บทที่ 2

ทฤษฎีและตัวสถิติที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงวิธีการประมาณสัมประสิทธิ์การถดถอยพหุคูณทั้ง 3 วิธี ซึ่งได้แก่ วิธีกำลังสองน้อยสุด วิธีการถดถอยองค์ประกอบหลัก และวิธีกำลังสองน้อยสุดแบ่งส่วน ซึ่งรายละเอียดต่างๆ มีดังนี้

2.1 การประมาณค่าสัมประสิทธิ์การถดถอยพหุคูณด้วยวิธีกำลังสองน้อยสุด

ตัวแบบทั่วไปของการวิเคราะห์การถดถอยพหุคูณเชิงเส้น สามารถเขียนได้ดังนี้

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

เมื่อ $\underset{\sim}{y}$ เป็นเวกเตอร์ของตัวแปรตามขนาด $n \times 1$

X เป็นเมทริกซ์ของตัวแปรอิสระขนาด $n \times (p+1)$

$\underset{\sim}{\beta}$ เป็นเวกเตอร์ของสัมประสิทธิ์การถดถอยพหุคูณขนาด $(p+1) \times 1$

$\underset{\sim}{\varepsilon}$ เป็นเวกเตอร์ของความคลาดเคลื่อนขนาด $n \times 1$

n เป็นขนาดตัวอย่างของตัวแปรแต่ละตัว

p เป็นจำนวนตัวแปรอิสระ

โดยที่ $E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0}$, $\mathrm{cov}(\underset{\sim}{\varepsilon}) = \sigma^2 I_n$

จากตัวแบบในสมการ (2.1) เราประมาณค่าเวกเตอร์พารามิเตอร์ $\underset{\sim}{\beta}$ ด้วยวิธีกำลังสองน้อยสุด ซึ่งจะได้ค่าประมาณคือ $\hat{\underset{\sim}{\beta}} = (X'X)^{-1}X'\underset{\sim}{y}$ โดย $\hat{\underset{\sim}{\beta}}$ เป็นตัวประมาณไม่เอนเอียงซึ่งมีค่าเฉลี่ยความคลาดเคลื่อนต่ำสุดในบรรดาตัวประมาณไม่เอนเอียง การประมาณค่า $\underset{\sim}{\beta}$ ด้วยวิธีกำลังสองน้อยสุดมีข้อสมมติที่จำเป็นข้อหนึ่งคือ ตัวแปรอิสระต้องไม่มีความสัมพันธ์กันซึ่งในภาคปฏิบัติเป็นไปได้น้อยมาก ในกรณีที่ตัวแปรอิสระมีความสัมพันธ์กัน คือ เกิดเงื่อนไขที่ไม่ดี (ill-condition) การประมาณค่า $\underset{\sim}{\beta}$ ด้วยวิธีกำลังสองน้อยสุดอาจจะไม่ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองน้อยที่สุด ซึ่งการพิจารณาผลของตัวแปรอิสระที่มีพหุสัมพันธ์กัน สามารถพิจารณาจากคุณสมบัติ 2 ประการ คือ เมทริกซ์

กำหนดให้ความแปรปรวนของเวกเตอร์ตัวประมาณ $\hat{\underset{\sim}{\beta}}$ มีค่าเป็น $\mathrm{cov}\left(\hat{\underset{\sim}{\beta}}\right)$ จะได้ว่า

(2.1.2) $$\mathrm{cov}(\hat{\underset{\sim}{\beta}}) = \sigma^2 (X'X)^{-1}$$

และให้ L_1 คือความแตกต่างระหว่าง $\hat{\underset{\sim}{\beta}}$ และ $\underset{\sim}{\beta}$ ดังนั้นค่าความแตกต่างกำลังสองระหว่าง $\hat{\underset{\sim}{\beta}}$ และ $\underset{\sim}{\beta}$ มีค่าเป็น

$$L_1^2 = \left(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta}\right)' \left(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta}\right)$$

ค่าเฉลี่ยความแตกต่างกำลังสองระหว่าง $\hat{\underset{\sim}{\beta}}$ และ $\underset{\sim}{\beta}$ คือ $E(L_1^2)$ ซึ่งมีค่าดังนี้

$$E(L_1^2) = \sigma^2 \mathrm{trace}(X'X)^{-1}$$

โดยที่ $E(L_1^2)$ จะมีค่าสมมูล (Equivalent) กับ $E\left[\left(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta}\right)' \left(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta}\right)\right]$

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

(2.1.3) $$= E\left[\hat{\underset{\sim}{\beta}}'\hat{\underset{\sim}{\beta}} - 2\hat{\underset{\sim}{\beta}}\underset{\sim}{\beta} + \underset{\sim}{\beta}'\underset{\sim}{\beta}\right]$$

$$= E\left[\hat{\underset{\sim}{\beta}}'\hat{\underset{\sim}{\beta}}\right] - \underset{\sim}{\beta}'\underset{\sim}{\beta}$$

เมื่อแทนค่า $E(L_1^2)$ ในสมการ (2.1.3) เราจะสามารถหาค่า $E(\hat{\underset{\sim}{\beta}}'\hat{\underset{\sim}{\beta}})$ ได้คือ

(2.1.4) $$E(\hat{\underset{\sim}{\beta}}'\hat{\underset{\sim}{\beta}}) = \underset{\sim}{\beta}'\underset{\sim}{\beta} + \mathrm{trace}(X'X)^{-1}$$

ถ้า $\underset{\sim}{\varepsilon}$ มีการแจกแจงแบบปกติจะได้ว่า

(2.1.5) $$\mathrm{Var}(L_1^2) = 2\sigma^4 \mathrm{trace}(X'X)^{-2}$$

เราเห็นได้ว่าจากสมการ (2.1.4) และ (2.1.5) จะอยู่ในรูปฟังก์ชันซึ่งเป็นผลบวกของสมาชิกในแนวทแยงของเมทริกซ์ $X'X$ ($\mathrm{trace}(X'X)$) ดังนั้นเพื่อความสะดวกในการเปรียบเทียบการประมาณค่า

$\hat{\underset{\sim}{\beta}}$ จึงควรแปลงให้อยู่ในรูปฟังก์ชันของค่าเจาะจง (eigenvalue) ของเมทริกซ์ $X'X$ โดยใช้คุณสมบัติที่สำคัญข้อหนึ่งของค่าเจาะจงของเมทริกซ์คือ ถ้า λ_i เป็นค่าเจาะจงของเมทริกซ์ $X'X$ เมื่อ $i = 1,2,\ldots,p$ แล้ว $\sum_{i=1}^{p} \lambda_i = \text{trace}(X'X)$

สมมติให้ค่า λ_i เป็นค่าเจาะจงของเมทริกซ์ $X'X$ ซึ่ง λ_{max} คือ ค่าเจาะจงที่มีค่ามากที่สุดและ λ_{min} คือ ค่าเจาะจงที่มีค่าน้อยที่สุด ดังนั้น $\lambda_{max} = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$

จากสมการ (2.1.4) และ (2.1.5) เราสามารถเขียนในรูปฟังก์ชันของค่าเจาะจง ได้ดังนี้

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

และ

$$Var(L_1^2) = 2\sigma^4 \sum_{i=1}^{p} \left(\frac{1}{\lambda_i}\right)^2$$

ดังนั้นในกรณีที่ตัวแปรอิสระมีสภาพที่ไม่เหมาะสม กล่าวคือ ค่าเจาะจงบางค่าของเมทริกซ์ $X'X$ จะมีค่าน้อยมากๆ ซึ่งมีผลทำให้ความแตกต่างระหว่าง $\hat{\underset{\sim}{\beta}}$ และ $\underset{\sim}{\beta}$ มีค่ามาก นั่นคือ การประมาณค่า $\underset{\sim}{\beta}$ ด้วยวิธีกำลังสองน้อยสุด จะให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองที่มีค่าสูงขึ้น ดังนั้นในกรณีที่เกิดปัญหาพหุสัมพันธ์ตัวประมาณ $\hat{\underset{\sim}{\beta}}$ จึงเป็นตัวประมาณที่ไม่เหมาะสม

