บทที่ 2

ตัวสถิติและผลงานวิจัยที่เกี่ยวข้อง

บทนี้จะกล่าวถึงการประมาณค่าสัมประสิทธิ์การถดถอยเชิงเส้นพหุคูณของ 3 วิธีได้แก่ วิธีกำลังสองน้อยสุด (Ordinary Least Square method) วิธีที่ได้จากสมการการถดถอยริดจ์โดยใช้วิธีของบลีแมน (Ridge Regression by Breiman method) และวิธีสมการถดถอยเชิงเส้นการ์ล็อต (Garrote Linear Regression method) ซึ่งจะศึกษาในกรณีที่ความคลาดเคลื่อนมีการแจกแจงแบบปกติ ปกติปลอมปน ไวบูลล์ และลอกนอร์มอล ซึ่งมีรายละเอียดต่าง ๆ ดังนี้

การประมาณค่าสัมประสิทธิ์ด้วยวิธีกำลังสองน้อยสุด (OLS)

ตัวแบบเชิงเส้นโดยทั่วไปของการวิเคราะห์การถดถอยเชิงเส้นพหุคูณสามารถเขียนได้ดังนี้

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

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

X เป็นเมทริกซ์ของตัวแปรอิสระขนาด n x (p+1)

$\underset{\sim}{\beta}$ เป็นเวกเตอร์ของสัมประสิทธิ์การถดถอยขนาด (p+1) x 1

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

n เป็นขนาดตัวอย่างของตัวแปรอิสระ

และ p เป็นจำนวนตัวแปรอิสระ

โดยที่ $E(\underset{\sim}{\varepsilon}) = 0$ และ $Cov(\underset{\sim}{\varepsilon}) = \sigma^2 I_n$

จากตัวแบบในสมการที่(2.1)จะได้ผลรวมความคลาดเคลื่อนกำลังสองอยู่ในรูปแบบดังนี้

(2.2) $$SSE(\underset{\sim}{\beta}) = (\underset{\sim}{y} - X\underset{\sim}{\beta})'(\underset{\sim}{y} - X\underset{\sim}{\beta})$$

วิธีกำลังสองน้อยสุดมีหลักการคือทำให้ผลรวมความคลาดเคลื่อนกำลังสองมีค่าต่ำสุด โดยจะได้

(2.3) $$(X'X)\underset{\sim}{\hat{\beta}} = X'\underset{\sim}{y}$$

จากสมการ (2.3) เป็นสมการปกติ (normal equation) และจะได้ตัวประมาณสัมประสิทธิ์การถดถอยพหุคูณของวิธีกำลังสองน้อยสุดอยู่ในรูปของ

(2.4) $$\hat{\beta} = (X'X)^{-1}X'y$$

จากตัวประมาณสัมประสิทธิ์การถดถอยจากสมการ (2.4) มีคุณสมบัติเป็นตัวประมาณที่ไม่เอนเอียงของ β กล่าวคือ

$$\begin{aligned} E(\hat{\beta}) &= E\left[(X'X)^{-1}X'y\right] \\ &= (X'X)^{-1}X'E(y) \\ &= (X'X)^{-1}X'X\beta \\ &= \beta \end{aligned}$$

ซึ่งตัวประมาณ β ที่ได้เป็นตัวประมาณไม่เอนเอียงและมีค่าเฉลี่ยความคลาดเคลื่อนกำลังสองน้อยสุด แต่ในการประมาณ β ด้วยวิธีกำลังสองน้อยสุด มีข้อกำหนดข้อหนึ่งคือ ตัวแปรอิสระจะต้องไม่มีความสัมพันธ์กัน แต่ในทางปฏิบัติเป็นไปได้น้อยมากจึงทำให้สมาชิกของเมทริกซ์สหสัมพันธ์ของตัวแปรอิสระเกิดเงื่อนไขที่ไม่ดี (ill-condition) ซึ่งทำให้การประมาณค่า β ด้วยวิธีกำลังสองน้อยสุดไม่สามารถให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองมีค่าต่ำสุดได้ ดังนั้นจึงควรพิจารณาคุณสมบัติของตัวประมาณค่า β สองประการ คือ เมทริกซ์ความแปรปรวนร่วมของตัวประมาณ $\hat{\beta}$ และ ค่าเฉลี่ยยกกำลังสองของความแตกต่างระหว่าง $\hat{\beta}$ และ β ซึ่งสามารถเขียนให้อยู่ในรูปของเมทริกซ์ $X'X$ และ σ^2 ดังต่อไปนี้

