บทที่ 2
แนวคิด ทฤษฎีและสถิติที่เกี่ยวข้อง

การหาจุดแบ่งที่เหมาะสมที่สุดสำหรับการพยากรณ์การจำแนกข้อมูลไม่จัดกลุ่ม สำหรับการวิจัยครั้งนี้จะใช้หาจุดแบ่งที่เหมาะสมที่สุดในตัวแบบโพรบิตแบบ 2 ประเภท ซึ่งจุดแบ่งที่เหมาะสมที่สุดจะให้ค่าอัตราความผิดพลาดในการจำแนกกลุ่มมีค่าต่ำสุด และใช้ตัวแบบการถดถอยพหุคูณเพื่อประมาณค่าพารามิเตอร์สำหรับใช้ในการประมาณค่าของจุดแบ่งที่เหมาะสมที่สุดในสถานการณ์อื่นๆ ต่อไป ซึ่งในบทนี้จะกล่าวถึงรายละเอียดเกี่ยวกับตัวแบบโพรบิตแบบ 2 ประเภท การประมาณค่าสัมประสิทธิ์การถดถอยในตัวแบบโพรบิตแบบ 2 ประเภท ด้วยวิธีภาวะความน่าจะเป็นสูงสุด (Maximum Likelihood Estimation) การหาจุดแบ่งโดยใช้ทฤษฎี Hadjicostas P. (2006) การหาอัตราความผิดพลาดในการจำแนกกลุ่มหรือสัดส่วนความถูกต้องในการจำแนกกลุ่ม

2.1 ตัวแบบโพรบิตแบบ 2 ประเภท (Binary Probit Model)

ตัวแบบที่ศึกษาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ เมื่อตัวแปรตามเป็นตัวแปรเชิงคุณภาพที่มี 2 ลักษณะ และตัวแปรอิสระเป็นตัวแปรเชิงปริมาณหรือตัวแปรหุ่น เมื่อได้รูปแบบความสัมพันธ์ระหว่างตัวแปรแล้ว จะนำไปใช้ในการประมาณค่าตัวแปรตาม และพยากรณ์โอกาสที่แต่ละหน่วยจะอยู่ในกลุ่มใดกลุ่มหนึ่งได้

จากตัวแปรตาม(Y_i) ซึ่งเป็นตัวแปรเชิงคุณภาพที่เป็นได้ 2 ค่า คือ 0 กับ 1 และ $X_{i1}, X_{i2}, ..., X_{ik}$ เป็นตัวแปรอิสระของค่าสังเกตที่ได้จากหน่วยตัวอย่างที่ i โดยที่ $i = 1, 2, ..., n$

โดยมี ตัวแปรแฝง คือ Y_i^* ซึ่งเป็นค่าที่วัดไม่ได้ จึงไม่ทราบค่าที่แท้จริง ทราบเพียงแต่ผลที่เกิดขึ้น โดย Y_i^* ของหน่วยตัวอย่างที่ i เป็นฟังก์ชันเชิงเส้นของตัวแปรอิสระ $X_{i1}, X_{i2}, ..., X_{ik}$ และนั่นคือ

$$Y_i^* = \beta X_i' + \varepsilon_i$$

เมื่อ ε_i คือค่าความคลาดเคลื่อนของหน่วยตัวอย่าง และ $\varepsilon_i \overset{iid}{\sim} N(0,1)$

จะได้

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \text{ or } \varepsilon_i < \beta X_i' \\ 0 & \text{if } Y_i^* \le 0 \text{ or } \varepsilon_i \ge \beta X_i' \end{cases}$$

จาก

$$\begin{aligned}\pi_i &= P(Y_i = 1)\\ &= P(Y_i^* > 0)\\ &= P(\varepsilon_i < \beta' X_i)\\ &= \Phi(\beta' X_i)\\ &= \int_{-\alpha}^{\beta X_i'} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-\varepsilon^2}{2}\right) d\varepsilon\end{aligned}$$

