บทที่ 3

วิธีดำเนินการวิจัย

ในบทนี้จะแบ่งออกเป็น 2 ส่วนคือ

ส่วนที่ 1 อธิบายวิธีการจำลองข้อมูลจากตัวแบบแฝงของการถดถอยโลจิสติก ทวิภาคซึ่งแบ่งเป็น

- วิธีการจำลองข้อมูลภายใต้สัดส่วนผลตอบสนอง จำนวนตัวแปรอธิบาย ขนาดตัวอย่าง และสัมประสิทธิ์การตัดสินใจที่แท้จริง
- วิธีการจำลองข้อมูลภายใต้ระดับความเชื่อถือได้ในตัวแปรอธิบาย จำนวนตัวแปรอธิบาย และขนาดตัวอย่าง
- วิธีการจำลองข้อมูลภายใต้อัตราการจำแนกผิดในตัวแปรตาม จำนวนตัวแปรอธิบาย และขนาดตัวอย่าง

ส่วนที่ 2 อธิบายขั้นตอนในการดำเนินการวิจัย

3.1 วิธีการจำลองข้อมูลจากตัวแบบตัวแปรแฝงของการถดถอยโลจิสติกทวิภาค

ตัวแปรแฝง (Latent Variable, Y^*) เป็นตัวแปรตามชนิดต่อเนื่อง ดังนั้นจึงต้องทำการแปลงให้ Y^* เป็นตัวแปรตาม Y ที่มีค่าเป็นไปได้ 2 ค่า โดยพิจารณาในรูปของ “เหตุการณ์ที่สนใจ” และ “เหตุการณ์ที่ไม่สนใจ” โดยที่ $Y = 1$ เมื่อเกิดเหตุการณ์ที่สนใจ และ $Y = 0$ เมื่อเกิดเหตุการณ์ที่ไม่สนใจ (Sharma, 2006) โดยใช้สมการ

$$Y = \begin{cases} 1 & \text{ถ้า } Y^* \geq c \;\; ; \;\; c \in R \\ 0 & \text{ในกรณีอื่นๆ} \end{cases} \tag{3.1}$$

เมื่อ c คือค่าคงที่ใดๆ ที่ใช้ในการกำหนดค่าให้ Y เป็น 1 หรือ 0 ซึ่งความน่าจะเป็นที่ Y จะเกิดเหตุการณ์ที่สนใจ แสดงได้รูปของสมการคือ

$$\pi = P(Y = 1) = P(y^* \geq c)$$

ถ้า Y^* มีความสัมพันธ์กับตัวแปรอิสระ ที่สามารถแสดงสมการของความสัมพันธ์ได้ดังนี้

$$Y^* = a + \sum_{j=1}^{k} b_j x_{ij} + \varepsilon_i \qquad ; \quad i = 1,2,...,n \tag{3.2}$$

เมื่อ a คือค่าคงที่ใดๆ และ ε คือค่าความคลาดเคลื่อนภายใต้ข้อสมมติที่มีการแจกแจงแบบสมมาตร ดังนั้นจากตัวแบบตัวแปรแฝงสามารถหาค่าความน่าจะเป็นที่จะเกิดเหตุการณ์ที่สนใจได้จากสมการ

$$\begin{aligned} \pi = P(Y=1) &= P(y^* \geq c) = P(\lambda + \varepsilon \geq c) \\ &= P(\varepsilon \geq c - \lambda) \\ &= 1 - F(c - \lambda) \\ &= F(\lambda - c) \end{aligned}$$

เมื่อ $\lambda = a + \sum_{j=1}^{k} b_j x_{ij}$ และ F เป็นฟังก์ชันการแจกแจงสะสมของ ε โดยทั่วไปจะกำหนดให้มีการแจกแจงแบบปรกติ การวัดระดับความสัมพันธ์เชิงเส้นของตัวแบบ (3.2) สามารถใช้ค่า R^2_{true} เป็นตัววัด ซึ่งในการวิจัยนี้จะใช้ค่า R^2_{est} ที่ได้จากการสร้างสมการถดถอยระหว่างค่า Y^* ซึ่งเป็นตัวแปรที่แปลงมาจาก Y โดยใช้สมการ (3.1) กับตัวแปรอธิบายโดยการใช้ตัวแบบการถดถอยโลจิสติกทวิภาค เป็นตัววัดระดับความสัมพันธ์เชิงเส้นของตัวแบบ

