บทที่ 3

วิธีการวิจัย

เทคนิคหรือวิธีการทางสถิติที่ใช้ในการวิเคราะห์ตัวแปรหลายตัวนั้น มีอยู่ด้วยกันหลายวิธี ในแต่ละวิธีต่างก็มีข้อจำกัด ความเหมาะสมและวัตถุประสงค์แตกต่างกันออก การเลือกใช้เทคนิคหรือวิธีการวิเคราะห์ปัญหาที่เหมาะสมจึงเป็นสิ่งจำเป็นยิ่ง อย่างไรก็ตาม บ่อยครั้งที่เราพบว่าในปัญหาเรื่องเดียวกัน มีเทคนิคหรือวิธีการทางสถิติที่ใช้ได้หลายวิธี ผู้ใช้จึงต้องตัดสินใจว่าจะเลือกใช้วิธีใด จึงเหมาะสมดีที่สุด ประหยัดค่าใช้จ่ายหรือใช้ทรัพยากรน้อยที่สุด แต่ในขณะเดียวกันก็ให้ผลชัดเจนและตรงตามวัตถุประสงค์ที่ได้กำหนดไว้

สำหรับในการศึกษาเพื่อหาตัวแบบที่เหมาะสมในการจัดลำดับความสำคัญของพื้นที่ที่มีปัญหายาเสพติดครั้งนี้ ผู้ศึกษาจะทำการเปรียบเทียบวิธีการทางสถิติ 4 วิธีด้วยกันคือ วิธีการวิเคราะห์องค์ประกอบหลัก วิธีการวิเคราะห์คาโนนิคอล วิธีการวิเคราะห์ถดถอยพหุ และวิธีการวิเคราะห์จำแนกประเภท ขั้นตอนของการวิเคราะห์ในแต่ละวิธีจะปรากฏรายละเอียดในส่วนแรกของบทนี้ สำหรับส่วนที่สองจะกล่าวถึงวิธีการคัดเลือกตัวแบบที่เหมาะสมในการจัดลำดับความสำคัญของพื้นที่ที่มีปัญหายาเสพติดและสถิติที่ใช้ในการคัดเลือก และในส่วนสุดท้ายของบทนี้จะเป็นวิธีการเปรียบเทียบผลการวิเคราะห์ประกอบหลักในกรณีที่ไม่ได้มีการตัดตัวแปรกับกรณีที่มีการตัดตัวแปรก่อนทำการวิเคราะห์ตลอดจนวิธีการตัดตัวแปรก่อนที่จะทำการวิเคราะห์องค์ประกอบหลักด้วยวิธีมัลติเพิลคอร์รีเลชั่น รายละเอียดและขั้นตอนของการวิเคราะห์ของวิธีการต่าง ๆ มีดังต่อไปนี้

3.1 วิธีการวิเคราะห์ของแต่ละวิธี

3.1.1 วิธีวิเคราะห์องค์ประกอบหลัก (Principal Component Method)

เป็นวิธีหนึ่งในการแสวงหาปัจจัยแรก ซึ่งเป็นองค์ประกอบหลักหรือเป็นตัวแปรตัวใหม่ที่แสดงความสัมพันธ์เชิงเส้นตรงกับตัวแปรต่าง ๆ ที่นำมาศึกษา การวิเคราะห์โดยเทคนิคนี้ เราต้องรู้ว่ามีปัจจัยใดที่ส่งผลต่อลักษณะที่เราจะศึกษา ผลจากการวิเคราะห์เราจะได้ตัวแปรประกอบใหม่ที่อยู่ในรูปค่าถ่วงน้ำหนัก ที่สามารถใช้จัดลำดับความมากน้อยหรือความสำคัญ

ที่สามารถใช้เปรียบเทียบกันได้ ตัวแบบหรือรูปแบบสมการเป็นดังนี้คือ

$$Y_i = \sum_{j=1}^{k} a_{ij} x_j \qquad i = 1,2 \ldots k$$
$$j = 1,2 \ldots k$$

โดยที่ Y_i = ค่าตัวแปรประกอบขององค์ประกอบ (Component) ที่ i

a_{ij} = ค่าน้ำหนัก (Loading) บนองค์ประกอบที่ i ของตัวแปรที่ j

x_j = ตัวแปรตัวที่ j

ขั้นตอนในการคำนวณเพื่อหาตัวแปรใหม่ซึ่งเป็นตัวแปรประกอบ มีดังนี้

1) จาก $\underset{\sim}{X} = [x_1 \ldots x_k]$ ซึ่งเป็นตัวแปร k ตัวที่นำมาศึกษา

2) หาเมตริกซ์ความแปรปรวนร่วม (Covariance Matrix; S_x)

$$S_x = E[(\underset{\sim}{X} - E(\underset{\sim}{X}))(\underset{\sim}{X} - E(\underset{\sim}{X}))']$$

3) หา Characteristic Roots (λ) จากสมการ $[S-\lambda I] = 0$

โดยที่ $\lambda = \lambda_1, \lambda_2 \ldots \lambda_k$

λ_i = Characteristic root ที่ i ของ S_x

และ $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$

I = Identity matrix

4) หา Characteristic Vectors ($\underset{\sim}{a}$) จากสมการ

$$(S - \lambda I)\underset{\sim}{a} = 0$$

$\underset{\sim}{a}_i$ = Characteristic vector ที่ i ที่สอดคล้อง

กับ Characteristic root ที่ i (λ_i)

$\underset{\sim}{a}_i'$ = $[a_1, a_2 \ldots a_k]$

5) หาค่า Characteristic vector ที่ normalized แล้ว โดยการหาร Characteristic vectors ด้วยความยาว (length) ของ Characteristic vectors จะได้ Principal factor.

6) หมุนแกนองค์ประกอบให้แกนตั้งฉากซึ่งกันและกัน (Orthogonal) จะได้องค์ประกอบที่เป็นอิสระต่อกัน ค่าที่ได้จากการหมุนแกนคือค่าตัวประกอบที่จะนำไปใช้วิเคราะห์เพื่อหาตัวแปรที่สำคัญ โดยเราจะพิจารณาจากตัวประกอบที่ 1 ซึ่งเป็นตัวประกอบที่มีค่าความแปรปรวน (Variance) สูงสุด และจากองค์ประกอบที่ 1 นี้ตัวแปรที่มีความสำคัญสูงสุดคือตัวแปรที่มีค่าน้ำหนักสูงสุดในองค์ประกอบที่ 1

ดังนั้น เราจะใช้องค์ประกอบที่ 1 หรือองค์ประกอบที่มีความแปรปรวนสูงสุด เพื่อหาตัวแปรใหม่ ซึ่งก็คือ

$$Y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1k}x_k$$

อย่างไรก็ตาม เราจะเห็นได้ว่าตัวประกอบที่ได้จะแสดงความสัมพันธ์เชิงเส้นกับตัวแปรเดิมที่นำมาศึกษาทั้งหมด ดังนั้น ในกรณีที่มีตัวแปรจำนวนมาก ย่อมจะทำให้เสียเวลาและค่าใช้จ่ายทั้งในการเก็บรวบรวมข้อมูลและการประมวลผล โดยเฉพาะในกรณีที่ตัวแปรบางตัวยากแก่การเก็บรวบรวมและไม่ช่วยให้ผลการวิเคราะห์มีความหมายมากขึ้นแต่อย่างใด ดังนั้น บ่อยครั้งที่เราพบว่าการตัดตัวแปรบางตัวออกก่อนที่จะทำการวิเคราะห์องค์ประกอบหลัก ไม่ทำให้ผลที่ได้รับเปลี่ยนแปลงไปแต่อย่างใด หรือถ้าจะเปลี่ยนแปลงก็เพียงเล็กน้อยเท่านั้น[1] จากแนวความคิดดังกล่าวในปี 1982 โจลลิฟเฟ่ (Jolliffe) ได้เสนอวิธีการตัดตัวแปร

---

[1]JOLLIFFE, I.T., "Discarding Variables in a Principal Component Analysis I : Artificial Data", Applied Statistics, 21 1972, pp 160-173.

