บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

วิธีที่ใช้ในการประมาณค่าสัมประสิทธิ์ในสมการถดถอยเชิงเส้นมีอยู่ด้วยกันหลายวิธี แต่ผู้วิจัยควรเลือกใช้วิธีที่เหมาะสมกับลักษณะของข้อมูล และเป็นไปข้อตกลงเบื้องต้น (Assumption) ของแต่ละวิธีเพราะจะทำให้ได้ตัวประมาณที่ดีมีประสิทธิภาพ ปัญหาส่วนใหญ่จะพบว่า ลักษณะการแจกแจงของความคลาดเคลื่อน (error) ไม่เป็นไปตามข้อตกลงเบื้องต้น คือ การแจกแจงไม่เป็นแบบปกติ

ในการศึกษาสมการถดถอยเชิงเส้น ซึ่งมีตัวแบบ (model) ดังนี้

$$Y = X\beta + \varepsilon$$

เมื่อ Y เป็นเวคเตอร์ของตัวแปรตามขนาด $n \times 1$

X เป็นเมตริกซ์ของตัวแปรอิสระขนาด $n \times p$ ซึ่งมี $rank = p$

β เป็นเวคเตอร์ของพารามิเตอร์ที่ไม่ทราบค่า ซึ่งเป็นค่าสัมประสิทธิ์ความถดถอยขนาด $p \times 1$

ε เป็นเวคเตอร์ของความคลาดเคลื่อนขนาด $n \times 1$

n เป็นขนาดของข้อมูลตัวอย่าง

p เป็นจำนวนตัวแปรอิสระที่ใช้ในสมการถดถอยเชิงเส้น

โดยมีข้อตกลงเบื้องต้น (Assumption) เกี่ยวกับความคลาดเคลื่อนดังนี้ คือ

1. $E(\varepsilon_i) = 0 \quad ; i = 1,2,\ldots,n$
2. $E(\varepsilon_i \varepsilon_j) = 0 \quad ; i \neq j, i,j = 1,2,\ldots,n$

$$V(\varepsilon_i) = \sigma^2 \quad ; \quad \sigma^2 \text{ ไม่ทราบค่า}$$

3. $\varepsilon_i \sim NID(0, \sigma^2)$

วิธีที่นิยมใช้กันมากในการประมาณค่าสัมประสิทธิ์ความถดถอย คือ วิธีกำลังสองต่ำสุด (Least Squares Method) ซึ่งจะได้ $\hat{\beta} = (X'X)^{-1}X'Y$ เป็นตัวประมาณที่ไม่เอนเอียงสำหรับ β และให้ค่าความแปรปรวนต่ำสุด $= \sigma^2(X'X)^{-1}$

นั่นก็คือ $\hat{\beta}$ มีคุณสมบัติเป็นตัวประมาณเชิงเส้นที่ดีที่สุดและไม่เอนเอียง(Best Linear Unbiased Estimator (BLUE)) สำหรับ ตามทฤษฎีของ Gauss - Markov และจะประมาณค่า σ^2 ด้วย $\hat{\sigma}^2$ ซึ่ง $\hat{\sigma}^2 = \frac{1}{n-p}(Y-X\beta)'(Y-X\beta)$ เป็นตัวประมาณที่ไม่เอนเอียงสำหรับ σ^2 แต่ $\hat{\beta}$ และ $\hat{\sigma}^2$ จะเป็นตัวประมาณที่มีประสิทธิภาพ (Efficiency Estimator) คือมีค่าความแปรปรวนต่ำสุดเพียงตัวเดียวในบรรดาตัวประมาณที่ไม่เอนเอียง (Uniformly Minimum Variance Estimator (UMVUE)) ก็ต่อเมื่อ ε มีการแจกแจงแบบปกติและมีคุณสมบัติตามข้อกำหนดข้างต้น ในกรณีที่ไม่ทราบการแจกแจงของ ε จึงควรพิจารณาหาวิธีการอื่นในการประมาณค่าของ β และ σ^2 ที่ดีกว่าตัวประมาณที่ได้จากวิธีกำลังสองต่ำสุดโดยไม่จำเป็นต้องมีข้อตกลงเบื้องต้นเกี่ยวกับลักษณะการแจกแจงของความคลาดเคลื่อน

