บทที่ 2

สถิติที่ใช้ในการวิจัย

การศึกษาวิธีการตรวจสอบค่าผิดปกติในการวิเคราะห์ความถดถอยเชิงเส้นในครั้งนี้ เป็นการศึกษาเพื่อเปรียบเทียบความน่าจะเป็นของความผิดพลาดประเภทที่ 1, ความน่าจะเป็นซึ่งค่าผิดปรกติที่ถูกตรวจพบเป็นค่าผิดปรกติจริงทุกค่า$(p_1)^*$, ความน่าจะเป็นซึ่งทำให้เกิดมาซคกิงเอฟเฟ็ค$(p_2)^*$ และความน่าจะเป็นซึ่งทำให้เกิดชวอมพิงเอฟเฟ็ค$(p_3)^*$ ของตัวสถิติ ของ 4 วิธีการ ซึ่งได้แก่ **วิธีของเมอร์วิน จี มาราชิง** (Mervyn G. Marasinghe, 1985), **วิธีของฮาไดและไซมันนอฟฟ์** (Hadi and Simonoff, 1993), **วิธีเวียนเกิดโดยลำดับ** (Sequential recursive method) และ**วิธีเวียนเกิดดัดแปร** (Modified recursive method) ผู้วิจัยจะศึกษาในกรณีที่มีค่าผิดปรกติ 3 กรณีคือ 1, 2 และ 3 ค่า ตามลำดับ เมื่อความคลาดเคลื่อนมีการแจกแจงเบ้ผู้วิจัยจะทำการแปลงข้อมูลให้อยู่เข้าสู่การแจกแจงปรกติโดยใช้ **วิธีการแปลงภายใต้การแจกแจงข้อมูลของฟาเบียน แฮร์นันเดซและริชาร์ด เอ จอห์นสัน** (Fabian Hernandez and Richard A. Johnson, 1980) และจะทำการประมาณสัมประสิทธิ์การถดถอย $\underset{\sim}{b}$ ด้วยวิธีกำลังสองน้อยสุด

สมการการถดถอยที่ศึกษามีรูปแบบดังนี้

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

เมื่อ n คือขนาดตัวอย่าง

$\underset{\sim}{y}$ คือเวกเตอร์ของตัวแปรตามซึ่งมีขนาด n x 1

X คือเมทริกซ์ของตัวแปรอิสระขนาด n x p และมี rank = p

p คือจำนวนพารามิเตอร์ที่ต้องการประมาณ

$\underset{\sim}{\beta}$ คือเวกเตอร์ของสัมประสิทธิ์การถดถอยที่ไม่ทราบค่าขนาด p x 1

และ $\underset{\sim}{\varepsilon}$ คือเวกเตอร์ของความคลาดเคลื่อนซึ่งมีขนาด n x 1

---

* ดูคำจำกัดความในบทที่ 1 หน้า 13

โดยมีข้อสมมติว่า

1. $E(\underset{\sim}{\varepsilon}) = 0$
2. $\text{cov}(\underset{\sim}{\varepsilon}) = E\left(\underset{\sim}{\varepsilon}\,\underset{\sim}{\varepsilon}'\right) = \sigma^2 I_n$

การวิเคราะห์ความถดถอยเชิงเส้นเป็นการศึกษาและอธิบายแนวโน้มเวกเตอร์ค่าสังเกต $\underset{\sim}{y}$ ด้วยเวกเตอร์การประมาณ $\underset{\sim}{\hat{y}}$ ซึ่ง $\underset{\sim}{\hat{y}} = X\underset{\sim}{b}$ เมื่อ $\underset{\sim}{b}$ เป็นค่าประมาณของสัมประสิทธิ์การถดถอยด้วยวิธีกำลังสองน้อยสุด ดังนั้นเราจะได้ว่า

$$\underset{\sim}{b} = \left(X'X\right)^{-1}X'\underset{\sim}{y} \tag{2.1}$$

$\therefore$ เราแทนค่า $\underset{\sim}{b}$ ของสมการ(2.1) ใน $\underset{\sim}{\hat{y}} = X\underset{\sim}{b}$ จะได้ว่า

$$\underset{\sim}{\hat{y}} = X\left(X'X\right)^{-1}X'\underset{\sim}{y} = H\underset{\sim}{y} \tag{2.2}$$

เมื่อ $H = X\left(X'X\right)^{-1}X'$ เป็นเมทริกซ์ฉายาเชิงตั้งฉาก(projection matrix) หรือเรียกอีกอย่างหนึ่งว่าแฮทเมทริกซ์ (hat matrix) ซึ่ง H มีคุณสมบัติดังนี้

