

การเปรียบเทียบวิธีคัดเลือกตัวแปรอิสระในแบบการถดถอยเชิงเส้นพหุโดยใช้ ขั้นตอนวิธีเชิงพันธุกรรมและการจำลองการอบเหนียว

A Comparison of Subset Selection in Multiple Linear Regression Model using Genetic Algorithm and Simulated Annealing

พิรภัทร์ ภาคภูมิภมมิลเลิศ¹, จรัสศรี รุ่งรัตนอุบล², อนามัย นาอุดม^{1*}

Peraphat Phakpoomkamonloest¹, Jaratsri Rungrattanaubol², Anamai Na-udom^{1*}

¹ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร

¹Department of Mathematics, Faculty of Science, Naresuan University

²ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยนเรศวร

²Department of Computer Science and Information Technology, Faculty of Science,
Naresuan University

*E-mail: anamain@nu.ac.th

บทคัดย่อ

การวิจัยนี้ มีวัตถุประสงค์เพื่อศึกษาการคัดเลือกตัวแปรอิสระเข้าสู่แบบการถดถอยเชิงเส้นพหุด้วยวิธีการค้นหาแบบขั้นตอนเชิงพันธุกรรม (Genetic Algorithm: GA) และการค้นหาแบบจำลองการอบเหนียว (Simulated Annealing: SA) และทำการเปรียบเทียบประสิทธิภาพและความรวดเร็วในการทำงานของทั้ง GA และ SA ในสถานการณ์ที่ไม่เกิดและเกิดปัญหาพหุสัมพันธ์เชิงเส้นต่อกันของตัวแปรอิสระ ทำการศึกษาโดยการจำลองข้อมูลและกำหนดจำนวนตัวแปรอิสระเท่ากับ 5 10 และ 20 ตามลำดับ และแต่ละขนาดของตัวแปรอิสระ กำหนดขนาดตัวอย่างเท่ากับ 50 100 200 และ 500 ตามลำดับ ทำการเปรียบเทียบประสิทธิภาพของ GA และ SA โดยพิจารณาค่าร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบและพิจารณาจำนวนรอบในการทำงานเป็นเกณฑ์ในการเปรียบเทียบความรวดเร็วในการทำงาน ผลการศึกษาพบว่า GA และ SA สามารถทำงานได้ดีทั้งในสถานการณ์ที่ไม่เกิดและเกิดปัญหาพหุสัมพันธ์เชิงเส้นต่อกันของตัวแปรอิสระ โดยที่ GA ให้ค่าร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบสูงกว่า SA ในกรณีที่ตัวแปรอิสระมีจำนวนน้อย และทั้ง 2 วิธีสามารถคัดเลือกตัวแปรอิสระได้ถูกต้องร้อยละ 100 เมื่อจำนวนตัวแปรอิสระมีค่าเพิ่มขึ้น นอกจากนี้ยังพบว่าทั้ง GA และ SA มีความรวดเร็วในการทำงานใกล้เคียงกันเมื่อตัวแปรอิสระมีจำนวนน้อย ในขณะที่ SA สามารถทำงานได้รวดเร็วมากกว่า GA เมื่อตัวแปรอิสระมีจำนวนมากขึ้น

คำสำคัญ: การค้นหาแบบขั้นตอนเชิงพันธุกรรม การค้นหาแบบจำลองการอบเหนียว

การคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ พหุสัมพันธ์เชิงเส้น

Abstract

The objective of this research is to study the subset selection in multiple linear regression model by using Genetic Algorithm (GA) and Simulated Annealing (SA) and compare the efficiency and rate of convergence of GA and SA in the case of with and without multicollinearity problem between input variables. The data used in this study were simulated under various situations. The numbers of input variables are 5 10 and 20 respectively, for each input variable, the sample sizes are 50 100 200 and 500 respectively. The efficiency is compared by using correct specification percentage of subset selection while rate of convergence is validated through number of loops spent for each case. The results show that GA and SA perform similarly in the case of with and without multicollinearity problem. Further, GA provide higher percentage of correct specification than SA for small number of input variable while both methods provide 100% of correct specification when the numbers of input variables are increased. In addition, the results reveal that both GA and SA have similar rate of convergence when the number of input variables is small while SA can work much faster than GA when the number of input variables or the dimension of the problem becomes larger.

Keywords: Genetic Algorithm, Simulated Annealing, Subset Selection, Multicollinearity

บทนำ

การถดถอยเชิงเส้นพหุ (Multiple Linear Regression) เป็นวิธีการทางสถิติที่ได้รับความนิยมอย่างแพร่หลาย ในการศึกษารูปแบบความสัมพันธ์ระหว่างตัวแปร โดยวัตถุประสงค์หลัก คือ การประมาณค่าตัวแปรตามหรือตัวแปรตอบสนอง (Dependent Variable or Response Variable) โดยการสร้างตัวแบบที่แสดงความสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระ (Independent Variable) ตั้งแต่ 2 ตัวขึ้นไปกับตัวแปรตาม ซึ่งในกรณีที่ตัวแปรอิสระที่เกี่ยวข้องจำนวนมากนั้น อาจก่อให้เกิดปัญหาที่ตัวแปรอิสระบางตัวมีความสัมพันธ์กันเอง หรือเรียกว่าเกิดปัญหาพหุสัมพันธ์เชิงเส้น (Multicollinearity) ส่งผลให้การประมาณค่าสัมประสิทธิ์ถดถอยมีความคลาดเคลื่อนและตัวประมาณที่ได้จะไม่มีความสมบัติตัวประมาณเชิงเส้นไม่เอนเอียงดีที่สุด (Best Linear Unbiased Estimator: BLUE)

อย่างไรก็ตามในปัจจุบันมีวิธีแก้ปัญหการประมาณค่าสัมประสิทธิ์ถดถอยเมื่อตัวแปรอิสระมีพหุสัมพันธ์เชิงเส้นต่อกันหลายวิธี เช่น วิธีการประมาณค่าสัมประสิทธิ์ถดถอยด้วยวิธีวิเคราะห์การถดถอยแบบบริดจ์ (Ridge Regression) วิธีวิเคราะห์การถดถอยแบบเลซโซ (Lasso Regression) และวิธีวิเคราะห์การถดถอยแบบยืดหยุ่น (Elastic net Regression) วิธีเหล่านี้จะช่วยลดปัญหาการประมาณค่าสัมประสิทธิ์ถดถอยของตัวแบบในกรณีที่ตัวแปรอิสระมีพหุสัมพันธ์เชิงเส้นต่อกันลงได้ นอกจากนี้การที่มีตัวแปรอิสระจำนวนมากที่อยู่ในตัวแบบ อาจก่อให้เกิดปัญหาตัวแบบที่มีความไม่เหมาะสม เนื่องจากอาจมีตัวแปรอิสระบางตัวที่ไม่มีความสัมพันธ์เชิงเส้นกับตัวแปรตาม ดังนั้นจึงต้องมีการค้นหาชุดของตัวแปรอิสระที่เหมาะสมที่สุดที่มีความสัมพันธ์เชิงเส้นกับตัวแปรตามอย่างแท้จริง ซึ่งหากมีการคัดเลือกตัวแปรอิสระที่มี

