บทที่ 4
การวิเคราะห์การถดถอยโลจิสติกสำหรับข้อมูลทวิ

4.1 กล่าวนำ

ในบทที่ 3 ได้อธิบายถึงการประยุกต์ใช้การถดถอยเชิงเส้น กับตัวแปรตามที่มีข้อมูลเป็นข้อมูลทวิแล้ว พบว่ายังไม่สามารถแก้ปัญหาเกี่ยวกับข้อจำกัดที่ว่า $0 \leq E(Y_i)=P_i \leq 1$ ได้ ในบทนี้ จะได้กล่าวถึงการพิจารณาโมเดลสำหรับ $E(Y_i) = P_i$ ที่เหมาะสม เพื่อให้ได้โมเดลการถดถอยที่มีคุณสมบัติเป็นไปตามข้อจำกัดดังกล่าว

โมเดลการถดถอยสำหรับข้อมูลทวิ ที่จะให้ค่าจากโมเดลมีค่าอยู่ระหว่าง 0 และ 1 นั้น พบว่า ควรจะเป็นโมเดลเส้นโค้งที่มีค่าลู่เข้าสู่ 0 และ 1 โมเดลที่มีลักษณะดังกล่าวนี้ ได้แก่

1. โมเดลโลจิสติก (logistic model)

$$P_i = [1 + \exp(-U)]^{-1}$$

ซึ่งคือ ฟังก์ชั่นความน่าจะเป็นสะสม (cumulative probability density function) ของการแจกแจงแบบโลจิสติก (logistic distribution) ที่มีฟังก์ชั่นการแจกแจงความน่าจะเป็น (probability density function ;p.d.f.) คือ

$$f(U) = e^{-U}/[1 + e^{-U}]^2 \qquad ; -\infty < U < \infty$$

เมื่อ $U = \sum_{j=0}^{q-1} \beta_j X_j \qquad ; j = 0,1,2,\ldots q-1$

$X_0 = 1$

และสามารถแปลงโมเดลโลจิสติก ได้เป็น $\text{logit}(P_i) = \ln (P_i/Q_i) =$

$U = \sum_{j=0}^{q-1} \beta_j X_j$ เมื่อ $Q_i = 1 - P_i$

2. โมเดลโพรบิท(probit model)

$$P_i = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{U} \exp(-1/2z^2)\, dz \quad ; \quad -\infty < z < \infty$$

ซึ่งเป็นฟังก์ชั่นการแจกแจงความน่าจะเป็นสะสม ของการแจกแจงปกติมาตรฐาน ที่มี p.d.f. เป็น

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp(-1/2z^2)$$

การใช้โมเดลโลจิสติกและโมเดลโพรบิท เป็นโมเดลการถดถอยสำหรับข้อมูลทวิ จะให้ผลคล้ายคลึงกัน แต่ความนิยมในโมเดลโลจิสติก มีมากกว่าโมเดลโพรบิท เนื่องจากสามารถนำไปประยุกต์ใช้ในงานทดลองวิจัยได้มากกว่า โดยเฉพาะอย่างยิ่งงานทดลองวิจัยเกี่ยวกับสิ่งมีชีวิต (bioassay) ดังนั้นในที่นี้ จึงจะกล่าวถึงเฉพาะการวิเคราะห์การถดถอยสำหรับข้อมูลทวิ ที่ใช้โมเดลโลจิสติกเป็นโมเดลการถดถอย ซึ่งเรียกว่าการวิเคราะห์การถดถอยโลจิสติก (logistic regression analysis) โดยมีข้อกำหนดว่า ข้อมูลทวิของตัวแปรตามที่ใช้ต้องมาจากการแจกแจงแบบทวินามเท่านั้น

4.2 โมเดลโลจิสติกเชิงเส้น(The linear logistic model)

ถ้า P_i คือความน่าจะเป็นที่จะเกิดผลสำเร็จ ของการแจกแจงแบบทวินาม ขนาด n_i ซึ่งมีความสัมพันธ์กับตัวแปรอิสระ X_{ji} แสดงด้วยโมเดลโลจิสติก เป็น

$$P_i = [1 + e^{-U_i}]^{-1}$$

$$= e^{U_i}[1 + e^{U_i}]^{-1}$$

เมื่อ $U_i = \sum_{j=0}^{q-1} \beta_j X_{ji}$; $j = 0, 1, \ldots, q-1$

$X_{0i} = 1$

และมี Q_i คือ ความน่าจะเป็นที่จะไม่เกิดผลสำเร็จ

โดย $Q_i = 1 - P_i$

$= e^{-U_i} [1 + e^{-U_i}]^{-1}$

$= [1 + e^{U_i}]^{-1}$

นิยาม odds of success หรือ Odds คืออัตราส่วนระหว่างความน่าจะเป็นที่จะเกิดผลสำเร็จ ต่อความน่าจะเป็นที่จะไม่เกิดผลสำเร็จ ; P_i / Q_i จะได้

$$\text{Odds} = P_i/Q_i = e^{U_i}$$

และ $$\ln \text{Odds} = \text{logit}(P_i) = U_i = \sum_{j=0}^{q-1} \beta_j X_{ji} \qquad (4.2.1)$$

เรียก $\text{logit}(P_i)$ ว่าเป็น link function ซึ่งแสดงว่าการแปลงโมเดลโลจิสติกของ P ให้อยู่ในรูปของ ln Odds ได้เป็นสมการเส้นตรง และเรียกสมการ (4.2.1) ว่าเป็นโมเดลโลจิสติกเชิงเส้น สำหรับตัวแปรอิสระ หรือโมเดลโลจิท (logit model)

