

บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันนี้ไม่ว่าจะเป็น ด้านเศรษฐศาสตร์ ด้านธุรกิจ ด้านสังคมศาสตร์ หรือแม้แต่ทางการแพทย์ล้วนแล้วแต่จะต้องอาศัยวิธีการทางสถิติมาเป็นเครื่องมือในการวิเคราะห์ข้อมูลเพื่อประกอบการตัดสินใจทั้งสิ้น โดยการเลือกใช้วิธีการทางสถิตินั้นจะต้องพิจารณาจากชนิดของข้อมูล วัตถุประสงค์ที่ต้องการและข้อสมมติเบื้องต้น(Basic Assumption) อีกด้วย เพื่อผลการวิเคราะห์ทางสถิติที่ถูกต้องและแม่นยำ

แต่ในการวิจัยจำนวนไม่น้อยจะพบว่าตัวแปรตาม(Dependent Variable)ที่สนใจศึกษามักมีลักษณะข้อมูลเป็นเชิงคุณภาพ กล่าวคือข้อมูลไม่ต่อเนื่อง เช่น การเป็นโรคหัวใจหรือการไม่เป็นโรคหัวใจ การสอบได้หรือสอบตก และการเลือกซื้อสินค้าหรือไม่เลือกซื้อสินค้า เป็นต้น ดังนั้นตัวแบบโพรบิตแบบ 2 ประเภท (Binary Probit Model) จึงเป็นอีกตัวแบบที่นิยมกันแพร่หลายสำหรับตัวแบบที่มีข้อจำกัดเกี่ยวกับตัวแปรตามที่เป็นตัวแปรเชิงคุณภาพมีค่าได้เพียง 2 ค่า (Dichotomy or Binary Variable) เท่านั้น คือค่า 0 และ 1 โดยกำหนดให้ตัวแปรตามมีค่าเท่ากับ 1 เมื่อเกิดเหตุการณ์ที่สนใจในหน่วยที่ i และตัวแปรตามมีค่าเท่ากับ 0 เมื่อไม่เกิดเหตุการณ์ที่สนใจในหน่วยที่ i และยิ่งเนื่องมาจากการวิเคราะห์การถดถอยที่มีข้อจำกัดเกี่ยวกับตัวแปรตามลักษณะดังกล่าวข้างต้นนี้จะไม่สามารถนำมาทำการวิเคราะห์การถดถอยโดยใช้ตัวแบบความน่าจะเป็นเชิงเส้น (Linear Probability Model) เพื่อประมาณค่าความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจได้ ด้วยวิธีกำลังสองน้อยที่สุด (Method of Least Squares) ได้ เนื่องจากข้อสมมติเบื้องต้นของตัวแบบไม่เป็นไปตามข้อกำหนด คือ การแจกแจงของค่าความคลาดเคลื่อนไม่เป็นแบบปกติ (Nonnormality) ค่าความแปรปรวนของพจน์ความคลาดเคลื่อน(Error Terms)มีลักษณะแตกต่างกัน (Heteroscedastic) ค่าประมาณของการเกิดเหตุการณ์ที่สนใจอาจอยู่นอกช่วง 0 ถึง 1 และการประมาณค่าความชัน (Slope) ที่สูงเกินจริง (Overestimated Slope) หรือต่ำเกินจริง (Underestimated Slope) (Guajarati ,1995,Pindyck and Rubinfeld,1998,Johnston and Dinardo,1997) อีกด้วย

ตัวแบบโพรบิตแบบ 2 ประเภท เป็นตัวแบบที่นำมาใช้วิเคราะห์ความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ แล้วยังสามารถนำมาใช้ในการพยากรณ์ตัวแปรตามเชิงคุณภาพว่าจะอยู่ในกลุ่มใดใน 2 กลุ่ม คือ กลุ่มที่เกิดเหตุการณ์ที่สนใจ (Success) และกลุ่มไม่เกิดเหตุการณ์ที่สนใจ (Failure) ของสิ่งที่สนใจศึกษา โดยส่วนใหญ่งานวิจัยเกือบทั้งหมดจะมีการจัดให้แต่ละหน่วยของสิ่งที่สนใจ อยู่ในกลุ่มใดกลุ่มหนึ่ง มักจะใช้จุดแบ่ง (Cut-off Point) หรือระดับความน่าจะเป็นที่

