บทที่ 2

แนวคิดและทฤษฎี

งานวิจัยนี้ต้องการศึกษาการจำแนกกลุ่มหรือการพยากรณ์กลุ่มแบบอันดับระหว่างวิธีวิเคราะห์จำแนกประเภทและวิธีวิเคราะห์ความถดถอยโลจิสติกอันดับซึ่งมีทฤษฎีที่เกี่ยวข้องดังนี้

2.1 สเกลอันดับ (Ordinal Scale)

เป็นการแบ่งกลุ่มเช่นเดียวกับสเกลแบ่งกลุ่มหรือนามกำหนด (Nominal Scale) แต่ให้รายละเอียดที่มากกว่านั่นคือ สามารถแสดงความแตกต่างระหว่างกลุ่ม หรือสามารถเปรียบเทียบกลุ่มหรือลำดับกลุ่มได้ว่ากลุ่มใดดีกว่า มากกว่า น้อยกว่า พอใจมากกว่า เห็นด้วยน้อยกว่ากลุ่มอื่นๆหรือไม่ เช่น ระดับการศึกษา รายได้ที่แบ่งเป็นช่วงๆ อายุที่แบ่งเป็นช่วงๆ เป็นต้น

2.2 กลุ่มอันดับ (Ordered Group)

กลุ่มอันดับคือ กลุ่มที่ถูกแบ่งด้วยสเกลอันดับ กล่าวคือ สามารถเปรียบเทียบความแตกต่างหรือความสำคัญระหว่างกลุ่มได้ เช่น ระดับการศึกษา อาจแบ่งออกเป็น 5 กลุ่ม ดังนี้

1. ประถมศึกษา
2. มัธยมศึกษาตอนต้น
3. มัธยมศึกษาตอนปลาย / ปวช. / ปวส.
4. ปริญญาตรี
5. สูงกว่าปริญญาตรี

ในกรณีนี้เราสามารถบอกได้ว่าคนที่อยู่ในกลุ่มที่ 1 คือ จบประถมศึกษาจะมีการศึกษาน้อยกว่าคนที่อยู่ในกลุ่มที่ 2 ถึง กลุ่มที่ 5

ตัวอย่างกรณีที่กลุ่มไม่ใช่กลุ่มแบบอันดับ เช่น การแบ่งรสชาติของนมกล่องที่ชื่นชอบ อาจแบ่งออกเป็น 5 กลุ่มดังนี้

1. ชอบรสหวาน
2. ชอบรสจืด
3. ชอบรสสตรอเบอรี่
4. ชอบรสกาแฟ
5. ชอบรสช็อคโกแลต

จะสังเกตว่าเราไม่สามารถบอกได้ว่าคนที่อยู่ในกลุ่มที่ 1 ซึ่งชอบรสหวาน มีความสำคัญแตกต่างหรือความสำคัญจากคนที่อยู่ในกลุ่มที่ 2 ถึงกลุ่มที่ 5 อย่างไร

2.3 การวิเคราะห์หลายตัวแปร (Multivariate Analysis)

ในการวิเคราะห์หลายตัวแปร เป็นการเก็บรวบรวมข้อมูลหลายๆ ลักษณะหรือหลายๆตัวแปรจากแต่ละหน่วย ในที่นี้กำหนดให้มีตัวแปร p ตัว คือ $X_1, X_2, ..., X_p$ โดยที่ $p > 1$ และกำหนดให้ X_{ij} = ตัวแปรที่ j ของหน่วยที่ i ; $i = 1,2,...,n$; $j = 1,2,...,p$

จากตัวอย่าง n หน่วย จะได้ข้อมูลของ n หน่วย จากตัวแปร p ตัว เขียนในรูปเมทริกซ์ X ที่มี n แถวนอน และ p แถวตั้ง

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & & & \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

2.4 การแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution)

