

## การเปรียบเทียบประสิทธิภาพระหว่างตัวแบบถดถอยโลจิสติกแบบสถิตและแบบพลวัต

### Comparison Efficiency between a Static and a Dynamic Logistic Regression Model

วรุณี มุริกา

#### บทคัดย่อ

วัตถุประสงค์ของการวิจัยครั้งนี้ เพื่อเปรียบเทียบประสิทธิภาพการพยากรณ์ระหว่างตัวแบบถดถอยโลจิสติกแบบสถิตและแบบพลวัต โดยตัวแบบสถิตจะพิจารณาข้อมูลเป็นเพียงหนึ่งช่วงเวลาเท่านั้น แต่ตัวแบบพลวัตจะพิจารณาข้อมูลที่เปลี่ยนไปในหลายช่วงเวลาย่อย ในรูปแบบของ Hazard model โดยศึกษาทั้งข้อมูลจำลองและข้อมูลจริง ซึ่งข้อมูลจำลองจะใช้ 3 ตัวแปรอิสระที่มีการแจกแจงแตกต่างกัน แบ่งเป็น 24 ช่วงเวลา สำหรับข้อมูลจริงเป็นข้อมูลทุติยภูมิเกี่ยวกับการเช่าซื้อสินค้าชนิดหนึ่ง โดยสนใจการเกิดหนี้ NPL (Non Performing Loan) แบ่งเป็น 4 ช่วงเวลา ผลการศึกษา สำหรับข้อมูลจำลอง ตัวแบบพลวัตมีประสิทธิภาพดีกว่าตัวแบบสถิต เนื่องจากพื้นที่ใต้โค้ง ROC ของตัวแบบพลวัตมากกว่าตัวแบบสถิตอย่างมีนัยสำคัญทางสถิติที่ระดับนัยสำคัญ 0.05 และสำหรับข้อมูลจริง เราไม่สามารถสรุปได้ว่า ตัวแบบพลวัตมีประสิทธิภาพดีกว่าตัวแบบสถิต ที่ระดับนัยสำคัญ 0.05

คำสำคัญ : ฟังก์ชันอัตราภาวะภัย, การวิเคราะห์ความถดถอยโลจิสติก, ตัวแบบพยากรณ์, เส้นโค้ง ROC

#### Abstract

The objective of this research is to compare the prediction power between two logistic regression models: static model and dynamic model. A static model considers only one period of explanatory variable values. A dynamic model, or a hazard model consider data of explanatory variables in several periods for each sample unit. We perform the comparisons on simulated data and real data. The simulated data, the model assumes 3 independent variables with different distributions. The simulation is divided into 24 periods. The real data is a secondary data from a credit activity. The interested event is NPL of customers and divided into 4 periods. The conclusions of this study are: for simulated data, the dynamic model is better than the static model at significance level of 0.05. For real data, the dynamic model is not different from the static model at significance level of 0.05.

**Key words:** Hazard function, Logistic regression model, Credit scoring model, ROC curve

## 1. บทนำ

การวิเคราะห์ความถดถอยโลจิสติก (Logistic Regression Analysis) เป็นที่นิยมมากขึ้นในปัจจุบัน ซึ่งเป็นการหาตัวแบบประมาณเพื่อนำมาพยากรณ์โอกาสของการเกิดเหตุการณ์ที่สนใจ สำหรับการวิเคราะห์สามารถทำได้ 2 แบบ คือ แบบสถิติหรือแบบปกติที่ใช้กันทั่วไป และแบบพลวัต เป็นการสร้างตัวแบบที่ได้มีการแบ่งเวลาออกเป็นหลายช่วงและพิจารณาข้อมูลของหน่วยตัวอย่างในแต่ละช่วงเวลาเหล่านั้น ซึ่งมาจากแนวคิดที่ว่าลักษณะของหน่วยตัวอย่างมีการเปลี่ยนแปลงไปตามเวลาในลักษณะของตัวแบบอัตราภาวะภัย (Hazard Model)

โดยส่วนใหญ่ นักวิจัยมักใช้การวิเคราะห์ความถดถอยโลจิสติกแบบสถิติเพราะเป็นวิธีที่ง่าย และให้ผลการพยากรณ์ที่น่าเชื่อถือได้ดีพอสมควร แต่สำหรับเหตุการณ์ที่เกิดขึ้นได้ยาก การพยากรณ์ต้องพิจารณาข้อมูลเป็นระยะเวลาหลายปี การวิเคราะห์แบบสถิติจะละเลยความจริงที่ลักษณะของหน่วยตัวอย่างมีการเปลี่ยนแปลงปีต่อปี แต่จะสนใจเพียงข้อมูลก่อนเกิดเหตุการณ์ โดยไม่ได้สนใจข้อมูลของตัวอย่างในช่วงที่ยังปลอดภัยอยู่ ซึ่งจะทำให้การประมาณค่าไม่ถูกต้อง มีความเอนเอียงจากการเลือกช่วงเวลาได้ จากงานวิจัยของ Tyler Shumway [1] ได้ศึกษาการวิเคราะห์ความถดถอยโลจิสติกกับข้อมูลตลาดหุ้นของกลุ่มอุตสาหกรรมในนิวยอร์กและอเมริกา ปี ค.ศ.1962 -1992 โดยกำหนดให้แต่ละช่วงเวลาเท่ากับ 1 ปี ซึ่งโดยธรรมชาติแล้วการเกิดเหตุการณ์ล้มละลายของบริษัทใหญ่จะเกิดขึ้นได้ไม่บ่อยนักและการพยากรณ์ต้องพิจารณาข้อมูลในหลายปี ผลการศึกษา พบว่าตัวแบบพลวัตให้เปอร์เซ็นต์ความถูกต้องของการพยากรณ์สูงกว่าตัวแบบสถิติ และจากงานวิจัยของ Sunti Tirapat and Seksan Kiatsupaibul [2] ได้ศึกษาการพยากรณ์มูลค่าความเสี่ยง (Credit VaR) ซึ่งเป็นมูลค่าความเสียหายสูงสุดที่คาดว่าจะเกิดขึ้นจากการผิวนัดชำระหนี้ของลูกค้า ภายใต้ระดับความเชื่อมั่นหนึ่งที่นักลงทุนยอมรับได้ โดยใช้ตัวแบบถดถอยโลจิสติกแบบสถิติ พบว่า ตัวแบบประมาณที่ได้ให้มูลค่าความเสี่ยงของการประมาณต่ำเกินไป ซึ่งทำให้ความถูกต้องและความน่าเชื่อถือของตัวแบบลดน้อยลง

