บทที่ 3
วิธีดำเนินการวิจัย

การวิจัยครั้งนี้ต้องการเปรียบเทียบประสิทธิภาพระหว่างตัวแบบถดถอยโลจิสติคแบบสถิตย์และตัวแบบถดถอยโลจิสติคแบบพลวัตร โดยศึกษากับการวิเคราะห์ความถดถอยโลจิสติคแบบ 2 กลุ่ม (Binary Logistic Regression) และพิจารณาตัวแบบถดถอยโลจิสติคแบบพลวัตรสำหรับช่วงเวลาไม่ต่อเนื่อง โดยจะศึกษากับข้อมูลที่ได้จากการจำลองขึ้นด้วยวิธีมอนติคาร์โล ซึ่งการเปรียบเทียบประสิทธิภาพของตัวแบบที่ได้จากการวิเคราะห์ทั้ง 2 แบบในแต่ละกรณีจะพิจารณาจากการทดสอบเพื่อวัดความถูกต้องของการพยากรณ์ด้วยพื้นที่ใต้โค้ง ROC ซึ่งจะได้แสดงขั้นตอนของการวิจัยในลำดับต่อไป

เพื่อความสะดวกในการกล่าวถึงผู้วิจัยจะขอใช้ว่า

1. ตัวแบบสถิตย์ (Static model) แทนตัวแบบถดถอยโลจิสติคแบบสถิตย์
2. ตัวแบบพลวัตร (Dynamic model) แทนตัวแบบถดถอยโลจิสติคแบบพลวัตร

แผนการดำเนินการวิจัย

ข้อกำหนดของข้อมูลจำลอง

ในการศึกษาครั้งนี้ได้กำหนดสถานการณ์ต่างๆ สำหรับการเปรียบเทียบประสิทธิภาพของตัวแบบสถิตย์และตัวแบบพลวัตร ดังนี้

1. จำนวนตัวแปรอิสระที่ใช้ในแต่ละตัวแบบเท่ากับ 3 ตัว และมีการแจกแจงที่แตกต่างกัน ดังนี้

1.1 X_1 มีการแจกแจงแบบปกติ (Normal Distribution) ด้วย $\mu = 0$ และ $\sigma^2 = 1$ นั่นคือ $X_1 \sim N(0,1)$ และมีฟังก์ชันความหนาแน่นอยู่ในรูปของ

$$f(x_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x_1^2} \qquad (3.1)$$

เมื่อ $E[X_1] = 0$, $V[X_1] = 1$

1.2 X_2 มีการแจกแจงแบบเบอร์นูลี (Bernoulli Distribution) ด้วยความน่าจะเป็นของความสำเร็จ $p = 1/2$ นั่นคือ $X_2 \sim Ber(1/2)$ และมีฟังก์ชันความหนาแน่นอยู่ในรูปของ

$$f(x_2) = p^{x_2}(1-p)^{1-x_2} = \left(\frac{1}{2}\right)^{x_2}\left(1-\frac{1}{2}\right)^{1-x_2} \quad ; x_2 = 0,1 \qquad (3.2)$$

เมื่อ $E[X_2] = \frac{1}{2}$, $V[X_2] = \frac{1}{4}$

1.3 X_3 มีการแจกแจงแบบเอกซโพเนนเชียล (Exponential Distribution) ด้วยพารามิเตอร์ $\lambda = 1/36$ นั่นคือ $X_3 \sim Exp(1/36)$ และมีฟังก์ชันความหนาแน่นอยู่ในรูปของ

$$f(x_3) = \begin{cases} \frac{1}{36}e^{-\frac{1}{36}x_3} & ; x > 0 \\ 0 & ; \text{อื่นๆ} \end{cases} \qquad (3.3)$$

เมื่อ $E[X_3] = \frac{1}{\lambda} = 36$, $V[X_3] = \frac{1}{\lambda^2} = (36)^2$ และเมื่อได้ค่า x_3 จากการจำลองแล้วจะทำการแปลงเป็นค่าจำนวนเต็ม โดยที่ $x_3 > 0$