4.3 การประมาณค่าพารามิเตอร์

ในการประมาณค่าพารามิเตอร์ β_j สำหรับโมเดลโลจิสติกเชิงเส้น $\text{logit}(P_i) = \sum_{j=0}^{q-1} \beta_j X_{ji}$ จะใช้วิธีประมาณค่าแบบภาวะน่าจะเป็นสูงสุด ซึ่งมีฟังก์ชั่นภาวะน่าจะเป็นสูงสุด คือ

$$L(\beta) = \prod_{i=1}^{n} {}^{n_i}C_{y_i} \; P_i^{y_i} \; Q^{n_i - y_i}$$

$$\ln L(\beta) = \sum_{i=1}^{n} [\ln {}^{n_i}C_{y_i} + y_i \ln P_i + (n_i - y_i) \ln Q_i] \quad (4.3.1)$$

$$= \sum_{i=1}^{n} [\ln {}^{n_i}C_{y_i} + y_i \ln P_i/Q_i + n_i \ln Q_i]$$

$$= \sum_{i=1}^{n} \{\ln {}^{n_i}C_{y_i} + y_i U_i + n_i \ln[1 + e^{U_i}]^{-1}\}$$

$$\frac{\partial \ln L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} y_i X_{ji} - \sum_{i=1}^{n} n_i X_{ji} e^{U_i} [1 + e^{U_i}]^{-1}$$

$$= \sum_{i=1}^{n} y_i X_{ji} - \sum_{i=1}^{n} n_i X_{ji} \hat{p}_i \quad ; j = 0,1,\ldots,q-1 \quad (4.3.2)$$
$$X_{0i} = 1$$

เมื่อทำสมการ (4.3.2) ให้เท่ากับ 0 จะให้สมการที่ไม่ใช่เส้นตรงจำนวน q สมการ ที่มีค่าประมาณของ β_j ซึ่งไม่ทราบค่าติดอยู่ในสมการ การหาค่าประมาณของ β_j จากสมการ q สมการนี้ จะใช้การประมาณค่าแบบภาวะน่าจะเป็นสูงสุดที่ทำซ้ำ (iterative maximum likelihood estimation โดย Newton-Raphson procedure และ Fisher's method of scoring และวิธีกำลังสองต่ำสุดถ่วงน้ำหนักที่ทำซ้ำ (iteratively weighted least squares ซึ่งขบวนการประมาณค่าพารามิเตอร์ดังกล่าวนี้จะต้องอาศัยคอมพิวเตอร์ช่วยในการประมวลผล

4.3.1 การประมาณค่าแบบภาวะน่าจะเป็นสูงสุดที่ทำซ้ำ

1. โดยวิธี Newton-Raphson

ให้ $U(\beta)$ คือ คอลัมน์เวคเตอร์ขนาด qx1 ของสมการ (4.3.2) จำนวน q

สมการ โดยค่าในแถวที่ j คือ $\partial \ln L(\beta)/\partial \beta_j$ ซึ่งเรียกว่า coefficient score และให้ $H(\beta)$ คือแมทริกซ์ขนาด q x q ที่ได้จากการหาอนุพันธ์ครั้งที่สองของสมการ (4.3.1) เทียบกับค่า β_k โดยค่าในแถวที่ j คอลัมน์ที่ k คือ $\partial^2 \ln L(\beta)/\partial \beta_j \partial \beta_k$ สำหรับ j, k = 0,1,...,q-1 แมทริกซ์ $H(\beta)$ บางครั้งเรียกว่า Hessian matrix

ให้ $U(\beta)$ คือ คอลัมน์เวคเตอร์ขนาด q x 1 ของ coefficient scores ที่ได้จากการแทนค่าประมาณของพารามิเตอร์ β, $\hat{\beta}$ แล้ว และโดย Taylor series กระจาย $U(\hat{\beta})$ รอบค่า β^* เมื่อ β^* คือค่าที่ใกล้เคียง $\hat{\beta}$ จะได้

$$U(\hat{\beta}) \approx U(\beta^*) + H(\beta^*)\ (\hat{\beta} - \beta^*) \qquad (4.3.1.1)$$

โดยนิยามของการประมาณค่าพารามิเตอร์ β_j โดยวิธีภาวะน่าจะเป็นสูงสุด จะต้องได้ว่า

$$\left.\frac{\partial \ln L(\beta)}{\partial \beta_j}\right|_{\hat{\beta}_j} = 0$$

ซึ่งจะได้ $U(\beta) = 0$ ดังนั้น จะได้ (4.3.1.1) เป็น

$$0 \approx U(\beta^*) + H(\beta^*)\ (\hat{\beta} - \beta^*)$$

$$\hat{\beta} \approx \beta^* - [H(\beta^*)]^{-1}\ U(\beta^*) \qquad (4.3.1.2)$$

จาก (4.3.1.2) เมื่อมีการทำซ้ำ ๆ กัน จนถึงรอบที่ r+1 จะได้ค่าประมาณของ β ในรอบที่ r+1 เป็น

$$\hat{\beta}_{r+1} = \hat{\beta}_r - [H(\hat{\beta}_r)]^{-1}\ U(\hat{\beta}_r) \qquad (4.3.1.3)$$

เมื่อ r = 0, 1, ... และ $\hat{\beta}_0$ คือคอลัมน์เวคเตอร์ของค่าประมาณเริ่มต้นของพารามิเตอร์ β

2. โดย Fisher's method of scoring

ให้ $I(\beta)$ คือค่าคาดหวังของ $H(\beta)$ คูณด้วย -1 และเรียก $I(\beta)$ ว่า information matrix โดย $I(\beta)$ เป็นแมทริกซ์ขนาด q x q ซึ่งมีค่าในแถวที่ j คอลัมน์ที่ k เป็น $-E\left[\frac{\partial^2 \ln L(\beta)}{\partial \beta_j \partial \beta_k}\right]$

inverse ของ $I(\beta)$ คือ $[I(\beta)]^{-1}$ เรียกว่า asymptotic variance-covariance matrix ของค่าประมาณ โดยวิธีภาวะน่าจะเป็นสูงสุด

และโดยขบวนการทำซ้ำ ๆ กัน เช่นเดียวกับวิธี Newton-Raphson จะได้ (4.3.1.3) เป็น

$$\hat{\beta}_{r+1} = \hat{\beta}_r + [I(\hat{\beta}_r)]^{-1} U(\hat{\beta}_r) \qquad (4.3.1.4)$$