จากการศึกษางานวิจัยดังที่กล่าวมา ทำให้ผู้วิจัยสนใจจะศึกษาและเปรียบเทียบประสิทธิภาพระหว่างตัวแบบถดถอย

โลจิสติกแบบสถิติและแบบพลวัต กับข้อมูลระดับหน่วยย่อย (Micro unit) ซึ่งจะแตกต่างจากงานวิจัยของ Tyler Shumway [1] ที่ศึกษากับข้อมูลระดับหน่วยใหญ่ (Macro unit) ซึ่งความซับซ้อนของข้อมูลหน่วยย่อยจะมีมากกว่าข้อมูลหน่วยใหญ่ และเกิดเหตุการณ์ที่สนใจได้ง่ายกว่า ไม่ต้องใช้เวลานานนัก โดยในการเปรียบเทียบจะพิจารณาจากพื้นที่ใต้โค้ง ROC ของการพยากรณ์

## 2. ทฤษฎี

ข้อมูลสำหรับการวิเคราะห์ความถดถอยโลจิสติกจะอยู่ในรูป  $(y_i, x_i)$  ;  $i = 1, 2, \dots, n$  เมื่อตัวแปรตาม  $y_i$  มีการแจกแจงแบบเบอร์นูลลีด้วยความน่าจะเป็นของการเกิดเหตุการณ์ที่สนใจ  $p_i = P(Y_i = 1 | x_i)$   $x_i$  แทน  $x_{1i}, x_{2i}, \dots, x_{pi}$  เป็นตัวแปรอิสระร่วมกัน  $p$  ตัวแปร และข้อมูลตัวอย่าง  $n$  หน่วยเป็นอิสระต่อกัน นั่นคือ

$$y_i = \begin{cases} 1 & ; \text{เกิดเหตุการณ์ที่สนใจ ด้วย } p_i \\ 0 & ; \text{ไม่เกิดเหตุการณ์ที่สนใจ ด้วย } 1-p_i \end{cases}$$

ตัวแบบถดถอยโลจิสติกอยู่ในรูปแบบดังนี้

$$E[Y | X] = \frac{e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}}}{1 + e^{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}}} = \pi(x_i) = p_i \quad (1)$$

และตัวแบบโลจิทมีรูปแบบดังนี้

$$\begin{aligned} \logit(\pi(x_i)) &= \ln\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right) \\ &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} \end{aligned} \quad (2)$$

การประมาณค่าพารามิเตอร์  $\beta$  ในตัวแบบถดถอยโลจิสติกจะใช้วิธีภาวะน่าจะเป็นสูงสุด โดยการหาค่าประมาณที่ทำให้ลอการิทึมของฟังก์ชันภาวะน่าจะเป็น  $L(\beta)$  มีค่ามากที่สุด ซึ่งเรียกตัวประมาณนี้ว่า “ตัวประมาณภาวะน่าจะเป็นสูงสุด” [3]

### 2.1 การวิเคราะห์ความถดถอยโลจิสติกแบบสถิติ

เป็นการพิจารณาข้อมูลของแต่ละหน่วยตัวอย่างเพียง 1 ช่วงเวลา ซึ่งจะได้จำนวนข้อมูลทั้งหมดเท่ากับ  $n$  ชุดสำหรับ  $n$  หน่วยตัวอย่าง โดยมี ฟังก์ชันภาวะน่าจะเป็นและลอการิทึมของฟังก์ชันภาวะน่าจะเป็น ดังนี้

$$l(\beta) = \prod_{i=1}^n (\pi(x_i))^{y_i} (1 - \pi(x_i))^{1-y_i} \quad (3)$$

$$L(\beta) = \ln l(\beta) = \sum_{i=1}^n \left\{ y_i \ln \left( \frac{\pi(x_i)}{1 - \pi(x_i)} \right) + \ln(1 - \pi(x_i)) \right\} \quad (4)$$

## 2.2 การวิเคราะห์ความถดถอยโลจิสติกแบบพลวัต

จะอาศัยฟังก์ชันอัตราภาวะภัย (Hazard Function) ซึ่งจะทำให้การแบ่งเวลาออกเป็นหลายช่วงเวลาย่อย และเก็บค่าของตัวแปรต่างๆ ที่เปลี่ยนแปลงไปในช่วงเวลาย่อยนี้ นั่นคือ แต่ละหน่วยตัวอย่างจะมีข้อมูลของตัวแปรที่เกี่ยวข้องสำหรับหลายช่วงเวลา คือ  $(x_{1t}, x_{2t}, \dots, x_{pt}, y_t); t = 1, 2, \dots$  การวิจัยครั้งนี้จะศึกษากรณีที่ช่วงเวลาเป็นแบบไม่ต่อเนื่อง ฟังก์ชันภาวะน่าจะเป็นและลอการิทึมของฟังก์ชันภาวะน่าจะเป็น มีรูปแบบดังนี้

$$\begin{aligned} l(\beta) &= \prod_{i=1}^n \left( F(t_i, x_i; \beta)^{y_i} \prod_{j < t_i} [1 - F(j, x_j; \beta)] \right) \\ &= \prod_{i=1}^n \left( \lambda(t_i, x_i; \beta)^{y_i} \prod_{j < t_i} [1 - \lambda(j, x_j; \beta)] \right) \\ &= \prod_{i=1}^n \lambda(t_i, x_i; \beta)^{y_i} S(t_i, x_i; \beta) \end{aligned} \quad (5)$$

$$\begin{aligned} L(\beta) &= \ln l(\beta) \\ &= \ln \left( \prod_{i=1}^n \lambda(t_i, x_i; \beta)^{y_i} S(t_i, x_i; \beta) \right) \end{aligned} \quad (6)$$

