บทที่ 2
ทฤษฎีที่เกี่ยวข้อง

2.1 วิธีวิเคราะห์การถดถอยโลจีสติค (Logistic Regression)

วิธีวิเคราะห์การถดถอยโลจีสติคเป็นหนึ่งในวิธีทางสถิติหลาย ๆ วิธีที่มีการนำมาใช้ในงานวิจัยทางการแพทย์และงานวิจัยทางสังคมศาสตร์ค่อนข้างมากโดยใช้ในการประมาณค่าความน่าจะเป็นของตัวแปรตอบสนอง (King, and Ryan, 2002: 163-170) ให้ y_i แทนค่าของตัวแปรตอบสนองของหน่วยตัวอย่างที่ i ซึ่งมีค่าเท่ากับ 1 ถ้าได้ผลลัพธ์ที่สนใจและมีค่าเท่ากับ 0 ถ้าไม่ได้ผลลัพธ์ที่สนใจให้ $\pi(X) = E(Y \mid X)$ 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 หน่วย โดยตัวแปรอื่น ๆ มีค่าคงที่ ตัวแปรตอบสนองมีการแจกแจง Bernoulli ความน่าจะเป็นของความสำเร็จประมาณได้จากตัวแบบ ค่าลอจิกของแต่ละหน่วยตัวอย่างจะเป็นผลรวมเชิงเส้นของตัวแปรที่ใช้ทำนายซึ่งได้จากการแทนค่าตัวแปรที่ใช้ทำนายในสมการ $g(x_i) = b_0 + b_1 x_{i1} + ... + b_k x_{ik}$ ค่าลอจิกที่คำนวณได้มีค่าอยู่ในช่วง $(-\infty, \infty)$ ซึ่งจะทำให้ได้ค่าความน่าจะเป็นอยู่ในช่วง (0,1) การทดสอบค่าสัมประสิทธิ์การถดถอยด้วยวิธีวิเคราะห์การถดถอยโลจีสติคจะใช้ค่าสถิติทดสอบ Wald ซึ่งมีการแจกแจงไคสแควร์ (Menard, 1995: 39) โดยค่าสถิติที่ได้จะขึ้นกับการเปลี่ยนแปลงของ Likelihood Function เมื่อเพิ่มตัวแปรที่ใช้ทำนายเข้าไปในตัวแบบ การใช้ค่าสถิติทดสอบ Wald มีกฎเกณฑ์เดียวกับการใช้ค่าสถิติทีและค่าสถิติเอฟจากวิธีวิเคราะห์การถดถอยพหุคูณซึ่งใช้ค่าสถิติจาก Likelihood Function ในการทดสอบความเหมาะสมของตัวแบบเช่นเดียวกัน (Cox and Snell, 1989: 71) ในทางปฏิบัติการใช้วิธีวิเคราะห์การถดถอยโลจีสติคจำเป็นต้องใช้จำนวนตัวอย่างมากกว่าวิธีวิเคราะห์การถดถอยพหุคูณเนื่องจากการทดสอบค่าสัมประสิทธิ์การถดถอยที่ได้จากวิธี Maximum Likelihood อาจจะให้ผลการทดสอบผิดพลาดได้หากจำนวนตัวอย่างน้อยกว่า 100 (Pampel, 2000:30) เมื่อข้อมูลที่ต้องการวิเคราะห์มีตัวแปรที่ใช้ทำนายมากขึ้นจำเป็นต้องใช้จำนวนตัวอย่างมากขึ้นและอย่างน้อยที่สุดควรจะใช้จำนวนตัวอย่าง 50 รายต่อตัวแปรที่ใช้ทำนาย 1 ตัว (Wright,1995: 221) โดยทั่วไปจำนวนตัวอย่างที่เหมาะสมจะขึ้นกับระดับความคลาดเคลื่อนชนิดที่ 1 และ ชนิดที่ 2 ที่ยอมรับได้ ระดับความสัมพันธ์โดยเฉลี่ยระหว่างตัวแปรที่ใช้ทำนายและตัวแปรตอบสนอง ความเชื่อถือได้ของการวัดและการแจกแจงความถี่ของตัวแปรตอบสนอง ในการใช้วิธีวิเคราะห์การถดถอยโลจีสติค หากตัวแปรตอบสนองมีจำนวนข้อมูล

ในแต่ละกลุ่มไม่เท่ากันก็จำเป็นต้องใช้จำนวนตัวอย่างมากขึ้นโดยตัวแปรที่ใช้ทำนายอาจเป็นตัวแปรเชิงปริมาณหรือตัวแปรเชิงคุณภาพก็ได้เช่นเดียวกับวิธีวิเคราะห์การถดถอยพหุคูณ

