บทที่ 1

บทนำ

1.1 ความสำคัญและความเป็นมาของปัญหา

ในการศึกษาการเปลี่ยนแปลงของตัวแปรตาม (dependent variable) ว่ามีผลมาจากตัวแปรอิสระ (indepenent variables) ชุดหนึ่งอย่างไรนั้น วิธีหนึ่งของการวิเคราะห์ความสำคัญของตัวแปรเหล่านั้นก็คือ การวิเคราะห์ความถดถอยพหุ (multiple regression analysis) ซึ่งเป็นกรณีหนึ่งของการวิเคราะห์ความถดถอยเชิงเส้นในการวิเคราะห์ความถดถอยพหุถือว่าการใช้ตัวแปรอิสระที่เหมาะสมมากกว่า 1 ตัว โดยทั่วๆไปย่อมทำให้ผลการประมาณค่าตัวแปรตามมีความถูกต้องมากกว่าการใช้ตัวแปรอิสระเพียงตัวเดียวสำหรับตัวแบบทั่วไป (general model) ของความสัมพันธ์ระหว่างตัวแปรอิสระ และ ตัวแปรตามเชิงเส้น ซึ่งเราสามารถเขียนสมการได้ดังนี้

(1.1) $$\underset{\sim}{y} = X \underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

เมื่อ $\underset{\sim}{y}$ เป็นเวกเตอร์ของตัวแปรตามขนาด $n \times 1$ โดยที่ n คือ จำนวนค่าสังเกต

X เป็นเมตริกซ์ของตัวแปรอิสระขนาด $n \times (p+1)$ โดยที่ p คือจำนวนตัวแปรอิสระ และ $\underset{\sim}{\varepsilon}$ เป็นเวกเตอร์ของค่าความคลาดเคลื่อนที่เกิดขึ้นซึ่งมีขนาด $n \times 1$ และมีการแจกแจงแบบปกติ โดยที่ $E(\underset{\sim}{\varepsilon}) = 0$ และ $Cov(\underset{\sim}{\varepsilon}) = \sigma^2 I_n$

วิธีที่นิยมมากที่สุดวิธีหนึ่งในการประมาณค่า ส.ป.ส.การถดถอยพหุจากรูปแบบดังกล่าวนี้ คือ วิธีกำลังสองน้อยที่สุด (least square method) ซึ่งจะได้ตัวประมาณของ $\underset{\sim}{\beta}$ อยู่ในรูปของ

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

$\hat{\underset{\sim}{\beta}}$ จะเป็นตัวประมาณที่ไม่เอนเอียงและให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำสุด ในการประมาณค่าส.ป.ส.การถดถอยพหุด้วยวิธีกำลังสองน้อยที่สุดมีสมมติฐานที่จำเป็นข้อหนึ่ง คือ ตัวแปรอิสระต้องไม่มีความสัมพันธ์กันในลักษณะเชิงเส้น ซึ่งในทางปฏิบัติเป็นไปได้น้อยมากเพราะตัวแปรอิสระบางตัวที่นำมาศึกษาอาจมีความสัมพันธ์กันอยู่โดยตัวแปรอิสระบางตัวอาจเป็นฟังก์ชันของตัวแปรอิสระตัวอื่น นั่นคือตัวแปรอิสระมีพหุสัมพันธ์ (multicolinearity)

ทำให้การประมาณค่าตัวแปรตามที่ได้อาจไม่เหมาะสม และมีผลทำให้ค่าเฉลี่ยความคลาด-เคลื่อนกำลังสองของค่าประมาณส.ป.ส.การถดถอยพหุมีค่ามากกล่าวคือค่าประมาณ ส.ป.ส. การถดถอยพหุที่ได้ขาดความเที่ยงตรง (accurancy)

