บทที่ 2

แบบจำลองเชิงเส้น
(Linear Models)

บทที่ 2
แบบจำลองเชิงเส้น
(Linear Models)

2.0 คำนำ

ความสัมพันธ์ในรูปฟังก์ชัน (Functional Relationship)

ถ้าให้ Y เป็นปริมาณหนึ่งซึ่งเราต้องการทำนาย

$X_1, X_2, \ldots, X_n$ เป็นตัวแปรจำนวน n ตัว

และ ฟังก์ชัน g ซึ่งทำให้ $Y = g(X_1, X_2, \ldots, X_n)$

ถ้าเราสามารถสังเกตค่า $X_1, X_2, \ldots, X_n$ ได้ เราจะสามารถทำนายค่า Y ได้อย่างถูกต้องแม่นยำ แล้วเราจะกล่าวว่า

"$Y, X_1, X_2, \ldots, X_n$ **มีความสัมพันธ์ในรูปฟังก์ชัน**"

ตัวอย่างเช่น เราต้องการทำนายความสูง (h) ของคนคนหนึ่ง เราทราบว่าเราไม่อาจทำนายความสูงของคนคนนั้นได้อย่างแม่นยำจากการทราบน้ำหนัก (w) ของเขาเพียงอย่างเดียว อย่างไรก็ดีเราอาจใช้ปริมาณอื่น ๆ มาช่วยได้ เช่น น้ำหนักของบิดา (X_1) ความสูงของบิดา (X_2) น้ำหนักของมารดา (X_3) เป็นต้น

เราสมมุติว่าสมการอยู่ในรูป $h = g(w, X_1, X_2, \ldots, X_n)$
หรืออาจสมมุติว่าเราสามารถเขียนสมการได้ในรูป

$$h = \alpha + \beta w + f(X_1, X_2, \ldots, X_n)$$

และสมมุติอีกว่า สำหรับค่า w ที่กำหนดให้ ปริมาณ $X_1, X_2, \ldots, X_n$ จะเปลี่ยนไป และ $f(X_1, X_2, \ldots, X_n)$ ทำหน้าที่เป็นความคลาดเคลื่อนสุ่ม (random error)

ถึงแม้ว่าเราทราบฟังก์ชัน g ข้างต้น และทราบว่า X_i ตัวใดบ้างที่ใช้ทำนายค่า Y เราก็ยังไม่สามารถทำนาย Y ได้อย่างแม่นยำ ทั้งนี้เพราะเราไม่สามารถวัดค่า X_i บางตัวหรือทุกตัวได้อย่างแม่นยำ นั่นคือมีความคลาดเคลื่อนจากการวัด

ความคลาดเคลื่อนที่สนใจ 2 ชนิดคือ

1) **ความคลาดเคลื่อนจากการใช้สมการ (Equation Error)**
เกิดจากการไม่ทราบฟังก์ชัน g หรือการใช้ฟังก์ชัน g ที่ผิด เนื่องจากไม่ทราบว่า X_i ใดบ้างที่จะใช้ทำนาย Y จึงทำให้ใช้ X_i ไม่ถูกต้องหรือไม่ครบถ้วน

2) **ความคลาดเคลื่อนจากการวัด (Measurement Error)**
เกิดจากการที่ไม่สามารถวัด X_i ได้อย่างแม่นยำ

2.1 คำจำกัดความ

คำจำกัดความ 2.1 เมื่อกล่าวถึง **แบบจำลอง (Model)** เราหมายถึง สมการทางคณิตศาสตร์ (Mathematical equation) ซึ่งประกอบด้วยตัวแปรเชิงสุ่ม (Random variables) ตัวแปรทางคณิตศาสตร์ (Mathematical variables) และ พารามิเตอร์ (Parameters)

คำจำกัดความ 2.2 เมื่อกล่าวถึง **แบบจำลองเชิงเส้น** (Linear Model) เราหมายถึง สมการซึ่งประกอบตัวแปรเชิงสุ่ม ตัวแปรทางคณิตศาสตร์ และ พารามิเตอร์ และเป็น**เชิงเส้นในพารามิเตอร์**

