บทที่ 3

วิธีดำเนินงานวิจัย

ข้อมูลที่ใช้ในการวิจัยครั้งนี้ได้จากการจำลองด้วยเทคนิคมอนติคาร์โล โดยใช้โปรแกรมภาษาฟอร์แทรน 77 ซึ่งขั้นตอนและโปรแกรมที่ใช้ในการวิจัยมีรายละเอียดดังต่อไปนี้

แผนการทดลอง

ในการวิจัยครั้งนี้กำหนดสถานการณ์ต่างๆ ที่ต้องการศึกษาดังนี้

1. กำหนดจำนวนตัวแปรอิสระ มี 2 ระดับ คือ 3 และ 5

2. กำหนดขนาดตัวอย่าง มี 3 ระดับ คือ 30, 50 และ 100 โดยที่ในแต่ละขนาดตัวอย่างจะสร้างข้อมูลเพิ่มขึ้นอีก 20 ตัวอย่าง เพื่อใช้ในการทดสอบตัวแบบ

3. กำหนดระดับพหุสัมพันธ์ระหว่างตัวแปรอิสระ

กรณีที่จำนวนตัวแปรอิสระเท่ากับ 3 ผู้วิจัยสนใจศึกษา เมื่อระดับพหุสัมพันธ์เท่ากับ 0.1, 0.3, 0.5, 0.7, 0.9 และ 0.99

กรณีที่จำนวนตัวแปรอิสระเท่ากับ 5 ผู้วิจัยสนใจศึกษา เมื่อระดับพหุสัมพันธ์เท่ากับ (0.1,0.1), (0.3,0.3), (0.5,0.5), (0.7,0.7), (0.9,0.9) และ (0.99,0.99)[1]

---

[1] กรณีที่จำนวนตัวแปรอิสระเท่ากับ 3 ระดับพหุสัมพันธ์เท่ากับ 0.1, 0.3, 0.5, 0.7, 0.9 และ 0.99 หมายถึง ระดับความสัมพันธ์ระหว่างตัวแปร X_1 กับ X_2, X_1 กับ X_3 และ X_2 กับ X_3

กรณีที่จำนวนตัวแปรอิสระเท่ากับ 5 ระดับพหุสัมพันธ์เท่ากับ (0.1,0.1), (0.3,0.3), (0.5,0.5), (0.7,0.7), (0.9,0.9) และ (0.99,0.99) ค่าแรกในวงเล็บ หมายถึง ระดับความสัมพันธ์ระหว่างตัวแปร X_1 กับ X_2, X_1 กับ X_3 และ X_2 กับ X_3 ค่าหลังในวงเล็บ หมายถึง ระดับความสัมพันธ์ระหว่างตัวแปร X_4 กับ X_5

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

4. กำหนดความคลาดเคลื่อนมีการแจกแจงดังต่อไปนี้

4.1 การแจกแจงปกติ โดยที่พารามิเตอร์ μ เท่ากับ 1 และ σ เท่ากับ 0.1, 0.3 และ 0.5 ตามลำดับ

4.2 การแจกแจงปกติปลอมปน โดยที่ สเกลแฟคเตอร์ (c) เท่ากับ 3 และ 10 และ เปอร์เซ็นต์การปลอมปน (p) เท่ากับ 5 และ 10

4.3 การแจกแจงลอกนอร์มอล โดยที่พารามิเตอร์ μ เท่ากับ 1 และ σ เท่ากับ 0.2264, 0.5915 และ 1.0069

