ผลกระทบของการกำหนดตัวแบบไม่ถูกต้อง ที่มีต่อตัววัด R^2 ของการถดถอยลอจิสติก

Misspecified Model Effects on R-squared Measures of Logistic Regression

แสงหล้า ชัยมงคล*

ภาควิชาคณิตศาสตร์และสถิติ คณะวิทยาศาสตร์และเทคโนโลยี มหาวิทยาลัยธรรมศาสตร์ ศูนย์รังสิต ตำบลคลองหนึ่ง อำเภอคลองหลวง จังหวัดปทุมธานี 12120

บทคัดย่อ

ในการวิจัยนี้ ผู้วิจัยศึกษาผลกระทบของการกำหนดตัวแบบไม่ถูกต้องที่มีต่อตัววัดสัมประสิทธิ์การตัดสินใจ (R^2) ของการถดถอยลอจิสติก โดยการกำหนดตัวแบบไม่ถูกต้องใน 3 ลักษณะ คือ (1) กำหนดรูปแบบของตัวแปรอิสระไม่ถูกต้อง (2) กำหนดฟังก์ชันเชื่อมโยงของตัวแบบไม่ถูกต้อง และ (3) กำหนดให้ตัวแบบการถดถอยมีตัวแปรอิสระขาดหายไป ตัววัด R^2 ที่นำมาพิจารณาคือตัววัดที่คำนวณด้วยหลักการของสัดส่วนการลดลงของการกระจาย (R^2_{OLS}, R^2_G) และ R^2 ที่คำนวณโดยอาศัยฟังก์ชันควรจะเป็น (R^2_L, R^2_M, R^2_N, R^2_C) ผลกระทบของการกำหนดตัวแบบไม่ถูกต้องจะประเมินเชิงตัวเลขที่ได้จากการจำลองข้อมูลโดยอาศัยเทคนิคมอนติคาร์โลที่กำหนดให้มีขนาดตัวอย่างเท่ากับ 100, 250, 500, และ 1000 และจำนวนตัวแปรอิสระเท่ากับ 1, 5, และ 10 ตัว การประเมินจะพิจารณาค่าความเอนเอียงสัมพัทธ์ และค่ารากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ย ผลการวิจัยสรุปได้ว่า การกำหนดตัวแบบไม่ถูกต้องทั้ง 3 รูปแบบ มีผลกระทบต่อ R^2_N อย่างมากและเป็นไปทิศทางที่ตรงกันข้ามกับ R^2 ตัวอื่นๆ โดยการกำหนดรูปแบบของตัวแปรอิสระไม่ถูกต้องและตัวแบบมีตัวแปรอิสระที่สำคัญขาดหายไปนั้นทำให้ R^2_N มีค่าความเอนเอียงสัมพัทธ์ และรากที่สองของความคลาดเคลื่อนกำลังสองเฉลี่ยสูง ในขณะที่ R^2 ตัวอื่นๆ นั้นจะเป็นตัวประมาณที่เอนเอียงเฉพาะในกรณีที่ตัวอย่างมีขนาดเล็ก และตัวแบบมีจำนวนตัวแปรอิสระน้อย สำหรับการกำหนดฟังก์ชันเชื่อมโยงไม่ถูกต้องในรูปฟังก์ชันโพรบิทนั้น มีผลกระทบต่อ R^2_N และ R^2_M เท่านั้น โดย R^2_N ให้ค่าความเอนเอียงสัมพัทธ์สูงกว่า R^2_M ในทางตรงกันข้ามกับการกำหนดฟังก์ชันเชื่อมโยงไม่ถูกต้องในรูปฟังก์ชันคอมพลีเมนทารีล็อก-ล็อกนั้นจะมีผลกระทบต่อ R^2 ทุกตัวที่นำมาศึกษา

คำสำคัญ : สัมประสิทธิ์การตัดสินใจเทียม การถดถอยลอจิสติก การกำหนดตัวแบบไม่ถูกต้อง

***ผู้รับผิดชอบบทความ :** saenglaplus@gmail.com

Abstract

In this study, we evaluated the three types of misspecified model on R-squared measures that have been suggested to measure the explained variation in the logistic regression models. The first type of misspecification is incorrect link function. The second type occurs when the explanatory variables are wrong functional form. The third type occurs when the important explanatory variables are omitted from the model. The R-squared measures based on proportional reduction in dispersion (R^2_{OLS}, R^2_G) and the R-squared measures based on likelihood function (R^2_L, R^2_M, R^2_N, R^2_C) have been explored. We report the results of a Monte Carlo simulation and evaluate the effects of misspecification based on the relative bias and root mean square error. We found that the R^2_N is substantially sensitive to all of types of misspecifications and it performs inversely of the other measures. The R^2_N is a bias estimate with highest relative bias on the misspecification of wrong functional form of covariate and with missing one covariate from the model. While the other measures are bias estimates only with small sample and a few covariate. For the misspecified link function of probit, the R^2_N yields higher relative bias than the R^2_M, while the other measures do not affect on this misspecification. Inversely, the complementary-log-log link function does affect on all R-squared measures under study.

Keywords: pseudo-R-squared measure, logistic regression, model misspecification

1. ที่มาและความสำคัญของปัญหา

การวิเคราะห์การถดถอยลอจิสติกทวิภาค (binary logistic regression analysis) หรือเรียกย่อๆ ว่า การวิเคราะห์การถดถอยลอจิสติก เป็นวิธีการทางสถิติที่ใช้ศึกษาความสัมพันธ์ระหว่างตัวแปรตามเชิงคุณภาพที่จำแนกข้อมูลออกเป็นสองกลุ่มที่เรียกว่าตัวแปรทวิภาค (binary responses variable) โดยที่กำหนดให้มีค่าเป็น 1 เมื่อตัวแปรตาม (Y_i) เกิดลักษณะที่สนใจและมีค่าเป็น 0 ในกรณีอื่นๆ กับตัวแปรอิสระอื่นๆ โดยที่ตัวแปรอิสระอาจเป็นตัวแปรเชิงคุณภาพหรือเชิงปริมาณก็ได้ เพื่อนำมาใช้ในการพยากรณ์ค่าความน่าจะเป็นของสิ่งที่เราสนใจ ความสัมพันธ์ระหว่างความน่าจะเป็นที่ตัวแปรตามเกิดลักษณะที่สนใจ π_i กับตัวแปรอิสระ x_{ij} เหล่านี้สามารถแสดงในรูปของตัวแบบการถดถอยลอจิสติกได้ดังนี้

$$\text{logit}(\pi_i)=\log\left(\frac{\pi_i}{1-\pi_i}\right)$$

$$=\beta_0+\sum_{j=1}^{p}\beta_j x_{ij} \qquad (1)$$

เมื่อ β_0, β_j คือ พารามิเตอร์ของตัวแบบ n คือ ขนาดตัวอย่าง และ p คือ จำนวนตัวแปรอิสระ

