บทที่ 2

ทฤษฎีที่ใช้ในการวิจัย

สมการถดถอยซึ่งประมาณโดยวิธี OLS ที่นำมาเปรียบเทียบมี 6 สมการด้วยกัน คือ

1. สมการถดถอย เมื่อไม่มีข้อมูลสูญหายเลยโดยจะเรียกว่าวิธีวิเคราะห์สมบูรณ์
2. สมการถดถอย เมื่อมีข้อมูลสูญหายโดยตัดชุดข้อมูลที่มีค่าสูญหายออกโดยจะเรียกว่าวิธีสูญหาย
3. สมการถดถอย เมื่อมีข้อมูลสูญหายโดยประมาณข้อมูลสูญหายจากสมการถดถอยเชิงเส้นโดยจะเรียกว่าวิธีวิเคราะห์ความถดถอย
4. สมการถดถอย เมื่อใช้ตัวประมาณ maximum likelihood ประมาณข้อมูลสูญหายโดยจะเรียกว่า วิธี MAXIMUM LIKELIHOOD
5. สมการถดถอย เมื่อใช้ค่าเฉลี่ยประมาณข้อมูลสูญหาย โดยจะเรียกว่าวิธีค่าเฉลี่ย
6. สมการถดถอย เมื่อใช้ค่ามัธยฐานประมาณข้อมูลสูญหาย โดยจะเรียกว่าวิธีค่ามัธยฐาน

<u>การประมาณค่าสัมประสิทธิ์ความถดถอยด้วยวิธีกำลังสองน้อยที่สุด</u> (ordinary least square : OLS)

วิธีการหาค่าประมาณพารามิเตอร์โดยวิธี OLS เป็นวิธีที่มีรากฐานมาจากทฤษฎีการประมาณค่าเชิงเส้นซึ่งเป็นวิธีที่คิดค้นโดย คาร์ล เฟรดริช เกาส์ (Karl Friedrich Gauss 1777 - 1855) และอังเดร แอนดรีวิช มาร์คอฟ (Andrei Andreevich Markov 1856 - 1922) โดยมีหลักเกณฑ์ดังนี้คือให้หาค่าตัวประมาณค่าพารามิเตอร์ที่ทำให้ผลบวกกำลังสองของผลต่างของค่าสังเกตกับค่าคาดหวังของตัวแปรมีค่าต่ำที่สุด

รูปแบบจำลองทั่วไปของสมการถดถอยเชิงเส้นคือ

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{e} \qquad \ldots (2.3.1)$$

เมื่อ $\underset{\sim}{Y}$ คือ เมตริกซ์ของตัวแปรตามขนาด $N \times 1$

X คือ เมตริกซ์ของตัวแปรอิสระขนาด $N \times M$

$\underset{\sim}{\beta}$ คือ เมตริกซ์สัมประสิทธิ์ความถดถอยเชิงเส้นขนาด $M \times 1$

$\underset{\sim}{e}$ คือ เมตริกซ์ของความคลาดเคลื่อนขนาด $N \times 1$ โดยที่

$\underset{\sim}{e} \sim N(0, \sigma^2 I_N)$ และ

N คือ ขนาดตัวอย่าง

M คือ จำนวนตัวแปรอิสระ + 1

ตัวประมาณค่าสัมประสิทธิ์ความถดถอยเชิงเส้น โดยวิธีกำลังสองน้อยที่สุดของ $\underset{\sim}{\beta}$ คือ $\hat{\underset{\sim}{\beta}}$ ที่ทำให้ผลบวกกำลังสองของความคลาดเคลื่อน (sum of squares error) มีค่าน้อยที่สุด

