

บทที่ 1

บทนำ

1.1 ที่มาและความสำคัญของปัญหา

ในการวิเคราะห์การถดถอยเชิงเส้นพหุคูณ (Multiple Linear Regression Analysis) เป็นการหา รูปแบบความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระ (1) โดยการสร้างตัวแบบเพื่ออธิบาย ความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอิสระซึ่งมีความสัมพันธ์ในรูปเชิงเส้น รวมถึงการ พยากรณ์ค่าตัวแปรตามจากตัวแปรอิสระ ซึ่งการใช้ตัวแบบการถดถอยเชิงเส้นในการพยากรณ์ให้มีความ ถูกต้องและแม่นยำที่สุดนั้น ขึ้นอยู่กับตัวแบบที่มีความเหมาะสม การได้ตัวแบบที่มีความ เหมาะสมต้องมาจากการคัดเลือกตัวแปรอิสระที่มีอิทธิพลต่อตัวแปรตาม ในทางปฏิบัติอาจพบว่าตัว แปรอิสระที่ใช้ในการอธิบายตัวแปรตามนั้นมีอยู่เป็นจำนวนมาก แต่ตัวแปรเหล่านั้นมีความสามารถ ในการอธิบายตัวแปรตามได้ไม่เท่ากัน ตัวแปรอิสระบางตัวอาจมีความสามารถในการอธิบายตัวแปร ตามได้สูงมาก ในขณะที่ตัวแปรอิสระบางตัวสามารถอธิบายตัวแปรตามได้ค่อนข้างต่ำในกรณีเช่นนี้ ควรตัดตัวแปรดังกล่าวออกจากตัวแบบการถดถอย นอกจากนี้ตัวแปรอิสระบางตัวที่นำมาใช้ในการ วิเคราะห์การถดถอย อาจมีความสัมพันธ์กันเอง หรือเรียกว่าเกิดปัญหาสหสัมพันธ์เชิงเส้น พหุ (Multicollinearity) ซึ่งจะส่งผลให้ความแปรปรวนของค่าประมาณพารามิเตอร์ของตัวแบบมีค่าสูง ดังนั้นในการวิเคราะห์จึงต้องกระทำอย่างรอบคอบเพื่อหาตัวแปรอิสระที่เหมาะสมที่สุดในการอธิบาย ตัวแปรตาม อนึ่ง การตัดตัวแปรอิสระที่สำคัญออกไปจากตัวแบบการถดถอย จะส่งผลให้ตัวแบบที่ใช้ นั้นเป็นตัวแบบที่มีประสิทธิภาพต่ำ และทำให้ค่าพยากรณ์ที่ได้มีความคลาดเคลื่อนสูง (2, 3)

ในปัจจุบันมีวิธีการคัดเลือกตัวแบบการถดถอยเชิงเส้นพหุคูณอยู่หลายวิธีโดยวิธีที่นิยมใช้กัน อย่างแพร่หลาย เช่น วิธีพิจารณาทุกตัวแบบที่เป็นไปได้ (All Possible Regression) เป็นการเลือกตัว แปรอิสระจากตัวแบบการถดถอยทั้งหมดที่เป็นไปได้ โดยอาศัยหลักเกณฑ์ต่างๆที่กำหนดขึ้น เช่น พิจารณาจากค่าความคลาดเคลื่อนหรือค่าสัมประสิทธิ์การตัดสินใจ เป็นต้น (4) วิธี FS เป็นวิธีการเลือก ตัวแปรอิสระเข้าในตัวแบบการถดถอยที่เริ่มแรกยังไม่มีตัวแปรอิสระอยู่เลย พิจารณาเข้าตัวแบบครั้ง ละหนึ่งตัวแปร ทำการคัดเลือกไปจนกว่าจะไม่มีตัวแปรอิสระตัวใดเข้าสู่ตัวแบบได้อีก วิธีการนี้เมื่อตัว