โดยที่ π_i คือ ค่าความน่าจะเป็นเมื่อเกิดเหตุการณ์ที่สนใจในหน่วยที่ i

Y_i คือ ตัวแปรตามเชิงคุณภาพที่มีค่าได้เพียง 2 ค่า คือ 0 และ 1

Y_i^* คือ ตัวแปรแฝง เป็นค่าที่วัดไม่ได้ ไม่ทราบค่าที่แท้จริง ทราบเพียงแต่ผล

$\Phi(.)$ คือ ฟังก์ชันการแจกแจงสะสมปกติมาตราฐาน

ε คือ ตัวแปรสุ่มปกติมาตราฐาน 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 $(k+1) \times 1$

ดังนั้น Probit Model สามารถเขียนแปลงให้อยู่ในรูป

$$\begin{aligned}\Phi^{-1}(\pi_i) &= \boldsymbol{\beta}' \mathbf{X}_i\\ &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} \quad \ldots\ldots\ldots\ldots\ldots (1)\end{aligned}$$

จากสมการที่(1) พบว่า การแปลงดังกล่าว จะทำให้เราได้ความสัมพันธ์เชิงเส้นตรงระหว่างตัวแปร

2.2 ฟังก์ชันความน่าจะเป็น (Likelihood function) ของข้อมูลการถดถอยโพรบิต

จากประชากรที่ทำการศึกษาเป็นตัวแปรตาม (Y_i) ที่มีเพียง 2 ค่า คือ 0 กับ 1 จึงใช้ฟังก์ชันการแจกแจงแบบเบอร์นูลลี่

$$P(Y_i = y_i) = p^{y_i}(1-p)^{1-y_i} \quad ; \; y_i = 0,1$$

สร้างฟังก์ชันของการแจกแจงความน่าจะเป็นร่วม (Joint Probability Density Function) ของหน่วยตัวอย่างอิสระ n ค่า โดยการคูณฟังก์ชันการแจกแจงความน่าจะเป็นของทุกหน่วยตัวอย่าง ($g(Y_i)$)

$$\begin{aligned}g(Y_1, Y_2, \ldots, Y_n) &= \prod_{i=1}^{n} g(Y_i)\\ &= \prod_{i=1}^{n} P_i^{y_i}(1-P_i)^{1-y_i}\\ &= \prod_{i=1}^{n} [\Phi(\beta X_i')]^{y_i} [1-\Phi(\beta X_i')]^{1-y_i}\end{aligned}$$

ดังนั้น $$L(\beta) = \prod_i^n \Phi(x_i'\beta)^{y_i} [1-\Phi(x_i'\beta)]^{1-y_i} \quad \ldots\ldots\ldots\ldots\ldots (2)$$

2.3 การประมาณค่าพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimation)

การหาค่าตัวประมาณค่าสัมประสิทธิ์การถดถอย ($\hat{\beta}$) ด้วยวิธีภาวะความน่าจะเป็นสูงสุด คือต้องทำให้ L มีค่ามากที่สุดโดยทำการหาอนุพันธ์เทียบกับ β ต่างๆ โดยเราสามารถเขียนให้อยู่ในรูปของลอการิทึม (Logarithm) หรือความควรจะเป็นลอการึทึม (Log-Likelihood) ได้ดังนี้

$$\ln L(\beta) = \sum_{i=1}^{n} \{y_i \cdot \ln[\Phi(x_i'\beta)] + (1 + y_i) \cdot \ln[1 - \Phi(x_i'\beta)]\} \quad \text{..... (3)}$$

$$= \sum_{y_i=0} \ln[1 - \Phi(x_i'\beta)] + \sum_{y_i=1} \ln \Phi(x_i'\beta) \quad \text{..... (4)}$$

