

บทที่ 3

วิธีดำเนินการวิจัย

การวิจัยนี้สนใจศึกษาเปรียบเทียบประสิทธิภาพของสัมประสิทธิ์การตัดสินใจ (R_o^2 , R_l^2) และสัมประสิทธิ์การตัดสินใจที่ปรับค่า ($R_{O,adj,MS}^2$, $R_{l,adj,MS}^2$, $R_{O,adj,LM}^2$, $R_{l,adj,LM}^2$, $R_{l,adj,SAS_{DEV}}^2$, $R_{l,adj,SAS_{AIC}}^2$) สำหรับตัวแบบการถดถอยโลจิสติก ภายใต้ขนาดตัวอย่าง (n) เท่ากับ 50, 100, 1000 และ 2000 และจำนวนตัวแปรอธิบายในตัวแบบการถดถอยเท่ากับ 1, 5 และ 10 ตัว โดยที่ตัวแปรอธิบายเป็นทั้งตัวแปรเชิงคุณภาพและเชิงปริมาณ นอกจากนั้นยังสนใจศึกษาความเอนเอียงสัมพัทธ์ (relative bias) และรากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (root mean square error, RMSE) ของสัมประสิทธิ์การตัดสินใจและสัมประสิทธิ์การตัดสินใจที่ปรับค่าที่กล่าวข้างต้น เมื่อกำหนดตัวแบบการถดถอยไม่ถูกต้องใน 3 ลักษณะคือ ลักษณะที่รูปแบบของตัวแปรอธิบายไม่ถูกต้อง (wrong functional form of the covariate) ลักษณะที่กำหนดฟังก์ชันเชื่อมโยงของตัวแบบไม่ถูกต้อง (misspecified link function) และลักษณะที่กำหนดให้ตัวแบบมีตัวแปรอธิบายที่มีความสัมพันธ์กับตัวแปรตามขาดหายไป (missing covariate)

ขั้นตอนในการวิจัย

งานวิจัยนี้เป็นการศึกษาโดยใช้วิธีการจำลองข้อมูลด้วยเทคนิคมอนติคาร์โล โดยการจำลองข้อมูลและการประมวลผลข้อมูลทำได้ด้วยการใช้โปรแกรม SAS[®] เวอร์ชัน 8.0 โดยมีขั้นตอนในการวิจัยดังต่อไปนี้

การจำลองชุดข้อมูล

จำลองชุดข้อมูลที่ใช้ในการศึกษาโดยที่

1. สร้างชุดข้อมูลของตัวแปรอธิบาย X_j เมื่อ $j=1, 2, \dots, 10$ โดยกำหนดให้ X_1 มีการแจกแจงปกติที่มีค่าเฉลี่ยเท่ากับ 2 และมีความแปรปรวนเท่ากับ 4 X_2 มีการแจกแจงเอกรูปไม่ต่อเนื่องที่มีค่าเป็นไปได้อีกคือ 0 และ 1 X_3 มีการแจกแจงปกติที่มีค่าเฉลี่ยเท่ากับ 20 และมีความแปรปรวนเท่ากับ 4

- X_4 มีการแจกแจงเอกกรุปไม่ต่อเนื่องที่มีค่าเป็นไปได้คือ 0, 1, 2 และ 3
 X_5 มีการแจกแจงปกติที่มีค่าเฉลี่ยเท่ากับ 8 และมีความแปรปรวนเท่ากับ 3
 X_6 มีการแจกแจงเอกกรุปไม่ต่อเนื่องที่มีค่าเป็นไปได้คือ 0, 1 และ 2
 X_7 มีการแจกแจงปกติที่มีค่าเฉลี่ยเท่ากับ 0 และมีความแปรปรวนเท่ากับ 1
 X_8 มีการแจกแจงเอกกรุปไม่ต่อเนื่องที่มีค่าเป็นไปได้คือ 0 และ 1
 X_9 มีการแจกแจงปกติที่มีค่าเฉลี่ยเท่ากับ 15 และมีความแปรปรวนเท่ากับ 2
 X_{10} มีการแจกแจงเอกกรุปไม่ต่อเนื่องที่มีค่าเป็นไปได้คือ 0, 1, 2, 3 และ 4

2. สร้างชุดข้อมูลของตัวแปรตามแบบทวิภาคจากตัวแบบ 3 ตัวแบบนี้

$$\text{logit}(p_i) = 0.5 - 0.45x_{i1} \quad (M01)$$

$$\text{logit}(p_i) = 0.5 - 0.45x_{i1} + 3.525x_{i2} - 0.1x_{i3} - 0.5x_{i4} + 0.2x_{i5} \quad (M05)$$

$$\begin{aligned} \text{logit}(p_i) = 0.5 - 0.45x_{i1} + 3.525x_{i2} - 0.1x_{i3} - 0.5x_{i4} + 0.2x_{i5} \\ - 0.4x_{i6} + 0.3x_{i7} - 0.3x_{i8} - 0.05x_{i9} + 0.8x_{i10} \end{aligned} \quad (M010)$$

