บทที่ 3

วิธีดำเนินการวิจัย

ในการศึกษาครั้งนี้สิ่งที่ต้องการคือการประมาณค่าตัวแปรตามในสมการถดถอยเชิงเส้นพหุ เมื่อตัวแปรตามบางค่าถูกตัดทิ้งทางขวาประเภทที่ 1 ซึ่งประกอบไปด้วยการประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอย 4 วิธี คือวิธีกำลังสองต่ำสุด วิธีการของแรตเตอร์จีและแมคลีช วิธีการของบัคเลย์และเจมส์ และวิธีภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนวิธีอีเอ็ม ในการศึกษาครั้งนี้เกณฑ์ในการเปรียบเทียบการประมาณพารามิเตอร์ด้วยวิธีใดจะให้ค่าการประมาณค่าตัวแปรตามที่ถูกตัดทิ้งได้ดีกว่า จะพิจารณาโดยการเปรียบเทียบค่าประมาณของตัวแปรตามกับค่าจริงก่อนถูกตัดทิ้ง ด้วยรากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (RMSE) เป็นตัววัดวิธีไหนจะให้ค่า RMSE น้อยที่สุดในการประมาณค่าตัวแปรตามในสมการถดถอย โดยจะศึกษาเมื่อค่าคลาดเคลื่อนมีการแจกแจงปกติ ดับเบิลเอกซโพเนนเชียล และแบบล็อกนอร์มอล ทั้งนี้เพราะว่าเมื่อค่าคลาดเคลื่อนแจกแจงปกติเป็นการแจกแจงที่เป็นไปตามข้อกำหนดพื้นฐาน จึงน่าสนใจศึกษาว่าเมื่อเปลี่ยนแปลงการแจกแจงความคลาดเคลื่อนเป็นการแจกแจงที่มีลักษณะการกระจายไปทางหางมากหรือหางยาวกว่าปกติ หรือการกระจายที่เบ้ไปทางขวามาก ๆ แล้วจะมีแนวโน้มออกมาเหมือนหรือใกล้เคียงกับการกระจายปกติหรือไม่ เนื่องจากในการนำไปประยุกต์ใช้กับข้อมูลจริง ซึ่งมักจะไม่ทราบการแจกแจงของความคลาดเคลื่อน แต่ผู้ใช้จะตั้งข้อสมมติว่าค่าคลาดเคลื่อนมีการแจกแจงปกติสำหรับรูปแบบของการแจกแจงปกติ และการแจกแจงดับเบิลเอกซโพเนนเชียล จะศึกษาที่ค่าสูงสุดของตัวแปรตามถูกตัดทิ้งที่ 53 , 58 และ 63 เมื่อความคลาดเคลื่อนแจกแจงล็อกนอร์มอล จะศึกษาที่ค่าสูงสุดของตัวแปรตามถูกตัดทิ้งที่ 55, 58 และ 61 ขนาดตัวอย่างที่สนใจศึกษาครั้งนี้เป็น 20 , 30, 40, 50, 60 และ 70 และจะศึกษาเมื่อเปอร์เซ็นต์การถูกตัดทิ้งเป็น 10% , 20%, 30% และ 40% ของขนาดตัวอย่าง และจำนวนตัวแปรอิสระ 2 ตัวแปร สำหรับการวิจัยครั้งนี้จะใช้วิธีมอนติคาร์โลในการสร้างข้อมูลให้มีสภาพการแจกแจงตามที่ต้องการ

การวิจัยครั้งนี้ใช้เทคนิคการจำลองแบบมอนติคาร์โล (Monte Carlo Simulation Technique) สร้างสถานการณ์ต่างๆ ดังนั้นในตอนแรกของบทนี้จะกล่าวถึงวิธีการจำลองโดยใช้เทคนิคมอนติคาร์โล แล้วจึงแสดงรายละเอียดของแผนการทดลอง ขั้นตอนการวิจัย และโปรแกรมที่ใช้ในการวิจัย ตามลำดับ

3.1 วิธีมอนติคาร์โล

เทคนิคในการจำลองตัวแบบทางคณิตศาสตร์นั้นมีอยู่หลายวิธีด้วยกัน ซึ่งวิธีของมอนติคาร์โลเป็นวิธีหนึ่งที่เป็นที่นิยมใช้ในปัจจุบัน ซึ่งหลักการของมอนติคาร์โลนั้นต้องจำลองตัวเลขสุ่ม (Random Number) เพื่อช่วยในการหาคำตอบของปัญหาที่ต้องการศึกษา