จากแบบจำลอง (2.3.1) ผลรวมของความคลาดเคลื่อนกำลังสองคือ

$$
\begin{aligned}
S(\beta) = \underset{\sim}{e}'\underset{\sim}{e} &= (\underset{\sim}{Y} - X\underset{\sim}{\beta})'(\underset{\sim}{Y} - X\underset{\sim}{\beta}) \\
&= \underset{\sim}{Y}'\underset{\sim}{Y} - \underset{\sim}{\beta}'X'\underset{\sim}{Y} - \underset{\sim}{Y}'X\beta + \underset{\sim}{\beta}'\underset{\sim}{X}'X\beta \\
&= \underset{\sim}{Y}'\underset{\sim}{Y} - 2\underset{\sim}{\beta}'X'\underset{\sim}{Y} + \underset{\sim}{\beta}'X'X\underset{\sim}{\beta} \qquad \ldots(2.3.2)
\end{aligned}
$$

การประมาณค่าสัมประสิทธิ์ความถดถอยโดยใช้วิธีกำลังสองน้อยที่สุด จากตัวแบบ (2.3.1) อาศัยวิธีการทางแคลคูลัสโดยการดิฟเฟอเรนชิเอท (Differentiate) สมการ (2.3.2) เทียบกับ β แล้วให้เท่ากับศูนย์

$$
\left.\frac{\partial S}{\partial \underset{\sim}{\beta}}\right|_{\hat{\underset{\sim}{\beta}}} = -2X'\underset{\sim}{Y} + 2X'X\underset{\sim}{\beta} = 0
$$

$$
X'X\underset{\sim}{\beta} = X'\underset{\sim}{Y} \qquad \ldots(2.3.3)
$$

สมการ (2.3.3) เรียกว่าสมการปกติและตัวประมาณกำลังสองน้อยที่สุดคือ

$$
\hat{\underset{\sim}{\beta}} = (X'X)^{-1}X'\underset{\sim}{Y} \qquad \ldots(a)
$$

ดังนั้นสมการถดถอยที่ใช้ประมาณ คือ

$$\hat{\underset{\sim}{Y}} = X\hat{\underset{\sim}{\beta}}$$

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

รายละเอียดเกี่ยวกับวิธีประมาณค่าสูญหายในการวิเคราะห์การถดถอยแต่ละวิธีเป็นดังนี้

2.1 <u>วิธีวิเคราะห์สมบูรณ์</u>

จะทำการประมาณค่าสัมประสิทธิ์ความถดถอยจากชุดข้อมูลครบถ้วนโดยวิธี OLS ดังนี้

$$\hat{\underset{\sim}{\beta}} = (X'X)^{-1}(X'\underset{\sim}{Y})$$

2.2 <u>วิธีสูญหาย</u>

จะทำการประมาณค่าสัมประสิทธิ์ความถดถอยจากชุดข้อมูลที่ตัดค่าสูญหายออกแล้วโดยวิธี OLS ดังนี้

$$\hat{\underset{\sim}{\beta}} = (X^{*\prime}X^{*})^{-1}(X^{*\prime}\underset{\sim}{Y}^{*})$$

เมื่อ X^{*} และ Y^{*} คือชุดข้อมูลที่ตัดค่าสูญหายทิ้ง

2.3 <u>วิธีวิเคราะห์ความถดถอย</u>

จะทำการประมาณค่าสูญหายของข้อมูลตัวแปรอิสระตามขั้นตอนต่อไปนี้

2.3.1 ตัดข้อมูลสูญหายทิ้งแล้วประมาณค่าสัมประสิทธิ์ความถดถอยโดยวิธี OLS ดังนี้

$$\hat{\beta}^{(LS)} = (X'X)^{-1}(X'\underset{\sim}{Y})$$

2.3.2 นำสมการถดถอยที่ได้จาก 2.3.1 มาประมาณข้อมูลสูญหายโดยพิจารณาจากสมการถดถอยเชิงเส้น

$$\hat{Y}_{j} = \hat{\alpha} + \hat{\beta}_{i}^{(LS)} X_{ij}$$

ดังนั้น $$X_{ij}^{(EST)} = \frac{\hat{Y}_{j} - \hat{\alpha}}{\hat{\beta}_{i}^{(LS)}}$$