และทำการแปลงค่า $\text{logit}(p_i)$ ที่ได้จากตัวแบบ (M01), (M05) และ (M010) ให้เป็นค่าตัวแปรสุ่ม Y_i โดยกำหนดให้ตัวแปร U_i มีการแจกแจงเอกกรุปที่ต่อเนื่องบนช่วง (0,1) ($U_i \sim \text{Uniform}(0,1)$) และให้ P_i คือผลต่างระหว่างค่า p_i และ U_i นั่นคือ $P_i = p_i - U_i$ จะได้ค่า P_i ที่มีค่าอยู่ระหว่าง -1 และ 1 และจะทำการปรับค่าของ P_i ให้เท่ากับจำนวนเต็มที่อยู่ใกล้ที่สุด คือถ้า $-1 \leq P_i \leq 0$ ให้ค่า $y = 0$ และ ถ้า $0 < P_i \leq 1$ ให้ค่า $y = 1$ จะทำให้ได้ตัวแปรตาม Y_i ที่เป็นตัวแปรสุ่มซึ่งมีค่า 0 และ 1 ในสัดส่วนที่ใกล้เคียงกัน

ตัวแบบที่ใช้ในการศึกษา

กำหนดตัวแบบไม่ถูกต้องใน 3 ลักษณะ ได้แก่

1. กำหนดตัวแบบเมื่อรูปแบบของตัวแปรอธิบายไม่ถูกต้อง กรณีนี้จะกำหนดให้ตัวแปรอธิบายเป็นตัวแปรเชิงปริมาณ (X_1) มีรูปแบบเป็นกำลังสอง (X_1^2) นั่นคือ

$$\text{logit}(p_i) = b_0 + b_1 x_{i1}^2 \quad (M11)$$

$$\text{logit}(p_i) = b_0 + b_1 x_{i1}^2 + b_2 x_{i2} + b_3 x_{i3} + b_4 x_{i4} + b_5 x_{i5} \quad (M15)$$

$$\begin{aligned} \text{logit}(p_i) = & b_0 + b_1 x_{i1}^2 + b_2 x_{i2} + b_3 x_{i3} + b_4 x_{i4} + b_5 x_{i5} \\ & + b_6 x_{i6} + b_7 x_{i7} + b_8 x_{i8} + b_9 x_{i9} + b_{10} x_{i10} \end{aligned} \quad (M110)$$

2. กำหนดฟังก์ชันเชื่อมโยงของตัวแบบไม่ถูกต้อง กรณีนี้จะกำหนดให้ฟังก์ชันเชื่อมโยงไม่ถูกต้อง 2 รูปแบบนั้นคือ

2.1 ฟังก์ชันคอมพลีเมนต์ลอจิสติก (complementary log-log)

จะได้ตัวแบบดังต่อไปนี้

$$\log[-\log(1-p_i)] = b_0 + b_1 x_{i1} \quad (M21)$$

$$\log[-\log(1-p_i)] = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + b_4 x_{i4} + b_5 x_{i5} \quad (M25)$$

$$\begin{aligned} \log[-\log(1-p_i)] = & b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + b_4 x_{i4} + b_5 x_{i5} \\ & + b_6 x_{i6} + b_7 x_{i7} + b_8 x_{i8} + b_9 x_{i9} + b_{10} x_{i10} \end{aligned} \quad (M210)$$

2.2 ฟังก์ชันโพรบิท (probit) จะได้ตัวแบบดังต่อไปนี้

$$\Phi^{-1}(p_i) = b_0 + b_1 x_{i1} \quad (M31)$$

$$\Phi^{-1}(p_i) = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + b_4 x_{i4} + b_5 x_{i5} \quad (M35)$$

$$\begin{aligned} \Phi^{-1}(p_i) = & b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + b_4 x_{i4} + b_5 x_{i5} \\ & + b_6 x_{i6} + b_7 x_{i7} + b_8 x_{i8} + b_9 x_{i9} + b_{10} x_{i10} \end{aligned} \quad (M310)$$

3. กำหนดให้ตัวแบบการถดถอยที่มีตัวแปรอธิบายที่มีความสัมพันธ์กับตัวแปรตาม
ขาดหายไป นั่นคือ

