# บทที่ 2

## ระเบียบวิธีที่ใช้ในการวิจัย

สิ่งที่สนใจศึกษาในการวิจัยครั้งนี้ คือ การเปรียบเทียบการจำแนกกลุ่มโดยวิธีการวิเคราะห์การถดถอยทวิและการวิเคราะห์จำแนกประเภท เมื่อตัวแปรอิสระมีการแจกแจงแบบเบ้ ในการศึกษาเปรียบเทียบจะใช้ค่าเฉลี่ยของร้อยละที่พยากรณ์ถูกต้อง เป็นเกณฑ์เปรียบเทียบ โดยจะกล่าวถึงรายละเอียดสำหรับแต่ละวิธีดังต่อไปนี้

2.1 ข้อสมมติเบื้องต้นของตัวแบบการวิเคราะห์การถดถอยเชิงเส้น

2.2 วิธีกำลังสองน้อยที่สุดแบบธรรมดา(Ordinary Least Square method)

2.3 การวิเคราะห์การถดถอยแบบทวิ (Binary Regression)

2.4 การวิเคราะห์จำแนกประเภท (Discriminant Analysis)

พิจารณาจาก 2 กรณี คือ

2.4.1 ประชากรทั้ง 2 กลุ่ม มีปัจจัยต่อไปนี้เท่ากัน

- Cost of misclassification
- Prior Probability

2.4.2 ประชากรทั้ง 2 กลุ่ม มีปัจจัยต่อไปนี้แตกต่างกัน

- Cost of misclassification ทั้ง 2 กลุ่มเท่ากัน
- Prior Probability ในแต่ละกลุ่มแตกต่างกัน

2.1 <u>ข้อสมมติทั่วไปของตัวแบบการถดถอยเชิงเส้น</u>

ในการสมมติความสัมพันธ์เชิงเส้นระหว่างตัวแปร y และ ตัวแปรอิสระ m ตัว ของ $x_1, x_2, \ldots, x_m$ และค่าผิดพลาด $\varepsilon$ นั่นคือ ถ้ามีตัวอย่าง n ค่าสังเกตของ y และค่า x เราสามารถเขียนสมการได้เป็น

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_m x_{im} + \varepsilon_i \; ; \; i = 1, \ldots, n \qquad (2.1)$$

ค่าสัมประสิทธิ์ $\beta$ และค่าพารามิเตอร์ของการแจกแจงของ $\varepsilon$ ไม่ทราบค่า ปัญหาคือเราต้องการประมาณ $\beta$ ที่ไม่ทราบเหล่านี้เมื่อมี n สมการของ (2.1) ซึ่งเราสามารถเขียนในรูปของเมตริกซ์ดังนี้

$$\underline{y} = x\underline{\beta} + \underline{\varepsilon} \qquad (2.2)$$

เมื่อ

$$\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix}, \quad x = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}, \quad \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_m \end{bmatrix}, \quad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}$$

โดยมีข้อสมมติคือ

1. $E(\underline{\varepsilon}) = \underline{0}$
2. $E(\underline{\varepsilon}'\underline{\varepsilon}) = \sigma^2 I_n$
3. X เป็นชุดของค่าตัวเลขคงที่
4. และมีอันดับ $m<n$

จากข้อสมมติที่ 1 $E(\underline{\varepsilon}) = \underline{0}$ นั่นคือ ตัวแปร $\varepsilon_i$ จะมีค่าคาดหวังเป็นศูนย์สำหรับทุกค่าของ $i, i = 1,\ldots,n$

จากข้อสมมติที่ 2 จะได้ว่า

$$E(\underline{\varepsilon}'\underline{\varepsilon}) = \begin{bmatrix} E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2) & \cdots & E(\varepsilon_1\varepsilon_n) \\ E(\varepsilon_2\varepsilon_1) & E(\varepsilon_2^2) & \cdots & E(\varepsilon_2\varepsilon_n) \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ E(\varepsilon_n\varepsilon_1) & E(\varepsilon_n\varepsilon_2) & \cdots & E(\varepsilon_n^2) \end{bmatrix} \quad n \times n$$

$$E(\underline{\varepsilon}'\underline{\varepsilon}) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & & \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}_{n \times n}$$