กำหนดให้ความแปรปรวนร่วมของตัวประมาณ $\hat{\beta}$ มีค่าเป็น $\text{cov}(\hat{\beta})$ ดังนี้

(2.5) $$\text{cov}(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

ให้ $d(\hat{\beta},\beta)$ เป็นระยะทางระหว่าง $\hat{\beta}$ และ β

(2.6) ดังนั้น $$\left\{d(\hat{\beta},\beta)\right\}^2 = (\hat{\beta}-\beta)'(\hat{\beta}-\beta)$$

เพราะฉะนั้นค่าเฉลี่ยยกกำลังสองของระยะทางระหว่าง $\hat{\beta}$ และ β จะอยู่ในรูปของ

(2.7)
$$E\left[d(\hat{\beta},\beta)\right]^2 = \sigma^2 trace(\mathrm{X}'\mathrm{X})^{-1}$$
$$= E\left[(\hat{\beta}-\beta)'(\hat{\beta}-\beta)\right]$$
$$= E\left[\hat{\beta}'\hat{\beta}-2\hat{\beta}\beta'+\beta'\beta\right]$$
$$= E\left[\hat{\beta}'\hat{\beta}\right]-\beta'\beta$$

ค่าของ $E\left[d(\hat{\beta},\beta)\right]^2$ จะมีค่าสมมูล (equivalent) กับค่าเฉลี่ยของ $\hat{\beta}$ และ β ซึ่งสามารถเขียนให้อยู่ในรูปของ

(2.8)
$$E\left[\hat{\beta}'\hat{\beta}\right] = \beta'\beta+\sigma^2 trace(\mathrm{X}'\mathrm{X})^{-1}$$

เมื่อ ε มีการแจกแจงแบบปกติจะได้ว่า

(2.9)
$$Var\left(d(\hat{\beta},\beta)\right)^2 = 2\sigma^4 trace(\mathrm{X}'\mathrm{X})^{-2}$$

จะเห็นได้ว่าจากสมการที่ (2.5),(2.7) และ (2.9) ต่างก็อยู่ในรูปของฟังก์ชันเมทริกซ์ $\mathrm{X}'\mathrm{X}$ ดังนั้นเพื่อความสะดวกในการเปรียบเทียบการประมาณค่า β จึงควรแปลงค่าของเมทริกซ์ $(\mathrm{X}'\mathrm{X})^{-1}$ ให้อยู่ในรูปฟังก์ชันของค่าเฉพาะ (eigenvalue) ของเมทริกซ์ $\mathrm{X}'\mathrm{X}$ โดยใช้คุณสมบัติข้อหนึ่งของค่าเฉพาะในเมทริกซ์ $\mathrm{X}'\mathrm{X}$ กล่าวคือค่า λ_i เป็นค่าเฉพาะของเมทริกซ์ $\mathrm{X}'\mathrm{X}$ เมื่อ i = 1,...,p และ $\sum_{i=1}^{p}\lambda_i = trace(\mathrm{X}'\mathrm{X})$ เมื่อ p เป็นจำนวนตัวแปรอิสระ

สมมติให้ค่าเฉพาะของเมทริกซ์ $\mathrm{X}'\mathrm{X}$ มีค่าเป็น

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

จากสมการ (2.7) ค่าเฉลี่ยยกกำลังสองของระยะทางระหว่าง $\hat{\beta}$ และ β อาจเขียนอยู่ในรูปฟังก์ชันของค่าเฉพาะของเมทริกซ์ $\mathrm{X}'\mathrm{X}$ ดังนี้

(2.10)
$$E\left[d(\hat{\beta},\beta)\right]^2 = \sigma^2 \sum_{i=1}^{p}\left(\frac{1}{\lambda_i}\right)$$

จากสมการ (2.9) ค่าความแปรปรวนยกกำลังสองของระยะทางระหว่าง $\hat{\beta}$ และ β อาจเขียนอยู่ในรูปฟังก์ชันของค่าเฉพาะของเมทริกซ์ ดังนี้

(2.11) $$Var\left(d(\hat{\beta},\beta)\right)^2 = 2\sigma^4 \sum_{i=1}^{p}\left(\frac{1}{\lambda_i}\right)^2$$