3.1 สำหรับตัวแบบการถดถอยที่มีตัวแปรอธิบาย 5 ตัว

กำหนดให้ตัวแปรอธิบายที่ขาดหายไปจากตัวแบบเป็นตัวแปรเชิงปริมาณ
(X_5) และตัวแปรเชิงคุณภาพ (X_2) ในตัวแบบ (M45) และ (M55) ตามลำดับ

$$\text{logit}(p_i) = b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + b_4x_{i4} \quad (M45)$$

$$\text{logit}(p_i) = b_0 + b_1x_{i1} + b_3x_{i3} + b_4x_{i4} + b_5x_{i5} \quad (M55)$$

กำหนดให้ตัวแปรอธิบายที่ขาดหายไปจากตัวแบบมีทั้งตัวแปรเชิง
ปริมาณและเชิงคุณภาพ โดยกำหนดให้ตัวแปรอธิบายที่ขาดหายไปคือ X_1, X_2, X_4, X_5

$$\text{logit}(p_i) = b_0 + b_3x_{i3} \quad (M65)$$

3.2 สำหรับตัวแบบการถดถอยที่มีตัวแปรอธิบาย 10 ตัว

กำหนดให้ตัวแปรอธิบายที่ขาดหายไปจากตัวแบบเป็นตัวแปรเชิงปริมาณ
(X_3, X_7) และกำหนดให้ตัวแปรอธิบายที่ขาดหายไปเป็นตัวแปรเชิงคุณภาพ (X_8, X_{10}) ใน
ตัวแบบ (M410) และ (M510) ตามลำดับ

$$\begin{aligned} \text{logit}(p_i) = & b_0 + b_1x_{i1} + b_2x_{i2} + b_4x_{i4} + b_5x_{i5} \\ & + b_6x_{i6} + b_8x_{i8} + b_9x_{i9} + b_{10}x_{i10} \end{aligned} \quad (M410)$$

$$\begin{aligned} \text{logit}(p_i) = & b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + b_4x_{i4} \\ & + b_5x_{i5} + b_6x_{i6} + b_7x_{i7} + b_9x_{i9} \end{aligned} \quad (M510)$$

กำหนดให้ตัวแปรอธิบายที่ขาดหายไปจากตัวแบบมีทั้งตัวแปรเชิง
ปริมาณและเชิงคุณภาพ โดยกำหนดให้ตัวแปรอธิบายที่ขาดหายไปคือ $X_1, X_3, X_5, X_6,$
 X_7, X_8, X_9, X_{10}

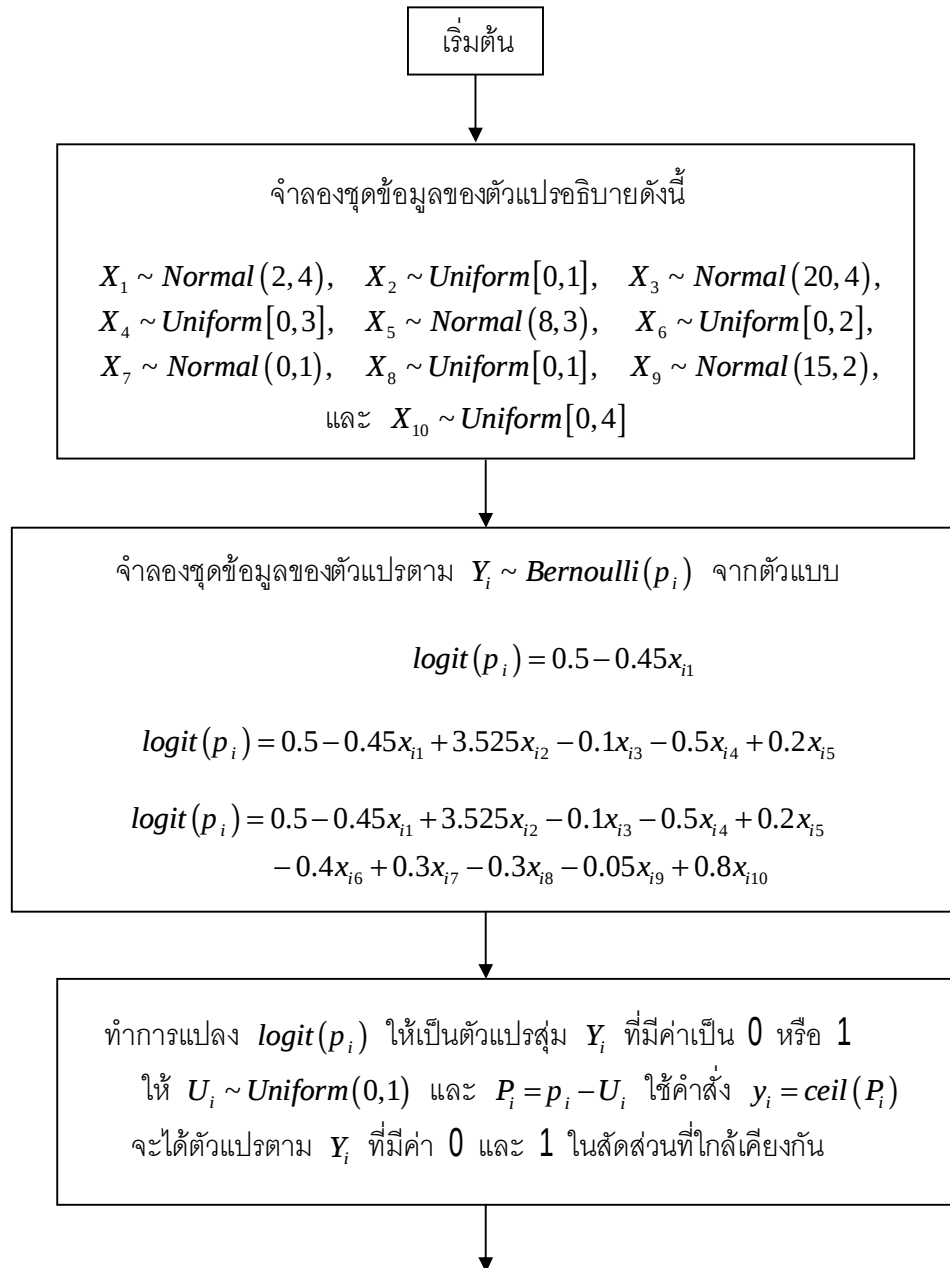
$$\text{logit}(p_i) = b_0 + b_2x_{i2} + b_4x_{i4} \quad (M610)$$

