บทที่ 1

บทนำ

1.1 ความเป็นมาและความสำคัญของปัญหา

ในปัจจุบันการนำความรู้ทางสถิติไปประยุกต์ใช้กับงานในสาขาต่างๆ ได้รับความนิยมและมีผู้ให้ความสำคัญมากขึ้น ทั้งนี้เพราะวิธีการทางสถิติเป็นระเบียบวิธีการดำเนินการอย่างมีระบบภายใต้เหตุและผล ดังนั้นการจะนำวิธีทางสถิติมาประยุกต์ใช้นั้นจำเป็นต้องอาศัยวิธีที่เหมาะสมกับวัตถุประสงค์ของงานวิจัย และลักษณะของข้อมูลที่จะนำมาวิเคราะห์ของสาขาต่างๆเหล่านั้น สำหรับการนำสถิติมาใช้เพื่อวิเคราะห์หาความสัมพันธ์ของตัวแปรมีวิธีการหลายวิธี ซึ่งวิธีการทางสถิติที่นิยมใช้กันมากวิธีหนึ่งคือการวิเคราะห์การถดถอย (Regression Analysis)

การวิเคราะห์การถดถอยเป็นวิธีการทางสถิติอย่างหนึ่งที่ใช้ศึกษาถึงลักษณะความสัมพันธ์ระหว่างตัวแปรตั้งแต่ 2 ตัวขึ้นไป โดยพิจารณาถึงรูปแบบความเป็นไปได้ของความสัมพันธ์ที่มีเป้าหมายของการวิเคราะห์คือหาสมการทางคณิตศาสตร์ที่ใช้เป็นตัวแทนแสดงความสัมพันธ์ระหว่างตัวแปรตั้งแต่ 2 ตัวขึ้นไป เพื่อใช้ในการประมาณค่าหรือพยากรณ์ค่าของตัวแปรหนึ่งจากค่าที่กำหนดให้ของตัวแปรอื่นๆ ที่เกี่ยวข้องกัน ตัวแปรในตัวแบบการถดถอยนี้สามารถแบ่งออกเป็น 2 ประเภท ประเภทแรกคือตัวแปรที่ต้องการประมาณหรือพยากรณ์เรียกว่าตัวแปรตาม (Dependent Variable) หรือตัวแปรตอบสนอง (Response Variable) ตัวแปรประเภทที่สองคือตัวแปรที่มีอิทธิพลต่อตัวแปรตาม เรียกว่าตัวแปรอธิบาย (Explanatory Variable) หรือตัวแปรอิสระ (independent variable) การวิเคราะห์การถดถอยประกอบด้วยขั้นตอนต่างๆ หลายขั้นตอน เช่น การคัดเลือกตัวแปรอธิบายที่มีอิทธิพลต่อการอธิบายความแปรผันของตัวแปรตาม การประมาณค่าพารามิเตอร์ เป็นต้น

การวิเคราะห์การถดถอยจะต้องทำภายใต้ข้อสมมุติ (Assumptions) เกี่ยวกับค่าคลาดเคลื่อนสุ่ม (Random Error) ดังนั้นเมื่อได้ตัวแบบการถดถอยแล้วจึงจำเป็นอย่างยิ่งที่จะต้องตรวจสอบค่าคลาดเคลื่อนสุ่มว่าสอดคล้องกับสมมุติฐานเบื้องต้นหรือไม่ เนื่องจากไม่อาจทราบค่าคลาดเคลื่อนสุ่มจริงได้ ดังนั้นจึงใช้ส่วนตกค้าง (Residuals) ที่เป็นค่าประมาณของค่าคลาดเคลื่อนสุ่มมาตรวจสอบแทน บ่อยครั้งที่ผลการตรวจสอบเหล่านี้ถูกนำไปใช้ในการพิจารณา

ปรับหาตัวแบบการถดถอยใหม่เพื่อให้ได้ตัวแบบที่เหมาะสมยิ่งขึ้น ตัวแบบการถดถอยเชิงเส้นมีรูปแบบดังนี้

$$Y = E(Y \mid X = \underline{x}_i) + \varepsilon = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i \quad ; \ i = 1,2,...,n$$

โดยที่ $E(Y \mid X = \underline{x}_i) = E(Y \mid x_{i1}, x_{i2},..., x_{ik})$ เป็นค่าคาดหวังของตัวแปรตามเมื่อกำหนดตัวแปรอธิบายต่างๆ

x_{ij} เป็นค่าของตัวแปรอธิบาย

β_0, β_j เป็นพารามิเตอร์ของตัวแบบ

ε_i เป็นความคลาดเคลื่อนสุ่ม มีสมมติฐานเบื้องต้นคือ การแจกแจงแบบปรกติที่มีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนคงที่