เมื่อ $$\hat{\alpha} = \bar{Y} - \bar{X}_{i}\hat{\beta}_{i}^{(LS)}$$

2.3.3 นำค่า $X_{ij}^{(EST)}$ แทนในค่าสูญหายของแต่ละตัวแปรอิสระ แล้วทำการประมาณค่าสัมประสิทธิ์ความถดถอยใหม่ด้วยวิธี OLS

$$\underset{\sim}{\hat{\beta}} = (X'X)^{-1}(X'\underset{\sim}{Y})$$

ดังนั้นจะได้สมการถดถอยที่จะใช้ประมาณ คือ

$$\underset{\sim}{\hat{Y}} = X\underset{\sim}{\hat{\beta}}$$

2.4 <u>วิธี MAXIMUM LIKELIHOOD</u>

Anderson ได้ทำการประมาณตัวพารามิเตอร์ของตัวแปรที่มีการแจกแจงแบบปกติและมีข้อมูลบางตัวของตัวแปรอิสระสูญหายโดยวิธี maximum likelihood ดังนี้

ถ้าข้อมูลมีลักษณะ

$$y_1, \ldots, y_n, y_{n+1}, \ldots, y_N$$
$$x_1, \ldots, x_n$$

Bivariate density function ของ x และ y สามารถเขียนในรูป

$$n(y, x \mid \mu_y, \mu_x ; \sigma^2_y, \sigma^2_x ; \rho) = n(y \mid \mu_y', \sigma^2_y) \cdot n(x \mid \alpha + \beta_{xy} y, \sigma^2_{xy})$$

เมื่อ
$$\left.\begin{aligned} \gamma &= \mu_x - \beta_{xy}\mu_y \\ \beta_{xy} &= \rho\sigma_x / \sigma_y \\ \sigma^2_{x.y} &= \sigma^2_x(1-\rho^2) \end{aligned}\right\} \qquad \text{... (A)}$$

Likelihood function คือ

$$\prod_{\alpha=1}^{n} n(y_\alpha, x_\alpha \mid \mu_y, \mu_x ; \sigma^2_y, \sigma^2_x ; \rho) \prod_{\alpha=1}^{N} n(y_\alpha \mid \mu_y, \sigma^2_y)$$

$$= \prod_{\alpha=1}^{N} n(y_\alpha \mid \mu_y, \sigma^2_y) \prod_{\alpha=1}^{n} n(x_\alpha \mid \gamma + \beta_{xy}y_\alpha, \sigma^2_{x.y}) \qquad \text{....(1)}$$

ตัวประมาณ maximum likelihood ของ μ_y, σ^2_y, γ, β_{xy} และ $\sigma^2_{x.y}$ สามารถหาได้โดยการ differentiate สมการที่ (1) เทียบกับ μ_y, σ^2_y, γ, β_{xy}, $\sigma^2_{x.y}$ แล้วเทียบให้เท่ากับ 0 ซึ่งจะได้ผลลัพธ์ดังนี้

$$\hat{\mu}_y = \bar{y} = \sum_{\alpha=1}^{N} y_\alpha / N$$

$$\hat{\sigma}^2_y = \sum_{\alpha=1}^{N} (y_\alpha - \bar{y})^2 / N$$

$$\hat{\beta}_{xy} = \frac{\sum_{\alpha=1}^{n} (x_\alpha - \bar{x}^*)(y_\alpha - \bar{y}^*)}{\sum_{\alpha=1}^{n} (y_\alpha - \bar{y}^*)^2}$$

$$\hat{\gamma} = \bar{x}^* - \hat{\beta}_{xy}\bar{y}^*$$

$$\hat{\sigma}^2_{x.y} = \frac{\sum_{\alpha=1}^{n} (x_\alpha - \bar{x}^*) - \hat{\beta}^2_{xy}\sum_{\alpha=1}^{n} (y_\alpha - \bar{y}^*)^2}{n}$$

$$\bar{y}^* = \sum_{\alpha=1}^{n} y_\alpha / n \quad , \quad \bar{x}^* = \sum_{\alpha=1}^{n} y_\alpha / n$$

