บทที่ 9
ตัวแปรตามที่เป็นตัวแปรเชิงคุณภาพหรือมีค่าจำกัด
(Qualitative and Limited Dependent Variables)

9.1 ความหมายและความสำคัญ

คำว่า Qualitative Dependent Variable หมายถึงกรณีที่ตัวแปรตาม Y เป็นตัวแปรเชิงคุณภาพ หรือ Categorical Variable ซึ่งนักวิจัยจำเป็นต้องกำหนดดัชนีสำหรับค่าตัวแปรดังกล่าวในรูปของตัวแปรดัมมี่ หรือกล่าวอย่างไม่เป็นทางการได้ว่า Qualitative Dependent Variable ก็คือตัวแปรตามที่เป็นตัวแปรดัมมี่นั่นเอง เราเรียกสมการ $Y = f(X\text{'s})$ ที่ Y เป็น Qualitative Dependent Variable ว่า Quantal Response Model

คำว่า Limited Dependent Variable คือตัวแปรตาม Y ที่มีค่าจำกัดอยู่เฉพาะในบางช่วงของ Real Line เช่น $Y > 0$ (เรียกว่าปัญหา Truncated Sample) หรือตัวแปร Y มีค่าได้ตลอดช่วงของ Real Line ยกเว้นบางช่วง (เรียกว่าปัญหา Censored Sample)

ทั้งกรณี Qualitative Dependent Variable และ Limited Dependent Variable เป็นสถานการณ์ที่มีผลให้ไม่อาจใช้ OLS -Estimator ตามปกติได้เพราะ OLS - Estimator ในกรณีเหล่านี้จะเป็น Biased Estimator และ Inconsistent สำหรับการศึกษาในที่นี้เป็นการศึกษาถึงกรณีของ Qualitative Dependent Variable เท่านั้น ผู้สนใจสามารถศึกษาเรื่อง Limited Dependent Variable ได้จากเอกสารอ้างอิง

พิจารณาสมการถดถอย $Y_i = \alpha + \beta X_i + u_i \quad ; i = 1,2,\ldots,M$ โดยที่ Y เป็นตัวแปรเชิงสุ่มที่มีค่าเป็น 0 และ 1 เท่านั้น เช่น $Y_i = 1$ ถ้าหน่วยสำรวจที่ i คือ ผู้ป่วยที่เป็นมะเร็งบริเวณศีรษะและลำคอ และ $Y_i = 0$ ถ้าหน่วยสำรวจคือผู้ป่วยที่เป็นมะเร็งบริเวณอื่น ๆ หรือ $Y_i = 0$ ถ้าหน่วยสำรวจคือผู้มีเงินได้ที่มีการศึกษาระดับอุดมศึกษา และ $Y_i = 1$ ถ้าหน่วยสำรวจคือผู้มีเงินได้ที่มีการศึกษาระดับอื่น ๆ

จากตัวอย่างนี้จะเห็นได้ว่า Y_i มีค่าได้ 2 ค่าคือ 0 และ 1 ผลสะท้อนที่ตามมาก็คือ u_i จะมีค่าได้ 2 ค่าคือ $u_i = Y-(\alpha+\beta X_i) = 1-(\alpha+\beta X_i)$ ถ้า $Y_i = 1$ และ $u_i = -(\alpha+\beta X_i)$ ถ้า

ปัญหาที่ปรากฏอย่างเด่นชัดก็คือ

1. เราไม่อาจควบคุมให้ $E(u_i) = 0$ ได้โดยง่าย กล่าวคือถ้าต้องการให้ข้อตกลงว่า $E(u_i) = 0$ เป็นจริงเรามีความจำเป็นต้องควบคุมให้ u_i ปรากฏค่าด้วยความน่าจะเป็น $(\alpha+\beta X_i) = p_i$ หรือ $1-(\alpha+\beta X_i) = q_i$ เท่านั้น[1] แต่การควบคุมให้ $0 < p_i < 1$ และ $0 < q_i < 1$ เป็นเรื่องที่เป็นไปได้ยากเพราะโดยปกติ $(\alpha + \beta X_i)$ อาจมีค่ามากกว่า 1 หรือแม้กระทั่งเป็นประมาณติดลบ คือ น้อยกว่า 0 ได้

2. $V(u_i)$ จะไม่คงที่ เพราะในกรณีนี้ตัวแปรสุ่ม Y มีการแจกแจงแบบเบอร์นุลลิ (Bernoulli Distribution) ดังนั้น $V(u_i) = p_i q_i = (\alpha + \beta X_i)(1 - \alpha - \beta X_i)$ ซึ่งจะไม่คงที่แต่จะแปรเปลี่ยนไปตามค่าของตัวแปร X ปัญหา Heteroscedasticity จึงเกิดขึ้น

3. สมการประมาณค่า $Y_i = \hat{\alpha} - \hat{\beta} X_i$ จะหวั่นไหว (Sensitive) ได้ง่ายเมื่อ X มีค่าเปลี่ยนแปลงไป

