บทที่ 3

ทฤษฎี Multiple Linear Regression

ทฤษฎี (Multiple Linear Regression) เป็นทฤษฎีที่กล่าวถึงการหาความสัมพันธ์ระหว่างข้อมูล ตั้งแต่ 2 ชุดขึ้นไป เมื่อกำหนดให้ข้อมูลชุดหนึ่งเป็นตัวแปรตาม (Dependent Variable) และอีกชุดหนึ่งเป็นตัวแปรอิสระ (Independent Variable) แล้วศึกษาหาความสัมพันธ์ของข้อมูลเหล่านั้นในรูปสมการกำลังหนึ่ง

$$Y = a + b_1 x_1 + b_2 x_2 + \dots + b_k x_k$$

เมื่อกำหนดให้ Y เป็นตัวแปรตาม (Dependent Variable)

X_1 เป็นตัวแปรอิสระ(Independent Variable)ตัวที่ 1

.....

X_k เป็นตัวแปรอิสระ(Independent Variable)ตัวที่ k

b_1 เป็น Partial Regression Coefficient ของ y ต่อ x_1 เมื่อให้ $x_1 \dots x_k$ คงที่ นั่นคือ เมื่อ x_1 เปลี่ยนแปลง 1 หน่วยจะมีผลทำให้ y เปลี่ยนแปลง b_1 หน่วย เมื่อให้ตัวแปรอิสระตัวอื่นคงที่

b_k เป็น Partial Regression Coefficient ของ y ต่อ x_k เมื่อให้ $x_1, \dots, x_{k-1}$ คงที่ นั่นคือ เมื่อ x_k เปลี่ยนแปลง 1 หน่วยจะมีผลให้ y เปลี่ยนแปลง b_k หน่วย เมื่อให้ตัวแปรอิสระตัวอื่นคงที่

a เป็นค่า intercept ของสมการเส้นตรง

การคำนวณค่า Regression Coefficient $(b_1,\ldots,b_k)$

ค่าประมาณของ Y คำนวณจากสมการ

$$Y = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_k X_k$$

ให้ Y เป็นค่าจริง

$\hat{Y}$ เป็นค่าประมาณจากสมการ

e เป็นความแตกต่างระหว่างค่าจริงกับค่าประมาณเรียกว่าความคลาดเคลื่อน (random error)

$$e = Y - \hat{Y}$$

$$= Y - a - b_1 X_1 - b_2 X_2 - b_3 X_3 - \ldots - b_k X_k$$

เมื่อข้อมูลมีจำนวน n ข้อมูล

$$\sum_{i=1}^{n} e^2 = \sum_{i=1}^{n} (Y-a-b_1X_1-b_2X_2-b_3X_3-\ldots-b_kX_k)^2 \quad \ldots (1)$$

เพื่อให้ความคลาดเคลื่อนมีค่าน้อยที่สุด จะใช้ Method of Least Square โดยให้ $\sum e^2$ มีค่าต่ำสุด จะได้ค่า $a, b_1, b_2 \ldots b_k$ เป็นค่า Unbiased Estimated และมี Standard Errors น้อยที่สุด เมื่อเปรียบเทียบกับ Unbiased Estimated อื่น ๆ ที่คำนวณจาก Model เดียวกัน

1) การหาค่า a โดยหา Partial Derivative ของสมการ (1) เมื่อเทียบกับ a แล้วกำหนดให้ = 0

$$\frac{\partial(\sum e^2)}{a} = 0$$

$$\frac{\partial(\sum e^2)}{a} = -2 \sum_{i=1}^{n} (Y-a-b_1X_1-b_2X_2-b_3X_3-\ldots-b_kX_k)=0$$

$$\therefore \quad \sum_{i=1}^{n} a = \sum_{i=1}^{n} (Y - b_1X_1 - b_2X_2 - b_3X_3 - \cdots - b_kX_k)$$

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 - b_3\bar{X}_3 - \cdots - b_k\bar{X}_k$$

แทนค่า a ในสมการ (1) จะได้

$$\sum_{i=1}^{n} e^2 = (Y - \bar{Y} - b_1(X_1 - \bar{X}_1) - b_2(X_2 - \bar{X}_2) - \cdots - b_k(X_k - \bar{X}_k))^2$$

$$y = (Y - \bar{Y}),\ x_1 = (X_1 - \bar{X}_1),\ x_2 = (X_2 - \bar{X}_2),\ \ldots\ x_k = (X_k - \bar{X}_k)$$

$$\therefore \quad \sum_{i=1}^{n} e^2 = \sum_{i=1}^{n} (y - b_1x_1 - b_2x_2 - \cdots - b_kx_k)^2 \quad \ldots (2)$$

