วิธีการคัดเลือกตัวแปรในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ

Variable Selection in Multiple Linear Regression Analysis

รัฐพงศ์ ชัยเอิก[1] ปริม ชูคากร[1]

วรรณพร จันโทภาส[2]

บทคัดย่อ

บทความนี้ได้ศึกษาเกี่ยวกับวิธีการคัดเลือกตัวแปรในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณมีอยู่หลายวิธีในวิธีการคัดเลือกทั่วไป ประกอบไปด้วย วิธีนำตัวแปรเข้าทั้งหมด (Enter Regression) วิธีเพิ่มตัวแปร (Forward Selection) วิธีลดตัวแปร (Backward Elimination) วิธีเพิ่มตัวแปรอิสระแบบขั้นตอน (Stepwise Selection) ในแต่ละวิธีมีขั้นตอนแตกต่างกันไป ทั้งนี้วิธีการคัดเลือกตัวแปรที่ได้รับความนิยมใช้กันมาก คือ วิธีเพิ่มตัวแปรอิสระแบบขั้นตอน (Stepwise Selection) เพราะวิธีนี้จะพิจารณาทั้งตัวแปรเข้าและตัวแปรออก นอกจากนี้ยังมีวิธีการคัดเลือกที่เกิดการประยุกต์จากวิธีการที่ใช้ในการแก้ไขปัญหาที่มีกลุ่มคำตอบที่แน่นอน เป็นกลุ่มวิธีการที่เรียกว่า ฮิวริสติก (Heuristic) ซึ่งในบทความนี้ได้นำเสนอวิธีเจเนติกอัลกอริทึม (Genetic Algorithm: GA) และวิธีทาบูเสิร์ช (Tabu Search: TS)

คำสำคัญ: วิธีการคัดเลือกตัวแปร, การถดถอยเชิงเส้นพหุคูณ, วิธีเจเนติกอัลกอริทึม, วิธีการค้นหาแบบต้องห้าม

Abstract

This article reviews about variable selection methods in Multiple linear regression analysis. There are many methods are provided in variable selection; The general methods are enter method, forward selection, backward elimination and stepwise selection. There is no recommendation that the method should be use. However, the most widely used method is stepwise selection because this method watches the order in which variable are removed or added.

Moreover, this article reviews the other method for variable selection applying the heuristic methods; genetic algrorithm and tabu search.

Keywords : Variable Selection, Multiple Linear Regression Analysis, Genetic Algorithm, Tabu Search

[1] นักศึกษาระดับปริญญาตรี สาขาสถิติ ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น

[2] อาจารย์ ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยขอนแก่น

1. บทนำ

สถิติมีอยู่หลายเทคนิคที่จะใช้ในการหาความสัมพันธ์ ซึ่งการเลือกใช้ในแต่ละเทคนิคทางสถิตินั้นขึ้นอยู่กับวัตถุประสงค์ของงานวิจัยที่ต้องการจะศึกษา การวิเคราะห์การถดถอย (Regression Analysis) เป็นวิธีการวิเคราะห์ทางสถิติและมีส่วนเกี่ยวข้องในการสร้างตัวแบบทางคณิตศาสตร์ ตัวแบบที่ได้นั้นจะแสดงความสัมพันธ์ของตัวแปรตาม (Dependent Variable) กับตัวแปรอิสระ (Independent Variable) โดยเรียกความสัมพันธ์ของตัวแปรตาม 1 ตัวกับตัวแปรอิสระ 1 ตัวว่า การวิเคราะห์ถดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression Analysis) และหากว่ามีตัวแปรอิสระมากกว่า 1 ตัวกับตัวแปรตามเพียงตัวเดียวจะ เรียกว่า การวิเคราะห์การถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression) (วิชิต หล่อจีระชุณห์กุล และจิราวัลย์ จิตรถเวช, 2548) ซึ่งรูปแบบที่ใช้กันอย่างแพร่หลายและเป็นที่นิยมในการนำมาวิเคราะห์ คือ การวิเคราะห์การถดถอยเชิงเส้นพหุคูณ (Multiple linear regression analysis) โดยตัวแบบนั้นเป็นความสัมพันธ์ระหว่างตัวแปรอิสระที่มีมากกว่า 1 ตัวกับตัวแปรตาม 1 ตัว ทั้งนี้ความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามอยู่ในรูปเชิง-เส้น โดยวัตถุประสงค์ของการวิเคราะห์เพื่อการทำนายค่าของตัวแปรตามที่ได้จากตัวแปรอิสระตลอดจนการอนุมานต่างๆ ที่สามารถทำได้เกี่ยวกับตัวแปรตาม การใช้ตัวแบบการถดถอยเชิงเส้นในการทำนายให้มีประสิทธิภาพนั้นจะขึ้นอยู่กับการเลือกตัวแบบที่มีความเหมาะสมสูงสุด โดยตัวแบบที่จะมีความเหมาะสมควรเกิดจากการคัดเลือกตัวแปรอิสระที่มีอิทธิพลมากต่อตัวแปรตามและจำนวนตัวแปรอิสระที่อยู่ในสมการทำนายต้องไม่มากและไม่น้อยเกินไป เนื่องจากว่าตัวแปรอิสระที่มีมากเกินไปจะทำให้ค่าทำนายที่ได้มีความคลาดเคลื่อนสูงและอาจทำให้เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ (Multicollinearity) นอกจากนั้นแล้วหากตัวแบบที่สร้างขึ้นขาดตัวแปรอิสระที่สำคัญไปจะทำให้ค่าทำนายมีความคลาดเคลื่อนสูงได้เช่นเดียวกัน

