บทที่ 3

วิธีดำเนินการวิจัย

ในการศึกษาครั้งนี้สิ่งที่ต้องการคือการประมาณค่าตัวแปรตามในสมการถดถอยเชิงเส้นอย่างง่าย เมื่อตัวแปรตามบางค่าถูกตัดทิ้งทางขวา ประเภทที่ 1 โดยการประมาณค่าพารามิเตอร์ด้วย 4 วิธี คือ วิธีกำลังสองต่ำสุด วิธีตัวประมาณของมิลเลอร์ วิธีกำลังสองต่ำสุดแบบตัดแปลงแค็พแลน-ไมเออร์ และวิธีการของบัคเลย์และเจมส์ ซึ่ง 3 วิธีหลังเป็นวิธีที่ใช้การกระทำวนซ้ำ(Iterative) จนกระทั่งค่าประมาณพารามิเตอร์ในรอบปัจจุบันเท่ากับค่าประมาณพารามิเตอร์ในรอบที่ผ่านมา ในการศึกษาครั้งนี้เกณฑ์ในการเปรียบเทียบการประมาณพารามิเตอร์ด้วยวิธีใดจะให้ค่าการประมาณค่าตัวแปรตามได้ดีกว่า จะพิจารณาโดยการเปรียบเทียบค่าประมาณของตัวแปรตามกับค่าจริงก่อนถูกตัดทิ้ง ด้วยรากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง(RMSE) โดยวิธีใดให้ค่า RMSE น้อยที่สุด จะเป็นวิธีที่ดีกว่า การวิเคราะห์การถดถอยโดยทั่วไปจะทำการศึกษาเมื่อค่าคลาดเคลื่อนแจกแจงแบบปกติ ซึ่งเป็นการแจกแจงที่เป็นไปตามข้อกำหนดพื้นฐาน จึงน่าสนใจศึกษาว่าเมื่อค่าคลาดเคลื่อนมีการแจกแจงเป็นแบบเบ้แล้วจะมีแนวโน้มเหมือนหรือต่างไปจากเมื่อค่าคลาดเคลื่อนมีการแจกแจงแบบปกติหรือไม่ และเนื่องจากข้อมูลทางด้านการประกันภัยส่วนใหญ่จะเป็นข้อมูลที่ถูกตัดทิ้งบางส่วน เช่น ข้อมูลเกี่ยวกับค่ารักษาพยาบาลของผู้เอาประกันภัย เป็นต้น และข้อมูลดังกล่าวส่วนใหญ่จะมีการแจกแจงแบบเบ้ ดังนั้นจึงทำการศึกษาเมื่อค่าคลาดเคลื่อนมีการแจกแจงปกติ ลอกนอร์มอล และไวบูลล์ และเมื่อค่าคลาดเคลื่อนแจกแจงปกติและแจกแจงลอกนอร์มอลกำหนดค่าสูงสุดของตัวแปรตามถูกตัดทิ้งที่ 8.5, 11.5 และ 14.5 ส่วนเมื่อค่าคลาดเคลื่อนแจกแจงไวบูลล์กำหนดค่าสูงสุดของตัวแปรตามถูกตัดทิ้งที่ 10, 12 และ 14 ขนาดตัวอย่างที่ศึกษาเป็น 20 , 30 , 40 , 50 และ 60 เปอร์เซ็นต์การถูกตัดทิ้งเป็น 10% , 20% , 30% และ 40% ของขนาดตัวอย่าง และมีตัวแปรอิสระ 1 ตัวซึ่งศึกษาการแจกแจง 2 แบบ คือ การแจกแจงปกติและแจกแจงไวบูลล์ ซึ่งสถานการณ์ต่างๆ ในการวิจัยนี้สร้างขึ้นจากการจำลองด้วยเทคนิคการจำลองแบบมอนติคาร์โล(Monte Carlo Simulation Technique) โดยมีรายละเอียดของแผนการทดลอง ขั้นตอนในการวิจัย และขั้นตอนการทำงานของโปรแกรมดังต่อไปนี้

3.1 แผนการทดลอง

ในการศึกษาครั้งนี้ต้องการเปรียบเทียบการประมาณค่าตัวแปรตามในสมการถดถอยเชิงเส้นอย่างง่าย เมื่อตัวแปรตามบางค่าถูกตัดทิ้งทางขวา ประเภทที่ 1 โดยการประมาณค่าพารามิเตอร์ด้วย 4 วิธี คือ วิธีกำลังสองต่ำสุด วิธีตัวประมาณของมิลเลอร์ วิธีกำลังสองต่ำสุดแบบตัดแปลงแค็พแลน-ไมเออร์ และวิธีการของบัคเลย์และเจมส์ ซึ่ง 3 วิธีหลังจะใช้การกระทำวนซ้ำ (Iterative) จนกระทั่งค่าประมาณพารามิเตอร์ในรอบปัจจุบันเท่ากับค่าประมาณพารามิเตอร์ในรอบที่ผ่านมา ในการศึกษาครั้งนี้ได้ทำการศึกษาเมื่อขนาดตัวอย่างมีค่าเป็น 20 , 30 , 40 , 50 และ 60 และเปอร์เซ็นต์การถูกตัดทิ้งเป็น 10% , 20% ,30% และ 40% ของขนาดตัวอย่าง โดยศึกษาภายใต้ค่าคลาดเคลื่อนมีการแจกแจง 3 แบบ คือ แจกแจงปกติ ลอกนอร์มอล และไวบูลล์ โดยที่ การแจกแจงปกติและลอกนอร์มอลกำหนดค่าสูงสุดของตัวแปรตามถูกตัดทิ้ง 3 ระดับ ที่ 8.5 , 11.5 และ 14.5 ส่วนการแจกแจงไวบูลล์ค่าสูงสุดของตัวแปรตามถูกตัดทิ้งที่ 10, 12 และ 14 ศึกษาเมื่อมีตัวแปรอิสระ 1 ตัวมีการแจกแจง 2 แบบ คือ การแจกแจงปกติ และ แจกแจงไวบูลล์ ดังนั้นรวมทั้งสิ้นเป็น 360 สถานการณ์ และทำการเปรียบเทียบความคลาดเคลื่อนระหว่างค่าประมาณของตัวแปรตามกับค่าจริงก่อนถูกตัดทิ้ง ด้วยค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง(RMSE) ของทั้ง 4 วิธี เพื่อหาวิธีที่ดีที่สุดของแต่ละสถานการณ์ โดยมีข้อกำหนดดังนี้

