บทที่ 2

ทฤษฎีและสถิติที่เกี่ยวข้อง

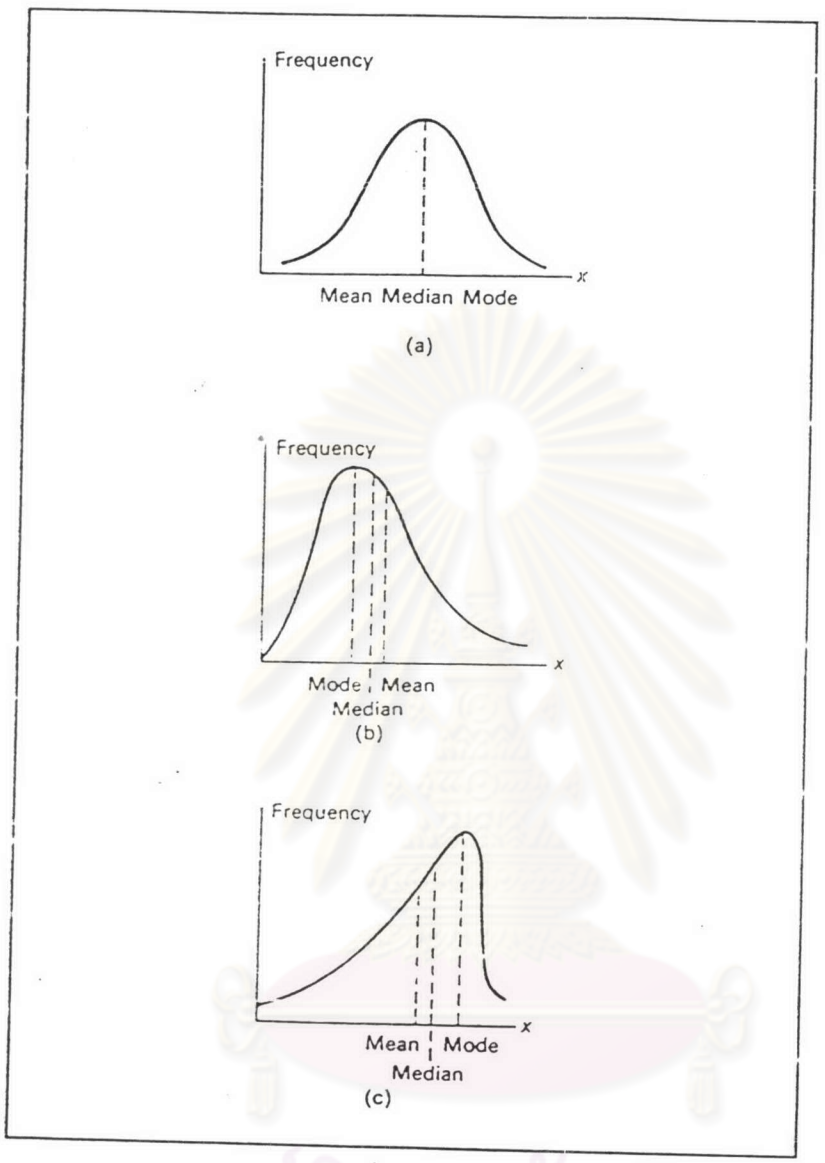
การประมาณค่าแบบช่วงที่เหมาะสมนั้นช่วงความเชื่อมั่นที่ประมาณได้จะต้องครอบคลุมค่าพารามิเตอร์ที่สนใจด้วยระดับความเชื่อมั่นที่กำหนดและช่วงความเชื่อมั่นที่ประมาณได้นั้นควรเป็นช่วงที่แคบ ดังนั้นในงานวิจัยนี้จึงมีจุดมุ่งหมายเพื่อเปรียบเทียบวิธีการประมาณค่าแบบช่วงสำหรับค่าสัมประสิทธิ์การถดถอยในสมการถดถอยเชิงเส้นตรงที่มีค่าคลาดเคลื่อนแจกแจงแบบเบ้ขวา โดยจะพิจารณาการแจกแจงของค่าคลาดเคลื่อน 4 การแจกแจง คือ การแจกแจงแลมดาของตูกีร์ การแจกแจงแกมมา การแจกแจงปกติ และการแจกแจงลอกนอร์มอล ภายใต้ระดับความเบ้ต่างๆ ส่วนเกณฑ์ในการพิจารณานั้นจะพิจารณาการผ่านระดับความเชื่อมั่นที่กำหนด และค่าเฉลี่ยความยาวของช่วงความเชื่อมั่น สำหรับในบทนี้จะกล่าวถึงทฤษฎีต่างๆที่เกี่ยวข้องกับการวิจัยซึ่งมีรายละเอียดดังนี้

2.1 ความเบ้และความโด่ง

2.1.1 ความเบ้ ประชากรที่มีการแจกแจงสมมาตรนั้น ตัวแปรสุ่ม X จะมีการแจกแจงสมมาตรรอบจุดใดจุดหนึ่ง ในที่นี้ขอกำหนดให้ชื่อจุด a กล่าวคือตัวแปรสุ่ม X จะสมมาตรรอบจุด a ถ้า

$$P(X \leq a - x) = P(X \geq a + x) \quad \text{สำหรับทุกค่า } x$$

แต่ถ้าประชากรมีการแจกแจงแบบไม่สมมาตร จะมีลักษณะของกราฟเบ้ไปข้างใดข้างหนึ่งดังตัวอย่างในรูปต่อไปนี้



ภาพประกอบที่ 2.1 แสดงตัวอย่างเส้นโค้งของการแจกแจงที่ไม่มีความเบ้ เบ้ขวา และเบ้ซ้าย

จะเห็นว่าในรูปที่ 2.1 (a) ประชากรมีการแจกแจงสมมาตร ซึ่งค่าเฉลี่ย ค่ามัธยฐาน และฐานนิยมจะมีค่าเท่ากัน ส่วนในรูปรูปที่ 2.1 (b) ประชากรมีการแจกแจงแบบเบ้ไปทางขวาเพราะพื้นที่ใต้เส้นโค้งทางด้านขวาของฐานนิยมมากกว่าพื้นที่ใต้เส้นโค้งทางด้านซ้าย และในรูปรูปที่ 2.1 (c) ประชากรมีการแจกแจงเบ้ไปทางซ้ายเพราะพื้นที่ใต้เส้นโค้งทางด้านซ้ายของฐานนิยมมากกว่าพื้นที่ใต้เส้นโค้งทางด้านขวา

การวัดความเบ้จะใช้การวัดโดยวิธีโมเมนต์ สูตรสำหรับหาค่าวัดสมมาตร หรือสัมประสิทธิ์ความเบ้คือ

$$\alpha_3 = \frac{E((X-\mu)^3)}{(V(X))^{3/2}}$$

การวัดความเบ้ด้วยโมเมนต์ศูนย์กลางอันดับที่ 3 จะให้ผลต่างๆดังนี้

1. ถ้าการแจกแจงสมมาตร ค่าสัมประสิทธิ์ความเบ้จะเท่ากับศูนย์
2. ถ้าการแจกแจงเบ้ขวา ค่าสัมประสิทธิ์ความเบ้จะเป็นบวก
3. ถ้าการแจกแจงเบ้ซ้าย ค่าสัมประสิทธิ์ความเบ้จะเป็นลบ

2.1.2 **ความโด่ง** ความโด่งของการแจกแจงมี 3 ลักษณะดังนี้

1. เส้นโค้งมีความโด่งเป็นปกติ (Mesokurtic)
2. เส้นโค้งที่แบนราบกว่าปกติ (Platykurtic)
3. เส้นโค้งที่โด่งกว่าปกติ (Leptokurtic)

การวัดความโด่ง หรือการหาค่าสัมประสิทธิ์ความโด่ง มีสูตรดังนี้

$$\alpha_4 = \frac{E((X-\mu)^4)}{(V(X))^2}$$

การวัดความโด่งด้วยโมเมนต์ศูนย์กลางอันดับที่ 4 จะให้ผลต่างๆดังนี้

1. ถ้าเส้นโค้งมีความโด่งเป็นปกติ ค่าสัมประสิทธิ์ความโด่งจะเท่ากับ 3
2. ถ้าเส้นโค้งแบนราบกว่าปกติ ค่าสัมประสิทธิ์ความโด่งจะน้อยกว่า 3
3. ถ้าเส้นโค้งโด่งกว่าปกติ ค่าสัมประสิทธิ์ความโด่งจะมากกว่า 3

2.2 การแจกแจงแลมดาของตูกีร์

Ramberg และ Sohmeiser ได้เสนอวิธีการสร้างตัวแปรสุ่มที่ขึ้นอยู่กับความเบ้และความโด่ง โดยตัวแปรสุ่มนี้มีการแจกแจงที่เรียกว่า"การแจกแจงแลมดาของตูกีร์" โดยตัวแปรสุ่มนี้จะถูกกำหนดจากค่าพารามิเตอร์ 4 ตัว ซึ่งสัมพันธ์กับค่าความเบ้และความโด่งดังนี้

$$X = R(p) = f^{-1}(X) = \lambda_1 + \frac{p^{\lambda_3} - (1-p)^{\lambda_4}}{\lambda_2} \quad , 0 \le p \le 1 \ , \ -\infty < X < \infty$$