จากที่ได้กล่าวไปแล้วนั้นเป็นการวิเคราะห์การถดถอยเชิงเส้นที่ตัวแปรตามเป็นตัวแปรเชิงปริมาณ แต่ในกรณีที่ตัวแปรตามเป็นตัวแปรเชิงคุณภาพที่จำแนกข้อมูลออกเป็นสองกลุ่ม เรียกว่าตัวแปรทวิภาค (Binary Variable) โดยที่กำหนดให้มีค่าเป็น 1 เมื่อตัวแปรตามเกิดเหตุการณ์ที่สนใจและมีค่าเป็น 0 ในกรณีอื่นๆ วิธีการทางสถิติที่ใช้หารูปแบบความสัมพันธ์ระหว่างตัวแปรตามทวิภาคนี้กับตัวแปรอธิบายอื่นๆ ที่อาจเป็นตัวแปรเชิงคุณภาพหรือเชิงปริมาณจะเรียกว่า การวิเคราะห์การถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Analysis) โดยมีวัตถุประสงค์ในการวิเคราะห์คือการพยากรณ์ความน่าจะเป็นของสิ่งที่เราสนใจและไม่สนใจ ความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอธิบายเหล่านี้สามารถแสดงในรูปของตัวแบบการถดถอยโลจิสติกทวิภาคซึ่งมีรูปแบบดังนี้

$$logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \quad ; \ i = 1,2,...,n$$

เมื่อ $\pi_i = P(Y_i = 1 \mid X = \underline{x}_i)$ แทนความน่าจะเป็นแบบมีเงื่อนไขเมื่อตัวแปรตามเกิดเหตุการณ์ที่สนใจ

β_0, β_j คือพารามิเตอร์ของตัวแบบ

Y_i คือตัวแปรตามแบบทวิภาค

x_{ij} คือค่าของตัวแปรอธิบาย

บ่อยครั้งพบว่าตัวแปรตาม (Y) ในตัวแบบการถดถอยโลจิสติกทวิภาคเป็นตัวแปรที่แปลงมาจากตัวแปรต่อเนื่อง เช่น การระบุว่าคนจะเป็นโรคเบาหวานหรือไม่ พิจารณาได้จากระดับน้ำตาลในเลือด ตัวแปรที่สนใจคือการเป็นหรือไม่เป็นโรคเบาหวาน ซึ่งเป็นตัวแปรที่แปลงมาจากตัวแปรต่อเนื่องคือ ระดับน้ำตาลในเลือด เป็นต้น โดยตัวแปรต่อเนื่องนี้มีชื่อเรียกว่าตัวแปรแฝง (Latent Variable) และเรียกตัวแบบการถดถอยโลจิสติกทวิภาคที่นำมาใช้วิเคราะห์ข้อมูลลักษณะดังกล่าวว่าตัวแบบตัวแปรแฝง (Latent Variable Model)

เมื่อได้ตัวแบบการถดถอยแล้ว ขั้นตอนต่อมาคือการตรวจสอบความเหมาะสมของตัวแบบ ค่าสัมประสิทธิ์การตัดสินใจ (Coefficient of Determination หรือ R^2) เป็นค่าที่ใช้กันอย่างแพร่หลายในการตรวจสอบความเหมาะสมของตัวแบบถดถอย และสำหรับตัวแบบการถดถอยโลจิสติกนั้น ได้มีผู้คิดวิธีการคำนวณค่า R^2 ที่คล้ายกับ ค่า R^2 ของสมการถดถอยในกรณีที่ตัวแปรตามเป็นตัวแปรต่อเนื่องไว้ด้วยกันหลายค่า โดย ควาลซีท (Kvalseth, 1985) ได้เสนอคุณสมบัติที่ดีของค่า R^2 ของการวิเคราะห์การถดถอยเชิงเส้นไว้ 8 ประการ โดยเฉพาะคุณสมบัติที่ว่า ค่า R^2 ที่ดีควรจะสามารถใช้เปรียบเทียบความเหมาะสมของตัวแบบต่างๆ ที่ได้มาจากข้อมูลชุดเดียวกันได้ และต่อมา มินาร์ด (Menard, 2000) ได้แนะนำเพิ่มเติมว่า ค่า R^2 ที่ได้ควรจะสามารถใช้เปรียบเทียบตัวแบบต่างๆที่ได้มาไม่ว่าตัวแบบนั้นจะมีตัวแปรอธิบาย หรือ ตัวแปรตามที่ต่างกัน หรือ แม้กระทั่งตัวแบบที่ได้มาจากข้อมูลที่ต่างกัน