เพราะฉะนั้น ตัวประมาณ maximum likelihood ของ μ_x, σ^2_x และ ρ หาได้จากการแทนค่า $\rho = \hat{\rho}$, $\beta_{xy} = \hat{\beta}_{xy}$, $\sigma^2_{xy} = \hat{\sigma}^2_{xy}$ ในสมการ (A) และแก้สมการได้ดังนี้

$$\hat{\mu}_x = \bar{x}^* + \beta_{xy}(y - \bar{y}^*)$$

$$\hat{\sigma}^2_x = \frac{\sum_{\alpha=1}^{n}(x_\alpha - \bar{x}^*)^2}{n} + \hat{\beta}^2_{xy}\left[\hat{\sigma}^2_y - \frac{\sum_{\alpha=1}^{n}(y_\alpha - \bar{y}^*)^2}{n}\right]$$

$$\hat{\rho} = \hat{\beta}_{xy}\, \hat{\sigma}_y / \hat{\sigma}_x$$

ถ้าในกรณีที่มีตัวแปรอิสระมากกว่า 2 ตัวแปร และมีข้อมูลอยู่ในลักษณะดังนี้

Y_1	X_{11}		X_{31}	
Y_2	X_{12}		X_{32}	
.	.		.	
.	.		.	
Y_n	X_{1n}		X_{3n}	
Y_{n+1}	X_{1n+1}			X_{4n+1}
.	.			.
.	.			.
Y_{n+m}	X_{1n+m}			X_{4n+m}
Y_{n+m+1}		X_{2n+m+1}		
.		.		
.		.		
Y_N		X_{2N}		

การประมาณตัวพารามิเตอร์ μ_{x1}, μ_{x2}, μ_{x3}, μ_{x4} โดยวิธี maximum likelihood จะทำในทำนองเดียวกับวิธีดังกล่าวข้างต้นโดยพิจารณาตัวแปรทีละคู่

ในการวิจัยครั้งนี้จะนำค่าประมาณพารามิเตอร์ไปแทนในข้อมูลสูญหายแล้วจึงทำการประมาณสมการถดถอยด้วยวิธี OLS

2.5 <u>วิธี OLS</u> เมื่อพิจารณาจาก ค่าเฉลี่ย ซึ่งมีขั้นตอนการดำเนินการดังนี้

<u>ขั้นที่ 1</u> ประมาณค่าสูญหายของแต่ละตัวแปรโดยใช้ค่าเฉลี่ยของข้อมูลที่ไม่สูญหายของแต่ละตัวแปรนั้น ๆ

$$\bar{x}_i = \frac{\sum_{j=1}^{k_i} x_{ij}}{k_i} \quad ; i = 1, 2, \ldots, M$$

โดยที่ $\bar{x}_i$ คือ ค่าเฉลี่ยของข้อมูลที่มีอยู่ของตัวแปรที่ i

k_i คือ จำนวนข้อมูลที่ไม่สูญหายของตัวแปรที่ i

M คือ จำนวนตัวแปรอิสระ

x_{ij} คือ ค่าของข้อมูลที่ไม่สูญหายของตัวแปรที่ i ค่าสังเกตที่ j

<u>ขั้นที่ 2</u> นำค่าเฉลี่ยแทนในข้อมูลที่สูญหายแล้วทำการประมาณสัมประสิทธิ์ความถดถอยด้วยวิธีกำลังสองน้อยที่สุด

