บทที่ 1

บทนำ

ความสำคัญและความเป็นมาของปัญหา

ในปัจจุบันการนำความรู้ทางด้านสถิติไปประยุกต์ใช้ในงานสาขาต่างๆ มีมากขึ้น ทั้งนี้เนื่องจากวิธีทางสถิติเป็นระเบียบวิธีดำเนินการอย่างมีระบบภายใต้เหตุและผล การนำวิธีการทางสถิติไปประยุกต์ใช้ เพื่อการวิเคราะห์ วิจัย และการนำเสนอ จำเป็นต้องอาศัยวิธีทางสถิติที่เหมาะสมกับวัตถุประสงค์ของงาน และเหมาะสมกับข้อมูลของสาขานั้นๆ ข้อมูลอาจเป็นข้อมูลเชิงปริมาณ (Quantitative Data) หรือข้อมูลเชิงคุณภาพ (Qualitative Data)

การวิเคราะห์การถดถอยเชิงเส้น เป็นการวิเคราะห์ที่มักมีผู้นิยมนำไปใช้กับข้อมูลเชิงปริมาณตั้งแต่ 2 ชนิดขึ้นไป ซึ่งเรียกว่า มีตัวแปรตั้งแต่ 2 ตัวขึ้นไป ใช้ตัวแปรหนึ่งตัวหรือมากกว่าเป็นตัวแปรอธิบาย ซึ่งเป็นปัจจัยที่ทำให้เกิดตัวแปรตามซึ่งมีตัวเดียว ความสัมพันธ์ของตัวแปรตามและตัวแปรอธิบายกำหนดไว้ในรูปสมการเชิงเส้น เรียกว่า สมการถดถอยเชิงเส้น (Linear Regression Model) การวิเคราะห์การถดถอยเชิงเส้นนี้ มีวัตถุประสงค์เพื่อคาดคะเนแนวโน้มของปริมาณชนิดหนึ่ง เมื่อปริมาณอื่นๆ ที่เป็นปัจจัยเปลี่ยนไป หรือเพื่อพยากรณ์ค่าตัวแปรตามเมื่อกำหนดค่าตัวแปรอธิบายนั้นๆ

ในวงการแพทย์ ทางชีวสถิติ ทางการศึกษา ทางเศรษฐศาสตร์ และทางสังคม มักพบข้อมูลเชิงคุณภาพเข้ามาเกี่ยวข้องด้วย เช่น อาการเกิดโรค (เกิด,ไม่เกิด) , เพศ (หญิง,ชาย) , ผลการสอบเข้ามหาวิทยาลัย (ได้,ไม่ได้) , การตัดสินใจซื้อสินค้า(ซื้อ,ไม่ซื้อ) เป็นต้น ข้อมูลเชิงคุณภาพในลักษณะเช่นนี้มีค่าที่สามารถวัดได้เพียง 2 ค่า

ถ้าข้อมูลที่สนใจจะศึกษาประกอบด้วยตัวแปรตามที่เป็นตัวแปรเชิงคุณภาพที่มี 2 ค่า (Dichotomous Dependent Variable) กับตัวแปรอธิบายหนึ่งตัว หรือ มากกว่าหนึ่ง เช่น การศึกษาเรื่อง การมีชีวิตรอดของคนไข้ผู้ได้รับการผ่าตัดเนื้อร้าย ตัวแปรที่ศึกษาคือ

$$y = \begin{cases} 1 & \text{ถ้าคนไข้ยังมีชีวิตรอดหลังการผ่าตัดแล้ว 5 ปี} \\ 0 & \text{อื่นๆ} \end{cases}$$