การประมาณค่าพารามิเตอร์ในตัวแบบพลวัตจะทำได้วิธีภาวะน่าจะเป็นสูงสุดเช่นเดียวกับตัวแบบสถิต แต่เนื่องจากจำนวน  $n$  ที่ใช้ในตัวแบบพลวัตเป็นจำนวนข้อมูลของหน่วยตัวอย่างในหลายช่วงเวลาหรือมีค่าสังเกตหลายค่าในหน่วยตัวอย่างเดียว ซึ่งทำให้ข้อมูลมีความบกพร่องในคุณสมบัติของความไม่เป็นอิสระต่อกัน ดังนั้น เราจำเป็นต้องทำการปรับแก้ค่าสถิติ  $\chi^2$  โดยการหารด้วยค่าเฉลี่ยของจำนวนตัวอย่างที่ใช้ในตัวแบบพลวัตต่อจำนวนหน่วยตัวอย่างจริง [1]

### 2.1.1 ฟังก์ชันการอยู่รอด (Survival Function)

เป็นความน่าจะเป็นที่ค่าสังเกตจะอยู่รอดต่อไปจากช่วงเวลา  $t$  โดยที่อยู่รอดมาแล้วจนถึงเวลา  $t$  นั่นคือ [4]

$$S(t, x; \beta) = P(T \geq t) = 1 - P(T < t) \quad (7)$$

### 2.1.2 ฟังก์ชันอัตราภาวะภัย (Hazard function)

เป็นลิมิตของความน่าจะเป็นที่ค่าสังเกตจะสูญเสียในช่วงเวลาสั้นๆ จาก  $t$  ถึง  $t + \Delta t$  เมื่อค่าสังเกตมีการอยู่รอดถึงเวลา  $t$  นั่นคือ [4]

$$\lambda(t, x; \beta) = \lim_{\Delta t \rightarrow 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t} \quad (8)$$

ความสัมพันธ์ระหว่างฟังก์ชันการอยู่รอดและฟังก์ชันอัตราภาวะภัย เป็นดังนี้

$$\lambda(t, x; \beta) = \frac{f(t, x; \beta)}{S(t, x; \beta)} = -\frac{S'(t, x; \beta)}{S(t, x; \beta)} \quad (9)$$

### 2.1.3 การประมาณค่า $\lambda(t, x; \beta)$ และ $S(t, x; \beta)$

สำหรับแต่ละช่วงเวลาที่ไม่ต่อเนื่อง สามารถคำนวณได้จากสมการดังต่อไปนี้ [5]

$$\hat{\lambda}(t, x; \beta) = \frac{h_t}{n_t} \quad (10)$$

$$\hat{S}(t, x; \beta) = \prod_{i=1}^t \left( \frac{n_i - h_i}{n_i} \right) = \prod_{i=1}^t [1 - \hat{\lambda}(t, x; \beta)] \quad (11)$$

เมื่อ  $h_t$  คือ จำนวนตัวอย่างที่เกิดเหตุการณ์ที่สนใจในช่วงเวลา  $t$

$n_t$  คือ จำนวนตัวอย่างที่ถูกสังเกตทั้งหมดในช่วงเวลา  $t$

## 2.2 โคลง ROC และ พื้นที่ใต้โคลง ROC

เป็นกราฟที่พล็อตระหว่างสัดส่วนของการพยากรณ์ได้ถูกต้องของการเกิดเหตุการณ์ (Sensitivity) และสัดส่วนของการพยากรณ์ผิดของการเกิดเหตุการณ์ (1-Specificity) พื้นที่ใต้โคลง ROC เป็นค่าที่แสดงถึงความเชื่อถือได้ของตัวแบบ หรือความถูกต้องของการพยากรณ์ ซึ่งมีพิสัยอยู่ระหว่าง 0 ถึง 1 และมีเกณฑ์ทั่วไปดังนี้ [6]

$0.5 \leq ROC < 0.6$  แสดงว่าเชื่อถือไม่ได้

$0.6 \leq ROC < 0.7$  แสดงว่าเชื่อถือได้น้อย

$0.7 \leq ROC < 0.8$  แสดงว่าเชื่อถือสามารถยอมรับได้

$0.8 \leq ROC < 0.9$  แสดงว่าเชื่อถือได้ในระดับดี

$ROC \geq 0.9$  แสดงว่าเชื่อถือได้ในระดับดีมาก

### 3. วิธีการดำเนินงานวิจัย

#### 3.1 กรณีข้อมูลจำลอง

1) จำนวนตัวแปรอิสระ ( $X$ ) ที่ใช้ในแต่ละตัวแบบเท่ากับ 3 ตัว และมีการแจกแจงที่แตกต่างกัน ดังนี้