2.2 การประมาณสัมประสิทธิ์การถดถอยด้วยวิธีการถดถอยองค์ประกอบหลัก

การถดถอยองค์ประกอบหลักเป็นเทคนิคการประมาณที่เอนเอียงวิธีหนึ่งที่ใช้แก้ปัญหาพหุสัมพันธ์ระหว่างตัวแปรอิสระ ในวิธีนี้เราใช้การประมาณด้วยวิธีกำลังสองน้อยสุดกับกลุ่มของตัวแปรที่เรียกว่า องค์ประกอบหลัก (principle component) ของเมทริกซ์สหสัมพันธ์ โดยมีการจำกัดจำนวนองค์ประกอบหลักที่แน่นอนซึ่งมีผลต่อการลดความแปรปรวนลง แต่เป็นตัวประมาณที่เอนเอียง ในกรณีที่มีพหุสัมพันธ์ระหว่างตัวแปรอิสระ วิธีนี้จะให้ผลลัพธ์ในการประมาณค่าและการพยากรณ์ค่าได้ดีกว่าวิธีกำลังสองน้อยสุด

2.2.1 <u>ตัวแบบและวิธีการวิเคราะห์การถดถอยองค์ประกอบหลัก</u>

พิจารณาเมทริกซ์ของเวกเตอร์เจาะจงที่สัมพันธ์กับค่าเจาะจง $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$

ของ X'X เราทราบว่า V' V = I เนื่องจาก V เป็นเมทริกซ์ตั้งฉาก (Orthogonal matrix) ดังนั้น เราสามารถเขียนตัวแบบถดถอยดั้งเดิมในรูปแบบของ

$$\underset{\sim}{y} = \beta_0 \underset{\sim}{1} + X^* VV' \underset{\sim}{\beta} + \underset{\sim}{\varepsilon} \tag{2.2.1}$$

$$\underset{\sim}{y} = \beta_0 \underset{\sim}{1} + Z \underset{\sim}{\alpha} + \underset{\sim}{\varepsilon} \tag{2.2.2}$$

เมื่อ $Z = X^* V$ และ $\underset{\sim}{\alpha} = V' \underset{\sim}{\beta}$

Z เป็นเมทริกซ์ขนาด $n \times 1$

และ $\underset{\sim}{\alpha}$ เป็นเวกเตอร์ของสัมประสิทธิ์ตัวใหม่ $\alpha_1, \alpha_2, \ldots, \alpha_p$ ขนาด $p \times 1$

เราสามารถพิจารณาแนวตั้งของเมทริกซ์ Z (สมาชิกของเมทริกซ์ Z_{ij}) เป็นตัวแปรใหม่ p ตัว หรือเป็นองค์ประกอบหลัก p ตัว และสามารถแสดงได้ว่าองค์ประกอบเหล่านี้ตั้งฉากซึ่งกันและกัน เราจะได้

$$\begin{aligned} Z'Z &= (X^* V)'(X^* V) \\ &= V' X^{*\prime} X^* V \\ &= \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p) \end{aligned} \tag{2.2.3}$$

ดังนั้นถ้าการถดถอยอยู่ในรูปแบบของตัวแบบ Z ในสมการที่ (2.2.2) ความแปรปรวนของสัมประสิทธิ์ (สมาชิกในแนวทแยงของ $(Z'Z)^{-1}$ หารด้วย σ^2) เป็นส่วนกลับของค่าเจาะจง ดังนี้

$$\frac{\mathrm{Var}(\hat{\alpha}_j)}{\sigma^2} = \frac{1}{\lambda_j} \quad ; j = 1, 2, \ldots, p \tag{2.2.4}$$

เราสังเกตได้ว่า $\hat{\underset{\sim}{\alpha}}$ เป็นตัวประมาณด้วยวิธีกำลังสองน้อยสุด ถ้าทุกองค์ประกอบหลักยังคงอยู่ในตัวแบบถดถอยแล้ว การแปลงตัวแปรเป็นองค์ประกอบหลักเหล่านั้นเป็นการหมุนแกนของตัวแปรถดถอย อย่างไรก็ตามตัวแปรใหม่เหล่านี้ตั้งฉากกัน ขนาดของความแปรปรวนยังคงเท่าเดิม เนื่องจากความแปรปรวนทั้งหมดถูกแจกแจงใหม่ ถ้าตัวแปรอิสระมีพหุสัมพันธ์กันจะมีอย่างน้อย 1 องค์ประกอบที่มีค่าเจาะจงน้อยที่สุด การตัดองค์ประกอบหลักออกไปอย่างน้อย 1 องค์ประกอบอาจจะทำให้ลดความแปรปรวนทั้งหมดในตัวแบบให้น้อยลงและสามารถปรับปรุงสมการที่ใช้พยากรณ์ให้เหมาะสมยิ่งขึ้นได้

2.2.2 เกณฑ์ที่ใช้ในการพิจารณาจำนวนองค์ประกอบหลัก

หลักการและความคิดพื้นฐานของการถดถอยองค์ประกอบหลักเหมือนกับหลักการของกำลังสองน้อยสุดในมุมมองของตัวแบบทั่วไป คือเป็นการประมาณค่าด้วยวิธีกำลังสองน้อยสุดที่ดำเนินบนองค์ประกอบและถ้าองค์ประกอบหนึ่งถูกตัดทิ้ง ผลลัพธ์ของตัวประมาณของสัมประสิทธิ์การถดถอยของตัวแปรดั้งเดิม X จะเป็นตัวประมาณที่เอนเอียง ความแปรปรวนที่เกิดจากการสร้างองค์ประกอบนั้นๆ จะถูกตัดทิ้งด้วย ทำให้ความแปรปรวนลดลง เช่นเดียวกับในมุมมองของตัวแปรด้วยวิธีกำลังสองน้อยสุด การพิจารณาจำนวนขององค์ประกอบหลักที่จะต้องตัดทิ้งจึงเป็นเรื่องที่สำคัญมาก ซึ่งเกณฑ์ที่ใช้ในการพิจารณามีหลายเกณฑ์ด้วยกัน ในการวิจัยนี้ เราจะใช้ค่าเฉลี่ยของค่าเจาะจงทั้งหมดเป็นเกณฑ์ นั่นคือ

$$\text{ค่าเฉลี่ยของค่าเจาะจงทั้งหมด } \overline{\lambda} = \frac{1}{p}\sum_{i=1}^{p} \lambda_i$$

โดยที่เราจะตัดองค์ประกอบหลักที่มีจำนวนเท่ากับจำนวนของค่าเจาะจะที่มีค่าน้อยกว่าค่าเฉลี่ยของค่าเจาะจงทั้งหมดทิ้งไป จำนวนองค์ประกอบที่เหลือจะเป็นตัวแปรถดถอยที่นำมาสร้างตัวแบบถดถอยด้วยวิธีกำลังสองน้อยสุด

