บทที่ 2

สถิติที่ใช้ในการวิจัย

ในบทนี้จะกล่าวถึงวิธีในการประมาณค่าที่สูญหายของตัวแปรตาม ในการวิเคราะห์การถดถอยเชิงเส้นพหุ และใช้วิธีกำลังสองน้อยที่สุดหาสัมประสิทธิ์การถดถอย เพื่อหาสมการถดถอยเชิงเส้นพหุในการพยากรณ์ ซึ่งวิธีการประมาณค่าสูญหายทั้ง 5 วิธีมีดังนี้คือ

1. วิธีสูญหาย คือวิธีที่ตัดชุดข้อมูลที่มีค่าสูญหายออก
2. วิธีค่าเฉลี่ย คือวิธีที่ใช้ค่าเฉลี่ยประมาณข้อมูลสูญหาย
3. วิธีสมการถดถอย คือวิธีที่ประมาณข้อมูลสูญหายจากสมการถดถอยเชิงเส้นพหุ
4. วิธีอีเอ็ม (EM Algorithm)
5. วิธีการของฮันท์ (Hunt's Method)

<u>วิธีกำลังสองน้อยที่สุดแบบสามัญ</u> (Ordinary Least Squares Method : OLS Method)

วิธีการหาสัมประสิทธิ์การถดถอยโดยวิธีกำลังสองน้อยที่สุด เป็นวิธีที่มีรากฐานมาจากทฤษฎีการประมาณค่าเชิงเส้น (Theory of Linear Estimation) ซึ่งเป็นวิธีที่คิดค้นโดย คาร์ล เฟรดริช เกาส์ (Karl Friedrich Gauss 1777-1855) และ อังเดร แอนดรีวิช มาร์คอฟ (Andrie Andreevich Markov 1856-1922)[1] โดยมีหลักเกณฑ์ดังนี้ คือหาค่าประมาณของพารามิเตอร์ที่ทำให้ผลบวกกำลังสองของผลต่างระหว่างค่าสังเกตกับค่าคาดหวังของตัวแปรมีค่าต่ำที่สุด

---

[1]ประชุม สุวัตถี,ดร., ทฤษฎีการอนุมานเชิงสถิติ. (กรุงเทพมหานคร: 2527), หน้า

การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยที่สุด

จากสมการความสัมพันธ์ระหว่างตัวแปรตาม y และตัวแปรอิสระ X คือ

$$y = X\beta + \varepsilon \qquad \text{เมื่อ } \varepsilon \sim N(0,\sigma^2 I_n)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n+m} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{(n+m)1} & x_{(n+m)2} & \cdots & x_{(n+m)p} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{n+m} \end{bmatrix}$$

เมื่อ y คือเวกเตอร์ของตัวแปรตามขนาด (n+m)x1

X คือเมตริกซ์ของตัวแปรอิสระขนาด (n+m)x(p+1) และแรงค์เต็ม (Full Rank)

β คือเวกเตอร์ของพารามิเตอร์ที่ไม่ทราบค่าขนาด (p+1)x1

ε คือเวกเตอร์ของความผิดพลาดขนาด (n+m)x1

n+m คือจำนวนค่าสังเกตทั้งหมด

p คือจำนวนตัวแปรอิสระ

โดยทั่วไปเมื่อมีข้อมูลอยู่อย่างครบถ้วน วิธีกำลังสองน้อยที่สุดในการประมาณสัมประสิทธิ์ของการถดถอย โดยทำให้ผลบวกกำลังสองของความคลาดเคลื่อน (Sum of Square of Errors : SSE) มีค่าน้อยที่สุด

$$\begin{aligned} SSE &= \varepsilon^T\varepsilon \\ &= (y - X\hat{\beta})^T (y - X\hat{\beta}) \\ &= y^Ty - y^TX\hat{\beta} - \hat{\beta}^TX^Ty + \hat{\beta}^TX^TX\hat{\beta} \\ &= y^Ty - 2\hat{\beta}^TX^Ty + \hat{\beta}^TX^TX\hat{\beta} \end{aligned}$$

