บทที่ 2

ทฤษฎีที่เกี่ยวข้องกับการวิจัย

การวิเคราะห์ความถดถอยพหุ คือ การวิเคราะห์ รูปแบบหรือสมการของความสัมพันธ์ระหว่างตัวแปรอิสระกับตัวแปรตาม โดยแสดงให้เห็นว่า ตัวแปรอิสระแต่ละตัวมีความสัมพันธ์กับตัวแปรตามหรือไม่มากน้อยเพียงใด และความสัมพันธ์นั้นเป็นไปในเชิงบวก หรือ เป็นไปในเชิงลบ เทคนิคการวิเคราะห์จะทดสอบว่าตัวแปรอิสระแต่ละตัวมีผลต่อตัวแปรตามหรือไม่และตัวแปรอิสระทุกตัวมีผลต่อตัวแปรตามมากน้อยเพียงใด ในการใช้เทคนิคดังกล่าว ผู้วิจัยจะต้องมีข้อสมมติว่า ตัวแปรอิสระแต่ละตัวเป็นอิสระซึ่งกันและกัน

เนื่องจาก การวิเคราะห์ความถดถอยพหุมีเป้าหมายที่จะศึกษาอิทธิพลของตัวแปรอิสระแต่ละตัวที่มีต่อตัวแปรตาม เมื่อพิจารณาร่วมกับตัวแปรอิสระอื่น ๆ ที่นำมาวิเคราะห์ โดยดูจากค่าสัมประสิทธิ์ความถดถอยพหุ และเพื่อใช้ในการประมาณค่าของตัวแปรตามเมื่อกำหนดให้ตัวแปรอิสระเปลี่ยนแปลงไป ความไม่คลาดเคลื่อนและความแม่นยำของค่าประมาณสัมประสิทธิ์ความถดถอยพหุเมื่อใช้วิธีกำลังสองน้อยที่สุด จึงเป็นสิ่งที่มีความสำคัญมาก

เมื่อตัวแปรอิสระมีพหุสัมพันธ์กันสูง ซึ่งดูได้จากค่าสัมประสิทธิ์ความสัมพันธ์ระหว่างตัวแปรอิสระแต่ละคู่ จะทำให้ค่าประมาณสัมประสิทธิ์ความถดถอยที่ได้จากวิธีกำลังสองน้อยที่สุด ขาดความแม่นตรง การหลีกเลี่ยงปัญหาการเกิดพหุสัมพันธ์ในตัวแปรอิสระโดยการตัดตัวแปรใดตัวแปรหนึ่งออกไป อาจทำให้ค่าประมาณสัมประสิทธิ์ความถดถอยพหุที่ได้คลาดเคลื่อนไปและความคลาดเคลื่อนนี้จะมากหากตัวแปรที่ถูกตัดทิ้งไปมีความสัมพันธ์สูงกับตัวแปรอื่น ๆ ที่มีอยู่ในสมการ

โดยสรุปเทคนิคการวิเคราะห์ความถดถอยพหุ เป็นเทคนิคที่มีประโยชน์ในการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระหลาย ๆ ตัวโดยค่าสัมประสิทธิ์ความถดถอยพหุที่ได้จะบอกให้ทราบว่าตัวแปรอิสระแต่ละตัวมีผลต่อตัวแปรตามมากน้อยเพียงใด ทั้งนี้ต้องคำนึงถึงตัวแปรอิสระตัวอื่น ๆ ด้วย และค่าความผิดพลาดมาตรฐานของค่าสัมประสิทธิ์ความถดถอย จะบอกให้ทราบว่าค่าสัมประสิทธิ์ความถดถอยที่ได้นั้นมีค่าแตกต่างจากศูนย์อย่างมีนัยสำคัญจริงหรือไม่

หากผู้วิจัยจะเปรียบเทียบอิทธิพลของตัวแปรอิสระแต่ละตัวก็จะดูได้จากค่าสัมประสิทธิ์ความถดถอยมาตรฐาน (Standard regression coefficient) วิธีที่นิยมใช้ในการแปลงจากสมการความถดถอยพหุโดยทั่วไป เป็นสมการมาตรฐานมี 2 วิธีคือ

<u>**วิธีที่**</u> 1 Unit normal scaling

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \quad ; \quad i = 1, 2, \ldots\ldots, n \quad j = 1, 2, \ldots\ldots, p$$

$$y_i = \frac{y_i - \bar{y}}{s_y} \quad ; \quad i = 1, 2, \ldots\ldots, n$$

โดยที่ x_{ij} คือค่าของตัวแปรอิสระตัวที่ j จากตัวอย่างชุดที่ i

y_i คือค่าของตัวแปรตามจากตัวอย่างชุดที่ i

$$s_j^2 = \frac{\sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}{n-1}$$ เป็นความแปรปรวนของตัวแปร x_j

$$s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}$$ เป็นความแปรปรวนของตัวแปรตาม

ตัวแปรอิสระและตัวแปรตามที่ถูกแปลงค่าแล้วจะมีค่าเฉลี่ยเป็นศูนย์และความแปรปรวนเป็นหนึ่ง เมื่อใช้ตัวแปรใหม่เหล่านี้ จะได้ตัวแบบความถดถอยพหุ เป็น

$$y_i = b_1 z_{i1} + b_2 z_{i2} + \ldots + b_p z_{ip} + \varepsilon_i$$

หากประมาณค่าสัมประสิทธิ์ความถดถอยพหุด้วยวิธีกำลังสองน้อยที่สุด จะได้

$$\hat{b} = (Z'Z)^{-1} Z'Y$$

<u>วิธีที่</u> 2 Unit length scaling

$$w_{ij} = \frac{x_{ij} - \bar{x}_j}{s_{jj}^{\frac{1}{2}}} \quad ; \quad i = 1, 2, \ldots\ldots, n \qquad j = 1, 2, \ldots\ldots, p$$

$$y_i = \frac{y_i - \bar{y}}{s_{yy}^{\frac{1}{2}}} \quad ; \quad i = 1, 2, \ldots\ldots, n$$

โดยที่ x_{ij} คือค่าของตัวแปรอิสระตังที่ j จากตัวอย่างชุดที่ i

y_i คือค่าของตัวแปรตามจากตัวอย่างชุดที่ i

$$s_{jj} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

ในการแปลงค่าตัวแปรแบบนี้ จะได้ค่าตัวแปรอิสระใหม่ w_j มีค่าเฉลี่ยเท่ากับศูนย์ และ $\sqrt{\sum_{i=1}^{n} (w_{ij} - \bar{w}_i)^2} = 1$ เมื่อเขียนสมการความถดถอยพหุ ในเทอมของตัวแปรใหม่เหล่านี้ จะได้

$$y_i = b_1 w_{i1} + b_2 w_{i2} + \ldots + b_p w_{ip} + \varepsilon_i \; ; \quad i = 1,2,\ldots\ldots, n$$

หากประมาณค่าสัมประสิทธิ์ความถดถอยพหุด้วยวิธีกำลังสองน้อยที่สุด จะได้ $\hat{b} = (W'W)^{-1} W'Y$ ในการใช้ Unit length scaling จะได้เมตริกซ์ความสัมพันธ์ $W'W$ ดังนี้

$$w'w = \begin{bmatrix} 1 & r_{12} & r_{13} & r_{14} \cdots\cdots r_{1p} \\ r_{12} & 1 & r_{23} & r_{24} \cdots\cdots r_{2p} \\ r_{13} & r_{23} & 1 & r_{34} \cdots\cdots r_{3p} \\ \vdots & & & \vdots \\ r_{1p} & r_{2p} & & 1 \end{bmatrix}$$