การประมาณค่าพารามิเตอร์ β โดยวิธี Newton-Raphson และ Fisher's method of scoring จะให้ค่าประมาณลู่เข้าสู่ค่าประมาณแบบภาวะน่าจะเป็นสูงสุดของ β ความคลาดเคลื่อนมาตรฐานของค่าประมาณ คือรากที่สองของค่าในแนวทะแยงของ $-[H(\beta)]^{-1}$ หรือ $[I(\beta)]^{-1}$ ซึ่งค่าประมาณและความคลาดเคลื่อนมาตรฐานของค่าประมาณโดยวิธีทั้งสองนี้จะให้ค่าเหมือนกัน เมื่อการวิเคราะห์การถดถอยนี้มีโมเดลเป็นโมเดลโลจิสติกเชิงเส้นซึ่งใช้กับข้อมูลทวิที่มีการแจกแจงแบบทวินาม และค่าประมาณที่ได้นี้จะมีการแจกแจงใกล้เคียงการแจกแจงปกติ

4.3.2 การประมาณค่าด้วยวิธีกำลังสองต่ำสุดถ่วงน้ำหนักที่ทำซ้ำ

ในการประมาณค่าพารามิเตอร์ b ด้วยวิธีกำลังสองต่ำสุดถ่วงน้ำหนัก กระทำจากโมเดลโลจิท ; $\text{logit}(P_i) = \ln P_i/Q_i = U_i = \sum_{j=0}^{q-1} \beta_j X_{ji}$ ได้

$$\hat{\beta} = (X'WX)^{-1}X'WU$$

เพื่อให้ค่า $\hat{\beta}$ ลู่เข้าสู่ค่าคงที่ จะทำการประมาณค่าด้วยวิธีกำลังสองต่ำสุดถ่วงน้ำหนักที่ทำซ้ำ ซึ่งต้องมีการปรับค่า W และ U ได้เป็น W^* และ U^* ตามลำดับ โดย McCullagh and Nelder (1983) แสดงวิธีการปรับค่า ดังนี้

$$U^*_{ir} = U_{ir} + \frac{(y_i - n_i\hat{p}_{ir})}{n_i}\left[\frac{\partial \ln P_i/Q_i}{\partial P_i}\right]_r$$

แต่ $$\frac{\partial \ln P_i/Q_i}{\partial P_i} = \frac{\partial \ln P_i}{\partial P_i} - \frac{\partial \ln (1-P_i)}{\partial P_i}$$

$$= \frac{1}{\hat{p}_i} + \frac{1}{1-\hat{p}_i} = \frac{1}{\hat{p}_i(1-\hat{p}_i)} = \frac{1}{\hat{p}_i\hat{q}_i}$$

ดังนั้น เมื่อทำซ้ำ r รอบจะได้

$$U^*_{ir} = X\hat{\beta}_r + \frac{(y_i - n_i\hat{p}_{ir})}{n_i\hat{p}_{ir}\hat{q}_{ir}}$$

และ $$W^{*-1}_{ir} = V(\hat{p}_{ir})\left[\frac{\partial \ln P_i/Q_i}{\partial P_i}\right]^2_r$$

$$= \frac{\hat{p}_{ir}\hat{q}_{ir}}{n_i}\cdot\left[\frac{1}{\hat{p}_{ir}\hat{q}_{ir}}\right]^2$$

$$= \frac{1}{n_i\hat{p}_{ir}\hat{q}_{ir}}$$

$$W^*_{ir} = n_i\hat{p}_{ir}\hat{q}_{ir}$$

ดังนั้น ค่าประมาณของ β ในรอบที่ r+1 เป็น

$$\hat{\beta}_{r+1} = (X'W^*_rX)^{-1}X'W^*_rU^*_r$$

ซึ่งจะให้ค่าประมาณเป็นเช่นเดียวกับการประมาณค่าด้วยวิธีภาวะน่าจะเป็นสูงสุดที่ทำซ้ำ โดย Fisher's method of scoring และสามารถแสดงให้เห็นได้ดังนี้

จากสมการ (4.3.1)

$$\ln L(\beta) = \sum_{i=1}^{n} \{\ln {}^{n_i}C_{y_i} + y_i \ln P_i + (n_i - y_i) \ln Q_i\}$$

$$\frac{\partial \ln L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{\partial \ln L(\beta)}{\partial P_i} \cdot \frac{\partial P_i}{\partial U_i} \cdot \frac{\partial U_i}{\partial \beta_j}\right] \qquad (4.3.2.1)$$

$$\frac{\partial \ln L(\beta)}{\partial P_i} = \frac{y_i}{\hat{p}_i} - \frac{n_i - y_i}{1 - \hat{p}_i} = \frac{y_i - n_i\hat{p}_i}{\hat{p}_i\hat{q}_i}$$

$$\frac{\partial U_i}{\partial P_i} = \frac{\partial \ln(P_i/Q_i)}{\partial P_i} = \frac{1}{\hat{p}_i\hat{q}_i}$$

$$\frac{\partial U_i}{\partial \beta_j} = \frac{\partial \left(\sum_{j=0}^{q-1} \beta_j X_{ji}\right)}{\partial \beta_j} = X_{ji}$$

ดังนั้นจะได้ (4.3.2.1) เป็น

$$\frac{\partial \ln L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{(y_i - n_i\hat{p}_i)}{\hat{p}_i\hat{q}_i} \cdot \hat{p}_i\hat{q}_i \cdot X_{ji}\right]$$

$$= \sum_{i=1}^{n} (y_i - n_i\hat{p}_i) X_{ji} \qquad (4.3.2.2)$$

ให้ $y_i^* = \frac{y_i - n_i\hat{p}_i}{n_i\hat{p}_i\hat{q}_i}$

และ $W_i^* = n_i\hat{p}_i\hat{q}_i$

จะได้ $\frac{\partial \ln L(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} y_i^* W_i^* X_{ji}$

ให้ X คือ แมทริกซ์ขนาด nxq ของตัวแปรอิสระ X จำนวน q - 1 ตัว

W^* คือ แมทริกซ์ขนาด nxn ที่มีค่าในแนวทะแยง คือ W_i^*

และ Y^* คือ แมทริกซ์ขนาด nx1 ที่มีค่าในแถวที่ i คือ y_i^*

ดังนั้นจะได้ $\frac{\partial \ln L(\beta)}{\partial \beta_j} = U(\beta) = X'W^*Y^*$