และเงื่อนไขอันดับแรก(First Order) สำหรับการให้สมการที่ (4) มีค่าสูงสุด (Maximization) และสามารถแก้สมการหาค่า First Order Condition ของแบบจำลองโพรบิต ได้ดังนี้

$$\frac{\partial \ln L(\beta)}{\partial \beta} = \sum_{i=1}^{n} \left\{ \frac{y_i \phi(\cdot)}{\Phi(\cdot)} + (1 - y_i) \left[\frac{-\phi(\cdot)}{1 - \Phi(\cdot)} \right] \right\} x_i = 0$$

$$= \sum_{y_i=0} \left[\frac{-\phi(x_i'\beta)}{1 - \Phi(x_i'\beta)} \right] x_i + \sum_{y_i=1} \left[\frac{\phi(x_i'\beta)}{\Phi(x_i'\beta)} \right] x_i = 0 \quad \text{..... (5)}$$

เมื่อ ϕ_i คือ ฟังก์ชันความหนาแน่นของตัวแปรสุ่มที่มีการแจกแจงแบบปกติ

Φ_i คือ ฟังก์ชันการแจกแจงสะสมของตัวแปรสุ่มที่มีการแจกแจงแบบปกติ

เมื่อประมาณค่าพารามิเตอร์ด้วยวิธีความควรจะเป็นสูงสุดในตัวแบบโพรบิตแบบ 2 ประเภท แล้ว จะสามารถนำไปใช้ในการพยากรณ์การจำแนกกลุ่มของตัวแบบ ดังนี้

- หน่วยที่ i จะถูกจัดให้อยู่ในกลุ่มที่การเกิดที่สนใจ (Y=1) ถ้า

$$\hat{\pi}_i = \Phi(\beta' X_i) \quad > c \qquad ;0 \le c \le 1$$

- หน่วยที่ i จะถูกจัดให้อยู่ในกลุ่มที่ไม่เกิดเหตุการณ์ที่สนใจ (Y=0) ถ้า

$$\hat{\pi}_i = \Phi(\beta' X_i) \quad \le c \qquad ;0 \le c \le 1$$

เมื่อ c คือ จุดแบ่ง หรือระดับของความน่าจะเป็นที่ใช้ในการพิจารณาการจำแนกกลุ่มว่าแต่ละหน่วยจะอยู่ในกลุ่มใดระหว่างกลุ่มการเกิดเหตุการณ์ที่สนใจ และกลุ่มไม่เกิดเหตุการณ์ที่สนใจ

2.4 **สถิติ** Kaiser-Meyer-Olkin (KMO)

Kaiser (1970) ได้เสนอสถิติ KMO ในการศึกษานี้ สถิติ KMO เป็นสถิติที่ใช้วัดระดับความสัมพันธ์ระหว่างตัวแปรอิสระ (Multicollinearity)

$$0 \le KMO = \frac{\sum_{i=1}^{p}\sum_{i<j=1}^{p}\rho_{ij}^2}{\sum_{i=1}^{p}\sum_{i<j=1}^{p}\rho_{ij}^2 + \sum_{i=1}^{p}\sum_{i<j=1}^{p}\rho_{ij|1,2,\ldots,(i-1),(i+1),\ldots,(j-1),(j+1),\ldots,p}^2} \le 1$$

โดยที่ ρ_{ij} คือ สัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระตัวที่ i และตัวแปรอิสระตัวที่ j

$\rho_{ij|1,2,\ldots,(i-1),(i+1),\ldots,(j-1),(j+1),\ldots,p}$ คือ สัมประสิทธิ์สหสัมพันธ์บางส่วนระหว่างตัวแปรอิสระตัวที่ i และตัวแปรอิสระตัวที่ j สำหรับ $i < j = 1,2,\ldots,p$

เกณฑ์สำหรับการแบ่งระดับความสัมพันธ์ระหว่างตัวแปรอิสระ แสดงดังนี้

- ไม่มีระดับความสัมพันธ์ระหว่างตัวแปรอิสระ: $KMO = 0$
- ตัวแปรอิสระมีความสัมพันธ์กันในระดับอย่างต่ำ: $0 < KMO \le 0.50$
- ตัวแปรอิสระมีความสัมพันธ์กันในระดับปานกลาง: $0.50 < KMO \le 0.75$
- ตัวแปรอิสระมีความสัมพันธ์กันในระดับสูง: $0.75 < KMO < 1.00$
- ตัวแปรอิสระมีความสัมพันธ์กันอย่างสมบูรณ์: $KMO = 1$

