บทที่ 3

วิธีดำเนินการวิจัย

การวิจัยครั้งนี้ต้องการหาจุดแบ่งที่เหมาะสมที่สุดสำหรับการพยากรณ์การจำแนกข้อมูลไม่จัดกลุ่มในตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภทสำหรับแต่ละสถานการณ์ที่ต้องการศึกษา โดยจุดแบ่งที่เหมาะสมที่สุดจะให้อัตราความผิดพลาดในการจำแนกกลุ่มมีค่าต่ำสุด จากนั้นจะนำผลลัพธ์ของทุกสถานการณ์มาวิเคราะห์ด้วยตัวแบบการถดถอยพหุคูณที่มีผลอันตรกิริยา (Interaction) เพื่อนำไปใช้ประมาณหาค่าจุดแบ่งที่เหมาะสมที่สุดในสถานการณ์อื่นๆ ต่อไป

การจำลองข้อมูลในแต่ละสถานการณ์จะจำลองขึ้นด้วยการทำงานของเครื่องคอมพิวเตอร์ โดยใช้เทคนิคมอนติคาร์โล ด้วยโปรแกรม R เนื่องจากวิธีมอนติคาร์โลเป็นเทคนิคที่ใช้ในการวิจัยครั้งนี้ ดังนั้นในตอนแรกของบทนี้จะกล่าวถึงวิธีมอนติคาร์โลก่อน แล้วจึงแสดงรายละเอียดของแผนการดำเนินการวิจัย ขั้นตอนในแผนการดำเนินการวิจัย ตลอดจนโปรแกรมที่ใช้ในการวิจัย ซึ่งรายละเอียดต่างๆเป็นดังนี้

3.1 เทคนิคมอนติคาร์โล

เทคนิคมอนติคาร์โลเป็นการจำลองระบบที่ไม่เปลี่ยนแปลงตามเวลา ซึ่งตัวแบบของการจำลองจะมีลักษณะเป็นตัวแบบทางคณิตศาสตร์ โดยการนำตัวเลขสุ่ม มาประยุกต์ใช้ในการแก้ปัญหาหรือหาคำตอบให้กับระบบที่ยังไม่แน่ใจในผลที่จะเกิดขึ้น ซึ่งมีขั้นตอนที่สำคัญ 3 ขั้นตอน ดังนี้

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

ขั้นตอนที่ 2 การประยุกต์ใช้เลขสุ่มในการแก้ปัญหา ขั้นตอนนี้เป็นการนำตัวแปรที่ได้จากขั้นตอนแรกมาใช้ในการหาค่าต่างๆ ตามปัญหาที่ต้องการศึกษา

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

จากหลักการของเทคนิคมอนติคาร์โล จะเห็นว่าการใช้เลขสุ่มเพื่อเป็นพื้นฐานในการหาคำตอบของปัญหา เป็นวิธีที่จะนำไปสู่แนวคิดในทางทฤษฎีที่เกี่ยวข้องกับการคำนวณโดยเฉพาะทฤษฎีความน่าจะเป็นที่จะนำไปสู่การอ้างอิงผลสรุปในสถานการณ์ของข้อมูลจริงเพราะไม่มีผลกระทบจากเรื่องอื่นๆ เข้ามาเกี่ยวข้องในการทดลอง เมื่อทำซ้ำๆ กันเป็นจำนวนมากแล้ว ความ

คลาดเคลื่อนอย่างสุ่มที่เกิดขึ้นในการวิเคราะห์หาค่าต่างๆ ในแต่ละครั้งให้หมดไป

3.2 แผนการดำเนินการวิจัย

ในการวิจัยครั้งนี้กำหนดสถานการณ์ต่างๆ ดังนี้

3.2.1 กำหนดจำนวนตัวแปรอิสระ คือ 1, 2, 3, 4, 5 และ 6 ตัว ที่เริ่มต้นมีการแจกแจงแบบยูนิฟอร์ม

3.2.2 กำหนดระดับความสัมพันธ์ระหว่างตัวแปรอิสระ คือ

3.2.2.1 กรณีไม่มีความสัมพันธ์กัน ($\rho = 0$)

3.2.2.2 กรณีมีความสัมพันธ์ระดับต่ำ ($\rho = 0.33$)

