บทที่ 2

ทฤษฎีและผลงานวิจัยที่เกี่ยวข้อง

ทฤษฎีที่เกี่ยวข้อง

การวิเคราะห์การถดถอยโลจีสติก (logistic regression analysis) เป็นวิธีทางสถิติที่ใช้ศึกษาความสัมพันธ์ระหว่างตัวแปร 2 กลุ่ม ได้แก่กลุ่มตัวแปรอธิบาย (explain variable) หรือตัวแปรอิสระ (independent variable) ซึ่งอาจเป็นตัวแปรเชิงคุณภาพหรือเชิงปริมาณก็ได้ และตัวแปรตาม (dependent variable) ที่เป็นตัวแปรเชิงคุณภาพ โดยมีวัตถุประสงค์เพื่อศึกษาความสัมพันธ์ระหว่างตัวแปรเหล่านี้และนำค่าตัวแปรอธิบายไปพยากรณ์ค่าความน่าจะเป็นที่จะเกิดความสำเร็จและไม่สำเร็จของตัวแปรตาม ประเภทของการวิเคราะห์การถดถอยโลจีสติกสามารถแบ่งตามค่าที่เป็นไปได้ของตัวแปรตาม กรณีที่ตัวแปรตามมีค่าเพียง 2 ค่าจะเรียกว่าการวิเคราะห์การถดถอยโลจีสติกทวิภาค (binary logistic regression analysis) แต่ถ้าตัวแปรตามมีค่ามากกว่า 2 ค่า จะเรียกว่าการวิเคราะห์การถดถอยโลจีสติกอเนกนาม (multinomial logistic regression analysis) ในงายวิจัยนี้คำว่าการถดถอยโลจีสติกจะหมายถึงการถดถอยโลจีสติกทวิภาค

ตัวแบบการถดถอยโลจีสติก (logistic regression model)

ตัวแบบการถดถอยโลจีสติก คือตัวแบบที่แสดงความสัมพันธ์ระหว่างตัวแปร 2 กลุ่ม ได้แก่กลุ่มตัวแปรอธิบายและตัวแปรตามแบบทวิภาคที่มีค่าเป็นไปได้ 2 ค่า โดยพิจารณาในรูปของ “ความสำเร็จ” และ “ความไม่สำเร็จ” โดยที่ $Y=1$ เมื่อพบความสำเร็จ และ $Y=0$ เมื่อพบความไม่สำเร็จ (Agresti, 1996) เนื่องจากการวิเคราะห์การถดถอยเชิงเส้นมีข้อสมมุติว่า $E(e)=0$ นั่นคือ $y=E(Y|X=x)+e$ และ $E(Y|X=x)=b_0+\sum_{j=1}^{p} b_j x_{ij}$ ซึ่งเป็นค่าเฉลี่ยของ Y มีค่าอยู่ในช่วง $(-\infty,\infty)$ แต่การวิเคราะห์การถดถอยโลจีสติกเป็นการศึกษาความสัมพันธ์ระหว่างค่าความน่าจะเป็นที่จะเกิดความสำเร็จ $P(Y=1|X=x)=p(X)$ กับตัวแปรอธิบาย โดยที่ $p(X)$ ก็คือค่าความน่าจะเป็นแบบมีเงื่อนไขที่ $Y=1$ เมื่อ Y_i

มีการแจกแจงแบร์นูลลีที่มีโอกาสเกิดความสำเร็จเท่ากับ $p(X_i)$ พบว่าค่าของ $p(X_i)$ จะอยู่ในช่วง $(0,1)$ ซึ่งแตกต่างจากตัวแบบการถดถอยเชิงเส้นที่จะมีค่าเฉลี่ยอยู่ในช่วง $(-\infty,\infty)$ สำหรับการวิเคราะห์การถดถอยโลจีสติกนั้น จะสนใจการพยากรณ์ค่าที่ถูกหรือผิดของตัวแปรตามมากกว่าค่าพยากรณ์ที่เป็นตัวเลข ดังนั้นถ้าใช้ตัวแบบการถดถอยเชิงเส้นจะทำให้ค่าพยากรณ์ที่ได้มีค่าไม่เหมาะสม นั่นคือมีค่าน้อยกว่า 0 หรือมากกว่า 1 ดังนั้นจึงนำตัวแบบการถดถอยโลจีสติกมาใช้แทนตัวแบบการถดถอยเชิงเส้น ตัวแบบการถดถอยโลจีสติกมีรูปแบบดังนี้

$$P\left(Y_i=1\middle|X=x_i\right)=p\left(X_i\right)=\frac{\exp\left(b_0+\sum_{j=1}^{p}b_jx_{ij}\right)}{1+\exp\left(b_0+\sum_{j=1}^{p}b_jx_{ij}\right)}$$

เมื่อ

$$P\left(Y=0\middle|X=x_i\right)=1-p\left(X_i\right)=\frac{1}{1+\exp\left(b_0+\sum_{j=1}^{p}b_jx_{ij}\right)}$$

และ

$$odds\equiv\frac{p\left(X_i\right)}{1-p\left(X_i\right)}=\exp\left(b_0+\sum_{j=1}^{p}b_jx_{ij}\right)$$

หรือ

$$logit\left(p_i\right)=\log\left[\frac{p\left(X_i\right)}{1-p\left(X_i\right)}\right]=b_0+\sum_{j=1}^{p}b_jx_{ij}$$

เมื่อ	$p\left(X_i\right)=P\left(Y_i=1\middle\|X=x_i\right)$	แทนความน่าจะเป็นแบบมีเงื่อนไขที่ $Y_i=1$
	b_0,b_j	คือพารามิเตอร์ของตัวแบบเมื่อ $j=1,\ldots,p$
	Y_i	คือตัวแปรตามแบบทวิภาค $i=1,\ldots,n$
	x_{ij}	คือค่าของตัวแปรอธิบาย

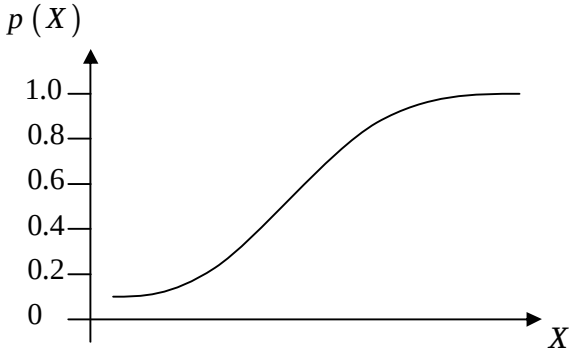
ในการวิเคราะห์การถดถอยโลจีสติกจะสามารถเขียนเทอมของค่าคลาดเคลื่อนสุ่ม (e_i) ได้ในรูปของ $y=p\left(X\right)+e$ ซึ่งถ้า $y=0$ แล้ว $e=-p\left(X\right)$ และถ้า $y=1$ แล้ว $e=1-p\left(X\right)$ ดังนั้นค่าคลาดเคลื่อนสุ่ม (e) ในการถดถอยโลจีสติกจึงมีการแจกแจงทวินาม (binomial distribution) ที่มีค่าเฉลี่ยเท่ากับ 0 และความแปรปรวนเท่ากับ $p\left(X\right)\left[1-p\left(X\right)\right]$

ตัวแบบการถดถอยโลจีสติกกรณีที่ตัวแปรตามเป็นแบบทวิภาค (binary responses variable) จะมีลักษณะที่สำคัญ 5 ประการคือ

1. ค่าเฉลี่ยแบบมีเงื่อนไขของเส้นการถดถอยโลจีสติก $p(X)=P(Y=1|x)$ ต้องมีค่าอยู่ระหว่าง 0 และ 1

2. ค่าคลาดเคลื่อนสุ่มมีการแจกแจงทวินาม (binomial distribution) ซึ่งถ้าตัวอย่างมีขนาดใหญ่ การแจกแจงดังกล่าวจะใกล้เคียงกับการแจกแจงปรกติ (normal distribution)

3. หลักเกณฑ์ที่เป็นแนวทางของการวิเคราะห์การถดถอยโลจีสติก สามารถใช้หลักเกณฑ์ของการวิเคราะห์การถดถอยเชิงเส้นทั้งแบบเชิงเดียวและแบบพหุคูณได้ เนื่องจากฟังก์ชันโลจีสติกมีลักษณะเส้นโค้งเป็นรูปตัว S หรือที่เรียกว่าโค้งซิกมอยด์ (sigmoid curve) โดยที่ความสัมพันธ์มีลักษณะเหมือนการถดถอยเชิงเส้นและเป็นเส้นตรงเมื่อ $p(X)$ อยู่ระหว่าง 0.2 ถึง 0.8 ดังนั้นในการคำนวณใดๆ เกี่ยวกับตัวแบบการถดถอยโลจีสติก สามารถใช้คุณสมบัติของการถดถอยเชิงเส้นได้เช่น การวิเคราะห์ส่วนตกค้าง (residuals analysis)