2.6 <u>วิธี OLS</u> เมื่อพิจารณาจาก ค่ามัธยฐาน ซึ่งมีขั้นตอนการดำเนินการ ดังนี้

<u>ขั้นที่ 1</u> ประมาณค่าสูญหายของแต่ละตัวแปรอิสระโดยใช้ค่ามัธยฐานของข้อมูลที่มีอยู่ของแต่ละตัวแปรนั้น ๆ โดยแยกพิจารณา 2 กรณี

1. เมื่อตัวแปรอิสระมีจำนวนข้อมูลที่ไม่สูญหายเป็นเลขคู่

$$X_{M_i} = \frac{1}{2}\left[X_{(i,\frac{k_i}{2})} + X_{(i,\frac{k_i}{2}+1)} \right]$$

โดยที่ $X_{(i,1)} < X_{(i,2)} < \ldots < X_{(i,k_i)}$

เมื่อ X_{M_i} คือ ค่ามัธยฐานของข้อมูลที่มีอยู่ของตัวแปรที่ i

k_i คือ จำนวนข้อมูลที่ไม่สูญหายของตัวแปร X_i

M คือ จำนวนตัวแปรอิสระ

$X_{(i,\frac{k_i}{2})}$ คือ ค่าของข้อมูลที่ไม่สูญหายของตัวแปรที่ i ค่าสังเกตที่ $\frac{k_i}{2}$

$X_{(i,\frac{k_i}{2}+1)}$ คือ ค่าของข้อมูลที่ไม่สูญหายของตัวแปรที่ i ค่าสังเกตที่ $\frac{k_i}{2}+1$

2. เมื่อตัวแปรอิสระมีจำนวนข้อมูลที่ไม่สูญหายเป็นเลขคี่

$$X_{M_i} = X_{(i,\frac{k_i+1}{2})}$$

โดยที่ $X_{(i,1)} < X_{(i,2)} < \ldots < X_{(i,k_i)}$

เมื่อ X_{M_i} คือ ค่ามัธยฐานของข้อมูลที่มีอยู่ของตัวแปรที่ i

k_i คือ จำนวนข้อมูลที่ไม่สูญหายของตัวแปร X_i

M คือ จำนวนตัวแปรอิสระ

$X_{(i,\frac{k_i+1}{2})}$ คือ ค่าของข้อมูลที่ไม่สูญหายของตัวแปรที่ i ค่าสังเกตที่ $\frac{k_i+1}{2}$

<u>ขั้นที่ 2</u> นำค่ามัธยฐานแทนในข้อมูลสูญหายแล้วทำการประมาณค่าสัมประสิทธิ์ความถดถอยด้วยวิธีกำลังสองน้อยที่สุด

<u>เกณฑ์ที่ใช้ในการเปรียบเทียบสมการถดถอยเชิงซ้อน</u> เมื่อใช้การประมาณข้อมูลสูญหายของตัวแปรอิสระโดยวิธีต่าง ๆ คือ ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) ของสมการถดถอยเมื่อไม่มีข้อมูลสูญหาย

ค่าคลาดเคลื่อนกำลังสองเฉลี่ย (mean square error : MSE) คำนวณจาก

$$MSE = \frac{\sum_{j=1}^{N} (Y_{ij} - \hat{Y}_{ij})^2}{N-M-1}$$

โดยที่ Y_{ij} คือ ค่าจริงของข้อมูลของตัวแปรตามตัวที่ i

$\hat{Y}_{ij}$ คือ ค่าประมาณของข้อมูลของตัวแปรตามตัวที่ i

N คือ ขนาดตัวอย่าง

M คือ จำนวนตัวแปรอิสระที่ใช้ในการวิเคราะห์

นอกจากการใช้เกณฑ์ดังกล่าวข้างต้น ในการวิจัยครั้งนี้ยังใช้อัตราส่วนของความคลาดเคลื่อนรวมของค่าประมาณที่ได้จากสมการถดถอยเมื่อแทนค่าข้อมูลสูญหายของตัวแปรอิสระด้วยวิธีต่าง ๆ เมื่อเทียบกับวิธีที่ 1 ซึ่งอัตราส่วนดังกล่าวสามารถบอกได้ถึงความใกล้เคียงของค่าประมาณของตัวแปรตามของชุดข้อมูลครบถ้วน และชุดข้อมูลที่ได้ตัดค่าสูญหายทิ้งและประมาณข้อมูลสูญหายในการวิเคราะห์การถดถอยด้วยวิธีการต่าง ๆ ซึ่งคำนวณจาก