ในการศึกษาวิจัยครั้งนี้จะใช้เทคนิคของมอนติคาร์โลดังกล่าวในการสร้างข้อมูลให้มีสภาพการแจกแจงตามที่ต้องการศึกษา ซึ่งขั้นตอนของวิธีมอนติคาร์โลที่ใช้กันอยู่ในปัจจุบันแบ่งได้เป็น 3 ขั้นตอนดังนี้

3.1.1 การสร้างตัวเลขสุ่ม การใช้ตัวเลขสุ่มเป็นสิ่งสำคัญมากในวิธีมอนติคาร์โลทั้งนี้ก็เพราะว่าหลักการของมอนติคาร์โลนั้นจะใช้ตัวเลขสุ่มมาช่วยในการหาคำตอบของปัญหา ซึ่งลักษณะของตัวเลขสุ่มจะมีการแจกแจงแบบยูนิฟอร์มในช่วง (0,1) สำหรับวิธีการสร้างตัวเลขสุ่มนั้นมีผู้เสนอไว้หลายวิธีด้วยกัน แต่วิธีหนึ่งที่ดีคือ ไวท์และสมิดท์ (White and Schmidt 1975 : 421) เสนอไว้คือ ลักษณะของตัวเลขสุ่มที่เกิดขึ้นมีการแจกแจงแบบยูนิฟอร์มในช่วง (0,1) และเป็นอิสระกัน ซึ่งรายละเอียดแสดงไว้ในภาคผนวก ก.

3.1.2 การประยุกต์ของปัญหาที่ต้องการศึกษามาใช้กับตัวเลขสุ่ม ซึ่งขั้นตอนนี้ขึ้นอยู่กับลักษณะของปัญหาที่ต้องการศึกษา บางปัญหาอาจไม่ใช้กับตัวเลขสุ่มโดยตรง แต่อาจมีขั้นตอนอื่นอีกหลายๆ ขั้นตอนซึ่งขั้นตอนเหล่านี้บางขั้นตอนต้องใช้ตัวเลขสุ่ม

3.2 แผนการทดลอง

ในการศึกษาครั้งนี้ต้องการเปรียบเทียบการประมาณค่าตัวแปรตามในสมการถดถอยเชิงเส้นพหุ เมื่อค่าสังเกตของตัวแปรตามเป็นค่าที่ถูกตัดทิ้งทางขวาประเภทที่ 1 ด้วยวิธีการประมาณพารามิเตอร์ 4 วิธี คือวิธีกำลังสองต่ำสุด วิธีการของแฮคเกอร์ตีและแมคลีช วิธีการของบัคเลย์และเจมส์ และวิธีภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนวิธีอีเอ็ม ในการศึกษาครั้งนี้ ได้ทำการศึกษาเมื่อขนาดตัวอย่างมีค่าเป็น 20, 30, 40, 50, 60 และ 70 ซึ่งในแต่ละขนาดตัวอย่างจะเกิดเปอร์เซ็นต์การถูกตัดทิ้งเป็น 4 ระดับคือ 10%, 20%, 30% และ 40% โดยศึกษาภายใต้ค่าคลาดเคลื่อนมีการแจกแจง 3 แบบ คือ การแจกแจงปกติ แจกแจงคับเบิลเอกซโพเนนเชียล และแจกแจงล็อกนอร์มอล และแต่ละการแจกแจงกำหนดค่าสูงสุดที่ถูกตัดทิ้งของตัวแปรตาม 3 ระดับ รวมทั้งสิ้น 216 สถานการณ์ และทำการเปรียบเทียบความคลาดเคลื่อนระหว่างค่าประมาณของตัวแปรตามกับค่าจริงก่อนถูกตัดทิ้งในรูปของค่ารากที่สองของค่าเฉลี่ยของความคลาดเคลื่อนกำลังสอง (RMSE) ของทั้ง 4 วิธี เพื่อหาวิธีที่ดีที่สุดของแต่ละสถานการณ์ โดยมีข้อกำหนดดังนี้

