บทที่ 1
บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ปัจจุบันวิธีการทางสถิติได้ถูกนำมาใช้อย่างแพร่หลาย โดยเฉพาะการวิเคราะห์ความถดถอยเป็นวิธีการวิเคราะห์ทางสถิติที่นิยมนำมาใช้ในการวิจัยด้านต่างๆ ซึ่งการวิเคราะห์ความถดถอยนี้เป็นวิธีการวิเคราะห์ข้อมูล เพื่ออธิบายถึงความสัมพันธ์ระหว่างตัวแปรตาม (Dependent variable) และตัวแปรอิสระ (Independent variables) ซึ่งในการวิเคราะห์ความถดถอย ตัวแปรอิสระจะเป็นตัวแปรเชิงปริมาณเพียงอย่างเดียว หรืออาจจะมีตัวแปรบางตัวเป็นตัวแปรเชิงปริมาณและตัวแปรบางตัวเป็นตัวแปรเชิงกลุ่ม ในขณะที่ตัวแปรตามจะเป็นตัวแปรเชิงปริมาณ แต่มีข้อมูลในบางด้านตัวแปรตามที่ทำการศึกษาไม่ได้เป็นตัวแปรเชิงปริมาณแต่เป็นตัวแปรเชิงกลุ่ม เช่น ด้านการควบคุมคุณภาพ ในการศึกษาสินค้าเสียที่ผลิตว่าในแต่ละกล่องมีสินค้าเสียจำนวนกี่ชิ้น โดยทำการศึกษาสินค้าที่ผลิตทั้งหมดจำนวน 90 กล่อง ในแต่ละกล่อง มีจำนวนสินค้า 10 ชิ้น ซึ่งสนใจว่าในจำนวน 10 ชิ้นมีสินค้าเสียกี่ชิ้น นอกจากนี้ยังมี ด้านเศรษฐศาสตร์ ด้านสังคมศาตร์ และ ด้านการตลาด เป็นต้น จะเห็นได้ว่าตัวแปรตามจะมีการแจกแจงแบบทวินาม

ดังนั้นจึงไม่สามารถนำการวิเคราะห์ความถดถอยแบบปกติมาอธิบายได้ เนื่องจากกราฟที่ได้ไม่เป็นเส้นตรงและทำให้ค่าคลาดเคลื่อนไม่มีการแจกแจงแบบปกติด้วย โดยจะใช้การวิเคราะห์ความถดถอยโลจิสติค (Logistic regression analysis) เข้ามาอธิบายแทน ซึ่งวัตถุประสงค์และแนวคิดยังคงเหมือนกับการวิเคราะห์ความถดถอยแบบปกติ คือ เพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ นำสมความถดถอยที่ได้ไปประมาณหรือพยากรณ์ค่าตัวแปรตามเมื่อกำหนดค่าตัวแปรอิสระ

เนื่องจากตัวแบบที่นำมาใช้แสดงความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระในการวิเคราะห์ความถดถอย คือ

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi} + \varepsilon_i \quad ; i = 1, 2, ..., m$$

และค่าเฉลี่ยของ y เมื่อกำหนด x คือ

$$E(y_i|x_i) = \beta_0 + \beta_1 x_{1i} \ldots + \beta_p x_{pi}$$

เมื่อ $-\infty < E(y_i|x_i) < \infty$

แต่ในการวิเคราะห์ความถดถอยโลจิสติค เมื่อตัวแปรตามมีการแจกแจงแบบทวินามนั่นคือ $y_i|x_i \sim Bin(n_i, \pi(x_i))$ จะได้

$$E(y_i|x_i) = n_i\pi(x_i)$$

ดังนั้นจะทำการแปลงค่า $E(y_i|x_i)$

$$\eta = E(y_i|x_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$$

ด้วยฟังก์ชันเชื่อมโยงแบบโลจิสติค

$$\eta = \ln\left[\frac{n_i\pi(x_i)}{n_i - n_i\pi(x_i)}\right] = \ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right]$$

เมื่อ η แทนฟังก์ชันเชื่อมโยงจะได้ว่า

$$\text{logit}\ (\pi(x_i)) = \ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$$

$$\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = e^{\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}}$$

$$\pi(x_i) = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}}$$

และเรียกตัวแบบที่ได้ว่าตัวแบบถดถอยโลจิสติค (Logistic regression model)

ดังนั้น ถ้ามีความเข้าใจในเรื่องตัวแบบถดถอยโลจิสติคเป็นอย่างดีแล้วก็สามารถวิเคราะห์ข้อมูลได้อย่างมีประสิทธิภาพ แต่เนื่องจาก $\beta_0, \beta_1, ..., \beta_p$ เป็นพารามิเตอร์ที่ไม่ทราบค่า ซึ่งถ้าสามารถประมาณค่าพารามิเตอร์ได้ใกล้เคียงค่าจริงแล้วก็ยิ่งทำให้การพยากรณ์ถูกต้องมากยิ่งขึ้น

ในการประมาณค่าพารามิเตอร์ของตัวแบบถดถอยโลจิสติคสามารถทำได้หลายวิธี คือ วิธีความควรจะเป็นสูงสุด (Maximum Likelihood Method) วิธีการถ่วงน้ำหนัก(Weighting Method) และวิธีปรับแก้เบื้องต้น (Prior Correction Method) ซึ่งผู้วิจัยสนใจศึกษาและเปรียบเทียบการประมาณค่าพารามิเตอร์ทั้ง 3 วิธี

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาวิธีการประมาณค่าพารามิเตอร์ของตัวประมาณของตัวแบบถดถอยโลจิสติค 3 วิธี คือ
   1.1 วิธีความควรจะเป็นสูงสุด (Maximum Likelihood Method)
   1.2 วิธีการถ่วงน้ำหนัก (Weighting Method)
   1.3 วิธีปรับแก้เบื้องต้น (Prior Correction Method)
2. เพื่อเปรียบเทียบวิธีการประมาณค่าทั้ง 3 วิธี

1.3 ขอบเขตการวิจัย

1. ตัวแปรอิสระ(X) ที่ศึกษามี 3 ระดับ คือ 3 , 5 และ 7 ตัว
2. ตัวแปรอิสระแต่ละตัวเป็นค่าคงที่
3. ตัวแปรอิสระแต่ละตัวไม่มีความสัมพันธ์กัน
4. ตัวแปรอิสระเป็นข้อมูลเชิงปริมาณที่มีการแจกแจงดังนี้
   - การแจกแจงปกติ (Normal Distribution)
     ฟังก์ชันความหนาแน่นอยู่ในรูปของ

$$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}(x_i-\mu)^2}$$

เมื่อ $E(X)=\mu, V(X)=\sigma^2$

5. ตัวแปรตาม Y_i เป็นอิสระซึ่งกันและกันและมีการแจกแจงแบบทวินามด้วยพารามิเตอร์ n_i และ $\pi(x_i)$ ซึ่งกำหนดดังนี้

   5.1 กำหนด n_i=n เท่ากับ 10 20 และ 30

   5.2 $\pi(x_i)=\frac{e^{\beta_0+\beta_1 x_{1i}+\ldots+\beta_p x_{pi}}}{1+e^{\beta_0+\beta_1 x_{1i}+\ldots+\beta_p x_{pi}}}$; i=1,2,...,m

6. จำนวนกลุ่มที่ทำการศึกษา(m) คือ 30 , 90 , 150 และ 210

7. กำหนดค่าเฉลี่ยความน่าจะเป็นของเหตุการณ์ที่สนใจของประชากร $\bar{\pi}(x_i)$ คือ 0.1 , 0.3 , 0.5 และ 0.8
8. กำหนด $\beta = 1$
9. กำหนดจำนวนการกระทำซ้ำในแต่ละสถานการณ์เป็น 500 ครั้ง
10. การวิจัยครั้งนี้ได้ใช้วิธีการจำลองสุ่ม (Simulation) ให้มีสถานการณ์ตามที่กำหนดด้วยโปรแกรม S – plus

1.4 เกณฑ์การตัดสินใจ

การเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ว่าวิธีการใดน่าจะมีความถูกต้องมากที่สุด จะพิจารณาจากเกณฑ์เปรียบเทียบ คือ ค่าระยะทางมาหาลาโนบิสเฉลี่ย (Average Mahalanobis Distance) เป็นเกณฑ์ประกอบการตัดสินใจ

- **ค่าระยะทางมาหาลาโนบิสเฉลี่ย**

$$AMH = \sum_{i=1}^{500} \frac{\sqrt{\left(\hat{\beta}-\beta\right)'\left(Cov^{-1}\left(\hat{\beta}_i,\hat{\beta}_j\right)\right)\left(\hat{\beta}-\beta\right)}}{500}$$

เมื่อ β_j หมายถึง ค่าจริงของพารามิเตอร์ในตัวแบบถดถอยโลจิสติค ในการจำลองรอบที่ j

$\hat{\beta}_j$ หมายถึง ค่าประมาณของพารามิเตอร์ในตัวแบบถดถอยโลจิสติค ในการจำลองรอบที่ j

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. เพื่อเป็นแนวทางในการตัดสินใจว่าควรใช้วิธีการใดในการประมาณค่าพารามิเตอร์ของตัวแบบถดถอยโลจิสติค
2. เพื่อเป็นแนวทางในการศึกษาเกี่ยวกับตัวแบบถดถอยโลจิสติค
3. เพื่อเป็นแนวทางในการนำไปประยุกต์ใช้กับข้อมูลจริง