แต่ในกรณีที่ตัวแปรอิสระมีสภาพไม่เหมาะสม ค่าเฉพาะของเมทริกซ์ $X'X$ จะมีค่ามาก ทำให้ระยะทางจาก $\hat{\beta}$ ไปยัง β มีค่ามาก ดังนั้นค่าจากสมการ (2.10) และ (2.11) ซึ่งก็คือค่า $E\left[d(\hat{\beta},\beta)\right]^2$ และ $Var\left(d(\hat{\beta},\beta)\right)^2$ จะมีค่าสูงตามไปด้วย

การประมาณค่าสัมประสิทธิ์การถดถอยพหุคูณด้วยวิธีริดจ์รีเกรชชัน(RID)

ในปี ค.ศ.1970 โฮเอิล(Hoel) และเคนนาร์ด(Kennard) ได้เสนอวิธีการริดจ์รีเกรชชันเพื่อแก้ปัญหาการเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ ซึ่งวิธีนี้จะลดค่าเฉลี่ยความคลาดเคลื่อนกำลังสองให้มีค่าต่ำลง จากหลักการของวิธีกำลังสองน้อยสุดพบว่าถ้าต้องการลดค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต้องทำให้เมทริกซ์ $X'X$ มีค่าต่ำลงหรือก็คือต้องทำให้ $|X'X|$ มีค่าเพิ่มขึ้นโดยบวกค่าคงที่ k ซึ่งมากกว่าศูนย์เข้ากับสมาชิกทุกตัวบนเส้นทแยงมุมของเมทริกซ์ $X'X$ ดังนั้นตัวประมาณสัมประสิทธิ์การถดถอยพหุคูณด้วยวิธีริดจ์รีเกรชชันอยู่ในรูปดังต่อไปนี้

(2.12) $$\hat{\beta}_R = (X'X + kI)^{-1}X'y \qquad ; k > 0$$

ให้ $(X'X + kI)^{-1} = W$ ดังนั้นจากสมการ (2.12) จะได้ว่า

(2.13) $$\hat{\beta}_R = W X'y$$

หรือสามารถเขียนให้อยู่ในรูปฟังก์ชันตัวประมาณสัมประสิทธิ์โดยวิธีกำลังสองน้อยสุด ($\hat{\beta}$) เพื่อสะดวกในการเปรียบเทียบวิธีการประมาณค่าสัมประสิทธิ์ ดังนี้

จากสมการที่ (2.13) อาจเขียนได้ว่า

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

(2.14) $$= Z\hat{\beta}$$

โดยที่ $Z = [\mathrm{I} + k(\mathrm{X}'\mathrm{X})^{-1}]^{-1}$ และ $W = (\mathrm{X}'\mathrm{X} + k\mathrm{I})^{-1}$ ค่า $\hat{\beta}_R$,Z และ W เหล่านี้จะนำไปพิจารณาเปรียบเทียบคุณสมบัติต่าง ๆ ในการประมาณค่าสัมประสิทธิ์ของวิธีริดจ์รีเกรชชัน คุณสมบัติที่จำเป็นของ $\hat{\beta}_R$,Z และ W มีดังนี้

1) ให้ $E_i(W)$ และ $E_i(Z)$ เป็นค่าเฉพาะของ W และ Z ตามลำดับ ซึ่งคำนวณได้จากสมการเฉพาะ (characteristic equation)

$$|W - E_i\mathrm{I}| = 0$$

และ $$|Z - E_i\mathrm{I}| = 0$$

จะได้ค่าเจาะจงของ W และ Z ดังนี้

(2.15) $$E_i(W) = \frac{1}{(\lambda_i + k)}$$

และ

(2.16) $$E_i(Z) = \frac{\lambda_i}{(\lambda_i + k)}$$

2) เราอาจเขียน Z ให้อยู่ในรูปฟังก์ชันของ W ได้ดังนี้

$$Z = [\mathrm{I} + k(\mathrm{X}'\mathrm{X})^{-1}]^{-1}$$
$$= \mathrm{I} - kW$$

3) ตัวประมาณ $\hat{\beta}_R$ จากสมการ (2.14) จะมีค่าน้อยกว่าตัวประมาณ $\hat{\beta}$ เมื่อ k > 0 กล่าวคือ