4. เราไม่อาจใช้ t-test, F-test ได้รวมทั้ง R^2 ก็ขาดความหมายที่แท้จริง และไม่ใช้เป็นดัชนีในการตัดสินใจได้อีกทั้ง $\hat{\sigma}^2$ ยังเป็น Inconsistent Estimator ของ σ^2 นอกจากนี้แล้วเรายังไม่อาจใช้วิธีการประมาณค่าต่าง ๆ ที่อาศัย Linear Function ของ Y_i มาประมาณค่าพารามิเตอร์ เช่น OLS, GLS ได้เพราะ Y มิได้แจกแจงแบบปกติ เว้นแต่จะได้รับการปรับโครงสร้างและสภาพทั่วไปเสียก่อน

9.2 **ตัวอย่าง**

เพื่อความเข้าใจปัญหาและติดตามการพัฒนาทางทฤษฎีได้โดยไม่ติดขัดผู้เขียนขอยกตัวอย่างงานวิจัยที่ดำเนินการโดยที่ตัวแปร Y เป็นตัวแปรเชิงคุณภาพ (Dichotomous หรือ Binary Choice Model) ดังนี้

จากข้อมูลเกี่ยวกับความพอใจของทหารว่าพึงพอใจที่จะประจำกองพลในภาคเหนือ หรือภาคใต้ ($Y_i = 1$ ถ้าทหารนายที่ i พอใจจะประจำกองพลในภาคเหนือ $Y_i = 0$ ถ้าทหารนายที่ i พอใจจะประจำกองพลในภาคใต้) โดยที่ตัวแปรอิสระคือ X_2 = ภูมิลำเนาเดิม ($X_2 = 1$ ถ้ามีภูมิลำเนาเดิมอยู่ภาคเหนือ $X_2 = 0$ ถ้าภูมิลำเนาเดิมอยู่ภาคใต้) X_3 = ที่ตั้งปัจจุบันของกองพลที่สังกัด ($X_3 = 1$ ถ้าสังกัดปัจจุบันอยู่ในภาคเหนือ $X_3 = 0$ ถ้าสังกัดปัจจุบันอยู่ในภาคใต้)

---

[1] ตัวแปร Y_i มีการแจกแจงแบบเบอร์นุลลิ กล่าวคือ โดยทั่วไปถ้า X มี pdf เป็น $f(x) = p^x q^{1-x}$; $x = 0, 1$ เราเรียกว่า f(x) ว่า Bernoulli Distribution โดยที่ถ้า $x = 0$ จะได้ $f(x) = q$ และถ้า $x = 1$ จะได้ $f(x) = p$

ผลการบันทึกข้อมูลจากทหาร 8,036 นาย ปรากฏข้อมูลดังตาราง [1]

j	k	ℓ	X_2 ภูมิลำเนาเดิม	X_3 ที่ตั้งสังกัดปัจจุบัน	Y ที่ตั้งสังกัดที่ต้องการ	ความถี่ $(f_{jk\ell})$
1	1	1	เหนือ	เหนือ	เหนือ	1342
1	1	0	เหนือ	เหนือ	ใต้	198
1	0	1	เหนือ	ใต้	เหนือ	1750
1	0	0	เหนือ	ใต้	ใต้	760
0	1	1	ใต้	เหนือ	เหนือ	487
0	1	0	ใต้	เหนือ	ใต้	446
0	0	1	ใต้	ใต้	เหนือ	472
0	0	0	ใต้	ใต้	ใต้	2581

และเมื่อจำแนกกลุ่มตัวอย่างจากตารางข้างต้นเพื่อพิจารณาความน่าจะเป็นจะพบว่าเราสามารถเสนอตารางได้ดังนี้

j k	X_2 ภูมิลำเนาเดิม	X_3 ที่ตั้งสังกัดปัจจุบัน	Y : ที่ตั้งสังกัดที่ต้องการ						รวม
			ภาคเหนือ (ℓ=1)			ภาคใต้ (ℓ=0)			
			f_{jk1}	p_{jk}	ω_{jk}	f_{jk0}	q_{jk}	$\frac{1}{\omega_{jk}}$	(n_{jk})
1 1	เหนือ	เหนือ	1342	.871	*6.778*	198	.129	.148	1540
1 0	เหนือ	ใต้	1750	.697	2.303	760	.303	.434	2510
0 1	ใต้	เหนือ	487	.522	1.092	446	.478	.916	933
0 0	ใต้	ใต้	472	.155	0.183	2581	.845	5.468	3053
			4051	.504	1.020	3985	.496	.984	8036

[1] **Goodman, Leo, A., Analyzing Qualitative Categorical Date . Log - Linear Models and Latent - Structure Analysis. (Addison , Wesley Publishing Company Advanced Book Program. London,** 1978), p.27-56

ข้อสังเกต

1. P_{jk} (หรือ $P_{jk}^{X_2X_3}$) = Prob(ต้องการสังกัดภาคเหนือ $| X_2 = X_{2j}$ และ $X_3 = X_{3k}$)
$= \frac{f_{jk1}}{n_{jk}}$ เช่น

$\hat{P}_{11}$ = Pr (ต้องการสังกัดภาคเหนือ $| X_2 = 1, X_3 = 1) = \frac{1342}{1540} = .871$