และ x_1 = อายุคนไข้เมื่อได้รับการผ่าตัด

x_2 = จำนวนปีที่คนไข้เข้ารับการรักษา

x_3 = จำนวนก้อนเนื้อร้ายที่ตรวจพบ

มีคำถามว่า จะนำการวิเคราะห์การถดถอยเชิงเส้นซึ่งสะดวกและง่ายต่อการแปลความหมาย มาประยุกต์ใช้กับข้อมูลตัวแปรตาม 2 ค่านี้ได้หรือไม่ คำตอบ คือ ไม่เป็นการเหมาะสมที่จะแสดงความสัมพันธ์ของตัวแปรตาม y กับตัวแปรอธิบาย x_1, x_2 และ x_3 ด้วยสมการถดถอยเชิงเส้น โดยมีเหตุผลที่สำคัญประการหนึ่ง คือ ตัวแปรตาม y ไม่มีการแจกแจงปกติตรงตามข้อสมมุติพื้นฐานของสมการถดถอยเชิงเส้น

วิธีหนึ่งที่จะใช้วิเคราะห์ความสัมพันธ์ของข้อมูลประเภทนี้ คือ *Cox* (1970) เสนอการสร้างสมการถดถอยโลจิสติคเชิงเส้น เพื่อให้เข้าใจง่ายเกี่ยวกับที่มาของสมการถดถอยโลจิสติคเชิงเส้น ขอให้ดู ตารางที่ 1.1 ในภาคผนวก ก ซึ่งแสดงข้อมูลเรื่องอายุกับการเป็นโรคหัวใจของคน 100 คน นำข้อมูลใน ตารางที่ 1.1 ในภาคผนวก ก มาจัดกลุ่มโดยจำแนกข้อมูลตามกลุ่มอายุ ดังแสดงในตารางที่ 1.2

ตารางที่ 1.2 แสดงสัดส่วนการเป็นโรคหัวใจจำแนกตามกลุ่มอายุ

กลุ่มอายุ (ปี)	จำนวนคน (n)	โรคหัวใจ (CHD) ไม่เป็น (y=0)	เป็น(y=1)	ค่าสัดส่วนตัวอย่าง $\left(p=\frac{y=1}{n}\right)$
20-29	10	9	1	0.10
30-34	15	13	2	0.13
35-39	12	9	3	0.25
40-44	15	10	5	0.33
45-49	13	7	6	0.46
50-54	8	3	5	0.63
55-59	17	4	13	0.76
60-69	10	2	8	0.80
รวม	100	57	43	0.43

ข้อมูลนี้เป็นข้อมูลประเภท 2 ค่าที่จัดกลุ่มแล้ว(group binary data) ให้ตัวแปรอธิบาย คือกึ่งกลางอายุ แทนด้วย x ตัวแปรตาม คือโรคหัวใจ แทนด้วย $y = 1$ ถ้าเป็น และ $y = 0$ ถ้าไม่เป็น ตัวแปรที่ปรากฏใหม่ คือ สัดส่วนของการเป็นโรคหัวใจของแต่ละกลุ่ม สัดส่วนของค่าสังเกตนี้คือ $\tilde{p}_i = \frac{y_i}{n_i}$