พิจารณา $-E\left[\frac{\partial^2 \ln L(\beta)}{\partial \beta_j \partial \beta_k}\right] = E\left[\frac{\partial \ln L(\beta)}{\partial \beta_j} \cdot \frac{\partial \ln L(\beta)}{\partial \beta_k}\right] = I(\beta)$

จาก (4.3.2.2)

จะได้ $-E\left[\frac{\partial^2 \ln L(\beta)}{\partial \beta_j \partial \beta_k}\right] = E\left[\sum_{i=1}^{n} (y_i - n_i\hat{p}_i)(y_{i'} - n_{i'}\hat{p}_{i'}) X_{ji}X_{ki'}\right]$

$$= \sum_{i=1}^{n} [E(y_i - n_i\hat{p}_i)(y_{i'} - n_{i'}\hat{p}_{i'})] X_{ji}X_{ki'}$$

$$= \sum_{i=1}^{n} [Cov(Y_i, Y_{i'})] X_{ji}X_{ki'}$$

$$= \sum_{i=1}^{n} [V(Y_i)] X_{ji}X_{ki} \quad ; \ y_i, y_{i'} \text{ เป็นอิสระกัน}$$

$$= \sum_{i=1}^{n} n_i\hat{p}_i\hat{q}_i X_{ji}X_{ki}$$

$$= \sum_{i=1}^{n} W_i^* X_{ji} X_{ki}$$

$$= X'W^*X = I(\beta)$$

จาก (4.3.1.4) จะได้ $\hat{\beta}_{r+1} = \hat{\beta}_r + (X'W_r^*X)^{-1} X'W_r^*Y_r^*$

$$= (X'W_r^*X)^{-1} [X'W_r^*(X\hat{\beta}_r + Y_r^*)]$$

เมื่อให้ $U_r^* = X\hat{\beta}_r + Y_r^*$

จะได้ $\hat{\beta}_{r+1} = (X'W_r^*X)^{-1} X'W_r^*U_r^*$

4.3.3 ขั้นตอนของขบวนการประมาณค่าด้วยวิธีกำลังสองต่ำสุดถ่วงน้ำหนักที่ทำซ้ำ

Collett (1991) ได้แสดงขั้นตอนของขบวนการประมาณค่าด้วยวิธีกำลังสองต่ำสุดถ่วงน้ำหนักที่ทำซ้ำ ดังนี้

1. ให้ค่าเริ่มต้นของ $\hat{p}_i$ คือ $\hat{p}_{i0} = \frac{y_i + 0.5}{n_i + 1}$

คำนวณ $W_{i0}^* = n_i \hat{p}_{i0} \hat{q}_{i0}$

และ $U_{i0}^* = \text{logit}(\hat{p}_{i0}) = \ln \hat{p}_{i0}/\hat{q}_{i0}$

2. คำนวณค่า $\hat{\beta}_1$ จาก $\hat{\beta}_1 = (X'W_0^*X)^{-1} X'W_0^*U_0^*$

3. นำค่า $\hat{\beta}_1$ จาก 2 เพื่อประมาณค่า $\hat{p}_{i1} = [1+\exp(-\sum_{j=0}^{q-1} \hat{\beta}_{j1} X_{ji})]^{-1}$

คำนวณค่า $W^*_{i1} = n_i \hat{p}_{i1} \hat{q}_{i1}$

และ $U^*_{i1} = \sum_{j=0}^{q-1} \hat{\beta}_{j1} X_{ji} + \frac{(y_i - n_i \hat{p}_{i1})}{n_i \hat{p}_{i1} \hat{q}_{i1}}$

4. คำนวณค่า $\hat{\beta}_2$ จาก $\hat{\beta}_2 = (X' W^*_1 X)^{-1} X' W^*_1 U^*_1$

5. ดำเนินการตามขั้นตอน 3-4 ซ้ำ ๆ จนได้ค่า $\hat{\beta}$ ลู่เข้าสู่ค่าคงที่ค่าหนึ่ง ซึ่งจะได้เป็น $\hat{\beta}$ ตามที่ต้องการ

4.4 การทดสอบภาวะสารูปดีของโมเดลโลจิสติกเชิงเส้น

เมื่อทำการประมาณค่าพารามิเตอร์ $\hat{\beta}$ จนได้โมเดลโลจิสติกเชิงเส้นตามที่ต้องการแล้ว ควรจะทำการทดสอบดูว่า โมเดลที่ได้นั้นมีภาวะสารูปดีกับค่าสังเกตุ (test of goodness of fit) หรือไม่ ซึ่งถ้าพบว่า ค่าที่ประมาณได้จากโมเดลมีภาวะสารูปดีกับค่าสังเกตุ ก็แสดงว่าโมเดลนั้นใช้ได้ แต่ถ้าไม่เป็นดังเช่นที่กล่าว ก็จำเป็นที่จะต้องหาโมเดลที่เหมาะสมต่อไป

4.4.1 การทดสอบแบบอัตราส่วนภาวะน่าจะเป็น

การทดสอบแบบอัตราส่วนภาวะน่าจะเป็น (likelihood ratio test) เป็นการเปรียบเทียบระหว่างฟังก์ชั่นภาวะน่าจะเป็นสูงสุด 2 ฟังก์ชั่น คือ

1. ฟังก์ชั่นภาวะน่าจะเป็นสูงสุดของโมเดล ที่ใช้ค่าพารามิเตอร์ซึ่งได้มาจากการประมาณค่าโดยวิธีภาวะน่าจะเป็นสูงสุด เรียกว่าฟังก์ชั่นภาวะน่าจะเป็นสูงสุดของโมเดลที่ประมาณได้ โดยเรียกโมเดลที่ประมาณได้ว่า current model ให้ฟังก์ชั่นภาวะน่าจะเป็นของโมเดลนี้แทนด้วย L_c

2. ฟังก์ชั่นภาวะน่าจะเป็นสูงสุดของโมเดล ที่พอดีกับค่าสังเกต หรือ full model ให้ฟังก์ชั่นภาวะน่าจะเป็นสูงสุดของโมเดลนี้ แทนด้วย L_f

