บทที่ 2

ระเบียบวิธีการวิจัย

วิธีการประมาณค่าสังเกตสูญหายที่นำมาเปรียบเทียบในการศึกษาครั้งนี้เป็นการศึกษาเพียง 3 วิธี วิธีแรกคือวิธีกำลังสองน้อยสุด วิธีนี้เป็นการประมาณค่าที่ทำให้ผลรวมกำลังสองของความคลาดเคลื่อนมีค่าน้อยที่สุด วิธีที่สองคือ วิธี EM algorithm วิธีนี้อาศัยฟังก์ชันไลลิฮูดเข้ามาช่วยในการประมาณค่าพารามิเตอร์ และหาค่าตัวสถิติเพียงพอ ซึ่งทำให้การวิเคราะห์ง่ายขึ้น และวิธีที่สามคือวิธี Imputation Method วิธีนี้เป็นการสร้างชุดข้อมูลขึ้นมาใหม่จากชุดที่มีอยู่เดิมและนำค่าความแปรปรวนมาประกอบในการพิจารณาเลือกค่าประมาณ จากหลักการดังกล่าวข้างต้น ผู้วิจัยจึงสนใจที่จะศึกษาและเปรียบเทียบวิธีการประมาณค่าสังเกตที่สูญหายทั้ง 3 วิธี

2.1 การประมาณค่าสูญหายโดยวิธีกำลังสองน้อยสุด (least square method)

การวิเคราะห์ข้อมูลโดยการประมาณค่าสังเกตสูญหาย เป็นการวิเคราะห์โดยประมาณ ทำได้โดยประมาณค่าสังเกตสูญหายขึ้นมาแล้ววิเคราะห์ผลด้วยเทคนิคตามแผนแบบการทดลองที่วางไว้ไปตามปกติ

การประมาณค่าสังเกตสูญหายเมื่อมีค่าสังเกตสูญหายเพียง 1 ค่า อลันและวิสชาร์ท (Allan and Wishart ;1930) เป็นผู้ริเริ่มเอามาใช้ และต่อมายทส์ (Yate ; 1937) ได้แสดงให้เห็นว่าสูตรดังกล่าวได้มาจากวิธีกำลังสองน้อยที่สุด คือค่าประมาณขึ้นเป็นค่าที่ทำให้ผลรวมของกำลังสองของความคลาดเคลื่อน (SSE) มีค่าน้อยที่สุด และวิธีนี้ยังสามารถนำไปใช้หาค่าประมาณของค่าสังเกตสูญหายหลายค่าโดยวิธีวนซ้ำ

เนื่องจากค่าประมาณของค่าสังเกตที่สูญหายเป็นค่าประมาณที่ทำให้ ผลรวมกำลังสองของความคลาดเคลื่อน ต่ำที่สุดที่คิดคำนวณจากค่าสังเกตที่มีอยู่เท่านั้น จึงมีผลต่อการวิเคราะห์ความแปรปรวนดังนี้

1. ค่าผลรวมกำลังสองของความคลาดเคลื่อน เป็นค่าที่ถูกต้อง

2. ค่าผลรวมกำลังสองของวิธีทดลอง (SSTr) ที่คำนวณได้สูงกว่าปกติ โดยที่ ผลรวมกำลังสองของวิธีทดลอง จะมีค่าสูงกว่าที่ควรจะเป็นเท่ากับผลต่างระหว่างกำลังสองของผลรวมกำลังสองภายในบล็อคเมื่อรวมค่าประมาณของค่าสังเกตสูญหายและเมื่อไม่มีค่าสังเกตสูญหาย ซึ่งในการวิเคราะห์ข้อมูลควรจะต้องนำไปหักออกจาก ผลรวมกำลังสองของวิธีทดลอง

