บทที่ 3

การดำเนินการวิจัย

การวิจัยครั้งนี้มีลักษณะเป็นการวิจัยเชิงทดลอง เพื่อต้องการเปรียบเทียบวิธีการประมาณค่าสูญหายของตัวแปรตามในสมการถดถอยเชิงเส้นพหุเพื่อการพยากรณ์ เมื่อค่าสังเกตของตัวแปรตามเป็นค่าที่สูญหายแบบสุ่ม โดยทำการประมาณค่าสูญหายของตัวแปรตามทั้ง 5 วิธีดังนี้

1. วิธีสูญหาย คือวิธีที่ตัดชุดข้อมูลที่มีค่าสูยหายออก
2. วิธีค่าเฉลี่ย คือวิธีที่ใช้ค่าเฉลี่ยของข้อมูลไม่สูญหายประมาณข้อมูลสูญหาย
3. วิธีสมการถดถอย คือวิธีที่ประมาณข้อมูลสูญหายจากสมการถดถอยเชิงเส้นพหุ
4. วิธีอีเอ็ม (EM Algorithm)
5. วิธีการของฮันท์ (Hunt's Method)

การเปรียบเทียบจะเปรียบเทียบความคลาดเคลื่อนของแต่ละวิธี ด้วยค่ารากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง ของค่าพยากรณ์ที่ได้จากการประมาณข้อมูลสูญหายด้วยวิธีต่าง ๆ ที่ ขนาดตัวอย่าง 5 ระดับ , ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน 5 ระดับ , สัดส่วนการสูญหายของตัวแปรตาม 7 ระดับ และลักษณะของตัวแปรอิสระตามที่กำหนด 3 รูปแบบ โดยที่ในแต่ละสถานการณ์มีการทำซ้ำ 200 ครั้ง

เนื่องจากในการวิจัยครั้งนี้ ผู้วิจัยได้อาศัยเทคนิคการจำลอง แบบมอนติคาร์โล (Monte Carlo Simulation Technique) มาสร้างสถานการณ์ต่าง ๆ ดังนั้นจะขอเริ่มกล่าวถึงวิธีการจำลองโดยใช้เทคนิคมอนติคาร์โล แล้วจึงแสดงรายละเอียดของขั้นตอนการวิจัย และและโปรแกรมที่ใช้ในการวิจัยตามลำดับ

วิธีจำลองโดยใช้เทคนิคมอนติคาร์โล

เทคนิคที่ใช้แก้ปัญหาในการคำนวณทางสถิตินั้นมีอยู่หลายวิธี วิธีการจำลองโดยใช้เทคนิคมอนติคาร์โล เป็นวิธีหนึ่งที่นิยมนำมาใช้แก้ปัญหากันอย่างแพร่หลายในปัจจุบัน ซึ่งหลักการของการจำลองโดยใช้เทคนิคดังกล่าว จะใช้เลขสุ่ม (Random Numbers) มาช่วยในการหาคำตอบของปัญหาที่ต้องการศึกษา

ขั้นตอนของวิธีการจำลองด้วยเทคนิคมอนติคาร์โลที่ใช้กันในปัจจุบันแบ่งได้เป็น 3 ขั้นตอนดังนี้คือ

1. การสร้างตัวเลขสุ่ม การใช้ตัวเลขสุ่มเป็นสิ่งที่สำคัญมากในเทคนิคนี้ ทั้งนี้เพราะว่าหลักการของการจำลองแบบมอนติคาร์โลนั้น จะใช้ตัวเลขสุ่มมาช่วยในการหาคำตอบของปัญหาโดยลักษณะของตัวเลขสุ่มที่นำมาใช้ จะมีการแจกแจงแบบสม่ำเสมอในช่วง (0,1) สำหรับวิธีการสร้างตัวเลขสุ่ม มีผู้เสนอไว้หลายวิธี แต่วิธีที่ดีนั้นลักษณะของตัวเลขสุ่มที่ถูกสร้างขึ้นจะต้องมีการแจกแจงแบบสม่ำเสมอในช่วง (0,1) ตัวเลขสุ่มแต่ละตัวเป็นอิสระต่อกัน และมีช่วงยาวก่อนจะเกิดเลขสุ่มซ้ำ (มีวัฏจักรยาว)

