การเปรียบเทียบประสิทธิภาพการทำนายการเป็นสมาชิกกลุ่มโดยใช้วิธีวิเคราะห์การถดถอยพหุคูณ วิธีริดจ์รีเกรสชัน วิธีวิเคราะห์การถดถอยโลจีสติค และวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง ในกรณีที่ตัวแปรตอบสนองมี 2 ค่า

A Comparison of the Efficiency of Multiple Regression, Ridge Regression, Logistic Regression and Linear Discriminant Analysis for Classifying Binary Outcomes

รองศาสตราจารย์ ดวงพร หัชชะวณิช
สาขาสถิติประยุกต์ คณะวิทยาศาสตร์และเทคโนโลยี
มหาวิทยาลัยหอการค้าไทย
Associate Professor Doungporn Hatchavanich
Department of Applied Statistics
School of Science and Technology
University of the Thai Chamber of Commerce
E-mail: doungporn__hat@utcc.ac.th

บทคัดย่อ

วิธีทางสถิติที่นิยมใช้ในการทำนายการเป็นสมาชิกกลุ่มที่ถูกต้องในกรณีที่ตัวแปรตอบสนองมี 2 ค่า คือ วิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง วิธีวิเคราะห์การถดถอยโลจีสติค และวิธีวิเคราะห์การถดถอย

พหุคูณ ซึ่งข้อสมมติในการใช้วิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรงและวิธีวิเคราะห์การถดถอยพหุคูณจะคล้ายคลึงกัน การใช้วิธีวิเคราะห์การถดถอยพหุคูณในกรณีที่เกิดพหุสัมพันธ์ระหว่างตัวแปรที่ใช้ทำนายจะทำให้ตัวประมาณค่าที่ได้จากวิธีกำลังสองน้อยที่สุดมีความแปรปรวนมาก วิธีที่สามารถแก้ปัญหานี้ได้ คือ วิธีริดจ์รีเกรสชัน ในขณะที่วิธีวิเคราะห์การถดถอยโลจีสติคไม่มีข้อสมมติเกี่ยวกับการแจกแจงปกติ ความสัมพันธ์เชิงเส้นและความแปรปรวนที่เท่ากันของตัวแปรที่ใช้ทำนาย ในการศึกษาครั้งนี้มีวัตถุประสงค์เพื่อเปรียบเทียบประสิทธิภาพในการทำนายการเป็นสมาชิกกลุ่มที่ถูกต้องด้วยวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง วิธีวิเคราะห์การถดถอยโลจีสติค วิธีวิเคราะห์การถดถอยพหุคูณ และวิธีริดจ์รีเกรสชัน โดยใช้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่ม ค่าเฉลี่ยของ B และค่าเฉลี่ยของ C ผลการศึกษา พบว่า ในทุกกรณีวิธีวิเคราะห์การถดถอยโลจีสติคและวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง ให้ผลการทำนายไม่แตกต่างกันและสามารถทำนายได้ดีกว่าวิธีอื่น ๆ

คำสำคัญ: วิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง วิธีวิเคราะห์การถดถอยโลจีสติค วิธีวิเคราะห์การถดถอยพหุคูณ วิธีริดจ์รีเกรสชัน พหุสัมพันธ์

Abstract

The most widely used statistical methods for analyzing categorical outcome variables are Linear Discriminant Analysis and Logistic Regression. If a dependent variable is a binary outcome, an analyst can choose among Discriminant Analysis, Logistic and Multiple Regression. The statistical assumptions required for Discriminant Analysis are essentially the same as for Multiple Regression. In the presence of multicollinearity, the ordinary least squares (OLS) estimator could become unstable due to their large variance, which leads to poor prediction. One of the popular solutions to this problem is Ridge Regression. Logistic Regression makes no assumption about the distribution of the independent variables, which do not have to be normally distributed, linearly related or of equal variance within each group. The purpose of this study was to compare the accuracy of the classifications of group membership of Multiple Regression, Ridge Regression, Logistic Regression and Linear Discrimination Analysis with the mean of classification error, mean of B and mean of C. The results showed that Logistic Regression and Linear Discrimination Analysis performed better than the others.

Keywords: Linear Discriminant Analysis, Logistic Regression, Multiple Regression, Ridge Regression, Multi-Collinearity

บทนำ

โดยทั่วไปในการทำนายการเป็นสมาชิกกลุ่มของตัวแปรตอบสนองที่มี 2 ค่า ผู้วิจัยอาจใช้วิธีวิเคราะห์การถดถอยพหุคูณ ในการสร้างตัวแบบความน่าจะเป็นเชิงเส้น (Menard, 1995) ซึ่งแสดงความสัมพันธ์ระหว่างตัวแปรตอบสนองและตัวแปรที่ใช้ทำนาย โดยตัวแปรตอบสนองเป็นผลรวมเชิงเส้นของตัวแปรที่ใช้ทำนาย $Y_i = \beta_0 + \beta_1 X_{it} + ... + \beta_k X_{ik} + \varepsilon_i$ ซึ่ง β_i เป็นค่าสัมประสิทธิ์การถดถอย X_i เป็นตัวแปรที่ใช้ทำนายและ ε_i เป็นค่าความคลาดเคลื่อนในการประมาณค่าสัมประสิทธิ์การถดถอย ซึ่งอธิบายการเปลี่ยนแปลงของค่าเฉลี่ยของตัวแปรตอบสนองเมื่อตัวแปรที่ใช้ทำนายตัวที่สนใจเพิ่มขึ้น 1 หน่วย ในขณะที่ตัวแปรที่ใช้ทำนายตัวอื่น ๆ มีค่าคงที่ โดยมีข้อสมมติว่าความคลาดเคลื่อนมีการแจกแจงปกติที่มีค่าเฉลี่ยเท่ากับศูนย์และมีค่าความแปรปรวนคงที่ ค่าประมาณตัวแปรตอบสนองจากตัวอย่างสุ่มเขียนเป็นสมการได้ดังนี้ $E(Y_i) = b_0 + b_1 x_{i1} + ... + b_k x_{ik}$ โดย Y_i เป็นค่าประมาณความน่าจะเป็น เมื่อ $Y_i = 1$ โดยทั่วไปความน่าจะเป็นที่ประมาณได้ควรมีค่าอยู่ในช่วง (0,1) แต่ในการประมาณค่าความน่าจะเป็นด้วยวิธีนี้อาจจะให้ค่าประมาณที่อยู่นอกช่วง (0,1) ได้ ซึ่งอาจเป็นเพราะความสัมพันธ์ระหว่างตัวแปรตอบสนองและตัวแปรที่ใช้ในการทำนายไม่เป็นเส้นตรง (Hocking, 1985) วิธีวิเคราะห์การถดถอยโลจีสติค (Logistic Regression) เป็นหนึ่งในวิธีทางสถิติหลาย ๆ วิธีที่มีการนำมาใช้ในงานวิจัยทางการแพทย์และงานวิจัยทางสังคมศาสตร์ค่อนข้างมาก โดยใช้ในการประมาณค่าความน่าจะเป็นของตัวแปรตอบสนอง (King and Ryan, 2002) ให้ Y_i แทนค่าของตัวแปรตอบสนองของหน่วยตัวอย่างที่ i ซึ่งมีค่าเท่ากับ 1 ถ้าได้ผลลัพธ์ที่สนใจและมีค่าเท่ากับ 0 ถ้าไม่ได้ผลลัพธ์ที่สนใจ ในการใช้วิธีวิเคราะห์การถดถอยโลจีสติคนั้นหากตัวแปรตอบสนองมีจำนวนข้อมูลในแต่ละกลุ่มไม่เท่ากันก็จำเป็นต้องใช้จำนวนตัวอย่างมากขึ้น โดยตัวแปรที่ใช้ทำนายอาจเป็นตัวแปรเชิงปริมาณหรือตัวแปรเชิงคุณภาพก็ได้ เช่นเดียวกับวิธีวิเคราะห์การถดถอยพหุคูณ

