บทที่ 3

วิธีดำเนินการวิจัย

3.1 วิธีการวิจัย

การวิจัยครั้งนี้ใช้เทคนิคมอนติคาร์โล (Monte Carlo Tichniue) สร้างตัวแปรอิสระที่มีการแจกแจงแบบปกติ โดยที่ควบคุมค่าสัมพันธ์ ρ เพื่อบังคับให้ตัวแปรอิสระมีปัญหา Multicolinearity ในระดับที่ต้องการ ตัวแปรตามจะมีค่าที่ผันแปรไปตามค่าของตัวแปรอิสระและ Disturbance term, u กล่าวคือ

$$Y_j = \sum_{i=1}^{p} \beta_i X_{ij} + u_j \quad , \quad j = 1,2,\ldots, n$$

โดยที่ β_j คือค่าจริงของพารามิเตอร์ ค่าจริงดังกล่าวกำหนดขึ้นมาจาก Normalized eigen vector ของ Correlation matrix และ $U \sim N(0, \sigma^2)$; $\sigma^2 = .01$, .10 , .50 , 1.0 , 5.0

การศึกษาเปรียบเทียบมุ่งศึกษาดูว่าวิธีประมาณค่าแบบใดเหมาะสมที่สุดในระหว่างวิธีประมาณค่าแบบต่าง ๆ 5 วิธี วิธีที่ให้ค่า Total Mean Square Error (TMSE) ต่ำกว่าถือว่าเป็นวิธีที่ดีกว่า

ขั้นตอนการวิจัยโดยสรุปปรากฏดังนี้

1. สร้างข้อมูลของตัวแปรอิสระ X_1 , X_2, ..., X_5 ให้มีการแจกแจงเป็นแบบปกติ และข้อจำกัดของสหสัมพันธ์ที่ได้กำหนดไว้

2. สร้างข้อมูลของตัวแปร Y จากสมการ $Y_j = \sum_{i=1}^{p} \beta_i X_{ij} + u_j$; $j = 1,2,\ldots, n$

ทั้งนี้กำหนดให้ $\beta = (\beta_1 , \beta_2,\ldots, \beta_p)$ คือ Normalized eigen vector ของ แมตริกซ์สหสัมพันธ์(Correlation matrix)และ $U \sim N(0,\sigma^2)$; $\sigma^2 = .01, .10, .50, 1.0, 5.0$

การศึกษาจำกัดเฉพาะกรณี β_S และ β_L เมื่อ β_S คือ Normalized eigen vector ที่สอดคล้องกับ λ_{min} และ β_L คือ Normalized eigen vector ที่สอดคล้องกับ λ_{max}

3. กำหนดให้ตัวแปรอิสระทุกตัวมีอัตราการสูญหายของข้อมูลผันแปรโดยสุ่ม ตั้งแต่ 5-15%

4. ประมาณค่าข้อมูลที่สูญหายโดยวิธี Mean Substitution และวิธี Regression

5. ประมาณค่าพารามิเตอร์ตามวิธี Ridge Regression

6. คำนวณหา Total Mean Square Error (TMSE)

$$TMSE(m) = \sum_{i=1}^{p} \left[\frac{1}{r} \sum_{j=1}^{r} \{\hat{\beta}^{*}_{ij}(k) - \beta_i\}^2 \right] ; m = 1,2,..., 5 , r = 15$$

7. สรุปผลและอภิปรายผล

3.2 สรุปวิธีประมาณค่าพารามิเตอร์

รายละเอียดของขั้นตอนวิธีประมาณค่าพารามิเตอร์ทั้ง 5 วิธีเป็นดังนี้

3.2.1 Mean-Hoerl, Kennard and Baldwin (Mean-HKB)

(1) ประมาณค่าข้อมูลของตัวแปรอิสระที่สูญหายไปโดยแทนด้วยค่าเฉลี่ย

(2) จากตัวอย่างที่สมบูรณ์ วิเคราะห์สมการถดถอยตามวิธี Ordinary Least Square กล่าวคือ

