บทที่ 3

วิธีดำเนินการวิจัย

ข้อมูลที่ใช้ในการวิจัยเรื่องนี้ ได้จากการจำลอง (Simulation) ขึ้นโดยเครื่องคอมพิวเตอร์ ทั้งนี้เพื่อจะสามารถกำหนดความสัมพันธ์ระหว่างตัวแปรอิสระแต่ละคู่ได้ในลักษณะต่าง ๆ โดยใช้วิธีจำลองของ Wichern and Churchill (1978 : 304) ซึ่งจะได้ค่าตัวแปรอิสระที่มีความสัมพันธ์กันในระดับต่าง ๆ

ตัวอย่างเช่น ถ้าต้องการสร้างข้อมูลชุดหนึ่งให้มีตัวแปรอิสระ 4 ตัว และขนาดตัวอย่างเท่ากับ 30 จะได้ค่าต่าง ๆ ของตัวแปรอิสระจากสมการ

$$X_{ij} = (1 - \alpha^2)^{\frac{1}{2}} Z_{ij} + \alpha Z_{i5} \quad ; \quad j = 1, 2$$

$$i = 1,2,3,\ldots, 30$$

$$X_{ij} = (1 - \alpha_*^2)^{\frac{1}{2}} Z_{ij} + \alpha_* Z_{i5} ; \quad j = 3,4$$

$$i = 1,2,3,\ldots, 30$$

โดยที่ Z_{i1} , $Z_{i2}, \ldots,$ Z_{i5} เป็นค่าตัวแปรอิสระ ซึ่งสร้างขึ้นโดยให้มีการแจกแจงแบบปกติ มีค่าเฉลี่ยเท่ากับศูนย์ และความแปรปรวนเท่ากับหนึ่ง (N(0 , 1))

α^2 เป็นความสัมพันธ์ (ทางทฤษฎี)ระหว่างตัวแปร X_1 และ X_2

α_*^2 เป็นความสัมพันธ์ (ทางทฤษฎี) ระหว่างตัวแปร X_3 และ X_4

$\alpha\alpha_*$ เป็นความสัมพันธ์ (ทางทฤษฎี) ระหว่าง X_j $(j = 1,2)$ และ X_3 หรือ X_4

ถ้าต้องการสร้างข้อมูลชุดหนึ่ง ให้มีจำนวนตัวแปรอิสระเท่ากับ 4 และขนาดตัวอย่างเท่ากับ 1000 จากการทดลองกำหนดให้ $\alpha^2 = 0.9$ และ $\alpha_*^2 = 0.3$ จะได้ค่าต่าง ๆ ของตัวแปรอิสระจากสมการ

$$X_{ij} = (1-0.9^2)^{\frac{1}{2}} Z_{ij} + 0.9\, Z_{i5} \quad ; \quad j = 1,2$$
$$i = 1,2,3,\ldots, 1000$$

$$X_{ij} = (1-0.3^2)^{\frac{1}{2}} Z_{ij} + 0.3\, Z_{i5} \quad ; \quad j = 3,4$$
$$i = 1,2,3,\ldots, 1000$$

จากข้อมูลตัวอย่างที่สร้างขึ้นนี้ เมื่อนำมาหาเมตริกซ์ความสัมพันธ์ จะได้

$$R = \begin{bmatrix} 1 & 0.8037 & 0.2551 & 0.2861 \\ 0.8037 & 1 & 0.2463 & 0.2591 \\ 0.2551 & 0.2463 & 1 & 0.1103 \\ 0.2861 & 0.2591 & 0.1103 & 1 \end{bmatrix}$$

ในการวิจัยครั้งนี้ กำหนดให้ตัวแปรตามและตัวแปรอิสระมีความสัมพันธ์กันแบบเชิงเส้น ซึ่งมีตัวแบบดังนี้