จากการศึกษาของ Pohlmann และ Dennis (2003) เปรียบเทียบวิธีวิเคราะห์การถดถอยพหุคูณและวิธีวิเคราะห์การถดถอยโลจีสติค โดยใช้ข้อมูล 2 ชุด ซึ่งได้ผลการทดสอบที่คล้ายคลึงกันเมื่อทดสอบที่ระดับนัยสำคัญ 0.05 ทั้งสองวิธีให้ค่าประมาณที่มีความสัมพันธ์กันค่อนข้างมาก เมื่อพิจารณาจากค่ากำลังสองของผลต่างระหว่างค่าความน่าจะเป็นของค่าสังเกตและค่าความน่าจะเป็นที่ประมาณได้ พบว่า วิธีการถดถอยโลจีสติคให้ผลการทำนายสมาชิกกลุ่มที่มีความแม่นยำกว่า ซึ่งวิธีวิเคราะห์การถดถอยโลจีสติคสามารถใช้ประมาณค่าความน่าจะเป็นได้ดีกว่าวิธีวิเคราะห์การถดถอยพหุคูณ

วิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง (Linear Discriminant Analysis) เป็นวิธีที่นำหลักการของวิธีวิเคราะห์การถดถอยและวิธีวิเคราะห์ความแปรปรวนมาใช้หาสมการเชิงเส้นซึ่งแสดงความสัมพันธ์ระหว่างตัวแปรตอบสนองและตัวแปรที่ใช้ทำนายที่ทำให้อัตราส่วนความผันแปรระหว่างกลุ่มและความผันแปรภายในกลุ่มมีค่าสูงสุดหรือทำให้ค่าร้อยละของการทำนายสมาชิกกลุ่มที่ผิดพลาดมีค่าน้อยที่สุด โดยตัวแปรที่ใช้ทำนายเป็นตัวแปรเชิงปริมาณและมีการแจกแจงปกติ แต่อาจมีบางตัวที่เป็นตัวแปรเชิงคุณภาพได้ มีเมทริกซ์ความแปรปรวนและความแปรปรวนร่วมของตัวแปรที่ใช้ทำนายแต่ละกลุ่มมีค่าเท่ากัน ส่วนตัวแปรตอบสนองต้องเป็นตัวแปร

เชิงคุณภาพ หากพบว่าอัตราส่วนของจำนวนข้อมูลทั้งหมดต่อจำนวนตัวแปรที่ใช้ทำนายน้อยกว่า 20 รายต่อ 1 ตัวแปร หรือจำนวนข้อมูลในกลุ่มน้อยของตัวแปรตอบสนองน้อยกว่า 20 หรือในกรณีที่ตัวแปรที่ใช้ทำนายมีมาตรการวัดในระดับช่วง ผู้วิจัยจำเป็นต้องระมัดระวังในการอธิบายผล (Schwab, 2003)

วิธีริดจ์รีเกรสชัน (Ridge Regression) เป็นวิธีประมาณค่าสัมประสิทธิ์การถดถอยพหุคูณในกรณีที่ตัวแปรที่ใช้ทำนายมีความสัมพันธ์กันเนื่องจากค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองที่ได้จากการประมาณค่าสัมประสิทธิ์การถดถอยพหุคูณด้วยวิธีกำลังสองน้อยที่สุด (Least Square Method) มีค่าเท่ากับ $\sigma^2(X'X)^{-1}$ ดังนั้น การลดค่าเฉลี่ยความคลาดเคลื่อนกำลังสองให้มีค่าน้อยลงจึงต้องลดค่า $(X'X)^{-1}$ ให้มีค่าน้อยลง 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 จะนำค่าเฉลี่ยความคลาดเคลื่อนกำลังสองที่ได้จากวิธีริดจ์รีเกรสชันและค่าเฉลี่ยความคลาดเคลื่อนกำลังสองที่คำนวณจากวิธีกำลังสองน้อยที่สุดมาเปรียบเทียบกัน คำนวณเช่นนี้เรื่อยไปจนกว่าจะได้ค่า k ซึ่งให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองที่ได้จากวิธีริดจ์รีเกรสชันน้อยกว่าวิธีกำลังสองน้อยที่สุด Hoerl, Kennard และ Baldwin (1975) ได้คิดค่า k เริ่มต้นที่เหมาะสม คือ $k = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$ โดย p เป็นจำนวนตัวแปรอิสระ $\hat{\beta}$ และ $\hat{\sigma}^2$ เป็นค่าประมาณซึ่งได้จากวิธีกำลังสองน้อยที่สุดจากนั้นคำนวณค่า

$$k_1 = \frac{p\hat{\sigma}^2}{\hat{\beta}'_R(k_0)\hat{\beta}_R(k_0)};\quad k_2 = \frac{p\hat{\sigma}^2}{\hat{\beta}'_R(k_1)\hat{\beta}_R(k_1)}$$