โดยที่ $r_{ij} = \dfrac{\sum_{k=1}^{n}(x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{(s_{ii}\, s_{jj})^{\frac{1}{2}}}$

$= \dfrac{s_{ij}}{(s_{ii}\, s_{jj})^{\frac{1}{2}}}$ เป็นสหสัมพันธ์ระหว่างตัวแปรอิสระ x_i และ x_j

ในทำนองเดียวกัน

$$w'y = \begin{bmatrix} r_{1y} \\ r_{2y} \\ \vdots \\ r_{ky} \end{bmatrix}$$

โดยที่ $r_{jy} = \dfrac{\sum_{u=1}^{n} (x_{uj} - \bar{x}_j)(y_u - \bar{y})}{(s_{jj}\, s_{yy})^{\frac{1}{2}}}$

$= \dfrac{s_{jy}}{(s_{jj}\, s_{yy})^{\frac{1}{2}}}$ เป็นสหสัมพันธ์ระหว่างตัวแปรอิสระ x_j และตัวแปรตาม y

ถ้าใช้ Unit normal scaling จะได้ $z'z = (n-1)\, w'w$ ค่าสัมประสิทธิ์ความถดถอยที่ได้จากตัวแปรเหล่านี้จะเรียกว่า สัมประสิทธิ์ความถดถอยมาตรฐาน

ตัวแบบทั่วไปของความสัมพันธ์ระหว่างตัวแปรอิสระและตัวแปรตามเป็นลักษณะเชิงเส้น จะมีลักษณะดังนี้

$$Y = X\beta + \varepsilon \quad ; \quad \varepsilon \sim N(0\ ,\ \sigma^2 I) \qquad \text{..............(2.1)}$$

โดยที่ Y คือ เมตริกซ์ของตัวแปรตามขนาด $n \times 1$

X คือ เมตริกซ์ของตัวแปรอิสระขนาด $n \times p$

β คือ เมตริกซ์ของสัมประสิทธิ์ความถดถอยพหุขนาด $p \times 1$

ε คือ เมตริกซ์ของความคลาดเคลื่อนขนาด $n \times 1$

n คือ ขนาดของตัวอย่าง

p คือ จำนวนตัวแปรอิสระ

จากตัวแบบ (2.1) ผลรวมความคลาดเคลื่อนกำลังสองจะมีค่าเป็น

$$\begin{aligned} s(\beta) = \varepsilon'\varepsilon &= (Y - X\beta)'(Y - X\beta) \\ &= Y'Y - \beta'X'Y - Y'X\beta + \beta'X'X\beta \\ &= Y'Y - 2\beta'X'Y + \beta'X'X\beta \qquad \text{..................(2.2)} \end{aligned}$$

ในการประมาณค่าสัมประสิทธิ์ความถดถอยพหุ จากตัวแบบ (2.1) วิธีที่นิยมใช้กันมากที่สุด คือวิธีกำลังสองน้อยที่สุด ซึ่งจะได้ค่าประมาณสัมประสิทธิ์ความถดถอยพหุจากการดิฟเฟอเรนชิเอท (Differentiate) สมการ (2.2) เทียบกับ β แล้วให้เท่ากับศูนย์

$$\therefore \quad \left.\frac{\partial s}{\partial \beta}\right|_{\hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0$$

ดังนั้น $X'X\hat{\beta} = X'Y$ ซึ่งจะได้ $\hat{\beta} = (X'X)^{-1}X'Y$ เป็นตัวประมาณที่ไม่เอนเอียง และให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองน้อยที่สุด ในบรรดาตัวประมาณที่ไม่เอนเอียงทั้งหลาย แต่ในการประมาณค่าสัมประสิทธิ์ความถดถอยพหุด้วยวิธีกำลังสองน้อยที่สุดนี้ มีข้อสมมติฐานที่จำเป็นข้อหนึ่งคือ ตัวแปรอิสระต้องไม่มีความสัมพันธ์กันในลักษณะเชิงเส้น ซึ่งในทางปฏิบัติเป็นไปได้น้อยมาก เพราะตัวแปรต่าง ๆ ที่นำมาศึกษาอาจมีความสัมพันธ์กัน ในกรณีที่ตัวแปรอิสระในตัวแบบมีสภาพไม่เหมาะสม (Ill-condition) เช่นนี้ การประมาณค่าสัมประสิทธิ์ความถดถอยพหุด้วยวิธีกำลังสองน้อยที่สุด จะไม่ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำสุด

สุชาติ ประสิทธิ์รัฐสินธุ์ และ ลัดดาวัลย์ รอดมณี (2527 , 23-26) ได้กล่าวถึงปัญหาและการแก้ไขในการใช้วิธีกำลังสองน้อยที่สุดไว้ว่า เมื่อตัวแปรอิสระมีพหุสัมพันธ์กันสูง จะทำให้ค่าสัมประสิทธิ์ที่ได้จากการประมาณขาดความแม่นตรง ทางออกในการแก้ไขมีอย่างน้อย 2 วิธีคือ

1. ผู้วิจัยต้องทบทวนเกี่ยวกับตัวแปรต่าง ๆ ที่มีอยู่ในสมการเสียใหม่ว่ามีการใช้ตัวแปรหลายตัวในเรื่องเดียวกันหรือไม่ โดยเฉพาะอย่างยิ่งการใช้ตัวแปรแทน (proxy variable) ซึ่งหมายถึงตัวแปรที่ผู้วิจัยคิดว่ามีความสัมพันธ์กับตัวแปรตัวใดตัวหนึ่งมาก แต่ขาดข้อมูลเกี่ยวกับตัวแปรตัวนั้น จึงนำตัวแปรอื่นมาใช้แทนที่มีความหมายเชิงทฤษฎีเหมือนกัน เช่น การศึกษาของสามีแทนความสามารถในการหารายได้ และใช้การศึกษาของภรรยาแทนความสามารถในการรับรู้การเปลี่ยนแปลงทางวัฒนธรรม ในกรณีเช่นนี้ ตัวแปรแทนทั้ง 2 ตัวแม้จะเป็นคนละเรื่องกัน แต่ก็มีความสัมพันธ์กันสูง เพราะในทางปฏิบัติ คู่สมรสมักจะเลือกคู่ หรือได้คู่สมรสที่มีระดับการศึกษาใกล้เคียงกัน การที่จะหลีกเลี่ยงปัญหาการเกิดพหุสัมพันธ์ของตัวแปรอิสระ โดยการตัดตัวแปรตัวใดตัวหนึ่งใน 2 ตัวนี้ทิ้งไปจะมีผลทำให้ค่าสัมประสิทธิ์ของตัวแปรต่าง ๆ ในสมการคลาดเคลื่อนไป ผู้วิจัยจึงควรทบทวนตัวแปรแทนอื่น ๆ ที่ดีกว่า โดยมีเหตุผลสนับสนุนที่เพียงพอมาใช้แทน สำหรับกรณีที่ตัวแปรแทน 2 ตัว อาจมีความหมายเชิงทฤษฎีอย่างเดียวกัน ในกรณีเช่นนี้ผู้วิจัยอาจจะนำตัวแปรทั้ง 2 มารวมกัน (โดยการหาค่าเฉลี่ยหรือวิธีการอื่น ๆ) หรือตัดตัวแปรตัวใดตัวหนึ่งไป แต่ทั้งนี้จะต้องระลึกอยู่เสมอว่าการหลีกเลี่ยงปัญหาการเกิดพหุสัมพันธ์ของตัวแปรอิสระโดยการตัดตัวแปรตัวใดตัวหนึ่งทิ้งไปอาจทำให้ค่าสัมประสิทธิ์ความสัมพันธ์ของตัวแปรต่าง ๆ ในสมการนั้น ๆ คลาดเคลื่อนไปและความคลาดเคลื่อนนี้จะมากหากตัวแปรที่ถูกตัดทิ้งไปมีความสัมพันธ์สูงกับตัวแปรอื่น ๆ ที่มีอยู่ในสมการ