2. การนำตัวเลขสุ่มมาประยุกต์ใช้กับปัญหาที่ต้องการศึกษา ซึ่งขั้นตอนนี้ขึ้นอยู่กับลักษณะของปัญหา บางปัญหาอาจจะไม่ใช้ตัวเลขสุ่มโดยตรง แต่จะนำไปผลิตเลขสุ่มที่มีการแจกแจงแบบอื่นต่อไป

3. การทดลองกระทำ เมื่อนำตัวเลขสุ่มมาประยุกต์ให้เข้ากับปัญหาที่ต้องการศึกษาได้แล้ว ขั้นต่อไปคือ การทดลองโดยใช้กระบวนการของการสุ่ม (Random Process) มากระทำในลักษณะซ้ำ ๆ กันหลาย ๆ ครั้ง เพื่อหาคำตอบที่ต้องการ

แผนการทดลอง

การวิจัยครั้งนี้ต้องการประมาณค่าสูญหายของตัวแปรตาม เพื่อการพยากรณ์ด้วยสมการถดถอยเชิงเส้นพหุ เมื่อค่าสังเกตของตัวแปรตามเป็นค่าที่สูญหายแบบสุ่ม โดยทำการประมาณค่าสูญหายของตัวแปรตามทั้ง 5 วิธี ที่ ขนาดตัวอย่าง 5 ระดับ , ค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน 5 ระดับ , สัดส่วนการสูญหายของตัวแปรตาม 7 ระดับ และลักษณะของตัวแปรอิสระตามที่กำหนด 3 รูปแบบ และใช้วิธีกำลังสองน้อยที่สุดหาสัมประสิทธิ์การถดถอยในแต่ละวิธี เพื่อนำมาพยากรณ์ค่าสังเกตของตัวแปรตาม จากนั้นจะทำการเปรียบเทียบค่าความคลาดเคลื่อนระหว่างค่าพยากรณ์ของตัวแปรตามกับค่าจริง ในรูปของค่ารากที่สองของ

ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของค่าพยากรณ์ที่ได้จากการประมาณค่าที่สูญหายทั้ง 5 วิธี เพื่อหาวิธีที่ดีที่สุดในแต่ละสถานการณ์ต่อไป

<u>ขั้นตอนในการวิจัย</u>

สมการถดถอยที่ใช้ในการวิจัยครั้งนี้จะใช้สมการถดถอยเชิงเส้นพหุ (Multiple Linear Regression Equation) โดยมีรูปแบบสมการดังนี้

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t \quad ; t = 1,..,n+m$$

เมื่อ y_t เป็นตัวแปรตาม

x_{1t}, x_{2t} เป็นตัวแปรอิสระตัวที่ 1 และ 2

β_i เป็นพารามิเตอร์ที่ไม่ทราบค่า $i = 0,1,2$

ε_t เป็นค่าความคลาดเคลื่อน

$n+m$ เป็นจำนวนค่าสังเกตทั้งหมด

n เป็นจำนวนค่าสังเกตที่ไม่สูญหาย

m เป็นจำนวนค่าสังเกตที่สูญหาย

หอสมุดกลาง สถาบันวิทยบริการ จุฬาลงกรณ์มหาวิทยาลัย

การจำลองการทดลองตามสถานการณ์ต่าง ๆ จะทำได้โดยการสร้างโปรแกรมคอมพิวเตอร์ด้วยภาษาฟอร์แทรน (FORTRAN 77) โดยใช้กับเครื่องคอมพิวเตอร์ AMDHAL 5850 เพื่อสร้างข้อมูลให้เป็นไปตามการวิจัย ซึ่งขั้นตอนของการวิจัยจะแบ่งออกเป็น 6 ขั้นตอน

1. สร้างตัวแปรอิสระ X ตามรูปแบบที่กำหนด
2. สร้างค่าความคลาดเคลื่อน ε_t ตามลักษณะที่กำหนด
3. สร้างค่าตัวแปรตาม Y_t
4. สุ่มตำแหน่งที่สูญหายของข้อมูล
5. ประมาณค่าที่สูญหายด้วยวิธีการทั้ง 5 วิธี
6. ประมาณค่าพารามิเตอร์เมื่อได้แทนข้อมูลที่สูญหายด้วยวิธี OLS
7. หาค่ารากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองของค่าพยากรณ์