1. สมมาตร (symetry) กล่าวคือ $H = X\left(X'X\right)^{-1}X' = H'$
2. นิจพล (idempotent) กล่าวคือ $H^2 = H$
3. เมทริกซ์ H มี h_{ii} เป็นสมาชิกในเส้นทแยงมุม(diagonal) ซึ่งมีค่าระหว่าง 0 ถึง 1 และ $h_{ii} = \sum_{j=1}^{n} h_{ij}^2 = h_{ii}^2 + \sum_{j\neq i} h_{ij}^2$
4. rank(H) = rank(X) = p

จากสัมประสิทธิ์การถดถอยที่ประมาณได้เราสามารถนำมาหาค่าตกค้าง (residual) $\underset{\sim}{e}$ ซึ่งเป็นผลต่างระหว่างเวกเตอร์ค่าสังเกต $\underset{\sim}{y}$ กับเวกเตอร์การประมาณ $\underset{\sim}{\hat{y}}$ และเป็นค่าประมาณของความคลาดเคลื่อน $\underset{\sim}{\varepsilon}$ ได้ดังนี้

$$\begin{aligned}
\underset{\sim}{e} &= \underset{\sim}{y} - \underset{\sim}{\hat{y}} \\
&= \underset{\sim}{y} - X\underset{\sim}{b} \\
&= \underset{\sim}{y} - X\left(X'X\right)^{-1}X'\underset{\sim}{y} \\
&= \underset{\sim}{y} - H\underset{\sim}{y} \\
&= (I - H)\underset{\sim}{y}
\end{aligned}$$

โดยที่ความแปรปรวนร่วมเกี่ยวของค่าตกค้าง (residual covariance) คือ

$$\text{cov}(\underset{\sim}{e}) = (I - H)\, s^2$$

และมีความแปรปรวนค่าตกค้างของค่าสังเกตที่ i คือ

$$Var(e_i) = (1 - h_{ii})\, s^2$$

เมื่อ s^2 เป็นตัวประมาณที่ไม่เอนเอียงของ σ^2 ซึ่ง $s^2 = \sum_{i=1}^{n} e_i^2 / (n - p)$; $i = 1, 2, \ldots, n$

จากการวิเคราะห์ความถดถอยข้างต้น เราสามารถนำมาหาค่าตกค้าง(residual)ต่างๆได้ดังนี้

ก) ค่าตกค้างมาตรฐานภายใน (Internally Studentized Residual)

$$R_i = e_i / \left(s\sqrt{1 - h_{ii}} \right) \quad ; i = 1, 2, \ldots, n$$

ข) ค่าตกค้างมาตรฐานภายนอก (Externally Studentized Residual)

$$R_i^* = e_i / \left(s_{(-i)}\sqrt{1 - h_{ii}} \right) \quad ; i = 1, 2, \ldots, n$$

เมื่อ $s_{(-i)}$ คือค่าตกค้างกำลังสองเฉลี่ย(residual mean square) เมื่อตัดค่าสังเกตที่ i ออกไป(วิธีการหาค่า $s_{(-i)}$ ดูที่ภาคผนวก ค.)

ค) ค่าตกค้างที่ปรับแล้ว (Adjusted Residual)

$$A_i = e_i / \sqrt{1 - h_{ii}} \quad ; i = 1, 2, \ldots, n$$

ซึ่งเราจะนำค่าตกค้างเหล่านี้ไปใช้ในการตรวจสอบค่าผิดปรกติในวิธีต่างๆ ดังนี้

1. วิธีของเมอร์วิน จี มาราซิง (Mervyn G. Marasinghe's method(MV))

ในปี ค.ศ. 1985 เมอร์วิน จี มาราซิง ได้เสนอวิธีการตรวจสอบค่าผิดปรกติกรณีที่มีค่าผิดปกติหลายค่าในการวิเคราะห์ความถดถอยเชิงเส้น โดยมีหลักเกณฑ์คือ เราจะใช้ค่าตกค้างที่ปรับแล้ว (A_i) เป็นค่าบ่งชี้ค่าผิดปรกติ โดยการนำค่าตกค้างที่ปรับแล้วซึ่งมีค่าสูงสุดในแต่ละรอบ (จำนวนค่าสูงสุดที่เป็นไปได้คือ k^* รอบ) จำนวน k^* ค่า โดยเปรียบเสมือนว่าค่าสังเกต k^* ค่านี้เป็นค่าผิดปรกติ จากนั้นจึงทำการคำนวณตัวสถิติทดสอบ

$$F_k^* = (S - Q_k^*) / S$$

เมื่อ $S = (n - p)\, s^2$ เป็นผลบวกค่าตกค้างกำลังสอง (residual sum of square)

$s^2 = \sum_{i=1}^{n} e_i^2 / (n - p)$ เป็นค่าตกค้างกำลังสองเฉลี่ย (residual mean square)

และ $Q_k^* = \sum_{i=1}^{k^*} A_i^2$ เป็นผลบวกค่าตกค้างที่ปรับแล้วสูงสุดในแต่ละรอบ

โดยที่ $|A_i|$ คือค่าตกค้างที่ปรับแล้วสูงสุดรอบที่ i (maximum absolute adjusted residual) และ k^* คือ จำนวนรอบสูงสุดที่เป็นไปได้ซึ่งเป็นจำนวนเต็มตั้งแต่ 2 ถึง 5 และ $k^* \geq k$ เมื่อ k คือ จำนวนค่าผิดปรกติ (ในการครั้งนี้เราจะศึกษากรณีที่ k = 0, 1, 2 และ 3 ตามลำดับ)

ขั้นตอนการทดสอบ

1. เราจะหาเซตของค่าสังเกตที่มี k^* ค่า โดยพิจารณาจากค่าสัมบูรณ์ของค่าตกค้างที่ปรับแล้วสูงสุดในแต่ละรอบ ($|A_i|$) กล่าวคือ

1.1 คำนวณค่าตกค้างที่ปรับแล้วของแต่ละค่าสังเกตแล้วเลือกค่าที่ $\max |A_i|$ ตัดออกจากข้อมูลและนำค่าสังเกตดังกล่าวไปรวมไว้ในเซตย่อย K

1.2 คำนวณหา $\max |A_i|$ จากข้อมูลที่เหลือ n - 1 ค่าสังเกต แล้วตัดค่าสังเกตนั้นออกไปรวมไว้ในเซตย่อย K

1.3 ทำการคำนวณหา $\max |A_i|$ จากข้อมูลที่เหลือต่อไป จนกระทั่งได้เซตย่อย K ที่มีขนาด k^*

2. คำนวนผลบวกค่าตกค้างกำลังสอง (residual sum of square) ในแต่ละรอบ (Q_k^*)

3. คำนวนค่าสถิติทดสอบ F_k^*

4. ตรวจสอบค่าผิดปรกติโดยมีสมมติฐานคือ

H : ไม่มีข้อมูลผิดปรกติ

เทียบกับ K : มีข้อมูลผิดปรกติอย่างมากที่สุด k^* ค่า

โดยนำค่า F_k^* ที่คำนวณได้เปรียบเทียบกับค่าขอบเขตวิกฤติในตารางของเมอร์วิน จี มาราชิง (1985) ถ้า F_k^* คำนวณ < F_k^* ตาราง เราจะปฏิเสธสมมติฐานว่าง (H) ซึ่งแสดงว่าค่าสังเกตที่ i ที่มีค่าตรงกับค่า A_i เป็นค่าผิดปรกติ

5. กรณีที่ปฏิเสธสมมติฐานว่าง เราจะตัดค่าสังเกตดังกล่าวในข้อ 4 ออกไปแล้ววิเคราะห์ข้อมูลขนาด n - 1 เช่นเดียวกับข้อ 1 ถึง 4 ทำซ้ำจนกว่าจะยอมรับสมมติฐานว่างจึงหยุดทำการทดสอบ

6. สรุปผลการทดสอบพร้อมสรุปค่าผิดปรกติ

2. วิธีของฮาไดและไซมันนอฟฟ์ (Hadi and Simonoff's method(HS))

วิธีนี้เป็นวิธีการตรวจสอบค่าผิดปรกติหลายค่าในตัวแบบเชิงเส้น โดยการพยายามจัดข้อมูลออกเป็น 2 เซต คือ เซตของข้อมูลสะอาด* (set of clean data point) และเซตของข้อมูลที่น่าจะเป็นค่าผิดปรกติ (set of points that contain the potential outliers) และจะตรวจสอบข้อมูลที่น่าจะเป็นค่าผิดปรกติโดยดูจากระยะที่มันเบี่ยงเบนไปจากเซตของข้อมูลสะอาด

ขั้นตอนการตรวจสอบ

1. เราจะหาเซตของข้อมูลสะอาด (แทนสัญลักษณ์ด้วยเซต M) ที่มีขนาด h = [(n + p - 1) / 2] ดังนี้

1.1 คำนวณหาค่าสัมบูรณ์ของค่าตกค้างที่ปรับแล้ว ($|A_i|$) แล้วจัดเรียงค่าสังเกต ตาม $|A_i|$ จากน้อยไปมาก

1.2 แบ่งข้อมูลออกเป็น 2 เซตย่อยคือ

ก. <u>เซตย่อยพื้นฐาน</u> (basic subset : B) ซึ่งมีขนาด s = p + 1 ค่าสังเกตแรกที่ได้จากข้อ 1.1 มี X_B เป็นเมทริกซ์ไม่เอกฐาน แต่ถ้า X_B เป็นเมทริกซ์เอกฐาน เราจะทำให้เป็นเมทริกซ์ไม่เอกฐานโดยการรวมค่าสังเกตที่อยู่ถัดจาก p + 1 ค่าสังเกตแรกเข้าไปในเซตย่อยพื้นฐานจนกว่า X_B จะเป็นเมทริกซ์ไม่เอกฐาน และกำหนดให้ $\underset{\sim B}{b}$ เป็นตัวประมาณสัมประสิทธิ์การถดถอยของเซตย่อยพื้นฐาน

ข. <u>เซตย่อยไม่พื้นฐาน</u> (non-basic subset) มีขนาด n - p - 1 ค่าสังเกตหลังของข้อ 1.1 หรือมีขนาดเท่ากับจำนวนค่าสังเกตที่เหลือจากการจัดเซตย่อยพื้นฐานแล้ว

1.3 ทำการวิเคราะห์ความถดถอยเชิงเส้นของเซตย่อยพื้นฐานจะได้

$$e_{Bi} = y_i - x_i' \underset{\sim B}{b}$$

ซึ่งมี $$Var(e_{Bi}) = \begin{cases} 1 - x_i'\left(X_B'X_B\right)^{-1}x_i & ; i \in B \\ 1 + x_i'\left(X_B'X_B\right)^{-1}x_i & ; i \notin B \end{cases}$$

เมื่อ e_{Bi} คือค่าตกค้างที่เกิดจากเซตย่อยพื้นฐาน

X_B คือเมทริกซ์ของตัวแปรอิสระในเซตย่อยพื้นฐาน

และ x_i' คือสมาชิกในแถวที่ i ของ X

---

* เซตของข้อมูลสะอาด หมายถึง เซตที่ไม่ข้อมูลผิดปรกติรวมอยู่

1.4 คำนวณค่าสัมบูรณ์ค่าตกค้างมาตรฐาน ดังนี้

$$|B_i| = \begin{cases} \dfrac{\left|y_i - x_i' \underset{\sim}{b}_{B}\right|}{\sqrt{1 - x_i'\left(X_B' X_B\right)^{-1} x_i}} & ; i \in B \\ \dfrac{\left|y_i - x_i' \underset{\sim}{b}_{B}\right|}{\sqrt{1 + x_i'\left(X_B' X_B\right)^{-1} x_i}} & ; i \notin B \end{cases}$$

1.5 จัดเรียงค่าสังเกตตามค่า $|B_i|$ จากน้อยไปมาก แล้วจัดเป็นเซตย่อยพื้นฐานที่มีขนาด $s + 1$ ค่าสังเกต

1.6 คำนวณหาเซตย่อยพื้นฐานต่อไปเหมือนข้อ 1.3 ถึง 1.5 จนกระทั่งเซตย่อยพื้นฐานมีขนาดเท่ากับเซต M คือมีขนาด $h = (n + p - 1) / 2$

2. หาตัวแบบที่เหมาะสมของเซต M แล้วคำนวณหาค่าตกค้างมาตรฐาน (M_i) ของเซต M โดยที่

$$|M_i| = \begin{cases} \dfrac{\left|y_i - x_i' \underset{\sim}{b}_{M}\right|}{s_M \sqrt{1 - x_i'\left(X_M' X_M\right)^{-1} x_i}} & ; i \in M \\ \dfrac{\left|y_i - x_i' \underset{\sim}{b}_{M}\right|}{s_M \sqrt{1 + x_i'\left(X_M' X_M\right)^{-1} x_i}} & ; i \notin M \end{cases}$$

3. จัดเรียงค่าสังเกตตาม ($|M_i|$) จากน้อยไปมาก และกำหนด $M_{(h+1)}$ เป็นตัวสถิติอันดับที่ $h + 1$ ของ $|M_i|$ โดยที่ h คือขนาดของเซต M ปัจจุบัน เราจะทำการทดสอบสมมติฐานว่าง

H : ค่าสังเกตที่ i เป็นค่าผิดปรกติ ; $i = h + 1, h + 2, \ldots, n$

เทียบกับ K : ค่าสังเกตที่ i ไม่เป็นค่าผิดปรกติ

ก) ถ้า $|M_{(h+1)}| \geq t_{(\alpha/2\,(h+1),\,h-p)}$ เราจะสรุปว่าค่าสังเกตทั้งหมดที่ $|M_{(i)}| \geq t_{(\alpha/2\,(h+1),\,h-p)}$ เป็นค่าผิดปรกติและหยุดทำการทดสอบ

ข) กรณีอื่นให้จัดเซตย่อย M ใหม่ โดยให้มีขนาด $h+1$ ค่าสังเกต ถ้า $n = h+1$ สรุปว่าไม่มีค่าผิดปรกติเกิดขึ้น จึงหยุดทำการทดสอบ แต่ถ้า $n \neq h + 1$ กลับไปเริ่มทำจากข้อ 2 ใหม่

3. วิธีของไคนิฟาร์ดและซวอลโล (Kianifard and Swallow's method)

ในปี ค.ศ. 1990 ไคนิฟาร์ด(Kianifard) และซวอลโล(Swallow) ได้เสนอวิธีการตรวจสอบค่าผิดปรกติโดยใช้เกณฑ์ค่าตกค้างเวียนเกิด (recursive residual) ซึ่งค่าตกค้างเวียนเกิดคือ

$$|W_i| = \frac{\left| y_i - x_i' \underset{\sim}{b}_{i-1} \right|}{\sqrt{1 + x_i' \left(X_{i-1}' X_{i-1} \right)^{-1} x_i}} \quad ; i = p+1, p+2, \ldots, n$$

โดยที่ $\underset{\sim}{b}_{i-1} = \left(X_{i-1}' X_{i-1} \right)^{-1} X_{i-1}' \underset{\sim}{y}_{i-1}$

X_{i-1} คือเมตริกซ์ขนาด (i-1) x p ที่ประกอบด้วยสมาชิก i-1 แถวแรกของ X

และ $\underset{\sim}{y}_{i-1}$ คือเวกเตอร์ย่อยที่ประกอบด้วยสมาชิก i-1 แถวแรกของ $\underset{\sim}{y}$

ในปี ค.ศ. 1950 แพลคเกตต์ (Plackett), และในปี ค.ศ. 1975 บราวน์และคณะ (Brown et al.) ได้เสนอวิธีการหาค่า W_i โดยการหา $\underset{\sim}{b}_i$ จาก $\underset{\sim}{b}_{i-1}$ ดังนี้

$$\underset{\sim}{b}_i = \underset{\sim}{b}_{i-1} + \frac{\left(X_{i-1}' X_{i-1} \right)^{-1} x_i \left(y_i - x_i' \underset{\sim}{b}_{i-1} \right)}{1 + x_i' \left(X_{i-1}' X_{i-1} \right)^{-1} x_i} \quad ; i = p+1, p+2, \ldots, n$$

เมื่อ $\left(X_i' X_i \right)^{-1} = \left(X_{i-1}' X_{i-1} \right)^{-1} - \frac{\left(X_{i-1}' X_{i-1} \right)^{-1} x_i x_i' \left(X_{i-1}' X_{i-1} \right)^{-1}}{1 + x_i' \left(X_{i-1}' X_{i-1} \right)^{-1} x_i}$

และ $S_i = S_{i-1} + W_i^2 = (\underset{\sim}{y}_i - X_i \underset{\sim}{b}_i)'(\underset{\sim}{y}_i - X_i \underset{\sim}{b}_i)$; S_i คือผลบวกค่าตกค้างกำลังสองของค่าสังเกต i ค่า

วิธีการตรวจสอบค่าผิดปรกติโดยใช้เกณฑ์ค่าตกค้างเวียนเกิดที่ ไคนิฟาร์ดและซวอลโลเสนอในปี 1990 มี 2 วิธี ได้แก่ วิธีเวียนเกิดโดยลำดับ (Sequential recursive method) และวิธีเวียนเกิดดัดแปร (modified recursive method) ซึ่งแต่ละวิธีมีขั้นตอนดังนี้

3.1 วิธีเวียนเกิดโดยลำดับ (sequential recursive method (SRM))

1. คำนวณหาค่าสัมบูรณ์ของค่าตกค้างที่ปรับแล้ว($|A_i|$) แล้วจัดเรียงค่าสังเกตตาม $|A_i|$ จากน้อยไปมาก

2. ใช้ p ค่าสังเกตแรกในข้อ 1 มาจัดเป็นเซตพื้นฐานเพื่อใช้คำนวณค่าตกค้างเวียนเกิด (W_i)

3. คำนวณค่า W_i และคำนวณหาค่า $W_i / s_{(-i)}$; i = p+1, p+2, ... , n

4. คำนวณหาค่า $\max|W_i / s_{(-i)}|$

5. ตรวจสอบค่าผิดปรกติโดยใช้สมมติฐาน

H : ไม่มีข้อมูลผิดปรกติ

เทียบกับ K : มีข้อมูลผิดปรกติอย่างน้อย 1 ค่า

โดยเราจะทำการเปรียบเทียบค่าสถิติที่คำนวณได้กับขอบเขตวิกฤติจากตารางการแจกแจงที ถ้า $\max|W_i / s_{(-i)}| > t_{\alpha/2n,n-p-1}$ เราจะปฏิเสธสมมติฐานว่างและสรุปว่าค่าสังเกตที่มี $\max|W_i / s_{(-i)}|$ เป็นค่าผิดปรกติ

6. กรณีปฏิเสธสมมติฐานว่าง เราจะตัดค่าสังเกตดังกล่าวในข้อ 5 ออกแล้ววิเคราะห์ข้อมูลขนาด n - 1 จากข้อ 1 ถึง 5 อีกครั้ง และทำไปจนกว่าจะยอมรับสมมติฐานว่างจึงหยุดทำการทดสอบ

