

การเปรียบเทียบวิธีการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบการถดถอยลอจิสติกโดยใช้วิธีเบย์เซียนและวิธีการถดถอยแบบขั้นตอน

A comparison of independent variable selection in a logistic regression model using Bayesian variable selection and stepwise regression

กานต์ณัฐ ณ บางช้าง^{1*}, ภัทรภร ศรีสอด¹, เมธาพร สารศรี¹ และภูริภัทร สุขสี¹

Kannat Na Bangchang^{1*}, Phattharaphon Srisod¹, Methapohn Sarasri¹ and Phuripat Suksee¹

Received: 19 February 2025 ; Revised: 8 May 2025 ; Accepted: 20 June 2025

บทคัดย่อ

การคัดเลือกตัวแปรอิสระที่เหมาะสม จะทำให้ตัวแบบที่ได้มีประสิทธิภาพสูง โดยเฉพาะอย่างยิ่งในการวิเคราะห์การถดถอย งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาวิธีการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบการถดถอยลอจิสติกด้วยวิธีเบย์เซียน (Bayesian Method) โดยใช้วิธีการสุ่มแบบกิบส์ (Gibbs Sampling) กับวิธีการถดถอยแบบขั้นตอน (Stepwise Regression) และทำการเปรียบเทียบวิธีการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ ทั้งในกรณีที่มีความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระในระดับต่ำมากและสูงมาก ศึกษาโดยการจำลองข้อมูลและประยุกต์ใช้กับข้อมูลลักษณะพฤติกรรมการเล่นออนไลน์เพื่อวินิจฉัยการเป็นโรคอัลไซเมอร์ ในการจำลองข้อมูลเพื่อใช้ในการวิจัยนี้กำหนดขนาดตัวอย่างเท่ากับ 25 และ 100 โดยกระทำซ้ำ 100 ครั้งในแต่ละกรณี และทำการเปรียบเทียบประสิทธิภาพของวิธีการสุ่มแบบกิบส์และวิธีการถดถอยแบบขั้นตอน โดยเกณฑ์พิจารณา คือ ค่าพิกัดค่าเกณฑ์สารสนเทศของอะกะอิเกะ (AIC) , ค่าเกณฑ์สารสนเทศของของเบส์ (BIC) จำนวนแปรอิสระที่ถูกต้องค่าประมาณสัมประสิทธิ์ของตัวแปรอิสระที่ถูกต้องโดยเฉลี่ยและร้อยละความถูกต้อง ผลศึกษาพบว่าทั้งกรณีตัวอย่างขนาดเล็ก ($n = 25$) และตัวอย่างขนาดใหญ่ ($n = 100$) กรณีที่ข้อมูลมีความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระในระดับต่ำมาก วิธีการสุ่มแบบกิบส์และวิธีการถดถอยแบบขั้นตอนมีประสิทธิภาพในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบไม่แตกต่างกัน แต่ในกรณีที่ข้อมูลมีความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระในระดับสูงมาก วิธีการสุ่มแบบกิบส์มีประสิทธิภาพในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบดีกว่าวิธีการถดถอยแบบขั้นตอน โดยผลลัพธ์ที่ได้จากการวิจัยสามารถนำวิธีการคัดเลือกตัวแปรอิสระไปประยุกต์ใช้กับข้อมูลจริงที่ตัวแปรอิสระมีความสัมพันธ์เชิงเส้นพหุได้

คำสำคัญ การถดถอยลอจิสติก, ความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระ , วิธีการคัดเลือกตัวแปรโดยใช้วิธีเบย์เซียน, วิธีการถดถอยแบบขั้นตอน, วิธีการสุ่มแบบกิบส์

Abstract

Selecting appropriate independent variables yields a highly efficient model, particularly in regression analysis. This study aims to examine the selection of independent variables in a logistic regression model using Bayesian variable selection via Gibbs sampling and stepwise regression. It compares these two selection methods under conditions of very low and very high multicollinearity among independent variables. The study is conducted through data simulation and applied to online writing behavior data for diagnosing Alzheimer's disease. The sample sizes for the simulation were set to 25 and 100, with 100 replications for each case. The performances of Gibbs sampling and stepwise regression were compared based on evaluation criteria, including the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), number of correctly selected independent variables, average estimated coefficients of

¹ สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์

¹ Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University

* Corresponding author; E-mail: kannat@mathstat.sci.tu.ac.th

correctly selected variables, and accuracy percentage. The results indicate that for both small ($n = 25$) and large ($n = 100$) sample sizes, when the multicollinearity among independent variables is very low, Gibbs sampling and stepwise regression show no significant difference in their efficiency in selecting independent variables. However, in cases where multicollinearity is very high, Gibbs sampling demonstrates superior performance in selecting independent variables compared to stepwise regression. The findings of this research can be applied to select independent variables in real-world data exhibiting multicollinearity.

Keywords: Logistic regression, multicollinearity, Bayesian variable selection, stepwise regression, Gibbs sampling

บทนำ

การวิเคราะห์การถดถอยลอจิสติกทวิภาค (Binary Logistic Regression Analysis) เป็นเทคนิคการวิเคราะห์ตัวแปรเชิงพหุที่มีวัตถุประสงค์เพื่อประมาณค่าหรือทำนายเหตุการณ์ที่สนใจว่าจะเกิดหรือไม่เกิดเหตุการณ์นั้นภายใต้อิทธิพลของตัวแปรอิสระ โดยตัวแบบประกอบด้วยตัวแปรตามที่เป็นตัวแปรแบบทวิภาค คือ มีได้สองค่า เช่น “ใช่” กับ “ไม่ใช่” เป็นต้น ซึ่งอาจจะมีตัวแปรอิสระที่เกี่ยวข้องทั้งแบบตัวเดียวหรือหลายตัว ที่เป็นไปได้ทั้งตัวแปรเชิงกลุ่มหรือตัวแปรแบบต่อเนื่อง (ภคสุภางค์ มาปรีดา, 2560)