อย่างไรก็ตาม มีนักวิจัยหลายท่านพบว่า ค่า R^2 ของตัวแบบการถดถอยโลจิสติกทวิภาคหลายค่าที่ขึ้นอยู่กับสัดส่วนการเกิดเหตุการณ์ที่สนใจภายใต้ขนาดตัวอย่างที่ศึกษาหรือสัดส่วนผลตอบสนอง (Response Proportion) เช่น โซเดอร์สตรอม (Soderstrom, 1997) และชาร์มา (Sharma, 2006) เป็นต้น การที่ R^2 ขึ้นอยู่กับ สัดส่วนผลตอบสนอง ทำให้ไม่สามารถที่จะใช้ R^2 เปรียบเทียบความเหมาะสมของตัวแบบที่เหมือนกันหรือต่างกันที่ได้มาจากข้อมูลที่ต่างชุดกัน หรือมีตัวแปรตามที่ต่างกัน ปัญหาของสัดส่วนผลตอบสนองสำหรับตัวแบบการถดถอยโลจิสติกสามารถอธิบายได้ดังต่อไปนี้ พิจารณาข้อมูล $y_i^{(j)}, x_{i1}^{(j)}, x_{i2}^{(j)}, ..., x_{ik}^{(j)}$ ($i = 1,2,..., n_{(j)}$ และ $n_{(j)}$ คือขนาดตัวอย่างจากกลุ่มที่ j โดยที่ $j = 1,2$) ที่รวบรวมจากตัวแปรตาม Y และตัวแปรอธิบาย $X_1, ..., X_k$ โดยที่ตัวห้อย $j = 1,2$ อาจหมายถึงข้อมูลที่แตกต่างกัน 2 ชุด หรืออาจหมายถึง ตัวแปรตามที่แตกต่างกันแต่มีความเกี่ยวข้องกัน หรือ อาจจะหมายถึงข้อมูลชุดย่อย 2 ชุดที่แตกต่างกันในชุดข้อมูล เมื่อกำหนดให้ $\hat{\pi}_1 \neq \hat{\pi}_2$ เป็นค่าประมาณของ สัดส่วนผลตอบสนองของตัวอย่างกลุ่มที่ 1 และ 2 ตามลำดับ ดังนั้นถ้านักวิจัยสร้างตัวแบบสำหรับกลุ่ม 2 กลุ่มได้ดังนี้

$$g^{(1)}(X) = \beta_0 + \beta'^{(1)}X \quad \text{สำหรับกลุ่มที่ 1 และ}$$

$$g^{(2)}(X) = \beta_0 + \beta'^{(2)}X \quad \text{สำหรับกลุ่มที่ 2}$$

และคำนวณค่า $R^2_{(1)}$ สำหรับกลุ่มที่ 1 และ $R^2_{(2)}$ สำหรับกลุ่มที่ 2 และใช้เป็นค่าที่จะเปรียบเทียบความเหมาะสมของตัวแบบ ถ้าพบว่า $R^2_{(1)} > R^2_{(2)}$ นักวิจัยสามารถทำการสรุปของผลลัพธ์ที่ได้ดังนี้

(1) ตัวแปรอธิบาย $\{X_1,..., X_k\}$ สามารถอธิบายความผันแปรที่เกิดขึ้นในตัวแปรตามในกลุ่มที่ 1 ได้ดีกว่าในกลุ่มที่ 2 หรือ

(2) ความแตกต่างระหว่าง $R^2_{(1)}$ และ $R^2_{(2)}$ อาจเนื่องมาจากความแตกต่างระหว่างค่าสัดส่วนผลตอบสนองของกลุ่มทั้งสอง

ดังนั้นการเปรียบเทียบตัวแบบที่ตัวแปรตามที่มีค่าสัดส่วนผลตอบสนองต่างกัน จึงไม่สามารถสรุปได้ว่า ตัวแบบใดสามารถอธิบายความผันแปรที่เกิดขึ้นในตัวแปรตามได้ดีกว่ากัน เนื่องจากมีปัญหาของสัดส่วนผลตอบสนองเข้ามาเกี่ยวข้อง