7. สรุปผลการทดสอบพร้อมสรุปค่าผิดปรกติ

3.2 วิธีเวียนเกิดดัดแปร (modified recursive method (MRM))

1. ในขั้นแรกเราจะทำเช่นเดียวกับข้อ 1 ถึง 5 ในวิธี SRM

2. ถ้าปฏิเสธสมมติฐานว่าง เราจะเปรียบเทียบ $|W_i / s_{(-i)}|$ ที่เหลือกับ $t_{\alpha/2\,,\,n-p-1}$ ถ้าค่าสังเกตใดที่มี $|W_i / s_{(-i)}| > t_{\alpha/2\,,\,n-p-1}$ แสดงว่าค่าสังเกตนั้นเป็นค่าผิดปรกติ

3. สรุปผลการทดสอบพร้อมกับสรุปค่าผิดปรกติ

การแปลงข้อมูลภายใต้การแจกแจงความน่าจะเป็น

ในกรณีที่การแจกแจงของความคลาดเคลื่อนไม่มีการแจกแจงปรกติหรือกรณีทั่วไปที่สเกลการวัดของตัวแปรไม่เหมาะสมกับสถิติวิเคราะห์ โดยทั่วไปนักสถิติจะแก้ปัญหาโดยการแปลงข้อมูลเพื่อให้มีคุณสมบัติที่เหมาะสม การแปลงหลักๆ ที่ใช้กันอยู่คือ การแปลงแบบยกกำลัง (Power transformation) ของ Box และ Cox(1964) ซึ่งมีรูปแบบการแปลง คือ

$$y^{(\lambda)} = \begin{cases} y^{\lambda} - 1 & ; \lambda \neq 0 \\ \log y & ; \lambda = 0 \end{cases} \qquad ; y > 0$$

ซึ่งจะทำให้ $y^{(\lambda)}$ ลู่เข้าสู่การแจกแจงปรกติ สำหรับวิธีการเลือกค่า λ ที่จะใช้ในการแปลงข้อมูลของ Box และ Cox จะเลือกจากค่า λ ที่ทำให้ $L(\lambda)$ มีค่าสูงสุด โดยที่

$$L(\lambda) = \begin{cases} \frac{n}{2}\log(\lambda^2) - \frac{n}{2}\log(RSS_\lambda) + (\lambda - 1)\sum_{i=1}^{n}\log(y_i) & ; \lambda \neq 0 \\ -\frac{n}{2}\log(RSS_\lambda) - \sum_{i=1}^{n}\log(y_i) & ; \lambda = 0 \end{cases}$$

เมื่อ RSS_λ คือผลบวกตกค้างกำลังสอง ซึ่งเราสามารถเลือกค่า λ โดยเลือกจากค่า λ ที่ทำให้ RSS_λ ต่ำสุดได้เช่นกัน

วิธีการแปลงของ Box และ Cox เป็นการแปลงข้อมูลภายใต้ตัวอย่างสุ่มของตัวแปรสุ่มเชิงบวกขนาด n โดยเราทำการเลือกพารามิเตอร์สำหรับการแปลงที่เหมาะสมกับตัวอย่างสุ่มชุดนั้น ดังนั้นตัวอย่างสุ่มแต่ละชุดในการแจกแจงเดียวกันก็จะทำให้ได้พารามิเตอร์ในการแปลงที่แตกต่างกัน ในปี ค.ศ. 1980 ฟาเบียน เฮอร์นันเดซและริชาร์ด เอ จอห์นสัน (Fabian Hernandez and Richard A. Johnson) ได้เสนอวิธีการแปลงข้อมูลเพื่อให้มีการแจกแจงปรกติเมื่อทราบการแจกแจงของข้อมูล โดยในการแปลงจะเป็นการแปลงตัวอย่างสุ่มภายใต้การแจกแจงความน่าจะเป็น และใช้จำนวนสารนิเทศคูลแบค-เลเบอร์ (Kullback - Leibler , 1968) เป็นหลักในการเลือกพารามิเตอร์ λ ที่จะใช้ในการแปลง ซึ่งมีรายละเอียดดังนี้

กำหนดให้ h_1 และ h_2 เป็นฟังก์ชันความหนาแน่นของความน่าจะเป็นและต่อเนื่องของตัวแปร Z มีค่าความแตกต่างระหว่างการแจกแจง คือ

$$I[h_1, h_2] = \int h_1(t)\log\left\{\frac{h_1(t)}{h_2(t)}\right\} dt$$

เราเรียกความแตกต่างระหว่างการแจกแจงดังกล่าวว่า **จำนวนสารนิเทศคูลแบค-เลเบอร์**

ถ้า y เป็นตัวแปรสุ่มเชิงบวก (positive random variable) ที่มีฟังก์ชันความหนาแน่น (probability density function (pdf)) คือ $g(\cdot)$ ในการเลือก λ ที่จะใช้ในการแปลง y ให้มีการแจกแจงปรกติ เราจะเลือก λ ที่ทำให้จำนวนสารนิเทศ $I[f_\lambda; \phi_{\mu\sigma}]$ ซึ่งแสดงค่าความแตกต่างระหว่างการแจกแจงของข้อมูลที่แปลงด้วย λ กับข้อมูลที่มีการแจกแจงปกติด้วยค่าเฉลี่ย μ และส่วนเบี่ยงเบนมาตรฐาน σ มีค่าต่ำสุดโดยที่

$$I[f_\lambda; \phi_{\mu\sigma}] = \int f_\lambda(a)\log\left\{\frac{f_\lambda(a)}{\phi_{\mu\sigma}(a)}\right\} da$$

เมื่อ $f_\lambda(a)$ เป็นฟังก์ชันความหนาแน่นของความน่าจะเป็นหนาแน่นของตัวแปร $A = y^{(\lambda)}$
และ $\phi_{\mu\sigma}$ เป็นฟังก์ชันความหนาแน่นของความน่าจะเป็นหนาแน่นของการแจกแจงปรกติด้วยค่าเฉลี่ย μ และส่วนเบี่ยงเบนมาตรฐาน σ

ในขั้นแรกเราจะเลือกค่า μ และ σ ด้วย f_λ ซึ่งจะทำให้ A มีการแจกแจงปรกติโดยประมาณก่อนแล้วจึงจะหาค่า λ ที่ทำให้จำนวนสารนิเทศมีค่าต่ำสุด หรือมีระยะทางน้อยที่สุด กล่าวคือ

กำหนด λ คงที่ และ ให้

$$y^{(\lambda)} = \begin{cases} y^\lambda - 1 & ; \lambda \neq 0 \\ \log y & ; \lambda = 0 \end{cases} \quad ; y > 0$$

สมมติว่าค่าคาดหวัง $E_g(y^{2\lambda})$ และ $E_g[(\log y)^2]$ มีค่าจำกัด ดังนั้นค่า μ และ σ ที่ทำให้จำนวนสารนิเทศ $I[f_\lambda; \phi_{\mu\sigma}]$ มีค่าต่ำสุดคือ

$$\mu_*(\lambda) = E_g(y^{(\lambda)})$$

และ $$\sigma_*^2(\lambda) = E_g[y^{(\lambda)} - E_g(y^{(\lambda)})]^2 = V_g(y^{(\lambda)})$$

สำหรับ $\mu_*(\lambda)$ และ $\sigma_*^2(\lambda)$ ที่เลือกแล้วเราจะได้

$$G(\lambda) = \min_{\mu,\sigma^2} I\left[f_\lambda ; \phi_{\mu\sigma}\right] \qquad (2.3)$$
$$= \frac{1}{2}\left[\log(2\pi)+1\right] + E_g\left\{\log\left[g(y)\right]\right\}$$
$$+ (1-\lambda)E_g\left[\log(y)\right] + \frac{1}{2}\log\left[Var_g\left(y^{(\lambda)}\right)\right]$$

ซึ่ง λ_* ที่ทำให้ G(·) มีค่าต่ำสุดเป็นพารามิเตอร์ที่เราจะใช้ในการแปลง y ให้มีการแจกแจงปรกติ

ตัวอย่างที่ 1 จงหาจำนวนสารนิเทศ $I\left[f_\lambda ; \phi_{\mu\sigma}\right]$ เพื่อใช้หาค่า λ_* สำหรับการแจกแจงแกมมา

จำนวนสารนิเทศฯ สำหรับการแจกแจงแกมมา หาได้ดังนี้

การแจกแจงแกมมามีสูตรของความหนาแน่นดังนี้

$$g(y) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} \quad ; y > 0 \; ; \; \alpha, \beta > 0$$

ถ้าเรากำหนดให้ $\beta = 1$ จะได้ฟังก์ชันความหนาแน่นของการแจกแจงแกมมา

อยู่ในรูปของ $g(y) = \frac{1}{\Gamma(\alpha)} y^{\alpha-1} e^{-y} \quad ; y > 0 \; ; \; \alpha > 0$

จากสมการที่ (2.3) เราจะหาค่าของแต่ละพจน์ดังนี้

$$E_g\left\{\log\left[g(y)\right]\right\} = E_g\left\{(\alpha-1)\log(y) - y - \log[\Gamma(\alpha)]\right\}$$
$$= E_g\left\{(\alpha-1)\log(y)\right\} - E_g(y) - E_g\left\{\log[\Gamma(\alpha)]\right\}$$
$$= (\alpha-1)\int_0^{\infty} \frac{\log(y) y^{\alpha-1} e^{-y}}{\Gamma(\alpha)}\, dy - \int_0^{\infty} \frac{y\, y^{\alpha-1}}{\Gamma(\alpha)} dy - \log[\Gamma(\alpha)]\, dy$$
$$= (\alpha-1)\left[\frac{\Gamma'(\alpha)^*}{\Gamma(\alpha)}\right] - \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)} - \log[\Gamma(\alpha)]$$
$$= (\alpha-1)\psi(\alpha)^{**} - \alpha\frac{\Gamma(\alpha)}{\Gamma(\alpha)} - \log[\Gamma(\alpha)]$$
$$= \alpha\psi(\alpha) - \psi(\alpha) - \alpha - \log[\Gamma(\alpha)]$$

$$E_g\left[\log(y)\right] = \int_0^{\infty} \frac{\log(y) y^{\alpha-1} e^{-y}}{\Gamma(\alpha)} dy$$
$$= \psi(\alpha)$$

---

* $\int_0^{\infty} \log(y) y^{\alpha-1} e^{-y} dy = \Gamma'(\alpha)$

** $\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = \psi(\alpha)$

$$Var_g\left(y^{(\lambda)}\right) = V_g\left(\frac{y^\lambda - 1}{\lambda}\right)$$

$$= \frac{1}{\lambda^2} V\left(y^\lambda\right)$$

จาก $Var(y) = E(y^2) - \{E(y)\}^2$ เราจะได้ว่า

$$Var_g\left(y^\lambda\right) = E_g\left(y^{2\lambda}\right) - \left\{E_g\left(y^\lambda\right)\right\}^2$$

$$= \int_0^\infty \frac{y^{2\lambda} y^{\alpha-1} e^{-y}}{\Gamma(\alpha)}\, dy - \left\{\int_0^\infty \frac{y^\lambda y^{\alpha-1} e^{-y}}{\Gamma(\alpha)}\, dy\right\}^2$$

$$= \frac{\Gamma(2\lambda+\alpha)}{\Gamma(\alpha)} - \left\{\frac{\Gamma(\lambda+\alpha)}{\Gamma(\alpha)}\right\}^2$$

$$= \frac{\Gamma(\alpha)\Gamma(2\lambda+\alpha) - \{\Gamma(\lambda+\alpha)\}^2}{\{\Gamma(\alpha)\}^2}$$

ดังนั้นเราจะได้ $Var_g\left(y^{(\lambda)}\right) = \dfrac{\Gamma(\alpha)\Gamma(2\lambda+\alpha) - \{\Gamma(\lambda+\alpha)\}^2}{\lambda^2\{\Gamma(\alpha)\}^2}$

เมื่อเราแทนค่าผลลัพธ์ของพจน์ต่างๆในสมการ(2.3) จะได้ว่า

$$G(\lambda) = \frac{1}{2}[\log(2\pi)+1] + \{\alpha\psi(\alpha) - \psi(\alpha) - \alpha - \log[\Gamma(\alpha)]\}$$

$$+ \{(1-\lambda)\psi(\alpha)\} + \left\{\frac{1}{2}\log\left[\frac{\Gamma(\alpha)\Gamma(2\lambda+\alpha) - [\Gamma(\lambda+\alpha)]}{\lambda^2[\Gamma(\alpha)]^2}\right]\right\}$$

$$= \frac{1}{2}[\log(2\pi)+1] + \alpha\psi(\alpha) - \psi(\alpha) - \alpha - \log[\Gamma(\alpha)]$$

$$+ \psi(\alpha) - \lambda\psi(\alpha) + \frac{1}{2}\log\left[\frac{\Gamma(\alpha)\Gamma(2\lambda+\alpha) - [\Gamma(\lambda+\alpha)]}{\lambda^2}\right]$$

$$- \frac{1}{2}\log\{[\Gamma(\alpha)]^2\}$$

$$= \frac{1}{2}\log[(2\pi)+1] - 2\log[\Gamma(\alpha)] + \alpha[\psi(\alpha) - 1]$$

$$- \lambda\psi(\alpha) + \frac{1}{2}\log\left\{\frac{\Gamma(\alpha)\Gamma(2\lambda+\alpha) - [\Gamma(\lambda+\alpha)]^2}{\lambda^2}\right\}$$

หมายเหตุ ψ คือฟังก์ชันไดแกมมา*(digamma function) ซึ่งมีรูปแบบฟังก์ชัน คือ

$$\psi(\alpha) = \frac{d}{d\alpha}\log[\Gamma(\alpha)]$$

$$= \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$$

$$= \ln\alpha - \frac{1}{2\alpha} - \frac{1}{12\alpha^2} + \frac{1}{120\alpha^4} + \frac{1}{252\alpha^6} + \cdots$$

ค่าประมาณของฟังก์ชันไดแกมมาเราจะดูจากตารางที่ ง.1 ในภาคผนวก

---

* William H. Beyer, CRC Standard Methematical Tables, 26th ed. (Florida: CRC Press, 1981) , p. 399