การคัดเลือกตัวแบบในการถดถอยเชิงเส้นพหุคูณโดยใช้วิธีดับเบิลเจเนติกอัลกอริทึม

Model Selection in Multiple Linear Regression Using Double Genetic Algorithms Method

ศิรินทิพย์ หมื่นจันทร์* และวฐา มินเสน

Sirintip Muenjan* and Watha Minsan

ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่

บทคัดย่อ

งานวิจัยฉบับนี้มีวัตถุประสงค์เพื่อประยุกต์วิธีดับเบิลเจเนติกอัลกอริทึม (Double Genetic Algorithms: DGA) สำหรับการคัดเลือกตัวแบบในการถดถอยเชิงเส้นพหุคูณ และทำการเปรียบเทียบกับวิธีเจเนติกอัลกอริทึม (Genetic Algorithm: GA) และวิธีการถดถอยแบบขั้นตอน (Stepwise Regression: SR) โดยอาศัยเกณฑ์ความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Squares Error: MSE) และค่าเฉลี่ยของ MSE (Average Mean Squares Error: AMSE) เป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพ จำลองข้อมูลภายใต้สถานการณ์ที่ตัวแปรอิสระเท่ากับ 10 ตัวแปร ขนาดตัวอย่างเท่ากับ 50, 100 และ 500 ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 50, 80, 150 และ 250 เมื่อไม่เกิดและเกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ ผลการวิจัยพบว่าในสถานการณ์ที่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ วิธี DGA ให้ค่า AMSE น้อยที่สุด ส่วนวิธี GA และ SR ให้ค่าที่ใกล้เคียงกัน สำหรับกรณีที่ไม่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ วิธี DGA GA และวิธี SR ให้ค่า MSE ที่ต่ำที่สุดเท่ากัน

คำสำคัญ : การคัดเลือกตัวแบบ วิธีดับเบิลเจเนติกอัลกอริทึม วิธีการแบบขั้นตอน การถดถอยเชิงเส้นพหุคูณ

Abstract

This research was aimed to apply Double Genetic Algorithms (DGA) method for model selection in multiple linear regression and compare with Genetic Algorithm (GA), Stepwise Regression (SR) method by MSE and AMSE criterion. The simulations data under situation are 10 variables, sample sizes are 50, 100 and 500, standard deviation of errors are 50, 80, 150 and 250, without and with multicollinearity problems. Findings are found that with multicollinearity problem, DGA method to the AMSE minimum but GA and SR are similarly. Without multicollinearity problem, the lowest MSE of DGA, GA, and SR method are almost the same.

Keywords : Model Selection, Double Genetic Algorithms, Stepwise Regression, Multiple Linear regression

*Corresponding author. Email : jookjik072@gmail.com

บทนำ

ในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ เป็นการหารูปแบบความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ (พลากร สีน้อย และจิราวัลย์ จิตรถเวช, 2553) โดยการสร้างตัวแบบเพื่ออธิบายความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระซึ่งมีความสัมพันธ์ในรูปเชิงเส้น รวมถึงการพยากรณ์ค่าตัวแปรตาม จากตัวแปรอิสระ ซึ่งการใช้ตัวแบบการถดถอยเชิงเส้นในการพยากรณ์ให้มีความถูกต้องและแม่นยำที่สุดนั้น ขึ้นอยู่กับตัวแบบที่มีความเหมาะสม การได้ตัวแบบที่มีความเหมาะสมต้องมาจากการคัดเลือกตัวแปรอิสระที่มีอิทธิพลต่อตัวแปรตาม ในทางปฏิบัติอาจพบว่าตัวแปรอิสระที่ใช้ในการอธิบายตัวแปรตามนั้นมีอยู่เป็นจำนวนมาก แต่ตัวแปรเหล่านั้นมีความสามารถในการอธิบายตัวแปรตามได้ไม่เท่ากันและบางตัวที่นำมาใช้ในการวิเคราะห์การถดถอยอาจมีความสัมพันธ์กันเอง หรือเรียกว่าเกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ (Multicollinearity) ซึ่งจะส่งผลให้ความแปรปรวนของค่าประมาณพารามิเตอร์ของตัวแบบมีค่าสูง (พิษณุ เจียวคุณ , 2548; เกตุจันทร์ จำปาไชยศรี, 2550) ในปัจจุบันมีวิธีการคัดเลือกตัวแบบการถดถอยเชิงเส้นพหุคูณอยู่หลายวิธี โดยวิธีที่นิยมใช้กันอย่างแพร่หลาย เช่น วิธีพิจารณาทุกตัวแบบที่เป็นไปได้ (All Possible Regression) วิธีการเพิ่มตัวแปร (Forward Selection: FS) วิธีการลดตัวแปร (Backward Elimination: BE) และวิธีการถดถอยแบบขั้นตอน (Stepwise Regression: SR) เป็นต้น นอกจากนี้ยังมีวิธีการคัดเลือกตัวแบบที่ประยุกต์มาจากวิธีการที่ใช้ในการแก้ปัญหาที่มีชุดคำตอบที่เป็นไปได้แน่นอน (Combinatorial Optimization: CO) เป็นกลุ่มวิธีการที่เรียกว่า เมตาฮิวริสติก (Metaheuristic) เช่น วิธีเจเนติกอัลกอริทึม (Genetic Algorithm: GA) วิธีทาบูเสิร์ช (Tabu Search: TS) และวิธีซิมมูเลเตดแอลเนลลิ่ง (Simulated Annealing: SA) (Drezner, 1999) โดยวิธีการเหล่านี้เป็นกระบวนการสำหรับแก้ปัญหาเพื่อหาค่าเหมาะสมที่สุดซึ่งมีขั้นตอนมาตรฐานที่แน่นอน รวมถึงมีเกณฑ์ในการหยุดกระบวนการทำงานเหมือนกัน นั่นคือ ครบจำนวนรอบที่กำหนด ได้ค่าฟังก์ชันวัตถุประสงค์ (Objective Function) ที่ต้องการหรือเหตุผลอื่นๆ ที่ได้กำหนดไว้ ดังนั้นจึงสามารถนำแนวคิดของวิธีการเหล่านี้มาช่วยในการคัดเลือกตัวแบบได้เช่นกัน โดยมีเป้าหมายในการคัดเลือกตัวแบบการถดถอยที่เหมาะสมที่สุดในการอธิบายตัวแปรตาม และมีความคลาดเคลื่อนในการพยากรณ์ต่ำ รวมถึงช่วยลดค่าใช้จ่ายและเวลาในการเก็บรวบรวมข้อมูลเกี่ยวกับตัวแปรอิสระในตัวแบบการถดถอยเชิงเส้นพหุคูณที่จะนำมาใช้ในการพยากรณ์ตัวแปรตาม (Renner และ Ekart, 2003; Thomas, 2008)

จากผลงานวิจัยของ Wasserman และ Sudjianto (1994) ได้ทำการศึกษาการคัดเลือกตัวแบบในการวิเคราะห์การถดถอยเชิงเส้น โดยวิธี GA และนำมาเปรียบเทียบกับวิธี SR เกณฑ์ที่ใช้ในการพิจารณาเลือกตัวแบบคือค่า AIC และพิจารณาเปรียบเทียบทั้ง 2 วิธี โดยใช้ค่า MSE เป็นการศึกษาที่อาศัยวิธี GA ในการคัดเลือกตัวแบบแล้วทำการประมาณค่าสัมประสิทธิ์การถดถอยของตัวแบบด้วยวิธีกำลังสองน้อยที่สุด (Ordinary Least Squares: OLS) ในกรณีที่ตัวแปรอิสระไม่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ ข้อมูลที่ใช้ในการวิจัยมาจากการจำลองข้อมูล โดยมีตัวแปรอิสระทั้งหมด 24 ตัวแปร ผลการศึกษาพบว่ามีตัวแปรอิสระอยู่ในตัวแบบทั้งหมด 7 ตัวแปร คือ $X_4, X_8, X_{14}, X_{15}, X_{17}, X_{20}$ และ X_{22} โดยตัวแบบที่ได้จากวิธี GA จะให้ค่า MSE เป็น 1,417.49 ในขณะที่วิธี SR จะให้ค่า MSE เป็น 1,548.52 ดังนั้นในกรณีที่มีจำนวนตัวแปรอิสระมาก วิธี GA จะมีความแม่นยำสูงกว่าวิธี SR รวมถึง Pasha (2002) ที่ทำการศึกษาเปรียบเทียบวิธีการคัดเลือกตัวแปรในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ โดยวิธี FS BE และ SR ผลการศึกษาพบว่าวิธี FS และ BE ให้ผลลัพธ์ในการคัดเลือกตัวแปรเหมือนกันในขณะที่วิธี SR คัดเลือกตัวแปรได้ดีกว่า 2 วิธี ข้างต้น และ Kapetanious (2007) ได้ทำการศึกษาการคัดเลือกตัวแบบการถดถอยในกรณีที่เกิดปัญหาทางเศรษฐมิติ โดยใช้วิธี SA และ GA เปรียบเทียบกับวิธีเกณฑ์ข้อสนเทศอาไคเคะ (Akaike Information Criterion: AIC) และเกณฑ์ข้อสนเทศเบส์ (Bayesian Information Criterion: BIC) ผลการศึกษาพบว่าวิธี SA และ GA ให้ความแม่นยำในการพยากรณ์สูงกว่าวิธีอื่นๆ โดยเฉพาะอย่างยิ่งในกรณีที่มีจำนวนตัวแปรอิสระสูง นอกจากนี้ Hasan (2013) ได้ทำการศึกษาการคัดเลือกตัวแบบในการวิเคราะห์การถดถอย