2.2.3 ความเอนเอียงของสัมประสิทธิ์การถดถอยองค์ประกอบหลัก

สมมติว่าเราพิจารณาองค์ประกอบหลักที่ถูกตัดทิ้งไป r องค์ประกอบ และเหลืออยู่ s องค์ประกอบ โดยที่ $s + r = p$ และเราพิจารณาเมทริกซ์เวกเตอร์เจาะจง $V = \left[V_1, V_2, \ldots, V_p\right]$ ของเมทริกซ์ $X'X$ ซึ่งสามารถแบ่งส่วนออกเป็น

$$V = \left[V_r : V_s\right]$$

และในทำนองเดียวกัน พิจารณาเมทริกซ์ Λ เป็นเมทริกซ์แนวทะแยงของค่าเจาะจงของ $X'X$ เราสามาถแบ่งส่วน Λ เป็น

$$\Lambda = \begin{bmatrix} \Lambda_r & 0 \\ 0 & \Lambda_s \end{bmatrix}$$

โดยที่ Λ_r และ Λ_s เป็นเมทริกซ์แนวทะแยงที่มี Λ_r เป็นเมทริกซ์ที่ประกอบด้วยค่าเจาะจงที่สัมพันธ์กับองค์ประกอบที่ถูกตัดทิ้ง เนื่องจาก $V(X'X)V = Z'Z = \Lambda$ การประมาณด้วยวิธีกำลังสองน้อยสุดของ $\underset{\sim}{\alpha}$ สามารถเขียนได้ดังนี้

(2.2.5) $$\hat{\underset{\sim}{\alpha}} = (Z'Z)^{-1}Z'y = \Lambda^{-1}V'X'y$$

ซึ่งแสดงได้ว่า ตัวประมาณสำหรับ $\underset{\sim}{\alpha}$ ที่อยู่ในตัวแบบ คือ

$$\hat{\underset{\sim}{\alpha}}_s = (Z'Z)^{-1}Z'y = \Lambda_s^{-1}V_s'X'y$$

ดังที่ได้กล่าวในตอนต้น เราสามารถมองได้ว่า การถดถอยองค์ประกอบหลักเป็นเหมือนกับการถดถอยด้วยกำลังสองน้อยที่สุดบนตัวแบบที่สร้างขึ้นด้วยองค์ประกอบหลักและเนื่องจากองค์ประกอบเหล่านี้ตั้งฉากกัน เราสามารถแสดงได้ว่า $\hat{\underset{\sim}{\alpha}}_s$ เป็นตัวประมาณไม่เอนเอียงของ $\underset{\sim}{\alpha}_s$ และเราสามารถแปลงองค์ประกอบหลักในตัวแบบกลับเป็นตัวแปรดั้งเดิมได้ ดังนี้

จาก $$\underset{\sim}{\alpha} = V'\underset{\sim}{\beta}$$

(2.2.6) ดังนั้น $$\underset{\sim}{\beta} = V\underset{\sim}{\alpha}$$

ถ้าเราตัดองค์ประกอบสุดท้าย r องค์ประกอบทิ้งไป ตัวประมาณกำลังสองน้อยสุดของสัมประสิทธิ์การถดถอยสำหรับพารามิเตอร์ p ตัว สามารถเขียนเป็น

$$b_{pc} = \begin{bmatrix} b_{1,pc} \\ b_{2,pc} \\ \vdots \\ b_{p,pc} \end{bmatrix} = \begin{bmatrix} V_1 V_2 \ldots V_{p-r} \end{bmatrix} \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \\ \vdots \\ \hat{\alpha}_r \end{bmatrix}$$

(2.2.7) หรือ $$b_{pc} = V_s \hat{\underset{\sim}{\alpha}}_s$$

ดังนั้น $$E(b_{pc}) = V_s \underset{\sim}{\alpha}_s = V_s V_s' \underset{\sim}{\beta}$$

เนื่องจาก $VV' = I = V_r V_r' + V_s V_s'$

$$E(b_{pc}) = \left[I - V_r V_r'\right] \underset{\sim}{\beta}$$
$$= \underset{\sim}{\beta} - V_r V_r' \underset{\sim}{\beta}$$
$$= \underset{\sim}{\beta} - V_r \underset{\sim}{\alpha_r}$$

ดังนั้น ตัวประมาณสัมประสิทธิ์การถดถอยมีความเอนเอียงเท่ากับ $V_r \underset{\sim}{\alpha_r}$ โดยที่ $\underset{\sim}{\alpha_r}$ เป็นเมทริกซ์ขององค์ประกอบหลักที่ถูกตัดทิ้งไป

2.2.3 <u>ความแปรปรวนของสัมประสิทธิ์การถดถอยองค์ประกอบหลัก</u>

การตัดองค์ประกอบเป็นผลให้ความแปรปรวนของสัมประสิทธิ์การถดถอย b_{pc} ขนาดของการลดลงของความแปรปรวนขึ้นอยู่กับขนาดของพหุสัมพันธ์ ถ้าทุกองค์ประกอบอยู่ในตัวแปร b_{pc} ถูกเปลี่ยนไปเป็นกำลังสองน้อยที่สุด ดังนั้น

$$\frac{\mathrm{Var}\, b}{\sigma^2} = (X'X)^{-1}$$
$$= V\Lambda^{-1}V' \qquad (2.2.8)$$
$$= V_r\Lambda_r^{-1}V_r' + V_s\Lambda_s^{-1}V_s'$$

จากสมการ (2.2.7) และในความจริงที่ว่า $\mathrm{Var}\, \hat{\alpha}_s = \Lambda_s^{-1}$ เราจะได้เมทริกซ์ความแปรปรวนร่วม

$$\frac{\mathrm{Var}\, b_{pc}}{\sigma^2} = V_s\Lambda_s^{-1}V_s'$$

ดังนั้น ความแตกต่างในเมทริกซ์ความแปรปรวนและความแปรปรวนร่วมสำหรับตัวประมาณกำลังสองน้อยสุดและตัวประมาณองค์ประกอบหลัก คือ $V_r\Lambda_r^{-1}V_r'$

2.3 การประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยสุดแบ่งส่วน

วิธีกำลังสองน้อยสุดแบ่งส่วนเป็นวิธีการใหม่ที่ใช้สร้างสมการการถดถอยที่ได้รับความสนใจอย่างมากที่ปรากฏในเอกสารหลายฉบับ เช่น เฮลแลนด์ (1988,1990), Hoskuldsson (1988), สโตนและบรู๊คส์ (1990) เป็นต้น วิธีนี้สามารถใช้กับการถดถอยที่มีตัวแปรตามตัวแปรเดียว (univariate partial least square) และการถดถอยที่มีตัวแปรตามมากกว่า 1 ตัว (multivariate partial least square) นั่นคือ เราสามารถใช้วิเคราะห์ตัวแปรตาม $y_1, y_2, \ldots, y_m$ ในการสร้างความสัมพันธ์ระหว่างตัวแปรตาม y และตัวแปรอิสระ $x_1, x_2, \ldots, x_p$ วิธีกำลังสองน้อยสุดแบ่งส่วนจะสร้างตัวแปรอิสระขึ้นใหม่เรียกว่าปัจจัย (factor) หรือ องค์ประกอบ (component) โดยที่แต่ละองค์ประกอบเป็นการรวมกันเชิงเส้นตรงของ $x_1, x_2, \ldots, x_p$ และใช้วิธีการถดถอยมาตรฐาน (standard regression) ในการสร้างสมการความสัมพันธ์ระหว่างองค์ประกอบกับตัวแปรตาม y

