

ANALYSING QUANTITATIVE MODELS ABOUT FACTORS OF ACTIVITIES OF DAILY LIVING IMPACTING ON THE STUDENT ASSESSMENT RESULT IN THAI NGUYEN UNIVERSITY OF TECHNOLOGY

Pham Thi Thu Hang^{1*}, Pham Thi Thuy²

¹TNU - University of Technology

²TNU - University of Education

ARTICLE INFO	ABSTRACT
Received: 06/5/2022	In the paper, we aim to present the construction of regression models based on the data collected on the underline factors of activities of daily living directly effecting to the student's assessment result of students in Thai Nguyen University in the recent years. Basing on this data, some approaches are proposed to construct useful quantitative models of important factors to the response, the student's assessment result, by using various techniques of statistics and machine learning including the multiple linear regression, Anova, k-folds. The analysis of these models shows how good they are and provide us some important and helpful conclusions relating the concerned problem. Through this, we expect to approve the result of the paper for referencing to the construction of the training and studying plan of the education managers as well as the learners in order to have a better quality training and studying.
Revised: 31/5/2022	
Published: 31/5/2022	
KEYWORDS	
Quantitative model	
Multiple linear regression	
Anova	
Education Science	
Statistical analysis	

PHÂN TÍCH MÔ HÌNH ĐỊNH LƯỢNG VỀ CÁC NHÂN TỐ CỦA CÁC HOẠT ĐỘNG HÀNG NGÀY ẢNH HƯỞNG ĐẾN KẾT QUẢ HỌC TẬP CỦA NGƯỜI HỌC TẠI TRƯỜNG ĐẠI HỌC KỸ THUẬT CÔNG NGHIỆP THÁI NGUYÊN

Phạm Thị Thu Hằng^{1*}, Phạm Thị Thủy²

¹Trường Đại học Kỹ thuật Công nghiệp - ĐH Thái Nguyên

²Trường Đại học Sư phạm - ĐH Thái Nguyên

THÔNG TIN BÀI BÁO	TÓM TẮT
Ngày nhận bài: 06/5/2022	Bài báo này giới thiệu việc xây dựng mô hình hồi quy dựa trên bộ dữ liệu quan sát được về các nhân tố của các hoạt động hàng ngày được xem là có ảnh hưởng trực tiếp đến kết quả học tập của người học tại trường Đại học Kỹ thuật Công nghiệp, Đại học Thái Nguyên trong những năm gần đây. Từ bộ dữ liệu này, một số hướng tiếp cận được đặt ra để xây dựng mô hình định lượng hữu ích về những nhân tố quan trọng, ảnh hưởng đến biến phụ thuộc, kết quả của người học, bằng nhiều phương pháp cơ bản của thống kê và máy học gồm hồi quy đa tuyến tính, phân tích Anova, phương pháp k-cụm. Sự phân tích của các mô hình này cho thấy mức độ phù hợp và hữu ích của chúng và cung cấp cho chúng ta một số kết luận có ý nghĩa, quan trọng, liên quan đến bài toán thực tiễn. Qua đây, các tác giả mong muốn đóng góp về mặt học thuật cho việc tham khảo xây dựng kế hoạch đào tạo của các nhà quản lý giáo dục cũng như việc xây dựng kế hoạch học tập người học nhằm mang lại hiệu quả tốt hơn.
Ngày hoàn thiện: 31/5/2022	
Ngày đăng: 31/5/2022	
TỪ KHÓA	
Mô hình định lượng	
Hồi quy đa tuyến tính	
Phân tích phương sai	
Khoa học giáo dục	
Phân tích thống kê	

DOI: <https://doi.org/10.34238/tnu-jst.5951>

* Corresponding author. Email: phamthuhang0201@gmail.com

1. Mở đầu

Kết quả học tập của người học là một nhân tố khách quan đánh giá năng lực của một cá nhân về khả năng tiếp thu tri thức khoa học và năng lực vận dụng những tri thức, kỹ năng chuyên môn được đào tạo, năng lực tư duy của họ. Một môi trường giáo dục đại học tốt mang đến cơ hội rèn luyện, cải tạo tư duy, trí tuệ và bồi đắp các kỹ năng cần thiết cho người học. Song tác động này không phải là một yếu tố có tính thụ động, mà cần có sự tích cực và chủ động tiếp nhận của người học. Chất lượng đào tạo cho người học có thực sự được nâng cao ở một mức độ cập nhật được với một nền sản xuất tiên tiến hay không còn phụ thuộc vào yếu tố chủ quan của người học. Do đó, việc nghiên cứu các nhân tố về lịch sinh hoạt hàng ngày, những nhân tố có vai trò quan trọng đối với sự hình thành thái độ và động cơ học tập, sẽ rất có ích đối với bản thân quá trình đào tạo cũng như quá trình học tập của họ. Trong nghiên cứu này, tác giả ứng dụng lý thuyết thống kê vào việc nghiên cứu định lượng những tác động của các thói quen sinh hoạt hàng ngày của người học trong một trường kỹ thuật đến kết quả học tập của họ thông qua một mô hình hồi quy đa tuyến tính được xây dựng từ bộ dữ liệu thu thập được trong những năm gần đây về người học của trường Đại học Kỹ thuật Công nghiệp, Đại học Thái Nguyên. Các phương pháp đánh giá chất lượng của mô hình được đưa ra để tìm ra một mô hình có ích, phù hợp nhất với bộ dữ liệu [1]-[6]. Từ kết quả đạt được đó, các nhân tố quan trọng được đánh giá bằng thang đo định lượng để mang lại những kết luận khách quan của chủ đề nghiên cứu [7], [8].

