การวิเคราะห์การถดถอยเชิงเส้นกรณีที่ข้อมูลมีมิติสูงโดยใช้ขั้นตอนวิธีเชิงพันธุกรรม

High-dimensional Linear Regression Analysis by using Genetic Algorithm

ปณัช อาภาวุฒิชัย* วิชิต หล่อจีระชุณห์กุล และ จิราวัลย์ จิตรถเวช

Panaj Abhavudhichai*, Vichit Lorchirachoonkul and Jirawan Jitthavech

สาขาสถิติ คณะสถิติประยุกต์ สถาบันบัณฑิตพัฒนบริหารศาสตร์

Department of Statistics, School of Applied Statistics, National Institute of Development Administration

Received : 14 December 2017

Accepted : 24 January 2018

Published online : 31 January 2018

บทคัดย่อ

งานวิจัยมีวัตถุประสงค์เพื่อศึกษาวิธีการประมาณค่าพารามิเตอร์และคัดเลือกตัวแปรในการวิเคราะห์การถดถอยเชิงเส้นกรณีที่ข้อมูลมีมิติสูงโดยใช้ขั้นตอนวิธีเชิงพันธุกรรม และนำผลของวิธีที่เสนอไปเปรียบเทียบกับวิธีที่รู้จักกันอย่างแพร่หลาย 3 วิธี ได้แก่วิธีลาสโซ่ วิธีอีลาสติกเน็ต และวิธีการถดถอยแบบขั้นตอน โดยใช้วิธีการจำลอง เกณฑ์ที่ใช้ในการพิจารณาเปรียบเทียบวิธีที่ศึกษาคือร้อยละของการคัดเลือกตัวแปรอิสระได้ถูกต้อง ร้อยละของการคัดเลือกตัวแปรอิสระมากเกินไป ร้อยละของการคัดเลือกตัวแปรอิสระน้อยเกินไป และร้อยละของการคัดเลือกตัวแปรอิสระไม่ถูกต้อง รวมทั้งค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของสมการถดถอยและความถูกต้องของค่าประมาณพารามิเตอร์ของตัวแบบ ผลการศึกษาสรุปได้ว่าขั้นตอนวิธีเชิงพันธุกรรมสามารถประมาณค่าพารามิเตอร์และคัดเลือกตัวแปรอิสระได้ดีที่สุดเมื่อเทียบกับ 3 วิธีดังกล่าวเกือบทุกกรณี

คำสำคัญ : การคัดเลือกตัวแปร ขั้นตอนวิธีเชิงพันธุกรรม ข้อมูลที่มีมิติสูง การวิเคราะห์การถดถอยเชิงเส้น

Abstract

The research objective is to study the effectiveness of parameter estimation and variable selection by using genetic algorithm in the high-dimensional linear regression analysis. The results of the proposed method from the simulation are compared with the other three well-known methods: lasso, elastic net, and stepwise regression. The comparison criteria are the percentage of the number of correct fitting models, the percentage of the number of over-fitting models, the percentage of the number of under-fitting models, the percentage of the number of incorrect fitting models including mean squared error and the accuracy of the parameter estimates. It can be concluded that the direct selection by genetic algorithm yields the best results when compared with the other three methods in nearly all cases.

Keywords : variable selection, genetic algorithm, high-dimensional data, linear regression analysis

* **Corresponding author. E-mail** : p.abhavudhichai@gmail.com

บทนำ

ในปัจจุบันเทคโนโลยีการจัดเก็บข้อมูลพัฒนาไปอย่างรวดเร็วและมีประสิทธิภาพ ปริมาณข้อมูลรวมทั้งจำนวนตัวแปรที่ต้องพิจารณามีมากขึ้นกว่าในอดีตอย่างมหาศาล จนอาจก่อให้เกิดปัญหาของมิติข้อมูล เนื่องจากการมีตัวแปรจำนวนมากทำให้มีความยากในการวิเคราะห์และสิ้นเปลืองทรัพยากรในการคัดเลือกตัวแปรอิสระเกี่ยวข้อง (Relevant Variables) โดยเฉพาะในกรณีที่มีจำนวนตัวแปรอิสระมากกว่าขนาดตัวอย่าง หรือที่เรียกว่าข้อมูลมีมิติสูง (High-Dimensional Data) ซึ่งวิธีการวิเคราะห์ทางสถิติจำนวนไม่น้อยได้รับผลกระทบและไม่สามารถนำมาใช้ได้ รวมถึงการวิเคราะห์การถดถอยเชิงเส้น อันเป็นวิธีการวิเคราะห์ทางสถิติที่นิยมใช้กันอย่างแพร่หลายซึ่งเกี่ยวข้องกับการสร้างตัวแบบเพื่อการอธิบายความสัมพันธ์เชิงเส้นของตัวแปรอิสระที่มีต่อตัวแปรตาม และเพื่อการทำนายและสถิติอนุมานอื่น ๆ ของตัวแปรตาม (Jitthavech, 2015) ในกรณีที่ข้อมูลมีมิติสูง ปัญหาที่เกิดขึ้นคือไม่สามารถประมาณค่าพารามิเตอร์ของตัวแบบการถดถอยด้วยวิธีกำลังสองน้อยที่สุดหรือวิธีภาวะน่าจะเป็นสูงสุด เพราะว่าเมทริกซ์จัตุรัสในสมการปรกติกลายเป็นเมทริกซ์เอกฐาน ส่งผลให้วิธีการคัดเลือกตัวแปรอิสระที่มีอยู่ในโปรแกรมสำเร็จรูปทางสถิติทั่วไปหลายวิธี เช่น วิธีพิจารณาทุกตัวแบบที่เป็นไปได้ วิธีเลือกตัวแปรอิสระแบบไปข้างหน้า วิธีตัดตัวแปรอิสระออกแบบถอยหลัง และวิธีการถดถอยแบบขั้นตอน (Montgomery, Peck, & Vining, 2006) มีปัญหาหรือข้อจำกัดในการคัดเลือกตัวแปรอิสระ ณ ขั้นตอนที่มีจำนวนพารามิเตอร์ที่ต้องประมาณค่ามากกว่าขนาดตัวอย่าง นอกจากนี้ยังไม่สามารถทดสอบสมมุติฐานเกี่ยวกับพารามิเตอร์ของตัวแบบโดยใช้ตัวสถิติทดสอบทีและเอฟ เนื่องจากตัวสถิติทดสอบดังกล่าวถูกพัฒนาขึ้นจากแนวคิดการทดสอบด้วยอัตราส่วนภาวะน่าจะเป็น (Likelihood Ratio Test) ซึ่งอ้างอิงตัวประมาณจากวิธีกำลังสองน้อยที่สุด (Pungpapong, 2015) และการมีจำนวนตัวแปรอิสระมากทำให้มีโอกาสสูงที่ตัวแปรอิสระเหล่านั้นจะมีความสัมพันธ์กันจนเกิดปัญหาความสัมพันธ์เชิงเส้นแบบพหุ (Multicollinearity) ส่งผลให้ค่าประมาณพารามิเตอร์ของตัวแบบไม่เสถียรและความแปรปรวนของตัวประมาณมีค่ามาก แม้ไม่ใช่กรณีที่ข้อมูลมีมิติสูง ดังนั้นในสองทศวรรษที่ผ่านมาจึงมีผู้เสนอแนวทางการแก้ไขหรือลดปัญหากรณีที่ข้อมูลมีมิติสูงสำหรับการวิเคราะห์การถดถอยตลอดจนสามารถขยายไปสู่ตัวแบบเชิงเส้นนัยทั่วไป (Generalized Linear Model) แนวทางที่ได้รับความนิยมในปัจจุบัน ได้แก่ การทำเร็กกิวลาร์ไรซ์เซชัน (Regularization) เช่น วิธีการถดถอยที่ปรับด้วยฟังก์ชันการลงโทษ (Penalized Regression) ซึ่งแต่ละวิธีใช้ฟังก์ชันการลงโทษ (Penalty Function) แตกต่างกัน หรือวิธีตัวคัดเลือกแดนท์ซิก (Dantzig Selector) (Candès & Tao, 2007) และอีกแนวทางหนึ่งคือวิธีแบบเบส์ (Bayesian Approach) ที่ใช้ขั้นตอนวิธี ลูกโซ่มาร์คอฟ มอนติ คาร์โล (Markov Chain Monte Carlo) หรือการทำซ้ำแบบมีเงื่อนไขฐานนิยม/มัธยฐาน (Iterated Conditional Modes/Medians) (Pungpapong, 2012) เป็นเครื่องมือในการวิเคราะห์ ในทางปฏิบัติวิธีเหล่านี้สามารถแก้ไขปัญหาได้ในระดับหนึ่ง แต่มีความเหมาะสมของการนำไปใช้แตกต่างกันขึ้นอยู่กับขนาดของข้อมูลและจำนวนตัวแปรอิสระ และมีเพียงบางวิธีเท่านั้นที่มีความสามารถในการคัดเลือกตัวแปรอิสระ เช่น วิธีลาสโซ่ (Lasso) (Tibshirani, 1996) วิธีอีลาสติกเน็ต (Elastic Net) (Zou & Hastie, 2005) และวิธีแบบเบส์ที่ใช้การแจกแจงก่อน (Prior Distribution) แบบสไปค์แอนด์สแลบ (Spike-and-slab) (Ishwaran & Rao, 2005) เป็นต้น

นอกจากการใช้วิธีเชิงคณิตศาสตร์ที่เป็นวิธีแม่นตรง (Exact Method) เช่นในวิธีกำลังสองน้อยที่สุดหรือวิธีภาวะน่าจะเป็นสูงสุด การประมาณค่าพารามิเตอร์ของตัวแบบการถดถอยสามารถใช้วิธีการค้นหาโดยตรง (Direct Search) ซึ่งเป็นวิธีหาคำตอบแบบเมตาฮิวริสติก (Metaheuristic) เพื่อแก้ปัญหาการหาค่าเหมาะที่สุด (Optimization Problem) ที่ไม่ได้อาศัยข้อสนเทศจากการหาเกรเดียนท์ (Gradient) หรืออนุพันธ์ ของฟังก์ชันเป้าหมาย (Objective Function) ในการค้นหาจุดเหมาะ