2) การหาค่า b โดยหา Partial Derivative ของสมการ (2) เทียบกับค่า b_i ทีละตัว แล้วกำหนดให้เท่ากับ 0 จะได้สมการทั้งหมด k สมการ ซึ่งเป็น Normal Equation สำหรับ Independent Variable k ตัวคือ

$$\frac{\partial \sum e^2}{\partial b_1} = 0, \quad b_1\sum x_1^2 + b_2\sum x_1x_2 + \cdots + b_k\sum x_1x_k = \sum x_1y$$

$$\frac{\partial \sum e^2}{\partial b_2} = 0, \quad b_1\sum x_2x_1 + b_2\sum x_2^2 + \cdots + b_k\sum x_2x_k = \sum x_2y$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$\frac{\partial \sum e^2}{\partial b_k} = 0, \quad b_1\sum x_kx_1 + b_2\sum x_kx_2 + \cdots + b_k\sum x_k^2 = \sum x_ky$$

ขั้นต่อไปดำเนินการหาค่า $b_1, b_2, \ldots b_k$ จาก Normal Equation ทั้ง k สมการ

<u>การคำนวณหาค่า $b_1, b_2, \ldots b_k$ โดยใช้ Matrix</u>

เมื่อข้อมูลมี Independent Variable 3 ตัว ($k = 3$) และมีจำนวน n ข้อมูล จะมีลักษณะดังนี้คือ

ลำดับที่	Y	X_1	X_2	X_3
1	Y_1	X_{11}	X_{21}	X_{31}
2	Y_2	X_{12}	X_{22}	X_{32}
.....				
n	Y_n	X_{1n}	X_{2n}	X_{3n}

เราสามารถหาความสัมพันธ์ระหว่างตัวแปรต่าง ๆ ในรูปของสมการกำลังหนึ่ง ได้ดังนี้

ถ้าเขียนสมการแสดงความสัมพันธ์ระหว่างตัวแปรต่าง ๆ ในรูปของ Matrix จะได้

$$Y = X\beta + \epsilon$$

โดยกำหนดให้

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & x_{n3} \end{bmatrix}$$

$$\beta = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}, \quad \epsilon = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

จะเห็นว่า

$$X'X = \begin{bmatrix} \sum x_{i1}^2 & \sum x_{i1}x_{i2} & \sum x_{i1}x_{i3} \\ \sum x_{i2}x_{i1} & \sum x_{i2}^2 & \sum x_{i2}x_{i3} \\ \sum x_{i3}x_{i1} & \sum x_{i}x_{i2} & \sum x_{i3}^2 \end{bmatrix}$$ เรียกว่า Coefficient Matrix

และ $X'Y = \begin{bmatrix} \sum x_{i1}y_i \\ \sum x_{i2}y_i \\ \sum x_{i3}y_i \end{bmatrix}$

ดังนั้น Normal Equation ในกรณีนี้คือ

$$X'XB = X'Y$$

$$(X'X)^{-1}(X'X)B = (X'X)^{-1}X'Y$$

$$B = (X'X)^{-1}X'Y$$

$$B = C\,(X'Y)$$

เมื่อกำหนดให้ $C = (X'X)^{-1} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix}$

C เรียกว่า Variance Covariance Matrix

$$B_1 = c_{11}\sum x_{i1}y_i + c_{12}\sum x_{i2}y_i + c_{13}\sum x_{i3}y_i$$

$$B_2 = c_{21}\sum x_{i1}y_i + c_{22}\sum x_{i2}y_i + c_{23}\sum x_{i3}y_i$$

$$B_3 = c_{31}\sum x_{i1}y_i + c_{32}\sum x_{i2}y_i + c_{33}\sum x_{i3}y_i$$

Sum of Squares เนื่องมาจากความถดถอย ซึ่งใช้สัญญลักษณ์ว่า $= \sum_{j=1}^{3} \beta_j \left(\sum_{i=1}^{n} x_{ij} y_i\right)$

$$= \beta_1 (\sum x_{i1} y_i) + \beta_2 (\sum x_{i2} y_i) + \beta_3 (\sum x_{i3} y_i)$$

ในการหา Matrix Inversion ของ $(X'X)^{-1}$ เพื่อหา β_j นั้น ในกรณีที่ข้อมูลมีจำนวนน้อย ๆ จะใช้วิธีที่ใช้กันโดยทั่วไปเรียกว่า Abbreviated Doolittle Method [1/] ซึ่งเป็นวิธีที่สลับซับซ้อนพอประมาณ ในทางปฏิบัติสำหรับข้อมูลที่มีจำนวนมากและมีตัวแปรอิสระหลายตัว จะใช้เครื่องคอมพิวเตอร์ในการคำนวณเพื่อความรวดเร็วและถูกต้อง