$$E(\underline{\varepsilon}'\underline{\varepsilon}) = \sigma^2 I_n$$

ตามข้อสมมติที่ 3 เมตริกซ์ X ต้องเป็นตัวเลขคงที่ หมายถึง การสุ่มตัวอย่างที่ซ้ำกันจะเกิดค่าหลายค่าในเวกเตอร์ $\underline{y}$ ซึ่งแปรผันในเวกเตอร์ $\underline{\varepsilon}$ และจำนวนค่าสังเกตของ X จะต้องมีจำนวนเกินจำนวนพารามิเตอร์ที่จะประมาณ ตามข้อสมมติที่ 4 ทั้งนี้ จะต้องไม่มีความสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระด้วย เราเรียกความสัมพันธ์เชิงเส้นระหว่างตัวแปรอิสระว่า multicollinearity

## 2.2 วิธีกำลังสองน้อยที่สุดแบบธรรมดา (Ordinary least square method)

วิธีการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นวิธีนี้ มีหลักในการประมาณค่าสัมประสิทธิ์คือ ทำให้ผลบวกกำลังสองของค่าผิดพลาดมีค่าน้อยที่สุด ซึ่งแสดงรายละเอียดดังนี้

นิยาม จากสมการ $\underline{y} = x\underline{\beta} + \underline{\varepsilon}$ เมื่อ $\underline{\varepsilon} \sim N(\underline{0}, \sigma^2 I_n)$ ตัวประมาณกำลังสองน้อยที่สุดของ $\underline{\beta}$ คือ $\hat{\underline{\beta}}$ ที่ทำให้ผลบวกกำลังสองของค่าผิดพลาด (Sum square errors) หรือ SSE มีค่าน้อยที่สุด

จากนิยาม จะทำการหาตัวประมาณค่ากำลังสองน้อยที่สุดได้ดังนี้

กำหนด $\hat{\underline{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_n)'$

จะได้ว่า $\hat{\underline{y}} = x\hat{\underline{\beta}}$ (2.3)

ดังนั้น $\underline{\varepsilon}$ เป็นเวกเตอร์แนวตั้งของค่าผิดพลาด n ค่าจาก (2.3) ผลบวกกำลังสองของค่าผิดพลาดคือ

$$SSE = \underline{\varepsilon}'\underline{\varepsilon}$$

$$SSE = (\underline{y} - x\hat{\underline{\beta}})'(\underline{y} - x\hat{\underline{\beta}})$$

$$SSE = \underline{y}'\underline{y} - 2\hat{\underline{\beta}}'x'\underline{y} + \hat{\underline{\beta}}'x'x\hat{\underline{\beta}}$$

การหาค่าน้อยที่สุดของผลบวกกำลังสองของค่าผิดพลาดทำได้โดยการหาอนุพันธ์ (differentiate) เทียบกับ $\underline{\beta}$ แล้วกำหนดให้เท่ากับ $\underline{0}$ ดังนี้

$$\frac{\partial}{\partial \underline{\beta}} (\underline{y}'\underline{y} - 2\underline{\beta}'\underline{x}'\underline{y} + \underline{\beta}'\underline{x}'\underline{x}\underline{\beta}) = \underline{0}$$

$$-2\underline{x}'\underline{y} + 2\underline{x}'\underline{x}\underline{\beta} = \underline{0}$$

$$\underline{\beta} = (\underline{x}'\underline{x})^{-1}\underline{x}\underline{y} \qquad (2.4)$$

เมื่อ $(\underline{x}'\underline{x})^{-1}\underline{x}$ เป็นเมตริกซ์ของค่าคงที่ สมาชิกใน $\underline{\beta}$ เป็นฟังก์ชันเชิงเส้นของ $\underline{y}$ นั่นคือ $\underline{\beta}$ เป็นตัวประมาณเชิงเส้น และแทนสมการ (2.2) ในสมการ (2.4) จะได้สมการในรูปใหม่ คือ

$$\underline{\beta} = (\underline{x}'\underline{x})^{-1}\underline{x}'(\underline{x}\underline{\beta} + \underline{\varepsilon})$$

$$= \underline{\beta} + (\underline{x}'\underline{x})^{-1}\underline{x}'\underline{\varepsilon}$$

