บทที่ 1

บทนำ

1.1 ความสำคัญและความเป็นมาของปัญหา

ในงานวิจัยที่ต้องอาศัยเทคนิคการวิเคราะห์ความถดถอยเป็นเครื่องมือช่วยในการเสาะหาคำตอบนั้น ผู้วิจัยมักจะเลือกใช้วิธีการกำลังสองต่ำสุดเป็นวิธีการในการศึกษา เนื่องจากเป็นวิธีการที่แพร่หลายและคุ้นเคยมากกว่า และยังเป็นวิธีการที่ให้ตัวประมาณ (Estimator) ที่ดีคือเป็นวิธีการที่ให้ตัวประมาณเชิงเส้นที่ไม่เอนเอียง และมีความแปรปรวนต่ำสุด เรียกคุณสมบัตินี้ว่า BLUE (Best Linear Unbiased Estimator) แต่ทั้งนี้ขึ้นอยู่กับข้อมูลที่จะนำมาวิเคราะห์ด้วยว่าต้องมีคุณสมบัติตรงตามข้อตกลงเบื้องต้นของการวิเคราะห์ความถดถอย คือ

1. ค่าความคลาดเคลื่อนจะต้องมีการแจกแจงเป็นแบบปกติที่มีค่าเฉลี่ยเป็น 0 และค่าความแปรปรวนเป็น σ^2

2. ค่าความคลาดเคลื่อนจะต้องเป็นอิสระต่อกัน หรือ ε_i และ ε_j จะต้องไม่มีความสัมพันธ์ต่อกัน เมื่อ $i \neq j$; $i = 1,\ldots, n$ $j = 1,\ldots, n$ เมื่อ n คือขนาดตัวอย่าง

3. ค่าความคลาดเคลื่อน (ε) จะต้องเป็นอิสระกับตัวแปรอิสระ (X) หรือ Cov ($\varepsilon_i X_i$) เท่ากับ 0; $i = 1,\ldots, n$ เมื่อ n คือขนาดตัวอย่าง

และตัวแปรตามจะมีค่าสังเกตของตัวเองและสามารถผันแปรค่าไปในลักษณะที่เราสามารถจะควบคุมได้

แต่ปัญหาหนึ่งที่มักจะพบได้คือ ปัญหาเมื่อตัวแปรตาม มีค่าจำกัด (Limited Dependent Variable) จึงทำให้ในบางค่าของข้อมูลเราไม่อาจทราบค่าสังเกตของตัวแปรตามได้อย่างแน่นอน หรือตัวแปรตามบางค่าจะมีค่าขาดหายไป ปัญหานี้เรียกว่า เกิด "Censored Data"

เมื่อเกิดปัญหาตัวแปรตามบางค่ามีค่าขาดหาย (Consored Data) ลักษณะข้อมูลของตัวแปรตามจะมีค่าปนกันระหว่างข้อมูลส่วนที่ไม่ขาดหาย (Survival Time or Uncensored Data) และข้อมูลส่วนที่ขาดหาย (Censoring Time or Censored Data) การทดลองหรือการศึกษาที่เกิดปัญหาตัวแปรตามบางค่ามีค่าขาดหาย เช่น

การทดลองเกี่ยวกับความทนทานหรืออายุการใช้งานของฉนวนกันความร้อนว่าจะขึ้นอยู่กับอุณหภูมิหรือความร้อนที่ได้รับหรือไม่ ในการทดลองนี้ ตัวแปรตามคือ อายุการใช้งานของฉนวน และตัวแปรอิสระ คือ อุณหภูมิหรือความร้อนที่ให้ เมื่อทำการทดลอง โดยให้อุณหภูมิหรือความร้อนแก่ฉนวน แล้วนับจำนวนเวลาหรือจำนวนชั่วโมงที่ฉนวนนั้นจะเสื่อมสภาพ เมื่อสิ้นสุดการทดลองฉนวนบางอันจะเสื่อมสภาพ ซึ่งข้อมูลนี้จะเป็นข้อมูลที่ไม่ขาดหาย (Uncensored Data) แต่ฉนวนบางอันจะยังคงอยู่ในสภาพที่ยังใช้งานได้ดี ซึ่งทำให้เราไม่สามารถที่จะทราบอายุการใช้งานที่แน่นอนได้ จะทราบก็แต่อายุการใช้งานเมื่อสิ้นสุดการทดลองเท่านั้น ข้อมูลนี้จะเป็นข้อมูลที่ขาดหาย (Censored Data)