2. การแก้ปัญหาความไม่แม่นตรงของค่าสัมประสิทธิ์ความถดถอยพหุ อาจทำได้โดยผู้วิจัยเก็บตัวอย่างให้มากขึ้น

2.1 <u>คุณสมบัติของค่าประมาณสัมประสิทธิ์ความถดถอยพหุ (β) ด้วยวิธีกำลังสองน้อยที่สุด</u>

ถ้าประมาณค่าสัมประสิทธิ์ความถดถอยพหุด้วยวิธีกำลังสองน้อยที่สุด จะได้ค่าประมาณ $\hat{\beta} = (X'X)^{-1}X'Y$; ซึ่ง $\hat{\beta}$ จะมีค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำสุด ในบรรดาตัวประมาณค่า $\hat{\beta}$ ที่ไม่เอนเอียงทั้งหลาย

จากตัวแบบทั่วไปของการวิเคราะห์ความถดถอยพหุ $Y = X\beta + \varepsilon$ อาจเขียนค่าความคลาดเคลื่อนในรูปของฟังก์ชั่น X และ Y ได้ดังนี้

$$\varepsilon = Y - X\hat{\beta}$$

$$\therefore \quad \varepsilon'\varepsilon = (Y - X\hat{\beta})'(Y - X\hat{\beta})$$

ถ้าตัวแปรอิสระของข้อมูลตัวอย่างที่นำมาศึกษาการวิเคราะห์ความถดถอยพหุมีพหุสัมพันธ์กัน การที่จะประมาณค่าสัมประสิทธิ์ความถดถอยพหุด้วยวิธีกำลังสองน้อยที่สุดอาจได้ เมตริกซ์ความแปรปรวนของตัวประมาณค่าสัมประสิทธิ์ความถดถอยพหุ ไม่มีค่าต่ำสุด ดังนั้นจึงควรพิจารณาคุณสมบัติของตัวประมาณค่าสัมประสิทธิ์ความถดถอยพหุด้วยวิธีกำลังสองน้อยที่สุด 2 ประการคือ

1. เมตริกซ์ความแปรปรวนของตัวประมาณค่าสัมประสิทธิ์ความถดถอยพหุด้วยวิธีกำลังสองน้อยที่สุด

2. ค่าเฉลี่ยกำลังสองของความแตกต่าง $\hat{\beta}$ และ β ซึ่งสามารถเขียนให้อยู่ในรูปของ $X'X$ และ σ^2

$$\mathrm{Var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

ให้ L_1 คือความแตกต่างของ $\hat{\beta}$ และ β ดังนั้น

$$L_1^2 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta)$$

ค่าเฉลี่ยกำลังสองของความแตกต่าง ระหว่าง $\hat{\beta}$ และ β มีค่าเป็น

$$E\left[L_1^2\right] = \sigma^2 \text{ Trace } (X'X)^{-1}$$

$$E\left[L_1^2\right] = E\left[(\hat{\beta}-\beta)' (\hat{\beta}-\beta)\right]$$

$$= E\left[\hat{\beta}'\beta - 2\hat{\beta}'\beta + \beta'\beta\right]$$

$$= E(\hat{\beta}'\hat{\beta}) - \beta'\beta$$

ดังนั้น $$E(\hat{\beta}'\hat{\beta}) = \beta'\beta + \sigma^2 \text{Trace } (X'X)^{-1}$$

เมื่อ ε มีการกระจายแบบปกติ จะได้

$$\text{Var } (L_1^2) = 2\,\sigma^4 \text{Trace } (X'X)^{-2}$$

ให้ λ_i เป็น eigenvalue ของ $X'X$ เมื่อ $i = 1, 2, \ldots\ldots, p$; p คือจำนวนตัวแปรอิสระ (ขนาดของเมตริกซ์ $X'X$)

$$\sum_{i=1}^{p} \lambda_i = \text{Trace } (X'X)$$

$$(\lambda_{max} = \lambda_1) \geqslant \lambda_2 \geqslant \lambda_3 \geqslant \ldots\ldots\ldots \geqslant (\lambda_p = \lambda_{min}) \geqslant 0$$

$$E(L_1^2) = \sigma^2 \sum_{i=1}^{p} (1/\lambda_i)$$

$$\text{Var}(L_1^2) = 2\sigma^4 \sum_{i=1}^{p} (1/\lambda_i)^2$$

ดังนั้น ในกรณีที่ตัวแปรอิสระมีสภาพไม่เหมาะสม eigenvalue บางค่าของเมตริกซ์ $X'X$ จะมีค่าน้อยมาก ๆ ซึ่งจะมีผลทำให้ความแตกต่างระหว่าง $\hat{\beta}$ และ β มีค่ามาก

2.2 การประมาณค่าสัมประสิทธิ์ความถดถอยพหุ โดยวิธี ริดจ์รีเกรสชั่น

Hoerl and Kennard (1970 : 55-67) ได้เสนอวิธี ริดจ์ รีเกรสชั่น โดยเขากล่าวว่า วิธีนี้จะให้ค่าประมาณสัมประสิทธิ์ความถดถอยพหุที่ให้ค่า เฉลี่ยความคลาดเคลื่อนกำลังสองต่ำสุด ซึ่งวิธีนี้จะบวกค่าคงที่มากกว่าศูนย์กับสมาชิกทุกตัว บนเส้นทะแยงมุมของ เมตริกซ์ $X'X$ เนื่องจากการประมาณค่าสัมประสิทธิ์ความถดถอยพหุด้วยวิธีกำลังสองน้อยที่สุดจะได้ค่า เฉลี่ยความคลาดเคลื่อนกำลังสองมีค่า เท่ากับ $\sigma^2(X'X)^{-1}$ ด้วย เหตุผลที่ว่าค่า เฉลี่ยความคลาดเคลื่อนกำลังสอง เป็นฟังก์ชั่นของ $(X'X)^{-1}$ ฉะนั้นการที่จะลดค่า เฉลี่ยความคลาดเคลื่อนกำลังสองให้ต่ำลงจึงต้องมุ่งไปที่การลดค่า $(X'X)^{-1}$ ให้ต่ำลง ซึ่งจะทำได้โดยการบวกค่าคงที่มากศูนย์กับสมาชิกทุกตัวบนเส้นทะแยงมุมของเมตริกซ์ $(X'X)$ ดังนั้นจะได้ค่าประมาณสัมประสิทธิ์ความถดถอยพหุด้วยวิธี ริดจ์ รีเกรสชั่น คือ