$\hat{P}_{13}$ = Pr (ต้องการสังกัดภาคเหนือ $| X_2 = 0, X_3 = 1) = \frac{487}{933} = .522$

2. ω_{jk} (หรือ $\omega_{jk}^{X_2X_3}$) = Odds (ต้องการสังกัดภาคเหนือ$| X_2 = X_{2j}, X_3 = X_{3k}$)

$= \frac{f_{jk1}}{f_{jk0}}$ = อัตราส่วน (Ratio)

$\hat{\omega}_{11}$ = Odds (ต้องการสังกัดภาคเหนือ $| X_2 = 1, X_3 = 1) = \frac{1342}{198} = 6.778$

= 6.778 : 1 หรือทหารที่มีภูมิลำเนาอยู่ภาคเหนือที่สังกัดกองพลในภาคเหนือ แต่ต้องการสังกัดเหนือบ้างภาคใต้บ้างคิดเป็นอัตราส่วน 6.778 ต่อ 1

$\hat{\omega}_{14}$ = Odds (ต้องการสังกัดภาคเหนือ $X_2 = 0, X_3 = 0) = \frac{472}{2581} = .183$

3. Odds และ Prob สามารถใช้คำนวณย้อนสู่กันและกันได้ดังนี้

$$\text{Odds} = \frac{\text{Prob}}{(1-\text{Prob})}$$

$$\text{Prob} = \frac{\text{Odds}}{(1+\text{Odds})}$$

เช่น

$\hat{\omega}_{11} = 6.778 \qquad \hat{P}_{11} = \hat{\omega}_{11} / (1+\hat{\omega}_{11}) = 6.778/7.778 = .871$

$\hat{\omega}_{14} = .183 \qquad \hat{P}_{14} = \hat{\omega}_{14} / (1+\hat{\omega}_{14}) = .183/1.183 = .155$

$\hat{P}_{11} = .871 \qquad \hat{\omega}_{11} = \hat{P}_{11} / (1-\hat{P}_{11}) = .871/.129 = 6.751$

$\hat{P}_{13} = .552 \qquad \hat{\omega}_{13} = \hat{P}_{13} / (1-\hat{P}_{13}) = .552/.478 = 1.155$

หมายเหตุ ค่า $\hat{\omega}_{11}$ และ $\hat{\omega}_{13}$ ตามตัวอย่างนี้คลาดเคลื่อนไปเพราะการปัดเศษ ถ้าใช้ค่าที่ไม่ปัดเศษจะได้เท่ากับในตารางซึ่งคำนวณโดยตรงตามข้อ 1 และ 2

ตัวอย่างนี้เป็นตัวอย่างที่ชี้ให้เห็นว่าในการวิเคราะห์สมการถดถอยที่ตัวแปร Y เป็น *Quantal* เราจำเป็นต้องจำแนกค่าสังเกตออกเป็นกลุ่ม ๆ (*Subsample*) เรียกว่า *Setting* ของตัวแปรอิสระ จากนั้นจึงคำนวณหาค่าความน่าจะเป็น p และ q หรือ Odds เพื่อนำไปใช้ประโยชน์ในการวิเคราะห์ต่อไป ในชั้นต้นนี้ผู้เขียนขอแสดงตัวอย่างนี้เพียงระดับการกำหนด Setting ของตัวแปรอิสระ[1] และการหาค่าความน่าจะเป็นเท่านั้น ซึ่งคิดว่าน่าจะมีประโยชน์เพียงพอที่จะใช้เป็นพื้นความรู้ในการศึกษาเรื่องของ Probitและ Logit ต่อไป และจะเสนอการประมาณค่าพารามิเตอร์ต่อไปเมื่อศึกษาถึงเรื่องนั้น

9.3 **การประมาณค่าสำหรับ** Binary Choice Model

เมื่อ Y_i ในสมการ $Y = f(X's)$ เป็นตัวแปรดัมมี่ซึ่งมีค่าได้ 2 ค่า (Binary Choice หรือ *Dichotomous*) คือ 0, 1 (หรือระหัสอื่น ๆ เช่น 1, 2 หรือ -1, 1 ตามที่ผู้วิจัยเห็นสมควร โดยที่ X's อาจเป็น Quantal หรือมิใช่ก็ได้[2] เรามีวิธีประมาณค่าได้หลายวิธีดังนี้

9.3.1 Linear Probability Model

วิธีนี้เสนอโดยโกลด์เบอร์เกอร์, เซลล์เนอร์ และ ลี (Goldberger, A.S., 1964: Zellner, A and Lee, T.H., 1965)[3] วิธีปฏิบัติคือจากสมการ $Y = X\beta + U$

ให้กำหนด Setting ของตัวแปรอิสระ $X_2, X_3, \ldots, X_k$ ออกเป็นส่วน ๆ รวม M ส่วน ๆ ละ n_i หน่วย ให้ $Y_i = \sum_{j}^{n_i} Y_{ij}$ เมื่อ $Y_{ij} = 0,1$; $i = 1,2,\ldots,M$; $j = 1,2,\ldots,n_i$ และ