จากตัวแบบตัวแปรแฝง (3.2) ค่า R^2_{true} สามารถคำนวณได้โดยการแบ่งความแปรปรวนของ Y^* ออกเป็น 2 ส่วนคือ ส่วนความแปรปรวนที่เกิดมาจากตัวแบบ และส่วนความแปรปรวนเฉลี่ยที่เนื่องมาจากรูปแบบของตัวแปรอธิบายทั้งหมด ดังแสดงได้ในสมการต่อไปนี้

$$Var(Y^*) = Var[E(Y^* \mid \underline{x}_i)] + E[Var(Y^* \mid \underline{x}_i)] \tag{3.3}$$

$$1 = \frac{Var[E(Y^* \mid \underline{x}_i)]}{Var(Y^*)} + \frac{E[Var(Y^* \mid \underline{x}_i)]}{Var(Y^*)} \tag{3.4}$$

เมื่อ $$\frac{Var[E(Y^* \mid \underline{x}_i)]}{Var(Y^*)} = R^2_{true}$$

ดังนั้น $$R_{true}^{2} = 1 - \frac{E\left[Var\left(Y^{*} \mid \underline{x}_{i}\right)\right]}{Var\left(Y^{*}\right)} \qquad (3.5)$$

จากข้อสมมติของการประมาณค่าพารามิเตอร์ด้วยวิธีกำลังสองน้อยสุดแบบสามัญ (Ordinary Least Squares Method, OLS) ก็คือค่าความแปรปรวนของความคลาดเคลื่อน นั่นคือ

$$E\left[Var\left(Y^{*} \mid \underline{x}_{i}\right)\right] = \sigma^{2} \qquad (3.6)$$

ดังนั้นค่า R_{true}^{2} สามารถใช้เป็นเกณฑ์ในการเปรียบเทียบค่า R_{est}^{2} ของการถดถอยโลจิสติกทวิภาคภายใต้ระดับต่างๆ ของสัดส่วนผลตอบสนองและถ้ากำหนดให้ $a = 0$ แล้วตัวแบบ (3.2) สามารถเขียนได้เป็น

$$Var\left(Y^{*}\right) = \sum_{j=1}^{k} b_{j}^{2}\sigma_{x_{j}}^{2} + \sigma_{\varepsilon}^{2} \qquad (3.7)$$

เมื่อ $\sigma_{x_j}^{2}$ คือความแปรปรวนของตัวแปรอธิบายตัวที่ j ; $j = 1,2,\ldots,p$ และ σ_{ε}^{2} คือความแปรปรวนของความคลาดเคลื่อน

แทนค่าสมการ (3.7) ในสมการ (3.5) จะได้ว่า

$$R_{true}^{2} = 1 - \frac{\sigma_{\varepsilon}^{2}}{\sum_{j=1}^{k} b_{j}^{2}\sigma_{x_{j}}^{2} + \sigma_{\varepsilon}^{2}} \qquad (3.8)$$

นั่นคือ $$\sum_{j=1}^{k} b_{j}^{2}\sigma_{x_{j}}^{2} = \frac{\sigma_{\varepsilon}^{2}R^{2}}{1 - R^{2}} \qquad (3.9)$$

จะได้ว่า $$b_{j} = \frac{1}{k}\left(\frac{\sigma_{\varepsilon}}{\sigma_{x_{j}}}\sqrt{\frac{R_{true}^{2}}{1 - R_{true}^{2}}}\right) \quad \text{เมื่อ} \quad j = 1,2,\ldots,p \qquad (3.10)$$

แทนค่าสมการ (3.9) ในสมการ (3.7) จะได้ว่า

$$Var\left(Y^{*}\right) = \frac{\sigma_{\varepsilon}^{2}R_{true}^{2}}{1 - R_{true}^{2}} + \sigma_{\varepsilon}^{2}$$

$$= \frac{\sigma_{\varepsilon}^{2}}{1 - R_{true}^{2}}$$

$$\equiv \sigma_{Y^*}^2 \tag{3.11}$$

ดังนั้น $$\sigma_\varepsilon^2 = \sigma_{Y^*}^2\left(1 - R_{true}^2\right) \tag{3.12}$$

