บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

วิธีการทางสถิติถูกนำมาใช้เป็นเครื่องมือในการวิเคราะห์ข้อมูลเพื่อประกอบการตัดสินใจในงานด้านต่างๆ อย่างแพร่หลาย ซึ่งในการวิเคราะห์ให้มีประสิทธิภาพนั้น จำเป็นต้องเลือกใช้วิธีการทางสถิติที่เหมาะสมกับข้อมูลและวัตถุประสงค์ของงานในด้านนั้นๆ

ปัจจุบันงานในด้านต่างๆ ไม่ว่าจะเป็นด้านเศรษฐศาสตร์ ด้านธุรกิจ ด้านสังคมศาสตร์ และด้านการแพทย์ มักจะมีข้อมูลเชิงคุณภาพเข้ามาเกี่ยวข้อง ซึ่งก็คือ ความสำเร็จ (Success) และความล้มเหลว (Failure) ของลักษณะที่สนใจศึกษา ดังนั้น ตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภท (Binary Logistic Regression Model) ซึ่งใช้ในการพยากรณ์ตัวแปรตามเชิงคุณภาพที่มีค่าได้เพียง 2 ค่า (Dichotomy or Binary Variable) คือค่า 0 และ 1 ว่าจะอยู่ในกลุ่มใดกลุ่มหนึ่งใน 2 กลุ่มโดยใช้ตัวแปรอิสระเป็นตัวพยากรณ์ จึงถูกนำมาใช้ประโยชน์อย่างแพร่หลาย ตัวอย่างเช่น

จากงานวิจัยของประพิม ศุภศันสนีย์และสุชาดา รัชชุกูล (2548) ได้ทำการศึกษาความสัมพันธ์ระหว่างลักษณะทางสังคมของบุคคลกับการเจ็บป่วยเป็นโรคหลอดเลือดสมองโดยใช้ตัวแบบการถดถอยโลจิสติกเพื่อพยากรณ์โอกาสของการป่วยเป็นโรคหลอดเลือดสมองของบุคคลนั้น

ฐิติพร อนุรักษ์กมลกุล (2548) ได้ทำการศึกษาความมีคุณค่าของรายงานการสอบบัญชีในการพยากรณ์การเข้าสู่การฟื้นฟูกิจการของบริษัทจดทะเบียนในตลาดหลักทรัพย์

Theodossiou P., Kahya E., Saidi R. และ Philippatos G (1996) ได้ทำการวิจัยทางการเงินโดยใช้กลุ่มของปัญหาทางการเงินในการพยากรณ์โอกาสที่จะเข้าถือสิทธิ์ในบริษัท

Hwa H.L.,Ko T.M.,HsiehF.J., Yen M.F.,Chou K.P. และ Tony H.H. (2007) ได้ทำการวิจัยทางการแพทย์โดยใช้ในการพยากรณ์โอกาสในการเกิดดาวน์ซินโดรมในเด็กของหญิงตั้งครรภ์โดยใช้น้ำเหลืองของมารดาเป็นตัวพยากรณ์

งานวิจัยส่วนใหญ่จัดให้แต่ละหน่วย, แต่ละบุคคลหรือแต่ละวัตถุ อยู่ในกลุ่มใดกลุ่มหนึ่งใน 2 กลุ่ม โดยใช้จุดแบ่ง (cut-off point) หรือระดับของความน่าจะเป็นที่ 0.5 โดยให้เหตุผลว่ากลุ่มของความสำเร็จมีโอกาสเกิดขึ้นเท่ากับกลุ่มของความล้มเหลวของลักษณะที่สนใจศึกษา หรือบางงานวิจัยอาจใช้จุดแบ่งค่าหนึ่งที่จะทำให้สัดส่วนของความถูกต้องในการจำแนกกลุ่มมีค่าสูงสุดหรือทำให้อัตราความผิดพลาดในการจำแนกกลุ่ม (Classification error rate) มีค่าต่ำสุดโดยให้เหตุผลว่า ข้อมูลถูกเลือกอย่างสุ่มจากลักษณะที่สนใจศึกษาแล้ว

ในการพยากรณ์โอกาสที่แต่ละหน่วยจะอยู่ในกลุ่มใดกลุ่มหนึ่งใน 2 กลุ่มของตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภท ประเด็นสำคัญที่สุดประเด็นหนึ่ง คือ จุดแบ่งที่เหมาะสมที่สุดคือจุดใดที่ทำให้สัดส่วนหรือร้อยละของความถูกต้องในการพยากรณ์การจำแนกกลุ่มมีค่าสูงสุด ซึ่งไม่มีงานวิจัยใดที่เคยทำการศึกษาการคัดเลือกจุดแบ่งของตัวแบบการถดถอยโลจิสติกที่จะทำให้การจำแนกกลุ่มมีความถูกต้องสูงสุดโดยพิจารณาถึงจำนวนของตัวแปรอิสระ ขนาดตัวอย่าง สัดส่วนของความล้มเหลวของลักษณะที่สนใจศึกษาและระดับความสัมพันธ์ระหว่างตัวแปรอิสระ (Multicollinearity) ของชุดข้อมูล ซึ่งเป็นประเด็นใหญ่ในมุมมองของนักสถิติ ซึ่งลักษณะของชุดข้อมูลเหล่านี้จะมีผลกระทบต่อจุดแบ่งสำหรับการประเมินการพยากรณ์การจำแนกกลุ่มหรือไม่ และถ้ามีผลต่อการคัดเลือกจุดแบ่งแล้วรูปแบบเหล่านั้นคืออะไร ตามที่ได้กล่าวมานั้น ยังไม่มีการหาคำตอบที่ชัดเจนสำหรับคำถามนี้และมีงานวิจัยเพียงเล็กน้อยที่เคยทำการตอบคำถามเหล่านี้ จากงานวิจัยของ Hadjicostas P. (2006) ได้ทำการศึกษาหาจุดแบ่งในตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภทที่ทำให้สัดส่วนความถูกต้องในการจำแนกกลุ่มมีค่าสูงสุด แต่ไม่ได้คำนึงถึงในส่วนของลักษณะของชุดข้อมูล

ผู้วิจัยจึงสนใจทำการศึกษาหาจุดแบ่งที่เหมาะสมที่สุดสำหรับตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภทที่ใช้ในการพยากรณ์โอกาสที่แต่ละหน่วย แต่ละบุคคลหรือแต่ละวัตถุ จะอยู่ในกลุ่มใดกลุ่มหนึ่งใน 2 กลุ่ม ซึ่งทำให้อัตราความผิดพลาดในการจำแนกกลุ่มมีค่าต่ำสุด โดยใช้ทฤษฎีของ Hadjicostas P. (2006) และพิจารณาปัจจัยต่างๆ คือ จำนวนของตัวแปรอิสระ ขนาดตัวอย่าง สัดส่วนของความล้มเหลวของลักษณะที่สนใจศึกษาและระดับความสัมพันธ์ระหว่างตัวแปรอิสระ

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อหาจุดแบ่งที่เหมาะสมที่สุดสำหรับการพยากรณ์การจำแนกข้อมูลไม่จัดกลุ่มในตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภท เมื่อพิจารณาลักษณะของชุดข้อมูล ดังนี้
   - เมื่อสัดส่วนของความล้มเหลวของลักษณะที่สนใจศึกษาเพิ่มขึ้น
   - เมื่อระดับความสัมพันธ์ระหว่างตัวแปรอิสระมีค่าเพิ่มขึ้น
   - เมื่อขนาดตัวอย่างเพิ่มขึ้น
   - เมื่อจำนวนตัวแปรอิสระเพิ่มขึ้น

2. เพื่อหาสมการการถดถอยพหุคูณ (Multiple Regression Model) สำหรับใช้ในการประมาณค่าจุดแบ่งที่เหมาะสมสำหรับการจำแนกข้อมูลไม่จัดกลุ่มในตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภท ในสถานการณ์อื่นๆ ต่อไป

1.3 ขอบเขตของการวิจัย

การวิจัยครั้งนี้มีขอบเขตของการวิจัยสำหรับการดำเนินงานวิจัย ดังนี้

1. ทำการศึกษาตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภทเพื่อหาจุดแบ่งที่เหมาะสมที่สุด
2. ตัวแปรตาม (Y) เป็นข้อมูลเชิงคุณภาพที่มี 2 ค่า คือ 0 และ 1 โดยสัดส่วนของความล้มเหลว ($Y=0$) ของลักษณะที่สนใจศึกษา (a) ในการศึกษาครั้งนี้แบ่งเป็น 3 ระดับ คือ 0.1, 0.5 และ 0.9
3. จำนวนของตัวแปรอิสระ (p) ในการศึกษาครั้งนี้แบ่งเป็น 3 ระดับ คือ
   - จำนวนตัวแปรอิสระน้อย คือ 1 และ 2 ตัว
   - จำนวนตัวแปรอิสระปานกลาง คือ 3 และ 4 ตัว
   - จำนวนตัวแปรอิสระมาก คือ 5 และ 6 ตัว
4. ขนาดตัวอย่าง (n) ในการศึกษาครั้งนี้แบ่งเป็น 3 ระดับ คือ
   - ขนาดตัวอย่างเล็ก คือ 20 และ 40
   - ขนาดตัวอย่างปานกลาง คือ 60 และ 80
   - ขนาดตัวอย่างใหญ่ คือ 100 และ 120

ความสัมพันธ์ระหว่างขนาดตัวอย่างและจำนวนตัวแปรอิสระที่ใช้ในการสร้างชุดข้อมูล คือ

$$\frac{n}{p} \geq 20$$

5. ระดับความสัมพันธ์ระหว่างตัวแปรอิสระ (M) ในการศึกษาครั้งนี้แบ่งเป็น 4 ระดับ คือ

   5.1 กรณีไม่มีความสัมพันธ์กัน $(\rho = 0)$

   5.2 กรณีมีความสัมพันธ์ระดับต่ำ $(\rho = 0.33)$

   5.3 กรณีมีความสัมพันธ์ระดับปานกลาง $(\rho = 0.67)$

   5.4 กรณีมีความสัมพันธ์ระดับสูง $(\rho = 0.99)$

   โดยกำหนดให้รูปแบบเมทริกซ์สหสัมพันธ์ (Correlation Matrix) คือ

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

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

6. ตัวแปรอิสระ (X) มีการแจกแจงเริ่มต้น เป็นการแจกแจงแบบยูนิฟอร์ม
7. กำหนดค่าพารามิเตอร์เริ่มต้นของสมการการถดถอยเป็นค่าใดๆ ในการศึกษาครั้งนี้ คือ $\beta_i = 0.1\,; i = 0,1,2,...,p$ และ $\varepsilon_i \sim N(0,25); i = 1,2,...,n$
8. กำหนดระดับนัยสำคัญ (α)ในการศึกษาครั้งนี้ที่ระดับ 0.05 (α = 0.05)
9. ในการศึกษาครั้งนี้ทำการจำลองข้อมูลโดยใช้เทคนิคมอนติคาร์โล (Monte Carlo Simulation) โดยการจำลองในแต่ละสถานการณ์จะกระทำซ้ำ 500 รอบ (N=500)

1.4 ข้อตกลงเบื้องต้น

การวิจัยครั้งนี้มีข้อตกลงเบื้องต้นสำหรับการดำเนินงานวิจัย ดังนี้

1. ศึกษาตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภท (Binary Logistic Regression Model) โดยมีรูปแบบ คือ

$$\pi_i = E\left(Y_i = 1 \middle| X_{1i} = x_{1i}, X_{2i} = x_{2i}, ..., X_{pi} = x_{pi}\right)$$
$$= \frac{\exp\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi}\right)}{1 + \exp\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi}\right)} \quad ; i = 1,2,...,n$$

โดยที่ π_i คือ ความน่าจะเป็นที่เกิดความสำเร็จ (Y=1) ของลักษณะที่สนใจศึกษาเมื่อมีตัวแปรอิสระ $X_1, X_2, ..., X_p$

$\beta_0, \beta_1, \beta_2, ..., \beta_p$ คือ ค่าสัมประสิทธิ์การถดถอยโลจิสติก

$X_{1i}, X_{2i}, ..., X_{pi}$ คือ ตัวแปรอิสระตัวที่ 1, 2, ..., p

exp คือ ค่า exponential

n คือ ขนาดตัวอย่าง

2. ศึกษาตัวแบบที่ตัวแปรอิสระมีการแจกแจงแบบยูนิฟอร์ม (Uniform Distribution) ซึ่งฟังก์ชันความน่าจะเป็นของ X อยู่ในรูปของ

$$f(x;a,b) = \frac{1}{b-a} \quad ; a < x < b$$ โดยที่ a และ b เป็นค่าคงที่และ $a < b$

ซึ่งมีค่าเฉลี่ยและค่าความแปรปรวน คือ

$$E(X) = \mu = \frac{a+b}{2}$$

$$Var(X) = \sigma^2 = \frac{(b-a)^2}{12}$$

1.5 คำจำกัดความที่ใช้ในการวิจัย

1. **ตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภท** (Binary Logistic Regression Model) หมายถึง ตัวแบบที่ศึกษาความสัมพันธ์ระหว่างตัวแปรอิสระหลายตัวกับตัวแปรตามซึ่งเป็นตัวแปรเชิงคุณภาพที่มีค่าได้เพียง 2 ค่า (dichotomous หรือ binary variable) ส่วนตัวแปรอิสระเป็นตัวแปรเชิงปริมาณหรือตัวแปรเชิงคุณภาพหรืออาจจะมีทั้งตัวแปรเชิงปริมาณหรือตัวแปรเชิงคุณภาพก็ได้ เมื่อได้รูปแบบความสัมพันธ์ระหว่างตัวแปรแล้ว จะนำไปใช้ในการประมาณค่าตัวแปรตามหรือการพยากรณ์โอกาสที่แต่ละหน่วยจะอยู่ในกลุ่มใดกลุ่มหนึ่ง
2. **จุดแบ่ง** (Cut-off point) หมายถึง ค่าความน่าจะเป็นที่ใช้ในการพิจารณาการจำแนกกลุ่มว่าแต่ละหน่วยจะอยู่ในกลุ่มใดระหว่างกลุ่มของความสำเร็จและกลุ่มของความล้มเหลวของลักษณะที่สนใจศึกษา
3. **ข้อมูลไม่จัดกลุ่ม** (ungrouped data) หมายถึง ข้อมูลที่เพิ่งได้จากการเก็บรวบรวมข้อมูลยังไม่ได้จัดแบ่งกลุ่ม จำแนกกลุ่ม แยกประเภทและไม่อยู่ในรูปตารางความถี่
4. **ฟังก์ชันโลจิท** (logit function) หมายถึง ฟังก์ชันการแปลงแบบโลจิสติก ซึ่งแปลงค่าความน่าจะเป็นจากช่วง (0,1) เป็นสมการ logit (π_i) ในช่วง $(-\infty,\infty)$
5. **การแจกแจงแบบเบอร์นูลลี** (Bernoulli Distribution) หมายถึง

   ตัวแปรสุ่ม Y เรียกว่า ตัวแปรสุ่มเบอร์นูลลี กล่าวคือ

   ถ้า $Y=0$ เมื่อ การทดลองไม่เกิดเหตุการณ์ที่สนใจ (ความล้มเหลว)

   และถ้า $Y=1$ เมื่อ การทดลองเกิดเหตุการณ์ที่สนใจ (ความสำเร็จ)

   โดยที่ $P(Y=1)=p$

   และ $P(Y=0)=1-p$, $0<p<1$

   เราอาจเขียนแทนด้วย $Y \sim Ber(p)$ ซึ่งฟังก์ชันความน่าจะเป็นอยู่ในรูปของ

   $$P(Y=y)=p^{Y}(1-p)^{1-Y}$$
6. **ความสัมพันธ์ระหว่างตัวแปรอิสระ** (multicollinearity) หมายถึง สถานการณ์ที่ตัวแปรอิสระมีความสัมพันธ์กัน
7. **ช่วงความเชื่อมั่น** (Confidence Interval) คือ ช่วงของการประมาณค่าเป็นช่วงที่อยู่รอบจุดของค่าประมาณ ช่วงของค่าเหล่านี้เป็นค่าเฉพาะที่บ่งบอกความเชื่อมั่นว่ามีค่าพารามิเตอร์อยู่ในช่วงนี้

1.6 ประโยชน์ที่คาดว่าจะได้รับ

1. เพื่อทราบจุดแบ่งที่เหมาะสมที่สุดสำหรับการพยากรณ์การจำแนกข้อมูลไม่จัดกลุ่มในตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภท เมื่อพิจารณาลักษณะของชุดข้อมูล ดังนี้
    - เมื่อสัดส่วนของความล้มเหลวของลักษณะที่สนใจศึกษาเพิ่มขึ้น
    - เมื่อระดับความสัมพันธ์ระหว่างตัวแปรอิสระมีค่าเพิ่มขึ้น
    - เมื่อขนาดตัวอย่างเพิ่มขึ้น
    - เมื่อจำนวนตัวแปรอิสระเพิ่มขึ้น
2. เพื่อทราบสมการการถดถอยพหุคูณ (Multiple Regression Model) สำหรับใช้ในการประมาณค่าจุดแบ่งที่เหมาะสมสำหรับการจำแนกข้อมูลไม่จัดกลุ่มในตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภท ในสถานการณ์อื่นๆ ต่อไป
3. เพื่อใช้เป็นแนวทางในการศึกษาวิจัยต่อไป

1.7 วิธีดำเนินการวิจัย

1. ศึกษาค้นคว้าเอกสารและข้อมูลที่เกี่ยวข้องกับงานวิจัย
2. จำลองข้อมูลตามขอบเขตที่ต้องการศึกษา
3. คำนวณหาจุดแบ่งที่เหมาะสมที่สุดสำหรับข้อมูลที่มีลักษณะตามที่ต้องการศึกษา
4. ทำการทดลองซ้ำ 500 รอบในแต่ละสถานการณ์
5. คำนวณหาค่าเฉลี่ยของจุดแบ่งที่เหมาะสมที่สุดและค่าร้อยละ (Percent) พร้อมทั้งช่วงความเชื่อมั่น (Confidence Interval)
6. ใช้ตัวแบบการถดถอยพหุคูณ (Multiple regression model) เพื่อประมาณค่าพารามิเตอร์สำหรับใช้ในการประมาณค่าจุดแบ่งที่เหมาะสมที่สุดในสถานการณ์อื่นๆ ต่อไป
7. สรุปผลที่ได้จากการวิจัย