บทที่ 1

บทนำ

1. ความสำคัญและความเป็นมาของปัญหา

ในการประมาณค่าเพื่อคาดคะเนเหตุการณ์ล่วงหน้าหรือการพยากรณ์ ผู้วิจัยมักเลือกวิธีการวิเคราะห์ความถดถอยพหุ (Mutliple regression analysis) เมื่อตัวแปรตามที่ใช้ในการศึกษามีความสัมพันธ์กับปัจจัยอื่น ๆ เราเรียกว่าตัวแปรอิสระ มากกว่า 1 ตัว ซึ่งสามารถเขียนในรูปของตัวแบบทั่วไปของการวิเคราะห์ความถดถอยพหุเชิงเส้นดังนี้

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

เมื่อ $\underset{\sim}{y}$ เป็นเวกเตอร์ของตัวแปรตามขนาด n X 1 โดยที่ n เป็นจำนวนค่าสังเกต

X เป็นเมตริกซ์ของตัวแปรอิสระขนาด n X p (p < n) และมี full rank p

$\underset{\sim}{\beta}$ เป็นเวกเตอร์ของพารามิเตอร์ที่ไม่ทราบค่า ขนาด p X 1

และ $\underset{\sim}{\varepsilon}$ เป็นเวกเตอร์ของความผิดพลาด ขนาด n X 1

โดยมีข้อตกลงเบื้องต้น (Assumption) ของความผิดพลาดดังนี้

$$E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$$

$$E(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}') = \sigma^2 I_n$$

วิธีที่นิยมใช้ในการประมาณค่าสัมประสิทธิ์การถดถอยพหุกันมาก คือ วิธีกำลังสองน้อยที่สุด (least square method) ซึ่งตัวประมาณกำลังสองน้อยที่สุดอยู่ในรูปของ

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

โดยจะเป็นตัวประมาณที่ไม่เอนเอียง (unbiased estimator) และมีความแปรปรวนต่ำสุดในบรรดาตัวประมาณที่ไม่เอนเอียง

ในกรณีที่ค่าสุ่มตัวอย่างที่ได้มาจากประชากรที่มีการแจกแจงแบบเบ้ และการแจกแจงแบบหางยาวกว่าปกติ (skewed distribution and long tailed distribution) วิธีกำลังสองน้อยที่สุดอาจจะไม่เหมาะสม เพราะวิธีนี้มีความไวต่อข้อมูลที่ผิดปกติ (Outliers) และสูญเสียประสิทธิภาพไปเมื่อการแจกแจงของความผิดพลาดไม่เป็นแบบปกติ ซึ่งได้มีผู้ค้นหาวิธีการแก้ปัญหาที่เกิดขึ้นนี้กันหลายวิธี แต่ที่รู้จักกันดี คือ วิธีการทางนอนพาราเมตริก โดยวิธีการประมาณค่าความผิดพลาดมาตรฐาน (Standard error) ในกรณีที่ไม่ทราบลักษณะการแจกแจงของประชากรและไม่สามารถหาได้จากสูตรทั่วไป โดยใช้เทคนิคการสุ่มตัวอย่างซ้ำ (Resampling) ซึ่งมีอยู่หลายวิธี ได้แก่

The Jacknife

The Bootstrap

Half-sampling

Subsampling

Balanced repeated replication

Influence function techniques

The delta method

แต่ละวิธีมาจากแนวความคิดพื้นฐานคล้ายกัน คือ การหาค่าประมาณของค่าความผิดพลาดมาตรฐาน โดยการสุ่มซ้ำจากข้อมูลที่เก็บรวบรวมมา ซึ่งพบว่าวิธีบูตสแตรป (The Bootstrap Method) เป็นวิธีที่ให้ผลดีที่สุด เพราะว่าการหาค่าประมาณโดยวิธีนี้เป็นการประมาณค่าภาวะน่าจะเป็นสูงสุดของนอนพาราเมตริก ทำให้ตัวประมาณที่ได้เป็นตัวประมาณแบบภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimator (MLE)) และจากการศึกษาของ Bradley Eforn (1982; 27) ได้ศึกษาวิธีบูตสแตรป พบว่าสามารถนำมาใช้ในการประมาณค่าคร่าว ๆ ที่ไม่สามารถหาค่าได้โดยตรงทางพาราเมตริก เช่น ไม่ทราบลักษณะการแจกแจงของประชากรไม่ทราบลักษณะการแจกแจงของความผิดพลาด