จากสมการการถดถอยลอจิสติก จะเห็นว่าเรามีข้อสมมติว่า ฟังก์ชันเชื่อมโยงจะอยู่ในรูปฟังก์ชันลอจิต ซึ่งถือว่าเป็นฟังก์ชันเชื่อมโยงที่ถูกต้อง ตัวแบบจะต้องมีตัวแปรอิสระทุกตัวที่เกี่ยวข้อง และฟังก์ชันลอจิตจะอยู่ในรูปฟังก์ชันเชิงเส้นของตัวแปรอิสระเหล่านี้ แต่บางครั้งอาจเป็นไปได้ว่าฟังก์ชันลอจิตที่ใช้เป็นฟังก์ชันเชื่อมโยงเป็นฟังก์ชันที่ไม่ถูกต้อง หรือตัวแบบมีตัวแปรอิสระที่สำคัญไม่ครบถ้วน หรือ

ความสัมพันธ์ระหว่างฟังก์ชันลอจิตของตัวแปรตามและตัวแปรอิสระไม่อยู่ในรูปเชิงเส้น ไม่ว่าจะเป็นกรณีใด เราจะเรียกว่ามีการกำหนดตัวแบบไม่ถูกต้อง (model misspecification)

เมื่อได้ตัวแบบการถดถอยแล้ว ขั้นตอนต่อมาคือการตรวจสอบความเหมาะสมของตัวแบบ ค่าสัมประสิทธิ์การตัดสินใจ (coefficient of determination; R^2) เป็นค่าที่ใช้กันอย่างแพร่หลายในการตรวจสอบความเหมาะสมของตัวแบบ โดยตัววัด R^2 ของการวิเคราะห์การถดถอยลอจิสติก (logistic regression) จะมีค่าที่คล้ายคลึงกับ R^2 ของการวิเคราะห์การถดถอยเชิงเส้น หรือที่เรียกว่าค่า R^2 เทียม (pseudo-R^2) ที่ใช้เป็นค่าที่ระบุถึงสัดส่วนในแง่ของฟังก์ชันความควรจะเป็นของตัวแปรตามที่อธิบายได้ด้วยตัวแปรอิสระ แทนที่จะเป็นค่าสัดส่วนของความแปรปรวนเช่นเดียวกับค่า R^2 ของการถดถอยเชิงเส้น แต่งานวิจัยชิ้นนี้ ต่อไปจะเรียก Pseudo-R^2 ย่อๆ เพียง R^2 เท่านั้นเพื่อให้ง่ายและสั้น

R^2 ของการวิเคราะห์การถดถอยลอจิสติก มีการคำนวณได้หลายวิธีที่แตกต่างกัน Menard [1] ได้จัดกลุ่มของ R^2 ของการวิเคราะห์การถดถอยลอจิสติกตามวิธีการคำนวณ เช่น R^2 ที่คำนวณด้วยหลักการของสัดส่วนการลดลงของการกระจาย (proportional reduction in dispersion) R^2 ที่คำนวณโดยอาศัยฟังก์ชันควรจะเป็น (likelihood function) และ R^2 ที่คำนวณด้วยข้อมูลที่จัดอันดับ (rank information) เป็นต้น ซึ่ง R^2เหล่านี้บางตัวสามารถเขียนความสัมพันธ์เชิงฟังก์ชันกันได้ [2] อย่างไรก็ตาม ยังไม่มีข้อสรุปที่ชัดเจนว่า ควรใช้ R^2 ตัวใดในการวัดสัดส่วนของความผันแปรของตัวแปรตามที่อธิบายได้ด้วยตัวแปรอิสระ โดยเฉพาะกรณีที่มีการกำหนดตัวแบบที่ไม่ถูกต้อง ซึ่ง R^2 ที่ดีควรจะได้รับผลกระทบจากการกำหนดตัวแบบที่ไม่ถูกต้องน้อยๆ

ดังนั้น ผู้วิจัยจึงสนใจศึกษาผลกระทบของการกำหนดตัวแบบไม่ถูกต้องที่มีต่อ R^2 ของตัวแบบการถดถอยลอจิสติก ที่คำนวณด้วยหลักการของสัดส่วนการลดลงของการกระจาย (proportional reduction in dispersion) และ R^2 ที่คำนวณโดยอาศัยฟังก์ชันควรจะเป็น (likelihood function) โดยจะศึกษาการกำหนดตัวแบบการถดถอยลอจิสติกไม่ถูกต้อง 3 กรณี ได้แก่ (1) กำหนดรูปแบบของตัวแปรอิสระไม่ถูกต้อง (wrong functional form of the covariate) (2) กำหนดฟังก์ชันเชื่อมโยงของตัวแบบไม่ถูกต้อง (misspecified link function) และ (3) กำหนดให้ตัวแบบมีตัวแปรอิสระที่มีความสัมพันธ์กับตัวแปรตามขาดหายไป (missing covariate)

2. ค่า Psuedo-R^2 หรือ R^2 ของการถดถอยลอจิสติก

การคำนวณค่าสัมประสิทธิ์การตัดสินใจ (R^2) ด้วยหลักการของสัดส่วนการลดลงของการกระจาย จะกำหนดให้ (y_i,x_i); I = 1, 2, …, n คือข้อมูลค่าสังเกตจำนวน n ค่า เมื่อ y_i คือค่าตัวแปรตามที่มีค่า 0 หรือ 1 ของค่าสังเกตที่ i และ x_i คือเวกเตอร์ของตัวแปรอิสระ และ $\hat{P}(y_i=1|x_i)=\hat{\pi}_i$ และ $\hat{P}(y_i=1)\equiv\overline{\pi}=\sum_{i=1}^{n}\frac{y_i}{n}=\overline{y}$ นอกจากนั้นยังกำหนดให้ $D(y_i)$ คือค่าที่วัดการกระจายสำหรับค่าสังเกตที่ i และ $D(y_i|x_i)$ คือค่าที่วัดการกระจายที่คำนวณได้จากการกำหนดเงื่อนไขของตัวแบบและเวกเตอร์ของตัวแปรอิสระ สำหรับค่า R^2 ที่นำมาศึกษาในงานวิจัยนี้ ได้แก่