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

$$s^2 = \frac{1}{n-p} [Y'Y - \hat{\beta}'X'Y]$$

(3) คำนวณหาค่า k ตามวิธี HKB คือ $k = \frac{ps^2}{\hat{\beta}'\hat{\beta}}$

(4) คำนวณหา Ridge Estimator ของ β

$$\hat{\beta}^{*}(1) = (X'X + kI_p)^{-1} X'Y$$

3.2.2 Mean-Lawless and Wang (Mean-LW)

(1) ประมาณค่าข้อมูลของตัวแปรอิสระที่สูญหายไป โดยวิธีแทนด้วยค่าเฉลี่ย

(2) จากค่าสังเกตที่สมบูรณ์ คำนวณหา Orthogonal modal matrix P ที่ $P'P = PP' = I_p$ ที่มีผลให้ $P'X'XP = \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$

(3) คำนวณหา OLS estimator ของ α จากสมการ $Y = X^*\alpha + U$, $X^* = XP$ และ $\alpha = P'\beta$

$$\hat{\alpha} = (X^{*\prime}X^*)^{-1}X^{*\prime}Y = \Lambda^{-1}X^{*\prime}Y$$

(4) คำนวณหาค่า k ตามวิธี LW คือ

$$k = \frac{PS^2}{\hat{\alpha}'\Lambda\hat{\alpha}}$$

$$S^2 = \frac{1}{n-p}\left[Y'Y - \hat{\beta}'X'Y\right]$$

(5) คำนวณหา Ridge Estimator ของ β

$$\hat{\beta}^*(2) = (X'X + kI_p)^{-1}X'Y$$

3.2.3 Regression-Hoerl, Kennard and Baldwin (Regression-HKB)

(1) ประมาณค่าข้อมูลของตัวแปรอิสระที่สูญหายไปโดยวิธีสมการทดถอย

(2) จากค่าสังเกตที่สมบูรณ์ วิเคราะห์สมการถดถอยตามวิธี Ordinary Least Square กล่าวคือ

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

$$S^2 = \frac{1}{n-p}\left[Y'Y - \hat{\beta}X'Y\right]$$

(3) คำนวณหาค่า k ตามวิธี HKB คือ $k = \frac{PS^2}{\hat{\beta}'\hat{\beta}}$

(4) คำนวณหา Ridge Estimator ของ β

$$\hat{\beta}^{*}(3) = (X'X + kI_p)^{-1} X'Y$$

3.2.4 Regression-Lawless and Wang (Regression-LW)

(1) ประมาณค่าข้อมูลของตัวแปรอิสระที่สูญหายไปโดยวิธีสมการถดถอย

(2) จากตัวอย่างที่สมบูรณ์ คำนวณหา Orthogonal modal matrix P

ที่ $P'P = PP' = I_p$ ที่มีผลให้ $P'X'XP = \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$

(3) คำนวณหา OLS estimator ของ α จากสมการ

$Y = X^{*}\alpha + U$, $X^{*} = XP$, $\alpha = P'\beta$

$$\hat{\alpha} = (X^{*\prime}X^{*})^{-1} X^{*\prime}Y = \Lambda^{-1} X^{*\prime}Y$$