$$\hat{\beta}_R = (X'X + kI)^{-1} X'Y \quad ; k > 0 \text{ ....} \quad (2.2.1)$$

ในระยะแรกที่ Hoerl and Kennard ได้เสนอการประมาณค่าสัมประสิทธิ์ความถดถอยพหุด้วยวิธี ริดจ์ รีเกรสชั่นนั้น ยังไม่สามารถหาค่า k ที่แน่นอนได้ นอกจากจะต้องมีการทดลองให้ k มีค่าเพิ่มขึ้น จากศูนย์ทีละน้อยและแทนค่า k ลงในสมการ (2.2.1)จนกว่าจะได้ค่า k ที่เหมาะสม และทุกครั้งที่กำหนดค่า k เพื่อคำนวณค่า $\hat{\beta}_R$ จะนำค่า เฉลี่ยความคลาดเคลื่อนกำลังสองของค่าประมาณที่ได้จากวิธี ริดจ์ รีเกรสชั่น และค่า เฉลี่ยความคลาดเคลื่อนกำลังสองของค่าประมาณที่ได้จากวิธีกำลังสองน้อยที่สุด มาเปรียบเทียบกัน ซึ่งจะทำเช่นนี้ไปเรื่อย ๆ จนกว่าค่า k ใดที่จะทำให้ค่า เฉลี่ยความคลาดเคลื่อนกำลังสองของค่าประมาณ $\hat{\beta}_R$ มีค่าน้อยกว่าค่า เฉลี่ยความคลาดเคลื่อนกำลังสองของค่าประมาณ β และจะถือว่าค่า k นั้นเริ่มจะเป็นค่า k ที่เหมาะสม จากนั้นจะค่อย ๆ เพิ่มค่า k ไปทีละน้อยอีกจนกว่าจะได้ k ที่เหมาะสมที่สุด ซึ่งจะได้ค่า เฉลี่ยความคลาดเคลื่อนกำลังสองของค่าประมาณสัมประสิทธิ์ความถดถอยพหุมีค่าต่ำสุด จะเห็นว่าจุดอ่อนของริดจ์ รีเกรสชั่น ก็คือ ไม่สามารถกำหนดค่า k ที่แน่นอนได้ แต่ต่อมาได้มีผู้พยายามจะประมาณค่า k ให้เหมาะสมยิ่งขึ้น ซึ่งวิธีนี้ก็ต้องเอาค่า k ไปปรับอีกเช่นกันเพียงแต่ต่างกันที่ค่าเริ่มต้นของ k เท่านั้น ซึ่งมีผู้เสนอวิธีการหาไว้หลายวิธี แต่ในการวิจัยเรื่องนี้ จะทำการศึกษา ริดจ์ รีเกรสชั่น โดยการหาค่าเริ่มต้นของ k ด้วยวิธีของ Hoerl, Kennard and Baldwin (1975: 105-123) ซึ่งกล่าวไว้ว่า ค่า k ที่เหมาะสมจะเป็น

$$k = \frac{p\,\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2.2.2)$$

โดยที่ p คือจำนวนตัวแปรอิสระ $\hat{\beta}$ และ $\hat{\sigma}^2$ เป็นค่าประมาณซึ่งได้จากวิธีกำลังสองน้อยที่สุด เขาได้แสดงให้เห็นว่าค่าเฉลี่ยคลาดเคลื่อนกำลังสองของ วิธี ริดจ์ รีเกรสชั่น ที่ลดลงจากวิธีกำลังสองน้อยที่สุด จะมีนัยสำคัญ ต่อมาในปี ค.ศ. 1976 Hoerl and Kennard ได้ปรับปรุงวิธีดังกล่าว โดยการนำค่า k ที่ได้จากสมการ (2.2.2) เป็นค่าเริ่มต้น จากนั้นก็จะแทนค่ากลับ (Iteration) เพื่อหาค่า $\hat{\beta}_R$

$$\hat{\beta} \qquad k_o = \frac{p\,\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$$

$$\hat{\beta}_R(k_o) \qquad k_1 = \frac{p\,\hat{\sigma}^2}{\hat{\beta}_R(k_o)'\,\hat{\beta}_R(k_o)}$$

$$\hat{\beta}_R(k_1) \qquad k_2 = \frac{p\,\sigma^2}{\hat{\beta}_R'(k_1)\,\hat{\beta}_R(k_1)}$$

$$\vdots \qquad \vdots$$

จะหยุดแทนค่า ถ้า $\frac{k_{j+1} - k_j}{k_j} < 20T^{-1.3}$,โดยที่ $T = \frac{\text{Trace }(x'x')^{-1}}{p}$

เมื่อตัวแปรอิสระมีพหุสัมพันธ์กันมาก ค่า T ก็จะมีค่าเพิ่มขึ้น ซึ่งทำให้ $20T^{-1.3}$ มีค่าน้อยลง โอกาสที่จะหยุดแทนค่าก็จะช้าขึ้น

Hoerl and Kennard ได้ใช้วิธีการทางคณิตศาสตร์มาควบคุมค่าความแปรปรวนของตัวประมาณค่าสัมประสิทธิ์ความถดถอยพหุ โดยวิธีกำลังสองน้อยที่สุดให้มีค่าต่ำลงได้ดังนี้

$$\hat{\beta}_R = (X'X + kI)^{-1} X'Y \qquad \ldots\ldots\ldots\ldots\ldots (2.2.3)$$

ให้ $$W = (X'X + kI)^{-1}$$

$$\therefore \quad \hat{\beta}_R = W X'Y \qquad \ldots\ldots\ldots\ldots\ldots (2.2.4)$$

ตัวประมาณค่าสัมประสิทธิ์ความถดถอยพหุ โดยวิธี ริดจ์ รีเกรสชั่น สามารถเขียนให้อยู่ในรูปฟังก์ชั่นของตัวประมาณ ค่าสัมประสิทธิ์ความถดถอยพหุ โดยวิธีกำลังสองน้อยที่สุด $(\hat{\beta})$ เพื่อความสะดวกในการเปรียบเทียบวิธีการประมาณค่าสัมประสิทธิ์ความถดถอยพหุที่คำนวณได้จากสองวิธีนี้

จากสมการ (2.2.4) อาจเขียนได้ว่า

$$\hat{\beta}_R = \left[I + k\,(X'X)^{-1}\right]^{-1} \hat{\beta} = z\,\hat{\beta}$$

โดยที่ $Z = \left[I + k\,(X'X)^{2}\right]^{-1}$ และ $W = \left[X'X + kI\right]^{-1}$

ได้ $E_i(W)$ และ $e_i(Z)$ เป็น eigenvalue ของ W และ Z ตามลำดับ ซึ่งได้จากการแก้สมการ Characteristic equation

$$|W - E_i I| = 0 \quad \text{และ} \quad |Z - E_i I| = 0$$

$$E_i(Z) = 1/(\lambda_i + k).$$

$$E_i(W) = 1/(\lambda_i + k)$$

อาจเขียน Z ให้อยู่ในรูปฟังก์ชั่นของ kW ได้ดังนี้

$$\begin{aligned} z &= I - k\,(X'X + kI)^{-1} \\ &= I - k\,W \end{aligned}$$

ค่าประมาณ $\hat{\beta}_R$ จะมีค่าน้อยกว่า β เมื่อ $k > 0$

$$\hat{\beta}_R' \beta_R \quad < \quad \hat{\beta}' \beta$$

<u>พิสูจน์จากนิยาม</u> : $\hat{\beta}_R = Z\hat{\beta}$; $X'X$ และ Z มีคุณสมบัติเป็น Symmetric positive definite จะนั้นจะได้

$$\hat{\beta}_R' \beta_R = (Z\hat{\beta})' (Z\hat{\beta})$$

$$= \sum_{i=1}^{p} E_i^2 (Z) \hat{\beta}' \hat{\beta}$$

$$< E_{i(max)}^2 (Z) \hat{\beta}' \beta$$

$E_{max}(z) = \lambda_1 / (\lambda_1 + k)$; λ_1 เป็น eigenvalueที่มีค่ามากที่สุดจะของเมตริกซ์ $X'X$ดังนั้น $\hat{\beta}_R' \hat{\beta}_R < \hat{\beta}' \hat{\beta}$

ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของตัวประมาณค่าสัมประสิทธิ์ความถดถอยพหุ โดยวิธีริดจ์ รีเกรสชั่นจะมีค่าเป็น

$$\text{MSE} (\hat{\beta}_R) = E\left[(\hat{\beta}_R - \beta)^2\right]$$

$$= \text{Var} (\hat{\beta}_R) + \left[E(\hat{\beta}_R) - \beta\right]^2 \quad \ldots\ldots\ldots (2.2.5)$$

การที่ยอมให้เกิดความเอนเอียงขึ้นใน $\hat{\beta}_R$ จะทำให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของ$\hat{\beta}_R$ มีค่าน้อยกว่าความแปรปรวนของตัวประมาณค่าสัมประสิทธิ์ความถดถอยพหุที่ไม่เอนเอียง สิ่งซึ่งส่งผลตามมาก็คือช่วงความเชื่อมั่นของตัวประมาณ β_Rจะแคบกว่า ช่วงความเชื่อมั่นของตัวประมาณ $\hat{\beta}$ แสดงว่า $\hat{\beta}_R$ เป็นตัวประมาณที่ค่อนข้างคงที่ (Stable) มากกว่า $\hat{\beta}$

จากสมการ (2.2.5) อาจเขียนได้ว่า

$$\text{MSE}\ (\hat{\beta}_R) = \sigma^2 \text{ Trace } (X'X + kI)^{-1} X'X (X'X + kI)^{-1} + k^2\beta'(X'X + kI)^{-2}\beta$$

$$= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} + k^2\beta'(X'X + kI)^{-2}\beta$$

ในการประมาณค่าสัมประสิทธิ์ความถดถอยพหุด้วย วิธี ริดจ์ รีเกรสชั่น จะต้องเลือกค่า k ซึ่งให้ค่าความแปรปรวนของตัวประมาณค่าสัมประสิทธิ์ความถดถอยพหุลดลงมากกว่าการเพิ่มขึ้นของความเอนเอียงยกกำลังสอง ถ้าทำได้เช่นนี้จะทำให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของตัวประมาณค่าสัมประสิทธิ์ความถดถอยพหุโดยวิธี ริดจ์ รีเกรสชั่น มีค่าน้อยกว่า ความแปรปรวนของตัวประมาณค่าสัมประสิทธิ์ความถดถอยพหุ โดยวิธีกำลังสองน้อยที่สุด

ค่าความคลาดเคลื่อนกำลังสองของตัวประมาณค่าตัวแปรตาม จะมีค่าเป็น

$$\text{SSE} = (Y - X\hat{\beta}_R)'(Y - X\hat{\beta}_R)$$

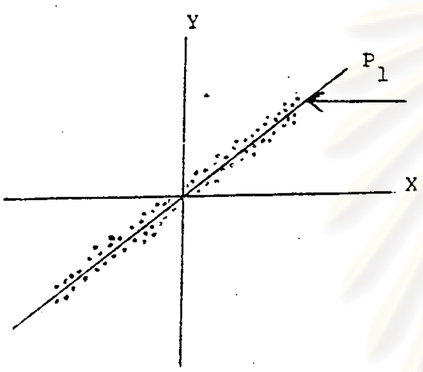
$$= (Y - X\hat{\beta})'(Y - X\hat{\beta}) + (\hat{\beta}_R - \hat{\beta})(X'X)(\hat{\beta}_R - \hat{\beta})$$

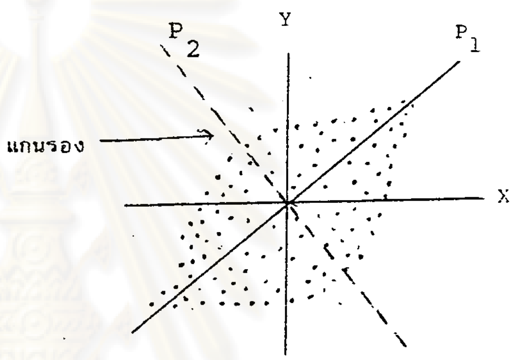
นอกจากจะใช้วิธี ริดจ์ รีเกรสชั่น ในการประมาณค่าสัมประสิทธิ์ความถดถอยพหุ เมื่อตัวแปรอิสระมีพหุสัมพันธ์กันแล้ว ก็อาจใช้วิธีเกรสชั่นพรินซิเปิ้ลคอมโพเน้นท์ (Regression principal component) ซึ่งจะได้ตัวประมาณที่เอนเอียง และให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำกว่าวิธีกำลังสองน้อยที่สุด

2.3 การประมาณค่าสัมประสิทธิ์ความถดถอยพหุ โดยวิธีรีเกรสชั่นพรินซิเปิ้ลคอมโพเน้นท์

วิธีนี้ได้นำการวิเคราะห์องค์ประกอบหลัก (Principal component analysis) มาช่วยในการประมาณค่าสัมประสิทธิ์ความถดถอยพหุ เมื่อเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ ด้วยการจัดรูปแบบตัวแปรอิสระเสียใหม่ก่อนที่จะนำไปหาค่าสัมประสิทธิ์ความถดถอยพหุ โดยให้ตัวแปรใหม่เป็นผลบวกเชิงเส้นของตัวแปรเดิม

การวิเคราะห์องค์ประกอบหลัก เป็นวิธีการลดตัวแปรให้น้อยลง โดยอาศัยหลักความสัมพันธ์เชิงเส้นระหว่างตัวแปรที่ใช้เป็นข้อมูลในการวิเคราะห์ แต่ไม่มีข้อสมมติเกี่ยวกับความสัมพันธ์เชิงสาเหตุ และผลระหว่างปัจจัยและตัวแปร เช่นการวิเคราะห์องค์ประกอบหลักที่ใช้ตัวแปร 2 ตัว คือ X และ Y ก่อนการวิเคราะห์องค์ประกอบหลักเริ่ม จากการศึกษาความสัมพันธ์ระหว่างตัวแปร สมมติว่า X และ Y มีความสัมพันธ์กัน ซึ่งความสัมพันธ์นี้ดูได้จากการลงจุดบนแกน X และแกน Y ดังที่แสดงไว้ในรูปที่ 2.1 และรูปที่ 2.2

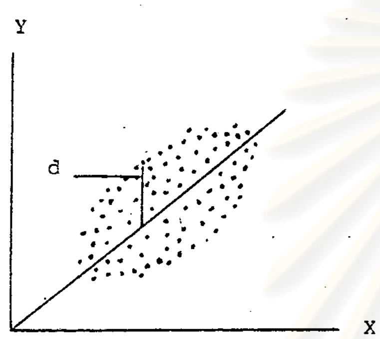




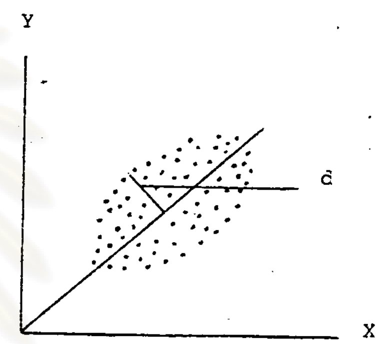
รูปที่ 2.1 X และ Y มีความสัมพันธ์กันสูง รูปที่ 2.2 X และ Y มีความสัมพันธ์กันต่ำกว่า