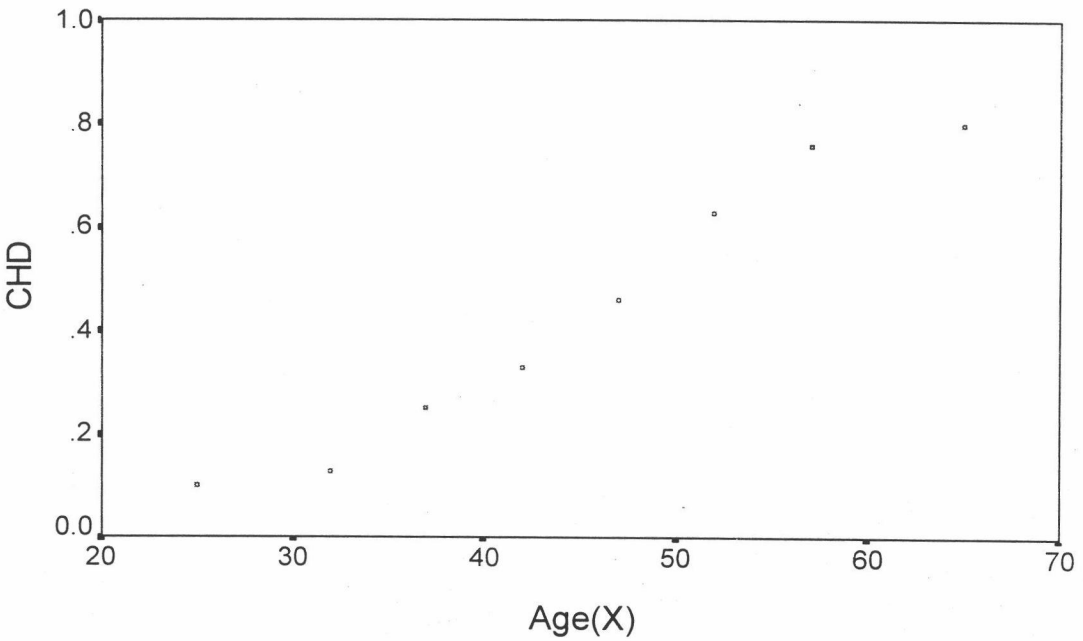
เมื่อ n_i เป็นจำนวนคนในกลุ่มที่ i

y_i เป็นความถี่ของคนเป็นโรคหัวใจในกลุ่มที่ i

ตัวแปรตาม y เป็นข้อมูล 2 ค่า ที่จัดกลุ่มแล้วมีการแจกแจงทวินาม (Binomial Distribution) พารามิเตอร์ n_i และ $\tilde{p}_i$ สำหรับตัวแปรตาม y ในข้อมูล 2 ค่า ที่ยังไม่ได้จัดกลุ่มมีการแจกแจงทวินาม พารามิเตอร์ 1 และ p_i (หรือที่เรียกว่ามีการแจกแจงแบร์นูลลี (Bernoulli Distribution) p_i แปรเปลี่ยนไปแต่ละคน) แผนภาพจุดตัวอย่างของข้อมูล 2 ค่า ที่ยังไม่ได้จัดกลุ่มไม่ได้แสดงธรรมชาติของความสัมพันธ์ของอายุ (x) กับโรคหัวใจ (y) มากนัก เห็นเพียงแต่กลุ่มของจุด 2 กลุ่ม กลุ่มหนึ่งที่ y=0 และอีกกลุ่มหนึ่งที่ y=1 (ดังแสดงในรูปที่ 1.1 ในภาคผนวก ข) ซึ่งแสดงว่าสำหรับ x ทุกระดับ y มีความผันแปรมาก ทำให้ยากต่อการให้ความหมายต่อความสัมพันธ์ระหว่างอายุ (x) กับโรคหัวใจ (y)

ถ้าพิจารณาดูความสัมพันธ์ของอายุ (x) กับสัดส่วนของค่าสังเกต ($\tilde{p}$) ที่เรียกว่าค่าเฉลี่ยของการเป็นโรคหัวใจของแต่ละกลุ่ม เมื่อพิจารณาดูกราฟเส้นโค้งจะเห็นธรรมชาติของความสัมพันธ์ ได้มากขึ้น ได้แก่ อายุมากขึ้นสัดส่วนการเป็นโรคหัวใจสูงขึ้น กราฟเส้นโค้งมีรูปเป็น s หรือที่เรียกว่าโค้งซิกมอยด์ (Sigmoid Curve) และความสัมพันธ์มีลักษณะใกล้การถดถอยเชิงเส้น

รูปที่ 1.2 แสดงความสัมพันธ์ระหว่างอายุ (x) กับสัดส่วนของการเป็นโรคหัวใจ (CHD)



