



KKU SCIENCE JOURNAL

Journal Home Page : <https://ph01.tci-thaijo.org/index.php/KKUSciJ>

Published by the Faculty of Science, Khon Kaen University, Thailand



การเปรียบเทียบวิธีการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบการถดถอยเชิงเส้นพหุคูณ โดยใช้วิธีเบส์และวิธีการถดถอยแบบขั้นตอน

A Comparison of Subset Selection in Multiple Linear Regression Model by Using Bayesian Variable Selection and Stepwise Regression

กานต์ณัฐ ฌ บางช้าง^{1*}

Kannat Na Bangchang^{1*}

¹สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ จังหวัดปทุมธานี 12000

¹Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathumthani, 12000, Thailand

บทคัดย่อ

งานวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบการถดถอยเชิงเส้นพหุคูณ ด้วยวิธีการคัดเลือกตัวแปรของเบส์ โดยใช้วิธีการสุ่มแบบกิบส์ และทำการเปรียบเทียบประสิทธิภาพในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบของเบส์โดยใช้วิธีการสุ่มแบบกิบส์กับวิธีการถดถอยแบบขั้นตอน โดยการจำลองข้อมูล ทั้งกรณีที่ไม่มีความสัมพันธ์เชิงเส้นแบบพหุระหว่างตัวแปรอิสระ และกำหนดขนาดตัวอย่างเท่ากับ 25 และ 100 โดยกระทำซ้ำ 500 ครั้งในแต่ละกรณีเกณฑ์ที่ใช้ในการเปรียบเทียบ คือ ร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบและค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ผลการวิจัยพบว่าเมื่อชุดข้อมูลจำลองเป็นขนาดตัวอย่างเล็ก ร้อยละความถูกต้องของการคัดเลือกตัวแปรด้วยวิธีการคัดเลือกตัวแปรของเบส์โดยใช้วิธีการสุ่มแบบกิบส์สูงกว่าวิธีการถดถอยแบบขั้นตอน ทั้งกรณีที่ไม่มีความสัมพันธ์เชิงเส้นแบบพหุระหว่างตัวแปรอิสระ แต่เมื่อขนาดตัวอย่างมากขึ้น กรณีที่ไม่มีความสัมพันธ์เชิงเส้นแบบพหุระหว่างตัวแปรอิสระ วิธีการถดถอยแบบขั้นตอนได้ร้อยละความถูกต้องของการคัดเลือกตัวแปรมากขึ้น ซึ่งเท่ากับกับวิธีการคัดเลือกตัวแปรของเบส์โดยใช้วิธีการสุ่มแบบกิบส์ และ MSE ของทั้งสองวิธีมีค่าเท่ากัน ซึ่งมีค่าน้อยมาก อันเนื่องมาจากค่าจริงและค่าพยากรณ์ของตัวแปรตามที่ได้จากแต่ละวิธีมีค่าใกล้เคียงกัน ในทางกลับกันเมื่อมีความสัมพันธ์เชิงเส้นแบบพหุระหว่างตัวแปรอิสระ วิธีการคัดเลือกตัวแปรของเบส์โดยใช้วิธีการสุ่มแบบกิบส์และวิธีการถดถอยแบบขั้นตอนไม่สามารถคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้อย่างถูกต้อง

ABSTRACT

The purpose of this research is to study variable selection for multiple linear regression models with Bayesian Variable Selection by using Gibbs Sampling (GS) and compare the efficiency of variable

*Corresponding Author, E-mail: kannat@mathstat.sci.tu.ac.th

Received date: 2 September 2025 | Revised date: 16 November 2025 | Accepted date: 19 November 2026

doi: 10.14456/kkuscij.2026.15

selection between Gibbs Sampling (GS) and Stepwise Regression (SR). The study employs simulation data which compared both cases with and without multicollinearity. A sample size of 25 and 100 is utilized, and repeated 500 times for each sample size. A significance level of 0.05 is set for selecting independent variables using SR. The comparison criterion includes the percentage of accurate variable selection and the mean squared error (MSE). When simulated dataset is small sample size, the percentage of accuracy with GS is higher than SR in both cases with and without multicollinearity. However, as the sample size increases, percentage of accurate of SR also increases for case without multicollinearity which is equivalent to GS. The MSE for both methods is equal, which is very low. This outcome is due to the similarity between true values and predicted values of the dependent variables obtained from each method. Conversely, when there is multicollinearity, both GS and SR are unable to accurately select variables for the model.