1 ค่าสังเกต $T_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$ $i=1,2,...,N$ เมื่อ N คือ จำนวนค่าสังเกต โดยที่ X_{i1} เป็นค่าคงที่จำลองจากการแจกแจงไวบูลล์ (Weibull Distribution) $W(\alpha,\beta)$ เมื่อ α มีค่าเป็น 2 และ β มีค่าเป็น 2 และ X_{i2} เป็นค่าคงที่จำลองจากการแจกแจงปกติ (Normal Distribution) $N(\mu,\sigma^2)$ เมื่อ μ มีค่าเป็น 10 และ σ^2 มีค่าเป็น 12 และ ε_i มี 3 การแจกแจงคือ แจกแจงปกติ $N(0,100)$ แจกแจงดับเบิลเอกซโพเนนเชียล $DB(0,\sqrt{50})$ และแจกแจงล็อกนอร์มอล $LN(1,1.1025)$ โดยกำหนด $\beta_0 = 2$ $\beta_1 = 3.5$ และ $\beta_2 = 3.5$

2 กำหนดค่าสูงสุด (Tc) ของตัวแปรตามถูกตัดทิ้ง 3 ระดับ โดยเมื่อค่าคลาดเคลื่อนมีการแจกแจงปกติ $N(0,100)$ และแจกแจงดับเบิลเอกซโพเนนเชียล $DB(0,\sqrt{50})$ กำหนดค่าสูงสุด Tc ที่ 53, 58 และ 63 และเมื่อแจกแจงแบบล็อกนอร์มอล $LN(1,1.1025)$ กำหนดค่าสูงสุด Tc ที่ 55, 58 และ 61

3 ขนาดตัวอย่างที่สนใจศึกษา คือ 20, 30, 40, 50, 60 และ 70

4 เปอร์เซ็นต์ของข้อมูลที่ถูกตัดทิ้งมี 4 ระดับ คือ 10%, 20%, 30% และ 40%

3.3 ขั้นตอนในการศึกษาวิจัย

ขั้นตอนในการศึกษาวิจัยมีดังนี้

1. จำลองค่าคงที่เป็นตัวแปรอิสระ X_{i1} และ X_{i2} โดย X_{i1} จำลองจากการแจกแจงไวบูลล์ $W(\alpha,\beta)$ และ X_{i2} การแจกแจงปกติ $N(\mu,\sigma^2)$ และจำลองค่าคลาดเคลื่อนตามการแจกแจงที่ต้องการศึกษาและจำลอง T_i จากความสัมพันธ์เชิงเส้น

$$T_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

2. กำหนดค่าสูงสุดของ T_i ที่ถูกตัดทิ้งที่ Tc

3. หาค่าสังเกตของตัวแปรตาม Y_i : $i = 1,2,3,\ldots,n+m$

$$Y_i = \begin{cases} T_i & ; T_i \le Tc \quad ; i = 1,2,\ldots,n \\ Tc & ; T_i > Tc \quad ; i = n+1, n+2,\ldots,n+m \end{cases}$$

n เป็นจำนวนตัวอย่างที่ไม่ถูกตัดทิ้ง

m เป็นจำนวนตัวอย่างที่ถูกตัดทิ้ง

4. นำค่า Y_i, X_{i1} และ X_{i2} ประมาณค่าพารามิเตอร์สัมประสิทธิ์ความถดถอยเชิงเส้นพหุด้วยวิธี

4.1 วิธีการของบัคเลย์และเจมส์

4.2 วิธีการของแชดเทอร์จีและแมคลีช

4.3 วิธีกำลังสองต่ำสุด

4.4 วิธีภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนวิธีอีเอ็ม

5. หาความคลาดเคลื่อนจากการประมาณค่าตัวแปรตาม และทำการเปรียบเทียบ

สำหรับรายละเอียดแต่ละขั้นตอนเป็นดังนี้

3.3.1 การจำลองตัวแปรอิสระ X_{i1} และ X_{i2} ซึ่งเป็นค่าคงที่ โดยที่ X_{i1} จำลองจากการแจกแจงไวบูลล์ (Weibull Distribution) $W(\alpha,\beta)$ โดย α มีค่าเป็น 2 และ β มีค่าเป็น 2 และ X_{i2} จำลองจากการแจกแจงปกติ $N(\mu,\sigma^2)$ มีค่าเฉลี่ยเป็น 10 ความแปรปรวนเป็น 12 และจำลองค่าคลาดเคลื่อน ε_i จากการแจกแจง 3 แบบคือ

ก. การแจกแจงปกติ $N(\mu,\sigma^2)$ ด้วยค่าเฉลี่ยเป็น 0 และความแปรปรวนเท่ากับ 100 (รายละเอียดในภาคผนวก ก.)