ที่สุด (Optimum Point) และพัฒนาขึ้นเพื่อจัดการกับปัญหาที่มีความซับซ้อนไม่สามารถนำวิธีแม่นตรงมาใช้แก้ปัญหาได้โดยง่ายหรือมีข้อจำกัดบางประการ เช่น ขั้นตอนวิธีเชิงพันธุกรรม (Genetic Algorithm) (Holland, 1973, 1975) วิธีการจำลองการอบเหนียว (Simulated Annealing) (Kirkpatrick, Gelatt, & Vecchi, 1983) วิธีการค้นหาแบบต้องห้าม (Tabu Search) (Glover, 1986, 1989, 1990) เป็นต้น วิธีการหาค่าเหมาะที่สุดเหล่านี้สามารถใช้คัดเลือกตัวแปรอิสระที่เหมาะสมได้ โดยมีผู้สนใจนำมาใช้ในการคัดเลือกตัวแปรอิสระสำหรับการวิเคราะห์การถดถอยเชิงเส้น เช่นเดรซเนอร์และจอร์จ (Drezner & George, 1999) ใช้วิธีการค้นหาแบบต้องห้ามเทียบกับวิธีการถดถอยแบบขั้นตอนและวิธีการปรับปรุงค่า R^2 สูงสุด (Maximum R^2 Improvement) กานต์ณัฐ ณ บางช้าง (Na Bangchang, 2011) ศึกษากรณีที่มีปัญหาความสัมพันธ์เชิงเส้นแบบพหุโดยเปรียบเทียบวิธีการถดถอยแบบขั้นตอนกับวิธีการค้นหาแบบต้องห้ามที่มีฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยและค่าเฉลี่ยความคลาดเคลื่อนสัมบูรณ์ นิศาชล งามประเสริฐสิทธิ์ (Ngamprasertsit, 2012) ศึกษากรณีที่มีปัญหาความสัมพันธ์เชิงเส้นแบบพหุโดยเปรียบเทียบวิธีการถดถอยแบบขั้นตอนซึ่งประมาณค่าพารามิเตอร์ของตัวแบบการถดถอยด้วยวิธีกำลังสองน้อยที่สุดและวิธีการถดถอยแบบริดจ์ (Ridge Regression) กับวิธีการค้นหาแบบต้องห้ามที่มีฟังก์ชันเป้าหมายเป็นค่าความคลาดเคลื่อนกำลังสองเฉลี่ยและค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่ปรับด้วยฟังก์ชันการลงโทษให้คล้ายกับการทำ ℓ_2 เร็กกิวลาร์ไรซ์เซชันในวิธีการถดถอยแบบริดจ์ และสำหรับงานวิจัยนี้ผู้วิจัยต้องการศึกษาการใช้ขั้นตอนวิธีเชิงพันธุกรรมประมาณค่าพารามิเตอร์และคัดเลือกตัวแปรอิสระ เพื่อเพิ่มแนวทางการจัดการกับปัญหาในกรณีที่ข้อมูลมีมิติสูง

วิธีดำเนินการวิจัย

งานวิจัยนี้ ศึกษาตัวแปรตามและตัวแปรอิสระในรูปความสัมพันธ์เชิงเส้น ภายใต้ข้อสมมุติ $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ โดยมีตัวแบบการถดถอยเป็น

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad (1)$$

เมื่อ $\mathbf{y} = (y_1, y_2, ..., y_n)'$ เป็นเวกเตอร์ของตัวแปรตามขนาด $n \times 1$

$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)'$ เป็นเมทริกซ์ของตัวแปรอิสระขนาด $n \times p$ โดยที่ $\mathbf{x}_i = (1, x_{i1}, x_{i2}, ..., x_{ik})'; i = 1, 2, ..., n$

$\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_k)'$ เป็นเวกเตอร์ของพารามิเตอร์ของตัวแบบขนาด $p \times 1$

$\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)'$ เป็นเวกเตอร์ของความคลาดเคลื่อนเชิงสุ่มขนาด $n \times 1$

และ n เป็นขนาดตัวอย่าง k เป็นจำนวนตัวแปรอิสระในตัวแบบ p เป็นจำนวนพารามิเตอร์ของตัวแบบ ซึ่ง $p = k + 1$

ข้อมูลที่ใช้ในการวิจัยได้มาจากการจำลองแบบมอนติ คาร์โล (Monte Carlo Simulation) ซึ่งทำซ้ำจำนวน 100 ครั้งในแต่ละกรณีโดยใช้โปรแกรม SAS เวอร์ชัน 9.4M3 มีขั้นตอนในการดำเนินการวิจัยดังนี้

1) สร้างประชากรที่มีตัวแปรตามและตัวแปรอิสระ ขนาด $N = 1,000,000$

2) กำหนดตัวแปรอิสระจำนวน 79 ตัวแปร ให้ไม่มีความสัมพันธ์กันและมีการแจกแจงเอกรูป โดย $X_1 \sim U(2,11)$, $X_2 \sim U(6.5,13.5)$, $X_3 \sim U(3.5,9.5)$, $X_4 \sim U(4,9)$, $X_5 \sim U(4.5,12.5)$, $X_6 \sim U(9,18)$, $X_7 \sim U(3,8)$, $X_8 \sim U(5.5,10.5)$, $X_9 \sim U(9.5,19.5)$, $X_{10} \sim U(6.5,15.5)$, $X_{11} \sim U(6,15)$, $X_{12} \sim U(7.5,14.5)$, $X_{13} \sim U(10,17)$, $X_{14} \sim U(8,16)$, $X_{15} \sim U(5.5,13.5)$, $X_{16} \sim U(9.5,16.5)$, $X_{17} \sim U(9,18)$, $X_{18} \sim U(2.5,11.5)$, $X_{19} \sim U(4.5,13.5)$, $X_{20} \sim U(8,14)$, $X_{21} \sim U(9,16)$, $X_{22} \sim U(6,12)$, $X_{23} \sim U(6.5,12.5)$, $X_{24} \sim U(1.5,8.5)$, $X_{25} \sim U(6,14)$,

$X_{26} \sim U(3,9), X_{27} \sim U(9,16), X_{28} \sim U(4.5,14.5), X_{29} \sim U(7,15), X_{30} \sim U(3,9), X_{31} \sim U(2,10), X_{32} \sim U(8.5,15.5), X_{33} \sim U(9,19),$
$X_{34} \sim U(2.5,8.5), X_{35} \sim U(8.5,16.5), X_{36} \sim U(4,12), X_{37} \sim U(6,12), X_{38} \sim U(4.5,14.5), X_{39} \sim U(3,9), X_{40} \sim U(4.5,13.5), X_{41} \sim U(3.5,9.5),$
$X_{42} \sim U(3,13), X_{43} \sim U(7,14), X_{44} \sim U(4,10), X_{45} \sim U(7.5,12.5), X_{46} \sim U(6,15), X_{47} \sim U(8,17), X_{48} \sim U(1.5,9.5), X_{49} \sim U(6,14),$
$X_{50} \sim U(8,14), X_{51} \sim U(9,19), X_{52} \sim U(3.5,9.5), X_{53} \sim U(2.5,11.5), X_{54} \sim U(5.5,13.5), X_{55} \sim U(4.5,12.5), X_{56} \sim U(5.5,12.5), X_{57} \sim U(9.5,17.5),$
$X_{58} \sim U(9.5,16.5), X_{59} \sim U(8.5,15.5), X_{60} \sim U(1.5,9.5), X_{61} \sim U(8,15), X_{62} \sim U(9,18), X_{63} \sim U(5.5,13.5), X_{64} \sim U(3,13), X_{65} \sim U(9.5,19.5),$
$X_{66} \sim U(3.5,10.5), X_{67} \sim U(4,14), X_{68} \sim U(5,15), X_{69} \sim U(5.5,13.5), X_{70} \sim U(8.5,17.5), X_{71} \sim U(9.5,18.5), X_{72} \sim U(3.5,10.5), X_{73} \sim U(3.5,13.5),$
$X_{73} \sim U(3.5,13.5), X_{74} \sim U(9,15), X_{75} \sim U(5,11), X_{76} \sim U(6.5,14.5), X_{77} \sim U(2,12), X_{78} \sim U(7.5,13.5)$ และ $X_{79} \sim U(2,8)$

3) สร้างความคลาดเคลื่อนเชิงสุ่มที่เป็นอิสระต่อกัน มีการแจกแจงปรกติ ค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนคงที่เท่ากับ 10^2 นั่นคือ $\varepsilon \sim N(0,10^2)$

4) กำหนดกรณีต่าง ๆ 8 กรณี โดยมี n, p และจำนวนตัวแปรอิสระเกี่ยวข้องรวมกับพจน์คงที่ z_1 ซึ่งเป็นจำนวนเต็มที่มีค่าเท่ากับร้อยละ 20 และร้อยละ 80 ของ p โดยที่อัตราส่วน $n:p$ เป็น 1:2 และ 1:4 ดังนี้ $n=5, p=10, z_1=2\,(z_0=8)$; $n=5, p=10, z_1=8\,(z_0=2)$; $n=5, p=20, z_1=4\,(z_0=16)$; $n=5, p=20, z_1=16\,(z_0=4)$; $n=20, p=40, z_1=8\,(z_0=32)$; $n=20, p=40, z_1=32\,(z_0=8)$; $n=20, p=80, z_1=16\,(z_0=64)$ และ $n=20, p=80, z_1=64\,(z_0=16)$ เมื่อ z_0 เป็นจำนวนตัวแปรอิสระไม่เกี่ยวข้องซึ่งมีค่าเท่ากับ $p-z_1$

5) กำหนดค่าพารามิเตอร์ของตัวแบบ $\beta_j = b_j\,;\, j=0,1,2,...,k-z_0$ และ $\beta_j = 0\,;\, j=z_1, z_1+1, z_1+2, ..., k$ เมื่อ $(b_j)_{j=0}^{63}$ มีค่า $(100,16,-10,15,18,-6,-7,-12,20,-2,-10,6,2,19,4,-20,-11,14,13,-9,7,-3,5,-16,17,-18,12,-11,13,14,-19,-12,8,7,10,-16,-13,6,3,20,5,9,-14,5,6,-11,10,-15,16,8,1,13,-5,17,7,-11,2,10,19,-6,-14,6,20,-17)$