วิธีการแก้ปัญหาตัวแปรอิสระมีพหุสัมพันธ์สามารถกระทำได้หลายวิธี วิธีที่น่าสนใจวิธีหนึ่งคือ วิธีริดจ์รีเกรสชัน (ridge regression method) ซึ่ง Hoerl and Kennard (1970 : 55-67) ได้เสนอวิธีริดจ์รีเกรสชัน โดยที่วิธีนี้จะให้ค่าประมาณสัมประสิทธิ์การถดถอยพหุที่ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำสุด โดยจะบวกค่าคงที่ค่าหนึ่งที่มากกว่าศูนย์กับสมาชิกทุกตัวบนเส้นทแยงมุมของเมตริกซ์ X′X เนื่องจากค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของวิธีกำลังสองน้อยที่สุดเป็นฟังก์ชันของ (X′X) ดังนั้น การที่จะพยายามลดค่าเฉลี่ยความคลาดเคลื่อนกำลังสองให้ต่ำลงจึงต้องพยายามลดค่า (X′X) ให้ต่ำลง โดยการบวกค่าคงที่ที่มากกว่าศูนย์กับสมาชิกทุกตัวบนเส้นทแยงมุมดังที่กล่าวข้างต้นวิธีนี้ไม่ต้องตัดตัวแปรอิสระออกจากตัวแบบถึงแม้จะเกิดพหุสัมพันธ์ในระหว่างตัวแปรอิสระ แต่ค่าประมาณสัมประสิทธิ์ความถดถอยพหุที่ได้จากริดจ์รีเกรสชันจะเอนเอียง (bias) จากวิธีกำลังน้อยที่สุดจะได้ $\hat{\beta}$ เป็นค่าประมาณของ β ซึ่ง

$$\hat{\beta} = (X'X)^{-1} X'\underline{y}$$

และ $$E(\hat{\beta}'\hat{\beta}) = \beta'\beta + \sigma^2 \,\text{trace}\, (X'X)^{-1}$$

(1.2) $$= \beta'\beta + \sigma^2 \sum (1/\lambda_i)$$

จะเห็นได้ว่าการที่จะทำให้ความยาวของ $\hat{\beta}'\hat{\beta}$ ลดลงก็โดยการเพิ่มค่า eigenvalue (λ) ของเมตริกซ์ (X′X) ให้มากขึ้น ซึ่งทำได้โดยการบวกค่าคงที่ k ซึ่งมีค่ามากกว่า 0 กับค่าบนเส้นทแยงมุมของเมตริกซ์ (X′X) ดังที่ได้กล่าวแล้วข้างต้น

$$\text{trace}(X'X)^{-1} = \sum (1/\lambda_i)$$

ดังนั้น $$\text{trace}(X'X + kI)^{-1} = \sum \lambda_i \,/(\lambda_i+k)^2$$

และ $$\begin{aligned} MSE(\hat{\beta}^*(k)) &= \text{Variance} + \text{Bias}^2 \\ &= \sigma^2 \sum \lambda_i/(\lambda_i+k)^2 + k^2\beta'(X'X + kI)^{-2}\beta \\ &= \sigma^2 \sum \lambda_i/(\lambda_i+k)^2 + k^2 \sum \hat{\beta}_i^{\,2}/(\lambda_i+k)^2 \end{aligned}$$

เมื่อค่าประมาณ ส.ป.ส.การถดถอยพหุโดย Ridge Regression คือ

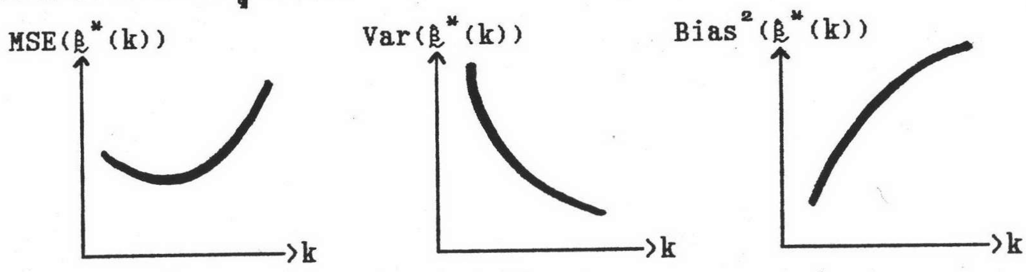
(1.3) $$\hat{\beta}^* = (X'X + kI)^{-1} X'\underline{y} \quad , \quad k > 0$$

ข้อสังเกต เพราะว่า MSE($\hat{\beta}^*$(k)) เป็นค่าเชิงปริมาณ จะได้ว่าค่าความแปรปรวน จะมีค่าลดลงเมื่อ ค่า k มีค่าสูงขึ้น แต่ค่าความเอนเอียงจะมีค่าสูงขึ้นเมื่อ k