ขั้นตอนในการประมวลผล

1. ในแต่ละตัวแบบกำหนดให้มีขนาดตัวอย่างเท่ากับ 50, 100, 1000 และ 2000
2. นำข้อมูลที่จำลองมาประมาณค่าสัมประสิทธิ์การถดถอยในตัวแบบการถดถอยโลจิสติกตัวแบบ (M01), (M05), ..., (M610) ด้วยวิธีความควรจะเป็นสูงสุดของฟิชเชอร์สกอรिंग (Fisher's scoring method) โดยใช้คำสั่ง PROC LOGISTIC ในโปรแกรม SAS[®] เวอร์ชัน 8.0
3. คำนวณค่า $R_{O,true}^2$ และ $R_{I,true}^2$ ด้วยสมการที่ (1.7) และ (1.8) โดยใช้ค่าเริ่มต้นของ p_i ที่ได้จากการจำลอง $\text{logit}(p_i)$ ซึ่งจะถือว่าเป็นค่าจริงของสัมประสิทธิ์การตัดสินใจ
4. คำนวณค่าประมาณของ R_O^2 และ R_I^2 ด้วยสมการที่ (2.4) และ (2.5) ซึ่งเป็นค่าสัมประสิทธิ์การตัดสินใจที่ยังไม่ปรับค่า
5. คำนวณค่าประมาณของ R_{adj}^2 ที่นำมาศึกษาได้แก่ $R_{O,adj,MS}^2$, $R_{I,adj,MS}^2$, $R_{O,adj,LM}^2$, $R_{I,adj,LM}^2$, $R_{I,adj,SAS_{DEV}}^2$ และ $R_{I,adj,SAS_{AIC}}^2$ ด้วยสมการที่ (1.1), (1.2), (1.3), (1.4), (1.5) และ (1.6) ตามลำดับ
6. ทำซ้ำในขั้นตอนที่ 1 ถึงขั้นตอนที่ 5 จำนวน 1,000 ครั้ง
7. คำนวณค่ามัธยฐาน ($\bar{R}^2$) ความเอนเอียง (bias) ความเอนเอียงสัมพัทธ์ (relative bias) ค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) ของค่าประมาณ R^2 และ R_{adj}^2 และคำนวณร้อยละของค่าประมาณ R^2 และ R_{adj}^2 ที่มีค่าอยู่นอกช่วง $[0,1]$ จากการทำซ้ำ 1,000 ครั้ง
8. วิเคราะห์ผลและรายงานผล
9. สรุปผลของการวิจัย

ขั้นตอนในการวิจัยสามารถเขียนเป็นแผนภาพได้ดังแผนภาพที่ 3.1

แผนภาพ 3.1
แสดงขั้นตอนของการวิจัย



แผนภาพ 3.1 (ต่อ)