เชิงเส้นพหุคูณโดยเสนอวิธีการแบบผสม (Hybrid GSA) ระหว่างวิธี GA และ SA ในการคัดเลือกตัวแปรของตัวแบบพร้อมทั้งเปรียบเทียบวิธีดังกล่าวกับวิธี GA FS และ BE ศึกษาในกรณีที่ไม่เกิดสหสัมพันธ์เชิงเส้นพหุ และศึกษาตัวแปรอิสระตั้งแต่ 11 ถึง 25 ตัวแปร ที่ขนาดตัวอย่างเท่ากับ 100 และ 200 เกณฑ์ที่ใช้ในการพิจารณาเลือกตัวแบบคือค่า AIC โดยใช้ข้อมูลในการวิเคราะห์ 12 ชุดข้อมูล ผลการศึกษาพบว่าการเปรียบเทียบค่า AIC วิธี Hybrid GSA ให้ค่าต่ำกว่าวิธี FS และ BE แต่ให้ค่าที่ใกล้เคียงกับวิธี GA เมื่อมีตัวแปรเพิ่มขึ้น ซึ่งให้ผลที่คล้ายกันทั้งสองขนาดตัวอย่าง

ดังนั้นผู้วิจัยจึงนำแนวคิดของวิธี GA สำหรับการคัดเลือกตัวแบบทำงานร่วมกับวิธี GA สำหรับหาค่าสัมประสิทธิ์การถดถอยของตัวแบบ โดยเรียกว่า วิธีดับเบิลเจเนติกอัลกอริทึม (Double Genetic Algorithms: DGA) มาประยุกต์ใช้ในการคัดเลือกตัวแบบการถดถอยเชิงเส้นพหุคูณ เพื่อให้ได้ตัวแบบที่เหมาะสมที่สุดในการพยากรณ์ตัวแปรตาม โดยให้ค่าความคลาดเคลื่อนในการพยากรณ์น้อยที่สุด และนำวิธี DGA เปรียบเทียบกับวิธี GA และ SR ที่ประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธี OLS อาศัยเกณฑ์ AIC เป็นเกณฑ์ในการพิจารณาคัดเลือกตัวแบบ รวมถึงใช้เกณฑ์ความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) และค่าเฉลี่ยของ MSE (Average Mean Squares Error: AMSE) เป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพของทั้ง 3 วิธี

ทฤษฎีที่เกี่ยวข้อง

(1) ตัวแบบการถดถอยเชิงเส้นพหุคูณมีรูปแบบทั่วไปดังนี้

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

โดยที่ $\mathbf{y}$ แทน เวกเตอร์ของตัวแปรตาม $\mathbf{y}$ ขนาด $n\times 1$ เมื่อ n แทนขนาดตัวอย่าง

$\mathbf{X}$ แทน เมตริกซ์ของตัวแปรอิสระขนาด $n\times(k\times 1)$ เมื่อ k แทนจำนวนตัวแปรอิสระ

$\boldsymbol{\beta}$ แทน เวกเตอร์ของพารามิเตอร์ของตัวแบบขนาด $(k\times 1)\times 1$

$\boldsymbol{\varepsilon}$ แทน เวกเตอร์ของความคลาดเคลื่อน ขนาด $n\times 1$

ข้อตกลงเบื้องต้น (Assumptions) ของการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ ได้แก่

1) $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ และความคลาดเคลื่อนแต่ละตัวเป็นอิสระต่อกัน โดย $\mathbf{I}_n$ แทน เมทริกซ์เอกลักษณ์ขนาด $n\times n$

2) ตัวแปรอิสระแต่ละตัว เป็นตัวแปรที่ทราบค่า และเป็นอิสระต่อกัน

3) ตัวแบบการถดถอยเป็นตัวแบบเชิงเส้นของพารามิเตอร์ (นันทพร บุญสุข, 2555)

การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธี OLS มีแนวคิดโดยเลือกค่าประมาณ $\boldsymbol{\beta}$ ในพจน์ของ $\mathbf{y}$ และ $\mathbf{X}$ เพื่อที่จะทำให้ผลบวกกำลังสองของความคลาดเคลื่อนมีค่าน้อยที่สุด จากตัวแบบ (1) เมื่อ $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$ ตัวประมาณค่ากำลังสองเฉลี่ยน้อยที่สุดของ $\boldsymbol{\beta}$ คือ $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ ซึ่งเป็นตัวประมาณค่าที่ไม่เอนเอียง และมีความแปรปรวนเท่ากับ $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ โดย $\mathrm{E}(\mathbf{b}) = \boldsymbol{\beta}$ และ $\mathrm{Var}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ (วิชิต หล่อจีระชุณห์กุล, 2524)

(2) เกณฑ์ AIC สร้างจากการประมาณความแปรปรวนของข้อสนเทศคูลส์แบล็ค-ไลท์เบอร์ (Kullback-Leibler Information) ระหว่างตัวแบบจริงกับตัวแบบที่เหมาะสมที่มีคุณสมบัติไม่เอนเอียง การคัดเลือกตัวแบบโดยใช้เกณฑ์ AIC จะเลือกตัวแบบที่ให้ค่า AIC ต่ำที่สุดเป็นตัวแบบที่เหมาะสมที่สุด ซึ่งจะคัดเลือกตัวแบบได้ดีเมื่อตัวอย่างมีขนาดใหญ่ สมการของเกณฑ์ AIC มีดังต่อไปนี้

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

เมื่อ n แทน ขนาดตัวอย่าง

SSE แทน ผลรวมความคลาดเคลื่อนกำลังสอง (Sum Squares Error) ของตัวแบบการถดถอย

p แทน จำนวนพารามิเตอร์ในตัวแบบการถดถอย

$\ln$ แทน ลอการิทึมฐานอี

(3) เกณฑ์ MSE สำหรับการวิจัยครั้งนี้ มีรูปแบบดังต่อไปนี้

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1} \qquad (3)$$

เมื่อ y_i แทน ตัวแปรตาม เมื่อ $i = 1,...,n$

$\hat{y}_i$ แทน ตัวแปรตามที่ได้จากการพยากรณ์ เมื่อ $i = 1,...,n$

n แทน ขนาดตัวอย่าง

k แทน จำนวนตัวแปรอิสระ

(4) ค่าความคลาดเคลื่อนมาตรฐาน (Standard Error, SE) ของการประมาณ β_j

$$\mathrm{SE}(\mathrm{b}_j) = \sqrt{S^2(\mathbf{X}'\mathbf{X})^{-1}_{jj}} \qquad (4)$$

เมื่อ j แทน ตำแหน่งของสมาชิกในแนวทแยงมุมของเมทริกซ์

S^2 แทน ความแปรปรวนของการประมาณค่า $\mathbf{y}$

$\mathbf{X}$ แทน เมตริกซ์ของตัวแปรอิสระขนาด $n \times (k \times 1)$ เมื่อ k แทนจำนวนตัวแปรอิสระ

(5) วิธีการ Stepwise Regression (SR)

เป็นวิธีการผสมระหว่างวิธี FS และ BE โดยเลือกตัวแปรอิสระเข้าในตัวแบบการถดถอยครั้งละหนึ่งตัว โดยมีขั้นตอนดังต่อไปนี้