บรรดาวิธีการทางนอนพาราเมตริก ได้มีผู้ศึกษาวิธีการประมาณค่า standard error ในกรณีที่ไม่ทราบลักษณะการแจกแจงของประชากรและไม่สามารถหาได้จากสูตรทั่วไป โดยใช้เทคนิคของการสุ่มตัวอย่าง (resampling) ซึ่งมีอยู่ด้วยกันหลายวิธี ได้แก่

- The jackknife
- The bootstrap
- Half - sampling
- Subsampling
- Balanced repeated replication
- The infinistesimal jackknife
- Influence function techniques

- The delta method

ซึ่งแต่ละวิธีมาจากแนวความคิดพื้นฐานคล้ายกันคือ หาค่าประมาณของ standard error โดยการสุ่มตัวอย่างซ้ำจากข้อมูลที่เก็บรวบรวมมา และพบว่าวิธีบูตสแตรป(bootstrap) เป็นวิธีที่ให้ผลดีที่สุดเพราะว่า การหาค่าประมาณโดยวิธีนี้เป็น nonparametric maximum likelihood estimate ทำให้ตัวประมาณที่ได้เป็นตัวประมาณแบบภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimator(MLE))[1] นอกจากนั้นวิธีบูตสแตรปยังสามารถนำไปใช้ในการประมาณค่าพารามิเตอร์อื่น ๆ ที่สนใจ เมื่อไม่ทราบลักษณะการแจกแจงของประชากร

ในปี ค.ศ. 1979 Bradley Efron ได้ศึกษาวิธีบูตแสตรปซึ่งมีแนวคิดมาจากวิธี Jackknife ของ Quenouille (1949) และ Turkey (1958) มาใช้ในการประมาณค่าต่าง ๆ ที่ไม่สามารถหาได้โดยตรงในทางนอนพาราเมตริก เช่นค่าส่วนเบี่ยงเบนมาตรฐาน (standard deviation) ของค่าสัมประสิทธิ์สหสัมพันธ์ (correlation coefficient) ลักษณะการแจกแจงของตัวประมาณ ช่วงความเชื่อมั่น ความเอนเอียง รวมทั้งนำมาใช้ในการประมาณค่าส่วนเบี่ยงเบนมาตรฐานของสัมประสิทธิ์ความถดถอยเชิงเส้น ในกรณีที่ไม่ทราบลักษณะการแจกแจงของความคลาดเคลื่อน จึงสนใจศึกษาและนำเอาวิธีบูตสแตรปมาใช้ในการประมาณประมาณค่าพารามิเตอร์พร้อมทั้งหาค่าความแปรปรวน ในการวิเคราะห์ความถดถอยเชิงเส้นที่มีรูปแบบทั่วไป (Ordinary linear regression) คือ $Y = X\beta + \varepsilon$ ในกรณีที่ไม่ทราบลักษณะการแจกแจงของความคลาดเคลื่อน โดยใช้เทคนิคการจำลองแบบที่เรียกว่า เทคนิคมอนติคาร์โลซิมูเลชั่น (Monte Carlo Simulation Technique) ในการสร้าง (generate) ข้อมูลตามขนาดและลักษณะที่ต้องการ พร้อมทั้งศึกษาคุณสมบัติของตัวประมาณและเปรียบเทียบประสิทธิภาพของตัวประมาณในรูปของค่าประสิทธิภาพสัมพัทธ์ (Relative Efficiecy) หรือ RE ของวิธีบูตสแตรปเทียบกับวิธีกำลังสองต่ำสุด โดยใช้ความคลาดเคลื่อนกำลังสองเฉลี่ย

---

[1] Efron, B. "The Bootstrap" The Jackknife,the Bootstraps and Other Resampling plans (1982):27.