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 เริ่มต้นที่เหมาะสม คือ $k = \frac{p\hat{\sigma}^2}{\sum \lambda_i \hat{\alpha}_i^2}$; $\hat{\alpha} = (Z'Z)^{-1}Z'Y$; λ_i เป็นค่าไอเกน (Eigen Value) ที่ i ของเมทริกซ์ $X'X$ จากการที่ยอมให้เกิดความเอนเอียงในการประมาณค่าสัมประสิทธิ์การถดถอยพหุคูณ จึงทำให้ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองของค่าประมาณสัมประสิทธิ์การถดถอยพหุคูณที่ได้จากวิธีริดจ์รีเกรสชันมีค่าน้อยกว่าค่าความแปรปรวนของค่าประมาณสัมประสิทธิ์การถดถอยพหุคูณจากวิธีกำลังสองน้อยที่สุด

ในการศึกษาสนใจเปรียบเทียบประสิทธิภาพการทำนายสมาชิกกลุ่มด้วยวิธีวิเคราะห์การถดถอยพหุคูณ วิธีริดจ์รีเกรสชัน วิธีวิเคราะห์การถดถอยโลจีสติค และวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง ในกรณีที่ตัวแปรที่ใช้ทำนายมีความสัมพันธ์กันในระดับต่าง ๆ โดยใช้โปรแกรม R จำลองข้อมูลซึ่งกำหนดให้ตัวแปรที่ใช้ทำนายมีทั้งตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพและกำหนดให้ตัวแปรที่ใช้ทำนายมีความสัมพันธ์กันตั้งแต่ระดับน้อยไปจนถึงระดับมากเพื่อหาข้อสรุปสำหรับเสนอเป็นแนวทางในการเลือกใช้วิธีที่เหมาะสมกับข้อมูล

วัตถุประสงค์ของงานวิจัย

เพื่อเปรียบเทียบประสิทธิภาพการทำนายการเป็นสมาชิกกลุ่มที่ถูกต้องด้วยวิธีวิเคราะห์การถดถอยพหุคูณ วิธีริดจ์รีเกรสชัน วิธีวิเคราะห์การถดถอยโลจีสติค และวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง

สมมติฐานของการวิจัย

ในกรณีที่ตัวแปรที่ใช้ทำนายมีความสัมพันธ์กัน วิธีริดจ์รีเกรสชัน วิธีวิเคราะห์การถดถอยโลจีสติค และวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรงให้ผลการจำแนกสมาชิกกลุ่มที่มีประสิทธิภาพมากกว่าวิธีวิเคราะห์การถดถอยพหุคูณ

ขอบเขตของการวิจัย

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

2) กำหนดให้ตัวแปรตอบสนองและตัวแปรที่ใช้ทำนายแต่ละตัวมีความสัมพันธ์กัน จากนั้นจึง generate ตัวแปรตอบสนอง Y จากตัวแปรสุ่มที่มีการแจกแจงทวินามซึ่งมีค่า

$P[X_i = 1] = \pi$ โดย $\pi(X) = \dfrac{\exp[g(X)]}{1+\exp[g(X)]}$ และ

$\beta_1, \beta_2, \ldots \beta_k$ เป็นค่าสัมประสิทธิ์การถดถอยโดย

$$g(X) = \ln\frac{\pi(X)}{1-\pi(X)} = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

และกำหนดให้ $\beta_0 = \beta_1 = \beta_3 \ldots = \beta_k = 1$ ซึ่งอธิบายได้ว่า ตัวแปรที่ใช้ทำนายแต่ละตัวมีผลต่อตัวแปรตอบสนองไม่แตกต่างกัน

3) กำหนดให้มีตัวแปรที่ใช้ทำนาย 4 6 และ 10 ตัว ซึ่งมีทั้งตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพ โดยมีกรณีศึกษา 12 กรณี ในตารางที่ 1

ตารางที่ 1 จำนวนตัวแปรที่ใช้ในการศึกษาจำนวนตัวแปรเชิงปริมาณ จำนวนตัวแปรเชิงคุณภาพ และร้อยละของจำนวนตัวแปรเชิงคุณภาพ ของกรณีศึกษา 12 กรณี

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

4) กำหนดให้ตัวแปรเชิงปริมาณแต่ละคู่มีความสัมพันธ์กันตั้งแต่ระดับน้อยไปจนถึงระดับมาก วัดระดับความสัมพันธ์ระหว่างตัวแปรโดยใช้ค่า Variance Inflation Factor (VIF) ของตัวแปรที่ใช้ทำนายซึ่งมีค่ามากที่สุด กำหนดให้มี 3 ระดับ คือ (1) $VIF \le 4$ (2) $4 < VIF \le 10$ และ (3) $VIF > 10$

5) ในการวัดประสิทธิภาพการจำแนกกลุ่ม Harrell และ 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 อธิบายได้ว่าสามารถจำแนกได้ถูกต้องสมบูรณ์ 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 อธิบายได้ว่าสามารถจำแนกได้ถูกต้องสมบูรณ์

5.3) ความคลาดเคลื่อนในการจำแนก คำนวณจากค่าร้อยละของข้อมูลที่ให้ผลการจำแนกแตกต่างจากข้อมูลที่ได้จากการ generate

6) กำหนดจำนวนรอบของการทำซ้ำโดยทดสอบความแตกต่างของค่าเฉลี่ยของ B ค่าเฉลี่ยของ C และค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มที่ได้จากการทำซ้ำ 1000 รอบและ 5,000 รอบ ซึ่งหากพบว่ามีค่าไม่แตกต่างกัน จะเลือกทำซ้ำ 1,000 รอบ

ประโยชน์ที่คาดว่าจะได้รับ

การศึกษาครั้งนี้ได้แนวทางในการเลือกใช้วิธีการทางสถิติในการทำนายการเป็นสมาชิกกลุ่มที่เหมาะสมกับข้อมูล