6) สร้างตัวแปรตามของตัวแบบเต็มรูป (Full Model) $y=\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_kx_k+\varepsilon$ ทั้งหมด 8 กรณี ซึ่งทำให้ได้ตัวแบบจริง (True Model) ในแต่ละกรณีเป็น $y=\beta_0+\beta_1x_1+\beta_2x_2+...+\beta_{z_1-1}x_{z_1-1}+\varepsilon$

7) สุ่มตัวอย่างขนาดตามที่กำหนดในข้อ 4)

8) นำข้อมูลที่ได้จากข้อ 7) มาประมาณค่าพารามิเตอร์ของตัวแบบและคัดเลือกตัวแปรอิสระเพื่อสร้างสมการถดถอย โดยใช้วิธีลาสโซ่และวิธีอีลาสติกเน็ตซึ่งหาค่าที่เหมาะสมของพารามิเตอร์ของฟังก์ชันการลงโทษ (Tuning Parameter: λ) ด้วยวิธีตรวจสอบไขว้ 5 ทบ (5-fold Cross Validation) วิธีการถดถอยแบบขั้นตอนซึ่งประมาณค่าพารามิเตอร์ของตัวแบบด้วยวิธีกำลังสองน้อยที่สุด และวิธีคัดเลือกตัวแปรอิสระที่ค้นหาค่าประมาณพารามิเตอร์ของตัวแบบด้วยขั้นตอนวิธีเชิงพันธุกรรม

9) ทำซ้ำในแต่ละกรณีจำนวน 100 ครั้ง แล้วคำนวณเกณฑ์ที่ใช้ในการพิจารณา เพื่อเปรียบเทียบวิธีที่ศึกษา

วิธีการประมาณค่าพารามิเตอร์และคัดเลือกตัวแปรอิสระ

1. วิธีกำลังสองน้อยที่สุด

การประมาณค่าพารามิเตอร์ของตัวแบบการถดถอยด้วยวิธีกำลังสองน้อยที่สุดเป็นการแก้ปัญหาการหาค่าเหมาะที่สุดแบบไม่มีเงื่อนไขบังคับ (Unconstrained Optimization Problem) ในรูป $\text{minimize } \|\boldsymbol{\varepsilon}\|_2^2$ หรือ $\text{minimize } \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2$ โดยใช้เงื่อนไขการหาอนุพันธ์เท่ากับศูนย์ ณ จุดต่ำสุดของฟังก์ชันเป้าหมาย ซึ่งสามารถทำให้อยู่ในรูปสมการปรกติ และเมื่อ $\mathbf{X'X}$ เป็นเมทริกซ์ไม่เอกฐาน ได้ตัวประมาณค่าพารามิเตอร์เท่ากับ

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{2}$$

2. วิธีลาสโซ่

การประมาณค่าพารามิเตอร์ของตัวแบบการถดถอยที่ปรับด้วยฟังก์ชันการลงโทษเป็นการทำเร็กกิวลาร์ไรซ์เซชันโดยเพิ่มเงื่อนไขบังคับ (Constraint) ของฟังก์ชันการลงโทษ นั่นคือ $\text{minimize}\ \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2$ ภายใต้เงื่อนไขบังคับ $\mathrm{P}(\boldsymbol{\beta}) < t$ เมื่อ $\mathrm{P}(\boldsymbol{\beta})$ แทนฟังก์ชันการลงโทษ ซึ่งไม่นิยมให้มีการลงโทษพจน์คงที่ (Intercept: β_0) ใน $\mathrm{P}(\boldsymbol{\beta})$ (ถ้าตัวแบบมี β_0) และค่า t มีความสัมพันธ์แบบหนึ่งต่อหนึ่งกับค่า λ วิธีลาสโซ่ก็คือวิธีการถดถอยที่ปรับด้วยฟังก์ชันการลงโทษซึ่งใช้ ℓ_1 นอร์ม (Norm) ใน $\mathrm{P}(\boldsymbol{\beta})$ มีตัวประมาณค่าพารามิเตอร์อยู่ในรูป

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\ ;\ \lambda \geq 0 \tag{3}$$

โดยการหาค่า $\hat{\boldsymbol{\beta}}$ ใน (3) นิยมใช้วิธีการถดถอยมุมน้อยที่สุด (Least Angle Regression: LARS) (Efron *et al.*, 2004) และด้วยธรรมชาติของ ℓ_1 นอร์ม $\hat{\boldsymbol{\beta}}$ ที่ได้จาก (3) มีลักษณะเป็นเวกเตอร์เบาบาง (Sparse Estimatior) ทำให้วิธีลาสโซ่สามารถคัดเลือกตัวแปรอิสระไปพร้อมกับประมาณค่าพารามิเตอร์ของตัวแบบ

3. วิธีอีลาสติกเน็ต

วิธีลาสโซ่มีข้อจำกัดคือสามารถคัดเลือกตัวแปรอิสระได้มากที่สุด n ตัวแปรจากทั้งหมด k ตัวแปร เมื่อเป็นกรณีที่ $k > n$ และหากตัวแปรอิสระมีความสัมพันธ์กันสูงเป็นกลุ่ม ๆ ก็มีแนวโน้มที่จะเลือกตัวแปรอิสระตัวใดตัวหนึ่งในกลุ่มเพียงตัวเดียว ดังนั้นวิธีอีลาสติกเน็ตจึงถูกพัฒนาขึ้นเพื่อแก้ไขข้อจำกัดดังกล่าว ด้วยการใช้ ℓ_1 นอร์มร่วมกับ ℓ_2 นอร์มใน $\mathrm{P}(\boldsymbol{\beta})$ ได้ตัวประมาณค่าพารามิเตอร์อยู่ในรูป

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\ \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2;\ \lambda_1 \geq 0, \lambda_2 \geq 0 \tag{4}$$

วิธีอีลาสติกเน็ตเป็นการรวมข้อดีของวิธีลาสโซ่กับวิธีการถดถอยแบบริดจ์เข้าด้วยกัน ทำให้สามารถคัดเลือกตัวแปรอิสระไปพร้อมกับประมาณค่าพารามิเตอร์ของตัวแบบ รวมทั้งมีความเหมาะสมในกรณี $k >> n$ และ/หรือตัวแปรอิสระมีความสัมพันธ์กันสูง โดยการหาค่า $\hat{\boldsymbol{\beta}}$ ใน (4) สามารถใช้วิธีการถดถอยมุมน้อยที่สุดสำหรับวิธีอีลาสติกเน็ต (LARS-EN) ได้

4. วิธีการถดถอยแบบขั้นตอน

วิธีการถดถอยแบบขั้นตอนเป็นการผสมผสานระหว่างวิธีเลือกตัวแปรอิสระแบบไปข้างหน้ากับวิธีตัดตัวแปรอิสระออกแบบถอยหลัง โดยพิจารณาเลือกตัวแปรอิสระเข้าสู่ตัวแบบการถดถอยครั้งละ 1 ตัวจากค่าสัมประสิทธิ์สหสัมพันธ์กับตัวแปรตาม และอาจมีการตัดตัวแปรอิสระที่เลือกไว้ออกภายหลังถ้าพบว่าไม่มีนัยสำคัญ ในงานวิจัยนี้กำหนดระดับนัยสำคัญของการนำตัวแปรอิสระเข้า α_1 เท่ากับ 0.15 เพื่อให้สามารถนำตัวแปรอิสระเข้าได้ง่าย และกำหนดระดับนัยสำคัญของการตัดตัวแปรอิสระออก α_2 เท่ากับ 0.05 ซึ่งเท่ากับระดับนัยสำคัญที่กำหนดในขั้นตอนวิธีเชิงพันธุกรรม เพื่อให้การเปรียบเทียบอยู่บน

พื้นฐานเดียวกัน อย่างไรก็ตาม กระบวนการคัดเลือกดังกล่าวจะหยุดลงทันทีหากได้สมการถดถอยที่มีจำนวนตัวแปรอิสระเท่ากับจำนวนค่าสังเกต

5. ขั้นตอนวิธีเชิงพันธุกรรม

ขั้นตอนวิธีเชิงพันธุกรรมเป็นวิธีการทางปัญญาประดิษฐ์ (Artificial Intelligence) วิธีหนึ่งที่เลียนแบบกระบวนการคัดเลือกตามธรรมชาติ (Natural Selection) ในทฤษฎีวิวัฒนาการของดาร์วิน (Darwin, 1859) เพื่อค้นหาคำตอบเหมาะที่สุด ซึ่งเป็นวิธีที่มีความแกร่ง (Robust) สามารถนำไปประยุกต์ใช้ในการแก้ปัญหาได้หลากหลายรูปแบบ ไม่มีความจำเพาะกับตัวแบบหรือลักษณะข้อมูลแบบใดแบบหนึ่ง (Goldberg, 1989) นอกจากนี้ยังเป็นวิธีการค้นหาโดยตรงที่ไม่ได้พิจารณาคำตอบที่เป็นไปได้ (Candidate Solutions) ทีละจุดเพื่อปรับทิศทางการค้นหาเฉพาะบริเวณข้างเคียง (Neighbour) ของคำตอบเดิมเท่านั้น แต่เป็นการค้นหาแบบขนาน (Parallel Search) ที่พิจารณาคำตอบที่เป็นไปได้หลายจุดพร้อมกัน ทำให้มีความเหมาะสมในการค้นหาคำตอบที่มีปริภูมิการค้นหา (Search Space) กว้าง และช่วยลดโอกาสที่กระบวนการค้นหาจะลู่เข้า (Converge) ไปสู่จุดเหมาะที่สุดเฉพาะที่ (Local Optimum) โดยมีขั้นตอนพื้นฐานในการทำงานสรุปได้ดังนี้

5.1 ขั้นเริ่มแรก (Initialization) สร้างประชากรของคำตอบเริ่มต้นโดยสุ่มค่าภายในปริภูมิการค้นหาจนครบจำนวนที่กำหนด และคำนวณค่าฟังก์ชันเป้าหมายของสมาชิกทุกตัวในประชากรของคำตอบ