เมื่อ p เป็นเลขสุ่มที่มีค่าระหว่าง 0 และ 1

X เป็นตัวแปรสุ่มที่มีการแจกแจงแบบแลมดาของตูกีร์

λ_1 เป็นพารามิเตอร์ที่กำหนดตำแหน่ง (Location Parameter)

λ_2 เป็นพารามิเตอร์มาตราส่วน (Scale Parameter)

λ_3 และ λ_4 เป็นพารามิเตอร์สัณฐาน (Shape Parameter) และเป็นฟังก์ชันของ ค่าเฉลี่ย ค่าเบี่ยงเบนมาตรฐาน ค่าความโด่ง และค่าความเบ้ ถ้าการแจกแจงเป็นแบบสมมาตรจะได้ว่า $\lambda_3 = \lambda_4$

ดังนั้นฟังก์ชันความหนาแน่นของตัวแปรสุ่ม X คือ

$$f(X) = f(R(p))$$

$$= 1/R(p) \quad , \quad R(p) = \frac{dR(p)}{dp}$$

$$= \lambda_2[\lambda_3 p^{\lambda_3 - 1} + \lambda_4(1-p)^{\lambda_4 - 1}]^{-1}$$

และโมเมนต์ที่ k เมื่อ $\lambda_1 = 0$ คือ

$$E(X)^k = \lambda_2^{-k} \sum_{i=0}^{k} \binom{k}{i} (-1)^i \beta(\lambda_3(k-i)+1, \lambda_4 i + 1)$$

จากสมการ โมเมนต์ที่ k จะหาค่าไม่ได้เมื่อค่าเบต้าฟังก์ชัน(β) มีค่าเป็นลบ ดังนั้นโมเมนต์ที่ k จะหาค่าได้ก็ต่อเมื่อ $\frac{-1}{k} < \min(\lambda_3, \lambda_4)$

จากสมการโมเมนต์ที่ k ทำให้สามารถหาค่าต่างๆต่อไปนี้ได้

ค่าเฉลี่ย $E(X) = \lambda_1 + \frac{A}{\lambda_2}$

ค่าความแปรปรวน $Var(X) = \frac{(B - A^2)}{\lambda_2^2}$

สัมประสิทธิ์ความเบ้ $\alpha_3 = \frac{(C - 3AB + 2A^3)}{(\lambda_2 \sigma)^3}$

สัมประสิทธิ์ความโด่ง $\alpha_4 = \frac{(D - 4AC + 6A^2B - 3A^4)}{(\lambda_2 \sigma)^4}$

กำหนด

$$A = \frac{1}{1+\lambda_3} - \frac{1}{1+\lambda_4}$$

$$B = \frac{1}{1+2\lambda_3} + \frac{1}{1+2\lambda_4} - 2\beta(1+\lambda_3, 1+\lambda_4)$$

$$C = \frac{1}{1+3\lambda_3} - \frac{1}{1+3\lambda_4} - 3\beta(1+2\lambda_3, 1+\lambda_4) + 3\beta(1+\lambda_3, 1+2\lambda_4)$$

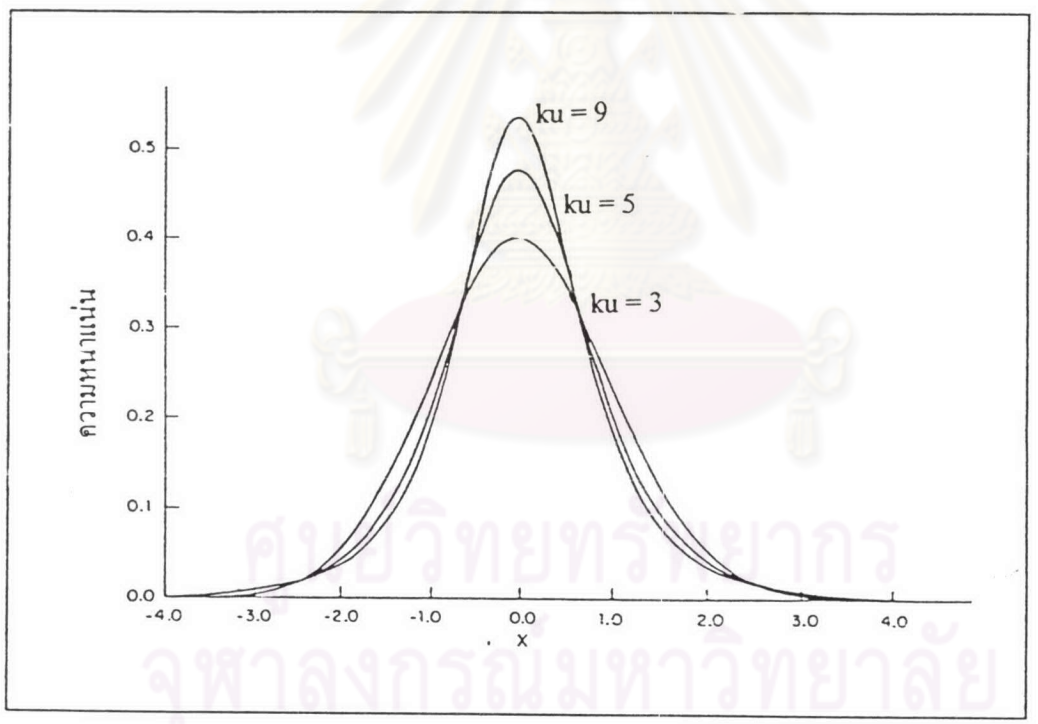
$$D = \frac{1}{1+4\lambda_3} + \frac{1}{1+4\lambda_4} - 4\beta(1+3\lambda_3, 1+\lambda_4) + 6\beta(1+2\lambda_3, 1+2\lambda_4) - 4\beta(1+\lambda_3, 1+3\lambda_4)$$

เราสามารถหาค่า λ_1 , λ_2 , λ_3 , λ_4 เมื่อกำหนดค่าความเบ้และความโด่งในระดับต่างๆได้จากตาราง Ramberg ที่แสดงในภาคผนวก โดยค่าพารามิเตอร์ที่ได้เป็นกรณีที่ค่าเฉลี่ยเป็นศูนย์และความแปรปรวนเท่ากับ 1 แต่ถ้าต้องการค่าพารามิเตอร์ในกรณีที่ค่าเฉลี่ยเท่ากับ μ และค่าความแปรปรวนเท่ากับ σ^2 จะต้องแปลงค่า λ_1 , λ_2 ดังนี้

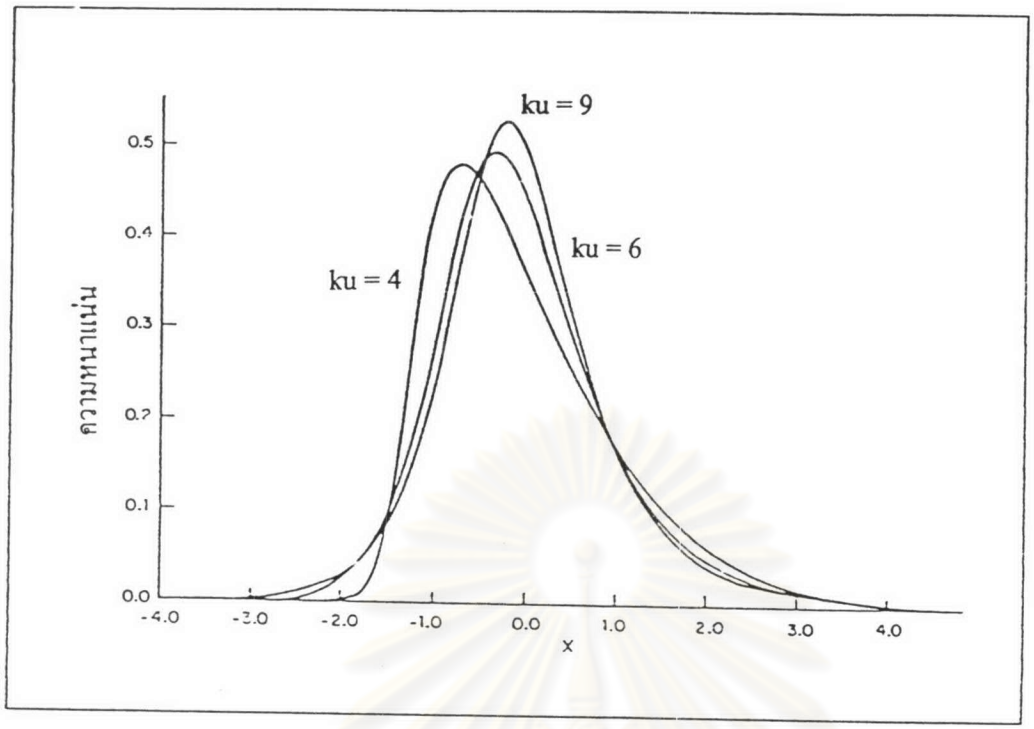
$$\lambda_1(\mu,\sigma) = \lambda_1(0,1)\sigma + \mu$$

