บทที่ 3
วิธีดำเนินการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อต้องการหาจุดแบ่งที่เหมาะสมที่สุดสำหรับการพยากรณ์การจำแนกข้อมูลไม่จัดกลุ่มในตัวแบบโพรบิตแบบ 2 ประเภทสำหรับแต่ละสถานการณ์ที่ต้องการศึกษา โดยจุดแบ่งที่เหมาะสมที่สุดจะให้อัตราความผิดพลาดในการจำแนกกลุ่มมีค่าต่ำสุดหรือสัดส่วนความถูกต้องในการจำแนกกลุ่มมีค่าสูงสุด จากนั้นจะนำผลลัพธ์ของทุกสถานการณ์มาวิเคราะห์ด้วยตัวแบบการถดถอยพหุคูณที่มีผลอันตรกิริยา (Interaction) เพื่อประมาณค่าพารามิเตอร์สำหรับใช้ในการประมาณค่าของจุดแบ่งที่เหมาะสมที่สุดในสถานการณ์อื่นๆ ต่อไป

การจำลองข้อมูลในแต่ละสถานการณ์จะจำลองขึ้นด้วยการทำงานของเครื่องคอมพิวเตอร์โดยใช้เทคนิคมอนติคาร์โล ด้วยโปรแกรม R เนื่องจากวิธีมอนติคาร์โลเป็นเทคนิคที่ใช้ในการวิจัยครั้งนี้ ดังนั้นในตอนแรกของบทนี้จะกล่าวถึงวิธีมอนติคาร์โลก่อน แล้วจึงแสดงรายละเอียดของแผนการดำเนินการวิจัย ขั้นตอนในแผนการดำเนินการวิจัย ตลอดจนโปรแกรมที่ใช้ในการวิจัย ซึ่งรายละเอียดต่างๆเป็นดังนี้

3.1 เทคนิคมอนติคาร์โล

เทคนิคมอนติคาร์โลเป็นการจำลองระบบที่ไม่เปลี่ยนแปลงตามเวลา ซึ่งตัวแบบของการจำลองจะมีลักษณะเป็นตัวแบบทางคณิตศาสตร์ โดยการนำตัวเลขสุ่ม มาประยุกต์ใช้ในการแก้ปัญหาหรือหาคำตอบให้กับระบบที่ยังไม่แน่ใจในผลที่จะเกิดขึ้น ซึ่งมีขั้นตอนที่สำคัญ 3 ขั้นตอนดังนี้

ขั้นตอนที่ 1 การสร้างเลขสุ่ม (Generate Random Number) การสร้างเลขสุ่มจะกำหนดให้มีการแจกแจงแบบยูนิฟอร์มในช่วง [0, 1] และเป็นอิสระซึ่งกันและกัน จากนั้นนำเลขสุ่มนี้ไปสร้างตัวแปรตามลักษณะการแจกแจงที่ต้องการศึกษา เพื่อเป็นข้อมูลของปัญหานั้นๆ

ขั้นตอนที่ 2 การประยุกต์ใช้เลขสุ่มในการแก้ปัญหา ขั้นตอนนี้เป็นการนำตัวแปรที่ได้จากขั้นตอนแรกมาใช้ในการหาค่าต่างๆ ตามปัญหาที่ต้องการศึกษา

ขั้นตอนที่ 3 การทดลอง ขั้นตอนนี้เป็นการทำวิธีนั้นซ้ำๆกัน (Replication) จำนวนหลายครั้ง โดยถือว่าการทำซ้ำๆ กันนั้น เป็นวิธีการเก็บรวบรวมข้อมูลให้มีจำนวนมาก เพื่อลดความไม่แน่นอนของคำตอบ ในการวิเคราะห์หาค่าต่างๆ ได้