วิธีดำเนินการวิจัย

ในการศึกษามีขั้นตอนดำเนินงาน ดังนี้

1) ข้อมูลที่ใช้ในการศึกษาได้จากการ generate ตัวแปรที่ใช้ทำนาย (ตัวแปรอิสระ) ให้มีทั้งตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพ โดยตัวแปรเชิงปริมาณ ได้จากการ generate ให้ตัวแปรมีการแจกแจงปกติมาตรฐานจำนวน n_1 ตัว ส่วนตัวแปรเชิงคุณภาพได้จากการ generate ให้ตัวแปรมีการแจกแจงทวินามโดยมีค่าของตัวแปรเท่ากับ 0 และ 1 โดย $P[X_i = 0] = P[X_i = 1] = 0.5$ กำหนดให้ตัวแปรที่ใช้ทำนายซึ่งเป็นตัวแปรเชิงปริมาณมีความสัมพันธ์กันตั้งแต่ระดับน้อยถึงมาก

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

3) สุ่มตัวอย่างจากประชากรโดยให้มีอัตราส่วนของขนาดตัวอย่างต่อจำนวนตัวแปรที่ใช้ทำนายเท่ากับ 50 ต่อ 1 30 ต่อ 1 20 ต่อ 1 และ 15 ต่อ 1 เพื่อหาตัวแบบในการทำนายการเป็นสมาชิกกลุ่ม

4) แบ่งข้อมูลออกเป็น 2 ชุด โดยข้อมูลชุดที่ 1 ใช้ในการสร้างสมการสำหรับการพยากรณ์ ส่วนข้อมูลชุดที่ 2 ใช้ในการหาค่าความคลาดเคลื่อนในการจำแนกกลุ่ม

5) ใช้วิธีวิเคราะห์การถดถอยพหุคูณ วิธี

วิเคราะห์การถดถอยโลจีสติค วิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง และวิธีริดจ์รีเกรสชัน (โดยวิธี Ridge 1 กำหนดค่า k เริ่มต้นด้วยวิธีของ Hoerl และ Kennard (1970) ส่วนวิธี Ridge 2 กำหนดค่า k เริ่มต้นด้วยวิธีของ Lawless และ Wang (1976)) ประมาณค่าความน่าจะเป็นที่ค่าสังเกตแต่ละค่าจะถูกจำแนกอยู่ในกลุ่มใดกลุ่มหนึ่งซึ่งหากค่าความน่าจะเป็นที่ค่าสังเกตจะถูกจำแนกเป็นกลุ่ม 1 มีค่ามากกว่าค่าความน่าจะเป็นที่ค่าสังเกตจะถูกจำแนกเป็นกลุ่ม 0 ค่าสังเกตค่านั้นจะถูกจำแนกให้อยู่กลุ่ม 1

6) ในการเปรียบเทียบประสิทธิภาพการทำนายการเป็นสมาชิกกลุ่มที่ถูกต้องด้วยวิธีวิเคราะห์การถดถอยพหุคูณ วิธีริดจ์รีเกรสชัน วิธีวิเคราะห์การถดถอยโลจีสติคและวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรงโดยหาค่าเฉลี่ยและค่าส่วนเบี่ยงเบนมาตรฐานของตัวบ่งชี้ B ตัวบ่งชี้ C และความคลาดเคลื่อนในการจำแนกกลุ่ม จากการทำซ้ำ 1,000 รอบ และ 5,000 รอบ

7) ทดสอบความแตกต่างระหว่างค่าเฉลี่ยของตัวบ่งชี้ B ตัวบ่งชี้ C และค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มโดยใช้ระดับนัยสำคัญ 0.05 พบว่า ไม่มีความแตกต่างระหว่างค่าเฉลี่ยจากการทำซ้ำ 1,000 รอบ และ 5,000 รอบ ในครั้งที่ 1 และครั้งที่ 2

8) ทดสอบความแตกต่างระหว่างค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกค่าเฉลี่ยของ B และค่าเฉลี่ยของ C ซึ่งได้จากการใช้วิธีวิเคราะห์การถดถอยพหุคูณ วิธีริดจ์รีเกรสชัน วิธีวิเคราะห์การถดถอยโลจีสติค และวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรงโดยทำซ้ำ 1,000 ครั้ง จากข้อมูลที่ได้จำลองขึ้น

9) สรุปผลการวิจัย

ผลการศึกษา

ผลการเปรียบเทียบประสิทธิภาพในการจำแนกกลุ่มโดยใช้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่ม

จากการศึกษา พบว่า ร้อยละ 94.44 ของข้อมูล (34 กรณี จาก 36 กรณี) ที่วิธีวิเคราะห์การถดถอยโลจีสติคให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มต่ำกว่าวิธีอื่น ๆ นอกจากนี้ ยังพบว่า เมื่อจำนวนตัวแปรที่ใช้ทำนายเพิ่มขึ้นทุกวิธีจะให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มลดลง เมื่อ VIF ของตัวแปรที่ใช้ทำนายที่มากที่สุดเพิ่มขึ้นทุกวิธีจะให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มมีค่าลดลง เมื่อจำนวนตัวแปรเชิงคุณภาพเพิ่มขึ้นทุกวิธีจะทำให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มมีค่าลดลง

เมื่อพิจารณาในแต่ละระดับของค่า VIF และแต่ละระดับของจำนวนตัวแปรเชิงคุณภาพ พบว่า หากอัตราส่วนขนาดตัวอย่างต่อตัวแปรที่ใช้ทำนาย 1 ตัวมีค่าลดลงทุกวิธีจะให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มเพิ่มขึ้นเมื่อพิจารณาในแต่ละระดับของจำนวนตัวแปรที่ใช้ทำนายและค่า VIF ที่มากที่สุด พบว่า กรณีที่ VIF ที่มากที่สุดของตัวแปรที่ใช้ทำนายเพิ่มขึ้นทุกวิธีจะให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกลดลง นอกจากนี้ยังพบว่า เมื่อจำนวนตัวแปรเชิงคุณภาพเพิ่มขึ้นจะให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกลดลง จากการทดสอบสมมติฐานความแตกต่างระหว่างค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มที่ได้จากวิธีวิเคราะห์การถดถอยโลจีสติคกับวิธีวิเคราะห์การถดถอยพหุคูณ วิธีริดจ์รีเกรสชัน (Ridge 1 และ Ridge 2) และวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรงของข้อมูลที่มี VIF < 4 กับ 4 < VIF ≤

10 และ 4 < VIF ≤ 10 กับ VIF > 10 หากใช้ระดับนัยสำคัญ 0.05 ได้ผลสรุป ดังนี้

(1) กรณีที่มีตัวแปรที่ใช้ทำนาย 4 ตัว และ 6 ตัว เมื่อพิจารณาในทุกระดับของจำนวนตัวแปรเชิงคุณภาพ พบว่า ทุกวิธีให้ผลการทดสอบเหมือนกัน กล่าวคือ มีความแตกต่างระหว่างค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มของข้อมูลที่มี VIF < 4 กับ 4 < VIF ≤ 10 และ 4 < VIF ≤ 10 กับ VIF > 10 นอกจากนี้ ยังพบว่า ในกรณีที่ตัวแปรที่ใช้ทำนายคู่ใดคู่หนึ่งในประชากรมีความสัมพันธ์กันมากขึ้นจะทำให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มลดลง

(2) กรณีที่มีตัวแปรที่ใช้ทำนาย 10 ตัว สามารถสรุปผลการทดสอบได้ ดังนี้

(2.1) กรณีที่ไม่มีตัวแปรเชิงคุณภาพและมีตัวแปรเชิงคุณภาพ 2 ตัวและ 4 ตัว พบว่า ทุกวิธีให้ผลการทดสอบเหมือนกัน กล่าวคือ มีความแตกต่างระหว่างค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มของข้อมูลที่มี VIF < 4 กับ 4 < VIF ≤ 10 และ 4 < VIF ≤ 10 กับ VIF > 10

(2.2) กรณีที่มีตัวแปรเชิงคุณภาพ 5 ตัว พบว่า วิธีวิเคราะห์การถดถอยพหุคูณให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มไม่แตกต่างกัน ในกรณีที่ข้อมูลมี VIF < 4 กับ 4 < VIF ≤ 10 วิธีวิเคราะห์การถดถอยโลจีสติคให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มไม่แตกต่างกัน ในกรณีที่ข้อมูลมี 4 < VIF ≤ 10 กับ VIF > 10 วิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรงให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มไม่แตกต่างกันในทุกระดับของค่า VIF ส่วนวิธีริดจ์รีเกรสชันให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มแตกต่างกันในทุกระดับของค่า VIF นอกจากนี้ ยังพบว่า เมื่อตัวแปรที่ใช้ทำนายคู่ใดคู่หนึ่งในประชากรมีความสัมพันธ์กันมากขึ้นจะให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มลดลง

(2.3) กรณีที่มีตัวแปรเชิงคุณภาพ 6 ตัว วิธีวิเคราะห์การถดถอยพหุคูณ วิธีวิเคราะห์การถดถอยโลจีสติค และวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรงให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มไม่แตกต่างกันในกรณีที่ข้อมูลมี VIF< 4 กับ 4 < VIF ≤ 10 ส่วนวิธีริดจ์รีเกรสชันให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มไม่แตกต่างกันในทุกระดับของค่า VIF

ผลการทดสอบความแตกต่างระหว่างค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มด้วยวิธีวิเคราะห์การถดถอยโลจีสติคกับวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง วิธีวิเคราะห์การถดถอยพหุคูณและวิธีริดจ์รีเกรสชันโดยใช้ระดับนัยสำคัญ 0.05 สรุปได้ ดังนี้

(1) กรณีที่มีตัวแปรที่ใช้ทำนาย 4 ตัว วิธีวิเคราะห์การถดถอยโลจีสติค ให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มไม่มีความแตกต่างจากวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง วิธีวิเคราะห์การถดถอยพหุคูณและวิธีริดจ์รีเกรสชัน ดังนั้น ในกรณีนี้จึงสามารถเลือกใช้วิธีใดวิธีหนึ่ง

(2) กรณีที่มีตัวแปรที่ใช้ทำนาย 6 ตัวสามารถสรุปผลได้ ดังนี้

(2.1) กรณีที่ไม่มีตัวแปรเชิงคุณภาพและกรณีที่มีตัวแปรเชิงคุณภาพ 2 ตัว พบว่า วิธีวิเคราะห์การถดถอยโลจีสติคให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มไม่แตกต่างจากวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรงวิธีวิเคราะห์การถดถอยพหุคูณและวิธีริดจ์รีเกรสชัน (Ridge 1 และ Ridge 2) ดังนั้น ในกรณีนี้จึงสามารถเลือกใช้วิธีใดวิธีหนึ่ง

(2.2) กรณีที่มีตัวแปรเชิงคุณภาพ 3 ตัว พบว่า วิธีวิเคราะห์การถดถอยโลจีสติคให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มไม่แตกต่างจากวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง วิธีวิเคราะห์การถดถอยพหุคูณ วิธีริดจ์รีเกรสชันซึ่งกำหนดค่า k เริ่มต้นด้วยวิธีที่ 2 (Ridge 2) ในกรณี VIF > 4 และวิธีริดจ์รีเกรสชันซึ่งกำหนดค่า k เริ่มต้นด้วยวิธีที่ 1 (Ridge 1) ในกรณีมี VIF >10

(2.4) กรณีที่มีตัวแปรเชิงคุณภาพ 4 ตัว วิธีวิเคราะห์การถดถอยโลจีสติคให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มไม่แตกต่างจากวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง ดังนั้น ในกรณีนี้จึงควรใช้วิธีวิเคราะห์การถดถอยโลจีสติค หรือวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรงเพราะให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มน้อยที่สุด

(3) กรณีที่มีตัวแปรที่ใช้ทำนาย 10 ตัว สามารถสรุปผลได้ ดังนี้

(3.1) กรณีที่ไม่มีตัวแปรเชิงคุณภาพและกรณีที่มีตัวแปรเชิงคุณภาพ 2 ตัว พบว่า วิธีวิเคราะห์การถดถอยโลจีสติค ให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มไม่แตกต่างกับวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง วิธีวิเคราะห์การถดถอยพหุคูณและวิธีริดจ์รีเกรสชัน (Ridge 1 และ Ridge 2)

(3.2) กรณีที่มีตัวแปรเชิงคุณภาพ 4 ตัว 5 ตัว และ 6 ตัว พบว่า วิธีวิเคราะห์การถดถอยโลจีสติคให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มไม่แตกต่างกับวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง

ผลการเปรียบเทียบประสิทธิภาพในการจำแนกกลุ่มโดยใช้ค่าเฉลี่ยของ B

จากการศึกษา พบว่า วิธีวิเคราะห์การถดถอยโลจีสติคให้ค่าเฉลี่ยของ B มากที่สุด คิดเป็นร้อยละ 83.33 และวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรงให้ค่าเฉลี่ยของ B มากที่สุดคิดเป็นร้อยละ 47.22 จากการทดสอบความแตกต่างระหว่างค่าเฉลี่ยของ B ที่ได้จากวิธีวิเคราะห์การถดถอยโลจีสติคกับวิธีอื่น ๆ โดยใช้ระดับนัยสำคัญ 0.05 พบว่า ในทุกกรณีที่ศึกษาวิธีวิเคราะห์การถดถอยโลจีสติคให้ค่าเฉลี่ยของ B ไม่แตกต่างกับวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรงแต่ให้ผลการทดสอบที่แตกต่างกับวิธีอื่น ๆ ดังนั้น จึงสรุปได้ว่าวิธีวิเคราะห์การถดถอยโลจีสติคและวิธีวิเคราะห์จำแนกกลุ่มมีประสิทธิภาพในการจำแนกกลุ่มไม่แตกต่างกันและมีประสิทธิภาพดีกว่าวิธีวิเคราะห์การถดถอยพหุคูณและวิธีริดจ์รีเกรสชัน

ผลการทดสอบความแตกต่างระหว่างค่าเฉลี่ยของ B ของข้อมูลที่มี VIF $\leq$ 4 กับ 4 < VIF $\leq$ 10 และ 4 < VIF $\leq$ 10 กับ VIF > 10 โดยใช้ระดับนัยสำคัญ 0.05 สามารถสรุปได้ ดังนี้

(1) กรณีที่มีตัวแปรที่ใช้ทำนาย 4 ตัว ทุกวิธีให้ผลการทดสอบเหมือนกันในทุกกรณี กล่าวคือ มีความแตกต่างระหว่างค่าเฉลี่ยของ B ของข้อมูลที่มีระดับ VIF ที่ต่างกัน โดยเมื่อตัวแปรที่ใช้ทำนายมีค่า VIF ที่มากที่สุดเพิ่มขึ้นทุกวิธีจะให้ค่าเฉลี่ยของ B เพิ่มขึ้น

(2) กรณีที่มีตัวแปรที่ใช้ทำนาย 6 ตัว ผลการทดสอบสามารถสรุปได้ ดังนี้

(2.1) วิธีวิเคราะห์การถดถอยโลจีสติคและวิธีวิเคราะห์จำแนกกลุ่มให้ผลการทดสอบที่เหมือนกันในทุกกรณี โดยมีความแตกต่างระหว่างค่าเฉลี่ยของ B เมื่อมีระดับค่า VIF ที่แตกต่างกัน เมื่อตัวแปรที่

ใช้ทำนายมีค่า VIF ที่มากที่สุดเพิ่มขึ้นจะให้ค่าเฉลี่ยของ B เพิ่มขึ้น

(2.2) วิธีวิเคราะห์การถดถอยพหุคูณและวิธีริดจ์รีเกรสชันให้ผลการทดสอบที่เหมือนกัน โดยในกรณีที่มีตัวแปรเชิงคุณภาพ 4 ตัว ไม่มีความแตกต่างระหว่างค่าเฉลี่ยของ B ในกรณีที่มีค่า VIF ≤ 4 และ $4 < \text{VIF} \leq 10$ ส่วนกรณีอื่น ๆ ให้ผลการทดสอบที่แตกต่างกัน โดยเมื่อตัวแปรที่ใช้ทำนายมีค่า VIF ที่มากที่สุดเพิ่มขึ้นจะให้ค่าเฉลี่ยของ B เพิ่มขึ้น

(3) กรณีที่มีตัวแปรที่ใช้ทำนาย 10 ตัว ให้ผลการทดสอบ ดังนี้

(3.1) วิธีวิเคราะห์การถดถอยโลจีสติคและวิธีวิเคราะห์จำแนกกลุ่มให้ผลการทดสอบไม่แตกต่างกันในทุกกรณี โดยในกรณีที่มีตัวแปรเชิงคุณภาพ 6 ตัว สองวิธีนี้ให้ค่าเฉลี่ยของ B ไม่แตกต่างกันแม้ว่าจะมีระดับ VIF แตกต่างกัน ส่วนในกรณีอื่น ๆ ให้ผลการทดสอบที่แตกต่างกันโดยเมื่อระดับ VIF เพิ่มขึ้นจะให้ค่าเฉลี่ยของ B เพิ่มขึ้น

(3.2) วิธีวิเคราะห์การถดถอยพหุคูณและวิธีริดจ์รีเกรสชันให้ผลการทดสอบที่เหมือนกันทุกกรณี โดยในกรณีที่มีตัวแปรเชิงคุณภาพ 5 ตัว สองวิธีนี้ให้ค่าเฉลี่ยของ B ไม่แตกต่างกันแม้ว่าจะมีระดับ VIF ที่แตกต่างกัน ส่วนในกรณีอื่น ๆ ให้ผลการทดสอบที่แตกต่างกันโดยเมื่อมีระดับ VIF เพิ่มขึ้นจะให้ค่าเฉลี่ยของ B เพิ่มขึ้น ซึ่งสามารถสรุปได้ว่าเมื่อตัวแปรที่ใช้ทำนายตัวใดตัวหนึ่งมีความสัมพันธ์กับตัวแปรอื่นเพิ่มขึ้นจะทำให้ประสิทธิภาพการจำแนกกลุ่มเพิ่มขึ้น

ผลการเปรียบเทียบประสิทธิภาพในการจำแนกกลุ่มโดยใช้ค่าเฉลี่ยของ C

จากการศึกษา พบว่า วิธีวิเคราะห์การถดถอยโลจีสติคให้ค่าเฉลี่ยของ C มากที่สุด (ร้อยละ 97.22) ผลการทดสอบความแตกต่างระหว่างค่าเฉลี่ยของ C ที่ได้จากวิธีวิเคราะห์การถดถอยโลจีสติคกับวิธีอื่น ๆ โดยใช้ระดับนัยสำคัญ 0.05 สามารถสรุปได้ ดังนี้

(1) ในทุกกรณีที่ศึกษาไม่มีความแตกต่างระหว่างค่าเฉลี่ยของ C ที่ได้จากวิธีวิเคราะห์การถดถอยโลจีสติคกับวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรงและวิธีริดจ์รีเกรสชัน

(2) วิธีวิเคราะห์การถดถอยโลจีสติค มีประสิทธิภาพในการจำแนกกลุ่มไม่แตกต่างกับวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรงและวิธีริดจ์-รีเกรสชันในทุกกรณี ในขณะที่วิธีวิเคราะห์การถดถอยโลจีสติคมีประสิทธิภาพในการจำแนกกลุ่มมากกว่าวิธีวิเคราะห์การถดถอยพหุคูณในกรณีที่มีตัวแปรที่ใช้ทำนาย 10 ตัว โดยเป็นตัวแปรเชิงคุณภาพ 5 ตัว และมีค่า VIF > 4 ส่วนกรณีอื่น ๆ มีประสิทธิภาพในการจำแนกกลุ่มไม่แตกต่างกัน

จากการเปรียบเทียบค่าเฉลี่ยของ C ตามระดับของ VIF ซึ่งแบ่งเป็น 3 ระดับ คือ VIF ≤ 4 $4 < \text{VIF} \leq 10$ และ VIF > 10 พบว่า เมื่อมีระดับ VIF เพิ่มขึ้นทุกวิธีจะให้ค่าเฉลี่ยของ C เพิ่มขึ้น นอกจากนี้ ยังพบว่า เมื่อจำนวนตัวแปรเชิงคุณภาพเพิ่มขึ้นทุกวิธีจะให้ค่าเฉลี่ยของ C ลดลง เมื่อจำนวนตัวแปรที่ใช้ทำนายเพิ่มขึ้นทุกวิธีจะให้ค่าเฉลี่ยของ C เพิ่มขึ้น จากการทดสอบความแตกต่างระหว่างค่าเฉลี่ยของ C ในกรณี VIF ≤ 4 กับ $4 < \text{VIF} \leq 10$ และ $4 < \text{VIF} \leq 10$ กับ VIF > 10 พบว่า ทุกวิธีให้ผลการทดสอบเหมือนกัน กล่าวคือ มีความแตกต่างระหว่างค่าเฉลี่ยของ C ในแต่ละระดับของ VIF โดยเมื่อ VIF เพิ่มขึ้นจะให้ค่าเฉลี่ยของ C เพิ่มขึ้น ยกเว้นในกรณีที่มีตัวแปรที่ใช้ทำนาย 10 ตัว โดยเป็นตัวแปรเชิงคุณภาพ 6 ตัว จะไม่มีความแตกต่างระหว่างค่าเฉลี่ยของ C ในกรณี VIF ≤ 4 กับ $4 < \text{VIF}$

≤ 10 ดังนั้น จึงสรุปได้ว่าทุกวิธีให้ผลการทดสอบที่เหมือนกัน กล่าวคือ เมื่อระดับ VIF เพิ่มขึ้นหรือเมื่อมีจำนวนตัวแปรที่ใช้ทำนายเพิ่มขึ้น จะทำให้ประสิทธิภาพในการจำแนกกลุ่มของ C จะเพิ่มขึ้น แต่หากมีจำนวนตัวแปรเชิงคุณภาพเพิ่มขึ้นจะทำให้ประสิทธิภาพในการจำแนกกลุ่มของ C ลดลง

สรุปผลและข้อเสนอแนะ

ในการทำนายการเป็นสมาชิกกลุ่มโดยทั่วไปหากตัวแปรที่ใช้ทำนายบางตัวมีความสัมพันธ์กันอย่างชัดเจนกับตัวแปรอื่นก็ควรพิจารณาตัดตัวแปรที่ใช้ทำนายบางตัวออกไป ซึ่งอาจจะมีตัวแปรที่เหลือบางตัวยังคงมีความสัมพันธ์กับตัวแปรอื่น ๆ ในกรณีเช่นนี้จึงจำเป็นต้องใช้วิธีวิเคราะห์การถดถอยพหุคูณ วิธีริดจ์รีเกรสชัน วิธีวิเคราะห์การถดถอยโลจีสติค และวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง ในการจำแนกการเป็นสมาชิกกลุ่มโดยที่ยังมีตัวแปรที่ใช้ทำนายบางตัวมีความสัมพันธ์กัน

ผลการเปรียบเทียบประสิทธิภาพในการทำนายการเป็นสมาชิกกลุ่มจากการใช้วิธีวิเคราะห์การถดถอยโลจีสติค วิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง วิธีวิเคราะห์การถดถอยเชิงพหุ และวิธีริดจ์รีเกรสชัน โดยพิจารณาจากค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มซึ่งพบว่าวิธีวิเคราะห์การถดถอยโลจีสติคให้ค่าเฉลี่ยของความคลาดเคลื่อนน้อยที่สุด แต่จากการทดสอบสมมติฐานความแตกต่างระหว่างค่าเฉลี่ยของความคลาดเคลื่อนที่ได้จากวิธีวิเคราะห์การถดถอยโลจีสติคกับวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง วิธีริดจ์รีเกรสชัน และวิธีวิเคราะห์การถดถอยพหุคูณโดยใช้ระดับนัยสำคัญ 0.05 พบว่า ในทุกกรณีวิธีวิเคราะห์การถดถอยโลจีสติคให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกไม่แตกต่างจากวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง ซึ่งอธิบายได้ว่ามีความสามารถในการจำแนกกลุ่มไม่แตกต่างกัน แต่ในบางกรณีวิธีวิเคราะห์การถดถอยโลจีสติคก็มีความสามารถในการจำแนกไม่แตกต่างจากวิธีริดจ์รีเกรสชันและวิธีวิเคราะห์การถดถอยพหุคูณ ซึ่งสามารถสรุปผลการศึกษาได้ในตารางที่ 2

ผลการเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มโดยใช้ค่าเฉลี่ยของ B สามารถสรุปได้ว่าวิธีวิเคราะห์การถดถอยโลจีสติคให้ค่าเฉลี่ยของ B ไม่แตกต่างจากวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง นอกจากนี้ ยังพบว่า สองวิธีนี้ให้ค่าเฉลี่ยของ B มากที่สุดในทุกกรณี เมื่อพิจารณาในแต่ละระดับของจำนวนตัวแปรที่ใช้ทำนายและจำนวนตัวแปรเชิงคุณภาพ พบว่า เมื่อข้อมูลมีค่า VIF เพิ่มขึ้นจะทำให้ค่าเฉลี่ยของ B เพิ่มขึ้น เมื่อจำนวนตัวแปรที่ใช้ทำนายเพิ่มขึ้นจะทำให้ค่าเฉลี่ยของ B เพิ่มขึ้นเมื่อพิจารณาที่แต่ละระดับของจำนวนตัวแปรที่ใช้ทำนายและค่า VIF พบว่า ในกรณีที่ประชากรมีจำนวนตัวแปรเชิงคุณภาพเพิ่มขึ้นจะทำให้ค่าเฉลี่ยของ B เพิ่มขึ้น

ตารางที่ 2 ข้อเสนอแนะในการเลือกใช้วิธีที่ให้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่มน้อยที่สุดจำแนกตามจำนวนตัวแปรที่ใช้ทำนาย จำนวนตัวแปรเชิงคุณภาพ และระดับ VIF

จำนวนตัวแปรที่ใช้ทำนาย	จำนวนตัวแปรเชิงคุณภาพ	VIF	วิธีที่ควรเลือกใช้				
			Logit	LDA	OLS	Ridge 1	Ridge 2
4	0, 1, 2	ทุกระดับ	√	√	√	√	√
6	0, 2	ทุกระดับ	√	√	√	√	√
	3	VIF $\leq$ 4	√	√	√		
		4 < VIF $\leq$ 10	√	√	√		
		VIF > 10	√	√	√	√	√
	4	ทุกระดับ	√	√			
10	0, 2	ทุกระดับ	√	√			
	4, 5, 6	ทุกระดับ	√	√			

ผลการเปรียบเทียบประสิทธิภาพการจำแนกกลุ่มโดยใช้ค่าเฉลี่ยของ C สามารถสรุปได้ว่าวิธีวิเคราะห์การถดถอยโลจีสติคให้ค่าเฉลี่ยของ C ไม่แตกต่างจากวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง วิธีริดจ์รีเกรสชัน และวิธีวิเคราะห์การถดถอยพหุคูณในทุกกรณี เมื่อพิจารณาในแต่ละระดับของจำนวนตัวแปรที่ใช้ทำนายและจำนวนตัวแปรเชิงคุณภาพ พบว่า เมื่อข้อมูลมีค่า VIF เพิ่มขึ้นจะทำให้ค่าเฉลี่ยของ C เพิ่มขึ้น เมื่อจำนวนตัวแปรที่ใช้ทำนายเพิ่มขึ้นจะทำให้ค่าเฉลี่ยของ C เพิ่มขึ้น เมื่อพิจารณาที่แต่ละระดับของจำนวนตัวแปรที่ใช้ทำนายและค่า VIF พบว่า ในกรณีที่ประชากรมีจำนวนตัวแปรเชิงคุณภาพเพิ่มขึ้นจะทำให้ค่าเฉลี่ยของ C ลดลง

ผลการเปรียบเทียบประสิทธิภาพการทำนายการเป็นสมาชิกกลุ่มที่ถูกต้องโดยใช้ค่าเฉลี่ยของความคลาดเคลื่อนในการจำแนกกลุ่ม ค่าเฉลี่ยของ B และค่าเฉลี่ยของ C พบว่า ในทุกกรณีวิธีวิเคราะห์การถดถอยโลจีสติคและวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรงมีประสิทธิภาพการทำนายการเป็นสมาชิกกลุ่มที่ถูกต้องไม่แตกต่างกัน และมีประสิทธิภาพในการจำแนกกลุ่มดีกว่าวิธีวิเคราะห์การถดถอยพหุคูณและวิธีริดจ์รีเกรสชันถึงแม้จะตัวแปรที่ใช้ในการทำนายบางตัวจะมีความสัมพันธ์กันก็ตาม

กิตติกรรมประกาศ

ขอขอบคุณมหาวิทยาลัยหอการค้าไทยที่ให้ทุนสนับสนุนการวิจัย

บรรณานุกรม

Harrell, F.E. , and Lee, K.L. 1985. “A Comparison of the Discrimination of Discriminant Analysis and Logistic Regression under Multivariate Normallity.” In P.K. Send (Ed.), **Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences,** pp. 333-343. Amsterdam: Elsevier Science.

Hocking, R.R. 1985. **The Analysis of Linear Models.** Monterey, CA: Brooks.

Hoerl, E., and Kennard, Robert W. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." **Technometrics** 12, 1: 55-67.

Hoerl, E., Kennard, Robert W., and Baldwin, Kent F. 1975. "Ridge Regression: Some Simulations." **Communication in Statistics** 4, 2: 105-123.

King, E.N., and Ryan, T.P. 2002. "A Preliminary Investigation of Maximum Likelihood Logistic Regression Versus Exact Logistic Regression." **American Statistician** 56, 3: 163-170.

Lawless, J.F., and Wang, P. 1976. "A Simulation Study of Ridge and other Regression Estimators." **Communications in Statistics** 5, 4: 307-323.

Menard, S. 1995. **Applied Logistic Regression Analysis.** Thousand Oaks, CA: Sage.

Pohmann, John T., and Dennis, W. 2003. **A Comparison of Ordinary Least Squares and Logistic Regression** [Online]. Available: https://kb.osu.edu/dspace/bitstream/handle/1811/23983 V103N5__118.pdf

Schwab, A.J. 2003. **Strategy for Complete Discriminant Analysis** [Online]. Available: sw38847/SolvingProblems

Associate Professor Doungporn Hatchavanich received her Master of Science Degree in Statistics from Chulalongkorn University, Thailand. She is currently a lecturer at the School of Science and Technology, University of the Thai Chamber of Commerce. Her main interests are in Sampling Techniques and Regression Analysis.