ขั้นตอนการวิจัย

ขั้นตอนในการวิจัยมีดังนี้

1. สร้างข้อมูลของความคลาดเคลื่อนให้มีลักษณะตามที่ต้องการศึกษา

2. การสร้างข้อมูลของตัวแปรอิสระ (X) ให้มีความสัมพันธ์กันตามที่กำหนดและสร้างข้อมูลตัวแปรตาม (y) จากรูปแบบความสัมพันธ์ $\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\epsilon}$ พร้อมทั้งทำข้อมูลให้อยู่ในรูปแบบมาตรฐาน (standard form) ในกรณีที่ความคลาดเคลื่อนมีการแจกแจงลอกนอร์มอล จะทำการแปลงข้อมูลของตัวแปรตาม โดยใช้การแปลงที่ใช้การยกกำลังของบอกซ์ (Box) และคอกซ์ (Cox) แล้วจึงประมาณค่าสัมประสิทธิ์การถดถอย บันทึกข้อมูลลงแฟ้ม (file) และนำไปวิเคราะห์ด้วยวิธีที่ใช้หลักการของโครงข่ายประสาทเทียมโดยใช้โปรแกรมสำเร็จรูป "Trajan 2.1 Shareware Neural Network Simulator, March 1997"

3. ประมาณค่าสัมประสิทธิ์การถดถอยพหุคูณด้วยวิธีริดจ์รีเกรสชัน พร้อมทั้งหาค่าความคลาดเคลื่อนกำลังสองเฉลี่ย

4. เปรียบเทียบประสิทธิภาพของวิธีริดจ์รีเกรสชัน และวิธีที่ใช้หลักการของโครงข่ายประสาทเทียม พร้อมทั้งสรุปผล

รายละเอียดแต่ละขั้นตอนมีดังนี้

1. สร้างข้อมูลของความคลาดเคลื่อนให้มีลักษณะตามที่ต้องการศึกษา

โดยการสร้างข้อมูลที่มีลักษณะการแจกแจงต่างๆ จะต้องใช้เลขสุ่ม (random number) ซึ่งมีการแจกแจงแบบสม่ำเสมอในช่วง (0,1) เป็นพื้นฐาน

1.1 การสร้างเลขสุ่มที่มีการแจกแจงปกติ

โดยใช้วิธีการของ บอกซ์ (Box) และมูลเลอร์ (Muller) (ค.ศ. 1978) โดยใช้ตัวเลขที่มีการแจกแจงสม่ำเสมอ 2 ค่า คือ R_1 และ R_2 ผลิตเลขสุ่มที่มีการแจกแจงปกติมาตรฐาน มีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนเท่ากับ 1 พร้อมกัน 2 ค่า คือ Z_1 และ Z_2 ได้ดังนี้

$$Z_1 = (-2\ln(R_1))^{1/2} \cos(2\pi R_2)$$

$$Z_2 = (-2\ln(R_1))^{1/2} \sin(2\pi R_2)$$

ทำการแปลงค่าเลขสุ่มดังกล่าวโดย

$$Normal_1 = \mu + \sigma Z_1$$

$$Normal_2 = \mu + \sigma Z_2$$

ดังนั้นเราจะได้ว่า $Normal_1$ และ $Normal_2$ มีการแจกแจงปกติด้วยค่าเฉลี่ยเท่ากับ μ และส่วนเบี่ยงเบนมาตรฐานเท่ากับ σ

1.2 การสร้างเลขสุ่มที่มีการแจกแจงปกติปลอมปน

โดยใช้วิธีแรมเซย์ (Ramsay) (ค.ศ. 1977) เสนอไว้ โดยสร้างการแจกแจงที่แปลงมาจากการแจกแจงปกติ ที่มีฟังก์ชันการแจกแจงอยู่ในรูปของ

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

โดยที่ตัวแปรสุ่ม X มาจากการแจกแจง $N(\mu,\sigma^2)$ ด้วยความน่าจะเป็น (1-p) และการแจกแจง $N(\mu,c^2\sigma^2)$ ด้วยความน่าจะเป็น p ซึ่ง μ, σ^2, c และ p เป็นค่าเฉลี่ย ค่าความแปรปรวน สเกลแฟคเตอร์ และเปอร์เซ็นต์การปลอมปน ตามลำดับ