วิธีกำลังสองน้อยสุดแบ่งส่วนนี้มีความคล้ายคลึงกับวิธีการถดถอยองค์ประกอบหลัก คือ องค์ประกอบจะถูกสร้างขึ้นจากตัวแปรอิสระในการถดถอย แต่มีความแตกต่างกันที่สำคัญระหว่าง 2 วิธีนี้ คือ การถดถอยองค์ประกอบหลักจะพิจารณาจากข้อมูลของตัวแปรอิสระ X เท่านั้น ในขณะที่วิธีกำลังสองน้อยสุดแบ่งส่วนนั้นค่าข้อมูลของตัวแปรอิสระ X และตัวแปรตาม y จะมีผลต่อการสร้างองค์ประกอบ ซึ่งคล้ายกับวิธีลาเรนท์รูธรีเกรสชัน (Larent Root regression) แต่แตกต่างกันในเรื่องการสร้างองค์ประกอบ จุดประสงค์หลักของวิธีกำลังสองน้อยสุดแบ่งส่วนนี้ คือ การสร้างองค์ประกอบที่ได้จากตัวแปรอิสระ X ที่ให้ข้อมูลในการพยากรณ์ค่า y มากที่สุด ในขณะเดียวกันสามารถลดปัญหาที่เกิดจากสมการการถดถอยที่มีจำนวนตัวแปรอิสระมากโดยใช้องค์ประกอบจำนวนไม่มากเป็นตัวแปรอิสระในสมการแทนจำนวนตัวแปรอิสระ X ซึ่ง วิธีนี้สามารถใช้ได้ดีเมื่อสมการการถดถอยมีจำนวนตัวแปรอิสระมากเมื่อเทียบกับขนาดตัวอย่างของตัวแปร มีขนาดเล็ก

2.3.1 <u>หลักการเบื้องต้นของวิธีกำลังสองน้อยสุดแบ่งส่วน</u>

จากรายละเอียดทั้งหมดที่กล่าวมาข้างต้นในหัวข้อ 2.3.1 เราสามารถสรุปเป็นขั้นตอนในการคำนวณ (Algorithm) โดยใช้หลักการทางคณิตศาสตร์ของตัวแบบเชิงเส้น (Linear Model) ได้ดังนี้

กำหนดเมทริกซ์ของข้อมูลอิสระ X (ค่าสังเกต × ตัวแปร) และเวกเตอร์ตัวแปรตาม $\underset{\sim}{y}$ นำมาลบด้วยค่าเฉลี่ยของตัวแปรอิสระ X และค่าเฉลี่ยของตัวแปรตาม y จะได้

$$E = X - 1\bar{x}$$
$$f = \underset{\sim}{y} - 1\bar{\underset{\sim}{y}}$$

โดยใช้หลักการทางคณิตศาสตร์ของตัวแบบเชิงเส้นของวิธีกำลังสองน้อยสุดแบ่งส่วน ซึ่งมีขั้นตอนดังนี้

1) การประมาณของเมทริกซ์ P ซึ่งเป็นเมทริกซ์ปริภูมิย่อย A มิติของค่าสังเกตของค่ามูล
2) การหาฉายาของเมทริกซ์ E บนปริภูมิดังกล่าวข้างต้น
3) การประมาณของสัมประสิทธิ์การถดถอย q ในการถดถอยของ f บน ค่าของเมทริกซ์ E ที่มีความสัมพันธ์กับ $\hat{P}$

เมทริกซ์ $\hat{P}$ เรียกว่าตัวประมาณของเมทริกซ์น้ำหนักถ่วง (loading matrix), $\hat{T}$ เป็นสัญลักษณ์ของเมริกซ์ของค่าของภาพฉายของเมทริกซ์ E บน $\hat{P}$ และ $\hat{q}$ เป็นตัวประมาณของเวกเตอร์น้ำหนักถ่วง (loading-vector)

ในรายละเอียดของขั้นตอนการสร้างปัจจัย a = 1,2,…, A ด้วยวิธีกำลังสองน้อยสุดแบ่งส่วน แสดงในขั้นตอน a,b,c และ d ดังนี้

a) เราใช้ข้อมูลส่วนที่เหลือของ f และประมาณเวกเตอร์แถว w_a โดยวิธีกำลังสองน้อยสุดจากตัวแบบของ

(2.3.1.1) $$E = fw_a + E_{new}$$

ผลลัพธ์ที่ได้จะมีขนาดเท่ากับ 1 โดยจะได้

(2.3.1.2) $$\hat{w}_a = Kf'E$$

เมื่อ K เป็นปัจจัยกำหนดขนาดที่ให้ขนาดของเวกเตอร์ $\hat{w}_a$ เท่ากับ 1

b) เราประมาณคะแนนขององค์ประกอบ $\underset{\sim}{t_a}$ เป็นเวกเตอร์แถวตั้ง โดยใช้วิธีกำลังสองน้อยสุดจากตัวประมาณ $\hat{w}_a$ ในตัวแบบ

$$E = f\hat{w}_a + E_{new}$$

ซึ่งจะได้ผลลัพธ์คือ

$$\hat{\underset{\sim}{t}}_a = E\,\hat{w}_a$$

เวกเตอร์น้ำหนักถ่วงตัวที่สอง $\hat{p}_a$ ประมาณได้จากตัวแบบ

$$E = \hat{\underset{\sim}{t}}_a\, p_a + E_{new}$$

ซึ่งจะได้ผลลัพธ์ คือ

$$\hat{p}_a = (\underset{\sim}{t_a}'\,\underset{\sim}{t_a})^{-1}\,\underset{\sim}{t_a}'\,E$$

c) เราประมาณเวกเตอร์น้ำหนักถ่วง q โดยใช้วิธีการถดถอยพหุคูณของ$(y-\bar{y})$บนค่าของ $\hat{T}\{\hat{t}_1,\hat{t}_2,\ldots\hat{t}_a\}$ จากตัวแบบ

(2.3.1.3)
$$f = \hat{\underset{\sim}{t}}_a q_a + E_{new}$$

และจะได้ผลลัพธ์ คือ

(2.3.1.4)
$$\hat{q}_a = (\underset{\sim}{t}_a{}' \underset{\sim}{t}_a)^{-1} \underset{\sim}{t}_a{}' f$$

d) คำนวณหาส่วนที่เหลือตัวใหม่

$$f = f - \hat{\underset{\sim}{t}}_a \hat{q}_a$$

$$E = E - \hat{\underset{\sim}{t}}_a \hat{p}_a$$

กลับไปที่ขั้นตอนที่ a) ถ้า a < A

จำนวนองค์ประกอบ A องค์ประกอบที่อยู่ในตัวแบบสามารถพิจารณาได้จากวิธีการทดสอบการพยากรณ์หรือที่เรียกว่าคลอส-วาลิดเดชัน (cross-validation) ซึ่งจะกล่าวรายละเอียดในหัวข้อ 2.3.4 และเราสามารถแปลงสัมประสิทธิ์การถดถอยขององค์ประกอบกลับไปเป็นสัมประสิทธิ์การถดถอยดั้งเดิมของตัวแปรอิสระ X ได้โดย

$$\hat{b} = \hat{P}' \hat{q}_a \quad \text{และ} \quad \hat{b}_0 = \bar{y} - \bar{x}\hat{b}$$

และเราสามารถพยากรณ์ค่าได้โดย

กำหนด x_i เป็นเวกเตอร์แถวของค่าตัวอย่างที่เพิ่มเข้ามาตัวที่ i และเราจะได้ค่าพยากรณ์ $\hat{y}_i$ จาก

$$\hat{y}_i = \hat{b}_0 + x_i \hat{b}$$

2.3.2 <u>ตัวแบบและวิธีการวิเคราะห์การถดถอยด้วยวิธีกำลังสองน้อยสุดแบ่งส่วน</u>

พิจารณาข้อมูลขนาดตัวอย่าง n ในการประมาณค่าความสัมพันธ์เชิงเส้นระหว่าง y และ $X_1, X_2, \ldots, X_p$ และข้อมูลตัวอย่างที่ i สามารถเขียนได้เป็น $X_1(i), X_2(i), \ldots, X_p(i)$ สำหรับ $i = 1,2,\ldots,n$

