บทที่ 2

ทฤษฎีและผลงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะแบ่งออกเป็น 2 ส่วนคือ

ส่วนที่ 1 เป็นการทบทวนทฤษฎีที่เกี่ยวข้อง โดยจะทบทวนเกี่ยวกับการวิเคราะห์การถดถอยโลจิสติกทวิภาคที่ครอบคลุมถึงตัวแบบ ข้อสมมติของตัวแบบ การประมาณค่าพารามิเตอร์ของตัวแบบ และการตรวจสอบความเหมาะสมของตัวแบบ

ส่วนที่ 2 เป็นการทบทวนผลงานที่เกี่ยวข้อง โดยจะทบทวนเกี่ยวกับผลงานที่เสนอค่า R^2_{adj} อิทธิพลของสัดส่วนผลตอบสนอง ระดับความเชื่อถือได้ในตัวแปรอธิบาย และอัตราการจำแนกผิดในตัวแปรตาม

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 ตัวแบบการถดถอยโลจิสติกทวิภาค (Binary Logistic Regression Model)

ในการวิเคราะห์การถดถอยของตัวแบบที่ตัวแปรตาม (Y) เป็นแบบทวิภาคคือ มีค่าเป็น 0 เมื่อไม่เกิดเหตุการณ์ที่สนใจหรือมีค่าเป็น 1 เมื่อเกิดเหตุการณ์ที่สนใจ ส่วนตัวแปรอธิบาย (X) นั้นอาจเป็นได้ทั้งตัวแปรเชิงปริมาณและตัวแปรเชิงคุณภาพ เรียกการวิเคราะห์การถดถอยของตัวแบบลักษณะนี้ว่า การวิเคราะห์การถดถอยโลจิสติกทวิภาค (Binary Logistic Regression) ซึ่งจะมีวิธีการวิเคราะห์ที่แตกต่างไปจากการวิเคราะห์การถดถอยเชิงเส้นทั่วไป เนื่องจากการวิเคราะห์การถดถอยเชิงเส้นมีข้อสมมุติว่า $E(\varepsilon)=0$ นั่นคือ $Y = E(Y \mid X = \underline{x}_i)+\varepsilon$ แสดงได้ในรูปของตัวแบบการถดถอยเชิงเส้นคือ $E(Y \mid X = \underline{x}_i)+\varepsilon = \beta_0 + \sum_{j=1}^{k}\beta_j x_{ij} + \varepsilon_i$ ซึ่งเป็นค่าเฉลี่ยของ Y มีค่าอยู่ในช่วง $(-\infty,\infty)$ แต่การวิเคราะห์การถดถอยโลจิสติกทวิภาค เป็นการศึกษาความสัมพันธ์เชิงเส้นระหว่างความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ $P(Y_i=1 \mid X=\underline{x}_i)=\pi(x_i)$ กับตัวแปรอธิบาย โดยที่ $\pi(x_i)$ ก็คือค่าความน่าจะเป็นแบบ มีเงื่อนไขเมื่อ $Y_i=1$ ณ ที่ระดับของตัวแปรอิสระ x_i และพบว่าค่าของ $\pi(x_i)$ จะอยู่ในช่วง $(0,1)$ ซึ่งแตกต่างจากตัวแบบการถดถอยเชิงเส้นที่จะมีค่าเฉลี่ยอยู่ในช่วง $(-\infty,\infty)$ ดังนั้นถ้าใช้ตัวแบบการถดถอยเชิงเส้นจะทำให้ค่าพยากรณ์ที่ได้มีค่าไม่เหมาะสม นั่นคือมีค่าน้อยกว่า 0 หรือ

มากกว่า 1 ด้วยเหตุนี้จึงนำตัวแบบการถดถอยโลจิสติกทวิภาคมาใช้แทนตัวแบบการถดถอยเชิงเส้น ซึ่งตัวแบบการถดถอยโลจิสติกทวิภาคมีรูปแบบดังนี้

$$P(Y_i = 1 \mid X = \underline{x}_i) = \pi(x_i) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)}$$

หรือ

$$P(Y_i = 0 \mid X = \underline{x}_i) = 1 - \pi(x_i) = \frac{1}{1 + \exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)}$$

หรือ

$$logit(\pi_i) = \log\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$$

เมื่อ $\pi(x_i) = P(Y_i = 1 \mid X = \underline{x}_i)$ แทนความน่าจะเป็นแบบมีเงื่อนไขที่ $Y_i = 1$

β_0, β_j คือพารามิเตอร์ของตัวแบบ

Y_i คือตัวแปรตามแบบทวิภาค

x_{ij} คือค่าของตัวแปรอธิบาย

โดยมีข้อสมมติที่สำคัญ คือ

(1) Y_i มีการแจกแจงแบบแบร์นูลลี ; $Y_i \in \{0,1\}$ เมื่อ $i = 1,2,...,n$

(2) Y_i , $i = 1,2,...,n$ เป็นอิสระแก่กัน

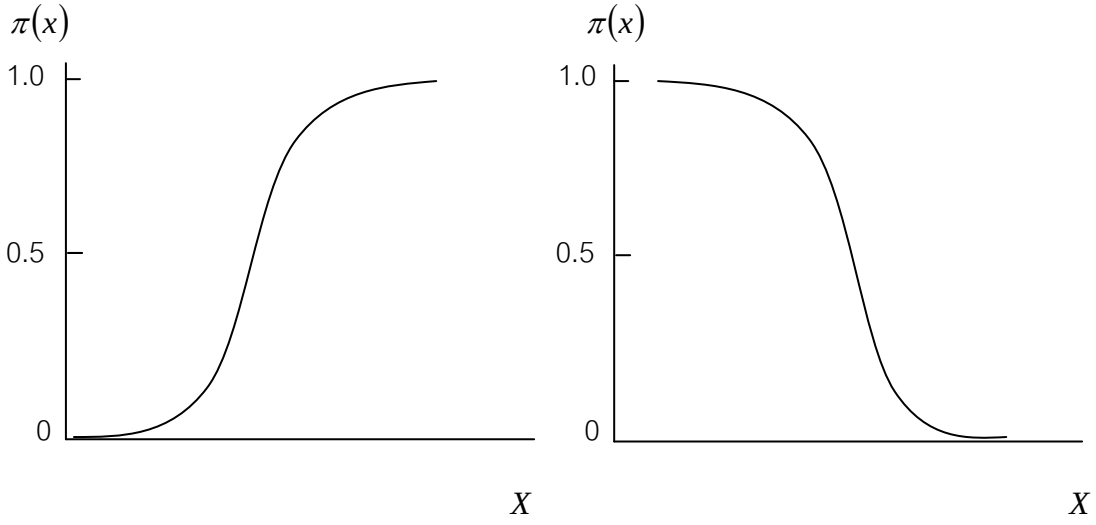
(3) ตัวแปรอิสระ x_{ij} ไม่มีความสัมพันธ์เชิงเส้นระหว่างกัน เมื่อ $i = 1,2,...,n$ และ $j = 1,2,...,k$

ตัวแบบการถดถอยโลจิสติกกรณีที่ตัวแปรตามเป็นแบบทวิภาค (Binary Responses Variable) จะมีลักษณะที่สำคัญ 5 ประการ (วีรานันท์, 2541, น. 53-57) คือ

(1) ค่าเฉลี่ยแบบมีเงื่อนไขของตัวแบบการถดถอยโลจิสติกทวิภาคต้องมีค่าอยู่ระหว่าง 0 และ 1

(2) การแจกแจงของค่าคลาดเคลื่อนสุ่มสามารถอธิบายได้ด้วยการแจกแจงแบบ ทวินาม (Binomial Distribution) ซึ่งถ้าตัวอย่างมีขนาดใหญ่ การแจกแจงดังกล่าวจะใกล้เคียง กับการแจกแจงปรกติ (Normal Distribution)

(3) หลักเกณฑ์ที่เป็นแนวทางของการวิเคราะห์การถดถอยโลจิสติกทวิภาค สามารถใช้หลักเกณฑ์ของการวิเคราะห์การถดถอยเชิงเส้นทั้งแบบเชิงเดียวและแบบพหุคูณได้ เนื่องจากฟังก์ชันการถดถอยโลจิสติกทวิภาคมีลักษณะเส้นโค้งเป็นรูปตัว S (S – Shape or Sigmoid curve) ที่เป็นได้ทั้ง Monotonic Increasing และ Monotonic Decreasing ขึ้นอยู่กับเครื่องหมายของสัมประสิทธิ์การถดถอยดังแสดงในภาพ 2.1 เส้นโค้งของฟังก์ชันการถดถอยโลจิสติกทวิภาคมีลักษณะสมมาตรที่ $\pi(x)$= 0.5 และจะมีลักษณะความสัมพันธ์เหมือนการถดถอยเชิงเส้นและใกล้เคียงกับเส้นตรงเมื่อ $\pi(x)$ อยู่ระหว่าง 0.2 ถึง 0.8 ซึ่งมีค่าภายในช่วง 0 และ 1 เท่านั้นไม่ว่าตัวแปรอธิบายจะเป็นค่าใดๆ