การหาค่ากำลังสองน้อยที่สุดของผลบวกกำลังสองของความคลาดเคลื่อน ทำได้โดยหาอนุพันธ์ (Differentiate) เทียบกับ $\hat{\beta}$ แล้วกำหนดให้เท่ากับศูนย์

$$\frac{\partial}{\partial \hat{\beta}} (y^T y - 2\hat{\beta}^T X^T y + \hat{\beta}^T X^T X \hat{\beta}) = 0$$

$$-2X^T y + 2X^T X \hat{\beta} = 0$$

$$(X^T X)\hat{\beta} = X^T y$$

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

ดังนั้นสมการถดถอยที่ใช้พยากรณ์คือ

$$\hat{y} = X\hat{\beta}$$

โดยที่ $E(\hat{\beta}) = \beta$ และ $V(\hat{\beta}) = (X^T X)^{-1}\sigma^2$

สำหรับรายละเอียดเกี่ยวกับวิธีประมาณค่าสูญหายของตัวแปรตาม ในการวิเคราะห์การถดถอยเชิงเส้นพหุแต่ละวิธีเป็นดังนี้

<u>วิธีสูญหาย</u>

วิธีการนี้จะเป็นวิธีการที่เกิดขึ้นเสมอ ๆ เมื่อผู้วิจัยต้องการแก้ปัญหาเฉพาะหน้า โดยการตัดชุดข้อมูลที่มีค่าสูญหายออกไป หลังจากนั้นก็จะประมาณค่าสัมประสิทธิ์การถดถอยโดยวิธี OLS ดังนี้

$$\hat{\beta}^* = (X^{*T} X^*)^{-1} X^{*T} y^*$$

เมื่อ X^* และ y^* คือชุดข้อมูลที่เหลืออยู่ทั้งของ X และ y

<u>วิธีค่าเฉลี่ย</u>

วิธีการนี้จะเป็นวิธีประมาณค่าสูญหายของตัวแปรตามโดยใช้ค่าเฉลี่ยของข้อมูลที่ไม่สูญหายของตัวแปรตาม นั่นคือ

$$\bar{y}^* = \frac{\sum_{t=1}^{n} y_t}{n}$$

โดยที่ $\bar{y}^*$ คือ ค่าเฉลี่ยของข้อมูลที่ไม่สูญหายของตัวแปรตาม

n คือ จำนวนข้อมูลที่ไม่สูญหายของตัวแปรตาม

เมื่อแทนข้อมูลที่สูญหายด้วยค่าเฉลี่ยแล้ว จะทำการประมาณค่าสัมประสิทธิ์การถดถอยโดยวิธี OLS เพื่อจะได้สมการถดถอยเชิงเส้นพหุมาใช้ในการพยากรณ์ต่อไป

<u>วิธีสมการถดถอย</u>

วิธีการนี้จะเป็นวิธีประมาณค่าสูญหายของตัวแปรตามดังขั้นตอนดังต่อไปนี้

1. จากข้อมูล $(\mathbf{X}, y)$ ที่เหลืออยู่ประมาณค่าสัมประสิทธิ์การถดถอย โดยวิธี OLS ดังนี้

$$\hat{\beta}^* = (\mathbf{X}^{*T}\mathbf{X}^*)^{-1}\mathbf{X}^{*T}y^*$$

เมื่อ $\mathbf{X}^*$ และ y^* คือชุดข้อมูลของ $(\mathbf{X}, y)$ ที่เหลืออยู่

2. นำค่าหรือสมการที่ได้จาก 1. มาประมาณค่าสูญหายของตัวแปรตามโดยพิจารณาจากสมการถดถอยเชิงเส้นพหุ

$$\hat{y}_t = \mathbf{x}_t^T \hat{\beta}^*$$
$$= [1 \;\; x_{1t} \;\; x_{2t}]\hat{\beta}^* \quad ; t = 1,2,..,m$$