บทความนี้จึงขอนำเสนอวิธีการคัดเลือกตัวแปรในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ ซึ่งประกอบด้วย การเลือกตัวแปรโดยวิธีนำตัวแปรเข้าทั้งหมด (Enter Regression) การเลือกตัวแปรโดยวิธีเพิ่มตัว-แปร (Forward Selection) การเลือกตัวแปรโดยวิธีลดตัวแปร (Backward Elimination) การเลือกตัวแปรโดยวิธีเพิ่มตัวแปรอิสระแบบขั้นตอน (Stepwise Regression) ซึ่ง 4 วิธีข้างต้นอาจเรียกว่า วิธีการคัดเลือกตัวแปรโดยวิธีทั่วไป นอกจากนี้จะขอนำเสนอวิธีการคัดเลือกตัวแปรที่ประยุกต์ใช้จากวิธีการหาคำตอบแบบมีเหตุผล ซึ่งเป็นวิธีการ ฮิวริสติก (Heuristic) ที่จะใช้หลักการของอัลกอริทึมในการหาคำตอบที่เหมาะสมที่สุด ได้แก่ การเลือกตัวแปรใช้วิธีการค้นหาแบบต้องห้าม (Tabu Search: TS) และการเลือกตัวแปรโดยใช้วิธีเจเนติกอัลกอริทึม (Genetic Algorithm: GA)

2. วิธีการคัดเลือกตัวแปรในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ

การคัดเลือกตัวแปรในแต่ละวิธีนั้นจะมีความแตกต่างกันออกไปทั้งในวิธีการคัดเลือกตัวแปรและขั้นตอนในการคัดเลือกตัวแปร ซึ่งในบทความนี้จะแบ่งเป็นวิธีการคัดเลือกตัวแปรวิธีการทั่วไปและวิธีหาคำตอบแบบมีเหตุผล

2.1 วิธีการคัดเลือกตัวแปรวิธีการทั่วไป

การพิจารณาการนำตัวแปรเข้าหรือออกของวิธีการทั่วไปนี้จะใช้การทดสอบสมมติฐานเกี่ยวกับพารามิเตอร์ในตัวแบบและพิจารณาค่าเกณฑ์ในการเปรียบเทียบ ซึ่งจะแสดงรายละเอียดดังต่อไปนี้

2.1.1 การเลือกตัวแปรโดยวิธีนำตัวแปรเข้าทั้งหมด (Enter Regression)

เป็นวิธีการเอาตัวแปรอิสระทุกตัวทั้งตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามอย่างมีนัยสำคัญทางสถิติและไม่มีนัยสำคัญทางสถิติเข้าไปวิเคราะห์ในสมการถดถอย ซึ่งการคัดเลือกตัวแปรอิสระด้วยวิธีพิจารณาทุกตัวแบบการถดถอยสรุปเป็นขั้นตอนดังนี้

1. จากตัวแปรอิสระทั้งหมด k ตัว ที่คาดว่ามีความสามารถในการทำนายค่าของตัวแปรตาม (Y) นำตัวแปรอิสระทีละ 1 ตัวแปรสร้างตัวแบบการถดถอยด้วยวิธีกำลังสองน้อยที่สุดแล้วเพิ่มตัวแปรอิสระเข้าไปในตัวแบบการถดถอยทีละ 1 ตัวแปร ได้ตัวแบบการถดถอยทั้งหมด $2^k - 1$ ตัวแบบ

2. คำนวณเกณฑ์ต่างๆ ของแต่ละตัวแบบการถดถอย ได้แก่ ค่าความคลาดเคลื่อนมาตรฐาน (S)โดย $S = \sqrt{MSE}$, ค่าสัมประสิทธิ์การกำหนดพหุคูณ $\left(R^2\right)$, ค่าสัมประสิทธิ์การกำหนดพหุคูณปรับแก้ $\left(R^2_{adj}\right)$ และค่า Mallow's C_p และสามารถเขียนตารางหรือกราฟแสดงค่าเกณฑ์ต่างๆ ของแต่ละตัวแบบการถดถอยเพื่อสะดวกแก่การเปรียบเทียบและการคัดเลือก

3. จากตารางหรือกราฟแสดงค่าเกณฑ์ในขั้นตอนที่ 2 พิจารณาว่ามีตัวแบบการถดถอยใดบ้างที่อยู่ในข่ายของการถูกเลือก อาจพิจารณาจากเกณฑ์เดียวหรือใช้หลายเกณฑ์ประกอบกัน ตัวแบบที่เลือกอาจมีตัวแบบเดียวหรือหลายตัวแบบขึ้นอยู่กับความเหมาะสมในการใช้งาน เช่น เมื่อใช้เกณฑ์ของค่า $\left(R^2\right)$ จะเลือกตัวแบบที่ให้ค่า R^2 สูงสุด หากใช้เกณฑ์ค่า (S) จะเลือกตัวแบบที่ให้ค่า S ต่ำสุดหรือใกล้เคียงต่ำสุด เป็นต้น จุดบกพร่องของวิธีการนี้ คือ ทำให้ได้ตัวแบบที่จะพิจารณาค่อนข้างหลากหลายเป็นการคัดเลือกตัวแปรเข้าสู่สมการถดถอยที่ไม่เหมาะสม

2.1.2 การเลือกตัวแปรโดยวิธีเพิ่มตัวแปร (Forward Selection)

เป็นวิธีการคัดเลือกตัวแปรอิสระเข้าสู่สมการทีละตัวตามลำดับความสัมพันธ์กับ Y โดยตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามมากที่สุดจะถูกคัดเลือกเข้าก่อน โดยพิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์อย่างง่าย (Simple correlation coefficient) สูงสุด เมื่อตัวแปรถูกคัดเข้าสมการแล้วจะทำการทดสอบว่าตัวแปรอิสระนั้นสามารถทำนายตัวแปรตามได้อย่างมีนัยสำคัญทางสถิติหรือไม่ ซึ่งจะสามารถคัดตัวแปรอิสระนี้เข้าไปในตัวแบบการถดถอยได้ก็ต่อเมื่อผลการทดสอบสมมุติฐานด้วยค่าสถิติ F หรือค่าสถิติ t พบว่ามีนัยสำคัญทางสถิติ หมายถึง ตัวแปรอิสระนี้มีความสำคัญในการทำนายค่าของ Y แต่หากผลสรุปพบว่าไม่มีนัยสำคัญทางสถิติ หมายถึงตัวแปรอิสระนี้ไม่มีความสำคัญในการทำนายค่าของ Y จากนั้นจะทำการคัดเลือกตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตาม อันดับถัดไปเข้าสมการแล้วทำการทดสอบว่าตัวแปรอิสระที่อยู่ในสมการสามารถร่วมกันทำนายตัวแปรตามได้อย่างมีนัยสำคัญทางสถิติหรือไม่ ทำเช่นนี้ไปจนกว่าจะไม่มีตัวแปรอิสระใดเข้าไปในสมการได้อีกจึงหยุดการคัดเลือกตัวแปรอิสระ ถือว่าสมการที่ได้นั้นมีความเหมาะสม จุดบกพร่องของวิธีนี้คือ ไม่ได้ตรวจสอบผลกระทบที่เกิดจากการเพิ่มตัวแปรพยากรณ์ตัวใหม่เข้าไปในตัวแบบต่อตัวแปรพยากรณ์ที่เข้าไปในตัวแบบก่อนหน้านี้แล้ว

2.1.3 การเลือกตัวแปรโดยวิธีลดตัวแปร (Backward Elimination)

เป็นวิธีการคัดเลือกตัวแปรอิสระออกจากสมการทีละตัวแปร โดยเริ่มจากการสร้างสมการถดถอยที่นำตัวแปรอิสระทุกตัวเข้าสู่สมการ แล้วจึงคัดเลือกตัวแปรอิสระออกทีละตัว โดยตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามน้อยที่สุดจะถูกคัดออก ตัวแปรอิสระตัวแรกที่ถูกพิจารณาคัดออกจะมีความสำคัญต่อการทำนายค่าของ Y น้อยที่สุด โดยมีค่าสถิติ $Partial\ F$ ต่ำที่สุด หรือมีค่าสถิติ t แบบไม่คิดเครื่องหมาย $\left|(t_0)\right|$ ต่ำที่สุด แต่จะสามารถนำตัวแปรอิสระแรกออกจากตัวแบบการถดถอยได้ต้องผ่านการทดสอบสมมุติฐานว่าตัวแปรอิสระนี้ไม่มีความสำคัญต่อการทำนายค่าของ Y คือ มีค่า $F_0 < F_{out}$ หรือมีค่า $|t_0| < t_{out}$ หรือมี $p-value > p_{out}$ เมื่อค่า F ที่ใช้คัดตัวแปรอิสระออก (F_{out}) คือค่าวิกฤต $F_{\alpha,(1,n-p)}$ ค่า t ที่ใช้คัดตัวแปรอิสระออก (t_{out}) คือค่าวิกฤต $t_{\frac{\alpha}{2},(n-p)}$ โดยที่ $p = k+1$ คือ จำนวนพารามิเตอร์ในตัวแบบการถดถอยเริ่มต้นที่มีตัวแปรอิสระครบเท่ากับ k ตัว และค่า $p-value$ ที่ใช้คัดตัวแปรอิสระออก (p_{out}) คือ ระดับนัยสำคัญจะดำเนินการลดตัวแปรอิสระในตัวแบบการถดถอย ต่อไปทำการทดสอบว่าตัวแปรที่ยังคงอยู่สามารถร่วมกันทำนายตัวแปรตามได้อย่างมีนัยสำคัญทางสถิติหรือไม่ ถ้าไม่ได้ก็จะคัดเลือกตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามน้อยอันดับถัดมาออกจากสมการ แล้วดูว่าสมการที่เหลือตัวแปรอิสระอยู่มีนัยสำคัญทางสถิติหรือไม่ ถ้ามีนัยสำคัญทางสถิติก็จะหยุดการคัดออก แต่ถ้าไม่มีนัยสำคัญทางสถิติก็จะทำการคัดเลือกตัวแปรอิสระออกต่อไปเรื่อยๆจนกว่าจะไม่มีตัวแปรอิสระที่ถูกคัดออกอีกการคัดเลือกจะสิ้นสุดเมื่อตัวแปรอิสระที่เหลืออยู่ในสมการทุกตัวแปรมีความสัมพันธ์กับตัวแปรตามอย่างมีนัยสำคัญทางสถิติ จุดบกพร่องของวิธีนี้ คือ ไม่ทราบว่าตัวแปรอิสระที่อยู่

ในสมการแต่ละตัวสามารถทำนายตัวแปรตามได้มากหรือไม่บอกได้เพียงว่าตัวแปรอิสระในตัวแบบนั้นสามารถร่วมกันทำนายตัวแปรตามได้ (ธิดาเดียว มยุรีสวรรค์, 2559)

2.1.4 การเลือกตัวแปรโดยวิธีเพิ่มตัวแปรอิสระแบบขั้นตอน (Stepwise Selection)

เป็นวิธีการคัดเลือกผสมผสานระหว่างวิธีการคัดเลือกตัวแปรอิสระทั้งแบบการเพิ่มตัวแปรและการลดตัวแปรเข้าด้วยกัน วิธีการถดถอยแบบขั้นตอนสามารถทำได้ ดังนี้

1) เหมือนกับวิธีคัดเลือกตัวแปรโดยเพิ่มตัวแปร คือคัดเลือกตัวแปรอิสระ 1 ตัวเข้าไว้ในตัวแบบการถดถอย โดยเลือกตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตาม Y สูงสุด หรือที่ทำให้ตัวแบบการถดถอยมีค่า SSE ต่ำสุดไว้ในตัวแบบการถดถอยเป็นตัวแปรแรก จากนั้นคำนวณค่าสถิติ $Partial\ F$ หรือค่าสถิติ t เพื่อทดสอบสมมุติฐานว่าตัวแปรอิสระมีความสำคัญต่อการทำนายค่าของ Y หรือไม่ โดยถ้า $F_0 \le F_{IN} = F_{\alpha,(1,n-2)}$ หรือ $|t_0| \le t_{IN} = t_{\frac{\alpha}{2},(n-p)}$ หรือ $p-value \ge p_{IN}$ เมื่อ p_{IN} คือระดับนัยสำคัญสำหรับการคัดตัวแปรอิสระเข้า แสดงว่าตัวแปรอิสระนั้นไม่มีความสำคัญต่อการทำนายค่าของ Y จะสิ้นสุดการคัดเลือกตัวแปรอิสระเข้าไว้ในตัวแบบการถดถอย แต่ถ้า $F_0 > F_{IN}$ หรือ $|t_0| > t_{IN}$ หรือ $p-value < p_{IN}$ แสดงว่าตัวแปรอิสระนั้นมีความสำคัญต่อการทำนายค่าของ Y ซึ่งจะคัดเข้าสู่ตัวแบบการถดถอยเป็นตัวแรก

2) ใช้วิธี Blackward elimination คัดตัวแปรอิสระนั้นออกจากตัวแบบการถดถอย หากเป็นไปได้ โดยจะสามารถคัดตัวแปรอิสระที่อยู่ในตัวแบบออกได้ หาก $F_0 < F_{out} = F_{\alpha,(1,n-2)}$ หรือ $|t_0| < t_{out} = t_{\frac{\alpha}{2},(n-2)}$ หรือ $p-value > p_{out}$ เมื่อ p_{out} คือระดับนัยสำคัญสำหรับการคัดตัวแปรอิสระออก การดำเนินการคัดเลือกตัวแปรอิสระจะสิ้นสุดลงและสรุปว่าไม่มีตัวแปรอิสระใดเลยที่มีความสำคัญต่อการทำนายค่าของตัวแปรตาม Y แต่ถ้า $F_0 \ge F_{out} = F_{\alpha,(1,n-2)}$ หรือ $|t_0| \ge t_{out} = t_{\frac{\alpha}{2},(n-2)}$ หรือ $p-value \le p_{out}$ แสดงว่าตัวแปรอิสระนั้นมีความสำคัญต่อการทำนายจะไม่สามารถคัดออกจากตัวแบบการถดถอยได้

3) พิจารณาค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนระหว่างตัวแปรอิสระกับ Y โดยจะเลือกตัวแปรอิสระที่ให้ค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนกับตัวแปรตามสูงที่สุด โดยถ้า $F_0 > F_{IN}$ หรือ $|t_0| > t_{IN}$ หรือ $p-value < p_{IN}$ แสดงว่าตัวแปรอิสระที่ถูกพิจารณาตัวต่อมามีความสำคัญต่อการทำนายค่าของ Y เมื่อมีตัวแปรอิสระก่อนหน้าอยู่ในตัวแบบการถดถอยก่อนแล้ว จะคัดตัวแปรอิสระที่ถูกพิจารณาเข้ามาอยู่ในตัวแบบเป็นตัวต่อมา

4) หลังจากนั้นจึงทำการทดสอบความมีนัยสำคัญของตัวแปรที่เข้ามาเป็นตัวสุดท้ายก่อน โดยพิจารณาจากค่าสถิติเอฟบางส่วน ถ้ามีค่าน้อยกว่าเกณฑ์ที่กำหนดไว้จะต้องนำตัวแปรนั้นออกจากตัวแบบ แต่ถ้าค่าสถิติเอฟบางส่วนมีค่ามากกว่าเกณฑ์ที่กำหนดไว้ ก็จะย้อนกลับไปทดสอบตัวแปรอิสระตัวก่อนหน้าที่เข้ามาอยู่ในตัวแบบโดยพิจารณาจากค่าสถิติเอฟบางส่วน เช่นเดียวกัน ทำซ้ำในขั้นตอนที่สามและสี่ต่อไปเรื่อยๆ จนกระทั่งไม่สามารถ

นำตัวแปรอิสระใดออกจากตัวแบบได้ หรือไม่สามารถนำตัวแปรอิสระใดเข้าสู่ตัวแบบได้อีก ข้อดีของวิธีการนี้คือ สามารถแก้จุดบกพร่องของวิธีการเลือกตัวแปรโดยวิธีเพิ่มตัวแปร (Forward Selection) และวิธีการเลือกตัวแปรโดยวิธีลดตัวแปร (Backward Elimination) ได้ (Montgomery, Peck and Vining, 2012)

2.2 วิธีการคัดเลือกตัวแปรที่มีการประยุกต์จากวิธีการหาคำตอบแบบมีเหตุผล

2.2.1 การเลือกตัวแปรใช้วิธีการค้นหาแบบต้องห้าม

Tabu search เป็นกระบวนการหาคำตอบแบบมีเหตุผลเหมาะกับในกรณีที่ต้องการหาคำตอบที่มีความซับซ้อนและมีตัวแปรที่ใช้ในการศึกษาเป็นจำนวนมาก ซึ่งวิธีการนี้จะอาศัยอัลกอริทึมในการทำงานของจำนวนรอบให้ได้สูงสุดและพิจารณาการลดลงของฟังก์ชันเป้าหมาย ดังนั้นการวิธีแบบต้องห้ามประยุกต์ใช้กับการคัดเลือกตัวแปรในตัวแบบถดถอยเชิงเส้นพหุได้โดยการจัดการของอัลกอริทึมเพื่อให้ได้ตัวแปรที่เข้าสู่ตัวแบบมีจำนวนน้อยที่สุดและมีฟังก์ชันเป้าหมาย คือ ต้องการให้ความคลาดเคลื่อนต่ำสุด อีกทั้งยังประหยัดเวลาในการคัดเลือกตัวแปรและวิธีการค้นหาแบบต้องห้ามนี้เหมาะแก่การคัดเลือกตัวแปรได้ทั้งกรณีที่ข้อมูลไม่มีและมีสหสัมพันธ์เชิงเส้นพหุ (Multicollinearity) เนื่องจากมีการกำหนดค่าของสัมประสิทธิ์ β ที่ทำให้ฟังก์ชันเป้าหมายมีค่าต่ำสุดจึงไม่ต้องมีการคำนวณ $\left(x'x\right)^{-1}$ ที่หากมีการเกิดปัญหาสหสัมพันธ์เชิงเส้นพหุแล้วหากประมาณค่า β ด้วยวิธีกำลังสองน้อยที่สุด (OLS) จะทำให้ $\beta=\left(x'x\right)^{-1}x'y$ มีความคลาดเคลื่อนสูง ซึ่งมีผลต่อการทดสอบสมมติฐานเกี่ยวกับพามิเตอร์ในตัวแบบจึงมีผลต่อการคัดเลือกตัวแบบโดยวิธีดั้งเดิม วิธีค้นหาแบบต้องห้ามสำหรับการคัดเลือกตัวแปรในตัวแบบการถดถอยเชิงพหุคูณสามารถสรุปขั้นตอนได้ดังนี้

1. กำหนดการหาคำตอบเริ่มต้นโดยการกำหนดตัวแบบการถดถอยการกำหนดค่าสัมประสิทธิ์การถดถอยเพื่อนำมาจัดทำโครงสร้างหน่วยความจำ

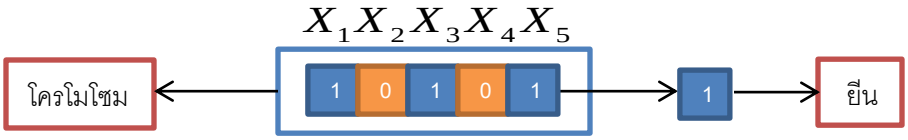
2. สร้าง Candidate โดยการสร้างเซตคำตอบข้างเคียงจากตัวแบบที่กำหนดในขั้นที่ 1 ภายใต้ความเหมาะสมที่สุดที่ถูกกำหนดจากฟังก์ชันเป้าหมาย ซึ่งในที่นี้คือ MSE และ $|MSE|$

3. ทำการปรับปรุง (Update) ตามลำดับ โดยการแทนคำตอบที่ดีที่สุดในปัจจุบันด้วยคำตอบที่ดีกว่าจากนั้นดำเนินการทำซ้ำในขั้นที่ 1 ตามลำดับ จนกระทั่งครบตามจำนวนรอบที่กำหนด

ข้อดีของวิธีแบบต้องห้าม คือ ช่วยป้องกันการคัดเลือกตัวแปรอิสระเข้ามาในตัวแบบมากเกินไปและสามารถใช้ในการคัดเลือกตัวแปรในกรณีที่ข้อมูลเกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ (Multicollinearity) ได้

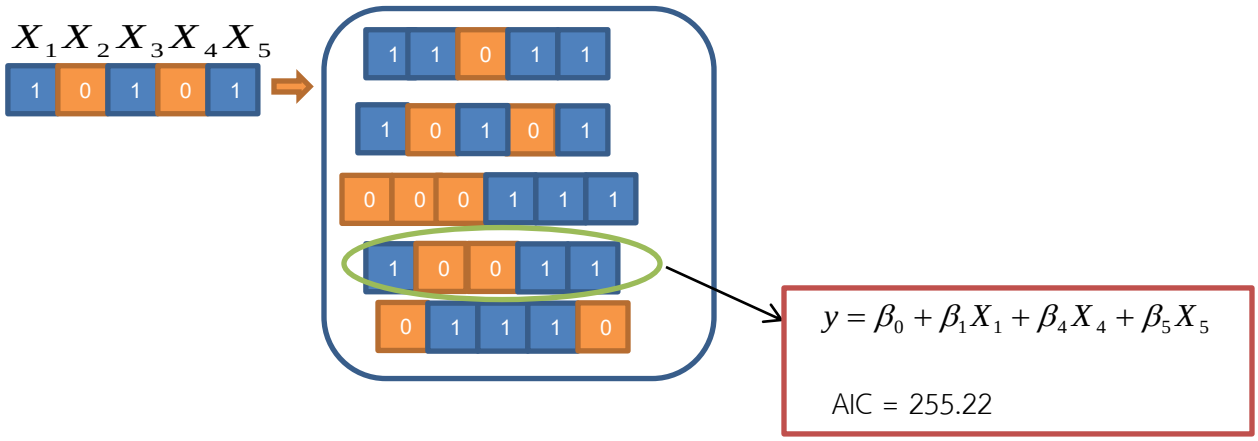
2.2.2 การเลือกตัวแปรโดยใช้วิธีเจเนติกอัลกอริทึม (Genetic Algorithm: GA)

เป็นวิธีการแก้ปัญหารูปแบบหนึ่งเพื่อให้ได้คำตอบเหมาะสมที่สุด บนพื้นฐานของทฤษฎีอธิบายเกี่ยวกับวัฏจักรทางพันธุกรรมของสิ่งมีชีวิตโดยจะอธิบายการเปลี่ยนแปลงกระบวณการทางธรรมชาติของพันธุกรรมและนำกลไกการเปลี่ยนแปลงมาประยุกต์ใช้ในการแก้ปัญหาค่าเหมาะสมที่สุด หลักการทำงานของ GA จะถูกนำเสนอข้อมูลในรูปแบบโครโมโซม โดยมีขั้นตอนเริ่มต้นจากการสร้างประชากรของคำตอบ เรียกว่า โครโมโซมจากการสุ่ม จากนั้นทำการถอดรหัสโครโมโซมและคำนวณค่าเหมาะสม ประชากรเหล่านี้จะต้องผ่านตัวดำเนินการทางพันธุกรรมเพื่อให้เกิดการปรับเปลี่ยนสายพันธุ์ ซึ่งได้แก่ การคัดเลือกสายพันธุ์ (Selection) การสลับสายพันธุ์ (Crossover) และการกลายพันธุ์ (Mutation) โดยกระบวนการจะถูกทำซ้ำไปเรื่อยๆ จนกว่าจะตรงเงื่อนไขในการหยุดค้นหา (ศิรินทิพย์ หมื่นจันทร์ และวฐา มินเสน, 2557) สำหรับการคัดเลือกตัวแปรในตัวแบบถดถอยเชิงพหุคูณที่ทำวิธี GA มาประยุกต์ใช้นั้น การกำหนดโครโมโซมจะหมายถึง ตัวแบบการถดถอยเชิงพหุคูณและยีนแต่ละยีนที่ประกอบเป็นโครโมโซมจะหมายถึง ตัวแปรอิสระในตัวแบบการถดถอยและค่าเหมาะสมที่ใช้ในการพิจารณา คือ ค่า Akaike's Information Criterion : AIC โดยมีขั้นตอนดังนี้ กำหนดจำนวนตัวแปรอิสระที่ใช้ในการพิจารณาจำนวน 5 ตัวแปร ตัวแบบถดถอย คือ $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$



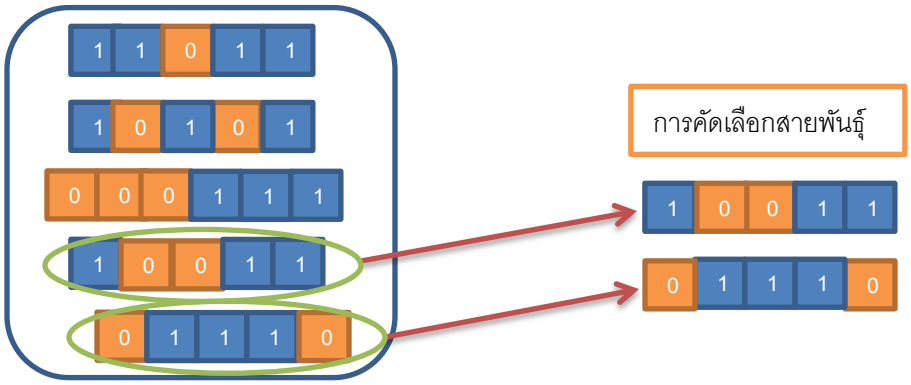
รูปที่ 1 ลักษณะของสายโครโมโซม

ขั้นที่ 1 สุ่มประชากรเพื่อเป็นประชากรเริ่มต้นของประชากรรุ่นที่ 1 โดยอยู่ในลักษณะสายโครโมโซม ซึ่งเปรียบได้กับตัวแบบการถดถอย ประกอบด้วยยีนแบบไบนารี (รูปที่ 1) โดยยีนเปรียบได้กับตัวแปรอิสระ พร้อมแปลงรหัสเพื่อคำนวณค่า AIC ของแต่ละโครโมโซม ดังรูปที่ 2 ยกตัวอย่าง เช่น โครโมโซมแท่งที่ 4 ทำการคัดเลือกตัวแปรแล้วสามารถเขียนตัวแบบถดถอยได้ คือ $y = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + \beta_5 X_5$ โดยมีค่า AIC = 255.22



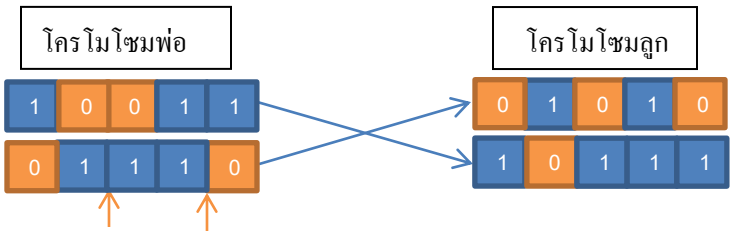
รูปที่ 2 แสดงรูปแบบสายโครโมโซม จำนวนโครโมโซมที่สร้างขึ้นและการแปลงรหัสโครโมโซม