ของตัวประมาณแต่ละตัว และฟังก์ชั่นเชิงเส้นของตัวประมาณในแต่ละวิธีเป็นค่าเปรียบเทียบ เนื่องจากในการเปรียบเทียบเชิงทฤษฎีเราทราบค่าพารามิเตอร์ ดังนั้นสามารถใช้ความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error) หรือ MSE เป็นเกณฑ์ในการเปรียบเทียบได้ ซึ่งค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวประมาณ จะเป็นค่าที่แสดงถึงความเที่ยงตรง (accuracy) ของตัวประมาณ โดยตัวประมาณที่มีความเที่ยงตรงที่สุดก็คือ ตัวประมาณที่มีความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาและนำเอาวิธีบูตสแตรปมาใช้ประมาณค่าพารามิเตอร์ในการวิเคราะห์ความถดถอยเชิงเส้นและค่าความแปรปรวนของตัวประมาณ พร้อมทั้งศึกษาคุณสมบัติของตัวประมาณ

2. เพื่อศึกษาเปรียบเทียบประสิทธิภาพของตัวประมาณที่ได้จากวิธีกำลังสองต่ำสุดและวิธีบูตสแตรปโดยใช้ความคลาดเคลื่อนกำลังสองเฉลี่ยเป็นตัวเปรียบเทียบ

1.3 ข้อตกลงเบื้องต้น

1. รูปแบบ (model) ที่ใช้ถูกต้อง

2. ความคลาดเคลื่อนเป็นตัวแปรสุ่มแบบต่อเนื่องที่มีการแจกแจงแบบเดียวกันและเป็นอิสระซึ่งกันและกัน (identically independent distribution)

นั่นคือ $\varepsilon_i \sim iid \quad F \quad ; i = 1,2,\ldots,n$

F เป็น probability distribution ที่ไม่ทราบ

$E(\varepsilon_i) = 0 \quad ; i = 1,2,\ldots,n$

$E(\varepsilon_i \varepsilon_j) = 0 \quad ; i \neq j$

$V(\varepsilon_i) = \sigma^2 \quad ; \sigma^2 > 0$ และไม่ทราบค่า

3. ตัวแปรอิสระเป็นอิสระแก่กันและอิสระจากความคลาดเคลื่อน

4. ตัวแปรอิสระแต่ละตัวเป็นค่าคงที่ที่เป็นอิสระซึ่งกันและกัน

5. จำนวนข้อมูลตัวอย่างต้องมากกว่าจำนวนตัวแปรอิสระ ($n > p$)

6. ในการวิจัยครั้งนี้จะใช้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยเป็นดัชนีสำคัญที่จะใช้เป็นเกณฑ์ในการเลือกตัวประมาณ โดยจะเลือกตัวประมาณที่มีความคลาดเคลื่อนกำลังสองเฉลี่ยต่ำสุด

1.4 ขอบเขตของการวิจัย

1. ศึกษาวิธีการประมาณค่าสัมประสิทธิ์ในสมการถดถอยเชิงเส้นและหาค่าความแปรปรวนของตัวประมาณโดยวิธีกำลังสองต่ำสุดและวิธีบูตสแตรป

2. ศึกษาคุณสมบัติของตัวประมาณที่ได้จาก 2 วิธีดังกล่าว

3. ข้อมูลที่ใช้ในการศึกษาได้จากการจำลอง(simulaton) โดยใช้เทคนิคมอนติคาร์โลจากเครื่องคอมพิวเตอร์ VAX-11/750

4. ขนาดตัวอย่างที่ใช้มี 3 ชุด ดังนี้

- ชุดที่ 1 (n_1) = 50 จำนวนตัวแปรอิสระ(p_1) = 5
- ชุดที่ 2 (n_2) = 10 จำนวนตัวแปรอิสระ(p_2) = 4
- ชุดที่ 3 (n_3) = 5 จำนวนตัวแปรอิสระ(p_3) = 3

5. ค่าคงที่ X แต่ละตัวมีการแจกแจงแบบปกติที่มีค่าเฉลี่ยและความแปรปรวนดังนี้

ค่าคงที่	ค่าเฉลี่ย	ความแปรปรวน
X_1	1.00	0.00
X_2	15.00	20.00
X_3	30.00	100.00
X_4	60.00	150.00
X_5	100.00	200.00

เมื่อ X_i แทน ค่าคงที่ตัวที่ i ; i = 1,2,3,4,5

<u>หมายเหตุ</u> ในการสร้างค่าคงที่ X ให้มีการแจกแจงแบบปกติ โดยมีค่าเฉลี่ยและความแปรปรวนดังกล่าว เพื่อให้ได้ค่าที่เป็นธรรมชาติและมีค่าห่างกันพอควร

