บทที่ 1

บทนำ

1.1 ความสำคัญและที่มาของปัญหาที่ทำวิจัย

โดยทั่วไปในการทำนายการเป็นสมาชิกกลุ่มของตัวแปรตอบสนองที่มี 2 ค่า ผู้วิจัยอาจใช้วิธีวิเคราะห์การถดถอยพหุคูณ ในการสร้างตัวแบบความน่าจะเป็นเชิงเส้น (Menard, 1995: 98) ซึ่งแสดงความสัมพันธ์ระหว่างตัวแปรตอบสนองและตัวแปรที่ใช้ทำนาย โดยตัวแปรตอบสนองเป็นผลรวมเชิงเส้นของตัวแปรที่ใช้ทำนาย $Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_k X_{ik} + \varepsilon_i$ ซึ่ง β_i เป็นค่าสัมประสิทธิ์การถดถอย X_i เป็นตัวแปรที่ใช้ทำนายและ ε_i เป็นค่าความคลาดเคลื่อนในการประมาณค่าสัมประสิทธิ์การถดถอยซึ่งอธิบายการเปลี่ยนแปลงของค่าเฉลี่ยของตัวแปรตอบสนองเมื่อตัวแปรที่ใช้ทำนายตัวที่สนใจเพิ่มขึ้น 1 หน่วยในขณะที่ตัวแปรที่ใช้ทำนายตัวอื่นๆมีค่าคงที่ โดยมีข้อสมมติว่าความคลาดเคลื่อนมีการแจกแจงปกติที่มีค่าเฉลี่ยเท่ากับศูนย์และมีค่าความแปรปรวนคงที่ ค่าประมาณตัวแปรตอบสนองจากตัวอย่างสุ่มเขียนเป็นสมการได้ดังนี้

$\hat{y}_i = b_0 + b_1 x_{i1} + ... + b_k x_{ik}$ โดย $\hat{y}_i$ เป็นค่าประมาณความน่าจะเป็นเมื่อ $y_i = 1$ โดยทั่วไปความน่าจะเป็นที่ประมาณได้ควรมีค่าอยู่ในช่วง (0,1) แต่ในการประมาณค่าความน่าจะเป็นด้วยวิธี นี้อาจจะให้ค่าประมาณที่อยู่นอกช่วง (0,1) ได้ซึ่งอาจเป็นเพราะความสัมพันธ์ระหว่างตัวแปรตอบสนองและตัวแปรที่ใช้ในการทำนายไม่เป็นเส้นตรง (Hocking,1985: 127) วิธีวิเคราะห์การถดถอยโลจีสติค (Logistic Regression) เป็นหนึ่งในวิธีทางสถิติหลายๆวิธีที่มีการนำมาใช้ในงานวิจัยทางการแพทย์และงานวิจัยทางสังคมศาสตร์ค่อนข้างมากโดยใช้ในการประมาณค่าความน่าจะเป็นของตัวแปรตอบสนอง (King, and Ryan, 2002: 163-170) ให้ y_i แทนค่าของตัวแปรตอบสนองของหน่วยตัวอย่างที่ i ซึ่งมีค่าเท่ากับ 1 ถ้าได้ผลลัพธ์ที่สนใจและมีค่าเท่ากับ 0 ถ้าไม่ได้ผลลัพธ์ที่สนใจ

ให้ $\pi(X) = E(Y|X)$ เป็นค่าเฉลี่ยแบบมีเงื่อนไขของ Y เมื่อกำหนดค่า X

ค่าลอจิกของตัวแบบการถดถอยโลจีสติคสามารถเขียนได้ดังนี้ $g(X) = \ln\frac{\pi(X)}{1-\pi(X)} = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$ โดย $\pi(X) = \frac{\exp[g(X)]}{1+\exp[g(X)]}$ และ $\beta_1, \beta_2, ..., \beta_k$ เป็นค่าสัมประสิทธิ์การถดถอยซึ่งประมาณได้จากวิธี Maximum Likelihood ค่าประมาณแต่ละตัวจะอธิบายการเปลี่ยนแปลงในรูปของค่าลอจิกเมื่อตัวแปรที่ใช้ทำนายตัวใดตัวหนึ่งมีค่าเพิ่มขึ้น 1 หน่วยโดยตัวแปรอื่นๆมีค่าคงที่ ตัวแปรตอบสนองมีการแจกแจง Bernoulli ความน่าจะเป็นของความสำเร็จประมาณได้จากตัวแบบ ค่าลอจิกของแต่ละหน่วยตัวอย่างจะเป็นผลรวมเชิงเส้นของตัวแปรที่ใช้ทำนายซึ่งได้จากการ

