บทที่ 2

ตัวสถิติและผลงานวิจัยที่เกี่ยวข้อง

การศึกษาวิธีการตรวจสอบค่าผิดปกติ และอิทธิพลของค่าสังเกต ในสมการการถดถอยเชิงเส้นพหุ ในงานวิจัยครั้งนี้ เป็นการศึกษาความน่าจะเป็นความคลาดเคลื่อนประเภทที่ 1 และอำนาจการทดสอบของตัวสถิติทดสอบ ทั้ง 3 วิธี คือ วิธีของ จีแบร์รี่ วิธีของคุ๊ก วิธีของแอนดูร์และเพรตจิบอน โดยศึกษา 2 กรณีคือ กรณีที่ค่าผิดปกติมี 1 ค่าและกรณีที่ค่าผิดปกติมี 2 ค่า ซึ่งจะกล่าวรายละเอียดต่อไป

ในการศึกษาสมการการถดถอยเชิงเส้นพหุ

$$\underset{\sim}{Y}_{(nxl)} = X_{(nxp)}\,\underset{\sim}{\beta}_{(pxl)} + \underset{\sim}{\varepsilon}_{(nxl)}$$

โดยที่

$\underset{\sim}{Y}$ เป็นเวคเตอร์ของตัวแปรตามที่มีขนาด nxl

X เป็นเมตริกซ์คงที่ของตัวแปรอิสระที่มีขนาด nxp และมี rank = ρ

$\underset{\sim}{\beta}$ เป็นเวคเตอร์ของสัมประสิทธิ์การถดถอยที่ไม่ทราบค่าและมีขนาด pxl

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

$$E(\underset{\sim}{\varepsilon}) = 0, V(\underset{\sim}{\varepsilon}) = \sigma^2 I_n \text{ , } COR(\varepsilon_i,\varepsilon_j) = 0 \text{ , } i \neq j$$

การประมาณค่าสัมประสิทธิ์การถดถอย โดยวิธีกำลังสองน้อยที่สุด จะได้ว่า

$$\hat{\underset{\sim}{\beta}} = (X^TX)^{-1}X^T\underset{\sim}{Y}$$

$$\hat{Y} = X\hat{\underset{\sim}{\beta}} = X(X^TX)^{-1}X^TY = H\underset{\sim}{Y}$$

เมื่อ $H_{(nxn)} = X(X^TX)^{-1}X^T$ เรียกว่า แฮทเมตริกซ์ (hat matrix) ซึ่งมีบทบาทสำคัญ ในการศึกษาค่าผิดปกติ และอิทธิพลของค่าสังเกต ความคลาดเคลื่อน

$$\underset{\sim}{R} = \underset{\sim}{Y} - \hat{\underset{\sim}{Y}}$$

$$= \underset{\sim}{Y} - X(X^TX)^{-1}X^T\underset{\sim}{Y} = (I - H)\underset{\sim}{Y}$$

ความแปรปรวนของความคลาดเคลื่อน คือ

$$V(\underset{\sim}{R}) = (I - H)\, s^2$$

ความแปรปรวนของความคลาดเคลื่อนที่ค่าสังเกต i คือ

$$V(\underset{\sim}{R}_i) = (1 - h_{ii})\, S^2_{(i)}$$

เมื่อ h_{ii} เป็นสมาชิกแนวทะแยงของ H

s^2 เป็นตัวประมาณที่ไม่เอนเอียงของ σ^2

ความคลาดเคลื่อนมาตรฐาน คือ

$$t_i = \frac{R_i}{s\sqrt{1-h_{ii}}}$$

ในการศึกษาค่าผิดปกติ ค่าความคลาดเคลื่อนมาตรฐาน เป็นตัวสำคัญในการบ่งชี้ค่าผิดปกติของวิธีจีแบร์รี่และวิธีของคุ๊ก

2.1 การคำนวณค่าแฮทเมตริกซ์

ในการศึกษาสมการการถดถอยเชิงเส้นพหุ เพื่อสะดวกในการคำนวณและประหยัดเวลาการคำนวณจะทำการแปลงตัวแปรอิสระ เป็นเมตริกซ์ตั้งฉาก (orthogonal matrix) โดยวิธี โคลสกี้ (Graybill 1976:229) ซึ่งจะทำให้ตัวแปรอิสระเป็นอิสระกัน การทำให้ตัวแปรอิสระตั้งฉากกันมีขั้นตอนดังนี้

