บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การวิเคราะห์การถดถอย เป็นวิธีการวิเคราะห์ข้อมูลที่ใช้กันแพร่หลายทั้งทางด้านวิทยาศาสตร์ และสังคมศาสตร์ การวิเคราะห์การถดถอยนี้เป็นวิธีการทางสถิติที่จะนำมาใช้เพื่อศึกษาหาความสัมพันธ์ระหว่างตัวแปรตาม กับตัวแปรอิสระ และประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุ วิธีที่นิยมใช้กันมากที่สุดในการประมาณค่าสัมประสิทธิ์ คือ วิธีกำลังสองน้อยที่สุด (Ordinary Least Square Method : OLS) ซึ่งเป็นวิธีที่ให้ตัวประมาณเชิงเส้นที่ไม่เอนเอียงและมีความแปรปรวนต่ำที่สุด (Minimum Variance Unbiased Estimator) ภายใต้ข้อตกลงเบื้องต้น คือ ตัวแปรอิสระแต่ละตัวจะต้องไม่มีความสัมพันธ์เชิงเส้นซึ่งกันและกัน

เมื่อตัวแปรอิสระต่าง ๆ ที่นำมาศึกษามีความสัมพันธ์กัน คือเกิดปัญหาพหุสัมพันธ์ (Multicollinearity) จะส่งผลให้ค่าประมาณสัมประสิทธิ์การถดถอยเชิงเส้นพหุด้วยวิธีกำลังสองน้อยที่สุดมีความเอนเอียงและค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Mean Square Error: MSE) มีค่าสูง เพื่อแก้ไขปัญหาดังกล่าวข้างต้นจึงได้มีผู้เสนอวิธีการ และพัฒนาวิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุเมื่อตัวแปรอิสระมีพหุสัมพันธ์กัน

กันสท์ และ เมสัน (Gunst and Mason, 1977) ได้เสนอวิธีการถดถอยองค์ประกอบหลัก (Principal Component Regression : PC) มาใช้ในการคำนวณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุเมื่อเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ โดยการจัดรูปแบบตัวแปรอิสระใหม่ก่อนที่จะนำไปหาค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุโดยให้ตัวแปรใหม่เป็นผลบวกเชิงเส้นของตัวแปรเดิม ซึ่งเป็นการลดผลกระทบของการเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ มีผลทำให้ตัวประมาณสัมประสิทธิ์การถดถอยเชิงเส้นพหุมีความถูกต้องมากขึ้นจึงช่วยให้ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองของสัมประสิทธิ์มีค่าต่ำกว่าวิธีกำลังสองน้อยที่สุด ถึงแม้ว่าตัวประมาณที่ได้จากวิธีนี้มีคุณสมบัติเป็นตัวประมาณที่เอนเอียง (biased estimator)

ครอส, จิน และฮานูมารา (Crouse, Jin and Hanumara, 1995) ได้พัฒนาวิธีริดจ์ รีเกรสชัน ที่ใช้ร่วมกับค่าเบื้องต้น ของสวินเดลให้มีประสิทธิภาพ โดยเสนอการประมาณค่า

k นำมาปรับค่าในสูตรประมาณค่าสัมประสิทธิ์กรณีตัวแปรอิสระมีพหุสัมพันธ์กัน ซึ่งสามารถหาค่า k ที่แน่นอนได้ และค่าประมาณที่ได้มีความแกร่ง ซึ่งนำไปใช้ในการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุในกรณีที่ตัวแปรอิสระมีพหุสัมพันธ์กัน นอกจากนี้ยังได้เปรียบเทียบวิธีการประมาณค่าระหว่าง วิธีกำลังสองน้อยที่สุด , วิธีริดจ์ รีเกรสชัน ที่ประมาณค่า k โดย โฮเอร์ล เคนนาร์ค และบาร์ดวิน วิธีริดจ์ รีเกรสชัน ที่ใช้ร่วมกับค่าเบื้องต้น (J) เมื่อ J คือเวกเตอร์ของค่าประมาณเบื้องต้นของค่าเฉลี่ยสัมประสิทธิ์การถดถอยเชิงเส้นพหุจากวิธีกำลังสองน้อยที่สุด

$$J=\left(\sum_{j=1}^{p}\frac{\hat{\beta}_{LS_j}}{p}\right)1_{p\times 1}$$ พบว่าวิธีริดจ์ รีเกรสชัน จะให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของสัมประสิทธิ์การถดถอยมีค่าน้อยกว่าวิธีกำลังสองน้อยที่สุด และตัวประมาณค่าที่ได้มีคุณสมบัติไม่เอนเอียง