จากรูปที่ 2.1 ตัวแปร X และ Y มีความสัมพันธ์กันสูง และเป็นความสัมพันธ์เชิงเส้นตรงในทางบวก ถ้าสามารถกำหนดความลาดชันของเส้นตรงได้ ดังเช่นที่แสดงไว้ในรูปก็จะได้เส้นที่สามารถหาค่าของ X เมื่อรู้ค่าของ Y และสามารถหาค่าของ X เมื่อรู้ค่าของ Y ได้ ซึ่งเส้นตรงที่ได้นี้จะเรียกว่า แกนหลัก (Principal axis) ถ้าจุดต่าง ๆ อยู่บนเส้นตรงแกนหลักทั้งหมด แกนหลักก็สามารถที่จะให้ข้อมูลเกี่ยวกับ X และ Y ได้อย่างถูกต้องสมบูรณ์ทุกค่า แต่ถ้าจุดแสดงค่า x และ y กระจายออกไปมาก ก็ต้องอาศัยแกนเพิ่มอีกหนึ่งแกน ซึ่งแกนที่เพิ่มขึ้นนี้จะต้องมีจุดเริ่มต้นตั้งฉากกับแกนหลักดังที่แสดงไว้ในรูปที่ 2.2 แกนหลักจะลากผ่านจุดต่าง ๆ ที่ทำให้ระยะทางระหว่างจุดกับแกนหลัก (โดยการลากเส้นจากจุดมาตั้งฉากกับแกนหลัก) สั้นที่สุด และทำให้ผลรวมของระยะทางยกกำลังสองมีค่าต่ำสุด

การหาค่าต่ำสุดของแกนหลักนี้ แตกต่างจากวิธีกำลังสองน้อยที่สุด ซึ่งพยายามที่จะหาเส้นตรง $y = a + bX$ และพยายามที่จะหาค่า $\hat{Y}$ ที่ประมาณได้ต่างจาก Y เดิมให้น้อยที่สุด กล่าวคือ $\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$ มีค่าน้อยที่สุด การลากเส้นระยะทางตามวิธีกำลังสองน้อยที่สุด เป็นการลากเส้นขนานกับแกน แต่การลากเส้นระยะทางตามวิธีการวิเคราะห์องค์ประกอบหลักจะลากเส้นตั้งฉากกับแกนหลัก



รูปที่ 2.3 การลากเส้นระยะทางต่ำสุดของวิธีกำลังสองน้อยที่สุด



รูปที่ 2.4 การลากเส้นระยะทางต่ำสุดของแกนหลักของวิธีการวิเคราะห์องค์ประกอบหลัก

รูปที่ 2.3 และรูปที่ 2.4 แสดงให้เห็นความแตกต่างระหว่างการหาค่าต่ำสุดของวิธีกำลังสองน้อยที่สุด และวิธีวิเคราะห์องค์ประกอบหลัก ถึงแม้ว่าทั้งสองวิธีจะพยายามทำให้ระยะทางต่ำสุดเช่นกัน ($\Sigma d_i^2 = 0$)

ถ้ามีจำนวนตัวแปรเพิ่มขึ้น จำนวนมิติของกราฟก็จะเพิ่มขึ้น เช่นถ้ามี 3 ตัวแปรก็ต้องเพิ่มเส้นแสดงมิติเพิ่มอีกหนึ่งเส้น และในการลงจุดก็ต้องคำนึงถึงค่าของตัวแปร 3 ตัวพร้อม ๆ กัน ในการหาแกนหลักก็ต้องหาแกนซึ่งสามารถอธิบายความผันแปรของตัวแปรทั้ง 3 ตัวให้ได้มากที่สุด และแกนต่อ ๆ ไป ก็จะต้องอธิบายความผันแปรที่เหลือให้ได้มากที่สุด

วิธีการแยกแกนตามลำดับจากแกนหลักไปยังแกนรอง หรือการลดตัวแปรตาม วิธีการวิเคราะห์องค์ประกอบหลักนี้ อาศัยสมการที่เรียกว่า Eigen-equation ซึ่งสามารถเขียนในรูปเมตริกซ์ได้ดังนี้

$$RV = \lambda V \qquad \text{.......................} \quad (2.3.1)$$

โดยที่ R คือ เมตริกซ์ที่ต้องการหา

V คือ เมตริกซ์ของ eigenvector

λ คือ eigenvalue

การแก้สมการ (2.3.1) จะได้ eigenvalues และ eigenvectors ที่สัมพันธ์กับเมตริกซ์ข้อมูล ซึ่งวิธีการแก้สมการนี้ อาศัยสมการที่เรียกว่า Determinant equation

$$\text{Det } (R - I\lambda) = 0 \qquad \text{...................} \quad (2.3.2)$$

ถ้าเมตริกซ์ประกอบด้วย 2 ตัวแปร จะได้

$$\text{Det} \begin{bmatrix} 1-\lambda & r_{12} \\ r_{12} & 1-\lambda \end{bmatrix} = 0 \qquad \text{..............} \quad (2.3.3)$$

ซึ่งสามารถกระจายเป็นสมการ ได้ดังนี้

$$(1-\lambda)\ (1-\lambda) - r_{12}\ (r_{12}) = 0 \text{.............} \quad (2.3.4)$$

$$\lambda^2 + 2\lambda + 1 - r_{12}^2 = 0 \text{............} (\ 2.3.5)$$

ในการแก้สมการ (2.3.5) ก็ทำได้ในทำนองเดียวกับการแก้สมการ $ax^2 + bx + c = 0$ ได้ eigenvalues ของเมตริกซ์ ความสัมพันธ์ระหว่างตัวแปร 2 ตัวคือ $\lambda = 1+r_{12}$และ $\lambda = 1-r_{12}$ ถ้าตัวแปร 2 ตัวมีความสัมพันธ์กันแบบสมบูรณ์ ($r_{12} = 1$) eigenvalue ค่าที่หนึ่ง (λ_1) จะมีค่าเป็นสอง และ eigenvalue ค่าที่สอง (λ_2) มีค่าเป็นศูนย์ แต่ถ้าตัวแปรอิสระไม่มีความสัมพันธ์กันเลย ($r_{12} = 0$) eigenvalue ทั้งสองค่า (λ_1 และ λ_2) มีค่าเท่ากับหนึ่ง ซึ่งจะพบว่าผลรวมของ λ_1 และ λ_2 จะเท่ากับจำนวนตัวแปรอิสระ ส่วนผลคูณของ λ_1 และ λ_2 จะเท่ากับ $(1 - r_{12}^2)$ ก็คือค่า Determinant ของเมตริกซ์ความสัมพันธ์

แกนหลักคือ แกนที่ eigenvalue มีค่ามากที่สุด ส่วนแกนของ eigenvalue จะมีค่ารองลงมาตามลำดับ ผลรวมของ eigenvalue ทั้งหมดจะเท่ากับจำนวนตัวแปร ถ้าหาร eigenvalue ค่าแรกด้วยจำนวนตัวแปรจะได้ค่าของความผันแปรของตัวแปรที่อธิบายได้โดยแกนหลัก และค่าผลหารต่อ ๆ ไปคือค่าของความผันแปรที่อธิบายได้โดยแกนรองตามลำดับ อาจสรุปได้ว่า