$Y = X\beta + \varepsilon$; $\quad Y$ = เมตริกซ์ของตัวแปรตาม

X = เมตริกซ์ของตัวแปรอิสระ

β = เมตริกซ์ของค่าสัมประสิทธิ์ความถดถอยพหุ

ε = เมตริกซ์ความคลาดเคลื่อน

ดังนั้น จะต้องกำหนดค่า β เพื่อจะสร้างค่า Y ขึ้นจากตัวแบบดังกล่าว การกำหนดค่า β จะกำหนดตามแนวความคิดของ Macdonald and Galaneau (1975; 407-416) ซึ่งจะกำหนดค่า β ขึ้น 2 ชุดคือ

1) กำหนดให้ β คือ eigenvector ที่สอดคล้องกับ eigenvalue ซึ่งมีค่ามากที่สุดของเมตริกซ์ความสัมพันธ์ของตัวแปรอิสระ ด้วยมีเหตุผลว่า ถ้ากำหนดค่า β เช่นนี้ จะทำให้ $E\left[L(R)\right]$ และ $E\left[L(PC)\right]$ มีค่าน้อยที่สุด โดยที่ $E\left[L(R)\right] = E\left[(\hat{\beta}_R - \beta)'(\hat{\beta}_R - \beta)\right]$; $\hat{\beta}_R$ คือค่าประมาณสัมประสิทธิ์ความถดถอยพหุด้วยวิธี ริดจ์ รีเกรสชั่น และ $E\left[L(PC)\right]$ =

$E\left[(\hat{\beta}_{PC}-\beta)'(\hat{\beta}_{PC}-\beta)\right]$; $\hat{\beta}_{PC}$ คือค่าประมาณสัมประสิทธิ์ความถดถอยพหุด้วยวิธีรีเกรสชั่นพรินซิเปิ้ลคอมโพเน้นท์

2) กำหนดให้ β คือ eigenvector ที่สอดคล้องกับ eigenvalue ที่มีค่าน้อยที่สุดของเมตริกซ์ความสัมพันธ์ของตัวแปรอิสระ ด้วยมีเหตุผลว่าถ้ากำหนดค่า β เช่นนี้จะทำให้ $E\left[L(R)\right]$ และ $E\left[L(PC)\right]$ มีค่ามากที่สุด

จากตัวแบบ $Y = X\beta + \varepsilon$; $E(\varepsilon) = 0$; $E(\varepsilon\varepsilon') = \sigma^2 I_n$ ในการวิจัยครั้งนี้จึงต้องสร้างความคลาดเคลื่อนของข้อมูลแต่ละชุดให้มีการแจกแจงแบบปกติ มีค่าเฉลี่ยเป็นศูนย์ และความแปรปรวนของความคลาดเคลื่อนมีค่าเป็น σ^2 โดยกำหนดให้ σ = 0.1 , 1 ,5 และ 10 ค่าของตัวแปรตาม Y_i จะได้จากการแทนค่าในสมการ

$$Y_i = X_{i1}\beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3 + \ldots\ldots\ldots + X_{ip}\beta_p + \varepsilon_i$$

ในที่สุดก็จะได้ข้อมูลชุดหนึ่ง ซึ่งประกอบไปด้วยตัวแปรอิสระและตัวแปรตาม โดยที่ความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามเป็นแบบเชิงเส้น ข้อมูลที่ได้จากตัวแบบนี้เป็นข้อมูลมาตรฐาน ซึ่งในการใช้ข้อมูลมาตรฐานนี้ Gunst and Mason (1977 : 616-628) ได้กล่าวว่า เป็นการลดความคลาดเคลื่อนเนื่องจากการคำนวณลงเท่านั้น ไม่ได้ทำให้ความคลาดเคลื่อนเนื่องจากการเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระลดลงแต่อย่างใด