นอกจากสัดส่วนผลตอบสนอง แล้วพบว่ายังมีปัจจัยอีกหลายๆ ปัจจัยที่มีอิทธิพล R^2 เช่น ความเชื่อถือได้ในตัวแปรอธิบาย และ/หรือ ตัวแปรตาม (Reliability of Explained Variables and/or Dependent Variable) เช่น จากการศึกษาของก็อตเฟรดสัน (Gottfredson, 1987) พบว่าระดับความเชื่อถือได้ในตัวแปรอธิบายและตัวแปรตามมีอิทธิพลต่อค่า R^2 ของตัวแบบการถดถอยโลจิสติกทั้งนี้เพราะตัวแปรในตัวแบบเป็นสาเหตุสำคัญ ที่ทำให้ค่าความคลาดเคลื่อนในการพยากรณ์มีค่าสูงหรือต่ำ ถ้าตัวแบบมีความคลาดเคลื่อนสูงอาจแก้ปัญหาโดยการเพิ่มขนาดตัวอย่างซึ่งทำให้มีผลกระทบหลายอย่างตามมา ดังนั้นตัวแปรที่ดีต้องมีระดับความเชื่อถือได้ที่มีค่าสูงๆ และโซเดอร์สตรอม (Soderstrom, 1997) ได้ศึกษาเกี่ยวกับอิทธิพลของระดับความเชื่อถือได้ในตัวแปรอธิบาย ที่มีต่อดัชนีของประสิทธิภาพในการทำนายสำหรับตัวแบบการถดถอยโลจิสติกซึ่งตัวแปรอธิบายในตัวแบบจะกำหนดให้มีระดับความเชื่อถือได้ 2 ระดับ คือ ระดับสูง (0.9) และระดับต่ำ (0.6) พบว่าการกำหนดความเชื่อถือได้ระดับสูงในตัวแปรอธิบายทำให้ดัชนีของประสิทธิภาพในการทำนายมีประสิทธิภาพดีกว่าการกำหนดความเชื่อถือได้ระดับต่ำ

อัตราการจำแนกผิดในตัวแปรอธิบาย และ/หรือ ตัวแปรตาม (Misclassification of Explained Variables and/or Dependent Variable) ก็ถือเป็นอีกปัจจัยที่มีอิทธิพลต่อ R^2 เช่น จากการศึกษาของ ฮัสแมน (Hausman, 1998) พบว่า อัตราการจำแนกผิดในตัวแปรตามมีผลต่อ

ประสิทธิภาพในการพยากรณ์คือ ทำให้ค่าสัมประสิทธิ์ในการพยากรณ์เกิดความเอนเอียง (Bias) และไม่คงเส้นคงวา (Inconsistent)

การใช้ R^2 ในการวัดสัดส่วนที่อธิบายความผันแปรในตัวแปรตามด้วยตัวแปรอธิบายนั้น บางครั้งจะให้ขนาดของความสัมพันธ์มากเกินไป แม้ว่าตัวแปรอธิบายที่เพิ่มเข้าไปในตัวแบบนั้นจะไม่มีความสัมพันธ์กับตัวแปรตาม ดังนั้นสำหรับตัวแบบการถดถอยโลจิสติกที่ประกอบด้วยตัวแปรอธิบายที่มีมากกว่าหนึ่งตัวแปรจึงจำเป็นต้องมีการปรับค่า R^2 ที่เรียกว่าสัมประสิทธิ์การตัดสินใจที่ปรับค่า (Adjusted R^2) หรือ R^2_{adj} การปรับค่าสัมประสิทธิ์การตัดสินใจนั้น โดยส่วนใหญ่จะใช้วิธีการปรับ 2 แบบด้วยกัน แบบแรกจะใช้จำนวนตัวแปรอธิบายที่มีในตัวแบบและขนาดตัวอย่างมาช่วยในการคำนวณ หรือเรียกว่าการใช้องศาเสรี (Degrees of Freedom) และแบบที่สองใช้ผลต่างของค่าฟังก์ชันควรจะเป็นของตัวแบบ 2 ตัวแบบ หรือที่เรียกว่าสถิติดีเวียนซ์ (Deviance) ได้มีผู้เสนอค่า R^2_{adj} ของการถดถอยโลจีสติกไว้หลายท่าน เช่น มิทเบิกค์และสเชมเปอร์ (Mittlbock and Schemper, 1996) ได้เสนอค่า R^2 ที่ปรับค่าโดยการใช้รูปแบบทั้งสองที่กล่าวข้างต้น ผลการศึกษาพบว่าค่า R^2_{adj} ทั้งสองตัวนี้ให้ผลเป็นที่น่าพอใจกว่าค่า R^2 ที่ยังไม่ได้ทำการปรับค่า ชเตทแลนด์ (Shtatland, 2000) ได้เสนอ R^2_{adj} ที่ปรับค่าด้วยหลักเกณฑ์สารสนเทศของ AIC เมื่อคำนวณโดยใช้วิธีความควรจะเป็นสูงสุดที่ใช้ใน PROC GENMOD ของโปรแกรม $SAS^{®}$ เลียโอะและแมคกี (Liao and McGee, 2003) ได้เสนอ R^2_{adj} ของการถดถอยโลจิสติกเพื่อใช้แก้ปัญหาของ R^2 ที่ให้ขนาดความสัมพันธ์ที่มากเกินไป และมีค่าสูงขึ้นเมื่อขนาดตัวอย่างลดลง โดยมีการปรับค่า R^2 ด้วยค่าคลาดเคลื่อนที่คำนวณโดยใช้ขั้นตอน 2 ขั้นตอนคือ ขั้นที่ 1 คำนวณค่าคลาดเคลื่อนของการพยากรณ์ โดยจะเรียกค่าคลาดเคลื่อนนี้ว่าค่าคลาดเคลื่อนของการพยากรณ์ที่ไม่สามารถแยกได้ (Inherent Prediction Error, IPE) ของตัวแบบโลจิสติก ขั้นที่ 2 ปรับค่าคลาดเคลื่อนของการพยากรณ์ที่ได้ในขั้นที่ 1 โดยใช้หลักการเช่นเดียวกับการประมาณค่าความแปรปรวน $\left(\sigma^2\right)$ ในตัวแบบการถดถอยด้วยวิธี $REML$ (Restricted or Residual Maximum Likelihood) เป็นต้น

