บทที่ 3

ระเบียบวิธีวิจัย

ในการคัดเลือกพื้นที่เพื่อประกาศเขตปฏิรูปที่ดินนั้น ผู้วิจัยได้นำข้อมูลที่เก็บรวบรวมได้มาทำการวิเคราะห์ตามระเบียบวิธีทางสถิติโดยใช้โปรแกรมสำเร็จรูป SPSS (Statistical Package for the Social Science) จะสร้างแบบจำลองที่ใช้แสดงสภาพการพัฒนาในแต่ละอำเภอ โดยพิจารณาจากตัวแปรต่าง ๆ ทางด้านเศรษฐกิจและสังคม ซึ่งได้กล่าวถึงตัวแปร และความหมายของตัวแปรไว้ในบทที่ 2 แล้ว ส่วนวิธีการวิเคราะห์ที่ใช้มีดังนี้

3.1 การคัดเลือกตัวแปร เพื่อจะนำไปใช้ในการวิเคราะห์

ในขั้นแรกจะคัดเลือกตัวแปรที่จะนำไปใช้ในการวิเคราะห์ด้วยวิธีทางสถิติต่าง ๆ โดยพิจารณาตัวแปรที่มีความสำคัญพอที่จะอธิบายถึงการจำแนกพื้นที่เพื่อประกาศเขตปฏิรูปที่ดินได้ ซึ่งจะพิจารณาจากความสัมพันธ์ของตัวแปรแต่ละตัวในเมตริกความสัมพันธ์ (Correlation Matrix) เพื่อขจัดอิทธิพลของพหุสัมพันธ์ในตัวแปรอิสระ (Multicollinearity) ของตัวแปรอิสระออกไปบ้าง ได้กำหนดสัญลักษณ์ของตัวแปรต่าง ๆ ดังนี้

Y_i = รายได้เฉลี่ยต่อครัวเรือนต่อปีของประชากรในอำเภอที่ i (บาท/ครัวเรือน/ปี)

X_1 = เนื้อที่ถือครองเพื่อการเกษตรเฉลี่ยต่อครัวเรือน (ไร่/ครัวเรือน)

X_2 = ร้อยละของเนื้อที่ถือครองเพื่อการเกษตร

X_3 = ร้อยละของเนื้อที่เช่า

X_4 = ร้อยละของผู้ถือครองที่มีการเช่า

X_5 = ร้อยละของผู้ถือครองที่ปราศจากที่ดิน

X_6 = ร้อยละของผู้ถือครองที่มีขนาดการถือครองตั้งแต่ 100 ไร่

X_7 = ร้อยละของเนื้อที่ชลประทาน

X_8 = ร้อยละของเนื้อที่ที่ปลูกพืชได้มากกว่า 1 ครั้ง

X_9 = เนื้อที่ปลูกข้าวเฉลี่ยต่อครัวเรือน

X_{10} = ร้อยละของครัวเรือนที่ปลูกข้าว

X_{11} = ร้อยละของครัวเรือนที่มีอาชีพหลักทางการเกษตร

X_{12} = ร้อยละของครัวเรือนที่มีรายได้เฉลี่ยต่อปีตั้งแต่ 5,168 บาทขึ้นไป

X_{13} = ร้อยละของครัวเรือนที่นำที่ดินไปจำนองหรือขายฝากพ่อค้าเอกชน

X_{14} = ราคาที่ดินเฉลี่ยต่อไร่

X_{15} = จำนวนธนาคารข้าว

X_{16} = จำนวนธนาคารโคกระบือ

X_{17} = จำนวนกลุ่มเกษตรกร

X_{18} = จำนวนกลุ่มยุวเกษตรกร

X_{19} = จำนวนกลุ่มสหกรณ์ต่าง ๆ

X_{20} = จำนวนกลุ่มชลประทานราษฎร์

X_{21} = จำนวนธนาคารต่าง ๆ ได้แก่ ธนาคารออมสิน ธนาคารเพื่อการเกษตรและสหกรณ์การเกษตร ธนาคารพาณิชย์

X_{22} = จำนวนโรงสีข้าวขนาดใหญ่

X_{23} = จำนวนโรงสีข้าวขนาดกลาง

X_{24} = จำนวนโรงสีข้าวขนาดเล็ก

X_{25} = อัตราการเพิ่มของประชากร

X_{26} = ความหนาแน่นของประชากร

X_{27} = จำนวนนักเรียนต่อครู 1 คน

X_{28} = จำนวนประชากรในอำเภอเฉลี่ยต่อสถานีอนามัย 1 โรง

X_{29} = ร้อยละของครัวเรือนที่มีไฟฟ้าใช้

X_{30} = ร้อยละของครัวเรือนที่มีเครื่องรับวิทยุ

X_{31} = ร้อยละของครัวเรือนที่มีเครื่องรับโทรทัศน์

X_{32} = ผลผลิตข้าวเฉลี่ยต่อไร่ (กก./ไร่)

X_{33} = เนื้อที่เช่าเฉลี่ยต่อครัวเรือน (ไร่/ครัวเรือน)

X_{34} = ร้อยละของผู้ถือครองที่เป็นเจ้าของที่ดิน

โดยจะได้แปลงข้อมูลทั้งหมดให้อยู่ในรูปของค่ามาตรฐาน Z (Standardized data) เพื่อให้ข้อมูลอยู่ในหน่วยเดียวกัน และสะดวกในการเปรียบเทียบ

$$ZY = \frac{Y - \mu_y}{\sigma_y}$$

$$ZX_i = \frac{X_i - \mu_{x_i}}{\sigma_{x_i}}$$

ในทางปฏิบัติจะไม่ทราบค่า μ_y , μ_{x_i} , σ_y และ σ_{x_i} ดังนั้นจากตัวอย่างที่ได้จะได้ค่าประมาณของ μ_y , μ_{x_i} , σ_y และ σ_{x_i} เป็น $\bar{Y}$, $\bar{X}_i$, S_y และ S_{x_i} ตามลำดับ

โดยที่ $$\bar{Y} = \frac{1}{n} \sum_{j=1}^{n} Y_j$$

$$\bar{X}_i = \frac{1}{n} \sum_{j=1}^{n} X_{ij}$$

$$S_y = \sqrt{\frac{1}{n-1} \sum_{j=1}^{n} (Y_j - \bar{Y})^2}$$

$$S_{x_i} = \sqrt{\frac{1}{n-1} \sum_{j=1}^{n} (X_{ij} - \bar{X}_i)^2}$$

ในการคัดเลือกตัวแปร เพื่อจะใช้ในการวิเคราะห์ จะพิจารณาเป็น 2 ตอนคือ

1) พิจารณาค่าความสัมพันธ์ของตัวแปรตาม (Dependent Variable ; ZY) กับตัวแปรอิสระใด ๆ(Independent Variable ; ZX_i) ถ้าค่าสัมบูรณ์ของค่าความสัมพันธ์ระหว่าง ZY กับ ZX_i ใด ๆ $|\gamma_{zy,zx_i}| < .10$ จะตัด ZX_i นั้นออกจากการวิเคราะห์ เพราะถือว่า ZX_i ตัวนั้นมีความสัมพันธ์กับ ZY น้อยมาก

2) พิจารณาค่าความสัมพันธ์ของตัวแปรอิสระด้วยกัน (Independent Variables ; ZX_i , ZX_j) จะพิจารณาจากเมตริกความสัมพันธ์[1] (Examination of Correlation Matrix) ซึ่งเป็นวิธีขจัดพหุสัมพันธ์ในตัวแปรอิสระวิธีหนึ่ง โดยดูที่ค่าของความสัมพันธ์ของ ZX_i กับ ZX_j ในเมตริกความสัมพันธ์ ถ้าตัวแปรอิสระ ZX_i กับ ZX_j คู่ใดมีค่าสัมบูรณ์ของค่าความสัมพันธ์ $|\gamma_{ZX_i,ZX_j}|$ เข้าใกล้ 1 ถือว่าตัวแปรอิสระคู่นั้นมีความสัมพันธ์กันค่อนข้างสูง จะใช้ตัวแปรอิสระตัวใดตัวหนึ่งแทนโดยตัดอีกตัวหนึ่งออกไปจากการวิเคราะห์ได้ ในการวิจัยครั้งนี้ ถ้า $|\gamma_{ZX_i\ ZX_j}| > .70$ จะตัดตัวแปรอิสระตัวใดตัวหนึ่งออกไป และยังอาจถือได้ว่าค่าสัมประสิทธิ์ของการตัดสินใจ (Coefficient of Determination; $\gamma^2_{ZX_i,ZX_j}$) ซึ่งเป็นค่าแสดงอิทธิพลของ ZX_i, ZX_j ที่มีต่อกันถึงประมาณ 50% ขึ้นไป และเป็นค่าที่มากพอ

<u>การทดสอบการกระจายของข้อมูล</u> (Test for Goodness of fit) ว่าข้อมูลมีการกระจายแบบปกติ และมีค่าเฉลี่ย = 0 มีค่าความแปรปรวน = 1 โดยใช้การทดสอบโคโมโกรอฟ-สมินอฟ (Kolmogorov-Smirnov Test)[2]

คุณสมบัติของการทดสอบโคโมโกรอฟ-สมินอฟ

(1) ข้อมูลที่จะใช้ทดสอบ จะต้องเป็นข้อมูลเชิงปริมาณ สามารถให้ลำดับได้ (Ordinal)

(2) ทราบฟังก์ชั่นความน่าจะเป็นของข้อมูล (probability density function ; pdf.) และ pdf. เป็นแบบต่อเนื่อง และถ้าไม่ทราบค่าพารามิเตอร์ จะประมาณ

---

[1]Douglas C. Montgomery, <u>Introduction to linear regression analysis</u>, (New York : John Wiley & Son, 1981), p. 296-299.

[2]Sidney Seigel, <u>Nonparametric Statistics for Behavioral Sciences</u>, (Japan : McGraw Hill. Kogakusha, LTD., 1956). p. 47-52.

ค่าพารามิเตอร์

(3) ใช้เมื่อจำนวนตัวอย่างมีน้อย

หลักการของวิธีการนี้คือ วิธีการทดสอบโคโมโกรอฟ-สมินอฟ เป็นการเปรียบเทียบระหว่างความถี่สะสม ที่ได้ทางทฤษฎีกับการที่ได้จากการสังเกตในกรณีนี้มีสมมติฐานดังนี้

H_o : ข้อมูลที่ใช้มีการกระจายแบบปกติ

ตัวสถิติที่ใช้ทดสอบคือ $D = \text{maximum} \mid F_o(X) - S_n(X) \mid$

โดยที่ $F_o(X)$ แทนฟังก์ชั่นความน่าจะเป็นสะสมของ X ภายใต้ H_o

$S_n(X)$ แทนความถี่สะสมของ X ในการสุ่มตัวอย่าง n ครั้ง

$= \frac{k}{n}$ เมื่อ k = จำนวนครั้งที่สังเกตได้

ซึ่งจะมีค่าน้อยกว่าหรือเท่ากับ X

อาณาเขตวิกฤตที่ระดับนัยสำคัญ α คือ $D \geq D_{n,\alpha}$

เมื่อ $D_{n,\alpha}$ คือค่าที่เปิดจากตาราง

n คือจำนวนตัวอย่าง

α คือระดับนัยสำคัญ

ดังนั้น จะปฏิเสธ H_o เมื่อ $D \geq D_{n,\alpha}$

3.2 วิธีวิเคราะห์การถดถอยเชิงซ้อน

จากแนวความคิดเรื่องการวัดความยากจนและจากการศึกษาของธนาคารโลกเกี่ยวกับความยากจน ซึ่งใช้รายได้ของประชากรมาเป็นตัววัดสภาพความเป็นอยู่ของประชากร จะกำหนดน้ำหนักของตัวแปรต่าง ๆ ที่จะนำมาสร้างแบบจำลองด้วยการใช้วิธีวิเคราะห์การถดถอยเชิงซ้อน โดยให้รายได้เฉลี่ยต่อครัวเรือนต่อปีของประชากรในอำเภอเป็นตัวแปรตาม ซึ่งมีรูปแบบดังนี้

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

โดยที่ β_i เป็นค่าสัมประสิทธิ์การถดถอยเชิงซ้อน ไม่ทราบค่าของตัวแบบ

$i = 0,1,2,\ldots,k$

และถ้าหากว่ามีจำนวนข้อมูล n ตัว ต่อตัวแปรแต่ละตัวแล้วจะได้รูปของข้อมูลเป็นดังนี้คือ

ข้อมูล	ตัวแปรตาม	ตัวแปรอิสระที่ 1	ที่ 2	---	ที่ k
1	Y_1	X_{11}	X_{21}		X_{k1}
2	Y_2	X_{12}	X_{22}		X_{k2}
3	Y_3	X_{13}	X_{23}		X_{k3}
.	.	.	.		.
.	.	.	.		.
.	.	.	.		.
n	Y_n	X_{1n}	X_{2n}		X_{kn}

จากข้อมูลข้างต้นนี้ หากนำมาเขียนในรูปของสมการเมตริกจะได้เป็น

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ . \\ . \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ 1 & X_{12} & X_{22} & \cdots & X_{k2} \\ 1 & X_{13} & X_{23} & \cdots & X_{k3} \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_o \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

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

โดยที่ $\underset{\sim}{Y}$ คือข้อมูลของตัวแปรตาม โดยมีเวคเตอร์อันดับที่ nxl

X คือข้อมูลของตัวแปรอิสระ โดยมีเมตริกซ์อันดับที่ nx(k+1)

$\underset{\sim}{\beta}$ คือสัมประสิทธิ์ของการถดถอย โดยมีเวคเตอร์อันดับที่ (k+1)xl

$\underset{\sim}{\varepsilon}$ คือค่าความคลาดเคลื่อนของค่าประมาณจากค่าจริงโดยมีเวคเตอร์อันดับที่ nxl

3.2.1 <u>ข้อสมมติเบื้องต้น</u>

ในการที่เราจะศึกษาถึงวิธีการประมาณและคุณสมบัติของรูปแบบ ซึ่งเราไม่ทราบค่านั้น เราจะต้องตั้งข้อสมมติ ซึ่งจำเป็นสำหรับการประมาณเสียก่อน และข้อสมมติที่จะกล่าวถึงต่อไปนี้ก็มีความสำคัญมาก เพราะคุณสมบัติของตัวประมาณโดยวิธีกำลังสองน้อยที่สุด (Least Squares Method) ซึ่งเป็นวิธีที่เรานำมาใช้ จะขึ้นอยู่กับข้อสมมติเหล่านั้น ซึ่งจะต้องนำมาพิจารณาประกอบก่อนลงมือทำการวิเคราะห์คือ

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

นั่นคือ Y_i เป็นฟังก์ชั่นเชิงเส้นของ X_{ij} กับ ε_i โดยที่ $i = 1,2,\ldots, n$

$j = 0,1,\ldots, k \quad X_{0i} = 1$

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

นั่นคือ $E(\varepsilon) = 0$ สำหรับทุกค่า i

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

นั่นคือ $E(\varepsilon_i^2) = \sigma^2$ สำหรับทุกค่า i นั่นคือ ε_i มีความแปรปรวนคงที่

$E(\varepsilon_i\varepsilon_j) = 0$ เมื่อ $i \neq j$ แสดงว่าแต่ละคู่ของ ε ไม่เกี่ยวข้องกัน (pairwise uncorrelated)

4) ค่าความคลาดเคลื่อนของค่าประมาณจากค่าจริง ε_i เป็นตัวแปรเชิงสุ่มที่มีการกระจายแบบปกติ $N(0,\sigma^2)$ ซึ่งข้อนี้จำเป็นต้องใช้ในกรณีที่ต้องการสร้างช่วงความเชื่อมั่นของ β_0 , β_1 , β_2 ,..., β_k หรือเมื่อต้องการทดสอบสมมติฐานใด ๆ เกี่ยวกับตัวประมาณ

5) X เป็นเมตริกอันดับที่ nx(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 <n หมายความว่าจำนวนค่าสังเกต n จะมากกว่าจำนวนพารามิเตอร์ (k+1) ที่จะต้องประมาณค่า

จากข้อสมมติดังกล่าวมาแล้วข้างต้นเราจะเห็นได้ว่า

1) $E(\underset{\sim}{Y}/X) = X\underset{\sim}{\beta} + E(\underset{\sim}{\varepsilon}/X) = X\underset{\sim}{\beta}$ เช่น

$$E(Y_i/X_{1i},X_{2i},\ldots,X_{ki}) = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki}$$

2) $E\left[(\underset{\sim}{Y} - E(\underset{\sim}{Y}))(\underset{\sim}{Y} - E(\underset{\sim}{Y}))'/X\right] = E(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}'/X) = \sigma^2 I$

เช่น $E\left[(Y_i - E(Y_i))(Y_i - E(Y_i))'/X_{1i},X_{2i},\ldots,X_{ki}\right] = \sigma^2$ สำหรับทุกค่า i

3.2.2 การประมาณค่าพารามิเตอร์

วิธีการในการประมาณค่าพารามิเตอร์การถดถอยนั้น กระทำได้ 2 วิธีคือ

1) วิธีกำลังสองน้อยที่สุด

2) วิธีน่าจะเป็นสูงสุด (Maximum Likelihood Method)

ทั้ง 2 วิธีดังกล่าวเป็นวิธีในการคำนวณหาค่าเส้นถดถอยจากข้อมูลที่กำหนดให้หรือกล่าวอีกนัยหนึ่งก็คือ ถ้ากำหนดสมการ

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

$\hat{\underset{\sim}{Y}} = X\underset{\sim}{b}$ นั่นเอง หากแต่ว่าแต่ละวิธีก็มีคุณสมบัติของตัวเอง

แต่ในที่นี้จะนำมากล่าวถึงเพียงวิธีเดียวเท่านั้นคือ วิธีที่ 1 ซึ่งถือว่าเป็นวิธีที่สำคัญและนิยมใช้กันมาก โดยมีหลักการของการประมาณแบบกำลังสองน้อยที่สุด คือการทำให้ผลรวมกำลังสองของส่วนเบี่ยงเบนของค่าสังเกตต่างจากค่าเฉลี่ยมีค่าน้อยที่สุด และจะทำให้ได้ตัวประมาณที่ไม่เอนเอียง

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

และ $E(\underset{\sim}{Y}) = X\underset{\sim}{\beta}$ $\quad \because \quad E(\underset{\sim}{\varepsilon}) = 0$

ซึ่ง $E(\underset{\sim}{Y})$ คือค่าเฉลี่ยของ $\underset{\sim}{Y}$

ดังนั้น $\underset{\sim}{\varepsilon} = \underset{\sim}{Y} - E(\underset{\sim}{Y})$

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

โดยวิธีกำลังสองน้อยที่สุด จะต้องหาก่อนว่ากำลังสองของค่าความคลาดเคลื่อนเท่ากับ

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

$$= \underset{\sim}{Y}'\underset{\sim}{Y} - \underset{\sim}{\beta}'X'\underset{\sim}{Y} - \underset{\sim}{Y}'X\underset{\sim}{\beta} + \underset{\sim}{\beta}'X'X\underset{\sim}{\beta}$$

$$= \underset{\sim}{Y}'\underset{\sim}{Y} - 2\underset{\sim}{\beta}'X'Y + \underset{\sim}{\beta}'X'X\underset{\sim}{\beta}$$

[เนื่องจาก $\underset{\sim}{Y}'X\underset{\sim}{\beta}$ เป็นสเกลล่า ดังนั้น ทรานสโพสของมันจึงมีค่าเท่าเดิม จึงได้ว่า $\underset{\sim}{Y}'X\underset{\sim}{\beta} = \underset{\sim}{\beta}'X'\underset{\sim}{Y}$]

จากนั้นจึงเริ่มหาตัวประมาณโดยทำให้กำลังสองของค่าความคลาดเคลื่อนที่ได้มานั้นมีค่าน้อยที่สุดโดยใช้ การดิฟเฟอเรนซิเอท

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

$$-2X'\underset{\sim}{Y} + 2X'X\underset{\sim}{b} = 0$$

$$X'\underset{\sim}{Y} = X'X\underset{\sim}{b}$$

จากนี้จะหาตัวประมาณของ $\underset{\sim}{\beta}$ คือ $\underset{\sim}{b}$

ซึ่งเป็นชุดสมการปกติ (normal equations) และถ้าหากจะนำมาเขียนในรูปสเกลล่าจะได้ว่า

$$\Sigma Y_i = b_o n + b_1 \Sigma X_{1i} + b_2 \Sigma X_{2i} + \ldots + b_k \Sigma X_{ki}$$

$$\Sigma X_{1i} Y_i = b_o \Sigma X_{1i} + b_1 \Sigma X_{1i}^2 + b_2 \Sigma X_{1i} X_{2i} + \ldots + b_k \Sigma X_{1i} X_{ki}$$

$$\Sigma X_{2i} Y_i = b_o \Sigma X_{2i} + b_1 \Sigma X_{1i} X_{2i} + b_2 \Sigma X_{2i}^2 + \ldots + b_k \Sigma X_{2i} X_{ki}$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$\Sigma X_{ki} Y_i = b_o \Sigma X_{ki} + b_1 \Sigma X_{1i} X_{ki} + b_2 \Sigma X_{2i} X_{ki} + \ldots + b_k \Sigma X_{ki}^2$$

การจะหา $\underset{\sim}{b}$ ได้จะต้องมีเงื่อนไขว่าชุดสมการปกติทั้งหมด k+1 สมการนี้ต้องเป็นอิสระซึ่งกันและกันโดยที่ $X'X$ จะต้องเป็นเมตริกที่ไม่เอกเทศ เพราะจะทำให้สามารถหา $(X'X)^{-1}$ ได้

$$\underset{\sim}{b} = (X'X)^{-1} X'\underset{\sim}{Y}$$

และเมื่อทำการดิฟเฟอเรนชิเอทอันดับที่ 2 จะได้

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

ซึ่งเป็นเดฟฟินิทท่างบวกเมื่อ X มี rank เต็มตามจำนวนคอลัมน์คือ k+1

เพราะฉะนั้น $\underset{\sim}{b} = (X'X)^{-1} X'\underset{\sim}{Y}$ จึงเป็นตัวประมาณที่ได้จากการทำให้กำลังสองของความคลาดเคลื่อนมีค่าน้อยที่สุด

จึงได้ว่า $\hat{\underset{\sim}{Y}} = X\underset{\sim}{b}$

ต่อไปจะพิจารณาค่าความแปรปรวนของ $\underset{\sim}{b}$

จาก
$$\begin{aligned} \underset{\sim}{b} &= (X'X)^{-1} X'Y \\ &= (X'X)^{-1} X' (X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}) \\ &= \underset{\sim}{\beta} + (X'X)^{-1} X'\underset{\sim}{\varepsilon} \\ \underset{\sim}{b} - \underset{\sim}{\beta} &= (X'X)^{-1} X'\underset{\sim}{\varepsilon} \end{aligned}$$

$$E(\underset{\sim}{b} - \underset{\sim}{\beta})(\underset{\sim}{b} - \underset{\sim}{\beta})' = E\left[(X'X)^{-1} X'\underset{\sim}{\varepsilon}\right]\left[(X'X)^{-1} X'\underset{\sim}{\varepsilon}\right]'$$

$$= E\left[(X'X)^{-1} X'\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}' X(X'X)^{-1}\right]$$

$$= (X'X)^{-1} X' E(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}') X(X'X)^{-1}$$

$$= \sigma^2 (X'X)^{-1} \quad (\because E(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}') = \sigma^2 I)$$

นั่นคือ $\text{Var}(\underset{\sim}{b}) = \sigma^2 (X'X)^{-1}$

จากข้างต้นจึงได้ว่า $E(\underset{\sim}{b} - \underset{\sim}{\beta})(\underset{\sim}{b} - \underset{\sim}{\beta})'$ เป็นวาเรียนซ์-โควาเรียนซ์ (Variance-Covariance) ของ $\underset{\sim}{b}$ แต่เนื่องจากไม่ทราบค่าของ σ^2 จึงต้องประมาณค่าของ σ^2 ก่อน

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

และ $\underset{\sim}{Y} = X\underset{\sim}{b}$

ให้ $\underset{\sim}{e} = \underset{\sim}{Y} - \hat{\underset{\sim}{Y}}$

$$= \underset{\sim}{Y} - X\underset{\sim}{b}$$

$$= \underset{\sim}{Y} - X(X'X)^{-1} X'\underset{\sim}{Y}$$

$$= (I - X(X'X)^{-1} X')\underset{\sim}{Y}$$

$$= M\underset{\sim}{Y} \quad (\text{ให้ } M = I - X(X'X)^{-1} X')$$

$$= M(X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon})$$

$$= MX\underset{\sim}{\beta} + M\underset{\sim}{\varepsilon}$$

$$= M\underset{\sim}{\varepsilon}$$

ทั้งนี้เนื่องจาก $MX = 0$

โดย $MX = (I - X(X'X)^{-1}X')X$

$= X - X(X'X)^{-1}X'X$

$= 0$

หา $\underset{\sim}{e}'\underset{\sim}{e} = \underset{\sim}{\varepsilon}'MM\underset{\sim}{\varepsilon}$

$= \underset{\sim}{\varepsilon}'M\underset{\sim}{\varepsilon}$

ทั้งนี้เนื่องจาก $MM = (I - X(X'X)^{-1}X')(I - X(X'X)^{-1}X')$

$= I - X(X'X)^{-1}X' - X(X'X)^{-1}X + X(X'X)^{-1}X'X(X'X)^{-1}X'$

$= I - X(X'X)^{-1}X'$

$= M$

$\therefore \underset{\sim}{e}'\underset{\sim}{e} = tr\,\underset{\sim}{\varepsilon}'M\underset{\sim}{\varepsilon}$

ทั้งนี้เนื่องจาก M เป็นสเกลล่าจึงเท่ากับ Trace ของตัวมันเอง

$\underset{\sim}{e}'\underset{\sim}{e} = tr\,M\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}'$

$\therefore tr(AB) = tr(BA)$

$\therefore E(\underset{\sim}{e}'\underset{\sim}{e}) = E(tr\,M\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}')$

$= tr\,M.\,E(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}')$

เพราะว่า trace เป็นฟังก์ชั่นเชิงเส้น

$E(\underset{\sim}{e}'\underset{\sim}{e}) = \sigma^2\,tr\,M \quad (\because E(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}') = \sigma^2 I)$

$= \sigma^2(n-k-1)$

ทั้งนี้เนื่องจาก $\text{tr } M = \text{tr } (I - X(X'X)^{-1} X')$

$= \text{tr } (I_n) - \text{tr } (X(X'X)^{-1} X')$

$= n - \text{tr} X'X(X'X)^{-1}$ $(\text{tr } (AB) = \text{tr } (BA))$

$= n - \text{tr } (I_{k+1})$

$= n - k - 1$

ดังนั้นจึงได้ว่า $s^2 = \frac{\underset{\sim}{e}' \underset{\sim}{e}}{n - k - 1}$

$= \frac{\underset{\sim}{Y}' M \underset{\sim}{Y}}{n - k - 1}$ ซึ่งเป็นตัวประมาณที่ไม่เอนเอียงของ σ^2

และ $s^2(X'X)^{-1}$ ก็เป็นตัวประมาณที่ไม่เอนเอียงของ $\sigma^2(X'X)^{-1}$

3.2.3 <u>คุณสมบัติของตัวประมาณโดยวิธีกำลังสองน้อยที่สุด</u>

จากข้อสมมติที่กำหนดให้ในหัวข้อข้างต้น ถ้าหากเป็นจริงแล้ว จะได้ว่า การหาค่าตัวประมาณโดยวิธีกำลังสองน้อยที่สุดจะมีคุณสมบัติดังนี้คือ

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

พิสูจน์

$\underset{\sim}{b} = (X'X)^{-1} X'Y$

$= (X'X)^{-1} X'(X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon})$

$= (X'X)^{-1} X'X\underset{\sim}{\beta} + (X'X)^{-1} X'\underset{\sim}{\varepsilon}$

$= \underset{\sim}{\beta} + (X'X)^{-1} X'\underset{\sim}{\varepsilon}$

$E(\underset{\sim}{b}) = \underset{\sim}{\beta}$

นั่นคือ จะได้ว่า $\underset{\sim}{b} = (X'X)^{-1} X'Y$ เป็นตัวประมาณที่ไม่เอนเอียงของ $\underset{\sim}{\beta}$

2. $\underset{\sim}{b} = (X'X)^{-1}X'Y$ จะเป็นตัวประมาณที่ไม่เอนเอียงเชิงเส้นที่ดีที่สุด (Best Linear Unbiased Estimator) ในความหมายคือ ตัวประมาณอื่น ๆ ของ $\underset{\sim}{\beta}$ ซึ่งเป็นเชิงเส้นกับ $\underset{\sim}{Y}$ และไม่เอนเอียงเช่นกัน จะมีค่าวาเรียนซ์-โควาเรียนซ์เมตริก ซึ่งมีค่าเกินกว่าวาเรียนซ์-โควาเรียนซ์ของตัวประมาณ $\underset{\sim}{b}$ หรืออาจจะกล่าวได้อย่างง่าย ๆ ก็คือในบรรดาตัวประมาณในจำพวกของตัวประมาณที่ไม่เอนเอียงเชิงเส้นด้วยกันแล้วตัวประมาณโดยวิธีกำลังสองน้อยที่สุดนั้นจะมีค่าความแปรปรวนต่ำสุด โดยทฤษฎีของกอลล์-มาคอฟฟ์และเป็นที่นิยมใช้กันอย่างแพร่หลาย

จากการที่ตัวประมาณของ $\underset{\sim}{\beta}$ เป็นเชิงเส้นกับ Y เราจึงสามารถเขียนตัวประมาณนั้นได้เป็น $A^{*}\underset{\sim}{Y}$ โดย A^{*} เป็นเมตริก $(k+1)\times n$ ซึ่งไม่ขึ้นอยู่กับค่าของ Y

$$\text{ให้ } A^{*} = (X'X)^{-1}X' + A$$

$$\therefore A^{*}Y = \left[(X'X)^{-1}X' + A\right]\underset{\sim}{Y}$$

$$= \left[(X'X)^{-1}X' + A\right](X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon})$$

$$= \left[I + AX\right]\underset{\sim}{\beta} + \left[(X'X)^{-1}X + A\right]\underset{\sim}{\varepsilon}$$

$$E(A^{*}Y) = (I + AX)\underset{\sim}{\beta} + 0$$

$$= \underset{\sim}{\beta} + AX\underset{\sim}{\beta}$$

ตัวประมาณ $A^{*}Y$ จะเป็นตัวประมาณที่ไม่เอนเอียง ถ้า $AX = 0$

$$\therefore \text{จาก } A^{*}\underset{\sim}{Y} = \underset{\sim}{\beta} + \left[(X'X)^{-1}X' + A\right]\underset{\sim}{\varepsilon}$$

$$A^{*}\underset{\sim}{Y} - \underset{\sim}{\beta} = \left[(X\ X)^{-1}X + A\right]\underset{\sim}{\varepsilon}$$

$$(A^{*}\underset{\sim}{Y}-\underset{\sim}{\beta})(A^{*}\underset{\sim}{Y}-\underset{\sim}{\beta})' = \left[(X'X)^{-1}X' + A\right]\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}'\left[(X'X)^{-1}X' + A\right]'$$

$$E(A^{*}\underset{\sim}{Y}-\underset{\sim}{\beta})(A^{*}\underset{\sim}{Y}-\underset{\sim}{\beta})' = \left[(X'X)^{-1}X' + A\right]E(\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}')\left[X(X'X)^{-1} + A'\right]$$

$$= \sigma^{2}\left[(X'X)^{-1}X' + A\right]\left[X(X'X)^{-1} + A'\right]$$

$$= \sigma^{2}\left[(X'X)^{-1} + AX(X'X)^{-1} + (X'X)^{-1}X'A' + AA'\right]$$

$$= \sigma^{2}\left[(X'X)^{-1} + AA'\right]$$

เนื่องจากเรากำหนดให้ $AX = 0 \rightarrow X'A' = 0$ ด้วย

ดังได้ว่า $Var(A^*\underset{\sim}{Y}) = \sigma^2 (X'X)^{-1} + \sigma^2 AA'$

$= Var(\underset{\sim}{b}) + \sigma^2 AA'$

นั่นคือ วาเรียนซ์-โควาเรียนซ์ของตัวประมาณจะมีค่าเกิน วาเรียนซ์-โควาเรียนซ์ของตัวประมาณ ซึ่งประมาณโดยวิธีกำลังสองน้อยที่สุด

$\rightarrow \underset{\sim}{b} = (X'X)^{-1} X'Y$ จะเป็นตัวประมาณที่ไม่เอนเอียงเชิงเส้นที่ดีที่สุด

3.3 <u>วิธีวิเคราะห์องค์ประกอบหลัก</u>

เป็นวิธีที่ให้แบบแผนความสัมพันธ์ระหว่างปัจจัยต่าง ๆ และแยกปัจจัยเหล่านี้เป็นกลุ่ม ๆ ตามความสัมพันธ์ที่เกี่ยวข้อง ตัวประกอบที่จะให้ค่าถ่วงน้ำหนัก (Loading) ซึ่งสามารถแปลงคะแนนดิบของปัจจัยนั้น ๆ ให้อยู่ในรูปค่าถ่วงน้ำหนัก เพื่อจัดลำดับความสำคัญให้ได้ชัดเจนถูกต้องยิ่งขึ้น

หลักการของวิธีวิเคราะห์องค์ประกอบโดยทั่วไป คือหาตัวแปรใหม่ที่แสดงความสัมพันธ์เชิงเส้นตรง (Linear Combination) ของตัวแปรเดิมหลาย ๆ ตัวแปร

ขั้นตอนของการคำนวณหารูปแบบของตัวแปรใหม่ในองค์ประกอบ (component)ต่าง ๆ

1) กำหนดตัวแปรที่ใช้ในการศึกษาวัดความสำคัญของพื้นที่ ให้ X เป็นตัวแปรต่าง ๆ ที่ใช้วัด

$$X = \begin{bmatrix} X_{11} & X_{12} \cdots X_{1k} \\ X_{21} & X_{22} \cdots X_{2k} \\ \vdots & \vdots \quad \vdots \\ X_{n1} & X_{n2} \cdots X_{nk} \end{bmatrix}$$

มีตัวแปร k ตัวเช่นเดียวกับการวิเคราะห์การถดถอยเชิงซ้อนและมีพื้นที่ n พื้นที่

2) หาความแปรปรวนร่วม (Covariance) ของตัวแปรทั้งหมด

$$S = E\left[(X - E(X))\ (X - E(X)) \right]$$

S = Covariance matrix ของ X

3) หาค่าคาแรคเตอรีสติค รูท (Characteristic roots ; λ)

$$| S - \lambda I | = 0$$

I = Identity matrix

$$\lambda = \left[\lambda_1\ \lambda_2\ \ldots\ \lambda_k \right]$$

λ_i = คาแรคเตอรีสติค รูทที่ i ของ S

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \ldots \geq \lambda_k$$

4) หาค่าคาแรคเตอรีสติค เวคเตอร์ (Characteristic vectors; $\underset{\sim}{a}$)

$$S\underset{\sim}{a} = \lambda\, \underset{\sim}{a}$$

a_i = คาแรคเตอรีสติคเวคเตอร์ที่ i ที่สอดคล้องกับคาแรคเตอรีสติค รูท λ_i

$$a_i = \begin{Bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ \cdot \\ a_k \end{Bmatrix}$$

5) หา normalized ของคาแรคเตอรีสติคเวคเตอร์

$$\text{normalized} = \begin{bmatrix} a_1 / \sqrt{\text{lenght}} \\ a_2 / \sqrt{\text{lenght}} \\ \vdots \\ a_k / \sqrt{\text{lenght}} \end{bmatrix}$$

$$\text{lenght} = a_1^2 + a_2^2 + \ldots + a_k^2$$

นำค่า normalizedของคาแรคเตอรีสติคเวคเตอร์ มาเรียงกันจะได้พรินสิปอลคอมโพเนนท์เวคเตอร์ (Principal Component Vector)

6) หมุนแกนองค์ประกอบให้แกนตั้งฉากซึ่งกันและกัน (Orthogonal) จะได้องค์ประกอบที่เป็นอิสระต่อกัน ค่าที่ได้จากการหมุนแกนคือค่าองค์ประกอบ (Principal Component) ที่จะนำไปใช้ในการวิเคราะห์เพื่อหาตัวแปรที่สำคัญ โดยพิจารณาจากองค์ประกอบที่มีค่าความแปรปรวนสูงสุด

องค์ประกอบที่ i (Principal Component ที่ i)

$$Y_i = \underset{\sim}{a}_i' \underset{\sim}{X}_i = \sum_{j=1}^{k} a_{ij} X_j$$

$$Y_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ik}X_k$$

จะเลือกองค์ประกอบที่ให้ความแปรปรวนสูงสุด โดยพิจารณา

$$\underset{\sim}{a}_i' \underset{\sim}{a}_i = 1$$

$$\underset{\sim}{a}_i' \underset{\sim}{a}_j = 0 \qquad i \neq j$$

Y_1 จะมีค่าวาเรียนซ์สูงที่สุด

Y_2 จะมีค่าวาเรียนซ์รองลงมา

.
.
.

Y_k จะมีวาเรียนซ์ต่ำที่สุด

ดังนั้นจึงใช้องค์ประกอบที่ 1 (Principal Component ที่ 1)

รูปแบบที่ใช้ $Y_1 = a_{11}X_1 + a_{12}X_2 + \ldots + a_{1k}X_k$

หรือ $Y = a_1X_1 + a_2X_2 + \ldots + a_kX_k$

3.4 <u>วิธีวิเคราะห์สหสัมพันธ์คาโนนิคอล</u>

ตามปกติเมื่อมีตัวแปรสุ่มอยู่ 2 ตัวคือ X และ Y จะสามารถหาความสัมพันธ์ของตัวแปรทั้งสองได้ โดยพิจารณาจากสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) หรือ ρ เมื่อ

$$\rho = \frac{Cov\ (X\ ,\ Y)}{\sqrt{Var(X)\ Var(Y)}}$$

ถ้าค่า Y ขึ้นอยู่กับค่า X หลาย ๆ ตัว วิธีการที่จะสามารถหาความสัมพันธ์ระหว่าง Y และ X เหล่านั้นจะใช้วิธีวิเคราะห์สหสัมพันธ์พหุคูณ (Multiple Correlation)

โดยให้ X เป็นปัจจัยหรือตัวแปรที่ต้องการศึกษาซึ่งมีอยู่ P ตัว

ดังนั้น $X' = \left[X_1 \quad X_2 \ \ldots \ X_p \right]$

และ $\hat{Y}$ เป็นค่าประมาณของผลรวมเชิงเส้น

$$\hat{Y}_j = \hat{\beta}_1X_{1j} + \hat{\beta}_2X_{2j} + \ldots + \hat{\beta}_pX_{pj}, \quad j = 1,2,\ldots,n$$

เมื่อ $\hat{\beta}_s$ = (S = 1,2,..., P) เป็นค่าสัมประสิทธิ์การถดถอยเชิงซ้อน (Multiple Regression Coefficient) ซึ่งหาค่าได้จากการประมาณโดยวิธี Least-Squareอันจะทำให้สามารถหาค่าสัมประสิทธิ์สหสัมพันธ์โดยการหาค่าความสัมพันธ์อย่างง่าย (Simple Correlation) ระหว่าง Y และ X นั่นเอง ยังมีอีกกรณีหนึ่ง ซึ่งนอกจากจะมีตัวแปร X หลาย ๆ ตัวแล้วยังมี Y หลาย ๆ ตัวคือ

$$Y' = \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_q \end{bmatrix}$$

วิธีการที่จะหาความสัมพันธ์ระหว่าง X หลาย ๆ ตัว และ Y หลาย ๆ ตัวก็คือ วิธีวิเคราะห์สหสัมพันธ์คาโนนิคอล จะได้ค่าความสัมพันธ์คาโนนิคอล (Canonical Correlation) ซึ่งจะหาได้จากความสัมพันธ์ระหว่างผลรวมเชิงเส้นของกลุ่มตัวแปร X และผลรวมเชิงเส้นของกลุ่มตัวแปร Y

เมื่อกำหนดค่า

$$X_j^* = \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_p X_p \quad \text{.....(1)}$$

$$Y_j^* = \beta_1 Y_2 + \beta_2 Y_2 + \cdots + \beta_q Y_q$$

ส่มการนี้ยังไม่ทราบค่าของ α_i และ β_i ที่จะทำให้ความสัมพันธ์ระหว่าง Y^* และ X^* มีค่ามากที่สุดที่จะเป็นไปได้

ถ้าให้ ρ_c Canonical Correlation Coefficient ระหว่าง X^* และ Y^*

ซึ่ง $$\rho_c = \frac{\text{Cov}(X^*, Y^*)}{\sqrt{\text{Var}(X^*)\ \text{Var}(Y^*)}} \quad \text{.....(2)}$$

และ $\rho_c^{(1)}$ เป็น Canonical Correlation Coefficient ที่มีค่ามากที่สุด เรียกว่า First Canonical Correlation ซึ่งเกิดจาก X_1^* และ Y_1^*

$\rho_c^{(2)}$ เป็น Canonical Correlation Coefficient ที่มีค่ามากที่สุดเป็นอันดับสองเรียกว่า Second Canonical Correlation ซึ่งเกิดจาก X_2^* และ Y_2^*

$\rho_c^{(3)}$ เป็น Canonical Correlation Coefficient ที่มีค่ามากที่สุดเป็นอันดับสามเรียกว่า Third Canonical Correlation ซึ่งเกิดจาก X_3^* และ Y_3^*

. . . .
. . . .
. . . .
. . . .

X_1^* และ Y_1^* เรียกว่า first canonical variables

X_2^* และ Y_2^* เรียกว่า second canonical variables

X_3^* และ Y_3^* เรียกว่า third canonical variables

. .
. .
. .
. .

คู่ของ canonical variable ต่าง ๆ นั้น จะมีหลักการเหมือนกัน

กรณี $q < p$ ก็จะมี Canonical Correlation ทั้งหมด q ค่า

และ Canonical Variable ทั้งหมด q คู่

ผลลัพธ์ต่าง ๆ เหล่านี้ได้มาจากหลักการต่อไปนี้คือ

ถ้า X มีค่าเฉลี่ยเป็น μ_1 และความแปรปรวนเท่ากับ Σ_{11}

Y มีค่าเฉลี่ยเป็น μ_2 และความแปรปรวนเท่ากับ Σ_{22}

ให้ $\underset{\sim}{\Sigma} = \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right]_{(p+q)\,x\,(p+q)}$

โดยที่ Σ_{11} เป็นเมตริกของความแปรปรวน (Variable matrix) ของ X

ซึ่งมีขนาด pxp

Σ_{22} เป็นเมตริกของความแปรปรวน (Variable matrix) ของY

ซึ่งมีขนาด qxq

Σ_{12} เป็นเมตริกของความแปรปรวนร่วม (Covariance matrix) ของ X และ Y ซึ่งมีขนาด pxq

$\Sigma_{21} = \Sigma_{12}'$

$$\Sigma = \left[\begin{array}{ccc|ccc} \sigma_{X_1X_1} & \sigma_{X_1X_2} & \cdots \sigma_{X_1X_p} & \sigma_{X_1Y_1} & \sigma_{X_1Y_2} & \cdots \sigma_{X_1Y_q} \\ \sigma_{X_2X_1} & \sigma_{X_2X_2} & \cdots \sigma_{X_2X_p} & \sigma_{X_2Y_1} & \sigma_{X_2Y_2} & \cdots \sigma_{X_2Y_q} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \sigma_{X_pX_1} & \sigma_{X_pX_2} & \sigma_{X_pX_p} & \sigma_{X_pY_1} & \sigma_{X_pY_2} & \sigma_{X_pY_q} \\ \hline \sigma_{X_1Y_1} & \sigma_{X_2Y_1} & \cdots \sigma_{X_pY_1} & \sigma_{Y_1Y_1} & \sigma_{Y_1Y_2} & \cdots \sigma_{Y_1Y_q} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \sigma_{X_1Y_q} & \sigma_{X_2Y_q} & \cdots \sigma_{X_pY_q} & \sigma_{Y_qY_1} & \sigma_{Y_qY_2} & \cdots \sigma_{Y_qY_q} \end{array}\right]$$

Canonical variables ในรูปของ matrix คือ

$$\underset{\sim}{X}^* = X\underset{\sim}{\alpha}$$

$$\underset{\sim}{Y}^* = Y\underset{\sim}{\beta}$$

เมื่อ $\underset{\sim}{X}^*$ และ $\underset{\sim}{Y}^*$ เป็นเวคเตอร์ของ Canonical variable ขนาด pxl และ qxl ตามลำดับ

$\underset{\sim}{X}$ และ $\underset{\sim}{Y}$ เป็นเมตริกของตัวแปรสุ่มแรกเริ่ม (Original random variables) ขนาด nxp และ nxq ตามลำดับ

$\underset{\sim}{\alpha}$ เป็น Coefficient vectors ขนาด pxl

$\underset{\sim}{\beta}$ เป็น Coefficient vectors ขนาด qxl

ดังนั้นจากสมการ (2) สามารถเขียน Canonical Correlation Coefficient ได้เป็น

$$\rho_c = \frac{\text{Cov}(X\underset{\sim}{\alpha}, Y\underset{\sim}{\beta})}{\sqrt{\text{Var}(X\underset{\sim}{\alpha})\ \text{Var}(Y\underset{\sim}{\beta})}}$$

หรือ
$$\rho_c = \frac{\underset{\sim}{\alpha}' \Sigma_{12} \underset{\sim}{\beta}}{\sqrt{\underset{\sim}{\alpha}' \Sigma_{11} \underset{\sim}{\alpha}\ \ \underset{\sim}{\beta}' \Sigma_{22} \underset{\sim}{\beta}}} \qquad \text{.....(2a)}$$

หาค่า ρ_c จากสมการ (2a) โดยที่ต้องการได้ ρ_c ที่มีค่าสูงสุด ดังนั้นจะ Maximize ρ_c เทียบกับ $\underset{\sim}{\alpha}$ และ $\underset{\sim}{\beta}$ ซึ่งวิธีดังกล่าวนี้จะง่ายขึ้น เมื่อ Canonical Variable แต่ละตัวมีคุณสมบัติที่ว่าความแปรปรวนเป็น 1 นั่นคือ

$$\text{Var}(X_j^*) = \underset{\sim}{\alpha}\ \Sigma_{11}\ \underset{\sim}{\alpha} = 1 \qquad \text{.....(3)}$$

$$\text{Var}(Y_j^*) = \underset{\sim}{\beta}\ \Sigma_{22}\ \underset{\sim}{\beta} = 1 \qquad \text{.....(4)}$$

ซึ่งเงื่อนไขนี้จะทำให้ได้ $\rho_c = \underset{\sim}{\alpha}' \Sigma_{12} \underset{\sim}{\beta}$ เมื่อ Maximize ρ_c โดยวิธี Least Square ซึ่งจะได้ดังนี้

$$\Sigma_{12}\ \underset{\sim}{\beta} - \lambda\ \Sigma_{11}\ \underset{\sim}{\alpha} = 0 \qquad \text{.....(5)}$$

$$\Sigma_{21}\ \underset{\sim}{\alpha} - \lambda\ \Sigma_{22}\ \underset{\sim}{\beta} = 0 \qquad \text{.....(6)}$$

เมื่อ λ เป็น Lagrange multiplier

คูณสมการ (5) ด้วย λ ได้

$$\Sigma_{12}\ \lambda\underset{\sim}{\beta} = \lambda^2\ \Sigma_{11}\ \underset{\sim}{\alpha} \qquad \text{.....(7)}$$

คูณสมการ (6) ด้วย Σ_{22}^{-1}

$$\Sigma_{22}^{-1}\ \Sigma_{21}\ \underset{\sim}{\alpha} = \lambda\underset{\sim}{\beta} \qquad \text{.....(8)}$$

นำสมการ (8) แทนในสมการ (7)

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

คูณทั้งสองข้างด้วย Σ_{11}^{-1} และจัดเทอมใหม่ จะได้

$$(\Sigma_{11}^{-1}\ \Sigma_{12}\ \Sigma_{22}^{-1}\ \Sigma_{21} - \lambda^2 I)\ \underset{\sim}{\alpha} = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots(9)$$

ทำนองเดียวกันจะได้

$$(\Sigma_{22}^{-1}\ \Sigma_{21}\ \Sigma_{11}^{-1}\ \Sigma_{12} - \lambda^2 I) = 0 \quad \ldots\ldots\ldots\ldots\ldots\ldots(10)$$

ทั้งสมการ (9) และ (10) เป็น homogeneous equations ซึ่งสามารถหาคำตอบได้ แต่จากการที่ได้กำหนดไว้ก่อนแล้วว่า $q \leqslant p$ ดังนั้น ถ้าหากทราบค่า λ^2 และ $\underset{\sim}{\beta}$ แล้วก็จะสามารถหา $\underset{\sim}{\alpha}$ ได้จากสมการ (5)

เนื่องจาก

$$\underset{\sim}{\alpha} = \frac{\Sigma_{11}^{-1}\ \Sigma_{12}\ \underset{\sim}{\beta}}{\lambda} \quad \ldots\ldots\ldots\ldots\ldots\ldots(11)$$

จากสมการ (5) เมื่อคูณด้วย $\underset{\sim}{\alpha}'$ จะได้

$$\underset{\sim}{\alpha}'\ \Sigma_{12}\ \underset{\sim}{\beta} = \lambda \underset{\sim}{\alpha}'\ \Sigma_{11}\ \underset{\sim}{\alpha}$$

จากสมการ (6) เมื่อคูณด้วย $\underset{\sim}{\beta}'$ จะได้

$$\underset{\sim}{\beta}'\ \Sigma_{21}\ \underset{\sim}{\alpha} = \lambda \underset{\sim}{\beta}'\ \Sigma_{22}\ \underset{\sim}{\beta}$$

เนื่องจาก $\underset{\sim}{\alpha}'\ \Sigma_{11}\ \underset{\sim}{\alpha} = \underset{\sim}{\beta}'\ \Sigma_{22}\ \underset{\sim}{\beta} = 1$

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

นั่นคือ จากสมการ (2a) จะได้ว่า $\underset{\sim}{\alpha}'\ \Sigma_{12}\ \underset{\sim}{\beta} = \lambda$ เป็น Canonical Correlation Coefficient ภายใต้เงื่อนไขของสมการ (3) และสมการ (4)

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

ถ้าหากถอดรากที่สองของ Characteristic root coefficient ที่มีค่ามากที่สุด ของสมการ (9) หรือ (10) จะได้ first canonical correlation

จากสมการ (10) หาค่า λ^2 และ $\underset{\sim}{\beta}$ ได้ และจาก (11) หาค่า $\underset{\sim}{\alpha}$ ได้ ให้

$\lambda_1^2 > \lambda_2^2 > \ldots > \lambda_q^2$ เป็น Characteristic root ของสมการ (10)

$\underset{\sim}{\beta}_1, \underset{\sim}{\beta}_2, \ldots, \underset{\sim}{\beta}_q$ เป็น Characteristic vectors ที่สัมพันธ์กับ Characteristic root λ_i^2 ดังกล่าว

$\alpha_1, \alpha_2, \ldots, \alpha_q$ เป็น characteristic vectors ซึ่งคำนวณได้จากการแทนค่า λ_i และ $\underset{\sim}{\beta}_i$ ลงใน (11)

ดังนั้น Canonical Correlation Coefficient ที่ i คือ $\hat{\lambda}_i$

และเซทของ Canonical variables ก็คือ $X_i^* = X \underset{\sim}{\alpha}_i$

$Y_i^* = Y \underset{\sim}{\beta}_i$

ค่าสัมประสิทธิ์สหสัมพันธ์คาโนนิคอล (Canonical Correlation Coefficient) ที่ได้มีคุณสมบัติเช่นเดียวกับค่าสัมประสิทธิ์สหสัมพันธ์อย่างง่าย (Simple Correlation Coefficient) กล่าวคือค่าสัมประสิทธิ์สหสัมพันธ์คาโนนิคอลจะมีค่าอยู่ระหว่าง 0 และ 1 และจะไม่เปลี่ยนค่าแม้ว่าจะเปลี่ยนหน่วยการวัดก็ตาม

การคำนวณ เนื่องจากคุณสมบัติดังกล่าวข้างต้นของ Canonical Correlation Coefficient ดังนั้นเมื่อไม่ทราบค่า Σ เราจะประมาณ Σ ด้วย Covariance Matrix (S)

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

เมื่อ S เป็น Covariance Matrix ของตัวอย่างหรืออาจจะคำนวณได้จาก Correlation matrix (R)

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

เมื่อ R เป็น Correlation Matrix ของตัวอย่าง

R_{11} เป็น matrix ขนาด pxp ซึ่งแสดงความสัมพันธ์ระหว่างตัวแปร X ซึ่งมี p ตัว

R_{22} เป็น matrix ขนาด qxq ซึ่งแสดงความสัมพันธ์ระหว่างตัวแปร Y ซึ่งมี q ตัว

R_{12} เป็น matrix ขนาด pxq ซึ่งแสดงความสัมพันธ์ระหว่างตัวแปร X และตัวแปร Y

$$R_{21} = R'_{12}$$

ค่าของ Canonical Correlation Coefficients ที่คำนวณจาก S หรือ R นั้น จะได้ผลเช่นเดียวกัน แต่ค่า α_i และ β_i นั้นจะได้ไม่เหมือนกัน กล่าวคือ

ถ้าหากคำนวณ α_i และ β_i จาก S matrix แล้ว ค่าที่ได้นั้นจะเป็นค่าตัวแปรแรกเริ่ม (Original Variables) ถ้าหากคำนวณจาก R matrix ค่าที่ได้จะเปลี่ยนอยู่ในรูปตัวแปรแรกเริ่มมาตรฐาน (Standardized Original Variables) อย่างไรก็ตามก็สามารถที่จะเปลี่ยนค่าสัมประสิทธิ์ที่คำนวณได้จากการใช้ S นั้นไปเป็นค่าสัมประสิทธิ์ที่ได้จาก R ดังจะเห็นได้จาก

$$\begin{bmatrix} \frac{1}{\sqrt{s_{11}}} & 0 \\ 0 & \frac{1}{\sqrt{s_{22}}} \end{bmatrix} \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{s_{11}}} & 0 \\ 0 & \frac{1}{\sqrt{s_{22}}} \end{bmatrix} = \begin{bmatrix} 1 & R_{12} \\ R_{21} & 1 \end{bmatrix}$$

ดังนั้น เมื่อนำ diagonal matrix ของส่วนกลับของค่าส่วนเบี่ยงเบนมาตรฐานมาคูณทั้งหน้าและหลังของ Covariance Matrix แล้ว ก็จะสามารถหาค่าของ Correlation Matrix α , β ได้ กล่าวคือ

ถ้า α^*_{i1} , α^*_{i2} ,..., α^*_{ip} และ β^*_{i1} , β^*_{i2} ,..., β^*_{iq} ;

(i = 1,2,..., q)

เป็นค่าที่คำนวณได้จาก matrix S แล้วจะสามารถหาค่า α β ซึ่งคำนวณได้จาก R ได้ โดยการหารแต่ละค่าสัมประสิทธิ์เหล่านั้นด้วยค่าส่วนเบี่ยงเบนมาตรฐานของตัวแปรแรกเริ่ม (Original Variable) นั่นคือ

$$\frac{\alpha^*_{i1}}{s_{x_{11}}} , \frac{\alpha^*_{i2}}{s_{x_{22}}} , \frac{\alpha^*_{i3}}{s_{x_{35}}} ,..., \frac{\alpha^*_{ip}}{s_{x_{pp}}}$$

และ

$$\frac{\beta^*_{i1}}{s_{y_{11}}} , \frac{\beta^*_{i2}}{s_{y_{22}}} , \frac{\beta^*_{i3}}{s_{y_{33}}} ,..., \frac{\beta^*_{iq}}{y_{qq}}$$

<u>การทดสอบนัยสำคัญของการวิเคราะห์สหสัมพันธ์คาโนนิคอล</u> เมื่อทำการวิเคราะห์สหสัมพันธ์คาโนนิคอลแล้ว สิ่งที่จะได้คือ

(1) ρ_c

(2) α และ β

ค่าของ α และ β ซึ่งสอดคล้องกับ λ^2 ของเมตริก $\Sigma_{11}^{-1}\ \Sigma_{12}\ \Sigma_{22}^{-1}\ \Sigma_{21}$ โดยที่ λ^2 ก็คือความแปรปรวนที่ร่วมกันของตัวแปรทั้งสองชุดนั่นเอง ด้วยเหตุนี้ก็จะมี α และ β อยู่หลายชุดด้วยกัน ดังนั้นสิ่งที่จะต้องคำนึงถึงคือ Canonical Variables ทั้ง q คู่นั้นมีความสัมพันธ์กันหรือไม่ คือต้องมีการทดสอบสมมติฐานที่ว่า ตัวแปรทั้งสองชุดนั้นมีความสัมพันธ์กันหรือไม่ นั่นคือทดสอบว่า Σ_{12} หรือ $\Sigma_{21} = 0$ หรือไม่

$H_o : \Sigma_{12} = 0$ $\qquad$ $H_o : \Sigma_{21} = 0$

$H_a : \Sigma_{12} \neq 0$ $\qquad$ $H_a : \Sigma_{21} \neq 0$

ใช้วิธีการทดสอบของ Bartlett ดังนี้

วิธีทดสอบของ Bartlett ใช้ตัวสถิติ V_o โดยที่

$$V_o = -\left[n-1-\frac{1}{2}(p+q+1)\right]\ln \Lambda_o$$

เมื่อ $$\Lambda_o = \frac{1}{\prod_{i=1}^{q}\left[1+\frac{\lambda_i^2}{(1-\lambda_i^2)}\right]}$$

$$= \frac{1}{\prod_{i=1}^{q}\left[\frac{1}{(1-\lambda_i^2)}\right]}$$

$$= \prod_{i=1}^{q}(1-\lambda_i^2)$$

V_o จะมีการกระจายแบบไคล่แควร์ที่องค่าความเป็นอิล่ระเท่ากับ pq ที่ระดับนัยล่ำคัญที่ต้องการ ถ้าค่า V_o ที่คำนวณได้มากกว่าค่าไคล่แควร์จากตารางจะล่รุปได้ว่า $\rho_c^{(1)}$ ซึ่งเป็น first canonical correlation coefficient นั้นไม่เท่ากับ 0 และจะทดล่อบความสัมพันธ์ของ ρ_c ของคู่ที่เหลือต่อไปคือ

$$\Lambda_1 = \prod_{i=2}^{q} (1- \lambda_i^2)$$

ซึ่งจะนำไปสู่ Bartlett V_1 ดังนี้

$$V_1 = - \left[n-2- \frac{1}{2} (p+q+1)\right] \ln \Lambda_1$$

Λ_1 มีการกระจายแบบไคล่แควร์ที่องค่าความเป็นอิล่ระเท่ากับ (p-1)(q-1) การล่รุปผลจะทำเหมือน V_o ดังนั้น ถ้าหากมี k canonical correlation coefficients ที่ทดล่อบแล้วปรากฏว่ามีนัยล่ำคัญล่วนที่เหลือก็จะถูกทดล่อบว่ำ (q-k) Canonical Correlation Coefficient เหล่านั้นจะเท่ากับ 0 หรือไม่กล่าวคือ

$$\Lambda_k = \prod_{i=(k+1)}^{q} (1- \lambda_i^2)$$

$$V_k = - \left[n-1-k- \frac{1}{2} (p+q+1)\right] \ln \Lambda_k$$

V_k มีการกระจายแบบไคล่แควร์ที่มีองค่าความเป็นอิล่ระเท่ากับ (p-k)(q-k)

ในการประมวลผลด้วยโปรแกรมล่ำเร็จรูป SPSS นั้น ผลลัพธ์ที่ได้จะมีล่ำคัญหรือค่าของความน่าจะเป็นที่จะไม่ยอมรับ H_o เมื่อ H_o จริง ในกรณีที่จะทำล่อบล่มมติฐาน ณ ระดับนัยล่ำคัญ $\alpha = 0.05$ นั้น อาจใช้ค่าของความน่าจะเป็นที่จะไม่ยอมรับ H_o เมื่อ H_o จริง มาเปรียบเทียบกับ $\alpha = 0.05$ เพื่อใช้เป็นเกณฑ์ในการตัดสินใจว่าจะยอมรับล่มมติฐาน H_o หรือไม่โดยที่

ถ้าค่าความน่าจะเป็นที่จะไม่ยอมรับ H_o หรือ H_o จริง < 0.05 จะไม่ยอมรับ H_o

ถ้าค่าความน่าจะเป็นที่จะไม่ยอมรับ H_o เมื่อ H_o จริง > 0.005 จะยอมรับ H_o

3.5 วิธีวิเคราะห์จำแนกประเภท

เป็นวิธีวิเคราะห์ทางสถิติโดยมีจุดมุ่งหมายที่จะคัดเลือกตัวแปรชุดหนึ่ง ซึ่งนักวิจัยคิดว่าตัวแปรชุดนี้มีความสัมพันธ์กับสิ่งที่ต้องการศึกษา จนถึงขั้นที่ตัวแปรชุดนี้เป็นตัวแบ่งแยกประชากรออกเป็นกลุ่มต่าง ๆ ได้อย่างชัดเจน

3.5.1 ขั้นตอนในการวิเคราะห์จำแนกประเภท มี 2 ขั้นตอนคือ

ขั้นที่ 1 การคัดเลือกตัวแปรชุดหนึ่ง เพื่อสร้างสมการที่ใช้ในการจำแนกประชากรออกเป็นกลุ่มต่าง ๆ กัน ได้อย่างชัดเจน สมการนี้คือ สมการจำแนกประเภท (Discriminant equation)

ขั้นที่ 2 การจำแนกตัวอย่างที่ได้ศึกษามานั้นเข้าเป็นสมาชิกของประชากรแต่ละกลุ่มโดยอาศัยสมการจำแนกประเภท

ในกรณีที่มีประชากร 2 กลุ่ม จะใช้วิธีการของ Fisher มาใช้ในการจำแนกประเภท กล่าวคือ ถ้ากำหนดให้

Π_1 เป็นประชากรกลุ่มที่ 1

Π_2 เป็นประชากรกลุ่มที่ 2

ตัวแปรที่ศึกษา คือ ตัวแปร X ซึ่งมีทั้งหมด p ตัวคือ

$$X = \left[X_1 \quad X_2 \quad \ldots \quad X_p \right]$$

วิธีการของ Fisher นั้น จะแปลงตัวแปร X เหล่านี้ไปเป็นค่าของตัวแปรเพียงตัวเดียวคือ Y โดยที่ Y_1 และ Y_2 เป็นค่าสังเกตที่ได้จากประชากร Π_1 และ Π_2 ตามลำดับ

ถ้า μ_{1y} = ค่าเฉลี่ยของค่า Y ซึ่งได้มาจากค่า X ของประชากร Π_1

μ_{2y} = ค่าเฉลี่ยของค่า Y ซึ่งได้มาจากค่า X ของประชากร Π_2

$\left.\begin{array}{l} \mu_1 = E(X/\Pi_1) = \text{ค่าคาดหวังของตัวแปร X ที่มาจากประชากร } \Pi_1 \\ \mu_2 = E(X/\Pi_2) = \text{ค่าคาดหวังของประชากร X ที่มาจากประชากร } \Pi_2 \end{array}\right\} \ldots(1)$

และโควาเรียนซ์เมตริก

$$\Sigma = E(\underset{\sim}{X} - \underset{\sim}{\mu}_1)(\underset{\sim}{X} - \underset{\sim}{\mu}_i)' \; ; \; i = 1,2,\ldots\ldots\ldots (2)$$

พิจารณาผลรวมเชิงเส้น

$$\underset{(1x\ell)}{Y} = \underset{(1xp)}{\underset{\sim}{\beta}'} \underset{(px\ell)}{\underset{\sim}{X}} \qquad \ldots\ldots\ldots\ldots\ldots (3)$$

เมื่อ $\underset{\sim}{Y}$ คือ Fisher's Linear Discriminant Function (ผลรวมเชิงเส้น)

$\underset{\sim}{\beta}$ คือ ค่าที่แสดงถึงความสำคัญของตัวแปร $X_1, X_2, \ldots, X_p$

$$\underset{\sim}{\beta}' = \begin{bmatrix} \beta_1 & \beta_2 \ldots, \beta_p \end{bmatrix}$$

$\underset{\sim}{X}$ คือตัวแปรที่ต้องการศึกษา ซึ่งมีทั้งหมด p ตัว

และ $\underset{\sim}{X}' = \begin{bmatrix} X_1 & X_2 \ldots, X_p \end{bmatrix}$

ดังนั้น จะได้

$$\left.\begin{aligned} \mu_{1y} &= E(\underset{\sim}{Y}/\Pi_1) = E(\underset{\sim}{\beta}' \underset{\sim}{X}/\Pi_1) = \underset{\sim}{\beta}' \underset{\sim}{\mu}_1 \\ \underset{\sim}{\mu}_{2y} &= E(\underset{\sim}{Y}/\Pi_2) = E(\underset{\sim}{\beta}' \underset{\sim}{X}/\Pi_2) = \underset{\sim}{\beta}' \underset{\sim}{\mu}_2 \end{aligned}\right\} \ldots\ldots\ldots (4)$$

และค่าความแปรปรวนของ Y จากทั้ง 2 ประชากรนั้นคือ

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

$$= \underset{\sim}{\beta}' \Sigma \underset{\sim}{\beta} \qquad \ldots\ldots\ldots\ldots\ldots (5)$$

วิธีการของ Fisher คือพยายามหาผลรวมเชิงเส้นของค่า X ซึ่งจะทำให้ระยะทางระหว่าง μ_{1y} และ μ_{2y} มีค่ามากที่สุดที่จะเป็นไปได้

ค่าผลรวมเชิงเส้นซึ่งดีที่สุดจะต้องมีคุณสมบัติในการแบ่งแยกประชากรทั้ง 2 กลุ่มออกจากกันได้มากที่สุด ซึ่งวิธีการที่จะได้ผลรวมเชิงเส้นที่ดีนั้นก็โดยการหาค่า β ที่ทำให้อัตราส่วน

$\dfrac{(\text{ระยะทางระหว่างค่าเฉลี่ยของ } Y)^2}{\text{ความแปรปรวนของ } Y}$ มีค่ามากที่สุด

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

$$= \frac{(\mu_{1Y} - \mu_{2Y})^2}{\sigma_Y^2}$$

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

$$= \frac{\underset{\sim}{\beta}'(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\underset{\sim}{\beta}}{\underset{\sim}{\beta}'\Sigma\underset{\sim}{\beta}} \qquad \ldots\ldots\ldots\ldots\ldots(6)$$

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

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

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

$$\underset{\sim}{\beta} = C\Sigma^{-1}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)$$

กำหนดให้ $C = 1$

จะได้ผลรวมเชิงเส้นเป็น

$$Y = \underset{\sim}{\beta}'\underset{\sim}{X} = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{X} \qquad \ldots\ldots\ldots\ldots\ldots(7)$$

สมการ $Y = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{X}$ สามารถใช้เป็นตัวจำแนกค่าสังเกตที่ได้มานั้นว่าจะอยู่ในประชากรกลุ่ม Π_1 หรือ Π_2 ได้โดยให้

$$Y_o = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)\Sigma^{-1}\underset{\sim}{X}_o$$

เป็นค่าของ discriminant function ของค่าสังเกต $\underset{\sim}{X}_o$ และให้

$$m = \frac{1}{2}(\mu_{1y} + \mu_{2y})$$

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

$$= \frac{1}{2}(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1}(\underset{\sim}{\mu}_1 + \underset{\sim}{\mu}_2) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(8)$$

เป็นจุดกึ่งกลาง (centroid) ระหว่างค่าเฉลี่ยของค่า Y จากทั้ง 2 ประชากร

สามารถเขียนกฎการจำแนกประเภทได้ดังนี้คือ

$$\left.\begin{array}{l} \text{จัด } \underset{\sim}{X}_o \text{ ให้อยู่ } \Pi_1 \text{ ถ้าหากว่า } Y_o = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} \underset{\sim}{X}_o < m \\ \text{จัด } \underset{\sim}{X}_o \text{ ให้อยู่ } \Pi_2 \text{ ถ้าหากว่า } Y_o = (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \Sigma^{-1} \underset{\sim}{X}_o > m \end{array}\right\} \quad \ldots\ldots\ldots(9)$$

ในทางปฏิบัติจะไม่ทราบค่า $\underset{\sim}{\mu}_1$, $\underset{\sim}{\mu}_2$ และ Σ ดังนั้นถ้ากลุ่มตัวอย่างจาก Π_1 และ Π_2 มาเป็นจำนวน n_1 และ n_2 แล้วโดยที่วัดค่าสังเกต

$$\underset{\sim}{X}' = \begin{bmatrix} X_1 & X_2 & \ldots & X_p \end{bmatrix}$$ จะได้ว่า

$$\underset{\sim}{X}_1 = \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1n_1} \end{bmatrix}$$
$$(pxn_1)$$

$$\underset{\sim}{X}_2 = \begin{bmatrix} X_{21} & X_{22} & \ldots & X_{2n_2} \end{bmatrix}$$
$$(pxn_2)$$

จะได้ค่าประมาณของ $\underset{\sim}{\mu}_1$, $\underset{\sim}{\mu}_2$ และ Σ^{-1} เป็น $\bar{\underset{\sim}{X}}_1$, $\bar{\underset{\sim}{X}}_1$ และ S^{-1}_{pooled} ตามลำดับโดยที่

$$\underset{\sim}{\bar{X}}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} \underset{\sim}{X}_{ij} \quad \text{.....(10)} \; ; \; \underset{\sim}{\bar{X}}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} \underset{\sim}{X}_{2j} \quad \text{.....(11)}$$

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

$$= \frac{(n_1 - 1)\, S_1 + (n_2-1)\, S_2}{n_1+n_2-2} \quad \text{.....(12)}$$

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

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

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

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

$$= (\underset{\sim}{\bar{X}}_1 - \underset{\sim}{\bar{X}}_2)^1 \; S^{-1}_{pooled} \; \underset{\sim}{X} \quad \text{.....(13)}$$

และค่าจุดกึ่งกลางระหว่างตัวแปร $\underset{\sim}{\bar{Y}}_1 = \hat{\underset{\sim}{\beta}}'\underset{\sim}{\bar{X}}_1$ $\underset{\sim}{\bar{Y}}_2 = \hat{\underset{\sim}{\beta}}'\underset{\sim}{\bar{X}}_2$

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

$$= \frac{1}{2}(\underset{\sim}{\bar{X}}_1 - \underset{\sim}{\bar{X}}_2)' \; S^{-1}_{pooled} \; (\underset{\sim}{\bar{X}}_1 + \underset{\sim}{\bar{X}}_2) \quad \text{.....(14)}$$

กฎการจำแนกประเภทจะเขียนดังนี้

จัดให้ X_o อยู่ใน Π_1 ถ้า $Y_o < m$

จัดให้ X_o อยู่ใน Π_2 ถ้า $Y_o > m$

เมื่อ $Y_o = (\bar{X}_1 - \bar{X}_2)' S^{-1}_{pooled} X_o$

จากสมการสามารถเขียนได้เป็น

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

ค่า $\beta_1, \beta_2, \ldots, \beta_p$ จะเป็นค่าแสดงถึงความสำคัญของตัวแปร $X_1, X_2, \ldots, X_p$ ซึ่งโดยปกติ X_i แต่ละค่านั้นจะมีหน่วยไม่เหมือนกัน การเปรียบเทียบค่า β_i เหล่านี้จะทำได้โดยการปรับค่า β_i เหล่านี้ให้เป็นมาตรฐาน (Standardized) เสียก่อน

การแปลงค่า β_i เหล่านี้สามารถทำได้โดยเอาสมาชิกในแนวเส้นทแยงมุม (diagonal) ของเมตริก W มาถอดรากที่สอง แล้วนำไปคูณกับค่า β_i ทุก ๆ ค่าตามสูตรต่อไปนี้

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

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

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

กลุ่มที่ 1 | กลุ่มที่ 2

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

β_i^* คือ Standardized discriminant weight ของ discriminant function

β_i คือ Raw discriminant weight ของ discriminant function

W_{ii} คือ Diagonal element ของ W เมตริก โดยที่ $i = 1,2,\ldots,p$

สมการจำแนกประเภทที่ได้มานั้นจะมีความสามารถในการแบ่งแยกกลุ่มได้ดีหรือไม่นั้นสามารถดูได้จากค่าความผิดพลาด เนื่องจากการจำแนกผิด (miss classification error) ซึ่งทราบได้จากการที่เปรียบเทียบว่า สิ่งที่ศึกษาอยู่นั้นเป็นสมาชิกของประชากรกลุ่มหนึ่งซึ่งทำให้ได้ผลที่ผิดจากความเป็นจริงไป

3.5.2 <u>วิธีการคัดเลือกตัวแปรเข้าไปในสมการจำแนกประเภท</u>

จากการที่กล่าวมาแล้วข้างต้นนั้นเป็นหลักการโดยทั่ว ๆ ไปของวิธีการสร้างสมการจำแนกประเภทในกรณีที่หากว่าในการวิจัยนั้น มีตัวแปรเป็นจำนวนมาก การคัดเลือกตัวแปรให้เหลือจำนวนน้อยที่สุดแต่มีความสามารถในการใช้เป็นตัวจำแนกมากที่สุดนั้น สามารถทำได้โดยใช้การคัดเลือกตัวแปรทีละตัวโดยที่จะหาตัวแปรที่ดีที่สุดตัวแรกและตัวแปรที่ดีที่สุดที่สองที่จะทำให้สมการจำแนกประเภทมีความสามารถจำแนกประเภทได้ดีที่สุด จากนั้นก็จะเลือกตัวที่สามและตัวต่อไปที่จะช่วยการจำแนกให้ดีขึ้นตามลำดับ ในแต่ละขั้นตอนตัวแปรที่ได้รับการคัดเลือกมาก่อนแล้วนั้นอาจถูกตัดทิ้งไปหากพบว่าเมื่อนำมารวมกับตัวแปรอื่น ๆ แล้วไม่ช่วยให้สมการจำแนกประเภทได้ดีขึ้น วิธีการนี้เรียกว่า วิธีการสร้างสมการจำแนกประเภทแบบขั้นตอน (Stepwise discriminant analysis) ซึ่งในการวิจัยครั้งนี้ไม่ได้ใช้วิธีนี้ แต่จะกำหนดตัวแปรเข้าไปในสมการโดยใช้ตัวแปรชุดเดียวกับตัวแปรที่ใช้ในการวิเคราะห์การถดถอยเชิงซ้อน เนื่องจากต้องการจะเปรียบเทียบความเหมาะสมของแต่ละวิธีการวิเคราะห์ด้วยตัวแปรชุดเดียวกัน

จากวิธีการดังกล่าวข้างต้น จะทำให้ได้สมการจำแนกประเภทและค่าคาแรคเตอร์สติก รูท ซึ่งเป็นค่าที่ได้จากขั้นตอนการหาสมการวิเคราะห์จำแนกประเภทโดยเป็นค่าความแปรปรวนของค่า Y ซึ่งได้จากการแปลงรูป (Transform) จากค่า X ต่าง ๆ $[X_1 \quad X_2 \quad \ldots \quad X_p]$ เป็นค่าที่ใช้วัดความสำคัญเชิงเปรียบเทียบของสมการได้ โดยที่สมการวิเคราะห์จำแนกประเภทที่ได้มานั้นได้ตามลำดับความสำคัญของค่า คาแรคเตอรีสติค รูท ต่าง ๆ จากมากที่สุดและรองลงไปตามลำดับ ค่าผลรวมของ คาแรคเตอรีสติค รูท ทั้งหมดนั้นเป็นค่าความแปรปรวนทั้งหมดของตัวแปรจำแนกประเภท (ตัวแปร X) ค่าคาแรคเตอรีสติค รูท แต่ละค่าจึงคิดเป็นอัตราส่วนร้อยของค่ารวมของ คาแรคเตอรีสติค รูท ทั้งหมด ทำให้สามารถใช้ค่านี้ไปอ้างอิงถึงความสำคัญเชิงเปรียบเทียบของสมการจำแนกประเภทได้

3.5.3 <u>การทดสอบนัยสำคัญของสมการจำแนกประเภท</u>

เมื่อได้สมการจำแนกประเภทมาแล้วและอยากทราบว่าสมการนี้สามารถมีอำนาจในการจำแนกกลุ่มได้อย่างมีนัยสำคัญทางสถิติหรือไม่ เราสามารถทดสอบได้จาก

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

โดยที่ V_m คือ Bartlett V Statisticซึ่งมีการกระจายเป็นแบบ Chi-Square ที่มี d.f. = p+k-2m

N คือ จำนวนตัวอย่างทั้งหมดจากประชากรทั้ง 2 กลุ่ม

p คือ จำนวนตัวแปรทั้งหมด

k คือ จำนวนกลุ่ม

λ_m คือ คาแรคเตอรีสติค รูท ต่าง ๆ เช่น ค่าตัวที่ 1,2,3,...,m

ค่า λ_m นี้คำนวณได้จาก

$$|A - \lambda_m I| = 0$$

เมื่อ A คือโควาเรียนซ์เมตริกของ X

ค่า λ_m นี้จะเลือกเอาเฉพาะค่าที่ไม่เป็นศูนย์

ถ้าค่า V_m ที่คำนวณได้มากกว่าค่าไคสแควร์ที่ระดับนัยสำคัญที่กำหนดไว้แล้ว แสดงว่าสมการนี้สามารถใช้จำแนกกลุ่มได้อย่างมีนัยสำคัญทางสถิติ

การคำนวณค่าอำนาจการจำแนกของกลุ่มของตัวแปร (Total Discriminatory Power)

$$\hat{W}^2 = 1 - \frac{N}{(m-k)(1+\lambda_1) + (1+\lambda_2)\ldots(1+\lambda_m)+1}$$

เมื่อ $\hat{W}^2$ = ค่าอำนาจในการแยกตัวแปรได้จากการประมาณค่า

N = จำนวนตัวอย่างทั้งหมดจากประชากรทั้ง 2 กลุ่ม

k = จำนวนกลุ่ม

λ_m = ค่าคาแรคเตอรีสติค รูท ต่าง ๆ

m = จำนวนค่าคาแรคเตอรีสติค รูท

3.6 <u>การคัดเลือกแบบจำลองทางคณิตศาสตร์ที่เหมาะสมสำหรับการคัดเลือกพื้นที่เพื่อประกาศเขตปฏิรูปที่ดิน</u>

จากแบบจำลองทั้ง 4 ชุดที่ได้จากการวิเคราะห์การถดถอยเชิงซ้อน การวิเคราะห์องค์ประกอบหลัก การวิเคราะห์คาโนนิคอล และการวิเคราะห์จำแนกประเภท ใช้ข้อมูลจาก 37 อำเภอ ใน 5 จังหวัดที่คัดเลือกมาวิจัยครั้งนี้ นำค่าตัวแปรทั้งหมดของแต่ละอำเภอมาแทนลงในแบบจำลองที่ได้จากการวิเคราะห์ทั้ง 4 วิธีนี้จะได้คะแนนความสำคัญของแต่ละอำเภอ ทั้ง 37 อำเภอ ต่อจากนั้นทำการจัดลำดับของอำเภอที่สุ่มตัวอย่างมา 37 อำเภอนี้ตามคะแนนความสำคัญของแต่ละอำเภอ ก็จะได้ลำดับที่ความสำคัญของพื้นที่ (อำเภอ) ที่แสดงถึงระดับการพัฒนาแล้วทั้ง 4 ชุด

จากการศึกษาเรื่องเส้นวัดความยากจนของนักเศรษฐศาสตร์ และการศึกษาเกี่ยวกับความยากจนของประชากรในชนบทของประเทศไทยของธนาคารโลก ปรากฏว่าต่างก็ใช้รายได้ของประชากรมาเป็นตัวพิจารณา แสดงว่ารายได้เป็นตัวบ่งชี้ที่ดีในการวัดสภาพความเป็นอยู่ของประชากร แต่สิ่งที่ไม่สามารถบอกได้คือความสำคัญของปัญหาในแต่ละตัวที่ประชากรในชนบทกำลังเผชิญอยู่ ดังนั้นจึงต้องดำเนินการวิเคราะห์การถดถอยเชิงซ้อน การวิเคราะห์องค์ประกอบหลัก การวิเคราะห์คาโนนิคอล และการวิเคราะห์จำแนกประเภท ซึ่งได้กำหนดน้ำหนักความสำคัญของตัวแปรแต่ละตัวขึ้นมา ก็จะสามารถชี้ให้เห็นถึงความสัมพันธ์ของปัญหาในแต่ละด้าน ซึ่งจะมีประโยชน์ในการคัดเลือกพื้นที่เพื่อประกาศเขตปฏิรูปที่ดินต่อไป อย่างไรก็ตามแบบจำลองที่จะใช้ได้ดีก็ควรจะเป็นแบบจำลองที่เมื่อนำมาใช้แล้วให้อันดับที่ของอำเภอที่สอดคล้องกับลำดับที่ของอำเภอที่ใช้รายได้เฉลี่ยต่อครัวเรือนต่อปีของประชากรในแต่ละอำเภอ

3.6.1 <u>การวัดความสอดคล้องของตัวแปรประเภทจัดอันดับ</u>

ตัวสถิติที่ใช้สำหรับการหาความสัมพันธ์ระหว่างตัวแปรประเภทจัดอันดับมีหลายตัวด้วยกัน เช่น สัมประสิทธิ์สหสัมพันธ์เชิงอันดับของสเปียร์แมน (γ_s) และสัมประสิทธิ์ความสัมพันธ์ของครัสคาลและกูดแมน (G) และสัมประสิทธิ์สหสัมพันธ์เชิงอันดับของเคนดาลล์ (τ) ในที่นี้จะใช้สัมประสิทธิ์สหสัมพันธ์เชิงอันดับของสเปียร์แมน (γ_s) เพราะมีขั้นตอนของการคำนวณที่ง่ายกว่า ดังนี้

สัมประสิทธิ์สหสัมพันธ์เชิงอันดับของสเปียร์แมน (Spearman's Coefficient of Rank Correlation ; γ_s) เป็นวิธีการที่ใช้วัดความสัมพันธ์ระหว่างตัวแปรประเภทจัดอันดับ วิธีการก็คือ การแบ่งข้อมูลออกเป็นสองอนุกรม แล้วหาผลต่างระหว่างข้อมูลทั้งสองชุดนั้น แต่ละคู่ของอันดับเดียวกันแล้วยกกำลังสอง จากนั้นนำค่าที่ได้มารวมกัน เพื่อหาค่าสัมประสิทธิ์ที่ต้องการ ในทางปฏิบัติ การหาความสัมพันธ์อาจเป็นความสัมพันธ์ระหว่างตัวแปรประเภทเดียวกันของประชากรกลุ่มเดียวกัน แต่มีการจัดอันดับสองครั้ง หรือเป็นความสัมพันธ์ระหว่างอันดับของบุคคลเดียวกันบนตัวแปรสองตัว เป็นต้น

สัมประสิทธิ์สหสัมพันธ์เชิงอันดับของสเปียร์แมนคำนวณได้จากสูตร*

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

โดยที่ γ_s = ค่าสัมประสิทธิ์สหสัมพันธ์เชิงอันดับของสเปียร์แมน

d_i = ค่าของผลต่างระหว่างอันดับ

n = จำนวนตัวอย่าง

เป็นการจัดอันดับโดยไม่มีตำแหน่งซ้ำกัน

ในกรณีที่เป็นการจัดอันดับโดยมีตำแหน่งที่ซ้ำกัน ก็ใช้วิธีเฉลี่ย เช่น ถ้ามีอันดับที่ 5 ซ้ำกัน 2 ตัว แต่ละตัวจะเป็นอันดับ 4.5 แต่ถ้ามีซ้ำกัน 3 ตัว ทุกตัวจะได้อันดับเดียวกัน คือตัวกลาง เช่น ถ้าตัวที่ 6 7 8 มีค่าเท่ากัน ทุกตัวจะได้อันดับ 7 หมด และการใช้สูตรข้างต้นจำเป็นต้องมีการปรับปรุงเพื่อให้ได้ค่าที่ถูกต้องยิ่งขึ้นดังนี้

$$\gamma_s = \frac{\Sigma X_i^2 - \Sigma Y_i^2 - \Sigma d_i^2}{2\sqrt{\Sigma X_i^2 \, \Sigma Y_i^2}}$$

---

*โดยทั่ว ๆ ไปถ้าค่า γ_s เข้าใกล้ + 1 หรือ -1 ถือว่ามีความสอดคล้องหรือมีความผกผันกันอย่างมาก แต่ถ้าค่า γ_s เข้าใกล้ 0 แสดงว่าการจัดลำดับไม่มีความสอดคล้องหรือผกผันกันแต่อย่างไร และถ้าค่า γ_s มีค่าต่ำกว่า + 0.5 แสดงว่ามีความสอดคล้องกันบ้างหรือค่า γ_s มีค่าน้อยกว่า -0.5 ก็แสดงว่ามีความผกผันกันบ้าง

โดยที่ $\Sigma X_i^2 = \frac{n^3 - n}{12} - \Sigma T_x$

$\Sigma Y_i^2 = \frac{n^3 - n}{12} = \Sigma T_y$

$T = \frac{t^3 - t}{12}$

X_i คือ ตัวแปรตัวหนึ่ง

Y_i คือ ตัวแปรตัวหนึ่ง

T_x คือ จำนวนครั้งที่ซ้ำกันในตัวแปร X

T_y คือ จำนวนครั้งที่ซ้ำกันในตัวแปร Y

n คือ จำนวนข้อมูล

d_i คือ ผลต่างระหว่างอันดับ

t คือ จำนวนครั้งที่ซ้ำกันในอันดับเดียวกัน

การใช้สูตรปรับปรุงควรจะใช้ในกรณีที่เป็นการจัดอันดับโดยมีตำแหน่งที่ซ้ำกันมาก ๆ เนื่องจากจะทำให้ค่าของ r_s ที่ได้มีความถูกต้องยิ่งขึ้น