เพิ่มขึ้นดังนั้นปัญหาสำคัญของวิธีนี้คือ เราจะประมาณค่า k ที่เหมาะสมได้อย่างไร ซึ่งมีวิธีการประมาณค่า k ด้วยกันหลายวิธี เช่น วิธี HKB (Hoerl,A.E.,Kennard, R.W. [2] and Baldwin,K.F.)(1975),วิธี Tze-San-Lee (Tze-San Lee Method) (1986) เป็นต้น

วิธีการประมาณค่า k เหล่านั้นอาจให้ผลของการประมาณค่า k ที่แตกต่างและค่า k ที่คำนวณได้ก็ยังอาจจะไม่เป็นค่า k ที่เหมาะสมที่สุด นอกจากนั้นวิธีการคำนวณค่อนข้างจะยุ่งยากซับซ้อน ดังนั้นจึงเป็นเรื่องที่น่าสนใจว่าจะมีวิธีประมาณค่า k ที่ให้ผลใกล้เคียงกับค่า k ที่เหมาะสมที่สุด โดยที่วิธีการคำนวณจะไม่ยุ่งยากเกินไป ในงานวิจัยนี้ผู้วิจัยขอเสนอวิธีในการหาค่า k ที่เหมาะสมซึ่งเรียกว่าวิธี "Binary Search" ดังมีรายละเอียดต่อไปนี้

"Binary Search" เป็นวิธีการหนึ่งในทางคอมพิวเตอร์ซึ่งใช้ในการค้นหาข้อมูลที่มีลักษณะเรียงลำดับ วิธีนี้จะช่วยทำให้การค้นหาข้อมูลเป็นไปอย่างมีประสิทธิภาพและรวดเร็ว ซึ่งในการวิจัยนี้พบว่าความสัมพันธ์ระหว่างค่า MSE($\beta^*(k)$) กับค่า k เมื่อนำมาเขียนกราฟแล้วจะมีลักษณะเป็นรูปโค้งหงาย(convere) กล่าวคือ เมื่อค่า k เพิ่มขึ้นค่า MSE($\beta^*(k)$) ซึ่งเป็นค่าเชิงปริมาณจะมีค่าลดลงตามลำดับซึ่งจะลดลงจนถึงค่า k ค่าหนึ่ง กล่าวคือเป็นค่า k ที่เหมาะสมซึ่งทำให้ MSE($\beta^*(k)$) มีค่าต่ำสุด แต่ถ้าค่า k มีค่ามากกว่าค่า k ที่เหมาะสมนั้นพบว่าค่า MSE($\beta^*(k)$) จะมีค่าเพิ่มขึ้นตามลำดับเมื่อ k เพิ่มขึ้น สาเหตุอาจเนื่องมาจากเมื่อ k มีค่าเพิ่มขึ้นทำให้ค่าความแปรปรวนลดลงแต่ค่าเอนเอียงมีค่าเพิ่มขึ้น ซึ่งอัตราการเพิ่มนี้มากกว่าอัตราการลดลงของค่าความแปรปรวนเมื่อค่า k เกินค่า k ที่เหมาะสมดังรูปต่อไปนี้



รูปที่ 1 รูปที่ 2 รูปที่ 3

โดย รูปที่ 1 แสดงความสัมพันธ์ระหว่างค่า k กับค่า MSE($\beta^*(k)$)
รูปที่ 2 แสดงความสัมพันธ์ระหว่างค่า k กับค่า Var($\beta^*(k)$)
รูปที่ 3 แสดงความสัมพันธ์ระหว่างค่า k กับค่า Bias^2($\beta^*(k)$)

สำหรับวิธี Binary Search ที่นำมาประยุกต์ใช้มีขั้นตอนดังนี้

<u>วิธี Binary Search</u>

กำหนด k(opt) คือ ค่า k ที่ทำให้ MSE($\hat{\beta}^*$(k)) มีค่าต่ำสุด

k(max) คือ ค่า k ที่มีค่ามากกว่า k(opt) ที่เล็กที่สุด

k(min) คือ ค่า k ที่มีค่าน้อยกว่า k(opt) ที่ใหญ่ที่สุด

C คือ ค่าคงที่ที่กำหนดขึ้นเพื่อเปรียบเทียบค่าผลต่างระหว่าง k(max) กับ k(min) เพื่อใช้เป็นเงื่อนไขยุติการประมวลผล ในที่นี้กำหนดเท่ากับ 0.0001