2.1 R^2 แบบกำลังสองน้อยสุดสามัญ (ordinary least squared; R^2_{OLS}) การคำนวณค่า R^2_{OLS} นี้ จะ

กำหนดให้ $D(y_i)=(y_i-\bar{y})^2$ และ $D(y_i|x_i)=(y_i-\hat{\pi}_i)^2$ โดยที่ $SST=\sum_i D(y_i)$ และ $SSE=\sum_i D(y_i \mid x_i)$ ดังนั้นค่า R^2_{OLS} สามารถเขียนได้ดังนี้

$$R^2_{OLS} = 1-\frac{SSE}{SST} = \frac{2\sum_{i=1}^{n} y_i\hat{\pi}_i - \sum_{i=1}^{n}\hat{\pi}_i^2 - n\bar{\pi}_i}{n\bar{\pi}(1-\bar{\pi})} \quad (2)$$

2.2 R^2 ของ Gini's (Gini's concentration; R^2_G) [3] คำนวณจากการใช้การวัดสู่ศูนย์กลาง (concentration) ของ Gini หรือ $C(\pi)=1-\sum_{j=1}^{s}\pi_j^2$ ในการวัดการกระจายของตัวแปรสุ่ม Y ที่มีมาตรวัดแบบนามสเกลโดยที่กำหนดให้ $Y=j$, $1\le j\le s$ ด้วยความน่าจะเป็น π_j ในกรณีที่ Y มีค่าเพียง 2 ค่า คือ 0 หรือ 1 ค่าของ $C(\pi)$ จะลดรูปเหลือเพียง $\pi(1-\pi)$ เมื่อ π คือความน่าจะเป็นที่ $Y=1$ และถ้าใช้แนวคิดของฟังก์ชันความแปรปรวนสำหรับตัวแบบลอจิสติกแล้วจะได้ว่า $D(y_i)=\bar{y}(1-\bar{y})$ และ $D(y_i|x_i)=\hat{\pi}_i(1-\hat{\pi}_i)$ ดังนั้นตัววัด R^2_G สามารถเขียนได้ดังสมการนี้

$$R^2_G = \frac{\sum_{i=1}^{n}\bar{y}(1-\bar{y}) - \sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i)}{\sum_{i=1}^{n}\bar{y}(1-\bar{y})} = \frac{\sum_{i=1}^{n}\hat{\pi}_i^2 - n\bar{y}^2}{n\bar{y}(1-\bar{y})} \quad (3)$$

2.3 R^2 แบบอัตราส่วนความควรจะเป็น (likelihood ratio; R^2_L) ค่า R^2_L นี้โดยทั่วไปจะถูกใช้ในโปรแกรมสำเร็จรูปต่างๆ ทางสถิติที่มีชื่อเรียกว่า McFadden's pseudo R^2 คำนวณโดยใช้ฟังก์ชันล็อกควรจะเป็น (-2log-likelihood) ของสองตัวแบบ โดยที่ $D_M=-2\log L_M$ จะมีค่าเท่ากับ SSE ของตัวแบบที่มีตัวแปรอิสระที่สนใจ p ตัว และ $D_0=-2\log L_0$ จะมีค่าเท่ากับ SSE ของตัวแบบที่มีเฉพาะค่าคงที่ซึ่งเปรียบเสมือนได้กับค่าผลรวมกำลังสองทั้งหมด (total sum of square; SST) ของ OLS [4] ดังนั้น R^2_L สามารถเขียนได้ดังนี้

$$R^2_L = 1-\frac{\log(L_M)}{\log(L_0)} = \frac{\log L_0 - \log L_M}{\log L_0} \quad (4)$$

2.4 R^2 จากการปรับปรุงค่ากำลังสองเฉลี่ยเรขาคณิต (Unadjusted and Adjusted Geometric Mean Square Improvement: R^2_M) Maddala [5] และ Mcgee [6] ได้เสนอค่า R^2_M ที่คำนวณได้จาก

$$R^2_M = 1-\exp\{-\frac{2}{n}[\log L_M - \log L_0]\} = 1-\left(\frac{L_0}{L_M}\right)^{2/n} \quad (5)$$

2.5 R^2 ที่ปรับค่าโดยคำนวณจากการปรับปรุงค่ากำลังสองเฉลี่ยเรขาคณิต (unadjusted and adjusted geometric mean square improvement; R^2_N) คำนวณจากการนำค่า $\max(R^2_M)=1-(L_0)^{2/n}$ ไปหารค่า R^2_M [7,8]

$$R^2_N = \frac{1-\left(\frac{L_0}{L_M}\right)^{2/n}}{1-L_0^{2/n}} \quad (6)$$

2.6 R^2 ที่คำนวณจากสัมประสิทธิ์ตารางการณ์จร (unadjusted and adjusted contingency coefficient; R^2_C) Aldrich และ Nelson [9] ได้เสนอ R^2 ที่คำนวณจากสัมประสิทธิ์ตารางการณ์จรโดยมีสูตรการคำนวณดังนี้

$$R^2_C = \frac{G_M}{G_M+n} \quad (7)$$

เมื่อ $G_M = -2\log\left(\frac{L_0}{L_M}\right)$

3. วิธีดำเนินการศึกษา

การจำลองข้อมูลและการประมาณค่าจะทำโดยเขียนชุดคำสั่งในโปรแกรม SAS® เวอร์ชัน 8.0 ซึ่งมีขั้นตอนการวิจัยดังต่อไปนี้

3.1 กำหนดให้มีขนาดตัวอย่าง (n) เท่ากับ 100, 250, 500 และ 1000 และจำนวนตัวแปรอิสระ (p) เท่ากับ 1, 5, และ 10 ตัว

3.2 จำลองชุดข้อมูลของตัวแปรอิสระ X_j เมื่อ j=1, 2, …, 10 โดยกำหนดให้ตัวแปรอิสระมีทั้งตัวแปรเชิงปริมาณและเชิงคุณภาพ ตัวแปรที่เป็นเชิงปริมาณ คือ X_1, X_3, X_5, X_7 และ X_9 โดยตัวแปรเหล่านี้จะถูกสุ่มจากการแจกแจงปกติดังต่อไปนี้ $X_1 \sim N(2,4)$, $X_3 \sim N(20,4)$, $X_5 \sim N(8,3)$, $X_7 \sim N(0,1)$ และ $X_9 \sim N(15,2)$ สำหรับตัวแปรเชิงคุณภาพนั้น จะถูกสุ่มจากการแจกแจงแบบแบ่งกลุ่ม โดยมีรายละเอียดดังนี้ $X_2 \sim Cat[0,1]$, $X_4 \sim Cat[0,1,2,3]$, $X_6 \sim Cat[0,1,2]$, $X_8 \sim Cat[0,1]$ และ $X_{10} \sim Cat[0,1,2,3,4]$

3.3 สร้างชุดข้อมูลของตัวแปรตามแบบทวิภาคจากตัวแบบ (M0), (M1), และ (M2) (ตารางที่ 1) โดยที่ค่าสัมประสิทธิ์ของตัวแบบจะกำหนดขึ้นเพื่อให้ค่าของตัวแปรตามที่มีค่า 0 และ 1 มีสัดส่วนเท่าๆ กัน