เวกเตอร์ค่าสังเกตของ y และ X_j เขียนแทนด้วย $\underset{\sim}{y}$ และ $\underset{\sim j}{x}$ โดยที่

$\underset{\sim}{y} = (y(1), y(2), \ldots, y(n))'$ และสำหรับ $j = 1,2,\ldots,p$; $\underset{\sim j}{x} = (x_j(1), x_j(2), \ldots, x_j(n))'$ และค่าเฉลี่ยตัวอย่างเขียนแทนด้วย $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y(i)$ และ $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_j(i)$ เราสามารถเขียนสมการ

การถดถอยได้ดังนี้

(2.3.2.1) $$\underset{\sim}{y} = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \ldots + \beta_p T_p$$

เมื่อแต่ละองค์ประกอบ T_k เป็นผลบวกเชิงเส้นของ x_j และสหสัมพันธ์ตัวอย่างสำหรับแต่ละคู่ขององค์ประกอบมีค่าเท่ากับศูนย์

สมการการถดถอยที่มีจำนวนพารามิเตอร์มากมีความซับซ้อนมากกว่าสมการการถดถอยที่มีจำนวนพารามิเตอร์น้อยกว่า เราจะเห็นได้ว่าองค์ประกอบในสมการการถดถอยข้างต้นจะมีจำนวนน้อยกว่าจำนวนของตัวแปร X อย่างมาก วิธีกำลังสองน้อยสุดมีจุดมุ่งหมายที่จะหลีกเลี่ยงการใช้สมการการถดถอยที่มีพารามิเตอร์มาก เราจึงได้พิจารณาความสัมพันธ์ระหว่างตัวแปร y และ X ที่เจาะจงบางตัว ในขณะที่ตัวแปรอิสระ X อื่นๆ จะไม่มีผลกระทบต่อการประมาณความสัมพันธ์นี้ แต่ตัวแปรอิสระเหล่านี้จะมีผลกระทบต่อองค์ประกอบ T_k

กำหนดให้ตัวแปร y_1^* และ X_{1j}^* เป็นตัวแปรที่ได้จากข้อมูลของ y และ X_j โดยที่ $y_1^* = y - \bar{y}$ และสำหรับ $j = 1,2,\ldots,p$

(2.3.2.2) $$X_{1j}^* = X_j - \bar{x}_j$$

โดยที่ค่าเฉลี่ยตัวอย่างของ y_1^* และ x_{1j}^* มีค่าเท่ากับศูนย์ และค่าข้อมูลของตัวแปรสามารถ เขียนได้ในรูป $y_1^* = y - \bar{y}.\underset{\sim}{1}$ และ $x_{1j}^* = x_j - \bar{x}_j.\underset{\sim}{1}$ เมื่อ $\underset{\sim}{1}$ เป็นเวกเตอร์ หน่วยที่มีขนาด $n \times 1$

วิธีกำลังสองน้อยสุดแบ่งส่วนจะพิจารณาการสร้างองค์ประกอบอย่างเป็นลำดับ องค์ประกอบแรก T_1 เป็นองค์ประกอบที่ใช้ในการพยากรณ์ค่า y_1^* และเป็นองค์ประกอบที่สร้างจากผลบวกเชิงเส้นของ X_{1j}^* โดยที่จะเราจะไม่นำสหสัมพันธ์ตัวอย่างระหว่างตัวแปรอิสระ X_{1j}^* มาพิจารณาในการสร้างองค์ประกอบ เราจะได้องค์ประกอบแรก T_1 โดยการถดถอยค่า y_1^* ด้วยตัวแปรถดถอยตัวแรก X_{11}^* และตัวถดถอยตัวต่อไปคือ X_{12}^* และตัวถดถอย X_{1j}^* แต่ละตัวต่อไปเรื่อยๆ สำหรับ $j = 1,2,\ldots,p$ และค่าเฉลี่ยตัวอย่างของตัวแปรอิสระแต่ละตัวมีค่าเป็น 0 เราจะได้ผลลัพธ์ของสมการการถดถอยด้วยวิธีกำลังสองน้อยสุดดังนี้

(2.3.2.3) $$\hat{y}_{1(j)}^* = b_{1j} X_{1j}^*$$

โดยที่ $b_{1j} = x_{1j}^{*'} y_1^* / (x_{1j}^{*'} x_{1j}^*)$ เราจะได้สมการการถดถอยในสมการ (2.3.2.3) p สมการที่ประมาณค่า y_1^* และในการรวมสมการประมาณค่าเหล่านี้โดยไม่คำนึงถึงสหสัมพันธ์ระหว่างตัวแปรอิสระ X_{1j}^* สามารถทำได้โดยใช้ค่าเฉลี่ยอย่างง่าย $\sum_j b_{1j} X_{1j}^* / p$ หรือให้อยู่ในรูปแบบทั่วไปมากขึ้นโดยใช้ค่าเฉลี่ยถ่วงน้ำหนัก เราจึงกำหนดองค์ประกอบ T_1 เป็นค่าเฉลี่ยถ่วงน้ำหนัก ดังนั้น

(2.3.2.4)
$$T_1 = \sum_{j=1}^{p} w_{1j} b_{1j} X^*_{ij}$$

เมื่อ $\sum_j w_{1j} = 1$ (การกำหนดค่าคงที่ $\sum_j w_{1j} = 1$ มีจุดมุ่งหมายเพื่อให้เป็นไปตามลักษณะของวิธีกำลังสองน้อยสุดแบ่งส่วนแต่ไม่จำเป็นเสมอไป เนื่องจากการที่นำค่าคงที่ใดๆ คูณกับองค์ประกอบ T_1 จะไม่มีผลกระทบต่าค่าขององค์ประกอบย่อยอื่นๆ หรือการพยากรณ์ค่าของ y) จากสมการที่ (2.3.2.4) ทำให้เกิดค่าพิสัย (range) ของความเป็นไปได้ในการสร้างองค์ประกอบ ซึ่งขึ้นอยู่กับน้ำหนักที่ใช้

ในขณะที่ T_1 เป็นค่าเฉลี่ยถ่วงน้ำหนักของตัวพยากรณ์ (predictors) ของ y^*_1 ซึ่งควรจะเป็นตัวพยากรณ์ที่สามารถใช้ในการพยากรณ์ค่าของ y^*_1 และ y ได้ดี แต่ยังมีตัวแปรอิสระ X ที่ยังให้ข้อมูลที่มีประโยชน์ในการพยากรณ์ค่า y ข้อมูลของ X^*_{1j} ที่ไม่ได้อยู่ใน T_1 สามารถประมาณค่าได้จากส่วนเหลือ (residual) ของการถดถอยของ X^*_j บน T_1 ซึ่งก็คือส่วนที่เหลือจากการถดถอยของ X^*_{1j} บน T_1 นั่นเอง ในทำนองเดียวกัน ความแปรผันใน Y ที่ไม่สามารถอธิบายได้ด้วย T_1 สามารถประมาณได้ด้วยส่วนที่เหลือจากการถดถอยของ y^*_1บน T_1 ซึ่งส่วนที่เหลือเหล่านี้สามารถเขียนแทนได้ด้วย X^*_{2j} เป็นส่วนที่เหลือของ X^*_{1j} และ y^*_2 เป็นส่วนที่เหลือของ y^*_1 ดังนั้น องค์ประกอบถัดไป, องค์ประกอบ T_2 เป็นผลบวกเชิงเส้นของ X^*_{2j} ซึ่งใช้ในการพยากรณ์ y^*_2 โดยสามารถสร้างองค์ประกอบด้วยวิธีเช่นเดียวกับ T_1 แต่แทนที่ X^*_{1j} และ y^*_1 ด้วย X^*_{2j} และ y^*_2 ตามลำดับ