ข. การแจกแจงหางยาวกว่าการแจกแจงปกติ โดยจำลอง ε_i จากการแจกแจงดับเบิลเอกซโพเนนเชียล $DB(\mu,\sigma)$ ด้วยค่าพารามิเตอร์ μ เป็น 0 และ σ เป็น $\sqrt{50}$ (รายละเอียดในภาคผนวก ก.)

ค. การแจกแจงที่เบ้ขวา โดยจำลอง ε_i จากการแจกแจงล็อกนอร์มอล $LN(1,1.1025)$

3.3.2 กำหนดค่าสูงสุดของ T_i ที่จะถูกตัดทิ้ง ในการศึกษาครั้งนี้กำหนดค่าสูงสุดของ T_i ที่จะถูกตัดทิ้งไว้ที่ 3 ระดับทำได้โดยจากรูปแบบ $T_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$: $i=1,2,...,n+m$ เมื่อจำลอง X_{i1}, X_{i2} และกำหนด β_0, β_1 และ β_2ตามรายละเอียดในข้อ 3.3.1 และนำ T_i หาค่าเฉลี่ย ให้ μ_T คือค่าเฉลี่ยของ T_i กำหนดค่าสูงสุดที่จะถูกตัดทิ้งระดับที่ 1,2 และ 3 ที่ $\mu_T+\sigma_T$, $\mu_T+1.5\sigma_T$ และที่ $\mu_T+2\sigma_T$ ตามลำดับ เมื่อ σ_T^2 คือความแปรปรวนของค่าคลาดเคลื่อนที่มีการแจกแจงตามที่ต้องการศึกษา ดังนั้นค่าสูงสุดที่ถูกตัดทิ้งสรุปได้ดังนี้

ก. เมื่อ ε_i แจกแจงแบบปกติ ตัดทิ้งระดับที่ 1,2, 3 ที่ Tc=53, 58 และที่ Tc =63 ตามลำดับ

ข. เมื่อ ε_i แจกแจงแบบดับเบิลเอกซโพเนนเชียล ตัดทิ้งระดับที่ 1, 2, 3 ที่ Tc= 53, 58 และระดับที่ 3 ที่ Tc =63

ค. เมื่อ ε_i แจกแจงแบบล็อกนอร์มอล ตัดทิ้งระดับที่ 1, 2, 3 ที่ Tc =55, 58 และระดับที่ 3 ที่ Tc =61

3.3.3 หาค่าสังเกตของตัวแปรตาม Y_i ; $i=1,2,...,n+m$ เมื่อ $n+m$ คือจำนวนตัวอย่างทั้งหมด n คือจำนวนตัวอย่างที่ไม่ถูกตัด m คือจำนวนตัวอย่างที่ถูกตัดทิ้ง

$$Y_i = \begin{cases} T_i & \text{ถ้า } T_i \le Tc \quad \text{(ไม่ถูกตัดทิ้ง)} \\ Tc & \text{ถ้า } T_i > Tc \quad \text{(ถูกตัดทิ้ง)} \end{cases}$$

$$\delta_i = \begin{cases} 1 & \text{ถ้า } T_i \le Tc \quad \text{(ไม่ถูกตัดทิ้ง)} \\ 0 & \text{ถ้า } T_i > Tc \quad \text{(ถูกตัดทิ้ง)} \end{cases}$$

เมื่อ Tc คือค่าสูงสุดที่ถูกตัดทิ้ง

การหาค่าสังเกตของตัวแปรตาม Y_i ทำได้โดยจำลอง T_i แล้วนำค่า T_i มาเปรียบเทียบกับค่า Tc จะได้ค่า $Y_i = T_i$ เมื่อ $T_i \le Tc$ และให้ δ_i = 1 แล้วนับเป็นค่าสังเกตที่ไม่ถูกตัดทิ้ง และจะได้ $Y_i = Tc$ เมื่อ $T_i > Tc$ และให้ δ_i = 0 แล้วนับเป็นค่าสังเกตที่ถูกตัดทิ้ง และทำการทดลองเช่นนี้จนกระทั่งได้ค่าสังเกตที่ไม่ถูกตัดทิ้งและถูกตัดทิ้งครบตามจำนวน n และ m ตามที่กำหนด