3.4 ตัวแบบที่ใช้ศึกษาผลกระทบของการกำหนดตัวแบบไม่ถูกต้องที่มีต่อตัววัด R^2 ต่างๆ คือ

3.4.1 กำหนดรูปแบบของตัวแปรอิสระไม่ถูกต้อง กรณีนี้จะกำหนดให้ตัวแปรอิสระเป็นตัวแปรเชิงปริมาณ (X_1) มีรูปแบบเป็นกำลังสอง (X_1^2) นั่นคือ ตัวแบบ (M3), (M4) และ (M5) (ตารางที่ 1)

ตารางที่ 1 รหัสและนิยามของตัวแบบการถดถอยลอจิสติกที่ใช้ในการศึกษา

รหัสตัวแบบ	นิยามของตัวแบบการถดถอย
(M0)	$\text{logit}(\pi_i)=0.5-0.45x_{i1}$
(M1)	$\text{logit}(\pi_i)=0.5-0.45x_{i1}+3.525x_{i2}-0.1x_{i3}-0.5x_{i4}+0.2x_{i5}$
(M2)	$\text{logit}(\pi_i)=0.5-0.45x_{i1}+3.525x_{i2}-0.1x_{i3}-0.5x_{i4}+0.2x_{i5}-0.4x_{i6}+0.3x_{i7}-0.3x_{i8}-0.05x_{i9}+0.8x_{i10}$
(M3)	$\text{logit}(\pi_i)=\beta_0+\beta_1 x_{i1}^2$
(M4)	$\text{logit}(\pi_i)=\beta_0+\beta_1 x_{i1}^2+\beta_2x_{i2}+\beta_3x_{i3}+\beta_4x_{i4}+\beta_5x_{i5}$
(M5)	$\text{logit}(\pi_i)=\beta_0+\beta_1 x_{i1}^2+\beta_2x_{i2}+\beta_3x_{i3}+\beta_4x_{i4}+\beta_5x_{i5}+\beta_6x_{i6}+\beta_7x_{i7}+\beta_8x_{i8}+\beta_9x_{i9}+\beta_{10}x_{i10}$
(M6)	$\log[-\log(1-\pi_i)]=\beta_0+\beta_1x_{i1}$
(M7)	$\log[-\log(1-\pi_i)]=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i3}+\beta_4x_{i4}+\beta_5x_{i5}$
(M8)	$\log[-\log(1-\pi_i)]=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i3}+\beta_4x_{i4}+\beta_5x_{i5}+\beta_6x_{i6}+\beta_7x_{i7}+\beta_8x_{i8}+\beta_9x_{i9}+\beta_{10}x_{i10}$
(M9)	$\Phi^{-1}(\pi_i)=\beta_0+\beta_1x_{i1}$
(M10)	$\Phi^{-1}(\pi_i)=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i3}+\beta_4x_{i4}+\beta_5x_{i5}$
(M11)	$\Phi^{-1}(\pi_i)=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i3}+\beta_4x_{i4}+\beta_5x_{i5}+\beta_6x_{i6}+\beta_7x_{i7}+\beta_8x_{i8}+\beta_9x_{i9}+\beta_{10}x_{i10}$
(M12)	$\text{logit}(\pi_i)=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i3}+\beta_4x_{i4}$
(M13)	$\text{logit}(\pi_i)=\beta_0+\beta_1x_{i1}+\beta_3x_{i3}+\beta_4x_{i4}+\beta_5x_{i5}$
(M14)	$\text{logit}(\pi_i)=\beta_0+\beta_3x_{i3}$
(M15)	$\text{logit}(\pi_i)=\beta_0+\beta_4x_{i4}$
(M16)	$\text{logit}(\pi_i)=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_4x_{i4}+\beta_5x_{i5}+\beta_6x_{i6}+\beta_8x_{i8}+\beta_9x_{i9}+\beta_{10}x_{i10}$
(M17)	$\text{logit}(\pi_i)=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i3}+\beta_4x_{i4}+\beta_5x_{i5}+\beta_6x_{i6}+\beta_7x_{i7}+\beta_9x_{i9}$
(M18)	$\text{logit}(\pi_i)=\beta_0+\beta_1x_{i1}+\beta_3x_{i3}$
(M19)	$\text{logit}(\pi_i)=\beta_0+\beta_2x_{i2}+\beta_4x_{i4}$

3.4.2 กำหนดฟังก์ชันเชื่อมโยงของตัวแบบม่ถูกต้อง กรณีนี้จะกำหนดให้ฟังก์ชันเชื่อมโยงไม่ถูกต้อง 2 รูปแบบ

- รูปแบบแรกคือฟังก์ชันคอมพลีเมนทารีล็อก-ล็อก (complementary log-log) นั่นคือ ตัวแบบ (M6), (M7) และ (M8) (ตารางที่ 1)

- รูปแบบที่สองคือฟังก์ชันโพรบิท (probit) นั่นคือ ตัวแบบ (M9), (M10) และ (M11) (ตารางที่ 1)

3.4.3 กำหนดให้ตัวแบบการถดถอยมีตัวแปรอิสระขาดหายไป นั่นคือ

- สำหรับตัวแบบการถดถอยที่มีตัวแปรอิสระ 5 ตัว กำหนดให้ตัวแปรอิสระที่ขาดหายไปจากตัวแบบเป็นตัวแปรเชิงปริมาณ (X_5) และตัวแปรเชิงคุณภาพ (X_2) ในตัวแบบ (M12) และ (M13) ตามลำดับ (ตารางที่ 1) และกำหนดให้ตัวแปรอิสระที่ขาดหายไปจากตัวแบบมีทั้งตัวแปรเชิงปริมาณและเชิงคุณภาพ โดยกำหนดให้ตัวแปรอิสระที่ขาดหายไปคือ X_1, X_2, X_4, X_5 และ X_1, X_2, X_3, X_5 สำหรับตัวแบบ (M14) และ (M15) ตามลำดับ (ตารางที่ 1)

- สำหรับตัวแบบการถดถอยที่มีตัวแปรอิสระ 10 ตัว กำหนดให้ตัวแปรอิสระที่ขาดหายไปจากตัวแบบเป็นตัวแปรเชิงปริมาณ (X_3, X_7) และกำหนดให้ตัวแปรอิสระที่ขาดหายไปเป็นตัวแปรเชิงคุณภาพ (X_8, X_{10}) ในตัวแบบ (M16) และ (M17) ตามลำดับ (ตารางที่ 1)

กำหนดให้ตัวแปรอิสระที่ขาดหายไปจากตัวแบบมีทั้งตัวแปรเชิงปริมาณและเชิงคุณภาพ โดยกำหนดให้ตัวแปรอิสระที่ขาดหายไป คือ X_2, X_4, X_5, X_6, X_7, X_8, X_9, X_{10} และ X_1, X_3, X_5, X_6, X_7, X_8, X_9, X_{10} สำหรับตัวแบบ (M18) และ (M19) ตามลำดับ (ตารางที่ 1)