ภาพที่ 2.1 เส้นโค้งของฟังก์ชันโลจีสติก

4. ตัวแปรตาม Y_i มีการแจกแจงแบร์นูลลีที่มี $P(Y_i=1|X=x_i)=p(X_i)$ โดยที่ $0\le p(X_i)\le 1$ แต่ $b_0+\sum_{j=1}^{p} b_j x_{ij}$ ไม่จำเป็นต้องอยู่ในช่วง (0,1) ดังนั้นจึงต้องหาฟังก์ชันเชื่อมโยง (link function) มาสร้างตัวแบบที่แทนความสัมพันธ์เชิงเส้นระหว่าง $p(X_i)$ กับตัวแปร X_i ในการถดถอยโลจีสติก โดยส่วนใหญ่จะทำการแปลงให้อยู่ในรูปโลจิท (logit) ของ $p(X_i)$ คือ

$$logit(p_i)=\log\left[\frac{p(X_i)}{1-p(X_i)}\right]$$

ซึ่ง $logit(p_i)$ ไม่จำเป็นต้องอยู่ในช่วง (0,1) และมีความสัมพันธ์เชิงเส้นในรูปของ

$$\log\left[\frac{p(X_i)}{1-p(X_i)}\right] = b_0 + \sum_{j=1}^{p} b_j x_{ij}$$

5. กรณีที่ตัวแบบการถดถอยมีตัวแปรอธิบายเพียง 1 ตัว

$\log\left[\frac{p(X_i)}{1-p(X_i)}\right] = b_0 + b_1 x_{i1}$ มีประโยชน์สำหรับการตีความหมายในเทอมของโอกาสความเป็นไปได้ (odds) โดย odds จะเพิ่มขึ้นเป็น $\exp(b_1)$ เท่า ณ ทุกๆ ค่าของ X_1 ที่เพิ่มขึ้น 1 หน่วย และ odds ของ $Y=1$ คือ $\frac{p(X_i)}{1-p(X_i)} = b_0 + b_1 x_{i1}$ เป็นฟังก์ชันเชื่อมโยงสำหรับตัวแบบการถดถอยโลจีสติก ให้มีรูปแบบเป็นตัวแบบเชิงเส้นวางนัยทั่วไป (generalized linear model, GLM)

นอกจากฟังก์ชันโลจิทแล้ว สำหรับตัวแบบเชิงเส้นวางนัยทั่วไป (GLM) ยังมีฟังก์ชันเชื่อมโยง (link function) ของตัวแบบการถดถอยอื่นอีก แต่ฟังก์ชันเชื่อมโยงที่นำมาปรับใช้กับตัวแบบโลจีสติกมี 2 แบบ โดยแบบแรกได้แก่ฟังก์ชันเชื่อมโยงโพรบิท (probit) ที่มีพื้นฐานมาจากฟังก์ชันความน่าจะเป็นสะสมของการแจกแจงปรกติมาตรฐาน (standard normal cumulative distribution function) ที่มีรูปแบบดังนี้คือ $F(z) = P(Z \le z)$ หรือ

$$\Phi(z) = \frac{1}{\sqrt{2p}} \int_{-\infty}^{z} \exp\left(-\frac{1}{2} z^2\right) dz$$

และฟังก์ชันโพรบิทสำหรับตัวแบบการถดถอยโลจีสติก แสดงได้ในรูป

$$p(X_i) = \Phi\left(b_0 + \sum_j b_j x_{ij}\right) \quad \text{และ} \quad \Phi^{-1}\left[p(X_i)\right] = b_0 + \sum_j b_j x_{ij}$$

หรือ

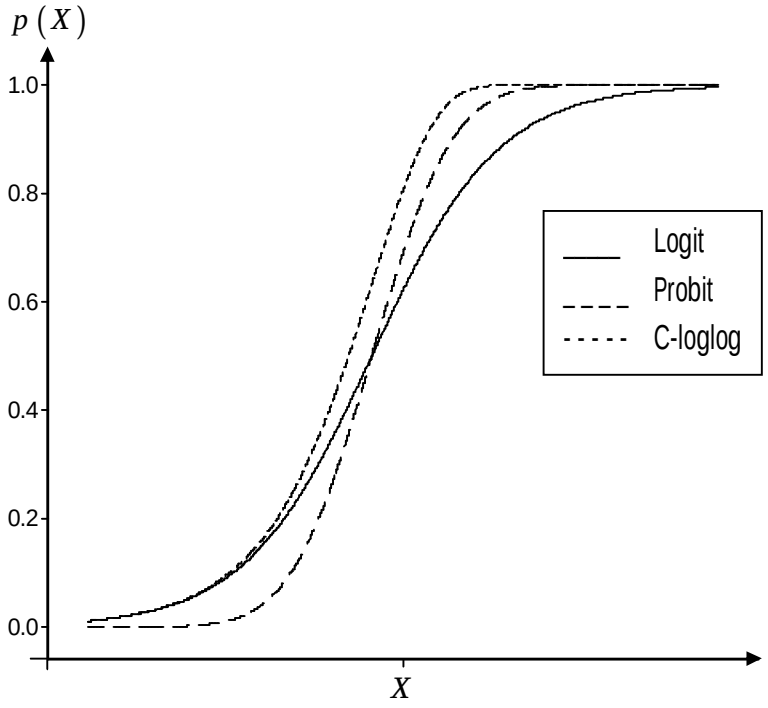
$$probit(p_i) = b_0 + \sum_j b_j x_{ij}$$

ฟังก์ชันโพรบิทนี้จะมีคุณสมบัติสมมาตร (symmetric) ที่ $p(X_i) = 0.5$ และมีลักษณะเส้นโค้งคล้ายกับฟังก์ชันโลจิท

ส่วนฟังก์ชันเชื่อมโยงแบบที่สองคือฟังก์ชันคอมพลีเมนทารีล็อก-ล็อก (complementary log-log) ซึ่งเป็นฟังก์ชันเชื่อมโยงที่ปรับจากฟังก์ชันโลจิทเมื่อค่าของ $p(X_i)$ เพิ่มจาก 0 ค่อนข้างช้าแต่มีค่าเข้าใกล้ 1 ค่อนข้างเร็ว (วีรานันท์, 2541) มีรูปแบบสำหรับตัวแบบการถดถอยโลจีสติกคือ

$$\log\left[-\log(1-p_i)\right] = b_0 + \sum_j b_j x_{ij}$$

ฟังก์ชันคอมพลีเมนทารีล็อก-ล็อกนี้จะไม่สมมาตรที่ $p(X_i) = 0.5$ ซึ่งไม่เหมือนกับฟังก์ชันโลจิทและฟังก์ชันโพรบิท



ภาพที่ 2.2 เส้นโค้งของฟังก์ชันโลจิท ฟังก์ชันโพรบิท และฟังก์ชันคอมพลีเมนทารีล็อก-ล็อก

การประมาณค่าพารามิเตอร์ของสัมประสิทธิ์การถดถอย

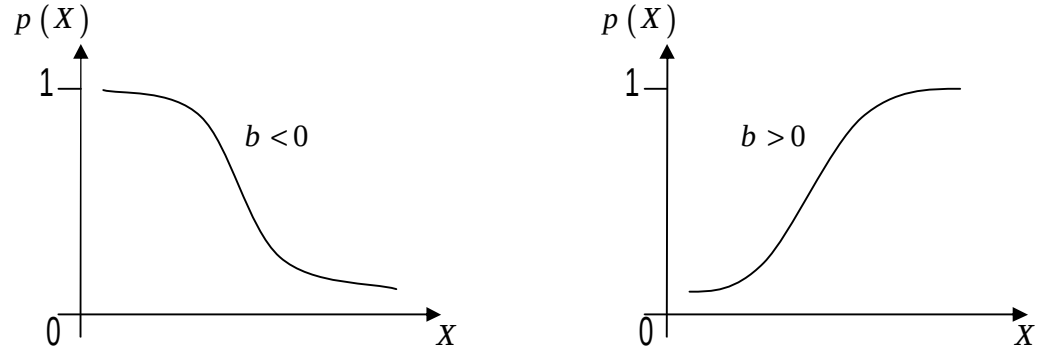
การประมาณค่าพารามิเตอร์ของสัมประสิทธิ์การถดถอยในตัวแบบเชิงเส้นทั่วไป โดยส่วนใหญ่นิยมใช้วิธีกำลังสองน้อยสุดแบบสามัญ (ordinary least square) และวิธีความควรจะเป็นสูงสุด (maximum likelihood) ซึ่งการประมาณค่าด้วยวิธีความควรจะเป็นสูงสุด เป็นวิธีที่ใช้ในการประมาณค่าพารามิเตอร์ของตัวแบบในกลุ่มเลขชี้กำลัง (exponential family models) แต่ในหลายสถานการณ์พบว่าไม่สามารถใช้วิธีการประมาณค่าดังกล่าวได้โดยตรง เนื่องจากสมการมีรูปแบบไม่เป็นเชิงเส้น (non linear) ในเทอมของพารามิเตอร์ การแก้สมการจึงต้องมีการคำนวณด้วยวิธีทำซ้ำเชิงตัวเลข (numerical iteration) และเมื่อนำมาใช้ร่วมกับวิธีความควรจะเป็นสูงสุดแล้ว ทำให้ได้วิธีที่มีประสิทธิภาพมากขึ้น วิธีอื่นๆ ที่ใช้ควบคู่กับวิธีความควรจะเป็นสูงสุดแบบทำซ้ำมีหลายวิธีแต่ในที่นี้จะกล่าวถึงเฉพาะวิธีของฟิชเชอร์สกอริง (Fisher's scoring method) ซึ่งเป็นวิธีที่ใช้สำหรับการประมาณค่าพารามิเตอร์ของตัวแบบการถดถอยในคำสั่ง PROC LOGISTIC ของโปรแกรม $SAS^{®}$

วิธีของฟิชเชอร์สกอริงได้จากการนำวิธีของนิวตัน ราฟสัน (maximum likelihood with Newton Raphson) มาปรับใช้ (Nelder and Wedderburn, 1972) โดยมีหลักการคำนวณดังนี้ ให้ตัวแปรตาม Y_i เป็นตัวแปรทวิภาคจำนวน n ค่าสังเกต และ $x_i = \left(x_{i0}, x_{i1}, ..., x_{ip}\right)$ แทนรูปแบบของค่าตัวแปรอธิบาย p ตัว ในกรณีที่ตัวแปรอธิบายทั้งหมดเป็นตัวแปรเชิงคุณภาพรูปแบบของ $x_i = \left(x_{i0}, x_{i1}, ..., x_{ip}\right)$ จะมีจำนวนน้อยกว่าจำนวนค่าสังเกต $\left(I < n\right)$ แต่ถ้าตัวแปรอธิบายเป็นตัวแปรเชิงปริมาณหลายๆ ตัว รูปแบบของ $x_i = \left(x_{i0}, x_{i1}, ..., x_{ip}\right)$ อาจมีจำนวนเท่ากับจำนวนค่าสังเกต $\left(I = n\right)$ (Agresti, 1990) และเมื่อ $x_{i0} = 1$; $i = 1, ..., n$ ตัวแบบการถดถอยโลจีสติกมีรูปแบบดังนี้

$$p\left(X_i\right) = \frac{\exp\left(\sum_{j=0}^{p} b_j x_{ij}\right)}{1 + \exp\left(\sum_{j=0}^{p} b_j x_{ij}\right)}$$

โดยที่ $p\left(X_i\right) = P\left(Y_i = 1 \middle| X_i = x_i\right)$ และ b_j คือพารามิเตอร์ของตัวแบบเมื่อ $j = 0, 1, ..., p$ b_j มีความสัมพันธ์กับ $p\left(X\right)$ คือถ้าค่า $b < 0$ ในขณะที่ x มีค่าเพิ่มมากขึ้น ค่า $p\left(X\right)$

จะมีค่าเข้าใกล้ 0 และถ้าค่า $b>0$ ในขณะที่ x มีค่าเพิ่มมากขึ้น ค่า $p(X)$ จะมีค่าเข้าใกล้ 1 และถ้าค่า b มีค่าเข้าใกล้ 0 มากๆ โค้งของ $p(X)$ จะเป็นเส้นตรงตามแนวแกน X



ภาพที่ 2.3 ความสัมพันธ์ของ $p(X)=P(Y=1|x)$ และ x เมื่อพิจารณาจาก b

สำหรับในแต่ละรูปแบบหรือชุดของ $x_i=(x_{i0}, x_{i1}, ..., x_{ip})$ ค่าตัวแปรตามที่ $Y_i=1$ อาจจะมีค่ามากกว่า 1 ค่าสังเกต ทำให้มีการนับจำนวนค่าสังเกตที่ $Y_i=1$ ในลักษณะใหม่ Y_i ในกรณีหลังนี้จึงเป็นตัวแปรสุ่มที่มีการแจกแจงทวินาม $bin(n_i, p(x_i))$, $i=1, 2, ..., I$ เมื่อ $I=n$ ด้วยค่าเฉลี่ย $E(Y_i)=n_i p(x_i)$ เมื่อ $n_1+n_2+...+n_I=n$ โดยที่ฟังก์ชันความน่าจะเป็นร่วม (joint probability mass function) ของ $(Y_1, ..., Y_I)$ คือ

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

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

แต่ $$\log\left(\frac{p(X_i)}{1-p(X_i)}\right)=\sum_{j=0}^{p} b_j x_{ij}$$ เมื่อ $j=0, 1, ..., p$

และ $$\left[1-p(X_i)\right]=\left[1+\exp\left(\sum_{j=0}^{p} b_j x_{ij}\right)\right]^{-1}$$

ดังนั้นฟังก์ชันควรจะเป็น (likelihood function) มีค่าเท่ากับ

$$\left\{\prod_{i=1}^{I}\left[1+\exp\left(\sum_{j=0}^{p} b_j x_{ij}\right)\right]^{-n_i}\right\}\exp\left[\sum_{i=1}^{I} y_i\left(\sum_{j=0}^{p} b_j x_{ij}\right)\right]$$

และฟังก์ชันล็อกควรจะเป็นเขียนแทนด้วย $L(b)$ ที่มีค่าเท่ากับ

$$L(b)=\left[\sum_j\left(\sum_i y_i x_{ij}\right)b_j\right]-\left\{\sum_i n_i \log\left[1+\exp\left(\sum_j b_j x_{ij}\right)\right]\right\}$$

จะเห็นว่าฟังก์ชัน $L(b)$ ขึ้นอยู่กับจำนวนนับแบบทวินาม (binomial count) โดยสังเกตจากตัวสถิติพอเพียง (sufficient statistics) นั่นคือ $\sum_i y_i x_{ij}; j = 0, 1, ..., p$

และ

$$\frac{\partial L(b)}{\partial b_a}=\sum_i y_i x_{ia}-\sum_i n_i x_{ia}\left[\frac{\exp\left(\sum_j b_j x_{ij}\right)}{1+\exp\left(\sum_j b_j x_{ij}\right)}\right]$$

นอกจากนี้

$$\frac{\partial^2 L(b)}{\partial b_a \partial b_b}=-\sum_i \frac{x_{ia}x_{ib}n_i \exp\left(\sum_j b_j x_{ij}\right)}{\left[1+\exp\left(\sum_j b_j x_{ij}\right)\right]^2}=-\sum_i x_{ia}x_{ib}n_i p_i\left(1-p_i\right)$$

และสมการปรกติจะอยู่ในรูป

$$\sum_i y_i x_{ia}-\sum_i n_i \hat{p}_i x_{ia}=0\ ; \qquad a=0, 1, ..., p \tag{2.1}$$

โดยที่

$$\hat{p}_i=\exp\left(\sum_j \hat{b}_j x_{ij}\right)\Big/\left[1+\exp\left(\sum_j \hat{b}_j x_{ij}\right)\right]$$

เป็นตัวประมาณความควรจะเป็นสูงสุด (MLE) ของ $p(X_i)$ ให้ X แทนเมทริกซ์ขนาด $I\times(p+1)$ ของค่าสังเกต x_{ij} และให้ $\hat{m}_i = n_i\hat{p}_i$ แล้วสมการ (2.1) จะสามารถเขียนได้ในรูป

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

สมการ (2.2) นี้จะคล้ายกับสมการปรกติจากตัวแบบการถดถอยที่ใช้วิธีกำลังสองน้อยสุด โดยที่

$$X'y = X'\hat{y} \qquad \text{เมื่อ} \qquad \hat{y} = X\hat{b} \qquad \text{และ} \qquad \hat{b} = (X'X)^{-1}X'y$$

การแก้สมการ (2.1) จะใช้วิธีของฟิชเชอร์สกอริง (Fisher's scoring method) โดยมีวิธีการคำนวณคือเริ่มต้นจะต้องมีการเดาค่าประมาณ $b^{(0)}$ ที่ทำให้ฟังก์ชันล็อกควรจะเป็น $L(b)$ มีค่าสูงสุด และนำไปใช้ในการทำซ้ำครั้งที่ 1 โดยมีหลักเกณฑ์ดังนี้
กำหนดให้

$$U = \left(\frac{\partial L(b)}{\partial b_1}, \frac{\partial L(b)}{\partial b_2}, ..., \frac{\partial L(b)}{\partial b_p}\right)$$

และ

$$H = \frac{\partial^2 L(b)}{\partial b_i \partial b_j} \qquad \text{ที่มีสมาชิกเป็น} \qquad h_{ij} = \left(\frac{\partial^2 L(b)}{\partial b_i \partial b_j}\right)$$

ซึ่ง H เป็นเมทริกซ์เอกฐาน (singular matrix) ของอนุพันธ์ที่สอง (second derivatives) ของ $L(b)$ และ $M = E(H)$ ให้ $L(b)^{(s)}$ และ $M^{(s)}$ แทนเทอมของ $L(b)$ และ M ของการประมาณค่า b ในครั้งที่ s ตามลำดับ เมื่อ $s = 0, 1, 2, ...$ และในการทำซ้ำของขั้นตอนที่ s พิจารณาให้เทอม $L(b)$ ใกล้เทอม $L(b)^{(s)}$ ณ $b^{(s)}$ ในลักษณะของการกระจายอนุกรมเทย์เลอร์ (Taylor's series expansion) กำลังที่สอง (2^{nd} order) รอบๆ ค่า $b = b^{(s)}$ ซึ่งมีรูปแบบดังนี้

$$L(b) = L\left[b^{(s)} + \left(b - b^{(s)}\right)\right]$$

$$L(b) = L\left(b^{(s)}\right) + \left[U^{(s)'}\left(b - b^{(s)}\right)\right] + \left[\frac{1}{2}\left(b - b^{(s)}\right)' M^{(s)}\left(b - b^{(s)}\right)\right]$$

แก้สมการ (2.1) จาก

$$\frac{\partial L(b)^{(s)}}{\partial b} = U^{(s)} + M^{(s)}\left(b - b^{(s)}\right) = 0$$

$$\hat{b} = b^{(s)} + \left[M^{(s)}\right]^{-1} U^{(s)} \qquad (2.3)$$

นำค่า $b^{(0)}$ ไปแทนที่ $b^{(s)}$ ใน (2.3) และคำนวณค่า $\hat{b}$ ใหม่เพื่อใช้สำหรับการทำซ้ำครั้งต่อไป เพราะฉะนั้นการทำซ้ำครั้งต่อไปจะมีลักษณะเป็น

$$b^{(s+1)} = b^{(s)} + \left(M^{(s)}\right)^{-1} U^{(s)}$$

การทำซ้ำจะทำในลักษณะแบบนี้ไปเรื่อยๆ จนกว่าจะได้ค่า $b^{(s)}$ ที่ลู่เข้าสู่ค่าประมาณความควรจะเป็นสูงสุด (MLE) นั่นคือการทำให้ผลต่างระหว่าง $b^{(s+1)}$ และ $b^{(s)}$ มีค่าน้อยมากๆ การทำซ้ำจึงเสร็จสิ้นและตัวประมาณความควรจะเป็นสูงสุดที่ได้คือ $b^{(s+1)}$

<u>การประเมินและการตรวจสอบตัวแบบการถดถอย</u>

การวิเคราะห์การถดถอยเชิงเส้นนั้นเป็นวิธีการที่ใช้ในการศึกษาความสัมพันธ์ระหว่างตัวแปรตามและตัวแปรอธิบาย เมื่อตัวแปรตามเป็นตัวแปรเชิงปริมาณและสนใจค่าพยากรณ์ของตัวแปรตามที่เป็นเชิงตัวเลข สำหรับการประเมินตัวแบบของการวิเคราะห์การถดถอยเชิงเส้น จะอยู่ภายใต้ผลบวกกำลังสอง 2 ค่าคือ SST และ SSE โดยที่ SST คำนวณจาก $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ เมื่อ $\bar{y}$ คือค่าพยากรณ์ของตัวแปรตามกรณีที่ไม่มีตัวแปรอธิบายอยู่ในตัวแบบการถดถอย 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 $\hat{y}_i = \hat{b}_0 + \hat{b}_j x_{ij}$ โดยค่า SSE จะวัดความเบี่ยงเบนในการพยากรณ์จากความแปรผันของค่าสังเกต y_i รอบเส้นถดถอย สำหรับการประมาณค่าพารามิเตอร์ของสัมประสิทธิ์การถดถอย $b_j, j = 1,..., p$ จะใช้วิธีกำลังสองน้อยสุดแบบสามัญ (ordinary least square, OLS) และสามารถหาค่าผลบวกกำลังสองของตัวแบบการ

ถดถอย (regression sum of squares, SSR) จากผลต่างระหว่าง SST และ SSE คือ $SSR = SST - SSE$

จากตัวแบบที่แสดงความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอธิบาย ก่อนการนำไปใช้ต้องทำการตรวจสอบตัวแบบที่ได้ โดยจะใช้การทดสอบด้วยค่าสถิติ F ซึ่งมีสมมุติฐานในการทดสอบที่สมมูลกัน 2 สมมุติฐาน คือ $H_0 : R^2 = 0$ และ $H_0 : b_1 = ... = b_p = 0$ ของการถดถอยเชิงเส้น ที่ได้จากการประมาณพารามิเตอร์ด้วยวิธีกำลังสองน้อยสุดแบบสามัญ สามารถคำนวณได้ดังนี้

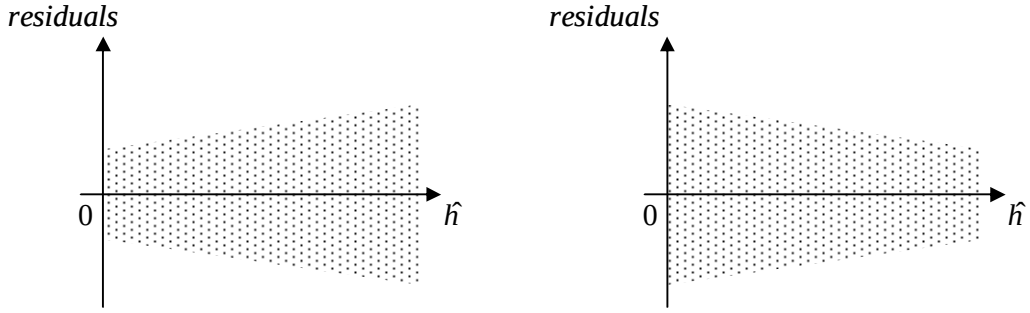
$$F = \left(\frac{SSR}{p} \right) \Big/ \left(\frac{SSE}{n-p-1} \right) = \left(\frac{n-p-1}{p} \right) \frac{SSR}{SSE}$$

เมื่อ n คือขนาดตัวอย่างและ p คือจำนวนตัวแปรอธิบาย ที่ระดับนัยสำคัญ a 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 $Y = b_0 + b_1 X_1 + ... + b_p X_p + e$ มีนัยสำคัญทางสถิติ

หรืออาจทำการตรวจสอบตัวแบบจากส่วนตกค้าง (residuals) ส่วนตกค้างหมายถึงผลต่างของ h และ $\hat{h}$ เมื่อ $\hat{h}$ คือตัวประมาณของ h ที่แสดงได้ในรูป $h = \underline{X} b$ มีขนาดเป็น $h = (h_1, ..., h_N)'$ โดยที่ $h_i = \sum_j b_j x_{ij}$; $i = 1, ..., N$, $j = 0, 1, ..., p$ และ $\underline{X}$ คือเมทริกซ์ของตัวแปรอธิบายที่มีค่าสังเกตเท่ากับ N การตรวจสอบเบื้องต้นพิจารณาว่ามีค่ามาตรฐานของส่วนตกค้าง (standardized residuals) ใดบ้างที่มากกว่า $|\pm 2 S.D|$ ถ้ามีอาจทำให้ตัวแบบการถดถอยที่ได้ไม่เหมาะสม ต้องทำการปรับตัวแบบใหม่ หรือวิเคราะห์ข้อมูลใหม่โดยอาจทดสอบสมมุติฐานเกี่ยวกับค่าผิดปรกติ เพื่อสร้างตัวแบบใหม่ที่ไม่รวมค่าผิดปรกติ นอกจากนี้อาจทำการตรวจสอบส่วนตกค้างของตัวแบบได้ดังนี้

ทำการลงจุดตกค้างเทียบกับ $\hat{h}$ หรือตัวแปรอธิบาย X เพื่อดูการเปลี่ยนแปลงของส่วนตกค้างเมื่อ $\hat{h}$ เปลี่ยนแปลงไปโดยที่ส่วนตกค้างควรมีค่าเฉลี่ยเป็นศูนย์ และความแปรปรวนคงที่สม่ำเสมอ ถ้าไม่พบรูปแบบผิดปรกติใดๆ แสดงว่าตัวแบบเหมาะสมในระดับหนึ่ง

แต่ถ้าพบรูปแบบผิดปรกติดังภาพที่ 2.4 จะเห็นได้ว่าความแปรปรวนของส่วนตกค้างจะมีค่าเพิ่มขึ้นเมื่อ $\hat{h}$ เพิ่มขึ้น และความแปรปรวนของส่วนตกค้างมีค่าลดลงเมื่อ $\hat{h}$ เพิ่มขึ้น แสดงว่าตัวแบบการถดถอยที่ได้นั้นยังไม่เหมาะสมที่จะใช้ในการพยากรณ์ หรือแสดงความสัมพันธ์ระหว่างตัวแปรอธิบายและตัวแปรตาม จะต้องทำการหาตัวแบบการถดถอยใหม่



ภาพที่ 2.4 การตรวจสอบความแปรปรวนของส่วนตกค้างจากความสัมพันธ์ของ $\hat{h}$

การตรวจสอบความเป็นอิสระของส่วนตกค้าง นอกจากจะพิจารณาจากกราฟของส่วนตกค้างเทียบกับค่าต่างๆ แล้ว ถ้าส่วนตกค้างขึ้นลงสลับกันเร็วรอบจุด 0 เมื่อค่า X เพิ่มขึ้น แสดงว่าค่าคลาดเคลื่อนสุ่มเป็นอิสระต่อกัน แต่ถ้าส่วนตกค้างมีค่าสูงกว่า 0 ติดๆ กันหลายค่า และมีค่าต่ำกว่า 0 ติดกันหลายๆ ค่าแสดงว่าค่าคลาดเคลื่อนสุ่มนั้นมีความสัมพันธ์กัน สามารถตรวจสอบความสัมพันธ์โดยใช้การทดสอบสหสัมพันธ์ของลำดับ (serial correlation) การทดสอบความมีนัยสำคัญที่นิยมใช้กันมากที่สุด คือการทดสอบของเดอร์บินวัทสัน (Durbin-Watson test)

หรืออาจตรวจสอบจากค่าสัมประสิทธิ์การตัดสินใจ (R^2) ในการถดถอยเชิงเส้นค่า R^2 เป็นค่าที่แสดงถึงสัดส่วนของความแปรผันของตัวแปรตามที่ตัวแบบสามารถอธิบายได้ โดยที่ R^2 มีค่าอยู่ระหว่าง 0 และ 1 สามารถคำนวณได้จาก

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

และสามารถเขียนในรูปของความสัมพันธ์ระหว่างค่า F และ R^2 ได้ดังนี้

$$F = \left(\frac{R^2}{p} \right) \Big/ \left(\frac{1-R^2}{n-p-1} \right)$$

ดังที่กล่าวไว้ในตอนต้นสำหรับค่าสัมประสิทธิ์การตัดสินใจ (R^2) ว่า R^2 จะไม่ลดลงแม้ว่ามีตัวแปรอธิบายที่ไม่มีความสัมพันธ์กับตัวแปรตามเพิ่มเข้าไปในตัวแบบ ดังนั้นถ้าตัวแบบการถดถอยมีตัวแปรอธิบายมากกว่า 1 ตัว สัมประสิทธิ์การตัดสินใจที่ปรับค่า (R^2_{adj}) เหมาะสำหรับใช้ในการตรวจสอบตัวแบบการถดถอยมากกว่า R^2 โดยที่

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-p-1} \right)\left(1-R^2\right)$$

หรือ

$$R^2_{adj} = 1 - \left(\frac{n-1}{n-p-1} \right)\frac{SSE}{SST} = 1 - \frac{(n-1)MSE}{SST}$$

นั่นคือ $MSE = SSE/(n-p-1)$ จะลดลงก็ต่อเมื่อ R^2_{adj} มีค่าเพิ่มขึ้น และการเพิ่มตัวแปรอธิบายเข้าไปในตัวแบบนั้นไม่จำเป็นเสมอไปว่า R^2_{adj} จะมีค่าเพิ่มขึ้น

สำหรับการประเมินตัวแบบการถดถอยโลจีสติก จะใช้วิธีในทำนองเดียวกันกับการทดสอบสถิติ F ของการถดถอยเชิงเส้น โดยที่เทอม SSE คือความแปรผันของ Y ที่ไม่สามารถอธิบายได้ของตัวแบบการถดถอยเชิงเส้น ส่วนฟังก์ชันล็อกควรจะเป็น (log-likelihood) คือความแปรผันของ Y ที่ไม่สามารถอธิบายได้ของตัวแบบการถดถอยโลจีสติก การนำเสนอค่า log-likelihood นั้น โดยทั่วไปจะใช้ -2log-likelihood เนื่องจากมีการแจกแจงใกล้เคียงกับการแจกแจงไคกำลังสอง (chi-square) ถ้าให้ D_0 หรือ $-2\left[l\left(y,\hat{p}^0\right)\right]$ เป็นตัวสถิติ -2log-likelihood ที่ตัวแบบมีเพียงเทอมค่าคงที่ (intercept) ซึ่งเปรียบได้กับ SST ในตัวแบบการถดถอยเชิงเส้น และให้ D_M หรือ $-2\left[l\left(y,\hat{p}\right)\right]$ เป็นตัวสถิติ -2log-likelihood ของตัวแบบการถดถอยโลจีสติกที่มีตัวแปรอธิบาย p ตัวที่ต้องการทดสอบ ซึ่งค่านี้เปรียบได้กับ SSE ในการถดถอยเชิงเส้น สำหรับตัวแปรตามแบบทวิภาค ให้ Y_i มีการแจกแจงแบร์นูลลีโดยที่ความน่าจะเป็นที่ $Y=1$ คือ p_i เมื่อ $\hat{p}^0 = \bar{y}$ ค่า D_0 และ D_M คำนวณจาก

$$D_0 = -2\left[l\left(y,\hat{p}^0\right)\right] = -2\sum_{i=1}^{n}\left[y_i \log \bar{y} + (1-y_i)\log(1-\bar{y})\right]$$

$$D_M = -2\left[l\left(y,\hat{p}\right)\right] = -2\sum_{i=1}^{n}\left[y_i \log \hat{p} + (1-y_i)\log(1-\hat{p})\right]$$

สำหรับการวิเคราะห์การถดถอยโลจีสติกสามารถหาผลบวกกำลังสองของตัวแบบการถดถอย (regression sum of squares) ได้จากผลต่างระหว่าง $-2\left[l\left(y,\hat{p}^0\right)\right]$ กับ $-2\left[l\left(y,\hat{p}\right)\right]$ หรือ $D_0 - D_M$ นั่นคือ $G_M = D_0 - D_M$ ซึ่งเป็นตัวสถิติไคกำลังสอง (chi-square) ของตัวแบบการถดถอยโลจีสติก ความหมายโดยตรงของ G_M คือความแตกต่างระหว่างตัวแบบที่มีเพียงเทอมค่าคงที่และตัวแบบที่มีตัวแปรอธิบายที่ต้องการทดสอบ โดย G_M จะใช้ทดสอบสมมุติฐาน $H_0 : b_1 = ... = b_p = 0$ ของตัวแบบการถดถอยโลจีสติก ถ้า G_M มีค่า $p-value \le a$ จะปฏิเสธสมมุติฐาน H_0 แสดงว่าตัวแบบควรประกอบด้วยตัวแปรอธิบาย การทดสอบด้วย G_M จะเหมือนกับการทดสอบด้วยตัวสถิติ F ในการถดถอยเชิงเส้น

วิธีที่ใช้ในการตรวจสอบความสัมพันธ์ระหว่างตัวแปรตามกับตัวแปรอธิบายในการถดถอยโลจีสติก วิธีหนึ่งที่เห็นได้ชัดเจนและมีความถูกต้อง คือการทดสอบอัตราส่วนควรจะเป็นของตัวแบบ 2 ตัวแบบ ซึ่งตัวสถิติที่ใช้ในการทดสอบก็คือ G_M ตัวอย่างเช่นให้ G_{M1} แทนตัวสถิติไคกำลังสองที่มีตัวแปร X_p ในตัวแบบและให้ G_{M2} แทนตัวสถิติไคกำลังสองที่ไม่มีตัวแปร X_p ในตัวแบบ จะได้ $G_p = G_{M1} - G_{M2}$ และมีองศาเสรีเท่ากับ 1 นอกจากนี้ยังสามารถใช้ตัวสถิติวาลด์ (Wald) ทดสอบระดับนัยสำคัญของสัมประสิทธิ์การถดถอยแต่ละตัวได้ ตัวสถิติวาลด์คำนวณจาก $W_p = \left[\frac{b_p}{S.E\left(b_p\right)}\right]^2$ ค่าของตัวสถิติวาลด์จะมีการแจกแจงเมื่อใกล้อนันต์แบบปรกติ (asymptotic normal distribution) และมีการคำนวณทำนองเดียวกับการคำนวณตัวสถิติ t ของสัมประสิทธิ์การถดถอยในการถดถอยเชิงเส้น

การคำนวณค่าสัมประสิทธิ์การตัดสินใจ $\left(R^2\right)$ ในการถดถอยโลจีสติกมีหลายวิธี ส่วนใหญ่ใช้หลักการคำนวณ 2 วิธี วิธีแรกคือวิธีกำลังสองน้อยสุดแบบสามัญ (ordinary least squared) สัมประสิทธิ์การตัดสินใจ $\left(R_O^2\right)$ ในวิธีกำลังสองน้อยสุดแบบสามัญนี้หมายถึงสัดส่วนความแปรปรวนของตัวแปรตามที่สามารถอธิบายได้ด้วยตัวแบบการถดถอย ค่า R_O^2 สามารถ

คำนวณได้จากค่าคลาดเคลื่อนของตัวแบบที่มีตัวแปรอธิบาย ซึ่งวัดความแตกต่างระหว่างค่าของตัวแปรตามและค่าพยากรณ์นั่นคือ $SSE=\sum_{i=1}^{n}(y_i-\hat{p}_i)^2$ เมื่อ $\hat{p}_i$ คือค่าพยากรณ์ของ y_i และ $SST=\sum_{i=1}^{n}(y_i-\bar{y})^2$ แทนค่าคลาดเคลื่อนของตัวแบบที่มีเพียงเทอมค่าคงที่ (intercept) เมื่อ $\bar{y}$ คือค่าเฉลี่ยของ y_i และจะใช้เป็นค่าพยากรณ์ของ y_i วิธีที่สองคือวิธีความควรจะเป็นสูงสุด (maximum likelihood) ที่คำนวณค่าสัมประสิทธิ์การตัดสินใจ (R_l^2) โดยใช้อัตราส่วนของฟังก์ชันควรจะเป็น (-2log-likelihood) ของสองตัวแบบ เมื่อตัวแบบแรกเป็นตัวแบบที่มีเพียงเทอมค่าคงที่ และตัวแบบที่สองเป็นตัวแบบที่มีตัวแปรอธิบายที่ต้องการทดสอบ p ตัว ซึ่งมีค่าฟังก์ชันล็อกควรจะเป็นคือ $-2l(y,\hat{p}^0)$ และ $-2l(y,\hat{p})$ ตามลำดับ โฮสเมอร์และเลมีโชว์ (Hosmer and Lemeshow, 1989) ได้เปรียบเทียบว่าตัวสถิติ $-2l(y,\hat{p})$ และ $-2l(y,\hat{p}^0)$ ของวิธีความควรจะเป็นสูงสุดเปรียบเสมือน SSE และ SST จากวิธีกำลังสองน้อยสุดแบบสามัญ ตามลำดับ ในขณะที่วิธีกำลังสองน้อยสุดพยายามทำให้ SSE มีค่าต่ำสุด และวิธีความควรจะเป็นสูงสุดพยายามทำให้ $-2l(y,\hat{p})$ มีค่าน้อยสุด ดังนั้นค่า R_l^2 จึงหมายถึงการลดสัดส่วนความแปรปรวนใน $-2l(y,\hat{p})$ รูปแบบของ R_O^2 และ R_l^2 แสดงได้ดังนี้

$$R_O^2=1-\frac{\sum_{i=1}^{n}(y_i-\hat{p}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \qquad (2.4)$$

$$R_l^2=\frac{-2\left[l(y,\hat{p}^0)-l(y,\hat{p})\right]}{-2\left[l(y,\hat{p}^0)\right]}=1-\frac{l(y,\hat{p})}{l(y,\hat{p}^0)} \qquad (2.5)$$

ดังที่ได้กล่าวไปแล้วว่าค่า R^2 ในตัวแบบการถดถอยโลจีสติกบางครั้งก็ให้ขนาดของความสัมพันธ์ที่มากเกินไปทั้งที่ตัวแปรอธิบายที่อยู่ในตัวแบบนั้นไม่มีความสัมพันธ์กับตัวแปรตาม เพราะฉะนั้นสำหรับตัวแบบการถดถอยโลจีสติกที่ประกอบด้วยตัวแปรอธิบายที่มีมากกว่าหนึ่งตัวแปร สัมประสิทธิ์การตัดสินใจที่ปรับค่า เหมาะสำหรับตรวจสอบตัวแบบการถดถอยโลจีสติกมากกว่า R^2 อย่างไรก็ตามในการปรับค่าของสัมประสิทธิ์การตัดสินใจ (R^2) นั้นจะต้องคำนึงถึงคุณลักษณะที่ดีของ R^2 ซึ่งมีดังต่อไปนี้ (Kvalseth, 1985)

1. R^2 ควรเป็นตัวชี้วัดความเหมาะสมของตัวแบบและสามารถที่จะอธิบายความหมายได้อย่างสมเหตุสมผลและเข้าใจง่าย

2. R^2 ควรจะเป็นอิสระจากหน่วยในการวัดของตัวแปรในตัวแบบการถดถอย

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

4. R^2 ควรจะประยุกต์ใช้ได้กับตัวแบบชนิดต่างๆ ไม่ว่าตัวแปรอธิบายจะเป็นตัวแปรสุ่มหรือตัวแปรไม่สุ่ม และไม่คำนึงถึงคุณสมบัติทางสถิติของตัวแปรในตัวแบบ (ซึ่งรวมถึงค่าคลาดเคลื่อน e)

5. R^2 ไม่ควรขึ้นกับเทคนิคที่ใช้ในการประมาณตัวแบบการถดถอย นั้นคือค่า R^2 จะวัดความเหมาะสมและไม่เหมาะสมของตัวแบบโดยไม่คำนึงถึงวิธีที่ได้มาของตัวแบบ

6. R^2 ของตัวแบบต่างๆ ที่มาจากข้อมูลชุดเดียวกันสามารถเปรียบเทียบกันได้โดยตรง

7. ค่าสัมพัทธ์ของ R^2 น่าจะสอดคล้องกับค่าวัดอื่นๆ ซึ่งเป็นที่ยอมรับในการวัดความเหมาะสมของตัวแบบการถดถอย เช่นค่าคลาดเคลื่อนมาตรฐานของค่าประมาณตัวแปรตาม $(\hat{Y})$ และรากที่สองของค่าเฉลี่ยคลาดเคลื่อนกำลังสอง (MSE)

8. ค่า R^2 ควรให้น้ำหนักของส่วนตกค้าง (residuals) ทางบวกและทางลบเท่าๆ กัน

อย่างไรก็ตามในการตรวจสอบความเหมาะสมของตัวแบบการถดถอยโลจีสติก โดยการใช้ค่าสัมประสิทธิ์การตัดสินใจที่ปรับค่า (R^2_{adj}) สามารถลดปัญหาที่เกิดจากการใช้ R^2 ลงได้ R^2_{adj} มีประโยชน์สำหรับใช้เปรียบเทียบตัวแบบการถดถอยที่มีจำนวนตัวแปรอธิบายแตกต่างกัน การที่ค่า R^2_{adj} ลดลงเมื่อเพิ่มตัวแปรอธิบายเข้าไปในตัวแบบแสดงให้เห็นว่าตัวแปรอธิบายที่เพิ่มเข้าไปนั้นมีความสัมพันธ์กับตัวแปรตามน้อยมาก มีงานวิจัยต่างๆ ที่สนับสนุนให้ใช้ R^2_{adj} มากกว่า R^2 ถ้าตัวแบบการถดถอยประกอบด้วยตัวแปรอธิบายตั้งแต่ 2 ตัวขึ้นไป

<u>งานวิจัยที่เกี่ยวข้อง</u>

อัลดริชและเนลสัน (Aldrich and Nelson, 1984) ได้เสนอการใช้ $pseudo - R^2$ โดยใช้ฟังก์ชันควรจะเป็นสูงสุด G_M คือ