จากหลักการของเทคนิคมอนติคาร์โล จะเห็นว่าการใช้เลขสุ่มเพื่อเป็นพื้นฐานในการหาคำตอบของปัญหา เป็นวิธีที่จะนำไปสู่แนวคิดในทางทฤษฎีที่เกี่ยวข้องกับการคำนวณโดยเฉพาะทฤษฎีความน่าจะเป็นที่จะนำไปสู่การอ้างอิงผลสรุปในสถานการณ์ของข้อมูลจริงเพราะไม่มี

ผลกระทบจากเรื่องอื่นๆ เข้ามาเกี่ยวข้องในการทดลอง เมื่อทำซ้ำๆกันเป็นจำนวนมากแล้ว ความคลาดเคลื่อนอย่างสุ่มที่เกิดขึ้นในการวิเคราะห์หาค่าต่างๆ ในแต่ละครั้งให้หมดไป

3.2 แผนการดำเนินการวิจัย

ในการวิจัยครั้งได้จำลองข้อมูลขึ้น โดยกำหนดสถานการณ์จำลองต่างๆ ดังนี้

1. กำหนดให้ข้อมูลของตัวแปรอิสระ (X) เริ่มต้นมีการแจกแจงแบบยูนิฟอร์ม
2. เมื่อข้อมูลตัวแปรตาม (Y^*) ให้มีความสัมพันธ์เชิงเส้นตรงกับตัวแปรอิสระและความผิดพลาด กำหนดให้ค่าพารามิเตอร์เริ่มต้นของสมการการถดถอยเป็นค่าใดๆ ในการวิจัยครั้งนี้ $\beta_i = 0.1\,;\, i = 0,1,2,...,p$ และ $\varepsilon_i \sim N(0,25)\,;\, i = 1,2,...,n$
3. ตัวแปรตาม (Y) เป็นข้อมูลเชิงคุณภาพที่มี 2 ค่า คือ 0 และ 1 โดยกำหนดสัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจในแต่ละชุดข้อมูล (a) เท่ากับ 0.1,0.5 และ0.9
4. กำหนดให้จำนวนของตัวแปรอิสระ (p) ในการวิจัยครั้งนี้แบ่งเป็น 3 ระดับ คือ
   - จำนวนตัวแปรอิสระน้อย คือ 1 และ 2 ตัว
   - จำนวนตัวแปรอิสระปานกลาง คือ 3 และ 4 ตัว
   - จำนวนตัวแปรอิสระมาก คือ 5 และ 6 ตัว
5. กำหนดให้ขนาดตัวอย่าง (n) ในการวิจัยครั้งนี้แบ่งเป็น 3 ระดับ คือ
   - ขนาดตัวอย่างเล็ก คือ 20 และ 40
   - ขนาดตัวอย่างปานกลาง คือ 60 และ 80
   - ขนาดตัวอย่างใหญ่ คือ 100 และ 120
6. กำหนดให้ระดับความสัมพันธ์ระหว่างตัวแปรอิสระ (M)ในการวิจัยครั้งนี้มี 4 กรณี คือ
   -ไม่มีความสัมพันธ์กันระหว่างตัวแปรอิสระ ($\rho = 0$)
   - ความสัมพันธ์ระหว่างตัวแปรอิสระในระดับต่ำ ($\rho = 0.33$)
   - ความสัมพันธ์ระหว่างตัวแปรอิสระในระดับปานกลาง ($\rho = 0.67$)
   - ความสัมพันธ์ระหว่างตัวแปรอิสระในระดับสูง ($\rho = 0.99$)
7. กำหนดระดับนัยสำคัญ (α)ในการวิจัยครั้งนี้ที่ระดับ 0.05
8. ในการวิจัยครั้งนี้ทำการจำลองข้อมูลโดยใช้เทคนิคมอนติคาร์โล โดยการจำลองในแต่ละสถานการณ์จะกระทำซ้ำ 500 รอบ