3.5 เกณฑ์ที่ใช้ในการเปรียบเทียบผลกระทบของการกำหนดตัวแบบไม่ถูกต้องที่มีต่อ R^2 ต่างๆ จะเปรียบเทียบความเอนเอียงสัมพัทธ์ (relative bias) ของค่าประมาณมัธยฐาน ($\tilde{R}^2$) ที่ได้จากการทำซ้ำจำนวน 1,000 ครั้ง โดยที่

$$\text{Relative Bias}=\left|\frac{\tilde{R}^2-R^2_{true}}{R^2_{true}}\right|\times 100\%$$

และ $R^2_{O,true}$ และ $R^2_{l,true}$ เป็นค่าสัมประสิทธิ์การตัดสินใจที่แท้จริงจากการคำนวณด้วยหลักการของสัดส่วนการลดลงของการกระจาย และจากการคำนวณด้วยอาศัยฟังก์ชันควรจะเป็นตามลำดับ โดยมีสูตรคำนวณดังนี้

$$R^2_{O,true}=1-\frac{\sum_{i=1}^{n}\left[\pi_i(1-\pi_i)\right]}{\sum_{i=1}^{n}\left[\bar{\pi}(1-\bar{\pi})\right]} \quad (8)$$

$$R^2_{l,true}=1-\frac{\sum_{i=1}^{n}\left[\pi_i\log\pi_i+(1-\pi_i)\log(1-\pi_i)\right]}{\sum_{i=1}^{n}\left[\bar{\pi}\log\bar{\pi}+(1-\bar{\pi})\log(1-\bar{\pi})\right]} \quad (9)$$

เมื่อ π_i คือค่าเริ่มต้นที่ได้จากการจำลอง logit(π_i) และ $\bar{\pi}$ คือค่าเฉลี่ยของ π_i เริ่มต้น

นอกจากนั้นยังเปรียบเทียบค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (root mean square error; RMSE) ของค่าประมาณสัมประสิทธิ์การตัดสินใจ ($R^2_{est,k}$) ที่คำนวณจากการทำซ้ำจำนวน 1,000 ครั้ง เมื่อ k=1, 2, …, 1000 โดยที่ ค่า RMSE คำนวณแยกตามวิธีการที่ใช้ในการคำนวณค่า R^2 ดังนี้

$$RMSE_O=\sqrt{\frac{\sum_{k=1}^{1,000}\left(R^2_{O,true}-R^2_{O,est,k}\right)^2}{1,000}} \quad (10)$$

$$RMSE_l=\sqrt{\frac{\sum_{k=1}^{1,000}\left(R^2_{l,true}-R^2_{l,est,k}\right)^2}{1,000}} \quad (11)$$

เมื่อ $R^2_{O,est,k}$ และ $R^2_{I,est,k}$ คือค่าประมาณ R^2 จากการคำนวณด้วยหลักการของสัดส่วนการลดลงของการกระจาย และจากการคำนวณด้วยอาศัยฟังก์ชันควรจะเป็นตามลำดับที่ได้จากการทำซ้ำครั้งที่ k

4. ผลการวิจัย

4.1 กำหนดรูปแบบของตัวแปรอิสระไม่ถูกต้อง

เมื่อพิจารณากรณีที่กำหนดรูปแบบของตัวแปรอิสระไม่ถูกต้องคือตัวแบบ (M3) (ตารางที่ 1) พบว่าค่า R^2_M ให้ค่าความเอนเอียงสัมพัทธ์ต่ำสุดในทุกขนาดตัวอย่าง สำหรับค่า R^2_{OLS}, R^2_G, R^2_L ให้ค่าความเอนเอียงสัมพัทธ์ใกล้เคียงกันและสูงกว่าค่า R^2 อื่นๆ และเมื่อพิจารณาค่ารากที่สองของค่าคลาดเคลื่อนกำลังสองเฉลี่ย (RMSE) ของค่า R^2 พบว่าค่า R^2_{OLS} และ R^2_G จะให้ค่า RMSE มากที่สุด ในขณะที่ค่า R^2_M จะให้ค่า RMSE น้อยที่สุด เมื่อขนาดตัวอย่างเท่ากับ 250, 500 และ 1000 แต่เมื่อตัวอย่างเท่ากับ 50 และ 100 ค่า RMSE ของ R^2_N จะมีค่ามากสุด ในขณะที่ค่า RMSE ของ R^2_M ไม่แตกต่างจากค่า RMSE ของค่า R^2 อื่นๆ

เมื่อพิจารณากรณีที่กำหนดรูปแบบของตัวแปรอิสระไม่ถูกต้องคือตัวแบบ (M4) พบว่าค่า R^2_N จะให้ค่าความเอนเอียงสัมพัทธ์สูงกว่า R^2 อื่นๆ ในทุกขนาดตัวอย่าง เมื่อพิจารณา RMSE ของค่า R^2 พบว่าค่า R^2_N จะให้ค่า RMSE มากกว่า R^2 อื่นๆ ในทุกขนาดตัวอย่างที่ทำการศึกษา

เมื่อพิจารณากรณีที่กำหนดรูปแบบของตัวแปรอิสระไม่ถูกต้องคือตัวแบบ (M5) พบว่าค่า R^2_N จะให้ค่าความเอนเอียงสัมพัทธ์สูงกว่า R^2 อื่นๆ ในทุกขนาดตัวอย่าง และเมื่อพิจารณา RMSE ของค่า R^2 พบว่าค่า R^2_N ให้ค่า RMSE มากกว่า R^2 อื่นๆ ในทุกขนาดตัวอย่าง

4.2 กำหนดฟังก์ชันเชื่อมโยงของตัวแบบไม่ถูกต้อง ซึ่งมีฟังก์ชันเชื่อมโยงไม่ถูกต้อง 2 รูปแบบ นั่นคือ

4.2.1 ฟังก์ชันคอมพลีเมนทารีล็อก-ล็อก (complementary log-log)