แปรอิสระถูกเลือกเข้าตัวแบบแล้วจะไม่มีพิจารณาตัดตัวแปรอิสระออกจากตัวแบบในภายหลัง วิธี BE เริ่มจากตัวแบบการถดถอยที่ประกอบด้วยตัวแปรอิสระทุกตัว หลังจากนั้นพิจารณาตัดตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามน้อยที่สุดออกครั้งละหนึ่งตัวแปร พิจารณาไปจนกระทั่งตัวแปรที่เหลืออยู่ในตัวแบบสามารถอธิบายความสัมพันธ์กับตัวแปรตามได้ทั้งหมด จึงหยุดกระบวนการ เมื่อตัวแปรอิสระตัวใดถูกตัดออกจากตัวแบบแล้ว จะไม่มีการพิจารณาเพิ่มเข้ามาอยู่ในตัวแบบอีกในภายหลัง (2) และวิธี SR เป็นวิธีการผสมระหว่างวิธี FS และวิธี BE โดยเลือกตัวแปรอิสระเข้าในตัวแบบการถดถอยครั้งละหนึ่งตัวแปร โดยตัวแปรอิสระที่อยู่ในตัวแบบแล้ว จะต้องมีการทดสอบว่าตัวแปรนั้นยังมีส่วนในการอธิบายความผันแปรของตัวแปรตาม ขณะที่ตัวแปรอิสระอื่นอยู่ในตัวแบบหรือไม่ นั่นคือ ตัวแปรอิสระใดที่ถูกเพิ่มเข้ามาในตัวแบบการถดถอยแล้ว อาจถูกตัดออกได้ในภายหลัง (4)

นอกจากวิธีดังกล่าวข้างต้นแล้วยังมีวิธีการคัดเลือกตัวแบบที่ประยุกต์มาจากวิธีที่ใช้ในการแก้ปัญหาที่มีชุดคำตอบที่เป็นไปได้แน่นอน (Combinatorial Optimization) เป็นกลุ่มวิธีการที่เรียกว่า เมตาฮิวริสติก (Metaheuristic) เช่น วิธี GA วิธี TS และวิธี SA เป็นต้น โดยวิธีการเหล่านี้เป็นกระบวนการสำหรับแก้ปัญหาเพื่อหาค่าเหมาะสมที่สุดซึ่งมีขั้นตอนมาตรฐานที่แน่นอน รวมถึงมีเกณฑ์ในการหยุดกระบวนการทำงานเหมือนกัน นั่นคือ ครบจำนวนรอบที่กำหนด ได้ค่าฟังก์ชันเป้าหมาย (Objective Function) ที่ต้องการ หรือเหตุผลอื่นๆ ที่ได้กำหนดไว้ ดังนั้นจึงสามารถนำแนวคิดของวิธีการเหล่านี้มาช่วยในการคัดเลือกตัวแบบได้เช่นกัน โดยมีเป้าหมายในการคัดเลือกตัวแบบการถดถอยที่เหมาะสมที่สุดในการอธิบายตัวแปรตาม และมีความคลาดเคลื่อนในการพยากรณ์ต่ำ นอกจากนี้ยังช่วยลดค่าใช้จ่ายและเวลาในการเก็บรวบรวมข้อมูลเกี่ยวกับตัวแปรอิสระในตัวแบบการถดถอยเชิงเส้นพหุคูณที่จะนำมาใช้ในการพยากรณ์ตัวแปรตาม

ในสถานการณ์ที่ไม่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ การประมาณค่าโดยใช้วิธี OLS สามารถประมาณค่าสัมประสิทธิ์การถดถอยของตัวแปรอิสระได้เป็นอย่างดี แต่ในสถานการณ์ที่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ วิธีการประมาณค่าพารามิเตอร์โดยวิธี OLS จะประมาณค่าสัมประสิทธิ์การถดถอยของตัวแปรอิสระผิดพลาดไปจากค่าจริง และทำให้ค่าความแปรปรวนของความคลาดเคลื่อนมีค่าสูงกว่าปกติ (5) ดังนั้นผู้วิจัยจึงนำแนวคิดของวิธี GA สำหรับการคัดเลือกตัวแปร ทำงานร่วมกับวิธี GA สำหรับหาค่าสัมประสิทธิ์การถดถอยของตัวแบบ โดยเรียกว่า วิธี DGA มาประยุกต์ใช้ในการคัดเลือกตัวแบบการถดถอยเชิงเส้นพหุคูณ เพื่อให้ได้ตัวแบบที่เหมาะสมที่สุดในการพยากรณ์ตัวแปรตาม โดยให้ค่าความคลาดเคลื่อนในการพยากรณ์น้อยที่สุด และนำวิธี DGA เปรียบเทียบกับวิธี SR