4. สุ่มตำแหน่งที่สูญหายของข้อมูล

ทำการสุ่มตำแหน่งที่สูญหายในสัดส่วนที่กำหนด โดยใช้วิธีสร้างของ White และ Schmidt รายละเอียดแสดงในภาคผนวก ก และใส่รหัสให้กับตำแหน่งที่สูญหายดังนี้

$$CODE(t) = \begin{cases} 0 & \text{เมื่อค่าสังเกตที่ t ไม่สูญหาย} \\ 1 & \text{เมื่อค่าสังเกตที่ t สูญหาย} \end{cases}$$

5. ประมาณค่าที่สูญหายด้วยวิธีการทั้ง 5 วิธี ดังต่อไปนี้

5.1 วิธีสูญหาย เป็นวิธีที่ตัดชุดข้อมูลที่มีค่าสูญหายออก

5.2 วิธีค่าเฉลี่ย เป็นวิธีที่ใช้ค่าเฉลี่ยของข้อมูลที่ไม่สูญหายของตัวแปรตามมาประมาณข้อมูลสูญหาย

5.3 วิธีสมการถดถอย เป็นวิธีที่ประมาณข้อมูลสูญหายจากสมการถดถอยเชิงเส้นพหุของข้อมูลที่ไม่สูญหาย

5.4 วิธีอีเอ็ม (EM Algorithm) มีขั้นตอนดังนี้

5.4.1 เฉพาะข้อมูลที่ไม่สูญหาย ประมาณพารามิเตอร์เริ่มต้นด้วยวิธี OLS

5.4.2 ประมาณข้อมูลสูญหาย ด้วยค่าคาดหวังของค่าที่สูญหายไปภายใต้เงื่อนไขชุดข้อมูลที่ไม่สูญหายและพารามิเตอร์ตัวปัจจุบัน ($\hat{\beta}^{\#}$) (โดยครั้งแรกพารามิเตอร์ตัวปัจจุบันคือพารามิเตอร์เริ่มต้น) : $E(y_t|X, y_1, \hat{\beta}^{\#})$

$$E(y_t|X, y_1, \hat{\beta}^{\#}) = \begin{cases} y_t & \text{ถ้า } y_t \text{ ไม่สูญหาย} \\ X_2\hat{\beta}^{\#} & \text{ถ้า } y_t \text{ สูญหาย} \end{cases}$$

เมื่อ X_2 คือเมตริกซ์ของตัวแปรอิสระขนาด mx(p+1) เมื่อข้อมูลของตัวแปรตามสูญหาย

5.4.3 ประมาณพารามิเตอร์ตัวปัจจุบันด้วยการแทนค่าที่สูญหายที่ได้จากข้อที่แล้วโดยวิธี OLS

5.4.4 ทำซ้ำจนกระทั่งพารามิเตอร์ตัวปัจจุบันคงที่ นั่นคือเราได้ค่าประมาณค่าที่สูญหายจาก $E(y_t|X, y_1, \hat{\beta}^{*})$

5.5 วิธีการของฮันท์ (Hunt's Method) มีขั้นตอนดังนี้

5.5.1 ประมาณค่าที่สูญหายเริ่มต้น ด้วยการเฉลี่ยเฉพาะข้อมูลที่ไม่สูญหายของตัวแปรตาม

5.5.2 หาเวกเตอร์ของความคลาดเคลื่อนตัวปัจจุบัน ดังนี้

ε(ตัวปัจจุบัน) = (**I** - **P**) y(ตัวก่อน)

เมื่อ ε คือเวกเตอร์ของความคลาดเคลื่อน

y คือเวกเตอร์ของข้อมูลตัวแปรตามทั้งหมด (แทนข้อมูลที่สูญหายแล้ว โดยครั้งแรกแทนข้อมูลที่สูญหายด้วยค่าเฉลี่ย)

I คือเมตริกซ์เอกลักษณ์ และ $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$

