การประเมินความเสี่ยงด้วยตัวแบบการถดถอยโลจิสติก
Risk Assessment through the Logistic Regression Model

*รศ.ร.อ.มานพ วราภักดิ์**

ความเสี่ยง (risk) เป็นความเป็นไปได้ของความเสียหาย หรือความสูญเสีย หรือภยันตราย ที่จะเกิดขึ้น เช่น ความเสี่ยงไม่สามารถชำระหนี้ได้ครบตามสัญญา หรือความเสี่ยงหนี้จะสูญ ความเสี่ยงติดเชื้อ ความเสี่ยงป่วยเป็นโรคมะเร็ง และความเสี่ยงสูญเสียชีวิต เป็นต้น มีปัจจัยต่าง ๆ ที่ก่อให้เกิดความเสี่ยงในแต่ละเรื่อง และพยายามป้องกัน หรือหลีกเลี่ยงความเสี่ยงหรือภัยที่จะเกิดขึ้น ซึ่งมีหลายวิธีการในการป้องกัน

การประเมินความเสี่ยงในเรื่องหนึ่ง ๆ เพื่อจะป้องกัน หรือหลีกเลี่ยงไม่ให้เกิดความเสียหาย ความสูญเสีย หรือภยันตรายต่าง ๆ มีหลายวิธีการแล้วแต่กรณี ในบทความนี้ ขอนำเสนอวิธีการเชิงสถิติที่นิยมใช้กันมากวิธีหนึ่ง มีชื่อเรียกว่า "การวิเคราะห์การถดถอยโลจิสติก" (logistic regression analysis) เพื่อหาตัวแบบสถิติ (statistical model) สำหรับประเมินความเสี่ยง ซึ่งผลลัพธ์หรือค่าวัดที่ได้จากตัวแบบ สามารถนำไปใช้ประโยชน์ประกอบการตัดสินใจ หรือในการพิจารณาป้องกันไม่ให้เกิดความความเสียหาย ความสูญเสีย หรือภยันตราย

1. ตัวแบบการถดถอยโลจิสติก

ตัวแบบการถดถอย (regression model) หรือสมการถดถอย เป็นตัวแบบที่นิยมใช้กันมากตัวแบบหนึ่งในการแสดงความสัมพันธ์เชิงคณิตศาสตร์ระหว่างตัวแปรตัวหนึ่ง ซึ่งเรียกทั่วไปว่า "ตัวแปรตาม" (dependent variable) กับตัวแปรอีกตัวหนึ่ง หรือหลายตัว เรียกตัวแปรเหล่านี้ว่า "ตัวแปรอิสระ" (independent variables) ซึ่งมีผลกระทบต่อค่าของตัวแปรตาม หรือค่าของตัวแปรตามจะขึ้นอยู่กับค่าของตัวแปรอิสระ ตัวอย่าง เช่น ตัวแปรตามที่เป็นจำนวนคนว่างงาน จะขึ้นอยู่กับปัจจัย หรือตัวแปรอิสระหลายตัว อาทิเช่น กำลังแรงงานรวม ผลิตภัณฑ์ในประเทศเบื้องต้น [gross domestic product (GDP)] ค่าใช้จ่ายภาครัฐ มูลค่าส่งออกรวม และมูลค่านำเข้ารวม เป็นต้น

---

* อาจารย์ประจำภาควิชาสถิติ คณะพาณิชยศาสตร์และการบัญชี จุฬาลงกรณ์มหาวิทยาลัย

ให้ Y เป็นตัวแปรตาม และให้ $x_1, x_2, \ldots, x_p$ เป็นตัวแปรอิสระ โดยทั่วไปจะได้ว่า ที่แต่ละค่าคงที่ชุดหนึ่ง ๆ ของ $(x_1, x_2, \ldots, x_p)$ จะมีการแจกแจงความน่าจะเป็นของ Y โดยมี $E(Y|x_1, x_2, \ldots, x_p)$ เป็นค่าเฉลี่ย ดังนั้น ได้ความสัมพันธ์

$$Y = E(Y|x_1, x_2, \ldots, x_p) + \varepsilon$$

โดยที่ ε เป็นค่าคลาดเคลื่อนสุ่ม (random error) แทนผลต่างระหว่างค่าจริง Y และค่าเฉลี่ย $E(Y|x_1, x_2, \ldots, x_p)$ ของ Y เมื่อกำหนดค่าของ $(x_1, x_2, \ldots, x_p)$

ค่าเฉลี่ย $E(Y|x_1, x_2, \ldots, x_p)$ จะเป็นฟังก์ชันของ $x_1, x_2, \ldots, x_p$ และรูปแบบของฟังก์ชันรูปแบบหนึ่งที่ใช้กันมากคือ ฟังก์ชันเชิงเส้นในเทอมของพารามิเตอร์ $\beta_0, \beta_1, \ldots, \beta_p$ เขียนได้เป็น

$$Y = E(Y|x_1, x_2, \ldots, x_p) + \varepsilon = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

และสำหรับค่าสังเกต $Y = Y_i$ และ $x_1 = x_{i,1}, x_2 = x_{i,2}, \ldots, x_p = x_{i,p}$ จะได้

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \varepsilon_i, \quad i = 1, 2, \ldots, n \tag{1}$$

โดยมีข้อสมมติพื้นฐานของค่าคลาดเคลื่อนสุ่ม $\varepsilon_i, \ i = 1, 2, \ldots, n$ ดังนี้คือ มีค่าเฉลี่ย $E(\varepsilon_i) = 0$ มีความแปรปรวน $Var(\varepsilon_i)$ คงที่ สมมติเท่ากับ σ^2 ไม่มีความแปรปรวนร่วม นั่นคือ $Cov(\varepsilon_i, \varepsilon_j) = 0$ สำหรับ $i \neq j$ (หรือ ε_i ไม่มีสหสัมพันธ์เชิงเส้นต่อกัน) และโดยทั่วไปจะสมมติด้วยว่า ε_i มีการแจกแจงปกติ เพราะฉะนั้น $Y_1, Y_2, \ldots, Y_n$ เป็นอิสระต่อกัน และ Y_i มีการแจกแจงปกติโดยมีค่าเฉลี่ย $E(Y_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p}$ และความแปรปรวน $Var(Y_i) = \sigma^2, \ i = 1, 2, \ldots, n$

เรียกตัวแบบ (1) ว่า "ตัวแบบการถดถอยเชิงเส้นพหุคูณ" (multiple linear regression model) (เชิงเส้นในเทอมของพารามิเตอร์) ในกรณีที่ตัวแบบ (1) มีตัวแปรอิสระเพียงตัวเดียว เรียกตัวแบบ (1) ว่า "ตัวแบบการถดถอยเชิงเส้นเชิงเดียว" (simple linear regression model)