จะได้ว่า $E(\underline{\beta}) = \underline{\beta} + (\underline{x}'\underline{x})^{-1}\underline{x}'E(\underline{\varepsilon}) = \underline{\beta}$ ถ้า $E(\underline{\varepsilon}) = \underline{0}$ นั่นคือ $\underline{\beta}$ เป็นค่าที่ไม่เอนเอียง และ $V(\underline{\beta}) = E(\underline{\beta} - \underline{\beta})(\underline{\beta} - \underline{\beta})'$

$$= (\underline{x}'\underline{x})^{-1}\underline{x}'E(\underline{\varepsilon}'\underline{\varepsilon})\underline{x}'(\underline{x}'\underline{x})^{-1}$$

$$= \sigma^2(\underline{x}'\underline{x})^{-1} \qquad \text{โดยใช้ } E(\underline{\varepsilon}'\underline{\varepsilon}) = \sigma^2 I_n$$

เนื่องจากข้อมูลของตัวแปรตามที่จะใช้ศึกษาในครั้งนี้มีเพียง 2 ค่า คือ 1 หรือ 0 จึงมีความจำเป็นจะต้องมีการปรับค่าตัวแปรตามให้เป็น 1 หรือ 0 ตามสัดส่วนที่ต้องการ สำหรับเกณฑ์ที่เหมาะสมในการจำแนกค่าพยากรณ์ในสมการที่(2.3)ออกเป็น 2 กลุ่ม โดยวิธีนี้ คือ

$$y_i = \begin{cases} 1 & \text{เมื่อ } y_i \geq y^* \\ 0 & \text{เมื่อ } y_i < y^* \end{cases}$$

โดยที่

$$y^* = \frac{1}{2}\left[\frac{\sum_{i=1}^{n_1} y_{i1}}{n_1} + \frac{\sum_{i=1}^{n_0} y_{i0}}{n_0}\right]$$

$$= \frac{1}{2}(\hat{y}_1 + \hat{y}_0)$$

$\hat{Y}_1$ = ค่าเฉลี่ยของค่าพยากรณ์ที่ได้จากสมการถดถอยพหุ เมื่อตัวแปรตามมีค่าเป็น 1

$\hat{Y}_0$ = ค่าเฉลี่ยของค่าพยากรณ์ที่ได้จากสมการถดถอยพหุ เมื่อตัวแปรตามมีค่าเป็น 0

$n_1$ = จำนวนตัวอย่าง เมื่อตัวแปรตามมีค่าเป็น 1

$n_0$ = จำนวนตัวอย่าง เมื่อตัวแปรตามมีค่าเป็น 0

$n_1+n_0 = n$

2.3 การวิเคราะห์การถดถอยแบบทวิ ( Binary regression )

การวิเคราะห์การถดถอยทวิเป็นเทคนิคที่พัฒนาขึ้นมาเพื่อแก้ไขปัญหาและข้อขัดแย้งที่เกิดจากการวิเคราะห์การถดถอย โดยวิธีกำลังสองน้อยที่สุดแบบธรรมดา สำหรับการวิเคราะห์การถดถอยทวิจะมีลักษณะคล้ายคลึงกับการวิเคราะห์ถดถอย โดยใช้วิธีกำลังสองน้อยที่สุดทั่ว ๆ ไป แต่เพิ่มเติมขั้นตอนการแปลงข้อมูลซึ่งได้จากวิเคราะห์ถดถอย โดยวิธีกำลังสองน้อยที่สุด เพื่อให้ค่าพยากรณ์ของตัวแปรตามที่ประมาณได้มีค่าอยู่ในช่วง [0,1]

Arnold Zellner และ Tong Hun Lee(1965) ได้เสนอวิธีการประมาณค่าสัมประสิทธิ์การถดถอยแบบทวิ และการแปลงค่าพยากรณ์ให้อยู่ในช่วง [0,1] วิธีการที่ Zellner นำเสนอเป็นวิธีที่สามารถนำมาประยุกต์ใช้ เพื่อแก้ไขปัญหาดังกล่าวที่กำลังประสบอยู่ขณะนี้ สำหรับการศึกษาวิจัยนี้จะเลือกรูปแบบการแปลงข้อมูลให้สอดคล้องกับการแจกแจงที่กำหนด ด้วยเส้นโค้งของการแจกแจงต่าง ๆ ที่สำคัญ 3 รูปแบบ คือ

1. Normit Model (Normal Curve)
2. Logit Model (Logistic Curve)
3. Gompit Model (Gompertz Curve)

ในการประมาณค่าสัมประสิทธิ์การถดถอยแบบทวินี้ มีความจำเป็นที่จะต้องอาศัยวิธีกำลังสองน้อยที่สุดแบบทั่วไป (Generalized least square) มาช่วยในการประมาณค่าพารามิเตอร์ ดังนี้

วิธีกำลังสองน้อยที่สุดแบบทั่วไป (Generalized least Square method: GLS)

ปัญหาหนึ่งที่เราพบในการศึกษาวิจัยครั้งนี้พบว่า เกิดปัญหา Heteroscedasticity เราพบว่า $E(e_i e_j) \neq 0$ สำหรับทุกค่าของ $i \neq j$ และ $E(e_i e_j) \neq \sigma^2$ สำหรับทุกค่าของ $i = j$

นั่นคือ $E(\varepsilon'\varepsilon) = \sigma^2 V^2$ โดยที่ V เป็นเมตริกซ์บวกแน่นอน ( Positive definite) จึงสามารถที่จะหาเมตริกซ์ P ที่ทำให้ $PP' = V$ ได้

จากสมการ $y = x\beta + \varepsilon$ สามารถแปลงให้อยู่ในรูปแบบใหม่ได้ดังนี้

$$\underset{\sim}{y}^{*} = x^{*}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}^{*} \qquad (2.5)$$

โดยที่ $\underset{\sim}{y}^{*} = P^{-1}\underset{\sim}{y}$

$x^{*} = P^{-1}x$

และ $\underset{\sim}{\varepsilon}^{*} = P^{-1}$

จากสมการ (2.5) จะได้ว่า

$$E(\underset{\sim}{\varepsilon}^{*}\,'\underset{\sim}{\varepsilon}^{*}) = \underset{\sim}{0}$$

$$E(\underset{\sim}{\varepsilon}^{*}\,'\underset{\sim}{\varepsilon}^{*}) = \sigma^{2}I_{n}$$

ดังนั้น เราจึงสามารถประมาณค่าพารามิเตอร์ $\underset{\sim}{\beta}$ ได้โดยวิธีกำลังสองน้อยที่สุดดังที่กล่าวมาแล้วในหัวข้อ 2.2 และ ค่าประมาณของ $\underset{\sim}{\beta}$ สามารถเขียนได้เป็น

$$\underset{\sim}{\hat{\beta}}_{GLS} = (x^{*}\,'x^{*})^{-1}x^{*}\,'y^{*}$$

$$= (x'V^{-1}x)^{-1}(x'V^{-1}\underset{\sim}{y})$$

$$V(\underset{\sim}{\hat{\beta}}_{GLS}) = \sigma^{2}(x^{*}\,'x^{*})^{-1}$$

$$= \sigma^{2}(x'V^{-1}x)^{-1}$$

โดยปกติใน GLS เราถือว่า $\sigma^{2}$ เป็นพารามิเตอร์ที่ไม่ทราบค่า ในขณะเดียวกันที่ V เป็นเมตริกซ์ที่เราต้องทราบค่าของสมาชิกทุกตัวเพื่อประมาณค่า $\underset{\sim}{\beta}$ และความแปรปรวนของ $\underset{\sim}{\beta}$ ในทางปฏิบัติเรามักไม่ทราบค่าของ V จะทราบได้ก็โดยการประมาณค่า V ด้วยวิธีที่เหมาะสมสำหรับการศึกษาครั้งนี้จะใช้ค่าประมาณที่เสนอโดยวิธีของ Arnold Zellner และ Tong Hun Lee (1985) การประมาณค่า V ด้วย $\hat{V}$ นี้เรียกว่า Estimated Generalized Least Square (EGLS)

สำหรับการประมาณค่าโดยวิธีกำลังสองน้อยที่สุดในหัวข้อ 2.2 นั้น ทำให้เกิดปัญหาหลายประการ เพื่อแก้ไขปัญหาดังกล่าว Zeller และ Lee จึงได้เสนอวิธีการแปลงค่าพยากรณ์ของตัวแปรตามในรูปของความน่าจะเป็นด้วยเส้นโค้งของการแจกแจงแบบต่าง ๆ 3 รูปแบบ โดยมีขั้นตอนในการประมาณค่าดังนี้

ทั้ง 3 รูปแบบมีขั้นตอนในการประมาณค่าต่าง ๆ ดังนี้

1.) ประมาณค่า $\underset{\sim}{\beta}$ ด้วยวิธี OLS จะได้ $\underset{\sim}{\hat{\beta}} = (X'X)^{-1}X'\underset{\sim}{y}$

2.) กำหนดให้ $\underset{\sim}{t}$ เป็นค่าตัวกลาง( Intermediate value )ที่ได้จากสมการประมาณค่า : $\underset{\sim}{\hat{y}} = X\underset{\sim}{\hat{\beta}}$ โดยแทนค่า $\underset{\sim}{y}$ ด้วย $\underset{\sim}{t}$ ในขั้นตอนนี้จะได้ค่า $t_{i}$ บางค่าอยู่นอกช่วง [0,1] ซึ่งไม่สอดคล้องกับค่าที่แท้จริงของตัวแปรตาม

3.) แปลงค่าพยากรณ์ $\underset{\sim}{t}$ จากขั้นตอนที่ 2 ให้อยู่ในช่วง [0,1] ในรูปแบบของความน่าจะเป็นด้วยเส้นโค้งของการแจกแจง ต่อไปนี้

3.1 Normit Model : $\hat{P}_i = 1/\sqrt{2\pi} \int^{t_i} e^{-1/2z^2} dz$

3.2 Logit Model : $\hat{P}_i = 1/(1+e^{-t_i})$

3.3 Gompit Model : $\hat{P}_i = e^{et_i}$

4). การนำค่า $\underset{\sim}{\beta}$ ที่ได้จากขั้นตอนที่ 1 มาใช้ในการแปลงค่าพยากรณ์ จะทำให้ค่าพยากรณ์ $\underset{\sim}{y}$ และ $\underset{\sim}{\beta}$ ที่ได้ยังไม่มีความเหมาะสม ดังนั้นจึงต้องประมาณค่า $\underset{\sim}{\beta}$ ค่าใหม่ขึ้นมาโดยอาศัยค่า $\underset{\sim}{t}$ จากขั้นตอนที่ 1 และค่าประมาณของ V ที่เสนอโดย Zellner และ Lee

ดังนั้น จะได้ค่าประมาณสัมประสิทธิ์การถดถอยพหุชุดใหม่ ในแต่ละวิธี ดังนี้

$$\underset{\sim}{\beta}^{*}_{EGLS} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}\underset{\sim}{t}$$

เนื่องจาก $E(e_i e_j) = \sigma^2 V$ และ V เป็น Unknown Matrix จำเป็นต้องประมาณค่า V ด้วย $\hat{V}$ ก่อนการวิเคราะห์

สำหรับเมตริกซ์ X และเวคเตอร์ $\underset{\sim}{t}$ จะเหมือนกันสำหรับแต่ละรูปแบบ แต่จะต่างกันที่เมตริกซ์ $\hat{V}$ (Diagonal Matrix) ค่าของ $\hat{V}_i$ ในแต่ละวิธีคำนวณได้จาก

Normit Model:

$$\hat{V}_i = (\hat{P}_i\hat{Q}_i)/n_i[Z(\hat{P}_i)]^2 \ldots, Z(\hat{P}_i) = 1/\sqrt{2\pi}\, e^{-1/2P_i^2}$$

Logit Model:

$$\hat{V}_i = 1/n_i\hat{P}_i\hat{Q}_i$$

Gompit Model:

$$\hat{V}_i = \hat{P}_i\hat{Q}_i/(\hat{P}_i \ln\hat{P}_i)^2$$

สำหรับค่า $t_i$, $\hat{P}_i$ และ $\hat{Q}_i$ ได้จากขั้นตอนที่ 2 และ 3

5). สมการในการประมาณค่าชุดใหม่ คือ $\underset{\sim}{t}^{*} = X\underset{\sim}{\beta}^{*}$ และนำค่า $\underset{\sim}{t}^{*}$ ที่ได้นี้มาหาค่าประมาณของ $\underset{\sim}{\hat{P}}$ โดยใช้การแปลงข้อมูลตามขั้นตอนที่ 3 กำหนดให้เป็นค่าเดิมคือ $\underset{\sim}{\hat{P}}$

6). นำค่า $\hat{\underset{\sim}{P}}$ ที่ได้มาใช้เพื่อพยากรณ์ตัวแปรตามว่ามีค่าเป็น 1 หรือ 0 โดยใช้เกณฑ์เดียวกับการวิเคราะห์การถดถอยโดยวิธีกำลังสองน้อยที่สุดแบบธรรมดา แต่เปลี่ยนค่าพยากรณ์จาก $\underset{\sim}{y}$ เป็น $\underset{\sim}{P}$ ดังนี้

$$y_i = \begin{cases} 1 \text{ เมื่อ } P_i \geq P^* \\ 0 \text{ เมื่อ } P_i < P^* \end{cases}$$

โดยที่ $$P^* = \frac{1}{2}\left(\hat{P}_1 + \hat{P}_0\right)$$

2.4 <u>การวิเคราะห์จำแนกประเภท</u> (Discriminant analysis)

การจำแนกกลุ่มและจัดเข้ากลุ่มกรณี 2 ประชากร เป็นวิธีการที่มุ่งจำแนกวัตถุออกเป็น 2 กลุ่ม ซึ่งเราทราบจำนวนกลุ่มชัดเจนแล้ว ในการแบ่งกลุ่มจะต้องมีตัวแปรอิสระที่จะมาช่วยในการสร้างเกณฑ์ เพื่อจำแนกกลุ่มและจัดสมาชิกใหม่ที่เพิ่มเข้ากลุ่มในกรณีนี้ถือว่ามี 2 ประชากร ดังนั้น จะเก็บข้อมูลจากประชากร $\pi_1$ มา $n_1$ หน่วยและจากประชากร $\pi_0$ มา $n_0$ หน่วย ($n_1+n_0 = n$) ในการเก็บข้อมูลจะบันทึกคุณลักษณะประจำตัวของวัตถุแต่ละหน่วย คุณลักษณะดังกล่าวอาจจะมีมากกว่า 1 คุณลักษณะ ซึ่งเรียกว่า จำนวนตัวแปรอิสระ (P) ในการศึกษาวิจัยนี้จะถือว่าตัวแปรอิสระมาจากประชากรที่มีการแจกแจงแบบเบ้ และการจัดเข้ากลุ่มจะนำเอาความสูญเสียจากการจัดเข้ากลุ่มผิด ( Cost of Misclassification ) และ ความน่าจะเป็นโดยหลักเกณฑ์ ( Prior Probability) มาเป็นเกณฑ์ร่วมกับ Fisher's Discriminant function (เป็นกรณีเฉพาะของเกณฑ์ใหม่) การจัดเกณฑ์ในลักษณะนี้เรียกว่า "Expected Cost of Misclassification (ECM)"

<u>การวิเคราะห์จำแนกกลุ่มในกรณีที่มีประชากร 2 กลุ่ม</u>

ถ้าการจำแนกกลุ่มเรามีตัวแปรที่ต้องการศึกษาอยู่ P ตัว มีประชากร 2 กลุ่ม คือ $\pi_1$ และ $\pi_0$ สามารถเขียนเวกเตอร์ของตัวแปรอิสระได้ดังนี้

$$X' = [x_1, x_2, x_3, \ldots, x_p]$$

สำหรับเกณฑ์การจัดกลุ่มนี้ฟิชเชอร์(Fisher, R.A.,1938) แนะนำให้ใช้ข้อมูลจากตัวแปรทุกตัว โดยถ่วงน้ำหนักด้วยวิธีเหมาะสม นั่นคือให้ใช้การประกอบกันของ

$x_1, x_2, x_3, \ldots x_p$ คือ

$$y = l_1 x_1 + l_2 x_2 + l_3 x_3 + \ldots + l_p x_p = l' x$$

โดยที่ l คือ เวคเตอร์ของน้ำหนักที่มีผลให้อัตราส่วนต่อไปนี้มีค่าสูงสุด

$$\delta = \frac{(\text{ระยะทางระหว่างค่าเฉลี่ยของ } Y)^2}{\text{ความแปรปรวนของ } Y}$$

$$\delta = \frac{(\mu_{1y} - \mu_{0y})^2}{l' \Sigma l}$$

$$= \frac{l' (\mu_1 - \mu_0)(\mu_1 - \mu_0) l}{l' \Sigma l} \qquad (2.6)$$

กำหนดให้ $\pi 1$ เป็นประชากรกลุ่มที่ตัวแปรตามมีค่าเป็น 1

$\pi 0$ เป็นประชากรกลุ่มที่ตัวแปรตามมีค่าเป็น 0

$\mu_1 = E(x/\pi 1)$ ค่าเฉลี่ยของเวคเตอร์ x ที่บันทึกจากประชากร $\pi 1$

$\mu_0 = E(x/\pi 0)$ ค่าเฉลี่ยของเวคเตอร์ x ที่บันทึกจากประชากร $\pi 0$

ดังนั้น ค่าเฉลี่ยของ Y ที่เกิดจาก linear combination ของ X's จากประชากร $\pi 1$ และ $\pi 0$ จึงมีค่าเป็น $\mu_{1y}$ และ $\mu_{0y}$ ตามลำดับ ดังนี้

$$\mu_{1y} = E(Y/\pi 1) = E(l' x/\pi 1) = l' \mu_1$$

$$\mu_{0y} = E(Y/\pi 0) = E(l' x/\pi 0) = l' \mu_0$$

ส่วนความแปรปรวนของ Y จากทั้ง 2 ประการ คือ

$$\sigma^2_y = \text{var}(l' x) = l' \text{Cov}(x) l$$

$$= l' \Sigma l$$

โดยที่

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2p} \\ \vdots & & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_{pp} \end{bmatrix} \quad ; \quad l' = [\, l_1, l_2, \ldots, l_p \,]$$

การหาสมการที่เหมาะสมในการแบ่งกลุ่มประชากรทั้งสองออกจากกันให้ได้มากที่สุดคือ การหาค่า l ที่ทำให้ $\theta$ มีค่ามากที่สุด ในทางปฏิบัติค่า $\mu_1$ , $\mu_0$ และ $\Sigma$ เป็นค่าที่เรามักจะไม่ทราบ ดังนั้น ถ้าสุ่มตัวอย่างจาก $\pi 1$ และ $\pi 0$ มาจำนวน n1 และ n0 ตามลำดับแล้ววัดค่าสังเกตเพื่อประมาณค่า $\mu_1$ $\mu_0$ และ $\Sigma$

ถ้าแบ่งเมตริกซ์ X ออกเป็น 2 ส่วน ดังนี้

$$X' = \left[\begin{array}{cccc:cccc} x_{111} & x_{112} & \cdots & x_{11n_1} & x_{211} & x_{212} & \cdots & x_{21n_0} \\ x_{121} & x_{122} & \cdots & x_{12n_1} & x_{221} & x_{222} & \cdots & x_{22n_0} \\ \vdots & \vdots & & & \vdots & \vdots & & \\ x_{1p1} & x_{1p2} & \cdots & x_{1pn_1} & x_{2p1} & x_{2p2} & \cdots & x_{2pn_0} \end{array}\right]$$

หรือ $X = [\, x_1' \mid x_0' \,]$

$x_1'$ และ $x_0'$ เป็นเมตริกซ์ของค่าสังเกต ซึ่งแต่ละเมตริกซ์จะได้เวกเตอร์ค่าเฉลี่ยดังนี้

$$x_1' = [\, x_{11} \;\; x_{12} \;\; \cdots \;\; x_{1p} \,]$$
$$x_0' = [\, x_{01} \;\; x_{02} \;\; \cdots \;\; x_{0p} \,]$$

และประมาณค่า Covariance Matrix $\Sigma$ ด้วย $S_p$ คือ Pooled Covariance Matrix

$$S_p = \frac{(n_1 - 1)\, S_1 + (n_0 - 1)\, S_0}{n_1 + n_0 - 2}$$

เมื่อ

$$S_1 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (x_{1j} - x_1)(x_{1j} - x_1)' \qquad (p \times p)$$

$$S_0 = \frac{1}{n_0 - 1} \sum_{j=1}^{n_0} (x_{0j} - x_0)(x_{0j} - x_0)' \qquad (p \times p)$$

จากสมการ (2.6) จะได้

$$= \frac{l'(\bar{x}_1 - \bar{x}_0)(\bar{x}_1 - \bar{x}_0)'l}{l'S_p l}$$

จะมีค่ามากที่สุด เมื่อ $\frac{\partial \delta}{\partial l} = 0$

$\therefore$ จะได้ $\hat{l} = c\, S_p^{-1} (\bar{x}_1 - \bar{x}_0)'$ ; $c = 0$

และเมื่อ $c = 1$ จะได้เวคเตอร์ $l$ เรียกว่า Fisher's linear discriminant function ดังนี้

$$y = l'x$$

$$= (\bar{x}_1 - \bar{x}_0)' S_p^{-1} x$$

ซึ่งจะมีผลให้อัตราส่วน $\delta$ มีค่าสูงสุด

ในการจำแนกกลุ่มโดยวิธีการวิเคราะห์จำแนกประเภทสามารถสรุปเป็นขั้นตอนในการจำแนกกลุ่มได้ดังต่อไปนี้ โดยจะพิจารณาค่า Prior Probability และ Cost of Misclassification ประกอบการจำแนกกลุ่มด้วย

1. คำนวณสมการจำแนกกลุ่ม

   Fisher's Linear Discriminant function

   $$y = l'x = (\bar{x}_1 - \bar{x}_0)' S_p^{-1} x$$

2. คำนวณค่า $y_{new}$ คือ ค่าของ discriminant function สำหรับค่าสังเกต $x_{new}$ หน่วยใหม่ที่ปรากฏภายหลัง หรือค่าสังเกตที่ต้องการพยากรณ์ โดยที่แทนค่า x ด้วย $x_{new}$ ในสมการขั้นตอนที่ 1

3. คำนวณค่า Midpoint ($\hat{m}$) คือ $\hat{m} = \frac{1}{2}(\bar{x}_1 - \bar{x}_0)' S_p^{-1} (\bar{x}_1 + \bar{x}_0)$

4. จัด $X_{new}$ เข้ากลุ่ม $\pi_1$ หรือ $\pi_0$ โดยใช้เกณฑ์ร่วมใหม่ที่พยายามทำให้ค่า EMC มีค่าต่ำที่สุด

   ถ้า $y_{new} - m \geq \ln \frac{c(1/0)}{c(0/1)} \cdot \frac{P_0}{P_1}$ จะจัด $x_{new}$ เข้ากลุ่ม $\pi_1$

ถ้า $y_{new} - m < \ln \frac{c(1/0)}{c(0/1)} \cdot \frac{P_0}{P_1}$ จะจัด $x_{new}$ เข้ากลุ่ม $\pi 0$

อสมการนี้ เรียกว่า Anderson's classification rule

$c(1/0)$ ความสูญเสียที่เกิดจากการจัดวัตถุเข้ากลุ่ม $\pi 1$ ทั้งที่ความจริงวัตถุนั้นมาจากกลุ่ม $\pi 0$

$c(0/1)$ ความสูญเสียที่เกิดจากการจัดวัตถุเข้ากลุ่ม $\pi 0$ ทั้งที่ความจริงวัตถุนั้นมาจากกลุ่ม $\pi 1$

การวิจัยครั้งนี้จะไม่นำค่า $c(1/0)$ และ $c(0/1)$ มาพิจารณา เพราะในทางปฏิบัติการที่จะกำหนดเท่าใดนั้น จะขึ้นอยู่กับผู้วิจัยและลักษณะของงานวิจัย ดังนั้นจึงถือว่า $c(1/0)$ และ $c(0/1)$ มีค่าเท่ากัน แต่ในทางปฏิบัติควรจะคำนึงความสูญเสีย ดังกล่าวเสมอ

---

**Prior Probability** : $P_1, P_0$ หมายถึง ความน่าจะเป็นโดยหลักเกณฑ์โอกาสที่จะเป็นสมาชิกของกลุ่ม $\pi 1$ และ $\pi 0$ ตามลำดับ จะนำมาพิจารณาในการวิจัยนี้เนื่องจากทราบความแตกต่างในขนาดของประชากร เพราะว่าถ้าประชากรกลุ่มใดมีขนาดใหญ่กว่าโอกาสที่วัตถุนั้นจะเป็นสมาชิกของกลุ่มนั้นย่อมมีมากกว่า