$$\lambda_2(\mu,\sigma) = \frac{\lambda_2(0,1)}{\sigma}$$

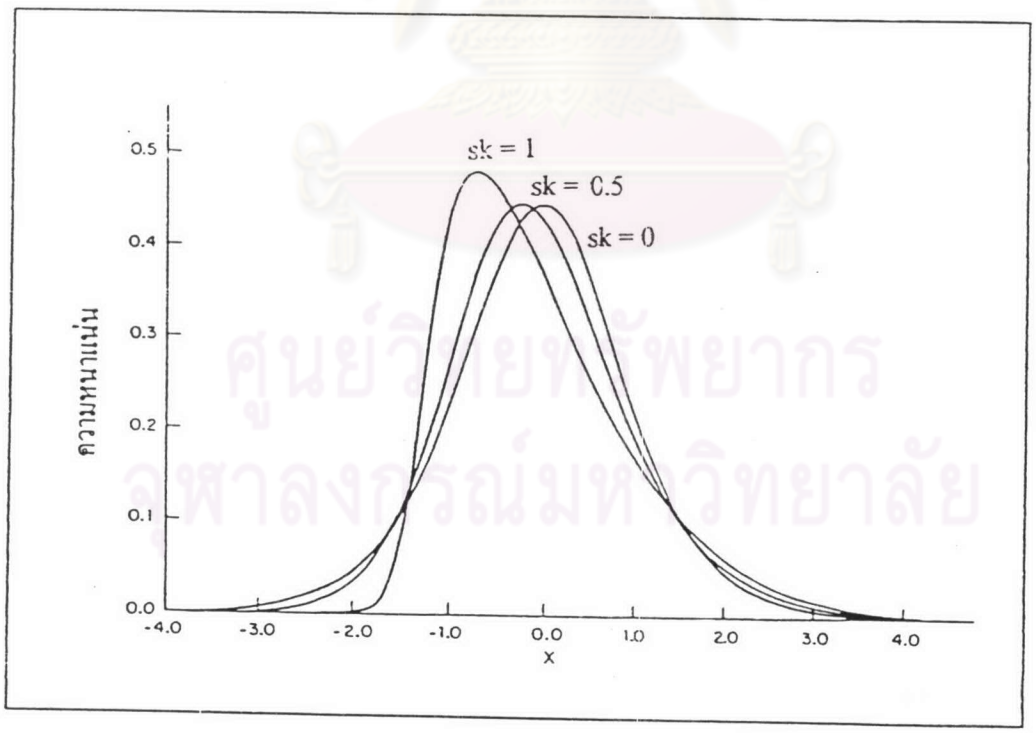
ฟังก์ชันความหนาแน่นของการแจกแจงแลมดาของตูกีร์จะมีลักษณะสมมาตรคล้ายการแจกแจงปกติ เมื่อพารามิเตอร์ λ_3 และ λ_4 เท่ากับ 0 และ 3 ตามลำดับ ซึ่งกราฟของฟังก์ชันความหนาแน่นของการแจกแจงจะขึ้นอยู่กับค่าพารามิเตอร์ดังที่แสดงในรูปต่อไปนี้



ภาพประกอบที่ 2.2 กราฟแสดงฟังก์ชันความหนาแน่นของการแจกแจงแลมดาของตูกีร์ ที่ความเบ้เท่ากับ 0 ความโด่งเท่ากับ 3 , 5 , 9



ภาพประกอบที่ 2.3 กราฟแสดงฟังก์ชันความหนาแน่นของการแจกแจงแลมดาของตูกีร์ ที่ความเบ้เท่ากับ 1 ความโด่งเท่ากับ 4 , 6 , 9



ภาพประกอบที่ 2.4 กราฟแสดงฟังก์ชันความหนาแน่นของการแจกแจงแลมดาของตูกีร์ ที่ความเบ้เท่ากับ 0 , 0.5 , 1 ความโด่งเท่ากับ 4

2.3 การแจกแจงแกมมา

ให้ X เป็นตัวแปรสุ่มต่อเนื่องที่มีการแจกแจงแกมมา ด้วยพารามิเตอร์ λ และ γ ดังนั้นฟังก์ชันความหนาแน่นจะอยู่ในรูปของ

$$f(x) = \frac{x^{\gamma-1}\exp(\frac{-x}{\lambda})}{\lambda^{\gamma}\Gamma(\gamma)} \quad , x \geq 0 \ , \ \gamma > 0 \ , \ \lambda > 0$$

$$= \quad 0 \qquad \text{อื่นๆ}$$

เมื่อ λ เป็นพารามิเตอร์เสกล

γ เป็นพารามิเตอร์รูปทร

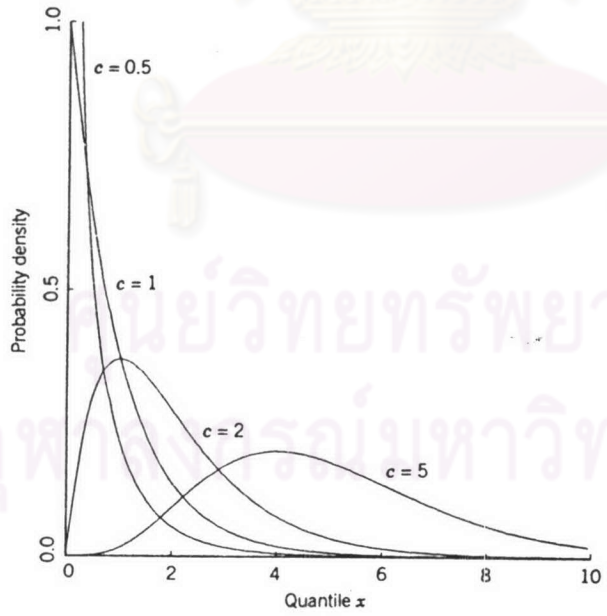
$\Gamma(\gamma)$ เป็นฟังก์ชันแกมมาของ γ

ค่าเฉลี่ยคือ $E(X) = \gamma\lambda$

ค่าความแปรปรวนคือ $Var(X) = \gamma\lambda^2$

สัมประสิทธิ์ความเบ้ $\alpha_3 = 2\gamma^{-1/2}$

สัมประสิทธิ์ความโด่ง $\alpha_4 = 3 + 6/\gamma$



ภาพประกอบที่ 2.5 กราฟแสดงการแจกแจงแกมมาที่ λ =1

γ = 0.5 , 1 , 2 , 5

2.4 การแจกแจงลอกนอร์มอล

ให้ X เป็นตัวแปรสุ่มต่อเนื่องที่มีการแจกแจงลอกนอร์มอล ด้วยพารามิเตอร์ μ และ σ^2 ดังนั้นฟังก์ชันความหนาแน่นจะอยู่ในรูปของ

$$f(x) = \frac{\exp\left(\frac{-(\ln x - \mu)^2}{2\sigma^2}\right)}{x\sqrt{2\pi\sigma^2}} = \frac{\exp\left(\frac{-\left(\ln\left(\frac{x}{M}\right)\right)^2}{2\sigma^2}\right)}{x\sqrt{2\pi\sigma^2}} \quad , x \geq 0 \, , \, \sigma > 0 \, , \, -\infty < \mu < \infty \, , \, M > 0$$

$$= 0 \qquad \text{อื่นๆ}$$

เมื่อ μ เป็นค่าเฉลี่ยของ ln x

σ^2 เป็นค่าความแปรปรวนของ ln x

M คือค่ามัธยฐานและ $M = e^{\mu}$

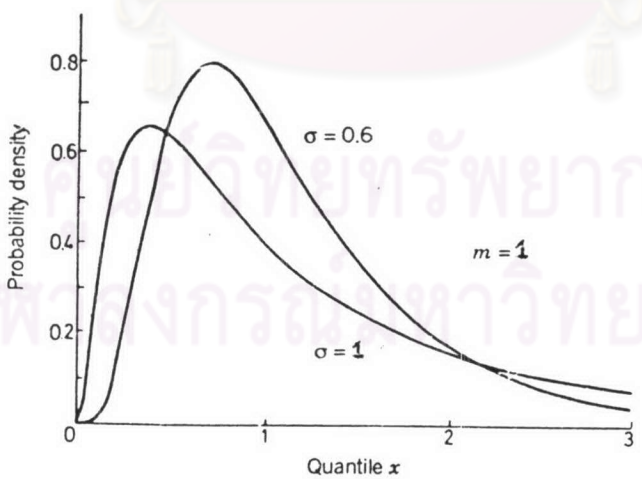
ค่าเฉลี่ย $E(X) = e^{(\mu+\frac{\sigma^2}{2})} = Me^{\frac{\sigma^2}{2}}$

ค่าความแปรปรวน $Var(X) = e^{(2\mu+\sigma^2)}(e^{\sigma^2}-1) = M^2w(w-1)$

สัมประสิทธิ์ความเบ้ $\alpha_3 = (w+2)(w-1)^{1/2}$

สัมประสิทธิ์ความโด่ง $\alpha_4 = w^4 + 2w^3 + 3w^2 - 3$

กำหนด $w = \exp(\sigma^2)$



ภาพประกอบที่ 2.6 กราฟแสดงการแจกแจงลอกนอร์มอลที่ M = 1 σ = 0.6 , 1

2.5 การแจกแจงปกติ

ให้ X เป็นตัวแปรสุ่มต่อเนื่องที่มีการแจกแจงปกติด้วยพารามิเตอร์ μ และ σ^2 ดังนั้นฟังก์ชันความหนาแน่นจะอยู่ในรูปของ

$$f(x) = \frac{\exp(\frac{-(x-\mu)^2}{2\sigma^2})}{\sqrt{2\pi\sigma^2}} \quad , -\infty < x < \infty \ , \ \sigma > 0 \ , \ -\infty < \mu < \infty$$