3.2.2.3 กรณีมีความสัมพันธ์ระดับปานกลาง ($\rho = 0.67$)

3.2.2.4 กรณีมีความสัมพันธ์ระดับสูง ($\rho = 0.99$)

3.2.3 กำหนดให้ความคลาดเคลื่อน (ε) มีการแจกแจงแบบปกติที่มีค่าเฉลี่ยเป็น 0 และความแปรปรวนมีค่าเท่ากับ 25

3.2.4 กำหนดขนาดตัวอย่างในแต่ละตัวแปร คือ 20, 40, 60, 80, 100 และ 120

3.2.5 ตัวแปรตาม (Y) เป็นข้อมูลเชิงคุณภาพที่มีค่าได้เพียง 2 ค่า คือ 0 และ 1 โดยกำหนดสัดส่วนของความล้มเหลว ($Y = 0$) ของลักษณะที่สนใจศึกษา คือ 0.1, 0.5 และ 0.9

3.2.6 กำหนดค่าพารามิเตอร์เริ่มต้นของสมการการถดถอยเป็นค่าใดๆ คือ $\beta_i = 0.1$ เมื่อ $i = 0, 1, 2, ..., p$

3.2.7 กำหนดระดับนัยสำคัญ (α) คือ 0.05 (α = 0.05)

3.2.8 กำหนดจำนวนการกระทำซ้ำในแต่ละสถานการณ์ คือ 500 รอบ

3.3 ขั้นตอนในการดำเนินงานวิจัย

สำหรับการดำเนินการวิจัยมีขั้นตอนดังนี้

1. สร้างข้อมูลที่ใช้ในการวิจัย
2. คำนวณหาค่าจุดแบ่งสำหรับตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภทโดยใช้ทฤษฎีของ Hadjicostas P. (2006) และคำนวณหาค่าเฉลี่ย ค่าร้อยละและช่วงความเชื่อมั่นของจุดแบ่งที่เหมาะสมที่สุดในแต่ละสถานการณ์ ซึ่งทำให้อัตราความผิดพลาดในการจำแนกกลุ่มมีค่าต่ำสุดหรือสัดส่วนความถูกต้องในการจำแนกกลุ่มมมีค่าสูงสุด

3. ใช้ตัวแบบการถดถอยพหุคูณ (Multiple regression model) เพื่อหาความสัมพันธ์ระหว่างค่าร้อยละของจุดแบ่งและปัจจัยต่างๆ คือจำนวนตัวแปรอิสระ ขนาดตัวอย่าง สัดส่วนของความล้มเหลวของลักษณะที่สนใจศึกษาและระดับความสัมพันธ์ระหว่างตัวแปรอิสระ
4. สรุปผลการวิจัยในแต่ละสถานการณ์

3.4 การจำลองข้อมูลที่ใช้ในงานวิจัย

การจำลองข้อมูลที่ใช้ในการวิจัย มีขั้นตอนต่างๆดังต่อไปนี้

1. สร้างข้อมูลตัวแปรอิสระโดยเริ่มต้นมีการแจกแจงแบบยูนิฟอร์มซึ่งมี equal space คือ กำหนดให้มีค่าเป็นช่วงลบและช่วงบวกเท่าๆ กัน ตามขนาดตัวอย่างที่กำหนดไว้ ดังนี้

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

สร้างจำนวนตัวแปรอิสระตามที่กำหนดไว้ และให้ตัวแปรอิสระดังกล่าวมีความสัมพันธ์กันตามระดับความสัมพันธ์ระหว่างตัวแปรอิสระ (M) ที่กำหนดไว้ 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