$X_1$  มีการแจกแจงแบบปกติด้วย  $\mu = 0$  และ  $\sigma^2 = 1$

$X_2$  มีการแจกแจงแบบเบอร์นูลลีด้วย  $p = 1/2$

$X_3$  มีการแจกแจงแบบเอกซ์โพเนนเชียลด้วย  $\lambda = 1/36$

โดยปรับค่า  $x_3$  เป็นจำนวนเต็ม โดยที่  $x_3 > 0$

2)  $x_1$  และ  $x_2$  เป็นค่าคงที่ตลอดช่วงการศึกษา ส่วน  $x_3$  มีค่าลดลงเรื่อยๆ ในช่วงเวลาถัดไป และตัวแปรอิสระแต่ละตัวไม่มีความสัมพันธ์กัน

3) ตัวแปรตาม ( $Y$ ) มีการแจกแจงแบบเบอร์นูลลีด้วย  $p = \pi(x)$  โดยกำหนดให้

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3}}$$

การคำนวณค่า  $y$  ในแต่ละช่วงเวลา ณ จุดสิ้นสุดของช่วงเวลานั้น และตัวอย่างใดที่มี  $y = 1$  หรือ  $x_3 = 0$  ในเวลา  $t$  จะหยุดการสังเกตหน่วยนั้นในช่วงเวลาถัดไป

4) กำหนดค่า  $\beta$  ใน  $\pi(x)$  กระทำโดยให้  $\bar{y} \approx 0.5$  เมื่อกำหนดให้  $\beta_1 = 1, \beta_2 = 1, \beta_3 = 1/12$  แล้วจึงคำนวณค่า  $\beta_0$  ได้  $\beta_0 = -4.2$

5) แบ่งช่วงเวลาเท่ากับ 24 ช่วงเวลา และใช้ขนาดตัวอย่างเท่ากับ 10,000 โดยแต่ละหน่วยตัวอย่างเป็นอิสระต่อกัน

6) จำนวนการกระทำซ้ำเท่ากับ 1,000 ครั้ง

7) วิเคราะห์ผลด้วยโปรแกรม R ซึ่งเป็นโปรแกรมเปิดรหัส (open source) สำหรับการคำนวณทางสถิติ

#### ขั้นตอนการวิเคราะห์ข้อมูลจำลอง

1) จำลองข้อมูลของตัวแปรต่างๆ ที่เกี่ยวข้องตามที่กำหนด

2) จัดข้อมูลให้เหมาะสมสำหรับการวิเคราะห์ความถดถอยโลจิสติกทั้งสองแบบ

3) แบ่งข้อมูลออกเป็น 2 ชุด ได้แก่ข้อมูลสำหรับการประมาณค่าพารามิเตอร์ และข้อมูลสำหรับการพยากรณ์

4) วิเคราะห์ความถดถอยโลจิสติก จะได้ 2 ตัวแบบ

ประมาณ คือ ตัวแบบสถิต และตัวแบบพลวัต

5) ทำการพยากรณ์ด้วยตัวแบบประมาณทั้งสอง

6) พล็อตโค้ง ROC และคำนวณพื้นที่ใต้โค้ง ROC

7) กระทำซ้ำจนครบตามจำนวนที่กำหนด คือ 1,000 ครั้ง

8) เปรียบเทียบประสิทธิภาพระหว่างตัวแบบทั้งสองจากพื้นที่ใต้โค้ง ROC โดยใช้สถิติทดสอบ T-test

#### 3.2 กรณีข้อมูลจริง

1) เป็นข้อมูลตัวอย่างที่สุ่มจากข้อมูลการเข้าซื้อสินค้าชนิดหนึ่งของลูกค้าประเภทบุคคลที่ทำสัญญาเช่าซื้อในระหว่างเดือนพฤษภาคม 2543 -เมษายน 2549 จำนวน 10,000 ตัวอย่าง ซึ่งลักษณะของการเช่าซื้อ คือ การที่ลูกค้ารับสินค้าไปใช้โดยการจ่ายเงินบางส่วนจำนวนหนึ่ง และส่วนที่เหลือจะแบ่งจ่ายในแต่ละเดือนพร้อมดอกเบี้ยด้วยจำนวนเงินที่เท่าๆ กัน จนครบตามจำนวนและเวลาที่กำหนด แต่ในบางครั้งอาจเกิดปัญหาลูกค้าไม่ชำระหนี้ตามกำหนดในบางเดือนหรืออาจไม่ชำระหนี้เลย ซึ่งจะทำให้เกิดเป็นหนี้ NPL ขึ้น

2) สุ่มข้อมูลตัวอย่าง 2 ชุดที่เป็นอิสระต่อกัน โดยสุ่มจากลูกค้าที่เริ่มทำสัญญาเช่าซื้อในช่วงเวลาที่ต่างกัน ชุดละ 5,000 ราย คือ ชุดที่ 1 ข้อมูลสำหรับการประมาณค่าพารามิเตอร์

ชุดที่ 2 เป็นข้อมูลสำหรับทดสอบการพยากรณ์

3) เก็บข้อมูลลูกค้าเป็นเวลา 1 ปี นับจากวันเริ่มทำสัญญาเช่าซื้อของลูกค้าแต่ละราย โดยแบ่งเป็น 4 ช่วงๆ ละ 3 เดือน

4) ตัวแปรอิสระที่ใช้ในแต่ละตัวแบบ ได้แก่

- PAY เป็นจำนวนเงินที่ต้องจ่ายต่อเดือน (Payment)
- LTV เป็นสัดส่วนของเงินที่กู้ต่อมูลค่าของสินค้า
- RTERM เป็นจำนวนเดือนที่เหลืออยู่ (Remaining Term)

เนื่องจากถ้าจำนวนเงินที่จ่ายต่อเดือนมากจะทำให้ยอดเงินหมดในระยะเวลาสั้นลง และถ้าจำนวนงวดที่เหลือน้อยอาจจะทำให้มีช่วงเวลาที่เสี่ยงต่อการเกิดหนี้ NPL น้อยลงได้ จึงเป็นเหตุผลที่ทำให้ผู้วิจัยเลือกใช้ตัวแปรเหล่านี้เป็นตัวแปรอิสระในการศึกษา