2.5 **การตรวจสอบความเหมาะสมของตัวแบบ**

ตัวแบบที่เป็นไปได้มีหลายตัวแบบ ซึ่งมีวิธีการวัดทางสถิติที่หลากหลาย สำหรับใช้ในการวัดว่าตัวแบบโพรบิตแบบ 2 ประเภทมีความสามารถในการจำแนกกลุ่มให้แต่ละหน่วยอยู่ในกลุ่มใดกลุ่มหนึ่งระหว่างกลุ่มที่เกิดเหตุการณ์การที่สนใจและกลุ่มที่ไม่เกิดเหตุการณ์ที่สนใจว่าเหมาะสมมากน้อยเพียงใด ซึ่งมีวิธีต่างๆ ดังนี้

- Chi-square Goodness of Fit Test และสถิติ Deviance
- HosMer-LoMeshow Test
- Classification Table
- Receiver Operating Characteristic Curve (ROC)
- สัมประสิทธิ์การตัดสินใจสำหรับการถดถอยโพรบิต $\left(R^2\right)$

- ตรวจสอบความถูกต้องของตัวแบบ (Model validation) ด้วยวิธีการใช้ชุดข้อมูลภายนอกหรือโดยแบ่งชุดข้อมูลออกเป็น 2 ส่วน

Classification table จะสอดคล้องกับแต่ละจุดแบ่ง ที่ถูกคัดเลือกซึ่งจะถูกใช้เป็นเกณฑ์สำหรับการคัดเลือกจุดที่ทำให้อัตราความผิดพลาดในการจำแนกกลุ่มมีค่าต่ำมีค่าต่ำมีค่าต่ำที่สุด โดย Classification table สำหรับแต่ละจุดแบ่ง แสดงดังนี้

		ค่าสังเกต		
		$y_i = 1$	$y_i = 0$	
ค่าพยากรณ์	$\hat{y}_i = 1$	A	C	A+C
(เกณฑ์ที่ใช้ในการตัดสินใจ คือ $\hat{\pi}_i \geq c$)	$\hat{y}_i = 0$	B	D	B+D
		A+B	C+D	A+B+C+D

โดย

"A" คือ จำนวนของค่าสังเกตที่ถูกจัดให้อยู่ในกลุ่มของการเกิดเหตุการณ์ที่สนใจและค่าสังเกตที่แท้จริงอยู่ในกลุ่มการเกิดเหตุการณ์ที่สนใจด้วย

"B" คือ จำนวนของค่าสังเกตที่ถูกจัดให้อยู่ในกลุ่มของการไม่เกิดเหตุการณ์ที่สนใจในขณะที่ค่าสังเกตที่แท้จริงอยู่ในกลุ่มการเกิดเหตุการณ์ที่สนใจด้วย

"C" คือ จำนวนของค่าสังเกตที่ถูกจัดให้อยู่ในกลุ่มของการเกิดเหตุการณ์ที่สนใจในขณะที่ค่าสังเกตที่แท้จริงอยู่ในกลุ่มของการไม่เกิดเหตุการณ์ที่สนใจด้วย

"D" คือ จำนวนของค่าสังเกตที่ถูกจัดให้อยู่ในกลุ่มของการเกิดเหตุการณ์ที่สนใจและค่าสังเกตที่แท้จริงอยู่ในกลุ่มของการไม่เกิดเหตุการณ์ที่สนใจด้วย

การคัดเลือกจุดแบ่ง ต้องทำให้อัตราความผิดพลาดในการจำแนกกลุ่มมีค่าต่ำดหรือสัดส่วนความถูกต้องในการจำแนกกลุ่มมีค่าสูงที่สุด โดยแสดงดังนี้