1.3 การสร้างเลขสุ่มที่มีการแจกแจงลอกนอร์มอล

การแจกแจงลอกนอร์มอลมีฟังก์ชันความหนาแน่นอยู่ในรูปของ

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma x} \exp\left[-\frac{1}{2\sigma^2}(\ln x - \mu)^2\right] \qquad ; x > 0$$

เมื่อ μ และ σ^2 เป็นค่าเฉลี่ยและความแปรปรวนของ Y โดยที่ $Y = \ln(x)$ แล้ว Y จะมีการแจกแจงปกติโดยมี $\exp(\sigma^2)$ เป็นพารามิเตอร์มาตราส่วน (scale parameter) และ μ เป็นพารามิเตอร์สัณฐาน (shape parameter)

ค่าเฉลี่ย ความแปรปรวน และสัมประสิทธิ์การแปรผันของการแจกแจงลอกนอร์มอล คือ

$$E(X) = \exp(\mu + \frac{\sigma^2}{2})$$

$$Var(X) = \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1]$$

$$C.V.(X) = \sqrt{[\exp(\sigma^2) - 1]}$$

สำหรับการสร้างเลขสุ่มที่มีการแจกแจงแบบลอกนอร์มอล ได้จากเลขชี้กำลัง (exponential) ของเลขสุ่มที่ได้จากการแจกแจงปกติ

2. การสร้างข้อมูลที่มีความสัมพันธ์เชิงเส้น

สร้างข้อมูลของตัวแปรอิสระ X ให้มีความสัมพันธ์กันดังนี้

กรณีจำนวนตัวแปรอิสระเท่ากับ 3 ระดับความสัมพันธ์ที่ศึกษา คือ 0.1, 0.3, 0.5, 0.7, 0.9 และ 0.99

กรณีจำนวนตัวแปรอิสระเท่ากับ 5 ระดับความสัมพันธ์ที่ศึกษา คือ (0.1,0.1), (0.3,0.3), (0.5,0.5), (0.7,0.7), (0.9,0.9) และ (0.99,0.99)

ผู้วิจัยได้ใช้วิธีการของวิเชิน (Wichern) และเชอร์ชิลล์ (Churchill) ซึ่งทำให้สามารถสร้างข้อมูลตัวแปรอิสระที่มีความสัมพันธ์ในระดับต่างๆ ดังตัวอย่างต่อไปนี้

<u>ตัวอย่าง</u> ต้องการสร้างข้อมูลที่มีจำนวนตัวแปรอิสระเท่ากับ 5 และขนาดตัวอย่างเท่ากับ 30 ได้จากความสัมพันธ์ต่อไปนี้

$$X_{ij} = (1-\alpha^2)^{1/2} Z_{ij} + \alpha Z_{i6} \qquad ; \ j = 1, 2, 3 \ ; i = 1, 2, ..., 30$$

$$X_{ij} = (1-\alpha_*^2)^{1/2} Z_{ij} + \alpha_* Z_{i6} \qquad ; \ j = 4, 5 \ ; i = 1, 2, ..., 30$$

เมื่อ $Z_{i1}, Z_{i2}, ..., Z_{i6}$ คือ ตัวแปรสุ่มที่ได้จากการแจกแจง N(0,1)

