บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในการศึกษาการเปลี่ยนแปลงของตัวแปรตาม(dependent variable) ว่ามีผลมาจากตัวแปรอิสระ(independent variable) ชุดหนึ่งอย่างไรนั้น วิธีหนึ่งที่ใช้ในการวิเคราะห์ความสำคัญของตัวแปรอิสระคือ การวิเคราะห์ความถดถอยพหุ(multiple regression analysis) ซึ่งเป็นกรณีหนึ่งของการวิเคราะห์ความถดถอยเชิงเส้น การวิเคราะห์ความถดถอยพหุมีหลักเกณฑ์ว่า การใช้ตัวแปรอิสระที่เหมาะสมมากกว่าหนึ่งตัว โดยทั่วๆ ไปย่อมทำให้ผลการประมาณค่าตัวแปรตามมีความถูกต้องมากกว่าการใช้ตัวแปรอิสระเพียงตัวเดียว สำหรับตัวแบบทั่วไป(general model) ของความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามเชิงเส้น สามารถเขียนได้ดังนี้

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

เมื่อ $\underset{\sim}{y}$ เป็นเวกเตอร์ของตัวแปรตามขนาด nx1, n คือจำนวนค่าสังเกต

X เป็นเมตริกซ์ของตัวแปรอิสระขนาด nx(p+1)และ $X'X$ มี full rank = p+1

$\underset{\sim}{\beta}$ เป็นเวกเตอร์ของสัมประสิทธิ์การถดถอยพหุขนาด (p+1)x1,

p = จำนวนตัวแปรอิสระ - 1

และ $\underset{\sim}{\varepsilon}$ เป็นเวกเตอร์ของค่าความคลาดเคลื่อนที่เกิดขึ้นขนาด nx1 ซึ่งมีการแจกแจงแบบปกติ โดยที่ $E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$ และ $E(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}') = \sigma^2 I_n$

ในการประมาณค่าสัมประสิทธิ์การถดถอยพหุจากตัวแบบดังกล่าวนี้ วิธีที่นิยมใช้กันมาก คือวิธีกำลังสองน้อยที่สุด(least square method) ซึ่งจะได้ว่า $\hat{\beta} = (X'X)^{-1}X'y$ เป็นตัวประมาณที่ไม่เอนเอียง ซึ่งมีค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำสุดในบรรดาตัวประมาณที่ไม่เอนเอียงเชิงเส้น แต่ในการประมาณค่าสัมประสิทธิ์การถดถอยพหุด้วยวิธีกำลังสองน้อยที่สุดมีสมมุติฐานที่จำเป็นข้อหนึ่งคือ ตัวแปรอิสระต้องไม่มีความสัมพันธ์กันในลักษณะเชิงเส้น ซึ่งกรณีนี้ในทางปฏิบัติเป็นไปได้น้อย เพราะตัวแปรอิสระที่นำมาศึกษาอาจมีความสัมพันธ์กันอยู่ โดยที่ตัวแปรอิสระบางตัวอาจเป็นฟังก์ชันของตัวแปรอิสระตัวอื่น กล่าวคือ ตัวแปรอิสระมีพหุสัมพันธ์ (multicollinearity) ทำให้การประมาณค่าตัวแปรตามที่ได้อาจไม่เหมาะสม และมีผลทำให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าประมาณสัมประสิทธิ์การถดถอยพหุมีค่ามาก นั่นคือ ค่าประมาณสัมประสิทธิ์การถดถอยพหุที่ได้ขาดความเที่ยงตรง(accuracy) ดังนั้นในการกำหนดรูปแบบของตัวแปรตามว่ามีผลมาจากตัวแปรอิสระใดบ้าง จำเป็นต้องคำนึงว่าตัวแปรอิสระมีพหุสัมพันธ์กันสูงหรือไม่ ถ้าตัวแปรอิสระมีพหุสัมพันธ์กันสูงอาจจะต้องคัดตัวแปรอิสระบางตัวออกไป ในบางกรณีความสัมพันธ์ระหว่างตัวแปรอิสระไม่ชัดเจน ทำให้การตัดตัวแปรอิสระตัวใดออกจากตัวแบบเป็นไปได้ยาก หรืออาจไม่ต้องการคัดตัวแปรอิสระตัวใดออกจากตัวแบบ เพราะถือว่าตัวแปรอิสระทุกตัวมีผลต่อการเปลี่ยนแปลงของตัวแปรตามมากพอสมควร

Hoerl และ Kennard(1970:55) ได้ศึกษาวิธีการประมาณค่าสัมประสิทธิ์ความถดถอยพหุที่ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำกว่าวิธีกำลังสองน้อยที่สุด ซึ่งเรียกว่าวิธีริดจ์รีเกรสชัน วิธีการนี้เป็นวิธีที่แก้ปัญหา multicollinearity โดยหลักการดังนี้ ถ้า $|X'X|$ มีค่าเข้าใกล้ศูนย์ซึ่งทำให้ $(X'X)^{-1}$ หาค่าไม่ได้และผลบวกกำลังสองของ $\hat{\beta}$ ซึ่งเท่ากับ $\hat{\beta}'\hat{\beta}$ มีค่ามากกว่าความจริง โดยที่ $\hat{\beta} = (X'X)^{-1}X'y$ ดังนั้นการที่จะทำให้ผลบวกกำลังสองของ $\hat{\beta}$ มีค่าลดลงโดยการทำให้ $(X'X)^{-1}$ มีค่าเพิ่มขึ้น กล่าวคือจะบวกค่าคงที่ k ที่มากกว่าศูนย์เข้ากับสมาชิกทุกตัวบนเส้นทแยงมุมของ $X'X$ ซึ่งจะทำให้ characteristic root ของ $X'X$ มีค่ามากขึ้น

และผลบวกกำลังสองของ $\hat{\beta}$ มีค่าลดลง* ดังนั้นเราจะได้ค่าประมาณสัมประสิทธิ์การถดถอยพหุของวิธีริดจ์รีเกรสชัน ดังนี้

$$(2) \qquad \tilde{\beta} = (X'X + kI)^{-1}X'y \quad ; k > 0$$

สมการ(2)ใช้หลักการคำนวณเหมือนกับวิธีกำลังสองน้อยที่สุด แต่จะแตกต่างกันที่เมตริกซ์ของตัวแปรอิสระ $(X'X)$

จากการศึกษาของ Wichern และ Churchill, Gibbon, McDonald และ Galarneau, TZE-SAN-LEE พบว่าวิธีที่ดีในการประมาณค่า k มีอยู่ 3 วิธีคือ

1. วิธี HKB(Hoerl-Kennard-Baldwin Method)

วิธีนี้จะใช้ $k = p\hat{\sigma}^2/\hat{\beta}'\hat{\beta}$

เมื่อ p = จำนวนตัวแปรอิสระ

$\hat{\beta} = (X'X)^{-1}X'y$ เป็นตัวประมาณกำลังสองน้อยที่สุด

และ $\hat{\sigma}^2 = (1/n-p)(y'y - \hat{\beta}'X'y)$

---

*เมื่อ $\hat{\beta}$ เป็นค่าประมาณของ β โดยวิธีกำลังสองน้อยที่สุดโดยที่ $\hat{\beta} = (X'X^{-1})X'y$ และ $E(\hat{\beta}'\hat{\beta}) = \beta'\beta + \sigma^2 \text{trace}(X'X)^{-1}$ นั่นคือ $E(\hat{\beta}'\hat{\beta}) = \beta'\beta + (\sigma^2/\lambda_1 + \sigma^2/\lambda_2 + \ldots + \sigma^2/\lambda_p)$ แสดงว่าผลบวกกำลังสองของ $\hat{\beta}$ มากกว่าผลบวกกำลังสองของ β อยู่ $\sigma^2 \text{trace}(X'X)^{-1}$ การที่จะทำให้ผลบวกกำลังสองของ $\hat{\beta}$ ลดลงก็โดยการเพิ่มค่า eigenvalue (λ) ของเมตริกซ์ $(X'X)$ โดยการบวกค่าคงที่ k ที่มากกว่าศูนย์กับค่าบนเส้นทแยงมุมของเมตริกซ์ $(X'X)$

2. วิธี TZE-SAN-LEE(TZE-SAN-LEE Method)

วิธีนี้จะใช้ $k = \lambda_p$

เมื่อ λ_p เป็นค่า eigenvalue ของ$(X'X)$ที่มีค่าน้อยที่สุด

3. วิธี McD&G(McDonald-Galarneau Method)

วิธีนี้จะเลือกค่า k จากสมการ

$$\underline{\beta}'(k)\underline{\beta}(k) = \underline{\hat{\beta}}'\underline{\hat{\beta}} - \sigma^2 \text{trace}(X'X)^{-1}$$

เมื่อ $\sigma^2 = (1/n-p)(\underline{y}'\underline{y}-\underline{\hat{\beta}}'X'\underline{y})$

ถ้าค่าสังเกตสุ่มมาจากประชากรที่มีการแจกแจงแบบเบ้ วิธีการประมาณค่า k ที่เหมาะสมอาจแตกต่างไปจากกรณีที่ค่าสังเกตของประชากรมีการแจกแจงแบบปกติ เนื่องจากวิธีริดจ์รีเกรสชันใช้หลักการคำนวณเหมือนกับวิธีกำลังสองน้อยที่สุด และวิธีกำลังสองน้อยที่สุดเป็นวิธีที่ไวต่อข้อมูลที่ผิดปกติ ดังนั้นจึงเป็นสิ่งน่าสนใจที่จะศึกษาวิธีการประมาณค่า k ในวิธีการของริดจ์รีเกรสชันว่าวิธีใดที่ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของตัวประมาณสัมประสิทธิ์การถดถอยพหุน้อยที่สุด

1.2 วัตถุประสงค์ของการวิจัย

1. ศึกษาและเปรียบเทียบ ตัวประมาณริดจ์ เมื่อความคลาดเคลื่อนมีการแจกแจงแบบปกติ

2. ศึกษาและเปรียบเทียบ ตัวประมาณริดจ์ เมื่อความคลาดเคลื่อนมีการแจกแจงแบบปกติปลอมปนและแบบเบ้

1.3 สมมุติฐานของการวิจัย

1. ภายใต้ลักษณะการแจกแจงของความคลาดเคลื่อนเป็นแบบปกติและขนาดตัวอย่างเดียวกัน การประมาณค่า k ด้วยวิธี HKB ดีที่สุด

2. ภายใต้ลักษณะการแจกแจงของความคลาดเคลื่อนเป็นแบบปกติปลอมปนและแบบเบ้และขนาดตัวอย่างเดียวกัน การประมาณค่า k ด้วยวิธี HKB ดีที่สุด

1.4 ข้อตกลงเบื้องต้น

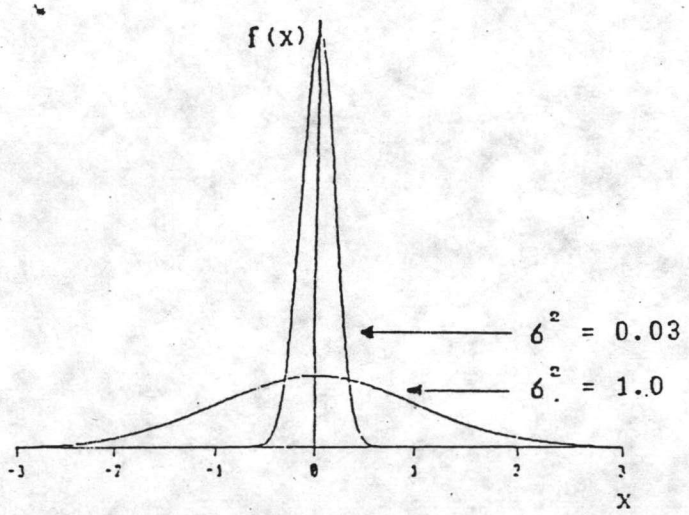
ในการวิจัยครั้งนี้ใช้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำสุดของการประมาณสัมประสิทธิ์การถดถอยพหุด้วยวิธีริดจ์รีเกรสชัน ซึ่งใช้เกณฑ์อัตราส่วนของผลต่างกำลังสองต่ำสุด(RDAMSE) ในการเปรียบเทียบ

1.5 ขอบเขตของการวิจัย

1. เมื่อความคลาดเคลื่อนมีการแจกแจงแบบปกติ(Normal distribution) ฟังก์ชันความหนาแน่นอยู่ในรูปของ

$$f(x) = \frac{1}{6\sqrt{2\pi}} \exp -\frac{1}{2}\frac{(x-\mu)^2}{6^2} \qquad ; 6>0$$

ในกรณีนี้จะศึกษาเมื่อ $\mu = 0$, $6^2 = 0.03$ และ 1.0 สาเหตุที่เลือกใช้ $6^2 = 0.03$ ด้วยเพราะต้องการพิจารณากรณีที่ข้อมูลมีการกระจายน้อย ตัวประมาณริดจ์วิธีใดจะให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำสุด



รูปที่ 1.5.1 แสดงเส้นโค้งของการแจกแจงแบบปกติ เมื่อ μ = 0 และ 6^2 = 0.03 และ 1.0

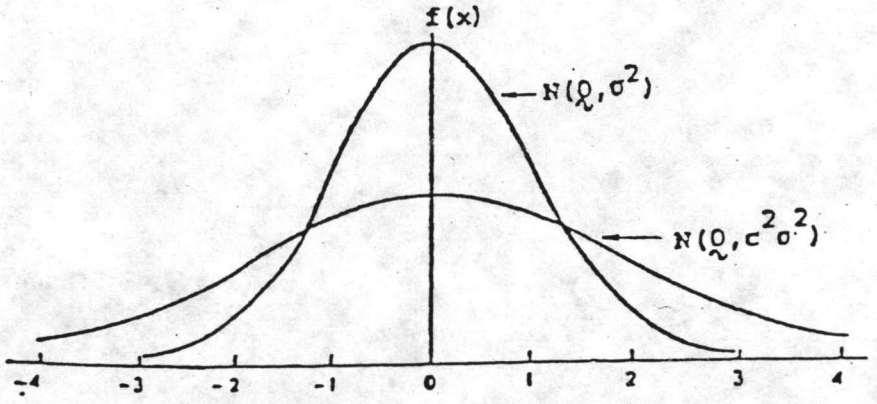
2. <u>เมื่อความผิดพลาดมีการแจกแจงแบบปกติปลอมปน</u>(Scale-contaminated normal distribution)

ฟังก์ชันการแจกแจงอยู่ในรูปของ

$$F = (1-p)N(\underline{0}, 6^2) + pN(\underline{0}, c^2 6^2)$$

เมื่อ c เป็นสเกลแฟคเตอร์(scale factor) โดยที่สเกลมีค่าสูงจะทำให้เกิดค่าสังเกตที่ผิดปกติมีค่าสูงด้วย ในที่นี้จะใช้ c = 3 และ 10

และ p คือเปอร์เซนต์การปลอมปน(percent of contamination) ซึ่งจะใช้ p = 5 และ 10



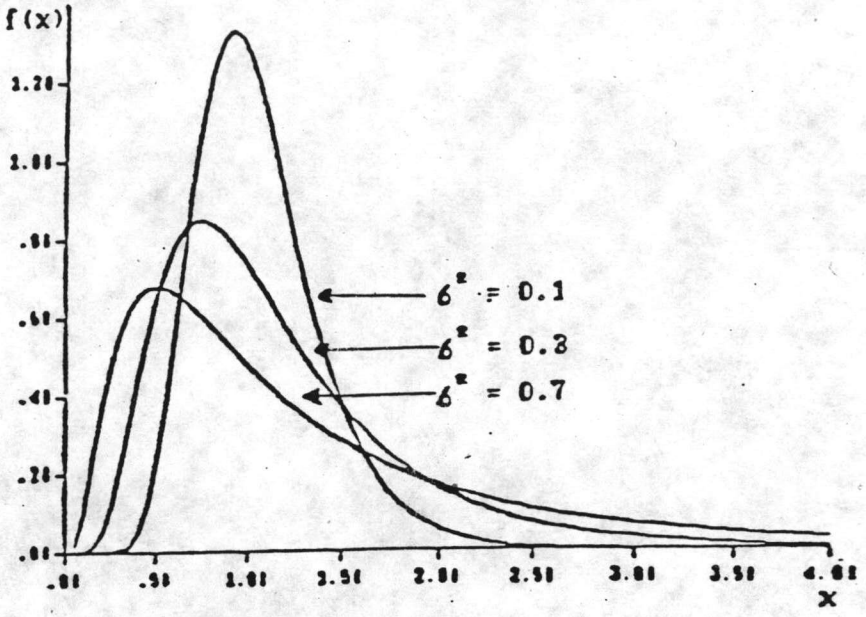
รูปที่ 1.5.2 แสดงเส้นโค้งการแจกแจงแบบปกติปลอมปน

3. <u>เมื่อความคลาดเคลื่อนมีการแจกแจงแบบลอกนอร์มอล</u>(Lognormal distribution)

ฟังก์ชันความหนาแน่นอยู่ในรูปของ

$$f(x) = \begin{cases} \dfrac{1}{x6\sqrt{2\pi}} \exp\left[-(\ln x - \mu)^2/26^2\right] & ; x>0, 6>0, \mu \in R \\ 0 & ; \text{อื่นๆ} \end{cases}$$

ผู้วิจัยสนใจศึกษา เมื่อค่าเฉลี่ย(μ)เท่ากับ 0 และความแปรปรวน(6^2)เท่ากับ 0.7, 0.3 และ 0.1 กล่าวคือ C.V.(x) = 100%,59% และ 22% ตามลำดับ และสาเหตุที่เลือกใช้ค่า C.V.=22% โดยที่ไม่เลือกค่า C.V.ที่มีค่าต่ำกว่านี้เพราะจากการพิจารณากราฟที่แสดงเส้นโค้งการแจกแจงแบบลอกนอร์มอล ถ้า C.V.มีค่าต่ำกว่านี้กราฟของการแจกแจงจะลู่เข้าสู่การแจกแจงแบบปกติมากขึ้น



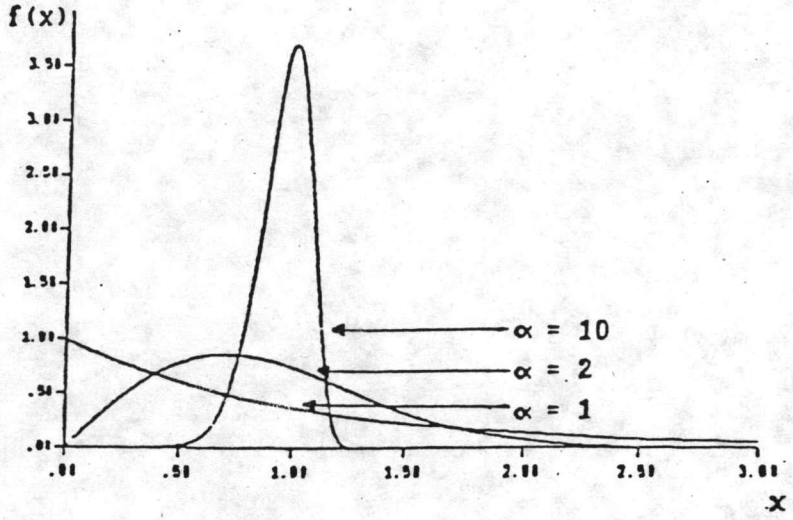
รูปที่ 1.5.3 แสดงเส้นโค้งการแจกแจงแบบลอกนอร์มอล เมื่อ $\mu = 0$ และ $\sigma^2 = 0.1, 0.3$ และ 0.7