อีกวิธีหนึ่ง คือ ตัวประมาณ M ซึ่งอยู่ในรูปของค่าน้อยที่สุดของฟังก์ชั่น ค่าผิดพลาดนี้สามารถเขียนได้ดังนี้

$$\min_{\underset{\sim}{\beta}} \sum_{i=1}^{n} \rho(\varepsilon_i/s) = \min_{\underset{\sim}{\beta}} \sum_{i=1}^{n} \rho[(y_i - \underset{\sim}{X}_i'\underset{\sim}{\beta})/s]$$

เมื่อ ε_i เป็นค่าผิดพลาดของค่าสังเกตที่ i และ s เป็นค่าที่เหมาะสมสำหรับการกระจายของ ε_i ซึ่ง Huber (1981) ได้เสนอข้อคิดเห็นเมื่อค่าสังเกตมีการกระจายแบบปลอมปน และค่าสังเกตที่ได้จากกรณีที่มีการแจกแจงแบบสมมาตร ว่าการเลือกตัวประมาณ s ที่เหมาะสมขึ้นกับ

1. ขอบเขตของ breakdown point มีขนาดใหญ่
2. มีความเอนเอียงไม่มาก
3. มีความแปรปรวนไม่มาก

ในการวิจัยครั้งนี้ ผู้วิจัยจึงสนใจที่จะศึกษาเปรียบเทียบการประมาณสัมประสิทธิ์การถดถอยของ 3 วิธี คือ วิธีกำลังสองน้อยที่สุด วิธีบูตสแตรป และวิธีตัวประมาณชนิด M ภายใต้สภาวการณ์ของลักษณะข้อมูลที่มีการแจกแจงแบบต่าง ๆ โดยใช้ค่าเฉลี่ยความผิดพลาดกำลังสอง (MSE) ของตัวประมาณแต่ละตัวเป็นตัวเปรียบเทียบ

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาวิธีการประมาณค่าสัมประสิทธิ์การถดถอยพหุ และคุณสมบัติของตัวประมาณของวิธีกำลังสองน้อยที่สุดวิธีบูตสแตรป และวิธีตัวประมาณชนิด M เมื่อค่าความผิดพลาดมีการแจกแจงแบบต่างๆ
2. เพื่อศึกษาวิธีการประมาณค่าสัมประสิทธิ์การถดถอยพหุวิธีกำลังสองน้อยที่สุด วิธีบูตสแตรป และวิธีตัวประมาณชนิด M เมื่อกำหนดขนาดของตัวอย่างต่างๆ กัน
3. เพื่อศึกษาเปรียบเทียบประสิทธิภาพของการประมาณค่าสัมประสิทธิ์การถดถอยพหุด้วยวิธีกำลังสองน้อยที่สุด วิธีบูตสแตรป และวิธีตัวประมาณชนิด M

1.3 ข้อตกลงเบื้องต้น

1. ค่าความผิดพลาด (ε_i) เป็นตัวแปรสุ่มที่มีการแจกแจงเหมือนกัน และเป็นอิสระซึ่งกันและกัน

2. การวิจัยครั้งนี้ใช้เกณฑ์การประมาณค่าสัมประสิทธิ์การถดถอยพหุภายใต้ค่าความผิดพลาดที่มีการแจกแจงแบบต่าง ๆ จากวิธีใดที่ให้ค่าเฉลี่ยความผิดพลาดกำลังสอง (Mean Square Error) น้อยที่สุด จะเป็นวิธีที่เหมาะสมของแต่ละสถานการณ์