เมื่อพิจารณากรณีที่กำหนดฟังก์ชันเชื่อมโยงไม่ถูกต้องเป็นฟังก์ชันคอมพลีเมนทารีล็อก-ล็อก และมีตัวแปรอิสระเพียงตัวเดียว (M6) เมื่อฟังก์ชันลอจิตเป็นฟังก์ชันเชื่อมโยงที่แท้จริง (M0) (ตารางที่ 1) พบว่าค่า R^2 ทุกตัวให้ค่าความเอนเอียงสัมพัทธ์ที่สูง โดยค่า R^2_G จะมีค่าความเอนเอียงสัมพัทธ์สูงที่สุด รองลงมาคือ R^2_L และ R^2_N จะให้ค่าความเอนเอียงสัมพัทธ์ต่ำสุดในทุกขนาดตัวอย่าง และเมื่อพิจารณา RMSE ของ R^2 พบว่าค่า RMSE จะมีค่าสูงเมื่อขนาดตัวอย่างเท่ากับ 50 และ 100 โดยค่า R^2_N จะให้ค่า RMSE สูงสุด แต่เมื่อขนาดตัวอย่างเท่ากับ 250, 500 และ 1000 แล้ว ค่า RMSE ของ R^2_G จะสูงสุด และค่า RMSE ของ R^2 ทุกตัวจะมีค่าลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น

เมื่อพิจารณากรณีที่กำหนดฟังก์ชันเชื่อมโยงไม่ถูกต้องเป็นฟังก์ชันคอมพลีเมนทารีล็อก-ล็อก และมีตัวแปรอิสระ 5 ตัว (M7) เมื่อฟังก์ชันลอจิตเป็นฟังก์ชันเชื่อมโยงที่แท้จริง (M1) พบว่าค่า R^2_G ให้ค่าความเอนเอียงสัมพัทธ์สูงที่สุด รองลงมาคือ R^2_L และเมื่อพิจารณาค่า RMSE ของค่า R^2 พบว่าค่า RMSE ของ R^2_G จะมีค่าสูงสุดสุด รองลงมาคือ R^2_L เมื่อขนาดตัวอย่างเท่ากับ 250, 500 และ 1000 แต่เมื่อขนาดตัวอย่างเท่ากับ 50 และ 100 แล้ว ค่า RMSE ของ R^2_C ให้ค่าสูงสุด

เมื่อพิจารณากรณีที่กำหนดฟังก์ชันเชื่อมโยงไม่ถูกต้องเป็นฟังก์ชันคอมพลีเมนทารีล็อก-ล็อก และมีตัวแปรอิสระ 10 ตัว (M8) เมื่อฟังก์ชันลอจิตเป็นฟังก์ชันเชื่อมโยงที่แท้จริง (M2) พบว่าค่า R_L^2 จะให้ค่าความเอนเอียงสัมพัทธ์สูงที่สุด รองลงมาคือ R_G^2 เมื่อขนาดตัวอย่างเท่ากับ 50 และ 100 แต่ในทางกลับกัน เมื่อตัวอย่างเท่ากับ 250, 500 และ 1000 ค่า R_G^2 จะให้ค่าความเอนเอียงสัมพัทธ์สูงที่สุด รองลงมาคือ R_L^2 เมื่อพิจารณาค่า RMSE ของค่า R^2 พบว่าค่า RMSE จะมีค่าสูงเมื่อขนาดตัวอย่างเท่ากับ 50, 100 และ 250 โดยเฉพาะค่า R_C^2 แต่เมื่อขนาดตัวเท่ากับ 500 และ 1000 แล้ว ค่า R_G^2 จะให้ค่า RMSE สูงสุด อย่างไรก็ตามค่า RMSE ของ R^2 ทุกตัวจะมีค่าลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น

4.2.2 ฟังก์ชันฟังก์ชันโพรบิท (probit)

เมื่อพิจารณากรณีที่กำหนดฟังก์ชันเชื่อมโยงไม่ถูกต้องเป็นฟังก์ชันโพรบิท (ตารางที่ 1) และมีตัวแปรอิสระในตัวแบบเพียงตัวเดียว (M9) เมื่อฟังก์ชันลอจิตเป็นฟังก์ชันเชื่อมโยงที่แท้จริง (M0) พบว่า R_N^2 ให้ค่าความเอนเอียงสัมพัทธ์สูงที่สุด รองลงมาคือ R_M^2 และ R_C^2 ตามลำดับ และเมื่อพิจารณาค่า RMSE ของค่า R^2 พบว่าค่า R_N^2 ให้ค่า RMSE มากสุดในทุกขนาดตัวอย่าง และค่า RMSE ของ R^2 ทุกตัวจะมีค่าลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น

เมื่อพิจารณากรณีที่กำหนดฟังก์ชันเชื่อมโยงไม่ถูกต้องเป็นฟังก์ชันโพรบิท และมีตัวแปรอิสระในตัวแบบ 5 ตัว (M10) เมื่อฟังก์ชันลอจิตเป็นฟังก์ชันเชื่อมโยงที่แท้จริง (M1) พบว่า ค่า R_N^2 ให้ค่าความเอนเอียงสัมพัทธ์สูงที่สุด รองลงมาคือ R_M^2 ในทุกขนาดตัวอย่างที่ทำการศึกษา และเมื่อพิจารณาค่า RMSE) ของค่า R^2 พบว่าค่า R_N^2 ให้ค่า RMSE มากสุดในทุกขนาดตัวอย่าง และค่า RMSE ของ R^2 ทุกตัวจะมีค่าลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น

เมื่อพิจารณากรณีที่กำหนดฟังก์ชันเชื่อมโยงไม่ถูกต้องเป็นฟังก์ชันเชื่อมโยงโพรบิทและมีตัวแปรอิสระในตัวแบบ 10 ตัว (M11) เมื่อฟังก์ชันลอจิตเป็นฟังก์ชันเชื่อมโยงที่แท้จริง (M2) พบว่าค่า R_N^2 ให้ค่าความเอนเอียงสัมพัทธ์ และค่า RMSE ลดลงเมื่อตัวอย่างมีขนาดเพิ่มขึ้น

4.3 กำหนดให้ตัวแบบการถดถอยมีตัวแปรอิสระขาดหายไป

เมื่อพิจารณากรณีที่กำหนดให้ตัวแปรอิสระ X_5 ขาดหายไปคือตัวแบบ (M12) (ตารางที่ 1) พบว่าค่า R_N^2 ให้ค่าความเอนเอียงสัมพัทธ์และค่า RMSE สูงกว่า R_M^2

เมื่อพิจารณากรณีที่กำหนดให้ตัวแปรอิสระ X_2 ขาดหายไปคือตัวแบบ (M13) พบว่าค่า R_M^2 มีความเอนเอียงสัมพัทธ์สูงสุดเมื่อขนาดตัวอย่างเท่ากับ 50 ตรงข้ามกับค่า R_N^2 ที่ให้ความเอนเอียงสัมพัทธ์และค่า RMSE สูงสุดเมื่อขนาดตัวอย่างเท่ากับ 50, 100 และ 250 ในขณะที่ค่า R_L^2 จะให้ความเอนเอียงสัมพัทธ์และค่า RMSE สูงสุดเมื่อขนาดตัวอย่างเท่ากับ 500 และ 1000