ขั้นที่ 1 เริ่มจากตัวแบบที่ไม่มีตัวแปรอิสระตัวใดอยู่ในตัวแบบ

ขั้นที่ 2 ทำการพิจารณาเลือกตัวแปรอิสระตัวแรกเข้าในตัวแบบการถดถอยโดยพิจารณาจากค่า AIC ที่ต่ำที่สุด นั้นคือ การพิจารณาค่า AIC ที่ตัวแปรเหล่านั้นอยู่ในตัวแบบ หากตัวแปรอิสระตัวใดอยู่ในตัวแบบแล้วทำให้ได้ค่า AIC ที่ต่ำที่สุด จะทำการเลือกตัวแปรนั้นเป็นตัวแรกของตัวแบบการถดถอย

ขั้นที่ 3 พิจารณาเลือกตัวแปรอิสระตัวถัดไปเข้าสู่ตัวแบบเมื่อมีตัวแปรอิสระก่อนหน้านี้อยู่ในตัวแบบด้วย โดยพิจารณาค่า AIC ของตัวแบบที่มีตัวแปรที่ยังเหลืออยู่แต่ละตัวในตัวแบบ แล้วให้ค่า AIC ต่ำที่สุด จะทำการเลือกตัวแบบนั้น เพื่อเข้าสู่ขั้นตอนต่อไป

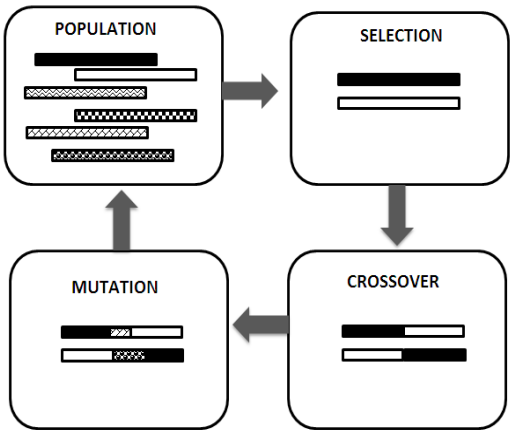
ขั้นที่ 4 จากตัวแบบการถดถอยที่ได้จะทำการพิจารณาค่า AIC ของตัวแบบเมื่อทำการตัดตัวแปรอิสระก่อนหน้านี้ออก หากค่า AIC ของตัวแบบนั้นมีค่าที่ลดลง จะทำการหยุดกระบวนการทำงาน ซึ่งแสดงว่าตัวแบบที่ได้นี้เป็นตัวแบบที่เหมาะสมที่สุดโดยให้ค่า AIC ต่ำที่สุดแล้ว หากมีค่ามากกว่า จะเข้าสู่ขั้นตอนต่อไป

ขั้นที่ 5 กระทำซ้ำในขั้นที่ 3 และขั้นที่ 4 จนกว่าจะไม่มีตัวแปรอิสระตัวใดเข้าสู่ตัวแบบหรือถูกตัดออกจากตัวแบบอีก

ขั้นที่ 6 นำตัวแบบที่ได้มาทำการประมาณค่าสัมประสิทธิ์การถดถอยโดยวิธี OLS จากนั้นคำนวณค่า MSE ของตัวแบบที่ได้

(6) วิธี Genetic Algorithms (GA)

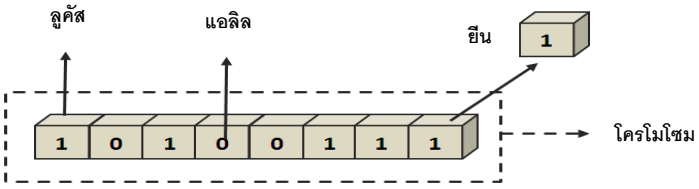
GA เป็นวิธีการแก้ปัญหารูปแบบหนึ่งเพื่อให้ได้คำตอบเหมาะสมที่สุดเป็นวิธีการที่คิดค้นโดย John Holland ซึ่งได้รับแรงบันดาลใจจากทฤษฎีวิวัฒนาการของ Charles Darwin ซึ่งเป็นทฤษฎีอธิบายเกี่ยวกับวัฏจักรทางพันธุกรรมของสิ่งมีชีวิต โดยมีจุดมุ่งหมายเพื่ออธิบายการเปลี่ยนแปลงกระบวนการทางธรรมชาติของพันธุกรรมและนำกลไกการเปลี่ยนแปลงเหล่านี้มาประยุกต์ใช้ในการแก้ปัญหาการหาค่าเหมาะสมที่สุด ซึ่งอาจเป็นค่าต่ำสุดหรือค่าสูงสุด ขึ้นอยู่กับรูปแบบของแต่ละปัญหา วัฏจักรของ GA แสดงให้เห็นถึงการอยู่รอดของสิ่งมีชีวิตในธรรมชาติ สิ่งมีชีวิตที่มีการปรับตัวให้เข้ากับสภาพแวดล้อมได้ดีกว่าจะสามารถอยู่รอดได้ ในขณะที่สิ่งมีชีวิตอื่นๆ ที่ไม่สามารถปรับตัวเองได้จะต้องสูญพันธุ์ไป การปรับตัวดังกล่าวแสดงว่าสิ่งมีชีวิตนั้นมีวิวัฒนาการเกิดขึ้นซึ่งเป็นการถ่ายทอดลักษณะทางพันธุกรรม ดังนั้นหลักการทำงานของ GA จึงถูกนำเสนอข้อมูลในรูปแบบโครโมโซม ขั้นตอนการทำงานของ GA เริ่มต้นจากการสร้างประชากรของคำตอบหรือที่เรียกว่าโครโมโซมจากการสุ่ม จากนั้นจึงทำการถอดรหัสโครโมโซม และคำนวณค่าความเหมาะสม (Fitness Value) ของแต่ละโครโมโซมจากฟังก์ชันวัตถุประสงค์ ประชากรเหล่านี้จะต้องผ่านตัวดำเนินการทางพันธุกรรม เพื่อให้เกิดการปรับเปลี่ยนสายพันธุ์ ซึ่งมีอยู่ 3 วิธี ได้แก่ การคัดเลือกสายพันธุ์ (Selection) การสลับสายพันธุ์ (Crossover) และการกลายพันธุ์ (Mutation) (Deep และ Thakur, 2007) ซึ่งกระบวนการดังกล่าวจะถูกทำซ้ำไปเรื่อยๆจนกว่าจะตรงกับเงื่อนไขในการหยุดค้นหา ซึ่งจะได้โครโมโซมที่เหมาะสมที่สุด หรือใกล้เคียงค่าที่ดีที่สุด ดังแสดงในภาพที่ 1 (Luca Scrucca, 2013)



ภาพที่ 1 แสดงขั้นตอนการทำงานทั่วไปของ GA

โดยมีขั้นตอนดังต่อไปนี้

ขั้นที่ 1 สร้างประชากรขึ้นมาจากการสุ่มเพื่อเป็นประชากรเริ่มต้น ของประชากรรุ่นที่ 1 โดยอยู่ในลักษณะสายโครโมโซมประกอบด้วยยีนแบบไบนารี ดังแสดงในภาพที่ 2 จำนวน 200 สายโครโมโซม โดยจำนวนยีนจะเท่ากับจำนวนตัวแปรอิสระ พร้อมทั้งแปลงรหัสเพื่อคำนวณค่า AIC ของแต่ละโครโมโซม



ภาพที่ 2 แสดงลักษณะของสายโครโมโซม

ขั้นที่ 2 ทำการคัดเลือกประชากรเพื่อเป็นโครโมโซมต้นแบบหรือเรียกว่าโครโมโซมพ่อแม่จากโครโมโซมที่มีค่า AIC ที่น้อยที่สุด 2 อันดับ

ขั้นที่ 3 กำหนดความน่าจะเป็นของการสลับสายพันธุ์ เท่ากับ 0.8 จากนั้นสุ่มตัวเลขในช่วง 0 ถึง 1 หากมีค่าน้อยกว่าหรือเท่ากับ 0.8 จะทำการสลับสายพันธุ์ ในที่นี้จะทำการ สลับสายพันธุ์ 2 จุด จากนั้นสุ่มเพื่อเลือกตำแหน่งที่จะสลับสายพันธุ์ พร้อมทั้งทำการสลับสายพันธุ์