อย่างไรก็ตามยังไม่มีงานวิจัยที่ศึกษาอิทธิพลของสัดส่วนผลตอบสนอง ระดับความเชื่อถือได้ในตัวแปรอธิบาย และอัตราการจำแนกผิดในตัวแปรตาม ที่มีต่อ R^2_{adj} ทั้ง 5 ค่าที่เสนอข้างต้น สำหรับตัวแบบตัวแปรแฝงของการถดถอยโลจิสติกทวิภาค ดังนั้นในงานวิจัยนี้ผู้วิจัยจึงสนใจศึกษาอิทธิพลของสัดส่วนผลตอบสนอง ระดับความเชื่อถือได้ในตัวแปรอธิบาย และอัตราการจำแนกผิดในตัวแปรตามที่มีต่อ R^2_{adj} ภายใต้ตัวแบบตัวแปรแฝงของการถดถอยโลจิสติกทวิภาค นอกจากนี้ผู้วิจัยยังสนใจศึกษาปฏิสัมพันธ์ 2 ทาง (Two-way Interaction) ระหว่าง

สัดส่วนผลตอบสนอง ระดับความเชื่อถือได้ในตัวแปรอธิบาย และอัตราการจำแนกผิดในตัวแปรตาม กับขนาดตัวอย่าง จำนวนตัวแปรอธิบายในตัวแบบ หรือปัจจัยอื่นที่กำหนดเป็นเงื่อนไขของ R^2_{adj} ภายใต้ตัวแบบตัวแปรแฝงของการถดถอยโลจิสติกทวิภาค

1.2 วัตถุประสงค์ของการศึกษา

(1) เพื่อศึกษาระดับความสัมพันธ์ระหว่างสัดส่วนผลตอบสนอง ระดับความเชื่อถือได้ในตัวแปรอธิบาย และอัตราการจำแนกผิดในตัวแปรตามกับค่า R^2_{adj} สำหรับตัวแบบตัวแปรแฝงของการถดถอยโลจิสติกทวิภาค

(2) เพื่อศึกษาปฏิสัมพันธ์ 2 ทางระหว่างสัดส่วนผลตอบสนอง ระดับความเชื่อถือได้ในตัวแปรอธิบาย อัตราการจำแนกผิดในตัวแปรตามกับจำนวนตัวแปรอธิบาย ขนาดตัวอย่าง และสัมประสิทธิ์การตัดสินใจที่แท้จริงที่กำหนดเป็นเงื่อนไขของ R^2_{adj} สำหรับตัวแบบตัวแปรแฝงของการถดถอยโลจิสติกทวิภาค

(3) เพื่อศึกษาเปรียบเทียบความเอนเอียงของ R^2_{adj} สำหรับตัวแบบตัวแปรแฝงของการถดถอยโลจิสติกทวิภาคภายใต้สัดส่วนผลตอบสนอง ระดับความเชื่อถือได้ในตัวแปรอธิบาย อัตราการจำแนกผิดในตัวแปรตาม จำนวนตัวแปรอธิบาย ขนาดตัวอย่าง และสัมประสิทธิ์การตัดสินใจที่แท้จริงที่กำหนดเป็นเงื่อนไขของ R^2_{adj} สำหรับตัวแบบตัวแปรแฝงของการถดถอยโลจิสติกทวิภาค

1.3 สมมุติฐานของการวิจัย

(1) ระดับความสัมพันธ์ระหว่างสัดส่วนผลตอบสนอง ระดับความเชื่อถือได้ในตัวแปรอธิบาย และอัตราการจำแนกผิดในตัวแปรตามกับค่า R^2_{adj} สำหรับตัวแบบตัวแปรแฝงของการถดถอยโลจิสติกทวิภาคแตกต่างกันขึ้นอยู่กับประเภทของ R^2_{adj}