การคัดเลือกตัวแปรเป็นกระบวนการหนึ่งในการวิเคราะห์ทางสถิติ และมีความจำเป็นต่องานวิจัย ช่วยให้นักวิจัยได้ค้นหาปัจจัยที่มีความสำคัญและมีอิทธิพลต่อเรื่องที่นักวิจัยสนใจ (พจนา แวสวีสต์, 2547) ดังนั้นการคัดเลือกตัวแปรที่เหมาะสมจึงเป็นสิ่งสำคัญในการพัฒนาตัวแบบให้มีประสิทธิภาพและมีความถูกต้องในการทำนายตัวแปรตาม อีกทั้งยังลดความซับซ้อน ทรัพยากร และเวลาในการวิเคราะห์การถดถอย เนื่องจากการที่ตัวแบบมีตัวแปรอิสระมาก ไม่ได้หมายความว่าตัวแปรอิสระจะมีความสัมพันธ์กับตัวแปรตามสูง อาจทำให้เกิดปัญหาความสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระ (Multicollinearity) และถ้าตัวแบบมีตัวแปรอิสระที่น้อยเกินไปก็อาจส่งผลให้เกิดความผิดพลาดในการทำนายตัวแปรตาม ดังนั้นตัวแบบการถดถอยลอจิสติกจะต้องประกอบไปด้วยตัวแปรอิสระที่มีความสัมพันธ์เฉพาะกับตัวแปรตาม และทำให้มีค่าประมาณความเอนเอียงน้อยที่สุด (ธนัญญ์ เกิดผล, 2559)

เทคนิคการคัดเลือกตัวแปรอิสระมีหลายวิธี เช่น วิธีการคัดเลือกตัวแปรแบบดั้งเดิม ประกอบไปด้วย วิธีนำตัวแปรเข้าทั้งหมด (Enter Regression) วิธีเพิ่มตัวแปร (Forward Selection) วิธีลดตัวแปร (Backward Elimination) และ วิธีการถดถอยแบบขั้นตอน (Stepwise Regression) เป็นต้น โดยเทคนิควิธีการข้างต้นใช้แนวคิดการหาค่าประมาณด้วยวิธีกำลังสองน้อยที่สุด (OLS) ซึ่งจะทำให้ได้ก็ต่อเมื่อข้อมูลมีจำนวนตัวแปรอิสระน้อยกว่าขนาดตัวอย่าง แต่เมื่อข้อมูลมีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง จะไม่สามารถประมาณค่าด้วยวิธีกำลังสองน้อยที่สุดได้ เนื่องจากเมทริกซ์จัตุรัส ($X'X$)

จะเป็นเมทริกซ์เอกฐาน (Singular Matrix) ทำให้ไม่สามารถหาเมทริกซ์ผกผัน (Inverse Matrix) ได้ นอกจากนี้ค่าประมาณที่ได้จะไม่เสถียรและไม่มีประสิทธิภาพ เนื่องจากปัญหาความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระ (Multicollinearity) อีกทั้งตัวแบบที่ได้จะมีจำนวนตัวแปรอิสระสูง จึงทำให้มีความซับซ้อนและอธิบายความสัมพันธ์ได้ยาก (Müller *et al.*, 2013) ดังนั้นกรณีที่มีตัวแปรอิสระมากกว่าขนาดตัวอย่าง วิธีคัดเลือกตัวแปรข้างต้นอาจจะยังไม่เหมาะสมกับข้อมูล

ผู้วิจัยจึงเลือกศึกษาการคัดเลือกตัวแปรด้วยวิธีเบย์เซียน (Bayesian Variable Selection) ซึ่งใช้วิธีการสุ่มแบบกิบส์ (Gibbs Sampling) เนื่องจากสนใจหลักการของเบย์เซียนเพื่อให้ได้ตัวแบบที่เหมาะสมกับข้อมูลโดยงานวิจัยนี้จะทำการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบการถดถอยลอจิสติกทวิภาค (Binary Logistics Regression) ด้วยวิธีของเบย์เซียน (Bayesian Variable Selection) โดยใช้วิธีการสุ่มแบบกิบส์ (Gibbs Sampling) กับวิธีการถดถอยแบบขั้นตอน (Stepwise Regression) และทำการเปรียบเทียบประสิทธิภาพในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบโดยจำลองข้อมูล ซึ่งกำหนดจำนวนตัวแปรอิสระในตัวแบบเต็มรูปแบบ (Full Model) เท่ากับ 200 ตัวแปร กำหนดจำนวนตัวแปรอิสระที่ส่งผลต่อตัวแปรตาม 10 ตัวแปร คือ เป็นตัวแปรที่มีค่าสัมประสิทธิ์การถดถอยเท่ากับ 2 และขนาดตัวอย่างที่ใช้ในการศึกษาเท่ากับ 25 และ 100 ตามลำดับ เพื่อให้ครอบคลุมทั้งกรณีที่ตัวแปรอิสระมากกว่าขนาดตัวอย่าง และขนาดตัวอย่างมากกว่าตัวแปรอิสระ โดยกระทำซ้ำ 100 ครั้งในแต่ละกรณี ทั้งกรณีที่ความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระในระดับต่ำมากและสูงมาก

วัตถุประสงค์ของการศึกษา

เพื่อศึกษาวิธีการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบการถดถอยลอจิสติกด้วยวิธีเบย์เซียนโดยใช้วิธีการสุ่มแบบกิบส์กับวิธีการถดถอยแบบขั้นตอน และเปรียบเทียบวิธีการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบทั้งในกรณีไม่มีและมีความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระ

วิธีการดำเนินงานวิจัย

1. แผนการดำเนินงานวิจัย

1.1 กำหนดตัวแบบแท้จริงโดยใช้ตัวแปรอิสระจำนวน 10 ตัวแปรที่มีความสัมพันธ์กับตัวแปรตามซึ่งตัวแบบเต็มรูปมีตัวแปรอิสระทั้งหมด 200 ตัวแปร โดยเริ่มกำหนดความคลาดเคลื่อนสุ่มที่เป็นอิสระต่อกัน โดยที่ $e \sim \text{Logistic}(0,1)$ ซึ่งกำหนดให้ตัวแปรอิสระมีการแจกแจงดังนี้

$$x_i \sim N(0_{p \times 1}, \Sigma_{p \times p}(\rho)) \quad (1)$$

1.2 การสุ่มข้อมูลของตัวแปรอิสระในแต่ละกรณี

กรณีที่ 1 ไม่มีความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระ ทำการสุ่มข้อมูลของตัวแปรอิสระ ทั้งกรณีขนาดตัวอย่างเท่ากับ 25 และ 100 ตามลำดับ

กรณีที่ 2 มีความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระ ทำการสุ่มข้อมูลของตัวแปรอิสระ ทั้งกรณีขนาดตัวอย่างเท่ากับ 25 และ 100 ตามลำดับ และสร้างตัวอิสระบางตัวให้มีความสัมพันธ์กัน (ประพันธ์ กล่อมพร, 2560) โดยกำหนดให้ตัวแปรอิสระ x_1 กับ x_2 และตัวแปรอิสระ x_{11} กับ x_{12} ความสัมพันธ์เชิงเส้นพหุในระดับสูงมาก ($r \geq 0.91$)

