu cũng chỉ ra Rào cản gia đình là yếu tố cản trở đáng kể việc học lên cao học, chỉ ra rằng các chính sách hỗ trợ người học cần chú trọng giảm bớt gánh nặng gia đình để tăng tỷ lệ tiếp tục học sau đại học.

Trên cơ sở kết quả phân tích, nghiên cứu đề xuất một số khuyến nghị nhằm nâng cao hiệu quả sử dụng mô hình hồi quy trong nghiên cứu hành vi học tập sau đại học. Về chính sách, các cơ sở đào tạo nên hỗ trợ người học gặp rào cản gia đình bằng các chương trình linh hoạt như lớp ngoài giờ, học online hoặc học bổng phù hợp. Về học thuật, các nhà nghiên cứu cần thận trọng khi phân đôi biến thứ bậc và nên ưu tiên hồi quy thứ bậc để tận dụng toàn bộ dữ liệu, đồng thời kiểm định giả định tỷ lệ odds tương đương. Về thực tiễn, mô hình nhị phân có thể dùng cho phân loại đơn giản, nhưng nên kết hợp mô hình thứ bậc để phân tích sâu, tránh bỏ sót những yếu tố quan trọng.

Dựa trên những kết quả đạt được và giới hạn đã nêu, nghiên cứu tiếp theo nên tiến hành khảo sát thực tế đối với đối tượng có nhu cầu học sau đại học tại Việt Nam, nhằm: (i) Thu thập dữ liệu thực tế về động lực, rào cản và yếu tố có thể tác động quyết định học; (ii) Kiểm định lại các phát hiện từ nghiên cứu giả lập, đặc biệt là vai trò của Rào cản gia đình; (iii) Ứng dụng mô hình hồi quy thứ bậc để tận dụng toàn bộ dữ liệu thứ tự, đồng thời kiểm tra giả định tỷ số OR tương đương bằng các phương pháp như kiểm định Brant (Brant, 1990). Nghiên cứu mở rộng có thể so sánh thêm các mô hình hiện đại khác như mô hình hồi quy thứ bậc từng phần hoặc mô hình hồi quy thứ bậc đa cấp để tăng độ chính xác dự báo, góp phần hoàn thiện bộ công cụ phân tích hành vi học tập trong bối cảnh giáo dục Việt Nam đang thúc đẩy chiến lược phát triển nguồn nhân lực chất lượng cao.

TÀI LIỆU THAM KHẢO

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179-211. DOI: https://doi.org/10.1016/0749-5978(91)90020-T

Agresti, A. (2010). *Analysis of ordinal categorical data* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Becker, G. S. (1964). *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*. University of Chicago Press.

Bender, R., & Grouven, U. (1998). Using binary logistic regression models for ordinal data with non-proportional odds. *Journal of Clinical Epidemiology*, 51(10), 809-816. DOI: https://doi.org/10.1016/S0895-4356(98)00074-8.

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, 46(4), 1171-1178. DOI: https://doi.org/10.2307/2532457.

Fullerton, A. S. (2009). A conceptual framework for ordered logistic regression models. *Sociological Methods & Research*, 38(2), 306-347. DOI: https://doi.org/10.1177/0049124109346162.

Gujarati, D. N., & Porter, D. C. (2009). *Basic econometrics* (5th ed.). New York, NY: McGraw-Hill.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Liu, I., & Agresti, A. (2005). The analysis of ordered categorical data: An overview and a survey of recent developments. *Test*, 14(1), 1-73.

Maslow, A. H. (1943). A theory of human motivation. *Psychological Review*, 50(4), 370-396.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105-142). Academic Press.

Nguyễn, T. N., & Phạm, M. Đ. (2021). Đánh giá mức độ sẵn sàng học cao học của sinh viên năm cuối các trường đại học công lập. *Tạp chí Giáo dục Đại học*, 35(2), 123-134.

Sankey, S. S., & Weissfeld, L. A. (1998). A study of the effect of dichotomizing ordinal data upon modeling. *Communications in Statistics - Simulation and Computation*, 27(4), 871-887.

Trần, V. H., & Nguyễn, T. T. (2020). Phân tích các yếu tố ảnh hưởng đến ý định học tiếp của sinh viên ngành kinh tế tại TP.HCM. *Tạp chí Kinh tế và Phát triển*, 27(3), 45-56.

Williams, R. (2006). Generalized ordered logit/partial proportional odds models for ordinal dependent variables. *The Stata Journal*, 6(1), 58-82. DOI: https://doi.org/10.1177/1536867X0600600104.

Comparing Binary and Ordered Logistic Regression in Studies of Education and Learner Behavior

Nguyen Ngoc Duy[1], Le Phan Thanh Hoa[2,*]

[1]Institute of Postgraduate Education, Van Lang University, Ho Chi Minh City, Vietnam

[2]Faculty of Accounting and Auditing, Van Lang University, Ho Chi Minh City, Vietnam

Abstract

This study compares the effectiveness of binary logistic regression and ordered logistic regression in analyzing office workers' readiness to pursue postgraduate education. A simulated survey dataset generated by ChatGPT includes 400 observations encompassing motivational, barrier, and support factors. Two models were applied: binary logistic regression (combining levels 4–5 as "ready" and levels 1–3 as "not ready") and ordered logistic regression (retaining all five ordinal levels). The ordered model identified a significant negative effect of family-related barriers on readiness ($\beta = -0.468$, OR = 0.63, $p = 0.021$), which was not detected in the binary model. Both models showed low overall fit (Pseudo $R^2 < 3\%$). The findings suggest that ordered logistic regression better utilizes ordinal information, while binary regression offers a simpler, more intuitive approach. The study recommends field surveys to validate these results.

Keywords: Social behavior, postgraduate education, quantitative analysis, binary logistic regression, ordered logistic regression.