3. องศาความเป็นอิสระของความคลาดเคลื่อน ลดลงเท่ากับจำนวนค่าสังเกตที่สูญหาย

การประมาณค่าสังเกตสูญหาย 1 ค่า

ให้ M_{ij} = ค่าสังเกตสูญหายซึ่งอยู่ในบล็อคที่ j และวิธีการทดลองที่ i

B_i = ผลรวมของบล็อคที่มีค่าสังเกตสูญหาย

T_i = ผลรวมของวิธีการทดลองที่มีค่าสังเกตสูญหาย

T = ผลรวมของค่าสังเกตทั้งหมด

เนื่องจากผลรวมกำลังสองของความคลาดเคลื่อน

$$SSE = \sum_{i=1}^{t}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$$

ให้ C เป็นค่าคงที่ใด ๆ จะเขียนผลรวมกำลังสองของความคลาดเคลื่อน อยู่ในรูปฟังก์ชันของ M_{ij} ได้ดังนี้

$$SSE = T + M_{ij}^2 - \frac{1}{b}(T_i + M_{ij})^2 - \frac{1}{t}(B_j + M_{ij})^2 + \frac{1}{tb}(T + M_{ij})^2$$

ค่าของ M_{ij} ซึ่งทำให้ผลรวมกำลังสองของความคลาดเคลื่อนน้อยที่สุดคือ M_{ij} ซึ่งทำให้

$$\frac{\partial SSE}{\partial M_{ij}} = 0 \quad \text{หรือ}$$

$$M_{ij} - \frac{1}{b}(T_i + M_{ij}) - \frac{1}{t}(B_j + M_{ij}) + \frac{1}{tb}(T + M_{ij}) = 0$$

ดังนั้น

$$M_{ij} = \frac{bB_j + tT_i - T}{(b-1)(t-1)}$$

การประมาณค่าสังเกตสูญหายตั้งแต่ 1 ค่าขึ้นไป

ให้ M_{ij} และ $M_{i'j'}$ เป็นค่าสังเกตสูญหาย 2 ค่า ซึ่งอยู่ต่างบล็อคและได้รับวิธีการทดลองต่างกัน

M_{ij} = ค่าสังเกตสูญหายซึ่งอยู่ในบล็อคที่ j และวิธีการทดลองที่ i

M_{ij} = ค่าสังเกตสูญหายซึ่งอยู่ในบล็อคที่ j และวิธีการทดลองที่ i

B_j และ $B_{j'}$ เป็นผลรวมของบล็อคที่ j และ j' โดยค่าสังเกตสูญหายมีค่าเป็นศูนย์

T_i และ $T_{i'}$ เป็นผลรวมของวิธีการทดลองi และ i' โดยค่าสังเกตสูญหายมีค่าเป็นศูนย์

T ผลรวมของค่าสังเกตทั้งหมดโดยค่าสังเกตสูญหายมีค่าเป็นศูนย์

เมื่อ C เป็นค่าคงที่ผลรวมกำลังสองของความคลาดเคลื่อนเขียนอยู่ในรูปฟังก์ชันของ M_{ij} และ $M_{ij'}$ ได้เป็น

$$SSE = T + M_{ij}^2 + M_{ij'}^2 - \frac{1}{b}(T_i + M_{ij})^2 - \frac{1}{b}(T_{i'} + M_{ij'})^2 - \frac{1}{t}(B_j + M_{ij})^2$$
$$- \frac{1}{t}(B_{j'} + M_{ij'})^2 + \frac{1}{tb}(T + M_{ij} + M_{ij'})^2$$

หาอนุพันธ์เทียบกับ M_{ij} และ $M_{ij'}$ เทียบกับศูนย์

$$M_{ij} - \frac{1}{b}(T_i + M_{ij}) - \frac{1}{t}(B_j + M_{ij}) + \frac{1}{tb}(T + M_{ij} + M_{ij'}) = 0 \text{.....(1)}$$

$$M_{ij'} - \frac{1}{b}(T_{i'} + M_{ij'}) - \frac{1}{t}(B_{j'} + M_{ij'}) + \frac{1}{tb}(T + M_{ij} + M_{ij'}) = 0 \text{.....(2)}$$

