บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

การจำแนก (classification) หรือการพยากรณ์ (forecasting) ว่าเหตุการณ์ใดจะเกิดขึ้นหรือไม่ เช่น ในการรักษาโรค ผู้ป่วยแต่ละรายจะมีอาการดีขึ้น เลวร้ายลง หรือมีโอกาสเสียชีวิตหรือไม่ ธนาคารต้องการปล่อยกู้ด้วยวงเงินหนึ่งแก่ลูกหนี้ โอกาสที่จะเป็นลูกหนี้ชั้นดี ลูกหนี้ธรรมดา หรือลูกหนี้ชั้นเลวเป็นอย่างไร เมื่อมีการออกสินค้าใหม่ผู้บริโภคจะมีการตอบรับอย่างไร ไม่ชอบ เฉยๆ หรือว่าชื่นชอบ ปัญหาต่างๆเหล่านี้ทำให้ความต้องการเกี่ยวกับการพยากรณ์เพื่อจำแนกกลุ่มตัวอย่างในปัจจุบันมีมากและมีบทบาทสำคัญในการตัดสินใจด้วย ดังนั้นเครื่องมือที่ใช้จำแนกกลุ่มที่อาจจะเกิดขึ้นจำเป็นต้องมีประสิทธิภาพ และให้ผลอย่างถูกต้อง

การเลือกเครื่องมือทางสถิติเพื่อการพยากรณ์นั้นต้องเลือกให้เหมาะสมกับงานและข้อมูลที่ต้องการวิเคราะห์ ข้อมูลอาจเป็นข้อมูลเชิงปริมาณ (Quantitative Data) หรือ ข้อมูลเชิงคุณภาพ (Qualitative Data) ในปัจจุบันการวิเคราะห์เพื่อการพยากรณ์หรือการจัดกลุ่มนั้นใช้อย่างกว้างขวางทั้งในวงการแพทย์ ชีวสถิติ เศรษฐศาสตร์ ศึกษาศาสตร์ และทางสังคมศาสตร์ มักพบข้อมูลเชิงคุณภาพเข้ามาเกี่ยวข้อง เช่น การเป็นโรค (รุนแรงมาก, ปานกลาง, ไม่รุนแรง), เพศ (ชาย, หญิง), ผลการสอบ (ผ่าน, ตก) เป็นต้น

กลุ่มอันดับ (Ordered Group) คือ กลุ่มของข้อมูลที่ถูกวัดด้วย สเกลอันดับ ซึ่งเมื่อข้อมูลเหล่านี้เป็นข้อมูลของตัวแปรอิสระในการจำแนกกลุ่มจะทำให้ได้เป็นกลุ่มลำดับที่สามารถบอกอันดับและความสำคัญได้ เช่น ตัวอย่างการจำแนกโอกาสการเกิดโรคของผู้ป่วย ซึ่งอาจจะมีกลุ่มที่สามารถจำแนกได้ดังนี้ ไม่เกิดโรค, มีโอกาสเกิดโรคต่ำ, มีโอกาสเกิดโรคปานกลาง ,มีโอกาสเกิดโรคสูง เราย่อมสามารถบอกลำดับความสำคัญของกลุ่มที่จะจำแนกได้ว่าถ้าเรียงตามอันดับความรุนแรงแล้ว ไม่เกิดโรคย่อมมีความรุนแรงต่ำที่สุด มีโอกาสเกิดโรคต่ำ, มีโอกาสเกิดโรคปานกลาง และมีโอกาสเกิดโรคสูง จะมีความรุนแรงเพิ่มขึ้นตามลำดับ หรือตัวอย่างการจำแนกกลุ่มลูกค้าที่จะให้ความสนใจในผลิตภัณฑ์ซึ่งจะเปิดตัวใหม่ อาจจะมีกลุ่มที่สามารถจำแนกได้เป็น ไม่สนใจเลย, สนใจเล็กน้อย, สนใจปานกลาง, สนใจมาก หรือในด้านการศึกษาผู้วิจัยอาจจะต้องการจำแนกลักษณะข้อสอบว่าข้อสอบมีการเบี่ยงเบนมากน้อยเพียงใดก็อาจจะมีกลุ่มที่ต้องการจำแนกได้เป็น ไม่มีการเบี่ยงเบน, เบี่ยงเบนในระดับที่รับได้, เบี่ยงเบนในระดับที่รับไม่ได้ (อรินทร์ น่วม-