ความสัมพันธ์เชิงเส้นกับตัวแปรตามสูงออกจากตัวแบบจะทำให้ตัวแบบที่จะนำไปประมาณค่าตัวแปรตามนั้นมีประสิทธิภาพต่ำ ส่งผลให้ค่าประมาณที่ได้มีความคลาดเคลื่อนสูง

ปัจจุบันมีวิธีการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบอยู่หลายวิธี เช่น วิธีตัดตัวแปร (Backward Elimination : BE) วิธีเพิ่มตัวแปร (Forward Selection : FS) และวิธีเพิ่มตัวแปรอิสระแบบขั้นตอน (Stepwise Regression : SR) อย่างไรก็ตามเมื่อมีจำนวนตัวแปรอิสระหรือจำนวนชุดข้อมูลมีค่ามากขึ้นจะส่งผลให้มีจำนวนตัวแบบที่เป็นไปได้มากขึ้นตามไปด้วย ซึ่งวิธีการดังกล่าวอาจไม่สามารถค้นหาตัวแบบที่ถูกต้องได้ ดังนั้นจึงมีการคิดค้นวิธีการคัดเลือกตัวแปรอิสระที่ประยุกต์มาจากวิธีการที่ใช้ในการแก้ปัญหาการหาค่าเหมาะที่สุดเชิงการจัด (Combinatorial Optimization Problem) หรือเมตาฮิวริสติก (Metaheuristic) ที่สามารถค้นหาตัวแบบที่ดีที่สุดได้โดยไม่ต้องพิจารณาตัวแบบที่เป็นไปได้ทั้งหมด และสามารถค้นหาคำตอบที่เหมาะสมที่สุดได้อย่างมีประสิทธิภาพในกรณีที่มีมิติของปัญหาที่สนใจศึกษามีขนาดใหญ่ วิธีการที่จัดอยู่ในกลุ่มของเมตาฮิวริสติก ที่ได้รับความนิยมในปัจจุบันมีอยู่หลายวิธี เช่น วิธีการค้นหาแบบจำลองการอบเหนียว (Simulated Annealing : SA) วิธีการค้นหาแบบต้องห้าม (Tabu Search : TS) และวิธีการค้นหาแบบขั้นตอนวิธีเชิงพันธุกรรม (Genetic Algorithm : GA) วิธีการเหล่านี้เป็นกระบวนการแก้ปัญหาเพื่อหาค่าหรือตัวแปรอิสระที่เหมาะสมที่สุดเข้าสู่ตัวแบบได้อย่างแม่นยำและมีความสามารถในการค้นหาตัวแบบที่เหมาะสมได้อย่างรวดเร็ว โดยเฉพาะในกรณีที่มีตัวแปรอิสระอยู่ในตัวแบบเป็นจำนวนมาก อย่างไรก็ตามยังไม่มีที่แน่ชัดว่าวิธีใดเป็นวิธีในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ดีที่สุด เนื่องจากแต่ละวิธีจะมีความเหมาะสมในการทำงานในแต่ละสถานการณ์ที่แตกต่างกันไป

ในปัจจุบันมีงานวิจัยที่ศึกษาเกี่ยวกับวิธีการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบด้วยวิธีการต่าง ๆ ภายใต้สถานการณ์ที่แตกต่างกันอยู่มากมาย ดังงานวิจัยของ พิศลย์ มีแก้ว (2546) ที่ทำการเปรียบเทียบประสิทธิภาพในการหาผลเฉลยที่เหมาะสมระหว่าง GA และ SA ในงานด้านกระบวนการพื้นผิวตอบสนอง (Response surface methodology) โดยพิจารณาสถานการณ์ที่มีและไม่มีปัจจัยรบกวนตามธรรมชาติ (Noise) ผลการวิจัยพบว่า GA สามารถหาผลเฉลยได้ดีกว่าในสถานการณ์ที่มีปัจจัยรบกวน ในขณะที่ SA สามารถทำงานได้ดีกว่าในสถานการณ์ที่ไม่มีปัจจัยรบกวน ต่อมา ศิรินทิพย์ หมื่นจันทร์ และ วฐา มินเสน (2557) ทำการประยุกต์ขั้นตอนการทำงานขอ GA โดยใช้ชื่อว่า ดับเบิลเจเนติกอัลกอริทึม (Double Genetic Algorithms: DGA) มาใช้ในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบและเปรียบเทียบประสิทธิภาพกับ GA และ SR โดยศึกษาในกรณีที่ไม่เกิดและเกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ ผลการวิจัยพบว่าในสถานการณ์ที่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ DGA ให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Squares Error: MSE) น้อยที่สุด ส่วน GA และ SR ให้ค่าที่ใกล้เคียงกัน สำหรับกรณีที่ไม่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุทั้ง 3 วิธีให้ค่า MSE ที่ต่ำที่สุดเท่ากัน และในงานวิจัยของ Kapetanious (2007) ได้ทำการพัฒนาตัวแบบถดถอยบนพื้นฐานของปัญหาทางเศรษฐมิติ (Econometrics) โดยใช้ GA SA และวิธีการเฉลี่ยตัวแบบของเบย์ (Bayesian Model Averaging Method) พร้อมทั้งเปรียบเทียบประสิทธิภาพในการพยากรณ์ตัวแบบที่ได้จากทั้ง 3 วิธี ผลการศึกษาพบว่า GA และ SA มีประสิทธิภาพในการพยากรณ์ตัวแปรตามได้ดีกว่าวิธีการเฉลี่ยตัวแบบของเบย์ โดยเฉพาะอย่างยิ่งในกรณีที่มีจำนวนตัวแปรอิสระเป็นจำนวนมากเป็นต้น

จากที่กล่าวมาข้างต้น ผู้วิจัยจึงสนใจศึกษาเปรียบเทียบวิธีการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบด้วยวิธีเมตาฮิวริสติก 2 วิธี คือ GA และ SA โดยพิจารณาตัวแบบการถดถอยเชิงเส้นพหุที่มีจำนวนตัวแปรอิสระต่าง ๆ รวมไปถึงขนาดตัวอย่างที่แตกต่างกันหลายระดับ ซึ่งในการศึกษาในครั้งนี้จะแบ่งออกเป็น 2 กรณี คือ กรณีที่ไม่เกิดปัญหาหุ้มพันซ์เชิงเส้นต่อกันของตัวแปรอิสระ จะทำการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบด้วย GA และ SA โดยจะทำการประมาณค่าสัมประสิทธิ์ถดถอยด้วยวิธีกำลังสองน้อยที่สุด และ กรณีที่เกิดปัญหาหุ้มพันซ์เชิงเส้นต่อกันของตัวแปรอิสระ จะทำการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบด้วย GA และ SA โดยจะทำการประมาณค่าสัมประสิทธิ์ถดถอยด้วยวิธีวิเคราะห์การถดถอยแบบบริดจ์ และกำหนดให้เกณฑ์ข้อสนเทศของอาไคเคะ (Akaike's Information Criterion : AIC) เป็นฟังก์ชันวัตถุประสงค์ของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบของทั้ง 2 วิธี โดยกำหนดให้ร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบเป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพของทั้ง 2 วิธี และกำหนดให้จำนวนรอบในการทำงานเป็นเกณฑ์ในการเปรียบเทียบความรวดเร็วของทั้ง 2 วิธี

