Research Article

Received: March 7, 2019; Accepted: September 2, 2019

ประสิทธิภาพของตัวดำเนินการข้ามสายพันธุ์ในขั้นตอนวิธีเชิงพันธุกรรมสำหรับการคัดเลือกตัวแปรในการวิเคราะห์การถดถอย

Performance of Crossover Operators in Genetic Algorithms for Variable Selection in Regression Analysis

พัทธ์ชนก ศรีสุรเดชชัย* และชญานิน อินกว่าง

สาขาวิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี

มหาวิทยาลัยธรรมศาสตร์ ศูนย์รังสิต ตำบลคลองหนึ่ง อำเภอคลองหลวง จังหวัดปทุมธานี 12120

Patchanok Srisuradetchai* and Chayanin Inkwarg

Department of Mathematics and Statistics, Faculty of Science and Technology,

Thammasat University, Rangsit Centre, Khlong Nueng, Khlong Luang, Pathum Thani 12120

บทคัดย่อ

การคัดเลือกตัวแปรในการวิเคราะห์การถดถอยเป็นขั้นตอนที่ท้าท้ายเมื่อตัวแปรอิสระมีจำนวนมากและคาดว่าตัวแปรบางคู่มีปฏิสัมพันธ์ (interaction) กันเพราะจำนวนตัวแบบทั้งหมดที่เป็นไปได้จะมีจำนวนมาก วิธีคัดเลือกตัวแปรแบบขั้นตอน (stepwise selection) ได้รับความนิยมและมีแนวโน้มให้ค่าดีที่สุดเฉพาะที่ (local optima) งานวิจัยนี้จึงใช้ขั้นตอนวิธีเชิงพันธุกรรม (genetic algorithm) กับตัวดำเนินการข้ามสายพันธุ์ทั้ง 6 วิธี กล่าวคือ การข้ามสายพันธุ์แบบจุดเดียว (1-PC) การข้ามสายพันธุ์แบบ 2 จุด (2-PC) การข้ามสายพันธุ์แบบ m/2 จุด [(m/2)-PC] การข้ามสายพันธุ์แบบ m-1 จุด [(m-1)-PC] การข้ามสายพันธุ์แบบเอกรูป (UNC) และการข้ามสายพันธุ์แบบสับเปลี่ยน (SHC) โดยใช้ข้อมูลจริงและข้อมูลที่ได้จากการจำลองจากทั้งตัวแบบการถดถอยเชิงเส้นและการถดถอยลอจิสติกทวิภาค โดยใช้เกณฑ์สารสนเทศของอะกะอิเกะ (Akaike's information criterion, AIC) ในการคัดเลือกตัวแปรอิสระ สำหรับข้อมูลที่จำลองจะสมมติให้ตัวแปรอิสระไม่มีความสัมพันธ์กันและมีความสัมพันธ์แบบ AR1 (first-order autoregressive) ที่มีสัมประสิทธิ์สหสัมพันธ์ 0.3, 0.5 และ 0.8 ผลลัพธ์ที่ได้นำมาเปรียบเทียบกับวิธีคัดเลือกตัวแปรแบบการเลือกแบบไปข้างหน้า การกำจัดแบบถอยหลัง และการกำจัดแบบสองทิศทาง ผลการศึกษาพบว่าขั้นตอนเชิงพันธุกรรมสามารถหาตัวแบบที่มีค่า AIC ต่ำกว่าวิธีคัดเลือกตัวแปรแบบขั้นตอน และเมื่อเปรียบเทียบตัวดำเนินการข้ามสายพันธุ์ทั้ง 6 พบว่า (m-1)-PC จะให้ตัวแบบที่มีความเหมาะสมน้อยกว่าตัวดำเนินการข้ามสายพันธุ์แบบอื่น ๆ อย่างมีนัยสำคัญทางสถิติ โดยทั่วไปตัวดำเนินการข้ามสายพันธุ์ 1-PC ให้ AIC โดยเฉลี่ยต่ำที่สุด รองลงมาเป็น 2-PC, (m/2)-PC, SHC และ UNC นอกจากนี้ยังพบว่า SHC และ UNC ไม่มีความแตกต่างอย่างมีนัยสำคัญในทุกกรณีที่ศึกษา

คำสำคัญ : ตัวดำเนินการข้ามสายพันธุ์; การถดถอยลอจิสติกทวิภาค; เกณฑ์สารสนเทศของอะกะอิเกะ; การคัดเลือกตัวแปร

***ผู้รับผิดชอบบทความ :** patchanok@mathstat.sci.tu.ac.th **doi:** 10.14456/tstj.2020.121

Abstract

Variable selection is a challenging procedure when there is a large number of explanatory variables and interaction effects are expected in a model. Because the number of possible models can be enormous, the stepwise selection, the most commonly used procedure, tends to give a local optimal model. This paper aims to apply the genetic algorithm with 6 types of crossover operators: single-point crossover (1-PC), 2-point crossover (2-PC), (m/2)-point crossover [(m/2)-PC], (m-1)-point crossover [(m-1)-PC], shuffle crossover (SHC), and uniform crossover (UNC) for real datasets and simulated data. Both linear and binomial logistic regressions are of interest and the Akaike's information criterion (AIC) is used for variable selections. For simulated data, the explanatory variables are set to have no correlation and to have correlations with the first-order autoregressive structure in which correlations equal to 0.3, 0.5, and 0.8. The results were compared with those from stepwise selections: forward selection, backward elimination, and bidirectional elimination. It is found that the genetic algorithm can find the model with a lower AIC than that by the stepwise selection. When the 6 crossover operators were compared, (m-1)-PC produced an inferior model which significantly different from other crossover operators. Generally, 1-PC gave a model with the lowest AIC, followed by 2-PC, (m/2)-PC, SHC and UNC. In addition, SHC and UNC were shown not to be statistically different.

Keywords: crossover operator; binomial logistic regression; Akaike's information criterion; variable selection

1. บทนำ

การวิเคราะห์การถดถอยนำไปใช้อย่างแพร่หลายในสาขาวิชาต่าง ๆ โดยมีวัตถุประสงค์หลักเพื่อประมาณสมการถดถอยระหว่างตัวแปรตามกับตัวแปรอิสระอย่างน้อย 1 ตัว ตัวแบบที่นิยมใช้ในการวิเคราะห์ข้อมูลที่มีตัวแปรตามเป็นตัวแปรสุ่มแบบต่อเนื่อง (continuous random variable) คือ ตัวแบบการถดถอยเชิงเส้น (linear regression model) [1] ซึ่งเขียนอยู่ในรูปเมทริกซ์ ดังนี้

$$\mathbf{Y}=\boldsymbol{\beta}\mathbf{X}+\boldsymbol{\varepsilon} \quad (1)$$

โดยที่ $\mathbf{Y}$ คือ เวกเตอร์ตัวแปรตามขนาด $n\times 1$; $\mathbf{X}$ คือ เมทริกซ์ของตัวแปรอิสระขนาด $n\times(p+1)$ $\boldsymbol{\beta}$ คือ เวกเตอร์ของพารามิเตอร์ที่ไม่ทราบค่าจริงหรือเวกเตอร์สัมประสิทธิ์การถดถอยขนาด $(p+1)\times 1$; $\boldsymbol{\varepsilon}$ คือ เวกเตอร์ของความคลาดเคลื่อนขนาด $n\times 1$ โดย $\varepsilon_i \sim N(0,\sigma^2)$; n คือ ขนาดตัวอย่าง ; p คือ จำนวนตัวแปรอิสระ

ตัวแบบที่นิยมในการวิเคราะห์ข้อมูลที่มีตัวแปรตามแบบจำแนกประเภทหรือเชิงกลุ่มที่แบ่งได้ 2 กลุ่ม (dichotomous response or binary) คือ การถดถอยลอจิสติกทวิภาค ให้ Y_i เป็นตัวแปรตามทวิภาคที่เป็นอิสระต่อกันของค่าสังเกตที่ $i, i=1,2,...,n$ เมื่อ $Y_i=1$ แทน การเกิดเหตุการณ์ที่สนใจและ $Y_i=0$ เมื่อไม่เกิดเหตุการณ์ที่สนใจ ดังนั้น Y_i มีการแจกแจง

แบร์นูลลี ในแต่ละค่าสังเกตอาจมีความน่าจะเป็นของพารามิเตอร์ที่เป็นเหตุการณ์ที่สนใจ π_i ที่แตกต่างกัน ดังนั้นฟังก์ชันการแจกแจงความน่าจะเป็นของ Y_i เขียนได้ในรูป $P(Y_i = y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$ และรูปแบบที่ง่ายที่สุดของฟังก์ชัน $\pi_i = \pi(x_{i1},...,x_{ip})$ คือฟังก์ชันเชิงเส้น $\pi_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$ แต่ฟังก์ชันดังกล่าวนี้อาจให้ค่า π_i นอกช่วง 0 ถึง 1 [2] ดังนั้นสมการที่ไม่เป็นเชิงเส้นที่ให้ π_i มีค่าระหว่าง 0 กับ 1 ที่นิยม คือ ตัวแบบการถดถอยลอจิสติก (logistic regression model) กล่าวคือ

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})} \quad (2)$$

โดยที่ $\exp(\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip})$ จะมีค่าเป็นบวกเสมอและตัวเศษของตัวแบบการถดถอยลอจิสติกจะมีค่าน้อยกว่าตัวส่วนส่งผลให้ π_i กล่าวคือ อยู่ในช่วง 0 ถึง 1 ตัวแบบใน (2) ยังสามารถเขียนได้ในอีกรูปได้ดังนี้