ขั้นที่ 4 กำหนดความน่าจะเป็นของการกลายพันธุ์ เท่ากับ 0.2 จากนั้นสุ่มตัวเลขในช่วง 0 ถึง 1 หากมีค่าน้อยกว่าหรือเท่ากับ 0.2 จะทำการกลายพันธุ์ ในกรณีที่มีการกลายพันธุ์ ตำแหน่งและจำนวนตำแหน่งที่จะทำการกลายพันธุ์ เป็นไปโดยสุ่ม จากนั้นจึงทำการกลายพันธุ์ พร้อมทั้งคำนวณค่า AIC ของโครโมโซมที่ได้

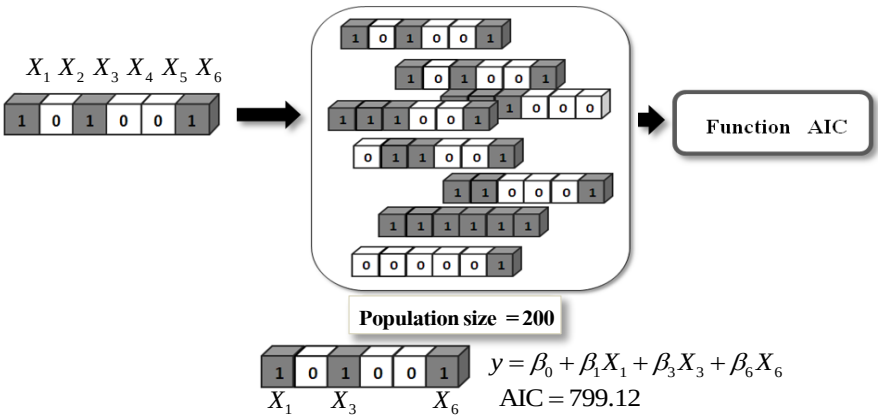
ขั้นที่ 5 ทำการแปลงรหัสจากโครโมโซมที่ได้มาเป็นตัวแปรที่เหมาะสมจะอยู่ในตัวแบบการถดถอยพร้อมทั้งคำนวณค่า AIC ทำการเลือกโครโมโซมที่ให้ค่า AIC น้อยที่สุด เพื่อนำมาเป็นตัวแบบที่เหมาะสมที่สุดสำหรับประชากรรุ่นที่ 1 โดยนำโครโมโซมพ่อแม่มาพิจารณาด้วย ดำเนินการทั้งสิ้นจำนวน 2,000 รุ่น แล้วนำโครโมโซมในรุ่นที่ 2,000 มาดำเนินการในขั้นต่อไป

ขั้นที่ 6 เมื่อได้โครโมโซมที่ให้ค่า AIC ต่ำที่สุดแล้ว ทำการประมาณค่าสัมประสิทธิ์การถดถอยโดยวิธี OLS จากนั้นคำนวณค่า MSE

วิธี Double Genetic Algorithms (DGA)

วิธี DGA เป็นวิธีการในการคัดเลือกตัวแบบการถดถอยเชิงเส้นพหุคูณซึ่งประกอบด้วย 2 ส่วน โดยส่วนที่ 1 คือ วิธี GA สำหรับการคัดเลือกตัวแบบทำงานร่วมกับส่วนที่ 2 คือ วิธี GA สำหรับการหาค่าสัมประสิทธิ์การถดถอยของตัวแบบ ซึ่งในส่วนที่ 1 จะมีกระบวนการทำงานเช่นเดียวกับวิธี GA ใน (6) โดยมีขั้นตอนดังนี้

ขั้นที่ 1 สร้างประชากรขึ้นมาจากการสุ่มเพื่อเป็นประชากรเริ่มต้น ของประชากรรุ่นที่ 1 โดยอยู่ในลักษณะสายโครโมโซมประกอบด้วยยีนแบบไบนารี ดังแสดงในภาพที่ 2 จำนวน 200 สายโครโมโซม โดยจำนวนยีนจะเท่ากับจำนวนตัวแปรอิสระ พร้อมทั้งแปลงรหัสเพื่อคำนวณค่า AIC ของแต่ละโครโมโซม ดังแสดงในภาพที่ 3

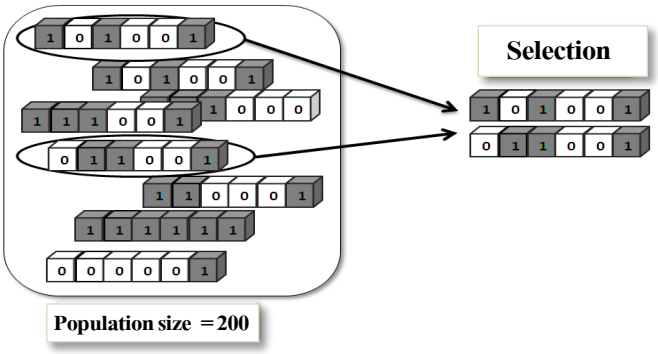


ภาพที่ 3 แสดงรูปแบบสายโครโมโซม จำนวนโครโมโซมที่สร้างขึ้นและการแปลงรหัสโครโมโซมของส่วนที่ 1

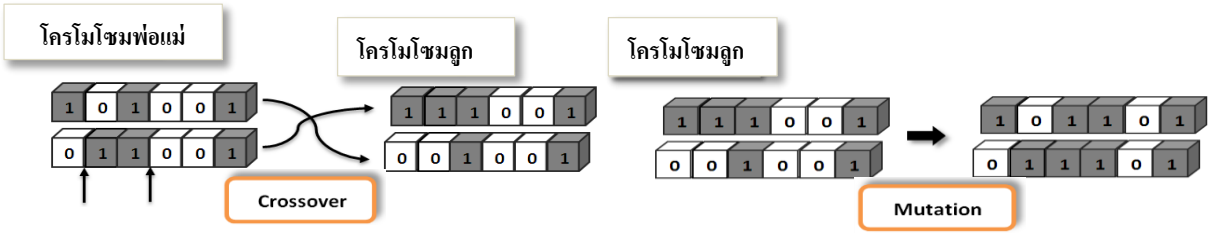
ขั้นที่ 2 ทำการคัดเลือกประชากรเพื่อเป็นโครโมโซมต้นแบบหรือเรียกว่าโครโมโซมพ่อแม่ จากโครโมโซมที่มีค่า AIC ที่น้อยที่สุด 2 อันดับแรก ดังแสดงในภาพที่ 4

ขั้นที่ 3 กำหนดความน่าจะเป็นของการสลับสายพันธุ์ เท่ากับ 0.8 จากนั้นสุ่มตัวเลขในช่วง 0 ถึง 1 หากมีค่าน้อยกว่า หรือเท่ากับ 0.8 จะทำการสลับสายพันธุ์ ในที่นี้จะทำการสลับสายพันธุ์ 2 จุด จากนั้นสุ่มเพื่อเลือกตำแหน่งที่จะสลับสายพันธุ์ พร้อมทั้งทำการสลับสายพันธุ์ ดังแสดงในภาพที่ 5

ขั้นที่ 4 กำหนดความน่าจะเป็นของการกลายพันธุ์ เท่ากับ 0.2 จากนั้นสุ่มตัวเลขในช่วง 0 ถึง 1 หากมีค่าน้อยกว่าหรือเท่ากับ 0.2 จะทำการกลายพันธุ์ ในกรณีที่มีการกลายพันธุ์ ตำแหน่งและจำนวนตำแหน่งที่จะทำการกลายพันธุ์ เป็นไปโดยสุ่ม จากนั้นจึงทำการกลายพันธุ์ พร้อมทั้งคำนวณค่า AIC ของโครโมโซมที่ได้ ดังแสดงในภาพที่ 5