1. ค่าสังเกต $T_i = \alpha + \beta x_i + \varepsilon_i \quad ; i = 1,2,3,...,N$, N คือ จำนวนค่าสังเกตกำหนด α = 4.43 และ β = 0.03 X_i เป็นค่าคงที่จำลองจากการแจกแจง 2 แบบ คือ การแจกแจงปกติ(Normal Distribution) N (μ,σ^2) เมื่อ μ = 34 และ σ^2 = 144 และการแจกแจงไวบูลล์(Weibull Distribution) W(α',β') เมื่อ α' = 3.3 และ β' = 38 และ ε_i มี 3 การแจกแจง คือ แจกแจงปกติ N(0,36), แจกแจงลอกนอร์มอล(Lognormal Distribution) LN (μ,σ^2) โดย μ = -5.5 และσ^2 = 7.3 ซึ่งจะทำให้มีค่าเฉลี่ยเท่ากับ 0.16 และความแปรปรวนเท่ากับ 36 ส่วนการแจกแจงไวบูลล์ W(0.5,1) จะมีค่าเฉลี่ยเท่ากับ 2 และความแปรปรวนเท่ากับ 20

2. เมื่อค่าคลาดเคลื่อนแจกแจงปกติและลอกนอร์มอล กำหนดค่าสูงสุด (T_c) ของตัวแปรตามถูกตัดทิ้งที่ 3 ระดับ คือ 8.5, 11.5 และ 14.5 ตามลำดับ และเมื่อค่าคลาดเคลื่อนแจกแจงไวบูลล์กำหนดค่าสูงสุดของตัวแปรตามถูกตัดทิ้งที่ 10, 12 และ 14

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

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

3.2 ขั้นตอนในการศึกษาวิจัย

ขั้นตอนในการศึกษาวิจัยมีดังนี้

3.3.1. จำลองตัวแปรอิสระ X_i เป็นค่าคงที่ โดยจำลองตามการแจกแจงที่ต้องการศึกษา และจำลองค่าคลาดเคลื่อนตามการแจกแจงที่ต้องการศึกษา และจำลอง T_i จากความสัมพันธ์เชิงเส้น

$$T_i = \alpha + \beta x_i + \varepsilon_i$$

3.3.2 กำหนดค่าสูงสุดของ T_i ที่ถูกตัดทิ้ง ที่ T_c

3.3.3 หาค่าสังเกตของตัวแปรตาม Y_i ; i = 1,2,3, ... ,N

$$Y_i = \begin{cases} T_i \; ; \; T_i \leq T_c \\ \\ T_c \; ; \; T_i > T_c \end{cases}$$

$$\delta_i = \begin{cases} 1 \; ; \; T_i \leq T_c \\ \\ 0 \; ; \; T_i > T_c \end{cases}$$

$\delta_i = 1$ หมายถึง เป็นค่าสังเกตที่ไม่ถูกตัดทิ้ง

$\delta_i = 0$ หมายถึง เป็นค่าสังเกตที่ถูกตัดทิ้ง

3.3.4 นำค่า Y_i และ X_i มาทำการประมาณค่าพารามิเตอร์ในสมการถดถอยเชิงเส้นอย่างง่าย ด้วยวิธี

1. วิธีกำลังสองต่ำสุด
2. วิธีตัวประมาณของมิลเลอร์
3. วิธีกำลังสองต่ำสุดแบบดัดแปลงแค็พแลน-ไมเออร์
4. วิธีการของบัคเลย์และเจมส์

3.3.5 หาค่าคลาดเคลื่อนจากการประมาณค่าตัวแปรตาม และทำการเปรียบเทียบ

สำหรับรายละเอียดแต่ละขั้นตอนเป็นดังนี้

3.3.1 การจำลองตัวแปรอิสระ X_i ซึ่งเป็นค่าคงที่ จากการแจกแจงตามที่ต้องการศึกษามี 2 แบบ คือ

ก. การแจกแจงปกติ(Normal Distribution) N (μ,σ^2) เมื่อ $\mu = 34$ และ $\sigma^2 = 144$

ข. การแจกแจงไวบูลล์(Weibull Distribution) W(α',β') เมื่อ $\alpha' = 3.3$ และ $\beta' = 38$

และจำลองค่าคลาดเคลื่อน ε_i จากการแจกแจงตามที่ต้องการศึกษามี 3 แบบ คือ

ก. การแจกแจงปกติ N(μ,σ^2) เมื่อ $\mu = 0$ และ $\sigma^2 = 36$

ข. การแจกแจงลอกนอร์มอล LN(-5.5,7.3)

ค. การแจกแจงไวบูลล์ W(0.5,1)