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} \quad (3)$$

เพื่อหาตัวแบบที่มีความเหมาะสมในการอนุมานและการพยากรณ์ตัวแปรตาม โดยการทบทวนวรรณกรรมมีผู้นำเสนอวิธีต่าง ๆ ในการคัดเลือกตัวแบบการถดถอย ได้แก่ การคัดเลือกแบบขั้นตอน (stepwise selection, SS) และขั้นตอนวิธีการค้นหาแบบสโตแคสติก (Stochastic search algorithm) เป็นต้น การเลือกตัวแบบถดถอยที่มีความเหมาะสมอาจพิจารณาว่าเป็นปัญหาการหาค่าที่เหมาะสมที่สุด (optimization problem) เนื่องจากการเลือกตัวแบบก็คือ การเลือกชุดของตัวแปรอิสระที่ทำให้ค่าของเกณฑ์ที่สนใจของตัวแบบมีค่าสูงสุด เช่น เกณฑ์สารสนเทศของอะกะอิเกะ (Akaike's information criterion, AIC) วิธีการถดถอยแบบขั้นตอน (stepwise regression) เป็นวิธีพื้นฐานที่นิยมที่แต่ละขั้นตอนจะสร้างและประเมินตัวแบบการถดถอยที่แตกต่างจากตัวแบบการถดถอยที่ดีที่สุดในปัจจุบันเพียงหนึ่งตัวแปร ถ้าหากมีตัวแบบการถดถอยใหม่มีค่าของเกณฑ์ที่สนใจสูงกว่าของตัวแบบที่ดีที่สุดในปัจจุบัน ตัวแบบใหม่นี้ก็จะเป็นตัวแบบตั้งต้นแล้วทำซ้ำไปเรื่อย ๆ จนไม่มีตัวแบบที่ดีกว่าตัวแบบในปัจจุบัน แล้วขั้นตอนวิธีจึงจะสิ้นสุดลง วิธีนี้เป็นกระบวนการหาค่าที่ดีที่สุดเฉพาะที่ (local search process) ซึ่งก็เป็นข้อเสียที่สำคัญของกระบวนการนี้ อย่างไรก็ตาม วิธี SS พบได้ในโปรแกรมทางสถิติทั่วไป ได้แก่ SPSS, Minitab, SAS และ R เป็นต้น

วิธีหนึ่งที่สามารถหลีกเลี่ยงค่าเหมาะสมที่สุดเฉพาะที่ คือ ใช้ขั้นตอนวิธีเชิงพันธุกรรม (genetic algorithm, GA) ซึ่งเป็นวิธีการที่เลียนแบบกระบวนการวิวัฒนาการทางธรรมชาติ กล่าวคือ โดยหลักของการถ่ายทอดพันธุกรรมแล้ว ในการอยู่รอดตามธรรมชาติ โครโมโซมที่มีลักษณะด้อยกว่าจะมีโอกาสที่จะสูญหายไปมากกว่าโครโมโซมที่มีลักษณะเด่นกว่า และมีแนวโน้มว่าโครโมโซมรุ่นต่อ ๆ ไปจะมีลักษณะที่ดีกว่ารุ่นบรรพบุรุษ [3] ซึ่ง GA จะค้นหาปริภูมิค้นหา (search space) ที่เป็นไปได้พร้อมกันทั้งหมด จึงสามารถหลีกเลี่ยงปัญหาการได้คำตอบที่ดีที่สุดเฉพาะที่ Paterlini และ Minerva [4] ศึกษาเปรียบเทียบประสิทธิภาพขั้นตอนวิธีเชิงพันธุกรรม 2 วิธี คือ genetic algorithm for regressors selection (GARS) และ genetic algorithm for regressors selection and transformation (GARST) แล้วเปรียบเทียบกับการคัดเลือกแบบขั้นตอน (SS) โดยใช้ข้อมูลปริมาณไขมัน (fat measurement) และข้อมูลระดับคลอเรสเตอรอล (cholesterol measurement) พิจารณาจากเกณฑ์ AIC เกณฑ์สารสนเทศของเบส์ (Bayesian information criterion, BIC) และ SIC (Schwarz information criterion) ในการเลือกตัวแบบที่มีความเหมาะสม พบว่าผลลัพธ์ของ GARS จะมี

ประสิทธิภาพดีกว่าในทุกเกณฑ์ที่พิจารณาเมื่อเทียบกับวิธี SS นอกจากนี้ Vinterbo และ Ohno-Machado [5] ใช้ข้อมูลตัวอย่างของกลุ่มโรคกล้ามเนื้อหัวใจตายเฉียบพลันในปี ค.ศ. 1999 เพื่อศึกษาการคัดเลือกตัวแปรอิสระในการถดถอยลอจิสติกโดยใช้เกณฑ์พื้นที่ใต้เส้นโค้งลักษณะเฉพาะดำเนินการตัวรับ (receiver operating characteristic curve, ROC) และใช้ GA ค้นหากลุ่มตัวแปรอิสระที่ดีที่สุดและใช้ตัวแบบที่ได้เปรียบเทียบกับตัวแบบที่ถูกสร้างด้วยวิธี SS ผลการศึกษาพบว่าตัวแบบที่สร้างด้วย GA ดีกว่าตัวแบบจากวิธี SS อย่างมีนัยสำคัญทางสถิติ Johnson และคณะ [6] ศึกษาปัจจัยที่ส่งผลต่อการลุกลามของโรคอัลไซเมอร์ (Alzheimer's disease) โดยใช้ SS ผลการศึกษาพบว่า SS จะเน้นเพียงปัจจัยเดียวมากกว่าหลายปัจจัยพร้อม ๆ กัน ซึ่งกรณีนี้ GA จะมีประสิทธิภาพสำหรับการค้นหาการรวมกันของตัวแปรอิสระเพื่อให้ได้ตัวแบบที่เหมาะสมมากกว่า

การคัดเลือกตัวแปรอิสระด้วย GA ผู้วิเคราะห์จะต้องกำหนดตัวดำเนินการข้ามสายพันธุ์ (crossover operator) ซึ่งมีบทบาทสำคัญในการเลียนแบบการรวม 2 โครโมโซมในทางชีววิทยา กล่าวคือ เป็นการแลกเปลี่ยนเพื่อให้เกิดความหลากหลาย โดยโครโมโซมจะเทียบเท่ากับตัวแปรอิสระในการวิเคราะห์การถดถอย ปัจจุบันมีตัวดำเนินการข้ามสายพันธุ์หลัก ๆ ที่แตกต่างกันที่สามารถใช้ใน GA ที่เข้ารหัสแบบทวิภาค ซึ่งแบ่งเป็น (1) การข้ามสายพันธุ์แบบจุดเดียว (2) การข้ามสายพันธุ์แบบ k จุด (3) การข้ามสายพันธุ์แบบสับเปลี่ยน และ (4) การข้ามสายพันธุ์แบบเอกรูป โดย Picek และ Golub [7] ศึกษาเปรียบเทียบตัวดำเนินการข้ามสายพันธุ์ที่แตกต่างกันใน GA ที่เข้ารหัสแบบทวิภาค โดยใช้ฟังก์ชันทดสอบทางคณิตศาสตร์ ซึ่งแสดงรูปทางเรขาคณิตที่มีความยากหลายระดับเพื่อพิจารณาประสิทธิภาพของตัวดำเนินการข้ามสายพันธุ์ ผลการศึกษาพบว่าการข้ามสายพันธุ์แบบเอกรูปและการข้ามสายพันธุ์แบบ 2 จุด ค่อนข้างมีประสิทธิภาพสูงเมื่อเทียบกับการข้ามสายพันธุ์แบบจุดเดียวเพราะสามารถค้นหาปริภูมิของปัญหา (problem space) ได้ทั่วถึงมากกว่า เนื่องจากมีจำนวนจุดสลับเปลี่ยนมากกว่า

การทบทวนวรรณกรรมที่เกี่ยวข้องยังไม่พบการศึกษาประสิทธิภาพของตัวดำเนินการข้ามสายพันธุ์ที่แตกต่างกันใน GA ว่ามีผลต่อการหาตัวแบบในการวิเคราะห์การถดถอยหรือไม่ งานวิจัยนี้จึงศึกษาการคัดเลือกตัวแปรอิสระโดยใช้ GA ที่ใช้ตัวดำเนินการข้ามสายพันธุ์แตกต่างกัน 6 วิธี พร้อมทั้งเปรียบเทียบกับวิธี SS โดยใช้ AIC เป็นเกณฑ์ในการเลือกตัวแบบ และเมื่อได้ตัวแบบที่เหมาะสมแล้วจะพิจารณาเปอร์เซ็นต์ความถูกต้องของจำนวนตัวแปรอิสระที่เข้าสู่ตัวแบบ สำหรับการข้ามสายพันธุ์แบบ k จุดจะขยายจากกรณี 1 และ 2 จุด เป็นจำนวนตัวแปรอิสระหารด้วย 2 และจำนวนตัวแปรอิสระลบออกด้วย 1 (ค่าสูงสุดที่เป็นไปได้ของ k) โดยใช้ทั้งข้อมูลจริงและข้อมูลจำลองที่มาจากตัวแบบการถดถอยเชิงเส้นและการถดถอยลอจิสติกทวิภาค