ที่ซ้ำซ้อน (Rejection Method) หลายวิธีด้วยกัน และผลจากการทดลองทั้งกับข้อมูลที่สร้างขึ้น (Artificial Data) ตลอดจนข้อมูลจริง (Real Data) พบว่า วิธีการตัดตัวแปรที่เรียกว่าวิธี มัลติเพิ่ลคอร์รีเลชั่น (Multiple Correlation Method) เป็นวิธีที่ง่ายและให้ผลดีหรือให้ผลใกล้เคียงกับผลที่ได้จากการวิเคราะห์ในกรณีที่ไม่ได้มีการตัดตัวแปร ในการศึกษาครั้งนี้จึงได้นำหลักการของวิธีการดังกล่าวมาทำการศึกษาเปรียบเทียบด้วย

สำหรับในการศึกษาครั้งนี้ ตัวแปรที่ใช้ในการวิเคราะห์องค์ประกอบหลักมี 11 ตัวดังต่อไปนี้

CASE	=	จำนวนคดียาเสพติด
ADDICT	=	จำนวนผู้ติดยาเสพติดที่สมัครใจเข้ารับการรักษา
WEIGHT	=	น้ำหนักของกลาง
FACTORY	=	จำนวนโรงงาน
DENSITY	=	ความหนาแน่นของประชากร
RATING	=	อัตราการเพิ่มของประชากร
TEMPLE	=	จำนวนวัด
STU-TEA	=	จำนวนนักเรียนต่อครูหนึ่งคน
BAR	=	จำนวนสถานเริงรมย์
BANK	=	จำนวนธนาคาร
TRUCK	=	จำนวนรถบรรทุก

3.1.2 วิธีวิเคราะห์คาโนนิคอล (Canonical Analysis)

เป็นเทคนิคการวิเคราะห์ตัวแปรหลายตัวอีกวิธีหนึ่ง มีจุดมุ่งหมายที่สำคัญคือหาแบบแผนความสัมพันธ์ระหว่างตัวแปร 2 ชุด และพิจารณาจากชุดที่มีความสัมพันธ์กันมากที่สุด หลักการของวิธีวิเคราะห์คาโนนิคอลก็คือเราจะแบ่งตัวแปรที่มีอยู่ออกเป็น 2 กลุ่มแล้วหาความสัมพันธ์ระหว่างตัวแปรทั้ง 2 กลุ่มนี้ ซึ่งถ้าตัวแปรทั้ง 2 กลุ่มนี้มีความสัมพันธ์กันแล้ว เราก็จะใช้ตัวแปรในกลุ่มที่จัดได้ง่ายหรือเก็บรวบรวมได้ง่ายกว่าไปใช้ ผลจากการวิเคราะห์เราจะ

ได้ตัวแปรตัวใหม่ ซึ่งเป็นตัวแปรประกอบที่สร้างขึ้นจากตัวแปรเดิมทั้ง 2 กลุ่ม ตัวแปรใหม่นี้เรียกว่า ตัวแปรคาโนนิคอล (Canonical Variate) มีขั้นตอนในการคำนวนดังนี้

ถ้าให้ $x_1, x_2, \ldots, x_p$ เป็นตัวแปรในกลุ่มที่ 1

และ $y_1, y_2, \ldots, y_q$ เป็นตัวแปรในกลุ่มที่ 2

กำหนดให้ $q \leqslant p$

ให้ U และ V เป็นตัวแปรใหม่ ซึ่งเป็นผลรวมเชิงเส้นของตัวแปรกลุ่มที่ 1 และ 2

เราจะได้ว่า

$$
\begin{aligned}
U_1 &= \alpha_{11}x_1 + \alpha_{12}x_2 + \ldots + \alpha_{1p}x_p \\
U_2 &= \alpha_{21}x_1 + \alpha_{22}x_2 + \ldots + \alpha_{2p}x_p \\
&\vdots \\
U_q &= \alpha_{q1}x_1 + \alpha_{q2}x_2 + \ldots + \alpha_{qp}x_p \\
\\
V_1 &= \beta_{11}y_1 + \beta_{12}y_2 + \ldots + \beta_{1q}y_q \\
&\vdots \\
V_q &= \beta_{q1}y_1 + \beta_{q2}y_2 + \ldots + \beta_{qq}y_q
\end{aligned}
$$

ถ้าให้ ρ_c เป็นค่าความสัมพันธ์ระหว่าง U และ V ค่า ρ_c นี้จะเรียกว่าค่าสัมประสิทธิ์สหสัมพันธ์คาโนนิคอล (Canonical Correlation Coefficient)

ซึ่ง
$$\rho_c = \frac{\text{Cov}(U, V)}{\sqrt{\text{Var}(U)\,\text{Var}(V)}}$$

วิธีวิเคราะห์คาโนนิคอลนี้เราต้องการหาแบบแผนความสัมพันธ์ที่ทำให้ความสัมพันธ์ระหว่าง U และ V สูงสุด ดังนั้นถ้าให้

$\rho_c^{(1)}$ เป็นสหสัมพันธ์คาโนนิคอลที่มีค่าสูงสุด ซึ่งจะเรียกว่าค่าสัมประสิทธิ์สหสัมพันธ์คาโนนิคอลอันดับที่หนึ่ง ที่เกิดจาก U_1 และ V_1

$\rho_c^{(2)} \; . \; . \; . \; \rho_c^{(q)}$ เป็นค่าสัมประสิทธิ์สหสัมพันธ์คาโนนิคอลที่มีค่ามากรองลงมาตามลำดับ

U_1 และ V_1 จะเรียกว่า First Canonical Variable

U_q และ V_q จะเรียกว่า q^{th} Canonical Variable

ซึ่งถ้าเขียนในรูปเมตริกซ์จะได้ว่า $\underset{\sim}{U} = x\underset{\sim}{\alpha}$

$$\underset{\sim}{V} = y\underset{\sim}{\beta}$$

กำหนดให้ $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}_{(p+q) \times (p+q)}$

โดยที่ Σ_{11} = เมตริกซ์ความแปรปรวนของกลุ่มตัวแปร x มีขนาด pxp

Σ_{22} = เมตริกซ์ความแปรปรวนของกลุ่มตัวแปร y มีขนาด qxq

Σ_{12} = เมตริกซ์ความแปรปรวนร่วมระหว่างกลุ่มตัวแปร x และ y มีขนาด pxq

และ $\Sigma_{21} = \Sigma_{12}'$

หรือเขียนให้เห็นได้ง่ายดังนี้

$$
\Sigma \quad = \quad \left[\begin{array}{cccc|cccc}
\alpha_{x_1x_1} & \alpha_{x_1x_2} & \cdots & \alpha_{x_1x_p} & \alpha_{x_1y_1} & \alpha_{x_1y_1} & \cdots & \alpha_{x_1y_q} \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
\alpha_{x_px_1} & \alpha_{x_py_2} & \cdots & \alpha_{x_px_p} & \alpha_{x_py_1} & \alpha_{x_py_2} & \cdots & \alpha_{x_py_q} \\
\hline
\alpha_{x_1y_1} & \alpha_{x_2y_1} & \cdots & \alpha_{x_py_1} & \alpha_{y_1y_1} & \alpha_{y_1y_2} & \cdots & \alpha_{y_1y_q} \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
\alpha_{x_1y_q} & \alpha_{x_2y_q} & \cdots & \alpha_{x_py_q} & \alpha_{y_qy_1} & \alpha_{y_qy_2} & \cdots & \alpha_{y_qy_q}
\end{array}\right]
$$

ดังนั้น $$\rho_c \quad = \quad \frac{\mathrm{Cov}(x\underset{\sim}{\alpha}, y\underset{\sim}{\beta})}{\sqrt{\mathrm{Var}\ (x\underset{\sim}{\alpha})\ .\ \mathrm{Var}\ (y\underset{\sim}{\beta})}}$$

หรือ $$\rho_c \qquad \frac{\underset{\sim}{\alpha}'\ \Sigma_{12}\ \underset{\sim}{\beta}}{\sqrt{\underset{\sim}{\alpha}'\ \Sigma_{11}\ \underset{\sim}{\alpha}\ \ \underset{\sim}{\beta}'\ \Sigma_{22}\ \underset{\sim}{\beta}}}$$

การหาค่า ρ_c ที่จะให้ได้ ρ_c มีค่าสูงสุด จะทำโดยการ maximize ρ_c เทียบกับ $\underset{\sim}{\alpha}$ และ $\underset{\sim}{\beta}$ วิธีการนี้จะง่ายขึ้นถ้าให้ Canonical Variable แต่ละตัวมีความแปรปรวนเป็น 1

นั่นคือ จะได้ว่า $\mathrm{Var}\ (U) = \mathrm{Var}\ (x\underset{\sim}{\alpha}) = \underset{\sim}{\alpha}'\ \Sigma_{11}\underset{\sim}{\alpha} = 1$ -----(1)

$\mathrm{Var}\ (V) = \mathrm{Var}\ (y\underset{\sim}{\beta}) = \underset{\sim}{\beta}'\ \Sigma_{22}\underset{\sim}{\beta} = 1$ -----(2)

ซึ่งเงื่อนไขดังกล่าวจะได้ว่า $\rho_c = \underset{\sim}{\alpha}'\ \Sigma_{12}\underset{\sim}{\beta}$ และเมื่อ maximize ρ_c โดยวิธี Least square จะได้ว่า

$$\Sigma_{12}\underset{\sim}{\beta} - \lambda\,\Sigma_{11}\underset{\sim}{\alpha} \quad = \quad 0 \qquad \text{------------ (3)}$$

$$\Sigma_{21}\underset{\sim}{\alpha} - \lambda\,\Sigma_{22}\underset{\sim}{\beta} \quad = \quad 0 \qquad \text{------------ (4)}$$