3.3.2 กำหนดค่าสูงสุดของ T_i ที่จะถูกตัดทิ้ง ในการศึกษาครั้งนี้กำหนดค่าสูงสุดของ T_i ที่จะถูกตัดทิ้งไว้ที่ 3 ระดับ คำนวณได้ จาก $T_i = \alpha + \beta x_i + \varepsilon_i \quad ; i = 1,2,3,...,N$ โดยการหาค่าเฉลี่ยของ T_i [$E(T_i)$] จากสมการข้างต้น และค่า α ,β ,ค่าเฉลี่ยของตัวแปรอิสระ[$E(X_i)$] และค่าเฉลี่ยของค่าคลาดเคลื่อน[$E(\varepsilon_i)$] กำหนดตามรายละเอียดและการแจกแจงตามข้อ 3.3.1 และกำหนดค่าสูงสุดที่จะถูกตัดทิ้งระดับที่ 1, 2 และ 3 ที่ $\mu_T + 0.5\sigma_T$, $\mu_T + \sigma_T$ และ $\mu_T + 1.5\sigma_T$ เมื่อ μ_T คือค่าเฉลี่ยของ T_i และ σ_T คือส่วนเบี่ยงเบนมาตรฐานของ T_i ซึ่งจะมีค่าเท่ากับส่วนเบี่ยงเบนมาตรฐานของค่าคลาดเคลื่อน ดังนั้นค่าสูงสุดที่ถูกตัดทิ้งจะเท่ากับ 8.5, 11.5 และ 14.5 เมื่อค่าคลาดเคลื่อนแจกแจงปกติและลอกนอร์มอล และเท่ากับ 10, 12 และ 14 เมื่อค่าคลาดเคลื่อนแจกแจงไวบูลล์

3.3.3 หาค่าสังเกตของตัวแปรตาม Y_i ; i = 1,2,3, ... ,N เมื่อ N คือจำนวนตัวอย่างทั้งหมดที่ทำการศึกษา

$$Y_i = \begin{cases} T_i \; ; \; T_i \leq T_c \\ T_c \; ; \; T_i > T_c \end{cases}$$

$$\delta_i = \begin{cases} 1 \; ; \; T_i \leq T_c \\ 0 \; ; \; T_i > T_c \end{cases}$$

$\delta_i = 1$ หมายถึง ค่าสังเกตที่ไม่ถูกตัดทิ้ง

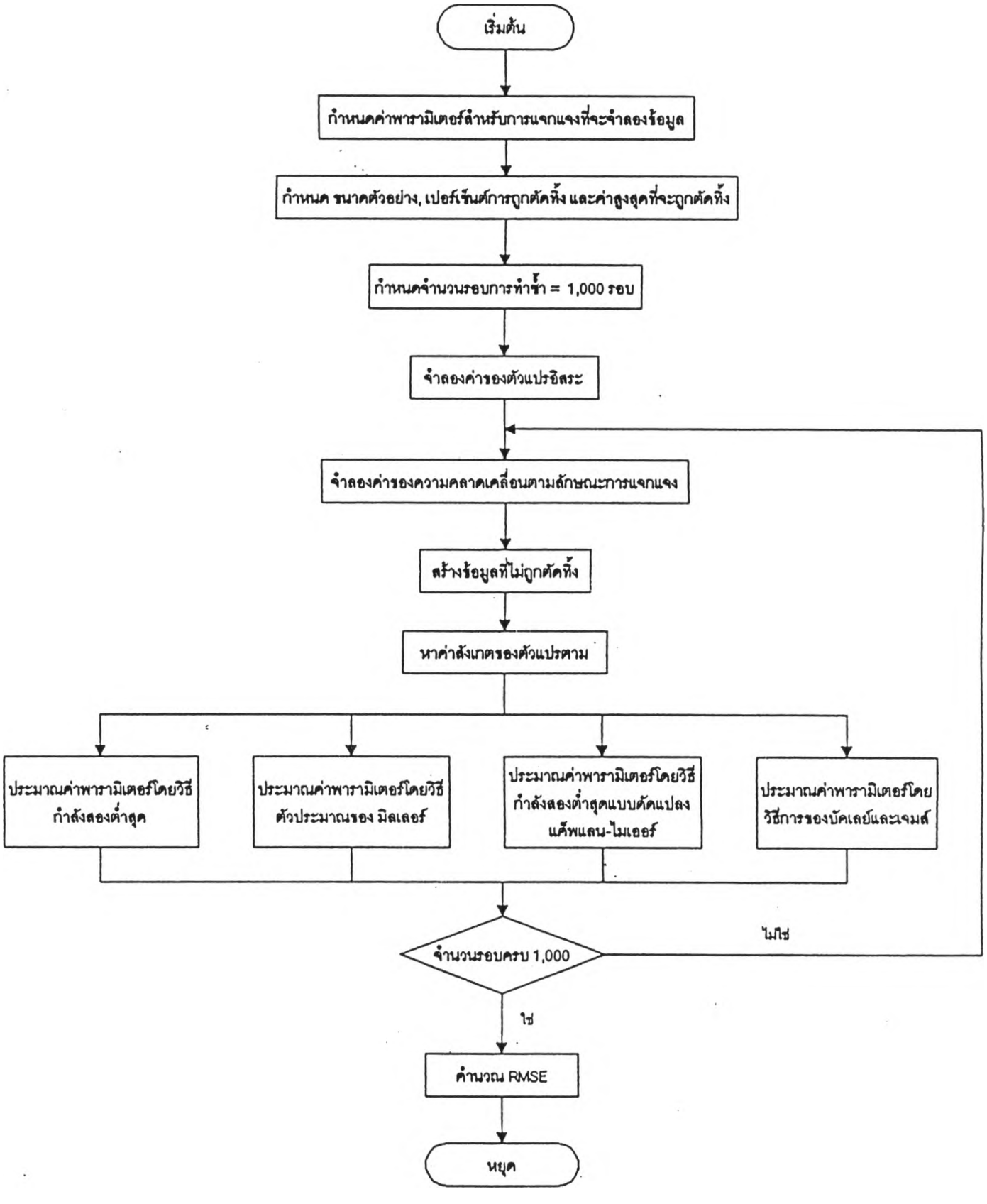
$\delta_i = 0$ หมายถึง ค่าสังเกตที่ถูกตัดทิ้ง