และให้ $$D = -2 \ln [L_c / L_f]$$

$$= -2 [\ln L_c - \ln L_f]$$

จากสมการ (4.3.1) จะได้ฟังก์ชั่นภาวะน่าจะเป็นสูงสุดของ current model เป็น

$$\ln L_c = \sum_{i=1}^{n} [\ln {}^{n_i}C_{y_i} + y_i \ln \hat{p}_i + (n_i - y_i) \ln \hat{q}_i]$$

และฟังก์ชั่นภาวะน่าจะเป็นสูงสุดของ full model เป็น

$$\ln L_f = \sum_{i=1}^{n} [\ln {}^{n_i}C_{y_i} + y_i \ln \tilde{p}_i + (n_i - y_i) \ln \tilde{q}_i]$$

เมื่อ $\tilde{p}_i = y_i/n_i$

ดังนั้น $$D = 2 \sum_{i=1}^{n} [y_i \ln (\tilde{p}_i/\hat{p}_i) + (n_i - y_i) \ln (\tilde{q}_i/\hat{q}_i)]$$

$$= 2 \sum_{i=1}^{n} [y_i \ln \frac{y_i}{\hat{y}_i} + (n_i - y_i) \ln \frac{n_i - y_i}{n_i - \hat{y}_i}]$$

เมื่อ $\hat{Y}_i = n_i\hat{p}_i$ และ $\hat{p}_i = [1 + \exp(-\sum_{j=0}^{q-1} b_j X_{ji})]^{-1}$

เรียก D ว่า deviance ซึ่งจะเป็นตัวสถิติที่ใช้ในการทดสอบ (test statistic)

เมื่อ D มีค่ามาก แสดงว่าโมเดลที่ประมาณได้จะไม่พอดีกับโมเดลของค่าสังเกต แต่ถ้าโมเดลทั้งสองมีความใกล้เคียงกัน ก็จะได้ D มีค่าน้อย โดยทฤษฎีจะได้ว่า D มีการแจก

แจงลู่เข้าสู่การแจกแจงแบบไคสแควส์ (chi-square distribution) ด้วย d.f.= n - q ดังนั้น เมื่อพบว่า D มีค่ามากกว่าค่า χ^2 จากตาราง แสดงว่ามีความแตกต่างระหว่างโมเดลประมาณ กับโมเดลของค่าสังเกตอย่างมีนัยสำคัญทางสถิติ ซึ่งหมายความว่าโมเดลที่ประมาณได้ยังไม่พอดีกับโมเดลของค่าสังเกต จะต้องมีการพิจารณาหาโมเดลใหม่ที่เหมาะสมต่อไป

ในการทดสอบภาวะสารูปดี โดยการทดสอบแบบอัตราส่วนภาวะน่าจะเป็นสูงสุดนี้ Collett (1991) พิจารณาว่า ถ้า deviance ที่คำนวณได้ มีค่าใกล้เคียงกับ d.f.= n - q สามารถสรุปได้ว่า โมเดลที่ประมาณได้มีความพอดีกับค่าสังเกต ทั้งนี้เนื่องจากข้อสมมติที่ว่าข้อมูลทวิที่มาจากการแจกแจงแบบทวินาม จะให้ค่าความแปรปรวนของความคลาดเคลื่อน ซึ่งในที่นี้คือ deviance หารด้วย d.f.= n - q มีค่าเท่ากับ 1[1]

4.4.2 Pearson's X^2-statistic

การทดสอบภาวะสารูปดี ระหว่างโมเดลประมาณกับโมเดลของค่าสังเกต สามารถทำได้อีกวิธีหนึ่ง คือ ใช้ Pearson's X^2-statistic ซึ่งนิยามค่าไว้ ดังนี้

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i\hat{q}_i}$$

โดย X^2 จะมีการแจกแจงลู่เข้าสู่การแจกแจงแบบไคสแควร์ด้วย d.f.= n-q เช่นเดียวกับการแจกแจงของ deviance

แม้ว่าการทดสอบภาวะสารูปดี ระหว่างโมเดลประมาณกับโมเดลของค่าสังเกต จะสามารถใช้ได้ทั้งค่า deviance และ Pearson's X^2-statistic เป็นตัวทดสอบสถิติ

---

[1] McCullagh and Nelder. Generalized liner models (1983) pp.83

ก็ตาม แต่โดยทั่วไปแล้ว การใช้ deviance จะมีความเหมาะสมกว่าการใช้ Pearson's X^2-statistic เพราะสามารถนำไปประยุกต์ใช้ในการเปรียบเทียบระหว่างโมเดลประมาณต่าง ๆ ได้ด้วย ซึ่งจะได้กล่าวต่อไป

4.5 การเปรียบเทียบโมเดลโลจิสติกเชิงเส้น(Comparing linear logistic model)

การวิเคราะห์การถอดถอยโลจิสติกโดยใช้โมเดลโลจิสติกเชิงเส้น เพื่อศึกษาถึงอิทธิพลของตัวแปรอิสระ จำนวนหลาย ๆ ตัวนั้น มีปัญหาเกิดขึ้นว่า ตัวแปรอิสระตัวใดบ้างที่จะมีอิทธิพลและตัวแปรอิสระใดบ้างที่ไม่มีอิทธิพล หรือกล่าวอีกนัยหนึ่งก็คือ โมเดลที่จะถูกนำไปใช้นั้นควรจะให้มีตัวแปรอิสระใดบ้างปรากฏอยู่ในโมเดล การพิจารณาเกี่ยวกับปัญหานี้ ก็คือการทดสอบสมมติฐานเกี่ยวกับ β_j; $j = 0, 1, \ldots, q-1$ นั่นเอง ทำได้โดยการเปรียบเทียบโมเดลโลจิสติกเชิงเส้นที่เป็นร่างแห (nested) กัน ทีละคู่ แล้วพิจารณาถึงผลกระทบของการนำตัวแปรเข้าหรือออกจากโมเดลว่า ทำให้ค่า deviance เปลี่ยนไปอย่างไร มีนัยสำคัญหรือไม่ วิธีการนี้เรียกว่าการวิเคราะห์ค่า deviance (analysis of deviance)