เมื่อ λ เป็น Laqrange multiplier

จาก (3) คูณด้วย λ จะได้ $\quad \Sigma_{12}\,\lambda\,\underset{\sim}{\beta} = \lambda^2\Sigma_{11}\,\underset{\sim}{\alpha}$ --------- (5)

จาก (4) คูณด้วย Σ_{22}^{-1} จะได้ $\quad \Sigma_{22}^{-1}\,\Sigma_{21}\,\underset{\sim}{\alpha} = \lambda\,\underset{\sim}{\beta}$ --------- (6)

นำค่า $\lambda\underset{\sim}{\beta}$ ที่ได้จาก (2) แทนลงใน (5) จะได้

$$\Sigma_{12}\,\Sigma_{22}^{-1}\,\Sigma_{21}\,\underset{\sim}{\alpha} \qquad \lambda^2\Sigma_{11}\,\underset{\sim}{\alpha}$$

คูณทั้งสองข้างด้วย Σ_{11}^{-1} , $(\Sigma_{11}^{-1}\,\Sigma_{12}\,\Sigma_{22}^{-1}\,\Sigma_{21} - \lambda^2 I)\,\underset{\sim}{\alpha} = 0$ ---(7)

ในทำนองเดียวกันจะได้ว่า $(\Sigma_{22}^{-1}\,\Sigma_{21}\,\Sigma_{11}^{-1}\,\Sigma_{12} - \lambda^2 I)\,\underset{\sim}{\beta} = 0$ ---(8)

ทั้งสมการ (7) และ (8) เป็น homogeneous equations ซึ่งสามารถหาคำตอบได้

จาก (3) $\quad \underset{\sim}{\alpha} = \dfrac{\Sigma_{11}^{-1}\,\Sigma_{12}\,\underset{\sim}{\beta}}{\lambda}$ ---------(9)

จาก (3) คูณด้วย $\underset{\sim}{\alpha}'$ จะได้ $\quad \underset{\sim}{\alpha}'\Sigma_{12}\,\underset{\sim}{\beta} = \lambda\,\underset{\sim}{\alpha}'\Sigma_{11}\,\underset{\sim}{\alpha}$

จาก (4) คูณด้วย $\underset{\sim}{\beta}'$ จะได้ $\quad \underset{\sim}{\beta}'\Sigma_{21}\,\underset{\sim}{\alpha} = \lambda\,\underset{\sim}{\beta}'\Sigma_{22}\,\underset{\sim}{\beta}$

จากเงื่อนไขที่กำหนดว่า $\quad \underset{\sim}{\alpha}'\Sigma_{11}\,\underset{\sim}{\alpha} \quad \underset{\sim}{\beta}'\Sigma_{22}\,\underset{\sim}{\beta} = 1$

ดังนั้น $\quad \underset{\sim}{\alpha}'\Sigma_{12}\,\underset{\sim}{\beta} \quad \underset{\sim}{\beta}'\Sigma_{21}\,\underset{\sim}{\alpha} = \lambda$ -----(10)

นั่นคือ $\quad \rho_c = \lambda = \underset{\sim}{\alpha}'\Sigma_{12}\,\underset{\sim}{\beta}$

และจาก (7) หรือ (8) รากที่สองของ Charactiristic root coefficient (λ^2) ที่ให้ค่ามากที่สุดจะเป็น First canonical correlation

จาก (8) เราสามารถหาค่า λ^2 และ β ได้โดยที่

$\lambda_1^2 \geq \lambda_2^2 \geq \ldots \lambda_q^2$ เป็น characteristic root

และ $\beta_1 , \beta_2 \ldots \beta_q$ เป็น characteristic vectors ที่สอดคล้องหรือสัมพันธ์กับ characteristic root ลำดับเดียวกัน

จาก (9) จะหาค่า α ได้ โดยที่

$\alpha_1 , \alpha_2 \ldots \alpha_q$ เป็น characteristic vectors ที่คำนวณได้จากการแทนค่า

λ_i และ β_i ลงใน (9)

และ canonical correlation coefficient ที่ i คือ $\rho_c^{(i)}$ ซึ่งมีค่าระหว่าง 0 ถึง 1

canonical variables $U_i = x\,\alpha_i$

$$U_i = y\,\beta_i$$

ในการคำนวณ เนื่องจากเราไม่ทราบค่า Σ จะเราจะประมาณ Σ ด้วย S

โดยที่

$$S = \left[\begin{array}{c|c} S_{11} & S_{12} \\ \hline S_{21} & X_{22} \end{array}\right]$$

หรืออาจจะคำนวณได้จาก Correlation matrix (R) โดยที่

$$R = \left[\begin{array}{c:c} R_{11} & R_{12} \\ \hdashline R_{21} & R_{22} \end{array}\right]$$

S และ R เป็น matrix ของความแปรปรวนร่วมและสหสัมพันธ์ของตัวอย่าง

S_{11} และ R_{11} เป็น matrix ของความแปรปรวนและสหสัมพันธ์ระหว่างตัวแปร X มีขนาด pxp

S_{22} และ R_{22} เป็น matrix ของความแปรปรวนและสหสัมพันธ์ระหว่างตัวแปร Y มีขนาด qxq

S_{12} และ S_{12} เป็น matrix ของความแปรปรวนและสหสัมพันธ์ระหว่างตัวแปร X และ Y มีขนาด pxq

และ $S_{21} = S'_{12}$, $R_{21} = R'_{12}$

ค่าสัมประสิทธิ์สหสัมพันธ์คาโนนิคอล ที่คำนวณได้ไม่ว่าจะจาก S หรือ R นั้นจะได้ผลเช่นเดียวกัน จะต่างกันที่ค่าของ $\underset{\sim}{\alpha}$ และ $\underset{\sim}{\beta}$ เท่านั้น กล่าวคือค่าของ $\underset{\sim}{\alpha}$ และ $\underset{\sim}{\beta}$ ที่คำนวณได้จากการใช้ S จะเป็นค่าที่ได้จากการใช้คะแนนดิบ แต่ค่า $\underset{\sim}{\alpha}$ และ $\underset{\sim}{\beta}$ ที่คำนวณได้จากการใช้ R จะเป็นค่าที่อยู่ในรูปคะแนนมาตรฐาน อย่างไรก็ตามเราสามารถเปลี่ยนค่าสัมประสิทธิ์ที่คำนวณได้จากการใช้ S ไปให้อยู่ในรูปคะแนนมาตรฐานได้ โดยการหารค่าสัมประสิทธิ์ที่หาได้จากการใช้ S ด้วยค่าเบี่ยงเบนมาตรฐาน

<u>การทดสอบนัยสำคัญของการวิเคราะห์คาโนนิคอล</u>

ผลจากการวิเคราะห์คาโนนิคอลเราจะได้ตัวแปรประกอบใหม่ และค่าสัมประสิทธิ์สหสัมพันธ์คาโนนิคอล ซึ่งแสดงความสัมพันธ์ระหว่างตัวแปร 2 กลุ่ม ซึ่งถ้าเราทราบว่าตัวแปร 2 กลุ่ม มีความสัมพันธ์กันแล้ว ในการวิเคราะห์คราวต่อไปเราจะใช้ตัวแปรของกลุ่มวัดได้ง่ายกว่าไปใช้ ดังนั้นเราจึงจำเป็นต้องทราบว่าตัวแปรทั้ง 2 กลุ่มนี้มีความสัมพันธ์กันหรือไม่ โดยทำการทดสอบว่าความแปรปรวนร่วม (Σ_{12} หรือ Σ_{21}) เท่ากับศูนย์หรือไม่

$$\therefore \quad H_o \; : \; \Sigma_{12} = 0 \text{ หรือ } \Sigma_{21} = 0$$

$$H_A \; : \; \Sigma_{12} \neq 0 \text{ หรือ } \Sigma_{21} \neq 0$$

ในการทดสอบเราจะใช้วิธีของ Wilks (Likelihood Ratio Test)

$$U_{cal} = \frac{[S]}{[S_{11}][S_{22}]} = \frac{[R]}{[R_{11}][R_{22}]}$$

เราจะเปรียบเทียบ U_{cal} กับ $U_{table}(p, q, n-1-q)$

และเราจะ Reject เมื่อ $U_{cal} < U_{table}$

ในกรณีที่ทำการประมวลผลด้วยโปรแกรมสำเร็จรูป SAS เราจะพิจารณาว่าตัวแปรทั้ง 2 กลุ่ม มีความสัมพันธ์กันอย่างมีนัยสำคัญทางสถิติหรือไม่ด้วยค่า PROB F หรือ P-VALUE.

