หอสมุดกลาง สถาบันวิทยบริการ
จุฬาลงกรณ์มหาวิทยาลัย

บทที่ 2

ระเบียบวิธีที่ใช้ในการวิจัย

ในการศึกษาเปรียบเทียบการวิเคราะห์เพื่อใช้ในการแยกกลุ่ม 2 กลุ่ม

โดยวิธีการวิเคราะห์ 2 วิธีด้วยกันคือ

1. การวิเคราะห์ความถดถอยเชิงซ้อน
2. การวิเคราะห์จำแนกกลุ่ม

2.1 <u>การวิเคราะห์ความถดถอยเชิงซ้อน</u>

เป็นการศึกษาความสัมพันธ์ระหว่างตัวแปรเพื่อประโยชน์ในการทำนาย (Predict) หรือคาดประมาณ (Estimate) ค่าของตัวแปรตามจากค่าของตัวแปรอิสระ ถ้ามีตัวแปรอิสระเพียงตัวเดียว เรียกว่า การวิเคราะห์ความถดถอยอย่างง่าย (Simple Regression Analysis) แต่ถ้ามีตัวแปรอิสระตั้งแต่ 2 ตัวขึ้นไป เรียกว่า การวิเคราะห์ความถดถอยเชิงซ้อน (Multiple Regression Analysis) ในทางปฏิบัติมักจะนิยมใช้ตัวแปรอิสระหลายตัวในการพยากรณ์ตัวแปรตาม ดังนั้นในการวิจัยนี้จึงมุ่งศึกษาการวิเคราะห์ความถดถอยเชิงซ้อน

ตัวแบบและสมการ การวิเคราะห์ความถดถอยเชิงซ้อน

ตัวแบบความถดถอยเชิงซ้อน $Y_i = \alpha + \beta_1X_{1i}+\beta_2X_{2i}+\ldots+\beta_kX_{ki}+\Sigma_i$ ....(1)

ตัวแบบความถดถอยเชิงซ้อนตัวอย่าง $Y_i = a + b_1X_{1i}+b_2X_{2i}+\ldots+b_kX_{ki}+e_i$ ....(2)

สมการถดถอยเชิงซ้อน $\mu_Y = \alpha + \beta_1X_1 +\beta_2X_2 +\ldots+\beta_kX_k$ ....(3)

สมการถดถอยเชิงซ้อนตัวอย่าง $\hat{Y} = a + b_1X_1 +b_2X_2 +\ldots+b_kX_k$ ....(4)

โดยที่ Y เป็นตัวแปรตาม

$X_1, X_2, \ldots, X_k$ เป็นตัวแปรอิสระ

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

$a, b_1, b_2, \ldots, b_k$ เป็นค่าประมาณของ $\alpha, \beta_1, \beta_2, \ldots\ldots, \beta_k$ ตามลำดับ ซึ่งเรียกว่าสัมประสิทธิ์ความถดถอย

2.1.1 การประมาณค่าสัมประสิทธิ์ความถดถอย

จากตัวแบบความถดถอยตัวอย่าง (2) จะเขียนในรูปของเมตริกซ์ได้ดังนี้

Sean

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{E} \qquad \ldots.(5)$$

เมื่อ

$$\underset{\sim}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{nX1} \qquad X = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & & & & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix}_{nx(k+1)}$$

$$\underset{\sim}{\beta} = \begin{bmatrix} a \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix}_{(k+1)x1} \qquad \underset{\sim}{E} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}_{nx1}$$

หาค่าเมตริกซ์ $\underset{\sim}{\beta}$ โดยใช้วิธีกำลังสองน้อยที่สุด (Ordinary Least Square)

$$\begin{aligned} \sum_{i=1}^{n} e_i^2 &= \underset{\sim}{E}'\underset{\sim}{E} \\ &= (\underset{\sim}{Y} - X\underset{\sim}{\beta})'(Y - X\underset{\sim}{\beta}) \\ &= \underset{\sim}{Y}'\underset{\sim}{Y} - 2\underset{\sim}{\beta}'X'\underset{\sim}{Y} + \underset{\sim}{\beta}'X'X\underset{\sim}{\beta} \end{aligned}$$