3.3 ขั้นตอนในการดำเนินการวิจัย

สำหรับการดำเนินการวิจัยมีขั้นตอนดังนี้

1. ศึกษาตัวแบบโพรบิตและวิธีการหาค่าของจุดแบ่งจากทฤษฎี Hadjicostas P.(2006)
2. สร้างข้อมูลตามลักษณะที่กำหนดเพื่อใช้ในการวิจัย
3. หาค่า Y^* จากการสร้างความสัมพันธ์เชิงเส้นตรงกับตัวแปรอิสระ
4. ทำการแปลงค่าตัวแปรตาม Y^* ที่ได้เป็น Y ที่มีค่าเป็น 0 หรือ 1 ตามสัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจและขนาดตัวอย่างตามที่กำหนดไว้
5. ประมาณค่าพารามิเตอร์ของตัวแบบโพรบิตแบบ 2 ประเภทโดยใช้วิธีภาวะน่าจะเป็นสูงสุดในกรณีที่ข้อมูลมีลักษณะที่ต้องการศึกษา
6. คำนวณหาจุดแบ่งที่เหมาะสมที่สุดในแต่สถานการณ์สำหรับตัวแบบโพรบิตแบบ 2 ประเภท ซึ่งทำให้สัดส่วนความถูกต้องในการจำแนกกลุ่มมีค่าสูงสุด
7. ทำการทดลองซ้ำ 500 รอบในแต่ละสถานการณ์
8. คำนวณหาค่าเฉลี่ยของจุดแบ่งที่เหมาะสมที่สุดของแต่ละสถานการณ์จาก 500 รอบ พร้อมทั้งหาค่าร้อยละและช่วงความเชื่อมั่น
9. ใช้ตัวแบบการถดถอยพหุคูณ เพื่อประมาณค่าพารามิเตอร์สำหรับใช้ในการประมาณค่าของจุดแบ่งที่เหมาะสมที่สุดในสถานการณ์อื่นๆ ต่อไป
10. สรุปผลการวิจัยในแต่ละสถานการณ์

3.4 การจำลองข้อมูลที่ใช้ในการวิจัย

การจำลองข้อมูลที่ใช้ในการวิจัย มีขั้นตอนต่างๆดังต่อไปนี้

1. สร้างข้อมูลตัวแปรอิสระเริ่มต้นให้มีการแจกแจงแบบยูนิฟอร์มซึ่งมี equal space คือ กำหนดให้มีค่าเป็นช่วงลบและช่วงบวกเท่าๆกัน ตามขนาดตัวอย่างที่กำหนดไว้ ดังนี้

$$x \sim U\left(-\frac{n}{2}, +\frac{n}{2}\right)$$

2. สร้างจำนวนตัวแปรอิสระตามที่กำหนดไว้ และให้ตัวแปรอิสระดังกล่าวมีความสัมพันธ์กันตามระดับความสัมพันธ์ระหว่างตัวแปรอิสระที่กำหนดไว้ โดยกำหนดให้รูปแบบเมทริกซ์สหสัมพันธ์ (Correlation Matrix) ดังนี้

$$\rho_{p\times p}=\begin{pmatrix}\rho_{11} & \rho_{12} & \cdots & \rho_{1p}\\ \rho_{21} & \rho_{22} & \cdots & \rho_{2p}\\ \vdots & \vdots & \ddots & \vdots\\ \rho_{p1} & \rho_{p2} & \cdots & \rho_{pp}\end{pmatrix}=\begin{pmatrix}1 & \rho & \cdots & \rho^{p-1}\\ \rho & 1 & \cdots & \rho^{p-2}\\ \vdots & \vdots & \ddots & \vdots\\ \rho^{p-1} & \rho^{p-2} & \cdots & 1\end{pmatrix}$$

โดยที่ $\rho_{ij}\ ;i,j=1,2,...,p$ คือสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระตัวที่ i และตัวแปรอิสระตัวที่ j ซึ่งจะมีทั้งหมด $\frac{p(p-1)}{2}$ คู่

- ตัวแปรอิสระไม่มีความสัมพันธ์กัน ($\rho=0$)