2. ทฤษฎีที่เกี่ยวข้อง

2.1 ขั้นตอนเชิงพันธุกรรม

ขั้นตอนวิธีเชิงพันธุกรรมหรือ GA เป็นวิธีการค้นหาแบบฮิวริสติกที่ใช้กับปัญหาการหาค่าที่เหมาะสมที่สุด คิดค้นโดย Holland [8] ซึ่งเลียนแบบกระบวนการการคัดเลือกโดยธรรมชาติของดาร์วิน คำตอบที่เป็นไปได้ (candidate solution) ของปัญหาการหาค่าสูงสุดจะแสดงโดยรหัสพันธุกรรม (genetic code) และมีความเหมาะสม (fitness) เปรียบเทียบได้กับคุณลักษณะของคำตอบ ในกระบวนการคัดเลือกโดยธรรมชาตินั้นจะมีแนวคิดว่าการผสมพันธุ์ระหว่างสิ่งที่มีความเหมาะสมสูงจะมีโอกาสในการถ่ายทอด

คุณลักษณะที่ต้องการให้กับลูกหลานในรุ่นถัดไป (future generation) มากกว่าการผสมพันธุ์ระหว่างสิ่งมีชีวิตที่มีความเหมาะสมต่ำ นอกจากนี้การกลายพันธุ์ในระดับยีน (genetic mutation) ซึ่งเกิดขึ้นได้ยากจะทำให้เกิดความหลากหลายของประชากร [9] โดย GA มีขั้นตอนย่อยที่สำคัญ ดังนี้

ขั้นที่ 1 การคัดเลือก (selection mechanism) เป็นกระบวนการที่เลือกพ่อแม่ให้ผลิตโครโมโซมลูกหลาน ซึ่งโดยทั่วไปจะเลือกโครโมโซมพ่อแม่ที่มีความเหมาะสม (fitness) สูงและเลือกพ่อแม่อีกโครโมโซมอย่างสุ่ม การเลือกแบบนี้จะทำให้โครโมโซมที่มีความเหมาะสมสูงมีโอกาสถูกเลือกเป็นโครโมโซมพ่อแม่มากกว่า และเรียกวิธีการคัดเลือกแบบนี้ว่าวิธีคัดเลือกแบบจัดอันดับ (rank-based method)

ขั้นที่ 2 กำหนดให้มีการข้ามสายพันธุ์เพื่อให้ได้คำตอบที่เหมาะสมโดยเร็ว พื้นฐานสำคัญในการจำแนกประเภทของตัวดำเนินการข้ามสายพันธุ์ คือ ประเภทการเข้ารหัสของโครโมโซม ซึ่งส่วนใหญ่ประเภทของตัวดำเนินการข้ามสายพันธุ์ขึ้นอยู่กับการแสดงปัญหาในรูปแบบโครโมโซม การเลือกตัวดำเนินการข้ามสายพันธุ์มีผลกระทบต่อประสิทธิภาพของ GA โดยตัวดำเนินการข้ามสายพันธุ์แบ่งเป็น

2.1.1 การข้ามสายพันธุ์แบบจุดเดียว (single point crossover, 1-PC) เป็นการข้ามสายพันธุ์แรกเริ่มและนิยมใช้มากที่สุด [8,10] มีการเลือกจุดข้ามสายพันธุ์เพียงจุดเดียวบนทั้งสองโครโมโซมพ่อแม่ โดยการเลือกเลขสุ่มระหว่าง (1,m-1) โดยที่ m เป็นความยาวของโครโมโซมหรือจำนวนตัวแปรอิสระ โครโมโซมพ่อแม่ทั้งสองจะถูกแบ่งที่จุดข้ามสายพันธุ์ที่เลือกไว้และสับเปลี่ยนข้อมูลทั้งหมดที่อยู่หลังจุดข้ามสายพันธุ์ (crossover point) เพื่อให้ได้โครโมโซมลูกหลานหรือคำตอบ [11] ซึ่งสามารถแสดงดังตัวอย่างในรูปที่ 1 และ 2

2.1.2 การข้ามสายพันธุ์แบบ k จุด (k-point crossover, k-PC) นำเสนอครั้งแรกโดย de Jong [10] โดยมีหลักเกณฑ์คล้ายกับ 1-PC แต่แตกต่างกันตรงจำนวนของจุดข้ามสายพันธุ์โดยค่าของ k มีค่าตั้งแต่ 1 ถึง m-1 การสับเปลี่ยนส่วนประกอบทางพันธุกรรมของพ่อแม่ทั้งสองจะเกิดด้านซ้ายของจุดการข้ามสายพันธุ์โดยสลับช่วงเว้นช่วงดังแสดงตัวอย่างในรูปที่ 3 และ 4 การทำเช่นนี้จะทำให้ได้โครโมโซมลูกหลานที่มีความเหมาะสมมากขึ้น [12] แต่การทบทวนวรรณกรรมยังไม่มีการศึกษาที่ชัดเจนว่าค่า k ที่เหมาะสมควรจะมีค่าเท่าใด ทราบแต่เพียงว่ายิ่ง k มีค่ามากก็จะสามารถค้นหาปริภูมิคำตอบได้ทั่วถึงมากขึ้น [7] ซึ่งงานวิจัยนี้จะให้ k เท่ากับ 1, 2 m/2 และ m-1 ทั้งนี้จะเห็นว่า 1-PC เป็นกรณีพิเศษของการข้ามสายพันธุ์แบบ k จุดเมื่อ k เท่ากับ 1

Parent 1:	1	0	0	1	1	1	0	0	1
Parent 2:	1	1	0	0	1	1	1	1	0

Figure 1 Points between 4th and 5th genes are selected as the crossover points

Offspring 1:	1	0	0	1	1	1	1	1	0
Offspring 2:	1	1	0	0	1	1	0	0	1

Figure 2 Two offspring are created by swapping the parent genes

Parent 1:	1	0	0	1	1	1	0	0	1
Parent 2:	1	1	0	0	1	1	1	1	0

Figure 3 Points between 2nd and 3rd, 4th and 5th and 7th and 8th genes are selected as crossover points

Offspring 1:	1	0	0	0	1	1	0	1	0
Offspring 2:	1	1	0	1	1	1	1	0	1

Figure 4 The resulting two offspring created by swapping the parent genes

2.1.3 การข้ามสายพันธุ์แบบสับเปลี่ยน (Shuffle crossover, SHC) พิจารณาโครโมโซมพ่อแม่ที่มีจุดการข้ามสายพันธุ์อย่างเป็นอิสระกัน เริ่มต้นด้วยสับเปลี่ยนคู่ยีนในโครโมโซมพ่อแม่ทั้งสองอย่างสุ่ม จากนั้นใช้วิธีการข้ามสายพันธุ์แบบจุดเดียวโดยการสุ่มเลือกจุดข้ามสายพันธุ์และรวมโครโมโซมพ่อแม่ทั้งสองเพื่อสร้างโครโมโซมลูกหลาน หลังจากทำ 1-PC แล้วยีนในโครโมโซมลูกหลานจะถูกสับเปลี่ยนกลับสู่ตำแหน่งเดิมก่อนการสับเปลี่ยนยีนตอนเริ่มต้น [13] ดังแสดงในรูปที่ 5, 6 และ 7 ตามลำดับ

Select the pairs of points for shuffle

	1	2	**3**	**4**	5	6	**7**	8	9
Parent 1:	1	0	0	1	1	1	0	0	1
Parent 2:	1	1	0	0	1	1	1	1	0

Shuffle selected pairs of points

	4	2	**7**	**1**	5	6	**3**	8	9
Parent 1:	1	0	0	1	1	1	0	0	1
Parent 2:	0	1	1	1	1	1	0	1	0

Figure 5 Random shuffle the pairs of genes in the both parents

	4	2	7	1	5	6	3	8	9
Parent 1:	1	0	0	1	1	1	0	0	1
Parent 2:	0	1	1	1	1	1	0	1	0
	4	2	7	1	5	6	3	8	9
Offspring 1:	1	0	0	1	1	1	0	1	0
Offspring 2:	0	1	1	1	1	1	0	0	1

Figure 6 Applying the 1-PC in each pair of genes (after the crossover point)

	1	2	3	4	5	6	7	8	9
Offspring 1:	1	0	0	1	1	1	0	1	0
Offspring 2:	1	1	0	0	1	1	1	0	1

Figure 7 The resulting chromosomes after reordering

2.1.4 การข้ามสายพันธุ์แบบเอกรูป (uniform crossover, UNC) จะตรงข้ามกับตัวดำเนินการข้ามสายพันธุ์ 3 วิธีก่อนหน้า UNC จะไม่แบ่งโครโมโซมพ่อแม่ออกเป็นส่วนเพื่อรวมแต่ละส่วนเข้าด้วยกันอีกครั้ง แต่ในการเลือกลูกหลาน วิธีการนี้จะปฏิบัติต่อแต่ละยีนของโครโมโซมอย่างเป็นอิสระกัน จำนวนของจุดไม่ได้กำหนดตั้งแต่เริ่มต้นแต่จะกำหนดความน่าจะเป็นที่จะสับเปลี่ยนยีน (probability of swapping, p_s) หน้ากากข้ามสายพันธุ์ (crossover mask) เป็นค่าจำนวนจริงที่สุ่มจากการแจกแจงเอกรูป แต่ละยีนของโครโมโซมลูกหลานจะสร้างโดยคัดลอกยีนจากหนึ่งในสองโครโมโซมพ่อแม่ โดยถ้าค่าหน้ากากข้ามสายพันธุ์มีค่าน้อยกว่าหรือเท่ากับ p_s จะสับเปลี่ยนยีนในโครโมโซมพ่อแม่ตำแหน่งเดียวกันและจะไม่สับเปลี่ยนยีนในโครโมโซมพ่อแม่ถ้ามีค่ามากกว่า p_s [12] ดังแสดงในรูปที่ 8 และ 9