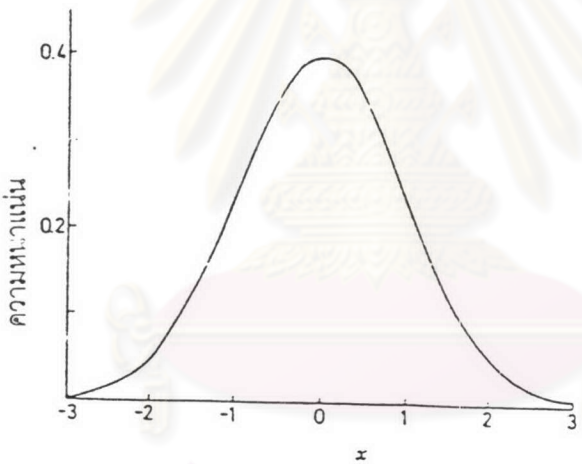
ค่าเฉลี่ย $E(X) = \mu$

ความแปรปรวน $Var(X) = \sigma^2$

สัมประสิทธิ์ความเบ้ $\alpha_3 = 0$

สัมประสิทธิ์ความโด่ง $\alpha_4 = 3$

ตัวอย่างกราฟแสดงฟังก์ชันความหนาแน่นของการแจกแจงปกติแสดงดังรูปที่ 2.7



ภาพประกอบที่ 2.7 กราฟแสดงการแจกแจงปกติมาตรฐาน N(0,1)

2.6 วิธีกำลังสองน้อยที่สุดแบบสามัญ(OLS)

จากสมการถดถอยเชิงเส้นเชิงเดียว $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ ซึ่งมีข้อสมมติของรูปแบบได้แก่ ε_i มีการแจกแจงที่เป็นอิสระกันแบบปกติที่ต่างก็มีค่าเฉลี่ย 0 และค่าความแปรปรวน σ^2 ซึ่งจะอธิบายได้ด้วย $\varepsilon_i \sim Nid(0, \sigma^2)$ จากข้อสมมติของ ε_i ดังกล่าวจะได้ว่า Y_i มีการแจกแจงที่เป็นอิสระกัน และที่ X_i ตัวแปร Y_i จะมีการแจกแจงแบบปกติที่มีค่าเฉลี่ย $\beta_0 + \beta_1 X_i$ และค่าความแปรปรวน σ^2 ซึ่งจะอธิบายได้ด้วย $Y_i \sim iNid(\beta_0 + \beta_1 X_i, \sigma^2)$

วิธีกำลังสองน้อยที่สุดจะให้สมการถดถอย $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ ซึ่ง $\hat{Y}_i$, $\hat{\beta}_0$ และ $\hat{\beta}_1$ เป็นค่าประมาณของ Y_i , β_0 และ β_1 ตามลำดับ โดย β_0 เป็นจุดที่เส้นการถดถอยตัดแกน Y หรือค่าของ Y เมื่อ X มีค่าเท่ากับ 0 และ β_1 เป็นอัตราการเพิ่มหรือลดของ Y เมื่อ X มีค่าเพิ่มขึ้นหนึ่งหน่วย จะเรียก $\hat{\beta}_1$ ว่าค่าความลาดชัน ทั้ง β_0 และ β_1 จะเรียกว่าค่าสัมประสิทธิ์การถดถอย

2.6.1 การหาค่าประมาณแบบจุด

ค่า β_0 และ β_1 จะหาได้จากการแก้สมการที่ได้จากการหาอนุพันธ์ย่อยของ SSE เทียบกับ β_0 และ β_1 แล้วกำหนดให้เท่ากับ 0 ดังนี้

$$\begin{aligned} \text{SSE} &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2 \\ &= \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i \right)^2 \end{aligned}$$

หาค่า β_0 และ β_1 ที่ทำให้ SSE มีค่าต่ำสุดด้วยวิธีการหาอนุพันธ์ย่อย

$$\frac{\partial SSE}{\partial \beta_0} = -2\sum \left(Y_i - \beta_0 - \beta_1 X_i \right) = 0$$

$$\frac{\partial SSE}{\partial \beta_1} = -2\sum \left(Y_i - \beta_0 - \beta_1 X_i \right) X_i = 0$$

เขียนสมการทั้งสองใหม่และเรียกสมการใหม่นี้ว่าสมการปกติ (Normal Equation)

$$n\beta_0 + \sum X_i \beta_1 = \sum Y_i$$

$$\sum X_i \beta_0 + \sum X_i^2 \beta_1 = \sum X_i Y_i$$

จากสมการปกติ แก้สมการหาตัวประมาณ OLS ของ β_0 และ β_1 ได้เป็น

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_{XX}} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

และ $\text{Var}(\hat{\beta}_1) = \text{Var}\left(\frac{\sum(X_i-\bar{X})Y_i}{\sum(X_i-\bar{X})^2}\right)$

$= \frac{1}{SS_{XX}^2}\sum(X_i-\bar{X})^2\,\text{Var}(Y_i)$

$= \frac{\sigma^2}{SS_{XX}}$ ซึ่งสามารถประมาณด้วย $\frac{\sum(Y_i-\hat{Y}_i)^2}{(n-2)\sum(X_i-\bar{X})^2}$

$\text{Var}(\hat{\beta}_0) = \text{Var}(\bar{Y}-\beta_1\bar{X})$

$= \text{Var}(\bar{Y}) + \bar{X}^2\,\text{Var}(\beta_1) - 2\bar{X}\,\text{Cov}(\bar{Y},\beta_1)$

$= \sigma^2\left(\frac{1}{n}+\frac{\bar{X}^2}{SS_{XX}}\right)$

$= \frac{\sigma^2\sum X_i^2}{nSS_{XX}}$ ซึ่งสามารถประมาณด้วย $\frac{\sum(Y_i-\hat{Y}_i)^2\sum X_i^2}{(n-2)nSS_{XX}}$

2.6.2 **การหาค่าประมาณแบบช่วงด้วยวิธีแบบฉบับ**(Classical Method)

การหาช่วงความเชื่อมั่นของพารามิเตอร์ θ จะทำได้จากการที่ทราบลักษณะการแจกแจงของตัวสถิติ $t = \frac{\hat{\theta}-\mu_{\hat{\theta}}}{S_{\hat{\theta}}}$ ที่มีการแจกแจงแบบ t นั่นคือจากความน่าจะเป็น

$$P(-t_{\frac{\alpha}{2},df} \le t \le t_{\frac{\alpha}{2},df}) = 1-\alpha$$

หาก $\hat{\theta}$ เป็นตัวประมาณที่ไม่อคติจะมี $\mu_{\hat{\theta}}$ เท่ากับ θ ดังนั้นเมื่อแทน t ด้วย $\frac{\hat{\theta}-\mu_{\hat{\theta}}}{S_{\hat{\theta}}}$ จะเขียนค่าความน่าจะเป็นใหม่ได้เป็น

$$P(\hat{\theta}-t_{\frac{\alpha}{2},df}S_{\hat{\theta}} \le \theta \le \hat{\theta}+t_{\frac{\alpha}{2},df}S_{\hat{\theta}}) = 1-\alpha$$

ดังนั้นขีดจำกัดของ (1 - α)100% ช่วงความเชื่อมั่นของ θ จึงเท่ากับ $\hat{\theta} \pm t_{\frac{\alpha}{2},df} S_{\hat{\theta}}$

ในการวิเคราะห์การถดถอยเชิงเส้นเชิงเดียว พารามิเตอร์ θ ได้แก่ β_0 และ β_1 จึงได้ว่า (1 - α)100% ช่วงความเชื่อมั่นของ β_1 คือ

$$\beta_1 \pm t_{\alpha/2,n-2} \sqrt{\frac{\sum_i (Y_i - \hat{Y}_i)^2}{(n-2)\sum_i (X_i - \bar{X})^2}}$$

ขนาดของช่วงความเชื่อมั่นของ β_1 ที่ได้จากวิธีแบบฉบับคือ

$$2t_{\alpha/2,n-2} \sqrt{\frac{\sum_i (Y_i - \hat{Y}_i)^2}{(n-2)\sum_i (X_i - \bar{X})^2}}$$

และได้ว่า (1 - α)100% ช่วงความเชื่อมั่นของ β_0 คือ

$$\beta_0 \pm t_{\alpha/2,n-2} \sqrt{\frac{\sum_i (Y_i - \hat{Y}_i)^2 \sum_i X_i^2}{(n-2)nSS_{XX}}}$$

ขนาดของช่วงความเชื่อมั่นของ β_0 ที่ได้จากวิธีแบบฉบับคือ

$$2t_{\alpha/2,n-2} \sqrt{\frac{\sum_i (Y_i - \hat{Y}_i)^2 \sum_i X_i^2}{(n-2)nSS_{XX}}}$$

2.7 วิธีกำลังสองน้อยที่สุดแบบถ่วงน้ำหนักและปรับให้เหมาะสม (Adaptive Weighted Least Squares)