1.4 ขอบเขตการวิจัย

1.4.1 ลักษณะการแจกแจงของความผิดพลาดที่ศึกษามีดังนี้

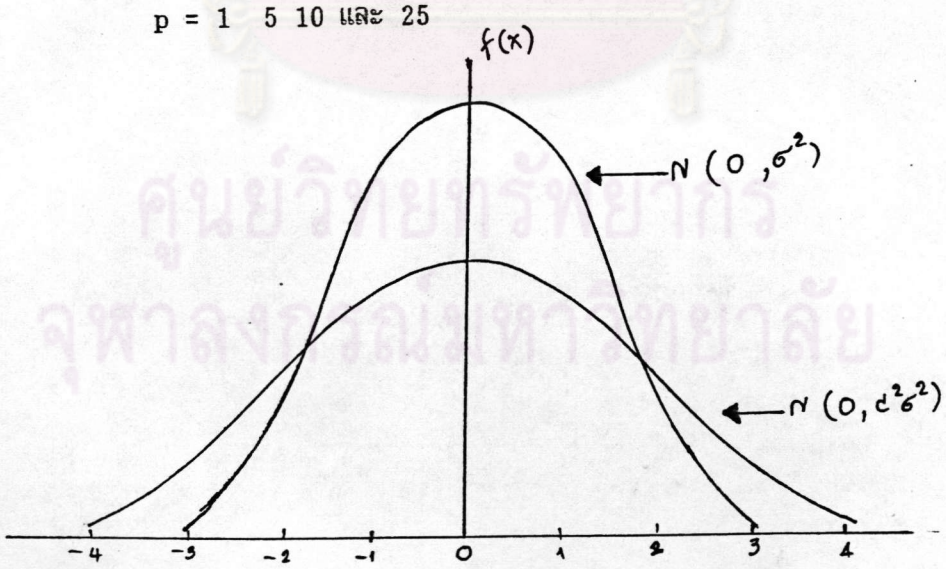
1.4.1.1 การแจกแจงแบบปกติปลอมปน (scale-contaminated normal distribution)

ฟังก์ชันการแจกแจงอยู่ในรูปของ

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

เมื่อ c คือสเกลแฟกเตอร์ (scale factor) ซึ่งถ้าค่าสูงจะทำให้ค่าสังเกตที่ผิดปกติมีค่าสูงด้วย ในที่นี้ใช้ c = 3 และ 10

และ p คือเปอร์เซนต์การปลอมปน (percent of contamination) ในที่นี้ใช้ p = 1 5 10 และ 25

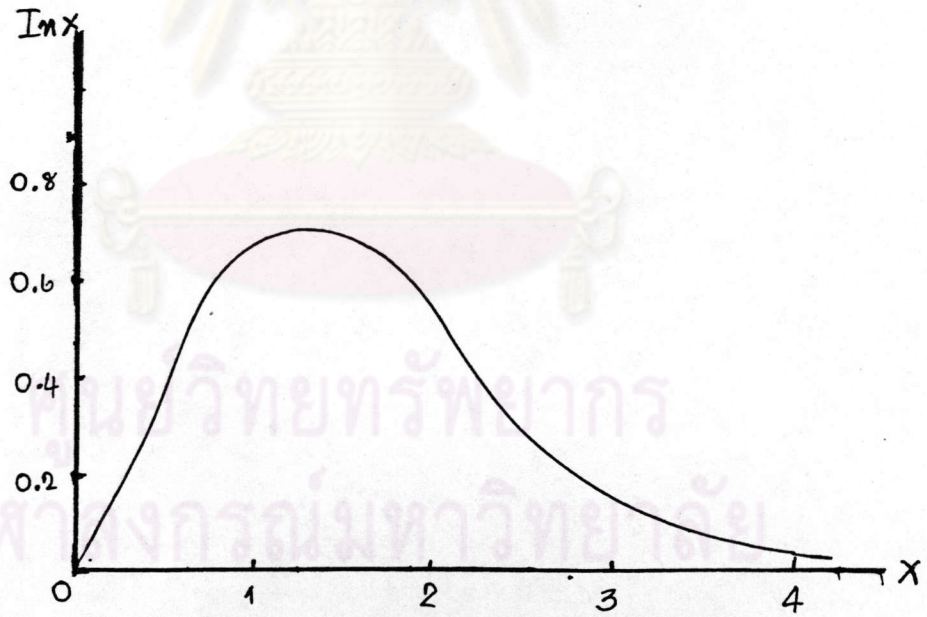


รูปที่ 1.4.1 แสดงเส้นโค้งของการแจกแจงแบบปกติปลอมปน ณ p และ c

1.4.1.2 การแจกแจงแบบลอกนอร์มอล (lognormal distribution)
ฟังก์ชันความหนาแน่นอยู่ในรูปของ

$$f(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} \exp - \{1(\ln x-u)^2/\sigma^2\} & ; X>0,\ \sigma^2 > 0, \\ 0 & ; \text{อื่น ๆ} \end{cases}$$

เมื่อ μ และ σ^2 เป็นค่าเฉลี่ยและความแปรปรวนของ Y โดยที่ $Y = \ln X$ และ Y มีการแจกแจงแบบปกติ ในที่นี้พิจารณาใช้ $\mu = 0, \sigma^2 = 1$