Parent 1:	1	0	0	1	1	1	0	0	1
Crossover mask	0.9	0.1	0.2	0.8	0.1	0.7	0.8	0.1	0.3
Parent 2:	1	1	0	0	1	1	1	1	0

Figure 8 Generating the crossover masks from a uniform distribution

Offspring 1:	1	1	0	1	1	1	0	1	0
Offspring 2:	1	0	0	0	1	1	1	0	1

Figure 9 Resulting two offspring chromosomes when $p_s = 0.5$

ขั้นที่ 3 การกลายพันธุ์ เป็นกระบวนการเปลี่ยนส่วนประกอบของโครโมโซมภายหลังการข้ามสายพันธุ์ ทำให้เกิดความหลากหลายของคำตอบโดยเปลี่ยนโครโมโซมเล็กน้อยอย่างสุ่มเพื่อให้มีการผันแปรอย่างต่อเนื่อง จึงช่วยหลีกเลี่ยงคำตอบที่มากที่สุดเฉพาะที่ [14]

2.2 เกณฑ์สารสนเทศของอะกะอิเกะ

เกณฑ์สารสนเทศของอะกะอิเกะหรือ AIC นำเสนอโดย Hirotugu ซึ่งมีสูตรเป็น AIC = $-2\log\hat{L}+2p$ โดย $\hat{L}$ คือ ค่าสูงสุดของฟังก์ชันภาวะน่าจะเป็นและ p คือ จำนวนของพารามิเตอร์ในตัวแบบ โดยยิ่งค่า AIC ต่ำแสดงว่าตัวแบบดังกล่าวจะมีความเหมาะสมมากกว่า เมื่อเปรียบเทียบเกณฑ์ AIC กับเกณฑ์ BIC ซึ่งพัฒนามาจากเกณฑ์ AIC โดย Sawa ในปี ค.ศ. 1978 พบว่าหากพิจารณาตามแนวคิดของเบส์แล้ว AIC และ BIC จะมีความน่าจะเป็นแรกเริ่ม (prior distribution) ที่ต่างกันโดย AIC จะมีความเหมาะสมมากกว่า BIC [15] Vrieze [16] ศึกษา

เปรียบเทียบเกณฑ์ทั้งสองโดยการจำลอง แล้วพบว่า AIC จะเลือกตัวแบบได้ดีกว่า BIC ภายใต้หลายสถานการณ์ นอกจากนี้ Yang [17] พบว่าเกณฑ์ AIC มีความเหมาะสมที่สุดเชิงกำกับ (asymptotically optimal) ในการคัดเลือกตัวแบบโดยใช้ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อนน้อยที่สุด (least mean squared error) เป็นเกณฑ์ภายใต้ข้อสมมติที่ว่าตัวแบบที่แท้จริง (true model) ไม่ได้อยู่ในเซตของตัวแบบที่ใช้ในการคัดเลือก (candidate set) ในขณะที่เกณฑ์ BIC ไม่ได้มีความเหมาะสมที่สุดเชิงกำกับตามเกณฑ์นี้

3. วิธีการวิจัย

การศึกษาครั้งนี้ใช้ทั้งข้อมูลจริงและข้อมูลที่จำลอง ซึ่งการคำนวณทั้งหมดจะกระทำโดยใช้โปรแกรม R เวอร์ชั่น 3.5.1 ตัวแบบที่ใช้ในการวิเคราะห์ คือ การถดถอยเชิงเส้นและการถดถอยลอจิสติกทวิภาค โดยตัวแบบเต็ม (full model) จะเป็นตัวแบบที่มีอิทธิพลหลักและอิทธิพลร่วม (interaction effect) เช่น หากมีตัวแปรอิสระ x_1, x_2 และ x_3 จะได้ตัวแบบเต็มเป็น $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \varepsilon$

3.1 ข้อมูลจริง

3.1.1 ข้อมูลไขมัน (body fat) ในร่างกายในปี ค.ศ. 1996 โดย Johnson [18] เป็นข้อมูลเปอร์เซ็นต์ไขมันในร่างกายที่วัดจากผู้ชาย 252 คน ตัวแปรอิสระเป็นตัวเลข (numeric) ทั้งหมด 13 ตัวแปร เช่น อายุ น้ำหนัก ส่วนสูงและค่าสัดส่วนร่างกายอื่น ๆ ตัวแปรตาม คือ เปอร์เซ็นต์ของไขมันในร่างกาย สำหรับจำนวนตัวแบบทั้งหมดที่เป็นไปได้เท่ากับ $2^{13+{}^{13}C_2} = 2.476 \times 10^{27}$ ตัวแบบ

3.1.2 ข้อมูลคุณภาพของไวน์แดง (red wine) ในปี ค.ศ. 2009 โดย Cortez และคณะ [19] เป็นข้อมูลคะแนนคุณภาพของไวน์แดงโปรตุเกส ประกอบด้วยค่าสังเกต 1,599 ค่า และตัวแปรอิสระที่เป็นตัวเลขทั้งหมด 11 ตัวแปร ซึ่งได้จากการทดสอบทางเคมีกายภาพ ได้แก่ ปริมาณแอลกอฮอล์ ค่าความเป็นกรดด่าง เป็นต้น ตัวแปรตาม คือ คะแนนเฉลี่ยจากการประเมินของผู้เชี่ยวชาญด้านไวน์อย่างน้อย 3 คน สำหรับจำนวนตัวแบบทั้งหมดที่เป็นไปได้เท่ากับ $2^{11+{}^{11}C_2} = 7.379 \times 10^{19}$ ตัวแบบ

3.1.3 ข้อมูลราคาบ้าน (house price) ในปี ค.ศ. 2011 โดย Cock [20] เป็นข้อมูลการขายที่อยู่อาศัยในเมืองเอมส์ รัฐไอโอวา ตั้งแต่ปี ค.ศ. 2006 ถึง 2010 โดยมีบ้านทั้งหมด 1,460 หลัง และตัวแปรอิสระเป็นตัวเลขที่เกี่ยวข้องกับขนาดพื้นที่ 20 ตัว ได้แก่ พื้นที่รอบบ้านทั้งหมด (ตารางฟุต) พื้นที่ชั้นที่ 1 ของบ้าน (ตารางฟุต) เป็นต้น ตัวแปรตาม คือ ราคาขายบ้าน (sale price) หน่วยดอลลาร์สหรัฐ สำหรับจำนวนตัวแบบทั้งหมดที่เป็นไปได้เท่ากับ $2^{20+{}^{20}C_2} = 1.646 \times 10^{63}$ ตัวแบบ

3.1.4 ข้อมูลการเป็นเบาหวานของชนเผ่าอินเดียน (Pima Indian diabetes) ในปี ค.ศ. 1990 จากสถาบันแห่งหนึ่ง [21] โดยเก็บข้อมูลจากชาวพื้นเมืองชนเผ่าอินเดียนเพศหญิงอายุอย่างน้อย 21 ปี จำนวน 786 คน แบ่งเป็นผู้ที่เป็นเบาหวาน 268 คน และไม่เป็นเบาหวาน 500 คน ตัวแปรอิสระเป็นตัวเลขทั้งหมด 8 ตัว เช่น อายุ BMI ตัวแปรตาม คือ สภาวะเป็นหรือไม่เป็นเบาหวาน สำหรับจำนวนตัวแบบทั้งหมดที่เป็นไปได้เท่ากับ $2^{8+{}^{8}C_2} = 68,719,476,736$ ตัวแบบ

3.2 ข้อมูลจำลอง

ข้อมูลจำลองกำหนดให้มีตัวแปรอิสระ 10 ตัว ที่มีสัมประสิทธิ์การถดถอย (β) ของตัวแปรอิสระเป็น $\beta_0 = 2, \beta_1 = 0.001, \beta_2 = 0.01, \beta_3 = 0.5, \beta_4 = 0.75, \beta_5 = 1.5, \beta_6 = 2, \beta_{1,2} = -2.5, \beta_{1,3} = -1.5, \beta_{1,9} = -1,$