2. Dữ liệu và các đánh giá dữ liệu

Mẫu thống kê có kích thước 403 được xây dựng trên cơ sở dữ liệu thu thập được từ sinh viên của trường trong các khóa sinh viên gần thời gian nghiên cứu (trước thời điểm dịch Covid-19) gồm K49, K50, K5-VB2, K51, K52, K53, K54, K55, thuộc nhiều khoa chuyên môn khác nhau của trường. Các kiểm định mẫu được đưa ra ở đây để đánh giá tính khách quan của bộ dữ liệu. Bộ dữ liệu gồm các trường thông tin là các biến thống kê được như sau:

1. Kết quả học tập (theo thang điểm 4) học kỳ 1 năm học (2019-2020), ký hiệu: *TBHK*.
2. Khóa: ký hiệu *Khoa*, sinh viên tương ứng với khóa đào tạo của trường.
3. Số tín chỉ đăng ký trong học kỳ 1: ký hiệu *TCDK*.
4. Số tín chỉ dự thi học kỳ: ký hiệu *TCTHI*, là một yếu tố cấu thành trực tiếp đến *TBHK*.
5. Thời gian nghỉ học trong học kỳ, tính bằng số tiết nghỉ học: ký hiệu *TG_nghi*.
6. Thời gian dành cho việc tự học trung bình một ngày trong học kỳ: ký hiệu *Tu_hoc*.
7. Thời gian chơi thể thao trung bình trong ngày: ký hiệu *Sport*, các môn thể thao vận động.
8. Thời gian chơi Games trung bình trong ngày: ký hiệu *Xgames*.
9. Thời gian tham gia các mạng xã hội (Facebook, Zalo, Instagram,...): ký hiệu *Facebook*.
10. Thời gian làm thêm trung bình trong ngày: ký hiệu *PT_Job*.
11. Thời gian dành cho việc giao lưu với bạn bè trung bình trong ngày: ký hiệu *Chatting*.
12. Thời gian dành cho việc mua sắm đồ ăn, đồ tiêu dùng trung bình: ký hiệu *Market*.
13. Thời điểm đi ngủ hàng ngày: ký hiệu *Sleep*.
14. Biến định tính xem xét khung giờ đi ngủ hàng ngày, ký hiệu *Time*, được phân loại theo các khung giờ: $L1 = (Time)_1$: khung giờ từ trước đến 22h30 (sớm); $L2 = (Time)_2$: từ sau 22h30 đến trước 0h00 (muộn); $L3 = (Time)_3$: từ sau 0h00 đến trước 3h00 (khuya).
15. Thời gian ngủ trung bình một đêm: ký hiệu *Sleeptime*.
16. Biến định tính xác định người sống cùng với người được điều tra (ký hiệu *Part*): $P1 = (Part)_1$: sống một mình; $P2 = (Part)_2$: sống cùng với bạn; $P3 = (Part)_3$: sống cùng với người yêu; $P4 = (Part)_4$: sống cùng với gia đình.
17. Thời gian tham gia các hoạt động đoàn thể, tình nguyện, tính trung bình theo tuần: *Youth*.
18. Thời gian tham gia sinh hoạt các Club tiếng Anh, tính trung bình theo tuần: ký hiệu *E-CLUB*.

19. Thời gian thực hiện các cuộc nói chuyện với người thân, người trong gia đình thông qua mạng di động, mạng Internet, tính trung bình theo tuần: ký hiệu *Video_chat*.

20. Biến định tính xác định môi trường cư trú của sinh viên được điều tra (ký hiệu *Residence*):

$P1 = (Residence)_1$: cư trú tại ký túc xá trường; $P2 = (Residence)_2$: cư trú tại xóm trọ gần trường (khoảng cách dưới 1,5 km); $P3 = (Residence)_3$: cư trú tại xóm trọ xa trường (khoảng cách trên 1,5 km); $P4 = (Residence)_4$: cư trú cùng gia đình (bố, mẹ, anh, chị, ông, bà...).

Bộ dữ liệu thu thập trong ngắn hạn (short-term), được giới hạn trong thời gian của học kỳ nghiên cứu, nhằm mục tiêu lấy được thông tin đáng tin cậy nhất, đồng thời các tác động của các biến dự báo đến biến phụ thuộc được thể hiện rõ nhất, ít bị ảnh hưởng bởi các nhân tố dài hạn (long-term). Các biến dự báo định lượng (numerical predictors) được đo theo cùng đơn vị thời gian là giờ. Các biến định tính (categorical variables) được mã hóa theo các giá trị như sau:

1) Biến *Time* là: $L1$ nếu $\{L2 = 0, L3 = 0\}$; $L2$ nếu $\{L2 = 1, L3 = 0\}$; $L3$ nếu $\{L2 = 0, L3 = 1\}$.

2) Biến *Part* là: $P1$ nếu $\{P2 = 0, P3 = 0, P4 = 0\}$; $P2$ nếu $\{P2 = 1, P3 = 0, P4 = 0\}$; $P3$ nếu $\{P2 = 0, P3 = 1, P4 = 0\}$; $P4$ nếu $\{P2 = 0, P3 = 0, P4 = 1\}$.

3) Biến *Residence* là: $R1$ nếu $\{R2 = 0, R3 = 0, R4 = 0\}$; $R2$ nếu $\{R2 = 1, R3 = 0, R4 = 0\}$; $R3$ nếu $\{R2 = 0, R3 = 1, R4 = 0\}$; $R4$ nếu $\{R2 = 0, R3 = 0, R4 = 1\}$.