เทคนิคของการวิเคราะห์สถิติแบบใช้พารามิเตอร์เกือบทุกชนิด มีเงื่อนไขหรือข้อสมมติที่ว่าข้อมูลมาจากประชากรที่มีการแจกแจงแบบปกติ ไม่ว่าจะเป็นการทดสอบสมมติฐานเกี่ยวกับตัวแปรหนึ่งตัวที่ใช้สถิติทดสอบ Z, t หรือเทคนิคของตัวแปรหลายตัว เช่น การวิเคราะห์ความแปรปรวนหลายตัวแปร (MANOVA) หรือการวิเคราะห์จำแนกประเภท (Discriminant Analysis)

ถ้าเวกเตอร์ของตัวแปร X หรือ $\mathbf{X}' = (X_1, X_2, ..., X_p)$ มีการแจกแจงแบบปกติที่มีเวกเตอร์ของค่าเฉลี่ยเป็น $\boldsymbol{\mu}$ และเมทริกซ์ความแปรปรวนร่วม (covariance matrix) คือ $\boldsymbol{\Sigma}$ โดยมีฟังก์ชันการแจกแจงเป็น

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right]$$

โดยที่ p เป็นจำนวนตัวแปรอิสระ หรือเขียนย่อๆว่า $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

2.5 การวิเคราะห์จำแนกประเภท (Discriminant Analysis)

การวิเคราะห์จำแนกประเภทเป็นเทคนิคที่ใช้ในการแบ่งหน่วยตัวอย่างออกเป็นกลุ่มย่อยๆ ตั้งแต่ 2 กลุ่มขึ้นไปโดยหน่วยตัวอย่างที่อยู่ในกลุ่มเดียวกันจะคล้ายคลึงกัน หน่วยตัวอย่างที่อยู่ต่างกลุ่มกันจะแตกต่างกันและหน่วยตัวอย่างใดหน่วยตัวอย่างหนึ่งจะต้องอยู่เพียงกลุ่มเดียว การวิเคราะห์จำแนกประเภทต้องทราบจำนวนกลุ่มมาก่อนว่ามีกี่กลุ่ม และต้องเป็นกลุ่มที่มีจริงอยู่แล้ว สำหรับสมการถดถอยที่สร้างขึ้นในการวิเคราะห์จำแนกประเภทนั้น ตัวแปรตามจะต้องเป็นตัวแปรเชิงกลุ่ม เช่น ถ้าแบ่งลูกหนี้เป็น 3 กลุ่ม ตัวแปรตาม Y จะมี 3 ค่า คือ

$$Y = \begin{cases} 1 & \text{ถ้าเป็นลูกหนี้ชั้นดี} \\ 2 & \text{ถ้าเป็นลูกหนี้ปกติ} \\ 3 & \text{ถ้าเป็นลูกหนี้ที่มีปัญหา} \end{cases}$$

ส่วนตัวแปรอิสระหรือตัวแปรที่ทำให้กลุ่มต่างกันควรเป็นตัวแปรเชิงปริมาณ กรณีที่ตัวแปรอิสระเป็นตัวแปรเชิงกลุ่ม หรือตัวแปรเชิงคุณภาพจะต้องปรับให้อยู่ในรูปตัวแปรหุ่น (Dummy Variable) ตัวแปรอิสระนี้อาจจะมีเพียงหนึ่งตัว หรือตั้งแต่ 2 ตัวขึ้นไป ความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระจะอยู่ในรูปเชิงเส้น ดังนี้

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + e \qquad \text{\_\_\_\_\_(2.1)}$$

โดยที่ Y เป็นตัวแปรตามและเป็นข้อมูลเชิงกลุ่ม

$X_1, X_2, ..., X_p$ เป็นตัวแปรอิสระ ; $p \geq 1$

e เป็นค่าคลาดเคลื่อน

ในการวิเคราะห์จำแนกประเภทจะเรียกสมการที่ (2.1) ว่า ฟังก์ชันจำแนกกลุ่ม (Discriminant Function) หรือ เรียกสมการนี้ว่า Fisher Discriminant Function ซึ่ง R.A.Fisher เป็นผู้คิดค้นขึ้น