T_c = ค่าสูงสุดของข้อมูลที่ถูกตัดทิ้ง

การหาค่าสังเกตของตัวแปรตาม Y_i ทำได้โดยจำลอง T_i แล้วนำค่า T_i มาเปรียบเทียบกับค่า T_c จะได้ค่า $Y_i = T_i$ เมื่อ $T_i \leq T_c$ และให้ $\delta_i = 1$ แล้วนับเป็นค่าสังเกตที่ไม่ถูกตัดทิ้ง และจะได้ $Y_i = T_c$ เมื่อ $T_i > T_c$ และให้ $\delta_i = 0$ แล้วนับเป็นค่าสังเกตที่ถูกตัดทิ้งและทำการทดลองเช่นนี้จนกระทั่งได้ค่าสังเกตที่ไม่ถูกตัดทิ้งและที่ถูกตัดทิ้งครบตามจำนวนที่กำหนด

3.3.4 นำค่า Y_i และ X_i; $i = 1,2,3,\_,N$ คำนวณค่าประมาณพารามิเตอร์จากสมการถดถอยเชิงเส้นอย่างง่าย 4 วิธี คือ วิธีกำลังสองต่ำสุด วิธีตัวประมาณของมิลเลอร์ วิธีกำลังสองต่ำสุดแบบดัดแปลงแค็พแลน-ไมเออร์ และวิธีการของบัคเลย์และเจมส์ และ 3 วิธีหลังจะใช้การกระทำซ้ำจนกระทั่งค่าพารามิเตอร์คงตัว อาจจะมีบางชุดของข้อมูล เช่นข้อมูลที่มีเปอร์เซ็นต์การถูกตัดทิ้งของข้อมูลมาก ค่าประมาณพารามิเตอร์จะแกว่งอยู่ระหว่าง 2 ค่า กรณีนี้ให้ใช้ค่าเฉลี่ยระหว่าง 2 ค่านั้นเป็นค่าประมาณพารามิเตอร์ หรือค่าประมาณพารามิเตอร์ของรอบปัจจุบันกับรอบที่ผ่านมาต่างกันไม่เกิน 0.001 ถือว่าค่าประมาณพารามิเตอร์นั้นหาค่าได้ ถ้าไม่เข้าข่ายทั้ง 2 กรณีนี้ก็จะทิ้งชุดข้อมูลนั้น และทำการจำลองข้อมูลที่หาค่าประมาณพารามิเตอร์ได้ทั้ง 4 วิธี จำนวน 1000 ชุดข้อมูล ในแต่ละสถานการณ์

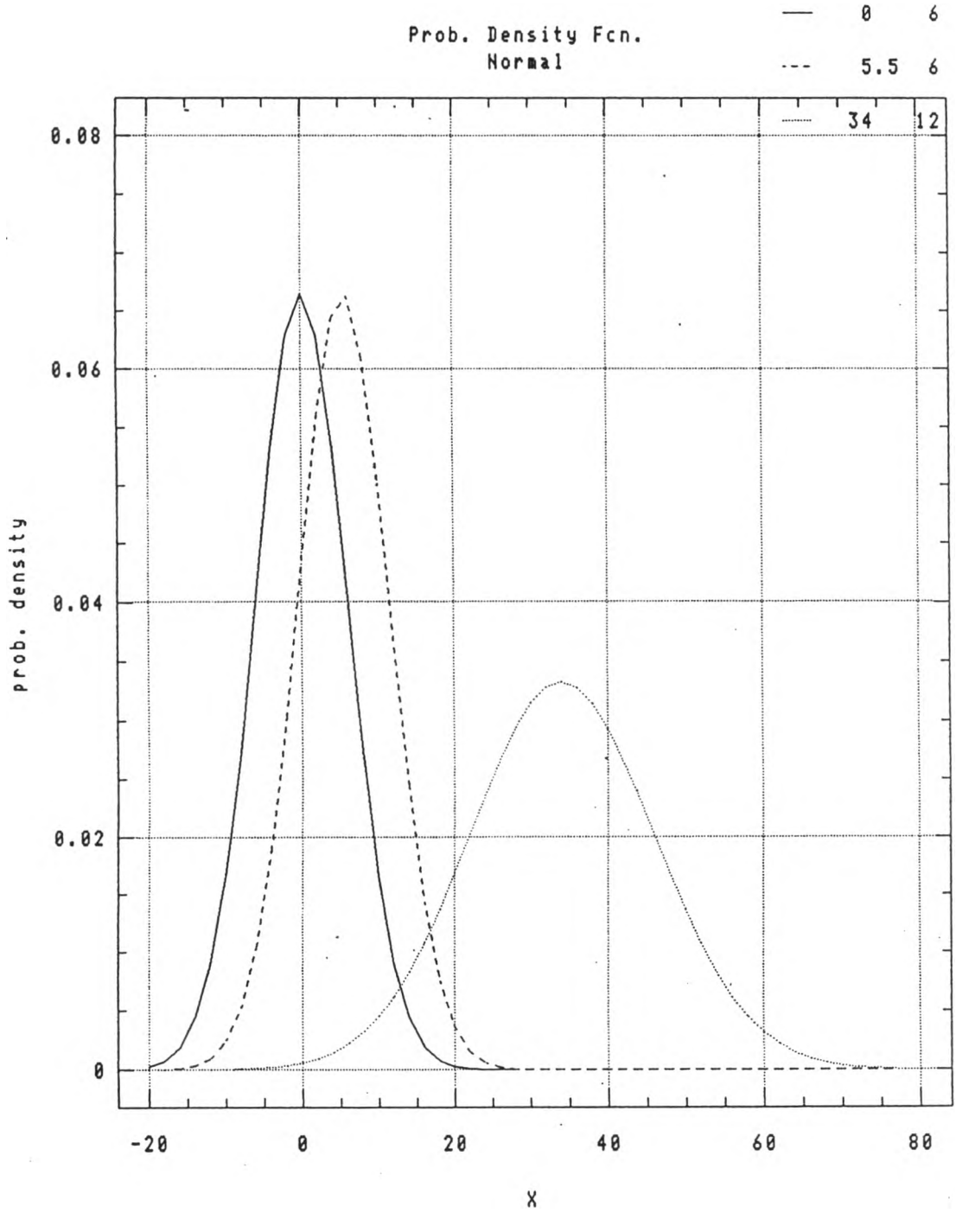
สรุปขั้นตอนการศึกษาเปรียบเทียบเป็นผังลำดับงานได้ดังรูปที่ 3.1