วัตถุประสงค์การวิจัย

เพื่อศึกษาวิธีการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบการถดถอยเชิงเส้นพหุด้วยวิธี GA และ SA เพื่อทำการเปรียบเทียบประสิทธิภาพโดยพิจารณาจากร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ และเปรียบเทียบความรวดเร็วในการทำงานของทั้ง 2 วิธีโดยพิจารณาจากจำนวนรอบในการทำงานในแต่ละมิติของปัญหาที่สนใจศึกษา

วิธีดำเนินการวิจัย

การวิจัยครั้งนี้ทำการเปรียบเทียบประสิทธิภาพและความรวดเร็วในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบถดถอยเชิงเส้นพหุด้วย GA และ SA ในกรณีที่ไม่มีปัญหาหุ้มพันซ์เชิงเส้น และเกิดปัญหาหุ้มพันซ์เชิงเส้นระดับสูง เมื่อมีการทำซ้ำ 500 รอบ โดยมีขั้นตอนดำเนินการวิจัยดังนี้

1. การจำลองข้อมูลเมื่อไม่เกิดและเกิดปัญหาหุ้มพันซ์เชิงเส้นต่อกันของตัวแปรอิสระ

1.1 กำหนดให้ตัวแปรอิสระมีการแจกแจงปกติหลายตัวแปร โดยมีรูปแบบทั่วไปคือ

$$\mathbf{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

โดยที่ $\mathbf{X}$ คือ เมทริกซ์ของตัวแปรอิสระขนาด $n \times k$ เมื่อ n คือ ขนาดตัวอย่าง และ k คือ จำนวนตัวแปรอิสระ

$\boldsymbol{\mu}$ คือ เวกเตอร์ค่าเฉลี่ยของตัวแปรอิสระ k ตัวแปร

$\boldsymbol{\Sigma}$ คือ เมทริกซ์ความแปรปรวนร่วม (Covariance matrix : Cov) ขนาด $k \times k$ ดังนี้

$$\boldsymbol{\Sigma} = [\text{Cov}[\mathbf{X}_{is}, \mathbf{X}_{it}]] \text{ เมื่อ } i = 1, 2, \dots, n \quad s = 1, 2, \dots, k$$

และ $t = 1, 2, \dots, k$

โดยกำหนดค่าเฉลี่ยและความแปรปรวนของตัวแปรอิสระแต่ละตัวดังนี้

1) กรณีที่จำนวนตัวแปรอิสระเท่ากับ 5 ตัวแปร กำหนดให้

$$X_{i1} \sim N(107,1), X_{i2} \sim N(77,1), X_{i3} \sim N(300,1), X_{i4} \sim N(88,1), X_{i5} \sim N(207,1)$$

2) กรณีที่จำนวนตัวแปรอิสระเท่ากับ 10 ตัวแปร กำหนดให้

$$X_{i1} \sim N(107,1), X_{i2} \sim N(77,1), X_{i3} \sim N(300,1), X_{i4} \sim N(88,1), X_{i5} \sim N(207,1), \\ X_{i6} \sim N(139,1), X_{i7} \sim N(620,1), X_{i8} \sim N(540,1), X_{i9} \sim N(715,1), X_{i10} \sim N(870,1)$$

3) กรณีที่จำนวนตัวแปรอิสระเท่ากับ 20 ตัวแปร กำหนดให้

$$X_{i1} \sim N(107,1), X_{i2} \sim N(77,1), X_{i3} \sim N(300,1), X_{i4} \sim N(88,1), X_{i5} \sim N(207,1), \\ X_{i6} \sim N(139,1), X_{i7} \sim N(620,1), X_{i8} \sim N(540,1), X_{i9} \sim N(715,1), X_{i10} \sim N(870,1), \\ X_{i11} \sim N(900,1), X_{i12} \sim N(845,1), X_{i13} \sim N(1080,1), X_{i14} \sim N(1400,1), \\ X_{i15} \sim N(2100,1), X_{i16} \sim N(3400,1), X_{i17} \sim N(4050,1), X_{i18} \sim N(3950,1), \\ X_{i19} \sim N(7500,1), X_{i20} \sim N(9840,1) \text{ เมื่อ } i = 1, 2, \dots, n$$

กรณีที่ก่อให้เกิดปัญหาพหุสัมพันธ์เชิงเส้นต่อกันของตัวแปรอิสระจะกำหนดให้ $Cov(X_{is}, X_{it}) = 0$ เมื่อ $s \neq t$

กรณีที่ก่อให้เกิดปัญหาพหุสัมพันธ์เชิงเส้นต่อกันของตัวแปรอิสระกำหนดให้ X_2 และ X_4 มีความสัมพันธ์กันในระดับสูง ($\rho = 0.90$) โดยจะกำหนดให้ $Cov(X_{is}, X_{it}) = 0$ เมื่อ $s \neq t$ และกำหนด $Cov(X_{i2}, X_{i4})$ เท่ากับ 0.90

1.2 กำหนดค่าสัมประสิทธิ์ถดถอยเมื่อจำนวนตัวแปรอิสระเท่ากับ 5 10 และ 20 ตามลำดับ ดังนี้

1) กรณีที่จำนวนตัวแปรอิสระเท่ากับ 5 ตัวแปร กำหนดให้

$$\beta_0 = 100, \beta_1 = -155, \beta_2 = 15, \beta_3 = 11.5, \beta_4 = 509, \beta_5 = 0$$

2) กรณีที่จำนวนตัวแปรอิสระเท่ากับ 10 ตัวแปร กำหนดให้

$$\beta_0 = 100, \beta_1 = -155, \beta_2 = 15, \beta_3 = 11.5, \beta_4 = 509, \beta_5 = 0, \beta_6 = 0, \beta_7 = 4, \beta_8 = 7, \beta_9 = 0, \\ \beta_{10} = 0$$

3) กรณีที่จำนวนตัวแปรอิสระเท่ากับ 20 ตัวแปร กำหนดให้

$$\beta_0 = 100, \beta_1 = -155, \beta_2 = 15, \beta_3 = 11.5, \beta_4 = 509, \beta_5 = 0, \beta_6 = 0, \beta_7 = 4, \beta_8 = 7, \beta_9 = 0, \\ \beta_{10} = 0, \beta_{11} = -1550, \beta_{12} = 900, \beta_{13} = 1000, \beta_{14} = 50, \beta_{15} = 0, \beta_{16} = 0, \beta_{17} = 40, \beta_{18} = 100, \\ \beta_{19} = 0, \beta_{20} = 0$$