α^2 คือ ความสัมพันธ์ระหว่างตัวแปร X_1, X_2 และ X_3

$\alpha\alpha_*$ คือ ความสัมพันธ์ระหว่าง X_1, X_2 และ X_3 กับ X_4 หรือ X_5

และ α_*^2 คือ ความสัมพันธ์ระหว่าง X_4 และ X_5

ผู้วิจัยสร้างข้อมูลตัวแปรตามจากความสัมพันธ์เชิงเส้น $\underset{\sim}{y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ โดยที่ $\underset{\sim}{y}$ คือ เวกเตอร์ของตัวแปรตาม X คือเมทริกซ์ของตัวแปรอิสระ $\underset{\sim}{\beta}$ คือค่าพารามิเตอร์ที่สร้างจากเวกเตอร์เฉพาะที่สอดคล้องกับค่าเฉพาะของเมทริกซ์ $X'X$ และ $\underset{\sim}{\varepsilon}$ คือค่าความคลาดเคลื่อนที่มีการแจกแจงต่างๆ แต่ในกรณีที่ความคลาดเคลื่อนมีการแจกแจงออกนอร์มอล ผู้วิจัยจะทำการแปลงข้อมูล Y เพื่อที่จะทำให้ข้อสมมติของความเป็นปกติใช้ได้ โดยใช้การแปลงที่ใช้การยกกำลังของบอกซ์และคอกซ์ เพื่อสร้างตัวแปร $Y^{(\lambda)}$ ที่มีรูปแบบความสัมพันธ์เชิงเส้น คือ $\underset{\sim}{y}^{(\lambda)} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ โดยที่ $\underset{\sim}{y}^{(\lambda)}$ คือเวกเตอร์ของตัวแปรตามที่แปลงโดยใช้การยกกำลังของบอกซ์ และคอกซ์ X คือเมทริกซ์ของตัวแปรอิสระ $\underset{\sim}{\beta}$ คือค่าพารามิเตอร์ที่สร้างจากเวกเตอร์เฉพาะที่สอดคล้องกับค่าเฉพาะของเมทริกซ์ $X'X$ และ $\underset{\sim}{\varepsilon}$ คือค่าความคลาดเคลื่อนที่มีการแจกแจงสมมาตรและเป็นปกติโดยประมาณ พร้อมทั้งทำข้อมูล X, $\underset{\sim}{y}$ และ $Y^{(\lambda)}$ ให้อยู่ในรูปแบบมาตรฐาน และทำการบันทึกข้อมูลที่สร้างได้เพื่อนำไปพยากรณ์ในการวิเคราะห์ถดถอยพหุคูณด้วยวิธีโครงข่ายประสาทเทียม

3. ผู้วิจัยใช้ข้อมูล (n-20) ตัวแรกในการประมาณค่าสัมประสิทธิ์การถดถอยพหุคูณด้วยวิธีริดจ์-รีเกรสชันจาก $\hat{\underset{\sim}{\beta}}_R = (X'X + kI)^{-1}X'\underset{\sim}{y}$ และทำการประมาณค่าพารามิเตอร์ k ด้วยวิธีการค้นหาลำดับ หลังจากนั้นผู้วิจัยใช้ข้อมูล 20 ตัวหลังแทนค่าในสมการพยากรณ์ได้ค่า $\hat{Y}_R$ นำมาหาค่าความคลาดเคลื่อนในการพยากรณ์ตัวแปรตาม และหาค่าความคลาดเคลื่อนกำลังสองเฉลี่ย

หลังจากนั้นผู้วิจัยได้ใช้ข้อมูลที่เก็บรวมแล้วข้อมูลไว้นำมาประมวลผลโดยใช้โปรแกรมสำเร็จรูป Trajan 2.1 เพื่อหาค่าความคลาดเคลื่อนกำลังสองเฉลี่ยโดยใช้วิธีโครงข่ายประสาทเทียม ซึ่งโครงข่ายประสาทเทียมที่ใช้ในการวิจัยครั้งนี้มีแบบจำลองเพอร์เซปตรอน (perceptron) 3 ชั้น คือ ชั้นที่หนึ่งเรียกว่า ชั้นนำเข้า มีจำนวนนิวรอนเท่ากับจำนวนตัวแปรอิสระ ชั้นที่สองเรียกว่า ชั้นซ่อน มีจำนวนนิวรอนเท่ากับ 1 และชั้นที่สามเรียกว่า ชั้นผลลัพธ์ มีจำนวนนิวรอนเท่ากับจำนวนตัวแปรตาม คือ 1 การเชื่อมต่อระหว่างชั้นแบบ Fully Connected Feedforward และใช้ขั้นตอนการเรียนรู้แบบแพร่ย้อนกลับ (backpropagation)