โดยสร้างตามเงื่อนไข $\rho_{12}:\rho_{13}:...:\rho_{p,p-1}=0$

- ความสัมพันธ์ระหว่างตัวแปรอิสระในระดับต่ำ ($\rho=0.33$)

โดยสร้างตามเงื่อนไข $\rho_{12}:\rho_{13}:...:\rho_{p,p-1}$ คือ $(0.33)^1:(0.33)^2:...:(0.33)^{\frac{p(p-1)}{2}}$

- ความสัมพันธ์ระหว่างตัวแปรอิสระในระดับปานกลาง ($\rho=0.67$)

โดยสร้างตามเงื่อนไข $\rho_{12}:\rho_{13}:...:\rho_{p,p-1}$ คือ $(0.67)^1:(0.67)^2:...:(0.67)^{\frac{p(p-1)}{2}}$

- ความสัมพันธ์ระหว่างตัวแปรอิสระในระดับสูง ($\rho=0.99$) โดย

สร้างตามเงื่อนไข $\rho_{12}:\rho_{13}:...:\rho_{p,p-1}$ คือ $(0.99)^1:(0.99)^2:...:(0.99)^{\frac{p(p-1)}{2}}$

3. สร้างค่าตัวแปรตาม (Y^*) จากการนำข้อมูลตัวแปรอิสระที่สร้างได้จากข้างต้น มาหาค่าตัวแปรตาม (Y^*) โดยสร้างให้มีความสัมพันธ์เชิงเส้นตรงกับตัวแปรอิสระและความคลาดเคลื่อน ซึ่งมีรูปแบบ ดังนี้

$$Y^*=\beta X'+\varepsilon$$

โดยที่ Y^* เป็นเมตริกซ์ของตัวแปรตามที่ทำการพยากรณ์เพื่อกำหนดค่าเบื้องต้น

เมื่อ X' เป็นเมตริกซ์ของตัวแปรอิสระ

β เป็นเวกเตอร์ของพารามิเตอร์ที่กำหนด กำหนดให้ β เริ่มต้น เท่ากับ 0.1

ε เป็นความคลาดเคลื่อนซึ่ง $\varepsilon\sim N(0,25)$

4. สร้างค่าตัวแปรตาม(Y) ที่มีค่าเป็น 0 หรือ 1 จากค่า Y^* ที่สร้างได้จากข้างต้นโดยทำการแปลงค่าตัวแปรตาม Y^* ที่ได้เป็น Y ที่มีค่าเป็น 0 หรือ 1 ตามสัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจของลักษณะที่สนใจศึกษา และค่า ขนาดตัวอย่างที่กำหนดไว้ข้างต้นดังนี้

4.1 หาค่าจำนวน Y ที่มีค่าเป็น 0 และ Y ที่มีค่าเป็น 1โดย

- Y ที่มีค่าเป็น 0 มีจำนวน =ขนาดตัวอย่าง× สัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจ

- Y ที่มีค่าเป็น1 มีจำนวน =ขนาดตัวอย่าง - [จำนวน Y ที่มีค่าเป็น 0]

4.2 การกำหนดค่า Y^* ให้เป็นค่า Y ที่มีค่าเป็น 0 และ Y ที่มีค่าเป็น 1 โดย

- เรียงลำดับค่าของ Y^* ทั้งหมดที่ได้ จากน้อยไปมาก
- ให้ Y^* ที่มีค่าน้อยที่สุด ตามจำนวน Y ที่มีค่าเป็น 0 เป็น Y ที่มีค่าเป็น 0 และ Y^* นอกนั้นคือ Y ที่มีค่าเป็น 1 ตามจำนวน Y ที่มีค่าเป็น 1