ในการวิเคราะห์การถดถอย เมื่อตัวแปรตาม Y มีค่าแบบทวิภาค (binary) คือมีค่าเป็นไปได้ 0 และ 1 เท่านั้น หรือนั่นคือ Y มีการแจกแจงแบร์นูลลี (Bernoulli distribution) $b(1, \pi), \ 0 \le \pi \le 1$ โดยมีฟังก์ชันความน่าจะเป็น $p_Y(y) = \pi^y (1-\pi)^{1-y}, \ y = 0, 1$ ต่างจากกรณี Y ในตัวแบบการถดถอยทั่วไป ซึ่งมีค่าเป็นจำนวนจริงใด ๆ ตัวอย่างค่าตัวแปรตาม Y ในการประยุกต์กรณีการแจกแจงแบร์นูลลี เช่น 1 แทนเป็นโรคหัวใจ 0 แทนไม่เป็นโรคหัวใจ 1

แทนหนี้ไม่สูญ 0 แทนหนี้สูญ 1 แทนพอใจ 0 แทนไม่พอใจ 1 แทนเห็นด้วย 0 แทนไม่เห็นด้วย 1 แทนได้กำไร 0 แทนไม่ได้กำไร และ 1 แทนสอบผ่าน 0 แทนสอบไม่ผ่าน เป็นต้น ในกรณีนี้ ไม่อาจจะใช้ตัวแบบการถดถอยทั่วไป (1) ได้ เพราะว่า ในกรณี Y_i มีการแจกแจง $b(1,\pi)$ อาจได้ค่า

$$E(Y_i) = E(Y_i|x_{i,1},\ldots,x_{i,p}) = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}$$

ซึ่งเท่ากับความน่าจะเป็น $\pi = \pi(\underline{x}'_i) = \pi(x_{i,1},\ldots,x_{i,p}) = P(Y_i = 1|x_{i,1},\ldots,x_{i,p})$ มีค่าไม่อยู่ในช่วง $[0,1]$ (ขัดแย้งความเป็นจริงของความน่าจะเป็น) ทั้งนี้เพราะ โดยทั่วไปไม่มีข้อจำกัดของค่า β_i และ $x_{i,j}$ ที่จะให้ได้ $E(Y_i)$ มีค่าอยู่ในช่วง $[0,1]$ เพราะฉะนั้น จำเป็นต้องพิจารณาการแปลงหรือฟังก์ชันแบบต่อเนื่องของ $(x_{i,1}, x_{i,2},\ldots,x_{i,p})$ ที่มีค่าอยู่ระหว่าง 0 และ 1 เสมอ ซึ่งได้ว่าฟังก์ชันหนึ่งคือ ฟังก์ชันการแจกแจงสะสม (cumulative distribution function) ของความน่าจะ-เป็นของตัวแปรสุ่มต่อเนื่อง และฟังก์ชันการแจกแจงสะสมที่นิยมใช้กันมากในทางปฏิบัติฟังก์ชันหนึ่งคือ ฟังก์ชันการแจกแจงสะสมของ "การแจกแจงโลจิสติก" (logistic distribution) มีรูปกราฟคล้ายตัวอักษร S เช่น ตัวอย่างรูปที่ 1 ซึ่งรูปแบบความสัมพันธ์ระหว่าง x กับ $\pi(x)$ สำหรับตัวแปรค่าต่อเนื่อง x หรือ ระหว่าง $\pi(h(x_1,\ldots,x_p))$ กับฟังก์ชันเชิงเส้น (ในเทอมของพารามิเตอร์) ต่อเนื่อง $h(x_1,\ldots,x_p)$ มักจะพบว่ามีรูปแบบความสัมพันธ์คล้ายตัวอักษร S จึงเป็นเหตุผลหนึ่งในการเลือกการแจกแจงโลจิสติก เขียนฟังก์ชันการแจกแจงสะสมของการแจกแจงโลจิสติกได้เป็น

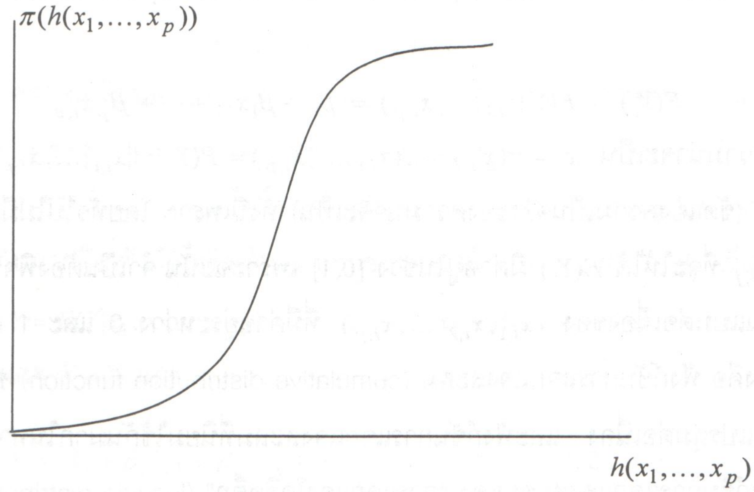
$$F[h(x_{i,1},\ldots,x_{i,p})] = \frac{\exp[h(x_{i,1},\ldots,x_{i,p})]}{1+\exp[h(x_{i,1},\ldots,x_{i,p})]} \qquad (2)$$

เพราะฉะนั้น เมื่อกำหนด $E(Y_i|x_{i,1},\ldots,x_{i,p})$ เป็นฟังก์ชันการแจกแจงสะสมของการแจกแจงโลจิสติก และถ้าเลือก

$$h(\underline{x}'_i) = h(x_{i,1},\ldots,x_{i,p}) = \underline{x}'_i\,\underline{\beta}$$

$$= \left(1,\ x_{i,1},\ x_{i,2},\ \ldots,\ x_{i,p}\right)\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \vdots\\ \beta_p\end{pmatrix}$$

$$= \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p}$$



รูปที่ 1

เป็นฟังก์ชันเชิงเส้นในเทอมของพารามิเตอร์ $\beta_0, \beta_1, \ldots, \beta_p$ จะได้ **ตัวแบบการถดถอยโลจิสติก** (logistic regression model) **ดังนี้**

$$Y_i = E(Y_i | x_{i,1}, \ldots, x_{i,p}) + \varepsilon_i$$

$$= \frac{\exp(\underline{x}'_i \, \underline{\beta})}{1 + \exp(\underline{x}'_i \, \underline{\beta})} + \varepsilon_i$$

$$= \frac{\exp(\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p})}{1 + \exp(\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p})} + \varepsilon_i \qquad (3)$$

โดยมี

$$E(Y_i) = E(Y_i | x_{i,1}, \ldots, x_{i,p}) = \pi(x_{i,1}, \ldots, x_{i,p}) = \frac{\exp(\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p})}{1 + \exp(\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p})} \qquad (4)$$