เมื่อใช้ข้อมูลตัวอย่างมาประมาณสมการที่ (2.1) จะได้เป็น

$$\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_p x_p$$

โดยที่ $\hat{Y}$ = Discriminant Score

b_i = สัมประสิทธิ์ของสมการจำแนกกลุ่ม

x_i = ตัวแปรอิสระหรือเรียกว่า ตัวแปรจำแนกกลุ่มตัวที่ i ; $i = 1,2,..., p$

p = จำนวนตัวแปรจำแนกกลุ่ม

k = จำนวนกลุ่ม

จำนวนสมการจำแนกกลุ่ม = $\min(p, k-1)$

การประมาณค่าสัมประสิทธิ์ $\beta_0, \beta_1, ...\beta_p$ ด้วย $b_0, b_1, ...b_p$ ตามลำดับ ใช้วิธีการของ Fisher โดยมีเป้าหมายเพื่อทำให้ความแตกต่างระหว่างกลุ่มมีค่ามากที่สุด นั่นคือ ทำให้ค่า $\frac{SSB}{SSW}$ มีค่ามากที่สุด

โดยที่ SSB (Sum Square Between group) หมายถึงผลบวกกำลังสองของความแตกต่างระหว่างกลุ่ม

SSW (Sum Square Within group) หมายถึงผลบวกกำลังสองของความแตกต่างภายในกลุ่ม

<u>วัตถุประสงค์ของการวิเคราะห์จำแนกประเภท</u> คือ

1. เพื่อศึกษาสาเหตุที่ทำให้กลุ่มมีความแตกต่าง โดยศึกษาว่าตัวแปรอิสระใดบ้างเป็นตัวแปรที่ทำให้กลุ่มต่างกัน โดยการสร้างฟังก์ชันการจำแนกกลุ่ม
2. เพื่อพยากรณ์หน่วยตัวอย่างใหม่ที่ยังไม่ทราบกลุ่มว่าควรจะอยู่กลุ่มใดในอนาคต

<u>เงื่อนไขของการวิเคราห์จำแนกประเภทเมื่อจำนวนกลุ่มมี k กลุ่ม</u> มีดังนี้

1. ตัวแปรอิสระ p ตัว $(X_1, X_2, ..., X_p)$ มีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution)
2. เมทริกซ์ค่าความแปรปรวนร่วม (Variance-Covariance matrix) ของแต่ละกลุ่มต้องเท่ากันนั่นคือ $\Sigma_1 = \Sigma_2 = ... = \Sigma_k$

<u>การตรวจสอบเงื่อนไขของการวิเคราะห์จำแนกประเภท</u>

1. การตรวจสอบเงื่อนไขเกี่ยวกับการแจกแจงของตัวแปรอิสระ p ตัวที่ต้องมีการแจกแจงแบบปกติหลายตัวแปร มีได้หลายวิธีเช่น Chi-Square plot และ วิธีของ Mardina ที่ใช้ตรวจสอบด้วยค่าความเบ้ และความโด่ง
2. การตรวจสอบเมทริกซ์ค่าความแปรปรวนร่วมของตัวแปรอิสระของแต่ละกลุ่มว่าเท่ากันหรือไม่ จะใช้สถิติทดสอบบ๊อกซ์ (Box's Test)

<u>วิธีการจำแนกกลุ่มโดยการวิเคราะห์จำแนกประเภท</u>

การจำแนกกลุ่มโดยวิธีวิเคราะห์จำแนกประเภทนั้นมีด้วยกันหลายวิธีแต่วิธีที่เลือกใช้ในการวิจัยครั้งนี้คือ วิธีความน่าจะเป็นก่อนและความน่าจะเป็นหลัง (Prior probability and Posterior probability) โดยวิธีนี้ผู้วิจัยจะเป็นผู้กำหนดความน่าจะเป็นหรือโอกาสที่หน่วยตัวอย่างจะอยู่ในกลุ่มใดกลุ่มหนึ่งขึ้นก่อนหลังจากนั้นจะทำการคำนวณหาค่าความน่าจะเป็นหลังของทุกกลุ่มแล้วเปรียบเทียบว่าความน่าจะเป็นหลังของกลุ่มใดมากสุดจะพยากรณ์ให้หน่วยตัวอย่างอยู่ในกลุ่มนั้น