ถนอม,2549) จากตัวอย่างที่ยกมานี้จะเห็นว่ากลุ่มที่ต้องการจำแนกหน่วยตัวอย่างในแต่ละสถานการณ์สามารถเรียงลำดับความสำคัญของกลุ่มได้

ทางด้านสถิตินั้นมีวิธีการที่สามารถนำมาช่วยในการพยากรณ์หรือบอกโอกาสเกิดของเหตุการณ์ได้ กล่าวคือ ผู้ศึกษาหรือผู้เกี่ยวข้องสามารถใช้วิธีการหรือเทคนิคทางสถิติ เช่น การวิเคราะห์หลายตัวแปร (Multivariate analysis) ช่วยในการสร้างตัวแบบตัวแบบ (model) ในการพยากรณ์ผลซึ่งในที่นี้จะเรียกว่าตัวแปรตาม (independent variable) โดยอาศัยข้อมูลจากตัวแปรอิสระ (dependent variable) ในกรณีตัวแปรตามซึ่งเป็นข้อมูลเชิงกลุ่ม (Categorical Data) โดยที่โดยที่ตัวแปรอิสระมีมาตรวัดแบบใดก็ได้วิธีการที่นำมาใช้ เช่น การวิเคราะห์จำแนกประเภท (Discriminant Analysis) และ การวิเคราะห์การถดถอยโลจิสติก (Logistic Regression Analysis)

การวิเคราะห์จำแนกประเภท เป็นวิธีการทางสถิติที่ใช้พยากรณ์เพื่อการแยกกลุ่มมีข้อตกลงเบื้องต้น (assumption) คือ ตัวแปรอิสระจะต้องมีการแจกแจงแบบปกติหลายตัวแปร (multivariate normal distribution) และเมทริกซ์ความแปรปรวนร่วม(variance-covariance matrix) ของแต่ละกลุ่มจะต้องเท่ากัน จากข้อตกลงเบื้องต้นดังกล่าวเป็นไปได้ยากที่ข้อมูลจะมีลักษณะดังที่ต้องการเพราะข้อมูลที่นำมาศึกษามักเป็นประเภทผสมมีทั้งข้อมูลต่อเนื่องและไม่ต่อเนื่อง (continuous and discrete variable) แต่อย่างไรก็ตาม ยังมีเทคนิคการวิเคราะห์การจำแนกกลุ่มอีกหลายวิธีที่ไม่จำเป็นต้องมีข้อตกลงเบื้องต้น

การวิเคราะห์ความถดถอยโลจิสติก เป็นวิธีการทางสถิติ ซึ่งใช้ประมาณความน่าจะเป็นของการเกิดเหตุการณ์จากกลุ่มของตัวแปรอิสระทั้งที่เป็นตัวแปรไม่ต่อเนื่องและตัวแปรต่อเนื่อง โดยการวิเคราะห์ความถดถอยโลจิสติกมีข้อตกลงเบื้องต้นแค่เพียงตัวแปรอิสระและค่าคลาดเคลื่อนเป็นอิสระกันและตัวแปรอิสระไม่มีความสัมพันธ์กัน(กัลยา วานิชย์บัญชา,2544)

การวิเคราะห์ความถดถอยโลจิสติกอันดับ เป็นการวิเคราะห์เชิงพหุชนิดหนึ่ง ใช้สำหรับวิเคราะห์ข้อมูลที่มีตัวแปรตาม 1 ตัวซึ่งเป็นตัวแปรแบบอันดับ (ordinal) และมีตัวแปรอิสระอย่างน้อยหนึ่งตัวซึ่งจะเป็นตัวแปรแบบใดก็ได้ ในอดีตการวิเคราะห์ทางสถิติเมื่อตัวแปรตามเป็นแบบอันดับ มักจะวิเคราะห์โดยยุบตัวแปรนั้นให้เป็นแบบนามกำหนด (nominal) แล้ววิเคราะห์ด้วย binary logistic regression ซึ่งทำให้รายละเอียดของข้อมูลขาดหายไปเนื่องจากตัวแปรถูกบังคับให้มีเพียงสองระดับเท่านั้นและยังทำให้ค่าสถิติเกิดความคลาดเคลื่อนอีกด้วย (สิริมาและจุฬาลักษณ์,2548) นักสถิติจึงได้คิดค้นวิธีทางสถิติที่เหมาะสมจนในที่สุดออกมาเป็นวิธี การวิเคราะห์ความถดถอยโลจิสติกอันดับ (ordinal logistic regression หรือ proportional odds model)