ภาพที่ 2.1 เส้นโค้งของฟังก์ชันการถดถอยโลจิสติกทวิภาคที่เป็นแบบ Monotonic Increasing และ Monotonic Decreasing ตามลำดับ

(4) ตัวแปรตาม Y_i มีการแจกแจงแบร์นูลลีที่มี $P(Y_i = 1 | X = \underline{x}_i) = \pi(x_i)$ โดยที่ $0 \le \pi(x_i) \le 1$ แต่ $\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$ ไม่จำเป็นต้องอยู่ในช่วง $(0,1)$ ดังนั้นจึงต้องหาฟังก์ชันเชื่อมโยง (Link Function) มาสร้างตัวแบบที่แทนความสัมพันธ์เชิงเส้นระหว่าง $\pi(x_i)$ กับ

ตัวแปร x_i ในการถดถอยโลจิสติก โดยส่วนใหญ่จะทำการแปลงให้อยู่ในรูปโลจิท (Logit) ของ $\pi(x_i)$ คือ

$$\log it\pi(x_i) = \log\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right]$$

ซึ่ง $logit(\pi_i)$ ไม่จำเป็นต้องอยู่ในช่วง $(0,1)$ และมีความสัมพันธ์เชิงเส้นในรูปของ

$$\log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}$$

(5) กรณีที่ตัวแบบการถดถอยมีตัวแปรอธิบายเพียง 1 ตัว $\log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \beta_0 + \beta_1 x_{i1}$ มีประโยชน์สำหรับการตีความหมายในเทอมของโอกาสความเป็นไปได้ (Odds) โดย Odds จะเพิ่มขึ้นเป็น $\exp(\beta_1)$ เท่า ณ ทุกๆ ค่าของ X_1 ที่เพิ่มขึ้น 1 หน่วย และ Odds ของ $Y=1$ คือ $\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \beta_0 + \beta_1 x_{i1}$ เป็นฟังก์ชันเชื่อมโยงสำหรับตัวแบบการถดถอยโลจิสติก ให้มีรูปแบบเป็นตัวแบบเชิงเส้นวางนัยทั่วไป (Generalized Linear Model, GLM)

2.1.2 การประมาณค่าพารามิเตอร์ของสัมประสิทธิ์การถดถอยโลจิสติกทวิภาค

(วีรานันท์, 2541 น. 170-196)

การประมาณค่าพารามิเตอร์ของตัวแบบเชิงเส้นทั่วไปในอดีต นิยมใช้วิธีภาวะน่าจะเป็นสูงสุดและวิธีกำลังสองน้อยสุดแบบถ่วงน้ำหนัก (Weighted Least Squares) อย่างไรก็ตามในหลายสถานการณ์พบว่าไม่สามารถใช้วิธีการประมาณค่าดังกล่าวได้โดยตรง เนื่องจากสมการปรกติที่พบอาจมีรูปแบบไม่เป็นเชิงเส้น (Non Linear) ในเทอมของพารามิเตอร์ การแก้สมการเหล่านี้จึงต้องมีการคำนวณเพิ่มเติมด้วยวิธีย้อนซ้ำเชิงตัวเลข (Numerical Iteration) และเมื่อนำมาใช้ร่วมกับวิธีภาวะน่าจะเป็นสูงสุดแล้ว ทำให้เป็นวิธีที่มีประสิทธิภาพมากขึ้น จึงเรียกวิธีการดังกล่าวว่า การประมาณค่าด้วยวิธีภาวะน่าจะเป็นสูงสุดแบบย้อนซ้ำ (Iterative Maximum – Likelihood) เป็นวิธีที่ใช้ในการประมาณค่าของพารามิเตอร์ของตัวแบบในกลุ่มเอกซ์โพเนนเชียล (Exponential Family Models) โดยสืบเนื่องมาจากนิวเดอร์และเวดเดอร์เบิร์น (Nelder and Wedderburn, 1972) ได้ขยายวิธีการประมาณค่าแบบย้อนซ้ำ (Iterative Estimation) วิธีหนึ่งที่

เรียกว่า วิธีฟิชเชอร์สกอริง (Fisher Scoring Method) ให้ใช้ควบคู่กับวิธีภาวะน่าจะเป็นสูงสุด (Maximum Likelihood) และใช้สำหรับการประมาณพารามิเตอร์ของตัวแบบเชิงเส้นวางนัยทั่วไป (Generalized Linear Model, GLM) ซึ่งเป็นตัวแบบที่รวมถึงตัวแบบเชิงเส้นที่มีพื้นฐานของการแจกแจงแบบปรกติและตัวแบบอื่นๆ ที่มีการแจกแจงในกลุ่มเอกซ์โพเนนเชียลด้วย

วิธีของฟิชเชอร์สกอริง (Fisher's Scoring Method) เป็นวิธีที่ใช้สำหรับการประมาณค่าพารามิเตอร์ของตัวแบบการถดถอยในคำสั่ง PROC LOGISTIC ของโปรแกรม *SAS* โดยที่วิธีของฟิชเชอร์สกอริงได้จากการนำวิธีของนิวตัน ราฟสัน (Maximum Likelihood with Newton Raphson) มาปรับใช้ โดยมีวิธีการดังต่อไปนี้ ให้ตัวแปรตาม Y_i เป็นตัวแปรทวิภาค (Binary) จำนวน n ค่าสังเกต และ $\underline{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$ แทนรูปแบบ ของค่าตัวแปรอธิบาย p ตัวในกรณีที่ตัวแปรอธิบายทั้งหมดเป็นตัวแปรเชิงคุณภาพรูปแบบของ $\underline{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$ จะมีจำนวนน้อยกว่าจำนวนค่าสังเกต แต่ถ้าตัวแปรอธิบายเป็นตัวแปรเชิงปริมาณหลายๆ ตัว รูปแบบของ $\underline{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$ อาจมีจำนวนเท่ากับจำนวนค่าสังเกต (Agresti, 1990) และเมื่อ $i = 1,2,..., n$ ตัวแบบการถดถอยโลจิสติกมีรูปแบบดังนี้

$$\pi(x_i) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}$$

โดยที่ $\pi(x_i) = P(Y_i = 1 \mid X = \underline{x}_i)$ และ β_0, β_j คือพารามิเตอร์ของตัวแบบเมื่อ $j = 1,2,..., p$ สำหรับในแต่ละรูปแบบหรือชุดของ $\underline{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})$ ค่าตัวแปรตามที่ $Y_i = 1$ อาจจะมีค่ามากกว่า 1 ค่าสังเกต ทำให้มีการนับจำนวนค่าสังเกตที่ $Y_i = 1$ ในลักษณะใหม่ Y_i ในกรณีหลังนี้จึงเป็นตัวแปรสุ่มที่มีการแจกแจงทวินาม $bin(n_i, \pi(x_i))$, $i = 1, 2, ..., I$ เมื่อ $I = n$ ด้วยค่าเฉลี่ย $E(Y_i) = n_i \pi(x_i)$ เมื่อ $n_1 + n_2 + ... + n_I = n$ โดยที่ฟังก์ชันความน่าจะเป็นร่วม (Joint Probability Mass Function) ของ $(Y_1, ..., Y_I)$ คือ

$$\prod_{i=1}^{I} \pi(x_i)^{y_i} [1 - \pi(x_i)]^{n_i - y_i} = \left\{\prod_{i=1}^{I} \left[1 - \pi(x_i)^{n_i}\right]\right\} \left\{\prod_{i=} \exp\left[\log\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right)^{y_i}\right]\right\}$$

$$= \left\{\prod_{i=1}^{I} \left[1 - \pi(x_i)^{n_i}\right]\right\} \exp\left[\sum_{i=1}^{I} y_i \log\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right)\right]$$

แต่ $$\log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij} \qquad \text{เมื่อ} \quad j = 1,2,\ldots, p$$

และ $$[1-\pi(x_i)] = \left[1+\exp\left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right]^{-1}$$

ดังนั้นฟังก์ชันควรจะเป็น (Likelihood Function) มีค่าเท่ากับ

$$\left\{\prod_{i=1}^{I}\left[1+\exp\left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right]^{-n_i}\right\}\exp\left[\sum_{i=1}^{I} y_i\left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right]$$

และฟังก์ชันล็อกควรจะเป็นเขียนแทนด้วย $L(\beta)$ ที่มีค่าเท่ากับ

$$L(\beta) = \left[\beta_0 + \sum_{j}\left(\sum_{i} y_i x_{ij}\right)\beta_j\right] - \left\{\sum_{i=1}^{I} n_i \log\left[1+\exp\left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right]\right\}$$

จะเห็นว่าฟังก์ชัน $L(\beta)$ ขึ้นอยู่กับจำนวนนับแบบทวินาม (Binomial Count) โดยสังเกตจากตัวสถิติพอเพียง (Sufficient Statistics) นั่นคือ $\sum_i y_i x_{ij}$ เมื่อ $j = 1,2,\ldots, p$ ดังนั้น

$$\frac{\partial L(\beta)}{\partial \beta_a} = \sum_i y_i x_{ia} - \sum_i n_i x_{ia}\left[\frac{\exp\left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)}{1+\exp\left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)}\right]$$