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

จากนิยามที่ว่า $\hat{\beta}_R = Z\hat{\beta}$ โดยที่เมทริกซ์ $\mathrm{X}'\mathrm{X}$ และ Z มีคุณสมบัติเป็นบวกแน่นอน (symmetric positive definition) จะได้ว่า

$$\hat{\beta}_R' \hat{\beta}_R = (Z\hat{\beta})'(Z\hat{\beta})$$
$$= \sum_{i=1}^{p} E_i^2(Z)\hat{\beta}'\hat{\beta}$$
$$< E_{i\,(\max)}^2(Z)\hat{\beta}'\hat{\beta}$$

เมื่อ $E_{\max}(Z) = \frac{\lambda_{\max}}{\lambda_{\max}+k}$ ซึ่ง $\lambda_{\max}$ เป็นค่าเฉพาะที่มีค่ามากที่สุดของเมทริกซ์ $X'X$ ดังนั้น จากคุณสมบัติของ Z ในข้อที่ 2) พบว่าเมื่อค่า k เท่ากับศูนย์แล้ว Z = I และค่า Z จะมีค่าเข้าใกล้ศูนย์เมื่อค่า $k \to \infty$ ดังนั้น

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

เนื่องจากผลบวกความคลาดเคลื่อนยกกำลังสองของตัวประมาณ $\hat{\beta}_R$ เป็นดังนี้

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

จากการประมาณค่าพารามิเตอร์ β ด้วยวิธีกำลังสองน้อยสุดเราสามารถเขียน $SSE(\hat{\beta})$ ซึ่งเป็นผลบวกของความคลาดเคลื่อนกำลังสองของตัวประมาณ $\hat{\beta}$ เป็น

(2.17) $$SSE(\hat{\beta}) = y'y - \hat{\beta}'X'y$$

เพื่อให้เห็นความแตกต่างระหว่าง $\hat{\beta}_R$ และ $\hat{\beta}$ สามารถเขียน $SSE(\hat{\beta}_R)$ ในเทอมของ $y'y$ และ $X'y$ ได้ดังนี้

(2.18) $$SSE(\hat{\beta}_R) = y'y - \hat{\beta}'X'y - k\hat{\beta}_R'\hat{\beta}_R$$

จากสมการที่ (2.17) และ (2.18) สามารถสรุปได้ว่าผลบวกความคลาดเคลื่อนยกกำลังสองของวิธีริดจ์รีเกรชชันให้ค่าน้อยกว่าผลบวกความคลาดเคลื่อนยกกำลังสองของวิธีกำลังสองน้อยสุด โดยขึ้นอยู่กับค่า k ว่าควรจะมีค่าเหมาะสมที่สุดจึงจะทำให้ค่าความคลาดเคลื่อนกำลังสองต่ำสุด

คุณสมบัติของค่าเฉลี่ยผลบวกความคลาดเคลื่อนยกกำลังสองของริดจ์รีเกรชชัน

จากการประมาณค่าสัมประสิทธิ์การถดถอย $\hat{\beta}$ ด้วยวิธีกำลังสองน้อยสุด จะได้ว่าค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองดังนี้

(2.19) $$MSE(\hat{\beta}) = Var(\hat{\beta}) + bias^2(\hat{\beta})$$

เนื่องจาก $E(\hat{\beta}) = \beta$ การประมาณค่าพารามิเตอร์ β ด้วยวิธีกำลังสองน้อยสุดได้ตัวประมาณ β ซึ่งเป็นตัวประมาณไม่เอนเอียงแต่ในการประมาณค่าพารามิเตอร์ $\hat{\beta}_R$ ด้วยวิธีริดจ์รีเกรชชั่น ตัวประมาณที่ได้เป็นตัวประมาณที่เอนเอียง ซึ่งก็คือ $E(\hat{\beta}_R) = Z\hat{\beta}$