จากนี้ทำการประมาณค่าพารามิเตอร์ $\beta_0, \beta_1, \ldots, \beta_p$ สมมติได้ค่าประมาณเป็น $\hat{\beta}_0$, $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ ตามลำดับ ดังนั้น **ตัวแบบประมาณความน่าจะเป็น** $\pi(\underline{x}_0) = P(Y = 1 | \underline{x}_0)$ สำหรับค่า $\underline{x} = \underline{x}_0$ ที่กำหนด คือ

$$\hat{\pi}(\underline{x}_0) = \hat{\pi}(x_{0,1}, \ldots, x_{0,p}) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{0,1} + \cdots + \hat{\beta}_p x_{0,p})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{0,1} + \cdots + \hat{\beta}_p x_{0,p})} \qquad (5)$$

ค่า $\hat{\pi}(\underline{x}_0)$ ที่ได้จะเป็นค่าประมาณความน่าจะเป็นของ $Y=1$ นั่นคือ ค่าประมาณความน่าจะเป็นที่จะเกิดเหตุการณ์ที่สนใจซึ่งแทนด้วย 1 เช่น ความน่าจะเป็นที่ไม่สามารถชำระหนี้ได้ครบตามสัญญา ดังนั้น ถ้าค่า $\hat{\pi}(\underline{x}_0)$ มีค่าสูงเข้าใกล้ 1 แสดงว่ามีความเป็นไปได้สูงที่จะเกิดเหตุการณ์นั้น และอาจพยากรณ์ว่าจะเกิดเหตุการณ์นั้น

ในทางปฏิบัติของการใช้ประโยชน์ของตัวแบบประมาณความน่าจะเป็น $\hat{\pi}(\underline{x})$ ตัวแปรอิสระ หรือปัจจัย x_i ต่าง ๆ จะผ่านการคัดเลือกด้วยวิธีการต่าง ๆ เชิงสถิติ เช่น การทดสอบนัยสำคัญของพารามิเตอร์ ดังนั้น ในตัวแบบอาจจะมีปัจจัย x_i บางตัวเท่านั้น (จากทั้งหมดที่นำมาวิเคราะห์) ปรากฏในตัวแบบ $\hat{\pi}(\underline{x})$

2. การประมาณค่าพารามิเตอร์

ในการประมาณค่าของพารามิเตอร์ $\beta_0, \beta_1, \ldots, \beta_p$ ในตัวแบบการถดถอยโลจิสติก มีหลายวิธีการ ในที่นี้จะกล่าวถึงแนวคิดของวิธีการความควรจะเป็นสูงสุด [maximum likelihood (ML) method] ในการประมาณค่า β_i ซึ่งเป็นวิธีการหนึ่งที่ใช้กันมาก

ให้ $Y_1, Y_2, \ldots, Y_n$ เป็นอิสระกัน และ $Y_i \sim b(1, \pi_i)$ โดยมีฟังก์ชันความน่าจะเป็น (probability function)

$$p(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \quad \text{สำหรับ} \quad y_i = 0, 1; \quad i = 1, 2, \ldots, n \tag{6}$$

โดยที่

$$\begin{aligned} \pi_i &= \pi(\underline{x}_i) = P(Y_i = 1 | \underline{x}_i) \\ &= F[h(\underline{x}'_i)] \\ &= F(\underline{x}'_i \underline{\beta}) \end{aligned}$$

เพราะฉะนั้น ฟังก์ชันความน่าจะเป็นร่วม (joint probability function) หรือ ฟังก์ชันความควรจะเป็น (likelihood function) ของ $Y_1, Y_2, \ldots, Y_n$ คือ

$$\begin{aligned} L(\underline{y}; \underline{\beta}) &= \prod_{i=1}^{n} p(y_i) \\ &= \prod_{i=1}^{n} [\pi(\underline{x}_i)]^{y_i} [1 - \pi(\underline{x}_i)]^{1-y_i} \end{aligned} \tag{7}$$

โดยทั่วไป จะใช้วิธี Newton-Raphson ในการหาค่าประมาณความควรจะเป็นสูงสุด (ML estimate) ของ $\underline{\beta}$ ซึ่งค่าประมาณความควรจะเป็นสูงสุดของ $\underline{\beta}$ เป็นค่าของ $\underline{\beta}$ ที่ได้ $L(\underline{y};\underline{\beta})$ มีค่ามากที่สุด หรือได้ $\ln L(\underline{y};\underline{\beta})$ มีค่ามากที่สุด โดยที่

$$\ln L(\underline{y};\underline{\beta}) = \sum_{i=1}^{n} y_i \ln[\pi(\underline{x}_i)] + \sum_{i=1}^{n}(1-y_i)\ln[1-\pi(\underline{x}_i)] \tag{8}$$

สมมติจากข้อมูล หรือค่าสังเกตจำนวน n จำแนกค่าของตัวแปรอิสระ p ตัวได้เป็นกลุ่มต่าง ๆ ไม่ซ้ำซ้อนกัน m กลุ่ม และสมมติกลุ่มที่ i มีจำนวนค่าสังเกต n_i ซึ่ง $\sum_{i=1}^{m} n_i = n$ และให้ W_i แทนจำนวนค่า Y เป็น 1 ในจำนวน n_i ฉะนั้น จะมีค่า Y เป็น 0 จำนวน $n_i - W_i$ และได้ว่า W_i มีการแจกแจงทวินาม (binomial distribution) ด้วยพารามิเตอร์ n_i และ $\pi(\underline{x}_i)$ และมีฟังก์ชันความน่าจะเป็น

$$p_{W_i}(w_i) = \binom{n_i}{w_i}[\pi(\underline{x}_i)]^{w_i}[1-\pi(\underline{x}_i)]^{n_i - w_i},\ w_i = 0,1,2,\ldots,n_i;\ i = 1,2,\ldots,m$$

และดังนั้น ได้ ln ของฟังก์ชันความควรจะเป็นของ $W_1, W_2, \ldots, W_m$ เป็น

$$\ln L(\underline{w};\underline{\beta}) = \sum_{i=1}^{m}\ln\binom{n_i}{w_i} + \sum_{i=1}^{m} w_i \ln[\pi(\underline{x}_i)] + \sum_{i-1}^{m}(n_i - w_i)\ln[1-\pi(\underline{x}_i)] \tag{9}$$