พิจารณาโมเดลโลจิสติกเชิงเส้น ที่เป็นร่างแหกัน 2 โมเดล ต่อไปนี้ คือ

โมเดล (1) : $\text{logit}(P_i) = \beta_0+\beta_1X_1+\ldots+\beta_hX_h$

โมเดล (2) : $\text{logit}(P_i) = \beta_0+\beta_1X_1+\ldots+\beta_hX_h+\beta_{h+1}X_{h+1}+\ldots+\beta_kX_k$

ให้ D_1 คือ deviance ของโมเดล (1) ซึ่งมี d.f.= $n-(h+1)$

และ D_2 คือ deviance ของโมเดล (2) ซึ่งมี d.f.= $n-(k+1)$

โมเดล (1) เป็นร่างแหของโมเดล (2) ซึ่งแน่นอนว่าค่า D_2 ย่อมมีค่าน้อยกว่า D_1 ผลต่างระหว่าง deviance ของ 2 โมเดล เป็น $D_1 - D_2$ คือการเปลี่ยนแปลงของ deviance ซึ่งเนื่องมาจากการนำตัวแปร $X_{h+1}, X_{h+2}, \ldots, X_k$ ใส่ในโมเดล ภายหลังจากในโมเดลมีตัวแปร $X_1, X_2, \ldots, X_h$ อยู่แล้ว

เนื่องจากได้กล่าวแล้วว่า deviance ของโมเดล จะมีการแจกแจงลู่เข้าสู่การแจกแจงแบบไคสแควร์ ดังนั้นจะได้ว่า การแจกแจงของ D_1-D_2 ก็จะลู่เข้าสู่การแจกแจงแบบไคสแควร์ด้วย ดังนี้

$$D_1 = -2[\ln L_{c1} - \ln L_f] \sim \chi^2 \quad ; \text{d.f.} = n-(h+1)$$

$$D_2 = -2[\ln L_{c2} - \ln L_f] \sim \chi^2 \quad ; \text{d.f} = n-(k+1)$$

$$D_1-D_2 = 2[\ln L_{c2} - \ln L_{c1}] \sim \chi^2 \quad ; \text{d.f.} = k-h$$

ดังนั้น ในการพิจารณาถึงผลกระทบของตัวแปรอิสระที่จะนำเข้าในโมเดล หรือนำออกจากโมเดล หรือการทดสอบ $H_o : \beta_{h+1} = \beta_{h+2} = \ldots = \beta_k = 0$ โดยมี H_1 : มีค่า β_j ; $j = h+1, h+2, \ldots, k$ อย่างน้อย 1 ค่าที่ไม่เท่ากับ 0 จะพิจารณาจากค่า D_1-D_2 ว่า มีนัยสำคัญหรือไม่ ซึ่งถ้าพบว่าไม่มีนัยสำคัญ คือ D_1 กับ D_2 มีค่าไม่แตกต่างกันทางสถิติ โมเดล (1) จะถูกนำไปใช้ ซึ่งหมายความว่า $\beta_{h+1} = \beta_{h+2} = \ldots = \beta_k = 0$ หรือตัวแปร $X_{h+1}, X_{h+2}, \ldots, X_k$ จะไม่มีอิทธิพลต่อโมเดล แต่ถ้าพบว่า D_1 กับ D_2 แตกต่างกันอย่างมีนัยสำคัญ ก็แสดงว่า มีค่า β_j ; $j = h+1, h+2, \ldots, k$ อย่างน้อย 1 ค่า ที่ไม่เท่ากับ 0 จะต้องมีการพิจารณาต่อไปว่า β_j ตัวใดที่มีค่าไม่เท่ากับ 0 ซึ่งถ้าพบว่า β_j ทุกค่ามีค่าไม่เท่ากับ 0 หมด โมเดล (2) จะถูกนำไปใช้

ในทางปฏิบัติ จะสร้าง all possible models ซึ่งอาจรวมเทอมต่าง ๆ ของตัวแปรอิสระที่มีการแปลงข้อมูลแล้ว และเทอมของอิทธิพลร่วมระหว่างตัวแปรอิสระที่เป็นกลุ่มด้วยขึ้นมาเปรียบเทียบกัน แล้วทำการวิเคราะห์ค่า deviance เพื่อเลือกโมเดลที่เหมาะสม พิจารณาได้จากตัวอย่างต่อไปนี้ คือ

จากการทดลองของ Hoblyn และ Palmer (1934) ซึ่งทดลองตัดรากของต้นพลัมพันธุ์ Common Mussel ในช่วงเวลาระหว่างเดือนตุลาคม 2477 และเดือนกุมภาพันธ์ 2478 ให้มีความยาว 2 ขนาด คือ ขนาด 12 และ 6 เซนติเมตร จำนวนขนาดละ 480 ส่วน และนำ

ครึ่งหนึ่งของแต่ละขนาดไปปลูกทันที สำหรับอีกครึ่งหนึ่งนำไปเพาะในทรายก่อนจนถึงฤดูใบไม้ผลิ แล้วจึงนำไปปลูก ทดลองจนถึงเดือนตุลาคม 2478 จึงนับจำนวนต้นพลัมที่มีชีวิตรอดอยู่ ได้ข้อมูลดังตาราง 4.1

ตาราง 4.1 อัตราการอยู่รอดของต้นพลัมจากการขยายพันธุ์โดยการตัดราก

ความยาวของรากที่ตัด	เวลาที่ปลูก	จำนวนที่อยู่รอด จากทั้งหมด 240 ส่วน	สัดส่วนที่อยู่รอด
6 เซนติเมตร	ปลูกทันที	107	0.45
	ปลูกในฤดูใบไม้ผลิ	31	0.13
12 เซนติเมตร	ปลูกทันที	156	0.65
	ปลูกในฤดูใบไม้ผลิ	84	0.35

Collett(1991) ได้นำข้อมูลนี้ไปวิเคราะห์การถดถอย ใช้โมเดลโลจิสติกเชิงเส้น จำนวน 5 โมเดล คือ