รูปที่ 1.4.3 แสดงเส้นโค้งของการแจกแจงแบบลอกนอร์มอล

1.4.1.3 การแจกแจงแบบแกมมา (gamma-distribution)
ฟังก์ชันความหนาแน่นอยู่ในรูปของ

$$f(x) = \begin{cases} \dfrac{x^{\alpha-1}\exp\{-x/\beta\}}{\beta^{\alpha}\Gamma(\alpha)} & ; x>0, \alpha>0, \beta>0 \\ 0 & ; \text{อื่น ๆ} \end{cases}$$

เมื่อ β เป็น scale parameter

และ α เป็น shape parameter

จะได้ว่า $E(X) = \beta\alpha$

$Var(X) = \beta^2\alpha$

สัมประสิทธิ์ความแปรปรวน (C.V.) $= 1/\sqrt{\alpha}$

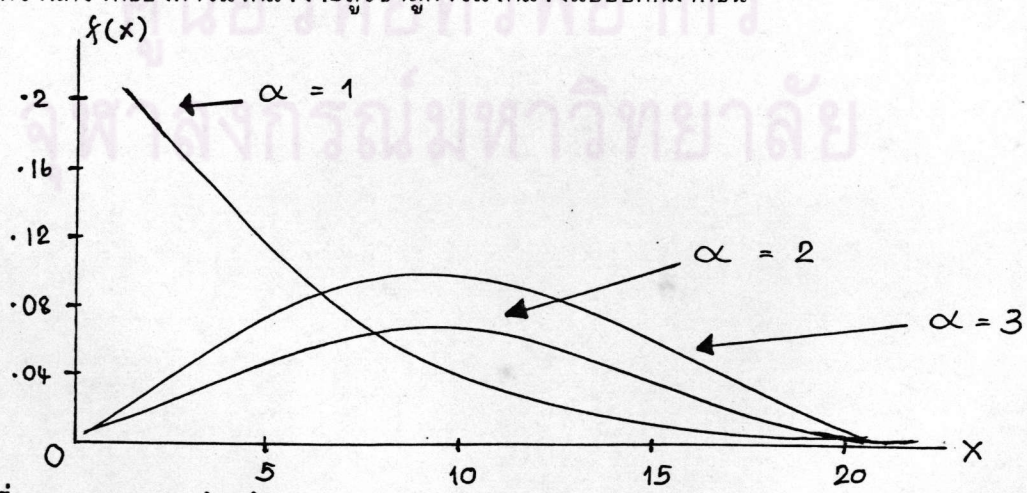
ในการวิจัยครั้งนี้จะศึกษา ณ ค่า α = 5 10 และ 150 เมื่อ β = 1 และ 2 กล่าวคือ

C.V.(X) = 100% (α = 1, β = 1)

C.V.(X) = 70% (α = 2, β = 1)

C.V.(X) = 58% (α = 3, β = 1)

และสาเหตุที่เลือกใช้ค่า C.V. = 100% 70% และ 58% โดยที่ไม่เลือกค่า C.V. ที่มีค่าต่ำกว่านี้เพราะจากการพิจารณากราฟที่แสดงเส้นโค้งของการแจกแจงแบบแกมมา ถ้าค่า C.V.ที่มีค่าต่ำกว่านี้กราฟของการแจกแจงจะลู่เข้าสู่การแจกแจงแบบปกติมากขึ้น