5.2 ขั้นการสืบพันธุ์ (Regeneration) เป็นขั้นตอนที่นำประชากรของคำตอบในรุ่น (Generation) ปัจจุบันไปผ่านตัวดำเนินการทางพันธุกรรม (Genetic Operators) ได้แก่ การคัดเลือก (Selection) การไขว้เปลี่ยน (Crossover) และการกลายพันธุ์ (Mutation) เพื่อสร้างประชากรของคำตอบในรุ่นถัดไปขึ้นมาแทนที่ โดยมีจุดประสงค์ในการวิวัฒน์ประชากรของคำตอบในแต่ละรุ่นให้มีความเหมาะสมมากยิ่งขึ้นจนสามารถลู่เข้าไปสู่จุดเหมาะที่สุดทั่วไป (Global Optimum) ได้ จากนั้นให้คำนวณค่าฟังก์ชันเป้าหมายของสมาชิกทุกตัวในประชากรของคำตอบรุ่นใหม่ที่สร้างขึ้น

5.3 ขั้นการทำซ้ำ ให้ตรวจสอบเกณฑ์ที่ใช้หยุด (Stopping Criteria) ถ้าเงื่อนไขไม่สอดคล้องกับเกณฑ์ดังกล่าวจะทำขั้นตอนที่ 5.2 ซ้ำจนกระทั่งเงื่อนไขสอดคล้องกับเกณฑ์จึงหยุดกระบวนการค้นหา และได้ผลลัพธ์เป็นสมาชิกตัวที่มีค่าฟังก์ชันเป้าหมายดีที่สุดจากประชากรของคำตอบในรุ่นสุดท้าย

การคัดเลือกตัวแปรอิสระโดยใช้ขั้นตอนวิธีเชิงพันธุกรรมในงานวิจัยนี้มีขั้นตอนคือ

1) ใช้ข้อมูลขนาด n สุ่มตัวอย่างซ้ำด้วยวิธีแจ็คไนฟ์ (Jackknife Resampling) (Quenouille, 1949, 1956; Tukey, 1958) จนได้ตัวอย่างสุ่มซ้ำขนาด $n-1$ จำนวน n ตัวอย่าง

2) ในแต่ละตัวอย่างขนาด $n-1$ นำไปหาค่าประมาณพารามิเตอร์ของตัวแบบด้วยขั้นตอนวิธีเชิงพันธุกรรม โดยใช้ตัวแปรอิสระ x_j ทั้งหมด k ตัวแปร และกำหนดค่าพารามิเตอร์ของขั้นตอนวิธีเชิงพันธุกรรมดังต่อไปนี้

2.1) ปริภูมิการค้นหาของ $\hat{\beta}_j$ แต่ละตัวคือช่วงปิด $[\beta_j-(\beta_j \bmod 5)-5, \beta_j-(\beta_j \bmod 5)+5]$ เนื่องจากไม่ต้องการให้ β_j เป็นจุดกึ่งกลางของปริภูมิการค้นหาสำหรับแต่ละ $\hat{\beta}_j$ เสมอไป

2.2) ขนาดประชากรของคำตอบเท่ากับ 500

2.3) ฟังก์ชันเป้าหมายใช้ค่าผลรวมกำลังสองของส่วนเหลือ (Residual Sum of Squares: RSS) (f) ซึ่งค้นหา $\hat{\boldsymbol{\beta}}$ ที่ทำให้ฟังก์ชันเป้าหมายมีค่าต่ำที่สุด โดยที่

$$f = \left\| \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \right\|_2^2 \quad (5)$$

2.4) ใช้ค่าจริงของคำตอบในตัวดำเนินการทางพันธุกรรม (Real-coded Genetic Algorithm)

2.5) การคัดเลือกใช้วิธีการคัดเลือกแบบทัวร์นาเมนท์ (Tournament Selection) (Miller & Goldberg, 1995) โดยมีขนาดทัวร์นาเมนท์ (Tournament Size) เป็น 2 ความน่าจะเป็นที่ผู้ชนะถูกคัดเลือก (Best-player-wins Probability) เท่ากับ 0.9

2.6) ใช้แนวคิดอติชนนิยม (Elitism) (De Jong, 1975) จำนวนสมาชิกอติชน (Elite Parameter) เท่ากับ 5

2.7) การไขว้เปลี่ยนใช้วิธีการแบบฮิวริสติก (Heuristic) (Wright, 1991) ความน่าจะเป็นที่จะเกิดการไขว้เปลี่ยน (Crossover Probability) เท่ากับ 0.9

2.8) การกลายพันธุ์ใช้วิธีการแบบเดลต้า (Delta) (Whitley, Mathias, & Fitzhorn, 1991) ค่าเดลต้า (δ) เท่ากับ 0.01 และความน่าจะเป็นที่จะเกิดการกลายพันธุ์ (Mutation Probability) เท่ากับ 0.05

2.9) เกณฑ์ที่ใช้หยุดคือจำนวนรอบสูงสุด ซึ่งกำหนดให้เท่ากับ $4np$ หรือค่าฟังก์ชันเป้าหมายไม่เปลี่ยนแปลงครบจำนวนรอบที่กำหนดคือ 5% ของจำนวนรอบสูงสุด โดยกำหนดความแม่นยำของค่าฟังก์ชันเป้าหมายเท่ากับ 10^{-4}

3) นำค่าประมาณพารามิเตอร์ของตัวแบบที่ได้จากตัวอย่างขนาด $n-1$ ทั้ง n ตัวอย่างไปหาค่าเฉลี่ยกับส่วนเบี่ยงเบนมาตรฐาน สำหรับแต่ละ $\hat{\beta}_j$ โดยที่ $\mathrm{SD}(\hat{\beta}_j) = \sqrt{\frac{\sum_{s=1}^{n} (\hat{\beta}_{js} - \bar{\hat{\beta}}_j)^2}{n-1}}$; $\bar{\hat{\beta}}_j = \frac{\sum_{s=1}^{n} \hat{\beta}_{js}}{n}$ จากนั้นคำนวณความคลาดเคลื่อนมาตรฐาน $\mathrm{SE}(\bar{\hat{\beta}}_j) = \frac{\mathrm{SD}(\hat{\beta}_j)}{\sqrt{n}}$ กับตัวสถิติทดสอบที $t = \frac{\bar{\hat{\beta}}_j}{\mathrm{SE}(\bar{\hat{\beta}}_j)}$ เพื่อทดสอบสมมุติฐาน $\mathrm{H}_0: \beta_j = 0$ เทียบกับ $\mathrm{H}_1: \beta_j \neq 0$ มีขอบเขตวิกฤตคือ $|t| > t_{\alpha/2, n-1}$ และหาค่าพี (p-Value) เพื่อใช้คัดเลือกตัวแปรอิสระ โดยให้พิจารณาค่าพีที่สมนัยกับ x_j แต่ละตัวแล้วตัด x_j ที่มีค่าพีมากที่สุดออกไป 1 ตัว หากค่าพีไม่น้อยกว่าระดับนัยสำคัญ 0.05 (การตัด x_j สมมูลกับการประมาณค่า β_j ด้วย 0)

4) ทำซ้ำตั้งแต่ขั้นตอนที่ 2) แต่พิจารณาเฉพาะ x_j ที่ยังไม่ถูกตัดออกไป จนกระทั่งค่าพีของ x_j ทุกตัวน้อยกว่าระดับนัยสำคัญ 0.05 กระบวนการคัดเลือกตัวแปรอิสระจึงหยุดลงและได้ $\bar{\hat{\beta}}_j$ ในขั้นสุดท้ายเป็นค่าประมาณของ β_j

เกณฑ์ที่ใช้ในการพิจารณาเพื่อเปรียบเทียบผลการจำลอง

1. เกณฑ์การคัดเลือกตัวแปรอิสระในสมการถดถอย

เกณฑ์การคัดเลือกตัวแปรอิสระในสมการถดถอยใช้ร้อยละของการคัดเลือกตัวแปรอิสระจากการทำซ้ำ 100 ครั้ง ดังนี้

1.1 การคัดเลือกตัวแปรอิสระได้ถูกต้อง

การคัดเลือกตัวแปรอิสระได้ถูกต้อง (Correct Fitting) หมายถึงการได้สมการถดถอยที่มีตัวแปรอิสระเกี่ยวข้องครบทุกตัวโดยไม่มีตัวแปรอิสระไม่เกี่ยวข้องอยู่ในสมการ

1.2 การคัดเลือกตัวแปรอิสระมากเกินไป

การคัดเลือกตัวแปรอิสระมากเกินไป (Over-fitting) หมายถึงการได้สมการถดถอยที่มีตัวแปรอิสระเกี่ยวข้องครบทุกตัวแต่มีตัวแปรอิสระไม่เกี่ยวข้องบางตัวรวมอยู่ในสมการ

1.3 การคัดเลือกตัวแปรอิสระน้อยเกินไป

การคัดเลือกตัวแปรอิสระน้อยเกินไป (Under-fitting) หมายถึงการได้สมการถดถอยที่มีตัวแปรอิสระเกี่ยวข้องไม่ครบทุกตัวแต่ไม่มีตัวแปรอิสระไม่เกี่ยวข้องอยู่ในสมการ

1.4 การคัดเลือกตัวแปรอิสระไม่ถูกต้อง

การคัดเลือกตัวแปรอิสระไม่ถูกต้อง (Incorrect Fitting) หมายถึงการได้สมการถดถอยที่มีตัวแปรอิสระเกี่ยวข้องไม่ครบทุกตัวและยังมีตัวแปรอิสระไม่เกี่ยวข้องบางตัวรวมอยู่ในสมการ