วิธีการนี้ได้ถูกสร้างขึ้นมาจากการรวบรวมแนวคิดหลายๆแนวคิดเข้าด้วยกันเพื่อใช้ในการประมาณค่าสัมประสิทธิ์การถดถอยอย่างมีประสิทธิภาพ แนวคิดดังกล่าวประกอบด้วย การปรับตัวแบบการถดถอย (Adaptive) การถ่วงน้ำหนัก(Weighted) การเรียงสับเปลี่ยน(Permutation) กระบวนการรอบบินส์ – มอนโร(Robbins – Monro search process) การเลือกค่าตั้งต้นที่เหมาะสมของกระบวนการรอบบินส์ – มอนโร (Starting values) เกณฑ์การสิ้นสุดกระบวนการรอบบินส์ – มอนโร(Stopping rule) การเลือกค่าปรับแก้ตัวประมาณในแต่ละขั้นตอนของกระบวนการรอบบินส์ – มอนโร(Choice of step length constant) โดยขั้นตอนหลักๆของวิธีการนี้จะมีอยู่ 4 ขั้นตอนซึ่งจะได้กล่าวถึงในหัวข้อ 2.7.1 , 2.7.2 , 2.6.3 และ 2.7.4 โดยตามลำดับ ส่วนในหัวข้อที่ 2.7.5 , 2.7.6 และ 2.7.7 จะเป็นส่วนประกอบที่ช่วยให้ได้ช่วงประมาณที่ดีขึ้น

2.7.1 <u>ขั้นตอนที่ 1</u> การปรับตัวแบบการถดถอย (Adaptive)

การปรับตัวแบบนั้นทำไปเพื่อให้ได้ตัวแบบลดรูปหรือที่เรียกว่า Reduce Model ซึ่งเป็นตัวแบบที่สามารถนำไปหาค่าน้ำหนักที่จะถ่วงให้กับแต่ละค่าสังเกตได้ การปรับตัวแบบมีวิธีการดังนี้

กำหนดตัวแบบเชิงเส้นคือ $Y = X_c\beta_c + \varepsilon$

โดย Y เป็นเวคเตอร์แนวตั้งขนาด n x 1 ของค่าสังเกต

X_c เป็นเมตตริกซ์ขนาด n x p

β_c เป็นเวคเตอร์แนวตั้งขนาด p x 1 ของพารามิเตอร์

ε เป็นเวคเตอร์แนวตั้งขนาด n x 1 ของ error

ขั้นที่ 1 แบ่ง X_c และ β_c ออกเป็น 2 ส่วน โดยให้ β_a เป็นพารามิเตอร์ตัวสุดท้ายใน β_c ส่วนที่เหลือคือ β_r ซึ่งมีขนาด (p - 1) x 1 จากนั้นก็แบ่ง X_c ให้สอดคล้องกัน โดยให้ X_a เป็นคอลัมภ์สุดท้ายของ X_c ส่วนที่เหลือคือ X_r ซึ่งมีขนาด n x (p – 1) จึงได้ตัวแบบใหม่คือ

$$Y = X_r\beta_r + X_a\beta_a + \varepsilon$$

ขั้นที่ 2 พิจารณา (1 - α)100% ช่วงความเชื่อมั่นของ β_a ค่า C จะอยู่ในช่วงความเชื่อมั่นเมื่อ p-value ของการทดสอบ H_0: $\beta_a = C$, H_1: $\beta_a > C$ มีค่ามากกว่า $\frac{\alpha}{2}$ และ p-value ของการทดสอบ H_0: $\beta_a = C$, H_1: $\beta_a < C$ มีค่ามากกว่า $\frac{\alpha}{2}$ จึงกำหนดให้ $\beta_{adj} = \beta_a - C$ และได้ตัวแบบเป็น

$$Y = X_r\beta_r + X_a(\beta_{adj} + C) + \varepsilon$$

ขั้นที่ 3 ให้ $Y_{adj} = Y - CX_a$ จะได้ว่า $Y_{adj} = X_r\beta_r + X_a\beta_{adj} + \varepsilon$

เห็นได้ว่าการทดสอบ H_0: $\beta_{adj} = 0$ จะเหมือนกับการทดสอบ H_0: $\beta_a = C$ และจะได้ตัวแบบลดรูป (Reduce Model) คือ

$$Y_{adj} = X_r\beta_r + \varepsilon$$

2.7.2 <u>ขั้นตอนที่ 2</u> **การถ่วงน้ำหนัก** (Weighted)

จุดประสงค์ของการถ่วงน้ำหนักในที่นี้คือ แปลงค่าสังเกตเพื่อทำให้ค่าเศษเหลือมีการแจกแจงใกล้เคียงกับการแจกแจงปกติ ดังนั้นการหาค่าน้ำหนักจึงต้องคำนวณจากค่าเศษเหลือไปตามลำดับ ดังนี้

ขั้นที่ 1 กำหนดเศษเหลือตัดทอน(Deleted Residual)คือ $d_i = Y_i - Y_{i(i)}$

($Y_{i(i)}$ คือค่าตัวแปรตามจากสมการถดถอยที่สร้างจากค่าสังเกต (n – 1) ค่าที่ไม่รวมค่าสังเกตที่ i)

ขั้นที่ 2 เศษเหลือตัดทอนมาตรฐาน(Studentized Deleted Residual) แทนด้วย g_i

คือ $g_i = \dfrac{d_i}{s(d_i)} = e_i\sqrt{\dfrac{(n-p-2)}{SSE_R(1-h_{ii}) - e_i^2}} = \dfrac{Y_i - Y_{i(i)}}{\sqrt{MSE_{(i)}h_{ii}}}$

โดย $s(d_i)$ คือค่าประมาณของส่วนเบี่ยงเบนมาตรฐานของ d_i

e_i คือค่าเศษเหลือ (Ordinary Residual)

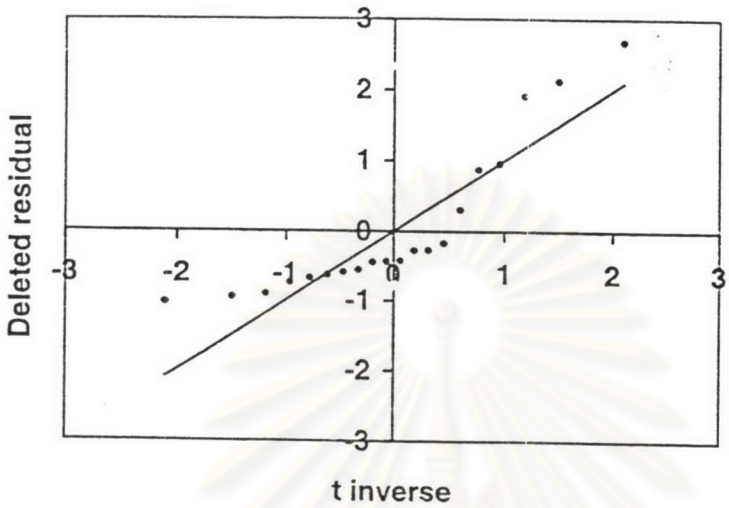
h_{ii} คือค่าที่ i ในแนวเฉียงของเมตตริกซ์ $X_R(X_R'X_R)^{-1}X_R'$

SSE_R คือผลรวมกำลังสองของเศษเหลือ(Sum of Square Residual) ของตัวแบบลดรูป(Reduce Model)

$MSE_{(i)}$ คือค่าเฉลี่ยกำลังสองของเศษเหลือ (Mean Square Error) ที่สร้างจากค่าสังเกต (n – 1) ค่าที่ไม่รวมค่าสังเกตที่ i

p คือจำนวนตัวแปรอิสระที่อยู่ในตัวแบบ

โดยทั่วไปแล้ว เศษเหลือตัดทอนมาตรฐานจะใช้วัดอิทธิพลของค่าสังเกตที่ i หรือ Y_i ที่มีต่อค่าประมาณ $\hat{Y}_i$ ถ้าค่าคลาดเคลื่อนมีการแจกแจงปกติแล้ว ค่าเศษเหลือตัดทอนมาตรฐานควรจะมีการแจกแจงแบบที่ด้วยองศาอิสระเท่ากับ n – p – 2 พิจารณาจากรูปต่อไปนี้



ภาพประกอบที่ 2.8 ตัวอย่าง probability plot ระหว่าง t – inverse กับ ค่าเศษเหลือตัดทอน

ค่าคลาดเคลื่อนจะมีการแจกแจงแบบปกติเมื่อค่าเศษเหลือตัดทอนมาตรฐานในรูปที่ 2.8 นี้เข้าใกล้เส้น 45 องศา หรือก็คือมีการแจกแจงแบบที่นั่นเอง ดังนั้นค่าน้ำหนักที่ต้องการ จึงต้องเป็นค่าที่สามารถปรับค่าเศษเหลือให้สอดคล้องกับเงื่อนไขดังกล่าวข้างต้น

ขั้นที่ 3 เศษเหลือตัดทอนมาตรฐานแบบปรับให้เรียบ(Smoothed Studentized Deleted Residual) หรือ S_i

เหตุผลของการปรับให้เรียบคือเพื่อทำให้ค่าน้ำหนักแต่ละตัวไม่ต่างกันมากจนเกินไป การปรับให้เรียบเริ่มด้วยการนำค่า g_i มาเรียงลำดับจากน้อยไปมาก และนำค่า g_2 , g_3 ,..., g_{n-1} มาคำนวณค่าเฉลี่ยเคลื่อนที่โดยใช้ 5 ค่า และนำมาใส่ในตำแหน่งที่ 3 , 4 , ... , n – 4 ซึ่งสามารถแสดงให้ดูง่ายได้ดังนี้

$$S_1 = g_1$$

$$S_2 = g_2$$

$$S_3 = \frac{g_2 + g_3 + g_4 + g_5 + g_6}{5}$$

$$S_4 = \frac{g_3 + g_4 + g_5 + g_6 + g_7}{5}$$

$$\vdots$$

$$S_{n-4} = \frac{g_{n-5} + g_{n-4} + g_{n-3} + g_{n-2} + g_{n-1}}{5}$$

$$S_{n-3} = g_{n-3}$$

$$S_{n-2} = g_{n-2}$$

$$S_{n-1} = g_{n-1}$$

$$S_n = g_n$$

ขั้นที่ 4 เศษเหลือตัดทอนมาตรฐานแบบศูนย์กลางและปรับให้เรียบ (Centered and Smoothed Studentized Deleted Residual) หรือ c