สัดส่วนของความผันแปรของแกนแต่ละแกนเท่ากับ eigenvalue ของแต่ละแกนหารด้วยจำนวนตัวแปร น้ำหนักของตัวแปรที่มีต่อองค์ประกอบ (Principal component loading) คือ อัตราความผันแปรของตัวแปรที่อธิบายได้โดยแกนหลักและแกนรอง ซึ่งเท่ากับผลคูณระหว่าง eigenvectors และรากที่สองของ eigenvalue ของแกน

ตัวประมาณที่เอนเอียงของค่าสัมประสิทธิ์ความถดถอยพหุ อาจหาได้จากวิธีรีเกรสชั่นพริ้นซิเปิ้ลคอมโพเน้นท์ จากตัวแบบทั่วไปของการวิเคราะห์ความถดถอยพหุ สมการ (2.1) อาจเขียนให้อยู่ในรูปคาโนนิคอล (Canonical form) ได้ดังนี้

$$Y = Z\alpha + \varepsilon \qquad \text{..................(2.3.6)}$$

โดยที่ $Z = XT$, $\alpha = T'\beta$ และ $T'X'XT = Z'Z = \Lambda$

X คือ เมตริกซ์ของตัวแปรอิสระขนาด $n \times p$

β คือ เมตริกซ์ ของสัมประสิทธิ์ความถดถอยพหุขนาด $p \times 1$

T คือ ออทอกอนอล เมตริกซ์ (Orthogonal matrix) $p \times p$ โดยที่แต่ละคอลัมน์ของเมตริกซ์เป็น eigenvectors ที่สอดคล้องกับค่า eigenvalues $\lambda_1, \lambda_2, \ldots\ldots, \lambda_p$ ตามลำดับ

Z คือ เมตริกซ์ซึ่งแต่ละคอลัมน์ของ Z คือตัวแปรอิสระชุดใหม่ ซึ่งเป็นผลรวมเชิงเส้น (linear combination) ของตัวแปรอิสระ หรืออาจกล่าวได้ว่า Z_1 คือ องค์ประกอบหลัก (องค์ประกอบที่ 1)

Z_2 คือ องค์ประกอบที่สอง

⋮

Z_p คือ องค์ประกอบที่ p

Λ คือ เมตริกซ์ในแนวเส้นทะแยงมุม ขนาด p x p โดยมีสมาชิกในแนวเส้นทะแยงมุมเป็น eigenvalues ของเมตริกซ์ $X'X$

จากสมการ (2.3.6) ใช้วิธีกำลังสองน้อยที่สุดเพื่อประมาณค่า α จะได้ตัวประมาณดังนี้

$$\hat{\alpha} = (Z'Z)^{-1} Z'Y = \Lambda^{-1} Z'Y$$

เมตริกซ์ความแปรปรวนของ $\hat{\alpha}$ มีค่าเป็น

$$Var(\hat{\alpha}) = \sigma^2 (Z'Z)^{-1} = \sigma^2 \Lambda^{-1}$$

ถ้า eigenvalue ของ $X'X$ มีค่าน้อย ทำให้ความแปรปรวนของตัวประมาณค่าสัมประสิทธิ์ความถดถอยพหุ ($\hat{\alpha}$) มีค่ามาก เนื่องจาก $Z'Z = \sum_{i=1}^{p} \sum_{j=1}^{p} Z_i Z_j = \Lambda$ จะนั้น λ_j ก็คือค่าความแปรปรวนขององค์ประกอบที่ j ซึ่งถ้า λ_j มีค่าเท่ากับหนึ่งทุกค่าของ j แสดงว่าตัวแปรอิสระไม่มีความสัมพันธ์กัน แต่ถ้า λ_j บางค่า มีค่าเท่ากับศูนย์แสดงว่าตัวแปรอิสระมีความสัมพันธ์กัน และถ้ามี λ_j บางค่ามีค่าเกือบเป็นศูนย์แสดงว่ามีพหุสัมพันธ์ระหว่างตัวแปรอิสระ

เนื่องจาก $\alpha = T'\beta$ ดังนั้น $\beta = T\alpha$ เพราะฉะนั้นความแปรปรวนของตัวประมาณค่าสัมประสิทธิ์ความถดถอยพหุมีค่าเป็น

$$Var(\hat{\beta}) = Var(T\hat{\alpha}) = T\Lambda^{-1}T'\sigma^2$$

$$\text{ดังนั้น } Var(\hat{\beta}_j) = \sigma^2 \left(\sum_{i=1}^{p} t_{ji}^2 / \lambda_i\right)$$

การที่ความแปรปรวนของ $\hat{\beta}$ เป็นผลรวมของส่วนกลับของ eigenvalue แสดงว่า ถ้ามี eigenvalue บางค่ามีค่าน้อย จะทำให้ความแม่นยำของตัวประมาณค่าสัมประสิทธิ์ความถดถอยพหุด้วยวิธีกำลังสองน้อยที่สุดลดลงไป

ในการใช้การวิเคราะห์องค์ประกอบหลัก เพื่อประมาณค่าสัมประสิทธิ์ความถดถอยพหุ เมื่อมีพหุสัมพันธ์ระหว่างตัวแปรอิสระนั้นจะนำองค์ประกอบมาใช้เพียงบางองค์ประกอบเท่านั้น

$$\text{ให้ } \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots\ldots \geq \lambda_s \geq \ldots\ldots \geq \lambda_p > 0$$

มี eigenvalues S ค่า ซึ่งมีค่าเกือบเป็นศูนย์ ในการวิเคราะห์องค์ประกอบหลักนั้น องค์ประกอบซึ่ง eigenvalue มีค่าน้อยจะถูกตัดออกจากตัวแบบ จากนั้นก็จะใช้วิธีกำลังสองน้อยที่สุด

เพื่อประมาณค่าสัมประสิทธิ์ความถดถอยพหุ ซึ่งจะได้

$$\hat{\alpha}_{PC} = B\hat{\alpha} \quad \text{โดยที่} \quad B = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{bmatrix}$$

$$b_1 = b_2 \cdots\cdots = b_{p-s} = 1 \qquad b_{p-S+1} = b_{p-S+2} = \cdots = b_p = 0$$

$$\hat{\alpha}_{PC} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_{P-S} \\ 0 \\ \vdots \\ 0 \end{bmatrix} \begin{array}{l} \\ \\ \text{P-S องค์ประกอบ} \\ \\ \\ \text{S องค์ประกอบ} \\ \\ \end{array}$$

ตัวประมาณค่าสัมประสิทธิ์ความถดถอยพหุ โดยวิธีรีเกรสชั่นพริ้นซิเปิ้ลคอมโพเนนท์ จะมีค่าเป็น

$$\hat{\beta}_{PC} = T\,\hat{\alpha}_{PC}$$

$$= \sum_{j=1}^{P-S} \lambda_j^{-1} t_j' X' Y t_j$$

ในการใช้การ วิธีรีเกรสชั่นพริ้นซิเปิ้ลคอมโพเน้นท์เพื่อประมาณค่าสัมประสิทธิ์ความถดถอยพหุ เมื่อมีพหุสัมพันธ์ระหว่างตัวแปรอิสระ ได้ลดผลกระทบของการเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ โดยการตัดองค์ประกอบบางองค์ประกอบ ออกจากตัวแบบ และในการตัดสินควรว่าควรจะตัดองค์ประกอบใดออกจากตัวแบบบ้างนั้น อาจจะใช้หลักเกณฑ์ดังนี้ (Duntman (1984 ; 175))