5.5.3 ประมาณค่าที่สูญหายตัวปัจจุบัน ดังนี้

y_t(ตัวปัจจุบัน) = y_t(ตัวก่อน) - ε_t(ตัวปัจจุบัน)

5.5.4 ทำซ้ำจนกระทั่งค่าประมาณค่าที่สูญหายคงที่

6. ประมาณค่าพารามิเตอร์ที่มีข้อมูลสมบูรณ์ด้วยวิธี OLS

นำค่า y_t และคู่อันดับ x_{1t} , x_{2t} ที่ครบสมบูรณ์ ; t = 1,..,n+m คำนวณค่าประมาณพารามิเตอร์หรือสัมประสิทธิ์การถดถอยจากสมการถดถอยเชิงเส้นพหุ ด้วยวิธีกำลังสองน้อยที่สุด เพื่อหาสมการถดถอยเชิงเส้นพหุในการพยากรณ์ทั้ง 5 วิธี

7. หาค่ารากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองของค่าพยากรณ์มีขั้นตอนต่าง ๆ ดังนี้

7.1 สร้างค่า y จริงตามวิธีการที่กล่าวมาข้างต้นไปล่วงหน้า 12 ค่าสังเกตหรือคาบเวลา

7.2 ประมาณค่า $\hat{y}$ พยากรณ์ล่วงหน้า 12 ค่าสังเกตหรือคาบเวลา ด้วยสมการถดถอยเชิงเส้นพหุที่ใช้ในการพยากรณ์ดังที่กล่าวมาข้างต้น

7.3 หาค่ารากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองของค่าพยากรณ์ของแต่ละค่าสังเกตหรือคาบเวลาใน 200 รอบ โดยมีสูตรการคำนวณดังนี้

$$RMSE_t = \sqrt{\frac{\sum_{i=1}^{200}(y_{ti} - \hat{y}_{ti})^2}{200}}$$

โดยที่ y_{ti} คือค่าสังเกตของข้อมูลตัวที่ t หรือคาบเวลาที่ t ในการทำซ้ำรอบที่ i

$\hat{y}_{ti}$ คือค่าพยากรณ์ของข้อมูลตัวที่ t หรือคาบเวลาที่ t ในการทำซ้ำรอบที่ i