คำสำคัญ: การจำลองข้อมูล การคัดเลือกตัวแปรของเบส์ วิธีการสุ่มแบบกิบส์ วิธีการถดถอยแบบขั้นตอน ความสัมพันธ์เชิงเส้นแบบพหุระหว่างตัวแปรอิสระ

Keywords: Simulation Data, Bayesian Variable Selection, Gibbs Sampling, Stepwise Regression, Multicollinearity

บทนำ

การวิจัยมีความสำคัญเป็นอย่างมากในปัจจุบัน เนื่องจากเป็นแนวทางในการแสวงหาความรู้ใหม่ ๆ และพัฒนาความรู้เดิมของสาขาวิชาต่าง ๆ ให้มีความเจริญก้าวหน้า ซึ่งวิธีการทางสถิติถูกนำมาประยุกต์ใช้ในงานวิจัยอย่างแพร่หลาย โดยเฉพาะการวิเคราะห์การถดถอย (Regression Analysis) เป็นวิธีการทางสถิติที่นิยมนำมาใช้ในงานวิจัยเพราะการวิจัยเป็นกระบวนการของการศึกษาตัวแปร จะสังเกตได้ว่าการวิจัยเกือบทุกเรื่องจะต้องเกี่ยวข้องกับตัวแปรอย่างน้อยที่สุด 1 ตัว ซึ่งการวิเคราะห์การถดถอยใช้เพื่อศึกษารูปแบบความสัมพันธ์ระหว่างตัวแปรหรือประมาณค่าของตัวแปรตาม โดยสร้างตัวแบบจากข้อมูลของตัวแปรอิสระ ดังนั้นการวิจัยจึงมีความเกี่ยวข้องกับการสร้างตัวแบบด้วย เพื่อใช้ในการพิจารณาความสัมพันธ์ของตัวแปรตามกับตัวแปรอิสระ แต่โดยส่วนใหญ่ในงานวิจัยมักจะเกี่ยวข้องกับการมีตัวแปรอิสระมากกว่า 1 ตัวแปร ที่มีความสัมพันธ์เชิงเส้นกับตัวแปรตามหรือเรียกว่า การถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression) อย่างไรก็ตามการมีตัวแปรอิสระอยู่ในตัวแบบเป็นจำนวนมากนั้นไม่ได้หมายความว่าตัวแปรอิสระทุกตัวจะมีความสัมพันธ์เชิงเส้นกับตัวแปรตาม กล่าวคืออาจมีตัวแปรอิสระบางตัวที่อธิบายความผันแปรได้น้อย นอกจากนี้อาจมีตัวแปรอิสระบางตัวที่มีความสัมพันธ์กันเอง หรือเรียกว่าเกิดปัญหาความสัมพันธ์เชิงเส้นแบบพหุ (Multicollinearity) นอกจากนี้การมีตัวแปรอิสระในตัวแบบถดถอยเชิงเส้นพหุคูณเป็นจำนวนมาก อาจก่อให้เกิดปัญหาตัวแบบมีความไม่เหมาะสม เนื่องจากอาจมีตัวแปรอิสระบางตัวที่ไม่มีความสัมพันธ์เชิงเส้นกับตัวแปรตาม ซึ่งถ้าเลือกตัวแบบไม่เหมาะสมก็จะส่งผลให้ผลการวิเคราะห์ไม่ถูกต้องด้วย เช่น การที่ตัวแบบมีตัวแปรอิสระมากเกินไป เมื่อนำตัวแบบไปใช้จะเกิดปัญหาเรื่องความแม่นยำในการประมาณค่า แต่ถ้าตัวแบบมีตัวแปรอิสระน้อยเกินไป เมื่อนำไปใช้ก็จะทำให้เกิดการประมาณค่าแบบเอนเอียง หรือถ้าคัดเลือกตัวแปรอิสระที่มีความสัมพันธ์เชิงเส้นกับตัวแปรตามสูงออกจากตัวแบบ จะทำให้ตัวแบบที่จะนำไปประมาณค่าตัวแปรตามนั้นมีประสิทธิภาพต่ำ ส่งผลให้ค่าประมาณที่ได้มีความคลาดเคลื่อนสูง จึงเห็นได้ว่าการเลือกตัวแบบถดถอยเชิงเส้นที่ดีที่สุดเป็นเรื่องที่ซับซ้อน จากกรณีศึกษาจำนวนมากพบว่าตัวแบบที่ไม่ซับซ้อนจะสามารถให้ผลการทำนายที่แม่นยำกว่า (Burnham and Anderson, 2002) เพราะยังมีความซับซ้อนมาก การปรับแต่งตัวแบบให้เหมาะกับข้อมูลก็ยังคงมีความยากมากยิ่งขึ้น ดังนั้นจึงควรเริ่มต้นที่ตัวแบบมีจำนวนตัวแปรอิสระน้อย ๆ แล้วจึงเพิ่มจำนวนตัวแปรอิสระลงในตัวแบบให้มากขึ้น ดังนั้นการเลือกตัวแบบที่มีความเหมาะสมสูงที่สุดจึงต้องอาศัยการคัดเลือกตัวแปรเป็นองค์ประกอบสำคัญ