5. ประมาณค่าพารามิเตอร์โดยใช้ตัวแบบโพรบิตแบบ 2 ประเภทด้วยวิธีภาวะความน่าจะเป็นสูงที่สุด
6. หาค่าประมาณของ $\hat{\pi}_i$ โดยนำค่าพารามิเตอร์ที่ได้จากข้อ 5 และ ค่าของตัวแปรอิสระที่สร้างขึ้น มาแทนค่ากลับลงไปในตัวแบบโพรบิตแบบ 2 ประเภท

3.5 **คำนวณค่าของจุดแบ่งโดยทฤษฎี**Hadjicostas P. (2006)

เมื่อได้ข้อมูลที่มีลักษณะตามต้องการแล้ว จากทฤษฎีของ Hadjicostas P. (2006) จะสามารถหาค่าของจุดแบ่งตามขั้นตอนต่อไปนี้

ขั้นที่ 1. เรียงลำดับค่า $\hat{\pi}_i$ จากน้อยไปหามาก $\hat{\pi}_1 < \hat{\pi}_2 < ... < \hat{\pi}_n$

ขั้นที่ 2. หาค่า M(i) สำหรับแต่ละ $i \in \{1,2,...,n\}$ โดย M(i) คือ อันดับของ $\hat{\pi}_i$

แต่ ถ้า $\hat{\pi}_i = \hat{\pi}_j$ จะเลือก M(i) ที่มีค่ามากที่สุดเป็นอันดับของค่า $\hat{\pi}_i$ และ $\hat{\pi}_j$

ขั้นที่ 3. หาค่า a_i สำหรับ $i = 0,1,2,...,n$ โดย $a_i = \sum_{k=1}^{M(i)} (-1)^{y_k}$ ซึ่งแบ่งเป็น 2 กรณี คือ

<u>กรณีที่ 1</u> $a_{i+1} = a_i + \sum_{k=M(i)+1}^{M(i+1)} (-1)^{y_k}$ ถ้า $M(i) < i+1$

<u>กรณีที่ 2</u> $a_{i+1} = a_i$ ถ้า $i+1 \le M(i)$

ขั้นที่ 4. หาค่า I_0 คือเซตของ j ทั้งหมด $j \in \{0,1,2,...,n\}$ ที่ซึ่ง $a_j = \max_{0 \le i \le n} a_i$

ขั้นที่ 5. หาค่า C_0 ซึ่งเป็นเซตของ c_0 ทั้งหมด จาก $C_0 = \bigcup_{i \in I_0} A_i$; $i \in \{0,1,2,...,n\}$

โดย พิจารณาตามเงื่อนไขดังนี้

- $A_i = [0, \hat{\pi}_1)$ ถ้า $i = 0$
- $A_i = [\hat{\pi}_i, \hat{\pi}_{i+1})$ ถ้า $\hat{\pi}_i < \hat{\pi}_{i+1}$ และ $1 \le i < n$
- $A_i = \{\hat{\pi}_i\} = \{\hat{\pi}_{M(i)}\}$ ถ้า $\hat{\pi}_i = \hat{\pi}_{i+1}$ และ $1 \le i < n$
- $A_i = [\hat{\pi}_n, 1]$ ถ้า $i = n$

ขั้นที่ 6. เลือกค่าของจุดแบ่ง (c) ที่เหมาะสมที่สุด ซึ่งคือค่า c ที่ทำให้สัดส่วนของความถูกต้องในการจำแนกกลุ่ม มีค่ามากที่สุด โดย $c \in C_0$ และ $c \in [0,1]$

สัดส่วนของความถูกต้อง $p(c) = \frac{N(c)}{n}$

โดย $p(c)$ คือ สัดส่วนของความถูกต้องในการจำแนกกลุ่มที่จุด c

$N(c)$ คือ จำนวนของความถูกต้องในการจำแนกกลุ่มที่จุด c

3.6 คำนวณค่าเฉลี่ยของจุดแบ่ง ค่าร้อยละของจุดแบ่ง และช่วงความเชื่อมั่นของจุดแบ่งของแต่ละสถานการณ์

- ค่าเฉลี่ยของจุดแบ่ง ($\hat{c}$)