และจาก $$E\left(Y^*\right) = \sum_{j=1}^{k} b_j \mu_{x_j}$$

$$= \frac{1}{k}\sum_{j=1}^{k}\left(\frac{\sigma_\varepsilon}{\sigma_{x_j}}\sqrt{\frac{R_{true}^2}{1-R_{true}^2}}\right)\cdot \mu_{x_j}$$

$$= \frac{1}{k}\sum_{j=1}^{k}\left(\frac{\mu_{x_j}}{\sigma_{x_j}}\sqrt{\sigma_{Y^*}^2 R_{true}^2}\right)$$

$$\equiv \quad \eta \tag{3.13}$$

เมื่อ μ_{x_j} และ σ_{x_j} คือ ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของตัวแปรอธิบายตัวที่ j เมื่อ $j = 1,2,..., p$

จากสมการ (3.1) สามารถเขียนความน่าจะเป็นที่ Y จะเกิดเหตุการณ์ที่สนใจ ในรูปของสมการดังนี้

$$\pi = P(Y=1) = P\left(y^* \geq c\right)$$

$$= P\left(\frac{Y^* - \eta}{\sigma_{Y^*}} > \frac{c-\eta}{\sigma_{Y^*}}\right)$$

$$= 1 - \Phi\left(\frac{c-\eta}{\sigma_{Y^*}}\right) \tag{3.14}$$

จากสมการจะเห็นได้ว่าสัดส่วนผลตอบสนอง (π) เป็นฟังก์ชันของค่าจุดตัด (c) ค่าเฉลี่ย (η) และส่วนเบี่ยงเบนมาตรฐาน (σ_{Y^*}) ของตัวแปรแฝง นั่นคือ ความแตกต่างของสัดส่วนผลตอบสนอง (π) ของประชากรที่แตกต่างอาจขึ้นอยู่กับการใช้จุดตัดของ Y^* ที่แตกต่างกันหรืออาจเนื่องมาจาก Y^* มีค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานแตกต่างกันในแต่ละกลุ่ม

3.1.1 วิธีการจำลองข้อมูลภายใต้สัดส่วนผลตอบสนอง จำนวนตัวแปรอธิบาย ขนาดตัวอย่าง และค่าสัมประสิทธิ์การตัดสินใจที่แท้จริง

ในที่นี้จะกำหนดให้ค่า $\sigma^2_{Y^*}$ และ c เป็นค่าคงที่ แต่ให้ค่า η แตกต่างกันในแต่ละชุดของข้อมูล ดังนั้นในการจำลองข้อมูลจะประกอบด้วย 2 ขั้นตอนคือ

ขั้นตอนที่ 1 การกำหนดค่าคงที่ของ $\sigma^2_{Y^*}$ และ c ทำได้โดยการกำหนดให้ค่า $\sigma^2_{Y^*}$ เป็นค่าคงที่ค่าหนึ่ง ๆ คือ V_0 แล้วคำนวณหาค่าคงที่ของ $c = c_0$ ที่ทำให้ได้สัดส่วนผลตอบสนองตามที่ต้องการ โดยมีรายละเอียดดังต่อไปนี้

(1.1) ในที่นี้กำหนดให้ค่า $V_0 = 10$

(1.2) คำนวณค่า $c = c_0$ โดย

(1.2.1) กำหนดค่าเริ่มต้นของ η หรือ η_0 จาก

$$\eta_0 = \frac{1}{k}\sum_{j=1}^{p}\left(\frac{\sigma_{\varepsilon(0)}}{\sigma_{x_j}}\sqrt{\frac{R^2_{true(0)}}{1-R^2_{true(0)}}}\right)\cdot\mu_{x_j}$$

โดยที่ $\sigma^2_{\varepsilon(0)} = V_0 \cdot \left(1 - R^2_{true(0)}\right)$

และ μ_{x_j} และ σ_{x_j} คือค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของตัวแปร x_j ; $j = 1,2,\dots,p$ ที่ได้กำหนดในขอบเขตการศึกษา (ข้อ 1 หน้า 7) และ Subscript (0) หมายถึง ค่าต่างๆที่คำนวณก่อนที่จะทำการทดลองข้อมูลตามเงื่อนไขต่างๆ ของการทดลอง

(1.2.2) คำนวณค่า c_0 ที่ทำให้ค่า $\pi = \pi_0$ ตามต้องการ เมื่อกำหนดให้ค่า $E\left(y^*\right) = \eta_0$ จะได้ว่า

$$c_0 = \sqrt{V_0}\,\phi^{-1}\left(1-\pi_0\right) + \eta_0$$

ในที่นี้ จะให้ค่า $R^2_{true(0)} = 0.1$ และ $\pi_0 = 0.10$

<u>ขั้นตอนที่ 2</u> จำลองข้อมูลตามเงื่อนไขของการทดลอง

ในที่นี้จะใช้ Subscript (l) แทนเงื่อนไขของการทดลอง $\left(n, R^2_{true}, \pi\right)$ ในการทดลองครั้งที่ l โดยมีขั้นตอนดังนี้

(2.1) คำนวณ $\sigma^2_{\varepsilon(l)}$ โดยที่

$$\sigma^2_{\varepsilon(l)} = V_0 \cdot \left(1 - R^2_{true(l)}\right)$$

(2.2) คำนวณ $b_{j(l)}$ โดยที่

$$b_{j(l)} = \frac{1}{p}\left(\frac{\sigma_{\varepsilon(l)}}{\sigma_{x_j}}\sqrt{\frac{R^2_{true(l)}}{1-R^2_{true(l)}}}\right)$$

(2.3) จำลองตัวแปรอธิบาย X_j ตามขอบเขตการวิจัยในบทที่ 1 หัวข้อ 1.3 และจำลอง $\varepsilon \sim N(0, \sigma^2_{\varepsilon(l)})$

(2.4) คำนวณค่า $\eta_{(l)}$ โดยมีขั้นตอนดังนี้

(2.4.1) คำนวณค่าของ η จากตัวแบบ ในที่นี้จะแทนด้วย $\eta_{m(l)}$

$$\eta_{m(l)} = \sum_{j=1}^{p} b_j \bar{x}_j + \bar{\varepsilon}$$

เมื่อ $\bar{x}_j$ และ $\bar{\varepsilon}$ เป็นค่าเฉลี่ยของตัวแปร x_j และ ε

(2.4.2) คำนวณค่าของ η ที่จะทำให้ได้ค่าสัดส่วนผลตอบสนองตามที่ต้องการเมื่อกำหนดสัดส่วนผลตอบสนอง $\left(\pi_{(l)}\right)$ ระดับต่างๆ ในที่นี้จะแทนด้วย $\eta_{d(l)}$

$$\eta_{d(l)} = c_0 - \sqrt{V_0}\,\phi^{-1}\left(1 - \pi_{(l)}\right)$$

(2.4.3) ปรับค่าของ η แทนด้วย $\eta_{adj(l)}$

$$\eta_{adj(l)} = \eta_{d(l)} - \eta_{m(l)}$$

(2.5) จำลองค่า Y^* จาก

$$Y^* = \eta_{adj(l)} + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p + \varepsilon$$

(2.6) แปลงค่า Y^* ให้เป็น Y ที่มีค่า 0 หรือ 1 โดยใช้

$$Y = \begin{cases} 1 & \text{ถ้า } Y^* \geq c_0 \\ 0 & \text{ในกรณีอื่นๆ} \end{cases}$$

3.1.2 วิธีการจำลองข้อมูลภายใต้ระดับความเชื่อถือได้ในตัวแปรอธิบาย จำนวนตัวแปรอธิบาย และขนาดตัวอย่าง

ในที่นี้จะศึกษาอิทธิพลของค่าความเชื่อถือได้สำหรับตัวแปรอธิบายเท่านั้น โดยการกำหนดค่าความเชื่อถือได้จะอาศัยวิธีการทางทฤษฎีการทดสอบแบบฉบับ (Classical Test Theory) ที่นิยามค่าความเชื่อถือได้ รูปอัตราส่วนของความแปรปรวนของค่าจริง (True Score Variance, σ_T^2) กับความแปรปรวนของค่าประมาณ (Observed Score Variance, σ_X^2) (Allen & Yen, 1979) ที่แสดงได้ในรูปของสมการคือ

$$\gamma = \sigma_T^2 / \sigma_X^2 \qquad \text{เมื่อ} \quad \sigma_T^2 = \sigma_X^2 - \sigma_\varepsilon^2$$