6. ค่าพารามิเตอร์ซึ่งเป็นค่าสัมประสิทธิ์ความถดถอยที่ใช้มีค่าดังนี้

$\beta_1 = 10.0$

$\beta_2 = 1.5$

$\beta_3 = 2.0$

$\beta_4 = 2.5$

$\beta_5 = 3.0$

เมื่อ β_i แทนค่าสัมประสิทธิ์ความถดถอยตัวที่ i ; i = 1,2,...,n

7. ลักษณะการแจกแจงของความคลาดเคลื่อนที่ศึกษามีดังนี้

- การแจกแจงแบบปกติ (Normal Distribution)
- การแจกแจงแบบยูนิฟอร์ม (Uniform Distribution)
- การแจกแจงแบบโลจิสติค (Logistic Distribution)

- การแจกแจงแบบดับเบิ้ลเอ็กซ์โปเนนเชียล (Double Exponential Distribution)
- การแจกแจงแบบปกติปลอมปน (Scale Contaminated Normal Distribution) ซึ่งจะศึกษาที่เปอร์เซ็นต์การปลอมปนเป็น 1% 5% 10% และ 25% สำหรับสเกลแฟคเตอร์ 2 ระดับ คือ 3 และ 10

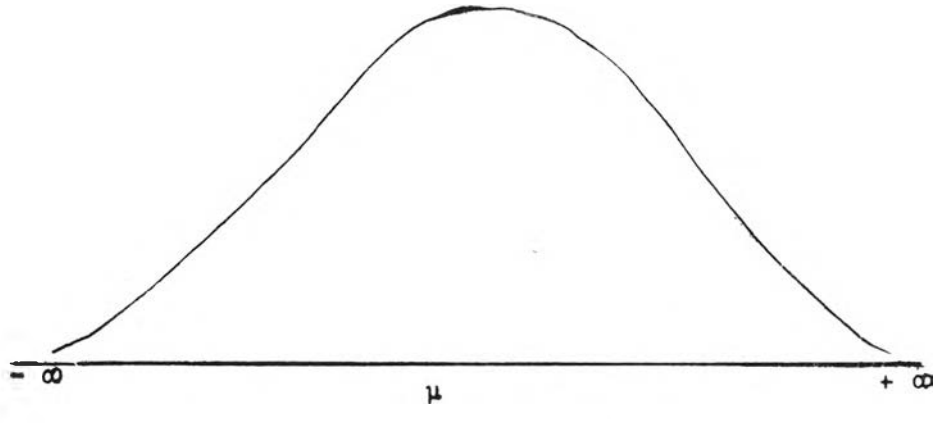
8. ทุกลักษณะการแจกแจงจะมีค่าเฉลี่ยของความคลาดเคลื่อนเป็น 0 นั่นคือ คือ $E(\varepsilon_i) = 0 \quad ; \quad i = 1,2,...n$ และค่าความแปรปรวน (6^2) เป็น 100

9. แต่ละลักษณะการแจกแจงจะมีค่าฟังก์ชั่นความน่าจะเป็น ค่าคาดหวัง ค่าความแปรปรวน ดังนี้

9.1 แจกแจงแบบปกติ (Normal Distribution)

ฟังก์ชั่นความน่าจะเป็น คือ

$$f(x) = \frac{1}{\sqrt{2\pi 6^2}} e^{-\frac{1}{26^2}(x-\mu)^2} \quad ; -\infty < x < \infty$$



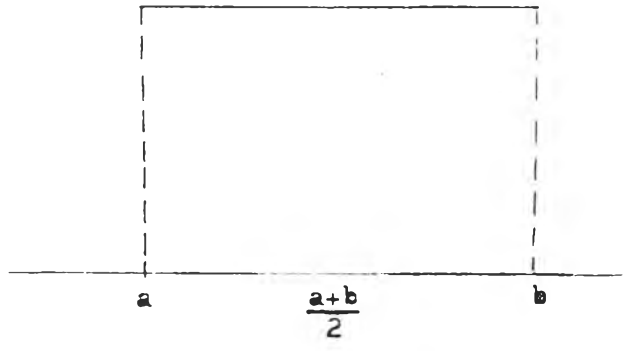
ค่าคาดหวัง $E(x) = \mu$

ค่าความแปรปรวน $V(x) = 6^2$

9.2 การแจกแจงแบบยูนิฟอร์ม (Uniform Distribution)