สำหรับค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองของตัวประมาณสัมประสิทธิ์การถดถอยอาจพิจารณาได้ดังนี้

$$
\begin{aligned}
E\left[d(\hat{\beta}_R, \beta)\right]^2 &= E\left[(\hat{\beta}_R - \beta)'(\hat{\beta}_R - \beta)\right] \\
&= E\left[(\hat{\beta} - \beta)' Z'Z(\hat{\beta} - \beta)\right] + (Z\beta - \beta)'(Z\beta - \beta) \\
&= \sigma^2 trace(\mathrm{X}'\mathrm{X})\, Z'Z + \beta'(Z - \mathrm{I})'(Z - \mathrm{I})\beta \\
&= \sigma^2[trace(\mathrm{X}'\mathrm{X} + k\mathrm{I})^{-1} - k\, trace(\mathrm{X}'\mathrm{X} + k\mathrm{I})^{-2}] \\
&\quad + k^2 \beta'(\mathrm{X}'\mathrm{X} + k\mathrm{I})^{-2}\beta \\
&= \sigma^2 \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i - k)^2} + k^2 \beta'(\mathrm{X}'\mathrm{X} + k\mathrm{I})^{-2}\beta \\
&= \gamma_1(k) + \gamma_2(k) \qquad (2.20)
\end{aligned}
$$

จะเห็นว่า $E\left[d(\hat{\beta}_R, \beta)\right]^2$ เป็นค่าเฉลี่ยความคลาดเคลื่อนยกกำลังสองซึ่งอยู่ในรูปผลบวกของฟังก์ชัน $\gamma_1(k)$ และ $\gamma_2(k)$

โดยที่เทอมแรก $\gamma_1(k)$ เป็นค่าความแปรปรวนของตัวประมาณ $\hat{\beta}_R$ และเทอมที่สอง $\gamma_2(k)$ เป็นระยะทางจาก $Z\beta$ ไปยัง β โดยจะมีค่าเท่ากับศูนย์เมื่อ k=0 ซึ่ง Z เป็นเมทริกซ์เอกลักษณ์การคูณซึ่งสอดคล้องกับค่าความเอนเอียงกำลังสองของ $\hat{\beta}$ ที่มีค่าเป็นศูนย์

สำหรับกรณีที่ k>0 อาจพิจารณาได้ว่า $\gamma_2(k)$ ในเทอมของความเอนเอียงของ $\hat{\beta}$ ยกกำลังสองซึ่งก็คือ $\gamma_2(k) = bias^2(\hat{\beta}_R)$ การที่เกิดความเอนเอียงทำให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของ $\hat{\beta}_R$ มีค่าน้อยกว่าตัวประมาณไม่เอนเอียง

ในการประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีริดจ์รีเกรชชันจะต้องเลือกค่า k ให้มีค่าเหมาะสมเพื่อที่จะทำให้ค่าความแปรปรวนของตัวประมาณสัมประสิทธิ์การถดถอยลดลง และค่าเฉลี่ยความคลาดเคลื่อนกำลังสองของตัวประมาณริดจ์รีเกรชชันมีค่าน้อยกว่าตัวประมาณจากวิธีกำลังสองน้อยสุด

การประมาณค่าพารามิเตอร์ k

ในการประมาณค่าสัมประสิทธิ์การของริดจ์รีเกรชชันต้องมีการบวกค่า k กับเส้นทแยงมุมของเมทริกซ์ $X'X$ ซึ่งมีผู้เสนอและทำการวิจัยเพื่อหาวิธีในการประมาณค่า k หลายวิธี เช่น ผลการวิจัยของ เจษฎาพร ยุทธนวิบูลย์ชัย (พ.ศ.2533) เป็นการนำค่า k จากวิธีของ โฮเอิล เคนนาร์ด (Hoel Kennard) และ บัลด์วิน (Baldwin) ในปี ค.ศ.1975 วิธีของ TZE-SAN-LEE และวิธีของ แมคโดนัลด์ การ์ลานิว (Macdonald Galarneau) ในปี ค.ศ.1975 เพื่อนำไปในการหาตัวประมาณของริดจ์รีเกรชชัน ผลงานวิจัยของจิรายุส พู่มนตรี (พ.ศ.2534) ได้เสนอวิธีการประมาณค่า k โดยใช้วิธีการ Binary Search ซึ่งมีข้อจำกัดคือใช้ได้กับข้อมูลที่มีลักษณะแบบไม่ต่อเนื่องและมีการเรียงลำดับ ซึ่งทำให้ต้องเสียเวลาเรียงลำดับข้อมูลโดยเปรียบเทียบกับวิธีการประมาณค่า k จากผลการวิจัยของเจษฎาพร และยังมีผลงานวิจัยของ ธันยากร ต้นชลขันธ์ (พ.ศ.2539) เป็นการหาค่า k แบบลำดับ (Sequential Search) เป็นการค้นหาค่า k ไปเรื่อย ๆจนกว่าจะได้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองที่มีค่าเหมาะสมโดยเปรียบเทียบกับวิธีหาตัวประมาณจากวิธีกำลังสองน้อยที่สุดและตัวประมาณที่ได้ใช้หลักการของริดจ์และสไตน์ ผู้วิจัยได้พยายามค้นหาข้อมูลเพื่อหาวิธีในการหาค่า k พบว่าในปี ค.ศ. 1995 บลีแมน (Breiman) ได้เสนอวิธีการประมาณค่าพารามิเตอร์ k จากค่าผลรวมของเส้นทแยงมุม (trace) ของเมทริกซ์ $(X'X(X'X+kI)^{-1})$ โดยทำการค้นหาข้อมูลแบบลำดับจนกระทั่งได้ค่าผลรวมของเส้นทแยงมุมเท่ากับจำนวนตัวแปรอิสระที่นำมาทำการวิจัย โดยมีหลักการในการทำงานดังนี้คือ

