การหาจุดแบ่งของตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภท สำหรับการพยากรณ์การจำแนกข้อมูลโดยใช้ฟังก์ชันโพรบิตเป็นฟังก์ชันเชื่อมโยง

The Cut-off Point Estimation of Binary Logistic Regression Model For Predictive Classification Using Probit Function as a Link Function

สุภิญญา คำมั่น[1*] และสุพล ดุรงค์วัฒนา[2]

Supinya Khammun and Supol Durongwatana

[1]นักศึกษาปริญญาโท สาขาวิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย

[2]รองศาสตราจารย์ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย

Abstract

The objective of this study is to find out the optimal cut-off point estimation of binary logistic regression model for predictive classification using probit function. The interesting factors are the numbers of independent variables (p) are 1, 2, 3, 4 and 5, the sample size (n) are 50, 100, 150, 200 and 250, the failure rate (a) are 0.1, 0.5 and 0.9 and the degree of multicollinearity among independent variables with 3 levels; low level (0 < $Max\{|r_{ij}|\} \leq$ 0.30), medium level (0.30 < $Max\{|r_{ij}|\} \leq$ 0.60) and high level (0.60 < $Max\{|r_{ij}|\} \leq$ 0.90). The data in all situations are generated using Monte Carlo technique through R-program. The cut-off point is captured using Hadjicostas P. (2006) theory. The results can be summarized as follow:

As the number of independent variables changed and the other factors are kept constant, with the failure rate equal to 0.5, the mean value of the cut-off point are switching swing near 0.5 and the mean value of the cut-off point increases when the sample size is big and the degree of multicollinearity among independent variables is high level and with other the failure rate, the mean value of the cut-off point mostly converge to value of 0.5. As sample size changed and the other factors are kept constant, with the failure rate equals to 0.5, the mean value of the cut-off point are converge to value of 0.5 and with other the failure rate, the mean value of the cut-off point mostly converge to value of 0.5. As the failure rate changed and the other factors are kept constant, with the number of independent variables equal to 1, and the mean value of the cut-off point increases and converges approximately to 0.5. When the number of independent variables equal to 2, 3, 4 and 5, the mean value of the cut-off point mostly converge to value of 0.5. As the degree of multicollinearity among independent variables changed and the other factors are kept constant, the mean value of the cut-off point mostly converge to value of 0.5.

Keywords: *cut-off point, binary logistic regression, classification error rate, probit function*

บทคัดย่อ

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อหาจุดแบ่งที่เหมาะสมที่สุดสำหรับตัวแบบถดถอยโลจิสติกแบบ 2 ประเภท สำหรับการจำแนกข้อมูลโดยใช้ฟังก์ชันโพรบิตเป็นฟังก์ชันเชื่อมโยง โดยปัจจัยที่สนใจศึกษาในงานการวิจัยครั้งนี้ คือ จำนวนตัวแปรอิสระเป็น 1, 2, 3, 4 และ 5 ขนาดตัวอย่างเป็น 50, 100, 150, 200 และ 250 สัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจเป็น 0.1, 0.5 และ 0.9 และระดับความสัมพันธ์ระหว่างตัวแปรอิสระเป็น 3 ระดับ คือ ความสัมพันธ์กันในระดับต่ำ ($0 < Max\{|r_{ij}|\} \leq 0.30$) ระดับปานกลาง ($0.30 < Max\{|r_{ij}|\} \leq 0.60$) และระดับสูง ($0.60 < Max\{|r_{ij}|\} \leq 0.90$) ซึ่งข้อมูลทั้งหมดจำลองโดยเทคนิคมอนติคาร์โล ด้วยโปรแกรม R การหาค่าจุดแบ่ง จะใช้ทฤษฎีของ Hadjicostas P. (2006) ผลการวิจัยสรุปได้ดังนี้