ขั้นตอนการสร้างองค์ประกอบ $T_2, T_3, \ldots, T_p$ จึงเป็นการทำซ้ำด้วยวิธีอย่างเดียวกัน โดยที่แต่ละองค์ประกอบจะพิจารณาจากส่วนที่เหลือของการถดถอยก่อนหน้านี้ โดยคำนึงถึงความแปรผันส่วนที่เหลือของตัวแปรตาม y ที่สัมพันธ์กับข้อมูลที่เหลือของตัวแปรอิสระ X ซึ่งสามารถเขียนขั้นตอนอยู่ในรูปทั่วไปได้ดังนี้

กำหนดให้ $T_i (i \geq 1)$ เป็นองค์ประกอบตัวล่าสุดที่ถูกสร้างขึ้นจากตัวแปร y^*_i และ X^*_{ij} $(j = 1,2,\ldots,p)$ และ T_i, y^*_iและ X^*_{ij} มีค่าตัวอย่างเป็น t_i, y^*_iและ x^*_{ij} ซึ่งค่าเฉลี่ยตัวอย่างของทั้งสามตัวมีค่าเท่ากับศูนย์ ในการสร้างองค์ประกอบ T_{i+1} เราจะพิจารณาจากตัวแปร $X^*_{(i+1)j}$ และ y^*_{i+1} สำหรับ $j = 1,2,\ldots,p$, X^*_{ij} ถดถอยด้วย T_i ได้สัมประสิทธิ์การถดถอย คือ$\left(t_i{}' t_i\right)^{-1} t_i{}' x^*_{ij}$ ดังนั้นเราสามารถหา $X^*_{(i+1)j}$ ได้ดังนี้

(2.3.2.5)
$$X^*_{(i+1)j} = X^*_{ij} - \left\{\left(t_i{}' t_i\right)^{-1} t_i x^*_{ij}\right\} T_i$$

ค่าตัวอย่าง $x^*_{(i+1)j}$ เป็นส่วนที่เหลือจากการถดถอย ในทำนองเดียวกัน เราสามารถหา y^*_{i+1} ได้จาก $y^*_{(i+1)j} = y^*_{ij} - \left\{\left(t_i{}' t_i\right)^{-1} t_i y^*_{ij}\right\} T_i$ และ y^*_{i+1} เป็นส่วนที่เหลือของการถดถอยของ y^*_i ด้วย T_i

ความผันแปรส่วนที่เหลือในตัวแปรตาม y คือ y^*_{i+1} และความผันแปรส่วนที่เหลือในตัวแปรอิสระ X^*_j คือ $X^*_{(i+1)j}$ ดังนั้นในขั้นตอนต่อไปคือ การถดถอย y^*_{i+1} ด้วย $X^*_{(i+1)j}$ แต่ละตัว ซึ่งการถดถอยตัวที่ j จะได้ $b_{(i+1)}X^*_{(i+1)j}$ เป็นตัวพยากรณ์ของ y^*_{i+1} โดยที่

(2.3.2.6) $$b_{(i+1)} = \left(x^{*\prime}_{(i+1)j}\, x^*_{(i+1)j}\right)^{-1} x^*_{(i+1)j} y^*_{(i+1)}$$

สร้างรูปแบบของผลบวกเชิงเส้นของตัวพยากรณ์เหล่านี้ เราจะได้องค์ประกอบตัวถัดไปดังนี้

(2.3.2.7) $$T_{i+1} = \sum_{j=1}^{p} w_{(i+1)j} b_{(i+1)j} X^*_{(i+1)j}$$

เราจะทำซ้ำวิธีการข้างต้นทั้งหมดเพื่อหาองค์ประกอบ T_{i+2} และหลังจากได้องค์ประกอบตามที่ต้องการแล้ว แต่ละองค์ประกอบที่ได้จะมีความสัมพันธ์กับตัวแปรตาม y ในตัวแบบถดถอยที่ได้กำหนดในตอนต้น คือ $\underset{\sim}{y} = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \Lambda + \beta_p T_p$ ซึ่งสัมประสิทธิ์การถดถอยในตัวแบบนี้หาได้โดยวิธีกำลังสองน้อยสุด

2.3.3 <u>คุณสมบัติของวิธีกำลังสองน้อยสุดแบ่งส่วน</u>

ข้อดีของวิธีกำลังสองน้อยสุด คือ สหสัมพันธ์ตัวอย่างในแต่ละคู่ขององค์ประกอบมีค่าเป็นศูนย์ เนื่องจาก (a) ส่วนที่เหลือของการถดถอยไม่มีความสัมพันธ์กับตัวถดถอยตัวอย่าง เช่น $X^*_{(i+1)j}$ ไม่สัมพันธ์กับ T_i สำหรับทุกค่าของ j (b) แต่ละองค์ประกอบ T_{i+1}, K, T_p เป็นผลบวกเชิงเส้นของ $X^*_{(i+1)j}, \ldots, X^*_{pj}$ ตามลำดับ ดังนั้นจากข้อ (a) องค์ประกอบเหล่านี้จึงไม่สัมพันธ์กับ T_i และผลจากการที่องค์ประกอบเหล่านี้ไม่มีความสัมพันธ์กันทำให้สัมประสิทธิ์การถดถอยจากตัวแบบถดถอยข้างต้นสามารถประมาณได้โดยการถดถอยอย่างง่ายที่มีตัวแปรอิสระเพียงตัวเดียว (simple regression) โดยที่สัมประสิทธิ์การถดถอย $\hat{\beta}_i$ ได้จากการถดถอย y ด้วย T_i และถ้าเราเพิ่มองค์ประกอบจำนวนหนึ่งเข้าไปในตัวแบบ ค่าของสัมประสิทธิ์การถดถอยขององค์ประกอบที่อยู่ในตัวแบบตัวก่อนๆ ก็จะไม่เปลี่ยนแปลงด้วยเช่นเดียวกัน ผลต่อมาจากการที่องค์ประกอบไม่มีความสัมพันธ์ คือ เราสามารถอธิบายความหมายของ y^*_{i+1} และ $X^*_{(i+1)j}$ ได้ว่า y^*_{i+1} และ $x^*_{(i+1)j}$ เป็นเวกเตอร์ของส่วนที่เหลือจากการถดถอยของ y และ X^*_j ด้วย T_1, K, T_i

2.3.4 <u>การหาจำนวนองค์ประกอบที่ใช้ในตัวแบบถดถอย</u>

การตัดสินใจว่าจะใช้จำนวนองค์ประกอบเท่าไหร่ในตัวแบบถดถอยเป็นเรื่องที่สำคัญมาก วิธีหนึ่งที่ใช้ คือ วิธีครอส-วาลิดเดชัน (cross-validation) ซึ่งเป็นวิธีที่ถูกนำมาใช้ในเอกสารของสโตน (Stone) และ บรู๊คส์ (Brooks) ในปีค.ศ.1990 และ โวล์ด เอ็ต. ออล. (Wold et. Al.) ในปี ค.ศ. 1984