ซึ่งใช้ประโยชน์ในการหาตัวประมาณ ML ของ $\underline{\beta}$ เช่นกัน

3. การทดสอบสมมติฐานสำหรับนัยสำคัญของพารามิเตอร์

ในการตรวจสอบนัยสำคัญของพารามิเตอร์ β_i ในตัวแบบการถดถอยโลจิสติก เพื่อการพิจารณาคัดเลือกตัวแปรอิสระ x_i เข้าในตัวแบบ จะทำการทดสอบสมมติฐาน $H_0 : \beta_i = 0$ เทียบกับ $H_1 : \beta_i \neq 0$ และทดสอบด้วยตัวสถิติ

$$Z = \frac{\hat{\beta}_i}{Se(\hat{\beta}_i)} \tag{10}$$

ซึ่ง Z มีการแจงแจงปกติมาตรฐาน $N(0,1)$ โดยประมาณ หรือใช้ตัวสถิติไคกำลังสอง ซึ่งเป็นแบบทดสอบวอลด์ (Wald test)

$$W = \left(\frac{\hat{\beta}_i}{Se(\hat{\beta}_i)}\right)^2 \tag{11}$$

โดยที่ตัวสถิติวอลด์ มีการแจกแจงไคกำลังสอง (chi-squares distribution) $\chi^2(1)$ โดยประมาณ ซึ่ง $Se(\hat{\beta}_i)$ คือ ค่าคลาดเคลื่อนมาตรฐาน (standard error) ของตัวประมาณ $\hat{\beta}_i$

เกณฑ์การทดสอบคือ ปฏิเสธสมมติฐาน $H_0: \beta_i = 0$ (ยอมรับ $H_1: \beta_i \neq 0$) ถ้าค่าของ W มีค่ามากกว่าค่าวิกฤติ $\chi^2_{\alpha/2}(1)$ หรือ W มีค่าน้อยกว่าค่าวิกฤติ $\chi^2_{1-\alpha/2}(1)$ หรือ ค่าความน่าจะเป็น p-value น้อยกว่าระดับนัยสำคัญ α [$\chi^2_{\alpha/2}(1)$ คือเปอร์เซ็นไทล์ (percentile) ที่ $100\left(1-\frac{\alpha}{2}\right)$ ของการแจกแจงไคกำลังสอง $\chi^2(1)$]

สำหรับการทดสอบสัมประสิทธิ์ p ตัวพร้อมกันโดยมี $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ เทียบกับ $H_1: \beta_i \neq 0$ อย่างน้อยหนึ่งค่า $i,\ i = 1,2,\ldots,p$ สามารถทดสอบด้วยตัวสถิติทดสอบอัตราส่วนความควรจะเป็น (likelihood ratio test statistic) :

$$LRTS = 2\left\{\sum_{i=1}^{m} w_i \ln[\hat{\pi}(\underline{x}_i)] + \sum_{i=1}^{m}(n_i - w_i)\ln[1-\hat{\pi}(\underline{x}_i)] - [w\ln(w) + (n-w)\ln(n-w) - n\ln(n)]\right\} \tag{12}$$

โดยที่ $\hat{\pi}(\underline{x}_i) = \dfrac{\exp(\underline{x}'_i\,\hat{\underline{\beta}})}{1+\exp(\underline{x}'_i\,\hat{\underline{\beta}})}$, $\hat{\underline{\beta}}$ เป็นตัวประมาณ ML และ w คือจำนวนค่าของ Y เป็น 1 ในจำนวนข้อมูลหรือค่าสังเกตทั้งหมด n ค่า

ตัวสถิติ $LRTS$ มีการแจกแจงไคกำลังสอง $\chi^2(p)$ และดังนั้น จะปฏิเสธ H_0 ที่ระดับนัยสำคัญ α ถ้าค่า $LRTS > \chi^2_\alpha(p)$

4. การวินิจฉัยตัวแบบ

ตัวแบบที่สร้างขึ้นอาจมีความไม่เหมาะสม หรือไม่พอเพียง (inadequate) เชิงสถิติ ในลักษณะที่ว่ารูปแบบของตัวแบบไม่เหมาะสม หรือตัวแบบยังไม่กลมกลืน (lack of fit) กับข้อมูลที่นำมาสร้างตัวแบบ หรือในลักษณะที่ว่า ตัวแบบไม่สอดคล้องข้อสมมติต่าง ๆ เช่น คุณสมบัติของค่าคลาดเคลื่อนสุ่ม และในบางกรณีอาจมีค่าผิดปกติ (outliers) ที่มีผลกระทบอย่างมากต่อความกลมกลืน (fit) ของตัวแบบ หรือต่อค่าประมาณของพารามิเตอร์ ความไม่เหมาะสม หรือความไม่พอเพียงเหล่านี้ อาจจะส่งผลให้เกิดความผิดพลาดสูงในผลสรุป หรือในการพยากรณ์ ดังนั้น จึง

ควรทำการวินิจฉัยตัวแบบ และถ้าพบตัวแบบมีความไม่เหมาะสม หรือไม่พอเพียง ควรปรับแก้ตัวแบบให้มีความเหมาะสมก่อนนำไปใช้ประโยชน์

ในการตรวจสอบความเหมาะสม หรือความพอเพียงของตัวแบบ มีค่าวัด หรือตัวสถิติหลายตัวที่อาจนำมาใช้ เช่น ค่าวัด หรือตัวสถิติต่อไปนี้

ส่วนตกค้าง (residuals) ซึ่งเป็นผลต่าง $Y_i - \hat{\pi}(\underline{x}_i),\ i = 1, 2, \ldots, n$ จะนำมาใช้ประโยชน์ในการวินิจฉัยตัวแบบ เพื่อตรวจสอบความพอเพียง (adequacy) เชิงสถิติของตัวแบบ หรือความสอดคล้อง หรือความกลมกลืน (goodness-of-fit) ของตัวแบบกับข้อมูล ส่วนตกค้างที่นิยมใช้กันคือ ส่วนตกค้างที่เป็น Studentized residuals, Deviance residuals หรือ Standardized residuals ค่าส่วนตกค้างเหล่านี้ควรมีค่าสัมบูรณ์ไม่มากกว่า 3 สำหรับตัวแบบที่มีความพอเพียง หรือมีความกลมกลืน ค่าสัมบูรณ์ที่มากกว่า 3 มักจะพบว่ามีค่าผิดปกติในกลุ่มข้อมูล

ในการทดสอบที่เป็นระเบียบวิธีทางสถิติสำหรับการทดสอบเทียบความกลมกลืน มีตัวสถิติทดสอบหลายตัว อาทิเช่น ตัวสถิติทดสอบ Hosmer-Lemeshow ตัวสถิติทดสอบ Pearson และตัวสถิติทดสอบ Deviance ตัวสถิติทดสอบเหล่านี้ต่างมีการแจกแจงไคกำลังสอง

นอกจากนี้ มีตัวสถิติอื่น ๆ ประกอบการพิจารณาความกลมกลืน ได้แก่ตัวสถิติ - 2 Log Likelihood และ Goodness of Fit ซึ่งตัวแบบที่สอดคล้อง หรือเหมาะสมกับข้อมูลมากกว่า จะมีค่าของตัวสถิติดังกล่าวค่าน้อยกว่า นอกจากนี้ มีตัวสถิติ Cox & Snell R^2 และ ตัวสถิติ Nagelkerke $\widetilde{R}^2$ ซึ่งตัวแบบที่สอดคล้อง หรือเหมาะสมกับข้อมูลมากกว่า จะมีค่าของตัวสถิติดังกล่าวค่ามากกว่า

ค่าวัดที่สำคัญมากสำหรับการประเมินความเหมาะสม หรือวัดความแม่นยำของตัวแบบการถดถอยโลจิสติก คือ เปอร์เซ็นต์ค่าของ Y ที่ได้จากการพยากรณ์ (เป็น 0 หรือ 1) ตรงกับค่า Y จริง (เป็น 0 หรือ 1) ที่ได้จากค่าสังเกต แสดงเป็นตารางการจำแนก (classification table) ได้ดังนี้

Observed	Predicted 0	1	Percentage Correct
0	$c_{0,0}$	$c_{0,1}$	$\dfrac{c_{0,0}}{c_{0,0}+c_{0,1}} \times 100\%$
1	$c_{1,0}$	$c_{1,1}$	$\dfrac{c_{1,1}}{c_{1,0}+c_{1,1}} \times 100\%$

เปอร์เซ็นต์ความถูกต้องรวม (overall percentage correct) เท่ากับ

$$\frac{c_{0,0}+c_{1,1}}{c_{0,0}+c_{0,1}+c_{1,0}+c_{1,1}} \times 100\% \tag{14}$$

การจำแนกค่าของ Y ที่ได้จากการพยากรณ์ ได้ความถี่ $c_{i,j}$ ข้างต้น จะต้องกำหนดจุดตัด (cut value) เป็นค่าอยู่ระหว่าง 0 และ 1 ซึ่งเป็นจุดจำแนกค่า Y หรือพยากรณ์ค่า Y เป็น 0 หรือ 1 เช่น กำหนดจุดตัดเท่ากับ 0.5 เพราะฉะนั้น ถ้าค่าความน่าจะเป็น $\hat{\pi}(\underline{x}_0)$ ณ ปัจจัย $\underline{x}_0$ มีค่ามากกว่าหรือเท่ากับ 0.5 จะให้ $Y = 1$ (พยากรณ์ค่า Y เป็น 1) ณ ปัจจัย $\underline{x}_0$ แต่ถ้าค่าความน่าจะเป็น $\hat{\pi}(\underline{x}_0)$ มีค่าน้อยกว่า 0.5 จะให้ $Y = 0$ (พยากรณ์ค่า Y เป็น 0) ณ ปัจจัย $\underline{x}_0$

จุดตัดอาจจะมีผลต่อเปอร์เซ็นต์ความแม่นยำ ดังนั้น ในทางปฏิบัติ นักวิเคราะห์ควรทดลองจุดตัดที่ค่าต่าง ๆ โดยอาจจะเริ่มจากค่า 0.5 แล้วทดลองเปลี่ยนเป็นค่าอื่น ๆ ที่มากกว่าและน้อยกว่า 0.5 และเลือกจุดตัดที่ให้เปอร์เซ็นต์แม่นยำมากสุดเป็นจุดตัดที่นำไปใช้คู่กับตัวแบบความน่าจะเป็น $\hat{\pi}(\underline{x})$ ตัวแบบที่มีความถูกต้อง หรือแม่นยำสูง ควรมีเปอร์เซ็นต์ความถูกต้องรวมไม่น้อยกว่า 90%

5. กรณีศึกษา

การวิเคราะห์ความเสี่ยงในการให้สินเชื่อเพื่อเช่าซื้อรถยนต์ของหน่วยธุรกิจแห่งหนึ่ง ด้วยตัวแบบการถดถอยโลจิสติก มีจุดประสงค์ที่จะสร้างตัวแบบการถดถอยโลจิสติกเพื่อนำไปใช้สำหรับการประเมินความเสี่ยงการให้สินเชื่อ

ความเสี่ยงที่สำคัญคือ ผู้เช่าซื้อไม่สามารถชำระหนี้ครบตามสัญญา ทำให้เกิดความเสียหาย หรือเกิดค่าใช้จ่ายแก่บริษัท เพราะฉะนั้น การตัดสินใจให้สินเชื่อหรือไม่ จึงเป็นเรื่องสำคัญซึ่งควรมีเกณฑ์ หรือกฎระเบียบประกอบการตัดสินใจ การวิเคราะห์ หรือการประเมินความเสี่ยง

ด้วยตัวแบบการถดถอยโลจิสติก จะเป็นเครื่องมือหนึ่งที่ให้ข้อมูลประกอบการตัดสินใจ

ด้วยตัวแบบการถดถอยโลจิสติก จะได้ค่าความน่าจะเป็น หรือโอกาสที่ผู้เช่าซื้อจะสามารถชำระหนี้ครบตามสัญญา ดังนั้น ถ้าผลออกมามีความน่าจะเป็นสูง ผู้ขอเช่าซื้อควรได้รับการพิจารณาได้สินเชื่อ

ในกรณีศึกษานี้จะทำการศึกษาเฉพาะการให้สินเชื่อเช่าซื้อรถยนต์ญี่ปุ่น เป็นรถยนต์นั่งส่วนบุคคล (7 ที่นั่ง) และ รถกระบะ และเป็นรถใหม่ โดยได้ข้อมูลจากหน่วยธุรกิจให้สินเชื่อแห่งหนึ่ง ในช่วงปี 2544 ถึง 2547 เป็นข้อมูลส่วนบุคคลของผู้ขอสินเชื่อ

ข้อมูล

การประเมินความเสี่ยงด้วยตัวแบบการถดถอยโลจิสติก จะใช้ข้อมูลในอดีตของผู้ขอสินเชื่อเช่าซื้อรถยนต์ทั้งในรายที่ชำระหนี้ได้ครบตามสัญญา และรายที่ไม่สามารถชำระหนี้ได้ครบตามสัญญา การเก็บข้อมูลในการวิเคราะห์ครั้งนี้เป็นข้อมูลทุติยภูมิ ได้ข้อมูลรายละเอียดส่วนบุคคลของผู้ที่มาขอสินเชื่อ โดยข้อมูลนั้นเป็นข้อมูลเกี่ยวกับสถานะความเป็นจริงของผู้ที่มาขอสินเชื่อ ณ วันที่ขอสินเชื่อ ได้แก่ อายุ สถานภาพสมรส อาชีพ รายได้ รายจ่าย ประเภทรถยนต์ ซีซีของรถ ราคารถ เปอร์เซ็นต์เงินดาวน์ ยอดเช่าซื้อรถยนต์ และผลการชำระหนี้ ข้อมูลเหล่านี้เป็นข้อมูลภายใน ซึ่งได้จากใบคำขอสินเชื่อเช่าซื้อรถยนต์