1.3 สุ่มขนาดตัวอย่างเท่ากับ 50 100 200 และ 500 ตามลำดับ

2. ตัวแบบการถดถอยเชิงเส้นพหุ

การวิเคราะห์การถดถอยเชิงเส้นพหุ คือ การศึกษาอิทธิพลของตัวแปรอิสระหลายตัวแปรที่มีอิทธิพลต่อตัวแปรตาม โดยสามารถแสดงรูปแบบความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตามได้จากตัวแบบการถดถอยเชิงเส้นพหุ ซึ่งสามารถนำรูปแบบความสัมพันธ์ที่สร้างได้ไปพยากรณ์ค่าตัวแปรตามได้ รูปแบบทั่วไปของสมการถดถอยเชิงเส้นพหุมีดังต่อไปนี้

$$y = X\beta + \varepsilon$$

โดยที่

y คือ เวกเตอร์ของตัวแปรตามขนาด $n \times 1$ เมื่อ n คือ ขนาดตัวอย่าง

X คือ เมทริกซ์ของตัวแปรอิสระขนาด $n \times (k+1)$ เมื่อ k คือ จำนวนตัวแปรอิสระ

β คือ เวกเตอร์ของสัมประสิทธิ์ถดถอยของตัวแบบขนาด $(k+1) \times 1$

ε คือ เวกเตอร์ของความคลาดเคลื่อนสุ่มขนาด $n \times 1$ และกำหนดให้ $\varepsilon_i \sim N(0, 5^2)$

เมื่อ $i = 1, 2, \dots, n$

3. วิธีการประมาณค่าพารามิเตอร์

3.1 วิธีกำลังสองน้อยที่สุด

คุณสมบัติของตัวประมาณค่า $\hat{\beta}$ โดยวิธีกำลังสองน้อยที่สุด คือ เป็นตัวประมาณค่าของ β ที่ทำให้ค่าความคลาดเคลื่อนกำลังสองต่ำที่สุด และ $\hat{\beta}$ เป็นตัวประมาณที่ไม่เอนเอียงและมีความแปรปรวนต่ำสุดของพารามิเตอร์ β และเป็นฟังก์ชันเชิงเส้นของ y ทำให้ตัวประมาณค่า $\hat{\beta}$ มีคุณสมบัติเป็นตัวประมาณเชิงเส้นที่ไม่เอนเอียงที่ดีที่สุด (Best Linear Unbiased Estimator: BLUE) ของ β การประมาณค่า β ด้วยวิธีกำลังสองน้อยที่สุดทำได้โดย

$$\hat{\beta} = (X'X)^{-1} X'y$$

3.2 วิธีการถดถอยแบบบริดจ์

ในการประมาณค่า β ด้วยวิธีกำลังสองน้อยที่สุดมีข้อจำกัดคือ ตัวแปรอิสระต้องไม่มีความสัมพันธ์กัน ซึ่งในทางปฏิบัติจะเป็นไปได้น้อยมาก ทำให้การประมาณค่า β ด้วยวิธีนี้มีความแปรปรวนมาก เพื่อแก้ปัญหานี้ โฮเอิร์ลและเคนนาร์ด (1970) ได้เสนอวิธีประมาณค่าสัมประสิทธิ์ถดถอยใหม่โดยการเพิ่มค่าที่เป็นบวกเล็ก ๆ r บนสมาชิกในแนวทแยง (Diagonal Elements) ของเมทริกซ์ $X'X$ ดังนั้นจะได้ตัวประมาณค่าสัมประสิทธิ์ถดถอยเป็น

$$\hat{\beta}_r = (X'X + rI)^{-1} X'y \quad \text{เมื่อ } r > 0$$

และเรียกค่า r ว่า พารามิเตอร์บริดจ์ (Ridge Parameter) หรือพารามิเตอร์เอนเอียง (Biased Parameter) และเรียกตัวประมาณ $\hat{\beta}_r$ ว่าตัวประมาณถดถอยแบบบริดจ์ (Ridge Regression Estimator)

ในการศึกษาครั้งนี้ผู้วิจัยใช้วิธีการประมาณค่า r ด้วยวิธีของ โฮเอิร์ล เคนนาร์ด และบาลด์วิน (1975) ซึ่งทำการประมาณค่า r ได้ดังนี้

$$r_{HKB} = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$$

เมื่อ p แทน จำนวนพารามิเตอร์ในตัวแบบ $(k+1)$

$\hat{\beta}$ แทน เวกเตอร์ของตัวประมาณค่าสัมประสิทธิ์ถดถอยด้วยวิธีกำลังสองน้อยที่สุด

และ $\hat{\sigma}^2 = \frac{(y - \hat{y})'(y - \hat{y})}{n - p}$ เมื่อ $\hat{y} = X\hat{\beta}_r$

4. ฟังก์ชันวัตถุประสงค์

ในงานวิจัยนี้เลือกใช้ฟังก์ชันวัตถุประสงค์ AIC เป็นเกณฑ์ในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบเพื่อให้ได้ตัวแบบที่เหมาะสม ทั้งนี้การคัดเลือกตัวแบบโดยใช้เกณฑ์ AIC จะเลือกตัวแบบที่ให้ค่า AIC ต่ำสุดเป็นตัวแบบที่เหมาะสมที่สุด โดยมีสูตรในการคำนวณดังต่อไปนี้

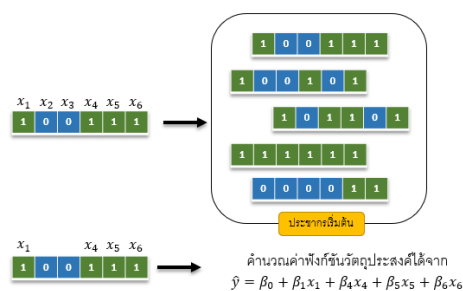
$$AIC = n \ln \left(\frac{SSE}{n} \right) + 2p$$

เมื่อ	n	แทน ขนาดตัวอย่าง
	SSE	แทน ผลบวกกำลังสองของความคลาดเคลื่อน (Error sum of squares) ของตัวแบบการถดถอย
	p	แทน จำนวนพารามิเตอร์ในตัวแบบ
	$\ln$	แทน ลอการิทึมธรรมชาติ

5. การค้นหาแบบขั้นตอนวิธีเชิงพันธุกรรม

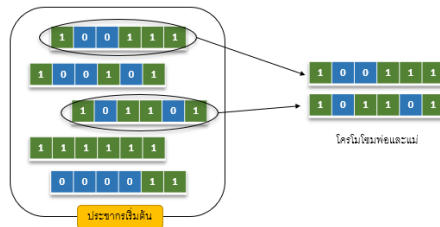
GA มีขั้นตอนในการทำงานดังนี้

ขั้นตอนที่ 1 สุ่มสร้างโครโมโซมประชากรเริ่มต้นจากเวกเตอร์ที่ใช้รหัสแบบไบนารีที่มีลักษณะเป็นเลขฐานสอง โดยมีความยาวเท่ากับจำนวนตัวแปรอิสระขึ้นมาจำนวนเท่ากับขนาดตัวอย่างที่กำหนด แล้วทำการแปลงรหัสเพื่อคำนวณค่าฟังก์ชันวัตถุประสงค์ (AIC) ของแต่ละโครโมโซม พร้อมทั้งกำหนดเกณฑ์การหยุดกระบวนการทำงานคือ $AIC_{1+ml_{GA}} - AIC_{1+(m+1)l_{GA}} < 0.001$ เมื่อ $m=0,1,2,\dots$ และ l_{GA} คือระยะห่างของรอบที่จะทำการตรวจสอบการหยุด (ในงานวิจัยนี้กำหนด l_{GA} ที่ทำให้จำนวนรอบในการทำงานของ GA มีขนาดใกล้เคียงกับ SA เพื่อทำการเปรียบเทียบประสิทธิภาพในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบของทั้ง 2 วิธี) โดยภาพที่ 1 แสดงถึงรูปแบบสายโครโมโซมที่สร้างขึ้นและการแปลงรหัส