โดย $c_i = s_i - \text{median}(s_1, \ldots, s_n)$ เมื่อ $i = 1, \ldots, n$

เหตุผลของการทำให้เข้าสู่ศูนย์กลางคือเพื่อทำให้ค่าเศษเหลือในรูปที่ 2.8 ผ่านจุดกำเนิด(Origin) และเพื่อทำให้ค่าน้ำหนักที่จะคำนวณหาต่อไปนั้นมีค่าเป็นบวกเสมอ

ขั้นที่ 5 การหาค่าน้ำหนักสำหรับข้อมูลแต่ละตัว หรือ w_i ($i = 1, \ldots, n$)

กำหนดให้ $t_i = t^{-1}_{(R_i - 0.5)/n, n-p-2}$ โดย R_i คือค่าอันดับ(Rank)ของ c_i

จึงได้ค่าน้ำหนักดังนี้

$$w_i = \min(t_i / c_i, 1.5) \quad \text{เมื่อ } c_i \neq 0$$

$$= 1 \quad \text{เมื่อ } c_i = 0$$

จึงสามารถสร้างตัวแบบถ่วงน้ำหนักได้โดยนำเมตตริกซ์ของค่าน้ำหนักไปคูณกับตัวแบบได้เป็น

$$wY_{adj} = wX_r\beta_r + wX_a\beta_{adj} + w\varepsilon$$

2.7.3 **ขั้นตอนที่ 3 การเรียงสับเปลี่ยน**(Permutation)

การเรียงสับเปลี่ยนในที่นี้เป็นการประยุกต์ใช้การทดสอบแบบสุ่ม (Randomization Test) กับการทดสอบสมมติฐานเพื่อเพิ่มประสิทธิภาพในการประมาณค่าขอบเขตบนและขอบเขตล่างของกระบวนการรอบบินส์-มอนโร งานวิจัยของแมนลี่ (Manly , 1991) ได้ให้ผลสรุปว่า การทดสอบแบบสุ่มเป็นวิธีการที่น่าสนใจสำหรับการทดสอบสมมติฐาน เพราะมีข้อสมมติเกี่ยวกับแจกแจงน้อยกว่าการทดสอบที่เกี่ยวกับพารามิเตอร์ (Parametric Test) และมีอำนาจการทดสอบที่ดี การเรียงสับเปลี่ยนที่ใช้ในงานวิจัยชิ้นนี้มีวิธีการดังนี้

จากตัวแบบที่ได้ทำการถ่วงน้ำหนักแล้ว นำมาเรียงสับเปลี่ยนสมาชิกของเวคเตอร์ X_a ทำให้ตัวแบบกลายเป็น

$$wY_{adj} = wX_r\beta_r + wX_{a,i}\beta_{adj} + w\varepsilon$$

โดย $X_{a,i}$ คือ X_a ที่ถูกเรียงสับเปลี่ยนสมาชิกอย่างสุ่ม

ส่วนค่าน้ำหนักนั้นไม่จำเป็นต้องคำนวณใหม่เพราะคิดมาจาก Y_{adj} และ X_r ซึ่งไม่ได้ถูกเรียงสับเปลี่ยน

2.7.4 **ขั้นตอนที่ 4 กระบวนการรอบบินส์-มอนโร** (Robbins – Monro Search Process)

วิธีการนี้ถูกสร้างขึ้นโดยรอบบินส์และมอนโร (Robbins and Monro , 1951) เพื่อใช้ในการหาขอบเขตบน (Upper Bound) และขอบเขตล่าง (Lower Bound) ของช่วงประมาณด้วยหลักการค้นหาแบบลำดับ (Sequential Search) ในแต่ละขั้นตอนของกระบวนการจะมีการสุ่มตัวอย่างซ้ำ (Resample) เพื่อปรับปรุงช่วงประมาณให้ดีขึ้น แต่ในงานวิจัยชิ้นนี้จะใช้การเรียงสับเปลี่ยนแทนการสุ่มตัวอย่างซ้ำ การเรียงสับเปลี่ยนหรือการสุ่มตัวอย่างซ้ำจะเป็นส่วนหนึ่งในการสร้างค่าปรับแก้ตัวประมาณในแต่ละขั้นตอน งานวิจัยของฮอดเจสและเลห์แมน (Hodges and Lehman , 1955) ได้ให้ผลสรุปว่าการเลือกใช้ค่าปรับแก้ตัวประมาณที่เหมาะสมจะทำให้ได้ช่วงประมาณที่ดี การใช้กระบวนการรอบบินส์ – มอนโรในการหาช่วงความเชื่อมั่นของสัมประสิทธิ์การถดถอยสำหรับงานวิจัยชิ้นนี้มีวิธีการดังนี้

กำหนดให้ L_i เป็นค่าประมาณของขอบเขตล่างของ β_a ในขั้นตอนลำดับที่ i

U_i เป็นค่าประมาณของขอบเขตบนของ β_a ในขั้นตอนลำดับที่ i

t_L คือค่าสถิติทดสอบสำหรับ H_0: $\beta_a = L_i$ ของตัวแบบถ่วงน้ำหนักที่ยังไม่ถูกเรียงสับเปลี่ยนสมาชิกใน X_a

t_U คือค่าสถิติทดสอบสำหรับ H_0: $\beta_a = U_i$ ของตัวแบบถ่วงน้ำหนักที่ยังไม่ถูกเรียงสับเปลี่ยนสมาชิกใน X_a

t_L^{per} คือค่าสถิติทดสอบสำหรับ H_0: $\beta_a = L_i$ ของตัวแบบถ่วงน้ำหนักที่ถูกเรียงสับเปลี่ยนสมาชิกใน X_a

t_U^{per} คือค่าสถิติทดสอบสำหรับ H_0: $\beta_a = U_i$ ของตัวแบบถ่วงน้ำหนักที่ถูกเรียงสับเปลี่ยนสมาชิกใน X_a

k_i คือค่าปรับแก้ตัวประมาณในขั้นตอนที่ i (จะได้กล่าวถึงรายละเอียดในหัวข้อ 2.7.7)

$$U_{i+1} = \begin{cases} U_i - k_i \dfrac{\alpha}{2i} & \text{ถ้า } t_U^{per} > t_U \\ U_i + k_i \dfrac{1-\alpha/2}{i} & \text{ถ้า } t_U^{per} \leq t_U \end{cases}$$

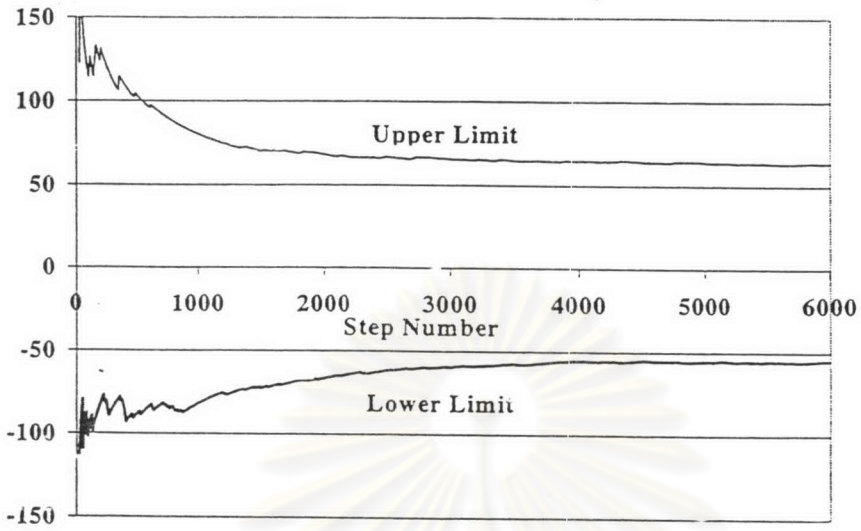
สำหรับค่า U เริ่มต้นจะใช้ค่าขอบเขตบนของช่วงความเชื่อมั่น (1- α)100% ที่ได้จากวิธีแบบฉบับ

ส่วนขอบเขตล่างจะหาจาก

$$L_{i+1} = \begin{cases} L_i + k_i \dfrac{\alpha}{2i} & \text{ถ้า } t_L^{per} < t_L \\ L_i - k_i \dfrac{1-\alpha/2}{i} & \text{ถ้า } t_L^{per} \geq t_L \end{cases}$$

ในทำนองเดียวกันค่า L เริ่มต้นจะใช้ค่าขอบเขตล่างของช่วงความเชื่อมั่น (1- α)100% ที่ได้จากวิธีแบบฉบับ ส่วนการเริ่มต้นของกระบวนการนั้นจะไม่นิยมเริ่มต้นที่ i = 1 ซึ่งรายละเอียดในส่วนนี้จะกล่าวถึงในหัวข้อ 2.7.5