โดยทั่วไปการคัดเลือกตัวแปรอิสระน้อยเกินไปเป็นปัญหาที่มีความรุนแรงในการวิเคราะห์มากกว่าการคัดเลือกตัวแปรอิสระไม่ถูกต้อง เพราะการคัดเลือกตัวแปรอิสระไม่ถูกต้องมีตัวแปรอิสระเกี่ยวข้องไม่ครบทุกตัวแต่ยังมีตัวแปรอิสระไม่เกี่ยวข้องบางตัวรวมอยู่ในสมการซึ่งช่วยในการอธิบายตัวแปรตามได้เพิ่มขึ้นบางส่วน อย่างไรก็ตามการคัดเลือกตัวแปรอิสระไม่ถูกต้องก็เป็นปัญหาที่มีความรุนแรงมากกว่าการคัดเลือกตัวแปรอิสระมากเกินไป (Gujarati, 2006) ดังนั้นในการเปรียบเทียบผลการคัดเลือกตัวแปรอิสระของแต่ละวิธีว่าวิธีใดเป็นวิธีที่ดีกว่า จึงพิจารณาวิธีที่มีร้อยละของการคัดเลือกตัวแปรอิสระได้ถูกต้องมากกว่าควบคู่กับร้อยละของการคัดเลือกตัวแปรอิสระมากเกินไป โดยให้ความสำคัญกับวิธีที่มีร้อยละของการคัดเลือกตัวแปรอิสระน้อยเกินไปน้อยกว่า และวิธีที่มีร้อยละของการคัดเลือกตัวแปรอิสระไม่ถูกต้องน้อยกว่า ตามลำดับ

2. เกณฑ์ความคลาดเคลื่อนกำลังสองเฉลี่ยของสมการถดถอย

ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของสมการถดถอย จากการทำซ้ำ 100 ครั้ง คำนวณได้จากสูตรดังนี้

$$\mathrm{MSE}(\hat{y}) = \frac{1}{100}\sum_{r=1}^{100}\left(\frac{1}{n}\sum_{i=1}^{n}(y_{ir}-\hat{y}_{ir})^2\right) \qquad (6)$$

เมื่อ y_{ir} เป็นค่าสังเกตที่ i ของตัวแปรตาม ในการทำซ้ำครั้งที่ r

n เป็นขนาดตัวอย่าง และ $\hat{y}_{ir} = \hat{\beta}_{0r} + \hat{\beta}_{1r}x_{1ir} + \hat{\beta}_{2r}x_{2ir} + ... + \hat{\beta}_{kr}x_{kir}$

โดยที่ $\hat{y}_{ir}$ เป็นค่าประมาณของตัวแปรตามเมื่อใช้ค่าสังเกตที่ i ในการทำซ้ำครั้งที่ r

$\hat{\beta}_{jr}$ เป็นค่าประมาณของพารามิเตอร์ของตัวแบบ β_j ในการทำซ้ำครั้งที่ r

x_{jir} เป็นค่าสังเกตที่ i ของตัวแปรอิสระ x_j ในการทำซ้ำครั้งที่ r

ค่า $\mathrm{MSE}(\hat{y})$ เป็นค่าที่ต้องพิจารณาด้วยความระมัดระวังเพราะวิธีต่าง ๆ มีการคัดเลือกตัวแปรอิสระที่แตกต่างกันดังที่ได้กล่าวไว้ในหัวข้อวิธีการประมาณค่าพารามิเตอร์และคัดเลือกตัวแปรอิสระ

3. เกณฑ์ความถูกต้องของค่าประมาณพารามิเตอร์ของตัวแบบ

เกณฑ์ความถูกต้องของค่าประมาณพารามิเตอร์ของตัวแบบเป็นเกณฑ์ที่ใช้ได้เฉพาะการจำลอง ในการเปรียบเทียบผลการประมาณค่าพารามิเตอร์ของแต่ละวิธี วิธีที่ดีกว่าคือวิธีที่ให้ค่าประมาณพารามิเตอร์ของตัวแบบ (โดยเฉลี่ย) ใกล้เคียงกับค่าพารามิเตอร์จริงที่ใช้ในการจำลองมากกว่า (และมีความคลาดเคลื่อนมาตรฐานต่ำกว่า)

ผลการวิจัยและวิจารณ์ผล

ตารางที่ 1-5 แสดงผลการประมาณค่าพารามิเตอร์และการคัดเลือกตัวแปรอิสระในแต่ละกรณีแบ่งตาม $n:p$ และ z_1

เมื่อ $n:p$ เป็น $1:2$ ในกรณี $n=5, p=10, z_1=2$ พบว่าขั้นตอนวิธีเชิงพันธุกรรมมีการคัดเลือกตัวแปรอิสระได้ถูกต้องเพียงร้อยละ 4 การคัดเลือกตัวแปรอิสระมากเกินไปมีร้อยละ 96 ซึ่งมากกว่าวิธีลาสโซ่ วิธีอีลาสติกเน็ต และวิธีการถดถอยแบบขั้นตอน แต่ไม่มีการคัดเลือกตัวแปรอิสระน้อยเกินไปและไม่มีการคัดเลือกตัวแปรอิสระไม่ถูกต้องแบบในวิธีอื่น ๆ และให้ค่าประมาณของ $\beta_0, \beta_1, \beta_2, ..., \beta_{z_1-1}$ โดยเฉลี่ยใกล้เคียงกับค่าจริงที่จำลองและมีความคลาดเคลื่อนมาตรฐานรวมทั้ง $\text{MSE}(\hat{y})$ ต่ำที่สุด อย่างไรก็ดี ในกรณีนี้วิธีการถดถอยแบบขั้นตอนมีการคัดเลือกตัวแปรอิสระได้ถูกต้องถึงร้อยละ 61 และให้ค่าประมาณของ β_1 โดยเฉลี่ยใกล้เคียงกับค่าจริงที่จำลองมากที่สุด ทำให้ไม่สามารถสรุปได้ว่าขั้นตอนวิธีเชิงพันธุกรรมประมาณค่าพารามิเตอร์และคัดเลือกตัวแปรอิสระได้ดีกว่าวิธีการถดถอยแบบขั้นตอน แต่ขั้นตอนวิธีเชิงพันธุกรรมประมาณค่าพารามิเตอร์และคัดเลือกตัวแปรอิสระในภาพรวมได้ดีกว่าวิธีลาสโซ่และวิธีอีลาสติกเน็ต และในกรณี $n=5, p=10, z_1=8$ พบว่าขั้นตอนวิธีเชิงพันธุกรรมมีการคัดเลือกตัวแปรอิสระได้ถูกต้องถึงร้อยละ 41 ในขณะที่วิธีอื่น ๆ ไม่มีการคัดเลือกตัวแปรอิสระได้ถูกต้อง การคัดเลือกตัวแปรอิสระมากเกินไปมีร้อยละ 52 ซึ่งมากกว่าวิธีอื่น ๆ การคัดเลือกตัวแปรอิสระน้อยเกินไปและการคัดเลือกตัวแปรอิสระไม่ถูกต้องมีเพียงร้อยละ 5 และร้อยละ 2 ตามลำดับ ซึ่งน้อยกว่าวิธีอื่น ๆ อย่างชัดเจน นอกจากนี้ยังให้ค่าประมาณของ $\beta_0, \beta_1, \beta_2, ..., \beta_{z_1-1}$ โดยเฉลี่ยใกล้เคียงกับค่าจริงที่จำลองมากที่สุดและมีความคลาดเคลื่อนมาตรฐานรวมทั้ง $\text{MSE}(\hat{y})$ ต่ำที่สุด สรุปได้ว่าขั้นตอนวิธีเชิงพันธุกรรมมีการประมาณค่าพารามิเตอร์และคัดเลือกตัวแปรอิสระได้ดีที่สุดเมื่อเทียบกับวิธีอื่น ๆ (ดูรายละเอียดได้ในตารางที่ 1-2)

เมื่อ $n:p$ เป็น $1:4$ ในกรณี $n=5, p=20, z_1=4$ พบว่าขั้นตอนวิธีเชิงพันธุกรรมมีการคัดเลือกตัวแปรอิสระได้ถูกต้องเพิ่มขึ้นเป็นร้อยละ 12 และร้อยละที่เหลือเป็นการคัดเลือกตัวแปรอิสระมากเกินไปโดยไม่มีการคัดเลือกตัวแปรอิสระน้อยเกินไปและไม่มีการคัดเลือกตัวแปรอิสระไม่ถูกต้อง ซึ่งดีกว่าวิธีอื่น ๆ ที่มีการคัดเลือกตัวแปรอิสระได้ถูกต้องและการคัดเลือกตัวแปรอิสระมากเกินไปลดลง แต่มีการคัดเลือกตัวแปรอิสระน้อยเกินไปหรือการคัดเลือกตัวแปรอิสระไม่ถูกต้องเพิ่มขึ้นแทน และในกรณี $n=5, p=20, z_1=16$ พบว่าขั้นตอนวิธีเชิงพันธุกรรมมีการคัดเลือกตัวแปรอิสระในลักษณะเดียวกับวิธีอื่น ๆ ดังที่กล่าวมา คือมีการคัดเลือกตัวแปรอิสระได้ถูกต้องลดลงเหลือเพียงร้อยละ 2 เช่นเดียวกับการคัดเลือกตัวแปรอิสระมากเกินไป การคัดเลือกตัวแปรอิสระน้อยเกินไปมีเพิ่มขึ้นแต่ยังคงน้อยกว่าวิธีอื่น ๆ คือร้อยละ 36 และมีการคัดเลือกตัวแปรอิสระไม่ถูกต้องเพิ่มขึ้นเป็นร้อยละ 60 อย่างไรก็ตาม หากพิจารณาร้อยละที่ตัวแปรอิสระแต่ละตัวอยู่ในสมการถดถอย (ดูตารางที่ 4) พบว่าขั้นตอนวิธีเชิงพันธุกรรมมีตัวแปรอิสระเกี่ยวข้องเพียงไม่กี่ตัวแปรเท่านั้นที่ถูกคัดเลือกออกไปบ่อยครั้งอันเป็นสาเหตุทำให้ได้สมการถดถอยที่มีตัวแปรอิสระน้อยเกินไปหรือมีตัวแปรอิสระไม่ถูกต้องเพิ่มขึ้นแต่ตัวแปรอิสระเกี่ยวข้องตัวอื่นยังคงมีอยู่ในสมการถดถอยมากกว่าวิธีอื่น ๆ อย่างชัดเจน โดยทั้ง 2 กรณีค่าประมาณ $\beta_0, \beta_1, \beta_2, ..., \beta_{z_1-1}$ โดยเฉลี่ยยังคงใกล้เคียงกับค่าจริงที่จำลองมากที่สุดและมีความคลาดเคลื่อนมาตรฐานรวมทั้ง $\text{MSE}(\hat{y})$ ต่ำที่สุด (ดูตารางที่ 1 และตารางที่ 3-4) จึงสรุปได้ว่าขั้นตอนวิธีเชิงพันธุกรรมมีการประมาณค่าพารามิเตอร์และคัดเลือกตัวแปรอิสระได้ดีที่สุดเมื่อเทียบกับวิธีอื่น ๆ