อัตราความผิดพลาดในการจำแนกกลุ่มมีค่าต่ำมีค่าต่ำ (Classification Error Rate: CER)	สัดส่วนความถูกต้องในการจำแนกกลุ่ม
$CER = \frac{B+C}{A+B+C+D}$	1 – CER
CER มีค่าต่ำกว่าแสดงว่ามีความเหมาะสมมากกว่า	อัตราความถูกต้องในการจำแนกกลุ่มมีค่ามากแสดงว่ามีความเหมาะสมมากกว่า

2.6 **ช่วงความเชื่อมั่น** (Confidence Interval)

การประมาณค่าแบบช่วงหรือช่วงความเชื่อมั่นเป็นการประมาณค่าพารามิเตอร์ของประชากรในรูปแบบช่วงโดยใช้ข้อมูลตัวอย่าง การประมาณแบบช่วงนั้นจะบอกถึงค่าต่ำสุดและค่าสูงสุดของพารามิเตอร์ที่เป็นไปได้

ระดับของค่าความเชื่อมั่นที่ใช้ในการสร้างช่วงความเชื่อมั่นนั้นจะกำหนดเป็นค่าควบคู่กับระดับนัยสำคัญ นั่นคือ $1-\alpha$ ที่เรียกว่าช่วงความเชื่อมั่น $(1-\alpha)100\%$ จะได้ว่า

$$P(L < \theta < U) = 1-\alpha$$

เรียก L ว่าขีดจำกัดความเชื่อมั่นล่าง (Lower Confidence Limit)

U ว่าขีดจำกัดความเชื่อมั่นบน (Upper Confidence Limit)

2.7 **เปอร์เซ็นต์ไทล์** (Percentile)

เป็นค่าที่แบ่งข้อมูลออกเป็น 100 ส่วนเท่าๆ กัน เมื่อข้อมูลถูกเรียงจากน้อยไปหามาก เนื่องจากค่าที่แบ่งจำนวนข้อมูลออกเป็น 100 ส่วนเท่าๆ กัน มีอยู่ 99 ค่า ดังนั้นเราจึงตั้งชื่อแต่ละค่าว่า เปอร์เซ็นต์ไทล์ที่หนึ่ง (P_1) เปอร์เซ็นต์ไทล์ที่สอง (P_2).... และเปอร์เซ็นต์ไทล์ที่เก้าสิบเก้า (P_{99}) ตามลำดับ

การหาเปอร์เซ็นต์ไทล์ คือต้องหาตำแหน่งของเปอร์เซ็นต์ไทล์ก่อน ให้ N เป็นจำนวนข้อมูลทั้งหมด ตำแหน่งต่างๆ ของเปอร์เซ็นต์ไทล์หาได้ดังนี้

P_1 อยู่ในตำแหน่งที่ คือ $\frac{1(N+1)}{100}$

P_2 อยู่ในตำแหน่งที่ คือ $\frac{2(N+1)}{100}$

$\vdots$

P_{99} อยู่ในตำแหน่งที่ คือ $\frac{99(N+1)}{100}$

โดยทั่วไป ตำแหน่งของเปอร์เซ็นต์ไทล์ที่ r คือ

P_r อยู่ในตำแหน่งที่ คือ $\frac{r(N+1)}{100}$

2.8 **การหาจุดแบ่ง โดยทฤษฎี**Hadjicostas P. (2006)

ทฤษฎี Hadjicostas P. (2006) เป็นวิธีการนี้ใช้ผลลัพธ์ทางคณิตศาสตร์ที่ง่ายแต่มีความถูกต้องแม่นยำในการหาจุดแบ่ง ที่ทำให้อัตราความผิดพลาดในการจำแนกกลุ่มมีค่าต่ำสุดหรือสัดส่วนความถูกต้องในการจำแนกกลุ่มมีค่าสูงสุดอีกด้วย โดยมีวิธีดังนี้