ฟังก์ชั่นความน่าจะเป็น คือ

$$f(x) = \frac{1}{b-a} \quad ; \quad a < x < b$$



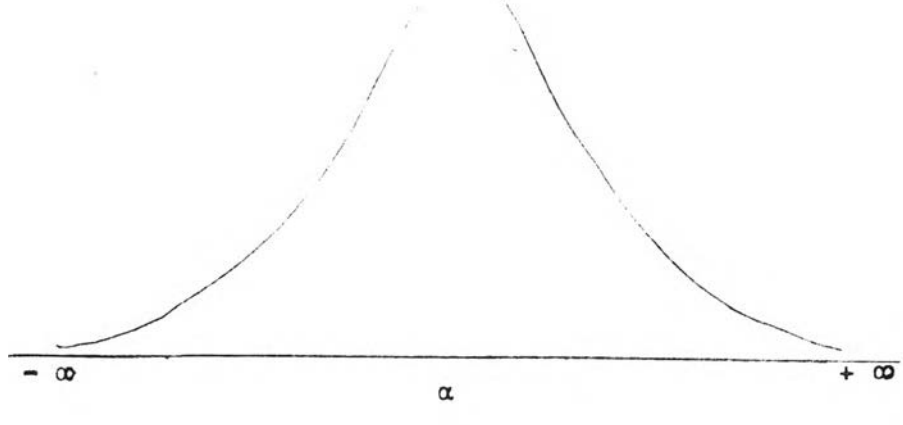
ค่าคาดหวัง $E(x) = \frac{(a + b)}{2}$

ค่าความแปรปรวน $V(x) = \frac{(b - a)^2}{12}$

9.3 การแจกแจงแบบโลจิสติค (Logistic Distribution)

ฟังก์ชั่นความน่าจะเป็น คือ

$$f(x) = \frac{1}{\beta} \cdot \frac{e^{-\left(\frac{x-\alpha}{\beta}\right)}}{\left\{1 + e^{-\left(\frac{x-\alpha}{\beta}\right)}\right\}^2} \quad ; \quad -\infty < x < \infty$$



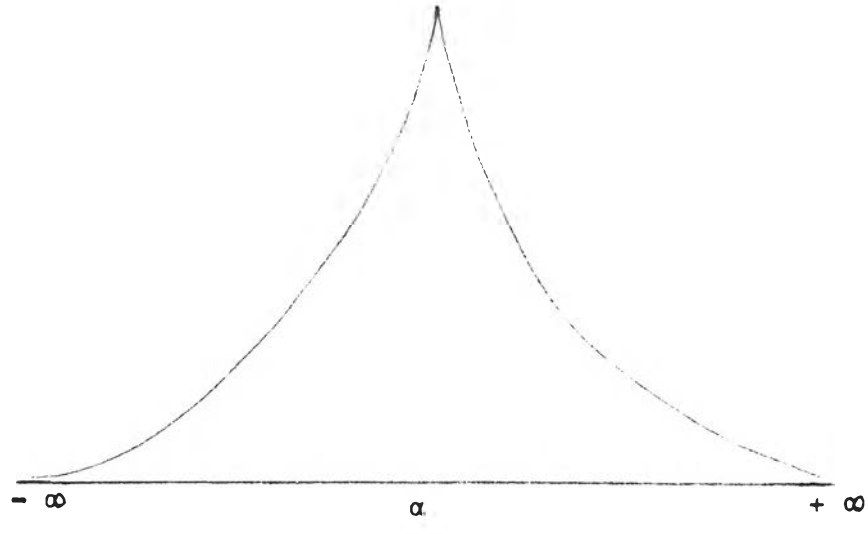
ค่าคาดหวัง $E(x) = \alpha$

ค่าความแปรปรวน $V(x) = \frac{1}{3}\pi^2\beta^2$

9.4 การแจกแจงแบบดับเบิ้ลเอ็กซ์โปเนเชียล (Double Exponential Distribution)