ข้อมูลที่นำมาศึกษา มีจำนวน 201 รายสำหรับสร้างตัวแบบการถดถอยโลจิสติก จำแนกเป็นข้อมูลของผู้ขอสินเชื่อที่แต่ละรายสามารถชำระหนี้ครบถ้วนตามสัญญาจำนวน 98 ราย และเป็นข้อมูลของผู้ขอสินเชื่อที่แต่ละรายไม่สามารถชำระหนี้ครบถ้วนตามสัญญาจำนวน 103 ราย

ตัวแปร และรายละเอียดข้อมูล

จากข้อมูลดังกล่าวข้างต้น ได้ทำการกำหนดตัวแปรสำหรับใช้ในการศึกษาดังนี้

<u>ตัวแปรตาม</u>

Result คือ ผลการชำระหนี้ มีค่าเป็นไปได้ 2 ค่า

0 หมายถึง ไม่สามารถชำระหนี้ได้ครบตามสัญญา

1 หมายถึง สามารถชำระหนี้ได้ครบตามสัญญา

<u>ตัวแปรอิสระ</u>

อาชีพ จำแนกเป็น 3 กลุ่ม คือ เจ้าของกิจการ รับราชการ/รัฐวิสาหกิจ และลูกจ้าง/พนักงาน ให้เป็นตัวแปร dummy โดยให้

Occ_1 = 1, Occ_2 = 0 แทนอาชีพเจ้าของกิจการ

Occ_1 = 0, Occ_2 = 1 แทนอาชีพรับราชการ/รัฐวิสาหกิจ

Occ_1 = 0, Occ_2 = 0 แทนอาชีพลูกจ้าง/พนักงาน

Age แทนอายุ (ปี)

Salary แทนรายได้ (พันบาทต่อเดือน)

Expense แทนรายจ่าย (พันบาทต่อเดือน)

Status แทนสถานภาพสมรส โดยให้

Status = 1 หมายถึง โสด/หม้าย

Status = 0 หมายถึง สมรส

Type แทนประเภทรถยนต์ โดยให้

Type = 1 ถ้าเป็นรถยนต์นั่งส่วนบุคคล

Type = 0 ถ้าเป็นรถยนต์กระบะ

CC แทนซีซีรถยนต์

Price แทนราคารถยนต์ (พันบาท)

Down แทนเปอร์เซ็นต์เงินดาวน์

Princ แทนยอดเช่าซื้อรถยนต์ (พันบาท)

ตัวอย่างข้อมูลที่นำมาวิเคราะห์สร้างตัวแบบการถดถอยโลจิสติก

no.	Occ_1	Occ_2	Age	Salary	Expense	Status	Type	cc	Price	Down	Princ	Result
1	0	0	45	13	4.29	1	0	2999	587	25	440	1
2	0	0	38	23.37	7.71	1	0	2499	517	25	387.7	1
3	0	1	60	23.26	11.16	0	1	1493	567.9	55	267.9	0
4	0	0	38	41	14.35	1	0	2892	476	25	357	1
5	0	0	29	25	8	1	1	1498	460	30	319	1
6	0	1	36	9.82	4.62	0	1	1498	453.9	45	259	0
7	0	1	46	100	49	0	1	2198	1289	30	902.3	0

no.	Occ_1	Occ_2	Age	Salary	Expense	Status	Type	cc	Price	Down	Princ	Result
8	0	0	32	13.4	4.56	1	0	2499	521	25	390	1
9	0	1	34	17.18	8.25	0	1	2999	748	30	527	0
10	0	1	46	16.8	8.06	0	1	1668	870	35	550	0
11	0	1	45	16.41	7.55	0	0	2494	481	25	360	0
...	...	...	...	...	...	...	...	...	...	...	...	...
112	0	0	43	30	10.5	1	1	1973	1105	25	828.75	1
113	0	0	33	8.29	2.74	1	1	1498	472	25	354	1
114	0	1	34	9.04	3.07	1	1	1668	729	25	546	1
115	1	0	28	8	2.56	1	0	2663	515	25	386.25	1
116	0	1	40	9.27	4.82	0	0	2499	497	45	267	0
117	0	0	35	4.1	1.23	1	1	1493	514.9	25	386	1
...	...	...	...	...	...	...	...	...	...	...	...	...

ผลการวิเคราะห์

ในการวิเคราะห์ข้อมูลหาตัวแบบการถดถอยโลจิสติก ใช้โปรแกรมสำเร็จรูป SPSS ในส่วน Analyze โดยเลือก Regression และเลือก Binary Logistic ในการใช้โปรแกรม SPSS

ผลลัพธ์ของการวิเคราะห์สรุปได้ดังนี้

ตารางที่ 1 ตัวแปร และสัมประสิทธิ์ในตัวแบบการถดถอยโลจิสติก:

```
---------------------- Variables in the Equation -------------------

Variable        B      S.E.      Wald     df     Sig      Exp(B)

SALARY     1.9643     .3466    32.1274     1    .0000     7.1299
EXPENSE   -3.9467     .6948    32.2697     1    .0000      .0193
STATUS     3.9662    1.4666     7.3132     1    .0068    52.7862
DOWN        .0763     .0360     4.5006     1    .0339     1.0793
Constant  -6.9951    1.7471    16.0308     1    .0001
---------------------------------------------------------------------
```

ผลการคัดเลือกตัวแปรโดยการพิจารณานัยสำคัญของพารามิเตอร์ β_i ได้ว่า มีตัวแปรหรือปัจจัยที่มีผลต่อค่าตัวแปรตาม (ผลการชำระหนี้) คือ รายได้ (salary) รายจ่าย (expense) สถานภาพสมรส (status) และเปอร์เซ็นต์เงินดาวน์ (down) จากนี้จะวินิจฉัยตัวแบบ เพื่อตรวจสอบความเหมาะสม หรือความพอเพียงเชิงสถิติของตัวแบบ

ค่าสถิติวัดความกลมกลืน หรือความพอเพียงเชิงสถิติของตัวแบบ:

```
Cox & Snell - R^2    .666
Nagelkerke - R^2     .888
```