ตารางที่ 1 ร้อยละของการคัดเลือกตัวแปรอิสระและค่าความคลาดเคลื่อนกำลังสองเฉลี่ย จากการทำซ้ำ 100 ครั้ง

วิธี	ร้อยละของการคัดเลือกตัวแปรอิสระ				MSE($\hat{y}$)
	ถูกต้อง	มากเกินไป	น้อยเกินไป	ไม่ถูกต้อง	
$n=5,\ p=10,\ z_1=2$					
ลาสโซ่	30	52	15	3	199.68
อีลาสติกเน็ต	31	64	5	-	91.95
การถดถอยแบบขั้นตอน	61	30	3	6	49.82
ขั้นตอนวิธีเชิงพันธุกรรม	4	96	-	-	17.48
$n=5,\ p=10,\ z_1=8$					
ลาสโซ่	-	-	80	20	1626.38
อีลาสติกเน็ต	-	26	46	28	810.34
การถดถอยแบบขั้นตอน	-	-	88	12	1570.63
ขั้นตอนวิธีเชิงพันธุกรรม	41	52	5	2	30.04
$n=5,\ p=20,\ z_1=4$					
ลาสโซ่	-	1	60	39	1356.06
อีลาสติกเน็ต	-	26	37	37	770.62
การถดถอยแบบขั้นตอน	1	-	41	58	421.23
ขั้นตอนวิธีเชิงพันธุกรรม	12	88	-	-	10.93
$n=5,\ p=20,\ z_1=16$					
ลาสโซ่	-	-	88	12	4998.92
อีลาสติกเน็ต	-	13	53	34	2568.00
การถดถอยแบบขั้นตอน	-	-	84	16	2763.49
ขั้นตอนวิธีเชิงพันธุกรรม	2	2	36	60	34.97
$n=20,\ p=40,\ z_1=8$					
ลาสโซ่	-	34	5	61	302.45
อีลาสติกเน็ต	-	47	3	50	168.71
การถดถอยแบบขั้นตอน	4	17	11	68	252.89
ขั้นตอนวิธีเชิงพันธุกรรม	-	100	-	-	0.44
$n=20,\ p=40,\ z_1=32$					
ลาสโซ่	-	-	50	50	7984.56
อีลาสติกเน็ต	-	45	11	44	1680.17
การถดถอยแบบขั้นตอน	-	-	46	54	2925.85
ขั้นตอนวิธีเชิงพันธุกรรม	-	14	2	84	4.52
$n=20,\ p=80,\ z_1=16$					
ลาสโซ่	-	-	28	72	3366.20
อีลาสติกเน็ต	-	22	10	68	1498.23
การถดถอยแบบขั้นตอน	-	-	-	100	285.10
ขั้นตอนวิธีเชิงพันธุกรรม	-	10	-	90	0.99
$n=20,\ p=80,\ z_1=64$					
ลาสโซ่	-	-	62	38	25895.64
อีลาสติกเน็ต	-	25	25	50	8903.89
การถดถอยแบบขั้นตอน	-	-	19	81	2965.87
ขั้นตอนวิธีเชิงพันธุกรรม	-	-	7	93	12.46

ตารางที่ 2 ค่าพารามิเตอร์ของตัวแบบร้อยละที่ตัวแปรอิสระอยู่ในสมการถดถอย ค่าเฉลี่ยของค่าประมาณพารามิเตอร์ของตัวแบบ ค่าความคลาดเคลื่อนมาตรฐาน และค่าพี ในกรณี $n=5,\ p=10,\ z_1=2$ และ $n=5,\ p=10,\ z_1=8$

j	β_j	ลาสโซ่			อีลาสติกเน็ต			การถดถอยแบบขั้นตอน			ขั้นตอนวิธีเชิงพันธุกรรม		
		ร้อยละ	$\bar{\hat{\beta}}_j$ (SE ($\bar{\hat{\beta}}_j$))	ค่าพี	ร้อยละ	$\bar{\hat{\beta}}_j$ (SE ($\bar{\hat{\beta}}_j$))	ค่าพี	ร้อยละ	$\bar{\hat{\beta}}_j$ (SE ($\bar{\hat{\beta}}_j$))	ค่าพี	ร้อยละ	$\bar{\hat{\beta}}_j$ (SE ($\bar{\hat{\beta}}_j$))	ค่าพี
$n=5,\ p=10,\ z_1=2$													
0	100	100	134.21 (5.36)	0.0000	100	129.77 (5.67)	0.0000	100	106.09 (4.96)	0.0000	100	100.16 (0.27)	0.0000
1	16	82	12.70 (0.37)	0.0000	95	11.31 (0.41)	0.0000	91	15.92 (0.26)	0.0000	100	15.79 (0.22)	0.0000
2	0	16	0.01 (1.66)	0.9941	28	0.01 (0.93)	0.9885	8	0.11 (3.18)	0.9726	44	0.44 (0.38)	0.2635
3	0	15	0.06 (0.89)	0.9483	23	0.04 (0.81)	0.9578	3	0.83 (5.95)	0.9016	28	-0.09 (0.63)	0.8901
4	0	23	0.76 (1.07)	0.4837	38	0.68 (0.96)	0.4820	6	6.12 (2.78)	0.0784	25	-0.17 (0.55)	0.7668
5	0	12	0.24 (0.71)	0.7411	22	1.14 (0.57)	0.0601	3	-0.64 (3.88)	0.8839	30	0.73 (0.50)	0.1561
6	0	19	0.79 (0.58)	0.1911	31	0.42 (0.42)	0.3274	6	1.31 (1.68)	0.4692	42	-0.17 (0.39)	0.6718
7	0	14	0.20 (1.94)	0.9214	21	0.02 (1.52)	0.9878	10	-0.67 (2.70)	0.8113	39	0.07 (0.49)	0.8847
8	0	12	0.37 (1.34)	0.7897	25	1.10 (0.78)	0.1699	6	3.82 (4.78)	0.4602	30	1.11 (0.51)	0.0375
9	0	18	-0.58 (0.59)	0.3411	29	-0.74 (0.51)	0.1587	6	-2.09 (1.99)	0.3418	45	-0.56 (0.32)	0.0870
$n=5,\ p=10,\ z_1=8$													
0	100	100	101.32 (11.74)	0.0000	100	112.46 (14.62)	0.0000	100	69.63 (23.13)	0.0033	100	100.21 (0.16)	0.0000
1	16	23	9.25 (1.68)	0.0000	63	6.30 (0.80)	0.0000	12	18.38 (5.10)	0.0041	100	15.30 (0.20)	0.0000
2	-10	18	-10.27 (2.58)	0.0010	46	-6.83 (1.35)	0.0000	7	-25.27 (3.29)	0.0003	100	-10.14 (0.18)	0.0000
3	15	15	11.42 (1.67)	0.0000	51	5.81 (1.03)	0.0000	10	30.55 (5.26)	0.0003	100	15.14 (0.19)	0.0000
4	18	15	21.98 (2.98)	0.0000	49	9.83 (1.49)	0.0000	14	38.02 (5.16)	0.0000	100	15.54 (0.17)	0.0000
5	-6	9	-8.53 (2.80)	0.0159	48	-3.88 (0.92)	0.0001	3	-14.31 (4.25)	0.0779	94	-5.54 (0.14)	0.0000
6	-7	15	-4.97 (3.46)	0.1728	42	-3.37 (1.14)	0.0050	7	-7.59 (15.37)	0.6391	99	-6.37 (0.18)	0.0000
7	-12	12	-13.51 (2.59)	0.0003	49	-4.90 (1.37)	0.0008	4	-18.69 (12.19)	0.2227	100	-10.45 (0.17)	0.0000
8	0	14	2.51 (3.44)	0.4776	47	0.48 (1.36)	0.7289	8	-0.06 (12.13)	0.9961	33	-0.43 (0.39)	0.2807
9	0	8	0.20 (2.80)	0.9459	42	-0.32 (0.70)	0.6492	4	8.24 (16.00)	0.6422	40	-0.40 (0.30)	0.1851