เมื่อ $\hat{y}_t$ คือค่าประมาณของค่าสูญหายตัวที่ t

x_{1t}, x_{2t} คือค่าสังเกตชุดที่ t ของตัวแปรอิสระตัวที่ 1 และ 2

m คือจำนวนข้อมูลที่สูญหาย

3. นำค่า $\hat{y}_t$ แทนในค่าสูญหายตัวที่ t ของตัวแปรตาม แล้วทำการประมาณค่าสัมประสิทธิ์ความถดถอย โดยวิธี OLS เพื่อจะได้สมการถดถอยเชิงเส้นพหุมาใช้ในการพยากรณ์

<u>วิธีอีเอ็ม</u>

Dempster และคณะ (1977) ได้สรุปแนวความคิดในการหาตัวประมาณภาวะน่าจะเป็นสูงสุด เมื่อมีข้อมูลบางส่วนสูญหายในการแจกแจงต่าง ๆ โดยอาศัยวิธีการทำซ้ำ (Iterative Method) จนกระทั่งได้ตัวประมาณภาวะน่าจะเป็นสูงสุด ให้ชื่อวิธีการนี้ว่าวิธีอีเอ็ม ซึ่งมีชื่อเต็ม ๆ ว่า Expectation-Maximization Algorithm หลักการของวิธีการนี้มี 2 ขั้นตอน คือ

<u>ขั้นตอนที่ 1</u> ขั้นตอน E (Expectation-Step) เป็นขั้นตอนที่หาค่าคาดหวังของค่าที่สูญหายไปภายใต้เงื่อนไขชุดข้อมูลที่ไม่สูญหายและพารามิเตอร์ตัวปัจจุบัน ค่าที่ได้นี้จะนำไปประมาณค่าที่สูญหาย

<u>ขั้นตอนที่ 2</u> ขั้นตอน M (Maximization-Step) เป็นขั้นตอนที่ประมาณค่าภาวะน่าจะเป็นสูงสุดของพารามิเตอร์ ด้วยการแทนค่าสูญหายที่ได้จากขั้นตอน E ทำซ้ำจนกระทั่งได้ตัวพารามิเตอร์ที่คงที่ นั่นคือตัวประมาณภาวะน่าจะเป็นสูงสุด

Little และ Rubin (1987) ได้ประยุกต์วิธีการของอีเอ็มมาใช้ในการประมาณค่าสูญหายในการวิเคราะห์การถดถอยเชิงเส้นพหุ ซึ่งลักษณะที่สนใจคือประมาณค่าสูญหายของตัวแปรตามในการวิเคราะห์การถดถอยเชิงเส้นพหุ โดยมีขั้นตอนการทำที่เข้าใจง่าย ๆ ดังนี้

1. จัดข้อมูลตามดังนี้

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \beta + \varepsilon$$

เมื่อ y_1 คือเวกเตอร์ของตัวแปรตามที่ไม่สูญหายขนาด nx1

y_2 คือเวกเตอร์ของตัวแปรตามที่สูญหายขนาด mx1

X_1 คือเมตริกซ์ของตัวแปรอิสระขนาด nx(p+1) เมื่อชุดข้อมูลของตัวแปรตามไม่สูญหาย

X_2 คือเมตริกซ์ของตัวแปรอิสระขนาด mx(p+1) เมื่อชุดข้อมูลของตัวแปรตามสูญหาย

2. เริ่มต้นด้วยการการประมาณค่าสัมประสิทธิ์การถดถอยโดยวิธี OLS จากชุดข้อมูลที่เหลืออยู่

$$\hat{\beta}^{(0)} = (X_1^T X_1)^{-1} X_1^T y_1$$

3. เข้าสู่ขั้นตอน E ในการทำซ้ำรอบที่ 1 เพื่อประมาณค่าที่สูญหาย

$$E(y_t | X, y_1, \hat{\beta}^{(0)}) = \begin{cases} y_t & \text{ถ้า } y_t \text{ ไม่สูญหาย } (t = 1,2,\ldots,n) \\ X_2 \hat{\beta}^{(0)} & \text{ถ้า } y_t \text{ สูญหาย } (t = n+1,\ldots,n+m) \end{cases}$$