การหาเฉลี่ยของจุดแบ่งของแต่ละสถานการณ์ จากการทดลองโดยการกระทำซ้ำ 500 รอบ ในแต่ละสถานการณ์ กำหนดให้ $\hat{c}$ เป็นตัวประมาณค่าของค่าพารามิเตอร์ c จะได้ว่า

$$\text{ค่าเฉลี่ยของจุดแบ่ง} \quad \hat{c} = \frac{\sum_{k=1}^{N} \hat{c}_{(k)}}{N} \qquad ;k = 1,2,\ldots\ldots,N$$

โดย N คือจำนวนรอบที่กระทำซ้ำในแต่ละสถานการณ์ (N=500)

- ค่าร้อยละของจุดแบ่ง (Percent)

ค่าร้อยละของจุดแบ่งของแต่ละสถานการณ์ จากการทดลองโดยการกระทำซ้ำ 500 รอบ ในแต่ละสถานการณ์

$$\text{ค่าร้อยละของจุดแบ่ง} \quad \hat{c} = \frac{\sum_{k=1}^{N} \hat{c}_{(k)}}{N} \times 100 \qquad ;k = 1,2,\ldots\ldots,N$$

โดย N คือจำนวนรอบที่กระทำซ้ำในแต่ละสถานการณ์ (N=500)

- ช่วงความเชื่อมั่นของจุดแบ่ง (Confidence Interval)

ช่วงความเชื่อมั่นของจุดแบ่งจะบอกถึงค่าต่ำสุดและค่าสูงสุดของค่าของจุดแบ่งที่เป็นไปได้ในแต่ละสถานการณ์

โดยในการวิจัยนี้ กำหนดให้ L คือค่าที่ตำแหน่งเปอร์เซ็นต์ไทล์ที่ $100\,(\alpha/2)$ และ U คือค่าที่ตำแหน่งเปอร์เซ็นต์ไทล์ที่ $100\,(1-\alpha/2)$ โดยกำหนด ค่า $\alpha = 0.05$ สามารถคำนวณช่วงความเชื่อมั่นของจุดแบ่งของ ดังนี้

จาก $P(L < \hat{c} < U) = 1 - \alpha$

โดย L คือ คำนวณจากค่าที่ตำแหน่งเปอร์เซ็นต์ไทล์ที่ 2.5

U คือ คำนวณจากค่าที่ตำแหน่งเปอร์เซ็นต์ไทล์ที่ 97.5

3.7 การวิเคราะห์โดยใช้ตัวแบบการถดถอยพหุคูณ

เมื่อคำนวณค่าร้อยละของจุดแบ่งครบทุกสถานการณ์ที่ต้องการศึกษาแล้ว จะใช้ตัวแบบการถดถอยพหุคูณ เพื่อประมาณค่าพารามิเตอร์ สำหรับใช้ในการหาค่าของจุดแบ่งที่เหมาะสมที่สุดในสถานการณ์อื่นๆ ต่อไป โดยตัวแบบการถดถอยพหุคูณ คือ

$$\text{Percent} = \theta_0 + \theta_1(p) + \theta_2(a) + \theta_3(n) + \theta_4(M) + \theta_5(ap) + \theta_6(an) + \theta_7(aM) + \theta_8(np) + \theta_9(nM) + \theta_{10}(pM) + \theta_{11}(apn) + \theta_{12}(apM) + \theta_{13}(pnM) + \theta_{14}(apnM) + \varepsilon$$