ขั้นที่ 2 Selection: ทำการคัดเลือกประชากรเพื่อนำมาเป็นโครโมโซมต้นแบบ เรียกว่า โครโมโซมพ่อแม่ โดยเลือกจากโครโมโซมทั้งหมดที่มีค่า AIC ที่น้อยที่สุด 2 อันดับ (รูปที่ 3) ซึ่งได้แก่ โครโมโซมแท่งที่ 4 และ 5



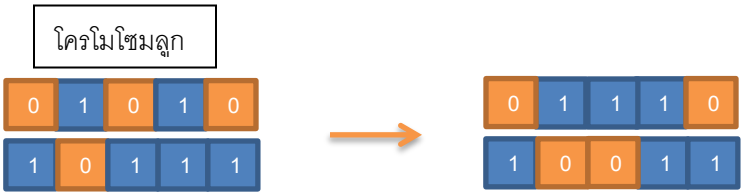
รูปที่ 3 แสดงการคัดเลือกโครโมโซม

ขั้นที่ 3 Crossover: สมมติกำหนดความน่าจะเป็นของการสลับสายพันธุ์ = 0.8 แล้วสุ่มตัวเลขช่วง 0 ถึง 1 หากมีค่าน้อยกว่า 0.8 จะทำการสลับสายพันธุ์ ในที่นี้ทำการสลับสายพันธุ์ชนิดกำหนดจำนวนจุด 2 จุด แต่ต้องทำไปโดยการสุ่มเพื่อเลือกตำแหน่งที่จะสลับสายพันธุ์ จากนั้นก็สลับสายพันธุ์โครโมโซมแล้วจะได้เป็นโครโมโซมลูก ดังรูปที่ 4



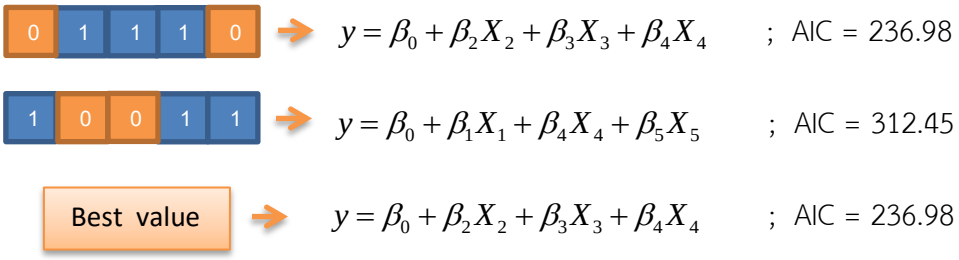
รูปที่ 4 แสดงการสลับสายพันธุ์

ขั้นที่ 4 Mutation: สมมติกำหนดความน่าจะเป็นของการกลายพันธุ์ = 0.2 แล้วสุ่มตัวเลขช่วง 0 ถึง 1 หากมีค่าน้อยกว่า 0.2 จะทำการกลายพันธุ์ จากนั้นทำการเปลี่ยนยีนบางตัวโดยสุ่มให้เกิดความแตกต่างของโครโมโซม ซึ่งในกรณีที่มีการกลายพันธุ์ตำแหน่งและจำนวนตำแหน่งที่จะทำการกลายพันธุ์เป็นไปโดยสุ่ม แล้วจึงทำการกลายพันธุ์จากนั้นคำนวณค่า AIC ของโครโมโซมที่ได้ ดังรูปที่ 5



รูปที่ 5 แสดงการกลายพันธุ์

ขั้นที่ 5 แปลงรหัสจากโครโมโซมที่ได้เป็นตัวแปรที่เหมาะสมจะอยู่ในตัวแบบถดถอยแล้วคำนวณค่า AIC เลือกโครโมโซมที่ให้ค่า AIC น้อยที่สุด (รูปที่6) เพื่อนำมาเป็นตัวแบบที่เหมาะสมที่สุดสำหรับประชากรรุ่นที่ 1 ในการพิจารณาโครโมโซมพ่อแม่ในรุ่นถัดไป จากนั้นดำเนินการในขั้นตอนที่ 1-5 จนครบจำนวนรุ่น (รอบ) ตามกำหนด



รูปที่ 6 แสดงการแปลงรหัสของโครโมโซมของประชากรรุ่นที่ 1 และเลือกโครโมโซม

3. งานวิจัยที่ทำการเปรียบเทียบการคัดเลือกตัวแปรโดยวิธีทั่วไปกับวิธีการค้นหาคำตอบแบบมีเหตุผล

การเปรียบเทียบวิธีการคัดเลือกตัวแปรอิสระระหว่างวิธี Tabu search กับวิธี Stepwise โดย กานต์ณัฐ ณ บางช้าง และจิราวัลย์ จิตรถเวช (2556) ศึกษาในกรณีค่าสัมประสิทธิ์สหสัมพันธ์ของตัวแปรอิสระอยู่ในระดับต่ำ (น้อยกว่า 0.1) วิธี Stepwise ตัวแปรอิสระทุกตัวจะให้ค่าสัมประสิทธิ์การถดถอยใกล้เคียงกับค่าที่กำหนดในการจำลองและตัวแบบที่ได้มีความถูกต้องสูงสุด แต่วิธี Tabu search จะให้ค่าสัมประสิทธิ์การถดถอยของตัวแปรอิสระแต่ละตัวเปลี่ยนไป ซึ่งการคัดเลือกตัวแบบมีความถูกต้องต่ำกว่าวิธี Stepwise ประมาณร้อยละ 1 กรณีค่าสัมประสิทธิ์สหสัมพันธ์ระดับปานกลาง (เท่ากับ 0.5) และระดับสูง (มากกว่า 0.96) เมื่อมีฟังก์ชันเป้าหมาย 2 ฟังก์ชัน ใช้วิธี Tabu search ตัวแปรอิสระทุกตัวจะให้ค่าสัมประสิทธิ์การถดถอยใกล้เคียงกับค่าที่กำหนดในการจำลองและตัวแบบที่ได้มีความถูกต้องมากกว่าวิธี Stepwise