ระดับ 0.5 โดยให้เหตุผลว่ากลุ่มที่เกิดเหตุการณ์ที่สนใจมีโอกาสเกิดขึ้นเท่ากับ กลุ่มที่ไม่เกิดเหตุการณ์ที่สนใจของสิ่งที่สนใจศึกษา หรือบางงานวิจัยอาจใช้จุดแบ่งค่าหนึ่งที่จะทำให้สัดส่วนของความถูกต้องของการจัดกลุ่มทั้งหมดมีค่าสูงสุดหรือทำให้อัตราความผิดพลาดในการจำแนกกลุ่มมีค่าต่ำ (Classification Error Rate) มีค่าต่ำสุดโดยให้เหตุผลว่า ข้อมูลถูกเลือกอย่างสุ่มจากสิ่งที่สนใจศึกษาแล้ว ผู้วิจัยส่วนใหญ่มักไม่คำนึงถึงจำนวนของตัวแปรอิสระ, ขนาดตัวอย่าง, สัดส่วนของการเกิดเหตุการณ์ที่สนใจและไม่เกิดเหตุการณ์ที่สนใจในชุดข้อมูล และระดับความสัมพันธ์ระหว่างตัวแปรอิสระ (Multicollinearity) ซึ่งผู้วิจัยได้เห็นถึงความสำคัญที่จะทำการวิจัยการคัดเลือกจุดแบ่งของตัวแบบโพรบิตแบบ 2 ประเภทที่จะทำให้อัตราส่วนความผิดพลาดในการจำแนกกลุ่มมีค่าต่ำสุดหรือสัดส่วนความถูกต้องในการจำแนกกลุ่มมีความถูกต้องสูงสุดสำหรับชุดข้อมูลในสถานการณ์ต่างๆ โดยลักษณะเหล่านั้นของชุดข้อมูล คือ จำนวนของตัวแปรอิสระ, ขนาดตัวอย่าง, สัดส่วนของการเกิดเหตุการณ์ที่สนใจและการไม่เกิดเหตุการณ์ที่สนใจในชุดข้อมูลและระดับความสัมพันธ์ระหว่างตัวแปรอิสระ ซึ่งลักษณะเหล่านี้จะมีผลกระทบต่อจุดแบ่งสำหรับการประเมินการพยากรณ์การจำแนกกลุ่มของข้อมูลหรือไม่ ถ้ามีผลต่อการคัดเลือกจุดแบ่งแล้วรูปแบบเหล่านั้นคืออะไร ที่จะทำให้สามารถหาจุดแบ่งที่เหมาะสมสำหรับตัวแบบโพรบิตแบบ 2 ประเภทมากที่สุด

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อหาของจุดแบ่ง ที่เหมาะสมสำหรับการจำแนกกลุ่มของข้อมูลในตัวแบบโพรบิตแบบ 2 ประเภท เมื่อ จำนวนตัวแปรอิสระเปลี่ยนแปลงไป, ขนาดตัวอย่างเปลี่ยนแปลงไป, สัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจในแต่ละชุดข้อมูลเปลี่ยนแปลงไป, ระดับความสัมพันธ์ของตัวแปรอิสระเปลี่ยนแปลงไป และนำปัจจัยต่างๆเหล่านี้มาพิจารณารวมกัน ซึ่งแปรเปลี่ยนไปพร้อมกัน

2. เพื่อหาสมการถดถอยพหุคูณ (Multiple Regression Model) สำหรับใช้ในการประมาณค่าของจุดแบ่งที่เหมาะสมสำหรับการจำแนกกลุ่มของข้อมูลในตัวแบบโพรบิตแบบ 2 ประเภท ในสถานการณ์อื่นๆ ต่อไป