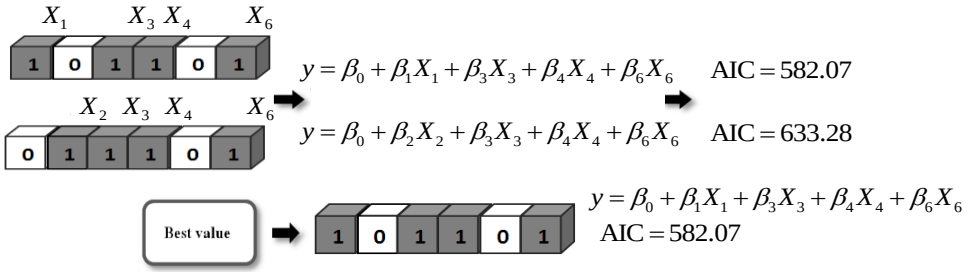


ภาพที่ 4 แสดงการคัดเลือกโครโมโซมของส่วนที่ 1



ภาพที่ 5 แสดงการสลับสายพันธุ์และการกลายพันธุ์โครโมโซมของส่วนที่ 1

ขั้นที่ 5 ทำการแปลงรหัสจากโครโมโซมที่ได้มาเป็นตัวแปรที่เหมาะสมจะอยู่ในตัวแบบถดถอยพร้อมทั้งคำนวณค่า AIC ทำการเลือกโครโมโซมที่ให้ค่า AIC น้อยที่สุด เพื่อนำมาเป็นตัวแบบที่เหมาะสมที่สุดสำหรับประชากรรุ่นที่ 1 โดยนำโครโมโซมพ่อแม่มาพิจารณาด้วย ดังแสดงในภาพที่ 6 โดยดำเนินการทั้งสิ้นจำนวน 2,000 รุ่น แล้วนำโครโมโซมในรุ่นที่ 2,000 มาดำเนินการในขั้นต่อไป

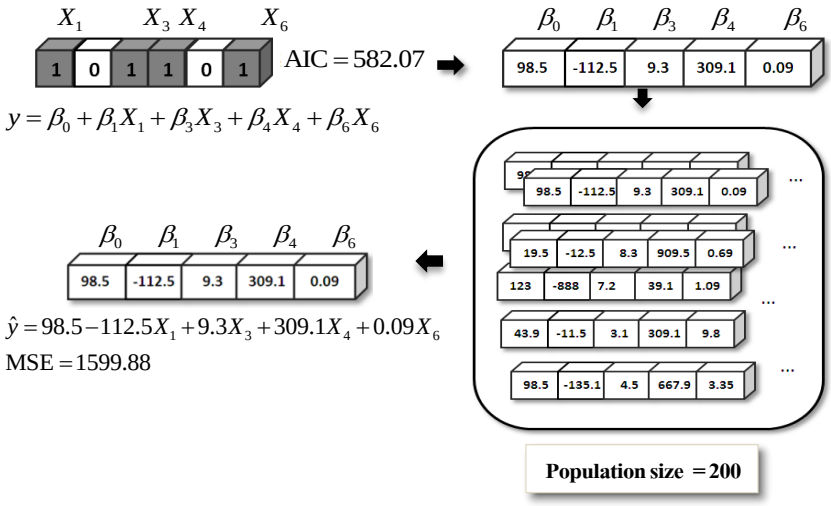


ภาพที่ 6 แสดงการแปลงรหัสของโครโมโซมของประชากรรุ่นที่ 1 และเลือกโครโมโซมของส่วนที่ 1

ขั้นที่ 6 เมื่อได้โครโมโซมที่ให้ค่า AIC ต่ำที่สุดแล้ว ทำการสร้างขอบเขตของการหาค่าสัมประสิทธิ์การถดถอยเพื่อทำการสุ่มหาค่าในการสร้างโครโมโซมในขั้นตอนต่อไป ซึ่งสร้างจากการประมาณค่าสัมประสิทธิ์การถดถอยของ OLS และค่า SE โดย ขอบล่างเท่ากับค่าสัมประสิทธิ์การถดถอยของ OLS – (2 x SE) ขอบบนเท่ากับค่าสัมประสิทธิ์การถดถอยของ OLS + (2 x SE)

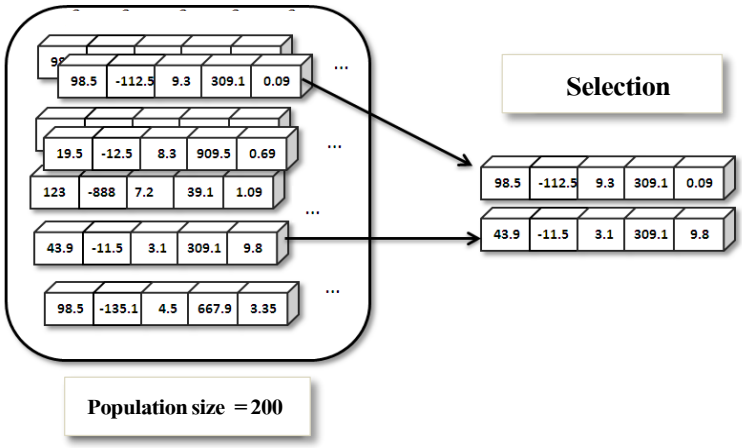
ส่วนที่ 2

ขั้นที่ 1 จากโครโมโซมที่ได้ในส่วนที่ 1 ขั้นที่ 6 นำมาสร้างโครโมโซมแบบค่าจริงจากการสุ่มในช่วงขอบเขตที่กำหนดไว้ เพื่อเป็นประชากรเริ่มต้น โดยจำนวนยีนจะเท่ากับจำนวนตัวแปรอิสระที่ถูกเลือกให้อยู่ในตัวแบบที่เหมาะสมและ β_0 พร้อมทั้งแปลงรหัสเพื่อคำนวณค่า MSE ของแต่ละโครโมโซม ดังแสดงในภาพที่ 7



ภาพที่ 7 แสดงการสร้างประชากรเริ่มต้น และการแปลงรหัสโครโมโซมของส่วนที่ 2

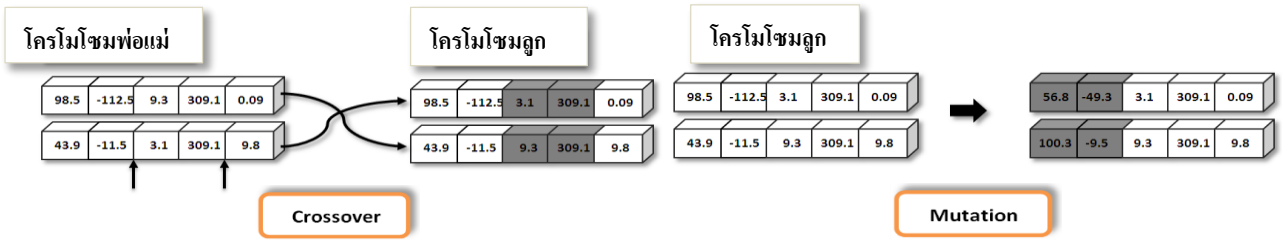
ขั้นที่ 2 ทำการคัดเลือกประชากรเพื่อเป็นโครโมโซมพ่อแม่ จากโครโมโซมที่มีค่า MSE ที่น้อยที่สุด 2 อันดับแรก ดังแสดงในภาพที่ 8



ภาพที่ 8 แสดงการคัดเลือกโครโมโซมของส่วนที่ 2

ขั้นที่ 3 กำหนดความน่าจะเป็นของการสลับสายพันธุ์ เท่ากับ 0.8 จากนั้นสุ่มตัวเลขในช่วง 0 ถึง 1 หากมีค่าน้อยกว่าหรือเท่ากับ 0.8 จะทำการสลับสายพันธุ์ ในที่นี้จะทำการสลับสายพันธุ์ 2 จุด จากนั้นสุ่มเพื่อเลือกตำแหน่งที่จะสลับสายพันธุ์ พร้อมทั้งทำการสลับสายพันธุ์ ดังแสดงในภาพที่ 9

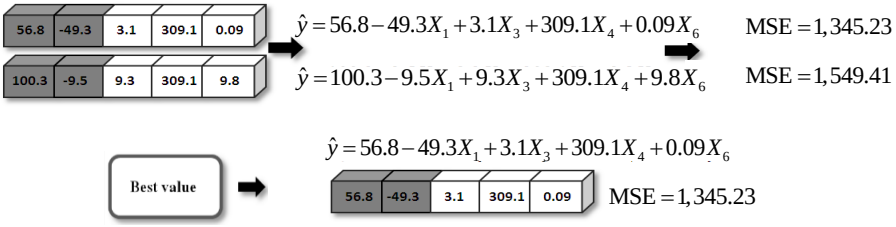
ขั้นที่ 4 กำหนดความน่าจะเป็นของการกลายพันธุ์ เท่ากับ 0.2 จากนั้นสุ่มตัวเลขในช่วง 0 ถึง 1 หากมีค่าน้อยกว่าหรือเท่ากับ 0.2 จะทำการกลายพันธุ์ ในกรณีที่มีการกลายพันธุ์ ตำแหน่งและจำนวนตำแหน่งที่จะทำการกลายพันธุ์เป็นไปโดยสุ่ม จากนั้นจึงทำการกลายพันธุ์ พร้อมทั้งคำนวณค่า MSE ของโครโมโซมที่ได้ ดังแสดงในภาพที่ 9