ประเด็นสำคัญที่ความสัมพันธ์ระหว่างสัดส่วน $\widetilde{p}$ กับอายุ x แตกต่างจากการถดถอยเชิงเส้น $\mathrm{E}[y \setminus x] = \beta_0 + \beta_1 x$ ตรงที่ $\mathrm{E}[y \setminus x]$ ที่จะเป็นไปได้มีค่าอยู่ในช่วง $(-\infty, \infty)$ แต่สำหรับข้อมูลประเภท 2 ค่านี้ สัดส่วน p หรือที่เรียกว่าค่าเฉลี่ยของ $y \setminus x$ มีค่าอยู่ระหว่าง $(0,1)$ ถ้าจะสร้างความสัมพันธ์ในรูป $p = \beta_0 + \beta_1 x$ และหาค่าพารามิเตอร์ β_0 และ β_1 โดยวิธีกำลังสองน้อยสุด (Method of Least Squares) ย่อมจะไม่ถูกต้องตามสมมุติฐานพื้นฐานของการถดถอยเชิงเส้น เพราะความแปรปรวนของสัดส่วนค่าสังเกต $\widetilde{p}_i$ ไม่คงที่ แต่ขึ้นอยู่กับสัดส่วนจริง p_i โดยที่ y_i (y=1) มีการแจกแจงทวินาม พารามิเตอร์ n_i และ p_i

$$\text{var}(y_i) = n_i\, p_i (1 - p_i) \quad \text{และ} \quad \text{var}(\widetilde{p}_i) = \text{var}\left(\frac{y_i}{n_i}\right)$$

$$= \frac{1}{n_i^2}\, \text{var}(y_i)$$

$$= \frac{p_i(1-p_i)}{n_i}$$

และค่าประมาณสัดส่วน $\hat{p}_i$ ที่จะคำนวณได้จากสมการถดถอย $\hat{p} = \hat{\beta}_0 + \hat{\beta}_1 x$ ก็ไม่อาจรับประกันได้ว่า มีค่าอยู่ในช่วง $(0,1)$ การจะนำสมการถดถอยไปใช้เพื่อการพยากรณ์ย่อมผิดพลาด

เมื่อต้องการประมาณค่าสัดส่วน p_i (ซึ่งเป็นความน่าจะเป็นอย่างมีเงื่อนไขของ $y=1$ ณ ระดับของ x) จากความสัมพันธ์ของตัวแปรตาม y กับตัวแปรอธิบาย x จึงจำเป็นต้องมีฟังก์ชันการแปลง เพื่อแปลง p_i ให้มีคุณสมบัติสอดคล้องตามสมมุติฐานพื้นฐานของการถดถอยเชิงเส้น $Cox(1970)$ เสนอการแปลงโลจิสติค (Logistic Transformation)

โดยกำหนดฟังก์โลจิต, $\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right)$

$$= \beta_0 + \beta_1 x$$

ซึ่งจะได้ $$p_i = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

$$= \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

เรียก p_i ว่าฟังก์ชันโลจิสติค โดเมน $-\infty < x < \infty$ และเรนจ์ $0 < p_i < 1$ กราฟมีรูปเป็นตัว s เรียกสุมการ $\text{logit}(p_i) = \beta_0 + \beta_1 x$ ว่าสมการถดถอยโลจิสติค p_i และ x เป็นตัวแปรที่ทราบค่า β_0 และ β_1 เป็นพารามิเตอร์ที่ไม่ทราบค่า

ตามตัวอย่างข้างต้น เมื่อ p เป็นความน่าจะเป็นของการเป็นโรคหัวใจ ณ ระดับอายุ x ความสัมพันธ์ระหว่าง p กับ x แสดงด้วยสมการถดถอยโลจิสติค

$$\text{logit}(p_i) = \beta_0 + \beta_1 x$$

ค่าประมาณของ p คือ $\hat{p} = \frac{1}{1+e^{-\hat{\beta}_0+\hat{\beta}_1 x}}$ โดยที่ $\hat{\beta}_0$ และ $\hat{\beta}_1$ คือ ค่าประมาณของ β_0 และ β_1 ตามลำดับ

เมื่อประมาณค่าพารามิเตอร์ β_0 และ β_1 ได้แล้ว จะหา ค่า p คือ ความน่าจะเป็นแบบมีเงื่อนไขของ $y=1$ ณ สถานการณ์ของตัวแปรอธิบาย x ได้ ปัญหาสำคัญคือ จะประมาณค่าพารามิเตอร์ β_0 และ β_1 ได้อย่างไร