2.6 การวิเคราะห์ความถดถอยโลจิสติกอันดับ (Ordinal Logistic Regression)

การวิเคราะห์ความถดถอยโลจิสติกอันดับ (Ordinal Logistic Regression หรือเรียกว่า Proportional odds model หรือ cumulative logit model) เป็นการวิเคราะห์หลายตัวแปรชนิดหนึ่ง ใช้สำหรับวิเคราะห์ข้อมูลที่มีตัวแปรตาม 1 ตัวซึ่งเป็นตัวแปรที่ถูกวัดด้วย ordinal scale และมีตัวแปรอิสระอย่างน้อย 1 ตัวซึ่งจะเป็นตัวแปรแบบเชิงปริมาณหรือตัวแปรเชิงคุณภาพก็ได้

ในอดีตการวิเคราะห์ทางสถิติเมื่อตัวแปรตามเป็นแบบอันดับมักจะวิเคราะห์โดยยุบตัวแปรนั้นให้เป็นแบบนามกำหนด(Nominal) แล้ววิเคราะห์ด้วย Binary Logistic Regression ซึ่งทำให้รายละเอียดของข้อมูลขาดหายไปเนื่องจากตัวแปรตามถูกบังคับให้มีเพียงสองระดับเท่านั้นและยังทำให้ค่าสถิติเกิดความคลาดเคลื่อนอีกด้วย จนในที่สุดได้มีการพัฒนาตัวแบบจนมาเป็น Ordinal Logistic Regression หรือ Proportional odds model ในที่สุด

การวิเคราะห์ด้วยวิธีความถดถอยโลจิสติกอันดับมีวัตถุประสงค์ 2 ประการ คือ

1. เพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตาม หรือศึกษาว่าตัวแปรใดบ้างที่มีอิทธิพลหรือผลกระทบต่อตัวแปรตาม
2. เพื่อพยากรณ์โอกาสที่จะเกิดเหตุการณ์หนึ่งๆที่สนใจ โดยใช้สมการที่สร้างขึ้น เมื่อทราบค่าตัวแปรอิสระ

ตัวแบบของ Proportional odds model นั้นจะเป็นการเปรียบเทียบความน่าจะเป็นที่ตัวแปรอิสระ Y จะมีค่าน้อยกว่าหรือเท่ากับ $k\ (Y \le j)$ กับ ความน่าจะเป็นที่ตัวแปรอิสระ Y จะมีค่ามากกว่าหรือเท่ากับ $k\ (Y \ge j)$

$$\text{logit}(\theta_j) = \ln\left[\frac{P(Y \le j|X)}{P(Y > j|X)}\right] \qquad \text{(2.2)}$$

$$= \ln\left[\frac{\pi_1(X)+\pi_2(X)+\ldots+\pi_j(X)}{\pi_{j+1}(X)+\pi_{j+2}(X)+\ldots+\pi_K(X)}\right]$$

$$= \alpha_k + \sum_{i=1}^{p}\beta_i X_i$$

เมื่อ j = จุดตัดที่ j ของตัวแปรตาม

α_j = ค่าคงที่ของสมการเมื่อตัวแปรตามมีจุดตัดที่ k

β_i = สัมประสิทธิ์ของตัวแปรอิสระตัวที่ i ; $i = 1,2,\ldots,p$

$\pi_j(X)$ = ความน่าจะเป็นที่ Y จะเท่ากับ j หรืออยู่กลุ่มที่ j

p = จำนวนตัวแปรอิสระทั้งหมด

K = จำนวนกลุ่ม

การวิเคราะห์ด้วยการวิเคราะห์ความถดถอยโลจิสติกอันดับนั้นเริ่มต้นด้วยแบ่งตัวแปรออกเป็นกลุ่มเพื่อเปรียบเทียบดังสมการที่ (2.2) ซึ่งจะแบ่งได้ $K-1$ กลุ่ม เมื่อ K คือจำนวนกลุ่มของ Y เช่น กรณีตัวแปรตาม Y มี 3 ค่า ($Y=1,2,3$) จะแบ่งตัวแปรตามออกได้เป็น 2 แบบเพื่อวิคราะห์ ดังนี้