เราจะได้ตัวแบบ ซึ่งอยู่ในรูปของ

$$\hat{\underset{\sim}{y}}_A = f(f(\sum_{i=1}^{p} x_i v_{ii})w_{11}$$

ซึ่งเราสามารถคำนวณค่า AMSE ของทั้งสองวิธีได้ดังนี้

$$MSE(R_j) = \frac{\sum_{i=n-19}^{n} (y_{ij} - \hat{y}_{R_{ij}})^2}{20}$$

$$MSE(A_j) = \frac{\sum_{i=n-19}^{n} (y_{ij} - \hat{y}_{A_{ij}})^2}{20}$$

เมื่อ i คือ ข้อมูลที่นำมาหาความคลาดเคลื่อน ; i = (n-19), (n-18), ..., n

j คือ จำนวนนิวรอนที่ทำการทดลองซ้ำ ; j = 1, 2, ..., 400

y_{ij} คือ ค่าสังเกตที่ i ในการทดลองครั้งที่ j

$\hat{y}_{R_{ij}}$ คือ ค่าพยากรณ์ที่ i ในการทดลองครั้งที่ j โดยใช้วิธีริดจ์รีเกรสชัน

$\hat{y}_{A_{ij}}$ คือ ค่าพยากรณ์ที่ i ในการทดลองครั้งที่ j โดยใช้วิธีที่ใช้หลักการของโครงข่ายประสาทเทียม

$MSE(R_j)$ คือ ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่ได้จากวิธีริดจ์รีเกรสชัน

และ $MSE(A_j)$ คือ ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยที่ได้จากวิธีโครงข่ายประสาทเทียม

เนื่องจากในงานวิจัยครั้งนี้กระทำซ้ำ 400 ครั้ง ในแต่ละสถานการณ์ ดังนั้นค่าเฉลี่ยของ MSE แต่ละวิธี คือ

$$AMSE(R) = \frac{1}{400}\sum_{j=1}^{400} MSE(R_j)$$

$$AMSE(A) = \frac{1}{400}\sum_{j=1}^{400} MSE(A_j)$$

4. เปรียบเทียบประสิทธิภาพของวิธีริดจ์รีเกรสชันและวิธีที่ใช้หลักการของโครงข่ายประสาทเทียม โดยใช้เปอร์เซ็นต์อัตราส่วนความแตกต่างระหว่างค่า AMSE ทั้ง 2 วิธี เป็นเกณฑ์ โดยจะเปรียบเทียบกับวิธีที่ให้ค่า AMSE ต่ำสุด คำนวณได้จาก

$$DIFF = \frac{AMSE_{(i)} - AMSE_{(min)}}{AMSE_{(min)}} \times 100 \qquad ; i = 1,2$$

เมื่อ $DIFF$ คือ เปอร์เซ็นต์อัตราส่วนผลต่างค่าเฉลี่ยของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย

$AMSE_{(min)}$ คือ ค่า AMSE ของวิธีที่ให้ค่าต่ำสุด

และ $AMSE_{(i)}$ คือ ค่า AMSE ของวิธีที่เหลือ

ผู้วิจัยได้แสดงตารางลักษณะการทำงานของโปรแกรมทั้งหมดที่ใช้ในการวิจัย รวมทั้งแผนผังการทำงานของโปรแกรมไว้ดังนี้