ในปัจจุบันสมการถดถอยโลจิสติคน่าจะมีความสำคัญมากขึ้นในวงการต่างๆ เช่น ในวงการธุรกิจซึ่งมักไม่เปิดเผยรายละเอียดของข้อมูลและมักพอใจให้ข้อมูลเพียงตอบว่าใช่หรือไม่ใช่ ถ้าผู้วิจัยมีความรู้เรื่องสมการถดถอยโลจิสติค มีความรู้เรื่องการออกแบบสอบถามชนิดไม่รบกวนผู้ให้คำตอบ โดยให้คำตอบเพียงว่า ใช่หรือไม่ใช่ ก็สามารถได้ข้อมูลประเภท

2 ค่าที่ละเอียดสอดคล้องกับปัญหาที่สนใจศึกษา

ในการวิจัยครั้งนี้ต้องการศึกษาสมการถดถอยโลจิสติค ศึกษาวิธีการประมาณค่าพารามิเตอร์ในสมการถดถอยโลจิสติคด้วยวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood) และวิธี ฟังก์ชันจำแนกประเภท (Discriminant Function) โดยจะนำวิธีกำลังสองน้อยสุดถ่วงน้ำหนัก (Weighted Least Squares) มาเปรียบเทียบด้วย ในการวิจัยครั้งนี้จะนำตัวแปรที่มีการแจกแจงแบบต่างๆ มาเป็นตัวแปรอธิบาย

วัตถุประสงค์ของการวิจัย

1. เพื่อศึกษาวิธีการประมาณค่าพารามิเตอร์ในสมการถดถอยโลจิสติคสำหรับข้อมูลประเภท 2 ค่า ที่มีตัวแปรอธิบาย 1 ตัว และ 2 ตัว ด้วยวิธี
   ก) ภาวะน่าจะเป็นสูงสุด (Maximum Likelihood)
   ข) ฟังก์ชันจำแนกประเภท (Discriminant Function)
   ค) กำลังสองน้อยสุดถ่วงน้ำหนัก (Weighted Least Squares)
2. เพื่อเปรียบเทียบผลของวิธีการประมาณค่าพารามิเตอร์ต่างๆ

สมมุติฐานของการวิจัย

การประมาณค่าพารามิเตอร์ในการสมการถดถอยโลจิสติคด้วยวิธีภาวะน่าจะเป็นสูงสุดมีประสิทธิภาพดีกว่าวิธีฟังก์ชันจำแนกประเภท และ วิธีกำลังสองน้อยสุดถ่วงน้ำหนัก

ขอบเขตของการวิจัย

1. ตัวแปรตาม (y) เป็นข้อมูลเชิงคุณภาพที่มี 2 ค่า คือ 1 และ 0 โดยในแต่ละขนาดตัวอย่างจะกำหนดสัดส่วนระหว่าง 1 และ 0 จำนวน 10 สัดส่วน ดังนี้ 0.50:0.50 , 0.55:0.45 , 0.60:0.40 , 0.65:0.35 , 0.7:0.3 , 0.75:0.25 , 0.80:0.20 , 0.85:0.15 , 0.90:0.10 และ 0.95:0.05

2. ตัวแปรอธิบาย (x) เป็นข้อมูลเชิงปริมาณที่มีการแจกแจง ดังนี้

2.1 กรณีตัวแปรอธิบาย 1 ตัว

2.1.1 การแจกแจงแบบปกติ (Normal Distribution)

ฟังก์ชันความหนาแน่นอยู่ในรูปของ

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

เมื่อ $\mathrm{E}[x] = \mu$

$\mathrm{V}[x] = \sigma^2$

ในการวิจัยครั้งนี้จะศึกษาที่ $\mu = 2$, $\sigma^2 = 0.25$, 1, 4

2.1.2 การแจกแจงแบบชี้กำลัง (Exponential Distribution)