1.3 กำหนดค่าสัมประสิทธิ์การถดถอย β_j ที่ไม่เท่ากับ 0 จำนวน 10 ตัว โดยให้มีค่า เป็น 2 เพื่อให้เห็นถึงผลกระทบของตัวแปรอิสระนั้นที่มีต่อตัวแปรตาม ดังนี้

$$\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = 2$$

และกำหนดให้สัมประสิทธิ์การถดถอยที่เหลือมีค่าเป็น 0

1.4 จำลองตัวแปรตาม (y) ให้มีค่าเป็น 0 กับ 1 โดยใช้ Latent Variable Model (Polson et al., 2013)

$$y = \begin{cases} 0 & \text{เมื่อ } \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon \leq 0 \\ 1 & \text{เมื่อ } \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon > 0 \end{cases}$$

$$\text{โดยที่ } \varepsilon \sim \text{Logistic}(0,1) \quad (2)$$

ตัวแปรอิสระและตัวแปรตามมีความสัมพันธ์กันภายใต้ตัวแบบการถดถอยลอจิสติกแบบ 2 กลุ่ม (Binary Logistic Regression) ซึ่งเขียนในรูปสมการดังนี้

$$P_y = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}} \quad (3)$$

โดยที่ P_y คือ ค่าความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ

1.5 ทำการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบการถดถอยลอจิสติกโดยวิธีการคัดเลือกตัวแปรอิสระ 2 วิธี คือวิธีการสุ่มแบบกิบส์ และวิธีการถดถอยแบบขั้นตอน

1.6 คำนวณค่า AIC ค่า BIC จากนั้นพิจารณาจำนวนแปรอิสระที่ถูกต้อง ค่าประมาณสัมประสิทธิ์ของตัวแปรอิสระที่ถูกต้องโดยเฉลี่ย และร้อยละความถูกต้องของตัวแบบที่ได้จากการคัดเลือกตัวแปรอิสระทั้ง 2 วิธี

1.7 ทำการจำลองข้อมูลซ้ำในแต่ละกรณีจำนวน 100 รอบ

2. ขั้นตอนการดำเนินงานวิจัย

2.1 ขั้นตอนการคัดเลือกตัวแปรด้วยวิธีการถดถอยแบบขั้นตอน (Stepwise Regression) (Arkes, 2023)

2.1.1 การคัดเลือกตัวแปรแบบนำเข้า (Forward Stepwise Selection) โดยใช้ระดับนัยสำคัญเท่ากับ 0.05

กระบวนการนี้เริ่มต้นโดยไม่มีตัวแปรอิสระใด ๆ อยู่ในตัวแบบ จากนั้นจึงทดสอบนำตัวแปรที่มีความสัมพันธ์กับตัวแปรตามเข้าในตัวแบบทีละตัว โดยพิจารณาตัวแปรที่สามารถลดค่า -2 Log-Likelihood (-2LL) ได้มากที่สุดอย่างมีนัยสำคัญ ตัวแปรที่ถูกเลือกเข้ามานี้จะต้องมีคุณสมบัติในการอธิบายความแปรปรวนของตัวแปรตามได้ดีขึ้น กระบวนการดังกล่าวจะทำซ้ำไปเรื่อย ๆ จนกระทั่งไม่มีตัวแปรใด ๆ ที่สามารถลดค่า -2LL ได้อย่างมีนัยสำคัญเพิ่มเติม

2.1.2 การคัดเลือกตัวแปรแบบคัดออก (Backward Stepwise Selection) โดยใช้ระดับนัยสำคัญเท่ากับ 0.05 กระบวนการนี้เริ่มต้นโดยการนำตัวแปรอิสระทั้งหมดเข้าสู่ตัวแบบ จากนั้นจึงทดสอบการคัดเลือกตัวแปรที่ควรถูกนำออก โดยพิจารณาตัวแปรที่ส่งผลต่อค่า -2LL น้อยที่สุดเมื่อนำออก หากการนำตัวแปรใดออกส่งผลให้ค่า -2LL ลดลงเล็กน้อยหรือไม่มีนัยสำคัญ ตัวแปรนั้นจะถูกคัดออกจากตัวแบบ กระบวนการนี้จะดำเนินต่อไปจนกว่าจะเหลือตัวแปรเฉพาะที่มีนัยสำคัญในการอธิบายตัวแปรตามและทำให้ค่า -2LL ต่ำที่สุด

2.1.3 การทำซ้ำคัดเลือกตัวแปรแบบนำเข้าและคัดออกทั้งสองวิธีดังกล่าวใช้กระบวนการทำซ้ำในการปรับปรุงตัวแบบ พิจารณาตัวแปรที่เพิ่มและลบเข้าออกจากตัวแบบอย่างต่อเนื่อง เพื่อลดค่า -2LL ให้ต่ำที่สุด การทำซ้ำนี้จะหยุดเมื่อไม่มีตัวแปรใด ๆ ที่สามารถลดค่า -2LL ได้เพิ่มเติมอย่างมีนัยสำคัญ ทำให้ได้ตัวแบบที่ดีที่สุด ซึ่งสามารถอธิบายความแปรปรวนของตัวแปรตามได้อย่างมีประสิทธิภาพ

2.2 ขั้นตอนการคัดเลือกตัวแปรด้วยวิธีของเบส์ โดยการสุ่มแบบกิบส์ (Tian *et al.*, 2019)

2.2.1 การหาตัวแปรตาม Y จำเป็นต้องทราบความน่าจะเป็นก่อน (Prior) ของ β และ τ ซึ่งต้องกำหนดความน่าจะเป็นก่อนของ y เพื่อใช้สำหรับการคัดเลือกตัวแปร และจึงจะสามารถกำหนดความน่าจะเป็นของ β และ τ ได้สมการ ดังนี้

$$\beta_i | y_i \sim (1 - y_i)N(0, \tau) + y_i N(0, \tau)$$

กำหนดการแจกแจงความน่าจะเป็นก่อนของ g

$$y \sim \text{Ber}(p_i); i = 1, 2, \dots, k \quad (5)$$

$$\text{โดยให้ } y = [y_1, \dots, y_j, \dots, y_p]^T$$

เมื่อ y_j เป็นตัวแปรที่ใช้เป็นตัวบ่งบอกว่าตัวแปรอิสระที่ j อยู่ในตัวแบบการถดถอยลอจิสติกหรือไม่ โดยที่