1.3 ขอบเขตของการวิจัย

1. ทำการศึกษาตัวแบบโพรบิตแบบ 2 ประเภทเพื่อหาจุดแบ่งที่เหมาะสมที่สุด
2. ตัวแปรตาม (Y) เป็นข้อมูลเชิงคุณภาพที่มี 2 ค่า คือ 0 และ 1 โดยกำหนดสัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจ(a) ของลักษณะที่สนใจศึกษา เท่ากับ 0.1 ,0.5 และ 0.9
3. จำนวนของตัวแปรอิสระ (p) ในการวิจัยครั้งนี้แบ่งเป็น 3 ระดับ คือ
 - จำนวนตัวแปรอิสระน้อย คือ 1 และ 2 ตัว
 - จำนวนตัวแปรอิสระปานกลาง คือ 3 และ 4 ตัว
 - จำนวนตัวแปรอิสระมาก คือ 5 และ 6 ตัว
4. ขนาดตัวอย่าง (n) ในการวิจัยครั้งนี้แบ่งเป็น 3 ระดับ คือ
 - ขนาดตัวอย่างเล็ก คือ 20 และ 40
 - ขนาดตัวอย่างปานกลาง คือ 60 และ 80
 - ขนาดตัวอย่างใหญ่ คือ 100 และ 120

โดยขนาดตัวอย่างและจำนวนตัวแปรอิสระที่ใช้ในการสร้างชุดข้อมูลจะขึ้นอยู่กับ

ความสัมพันธ์ ดังนี้ $\frac{n}{p} \geq 20$

5. ระดับความสัมพันธ์ระหว่างตัวแปรอิสระ (M)
 - ตัวแปรอิสระไม่มีความสัมพันธ์กัน ($p = 0$)
 - ความสัมพันธ์ระหว่างตัวแปรอิสระในระดับต่ำ ($p = 0.33$)
 - ความสัมพันธ์ระหว่างตัวแปรอิสระในระดับปานกลาง ($p = 0.67$)
 - ความสัมพันธ์ระหว่างตัวแปรอิสระในระดับสูง ($p = 0.99$)

โดยกำหนดให้รูปแบบเมทริกซ์สหสัมพันธ์ (Correlation Matrix) คือ

$$\rho_{p \times p} = \begin{pmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & \rho_{pp} \end{pmatrix} = \begin{pmatrix} 1 & \rho & \cdots & \rho^{p-1} \\ \rho & 1 & \cdots & \rho^{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \cdots & 1 \end{pmatrix}$$

โดยที่ ρ_{ij} ; $i, j = 1, 2, \dots, p$ คือสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระตัวที่ i

และตัวแปรอิสระตัวที่ j ซึ่งจะมีทั้งหมด $\frac{p(p-1)}{2}$ คู่

- 6. กำหนดค่าพารามิเตอร์เริ่มต้นของสมการการถดถอยเป็นค่าใดๆ ในการวิจัยครั้งนี้กำหนดให้ $\beta_i = 0.1; i = 0,1,2,...,p$ และ $\varepsilon_i \sim N(0,25); i = 1,2,...,n$
- 7. กำหนดระดับนัยสำคัญ (α) ในการวิจัยครั้งนี้ที่ระดับ 0.05
- 8. ตัวแปรอิสระเริ่มต้นมีการแจกแจงแบบยูนิฟอร์ม
- 9. ในการวิจัยครั้งนี้ทำการจำลองข้อมูลโดยใช้เทคนิคมอนติคาร์โล (Monte Carlo SiMulation) โดยการจำลองในแต่ละสถานการณ์จะกระทำซ้ำ 500 รอบ

1.4 ข้อตกลงเบื้องต้น

สำหรับการวิจัยครั้งนี้ศึกษามาจากตัวอย่างที่จำลองขึ้นมาจากตัวแบบ และการแจกแจงดังนี้

- 1. ศึกษาตัวแบบโพรบิตแบบ 2 โดยมีรูปแบบ คือ
$$\begin{aligned}\pi_i &= P(Y_i = 1) \\ &= P(Y_i^* > 0) \\ &= P(\varepsilon_i < \beta'X_i) \qquad \qquad \qquad ; i=1,2,...,n \\ &= \int_{-\alpha}^{\beta X_i} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\varepsilon^2}{2}\right) d\varepsilon\end{aligned}$$