และสมการปรกติจะอยู่ในรูป

$$\sum_i y_i x_{ia} - \sum_i n_i \hat{\pi}_i x_{ia} = 0 \quad ; \quad a = 1,2,\ldots, p \qquad (2.1)$$

โดยที่

$$\hat{\pi}_i = \frac{\exp\left(\hat{\beta}_0 + \sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)}{1+\exp\left(\hat{\beta}_0 + \sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)}$$

เป็นตัวประมาณความควรจะเป็นสูงสุด (MLE) ของ $\pi(x_i)$

สำหรับการแก้สมการ (2.1) อาจใช้วิธีนิวตัน ราฟสัน หรือใช้วิธีฟิชเชอร์สกอริงก็ได้ ถ้าให้ X แทนเมทริกซ์ขนาด $I\times(p+1)$ ของ x_{ij} และให้ $\hat{m}_i = n_i\hat{\pi}_i$ แล้วสมการ (2.1) จะสามารถเขียนได้ในรูปแบบใหม่คือ

$$X'Y = X'\hat{m} \qquad (2.2)$$

นั่นคือ ตัวสถิติที่พอเพียง เท่ากับตัวประมาณของค่าคาดหวังและสมการ (2.2) นี้จะคล้ายกับสมการปรกติจากตัวแบบการถดถอยที่ใช้วิธีกำลังสองน้อยสุดกล่าวคือ

$$X'y = X'\hat{y} \qquad \text{เมื่อ} \qquad \hat{y} = X\hat{\beta}$$

นอกจากนี้

$$\frac{\partial^2 L(\beta)}{\partial\beta_a\partial\beta_b} = -\sum_i \frac{x_{ia}x_{ib}n_i \exp\left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)}{\left[1+\exp\left(\beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}\right)\right]^2}$$

$$= -\sum_i x_{ia}x_{ib}n_i\hat{\pi}_i(1-\hat{\pi}_i)$$

ในการประมาณค่าพารามิเตอร์ของตัวแบบไม่ว่าจะใช้วิธีนิวตัน ราฟสัน หรือใช้วิธีฟิชเชอร์สกอริง พบว่ามีสิ่งที่เหมือนกันคือ กระบวนการย้อนซ้ำต่างก็อาศัยวิธีภาวะน่าจะเป็นสูงสุดก่อน แล้วใช้การประมาณค่าจากสมการปรกติในลักษณะของวิธีกำลังสองน้อยที่สุดแบบถ่วงน้ำหนัก โดยมีกระบวนการสร้างตัวถ่วงน้ำหนักขึ้นใหม่ในทุกๆ รอบของการย้อนซ้ำจนกระทั่งได้ตัวประมาณที่ลู่เข้าและมีคุณสมบัติตามต้องการ เช่น ความพอเพียง ความคงเว้นคงวา เป็นต้น วิธีดังกล่าวนี้จึงอาจเรียกรวมกันได้ว่า การประมาณค่าด้วยวิธีภาวะน่าจะเป็นสูงสุดแบบย้อนซ้ำ

2.1.3 การตรวจสอบความเหมาะสมของตัวแบบการถดถอยโลจิสติกทวิภาค

การวิเคราะห์การถดถอยเชิงเส้นนั้นเป็นวิธีการที่ใช้ในการศึกษาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอธิบาย เมื่อตัวแปรตามเป็นตัวแปรเชิงปริมาณและสนใจค่าพยากรณ์ของตัวแปรตามที่เป็นเชิงตัวเลข แต่สำหรับการวิเคราะห์การถดถอยโลจิสติกทวิภาคนั้นค่าพยากรณ์ที่สนใจคือการเกิดเหตุการณ์ที่สนใจและไม่สนใจมากกว่าการพยากรณ์เชิงตัวเลข

เนื่องจากการประเมินตัวแบบของการวิเคราะห์การถดถอยเชิงเส้นจะอยู่ภายใต้ผลบวกกำลังสอง 2 ค่าคือ SST และ SSE โดยที่ SST คำนวณจาก $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ เมื่อ $\bar{y}$ คือค่าพยากรณ์ของตัวแปรตามกรณีที่ไม่มีตัวแปรอธิบายอยู่ใน ตัวแบบการถดถอย SST จะวัดความเบี่ยงเบนในการพยากรณ์ของค่าสังเกต y_i รอบ $\bar{y}$ ส่วน SSE คำนวณจาก $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ เมื่อ $\hat{y}_i$ คือค่าพยากรณ์ของตัวแปรตามกรณีที่มีตัวแปรอธิบายในตัวแบบ ค่า $\hat{y}_i$ ได้จาก $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_j x_{ij}$ โดยค่า SSE จะวัดความเบี่ยงเบนในการพยากรณ์จากความแปรผันของค่าสังเกต y_i รอบเส้นถดถอย สำหรับการประมาณค่าพารามิเตอร์ของสัมประสิทธิ์การถดถอย $\beta_j ; j = 1,2,..., p$ จะใช้วิธีกำลังสองน้อยสุดแบบสามัญ (Ordinary Least Square, OLS) และสามารถหาค่าผลบวกกำลังสองของตัวแบบการถดถอย (Regression Sum of Squares, SSR) จากผลต่างระหว่าง SST และ SSE คือ $SSR = SST - SSE$

จากตัวแบบที่แสดงความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอธิบาย ก่อนการนำไปใช้ต้องทำการตรวจสอบความเหมาะสมของตัวแบบที่ได้ ซึ่งวิธีการตรวจสอบความเหมาะสมของตัวแบบการถดถอยโลจิสติกทวิภาคนั้นมีหลายวิธี แต่ในที่นี้จะกล่าวถึงเพียง 3 วิธีดังนี้

(1) ใช้การทดสอบด้วยค่าสถิติ F ซึ่งมีสมมุติฐานในการทดสอบที่สมมูลกัน 2 สมมุติฐาน คือ $H_0 : R^2 = 0$ และ $H_0 = \beta_1 = ... = \beta_p = 0$ ของการถดถอยเชิงเส้นที่ได้จากการประมาณพารามิเตอร์ด้วยวิธีกำลังสองน้อยสุดแบบสามัญสามารถคำนวณได้ดังนี้

$$F = \left(\frac{SSR}{p}\right) \Big/ \left(\frac{SSE}{n-p-1}\right) = \left(\frac{n-p-1}{p}\right)\frac{SSR}{SSE}$$

เมื่อ n คือขนาดตัวอย่างและ p คือจำนวนตัวแปรอธิบาย ที่ระดับนัยสำคัญ α ถ้าค่า F น้อยกว่า $F_{p,n-p-1}$ จากตารางจะยอมรับสมมุติฐาน H_0 หมายความว่าตัวแปรตามและตัวแปรอธิบายไม่มีความสัมพันธ์เชิงเส้นกัน ในทำนองกลับกันถ้าค่า F มากกว่า $F_{p,n-p-1}$ จากตารางจะปฏิเสธสมมุติฐาน H_0 หมายความว่ามีความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอธิบายอย่างน้อย 1 ตัว นั่นคือตัวแบบ $Y = \beta_1 X_1 + ... + \beta_p X_p + \varepsilon$ มีนัยสำคัญทางสถิติ

(2) ใช้ตัวสถิติที่คำนวณจากตารางจำแนก (Classification Table) ของการถดถอยโลจิสติกทวิภาค โดยดูว่าค่าสังเกตของตัวแปรตามมีค่าเป็น 0 หรือ 1 และค่าพยากรณ์ มีค่าเท่ากับค่าใด ถ้าค่าความน่าจะเป็นของการพยากรณ์มีค่าเท่ากับหรือมากกว่า p ที่กำหนด ให้ถือ

ว่าค่าพยากรณ์เป็น 1 และถ้าค่าความน่าจะเป็นของการพยากรณ์มีค่าน้อยกว่า p ที่กำหนดให้ถือว่าค่าพยากรณ์เป็น 0 และนับจำนวนค่าสังเกตและค่าพยากรณ์ที่ให้ผลตรงกัน หารด้วยจำนวนค่าสังเกตทั้งหมดแล้วคูณด้วย 100 หรืออาจคำนวณหาร้อยละของการพยากรณ์ถูกต้องของตัวแปรตามที่มีค่าเป็น 0 หรือ 1 แต่การจำแนกค่าสังเกตว่าพยากรณ์ได้ถูกต้อง หรือไม่นั้นจะไม่ถือว่าเป็นมาตรวัดที่ดีที่ใช้ในการตรวจสอบความเหมาะสมของตัวแบบ เนื่องจากไม่สามารถบอกความแตกต่างระหว่างค่าพยากรณ์กับค่าสังเกตได้