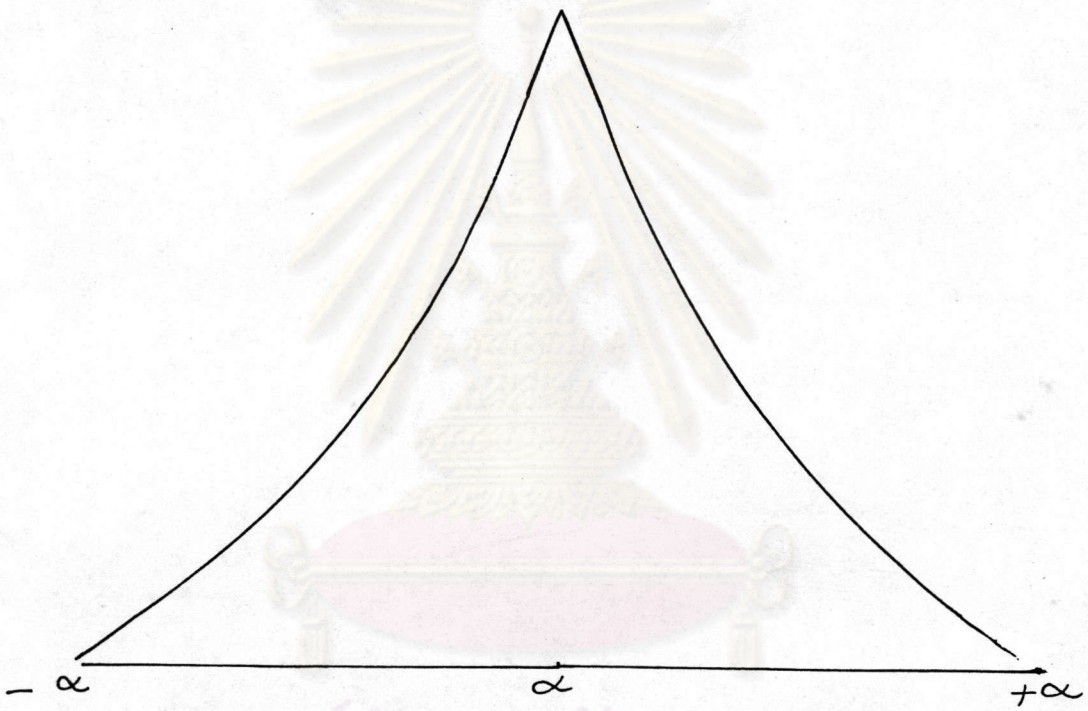


รูปที่ 1.4.4 แสดงเส้นโค้งของการแจกแจงแบบแกมมา ณ β = 1 และ α = 1, 2 และ 3

1.4.1.4 การแจกแจงแบบดับเบิ้ลเอ็กซ์โปเนนเชียล (Double Exponential Distribution)

ฟังก์ชันความน่าจะเป็น คือ

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



ค่าคาดหวัง $E(X) = \alpha$

ค่าความแปรปรวน $V(X) = 2\beta^2$

1.4.8 ขนาดตัวอย่างและตัวแปรอิสระ

จำนวนตัวแปรอิสระและขนาดตัวอย่างที่ใช้ในการวิจัยครั้งนี้แบ่งเป็น 3 กลุ่ม ดังนี้

1. ตัวแปรอิสระ p = 5 จะใช้ขนาดตัวอย่าง 50 และ 100
2. ตัวแปรอิสระ p = 3 จะใช้ขนาดตัวอย่าง 5 10 และ 20
3. ตัวแปรอิสระ p = 2 จะใช้ขนาดตัวอย่าง 4

ทั้งนี้ก็เพื่อจะให้ได้ผลของการวิจัยได้ครอบคลุมทั้งขนาดตัวอย่างและจำนวนตัวแปรอิสระ ที่มีปริมาณน้อย ปานกลาง และใหญ่ คือจะใช้ที่จำนวนตัวแปรอิสระ = 5 ขนาดตัวอย่าง 50 และ 100 เป็นตัวแทนของกลุ่มตัวอย่างขนาดใหญ่ สำหรับที่จำนวนตัวแปรอิสระ = 3 ขนาดตัวอย่าง = 5, 10, 20 เป็นตัวแทนของกลุ่มตัวอย่างขนาดปานกลาง และที่จำนวนตัวแปรอิสระ = 2 ขนาดตัวอย่าง = 4 เป็นตัวแทนของกลุ่มตัวอย่างที่มีขนาดน้อย

1.5 ประโยชน์ที่คาดว่าจะได้รับ

ผลการศึกษาวิธีการประมาณและคุณสมบัติของตัวประมาณจะทำให้สามารถบอกได้ว่าวิธีการประมาณค่าสัมประสิทธิ์ถดถอยพหุทั้ง 3 มีคุณสมบัติและวิธีการที่จะนำไปใช้ในการประมาณค่าสัมประสิทธิ์ถดถอยพหุได้อย่างไร

1.6 คำจำกัดความ

1.6.1 ค่าเฉลี่ยความผิดพลาดกำลังสองเฉลี่ย (Mean Square Error) หรือ MSE ของตัวประมาณ คือ ถ้า $\hat{\theta}$ เป็นตัวประมาณของพารามิเตอร์ แล้ว ความผิดพลาดกำลังสองเฉลี่ยของ θ คือ $E(\hat{\theta} - \theta)^2$