i คือจำนวนรอบของการทำซ้ำ ; i = 1,2,..,200

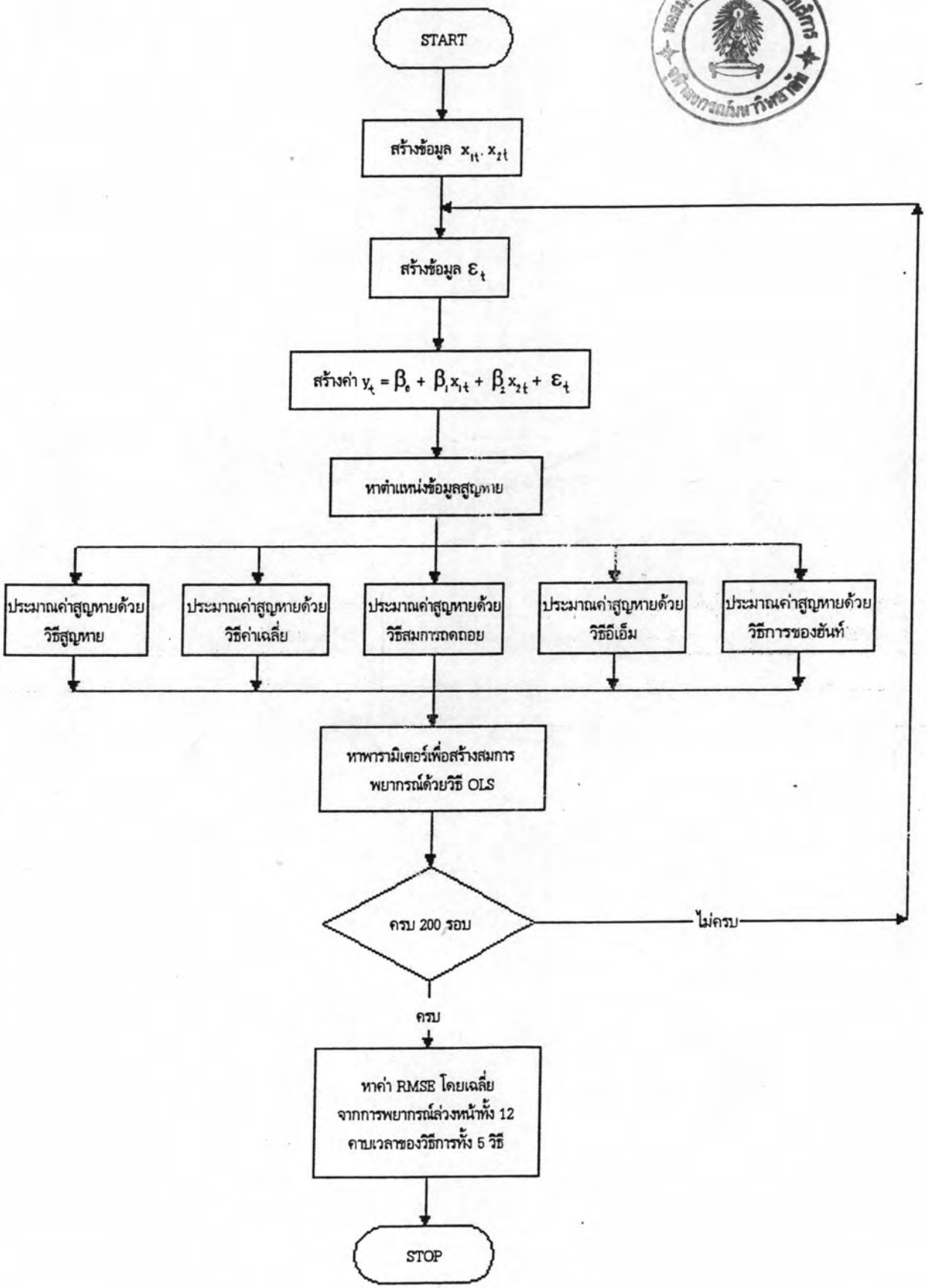
t คือค่าสังเกตของข้อมูลหรือคาบเวลาของการพยากรณ์ ; t = (n+m)+1,(n+m)+2,..,(n+m)+12

7.4 หาค่ารากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองของค่าพยากรณ์ โดยทำการเฉลี่ยในช่วง 12 ค่าสังเกตหรือคาบเวลา

$$RMSE = \frac{\sum_{t=1}^{12} RMSE_t}{12}$$

7.5 ทำการเปรียบเทียบค่ารากที่สองของค่าเฉลี่ย ของความคลาดเคลื่อนกำลังจากการพยากรณ์ (RMSE) ทั้ง 5 วิธี

การคำนวณ RMSE จากการพยากรณ์ เมื่อได้ประมาณข้อมูลที่สูญหายด้วยวิธีทั้ง 5 วิธีในการทดลองที่ขนาดตัวอย่างหนึ่ง ๆ จะเปลี่ยนค่าเบี่ยงเบนมาตรฐานของความคลาดเคลื่อน 5 ระดับ จะเปลี่ยนสัดส่วนการสูญหายของตัวแปรตาม 7 ระดับ และจะเปลี่ยนลักษณะของตัวแปรอิสระตามที่กำหนด 3 รูปแบบ โดยที่ในแต่ละสถานการณ์มีการทำซ้ำ 200 ครั้งจนครบทุกสถานการณ์ ซึ่งขั้นตอนของการทดลองดังกล่าวนี้ จะสรุปเป็นผังงานได้ดังรูป



รูปที่ 3.1 แสดงผังงานของขั้นตอนการดำเนินงาน

โปรแกรมที่ใช้ในการวิจัย

โปรแกรมที่ใช้ในการวิจัยทั้งหมด เขียนด้วยภาษาฟอร์แทรน (FORTRAN) โดยใช้กับเครื่อง AMDSHL 5860 ซึ่งในแต่ละสถานการณ์ของการทดลอง ลักษณะการทำงานของโปรแกรมจะเหมือนกัน สำหรับรายละเอียดของโปรแกรมจะแสดงไว้ในภาคผนวก ข ซึ่งจะเป็นโปรแกรมการทำงานของแต่ละวิธีการ