5) ตัวแปร PAY และ LTV เป็นค่าคงที่ตลอดช่วงการศึกษา ส่วนตัวแปร RTERM มีค่าเปลี่ยนแปลงไปตามแต่ละช่วงเวลา

6) ตัวแปรตาม STATUS เป็นตัวแปรที่แสดงสถานะของการชำระหนี้ของลูกค้า โดยที่ STATUS = 1 ลูกค้าเกิดหนี้ NPL (เป็นลูกค้าที่มีการค้างชำระติดต่อกันตั้งแต่ 3 เดือนขึ้นไป) และ STATUS = 0 ลูกค้าที่ไม่เกิดหนี้ NPL (เป็นลูกค้าที่ไม่มีการค้างชำระเลยหรือค้างชำระติดต่อกันน้อยกว่า 3 เดือน)

7) ทำการวิเคราะห์ผลด้วยโปรแกรม SAS

### 3.3 ขั้นตอนการวิเคราะห์ข้อมูลจริง

1) จัดข้อมูลให้เหมาะสมสำหรับการวิเคราะห์ทั้งสองแบบ พร้อมทั้งวิเคราะห์สถิติเบื้องต้นของข้อมูลตัวอย่าง

2) ทำการวิเคราะห์ความถดถอยโลจิสติกทั้งสองแบบ

3) ทดสอบการพยากรณ์ด้วยตัวแบบประมาณทั้งสอง

4) พล็อตโค้ง ROC และคำนวณพื้นที่ใต้โค้ง ROC

5) เปรียบเทียบประสิทธิภาพจากพื้นที่ใต้โค้ง ROC โดยใช้สมการ (12) ซึ่งเป็นการคำนวณหาขนาดตัวอย่างที่เหมาะสมที่จะทำให้เกิดความแตกต่างอย่างมีนัยสำคัญทางสถิติของพื้นที่ใต้โค้ง ROC ระหว่าง 2 ตัวแบบ [7]

$$n = \left[ \frac{Z_\alpha \sqrt{2V_1} + Z_\beta \sqrt{V_1 + V_2}}{\delta} \right]^2 \quad (12)$$

โดยที่  $\theta_1$  = พื้นที่ใต้โค้ง ROC ของตัวแบบสถิติ

$\theta_2$  = พื้นที่ใต้โค้ง ROC ของตัวแบบพลวัต

$\delta = \theta_2 - \theta_1$  ผลต่างของพื้นที่ทั้งสอง

$V_1 = Q_1 + Q_2 - 2\theta_1^2$  เมื่อ  $Q_1$  และ  $Q_2$  ได้จากการแทนค่า

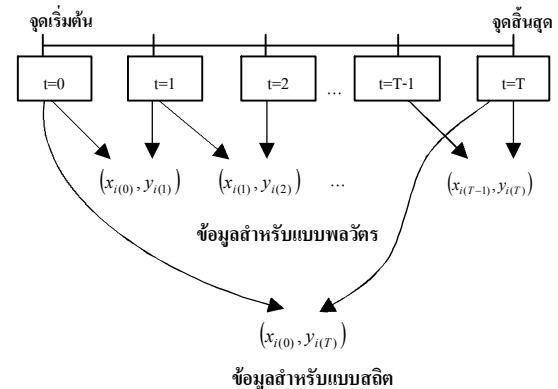
$\theta_1$  ในสมการ (13)

$V_2 = Q_1 + Q_2 - 2\theta_2^2$  เมื่อ  $Q_1$  และ  $Q_2$  ได้จากการแทน

ค่า  $\theta_2$  ในสมการ (13)

$$Q_1 = \frac{\theta}{2 - \theta} \quad , \quad Q_2 = \frac{2\theta^2}{1 + \theta} \quad (13)$$

### 3.4 การจัดข้อมูลสำหรับการวิเคราะห์



รูปที่ 1 การจัดข้อมูลสำหรับการวิเคราะห์ทั้งสองแบบ

### 3.5 การคำนวณ $\hat{p}_i$ สำหรับตัวแบบพลวัต

1) จากตัวแบบประมาณ คำนวณค่า  $P(Y_t = 1) = \hat{p}_i$  ในแต่ละช่วงเวลาที่ใช้ในการทดสอบ

2) คำนวณค่า  $P(Y_t = 0) = 1 - \hat{p}_i$  ในแต่ละช่วงเวลา

3) คำนวณค่า  $P(Y = 0) = \prod_{t=1}^T (1 - \hat{p}_i)$

4) คำนวณค่า  $P(Y = 1) = 1 - P(Y = 0) = \hat{p}_i$  ซึ่งเป็นความน่าจะเป็นที่หน่วยตัวอย่างจะเกิดเหตุการณ์ที่สนใจ

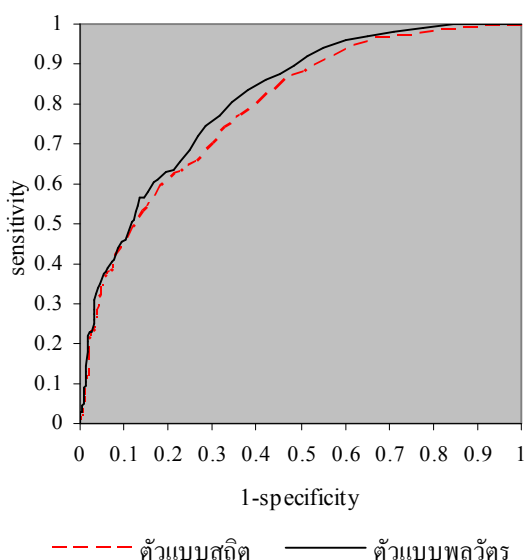
## 4. ผลการวิจัย

แยกเป็นกรณีย่อยตามลักษณะของการทดสอบการพยากรณ์ ได้ผลการวิจัยดังนี้

### 4.1 กรณีข้อมูลจำลอง

4.1.1 กรณีย่อย 1 : ใช้ข้อมูลตัวอย่าง 1 ชุด โดยใช้ข้อมูลในช่วงเวลาที่ 1 - 12 สำหรับการประมาณค่า และพยากรณ์กับข้อมูลในช่วงเวลาที่ 13 - 24