ในการคำนวณหาค่าความผิดพลาดกำลังสองเฉลี่ยของฟังก์ชั่นเชิงเส้นของตัวประมาณ θ ในที่นี้คือ $1'\hat{\beta}$; เมื่อ 1 เป็นเมตริกซ์ขนาด p X 1 ที่สมาชิกทุกตัวมีค่าเป็น 1 นั่นก็คือการหาค่าความผิดพลาดกำลังสองเฉลี่ยในรูปของผลบวกของตัวประมาณนั่นเอง

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

1.6.3 ความไม่เอนเอียง (Unbiased) ของตัวประมาณ คือถ้า $\hat{\theta}$ เป็นตัวประมาณที่ไม่เอนเอียงของ θ ก็ต่อเมื่อ $E(\hat{\theta}) = \theta$

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

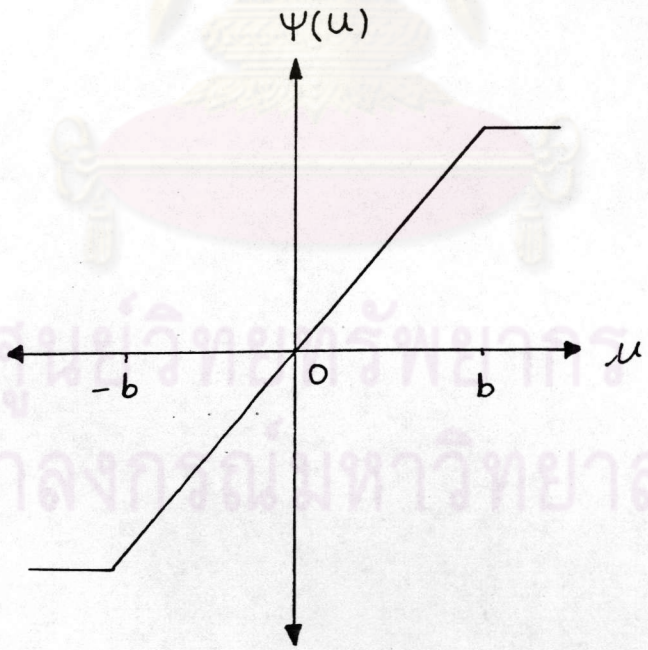
5.1 เป็นฟังก์ชั่นเชิงเส้นของตัวแปรสุ่ม

5.2 เป็นตัวประมาณที่ไม่เอนเอียง

5.3 เป็นตัวประมาณที่มีค่าความแปรปรวนต่ำสุด

1.6.5 ตัวประมาณที่มีค่าความแปรปรวนต่ำสุดเพียงตัวเดียวในบรรดาตัวประมาณที่ไม่เอนเอียง (Uniformly Minimum Variance Unbiased Estimator) หรือ UMVUE เป็นคุณสมบัติของตัวประมาณที่ผู้วิจัยต้องการ θ คือ ถ้ามีตัวประมาณที่ไม่เอนเอียงสำหรับ θ ซึ่งเป็นสถิติที่พอเพียงสำหรับ θ และมีค่าความแปรปรวนต่ำกว่าค่าความแปรปรวนของตัวประมาณอื่นๆ ที่ไม่เอนเอียงสำหรับ θ แล้ว ตัวประมาณดังกล่าวจะมีคุณสมบัติเป็นตัวประมาณที่ดีที่สุดหรือมีค่าความแปรปรวนต่ำสุดเพียงตัวเดียวในบรรดาตัวประมาณที่ไม่เอนเอียง

1.6.6 เกณฑ์ความแกร่งของ Huber คือสมการเชิงเส้นที่จุดกำเนิดและมีค่าคงที่ที่ปลายทั้งสองข้างพิจารณาจากรูปที่ 1.6.7



รูป 1.6.7 แสดงกราฟของเกณฑ์ความแกร่ง Huber ณ b เมื่อ b คือจุดเปลี่ยนเว้า

และมีฟังก์ชั่นความแกร่งดังนี้

$$(r) = \begin{cases} -r^2/2 & ; \text{ถ้า } |r| < k \\ k|r| = k^2/2 & ; \text{อื่น ๆ} \end{cases}$$

ซึ่งค่าประมาณของ Huber เป็นค่าประมาณภาวะน่าจะเป็นสูงสุด (MLE) ที่ดีสำหรับการแจกแจงแบบ ε -contaminate Gaussion และทำให้ค่าประมาณสเกลเป็นเงื่อนไขที่สำคัญของค่าสังเกตในการแปลงข้อมูลจากกำลังสองในรูปสมการเชิงเส้น