งานวิจัยของกาทเวต (Garthwaite ,1996) ได้เสนอว่าอย่างน้อยควรทำไปจนถึง i = 5,999 จึงจะได้ค่าประมาณที่ใกล้เคียงกับค่าขอบเขตที่แท้จริง และในงานวิจัยที่กล่าวถึงนี้ยังได้แสดงกราฟตัวอย่างซึ่งชี้ให้เห็นว่าเมื่อกระบวนการดำเนินไป ขอบเขตบนและขอบเขตล่างจะลู่เข้าสู่ค่าคงที่ที่เหมาะสม กราฟดังกล่าวเป็นดังนี้



ภาพประกอบที่ 2.9 ค่าประมาณของขอบเขตบนและขอบเขตล่างของระดับไขมันไตรกลีเซอไรด์(Triglyceride) ในเลือดเมื่อกระบวนการดำเนินไป

(สำหรับที่มาของกระบวนการรอบบินส์ – มอนโรจะแสดงไว้ในภาคผนวก)

2.7.5 การเลือกค่าตั้งต้นที่เหมาะสมของกระบวนการรอบบินส์ – มอนโร(Starting values)

การเริ่มต้นกระบวนการด้วยค่า i = 1 หรือ i = 2 นั้น เป็นการยากที่จะได้ค่าประมาณที่ใกล้เคียงค่าจริงของขอบเขตความเชื่อมั่น เอฟรอน (Efron ,1981) ได้เสนอว่าควรเลือกจุดตั้งต้นของกระบวนการโดยพิจารณาถึงเปอร์เซ็นไทล์ และได้สรุปวิธีเลือกค่าตั้งต้นไว้ดังนี้

1. ค่า i เริ่มต้น = $0.3\dfrac{2-\frac{\alpha}{2}}{\frac{\alpha}{2}}$ (ถ้าค่าที่ได้เป็นทศนิยมให้ปัดทศนิยมทิ้ง)

2. ค่าที่ได้จากข้อ 1 ไม่ควรเกิน 50 ถ้าเกินให้ใช้ค่า i ตั้งต้นเท่ากับ 50

จากเกณฑ์ดังกล่าวจึงสามารถคำนวณค่า i เริ่มต้นที่จะใช้ในงานวิจัยชิ้นนี้ได้ดังนี้

ที่ระดับความเชื่อมั่น 90% ค่า i เริ่มต้น = $0.3\dfrac{2-\frac{0.1}{2}}{\frac{0.1}{2}} = 11.7$ ปัดทศนิยมทิ้งได้เป็น 11

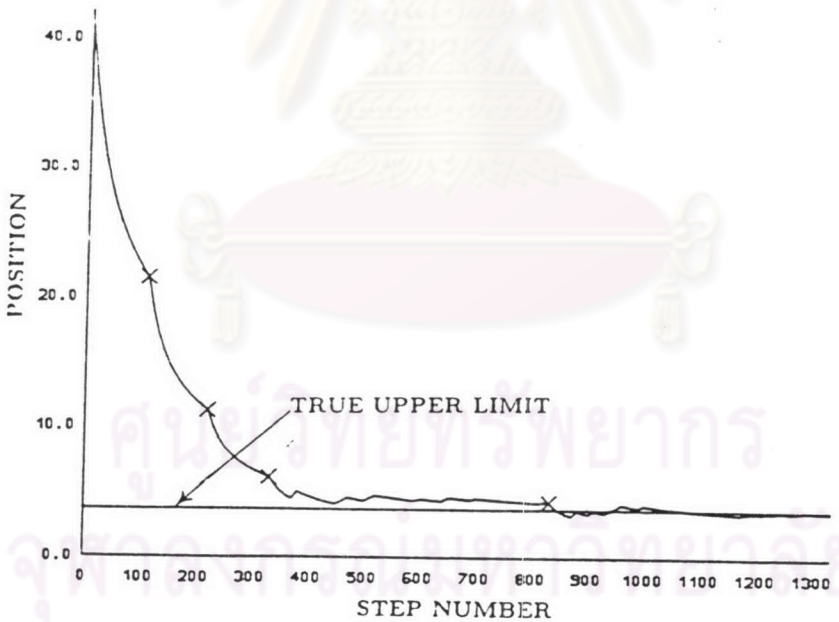
ที่ระดับความเชื่อมั่น 95% ค่า i เริ่มต้น = $0.3\dfrac{2-\frac{0.05}{2}}{\frac{0.05}{2}} = 23.7$ ปัดทศนิยมทิ้งได้เป็น 23

ที่ระดับความเชื่อมั่น 99% ค่า i เริ่มต้น = $0.3\dfrac{2-\frac{0.01}{2}}{\frac{0.01}{2}} = 119.7$ ค่าที่ได้เกินกว่า 50 จึงปัดเป็น 50

2.7.6 **เกณฑ์การสิ้นสุดกระบวนการรอบบินส์ – มอนโร**(Stopping rule)

เมื่อกระบวนการดำเนินไป (i เพิ่มขึ้น) ประสิทธิภาพของช่วงประมาณก็จะสูงขึ้น แต่เมื่อถึงจุดๆหนึ่ง แม้ว่าจะดำเนินกระบวนการต่อไปก็ไม่อาจเพิ่มประสิทธิภาพได้อีก มีแต่จะสิ้นเปลืองเวลาของคอมพิวเตอร์มากขึ้น จึงมีการกำหนดหลักเกณฑ์ในการหาจุดสิ้นสุดกระบวนการซึ่งแบ่งออกเป็น 2 หลักเกณฑ์ดังนี้

2.7.6.1 *พิจารณาจากกราฟ* ยกตัวอย่างกราฟต่อไปนี้



ภาพประกอบที่ 2.10 การพิจารณาจุดสิ้นสุดกระบวนการจากกราฟ

จะพบว่าค่าประมาณเข้าใกล้ค่าจริงมากที่สุดเมื่อ i = 830 เป็นต้นไป แม้ว่าจะดำเนินกระบวนการต่อไปก็ไม่อาจได้ค่าประมาณที่ดีกว่านี้ ดังนั้นควรใช้ i = 830 เป็นจุดสิ้นสุดกระบวนการสำหรับสถานการณ์นี้

2.7.6.2 *ใช้การทดสอบสมมุติฐาน* เนื่องจากหัวข้อนี้มีส่วนเกี่ยวเนื่องกับที่มาของกระบวนการรอบบินส์ – มอนโรเป็นอย่างมาก จึงขอนำไปอธิบายพร้อมกันกับที่มาของกระบวนการรอบบินส์ – มอนโรในภาคผนวก

สำหรับงานวิจัยชิ้นนี้จะใช้ i = 5,999 เป็นจุดสิ้นสุดกระบวนการตามผลสรุปจากงานวิจัยของกาทเวต (Garthwaite ,1996)

2.7.7 การเลือกค่าปรับแก้ตัวประมาณในแต่ละขั้นตอนของกระบวนการรอบบินส์ – มอนโร (Choice of step length constant)

จากการประยุกต์ใช้งานวิจัยของกาทเวตและบัคแลนด์ (Garthwaite & Buckland , 1992) ทำให้ได้ข้อสรุปว่าค่า k_i ที่เหมาะสมสำหรับการหาขอบเขตบนคือ $k_i = c(U_i - \hat{\beta}_a)$ ส่วนการหาขอบเขตล่างจะใช้ $k_i = c(\beta_a - L_i)$ และจะใช้ $c = \dfrac{2}{Z_\alpha (2\pi)^{-\frac{1}{2}} \exp(-\frac{Z_\alpha^2}{2})}$

โดย $\hat{\beta}_a$ เป็นตัวประมาณกำลังสองน้อยที่สุดที่คำนวณจากรูปแบบเต็ม(Complete Model) ที่ถ่วงน้ำหนักแล้ว

ดังนั้นในกรณี α = 0.01 ควรใช้ $k_i = 53.73(U_i - \hat{\beta}_a)$ สำหรับการประมาณค่าขอบเขตบน

และควรใช้ $k_i = 53.73(\beta_a - L_i)$ สำหรับการประมาณค่าขอบเขตล่าง

กรณี α = 0.05 ควรใช้ $k_i = 17.46(U_i - \hat{\beta}_a)$ สำหรับการประมาณค่าขอบเขตบน

และควรใช้ $k_i = 17.46(\beta_a - L_i)$ สำหรับการประมาณค่าขอบเขตล่าง

กรณี α = 0.1 ควรใช้ $k_i = 11.79(U_i - \hat{\beta}_a)$ สำหรับการประมาณค่าขอบเขตบน

และควรใช้ $k_i = 11.79(\beta_a - L_i)$ สำหรับการประมาณค่าขอบเขตล่าง

2.8 วิธีบูตสแตรป(BOOTSTRAP)

วิธีการหาค่าตัวประมาณของพารามิเตอร์วิธีนี้ถูกเสนอขึ้นโดย แบรดเลย์ เอฟรอน (Bradley Efron) ในปี ค.ศ. 1979 ซึ่งมีแนวคิดมาจากวิธีแจกไนฟ์ (Jackknife) ของเควโนอีล (Queneuille) และตูกี้ (Tukey) ซึ่งวิธีบูตสแตรปนี้สามารถนำมาแก้ปัญหาการไม่สามารถหาค่าประมาณในกรณีที่ข้อมูลไม่เป็นไปตามข้อตกลงเบื้องต้น (Assumptions) เช่น ความคลาดเคลื่อนมีการแจกแจงที่ไม่เป็นแบบปกติ หรือหาค่าประมาณได้ยาก เช่น การประมาณส่วนเบี่ยงเบนมาตรฐานของสัมประสิทธิ์สหสัมพันธ์