กรณีที่จำนวนตัวแปรอิสระเปลี่ยนแปลง แต่ปัจจัยอื่นๆ คงที่ พบว่า ที่สัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจเท่ากับ 0.5 ค่าเฉลี่ยของจุดแบ่งมีค่าสูงต่ำสลับไปมาแต่แกว่งอยู่ใกล้ ค่า 0.5 แต่ค่าเฉลี่ยจะมีแนวโน้มเพิ่มขึ้นเมื่อระดับความสัมพันธ์สูงและขนาดตัวอย่างใหญ่ และที่สัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจอื่นๆ ส่วนใหญ่ค่าเฉลี่ยของ จุดแบ่งมีค่าสูงต่ำสลับไปมาแต่แกว่งอยู่ใกล้ค่า 0.5 กรณีที่ขนาดตัวอย่างเปลี่ยนแปลง แต่ปัจจัยอื่นๆ คงที่ พบว่า ที่สัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจ เท่ากับ 0.5 ค่าเฉลี่ยของจุดแบ่งมีค่าสูงต่ำสลับไปมาแต่แกว่งอยู่ใกล้ค่า 0.5 และที่สัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจอื่นๆ ส่วนใหญ่ค่าเฉลี่ยของจุดแบ่งมีค่าสูงต่ำสลับไปมาแต่แกว่งอยู่ใกล้ ค่า 0.5 กรณีที่สัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจเปลี่ยนแปลง แต่ปัจจัยอื่นๆ คงที่ พบว่า ที่จำนวนตัวแปรอิสระ เท่ากับ 1 ค่าเฉลี่ยของจุดแบ่งมีค่าเพิ่มขึ้นลู่เข้าหาค่า 0.5 และที่จำนวนตัวแปรอิสระ เท่ากับ 2, 3, 4, 5 ส่วนใหญ่ค่าเฉลี่ยของจุดแบ่งมีค่าสูงต่ำสลับไปมาแต่แกว่งอยู่ใกล้ค่า 0.5 กรณีที่ระดับความสัมพันธ์ระหว่างตัวแปรอิสระเปลี่ยนแปลงไป แต่ปัจจัยอื่น ๆ คงที่ พบว่า ส่วนใหญ่ค่าเฉลี่ยของจุดแบ่งมีค่าสูงต่ำสลับไปมาแต่แกว่งอยู่ใกล้ค่า 0.5

คำสำคัญ: *จุดแบ่ง, การถดถอยโลจิสติกแบบ 2 ประเภท, อัตราความผิดพลาดในการจำแนกกลุ่ม, ฟังก์ชัน โพรบิต*

บทนำ

ปัจจุบันงานวิจัยในด้านต่างๆ ได้นำเอาเทคนิคการวิเคราะห์ทางสถิติมาเป็นเครื่องมือในการวิเคราะห์ข้อมูลเพื่อประกอบการตัดสินใจ ซึ่งข้อมูลส่วนใหญ่ที่นำมาวิเคราะห์มักจะมีข้อมูลเชิงคุณภาพเข้ามาเกี่ยวข้อง ซึ่งก็คือ การเกิดเหตุการณ์ที่สนใจ (Success) และการไม่เกิดเหตุการณ์ที่สนใจ (Failure) ดังนั้น ตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภท (Binary Logistic Regression Model) ซึ่งใช้ในการพยากรณ์ตัวแปรตาม เชิงคุณภาพที่มีค่าได้เพียง 2 ค่า (Dichotomy or Binary Variable) คือค่า 0 และ 1 ว่าจะอยู่ในกลุ่มใดกลุ่มหนึ่งใน 2 กลุ่มจึงถูกนำมาใช้ประโยชน์อย่างแพร่หลาย

งานวิจัยส่วนใหญ่ที่ใช้ตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภทในการวิเคราะห์ข้อมูล มักจะมีข้อตกลงเบื้องต้นโดยให้กลุ่มที่เกิดเหตุการณ์ที่สนใจมีโอกาสเกิดขึ้นเท่ากับกลุ่มที่ไม่เกิดเหตุการณ์ที่ โดยใช้จุดแบ่ง (cut-off point) ที่ 0.5 โดยผู้วิจัยส่วนใหญ่มักไม่คำนึงถึงจำนวนของตัวแปรอิสระ ขนาดตัวอย่าง สัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจและระดับความสัมพันธ์ระหว่างตัวแปรอิสระ ซึ่งลักษณะของชุดข้อมูลเหล่านี้อาจมีผลกระทบต่อจุดแบ่งสำหรับการประเมินการพยากรณ์การจำแนกกลุ่ม

ดังนั้น ผู้วิจัยจึงสนใจทำการศึกษาหาจุดแบ่งที่เหมาะสมที่สุดสำหรับตัวแบบการถดถอยโลจิสติก แบบ 2 ประเภทโดยใช้ฟังก์ชันโพรบิตเป็นฟังก์ชันเชื่อมโยงที่ทำให้อัตราความผิดพลาดในการจำแนกกลุ่มมีค่าต่ำสุด

โดยพิจารณาจากปัจจัยต่างๆ คือ จำนวนของตัวแปรอิสระ ขนาดตัวอย่าง สัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจ และระดับความสัมพันธ์ระหว่างตัวแปรอิสระ

วัตถุประสงค์

เพื่อหาจุดแบ่งที่เหมาะสมที่สุดสำหรับการพยากรณ์การจำแนกข้อมูลในตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภทโดยใช้ฟังก์ชันโพรบิตเป็นฟังก์ชันเชื่อมโยง เมื่อชุดข้อมูลมีลักษณะดังนี้

- จำนวนตัวแปรอิสระเพิ่มขึ้น
- ขนาดตัวอย่างเพิ่มขึ้น
- สัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจเพิ่มขึ้น
- ระดับความสัมพันธ์ระหว่างตัวแปรอิสระเพิ่มขึ้น