ซึ่งถ้ามีค่าสังเกตสูญหายเพียง 1 ค่าสมมติว่าเป็น M_{ij} ค่าประมาณของ M_{ij} หาได้โดยการแทนค่า $M_{ij'} = 0$ ในสมการที่ (1)จะได้

$$M_{ij} = \frac{bB_j + tT_i - T}{(b-1)(t-1)}$$

แต่ถ้าค่าสังเกตสูญหาย 2 ค่าแก้สมการ (1) และ (2) จะได้

$$M_{ij} = \frac{(b-1)(t-1)(bB_j + tT_i - T) - (bB_{j'} + tT_{i'} - T)}{(b-1)^2(t-1)^2 - 1}$$

$$M_{ij'} = \frac{(b-1)(t-1)(bB_{j'} + tT_i - T) - (bB_j + tT_{i'} - T)}{(b-1)^2(t-1)^2 - 1}$$

ในกรณีที่มีค่าสังเกตสูญหายมากกว่าค่าจะประมาณค่าสูญหายได้โดยวิธีเดียวกันคือเขียนผลรวมกำลังสองของความคลาดเคลื่อนในรูปฟังก์ชันของค่าสังเกตที่สูญหายแล้วหาอนุพันธ์เทียบกับค่าสังเกตเทียบเท่ากับศูนย์จะได้สมการจำนวนเท่ากับจำนวนค่าสังเกตที่สูญหายแก้สมการจะได้สูตรการประมาณค่าสังเกตสูญหายเหล่านั้น

ด้วยการประมาณค่าโดยวิธีดังกล่าวค่อนข้างจะยุ่งยากในทางปฏิบัติจึงมักจะหาค่าประมาณโดยวิธีวนซ้ำตามแบบของเยทส์ซึ่งทำได้ตามขั้นตอนต่อไปนี้

สมมติมีค่าสังเกตสูญหายจำนวน n ค่า

1) ประมาณค่าสังเกตสูญหายเบื้องต้นของค่าที่ 1 ถึง n-1 โดยใช้สูตร

$$M_{ij} = \frac{\bar{Y}_i + \bar{B}_j}{2}$$

เมื่อ M_{ij} = ค่าสังเกตสูญหายที่ต้องการประมาณ ซึ่งอยู่ในบล็อคที่ j ได้รับวิธีการทดลองที่ i

$\bar{X}_i$ และ $\bar{B}_j$ = ค่าเฉลี่ยของวิธีการทดลองและของบล็อคที่มีค่าสังเกตสูญหาย

2) ประมาณค่าสังเกตสูญหายค่าที่ n ด้วยสูตร

$$M_{ij} = \frac{bB_j + tT_i - T}{(b-1)(t-1)}$$

เมื่อ M_{ij} = ค่าสังเกตสูญหายซึ่งอยู่ในบล็อคที่ j และวิธีการทดลองที่ i

B_i = ผลรวมของบล็อคที่มีค่าสังเกตสูญหาย

T_i = ผลรวมของวิธีการทดลองที่มีค่าสังเกตสูญหาย

T = ผลรวมของค่าสังเกตทั้งหมด

b และ t = จำนวนบล็อคและจำนวนวิธีการทดลอง ตามลำดับ

3) ประมาณค่าสังเกตสูญหายค่าที่ 1 ถึง n ใหม่ โดยประมาณทีละค่าเริ่มจากค่าที่ 1 ด้วยสูตร

$$M_{ij} = \frac{bB_j + tT_i - T}{(b-1)(t-1)}$$

ทำซ้ำไปมาจนค่าที่ประมาณได้ทุกค่าคงตัว

ตัวอย่าง การวางแผนการทดลองแบบสุ่มในบล็อคสมบูรณ์ เมื่อค่าสังเกตสูญหาย 2 ค่า

วิธีการทดลอง	บล็อค			
	1	2	3	4
1	5.113	5.398	5.307	4.678
2	5.436	5.952	4.719	**M24**
3	5.272	5.713	5.483	4.749
4	5.164	4.831	**M43**	4.410
5	4.804	4.848	4.432	4.748
6	5.254	4.542	4.919	4.098