<u>ขั้นที่ 1</u> กำหนดค่าเริ่มต้น k(max) = 1 สำหรับการแจกแจงปกติ และปกติปลอมปน และ กำหนดค่าเริ่มต้น k(max) = 100 สำหรับการแจกแจงแบบเบ้ และ k(min) = 0

<u>ขั้นที่ 2</u> กำหนดให้

k(opt) = (k(max) + k(min)) / 2

และทำการวิเคราะห์ค่า MSE($\hat{\beta}^*$(k)) ที่ k เท่ากับ k(opt)-C , k(opt) , k(opt)+C

กรณีที่ 2.1 ถ้าค่า MSE($\hat{\beta}^*$(k)) ที่ k เท่ากับ k(opt)-C , k(opt) , k(opt)+C มีค่าเพิ่มขึ้นตามลำดับ นั่นคือค่า k(opt) ที่คำนวณได้มีค่ามากกว่า k ที่เหมาะสม ดังนั้นต้องทำการคำนวณหาค่า k(opt) ใหม่และปรับปรุงค่า k(max) โดยให้

k(max) = k(opt)

และทำการวิเคราะห์ขั้นที่ 2 ใหม่จนกว่าจะเข้ากรณีที่ 2.3 หรือ 2.4

กรณีที่ 2.2 ถ้าค่า MSE($\hat{\beta}^*$(k)) ที่ k เท่ากับ k(opt)-C , k(opt) , k(opt)+C มีค่าลดลงตามลำดับ นั่นคือค่า k(opt) ที่คำนวณได้มีค่าต่ำกว่า k ที่เหมาะสม ดังนั้นต้องทำการคำนวณหาค่า k(opt) ใหม่และปรับปรุงค่า k(min) โดยให้

k(min) = k(opt)

และทำการวิเคราะห์ขั้นที่ 2 ใหม่จนกว่าจะเข้ากรณีที่ 2.3 หรือ 2.4

กรณีที่ 2.3 ถ้าค่า MSE($\hat{\beta}^*$(k)) ที่ค่า k เท่ากับ k(opt) ให้ค่าต่ำกว่าที่ k(opt)-C และ k(opt)+C นั่นคือค่า k(opt) ที่คำนวณได้จะเป็นค่า k ที่เหมาะสม จะ

ยุติการประมวลผล

กรณีที่ 2.4 ถ้าผลต่างของ k(max) กับ k(min) มีค่าน้อยกว่าค่าคงที่ C ที่กำหนดขึ้น จะยุติการประมวลผล โดยถือว่า k(opt) ที่คำนวณได้เป็นค่าประมาณของค่า k ที่เหมาะสม

ค่า C ที่กำหนด	ความคลาดเคลื่อนของค่า k	จำนวนรอบที่ประมวลผลสำหรับ	
		ปกติ และ ปกติปลอมปน	แบบเบ้
0.1	|0.05|	4	11
0.01	|0.005|	7	14
0.001	|0.0005|	10	17
0.0001	|0.00005|	14	21

ถ้าค่าสังเกตสุ่มมาจากประชากรที่มีการแจกแจงแบบเบ้ วิธีการประมาณค่า k ที่เหมาะสมอาจแตกต่างไปจากกรณีที่ค่าสังเกตของประชากรมีการแจกแจงแบบปกติ และ ปกติปลอมปน เนื่องจากวิธีริดจ์รีเกรสชันใช้หลักการคำนวณเหมือนกับวิธีกำลังสองน้อยที่สุดและวิธีกำลังสองน้อยที่สุดเป็นวิธีที่ไวต่อข้อมูลที่ผิดปกติ ดังนั้นจึงเป็นสิ่งน่าสนใจที่จะศึกษาวิธีการประมาณค่า k ในวิธีการของริดจ์รีเกรสชันว่าวิธีใดที่ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของตัวประมาณสัมประสิทธิ์การถดถอยพหุน้อยที่สุด และข้อมูลต่างๆที่ได้มีลักษณะอย่างไร