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

โดยที่ n คือขนาดตัวอย่าง และ $G_M = -2\left[l\left(y,\hat{p}^0\right) - l\left(y,\hat{p}\right)\right]$ มีการแจกแจงไคกำลังสอง เนื่องจาก G_M ได้มาจากฟังก์ชันควรจะเป็น R_C^2 จึงเป็นตัวสถิติที่ดีที่พยายามทำให้ได้ค่าสูงสุด อย่างไรก็ตามค่า R_C^2 นี้จะมีค่าไม่เท่ากับ 1 ถึงแม้ว่าตัวแบบจะสามารถพยากรณ์ข้อมูลได้อย่างสมบูรณ์ก็ตาม เนื่องจากการเพิ่มขนาดตัวอย่างในตัวส่วน

มิทเบิกค์และสเชมเปอร์ (Mittlböck and Schemper, 1996) ได้กล่าวถึงข้อเสียของ R^2 2 ข้อ ข้อแรกคือ R^2 ขึ้นอยู่กับขอบเขตและการแจกแจงของตัวแปรตาม และข้อสองคือถ้าตัวแปรตามเป็นแบบทวิภาค R^2 จะมีค่าต่ำกว่าที่ควรจะเป็น แม้ว่าตัวแบบการถดถอยจะสามารถพยากรณ์ข้อมูลได้อย่างสมบูรณ์ก็ตาม นอกจากนั้นมิทเบิกค์และสเชมเปอร์ได้ทำการเปรียบเทียบสัมประสิทธิ์การตัดสินใจของตัวแบบการถดถอยโลจีสติกทั้งที่มีการปรับค่าและไม่ปรับค่าที่คำนวณด้วยวิธีกำลังสองน้อยสุดแบบสามัญ และวิธีความควรจะเป็นสูงสุด โดยทำการศึกษาภายใต้ขนาดตัวอย่างเท่ากับ 50, 100, 200 และ 1000 ตัวแปรอธิบายมีการแจกแจงแบร์นูลลีที่มีค่า 0 และ 1 ในสัดส่วนเท่าๆ กัน และสำหรับค่าสัมประสิทธิ์การถดถอย $\hat{b}_1$ ถึง $\hat{b}_p$ กำหนดให้มีค่าเท่ากันและสมมุติให้เท่ากับ 3.525, 2.268 และ 1.662 เมื่อ $p = 1, 5$ และ 10 ตามลำดับ จากการศึกษาพบว่าสัมประสิทธิ์การตัดสินใจที่คำนวณด้วยวิธีกำลังสองน้อยสุดแบบสามัญนั้น สัมประสิทธิ์การตัดสินใจที่ปรับค่าเหมาะสำหรับใช้ในการวัดความเหมาะสมของตัวแบบมากกว่าสัมประสิทธิ์การตัดสินใจที่ไม่ปรับค่า ถ้าอัตราส่วน $p/n = 0.2$ และทำนองเดียวกันสำหรับสัมประสิทธิ์การตัดสินใจที่คำนวณด้วยวิธีความควรจะเป็นสูงสุด พบว่าสัมประสิทธิ์การตัดสินใจที่ปรับค่าให้ผลที่ดีกว่าสัมประสิทธิ์การตัดสินใจที่ไม่ปรับค่า ด้วยการใช้ความถูกต้องของ $(p+1)/2$ และยังแนะนำว่าถ้าจะใช้สัมประสิทธิ์การตัดสินใจที่ปรับค่า สำหรับการเปรียบเทียบตัวแบบการถดถอยโลจีสติกที่ใช้ขั้นตอนการคัดเลือกตัวแปรอธิบายแบบขั้น (stepwise) ควรให้ค่า p เป็นจำนวนตัวแปรอธิบายที่เข้าไปทำการคัดเลือกในตัวแบบทั้งหมดมากกว่าที่จะเป็นจำนวนตัวแปรอธิบายในตัวแบบการถดถอยที่ได้ในขั้นสุดท้าย

ชีท (Shieh, 2001) ได้ศึกษาเกี่ยวกับการคำนวณขนาดตัวอย่าง สำหรับตัวแบบการถดถอยโลจีสติก และตัวแบบการถดถอยปัวซง (Poisson regression model) ซึ่งได้ปรับปรุงสูตรการคำนวณขนาดตัวอย่างด้วยการรวมค่าของการประมาณค่าพารามิเตอร์ต่างๆ ด้วยวิธี

ความควรจะเป็นสูงสุดไว้ในสูตร ภายใต้สมมุติฐานหลัก (null hypothesis) ที่ปรับปรุงจากวิธีของไวท์ติมอร์ (Whittemore, 1981) และของซิกนอรินิ (Signorini, 1991) เพื่อให้ได้สูตรที่ชัดเจนขึ้นที่ใช้สำหรับการพิจารณาขนาดตัวอย่าง ผลการศึกษาพบว่าวิธีดังกล่าวมีความถูกต้องมากกว่าวิธีของไวท์ติมอร์ (Whittemore) และของซิกนอรินิ (Signorini) ภายใต้เงื่อนไขที่สร้างขึ้นมา โดยตัวแบบที่พิจารณาประกอบด้วย 2 ตัวแบบได้แก่ $h = b_0 + X_1 b_1$ และ $h = b_0 + X_1 b_1 + X_2 b_2$ พบว่าขนาดตัวอย่างที่คำนวณได้ สำหรับข้อมูลที่ตัวแปรอธิบายเป็นตัวแปรเชิงคุณภาพแบบอเนกนาม (multinomial variable) มีขนาดตัวอย่างตั้งแต่ 1,700 ขึ้นไป

ชเตทแลนด์ (Shtatland, 2000) ได้แนะนำตัววัด R^2 2 ตัวในคำสั่ง PROC LOGISTIC และ PROC GENMOD ของโปรแกรม $SAS^{®}$ คือ $R^2_{SAS,DEV}$ และ $R^2_{SAS,AIC}$ โดยที่ $R^2_{SAS,DEV}$ จะเปรียบเทียบตัวแบบที่มีตัวแปรอธิบายกับตัวแบบที่มีเพียงเทอมค่าคงที่ (intercept) ซึ่งเขียนได้ในรูป

$$R^2_{SAS,DEV} = 1 - \exp\left\{-2\left[l(y,\hat{p}) - l\left(y,\hat{p}^0\right)\right]/n\right\}$$

โดยที่ $l(y,\hat{p})$ และ $l\left(y,\hat{p}^0\right)$ คือฟังก์ชันล็อกควรจะเป็นสูงสุด (maximized log likelihood) สำหรับตัวแบบที่มีตัวแปรอธิบายและตัวแบบที่มีเพียงเทอมค่าคงที่ (intercept) ตามลำดับ เมื่อ n คือขนาดตัวอย่าง อย่างไรก็ตามจากนิยามของ $R^2_{SAS,DEV}$ ไม่สามารถทำให้ $R^2_{SAS,DEV}$ มีค่าสูงสุดเป็น 1 ได้ ตัวอย่างเช่นสำหรับตัวแบบที่สมบูรณ์และส่วนตกค้าง (residuals) เท่ากับ 0 แต่ปรากฏว่าค่า $R^2_{SAS,DEV} = 0.75$ เท่านั้น ทำให้เป็นข้อเสียเปรียบที่เห็นได้ชัดของตัววัด $R^2_{SAS,DEV}$ นาเจลเคอร์เค (Nagelkerke, 1991) จึงได้ทำการปรับค่า $R^2_{SAS,DEV}$ ด้วยการใช้ค่าสูงสุดของ $R^2_{SAS,DEV}$ ที่มีค่าเท่ากับ $1 - \exp\left[2l\left(y,\hat{p}^0\right)/n\right]$ ซึ่งจะใช้สัญลักษณ์ $R^2_{l,adj,SAS_{DEV}}$ ที่สามารถเขียนได้ดังนี้

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

ถึงแม้ว่า $R^2_{l,adj,SAS_{DEV}}$ สามารถมีค่าสูงสุดเท่ากับ 1 แต่ปรากฏว่าทำให้ได้ค่าที่สูงจนเกินไป ในบางครั้งมีหลายสถานการณ์ที่ค่าของ $R^2_{SAS,DEV}$ ให้ค่าที่น้อย แต่กลับพบว่าค่า $R^2_{l,adj,SAS_{DEV}}$ ให้

ค่าที่สูงมากซึ่งไม่สอดคล้องกัน จึงจำเป็นต้องหา R^2 ตัวใหม่มาเป็นตัวชี้วัดเพิ่มเติม ชเตทแลนด์ (Shtatland, 2000) จึงได้แนะนำ $R^2_{SAS,AIC}$ ที่ได้มาจากการปรับค่าโดยอาศัยหลักเกณฑ์สารสนเทศของอาไคเคะ (Akaike's information criterion, AIC) โดยมีขึ้นตอนการปรับค่าดังนี้ ขั้นแรกปรับค่า R^2 โดยใช้หลักการปรับค่าเหมือนกับ R^2 ของการถดถอยเชิงเส้นด้วยการใช้จำนวนตัวแปรอธิบายหรือขนาดตัวอย่าง กรณีที่ใช้ทั้งจำนวนตัวแปรอธิบายและขนาดตัวอย่าง การปรับค่า R^2 ทำได้ดังนี้

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

ในกรณีที่ n มีขนาดใหญ่และ p มีจำนวนน้อย การปรับค่า R^2 อาจใช้แค่จำนวนตัวแปรอธิบายเท่านั้น โดยมีสูตรการปรับค่าดังนี้

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

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

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

ดังนั้นการปรับค่า $R^2_{l,adj,SAS_{AIC}}$ ที่ใช้หลักเกณฑ์สารสนเทศของอาไคเคะ (Akaike's information criterion, AIC) นั่นก็เพื่อป้องกันความเอนเอียงที่อาจเกิดขึ้นในกรณีที่มีขนาดตัวอย่างน้อย (Hurvich and Tsai, 1995) ค่า $R^2_{l,adj,SAS_{DEV}}$ และ $R^2_{l,adj,SAS_{AIC}}$ สามารถอธิบายความสัมพันธ์ของข้อมูลได้ดีเพราะ $R^2_{l,adj,SAS_{DEV}}$ และ $R^2_{l,adj,SAS_{AIC}}$ มีหลักการคำนวณมาจากวิธีความควรจะเป็นสูงสุด ตัววัด $R^2_{l,adj,SAS_{DEV}}$ และ $R^2_{l,adj,SAS_{AIC}}$ จึงเหมาะสำหรับช่วยในการคัดเลือกตัวแบบการถดถอยโลจีสติก

ธนิษฐา คำศรี (2546) ได้ทำการศึกษาเปรียบเทียบ R^2 ที่ใช้ในการตรวจสอบความเหมาะสมของตัวแบบการถดถอยโลจีสติก โดยค่า R^2 ที่นำมาศึกษาคือ R_O^2, R_l^2, R_C^2, $R_{SAS,DEV}^2$ และ $R_{l,adj,SAS_{DEV}}^2$ และกำหนดตัวแปรตามที่มีสัดส่วนของ $Y=1$ ณ ค่า p ในระดับต่างๆ ที่มีการแจกแจงแบร์นูลลี โดยที่ $p=0.01, 0.1, 0.2, 0.3, 0.4$ และ 0.5 และกำหนดตัวแปรอธิบาย X_i, $i=1, 2, 3$ ประกอบด้วยตัวแปรเชิงคุณภาพและเชิงปริมาณ โดยที่ X_1 มีการแจกแจงปรกติ X_2 และ X_3 มีการแจกแจงแบร์นูลลี ขนาดตัวอย่างที่ทำการศึกษาคือ $200, 400, 600, 1000, 2000$ และ 3000 พบว่าตัวแบบการถดถอยโลจีสติกที่ประกอบด้วยตัวแปรอธิบายทั้งที่เป็นตัวแปรเชิงคุณภาพและเชิงปริมาณ ตัวสถิติ $R_C^2, R_{SAS,DEV}^2$ และ R_O^2 ให้ค่าสัมบูรณ์ของสัมประสิทธิ์สหสัมพันธ์กับสัดส่วนของ $Y=1$ ต่ำสุด เมื่อเปรียบเทียบกับตัวสถิติ R_l^2 และ $R_{l,adj,SAS_{DEV}}^2$ ที่มีแนวโน้มลดลงเมื่อขนาดตัวอย่างเพิ่มขึ้น จึงชี้ให้เห็นว่า $R_C^2, R_{SAS,DEV}^2$ และ R_O^2 เป็นตัวสถิติที่สามารถวัดความเหมาะสมของตัวแบบการถดถอยโลจีสติกในแง่ที่ไม่ขึ้นอยู่กับสัดส่วนของ $Y=1$ ได้ดีกว่าตัวสถิติอื่นๆ

เลียโอะและแมคกี (Liao and McGee, 2003) ได้เสนอ R_{adj}^2 ของการถดถอยโลจีสติก เพื่อใช้แก้ปัญหาของ R^2 ที่ให้ขนาดความสัมพันธ์ที่มากเกินไปเมื่อมีจำนวนตัวแปรอธิบายเพิ่มขึ้นและขนาดตัวอย่างลดลง โดยมีการปรับค่า R^2 ด้วยค่าปรับของค่าคลาดเคลื่อนจากการพยากรณ์ที่ไม่สามารถแยกได้ (inherent prediction error, IPE) ในตัวแบบการถดถอยโลจีสติก โดยมีขั้นตอนการทำดังนี้ ขั้นแรกคำนวณค่า IPE หลังจากนั้นทำการประมาณค่าความเอนเอียง (bias-corrected estimator) ของค่า IPE โดยใช้หลักการเช่นเดียวกับการประมาณค่าความแปรปรวน (s^2) ในตัวแบบการถดถอยด้วยวิธี REML (restricted or residual maximum likelihood) ดังนั้นวิธีการของเลียโอะและแมคกีค่าตัวแปรตามจะถูกแยกออกเป็น 2 แบบ คือตัวแปรตามมีการแจกแจงแบร์นูลลี $(y_i \sim Bernoulli(p_i))$ และ $y_i^{new} \sim Bernoulli(\hat{p}_i)$ โดยค่า $\hat{p}_i$ มีค่าเท่ากับ $n^{-1}\sum_{i=1}^{n}(y_i-\hat{p}_i)^2$ และ $-n^{-1}l(y,\hat{p})$ สำหรับการคำนวณค่า R_{adj}^2 ด้วยวิธีกำลังสองน้อยสุดแบบสามัญ $\left(R_{O,adj,LM}^2\right)$ และวิธีความควรจะเป็นสูงสุด $\left(R_{l,adj,LM}^2\right)$ ตามลำดับ ซึ่งสูตรการคำนวณแสดงได้ดังต่อไปนี้

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

และ

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

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

y^{new} คือค่าของตัวแปรตามที่ใช้ในการวิเคราะห์ครั้งที่ 2

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

$l(y,\hat{p})$, $l\left(y,\hat{p}^0\right)$ และ n นิยามได้เช่นเดียวกับสมการ (1.2)

ผลการศึกษากรณีที่ตัวแปรอธิบายมีการแจกแจงเอกรูปที่ต่อเนื่อง พบว่าเมื่อเพิ่มตัวแปรอธิบายที่ไม่มีความสัมพันธ์กับตัวแปรตามเข้าในตัวแบบค่า R_l^2 และ R_O^2 เพิ่มขึ้นอย่างมากและให้ค่าความสัมพันธ์ที่สูง ส่วนค่า $R^2_{O,adj,MS}$ และ $R^2_{l,adj,MS}$ จะเพิ่มขึ้นอย่างต่อเนื่องและคงที่ ค่า $R^2_{l,adj,LM}$ และ $R^2_{O,adj,LM}$ จะให้ค่าใกล้เคียงกับค่าสัมประสิทธิ์การตัดสินใจที่แท้จริง $\left(R^2_{true}\right)$ และพบว่าเมื่อทำการเพิ่มขนาดตัวอย่าง ค่า R^2_{adj} ทั้ง 4 ตัวให้ค่าที่ใกล้เคียงกับค่า R^2_{true} สำหรับค่า $R^2_{l,adj,LM}$ ให้ผลที่ดีกว่าค่า $R^2_{O,adj,MS}$ และเหมาะที่จะใช้ในกรณีที่ตัวแบบมีจำนวนตัวแปรอธิบายมากหรือมีขนาดตัวอย่างน้อย

กรรณิกา โกวเครือ (2548) ได้ศึกษาเปรียบเทียบวิธีการประมาณค่าพารามิเตอร์ของตัวแบบการถดถอยโลจีสติกเมื่อตัวแปรตามเป็นแบบทวิภาค 4 วิธีได้แก่ วิธีฟังก์ชันจำแนกประเภทของฟิชเชอร์ (Fisher's discriminant function) วิธีความควรจะเป็นสูงสุดของนิวตัน ราฟสัน (maximum likelihood with Newton Raphson) วิธีควอดาติกวันเสต็ปบูตแสตรป (quadratic one-step bootstrap) และวิธีเอคแซค โลจีสติก รีเกรสชัน (exact logistic regression) ผลการศึกษาพบว่าในกรณีที่ตัวแปรอธิบายในตัวแบบมีการแจกแจงความน่าจะเป็นที่แตกต่างกัน วิธีฟังก์ชันจำแนกประเภทของฟิชเชอร์เป็นวิธีประมาณค่าพารามิเตอร์ที่มีประสิทธิภาพสูงสุด