โดย กรณีตัวแปรอิสระไม่มีความสัมพันธ์กัน $(\rho=0)$

กรณีตัวแปรอิสระมีความสัมพันธ์ระดับต่ำ $(\rho=0.33)$

กรณีตัวแปรอิสระมีความสัมพันธ์ระดับปานกลาง $(\rho=0.67)$

กรณีตัวแปรอิสระมีความสัมพันธ์ระดับสูง $(\rho=0.99)$

2. สร้างข้อมูลตัวแปรตาม (Y^*) ให้มีความสัมพันธ์เชิงเส้นตรงกับตัวแปรอิสระที่สร้างได้จากข้างต้นและความคลาดเคลื่อนซึ่งมีรูปแบบ ดังนี้

$$Y^*=X\widetilde{\beta}+\varepsilon$$

โดยที่ Y^* เป็นเมตริกซ์ของตัวแปรตามที่ทำการพยากรณ์เพื่อกำหนดค่าเบื้องต้น

เมื่อ X เป็นเมตริกซ์ของตัวแปรอิสระ

$\widetilde{\beta}$ เป็นเวกเตอร์ของพารามิเตอร์ที่กำหนด กำหนดให้ β เริ่มต้นเท่ากับ 0.1

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

3. สร้างค่าตัวแปรตาม (Y) ให้มีค่าเป็น 0 หรือ 1 จากค่า Y^* ที่สร้างจากความสัมพันธ์เชิงเส้นตรงกับตัวแปรอิสระ (X) ข้างต้นโดยทำการแปลงค่าตัวแปรตาม Y^* ที่ได้เป็น Y ที่มีค่าเป็น 0 หรือ 1 ตามสัดส่วนของความล้มเหลว (a) ของลักษณะที่สนใจศึกษา และค่า ขนาดตัวอย่าง (n) ที่กำหนดไว้ข้างต้น ดังนี้

   3.1 หาจำนวน Y ที่มีค่าเป็น 0 และ 1โดยจำนวนของ Y ที่มีค่าเป็น 0 เท่ากับผลคูณของขนาดตัวอย่าง (n) กับสัดส่วนของความล้มเหลว (a) ของลักษณะที่สนใจศึกษา ส่วนจำนวนของ Y ที่มีค่าเป็น 1 คือผลต่างของขนาดตัวอย่างกับจำนวน Y ที่มีค่าเป็น 0

   3.2 กำหนดค่า Y^* ให้เป็นค่า Y ที่มีค่าเป็น 0 และ 1 โดยเรียงลำดับค่า Y^* ที่ได้จากน้อยไปมาก จากนั้นกำหนดให้ Y^* ที่มีค่าน้อยที่สุดเป็น Y ที่มีค่าเป็น 0 ตามจำนวนที่คำนวณได้จากข้อ 3.1 และกำหนดให้ Y^* อื่นๆ นั้นเป็น Y ที่มีค่าเป็น 1 ตามจำนวนที่คำนวณได้จากข้อ 3.1

4. ประมาณค่าพารามิเตอร์โดยใช้ตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภทด้วยวิธีภาวะน่าจะเป็นสูงสุด

5. หาค่าประมาณของ $\hat{\pi}_i$ โดยนำค่าพารามิเตอร์ที่ได้จากข้อ 4 และ ค่าของตัวแปรอิสระที่สร้างขึ้น มาแทนค่ากลับลงไปในตัวแบบตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภท

3.5 คำนวณค่าจุดแบ่งโดยใช้ทฤษฎีของ Hadjicostas P. (2006)

วิธีการนี้ Hadjicostas P. (2006) ได้เสนอไว้โดยใช้ผลลัพธ์ทางคณิตศาสตร์ที่มีความถูกต้องแม่นยำในการหาจุดแบ่งที่ทำให้สัดส่วนความถูกต้องในการจำแนกกลุ่มมีค่าสูงสุด โดยเมื่อได้ข้อมูลที่มีลักษณะตามที่ต้องการแล้ว จะสามารถหาค่าจุดแบ่งตามขั้นตอน ดังนี้

ขั้นที่ 1. เรียงอันดับค่า $\hat{\pi}_i$ จากน้อยไปมาก $\hat{\pi}_1 < \hat{\pi}_2 < ... < \hat{\pi}_n$

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 $\hat{\pi}_j$

โดย $M(0) = 0$; $i \le M(i) \le n$

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

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

ii. $a_{i+1} = a_i$ ถ้า $i+1 \le M(i)$

ขั้นที่ 4. หา I_0 ซึ่งเป็นเซตของ j ทั้งหมด โดย $j \in \{0,1,2,...,n\}$ 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. เลือกค่าจุดแบ่ง (c) ที่เหมาะสมที่สุด ซึ่งคือค่า c ที่ทำให้สัดส่วนของความถูกต้องในการจำแนกกลุ่มมีค่ามากที่สุด โดย $c \in C_0$ และ $c \in [0,1]$