ดังนั้น $y_t^{(1)} = E(y_t | X, y_1, \hat{\beta}^{(0)})$

4. เข้าสู่ขั้นตอน M ในการทำซ้ำรอบที่ 1

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

5. หาค่าสัมบูรณ์ของผลต่างระหว่างค่าสัมประสิทธิ์การถดถอยเริ่มต้น กับค่าสัมประสิทธิ์การถดถอยรอบที่ 1 ของค่าสัมประสิทธิ์ทุกค่า

6. ถ้าค่าทุกค่าในขั้นที่ 5 มากกว่า 0.001 ให้ทำขั้นต่อไป แต่ถ้าไม่มากกว่า 0.001 จะได้ค่าประมาณค่าที่สูญหายด้วยขั้นตอน E รอบที่ 1

7. เข้าสู่ขั้นตอน E ในการทำซ้ำรอบที่ k ; k = 2,3,...

$$E(y_t|X, y_1, \hat{\beta}^{(k-1)}) = \begin{cases} y_t & \text{ถ้า } y_t \text{ ไม่สูญหาย } (t = 1,2,\ldots,n) \\ X_2\hat{\beta}^{(k-1)} & \text{ถ้า } y_t \text{ สูญหาย } (t = n+1,\ldots,n+m) \end{cases}$$

ดังนั้น $y_t^{(k)} = E(y_t|X, y_1, \hat{\beta}^{(k-1)})$

8. เข้าสู่ขั้นตอน M ในการทำซ้ำรอบที่ k

$$\hat{\beta}^{(k)} = (X^TX)^{-1}X^Ty^{(k)}$$

9. หาค่าสัมบูรณ์ของผลต่างระหว่างค่าสัมประสิทธิ์การถดถอยรอบที่ k-1 กับค่าสัมประสิทธิ์การถดถอยรอบที่ k ของค่าสัมประสิทธิ์ทุกค่า

10. ถ้าค่าทุกค่าในขั้นที่ 9 มากกว่า 0.001 ให้กลับไปทำซ้ำขั้นที่ 7 ถึง 9 เรื่อย ๆ จนกระทั่งค่าสัมบูรณ์ของผลต่างไม่มากกว่า 0.001 ดังนั้นจะได้ค่าประมาณค่าที่สูญหายด้วยขั้นตอน E สุดท้าย

11. นำค่าทุกค่าในขั้นที่ 6 หรือ 10 แทนในค่าที่สูญหายแล้วทำการประมาณค่าสัมประสิทธิ์การถดถอย โดยวิธี OLS เพื่อจะได้สมการถดถอยเชิงเส้นพหุมาใช้ในการพยากรณ์

<u>วิธีการของฮันท์</u>

Hunt (1987) ได้เสนอวิธีการนี้เพื่อประมาณค่าสูญหายของตัวแปรตามในสมการ $\mathcal{Y} = X\beta + \varepsilon$ เมื่อ $\varepsilon \sim N(0,\sigma^2 I_n)$ โดยใช้หลักการของ Healy และ Westmacott (1956) ซึ่งหลักการนั้นคือการใช้วิธีการทำซ้ำ (Iterative Method) เพื่อลดความคลาดเคลื่อนของค่าประมาณค่าที่สูญหายจนกระทั่งค่าประมาณของค่าที่สูญหายมีค่าคงที่ และวิธีนี้ได้ใช้หลักการของโพรเจกชันเชิงตั้งฉาก (Orthogonal Projection)* มาหาเวกเตอร์ของความคลาดเคลื่อน (ε) ดังนี้

$$\varepsilon = (I - P)\mathcal{Y}$$

เมื่อ I คือเมตริกซ์เอกลักษณ์ และ $P = X(X^TX)^{-1}X^T$

โดยจะสรุปวิธีการนี้ด้วยขั้นตอนต่อไปนี้

1. กำหนดค่าเริ่มต้นของขั้นตอนการทำซ้ำ โดยแทนค่าที่สูญหายด้วยค่าเฉลี่ยของข้อมูลที่ไม่สูญหายของตัวแปรตาม นั่นคือ