วิธีบูตสแตรปมีหลักการดังนี้คือ กำหนดให้ตัวอย่างที่ถูกเก็บรวบรวมมาจากประชากรเปรียบเสมือนประชากร แล้วทำการสุ่มแบบใส่คืน (Resampling with Replacement) จากตัวอย่างที่มีอยู่ด้วยจำนวนครั้งที่มากพอ เพื่อสร้างการแจกแจงของตัวสถิติตัวอย่าง (Sampling Distribution) และนำไปใช้ประมาณค่าพารามิเตอร์ที่สนใจ

ในการประมาณค่าแบบช่วงของ β_0 และ β_1 ในสมการถดถอยเชิงเส้นเชิงเดียว $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ เมื่อ ε_i ไม่ได้แจกแจงแบบปกติ ดังนั้น $t = \dfrac{\beta_i - \beta_i}{S(\beta_i)}$ $i = 0, 1$ จึงไม่ได้แจกแจงแบบที ที่องศาอิสระ (n – 2) จึงต้องประมาณค่าโดยการสร้างการแจกแจงใหม่ ด้วยวิธีบูตสแตรปดังขั้นตอนต่อไปนี้

1) จำลองข้อมูลชุดแรกเริ่มคือ $\{ (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \}$ โดยเป็นข้อมูลชุดเดียวกับที่นำไปหาค่าประมาณด้วยวิธีแบบฉบับ(CM) และวิธีกำลังสองน้อยที่สุดแบบถ่วงน้ำหนักและปรับให้เหมาะสม (AWLS)

2) จากข้อมูลชุดแรกเริ่ม นำมาสุ่มซ้ำแบบคืนที่ 2,000 ชุด โดยแต่ละชุดมีขนาด n ดังนี้

$\{ (x_{1,1}, y_{1,1}), (x_{2,1}, y_{2,1}), \ldots, (x_{n,1}, y_{n,1}) \}$ ชุดข้อมูลจากการบูตสแตรปครั้งที่ 1

$\{ (x_{1,2}, y_{1,2}), (x_{2,2}, y_{2,2}), \ldots, (x_{n,2}, y_{n,2}) \}$ ชุดข้อมูลจากการบูตสแตรปครั้งที่ 2

$\{ (x_{1,j}, y_{1,j}), (x_{2,j}, y_{2,j}), \ldots, (x_{n,j}, y_{n,j}) \}$ ชุดข้อมูลจากการบูตสแตรปครั้งที่ j

โดย $(x_{i,j}, y_{i,j})$ คือข้อมูลลำดับที่ i ในการบูตสแตรปครั้งที่ j และ $j = 1, 2, \ldots, 2{,}000$

3) จากข้อมูล $\{ (x_{1,j}, y_{1,j}), (x_{2,j}, y_{2,j}), \ldots, (x_{n,j}, y_{n,j}) \}$ แต่ละชุด นำมาคำนวณหาค่าประมาณ OLS

$\hat{\beta}_{0,j}$, $\hat{\beta}_{1,j}$ โดย $j = 1, 2, \ldots, 2{,}000$

4) จากค่า $\hat{\beta}^{*}_{0,j}$ และ $\hat{\beta}^{*}_{1,j}$ ทั้งหมด (j = 1 , 2 , ... , 2,000) นำมาหาค่าเฉลี่ยและค่าเบี่ยงเบนมาตรฐาน

$$\bar{\hat{\beta}}^{*}_{0} = \frac{1}{2000}\sum_{j=1}^{2000}\hat{\beta}^{*}_{0,j} \quad , \quad \bar{\hat{\beta}}^{*}_{1} = \frac{1}{2000}\sum_{j=1}^{2000}\hat{\beta}^{*}_{1,j}$$

$$S(\hat{\beta}^{*}_{0}) = \sqrt{\frac{\sum(\hat{\beta}^{*}_{0,j} - \bar{\hat{\beta}}^{*}_{0})^2}{2.000 - 1}}$$

$$S(\hat{\beta}^{*}_{1}) = \sqrt{\frac{\sum(\hat{\beta}^{*}_{1,j} - \bar{\hat{\beta}}^{*}_{1})^2}{2.000 - 1}}$$

5) คำนวณค่าสถิติ

$$t^{*}_{0,j} = \frac{\hat{\beta}^{*}_{0,j} - \hat{\beta}_{0}}{S(\hat{\beta}_{0})} \qquad \text{สำหรับทุกค่า } j = 1, 2, \ldots, 2{,}000$$

$$t^{*}_{1,j} = \frac{\hat{\beta}^{*}_{1,j} - \hat{\beta}_{1}}{S(\hat{\beta}^{*}_{1})} \qquad \text{สำหรับทุกค่า } j = 1, 2, \ldots, 2{,}000$$

โดย $\hat{\beta}_0$ และ $\hat{\beta}_1$ คือตัวประมาณที่ได้จากวิธี OLS โดยใช้ข้อมูลชุดแรกเริ่ม

6) จากค่า $\{ t^{*}_{0,1}, t^{*}_{0,2}, \ldots, t^{*}_{0,2000} \}$ นำมาเรียงค่าจากน้อยไปมาก หาค่าเปอร์เซ็นไทล์ที่ $100(\frac{\alpha}{2})$ กำหนดให้เป็น $t^{*}_{1-\frac{\alpha}{2}}$ สำหรับ β_0 และหาค่าเปอร์เซ็นไทล์ที่ $100(1 - \frac{\alpha}{2})$ กำหนดให้เป็น $t^{*}_{\frac{\alpha}{2}}$ สำหรับ β_0 ในทำนองเดียวกันสำหรับ β_1 นำค่า $\{ t^{*}_{1,1}, t^{*}_{1,2}, \ldots, t^{*}_{1,2000} \}$ มาเรียงลำดับเพื่อหาค่า $t^{*}_{1-\frac{\alpha}{2}}$ สำหรับ β_1 และ $t^{*}_{\frac{\alpha}{2}}$ สำหรับ β_1 การหาค่าเปอร์เซ็นไทล์จะทำได้ดังนี้

ก) การหา $t^{*}_{\frac{\alpha}{2}}$ สำหรับ β_0

กำหนดตัวแปร Location = $(2000+1)(1-\frac{\alpha}{2})$

Temp = Location – INT(Location) , INT( ) เป็นฟังก์ชันในการหาจำนวนเต็มโดยการปัดทศนิยมทิ้ง

Ptr = INT(Location) + 1

จะได้ $t^{*}_{\frac{\alpha}{2}}$ สำหรับ β_0 เท่ากับ (Temp X $t_{0,\,Ptr}$) + [(1-Temp) X $t_{0,\,Ptr-1}$]

ข) การหา $t^{*}_{1-\frac{\alpha}{2}}$ สำหรับ β_0

กำหนดตัวแปร Location = $(2000+1)(\frac{\alpha}{2})$

Temp = Location – INT(Location)

Ptr = INT(Location) + 1

จะได้ $t^{*}_{1-\frac{\alpha}{2}}$ สำหรับ β_0 เท่ากับ (Temp X $t_{0,\,Ptr}$) + [(1-Temp).X $t_{0,\,Ptr-1}$]

ค) การหา $t^{*}_{\frac{\alpha}{2}}$ สำหรับ β_1

กำหนดตัวแปร Location = $(2000+1)(1-\frac{\alpha}{2})$

Temp = Location – INT(Location)

Ptr = INT(Location) + 1

จะได้ $t^{*}_{\frac{\alpha}{2}}$ สำหรับ β_1 เท่ากับ (Temp X $t_{1,\,Ptr}$) + [(1-Temp) X $t_{1,\,Ptr-1}$]

ง) การหา $t^{*}_{1-\frac{\alpha}{2}}$ สำหรับ β_1

กำหนดตัวแปร Location = $(2000+1)(\frac{\alpha}{2})$

Temp = Location – INT(Location)

Ptr = INT(Location) + 1

จะได้ $t^{*}_{1-\frac{\alpha}{2}}$ สำหรับ β_1 เท่ากับ (Temp X $t_{1,\,Ptr}$) + [(1-Temp) X $t_{1,\,Ptr-1}$]

7) คำนวณช่วงความเชื่อมั่น 100(1 - α)% สำหรับ β_0 คือ

$$\left[\hat{\beta}_0 - t^{*}_{1-\frac{\alpha}{2}} S(\beta_0),\ \hat{\beta}_0 + t^{*}_{\frac{\alpha}{2}} S(\beta_0)\right]$$

และคำนวณช่วงความเชื่อมั่น 100(1 - α)% สำหรับ β_1 คือ

$$\left[\hat{\beta}_1 - t^{*}_{1-\frac{\alpha}{2}} S(\beta_1),\ \beta 1 + t^{*}_{\frac{\alpha}{2}} S(\beta_1)\right]$$

โดย $\hat{\beta}_0$ และ β_1 คือตัวประมาณที่ได้จากวิธี OLS โดยใช้ข้อมูลชุดแรกเริ่ม

8) กลับไปทำขั้นตอนที่ 1 จนถึงขั้นตอนที่ 7 จนครบ 1,000 รอบ

9) ตรวจสอบว่ามีกี่ช่วงที่ครอบคลุมค่าจริง β_i, i = 0 , 1 แล้วหาค่าสัดส่วนโดยหารด้วย 1,000 เป็นค่าสัมประสิทธิ์ความเชื่อมั่นจากตัวอย่าง(1 - α)