3.3.4 นำค่า Y_i และ X_{i1}, X_{i2} : $i=1,2,\ldots,n+m$ คำนวณค่าประมาณพารามิเตอร์จากสมการถดถอยเชิงเส้นพหุ ด้วยวิธีกำลังสองต่ำสุด วิธีการของแชดเทอร์จีและแมคลีช วิธีการของบัคเลย์และเจมส์ และวิธีการภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนวิธีอีเอ็ม เนื่องจากวิธีกำลังสองต่ำสุดและวิธีการของแชดเทอร์จีและแมคลีชสามารถประมาณพารามิเตอร์ได้ (Convergence) ทุกสถานการณ์ที่ศึกษา แต่วิธีการของบัคเลย์และเจมส์ และวิธีภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนวิธีอีเอ็ม จะมีบางชุดของข้อมูลเช่นข้อมูลที่มีเปอร์เซ็นต์การถูกตัดทิ้งจำนวนมาก ค่าประมาณพารามิเตอร์จะแกว่งอยู่ระหว่าง 2 ค่า ก็จะใช้ค่าเฉลี่ยระหว่าง 2 ค่านั้นเป็นค่าประมาณพารามิเตอร์หรือค่าประมาณของพารามิเตอร์ของรอบปัจจุบันกับรอบที่ผ่านมาต่างกันน้อยกว่า 0.005 ถือว่าค่าประมาณพารามิเตอร์นั้นหาค่าได้ ถ้าไม่เข้าข่ายทั้ง 2 กรณี ก็จะทิ้งข้อมูลชุดนั้น และทำการจำลองข้อมูลที่หาค่าประมาณพารามิเตอร์ได้ทั้ง 4 วิธีจำนวน 1000 ชุดข้อมูลในแต่ละสถานการณ์

การประมาณพารามิเตอร์แต่ละวิธีมีรายละเอียดดังนี้

3.3.4.1 กำลังสองต่ำสุด มีขั้นตอนเป็นดังนี้

ประมาณค่าพารามิเตอร์ $\hat{\beta}$ จากข้อมูลที่ถูกตัดทิ้งและไม่ถูกตัดทิ้ง โดยการหาอนุพันธ์ของผลบวกกำลังสองของความคลาดเคลื่อนเทียบกับ $\hat{\beta}$ แล้วกำหนดให้เท่ากับ $\underset{\sim}{0}$

จากนั้นแก้สมการหาค่า $\hat{\beta}$ ได้ดังนี้ $\frac{\partial}{\partial \hat{\beta}}\left(YY - 2\hat{\beta}X'Y + \hat{\beta}XX\hat{\beta}\right) = \underset{\sim}{0}$

เพราะฉะนั้น $\hat{\beta} = (XX)^{-1}XY'$

3.3.4.2 วิธีการของแชตเทอร์จีและเมคลีช มีขั้นตอนดังนี้

1. เฉพาะข้อมูลที่ไม่ถูกตัดทิ้ง ประมาณค่าพารามิเตอร์เริ่มต้น $\hat{\beta}$ โดยวิธีกำลังสองต่ำสุด

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

เมื่อ u หมายถึงค่าสังเกตที่ไม่ถูกตัดทิ้ง

X^{u} เป็นเมตริกซ์ของตัวแปรอิสระของค่าสังเกตที่ไม่ถูกตัดทิ้ง

Y^{u} เป็นเวกเตอร์ของตัวแปรตามของค่าสังเกตที่ไม่ถูกตัดทิ้ง

2. ประมาณค่าสังเกตที่ถูกตัดทิ้ง $Y_i(\hat{\beta})$ โดย

$$Y_i(\hat{\beta}) = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2}$$

3. เปรียบเทียบค่า $Y_i(\hat{\beta})$ กับค่า Tc ถ้าค่า $Y_i(\hat{\beta})$ มีค่ามากกว่าค่า Tc ให้แทนที่ค่าถูกตัดทิ้งด้วย $Y_i(\hat{\beta})$ แต่ถ้าค่า $Y_i(\hat{\beta})$ มีค่าน้อยกว่าหรือเท่ากับ Tc ก็ให้คงค่าสังเกตที่ถูกตัดทิ้งตัวเดิมไว้

4. นำค่าสังเกตที่ไม่ถูกตัดทิ้งและถูกตัดทิ้งมาประมาณพารามิเตอร์ด้วยวิธีกำลังสองต่ำสุด