4) Biến *Khoa* là: $K49 = (Khoa)_1$ nếu $\{K5VB2 = K50 = K51 = K52 = K53 = K54 = K55 = 0\}$, $K5VB2 = (Khoa)_2$ nếu $\{K5VB2 = 1, K50 = K51 = K52 = K53 = K54 = K55 = 0\}$, $K50 = (Khoa)_3$ nếu $\{K5VB2 = K51 = K52 = K53 = K54 = K55 = 0, K50 = 1\}$, $K51 = (Khoa)_4$ nếu $\{K5VB2 = K50 = K52 = K53 = K54 = K55 = 0, K51 = 1\}$, $K52 = (Khoa)_5$ nếu $\{K5VB2 = K50 = K51 = K53 = K54 = K55 = 0, K52 = 1\}$, $K53 = (Khoa)_6$ nếu $\{K5VB2 = K50 = K51 = K52 = K54 = K55 = 0, K53 = 1\}$, $K54 = (Khoa)_7$ nếu $\{K5VB2 = K50 = K51 = K52 = K53 = K55 = 0, K54 = 1\}$, $K55 = (Khoa)_8$ nếu $\{K5VB2 = K50 = K51 = K52 = K53 = K54 = 0, K55 = 1\}$.

2.1. Các kiểm định trên bộ dữ liệu

Các công cụ kiểm định trực quan cho phép đánh giá mức độ chuẩn hóa dữ liệu khách quan bao gồm phân tích tính chất phân phối chuẩn của biến phụ thuộc *TBHK* thông qua các test Shapiro–Wilk và non-constant error variance, các thông số skewness và kurtosis, ma trận tương quan giữa các cặp biến định lượng (*TBHK, TCTHI, TG_nghi, Tu_hoc, Sport, XGames, Facebook, PT_Job, Chatting, Market, Sleeptime, Youth, E_CLUB*), biểu đồ hồi quy tuyến tính của biến phụ thuộc trên từng biến dự báo cụ thể. Các kiểm định này được thể hiện cho mô hình cụ thể được đánh giá [1], [3], [4]. Ma trận hệ số tương quan giữa các cặp biến định lượng thể hiện trong Bảng 1.

Bảng 1. Ma trận hệ số tương quan giữa các cặp biến định lượng từ bộ dữ liệu

	<i>TBHK</i>	<i>TCTHI</i>	<i>TG_nghi</i>	<i>Tu_hoc</i>	<i>Sport</i>	<i>XGames</i>	<i>Facebook</i>	<i>PT_Job</i>	<i>Chatting</i>	<i>Market</i>	<i>Sleeptime</i>	<i>Youth</i>	<i>E_CLUB</i>
<i>TBHK</i>	1,00	0,17	-0,36	0,34	0,07	-0,52	-0,16	-0,14	0,13	0,04	-0,09	0,08	0,11
<i>TCTHI</i>	0,17	1,00	-0,12	0,04	0,00	-0,17	0,01	0,00	0,04	0,06	-0,05	0,04	0,09
<i>TG_nghi</i>	-0,36	-0,12	1,00	-0,11	0,01	0,30	0,05	0,13	-0,03	-0,05	0,11	-0,01	-0,07
<i>Tu_hoc</i>	0,34	0,04	-0,11	1,00	0,01	-0,25	-0,14	-0,02	0,03	-0,06	-0,07	0,11	0,09
<i>Sport</i>	0,07	0,00	0,01	0,01	1,00	-0,24	0,02	0,03	-0,02	0,09	-0,08	0,09	-0,04
<i>XGames</i>	-0,52	-0,17	0,30	-0,25	-0,24	1,00	0,13	0,15	-0,02	-0,06	0,01	-0,08	-0,06
<i>Facebook</i>	-0,16	0,01	0,05	-0,14	0,02	0,13	1,00	0,26	-0,03	0,06	0,00	0,00	-0,02
<i>PT_Job</i>	-0,14	0,00	0,13	-0,02	0,03	0,15	0,26	1,00	-0,04	-0,01	-0,07	0,07	0,07
<i>Chatting</i>	0,13	0,04	-0,03	0,03	-0,02	-0,02	-0,03	-0,04	1,00	0,06	-0,09	0,16	0,12
<i>Market</i>	0,04	0,06	-0,05	-0,06	0,09	-0,06	0,06	-0,01	0,06	1,00	-0,02	0,17	0,08
<i>Sleeptime</i>	-0,09	-0,05	0,11	-0,07	-0,08	0,01	0,00	-0,07	-0,09	-0,02	1,00	-0,04	0,13
<i>Youth</i>	0,08	0,04	-0,01	0,11	0,09	-0,08	0,00	0,07	0,16	0,17	-0,04	1,00	0,35
<i>E_CLUB</i>	0,11	0,09	-0,07	0,09	-0,04	-0,06	-0,02	0,07	0,12	0,08	0,13	0,35	1,00

Từ ma trận này, ta thấy các cặp biến đều có hệ số tương quan nhỏ (dưới 0,9 được coi là mức chấp nhận được). Điều này cho thấy ta không có vấn đề gặp phải về sự tương quan tuyến tính giữa chúng [2].

2.2. Xây dựng mô hình hồi quy dựa trên các phương pháp lựa chọn biến phụ thuộc

Trong mục này chúng ta sẽ đề xuất các mô hình hồi quy tuyến tính và phi tuyến, các phương pháp lựa chọn mô hình phù hợp với bộ dữ liệu thu thập. Việc kiểm chứng tính hữu dụng, chất lượng các mô hình này dựa trên các test Shapiro–Wilk và non-constant error variance (ncv), chỉ số variance inflation factor (VIF), chỉ số AIC (tương đương chỉ số BIC) [5], [7]. Các kiểm định khác đánh giá chất lượng mô hình cũng được trình bày trong Mục 3 thông qua các phương pháp kiểm định chéo, kiểm định phân nhóm k-cụm, kiểm định dựa trên tập huấn luyện (training data) và tập kiểm chứng (testing data).