มูนิซ และ ไคเบรีย (Muniz and Kibria, 2009) ได้เสนอวิธีการประมาณค่า k ของวิธีริดจ์ รีเกรสชัน ซึ่งนำไปใช้ในการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุในกรณีที่ตัวแปรอิสระมีพหุสัมพันธ์กัน นอกจากนี้ยังได้เปรียบเทียบวิธีการประมาณค่าระหว่าง วิธีกำลังสองน้อยที่สุด ,วิธีริดจ์รีเกรสชัน ที่ประมาณค่า k โดย โฮเอร์ล และ เคนนาร์ค วิธีริดจ์ รีเกรสชัน ที่ประมาณค่า k โดย ไคเบรีย (Kibria 2003) วิธีริดจ์ รีเกรสชัน ที่ประมาณค่า k โดย คาลาฟ และ ชูเคอร์ (Khalaf and Shukur, 2005) พบว่าวิธีริดจ์ รีเกรสชัน ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของสัมประสิทธิ์การถดถอยมีค่าน้อยกว่าวิธีกำลังสองน้อยที่สุด เมื่อใช้ค่า

$$k_M=\text{median}\left(\frac{1}{\sqrt{\frac{t_{max}\hat{\sigma}^2}{(n-p)\hat{\sigma}^2+t_{max}\hat{\alpha}_j^2}}}\right)$$ และให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของตัวประมาณค่าสัมประสิทธิ์การถดถอยมีค่าต่ำที่สุด

จะเห็นได้ว่าทั้ง วิธีการถดถอยองค์ประกอบหลัก วิธีริดจ์ รีเกรสชันที่มีค่าเบื้องต้น และวิธีมูนิซและไคเบรียริดจ์ รีเกรสชัน เป็นวิธีที่ใช้ในการแก้ปัญหาตัวแปรอิสระมีพหุสัมพันธ์กัน แต่ยัง

ไม่ได้มีการเปรียบเทียบวิธีการทั้ง 3 วิธีว่าวิธีการใดให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำสุด ดังนั้นผู้วิจัยจึงสนใจศึกษาเปรียบเทียบวิธีริดจ์ รีเกรสชัน ที่มีค่าเบื้องต้น, วิธีการถดถอยองค์ประกอบหลัก และวิธีมูนิซและไคเบรียริดจ์ รีเกรสชัน ว่าควรเลือกใช้วิธีการใดจึงเหมาะสม

1.2 วัตถุประสงค์ของการวิจัย

การศึกษาครั้งนี้มีวัตถุประสงค์เพื่อศึกษาเปรียบเทียบหาวิธีการประมาณค่าที่มีประสิทธิภาพสูงสุด ระหว่างวิธีการถดถอยองค์ประกอบหลัก วิธีริดจ์ รีเกรสชันที่มีค่าเบื้องต้น (Ridge Regression with Prior Information) และวิธีมูนิซและไคเบรียริดจ์ รีเกรสชัน ปัจจัยที่กำหนดในการศึกษามี 4 ปัจจัยคือความแปรปรวนของค่าคลาดเคลื่อนสุ่ม 4 ระดับ ขนาดตัวอย่าง 5 ขนาด ระดับพหุสัมพันธ์ระหว่างตัวแปรอิสระ 4 ระดับ และค่าสัมประสิทธิ์การถดถอย 6 แบบ

1.3 สมมติฐานของการวิจัย

สมมติฐานหลักของการศึกษามีดังนี้

1. ในกรณีที่ตัวแปรอิสระมีระดับพหุสัมพันธ์กันสูง วิธีมูนิซและไคเบรียริดจ์ รีเกรสชัน จะมีค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ยและค่าเฉลี่ยความเอนเอียงของตัวประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุน้อยกว่าวิธีอื่น ๆ
2. ในกรณีที่ความแปรปรวนมีค่ามาก วิธีมูนิซและไคเบรียริดจ์ รีเกรสชัน จะมีค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ยและค่าเฉลี่ยความเอนเอียงของตัวประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุน้อยกว่าวิธีอื่น ๆ
3. ในกรณีที่ขนาดตัวอย่างใหญ่ วิธีมูนิซและไคเบรียริดจ์ รีเกรสชัน จะมีค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ยและค่าเฉลี่ยความเอนเอียงของตัวประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุน้อยกว่าวิธีอื่น ๆ