และวิธี GA ที่ประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธี OLS อาศัยเกณฑ์ AIC เป็นเกณฑ์ในการพิจารณาคัดเลือกตัวแบบ และเกณฑ์ MSE เป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพของทั้ง 3 วิธี

1.2 วัตถุประสงค์ของการวิจัย

เพื่อประยุกต์วิธี DGA สำหรับการคัดเลือกตัวแบบการถดถอยเชิงเส้นพหุคูณ พร้อมทั้งทำการเปรียบเทียบกับวิธี SR และวิธี GA โดยใช้เกณฑ์ MSE เป็นเกณฑ์ในการเปรียบเทียบประสิทธิภาพของทั้ง 3 วิธีและใช้ค่า AIC เป็นเกณฑ์พิจารณาคัดเลือกตัวแบบ เมื่อข้อมูลเกิดและไม่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ

1.3 ขอบเขตของการวิจัย

1.3.1 ศึกษาภายใต้ตัวแบบการถดถอยเชิงเส้นพหุคูณที่อยู่ในรูปเชิงเส้นของพารามิเตอร์และตัวแปรอิสระโดยมีรูปแบบทั่วไปคือ

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i$$

โดยที่ y_i แทน ตัวแปรตาม เมื่อ $i = 1, \dots, n$ และ n แทนขนาดตัวอย่าง

$X_{i1}, X_{i2}, \dots, X_{ik}$ แทน ตัวแปรอิสระ เมื่อ k แทนจำนวนตัวแปรอิสระ

$\beta_0, \beta_1, \dots, \beta_k$ แทน สัมประสิทธิ์การถดถอย

ε_i แทน ความคลาดเคลื่อนสุ่ม

1.3.2 ทำการจำลองข้อมูล เพื่อสร้างตัวแบบถดถอยโดยกำหนดตัวแปรอิสระมีการแจกแจงแบบเอกรูป มีรูปแบบคือ $X_{ij} \sim U(a, b)$ กำหนดความคลาดเคลื่อนมีการแจกแจงแบบปกติ มีรูปแบบคือ $\varepsilon_i \sim N(0, \sigma^2)$ และกำหนดขนาดประชากร $N=10,000$

1.3.3 ทำการศึกษาภายใต้สถานการณ์ต่างๆ ดังนี้

1) การเกิดปัญหาสหสัมพันธ์เชิงเส้นพหุและไม่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ

2) ขนาดตัวอย่างเท่ากับ 50, 100 และ 500

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

4) กำหนดจำนวนตัวแปรอิสระเท่ากับ 10 และ 20 ตัวแปร ดังนี้

เมื่อตัวแปรอิสระมีจำนวน 10 ตัวแปร จะกำหนดให้

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

เมื่อตัวแปรอิสระมีจำนวน 20 ตัวแปร จะกำหนดให้

$$X_{i1} \sim U(25,200), X_{i2} \sim U(50,180), X_{i3} \sim U(15,250), X_{i4} \sim U(87,550), \\ X_{i5} \sim U(1,70), X_{i6} \sim U(35,90), X_{i7} \sim U(3,17), X_{i8} \sim U(8,22), \\ X_{i9} \sim U(15,80), X_{i10} \sim U(20,150), X_{i11} \sim U(1200,1550), \\ X_{i12} \sim U(800,900), X_{i13} \sim U(890,1000), X_{i14} \sim U(500,900) \\ X_{i15} \sim U(1300,1700), X_{i16} \sim U(3500,5000), X_{i17} \sim U(3200,17000), \\ X_{i18} \sim U(1900,2900), X_{i19} \sim U(15000,25000) \text{ และ } \\ X_{i20} \sim U(200000,500000)$$