ในการวิเคราะห์ข้อมูลเมื่อเกิดค่าขาดหายนี้ จะพบว่า ส่วนใหญ่ก็ยังคงจะใช้วิธีการกำลังสองต่ำสุดในการวิเคราะห์ข้อมูล ซึ่งการวิเคราะห์จะกระทำได้ใน 2 กรณี คือ

กรณีแรกจะถือว่า ค่าขาดหายเป็นค่าที่ไม่ขาดหาย นั่นคือจะทำการวิเคราะห์ข้อมูลตามที่มีอยู่ทั้งหมด

กรณีที่สอง ไม่สนใจข้อมูลส่วนที่เป็นค่าขาดหาย นั่นคือจะทำการวิเคราะห์ข้อมูลเฉพาะกับข้อมูลที่ไม่ขาดหาย

ซึ่งการวิเคราะห์โดยใช้วิธีการกำลังสองต่ำสุดในทั้ง 2 กรณีนี้ จะทำให้ได้ตัวประมาณที่เอนเอียง และโดยเฉลี่ยการประมาณค่า จะต่ำกว่าความเป็นจริง หรือจะทำให้ได้ช่วงแห่งความเชื่อมั่นแคบกว่าความเป็นจริง[1]

---

[1] Josef Schmee and Gerald J. Hahn. "A Simple Method for Regression Analysis with Censored Data." Technometrics 21(4); (1979): 417-418.

ด้วยเหตุนี้ จึงได้มีผู้คิดวิธีการในการประมาณค่าพารามิเตอร์ เมื่อเกิดปัญหาข้อมูลขาดหาย ในตัวแปรตามขึ้น คือ มิลเลอร์ (Rupert G. Miller : 1976) และบัคเลย์และเจมส์ (Jonathan Buckley & Ian James : 1979)

มิลเลอร์ ให้ทำการประมาณค่าพารามิเตอร์ โดยใช้ Weighted Least Squares เฉพาะกับข้อมูลที่ไม่ขาดหาย แล้วทำการวนซ้ำ (Iterative) กับค่าพารามิเตอร์ที่ประมาณจนกว่าค่าพารามิเตอร์ที่ประมาณนั้นจะคงตัว (Converge) สำหรับบัคเลย์และเจมส์นั้น ทำการประมาณค่าพารามิเตอร์ โดยให้ทำการประมาณค่าข้อมูลที่ขาดหายขึ้นมา แล้วทำการวิเคราะห์ข้อมูลรวมทั้งหมด คือรวมทั้งข้อมูลที่ไม่ขาดหาย และข้อมูลที่ขาดหายที่ได้ประมาณขึ้น แล้วทำการวนซ้ำกับค่าพารามิเตอร์ที่ประมาณจนกว่าค่าพารามิเตอร์ที่ประมาณนั้นจะคงตัวเช่นเดียวกับวิธีการของมิลเลอร์ ซึ่งวิธีการของบัคเลย์และเจมส์จะให้ตัวประมาณที่ไม่เอนเอียง ในขณะที่วิธีการของมิลเลอร์ จะให้ตัวประมาณที่เกือบจะไม่เอนเอียง (nearly unbiased)

มิลเลอร์และเจอรี่ (Rupert Miller and Jerry Halperm : 1982) ได้ทำการศึกษาเปรียบเทียบวิธีการของมิลเลอร์กับวิธีการของบัคเลย์และเจมส์ โดยใช้ข้อมูลจาก Stanford heart transplant data ซึ่งเป็นข้อมูลที่ได้จากการผ่าตัดเปลี่ยนหัวใจ ให้กับคนไข้ 157 คน ในจำนวนนี้ คนไข้ 55 คนยังมีชีวิตอยู่หลังจากสิ้นสุดการทดลอง ซึ่งข้อมูลของคนไข้ทั้ง 55 คน จะเป็นข้อมูลที่ขาดหาย และข้อมูลของคนไข้ 102 คนที่เหลือ ซึ่งเสียชีวิตหลังจากสิ้นสุดการทดลอง จะเป็นข้อมูลที่ไม่ขาดหาย จากการศึกษาเปรียบเทียบพบว่า วิธีการของบัคเลย์และเจมส์ จะให้ค่าประมาณพารามิเตอร์ที่น่าเชื่อถือ (reliable) กว่าวิธีการของมิลเลอร์

ดังนั้นจึงเป็นที่น่าสนใจว่า เมื่อเกิดปัญหาตัวแปรตามบางค่ามีค่าขาดหายในการวิเคราะห์ความถดถอย วิธีการกำลังสองต่ำสุด วิธีการของมิลเลอร์ และวิธีการของบัคเลย์และเจมส์ วิธีการใดจะเป็นวิธีการที่เหมาะสม ในการประมาณค่าพารามิเตอร์ และเหมาะสมในสถานการณ์ใด