ภาพที่ 9 แสดงการสลับสายพันธุ์และการกลายพันธุ์โครโมโซมของส่วนที่ 2

ขั้นที่ 5 ทำการแปลงรหัสจากโครโมโซมที่ได้มาเป็นตัวแปรที่เหมาะสมจะอยู่ในตัวแบบการถดถอยพร้อมทั้งคำนวณค่า MSE ทำการเลือกโครโมโซมที่ให้ค่า MSE น้อยที่สุด เพื่อนำมาเป็นตัวแบบที่เหมาะสมที่สุดสำหรับประชากรรุ่นที่ 1 โดยนำโครโมโซมพ่อแม่มาพิจารณาด้วย ดังแสดงในภาพที่ 10 โดยดำเนินการทั้งสิ้นจำนวน 2,000 รุ่น

ขั้นที่ 6 นำโครโมโซมในรุ่นที่ 2,000 มาเป็นคำตอบในแต่ละรอบของการทำซ้ำในแต่ละสถานการณ์ที่สร้างขึ้น



ภาพที่ 10 แสดงการแปลงรหัสของโครโมโซมของประชากรรุ่นที่ 1 และเลือกโครโมโซมของส่วนที่ 2

วิธีการวิจัย

การวิจัยครั้งนี้เป็นการศึกษาวิธีการในการคัดเลือกตัวแบบการถดถอยเชิงเส้นพหุคูณ โดยวิธี DGA และเปรียบเทียบวิธีดังกล่าวกับวิธี GA และ SR โดยใช้เกณฑ์ AIC เป็นเกณฑ์ในการพิจารณาคัดเลือกตัวแบบ และอาศัยค่า MSE ที่ต่ำที่สุดและค่า AMSE เป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพของทั้ง 3 วิธี สำหรับข้อมูลที่ใช้ในการวิจัยได้จากการจำลองสถานการณ์ โดยโปรแกรม R Version 3.0.1 และ R Commander Version 1.9-6 กำหนดตัวแปรอิสระเท่ากับ 10 ตัวแปร ภายใต้สถานการณ์ทั้งหมด 24 สถานการณ์ประกอบด้วย สถานการณ์ที่ไม่เกิดและเกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 50, 80, 150 และ 250 รวมถึงขนาดตัวอย่างเท่ากับ 50, 100 และ 500 กระทำซ้ำในแต่ละสถานการณ์จำนวน 100 รอบ ดังแสดงในภาพที่ 11 โดยมีขั้นตอนการวิจัยดังต่อไปนี้

(1) สร้างขนาดประชากร N = 10,000

(2) ตัวแปรอิสระจำนวน 10 ตัวแปร มีการแจกแจงแบบเอกรูป ดังนี้

$X_{i1} \sim U(25,200)$, $X_{i2} \sim U(50,180)$, $X_{i3} \sim U(15,250)$, $X_{i4} \sim U(87,550)$,
$X_{i5} \sim U(1,70)$, $X_{i6} \sim U(35,90)$, $X_{i7} \sim U(3,17)$, $X_{i8} \sim U(8,22)$, $X_{i9} \sim U(15,80)$
และ $X_{i10} \sim U(20,150)$

(3) ความคลาดเคลื่อนมีการแจกแจงปกติ โดยกำหนดให้มี 4 รูปแบบดังนี้

$\varepsilon_i \sim N(0,50^2)$, $\varepsilon_i \sim N(0,80^2)$, $\varepsilon_i \sim N(0,150^2)$ และ $\varepsilon_i \sim N(0,250^2)$

(4) ค่าสัมประสิทธิ์การถดถอยของตัวแปรอิสระ 10 ตัวแปร ดังนี้

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

กำหนดค่าสัมประสิทธิ์การถดถอยโดยพิจารณาจากช่วงของตัวแปรอิสระแต่ละตัวที่ทำการกำหนดขึ้นและจากผลงานวิจัยของ กานต์ณัฐ ณ บางช้าง (2554) ที่ทำการศึกษาเกี่ยวกับการคัดเลือกแปรในตัวแบบการถดถอยเชิงเส้นซึ่งมีลักษณะที่คล้ายกัน ดังนั้นจึงทำการกำหนดค่าสัมประสิทธิ์การถดถอย เมื่อตัวแปรอิสระมี 10 ตัวแปร ตามที่กำหนดข้างต้น

(5) สุ่มขนาดตัวอย่างจำนวน 50, 100 และ 500 จากประชากร N

(6) สร้างตัวแปรตาม จากตัวแบบการถดถอยที่มีตัวแปรอิสระ 10 ตัวแปร ดังต่อไปนี้

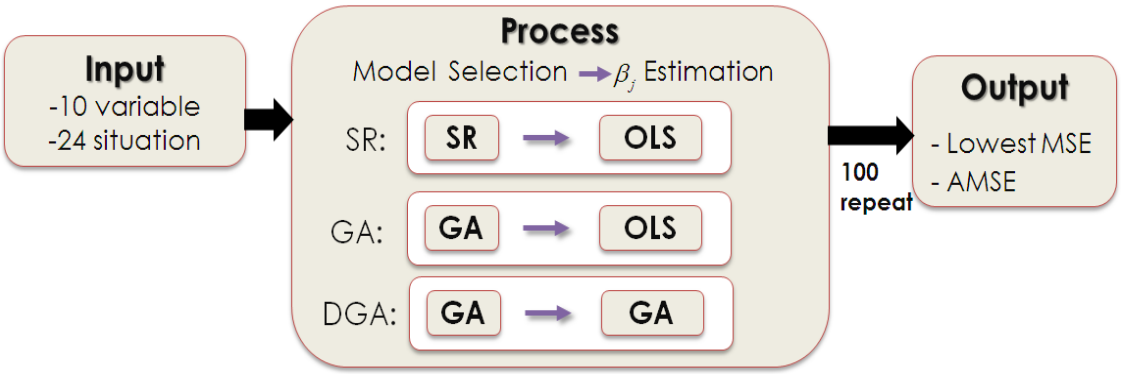
$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + ... + \beta_9 X_{i9} + \beta_{10} X_{i10} + \varepsilon_i$$

โดยที่ y_i แทน ตัวแปรตาม เมื่อ $i = 1,...,n$ และ n แทนขนาดตัวอย่าง
$X_{i1}, X_{i2}, ..., X_{i10}$ แทน ตัวแปรอิสระ
$\beta_0, \beta_1, ..., \beta_{10}$ แทน สัมประสิทธิ์การถดถอย ตามที่กำหนด
ε_i แทน ความคลาดเคลื่อนสุ่ม

(7) ในสถานการณ์ที่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุจะทำการกำหนดความสัมพันธ์ของตัวแปรอิสระ โดยกำหนดให้ X_1, X_3 และ X_2, X_3 มีความสัมพันธ์กัน ในรูปแบบ $X_{i3} = X_{i1} + X_{i2}$

(8) ทำการคัดเลือกตัวแบบการถดถอยด้วยวิธี DGA GA และ SR โดยการทำซ้ำ 100 รอบ

(9) ทำการเปรียบเทียบผลการคัดเลือกตัวแบบการถดถอยของทั้ง 3 วิธี และสรุปผลการวิจัย



ภาพที่ 11 แสดงโครงสร้างการวิจัย

ผลการวิจัย

ผลการวิจัยทำการจำแนกตามสถานการณ์ในการจำลองข้อมูล 24 สถานการณ์ ซึ่งมีการกำหนดตัวแปรอิสระเท่ากับ 10 ตัวแปร สถานการณ์ที่ไม่เกิดและเกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ ส่วนเบี่ยงเบนมาตรฐานของความ