เอ็ม. สโตน (M. Stone) ในปีค.ศ.1974 ได้เสนอวิธีคลอส-วาลิคเดชันเป็นเกณฑ์ในการเลือก (choices) จำนวนตัวแปรหรือองค์ประกอบ และการประเมินความใช้ได้ (assessment or validation) ของการพยากรณ์ที่ใช้ในวิธีการวิเคราะห์ข้อมูลที่เกี่ยวข้องกับตัวพยากรณ์ เช่น การประมาณค่าตัวแปรเดียว การถดถอยเชิงเส้น และการวิเคราะห์ความแปรปรวน เป็นต้น หลักการของการพยากรณ์ค่าจะทำการแบ่งขนาดตัวอย่างออกเป็น 2 กลุ่ม คือ กลุ่มตัวอย่างย่อยที่ใช้สร้างตัวแบบ (construction subsample) และกลุ่มตัวอย่างย่อยที่ใช้ในการพยากรณ์ (validation subsample) ซึ่งมีหลายวิธี กรอบของทฤษฎีโดยทั่วไปมีดังนี้

สมมติให้ S เป็นเซตของข้อมูล (x,y) ที่มีขนาดตัวอย่างเท่ากัน n โดยที่ x และ y เป็นข้อมูลทั่วไป ดังนั้นเราสามารถเขียนได้เป็น

$$S = \{(x_i, y_i) | i = 1,2,\ldots,n\}$$

พิจารณาข้อมูลตัวอย่างกลุ่มใหม่ที่มีเฉพาะค่า x และต้องการพยากรณค่า y โดยใช้ค่า $\hat{y}$ ซึ่งเป็นฟังก์ชันของ x และ S เราจะสมมติให้กลุ่มของตัวพยากรณ์ (class of predictors) ในขั้นเริ่มต้นเป็น

(2.3.4.1) $$\{\hat{y}(x;\alpha,S) | \alpha \in \mathscr{A}\}$$

โดยที่ $\hat{y}(x;\alpha,S)$ มีความสัมพันธ์ในเชิงพยากรณ์ซึ่งขึ้นอยู่กับ x และ S ซึ่งค่าของ $\hat{y}$ ที่มีความสัมพันธ์ขึ้นอยู่กับ S นั้นอาจได้จากขนาดตัวอย่างเจาะจง n, n-1 และ n-2 ในขณะที่ $\mathscr{A}$ ไม่ขึ้นอยู่กับขนาดตัวอย่าง n เราจะได้ทางเลือกในสมการที่ (2.3.4.1) คือการพิจารณาค่า α จากเซตของ S ดังนั้นทางเลือกของ α จึงเป็นฟังก์ชันของ S ซึ่งสามารถอธิบายว่าเป็น "การประมาณค่า α "

<u>ตัวอย่าง 2.3.1</u> ถ้า x และ y เป็นจำนวนจริง เราอาจจะได้ตัวพยากรณ์ ดังนี้

(2.3.4.2) $$\hat{y}(x;\alpha,S) = \bar{y} + \alpha \cdot \left\{ \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} \right\} (x_i - \bar{x})$$

ซึ่ง $\mathscr{A} = [0,1]$

เราปรับปรุงวิธีการของทางเลือกที่มีความเหมาะสมที่จะใช้ทางเลือกแบบคลอส-วาลิดโทรี (cross-validtory choice) ของ α และ วิธีการของการประเมินความใช้ได้ของตัวพยากรณ์ (cross-validtory assess- ment) ของทางเลือกเหล่านี้ มีรายละเอียดดังนี้

1. ทางเลือกอย่างง่าย (naive choice) ของ α คือ ค่า $\alpha^0(S) \in \mathscr{A}$ ที่ทำให้

$$\bar{L}(\alpha) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} L\left[y_i, \hat{y}(x_i;\alpha,S)\right]$$ มีค่าน้อยที่สุด

โดยที่ $L\left[y, \hat{y}\right]$ เป็นฟังก์ชันความสูญเสียของ $\hat{y}$ ที่เป็นค่าพยากรณ์ค่าจริงของ y

<u>ตัวอย่าง 2.3.4.2</u> จากตัวอย่างที่ 2.3.4.1 เราจะได้ $L\left[y, \hat{y}\right] = (y - \hat{y})^2$ และ $\alpha^0(S) = 1$ จะให้ค่าฟังก์ชันความสูญเสียน้อยที่สุด ซึ่งสอดคล้องกับการเลือกความเหมาะสมของตัวพยากรณ์โดยใช้วิธีกำลังสองน้อยสุดนั่นเอง

2. การประเมินความใช้ได้อย่างง่าย (naive assessment) ของทางเลือกอย่างง่าย (naive choice) ในข้อ 1 เรากำหนดให้ $\bar{L}(\alpha^0(S))$ เป็นค่าเฉลี่ย n ตัวของข้อมูลในเซต S ของ

$$L\left[y,\hat{y}(x_i;\alpha^0(S),S)\right]$$

<u>ตัวอย่าง 2.3.4.3</u> สำหรับตัวอย่าง 2.3.4.1 และค่ากำลังสองของ L เราจะได้ $\bar{L}(\alpha^0(S))$ มีค่าเท่ากับ SSE / n เมื่อ SSE เป็นผลบวกกำลังสองของความผิดพลาด (residual sum of squares)

3. การประเมินความใช้ได้โดยใช้คลอส-วาลิดโทรี่ (cross-validtory assessment) ของทางเลือกอย่างง่าย

กำหนดให้

$$C^0 \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} L\left[y_i,\hat{y}(x_i;\alpha^0(S_{-i}),S_{-i})\right]$$

เมื่อ S_{-i} คือ เซตของข้อมูลตัวอย่างที่ตัดข้อมูลตัวที่ i ทิ้งไป และ $\alpha^0(S_{-i})$ เป็นทางเลือกอย่างง่ายของ α ที่ขึ้นอยู่กับ S_{-i} นั่นคือ ค่าของ α ที่ทำให้

$$\bar{L}_{-i}(\alpha) \overset{def}{=} \frac{1}{(n-1)}\sum\nolimits_{-i} L\left[y_j,\hat{y}(x_j;\alpha,S_{-i})\right]$$ มีค่าน้อยที่สุด

เมื่อ Σ_{-i} คือ ผลบวกที่ไม่รวมค่าของข้อมูลตัวที่ i C^0 เกี่ยวข้องกับการพยากรณ์ค่าด้วยขนาดตัวอย่าง $n - 1$

4. ทางเลือกแบบคลอส-วาลิดโทรี่ (cross-validtory choice) ของ α คือ ค่า $\alpha^+(S) \in \mathscr{A}$ ที่ทำให้

$$C(\alpha) \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} L\left[y_{-i},\hat{y}(x_i;\alpha,S_{-i})\right]$$

5. การประเมินความใช้ได้ของการพยากรณ์แบบคลอส-วาลิดโทรี่ (cross-validtory assessment) ของทางเลือกแบบคลอสวาลิดโทรี่ (cross-validtory choice) เรากำหนดให้

$$C^+(\alpha) \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} L\left[y_i,\hat{y}(x_i;\alpha^+(S_{-i}),(,S_{-i})\right]$$

เมื่อ $\alpha^+(S_{-i})$ เป็นทางเลือกแบบคลอส-วาลิดโทรี่ของ α บนพื้นฐานของ S_{-i} นั่นคือ ค่าของ α ที่ทำให้

$$C_{-i}(\alpha) \stackrel{def}{=} \frac{1}{(n-1)} \sum_{-i} L\left[y_j, \hat{y}(x_j; \alpha, (S_{-ij})\right]$$