(2) มีปฏิสัมพันธ์ 2 ทางระหว่างสัดส่วนผลตอบสนอง ระดับความเชื่อถือได้ในตัวแปรอธิบาย และอัตราการจำแนกผิดในตัวแปรตามกับจำนวนตัวแปรอธิบาย ขนาดตัวอย่าง และสัมประสิทธิ์การตัดสินใจที่แท้จริงที่กำหนดเป็นเงื่อนไขของ R^2_{adj} สำหรับตัวแบบตัวแปรแฝงของการถดถอยโลจิสติกทวิภาค

(3) R^2_{adj} สำหรับตัวแบบตัวแปรแฝงของการถดถอยโลจิสติกทวิภาค ที่คำนวณโดยใช้หลักการเดียวกันมีความเอนเอียงไม่แตกต่างกัน

1.4 ขอบเขตการศึกษา

งานวิจัยนี้เป็นการศึกษาโดยใช้การจำลองข้อมูลด้วยเทคนิคมอนติคาร์โล ที่มีขอบเขตการศึกษาดังนี้

(1) กรณีศึกษาอิทธิพลของสัดส่วนผลตอบสนอง

(1.1) กำหนดสัดส่วนผลตอบสนอง (π) 3 ระดับ คือ 0.1 , 0.3 และ 0.5

(1.2) กำหนดค่าสัมประสิทธิ์การตัดสินใจที่แท้จริง $\left(R^2_{true}\right)$ 3 ระดับ คือ 0.1 , 0.5 และ 0.9

(1.3) กำหนดตัวแปรอธิบาย X_j เมื่อ $j = 1,2,...,10$ โดยมีการแจกแจงแบบปรกติที่มีอิสระต่อกัน ดังต่อไปนี้

$X_1 \sim N(2,4)$, $X_2 \sim N(0,1)$, $X_3 \sim N(-2,2)$, $X_4 \sim N(3,3)$, $X_5 \sim N(1,4)$, $X_6 \sim N(-4,5)$, $X_7 \sim N(0,6)$, $X_8 \sim N(-12,7)$, $X_9 \sim N(-6,8)$ และ $X_{10} \sim N(15,9)$

(2) กรณีศึกษาอิทธิพลของความเชื่อถือได้ในตัวแปรอธิบาย

(2.1) กำหนดระดับความเชื่อถือได้ในตัวแปรอธิบาย (γ) 4 ระดับ คือ 0.30 0.50 0.70 และ 0.90

(2.2) กำหนดสัดส่วนผลตอบสนอง (π) ให้มีค่าเท่ากับ 0.5

(2.3) กำหนดค่าสัมประสิทธิ์การตัดสินใจที่แท้จริง $\left(R^2\right)$ ให้มีค่าเท่ากับ 0.9

(2.4) กำหนดตัวแปรอธิบายให้มีการแจกแจงปรกติ ที่มีค่าเฉลี่ย เช่นเดียวกับ (1.3) แต่ความแปรปรวนจะเป็นค่าที่คำนวณขึ้นภายใต้เงื่อนไขของระดับความเชื่อถือได้ โดยจะกล่าวถึงในบทที่ 3

(3) กรณีศึกษาอิทธิพลของการจำแนกผิดในตัวแปรตาม

(3.1) กำหนดอัตราการจำแนกผิดในตัวแปรตาม (α) 2 ระดับ คือ 2% และ 10% ของขนาดตัวอย่าง โดยลักษณะการจำแนกผิดจะแยกออกเป็น 3 ลักษณะคือ การจำแนกค่า 0 ผิดทั้งหมด การจำแนกค่า 1 ผิดทั้งหมด และการจำแนกค่า 0 และ 1 ผิดอย่างละครึ่ง

(3.2) กำหนดค่าของสัดส่วนผลตอบสนอง (π) และสัมประสิทธิ์การตัดสินใจที่แท้จริง $\left(R^2_{true}\right)$ เช่นเดียวกับ (2.1) และ (2.2)

(3.3) กำหนดตัวแปรอธิบายเช่นเดียวกับ (1.3)

สำหรับขอบเขตอื่นๆ ที่จะกล่าวต่อไปนี้ เป็นขอบเขตการวิจัยที่จะใช้ร่วมกันทั้งใน 3 การศึกษา

(2) จำลองชุดข้อมูลของตัวแปรแฝงจากตัวแบบตัวแปรแฝง (Latent Variable Model) ที่ประกอบด้วยจำนวนตัวแปรอธิบาย (k) 1 5 หรือ 10 ตัว

$$Y^* = a + b_j x_{ij} + \varepsilon \quad \text{เมื่อ} \quad i = 1,2,...,n \quad \text{และ} \quad j = 1,2,...,k$$

(3) กำหนดขนาดตัวอย่าง (n) ที่ทำการศึกษาคือ 50 250 500 และ 1000

(4) การประมาณค่าพารามิเตอร์ของสัมประสิทธิ์การถดถอยของตัวแบบการถดถอยโลจิสติกทวิภาคจะใช้วิธีความควรจะเป็นสูงสุดของฟิชเชอร์สกอริง (Fisher's Scoring Method) โดยใช้โปรแกรม $SAS^{®}$ เวอร์ชัน 9.0 ในการคำนวณ