(David and Stanley,2000) ซึ่งมีเงื่อนไขเพียง odds ratio ของตัวแปรอิสระแต่ละตัวมีค่าคงที่ ไม่ว่าจะแบ่งตัวแปรตาม Y อย่างไร

1.2 วัตถุประสงค์ของการวิจัย

เพื่อศึกษาและเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มแบบอันดับ ด้วยวิธีการวิเคราะห์จำแนกประเภท (Discriminant Analysis) และการวิเคราะห์ความถดถอยโลจิสติกอันดับ (Ordinal Logistic Regression) เมื่อสมมติฐานความเท่ากันของเมตริกซ์ค่าความแปรปรวนร่วมของตัวแปรอิสระของทุกกลุ่มเป็นจริงและไม่เป็นจริง เมื่อตัวแปรอิสระไม่มีความสัมพันธ์กันและสัมพันธ์กัน เมื่อขนาดตัวอย่างมีขนาดมากขึ้น และเมื่อจำนวนตัวอย่างในแต่ละกลุ่มเท่ากันและไม่เท่ากันว่าวิธีการใดจะให้ประสิทธิภาพในการจำแนกกลุ่มมากกว่ากัน

1.3 ขอบเขตของการวิจัย

การวิจัยครั้งนี้มีการกำหนดขอบเขตของการวิจัย ดังนี้

1.ตัวแปรตามเป็นตัวแปรเชิงกลุ่มแบบอันดับจำนวน (k) 3, 4 และ 5 กลุ่ม โดยมีจำนวนตัวแปรอิสระ(p) เป็น 2, 3 และ 4 ตัวแปรอิสระ

2.ตัวแปรอิสระมีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal) ซึ่งแบ่งเป็นกรณีอิสระกัน (independent) และ ไม่เป็นอิสระกัน (dependent) 3 ระดับ ด้วยสหสัมพันธ์ (correlation: ρ) เท่ากับ 0.1 , 0.5 และ 0.9

3.ศึกษาเมื่อขนาดตัวอย่าง (n) มีขนาดเท่ากับ 120 , 240 และ 360

4.กำหนดขนาดตัวอย่างให้แต่ละกลุ่มอันดับ เป็น 3 ระดับดังนี้ ไม่สมดุล (unbalanced) , สมดุลปานกลาง (moderately balanced) และ สมดุล (balanced)

ตารางที่ 1.1 ตารางแสดงจำนวนตัวอย่างในแต่ละกลุ่มอันดับเมื่อขนาดตัวอย่างทั้งหมด (n) ถูกแบ่งในแต่ละกลุ่มเป็นระดับไม่สมดุล สมดุลปานกลาง และสมดุล กรณีตัวแปรตามมี 3 กลุ่ม

	ไม่สมดุล (unbalanced)	**สมดุลปานกลาง** (moderately-balanced)	**สมดุล** (balanced)
n = 120			
กลุ่มที่ 1	10	30	40
กลุ่มที่ 2	40	40	40
กลุ่มที่ 3	70	50	40
n = 240			
กลุ่มที่ 1	50	70	80
กลุ่มที่ 2	80	80	80
กลุ่มที่ 3	110	90	80
n = 360			
กลุ่มที่ 1	90	110	120
กลุ่มที่ 2	120	120	120
กลุ่มที่ 3	150	130	120

ตารางที่ 1.2 ตารางแสดงจำนวนตัวอย่างในแต่ละกลุ่มอันดับเมื่อขนาดตัวอย่างทั้งหมด (n) ถูกแบ่งในแต่ละกลุ่มเป็นระดับไม่สมดุล สมดุลปานกลาง และสมดุล กรณีตัวแปรตามมี 4 กลุ่ม

	ไม่สมดุล (unbalanced)	สมดุลปานกลาง (moderately-balanced)	สมดุล (balanced)
n = 120			
กลุ่มที่ 1	10	20	30
กลุ่มที่ 2	20	30	30
กลุ่มที่ 3	40	30	30
กลุ่มที่ 4	50	40	30
n = 240			
กลุ่มที่ 1	20	50	60
กลุ่มที่ 2	50	60	60
กลุ่มที่ 3	70	60	60
กลุ่มที่ 4	100	70	60
n = 360			
กลุ่มที่ 1	60	80	90
กลุ่มที่ 2	80	90	90
กลุ่มที่ 3	100	90	90
กลุ่มที่ 4	120	100	90