<u>ตัวแบบที่ 1</u> เปรียบเทียบ $Y=1$ กับ $[(Y=2)+(Y=3)]$

$$\pi_1 = \frac{e^{\alpha_1+\sum_{i=1}^{p}\beta_i X_i}}{1+e^{\alpha_1+\sum_{i=1}^{p}\beta_i X_i}}$$

$$\pi_2+\pi_3 = \frac{1}{1+e^{\alpha_1+\sum_{i=1}^{p}\beta_i X_i}}$$

$$\text{Odds} = \frac{\pi_1(X)}{\pi_2(X)+\pi_3(X)} = e^{\alpha_1+\sum_{i=1}^{p}\beta_i X_i}$$

$$\begin{aligned}\text{logit}(\theta_1) &= \ln\left[\frac{P(Y\le 1|X)}{P(Y>1|X)}\right]\\ &= \ln\left[\frac{\pi_1(X)}{\pi_2(X)+\pi_3(X)}\right]\\ &= \alpha_1+\sum_{i=1}^{p}\beta_i X_i\end{aligned}$$

<u>ตัวแบบที่ 2</u> เปรียบเทียบ $[(Y=1)+(Y=2)]$ กับ $Y=3$

$$\pi_1+\pi_2 = \frac{e^{\alpha_2+\sum_{i=1}^{p}\beta_i X_i}}{1+e^{\alpha_2+\sum_{i=1}^{p}\beta_i X_i}}$$

$$\pi_3 = \frac{1}{1+e^{\alpha_2+\sum_{i=1}^{p}\beta_i X_i}}$$

$$\text{Odds} = \frac{\pi_1(X)+\pi_2(X)}{\pi_3(X)} = e^{\alpha_2+\sum_{i=1}^{p}\beta_i X_i}$$

$$\begin{aligned}\text{logit}(\theta_2) &= \ln\left[\frac{P(Y\le 2|X)}{P(Y>2|X)}\right]\\ &= \ln\left[\frac{\pi_1(X)+\pi_2(X)}{\pi_3(X)}\right]\\ &= \alpha_2+\sum_{i=1}^{p}\beta_i X_i\end{aligned}$$

หลังจากจัดให้ตัวแปร Y เป็น 2 กลุ่มแล้วจะสามารถวิเคราะห์ด้วย Binary Logistic Regression ได้

จะสังเกตว่าตัวแบบของ Proportional odds model นั้น ค่าสัมประสิทธิ์ β_i จะเท่ากันในทุกตัวแบบไม่ว่าจะแบ่งค่า Y ที่จุดใด จะแตกต่างกันเฉพาะจุดตัด α_k จึงเป็นที่มาของสมมติฐานสำหรับการวิเคราะห์ด้วยวิธีความถดถอยโลจิสติกอันดับคือ odds ratio ของตัวแปรอิสระแต่ละตัวต้องมีค่าคงที่ ไม่ว่าจะแบ่งตัวแปรตาม Y อย่างไร ข้อตกลงนี้เป็นข้อตกลงสำคัญที่ต้องตรวจสอบทุกครั้ง ถ้าข้อตกลงนี้ไม่เป็นจริงจะต้องเลือกใช้ตัวแบบอื่นในการวิเคราะห์

การทดสอบข้อตกลงเบื้องต้นการเท่ากันของ Odds ratio นั้นจะใช้วิธีการทดสอบ Wald Test ซึ่งเป็นการเปรียบเทียบค่าสัมประสิทธิ์ของตัวแปรอิสระ X แต่ละตัวในทุกตัวแบบว่าแตกต่างกันหรือไม่

การจำแนกกลุ่มโดยวิธีวิเคราะห์ความถดถอยโลจิสติกอันดับนั้นจะทำได้โดยคำนวณหาค่าความน่าจะเป็นที่ตัวแปร Y จะมีค่าเท่ากับกลุ่มต่างๆ ถ้าค่าความน่าจะเป็นของ Y ในกลุ่มไหนมีค่ามากที่สุดจะพยากรณ์ให้หน่วยตัวอย่างอยู่ในกลุ่มนั้น