แทนค่าตัวแปรที่ใช้ทำนายในสมการ $g(x_i) = b_0 + b_1 x_{i1} + ... + b_k x_{ik}$ ค่าลอจิกที่คำนวณได้มีค่าอยู่ในช่วง $(-\infty, \infty)$ซึ่งจะทำให้ได้ค่าความน่าจะเป็นอยู่ในช่วง (0,1) การทดสอบค่าสัมประสิทธิ์การถดถอยด้วยวิธีวิเคราะห์การถดถอยโลจีสติคจะใช้ค่าสถิติทดสอบ Wald ซึ่งมีการแจกแจงไคสแควร์ (Menard, 1995: 39) โดยค่าสถิติที่ได้จะขึ้นกับการเปลี่ยนแปลงของ Likelihood Function เมื่อเพิ่มตัวแปรที่ใช้ทำนายเข้าไปในตัวแบบ การใช้ค่าสถิติทดสอบ Wald มีกฎเกณฑ์เดียวกับการใช้ค่าสถิติทีและค่าสถิติเอฟจากวิธีวิเคราะห์การถดถอยพหุคูณ ซึ่งสามารถใช้ค่าสถิติจาก Likelihood Function ในการทดสอบความเหมาะสมของตัวแบบเช่นเดียวกัน (Cox and Snell, 1989: 71) ในทางปฏิบัติการใช้วิธีวิเคราะห์การถดถอยโลจีสติคจำเป็นต้องใช้จำนวนตัวอย่างมากกว่าวิธีวิเคราะห์การถดถอยพหุคูณ เนื่องจากการทดสอบค่าสัมประสิทธิ์การถดถอยที่ได้จากวิธี Maximum Likelihood อาจจะให้ผลการทดสอบผิดพลาดได้หากจำนวนตัวอย่างน้อยกว่า 100 (Pampel, 2000:30) เมื่อข้อมูลที่ต้องการวิเคราะห์มีตัวแปรที่ใช้ทำนายมากขึ้นจำเป็นต้องใช้จำนวนตัวอย่างมากขึ้นและอย่างน้อยที่สุดควรจะใช้จำนวนตัวอย่าง 50 รายต่อตัวแปรที่ใช้ทำนาย 1 ตัว (Wright, 1995: 221) โดยทั่วไปจำนวนตัวอย่างที่เหมาะสมจะขึ้นกับระดับความคลาดเคลื่อนชนิดที่ 1 และชนิดที่ 2 ที่ยอมรับได้ ระดับความสัมพันธ์โดยเฉลี่ยระหว่างตัวแปรที่ใช้ทำนายและตัวแปรตอบสนอง ความเชื่อถือได้ของการวัดและการแจกแจงความถี่ของตัวแปรตอบสนอง ในการใช้วิธีวิเคราะห์การถดถอยโลจีสติคนั้นหากตัวแปรตอบสนองมีจำนวนข้อมูลในแต่ละกลุ่มไม่เท่ากันก็จำเป็นต้องใช้จำนวนตัวอย่างมากขึ้น โดยตัวแปรที่ใช้ทำนายอาจเป็นตัวแปรเชิงปริมาณหรือตัวแปรเชิงคุณภาพก็ได้เช่นเดียวกับวิธีวิเคราะห์การถดถอยพหุคูณ

จากการศึกษาของ Pohlmann John T ;Dennis W.(2003) ได้เปรียบเทียบวิธีวิเคราะห์การถดถอยพหุคูณและวิธีวิเคราะห์การถดถอยโลจีสติคโดยใช้ข้อมูล 2 ชุด ซึ่งได้ผลการทดสอบที่คล้ายคลึงกันเมื่อทดสอบที่ระดับนัยสำคัญ 0.05 ทั้งสองวิธีให้ค่าประมาณที่มีความสัมพันธ์กันค่อนข้างมาก เมื่อพิจารณาจากค่ากำลังสองของผลต่างระหว่างค่าความน่าจะเป็นของค่าสังเกตและค่าความน่าจะเป็นที่ประมาณได้พบว่าวิธีการถดถอยโลจีสติคให้ผลการทำนายสมาชิกกลุ่มที่มีความแม่นยำกว่า ซึ่งวิธีวิเคราะห์การถดถอยโลจีสติคสามารถใช้ประมาณค่าความน่าจะเป็นได้ดีกว่าวิธีวิเคราะห์การถดถอยพหุคูณ วิธีวิเคราะห์จำแนกกลุ่ม (Discriminant Analysis) เป็นวิธีที่ได้นำหลักการของวิธีวิเคราะห์การถดถอยและวิธีวิเคราะห์ความแปรปรวนมาใช้หาสมการเชิงเส้นซึ่งแสดงความสัมพันธ์ระหว่างตัวแปรตอบสนองและตัวแปรที่ใช้ทำนายที่ทำให้อัตราส่วนความผันแปรระหว่างกลุ่มและความผันแปรภายในกลุ่มมีค่าสูงสุดหรือทำให้ค่าร้อยละของการทำนายสมาชิกกลุ่มที่ผิดพลาดมีค่าน้อยที่สุด โดยตัวแปรที่ใช้ทำนายเป็นตัวแปรเชิงปริมาณและมีการแจกแจงปกติแต่

อาจมีบางตัวที่เป็นตัวแปรเชิงคุณภาพได้ มีเมทริกซ์ความแปรปรวนและความแปรปรวนร่วมของตัวแปรที่ใช้ทำนายแต่ละกลุ่มมีค่าเท่ากัน ส่วนตัวแปรตอบสนองต้องเป็นตัวแปรเชิงคุณภาพ ในการใช้วิธีวิเคราะห์จำแนกกลุ่มทำนายสมาชิกกลุ่มด้วยโปรแกรมสำเร็จรูป SPSS สามารถเลือกใช้วิธีการทำนายได้ 2 รูปแบบคือแบบ Original ซึ่งทำนายสมาชิกกลุ่มโดยใช้ข้อมูลทั้งหมดในการสร้างสมการจำแนกกลุ่มหรือแบบ Cross_Validation ซึ่งทำนายสมาชิกกลุ่มโดยใช้ข้อมูล n-1 ราย (เมื่อ n แทนจำนวนข้อมูลทั้งหมด) ในการสร้างสมการจำแนกกลุ่มและเหลือข้อมูลอีก 1 ราย ไว้สำหรับทำนายการเป็นสมาชิกกลุ่มจากสมการที่สร้างขึ้น ในกรณีที่ค่าร้อยละของผลการทำนายสมาชิกกลุ่มที่ถูกต้องซึ่งได้จากแบบ Cross_Validation มีค่าน้อยกว่าร้อยละ 25 ของผลรวมกำลังสองของความน่าจะเป็นก่อน (Prior Probability) ของตัวแปรตอบสนองแต่ละกลุ่มจะให้ผลการทำนายที่ผิดพลาดได้และหากพบว่าอัตราส่วนของจำนวนข้อมูลทั้งหมดต่อจำนวนตัวแปรที่ใช้ทำนายน้อยกว่า 20 รายต่อ 1 ตัวแปร หรือ จำนวนข้อมูลในกลุ่มน้อยของตัวแปรตอบสนองน้อยกว่า 20 หรือในกรณีที่ตัวแปรที่ใช้ทำนายมีมาตรการวัดในระดับช่วง ผู้วิจัยจำเป็นต้องระมัดระวังในการอธิบายผล (Schwab, 2003)

วิธีริดจ์รีเกรสชัน (Ridge Regression) เป็นวิธีประมาณค่าสัมประสิทธิ์การถดถอยพหุคูณในกรณีที่ตัวแปรที่ใช้ทำนายมีความสัมพันธ์กันเนื่องจากค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองที่ได้จากการประมาณค่าสัมประสิทธิ์การถดถอยพหุคูณด้วยวิธีกำลังสองน้อยที่สุด (Least Square Method) มีค่าเท่ากับ $\sigma^2(X'X)^{-1}$ ดังนั้นการลดค่าเฉลี่ยความคลาดเคลื่อนกำลังสองให้มีค่าต่ำลงจึงต้องลดค่า $(X'X)^{-1}$ ให้มีค่าต่ำลง Hoerl และ Kennard (1970) ได้เสนอวิธีประมาณค่าสัมประสิทธิ์การถดถอยพหุคูณที่ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำสุดโดยบวกค่าคงที่กับสมาชิกทุกตัวบนเส้นทแยงมุมของเมทริกซ์ $X'X$ ซึ่งได้ค่าประมาณสัมประสิทธิ์การถดถอยด้วยวิธีริดจ์รีเกรสชัน $\hat{\beta}_R = (X'X + kI)^{-1}X'Y$ โดยกำหนดให้ k มีค่าเพิ่มขึ้นทีละน้อย จนกว่าจะได้ค่า k ที่เหมาะสม ทุกครั้งที่เพิ่มค่า k จะนำค่าเฉลี่ยความคลาดเคลื่อนกำลังสองที่ได้จากวิธีริดจ์-รีเกรสชันและค่าเฉลี่ยความคลาดเคลื่อนกำลังสองที่คำนวณจากวิธีกำลังสองน้อยที่สุดมาเปรียบเทียบกัน คำนวณเช่นนี้เรื่อยไปจนกว่าจะได้ค่า k ซึ่งให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองที่ได้จากวิธีริดจ์รีเกรสชันน้อยกว่าวิธีกำลังสองน้อยที่สุด Hoerl, Kennard และ Baldwin (1975) ได้คิดค่า k เริ่มต้นที่เหมาะสมคือ $k = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$ โดย p เป็นจำนวนตัวแปรอิสระ $\hat{\beta}$ และ $\hat{\sigma}^2$ เป็นค่าประมาณซึ่งได้จากวิธีกำลังสองน้อยที่สุดจากนั้นคำนวณค่า $k_1 = \frac{p\hat{\sigma}^2}{\hat{\beta}'_R(k_0)\hat{\beta}_R(k_0)}$

$k_2 = \frac{p\hat{\sigma}^2}{\hat{\beta}'_R(k_1)\hat{\beta}_R(k_1)}$ โดยจะหยุดแทนค่าถ้า $\frac{k_{j+1} - k_j}{k_j} < 20T^{-1.3}$ ซึ่ง $T = \frac{\text{Trace}(X'X)^{-1}}{p}$ ในกรณีตัวแปรที่ใช้ทำนายมีพหุสัมพันธ์กันมาก ค่า T จะมีค่าเพิ่มขึ้น ซึ่งทำให้ $20T^{-1.3}$ มีค่าน้อยลงโอกาสที่จะหยุดแทนค่าจะช้าขึ้น ต่อมา Lawless และ Wang (1976) ได้คิดค่า k เริ่มต้นที่เหมาะสมคือ $k = \frac{p\hat{\sigma}^2}{\sum \lambda_i \hat{\alpha}_i^2}$; $\hat{\alpha} = (Z'Z)^{-1}Z'Y$; λ_i เป็นค่าไอเกน (eigen value) ที่ i ของเมทริกซ์ $X'X$ จากการที่ยอมให้เกิดความเอนเอียงในการประมาณค่าสัมประสิทธิ์การถดถอยพหุคูณจึงทำให้ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองของค่าประมาณสัมประสิทธิ์การถดถอยพหุคูณที่ได้จากวิธีริดจ์รีเกรสชันมีค่าน้อยกว่าค่าความแปรปรวนของค่าประมาณสัมประสิทธิ์การถดถอยพหุคูณจากวิธีกำลังสองน้อยที่สุด

จากการศึกษาเปรียบเทียบประสิทธิภาพการทำนายการเป็นสมาชิกกลุ่มโดยวิธีวิเคราะห์การถดถอยพหุคูณ วิธีวิเคราะห์การถดถอยโลจีสติคและวิธีการวิเคราะห์จำแนกกลุ่มในกรณีที่ตัวแปรตอบสนองมี 2 ค่า (ดวงพร,2551:149-167) พบว่าค่าร้อยละของผลการทำนายที่ถูกต้องซึ่งได้จากวิธีทั้งสามมีความสัมพันธ์เชิงบวกกับค่าสัมประสิทธิ์การกำหนดถึงแม้จะไม่ได้มีการตรวจสอบการแจกแจงปกติของตัวแปรที่ใช้ทำนายก็ตาม วิธีวิเคราะห์จำแนกกลุ่มจะให้ค่าเฉลี่ยของผลการทำนายที่ถูกต้องมากที่สุดในทุกระดับของค่า VIF (น้อยกว่า 4 และไม่น้อยกว่า 4) และค่าร้อยละของตัวแปรเชิงคุณภาพ (แบ่งเป็น 3 กลุ่ม คือ 1) ไม่มีตัวแปรเชิงคุณภาพ 2) มีตัวแปรเชิงคุณภาพร้อยละ 1 – ร้อยละ 50 3) มีตัวแปรเชิงคุณภาพมากกว่าร้อยละ 50) สำหรับกรณีที่มีค่าสัมประสิทธิ์การกำหนดมากกว่า 0.5 ทุกวิธีให้ค่าเฉลี่ยของผลการทำนายที่ถูกต้องไม่แตกต่างกัน ส่วนในกรณีที่มีค่าสัมประสิทธิ์การกำหนดน้อยกว่า 0.5 วิธีวิเคราะห์การถดถอยพหุคูณและวิธีวิเคราะห์การถดถอยโลจีสติคให้ค่าเฉลี่ยร้อยละของผลการทำนายที่ถูกต้องแตกต่างกัน เนื่องจากในการศึกษาผู้วิจัยได้ใช้ข้อมูลจากงานวิจัย 3 เรื่องและข้อมูลที่ใช้เป็นกรณีศึกษาในโปรแกรมสำเร็จรูป SPSS จึงไม่สามารถกำหนดระดับความสัมพันธ์ระหว่างตัวแปรที่ใช้ทำนายได้ครอบคลุมทุกระดับ ในการศึกษานี้จึงสนใจเปรียบเทียบประสิทธิภาพการทำนายสมาชิกกลุ่มด้วยวิธีวิเคราะห์การถดถอยพหุคูณ วิธีริดจ์รีเกรสชัน วิธีวิเคราะห์การถดถอยโลจีสติค และวิธีวิเคราะห์จำแนกกลุ่ม ในกรณีที่ตัวแปรที่ใช้ทำนายมีความสัมพันธ์กันในระดับต่างๆ โดยใช้โปรแกรม R ในการจำลองข้อมูลซึ่งกำหนดให้ตัวแปรที่ใช้ทำนายมีทั้งตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพ

และกำหนดให้ตัวแปรที่ใช้ทำนายมีความสัมพันธ์กันตั้งแต่ระดับน้อยไปจนถึงระดับมากเพื่อหาข้อสรุปสำหรับเสนอเป็นแนวทางในการเลือกใช้วิธีที่เหมาะสมกับข้อมูล

1.2 วัตถุประสงค์ของงานวิจัย

เพื่อเปรียบเทียบประสิทธิภาพการทำนายการเป็นสมาชิกกลุ่มที่ถูกต้องด้วยวิธีวิเคราะห์การถดถอยพหุคูณ วิธีริดจ์รีเกรสชัน วิธีวิเคราะห์การถดถอยโลจีสติค และวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง

1.3 สมมติฐานของการวิจัย

ในกรณีที่ตัวแปรที่ใช้ทำนายมีความสัมพันธ์กันวิธีริดจ์รีเกรสชัน วิธีวิเคราะห์การถดถอยโลจีสติค และวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง ให้ผลการจำแนกสมาชิกกลุ่มที่มีประสิทธิภาพมากกว่าวิธีวิเคราะห์การถดถอยพหุคูณ

1.4 ขอบเขตของการวิจัย

1) ข้อมูลที่ใช้ในการศึกษาได้จากการ Simulation โดยใช้โปรแกรม R ซึ่งกำหนดให้ตัวแปรตอบสนอง Y มี 2 ค่า คือ 0 และ 1

2) กำหนดให้ตัวแปรตอบสนองและตัวแปรที่ใช้ทำนายแต่ละตัวมีความสัมพันธ์กัน generate ตัวแปรตอบสนอง Y จากตัวแปรสุ่มที่มีการแจกแจงทวินาม ซึ่งมีค่า $P[X_i = 1] = \pi$ โดย $\pi(X) = \frac{\exp[g(X)]}{1+\exp[g(X)]}$ และ $\beta_1, \beta_2, ..., \beta_k$ เป็นค่าสัมประสิทธิ์การถดถอยโดย $g(X) = \ln\frac{\pi(X)}{1-\pi(X)} = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$ และกำหนดให้ $\beta_0 = \beta_1 = \beta_3 ... = \beta_k = 1$ ซึ่งอธิบายได้ว่าตัวแปรที่ใช้ทำนายแต่ละตัวมีผลต่อตัวแปรตอบสนองไม่แตกต่างกัน

3) กำหนดให้มีตัวแปรที่ใช้ทำนาย 4 6 และ 10 ตัว ซึ่งมีทั้งตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพ โดยมีกรณีศึกษา 12 กรณี ในตาราง 1.1

ตาราง 1.1 จำนวนตัวแปรที่ใช้ในการศึกษา จำนวนตัวแปรเชิงปริมาณ จำนวนตัวแปรเชิงคุณภาพ และร้อยละของจำนวนตัวแปรเชิงคุณภาพ ของกรณีศึกษา 12 กรณี

	จำนวนตัวแปรที่ใช้ทำนาย	จำนวนตัวแปรเชิงปริมาณ	จำนวนตัวแปรเชิงคุณภาพ	ร้อยละของจำนวนตัวแปรเชิงคุณภาพ
กรณีที่ 1	4	4	0	0
กรณีที่ 2	4	3	1	25
กรณีที่ 3	4	2	2	50
กรณีที่ 4	6	6	0	0
กรณีที่ 5	6	4	2	33.33
กรณีที่ 6	6	3	3	50
กรณีที่ 7	6	2	4	66.67
กรณีที่ 8	10	10	0	0
กรณีที่ 9	10	8	2	20
กรณีที่ 10	10	6	4	40
กรณีที่ 11	10	5	5	50
กรณีที่ 12	10	4	6	60

4) กำหนดให้ตัวแปรเชิงปริมาณแต่ละคู่มีความสัมพันธ์กันตั้งแต่ระดับน้อยไปจนถึงระดับมาก ซึ่งวัดระดับความสัมพันธ์จากค่า Variance Inflation Factor (VIF) ของตัวแปรที่ใช้ทำนายซึ่งมีค่ามากที่สุด โดยแต่ละกรณีกำหนดให้มี 3 ระดับ คือ (1) $VIF \leq 4$ (2) $4 < VIF \leq 10$ และ (3) VIF>10 (ในตาราง 1.2)

ตาราง 1.2 จำนวนตัวแปรที่ใช้ในการศึกษา จำนวนตัวแปรเชิงปริมาณ จำนวนตัวแปรเชิงคุณภาพ และค่า VIF ของข้อมูลที่ใช้ศึกษา

กรณีศึกษา	จำนวนตัวแปรที่ใช้ทำนาย	จำนวนตัวแปรเชิงปริมาณ	จำนวนตัวแปรเชิงคุณภาพ	ค่า VIF ของตัวแปรที่มีค่ามากที่สุด
1	4	4	0	$VIF \leq 4$
2	4	4	0	$4 < VIF \leq 10$
3	4	4	0	$VIF > 10$
4	4	3	1	$VIF \leq 4$
5	4	3	1	$4 < VIF \leq 10$
6	4	3	1	$VIF > 10$
7	4	2	2	$VIF \leq 4$
8	4	2	2	$4 < VIF \leq 10$
9	4	2	2	$VIF > 10$
10	6	6	0	$VIF \leq 4$
11	6	6	0	$4 < VIF \leq 10$
12	6	6	0	$VIF > 10$
13	6	4	2	$VIF \leq 4$
14	6	4	2	$4 < VIF \leq 10$
15	6	4	2	$VIF > 10$
16	6	3	3	$VIF \leq 4$
17	6	3	3	$4 < VIF \leq 10$
18	6	3	3	$VIF > 10$
19	6	2	4	$VIF \leq 4$
20	6	2	4	$4 < VIF \leq 10$
21	6	2	4	$VIF > 10$
22	10	10	0	$VIF \leq 4$
23	10	10	0	$4 < VIF \leq 10$
24	10	10	0	$VIF > 10$
25	10	8	2	$VIF \leq 4$
26	10	8	2	$4 < VIF \leq 10$
27	10	8	2	$VIF > 10$
28	10	6	4	$VIF \leq 4$

ตาราง 1.2 (ต่อ)

กรณีศึกษา	จำนวนตัวแปรที่ใช้ทำนาย	จำนวนตัวแปรเชิงปริมาณ	จำนวนตัวแปรเชิงคุณภาพ	ค่า VIF ของตัวแปรที่มีค่ามากที่สุด
29	10	6	4	$4 < VIF \leq 10$
30	10	6	4	$VIF > 10$
31	10	5	5	$VIF \leq 4$
32	10	5	5	$4 < VIF \leq 10$
33	10	5	5	$VIF > 10$
34	10	4	6	$VIF \leq 4$
35	10	4	6	$4 < VIF \leq 10$
36	10	4	6	$VIF > 10$

5) ในการวัดประสิทธิภาพการจำแนกกลุ่ม Harrell and Lee (1985) ได้เสนอให้ใช้ค่า B ค่า C และค่าความคลาดเคลื่อนในการจำแนก ดังนี้

5.1) ค่า B คำนวณจากสูตร

$B = 1 - \sum_{i=1}^{n} (P_i - Y_i)^2 / n$ โดย P_i เป็นค่าความน่าจะเป็นในการจำแนกตัวแปรตามให้อยู่ในกลุ่ม i Y_i เป็นค่าของตัวแปรตามซึ่งมีค่าเท่ากับ 0 หรือ 1 n เป็นจำนวนสมาชิกของตัวแปร Y_i B มีค่าอยู่ในช่วง [0,1] ซึ่งหากมีค่าเท่ากับ 1 อธิบายได้ว่าสามารถอธิบายการจำแนกได้ถูกต้องสมบูรณ์ ในกรณีที่ $n_0 = n_1$ (n_0 เป็นจำนวนตัวแปรสุ่ม Y_i ที่มีค่าเท่ากับ 0 n_1 เป็นจำนวนตัวแปรสุ่ม Y_i ที่มีค่าเท่ากับ 1) B มีค่าเท่ากับ 0.75 ค่า B นอกจากจะวัดความสามารถในการจำแนกแล้วยังสามารถวัดความแม่นยำในการพยากรณ์ (accuracy of prediction) ด้วย

5.2) ค่า C คำนวณจากสูตร

$$C = \sum_{\substack{i=1 \\ Y_i=0}}^{n_0} \sum_{\substack{j=1 \\ Y_j=1}}^{n_1} \left[I(P_j > P_i) + \frac{1}{2} I(P_j = P_i) \right] / n_0 n_1$$

เป็นค่าวัดความสามารถในการจำแนกค่าของตัวแปรตาม โดย n_0 เป็นจำนวนตัวแปรสุ่ม Y_i ที่มีค่าเท่ากับ 0 n_1 เป็นจำนวนตัวแปรสุ่ม

Y_i ที่มีค่าเท่ากับ 1 P_k เป็นค่าประมาณของ $P(Y_k=1|X_k)$ จาก $P(Y_i=1|X_i)=\frac{e^{\beta^T X_i}}{1+e^{\beta^T X_i}}$ โดย Y_i เป็นตัวแปรสุ่มที่เป็นอิสระกันและมีการแจกแจงเบอร์นูลลี β เป็นค่าประมาณสัมประสิทธิ์การถดถอย ซึ่งหาก C มีค่าเท่ากับ 1 อธิบายได้ว่าสามารถอธิบายการจำแนกได้ถูกต้องสมบูรณ์

5.3) ความคลาดเคลื่อนในการจำแนก คำนวณจากค่าร้อยละของข้อมูลที่ให้ผลการจำแนกแตกต่างจากข้อมูลที่ได้จากการ generate

6) กำหนดจำนวนการทำซ้ำโดยการทดลองทำซ้ำ 1000 รอบ และ 5000 รอบ ทดสอบความแตกต่างของค่าเฉลี่ยของ B ค่าเฉลี่ยของ C และค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนก ซึ่งหากพบว่ามีค่าไม่แตกต่างกัน จะเลือกทำซ้ำ 1000 รอบ

1.5 ประโยชน์ที่คาดว่าจะได้รับ

ได้แนวทางในการเลือกใช้วิธีการทางสถิติในการทำนายการเป็นสมาชิกกลุ่มที่เหมาะสมกับข้อมูล