บทที่ 1

บทนำ

1.1 ความสำคัญและความเป็นมาของปัญหา

ในการศึกษาว่าตัวแปรอิสระ (Independent Variables) มีผลทำให้ตัวแปรตามมีการเปลี่ยนแปลงอย่างไรนั้น อาจนำวิธีการทางสถิติมาใช้ ซึ่งวิธีหนึ่งที่ใช้ในการวิเคราะห์ความสำคัญของตัวแปรอิสระ คือ การวิเคราะห์ความถดถอยพหุ (Multiple regression analysis) โดยถือว่าการใช้ตัวแปรอิสระที่เหมาะสมมากกว่าหนึ่งตัว จะทำให้ผลของการประมาณค่าตัวแปรตามมีความถูกต้องมากกว่าการใช้ตัวแปรอิสระเพียงตัวเดียว แต่อย่างไรก็ตามจำเป็นที่จะต้องคำนึงถึงกรณีที่ตัวแปรอิสระมีพหุสัมพันธ์ (Multicollinearity) กัน ซึ่งหมายถึงการที่ตัวแปรอิสระมีความสัมพันธ์กันเอง ปัญหาดังกล่าวข้างต้นจะส่งผลกระทบรุนแรงต่อการประมาณค่าสัมประสิทธิ์ความถดถอยพหุ ($\underset{\sim}{\beta}$)

สำหรับตัวแบบทั่วไป (General Model) ที่แสดงความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามแบบเชิงเส้น (Linear Relationship) จะมีรูปแบบดังนี้

$$\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \qquad (1)$$

เมื่อ $\underset{\sim}{y}$ คือ เมตริกซ์ของตัวแปรตามขนาด n x 1

X คือ เมตริกซ์ของตัวแปรอิสระขนาด n x p

$\underset{\sim}{\beta}$ คือ เมตริกซ์ของสัมประสิทธิ์ความถดถอยพหุขนาด p x 1

$\underset{\sim}{\varepsilon}$ คือ เมตริกซ์ของความคลาดเคลื่อนขนาด n x 1

n คือ ขนาดตัวอย่าง

p คือ จำนวนตัวแปรอิสระ

ในการประมาณค่าสัมประสิทธิ์ความถดถอยพหุ โดยทั่วไปจะใช้วิธีกำลังสองน้อยที่สุด (Least squares method) ซึ่งเป็นวิธีที่นิยมใช้กันมาก ตัวประมาณ ($\underset{\sim}{\beta}$) ที่ได้เท่ากับ $(x'x)^{-1}x'y$ โดยตัวประมาณนี้จะเป็นตัวประมาณที่ไม่เอนเอียงและให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Mean square error) น้อยที่สุด ในบรรดาตัวประมาณที่ไม่เอนเอียงทั้งหลาย แต่ในการประมาณค่าสัมประสิทธิ์ความถดถอยพหุด้วยวิธีกำลังสองน้อยที่สุดนั้น ตัวแปรอิสระจะต้องไม่มีความสัมพันธ์ในลักษณะเชิงเส้น ซึ่งในทางปฏิบัติเป็นไปได้น้อยมาก เนื่องจากตัวแปรต่าง ๆ ที่นำมาศึกษาอาจมีความสัมพันธ์กัน ตัวแปรอิสระบางตัวอาจเป็นฟังก์ชันของตัวแปรอิสระตัวอื่น ๆ กรณีเช่นนี้กล่าวได้ว่าตัวแปรมีพหุสัมพันธ์กัน จะมีผลทำให้ตัวประมาณกำลังสองน้อยที่สุดที่ได้เอนเอียง และค่าเฉลี่ยความคลาดเคลื่อนกำลังสองไม่เป็นค่าที่ต่ำที่สุด นั่นคือ ค่าประมาณสัมประสิทธิ์การถดถอยพหุที่ได้ขาดความเที่ยงตรง (Accuracy) ดังนั้นวิธีการกำลังสองน้อยที่สุดจะแก้ไขปัญหานี้โดยการขจัดตัวแปรอิสระที่มีความสัมพันธ์กันออกไป แต่ในบางครั้งการตัดสินใจว่าตัวแปรอิสระตัวใดควรถูกคัดออกไปนั้นทำได้ยาก เนื่องจากลักษณะความสัมพันธ์ระหว่างตัวแปรอิสระที่เกิดขึ้นไม่ชัดเจนพอ และถือว่าตัวแปรอิสระทุกตัวต่างมีผลต่อการเปลี่ยนแปลงของตัวแปรตามมากพอสมควรซึ่งจากเหตุผลดังกล่าววิธีกำลังสองน้อยที่สุด จึงไม่เหมาะสมที่จะนำมาใช้ในกรณีที่เกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ

จากปัญหาข้างต้น นักสถิติได้ทำการศึกษาวิธีการประมาณค่าสัมประสิทธิ์ความถดถอยที่ให้ค่าเฉลี่ยความคลาดเคลื่อนของผลต่างกำลังสองต่ำกว่าวิธีกำลังสองน้อยที่สุด เมื่อเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ โดยให้ชื่อว่า ริดจ์รีเกรสชั่น (Ridge regression) ซึ่งทำการศึกษาโดย Hoerl and Kennard (1970) วิธีการนี้อาศัยหลักการที่ว่า ถ้า $|x'x|$ มีค่าเข้าใกล้ศูนย์ซึ่งทำให้มีปัญหาในการคำนวณ $(x'x)^{-1}$ และผลบวกกำลังสองของ $\underset{\sim}{\beta}$ ซึ่งเท่ากับ $\underset{\sim}{\beta}'\underset{\sim}{\beta}$ มีค่ามากกว่าความเป็นจริง ดังนั้นการที่จะทำให้ผลบวกกำลังสองของ $\underset{\sim}{\beta}$ มีค่าลดลง โดยการทำให้ $(x'x)^{-1}$ มีค่าเพิ่มขึ้นโดยการบวกค่าคงที่ k ที่มากกว่าศูนย์เข้ากับสมาชิกทุกตัวบนเส้นทแยงมุมของ $(x'x)$ ซึ่งมีผลทำให้คาแรคเตอร์ริสติกรูทธ์ (Characteristic Root) ของ $(x'x)$ มีค่ามากขึ้นและผลบวกกำลังสองของ $\underset{\sim}{\beta}$ มีค่าลดลง ตัวประมาณค่าสัมประสิทธิ์ความถดถอยพหุ โดยวิธีริดจ์รีเกรสชั่นเป็นได้ดังนี้

$$\underset{\sim}{\hat{\beta}}_R = (x'x + kI)^{-1}x'y \qquad ; k > 0 \qquad (2)$$

สมการ (2) ใช้หลักการคำนวณเหมือนกับวิธีกำลังสองน้อยที่สุด แต่จะแตกต่างกันที่เมตริกซ์ของตัวแปรอิสระ $(x'x)$ Hoerl and Kennard ได้กล่าวว่าตัวประมาณค่าสัมประสิทธิ์ความถดถอยพหุที่ได้จากวิธีริดจ์รีเกรสชั่นนี้ จะมีลักษณะค่อนข้างคงที่ ค่าสัมบูรณ์ของตัวประมาณมีค่าสมเหตุสมผลและเครื่องหมายของค่าประมาณสัมประสิทธิ์ความถดถอยพหุจะถูกต้อง

นอกจากวิธีริดจ์รีเกรสชั่นนี้แล้วยังมีอีกวิธีการหนึ่งที่น่าสนใจ คือ วิธีลาเท้นรูทธ์รีเกรสชั่น (Latent root regression) ซึ่งคิดค้นโดย Gunst , Webster และ Mason (1978) วิธีการนี้จะกำหนดให้เมตริกซ์ A = [y : x] และทำการคำนวณค่าคาแรคเตอร์ริสติกรูทธ์ หรือค่าลาเท้นรูทธ์ (Latent root) จากเมตริกซ์ ÁA นำค่าลาเท้นรูทธ์ และ ลาเท้นเวกเตอร์ (Latent vectors) ที่ได้มาเปรียบเทียบกับเกณฑ์ที่กำหนดไว้ ถ้าค่าลาเท้นรูทธ์ และ ลาเท้นเวกเตอร์ มีค่าต่ำกว่าเกณฑ์จะถูกคัดออก ซึ่งมีผลทำให้ตัวประมาณสัมประสิทธิ์ความถดถอยพหุมีความถูกต้องและใกล้เคียงกว่าตัวประมาณสัมประสิทธิ์ความถดถอยพหุที่ได้จากวิธีกำลังสองน้อยที่สุด