ในการวิจัยเรื่องนี้ต้องการเปรียบเทียบการประมาณค่า เมื่อเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ โดยใช้วิธีกำลังสองน้อยที่สุด วิธีริดจ์ รีเกรสชั่น และวิธีรีเกรสชั่นพรินซิเปิ้ลคอมโพเน้นท์ เนื่องจากวิธีริดจ์ รีเกรสชั่น และวิธีรีเกรสชั่นพรินซิเปิ้ลคอมโพเน้นท์จะให้ค่าประมาณสัมประสิทธิ์ความถดถอยพหุที่เอนเอียง แต่ทั้งสองวิธีนี้จะให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าประมาณสัมประสิทธิ์ความถดถอยพหุน้อยกว่าวิธีกำลังสองน้อยที่สุด ฉะนั้นในที่นี้จึงจะทำการศึกษาเพื่อดูว่าวิธีกำลังสองน้อยที่สุดจะให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองมากกว่าวิธี ริดจ์ รีเกรสชั่น และวิธีรีเกรสชั่นพรินซิเปิ้ลคอมโพเน้นท์ อยู่มากหรือน้อยเพียงใด เมื่อดัชนีพหุสัมพันธ์มีค่าต่าง ๆ กัน

ให้ $MSE(LS) = E(\hat{\beta}_{LS} - \beta)^2$; $\hat{\beta}_{LS}$ คือค่าประมาณสัมประสิทธิ์ความถดถอยพหุด้วยวิธีกำลังสองน้อยที่สุด

$MSE(R) = E(\hat{\beta}_R - \beta)^2$; $\hat{\beta}_R$ คือค่าประมาณสัมประสิทธิ์ความถดถอยพหุด้วยวิธี ริดจ์ รีเกรสชั่น

$$MSE(PC) = E(\hat{\beta}_{PC} - \beta)^2 \; ; \; \hat{\beta}_{PC}$$ คือค่าประมาณสัมประสิทธิ์ความถดถอยพหุด้วยวิธีรีเกรสชั่นพรินซิเปิ้ลคอมโพเน้นท์

ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าประมาณสัมประสิทธิ์ความถดถอยพหุจากวิธีกำลังสองน้อยที่สุดมีค่ามากกว่าวิธีริดจ์ รีเกรสชั่น

$$PR\% = \frac{(MSE(LS) - MSE(R)) \times 100}{MSE(LS)}$$

ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าประมาณสัมประสิทธิ์ความถดถอยพหุจากวิธีกำลังสองน้อยที่สุดมีค่ามากกว่าวิธีรีเกรสชั่นพรินซิเปิ้ลคอมโพเน้นท์

$$PPC\% = \frac{(MSE(LS) - MSE(PC)) \times 100}{MSE(LS)}$$

เมื่อดัชนีพหุสัมพันธ์มีค่าเพิ่มขึ้นเปอร์เซ็นต์ความแตกต่างนี้อาจเพิ่มขึ้นหรือลดลง นอกจากนี้จะศึกษาเพื่อดูว่าค่าประมาณตัวแปรตาม $(\hat{Y})$ ที่คำนวณได้จากวิธี ริดจ์ รีเกรสชั่น และวิธีรีเกรสชั่นพรินซิเปิ้ลคอมโพเน้นท์ จะใกล้เคียงกับค่าตัวแปรตามมากกว่าวิธีกำลังสองน้อยที่สุดหรือไม่ วิธีใดจะให้ค่าประมาณตัวแปรตามที่ใกล้เคียงกับค่าตัวแปรตามมากที่สุด โดยดูจาก

$$\hat{\sigma}^2_{LS} = \frac{\sum_{i=1}^{n} (\hat{Y}_{LSi} - Y_i)^2}{n-p}$$

$$\hat{\sigma}^2_{R} = \frac{\sum_{i=1}^{n} (\hat{Y}_{Ri} - Y_i)^2}{n-p}$$

$$\hat{\sigma}^2_{PC} = \frac{\sum_{i=1}^{n} (\hat{Y}_{PCi} - Y_i)^2}{n-p}$$

ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าประมาณตัวแปรตามจากวิธี ริดจ์ รีเกรสชั่น มีค่ามากกว่าวิธีกำลังสองน้อยที่สุด

$$VR = \frac{(\hat{\sigma}^2_{R} - \hat{\sigma}^2_{LS}) \times 100}{\hat{\sigma}^2_{LS}}$$

ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของค่าประมาณตัวแปรตามจากวิธีรีเกรสชั่นพรินซิเปิ้ลคอมโพ-เน้นท์ที่มีค่ามากกว่าวิธีกำลังสองน้อยที่สุด

$$VPC = \frac{(\hat{\sigma}^2_{PC} - \hat{\sigma}^2_{LS}) \times 100}{\hat{\sigma}^2_{LS}}$$

เนื่องจากในการวิจัยเรื่องนี้ มีการกำหนด ค่า β ขึ้น 2 ชุดคือ กำหนดให้ β คือ eigenvector ที่สอดคล้องกับ eigenvalue ที่มีค่ามากที่สุด ซึ่งจากการกำหนดเช่นนี้ จะทำให้ $E\left[L(R)\right]$ และ $E\left[L(PC)\right]$ มีค่าน้อยที่สุด เพราะฉะนั้นค่า PR และ PPC ที่คำนวณได้จากการกำหนดค่า β เช่นนี้ จะเป็นค่าขอบเขตบนของ PR และ PPC และเมื่อกำหนดให้ β คือ eigenvector ที่สอดคล้องกับ eigenvalue ที่มีค่าน้อยที่สุด ซึ่งจากการกำหนดเช่นนี้ จะทำให้ $E\left[L(R)\right]$ และ $E\left[L(PC)\right]$ มีค่ามากที่สุด เพราะฉะนั้นค่า PR และ PPC ที่คำนวณได้จากการกำหนดค่า β เช่นนี้ จะเป็นค่าขอบเขตล่างของ PR และ PPC

<u>แผนผังการเขียนโปรแกรม</u>

1. จำลอง (Simulate) ข้อมูลขึ้น โดยกำหนดความสัมพันธ์ของตัวแปรอิสระในลักษณะต่าง ๆ กัน
2. คำนวณ ค่าดัชนีพหุสัมพันธ์
3. กำหนดให้ $\underset{\sim}{\beta}$ คือ eigenvector ที่สอดคล้องกับ eigenvalue ที่มีค่าน้อยที่สุด
4. สุ่มค่าความคลาดเคลื่อน โดยที่ $\varepsilon_i \sim N(0 , \sigma^2)$
5. สร้างค่าตัวแปรตามจากสมการ $y_i = x_{i1} \beta_1 + x_{i2} \beta_2 + \ldots + x_{ip} \beta_p + \varepsilon_i$
6. ประมาณค่าสัมประสิทธิ์ความถดถอยพหุโดยวิธี กำลังสองน้อยที่สุด
   วิธี ริดจ์ รีเกรสชั่น
   วิธีรีเกรสชั่นพรินซิเปิ้ลคอมโพเนนท์
7. คำนวณค่า MSE(LS) , MSE(R) , MSE(PC) , $\hat{\sigma}^2_{LS}$, $\hat{\sigma}^2_{R}$, $\hat{\sigma}^2_{PC}$
8. สุ่มความคลาดเคลื่อน ชุดใหม่ กลับไปทำขั้นตอน 4 ซ้ำจนกว่าจะได้ 100 ครั้ง
9. หาค่าเฉลี่ยของ MSE(LS) , MSE(R) , MSE(PC), $\hat{\sigma}^2_{LS}$, $\hat{\sigma}^2_{R}$, $\hat{\sigma}^2_{PC}$
10. คำนวณค่า PR , PPC, VR, VPC
11. กำหนดให้ $\underset{\sim}{\beta}$ คือ eigenvector ที่สอดคล้องกับ eigenvalue ที่มีค่ามากที่สุด
12. กลับไปทำขั้นตอน 4 ใหม่