ให้

$$A = X^T X = U^T U$$

$$\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1p} \\ A_{21} & A_{22} & \cdots & A_{2p} \\ \vdots & & & \\ A_{p1} & A_{p2} & \cdots & A_{pp} \end{bmatrix} = \begin{bmatrix} u_{11} & 0 & 0 & \cdots & 0 \\ u_{21} & u_{22} & 0 & \cdots & 0 \\ \vdots & & & & \\ u_{p1} & u_{p2} & u_{p3} & \cdots & u_{pp} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{p} \\ 0 & u_{22} & u_{23} & \cdots & u_{2p} \\ & \vdots & & & \\ 0 & 0 & 0 & \cdots & u_{pp} \end{bmatrix}$$

สมาชิกของเมตริกซ์ U มีขั้นตอนการคำนวณ คือ

1) หาสมาชิกแถวที่ 1 คอลัมน์ที่ 1 ของเมตริกซ์โดย

$$u_{11} = \sqrt{A_{11}}$$

2) หาสมาชิกแถวที่ 1 คอลัมน์ 2 ถึงคอลัมน์ที่ p ของเมตริกซ์ U โดย

$$u_{1j} = \frac{A_{1j}}{u_{11}} = \frac{A_{1j}}{\sqrt{A_{11}}} \qquad j = 2,3,\ldots,p$$

3) หาสมาชิกบนเส้นทะแยงมุม (diagonal) ของเมตริกซ์โดย

$$u_{ii} = \sqrt{A_{ii} - \sum_{p=1}^{i-1} u_{pi}^2} \qquad i = 2,3,\ldots, p$$

4) หาสมาชิกที่เหลือของเมตริกซ์ U โดย

$$u_{ij} = \frac{1}{u_{ii}} \left[A_{ij} - \sum_{p=1}^{i-1} u_{pi} . u_{pj} \right] \quad j > i \text{ และ } i = 2,3,\ldots p-1$$

$$u_{ij} = 0 \quad j < i \text{ และ } i = 2,3,\ldots,p$$

จากนั้นตัวแปรอิสระ ที่ถูกทำให้ตั้งฉากจะมีความแปรปรวนเดียวกัน แล้วนำมาแปลงให้อยู่ในรูปใหม่โดย

$$Q = XU^{-1}$$

ดังนั้นการคำนวณ $H = X(X^T X)^{-1} X^T = QQ^T$

2.2 วิธีการตรวจสอบวิธีของจีแบร์รี่

วิธีการตรวจสอบค่าผิดปกติ ของวิธีจีแบร์รี่ เป็นการตรวจสอบค่าผิดปกติในกรณีที่ค่าผิดปกติและค่าอิทธิพล มี 1 ค่าหรือมากกว่าหนึ่งค่า ซึ่งเสนอโดย จี แบร์รี่ (G. Barrie Wetherill - 1986) โดยมีหลักเกณฑ์คือ ค่าความคลาดเคลื่อนมาตรฐานเป็นตัวบ่งชี้ค่าผิดปกติ ถ้าค่าความคลาดเคลื่อนมาตรฐานมีค่าสูงมาก แสดงว่า ตัวแปรตาม Y เป็นค่าผิดปกติ ทำการตรวจสอบทีละค่า ตัวสถิติที่ทดสอบ คือ $d_i = \frac{t_i^2}{n-p}$ ซึ่งมีขั้นตอนการคำนวณดังนี้

1) คำนวณค่าความคลาดเคลื่อนมาตรฐาน t_i

2) เลือกค่าความคลาดเคลื่อนมาตรฐาน t_i ที่มีค่าสูงสุด $t = \max|t_i|$

3) ทำการตรวจสอบค่าสังเกตที่ตรงกับ t ว่าเป็นค่าผิดปกติ จากสมมติฐาน

$$H_o : E(\underset{\sim}{y}) = X\underset{\sim}{\beta}$$

$$H_a : E(\underset{\sim}{y}) = X\underset{\sim}{\beta} + d$$

ตัวสถิติทดสอบ

$$d_i = \frac{t_i^2}{n-p}$$

ซึ่ง d_i เป็นตัวสถิติที่มีการแจกแจงเบต้า ที่มีพารามิเตอร์เป็น

$$\left(\frac{1}{2}, \frac{n-p-1}{2}\right)$$

4) นำค่า d_i เปรียบเทียบกับค่าขอบเขตวิกฤต จากตาราง Lund (1975): ซึ่งมีความสัมพันธ์ คือ

$$t_o = |(n-p)\ d_o|^{\frac{1}{2}}$$

<u>เกณฑ์การตัดสินใจ</u> : ถ้า $d_i > d_o$ ปฏิเสธสมมติฐาน H_o แสดงว่า ค่าสังเกตที่ i เป็นค่าผิดปกติ

5) ในกรณีที่ปฏิเสธ H_o :จะตัดค่าสังเกตที่ตรงกับค่า t แล้ววิเคราะห์ข้อมูลขนาด n-1 เช่นเดียวกับข้อ 1-4

2.3 <u>วิธีการตรวจสอบวิธีของคุ๊ก</u>

วิธีการตรวจสอบค่าผิดปกติ และอิทธิพลของค่าสังเกต วิธีของคุ๊ก เสนอโดย อาร์ เดนนิส คุ๊ก (R. Dennis Cook 1977 1979 1980) โดยมีหลักเกณฑ์ คือ ค่าผิดปกติย่อมมีอิทธิพลต่อสัมประสิทธิ์การถดถอยโดยตรง เมื่อพิจารณาผลต่างของกำลังสองค่าสัมประสิทธิ์การถดถอยจากข้อมูลขนาด n $(\hat{\beta})$ กับสัมประสิทธิ์การถดถอยจากข้อมูล n-1 $(\hat{\beta}_{(i)})$ ถ้าค่าสังเกตที่ตัดไปมีผลต่อสัมประสิทธิการถดถอยมาก แสดงว่าค่าสังเกตนั้นย่อมเป็นค่าผิดปกติ

ตัวสถิติที่ทดสอบคือ

$$C_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})^T X^T X(\hat{\beta}_{(i)} - \hat{\beta})}{ps^2} \qquad (2.3.1)$$

เมื่อ $\hat{\beta}$ เป็นสัมประสิทธิ์การถดถอยที่ได้จากการประมาณค่าข้อมูลขนาด n

$\hat{\beta}_{(i)}$ เป็นสัมประสิทธิ์การถดถอยที่ได้จากการประมาณค่าข้อมูลขนาด n-1

จากสมการ (2.3.1) สามารถเขียนใหม่ได้เป็น

$$C_i = \frac{t_i^2}{p} . w_i \qquad \text{เมื่อ} \quad w_i = \frac{h_{ii}}{1-h_{ii}} \tag{2.3.2}$$

(การพิสูจน์ดูจากภาคผนวก)

ตัวสถิติ C_i มีการแจกแจงแบบเอฟ (Cook R.D. : 1980) ที่มีดีกรีแห่งความอิสระเป็น $(1, n-p-1, \alpha)$ ในการตรวจสอบสมมติฐาน $H_o : E(\underset{\sim}{y}) = X\underset{\sim}{\beta}$ ตัวสถิติที่คำนวณคือ

$$F = \frac{t_i^2 (n-p-1)}{(n-p-t_i^2)} \tag{2.3.3}$$

นำมาเปรียบเทียบกับค่าวิกฤต $F_{(1,n-p-1,\alpha)}$ ในการคำนวณมีขั้นตอนคือ

1) คำนวณค่า t_i

2) เลือกค่า $t = \max |t_i|$

3) คำนวณค่าสถิติ F จากสมการ (2.3.3)

4) เปรียบเทียบค่า F กับ $F_{(1,n-p-1)}$ ซึ่งมีเกณฑ์การพิจารณาดังนี้

   ปฏิเสธ H_o : เมื่อ $F > F_{(1,n-p-1)}$

5) ในกรณีที่ปฏิเสธ จะทำการลบค่าสังเกตที่ตรงกับ t_i แล้วคำนวณค่าสถิติเช่นเดียวกับ ข้อ 1 - 4

2.4 วิธีการตรวจสอบของแอนดรูร์และเพรตจิบอน

เดวิด เอฟ แอนดรูร์ และ ดาร์ลิล เพรตจิบอน เสนอในปี พ.ศ. 2521 (David F. Andrews & Daryl Pregibon : 1978) โดยพิจารณาจากหลักการที่ว่า ค่าสังเกตใดมีผลต่อค่าดีเทอร์มินัลของ $X^T X$ $(|X^T X|)$ มากหรือน้อย เมื่อลบค่าสังเกตนี้ออกไป ย่อมมีผลต่อการเปลี่ยนแปลงค่าดีเทอร์มินัลมากหรือน้อยตามลำดับ ซึ่งนำมารวมกับแนวคิดที่ว่า ค่าผิดปกติย่อมมีผลต่อผลบวกกำลังสองของความคลาดเคลื่อนลดลง เมื่อลบค่าผิดปกติ สามารถแสดงความสัมพันธ์ในรูปสมการ

$$X^{*T} X^* = RSS \times |X^T X| \qquad (2.4.1)$$

$$X^*_{(nx(p+1))} = (X : y)$$ เป็นเมตริกซ์อิสระที่เพิ่มคอลัมน์เวกเตอร์ของตัวแปรตาม $\underset{\sim}{y}$ ซึ่งมี rank = p + 1

ตัวสถิติของแอนดรูร์และเพรตจิบอน คือ

$$R^{(k)}_{ij..}(X^*) = \frac{D^{(k)}_{ij..} \; X^{*T} X^*}{X^{*T} X^*} \qquad (2.4.2)$$

$D^{(k)}_{ij..}(*)$ หมายถึง กรณีที่ลบค่าสังเกต ที่ i, j.. จำนวน k ค่า

$\bar{D}^{(k)}_{ij..}(*)$ หมายถึง คอมพลีเมนต์ของ $D^{(k)}_{ij..}(*)$

ในการคำนวณค่าสถิติวิธีของแอนดรูร์และเพรตจิบอน อาศัยคุณสมบัติของเมตริกซ์ และดีเทอร์มินัล คือ

$$X^{*T} X^* = D^{(1)}_i \; (X^{*T} X^*) + \underset{\sim}{x}^{*T}_i \; \underset{\sim}{x}^*_i$$

สามารถคำนวณค่า $R^{(k)}_{ij..}(X^*)$ ได้จาก

1) กรณี $k > 1$

$$R_{ij}^{(k)} = \bar{D}_{ij..}^{(k)} \left| w^* \right| \qquad (2.4.3)$$

2) กรณี $k = 1$

$$R_{i}^{(1)} = w_{ii}^{*} \qquad (2.4.4)$$

เมื่อ

$$w^* = I - X^* (X^{*T} X^*)^{-1} X^{*T}$$

$$= I - Q^* Q^{*T}$$

(การพิสูจน์ดูจากภาคผนวก)

การคำนวณตัวสถิติทดสอบ วิธีของแอนดูร์และเพรตจิบอน มีขั้นตอนการคำนวณดังนี้

1) คำนวณค่า $R_{i}^{(k)}$ โดยคำนวณจากสมการ (2.4.3) และ (2.4.4)

2) เลือก $R_{ij..}^{(k)}(X)^*$ ที่มีค่าต่ำสุด

3) คำนวณค่า $R_{ij..}^{(k)}$ (X) (คำนวณเฉพาะค่าของตัวแปรอิสระ)

4) คำนวณค่าความน่าจะเป็นที่แน่นอน (Exact probability) : α_k

5) นำค่า α_k เปรียบเทียบกับค่าระดับนัยสำคัญที่กำหนด (α) ซึ่งกำหนด 3 ระดับ คือ α = 0.01 0.05 0.10

6) เกณฑ์การตัดสินใจ : จากสมมติฐาน

H_o : ค่าสังเกตที่ k เป็นค่าผิดปกติ

H_a : ค่าสังเกตที่ k ไม่เป็นค่าผิดปกติ

ปฏิเสธ H_o: เมื่อ $\alpha_k < \frac{\alpha}{2}$