1.2 วัตถุประสงค์ของการวิจัย

การวิจัยนี้มีวัตถุประสงค์ที่จะศึกษาเปรียบเทียบวิธีประมาณตัวประมาณริดจ์ (Ridge estimators) ต่างๆในการวิเคราะห์ความถดถอยแบบริดจ์ ซึ่งในที่นี้มีด้วยกัน 4 วิธี

1. วิธี HKB
2. วิธี Tze-San Lee

3. วิธี Hoerl and Kennard

4. วิธีประมาณแบบทวิ (Binary Search)

โดยศึกษาเปรียบเทียบตัวประมาณบริดจ์ เมื่อการแจกแจงของค่าความคลาดเคลื่อนเป็นแบบปกติ ปกติปลอมปนและแบบเบ้

1.3 สมมติฐานของการวิจัย

เมื่อความคลาดเคลื่อนมีการแจกแจงปกติ ปกติปลอมปน และแบบเบ้ วิธีการประมาณค่า k ด้วยวิธี "Binary Search" จะให้ผลดีที่สุดภายใต้จำนวนตัวแปรอิสระ ขนาดตัวอย่าง ระดับของตัวแปรอิสระมีพหุสัมพันธ์ (degree of multicollinearity) ระดับของส.ป.ส.ความแปรผัน (coefficient of variation) และชุดของส.ป.ส. การถดถอย โดยใช้ข้อกำหนดเดียวกัน

1.4 ขอบเขตของการวิจัย

4.1 <u>เมื่อความคลาดเคลื่อนมีการแจกแจงปกติ</u> โดยการแจกแจงความผิดพลาดเป็นไปตามข้อตกลงของสมการถดถอยคือ $\underset{\sim}{\varepsilon} \sim N(\underset{\sim}{1},\sigma^2 I_n)$

4.1.1 จำนวนตัวแปรอิสระที่ใช้ศึกษามี 2 ระดับ คือ 3,5 ตามลำดับ

4.1.2 ขนาดตัวอย่างที่ศึกษามี 3 ขนาดคือ 30,50,100

4.1.3 ระดับของส.ป.ส.ความแปรผัน (C.V.) ที่ศึกษามี 3 ระดับ คือ 5%,10% และ 15%

สาเหตุที่ศึกษาใน 3 ระดับนี้เพราะในกรณีที่ค่า C.V. สูงความคลาดเคลื่อนที่ศึกษาก็มีการแจกแจงแบบอื่นที่ไม่ใช่การแจกแจงปกติ

4.1.4 ก) กรณีตัวแปรอิสระเท่ากับ 3 ระดับของพหุสัมพันธ์ของตัวแปรอิสระที่ศึกษามีด้วยกัน 3 ระดับ

คือ ระดับรุนแรงมาก ρ = (0.99)

ระดับรุนแรง ρ = (0.90)

ระดับปานกลาง ρ = (0.70)

โดยที่ ρ คือความสัมพันธ์ระหว่าง X1 กับ X2 , X1 กับ X3 และ X2 กับ X3

ข) กรณีตัวแปรอิสระเท่ากับ 5 ระดับของพหุสัมพันธ์ของตัวแปรอิสระที่ศึกษามีด้วยกัน 3 ระดับ

คือ ระดับรุนแรงมาก ρ = (0.99),(0.99),(0.99),(0.99)

ระดับรุนแรง ρ = (0.90),(0.90),(0.90),(0.90)

ระดับปานกลาง ρ = (0.70),(0.70),(0.70),(0.30)

โดยที่ ρ คือความสัมพันธ์ระหว่าง X1 กับ X2 , X1 กับ X3 , X2 กับ X3 และ X4 กับ X5

4.1.5 ค่าส.ป.ส.การถดถอยที่ใช้ศึกษามี 2 ชุด คือ

ก) ค่า eigenvector ซึ่งสอดคล้องกับ eigenvalue ที่มีค่ามากที่สุด ซึ่งทำให้ $E[L(R)] = E[(\hat{\beta}_r-\beta)'(\hat{\beta}_r-\beta)]$ มีค่าน้อยที่สุด