$$y_t^{(0)} = \begin{cases} y_t & \text{ถ้า } y_t \text{ ไม่สูญหาย } (t = 1,2,\ldots,n) \\ \bar{y}^* & \text{ถ้า } y_t \text{ สูญหาย } (t = n+1,\ldots,n+m) \end{cases}$$

$$\bar{y}^* = \frac{\sum_{t=1}^{n} y_t}{n}$$

โดยที่ $\bar{y}^*$ คือ ค่าเฉลี่ยของข้อมูลที่ไม่สูญหายของตัวแปรตาม

n คือ จำนวนข้อมูลที่ไม่สูญหายของตัวแปรตาม

---

*โพรเจกชันเชิงตั้งฉาก เป็นการแปลงเชิงเส้นจากเวกเตอร์ค่าสังเกต ($\mathcal{Y}$) ไปยังเวกเตอร์ค่าประมาณ ($\hat{\mathcal{Y}}$) เพื่อทำให้เกิดความคลาดเคลื่อนในการประมาณน้อยที่สุด นั่นคือ คือ $P\mathcal{Y} = \hat{\mathcal{Y}}$ และจะได้ $\mathcal{Y} - \hat{\mathcal{Y}} = (I - P)\mathcal{Y} = \varepsilon$

2. หาเวกเตอร์ของความคลาดเคลื่อนรอบที่ 1

$$\mathcal{E}^{(1)} = (\mathbf{I} - \mathbf{P})\mathcal{Y}^{(0)}$$

3. ประมาณค่าสูญหายรอบที่ 1 ซึ่งก็คือการลดความคลาดเคลื่อนของค่าประมาณค่าที่สูญหายนั่นเอง

$$y_t^{(1)} = \begin{cases} y_t & \text{ถ้า } y_t \text{ ไม่สูญหาย } (t = 1,2,\ldots,n) \\ y_t^{(0)} - \varepsilon_t^{(1)} & \text{ถ้า } y_t \text{ สูญหาย } (t = n+1,\ldots,n+m) \end{cases}$$

4. หาค่าสัมบูรณ์ของผลต่างระหว่างค่าประมาณค่าที่สูญหายเริ่มต้นกับค่าประมาณค่าที่สูญหายรอบที่ 1 ของค่าที่สูญหายทุกตัว

5. ถ้าค่าทุกตัวในขั้นที่ 4 มากกว่า 0.001 ให้ทำขั้นต่อไป แต่ถ้าไม่มากกว่า 0.001 จะได้ค่าประมาณค่าที่สูญหาย

6. หาเวกเตอร์ของความคลาดเคลื่อนรอบที่ k ; k = 2,3,...

$$\mathcal{E}^{(k)} = (\mathbf{I} - \mathbf{P})\mathcal{Y}^{(k-1)}$$

7. ประมาณค่าที่สูญหายรอบที่ k

$$y_t^{(k)} = \begin{cases} y_t & \text{ถ้า } y_t \text{ ไม่สูญหาย } (t = 1,2,\ldots,n) \\ y_t^{(k-1)} - \varepsilon_t^{(k)} & \text{ถ้า } y_t \text{ สูญหาย } (t = n+1,\ldots,n+m) \end{cases}$$

8. หาค่าสัมบูรณ์ของผลต่างระหว่างค่าประมาณค่าที่สูญหายรอบที่ k-1 กับค่าประมาณค่าที่สูญหายรอบที่ k ของค่าที่สูญหายทุกตัว

9. ถ้าค่าทุกตัวในขั้นที่ 8 มากกว่า 0.001 ให้กลับไปทำซ้ำขั้นที่ 6 ถึง 8 เรื่อย ๆ จนกระทั่งค่าสัมบูรณ์ของผลต่างไม่มากกว่า 0.001 ดังนั้นจะได้ค่าประมาณค่าที่สูญหาย

10. นำค่าทุกตัวในขั้นที่ 5 หรือ 9 แทนในค่าที่สูญหาย แล้วทำการประมาณค่าสัมประสิทธิ์การถดถอย โดยวิธี OLS เพื่อจะได้สมการถดถอยเชิงเส้นพหุมาใช้ในการพยากรณ์