2. ค่าของ x_1 และ x_2 เป็นค่าคงที่ตลอดช่วงของการศึกษา ส่วนค่าของ x_3 จะมีค่าที่เปลี่ยนแปลงไปตามเวลาโดยจะมีค่าลดลงทีละ 1 ในช่วงเวลาถัดไป

3. ตัวแปรอิสระแต่ละตัวไม่มีความสัมพันธ์กัน

4. ตัวแปรตามมีการแจกแจงแบบเบอร์นูลี (Bernoulli distribution) ด้วยความน่าจะเป็นของเหตุการณ์ที่สนใจ $P(Y=1 \mid \mathbf{x}) = \pi(\mathbf{x})$ นั่นคือ $Y \sim Ber(\pi(\mathbf{x}))$ โดยกำหนดค่าพารามิเตอร์ $\boldsymbol{\beta}$ ในสมการ 4.1 กระทำโดยให้ $\bar{Y} \approx 0.5$ เมื่อกำหนดให้ $\beta_1 = 1$, $\beta_2 = 1, \beta_3 = 1/12$ แล้วจึงคำนวณค่า β_0 ได้ $\beta_0 = -4.2$ ดังนั้นจะได้

$$4.1 \quad \pi(\mathbf{x}_{it}) = \frac{e^{-4.2+1x_{1it}+1x_{2it}+(1/12)x_{3it}}}{1+e^{-4.2+1x_{1it}+1x_{2it}+(1/12)x_{3it}}} = p_{it} \quad ; i = 1,2,\ldots,n, \quad t = 1,2,\ldots,T$$

4.2 สำหรับแต่ละหน่วยตัวอย่าง การคำนวณค่าตัวแปรตาม Y ในแต่ละช่วงเวลาจะพิจารณาที่จุดสิ้นสุดของช่วงเวลานั้นหรือที่จุดเริ่มต้นของช่วงเวลาถัดไป โดยหน่วยตัวอย่างใดมี p_{it} จากสมการ 4.1 มากกว่าตัวเลขสุ่มแบบยูนิฟอร์ม [0,1] จะให้เป็น $y_{it} = 1$ นอกนั้นให้มีค่าเป็น $y_{it} = 0$

4.3 หน่วยตัวอย่างใดมี $y_{it}=1$ (เกิดเหตุการณ์ที่สนใจในช่วงเวลา t) หรือมี $x_{3it}=0$ (สิ้นสุดเวลาของหน่วยตัวอย่าง i ที่ช่วงเวลา t) เราจะหยุดการจำลองหน่วยตัวอย่างนั้นในช่วงเวลาต่อไป $t+1,\ldots,T$

5. ทำการจำลองตัวแปรตาม ตามแบบการวิเคราะห์แบบพลวัตรสำหรับช่วงเวลาไม่ต่อเนื่อง โดยแบ่งจำนวนช่วงเวลาย่อยเป็น 12 และ 24 ช่วง ที่มีขนาดความยาวแต่ละช่วงเท่ากัน

6. ใช้ขนาดตัวอย่างเท่ากับ 10,000 หน่วยตัวอย่างที่เป็นอิสระต่อกัน และแบ่งข้อมูลออกเป็น 2 ชุด ได้แก่ ข้อมูลสำหรับการประมาณค่าพารามิเตอร์และข้อมูลสำหรับการพยากรณ์

7. จำนวนการกระทำซ้ำเท่ากับ 1,000 ครั้ง และทำการจำลองแยกกันในแต่ละกรณีที่ศึกษา

8. ทำการจำลองสุ่มให้มีสถานการณ์ตามที่กำหนดและวิเคราะห์ผลด้วยโปรแกรม R ซึ่งเป็นโปรแกรมเปิดรหัส (open source) สำหรับการคำนวณทางสถิติ

ขั้นตอนการดำเนินงาน

1. ที่ช่วงเวลาเริ่มต้น $t=0$ สำหรับแต่ละหน่วยตัวอย่าง ทำการจำลองค่าของตัวแปรอิสระทั้ง 3 ตัว จากการแจกแจงที่กำหนด