1.4 ขอบเขตของการวิจัย

1. ศึกษากรณีที่มีตัวแปรอิสระ 6 ตัวแปร
2. ศึกษากรณีที่มีขนาดตัวอย่าง (n) เป็น 10,20,30,50,100
3. ศึกษาภายใต้ลักษณะการแจกแจงของความคลาดเคลื่อนสุ่มมีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเป็น 0 ความแปรปรวนเท่ากับ 0.25, 1, 9 และ 25
4. กำหนดสัมประสิทธิ์การถดถอยเชิงเส้นพหุ (β)

$$\beta = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 5 \\ 5 \\ 5 \end{bmatrix}, \begin{bmatrix} 5 \\ 1 \\ 1 \\ 1 \\ 5 \\ 5 \\ 5 \end{bmatrix}, \begin{bmatrix} 1 \\ 5 \\ 1 \\ 1 \\ 5 \\ 5 \\ 5 \end{bmatrix}, \begin{bmatrix} 5 \\ 5 \\ 5 \\ 5 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 5 \\ 5 \\ 5 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 5 \\ 1 \\ 5 \\ 5 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

5. ระดับพหุสัมพันธ์ระหว่างตัวแปรอิสระที่สนใจศึกษาแบ่งเป็น 3 ระดับคือ
   - 5.1 ระดับต่ำ $\rho_{ij} = 0.3$
   - 5.2 ระดับกลาง $\rho_{ij} = 0.5, 0.7$
   - 5.3 ระดับสูง $\rho_{ij} = 0.9$

   เมื่อ ρ_{ij} แทนสัมประสิทธิ์สหสัมพันธ์ระหว่าง X_i กับ X_j $(1 \leq i \neq j \leq 3)$
6. การจำลองค่าทำซ้ำ 1,000 ครั้ง ในแต่ละสถานการณ์

1.5 เกณฑ์ที่ใช้ในการเปรียบเทียบ

เกณฑ์ที่ใช้ในการเปรียบเทียบว่าวิธีในการหาตัวประมาณสัมประสิทธิ์การถดถอยเชิงเส้นพหุวิธีใดจะมีประสิทธิภาพมากที่สุด คือค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ย (Average Mean Square Error : AMSE) และค่าเฉลี่ยความเอนเอียงของตัวประมาณค่าสัมประสิทธิ์การ

ถดถอยเชิงเส้นพหุ โดยวิธีที่ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองเฉลี่ยและค่าเฉลี่ยความเอนเอียงต่ำที่สุดจะเป็นวิธีที่มีประสิทธิภาพมากที่สุด

1.6 ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

เพื่อใช้เป็นเกณฑ์ในการตัดสินใจเลือกวิธีการประมาณค่าสัมประสิทธิ์การถดถอยที่มีประสิทธิภาพมากที่สุดให้เหมาะกับสถานการณ์ เมื่อเกิดปัญหาตัวแปรอิสระมีพหุสัมพันธ์กันในตัวแบบเชิงเส้น

1.7 สัญลักษณ์และนิยามศัพท์

$tr(X'X)$ คือ ผลบวกของสมาชิกทุกตัวบนแนวทแยงมุมหลักของเมทริกซ์ $X'X$

λ คือ ค่าไอเกนของ $X'X$ ที่คำนวณจาก $|X'X - \lambda I| = 0$ โดยที่ $\lambda_1, \lambda_2, ..., \lambda_p$ คือ ค่าไอเกนของ $X'X$ เรียงลำดับจากมากไปน้อย

$\hat{\beta}_{LS}$ คือ เวกเตอร์ของตัวประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุของวิธีกำลังสองน้อยที่สุด

$\hat{\beta}_{PC}$ คือ เวกเตอร์ของตัวประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุของวิธีการถดถอยองค์ประกอบหลัก

$\hat{\beta}_{RJ}$ คือ เวกเตอร์ของตัวประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุของวิธีริดจ์รีเกรสชันที่มีค่าเบื้องต้น เสนอโดย ครอส, จิน และฮานูมารา

$\hat{\beta}_{RM}$ คือ เวกเตอร์ของตัวประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุของวิธีมูนิซและไคเบรียริดจ์รีเกรสชัน