อัตราส่วนของค่าคลาดเคลื่อนรวมของวิธีที่ i เมื่อเทียบกับวิธีที่ 1 $= \sum_{j=1}^{N} \left| (\hat{Y}_{ij} - \hat{Y}_{1j}) / \hat{Y}_{1j} \right|$

เมื่อ i = 2, 3, 4, 5, 6

บทที่ 3

วิธีดำเนินการวิจัย

การวิจัยครั้งนี้ต้องการศึกษาเปรียบเทียบวิธีการประมาณค่าสูญหายในการวิเคราะห์การถดถอยเมื่อมีสมการถดถอย 6 สมการ คือ

ก. สมการถดถอย เมื่อไม่มีข้อมูลสูญหายเลย

ข. สมการถดถอย เมื่อมีข้อมูลสูญหายโดยตัดชุดข้อมูลที่มีค่าสูญหายออก

ค. สมการถดถอย เมื่อประมาณข้อมูลสูญหายจากสมการถดถอยเชิงเส้น

ง. สมการถดถอย เมื่อประมาณข้อมูลสูญหายด้วยวิธี MAXIMUM LIKELIHOOD

จ. สมการถดถอย เมื่อใช้ค่าเฉลี่ยประมาณข้อมูลสูญหาย

ฉ. สมการถดถอย เมื่อใช้ค่ามัธยฐานประมาณข้อมูลสูญหาย

โดยจะพิจารณาเปรียบเทียบค่าคลาดเคลื่อนกำลังสองเฉลี่ยของสมการถดถอยเมื่อได้ประมาณข้อมูลสูญหายด้วยวิธีต่าง ๆ กับของสมการถดถอยเมื่อไม่มีข้อมูลสูญหายเมื่อประชากรมีการแจกแจงปกติ (normal distribution) ที่มีการกระจาย 3 ระดับคือ C.V. = 0.05, 0.20 และ 1.00, ขนาดตัวอย่างที่สนใจในการศึกษามี 3 ขนาด คือ 30, 70 และ 100, ค่าเบี่ยงเบนมาตรฐานมี 4 ระดับคือ 5, 10, 20 และ 25 ,จำนวนตัวแปรมี 4 ขนาดคือ 2, 3, 5 และ 7 และการสูญหายของข้อมูลมี 3 ระดับ คือ 5%, 10% และ 15% ทั้งนี้เทคนิคที่ใช้ในการสร้างแบบจำลองข้อมูล คือ วิธีมอนติคาร์โล

เนื่องจากวิธีมอนติคาร์โลเป็นเทคนิคที่ใช้ในการวิจัยครั้งนี้ ดังนั้นในตอนแรกของบทนี้จะกล่าวถึงวิธีมอนติคาร์โลก่อน แล้วจึงเป็นขั้นตอนการวิจัย และโปรแกรมที่ใช้ในการวิเคราะห์สมการถดถอยเชิงซ้อนตามลำดับ

3.1 วิธีมอนติคาร์โล (Monte Carlo Method)

เทคนิคที่ใช้สำหรับแก้ปัญหาในการคำนวณทางคณิตศาสตร์นั้นมีอยู่หลายวิธี วิธีมอนติคาร์โลเป็นเทคนิคอย่างหนึ่งที่ใช้แก้ปัญหานี้ Hammevsley และ Handscomb (1964 : 2) กล่าวว่า