$y_j = 1$ เมื่อตัวแปรอิสระตัวที่ j อยู่ในตัวแบบการถดถอยลอจิสติก

$y_j = 0$ เมื่อตัวแปรอิสระตัวที่ j ไม่อยู่ในตัวแบบการถดถอยลอจิสติก

กำหนดการแจกแจงก่อนของ β ให้มีการแจกแจงแบบปกติโดยมีเงื่อนไขอยู่บน τ, y, y

$$\beta | \tau, y, y \sim N\left(\Sigma(\tau)\left(x_r'(y - \frac{1}{2} \mathbf{1}_n)\right), (x_r' \tau x_r + \tau^{-1})^{-1}\right) \quad (6)$$

กำหนดการแจกแจงก่อนของ τ ที่มีการแจกแจงแกมมา (Gamma distribution) ดังนี้

$$\tau = \begin{bmatrix} \tau_0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \tau_p \end{bmatrix}$$

$$\tau_j | \lambda, d, y \sim \text{Gamma}\left(\frac{3}{p}, \frac{1}{d^2}\right) \quad (7)$$

2.2.2 ใช้การสุ่มแบบกิบส์เริ่มต้นการสุ่มแบบกิบส์โดยสร้างการสุ่มของ y แล้วจึงสร้างลำดับการสุ่มเป็นดังนี้

$$y^0, \beta^0, \tau^0, y^1, \beta^1, \tau^1, \dots, y^j, \beta^j, \tau^j$$

ซึ่งจะได้ค่าออกมาทีละตัวตามลำดับ หลังจากนั้นคำนวณความน่าจะเป็นภายหลังของ y ที่มีความน่าจะเป็นหลังสูงสุดเข้าสู่ตัวแบบเป็นอันดับแรก และมีค่าสูงรองลงมาเข้าสู่ตัวแบบ ตามลำดับ

3. เกณฑ์ที่ใช้ในการเปรียบเทียบ

3.1 เกณฑ์สารสนเทศของอะกะอิเกะถูกนำเสนอโดย Akaike (1978) เพื่อคัดเลือกตัวแบบที่ให้ค่าพยากรณ์แม่นยำที่สุด พิจารณาจากการประมาณค่าความคลาดเคลื่อนรวมกับข้อสนเทศ (Information) ของค่าสังเกตและใช้แนวคิดจากการหาค่าต่ำสุดของข้อสนเทศด้วยหลักการคูลส์เบ็ค-ไลท์เบอร์ (Kullback-Leibler Information) ข้อตกลงเบื้องต้นคือตัวประมาณของสัมประสิทธิ์การถดถอยได้มาจากวิธีภาวะน่าจะเป็นสูงสุด และยังเป็นวิธีการคัดเลือกตัวแบบที่เหมาะสมภายใต้ตัวแบบที่มีการใช้ข้อมูลชุดเดียวกัน แต่มีจำนวนตัวแปรอิสระที่แตกต่างกันใช้อธิบายความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระชุดหนึ่ง ตัวแบบที่เหมาะสมกว่าจะมีค่า AIC ต่ำกว่า โดยมีสูตรดังนี้

$$\text{AIC} = -2(\log \text{likelihood}) + 2p \quad (8)$$

โดยที่ $\log$ = ลอการิทึมธรรมชาติ (ln)

likelihood = ฟังก์ชันภาวะน่าจะเป็นสูงสุด (Maximum

Likelihood Function)

p = จำนวนตัวแปรอิสระที่ถูกเลือกเข้าตัวแบบ

3.2 เกณฑ์ข้อสนเทศของเบส์ (BIC) ถูกนำเสนอโดยซาว่า (Sawa, 1978) โดยพัฒนามาจากเกณฑ์ AIC ทั้งนี้เกณฑ์ BIC จะเลือกตัวแบบที่ให้ค่า BIC ต่ำที่สุดเป็นตัวแบบที่ถูกต้อง

$$\text{BIC} = -2(\log \text{likelihood}) + p \cdot \log(n) \quad (9)$$

โดยที่ $\log$ = ลอการิทึมธรรมชาติ (ln)

likelihood = ฟังก์ชันภาวะน่าจะเป็นสูงสุด

(Maximum Likelihood Function)

p = จำนวนตัวแปรอิสระที่ถูกเลือกเข้าตัวแบบ

n = จำนวนขนาดตัวอย่าง

3.3 การนับจำนวนแปรอิสระที่ถูกต้องที่เข้าสู่ตัวแบบ โดยพิจารณาจำนวนตัวแปรอิสระที่คัดเลือกเข้าสู่ตัวแบบได้ตรงกับเงื่อนไขที่กำหนดไว้ได้แก่

$$X_1, X_3, X_4, X_5, X_{11}, X_{12}, X_{13}, X_{14}, X_{15}$$

เช่น ภายในตัวแบบคัดเลือกตัวแปรอิสระมาทั้งหมด

5 ตัวแปร ได้แก่ $X_1, X_4, X_7, X_{10}, X_{11}$ จะนับจำนวนตัวแปรอิสระที่ถูกต้อง (ส่งผลต่อตัวแปรตาม) ได้ 3 ตัวแปร เป็นต้น

3.4 ค่าประมาณสัมประสิทธิ์ของตัวแปรอิสระที่ถูกตัดโดยเฉลี่ย

ค่าสัมประสิทธิ์ที่ประมาณได้ของตัวแปรอิสระที่ถูกต้องนำมาหาค่าเฉลี่ย เพื่อแสดงความสำคัญของตัวแปรอิสระต่อตัวแปรตาม ตัวอย่างเช่น หากพิจารณาค่าสัมประสิทธิ์ของตัวแปร $X_1, X_2, X_3, X_{11}, X_{20}$ ซึ่งมีค่าเท่ากับ 49.567, 17.753, 98.665, 84.467, 46.643 และ 78.123 ตามลำดับ จะได้ค่าประสิทธิภาพของตัวแปรอิสระที่ถูกต้องโดยเฉลี่ย 50.595

3.5 จำนวนร้อยละความถูกต้อง

กรณีที่ไม่มีความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระ จะพิจารณาแค่กรณีเดียว คือ เมื่อตัวแบบเลือกตัวแปรอิสระ ได้แก่

$$\overline{X_1, X_2, X_3, X_4, X_5, X_{11}, X_{12}, X_{13}, X_{14}, X_{15}}$$

เข้าสู่ตัวแบบอย่างน้อยหนึ่งตัวแปร จึงจะถือว่าคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้อย่างถูกต้อง

กรณีที่มีความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระ จะพิจารณาตัวแบบที่เลือกตัวแปรอิสระเข้าสู่ตัวแบบอย่างน้อยหนึ่งตัวแปร โดยที่ตัวแบบไม่เลือกตัวแปรอิสระ x_1x_1 กับ x_2x_2 และตัวแปรอิสระ $x_{11}x_{11}$ กับ $x_{12}x_{12}$ เข้าสู่ตัวแบบพร้อมกัน เนื่องจากตัวแปรอิสระข้างต้น มีความสัมพันธ์เชิงเส้นพหุในระดับสูงมาก จึงจะถือว่าคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้อย่างถูกต้อง