ฟังก์ชั่นความน่าจะเป็น คือ

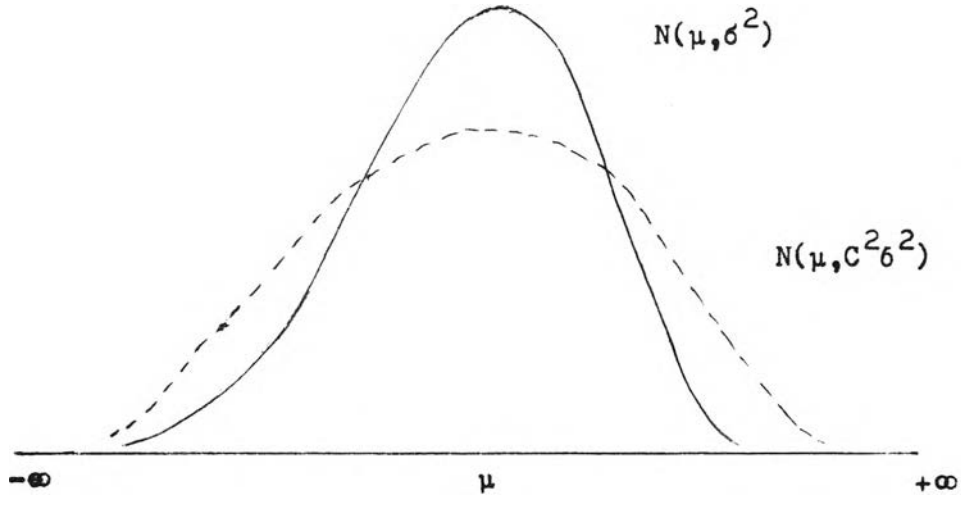
$$f(x) = \frac{1}{2\beta} \cdot e^{-\left|\frac{x-\alpha}{\beta}\right|} \quad ; -\infty < x < \infty$$



ค่าคาดหวัง $E(x) = \alpha$

ค่าความแรปรวน $V(x) = 2\beta^2$

9.5 การแจกแจงแบบปกติปลอมปน (Scale Contaminated Normal Distribution)



ลักษณะการแจกแจงแบบปกติปลอมปน เป็นการแจกแจงที่แปลงมาจากการแจกแจงแบบปกติ ซึ่งมีฟังก์ชั่นการแปลงเป็นดังนี้

$$F = (1-p)\, N(\mu,\sigma^2) + p\, N(\mu, c^2\sigma^2), \quad c > 0$$

หมายความว่าค่า x จะมาจากการแจกแจงแบบ $N(\mu, \sigma^2)$ ด้วยความน่าจะเป็น 1-p และจากการแจกแจงแบบ $N(\mu, c^2\sigma^2)$ ด้วยความน่าจะเป็น p

เมื่อ μ และ σ^2 เป็นพารามิเตอร์ที่กำหนดค่าเฉลี่ย และค่าความแปรปรวนของความคลาดเคลื่อน

p และ c เป็นค่าคงที่ (Fixed constant) ที่กำหนดเปอร์เซ็นต์การปลอมปนและสเกลแฟคเตอร์

10. การจำลองในแต่ละการทดลองกระทำซ้ำ 200 ครั้ง

11. การสุ่มตัวอย่างแบบใส่คืน (with replacement) ในวิธีบูตสแตรปกระทำ 100 ครั้ง

1.5 คำจำกัดความ

1. ความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error) หรือ MSE ของตัวประมาณ คือ ถ้า $\hat{\theta}$ เป็นตัวประมาณของพารามิเตอร์ θ แล้ว ความคลาดเคลื่อนกำลังสองเฉลี่ยของ θ คือ $E(\hat{\theta}-\theta)^2$

ในการคำนวณหาค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของฟังก์ชั่นเชิงเส้นของตัวประมาณ θ ในที่นี้คือ $l'\hat{\beta}$ เมื่อ l เป็นเมตริกซ์ขนาด $p \times 1$ ที่สมาชิกทุกตัวมีค่าเป็น 1 นั่นก็คือการหาค่าความคลาดเคลื่อนกำลังสองเฉลี่ยในรูปของผลบวกของตัวประมาณนั่นเอง

2. ความแปรปรวน (Variance) ของตัวประมาณ คือ ถ้า $\hat{\theta}$ เป็นตัวประมาณของพารามิเตอร์ θ แล้ว ความแปรปรวนของ $\hat{\theta}$ คือ $E(\hat{\theta}-E(\hat{\theta}))^2$