2.2.1. Mô hình hồi quy đa tuyến tính thô sơ dựa trên toàn bộ tập dữ liệu ban đầu: Nmod

Đối với mô hình này, ta xem xét sự đóng góp của tất cả các biến dự báo trong tập dữ liệu ban đầu. Mô hình được biểu diễn trong phương trình (1). Thông qua các kết quả hệ số thu được của mô hình này, thu được trên phần mềm R, ta có thể thấy mức độ đóng góp không đạt mức độ đáng tin cậy cao của các biến *Time*, *Sport*, *Facebook*, *PT_Job*, *Chatting*, *Sleeptime*, *Part*, *Youth*, *E_CLUB*, *Market*, *Khoa*. Điều này cho thấy mức độ overfitting của mô hình và sự cần thiết loại bỏ các yếu tố này để cải thiện mô hình.

$$TBHK = b_0 + b_1(TCTHI) + b_{2,2}(Time)_2 + b_{2,3}(Time)_3 + b_3(TG_nghe) + b_4(Tu_hoc) + b_5(Sport) + b_6(XGames) + b_7(Facebook) + b_8(PT_Job) + b_9(Chatting) + b_{10}(Sleeptime) + b_{11,2}(Part)_2 + b_{11,3}(Part)_3 + b_{11,4}(Part)_4 + b_{12,2}(Residence)_2 + b_{12,3}(Residence)_3 + b_{12,4}(Residence)_4 + b_{13}(Youth) + b_{14}(E_CLUB) + b_{15}(Market) + b_{16,2}(Khoa)_2 + b_{16,3}(Khoa)_3 + b_{16,4}(Khoa)_4 + b_{16,5}(Khoa)_5 + b_{16,6}(Khoa)_6 + b_{16,7}(Khoa)_7 + b_{16,8}(Khoa)_8. \quad (1)$$

Các thông số skewness và kurtosis của biến phụ thuộc *TBHK* là 0.1114761 và 0.7941757 so với 0 và 3 của các thông số này đối với phân phối chuẩn. Các test cho tiêu chuẩn phân phối chuẩn của thặng dư (residual) cho mô hình là Shapiro–Wilk và ncv cho các kết quả $W = 0.98597$, $p\text{-value} = 0.0006266$ và $Chisquare = 18.42896$, $Df = 1$, $p = 1.7636e-05$ tương ứng. Các $p\text{-value}$ của hai test này đều nhỏ hơn nhiều mức tin cậy chấp nhận được tối thiểu 0.10. Thông số (Akaike) $AIC = 619.2596$ (hoặc thông số tương đương là Schwarz Bayes $BIC = 735.1567$). Đối với thông số này, mô hình con nào có thông số nhỏ hơn là tốt hơn. Ta sẽ dùng tiêu chuẩn này để lựa chọn mô hình [5], [7].

2.2.2. Mô hình hồi quy đa tuyến tính dựa trên việc lựa chọn biến dự báo từ mô hình Nmod

Việc lựa chọn biến dự báo dựa trên thông số AIC đưa đến mô hình M0 biểu diễn ở phương trình (2). Tuy vậy, các thông số từ hai test Shapiro–Wilk và ncv cũng cho kết quả nhỏ, không đạt mức đáng tin cậy yêu cầu.

$$TBHK = 1.97197 + 0.01495 * TCTHI - 0.02475 * TG_nghe + 0.10461 * Tu_hoc - 0.20258 * XGames - 0.02116 * Facebook + 0.03520 * Chatting - 0.19134 * ResidenceR2 - 0.30077 * ResidenceR3 - 0.15897 * ResidenceR4. \quad (2)$$

Lựa chọn mô hình dựa trên tiêu chí hai test kiểm định thặng dư của mô hình có tuân theo phân phối chuẩn hay không dẫn đến một tập các biến dự báo và mô hình tương ứng, ký hiệu LM0 được cho bởi phương trình (3). Trong mô hình này, hai test trên cho các kết quả kiểm định $p\text{-value}$ tốt ở mức chấp nhận được, lần lượt là $W = 0.99733$, $p\text{-value} = 0.7651$ và $Chisquare = 2.410272$, $Df = 1$, $p = 0.12054$. Tuy vậy, mô hình này còn các biến dự báo là *Time*, *Sleeptime*, *PT_Job* có $p\text{-value}$ của ước lượng hệ số theo phân phối t (phân phối Student) không đạt (tức cao hơn) mức đáng tin cậy mặc định. Điều này dẫn đến khả năng về sự dư thừa thông tin (overfitting) của mô hình LM0.

$$TBHK = 1.515986 + 0.075029 * TimeL2 - 0.007532 * TimeL3 + 0.034612 * TCTHI - 0.199506 * ResidenceR2 - 0.252170 * ResidenceR3 - 0.282932 * ResidenceR4 - 0.029349 * Facebook - 0.025431 * Sleeptime + 0.068770 * Sport - 0.022309 * PT_Job + 0.156594 * Tu_hoc. \quad (3)$$

Từ kết quả này, một lần nữa sử dụng chỉ số AIC cho việc lựa chọn mô hình. Chỉ số AIC nhỏ nhất xây dựng được từ LM0, ký hiệu LM1, cho bởi phương trình (4).

$$TBHK = 1.41032 + 0.03544 * TCTHI + 0.16253 * Tu_hoc - 0.02888 * Facebook - 0.02089 * PT_Job - 0.21045 * ResidenceR2 - 0.27050 * ResidenceR3 - 0.30135 * ResidenceR4 \quad (4)$$

Thông số hai test Shapiro–Wilk và ncv lần lượt là $W = 0.99777$, $p\text{-value} = 0.8745$ và $\text{Chisquare} = 3.114972$, $Df = 1$, $p = 0.077576$. AIC của LM1 là 720.343 so với 722.8904 của LM0 có chút cải thiện. Chỉ số Multiple R-squared: 0.1923, Adjusted R-squared: 0.178, F-statistic: 13.4 on 7 và 394 DF, $p\text{-value}: 1.531e-15$. Các kiểm nghiệm mô hình LM1 trực quan hóa bởi các Hình 1-4.