$p_i = \frac{Y_i}{n_i}$; $i = 1,2,\ldots,M$ คือความน่าจะเป็นที่ $Y_i = 1$ ใน Setting ที่ i ; $i = 1,2,\ldots,M$ หรือ $p_i = \frac{Y_i}{n_i}$ ก็คือ Sample Proportion ใน Setting ที่ i ของ X's

---

[1] การกำหนด Setting ของตัวแปรอิสระก็คือการจัดกลุ่ม (Combination) ของระหัสของตัวแปรตามตัวอย่างนี้ $X_2 = 0,1$ ขณะที่ $X_3 = 0,1$ เราจึงสามารถจัดกลุ่มค่าสังเกตของ X_2 และ X_3 ได้ 2^2 กลุ่ม คือ (0, 0), (0, 1) (1, 0), (1, 1) ด้วยเหตุนี้ถ้ามีตัวแปรอิสระที่เป็นตัวแปรดัมมี่ทั้งสิ้น m ตัว เราสามารถจัดกลุ่มค่าสังเกตของตัวแปรอิสระได้ทั้งสิ้น 2^m กลุ่ม

[2] คำว่า Quantal มีความหมายเดียวกับคำว่า Discrete ในทางปฏิบัติเราใช้ในความหมายของ Dummy Variable

[3] Judge, G.G., et. al., **Op. Cit.**, p.587

ดังนั้นสมการถดถอยที่ต้องการคือ $p_i = X_i'\beta + u_i$ เมื่อ X_i' คือ Row ที่ i ของแมตริกซ์ X กล่าวคือ $p_i = \beta_1+\beta_2 X_{2i}+\ldots+\beta_k X_{ki}+ u_i$; $i = 1,2,\ldots,M$ หรือเสนอในรูปแมตริกซ์ได้เป็น $p = X\beta + u$ โดยที่ $p = [p_1,p_2,\ldots,p_M]'$ คือเวคเตอร์ขนาด $M\times 1$, X คือแมตริกซ์ขนาด $M\times k$ และ $\beta = [\beta_1,\beta_2,\ldots,\beta_k]'$

สำหรับการตรวจสอบคุณสมบัติของ u_i ในสมการ $p_i = X_i'\beta + u_i$ ให้พิจารณาดังนี้

ให้ true proportion P_i สัมพันธ์กับ Sample Proportion p_i ในลักษณะต่อไปนี้คือ

$$p_i = P_i + u_i \quad ; \quad i = 1,2,\ldots,M$$ [1]

จะพบว่า $u_i = p_i - P_i$; $i = 1,2,\ldots,M$ ซึ่งโดยอาศัยทฤษฎีความน่าจะเป็นเราทราบว่าตัวแปรสุ่ม $Y_i = 0,1$ มีการแจกแจงแบบเบอร์นุลลิ ดังนั้น u_i จึงมีการแจกแจงแบบทวินามที่มีค่าเฉลี่ยเท่ากับ 0 และมีความแปรปรวนเท่ากับ $\frac{P_iQ_i}{n_i}$ กล่าวคือ [2]

$$E(u_i) = E(p_i) - P_i = P_i - P_i = 0$$

$$V(u_i) = V(p_i) = V\left(\frac{y_i}{n_i}\right) = \frac{n_iP_iQ_i}{n_i^2} = \frac{\sigma_i^2}{n_i^2} = \frac{P_iQ_i}{n_i} \quad ; \ i = 1,2,\ldots,M$$

ดังนั้น

$$V(U) = \begin{vmatrix} \frac{P_1Q_1}{n_1} & & & & \\ & \frac{P_2Q_2}{n_2} & & & \\ & & \frac{P_3Q_3}{n_3} & & \\ & & & \ddots & \\ & & & & \frac{P_MQ_M}{n_M} \end{vmatrix} = \Phi \neq \sigma^2 I_M$$

---

[1] จากสมการ $p = X\beta + u$ แสดงว่า $X\beta$ ทำหน้าที่เสมือน true proportion ถ้ากำหนดให้ $X\beta = P$ จึงพบว่า $p = P + u$ หรือ $p_i = P_i + u_i$; $i = 1,2,\ldots,M$

[2] ตามทฤษฎีความน่าจะเป็น ถ้า $X_i = 0,1$ แล้ว $\sum_{i}^{n} X_i$ จะมีการแจกแจงแบบทวินาม คือ $b(n, p)$

ดังนั้นเมื่อพิจารณาแมตริกซ์ Φ เราจึงพบว่า $\Phi \neq \sigma^2 I_M$ หรือเกิดปัญหา Heteroscedasticity ทางออกก็คือประมาณค่า β โดยอาศัย GLS เมื่อทราบค่าของ Φ กล่าวคือ

$$\hat{\beta} = (X'\Phi^{-1}X)^{-1}(X'\Phi^{-1}p)$$

และในกรณีที่ไม่ทราบค่าของ Φ หรือนัยหนึ่งก็คือไม่ทราบค่าของ P_i ให้ดำเนินการประมาณค่าของ P_i เสียก่อนแล้วจึงประมาณค่าเวคเตอร์ β โดยวิธี EGLS ตามขั้นตอนต่อไปนี้

ขั้นที่ 1 จากสมการ $p = X\beta + U$ โดยที่ $p = [p_1, p_2, \ldots, p_M]'$ เมื่อ $p_i = \frac{y_i}{n_i}$ $i = 1,2,\ldots,M$ ให้ประมาณค่าเวคเตอร์ β ตามวิธี OLS ได้

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

ขั้นที่ 2 แทนที่ β ด้วย $\hat{\beta}$ จะได้สมการประมาณค่า $p = X\hat{\beta}$ หรือ $\hat{p} = X\hat{\beta}$ สมาชิกของเวคเตอร์ $\hat{p}$ คือ $\hat{p}_i$; $i = 1,2,\ldots,M$

ขั้นที่ 3 ประมาณค่า $V(u_i)$ โดยใช้ $\hat{p}_i$ แทน P_i ได้ $V(u_i) = \frac{\hat{p}_i\hat{q}_i}{n_i}$; $i = 1,2,\ldots,M$ เมื่อ $\hat{q}_i = 1 - \hat{p}_i$ และพบว่า

$$\hat{\Phi} = \begin{bmatrix} \frac{\hat{p}_1\hat{q}_1}{n_1} & & & \\ & \frac{\hat{p}_2\hat{q}_2}{n_2} & & \\ & & \ddots & \\ & & & \frac{\hat{p}_M\hat{q}_M}{n_M} \end{bmatrix} \qquad \hat{\Phi}^{-1} = \begin{bmatrix} \frac{n_1}{\hat{p}_1\hat{q}_1} & & & \\ & \frac{n_2}{\hat{p}_2\hat{q}_2} & & \\ & & \ddots & \\ & & & \frac{n_M}{\hat{p}_M\hat{q}_M} \end{bmatrix}$$

ขั้นที่ 4 ใช้ $\hat{\Phi}$ ในการประมาณค่าเวคเตอร์ β ตามวิธี EGLS กล่าวคือ

$$\hat{\hat{\beta}} = (X'\hat{\Phi}^{-1}X)^{-1}(X'\hat{\Phi}^{-1}p), \quad \hat{V}(\hat{\hat{\beta}}) = (X'\hat{\Phi}^{-1}X)^{-1}, \quad p = X\hat{\hat{\beta}}$$

วิธีนี้เป็นวิธีที่ง่ายที่สุดแต่มีปัญหาซึ่งนับเป็นจุดอ่อน 2 ประการคือ

1. เราไม่อาจรับประกันได้ว่าค่า $\hat{p}_i$ จากขั้นที่ 2 จะมีค่าปรากฏในช่วง $[0,1]$ เพราะอาจเป็นไปได้ที่ $\hat{p}_i > 1$ หรือ $\hat{p}_i < 0$

2. เราไม่อาจรับประกันได้ว่าค่าพยากรณ์ของ p_i จากสมการประมาณค่า $p = X\hat{\beta}$ จะมีค่าปรากฏในช่วง [0,1]

วิธีแก้ปัญหาอันเป็นจุดอ่อนก็คือ ให้เปลี่ยนไปใช้เทคนิคการประมาณค่าแบบอื่น ขอให้สังเกตว่าเมื่อ Y เป็นตัวแปรดัมมี่เราจะไม่พยากรณ์ค่าของ Y แต่จะพยากรณ์สัดส่วนของ Y ในแต่ละ Setting ของ X's

9.3.2 Transformation Approach

ทางออกสำหรับปัญหาค่า $\hat{p}$ ไม่ตกอยู่ใน Unit Internal [0,1] นั้นเราสามารถแก้ได้ 3 วิธีคือ วิธี Inequality - Restricted Least Square (IRLS) วิธี Normit analysis (หรือ Probit Analysis) และวิธี Logit Analysis

วิธี IRLS ก็คือวิธีกำหนดเงื่อนไขว่า $P_i \leqslant 1$ และ $P_i \geqslant 0$ ซึ่งก็คือ $X_i'\beta \leqslant 1$ และ $X_i'\beta \geqslant 0$ เมื่อ X_i' คือแถวที่ i ของแมตริกซ์ X ในสมการ $p = X\beta + U$ นั่นเอง ค่าประมาณของ β ตามวิธี IRLS ก็คือค่าประมาณที่เกิดขึ้นจาก

minimize $e'e$ ภายใต้ข้อจำกัดว่า

(1) $X\beta \leqslant 1_M$ เมื่อ $1_M = [1,1,\ldots,1]'$

(2) $X\beta \geqslant 0_M$ เมื่อ $0_M = [0,0,\ldots,0]'$

แต่ค่าประมาณของ β ตามวิธี IRLS ยังมีปัญหาคือ เรายังไม่ได้พัฒนา Sampling Property ของ $\hat{\beta}$ ไว้อย่างกว้างขวางเพียงพอแก่งานวิเคราะห์[1]

วิธีประมาณค่า β อีกวิธีหนึ่งที่สามารถแก้ปัญหาเรื่องค่าประมาณ $\hat{p}$ ไม่ปรากฏใน Unit Interval [0,1] ก็คือ Normit Analysis และ Logit Analysis

[1] อ่าน Judge, G.G., et al., Ibid., p.78-94, 588-591

หลักการโดยสรุปของทั้งวิธี Normit Analysis และ Logit Analysis ก็คือการแปลงค่า P_i และ p_i ให้เป็น Quantile[1] ของ Distribution ใดๆ ที่พิจารณาเห็นว่าเหมาะสมแล้วนำค่า Quantile ดังกล่าวไปใช้วิเคราะห์สมการถดถอย จากนั้นจึงแปลงค่า Quantile กลับสู่รูปความน่าจะเป็น P_i และ p_i ซึ่งมีผลให้ค่าประมาณและค่าพยากรณ์ของ P_i และ p_i ปรากฏอยู่ใน Unit Interval

9.3.2.1 Normit Analysis[2]

จากสมการความสัมพันธ์ $p_i = P_i + u_i \; ; \; i = 1,2,\ldots,M$ ....(1)

ให้ Normit ของ P_i คือ

Normit $P_i = I_i$ เมื่อ $I_i = X_i'\beta = \beta_1+\beta_2 X_{2i}+\ldots+\beta_k X_{ki}$ โดยถือว่า P_i มีความสัมพันธ์กับ I_i ดังนี้

$$P_i = F(I_i) = \int_{-\infty}^{I_i} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}t^2)dt^{3}$$

หรือ $I_i = F^{-1}(P_i)$ ....(2)

จาก (1) คือ $p_i = P_i + u_i \; ; \; i = 1,2,\ldots,M$ จะพบว่า

$$F^{-1}(p_i) = F^{-1}(P_i + u_i) \qquad \ldots(3)$$

โดยอาศัย Taylar's Series ของ f(x) ที่กระจายรอบ a คือ

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)}{2!} f''(a) + R$$

ดังนั้นเมื่อกระจาย $F^{-1}(P_i + u_i)$ รอบ P_i จะพบว่า

---

[1] ถ้า $P_i = \int_{-\infty}^{I_i} f_x(x)dx = Pr(X < I_i) = F(I_i)$ เราเรียก I_i ว่าเป็น Quantile of Order P_i แสดงว่าการแปลงรูป P_i เป็น I_i เป็นกระบวนการที่อาศัย CDF ของตัวแปรสุ่มเป็นตัวกลาง และเนื่องจาก CDF คือฟังก์ชั่นของตัวแปรสุ่มที่มี Domain (ในที่นี้คือค่าของ I_i) ปรากฏอยู่ใน Real Line ขณะที่ Counter Domain (ในที่นี้คือ P_i หรือ p_i) มีค่าปรากฏอยู่ช่วง [0,1] ดังนั้นปัญหา P_i และ p_i ไม่ปรากฏค่าใน Unit Interval จึงหมดไป

[2] Zellner, A. and Lee, T.H., "Joint Estimation of Relationships Involving Discrete Random Variables", (Econometrica, 33, April, 1965) p. 382-394

[3] Probit $P_i = I_i + 5$ และขอให้สังเกตว่า I_i คือ Quantile of Order P_i ของ Normal CDF

$$F^{-1}(P_i+u_i) = F^{-1}(P_i)+u_i \frac{d}{dP_i} F^{-1}(P_i)+R_i \qquad \ldots.(4)$$

นั่นคือ $F^{-1}(p_i) = F^{-1}(P_i) + u_i \frac{d}{dP_i} F^{-1}(P_i)+ R_i$

$$I_i^0 = I_i+u_i \frac{d}{dP_i} F^{-1}(P_i)+ \quad R_i \qquad \ldots.(5)$$

โดยที่ $I_i^0 = F^{-1}(p_i)$ เรียกว่า Observed Normit และ $I_i = F^{-1}(P_i)$ เรียกว่า True Normit แต่ R_i (R_i คือRemainderของTayler'sSeries)มีค่าต่ำและสามารถตัดทิ้งไปได้ ขณะเดียวกันเราสามารถพิสูจน์ได้ว่า $u_i \frac{d}{dP_i} F^{-1}(P_i) = \frac{u_i}{Z(P_i)}$ เมื่อ $Z(P_i)$ คือค่าของ $f_Z(P_i)$ กล่าวคือ

$$Z(P_i) = \frac{1}{\sqrt{2\P}} \text{-PI-} \frac{1}{2} P_i^2\}$$

ดังนั้นสมการที่ 5 จึงเปลี่ยนรูปเป็น

$$I_i^0 = I_i \ t \ \frac{u_i}{Z(P_i)} \quad ; i = 1,2,\ldots,M$$

หรือสมการถดถอยที่แปลงรูปแล้วก็คือ

$$I_i^0 = \beta_1+\beta_2X_{2i}+\beta_3X_{3i}+\ldots+\beta_kX_{ki}+ \frac{u_i}{Z(P_i)} \ ; i = 1,2,..,M \qquad \ldots.(6)$$