โดยที่ π_i คือ ค่าความน่าจะเป็นเมื่อเกิดเหตุการณ์ที่สนใจในหน่วยที่ i
 Y_i คือ ตัวแปรตามเชิงคุณภาพที่มีค่าได้เพียง 2 ค่า คือ 0 และ 1
 Y_i^* คือ ตัวแปรแฝง (Latent Variable) เป็นค่าที่วัดไม่ได้ จึงไม่ทราบค่าที่แท้จริงทราบเพียงแต่ผลที่เกิดขึ้น

$\Phi(\cdot)$ คือ ฟังก์ชันการแจกแจงสะสมปกติมาตรฐาน
 ε คือ ตัวแปรสุ่มปกติมาตรฐาน โดย $\varepsilon_i \overset{iid}{\sim} N(0,1)$
 β' คือ เวกเตอร์ของสัมประสิทธิ์การถดถอย จำนวน $k+1$ ตัว ขนาด $1 \times (k+1)$
 X_i คือ เวกเตอร์ของตัวแปรอิสระที่เป็นตัวแปรเชิงปริมาณจำนวน k ตัว ขนาด $(k+1) \times 1$

- 2. ตัวแปรอิสระเริ่มต้นมีการแจกแจงแบบยูนิฟอร์ม ที่มีฟังก์ชันความน่าจะเป็นของ x อยู่ในรูปของ

$$f(x)=\frac{1}{b-a} \quad ; a < x < b \quad \text{เมื่อ} \quad a < b \text{ และ } a, b \text{ เป็นค่าคงที่}$$

โดยมี

$$\text{ค่าเฉลี่ย } E(X) = \mu = \frac{a+b}{2}$$

$$\text{ค่าความแปรปรวน } \text{Var}(X) = \sigma^2 = \frac{(b-a)^2}{12}$$

1.5 คำจำกัดความที่ใช้ในการวิจัย

1. พารามิเตอร์ (Parameter) คือค่าคงที่ที่แสดงคุณลักษณะบางประการของประชากร
2. ค่าเฉลี่ย (Mean) คือ ค่าที่แสดงว่าข้อมูลส่วนใหญ่อยู่ที่ไหน
3. ความแปรปรวน (Variance) คือ ค่าที่บอกว่าค่าของตัวแปรสุ่ม จะมีความผิดพลาดกาลังสอง โดยเฉลี่ยจากค่าของข้อมูลส่วนใหญ่ เท่ากับเท่าไร กล่าวคือ เป็นค่าที่วัดแนวโน้มที่ตัวแปรสุ่ม จะห่างจากค่าเฉลี่ยของมัน
4. ตัวแบบโพรบิตแบบ 2 ประเภท (Binary Probit Model) หมายถึง ตัวแบบที่ศึกษาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ เมื่อตัวแปรตามเป็นตัวแปรเชิงคุณภาพที่มี 2 ลักษณะ คือมีค่าเท่ากับ 1 หรือ 0 และตัวแปรอิสระเป็นตัวแปรเชิงปริมาณหรือตัวแปรหุ่น เมื่อได้รูปแบบความสัมพันธ์ระหว่างตัวแปรแล้ว จะนำไปใช้ในการประมาณค่าตัวแปรตาม
5. จุดแบ่ง (Cut-off Point) หมายถึง ค่าความน่าจะเป็นที่ใช้ในการพิจารณาการจำแนกกลุ่มของข้อมูลว่าแต่ละหน่วยจะอยู่ในกลุ่มใดระหว่างกลุ่มของเหตุการณ์ที่สนใจ (Success) และกลุ่มไม่เกิดเหตุการณ์ที่สนใจ(Failure)
6. ข้อมูลดิบ (Ungrouped Data) คือ ข้อมูลที่เพิ่งได้จากการเก็บรวบรวมข้อมูล ยังไม่ได้จัดแบ่งกลุ่ม จำแนกกลุ่ม แยกประเภทและไม่อยู่ในรูปตารางความถี่
7. ตัวแปรแฝง (Latent Variable) เป็นค่าที่วัดไม่ได้ จึงไม่ทราบค่าที่แท้จริง ทราบเพียงแต่ผลที่เกิดขึ้น
8. ความสัมพันธ์ระหว่างตัวแปรอิสระ (Multicollinearity) คือสถานการณ์ที่ตัวแปรอิสระมีความสัมพันธ์กัน โดยแบ่งเป็น 5 ระดับ ดังนี้
 - ตัวแปรอิสระไม่มีความสัมพันธ์กัน
 - ตัวแปรอิสระมีความสัมพันธ์ในระดับต่ำ
 - ตัวแปรอิสระมีความสัมพันธ์ในระดับปานกลาง
 - ตัวแปรอิสระมีความสัมพันธ์ในระดับสูง
 - ตัวแปรอิสระมีความสัมพันธ์กันอย่างสมบูรณ์