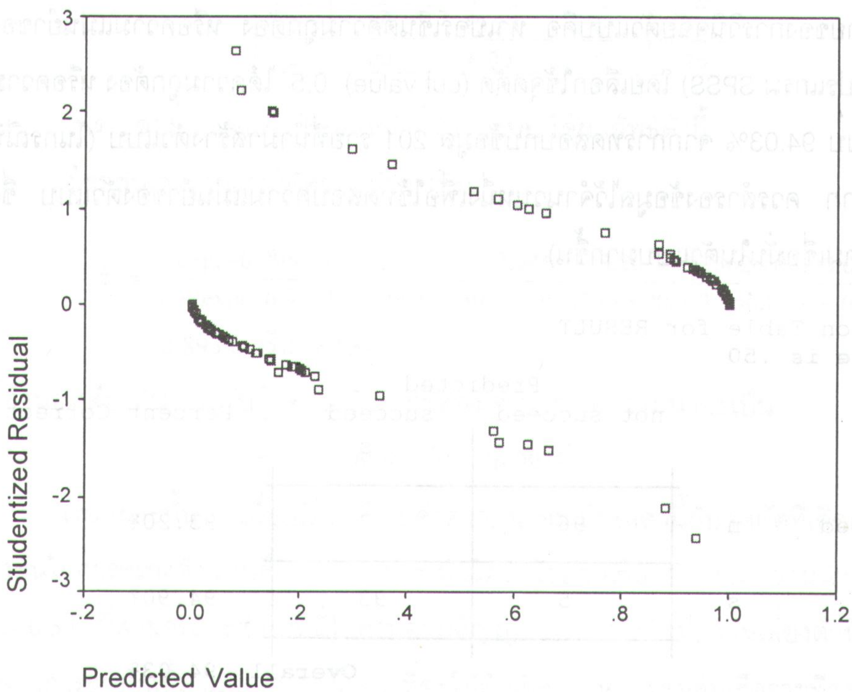
วินิจฉัยตัวแบบ:

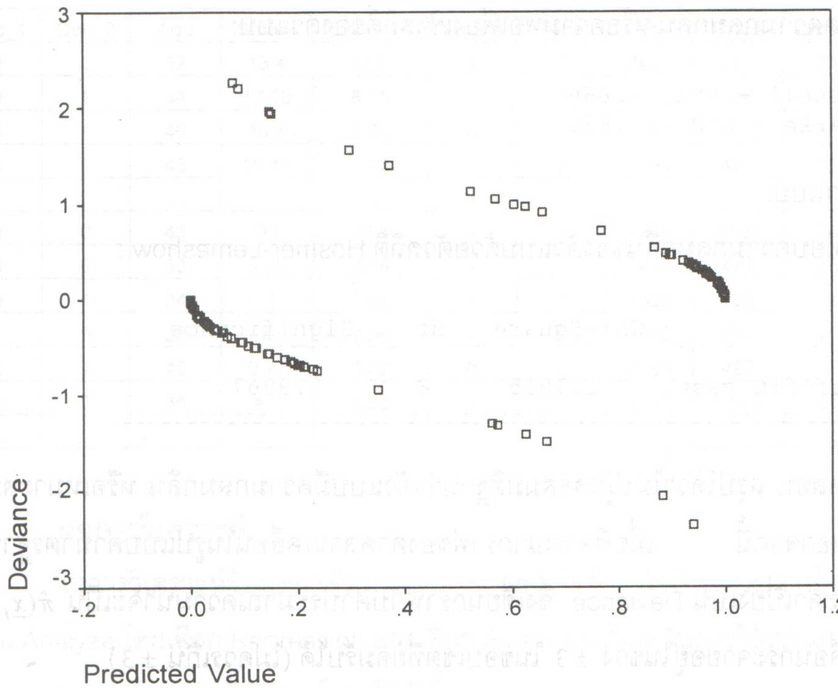
ทดสอบเทียบความกลมกลืนของตัวแบบด้วยตัวสถิติ Hosmer-Lemeshow :

```
                        Chi-Square    df    Significance

Goodness-of-fit test      1.7935       8       .9867
---------------------------------------------------------------
```

ผลการทดสอบ สรุปได้ว่าไม่ปฏิเสธสมมติฐานว่าตัวแบบมีความกลมกลืน หรือเหมาะสมกับข้อมูล นอกจากนี้ เมื่อพิจารณากราฟของค่าคลานเคลื่อนในรูปแบบค่ามาตรฐาน Studentized และค่าเบี่ยงเบน Deviance ซึ่งเขียนกราฟกับค่าประมาณความน่าจะเป็น $\hat{\pi}(\underline{x}_i)$ พบว่าค่าคลาดเคลื่อนกระจายอยู่ในช่วง ± 3 ในขอบเขตที่ยอมรับได้ (ไม่ควรเกิน ± 3)





ขั้นสุดท้ายของการวินิจฉัยตัวแบบคือ หาเปอร์เซ็นต์ความถูกต้อง หรือความแม่นยำของตัวแบบ (โดยใช้โปรแกรม SPSS) โดยเลือกใช้จุดตัด (cut value) 0.5 ได้ความถูกต้อง หรือความแม่นยำของตัวแบบ 94.03% จากการทดสอบกับข้อมูล 201 รายที่นำมาสร้างตัวแบบ (ในกรณีที่มีข้อมูลจำนวนมาก ควรสำรองข้อมูลไว้จำนวนหนึ่งเพื่อใช้ทดสอบความแม่นยำของตัวแบบ ซึ่งจะช่วยให้เกิดความเชื่อมั่นในตัวแบบมากขึ้น)

Classification Table for RESULT
The Cut Value is .50

		Predicted not succeed n	Predicted succeed s	Percent Correct
Observed				
not succeed	n	96	7	93.20%
succeed	s	5	93	94.90%
			Overall	**94.03%**

เพราะฉะนั้น ในการประเมินความเสี่ยงของการให้สินเชื่อเช่าซื้อรถยนต์ด้วยตัวแบบการถดถอยโลจิสติก สำหรับสถานประกอบการที่นำมาเป็นกรณีศึกษานี้ ได้ตัวแบบประมาณความน่าจะเป็นที่ผู้ขอสินเชื่อจะสามารถชำระหนี้ได้ครบตามสัญญา ดังนี้

$$\hat{\pi} = \frac{\exp(-6.9951 + 1.9643Salary - 3.9467Expense + 3.9662Status + 0.0763Down)}{1 + \exp(-6.9951 + 1.9643Salary - 3.9467Expense + 3.9662Status + 0.0763Down)}$$

หรือในรูปแบบของ "ตัวแบบล็อกของอัตราส่วนความน่าจะเป็น" (log odds ratio model) หรือ "ตัวแบบโลจิท" (logit model):

$$\ln(\hat{OR}_{\underline{x}}) = \ln\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$$

$$= -6.9951 + 1.9643Salary - 3.9467Expense + 3.9662Status + 0.0763Down$$

จากตัวแบบความน่าจะเป็น $\hat{\pi}$ นำไปใช้ประโยชน์ได้ดังนี้

สมมติมีลูกค้ามาขอสินเชื่อสองราย โดยมีข้อมูลที่เป็นปัจจัยในตัวแบบคือ

no.	Salary	Expense	Status	Down
1	30	13.20	0	30
2	20	8.80	0	25

เพราะฉะนั้น แทนค่าปัจจัยเหล่านี้ในตัวแบบ ได้ผลลัพธ์ดังนี้

ผู้ขอสินเชื่อรายแรกได้ความน่าจะเป็น

$$\hat{\pi} = \frac{\exp(-6.9951 + 1.9643\times30 - 3.9467\times13.20 + 3.9662\times0 + 0.0763\times30)}{1 + \exp(-6.9951 + 1.9643\times30 - 3.9467\times13.20 + 3.9662\times0 + 0.0763\times30)}$$