จากการเปรียบเทียบเส้นโค้ง ROC พบว่า เกือบทุกครั้งของการทดลองเส้นโค้ง ROC ของตัวแบบพลวัต อยู่เหนือเส้นโค้ง ROC ของตัวแบบสถิต ดังตัวอย่างการเปรียบเทียบในรูปที่ 2



รูปที่ 2 ตัวอย่างเปรียบเทียบเส้นโค้ง ROC ของตัวแบบสถิตและตัวแบบพลวัตสำหรับกรณีย่อย 1

เมื่อทดสอบสมมติฐานของความแตกต่างระหว่างตัวแบบทั้งสอง สรุปได้ว่า ตัวแบบพลวัตมีพื้นที่ใต้โค้งเฉลี่ยมากกว่าตัวแบบสถิตอย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05 โดยมีพื้นที่ใต้โค้งเฉลี่ยเท่ากับ 0.8035 และตัวแบบสถิตมีพื้นที่ใต้โค้งเฉลี่ยเท่ากับ 0.7901 และมีความแตกต่างของพื้นที่เฉลี่ยเท่ากับ 0.0135 ดังตารางที่ 1 ซึ่งสมมติฐานของการทดสอบ คือ

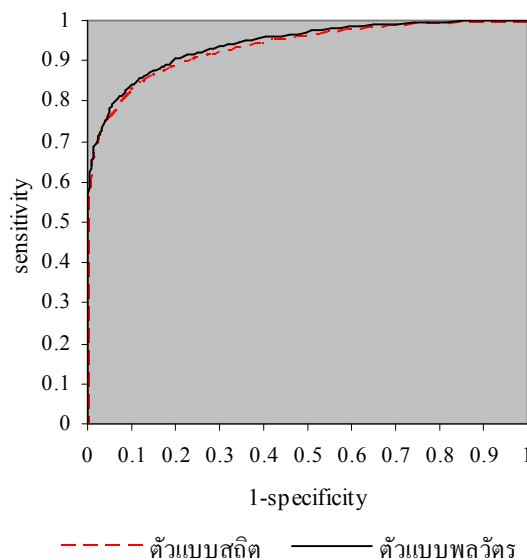
$$\begin{aligned} H_0 : \mu(ROC)_{Dynamic} &\leq \mu(ROC)_{Static} \\ H_1 : \mu(ROC)_{Dynamic} &> \mu(ROC)_{Static} \end{aligned} \quad (14)$$

ตารางที่ 1 พื้นที่ใต้โค้ง ROC และการเปรียบเทียบ

| ตัวแบบ         | พื้นที่, ความแตกต่างของพื้นที่ใต้โค้ง ROC |        |        |       |         |
|----------------|-------------------------------------------|--------|--------|-------|---------|
|                | เฉลี่ย                                    | SD.    | SE.    | t     | p-value |
| สถิต           | 0.7901                                    | 0.0158 | 0.0005 |       |         |
| พลวัต          | 0.8035                                    | 0.0150 | 0.0005 |       |         |
| พลวัต vs. สถิต | 0.0135                                    | 0.0053 | 0.0002 | 81.23 | 0.00    |

**4.1.2 กรณีย่อย 2 :** ใช้ข้อมูลตัวอย่าง 2 ชุดที่เป็นอิสระต่อกัน โดยใช้ชุดที่ 1 ในช่วงเวลาที่ 1 - 12 สำหรับการประมาณค่า และพยากรณ์กับข้อมูลชุดที่ 2 ในช่วงเวลาที่ 1 - 12

จากการเปรียบเทียบเส้นโค้ง ROC พบว่า เกือบทุกครั้งของการทดลองจะได้เส้นโค้ง ROC ของตัวแบบพลวัตอยู่เหนือเส้นโค้ง ROC ของตัวแบบสถิตเล็กน้อย ดังตัวอย่างการเปรียบเทียบในรูปที่ 3



รูปที่ 3 ตัวอย่างเปรียบเทียบเส้นโค้ง ROC ของตัวแบบสถิตและตัวแบบพลวัตสำหรับกรณีย่อย 2

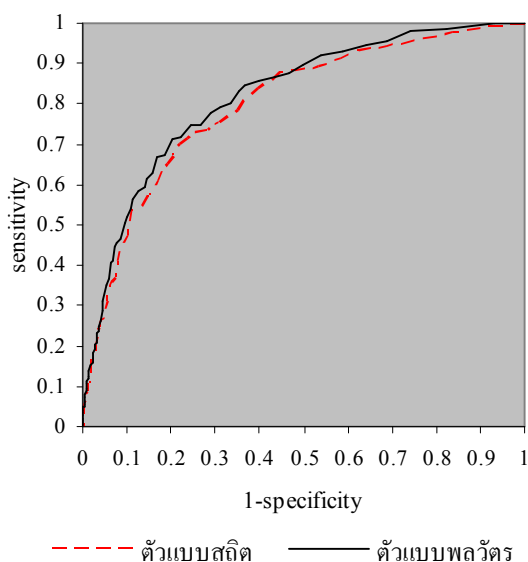
เมื่อทดสอบสมมติฐาน (14) สรุปได้ว่า ตัวแบบพลวัตมีพื้นที่ใต้โค้งเฉลี่ยมากกว่าตัวแบบสถิตอย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05 โดยมีพื้นที่ใต้โค้งเฉลี่ยเท่ากับ 0.9441 และตัวแบบสถิตมีพื้นที่ใต้โค้งเฉลี่ยเท่ากับ 0.9404 และมีความแตกต่างของพื้นที่เฉลี่ยเท่ากับ 0.0037 ดังตารางที่ 2