(4) คำนวณหาค่า k ตามวิธี LW คือ

$$k = \frac{PS^2}{\hat{\alpha}'\Lambda\hat{\alpha}}$$

โดยที่ $S^2 = \frac{1}{n-p}\left[Y'Y - \hat{\beta}'X'Y\right]$

(5) คำนวณหา Ridge Estimator ของ β

$$\hat{\beta}^{*}(4) = (X'X + kI_p)^{-1} X'Y$$

3.2.5 Ordinary Least Square (OLS)

(1) ตัดตัวอย่างที่ไม่สมบูรณ์ในชุดนั้นทิ้งไป

(2) นำตัวอย่างที่เหลือมาวิเคราะห์สมการถดถอยตามวิธี

$$\hat{\beta}^{*}(5) = (X'X)^{-1} X'Y$$

หรือ $\hat{\beta}^{*}(5) = (X'X + kI_p)^{-1} X'Y$ เมื่อ $k = 0$

3.3 การสร้างข้อมูลของตัวแปรอิสระ

1. $X_{ij} = (1 - \rho^2)^{1/2} Z_{ij} + \rho Z_{6j}$; i = 1,2,3, j = 1,2,..., n

รวม 15 ชุด โดยที่ Z_{1j} , Z_{2j}, Z_{3j} และ Z_{6j} คือ Independent pseudo N(0,1) และ ρ คือ ค่าสหสัมพันธ์ที่แท้จริงระหว่างตัวแปรอิสระ คือ ระหว่าง X_1 กับ X_2 , X_1กับ X_3 และ X_2 กับ X_3 ทั้งนี้ $\rho^2 = (.99)^2$, $(.90)^2$, $(.70)^2$

2. $X_{ij} = (1 - \rho_*^2)^{1/2} Z_{ij} + \rho_* Z_{6j}$; i = 4,5,j=1,2,..., n

รวม 15 ชุด โดยที่ Z_{4j} และ Z_{5j} คือ Independent pseudo N(0,1) และ ρ_* คือค่าสหสัมพันธ์ที่แท้จริงระหว่าง X_4 กับ X_5 ทั้งนี้ $\rho_*^2 = (.99)^2$, $(.90)^2$, $(.30)^2$ $(.10)^2$

3. แปลงข้อมูลทุกรายการให้อยู่ในรูปของคะแนนมาตรฐานคือแปลงจาก X_{ij} เป็น $(X_{ij} - \bar{X}_i)/S_i$; i = 1,2,..., 5

3.4 การสร้างข้อมูลของตัวแปรตาม

1. จากข้อมูลของตัวแปรอิสระในตอน 3.3 เมื่อแปลงเป็นคะแนนมาตรฐาน จะมีผลให้เมตริกซ์ $X'X$ เปลี่ยนเป็นเมตริกซ์สหสัมพันธ์ดังนี้

ตัวแปรที่	1	2	3	4	5
1	1				
2	ρ^2	1		SYM	
3	ρ^2	ρ^2	1		
4	$\rho\rho_*$	$\rho\rho_*$	$\rho\rho_*$	1	
5	$\rho\rho_*$	$\rho\rho_*$	$\rho\rho_*$	ρ_*^2	1

2. จากเมตริกซ์สหสัมพันธ์ในแต่ละกรณีของข้อ 1 คำนวณหา Eigen value และ normalized eigen vector ดังนี้

(1) $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4 > \lambda_5$ คือ Ordered eigen value ของเมตริกซ์ $X'X$

(2) V_i ; i = 1,2,..., 5 คือ normalized eigen vector ที่สอดคล้องกับ λ_i ; i = 1,2,..., 5 และให้ β_L และ β_S คือ เวกเตอร์ V ที่สอดคล้องกับ λ_1 และ λ_5 ตามลำดับ

3. กำหนดค่าของตัวแปรตามเป็น 2 ชุดตามค่าของพารามิเตอร์และค่าของตัวแปรอิสระดังนี้

ก. เมื่อ $\beta = \beta_L$

$$Y_j = \beta_{1L} X_{1j} + \beta_{2L}X_{2j} + \beta_{3L}X_{3j} + \beta_{4L}X_{4j} + \beta_{5L}X_{5j} \text{ ; } j = 1,2,..., n$$

ข. เมื่อ $\beta = \beta_S$

$$Y_j = \beta_{1S} X_{1j} + \beta_{2S}X_{2j} + \beta_{3S}X_{3j} + \beta_{4S}X_{4j} + \beta_{5S}X_{5j} \text{ ; } j = 1,2,..., n$$