คลาดเคลื่อนเท่ากับ 50, 80, 150 และ 250 รวมถึงขนาดตัวอย่างเท่ากับ 50, 100 และ 500 กระทำซ้ำในแต่ละสถานการณ์จำนวน 100 รอบ โดยผลการวิจัยพบว่า ในสถานการณ์ที่ไม่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ และส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 50 ในแต่ละขนาดตัวอย่าง วิธีการคัดเลือกตัวแบบการถดถอยทั้ง 3 วิธี ให้ค่า MSE ที่ต่ำที่สุดเท่ากัน ในส่วนของค่า AMSE วิธี DGA ให้ค่าที่น้อยกว่าวิธี GA และ SR แต่มีความแปรปรวนที่มากกว่าวิธีอื่น โดยวิธี SR มีความแปรปรวนที่น้อยที่สุดในทุกขนาดตัวอย่าง และสำหรับสถานการณ์ที่ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 80 พบว่าที่ขนาดตัวอย่างเท่ากับ 100 และ 500 วิธีการทั้ง 3 วิธีให้ค่า MSE ที่ต่ำที่สุดเท่ากัน ส่วนขนาดตัวอย่างเท่ากับ 50 วิธี DGA ให้ค่า MSE ที่ต่ำที่สุด น้อยกว่าวิธีอื่น รวมถึงให้ค่า AMSE น้อยกว่าวิธีอื่นด้วย ในขณะที่วิธี SR มีความแปรปรวนน้อยที่สุด รวมถึงสถานการณ์ที่ส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 150 และ 250 ขนาดตัวอย่างเท่ากับ 500 วิธีการคัดเลือกตัวแบบทั้ง 3 วิธีให้ค่า MSE ที่ต่ำที่สุดเท่ากัน ส่วนขนาดตัวอย่างเท่ากับ 100 วิธี DGA ให้ค่าน้อยที่สุด รวมถึงให้ค่า AMSE น้อยกว่าวิธีอื่นในทุกขนาดตัวอย่าง โดยแสดงผลดังตารางที่ 1 และตารางที่ 2

สำหรับสถานการณ์ที่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ และส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 50 เมื่อขนาดตัวอย่างเท่ากับ 50 และ 500 วิธี DGA และ GA ให้ค่า MSE ที่ต่ำที่สุดเท่ากัน รวมถึงวิธี DGA ยังให้ค่า AMSE และความแปรปรวนน้อยกว่าวิธีอื่น ในทุกขนาดตัวอย่าง และเมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเท่ากับ 80 วิธี DGA ให้ค่า MSE ที่ต่ำที่สุด ค่า AMSE และความแปรปรวนน้อยกว่าวิธีอื่น ในทุกขนาดตัวอย่าง ขณะที่วิธี GA และ SR ให้ค่า MSE ที่ต่ำที่สุดเท่ากันที่ขนาดตัวอย่าง 50 และ 500 เมื่อทำการเพิ่มส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนเป็น 150 และ 250 วิธี DGA ให้ค่า MSE ที่ต่ำที่สุด น้อยกว่าวิธี GA และ SR ในขณะที่ทั้ง 2 วิธีนี้ให้ค่าที่เท่ากัน รวมถึงค่า AMSE และความแปรปรวน วิธี DGA ยังคงให้ค่าที่น้อยกว่า ในทุกขนาดตัวอย่าง ดังแสดงผลในตารางที่ 3 และตารางที่ 4

ตารางที่ 1 ผลการคัดเลือกตัวแบบการถดถอยจำแนกตามวิธีการคัดเลือกตัวแบบและตามขนาดของ n เมื่อ $\varepsilon_i \sim N(0,50^2)$, $\varepsilon_i \sim N(0,80^2)$ และไม่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ

		$\varepsilon_i \sim N(0,50^2)$			$\varepsilon_i \sim N(0,80^2)$		
n	วิธีการ	MSE ที่ต่ำที่สุดจากการทำซ้ำ 100 รอบ	AMSE	ความแปรปรวน	MSE ที่ต่ำที่สุดจากการทำซ้ำ 100 รอบ	AMSE	ความแปรปรวน
50	SR	1,316	2,732.62	184,169.08	3,331	6,917.11	2,724,109.43
	GA	1,316	2,720.08	184,732.12	3,331	6,952.09	2,738,497.55
	DGA	1,316	2,550.94	185,044.56	3,117	6,623.20	2,833,902.20
100	SR	1,873	2,501.52	56,189.84	4,275	6,480.58	1,041,945.10
	GA	1,873	2,512.81	56,488.19	4,275	6,561.10	1,179,551.63
	DGA	1,873	2,466.65	56,509.44	4,275	6,379.12	1,269,540.40
500	SR	2,043	2,430.98	13,229.08	5,664	6,354.54	101,286.48
	GA	2,043	2,431.01	13,228.92	5,664	6,367.10	104,109.56
	DGA	2,043	2,422.94	13,253.76	5,664	6,335.58	106,462.66

ตารางที่ 2 ผลการคัดเลือกตัวแบบการถดถอยจำแนกตามวิธีการคัดเลือกตัวแบบและตามขนาดของ n เมื่อกำหนด $\varepsilon_i \sim N(0,150^2)$, $\varepsilon_i \sim N(0,250^2)$ และไม่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ

		$\varepsilon_i \sim N(0,150^2)$			$\varepsilon_i \sim N(0,250^2)$		
n	วิธีการ	MSE ที่ต่ำที่สุดจากการทำซ้ำ 100 รอบ	AMSE	ความแปรปรวน	MSE ที่ต่ำที่สุดจากการทำซ้ำ 100 รอบ	AMSE	ความแปรปรวน
50	SR	12,418	22,970.93	24,090,695.60	34,191	64,814.95	265,264,200.31
	GA	11,927	22,852.20	24,795,157.30	34,191	66,164.50	265,980,849.10
	DGA	11,927	22,702.15	24,545,937.30	34,191	63,638.86	295,043,916.10
100	SR	18,299	23,558.90	8,611,363.79	53,978	62,743.86	84,885,497.08
	GA	18,299	23,944.58	8,602,900.97	52,827	63,252.37	93,971,197.16
	DGA	17,129	23,104.09	9,415,100.80	52,515	62,166.76	92,946,183.98
500	SR	20,203	22,232.29	2,615,703.67	54,808	61,289.62	14,406,714.55
	GA	20,203	22,304.53	2,620,471.73	54,808	61,812.61	14,192,084.20
	DGA	20,203	22,168.97	2,619,299.19	54,808	60,924.36	14,851,963.58

ตารางที่ 3 ผลการคัดเลือกตัวแบบการถดถอยจำแนกตามวิธีการคัดเลือกตัวแบบและตามขนาดของ n เมื่อกำหนด $\varepsilon_i \sim N(0,50^2)$, $\varepsilon_i \sim N(0,80^2)$ และเกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ

		$\varepsilon_i \sim N(0,50^2)$			$\varepsilon_i \sim N(0,80^2)$		
n	วิธีการ	MSE ที่ต่ำที่สุดจากการทำซ้ำ 100 รอบ	AMSE	ความแปรปรวน	MSE ที่ต่ำที่สุดจากการทำซ้ำ 100 รอบ	AMSE	ความแปรปรวน
50	SR	1,713	2,782.91	591,462.61	3,702	7,067.10	3,455,654.16
	GA	1,668	2,712.51	472,601.08	3,702	7,170.72	3,603,663.56
	DGA	1,668	2,682.12	467,331.39	3,554	6,800.40	2,968,835.76
100	SR	1,988	2,586.68	100,781.59	5,263	6,888.88	1,138,692.36
	GA	1,780	2,563.47	100,635.74	4,987	6,768.68	1,068,596.26
	DGA	1,744	2,537.15	100,023.70	4,863	6,684.40	972,102.24
500	SR	2,201	2,495.08	13,537.03	5,607	6,414.96	152,561.06
	GA	2,194	2,493.76	13,548.48	5,607	6,391.62	145,918.68
	DGA	2,194	2,489.46	13,502.94	5,520	6,373.88	146,874.64

ตารางที่ 4 ผลการคัดเลือกตัวแบบการถดถอยจำแนกตามวิธีการคัดเลือกตัวแบบและตามขนาดของ *n* เมื่อกำหนด $\varepsilon_i \sim N(0,150^2)$, $\varepsilon_i \sim N(0,250^2)$ และเกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ

		$\varepsilon_i \sim N(0,150^2)$			$\varepsilon_i \sim N(0,250^2)$		
n	วิธีการ	MSEที่ต่ำที่สุดจากการทำซ้ำ 100 รอบ	AMSE	ความแปรปรวน	MSE ที่ต่ำที่สุดจากการทำซ้ำ 100 รอบ	AMSE	ความแปรปรวน
50	SR	16,194	25,497.23	30,189,481.79	42,040	68,061.99	377,846,831.91
	GA	16,194	25,576.63	28,192,883.94	42,040	67,377.56	365,578,788.63
	DGA	15,410	24,121.20	27,726,952.11	40,909	66,190.49	359,704,682.16
100	SR	15,925	23,043.33	19,174,448.78	46,992	67,579.62	95,116,762.75
	GA	15,925	23,804.13	19,292,288.08	46,992	65,303.79	95,880815.92
	DGA	15,503	22,386.95	18,629,084.97	45,679	64,638.86	91,229,827.85
500	SR	20,283	22,700.05	1,736,134.22	56,328	62,715.98	17,777,781.68
	GA	20,283	22,634.51	1,725,071.29	56,328	62,537.12	18,365,583.93
	DGA	20,103	22,552.31	1,723,757.77	55,437	62,227.59	17,009,459.19