<u>การวิเคราะห์ข้อมูลโดยวิธี Multiple Linear Regression</u>

ตามทฤษฎี Multiple Linear Regression ถือว่า Deviation ของ Y มีการกระจายแบบ Normal ซึ่งมี Mean = 0 และ Variance = σ^2

$$\varepsilon \sim N(0, \sigma^2)$$

Unbiased Estimated ของ σ^2 คือ $s^2 = \frac{\sum (Y-\hat{Y})^2}{n-k}$

เมื่อ n = จำนวนข้อมูล

k = จำนวน parameter ที่ใช้ในการคำนวณ

ถ้าให้ $e = (Y-\hat{Y})$ เป็นความเบี่ยงเบนเนื่องจากความถดถอย

และ $\hat{y} = \hat{Y} - \bar{Y}$

$y = Y - \bar{Y}$

1/ ดูรายละเอียดได้จากหนังสือ Principles and Procedures of Statistics ของ Steel & Torrie หน้า 290 หัวข้อ 14.10

เมื่อ $y = Y-\bar{Y}$

$= (\hat{Y}-\bar{Y}) + (Y-\hat{Y})$

$y = \hat{y} + e$

ดังนั้น $\sum y^2 = \sum \hat{y}^2 + \sum e^2$

จะเห็นว่าผลรวมกำลังสองของ deviation Y จาก mean ($\sum y^2$) ประกอบด้วย

ส่วนที่ 1 $\sum \hat{y}^2$ ซึ่งเรียกว่า sum of squares เนื่องมาจากความถดถอยอันเป็นผลรวมกำลังสองของ deviation Y จาก mean สำหรับข้อมูลที่มีตัวแปรอิสระ 3 ตัว

$$\sum \hat{y}^2 = b_1 \sum x_1 y + b_2 \sum x_2 y + b_3 \sum x_3 y$$

ส่วนที่ 2 $\sum e^2$ สิ่งเป็นผลรวมกำลังสองของ deviation Y ค่าจริงจากค่า $\hat{Y}$ ที่คำนวณได้

ตาราง Analysis of Multiple Regression

แหล่งความแปรปรวน	Degrees of freedom	Sum of Squares จากนิยาม	Sum of Squares สูตรในการคำนวน	Mean Square	F
Regression	k	$\sum(\hat{Y}-\bar{Y})^2 = \sum \hat{y}^2$	$b_1 \sum x_1 y \ldots + b_k \sum x_k y$	$\frac{SS}{k} = A$	A/B
Deviation	n-k-1	$\sum(Y-\hat{Y})^2 = \sum e^2$	Total SS-Regression SS	$\frac{SSR}{n-k-1} = B$	
Total	n-1	$\sum(Y-\bar{Y})^2 = \sum y^2$	$\sum Y^2 - \frac{(\sum Y)^2}{n}$		

เมื่อ n = จำนวนข้อมูล ,k = จำนวนตัวแปรอิสระ(Independent Variable)

จาก F-test โดยการเปรียบเทียบค่า F ที่คำนวนได้กับ F(k,n-k-1) จากตาราง

ณ ระดับความเชื่อมั่น α =.05,.01 ทำให้เราสามารถสรุปผลการวิเคราะห์ว่า X_1, X_2, X_3 มีอิทธิพลต่อ Y อย่างมีนัยสำคัญ หรือไม่

Standard Errors และการทดสอบค่า b_i

สูตร $s_{y.1...k} = \sqrt{\frac{\sum(Y-\hat{Y})^2}{n-k-1}}$

ซึ่ง $s_{y.1...k}$ เป็นค่า standard error ของการประมาณค่า $\hat{Y}$ ซึ่งเป็นค่าของ Mean square of deviation ในตาราง Analysis of Multiple Linear Regression

$$S_{bi} = \sqrt{c_{ii} s^2_{y.1...k}}$$

เมื่อ S_{bi} เป็นค่า Standard error ของ b_i

c_{ii} เป็นค่าของ diagonal ของ variance covariance Matrix เพื่อทดสอบสมมติฐาน $H_o : B_i = B_{io} = 0; i=1,2,3,...,k$