ผู้วิจัยจึงศึกษาวิธีการคัดเลือกตัวแปรของเบส์ (Bayesian Variable Selection) โดยใช้วิธีการสุ่มแบบกิบส์ (Gibbs Sampling) และเปรียบเทียบประสิทธิภาพในการคัดเลือกตัวแปรด้วยวิธีการของเบส์โดยใช้วิธีการสุ่มแบบกิบส์ กับวิธีการถดถอยแบบขั้นตอน (Stepwise Regression)

วิธีการดำเนินการวิจัย

งานวิจัยนี้ศึกษาการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบการถดถอยเชิงเส้นพหุคูณด้วยวิธีการคัดเลือกตัวแปรของเบส์ โดยใช้วิธีการสุ่มแบบกิบส์ และวิธีการถดถอยแบบขั้นตอนโดยใช้ข้อมูลจำลองที่มีจำนวนตัวแปรอิสระในตัวแบบเต็มรูปแบบเท่ากับ 24 ตัวแปร และจำนวนตัวแปรอิสระในตัวแบบแท้จริงเท่ากับ 10 ตัวแปร ขนาดตัวอย่างเท่ากับ 25 และ 100 ทั้งกรณีที่ไม่มีและมีความสัมพันธ์เชิงเส้นแบบพหุระหว่างตัวแปรอิสระ โดยมีขั้นตอนการดำเนินงานดังนี้

1. ทำการจำลองข้อมูลที่จะใช้ในการศึกษา โดยกำหนดให้ตัวแปรอิสระ 24 ตัวแปร มีการแจกแจงแบบเอกรูป $X \sim \text{Uniform}(-200, 200)$ ซึ่งมีตัวแปรอิสระในตัวแบบแท้จริงเพียง 10 ตัวเท่านั้นที่ใช้ในการกำหนดค่าตัวแปรตาม Y และกำหนดความคลาดเคลื่อนสุ่มที่เป็นอิสระต่อกันและมีการแจกแจงแบบปกติ $\epsilon \sim N(0,1)$

2. การสุ่มข้อมูลของตัวแปรอิสระสำหรับแต่ละกรณี

กรณีที่ไม่มีความสัมพันธ์เชิงเส้นแบบพหุระหว่างตัวแปรอิสระ

ทำการสุ่มข้อมูลของตัวแปรอิสระ ทั้งกรณีขนาดตัวอย่างเท่ากับ 25 และ 100

กรณีที่มีความสัมพันธ์เชิงเส้นแบบพหุระหว่างตัวแปรอิสระ

ทำการสุ่มข้อมูลของตัวแปรอิสระ ทั้งกรณีขนาดตัวอย่างเท่ากับ 25 และ 100 แล้วกำหนดตัวแปรอิสระบางตัวใหม่ให้มีความสัมพันธ์กัน ในการศึกษาสร้าง X_8 ให้มีความสัมพันธ์กับ X_1 และสร้าง X_9 ให้มีความสัมพันธ์กับ X_2 ซึ่งกำหนดให้มีความสัมพันธ์เชิงเส้นแบบพหุในระดับสูงมาก ($r \geq 0.91$) และกรณีที่ไม่มีความสัมพันธ์เชิงเส้นแบบพหุ จะมีระดับความสัมพันธ์ระหว่างตัวแปรอิสระไม่เกิน 0.3

วิธีจำลองค่าของแมคโดนัลด์และกาลารนู (Macdonal and Galaneau)

$$X_{ij} = \sqrt{(1 - \rho^2)}U_{ij} + \rho U_{i(p+1)} ; i = 1, 2, \dots, n, j = 1, 2, \dots, p \quad (1)$$

โดยที่ ρ คือ ค่าสัมประสิทธิ์สหสัมพันธ์ที่ต้องการกำหนด p คือ จำนวนตัวแปรอิสระ และ

$U_{i1}, U_{i2}, \dots, U_{i(p+1)}$ เป็นค่าของตัวแปรอิสระที่สร้างขึ้นโดยการให้มีการแจกแจงเอกรูป

3. ตัวแปรอิสระที่เกี่ยวข้องกับตัวแปรตาม Y คือ $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$ และตัวแปรอิสระที่ไม่เกี่ยวข้องกับตัวแปรตาม Y คือ $X_{11}, X_{12}, X_{13}, X_{14}, X_{15}, X_{16}, X_{17}, X_{18}, X_{19}, X_{20}, X_{21}, X_{22}, X_{23}, X_{24}$ กำหนดค่าสัมประสิทธิ์การถดถอยดังนี้ $\beta' = [0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$