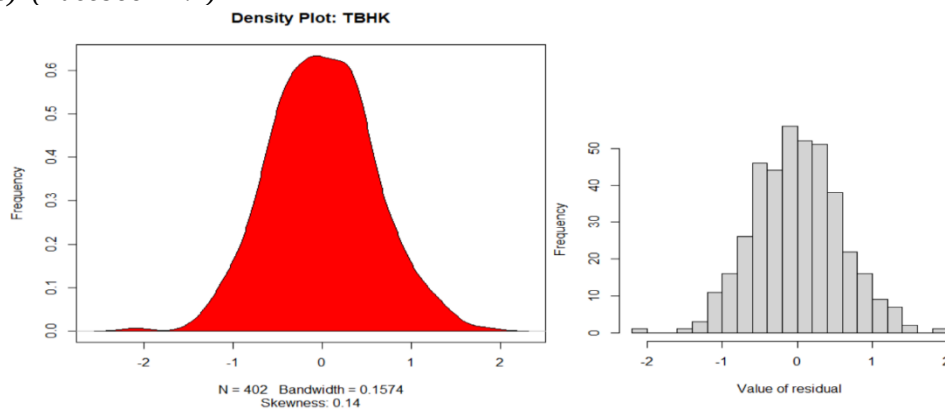
Các biểu đồ trong Hình 1-2 thể hiện tính phân phối chuẩn tắc của thặng dư cho mô hình LM1. Biểu đồ Hình 3 (phải) biểu diễn tính đúng đắn của mô hình trên toàn bộ tập dữ liệu quan sát so với những giá trị phù hợp (fitted values). Các biểu đồ trong Hình 3 (trái) và Hình 4 thể hiện tầm ảnh hưởng những quan sát ngoại lệ trên tổng thể mô hình LM1. Những outliers (quan sát ngoại lệ) có độ đo Cook lớn hơn $4/394 \approx 0.01$. Từ các biểu đồ này, có thể thấy LM1 không có những vi phạm nghiêm trọng, là mô hình chấp nhận được.

2.2.3. Mô hình hồi quy chứa các nhân tố tương tác (interaction terms)

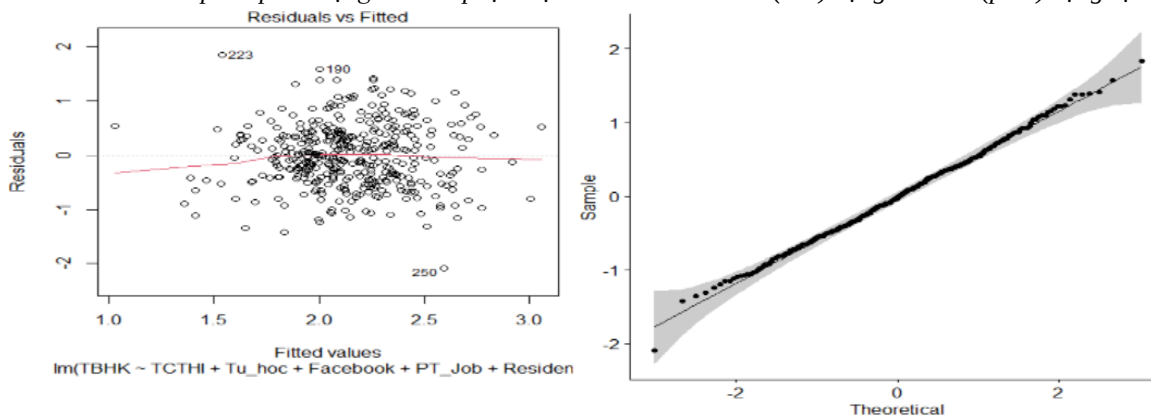
Xét hai mô hình chứa nhân tố tương tác, ký hiệu M0 và M1, cho bởi phương trình (5) và (6):

$$\begin{aligned} TBHK = & 1.206 + 0.02721 * TCTHI - 0.04355 * (Sport^{0.7}) - 0.01599 * (Tu_hoc^{2.7}) - \\ & 1.199e-05 * (XGames^{6.3}) + 0.3952 * (Tu_hoc) + 0.04.798 * (Facebook) - 0.2074 * ResidenceR2 \\ & - 0.2944 * ResidenceR3 - 0.2808 * ResidenceR4 + 9.213e-03 * (Sport^{0.7}) * (Tu_hoc^{2.7}) - \\ & 1.161e-05 * (XGames^{6.3}) * (Tu_hoc) - 3.920e-02 * (Tu_hoc) * (Facebook), \end{aligned} \quad (5)$$