(3) ใช้ค่าสัมประสิทธิ์การตัดสินใจ $\left(R^2\right)$ ในการวิเคราะห์การถดถอยเชิงเส้น ซึ่งเป็นค่าที่แสดงถึงสัดส่วนของความแปรผันของตัวแปรตามที่อธิบายได้ด้วยตัวแปรอธิบาย และเช่นเดียวกับการวิเคราะห์การถดถอยโลจิสติกทวิภาคที่ค่า R^2 จะใช้ในการวัดระดับความสัมพันธ์เพียงแต่การใช้ค่า R^2 สำหรับการวิเคราะห์การถดถอยโลจิสติกทวิภาคนั้นจะมีความยุ่งยากมากกว่าค่า R^2 ของการวิเคราะห์การถดถอยเชิงเส้น เนื่องจากสามารถคำนวณได้หลายวิธี โดยที่แต่ละวิธีอาจมีแนวคิดหรือวิธีการคำนวณที่แตกต่างกัน นอกจากนี้ยังก่อให้เกิดความสับสนในการอธิบายและการนำไปใช้ประโยชน์ ดังนั้นจึงควรศึกษาลักษณะที่ดีของตัววัด R^2 โดยควาลซีท (Kvalseth, 1985) ได้เสนอคุณสมบัติที่ดีของค่า R^2 ของสมการถดถอยเชิงเส้น (Linear Regression) ไว้ 8 ประการ ดังต่อไปนี้

1) R^2 ต้องเป็นเครื่องวัดความเหมาะสมของตัวแบบและสามารถอธิบายความหมายได้อย่างสมเหตุสมผล

2) R^2 ควรเป็นอิสระจากหน่วยในการวัดของตัวแปรในตัวแบบ

3) ขอบเขตของ R^2 ควรมีค่าอยู่ระหว่าง 0 และ 1 โดย $R^2=1$ เมื่อตัวแบบสามารถพยากรณ์ข้อมูลได้อย่างสมบูรณ์ แต่ถ้า $R^2=0$ แสดงว่าตัวแปรอธิบายไม่มีความสัมพันธ์กับตัวแปรตาม

4) R^2 ควรจะประยุกต์ใช้ได้กับตัวแบบชนิดต่างๆ และไม่คำนึงถึงคุณสมบัติทางสถิติของตัวแปรในตัวแบบ (ครอบคลุมถึงค่าคลาดเคลื่อน ε)

5) R^2 ไม่ควรขึ้นกับเทคนิคที่ใช้ในการประมาณตัวแบบการถดถอย นั่นคือ ค่า R^2 จะวัดความเหมาะสมของตัวแบบโดยไม่คำนึงถึงวิธีที่ได้มาของตัวแบบ

6) R^2 ของตัวแบบต่างๆ ที่มาจากข้อมูลชุดเดียวกันสามารถเปรียบเทียบกันได้โดยตรง

7) ค่าสัมพัทธ์ของ R^2 น่าจะสอดคล้องกับค่าวัดอื่นๆ ซึ่งเป็นที่ยอมรับในการวัดความเหมาะสมของตัวแบบการถดถอย เช่น ค่าคลาดเคลื่อนมาตรฐานของค่าประมาณของตัวแปรตาม $\left(\hat{Y}\right)$ และรากที่สองของค่าเฉลี่ยคลาดเคลื่อนกำลังสอง $\left(MSE\right)$

8) ค่า R^2 ควรให้น้ำหนักของส่วนตกค้าง (Residuals) ทางบวกและทางลบเท่าๆ กัน

ในปัจจุบันมีโปรแกรมสำเร็จรูปทางสถิติที่สำคัญ เช่น SPSS SAS และ STATA เป็นต้น ที่รายงานค่า R^2 สำหรับการถดถอยโลจิสติกทวิภาค นอกจากนี้ได้มีผู้เสนอตัววัด R^2 สำหรับการถดถอยโลจิสติกทวิภาคไว้หลายแบบ ซึ่งสามารถจัดกลุ่มได้ตามวิธีการคำนวณ เช่น R^2 ที่คำนวณด้วยหลักการของสัดส่วนการลดลงของการกระจาย (Proportional Reduction in Dispersion) และ R^2 ที่คำนวณด้วยอาศัยฟังก์ชันควรจะเป็น (Likelihood Function) เป็นต้น โดยรายละเอียดของตัววัด R^2 เหล่านี้มีดังต่อไปนี้

(3.1) ตัววัด R^2 ที่คำนวณด้วยหลักการของสัดส่วนการลดลงของการกระจาย (R^2 Measure Based on Proportional Reduction in Dispersion)

ให้ $(y_i, x_i),\ i = 1, 2, ..., n$ คือข้อมูลค่าสังเกตจำนวน n ค่า เมื่อ y_i คือค่า ตัวแปรตามที่มีค่า 0 หรือ 1 ของค่าสังเกตที่ *i* และ x_i คือเวกเตอร์ของตัวแปรอิสระ และกำหนดให้

$$\hat{P}(y_i = 1 \mid x_i) = \hat{\pi}_i$$

และ
$$\hat{P}(y_i = 1) \equiv \bar{\pi} = \sum_{i=1}^{n} \frac{y_i}{n} = \bar{y}$$

นอกจากนั้น ยังกำหนดให้ $D(y_i)$ คือค่าที่วัดการกระจายสำหรับค่าสังเกตที่ i และ $D(y_i \mid x_i)$ คือค่าที่วัดการกระจายที่คำนวณได้จากการกำหนดเงื่อนไขของตัวแบบและเวกเตอร์ของตัวแปรอิสระ ตัววัด R^2 เหล่านี้จะระบุสัดส่วนของความผันแปรของตัวแปรตามที่อธิบายได้ด้วยตัวแปรอิสระ (Proportion of Explained Variation, PEV) ซึ่งมีรูปแบบโดยทั่วๆ ไปคือ

$$PEV = \frac{\sum_{i=1}^{n} D(y_i) - \sum_{i=1}^{n} D(y_i \mid x_i)}{\sum_{i=1}^{n} D(y_i)}$$

ค่า $D(y_i)$ และ $D(y_i \mid x_i)$ จะถูกกำหนดแตกต่างกันขึ้นอยู่กับชนิดของตัววัด R^2 ซึ่งสามารถแบ่งได้ดังนี้

(3.1.1) R^2 แบบกำลังสองน้อยสุดสามัญ (Ordinary Least Squared, R_O^2)

ในกรณีนี้จะกำหนดให้ $D(y_i) = (y_i - \bar{y})^2$ และ $D(y_i \mid x_i) = (y_i - \hat{\pi}_i)^2$ และ $R_O^2 = 1 - \frac{SSE}{SST}$ เมื่อ $SST = \sum_i D(y_i)$ และ $SSE = \sum_i D(y_i \mid x_i)$ ดังนั้นค่า R_O^2 สามารถเขียนได้ดังนี้

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

มิทเบิกค์และสเชมเปอร์ (Mittlbock and Schemper, 1996) พบว่า R_O^2 จะมีค่าเท่ากับค่ากำลังสองของสัมประสิทธิ์สหสัมพันธ์ระหว่างค่าพยากรณ์ของความน่าจะเป็นกับค่าสังเกตของการเป็นสมาชิกหรือไม่เป็นสมาชิกในแต่ละระดับ และยังพบว่า ค่า R_O^2 นี้จะสอดคล้องกับค่า R^2 ของตัวแบบการถดถอยเชิงเส้น อย่างไรก็ตาม มินาร์ด (Menard, 2000) ให้ความเห็นว่าแนวคิดของ R_O^2 นั้นไม่เหมือนกับแนวคิดของ R^2 ของตัวแบบการถดถอยเชิงเส้น ทั้งนี้เพราะการประมาณค่าพารามิเตอร์ของตัวแบบการถดถอยโลจิสติกทวิภาคนั้น จะใช้วิธีความควรจะเป็นสูงสุด (Maximum Likelihood Method, ML) ในขณะที่ตัวแบบการถดถอยเชิงเส้นนั้น จะใช้วิธีกำลังสองน้อยสุดสามัญ (Ordinary Least Square, OLS)

(3.1.2) R^2 แบบอัตราส่วนความควรจะเป็น (Likelihood Ratio, R_L^2)

ให้ L_0 แทนฟังก์ชันควรจะเป็นสำหรับตัวแบบที่ประกอบด้วยค่าคงที่เท่านั้น

L_M แทน ฟังก์ชันควรจะเป็นสำหรับตัวแบบที่ประกอบด้วยตัวแปรอิสระที่สนใจ

L_S แทนฟังก์ชันควรจะเป็นสำหรับตัวแบบเต็ม (Saturated Model) หรือตัวแบบที่มีจำนวนพารามิเตอร์เท่ากับจำนวนค่าสังเกต และกำหนดสถิติดีเวียนซ์ (Deviance) ของตัวแบบที่มีตัวแปรอิสระ (D_M) และตัวแปรที่มีเฉพาะค่าคงที่ (D_0) เป็นดังต่อไปนี้

$$D_M \equiv -2(\log L_M - \log L_S) \equiv -2\log(L_M / L_S)$$
$$D_0 \equiv -2(\log L_0 - \log L_S) \equiv -2\log(L_0 / L_S)$$