และนำปัจจัยที่กล่าวข้างต้นมาพิจารณารวมกัน ซึ่งแปรเปลี่ยนไปพร้อมกัน

แนวคิด ทฤษฎี กรอบแนวคิด

1. ตัวแบบโพรบิตและการประมาณค่า

ตัวแบบโพรบิตเป็นตัวแบบที่ศึกษาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ เมื่อตัวแปรตามเป็นตัวแปรเชิงคุณภาพที่มี 2 ลักษณะ และตัวแปรอิสระเป็นตัวแปรเชิงปริมาณหรือตัวแปรหุ่น นำไปใช้หาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอิสระ และใช้ในการพยากรณ์โอกาสที่แต่ละหน่วยจะอยู่ใน กลุ่มใดกลุ่มหนึ่งได้

จากตัวแปรตาม (Y_i) ซึ่งเป็นตัวแปรเชิงคุณภาพที่เป็นได้ 2 ค่า คือ 0 กับ 1 และ $X_{i1}, X_{i2},..., X_{ik}$ เป็นตัวแปรอิสระของค่าสังเกตที่ได้จากหน่วยตัวอย่างที่ i โดยที่ $i = 1,2,...,n$ โดยมี ตัวแปรแฝง คือ Y_i^* ซึ่งเป็นค่าที่วัดไม่ได้ จึงไม่ทราบค่าที่แท้จริง ทราบเพียงแต่ผลที่เกิดขึ้น โดย Y_i^* ของหน่วยตัวอย่างที่ i เป็นฟังก์ชันเชิงเส้นของตัวแปรอิสระ $X_{i1}, X_{i2},..., X_{ik}$ และนั่นคือ

$$Y_i^* = \beta' X_i - \varepsilon_i$$

เมื่อ ε_i คือค่าความคลาดเคลื่อนของหน่วยตัวอย่าง และ $\varepsilon_i \overset{iid}{\sim} N(0, 1)$

จะได้ $$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \text{ or } \varepsilon_i < \beta' X_i \\ 0 & \text{if } Y_i^* \le 0 \text{ or } \varepsilon_i \ge \beta' X_i \end{cases}$$

จาก
$$\begin{aligned} \pi_i &= P(Y_i = 1) \\ &= P(Y_i^* > 0) \\ &= P(\varepsilon_i < \beta' X_i) \quad ; i = 1,2,...,n \\ &= \Phi(\beta' X_i) \end{aligned}$$

$$= \int_{-\infty}^{\beta X_i'} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-\varepsilon^2}{2}\right) d\varepsilon$$

ดังนั้น ตัวแบบโพรบิต สามารถเขียนแปลงให้อยู่ในรูป

$$\Phi^{-1}(\pi_i) = \beta' X_i$$

$$= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} \quad \text{..... (1)}$$

จากสมการที่ (1) พบว่า การแปลงดังกล่าว จะทำให้เราได้ความสัมพันธ์เชิงเส้นตรงระหว่างตัวแปร

ฟังก์ชันโพรบิต คือ $Probit\left(E[Y_i|\tilde{X}_i^T = \tilde{x}_i^T]\right) = \Phi^{-1}\left(E[Y_i|\tilde{X}_i^T = \tilde{x}_i^T]\right) = \Phi^{-1}(\pi_i)$

ซึ่งฟังก์ชันโพรบิตจะทำการแปลงค่า π_i จากช่วง (0, 1) เป็นค่าที่อยู่ในช่วง $(-\infty, \infty)$

1.1 ฟังก์ชันความน่าจะเป็น (Likelihood function)

เนื่องจากตัวแปรตาม (Y_i) ที่ทำการศึกษามีเพียง 2 ค่า คือ 0 กับ 1 จึงใช้ฟังก์ชันการแจกแจงแบบเบอร์นูลลี่

$$P(Y_i = y_i) = p^{y_i}(1-p)^{1-y_i} \quad ; y_i = 0,1$$

สร้างฟังก์ชันของการแจกแจงความน่าจะเป็นร่วม (Joint Probability Density Function) ของหน่วยตัวอย่างอิสระ n ค่า โดยการคูณฟังก์ชันการแจกแจงความน่าจะเป็นของทุกหน่วยตัวอย่าง $(g(Y_i))$

$$g(Y_1, Y_2, ..., Y_n) = \prod_{i=1}^{n} g(Y_i)$$

$$= \prod_{i=1}^{n} P_i^{y_i}(1-P)^{1-y_i}$$

$$= \prod_{i=1}^{n} [\Phi(\beta X_i')]^{y_i} [1 - \Phi(\beta X_i')]^{1-y_i}$$

ดังนั้น $$L(\beta) = \prod_{i=1}^{n} [\Phi(x_i'\beta)]^{y_i} [1 - \Phi(x_i'\beta)]^{1-y_i} \quad \text{..... (2)}$$

1.2 การประมาณค่าพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimation)

การหาค่าตัวประมาณค่าสัมประสิทธิ์การถดถอย $(\hat{\beta})$ ด้วยวิธีภาวะความน่าจะเป็นสูงสุดสามารถเขียนให้อยู่ในรูปของลอการิทึม (Logarithm) หรือความควรจะเป็นลอการึทึม (Log-Likelihood) ได้ดังนี้

$$\ln L(\beta) = \sum_{i=1}^{n} \{y_i \ln[\Phi(x_i'\beta)] + (1-y_i)\ln[1 - \Phi(x_i'\beta)]\} \quad \text{.....(3)}$$

$$= \sum_{y_i=0} \ln[1-\Phi(x_i'\beta)] + \sum_{y_i=1} \ln[\Phi(x_i'\beta)] \qquad \text{.....(4)}$$

และเงื่อนไขอันดับแรก (First Order) สำหรับการให้สมการที่ (4) มีค่าสูงสุด (Maximization) และสามารถแก้สมการหาค่าเงื่อนไขอันดับแรกของฟังก์ชันโพรบิต ได้ดังนี้

$$\frac{\partial \ln L(\beta)}{\partial \beta} = \sum_{i=1}^{n} \left\{ \frac{y_i \phi(.)}{\Phi(.)} + (1-y_i) \left[\frac{-\phi(.)}{1-\Phi(.)} \right] \right\} X_i = 0$$

$$= \sum_{y_i=0} \left[\frac{-\phi(X_i'\beta)}{1-\Phi(X_i'\beta)} \right] X_i + \sum_{y_i=1} \left[\frac{\phi(X_i'\beta)}{\Phi(X_i'\beta)} \right] X_i = 0 \qquad \text{.....(5)}$$

เมื่อ $\phi(.)$ คือ ฟังก์ชันความหนาแน่นของตัวแปรสุ่มที่มีการแจกแจงแบบปกติ

$\Phi(.)$ คือ ฟังก์ชันการแจกแจงสะสมของตัวแปรสุ่มที่มีการแจกแจงแบบปกติ

เมื่อประมาณค่าพารามิเตอร์ด้วยวิธีความควรจะเป็นสูงสุดในตัวแบบการถดถอยโลจิสติก แบบ 2 ประเภทโดยใช้ฟังก์ชันโพรบิตเป็นฟังก์ชันเชื่อมโยงแล้วจะสามารถนำไปใช้ในการพยากรณ์การจำแนกกลุ่มของตัวแบบ ดังนี้

- หน่วยที่ i จะถูกจัดให้อยู่ในกลุ่มที่เกิดเหตุการณ์ที่สนใจ (Y=1) ถ้า

$$\hat{\pi}_i = \Phi(\beta' X_i) > c \qquad ;0 \le c \le 1$$

- หน่วยที่ i จะถูกจัดให้อยู่ในกลุ่มที่ไม่เกิดเหตุการณ์ที่สนใจ (Y=0) ถ้า

$$\hat{\pi}_i = \Phi(\beta' X_i) \le c \qquad ;0 \le c \le 1$$

เมื่อ c คือ จุดแบ่งหรือระดับของความน่าจะเป็นที่ใช้ในการพิจารณาการจำแนกกลุ่มว่าแต่ละหน่วยจะอยู่ในกลุ่มใดระหว่างกลุ่มที่เกิดเหตุการณ์ที่สนใจ และกลุ่มที่ไม่เกิดเหตุการณ์ที่สนใจ

2. การหาจุดแบ่ง โดยทฤษฎีของ Hadjicostas P. (2006)

<u>บทตั้ง</u> $N(c) = \sum_{j=1}^{M(i)} (1-y_j) + \sum_{j=M(i)+1}^{n} y_j$ สำหรับ $i \in \{0,1,2,...,n\}$ ใดๆ และ $c \in A_i$

<u>ทฤษฎีบท</u> ให้ $a_i = \sum_{k=1}^{M(i)} (-1)^{y_k}$ สำหรับ $i = 0,1,2,\ldots,n$ ให้ I_0 เป็นเซตของ j ทั้งหมด $j \in \{0,1,2,...,n\}$ ซึ่ง $a_j = \max_{0 \le i \le n} a_i$ และให้ C_0 เป็นเซตของ c_0 ทั้งหมด $c_0 \in [0,1]$ ซึ่ง $p(c_0) = \max_{c \in [0,1]} p(c)$ แล้ว $C_0 = \bigcup_{i \in I_0} A_i$

จากบทตั้งและทฤษฎีข้างต้น จะได้ขั้นตอนการหาจุดแบ่งดังนี้

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

2. สำหรับแต่ละ $i \in \{1,2,...,n\}$ ใดๆ หาค่า $M(i)$ ซึ่ง $M(i)$ คือ $\max\ j \in \{1,2,...,n\}$ ถ้า $\hat{\pi}_i = \hat{\pi}_j$ โดย $M(0) = 0\ ; i \le M(i) \le n$

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

3.1 $a_{i+1} = a_i + \sum_{k=M(i)+1}^{M(i+1)} (-1)^{y_k}$ ถ้า $M(i) < i+1$

3.2 $a_{i+1} = a_i$ ถ้า $i+1 \le M(i)$

4. หา I_0 ซึ่งเป็นเซตของ j ทั้งหมด โดย $j \in \{0,1,2,...,n\}$ ซึ่ง $a_j = \max_{0 \le i \le n} a_i$

5. หา C_0 ซึ่งเป็นเซตของ c_0 ทั้งหมด โดย $c_0 \in [0,1]$ ซึ่ง $p(c_0) = \max_{c \in [0,1]} p(c)$ แล้ว $C_0 = \bigcup_{i \in I_0} A_i$ สำหรับ $i \in \{0,1,2,...,n\}$ จะได้ว่า

$A_i = [0, \hat{\pi}_1)$ 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 $i = n$

ข้อสังเกต $\bigcup_{i=0}^{n} A_i = [0,1]$

6. เลือกค่าของจุดแบ่ง c โดย $c \in C_0$ และ $c \in [0,1]$ ซึ่งเป็นค่า c ที่ทำให้สัดส่วนของ ความถูกต้องในการจำแนกกลุ่มที่จุด c มีค่ามากที่สุด

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

เมื่อ $p(c)$ คือ สัดส่วนความถูกต้องในการจำแนกกลุ่มที่จุด c

$N(c)$ คือ จำนวนของความถูกต้องในการจำแนกกลุ่มที่จุด c

วิธีการวิจัย

การวิจัยครั้งนี้ มีวิธีดำเนินการวิจัยสำหรับการดำเนินการวิจัย ดังนี้

1. ศึกษาค้นคว้าเอกสารและข้อมูลที่เกี่ยวข้องกับงานวิจัย
2. จำลองข้อมูลตามขอบเขตการวิจัย ดังนี้
   - สร้างข้อมูลตัวแปรอิสระโดยให้มีการแจกแจงเริ่มต้นเป็นแบบยูนิฟอร์ม ตามขนาดตัวอย่างที่กำหนดไว้ คือ 50, 100, 150, 200 และ 250 และให้ตัวแปรอิสระดังกล่าวมีความสัมพันธ์กันตามระดับความสัมพันธ์ระหว่างตัวแปรอิสระที่กำหนดไว้
   - สร้างค่าตัวแปรตาม (Y^*) โดยสร้างให้มีความสัมพันธ์เชิงเส้นตรงกับตัวแปรอิสระ ที่สร้างได้จากข้างต้นและความคลาดเคลื่อน ซึ่งมีรูปแบบ ดังนี้ $Y_i^* = \beta' X_i - \varepsilon_i$

- สร้างค่าตัวแปรตาม (Y) โดยแปลงค่าตัวแปรตาม Y^* ที่ได้เป็น Y ที่มีค่าเป็น 0 หรือ 1 ตามสัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจ และขนาดตัวอย่างที่กำหนดไว้

3. คำนวณหาจุดแบ่งที่เหมาะสมที่ทำให้อัตราความผิดพลาดในการจำแนกกลุ่มมีค่าต่ำสุดในแต่ละสถานการณ์
4. ทำการทดลองซ้ำ 500 รอบในแต่ละสถานการณ์
5. คำนวณหาค่าเฉลี่ยของจุดแบ่งที่เหมาะสมที่สุด เพื่อนำมาคำนวณค่าร้อยละ (Percent) และช่วงความเชื่อมั่น (Confidence Interval)
6. สรุปผลที่ได้จากการวิจัย

ผลการวิจัย

การวิจัยครั้งนี้มีวัตถุประสงค์เพื่อหาจุดแบ่งของตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภท สำหรับการจำแนกข้อมูลโดยใช้ฟังก์ชันโพรบิตเป็นฟังก์ชันเชื่อมโยง โดยมีผลของการวิจัยเป็นดังนี้

1. กรณีจำนวนตัวแปรอิสระเพิ่มขึ้น

เมื่อสัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจมีค่าเท่ากับ 0.1 ที่ระดับความสัมพันธ์ระหว่างตัวแปรอิสระอยู่ในระดับปานกลาง ขนาดตัวอย่าง เท่ากับ 100 พบว่า ค่าเฉลี่ยของจุดแบ่งมีแนวโน้มเพิ่มขึ้น และที่ระดับความสัมพันธ์ระหว่างตัวแปรอิสระอยู่ในระดับสูง ขนาดตัวอย่าง เท่ากับ 100 พบว่า ค่าเฉลี่ยของ จุดแบ่งมีแนวโน้มเพิ่มขึ้นแต่มีค่าลดลงเมื่อตัวแปรอิสระ เท่ากับ 5 สำหรับสถานการณ์อื่นๆ พบว่า ค่าเฉลี่ยของจุดแบ่งมีค่าสูงต่ำสลับไปมาแต่แกว่งอยู่ใกล้ค่า 0.5 เมื่อจำนวนตัวแปรอิสระเพิ่มขึ้น

เมื่อสัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจมีค่าเท่ากับ 0.5 ที่ระดับความสัมพันธ์ระหว่างตัวแปรอิสระอยู่ในระดับสูง ขนาดตัวอย่าง เท่ากับ 250 พบว่า ค่าเฉลี่ยของจุดแบ่งมีแนวโน้มเพิ่มขึ้น สำหรับสถานการณ์อื่นๆ พบว่า ค่าเฉลี่ยของจุดแบ่งมีค่าสูงต่ำสลับไปมาแต่แกว่งอยู่ใกล้ค่า 0.5 เมื่อจำนวนตัวแปรอิสระเพิ่มขึ้น

เมื่อสัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจมีค่าเท่ากับ 0.9 ที่ระดับความสัมพันธ์ระหว่าง ตัวแปรอิสระอยู่ในระดับต่ำ ขนาดตัวอย่าง เท่ากับ 50 และระดับปานกลาง ขนาดตัวอย่าง เท่ากับ 200 ตามลำดับ พบว่า ค่าเฉลี่ยของจุดแบ่งมีแนวโน้มลดลง และที่ระดับความสัมพันธ์ระหว่างตัวแปรอิสระอยู่ในระดับสูง ขนาดตัวอย่าง เท่ากับ 100 พบว่า ค่าเฉลี่ยของจุดแบ่งมีแนวโน้มเพิ่มขึ้นแต่มีค่าลดลงเมื่อตัวแปรอิสระ เท่ากับ 5 สำหรับสถานการณ์อื่นๆ พบว่า ค่าเฉลี่ยของจุดแบ่งมีค่าสูงต่ำสลับไปมาแต่แกว่งอยู่ใกล้ค่า 0.5 เมื่อจำนวน ตัวแปรอิสระเพิ่มขึ้น

2. กรณีขนาดตัวอย่างเพิ่มขึ้น

เมื่อสัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจมีค่าเท่ากับ 0.1 ที่ระดับความสัมพันธ์ระหว่าง ตัวแปรอิสระอยู่ในระดับสูง จำนวนตัวแปรอิสระ เท่ากับ 4 พบว่า ค่าเฉลี่ยของ จุดแบ่งมีแนวโน้มเพิ่มขึ้น สำหรับสถานการณ์อื่นๆ พบว่า ค่าเฉลี่ยของจุดแบ่งมีค่าสูงต่ำสลับไปมาแต่แกว่งอยู่ใกล้ค่า 0.5 เมื่อขนาดตัวอย่างเพิ่มขึ้น

เมื่อสัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจมีค่าเท่ากับ 0.5 พบว่า ทุกสถานการณ์ให้ค่าเฉลี่ยของจุดแบ่งมีค่าสูงต่ำสลับไปมาแต่แกว่งอยู่ใกล้ค่า 0.5 เมื่อขนาดตัวอย่างเพิ่มขึ้น

เมื่อสัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจมีค่าเท่ากับ 0.9 ที่ระดับความสัมพันธ์ระหว่างตัวแปรอิสระอยู่ในระดับปานกลาง จำนวนตัวแปรอิสระ เท่ากับ 5 พบว่า ค่าเฉลี่ยของจุดแบ่งมีแนวโน้มลดลง สำหรับสถานการณ์อื่นๆ พบว่า ค่าเฉลี่ยของจุดแบ่งมีค่าสูงต่ำสลับไปมาแต่แกว่งอยู่ใกล้ค่า 0.5 เมื่อขนาดตัวอย่างเพิ่มขึ้น

3. กรณีสัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจเพิ่มขึ้น

เมื่อจำนวนตัวแปรอิสระ เท่ากับ 1 พบว่า ทุกสถานการณ์ให้ค่าเฉลี่ยของจุดแบ่งมีค่าสูงต่ำสลับไปมาแต่แกว่งอยู่ใกล้ค่า 0.5

เมื่อจำนวนตัวแปรอิสระ เท่ากับ 2 ที่ระดับความสัมพันธ์ระหว่างตัวแปรอิสระอยู่ในระดับต่ำ ขนาดตัวอย่าง เท่ากับ 250 และระดับปานกลาง ขนาดตัวอย่าง เท่ากับ 150 พบว่า ค่าเฉลี่ยของจุดแบ่งมีแนวโน้มลดลง และที่ระดับความสัมพันธ์ระหว่างตัวแปรอิสระอยู่ในระดับปานกลาง ขนาดตัวอย่าง เท่ากับ 50, 200 และระดับสูง ขนาดตัวอย่าง เท่ากับ 50 พบว่า ค่าเฉลี่ยของจุดแบ่งมีแนวโน้มเพิ่มขึ้นแต่มีค่าเพิ่มขึ้น สำหรับสถานการณ์อื่นๆ พบว่า ค่าเฉลี่ยของจุดแบ่งมีค่าสูงต่ำสลับไปมาแต่แกว่งอยู่ใกล้ค่า 0.5 เมื่อสัดส่วนของการ ไม่เกิดเหตุการณ์ที่สนใจเพิ่มขึ้น

เมื่อจำนวนตัวแปรอิสระ เท่ากับ 3 ที่ระดับความสัมพันธ์ระหว่างตัวแปรอิสระอยู่ในระดับต่ำและระดับปานกลาง ขนาดตัวอย่าง เท่ากับ 50 พบว่า ค่าเฉลี่ยของจุดแบ่งมีแนวโน้มลดลง และที่ระดับความสัมพันธ์ระหว่างตัวแปรอิสระอยู่ในระดับปานกลาง ขนาดตัวอย่าง เท่ากับ 150, 200 และระดับสูง ขนาดตัวอย่าง เท่ากับ 50 พบว่า ค่าเฉลี่ยของจุดแบ่งมีแนวโน้มเพิ่มขึ้น สำหรับสถานการณ์อื่นๆ พบว่า ค่าเฉลี่ยของ จุดแบ่งมีค่าสูงต่ำสลับไปมาแต่แกว่งอยู่ใกล้ค่า 0.5 เมื่อสัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจเพิ่มขึ้น

เมื่อจำนวนตัวแปรอิสระ เท่ากับ 4 ที่ระดับความสัมพันธ์ระหว่างตัวแปรอิสระอยู่ในระดับต่ำ ขนาดตัวอย่าง เท่ากับ 250 พบว่า ค่าเฉลี่ยของจุดแบ่งมีแนวโน้มเพิ่มขึ้น และที่ระดับความสัมพันธ์ระหว่าง ตัวแปรอิสระอยู่ในระดับสูง ขนาดตัวอย่าง เท่ากับ 200, 250 พบว่า ค่าเฉลี่ยของจุดแบ่งมีแนวโน้มลดลง สำหรับสถานการณ์อื่นๆ พบว่า ค่าเฉลี่ยของจุดแบ่งมีค่าสูงต่ำสลับไปมาแต่แกว่งอยู่ใกล้ค่า 0.5 เมื่อสัดส่วนของการ ไม่เกิดเหตุการณ์ที่สนใจเพิ่มขึ้น

เมื่อจำนวนตัวแปรอิสระ เท่ากับ 5 ที่ระดับความสัมพันธ์ระหว่างตัวแปรอิสระอยู่ในระดับต่ำและระดับปานกลาง ขนาดตัวอย่าง เท่ากับ 150, 200 พบว่า ค่าเฉลี่ยของจุดแบ่งมีแนวโน้มลดลง และที่ระดับความสัมพันธ์ระหว่างตัวแปรอิสระอยู่ในระดับต่ำ ขนาดตัวอย่าง เท่ากับ 50 และระดับสูง ขนาดตัวอย่าง เท่ากับ 100, 150 พบว่า ค่าเฉลี่ยของจุดแบ่งมีแนวโน้มเพิ่มขึ้น สำหรับสถานการณ์อื่นๆ พบว่า ค่าเฉลี่ยของจุดแบ่ง มีค่าสูงต่ำสลับไปมาแต่แกว่งอยู่ใกล้ค่า 0.5 เมื่อสัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจเพิ่มขึ้น

4. กรณีระดับความสัมพันธ์ระหว่างตัวแปรอิสระเพิ่มขึ้น

เมื่อสัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจมีค่าเท่ากับ 0.1 ที่จำนวนตัวแปรอิสระ เท่ากับ 2 ขนาดตัวอย่าง เท่ากับ 50, 100, 250 และจำนวนตัวแปรอิสระ เท่ากับ 5 ขนาดตัวอย่าง เท่ากับ 200 พบว่า ค่าเฉลี่ยของจุดแบ่งมีแนวโน้มลดลง และที่จำนวนตัวแปรอิสระ เท่ากับ 2 ขนาดตัวอย่าง เท่ากับ 150 จำนวนตัวแปรอิสระ เท่ากับ 4 ขนาดตัวอย่าง เท่ากับ 250 และจำนวนตัวแปรอิสระ เท่ากับ 5 ขนาดตัวอย่าง เท่ากับ 50, 250 พบว่า ค่าเฉลี่ยของจุด

แบ่งมีแนวโน้มเพิ่มขึ้น สำหรับสถานการณ์อื่นๆ พบว่า ค่าเฉลี่ยของจุดแบ่งมีค่าสูงต่ำ สลับไปมาแต่แกว่งอยู่ใกล้ค่า 0.5 เมื่อระดับความสัมพันธ์ระหว่างตัวแปรอิสระเพิ่มขึ้น

เมื่อสัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจมีค่าเท่ากับ 0.5 ที่จำนวนตัวแปรอิสระ เท่ากับ 2 ขนาดตัวอย่าง เท่ากับ 150 จำนวนตัวแปรอิสระ เท่ากับ 4 ขนาดตัวอย่าง เท่ากับ 50 และจำนวนตัวแปรอิสระ เท่ากับ 5 ขนาดตัวอย่าง เท่ากับ 200 พบว่า ค่าเฉลี่ยของจุดแบ่งมีแนวโน้มลดลง และที่จำนวนตัวแปรอิสระ เท่ากับ 2 ขนาดตัวอย่าง เท่ากับ 50 จำนวนตัวแปรอิสระ เท่ากับ 3 ขนาดตัวอย่าง เท่ากับ 150, 200 จำนวน ตัวแปรอิสระ เท่ากับ 4 ขนาดตัวอย่าง เท่ากับ 100, 200, 250 และจำนวนตัวแปรอิสระ เท่ากับ 5 ขนาดตัวอย่าง เท่ากับ 250 พบว่า ค่าเฉลี่ยของจุดแบ่งมีแนวโน้มเพิ่มขึ้น สำหรับสถานการณ์อื่นๆ พบว่า ค่าเฉลี่ยของจุดแบ่งมีค่าสูงต่ำสลับไปมาแต่แกว่งอยู่ใกล้ค่า 0.5 เมื่อระดับความสัมพันธ์ระหว่างตัวแปรอิสระเพิ่มขึ้น

เมื่อสัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจมีค่าเท่ากับ 0.9 ที่จำนวนตัวแปรอิสระ เท่ากับ 3 ขนาดตัวอย่าง เท่ากับ 200, 250 และจำนวนตัวแปรอิสระ เท่ากับ 4 ขนาดตัวอย่าง เท่ากับ 100, 200 พบว่า ค่าเฉลี่ยของจุดแบ่งมีแนวโน้มลดลง และที่จำนวนตัวแปรอิสระ เท่ากับ 4 ขนาดตัวอย่าง เท่ากับ 50 จำนวน ตัวแปรอิสระ เท่ากับ 5 ขนาดตัวอย่าง เท่ากับ 50, 250 พบว่า ค่าเฉลี่ยของจุดแบ่งมีแนวโน้มเพิ่มขึ้น สำหรับสถานการณ์อื่นๆ พบว่า ค่าเฉลี่ยของจุดแบ่งมีค่าสูงต่ำสลับไปมาแต่แกว่งอยู่ใกล้ค่า 0.5 เมื่อระดับความสัมพันธ์ระหว่างตัวแปรอิสระเพิ่มขึ้น

การอภิปรายผล

จากผลการวิจัยจะสามารถสรุปได้ว่า จำนวนตัวแปรอิสระ, ขนาดตัวอย่าง, สัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจ และระดับความสัมพันธ์ระหว่างตัวแปรอิสระมีผลต่อค่าของจุดแบ่งเพียงเล็กน้อย เนื่องจากค่าจุดแบ่งในเกือบทุกสถานการณ์มีค่าแกว่งในจุด 0.5 ไม่ขึ้นอยู่กับปัจจัยอื่นๆ แสดงว่าค่าจุดแบ่งที่เหมาะสมสำหรับการจำแนกกลุ่มของข้อมูลในตัวแบบถดถอยโลจิสติกแบบ 2 ประเภทโดยมีฟังก์ชันโพรบิตเป็นฟังก์ชันเชื่อมโยง คือ ค่า 0.5

ข้อเสนอแนะ

1. ในงานวิจัยครั้งนี้ได้ศึกษาปัจจัยที่มีผลต่อค่าจุดแบ่งเพียง 4 ปัจจัย คือ จำนวนตัวแปรอิสระ, ขนาดตัวอย่าง, สัดส่วนของการไม่เกิดเหตุการณ์ที่สนใจ และระดับความสัมพันธ์ระหว่างตัวแปรอิสระในการวิจัยครั้งต่อไปอาจหาปัจจัยอื่นๆ เพิ่มเติม

2. ในการวิจัยครั้งนี้ได้กำหนดให้แต่ละปัจจัยคงที่ เมื่อมีปัจจัยหนึ่งเปลี่ยนแปลง ในงานวิจัย ครั้งต่อไปอาจใช้กรณีที่ปัจจัยเปลี่ยนไปพร้อมกันหลายๆ ปัจจัย

3. เมื่อต้องการหาค่าจุดแบ่งที่เหมาะสมสำหรับการจำแนกกลุ่มของข้อมูลในตัวแบบถดถอยโลจิสติกแบบ 2 ประเภทโดยมีฟังก์ชันโพรบิตเป็นฟังก์ชันเชื่อมโยง สามารถนำค่าจุดแบ่งนี้ไปใช้ได้ตามแต่ละสถานการณ์ที่ต้องการศึกษา ภายใต้สถานการณ์ที่เหมือนกัน

เอกสารอ้างอิง

Hadjicostas, P. Maximizing proportions of correct classifications in binary logistic regression. ***Journal of Applied Statistics*** 33 (2006) : 629-640.