ตารางที่ 2 พื้นที่ใต้โค้ง ROC และการเปรียบเทียบ

| ตัวแบบ         | พื้นที่, ความแตกต่างของพื้นที่ใต้โค้ง ROC |        |         |        |         |
|----------------|-------------------------------------------|--------|---------|--------|---------|
|                | เฉลี่ย                                    | SD.    | SE.     | t      | p-value |
| สถิต           | 0.9404                                    | 0.0022 | 0.0001  |        |         |
| พลวัต          | 0.9441                                    | 0.0021 | 0.0001  |        |         |
| พลวัต vs. สถิต | 0.0037                                    | 0.0006 | 0.00002 | 208.70 | 0.00    |

**4.1.3 กรณีย่อย 3 :** ใช้ข้อมูลตัวอย่าง 2 ชุดที่เป็นอิสระต่อกัน โดยใช้ข้อมูลชุดที่ 1 ในช่วงเวลาที่ 1 - 12 สำหรับการประมาณค่า และทำการพยากรณ์กับข้อมูลชุดที่ 2 ในช่วงเวลาที่ 13 - 24

จากผลการเปรียบเทียบ พบว่า เกือบทุกครั้งของการทดลอง จะได้เส้นโค้ง ROC ของตัวแบบพลวัตอยู่เหนือเส้นโค้ง ROC ของตัวแบบสถิต ดังตัวอย่างการเปรียบเทียบในรูปที่ 4



**รูปที่ 4** ตัวอย่างเปรียบเทียบเส้นโค้ง ROC ของตัวแบบสถิตและตัวแบบพลวัตสำหรับกรณีย่อย 3

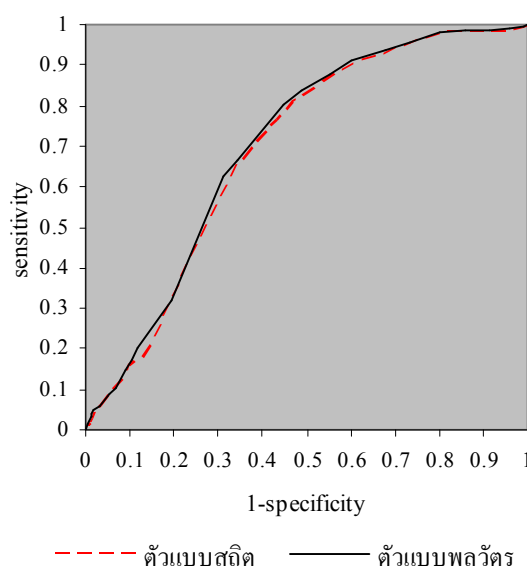
เมื่อทดสอบสมมติฐาน (14) สรุปได้ว่า ตัวแบบพลวัตมีพื้นที่ใต้โค้งเฉลี่ยมากกว่า ตัวแบบสถิต อย่างมีนัยสำคัญทางสถิติที่ระดับ 0.05 โดยมีพื้นที่ใต้โค้งเฉลี่ยเท่ากับ 0.8015 และตัวแบบสถิตมีพื้นที่ใต้โค้งเฉลี่ยเท่ากับ 0.7882 และมีความแตกต่างของพื้นที่เฉลี่ยเท่ากับ 0.0133 ดังตารางที่ 3

**ตารางที่ 3** พื้นที่ใต้โค้ง ROC และการเปรียบเทียบ

| ตัวแบบ         | พื้นที่, ความแตกต่างของพื้นที่ใต้โค้ง ROC |        |        |       |         |
|----------------|-------------------------------------------|--------|--------|-------|---------|
|                | เฉลี่ย                                    | SD.    | SE.    | t     | p-value |
| สถิต           | 0.7882                                    | 0.0168 | 0.0005 |       |         |
| พลวัต          | 0.8015                                    | 0.0160 | 0.0005 |       |         |
| พลวัต vs. สถิต | 0.0133                                    | 0.0055 | 0.0002 | 76.72 | 0.00    |

## 4.2 กรณีข้อมูลจริง

ในการวิเคราะห์จะใช้ข้อมูลตัวอย่าง 2 ชุดที่เป็นอิสระกัน โดยใช้ข้อมูลชุดที่ 1 สำหรับการประมาณค่าพารามิเตอร์และทำการพยากรณ์กับข้อมูลชุดที่ 2 จากผลการวิเคราะห์ พบว่า พื้นที่ใต้โค้ง ROC ของทั้งสองตัวแบบมีค่าใกล้เคียงกัน โดยที่ตัวแบบพลวัตมีพื้นที่ใต้โค้ง ROC มากกว่าตัวแบบสถิตเล็กน้อย แต่จากการทดสอบสมมติฐาน  $H_0: \theta_2 \leq \theta_1$  vs.  $H_1: \theta_2 > \theta_1$  โดยใช้สมการ (12) ที่ระดับนัยสำคัญ 0.05 และอำนาจการทดสอบ 80% เราไม่สามารถสรุปได้ว่า ตัวแบบพลวัตมีพื้นที่ใต้โค้ง ROC สูงกว่าตัวแบบสถิต ดังรูปที่ 5