รูปที่ 3.1 แสดงผังงานสำหรับหาค่าความคลาดเคลื่อนจากการประมาณทั้ง 4 วิธี

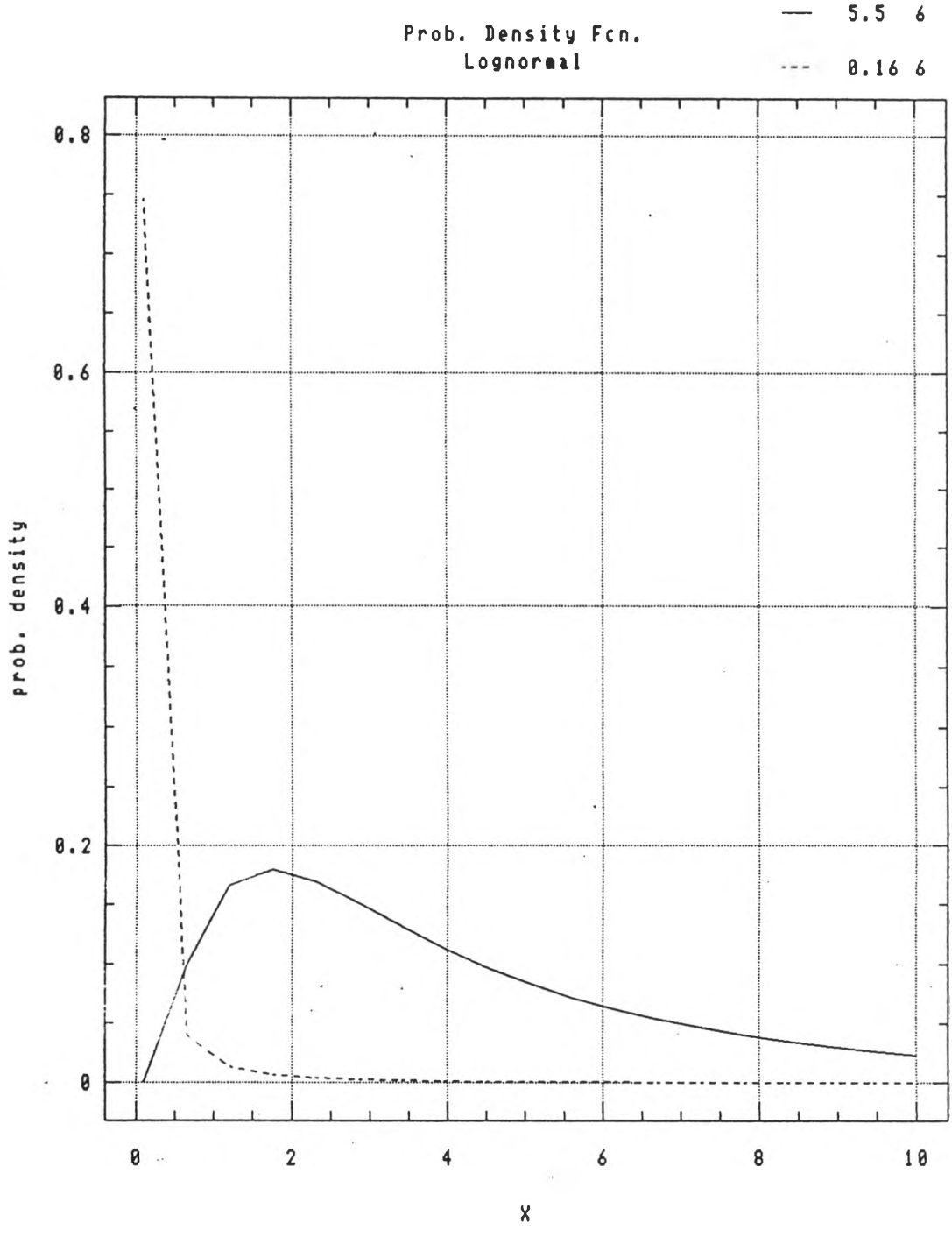
กราฟแสดงการแจกแจงรูปแบบต่างๆ ที่ใช้ในการวิจัยครั้งนี้

1. รูปที่ 3.2 แสดงการแจกแจงปกติของตัวแปรอิสระ X ที่มี μ เท่ากับ 34 และ σ^2 เท่ากับ 144 ,ค่าคลาดเคลื่อนที่มี μ เท่ากับศูนย์ และ $\sigma^2 = 36$ และ ตัวแปรตาม T ที่มี μ เท่ากับ 5.5 และ $\sigma^2 = 36$