จากการศึกษาของ Pohlmann John T ; Dennis W.(2003) ได้เปรียบเทียบวิธีวิเคราะห์การถดถอยพหุคูณและวิธีการถดถอยโลจีสติคโดยใช้ข้อมูล 2 ชุด ซึ่งได้ผลการทดสอบคล้ายคลึงกันเมื่อทดสอบที่ระดับนัยสำคัญ 0.05 ทั้งสองวิธีให้ค่าประมาณที่มีความสัมพันธ์กันค่อนข้างมาก เมื่อพิจารณาจากค่ากำลังสองของผลต่างระหว่างค่าความน่าจะเป็นของค่าสังเกตและค่าความน่าจะเป็นที่ประมาณได้พบว่าวิธีวิเคราะห์การถดถอยโลจีสติคให้ผลการทำนายสมาชิกกลุ่มที่มีความแม่นยำกว่า วิธีวิเคราะห์การถดถอยโลจีสติคจึงสามารถใช้ประมาณค่าความน่าจะเป็นได้ดีกว่าวิธีวิเคราะห์การถดถอยพหุคูณ

2.2 วิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง (Linear Discriminant Analysis : LDA)

วิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง สามารถใช้ในการหาตัวแปรที่ใช้ในการจำแนกกลุ่มของตัวแปรตามซึ่งมีตั้งแต่ 2 กลุ่มขึ้นไปเพื่อสร้างตัวแบบในการทำนายการเป็นสมาชิกกลุ่มของข้อมูลใหม่ (Worth and Cronin,2003) ในการใช้วิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรง ตัวแปรที่ใช้ทำนายในแต่ละกลุ่มต้องมีการแจกแจงปกติซึ่งมีค่าความแปรปรวนไม่แตกต่างกัน โดยวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรงที่ง่ายที่สุดเป็นวิธีที่ใช้จำแนกกลุ่มในกรณีที่ตัวแปรตามมี 2 กลุ่มซึ่งฟังก์ชันที่ใช้ในการจำแนกกลุ่มในเชิงเส้นตรงจะผ่านจุดเซนทรอยด์ (centrods) ของกลุ่มทั้งสอง ในตัวแบบมาตรฐานของวิธีวิเคราะห์จำแนกในเชิงเส้นตรง (Standard Linear Discriminant Analysis Model) มีข้อสมมติว่าการแจกแจงแบบมีเงื่อนไขของ X|y เป็นแบบ multivariate normal ซึ่งมีค่าเวคเตอร์เฉลี่ย μ_y โดยมีเมทริกซ์ค่าความแปรปรวน $\sum$ ซึ่งความน่าจะเป็นที่ค่าสังเกตใดๆจะถูกจำแนกให้เป็นกลุ่ม 1 เป็นดังนี้ $P(1|x)=\frac{1}{1+(e^{\alpha+\beta x})^{-1}}$ โดยค่าสัมประสิทธิ์ α และ β คำนวณได้ ดังนี้ $\alpha=-\log\frac{\pi_1}{\pi_0}+\frac{1}{2}(\mu_1+\mu_0)^T\sum^{-1}(\mu_1-\mu_0)$

$\beta=(\mu_1-\mu_0)^T\sum^{-1}$ โดย π_1 และ π_0 เป็นความน่าจะเป็นก่อนของการจำแนกกลุ่ม 1 และกลุ่ม 0 ในทางปฏิบัติจะไม่ทราบค่าพารามิเตอร์ π_1,π_0,μ_1,μ_0 และ $\sum$ ซึ่งจะต้องแทนค่าด้วยค่าที่ประมาณได้จากตัวอย่าง

$$\hat{\pi}_1=\frac{n_1}{n},\hat{\pi}_0=\frac{n_0}{n}\ ,\ \hat{\mu}_1=\bar{x}_1=\frac{1}{n_1}\sum_{Y_i=1}x_i\ ,\ \hat{\mu}_0=\bar{x}_0=\frac{1}{n_0}\sum_{Y_i=0}x_i$$

$$\hat{\Sigma}=\left[\sum_{Y_i=1}(x_i-\bar{x}_1)(x_i-\bar{x}_1)^T+\sum_{Y_i=0}(x_i-\bar{x}_0)(x_i-\bar{x}_0)^T\right]/n$$

เป็นรูปแบบเดียวกับวิธีวิเคราะห์การถดถอยโลจีสติค ดังนั้นวิธีวิเคราะห์การถดถอยโลจีสติคและวิธีวิเคราะห์จำแนกกลุ่มในเชิงเส้นตรงมีรูปแบบของฟังก์ชันที่ไม่แตกต่างกัน แต่จะแตกต่างกันในส่วนของวิธีการประมาณค่าสัมประสิทธิ์การถดถอยเท่านั้น