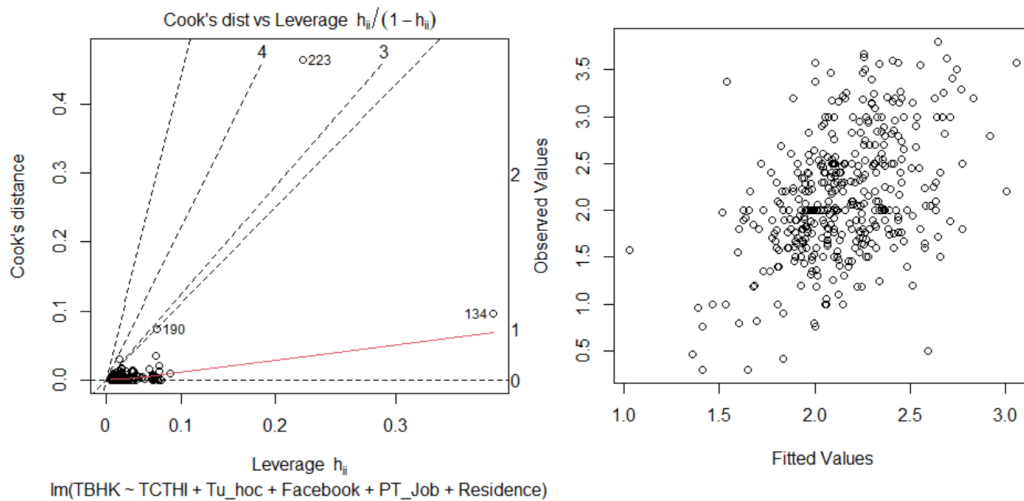
$$\begin{aligned} TBHK = & 1.207 + 2.769e-02 * TCTHI - 3.671e-02 * (Sport^{0.67}) - 1.472e-02 * (Tu_hoc^{2.75}) \\ & - 1.211e-05 * (XGames^{6.3}) + 3.676e-01 * (Tu_hoc) + 2.496e-02 * (Facebook^{1.4}) - 2.055e-01 \\ & * ResidenceR2 - 2.962e-01 * ResidenceR3 - 2.822e-01 * ResidenceR4 + 8.917e-03 * \\ & (Sport^{0.67}) * (Tu_hoc^{2.75}) - 1.164e-05 * (XGames^{6.3}) * (Tu_hoc) - 1.853e-02 * \\ & (Tu_hoc) * (Facebook^{1.4}) \end{aligned} \quad (6)$$



Hình 1. Biểu đồ phân phối thặng dư biến phụ thuộc của mô hình LM1: (trái) dạng miền và (phải) dạng cột

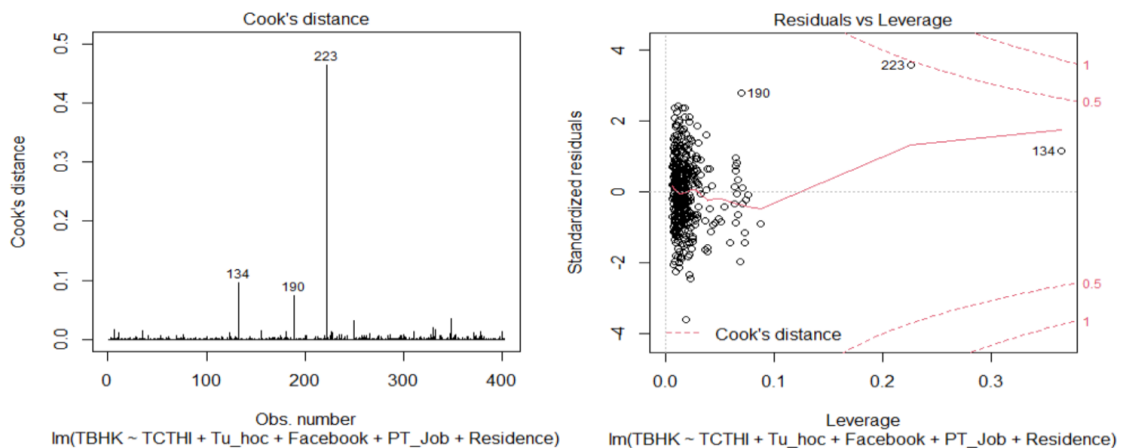


Hình 1. Biểu đồ giá trị phù hợp và thặng dư (trái) và biểu đồ Q-Q chuẩn tắc (phải) của thặng dư mô hình LM1



Hình 3. Biểu đồ khoảng cách Cook và leverage (trái) và biểu đồ giữa giá trị quan sát và giá trị phù hợp (phải) của LM1

Các thông số 2 test Shapiro–Wilk và ncv của hai mô hình này đều được đáp ứng tốt: với $W = 0.99678$, $p\text{-value} = 0.6093$, $\text{Chisquare} = 2.348949$, $Df = 1$, $p = 0.12537$, và $W = 0.99702$, $p\text{-value} = 0.6763$, $\text{Chisquare} = 2.769762$, $Df = 1$, $p = 0.09606$ tương ứng cho từng mô hình. Các thông số Multiple R-squared: 0.2882, Adjusted R-squared: 0.2662, F-statistic: 13.12 với 12 và 389 DF, $p\text{-value} < 2.2e-16$ đối với mô hình M0, và Multiple R-squared: 0.2886, Adjusted R-squared: 0.2666, F-statistic: 13.15 với 12 và 389 DF, $p\text{-value} < 2.2e-16$ cho mô hình M1. Tuy nhiên, hai mô hình này cho các thông số VIF (Variance inflation factor) ở các số hạng liên quan biến Tu_hoc .



Hình 4. Biểu đồ khoảng cách Cook (trái) và biểu đồ leverage và thặng dư chuẩn hóa (phải)

3. Kiểm định mô hình

Trong mục này, ta sẽ xem xét các phương pháp kiểm định nhằm đánh giá sự không phù hợp, **quá khớp** (overfitting) của mô hình. Phương pháp ở đây gồm hai hướng chính: Kiểm định sử dụng hai nhóm: Dữ liệu huấn luyện (Training data) và Dữ liệu kiểm chứng (Testing data); Kiểm định chéo (Cross Validation) sử dụng phương pháp k-cụm (k-folds) để chia dữ liệu thu thập thành k nhóm và xây dựng mô hình trên dữ liệu có nhóm thứ k bị loại đi và được dùng để kiểm chứng, cuối cùng dựa trên việc lấy giá trị trung bình kết quả thu được.

3.1. Phân chia hai nhóm dữ liệu: Training data và Testing data

Ta chia dữ liệu ban đầu làm 2 nhóm theo tỷ lệ 80% (training data dùng để xây dựng mô hình) và 20% (testing data dùng để kiểm nghiệm mô hình) một cách ngẫu nhiên. Ta tiến hành thực thi nhờ công cụ R, với tham số set.seed (1). Các test cơ bản Shapiro–Wilk và ncv cho kết quả đạt yêu cầu (thể hiện trong Bảng 2). Với các lựa chọn set.seed khác cho kết quả tương tự.

Bảng 2. Bảng các kết quả ncv Test và Shapiro Test cho các mô hình

Mô hình	Shapiro test	ncv test
LM0	W = 0.99699, p-value = 0.8175	Chisquare = 1.880531, Df = 1, p = 0.17027
LM1	W = 0.99717, p-value = 0.854	Chisquare = 2.705386, Df = 1, p = 0.10001
M1	W = 0.99727, p-value = 0.8721	Chisquare = 2.105556, Df = 1, p = 0.14676

Kiểm định mô hình trên Dữ liệu kiểm chứng được thực hiện thông qua hàm predict trong R. Ma trận tương quan (correlation) của độ chính xác giữa giá trị dự báo (predicted) từ mô hình và dữ liệu kiểm chứng (actuals) cho bởi hàm cor đối với mô hình LM1 cho giá trị $\text{cor}(\text{predicted}, \text{actuals}) = 0.4318$ tuy không quá lớn song phản ánh mức tin cậy đáng kể của tính đúng đắn cho mô hình LM1. Mô hình LM0, M0, M1 cũng cho kết quả khá tương đồng. *Độ chính xác min-max* (min-max accuracy): cho ta biết mô hình dự báo sai lệch bao nhiêu so với dữ liệu thực tế (Bảng 3). Các giá trị này gần với 1 cho thấy tính đúng đắn của các mô hình. *Độ lệch phần trăm trung bình tuyệt đối*: (Mean Absolute Percentage Error, MAPE) $\text{MAPE} = 100 \times \sum_{i=1}^n \frac{|y_{\text{predicted}} - y_{\text{actuals}}|}{n|y_{\text{actuals}}|}$, có miền giá trị từ 0 đến ∞ , giá trị càng bé càng phản ánh tính đúng đắn của mô hình (Bảng 3). Các giá trị khá bé cho thấy mức đáng tin cậy của các mô hình.

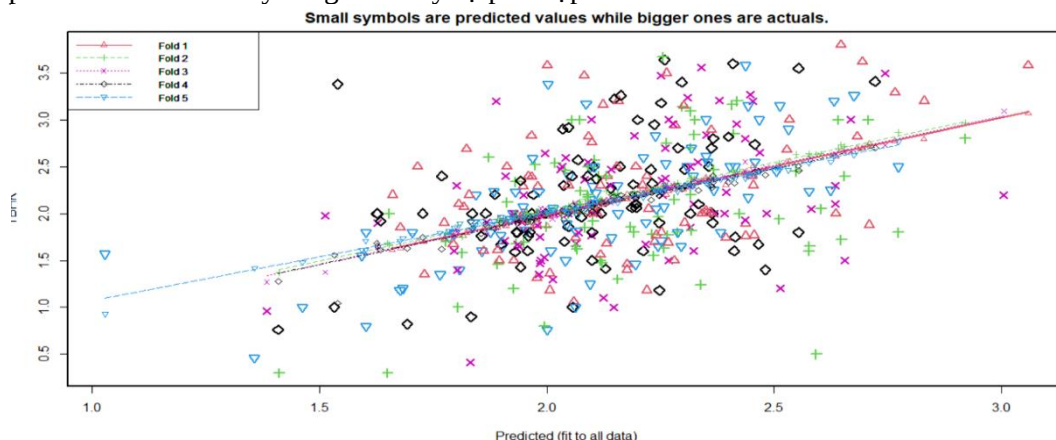
Bảng 3. Độ đo tính chính xác Min-max và Độ lệch phần trăm trung bình tuyệt đối các mô hình

Mô hình	Min-max accuracy	MAPE (%)
LM0	0.804943	29.95281
LM1	0.8028475	30.41654
M1	0.8210835	26.06734

Các thông số MAE, MSE, RMSE với mô hình LM1 lần lượt là 0.4629594, 0.3177434, 0.5636873 đều nhỏ. Các kết quả thu được cũng tương đồng cho mô hình LM0 và M0, M1.

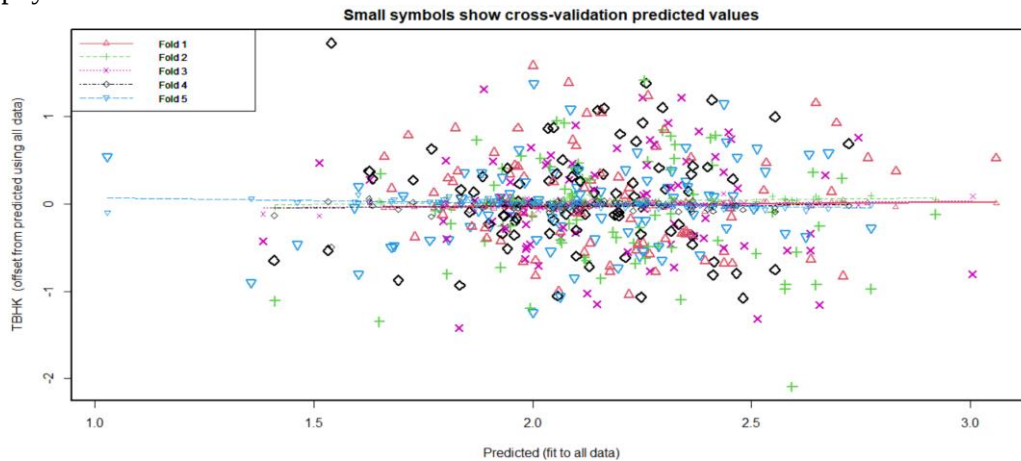
3.2. Kiểm định chéo k-folds

Kiểm định chéo 5-cụm (5-folds) được biểu diễn trong Hình 5. Ký hiệu nhỏ biểu thị giá trị thu được từ mô hình được xây dựng từ fold thứ k tương ứng và ký hiệu lớn là dữ liệu thực tế. Các ký hiệu cùng màu biểu thị chúng thuộc cùng một nhóm thứ k (fold thứ k). Các đường thẳng tương đối song song với nhau cho thấy các mô hình từ k folds khá giống nhau. Biểu diễn trực quan này cho thấy mô hình LM1 khá tốt. Ngoài ra, các ký hiệu nhỏ và các ký hiệu lớn cùng màu không có sự phân tán lớn. Điều này càng cho thấy sự phù hợp của mô hình LM1.