**รูปที่ 5** เปรียบเทียบเส้นโค้ง ROC ของตัวแบบสถิตและตัวแบบพลวัตสำหรับกรณีข้อมูลจริง

## 5. สรุปผล

จากการศึกษาเพื่อเปรียบเทียบประสิทธิภาพระหว่างตัวแบบถดถอยโลจิสติกแบบสถิตและแบบพลวัต เราพบว่า กรณีข้อมูลจำลอง ตัวแบบพลวัตมีพื้นที่ใต้โค้ง ROC เฉลี่ยมากกว่าตัวแบบสถิตอย่างมีนัยสำคัญทางสถิติที่ระดับนัยสำคัญ 0.05 ดังนั้น สามารถสรุปได้ว่าตัวแบบพลวัตเป็นตัวแบบที่มีประสิทธิภาพในการพยากรณ์ดีกว่าตัวแบบสถิต ส่วนกรณีข้อมูลจริง ถึงแม้ว่าตัวแบบพลวัตจะมีพื้นที่ใต้โค้ง ROC มากกว่าตัวแบบสถิต แต่ไม่มีนัยสำคัญทางสถิติที่ระดับ 0.05

ดังนั้น จึงไม่สามารถสรุปได้ว่าตัวแบบพลวัตมีประสิทธิภาพในการพยากรณ์ดีกว่าตัวแบบสถิต

เนื่องจาก ในการเปรียบเทียบประสิทธิภาพของตัวแบบถดถอยโลจิสติกแบบสถิตและแบบพลวัตสำหรับข้อมูลจริงนั้น เราใช้ตัวแปรอิสระในแต่ละตัวแบบ 3 ตัวแปร แต่มีเพียง 1 ตัวแปร คือ เวลาที่เหลืออยู่ของลูกหนี้ เท่านั้นที่มีค่าผันแปรกับเวลาและมีอัตราผันแปรค่อนข้างน้อย จึงทำให้ข้อมูลที่ใช้ในการวิเคราะห์แบบพลวัตไม่แตกต่างจากข้อมูลที่ใช้วิเคราะห์แบบสถิตที่เป็นเสมือนการใช้ค่าเฉลี่ยของข้อมูล จึงทำให้ไม่สามารถสรุปได้ว่าตัวแบบพลวัตมีประสิทธิภาพในการพยากรณ์ดีกว่าตัวแบบสถิต และเนื่องจากข้อมูลจริงที่ใช้ในการวิเคราะห์เป็นข้อมูลทุติยภูมิที่มีการเก็บข้อมูลโดยมุ่งตามแบบสถิตมาตั้งแต่ต้น การปรับข้อมูลสำหรับการวิเคราะห์แบบพลวัตจึงทำได้ในขอบเขตที่จำกัดและอาจยังไม่เพียงพอที่จะเห็นประสิทธิภาพที่เหนือกว่าของตัวแบบพลวัตเมื่อเทียบกับตัวแบบสถิต ดังนั้น ธุรกิจหรืองานวิจัยใดที่ต้องการจะวิเคราะห์ข้อมูลด้วยวิธีการแบบพลวัตอาจจะต้องมีการวางแผนการเก็บข้อมูลให้เข้ากับการวิเคราะห์แบบพลวัตตั้งแต่เริ่มต้น ซึ่งถ้าข้อมูลของตัวแปรอิสระที่มีค่าผันแปรไปตามเวลาอย่างมาก หรือค่าของตัวแปรอิสระทั้งหมดมีค่าผันแปรไปตามเวลาแล้ว การวิเคราะห์แบบพลวัตควรจะมีความมีประสิทธิภาพที่ดีกว่าการวิเคราะห์แบบสถิตอย่างเห็นได้ชัดเจน

การวิเคราะห์แบบพลวัตเป็นวิธีที่น่าสนใจมาก เพราะมีวิธีการวิเคราะห์ที่ไม่ซับซ้อน มีหลักการประมาณค่าเหมือนแบบปกติที่เคยปฏิบัติมา และยังสามารถใช้โปรแกรมการคำนวณแบบเดิมได้ ซึ่งอาจใช้เวลาในการคำนวณเพิ่มขึ้นบ้าง แต่เนื่องจากอุปกรณ์การคำนวณในปัจจุบันมีประสิทธิภาพสูงขึ้นมากเมื่อเทียบกับอดีต จึงทำให้ผู้วิจัยไม่สังเกตเห็นความแตกต่างด้านเวลาการคำนวณระหว่างวิธีการวิเคราะห์แบบสถิตและแบบพลวัต แต่ทั้งนี้เราอาจจะต้องเสียเวลาในการจัดข้อมูลบ้าง ซึ่งก็ไม่ได้ยุ่งยากจนเกินไป

## เอกสารอ้างอิง

- 1 T. Shumway, "Forecasting Bankruptcy More Accurately: A Simple Hazard Model." *Journal of Business*, Vol.74, January 2001, pp. 101-124.
- 2 S. Tirapat and S. Kiatsupaibul, "Credit value at risk via credit scoring model". *Simulation Society Research Workshop*, 2007, pp. 1-4.
- 3 กัลยา วานิชย์บัญชา. การวิเคราะห์ข้อมูลหลายตัวแปร, โรงพิมพ์บริษัท ธรรมสาร จำกัด กรุงเทพมหานคร, พิมพ์ครั้งที่ 1, 2548.
- 4 นุชymas นันตสุขนซ์. การประมาณค่าฟังก์ชันอัตราภาวะภัยสำหรับข้อมูลสมบูรณ. วิทยานิพนธ์ปริญญาโทบริหารศาสตรบัณฑิต มหาวิทยาลัย, 2545.
- 5 N. M. Kiefer, "Economic Duration Data and Hazard Function." *Journal of Economic Literature*, Vol.26, No.2, June 1988, pp. 646-679.
- 6 D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*. New York, Second Edition, 2000.
- 7 J.A. Hanley, and B.J. McNeil, "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve." *Radiology*, Vol. 143, April 1982, pp. 29-36.