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

วิธีสร้างสมการจำแนกกลุ่มมี 2 วิธี คือ วิธีตรง (Direct Method) และวิธีขั้นตอน (Stepwise Method) วิธีตรงเป็นวิธีที่ผู้วิจัยต้องการใช้ตัวแปรทุกตัวในการวิเคราะห์โดยไม่ต้องการดูผลจากแต่ละขั้นตอนจึงเป็นวิธีที่ประหยัดเวลา ส่วนวิธีขั้นตอนเป็นวิธีที่เลือกตัวแปรทีละตัวเข้าสมการโดยจะหาตัวแปรที่ดีที่สุดในการจำแนกมาเข้าสมการเป็นตัวแรกจากนั้นหาตัวแปรที่ดีเป็นลำดับที่สองเข้าสมการเพื่อปรับปรุงแก้ไขทำให้สมการสามารถจำแนกได้ดีขึ้น ในขั้นตอนต่อๆไปจะหาตัวแปรที่ดีที่สุดจากตัวแปรที่เหลือมาเข้าสมการในแต่ละขั้นตอนโดยตัวแปรที่ถูกคัดเลือกก่อนหน้านี้อาจจะถูกเลือกออกไปหากพบว่าเมื่อนำมารวมกับตัวแปรอื่นๆแล้วไม่ช่วยให้สมการสามารถทำนายสมาชิกกลุ่มได้ดีขึ้น จึงอาจกล่าวได้ว่าการคัดเลือกตัวแปรที่ละขั้นตอนมีความคล้ายคลึงกับวิธีวิเคราะห์การถดถอยเชิงพหุแบบขั้นตอน โดยเทคนิคที่ใช้ในการคัดเลือกตัวแปรมีหลายวิธีแต่ที่นิยมใช้มี 5 วิธี ดังนี้

1) วิธี Wilk เป็นวิธีที่พิจารณาจากค่าสถิติ F ของตัวแปรทั้งหมด (Overall Multivariate F Ratio) ในการทดสอบความแตกต่างระหว่างกลุ่ม ตัวแปรที่ให้ค่าสถิติ F มากที่สุด

จะทำให้ค่า Wilk's Lambda (เป็นมาตรวัดความแตกต่าง) มีค่าลดลง ซึ่งหาก Wilk's Lambda มีค่าน้อยแสดงว่าสมการที่ได้สามารถทำนายสมาชิกกลุ่มได้ดี

2) วิธี Maha เป็นวิธีที่ทำให้ระยะห่างของความแตกต่างระหว่างกลุ่มที่ใกล้กันที่สุดมีค่ามากที่สุด

3) วิธี Maximin F เป็นวิธีที่เลือกตัวแปรจากค่าสถิติ F ระหว่างกลุ่มที่มีความแตกต่างกันมากที่สุดมีค่าน้อยที่สุด

4) วิธี MINRESID เป็นวิธีที่พยายามแยกกลุ่มที่ใกล้ชิดกันให้มีค่าแตกต่างกันโดยพยายามทำให้ค่าความผันแปรที่เหลือมีค่าน้อยที่สุด

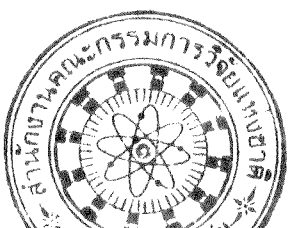
5) วิธี Rao เป็นวิธีที่วัดระยะห่างของสมการทั่วไป (Generalized Distance Measure) ตัวแปรที่ได้รับการคัดเลือกจะเป็นตัวแปรที่ให้ระยะห่างมากที่สุด วิธีนี้จะทำให้มีความแตกต่างระหว่างกลุ่มมากที่สุด

ในการเลือกตัวแปรเข้าสมการจำแนกประเภทไม่ว่าจะใช้วิธีใดก็ตามจะใช้หลักเกณฑ์เดียวกันกล่าวคือตัวแปรจะได้รับการคัดเลือกก็ต่อเมื่อค่าสถิติ F เชิงส่วน (Partial Multivariate F Ratio) มีค่ามากกว่าที่กำหนดไว้

2.3 วิธีริดจ์รีเกรสชัน (Ridge Regression)