ดังนั้นรูปแบบทั่วไปของตัววัดที่คำนวณด้วยฟังก์ชันควรจะเป็นคือ

$$G=\frac{\log L_M - \log L_0}{\log L_S - \log L_0}=\frac{D_0 - D_M}{D_0}$$

สำหรับตัวแบบการถดถอยโลจิสติกทวิภาค ดังนั้นสำหรับตัวแบบเต็ม (Saturated Model) ค่าประมาณความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจเมื่อกำหนดเวกเตอร์ของตัวแปรอิสระจะมีค่าเท่ากับค่าสังเกตนั้นๆหรือ $\hat{P}(y_i=1|x_i)\equiv\hat{\pi}_i=y_i$ และฟังก์ชันควรจะเป็นจะมีค่าเป็น 1 นั่นคือ

$$L_S=\prod_{i=1}^{n} y_i^{y_i}\times(1-y_i)^{(1-y_i)}=1$$

นอกจากนั้น โฮสเมอร์และเลมโชว์ (Hosmer and Lemeshow ,2000) ให้ข้อสังเกตว่าค่าสถิติดีเวียนซ์สำหรับตัวแบบโลจิสติกทวิภาคจะมีค่าเท่ากับผลรวมความคลาดเคลื่อนกำลังสอง (Sum of Square Error: SSE) ที่คำนวณโดยวิธีกำลังสองน้อยสุดแบบสามัญ (Ordinary Least Square, OLS) นั่นคือ จะได้ว่าค่า $D_M=-2\log L_M$ จะมีค่าเท่ากับ SSE ของตัวแบบที่มีตัวแปรอิสระ และ $D_0=-2\log L_0$ จะมีค่าเท่ากับ SSE ของตัวแบบที่มีเฉพาะค่าคงที่ซึ่งเปรียบเสมือนได้กับค่าผลรวมกำลังสองทั้งหมด (Total Sum of Square, SST) ของ OLS ดังนั้น R_L^2 สำหรับตัวแบบถดถอยโลจิสติกทวิภาค สามารถเขียนได้ดังนี้

$$\begin{aligned}R_L^2 &= 1-\frac{\log(L_M)}{\log(L_0)}\\ &=\frac{\log L_0 - \log L_M}{\log L_0}\end{aligned}$$

ค่า R_L^2 นี้โดยทั่วไปจะถูกใช้ในโปรแกรมสำเร็จรูปต่างๆทางสถิติ ที่มีชื่อเรียกว่า McFadden's pseudo R^2

(3.2) ตัววัด R^2 ที่คำนวณด้วยฟังก์ชันควรจะเป็น (R^2 Measures Based on Likelihood Function)

ตัววัด R^2 ที่คำนวณด้วยฟังก์ชันควรจะเป็น นอกจากตัววัด R_L^2 ที่ได้กล่าวไปในหัวข้อ (3.1.3) แล้ว ยังมีตัววัด R^2 อีกหลายตัวดังต่อไปนี้

(3.2.1) R^2 ที่ไม่ปรับและที่ปรับค่าโดยคำนวณจากการปรับปรุงค่ากำลังสองเฉลี่ยเรขาคณิต (Unadjusted and Adjusted Geometric Mean Square Improvement, R_M^2, R_N^2)

แมดดาลาร์ (Maddala, 1983) และแมกกี (Mcgee, 1990) ได้เสนอค่า R_M^2 ที่คำนวณได้จาก

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

เมื่อ L_0 และ L_M นิยามได้เช่นเดียวกับข้างต้น เนื่องจากค่า L_M เป็นฟังก์ชันผลคูณของความน่าจะเป็น ดังนั้นจึงทำให้ค่า R_M^2 มีค่าต่ำกว่า 1 และมีค่าสูงสุดเท่ากับ

$$\max(R_M^2) = 1 - (L_0)^{2/n}$$

นาเกลเคิร์ก (Nagelkerke, 1991) พบว่าสำหรับตัวแบบถดถอยโลจิสติกทวิภาคที่มีค่า $y = 1$ และ $y = 0$ ในสัดส่วนที่เท่าๆกันแล้ว ค่าสูงสุดของค่า R_M^2 จะมีค่าเพียง 0.75 เท่านั้น ดังนั้นนาเกลเคิร์กจึงได้เสนอให้มีการปรับค่า R_M^2 โดยการนำค่าสูงสุดไปหาค่า R_M^2 เดิมและเรียกว่าที่ได้จากการปรับนี้ว่า R_N^2 ซึ่งสามารถคำนวณได้จากสูตร

$$R_N^2 = \frac{1 - \left(\frac{L_0}{L_M}\right)^{2/n}}{1 - L_0^{2/n}}$$

(3.2.2) R^2 ที่ไม่ปรับค่าและที่ปรับค่าโดยคำนวณจากสัมประสิทธิ์ตารางการณ์จร (Unadjusted and Adjusted Contingency Coefficient, R_C^2 , R_{CS}^2)

อัลดริชและเนลสัน (Aldrich and Nelson, 1984) ได้เสนอ R^2 ที่คำนวณจากสัมประสิทธิ์ตารางการณ์จรโดยมีสูตรการคำนวณดังนี้

$$R_C^2 = \frac{G_M}{G_M + n}$$

เมื่อ $G_M = -2\log\left(\frac{L_0}{L_M}\right)$ ซึ่งค่า R_C^2 มีขอบเขตบนไม่ถึง 1 ถึงแม้ว่าตัวแบบมีความเหมาะสมกับข้อมูลอย่างสมบูรณ์ก็ตาม นอกจากนั้นค่า R_C^2 ยังขึ้นอยู่กับขนาดตัวอย่าง นั่นหมายความว่า

ขอบเขตบนของ R_C^2 นั้นจะเปลี่ยนไปเมื่อชุดของข้อมูลเปลี่ยนไป แฮกเกิ้ลและมิทเชล (Haggle and Mitchell) พบว่าค่าสูงสุดของ R_C^2 มีค่าเท่ากับ

$$\max(R_C^2) = \frac{-2[\bar{y}\log\bar{y} + (1-\bar{y})\log(1-\bar{y})]}{1-2[\bar{y}\log\bar{y} + (1-\bar{y})\log(1-\bar{y})]}$$

เมื่อ $\bar{y} = \sum_{i=1}^{n} y_i / n$ คือสัดส่วนของ $y = 1$ ในตัวอย่าง ดังนั้น แฮกเกิ้ลและมิทเชลได้ปรับค่าของ R_C^2 ด้วยค่าสูงสุดของมันทำให้ได้ค่า R_{CS}^2 ที่เขียนเป็นสมการได้ดังนี้

$$R_{CS}^2 = \frac{R_C^2}{\max(R_C^2)}$$

แต่พบว่าค่า R^2 ในตัวแบบการถดถอยโลจิสติกทวิภาคบางครั้งก็ให้ขนาดของความสัมพันธ์ที่มากเกินไปทั้งที่ตัวแปรอธิบายที่อยู่ในตัวแบบนั้นไม่มีความสัมพันธ์กับตัวแปรตาม เพราะฉะนั้นสำหรับตัวแบบการถดถอยโลจิสติกทวิภาคที่ประกอบด้วยตัวแปรอธิบายที่มีมากกว่าหนึ่งตัวแปร การใช้สัมประสิทธิ์การตัดสินใจที่ปรับค่า $\left(R_{adj}^2\right)$ จึงเหมาะสำหรับใช้ในการตรวจสอบความเหมาะสมของตัวแบบมากกว่าการใช้ R^2

อย่างไรก็ตามในการตรวจสอบความเหมาะสมของตัวแบบการถดถอยโลจิสติกทวิภาค โดยการใช้ R_{adj}^2 สามารถลดปัญหาที่เกิดจากการใช้ R^2 ลงได้ เนื่องจาก R_{adj}^2 มีประโยชน์สำหรับใช้เปรียบเทียบตัวแบบการถดถอยที่มีจำนวนตัวแปรอธิบายแตกต่างกัน และการที่ค่า R_{adj}^2 ลดลงเมื่อเพิ่มตัวแปรอธิบายเข้าไปในตัวแบบนั้น แสดงให้เห็นว่าตัวแปรอธิบายที่เพิ่มเข้าไปมีความสัมพันธ์กับตัวแปรตามน้อยมาก มีงานวิจัยต่างๆ ที่สนับสนุนให้ใช้ R_{adj}^2 มากกว่า R^2 สำหรับตัวแบบการถดถอยประกอบด้วยตัวแปรอธิบายตั้งแต่ 2 ตัวขึ้นไป โดยจะนำเสนอในหัวข้อต่อไป