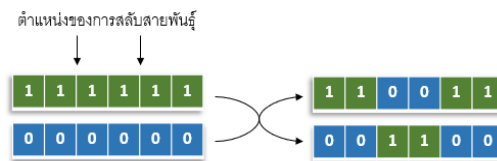
ภาพที่ 1 แสดงรูปแบบสายโครโมโซมที่สร้างขึ้นและการแปลงรหัส

ขั้นตอนที่ 2 ทำการคัดเลือกผลเฉลยเริ่มต้นจากประชากรเริ่มต้นโดยพิจารณาจากค่าฟังก์ชันวัตถุประสงค์ที่น้อยที่สุดมา 1 คู่ เรียกว่าโครโมโซมพ่อและโครโมโซมแม่ ดังภาพที่ 2



ภาพที่ 2 แสดงการคัดเลือกโครโมโซม

ขั้นตอนที่ 3 พิจารณาการสลับสายพันธุ์จากการกำหนดค่าความน่าจะเป็นเท่ากับ 0.8 เทียบกับตัวเลขสุ่ม $r \sim U(0,1)$ ถ้าค่า r มีค่าน้อยกว่าค่าความน่าจะเป็นที่กำหนดจะทำการสลับสายพันธุ์แบบ 2 จุด จากนั้นสุ่มเพื่อเลือกตำแหน่งของยีนที่จะสลับสายพันธุ์พร้อมทั้งทำการสลับสายพันธุ์ ดังภาพที่ 3



ภาพที่ 3 แสดงการสลับสายพันธุ์

ขั้นตอนที่ 4 พิจารณาการกลายพันธุ์แบบมาตรฐาน ดังภาพที่ 4 จากการกำหนดค่าความน่าจะเป็นของแต่ละยีนเท่ากับ 0.2 เทียบกับตัวเลขสุ่ม $r \sim U(0,1)$ ถ้าค่า r มีค่าน้อยกว่าค่าความน่าจะเป็นที่กำหนดจะทำการกลายพันธุ์ในยีนนั้นเมื่อทำการกลายพันธุ์เสร็จสิ้นจะคำนวณค่าฟังก์ชันวัตถุประสงค์ของโครโมโซมที่ได้



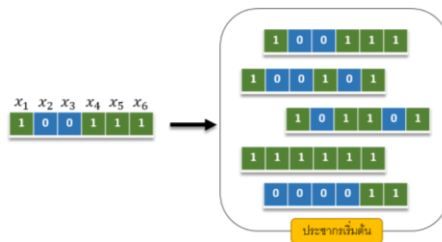
ภาพที่ 4 แสดงการกลายพันธุ์

ขั้นตอนที่ 5 พิจารณาการแทนที่ด้วยการแปลงรหัสจากโครโมโซมที่ได้ให้เป็นตัวแปรที่เหมาะสม แล้วคำนวณค่าฟังก์ชันวัตถุประสงค์หากมีค่าฟังก์ชันวัตถุประสงค์ที่ดีกว่าโครโมโซมที่อยู่ในประชากรเริ่มต้นรวมถึงโครโมโซมพ่อแม่และโครโมโซมแม่ด้วยจะทำการแทนที่โครโมโซมดังกล่าวเข้าไปในประชากรนั้นแทน ทำการตรวจสอบการหยุดกระบวนการทำงานหากเป็นไปตามเงื่อนไขที่กำหนดไว้ข้างต้นจะหยุดการทำงานทั้งหมด

6. การค้นหาแบบจำลองการอบเหี่ยว

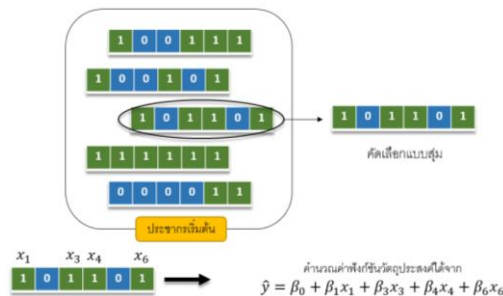
SA มีขั้นตอนในการทำงานดังนี้

ขั้นตอนที่ 1 สุ่มสร้างประชากรเริ่มต้นจากเวกเตอร์ที่ใช้รหัสแบบไบนารีที่มีลักษณะเป็นเลขฐานสอง โดยมีความยาวเท่ากับจำนวนตัวแปรอิสระขึ้นมา 1 ชุด ดังภาพที่ 5 พร้อมทั้งกำหนดค่าเริ่มต้นของอุณหภูมิไว้ที่ค่า 1000 องศา และกำหนดเกณฑ์การหยุดกระบวนการทำงาน คือการกำหนดรอบของการทำงานดังกล่าวไว้ในขั้นตอนที่ 8



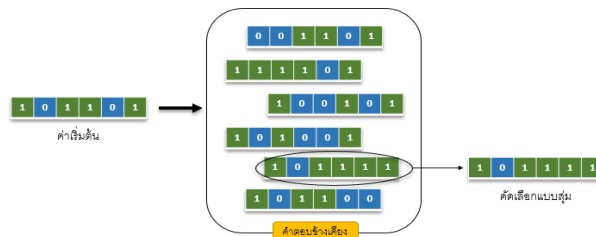
ภาพที่ 5 แสดงการสร้างประชากรเริ่มต้นและการสุ่มผลเฉลยเริ่มต้น

ขั้นตอนที่ 2 สุ่มเวกเตอร์ผลเฉลยเริ่มต้นจากประชากรเริ่มต้นมา 1 เวกเตอร์ ดังภาพที่ 6 พร้อมทั้งคำนวณค่าฟังก์ชันวัตถุประสงค์และกำหนดให้เป็นผลเฉลยที่ดีที่สุด



ภาพที่ 6 แสดงการสร้างประชากรเริ่มต้นและการสุ่มผลเฉลยเริ่มต้น

ขั้นตอนที่ 3 สร้างเซตของเวกเตอร์ผลเฉลยข้างเคียงของผลเฉลยเริ่มต้น โดยเวกเตอร์ผลเฉลยข้างเคียงจะมีความแตกต่างจากเวกเตอร์ผลเฉลยเริ่มต้นเวกเตอร์ละ 1 ตำแหน่ง แล้วทำการสุ่มเวกเตอร์ผลเฉลยข้างเคียงมา 1 ค่า พร้อมทั้งคำนวณค่าฟังก์ชันวัตถุประสงค์ ดังภาพที่ 7



ภาพที่ 7 แสดงการสร้างประชากรเวกเตอร์ผลเฉลยข้างเคียงและการสุ่มเวกเตอร์ผลเฉลยข้างเคียง

ขั้นตอนที่ 4 คำนวณผลต่างของค่าฟังก์ชันวัตถุประสงค์ระหว่างผลเฉลยเริ่มต้นและผลเฉลยข้างเคียง