เป็นวิธีประมาณค่าสัมประสิทธิ์การถดถอยพหุคูณในกรณีที่ตัวแปรที่ใช้ทำนายมีความสัมพันธ์กันเนื่องจากค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองที่ได้ จากการประมาณค่าสัมประสิทธิ์การถดถอยพหุคูณด้วยวิธีกำลังสองน้อยที่สุด (Least Square Method)มีค่าเท่ากับ $\sigma^2(X'X)^{-1}$ ดังนั้นการลดค่าเฉลี่ยความคลาดเคลื่อนกำลังสองให้มีค่าต่ำลงจึงต้องลดค่า $(X'X)^{-1}$ ให้มีค่าต่ำลง Hoerl และ Kennard (1970) ได้เสนอวิธีประมาณค่าสัมประสิทธิ์การถดถอยพหุคูณที่ให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองต่ำสุดโดยบวกค่าคงที่กับสมาชิกทุกตัวบนเส้นทแยงมุมของเมทริกซ์ $X'X$ ซึ่งได้ค่าประมาณสัมประสิทธิ์การถดถอยด้วยวิธีริดจ์รีเกรสชัน $\hat{\beta}_R = (X'X + kI)^{-1}X'Y$ ซึ่งต้องกำหนดให้ k มีค่าเพิ่มขึ้นทีละน้อย จนกว่าจะได้ค่า k ที่เหมาะสม ทุกครั้งที่กำหนดค่า k จะนำค่าเฉลี่ยความคลาดเคลื่อนกำลังสองที่ได้จากวิธีริดจ์รีเกรสชันและค่าเฉลี่ยความคลาดเคลื่อนกำลังสองที่คำนวณจากวิธีกำลังสองน้อยที่สุดมาเปรียบเทียบกัน คำนวณเช่นนี้เรื่อยไปจนกว่าจะได้ค่า k ซึ่งให้ค่าเฉลี่ยความคลาดเคลื่อนกำลังสองที่ได้จากวิธีริดจ์รีเกรสชันน้อยกว่าวิธีกำลังสองน้อยที่สุด ซึ่งในการศึกษาได้กำหนดค่า k เริ่มต้น 2 วิธีดังนี้

1) วิธีของ Hoerl, Kennard และ Baldwin (1975) ได้คิดค่า k เริ่มต้นที่เหมาะสมคือ $k = \frac{p\hat{\sigma}^2}{\hat{\beta}'\hat{\beta}}$ โดย p เป็นจำนวนตัวแปรอิสระ $\hat{\beta}$ และ $\hat{\sigma}^2$ เป็นค่าประมาณซึ่งได้จากวิธีกำลังสองน้อยที่สุด จากนั้นคำนวณค่า $k_1 = \frac{p\hat{\sigma}^2}{\hat{\beta}'_R(k_0)\hat{\beta}_R(k_0)}$ $k_2 = \frac{p\hat{\sigma}^2}{\hat{\beta}'_R(k_1)\hat{\beta}_R(k_1)}$



สำนักงานคณะกรรมการวิจัยแห่งชาติ
ห้องสมุดงานวิจัย
วันที่ 17 ส.ค. 2555
243307

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

2) วิธีของ 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 $X'X$ จากการที่ยอมให้เกิดความเอนเอียงในการประมาณ ค่าสัมประสิทธิ์การถดถอยพหุคูณจึงทำให้ค่าเฉลี่ยของความคลาดเคลื่อนกำลังสองของค่าประมาณสัมประสิทธิ์การถดถอยพหุคูณที่ได้จากวิธีริดจ์รีเกรสชันมีค่าน้อยกว่าค่าความแปรปรวนของค่าประมาณสัมประสิทธิ์การถดถอยพหุคูณจากวิธีกำลังสองน้อยที่สุด

2.4 วิธีวิเคราะห์การถดถอยพหุคูณ

โดยทั่วไปในการทำนายการเป็นสมาชิกกลุ่มของตัวแปรตอบสนองที่มี 2 ค่า ผู้วิจัยอาจใช้วิธีวิเคราะห์การถดถอยพหุคูณ ในการสร้างตัวแบบความน่าจะเป็นเชิงเส้น (Menard, 1995: 98) แสดงความสัมพันธ์ระหว่างตัวแปรตอบสนองและตัวแปรที่ใช้ทำนาย 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 หน่วย ในขณะที่ตัวแปรที่ใช้ทำนายตัวอื่นๆมีค่าคงที่โดยมีข้อสมมติว่าความคลาดเคลื่อนมีการแจกแจงปกติที่มีค่าเฉลี่ยเท่ากับศูนย์และมีค่าความแปรปรวนคงที่ 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) แต่ในการประมาณค่าความน่าจะเป็นด้วยวิธีนี้อาจจะให้ค่าประมาณที่อยู่นอกช่วง (0,1)ได้ ซึ่งอาจเป็นเพราะความสัมพันธ์ระหว่างตัวแปรตอบสนองและตัวแปรที่ใช้ในการทำนายไม่เป็นเส้นตรง (Hocking, 1985: 127) Samprit Chatterjee,Ali S. Hadi (2006) ได้ใช้วิธีวิเคราะห์การถดถอยพหุคูณและวิธีวิเคราะห์การถดถอย-โลจีสติคในการพยากรณ์การล้มละลายของบริษัท ซึ่งพบว่าวิธีวิเคราะห์การถดถอยโลจีสติคให้ผลการจำแนกกลุ่มที่ถูกต้องมากกว่าวิธีวิเคราะห์การถดถอยพหุคูณ