2. คำนวณความน่าจะเป็นของเหตุการณ์ที่สนใจ p_{it} ของหน่วยตัวอย่าง i ในช่วงเวลา t นั้น จากสมการ 4.1 ด้วยค่าของตัวแปรอิสระจากขั้นตอน 1 โดยให้ค่าของ x_3 ลดลง 1 หน่วย เพื่อทำการจำลองค่าของตัวแปรตามของหน่วยตัวอย่างในช่วงเวลานั้น y_{it}

3. จำลองเลขสุ่มแบบยูนิฟอร์ม [0,1] จำนวนเท่ากับขนาดตัวอย่าง เพื่อเปรียบเทียบกับค่า p_{it} จากขั้นตอน 2 แล้วแปลงค่า p_{it} เป็นค่าสังเกตของตัวแปรตาม โดยถ้าหน่วยตัวอย่างใดมีค่า p_{it} มากกว่าเลขสุ่มจะกำหนดให้เป็น $y_{it}=1$ นอกนั้นให้มีค่าเป็น $y_{it}=0$

4. ที่ช่วงเวลาถัดไป จำลองค่าสังเกต y_{it} เช่นเดียวกันกับขั้นตอนที่ 2 และ 3 โดยที่หน่วยตัวอย่างใดมี $y_{it}=1$ หรือ มี $x_{3it}=0$ เราจะหยุดการจำลองข้อมูลของหน่วยตัวอย่างนั้นในช่วงเวลาต่อไป กระทำเช่นนี้จนครบทุกช่วงเวลา

5. เมื่อได้ข้อมูลตัวอย่างครบตามจำนวนและช่วงเวลาที่กำหนด ทำการแบ่งข้อมูลออกเป็น 2 ชุด ได้แก่

- ข้อมูลชุดที่ 1 : ใช้สำหรับการประมาณค่าพารามิเตอร์ (Training set)
- ข้อมูลชุดที่ 2 : ใช้สำหรับทดสอบการพยากรณ์ (Validation set)

6. จัดข้อมูลตัวอย่างทั้ง 2 ชุดให้เหมาะสมสำหรับการวิเคราะห์ความถดถอยโลจิสติคทั้งแบบสถิตย์และแบบพลวัตร

7. วิเคราะห์ความถดถอยโลจิสติคทั้งสองแบบด้วยข้อมูลที่ใช้สำหรับการประมาณค่าพารามิเตอร์ของแต่ละแบบ จากการวิเคราะห์จะได้ 2 ตัวแบบประมาณ คือ ตัวแบบสถิตย์และตัวแบบพลวัตร พร้อมทั้งค่าสถิติที่เกี่ยวข้อง

8. ทดสอบการพยากรณ์ของตัวแบบทั้งสองด้วยข้อมูลที่ใช้สำหรับทดสอบการพยากรณ์ของแต่ละแบบ ทำการคำนวณค่าความน่าจะเป็นของเหตุการณ์ที่สนใจจากการพยากรณ์ของแต่ละหน่วยตัวอย่าง $(\hat{p}_i)$ คำนวณค่า Sensitivity และ 1-Specificity ของแต่ละตัวแบบ