การหาค่า t., $t = \frac{b_i - B_{io}}{S_{bi}}$ degree of freedom = n-k-1

$$= \frac{b_i}{S_{bi}}$$

เมื่อ b_i เป็นค่าของ partial regression coefficient ที่ต้องการทดสอบ นำค่า t ที่คำนวณได้เปรียบเทียบกับ t ที่เปิดจากตารางที่ α=.01,.05 ด้วย df = n-k-1

แล้วสรุปผลที่จะยอมรับหรือปฏิเสธสมมติฐาน $H_o : B_i = B_{io} = o$

Partial และ Multiple Correlation

โดยปกติสหสัมพันธ์แบบง่าย (Simple Correlation) ใช้สัญลักษณ์ r_{xy} เพื่อแสดงความสัมพันธ์ของตัวแปร (Variable) 2 ตัว x และ y

โดยที่ $r_{xy} = \sqrt{b_{y.x} \cdot b_{x.y}}$

$b_{y.x}$ คือ Simple Regression Coeffcient of y on x

$b_{x.y}$ คือ Simple Regression Coefficient of x on y

r มีค่าอยู่ระหว่าง -1 กับ 1 เครื่องหมายของ r จะแสดงทิศทางของความสัมพันธ์ระหว่างตัวแปรที่ต้องการหาว่า มีความสัมพันธ์อย่างไร ความสัมพันธ์ตามกัน ;r มีเครื่องหมาย + แสดงว่าเมื่อ x มีค่าเพิ่ม y มีค่าเพิ่มขึ้นด้วย ความสัมพันธ์กลับกัน ;r มีเครื่องหมาย - แสดงว่าเมื่อ x มีค่าลดลง y จะมีค่าเพิ่มขึ้น และเมื่อ y มีค่าลดลง x จะมีค่าเพิ่มขึ้น

ค่าสัมบูรณ์ของ r ($|r|$)ที่ใกล้ 1 แสดงว่าข้อมูลทั้งสองมีความสัมพันธ์กันมาก

" ($|r|$)ที่ใกล้ 0 " " " น้อย

<u>การหา Partial Correlation</u>

ในกรณีที่มีตัวแปร(variable) ตั้งแต่ 3 ตัวขึ้นไป จนเห็นว่าโดยธรรมดาจะมีสหสัมพันธ์แบบง่าย r_{12}, r_{13} และ r_{23}

โดยที่ r_{12} เป็นสหสัมพันธ์ระหว่าง variable ที่ 1 และ 2

r_{13} " " " 1 และ 3

r_{23} เป็นสหสัมพันธ์ระหว่าง variable ที่ 2 และ 3

สำหรับกรณีที่ variable ทั้ง 3 ตัวผันแปรร่วมกันแบบ Normal Distribution เราคำนวณ Partial Correlation Coefficient $r_{12.3}$, $r_{13.2}$ หรือ $r_{23.1}$

เมื่อ $r_{12.3}$ เป็นสหสัมพันธ์ระหว่าง variable 1 และ 2 เมื่อ variable ที่ 3 คงที่ ที่ค่าใดค่าหนึ่ง โดยทฤษฎี $r_{12.3}$ จะเท่ากันที่ทุก ๆ ค่าของ variable ที่ 3

$r_{13.2}$ และ $r_{12.3}$ ก็มีความหมายในทำนองเดียวกัน

โดยที่ $r_{12.3} = \dfrac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1-r_{13}^2)\,(1-r_{23}^2)}}$

กำหนดให้ $r_{12.3}$ เป็นความสัมพันธ์ระหว่าง variable ตัวที่ 1 และตัวที่ 2 เมื่อ variable ตัวที่ 3 มีค่าคงที่ ค่าใดค่าหนึ่ง เรียกว่า PARTIAL CORRELATION

โดยทฤษฎี $r_{12.3}$ จะเท่ากันทุก ๆ ค่าของ variable ค่าของ r และเครื่องหมาย มีความหมายเช่นเดียวกับ SIMPLE CORRELATION

การหา MULTIPLE CORRELATION

MULTIPLE CORRELATION โดยทั่วไปเป็นสหสัมพันธ์ระหว่าง Y และ $X_1, X_2, X_3, \ldots X_k$ หรือเป็นสหสัมพันธ์อย่างง่ายระหว่าง Y กับ $\hat{Y}$ ใช้สัญลักษณ์ R

$$R = r_{y\hat{y}}$$

$$= \frac{COV\ (y, \hat{y})}{\sqrt{V(y)\ V\ (\hat{y})}} \quad , \ y = \hat{y} + e$$

$0 \leqslant R \leqslant 1$; ค่า R ที่ได้จะแสดงว่า การ estemated ค่า y นั้นทำได้ดีเพียงไร

* * * * * * *