4. <u>เมื่อความคลาดเคลื่อนมีการแจกแจงแบบไวบูลล์</u>(Weibull distribution) ฟังก์ชันความหนาแน่นอยู่ในรูปของ

$$f(x) = \begin{cases} \dfrac{\alpha x^{\alpha-1} \exp\left[(-x/\beta)^{\alpha}\right]}{\beta^{\alpha}} & ; x>0,\ \alpha>0,\ \beta>0 \\ 0 & ; \text{อื่นๆ} \end{cases}$$

เมื่อ β เป็น scale parameter

และ α เป็น shape parameter



รูปที่ 1.5.4 แสดงเส้นโค้งการแจงแจงแบบไวบูลล์ เมื่อ β = 1 และ α = 1,2 และ 10

ในการวิจัยครั้งนี้สนใจศึกษาเมื่อ β = 1 และ α = 1,2 และ 10 กล่าวคือ C.V.(x)= 100%,52% และ 17% ตามลำดับเหตุที่เลือกใช้ C.V. = 17% โดยไม่เลือกค่า C.V.ต่ำกว่านี้ด้วยเหตุผลเดียวกับการเลือกค่า C.V.ของการแจกแจงแบบลอกนอร์มอล

5. จำนวนตัวแปรอิสระและขนาดตัวอย่าง

จะศึกษาในกรณีที่ขนาดตัวอย่าง(n)เท่ากับ 10,30,50,100 และจำนวนตัวแปรอิสระเท่ากับ 3,5

6. การจำลองประชากร

ข้อมูลที่ใช้ในการศึกษาใช้วิธีจำลองของ Wichern และ Churchill(1978:304) เพื่อใช้กำหนดความสัมพันธ์ของตัวแปรอิสระ ในกรณีที่มีตัวแปรอิสระเท่ากับ 5 จะศึกษา ที่ระดับ

ความสัมพันธ์(.99,.99),(.99,.90),(.70,.30) และเมื่อตัวแปรอิสระเท่ากับ 3 จะศึกษาที่ระดับความสัมพันธ์(.99),(.90),(.70) ในกรณีที่การแจกแจงของความคลาดเคลื่อนเป็นแบบปกติและแบบปกติปลอมปนจะจำลองประชากรจากตัวแบบ

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}$$

โดยที่ $\underline{\beta}$ เป็นเวกเตอร์ของสัมประสิทธิ์การถดถอยพหุ และ $\underline{\varepsilon}$ มีการแจกแจงตามที่ต้องการ เมื่อการแจกแจงของความคลาดเคลื่อนเป็นแบบเบ้ จะสร้าง $\underline{y}$ ให้มีการแจกแจงแบบเบ้โดยตรง โดยที่เมตริกซ์ X และ $\underline{\beta}$ จะถูกสร้างในทำนองเดียวกันกับที่กล่าวมา

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1. ผลการศึกษาเป็นแนวทางในการศึกษาเกี่ยวกับการประมาณค่าสัมประสิทธิ์การถดถอยพหุในกรณีที่ตัวแปรอิสระมีพหุสัมพันธ์กัน

2. ผลการศึกษาเปรียบเทียบจะสามารถสรุปได้ว่าควรเลือกใช้วิธีการประมาณค่า k(ridge estimators)วิธีใด