(5) R^2_{adj} ที่นำมาศึกษาคือ R^2_{adj} ที่เสนอโดย มิทเบิกค์และสเชมเปอร์ (Mittlbock and Schemper, 1996) ชเตทแลนด์ (Shtatland, 2000) และ เลียโอะและแมคกี (Liao and McGee, 2003)

(6) ในการจำลองจะทำซ้ำ $1,000$ ครั้ง ในแต่ละตัวแบบการถดถอยภายใต้เงื่อนไขต่างๆ ที่กำหนดในการศึกษาครั้งนี้

(7) ในงานวิจัยนี้ ผู้วิจัยสนใจศึกษาเฉพาะ อิทธิพลของปฏิสัมพันธ์ 2 ทาง

1.5 เกณฑ์ที่ใช้ในการศึกษา

(1) กรณีศึกษาระดับความสัมพันธ์ระหว่างสัดส่วนผลตอบสนอง ระดับความเชื่อถือได้ในตัวแปรอธิบาย และอัตราการจำแนกผิดในตัวแปรตามกับค่า R^2_{adj} ที่นำมาศึกษา จะพิจารณาจากค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) ระหว่าง R^2_{adj} กับระดับต่างๆ ของสัดส่วนผลตอบสนอง ระดับความเชื่อถือได้ในตัวแปรอธิบาย และอัตราการจำแนกผิดในตัวแปรตาม

(2) กรณีศึกษาปฏิสัมพันธ์ 2 ทางระหว่างสัดส่วนผลตอบสนอง ระดับความเชื่อถือได้ในตัวแปรอธิบาย และอัตราการจำแนกผิดในตัวแปรตามกับเงื่อนไขอื่นๆ ของ R^2_{adj} สำหรับตัวแบบตัวแปรแฝงของการถดถอยโลจิสติกทวิภาคนั้น จะทดสอบความมีนัยสำคัญโดยใช้การวิเคราะห์ความแปรปรวน (Analysis of Variance) ที่ใช้ตัวสถิติทดสอบ F

(3) ศึกษาเปรียบเทียบความเอนเอียงของ R^2_{adj} จะพิจารณาเปรียบเทียบจาก

(3.1) ค่าเฉลี่ย (Mean, $\overline{R}^2_{est}$) ของค่าประมาณสัมประสิทธิ์การตัดสินใจที่ปรับค่า (R^2_{adj}) เมื่อ $m = 1,2,...,1000$ มีสูตรการคำนวณดังนี้

$$\overline{R}^2_{est} = \frac{\sum_{m=1}^{1000} R^2_{est,m}}{1{,}000} \tag{1.1}$$

เมื่อ $R^2_{est,m}$ คือค่าประมาณของ R^2_{adj} จากการคำนวณด้วยวิธีกำลังสองน้อยสุดและวิธีความควรจะเป็นสูงสุดตามลำดับที่ได้จากการทำซ้ำครั้งที่ m

(3.2) ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error, MSE) ของค่าประมาณสัมประสิทธิ์การตัดสินใจที่ปรับค่า (R^2_{adj}) เมื่อ $m = 1,2,...,1000$ มีสูตรการคำนวณดังนี้

$$MSE = \frac{\sum_{m=1}^{1000} \left(R^2_{true} - R^2_{est,m}\right)^2}{1000} \tag{1.2}$$

เมื่อ R^2_{true} คือค่าสัมประสิทธิ์การตัดสินใจที่แท้จริง และ $R^2_{est,m}$ คือค่าประมาณของ R^2_{adj} จากการคำนวณด้วยวิธีกำลังสองน้อยสุดและวิธีความควรจะเป็นสูงสุดตามลำดับที่ได้จากการทำซ้ำครั้งที่ m

1.6 ประโยชน์ที่คาดว่าจะได้รับ

(1) ทราบอิทธิพลของสัดส่วนผลตอบสนอง ระดับความเชื่อถือได้ในตัวแปรอธิบาย และอัตราการจำแนกผิดในตัวแปรตามที่มีต่อ R^2_{adj} สำหรับตัวแบบตัวแปรแฝงของการถดถอยโลจิสติกทวิภาค

(2) เป็นแนวทางในการเลือกใช้ R^2_{adj} สำหรับตัวแบบตัวแปรแฝงของการถดถอยโลจิสติกทวิภาคให้เหมาะสมกับสัดส่วนผลตอบสนอง ระดับความเชื่อถือได้ในตัวแปรอธิบาย และอัตราการจำแนกผิดในตัวแปรตาม