I1743306X

วิธีการของแชดเทอร์จีและแมคลีชใช้การวนทำซ้ำจนกระทั่งได้พารามิเตอร์ของรอบปัจจุบันเท่ากับรอบที่ผ่านมาถ้าพารามิเตอร์ยังไม่เท่ากันก็จะทำการประมาณค่าสังเกตที่ถูกตัดทิ้งและนำข้อมูลทั้งหมดคำนวณหาพารามิเตอร์รอบใหม่ หรือถ้าพารามิเตอร์แกว่งอยู่ระหว่าง 2 ค่าก็จะใช้ค่าเฉลี่ยของ 2 ค่านั้นเป็นค่าประมาณพารามิเตอร์

3.3.4.3 วิธีของบัคเลย์และเจมส์ มีขั้นตอนดังนี้

1. เฉพาะข้อมูลที่ไม่ถูกตัดทิ้งประมาณพารามิเตอร์เริ่มต้น $\hat{\underset{\sim}{\beta}}$ ด้วยวิธีกำลังสองต่ำสุด

$$\hat{\underset{\sim}{\beta}} = \left(X^{u'}X^{u}\right)^{-1}X^{u'}\underset{\sim}{Y}^{u}$$

เมื่อ u หมายถึงค่าสังเกตที่ไม่ถูกตัดทิ้ง

X^{u} เป็นเมตริกซ์ของตัวแปรอิสระของค่าสังเกตที่ไม่ถูกตัดทิ้ง

$\underset{\sim}{Y}^{u}$ เป็นเวกเตอร์ของตัวแปรตามของค่าสังเกตที่ไม่ถูกตัดทิ้ง

2. หาค่าความคลาดเคลื่อนบางส่วน (Partial Residuals) โดยคำนวณที่ $\hat{\beta}_0 = 0$, $\hat{e}_i = Y_i - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2}$; $i = 1,2,\ldots,n+m$ และนำ $\hat{e}_i$ มาเรียงลำดับจากน้อยไปมาก $\hat{e}_{(1)} < \hat{e}_{(2)} <,\ldots,< \hat{e}_{(n+m)}$ แล้วคำนวณค่าประมาณ $\hat{F}(\hat{e}_i)$ โดยใช้ตัวประมาณพีแอล

$$\hat{F}(\hat{e}_i) = 1 - \prod_{i:\hat{e}_{(i)} \le \hat{e}_i} \left(\frac{n+m-i}{n+m-i+1}\right)^{\delta(i)}$$

เมื่อ n คือจำนวนค่าสังเกตที่ไม่ถูกตัดทิ้ง

m คือจำนวนค่าสังเกตที่ถูกตัดทิ้ง

3. นำค่า $\hat{F}(\hat{e}_i)$ มาหาค่าถ่วงน้ำหนัก $W(e_i, \hat{\beta})$

4. ประมาณค่าสังเกตที่ถูกตัดทิ้งด้วยค่าคาดหวังอย่างมีเงื่อนไข

$$E\left(Y_i / Y_i > Tc, X\ \hat{\underset{\sim}{\beta}}\right) = Y_i(\hat{\beta})$$

$$Y_i(\hat{\beta}) = \left(\hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2}\right) + \frac{\sum_{u} W_k(e_k, \hat{\beta})(Y_k - \hat{\beta}_1 X_{k1} - \hat{\beta}_2 X_{k2})}{\left\{1 - \hat{F}(Tc - \hat{\beta}_1 X_{i1} - \hat{\beta}_2 X_{i2}\right\}}$$

5. นำค่าประมาณของค่าสังเกตที่ถูกตัดทิ้งรวมกับค่าสังเกตที่ไม่ถูกตัดทิ้ง มาคำนวณหาค่าประมาณพารามิเตอร์ด้วยวิธีกำลังสองต่ำสุด

วิธีการของบัคเลย์และเจมส์จะทำงานซ้ำจนกระทั่งได้ค่าประมาณพารามิเตอร์รอบปัจจุบันเท่ากับรอบที่ผ่านมาถ้าพารามิเตอร์ยังไม่เท่ากันก็จะทำการประมาณค่าสังเกตที่ถูกตัดทิ้งและนำข้อมูลทั้งหมดคำนวณหาพารามิเตอร์รอบใหม่ หรือถ้าพารามิเตอร์แกว่งอยู่ระหว่าง 2 ค่าก็จะใช้ค่าเฉลี่ยของ 2 ค่านั้นเป็นค่าประมาณพารามิเตอร์

3.3.4.4 วิธีการภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนวิธีอีเอ็ม มีขั้นตอนดังนี้