4. กำหนดให้ $U \sim N(0,\sigma^2)$ โดยที่ σ^2 = .01, .10, .50, 1.0, 5.0

5. จากค่าของตัวแปรอิสระในตอน 3.3 ค่าของ β_L และ β_S ในข้อ 3 ข้างต้น และค่าของ σ^2 ในข้อ 4 สร้างข้อมูลของตัวแปรตามในแต่ละกรณี ๆ ละ 15 ชุด ๆ ละ 20 หน่วยและ 30 หน่วยตามลำดับ

3.5 ขั้นตอนการประมาณค่าสัมประสิทธิ์ที่สูญหาย

เพื่อความเข้าใจในเทคนิคการประมาณค่าสัมประสิทธิ์ที่สูญหาย จะยกตัวอย่างในกรณีที่มีตัวแปรอิสระเท่ากับ 3 และขนาดตัวอย่างเท่ากับ 10 ดังนี้

ตัวแปรที่

		1	2	3
	1	0	1	1
	2	1	1	1
	3	1	0	0
	4	1	0	0
ค่าสังเกตที่	5	1	1	1
	6	1	1	1
	7	0	0	1
	8	1	1	1
	9	1	1	0
	10	1	1	1

โดยให้ 1 = ข้อมูลที่สมบูรณ์

0 = ข้อมูลที่สูญหาย

3.5.1 วิธีแทนด้วยค่าเฉลี่ย

จากตัวอย่างข้างต้นในการประมาณค่าสังเกตที่สูญหายในตัวแปรที่ 1 นำค่าสังเกตที่สมบูรณ์เฉพาะในตัวแปรที่ 1 คือ ค่าสังเกตที่ 2 3 4 5 6 8 9 และ 10 มาหาค่าเฉลี่ย ซึ่งค่าที่ได้จะถูกนำไปแทนในค่าสังเกตที่ 1 และ 7 เฉพาะในตัวแปรที่ 1

สำหรับในตัวแปรที่ 2 นำค่าสังเกตที่ 1 2 5 6 8 9 และ 10 มาหาค่าเฉลี่ยค่าที่ได้จะถูกนำไปแทนในค่าสังเกตที่ 3 4 และ 7 เฉพาะในตัวแปรที่ 2 ส่วนในตัวแปรที่ 3 ใช้ค่าสังเกตที่ 1 2 5 6 7 8 และ 10 มาหาค่าเฉลี่ย และนำไปแทนในค่าสังเกตที่ 3 4 และ 9 เช่นเดียวกัน

3.5.2 วิธีส่มการถดถอย

วิธีนี้มีขั้นตอนที่สำคัญโดยสรุปดังนี้

(1) เริ่มจากการพิจารณาค่าสังเกตที่มีค่าสูญหายเพียง 1 ค่า ซึ่งในที่นี้คือค่าสังเกตที่ 1 และ 9 โดยการนำค่าสังเกตที่สมบูรณ์คือ ค่าสังเกตที่ 2 5 6 8 และ 10 ไปสร้างสมการถดถอยขึ้น 3 สมการ คือ