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

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

ค่าเฉลี่ยของจุดแบ่ง $\hat{c} = \dfrac{\sum_{k=1}^{N} \hat{c}_{(k)}}{N} \quad ; k = 1,2,.....,N$

โดย N คือจำนวนรอบที่กระทำซ้ำในแต่ละสถานการณ์ (N=500)

- ค่าร้อยละของจุดแบ่ง (Percent)

ค่าร้อยละของจุดแบ่งของแต่ละสถานการณ์ จากการทดลองโดยการกระทำซ้ำ 500 รอบ ในแต่ละสถานการณ์

ค่าร้อยละของจุดแบ่ง $Percent\,of\ \hat{c} = \dfrac{\sum_{k=1}^{N} \hat{c}_{(k)}}{N} \times 100 \ ; k = 1,2,.....,N$

โดย N คือจำนวนรอบที่กระทำซ้ำในแต่ละสถานการณ์ (N=500)

- <u>ช่วงความเชื่อมั่นของจุดแบ่ง (Confidence Interval)</u>

ช่วงความเชื่อมั่นของจุดแบ่งจะบอกถึงค่าต่ำสุดและค่าสูงสุดของค่าจุดแบ่งที่เป็นไปได้ในแต่ละสถานการณ์

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 วิเคราะห์การถดถอยพหุคูณ

เมื่อคำนวณค่าร้อยละของจุดแบ่งครบทุกสถานการณ์ที่ต้องการศึกษาแล้ว จะใช้ตัวแบบการถดถอยพหุคูณ (Multiple regression model) เพื่อประมาณค่าพารามิเตอร์ สำหรับใช้ในการหาค่าจุดแบ่งที่เหมาะสมที่สุดในสถานการณ์อื่นๆ ต่อไป โดยตัวแบบการถดถอยพหุคูณ คือ

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

โดย Percent of $\hat{c}$ คือ ค่าร้อยละของจุดแบ่ง

p คือ จำนวนตัวแปรอิสระ

a คือ สัดส่วนของความล้มเหลวของลักษณะที่สนใจศึกษา

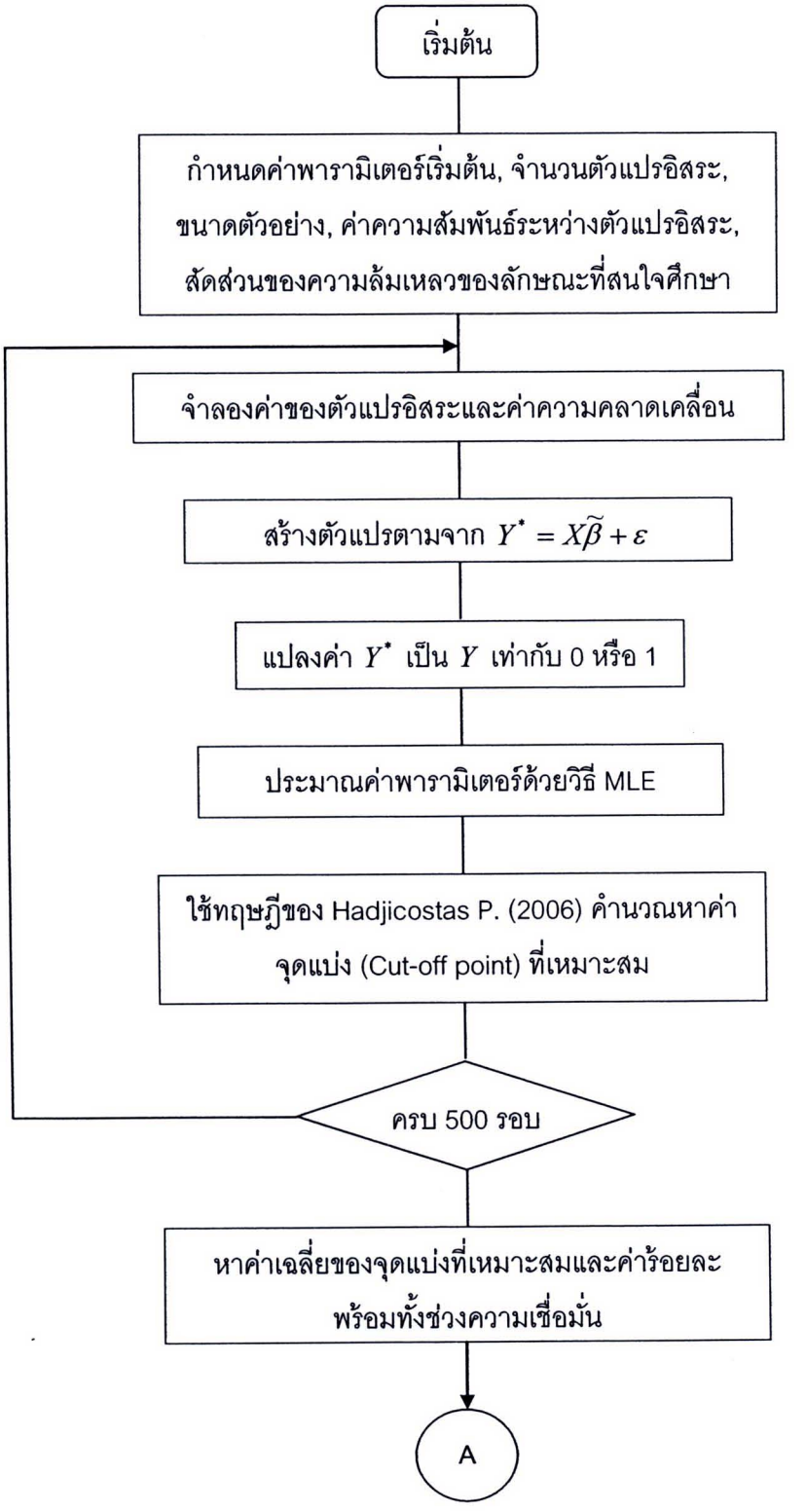
n คือ ขนาดตัวอย่าง

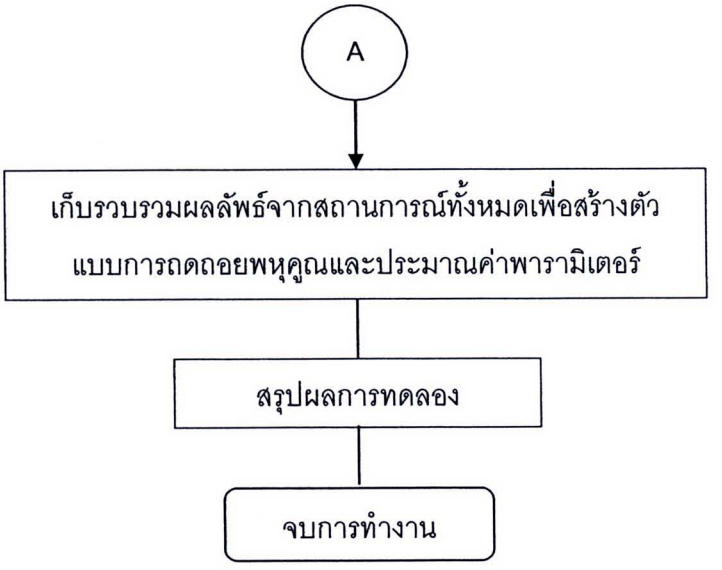
M คือ ระดับความสัมพันธ์ระหว่างตัวแปรอิสระ

3.8 สรุปผลการวิจัยในแต่ละสถานการณ์

เมื่อทำการหาค่าจุดแบ่งที่เหมาะสมที่สุดครบทุกสถานการณ์ที่ต้องการศึกษาแล้ว นำผลการทดลองมาสรุปในรูปตาราง เพื่อดูแนวโน้มว่าปัจจัยที่ต้องการศึกษามีผลต่อค่าจุดแบ่งอย่างไรในแต่ละสถานการณ์

3.9 ขั้นตอนการทำงานของโปรแกรม





รูปที่ 3.1 แสดงขั้นตอนการทำงานของโปรแกรม