<u>ทฤษฎีบท 1</u> ให้ $a_i = \sum_{k=1}^{M(i)} (-1)^{y_k}$ สำหรับ $i = 0,1,2,...,n$ ให้ I_0 เป็นเซตของ j ทั้งหมด $j \in \{0,1,2,...,n\}$ ที่ซึ่ง $a_j = \max_{0 \le i \le n} a_i$ และให้ C_0 เป็นเซตของ c_0 ทั้งหมด $c_0 \in [0,1]$ ที่ซึ่ง $p(c_0) = \max_{c \in [0,1]} p(c)$ แล้ว $C_0 = \bigcup_{i \in I_0} A_i$

<u>บทตั้ง 1</u> $N(c) = \sum_{j=1}^{M(i)} (1 - y_j) + \sum_{j=M(i)+1}^{n} y_j$ สำหรับ $i \in \{0,1,2,...,n\}$ ใดๆ และ $c \in A_i$

จากทฤษฎีและบทตั้งดังกล่าว จะได้ว่า

1. เรียงอันดับค่า $\hat{\pi}_i$ จากน้อยไปหามาก $\hat{\pi}_1 < \hat{\pi}_2 < ... < \hat{\pi}_n$
2. สำหรับแต่ละ $i \in \{1,2,...,n\}$ ใดๆ ให้ M(i) เป็น $\max j \in \{1,2,...,n\}$ ที่ซึ่ง $\hat{\pi}_i = \hat{\pi}_j$ โดย $M(0) = 0 \; ; i \le M(i) \le n$
3. หาค่า a_i โดย $a_i = \sum_{k=1}^{M(i)} (-1)^{y_k}$ สำหรับ $i = 0,1,2,...,n$ แบ่งเป็น 2 กรณี คือ

   <u>กรณีที่ 1</u> $a_{i+1} = a_i + \sum_{k=M(i)+1}^{M(i+1)} (-1)^{y_k}$ ถ้า $M(i) < i+1$

   <u>กรณีที่ 2</u> $a_{i+1} = a_i$ ถ้า $i+1 \le M(i)$
4. หา I_0 ซึ่งเป็นเซตของ j ทั้งหมด $j \in \{0,1,2,...,n\}$ ที่ซึ่ง $a_j = \max_{0 \le i \le n} a_i$
5. หา C_0 เป็นเซตของ c_0 ทั้งหมด $c_0 \in [0,1]$ ที่ซึ่ง $p(c_0) = \max_{c \in [0,1]} p(c)$ แล้ว $C_0 = \bigcup_{i \in I_0} A_i$ โดย $i \in \{0,1,2,...,n\}$ จะได้ว่า

   $A_i = [0, \hat{\pi}_1)$ ถ้า $i = 0$

   $A_i = [\hat{\pi}_i, \hat{\pi}_{i+1})$ ถ้า $\hat{\pi}_i < \hat{\pi}_{i+1}$ และ $1 \le i < n$

   $A_i = \{\hat{\pi}_i\} = \{\hat{\pi}_{M(i)}\}$ ถ้า $\hat{\pi}_i = \hat{\pi}_{i+1}$ และ $1 \le i < n$

   $A_i = [\hat{\pi}_n, 1]$ ถ้า $i = n$

   ข้อสังเกต $\bigcup_{i=0}^{n} A_i = [0,1]$
6. เลือกค่าของจุดแบ่ง c โดย $c \in C_0$ และ $c \in [0,1]$ ซึ่งเป็นค่า c ที่ทำให้สัดส่วนของความถูกต้องในการจำแนกกลุ่มที่จุด c มีค่ามากที่สุด

$$p(c) = \frac{N(c)}{n}$$

โดย $p(c)$ คือ สัดส่วนของความถูกต้องในการจำแนกกลุ่มที่จุด c

$N(c)$ คือ จำนวนของความถูกต้องในการจำแนกกลุ่มที่จุด c