ตาราง ก

ประมาณค่าสังเกตสูญหายโดยวิธีกำลังสองน้อยสุด

รอบที่ 1 ประมาณ M24

$$\bar{X}_2 = \frac{16.107}{3} = 5.369 \; ; \; \bar{B}_4 = \frac{22.683}{3} = 4.537$$

$$M24 = \frac{\bar{X}_2 + \bar{B}_4}{2} = 4.953$$

แทนค่า M24 ในตาราง ก จากนั้นประมาณ M43

$$M_{43} = \frac{bB_3 + tT_4 - T}{(b-1)(t-1)}$$

$$= \frac{4(24.860) + 6(14.405) - (109.870 + 4.953)}{(4-1)(6-1)} = 4.736$$

แทนค่า M43 = 4.736 ในตาราง ก ลบค่า M24 ทิ้ง

รอบที่ 2 ประมาณ M24

$$M_{24} = \frac{bB_4 + tT_2 - T}{(b-1)(t-1)}$$

$$= \frac{4(22.683) + 6(16.107) - (109.870 + 4.736)}{(4-1)(6-1)} = 4.851$$

แทนค่า M24 = 4.851 ในตาราง ก ลบค่า M43 ทิ้ง

ประมาณ M43

$$M_{43} = \frac{bB_3 + tT_4 - T}{(b-1)(t-1)}$$

$$= \frac{4(24.860) + 6(14.405) - (109.870 + 4.851)}{(4-1)(6-1)} = 4.743$$

แทนค่า M43 = 4.743 ในตาราง ก ลบค่า M24 ทิ้ง

รอบที่ 3 ประมาณ M24

$$M_{24} = \frac{bB_4 + tT_2 - T}{(b-1)(t-1)}$$

$$= \frac{4(22.683) + 6(16.107) - (109.870 + 4.743)}{(4-1)(6-1)} = 4.851$$

ซึ่งซ้ำกับค่าที่ได้ในรอบที่ 2

ดังนั้นค่าประมาณของ M24 = 4.851 และ M43 = 4.743

2.2 วิธีประมาณค่าสูญหายโดยวิธี EM algorithm (Expectation Maximization)

การประมาณค่าสังเกตสูญหายโดยวิธี EM algorithm (Expectation Maximization) เป็นทางเลือกหนึ่ง ซึ่งเสนอโดยเดมเสตอร์ ลายด์ และรูบิน(Dempster Laird and Rubin) ค่าที่ประมาณขึ้นเป็นค่าที่มาจากกระบวนการวนซ้ำเพื่อค้นหาค่าประมาณไลลิฮูดที่มากที่สุดของค่าพารามิเตอร์ โดยกำหนดเงื่อนไขของข้อมูลที่สูญหายในฟังก์ชันไลลิฮูดจะทำให้การวิเคราะห์นั้นง่ายขึ้น EM algorithm นี้แบ่งได้เป็น 2 ขั้นตอน คือขั้นหาค่าคาดหวัง E step และขั้นหาค่ามากที่สุด M step EM algorithm นี้ เราจะหาค่าตามลำดับของ $\theta^{(t)}$ ที่เพิ่มขึ้น ในฟังก์ชันไลลิฮูดของ θ ซึ่งมาจากการวนซ้ำ

สมมติว่าข้อมูล Y มีองค์ประกอบ 2 ส่วน คือ $Y = (Y_{mis}, Y_{obs})$ และสมมติว่ามีข้อมูลทั้งหมด n ค่า m ค่าเป็นค่าที่เก็บมาได้ ส่วนที่เหลืออีก n – m ค่าเป็นข้อมูลที่หายไป

พิจารณาการแจกแจงของข้อมูลที่เหลืออยู่ว่ามีการแจกแจงแบบใดหากมีการแจกแจงแบบปกติ ขั้นหาค่าคาดหวัง E step หาได้จาก

$$E(\sum_{i=1}^{n} y_i \mid \theta^{(t)}, Y_{obs}) = \sum_{i=1}^{m} y_i + (n-m)\mu^{(t)} \quad \text{.....(1)}$$

$$E(\sum_{i=1}^{n} y_i^2 \mid \theta^{(t)}, Y_{obs}) = \sum_{i=1}^{m} y_i^2 + (n-m)[(\mu^{(t)})^2 + (\sigma^{(t)})^2] \quad \text{.....(2)}$$