โดยที่ S_{-ij} คือ เซตของข้อมูลตัวอย่างที่ตัดข้อมูลตัวที่ i และ ตัวที่ j ทิ้งไป

6. ทางเลือกแบบคลอส-วาลิดโทรี่สองขั้นตอนของ α เราเรียกว่า ดับเบิ้ลคลอส (double-cross) พิจารณาในกรณีที่ α มี 2 องค์ประกอบคือ $\alpha = (a,b)$ โดยที่ $a \in A$ และ $b \in B(a)$ กำหนด $b^{+}(S_{-i},a)$ คือ ค่าของ b ที่ทำให้

$$\frac{1}{(n-1)} \sum_{-i} L\left[y_j, \hat{y}(x_j; (a,b), S_{-ij})\right] \quad \text{มีค่าน้อยที่สุด}$$

ดังนั้น $a^{++}(S)$ คือ ค่าของ a ที่ทำให้

$$C^{+}(a) \stackrel{def}{=} \frac{1}{n} \sum_{i=1}^{n} L\left[y_i, \hat{y}(x_i; a, b^{+}(S_{-i}, a)), (S_{-i})\right] \quad \text{มีค่าน้อยที่สุด}$$

ในขั้นตอนสุดท้ายเลือก $b = b^{++}(S)$ ใน $B(a^{++}(S))$ ที่ทำให้

$$\frac{1}{n} \sum_{i=1}^{n} L\left[y_i, \hat{y}(x_i; a^{++}(S), b), S_{-i})\right] \quad \text{มีค่าน้อยที่สุด}$$

เราสามารถเขียนได้ว่า $\alpha^{++}(S) = (a^{++}(S), b^{++}(S))$ และเรียกว่าเป็นทางเลือกแบบคลอส-วาลิดโทรี่ 2 ขั้นตอนของ α หรือ "ดับเบิ้ลคลอส" นั่นเอง การประเมินความใช้ได้แบบคลอส-วาลิดโทรี่ของ $\alpha^{++}(S)$ เป็นเช่นเดียวกับของ $\alpha^{+}(S)$ โดยคำนวณ

$$C^{++}(\alpha) \stackrel{def}{=} \frac{1}{n} \sum_{i=1}^{n} L\left[y_i, \hat{y}(x_i; \alpha^{++}(S_{-i}), S_{-i})\right]$$

เมื่อ $\alpha^{++}(S_{-i})$ ได้จากข้อมูล S_{-i} เช่นเดียวกับ $\alpha^{++}(S)$ ที่ได้จาก S

ในการวิจัยครั้งนี้ผู้วิจัยได้นำวิธีการของคลอส-วาลิดเดชันมาใช้ในการหาจำนวนองค์ประกอบที่ใช้ในสมการการถดถอยของวิธีกำลังสองน้อยสุดแบ่งส่วน ซึ่งสามารถอธิบายเป็นขั้นตอนง่ายๆได้โดยจะทำการแบ่งข้อมูลที่ใช้ในการสร้างตัวแบบออกเป็น 3 กลุ่มย่อย ข้อมูล 2 กลุ่มย่อยจะนำมาใช้ในการสร้างตัวแบบ ส่วนข้อมูล 1 กลุ่มย่อยที่เหลือจะนำมาหาค่าพยากรณ์ y แล้วหาผลรวมของความคลาดเคลื่อนกำลังสอง ทำซ้ำจนทุกกลุ่มย่อยได้ถูกตัดทิ้งไปแล้วหาผลรวมทั้งหมดของความคลาดเคลื่อนกำลังสองของทุกกลุ่ม องค์ประกอบตัวถัดไปจะถูกพิจารณาเพิ่มเข้าในสมการก็ต่อเมื่อให้ผลรวมทั้งหมดของความคลาดเคลื่อนกำลังสองของทุกกลุ่มน้อยลงและในทำนองเดียวกันจะหยุดทำการเพิ่มองค์ประกอบเมื่อองค์ประกอบนั้นให้ผลรวมทั้งหมดของความคลาดเคลื่อนกำลังสองของทุกกลุ่มมากขึ้น

2.4 ดัชนีพหุสัมพันธ์ (Degree of multicollinearlity)

2.4.1 สหสัมพันธ์อย่างง่ายระหว่างตัวแปรอิสระ (Simple correlation among the - regression variables)

พิจารณาเมทริกซ์สหสัมพันธ์ของ $X^{*'}X^{*}$ โดยที่สมาชิกแถวที่ i และแนวตั้งที่ j ของเมทริกซ์ $X^{*'}X^{*}$ คือ $X_{ij}^{*} = X_{ij} - \bar{x}$ สมาชิกแต่ละตัวของเมทริกซ์สหสัมพันธ์ $X^{*'}X^{*}$ แสดงถึงสหสัมพันธ์อย่างง่ายของตัวแปรอิสระ X_i และ X_j ถ้าสหสัมพันธ์อย่างง่ายของตัวแปรอิสระ X_i และ X_j ใดๆมีค่าสูงมาก แสดงว่าตัวแปรอิสระมีแนวโน้มที่จะมีพหุสัมพันธ์ อย่างไรก็ตามการพิจารณาด้วยวิธีนี้ไม่ได้กำหนดค่าสหสัมพันธ์อย่างง่ายของตัวแปรอิสระแต่ละคู่ว่าควรเป็นเท่าไร จึงจะเกิดพหุสัมพันธ์

2.4.2 Variance Inflation Factors (VIF)

$$VIF_i = \frac{1}{1-R_i^2} \quad \text{และ} \quad \overline{VIF} = \sum_{i=1}^{p} VIF_i$$

โดยที่ R_i เป็นสัมประสิทธิ์สหสัมพันธ์พหุของตัวแปรอิสระ X_i และตัวแปรอิสระอื่นๆ

ถ้า R_i^2 มีค่าใกล้ 1 ค่า VIF จะมีค่ามาก ซึ่งจะเกิดขึ้นในกรณีที่ตัวแปรอิสระ X_i มีความสัมพันธ์เชิงเส้นกับตัวแปรอิสระอื่นๆ ที่เหลือ โดยทั่วไปถือว่า ถ้า VIF มีค่าเกิน 10 ก็จะมีเหตุผลพอที่จะสรุปได้ว่าเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ ซึ่งควรจะตัดตัวแปรอิสระบางตัวออกหรือเลือกตัวประมาณอื่นแทนที่ตัวประมาณโดยวิธีกำลังสองน้อยที่สุด

2.4.3 ค่าเจาะจงของ $X'X$

ค่าเจาะจง (eigenvalues) เวกเตอร์เจาะจง (eigenvector) ของเมทริกซ์สหสัมพันธ์มีบทบาทสำคัญในการพิจารณาพหุสัมพันธ์ของตัวแปรอิสระในการวิเคราะห์ความถดถอย ถ้าค่าเจาะจงที่เล็กที่สุดมีค่าเข้าใกล้ศูนย์ แสดงว่าตัวแปรอิสระมีความสัมพันธ์กันเชิงเส้นอย่างมาก พหุสัมพันธ์สามารถวัดได้ในเทอมของอัตราส่วนระหว่างค่าเจาะจงที่ใหญ่ที่สุดและค่าเจาะจงที่เล็กที่สุด $\phi = \frac{\lambda_{\max}}{\lambda_{\min}}$ โดยที่ λ คือ ค่าเจาะจงของเมทริกซ์ และเรียก ϕ ว่าตัวเลขเงื่อนไข (condition number) ซึ่งถ้า ϕ มีค่ามากเกินกว่า 100 แสดงว่าเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ และถ้ามีค่ามากกว่า 1,000 แสดงว่าเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระมาก