บทที่ 2

สถิติทดสอบและการแจกแจงที่สำคัญ

ในการศึกษาเกี่ยวกับการตรวจสอบความไม่คงที่ของความคลาดเคลื่อนในตัวแบบความถดถอยเชิงเส้นนั้น ตัวสถิติที่ใช้ทดสอบมีอยู่หลายวิธี ในการศึกษาวิจัยครั้งนี้ คือ การทดสอบแจ็คไนฟ์ (Jackknife Test) ,การทดสอบด้วยอัตราส่วนภาวะน่าจะเป็น (Likelihood Ratio Test) ,การทดสอบของบาร์ตเลตต์ (Bartlette 's Test) และศึกษาในกรณีที่ความคลาดเคลื่อนมีการแจกแจงแบบปกติ (Normal Distribution) การแจกแจงปกติปลอมปน (Scale contaminate Normal distribution) เมื่อข้อมูลมีลักษณะเป็นกลุ่ม ภายใต้ข้อกำหนดดังได้กล่าวไว้ในขอบเขตของการวิจัยในบทที่ 1 ดังนั้นในบทนี้จะกล่าวถึงการแจกแจงปกติ และการแจกแจงปกติปลอมปน ก่อน แล้วจึงกล่าวถึงรายละเอียดของการทดสอบแต่ละวิธีซึ่งรายละเอียดต่าง ๆ เป็นดังนี้

การแจกแจงที่เกี่ยวข้องในการศึกษา

1. การแจกแจงปกติ (Normal Distribution) ฟังก์ชั่นความน่าจะเป็นของการแจกแจงปกติมีรูปแบบเป็น

$$f(x) = \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} \qquad ; \sigma^2 > 0$$

μ,σ เป็นพารามิเตอร์

คุณสมบัติของการแจกแจง

1. โค้งปกติมีฐานนิยมอยู่ที่ $x=\mu$ (โค้งสูงที่สุด)
2. โค้งมีสมมาตรกับแกนตั้งที่ลากผ่าน μ
3. โค้งปกติมีจุดเปลี่ยนเว้าที่ $x= \mu +\sigma$
4. ปลายโค้งเข้าใกล้แกน x เมื่อ x มีค่าห่างจาก μ ออกไปทุกที
5. พื้นที่ทั้งหมดที่อยู่ใต้เส้นโค้งและอยู่เหนือแกน x มีค่าเป็น 1
6. ค่าเฉลี่ยของ x คือ μ และความแปรปรวนคือ σ^2

2. การแจกแจงปกติปลอมปน (Scale contaminated Normal distribution) ฟังก์ชั่นความน่าจะเป็นมีรูปแบบเป็น

$$f(x) = (1-p)N(\mu, \sigma^2) + pN(\mu, c^2\sigma^2)$$

โดยที่ p คือ เปอร์เซ็นต์การปลอมปน

c^2 คือ ค่าสเกลคอนทามิเนต ซึ่งจะเป็นจำนวนเท่าของความแปรปรวน

การทดสอบที่ใช้ในการศึกษา

รูปแบบสมการถดถอยเชิงเส้นสำหรับข้อมูลภาคตัดขวาง มีข้อสมมติที่ว่าภายในกลุ่มย่อยของค่าสังเกตมีค่าความแปรปรวนเท่ากัน และต่างกันในทุกกลุ่ม ตัวอย่างโดยทั่วไปของข้อมูลลักษณะนี้เช่น ค่าสังเกตที่ถูกจับกลุ่มตามภาคภูมิศาสตร์ , รายได้ เป็นต้น ในการวิจัยจะศึกษาที่จำนวนกลุ่ม (m) = 4 แต่ละกลุ่มมีค่าสังเกต (n_i) เท่ากัน = 10,20,30 และขนาดตัวอย่าง (n) คือ 40,80,120 ตามลำดับโดยที่ $\sum_{i=1}^{m} n_i = n$ สมการถดถอยเชิงเส้นอย่างง่ายสำหรับกลุ่มที่ i สามารถเขียนได้ดังนี้

$$y_i = X_i\beta_i + \varepsilon_i \quad ; \quad i = 1,2,3,4 \qquad \text{......................................} (2.1)$$

โดยที่ $y_i = (y_{i1}, y_{i2}, \ldots, y_{in})'$ มีขนาดเวกเตอร์เป็น (n_ix1)

X_i เป็นเมตริกซ์ขนาด ($n_i$xk) ; k คือจำนวนพารามิเตอร์

β_i เป็นสัมประสิทธิ์การถดถอย มีขนาดเวกเตอร์เป็น (kx1)

และ ε_i เป็นความคลาดเคลื่อนอิสระ

สำหรับแต่ละกลุ่มสมมติว่า $n_i > k$ และ $E(\varepsilon_i) = 0$,$E(\varepsilon_i \varepsilon_j') = 0$; $i \neq j$ และ $E(\varepsilon_{ii}) = \sigma^2 I_n$

เมื่อกำหนดให้สัมประสิทธิ์เท่ากันในทุกกลุ่ม ($\beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta$)

รูปแบบสมการถดถอยตาม 2.1 จะเขียนใหม่ได้เป็น

$$Y = X\beta + \varepsilon$$

โดยที่ $Y = (y_1, y_2, \ldots, y_4)'$ มีขนาดเวกเตอร์เป็น (nx1)

$X = (X_1, X_2, X_3, X_4)'$ ขนาด (nxk) ; k คือจำนวนพารามิเตอร์

β มีขนาดเวกเตอร์เป็น (kx1)

$\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4)'$ ขนาด (nx1)

สมมติฐานที่พิจารณาสำหรับกรณีนี้คือ ความแปรปรวนของความคลาดเคลื่อน (σ_i^2) มีค่าคงที่ เขียนเป็นสัญญลักษณ์คือ

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma^2$$

ขั้นตอนการดำเนินการ โดย

1. Generate ค่า x_{it} ที่มีการแจกแจงปกติโดยให้ขนาดตัวอย่าง (n) ถูกแบ่งเป็นกลุ่ม ๆ แต่ละกลุ่มมีการแจกแจงปกติที่ค่าเฉลี่ยและความแปรปรวนแตกต่างกัน

2. Generate ค่าความคลาดเคลื่อนที่มีการแจกแจงต่าง ๆ กัน ตามค่าเฉลี่ยและความแปรปรวนที่กำหนด

3. กำหนดให้ ค่าพารามิเตอร์เริ่มต้น β_{0i} , $\beta_{1i} = 1$

4. คำนวณหา y_{it} จาก $y_{it} = \beta_{0i} + \beta_{1i}x_{it} + \varepsilon_{it}$

5. คำนวณหา $\hat{\beta}_0, \hat{\beta}_1$ จาก x_{it} และ y_{it} โดยวิธีการ OLS

6. คำนวณค่า $\hat{y}_{it}$ จาก $\hat{y}_{it} = \hat{\beta}_0 + \hat{\beta}_1 x_{it}$

7. คำนวณค่า e_{it} จาก $e_{it} = y_{it} - \hat{y}_{it}$

ค่า e_{it} ที่ได้จะนำไปคำนวณหาตัวสถิติแต่ละค่าตามวิธีการดังต่อไปนี้

1. **การทดสอบแจ็คไนฟ์ (Jackknife test)**

1. คำนวณหาความแปรปรวนของความคลาดเคลื่อน จาก

$$S_i^2 = \sum_{t=1}^{n_i} (e_{it} - \bar{e}_i)^2 / v_i$$

โดยที่ $\bar{e}_i = \sum_{t=1}^{n_i} \frac{e_{it}}{n_i}$

$$v_i = \frac{n-k}{w_i} - 1$$

$$w_i = \frac{n}{n_i}$$

2. ตัด j^{th} residual ใน i^{th} กลุ่ม แล้วคำนวณหา

$$S_{i(j)}^2 = \sum_{t \neq j}^{n_i} (e_{it} - \bar{e}_{i(j)})^2 / v_i$$

โดยที่ $\bar{e}_{i(j)} = \sum_{t \neq j}^{n_i} \frac{e_{it}}{(n_i - 1)}$

$$v_i = \frac{n-k-1}{w_i} - 1$$

$$w_i = \frac{n}{n_i}$$

3. จะได้ Pseudo-value : u_{ij} ดังนี้

$$u_{ij} = n_i \ln S_i^2 - (n_i - 1)\ln S_{i(j)}^2$$

4. จะได้ ตัวสถิติ JACK ดังนี้

$$JACK = \frac{\sum_{i=1}^{m} n_i(\bar{u}_{i.} - \bar{u}_{..})^2 / (m-1)}{\sum_{i=1}^{m}\sum_{j=1}^{n_i} (u_{ij} - \bar{u}_{i.})^2 / \sum_{i=1}^{m}(n_i - 1)}$$

โดยที่

$$\bar{u}_{i} = \sum_{i=1}^{n_i} \frac{u_{ij}}{n_i}$$

$$\bar{u} = \sum_{i=1}^{m}\sum_{i=1}^{n_i} \frac{u_{ij}}{n_i} / m$$

ภายใต้ H_0 , JACK จะมีการแจกแจงใกล้เคียงกับ $F_{m-1,d}$ โดยที่ $d = \sum_{i=1}^{m}(n_i - 1)$

2. การทดสอบด้วยอัตราส่วนภาวะน่าจะเป็น (Likelihood Ratio Test) :

พิจารณาจาก $\lambda = \prod_{i=1}^{m} (\tilde{\sigma}_i^2)^{n_i/2} / (\tilde{\sigma}^2)^{n/2}$

โดยที่ $\tilde{\sigma}^2 = e'e / n$

$\tilde{\sigma}_i^2 = \hat{\varepsilon}_i' \hat{\varepsilon}_i / n_i$

$e = y - X\hat{\beta}_{OLS}$ คือเวกเตอร์ของค่าประมาณของความคลาดเคลื่อน ทั้ง 4 กลุ่ม

$y = (y_1, y_2, y_3, y_4)'$

$y_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})'$; i=1,2,...,4

$X = (X_1, X_2, X_3, X_4)'$

$X_i = (X_{i1}, X_{i2}, \ldots, X_{in_i})'$; i=1,2,...,4

$\hat{\varepsilon}_i = y_i - X_i \hat{\beta}_{ML}$

$$\hat{\beta}_{ML} = \left[\sum_{i=1}^{4} \frac{1}{\hat{\sigma}_i^2} X'_i X_i\right] \sum_{i=1}^{4} \frac{1}{\hat{\sigma}_i^2} X'_i y_i$$

$\hat{\sigma}_i^2 = e_i' e_i / n_i$

จะได้ตัวสถิติ LR จาก $-2\ln\lambda$:

$$LR = n \ln \tilde{\sigma}^2 - \sum_{i=1}^{4} n_i \ln \tilde{\sigma}_i^2$$

เมื่อประมาณค่าของ $\tilde{\sigma}^2$ และ $\tilde{\sigma}_i^2$ ด้วยวิธีกำลังสองต่ำสุดสามัญ

ตัวสถิติจากการทดสอบด้วยอัตราส่วนภาวะน่าจะเป็นโดยอาศัยวิธีกำลังสองต่ำสุดสามัญในการประมาณค่าพารามิเตอร์ คือ

$$\text{LROLS} = n\ln S^2 - \sum_{i=1}^{4} n_i \ln S_i^2$$

โดยที่ $S^2 = \dfrac{e'e}{n-k}$

$$S_i^2 = \sum_{t=1}^{n_i} (e_{it} - \bar{e}_i)^2 / v_i$$

โดยที่ $\bar{e}_i = \displaystyle\sum_{t=1}^{n_i} \frac{e_{it}}{n_i}$

$$v_i = \frac{n-k}{w_i} - 1$$

$$w_i = \frac{n}{n_i}$$

ภายใต้ H_0 ,LROLS จะสามารถประมาณด้วยการแจกแจงไคสแควร์ ที่ระดับขั้นความเป็นเสรี m-1

3. การทดสอบของบาร์ตเลตต์ (Bartlette's test) :

ตัวสถิติจากการทดสอบของบาร์ตเลตต์ คือ

$$\text{BART} = \frac{(n-m)\ln S^2 - \sum_{i=1}^{m} n_i \ln S_i^2}{1 + \frac{1}{3(m-1)}\left(\sum_{i=1}^{m} \frac{1}{n_i - 1} - \frac{1}{n-m}\right)}$$

โดยที่ $S^2 = \dfrac{e'e}{n-k}$ และ

$$S_i^2 = \sum_{t=1}^{n_i} (e_{it} - \bar{e}_i)^2 / v_i$$

โดยที่ $\bar{e}_i = \displaystyle\sum_{t=1}^{n_i} \frac{e_{it}}{n_i}$

$$v_i = \frac{n-k}{w_i} - 1$$

$$w_i = \frac{n}{n_i}$$

ภายใต้ H_0 ,BART จะสามารถประมาณด้วยการแจกแจงไคสแควร์ ที่ระดับขั้นความเป็นเสรี m-1