4. สร้างตัวแปรตาม Y ตามรูปแบบความสัมพันธ์ $Y = X\beta + \epsilon$

5. ทำการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบด้วยวิธีการคัดเลือกตัวแปรของเบส์โดยใช้วิธีการสุ่มแบบกิบส์และวิธีการถดถอยแบบขั้นตอนในแต่ละกรณี

6. การนับผลการคัดเลือกตัวแปร

กรณีที่ไม่มีความสัมพันธ์เชิงเส้นแบบพหุระหว่างตัวแปรอิสระ จะพิจารณาแค่กรณีเดียว คือ $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$ ต้องอยู่ในตัวแบบ จึงจะถือว่าคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้อย่างถูกต้อง

กรณีที่มีความสัมพันธ์เชิงเส้นแบบพหุระหว่างตัวแปรอิสระ แบ่งออกเป็น 4 กรณี คือ

$$6.1) X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_{10}$$

$$6.2) X_1, X_3, X_4, X_5, X_6, X_7, X_9, X_{10}$$

$$6.3) X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_{10}$$

$$6.4) X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}$$

หากสามารถคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ตามกรณีใดกรณีหนึ่งใน 4 กรณีนี้ จะถือว่าสามารถคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้อย่างถูกต้อง กล่าวคือ จะไม่เลือกตัวแปรอิสระที่มีความสัมพันธ์เชิงเส้นแบบพหุในระดับสูงมากเข้าตัวแบบพร้อมกันทั้งสองตัวแปร

7. แต่จะรอบให้โปรแกรม R version 4.3.1 (<https://cran.r-project.org/bin/windows/base/old/4.3.1/>) ดำเนินการคำนวณร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ และค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของวิธีการคัดเลือกตัวแปรอิสระทั้ง 2 วิธี

8. ทำการคัดเลือกตัวแปรอิสระซ้ำ 500 ครั้ง ในแต่ละกรณี

9. เปรียบเทียบร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบและค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของแต่ละกรณี โดยจะถือว่าตัวแบบใดที่มีร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบสูง และค่าความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำ เป็นตัวแบบที่ดีกว่า

วิธีการทางสถิติที่ใช้ในงานวิจัย

วิธีการถดถอยแบบขั้นตอน เป็นวิธีการเลือกตัวแปรอิสระเข้าในตัวแบบการถดถอยครั้งละหนึ่งตัวแปร ตัวแปรอิสระใดที่ถูกเลือกเข้ามาอยู่ในตัวแบบการถดถอยแล้วจะต้องมีการทดสอบว่าตัวแปรอิสระที่อยู่ในตัวแบบการถดถอยแล้วมีส่วนในการอธิบายความผันแปรของตัวแปรตาม ขณะที่ตัวแปรอิสระอื่นอยู่ในตัวแบบการถดถอย (Efroymson, 1960) นั่นคือตัวแปรอิสระใดที่ถูกเลือกเข้ามาอยู่ในตัวแบบการถดถอยแล้วอาจจะถูกนำออกไปได้ภายหลังถ้าพบว่าตัวแปรอิสระนั้นไม่มีนัยสำคัญ โดยมีขั้นตอนดังนี้

1) เริ่มต้นพิจารณาค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตาม โดยเลือกตัวแปรอิสระซึ่งมีค่าสัมประสิทธิ์สหสัมพันธ์กับตัวแปรตามสูงที่สุดเข้าสู่ตัวแบบเป็นตัวแรก

2) ภายหลังจากที่ตัวแปรอิสระถูกนำเข้ามาในตัวแบบแล้วจะมีการทดสอบนัยสำคัญของตัวแปรอิสระที่นำเข้ามาเป็นตัวแรกโดยใช้ค่าสถิติ t (t -value) หรือค่าสถิติ F (F -value)

3) พิจารณาค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนระหว่างตัวแปรอิสระที่เหลือกับตัวแปรตาม โดยจะเลือกตัวแปรอิสระที่ให้ค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนกับตัวแปรตามสูงที่สุดเข้ามาอยู่ในตัวแบบเป็นตัวต่อมา

4) หลังจากนั้นจึงทำการทดสอบความมีนัยสำคัญของตัวแปรที่เข้ามาเป็นตัวสุดท้ายก่อนโดยพิจารณาจากค่าสถิติ F บางส่วน ถ้ามีค่าน้อยกว่าค่าสถิติ F จากตารางจะต้องนำตัวแปรนั้นออกจากตัวแบบ แต่ถ้าค่าสถิติ F บางส่วนมีค่ามากกว่าค่าสถิติ F จากตารางก็จะย้อนกลับไปทดสอบตัวแปรอิสระตัวก่อนหน้าที่เข้ามาอยู่ในตัวแบบโดยพิจารณาจากค่าสถิติ F บางส่วนเช่นเดียวกัน

5) ทำซ้ำในขั้นตอนที่ 3) และ 4) ต่อไปเรื่อย ๆ จนกระทั่งไม่สามารถนำตัวแปรอิสระใดออกจากตัวแบบได้หรือไม่สามารถนำตัวแปรอิสระใดเข้าสู่ตัวแบบได้อีก จะพบว่าวิธีการถดถอยแบบขั้นตอนใช้การพิจารณาค่าสถิติ F บางส่วน ดังนั้นการกำหนดระดับนัยสำคัญของค่าสถิติ F สำหรับเลือกตัวแปรอิสระเข้า (α_{enter}) และนำตัวแปรอิสระออก (α_{stay}) จึงมีความสำคัญ ตัวอย่างเช่นถ้ามีการกำหนดค่าระดับนัยสำคัญของค่าสถิติ F สำหรับเลือกตัวแปรอิสระเข้าลดลง จะทำให้การเลือกตัวแปรอิสระเข้ามาในตัวแบบลดลงหรือทำให้นำตัวแปรอิสระออกจากตัวแบบเพิ่มขึ้น

วิธีการคัดเลือกตัวแปรของเบส์ สำหรับตัวแบบการถดถอยเชิงเส้นพหุคูณ จะอาศัยหลักการคัดเลือกตัวแปรโดยการค้นหาด้วยความน่าจะเป็น (Stochastic Search Variable Selection : SSVS) ซึ่งใช้วิธีการสุ่มแบบกิบส์ มีหลักการคือ กำหนดการแจกแจงก่อน (Prior Distribution) ของสัมประสิทธิ์การถดถอย (β) ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนสุ่ม (σ) และตัวแปรตาม (Y) ในตัวแบบการถดถอย แล้วคำนวณความน่าจะเป็นภายหลัง (Posterior Probability) ของกลุ่มของตัวแปรอิสระที่เป็นไปได้ทั้งหมดเมื่อมีข้อมูลของตัวแปรตาม (Tippayawannakorn, 1997) ซึ่งมีขั้นตอนเป็นดังนี้

$$1) \text{ ให้ } Y = [Y_1, \dots, Y_j, \dots, Y_p]^T \quad (2)$$

เมื่อ Y_j เป็นตัวแปรที่ใช้เป็นตัวบ่งชี้ว่าตัวแปรอิสระที่ j อยู่ในตัวแบบการถดถอยหรือไม่ โดยที่ $Y_j = 1$ เมื่อตัวแปรอิสระตัวที่ j อยู่ในตัวแบบการถดถอย และ $Y_j = 0$ เมื่อตัวแปรอิสระตัวที่ j ไม่อยู่ในตัวแบบการถดถอย

2) สมมติให้ตัวแปรตาม Y มีการแจกแจงปกติ (Normal Distribution) โดยมีเงื่อนไขอยู่บน β, σ

$$Y|\beta, \sigma \sim N_n(X\beta, \sigma^2 I) \quad (3)$$

และได้ฟังก์ชันภาวะความน่าจะเป็น (Likelihood Function) ของตัวแปรตาม Y จาก $Y|\beta, \sigma \sim N_n(X\beta, \sigma^2 I)$ คือ

$$f(Y|\beta, \sigma) = (2\pi)^{-n/2} \sigma^{-n} \exp \left[-\frac{1}{2\sigma^2} (Y - X\beta)^T (Y - X\beta) \right] \quad (4)$$

เมื่อ Y คือ เวกเตอร์ขนาด $n \times 1$ ของตัวแปรตามที่ถูกปรับด้วยค่าเฉลี่ย

X คือ เมทริกซ์ขนาด $n \times p$ ของตัวแปรอิสระที่ถูกปรับด้วยค่าเฉลี่ย

β คือ เวกเตอร์ของสัมประสิทธิ์การถดถอยขนาด $p \times 1$

σ คือ ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนสุ่ม

วิธีการคัดเลือกตัวแปรอิสระของเบส์นี้ได้กำหนดการแจกแจงก่อนของ β เป็นแบบคอนจูเกต (Conjugate Prior) และอยู่ในรูปของการแจกแจงก่อนแบบลำดับขั้น (Hierarchical Prior Distribution) (Mitchell and Beauchamp, 1988) ดังนี้

การแจกแจงก่อนแบบลำดับขั้น สำหรับ β, σ, γ คือ

$$f(\beta, \sigma, \gamma) = f(\beta|\sigma, \gamma) f(\sigma|\gamma) f(\gamma) \quad (5)$$

3) กำหนดการแจกแจงก่อนแบบคอนจูเกตของ β ให้มีการแจกแจงปกติ โดยมีเงื่อนไขอยู่บน σ, γ คือ

$$\beta|\sigma, \gamma \sim N_p(0, \sigma^2 D_\gamma R_\gamma D_\gamma) \quad (6)$$

ดังนั้น ฟังก์ชันความหนาแน่นน่าจะเป็นก่อน (Prior Probability Density Function) ของ β ที่มีเงื่อนไขอยู่บน σ, γ คือ

$$f(\beta|\sigma, \gamma) = (2\pi)^{-\frac{p}{2}} \sigma^{-p} |D_\gamma R_\gamma D_\gamma|^{-\frac{1}{2}} \exp \left[-\frac{1}{2\sigma^2} \beta^T (D_\gamma R_\gamma D_\gamma)^{-1} \beta \right] \quad (7)$$

$$; -\infty < \beta_j < \infty, j = 1, \dots, p$$

เมื่อ D_γ คือ Diagonal Matrix ขนาด $p \times p$ ที่ใช้กำหนดสัดส่วนของเมทริกซ์ของความแปรปรวนร่วมก่อน (Prior Covariance Matrix) ของ β และให้สมาชิกในแนวทแยงมุมของ D_γ คือ $(D_\gamma)_{jj} = \sqrt{U_{0\gamma_j}}$ เมื่อ $\gamma_j = 0$; $j = 1, \dots, p$ และ $(D_\gamma)_{jj} = \sqrt{U_{1\gamma_j}}$ เมื่อ $\gamma_j = 1$; $j = 1, \dots, p$ โดยที่ $U_{0\gamma_j} < U_{1\gamma_j}$

ตัวอย่างของ D_Y ในกรณี $p = 3$ และ $\gamma = [\gamma_1, \gamma_2, \gamma_3]^T = [1, 0, 0]^T$ (Viadinugroho, 2021) คือ

$$D_Y = \begin{bmatrix} \sqrt{U_{1\gamma_1}} & 0 & 0 \\ 0 & \sqrt{U_{0\gamma_2}} & 0 \\ 0 & 0 & \sqrt{U_{0\gamma_3}} \end{bmatrix}$$

จาก $\beta|\sigma, \gamma \sim N_p(0, \sigma^2 D_Y R_Y D_Y)$ และ $U_{0\gamma_j} = 0$ จะได้

$$\beta_Y|\sigma, \gamma \sim N_{q_Y}(0, \sigma^2 D_{1Y} R_{1Y} D_{1Y}) \quad (8)$$

เมื่อ β_Y คือ เวกเตอร์ของสัมประสิทธิ์การถดถอยที่สอดคล้องกับ $\gamma_j = 1$

q_Y คือ ขนาดของสับเซต (subset) ของตัวแปรอิสระที่ $\gamma_j = 1$

จาก $\beta|\sigma, \gamma \sim N_p(0, \sigma^2 D_Y R_Y D_Y)$; R_Y คือ เมทริกซ์ของสหสัมพันธ์ก่อน (Prior Correlation Matrix) ของ β เมื่อมีเงื่อนไขอยู่บน γ ซึ่งมี 2 กรณี คือ $R_Y = I$ คือ สัมประสิทธิ์การถดถอยแต่ละตัวเป็นอิสระกัน และ $R_Y \propto (X^T X)^{-1}$ คือ สัมประสิทธิ์การถดถอยแต่ละตัวมีความสัมพันธ์กันเอง (Lee, 2021)

จาก $\beta_Y|\sigma, \gamma \sim N_{q_Y}(0, \sigma^2 D_{1Y} R_{1Y} D_{1Y})$; R_{1Y} คือ เมทริกซ์ของสหสัมพันธ์ก่อนของ β_Y เมื่อ $\gamma_j = 1$ และ $R_Y \propto (X_Y^T X_Y)^{-1}$ ซึ่งทำได้

$$D_{1Y} R_{1Y} D_{1Y} = c(X_Y^T X_Y)^{-1}; 0.1 \leq c \leq 1 \quad (9)$$

เมื่อ X_Y คือ เมทริกซ์ขนาด $n \times q_Y$ ของตัวแปรอิสระ

4) กำหนดการแจกแจงก่อนของ σ^2 มีการแจกแจงอินเวอร์สแกมมา (Inverse Gamma) ดังนี้

$$\sigma^2|\gamma \sim IG\left(\frac{v}{2}, \frac{v\lambda_v}{2}\right) \quad (10)$$

โดยที่ v คือ องศาเสรี (degrees of freedom) ของการแจกแจงอินเวอร์สแกมมา และ λ_v คือ พารามิเตอร์บ่งขนาด (Scale Parameter) ของการแจกแจงอินเวอร์สแกมมา

5) กำหนดการแจกแจงก่อนของ γ_j ให้มีการแจกแจงเบอร์นูลลี (Bernoulli Distribution) สมมติให้ γ_j แต่ละตัวเป็นอิสระกัน และ $p(\gamma_j = 1) = W_j$ จะได้ความน่าจะเป็นก่อนของ γ คือ

$$f(\gamma) = \prod_{j=1}^p W_j^{\gamma_j} (1 - W_j)^{(1-\gamma_j)} \quad (11)$$

เมื่อ W_j คือ ความน่าจะเป็นที่ตัวแปรอิสระตัวที่ j อยู่ในตัวแบบการถดถอย

6) การหาค่า $U_{0\gamma_j}, U_{1\gamma_j}$ ในข้อ 3) หาได้โดย

$$\left(\frac{U_{0\gamma_j}}{U_{1\gamma_j}}\right)^{\frac{v}{2}+1} = \frac{\left[U_{0\gamma_j} + \frac{\delta_j^2}{(v\lambda_v)}\right]}{\left[U_{1\gamma_j} + \frac{\delta_j^2}{(v\lambda_v)}\right]} \quad (12)$$

โดยที่ $\delta_j = \Delta Y / \Delta X_j$

เมื่อ $\Delta Y =$ ค่าสูงสุดของตัวแปรตาม - ค่าต่ำสุดของตัวแปรตาม

$\Delta X_j =$ ค่าสูงของตัวแปรอิสระที่ j - ค่าต่ำสุดของตัวแปรอิสระที่ j และ

$$U_{0\gamma_j} / U_{1\gamma_j} = 100 \quad (13)$$

7) การเลือกตัวแปรอิสระ

7.1) คำนวณความน่าจะเป็นภายหลังของ $\gamma_{(k)}$; $k = 1, 2, \dots, 2^p$ เมื่อมีข้อมูลของตัวแปรตาม Y ทั้ง 2^p เวกเตอร์

7.2) เลือก $\gamma_{(k)}$ ที่มีความน่าจะเป็นภายหลังสูงสุด ซึ่งจาก $\gamma_{(k)}$ นี้จะทำให้ได้กลุ่มของตัวแปรอิสระที่อยู่ในตัวแบบการถดถอย

ผลการวิจัยและวิจารณ์ผล

ผลการวิจัยนำเสนอโดยแบ่งพิจารณาตามขนาดตัวอย่าง คือ 25 และ 100 และแบ่งย่อยเป็นกรณีที่ไม่มีและมีความสัมพันธ์เชิงเส้นแบบพหุระหว่างตัวแปรอิสระ

เมื่อชุดข้อมูลจำลองเป็นขนาดตัวอย่างเล็ก ร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบด้วยวิธีการคัดเลือกตัวแปรของเบส์โดยใช้วิธีการสุ่มแบบกิบส์สูงกว่าวิธีการถดถอยแบบขั้นตอน ทั้งกรณีที่ไม่มีและมีความสัมพันธ์เชิงเส้นแบบพหุระหว่างตัวแปรอิสระ แต่เมื่อขนาดตัวอย่างมากขึ้น กรณีที่ไม่มีความสัมพันธ์เชิงเส้นแบบพหุระหว่างตัวแปรอิสระ วิธีการถดถอยแบบขั้นตอนได้ร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบมากขึ้น ดังแสดงในตารางที่ 1 ซึ่งร้อยละความถูกต้องของการคัดเลือกตัวแปรที่ได้มีค่าเท่ากับวิธีการคัดเลือกตัวแปรของเบส์โดยใช้วิธีการสุ่มแบบกิบส์ นอกจากนี้ MSE ของทั้งสองวิธียังคงมีค่าเท่ากันด้วย ซึ่งมีค่าน้อยมาก ดังแสดงในตารางที่ 2 อันเนื่องมาจากค่าจริงและค่าพยากรณ์ของตัวแปรตามที่ได้จากแต่ละวิธีมีค่าใกล้เคียงกัน ในทางกลับกันเมื่อข้อมูลเป็นขนาดตัวอย่างใหญ่และมีความสัมพันธ์เชิงเส้นแบบพหุระหว่างตัวแปรอิสระ วิธีการคัดเลือกตัวแปรของเบส์โดยใช้วิธีการสุ่มแบบกิบส์และวิธีการถดถอยแบบขั้นตอนไม่สามารถคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้อย่างถูกต้อง เนื่องจากร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบของทั้งสองวิธีมีค่าเป็นศูนย์