1. นำข้อมูลทั้งหมดทั้งที่ไม่ถูกตัดทิ้งและถูกตัดทิ้งประมาณพารามิเตอร์ $\hat{\beta}$, $\hat{\sigma}^2$ เริ่มต้น (k=1) ด้วยวิธีกำลังสองต่ำสุด

2. นำพารามิเตอร์รอบที่ k แทนค่าเพื่อประมาณค่าสังเกตที่ถูกตัดทิ้งด้วย $E\left(Y_t / Y_t > Tc, \beta, \sigma\right) = \hat{\mu}^{(k)} + \hat{\sigma}_{(k)} h\left(\hat{z}_t^{(k)}\right)$ แล้วนำค่าสังเกตที่ไม่ถูกตัดและถูกตัดประมาณ $\hat{\beta}$ รอบที่ k+1 ด้วยวิธีกำลังสองต่ำสุด

3. ประมาณ $\hat{\sigma}^2_{(k+1)}$ รอบที่ k+1 จาก

$$\hat{\sigma}^2_{(k+1)} = \left[\sum_{1}^{n}\left(y_t - \hat{\mu}_t^{(k)}\right) + \hat{\sigma}^2_{(k)} \sum_{n+1}^{n+m}\left(1 + \hat{z}_t^{(k)} h\left(\hat{z}_t^{(k)}\right)\right)\right] \Big/ (n+m)$$

การประมาณด้วยภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนวิธีอีเอ็มใช้กระบวนการวนทำซ้ำจนกระทั่งได้ค่าประมาณพารามิเตอร์รอบปัจจุบันเท่ากับรอบที่ผ่านมาถ้าพารามิเตอร์ยังไม่เท่ากันก็จะทำการประมาณค่าสังเกตที่ถูกตัดทิ้งและนำข้อมูลทั้งหมดคำนวณหาพารามิเตอร์รอบใหม่ หรือถ้าพารามิเตอร์แกว่งอยู่ระหว่าง 2 ค่าก็จะใช้ค่าเฉลี่ยของ 2 ค่านั้นเป็นค่าประมาณพารามิเตอร์

3.3.5 หาความคลาดเคลื่อนจากการประมาณค่าตัวแปรตาม โดยการเปรียบเทียบกับค่าจริงก่อนถูกตัดทิ้ง เพื่อนำมาคำนวณหาค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (RMSE) เนื่องจากการทดลองได้ทำซ้ำๆ กัน 1000 รอบ ดังนั้นในแต่ละสถานการณ์ให้ j แทนรอบที่ทำซ้ำ j=1,2, . . . , 1000 คำนวณรากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองหาได้จาก

$$MSE_j = \frac{1}{m}\sum_{i=1}^{m}\left(T_i - \hat{T}_i\right)^2$$

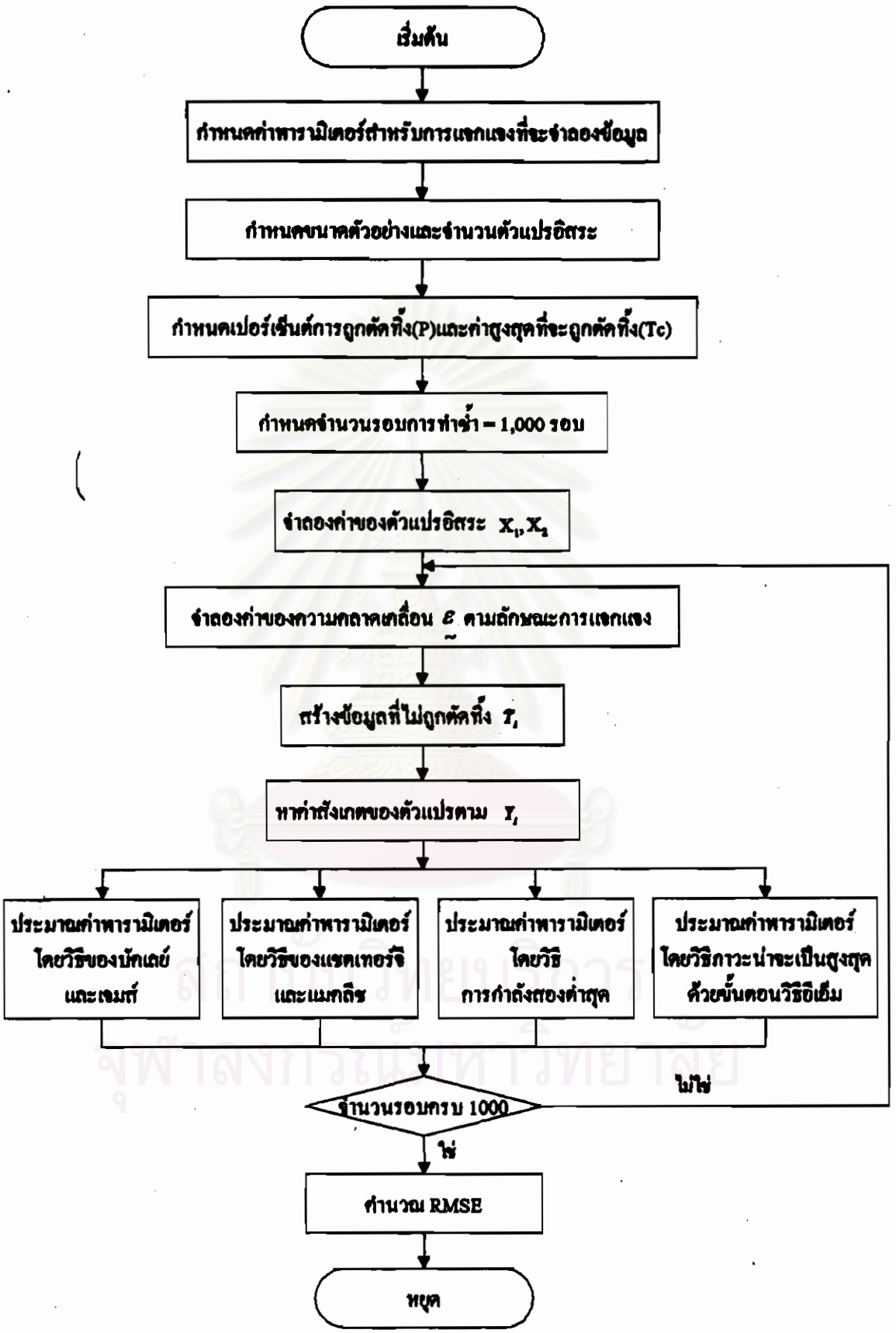
$$RMSE_j = \sqrt{MSE_j}$$

$$RMSE = \frac{1}{1000}\sum_{j=1}^{1000} RMSE_j$$

จากนั้นจึงนำ RMSE ของการประมาณตัวแปรตามเปรียบเทียบกันทั้ง 4 วิธี เพื่อหาว่าวิธีการใดให้ค่า RMSE ของการประมาณค่าตัวแปรตามต่ำที่สุดจะเป็นวิธีที่ประมาณค่าตัวแปรตามได้ดีที่สุดในแต่ละสถานการณ์

ในการหาค่า RMSE ของการประมาณทั้งหมด 4 วิธี ในขนาดตัวอย่างหนึ่งๆ จะเปลี่ยนเปอร์เซ็นต์การถูกตัดทิ้งเป็น 10%,20%,30% และ 40% และเปลี่ยนการแจกแจงของค่าคลาดเคลื่อน เป็น 3 การแจกแจงคือ แจกแจงปกติ แจกแจงดับเบิลเอกซโพเนนเชียล และแจกแจงล็อกนอร์มอล โดยแต่ละสถานการณ์ทำซ้ำๆ กัน 1000 รอบ ซึ่งขั้นตอนการทดลองสามารถสรุปเป็นผังงานได้ดังรูปที่ 3.1

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย



รูปที่ 3.1 แสดงผังงานสำหรับหาค่าความคลาดเคลื่อนจากการประมาณทั้ง 4 วิธีการ

3.4 โปรแกรมที่ใช้ในการวิจัย

โปรแกรมที่ใช้ในการวิจัยทั้งหมด เขียนด้วยภาษาฟอร์แทรน (FORTRAN) โดยใช้กับเครื่อง AMDAHL 5860 ซึ่งในแต่ละสถานการณ์ของการทดลองลักษณะการทำงานของโปรแกรมจะเหมือนกัน สำหรับรายละเอียดของโปรแกรมจะแสดงไว้ในภาคผนวก ข ซึ่งจะเป็นโปรแกรมการทำงานของแต่ละวิธีการ คือ วิธีการของบัคเลย์และเซมส์ วิธีการของแชตเทอร์จีและแมคลีช วิธีกำลังสองต่ำสุด และวิธีภาวะน่าจะเป็นสูงสุดด้วยขั้นตอนวิธีอีเอ็ม

สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย