บทที่ 2

เอกสารและงานวิจัยที่เกี่ยวข้อง

การหาจุดแบ่งที่เหมาะสมที่สุดสำหรับการพยากรณ์การจำแนกข้อมูลไม่จัดกลุ่ม สำหรับการวิจัยครั้งนี้จะใช้หาจุดแบ่งที่เหมาะสมที่สุดในตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภท ซึ่งจุดแบ่งที่เหมาะสมที่สุดจะให้ค่าอัตราความผิดพลาดในการจำแนกกลุ่มมีค่าต่ำสุด และใช้ตัวแบบการถดถอยพหุคูณเพื่อประมาณค่าพารามิเตอร์สำหรับใช้ในการประมาณค่าจุดแบ่งที่เหมาะสมที่สุดในสถานการณ์อื่นๆ ต่อไป ซึ่งในบทนี้จะกล่าวถึงรายละเอียดเกี่ยวกับตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภท การประมาณค่าสัมประสิทธิ์การถดถอยในตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภท ด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimation) การหาจุดแบ่งโดยใช้ทฤษฎี Hadjicostas P. (2006) การหาอัตราความผิดพลาดในการจำแนกกลุ่มหรือสัดส่วนความถูกต้องในการจำแนกกลุ่ม

แนวคิดและทฤษฎี

2.1 ตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภท (Binary Logistic Regression Model)

ตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภทเป็นตัวแบบที่ศึกษาถึงความสัมพันธ์ของตัวแปร 2 กลุ่ม ได้แก่ กลุ่มตัวแปรอิสระ (X) และกลุ่มของตัวแปรตาม (Y) ซึ่งตัวแปรตามนี้เป็นตัวแปรเชิงคุณภาพมีค่าได้เพียง 2 ค่า โดยพิจารณาในรูปของความสำเร็จและความล้มเหลว โดยที่ $Y=1$ เมื่อพบความสำเร็จ และ $Y=0$ เมื่อพบความล้มเหลว ซึ่งเมื่อได้ตัวแบบความสัมพันธ์ระหว่างตัวแปรแล้วจะสามารถนำไปใช้พยากรณ์โอกาสที่แต่ละหน่วยจะอยู่ในกลุ่มใดกลุ่มหนึ่งได้

ตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภท สำหรับการพยากรณ์การจำแนกกลุ่ม เป็นดังนี้

จาก $$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+...+\beta_pX_{pi}+\varepsilon_i \qquad ;i=1,2,...,n$$

เนื่องจากตัวแปรตาม Y มีค่าได้เพียง 2 ค่า คือ 0 และ 1 ดังนั้น จึงมีการแจกแจงแบบเบอร์นูลลี (Bernoulli Distribution)

โดย $$\Pr\left(Y_i=1\middle|X_{1i}=x_{1i},X_{2i}=x_{2i},...,X_{pi}=x_{pi}\right)=\pi_i$$

$$\Pr\left(Y_i=0\middle|X_{1i}=x_{1i},X_{2i}=x_{2i},...,X_{pi}=x_{pi}\right)=1-\pi_i \quad ;0<\pi_i<1$$

Y_i มีค่าเฉลี่ยและค่าความแปรปรวน ดังนี้

$$E\left(Y_i\middle|X_{1i}=x_{1i},X_{2i}=x_{2i},...,X_{pi}=x_{pi}\right)=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+...+\beta_pX_{pi}$$

$$Var\left(Y_i \middle| X_{1i}=x_{1i}, X_{2i}=x_{2i},...,X_{pi}=x_{pi}\right)=\pi_i(1-\pi_i) \qquad ;i=1,2,...,n$$

เมื่อ π_i คือความน่าจะเป็นที่เกิดความสำเร็จของลักษณะที่สนใจศึกษา

$1-\pi_i$ คือความน่าจะเป็นที่เกิดความล้มเหลวของลักษณะที่สนใจศึกษา

เนื่องจากถ้าตัวแบบการถดถอยเชิงเส้นคือ $Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+...+\beta_pX_{pi}+\varepsilon_i$; $i=1,2,...,n$ ซึ่งมีข้อตกลงเบื้องต้นคือ $E(\varepsilon)=0$ นั่นคือ

$$E\left(Y_i \middle| X_{1i}=x_{1i}, X_{2i}=x_{2i},...,X_{pi}=x_{pi}\right)=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+...+\beta_pX_{pi} \;;i=1,2,...,n$$

ฟังก์ชันโลจิท (logit function) จะอยู่ในรูปของ

$$Logit\left[E\left(Y_i \middle| X_{1i}=x_{1i}, X_{2i}=x_{2i},...,X_{pi}=x_{pi}\right)\right]=\ln\left(\frac{E\left(Y_i \middle| X_{1i}=x_{1i}, X_{2i}=x_{2i},...,X_{pi}=x_{pi}\right)}{1-E\left(Y_i \middle| X_{1i}=x_{1i}, X_{2i}=x_{2i},...,X_{pi}=x_{pi}\right)}\right)$$

$$=\ln\left(\frac{\pi_i}{1-\pi_i}\right)$$

ซึ่งฟังก์ชันโลจิทจะทำการแปลงค่า π_i จากช่วง (0,1) เป็นค่าที่อยู่ในช่วง $(-\infty,\infty)$ โดยฟังก์ชันผกผันของฟังก์ชันโลจิท จะแปลงค่า $\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+...+\beta_pX_{pi}$ เป็นค่าที่อยู่ในช่วง [0,1] ซึ่งเรียกว่า ฟังก์ชันโลจิสติก (logistic function) และจะได้ตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภท ดังนี้

$$\ln\left(\frac{E\left(Y_i \middle| X_{1i}=x_{1i}, X_{2i}=x_{2i},...,X_{pi}=x_{pi}\right)}{1-E\left(Y_i \middle| X_{1i}=x_{1i}, X_{2i}=x_{2i},...,X_{pi}=x_{pi}\right)}\right)=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+...+\beta_pX_{pi}$$

$$\left(\frac{E\left(Y_i \middle| X_{1i}=x_{1i}, X_{2i}=x_{2i},...,X_{pi}=x_{pi}\right)}{1-E\left(Y_i \middle| X_{1i}=x_{1i}, X_{2i}=x_{2i},...,X_{pi}=x_{pi}\right)}\right)=\exp\left(\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+...+\beta_pX_{pi}\right)$$

$$E\left(Y_i \middle| X_{1i}=x_{1i}, X_{2i}=x_{2i},...,X_{pi}=x_{pi}\right)=\frac{\exp\left(\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+...+\beta_pX_{pi}\right)}{1+\exp\left(\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+...+\beta_pX_{pi}\right)}$$

$$\pi_i=\frac{\exp\left(\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+...+\beta_pX_{pi}\right)}{1+\exp\left(\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+...+\beta_pX_{pi}\right)}$$

ซึ่งตัวแบบนี้สามารถทำให้อยู่ในรูปของเชิงเส้นได้ ดังนี้

$$Logit(\pi_i)=\ln\left[\frac{E\left(Y_i \middle| X_{1i}=x_{1i}, X_{2i}=x_{2i},...,X_{pi}=x_{pi}\right)}{1-E\left(Y_i \middle| X_{1i}=x_{1i}, X_{2i}=x_{2i},...,X_{pi}=x_{pi}\right)}\right]=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+...+\beta_pX_{pi}$$

$$Y_i^*=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+...+\beta_pX_{pi}+\varepsilon_i$$

$$Y_i^*=\ln\left(\frac{E\left(Y_i \middle| X_{1i}=x_{1i}, X_{2i}=x_{2i},...,X_{pi}=x_{pi}\right)}{1-E\left(Y_i \middle| X_{1i}=x_{1i}, X_{2i}=x_{2i},...,X_{pi}=x_{pi}\right)}\right) \text{ โดย } \varepsilon_i\sim N\left(0,\sigma^2\right);i=1,2,...,n$$

2.2 ฟังก์ชันภาวะน่าจะเป็น (Likelihood function) ของข้อมูลการถดถอยโลจิสติกแบบ 2 ประเภท

เมื่อแต่ละค่าสังเกต $(x_{1i}, x_{2i}, ..., x_{pi}, y_i)$; $i = 1,2,...,n$ คือตัวแปรสุ่มแบบเบอร์นูลลี ซึ่ง

$$\Pr\left(Y_i = 1 \middle| X_{1i} = x_{1i}, X_{2i} = x_{2i}, ..., X_{pi} = x_{pi}\right) = \pi_i(x_{1i}, x_{2i}, ..., x_{pi})$$

$$\Pr\left(Y_i = 0 \middle| X_{1i} = x_{1i}, X_{2i} = x_{2i}, ..., X_{pi} = x_{pi}\right) = 1 - \pi_i(x_{1i}, x_{2i}, ..., x_{pi})$$

การแจกแจงความน่าจะเป็น คือ

$$f_i\left(Y_i = y_i \middle| X_{1i} = x_{1i}, X_{2i} = x_{2i}, ..., X_{pi} = x_{pi}\right) = \left[\pi_i(x_{1i}, ..., x_{pi})\right]^{y_i}\left[1 - \pi_i(x_{1i}, ..., x_{pi})\right]^{1-y_i}$$

เมื่อ $y_i = 0,1$; $i = 1,2,...,n$

จะได้ฟังก์ชันภาวะน่าจะเป็น คือ

$$l\left(\beta_0, \beta_1, ..., \beta_p \middle| \left\{Y_i = y_i \middle| X_{1i} = x_{1i}, X_{2i} = x_{2i}, ..., X_{pi} = x_{pi}\right\}, i = 1,2,...,n\right) = \prod_{i=1}^{n}\left[\pi_i(x_{1i}, ..., x_{pi})\right]^{y_i}\left[1 - \pi_i(x_{1i}, ..., x_{pi})\right]^{1-y_i}$$

ลอการิทึมธรรมชาติของฟังก์ชันภาวะน่าจะเป็น (log-likelihood function) คือ

$$L(\beta_0, \beta_1, ..., \beta_p) = \ln\left\{l\left(\beta_0, \beta_1, \beta_2, ..., \beta_p \middle| \left\{Y_i = y_i \middle| X_{1i} = x_{1i}, X_{2i} = x_{2i}, ..., X_{pi} = x_{pi}\right\}; i = 1,2,...,n\right)\right\}$$

$$= \ln\left\{\prod_{i=1}^{n}\left[\pi_i(x_{1i}, ..., x_{pi})\right]^{y_i}\left[1 - \pi_i(x_{1i}, ..., x_{pi})\right]^{1-y_i}\right\}$$

$$= \sum_{i=1}^{n}\left\{y_i \ln\left[\pi_i(x_{1i}, ..., x_{pi})\right] + (1 - y_i)\ln\left[1 - \pi_i(x_{1i}, ..., x_{pi})\right]\right\}$$

$$= \sum_{i=1}^{n}\left\{y_i \ln\left[\pi_i(x_{1i}, ..., x_{pi})\right] + \ln\left[1 - \pi_i(x_{1i}, ..., x_{pi})\right] - y_i \ln\left[1 - \pi_i(x_{1i}, ..., x_{pi})\right]\right\}$$

$$= \sum_{i=1}^{n}\left\{y_i\left\{\ln\left[\pi_i(x_{1i}, ..., x_{pi})\right] - \ln\left[1 - \pi_i(x_{1i}, ..., x_{pi})\right]\right\} + \ln\left[1 - \pi_i(x_{1i}, ..., x_{pi})\right]\right\}$$

$$= \sum_{i=1}^{n} y_i \ln\left(\frac{\pi_i(x_{1i}, ..., x_{pi})}{1 - \pi_i(x_{1i}, ..., x_{pi})}\right) + \sum_{i=1}^{n}\ln\left[1 - \pi_i(x_{1i}, ..., x_{pi})\right]$$

เมื่อ $$\pi_i = \frac{\exp\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi}\right)}{1 + \exp\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi}\right)}$$

$$1 - \pi_i = \frac{1}{1 + \exp\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi}\right)}$$

$$\ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi}$$

ดังนั้น ลอการิทึมธรรมชาติของฟังก์ชันภาวะน่าจะเป็น คือ

$$L(\beta_0, \beta_1, ..., \beta_p) = \sum_{i=1}^{n} y_i\left(\beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi}\right) - \sum_{i=1}^{n}\ln\left[1 + \exp\left(\beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi}\right)\right]$$

2.3 การประมาณค่าพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood Estimation)

การหาค่าตัวประมาณค่าสัมประสิทธิ์การถดถอยด้วยวิธีภาวะน่าจะเป็นสูงสุด คือต้องทำให้ L มีค่ามากที่สุดโดยทำการหาอนุพันธ์เทียบกับ β ต่างๆ เมื่อเราทราบการแจกแจงของ Y เราจะสามารถสร้างฟังก์ชันภาวะน่าจะเป็นได้ เนื่องจากการประมาณค่าไม่ได้เป็นไปตามรูปแบบ เราจึงต้องใช้วิธีการประมาณเชิงตัวเลข (จำเป็นต้องทำซ้ำเพื่อให้ได้มาซึ่งตัวประมาณภาวะน่าจะเป็นสูงสุด (MLE)) โดยในการทำซ้ำ 1 ครั้งจะได้ตัวประมาณภาวะน่าจะเป็นสูงสุด $b_0, b_1, b_2, ..., b_p$ แล้วทำการคำนวณค่าประมาณของ π_i ; $i = 1,2,...,n$ ดังนี้

$$\hat{\pi}_i = \frac{\exp\left(b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + b_p x_{pi}\right)}{1 + \exp\left(b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + b_p x_{pi}\right)} \qquad ; i = 1,2,...,n$$

เมื่อประมาณค่าพารามิเตอร์ด้วยวิธีภาวะน่าจะเป็นสูงสุดในตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภทแล้ว จะสามารถนำไปใช้ในการพยากรณ์การจำแนกกลุ่มของตัวแบบ ดังนี้

- หน่วยที่ i จะถูกจัดให้อยู่ในกลุ่มของความสำเร็จของลักษณะที่สนใจศึกษา (Y=1) ถ้า

$$\hat{\pi}_i = \frac{\exp\left(b_0 + b_1 x_{10} + b_2 x_{20} + ... + b_p x_{p0}\right)}{1 + \exp\left(b_0 + b_1 x_{10} + b_2 x_{20} + ... + b_p x_{p0}\right)} > c \qquad ; 0 \le c \le 1$$

- หน่วยที่ i จะถูกจัดให้อยู่ในกลุ่มของความล้มเหลวของลักษณะที่สนใจศึกษา (Y=0) ถ้า

$$\hat{\pi}_i = \frac{\exp\left(b_0 + b_1 x_{10} + b_2 x_{20} + ... + b_p x_{p0}\right)}{1 + \exp\left(b_0 + b_1 x_{10} + b_2 x_{20} + ... + b_p x_{p0}\right)} \le c \qquad ; 0 \le c \le 1$$

เมื่อ c คือ จุดแบ่งหรือระดับของความน่าจะเป็นที่ใช้ในการพิจารณาการจำแนกกลุ่มว่าแต่ละหน่วยจะอยู่ในกลุ่มใดระหว่างกลุ่มของความสำเร็จและกลุ่มของความล้มเหลวของลักษณะที่สนใจศึกษา

2.4 การตรวจสอบความเหมาะสมของตัวแบบ

เนื่องจากตัวแบบที่เป็นไปได้มีหลายตัวแบบ ดังนั้นจึงมีวิธีการวัดทางสถิติที่หลากหลายสำหรับใช้ในการวัดว่าตัวแบบโลจิสติกมีความสามารถในการจำแนกกลุ่มให้แต่ละหน่วยอยู่ในกลุ่มใดกลุ่มหนึ่งระหว่างกลุ่มของความสำเร็จและกลุ่มของความล้มเหลวของลักษณะที่สนใจศึกษาเหมาะสมมากน้อยเพียงใด ดังนี้

- Chi-square goodness of fit test และสถิติ Deviance
- Hosmer-Lomeshow test
- Classification table

- Receiver Operating Characteristic Curve (ROC)
- สัมประสิทธิ์การตัดสินใจสำหรับการถดถอยโลจิสติก $\left(R^2\right)$
- การตรวจสอบความถูกต้องของตัวแบบ (Model validation) ด้วยวิธีการใช้ชุดข้อมูลภายนอกหรือโดยแบ่งชุดข้อมูลออกเป็น 2 ส่วน

Classification table จะสอดคล้องกับแต่ละจุดแบ่งที่ถูกคัดเลือกซึ่งจะถูกใช้เป็นเกณฑ์สำหรับการคัดเลือกจุดแบ่งที่ทำให้อัตราความผิดพลาดในการจำแนกกลุ่มมีค่าต่ำสุด

Classification table สำหรับแต่ละจุดแบ่งแสดงดังนี้

		ค่าสังเกต		
		$y_i = 1$	$y_i = 0$	
ค่าพยากรณ์ (เกณฑ์ที่ใช้ในการตัดสินใจ คือ $\hat{\pi}_i > c$)	$\hat{y}_i = 1$	A	C	A+C
	$\hat{y}_i = 0$	B	D	B+D
		A+B	C+D	A+B+C+D

โดย "A" คือ จำนวนของค่าสังเกตที่ถูกจัดให้อยู่ในกลุ่มของความสำเร็จและค่าสังเกตที่แท้จริงอยู่ในกลุ่มของความสำเร็จด้วย

"B" คือ จำนวนของค่าสังเกตที่ถูกจัดให้อยู่ในกลุ่มของความล้มเหลวในขณะที่ค่าสังเกตที่แท้จริงอยู่ในกลุ่มของความสำเร็จ

"C" คือ จำนวนของค่าสังเกตที่ถูกจัดให้อยู่ในกลุ่มของความสำเร็จในขณะที่ค่าสังเกตที่แท้จริงอยู่ในกลุ่มของความล้มเหลว

"D" คือ จำนวนของค่าสังเกตที่ถูกจัดให้อยู่ในกลุ่มของความล้มเหลวและค่าสังเกตที่แท้จริงอยู่ในกลุ่มของความล้มเหลวด้วย

สำหรับแต่ละจุดแบ่ง c การคัดเลือกจุดแบ่งจะต้องทำให้อัตราความผิดพลาดในการจำแนกกลุ่มมีค่าต่ำสุดหรืออัตราความถูกต้องในการจำแนกกลุ่มมีค่าสูงสุด โดยแสดงดังนี้

อัตราความผิดพลาดในการจำแนกกลุ่ม (Classification Error Rate: CER)

$$CER = \frac{B + C}{A + B + C + D}$$

ถ้า CER มีค่าต่ำกว่าแสดงว่ามีความเหมาะสมมากกว่า

สัดส่วนความถูกต้องในการจำแนกกลุ่ม ($p(c)$)

$$p(c) = \frac{A + D}{A + B + C + D} = 1 - CER$$

ถ้า $p(c)$ มีค่ามากแสดงว่ามีความเหมาะสมมากกว่า

2.5 **สถิติ** Kaiser-Meyer-Olkin (KMO)

Kaiser (1970) ได้เสนอสถิติ KMO ในการศึกษานี้ สถิติ KMO เป็นสถิติที่ใช้วัดระดับความสัมพันธ์ระหว่างตัวแปรอิสระ

$$0 \le KMO = \frac{\sum_{i=1}^{p}\sum_{i<j=1}^{p} r_{ij}^2}{\sum_{i=1}^{p}\sum_{i<j=1}^{p} r_{ij}^2 + \sum_{i=1}^{p}\sum_{i<j=1}^{p} a_{ij}^2} \le 1$$

โดยที่ r_{ij} คือสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปรอิสระตัวที่ i และตัวแปรอิสระตัวที่ j

a_{ij} คือสัมประสิทธิ์สหสัมพันธ์บางส่วนระหว่างตัวแปรอิสระตัวที่ i และตัวแปรอิสระตัวที่ j สำหรับ $i < j = 1,2,...,p$

ระดับความสัมพันธ์ระหว่างตัวแปรอิสระแบ่งเป็น 5 ระดับ ดังนี้

- ตัวแปรอิสระไม่มีความสัมพันธ์กัน เมื่อ $KMO = 0$
- ความสัมพันธ์ระหว่างตัวแปรอิสระอยู่ในระดับต่ำ เมื่อ $0 < KMO \le 0.50$
- ความสัมพันธ์ระหว่างตัวแปรอิสระอยู่ในระดับปานกลาง เมื่อ $0.50 < KMO \le 0.75$
- ความสัมพันธ์ระหว่างตัวแปรอิสระอยู่ในระดับสูง เมื่อ $0.75 < KMO < 1.00$
- ตัวแปรอิสระมีความสัมพันธ์กันอย่างสมบูรณ์ เมื่อ $KMO = 1$

2.6 **ช่วงความเชื่อมั่น** (Confidence Interval)

การประมาณค่าแบบช่วงหรือช่วงความเชื่อมั่นเป็นการประมาณค่าพารามิเตอร์ของประชากรในรูปแบบช่วงโดยใช้ข้อมูลตัวอย่าง การประมาณแบบช่วงนั้นจะบอกถึงค่าต่ำสุดและค่าสูงสุดของพารามิเตอร์ที่เป็นไปได้

ระดับของค่าความเชื่อมั่นที่ใช้ในการสร้างช่วงความเชื่อมั่นนั้นจะกำหนดเป็นค่าควบคู่กับระดับนัยสำคัญ นั่นคือ $1-\alpha$ ที่เรียกว่าช่วงความเชื่อมั่น $(1-\alpha)100\%$ จะได้ว่า

$$P(L < \theta < U) = 1-\alpha$$

เรียก L ว่าขีดจำกัดความเชื่อมั่นล่าง (lower confidence limit)

U ว่าขีดจำกัดความเชื่อมั่นบน (upper confidence limit)

ในงานวิจัยนี้กำหนดให้ L หาจากค่าที่ตำแหน่งเปอร์เซ็นต์ไทล์ที่ $100(\alpha/2)$

U หาจากค่าที่ตำแหน่งเปอร์เซ็นต์ไทล์ที่ $100(1-\alpha/2)$

2.7 เปอร์เซ็นต์ไทล์ (Percentile)

เป็นค่าที่แบ่งข้อมูลออกเป็น 100 ส่วนเท่าๆ กัน เมื่อข้อมูลถูกเรียงจากน้อยไปหามาก เนื่องจากค่าที่แบ่งจำนวนข้อมูลออกเป็น 100 ส่วนเท่าๆ กัน มีอยู่ 99 ค่า ดังนั้นเราจึงตั้งชื่อแต่ละค่าว่า เปอร์เซ็นต์ไทล์ที่หนึ่ง (P_1) เปอร์เซ็นต์ไทล์ที่สอง (P_2) .... และเปอร์เซ็นต์ไทล์ที่เก้าสิบเก้า (P_{99}) ตามลำดับ

การหาเปอร์เซ็นต์ไทล์ คือต้องหาตำแหน่งของเปอร์เซ็นต์ไทล์ก่อน

ให้ N เป็นจำนวนข้อมูลทั้งหมด ตำแหน่งต่างๆ ของเปอร์เซ็นต์ไทล์หาได้ดังนี้

P_1 อยู่ในตำแหน่งที่ คือ $\frac{1(N+1)}{100}$

P_2 อยู่ในตำแหน่งที่ คือ $\frac{2(N+1)}{100}$

$\vdots$

P_{99} อยู่ในตำแหน่งที่ คือ $\frac{99(N+1)}{100}$

โดยทั่วไป ตำแหน่งของเปอร์เซ็นต์ไทล์ที่ r คือ

P_r อยู่ในตำแหน่งที่ คือ $\frac{r(N+1)}{100}$

2.8 วิธีการหาจุดแบ่งที่เหมาะสมที่สุดสำหรับตัวแบบการถดถอยโลจิสติกแบบ 2 ประเภท โดยทฤษฎีของ Hadjicostas P. (2006)

วิธีการนี้ Hadjicostas P. (2006) ได้เสนอไว้โดยใช้ผลลัพธ์ทางคณิตศาสตร์ที่ง่ายแต่มีความถูกต้องแม่นยำในการหาจุดแบ่งที่ทำให้สัดส่วนความถูกต้องในการจำแนกกลุ่มมีค่าสูงสุด ดังนี้

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

ให้ $A_i = [0, \hat{\pi}_1)$ 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 $\bigcup_{i=0}^{n} A_i = [0,1]$

ให้ $p(c)$ คือสัดส่วนความถูกต้องในการจำแนกกลุ่มที่จุด c ที่ซึ่ง $c \in [0,1]$

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

บทตั้ง สำหรับ $i \in \{0,1,2,...,n\}$ ใดๆ และ $c \in A_i$

$$N(c) = \sum_{j=1}^{M(i)} (1 - y_j) + \sum_{j=M(i)+1}^{n} y_j \qquad .....(1)$$

พิสูจน์ บทตั้งนี้มาจากความจริงที่ว่า $N(c)$ คือผลรวมของจำนวนของค่าศูนย์(ความล้มเหลว)ใน $(y_1, y_2, ..., y_{M(i)})$ และจำนวนของค่าหนึ่ง (ความสำเร็จ) ใน $(y_{M(i)+1}, ..., y_n)$ (ถ้า $M(i) = 0$ แล้ว $(y_1, y_2, ..., y_{M(i)})$ จะไม่มี เช่นเดียวกับ ถ้า $M(i) = n$ แล้ว $(y_{M(i)+1}, ..., y_n)$ จะไม่มี)

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

พิสูจน์ ให้ e_i เป็นด้านขวาของสมการ (1) นั่นคือ $e_i = \sum_{j=1}^{M(i)} (1 - y_j) + \sum_{j=M(i)+1}^{n} y_j$ โดยการใช้เอกลักษณ์ $1 - 2y_j = (-1)^{y_j}$ จะได้ว่า $e_i = \sum_{j=1}^{n} y_j + a_i$ สำหรับ $i \in \{0,1,2,...,n\}$

จากบทตั้ง สำหรับ $c_0 \in [0,1]$ ใดๆ จะได้ว่า

$$c_0 \in C_0 \Leftrightarrow N(c_0) = \max_{c \in [0,1]} N(c) = \max_{0 \le i \le n} \max_{c \in A_i} N(c) = \max_{0 \le i \le n} e_i = \sum_{j=1}^{n} y_j + \max_{0 \le i \le n} a_i$$

(i) ให้ $c_0 \in \bigcup_{i \in I_0} A_i$ หา $i_0 \in I_0$ ที่ซึ่ง $c_0 \in A_{i0}$

ดังนั้น $N(c_0) = e_{i0} = \sum_{j=1}^{n} y_j + a_{i0} = \sum_{j=1}^{n} y_j + \max_{0 \le i \le n} a_i$

นั่นคือ $c_0 \in C_0$ ดังนั้น $\bigcup_{i \in I_0} A_i \subseteq C_0$

(ii) ให้ $c_0 \in C_0$ เนื่องจาก $\bigcup_{i=0}^{n} A_i = [0,1]$ นั่นคือ $i_1 \in \{0,1,2,...,n\}$ ที่ซึ่ง $c_0 \in A_{i1}$

แล้ว $N(c_0) = e_{i_1} = \sum_{j=1}^{n} y_j + a_{i_1}$ แต่ $N(c_0) = \sum_{j=1}^{n} y_j + \max_{0 \le i \le n} a_i$

ดังนั้น $a_{i_1} = \max_{0 \le i \le n} a_i$ แสดงว่า $i_1 \in I_0$

ดังนั้น $c_0 \in \bigcup_{i \in I_0} A_i$ และ $C_0 \subseteq \bigcup_{i \in I_0} A_i$

โดยมีรายละเอียดและวิธีการดังนี้

1. เรียงอันดับค่า $\hat{\pi}_i$ จากน้อยไปหามาก $\hat{\pi}_1 < \hat{\pi}_2 < ... < \hat{\pi}_n$ โดยถ้า $\hat{\pi}_i$ คือจุดแบ่งแล้วจะพยากรณ์ให้เป็นกลุ่มของความล้มเหลวของลักษณะที่สนใจศึกษา (Y=0)
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 $i = 0,1,2,...,n$ หาค่า $a_i = \sum_{k=1}^{M(i)} (-1)^{y_k}$ โดยแบ่งเป็น 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 $j \in \{0,1,2,...,n\}$ ที่ซึ่ง $a_j = \max_{0 \le i \le n} a_i$
5. หา C_0 ซึ่งเป็นเซตของ c_0 ทั้งหมด โดย $c_0 \in [0,1]$ ที่ซึ่ง $p(c_0) = \max_{c \in [0,1]} p(c)$ แล้ว $C_0 = \bigcup_{i \in I_0} A_i$