ดังนั้นในการศึกษาครั้งนี้ จึงสนใจที่จะศึกษาเปรียบเทียบวิธีการกำลังสองต่ำสุด วิธีการของบัคเลย์และเจมส์ วิธีการของมิลเลอร์ ในการวิเคราะห์ความถดถอยเมื่อตัวแปรตามบางค่ามีค่าขาดหาย

1.2 วัตถุประสงค์ของการวิจัย

ทำการเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ ในการวิเคราะห์ความถดถอยเมื่อตัวแปรตามบางค่ามีค่าขาดหาย โดยใช้วิธีการของ

1. วิธีการกำลังสองต่ำสุด (Least Squares Method)
2. วิธีการของมิลเลอร์ (Miller Method)
3. วิธีการของบัคเลย์และเจมส์ (Buckley and James Method)

1.3 ข้อตกลงเบื้องต้น

1. การแจกแจงของค่าขาดหายและค่าที่ไม่ขาดหาย จะเป็นอิสระต่อกัน
2. ตัวแปรตาม (Dependent Variable) เท่านั้นที่เป็นค่าขาดหาย
3. ในการศึกษาครั้งนี้ถือว่า ค่าความคลาดเคลื่อนกำลังสองเฉลี่ยเป็นดัชนีสำคัญที่จะใช้ในการเปรียบเทียบตัวประมาณ

1.4 ขอบเขตของการวิจัย

1. ในการศึกษาครั้งนี้จะทำการเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ ในการวิเคราะห์ความถดถอยเมื่อตัวแปรตามมีค่าขาดหาย โดยใช้วิธีการของกำลังสองต่ำสุด วิธีการของมิลเลอร์ วิธีการของบัคเลย์และเจมส์
2. ศึกษาเฉพาะกรณีของการวิเคราะห์ความถดถอยอย่างง่าย (Simple Linear Regression)
3. กำหนดค่าพารามิเตอร์ $\alpha = 30$ $\beta = 20$ ทุกเงื่อนไขของการศึกษา
4. ศึกษาเมื่อกรณีของค่าที่ไม่ขาดหาย T_i มีรูปแบบเป็น

$$T_i = \alpha + \beta X_i + \varepsilon_i \qquad ; i = 1,\ldots, n$$

เมื่อค่าความคลาดเคลื่อน ε มีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเป็น 0 และค่าความแปรปรวนเป็น 16

5. ศึกษาเมื่อกรณีของค่าขาดหาย C_i มีการแจกแจงเป็นแบบ

5.1 แบบยูนิฟอร์มในช่วง $(\beta X_i , \beta X_i + C)$ เมื่อ $\beta = 20$ และค่า C จะแปรเปลี่ยนไปเพื่อให้เกิดเปอร์เซ็นต์เฉลี่ยของค่าขาดหายตามที่กำหนดไว้

5.2 แบบแกมม่า เมื่อ $\alpha = 1$ และ β จะแปรเปลี่ยนไปเพื่อให้เกิดเปอร์เซ็นต์เฉลี่ยของค่าขาดหายตามที่กำหนดไว้

5.3 แบบปกติ โดยที่ค่าเฉลี่ยและค่าความแปรปรวนจะแปรเปลี่ยนไปเพื่อให้เกิดเปอร์เซ็นต์เฉลี่ยของค่าขาดหายตามที่กำหนดไว้

5.4 เมื่อค่าขาดหายเป็นฟังก์ชันเชิงเส้นกับค่าความคลาดเคลื่อน ในรูปแบบความสัมพันธ์เป็น

$$C_i = \alpha_1 + \beta_1 X_i + \varepsilon_i$$

เมื่อค่าความคลาดเคลื่อน ε มีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเท่ากับ 0 และค่าความแปรปรวน σ^2

โดยที่ค่า α_1 และ β_1 และ σ^2 จะแปรเปลี่ยนไปเพื่อให้เกิดเปอร์เซ็นต์เฉลี่ยของค่าขาดหายตามที่กำหนดไว้

6. ศึกษาเมื่อกรณีของประเภทค่าขาดหายเป็นแบบสุ่ม (Random Censoring) โดยที่ตัวแปรใหม่ที่จะใช้ในการวิเคราะห์จะมาจาก

$$Y_i = \min(T_i, C_i) \quad ; i = 1,\ldots, n$$

เมื่อมีตัวชี้ δ_i ที่สัมพันธ์กับค่า Y_i ดังนี้

$$\delta_i = \begin{cases} 1 & \text{ถ้า } T_i < C_i \text{ (เป็นข้อมูลที่มาจากค่าที่ไม่ขาดหาย)} \\ 0 & \text{ถ้า } T_i > C_i \text{ (เป็นข้อมูลที่มาจากค่าขาดหาย)} \end{cases}$$