3. ความไม่เอนเอียง (Unbiased) ของตัวประมาณ คือถ้า $\hat{\theta}$ เป็นตัวประมาณของพารามิเตอร์ θ แล้ว จะถือว่า $\hat{\theta}$ เป็นตัวประมาณที่ไม่เอนเอียงของ θ ก็ต่อเมื่อ $E(\hat{\theta}) = \theta$

4. สถิติที่มีความพอเพียง (Sufficient statistics) จะต้องมีคุณสมบัติดังนี้ คือ ถ้าให้ $x_1, x_2, \ldots, x_n$ เป็นตัวอย่างสุ่มจากประชากรที่มีฟังก์ชั่นความหนาแน่น $f(x,\theta)$, $\theta \in \Omega$ สถิติ $t = t(x_1, x_2, \ldots, x_n)$ เป็นสถิติที่มีความพอเพียง (sufficient statistics) ของพารามิเตอร์ θ ถ้าไม่ว่า $t_1, t_2, \ldots, t_{r-1}$ จะเป็นสถิติอื่นใด $r=2,3,\ldots,n$ ฟังก์ชั่นความหนาแน่นอย่างมีเงื่อนไข (conditional density function) ของ $t_1, t_2, \ldots, t_{r-1}$ เมื่อกำหนด t ให้ คือ $g(t_1, t_2, \ldots, t_{r-1}/t)$ เป็นฟังก์ชั่นที่ไม่ขึ้นอยู่กับ θ (คือไม่เป็นฟังก์ชั่นของ θ)

5. ตัวประมาณเชิงเส้นที่ดีที่สุดและไม่เอนเอียง (Best Linear Unbiased Estimator) หรือ BLUE เป็นคุณสมบัติหนึ่งของตัวประมาณ โดยตัวประมาณ θ จะมีคุณสมบัติเป็น BLUE ของพารามิเตอร์ θ ถ้า $\hat{\theta}$ มีคุณสมบัติครบ 3 ข้อต่างต่อไปนี้

5.1 เป็นฟังก์ชั่นเชิงเส้นของตัวอย่างสุ่ม

5.2 เป็นตัวประมาณที่ไม่เอนเอียง

5.3 เป็นตัวประมาณที่มีค่าความแปรปรวนต่ำสุด

6. ตัวประมาณที่มีค่าความแปรปรวนต่ำสุดเพียงตัวเดียวในบรรดาตัวประมาณที่ไม่เอนเอียง (Uniformly Minimum Variance Unbiased Estimator) หรือ UMVUE เป็นคุณสมบัติของตัวประมาณที่ผู้วิจัยต้องการ θ คือถ้ามีตัวประมาณที่ไม่เอนเอียงสำหรับ ซึ่งเป็นสถิติที่พอเพียงสำหรับ θ และมีค่าความแปรปรวนต่ำกว่าค่าความแปรปรวนของตัวประมาณอื่น ๆ ที่ไม่เอนเอียงสำหรับ θ แล้ว ตัวประมาณดังกล่าวจะมีคุณสมบัติเป็นตัวประมาณที่ดีที่สุดหรือมีค่าความแปรปรวนต่ำสุดเพียงตัวเดียวในบรรดาตัวประมาณที่ไม่เอนเอียง

7. ประสิทธิภาพสัมพัทธ์ (Relative Efficiency) หรือ RE ของตัวประมาณ θ_1 เมื่อเทียบกับ θ_2 เป็นการเปรียบเทียบค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวประมาณ 2 ตัว ที่ประมาณค่าพารามิเตอร์เดียวกัน ในรูปอัตราส่วนของค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของตัวประมาณหนึ่ง ต่อค่าความคลาดเคลื่อนกำลังสองเฉลี่ยของอีกตัวหนึ่ง หรืออาจจะพิจารณาจากสูตร

$$RE(\theta_1, \theta_2) = \frac{MSE(\theta_2)}{MSE(\theta_1)}$$

1.6 ประโยชน์ของการวิจัย

เพื่อช่วยให้ข้อสรุปในการเลือกใช้วิธีการประมาณค่าสัมประสิทธิ์ในสมการถดถอยเชิงเส้นได้อย่างมีประสิทธิภาพ เมื่อความคลาดเคลื่อนไม่เป็นไปตามข้อตกลงเบื้องต้น