พิจารณาสมการที่ 6 จะพบว่า

$$E\left(\frac{u_i}{Z(P_i)}\right) = \frac{E(u_i)}{Z(P_i)} = 0 \quad , i = 1,2,\ldots,M$$

$$V\left(\frac{u_i}{Z(P_i)}\right)= \frac{1}{[Z(P_i)]^2}V(u_i) = \frac{1}{[Z(P_i)]^2}\frac{P_iQ_i}{n_i} \quad ; i = 1,2,\ldots,M$$

$$V(U) = \phi = \begin{bmatrix} P_1Q_1/n_1[Z(P_1)]^2 & & & \\ & P_2Q_2/n_2[Z(P_2)]^2 & & \\ & & \ddots & \\ & & & P_MQ_M/n_M[Z(P_M)]^2 \end{bmatrix}_{M\times M}$$

$$= \sigma^2 I_M$$

และเนื่องจาก p_i เป็นค่าประมาณของ P_i ดังนั้น

$$\hat{\phi} = \begin{bmatrix} p_1q_1/n_1[Z(p_1)]^2 & & & \\ & p_2q_2/n_2[Z(p_2)]^2 & & \\ & & \ddots & \\ & & & p_Mq_M/n_M[Z(p_M)] \end{bmatrix}_{M\times M}$$

ดังนั้นโดยอาศัย EGLS เราสามารถประมาณค่า β ได้ดังนี้

$$\hat{\beta} = (X'\hat{\phi}^{-1}X)^{-1}X'\phi^{-1}I^0 \quad , \hat{V}(\hat{\beta}) \cong (X'\hat{\phi}^{-1}X)^{-1}$$ [1]

โดยที่ $I^0 = [F^{-1}(p_1), F^{-1}(p_2), \ldots, F^{-1}(p_M)]'$ คือเวคเตอร์ของ Observed Normit และ $\hat{\beta} = [\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_M]'$

9.3.2.2 Logit Analysis

Logit Analysis เป็นเทคนิคการแปลงค่า p_i สู่ดัชนี I_i โดยผ่าน Logistic Distribution คือ กำหนดให้

---

[1] อ่าน Zellner, A. and Lee, T.H., Ibid.. p.384

$$P_i = \frac{1}{1+\exp[-I_i]} = \frac{1}{1+\exp[-X_i'\beta]} = \frac{1}{1+\exp[-(\beta_1+\beta_2X_{2i}+\ldots+\beta_kX_{ki})]}; \quad i = 1,2,\ldots,M \quad \ldots(1)$$

การพัฒนาวิธีประมาณค่า β ดำเนินการดังนี้

จากสมการ $p_i = P_i + u_i$ และ $1-p_i = 1-(P_i+u_i)$ จะพบว่า

$$\frac{p_i}{1-p_i}[1] = \frac{P_i + u_i}{(1-P_i)-u_i}$$

$$= \frac{P_i(1+u_i/P_i)}{1-P_i(1-u_i/Q_i)} \text{ เมื่อ } Q_i = 1 - P_i$$

ดังนั้น $\ln(\frac{p_i}{1-p_i}) = \ln(\frac{P_i}{1-P_i}) + \ln(1+\frac{u_i}{P_i}) - \ln(1-\frac{u_i}{Q_i});\ i = 1,2,\ldots,M$

---

[1] p_i คือ Probability ที่จะได้ Setting ที่ i ของ X's และ $\frac{p_i}{1-p_i}$ เรียกว่า Observed Odd ของ Setting ที่ i ของ X's หรือ Observed Logit ขณะที่ $\frac{P_i}{1-P_i}$ เรียกว่า True Odd หรือ True Logit

สำหรับ Logistic Distribution มี CDF และคุณลักษณะต่าง ๆ ดังนี้

1. $F(x) = [1+e^{-(x-\alpha)/\beta}]^{-1}$; $-\infty<\alpha<\infty$, $\beta > 0$ ขอให้สังเกต Parameter Space ของ α และ β จะเห็นว่ามีลักษณะเดียวกับ μ และ σ^2 ของ $N(\mu,\sigma^2)$
2. $E(X) = \alpha$
3. $V(X) = \frac{\beta^2\pi^2}{3}$
4. $M_x(t) = e^{\alpha t}\pi\beta t \text{ Cosec }(\pi\beta t)$

ถ้าให้ $\alpha = 0, \beta = 1$ จะพบว่า

$F(x) = [1+e^{-x}]^{-1}$, $E(X) = 0$, $V(X) = 3.29$, $M_x(t) = \pi t \text{ Cosec}(\pi t)$

ขอให้สังเกตว่าในกรณีนี้ Logistic Distribution จะมีธรรมชาติใกล้เคียง N(0,1) มาก

โดยอาศัยTaylar's Series กระจายรอบ a = 0 เราพบว่า

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots$$

$$\ln(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} + \cdots$$

$$\ln\left(\frac{p_i}{1-p_i}\right) = \ln\left(\frac{P_i}{1-P_i}\right) + \frac{u_i}{P_i} + R_{1i} - \left(-\frac{u_i}{Q_i} + R_{2i}\right);\ i = 1,2,\ldots,M$$

เมื่อ R_{1i} และ R_{2i} คือ Remainder ของ $\ln(1+\frac{u_i}{P_i})$ และ $\ln(1-\frac{u_i}{Q_i})$ ตามลำดับ

นั่นคือ

$$\ln\left(\frac{p_i}{1-p_i}\right) = \ln\left(\frac{P_i}{1-P_i}\right) + \frac{u_i}{P_i} + \frac{u_i}{Q_i} + R_i\ ;\ i = 1,2,\ldots,M$$

$$= \ln\left(\frac{P_i}{1-P_i}\right) + \frac{u_i(P_i+Q_i)}{P_iQ_i} + R_i\ ;\ i = 1,2,\ldots,M$$

นั่นคือ $\ln\left(\frac{p_i}{1-p_i}\right) = \ln\left(\frac{P_i}{1-P_i}\right) + \frac{u_i}{P_iQ_i} + R_i\ ;\ i = 1,2,\ldots,M \quad \ldots.(2)$

จากสมการที่ 1 เราทราบว่า

$$P_i = \frac{1}{1+\exp[-(\beta_1+\beta_2X_{2i}+\ldots+\beta_kX_{ki})]}$$

และ

$$1 - P_i = \frac{\exp[-(\beta_1+\beta_2X_{2i}+\ldots+\beta_kX_{ki})]}{1+\exp[-(\beta_1+\beta_2X_{2i}+\ldots+\beta_kX_{ki})]}$$

ดังนั้น $\frac{P_i}{1-P_i} = \frac{1}{\exp[-(\beta_1+\beta_2X_{2i}+\ldots+\beta_kX_{ki})]}\ ;\ i = 1,2,\ldots,M$

และ $\ln\left(\frac{P_i}{1-P_i}\right) = \beta_1+\beta_2X_{2i}+\ldots+\beta_kX_{ki} \quad \ldots.(3)$

แทนที่สมการ 3 ในสมการที่ 2 จะพบว่าสมการถดถอยที่แปลงรูปแล้วคือ

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_1+\beta_2X_{2i}+\ldots+\beta_kX_{ki} + \frac{u_i}{P_iQ_i} + R_i\ ;\ i = 1,2,\ldots,M \quad \ldots(4)$$

พิจารณาสมการที่ 4 จะพบว่า

$$E\left(\frac{u_i}{P_iQ_i}\right) = 0 \text{ และ } V\left(\frac{u_i}{P_iQ_i}\right) = \frac{1}{P_i^2Q_i^2} \cdot \frac{P_iQ_i}{n_i} = \frac{1}{n_iP_iQ_i} \text{ ดังนั้น}$$

$$V(U) = \phi = \begin{bmatrix} 1/n_1P_1Q_1 & & & \\ & 1/n_2P_2Q_2 & & \\ & & \ddots & \\ & & & 1/n_MP_MQ_M \end{bmatrix}_{M\times M}$$

$$\text{และ} \quad \hat{V}(U) = \hat{\phi} = \begin{bmatrix} 1/n_1p_1q_1 & & & \\ & 1/n_2p_2q_2 & & \\ & & \ddots & \\ & & & 1/n_Mp_Mq_M \end{bmatrix}_{M\times M}$$

ดังนั้นโดยอาศัย EGLS เราสามารถประมาณค่าเวคเตอร์ β ได้ดังนี้คือ

$$\hat{\beta} = (X'\hat{\Phi}^{-1}X)^{-1}X'\hat{\Phi}I^0 \quad , \quad \hat{V}(\hat{\beta}) \cong (X'\hat{\Phi}^{-1}X)^{-1}$$

โดยที่ I^0 คือเวคเตอร์ของ Observed Logit กล่าวคือ $I^0 = [\ln\frac{p_i}{1-p_i}, \ln\frac{p_2}{1-p_2}, \ldots, \ln\frac{p_M}{1-p_M}]'$

สมการประมาณค่าของทั้ง Normit Model และ Logit Model คือ

$$I_i = \hat{\beta}_1+\hat{\beta}_2X_{2i}+\ldots+\hat{\beta}_kX_{ki}$$

แต่สิ่งที่เราต้องการพยากรณ์คือ p_i มิใช่ I_i ดังนั้น เมื่อแทนค่าในอนาคตของ X's ลงในสมการประมาณค่าซึ่งจะทำให้ได้ค่า $\hat{I}_i$ ให้แปลงค่า $\hat{I}_i$ สู่ $\hat{p}_i$ ดังนี้

$$\hat{p}_i = F(I_i)$$

นั่นคือ

$$\hat{p}_i = \begin{cases} F(\hat{I}_i) = Pr(Z < \hat{I}_i) & \text{สำหรับ Normit Model} \\ F(\hat{I}_i) = 1/[1+e^{-\hat{I}_i}] & \text{สำหรับ Logit Model} \end{cases}$$

จากนั้นให้ย้อนค่า $\hat{p}_i$ สู่ $\hat{Y}_i$ ดังนี้[1]

$$\text{Normit Model} : \hat{Y}_i = \begin{cases} 1 \text{ ถ้า } \hat{p}_i = Pr(Z \leq \hat{I}_i) \\ 0 \text{ ถ้า } p_i = \hat{P}r(Z > \hat{I}_i) \end{cases}$$

$$\text{Logit Model} : \hat{Y}_i = \begin{cases} 1 \text{ ถ้า } \hat{p}_i = 1/[1+e^{-\hat{I}_i}] \\ 0 \text{ ถ้า } \hat{p}_i = 1- 1/[1+e^{-\hat{I}_i}] \end{cases}$$

---

[1] Zellner, A and Lee, T.H., Ibid., p.384