ตารางที่ 1 ร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ เมื่อขนาดตัวอย่างเท่ากับ 25 และ 100

n	Multicollinearity	Method	Percentage of Accuracy
25	No Multicollinearity	Gibbs	100
		Stepwise	72.4
	With Multicollinearity	Gibbs	90.6
		Stepwise	6
100	No Multicollinearity	Gibbs	100
		Stepwise	100
	With Multicollinearity	Gibbs	0
		Stepwise	0

ตารางที่ 2 ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) เมื่อขนาดตัวอย่างเท่ากับ 25 และ 100

n	Multicollinearity	Method	MSE
25	No Multicollinearity	Gibbs	140.997
		Stepwise	112.255
	With Multicollinearity	Gibbs	658.908
		Stepwise	719.089
100	No Multicollinearity	Gibbs	0.936
		Stepwise	0.936
	With Multicollinearity	Gibbs	0.919
		Stepwise	0.919

สรุปผลการวิจัย

วิธีการคัดเลือกตัวแปรของเบส์โดยใช้วิธีการสุ่มแบบกิบส์และวิธีการถดถอยแบบขั้นตอนมีประสิทธิภาพแตกต่างกันไปในแต่ละสถานการณ์ สำหรับขนาดตัวอย่างเล็กวิธีการคัดเลือกตัวแปรของเบส์โดยใช้วิธีการสุ่มแบบกิบส์จะมีประสิทธิภาพในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบดีที่สุดทั้งกรณีที่ไม่มีความสัมพันธ์เชิงเส้นแบบพหุ เนื่องจากได้ร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบสูงกว่าวิธีการถดถอยแบบขั้นตอน แต่เมื่อเป็นขนาดตัวอย่างใหญ่ กรณีที่ไม่มีความสัมพันธ์เชิงเส้นแบบพหุระหว่างตัวแปรอิสระ วิธีการคัดเลือกตัวแปรของเบส์โดยใช้วิธีการสุ่มแบบกิบส์และวิธีการถดถอยแบบขั้นตอนจะมีประสิทธิภาพในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบเท่ากัน เนื่องจากได้ร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบและค่า MSE ที่เท่ากัน แต่ในกรณีที่มีความสัมพันธ์เชิงเส้นแบบพหุระหว่างตัวแปรอิสระทั้งสองวิธีไม่มีประสิทธิภาพในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ เนื่องจากไม่สามารถคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้อย่างถูกต้อง ซึ่งสอดคล้องกับ Yardimci and Erar (2002) ที่ชี้ว่าปัญหาความสัมพันธ์เชิงเส้นแบบพหุ ส่งผลรุนแรงต่อความสามารถในการเลือกตัวแปรอิสระไม่ว่าจะเป็นวิธีการคัดเลือกตัวแปรของเบส์ หรือ วิธีการถดถอยแบบขั้นตอน อย่างไรก็ตามผู้ที่สนใจในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ อาจนำเทคนิควิธีอื่น ๆ มาปรับใช้ เช่น วิธีการถดถอยแบบบริดจ์ (Ridge Regression) วิธีการถดถอยแบบลาซโซ (LASSO Regression) ในกรณีที่มีปัญหาความสัมพันธ์เชิงเส้นแบบพหุ

เอกสารอ้างอิง

- Burnham, K. P. and Anderson, D. R. (2002). Model selection and multimodel inference: A practical information-theoretic approach. 2nd Edition. New York: Springer-Verlag.
- Efroymson, M.A. (1960). Multiple Regression Analysis. In: Ralston A. and Wilf, H., Eds., Mathematical Methods for Digital Computers. New York: John Wiley.
- Lee, S. (2021). Bayesian linear regression with Gibbs sampling using R code [Internet]. Source: <https://shorturl.asia/vx6jB>. Retrieved date 17 August 2025.
- Mitchell, T. and Beauchamp, J. (1988). Bayesian variable selection in linear regression. Journal of the American Statistical Association. 83(404): 1023 - 1032. doi: 10.2307/2290129.
- Tippayawannakorn, N. (1997). Comparison of predicted values of dependent variables using model selection by Bayesian variable selection, backward elimination and stepwise regression with hierarchical polynomial regression. Dissertation, Chulalongkorn University. Bangkok.
- Viadinugroho, R. (2021). Generate simulated dataset for linear model in R. Source: <https://shorturl.asia/T8PRO>. Retrieved date 17 August 2025.
- Yardimci, A and Erar, A. (2002). Bayesian variable selection in linear regression and a comparison. Hacettepe Journal of Mathematics and Statistics. 31: 63 - 76.