ตารางที่ 1 แสดงลักษณะการทำงานของโปรแกรมที่ใช้ในการวิจัย

อันดับที่	ชื่อโปรแกรม	คุณสมบัติโปรแกรม	ชื่อโปรแกรมที่เรียกใช้
Main Program	-	1. รับค่าและกำหนดค่าพารามิเตอร์ในสถานการณ์ต่างๆ 2. คำนวณค่า AMSE ที่ได้จากวิธีริดจ์รีเกรสชัน	GENX, STNDRDX, CALXTX, EIGEN, NORMAL, SCALE, LOGNOR, CALY, BOXCOX, STNDRDY, WRTDT, YSD, CALK, CBTHG, CMSE
Subroutine			
1	GENX	สร้างเมทริกซ์ตัวแปรอิสระ X ที่สัมพันธ์กัน	NORMAL
2	STNDRDX	แปลงเมทริกซ์ให้อยู่ในรูปแบบมาตรฐาน	
3	STNDRDY	แปลงเวกเตอร์ $\underset{\sim}{y}$ ให้อยู่ในรูปแบบมาตรฐาน	
4	WRTDT	เขียนเมทริกซ์ X และเวกเตอร์ $\underset{\sim}{y}$ ที่อยู่ในรูปมาตรฐานลงแฟ้มข้อมูล	
5	CALXTX	คำนวณหาค่า $X'X$	
6	EIGEN	รับและส่งค่าเฉพาะและเวกเตอร์เฉพาะ	JACOBI
7	JACOBI	คำนวณค่าเฉพาะและเวกเตอร์เฉพาะโดยใช้วิธี JACOBI	
8	CALY	คำนวณหาค่าเวกเตอร์ตัวแปรตาม $\underset{\sim}{y}$	
9	BOXCOX	แปลงค่า $\underset{\sim}{y}$ โดยใช้การยกกำลังของ Box และ Cox เมื่อ $\underset{\sim}{\varepsilon}$ มีการแจกแจงลอกนอร์มอล	OLS

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

ตารางที่ 1 (ต่อ)

อันดับที่	ชื่อโปรแกรม	คุณสมบัติโปรแกรม	ชื่อโปรแกรมที่เรียกใช้
10	OLS	คำนวณ $\hat{\beta}$ โดยใช้วิธีกำลังสองน้อยสุด	BXTY, INVRS, BMSE
11	BXTY	คำนวณหาค่า $X'y$	
12	INVRS	คำนวณหาเมทริกซ์ผกผัน	
13	BMSE	คำนวณหาค่า MSE ในการพยากรณ์ $\hat{y}$ ที่ได้จากวิธีกำลังสองน้อยสุด	
14	CALK	คำนวณหาค่าพารามิเตอร์ k ที่เหมาะสมโดยวิธีค้นหาแบบลำดับ	MSEB
15	YSD	คำนวณค่าส่วนเบี่ยงเบนมาตรฐานของ y ชุดที่ 1 ถึง (n-20)	
16	CXTY	คำนวณหาค่า $X'y$ โดยใช้ข้อมูลชุดที่ 1 ถึง (n-20)	
17	CBTHG	คำนวณหาค่า $\hat{\beta}_R$ โดยใช้ข้อมูลชุดที่ 1 ถึง (n-20)	
18	CMSE	คำนวณหาค่า MSE ในการพยากรณ์ $\hat{y}_R$ ที่ได้จากวิธีริดจ์รีเกรสชัน	
Function			
1	MSEB	คำนวณหาค่า MSE ในการประมาณ $\hat{\beta}_R$ เพื่อหาค่า k ที่เหมาะสม	
2	NORMAL	สร้างเลขสุ่มที่มีการแจกแจงปกติ	RAND
3	SCALE	สร้างเลขสุ่มที่มีการแจกแจงปกติปลอมปน	NORMAL
4	LOGNOR	สร้างเลขสุ่มที่มีการแจกแจงลอกนอร์มอล	NORMAL
5	RAND	สร้างเลขสุ่มที่มีการแจกแจงสม่ำเสมอ	

รูปที่ 3.1 แผนผังการทำงานของโปรแกรม