2.2 งานวิจัยที่เกี่ยวข้อง

2.2.1 การปรับค่าสัมประสิทธิ์การตัดสินใจสำหรับการถดถอยโลจิสติกทวิภาค

มีผู้เสนอวิธีการปรับค่าสัมประสิทธิ์การตัดสินใจสำหรับการถดถอยโลจิสติกทวิภาคไว้หลายวิธี เช่น มิทเบิกค์และสเชมเปอร์ (Mittlbock and Schemper, 1996 ได้ศึกษาค่า R^2

ที่ปรับค่าที่คำนวณด้วยวิธีกำลังสองน้อยสุดแบบสามัญ (Ordinary Least Squares, OLS) และวิธีความควรจะเป็นสูงสุด (Maximum Likelihood, ML) ในที่นี้ใช้สัญลักษณ์ $R^2_{O,adj,MS}$ และ $R^2_{l,adj,MS}$ สำหรับการปรับค่าที่ใช้คำนวณด้วยวิธีกำลังสองน้อยสุดแบบสามัญและวิธีความควรจะเป็นสูงสุดตามลำดับ รูปแบบของ $R^2_{O,adj,MS}$ และ $R^2_{l,adj,MS}$ แสดงได้ดังนี้

$$R^2_{O,adj,MS} = 1 - \frac{(n-1)\sum_{i=1}^{n}(y_i - \hat{\pi}_i)^2}{(n-p-1)\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.3}$$

$$R^2_{l,adj,MS} = 1 - \frac{l(y,\hat{\pi}) - (p+1)/2}{l(y,\hat{\pi}^0) - 1/2} \tag{2.4}$$

เมื่อ y_i คือค่าสังเกตของตัวแปรตามแบบทวิภาค

n คือขนาดตัวอย่าง

p คือจำนวนตัวแปรอธิบายในตัวแบบการถดถอย

$\hat{\pi}_i$ คือค่าประมาณของ π_i กรณีที่ตัวแบบมีตัวแปรอธิบาย p ตัว

$\bar{y}$ คือค่าเฉลี่ยของตัวแปรตามแบบทวิภาค $\bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{n}$

$l(y,\hat{\pi})$ คือฟังก์ชันล็อกควรจะเป็นสูงสุดสำหรับตัวแบบที่มีตัวแปรอธิบาย p ตัว ซึ่งคำนวณจาก $\sum_{i=1}^{n}\left[y_i \log(\hat{\pi}_i) + (1-y_i)\log(1-\hat{\pi}_i)\right]$

$l(y,\hat{\pi}^0)$ คือฟังก์ชันล็อกควรจะเป็นสูงสุดสำหรับตัวแบบที่มีเพียงเทอมค่าคงที่ ซึ่งคำนวณจาก $\sum_{i=1}^{n}\left[y_i \log(\hat{\pi}_i^0) + (1-y_i)\log(1-\hat{\pi}_i^0)\right]$

นอกจากนี้มิทเบิกค์และสเชมเปอร์ ได้ทำการเปรียบเทียบสัมประสิทธิ์การตัดสินใจของตัวแบบการถดถอยโลจิสติกทั้งที่มีการปรับค่า และไม่ปรับค่าที่คำนวณด้วยวิธีกำลังสองน้อยสุดแบบสามัญและวิธีความควรจะเป็นสูงสุด เมื่อตัวแปรอธิบายมีการแจกแจงแบร์นูลลีที่มีค่า 0 และ 1 ในสัดส่วนเท่าๆ กัน จากการศึกษาพบว่า สัมประสิทธิ์การตัดสินใจที่คำนวณด้วยวิธีกำลังสองน้อยสุดแบบสามัญนั้นสัมประสิทธิ์การตัดสินใจที่ปรับค่าเหมาะสำหรับใช้ในการวัดความเหมาะสมของตัวแบบมากกว่าสัมประสิทธิ์การตัดสินใจที่ไม่ปรับค่า ถ้าอัตราส่วน $p/n = 0.2$ และทำนองเดียวกันสำหรับสัมประสิทธิ์การตัดสินใจที่คำนวณด้วยวิธีความควรจะเป็นสูงสุด พบว่าสัมประสิทธิ์การตัดสินใจที่ปรับค่าให้ผลที่ดีกว่าสัมประสิทธิ์การตัดสินใจที่ไม่ปรับค่าด้วยการใช้

ความถูกต้องของ $(p+1)/2$ และยังแนะนำว่าถ้าจะใช้สัมประสิทธิ์การตัดสินใจที่ปรับค่าสำหรับการเปรียบเทียบตัวแบบการถดถอยโลจิสติก ที่ใช้ขั้นตอนการคัดเลือกตัวแปรอธิบายแบบขั้น (Stepwise) ควรให้ค่า p เป็นจำนวนตัวแปรอธิบายที่เข้าไปทำการคัดเลือกในตัวแบบทั้งหมดมากกว่าที่จะเป็นจำนวนตัวแปรอธิบายในตัวแบบการถดถอยที่ได้ในขั้นสุดท้าย

เลียโอะและแมคกี (Liao and McGee, 2003) ได้เสนอ R^2_{adj} ของการถดถอยโลจิสติกเพื่อใช้แก้ปัญหาของ R^2 ที่ให้ขนาดความสัมพันธ์ที่มากเกินไป และมีค่าสูงขึ้นเมื่อขนาดตัวอย่างลดลงโดยที่ตัวแบบมีตัวแปรอธิบายมากใกล้ตัวแบบเต็ม (Saturated Model) โดยมีการปรับค่า R^2 ด้วยค่าคลาดเคลื่อนที่คำนวณโดยใช้ขั้นตอน 2 ขั้นตอนคือ ขั้นแรกคำนวณค่า IPE หลังจากนั้นทำการประมาณค่าความเอนเอียง (Bias-Corrected Estimator) ของค่า IPE โดยใช้หลักการเช่นเดียวกับการประมาณค่าความแปรปรวน (σ^2) ในตัวแบบการถดถอยด้วยวิธี $REML$ (Restricted or Residual Maximum Likelihood) ดังนั้นวิธีการของเลียโอะและแมคกีค่าตัวแปรตามจะถูกแยกออกเป็น 2 แบบ คือตัวแปรตามมีการแจกแจงแบร์นูลลี $(y_i \sim Bernoulli(\pi_i))$ และ $y_i^{new} \sim Bernoulli(\hat{\pi}_i)$ โดยค่า $\hat{\pi}_i$ มีค่าเท่ากับ $n^{-1}\sum_{i=1}^{n}(y_i-\hat{\pi}_i)^2$ และ $-n^{-1}l(y,\hat{\pi})$ ในที่นี้ใช้สัญลักษณ์ $R^2_{O,adj,LM}$ และ $R^2_{l,adj,LM}$ สำหรับการคำนวณด้วยวิธีกำลังสองน้อยสุดแบบสามัญและวิธีความควรจะเป็นสูงสุดตามลำดับ รูปแบบ $R^2_{O,adj,LM}$ และ $R^2_{l,adj,LM}$ แสดงได้ดังนี้

$$R^2_{O,adj,LM} = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{\pi}_i)^2 - \sum_{i=1}^{n}\hat{\pi}_i^{new}\left(1-\hat{\pi}_i^{new}\right) + \sum_{i=1}^{n}\hat{\pi}_i(1-\hat{\pi}_i)}{\sum_{i=1}^{n}(y_i-\bar{y})^2 - \sum_{i=1}^{n}\bar{y}^{new}\left(1-\bar{y}^{new}\right) + \sum_{i=1}^{n}\bar{y}(1-\bar{y})} \tag{2.5}$$

และ

$$R^2_{l,adj,LM} = 1 - \frac{l(y,\hat{\pi}) - \sum_{i=1}^{n}\left[\hat{\pi}_i^{new}\log\hat{\pi}_i^{new} - \left(1-\hat{\pi}_i^{new}\right)\log\left(1-\hat{\pi}_i^{new}\right)\right] + \sum_{i=1}^{n}\left[\hat{\pi}_i\log\hat{\pi}_i + (1-\hat{\pi}_i)\log(1-\hat{\pi}_i)\right]}{l(y,\hat{\pi}^0) - \sum_{i=1}^{n}\left[\bar{y}^{new}\log\bar{y}^{new} - \left(1-\bar{y}^{new}\right)\log\left(1-\bar{y}^{new}\right)\right] + \sum_{i=1}^{n}\left[\bar{y}\log\bar{y} + (1-\bar{y})\log(1-\bar{y})\right]} \tag{2.6}$$

เมื่อ $\hat{\pi}^{new}$ คือค่าประมาณของ $\hat{\pi}_i$

y^{new} คือค่าของตัวแปรตามที่ใช้ในการวิเคราะห์ครั้งที่ 2

$\bar{y}^{new}$ คือค่าเฉลี่ยของตัวแปรตาม $\bar{y}^{new} = \frac{y_1^{new} + y_2^{new} + \ldots + y_n^{new}}{n}$

และ $l(y,\hat{\pi})$ $l(y,\hat{\pi}^0)$ และ n นิยามได้เช่นเดียวกับสมการ (2.3) และ (2.4)