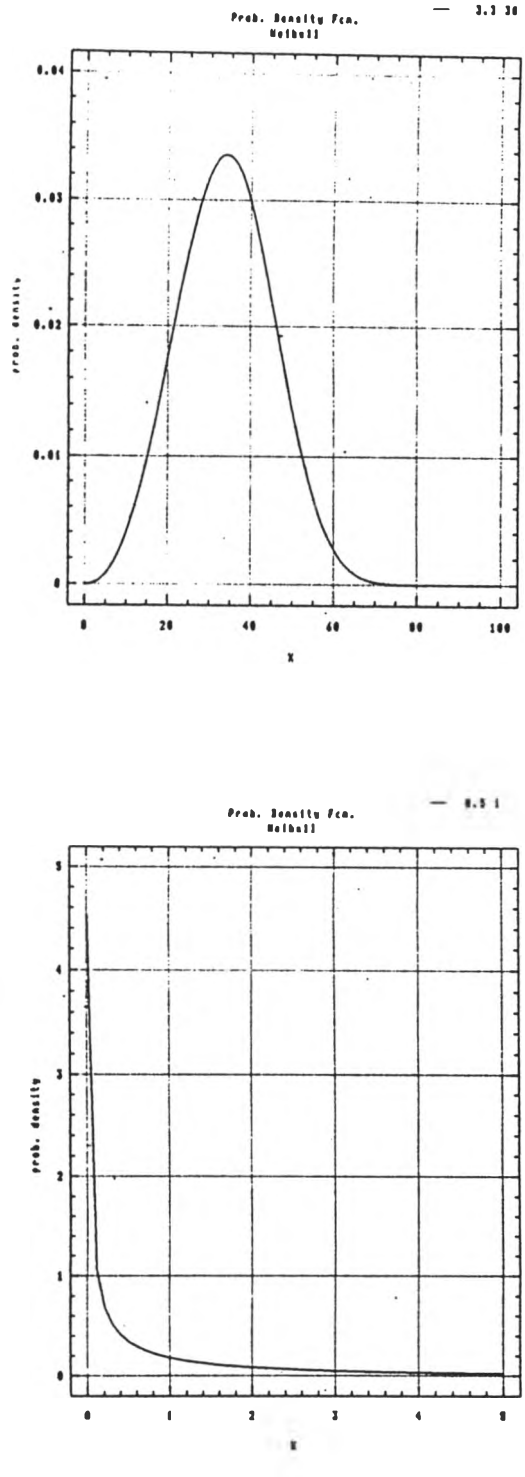
รูปที่ 3.2 แสดงการแจกแจงปกติ

2. รูปที่ 3.3 แสดงการแจกแจงลอกนอร์มอลของตัวแปรตาม T ที่มีค่าเฉลี่ย เท่ากับ 5.5 และ ความแปรปรวน เท่ากับ 36 (μ = 1.31 , σ^2 = 0.78) และค่าคลาดเคลื่อนที่มีค่าเฉลี่ย เท่ากับ 0.16 และ ความแปรปรวน เท่ากับ 36 (μ = -5.5 , σ^2 = 7.3)



รูปที่ 3.3 แสดงการแจกแจงลอกนอร์มอล

3. รูปที่ 3.4 แสดงการแจกแจงไวบูลล์ของตัวแปรอิสระ X [W(3.3,38)] และค่าคลาดเคลื่อนที่มีค่าเฉลี่ยเท่ากับ 2 และความแปรปรวนเท่ากับ 20 [W(0.5,1)]



รูปที่ 3.4 แสดงการแจกแจงไวบูลล์

การประมาณพารามิเตอร์แต่ละวิธีมีรายละเอียดดังนี้

3.3.4.1 **วิธีกำลังสองต่ำสุด**

ประมาณค่าพารามิเตอร์ $\hat{\beta}$ โดยการหาอนุพันธ์ของผลบวกกำลังสองของความคลาดเคลื่อนเทียบกับ $\hat{\beta}$ แล้วกำหนดสมการให้เท่ากับ $\underset{\sim}{0}$

$$\frac{\partial}{\partial \hat{\beta}} \quad (\underset{\sim}{Y}'\underset{\sim}{Y} - \quad 2\hat{\beta}'\underset{\sim}{X}'\underset{\sim}{Y} + \hat{\beta}'\underset{\sim}{X}'\underset{\sim}{X}\hat{\beta}) \quad = \quad \underset{\sim}{0}$$

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

3.3.4.2 **วิธีตัวประมาณของมิลเลอร์ มีขั้นตอนเป็นดังนี้**

1. **เฉพาะข้อมูลที่ไม่ถูกตัดทิ้ง ประมาณค่าพารามิเตอร์** $\hat{\beta}$ **เริ่มต้นด้วยวิธีกำลังสองต่ำสุด**

$$\hat{\beta} = \frac{\sum_u y_i (x_i - \bar{x}^u)}{\sum_u (x_i - \bar{x}^u)^2}$$

เมื่อ $\sum_u$ คือผลรวมของค่าสังเกตเฉพาะค่าที่ไม่ถูกตัดทิ้ง

$\bar{x}^u$ คือค่าเฉลี่ยของ x_i เฉพาะข้อมูลที่ไม่ถูกตัดทิ้ง

2. หาค่าคลาดเคลื่อนบางส่วน (Partial Residuals) โดยคำนวณที่ $\hat{\alpha} = 0$, $e_i = y_i - \hat{\beta} x_i$; $i = 1,2,3,...,N$ และนำ e_i มาเรียงลำดับจากน้อยไปหามาก จะได้ $e(1) < e(2) < ... < e(N)$ ถ้าค่าคลาดเคลื่อนที่มากที่สุดเป็นข้อมูลที่ถูกตัดทิ้งให้ทำการเปลี่ยนข้อมูลค่าคลาดเคลื่อนดังกล่าวเป็นข้อมูลที่ไม่ถูกตัดทิ้ง แล้วคำนวณค่าประมาณ $\hat{F}(e_i)$ โดยใช้ตัวประมาณพีแอล

$$\hat{F}(e_i) = 1 - \hat{S}(e_i)$$

$$\hat{S}(e_i) = \prod_{i\,;\,e(i) \le e_i}^{N} \left[\frac{(N-i)}{(N-i+1)}\right]^{\delta_i}$$