ฟังก์ชันความหนาแน่นอยู่ในรูปของ

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}} & ; \; x > o \\ 0 & ; \; \text{อื่นๆ} \end{cases}$$

เมื่อ $\mathrm{E}[x] = \theta$

$\mathrm{V}[x] = \theta^2$

ในการวิจัยครั้งนี้จะศึกษาที่ θ = 0.5 , 1 , 2

2.1.3 การแจกแจงแบบไวบูลล์ (Weibull Distribution)

ฟังก์ชันความหนาแน่นอยู่ในรูปของ

$$f(x) = \begin{cases} \frac{\alpha x^{\alpha} \exp\left(\frac{x}{\beta}\right)^{\alpha}}{\beta^{\alpha}} & ; \; x>0 \, , \, \alpha >0 \, , \, \beta > 0 \\ 0 & ; \; \text{อื่นๆ} \end{cases}$$

เมื่อ $E[x] = \beta\Gamma\left[1+\frac{1}{\alpha}\right]$

$V[x] = \beta^2\left[\Gamma\left(1+\frac{2}{\alpha}\right)-\Gamma^2\left(1+\frac{1}{\alpha}\right)\right]$

ในการวิจัยครั้งนี้จะศึกษาที่ $\alpha = 2$, $\beta = 0.5, 1, 2$

2.2 กรณีตัวแปรอธิบาย 2 ตัว

2.2.1 x_1 มีการแจกแจงปกติ , x_2 มีการแจกแจงแบบชี้กำลัง

2.2.2 x_1 มีการแจกแจงปกติ , x_2 มีการแจกแจงแบบไวบูลล์

2.2.3 x_1 มีการแจกแจงแบบชี้กำลัง , x_2 มีการแจกแจงแบบไวบูลล์

3. จำนวนตัวแปรอธิบาย 1 ตัวแปร และ 2 ตัวแปร

4. ขนาดตัวอย่าง N=20 , 40 , 60 และ 80

เกณฑ์การตัดสินใจ

เกณฑ์ในการตัดสินใจว่าวิธีประมาณค่าพารามิเตอร์วิธีใดจะมีความถูกต้องมากที่สุด จะพิจารณาจาก เกณฑ์ค่ารากที่สองของค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Square Root Mean Squares Error (RMSE)) และเกณฑ์ที่ใช้ประกอบการพิจารณาอีกเกณฑ์หนึ่ง คือ Deviance, D ค่า Deviance มีค่าขึ้นอยู่กับค่าความน่าจะเป็นที่ประมาณได้เท่านั้น ซึ่งมีสูตรดังนี้

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(\pi(\mathbf{x}_i)-\hat{\pi}(\mathbf{x}_i)\right)^2}{\text{N - m}}}$$

และ $$D = -2\sum_{i=1}^{N}\left\{\hat{\pi}(\mathbf{x}_i)\log it(\hat{\pi}(\mathbf{x}_i)) + \ln(1-\hat{\pi}(\mathbf{x}_i))\right\}$$

เมื่อ $\pi(\mathbf{x}_i)$ หมายถึงค่าสังเกตที่ i

$\hat{\pi}(\mathbf{x}_i)$ หมายถึงค่าพยากรณ์ที่ i

N หมายถึงขนาดตัวอย่าง

m หมายถึงจำนวนพารามิเตอร์

ประโยชน์ของการวิจัย

1. เพื่อเป็นแนวทางในการตัดสินใจว่าควรใช้วิธีการใดในการประมาณค่าพารามิเตอร์จึงจะทำให้ผลการประมาณมีความผิดพลาดน้อยที่สุด
2. เพื่อเป็นแนวทางในการศึกษาเพื่อเปรียบเทียบกับวิธีการทางสถิติอื่นๆ ที่เกี่ยวข้องต่อไป
3. เพื่อเป็นแนวทางให้ผู้สนใจนำไปประยุกต์ใช้กับข้อมูลจริง