7. ศึกษาเมื่อกรณีเกิดเปอร์เซ็นต์ของค่าขาดหาย (Average % Censored) เป็น 5%, 10%, 15%, 20%, 25%, 30%, 50%, 60%, 70%, 80%

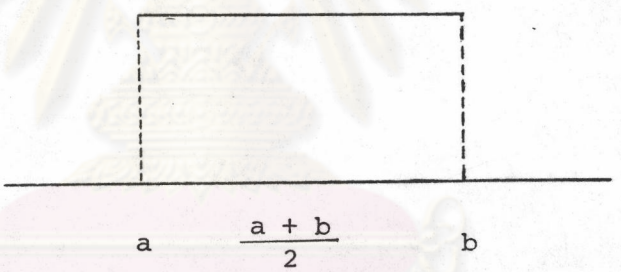
8. ศึกษาเมื่อขนาดตัวอย่างเป็น 10, 20, 50, 60, 100, 150

9. ข้อมูลที่ใช้ในการศึกษาได้มาจากการจำลองขึ้นในเครื่องคอมพิวเตอร์โดยจะกระทำซ้ำ 100 รอบ ในทุกสถานการณ์ที่ศึกษา

รูปแบบของการแจกแจงที่กำหนด

การแจกแจงแบบยูนิฟอร์ม (Uniform Distribution) มีฟังก์ชั่นความน่าจะเป็นคือ

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

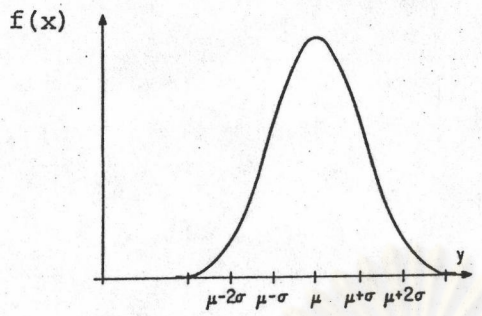


ค่าคาดหวัง $E(X) = \frac{(a+b)}{2}$

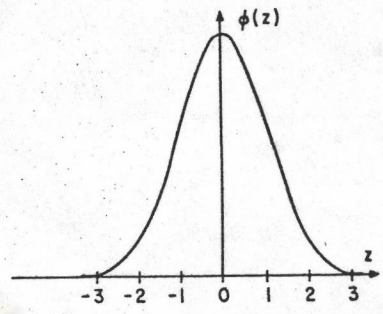
ค่าความแปรปรวน $V(X) = \frac{(b-a)^2}{12}$

การแจกแจงแบบปกติ (Normal Distribution) มีฟังก์ชั่นความน่าจะเป็น คือ

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



Normal probability density.



Standard normal density.

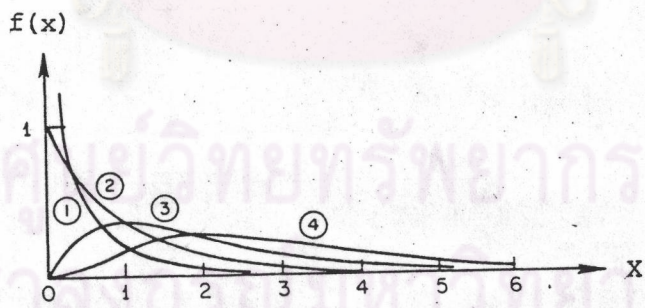
ค่าคาดหวัง $E(X) = \mu$

ค่าความแปรปรวน $V(X) = \sigma^2$

การแจกแจงแบบแกมม่า (Gamma Distribution) มีฟังก์ชั่นความน่าจะเป็น คือ

$$f(x) = \frac{1}{\Gamma(\beta)} x^{\beta-1} \frac{e^{-x/\alpha}}{\alpha^{\beta}} \quad ; \ x > 0$$

$$; \ \alpha, \beta > 0$$



The gamma density for $\alpha = 1$ and

① $\beta = 1/2$, ② $\beta = 1$, ③ $\beta = 2$, ④ $\beta = 3$.

ค่าคาดหมาย $E(X) = \beta\alpha$

ค่าความแปรปรวน $V(X) = \beta\alpha^2$

1.5 ประโยชน์ที่คาดว่าจะได้รับ

เพื่อช่วยให้นักวิจัยมีผลสรุปและหลักฐาน ในการเลือกใช้วิธีการวิเคราะห์ความถดถอย เมื่อตัวแปรตามบางค่ามีค่าขาดหาย