โดย Percent คือ ค่าร้อยละของจุดแบ่ง

p คือ จำนวนตัวแปรอิสระ

n คือ ขนาดตัวอย่าง

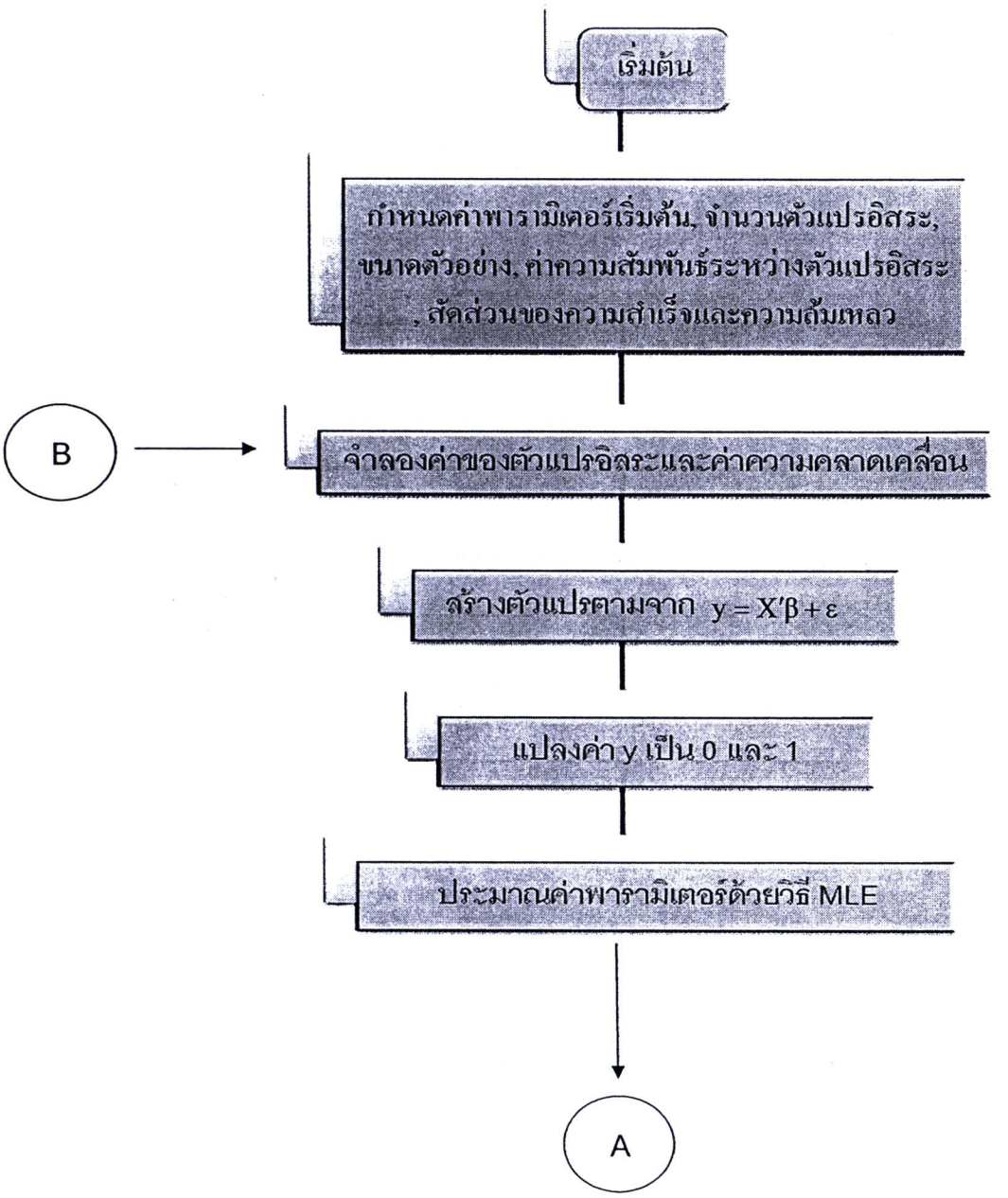
a คือ สัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจ

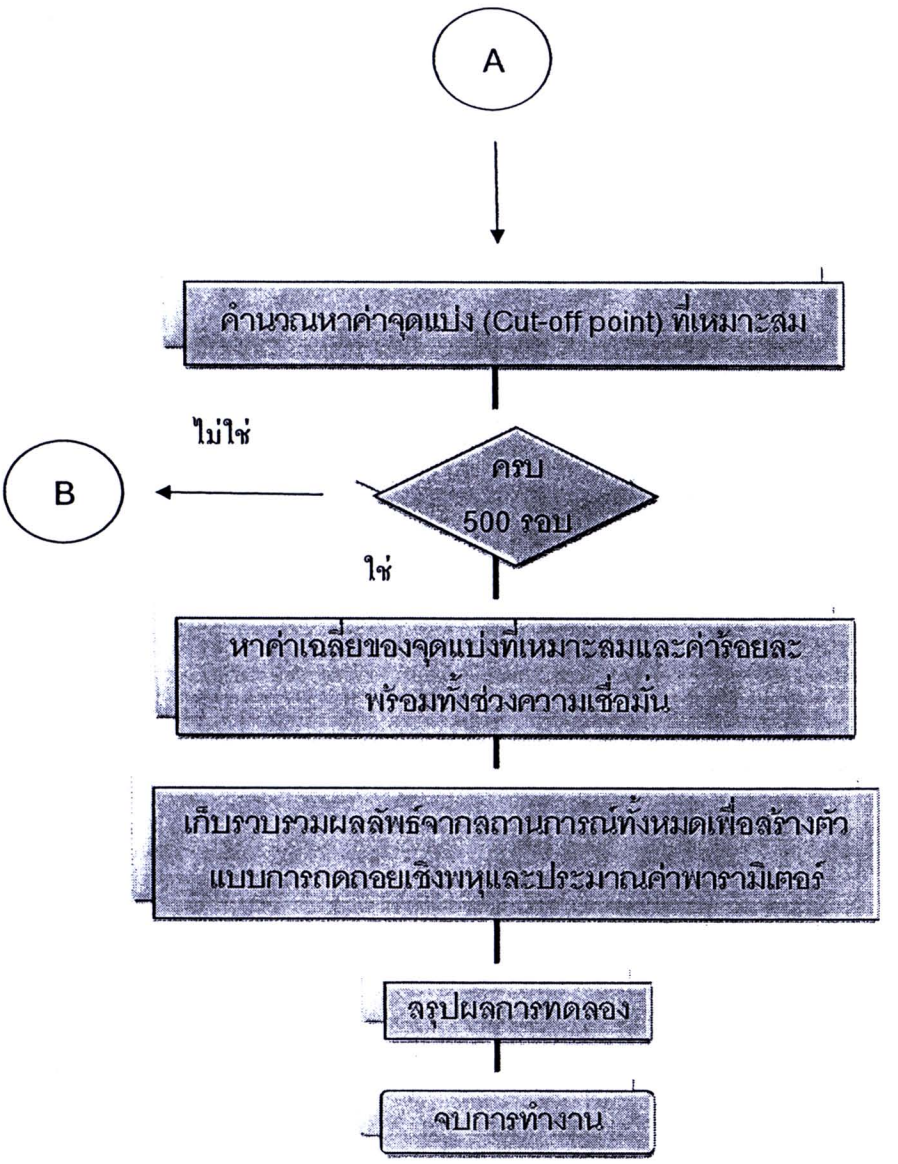
M คือ ระดับความสัมพันธ์ของตัวแปรอิสระ

3.8 สรุปผลการวิจัยในแต่ละสถานการณ์

เมื่อทำการหาค่าของจุดแบ่งที่เหมาะสมที่สุดครบทุกสถานการณ์ที่ต้องการศึกษาแล้ว นำผลการทดลองมาสรุปในรูปตาราง เพื่อดูแนวโน้มว่าปัจจัยที่ต้องการศึกษามีผลต่อค่าของจุดแบ่งอย่างไรในแต่ละสถานการณ์

3.9 ขั้นตอนการทำงานของโปรแกรม





รูปที่ 3.1 แสดงขั้นตอนการทำงานของโปรแกรม