ตารางที่ 1.3 ตารางแสดงจำนวนตัวอย่างในแต่ละกลุ่มอันดับเมื่อขนาดตัวอย่างทั้งหมด (n) ถูกแบ่งในแต่ละกลุ่มเป็นระดับไม่สมดุล สมดุลปานกลาง และสมดุล กรณีตัวแปรตามมี 5 กลุ่ม

	ไม่สมดุล (unbalanced)	สมดุลปานกลาง (moderately-balanced)	สมดุล (balanced)
n = 120			
กลุ่มที่ 1	12	20	24
กลุ่มที่ 2	18	22	24
กลุ่มที่ 3	24	24	24
กลุ่มที่ 4	30	26	24
กลุ่มที่ 5	36	28	24
n = 240			
กลุ่มที่ 1	32	40	48
กลุ่มที่ 2	40	42	48
กลุ่มที่ 3	48	48	48
กลุ่มที่ 4	56	52	48
กลุ่มที่ 5	64	56	48
n = 360			
กลุ่มที่ 1	48	64	72
กลุ่มที่ 2	60	68	72
กลุ่มที่ 3	72	72	72
กลุ่มที่ 4	84	76	72
กลุ่มที่ 5	96	80	72

5. ศึกษากรณีที่เมทริกซ์ความแปรปรวนร่วมของตัวแปรอิสระในแต่ละกลุ่มอันดับเท่ากันและไม่เท่ากัน โดยกำหนดค่าเพื่อสร้างข้อมูลดังนี้

สร้างตัวแปรอิสระให้มีการแจกแจงแบบปกติหลายตัวแปร (Multivariate Normal Distribution) ซึ่งมีฟังก์ชันการแจกแจงเป็น

$$f(x)=\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right]$$

โดยที่ p เป็นจำนวนตัวแปรอิสระ หรือเขียนย่อๆว่า $X \sim N_p(\mu,\Sigma)$ ซึ่ง $X=(x_1,x_2,...,x_n)'$, μ เป็นเวกเตอร์ของค่าเฉลี่ย $(\mu_1,\mu_2,...,\mu_n)'$ ซึ่ง $\mu_i=E(X_i)$, $i=1,2,...,p$ และ Σ เป็นเมทริกซ์ความแปรปรวนร่วม (covariance matrix) ขนาด $p\times p$

$$\Sigma=\begin{bmatrix}\sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,p}\\ \sigma_{2,1} & \sigma_{2,2} & \cdots & \sigma_{2,p}\\ \vdots & & & \\ \sigma_{p,1} & \sigma_{p,2} & \cdots & \sigma_{p,p}\end{bmatrix}$$

โดยที่ $\sigma_{i,j}=Cov(X_i,X_j)=\sigma_{j,i}=Cov(X_j,X_i)$ สำหรับ $i\neq j$ และ $\sigma_{i,i}=Var(X_i)$ $i=1,2,...,p$, $j=1,2,...,p$

การจำลองข้อมูลทำโดยการสร้างตัวแปรอิสระ X แต่ละตัวก่อนให้ได้ตามขนาดตัวอย่างที่จะทำการศึกษาแล้วจึงทำการแบ่งกลุ่มโดยสร้างตัวแปรอิสระให้มีการแจกแจงเป็น

$X \sim N_p(\mu,\Sigma)$, $p=2,3,4$

โดยที่ $\mu=(0,0,...,0)'_{1\times p}$ และ $\Sigma=\begin{bmatrix}1 & \sigma_{1,2} & \cdots & \sigma_{1,p}\\ \sigma_{2,1} & 1 & \cdots & \sigma_{2,p}\\ \vdots & & & \\ \sigma_{p,1} & \sigma_{p,2} & \cdots & 1\end{bmatrix}$

โดยที่ $\sigma_{i,j}$ เมื่อ $i\neq j$ เป็นไปตามขอบเขตของระดับความสัมพันธ์ระหว่างตัวแปรอิสระที่ต้องการศึกษา นั่นคือ

- กรณีตัวแปรอิสระเป็นอิสระกัน (independent) :

$$\sigma_{i,j}=0 \text{ ทุกๆ } i\neq j$$

- กรณีตัวแปรอิสระสัมพันธ์กัน (dependent) ด้วย correlation เท่ากับ 0.1:

$$\sigma_{i,j}=\frac{\rho_{i,j}}{Var(X_i)Var(X_j)}=\frac{0.1}{(1)(1)}=0.1 \quad \text{ทุกๆ } i\neq j$$