ตัวอย่าง ถ้า β_0, β_1, β_2 เป็นพารามิเตอร์ที่ไม่ทราบค่า

$\beta_0 + \beta_1 X + \beta_2 Y = 0$ เป็นแบบจำลองเชิงเส้น

และ $\beta_0 + \beta_1 e^x + \beta_2 Y \cos x = 0$ เป็นแบบจำลองเชิงเส้น

แต่ $\beta_0^2 + X \sin \beta_1 + \beta_2 = 0$ ไม่เป็นแบบจำลองเชิงเส้น

ในบทนี้สัญลักษณ์ต่าง ๆ ที่ใช้คือ

ใช้ U^*, V^*, X^*,... แทนตัวแปรทางคณิตศาสตร์ที่สังเกตค่าไม่ได้

U, V, X,... แทนตัวแปรทางคณิตศาสตร์ที่สังเกตค่าได้

y_i^*, u_i^*, v_i^*,... แทนตัวแปรเชิงสุ่มที่สังเกตค่าไม่ได้

y_i, u_i, v_i,... แทนตัวแปรเชิงสุ่มที่สังเกตค่าได้

a, b, d, e,... แทนตัวแปรเชิงสุ่มที่ไม่สามารถสังเกตค่าได้

และ อักษรกรีก แทนพารามิเตอร์ที่ไม่ทราบค่า

2.2 การจำแนกแบบจำลอง (ตามวิธีของนักสถิติ)

แบบจำลอง 1 General Linear Hypothesis of Full Rank

$$y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + e$$

หรือเขียนแบบจำลองในรูปเมตริกซ์ดังนี้คือ

$$Y = X\beta + e$$

โดยที่ $X_1, X_2, \ldots, X_k$ เป็นตัวแปรทางคณิตศาสตร์ที่ทราบค่า

y เป็นตัวแปรเชิงสุ่มที่สังเกตค่าได้

$\beta_o, \beta_1, \ldots, \beta_k$ เป็นพารามิเตอร์ที่ไม่ทราบค่า

e เป็นตัวแปรเชิงสุ่มที่สังเกตค่าไม่ได้ และมี mean เป็น 0

แบบจำลอง 2 Functionally Related Models

(with Variables Subject to Measurement Error)

แบบจำลองนี้แสดงความสัมพันธ์ในรูปฟังก์ชันของตัวแปรทางคณิตศาสตร์

ถ้า $Y^*, X_1^*, \ldots, X_k^*$ สังเกตค่าไม่ได้

$y, x_1, \ldots, x_k$ สังเกตค่าได้ โดยที่ $y = Y^* + e$

$$x_i = X_i^* + e_i \ , \ i = 1, \ldots, k$$

e และ e_i's คือความคลาดเคลื่อนจากการวัด (Measurement error)

จากความสัมพันธ์ $Y^* = \beta_0 + \beta_1 X_1^* + \ldots + \beta_k X_k^*$

หรือ $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + e - \beta_1 e_1 - \beta_2 e_2 - \ldots - \beta_k e_k$

ตัวอย่างเช่น กฎของโอห์ม (Ohm's Law) กล่าวว่า จำนวนโวลท์ V^* ในวงจรหนึ่งมีค่าเท่ากับผลคูณของความต้านทาน ρ ของเส้นลวด กับกระแสไฟฟ้า I สมการคือ $V^* = \rho I$ สมมุติว่าผู้ทดลองต้องการหาความต้านทาน ρ ของวงจร สมมุติว่าเขาไม่สามารถวัด V^* แต่เขาวัดได้ค่า $v = V^* + e$ แล้วสมการกลายเป็น

$v = \rho I + e$ ในที่นี้ v และ e เป็นตัวแปรเชิงสุ่ม

I เป็นตัวแปรทางคณิตศาสตร์ และ

ρ คือพารามิเตอร์ที่ไม่ทราบค่าซึ่งเราจะทำการประมาณค่า

<u>แบบจำลอง 3</u> Regression Models

$y, x_1, \ldots, x_k$ เป็นเซทของตัวแปรเชิงสุ่มซึ่งมีฟังก์ชันการแจกแจงร่วมที่มี

$$E(Y \mid x_i = X_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \quad , i = 1, \ldots, k$$

หรือ $y_x = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + e$ โดยที่ $E(e) = 0$

<u>หมายเหตุ</u> ข้อแตกต่างระหว่างแบบจำลอง 1 และแบบจำลอง 3 คือ

X_i ในแบบจำลอง 1 เป็นตัวแปรทางคณิตศาสตร์

X_i ในแบบจำลอง 3 เป็นค่าของตัวแปรเชิงสุ่ม

ตัวอย่างเช่น สมมุติว่าเราต้องการทำนายอุณหภูมิ y ในบริเวณหนึ่งโดยการใช้เพียงค่าความชื้น x เราสมมุติว่า x และ y มาจากฟังก์ชันการแจกแจงร่วมแบบ bivariate เราสามารถใช้แบบจำลองการถดถอยและทำนายอุณหภูมิเฉลี่ยสำหรับค่าความชื้นที่กำหนดให้ค่าหนึ่ง ๆ

<u>แบบจำลอง 4</u> **Experimental Design Models**

(General Linear Hypothesis of Less than Full Rank)

$$y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + e$$

โดยที่ y เป็นตัวแปรเชิงสุ่ม

$X_1, X_2, \ldots, X_k$ มีค่า 0 หรือ 1

$\beta_1, \beta_2, \ldots, \beta_k$ เป็นพารามิเตอร์ที่ไม่ทราบค่า

e เป็นความคลาดเคลื่อนสุ่ม

แบบจำลอง 4 เป็นกรณีพิเศษของแบบจำลอง 1 แต่เนื่องจากความสำคัญของแบบจำลองนี้ จึงแยกออกมาพิจารณาต่างหาก