9. การแจกแจงแบบเบอรฺนูลลี (Bernoulli Distribution) หมายถึงตัวแปรสุ่ม Y เรียกว่า ตัวแปรสุ่มเบอรฺนูลลี กล่าวคือ

ถ้า $Y = 0$ เมื่อ การทดลองไม่เกิดเหตุการณ์ที่สนใจ

และถ้า $Y = 1$ เมื่อ การทดลองเกิดเหตุการณ์ที่สนใจ

โดยที่ $P(Y = 1) = p$, $0 < p < 1$ และ $P(Y = 0) = 1 - p$

เราอาจเขียนแทนด้วย $Y \sim \text{Ber}(p)$ ซึ่งฟังก์ชันความน่าจะเป็นอยู่ในรูปของ

$$P(Y = y) = p^y (1 - p)^{1-y}$$

10. การประมาณค่าแบบช่วงหรือ ช่วงความเชื่อมั่น (Confidence Intervals) เป็นการประมาณค่าพารามิเตอร์ของประชากรในรูปแบบช่วงโดยใช้ข้อมูลตัวอย่าง การประมาณแบบช่วงนั้นจะบอกถึงค่าต่ำสุดและค่าสูงสุดของพารามิเตอร์ที่เป็นไปได้

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถหาจุดแบ่งที่เหมาะสมสำหรับการจำแนกกลุ่มของข้อมูลในตัวแบบโพบริตแบบ 2 ประเภท เมื่อ จำนวนตัวแปรอิสระเปลี่ยนแปลงไป, ขนาดตัวอย่างเปลี่ยนแปลงไป, สัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจในแต่ละชุดของข้อมูลเปลี่ยนแปลงไป, ระดับความสัมพันธ์ของตัวแปรอิสระเปลี่ยนแปลงไป และนำปัจจัยต่างๆเหล่านี้มาพิจารณารวมกัน ซึ่งแปรเปลี่ยนไปพร้อมกัน
2. สามารถหาสมการถดถอยพหุคูณ สำหรับใช้ในการประมาณค่าของจุดแบ่งที่เหมาะสมสำหรับการจำแนกกลุ่มของข้อมูลในตัวแบบโพบริตแบบ 2 ประเภท ในสถานการณ์อื่นๆ ต่อไป
3. สามารถใช้เพื่อเป็นแนวทางในการศึกษาวิจัยต่อไป

1.7 วิธีดำเนินการวิจัย

1. ศึกษาค้นคว้าเอกสารและข้อมูลที่เกี่ยวข้องกับการวิจัย
2. จำลองข้อมูลให้มีการแจกแจงตามลักษณะที่ต้องการศึกษาโดยใช้เทคนิคมอนติคาร์โล
3. คำนวณหาจุดแบ่งที่เหมาะสมที่สุดในแต่ละสถานการณ์สำหรับตัวแบบโพบริตแบบ 2 ประเภท ซึ่งทำให้อัตราความผิดพลาดในการจำแนกกลุ่มมีค่าต่ำสุดหรือสัดส่วนความถูกต้องในการจำแนกกลุ่มมีค่าสูงสุด
4. ทำการทดลองซ้ำ 500 รอบในแต่ละสถานการณ์
5. คำนวณหาค่าเฉลี่ยของจุดแบ่งที่เหมาะสมที่สุดเพื่อนำมาคำนวณค่าเปอร์เซ็นต์ (Percent) และช่วงความเชื่อมั่น

6. ใช้ตัวแบบการถดถอยพหุคูณ เพื่อประมาณค่าพารามิเตอร์ สำหรับใช้ในการหาจุดแบ่งที่เหมาะสมที่สุดในสถานการณ์อื่นๆ ต่อไป
7. สรุปผลที่ได้จากการวิจัย