วิธีการคำนวณค่าสัมประสิทธิ์ความถดถอยพหุทั้งสองวิธีที่กล่าวมานี้ไม่จำเป็นต้องคัดตัวแปรอิสระออกจากตัวแบบ แม้ว่าตัวแปรอิสระนั้นจะมีพหุสัมพันธ์กันก็ตาม แต่ตัวประมาณที่ได้จากทั้ง 2 วิธีจะเป็นตัวประมาณที่เอนเอียง

ถ้าค่าสังเกตสุ่มมาจากประชากรที่มีการแจกแจงแบบเบ้ ตัวประมาณริดจ์รีเกรสชั่นและวิธีลาเท้นรูทธ์รีเกรสชั่นอาจแตกต่างไปจากกรณีที่ค่าสังเกตของประชากรมีการแจกแจงแบบปกติ เนื่องจากทั้งสองวิธีการเป็นตัวประมาณที่ถูกดัดแปลงมาจากตัวประมาณกำลังสองน้อยที่สุด และวิธีการกำลังสองน้อยที่สุดเป็นวิธีการที่ไวต่อข้อมูลที่ผิดปกติ ดังนั้นจึงเป็นสิ่งที่น่าสนใจที่จะศึกษาว่าวิธีการริดจ์รีเกรสชั่นและวิธีลาเท้นรูทธ์รีเกรสชั่น วิธีใดให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของตัวประมาณสัมประสิทธิ์การถดถอยพหุน้อยที่สุด

1.2 วัตถุประสงค์ของการวิจัย

1. ศึกษาและเปรียบเทียบ ตัวประมาณกำลังสองน้อยที่สุด ตัวประมาณริดจ์รีเกรสชั่นและตัวประมาณลาเท้นรูทธ์รีเกรสชั่น เมื่อความคลาดเคลื่อนมีการแจกแจงแบบปกติ การแจกแจงแบบปกติปลอมปน และการแจกแจงแบบลอกนอร์มอล

2. เพื่อเป็นแนวทางในการเลือกใช้ตัวประมาณสัมประสิทธิ์ความถดถอยพหุ เมื่อข้อมูลเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ

1.3 สมมติฐานของการวิจัย

โดยทั่วไป ตัวประมาณค่าสัมประสิทธิ์ความถดถอยพหุ โดยวิธีลาเท้นรูทธ์ จะให้ค่าตัวประมาณได้ถูกต้องและแม่นยำกว่า วิธีริดจ์รีเกรสชั่น

1.4 ข้อตกลงเบื้องต้น

ในการวิจัยครั้งนี้เกณฑ์ที่ใช้ในการเปรียบเทียบตัวประมาณ คือ ค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Mean Square Error) ของการประมาณสัมประสิทธิ์การถดถอยพหุด้วยวิธีกำลังสองน้อยที่สุด วิธีการริดจ์รีเกรสชั่นและวิธีการลาเท้นรูทธ์รีเกรสชั่น

1.5 ขอบเขตของการวิจัย

1.5.1 ทำการศึกษาในกรณีที่รูปแบบของการแจกแจงของความคลาดเคลื่อนมีลักษณะการแจกแจง ดังนี้

1.5.1.1 เมื่อความคลาดเคลื่อนมีการแจกแจงแบบปกติ (Normal distribution) ฟังก์ชันความหนาแน่นอยู่ในรูปของ

$$f(x) = \begin{cases} \dfrac{1}{\sigma\sqrt{2\pi}} \exp \dfrac{-(x-\mu)^2}{2\sigma^2} & ; x > 0 \\ 0 & ; \text{อื่น ๆ} \end{cases}$$