$\beta_{1,10}=0.1, \beta_{1,2}=-2.5, \beta_{1,3}=-1.5, \beta_{1,9}=-1,$ $\beta_{1,10}=0.1, \beta_{2,3}=-2.5, \beta_{2,4}=-1, \beta_{2,9}=-1.5,$ $\beta_{2,10}=2, \beta_{3,4}=-0.8, \beta_{9,10}=3$ และสัมประสิทธิ์การถดถอยที่เหลือกำหนดให้เป็น 0 ในงานวิจัยนี้กำหนดให้ตัวแปรอิสระทั้งหมดเป็นอิสระต่อกัน และเพื่อจำลองให้ใกล้เคียงกับสถานการณ์ในชีวิตจริงมากที่สุด กำหนดให้ตัวแปรอิสระมีความสัมพันธ์โดยมีสัมประสิทธิ์สหสัมพันธ์ (ρ) เท่ากับ 0.3, 0.5, 0.8 และมีโครงสร้างความสัมพันธ์ของตัวแปรอิสระเป็นแบบอัตสัมพันธ์อันดับที่ 1 (first-order autoregressive, AR1) เนื่องจากเป็นโครงสร้างที่ค่อนข้างง่ายไม่ซับซ้อนและแสดงถึงสถานการณ์ที่ตัวแปรอิสระ 2 ตัวแปรใด ๆ ที่มีความสัมพันธ์กับตัวแปรตามสูงก็มักจะมีความสัมพันธ์กันเองสูง ส่วนตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามน้อยก็มักมีความสัมพันธ์ในระดับต่ำกับตัวแปรอิสระที่สำคัญต่อตัวแปรตาม [22] โครงสร้างความสัมพันธ์แบบ AR1 สำหรับตัวแปรอิสระจำนวน 10 ตัว เป็นดังนี้

$$\Sigma=\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^9 \\ \rho & 1 & \rho & \rho^2 & \cdots & \rho^8 \\ \rho^2 & \rho & 1 & \rho & \cdots & \rho^7 \\ \rho^3 & \rho^2 & \rho & 1 & \cdots & \rho^6 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^9 & \rho^8 & \rho^7 & \rho^6 & \cdots & 1 \end{bmatrix} \quad (4)$$

ในการจำลอง ตัวแปรอิสระทั้ง 10 ตัวจะมีการแจกแจงปรกติที่มีค่าเฉลี่ย $\mu_1=-2,$ $\mu_2=1, \mu_3=5, \mu_4=0, \mu_5=5, \mu_6=-2, \mu_7=1,$ $\mu_8=0, \mu_9=5, \mu_{10}=0$ และมีเมทริกซ์ความแปรปรวน-ความแปรปรวนร่วมเท่ากับ Σ ใน (4) สำหรับการสร้างตัวแปรสุ่ม $(X_1, X_2, ..., X_{10})$ นี้จะใช้ฟังก์ชัน mvrnorm ในไลบรารี MASS หลังจากได้ $(x_1, x_2, ..., x_{10})$ จะนำไปคำนวณ $\mu^*=\beta_0+\sum_{i=1}^{10}\beta_i x_i + \sum_{j=i+1}^{10}\sum_{i=1}^{9}\beta_{ij}x_i x_j$ หากเป็นตัวแบบการถดถอยเชิงเส้นจะสร้าง y จาก $N(\mu^*, \sigma^2=1)$ และหากเป็นตัวแบบการถดถอยลอจิสติกทวิภาคจะสร้าง y จาก Bernoulli (π) โดยที่ $\pi=e^{\mu^*}/\left(1+e^{\mu^*}\right)$ ในงานวิจัยนี้กำหนดขนาดตัวอย่างเท่ากับ 300 และ 1,000 ทั้งนี้ผู้วิจัยได้ทดลองตัวอย่างขนาด 500 พบว่าให้ผลใกล้เคียงกับขนาด 300 และตัวอย่างขนาด 2,000 จะให้ผลใกล้เคียงกับ 1,000 สำหรับตัวอย่างขนาดเล็กกว่า 300 เช่น 100 มีแนวโน้มที่จะเกิดปัญหาความไม่เสถียรในการประมาณค่า ในที่นี้กำหนดตัวแปรอิสระ 10 ตัว ดังนั้นตัวแบบทั้งหมดที่เป็นไปได้จะเท่ากับ $2^{10+{}^{10}C_2}=3.603\times10^{16}$ ตัวแบบ

3.3 การคัดเลือกตัวแปร

การคัดเลือกตัวแปรอิสระแบบ SS ใช้ฟังก์ชัน step ในไลบรารี stats ในการคัดเลือกตัวแบบโดยใช้ค่า AIC เป็นเกณฑ์ โดยพิจารณา 3 วิธี คือ การเลือกแบบไปข้างหน้า (forward selection, FWD) การกำจัดแบบถอยหลัง (backward elimination, BWD) และการกำจัดแบบสองทาง (bidirectional elimination, BDN) ซึ่งเป็นวิธีที่รวม FWD และ BWD เข้าไว้ด้วยกัน [2]

การคัดเลือกตัวแปรอิสระแบบ GA ผู้วิจัยได้เขียนโปรแกรม R ขึ้นมาเองโดยได้กำหนดจำนวนรุ่น (generation, itr) เท่ากับ 100 และ 150 ขนาดของแต่ละรุ่น (P) เท่ากับ 2× จำนวนตัวแปรอิสระในชุดข้อมูล (m) อัตราการกลายพันธุ์ (m.rate) เท่ากับ 0.01 จำนวนการทำซ้ำ (round) เท่ากับ 50,000 วิธีการคัดเลือกที่ใช้ คือ วิธีคัดเลือกแบบจัดอันดับ ตัวดำเนินการข้ามสายพันธุ์ 6 วิธี คือ 1-PC, 2-PC, (m/2)-PC, (m-1)-PC, SHC และ UNC ในที่นี้ตัวดำเนินการกลายพันธุ์ที่ใช้เป็นวิธีการกลายพันธุ์แบบพลิกกลับ

3.4 เกณฑ์คัดเลือกและประเมินตัวแบบ

นอกเหนือจากเกณฑ์สารสนเทศของอะกะ-อิเกะคำนวณได้จาก $AIC=-2\log\hat{L}+2k$ แล้ว งาน

วิจัยยังได้พิจารณาเปอร์เซ็นต์ความถูกต้อง (correctness percentage criteria, CPC) ซึ่งคือ เปอร์เซ็นต์ของจำนวนตัวแปรอิสระที่ถูกคัดเลือกเข้าสู่ตัวแบบได้ถูกต้องจากจำนวนตัวแปรทั้งหมดที่อยู่ในตัวแบบที่คัดเลือกได้และจำนวนตัวแปรในตัวแบบที่แท้จริง โดยมีหลักการคำนวณดังนี้ กำหนดให้ A แทน เซตของตัวแปรทั้งหมดที่ถูกคัดเลือกเข้าสู่ตัวแบบหรือตัวแปรที่มี $\hat{\beta} \neq 0$ และ B แทนเซตของตัวแปรที่มีความสัมพันธ์ต่อตัวแปรตามหรือตัวแปรที่มี $\beta \neq 0$ แล้ว CPC จะคำนวณจาก $n(A \cap B)/n(A \cup B) \times 100\%$ หรือพื้นที่แรเงาในรูปที่ 10(A) หากสมมติว่า มีตัวแปรอิสระทั้งหมด 12 ตัวให้เป็น $x_1, x_2, x_3, \ldots, x_{12}$ โดยตัวแปรอิสระที่มี $\beta \neq 0$ มีเพียง 6 ตัว ซึ่งกำหนดให้เป็นเซต $B = \{x_1, x_3, x_5, x_7, x_9, x_{11}\}$ และมีตัวแปรทั้งหมดที่ถูกคัดเลือกเข้าสู่ตัวแบบเป็น $A = \{x_1, x_2, x_3, x_4, x_5, x_7, x_9, x_{11}\}$ แล้ว CPC จะเท่ากับ $n(A \cap B)/n(A \cup B) \times 100\% = 6/8 \times 100\% = 75\%$ จะเห็นว่าเซต B เป็นสับเซตของ A ซึ่งแสดงดังรูปที่ 10(B) และหากสมมติว่ามี $A = B$ ดังรูปที่ 10(C) แล้ว CPC จะเท่ากับ 100% เกณฑ์เปอร์เซ็นต์ความถูกต้องนี้จะคำนึงถึงขนาดของตัวแบบ กล่าวคือ หากตัวแบบมีตัวแปรอิสระจำนวนมาก ตัวหาร $n(A \cup B)$ ก็จะมีค่ามาก และหากตัวแบบมีตัวแปรอิสระน้อยตัวหาร $n(A \cup B)$ ก็จะมีค่าน้อย สำหรับตัวเศษ $n(A \cap B)$ จะนับเฉพาะจำนวนตัวแปรอิสระที่ $\beta \neq 0$ และถูกคัดเลือกเข้าสู่ตัวแบบ ซึ่งถ้าตัวแบบที่คัดเลือกใกล้เคียงกับตัวแบบที่แท้จริง $n(A \cup B)$ จะใกล้เคียงกับ $n(A \cap B)$ ส่งผลให้ CPC มีค่าใกล้ 100 %

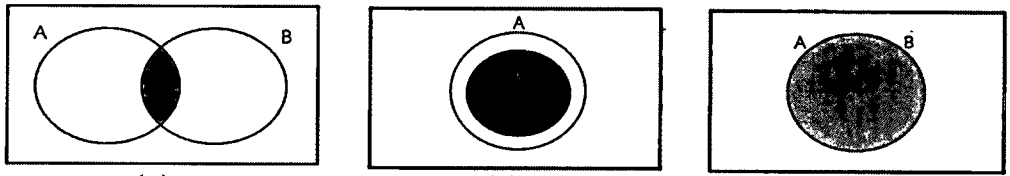

Figure 10 Area of intersection between two circles

4. ผลการวิจัย

ผลการวิจัยจะแบ่งออกเป็น 2 ส่วน คือ ผลการศึกษาจากการเปรียบเทียบ GA กับ SS และการเปรียบเทียบประสิทธิภาพของตัวดำเนินการข้ามสายพันธุ์ทั้ง 6 วิธี ในข้อมูลจริงและข้อมูลจำลอง