ดังนั้น

$$\gamma = \frac{\sigma_X^2 - \sigma_\varepsilon^2}{\sigma_X^2} = 1 - \frac{\sigma_\varepsilon^2}{\sigma_X^2}$$

เมื่อ σ_ε^2 คือค่าความแปรปรวนของความคลาดเคลื่อน

ดังนั้นเมื่อกำหนดค่า γ และ σ_ε^2 สามารถหาค่า σ_X^2 ได้จาก

$$\sigma_X^2 = \frac{\sigma_\varepsilon^2}{1-\gamma}$$

ในกรณีที่มีตัวแปรอธิบายจำนวน p ตัวที่เป็นอิสระต่อกันแล้ว จะได้ว่า

$$\sigma_{X_j}^2 = \frac{1}{p}\left(\frac{\sigma_\varepsilon^2}{1-\gamma}\right) \quad ; \quad j = 1,2,\ldots,p \tag{3.15}$$

ดังนั้นวิธีการจำลองข้อมูลจะทำเช่นเดียวกับการจำลองข้อมูลภายใต้สัดส่วนผลตอบสนอง และเงื่อนไขอื่นๆ ของ R_{adj}^2 ดังที่อธิบายในหัวข้อ (3.1.1) ต่างกันเพียงขั้นตอนในการจำลองค่าของตัวแปรอธิบาย x_j โดยที่ค่าความแปรปรวนจะเป็นค่าที่คำนวณได้จากสมการ (3.15) และค่าเฉลี่ยตามที่กำหนดในขอบเขตการศึกษา (ข้อ 1 หน้า 7)

3.1.3 วิธีการจำลองข้อมูลภายใต้อัตราการจำแนกผิดในตัวแปรตาม จำนวนตัวแปรอธิบาย และขนาดตัวอย่าง

การจำลองข้อมูลภายใต้อัตราการจำแนกผิดในตัวแปรตาม จะทำเช่นเดียวกับการจำลองข้อมูลภายใต้สัดส่วนผลตอบสนอง และเงื่อนไขอื่นๆ ของ R_{adj}^2 ดังที่อธิบายในหัวข้อ (3.1.1) จนกระทั่งได้ค่า Y ที่มีค่าเป็น 0 หรือ 1 แล้ว จึงทำการแทนค่า 0 ด้วย 1 และแทนค่า 1 ด้วย 0 ตามอัตราที่กำหนดไว้ในขอบเขตการศึกษา (ข้อ 3.1 หน้า 7)

3.2 ขั้นตอนในการดำเนินการวิจัย

การวิจัยครั้งนี้ ใช้วิธีการจำลองแบบมอนติคาร์โล (Monte Carlo Simulation Method) โดยการจำลองข้อมูลและการประมวลผลข้อมูลทำได้ด้วยการใช้โปรแกรม $SAS^{®}$ เวอร์ชัน 9.0 ขั้นตอนในการวิจัยมีดังนี้

(1) จำลองข้อมูลตามที่อธิบายในหัวข้อ 3.1

(2) นำข้อมูลที่จำลองมาประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีความควรจะเป็นสูงสุดของฟิชเชอร์สกอริง (Fisher's Scoring Method) โดยใช้คำสั่ง PROC LOGISTIC ในโปรแกรม $SAS^{®}$ เวอร์ชัน 9.0

(3) คำนวณค่าสัมประสิทธิ์การตัดสินใจที่ปรับค่า $\left(R_{adj}^2\right)$ ที่นำมาศึกษาได้แก่ $R_{O,adj,MS}^2$, $R_{l,adj,MS}^2$, $R_{O,adj,LM}^2$, $R_{l,adj,LM}^2$ และ $R_{l,adj,SAS_{AIC}}^2$ ด้วยสมการที่ (2.3), (2.4), (2.5), (2.6) และ (2.7) ตามลำดับ

(4) ทำซ้ำในขั้นตอนที่ 2 ถึงขั้นตอนที่ 3 จำนวน $1,000$ ครั้ง

(5) ทดสอบความสัมพันธ์ และปฏิสัมพันธ์ 2 ทางระหว่างปัจจัยและเงื่อนไขอื่นๆ ของ R_{adj}^2 รวมทั้งเปรียบเทียบความเอนเอียงของ R_{adj}^2 แต่ละค่า

(6) วิเคราะห์ผลและรายงานผล

(7) สรุปผลของการวิจัย