โมเดล (1) : $\text{logit}(\hat{p}_{jk}) = b_o$

โมเดล(2) : $\text{logit}(\hat{p}_{jk}) = b_o + b_1 X_j$

โมเดล(3) : $\text{logit}(\hat{p}_{jk}) = b_o + b_2 X_k$

โมเดล(4) : $\text{logit}(\hat{p}_{jk}) = b_o + b_1 X_j + b_2 X_k$

โมเดล(5) : $\text{logit}(\hat{p}_{jk}) = b_o + b_1 X_j + b_2 X_k + b_3 X_j X_k$

เมื่อ $\hat{p}_{jk}$ คือ ความน่าจะเป็นที่ต้นพลัมจากรากจะมีชีวิตอยู่รอด เมื่อรากที่ปลูกมีขนาด j และปลูกในเวลา k ; $j = 1, 2$

$k = 1, 2$

b_o คือ ค่าคงที่

b_1 คือ สัมประสิทธิ์การถดถอยเนื่องจากความยาวราก

b_2 คือ สัมประสิทธิ์การถดถอย เนื่องจากเวลาปลูก

b_3 คือ สัมประสิทธิ์การถดถอย เนื่องจากอิทธิพลร่วมระหว่างความยาวราก และเวลาปลูก

X_j คือ ความยาวรากขนาดที่ j ; $X_1 = 0$, $X_2 = 1$

X_k คือ เวลาปลูกที่ k ; $X_1 = 0$, $X_2 = 1$

ผลการวิเคราะห์การถดถอยโลจิสติก ด้วยโปรแกรมสำเร็จรูป GLIM แสดงค่า deviance ของแต่ละโมเดล ในตาราง 4.2 และการวิเคราะห์ deviance ในตาราง 4.3

ตาราง 4.2 ค่า deviance ของแต่ละโมเดล

โมเดล	Deviance	d.f.
b_o	151.02	3
$b_o + b_1X_j$	105.18	2
$b_o + b_2X_k$	53.44	2
$b_o + b_1X_j + b_2X_k$	2.29	1
$b_o + b_1X_j + b_2X_k + b_3X_jX_k$	0.00	0

ตาราง 4.3 การวิเคราะห์ deviance

Source of variation	Deviance	d.f.
ความยาวรากเมื่อไม่มีเวลาปลูก	151.02 - 105.18 = 45.84	1
เวลาปลูกเมื่อไม่มีความยาวราก	151.02 - 53.44 = 97.58	1
ความยาวรากเมื่อมีเวลาปลูก	53.44 - 2.29 = 51.15	1
เวลาปลูกเมื่อมีความยาวราก	105.18 - 2.29 = 102.89	1
มีอิทธิพลร่วม	2.29	1

เมื่อพิจารณาค่า deviance ที่ลดลง จากตาราง 4.3 พร้อมกับพิจารณาค่า deviance จากตาราง 4.2 ควบคู่กันพบว่า ควรจะเลือกใช้ โมเดล (4) ซึ่งมีเฉพาะอิทธิพลของปัจจัยหลัก คือความยาวรากกับเวลาที่ปลูก เพราะสามารถลดค่าของ deviance ลงได้มากที่สุด เท่ากับ 102.89 ซึ่งมีนัยสำคัญ และค่า deviance ของ โมเดล (4) เท่ากับ 2.29 จะไม่มีนัยสำคัญที่ระดับ 10% (มีค่าน้อยกว่าค่า χ^2 จากตารางที่ d.f.= 1, ระดับนัยสำคัญ = 0.10) ซึ่งหมายความว่าโมเดล (4) มีภาวะสารูปดีกับค่าสังเกต

ค่าประมาณของพารามิเตอร์ จากโปรแกรมสำเร็จรูป GLIM เป็นดังนี้

scale deviance = 2.2938 at cycle 3

d.f. = 1

	estimate	s.e.	parameter
1	-0.3039	0.1172	1
2	1.018	0.1455	LENG(2)
3	-1.428	0.1465	TIME(3)

แสดงว่า $b_o = -0.3039$

$b_1 = 1.018$

$b_2 = -1.428$

จากโมเดล (4) : $\text{logit}(\hat{p}_{jk}) = b_o + b_1X_j + b_2X_k$

จะได้ $\text{logit}(\hat{p}_{11}) = b_o + b_1(0) + b_2(0) = -0.3039$

$\text{logit}(\hat{p}_{12}) = b_o + b_1(0) + b_2(1) = -0.3039 - 1.428 = -1.732$

$\text{logit}(\hat{p}_{21}) = b_o + b_1(1) + b_2(0) = -0.3039 + 1.018 = 0.714$

$\text{logit}(\hat{p}_{22}) = b_o + b_1(1) + b_2(1) = -0.3039 + 1.018 - 1.428$

$= -0.714$

และ $\hat{p}_{11} = [1 + \exp-(-0.3039)]^{-1} = 0.425$

$\hat{p}_{12} = [1 + \exp-(-1.732)]^{-1} = 0.150$

$\hat{p}_{21} = [1 + \exp-(0.714)]^{-1} = 0.671$

$\hat{p}_{22} = [1 + \exp-(-0.714)]^{-1} = 0.329$

ซึ่งเปรียบเทียบกับค่าในตาราง 4.1 ได้ตาราง 4.4 เป็น

ตาราง 4.4 เปรียบเทียบความน่าจะเป็นระหว่างค่าที่สังเกตได้ กับค่าประมาณ

ความยาวของรากที่ตัด	เวลาที่ปลูก	ความน่าจะเป็นที่จะอยู่รอด	
		ค่าสังเกต	ค่าประมาณ
6 เซนติเมตร	ปลูกทันที	0.446	0.425
	ปลูกในฤดูใบไม้ผลิ	0.129	0.150
12 เซนติเมตร	ปลูกทันที	0.650	0.671
	ปลูกในฤดูใบไม้ผลิ	0.350	0.329