9. พล็อตโค้ง ROC พร้อมทั้งคำนวณพื้นที่ใต้โค้ง ROC ของตัวแบบทั้งสอง

10. กระทำซ้ำจนครบตามจำนวนที่กำหนด คือ 1,000 ครั้ง

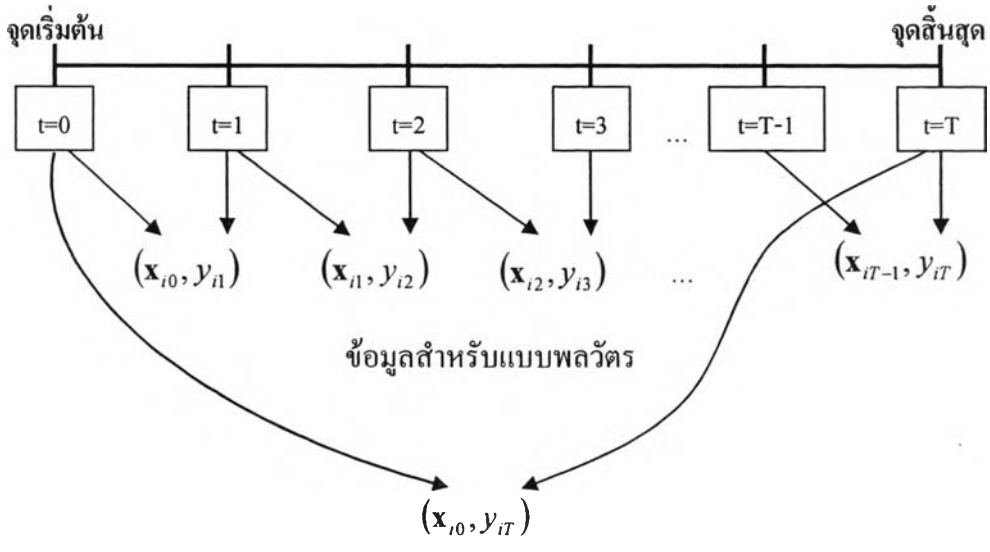
11. เมื่อกระทำซ้ำครบตามจำนวนที่กำหนด ทำการเปรียบเทียบประสิทธิภาพของการพยากรณ์ระหว่างตัวแบบทั้งสองจากพื้นที่ใต้โค้ง ROC ด้วยสถิติทดสอบ T-test

ซึ่งขั้นตอนของกระบวนการวิเคราะห์แสดงดังแผนผังที่ 1 ท้ายบท

การจัดข้อมูลสำหรับการวิเคราะห์แบบสถิตย์และแบบพลวัตร

ข้อมูลสำหรับการวิเคราะห์ความถดถอยโลจิสติคแบบพลวัตรจะเป็นข้อมูลในหลายช่วงเวลา โดยข้อมูลของตัวแปรอิสระเป็นค่าที่จุดเริ่มต้นของแต่ละช่วงเวลาย่อยและข้อมูลของตัวแปรตามเป็นค่าที่จุดสิ้นสุดของช่วงเวลาย่อยนั้นๆ ส่วนข้อมูลสำหรับการวิเคราะห์แบบสถิตย์นั้นแต่ละหน่วยตัวอย่างจะมีเพียง 1 ชุดหรือ 1 ค่าสังเกตของตัวแปรต่างๆ คือ ข้อมูลของตัวแปรอิสระที่จุดเริ่มต้นของการศึกษาและข้อมูลของตัวแปรตามที่จุดสิ้นสุดของการศึกษา ดังรูปที่ 3.1

รูปที่ 3.1 การจัดข้อมูลสำหรับการวิเคราะห์แบบสถิตและแบบพลวัตร



การคำนวณความน่าจะเป็นของเหตุการณ์ของหน่วยตัวอย่าง $(\hat{p}_i)$ สำหรับตัวแบบพลวัตร

จากการวิเคราะห์ความถดถอยโลจิสติคแบบพลวัตรจะได้ตัวแบบประมาณ คือ ตัวแบบพลวัตร ซึ่งเราจะนำมาคำนวณหาความน่าจะเป็นของเหตุการณ์ที่สนใจของการพยากรณ์สำหรับแต่ละหน่วยตัวอย่างในข้อมูลที่ใช้ทดสอบ โดยจะอาศัยหลักของความน่าจะเป็นของการอยู่รอดหรือไม่เกิดเหตุการณ์ และการสูญเสียหรือการเกิดเหตุการณ์ของหน่วยตัวอย่างในช่วงเวลาของการทดสอบ ซึ่งมีขั้นตอนการคำนวณ ดังนี้