ขั้นตอนที่ 5 ตรวจสอบว่าควรมีการแทนที่ผลเฉลยเดิมหรือไม่โดยใช้ความน่าจะเป็นในการตัดสินใจตามเงื่อนไขดังต่อไปนี้

$$P(\Delta E) = \begin{cases} 1 & , \Delta E < 0 \\ e^{-\Delta E/T} & , \Delta E \geq 0 \end{cases}$$

เมื่อ $P(\Delta E)$ คือ ความน่าจะเป็นในการยอมรับผลเฉลยใหม่

ΔE คือ ผลต่างของค่าฟังก์ชันวัตถุประสงค์ของผลเฉลยใหม่และผลเฉลยก่อนหน้า

T คือ อุณหภูมิปัจจุบัน

กรณีที่ ΔE มากกว่าเท่ากับศูนย์จะนำตัวเลขสุ่ม $r \sim U(0,1)$ มาช่วยในการตัดสินใจ โดยถ้า $P(\Delta E) > r$ แล้วเวกเตอร์ผลเฉลยข้างเคียงที่ถูกสุ่มมาจะถูกใช้เป็นผลเฉลยเริ่มต้นในการทำงานรอบถัดไป แต่ถ้า $P(\Delta E) \leq r$ แล้วจะคงผลเฉลยเริ่มต้นในการทำงานรอบถัดไปไว้ที่เวกเตอร์ผลเฉลยเริ่มต้นเดิม

ขั้นตอนที่ 6 นำผลเฉลยที่ได้จาก ขั้นตอนที่ 5 มาเปรียบเทียบกับผลเฉลยที่ดีที่สุดที่ได้ทำการบันทึกไว้ เพื่อพิจารณาการแทนที่ผลเฉลยที่ดีที่สุด โดยจะทำการเปรียบเทียบค่าฟังก์ชันวัตถุประสงค์ของผลเฉลยที่ดีที่สุดที่บันทึกไว้กับค่าฟังก์ชันวัตถุประสงค์ของผลเฉลยที่ได้จาก ขั้นตอนที่ 5 หากค่าฟังก์ชันวัตถุประสงค์ของผลเฉลยที่ดีที่สุดที่ถูกบันทึกไว้มีค่าที่ต่ำกว่าจะทำการบันทึกผลเฉลยที่ดีที่สุดเดิมไว้ แต่ถ้าค่าฟังก์ชันวัตถุประสงค์ของผลเฉลยที่ได้จาก ขั้นตอนที่ 5 มีค่าที่ต่ำกว่าจะทำการแทนที่ให้เป็นผลเฉลยที่ดีที่สุดใหม่แทน

ขั้นตอนที่ 7 ทำการปรับลดอุณหภูมิ โดยพิจารณาจาก $T_{k+1} = \alpha T_k$ เมื่อ T_k คือ อุณหภูมิในรอบการทำงานที่ k โดยกำหนดค่า $\alpha = 0.95$

ขั้นตอนที่ 8 ทำการตรวจสอบการหยุดกระบวนการทำงานโดยพิจารณาจากการแทนที่ของผลเฉลยที่ดีที่สุด โดยถ้าผลเฉลยที่ดีที่สุดไม่ได้รับการแทนที่เป็นจำนวน l_{SA} รอบ จะหยุดกระบวนการทำงานเมื่อกำหนดให้ l_{SA} คือจำนวนรอบที่จะทำการตรวจสอบการหยุดของ SA (ในงานวิจัยนี้กำหนด l_{SA} ที่ทำให้จำนวนรอบในการทำงานของ SA มีขนาดใกล้เคียงกับ GA เพื่อทำการเปรียบเทียบประสิทธิภาพในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบของทั้ง 2 วิธี)

6. เกณฑ์ที่ใช้ในการหยุดกระบวนการทำงาน

เกณฑ์ที่ใช้ในการหยุดกระบวนการทำงานในงานวิจัยนี้แบ่งออกเป็น 2 กรณี คือ

กรณีที่ 1 กำหนดเกณฑ์ที่ใช้ในการหยุดกระบวนการทำงานของ GA คือ $AIC_{1+ml_{GA}} - AIC_{1+(m+1)l_{GA}} < 0.001$ เมื่อ $m=0,1,2,\dots$ และ l_{GA} คือระยะห่างของรอบที่จะทำการตรวจสอบการหยุด และเกณฑ์ในการหยุดกระบวนการทำงานของวิธี SA คือ ตัวแบบที่ดีที่สุดที่ได้รับการคัดเลือกในการทำงานไม่มีการเปลี่ยนแปลงเป็นจำนวน l_{SA} รอบ เมื่อกำหนดให้ l_{SA} คือจำนวนรอบที่จะทำการตรวจสอบการหยุดกระบวนการทำงานของ SA เพื่อทำการเปรียบเทียบประสิทธิภาพในการคัดเลือก

ตัวแปรอิสระเข้าสู่ตัวแบบของทั้ง 2 วิธี โดยมีเงื่อนไข คือ ในกรณีที่มีจำนวนตัวแปรอิสระ 5 และ 10 ตัวแปร กำหนดให้ l_{SA} มีขนาดเท่ากับ 20 เท่าของจำนวนตัวแปรอิสระ ในขณะที่กรณีที่มีจำนวนตัวแปรอิสระ 20 ตัวแปร กำหนดให้ l_{SA} มีขนาดเท่ากับ 65 เท่าของจำนวนตัวแปรอิสระ และกำหนด l_{GA} ที่ทำให้จำนวนรอบในการทำงานของ GA มีขนาดใกล้เคียงกับ SA ในแต่ละกรณี

กรณีที่ 2 กำหนดเกณฑ์ที่ใช้ในการหยุดกระบวนการทำงานของ GA และ SA คือ แต่ละวิธีทำการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ถูกต้องร้อยละ 100 เพื่อทำการเปรียบเทียบความรวดเร็วในการทำงานของทั้ง 2 วิธี โดยพิจารณาจากจำนวนรอบในการทำงาน

7.เกณฑ์ที่ใช้ในการเปรียบเทียบ

กำหนดให้ร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบเป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพของ GA และ SA โดยมีสมการดังต่อไปนี้

$$\frac{C}{R} \times 100$$

เมื่อ	C	แทน จำนวนครั้งของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ถูกต้อง
	R	แทน จำนวนรอบของการทำซ้ำ

ผลการวิจัย

ผลการวิจัยจะนำเสนอแยกเป็น 2 กรณี คือ กรณีที่ไม่เกิดและเกิดปัญหาพหุสัมพันธ์เชิงเส้นต่อกันของตัวแปรอิสระ โดยแต่ละกรณีแบ่งผลการศึกษาออกเป็น 2 ส่วน คือ เมื่อมีการกำหนดเกณฑ์ในการหยุดกระบวนการทำงาน คือ กำหนดให้ l_{GA} และ l_{SA} ที่ทำให้จำนวนรอบในการทำงานของทั้ง 2 วิธีให้ใกล้เคียงกัน เพื่อทำการเปรียบเทียบประสิทธิภาพในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบของทั้ง 2 วิธี และในกรณีที่ GA และ SA ทำการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ถูกต้องร้อยละ 100 เพื่อทำการเปรียบเทียบความรวดเร็วในการทำงานของทั้ง 2 วิธี โดยพิจารณาจากจำนวนรอบในการทำงาน โดยกำหนดสัญลักษณ์แทนความหมายต่าง ๆ ดังนี้

x	แทน จำนวนตัวแปรอิสระในตัวแบบ
n	แทน ขนาดตัวอย่าง
CSP	แทน ร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบ

(Correct Specification Percentage)

ส่วนที่ 1 กำหนดเกณฑ์ในการหยุดกระบวนการทำงาน โดยกำหนดค่า l_{GA} ที่ทำให้จำนวนรอบในการทำงานของ GA ใกล้เคียงกับ SA ซึ่งในการวิจัยนี้กำหนดให้ l_{SA} มีขนาดเป็น 20 เท่าของจำนวนตัวแปรอิสระในกรณี 5 และ 10 ตัวแปร และกำหนดให้ l_{SA} มีขนาดเป็น 65 เท่าของจำนวนตัวแปรอิสระใน

กรณี 20 ตัวแปร เพื่อเปรียบเทียบร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบของ GA และ SA ผลการวิจัยแสดงดังต่อไปนี้

ตารางที่ 1 แสดงร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบของ GA และ SA และจำนวนรอบเฉลี่ยในการทำงานของแต่ละวิธี เมื่อจำนวนตัวแปรอิสระเท่ากับ 5 10 และ 20 ตัวแปร

x	n	วิธีการ	เมื่อไม่เกิดปัญหาพหุสัมพันธ์เชิงเส้น		เมื่อเกิดปัญหาพหุสัมพันธ์เชิงเส้น	
			CSP	จำนวนรอบเฉลี่ย	CSP	จำนวนรอบเฉลี่ย
5	50	GA	100	<u>114.8</u>	100	<u>115.36</u>
		SA	72	119.63	74.2	123.448
	100	GA	100	<u>116.34</u>	100	<u>116.62</u>
		SA	72.4	119.63	70.4	120.07
	200	GA	100	<u>114.24</u>	100	<u>116.2</u>
		SA	61.2	115.722	67	116.38
	500	GA	100	115.92	100	115.92
		SA	65.4	<u>113.698</u>	66	<u>113.532</u>
10	50	GA	100	<u>296.7</u>	100	<u>295.5</u>
		SA	100	299.216	100	297.642
	100	GA	100	297	100	295.2
		SA	100	<u>287.374</u>	100	<u>287.932</u>
	200	GA	100	295.5	100	296.1
		SA	100	<u>277.214</u>	100	<u>276.328</u>
	500	GA	100	294	100	292.5
		SA	100	<u>266.442</u>	100	<u>265.06</u>
20	50	GA	100	1484.82	100	1492.12
		SA	100	<u>1469.168</u>	100	<u>1471.454</u>
	100	GA	100	1492.12	100	1477.52
		SA	100	<u>1464.916</u>	100	<u>1466.192</u>
	200	GA	100	1484.82	100	1474.6
		SA	100	<u>1459.17</u>	100	<u>1461.028</u>
	500	GA	100	1483.36	100	1474.6
		SA	100	<u>1459.22</u>	100	<u>1456.428</u>

หมายเหตุ: ตัวเลขที่มีลักษณะหนาแสดงถึงร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบที่สูงที่สุดของแต่ละกรณี และตัวเลขที่ขีดเส้นใต้แสดงถึงจำนวนรอบเฉลี่ยในการทำงานที่น้อยที่สุดของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบของแต่ละกรณี

จากตารางที่ 1 พบว่าในสถานการณ์ที่ไม่เกิดและเกิดปัญหาพหุสัมพันธ์เชิงเส้นต่อกันของตัวแปรอิสระ ค่าร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบในกรณีที่จำนวนตัวแปรอิสระเท่ากับ 5 ตัว GA สามารถคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ถูกต้อง 100 เปอร์เซ็นต์ สำหรับทุกกรณีของขนาดตัวอย่าง ในขณะที่ SA สามารถคัดเลือกตัวแบบที่ถูกต้องได้เพียงร้อยละ 65 ถึงร้อยละ 72 เท่านั้น แต่เมื่อพิจารณาเวลาที่ใช้ในการค้นหาคำตอบของทั้ง 2 วิธี พบว่า ทั้ง GA และ SA ใช้จำนวนรอบ

เฉลี่ยใกล้เคียงกัน ส่วนในกรณีที่จำนวนตัวแปรอิสระเพิ่มขึ้น (10 และ 20 ตัวแปร) พบว่าทั้ง GA และ SA สามารถคัดเลือกตัวแปรอิสระได้ถูกต้องร้อยละ 100 ทุกกรณีภายใต้การศึกษา และเวลาที่ใช้ในการค้นหาคำตอบที่ดีที่สุดของทั้ง 2 วิธีใกล้เคียงกัน ผลที่ได้จากการศึกษาแสดงให้เห็นว่า ทั้ง GA และ SA สามารถทำงานได้ดีทั้งในกรณีที่เกิดและไม่เกิดปัญหาพหุสัมพันธ์เชิงเส้นต่อกันของตัวแปรอิสระ

ส่วนที่ 2 กำหนดเงื่อนไขในการหยุดกระบวนการทำงานของ GA และ SA เมื่อสามารถคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ถูกต้องร้อยละ 100 เมื่อจำนวนตัวแปรอิสระเท่ากับ 5 10 และ 20 ตัวแปร เพื่อเปรียบเทียบความรวดเร็วในการทำงานของทั้ง 2 วิธี โดยพิจารณาจากจำนวนรอบในการทำงาน ผลการวิจัยแสดงดังต่อไปนี้

ตารางที่ 2 แสดงร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบของ GA และ SA และจำนวนรอบเฉลี่ยในการทำงานของแต่ละวิธี เมื่อจำนวนตัวแปรอิสระเท่ากับ 5 10 และ 20 ตัวแปร

x	n	วิธีการ	เมื่อไม่เกิดปัญหาพหุสัมพันธ์เชิงเส้น		เมื่อเกิดปัญหาพหุสัมพันธ์เชิงเส้น	
			CSP	จำนวนรอบเฉลี่ย	CSP	จำนวนรอบเฉลี่ย
5	50	GA	100	<u>82.6</u>	100	<u>82.2</u>
		SA	100	203.58	100	204.756
	100	GA	100	<u>82.6</u>	100	<u>82.2</u>
		SA	100	199.192	100	200.428
	200	GA	100	<u>83</u>	100	<u>82.5</u>
		SA	100	196.474	100	195.918
	500	GA	100	<u>81.7</u>	100	<u>81.5</u>
		SA	100	192.408	100	193.102
	50	GA	100	296.4	100	296.4
		SA	100	<u>276.638</u>	100	<u>276.89</u>
10	100	GA	100	296.4	100	297
		SA	100	<u>268.952</u>	100	<u>269.004</u>
	200	GA	100	295.2	100	295.2
		SA	100	<u>258.898</u>	100	<u>261.098</u>
	500	GA	100	295.8	100	294.9
		SA	100	<u>248.59</u>	100	<u>248.49</u>
	50	GA	100	1490.66	100	1490.66
		SA	100	<u>418.962</u>	100	<u>419.446</u>
	100	GA	100	1471.68	100	1486.28
		SA	100	<u>414.556</u>	100	<u>413.046</u>
20	200	GA	100	1480.44	100	1478.98
		SA	100	<u>412.55</u>	100	<u>412.072</u>
	500	GA	100	1476.06	100	1490.66
		SA	100	<u>408.226</u>	100	<u>409.748</u>