ตารางที่ 3 ค่าพารามิเตอร์ของตัวแบบ ร้อยละที่ตัวแปรอิสระอยู่ในสมการถดถอย ค่าเฉลี่ยของค่าประมาณพารามิเตอร์ของตัวแบบ ค่าความคลาดเคลื่อนมาตรฐาน และค่าพี ในกรณี $n=5, p=20, z_1=4$

j	β_j	ลาสโซ่			อีลาสติกเน็ต			การถดถอยแบบขั้นตอน			ขั้นตอนวิธีเชิงพันธุกรรม		
		ร้อยละ	$\bar{\hat{\beta}}_j$ (SE ($\bar{\hat{\beta}}_j$))	ค่าพี	ร้อยละ	$\bar{\hat{\beta}}_j$ (SE ($\bar{\hat{\beta}}_j$))	ค่าพี	ร้อยละ	$\bar{\hat{\beta}}_j$ (SE ($\bar{\hat{\beta}}_j$))	ค่าพี	ร้อยละ	$\bar{\hat{\beta}}_j$ (SE ($\bar{\hat{\beta}}_j$))	ค่าพี
$n=5, p=20, z_1=4$													
0	100	100	182.06 (11.67)	0.0000	100	178.13 (12.35)	0.0000	100	173.86 (21.74)	0.0000	100	99.93 (0.19)	0.0000
1	16	26	8.14 (1.21)	0.0000	50	5.51 (0.61)	0.0000	30	17.46 (1.27)	0.0000	100	15.47 (0.19)	0.0000
2	-10	10	-11.80 (3.37)	0.0067	34	-4.62 (1.17)	0.0004	17	-16.72 (3.61)	0.0003	100	-10.01 (0.17)	0.0000
3	15	8	12.75 (3.06)	0.0042	35	4.86 (1.02)	0.0000	16	16.47 (3.46)	0.0003	100	14.67 (0.17)	0.0000
4	0	5	1.56 (7.07)	0.8360	27	0.51 (1.36)	0.7091	7	6.09 (8.79)	0.5142	8	1.48 (0.55)	0.0319
5	0	3	-2.75 (7.32)	0.7430	28	0.89 (0.66)	0.1889	5	-3.00 (7.50)	0.7101	14	0.29 (0.62)	0.6447
6	0	5	3.00 (4.55)	0.5454	26	0.26 (0.58)	0.6598	1	37.94 (-)*	-	15	0.61 (0.59)	0.3168
7	0	5	-0.80 (4.21)	0.8593	27	-0.23 (0.98)	0.8116	3	-13.94 (14.59)	0.4401	11	0.78 (0.61)	0.2337
8	0	5	3.86 (2.54)	0.2032	31	0.98 (0.80)	0.2283	7	10.80 (5.27)	0.0860	6	-0.30 (1.45)	0.8421
9	0	9	1.68 (2.36)	0.4964	29	-0.15 (0.77)	0.8431	12	-2.24 (3.35)	0.5172	21	-0.19 (0.47)	0.6851
10	0	4	-2.54 (3.11)	0.4743	27	-0.83 (0.70)	0.2439	7	-5.76 (5.98)	0.3729	9	0.43 (0.66)	0.5302
11	0	4	1.02 (2.91)	0.7501	27	0.41 (0.55)	0.4626	3	7.37 (3.67)	0.1819	20	0.58 (0.42)	0.1874
12	0	4	5.83 (3.77)	0.2201	28	0.52 (0.68)	0.4470	1	19.83 (-)*	-	12	-0.52 (0.65)	0.4383
13	0	5	3.95 (3.04)	0.2641	29	-0.69 (1.01)	0.5031	4	-0.70 (5.92)	0.9137	12	-0.39 (0.48)	0.4349
14	0	2	0.36 (0.01)	0.0130	28	0.08 (0.42)	0.8534	1	0.65 (-)*	-	20	-0.15 (0.41)	0.7116
15	0	4	-3.70 (2.18)	0.1889	30	0.03 (0.63)	0.9638	7	-0.22 (4.51)	0.9630	17	0.53 (0.57)	0.3679
16	0	8	0.32 (3.21)	0.9228	31	0.57 (0.84)	0.5072	11	0.24 (5.13)	0.9642	12	0.74 (0.56)	0.2133
17	0	7	-0.50 (4.46)	0.9142	25	0.15 (1.08)	0.8926	9	-8.74 (4.51)	0.0885	18	0.43 (0.62)	0.5036
18	0	4	-4.86 (2.15)	0.1084	29	-0.62 (0.60)	0.3076	9	2.93 (6.39)	0.6588	17	0.17 (0.60)	0.7753
19	0	5	3.27 (1.59)	0.1095	29	0.91 (0.56)	0.1106	4	13.96 (5.45)	0.0833	15	0.15 (0.59)	0.8053

ตารางที่ 4 ค่าพารามิเตอร์ของตัวแบบ ร้อยละที่ตัวแปรอิสระอยู่ในสมการถดถอย ค่าเฉลี่ยของค่าประมาณพารามิเตอร์ของตัวแบบ ค่าความคลาดเคลื่อนมาตรฐาน และค่าพี ในกรณี $n=5, p=20, z_1=16$

j	β_j	ลาสโซ่			อีลาสติกเน็ต			การถดถอยแบบขั้นตอน			ขั้นตอนวิธีเชิงพันธุกรรม		
		ร้อยละ	$\bar{\hat{\beta}}_j$ (SE $(\bar{\hat{\beta}}_j)$)	ค่าพี	ร้อยละ	$\bar{\hat{\beta}}_j$ (SE $(\bar{\hat{\beta}}_j)$)	ค่าพี	ร้อยละ	$\bar{\hat{\beta}}_j$ (SE $(\bar{\hat{\beta}}_j)$)	ค่าพี	ร้อยละ	$\bar{\hat{\beta}}_j$ (SE $(\bar{\hat{\beta}}_j)$)	ค่าพี
$n=5, p=20, z_1=16$													
0	100	100	332.98 (24.22)	0.0000	100	307.51 (22.47)	0.0000	100	291.86 (48.74)	0.0000	100	100.21 (0.11)	0.0000
1	16	6	12.36 (6.39)	0.1107	32	4.94 (1.42)	0.0015	9	22.52 (6.67)	0.0097	100	15.29 (0.12)	0.0000
2	-10	8	-14.25 (5.76)	0.0426	34	-4.23 (1.41)	0.0052	5	-25.20 (16.18)	0.1943	100	-9.42 (0.13)	0.0000
3	15	6	9.26 (3.71)	0.0549	34	5.25 (1.26)	0.0002	5	17.73 (17.94)	0.3788	100	15.63 (0.13)	0.0000
4	18	7	22.69 (7.34)	0.0213	34	8.74 (2.19)	0.0003	11	26.92 (14.85)	0.0999	100	15.69 (0.12)	0.0000
5	-6	4	-3.42 (5.13)	0.5522	30	-3.04 (1.13)	0.0116	4	-5.61 (9.29)	0.5883	99	-4.90 (0.13)	0.0000
6	-7	8	-2.02 (3.24)	0.5535	32	-1.22 (1.26)	0.3426	7	1.88 (13.62)	0.8946	98	-5.04 (0.16)	0.0000
7	-12	5	-8.61 (19.78)	0.6858	30	-0.54 (3.57)	0.8807	6	-32.68 (36.43)	0.4109	100	-9.94 (0.12)	0.0000
8	20	7	16.52 (6.95)	0.0548	37	6.48 (1.38)	0.0000	10	24.89 (17.63)	0.1915	100	20.21 (0.12)	0.0000
9	-2	6	5.98 (7.78)	0.4767	27	0.03 (1.37)	0.9822	6	28.49 (20.18)	0.2171	23	-0.24 (0.40)	0.5564
10	-10	3	-19.30 (9.17)	0.1699	30	-4.57 (1.16)	0.0005	5	-19.86 (13.93)	0.2269	100	-9.43 (0.14)	0.0000
11	6	4	4.64 (9.24)	0.6499	31	2.10 (1.23)	0.0985	2	19.21 (16.29)	0.4479	100	5.97 (0.14)	0.0000
12	2	2	12.19 (10.26)	0.4453	30	-0.03 (0.84)	0.9685	5	22.75 (12.58)	0.1447	29	1.87 (0.25)	0.0000
13	19	9	13.26 (7.63)	0.1202	34	4.59 (1.93)	0.0236	12	17.40 (14.89)	0.2673	100	16.46 (0.11)	0.0000
14	4	7	-7.69 (9.03)	0.4272	32	0.45 (2.22)	0.8416	7	-18.05 (15.32)	0.2834	35	2.22 (0.14)	0.0000
15	-20	14	-12.04 (4.18)	0.0128	43	-7.08 (1.56)	0.0000	15	-17.94 (8.75)	0.0595	100	-19.90 (0.15)	0.0000
16	0	6	-3.50 (5.81)	0.5737	33	0.32 (1.72)	0.8536	6	-4.93 (17.87)	0.7939	19	1.32 (0.36)	0.0019
17	0	3	-7.89 (9.81)	0.5054	31	0.95 (0.78)	0.2329	5	-6.59 (8.61)	0.4865	29	1.20 (0.29)	0.0002
18	0	3	-8.31 (6.93)	0.3534	28	0.77 (0.81)	0.3496	2	-14.40 (11.45)	0.4275	23	0.99 (0.40)	0.0221
19	0	1	-2.15 (-)*	-	28	-0.35 (0.97)	0.7210	5	-6.16 (11.95)	0.6333	18	1.37 (0.33)	0.0006

ในตารางที่ 2-4 ค่าเฉลี่ยของค่าประมาณพารามิเตอร์ของตัวแบบ $\bar{\hat{\beta}}_j$ ค่าความคลาดเคลื่อนมาตรฐาน $\mathrm{SE}(\bar{\hat{\beta}}_j)$ และค่าพี คำนวณเฉพาะครั้งที่ตัวแปรอิสระ x_j อยู่ในสมการถดถอย ซึ่งมีจำนวนครั้งเท่ากับค่าร้อยละในตาราง และ * หมายถึงไม่สามารถหา $\mathrm{SE}(\bar{\hat{\beta}}_j)$ ได้เนื่องจากมีจำนวนครั้งที่ตัวแปรอิสระ x_j อยู่ในสมการถดถอยเพียง 1 ครั้ง

ตารางที่ 5 จำนวนตัวแปรอิสระเกี่ยวข้องที่ตัดออกจากสมการ ในกรณี $n = 20$

กรณี	จำนวนตัวแปรอิสระเกี่ยวข้องที่ตัดออกจากสมการ (จำนวนสมการที่มีการตัดตัวแปรอิสระเกี่ยวข้อง)							
	การคัดเลือกตัวแปรอิสระน้อยเกินไป				การคัดเลือกตัวแปรอิสระไม่ถูกต้อง			
	ลาสโซ่	อีลาสติกเน็ต	การถดถอยแบบขั้นตอน	ขั้นตอนวิธีเชิงพันธุกรรม	ลาสโซ่	อีลาสติกเน็ต	การถดถอยแบบขั้นตอน	ขั้นตอนวิธีเชิงพันธุกรรม
$n=20, p=40, z_1=8$	3-7(5)	3-7(3)	2-7(11)	-	1-5(61)	1-4(50)	1-6(68)	-
$n=20, p=40, z_1=32$	19-31(50)	18-31(11)	22-30(46)	1-3(2)	13-30(50)	1-30(44)	18-31(54)	1-4(84)
$n=20, p=80, z_1=16$	12-15(28)	13-15(10)	-	-	5-15(72)	1-15(68)	7-15(100)	1-3(90)
$n=20, p=80, z_1=64$	48-63(62)	48-63(25)	47-63(19)	5-7(7)	46-63(38)	2-61(50)	47-62(81)	2-7(93)