สมการที่ 1 กำหนดให้ ตัวแปรที่ 1 เป็นตัวแปรตาม ตัวแปรที่ 2 และ 3 เป็นตัวแปรอิสระ

สมการที่ 2 กำหนดให้ ตัวแปรที่ 2 เป็นตัวแปรตาม ตัวแปรที่ 1 และ 3 เป็นตัวแปรอิสระ

สมการที่ 3 กำหนดให้ ตัวแปรที่ 3 เป็นตัวแปรตาม ตัวแปรที่ 1 และ 2 เป็นตัวแปรอิสระ

จากนั้นจึงประมาณค่าสังเกตที่สูญหายในค่าสังเกตที่ 1 และ 9 โดยใช้สมการที่ 1 และ 3 ตามลำดับ แล้วจึงหาผลต่างระหว่างค่าสังเกตที่ประมาณได้กับค่าสังเกตที่มีอยู่เดิม ซึ่งในที่นี้คือค่า 0 ถ้าผลต่างมากกว่า 0.1 ก็จะดำเนินการประมาณค่าสังเกตที่สูญหายโดยใช้วิธีการเดิมอีกครั้งหนึ่ง ซึ่งในที่นี้ก็จะนำค่าสังเกตที่สมบูรณ์รวมกับค่าสังเกตที่ประมาณได้จากครั้งแรก คือ ค่าสังเกตที่ 1 2 5 6 8 9 และ 10 ไปสร้างสมการถดถอยเพื่อประมาณค่าตัวแปรที่ 1 และ 3 อีกครั้งหนึ่ง และจะทำการหาผลต่างจากค่าที่ประมาณได้กับค่าที่ได้จากครั้งก่อน และจะทำซ้ำไปเรื่อย ๆ จนกระทั่งผลต่างไม่มากกว่า 0.1

(2) เริ่มพิจารณาค่าสังเกตที่มีค่าสังเกตสูญหาย 2 ค่า ซึ่งในที่นี้คือค่าสังเกตที่ 3 4 และ 7 โดยการนำค่าสังเกตที่สมบูรณ์ที่ได้จาก (1) คือค่าสังเกตที่ 1 2 5 6 8 9 และ 10 ไปสร้างสมการถดถอยขึ้น 3 สมการ เพื่อประมาณค่าสังเกตที่ 3 4 และ 7 ตามลำดับ โดยยึดหลักเกณฑ์เช่นเดียวกับ (1)

3.6 วิธีวิเคราะห์ข้อมูล

แนวทางในการวิเคราะห์ข้อมูลสำหรับการวิจัยจำแนกแผนการวิเคราะห์ออกเป็นดังนี้

1. เปรียบเทียบการประมาณค่าพารามิเตอร์ตามระดับความรุนแรงของปัญหา Multicolinearite

2. เปรียบเทียบการประมาณค่าพารามิเตอร์ตามระดับของความแปรปรวน

I17613078

3.7 โปรแกรมที่ใช้ในการวิจัย

โปรแกรมที่ใช้ในการวิจัยครั้งนี้ เขียนขึ้นด้วยภาษาเบสิก (BASICA) บนเครื่อง IBM PS/2 (IBM PERSONAL SYSTEM/2) MODEL 30 ซึ่งมีโปรแกรมทั้งหมด 9 โปรแกรม ดังนี้

3.7.1 โปรแกรมการสร้างตัวแปรอิสระและตัวแปรตาม ตามเงื่อนไขที่กำหนดรวมทั้งการกำหนดค่าสูญหายโดยสุ่มให้แก่ตัวแปรอิสระ

3.7.2 โปรแกรมประมาณค่าสังเกตที่สูญหายโดยวิธี Mean Substitution

3.7.3 โปรแกรมประมาณค่าพารามิเตอร์ตามวิธี Mean-HKB รวมทั้งคำนวณค่า Total Mean Square Error (TMSE)

3.7.4 โปรแกรมประมาณค่าพารามิเตอร์ตามวิธี Mean-LW รวมทั้งคำนวณค่า Total Mean Square Error (TMSE)

3.7.5 โปรแกรมประมาณค่าสังเกตที่สูญหายโดยวิธี Regression

3.7.6 โปรแกรมประมาณค่าพารามิเตอร์ตามวิธี Regression-HKB รวมทั้งคำนวณค่า Total Mean Square Error (TMSE)

3.7.7 โปรแกรมประมาณค่าพารามิเตอร์ตามวิธี Regression-LW รวมทั้งคำนวณค่า Total Mean Square Error (TMSE)

3.7.8 โปรแกรมประมาณค่าสังเกตที่สูญหายและประมาณค่าพารามิเตอร์ตามวิธี Ordinary Least Square (OLS)

3.7.9 โปรแกรมคำนวณค่า Total Mean Square Error (TMSE) ของวิธี Ordinary Least Square (OLS)

สำหรับขั้นตอนในการหาค่าสถิติต่าง ๆ เพื่อใช้ในการวิจัย สรุปเป็นผังงาน (Flowchart) ดังนี้

รูปที่ 1 ผังงานของโปรแกรมที่ใช้ในการวิจัย