- กรณีตัวแปรอิสระสัมพันธ์กัน (dependent) ด้วย correlation เท่ากับ 0.5:

$$\sigma_{i,j} = \frac{\rho_{i,j}}{Var(X_i)Var(X_j)} = \frac{0.5}{(1)(1)} = 0.5 \quad \text{ทุกๆ } i \neq j$$

- กรณีตัวแปรอิสระสัมพันธ์กัน (dependent) ด้วย correlation เท่ากับ 0.9:

$$\sigma_{i,j} = \frac{\rho_{i,j}}{Var(X_i)Var(X_j)} = \frac{0.9}{(1)(1)} = 0.9 \quad \text{ทุกๆ } i \neq j$$

ทั้งกรณีเมทริกซ์ความแปรปรวนร่วมของตัวแปรอิสระในแต่ละกลุ่มอันดับเท่ากันและไม่เท่ากันจะสร้างข้อมูลตัวแปรอิสระเหมือนกันดังที่กล่าวมาแล้วข้างต้น แต่จะทดสอบสมมติฐานการเท่ากันของเมทริกซ์ความแปรปรวนร่วมโดยสถิติทดสอบ Box's Test วนซ้ำให้ได้ข้อมูลทั้งสองกรณีที่ต้องการศึกษา

6.สร้างข้อมูลให้สมมติฐานการเท่ากันของ odds ratio ของทุกตัวแปรอิสระสำหรับการวิเคราะห์ความถดถอยโลจิสติกอันดับเป็นจริงโดยการทดสอบสมมติฐาน

7. เกณฑ์ในการจำแนกกลุ่มสำหรับวิธีวิเคราะห์จำแนกประเภทนั้นจะใช้วิธีความน่าจะเป็นก่อนและความน่าจะเป็นหลัง (Prior Probability and Posterior Probability) โดยงานวิจัยนี้กำหนดความน่าจะเป็นก่อนให้หน่วยตัวอย่างทุกหน่วยมีโอกาสอยู่ในแต่ละกลุ่มเท่ากันทุกกลุ่ม กล่าวคือ ความน่าจะเป็นก่อนที่หน่วยตัวอย่างจะอยู่ในกลุ่มที่ i มีค่าเท่ากับ $\frac{1}{k}$ เมื่อ k คือจำนวนกลุ่มทั้งหมด และ $i = 1,2,...,k$

1.4 เกณฑ์การตัดสินใจ

เกณฑ์ในการตัดสินใจคือ อัตราความผิดพลาดที่เห็นชัด (Apparent Rate Error: APER) หมายถึง จำนวนหน่วยวิเคราะห์ในทั้ง k กลุ่ม ที่ถูกจัดเข้ากลุ่มผิด หารด้วยจำนวนหน่วยวิเคราะห์ทั้งหมดคูณด้วย 100% (Johnson and Wichern:1992)

$$APER = \frac{n_{1M} + n_{2M} + ... + n_{kM}}{n_1 + n_2 + ... + n_k} \times 100\%$$

n_{1M} = จำนวนหน่วยวิเคราะห์ในกลุ่มที่ 1 แต่ถูกวิเคราะห์ให้อยู่ในกลุ่มอื่น

n_{2M} = จำนวนหน่วยวิเคราะห์ในกลุ่มที่ 2 แต่ถูกวิเคราะห์ให้อยู่ในกลุ่มอื่น

n_{kM} = จำนวนหน่วยวิเคราะห์ในกลุ่มที่ k แต่ถูกวิเคราะห์ให้อยู่ในกลุ่มอื่น

n_1 = จำนวนหน่วยวิเคราะห์ทั้งหมดในกลุ่มที่ 1

n_2 = จำนวนหน่วยวิเคราะห์ทั้งหมดในกลุ่มที่ 2

n_k = จำนวนหน่วยวิเคราะห์ทั้งหมดในกลุ่มที่ k

1.5 ประโยชน์ที่คาดว่าจะได้รับ

1. สามารถเปรียบเทียบวิธีการจำแนกกลุ่มแบบอันดับและเลือกใช้ตัวแบบพยากรณ์ได้เหมาะสมกับข้อมูลที่จะทำการวิเคราะห์และมีประสิทธิภาพ

2. สามารถเป็นแนวทางในการศึกษาตัวแบบจำแนกสำหรับการจำแนกกลุ่มแบบอันดับเมื่อข้อมูลมีลักษณะต่างๆออกไป