เมื่อพิจารณากรณีที่กำหนดให้ตัวแปรอิสระ X_1, X_2, X_4, X_5 ขาดหายไปคือตัวแบบ (M14) พบว่าค่า R^2 ทุกตัวให้ค่าความเอนเอียงสัมพัทธ์ของค่า R^2 ทุกตัวมีค่าสูงอย่างน้อย 95% ในทุกขนาดตัวอย่าง

เมื่อพิจารณากรณีที่กำหนดให้ตัวแปรอิสระ X_1, X_2, X_3, X_5 ขาดหายไปคือตัวแบบ (M15) พบว่าค่า R_N^2 ให้ความเอนเอียงสัมพัทธ์และค่า RMSE ต่ำสุดและแตกต่างจากค่า R^2 อื่นๆ อย่างเห็นได้ชัดเจน

เมื่อพิจารณากรณีที่กำหนดให้ตัวแปรอิสระ X_3 และ X_7 ขาดหายไปคือตัวแบบ (M16) พบว่า

ค่า R_N^2 ให้ค่าความเอนเอียงสัมพัทธ์ และค่า RMSE สูงสุดในทุกขนาดตัวอย่าง ในขณะที่ค่า R_{OLS}^2, R_G^2, R_L^2 และ R_M^2 ให้ค่าความเอนเอียงสัมพัทธ์เฉพาะในกรณีที่ตัวอย่างมีขนาดเท่ากับ 50 และ 100 เท่านั้น

เมื่อพิจารณากรณีที่กำหนดให้ตัวแปรอิสระ X_8 และ X_{10} ขาดหายไปคือตัวแบบ (M17) พบว่าผลสรุปที่ได้จะเหมือนกับตัวแบบ (M16) เพียงแต่ปริมาณของความเอนเอียงสัมพัทธ์และค่า RMSE จะมีค่าที่น้อยกว่าของตัวแบบ (M16) นั่นคือ ค่า R_N^2 ให้ค่าความเอนเอียงสัมพัทธ์และค่า RMSE สูงสุดในทุกขนาดตัวอย่าง

เมื่อพิจารณากรณีที่กำหนดให้ตัวแปรอิสระ X_2, X_4, X_5, X_6, X_7, X_8, X_9, และ X_{10} ขาดหายไปคือตัวแบบ (M18) พบว่า R^2 ทุกตัวให้ค่าความเอนเอียงสัมพัทธ์สูงมาก และ R_N^2 ให้ค่าความเอนเอียงสัมพัทธ์ต่ำสุดในทุกขนาดตัวอย่าง โดยที่มีค่าประมาณ 75-83% นอกจากนั้น R_N^2 ยังให้ค่า RMSE ต่ำสุดด้วย

เมื่อพิจารณากรณีที่กำหนดให้ตัวแปรอิสระ X_1, X_3, X_5, X_6, X_7, X_8, X_9, และ X_{10} ขาดหายไปคือตัวแบบ (M19) พบว่าค่า R_M^2 จะให้ค่าความเอนเอียงสัมพัทธ์ต่ำสุด เมื่อขนาดตัวอย่างเท่ากับ 50 และ 100 แต่เมื่อขนาดตัวอย่างเท่ากับ 250, 500 และ 1000 ค่า R_M^2 และ R_N^2 ให้ค่าความเอนเอียงสัมพัทธ์ไม่แตกต่างกัน

5. สรุปผลการทดลอง

ผลกระทบของการกำหนดตัวแบบไม่ถูกต้องที่มีต่อ R^2 ที่นำมาศึกษานั้น จะแตกต่างกันขึ้นอยู่กับลักษณะของการกำหนดตัวแบบไม่ถูกต้อง จำนวนตัวแปร และขนาดตัวอย่าง สามารถสรุปได้ดังนี้

5.1 กรณีกำหนดรูปแบบของตัวแปรอิสระไม่ถูกต้อง โดยกำหนดให้ตัวแปรอิสระ X_1 มีรูปแบบเป็นกำลังสอง สรุปได้ว่าเมื่อตัวแบบมีเพียงตัวแปรอิสระ 1 ตัว ค่า R_M^2 จะให้ค่าความเอนเอียงสัมพัทธ์ต่ำสุด ในขณะที่ R_{OLS}^2, R_G^2 และ R_L^2 จะให้ค่าความเอนเอียงสัมพัทธ์และค่า RMSE ใกล้เคียงกันและมากสุดในทุกขนาดตัวอย่าง แต่เมื่อมีจำนวนตัวแปรอิสระในตัวแบบเพิ่มเป็น 5 และ 10 ตัว พบว่ามีเฉพาะ R_N^2 เท่านั้นที่ให้ค่าความเอนเอียงสัมพัทธ์สูง และเมื่อเปรียบเทียบความเอนเอียงสัมพัทธ์ของ R_N^2 พบว่าค่าความเอนเอียงสัมพัทธ์กรณีที่มีตัวแปรจำนวน 5 และ 10 จะมีค่ามากกว่ากรณีที่มีตัวแปร 1 ตัว

ดังนั้นสามารถสรุปได้ว่าตัววัด R_N^2 จะได้รับผลกระทบจากการกำหนดรูปแบบของตัวแปรอิสระไม่ถูกต้องมากสุด โดยผลกระทบนั้นทำให้ค่าความเอนเอียงสัมพัทธ์และค่า RMSE สูงขึ้น เมื่อมีจำนวนตัวแปรอิสระในตัวแบบเพิ่มขึ้น แต่ผลกระทบจะลดลงเมื่อมีขนาดตัวอย่างเพิ่มขึ้น ในขณะที่การกำหนดรูปแบบของตัวแปรอิสระไม่ถูกต้องมีผลกระทบต่อตัววัดอื่นๆ เฉพาะในกรณีที่ตัวแบบมีตัวแปรอิสระเพียงตัวเดียว

5.2 กรณีกำหนดฟังก์ชันเชื่อมโยงไม่ถูกต้อง เมื่อฟังก์ชันเชื่อมโยงที่แท้จริงเป็นฟังก์ชันลอจิต ผลกระทบของการกำหนดฟังก์ชันเชื่อมโยงไม่ถูกต้องที่มีต่อค่า R^2 ของการถดถอยลอจิสติกนั้น นอกจากจะขึ้นอยู่กับประเภทของฟังก์ชันเชื่อมโยงแล้ว ยังพบว่าขึ้นอยู่กับจำนวนตัวแปรอิสระ และขนาดตัวอย่างที่อยู่ในตัวแบบ เมื่อพิจารณาค่าความเอนเอียงสัมพัทธ์ของค่า R^2 เมื่อกำหนดฟังก์ชันเชื่อมโยงคอมพลีเมนทารีล็อก-ล็อก พบว่า R^2 ทุกตัวให้ค่าความเอนเอียงสัมพัทธ์ที่สูงในทุกขนาดตัวอย่าง และทุกจำนวนตัวแปร โดยค่า R_N^2 จะมีค่าความเอนเอียงสัมพัทธ์ต่ำสุด