โดยทั่ว ๆ ไปอาจเขียนแบบจำลองได้ดังนี้

$$y_{ij\ldots k} = \mu + \beta_i + \alpha_j + \ldots + e_{ij\ldots k}$$

โดยที่ $\beta_i, \alpha_j, \ldots$ เป็นพารามิเตอร์ที่ไม่ทราบค่า

$e_{ij\ldots k}$ เป็นความคลาดเคลื่อนสุ่ม

<u>แบบจำลอง 5</u> **Component-of-Variance Models**

$a_1, a_2, \ldots, a_p$ เป็นตัวแปรเชิงสุ่มซึ่งสังเกตค่าไม่ได้จากการแจกแจงซึ่งมีค่าเฉลี่ย (mean) เท่ากับ 0 และ ความแปรปรวน (variance) เท่ากับ σ_a^2

$b_{11}, b_{12}, \ldots, b_{pq}$ เป็นตัวแปรเชิงสุ่มซึ่งสังเกตค่าไม่ได้จากการแจกแจงซึ่งมีค่าเฉลี่ยเท่ากับ 0 และ ความแปรปรวนเท่ากับ σ_b^2

y_{ij} เป็นตัวแปรเชิงสุ่มซึ่งสังเกตค่าได้

ทั้งนี้

$$y_{ij} = \mu + a_i + b_{ij}$$

โดยที่ μ เป็นพารามิเตอร์ที่ไม่ทราบค่า

ดังนั้น $Var(y_{ij}) = \sigma_a^2 + \sigma_b^2$

เราอาจเขียนแบบจำลองนี้ในรูป

$$y = aX_1 + bX_2 + e$$

แบบจำลองนี้คล้ายกับแบบจำลอง 4 ตรงที่ X_i's มีค่า 0 หรือ 1 เท่านั้น ส่วน a และ b ไม่ใช่พารามิเตอร์ แต่เป็นตัวแปรเชิงสุ่มจากการแจกแจงซึ่งมีความแปรปรวนเท่ากับ σ_a^2 และ σ_b^2 ตามลำดับ

จุดประสงค์ของแบบจำลองชนิดนี้คือ การสังเกตค่าของ y แล้วประมาณค่า σ_a^2, σ_b^2 และ $Var(e) = \sigma^2$

ตัวอย่างเช่น ในการวัดปริมาณไนโตรเจนของใบไม้ของต้นไม้ต้นหนึ่ง ที่มาหลัก ๆ ของความแปรปรวนมีอยู่ 2 ที่มา คือ จากความแปรปรวนของใบไม้บนต้นไม้ และจากความแปรปรวนเนื่องจากการวัดหรือการทดลอง สมมุติว่าเราเอาใบไม้มา n ใบจากต้นไม้ซึ่งมีปริมาณไนโตรเจนที่แท้จริงของใบไม้ใบที่ i เท่ากับ a_i ในกรณีนี้ a_i เป็นตัวแปรเชิงสุ่มจากการแจกแจงซึ่งมีความแปรปรวน σ_a^2 สมมุติว่า ทำการวัด m ครั้งจากใบไม้แต่ละใบเพื่อหาปริมาณไนโตรเจนของใบไม้ใบนั้น

ถ้าให้ y_{ij} แทนปริมาณไนโตรเจนที่วัดได้ในครั้งที่ j ของใบไม้ใบที่ i

e_{ij} เป็นตัวแปรเชิงสุ่มจากการแจกแจงซึ่งมีความแปรปรวน σ^2

เราเขียนแบบจำลองได้ในรูป

$$y_{ij} = \mu + a_i + e_{ij}$$

μ เป็นค่าคงที่ ซึ่งคือค่าเฉลี่ยของ y_{ij}

จุดประสงค์หนึ่งของแบบจำลองชนิดนี้คือ การสังเกตค่าของ y แล้วประมาณค่า σ_a^2 และ $Var(e) = \sigma^2$

โปรดสังเกตว่าความแปรปรวนของ y_{ij} เท่ากับ $\sigma_a^2 + \sigma^2$ ซึ่งเป็นฟังก์ชันเชิงเส้นของความแปรปรวน

ในบทที่ 3 - 5 จะกล่าวถึงแบบจำลอง 1 ในบทที่ 6 จะกล่าวถึงแบบจำลอง 3 ส่วนในบทที่ 7 จะกล่าวถึงแบบจำลอง 4 และกล่าวถึงแบบจำลอง 5 เพียงเล็กน้อย