กำหนดให้

k_{opt} คือค่า k ที่ทำให้ $trace(X'X(X'X+kI)^{-1})$ เท่ากับจำนวนตัวแปรอิสระ

k_{bef} คือค่า k เริ่มต้นในการทดสอบ

และ d คือค่าคงที่ที่กำหนดขึ้นเป็นระยะห่างในการทดสอบเพื่อหาค่า k

<u>ขั้นที่ 1</u> กำหนดให้ k_{bef} = 0.0 และ d = 0.001

<u>ขั้นที่ 2</u> คำนวณค่า $k_{opt} = k_{bef} + 0.001$

<u>ขั้นที่ 3</u> แทนค่า k_{opt} = 0.001 ลงไปใน $trace(X'X(X'X+kI)^{-1})$ เพื่อทำการทดสอบว่า

ก) ถ้า $trace(X'X(X'X+k_{opt}I)^{-1})$ = จำนวนตัวแปรอิสระที่ทำการทดสอบ ผู้วิจัยจึงทำการยุติการประมวลผล

ข) ถ้า $trace(X'X(X'X+k_{opt}I)^{-1}) \neq$ จำนวนตัวแปรอิสระที่ทำการทดสอบ ผู้วิจัยจะทำการวิเคราะห์ในขั้นตอนที่ 2 จนกว่าจะเข้ากรณี ก)

<u>ขั้นที่ 4</u> เมื่อได้ค่า k จากกรณี ก) แล้วนำให้หาค่าตัวประมาณสัมประสิทธิ์การถดถอยของริดจ์รี-เกรชชัน คือ $\hat{\beta}_R = (\mathrm{X}'\mathrm{X} + k\mathrm{I})^{-1}\mathrm{X}'y$