ยกเว้นเฉพาะค่า R^2_{OLS} ที่ให้ค่าความเอนเอียงสัมพัทธ์และค่า RMSE ต่ำเมื่อมีจำนวนตัวแปรเพิ่มขึ้นเป็น 5 และ 10 ตัว

แต่เมื่อกำหนดฟังก์ชันเชื่อมโยงโพรบิทมีเฉพาะค่า R^2_N เท่านั้นที่ให้ค่าความเอนเอียงสัมพัทธ์และค่า RMSE ที่สูงในทุกขนาดตัวอย่างและจำนวนตัวแปรอิสระ ในขณะที่ R^2_M ให้ค่าความเอนเอียงสัมพัทธ์และค่า RMSE สูงเฉพาะในกรณีที่มีตัวแปรอิสระจำนวน 10 และขนาดตัวอย่างตั้งแต่ 250 เท่านั้น แต่อย่างไรก็ตาม ค่าความเอนเอียงสัมพัทธ์มีค่าน้อยกว่าของ R^2_N อย่างเห็นได้ชัดเจน

5.3 กรณีกำหนดให้ตัวแบบการถดถอยมีตัวแปรอิสระขาดหายไป ผลกระทบที่มีต่อค่า R^2 นั้น จะขึ้นอยู่กับจำนวนและประเภทของตัวแปรอิสระที่ขาดหายไป นอกจากนั้นยังขึ้นอยู่กับขนาดตัวอย่าง โดยพบว่าการขาดหายไปของตัวแปรอิสระมีผลทำให้ค่าความเอนเอียงสัมพัทธ์และค่า RMSE ของ R^2 ทุกตัวมีค่าสูงในทุกกรณีของการกำหนดให้มีตัวแปรอิสระขาดหายไป ยกเว้นเฉพาะในกรณีที่กำหนดให้มีตัวแปรอิสระขายหายไป จำนวน 1 ตัว ในกรณีนี้พบว่าเมื่อตัวแปรที่ขาดหายไปเป็นตัวแปรเชิงปริมาณ R^2_M และ R^2_N เท่านั้นที่ค่าความเอนเอียงสัมพัทธ์และค่า RMSE ที่สูงในทุกกรณีของขนาดตัวอย่าง โดย R^2_N ให้ค่าความเอนเอียงสัมพัทธ์และค่า RMSE สูงกว่าค่าของ R^2_M อย่างเด่นชัด แต่เมื่อตัวแปรที่ขาดหายไปเป็นเชิงคุณภาพจะมีเฉพาะ R^2_N เท่านั้นที่ให้ค่าความเอนเอียงสัมพัทธ์และค่า RMSE สูง และเมื่อกำหนดให้ตัวแปรอิสระขายหายไปเกือบทั้งหมดโดยให้เหลือตัวแปรอิสระในตัวแบบเพียงตัวเดียวหรือ 2 ตัว นั้น พบว่า R^2 ทุกตัวให้ค่าความเอนเอียงสัมพัทธ์และค่า RMSE ที่สูงมาก มีเพียง R^2_N ที่ให้ค่าประมาณความเอนเอียงสัมพัทธ์ต่ำสุด

6. ข้อเสนอแนะ

การวิจัยครั้งนี้ ผู้วิจัยมีข้อเสนอแนะดังต่อไปนี้

6.1 ในการวิจัยครั้งนี้ผู้วิจัยได้ศึกษาเฉพาะการกำหนดตัวแบบการถดถอยลอจิสติกไม่ถูกต้องใน 3 ลักษณะเท่านั้น ยังมีลักษณะการกำหนดตัวแบบไม่ถูกต้องอีกหลายรูปแบบที่ควรจะพิจารณา เช่น ตัวแบบเกิดปัญหาการกระจายของข้อมูลที่มากกว่าการกระจายของข้อมูลที่มีการแจกแจงแบบทวินาม หรือที่เรียกว่าการเกิดปัญหา over-dispersion หรือตัวแปรอิสระมีการวัดที่ไม่ถูกต้อง หรือเกิดความผิดพลาดในการแบ่งกลุ่มของตัวแปรเชิงปริมาณ เป็นต้น ดังนั้นในการวิจัยครั้งต่อไปอาจจะพิจารณาการกำหนดตัวแบบไม่ถูกต้องเหล่านี้ด้วย

6.2 ในการวิจัยครั้งนี้ผู้วิจัยได้กำหนดให้การเกิดผลตอบสนองที่มีค่า 0 และ 1 ในสัดส่วนที่เท่าๆ กัน นั่นคือกำหนดให้ $P(Y=1)=0.5$ แต่จากผลงานวิจัยอื่นๆ พบว่าสัดส่วนการเกิดผลตอบสนอง หรือ ที่เรียกว่า base rate นั้น จะมีผลต่อค่าประมาณของ R^2 ของการถดถอยลอจิสติก ดังนั้นในการวิจัยต่อไป ควรจะเพิ่มการศึกษาเมื่อกำหนดให้ base rate มีอัตราที่แตกต่างกัน

6. เอกสารอ้างอิง

[1] Menard, S., 2000, Coefficient of determination for multiple logistic regression analysis, Amer. Stat. 54: 17-24.

[2] แสงหล้า ชัยมงคลม, 2551, ความสัมพันธ์เชิงฟังก์ชันระหว่างค่าสัมประสิทธิ์การตัดสินใจประเภทต่างๆ สำหรับการวิเคราะห์การถดถอยลอจิสติกที่ได้จาก SPSS, ว.วิทยาศาสตร์และเทคโนโลยี 16(1): 71-76.

[3] Haberman, S.J., 1982, Analysis of dispersion of multinomial responses, J. Ame. Stat. Ass. 77: 568-580.

[4] Hosmer, D.W. and Lemeshow, S., 2000, Applied Regression Analysis, 2nd Ed, Wiley &Sons, Inc., New York.

[5] Maddala, G.S., 1983, Limitted-Dependent and Qualitative Variables in Economics, Cambridge University Press, Cambridge.

[6] Mcgee, L., 1990, R^2 measures based on wald and likelihood ratio joint significance tests, Ame. Stat. 44: 250-253.

[7] Nagelkerke, N.J.D., 1991, A note on a general definition of coefficient of determination, Biometrika 78: 691-692.

[8] Agresti, A., 1986, Applying R^2-type measures to ordered categorical data, Technometrics 28: 133-138.

[9] Aldrich, J.H. and Nelson, F.D., 1984, Linear Probability, Logit, and Probit Models, Sage, Beverly Hills, CA.