เมื่อ i คือ ลำดับที่ของความคลาดเคลื่อน

N คือ จำนวนข้อมูลทั้งหมด

3. นำค่า $\hat{F}(e_i)$ มาหาค่าถ่วงน้ำหนัก $w_i(\hat{\beta})$

4. ประมาณค่าพารามิเตอร์ตามวิธีตัวประมาณของมิลเลอร์ ได้ดังนี้

$$\hat{\beta} = \frac{\sum_u w_i(\hat{\beta}) y_i (x_i - \bar{x}^w)}{\sum_u w_i(\hat{\beta}) (x_i - \bar{x}^w)^2}$$

$$\hat{\alpha} = \sum_u w_i(\hat{\beta}) (y_i - \hat{\beta} x_i)$$

เมื่อ $\bar{x}^w = \sum_u w_i(\hat{\beta}) x_i$

วิธีตัวประมาณของมิลเลอร์จะกระทำซ้ำจนกระทั่งได้ค่าประมาณพารามิเตอร์ $\hat{\beta}$ รอบปัจจุบันเท่ากับรอบที่ผ่านมา หรือถ้าค่าพารามิเตอร์แกว่งอยู่ระหว่าง 2 ค่า จะใช้ค่าเฉลี่ยของ 2 ค่านั้นเป็นค่าประมาณพารามิเตอร์ $\hat{\beta}$ หรือค่าประมาณพารามิเตอร์ในรอบที่ผ่านมากับรอบปัจจุบันมีผลต่างกันไม่เกิน 0.001 ก็จะใช้ค่าในรอบปัจจุบันเป็นค่าประมาณพารามิเตอร์ $\hat{\beta}$ และหลังจากนั้นจึงทำการประมาณค่าพารามิเตอร์ $\hat{\alpha}$

3.3.4.3 วิธีกำลังสองต่ำสุดแบบตัดแปลงแค็พแลน-ไมเออร์ มีขั้นตอนดังนี้

1. เฉพาะข้อมูลที่ไม่ถูกตัดทิ้ง ประมาณค่าพารามิเตอร์ $\hat{\beta}$ เริ่มต้นด้วยวิธีกำลังสองต่ำสุด

$$\hat{\beta} = \frac{\sum_u y_i (x_i - \bar{x}^u)}{\sum_u (x_i - \bar{x}^u)^2}$$

เมื่อ $\sum_u$ คือผลรวมของค่าสังเกตเฉพาะค่าที่ไม่ถูกตัดทิ้ง

$\bar{x}^u$ คือค่าเฉลี่ยของ x_i เฉพาะข้อมูลที่ไม่ถูกตัดทิ้ง

2. หาค่าคลาดเคลื่อนบางส่วน (Partial Residuals) โดยคำนวณที่ $\hat{\alpha} = 0$, $e_i = y_i - \hat{\beta} x_i$; $i = 1,2,3,...,N$ และนำ e_i มาเรียงลำดับจากน้อยไปหามาก จะได้ $e(1) < e(2) < ... < e(N)$ แล้วคำนวณค่าประมาณ $\hat{F}(e_i)$ โดยใช้ตัวประมาณพีแอล

$$\hat{F}(e_i) = 1 - \hat{S}(e_i)$$

$$\hat{S}(e_i) = \prod_{i\,;\,e(i) \le e_i} \left[\frac{(N-i)}{(N-i+1)} \right]^{\delta_i}$$

เมื่อ i คือ ลำดับที่ของความคลาดเคลื่อน

N คือ จำนวนข้อมูลทั้งหมด

3. นำค่า $\hat{F}(e_i)$ มาหาค่าถ่วงน้ำหนัก $w_i(\hat{\beta})$ ในกรณีที่ค่าคลาดเคลื่อนที่มากที่สุดเป็นข้อมูลที่ถูกตัดทิ้งให้ทำการปรับค่าถ่วงน้ำหนักเป็น $w_i'(\hat{\beta})$

$$w_i'(\hat{\beta}) = \frac{w_i(\hat{\beta})}{\sum_u w_j(\hat{\beta})}$$

4. ประมาณค่าพารามิเตอร์ตามวิธีกำลังสองต่ำสุดแบบดัดแปลงแค็พแลน-ไมเออร์ ได้ดังนี้

$$\hat{\beta} = \frac{\sum_u w_i'(\hat{\beta})\, y_i (x_i - \bar{x}^k)}{\sum_u w_i'(\hat{\beta})\, (x_i - \bar{x}^k)^2}$$

$$\hat{\alpha} = \sum_u w'_i(\hat{\beta})(y_i - \hat{\beta} x_i)$$

เมื่อ $\sum_u$ คือผลรวมของค่าสังเกตเฉพาะค่าที่ไม่ถูกตัดทิ้ง

$\bar{x}^k$ คือค่าเฉลี่ยแบบถ่วงน้ำหนักของค่าสังเกต x ซึ่งเท่ากับ $\sum_u w_i'(\hat{\beta}) x_i$

วิธีกำลังสองต่ำสุดแบบดัดแปลงแค็พแลน-ไมเออร์ จะกระทำซ้ำจนกระทั่งได้ค่าประมาณพารามิเตอร์ $\hat{\beta}$ รอบปัจจุบันเท่ากับรอบที่ผ่านมา หรือถ้าค่าพารามิเตอร์แกว่งอยู่ระหว่าง 2 ค่า จะใช้ค่าเฉลี่ยของ 2 ค่านั้นเป็นค่าประมาณพารามิเตอร์ $\hat{\beta}$ หรือค่าประมาณพารามิเตอร์ในรอบที่ผ่านมากับรอบปัจจุบันมีผลต่างกันไม่เกิน 0.001 ก็จะใช้ค่าในรอบปัจจุบันเป็นค่าประมาณพารามิเตอร์ $\hat{\beta}$ และหลังจากนั้นจึงทำการประมาณค่าพารามิเตอร์ $\hat{\alpha}$

3.3.4.4 วิธีการของบัคเลย์และเจมส์ มีขั้นตอนดังนี้

1. เฉพาะข้อมูลที่ไม่ถูกตัดทิ้ง ประมาณค่าพารามิเตอร์ $\hat{\beta}$ เริ่มต้นด้วยวิธีกำลังสองต่ำสุด

$$\hat{\beta} = \frac{\sum_u y_i (x_i - \bar{x}^u)}{\sum_u (x_i - \bar{x}^u)^2}$$

เมื่อ $\sum_u$ คือผลรวมของค่าสังเกตเฉพาะค่าที่ไม่ถูกตัดทิ้ง

$\bar{x}^u$ คือค่าเฉลี่ยของ x_i เฉพาะข้อมูลที่ไม่ถูกตัดทิ้ง

2. หาค่าคลาดเคลื่อนบางส่วน(Partial Residuals) โดยคำนวณที่

$\hat{\alpha} = 0$, $e_i = y_i - \hat{\beta} x_i \quad ; i = 1,2,3,...,N$ และนำ e_i มาเรียงลำดับจากน้อยไปหามาก จะได้ $e(1) < e(2) < ... < e(N)$ แล้วคำนวณค่าประมาณ $\hat{F}(e_i)$ โดยใช้ตัวประมาณพีแอล

$$\hat{F}(e_i) = 1 - \hat{S}(e_i)$$

$$\hat{S}(e_i) = \prod_{i\,;\,e(i) \le e_i}^{N} \left[\frac{(N-i)}{(N-i+1)} \right]^{\delta_i}$$

เมื่อ i คือ ลำดับที่ของความคลาดเคลื่อน

N คือ จำนวนข้อมูลทั้งหมด

3. นำค่า $\hat{F}(e_i)$ มาหาค่าถ่วงน้ำหนัก $w_i(\hat{\beta})$ ในกรณีที่ค่าคลาดเคลื่อนที่มากที่สุดเป็นข้อมูลที่ถูกตัดทิ้งให้ทำการปรับค่าถ่วงน้ำหนักเป็น $w_i'(\hat{\beta})$

$$w_i'(\hat{\beta}) = \frac{w_i(\hat{\beta})}{\sum_u w_j(\hat{\beta})}$$

4. ประมาณค่าสังเกตที่ถูกตัดทิ้งด้วยค่าคาดหวังอย่างมีเงื่อนไข $E[\, y_i / y_i > T_c, x_i, \hat{\beta}\,]$ ซึ่งจะใช้ $\bar{Y}_i(\hat{\beta})$ แทน

$$\bar{Y}_i(\hat{\beta}) = \hat{\beta} x_i + \frac{\sum_{k \in u} w_k'(\hat{\beta})(y_k - \hat{\beta} x_k)}{1 - \hat{F}(T_c - \hat{\beta} x_i)}$$

5. นำค่าประมาณของค่าสังเกตที่ถูกตัดทิ้งและค่าสังเกตที่ไม่ถูกตัดทิ้ง มาหาค่าประมาณพารามิเตอร์ ได้ดังนี้

$$\hat{\beta} = \frac{\sum_u y_i(x_i - \bar{x}) + \sum_C \bar{y}_i(\hat{\beta})(x_i - \bar{x})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

$$\hat{\alpha} = \left\{ \frac{\left[\sum_u y_i + \sum_c \bar{y}_i (\hat{\beta}) \right]}{N} \right\} - \hat{\beta}\, \bar{x}$$

เมื่อ $\sum_u$ คือ ผลรวมเฉพาะค่าสังเกตที่ไม่ถูกตัดทิ้ง

$\sum_c$ คือ ผลรวมเฉพาะค่าสังเกตที่ถูกตัดทิ้ง

วิธีการของบัคเลย์และเจมส์ จะกระทำซ้ำจนกระทั่งได้ค่าประมาณพารามิเตอร์ $\hat{\beta}$ รอบปัจจุบันเท่ากับรอบที่ผ่านมา หรือถ้าค่าพารามิเตอร์แกว่งอยู่ระหว่าง 2 ค่า จะใช้ค่าเฉลี่ยของ 2 ค่านั้นเป็นค่าประมาณพารามิเตอร์ $\hat{\beta}$ หรือค่าประมาณพารามิเตอร์ในรอบที่ผ่านมากับรอบปัจจุบันมีผลต่างกันไม่เกิน 0.001 ก็จะใช้ค่าในรอบปัจจุบันเป็นค่าประมาณพารามิเตอร์ $\hat{\beta}$ และหลังจากนั้นจึงทำการประมาณค่าพารามิเตอร์ $\hat{\alpha}$

3.3.5 หาค่าคลาดเคลื่อนจากการประมาณค่าตัวแปรตาม โดยการเปรียบเทียบกับค่าจริงก่อนการถูกตัดทิ้ง เพื่อนำมาคำนวณหาค่ารากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (RMSE) เนื่องจากการทดลองได้กระทำซ้ำๆ กัน 1000 รอบ ดังนั้นในแต่ละสถานการณ์ให้ j แทนรอบที่ทำซ้ำ j = 1,2,3,...,1000 คำนวณรากที่สองของค่าเฉลี่ยความคลาดเคลื่อนกำลังสองหาได้จาก

$$MSE_j = \frac{1}{N} \sum_{i=1}^{N} \left[T_i - \hat{Y}_i \right]^2$$

$$RMSE_j = \sqrt{MSE_j}$$

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

จากนั้นจึงนำ RMSE ของการประมาณตัวแปรตามเปรียบเทียบกันทั้ง 4 วิธี เพื่อหาว่าวิธีการใดให้ค่า RMSE ของการประมาณค่าตัวแปรตามต่ำที่สุด วิธีนั้นจะเป็นวิธีที่ประมาณค่าตัวแปรตามได้ดีที่สุดในแต่ละสถานการณ์

ในการหาค่า RMSE ของการประมาณทั้ง 4 วิธี ในขนาดตัวอย่างหนึ่งๆ จะเปลี่ยนเปอร์เซ็นต์การถูกตัดทิ้งเป็น 10%, 20% 30% และ 40% และเปลี่ยนการแจกแจงของตัวแปรอิสระเป็น 2 การแจกแจง คือการแจกแจงปกติและไวบูลล์ ส่วนการแจกแจงของค่าคลาดเคลื่อนมี 3 แบบ คือการแจกแจงปกติ ลอกนอร์มอล และไวบูลล์ โดยในแต่ละสถานการณ์ทำซ้ำๆ กัน 1000 รอบ