1) วาดกราฟ โดยให้แกน y แทน eigenvalue (λ_i) และแกน x แทนค่า i แล้วดูว่า eigenvalues จะเริ่มมีค่าน้อยเมื่อไหร่ (กฎข้อนี้ เป็นผลงานวิจัยของ Cattel (1966) ซึ่งให้ชื่อว่า "Scree graph")

2) โดยการกำหนดเปอร์เซ็นต์ที่สามารถอธิบายความผันแปรได้

3) ตัดองค์ประกอบซึ่งมี eigenvalue น้อยกว่าค่าเฉลี่ยของ eigenvalue

2.4 ดัชนีพหุสัมพันธ์ (Degree of multicollinearity)

ดัชนีพหุสัมพันธ์คือ ค่าซึ่งบอกให้ทราบว่าตัวแปรอิสระมีความสัมพันธ์กันมากหรือน้อยเพียงใด มีผู้เสนอค่าต่าง ๆ ซึ่งสามารถใช้เป็นดัชนีพหุสัมพันธ์ไว้หลายคน

1. Marquart (1970 ; 299) ได้เสนอให้ใช้ค่าสมาชิกในแนวเส้นทะแยงมุมของอินเวอร์สเมตริกซ์ความสัมพันธ์ ซึ่งให้ชื่อว่า Variance inflation factor (VIF) เป็นดัชนีพหุสัมพันธ์โดยที่

$$VIF_i = \frac{1}{(1 - R_i^2)} \quad ; \quad R_j^2 \text{ คือค่าสัมประสิทธิ์การตัดสินใจของตัวแปร } X_j \text{ ซึ่งถดถอยกับตัวแปรอิสระที่เหลือ}$$

ถ้า x เกือบจะเป็นอิสระกับตัวแปรอื่น ๆ R^2 จะมีค่าน้อย และ VIF จะมีค่าเข้าใกล้หนึ่ง แต่ถ้า x มีความสัมพันธ์เชิงเส้นกับตัวแปรที่เหลือบางตัว R_j^2 จะมีค่าใกล้เคียงกับหนึ่ง และ VIF จะมีค่ามาก ถ้าข้อมูลมี VIF บางค่าซึ่งมีค่ามากจะบอกได้ ว่าตัวแปรอิสระมีพหุสัมพันธ์กัน โดยทั่วไปในทางปฏิบัติถ้า VIF ค่าใดมีค่ามากกว่า 5 หรือ 10 แสดงว่าค่าประมาณสัมประสิทธิ์ความถดถอยพหุซึ่งสอดคล้องกับ VIF ค่านั้นเป็นตัวประมาณที่ไม่ดีเนื่องจากมีพหุสัมพันธ์ระหว่างตัวแปรอิสระ

2. Chaterjee and Price (1977, 1983) ได้กล่าวว่า VIF อาจใช้หาค่าคาดหวังของระยะห่างกำลังสองของค่าประมาณ โดยวิธีกำลังสองน้อยที่สุด จากค่าจริงซึ่งจะใช้วัดความแม่นยำของค่าประมาณจากวิธีกำลังสองน้อยที่สุด ถ้าระยะห่างมีค่าน้อยลงตัวประมาณจะยิ่งมีความแม่นยำมากขึ้น

ให้ L^2 คือ ระยะห่างกำลังสอง

$$\therefore L^2 = \sigma^2 \sum_{i=1}^{p} VIF_i \quad ; \quad p = \text{จำนวนตัวแปรอิสระ}$$

ถ้าตัวแปรอิสระไม่มีพหุสัมพันธ์กัน VIF จะมีค่าเท่ากับหนึ่งทุกค่า ซึ่งจะทำให้

$$L^2 = p\sigma^2 , \quad R_L = \frac{\sigma^2 \sum_{i=1}^{p} VIF_i}{p\sigma^2} = \frac{\sum_{i=1}^{p} VIF_i}{p}$$

R_L จะวัดความคลาดเคลื่อนกำลังสองของวิธีกำลังสองน้อยที่สุด เมื่อเปรียบเทียบกับขนาดของความคลาดเคลื่อนของข้อมูลนั้น ถ้าตัวแปรอิสระไม่มีพหุสัมพันธ์กัน ดังนั้นค่า R_L จึงอาจใช้เป็นดัชนีพหุสัมพันธ์ได้ ตัวอย่างเช่น ถ้า R_L = 124 หมายความว่า ความคลาดเคลื่อนกำลังสองของตัวประมาณวิธีกำลังสองน้อยที่สุด จะเป็น 124 เท่าของความคลาดเคลื่อนกำลังสองของตัวประมาณวิธีกำลังสองน้อยที่สุด ถ้าตัวแปรอิสระไม่มีพหุสัมพันธ์กัน

3. Montgomery (1982 : 301) ถ้า $\lambda_1 , \lambda_2 \ldots\ldots \lambda_p$ เป็น eigenvalue ของเมตริกซ์ความสัมพันธ์ให้ $m = \frac{\lambda max}{\lambda min}$ อาจใช้ค่า m เป็นดัชนีพหุสัมพันธ์ได้ ค่า m จะวัดการกระจายของ eigenvalue โดยทั่วไปถ้า $m < 100$ การที่ตัวแปรอิสระมีพหุสัมพันธ์กันจะไม่ก่อให้เกิดปัญหาจะไม่ก่อให้เกิดปัญหาในการประมาณมากนัก แต่ถ้า $m > 1000$ การที่ตัวแปรอิสระมีพหุสัมพันธ์กันจะก่อให้เกิดปัญหาในการประมาณมาก

4. Montgomery (1982, 305) กล่าวว่า $|R|^{\frac{1}{2}}$ จะเป็นตัววัดถึงการสูญเสียอำนาจในการประมาณเนื่องจากมีพหุสัมพันธ์ระหว่างตัวแปรอิสระ โดยที่ R คือเมตริกซ์ความสัมพันธ์

ในการวิจัยเรื่องนี้ ใช้ $|R|^{\frac{1}{2}}$ เป็นดัชนีพหุสัมพันธ์ เพราะมีวัตถุประสงค์ที่จะเปรียบเทียบการประมาณค่าด้วยวิธี ริดจ์ รีเกรสชั่น วิธีรีเกรสชั่นพรินซิเปิ้ลคอมโพเน้นท์ และวิธีกำลังสองน้อยที่สุด เมื่อตัวแปรอิสระมีพหุสัมพันธ์กัน ดังนั้นจึงต้องสร้างข้อมูลให้ตัวแปรอิสระมีพหุสัมพันธ์กันทุกระดับของค่าดัชนีพหุสัมพันธ์ การใช้ $|R|^{\frac{1}{2}}$ เป็นดัชนีพหุสัมพันธ์ ทำให้สามารถกำหนดช่วงของ $|R|^{\frac{1}{2}}$ ได้ว่า $0 < |R|^{\frac{1}{2}} < 1$; ถ้า $|R|^{\frac{1}{2}}$ มีค่าใกล้เคียงกับหนึ่ง แสดงว่าตัวแปรอิสระมีพหุสัมพันธ์กันน้อย ถ้า $|R|^{\frac{1}{2}} = 1$ แสดงว่าตัวแปรอิสระไม่มีพหุสัมพันธ์กัน แต่ถ้า $|R|^{\frac{1}{2}}$ มีค่าใกล้เคียงกับศูนย์ แสดงว่าตัวแปรอิสระมีพหุสัมพันธ์กันมาก