หมายเหตุ: ตัวเลขที่ขีดเส้นใต้แสดงถึงจำนวนรอบเฉลี่ยในการทำงานที่น้อยที่สุดของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบของแต่ละกรณี

จากตารางที่ 2 พบว่าในสถานการณ์ที่ไม่เกิดและเกิดปัญหาพหุสัมพันธ์เชิงเส้นต่อกันของตัวแปรอิสระ เมื่อจำนวนตัวแปรอิสระเท่ากับ 5 ตัวแปร GA ใช้จำนวนรอบในการทำงานน้อยกว่า SA สำหรับทุกขนาดตัวอย่าง ในขณะที่ SA ใช้จำนวนรอบในการทำงานน้อยกว่า GA เมื่อจำนวนตัวแปรอิสระเท่ากับ 10 และ 20 ตัวแปร สำหรับทุกขนาดตัวอย่าง แสดงให้เห็นว่า SA สามารถทำงานได้รวดเร็วกว่า GA อย่างเห็นได้ชัด โดยเฉพาะอย่างยิ่ง เมื่อมิติของปัญหามีขนาดใหญ่ และจะเห็นว่าความแตกต่างของจำนวนรอบเฉลี่ยในการทำงานระหว่าง GA และ SA จะเพิ่มมากขึ้นเมื่อจำนวนของตัวแปรอิสระเพิ่มมากขึ้น ดังจะเห็นได้จากกรณีที่มีจำนวนตัวแปรอิสระ 20 ตัวแปร SA มีจำนวนรอบเฉลี่ยในการทำงานที่น้อยกว่า GA ประมาณ 3.5 เท่าสำหรับทุกขนาดตัวอย่าง

สรุปผลการวิจัยและข้อเสนอแนะ

การวิจัยนี้มีวัตถุประสงค์เพื่อศึกษาวิธีการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบการถดถอยเชิงเส้นพหุด้วยวิธี GA และ SA และเพื่อเปรียบเทียบประสิทธิภาพของทั้ง 2 วิธีโดยพิจารณาร้อยละความถูกต้องของการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบและพิจารณาความเร็วในการทำงานของทั้ง 2 วิธี ผลจากการศึกษาพบว่าประสิทธิภาพในการทำงานของทั้ง GA และ SA ขึ้นอยู่กับการปรับขั้นตอนและการปรับพารามิเตอร์ต่าง ๆ ที่เกี่ยวข้องในการทำงานของแต่ละวิธีให้มีความเหมาะสม ในการศึกษาครั้งนี้ได้ทำการปรับขั้นตอนและพารามิเตอร์ของทั้ง 2 วิธี ดังที่กำหนดไว้ในส่วนของวิธีดำเนินการวิจัย ซึ่งได้ผลการวิจัยตามที่น่าเสนอในตารางที่ 1 และตารางที่ 2 ทำให้สามารถสรุปได้ว่า GA มีประสิทธิภาพในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ดีกว่า SA เมื่อจำนวนตัวแปรอิสระค่อนข้างน้อย ในขณะที่ทั้ง GA และ SA มีประสิทธิภาพเท่ากันเมื่อจำนวนตัวแปรอิสระเพิ่มมากขึ้น อย่างไรก็ตามในตารางที่ 2 แสดงให้เห็นว่า GA และ SA จะมีประสิทธิภาพในการคัดเลือกตัวแปรอิสระเข้าสู่ตัวแบบได้ถูกต้องร้อยละ 100 ในทุก ๆ กรณีที่ทำการศึกษา หากมีการกำหนดพารามิเตอร์ที่เกี่ยวข้องที่ทำให้ทั้ง GA และ SA สามารถทำงานได้เต็มประสิทธิภาพ และเมื่อพิจารณาความรวดเร็วในการทำงาน พบว่า GA และ SA ใช้เวลาในการทำงานใกล้เคียงกัน ในขณะที่ SA ใช้เวลาในการทำงานน้อยกว่า GA ในกรณีที่จำนวนตัวแปรอิสระเพิ่มขึ้น นอกจากนี้ยังพบว่าวิธีเมตาฮิวริสติกเป็นวิธีที่มีความแข็งแกร่งต่อปัญหาพหุสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระ อย่างไรก็ตาม เพื่อเป็นการขยายขอบเขตของการศึกษา จึงควรมีการประยุกต์วิธีเมตาฮิวริสติกอื่น ๆ เช่น ขั้นตอนวิธีหาค่าเหมาะสมที่สุดด้วยระบบอาณานิคม (Ant colony optimization) หรือขั้นตอนวิธีหาค่าเหมาะสมที่สุดแบบกลุ่มอนุภาค (Particle swarm optimization) มาใช้ในการเลือกตัวแปรอิสระเข้าสู่ตัวแบบ รวมไปถึงในสถานการณ์ที่เกิดปัญหาพหุสัมพันธ์เชิงเส้นต่อกันของตัวแปรอิสระมากกว่า 1 คู่ และกรณีที่เกิดค่าผิดปกติในตัวแปรอิสระและตัวแปรตาม

เอกสารอ้างอิง

- ศรินทิพย์ หมิ่นจันทร์. (2557). การคัดเลือกตัวแบบในการถดถอยเชิงเส้นพหุคูณโดยใช้วิธีดัดเบิ้ลเจเนติก อัลกอริทึม. วิทยานิพนธ์ปริญญาวิทยาศาสตรมหาบัณฑิต, มหาวิทยาลัยเชียงใหม่, เชียงใหม่.
- พิศลย์ มีแก้ว. (2546). การเปรียบเทียบวิธีการหาค่าตอบที่ดีที่สุดด้วยวิธีทางฮิวริสติก. วิทยานิพนธ์ วิศวกรรมศาสตรมหาบัณฑิต, มหาวิทยาลัยธรรมศาสตร์.

- Kapetanious, G. (2007). Variable Selection in Regression Models Using Nonstandard Optimization of Information Criteria. *Computational Statistics & Data Analysis*. 52 (1), 4-15.
- Hoerl, Arthur E. and Kennard, Robert W. (1970). Ridge Regression: Biased Estimation to Nonorthogonal Problem. *Technometrics*. 12 (1), 55-67.
- Hoerl, Arthur E. and Kennard, Robert W. (1970). Ridge Regression: Applications to Nonorthogonal Problem. *Technometrics*. 12 (1), 69-82.
- Hoerl, Arthur E.; Kennard, Robert W. and Baldwin, Kent F. (1975). Ridge Regression: Some Simulations. *Communications in Statistics*. 4 (2), 105-123.
- Douglas C. Montgomery, Elizabeth A. Peck, and G.Geoffrey Vining. (2006). *Introduction to Linear Regression Analysis* (4th ed.). New York: Wiley-Interscience.

วันที่รับบทความ 22 ก.ค. 63, วันที่แก้ไขบทความ 31 ต.ค. 63, วันที่ตอบรับบทความ 29 ธ.ค. 63