$$= 0.8934 \text{ หรือ } 89.34\%$$

ในทำนองเดียวกัน สำหรับผู้ขอสินเชื่อรายที่สอง คำนวณได้ความน่าจะเป็น

$$\hat{\pi} = 0.3699 \text{ หรือ } 36.99\%$$

เพราะฉะนั้น เมื่อเปรียบเทียบค่าความน่าจะเป็นเหล่านี้กับจุดตัดที่เลือกกำหนดไว้แล้วข้างต้นในการสร้างตัวแบบคือ 0.5 ประเมินได้ว่าผู้ขอสินเชื่อรายแรกมีความน่าจะเป็นสูง (ค่า $\hat{\pi} > 0.5$) ที่สามารถจะชำระหนี้ได้ครบตามสัญญา (หรือบริษัทมีความเสี่ยงต่ำที่จะสูญเสียรายได้จากการชำระหนี้ตามสัญญา) จึงควรที่จะให้สินเชื่อ สำหรับผู้ขอสินเชื่อรายที่สองมีความน่าจะ

เป็นต่ำ (ค่า $\hat{\pi} < 0.5$) ที่สามารถจะชำระหนี้ได้ครบตามสัญญา (หรือบริษัทมีความเสี่ยงสูงที่จะสูญเสียรายได้ หรือมีความเสี่ยงสูงที่จะเสียค่าใช้จ่ายเนื่องจากการชำระหนี้ไม่ครบตามสัญญา) จึงไม่ควรที่จะให้สินเชื่อ

ข้อมูลอื่นที่ได้จากตัวแบบ $\hat{\pi}$ ที่อาจจะเป็นประโยชน์ เช่น ผลกระทบของปัจจัย

จากสัมประสิทธิ์ของปัจจัยหรือตัวแปรอิสระในตัวแบบ พบว่าสัมประสิทธ์ของรายได้ [(salary) เท่ากับ 1.9643] เป็นค่าบวก แสดงถึงผลกระทบเชิงบวกคือ รายได้ของผู้ขอสินเชื่อยิ่งมากขึ้น (ขณะที่ปัจจัยอื่น ๆ คงที่ ค่าเดียวกัน) ความน่าจะเป็นที่ผู้ขอสินเชื่อจะชำระหนี้ได้ครบตามสัญญาจะมีมากขึ้น ส่วนปัจจัยรายจ่าย พบว่าสัมประสิทธ์ของรายจ่าย [(expense) เท่ากับ -3.9467] เป็นค่าลบ แสดงถึงผลกระทบเชิงลบคือ รายจ่ายของผู้ขอสินเชื่อยิ่งมากขึ้น (ขณะที่ปัจจัยอื่น ๆ คงที่ ค่าเดียวกัน) ความน่าจะเป็นที่ผู้ขอสินเชื่อจะชำระหนี้ได้ครบตามสัญญาจะลดลง ในทำนองเดียวกัน สำหรับสถานภาพสมรส ขณะที่ปัจจัยอื่น ๆ คงที่ ค่าเดียวกัน พบว่าคนโสดมีโอกาสหรือมีความน่าจะเป็นมากกว่าคนไม่โสดที่จะชำระหนี้ได้ครบ และสำหรับเงินดาวน์ ถ้าจ่ายมากขึ้น จะมีโอกาสมากขึ้นที่จะชำระหนี้ได้ครบ (ขณะที่ปัจจัยอื่น ๆ คงที่ ค่าเดียวกัน)

นอกจากนี้ มีอัตราส่วนความน่าจะเป็น (odds ratio) เป็นอัตราส่วนระหว่างความน่าจะเป็นที่จะชำระหนี้ครบ และความน่าจะเป็นที่จะชำระหนี้ไม่ครบ ตัวอย่าง เช่น สำหรับผู้ขอสินเชื่อรายที่หนึ่งในตัวอย่างสมมติข้างต้น ได้อัตราส่วนความน่าจะเป็น

$$\frac{\hat{\pi}}{1-\hat{\pi}} = \frac{0.8934}{1-0.8934} = 8.381$$

แสดงว่าผู้ขอสินเชื่อตามข้อมูลข้างต้นมีโอกาสหรือมีความน่าจะเป็นที่จะชำระหนี้ครบ เป็น 8.381 เท่าของโอกาสหรือความน่าจะเป็นที่จะชำระหนี้ไม่ครบ นอกจากนี้ ถ้าสถานภาพสมรส (status) เปลี่ยนจาก 0 คือสมรส เป็น 1 คือโสด/หม้าย อัตราส่วนความน่าจะเป็น จะมีค่าเปลี่ยนเป็นทวีคูณ คือจะมีค่าเท่ากับ $\exp(3.9662) = e^{3.9662} = 52.78$ เท่าของค่าเดิม (ค่านี้มีในผลลัพธ์ตารางที่ 1 จากโปรแกรม SPSS) และในกรณีนี้มีค่าเพิ่มขึ้น (เนื่องจากค่ามากกว่า 1 หรือสัมประสิทธิ์ของตัวแปรสถานภาพคือ 3.9662 เป็นค่าบวก) ซึ่งเพิ่มขึ้น $(e^{3.9662}-1)100\% = 5178.4\%$ นั่นคือ อัตราส่วนความน่าจะเป็นมีค่าเปลี่ยนเป็น $52.784 \times 8.381 = 442.4$ โดยประมาณ ดังนั้น สำหรับสถานภาพโสด (ขณะที่ปัจจัยหรือตัวแปรอื่น ๆ มีค่าเดียวกัน รายได้ = 30 รายจ่าย = 13.20 และเปอร์เซ็นต์เงินดาวน์ = 30) โอกาสหรือความน่าจะเป็นชำระหนี้ได้ครบเพิ่ม

ขึ้นเป็นหลายเท่าของโอกาสชำระหนี้ไม่ครบ สำหรับผลกระทบของปัจจัยอื่น ๆ อธิบายได้ในทำนองเดียวกัน

6. บรรณานุกรม

Abraham, B. and Ledolter, J. *Introduction to Regression Modeling*. CA: Thomson Brooks/Cole, 2006.

Hosmer, D. W. and Lemeshow, S. *Applied Logistic Regression* (2nd ed.). New York: Wiley, 2000.

Montgomery, D. C., Peck, E. A., and Vining, G. G. *Introduction to Linear Regression Analysis* (3rd ed.). New York: Wiley, 2001.

Pregibon, D. Logistic regression diagnostics. *Annals of Statistics*, 9, 705-724, 1981.

Ryan, T. P. Modern Regression Methods. New York: Wiley, 1997.