หมายเหตุ : ค่าในวงเล็บคือจำนวนสมการ

เมื่อ n เพิ่มเป็น 20 โดยอัตราส่วน $n:p$ และ $p:z_1$ คงเดิม พบว่าเกือบทุกกรณีที่ศึกษาไม่มีการคัดเลือกตัวแปรอิสระได้ถูกต้อง และส่วนใหญ่มีการคัดเลือกตัวแปรอิสระมากเกินไปหรือการคัดเลือกตัวแปรอิสระน้อยเกินไปลดลงโดยมีการคัดเลือกตัวแปรอิสระไม่ถูกต้องเพิ่มขึ้นแทน โดยที่ขั้นตอนวิธีเชิงพันธุกรรมมีการคัดเลือกตัวแปรอิสระไม่ถูกต้องเพิ่มขึ้นอย่างมาก แต่ยังคงมีการคัดเลือกตัวแปรอิสระน้อยเกินไปน้อยกว่าวิธีอื่น ๆ ยกเว้นในกรณี $n=20, p=40, z_1=8$ ที่ขั้นตอนวิธีเชิงพันธุกรรมมีการคัดเลือกตัวแปรอิสระมากเกินไปร้อยละ 100 และวิธีการถดถอยแบบขั้นตอนมีการคัดเลือกตัวแปรอิสระได้ถูกต้องร้อยละ 4 แต่มีการคัดเลือกตัวแปรอิสระน้อยเกินไปเพิ่มขึ้นเป็นร้อยละ 11 และมีการคัดเลือกตัวแปรอิสระไม่ถูกต้องเพิ่มขึ้นถึงร้อยละ 68 และเมื่อพิจารณารายละเอียดในตารางที่ 5 พบว่าสมการถดถอยที่ได้จากขั้นตอนวิธีเชิงพันธุกรรมมีการตัดตัวแปรอิสระเกี่ยวข้องออกไปจำนวนน้อยกว่าวิธีอื่น ๆ อย่างชัดเจน นอกจากนี้ค่าประมาณ $\beta_0, \beta_1, \beta_2, ..., \beta_{z_1-1}$ โดยเฉลี่ยยังคงใกล้เคียงกับค่าจริงที่จำลองมากที่สุดและมีความคลาดเคลื่อนมาตรฐานรวมทั้ง $\mathrm{MSE}(\hat{y})$ ต่ำที่สุด ดังนั้นจึงสามารถสรุปได้ว่าในกรณีที่ $n = 20$ ทุกกรณี ขั้นตอนวิธีเชิงพันธุกรรมมีการประมาณค่าพารามิเตอร์และคัดเลือกตัวแปรอิสระได้ดีที่สุดเมื่อเทียบกับวิธีอื่น ๆ

สรุปผลการวิจัย

การประมาณค่าพารามิเตอร์ของตัวแบบโดยใช้วิธีกำลังสองน้อยที่สุดให้ตัวประมาณค่าไม่เอนเอียงเชิงเส้นที่ดีที่สุด (Best Linear Unbiased Estimator: BLUE) แต่มีปัญหาในการวิเคราะห์การถดถอยเมื่อข้อมูลมีมิติสูง ทำให้วิธีการถดถอยแบบขั้นตอนมีข้อจำกัดในการคัดเลือกตัวแปรอิสระ วิธีการถดถอยที่ปรับด้วยฟังก์ชันการลงโทษซึ่งให้ค่าประมาณพารามิเตอร์ของตัวแบบที่เอนเอียงเช่น วิธีลาสโซ่และวิธีอีลาสติกเน็ต สามารถแก้ปัญหาดังกล่าวได้ในระดับหนึ่ง แต่อาจไม่มีความเหมาะสมเมื่อนำไปใช้กับข้อมูลบางกรณี โดยเฉพาะเมื่อในตัวแบบจริงมีตัวแปรอิสระเกี่ยวข้องจำนวนมากกว่าขนาดตัวอย่าง ขณะที่การประมาณค่าพารามิเตอร์โดยใช้ขั้นตอนวิธีเชิงพันธุกรรมไม่มีการหาเมทริกซ์ผกผันเหมือนกับในวิธีกำลังสองน้อยที่สุด ดังนั้นในกรณีที่ข้อมูลมีมิติสูง ขั้นตอนวิธีเชิงพันธุกรรมจึงไม่มีปัญหาการประมาณค่าพารามิเตอร์รวมทั้งไม่มีข้อจำกัดในการคัดเลือกตัวแปรอิสระ และจากผลการศึกษาสรุปได้ว่าเมื่อเทียบกับวิธีลาสโซ่ วิธีอีลาสติกเน็ต และวิธีการถดถอยแบบขั้นตอน ขั้นตอนวิธีเชิงพันธุกรรมสามารถคัดเลือกตัวแปรอิสระได้ดีที่สุดและสามารถประมาณค่าพารามิเตอร์ได้ใกล้เคียงกับค่าจริงมากที่สุดในเกือบทุกกรณีที่จำลอง นอกจากนี้ เมื่อมีจำนวนตัวแปรอิสระที่ต้องพิจารณาเพิ่มขึ้น การคัดเลือกตัวแปรอิสระจะมีความยุ่งยาก

มากขึ้น ทำให้ทุกวิธีที่ศึกษามีโอกาสสูงที่จะคัดเลือกตัวแปรอิสระได้แย่ลง แต่ขั้นตอนวิธีเชิงพันธุกรรมก็ได้รับผลกระทบน้อยที่สุด อย่างไรก็ดี เมื่อข้อมูลมีมิติสูงและในตัวแบบจริงมีจำนวนตัวแปรอิสระเกี่ยวข้องน้อยกว่าขนาดตัวอย่าง หรือในตัวแบบจริงมีตัวแปรอิสระเกี่ยวข้องจำนวนไม่มาก ขั้นตอนวิธีเชิงพันธุกรรมมีโอกาสที่จะคัดเลือกตัวแปรอิสระได้มากเกินไปสูงกว่าวิธีอื่น ๆ และเมื่อมีจำนวนตัวแปรอิสระที่ต้องพิจารณาเพิ่มขึ้นหรือมีการกำหนดปริภูมิการค้นหากว้างเกินไป ขั้นตอนวิธีเชิงพันธุกรรมยังมีข้อเสียคือใช้เวลามากในการค้นหาคำตอบ เนื่องจากมีจำนวนรอบในการค้นหาเพิ่มขึ้นอย่างทวีคูณ

กิตติกรรมประกาศ

ผู้เขียนขอขอบคุณผู้ทรงคุณวุฒิสำหรับข้อเสนอแนะที่สร้างสรรค์ เป็นประโยชน์ต่อการปรับปรุงการนำเสนอของบทความนี้ให้ดีขึ้น

เอกสารอ้างอิง

Candès, E. & Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, *35*(6), 2313–2351.

Darwin, C. (1859). *The origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: John Murray.

De Jong, K. A. (1975). *An analysis of the behavior of a class of genetic adaptive systems*. Doctoral dissertation, University of Michigan.

Drezner, Z., & George, A. (1999). Tabu search model selection in multiple regression analysis. *Communications in Statistics – Simulation and Computation*, *28*(2), 349–367.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression (with discussion). *The Annals of Statistics*, 32(2), 407-499.

Glover, F. (1986). Future paths for integer programming and links to artificial intelligence. *Computers and Operations Research*, *13*(5), 533–549.

Glover, F. (1989). Tabu search – part 1. *ORSA Journal on Computing*, *1*(2), 190–206.

Glover, F. (1990). Tabu search – part 2. *ORSA Journal on Computing*, *2*(1), 4–32.

Glover, F. (1990). Tabu search: A tutorial. *Interfaces*, *20*(1), 74-94.

Goldberg, D.E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Massachusetts: Addison-Wesley Publishing.

Gujarati, D. N. (2006). *Essentials of Econometrics (3rd ed.)*. New York: McGraw-Hill.

Holland, J. H. (1973). Genetic algorithms and the optimal allocation of trials. *SIAM Journal on Computing*, *2*(2), 88–105.

Holland, J. H. (1975). *Adaptation in natural and artificial systems*. Ann Arbor: University of Michigan Press.

Ishwaran, H., & Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, *33*(2), 730–773.

Jitthavech, J. (2015). *Regression analysis*. Bangkok: National Institute of Development Administration. (in Thai)

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671-680.

Miller, B. L. & Goldberg, D. E. (1995). Genetic algorithms, tournament selection, and the effects of noise. *Complex Systems*, *9*(3), 193–212.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2006). *Introduction to linear regression analysis (4th ed.)*. New Jersey: John Willey & Sons.

Na Bangchang, K. (2011). *A variable selection in multiple linear regression models based on tabu search*. Master's thesis, National Institute of Development Administration. (in Thai)

Ngamprasertsit, N. (2012). *A comparison of variable selection by ridge regression and tabu search with multicollinearity*. Master's thesis, National Institute of Development Administration. (in Thai)

Pungpapong, V. (2012). *Empirical Bayes variable selection for high-dimensional regression*. Doctoral dissertation, Purdue University.

Pungpapong, V. (2015). A brief review on high-dimensional linear regression. *Thammasat Journal of Science and Technology*, *23*(2), 212–223. (in Thai)

Quenouille, M. H. (1949). Problems in plane sampling. *The Annals of Mathematical Statistics*, *20*(3), 355–375.

Quenouille, M. H. (1956). Notes on bias in estimation. *Biometrika*, *43*(3–4), 353–360.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B, 58*(1), 267–288.

Tukey, J. W. (1958). Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics*, *29*(2), 614–623.

Whitley, D., Mathias, K., & Fitzhorn, P. (1991). Delta coding: An iterative search strategy for genetic algorithms. In *Proceeding of the 4th International Conference on Genetic Algorithms*. (pp. 77–84). CA: Morgan Kaufmann.

Wright, A. H. (1991). Genetic algorithms for real parameter optimization. *Foundations of Genetic Algorithms*, *1*, 205–218.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B, 67*(2), 301–320.