4.1 การเปรียบเทียบวิธีการคัดเลือกตัวแปรอิสระ

การเปรียบเทียบวิธีการคัดเลือกตัวแปรในกรณีข้อมูลจริงที่ใช้การถดถอยเชิงเส้นและการถดถอยลอจิสติกทวิภาค ซึ่งทำซ้ำ GA มากถึง 100 รอบ แต่ละรอบมีจำนวนรุ่น 100 รุ่น โดยตัวแบบที่ดีที่สุดใน GA จะพิจารณาค่า AIC ในแต่ละรุ่นจนมีแนวโน้มคงที่และค่า AIC ที่นำมาเป็นผลลัพธ์จะเป็นค่าที่ต่ำสุดดังแสดงตัวอย่างในรูปที่ 11 (ค่า AIC ที่ต่ำที่สุดอยู่ในวงกลม) ผลการศึกษาการคัดเลือกตัวแบบจากข้อมูลจริงทั้ง 4 ชุด พบว่า GA เลือกตัวแบบที่มีค่า AIC ต่ำกว่า SS ในทั้ง 3 วิธี (FWD BWD และ BDN) โดยผลการศึกษาสรุปไว้ในตารางที่ 1 ดังนั้น GA มีความสามารถในการหาตัวแบบที่มีความเหมาะสมมากกว่า SS ในทุกข้อมูลจริงที่ศึกษา ทั้งนี้ไม่สามารถหา CPC ได้เนื่องจากไม่ทราบตัวแบบที่แท้จริง

สำหรับข้อมูลจำลอง วิธี GA จะทำซ้ำ 50 รอบ แต่ละรอบมีจำนวนรุ่น 150 รุ่น โดยจะเลือกตัวแบบที่มีค่า AIC ต่ำที่สุด ผลการศึกษาสรุปแสดงดังใน

ตารางที่ 2 จะเห็นว่า GA ที่ใช้ตัวดำเนินการข้ามสายพันธุ์ทั้ง 6 วิธี สามารถเลือกตัวแบบที่มีค่า AIC ต่ำกว่า SS ในทั้ง 3 วิธี ในกรณีที่ $\rho = 0$ หรือตัวแปรอิสระเป็นอิสระต่อกัน การถดถอยเชิงเส้นที่ใช้วิธี GA ทุกตัวดำเนินการข้ามสายพันธุ์ให้ค่า AIC ต่ำที่สุดและเท่ากัน พร้อมทั้งยังให้ CPC สูงสุดเท่า SS ซึ่งให้ค่า AIC มากกว่า จึงกล่าวได้ว่าหากตัวแปรอิสระไม่มีความสัมพันธ์กันวิธี GA ให้ตัวแบบที่มี AIC ต่ำสุดและคัดเลือกตัวแปรอิสระได้ถูกต้องสูงสุด ในกรณีที่ตัวแปรอิสระเริ่มมีความสัมพันธ์ถึงแม้ว่า GA จะให้ค่า AIC ต่ำที่สุดแต่ CPC จะน้อยกว่าวิธีที่ดีที่สุดใน SS เช่น กรณีที่ $\rho = 0.8$ และ $n = 1,000$ ทั้งในการถดถอยเชิงเส้นและการถดถอยลอจิสติกทวิภาค พบว่า CPC สูงสุดมีค่าเท่ากับ 38.46 % (BWD) และ 25.81 % (BDN, FWD) ตามลำดับ ขณะที่วิธีที่ดีที่สุด (AIC ต่ำที่สุด) ของ GA ให้ CPC เท่ากับ 26.92 % และ 16.90 % ตามลำดับ การเพิ่มขนาดตัวอย่างจาก 300 เป็น 1,000 ผลสรุปด้านวิธีการคัดเลือกไม่ได้เปลี่ยนแปลง กล่าวคือ GA ยังคงให้ตัวแบบที่มี AIC ต่ำสุดแต่ CPC จะน้อยกว่า SS อย่างไรก็ตาม เมื่อเปรียบเทียบ CPC จะพบว่าที่ขนาดตัวอย่าง 1,000 ค่า CPC สูงกว่าขนาดตัวอย่าง 300 เสมอในทุกค่า ρ และทุกวิธีการคัดเลือก เช่น กรณี $\rho = 0.3$ วิธี BDN ในการถดถอยเชิงเส้นที่ n = 300 ให้ค่า CPC เท่ากับ 28.23 % ในขณะที่ n = 1,000 ค่า CPC มีค่าเพิ่มขึ้นเป็น 35.71 %

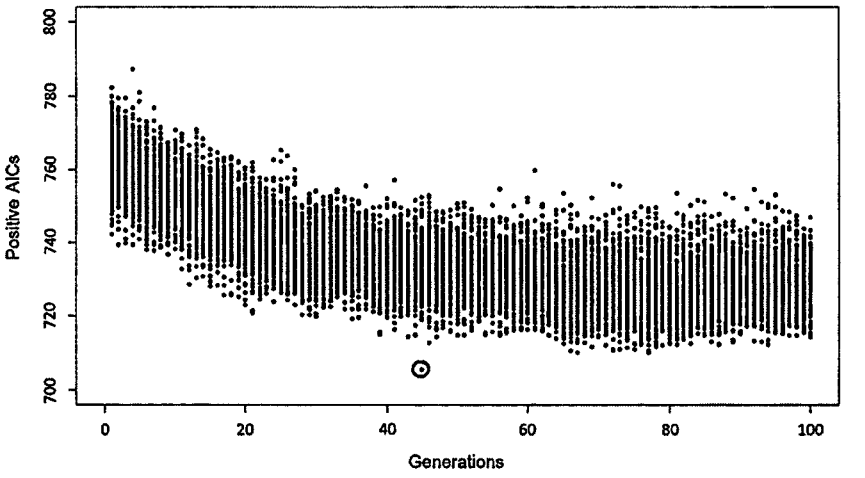

Figure 11 The result of the genetic algorithm with 100 generations for body fat dataset

4.2 การเปรียบเทียบประสิทธิภาพของตัวดำเนินการข้ามสายพันธุ์

การเปรียบเทียบความแตกต่างของประสิทธิภาพของตัวดำเนินการข้ามสายพันธุ์ที่แตกต่างกัน 6 วิธี จะใช้การทดสอบไม่อิงพารามิเตอร์ด้วยวิธี Kruskal-Wallis (H statistic) และการทดสอบรายคู่ (multiple comparisons) ด้วยการทดสอบดัน (Dunn's test) โดยประสิทธิภาพที่กล่าวถึงนี้จะพิจารณาจากค่า AIC ที่ได้จากการทำซ้ำ 50 รอบในทุก ๆ วิธีและในแต่ละสถานการณ์

กรณีข้อมูลจริง ผลของการทดสอบความแตกต่างของค่า AIC ที่ได้จากตัวดำเนินการข้ามสายพันธุ์ทั้ง 6 วิธี สรุปไว้ในตารางที่ 3 ซึ่งพบว่าค่าเฉลี่ยของลำดับ (mean ranks) ของ AIC อย่างน้อย 1 คู่

Table 2 The lowest AICs and CPC values obtained from 9 methods for the simulated datasets

ρ	Methods	Sample sizes							
		n = 300				n = 1,000			
		Linear regression		Logistic regression		Linear regression		Logistic regression	
		AIC	CPC	AIC	CPC	AIC	CPC	AIC	CPC
0	FWD	1823.83	17.86 %	250.28		6083.43	***52.17 %***	927.78	20.69 %
	BWD	1823.99	***28.57 %***	251.52		6080.39	43.48 %	925.07	***31.03 %***
	BDN	1822.33	17.86 %	250.28		6080.08	47.83 %	927.78	20.69 %
	1-PC	**1820.61**	***28.57 %***	**249.70**	***19.23 %***	**6079.80**	***52.17 %***	**922.42**	***31.03 %***
	2-PC								
	(m/2)-PC								
	(m-1)-PC								
	SHC								
	UNC								
0.3	FWD	1826.46		224.71	6.25 %	6082.50	35.71 %	793.07	***29.63 %***
	BWD	1827.85	***28.13 %***	222.34	***13.62 %***	6075.06	***42.86 %***	786.14	14.81 %
	BDN	1826.46		224.71	6.25 %	6082.50	35.71 %	791.14	18.52 %
	1-PC	**1823.08**	25.00 %	**218.96**	12.50 %	**6073.24**	35.71 %	**786.05**	14.81 %
	2-PC								
	(m/2)-PC								
	(m-1)-PC								
	SHC								
	UNC								
0.5	FWD	1824.70	23.33 %	261.47	8.11 %	6076.48	27.59 %	817.05	16.67 %
	BWD	1821.35	***26.67 %***	258.11	***21.62 %***	6079.80	***37.93 %***	805.34	***23.33 %***
	BDN	1824.70	23.33 %	261.47	8.11 %	6076.48	27.59 %	817.05	16.67 %
	1-PC	**1819.06**	20.00 %	**254.49**	10.81 %	**6073.32**	34.48 %	**805.13**	***23.33 %***
	2-PC			256.50	10.81 %				
	(m/2)-PC	1819.56	20.00 %	256.50	8.11 %				
	(m-1)-PC	1820.30	23.33 %	256.52	10.81 %	*6073.44*	34.48 %	805.26	
	SHC	**1819.06**	20.00 %	**254.49**	10.81 %	**6073.32**	34.48 %	**805.13**	
	UNC								
0.8	FWD	1829.57	11.11 %	238.16	12.50 %	5982.65	23.08 %	697.72	***25.81 %***
	BWD	1836.41	***25.00 %***	238.33	***18.13 %***	5975.20	***38.46 %***	704.61	19.35 %
	BDN	1829.57	11.11 %	237.17	9.38 %	5982.65	23.08 %	697.72	***25.81 %***
	1-PC	**1823.93**	16.67 %	230.69	12.50 %	**5973.09**	26.92 %	**693.55**	16.90 %
	2-PC			**230.52**	15.63 %				
	(m/2)-PC			230.62	15.63 %				
	(m-1)-PC	1824.80	19.44 %	230.89	12.50 %			694.84	16.93 %
	SHC	**1823.93**	16.67 %	230.69	15.63 %			**693.55**	16.90 %
	UNC								

The bold numbers represent the lowest AICs and bold-italic numbers represent the highest CPCs.

Table 1 The lowest AICs obtained from 9 methods for the real datasets

Methods	Body fat	Red wine quality	House prices	Diabetes
FWD	723.800	-1461.090	30464.020	728.380
BWD	708.100	-1484.080	30461.210	705.650
BDN	722.330	-1465.680	30457.820	726.660
1-PC	**705.330**	-1485.717	**30456.300**	**705.483**
2-PC	707.899	-1486.117	30458.700	
(m/2)-PC	707.176	-1486.626	30461.070	
(m-1)-PC	710.306	-1485.696	30462.530	
SHC	708.186	-1484.623	30461.300	
UNC	707.749	**-1486.991**	30459.100	

The bold numbers correspond the lowest AICs.

Table 3 Comparisons of the rank means of AICs obtained from 9 methods for the real datasets

Models	Datasets	H statistic	p-values
Linear regression	Body fat	108.350	< 0.0001
	Red wine quality	25.993	< 0.0001
	House price	71.716	< 0.0001
Binary logistic regression	Diabetes	42.864	< 0.0001

Table 4 Comparisons of the rank means of AICs obtained from 9 methods for the simulated datasets

Correlation coefficients	Linear regression		Binary logistic regression	
	H statistic	p-values	H statistic	p-values
$\rho = 0$	78.980	< 0.0001	62.481	< 0.0001
$\rho = 0.3$	89.269	< 0.0001	101.953	< 0.0001
$\rho = 0.5$	45.659	< 0.0001	28.786	< 0.0001
$\rho = 0.8$	34.792	< 0.0001	24.212	< 0.0010

แตกต่างกันเนื่องจากค่าพีน้อยกว่า 0.0001 และในข้อมูลจำลองของทั้งการถดถอยเชิงเส้นและการถดถอยลอจิสติกทวิภาคทุกระดับความสัมพันธ์พบว่า ค่าเฉลี่ยของลำดับของ AIC อย่างน้อย 1 คู่ แตกต่างกันโดยผล

การทดสอบสรุปไว้ในตารางที่ 4 เมื่อพิจารณาผลลัพธ์ของการทดสอบรายคู่ต่อไปด้วยการทดสอบดันซึ่งสรุปดังในตารางที่ 5 พบว่า (m-1)-PC มีประสิทธิภาพต่ำกว่าตัวดำเนินการข้ามสายพันธุ์อีก 5 วิธี ในทุกข้อมูลที่ศึกษา เช่น ในข้อมูลจำลองจากตัวแบบการถดถอยเชิงเส้นที่ $\rho = 0$ วิธี (m-1)-PC ซึ่งมีตัวอักษร d นี้แตกต่างจากวิธี 2-PC, 1-PC, SHC, UNC และ (m/2)-PC ซึ่งได้ตัวอักษรยก a a,b a,b,c b,c และ c ตามลำดับ จะเห็นว่า 2-PC, 1-PC และ SHC มีค่าเฉลี่ยของลำดับไม่แตกต่างอย่างมีนัยสำคัญทางสถิติเนื่องจากมีตัวอักษร a เหมือนกันและเมื่อพิจารณาตัวอย่างกราฟค่า AIC 50 ค่าของข้อมูลดังแสดงในรูปที่ 12 พบว่าเส้นกราฟค่า AIC ของ 54-PC หรือ (m-1)-PC ส่วนใหญ่จะแกว่งเหนือค่า AIC ของการข้ามสายพันธุ์แบบอื่น ๆ ในรูปที่ 13 แสดงค่า AIC ที่ได้จากทุกตัวดำเนินการข้ามสายพันธุ์ซึ่งเส้น 104-PC มีแนวโน้มที่แกว่งเหนือตัวดำเนินการข้ามสายพันธุ์อื่น ๆ นอกจากนี้ยังพบว่าในทุกข้อมูลที่ศึกษา (m-1)-PC จะให้ค่า AIC โดยเฉลี่ยสูงสุด รองลงมาโดยส่วนใหญ่จะเป็น UNC, SHC, (m/2)-PC, 2-PC และ 1-PC ตามลำดับ จะสังเกตว่า 1-PC มีแนวโน้มที่คัดเลือกตัวแบบที่มีค่า AIC ต่ำที่สุดแต่ไม่แตกต่างจาก 2-PC อย่างมีนัยสำคัญทางสถิติ และยังพบว่า UNC และ SHC ไม่แตกต่างกันอย่างมีนัยสำคัญทางสถิติในทุกกรณีที่ศึกษา ทั้งนี้ขนาดตัวอย่าง 300 และ 1,000 ให้ผลการสรุปไม่แตกต่างกัน กล่าวคือ ค่าพีมีค่าน้อยมากซึ่งเป็นไปในทำนองเดียวกันกับตารางที่ 3, 4 และ 5

Table 5 Multiple comparisons of the AIC means obtained from 6 crossover operators by Dunn's test

Datasets		Crossover operators in ascending order of AIC					
Linear regression	$\rho = 0$	2-PC^{a}	1-PC^{a,b}	SHC^{a,b,c}	UNC^{b,c}	(m/2)-PC^{c}	(m-1)-PC^{d}
	$\rho = 0.3$	1-PC^{a}	2-PC^{b}	(m/2)-PC^{b}	SHC^{b}	UNC^{b}	(m-1)-PC^{c}
	$\rho = 0.5$	1-PC^{a}	SHC^{a,b}	2-PC^{a,b}	(m/2)-PC^{a,b}	UNC^{b}	(m-1)-PC^{c}
	$\rho = 0.8$	1-PC^{a}	2-PC^{a}	(m/2)-PC^{a}	SHC^{a}	UNC^{a}	(m-1)-PC^{b}
	Body fat	1-PC^{a}	2-PC^{b}	(m/2)-PC^{b}	SHC^{b}	UNC^{b}	(m-1)-PC^{c}
	House price	1-PC^{a}	2-PC^{a}	(m/2)-PC^{a,b}	SHC^{b}	UNC^{b}	(m-1)-PC^{c}
	Red wine	1-PC^{a}	2-PC^{a}	(m/2)-PC^{a}	SHC^{a}	UNC^{a}	(m-1)-PC^{b}
Binary logistic regression	$\rho = 0$	1-PC^{a}	2-PC^{a}	SHC^{a,b}	UNC^{a,b}	(m/2)-PC^{b}	(m-1)-PC^{c}
	$\rho = 0.3$	1-PC^{a}	2-PC^{a,b}	(m/2)-PC^{b,c}	SHC^{b,c}	UNC^{c}	(m-1)-PC^{d}
	$\rho = 0.5$	1-PC^{a}	2-PC^{a}	(m/2)-PC^{a}	SHC^{a}	UNC^{a}	(m-1)-PC^{b}
	$\rho = 0.8$	2-PC^{a}	1-PC^{a,b}	(m/2)-PC^{a,b}	SHC^{a,b}	UNC^{b}	(m-1)-PC^{c}
	Diabetes	1-PC^{a}	2-PC^{a,b}	SHC^{a,b}	(m/2)-PC^{b}	UNC^{b}	(m-1)-PC^{c}

The methods with different superscript letters are significantly different ($p < 0.10$).

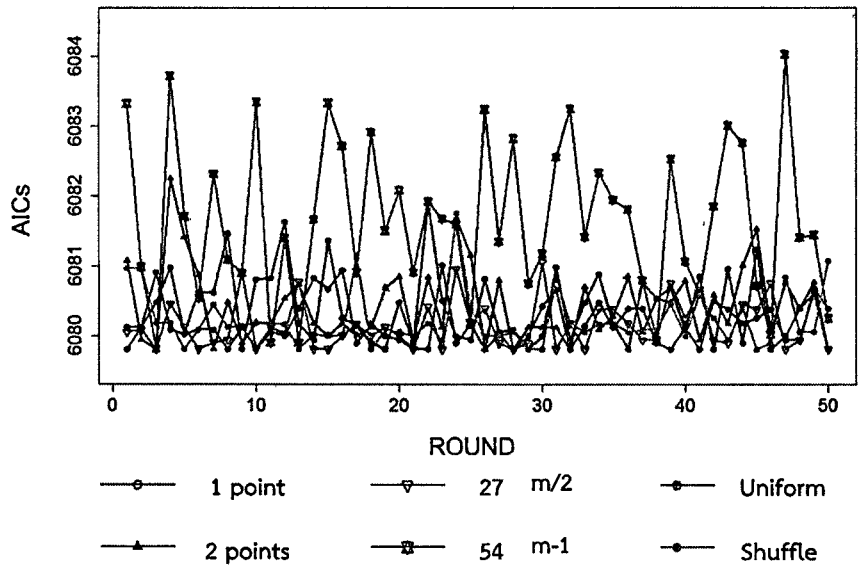

Figure 12 AIC values for the simulated data ($\rho = 0$) with 50 replications

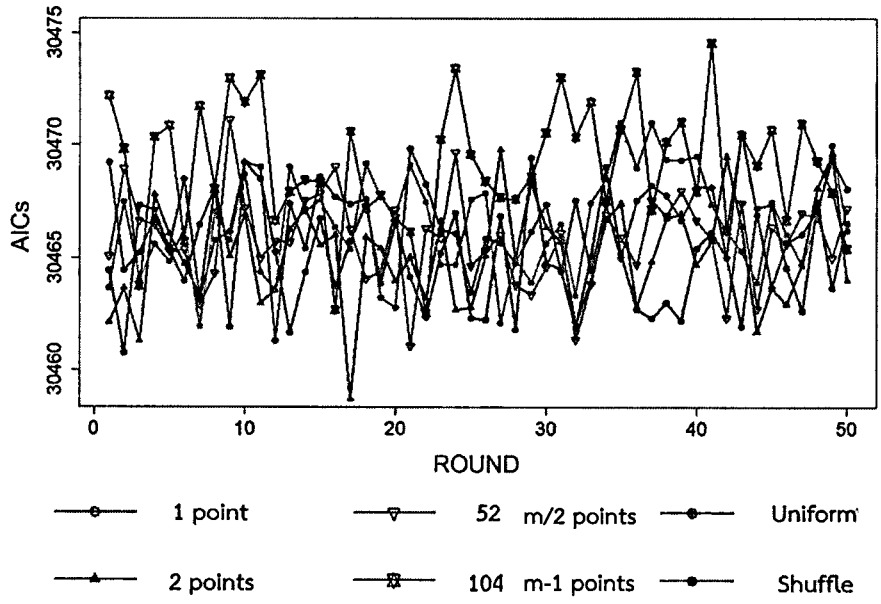

Figure 13 AIC values for the house price data with 50 replications

5. สรุปผลและอภิปราย

การเปรียบเทียบ SS กับ GA โดยใช้ทั้งข้อมูลจริงและข้อมูลจำลองทั้งในการถดถอยเชิงเส้นและการถดถอยลอจิสติกทวิภาค พบว่าในทุกกรณีที่ศึกษา GA สามารถหาตัวแบบที่มีค่า AIC ต่ำกว่า SS และเมื่อพิจารณาตัวดำเนินการข้ามสายพันธุ์ทั้ง 6 วิธี พบว่า (m-1)-PC จะมีประสิทธิภาพต่ำที่สุด (ให้ AIC โดยเฉลี่ยสูงสุด) เมื่อเปรียบเทียบกับตัวดำเนินการข้ามสายพันธุ์อื่น ๆ อย่างมีนัยสำคัญทางสถิติ แต่โดยทั่วไป 1-PC จะให้ AIC โดยเฉลี่ยต่ำที่สุดรองลงมาเป็น 2-PC,

(m/2)-PC, SHC และ UNC อย่างไรก็ตาม ตัวดำเนินการข้ามสายพันธุ์ทั้ง 5 นี้ [ยกเว้น (m-1)-PC] อาจไม่แตกต่างกันโดยจะขึ้นอยู่กับข้อมูลที่ศึกษา จึงสรุปได้ว่าหากเพิ่มจุดที่ใช้สับเปลี่ยนโครโมโซมสูงสุดเท่ากับ m-1 กลับให้ตัวแบบที่ไม่เหมาะสมแต่หากเป็น 1, 2 และ m/2 จุดยังคงให้ผลที่ดีและไม่แตกต่างกันอย่างชัดเจน เมื่อพิจารณาเกณฑ์ CPC พบว่า SS ซึ่งประกอบด้วย FWD, BWD และ BDN ถึงแม้จะเลือกตัวแบบที่มีค่า AIC สูงกว่าแต่มีแนวโน้มที่จะให้ค่า CPC สูงกว่า จึงเป็นข้อพึงสังเกตว่าถึงแม้จะได้ตัวแบบที่มีค่า AIC ต่ำที่สุดแต่ก็อาจไม่ใช้ตัวแบบที่ง่ายและมีตัวแปรอิสระที่สมควรจะอยู่ในตัวแบบมากที่สุด อย่างไรก็ตาม ในสถานการณ์จริง ผู้วิเคราะห์จะไม่ทราบตัวแบบที่แท้จริง การคัดเลือกตัวแบบที่มี AIC ต่ำที่สุดยังคงเป็นที่นิยม ทั้งนี้หากตัวแบบสุดท้าย (ที่มีค่า AIC ต่ำที่สุด) ที่ได้มีค่าพีจากการทดสอบสมมติฐาน $\beta = 0$ ของทุกตัวแปรที่อยู่ในตัวแบบน้อยกว่าระดับนัยสำคัญที่กำหนดก็จะทำให้ผู้วิเคราะห์มีความมั่นใจมากขึ้นว่าตัวแบบที่คัดเลือกได้น่าจะประมาณตัวแบบที่แท้จริงได้อย่างดี งานวิจัยนี้จึงแนะนำให้ใช้ GA ร่วมกับ 1-PC หรือ 2-PC และหากตัวแปรอิสระไม่มีความพันธ์กัน ($\rho = 0$) ผู้วิเคราะห์สามารถมั่นใจได้ว่าตัวแบบที่ได้นั้นมีค่า AIC ต่ำที่สุดและ CPC สูงที่สุด

6. References

[1] Agresti, A., 2015, Foundations of Linear and Generalized Linear Models, John Wiley & Sons, Inc., New Jersey.

[2] Bilder, C. R. and Loughin, T. M., 2015, Analysis of Categorical Data with R, CRC Press, Inc., Boca Raton.

[3] Simon, D., 2013, Evolutionary Optimization Algorithms, John Wiley & Sons, Inc., New Jersey.

[4] Paterlini, S. and Minerva, T., 2010, Regression Model Selection Using Genetic Algorithm, Available Source: http://www.wseas.us/e-library/conferences/2010/lasi/NNECFS/ NNECFS- 0 1 . pdf, October 19, 2018.

[5] Vinterbo, S. and Ohno-Machado, L., 1999, A genetic algorithm to select variable in logistic regression: Example in the domain of myocardial infarction, Proc. AMIA Symp. 1999: 984-988.

[6] Johnson, P., Vendewater, L., Wilson, W., Maruff, P., Savage, G., Graham, P. and Macaulay, L. S., 2014, Genetic algorithm with logistic regression for prediction of progression to Alzheimer' s disease, BMC Bioinfomatics 15(16): 1-14.

[7] Picek, S. and Golub, M., 2010, Comparison of a crossover operator in binary-coded genetic algorithms, WSEAS Transact. Comput. 9: 1064-1073.

[8] Holland, J.H., 1975, Adaptation in Natural and Artificial Systems, In Hoeting, J.A. and Givens, G.H., Computational Statistics, 2nd Ed., John Wiley & Sons, Inc., New Jersey.

[9] Hoeting, J. A. and Givens, G. H., 2013, Computational Statistics, 2th Ed., John Wiley & Sons, Inc., New Jersey, 469 p.

[10] de Jong, K.A., 1975, An Analysis of the Behavior of a Class of Genetic Adaptive Systems, Doctoral Dissertation, University of Michigan, Ann-Arbor, MI.

[11] Umbarkar, A. J. and Sheth, P. D., 2015, Crossover operators in genetic algorithms: A review, ICTACT J. Soft Comput. 6: 1083-1092.

[12] Gwiazda, T. D., 2006, Genetic Algorithms Reference, TomaszGwiazda E- Book, Poland, 410 p.

[13] Shodhganga a Reservoir of Indian Theses, Chapter 9: Crossover, Available Source: http://shodhganga.inflibnet.ac.in/bitstream/10603/32680/19/19_chapter%209.pdf, October 9, 2018.

[14] Soni, N. and Kumar, T., 2014, Study of various mutation operators in genetic algorithms, IJCSIT 5: 4519-4521.

[15] Burnham, K. P. and Anderson, D. R., 2004, Multimodel inference: Understanding AIC and BIC in model selection, Sociol. Methods Res. 33: 261-304.

[16] Vrieze, S. I., 2012, Model selection and psychological theory: A discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) , Psychol. Methods 17: 228-243.

[17] Yang, Y., 2005, Can the strengths of AIC and BIC be shared?, Biometrika, 92: 937-950.

[18] Johnson, R.W., 1996, Fitting percentage of body fat to simple body measurements, J. Stat. Edu. 4(1).

[19] Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. , 2009, Modeling wine preferences by data mining from physicochemical properties, Decis. Supp. Syst. 47: 547-553.

[20] Cock, D.D., 2011, Ames, Iowa: BDNernative to the Boston Housing Data as an End of Semester Regression Project, J. Stat. Edu. 19(3): 1-15.

[21] National Institute of Diabetes and Digestive and Kidney Diseases, Pima Indian Diabetes, Available Source: https://www.kaggle.com/rnmehta5/pima-indian-diabetes-binary-classification/data, November 25, 2018.

[22] Albright, J. , Introduction to Random Effects Models Including HLM, Available Source: https://www.methodsconsultants. com/ tutorial/ introduction- to- random-effects- models- including- hlm, February 19, 2019.