4. ขั้นตอนนำไปใช้กับข้อมูลจริง

งานวิจัยนี้ใช้ข้อมูลลักษณะพฤติกรรมกรรมการเขียนออนไลน์เพื่อวินิจฉัยการเป็นโรคอัลไซเมอร์ (Cilia et al., 2022) ศึกษาจากผู้เข้าร่วมทดสอบจำนวน 174 ราย ในประเทศอิตาลี ปี ค.ศ. 2022 โดยมีตัวแปรตามมี 1 ตัวแปร ที่มีค่าได้ 2 ค่า คือ การเป็นโรคอัลไซเมอร์ (y) โดยกำหนด $y = 1$ แทนเป็นโรคอัลไซเมอร์ มีผู้เข้าร่วมทดสอบทั้งหมดจำนวน 89 ราย และ $y = 0$ แทนไม่เป็นโรคอัลไซเมอร์ ($y = 1$) มีผู้เข้าร่วมทดสอบทั้งหมดจำนวน 85 ราย

ตัวแปรอิสระจำนวน 450 ตัวแปร คือ ลักษณะพฤติกรรมกรรมการเขียนออนไลน์ เช่น เวลาที่ปากกาอยู่ในอากาศ, ดัชนีการกระจายของปลายมือ, อัตราการเปลี่ยนแปลงความเร็วของปากกาเมื่ออยู่ในอากาศระหว่างการเขียน, อัตราการเปลี่ยนแปลงของความเร็วในการเคลื่อนที่ของปากกาเมื่อสัมผัสกับกระดาษระหว่างการเขียน, ค่าเฉลี่ยของแรงกดปากกา, ค่าเฉลี่ยของมุมกดปากกา เป็นต้น

แบ่งข้อมูลจริงของตัวแปรอิสระออกเป็น 2 กรณี

4.1 กรณีที่มีความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระในระดับที่ต่ำมาก

4.1.1 การปรับมาตรฐานข้อมูล

การปรับมาตรฐานข้อมูลเป็นการปรับค่าของข้อมูลให้อยู่ในช่วงมาตรฐาน โดยมีค่าเฉลี่ยเท่ากับ 0 และส่วนเบี่ยงเบนมาตรฐานเท่ากับ 1 ซึ่งช่วยให้สามารถเปรียบเทียบตัวแปรที่มีหน่วยวัดต่างกันได้ (บุญชม ศรีสะอาด, 2556) โดยคำนวณได้จากสูตร ดังนี้

$$Z_i = \frac{x_i - \mu}{\sigma} \quad (10)$$

โดย Z_i คือ ค่าที่ได้หลังจากการปรับมาตรฐานข้อมูล

x_i คือ ค่าของข้อมูลแต่ละตัว

μ คือ ค่าเฉลี่ยของข้อมูล

σ คือ ส่วนเบี่ยงเบนมาตรฐานของข้อมูล

4.1.2 คัดเลือกตัวแปรอิสระที่มีค่าสัมประสิทธิ์สหสัมพันธ์ต่อกันในระดับที่ต่ำมาก ($r \leq 0.01$ หรือ $r \geq 0.01$)

คัดเลือกเฉพาะตัวแปรอิสระคู่ที่มีค่าสัมประสิทธิ์สหสัมพันธ์ต่ำกว่า 0.01 โดยค่าสัมประสิทธิ์สหสัมพันธ์นั้นเป็นการวัดความสัมพันธ์เชิงเส้นระหว่างตัวแปรสองตัวซึ่งคำนวณได้จากสูตร ดังนี้

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (11)$$

โดย r_{xy} คือ ค่าสหสัมพันธ์ระหว่างตัวแปร X และ Y (correlation)

x_i คือ ค่าข้อมูลในแต่ละจุดของตัวแปร X

y_i คือ ค่าข้อมูลในแต่ละจุดของตัวแปร Y

$\bar{x}$ คือ ค่าเฉลี่ยของตัวแปร X

$\bar{y}$ คือ ค่าเฉลี่ยของตัวแปร Y

4.2 กรณีที่มีความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระในระดับสูงมาก นำข้อมูลของตัวแปรอิสระจากกลุ่มตัวอย่างผู้ป่วยโรคอัลไซเมอร์และกลุ่มที่ไม่มีผู้ป่วยโรคอัลไซเมอร์มาใช้ในการคัดเลือกตัวแปรอิสระ โดยไม่มีการปรับข้อมูล เนื่องจากชุดข้อมูลนี้มีจำนวนตัวแปรอิสระจำนวนมากที่มีความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระในระดับที่สูงมาก

ผลการวิจัย

ผลการวิจัยนำเสนอโดยใช้รูปแบบตารางซึ่งแบ่งเป็น 2 กรณีหลัก คือ กรณีที่มีขนาดตัวอย่าง (n) เท่ากับ 25 และ

100 และแบ่งพิจารณาตามความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระเพิ่มเติมเป็น 2 กรณีย่อย คือ กรณีที่มีความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระในระดับที่ต่ำมากและกรณีที่ความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระในระดับที่สูงมาก โดยกำหนดให้ตัวแปรอิสระ X_1 กับ X_2 และ X_{11} กับ X_{12} มีความสัมพันธ์เชิงเส้นพหุในระดับสูงมาก โดยใช้วิธี Cholesky Decomposition (ประพันธ์ กล่อมพร, 2560) โดยกำหนดค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระ X_1 กับ X_2 และ X_{11} กับ X_{12} มีค่าตั้งแต่ 0.91 ขึ้นไป เพื่อทำการเปรียบเทียบประสิทธิภาพในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบของทั้งสองวิธีสำหรับในกรณีที่ความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระในระดับต่ำมากและสูงมาก โดยใช้เกณฑ์สารสนเทศของอะกะอิเกะ (AIC) เกณฑ์สารสนเทศของ

เบส์ (BIC) จำนวนร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระ จำนวนตัวแปรอิสระที่เลือกถูกต้อง และค่าประมาณสัมประสิทธิ์การถดถอยโดยเฉลี่ย

ผลการวิจัย

กรณีข้อมูลจากการจำลอง

กรณีนี้นำเสนอผลลัพธ์ที่ได้จากข้อมูลจากการจำลอง โดยเปรียบเทียบค่าร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระระหว่างวิธีการถดถอยแบบขั้นตอนและวิธีการคัดเลือกตัวแปรโดยใช้วิธีเบส์เซียน โดยการสุ่มแบบกิบส์ ในกรณีที่ความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระในระดับต่ำมากและสูงมาก โดยแบ่งออกเป็น 2 ส่วน คือ ตัวอย่างขนาดเล็กและตัวอย่างขนาดใหญ่ แสดงดังใน Table 1