ขั้นหาค่ามากที่สุด M step หาได้จาก

$$\mu^{(t+1)} = \frac{E(\sum_{i=1}^{n} y_i \mid \theta^{(t)}, Y_{obs})}{n} \quad \text{..... (3)}$$

$$(\sigma^{(t+1)})^2 = \frac{E(\sum_{i=1}^{n} y_i^2 \mid \theta^{(t)}, Y_{obs})}{n} - (\mu^{(t+1)})^2 \quad \text{.....(4)}$$

กำหนดค่า $\mu^{(t)} = \mu^{(t+1)} = \hat{\mu}$ และ $\sigma^{(t)} = \sigma^{(t+1)} = \hat{\sigma}$ ในสมการ (1)-(4) และวนซ้ำจนลู่เข้าสู่ค่าคงที่

$$\hat{\mu} = \frac{\sum_{i=1}^{m} y_i}{m}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{m} y_i^2}{m} - \hat{\mu}^2$$

2.3 วิธีประมาณค่าสูญหายโดยวิธี Imputation (Imputation Method)

วิธี Imputation Method เป็นวิธีที่ใช้กับกรณีข้อมูลที่หายเป็นไปอย่างสุ่ม ถ้ามีข้อมูลสูญหาย m ค่า ให้สร้างชุดข้อมูลขึ้นมาใหม่โดยสุ่มจากข้อมูลที่เหลืออยู่ สร้างให้จำนวนแถวเท่ากับจำนวนที่ข้อมูลหาย จำนวนคอลัมน์มีค่าอยู่ระหว่าง 2 ถึง 10 คอลัมน์ จากนั้นคำนวณหาค่าเฉลี่ยในแต่ละชุดของข้อมูลโดยนำค่าเฉลี่ยของข้อมูลชุดที่มีความแปรปรวนต่ำสุดมาเป็นตัวแทนของข้อมูลที่สูญหาย

ค่าประมาณแบบจุดของ q^* คือ

$$q^* = \frac{\sum_{j=1}^{m} q_i}{m}$$

q^* = ค่าเฉลี่ยของข้อมูลแต่ละชุด

ตัวอย่าง กรณีข้อมูลสูญหาย 2 ค่า

จำนวนแถวที่สร้างขึ้นใหม่จะมีค่าเท่ากับจำนวนข้อมูลที่สูญหายไป

จำนวนคอลัมน์ที่สร้างขึ้นมีค่าอยู่ระหว่าง 2 ถึง 10 ค่า จะให้ผลการประมาณค่าสูญหายที่ดี

Y1
Y5

ค่าเฉลี่ยชุดที่ 1

Y6
Y9

ค่าเฉลี่ยชุดที่ 2

Y11
Y14

ค่าเฉลี่ยชุดที่ 3

พิจารณาเลือกค่าเฉลี่ยชุดที่มีความแปรปรวนต่ำสุดเป็นตัวแทนของข้อมูลที่สูญหาย

2.4 เกณฑ์ที่ใช้ในการเปรียบเทียบวิธีการประมาณ

เกณฑ์การพิจารณาว่า วิธีการประมาณค่าสูญหายใด จะให้ค่าประมาณที่ดีจะพิจารณาโดยเปรียบเทียบค่าเฉลี่ยกำลังสองของความคลาดเคลื่อน MSE (Mean Square Error)

$$MSE = \frac{\sum_{i=1}^{n}(Y - \hat{Y})^2}{n}$$

Y = ค่าจริงที่ได้จากการจำลอง

$\hat{Y}$ = ค่าประมาณจากการใช้วิธีการประมาณค่า

n = จำนวนรอบจากการทดลอง

วิธีใดให้ค่าเฉลี่ยกำลังสองของความคลาดเคลื่อน MSE ต่ำกว่าเป็นวิธีที่ดีกว่า นั่นแสดงว่าค่าประมาณที่ได้มีค่าใกล้เคียงค่าจริงที่สูญหายไปมากกว่า