ผลการศึกษาพบว่าเมื่อเพิ่มตัวแปรอธิบายที่เป็นอิสระกับตัวแปรตามเข้าในตัวแบบค่า $R^2_{l,adj,LM}$ ให้ผลที่ดีกว่าค่า $R^2_{O,adj,MS}$ โดยเฉพาะตัวแบบที่มีขนาดตัวอย่างเล็กหรือตัวแบบที่มีจำนวนตัวแปรอธิบายมาก

ชเตทแลนด์ (Shtatland, 2000) ได้แนะนำ $R^2_{SAS,AIC}$ ที่ได้มาจากการปรับค่าโดยอาศัยหลักเกณฑ์สารสนเทศของอาไคเคะ (Akaike's Information Criterion, AIC) ของโปรแกรม SAS ใน PROC GENMOD โดยมีขั้นตอนการปรับค่าดังนี้ ขั้นแรกปรับค่า R^2 โดยใช้หลักการปรับค่าเหมือนกับ R^2 ของการถดถอยเชิงเส้นด้วยการใช้จำนวนตัวแปรอธิบายหรือขนาดตัวอย่าง กรณีที่ใช้ทั้งจำนวนตัวแปรอธิบายและขนาดตัวอย่าง การปรับค่า R^2 ทำได้ดังนี้

$$ {}_1R^2_{adj,AIC} = 1 - \frac{l(y,\hat{\pi})}{l(y,\hat{\pi}^0)}\left(\frac{n-1}{n-p-1}\right) $$

ในกรณีที่ n มีขนาดใหญ่และ p มีจำนวนน้อย การปรับค่า R^2 อาจใช้แค่จำนวนตัวแปรอธิบายเท่านั้น โดยมีสูตรการปรับค่าดังนี้

$$ {}_2R^2_{adj,AIC} = 1 - \frac{l(y,\hat{\pi}) - (p+1)}{l(y,\hat{\pi}^0) - 1} $$

ขั้นตอนการปรับค่าขั้นต่อมา คือทำการปรับค่า ${}_1R^2_{adj,AIC}$ และ ${}_2R^2_{adj,AIC}$ พร้อมๆ กันโดยอาศัยวิธีการปรับแก้ของอาไคเคะ (Akaike's Type Correction) (Menard, 1995) ซึ่งจะได้ดังสมการต่อไปนี้

$$ R^2_{l,adj,SAS_{AIC}} = 1 - \frac{l(y,\hat{\pi}) - \left[(p+1)(n-1)/(n-p-1)\right]}{l(y,\hat{\pi}^0) - 1} \qquad (2.7) $$

ดังนั้นการปรับค่า $R^2_{l,adj,SAS_{AIC}}$ ที่ใช้หลักเกณฑ์สารสนเทศของอาไคเคะ (Akaike's Information Criterion, AIC) ก็เพื่อป้องกันความเอนเอียงที่อาจเกิดขึ้นในกรณีที่มีขนาดตัวอย่างน้อย

นารีรัตน์ ณ นุวงศ์ (2551) ได้ศึกษาเปรียบเทียบประสิทธิภาพของ R^2_{adj} ทั้งหมดที่ได้กล่าวไปข้างต้น เมื่อมีการกำหนดตัวแบบการถดถอยโลจิสติกไม่ถูกต้องใน 3 ลักษณะได้แก่ (1) รูปแบบของตัวแปรอธิบายไม่ถูกต้อง (2) กำหนดฟังก์ชันเชื่อมโยงของตัวแบบไม่ถูกต้อง (3)

กำหนดให้ตัวแบบมีตัวแปรอธิบายที่มีความสัมพันธ์กับตัวแปรตามขาดหายไป เมื่อกำหนดฟังก์ชันเชื่อมโยงที่แท้จริงเป็นฟังก์ชันโลจิท ผลการศึกษาพบว่า R^2_{adj} ที่คำนวณโดย ใช้วิธีความควรจะเป็นสูงสุดของเลียโอะและแมคกี มีความเหมาะสมมากสุดที่จะใช้ในการอธิบายสัดส่วนของความผันแปรที่เกิดขึ้นในความน่าจะเป็นที่จะเกิดความสำเร็จของตัวแปรตามเนื่องจากอิทธิพลของตัวแปรอธิบาย โดยเฉพาะในกรณีที่มีตัวแปรอธิบายจำนวนมากและขนาดตัวอย่างเล็ก

2.2.2 ปัญหาของสัดส่วนผลตอบสนอง

สัดส่วนผลตอบสนองของเหตุการณ์ใดๆ จะหมายถึงสัดส่วนของการเกิดเหตุการณ์นั้นๆ ในตัวอย่างที่สนใจศึกษา เขียนได้ในรูป $P(Y=1)=\bar{y}=\sum_{i=1}^{n}\frac{y_i}{n}$ และสำหรับการถดถอย โลจิสติกทวิภาคค่า $\bar{y}=\pi$ ซึ่งก็คือค่าเฉลี่ยของความน่าจะเป็นแบบมีเงื่อนไขของการเกิดเหตุการณ์ที่สนใจต่อรูปแบบของค่าสังเกตที่แตกต่างกันทั้งหมดที่เป็นไปได้นั่นเอง มีผู้ทำการศึกษาอิทธิพลสัดส่วนผลตอบสนองที่มีต่อค่า R^2 ไว้หลายท่าน เช่น อัชและชวาทส์ (Ahs and Shwartz, 1999) ได้ศึกษาอิทธิพลสัดส่วนผลตอบสนองที่มีต่อค่า R^2_O ซึ่งพิจารณาจากประชากร 2 กลุ่ม คือกลุ่มที่มีความเสี่ยงต่อการตายสูงและต่ำ โดยที่การตายนั้นสามารถเกิดได้เฉพาะกลุ่มใดกลุ่มหนึ่งเท่านั้น ผลการศึกษาพบว่าค่า R^2_O เป็นฟังก์ชันของสัดส่วนผลตอบสนอง

มิทเบิกค์และสเชมเปอร์ (Mittlbock and Schemper, 2002) ได้สนับสนุนงานของอัชและชวาทส์ (Ahs and Shwartz, 1999) ที่ว่าค่า R^2 ขึ้นอยู่กับค่าสัดส่วนผลตอบสนอง โดยถ้าสัดส่วนผลตอบสนองมีค่าน้อย (มีค่าเข้าใกล้ 0 หรือ 1) แล้วค่าตัวแปรตามจะถูกกำหนดค่อนข้างแน่นอน ทำให้ค่าความผันแปรที่เกิดขึ้นในตัวแปรตามมีค่าน้อยในทางตรงข้ามถ้าสัดส่วนผลตอบสนองมีค่ามาก (มีค่าเข้าใกล้ 0.5) ความผันแปรในตัวแปรตามจะมีค่ามากและตัวแปรอธิบายอาจจะอธิบายความผันแปรที่เกิดขึ้นในตัวแปรตามได้มาก

มินาร์ด (Menard, 2000) ได้ศึกษาอิทธิพลสัดส่วนผลตอบสนอง ที่มีต่อค่า R^2 5 วิธี คือ R^2_O, R^2_L, R^2_M, R^2_N และ R^2_C นอกจากนี้ยังได้ศึกษาดัชนีของประสิทธิภาพในการทำนายทั้ง 3 ตัว คือ λ_P , τ_P และ ϕ_P สำหรับตัวแบบการถดถอยโลจิสติกทวิภาค โดยกำหนดสัดส่วนผลตอบสนองระดับต่างๆ คือ 0.01, 0.021, 0.098, 0.133, 0.135, 0.40 และ 0.495 ผลการศึกษาพบว่า R^2_L มีสหสัมพันธ์กับสัดส่วนผลตอบสนองน้อยที่สุดและยังพบว่า R^2_L เป็นตัวสถิติที่สามารถอธิบายอัตราส่วนการลดลงของความคลาดเคลื่อนได้ดีกว่าค่า R^2 แบบอื่นๆ ดังนั้นจึงสรุปได้ว่า R^2_L เป็นตัววัดที่ดีที่สุด และสำหรับดัชนีของประสิทธิภาพในการทำนายนั้นพบว่า

τ_P ไม่ขึ้นกับสัดส่วนผลตอบสนองเมื่อเปรียบเทียบกับดัชนีของประสิทธิภาพในการทำนายแบบอื่นๆ

โซเดอร์สตรอม (Soderstrom, 1997) ได้ศึกษาศึกษาอิทธิพลของสัดส่วนผลตอบสนองที่มีต่อดัชนีของประสิทธิภาพในการทำนายทั้ง 3 ตัว คือ λ_P, τ_P และ ϕ_P สำหรับตัวแบบการถดถอยโลจิสติกทวิภาค 2 ตัวแบบ โดยที่ตัวแบบแรกประกอบด้วยตัวแปรอธิบาย 2 ตัว คือตัวแปรอธิบายที่เป็นแบบทวิภาคและแบบต่อเนื่อง และตัวแบบที่สองประกอบด้วยตัวแปรอธิบายที่ทั้งสองตัวเป็นแบบทวิภาค ภายใต้สัดส่วนผลตอบสนองระดับต่างๆ คือ 0.1, 0.3 และ 0.5 ซึ่งผลจากศึกษาพบว่า ϕ_P เป็นดัชนีที่ไม่เปลี่ยนแปลงตามสัดส่วนผลตอบสนองซึ่งให้ผลดีสุดสำหรับตัวแบบที่ประกอบด้วยตัวแปรอธิบายแบบทวิภาคทั้งคู่ แต่สำหรับ λ_P ยังคงมีความไวมากต่อการเปลี่ยนแปลงของสัดส่วนผลตอบสนองซึ่งสอดคล้องกับคำแนะนำของมินาร์ดในการเลือก τ_P เพื่อประเมินประสิทธิภาพในการทำนายของตัวแบบการถดถอยโลจิสติกทวิภาค

ธนิษฐา คำศรี (2546) ได้ศึกษาอิทธิพลสัดส่วนผลตอบสนองที่มีค่า R^2 5 วิธี คือ R_O^2, R_L^2, R_M^2, R_N^2 และ R_C^2 นอกจากนี้ยังได้ศึกษาดัชนีของประสิทธิภาพในการทำนายทั้ง 3 ตัว คือ λ_P, τ_P และ ϕ_P สำหรับตัวแบบการถดถอยโลจิสติกพหุคูณ ภายใต้สัดส่วนผลตอบสนองระดับต่างๆ คือ 0.01, 0.10, 0.20, 0.30, 0.40 และ 0.50 และกำหนดตัวแปรอธิบาย $X_i, i=1,2,3$ ประกอบด้วยตัวแปรเชิงคุณภาพและเชิงปริมาณ โดยที่ X_1 มีการแจกแจงปรกติ X_2 และ X_3 มีการแจกแจงแบร์นูลลี ผลการศึกษาพบว่าตัวแบบการถดถอยโลจิสติกพหุคูณ ที่ประกอบด้วยตัวแปรอธิบายทั้งที่เป็นตัวแปรเชิงคุณภาพและเชิงปริมาณตัวสถิติ R_C^2, R_M^2 และ R_O^2 ให้ค่าสัมบูรณ์ของสัมประสิทธิ์สหสัมพันธ์กับสัดส่วนผลตอบสนองต่ำสุด เมื่อเปรียบเทียบกับตัวสถิติอื่นๆ และมีแนวโน้มลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้นอีกด้วย จึงชี้ให้เห็นว่า R_C^2, R_M^2 และ R_O^2 เป็นตัวสถิติที่สามารถวัดความเหมาะสมของตัวแบบการถดถอยโลจิสติกในแง่ที่ไม่ขึ้นอยู่กับสัดส่วนผลตอบสนองได้ดีกว่าตัวสถิติอื่นๆ

ชาร์มา (Shama, 2006) ได้ศึกษาอิทธิพลสัดส่วนผลตอบสนองที่มีค่า R^2 ทั้งที่เป็นตัววัดแบบใช้พารามิเตอร์และไม่ใช้พารามิเตอร์ สำหรับตัวแบบการถดถอยโลจิสติก 2 แบบคือตัวแบบตัวแปรแฝงและตัวแบบตัวแปรกำหนด ในกรณีที่มีตัวแปรอธิบายเพียงตัวเดียวภายใต้ สัดส่วนผลตอบสนองระดับต่างๆ คือ 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45 และ 0.50 ผลการศึกษาพบว่าสำหรับตัวสถิติ R^2 แบบต่างๆ นั้น R_{CS}^2 เป็นตัวประมาณที่ดีที่สุด เนื่องจากให้ค่าความคลาดเคลื่อนกำลังสองเฉลี่ย (MSE) และความเอนเอียง (Bias) ต่ำสุด นอกจากนี้ยัง

พบว่าสัดส่วนผลตอบสนองไม่มีอิทธิพลต่อตัวสถิติ R_L^2 ดังนั้นจึงเป็นตัวสถิติที่สามารถวัดความเหมาะสมของตัวแบบการถดถอยโลจิสติกได้ดีกว่าตัวสถิติอื่นๆ

2.2.3 **ความเชื่อถือได้ในตัวแปรอธิบาย** (Reliability of Explanatory Variables)

ความเชื่อถือได้ของตัวแปรอธิบาย คือค่าที่ใช้แสดงถึงความน่าเชื่อถือในตัวแปรอธิบายของตัวแบบการถดถอยที่มีค่าอยู่ในช่วง(0,1) เช่น ถ้าความเชื่อถือได้ในตัวแปรอธิบายมีค่าเท่ากับ 0 หมายความว่า ตัวแปรอธิบายในตัวแบบนั้นไม่สามารถอธิบายความผันแปรที่เกิดขึ้นในตัวแปรตามได้ เนื่องจากเป็นตัวประมาณค่าที่ไม่มีเสถียรภาพและมีความคลาดเคลื่อนสูงมีผู้ศึกษาเกี่ยวกับอิทธิพลของความเชื่อถือได้ในตัวแปรอธิบายที่มีผลค่า R^2 เช่น ก็อตเฟรดสัน (Gottfredson, 1987) พบว่าระดับความเชื่อถือได้ของตัวแปรอธิบายและตัวแปรตามมีอิทธิพลต่อค่า R^2 ของตัวแบบการถดถอยโลจิสติก ทั้งนี้เพราะตัวแปรในตัวแบบเป็นสาเหตุสำคัญที่ทำให้ค่าความคลาดเคลื่อนในการพยากรณ์มีค่าสูงหรือต่ำ ถ้าตัวแบบมีความคลาดเคลื่อนสูงอาจแก้ปัญหาโดยการเพิ่มขนาดตัวอย่างซึ่งทำให้มีผลกระทบหลายอย่างตามมา ดังนั้นตัวแปรที่ดีควรมีระดับความเชื่อถือได้ที่มีค่าสูงๆ

โซเดอร์สตรอม (Soderstrom, 1997) ทำการศึกษาอิทธิพลของความเชื่อถือได้ในตัวแปรอธิบายที่มีต่อดัชนีของประสิทธิภาพในการทำนายทั้ง 3 ตัว คือ λ_P, τ_P และ ϕ_P สำหรับตัวแบบการถดถอยโลจิสติก 2 ตัวแบบ โดยที่ตัวแบบแรกประกอบด้วยตัวแปรอธิบาย 2 ตัว คือ ตัวแปรอธิบายที่เป็นแบบทวิภาคและแบบต่อเนื่อง และตัวแบบที่สองประกอบด้วยตัวแปรอธิบายที่ทั้งสองตัวเป็นแบบทวิภาค ซึ่งตัวแปรอธิบายในตัวแบบที่กล่าวมานั้นจะกำหนดให้มีระดับความเชื่อถือได้ 2 ระดับ คือสูง (0.9) และต่ำ (0.6) ผลการศึกษาพบว่า การกำหนดความเชื่อถือได้ในตัวแปรอธิบายระดับสูงทำให้ดัชนีของประสิทธิภาพในการทำนายมีประสิทธิภาพดีกว่าการกำหนดความเชื่อถือได้ในตัวแปรอธิบายระดับต่ำ เนื่องจากความเชื่อถือได้ในระดับสูงไม่มีอิทธิพลต่อดัชนีของประสิทธิภาพในการทำนายทั้ง 3 ตัว

2.2.4 **อัตราการจำแนกผิดในตัวแปรตาม** (Misclassification of Dependent Variables)

อัตราการจำแนกผิดในตัวแปรตาม Y ในที่นี้จะหมายถึง การที่ค่า Y ได้ถูกบันทึกค่าที่แตกต่างไปจากค่าที่แท้จริง เช่น Y ถูกบันทึกเป็น 0 เมื่อค่าจริงของ Y เป็น 1 ซึ่งพิจารณาได้จากค่าความน่าจะเป็นซึ่งเขียนได้ในรูปของ

$$\alpha_0 = P(y_\alpha = 1 \mid y = 0)$$

$$\alpha_1 = P(y_\alpha = 0 \mid y = 1)$$

เมื่อ α_0 และ α_1 คือ ค่าความน่าจะเป็นในการกำหนดให้ตัวแปรตาม (y_α) มีค่าเป็น 1 เมื่อตัวแปรตามที่แท้จริง (y) มีค่าเป็น 0 และค่าความน่าจะเป็นในการกำหนดให้ตัวแปรตาม (y_α) มีค่าเป็น 0 เมื่อตัวแปรตามที่แท้จริง (y) มีค่าเป็น 1 ตามลำดับ ฮัสแมน (Hausman, 1998) ได้ศึกษาอัตราการจำแนกผิดในตัวแปรตามสำหรับตัวแบบการถดถอยโลจิสติกมีผลต่อประสิทธิภาพในการพยากรณ์ พบว่าอัตราการจำแนกผิดในตัวแปรตามจะทำให้ค่าสัมประสิทธิ์ในการพยากรณ์เกิดความเอนเอียง (Bias) และไม่คงเส้นคงวา (Inconsistent)