สำหรับในการศึกษาครั้งนี้ ตัวแปรที่ใช้ในการวิเคราะห์คาโนนิคอลประกอบด้วยตัวแปร 11 ตัว ซึ่งในการวิเคราะห์เราจะแบ่งตัวแปรออกเป็น 2 กลุ่มคือกลุ่มตัวแปรทางด้านเศรษฐกิจ ซึ่งประกอบด้วยตัวแปร 3 ตัวคือจำนวนธนาคาร จำนวนสถานเริงรมย์ และจำนวนรถบรรทุก ตัวแปรอีกกลุ่มหนึ่งคือกลุ่มตัวแปรทางด้านสังคม ซึ่งประกอบด้วยตัวแปร 8 ตัวคือ จำนวนผู้ติดยาเสพติดที่สมัครใจเข้ารับการรักษา จำนวนคดียาเสพติด อัตราการเพิ่มของประชากร ความหนาแน่นของประชากร จำนวนนักเรียนต่อครูหนึ่งคน จำนวนวัด จำนวนโรงงาน และน้ำหนักของกลาง

3.1.3 <u>วิธีวิเคราะห์ถดถอยพหุ</u> (Multiple Regression Analysis)

เป็นวิธีทางสถิติที่ใช้ในการวิเคราะห์ตัวแบบหรือรูปแบบของความสัมพันธ์เชิงเส้นระหว่างตัวแปรตามหนึ่งตัวกับตัวแปรอิสระหลายตัว จากตัวแบบที่ได้เราจะทราบว่าตัวแปรอิสระแต่ละตัวมีความสัมพันธ์กับตัวแปรตามมากน้อยเพียงใด เป็นไปในทิศทางเดียวกันหรือตรงกัน

ข้ามในการใช้วิธีนี้จะต้องสมมติว่าตัวแปรอิสระแต่ละตัวเป็นอิสระซึ่งกันและกัน และปฏิกิริยาระหว่างตัวแปรอิสระด้วยกันไม่มีผลต่อตัวแปรตาม ซึ่งรูปแบบความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระเป็นดังนี้

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

โดยที่ β_0 = ค่าคงที่หรือจุดเริ่มต้น

$\beta_1 \ldots \beta_k$ = ค่าสัมประสิทธิ์ถดถอย

Y = ค่าตัวแปรตามที่ต้องการจะศึกษา

$X_1 \ldots X_k$ = ค่าตัวแปรอิสระ k ตัวที่นำมาศึกษา

ε = ค่าความคลาดเคลื่อน

ในการประมาณค่า β นั้น จะประมาณด้วยวิธีกำลังสองน้อยที่สุด (Least Square Method) ซึ่งโดยวิธีนี้ จะทำให้ผลรวมกำลังสองของส่วนเบี่ยงเบนของค่าสังเกตต่างจากค่าเฉลี่ยมีค่าน้อยที่สุด และทำให้ได้ตัวประมาณที่ไม่เอนเอียง ขั้นตอนในการคำนวณพอสรุปได้ดังต่อไปนี้

จากข้อมูลที่นำมาศึกษาในครั้งนี้มีจำนวนข้อมูล 64 รายการคือ n = 64 เขียนเป็นสมการได้ดังนี้

$$\begin{aligned}
Y_1 &= \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \ldots + \beta_k X_{k1} + \varepsilon_1 \\
Y_2 &= \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \ldots + \beta_k X_{k2} + \varepsilon_2 \\
&\vdots \\
Y_n &= \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \ldots \quad \beta_k X_{kn} + \varepsilon_n
\end{aligned}$$

จากสมการข้างต้นสามารถเขียนอยู่ในรูปของเมตริกซ์ ได้ดังนี้

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$$

หรือ

$$\underset{n x 1}{\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}} = \underset{(n,k+1)}{\begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix}} \underset{(k+1,1)}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}} + \underset{(n x 1)}{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}}$$

ถ้า $\hat{Y}$ เป็นตัวประมาณของ Y

$\hat{\beta}$ เป็นตัวประมาณของ β

e เป็นตัวประมาณของ ε

เราจะได้สมการ $\hat{\underset{\sim}{Y}} = X\hat{\underset{\sim}{\beta}} + \underset{\sim}{e}$

$$\underset{\sim}{e} = \underset{\sim}{Y} - X\hat{\underset{\sim}{\beta}}$$

$$\sum_{i=1}^{n} e_i^2 = e'e$$

$$= (\underset{\sim}{Y} - X\hat{\underset{\sim}{\beta}})'(\underset{\sim}{Y} - X\hat{\underset{\sim}{\beta}})$$

การหาค่า $\hat{\beta}$ จะทำโดยการ minimize ผลบวกกำลังสองของความคลาดเคลื่อน (e) ด้วยการหาอนุพันธ์ (differentiate) $\sum_{i=1}^{n} e_i^2$ เทียบกับค่า $\hat{\beta}$ แล้วให้เท่ากับศูนย์ จะได้ว่า

$$\frac{\partial}{\partial \beta}(\underset{\sim}{e}'\underset{\sim}{e}) = -2X'\underset{\sim}{Y} + 2X'X\hat{\underset{\sim}{\beta}} = 0$$

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

$$\therefore \quad \hat{\underset{\sim}{\beta}} = (X'X)^{-1}(X'\underset{\sim}{Y})$$

และเมื่อทำการ differentiate อันดับที่ 2 จะได้

$$\frac{\partial}{\partial \hat{\underset{\sim}{\beta}}}(\underset{\sim}{e}'\underset{\sim}{e}) = 2X'X$$

ซึ่งเป็น differentiate ทางบวก เพราะฉนั้น $\hat{\underset{\sim}{\beta}} = (X'X)^{-1}(X'\underset{\sim}{Y})$ จึงเป็นตัวประมาณที่ได้จากการทำให้กำลังสองของความคลาดเคลื่อนมีค่าน้อยที่สุด

และจากการคำนวณจะได้ค่าความแปรปรวนของ $\underset{\sim}{Y}$ (S^2) ดังนี้

$$S^2 = \underset{\sim}{e}'\underset{\sim}{e}/(n-k-1)$$

โดยที่ S^2 เป็นตัวประมาณที่ไม่เอนเอียงของ δ^2

ข้อสมมติเบื้องต้นในการวิเคราะห์ถดถอยพหุ

ข้อสมมติเบื้องต้นสำหรับการประมาณโดยใช้วิธีกำลังสองน้อยที่สุด มีดังต่อไปนี้

1) จาก $\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$

Y_i เป็นฟังก์ชันเชิงเส้นของ X_{ij} กับ ε_i

โดยที่ $i = 1,2 \ldots n$

$j = 0,1 \ldots k$, และ $X_{0i} = 1$

2) $E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$

นั่นคือ $E(\underset{\sim}{\varepsilon}) = 0$ สำหรับทุกค่าของ i

3) $E(\underset{\sim}{\varepsilon}\,\underset{\sim}{\varepsilon}') = \sigma^2 I$

นั่นคือ $E(\varepsilon_i^2) = \sigma^2$ สำหรับทุกค่าของ i นั่นคือ ε_i มีความแปรปรวนที่ $E(\varepsilon_i \varepsilon_j) = 0$ เมื่อ $i \neq j$ ซึ่งแสดงว่าแต่ละคู่ของ ε ไม่เกี่ยวข้องกัน

4) ε_i เป็นตัวแปรเชิงสุ่มที่มีการกระจายแบบปกติ $N(0, \sigma^2)$ ซึ่งเงื่อนไขข้อนี้จะใช้ในกรณีที่ต้องการสร้างช่วงความเชื่อมั่นของ β_0 , $\beta_1 \ldots \beta_k$ หรือเมื่อต้องการทดสอบสมมติฐานใด ๆ เกี่ยวกับตัวประมาณ

5) X เป็นเมตริกซ์อันดับที่ $n \times (k+1)$ ซึ่งมีค่าคงที่ จึงทำให้ X และ $\underset{\sim}{\varepsilon}$ เป็นตัวแปรอิสระกัน เราจึงได้ว่า

$$E(\underset{\sim}{\varepsilon}/X) = E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$$

$$E(X'\underset{\sim}{\varepsilon}) = X'E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$$

$$E(\underset{\sim}{\varepsilon}\,\underset{\sim}{\varepsilon}'/X) = E(\underset{\sim}{\varepsilon}\,\underset{\sim}{\varepsilon}') = \sigma^2 I$$

6) X มี rank $k + 1$ ซึ่ง $(k+1) < n$ หมายความว่าจำนวนค่าสังเกต n จะต้องมากกว่าจำนวนพารามิเตอร์ $(k+1)$ ที่จะต้องประมาณค่า

<u>เกณฑ์การคัดเลือกตัวแปรที่ใช้เป็นตัวบ่งชี้ถึงพื้นที่ที่มีปัญหายาเสพติด</u>

เนื่องจากในการวิเคราะห์ถดถอยพหุนั้น ผู้วิเคราะห์จำเป็นจะต้องรู้ว่าตัวแปรใดเป็นตัวแปรตามหรือเป็นตัวแปรที่ต้องการจะศึกษา ซึ่งในที่นี้ก็คือเป็นตัวแปรที่ใช้เป็นตัวบ่งชี้ถึงพื้นที่ที่มีปัญหายาเสพติด แต่ในทางปฏิบัติจากการศึกษาเท่าที่ผ่านมายังไม่สามารถระบุได้ว่าตัวแปรใดที่จะใช้เป็นตัวบ่งชี้ถึงพื้นที่ที่มีปัญหายาเสพติดได้ ในการศึกษาครั้งนี้ตัวแปรที่จะใช้เป็นตัวแปรตามหรือตัวแปรที่บ่งชี้ถึงพื้นที่ที่มีปัญหายาเสพติด คือตัวแปรที่มีน้ำหนักความสำคัญสูงสุดในสมการที่คัดเลือกได้จากการเปรียบเทียบระหว่างผลการวิเคราะห์องค์ประกอบหลักกับการวิเคราะห์

คาโนนิคอล ทั้งนี้ถือว่าตัวแปรนั้นสามารถอธิบายตัวแปรประกอบใหม่ ซึ่งจะใช้เป็นตัวบ่งชี้ถึงพื้นที่ที่มีปัญหายาเสพติดได้มากกว่าตัวแปรอื่น ๆ สำหรับการคัดเลือกสมการ ซึ่งเป็นผลจากการเปรียบเทียบระหว่างวิธีการวิเคราะห์องค์ประกอบหลักกับวิธีการวิเคราะห์คาโนนิคอล จะพิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์แบบอันดับ (Spearman's Coefficient of rank Correlation ; r_s) ระหว่างค่าคะแนนความสำคัญจากวิธีการทั้งสอง ถ้าค่า r_s ที่ได้มีค่าสูงเราจะเลือกใช้สมการที่ได้จากวิธีการวิเคราะห์ใดก็ได้ ทั้งนี้เพราะวิธีการทั้งสองให้ผลของลำดับความสำคัญของพื้นที่มีปัญหายาเสพติดสอดคล้องกัน แต่ถ้าค่า r_s ที่ได้มีค่าต่ำ (ต่ำกว่า 0.5) แสดงว่าวิธีการวิเคราะห์ทั้งสองให้ผลไม่สอดคล้องกัน ดังนั้นการจะเลือกใช้วิธีการวิเคราะห์ใดจะพิจารณาจากค่า C.V. (Coefficient of Variation) เป็นหลัก

3.1.4 <u>วิธีการวิเคราะห์จำแนกประเภท</u> (Discriminant Analysis)

เป็นวิธีการทางสถิติที่ใช้ในการคัดเลือกตัวแปรกลุ่มหนึ่งออกจากตัวแปรทั้งหมดที่นำมาศึกษา ซึ่งตัวแปรที่ถูกคัดออกมานี้จะเป็นตัวแปรที่มีความสัมพันธ์กับสิ่งที่ต้องการศึกษา ผลที่ได้จากการวิเคราะห์จะได้สมการจำแนกประเภท (Discriminant Function) ซึ่งแสดงความสัมพันธ์ระหว่างตัวแปรอิสระกลุ่มหนึ่งที่ถูกคัดเลือกว่ามีผลต่อสิ่งที่ต้องการศึกษา และค่าสัมประสิทธิ์ของตัวแปรที่อยู่ในสมการ จะบ่งชี้ถึงน้ำหนักหรือความสัมพันธ์ของตัวแปรอิสระที่มีต่อสิ่งที่ต้องการศึกษา ซึ่งเปรียบได้กับค่าสัมประสิทธิ์ถดถอยในการวิเคราะห์ถดถอยพหุ หรือเขียนเป็นตัวแบบได้ดังนี้

$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

โดยที่ Y = ค่าตัวแปรที่ใช้ในการจำแนกประเภทหรือที่ต้องการศึกษา

$\beta_1 \ldots \beta_p$ = ค่าสัมประสิทธิ์ของสมการจำแนกประเภทระหว่างตัวแปรอิสระที่ถูกเลือกกับตัวแปรที่ต้องการศึกษา (Discriminant Function Coefficients)

$X_1 \ldots X_p$ = ตัวแปรอิสระต่าง ๆ ที่ถูกคัดเลือกว่ามีความสัมพันธ์กับตัวแปรที่ต้องการศึกษา

อย่างไรก็ตาม การสร้างตัวแบบหรือสมการวิเคราะห์จำแนกประเภทมี 2 วิธีคือ วิธีตรงและวิธีแบบขั้นตอน

วิธีตรง (Direct Method) วิธีนี้จะใช้ตัวแปรทุกตัวที่ผู้ศึกษาคิดว่ามีความสัมพันธ์กับสิ่งที่ต้องการศึกษา สามารถจำแนกประชากรออกเป็นกลุ่มต่าง ๆ ได้ มาใช้ในการสร้างสมการจำแนกประเภท

วิธีแบบขั้นตอน (Stepwise Method) วิธีนี้จะทำการคัดเลือกตัวแปรชุดหนึ่งที่เหมาะสมเพื่อใช้ในการสร้างสมการจำแนกประเภท การคัดเลือกตัวแปรเข้าสู่สมการจะทำการคัดเลือกตัวแปรเข้าสู่สมการทีละหนึ่งตัว โดยตัวแปรตัวแรกที่ถูกเลือกจะเป็นตัวแปรที่มีอำนาจในการจำแนกกลุ่มมากที่สุด ตัวแปรตัวที่สองที่ถูกเลือกเป็นตัวแปรที่เมื่อรวมกับตัวแปรตัวแรกแล้วมีอำนาจในการจำแนกกลุ่มได้ดีขึ้น จากนั้นก็จะเลือกตัวแปรตัวต่อ ๆ ไป เข้าสู่สมการอีก โดยตัวแปรที่จะถูกเลือกเข้าสู่สมการ จะต้องเป็นตัวแปรที่ทำให้อำนาจในการจำแนกกลุ่มของสมการดีขึ้นตามลำดับและอย่างมีนัยสำคัญและก็เช่นเดียวกันการวิเคราะห์ถดถอยพหุแบบขั้นตอน กล่าวคือ ในแต่ละขั้นตอนตัวแปรที่ได้รับการคัดเลือกมาก่อนแล้วนั้น อาจจะถูกตัดทิ้งไปหากพบว่าเมื่อนำมารวมกับตัวแปรอื่น ๆ แล้วไม่ช่วยให้สมการจำแนกประเภทได้ดีขึ้น

ขั้นตอนในการวิเคราะห์จำแนกประเภท ในขั้นแรกจะทำการคัดเลือกตัวแปรชุดหนึ่งเพื่อสร้างสมการจำแนกประเภท (Discriminant Equation) ที่จะใช้ในการจำแนกประชากรออกเป็นกลุ่มต่าง ๆ ได้ จากนั้นก็จะใช้สมการจำแนกประเภทที่ได้จำแนกตัวอย่างเข้าเป็นสมาชิกของประชากรในแต่ละกลุ่ม

ในการศึกษาครั้งนี้จะใช้วิธีการของ Fisher มาใช้ในการจำแนกกลุ่ม ถ้าให้ π_1 เป็นประชากรกลุ่มที่ 1

π_2 เป็นประชากรกลุ่มที่ 2

จากตัวแปรที่ใช้ในการศึกษา p ตัว $x = [x_1 \ \ x_2 \ \dots \ x_p]$ ทำการแปลงตัวแปร X ไปเป็นค่าของตัวแปร Y ตัวเดียว โดยที่ Y_1 และ Y_2 เป็นค่าสังเกตที่ได้จากประชากร π_1 และ π_2 ตามลำดับ

ถ้า $\underset{\sim}{\mu}_{1y}$ = ค่าเฉลี่ยของค่า y ซึ่งได้มาจากค่า x ของประชากร π_1

$\underset{\sim}{\mu}_{2y}$ = ค่าเฉลี่ยของค่า y ซึ่งได้มาจากค่า x ของประชากร π_2

$\underset{\sim}{\mu}_1 = E(x/\pi_1)$ ค่าคาดหวังของตัวแปร x ที่ได้มาจากประชากร π_1

$\underset{\sim}{\mu}_2 = E(x/\pi_2)$ ค่าคาดหวังของตัวแปร x ที่ได้มาจากประชากร π_2

Σ = Covariance Matrix

$\Sigma = E(\underset{\sim}{X} - \underset{\sim}{\mu}_i)(\underset{\sim}{X} - \underset{\sim}{\mu}_i)' \; ; \; i = 1,2$

พิจารณาผลรวมเชิงเส้น

$$\underset{(1\times\ell)}{\underset{\sim}{Y}} = \underset{(1\times p)(p\times\ell)}{\underset{\sim}{\beta}'\underset{\sim}{X}}$$

$\underset{\sim}{Y}$ คือ Fisher's Linear Discriminant Function

$\underset{\sim}{\beta}$ คือ ค่าน้ำหนักความสำคัญของตัวแปร $x_1, x_2, \ldots, x_p$

$\underset{\sim}{\beta}' = [\beta_1, \beta_2 \cdots \beta_p]$

$\underset{\sim}{X}$ คือ ตัวแปรที่นำมาศึกษา ซึ่งมีทั้งหมด p ตัว

$\underset{\sim}{X}' = [x_1, x_2, \ldots, x_p]$

ดังนั้นจะได้ว่า

$$\underset{\sim}{\mu}_{1y} = E(\underset{\sim}{Y}/\pi_1) = E(\underset{\sim}{\beta}'\underset{\sim}{X}/\pi_1) = \underset{\sim}{\beta}'\underset{\sim}{\mu}_1$$

$$\underset{\sim}{\mu}_{2y} = E(\underset{\sim}{Y}/\pi_2) = E(\underset{\sim}{\beta}'\underset{\sim}{X}/\pi_2) = \underset{\sim}{\beta}'\underset{\sim}{\mu}_2$$

$$\sigma_y^2 = Var(\underset{\sim}{\beta}'\underset{\sim}{X}) = \underset{\sim}{\beta}' Cov(\underset{\sim}{X})\underset{\sim}{\beta} = \underset{\sim}{\beta}'\Sigma\underset{\sim}{\beta}$$

ตามหลักของวิธีการ Fisher ก็คือการพยายามหาผลรวมเชิงเส้นของค่า x ซึ่งจะทำให้ระยะทางระหว่าง μ_{1y} และ μ_{2y} มีค่ามากที่สุดที่จะเป็นไปได้

ค่าผลรวมเชิงเส้นซึ่งดีที่สุดจะต้องมีคุณสมบัติในการแบ่งแยกประชากรทั้ง 2 กลุ่มออกจากกันได้มากที่สุด ซึ่งวิธีการที่จะได้ผลรวมเชิงเส้นที่ดีนั้นก็โดยการหาค่า β ที่ทำให้อัตราส่วนระหว่าง (ระยะทางระหว่างค่าเฉลี่ยของ y $)^2$ กับ (ความแปรปรวนของ y) มีค่ามากที่สุด

$$\text{ถ้าให้ } \ell = \frac{(\text{ระยะทางระหว่างค่าเฉลี่ยของ } y)^2}{(\text{ความแปรปรวนของ } y)}$$

$$= \frac{(\underset{\sim}{\mu}_{1y} - \underset{\sim}{\mu}_{2y})^2}{\sigma_y^2}$$

$$= \frac{(\underset{\sim}{\beta}'\underset{\sim}{\mu}_1 - \underset{\sim}{\beta}'\underset{\sim}{\mu}_2)^2}{\underset{\sim}{\beta}'\Sigma\underset{\sim}{\beta}}$$

$$= \frac{\underset{\sim}{\beta}'(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\underset{\sim}{\beta}}{\underset{\sim}{\beta}'\Sigma\underset{\sim}{\beta}}$$

ค่า $\underset{\sim}{\beta}$ ที่ได้จะมีค่ามากที่สุดเมื่อ

$$\frac{\partial \ell}{\partial \underset{\sim}{\beta}} = 0$$

นั่นคือ $C(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2) - \Sigma\underset{\sim}{\beta} = 0$

$$\underset{\sim}{\beta} = C\Sigma^{-1}(\mu_1 - \mu_2)$$

กำหนดให้ $C = 1$ จะได้ผลรวมเชิงเส้นเป็น

$$\underset{\sim}{Y} = \underset{\sim}{\beta}'\underset{\sim}{X} = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{X} \quad \text{------} \quad (1)$$

สมการ (1) สามารถใช้เป็นตัวจำแนกค่าสังเกตที่ได้มานั้นว่าจะอยู่ใน ประชากรกลุ่ม π_1 หรือ π_2 ได้โดยให้

$$Y_0 = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} \underset{\sim}{X}_0$$

เป็นค่าของ Discriminant function ของค่าสังเกต x_o และให้

$$m = \frac{1}{2} (\underset{\sim}{\mu}_{1y} + \underset{\sim}{\mu}_{2y})$$

$$= \frac{1}{2} (\underset{\sim}{\beta}\underset{\sim}{\mu}_1 + \underset{\sim}{\beta}\underset{\sim}{\mu}_2)$$

$$= \frac{1}{2} (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} (\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2)$$

m จะเป็นจุดกึ่งกลาง (Centriod) ระหว่างค่าเฉลี่ยของค่า y จากทั้ง 2 ประชากรสามารถเขียนกฎการจำแนกประเภทได้ดังนี้คือ

จัด x_0 ให้อยู่ π_1 ถ้า $y_0 = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} \underset{\sim}{x}_0 < m$

จัด x_0 ให้อยู่ π_2 ถ้า $y_0 = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} \underset{\sim}{X}_0 > m$

ในทางปฏิบัติเราจะไม่ทราบค่า $\underset{\sim}{\mu}_1$, $\underset{\sim}{\mu}_2$ และ Σ ดังนั้นถ้าสุ่มตัวอย่าง จาก π_1 และ π_2 มาเป็นจำนวน n_1 และ n_2 แล้ว

$$\underset{\sim}{X}' = [x_1 \; x_2 \; \ldots \; x_p]$$

จะได้ว่า

$$\underset{(p \times n_1)}{\underset{\sim}{X}_1} = [x_{11} \; x_{12} \cdots x_{1n_1}]$$

$$\underset{(p \times n_2)}{\underset{\sim}{X}_2} = [x_{21} \; x_{22} \cdots x_{2n_2}]$$

$\bar{\underset{\sim}{X}}_1$, $\bar{\underset{\sim}{X}}_2$ และ S^{-1} pooled จะเป็นค่าประมาณของ $\underset{\sim}{\mu}_1$, $\underset{\sim}{\mu}_2$ และ Σ^{-1} ตามลำดับ

โดยที่ $\underset{\sim}{\bar{X}}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} \underset{\sim}{X}_{1j}$ ------------ (10)

$$\underset{\sim}{\bar{X}}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} \underset{\sim}{X}_{2j} \quad \text{------------ (11)}$$

$$S = \left[\frac{n_1 - 1}{(n_1 - 1)+(n_2 - 1)}\right] S_1 + \left[\frac{n_2 - 1}{(n_1 - 1)+(n_2 - 1)}\right] S_2$$

$$= \frac{(n_1 - 1)\, S_1 + (n_2 - 1)\, S_2}{n_1 + n_2 - 2}$$

$$\underset{(pxp)}{S_1} = \frac{1}{n_1-1} \sum_{j=1}^{n_1} (\underset{\sim}{X}_{1j} - \underset{\sim}{\bar{X}}_1)(\underset{\sim}{X}_{1j} - \underset{\sim}{\bar{X}}_1)'$$

$$\underset{(pxp)}{S_2} = \frac{1}{n_2-1} \sum_{j=1}^{n_2} (\underset{\sim}{X}_{2j} - \underset{\sim}{\bar{X}}_2)(\underset{\sim}{X}_{2j} - \underset{\sim}{\bar{X}}_2)'$$

เมื่อแทนค่า $\underset{\sim}{\bar{X}}_1$, $\underset{\sim}{\bar{X}}_2$ และ S^{-1} pooled ลงใน (7) จะได้ Fisher's Sample Linear Discriminant Function เป็น

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

$$= (\underset{\sim}{\bar{X}}_1 - \underset{\sim}{\bar{X}}_2)' \, S^{-1} \text{ pooled } \underset{\sim}{X}$$

และ $\hat{m} = \frac{1}{2}(\bar{Y}_1 + \bar{Y}_2)$

$$= \frac{1}{2}(\hat{\beta}'\bar{X}_1 + \hat{\beta}'\bar{X}_2)$$

$$= \frac{1}{2}(\bar{X}_1 + \bar{X}_2)' S^{-1} \text{pooled} (\bar{X}_1 + \bar{X}_2)$$

กฎการจำแนกประเภทจะเขียนได้เป็นดังนี้

จัด x_0 ให้อยู่ใน π_1 ถ้า $y_0 < m$

จัด x_0 ให้อยู่ใน π_2 ถ้า $y_0 > m$

เมื่อ $y_0 = (\bar{X}_1 - \bar{X}_2)' S^{-1} \text{pooled } x_0$

จากสมการสามารถเขียนได้เป็น $y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$

ค่า β_1 , $\beta_2 \ldots \beta_p$ จะเป็นค่าแสดงถึงน้ำหนักความสำคัญของตัวแปร x_1 , $x_2 \ldots x_p$ ซึ่งโดยปกติ x_i แต่ละค่านั้นจะมีหน่วยไม่เหมือนกัน การเปรียบเทียบค่า β_i เหล่านี้จะทำได้โดยการปรับค่า β_i เหล่านั้นให้เป็นค่ามาตรฐาน (Standardized) เสียก่อน การแปลงค่า β_i เหล่านี้สามารถทำได้โดยเอาสมาชิกในแนวเส้นทะแยงมุม (diagonal) ของเมตริกซ์ W มาถอดรากที่สอง แล้วนำไปคูณกับค่า β_i ทุก ๆ ค่าตามสูตรต่อไปนี้

$$\beta_i^* = (\sqrt{W_{i1}})(\beta_i)$$

เมื่อ β_i^* คือ Standardized Discriminant weight ของ Discriminant Function

$$W = (X - \bar{X}_{go})(X - \bar{X}_{go})'$$

เมื่อ W คือ Diagonal element ของเมตริกซ์ W โดยที่ $i = 1,2 \ldots p$

$$
\underset{\sim}{X} = \left[\begin{array}{cccc:cccc}
x_{111} & x_{121} & \cdots & x_{1n1} & x_{112} & x_{122} & \cdots & x_{1n1} \\
x_{211} & x_{221} & \cdots & x_{2n2} & x_{212} & x_{222} & \cdots & x_{2n2} \\
\vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
x_{p11} & x_{p21} & & x_{pn1} & x_{p12} & x_{p22} & & x_{pn2}
\end{array}\right]
$$

$$
\underset{\sim}{X}_{go} = \left[\begin{array}{ccc:ccc}
\bar{x}_{11} & \cdots & \bar{x}_{11} & \bar{x}_{12} & \cdots & \bar{x}_{12} \\
\bar{x}_{21} & \cdots & \bar{x}_{21} & \bar{x}_{22} & \cdots & \bar{x}_{22} \\
\vdots & & \vdots & \vdots & & \vdots \\
\bar{x}_{p1} & \cdots & \bar{x}_{p1} & \bar{x}_{p2} & \cdots & \bar{x}_{p2}
\end{array}\right]
$$

การพิจารณาว่าสมการจำแนกประเภทที่ได้มานั้นจะมีความสามารถในการแบ่งแยกกลุ่มได้ดีหรือไม่นั้นสามารถดูได้จากค่าความผิดพลาดเนื่องจากการจำแนกผิด (Mis Classification error)

การทดสอบนัยสำคัญของสมการจำแนกประเภท

เมื่อได้สมการจำแนกประเภทมาแล้ว วิธีการตรวจสอบว่าสมการที่ได้นี้มีอำนาจในการจำแนกกลุ่มได้อย่างมีนัยสำคัญทางสถิติหรือไม่ เราสามารถทดสอบได้จาก

$$V_m = \left[N - 1 -(p + k)/2)\right] \ln (1 + \lambda_m)$$

โดยที่ V_m คือ Bartlett V Statistics ซึ่งมีการกระจายแบบ Chi-Square ที่มี d.f เท่ากับ $(p + k - 2m)$

N = จำนวนตัวอย่างทั้งหมดจากประชากรทั้ง 2 กลุ่ม

p = จำนวนตัวแปรทั้งหมด

k = จำนวนกลุ่ม

λ_m = Characteristic root ตัวที่ m

ค่า λ_m นี้คำนวณได้จาก $|A - \lambda_m| = 0$

เมื่อ A = Covarience matrix ของ x

ค่า λ_m นี้จะเลือกเฉพาะค่าที่ไม่เป็นศูนย์เท่านั้น

สมการที่ได้จะเป็นสมการที่มีอำนาจในการจำแนกกลุ่มได้อย่างมีนัยสำคัญทางสถิติ ถ้าค่า V_m ที่คำนวณได้มีค่ามากกว่าค่า Chi-Square ที่ระดับนัยสำคัญที่กำหนดไว้

<u>เกณฑ์การคัดเลือกตัวแปรที่จะใช้ในการจำแนกประเภท</u>

ตัวแปรที่จะใช้ในการจำแนกประเภทคือตัวแปรซึ่งเป็นตัวบ่งชี้ถึงพื้นที่ที่มีปัญหายาเสพติด ซึ่งในทางปฏิบัติแล้วเราไม่สามารถทราบได้ ดังนั้นหลักเกณฑ์ในการคัดเลือกตัวแปรจำแนกประเภทก็ใช้เกณฑ์เช่นเดียวกันกับการคัดเลือกตัวแปรตามที่ใช้ในการวิเคราะห์ถดถอยพหุ กล่าวคือ ใช้ตัวแปรที่มีค่าน้ำหนักความสำคัญสูงสุดจากตัวแบบหรือสมการที่เลือกได้ระหว่างวิธีการวิเคราะห์องค์ประกอบหลักกับวิธีการวิเคราะห์คาโนนิคอล รายละเอียดดังได้กล่าวแล้วในหัวข้อเกณฑ์การคัดเลือกตัวแปรที่ใช้เป็นตัวบ่งชี้ถึงพื้นที่ที่มีปัญหายาเสพติดในหัวข้อที่ 3.1.3

3.2 <u>วิธีการคัดเลือกตัวแบบที่เหมาะสมในการเลือกหรือจัดลำดับความสำคัญของพื้นที่ที่มีปัญหายาเสพติด</u>

ในการคัดเลือกตัวแบบที่เหมาะสมกับการคัดเลือกหรือจัดลำดับความสำคัญของพื้นที่ที่มีปัญหายาเสพติด จะดำเนินตามขั้นตอนดังต่อไปนี้

1) วิเคราะห์ผลตามวิธีการวิเคราะห์องค์ประกอบหลัก โดยใช้ตัวแปรทั้งหมด 11 ตัว ที่นำมาศึกษาในครั้งนี้

2) วิเคราะห์ผลตามวิธีการวิเคราะห์คาโนนิคอล ถ้าค่าสัมประสิทธิ์สหสัมพันธ์คาโนนิคอลที่ได้มีค่าสูง จะใช้ตัวแบบที่ได้จากกลุ่มของตัวแปรที่วัดได้ง่ายกว่าไปเปรียบเทียบกับผลที่ได้จากวิธีการวิเคราะห์องค์ประกอบหลัก แต่ในกรณีที่ได้ค่าสัมประสิทธิ์สหสัมพันธ์คาโนนิคอลต่ำ จะใช้ตัวแบบที่ได้จากการนำตัวแปรทั้ง 2 กลุ่มมารวมกัน ซึ่งจะทำให้ผลที่ได้ถูกต้องมากยิ่งขึ้น

3) เปรียบเทียบผลการวิเคราะห์ระหว่างวิธีวิเคราะห์องค์ประกอบหลักกับวิธีวิเคราะห์คาโนนิคอล โดยพิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์แบบอันดับ (Spearman's Coefficient of rank Correlation ; r_s) ระหว่างค่าคะแนนของลำดับความสำคัญจากวิธีการทั้งสอง ถ้าค่า r_s ที่ได้มีค่าสูงจะเลือกใช้สมการใดก็ได้จาก 2 วิธีในการจัดลำดับความสำคัญของพื้นที่ที่มีปัญหายาเสพติดทั้งนี้จะขึ้นอยู่กับความสะดวกเป็นหลัก ในทางตรงข้ามถ้า r_s ที่ได้มีค่าต่ำ (ต่ำกว่า 0.5) แสดงว่าวิธีการทั้งสองให้ผลไม่สอดคล้องกัน ดังนั้นการจะเลือกใช้สมการที่ได้จากวิธีการวิเคราะห์ใดจะพิจารณาจากค่า C.V. ของค่าคะแนนความสำคัญที่ได้เป็นหลัก และจากสมการที่คัดเลือกได้จากวิธีการทั้งสอง ตัวแปรที่มีค่าน้ำหนักความสำคัญสูงที่สุดในสมการ จะเป็นตัวแปรตามและตัวแปรจำแนกประเภทที่ใช้ในการวิเคราะห์ถดถอยพหุและวิธีการวิเคราะห์จำแนกประเภทต่อไป

4) วิเคราะห์ผลตามวิธีการวิเคราะห์ถดถอยพหุและวิธีการวิเคราะห์จำแนกประเภท

5) เปรียบเทียบผลที่ได้จากการวิเคราะห์ เพื่อเลือกตัวแบบที่เหมาะสมกับการคัดเลือกหรือจัดลำดับความสำคัญของพื้นที่ที่มีปัญหายาเสพติด

6) เสนอตัวแบบที่เหมาะสม

<u>วิธีหาค่าสัมประสิทธิ์สหสัมพันธ์แบบอันดับของสเปียร์แมน</u> (Spearman's Coefficient of rank Correlation ; r_s)

ในการหาความสัมพันธ์ระหว่างตัวแปร 2 ตัวโดยทั่วไปเราจะใช้ค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) แต่ในกรณีที่ตัวแปรที่นำมาศึกษา เป็นตัวแปรประเภทจัดอันดับ ตัวสถิติที่ใช้ในการหาความสัมพันธ์จะมีหลายแบบด้วยกัน อาทิ สัมประสิทธิ์ความสัมพันธ์ครัสคาลและกูดแมน (G) สัมประสิทธิ์สหสัมพันธ์เชิงอันดับของเคนดาลล์ (τ) และสัมประสิทธิ์สหสัมพันธ์แบบอันดับของสเปียร์แมน (r_s) ในการศึกษาครั้งนี้เราจะใช้วิธีสุดท้ายที่กล่าวถึง ทั้งนี้เพราะมีขั้นตอนที่ใช้ในการคำนวณที่สะดวกและรวดเร็ว ตามสูตรที่ปรากฏข้างล่างนี้

$$r_s = \frac{1 - 6 \sum_{\lambda=1}^{n} d_i^2}{n(n^2 - 1)}$$

โดยที่ r_s = ค่าสัมประสิทธิ์สหสัมพันธ์แบบอันดับของสเปรียร์แมน

d_i = ค่าของผลต่างระหว่างอันดับระหว่างตัวแปร 2 กลุ่ม

n = จำนวนตัวอย่าง

3.3 <u>วิธีการเปรียบเทียบผลการวิเคราะห์องค์ประกอบหลักในกรณีที่ไม่ได้มีการตัดตัวแปรกับกรณีที่มีการตัดตัวแปร</u>

ในการเปรียบเทียบผลการวิเคราะห์องค์ประกอบหลักระหว่างกรณีที่ไม่ได้มีการตัดตัวแปรกับกรณีที่มีการตัดตัวแปร จะดำเนินการขั้นตอนต่อไปนี้

1) วิเคราะห์ผลตามวิธีการวิเคราะห์องค์ประกอบหลัก โดยใช้ตัวแปรทั้งหมด 11 ตัวที่นำมาศึกษาในครั้งนี้

2) ดำเนินการตัดตัวแปรด้วยวิธีมัลติเพิ่ลคอร์รีเลชั่น

3) จากตัวแปรที่คัดเลือกได้จากวิธีมัลติเพิ่ลคอร์รีเลชั่นแล้วนำตัวแปรเหล่านี้ไปวิเคราะห์ผลตามวิธีการวิเคราะห์องค์ประกอบหลักอีกครั้งหนึ่ง

4) เปรียบเทียบผลความเหมือนหรือความสัมพันธ์กันระหว่างเซทของององค์ประกอบที่ได้จากกรณีตัดและไม่ตัดตัวแปรโดยใช้สูตร

$$Q = \{ \sum_{i=1}^{m} q_i \, r(i) \} / \{ \sum_{i=1}^{m} q_i \}$$

ตามรายละเอียดที่ปรากฏในหัวข้อที่ 3.3.2 ซึ่งถ้าค่า Q ยิ่งมีค่าสูงแสดงว่า ผลที่ได้จากการวิเคราะห์องค์ประกอบหลักระหว่างกรณีที่ไม่ได้มีการตัดตัวแปรกับกรณีที่มีการตัดตัวแปรให้ผลสอดคล้องหรือสัมพันธ์กันยิ่งมาก กล่าวโดยสรุปก็คือความสอดคล้องหรือสัมพันธ์กันของผลการวิเคราะห์กรณีที่ไม่ได้มีการตัดตัวแปรกับกรณีที่มีการตัดตัวแปรแปรผันตามค่า Q

3.3.1 <u>การตัดตัวแปรด้วยวิธีมัลติเพิ่ลคอร์รีเลชั่น</u> (Multiple Correlation Method)

เป็นวิธีการตัดตัวแปรที่ซ้ำซ้อนก่อนที่จะดำเนินการวิเคราะห์องค์ประกอบหลัก โดยการตัดตัวแปรนั้นจะพิจารณาจากค่าสหสัมพันธ์พหุ (Multiple Correlation) เกณฑ์ในการตัดตัวแปร มีรายละเอียดดังต่อไปนี้

1) จากตัวแปร k ตัวที่นำมาศึกษา หาค่าสหสัมพันธ์พหุของตัวแปรแต่ละตัวกับตัวแปรที่เหลือ (k-1) ตัว จะได้ค่าสหสัมพันธ์พหุ k ค่า

2) จากค่าสหสัมพันธ์พหุ k ค่า พิจารณาตัดตัวแปรที่มีค่าสหสัมพันธ์พหุสูงสุดออกไป จะเหลือตัวแปร (k-1) ตัว

3) จากตัวแปรที่เหลือ (k-1) ตัว หาค่าสหสัมพันธ์พหุของตัวแปรแต่ละตัวกับตัวแปรที่เหลือ (k-2) ตัว จะได้ค่าสหสัมพันธ์พหุ (k-1) ค่า

4) จากค่าสหสัมพันธ์พหุ (k-1) ค่า พิจารณาตัดตัวแปรที่มีค่าสห-สัมพันธ์พหุสูงสุดออกไปอีก จะเหลือตัวแปร (k-2) ตัว

5) ดำเนินการหาค่าสหสัมพันธ์พหุของตัวแปรที่เหลือและตัดตัวแปรที่มีค่าสหสัมพันธ์พหุสูงสุดออกไป ทำซ้ำ ๆ ไปเรื่อย ๆ จนกระทั่งค่าสหสัมพันธ์พหุของตัวแปรทุกตัวที่เหลือมีค่าต่ำกว่า 0.15 จึงหยุดการตัดตัวแปร

แต่อย่างไรก็ตามโจลิฟเฟ่ (Jolliffe) ได้เสนอไว้ด้วยว่า ในการวิเคราะห์องค์ประกอบหลักนั้นควรจะเก็บตัวแปรไว้ในการวิเคราะห์ไม่น้อยกว่า 4 ตัว ดังนั้นในการศึกษาครั้งนี้ การตัดตัวแปรจะหยุดกระทำเมื่อกรณีใดกรณีหนึ่งดังกล่าวข้างต้นเกิดขึ้น

3.3.2 <u>หลักเกณฑ์ในการเปรียบเทียบผลระหว่างการวิเคราะห์องค์ประกอบหลักในกรณีที่ไม่ได้มีการตัดตัวแปรกับกรณีที่มีการตัดตัวแปร</u> (Comparison of Principal Components)

หลักเกณฑ์ในการเปรียบเทียบก็เช่นเดียวกับการหาความสัมพันธ์ของตัวแปรทั่วไป คือพิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์แบบอันดับ (Rank Correlation Coefficient)

จาก $$r_m = 1 - 6\{\sum_{i=1}^{n} (R_{mf_i} - R_{mr_i})^2\} / \{n(n^2 - 1)\}$$

โดยที่ r_m = ค่าสัมประสิทธิ์สหสัมพันธ์แบบอันดับขององค์ประกอบคู่ที่ m ระหว่างองค์ประกอบที่ได้จากกรณีที่ไม่ได้มีการตัดตัวแปรกับกรณีที่มีการตัดตัวแปร

R_{mf_i} = ค่าลำดับที่ของค่าสังเกตที่ i ในกรณีที่ไม่ได้มีการตัดตัวแปร

R_{mr_i} = ค่าลำดับที่ของค่าสังเกตที่ i ในกรณีที่มีการตัดตัวแปร

n = จำนวนค่าสังเกต

ค่า r_m นี้จะใช้วัดความเหมือนหรือความสัมพันธ์กันขององค์ประกอบคู่ที่ m ดังนั้นถ้าต้องการหาค่าเฉลี่ยของความเหมือนหรือความสัมพันธ์กันระหว่างเซทขององค์ประกอบที่ได้จากกรณีที่ไม่ได้มีการตัดตัวแปรกับกรณีที่มีการตัดตัวแปรจะต้องทำการถ่วงน้ำหนักด้วยความสามารถในการอธิบายความแปรปรวน ซึ่งเขียนเป็นสูตรได้ดังนี้

$$Q = \{\sum_{i=1}^{m} q_i \, r(i)\} / \{\sum_{i=1}^{m} q_i\}$$

โดยที่ q_i = ความสามารถในการอธิบายความแปรปรวนหรือสัดส่วนของความแปรปรวนที่ได้จากองค์ประกอบที่ i ในกรณีที่ไม่ได้มีการตัดตัวแปร

$r(i)$ = ค่าสูงสุดของค่าสัมประสิทธิ์สหสัมพันธ์แบบอันดับระหว่างองค์ประกอบที่ i ของกรณีที่ไม่ได้มีการตัดตัวแปรกับองค์ประกอบใด ๆ ของกรณีที่มีการตัดตัวแปร

m = จำนวนองค์ประกอบที่เราสนใจในกรณีการวิเคราะห์ที่ไม่ได้มีการตัดตัวแปร

และ Q = ค่าที่ใช้วัดความเหมือนหรือความสัมพันธ์กันระหว่างชุดขององค์ประกอบที่ได้จากการวิเคราะห์ในกรณีที่ไม่ได้มีการตัดตัวแปรกับกรณีที่มีการตัดตัวแปร ถ้าค่า Q ยิ่งมีค่ามากแสดงว่าผลจากการวิเคราะห์ในกรณีที่ไม่ได้มีการตัดตัวแปรกับกรณีที่มีการตัดตัวแปรก็ยิ่งใกล้เคียงกันมาก