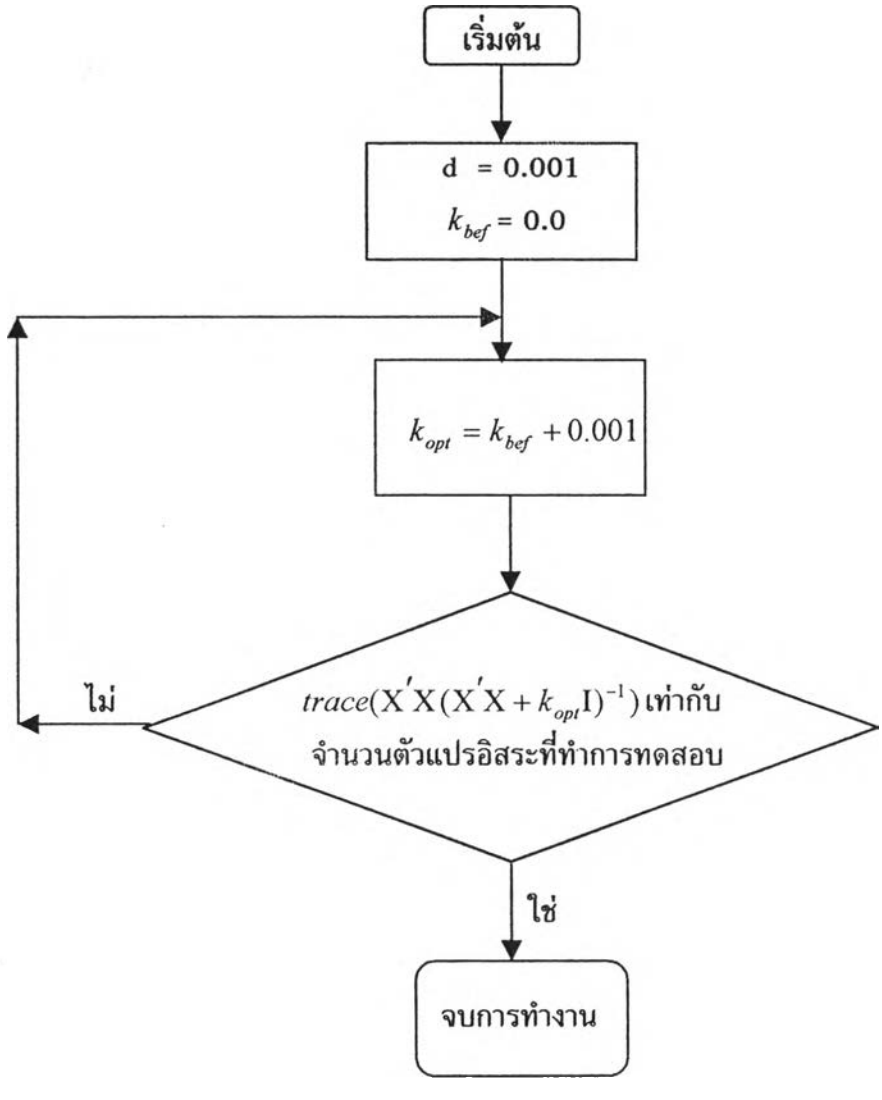
ในการกำหนดระยะห่างให้การทดสอบเพื่อหาค่า k ผู้วิจัยได้ทำการทดลองกำหนดค่า d ไว้หลายช่วงดังนี้

1. กำหนดค่า d = 0.1 และ 0.01 พบว่าจะให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำ แต่ช่วงห่างของระยะไม่ละเอียดพอจึงทำให้ในบางกรณีที่ทำการทดสอบไม่สามารถหาค่าของ

$trace(\mathrm{X}'\mathrm{X}(\mathrm{X}'\mathrm{X} + k\mathrm{I})^{-1})$ = จำนวนตัวแปรอิสระที่ทำการทดสอบ

2. กำหนดค่า d = 0.001 พบว่าจะให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองสูงขึ้นเล็กน้อย และสามารถหาค่าของ $trace(\mathrm{X}'\mathrm{X}(\mathrm{X}'\mathrm{X} + k\mathrm{I})^{-1})$ = จำนวนตัวแปรอิสระที่ทำการทดสอบ ได้ในทุกกรณี

ดังนั้นในการทำวิจัยครั้งนี้ผู้วิจัยได้เลือกค่า d = 0.001 เพราะว่าให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองไม่สูงจนเกินไปและสามารถหาค่า k ได้ในทุกกรณี โดยสามารถสรุปเป็นผังการทำงานได้ดังนี้



การประมาณค่าสัมประสิทธิ์ด้วยวิธีสมการถดถอยเชิงเส้นการ์ล็อต(GAR)

ในปี ค.ศ. 1995 บลีแมน (Breiman) ได้เสนอวิธีการประมาณค่าสัมประสิทธิ์การถดถอยพหุคูณในการแก้ปัญหาการเกิดพหุสัมพันธ์ระหว่างตัวแปรอิสระ โดยนำข้อดีของวิธีกำลังสองน้อยสุดและได้อาศัยหลักการของตัวประมาณสไตน์ (Stein estimator) เข้ามาช่วยในการประมาณค่าสัมประสิทธิ์การถดถอยของวิธี GAR โดยตัวประมาณสัมประสิทธิ์การถดถอยแต่ละตัวจะได้จากค่าคงที่ c แต่ละตัวคูณกับตัวประมาณสัมประสิทธิ์จากวิธีกำลังสองน้อยสุดแต่ละตัว ซึ่งสามารถเขียนได้อยู่ในรูป

(2.21) $$\underset{\sim}{\hat{\beta}_G}(k) = \hat{c}_k \underset{\sim}{\hat{\beta}_k} \quad ; k = 1,\ldots, p+1$$

โดยที่ $\sum_{k=1}^{p+1} c_k^2 \le (p+1)^2$ เมื่อ p เป็นจำนวนตัวแปรอิสระ

ค่าของ c แต่ละค่าหาได้จากรูปแบบของ

(2.22) $$y_i = \sum_{k=1}^{p+1} c_k \hat{\beta}_k X_{ki} + \varepsilon_i \quad ; i = 1,\ldots, n$$