Table 1 The comparison of AIC, BIC, the number of correctly selected independent variables, the percentage of correctly selected variables, and the average estimated coefficients for small sample sizes ($n = 25$) and large sample sizes ($n = 100$)

n	Correlation coefficient	Method	AIC	BIC	The average number of correctly selected variables	The average estimated coefficients	The percentage of correctly selected variables
25	0.01	Stepwise	12.42	20.04	4.19	306.21	99
		Gibbs	50.00	80.47	8.83	5.51	100
	$r \geq 0.91$	Stepwise	11.18	17.99	4.29	338.63	3
		Gibbs	47.60	76.60	8.96	3.91	17
100	0.01	Stepwise	42.93	99.48	4.19	354.52	100
		Gibbs	190.86	408.67	8.83	8.35	100
	$r \geq 0.91$	Stepwise	34	73.94	4.29	572.96	15
		Gibbs	171.87	368.78	8.96	6.87	93

ส่วนที่ 1 สำหรับข้อมูลขนาดเล็ก

กรณีที่ความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระในระดับต่ำมาก ผลจาก Table 1 พบว่าค่า AIC และ BIC ของวิธีการสุ่มแบบกิบส์เท่ากับ 50.00 และ 80.47 ตามลำดับ ส่วนค่า AIC และ BIC ของวิธีการถดถอยแบบขั้นตอนเท่ากับ 12.42 และ 20.40 ตามลำดับ แสดงให้เห็นว่า ค่า AIC และ BIC ของวิธีการสุ่มแบบกิบส์สูงกว่าวิธีการถดถอยแบบขั้นตอน โดยต่างกันที่ 37.58 สำหรับค่า AIC และ 60.07 สำหรับค่า BIC

ค่าประมาณสัมประสิทธิ์ของตัวแปรอิสระที่ถูกต้องโดยเฉลี่ยสำหรับวิธีการสุ่มแบบกิบส์อยู่ที่ 5.51 ส่วนวิธีการถดถอยแบบขั้นตอนมีค่าเฉลี่ยที่ 306.21 ซึ่งทั้งสองวิธีมีความแตกต่างจากค่าสัมประสิทธิ์ที่กำหนดไว้ คือ 2 โดยมีค่าส่วนต่าง 3.51 และ 304.21 ตามลำดับ

การคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบของวิธีการสุ่มแบบกิบส์มีความถูกต้องที่ร้อยละ 100 และวิธีการถดถอยแบบขั้นตอนมีความถูกต้องร้อยละ 99

กรณีที่ความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระในระดับสูงมาก พบว่าค่า AIC และ BIC ของวิธีการสุ่มแบบกิบส์เท่ากับ 47.60 และ 76.60 ตามลำดับ ขณะที่ค่า AIC และ BIC ของวิธีการถดถอยแบบขั้นตอนเท่ากับ 11.8 และ 17.99 ตามลำดับ แสดงให้เห็นว่าค่า AIC และ BIC ของวิธีการสุ่มแบบกิบส์ยังคงสูงกว่าวิธีการถดถอยแบบขั้นตอน โดยแตกต่างกัน 35.8 สำหรับค่า AIC และ 58.61 สำหรับค่า BIC

ค่าประมาณสัมประสิทธิ์ของตัวแปรอิสระที่ถูกต้องโดยเฉลี่ยสำหรับวิธีการสุ่มแบบกิบส์เท่ากับ 3.91 ส่วนวิธีการถดถอยแบบขั้นตอนมีค่าเฉลี่ยเท่ากับ 338.63 ซึ่งทั้งสองวิธีมีความแตกต่างจากค่าสัมประสิทธิ์ที่กำหนดไว้ คือ 2 โดยมีค่าส่วนต่าง 1.91 และ 336.63 ตามลำดับ

การคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบของวิธีการ
สุ่มแบบกิบส์มีความถูกต้องร้อยละ 3 ส่วนวิธีการถดถอยแบบ
ขั้นตอนมีความถูกต้องร้อยละ 17

ส่วนที่ 2 สำหรับข้อมูลขนาดใหญ่

กรณีที่มีความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปร
อิสระในระดับต่ำมาก ผลจาก Table 1 พบว่า ค่า AIC และ
BIC ของวิธีการสุ่มแบบกิบส์เท่ากับ 190.86 และ 408.67 ตาม
ลำดับ ส่วนค่า AIC และ BIC ของวิธีการถดถอยแบบขั้นตอน
เท่ากับ 42.93 และ 99.48 ตามลำดับ แสดงให้เห็นว่า ค่า AIC
และ BIC ของวิธีการสุ่มแบบกิบส์สูงกว่าวิธีการถดถอยแบบ
ขั้นตอน โดยต่างกันที่ 147.93 สำหรับค่า AIC และ 309.19
สำหรับค่า BIC

วิธีการสุ่มแบบกิบส์มีจำนวนตัวแปรอิสระที่ถูก
ต้องในตัวแบบเท่ากับ 8.83 ขณะที่วิธีการถดถอยแบบขั้นตอน
มีจำนวนตัวแปรอิสระที่ต้องเท่ากับ 4.19 โดยวิธีการสุ่มแบบ
กิบส์ให้จำนวนตัวแปรอิสระที่ต้องมากกว่า 4.64

ค่าประมาณสัมประสิทธิ์ของตัวแปรอิสระที่ต้อง
โดยเฉลี่ยสำหรับวิธีการสุ่มแบบกิบส์อยู่ที่ 8.35 ส่วนวิธีการ
ถดถอยแบบขั้นตอนมีค่าเฉลี่ยที่ 354.52 ซึ่งทั้งสองวิธีมีความ
แตกต่างจากค่าสัมประสิทธิ์ที่กำหนดไว้ คือ 2 โดยมีค่าส่วน
ต่าง 6.35 และ 352.52 ตามลำดับ

การคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบของทั้งสอง
วิธีมีความถูกต้องที่ร้อยละ 100

กรณีที่มีความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปร
อิสระในระดับสูงมาก พบว่าค่า AIC และ BIC ของวิธีการสุ่ม
แบบกิบส์เท่ากับ 171.87 และ 368.78 ตามลำดับ ขณะที่ค่า
AIC และ BIC ของวิธีการถดถอยแบบขั้นตอนเท่ากับ 34 และ
73.94 แสดงให้เห็นว่าค่า AIC และ BIC ของวิธีการสุ่มแบบ
กิบส์ยังคงสูงกว่าวิธีการถดถอยแบบขั้นตอน โดยแตกต่างกัน
137.87 สำหรับค่า AIC และ 294.84 สำหรับค่า BIC

วิธีการสุ่มแบบกิบส์ให้จำนวนตัวแปรอิสระที่ถูก
ต้องในตัวแบบเท่ากับ 8.96 ขณะที่วิธีการถดถอยแบบขั้นตอน
ให้ค่าเท่ากับ 4.29 โดยวิธีการสุ่มแบบกิบส์มีจำนวนตัวแปร
อิสระที่ต้องมากกว่า 4.67

ค่าประมาณสัมประสิทธิ์ของตัวแปรอิสระที่ต้อง
โดยเฉลี่ยสำหรับวิธีการสุ่มแบบกิบส์เท่ากับ 6.87 ส่วนวิธีการ
ถดถอยแบบขั้นตอนมีค่าเฉลี่ยเท่ากับ 572.96 ซึ่งทั้งสองวิธีมี
ความแตกต่างจากค่าสัมประสิทธิ์ที่กำหนดไว้ คือ 2 โดยมีค่า
ส่วนต่าง 4.87 และ 570.96 ตามลำดับ

การคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบของวิธีการ
สุ่มแบบกิบส์มีความถูกต้องร้อยละ 93 ส่วนวิธีการถดถอยแบบ
ขั้นตอนมีความถูกต้องร้อยละ 15

กรณีข้อมูลจริง

ในกรณีนี้ได้ศึกษาโดยใช้ข้อมูลจริงโดยแบ่งออกเป็น
คัดเลือกตัวแปรโดยใช้วิธีการถดถอยแบบขั้นตอนและวิธี
เบสเซียน โดยการสุ่มแบบกิบส์ ในกรณีที่มีความสัมพันธ์เชิง
เส้นพหุระหว่างตัวแปรอิสระในระดับต่ำมากและสูงมาก ดังแสดง
ใน Table 2

Table 2 The comparison of AIC, BIC, and the accuracy of independent variable selection for the real dataset.

Correlation coefficient	Method	AIC	BIC
0.01	Stepwise	60.00	154.77
	Gibbs	190.00	490.11
$r \geq 0.91$	Stepwise	316.00	815.13
	Gibbs	344.00	887.35

กรณีที่มีความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปร
อิสระในระดับต่ำมาก ผลจาก Table 2 พบว่า ค่า AIC และ
BIC ของวิธีการสุ่มแบบกิบส์เท่ากับ 190.00 และ 490.11 ตาม
ลำดับ ส่วนค่า AIC และ BIC ของวิธีการถดถอยแบบขั้นตอน
เท่ากับ 60.00 และ 154.77 ตามลำดับ แสดงให้เห็นว่า ค่า AIC
และ BIC ของวิธีการสุ่มแบบกิบส์สูงกว่าวิธีการถดถอยแบบ
ขั้นตอน โดยต่างกันที่ 130.00 สำหรับค่า AIC และ 335.34
สำหรับค่า BIC

ในกรณีที่มีความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปร
อิสระในระดับสูงมาก พบว่าค่า AIC และ BIC ของวิธีการสุ่ม

แบบกิบส์เท่ากับ 344.00 และ 887.35 ตามลำดับ ขณะที่ค่า
AIC และ BIC ของวิธีการถดถอยแบบขั้นตอนเท่ากับ 316.00
และ 887.35 ตามลำดับ แสดงให้เห็นว่าค่า AIC และ BIC ของ
วิธีการสุ่มแบบกิบส์ยังคงสูงกว่าวิธีการถดถอยแบบขั้นตอน

สรุปผลการวิจัยและข้อเสนอแนะ

การศึกษานี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพ
การคัดเลือกตัวแปรอิสระของทั้งสองวิธีสำหรับแต่ละกรณีโดย
ใช้เกณฑ์ในการเปรียบเทียบ คือ เกณฑ์ AIC และ BIC จำนวน
ร้อยละความถูกต้องของการคัดเลือกตัวแปร จำนวนตัวแปรที่

เลือกถูกต้อง และค่าประมาณสัมประสิทธิ์โดยเฉลี่ย โดยมีการจำลองข้อมูลและทดสอบตัวแบบกับข้อมูลจริง

จากการศึกษาพบว่าทั้งวิธีการสุ่มแบบกิบส์และวิธีการถดถอยแบบขั้นตอน สามารถทำงานได้อย่างมีประสิทธิภาพมากขึ้นเมื่อมีการปรับขั้นตอนและพารามิเตอร์ต่าง ๆ อย่างเหมาะสมกับข้อมูลที่กำลังวิเคราะห์ โดยการปรับพารามิเตอร์และการตั้งค่าที่เหมาะสมในแต่ละวิธีมีความสำคัญอย่างมาก เนื่องจากจะส่งผลโดยตรงต่อความแม่นยำและประสิทธิภาพของตัวแบบ

การเปรียบเทียบค่า AIC และ BIC ในกรณีที่มีความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระในระดับต่ำมากและสูงมากทั้งในขนาดตัวอย่าง 25 และ 100 พบว่าวิธีการถดถอยแบบขั้นตอนมีค่า AIC และ BIC ต่ำกว่าวิธีการสุ่มแบบกิบส์ในทุกกรณี แม้ว่าวิธีการสุ่มแบบกิบส์จะสามารถคัดเลือกตัวแปรได้ดีและมีความยืดหยุ่นมากกว่า แต่ค่า AIC และ BIC กลับสูงกว่าวิธีการถดถอยแบบขั้นตอนในทุกกรณี แสดงให้เห็นว่าแม้วิธีการสุ่มแบบกิบส์จะมีความซับซ้อนและมีความยืดหยุ่นในการคัดเลือกตัวแปร แต่วิธีการถดถอยแบบขั้นตอนยังคงเหมาะสมกว่าในแง่ของความเรียบง่ายของตัวแบบ

การเปรียบเทียบค่าต่าง ๆ ในกรณีที่มีความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระในระดับต่ำมาก ทั้งวิธีการสุ่มแบบกิบส์และวิธีการถดถอยแบบขั้นตอนสามารถคัดเลือกตัวแปรอิสระได้อย่างมีประสิทธิภาพได้อย่างใกล้เคียงกัน แต่เมื่อมีความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระสูงมาก วิธีการสุ่มแบบกิบส์มีประสิทธิภาพที่ดีกว่า โดยเฉพาะเมื่อขนาดตัวอย่างใหญ่ เนื่องจากวิธีการสุ่มแบบกิบส์มีความยืดหยุ่นในการเลือกตัวแปรและสามารถรักษาตัวแปรสำคัญไว้ได้มากกว่า ในขณะที่วิธีการถดถอยแบบขั้นตอนมักจะตัดตัวแปรสำคัญออกจากตัวแบบในกรณีที่ตัวแปรอิสระมีความสัมพันธ์ต่อกันสูง

เมื่อเปรียบเทียบจำนวนตัวแปรอิสระที่ถูกต้องในตัวแบบระหว่างวิธีการสุ่มแบบกิบส์และการถดถอยแบบขั้นตอน พบว่าวิธีการสุ่มแบบกิบส์สามารถคัดเลือกตัวแปรที่ถูกต้องได้มากกว่าในทุกกรณี โดยเฉพาะเมื่อขนาดตัวอย่างเพิ่มขึ้น แสดงได้ว่าวิธีการสุ่มแบบกิบส์มีความแม่นยำในการคัดเลือกตัวแปรที่ถูกต้องมากกว่าวิธีการถดถอยแบบขั้นตอน

เมื่อเปรียบเทียบค่าประมาณสัมประสิทธิ์ของตัวแปรอิสระระหว่างวิธีการสุ่มแบบกิบส์และการถดถอยแบบขั้นตอน พบว่า วิธีการสุ่มแบบกิบส์ให้ค่าประมาณที่ใกล้เคียงกับค่าสัมประสิทธิ์ที่แท้จริงมากกว่าทั้งในกรณีที่มีความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระในระดับต่ำมากและสูงมาก รวมถึงในขนาดตัวอย่างที่เท่ากับ 25 และ 100

โดยสรุปแล้ว วิธีการสุ่มแบบกิบส์มีประสิทธิภาพดีกว่าในการจัดการกับปัญหาความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระและในการเลือกตัวแปรที่ถูกต้อง โดยเฉพาะเมื่อขนาดตัวอย่างใหญ่ ขณะที่วิธีการถดถอยแบบขั้นตอนยังคงเป็นวิธีที่เหมาะสมเมื่อไม่มีความสัมพันธ์เชิงเส้นพหุระหว่างตัวแปรอิสระ เมื่อพิจารณาจากเกณฑ์ AIC และ BIC

อย่างไรก็ตาม เพื่อเป็นการขยายขอบเขตการศึกษาและหาตัวแบบที่มีประสิทธิภาพดียิ่งขึ้น อาจนำวิธีการคัดเลือกตัวแปรอิสระด้วยวิธีการสุ่มแบบกิบส์ไปประยุกต์ใช้กับข้อมูลจริง และนำวิธีการคัดเลือกตัวแปรอิสระที่ใช้ในงานวิจัยไปเปรียบเทียบกับเทคนิคแบบอื่นๆ เช่น การถดถอยแบบแลซโซ่ (Lasso Regression) เป็นต้น

กิตติกรรมประกาศ

งานวิจัยเรื่อง การเปรียบเทียบวิธีการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบการถดถอยลอจิสติกทวินามโดยใช้วิธีเบย์เซียนและวิธีการถดถอยแบบขั้นตอนนี้ สำเร็จลุล่วงได้ด้วยความร่วมมือและการสนับสนุนจากหลายฝ่าย คณะผู้วิจัยขอขอบคุณสาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ ที่ให้การสนับสนุนทุนในการศึกษาวิจัย รวมทั้งทรัพยากรต่าง ๆ ตลอดจนการเข้าถึงข้อมูลที่เป็นในการวิจัย สุดท้ายนี้ คณะผู้วิจัยหวังเป็นอย่างยิ่งว่างานวิจัยนี้จะมีส่วนในการพัฒนาและประยุกต์ใช้วิธีการคัดเลือกตัวแปรอิสระในงานวิเคราะห์เชิงสถิติ ตลอดจนเป็นประโยชน์ต่อผู้ที่สนใจในการนำวิธีเบย์เซียนและวิธีการถดถอยแบบขั้นตอนมาใช้ในการงานวิจัยและการวิเคราะห์ข้อมูล

เอกสารอ้างอิง

- ธนัญญ์ เกิดผล. (2559). *ตัวแบบการถดถอยลอจิสติกแลซโซ่ที่อธิบายความสัมพันธ์ของการเป็นหนี้สินและตัวชี้วัดระดับความเข้มแข็งของครัวเรือน* [สารนิพนธ์ปริญญาโทบริหารธุรกิจ, มหาวิทยาลัยธรรมศาสตร์].
- บุญชม ศรีสะอาด. (2556). *วิธีการทางสถิติสำหรับการทำวิจัย เล่ม 1* (พิมพ์ครั้งที่ 5). สุวีริยาสาส์น.
- ประพันธ์ กล่อมพร. (2560). *ขั้นตอนวิธีกำหนดเมทริกซ์สหสัมพันธ์แรกเริ่มโดยการจำลองสำหรับการแปลงนอร์ท* [วิทยานิพนธ์ปริญญาโทบริหารธุรกิจ, มหาวิทยาลัยธรรมศาสตร์].
- พจนา แวสวัสดิ์. (2547). เกณฑ์การคัดเลือกตัวแบบสำหรับความถดถอยเชิงเส้นตรง. *วารสารศรีปทุมปริทัศน์*, 4(1), 5–20.

- ภักธุภางค์ มาปรีดา. (2560). *ตัวแบบการถดถอยลอจิสติกในการพยากรณ์ความน่าจะเป็นของการชำระหนี้ได้ของครัวเรือน: กรณีศึกษา จังหวัดปทุมธานี* [สารนิพนธ์ปริญญา มหาบัณฑิต, มหาวิทยาลัยธรรมศาสตร์].
- Akaike, H. (1978). A Bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics*, 30(1), 9–14.
- Arkes, J. (2023). *Regression analysis: A practical introduction* (2nd ed.). Routledge.
- Cilia, M., De Stefano, C., Fontanella, F., & Di Freca, A. S. (2022). Diagnosing Alzheimer's disease from on-line handwriting: A novel dataset and performance benchmarking. *Engineering Applications of Artificial Intelligence*, 111, Article 104822. <https://doi.org/10.1016/j.engappai.2022.104822>
- Müller, S., Scealy, J. L., & Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science*, 28(2), 135–167.
- Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American Statistical Association*, 108(504), 1339–1349.
- Sawa, T. (1978). Information criteria for discriminating among alternative regression models. *Econometrica*, 46(6), 1273–1291.
- Tian, Y., Bondell, H. D., & Wilson, A. (2019). Bayesian variable selection for logistic regression. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(5), 378–393.