สรุปผลและอภิปรายผลการวิจัย

การเปรียบเทียบค่า MSE ที่ต่ำที่สุด วิธี DGA GA และ SR ในสถานการณ์ที่ไม่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ ทั้ง 3 วิธี ให้ค่า MSE ที่ต่ำที่สุดเท่ากัน สำหรับสถานการณ์ที่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ วิธี DGA ให้ค่า MSE ที่ต่ำที่สุด น้อยกว่าวิธีอื่นๆ และมีบางสถานการณ์ที่วิธี DGA ให้ค่า MSE ที่ต่ำที่สุด เท่ากับวิธี GA เมื่อพิจารณาความแปรปรวน พบว่าในสถานการณ์ที่ไม่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ วิธี SR มีความแปรปรวนที่ต่ำกว่าวิธี DGA และ GA ในทุกขนาดตัวอย่างและส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน แต่สำหรับสถานการณ์ที่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ วิธี DGA มีความแปรปรวนที่ต่ำกว่าวิธี GA และ SR ในทุกขนาดตัวอย่าง และส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน เมื่อทำการเปรียบเทียบค่า AMSE ในสถานการณ์ที่ไม่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ ทั้ง 3 วิธี มีค่า AMSE ใกล้เคียงกันในทุกๆสถานการณ์ที่ทำการศึกษา เมื่อพิจารณาเปรียบเทียบวิธี DGA กับ GA พบว่าวิธี DGA ให้ค่า AMSE ที่ต่ำกว่า เนื่องมาจาก วิธีการ DGA มีกระบวนการในการทำงาน 2 ขั้นตอน ซึ่งเป็นการเพิ่มประสิทธิภาพในการหาค่าสัมประสิทธิ์การถดถอยที่พยายามหาค่าของฟังก์ชันวัตถุประสงค์ ในที่นี้กำหนดเป็นฟังก์ชัน MSE เพื่อให้ได้ค่าที่ต่ำที่สุด

ค่า MSE ที่ต่ำที่สุดของวิธี DGA GA และ SR มีแนวโน้มที่เพิ่มขึ้นเมื่อขนาดตัวอย่างเพิ่มขึ้น ค่า AMSE มีแนวโน้มที่ลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น และมีแนวโน้มที่เพิ่มขึ้นเมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมีค่าสูงขึ้น รวมถึงความแปรปรวนของค่า MSE มีแนวโน้มที่ลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้นทั้งในสถานการณ์ที่ไม่เกิดและเกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ สำหรับวิธี GA และ SR ผลการเปรียบเทียบ ค่า AMSE มีความสอดคล้องกับผลงานวิจัยของ Wasserman และ Sudijanto (1994) ซึ่งอาศัยเกณฑ์การคัดเลือกตัวแบบ โดยใช้ค่า AIC และพิจารณาเปรียบเทียบ

ประสิทธิภาพของทั้ง 2 วิธี โดยใช้ AMSE เช่นเดียวกัน ซึ่งวิธี GA ให้ค่าที่ต่ำกว่าวิธี SR เมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนมีค่าสูง

ดังนั้นการวิจัยครั้งนี้ แสดงให้เห็นว่าวิธี DGA มีประสิทธิภาพในการหาค่า MSE ที่ต่ำที่สุดในขอบเขตที่กำหนดมากกว่าวิธี GA และ วิธี SR สำหรับทุกสถานการณ์ที่ทำการศึกษา และเมื่อกำหนดค่าสัมประสิทธิ์การถดถอยอยู่ในช่วงขอบบนและขอบล่างของตัวแปรอิสระแต่ละตัวที่มีการแจกแจงแบบเอกรูปทั้งค่าที่เป็นลบและบวก ผลการวิจัยที่ได้จะเป็นไปตามการสรุปผลข้างต้น

กิตติกรรมประกาศ

ผู้วิจัยขอขอบคุณ บัณฑิตวิทยาลัย มหาวิทยาลัยเชียงใหม่ ที่ให้การสนับสนุนค่าใช้จ่ายอันเกี่ยวเนื่องจากการทำวิจัยนี้ รวมถึงภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่ ที่อำนวยความสะดวกสำหรับสถานที่ในการวิเคราะห์ข้อมูลของงานวิจัยนี้ให้สำเร็จลุล่วงด้วยดี

เอกสารอ้างอิง

กานต์ณัฐ ณ บางช้าง. (2554). *การคัดเลือกตัวแปรในตัวแบบการถดถอยเชิงเส้นพหุโดยใช้วิธีการค้นหาแบบต้องห้าม*. วิทยานิพนธ์ปริญญามหาบัณฑิต, คณะสถิติประยุกต์, สถาบันบัณฑิตพัฒนบริหารศาสตร์.

เกตุจันทร์ จำปาไชยศรี. (2550). *การวิเคราะห์การถดถอย*, ภาควิชาคณิตศาสตร์ คณะวิทยาศาสตร์, มหาวิทยาลัยนเรศวร.

นันทพร บุญสุข. (2555). *เกณฑ์และสถิติทดสอบในการเลือกตัวแปรในตัวแบบการถดถอยเชิงเส้นแบบพหุกรณีที่ไม่สามารถสร้างตัวแบบเต็มรูป*. วิทยานิพนธ์ปริญญามหาบัณฑิต, คณะสถิติประยุกต์, สถาบันบัณฑิตพัฒนบริหารศาสตร์.

พลากร สีน้อย และจิราวัลย์ จิตรถเวช. (2553). สถิติทดสอบเพื่อคัดเลือกตัวแบบการถดถอยเชิงเส้นพหุคูณ. *วารสารวิทยาศาสตร์บูรพา*, *15*(2), 47-56.

พิษณุ เจียวคุณ. (2548). *การวิเคราะห์การถดถอย*, ภาควิชาสถิติ คณะวิทยาศาสตร์, มหาวิทยาลัยเชียงใหม่.

รุ่งรัตน์ ภิสัชเพ็ญ. (2553). *คู่มือการสร้างแบบจำลองด้วยโปรแกรม Arena (ฉบับปรับปรุง)*, กรุงเทพมหานคร.

วิชิต หล่อจีระชุณห์กุล. (2524). *เทคนิคการพยากรณ์*, กรุงเทพมหานคร, สถาบันบัณฑิตพัฒนบริหารศาสตร์.

Deep, K. and M. Thakur. (2007). A new mutation operator for real coded genetic algorithms. *Applied Mathematics and Computation*, *193*(1), 211-230.

Drezner, Z., Marcoulides,G. A. and Salhi,S. (1999). Tabu Search Model Selection in Multiple Regression Analysis. *Communication in Statistics-Simulation and Computation*, *28*(April), 349-367.

Hasan Örkcü, H. (2013). Subset selection in multiple linear regression models: A hybrid of genetic and simulated annealing algorithms. *Applied Mathematics and Computation*, *219*(23), 11018-11028.

Kapetanious, G. (2007). Variable Selection in Regression Models Usind Nonstandard Optimization of information Criteria. *Computational Statistics & Data Analysis*, 50(September), 4-15.

Luca Scrucca. (2013). GA : A Package for Genetic Algorithm in R. *Journal of Statistical Software*, *53*(4), 1-37.

Pasha G.R. (2002). Selection of Variables in Multiple Regression Using Stepwise Regression. *Journal of Research (Science)*, *13*(December), 119-127.

Renner, G. and A. Ekárt. (2003). Genetic algorithms in computer aided design. *Computer-Aided Design*, *35*(8), 709-726.

Thomas Ng, S., M. Skitmore, et al. (2008). Using genetic algorithms and linear regression analysis for private housing demand forecast. *Building and Environment*, *43*(6), 1171-1184.

Wasserman, G. S. and A. Sudjianto. (1994). All subsets regression using a genetic search algorithm. *Comput. Ind. Eng*, *27*(1-4), 489-492.