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

$$\frac{\partial \underset{\sim}{e}'\underset{\sim}{e}}{\partial \underset{\sim}{\beta}} = -2\ X'\underset{\sim}{Y} + 2X'X\underset{\sim}{\beta}$$ และกำหนดให้เท่ากับศูนย์

นั่นคือ $$X'X\underset{\sim}{\beta} = X'\underset{\sim}{Y}$$

scan

$$\underset{\sim}{\beta} = (X'X)^{-1}\ X'\underset{\sim}{Y} \qquad ....(6)$$

จากสมการ (6) ค่าสัมประสิทธิ์ความถดถอยที่ได้เป็นค่าแสดงการเปลี่ยนแปลงค่าเฉลี่ยของ Y เมื่อ X_i เปลี่ยนไป 1 หน่วยขณะที่ตัวแปรอิสระอื่น ๆ คงที่ และค่าสัมประสิทธิ์ความถดถอยคะแนนดิบ (Unstandardized Coefficient) นี้ เป็นค่าซึ่งใช้ในการประมาณค่า Y เท่านั้น ถ้าต้องการเปรียบเทียบความสำคัญของตัวแปรอิสระที่มีต่อตัวแปรตามจะทำได้โดยแปลงค่าสัมประสิทธิ์ความถดถอยคะแนนดิบ (b_i) ให้เป็นค่าสัมประสิทธิ์คะแนนมาตรฐาน (Standardized Coefficient)

โดย $$B_i = b_i \frac{S_{Xi}}{S_Y}$$ scan

เมื่อ B_i = ค่าสัมประสิทธิ์คะแนนมาตรฐาน

S_{X_i} = ส่วนเบี่ยงเบนมาตรฐานของ X_i

S_Y = ส่วนเบี่ยงเบนมาตรฐานของ Y

จากสมการความถดถอย

$$\hat{Y} = a + b_1\ X_1 + b_2\ X_2 + \ldots\ldots + b_k\ X_k$$

นำสัมประสิทธิ์ความถดถอย $b_1\ b_2\ ,\ldots.,b_k$ ที่ได้จากสมการ (6) มาทดสอบความมีนัยสำคัญ โดยทดสอบสมมติฐาน

$$H_o\ :\ \beta_1 = \beta_2 = \ldots. = \beta_k = 0$$

ค่าสถิติที่ใช้ในการทดสอบคือ $$F = \frac{MSR}{MSE}$$

ค่า MSR และ MSE คำนวณได้จากตารางวิเคราะห์ความแปรปรวน (Analysis of Variance) ดังนี้

ตารางวิเคราะห์ความแปรปรวน

แหล่งความแปรผัน	ระดับแห่งความเป็นอิสระ	ผลบวกกำลังสอง	ผลบวกกำลังสองเฉลี่ย	F
ความถดถอย	k	$\underset{\sim}{\beta}'X'\underset{\sim}{Y} - n\bar{Y}^2 = SSR$	$\frac{SSR}{k} = MSR$	$\frac{MSR}{MSE}$
ความคลาดเคลื่อน	$n-k-1$	$\underset{\sim}{Y}'\underset{\sim}{Y} - \underset{\sim}{\beta}'X'\underset{\sim}{Y} = SSE$	$\frac{SSE}{n-k-1} = MSE$	
ยอดรวม	$n-1$	$\underset{\sim}{Y}'\underset{\sim}{Y} - n\bar{Y}^2 = SST$		

และ $$R^2 = \frac{SSR}{SST}$$

R^2 นี้เรียกว่าสัมประสิทธิ์การตัดสินใจ (Coefficent of Determination) ซึ่งจะมีค่าอยู่ระหว่าง 0 กับ 1

$R^2 \times 100$ จะหมายถึงร้อยละของความแปรปรวนของตัวแปรตามที่สามารถอธิบายได้ด้วยตัวแปรอิสระ

2.1.2 <u>การคัดเลือกตัวแปรอิสระเข้าสมการความถดถอยเชิงซ้อน โดยวิธีถดถอยขั้นบันได</u>

<u>ขั้นที่ 1</u> 1. คำนวณหาค่าสัมประสิทธิ์สหสัมพันธ์อย่างง่าย r_{x_iY} (Simple Correlation Coefficient) ระหว่างตัวแปรอิสระทั้งหมด

เมื่อ $$r_{X_iY} = \frac{\sum_{i=1}^{n} X_iY_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right)\left(\sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}\right)}}$$

ตัวแปรอิสระใดที่ให้ค่าสัมประสิทธิ์สหสัมพันธ์กับตัวแปรตามสูงสุด ตัวแปรอิสระนั้นจะได้รับการคัดเลือกเข้าสมการก่อน สมมติว่า r_{X_1Y} มีค่าสูงสุด ดังนั้น X_1 จึงเป็นตัวแปรที่เข้าสมการ

ศูนย์วิทยทรัพยากร จุฬาลงกรณ์มหาวิทยาลัย

2. ทดสอบนัยสำคัญ เมื่อ X_1 เป็นตัวแปรอิสระที่เราเลือกมาพิจารณาสร้างสมการก่อน โดยทดสอบ $H_o : \beta_1 = 0$ ถ้ามีนัยสำคัญทางสถิติ ก็ได้สมการความถดถอยเป็น $\hat{Y} = a + b_1 X_1$ ถ้าไม่มีนัยสำคัญทางสถิติก็จะหยุดการคำนวณ

<u>ขั้นที่ 2</u> 1. คำนวณค่าสัมประสิทธิ์สหสัมพันธ์บางส่วน (Partial Correlation Coefficient) ระหว่างตัวแปรอิสระที่เหลือ X_j ($j \neq 1$) กับตัวแปรตาม Y โดยกำหนดให้ตัวแปรอิสระ X_1 ซึ่งเลือกไว้ในสมการแล้วคงที่ ($r_{X_j Y . X_1}$)

เมื่อ

$$r_{X_j Y . X_1} = \frac{r_{X_j Y} - r_{X_j X_1} r_{Y X_1}}{\sqrt{(1 - r^2_{X_j X_1})(1 - r^2_{Y X_1})}}$$

ตัวแปรอิสระที่เหลือ X_j ตัวใด ที่ให้ค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนกับตัวแปรตาม Y สูงสุดจะได้รับการคัดเลือกเข้าในสมการถดถอยต่อไป สมมติว่า X_2 เป็นตัวแปรอิสระที่ถูกเลือกเพิ่มเข้าไปในสมการ และทำการทดสอบ $H_o : \beta_2 = 0$ ถ้าพบว่าไม่มีนัยสำคัญทางสถิติ ก็จะตัด X_2 ออกจากสมการ และหยุดการคำนวณ แต่ถ้ามีนัยสำคัญทางสถิติ ก็แสดงว่า X_2 มีความสัมพันธ์กับ Y

2. ทดสอบ Partial F ของ X_1 และ X_2 ถ้าพบว่าตัวแปรอิสระใดไม่มีนัยสำคัญทางสถิติก็จะตัดตัวแปรนั้นออกจากสมการ แต่ถ้าผลการทดสอบมีนัยสำคัญทางสถิติ ก็จะรวมตัวแปรนั้นเข้าในสมการ

<u>ขั้นที่ 3</u> ทำเช่นเดียวกับขั้นที่ 2 คือ หาค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนระหว่างตัวแปรอิสระที่เหลือกับตัวแปรตาม Y โดยกำหนดให้ X_1 และ X_2 ซึ่งเลือกไว้แล้วในสมการถดถอยนั้นคงที่ ($r_{X_j Y . X_1 X_2}$)

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

เมื่อ

$$r_{X_jY.X_1X_2} = \frac{r_{X_jY.X_2} - r_{X_jX_1.X_2}\, r_{YX_1.X_2}}{\sqrt{(1 - r^2_{X_jX_1.X_2})(1-r^2_{YX_1.X_2})}}$$

$$= \frac{r_{X_jY.X_1} - r_{X_jX_2.X_1}\, r_{YX_2.X_1}}{\sqrt{(1 - r^2_{X_jX_2.X_1})(1-r^2_{YX_2.X_1})}}$$

ตัวแปรอิสระใด ให้ค่าสัมประสิทธิ์สหสัมพันธ์บางส่วนสูงสุด จะได้รับการคัดเลือกเข้าสมการถดถอยต่อไป และทำการทดสอบเช่นเดียวกับในขั้นที่ 2 จนกว่าจะไม่มีตัวแปรอิสระใดเป็นที่ยอมรับ

2.2 การวิเคราะห์จำแนกกลุ่ม

เป็นวิธีการวิเคราะห์ที่มีวัตถุประสงค์ที่จะคัดเลือกตัวแปรกลุ่มหนึ่งหรือชุดหนึ่ง ที่นักวิจัยคิดว่าตัวแปรเหล่านี้มีความสัมพันธ์กับสิ่งที่ต้องการศึกษา (ตัวแปรตาม) จนถึงขั้นที่สามารถแยกประชากรออกเป็นกลุ่มต่าง ๆ ได้ แล้วนำเอาตัวแปรชุดนี้มาใช้ในการประมาณการเป็นสมาชิกของกลุ่ม สมการที่ได้นี้เรียกว่า สมการจำแนกกลุ่ม (Discriminant Function) จำนวนของสมการที่ได้จะน้อยกว่าจำนวนกลุ่มของประชากรที่ต้องการจำแนก 1 กลุ่มเสมอ

2.2.1 การวิเคราะห์จำแนกกลุ่ม ในกรณีที่มีประชากร 2 กลุ่ม

ถ้าเรามีตัวแปรที่ต้องการศึกษาอยู่ P ตัว

$$X' = [X_1, X_2, X_3, \ldots\ldots X_P]$$

จะได้สมการจำแนกกลุ่มคือ

$$Y = V_1X_1+V_2X_2+V_3X_3+\ldots\ldots+V_PX_P \qquad \ldots\ldots(7)$$

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

หอสมุดกลาง สถาบันวิทยบริการ จุฬาลงกรณ์มหาวิทยาลัย

กำหนดให้ π_1 เป็นประชากรกลุ่มที่ 1

π_2 เป็นประชากรกลุ่มที่ 2

μ_{1y} = ค่าเฉลี่ยของ Y ซึ่งได้มาจากค่า X ของ π_1

μ_{2y} = ค่าเฉลี่ยของ Y ซึ่งได้มาจากค่า X ของ π_2

μ_1 = ค่าคาดหวังของตัวแปร X ที่มาจาก π_1

μ_2 = ค่าคาดหวังของตัวแปร X ที่มาจาก π_2

จากสมการ (7) พิจารณาผลรวมเชิงเส้น

$$\underset{(1\times1)}{Y} = \underset{(1\times P)}{V'}\ \underset{(P\times1)}{X} \qquad .....(8)$$

จะได้ $$\mu_{1y} = E(Y/\pi_1) = E(V'X/\pi_1) = V'\mu_1 \qquad .....(9)$$

$$\mu_{2y} = E(Y/\pi_2) = E(V'X/\pi_2) = V'\mu_2 \qquad .....(10)$$

ความแปรปรวนของ Y จากทั้ง 2 ประชากร คือ

$$\sigma_Y^2 = Var(V'X) = V'\ Cov(X)V$$
$$= V'\,\Sigma\,V \qquad .....(11)$$

การหาสมการที่เหมาะสมในการแบ่งกลุ่ม ประชากรทั้ง 2 กลุ่มออกจากกันได้มากที่สุด ก็คือ การหาค่า V ที่ทำให้ λ มีค่ามากที่สุด โดย

$$\lambda = \frac{(\text{ระยะทางระหว่างค่าเฉลี่ยของY})^2}{\text{ความแปรปรวนของ Y}}$$

$$= \frac{(\mu_{1y} - \mu_{2y})^2}{V'\,\Sigma\,V}$$

$$= \frac{V'(\mu_1-\mu_2)(\mu_1-\mu_2)'V}{V'\,\Sigma\,V} \qquad .....(12)$$

ในทางปฏิบัติค่า μ_1, μ_2 และ Σ เป็นค่าที่เรามักจะไม่ทราบ ดังนั้นถ้าเราสุ่มตัวอย่างจาก π_1 และ π_2 มาจำนวน n_1 และ n_2 ตามลำดับแล้ววัดค่าสังเกต

ศูนย์วิทยทรัพยากร จุฬาลงกรณ์มหาวิทยาลัย

$$X' = \left[\begin{array}{cccc|cccc} x_{111} & x_{112} & \cdots & x_{11n_1} & x_{211} & x_{212} & \cdots & x_{21n_2} \\ x_{121} & x_{122} & \cdots & x_{12n_1} & x_{221} & x_{222} & \cdots & x_{22n_2} \\ \vdots & \vdots & & \vdots & \vdots & & & \vdots \\ x_{1p1} & x_{1p2} & \cdots & x_{1pn_1} & x_{2p1} & x_{2p2} & \cdots & x_{2pn_2} \end{array}\right]$$

แบ่ง เมตริกซ์ X' ออกเป็น 2 ส่วนโดยใช้เส้นขีดตามแนวตั้งดังนี้

$$X' = [X_1' : X_2']$$

X_1' และ X_2' เป็นเมตริกซ์ของค่าสังเกต ซึ่งแต่ละเมตริกซ์จะได้เวคเตอร์ค่าเฉลี่ยดังนี้

$$\bar{x}_1' = [\bar{x}_{11}\ \bar{x}_{12}\ \cdots\cdots\ \bar{x}_{1p}]$$

$$\bar{x}_2' = [\bar{x}_{21}\ \bar{x}_{22}\ \cdots\cdots\ \bar{x}_{2p}]$$

และประมาณ Covariance Matrix Σ ด้วย S_* คือ Pooled Covariance Matrix

$$\underset{(P\times P)}{S_*} = \left[\frac{(n_1-1)}{(n_1-1)+(n_2-1)}\right]S_1 + \left[\frac{(n_2-1)}{(n_1-1)+(n_2-1)}\right]S_2$$

เมื่อ $$\underset{(P\times P)}{S_1} = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(x_{1j}-\bar{x}_1)(x_{1j}-\bar{x}_1)'$$

$$\underset{(P\times P)}{S_2} = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(x_{2j}-\bar{x}_2)(x_{2j}-\bar{x}_2)'$$

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

ดังนั้นจากสมการ (12) จะได้

$$\hat{\lambda} = \frac{V'(\bar{X}_1-\bar{X}_2)(\bar{X}_1-\bar{X}_2)'V}{V' S_* V} \qquad \ldots\ldots\ldots(13)$$

$\hat{\lambda}$ จะมีค่ามากที่สุดเมื่อ $\frac{\partial\hat{\lambda}}{\partial V} = 0$

$\therefore$ จะได้ $c(\bar{X}_1-\bar{X}_2) - S_* V = 0$

$$\hat{V} = c\, S_*^{-1}(\bar{X}_1-\bar{X}_2)$$

กำหนดให้ C =1 ดังนั้น $\hat{V} = S_*^{-1}(\bar{X}_1-\bar{X}_2)$

แทนค่า $\hat{V}$ ในสมการ (13)

$$\therefore \hat{\lambda} = (\bar{X}_1-\bar{X}_2)'\, S_*^{-1}\, (\bar{X}_1-\bar{X}_2)$$

$$= D^2$$

ซึ่งเรียกว่า Mahalanobis D^2 หรือ Generalized Distance

จากสมการที่ (8) จะได้ผลรวมเชิงเส้น

$$\hat{Y} = \hat{V}'\, X = (\bar{X}_1-\bar{X}_2)'\, S_*^{-1}\, X \qquad \ldots\ldots\ldots\ldots(14)$$

จากสมการ (14) ค่า $\hat{V}_i$ จะเป็นค่าแสดงถึงความสำคัญของตัวแปร $X_1, X_2, \ldots, X_p$ ซึ่งโดยปกติ X_i แต่ละค่านั้นจะมีหน่วยไม่เหมือนกัน การเปรียบเทียบค่า $\hat{V}_i$ เหล่านี้จะทำได้ก็โดยการปรับ $\hat{V}_i$ ให้เป็นคะแนนมาตรฐาน (Standardized) เสียก่อน โดยนำเอาสมาชิกในแนวเส้นทะแยงมุม (Diagonal) ของเมตริกซ์ W มาถอดรากที่สอง แล้วนำไปคูณกับค่า $\hat{V}_i$ ทุก ๆ ค่าตามสูตรต่อไปนี้

$$V_i^* = (\sqrt{W_{ii}})(\hat{V}_i)$$

เมื่อ W คือ $(\underset{\sim}{X} - \bar{\underset{\sim}{X}}_{g.})(X - \bar{X}_{g.})'$

$$\bar{X}_{g.} = \left[\begin{array}{ccc|ccc} \bar{X}_{11} & \cdots\cdots & \bar{X}_{11} & \bar{X}_{12} & \cdots\cdots & \bar{X}_{12} \\ \bar{X}_{21} & \cdots\cdots & \bar{X}_{21} & \bar{X}_{22} & \cdots\cdots & \bar{X}_{22} \\ \vdots & & \vdots & \vdots & & \vdots \\ \bar{X}_{p1} & \cdots\cdots & \bar{X}_{p1} & \bar{X}_{p2} & \cdots\cdots & \bar{X}_{p2} \end{array}\right]$$

เมื่อ V_i^* คือ Standardized Discriminant Weight

V_i คือ Raw Discriminant Weight

W_{ii} คือ Diagonal Element ของเมตริกซ์ W

จาก $\hat{Y} = (\bar{X}_1-\bar{X}_2)'S_*^{-1} X$

สมการนี้สามารถใช้เป็นเกณฑ์ในการแยกค่าสังเกตที่ได้มานั้นว่าจะอยู่ในประชากรกลุ่ม Π_1 หรือ Π_2 ได้โดยให้

$Y_o = (\bar{X}_1-\bar{X}_2)'\ S_*^{-1}\ X_o$ เป็นค่าของ Discriminant Function ของค่าสังเกต X_o และกำหนดให้ m เป็นค่าจุดกึ่งกลางระหว่าง μ_{1y} และ μ_{2y} สำหรับตัวอย่าง

ค่า $m = \frac{1}{2}(\bar{Y}_1+\bar{Y}_2)$ โดย $\bar{Y}_1 = \hat{V}'\bar{X}_1$ และ $\bar{Y}_2 = \hat{V}'\bar{X}_2$

$\therefore\ m = \frac{1}{2}(\bar{X}_1-\bar{X}_2)'\ S_*^{-1}\ (\bar{X}_1+\bar{X}_2)$

ดังนั้นจะได้กฎการจำแนกกลุ่ม คือ

จัดค่าสังเกต (X_o) อยู่ใน Π_1 ถ้าหาก $Y_o \geq m$

จัดค่าสังเกต (X_o) อยู่ใน Π_2 ถ้าหาก $Y_o < m$

สมการจำแนกกลุ่มที่ได้มานั้น จะมีความสามารถในการแยกกลุ่มได้ดีหรือไม่ เราสามารถดูได้จากค่า Missclassification error ซึ่งเป็นค่าที่ได้จากการเปรียบเทียบสิ่งที่เราศึกษานั้นเป็นสมาชิกของประชากรกลุ่มหนึ่ง แต่เมื่อใช้สมการจำแนกกลุ่มมาจัดการแยกกลุ่มแล้วกลับไปเป็นสมาชิกของอีกกลุ่มหนึ่ง ซึ่งทำให้ได้ผลที่ผิดจากความเป็นจริงไป

2.2.2 การคัดเลือกตัวแปรเข้าไปในสมการโดยวิธีแบบขั้นตอนของ Mahalanobis

เป็นวิธีการที่ทำให้ระยะความแตกต่างระหว่างกลุ่มที่ใกล้กันที่สุด มีค่ามากที่สุด โดยจะทำการคัดเลือกตัวแปรอิสระทีละตัวตามอำนาจการจำแนกของตัวแปรนั้น ตัวแปรตัวแรกที่ถูกเลือกเข้ามานั้นจะมีอำนาจในการจำแนกสูงสุด เมื่อเทียบกับตัวแปรอื่น ๆ จากนั้นก็จะเลือกตัวแปรที่ดีสุดตัวที่สองเข้าสมการเพื่อปรับปรุงทำให้สมการจำแนกกลุ่มดีขึ้น และเลือกตัวแปรตัวที่สามและตัวต่อ ๆ ไปที่จะช่วยให้สมการจำแนกกลุ่มดีขึ้นตามลำดับ จนกระทั่งถึงเกณฑ์ ๆ หนึ่ง ซึ่งเราได้กำหนดไว้ว่าตัวแปรที่เหลือนั้นไม่สามารถเข้าไปในสมการจำแนกกลุ่มได้อีกแล้ว

ขั้นตอนในการทดสอบกระทำดังนี้

1. คำนวณหาค่า Generalized Distance D_P^2

โดยที่ $D_P^2 = (\bar{X}_1-\bar{X}_2)' S_*^{-1} (\bar{X}_1-\bar{X}_2)$ เป็นค่าที่เกิดจากตัวแปร P ตัว

2. คำนวณหาค่า Generalized Distance D_{P+q}^2

$D_{P+q}^2 = (\bar{X}_1-\bar{X}_2) S_*^{-1} (\bar{X}_1-\bar{X}_2)$ เป็นค่าที่เกิดจากตัวแปร P+q ตัว

3. คำนวณหาค่า $D_{P+q}^2 - D_P^2$

4. สมมติฐานของการทดสอบ

H_o : จำนวนตัวแปร q ไม่ทำให้อำนาจในการจำแนกดีขึ้น

H_A : จำนวนตัวแปร q ทำให้อำนาจในการจำแนกดีขึ้น

5. ขอบเขตของการปฏิเสธสมมติฐาน

$$F = \frac{n_1+n_2-P-q-1}{q} \cdot \frac{m(D_{P+q}^2-D_P^2)}{1-mD_P^2} > F_{\alpha;\, q, n_1+n_2-P-q-1}$$

โดยที่ n_1 = จำนวนตัวอย่างจากประชากรกลุ่ม 1

n_2 = จำนวนตัวอย่างจากประชากรกลุ่ม 2

$$m = \frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-2)}$$

q = จำนวนตัวแปรที่จะเข้าไปในสมการจำแนก ในที่นี้ใช้วิธีแบบขั้นตอน q จึงมีค่าเป็น 1 เสมอ

P = จำนวนตัวแปรที่มีอยู่เดิมแล้วในสมการจำแนก

α = ระดับนัยสำคัญที่กำหนดไว้

ศูนย์วิทยทรัพยากร
จุฬาลงกรณ์มหาวิทยาลัย

013368