ต่อมา ศิรินทิพย์ หมื่นจันทร์ และวฐา มินเสน (2557) ได้ทำการศึกษาการเปรียบเทียบวิธีการคัดเลือกตัวแปรอิสระระหว่างวิธี Genetic กับวิธี Stepwise ในสถานการณ์ที่ไม่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุทั้ง 2 วิธี พบว่าให้ค่า MSE ต่ำที่สุดเท่ากัน เมื่อพิจารณาจากความแปรปรวนวิธี Genetic ในทุกขนาดตัวอย่างที่ศึกษามีความแปรปรวนมากกว่าวิธี Stepwise และเมื่อทำการเปรียบเทียบค่า AMSE วิธี Stepwise จะให้ค่าต่ำกว่าวิธี Genetic ส่วนสถานการณ์ที่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุพบว่าวิธี Genetic ให้ค่า MSE ค่อนข้างต่ำกว่าวิธี Stepwise เมื่อพิจารณาจากความแปรปรวนวิธี Genetic มีความแปรปรวนต่ำกว่า วิธี Stepwise และเมื่อทำการเปรียบเทียบค่า AMSE วิธี Genetic ให้ค่าต่ำกว่าวิธี Genetic ซึ่งเมื่อเปรียบเทียบแล้ววิธี Genetic ให้ค่าต่ำกว่าวิธีของ Stepwise เมื่อส่วนเบี่ยงเบนมาตรฐานของความคลาดเคลื่อนสูง

4.สรุป

การคัดเลือกตัวแปรอิสระโดยวิธีทั่วไปเข้าสมการถดถอยเชิงเส้นพหุคูณมีวิธีการคัดเลือกหลายวิธี แต่ละวิธีมีจุดเด่นและข้อบกพร่องแตกต่างกัน โดยวิธีที่นิยมนำมาใช้อย่างแพร่หลาย คือวิธี Stepwise เพราะ วิธีการนี้จะทำการทดสอบตัวแปรอิสระที่เข้าสมการไปแล้วทุกครั้งที่มีการนำตัวแปรอิสระใหม่เข้าในสมการ ซึ่งตัวแปรอิสระบางตัวที่เข้าไปในสมการแล้วก็สามารถถูกขจัดออกจากสมการได้ หากพบว่าตัวแปรอิสระตัวนั้นไม่ได้ส่งผลให้ค่า R^2 เพิ่มขึ้นอย่างมีนัยสำคัญทางสถิติ

นอกจากนี้ยังมีวิธีการคัดเลือกตัวแบบที่ประยุกต์มาจากวิธีการที่ใช้ในการแก้ปัญหาที่มีชุดคำตอบที่เป็นไปได้แน่นอน เป็นกลุ่มวิธีการที่เรียกว่า ฮิวริสติก (Heuristic) โดยวิธีนี้จะอาศัยหลักการของอัลกอริทึม ซึ่งในที่นี้ได้นำเสนอไป 2 วิธี คือ วิธีการค้นหาแบบต้องห้าม และวิธีเจเนติกอัลกอริทึม นอกจากนี้ยังมีวิธีซิมมูเลเตดแอ ล-เนลลิ่ง (Simulated Annealing : SA) ที่สามารถนำแนวคิดของวิธีการนี้มาช่วยในการคัดเลือกตัวแบบได้เช่นกัน

บรรณานุกรม

กานต์ณัฐ ณ บางช้าง, และจิราวัลย์ จิตรถเวช. (2556). **การคัดเลือกตัวแปรในตัวแบบการถดถอยเชิงเส้นพหุโดยใช้วิธีการค้นหาแบบต้องห้าม**. วารสารวิทยาศาสตร์ มข, 41(1), 250–261.

ธิดาเดียว มยุรีสวรรค์. (2559). **การวิเคราะห์การถดถอย :Regression Analysis** (พิมพ์ครั้งที่ 1) . ขอนแก่น: บริษัท เพ็ญพรินติ้ง จำกัด.

วิชิต หล่อจีระชุณห์กุล, และจิราวัลย์ จิตรถเวช. (2548). **เทคนิคการพยากรณ์** (พิมพ์ครั้งที่ 3). กรุงเทพมหานคร: โครงการส่งเสริมและพัฒนาเอกสารวิชาการ สถาบันบัณฑิตพัฒนบริหารศาสตร์.

ศิรินทิพย์ หมื่นจันทร์, และวฐา มินเสน. (2557). **การคัดเลือกตัวแบบในการถดถอยเชิงเส้นพหุคูณโดยใช้ วิธีดับเบิลเจเนติกอัลกอริทึม**. วารสารวิทยาศาสตร์บูรพา Burapha Science Journal, 19(2), 139–153.

Montgomery, D.C., Peck, E.A. And Vining G.G. (2012). **Introduction to Linear Regression Analysis**. 5th ed. New Jersey: John Wiley & Sons.