ข) ค่า eigenvector ซึ่งสอดคล้องกับ eigenvalue ที่มีค่าต่ำที่สุด ซึ่งทำให้ $E[L(R)] = E[(\hat{\beta}_r-\beta)'(\hat{\beta}_r-\beta)]$ มีค่ามากที่สุด

4.2 <u>เมื่อความคลาดเคลื่อนมีการแจกแจงแบบปกติปลอมปน</u>

(Scale Contaminated Normal Distribution)

ฟังก์ชันการแจกแจงอยู่ในรูปของ

$$F = (1-p)\, N(\underset{\sim}{1},\sigma^2) + N(\underset{\sim}{1},c^2\sigma^2)$$

เมื่อ c คือ สเกลแฟคเตอร์ (Scale factor) ถ้าสเกลแฟคเตอร์มีค่าสูงจะทำให้เกิดค่าสังเกตที่ผิดปกติมีค่าสูงด้วย ในที่นี้จะใช้ c = 3 และ c = 10

และ p คือ เปอร์เซนต์การปลอมปน (percent of contamination) ในกรณีนี้จะใช้ p = 5,10

4.3 <u>เมื่อความคลาดเคลื่อนมีการแจกแจงแบบเบ้</u>

ผู้วิจัยจะศึกษากรณีการแจกแจงลอกนอร์มอล (Lognomal Distribution) โดยที่ฟังก์ชันความหนาแน่นของการแจกแจงดังกล่าวอยู่ในรูปของ

$$f(x) = \begin{cases} (1/x\sigma\sqrt{2\pi})\exp(-[1/2(\ln(x) - \mu]/\sigma^2) & ;\ x>0,\sigma>0 \\ 0 & ;\ \text{อื่นๆ} \end{cases}$$

เมื่อ μ และ σ^2 เป็นค่าเฉลี่ยและความแปรปรวนของ Y โดยที่ Y = ln(x) Y จะมีการแจกแจงแบบปกติ ในที่นี้จะศึกษาโดยใช้ C.V = 100% , 59% , 22% โดยกำหนดค่า μ = 1 ในการศึกษา

1.5 เกณฑ์ที่ใช้พิจารณา

ในการวิจัยครั้งนี้จะใช้ค่าอัตราส่วนผลต่างของค่าเฉลี่ยความคลาดเคลื่อนกำลังสอง (Ratio of Different Average Mean Square Error) ของการประมาณสัมประสิทธิ์การถดถอยพหุแบบริดจ์ เป็นเกณฑ์ในการพิจารณา

1.6 วิธีดำเนินการวิจัย

1. สร้างข้อมูลของตัวแปรอิสระ X1,...,Xp โดยใช้เทคนิคมอนติคาร์โลตามระดับต่างๆที่กำหนดไว้ในขอบเขตการวิจัย โดยในแต่ละกรณีจะผลิตทั้งสิ้น 200 รอบ
2. สร้างข้อมูลตัวแปรตามจากตัวแปรอิสระที่สร้างขึ้นและตามระดับพหุสัมพันธ์ที่กำหนดไว้ในขอบเขตการวิจัย
3. ประมาณค่าพารามิเตอร์ตามวิธีริดจ์รีเกรสชัน
4. คำนวณหาค่าเฉลี่ยความผิดพลาดกำลังสอง (AMSE)
   (Average Mean Square Error)
5. เปรียบเทียบค่าอัตราส่วนผลต่างของค่าเฉลี่ยความผิดพลาดกำลังสอง (Ratio of Different Average Mean Square Error) (RDAMSE)

$$\text{RDAMSE} = \frac{[\text{AMSE}(i) - \text{AMSE}(\min)]}{\text{AMSE}(\min)} \times 100$$

เมื่อ $i = 1, \ldots, 4$

6. สรุปผลและอภิปรายผล

1.7 ประโยชน์ที่คาดว่าจะได้รับ

1. ผลการศึกษาเป็นแนวทางในการศึกษาเกี่ยวกับการประมาณค่าส.ป.ส.การถดถอยพหุในกรณีตัวแปรอิสระมีพหุสัมพันธ์เมื่อความคลาดเคลื่อนมีการแจกแจงแบบปกติ ปกติปลอมปน และ แบบเบ้
2. ผลการศึกษาเปรียบเทียบสามารถบอกได้ว่าวิธีการประมาณค่าแบบใดจะให้ผลดีกว่ากัน