(3) เป็นแนวทางในการเลือกใช้ R^2_{adj} สำหรับตัวแบบตัวแปรแฝงของการถดถอยโลจิสติกทวิภาคให้เหมาะสมในแต่ละสถานการณ์ของขนาดตัวอย่างและจำนวนตัวแปรอธิบาย

1.7 สัญลักษณ์และคำจำกัดความที่ใช้ในงานวิจัย

R^2_{true} แทนค่าสัมประสิทธิ์การตัดสินใจที่แท้จริงของตัวแบบการถดถอยโลจิสติกทวิภาคที่กำหนดในการวิจัยครั้งนี้

$R^2_{O,adj,MS}$ แทนสัมประสิทธิ์การตัดสินใจที่ปรับค่าด้วยองศาเสรี เมื่อคำนวณโดยใช้วิธีกำลังสองน้อยสุดแบบสามัญ ที่ศึกษาในงานของมิทเบิกค์และสเชมเปอร์

$R^2_{l,adj,MS}$ แทนสัมประสิทธิ์การตัดสินใจที่ปรับค่าด้วยสถิติดีเวียนซ์ เมื่อคำนวณโดยใช้วิธีความควรจะเป็นสูงสุด ที่ศึกษาในงานของมิทเบิกค์และสเชมเปอร์

$R^2_{O,adj,LM}$ แทนสัมประสิทธิ์การตัดสินใจที่ปรับค่าของเลียโอะและแมคกี เมื่อคำนวณโดยใช้วิธีกำลังสองน้อยสุดแบบสามัญ

$R^2_{l,adj,LM}$ แทนสัมประสิทธิ์การตัดสินใจที่ปรับค่าของเลียโอะและแมคกี เมื่อคำนวณโดยใช้วิธีความควรจะเป็นสูงสุด

$R^2_{l,adj,SAS_{AIC}}$ แทนสัมประสิทธิ์การตัดสินใจที่ปรับค่าด้วยหลักเกณฑ์สารสนเทศของ AIC เมื่อคำนวณโดยใช้วิธีความควรจะเป็นสูงสุด ที่ใช้ใน PROC GENMOD ของโปรแกรม $SAS^{®}$

$R^2_{est,m}$ แทนค่าประมาณของสัมประสิทธิ์การตัดสินใจที่ปรับค่าจากการคำนวณด้วยวิธีกำลังสองน้อยสุดและวิธีความควรจะเป็นสูงสุด สำหรับตัวแบบตัวแปรแฝงของการถดถอยโลจิสติกทวิภาคที่ได้จากการทำซ้ำครั้งที่ m เมื่อ $m = 1,2,...,1000$

$\overline{R}^2_{est}$ แทนค่าเฉลี่ย (Mean) ของ $R^2_{est,m}$ จากการทำซ้ำจำนวน $1,000$ ครั้ง

MSE แทนค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (Mean Square Error, MSE) ของ $R^2_{est,m}$ จากการทำซ้ำจำนวน $1,000$ ครั้ง

ตัวแบบการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Model) คือ ตัวแบบที่ศึกษาถึงความสัมพันธ์ระหว่างตัวแปรอธิบายและตัวแปรตามแบบทวิภาคที่มีค่าเป็นไปได้ 2 ค่า คือ 1 หรือ 0

ตัวแปรแฝง (Latent Variable) คือ ตัวแปรที่เป็นตัวแทนของตัวแปรต่อเนื่องและทำการแปลงตัวแปรต่อเนื่องนั้นให้มีค่าที่เป็นไปได้ 2 ค่า คือ 1 หรือ 0

ตัวแบบตัวแปรแฝง (Latent Variable Model) คือ ตัวแบบที่ศึกษาถึงความสัมพันธ์ระหว่างตัวแปรอธิบายและตัวแปรแฝง

สัดส่วนผลตอบสนอง (Response Proportion) คือสัดส่วนของ $Y=1$ ต่อขนาดตัวอย่างทั้งหมด

ระดับความเชื่อถือได้ในตัวแปรอธิบาย (Reliability in Explained Variables) คือ ค่าที่ใช้แสดงถึงความน่าเชื่อถือของตัวแปรอธิบายในตัวแบบการถดถอยที่มีค่าอยู่ในช่วง (0,1)

อัตราการจำแนกผิดในตัวแปรตาม (Misclassification Rate in Dependent - Variable) คือค่าความน่าจะเป็นในการกำหนดให้ตัวแปรตามมีค่าเป็น 1 เมื่อค่าของตัวแปรตามที่แท้จริงมีค่าเป็น 0 หรือค่าความน่าจะเป็นในการกำหนดให้ตัวแปรตามมีค่าเป็น 0 เมื่อค่าของตัวแปรตามที่แท้จริงมีค่าเป็น 1