Hình 5. Biểu diễn trực quan Kiểm định chéo 5-folds cho mô hình LM1

Biểu diễn thặng dư Kiểm định chéo 5-folds cho mô hình LM1 thể hiện trong Hình 6. Mức phân tán thặng dư của dữ liệu quanh đường thẳng nằm ngang một cách ngẫu nhiên cho thấy tính hợp lý của mô hình.



Hình 6. Biểu diễn thặng dư Kiểm định chéo 5-folds cho mô hình LM

4. Kết luận

Các mô hình chứa các nhân tố tương tác M0, M1 đáp ứng yêu cầu với hai test Shapiro và ncv; có hệ số VIF khá cao, điều này lý giải quan hệ đa tuyến tính (multilinearity) trong các biến dự báo của các mô hình này lớn. Thông tin các biến dự báo cung cấp trong mô hình là dư thừa nhiều hơn (overfitting) so với các mô hình tuyến tính [8]. Do đó, mô hình M0, M1 có tính dự báo kém hơn.

Đối với mô hình LM1, các nhân tố chủ quan *TCTHI* ảnh hưởng trực tiếp đến biến phụ thuộc *TBHK*, tuy nhiên nhân tố này ít mang ý nghĩa đối với mục đích nghiên cứu. Các nhân tố gián tiếp là *Residence*, *Facebook*, *PT_Job* và *Tu_hoc* thể hiện vai trò quan trọng trong mô hình. Trong các nhân tố gián tiếp này, chỉ có hệ số của *Tu_hoc* là dương, 0.16253. Yếu tố này thậm chí còn có biên độ lớn hơn biên độ của hệ số của nhân tố trực tiếp *TCTHI*. Điều này khẳng định một chân lý rằng, tự học quyết định sự thành công lớn của quá trình đào tạo, vượt trên bất kể rào cản nào, nếu giả thiết khối lượng yêu cầu lao động (ở đây là số tín chỉ dự thi học kỳ) không vượt quá một ngưỡng giới hạn quá tải (tức số môn học không quá nhiều đến ngưỡng bất thường). Các yếu tố còn lại xem như những yếu tố tiêu cực (hệ số âm), trong đó *Residence* là nhân tố ảnh hưởng tiêu cực nhất. Điều này nên được định hướng tư tưởng rõ cho người học từ những ngày đầu tiên nhập trường. Đáng lưu ý, với các sinh viên ở cùng gia đình, *Residence* = R4, sự ảnh hưởng theo chiều hướng này càng sâu sắc. Để trả lời cho lý do của sự tác động tiêu cực này, có thể có 2 hướng giả thiết đưa ra: a) Những sinh viên ở cùng gia đình thường có học lực nền không cao vì đa phần họ là người trong thành phố Thái Nguyên, đây là một xu hướng chọn trường gần với học lực; b) Những sinh viên này ở cùng gia đình, có nhiều mối quan hệ từ trước, điều này chi phối nhiều đến sự nỗ lực vươn lên trong học tập của các sinh viên này. Xu hướng này cũng cho thấy Kỳ túc xá chính là môi trường lý tưởng nhất giúp các sinh viên đạt thành tích tốt trong học tập (hệ số của tham số *Residence* = R1 tương ứng là 0). Làm thêm, biến *PT_Job*, cũng là một nhân tố tác động tiêu cực đến kết quả học tập.

TÀI LIỆU THAM KHẢO/ REFERENCES

- [1] N. H. Albahtiti, R. Khazaei, T. H. Sasa, E. Almuher, and W. Alahmad. "Effects of Daily Activities on Academic Performance of Applied Science University Students," *Journal of Sociology and Anthropology*, vol. 4, no. 1, pp. 8-22, 2020.
- [2] C. C. Gavin and L. C. Talbot Nicola, "On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation," *Journal of Machine Learning Research*, vol. 11, pp. 2079-2107, 2010.

-
- [3] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statist. Surv.*, vol. 4, pp. 40-79, 2010.
 - [4] A. Airola, T. Pahikkala, W. Waegeman, B. De Baets, and T. Salakoski, "An experimental comparison of cross-validation techniques for estimating the area under the ROC curve," *Computational Statistics & Data Analysis*, vol. 55, no. 4, pp. 1828-1844, 2011.
 - [5] C. Ma, R. L. V. Florecilla Russell, C. Judith, and O. Rosella, "Student engagement: Defining teacher effectiveness and teacher engagement," *Journal of Institutional Research South East Asia*, vol. 15, pp. 5-19, 2017.
 - [6] A. C. Rencher and W. F. Christensen, *Methods of Multivariate Analysis*, Third Edition, John Wiley & Sons, Inc, 2002.
 - [7] J. Ding, V. Tarokh, and Y. Yang, "Bridging AIC and BIC: A New Criterion for Autoregression," *IEEE Transactions on Information Theory*, vol. 64, no. 6, pp. 4024-4043, 2018.
 - [8] A. M. Molinaro, R. Simon, and R. M. Pfeiffer, "Prediction error estimation: a comparison of resampling methods," *Bioinformatics*, vol. 21, no. 15, pp. 3301-3307, 2005.