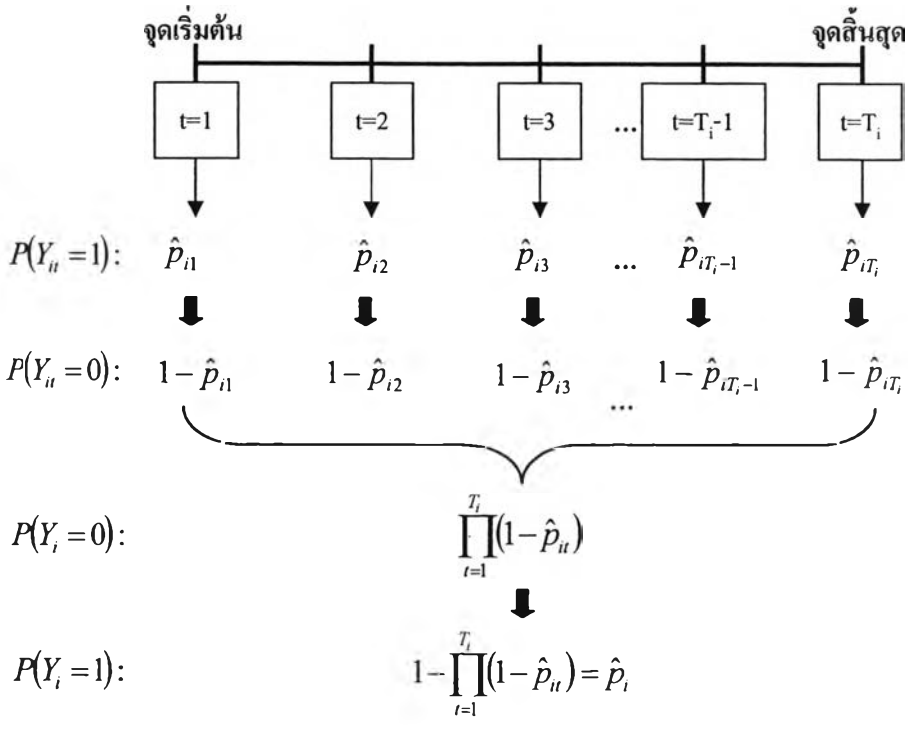
1. สำหรับแต่ละหน่วยตัวอย่าง $i = 1,2,\ldots,n$ คำนวณความน่าจะเป็นของเหตุการณ์ที่สนใจในแต่ละช่วงเวลา $t = 1,2,\ldots,T_i$ ที่ใช้ในการทดสอบ $P(Y_{it} = 1) = \hat{p}_{it}$ จากตัวแบบพลวัตรที่ได้จากการประมาณ

2. คำนวณความน่าจะเป็นของการไม่เกิดเหตุการณ์ที่สนใจในแต่ละช่วงเวลา นั่นคือ $P(Y_{it} = 0) = 1 - \hat{p}_{it}$

3. คำนวณความน่าจะเป็นที่หน่วยตัวอย่างจะไม่เกิดเหตุการณ์ตลอดช่วงเวลาของการทดสอบ นั่นคือ $P(Y_i = 0) = \prod_{t=1}^{T_i}(1 - \hat{p}_{it})$ เมื่อ T_i เป็นจำนวนช่วงเวลาทั้งหมดของหน่วยตัวอย่างนั้น

4. คำนวณความน่าจะเป็นที่หน่วยตัวอย่างจะเกิดเหตุการณ์ในช่วงเวลาของการทดสอบ นั่นคือ $P(Y_i = 1) = 1 - P(Y_i = 0) = \hat{p}_i$ ซึ่งก็คือ ความน่าจะเป็นของเหตุการณ์ที่สนใจของการพยากรณ์สำหรับหน่วยตัวอย่างนั้นในตัวแบบพลวัตรนั่นเอง ขั้นตอนแสดงดังรูปที่ 3.2

รูปที่ 3.2 การคำนวณความน่าจะเป็นของเหตุการณ์ของหน่วยตัวอย่าง $(\hat{p}_i)$ สำหรับตัวแบบพลวัตร



แผนผังที่ 1 ขั้นตอนของการวิเคราะห์ข้อมูลจำลอง