1.3.4 กำหนดค่าสัมประสิทธิ์การถดถอยโดยพิจารณาจากช่วงของตัวแปรอิสระแต่ละตัวที่ทำการกำหนดขึ้นและจากผลงานวิจัยของ กานต์ฉัฐ ณ บางช้าง (5) ที่ทำการศึกษาเกี่ยวกับการคัดเลือกแปรในตัวแบบการถดถอยเชิงเส้นซึ่งมีลักษณะที่คล้ายกัน ดังนั้นจึงทำการกำหนดค่าสัมประสิทธิ์การถดถอยดังต่อไปนี้

เมื่อตัวแปรอิสระมีจำนวน 10 ตัวแปร จะกำหนดให้

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

เมื่อตัวแปรอิสระมีจำนวน 20 ตัวแปร จะกำหนดให้

$$\beta_0 = 100, \beta_1 = -155, \beta_2 = 15, \beta_3 = 11.5, \beta_4 = 509, \beta_5 = 0, \beta_6 = 0, \\ \beta_7 = 4, \beta_8 = 7, \beta_9 = 0, \beta_{10} = 0, \beta_{11} = -1,550, \beta_{12} = 900, \beta_{13} = 1,000, \\ \beta_{14} = 50, \beta_{15} = 0, \beta_{16} = 0, \beta_{17} = 40, \beta_{18} = 100, \beta_{19} = 0 \text{ และ } \beta_{20} = 0$$

1.3.5 กำหนดตัวแปรตามจากตัวแบบการถดถอยต่อไปนี้

เมื่อตัวแปรอิสระมีจำนวน 10 ตัวแปร

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

เมื่อตัวแปรอิสระมีจำนวน 20 ตัวแปร

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \dots + \beta_{19} X_{i19} + \beta_{20} X_{i20} + \varepsilon_i$$

1.3.6 สำหรับสถานการณ์ที่เกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ กำหนดให้ตัวแปรอิสระ

$$X_{i3} = X_{i1} + X_{i2}$$

1.3.7 วิธีการคัดเลือกตัวแปรที่ใช้ ได้แก่ วิธี DGA วิธี SR และวิธี GA

1.3.8 ฟังก์ชันเป้าหมายของวิธี DGA ได้แก่ ฟังก์ชัน AIC และ MSE ส่วนฟังก์ชันเป้าหมายของวิธี GA ได้แก่ ฟังก์ชัน AIC

1.3.9 ทำการคัดเลือกตัวแบบโดยใช้เกณฑ์ AIC และทำการเปรียบเทียบวิธี DGA วิธี SR และวิธี GA โดยใช้เกณฑ์ MSE ซึ่งประกอบด้วยค่า MSE ที่ต่ำที่สุดและค่า AMSE

1.3.10 กำหนดจำนวนกระทำซ้ำ 100 รอบ และ จำนวนรุ่นประชากรของวิธี DGA และวิธี GA เท่ากับ 2,000 รุ่น ในแต่ละสถานการณ์

1.4 วิธีการดำเนินการวิจัย

1.4.1 จำลองข้อมูลตัวแปร $X_{ij} \sim U(a, b)$, $\varepsilon_i \sim N(0, \sigma^2)$ และ β_j พร้อมทั้งข้อมูลตัวแปร y_i จากตัวแบบการถดถอยที่กำหนด ในจำนวนประชากรขนาด $N=10,000$

1.4.2 ทำการสุ่มตัวอย่างตามที่ได้กำหนด

1.4.3 คัดเลือกตัวแบบด้วยวิธี DGA วิธี SR และ วิธี GA

1.4.4 ทำซ้ำจนกว่าจะสิ้นสุดการทำงานในแต่ละสถานการณ์

1.4.5 ทำการเปรียบเทียบผลการคัดเลือกตัวแบบของทั้ง 3 วิธี และสรุปผลการวิจัย

1.5 นิยามศัพท์

1.5.1 ตัวแบบจริง หมายถึง ตัวแบบการถดถอยเชิงเส้นพหุคูณที่มีตัวแปรอิสระครบทุกตัวและมีความคลาดเคลื่อนเชิงสุ่มที่กำหนดขึ้นที่ใช้สร้างข้อมูลตัวแปรตาม