การทดสอบสมมติฐานเกี่ยวกับ β_j อาจทำการทดสอบได้อีกวิธีหนึ่ง โดยการพิจารณาการแจกแจงของตัวประมาณ $\hat{\beta}_j = b_j$ ซึ่งมีการแจกแจงใกล้เคียงการแจกแจงแบบปกติ[2] จะได้ตัวทดสอบสถิติ เป็น

$$Z = \frac{b_j}{\sqrt{Var(b_j)}}$$

ซึ่งมีการแจกแจงใกล้เคียงการแจกแจงแบบปกติมาตรฐาน

หรือ $$Z^2 = \frac{b_j^2}{Var(b_j)}$$

ซึ่งมีการแจกแจงใกล้เคียงการแจกแจงแบบไคสแควร์ ด้วย d.f.= 1

[2] McCullagh and Nelder. Generalized liner models (1983) pp.83

4.6 การทำนายค่าความน่าจะเป็นจากโมเดล

วัตถุประสงค์หลักของการวิเคราะห์การถดถอยก็คือ การทำนายค่า (prediction) หรือการประมาณค่า (estimation) ของตัวแปรตามจากโมเดลประมาณที่วิเคราะห์ได้ ในการวิเคราะห์การถดถอยโลจิสติก ค่าทำนายที่ต้องการคือความน่าจะเป็นที่จะได้ข้อมูลทวิมีค่าเท่ากับ 1 หรือ $\hat{p}_i$

$$\text{โดย} \quad \hat{p}_i = [1 + \exp(-\sum_{j=0}^{q-1} b_j X_{ji})]^{-1}$$

เนื่องจากการวิเคราะห์การถดถอยโลจิสติก จะทำการแปลงค่า $\hat{p}_i$ ให้อยู่ในรูปของ $\text{logit}(\hat{p}_i)$ ซึ่งจะได้ว่า $\text{logit}(\hat{p}_i)$ มีความสัมพันธ์เชิงเส้นกับตัวแปรอิสระ X_{ji} มีสมการถดถอยโลจิสติกเชิงเส้น เป็น

$$\text{logit}(\hat{p}_i) = u_i = \sum_{j=0}^{q-1} b_i X_{ji} \quad ; \; X_{0i} = 1 \qquad (4.6.1)$$

การหาค่าประมาณของ P_o ณ ค่า X_{jo} ที่กำหนด กระทำได้โดยการประมาณค่า $\text{logit}(\hat{p}_i)$ ณ ค่า X_{jo} ที่กำหนด จากสมการ (4.6.1) ได้ค่า $\text{logit}(\hat{p}_o)$ หรือ u_o เป็นค่าประมาณของ $\text{logit}(P_o)$ หรือ U_o โดยความแปรปรวนของ u_o คือ $V(u_o)$ เป็น

$$V(u_o) = \sum_{j=0}^{q-1} X_{jo}^2 V(b_j) + 2 \sum_{j,h=0}^{q-1} X_{jo} X_{ho} \text{Cov}(b_j b_h) \; ; \; X_{oo} = 1$$

และได้ช่วงความเชื่อมั่น $(1 - \alpha)\%$ สำหรับค่าประมาณของ U_o มีค่าเป็น

$$u_o - Z_{\alpha/2} \sqrt{V(u_o)} < U_o < u_o + Z_{\alpha/2} \sqrt{V(u_o)}$$

ถ้าให้ $u_{oL} = u_o - Z_{\alpha/2}\sqrt{V(u_o)}$ = ขีดจำกัดล่างของค่าประมาณของ U_o

$u_{oU} = u_o + Z_{\alpha/2}\sqrt{V(u_o)}$ = ขีดจำกัดบนของค่าประมาณของ U_o

จะได้ $p_{oL} = [1 + \exp(-u_{oL})]^{-1}$ = ขีดจำกัดล่างของค่าประมาณของ P_o

$p_{oU} = [1 + \exp(-u_{oU})]^{-1}$ = ขีดจำกัดบนของค่าประมาณของ P_o

4.7 การตรวจสอบโมเดล

เมื่อประมาณโมเดลการถดถอยโลจิสติกเชิงเส้น และวิเคราะห์ค่า deviance เพื่อทดสอบพารามิเตอร์ β_j จนได้โมเดลที่ต้องการแล้ว อาจพบว่าค่า deviance ของการทดสอบภาวะสารูปดี ยังคงมีนัยสำคัญอยู่ ซึ่งหมายความว่าโมเดลที่ประมาณได้ยังไม่พอดีกับค่าสังเกต ควรทำการตรวจสอบโมเดล เพื่อหาสาเหตุของความไม่พอดี หรือความไม่เหมาะสมของโมเดล ซึ่งจะกระทำได้ในทำนองเดียวกับการพิจารณาหาสาเหตุของความไม่เหมาะสมของโมเดล ในการวิเคราะห์การถดถอยเชิงเส้น ที่กล่าวในบทที่ 2

สำหรับความคลาดเคลื่อนที่ใช้ในการตรวจสอบความเหมาะสมของโมเดล ในการวิเคราะห์การถดถอยโลจิสติกเชิงเส้น ได้แก่

1. Pearson residual

$$X_i = \frac{y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1 - \hat{p}_i)}}$$

ซึ่ง $\sum_{i=n}^{n} X_i^2$ ก็คือค่า Pearson's X^2- statistic

และมี standardized Pearson residual เป็น

$$r_{Pi} = \frac{y_i - n_i\hat{p}_i}{\sqrt{v_i(1 - h_i)}}$$

เมื่อ $v_i = n_i\hat{p}_i(1 - \hat{p}_i)$

h_i = คือค่าในแนวทะแยงของแมทริกซ์ H ขนาด n x n

$H = W^{1/2}X(X'WX)^{-1}X'W^{1/2}$

W = weighted matrix ขนาด n x n ที่ใช้ในการประมาณค่าพารามิเตอร์ β_j

2. Deviance residual

$$d_i = sgn(y_i - \hat{y}_i)\left[2\left\{y_i \ln\frac{y_i}{\hat{y}_i} + (n_i - y_i)\ln\frac{n_i - y_i}{n_i - \hat{y}_i}\right\}\right]^{1/2}$$

$sgn(y_i - \hat{y}_i)$ คือฟังก์ชั่นที่ทำให้ค่า d_i มีค่าเป็นบวก เมื่อ $y_i \geq \hat{y}_i$ และมีค่าเป็นลบ เมื่อ $y_i < \hat{y}_i$ ซึ่ง $\sum_{i=1}^{n} d_i^2$ ก็คือค่า deviance

และมี standardized deviance residual เป็น

$$r_{Di} = \frac{d_i}{\sqrt{1 - h_i}}$$