ในการวิจัยครั้งนี้ จะศึกษาเมื่อค่าเฉลี่ย $\mu = 0$ และ $\sigma^2 = 1$

1.5.1.2 เมื่อความคลาดเคลื่อนมีการแจกแจงแบบปกติปลอมปน (Scale - contaminated distribution) ฟังก์ชั่นการแจกแจงอยู่ในรูปของ

$$F(x) = (1-p)N(0, \sigma^2)+pN(0, c^2\sigma^2)$$

เมื่อ c คือ สเกลแฟคเตอร์ (scale factor) โดยที่สเกลมีค่าสูงจะทำให้เกิดค่าสังเกตที่ผิดปกติมีค่าสูงด้วย ในการวิจัยครั้งนี้จะศึกษาที่ค่า c = 3 และ 10

เมื่อ p คือ เปอร์เซนต์การปลอมปน (percent of contamination) ในการวิจัยครั้งนี้จะศึกษาที่ค่า p = 5 และ 10 %

1.5.1.2 เมื่อความคลาดเคลื่อนมีการแจกแจงแบบลอกนอร์มอล (Lognormal distribution) ฟังก์ชั่นความหนาแน่นอยู่ในรูปของ

$$f(x) = \begin{cases} \dfrac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\{\ln(x-\mu)^2/2\sigma^2\right] ; x > 0,\ \sigma > 0 \\ 0 \qquad\qquad ; \text{อื่น ๆ} \end{cases}$$

ในการวิจัยครั้งนี้ จะศึกษาเมื่อค่าเฉลี่ย $\mu = 0$ และความแปรปรวน เท่ากับ 1

1.5.2 ขนาดตัวอย่างที่ใช้ในการศึกษา กำหนดให้มีขนาดดังนี้ 10, 30 และ 50

1.5.3 จำนวนตัวแปรอิสระ ซึ่งนำมาพิจารณาในการวิจัย เท่ากับ 3 และ 5

1.5.4 กำหนดค่าสหสัมพันธ์ระหว่างตัวแปรอิสระ เพื่อบังคับให้ตัวแปรอิสระเกิดพหุสัมพันธ์กัน ในระดับที่ต้องการ โดยจัดแบ่งระดับของการเกิดพหุสัมพันธ์เป็น 9 ช่วงดังนี้ [0.11-0.20] , [0.21-0.30] , [0.31-0.40] , [0.41-0.50] , [0.51-0.60] , [0.61-0.70] , [0.71-0.80] , [0.81-0.90] และ [0.91-1.00] ทั้งนี้ เพื่อให้ได้ผลสรุปที่มีความละเอียดและมีความถูกต้องมากขึ้น อีกทั้งจะช่วยให้สามารถเห็นแนวโน้มของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของตัวประมาณแต่ละวิธี เมื่อระดับความสัมพันธ์เปลี่ยนแปลงไป

1.5.5 ในการวิจัยครั้งนี้ จะสร้างแบบจำลองข้อมูลให้มีสถานการณ์ตามต้องการโดยอาศัยเครื่องคอมพิวเตอร์เขียนด้วยโปรแกรม Fortran 77 และจำลองข้อมูลโดยอาศัยหลักการของ Multinormal ซึ่งจะได้ค่าตัวแปรอิสระที่มีความสัมพันธ์กันในระดับต่าง ๆ ทั้งนี้จะทำการทดลอง 1,000 รอบ ในแต่ละสถานการณ์

1.5.6 ในการคำนวณค่าเฉลี่ยความคลาดเคลื่อนกำลังสองและตัวประมาณลาเท้นรุทช์ รีเกรสชั่น เกณฑ์ที่ใช้ในการตัดค่าลาเท้นรุทช์ และลาเท้นเวกเตอร์ เพื่อคำนวณค่าดังกล่าว คือ เมื่อค่าลาเท้นรุทช์ มากกว่าหรือเท่ากับ 0.05 และค่าลาเท้นเวกเตอร์ มากกว่าหรือเท่ากับ 0.10

1.6 ประโยชน์ที่คาดว่าจะได้รับ

เพื่อเป็นแนวทางในการเลือกใช้ตัวประมาณสัมประสิทธิ์ความถดถอยที่เหมาะสมในกรณีเกิดพหุสัมพันธ์ ภายใต้การแจกแจงของความคลาดเคลื่อนแบบปกติหรือแบบเบ้ โดยที่ขนาดตัวอย่าง จำนวนตัวแปรอิสระ และค่าสหสัมพันธ์มีค่าต่าง ๆ กัน