1.5.2 ตัวแบบที่เหมาะสม หมายถึง ตัวแบบการถดถอยเชิงเส้นพหุคูณที่มีตัวแปรอิสระเพียงบางตัว ที่ได้รับการคัดเลือกจากวิธีการคัดเลือกตัวแปรของแต่ละวิธี

1.5.3 การเกิดปัญหาสหสัมพันธ์เชิงเส้นพหุ (Multicollinearity) หมายถึง การที่ตัวแปรอิสระของตัวแบบการถดถอยเชิงเส้นพหุคูณมีความสัมพันธ์กัน

1.5.4 จุดที่สูงที่สุด (Global Maximum) หมายถึง จุดที่อยู่สูงที่สุดของกราฟฟังก์ชันค่าตอบ

1.5.5 จุดสูงสุด (Local Maximum) หมายถึง จุดที่มีลักษณะคล้ายกับจุดที่สูงที่สุดของกราฟฟังก์ชันค่าตอบซึ่งจะเป็นจุดสูงสุดบริเวณเฉพาะที่

1.5.6 จุดที่ต่ำที่สุด (Global Minimum) หมายถึง จุดที่อยู่ต่ำที่สุดของกราฟฟังก์ชันค่าตอบ

1.5.7 จุดต่ำสุด (Local Minimum) หมายถึง จุดที่มีลักษณะคล้ายกับจุดที่ต่ำที่สุดของกราฟฟังก์ชันค่าตอบซึ่งจะเป็นจุดต่ำสุดบริเวณเฉพาะที่

1.5.8 โครโมโซม (Chromosome) หมายถึง สายรหัสที่เป็นตัวแทนของค่าตอบ

1.5.9 ยีน (Gene) หมายถึง รหัส ตัวอักษรหรือ ตัวเลขที่ใช้เพื่อจะเป็นค่าตอบ

1.5.10 แอลลีล (Allele) หมายถึง ค่าของรหัส

1.5.11 ลูคัส (Locus) หมายถึง ตำแหน่งของรหัสบนสายรหัส

1.6 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1.6.1 เรียนรู้ขั้นตอนในการประยุกต์ใช้เทคนิควิธี DGA สำหรับการคัดเลือกตัวแบบการถดถอยเชิงเส้นพหุคูณ

1.6.2 เป็นแนวทางสำหรับการคัดเลือกตัวแบบการถดถอย เมื่อตัวแปรอิสระมีจำนวนมาก รวมถึงเป็นแนวทางในการพัฒนาเทคนิควิธีอื่นๆ สำหรับการคัดเลือก ตัวแบบการถดถอยเชิงเส้นพหุคูณ

1.7 สถานที่ที่ใช้ในการดำเนินการวิจัยและรวบรวมข้อมูล

1.7.1 สถานที่ใช้ในการดำเนินการวิจัย

- 1) ภาควิชาสถิติ คณะวิทยาศาสตร์ มหาวิทยาลัยเชียงใหม่
- 2) สำนักบริการเทคโนโลยีสารสนเทศ มหาวิทยาลัยเชียงใหม่
- 3) สำนักหอสมุด มหาวิทยาลัยเชียงใหม่

1.7.2 อุปกรณ์ในการดำเนินการวิจัย

1) คอมพิวเตอร์ส่วนบุคคล จำนวน 1 เครื่อง

1.1) CPU Intel(R) Core(TM) 2 Duo @ 2.10 GHz

1.2) RAM ขนาดความจุ 4 GB

1.3) เชื่อมต่อกับอินเทอร์เน็ตได้

2) ซอฟต์แวร์

2.1) โปรแกรม R Version 3.0.1 และ R Commander Version 1.9-6

2.2) โปรแกรม Microsoft Office 2010