กำหนดให้ $\underset{\sim}{\hat{\beta}} X_{ki} = Z_{ki}$ ดังนั้นจะได้

(2.23) $$y_i = \sum_{k=1}^{p+1} c_k Z_{ki} + \varepsilon_i$$

หรือสามารถเขียนให้อยู่ในรูปเมทริกซ์

(2.24) $$\underset{\sim}{y} = Z\underset{\sim}{c} + \underset{\sim}{\varepsilon}$$

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

Z เป็นเมทริกซ์ของตัวแปรอิสระคูณกับตัวประมาณกำลังสองน้อยสุดขนาด n x (p+1)

$\underset{\sim}{c}$ เป็นเวกเตอร์ค่าคงที่ขนาด (p+1) x 1

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

n เป็นขนาดตัวอย่างของตัวแปรอิสระ

และ p เป็นจำนวนตัวแปรอิสระ

โดยที่ $E(\underset{\sim}{\varepsilon}) = 0$ และ $Cov(\underset{\sim}{\varepsilon}) = \sigma^2 I_n$

จากสมการ (2.24) จะได้ผลรวมความคลาดเคลื่อนกำลังสองอยู่ในรูปแบบดังนี้

$$\begin{aligned} \underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon} &= (\underset{\sim}{y}-\mathrm{Z}\underset{\sim}{c})'(\underset{\sim}{y}-\mathrm{Z}\underset{\sim}{c}) \\ &= \underset{\sim}{y}'\underset{\sim}{y}-\underset{\sim}{c}'\mathrm{Z}'\underset{\sim}{y}-\underset{\sim}{y}'\mathrm{Z}\underset{\sim}{c}+\underset{\sim}{c}'\mathrm{Z}'\mathrm{Z}\underset{\sim}{c} \\ &= \underset{\sim}{y}'\underset{\sim}{y}-2\underset{\sim}{c}'\mathrm{Z}'\underset{\sim}{y}+\underset{\sim}{c}'\mathrm{Z}'\mathrm{Z}\underset{\sim}{c} \end{aligned} \qquad (2.25)$$

ผู้วิจัยได้นำหลักการของการหาค่าสัมประสิทธิ์ของวิธีกำลังสองน้อยสุดมาใช้ กล่าวคือจะทำให้ผลรวมของค่าความคลาดเคลื่อนกำลังสองมีค่าต่ำสุด โดยจะได้ค่าประมาณสัมประสิทธิ์การถดถอยพหุคูณจากการหาอนุพันธ์อันดับที่1 ของสมการ (2.25) เทียบกับ $\underset{\sim}{c}$ แล้วให้เท่ากับศูนย์ ซึ่งผลที่ได้อยู่ในรูปแบบดังนี้

$$\frac{\partial}{\partial \underset{\sim}{c}}S(\underset{\sim}{c}) = -2Z'\underset{\sim}{y}+2Z'Z\underset{\sim}{\hat{c}} = 0$$

กล่าวคือ $$Z'Z\underset{\sim}{\hat{c}} = Z'\underset{\sim}{y}$$

จะได้ตัวประมาณ $\underset{\sim}{c}$ อยู่ในรูปของ

$$\underset{\sim}{\hat{c}} = (\mathrm{Z}'\mathrm{Z})^{-1}\mathrm{Z}'\underset{\sim}{y} \qquad (2.26)$$

ดังนั้นตัวประมาณสัมประสิทธิ์การถดถอยของวิธี GAR จะมีลักษณะดังนี้

$$\hat{\beta}_G(1) = \hat{c}_1\hat{\beta}_1 = \hat{\delta}_1$$

$$\vdots$$

$$\hat{\beta}_G(p+1) = \hat{c}_{p+1}\hat{\beta}_{p+1} = \hat{\delta}_{p+1}$$

หรือ $$\underset{\sim}{\hat{\beta}}_{(G)} = \underset{\sim}{\hat{\delta}}$$

ตัวประมาณที่ได้จากวิธี GAR เป็นค่าคงที่และเป็นตัวประมาณที่ไม่เอนเอียง โดยมีค่าใกล้เคียงกับตัวประมาณกำลังสองน้อยสุด